Analysis of genome-wide, cancer-associated mutation datasets in mouse and human

Jenny Mattison

Wellcome Trust Sanger Institute

Trinity College

University of Cambridge

This dissertation is submitted for the degree of Doctor of Philosophy

September 2008

Declaration

This thesis describes work carried out between April 2005 and September 2008 under the supervision of Dr Tim Hubbard and Dr David Adams at the Wellcome Trust Sanger Institute, while member of Trinity College, University of Cambridge. This dissertation is the result of my own work and contains nothing that is the outcome of work done in collaboration with others, except as specified in the text and Acknowledgements. No part of this dissertation nor anything substantially the same has been, or is being, submitted for any qualification at this or any other university. This dissertation does not exceed the page limit set by the Biology Degree Committee.

Jenny Mattison
September 2008

Abstract

The complexity of human cancer genomes complicates the identification of those mutations that drive the tumourigenic process. Integrative analyses, particularly cross-species comparisons, provide a means of distinguishing likely driver mutations from the background of passenger mutations that arise in unstable cancer genomes. This thesis describes the analysis of human and mouse experimental datasets to identify human cancer gene candidates.

In mice, candidate cancer genes can be 'tagged' using insertional mutagens such as retroviruses and transposons. The analysis of more than 1,000 mouse tumours generated by insertional mutagenesis is described. Insertion sites are mapped to the mouse genome and are used to identify candidate cancer genes. The distribution of insertions within and around candidate genes is analysed to predict the likely mechanisms of mutagenesis and, therefore, the possible structure and function of the mutated gene products. Candidates are also characterised by comparison with other human and mouse cancer-associated mutation datasets, and co-operating cancer genes are identified in an attempt to better understand cancer gene pathways.

The mouse insertional mutagenesis results are then compared to genome-wide copy number data for human cancers. The Wellcome Trust Sanger Institute has generated comparative genomic hybridisation (CGH) data for ~700 human cancer cell lines using the Affymetrix 10K SNP array and, more recently, for ~600 human cancer cell lines using the high resolution Affymetrix SNP 6.0 array. Regions of copy number change in human cancers often encompass many genes, and it can be difficult to determine which genes contribute to the cancer phenotype. In this thesis, the human CGH data are processed into regions of copy number change and the mouse candidate cancer genes identified by retroviral insertional mutagenesis are used to narrow down the candidates in amplicons and deletions. The over-representation of mouse candidate oncogenes in regions of copy number gain suggests that a significant proportion of genes contributing to retrovirus-induced tumourigenesis in the mouse are also amplified in, and contribute to the development of, human cancers. Candidate oncogenes and tumour suppressor genes that are recurrently mutated in both human tumours and murine lymphomas are identified as strong candidates for a role in tumourigenesis.

Acknowledgements

First and foremost I would like to thank my supervisors Tim Hubbard and David Adams for giving me the opportunity to carry out this project, and for all their advice, support and encouragement along the way. Many thanks also to my thesis committee, Paul Edwards, Andy Futreal and Julian Parkhill, for their guidance and advice. The project would not have been possible without the work of Anton Berns' lab at the Netherlands Cancer Institute (NKI), who generated the mouse tumours in the retroviral insertional mutagenesis screen, and the Wellcome Trust Sanger Institute Cancer Genome Project (CGP), led by Mike Stratton and Andy Futreal, who generated all of the copy number data used in this thesis. At the NKI, I would like to thank Anthony Uren, Jaap Kool, Jos Jonkers, Lodewyk Wessels, Maarten van Lohuizen and Anton Berns for enabling me to participate in their exciting study, and for providing me with the data that I needed for my project. I am particularly grateful to Anthony Uren and Jaap Kool for helping me to understand the intricacies of insertional mutagenesis, and Jeroen de Ridder for performing the statistical analysis required to identify common insertion sites in the data. Also thanks to Anthony Cox and Daoud Sie for discussions and advice on mapping the insertions. Among the CGP team, I am particularly grateful to Richard Wooster and Adam Butler for giving me access to the copy number data, and Graham Bignell for helping me to understand it. Thanks also to Chris Greenman for his helpful statistical advice, and Helen Davies, who provided the *TP53* and *CDKN2A* mutation statuses for the human cancer cell lines. For the *Sleeping Beauty* data, I extend my gratitude to Lara Collier at the University of Minnesota, who performed the screen. Many thanks also to Fanni Gergely at the Cambridge Research Institute, who carried out the functional validation of *QSK*. I would also like to thank all members of the Hubbard Research Group, past and present, who have opened my eyes to a strange new world and have been a pleasure to work with. Thanks particularly to Matias Piipari for extracting the information I needed from TRANSFAC, and Andreas Prlic for general IT support! On a personal note, I thank my parents and my sister Laura for lavishing affection on me, and my grandma Ethel, whose enthusiasm for science and research is as strong as ever at the age of 96. I would also like to acknowledge the little person-to-be who has made thesis writing something of a challenge, but whom I love already. Finally, I thank my husband Simon for supporting me in my endeavour, and particularly for putting up with me over the last few months. I guess it is my turn to vacuum.

Table of Contents

List of Figures

List of Tables

Appendices

Abbreviations

ALL	acute lymphoblastic leukaemia
AML	acute myeloid leukaemia
API	application programming interface
BAC	bacterial artificial chromosome
CGH	comparative genomic hybridisation
ChIP	chromatin immunoprecipitation
CIS	common insertion site
CML	chronic myelogenous leukaemia
CNV	copy number variation
COSMIC	Catalogue of Somatic Mutations in Cancer
CRUK	Cancer Research UK
DAS	distributed annotation system
ES	embryonic stem
EST	expressed sequence tag
ESP	end-sequence profiling
HGNC	HUGO Gene Nomenclature Committee
HMM	hidden Markov model
IARC	International Agency for Research on Cancer
IR/DR	inverted repeat/direct repeat
KC	kernel convolution
LINE	long interspersed nuclear element
LOH	loss of heterozygosity
LTR	long terminal repeat
MC	Monte Carlo
MCC	Matthew’s Correlation Coefficient
MCR	minimal common region of amplification or deletion
MGI	Mouse Genome Informatics
MMTV	mouse mammary tumour virus
MuLV	murine leukaemia virus
NCBI	National Center for Biotechnology Information
NKI	Netherlands Cancer Institute
PCR	polymerase chain reaction
PET	paired-end ditag sequencing
RTCGD	Retroviral Tagged Cancer Gene Database
SB	*Sleeping Beauty*
SNP	single nucleotide polymorphism
SINE	short interspersed nuclear element
TFBS	transcription factor binding site
UTR	untranslated region
VISA	viral insertion site amplification
WGSA	whole-genome sampling assay
WHO	World Health Organization
WTSI	Wellcome Trust Sanger Institute

Chapter 1 Introduction

1.1 Outline of introduction

This introduction presents the foundations of the work described in this thesis. Section 1.2 focuses on the importance of studying the genetic basis of cancer, beginning with an overview of the burden of cancer. This is followed by a synopsis of the major contributions to the current understanding of how cancer develops, and a description of the main classes of genes and some of the genetic pathways known to be involved in cancer development. The section concludes with a discussion of the contribution of cancer genetics to the development of drugs for cancer treatment. Section 1.3 discusses the use of genome-wide approaches in the identification of cancer genes in humans. Prior work on the analysis of mutations, gene expression and epigenetics in cancer genomes is outlined, and research into the analysis of copy number changes is described in greater detail. Methods to identify transcription factor binding sites, and therefore to elucidate regulatory pathways, are also discussed. Section 1.4 describes the role of the mouse in cancer research and focuses on the use of retroviral and transposon-mediated mutagenesis in the genome-wide discovery of novel cancer genes and collaborations between genes involved in cancer. A significant portion of the work presented in this thesis relates to the comparison of human and mouse datasets for cancer gene identification, and previous studies of this kind are discussed in Section 1.5. Finally, the aims and rationale of this thesis are presented in Section 1.6.

1.2 An introduction to cancer

1.2.1 Definition and classification

Cancer is a class of diseases manifesting as uncontrolled cell division that leads to invasion of surrounding tissues and spread to distant sites (metastasis). These malignant properties of cancers differentiate them from benign tumours, in which abnormal cell proliferation is usually confined locally. Most cancers are classified according to the tissue of origin. There are over 100 distinct types, and 4 broad categories: carcinoma, arising in epithelial cells; sarcoma, arising in connective or supportive tissue and soft tissue; leukaemia, arising in blood-forming tissues; and lymphoma, arising in cells of the immune system. See Pelengaris and Khan (2006).

1.2.2 Epidemiology

Cancer is a leading cause of death worldwide, accounting for 13% of all deaths in 2005 (WHO, 2008). In developed countries, it is the second greatest cause of death after cardiovascular disease, while in less developed countries, it is the third greatest after infectious and cardiovascular diseases. In 2002, 24% of all deaths in the UK were caused by cancer, compared with 12% in Asia and just 4% in Africa (CRUK, 2008; Ferlay *et al.*, 2004). Economic growth in Asia is expected to cause a rise in the proportion of deaths from cancer, and yet, due to its population size, more than half of all deaths from cancer already occur in Asia (Ferlay *et al.*, 2004). The global population is growing and ageing and, as cancer is predominantly a disease of older people (CRUK, 2008), the number of cancer deaths is expected to increase by 45% between 2007 and 2030 (WHO, 2008).

More than a quarter of a million new cases of cancer are diagnosed each year in the UK, and the four most common cancers - breast, lung, colorectal and prostate - account for half of these. In 2004, the most common cancers in men and women were breast and prostate, respectively. However, in both sexes, lung cancer was the biggest killer, accounting for 22% of all cancer deaths in 2005 (CRUK, 2008; Figure 1.1).

It is estimated that around 35% of all deaths from cancer are preventable, and 9 main modifiable risk factors have been identified (Danaei *et al.*, 2005). The leading risk factor is smoking, which is thought to contribute to 21% of all preventable cancers. Others include alcohol use, diet, and physical inactivity. Environmental risk factors account for much of the striking geographical variation in the incidence of certain cancers, and migration studies indicate that reducing exposure to these factors could eliminate a high proportion of deaths from cancer. There is, for example, a heightened risk of developing stomach cancer in Japan (Parkin *et al.*, 2005), where risk factors include infection by *Helicobacter Pylori* (IARC, 1994) and a diet rich in salted foods (Tsugane, 2005). However, within one generation of settling in Hawaii, the incidence of stomach cancer among Japanese immigrants declines to levels comparable with the surrounding population (Peto, 2001).

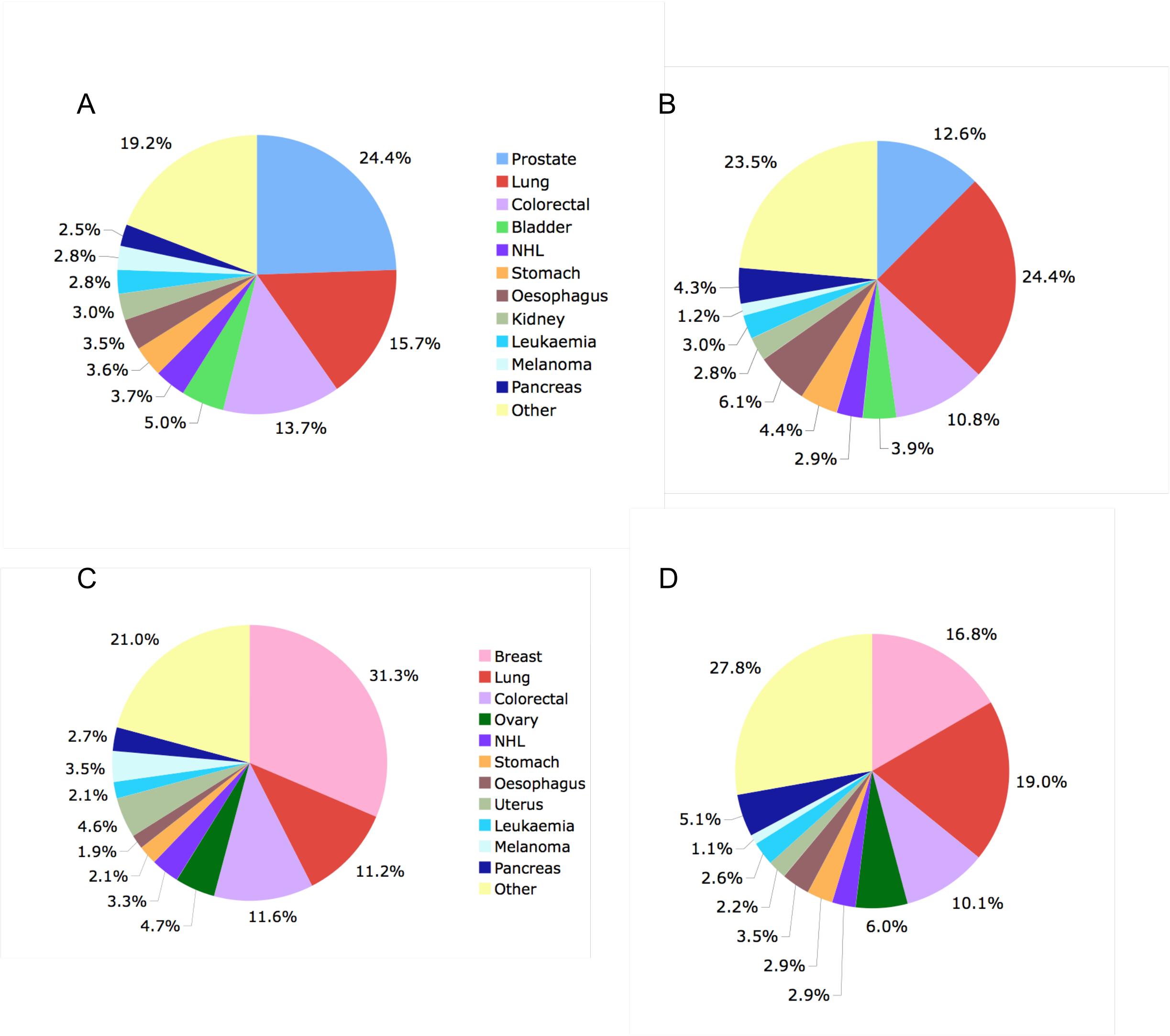

Figure 1.1. Summary of cancer incidence in 2004 and deaths from cancer in 2005 for the most common sites of cancer in males and females in the UK. Cancer incidence and mortality among males are shown in Figures A and B, respectively. Cancer incidence and mortality among females are shown in Figures C and D, respectively. The statistics for this figure were obtained from the Cancer Research UK CancerStats resource (CRUK, 2008).

While prevention could significantly reduce the burden of cancer, improvements in early diagnosis and treatment are also essential. Screening procedures that have reduced cancer mortality rates include the identification and removal of polyps in the colon (Weir *et al.*, 2003) and pre-cancerous cells in the uterus (Misra *et al.*, 1998), and widespread mammography screening for breast cancer (Shapiro, 1997). However, effective screening has been developed for only a handful of cancers, and advances in cancer treatment have been slower than for other chronic diseases, such as cardiovascular disease (Danaei *et al.*, 2005). A greater understanding of the genetic basis of cancers is essential for the development of effective treatments and diagnostic techniques.

1.2.3 The multi-stage theory of carcinogenesis

1.2.3.1 The somatic mutation theory

The theory that cancer is caused by somatic mutation can be traced back to Boveri (1926, 1914), who, extending the views of Hansemann (1890) and through his own work on aneuploidy in cancer cells, postulated that tumours originate from a single cell that has acquired chromosomal abnormalities. 35 years later, the multistage theory of carcinogenesis was borne, first postulated as two-stage carcinogenesis, in which an initiator and a promoter agent were proposed to be required for malignancy (Berenblum and Shubik, 1949), and later in the Armitage-Doll model, which suggested that six or seven independent, sequential, events were required (Armitage and Doll, 1954). Nowell (1976) proposed a model of clonal evolution, in which tumours evolve from a single cell through a series of stepwise genetic alterations within the original clone. He postulated that as the tumour progresses, genetically variant sublines emerge and the most favourable sublines, i.e. those with the greatest growth advantage, are selected (Figure 1.2).

An alternative theory for carcinogenesis, the tissue organisation field theory, proposes that rather than a cell acquiring the ability to proliferate uncontrollably through mutation, proliferation is in fact the default state of cells and cancer is caused by disruption to interactions between cells and tissues (Soto and Sonnenschein, 2004). There is, however, overwhelming support in favour of the somatic mutation theory for most cancer types.

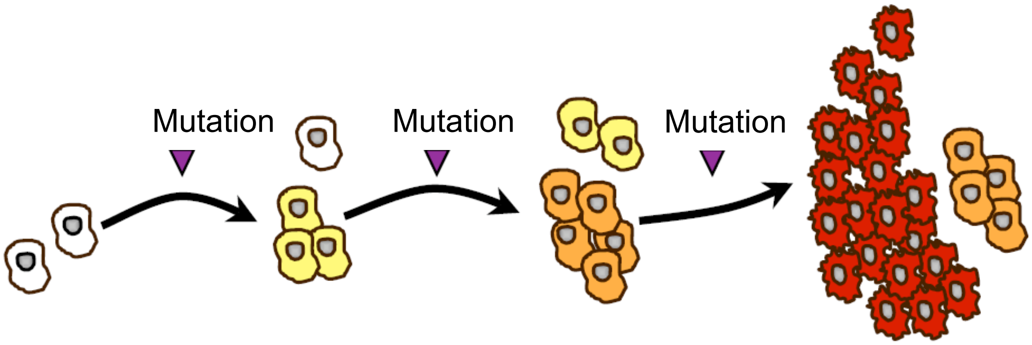

Figure 1.2. The clonal evolution of cancer. Tumours evolve from a single cell through a series of stepwise genetic changes within the original clone. Cells containing mutations that confer the greatest growth advantage are selected and become the dominant clone. Adapted from a figure supplied by D.J. Adams.

1.2.3.2 The cancer stem cell hypothesis

In the original model of clonal evolution, events in all tumour cells can participate in the evolution of the tumour. However, the cancer stem cell hypothesis proposes that only cells that are capable of self-renewal, i.e. stem cells, contribute to tumour evolution and that these give rise to most of the cells with a more differentiated phenotype (for review, see Shipitsin and Polyak, 2008). The theory has some inconsistencies, but it is clear that putative cancer stem cells exist in most, if not all, cancer types, and xenotransplant assays have shown that stem cell-like tumour cells have a significantly higher potential to form tumours in irradiated NOD-SCID mice than do other cells from the same human tumour (Shipitsin and Polyak, 2008). Compared with well-differentiated tumours, poorly differentiated tumours overexpress genes that are normally enriched in embryonic stem (ES) cells (Ben-Porath *et al.*, 2008). These genes include the transcriptional targets of NANOG, OCT4 and SOX2, which are key regulators of pluripotency and self-renewal in ES cells (see Loh *et al.*, 2006). Wong *et al.* (2008) constructed a "module map" of stem cell genes, and showed that a subset of adult tissue stem cells shares a core gene expression program with ES cells, and that the ES cell-like program is frequently activated in human epithelial cancers. Other recent research has shown that the epithelial-mesenchymal transition, which is often activated in tumour metastasis, is linked to the acquisition of epithelial stem cell-like properties (Mani *et al.*, 2008).

1.2.3.3 Rate-limiting events in tumourigenesis

While it is widely accepted that cancer is caused by stepwise mutations, there are conflicting theories about how these mutations arise. The Armitage-Doll model suggests that mutations arise gradually over time, and that the number of rate-limiting events required for carcinogenesis can be inferred from the age-specific incidence of cancer and the rate of successive mutations in cells (Armitage and Doll, 1954). Cancers will not fit the model if the mutation rate is not constant, e.g. in smokers, where the mutation rate increases at the onset of smoking, or if the incidence does not increase with age, e.g. in childhood cancers (for review, see Knudson, 2001). However, the estimate of 5 to 7 mutations in colorectal cancer is compatible with the genetic model for colorectal tumourigenesis, in which at least four or five genes were proposed to be required for malignancy (Ashley, 1969; Fearon and Vogelstein, 1990).

More recent research suggests that a single rate-limiting step may be required for epithelial carcinogenesis, and that telomere crisis is one of the processes responsible for this step (Frieboes and Brody, 2005). The telomere crisis hypothesis proposes that mutations occur suddenly in cells with telomere dysfunction (Chin *et al.*, 2004; Maser and DePinho, 2002). In cells without active telomerase, telomeres erode and eventually cease to function. At this point, cells show massive genomic instability, including end-to-end fusions, non-reciprocal translocations, amplifications and deletions (Artandi *et al.*, 2000; O'Hagan *et al.*, 2002). This results in rapid cell senescence but some cells may escape by reactivating telomerase, and further mutations accumulate, leading to tumour progression (Maser and DePinho, 2002). Genomic instability is discussed in further detail in Sections 1.2.5.1.3 and 1.3.

1.2.4 The hallmarks of cancer

Hanahan and Weinberg (2000) proposed that all genetic alterations in cancer can be represented by six essential changes in cell physiology. These are “self-sufficiency in growth signals, insensitivity to antigrowth signals, evading apoptosis, limitless replicative potential, sustained angiogenesis, and tissue invasion and metastasis”. The authors suggest that all tumours must acquire the same six capabilities, but that different genes may be mutated, and in a different order, even within cancers of the same type. The review by Hanahan and Weinberg is considered a seminal work, and the six “hallmarks of cancer” appear to be shared by most, if not all, malignancies.

1.2.5 Cancer genes

1.2.5.1 Classification

The term “cancer gene” will be used throughout this thesis to describe a gene for which mutations have been causally implicated in cancer. Cancer genes are often divided into 3 classes known as oncogenes, tumour suppressor genes and caretaker genes.

1.2.5.1.1 Oncogenes

In general, oncogenes play a role in accelerating cell growth and proliferation, but they may also contribute to loss of differentiation, avoidance of apoptosis, cell motility and invasion (see Pelengaris and Khan, 2006). The normal counterparts of oncogenes, known as proto-oncogenes, mainly encode growth factors, growth factor receptors, signal transducers, transcription factors and regulators of cell death. Proto-oncogenes may become oncogenes through increased protein activity resulting from intragenic mutations that affect critical residues; increased protein concentration resulting from gene amplification, misregulation of gene expression or an increase in protein stability; or chromosomal translocations that increase inappropriate gene expression or produce a constitutively active fusion protein (Vogelstein and Kinzler, 2004). Oncogenes are dominant at the cellular level. One of the best known oncogenes is *MYC*, which appears to be activated in most human cancers at some stage during their development (see Pelengaris and Khan, 2006).

1.2.5.1.2 Tumour suppressor genes

In contrast to oncogenes, tumour suppressor genes act to limit the growth of tumours and inactivating mutations in these genes can lead to tumour development. Tumour suppressors inhibit cell proliferation by inducing growth arrest or apoptosis in response to DNA damage or hyperproliferative signals induced by oncogenes (see Pelengaris and Khan, 2006). They may be inactivated by missense mutations that alter sites required for protein activity; nonsense mutations that result in an inactive truncated protein; intragenic deletions and insertions; or epigenetic silencing (Vogelstein and Kinzler, 2004). Most tumour suppressor genes follow Knudson's "two-hit hypothesis", which proposes that both copies of the gene must be inactivated to confer a selective growth advantage on the cell (Knudson, 1971). Knudson applied his hypothesis to the identification of the first tumour suppressor gene, *RB1*. Compared with sporadic retinoblastoma, the hereditary form of this rare eye cancer arises earlier and is more often bilateral because cells already harbour one germline *RB1* mutation and require only one additional somatic "hit" (Knudson, 1971). Some tumour suppressor genes are haploinsufficient, i.e. the loss of only one allele is required to confer a growth advantage. Haploinsufficiency of *PTEN* is sufficient for prostate cancer development, but progression is faster when both copies are inactivated (Trotman *et al.*, 2003).

1.2.5.1.3 Caretaker genes

Caretakers maintain DNA integrity and their inactivation results in an increased tendency to acquire mutations in other genes, including oncogenes and tumour suppressor genes (Vogelstein and Kinzler, 2004). Mutations in genes involved in repairing subtle mistakes during replication can cause microsatellite instability, which manifests as alterations in the length of short (1-4 bp) repetitive sequences called microsatellites (Loeb *et al.*, 2003). Cells with microsatellite instability are particularly prone to mutation in the *TGFBR2* tumour suppressor gene, and this is a common mechanism of disease in hereditary nonpolyposis colorectal cancer (HNPCC), in which patients have a germline mutation and a second, somatic, mutation in a mismatch repair gene, most often *MSH2* or *MLH1* (reviewed in Knudson, 2001). Much more common than microsatellite instability is chromosomal instability, which is caused by mutations in genes that are involved in large-scale processes such as recombination and double-strand repair (Lengauer *et al.*, 1998; Loeb *et al.*, 2003). Chromosomal instability is characterised by gross chromosomal alterations, such as duplication or deletion of entire chromosomes (aneuploidy) or parts of chromosomes, and chromosomal rearrangement. Microsatellite and chromosomal instability are collectively known as genomic instability.

1.2.5.1.4 Genes with dual roles in cancer

The terms oncogene and tumour suppressor gene will be used to characterise genes described in this thesis. However, it should be noted that these terms are somewhat simplistic as the role of a protein may be dependent on the cellular context. Some mitogenic proteins have an intrinsic tumour suppressor activity such that inappropriate activation of the protein results in apoptosis of the mutated cell (Cobleigh *et al.*, 1999). Activation of *Myc* in the pancreatic β cells of transgenic mice induces β cell proliferation but also induces apoptosis, which rapidly overwhelms the cell mass (Pelengaris *et al.*, 2002). Likewise, the NOTCH1 receptor plays both oncogenic and tumour suppressive roles that reflect the pleiotropic effects of NOTCH1 signalling in different tissues (for review, see Radtke and Raj, 2003). NOTCH1 signalling is essential for maintaining haematopoietic stem cells and for committing haematopoietic progenitors to the T-cell lineage (Radtke *et al.*, 1999). Aberrant *NOTCH1* expression contributes to over 50% of cases of human T-cell acute lymphoblastic leukaemia (Weng *et al.*, 2004). The involvement of *NOTCH1* was established through the discovery of a translocation between chromosomes 7 and 9 that brings the dominant active cytoplasmic domain under the control of the *TCRβ* locus (Ellisen *et al.*, 1991), but point mutations and deletions are also implicated (Weng *et al.*, 2004). In mice, Notch1 induces lymphomas by suppressing p53 (Beverly *et al.*, 2005). However, Notch1 also functions as a tumour suppressor in mouse skin, where it participates in terminal differentiation by inducing *Waf1* and repressing Shh and Wnt signalling (Radtke and Raj, 2003).

1.2.5.2 Cancer Gene Census

In 2004, a census of genes in which mutations have been causally implicated in human cancer was compiled from the literature (Futreal *et al.*, 2004). It lists genes that are mutated by insertions, deletions or base substitutions in the coding region or splice sites, or by chromosomal translocations or copy number changes. Stringent criteria were applied to exclude genes in which reported mutations could be "passenger" mutations that do not confer any growth advantage.

The census indicates that mutations in more than 1% of human genes are implicated in cancer. Of the 291 genes listed in the original census, 90% have somatic mutations in cancer, 20% have germline mutations, and 10% have both. Chromosomal translocations are the most common class of somatic mutation in human cancer and almost all are dominant at the cellular level. Excluding translocations, there are equal numbers of recessive and dominant somatic mutations within the census list. The protein kinase domain is the most common domain encoded by genes in the census. Domains in proteins involved in transcriptional regulation and DNA maintenance and repair are also over-represented. See Futreal *et al.* (2004).

The Cancer Gene Census is frequently updated and the working list can be downloaded from http://www.sanger.ac.uk/genetics/CGP/Census/. It represents a valuable source of "known" cancer genes that will be utilised in this thesis.

1.2.6 Pathways in cancer

It is often more sensible to focus on the pathways that have been disrupted in cancer, rather than on individual genes. The p53 and RB1 pathways are thought to be inactivated in most, if not all, cancers. However, while *TP53*, which encodes p53, and *RB1* are often mutated, the same effect can be achieved by mutating a different gene in the pathway (see Vogelstein and Kinzler, 2004 and Figure 1.3).

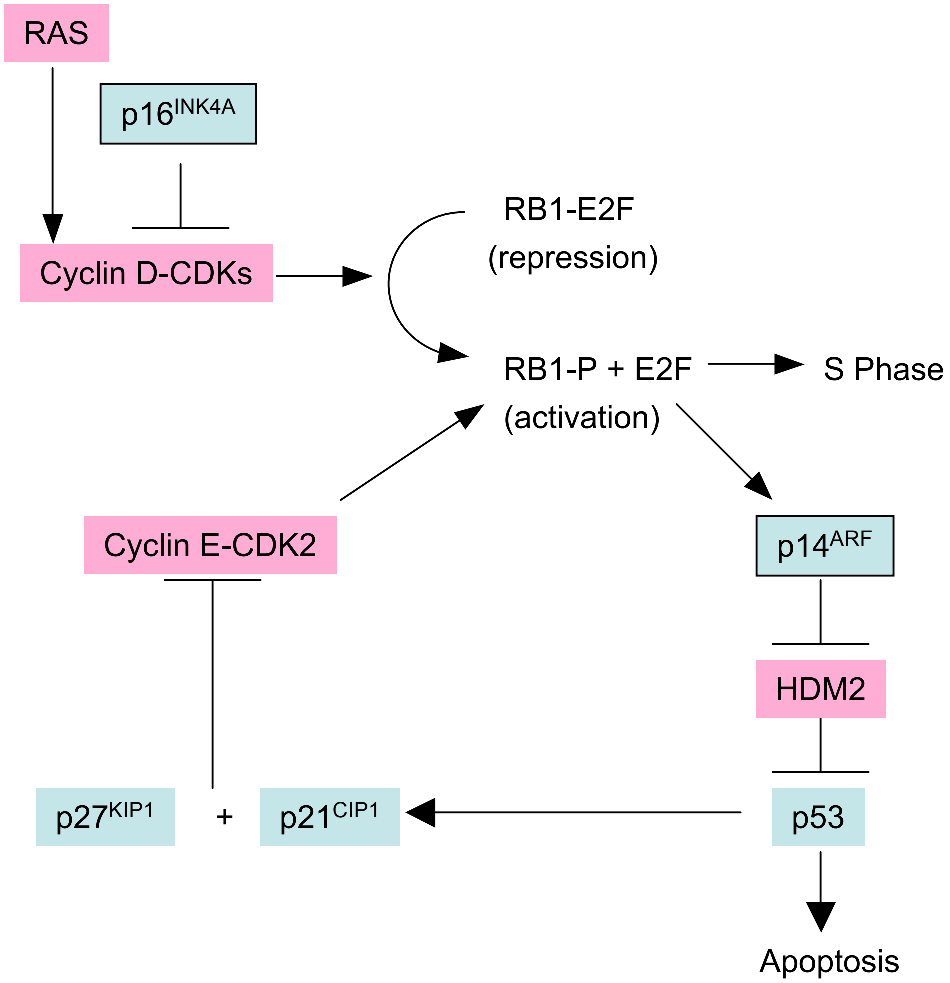

Figure 1.3. Mutations in different genes in the same pathway can have an equivalent effect. The figure shows a simple representation of the p53 and RB1 genetic pathways. The pathways are coupled through the *INK4A/ARF* locus, which encodes p16^{INK4A} and p14^{ARF}, shown here in black boxes, and through p21^{CIP1}, which is activated by p53 and inhibits Cyclin E-CDK2 complexes in the RB1 pathway. Genes that are frequently inactivated in cancer are shown in blue; genes that are frequently activated in cancer are shown in pink. Adapted from Figure 1 in Lowe & Sherr (2003).

p53 is a transcription factor that inhibits cell growth and induces apoptosis in response to cellular stress, such as DNA damage or hyperproliferative signals induced by oncogenes (for review, see Vogelstein *et al.*, 2000). Many p53-responsive genes are involved in arresting cell proliferation at the G1/S and G2/M cell cycle transitions so that cells with DNA damage can be repaired before proceeding to DNA replication or mitosis (Vogelstein *et al.*, 2000). p53 is inhibited by the binding of HDM2 (known as Mdm2 in the mouse) to its N-terminal transactivation domain (Momand *et al.*, 2000). HDM2 also acts as an E3 ubiquitin ligase that targets itself and p53 for degradation by the ubiquitin-dependent proteasome pathway (Momand *et al.*, 2000). Overexpression of *HDM2* may have an equivalent effect to underexpression of *TP53*, and amplification of *HDM2* has been observed in a variety of tumours, including breast, lung and gastric cancers (Gunther *et al.*, 2000; Marchetti *et al.*, 1995a; Marchetti *et al.*, 1995b).

The RB1 pathway regulates cell proliferation by repressing the transcription of genes required for progression through the G1 phase of the cell cycle and for entry into S phase (Figure 1.3 and for review, see Weinberg, 1990). In mid-G1 phase, mitogenic signals from the RAS/MAP kinase pathway activate transcription of D-type cyclins, which bind to the cyclin-dependent kinases CDK4 and CDK6 and initiate phosphorylation of RB1. This results in the release of E2F transcription factors and their subunit partners, DP, from complexes with RB1, and the E2Fs activate transcription of genes required for cell cycle progression. Cyclin E-CDK2 complexes complete the phosphorylation of RB1. A further level of regulation is provided by cyclin-dependent kinase inhibitory (CDKI) proteins, which consist of the INK4 and CIP/KIP protein families. The INK4 proteins ($p16^{INK4A}$, $p15^{INK4B}$, $p18^{INK4C}$ and $p19^{INK4D}$) inhibit CDKs, whereas the CIP/KIP proteins ($p27^{KIP1}$ and $p21^{CIP1}$) stimulate assembly of the cyclin D-CDK4-6 complexes and inhibit cyclin E-CDK2 (for review, see Sherr, 2001). Inactivating $p16^{INK4A}$, $p18^{INK4c}$, $p21^{CIP1}$ or $p27^{KIP1}$ has a similar effect to inactivating *RB1* (Sherr, 2001; Vogelstein and Kinzler, 2004). $p16^{INK4A}$ is inactivated by homozygous deletion, promoter methylation or, to a lesser extent, point mutation, in a large number of tumours (for review, see Liggett and Sidransky, 1998). Likewise, activation of *CDK4* and *cyclin D1* has an equivalent effect on the RB1 pathway, and these oncogenes are frequently amplified and overexpressed in cancer (Vogelstein and Kinzler, 2004).

Cancer pathways are not standalone entities. As well as regulating the RB1 pathway, $p21^{CIP1}$ is one of the major transcriptional targets of p53 (Vogelstein *et al.*, 2000). In addition, the p53 and RB1 pathways are coupled through the *INK4A/ARF* (or *CDKN2A*) locus, which uses alternative reading frames to encode two tumour suppressors: p16$^{\text{INK4A}}$, described above, and p14$^{\text{ARF}}$ (also known as p19$^{\text{ARF}}$ or ARF), which activates p53 by sequestering HDM2 (Quelle *et al.*, 1995; Sherr, 2001; Figure 1.3). The *INK4A/ARF* locus is frequently mutated in human cancer but mutations in *TP53* and *INK4A/ARF* are often mutually exclusive, e.g. in human glioblastoma (Fulci *et al.*, 2000). This suggests that inactivating both loci may not provide any additional growth advantage. However, expression and genotypic analysis of *Trp53*, *Arf* and *Mdm2* in Myc-induced murine lymphomas showed that *Mdm2* was overexpressed in a significant proportion of *Arf*-deficient tumours, while loss of both *Arf* and *Trp53* in primary pre-B cells results in a greater growth advantage than the loss of one gene alone (Eischen *et al.*, 1999).

1.2.7 Treatment of cancer

The main forms of cancer treatment, often used in combination, are surgery, radiotherapy and chemotherapy. Some cancers respond well to these treatments, e.g. testicular cancer has a high cure rate following chemotherapy, but others, such as lung cancer, show a much lower response (CRUK, 2008). Radiotherapy and chemotherapy can have considerable side effects as neither specifically targets cancer cells.

A greater understanding of the genetic basis of cancer has initiated the development of more effective therapies that specifically target deregulated gene expression and signalling pathways in cancer cells. Gleevec (imatinib) targets the BCR-ABL oncoprotein, which causes 95% of cases of chronic myelogenous leukaemia (CML) and ~20% of cases of acute lymphoblastic leukaemia (ALL) (Deininger and Druker, 2003; Faderl *et al.*, 1999). Gleevec stabilises a catalytically inactive form of BCR-ABL (Nagar *et al.*, 2002). It also inhibits four other tyrosine kinases (KIT, PDGFRA, PDGFRB and ARG) but shows minimal side effects (Buchdunger *et al.*, 1996; Druker *et al.*, 1996; Okuda *et al.*, 2001). Treatment has an 89% response rate in chronic CML after 5 years (Druker *et al.*, 2006), and an initial, but not durable, response rate of 52% in patients who have progressed to blast crisis, the terminal phase of the disease (Sawyers *et al.*, 2002). Gleevec has also proved effective in the treatment of gastrointestinal stromal tumours (GISTs) by targeting KIT (Joensuu *et al.*, 2001; van Oosterom *et al.*, 2001) and PDGFRA (Apperley *et al.*, 2002). Other tyrosine kinase inhibitors include Herceptin (trastuzumab), which targets the HER2/ERBB2 receptor in breast cancer (Cobleigh *et al.*, 1999), and

Iressa (gefitinib), which targets the epidermal growth factor receptor (EGFR) in lung adenocarcinomas and non-small cell lung cancers (Fukuoka *et al.*, 2003).

As with traditional therapies, there is evidence that cancer cells can develop resistance to targeted therapies (Balak *et al.*, 2006; Engelman *et al.*, 2007; Gorre *et al.*, 2001; Kobayashi *et al.*, 2005; Nagata *et al.*, 2004; Shattuck *et al.*, 2008), necessitating the development of new drugs for targeted combination therapy (Baselga, 2006). However, the results outlined above demonstrate that targeting a single, critical gene in a complex tumour can elicit a dramatic response. Success of such a treatment depends on the targeted kinase being required for growth and survival of the tumour throughout its evolution (a notion known as "oncogene addiction" (Weinstein, 2002)). The mutation status of other genes can also influence drug response. For example, breast tumours that harbour an amplification of *HER2/ERBB2* are less responsive to trastuzumab if they also harbour an oncogenic *PIK3CA* mutation or have low *PTEN* expression (Berns *et al.*, 2007). Likewise, lung cancers that contain *KRAS* mutations are resistant to treatment with EGFR inhibitors because KRAS acts further downstream in the EGFR pathway (Pao *et al.*, 2005). Due to huge variation in the genetic basis of different cancers, each targeted therapy will be effective against only a subset of cancers. This necessitates the identification of many different drug targets, and fundamentally relies on the identification and characterisation of mutated genes in cancer.

1.3 Genome-wide approaches for human cancer gene discovery

The elucidation of the human genome sequence and developments in high-throughput techniques for genome-wide analysis have allowed for profiling of entire cancer genomes. This section discusses the large-scale technologies that are available for detecting alterations and, ultimately, for identifying cancer genes in human cancer genomes.

1.3.1 Gene resequencing

Advances in DNA sequencing technology have enabled the identification of recurrent intragenic mutations across multiple cancer genomes. Davies and colleagues (2002) screened the coding sequence and intron-exon junctions of *BRAF* for mutations in more than 900 human cancer cell lines and primary tumours, and found somatic missense mutations in 66% of malignant melanomas and in a smaller proportion of many other human cancers. 80% of *BRAF*-mutated melanomas were found to contain a V599E substitution, which is thought to constitutively activate the kinase by mimicking phosphorylation (Davies *et al.*, 2002). An inhibitor has recently been developed that selectively targets the V599E gene product, and so selectively targets BRAF in tumour cells (Tsai *et al.*, 2008).

As the cost of sequencing has diminished, it has become possible to perform larger scale screens to look for mutations in multiple genes across multiple tumours. The first systematic mutational study of a complete gene family was performed by Bardelli and coworkers (2003), who identified 7 candidate cancer genes in a screen of the tyrosine kinase gene family in 182 colorectal cancers. A further study of mutations in the tyrosine phosphatase gene family identified 6 putative tumour suppressor genes that were mutated in 26% of the colorectal cancers analysed (Wang *et al.*, 2004). Resequencing of the phosphatidylinositol 3-kinase (PI3K) gene family revealed one member, *PIK3CA*, that is frequently mutated in tumours of the colon, breast, brain and lung, with most mutations clustering within the helical or catalytic domain (Samuels and Velculescu, 2004). Mutations have since been identified in additional tumour types, such as hepatocellular carcinomas (Bachman *et al.*, 2004) and ovarian cancers (Campbell *et al.*, 2004; Levine *et al.*, 2005). A screen of serine/threonine kinases showed that 40% of colorectal tumours harbour a mutation in 1 of 8 PI3K-pathway genes (Parsons *et al.*, 2005). The PI3K pathway regulates a wide range of cellular functions that are important in cancer, including growth, proliferation, survival, angiogenesis and migration (Brugge *et al.*, 2007).

Studies at the Wellcome Trust Sanger Institute have centred around the resequencing of coding regions from all 518 genes of the protein kinase family. A study of 25 breast cancers revealed diverse patterns of mutation, with variation in the number of mutations and in the identity of mutated genes, such that no commonly point-mutated kinase gene was identified (Stephens *et al.*, 2005). A study of 33 lung cancers reached similar conclusions (Davies *et al.*, 2005). While both studies showed an over-representation of nonsynonymous substitutions, as predicted for “driver” mutations that confer a selective growth advantage on the cell, most of the mutations are likely to be “passenger” mutations that do not contribute to tumourigenesis. Protein kinase resequencing at the Sanger Institute has culminated in the identification of 921 base substitution somatic mutations in 210 diverse human cancers (Greenman *et al.*, 2007). Putative driver mutations were identified in 119 genes but 83% of mutations were predicted to be passengers. Cancers showed variation in mutation prevalence, with many of the cancer types with highest prevalence originating from high turnover, surface epithelia that are most exposed to mutagens (Greenman *et al.*, 2007). Cancers also showed different "mutational signatures", which often reflect differences in mutagenic exposure. For example, most lung cancers have a high proportion of C:G > A:T transversions, which are caused by exposure to tobacco carcinogens (Davies *et al.*, 2005).

The first study to approach the scale of a genome-wide screen involved resequencing the coding regions of all (~13,000) consensus coding sequence (CCDS) genes in 11 breast and 11 colorectal cancers (Sjoblom *et al.*, 2006). Each cancer was found to harbour an average of 93 mutated genes, of which at least 11 (189 candidates in total) were thought to be driver mutations. Many of the functional groups and pathways enriched for candidate cancer genes were unique to one or other cancer type, suggesting differences in the tumourigenic process in breast and colorectal cancers (Lin *et al.*, 2007). There have been claims that the statistical analysis performed in this screen was flawed, in part because they used a different dataset to estimate background mutation rates, which can vary between and within cancer genomes, and because the sample size was small (Getz *et al.*, 2007). However, the findings of this study are in agreement with those of Greenman *et al.* (2007) in suggesting that the genomic landscape of human cancers is more complex than previously thought (Kaiser, 2006). The study has since been expanded to include all of the human RefSeq (Pruitt *et al.*, 2007) genes and a larger number of breast and colorectal cancers (Wood *et al.*, 2007). Each tumour contained an average of 15 potential driver mutations and most of these were in genes that were mutated in fewer than 5% of tumours, therefore recapitulating the conclusions of the previous studies.

Although statistical methods can provide a prediction of the likely driver and passenger mutations within a cancer, there is a strong rationale for using functional assays to test the predictions. Frohling and coworkers (2007) resequenced the coding exons and splice junctions of the receptor tyrosine kinase *FLT3* in samples from patients with acute myeloid leukaemia (AML). They found that out of 9 mutants with candidate driver mutations, only 4 were able to transform cells in culture (for review, see Futreal, 2007).

The Wellcome Trust Sanger Institute Catalogue of Somatic Mutations in Cancer (COSMIC) collates and displays somatic mutation information relating to human cancers

(Forbes *et al.*, 2006). At the time of writing (May 2008, COSMIC release 37), the database contained mutation data for around 4,770 genes from ~260,000 tumours. Gene resequencing is also a major component of the $50 million 3-year pilot phase of the Cancer Genome Atlas (http://cancergenome.nih.gov/), a large-scale collaboration between the National Cancer Institute (NCI) and the National Human Genome Research Institute (NHGRI).

1.3.2 Gene expression profiling

Gene expression arrays can be used to analyse the transcription of thousands of genes simultaneously. There are two main types: cDNA arrays, where clones corresponding to the transcripts to be analysed are spotted onto a matrix, and oligonucleotide arrays, where oligonucleotides corresponding to the transcripts are synthesised onto a matrix along with mismatch control oligonucleotides. A new approach has also been developed, in which the abundance of transcripts is measured directly using Illumina (http://www.illumina.com) sequencing technology. In two-colour microarray expression analysis, the sample of interest and a control sample are differentially labelled with fluorescent dyes and are hybridised onto the array, which is then scanned to determine the ratio of fluorescence intensities for each gene. The ratio represents the relative amounts of transcript in the sample. Unsupervised clustering of the expression data for multiple samples can be used to subcategorise cancers. For example, lung cancers cluster into known histological subtypes that are predictive of patient survival (Beer *et al.*, 2002; Bhattacharjee *et al.*, 2001; Garber *et al.*, 2001). Gene expression profiles may also provide an indication of the genes involved in oncogenesis in a given tumour. Lung cancers harbouring a mutation in *KRAS* have a characteristic expression profile that can be used in their identification (Sweet-Cordero *et al.*, 2005). Analysis of gene expression does not provide any insights into the underlying genetic changes and it can be affected by physiological variation, such as the degree of inflammatory response or hypoxia (Eden *et al.*, 2004). However, it is important as a complementary approach to other methods of cancer profiling, such as mutational and copy number analysis. Integrative approaches involving gene expression and copy number analysis are discussed in the following section.

1.3.3 Copy number analysis

1.3.3.1 DNA copy number changes

Changes in DNA copy number result from chromosomal aberrations such as deletions and duplications, non-reciprocal translocations and gene amplifications. Copy number variations (CNVs) have been identified in all humans studied (Feuk *et al.*, 2006), and a genome-wide study of 270 apparently healthy individuals from four diverse populations identified almost 1,500 germline copy number variable regions encompassing 12% of the human genome (Redon *et al.*, 2006). CNVs accounted for ~18% of the total detected variation in gene expression between individuals, suggesting that they make a considerable contribution to phenotypic variation (Stranger *et al.*, 2007). In the context of cancer, genomic instability results in the acquisition of somatic copy number aberrations that may contribute to tumourigenesis through the amplification of oncogenes and/or loss of tumour suppressor genes. Genomic instability is also referred to in Sections 1.2.3.3 and 1.2.5.1.3.

Chromosome instability, which manifests as alterations in chromosome number (aneuploidy), seems to arise early in tumourigenesis but increases with tumour progression (for review, see Lengauer *et al.*, 1998). Fridlyand and coworkers (2006) found that shorter or altered telomeres were associated with greater numbers of amplifications but that the frequency of low-level changes was associated with altered expression of genes involved in mitosis, cell cycle, DNA replication and repair, and included many genes that are direct targets of E2F (Fridlyand *et al.*, 2006). This suggests that the RB1 pathway (see Section 1.2.6) contributes to chromosome instability, as hypothesised by Hernando *et al.* (2004) (Fridlyand *et al.*, 2006). Advanced tumours tend to reach a stable state, which, in the form of cancer cell lines, are stable over many generations and in different laboratories, suggesting that they have evolved to an optimal state (Albertson *et al.*, 2003).

1.3.3.2 Using CGH to detect copy number changes

Large alterations in copy number were initially detected and quantified using metaphase spreads in a technique known as comparative genomic hybridisation (CGH) (Kallioniemi *et al.*, 1992). In CGH, cancer and normal genomic DNA are differentially labelled with fluorochromes and are co-hybridised to normal metaphase chromosomes. Cot-1 DNA is added to suppress hybridisation to repetitive elements in the genome. The ratio of fluorescence intensities at any chromosomal position is approximately proportional to the ratio of copy numbers of the cancer and normal DNA at that position (reviewed in Pinkel *et al.*, 1998). CGH profiles can be viewed and compared using the NCBI Cancer Chromosomes database, which integrates three databases of chromosomal aberrations in cancer: the SKY/M-FISH & CGH Database, the Mitelman Database of Chromosome Aberrations in Cancer, and the Recurrent Chromosome Aberrations in Cancer database (Knutsen *et al.*, 2005). Rearrangement breakpoints are linked to the underlying genome assembly. However, the tool is limited to cytogenetic resolution because CGH cannot detect changes of less than 20 Mb or distinguish changes that are close together, and it cannot determine exact genomic coordinates (Pinkel *et al.*, 1998).

Array CGH is a higher resolution, high-throughput version of conventional CGH, in which differentially labelled cancer and reference samples are hybridised to an array made from large genomic clones, e.g. bacterial artificial chromosomes (BACs), or cDNAs (for review, see Albertson and Pinkel, 2003; Pinkel *et al.*, 1998; Pollack *et al.*, 1999). The copy number is measured at each probe on the array, and can be mapped directly to the genome. A disadvantage of array CGH is that it cannot detect loss of heterozygosity (LOH), which has traditionally been identified using methods involving microsatellites and restriction fragment length polymorphisms (RFLPs) that are not suitable for large scale analyses (see Thomas *et al.*, 2006).

Single nucleotide polymorphism (SNP) arrays are the most recent development in copy number analysis. SNPs account for most of the genetic variation in the human genome (Stranger *et al.*, 2007) and they occur, on average, every 100-300 base pairs along the genome. The Affymetrix GeneChip Mapping Assay (http://www.affymetrix.com) is a commonly used procedure that combines a whole-genome sampling assay (WGSA) with high-density SNP arrays (Kennedy *et al.*, 2003; Matsuzaki *et al.*, 2004). WGSA is used to reduce the complexity of the sample, and involves ligating an adapter to restriction-digested DNA, which enables PCR amplification using a single primer that is complementary to the adapter (Figure 1.4B). The amplified DNA is then fragmented, labelled and hybridised to the array. SNPs within the amplified DNA are used as probes on the array, therefore ensuring that all probes are informative (Bignell *et al.*, 2004). In the Affymetrix GeneChip Mapping 10K assay, which uses an array containing 11,555 SNPs, WGSA involves a single restriction enzyme, *XbaI* (Kennedy *et al.*, 2003).

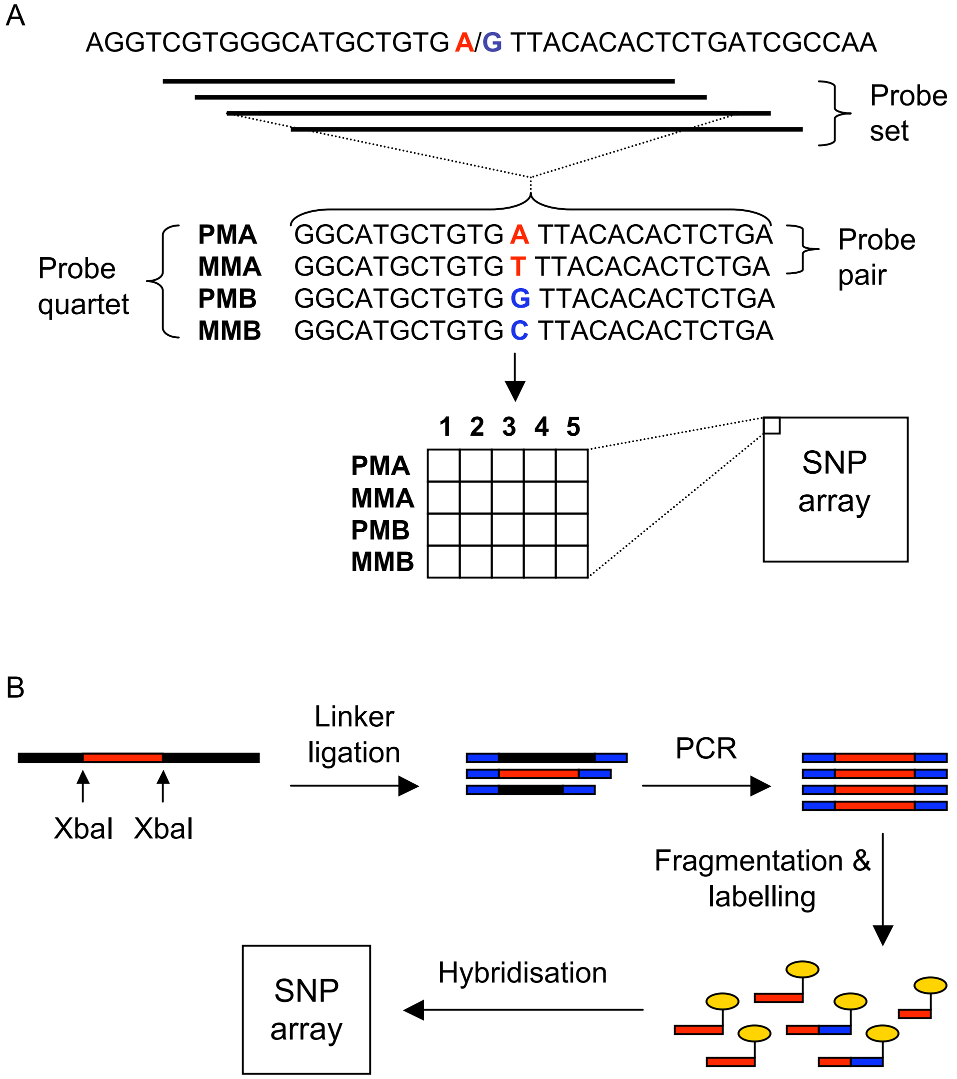
A
AGGTCGTGGGCATGCTGTG A/G TTACACACTCTGATCGCCAA
Probe set
Probe quartet
PMA GGCATGCTGTG A TTACACACTCTGA
MMA GGCATGCTGTG T TTACACACTCTGA
PMB GGCATGCTGTG G TTACACACTCTGA
MMB GGCATGCTGTG C TTACACACTCTGA
Probe pair
1 2 3 4 5
PMA
MMA
PMB
MMB
SNP array
B
XbaI XbaI
Linker ligation
PCR
Fragmentation & labelling
Hybridisation
SNP array

Figure 1.4. Array design (A) and whole-genome sampling assay (B) for the Affymetrix SNP array. A. A SNP in the DNA sequence is shown in red/blue. The SNP is represented in the array by a probe set, which comprises multiple probe quartets that differ from one another in the position of the polymorphic site relative to the centre of the probe. Each probe quartet consists of four 25mer oligonucleotides in the form of two probe pairs, which comprise a perfect match (PM) probe and a mismatch (MM) probe corresponding to each SNP allele (A and B). **B.** Genomic DNA is digested with a restriction enzyme, shown here as *XbaI*, and a linker (shown in blue) is ligated to the digested DNA. The DNA is PCR amplified using a primer that binds to the linker. Amplified DNA is fragmented, labelled and hybridised to the array.

Regions of the genome in which the *XbaI* site is rare will be under-represented in the array (Bignell *et al.*, 2004). The higher resolution 100K SNP array therefore use two restriction enzymes, *XbaI* and *HindIII*, which produce complementary SNP densities (Matsuzaki *et al.*, 2004). Each SNP in an Affymetrix array is represented by a “probe set” comprising multiple “probe quartets”. Each probe quartet consists of four 25mer oligonucleotides in the form of two “probe pairs” comprising a perfect match probe and a mismatch probe corresponding to each SNP allele (Figure 1.4A). Probe quartets differ from one another in offset, i.e. the position of the polymorphic site relative to the centre of the oligonucleotide, and orientation (reviewed in Xiao *et al.*, 2007). Normal and tumour DNA are hybridised to different arrays, therefore avoiding the need for matched samples and allowing for a pool of normal samples to be used as a control (Bignell *et al.*, 2004; Figure 1.4C). As in other forms of array CGH, the copy number at each probe can be inferred from the intensity of fluorescence of hybridised sample DNA (Bignell *et al.*, 2004; Zhao *et al.*, 2004).

Commercially available arrays now range in resolution from 10,000 to ~1 million SNPs across the genome. SNP arrays therefore provide the potential for fine mapping of copy number changes, enabling the identification of small aberrations and accurate mapping of chromosomal breakpoints. Furthermore, the SNPs can be genotyped and compared to a normal sample to identify regions of LOH. This permits the identification of complex changes such as LOH without decrease in copy number and decrease in copy number without LOH (Bignell *et al.*, 2004; Raghavan *et al.*, 2005; Zhao *et al.*, 2004). Such changes are common, as demonstrated in pancreatic and cervical cancer cell lines, where the proportion of LOH associated with copy-reduction was found to be just 32% (Calhoun *et al.*, 2006) and 25% (Kloth *et al.*, 2007), respectively.

CGH signal intensities must be normalised to account for technical bias while still retaining biologically relevant changes. Normalisation of array CGH data has generally involved the use of methods originally developed for normalising gene expression microarray data (for review, see Quackenbush, 2002). Cross-slide and within-slide normalisation are used to transform the data such that all arrays, and all the spots on each array, are comparable. In median normalisation, all values are multiplied by a constant factor so that all arrays have a median $\log_2$ ratio of 0. Lowess, or Loess, normalisation accounts for spot intensity biases and other dependencies such as the location of the spot on the array and the use of different print tips. The data are linearised by subtracting a

Lowess regression curve. A number of additional methods for dealing with spatial effects in expression microarray data are reviewed in Neuvial *et al.* (2006).

In general, array CGH must be more stringent than gene expression analysis because it is required to detect single copy changes and, while the copy number, unlike the expression level, of a gene is expected to be identical in two samples, this is often not the case due to tumour heterogeneity and the presence of contaminating stromal cells (Khojasteh *et al.*, 2005). Khojasteh and coworkers (2005) proposed a multi-step normalisation process specifically for dealing with array CGH data. A "spatial segmentation" algorithm has also been developed to account for array CGH-specific spatial effects designated "local spatial biases", where clusters of spots show a shift in signal, and "continuous spatial gradient", where there is a smooth gradient in signal across the array (Neuvial *et al.*, 2006). Staaf and coworkers (2007) showed that copy number imbalances correlate with intensity in array CGH data and that normalisation of expression data erroneously corrects for biologically relevant gains in copy number. They have therefore developed a normalisation algorithm that prevents suppression of copy number ratios by stratifying the data into separate populations representing discrete copy number levels (Staaf *et al.*, 2007). Array CGH data are also affected by a genome-wide technical artefact termed "spatial autocorrelation", or "wave", for which the peaks and troughs are aligned across samples but the amplitude, and for some samples, the direction, varies (Marioni *et al.*, 2007). Removal of the wave using a Lowess curve led to an increase in the number of biologically relevant CNVs detected in array CGH data from normal individuals (Marioni *et al.*, 2007).

Affymetrix have developed a number of procedures for normalising SNP array CGH data. As described above, each SNP on an Affymetrix array is represented by a probe set comprising multiple probe pairs (Figure 1.4A). Fluorescence on the mismatch probes represents non-specific hybridisation, and the data can be corrected by subtracting the mismatch from the perfect match intensity for each probe pair. The corrected intensities are then averaged across the probe set. The data can be globally normalised by multiplying the average intensity of the experimental array, i.e. the array to which the cancer sample is hybridised, by a normalisation factor to make it numerically equivalent to the average intensity of the control array, to which a normal sample is hybridised. Intensity ratios are calculated by dividing the average intensity for each SNP in the experimental array by the equivalent value in the control array. Three software packages that are commonly used for processing copy number data on Affymetrix SNP arrays are Copy Number Analyser for GeneChip arrays (CNAG, Nannya *et al.*, 2005), DNA-Chip Analyzer (dChip, Zhao *et al.*, 2004) and Affymetrix GeneChip Chromosome Copy Number Analysis Tool (CNAT, Huang *et al.*, 2004). These are compared and reviewed in Baross *et al.* (2007), who concluded that the detection of all real CNVs from a 100K array necessitated the combined use of multiple procedures.

The next step, following normalisation, is to identify regions of copy number change within the CGH data. Many different approaches have been developed for segmenting the genome into regions of homogeneous copy number. These include change-point analysis, where the genome is segmented at points where the copy number changes significantly (Olshen *et al.*, 2004; Venkatraman and Olshen, 2007), Hidden Markov Models (HMMs) (Engler *et al.*, 2006; Marioni *et al.*, 2006; Nannya *et al.*, 2005; Rueda and Diaz-Uriarte, 2007; Shah *et al.*, 2006; Stjernqvist *et al.*, 2007), hierarchical clustering along chromosomes (Wang *et al.*, 2005) and smoothing methods (Hsu *et al.*, 2005; Huang *et al.*, 2007). There are also a number of web-based applications, such as ADaCGH (Diaz-Uriarte and Rueda, 2007) and CGHweb (Lai *et al.*, 2008), for viewing and comparing outputs from multiple algorithms. Further methods have been developed to identify copy number changes specifically in SNP array CGH data, which has increased noise at the probe level compared with BAC array CGH (Yu *et al.*, 2007), and a number of these infer allele-specific copy numbers (Huang *et al.*, 2006a; LaFramboise *et al.*, 2005; Lamy *et al.*, 2007; Nannya *et al.*, 2005; Yu *et al.*, 2007). Some of the methods for detecting copy number changes are discussed in further detail in Section 4.6.

Finally, having identified regions of copy number change, the statistical power can be increased by examining the region across many samples. Unlike for CNVs in normal samples, cross-sample analysis of copy number changes in cancer is hampered by the large size of many rearrangements, variation in the location of breakpoints between samples, and sample heterogeneity that prevents accurate estimation of the copy number (Marioni *et al.*, 2007). A handful of methods have been developed to identify recurrent regions of copy number change in tumours: CMAR (Rouveirol *et al.*, 2006), STAC (Diskin *et al.*, 2006), H-HMM (Fiegler *et al.*, 2007) and KC-SMART (Klijn *et al.*, 2008). The latter is the only algorithm that does not discretise the data into 3 states (loss, gain and no change), which can lead to undetected copy number changes in heterogeneous tumours (Klijn *et al.*, 2008).

1.3.3.3 Analysis of copy number changes in cancer genomes

CGH can detect aneuploidy, gene amplifications and deletions, and non-reciprocal translocations in cancer genomes. Gene amplifications are gains in copy number of restricted regions of DNA (Bignell *et al.*, 2007) that contribute to tumourigenesis by increasing the transcript levels, and therefore the protein levels, of oncogenes (Schwab, 1999). Gene amplification is the major mechanism of oncogenesis for a number of cancer genes, including *MYCN*, which is amplified in ~30% of advanced neuroblastomas (Seeger *et al.*, 1985). Amplified genes represent a promising target for cancer therapy, as demonstrated in breast cancers harbouring an amplified *HER*/*ERBB2* receptor gene (Cobleigh *et al.*, 1999, see Section 1.2.7).

Deletions are an important mechanism for inactivating tumour suppressor genes, including *PTEN* (Li *et al.*, 1997) and *CDKN2A* (*INK4A/ARF*) (Orlow *et al.*, 1995). A genome-wide analysis of homozygous deletions in over 600 cancer cell lines showed that deletions occur in regions with fewer genes and repeat elements but higher flexibility compared with the rest of the genome (Cox *et al.*, 2005). A significant proportion occur in regions that are prone to chromosome breakage, and some of the genes in these "fragile sites", such as *WWOX* and *FHIT*, show similar mutational patterns to known tumour suppressor genes, so it is not clear whether or not these genes are causally implicated in cancer (Futreal *et al.*, 2004).

Like gene expression analysis, copy number profiling can be used to subcategorise cancers. It can distinguish three subtypes of glioblastoma (Maher *et al.*, 2006), and separates leiomyosarcomas into a distinct cluster from gastrointestinal stromal tumours, which, until recently, were classified as the same tumour type (Meza-Zepeda *et al.*, 2006). It also provides predictive power in breast cancer prognosis, where a poor prognosis is indicated by high-level amplification (Chin *et al.*, 2006), extensive chromosome instability (Fridlyand *et al.*, 2006) and/or the presence of multiple, closely spaced amplicons, or "firestorms", on a single chromosome arm (Hicks *et al.*, 2006). Copy number profiles can also help to stage a tumour, such as in cervical cancer, where gain of chromosome 3q is associated with the transition from severe dysplasia to invasive carcinoma (Kersemaekers *et al.*, 1998). Furthermore, studies in ovarian cancer have revealed an association between drug response and the presence of copy number changes associated with drug sensitivity or resistance (Bernardini *et al.*, 2005; Kim *et al.*, 2007a).

The amplification of genes involved in drug metabolism or inactivation is commonly observed in cultured cells as a means of acquiring drug resistance (Lengauer *et al.*, 1998).

While many cancer genomes have been analysed for copy number changes, there has been limited progress in determining the functional significance of altered regions. One successful approach involves identifying recurrently altered regions that are specific to particular tumour types. This enables the identification of “lineage addiction” cancer genes, which may target essential lineage-specific survival functions and therefore represent promising therapeutic targets (Garraway and Sellers, 2006). Two such genes are the melanoma-specific oncogene *MITF*, which is selectively amplified and overexpressed in 20% of melanomas (Garraway *et al.*, 2005), and *NKX2-1*, which lies in the minimal amplified region of a lung-cancer-specific amplicon on chromosome 14q13.3 found in up to 20% of lung cancers (Kendall *et al.*, 2007; Weir *et al.*, 2007). Genes *TTF1* and *NKX2-8* are usually co-amplified with *NKX2-1* in the 14q13.3 amplicon and all three genes have been shown to co-operate in lung tumourigenesis (Kendall *et al.*, 2007). The co-occurrence and mutual exclusivity of copy number alterations at different loci may also reflect co-operating and complementary cancer genes, respectively. For example, gains of *ERBB2* and *CCNE1* frequently co-occur in bladder cancer, while *CCND1* and *E2F1*, which function in the same pathway, are mutually exclusive (Veltman *et al.*, 2003).

The identification of cancer genes in regions of copy number change can be challenging because changes often span large regions of the genome that encompass many genes and may include many attractive candidates. Gains of more than one copy may have involved multiple evolutionary events and the critical gene may reside at the highest peak in copy number, as demonstrated for oncogenes *CYP24* and *ZNF217* in breast cancer (Albertson *et al.*, 2000). Measurement of gene expression is also important for evaluating candidate cancer genes. *SPANXB* was identified as the putative critical gene in an Xq duplication in acute lymphoblastic leukaemias with an *ETV6/RUNX1* translocation since it was the only gene with high and uniform overexpression across all samples (Lilljebjorn *et al.*, 2007). While gene expression and gene dosage are rarely perfectly correlated, many studies, such as the comparison of array CGH and gene expression data in breast cancers, have shown good correlation (Hyman *et al.*, 2002; Pollack *et al.*, 2002). However, genes that are amplified are not necessarily overexpressed, as demonstrated by Kloth and colleagues (2007), who did not observe a genome-wide correlation between copy number and gene expression in cervical cancer cell lines. Gene expression is influenced by factors other than gene dosage, such as the availability of transcription and regulatory factors, DNA methylation and chromatin conformation, and the presence of miRNAs (Kloth *et al.*, 2007).

The integration of copy number analysis with gene resequencing also facilitates cancer gene identification. Mullighan and colleagues (2007) performed a genome-wide analysis of genetic alterations in 242 paediatric acute lymphoblastic leukaemias (ALL) using 100K and 250K SNP arrays. They found mutations in genes that regulate late B lymphocyte development in 40% of B-progenitor ALL cases. *PAX5* mutations, which included deletions, point mutations and translocations, were identified in 32% of cases (Mullighan *et al.*, 2007). ALL genomes are relatively stable, but genomes harbouring different translocations show variability in the number of copy number changes, which may reflect differences in the number of events required for tumourigenesis (Mullighan *et al.*, 2007; Wang and Armstrong, 2007). The integration of resequencing data, and epigenetic data (see Section 1.3.4), can facilitate the identification of tumour suppressor genes in regions of LOH, where the other allele may be inactivated by point mutation or epigenetic changes.

The identification of human cancer genes is aided by the integration of complementary genome-wide analyses of human cancers, but the integration of cancer-associated mutation datasets from other species, particularly the mouse, provides an even more powerful approach for cancer gene discovery. Cross-species comparisons are discussed in Section 1.5.

1.3.3.4 Limitations of CGH and alternative strategies

Limitations of CGH-based approaches include difficulties in determining the ploidy of the sample and identifying the location of rearranged sequences in the cancer genome. However, the ploidy and location of larger rearrangements (> 10 Mb) can be discerned by combining CGH with G-banding or Spectral Karyotyping (SKY) (Watson *et al.*, 2007). CGH may also struggle to detect low level changes and changes in heterogeneous samples, e.g. primary cancers containing normal stromal cells, and it is affected by low-copy reiterated sequences, including gene paralogues (for full review, see Pinkel and Albertson, 2005).

A further limitation of CGH is that while it can detect nonreciprocal, or unbalanced, translocations, which result in the gain or loss of DNA and often cause the inactivation of tumour suppressor genes (Mitelman *et al.*, 2004), it cannot detect reciprocal, or balanced, translocations. These result in fusion transcripts or transcriptional deregulation due to the positioning of an intact gene next to promoter and/or enhancer elements of another gene. It has recently been discovered that cytogenetically balanced translocations are frequently associated with focal copy number alterations, suggesting that high-resolution array CGH may in fact be capable of detecting a proportion of balanced translocations in cancer (Watson *et al.*, 2007). However, truly balanced translocations cannot be identified.

Balanced translocations are often initiating events in tumourigenesis that are essential for tumour development, and they therefore represent promising therapeutic targets (see Section 1.2.7). Until recently, it was thought that balanced translocations predominated in haematopoietic tumours, but an assessment of data in the Mitelman Database of Chromosome Aberrations in Cancer suggests that they also play an important role in epithelial tumourigenesis (Mitelman *et al.*, 2004). Furthermore, human solid tumours appear to contain large numbers of gene fusions (Volik *et al.*, 2006) and a quarter of the breakpoints detected in 3 breast cancer cell lines were found to be balanced (Howarth *et al.*, 2008). The high-throughput identification of balanced translocations has been hindered because translocation breakpoints cannot be amplified by PCR (Howarth *et al.*, 2008). Genome-wide techniques for identifying translocations include array painting, in which chromosomes are sorted and DNA is amplified and hybridised to DNA microarrays (Howarth *et al.*, 2008), and informatics approaches, such as the algorithm developed by Tomlins and coworkers (2005) that used RNA expression data to identify candidate gene fusions in prostate cancers. The *EML4-ALK* fusion was identified in non-small cell lung cancers by paired-end sequencing (Soda *et al.*, 2007).

End-sequence profiling (ESP) can be used to precisely map all types of genomic rearrangements, including balanced translocations (Volik *et al.*, 2003). ESP involves constructing a BAC library from the cancer genome and sequencing the ends of clones to identify rearrangements, which map to locations in the reference genome that are of abnormal distance or orientation (Volik *et al.*, 2003; Figure 1.5). The method can also identify fusion transcripts (tESP) and can be targeted to specific amplicons (Volik *et al.*, 2006). Complete sequencing of the BACs enables detailed analysis of the structure of genomic rearrangements and elucidation of the mechanisms of rearrangement.

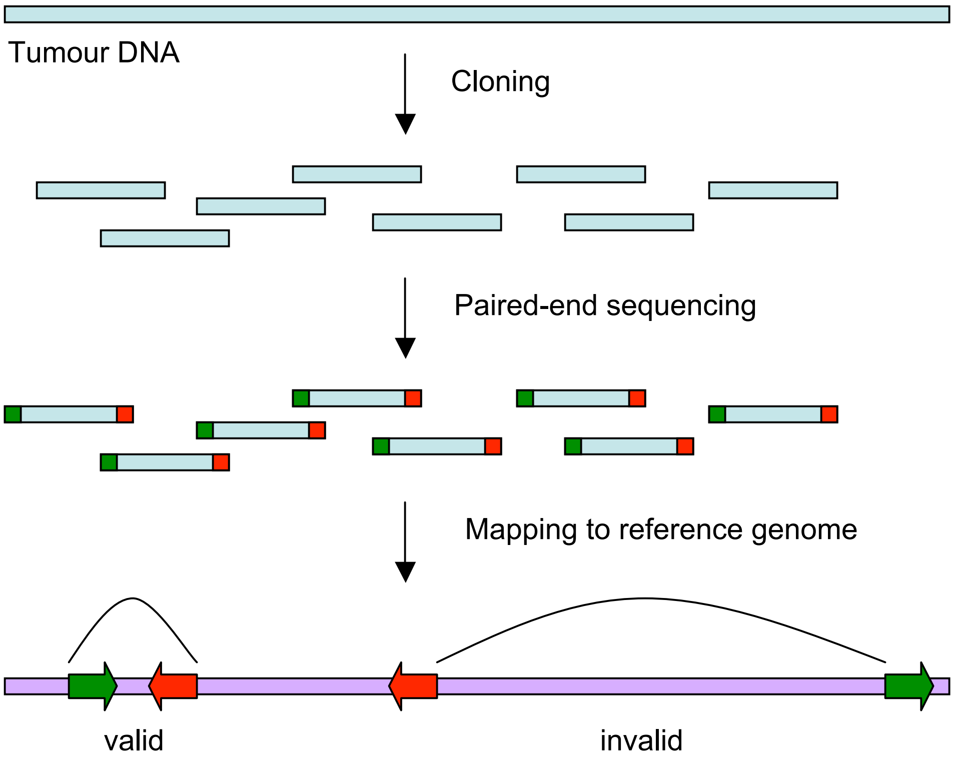

Figure 1.5. End sequence profiling of tumour DNA. 100-250 kb regions of the tumour genome are cloned and a 500 bp region at the end of each clone is sequenced. The ends are mapped to the human reference genome. Ends that are an abnormal distance apart or in an abnormal orientation, shown here as "invalid", are indicative of rearrangements within the tumour genome. Redrawn with minor modifications from Figure 1 of Raphael *et al*. (2008).

ESP-based analysis of 4 cancer amplicons revealed evidence for sister chromatid break-fusion-bridge cycles, the excision and reintegration of double minutes (extrachromosomal DNA), and more complex architectures involving clusters of small genomic fragments (Bignell *et al.*, 2007). Break-fusion-bridge cycles are initiated by a double-strand chromosomal break, which, following DNA synthesis, results in sister chromatids with identical free DNA ends that fuse to one another to prevent apoptosis. An anaphase bridge is formed during chromatid separation in mitosis, and this results in a new double-strand break and reinitiation of the cycle (McClintock, 1941).

ESP analysis of 6 epithelial cancers, including primary tumours from brain, breast and ovary, plus a metastatic prostate tumour and 2 breast cancer cell lines, revealed extensive chromosomal rearrangements, some of which appeared to be recurrent (Raphael *et al.*, 2008). Despite the benefits of this strategy, sequencing large numbers of clones across many cancer genomes is costly and impractical. However, Bashir and colleagues (2008) have derived a formula to maximise the probability of detecting fusion genes with the least amount of sequencing. The formula depends on the distribution of gene lengths and the parameters of the sequencing strategy used (Bashir *et al.*, 2008). A high-throughput alternative to ESP, which involves massively parallel sequencing of the ends of randomly sheared DNA, has recently been applied to the genome-wide analysis of somatic and germline rearrangements in 2 lung cancers (Campbell *et al.*, 2008). The analysis revealed a wide spectrum of rearrangements, as well as providing high-resolution copy number information. Paired-end sequencing is an attractive strategy for the complete characterisation of rearrangements in cancer.

1.3.4 Epigenetic profiling

Epigenetic changes are chemical modifications to the DNA or histones that change the structure of chromatin but do not alter the DNA sequence. If chromatin is in the condensed conformation, transcription factors cannot access the DNA and genes are therefore not expressed, whereas genes in open chromatin can be expressed as required. DNA methylation and changes in chromatin conformation have both been implicated in tumourigenesis. DNA methylation of CpG islands, which are located in promoter regions, can result in gene "silencing" by preventing transcription factor binding. It can also repress gene expression by recruiting methyl-binding domain proteins, which associate with histone deacetylases (HDACs). HDACs mediate chromatin condensation by deacetylating histones. See Pelengaris and Khan (2006).

Aberrant DNA methylation of *CDKN2A* has been observed in a wide range of common cancer types (Herman *et al.*, 1995; Merlo *et al.*, 1995), while *VHL* and *BRCA1* are silenced by methylation in a significant proportion of kidney (Herman *et al.*, 1994) and breast and ovarian cancers (Esteller *et al.*, 2000), respectively. *VHL* and *BRCA1* are also frequently mutated in cancer, but for other tumour suppressor genes, such as *RASSF1A*, promoter hypermethylation appears to be the principal mechanism for inactivation (for review, see Jones and Baylin, 2002).

Detection of DNA methylation relies on the ability to distinguish cytosine from 5-methylcytosine. This can be achieved using restriction enzymes that restrict only unmethylated DNA, or by using sodium bisulfite, which converts unmethylated cytosines to uracil, or by immunoprecipitation of methylated DNA using 5-methylcytosine-specific antibodies or methyl-binding domain proteins (see Down *et al.*, 2008). All three approaches can be applied to the genome-wide detection of DNA methylation through the use of oligonucleotide arrays. However, restriction enzyme-based methods are limited to the analysis of CpG sites that contain the recognition site for the enzyme in use, while bisulfite conversion reduces the complexity of the DNA and so reduces the number of unique probes that can be used on the array (Down *et al.*, 2008). Bisulfite conversion and methylated DNA immunoprecipitation have also been combined with next-generation sequencing in techniques known as BS-seq (Cokus *et al.*, 2008) and MeDIP-seq (Down *et al.*, 2008), respectively. Histone modifications can be detected using chromatin immunoprecipitation (ChIP), which is described in Section 1.3.5.

Large genomic regions, such as an entire chromosome arm, can show aberrant methylation in cancer (Frigola *et al.*, 2006), and there is evidence to suggest that some cancers show a CpG island methylator phenotype (CIMP). CIMP+ colorectal cancers have significantly more hypermethylation at CpG islands, including an increased incidence of *CDKN2A* and *THBS1* methylation (Toyota *et al.*, 1999), and they are characterised by a methylated mismatch repair gene, *MLH1*, which gives rise to microsatellite instability (Weisenberger *et al.*, 2006; see Section 1.2.5.1.3 for a description of microsatellite instability). Genes that are reversibly repressed by Polycomb proteins in embryonic stem cells are significantly over-represented amongst constitutively hypermethylated genes in colorectal cancers (Widschwendter *et al.*, 2007). This provides support for the theory of a stem cell origin of cancer (Section 1.2.3.2). A detailed discussion of the epigenomics of cancer is beyond the scope of this thesis, which focuses on changes in cancer that alter the DNA sequence. Epigenomics approaches are reviewed in Callinan and Feinberg (2006) and, for a detailed review of epigenomics and its relevance to the cancer stem cell hypothesis, see Jones and Baylin (2007).

1.3.5 Genome-wide mapping of transcription factor binding sites

The mapping of transcription factor binding sites (TFBS) across the whole genome can help to elucidate gene regulatory networks. Chromatin immunoprecipitation (ChIP) is a powerful approach for analysing TFBS in living cells (Wei *et al.*, 2006). Cells are treated with formaldehyde to mediate the formation of cross-links between DNA and proteins. The chromatin is then fragmented by sonication and an antibody against the transcription factor of interest is used to immunoprecipitate the transcription factor bound to DNA (see Loh *et al.*, 2006). The precipitated DNA can be used to probe a DNA microarray in a high-throughput method known as ChIP-chip. This approach has been used to map TFBS in the yeast genome (Ren *et al.*, 2000). For more complex genomes, it has been necessary to restrict analysis to specific regions, such as promoter regions or individual chromosomes (Boyer *et al.*, 2005; Cawley *et al.*, 2004; Horak *et al.*, 2002; Weinmann *et al.*, 2002), but more recent analyses have used ChIP-chip to survey the entire genome (Kim *et al.*, 2005b; Lee *et al.*, 2006).

An alternative approach involves cloning and sequencing the precipitated DNA fragments, and then mapping the sequences to the genome. Initially, this involved the sequencing of individual fragments sampled from the DNA pool (Hug *et al.*, 2004; Weinmann *et al.*, 2001). However, high coverage is required to distinguish real binding sites from background DNA, and this has been achieved at reduced cost by sequencing a “tag” from each DNA fragment by serial analysis of gene expression (SAGE) (Chen and Sadowski, 2005; Impey *et al.*, 2004; Kim *et al.*, 2005a; Roh *et al.*, 2005). To overcome the problems of ambiguity associated with mapping short tags, Wei and coworkers (2006) developed an approach called ChIP-PET, in which ChIP is coupled with paired-end ditag (PET) sequencing so that both the 5’ and 3’ ends of each DNA fragment are sequenced (Figure 1.6). This method was applied to the unbiased global mapping of 542 p53 binding sites in the human genome (Wei *et al.*, 2006). The functions of p53 target genes

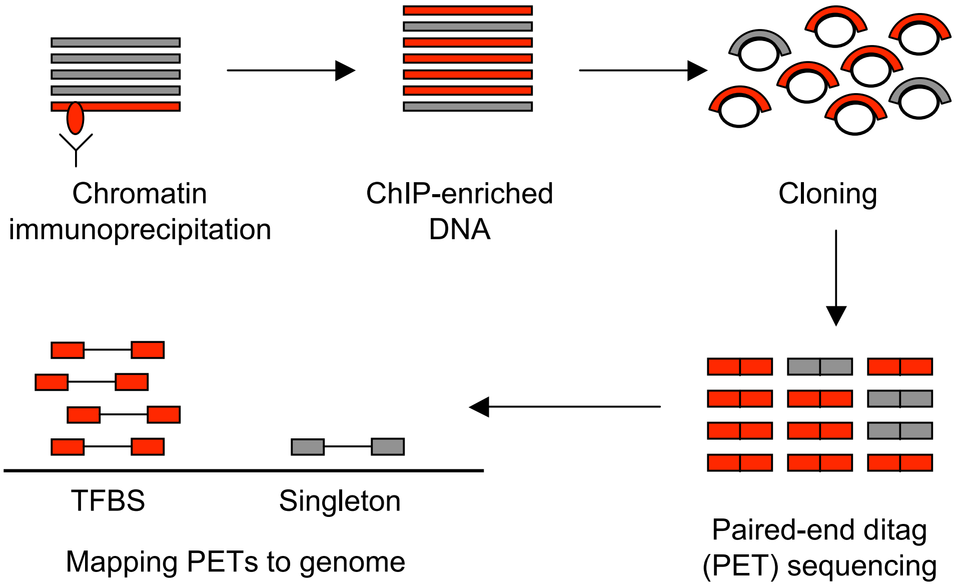
Chromatin immunoprecipitation
ChIP-enriched DNA
Cloning
Paired-end ditag (PET) sequencing
TFBS
Singleton
Mapping PETs to genome

Figure 1.6. Overview of ChIP-PET for mapping transcription factor binding sites. In chromatin precipitation (ChIP), the chromatin is fragmented by sonication and an antibody against the transcription factor of interest is used to precipitate the transcription factor bound to DNA. The ChIP-enriched DNA is cloned and the ends of each clone are sequenced to create a library of paired-end ditags (PETs). The PETs are mapped to the reference genome. Multiple PETs mapping to a single location indicate the presence of a transcription factor binding site (TFBS) at that location. Redrawn with modifications from Figure 1 of Loh *et al.* (2006) .

included known roles of p53, such as apoptosis, DNA repair and transcription regulation, but also novel functions, such as cell adhesion and mobility (see Wei *et al.*, 2006).

Loh and coworkers (2006) applied the ChIP-PET technology to the global mapping of Oct4 and Nanog binding sites within mouse embryonic stem (ES) cells. Oct4 and Nanog are required for the maintenance of ES cell pluripotency and self-renewal and may play an important role in cancer (see Section 1.2.3.2). Approximately 1,000 and 3,000 high confidence binding sites were identified for Oct4 and Nanog, respectively, and the presence of one or other binding site was found to be associated with genes that are repressed and induced during differentiation. The target genes include known effectors of ES cell fate, such as *Foxd3* and *Setdb1*, genes required for maintaining pluripotency, including *Esrrb* and *Rif1*, and *Mycn*, which is involved in ES cell self-renewal and proliferation. Most of the Oct4 binding sites also bind Sox2, suggesting that Oct4 and Sox2 co-operate in regulating gene expression.

ChIP-PET has also been used in human B cells to identify more than 4,000 potential binding sites for Myc, of which 668 were identified as direct targets of Myc regulation (Zeller *et al.*, 2006). Many of the target genes are involved in protein synthesis and cell metabolism, which is consistent with a role for Myc in controlling cell size. A large number of transcription factors were also identified. This study showed a weak overlap with other analyses of Myc binding sites, reflecting the current limitations of ChIP-PET, such as the limited sensitivity of PET detection, the experimental noise associated with ChIP, and the fact that the analysis only describes a snapshot of transcription factor binding at a particular moment in time (Zeller *et al.*, 2006). A comparative study of STAT1 binding sites identified by ChIP-chip and ChIP-PET found a considerable overlap between methods, but each method also identified unique sites, suggesting that higher accuracy could be achieved by using both techniques (Euskirchen *et al.*, 2007).

The most advanced method for identifying TFBS is ChIP-seq, in which the DNA fragments isolated by ChIP are amplified and sequenced using next-generation sequencing technology. ChIP-seq requires less starting material and involves fewer steps, making it faster and less prone to error. ChIP-seq using Solexa massively parallel sequence identified STAT1 binding sites in human HeLa S3 cells with an estimated sensitivity of 70-92% and specificity of at least 95% (Robertson *et al.*, 2007).

1.4 Cancer gene discovery in the mouse

1.4.1 The mouse as a model for studying cancer

1.4.1.1 Background

The mouse is a leading model system for cancer research because it has a rapid reproduction rate and breeds well in captivity and, owing to its small size, it can be maintained in large numbers in limited space (see Frese and Tuveson, 2007). It is also genetically and physiologically similar to human. In light of these factors, the mouse genome has been sequenced and annotated to a high standard, second only to that of human (Waterston *et al.*, 2002).

The mouse was initially used as a cancer model through tumour transplantation within inbred strains, but following the discovery of the immunodeficient "nude" mouse and, later, the severe combined immunodeficient (SCID) mouse, it became possible to transplant human tumours into the mouse, creating xenograft models. Such models can be used to rapidly assess tumour tissue and cell lines *in vivo* but they do not fully recapitulate the behaviour of an endogenous tumour because many features of the tumour microenvironment, such as stromal cells, vasculature and immune cells, are missing. The tumour xenograft is also likely to be less heterogeneous than the endogenous tumour because cells in culture are under high selective pressure. These factors have contributed to the limited success of xenograft models in drug development (for review, see Sharpless and Depinho, 2006)

Many inbred strains that spontaneously develop cancer at high frequency have been established, and these, as well as mice that have been treated with a mutagen, are useful for studying the properties of endogenous cancers *in vivo*. They have been used to identify cancer genes and to assess the effects of carcinogens and therapeutic compounds. However, these models may be biased towards specific types of tumour that show variable penetrance and latency and do not accurately reflect common human cancers (Frese and Tuveson, 2007).

1.4.1.2 Genetically engineered mouse models

Genetically engineered mouse models represent a major advance in cancer research that allows for the study of gene function *in vivo* and for the creation of models that more accurately recapitulate human cancers. Genetically engineered models can be classified as transgenic or endogenous (Frese and Tuveson, 2007).

1.4.1.2.1 Transgenic models

Transgenic mice can be created to study the effect of overexpressing an oncogene or a dominant-negative tumour suppressor gene, which encodes a mutant tumour suppressor that can inactivate the wildtype protein. Transgenic mice can be generated by pronuclear microinjection, in which a construct containing the gene of interest (transgene) is microinjected into the mouse oocyte after fertilisation and randomly integrates into the genome, usually in tandem copies. If the transgenic cells contribute to the germ line, the genetic change can be transmitted to the next generation, producing mice that are fully transgenic and establishing a strain. Many genes involved in cancer development are also essential for mouse development. Therefore, to prevent embryonic lethality and to restrict overexpression to specific tissues, the construct containing the gene of interest also contains promoter elements designed for spatial and temporal restriction of gene expression. For example, the Tet-On and Tet-Off systems (Baron and Bujard, 2000) promote gene expression in the presence and absence, respectively, of doxycycline, a non-toxic analogue of tetracycline, while fusing the gene of interest to a gene encoding the oestrogen receptor binding domain results in an inactive protein that is activated upon treatment with Tamoxifen (Eilers *et al.*, 1989).

Limitations of the microinjection method include the possibility that, because the transgene integrates randomly, it could disrupt other genes, resulting in a phenotype that does not reflect the function of the gene of interest (for review, see Muller, 1999). In addition, the tendency of the transgene to integrate in multiple copies could result in excessive overexpression that is toxic to the animal (Muller, 1999). However, transgenic mice have made a significant contribution to cancer research. In the earliest examples, mouse models were used to demonstrate the role of oncogenes in cancer. For example, tissue-specific overexpression of the *Myc* oncogene in mammary glands and B-cells resulted in the generation of mice prone to breast cancer (Stewart *et al.*, 1984) and lymphomas (Adams *et al.*, 1985), respectively. Overexpression of dominant-negative mutant tumour suppressor genes has also proved effective, e.g. a gene encoding mutant type II transforming growth factor beta (Tgfβ) receptor has been shown to accelerate chemically induced tumourigenesis in the mammary gland and lung (Bottinger *et al.*, 1997).

1.4.1.2.2 Endogenous models

A knockout mouse can be created to study the effect of inactivating a tumour suppressor gene. In this method, a targeting vector is transfected into embryonic stem (ES) cells, which are harvested from the inner cell mass of mouse blastocysts. The vector must share homology with the region of the mouse gene that is being targeted, i.e. the tumour suppressor gene of interest, and must also contain genes for selection, such that only cells in which the vector DNA has replaced the endogenous DNA by homologous recombination will survive. The surviving ES cells are injected back into a blastocyst, and will contribute to all cell lineages, including the germ line (Robertson *et al.*, 1986). The targeting vector can be engineered to knock out the whole gene or part of a gene, or small changes can be introduced into the gene sequence. Alternatively, the complete gene under the control of a strong promoter can be introduced to create a knockin mouse for overexpressing oncogenes. By targeting a single copy to the genome, this overcomes the problems associated with pronuclear microinjection. (For review, see Muller, 1999).

As with transgenic mice, mutations can be spatiotemporally regulated. Conditional mouse models frequently use the Cre-lox system from bacteriophage P1, in which Cre recombinase catalyses recombination between loxP sites (Sauer and Henderson, 1988), and the intervening DNA is deleted or inverted, depending on the orientation of the sites (Lakso *et al.*, 1992). loxP sites can therefore be placed on either side of a gene region to remove that region in the presence of Cre (Figure 1.7). Large-scale chromosomal deletions and inversions can also be generated by placing loxP sites further apart on the chromosome (Kmita *et al.*, 2000; Smith *et al.*, 2002), while chromosomal translocations can be created by placing a loxP site at each breakpoint (Forster *et al.*, 2003). Conditional oncogene expression can be achieved by inserting a stop cassette, which is flanked by loxP sites, between the promoter and the first exon such that Cre-mediated excision of the cassette results in expression of the gene (de Alboran *et al.*, 2001; Jackson *et al.*, 2001).

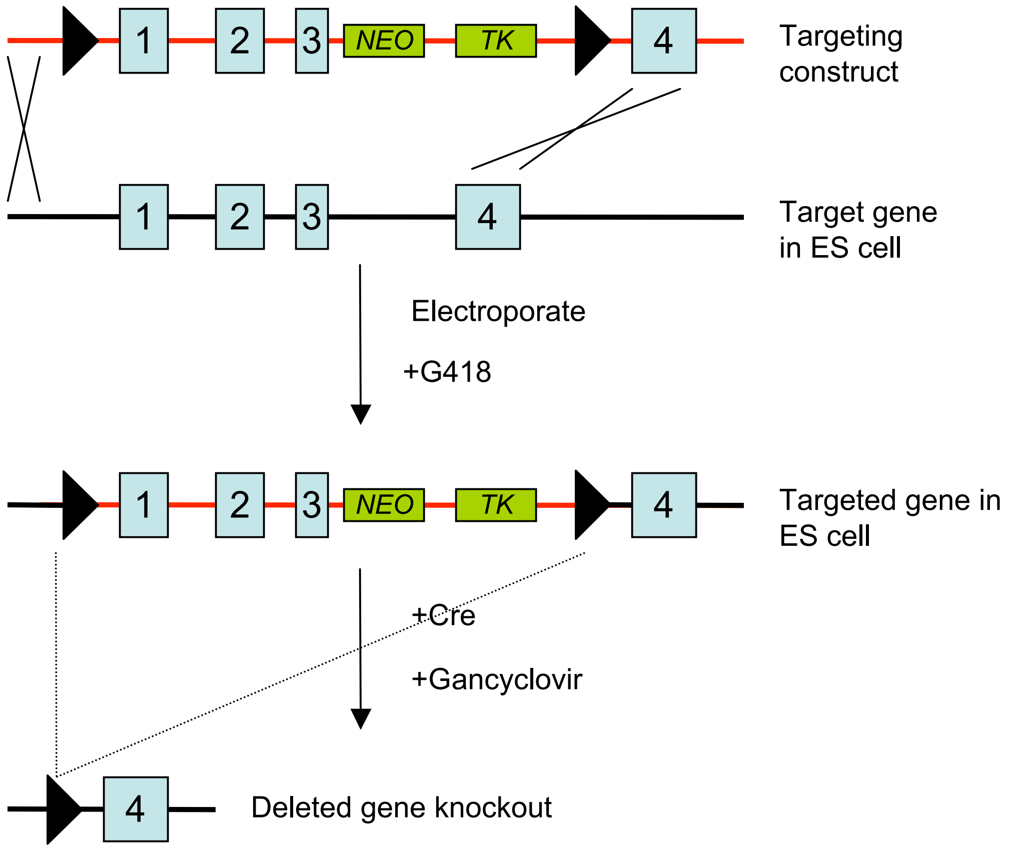

Figure 1.7. Generation of a conditional knockout allele in ES cells. A targeted gene construct is designed that contains loxP sites flanking the region of the gene to be deleted as well as genes for selection. Upon introduction into ES cells, DNA in the construct replaces endogenous DNA in the target gene by homologous recombination. The addition of G418 selects for cells that express the *Neomycin* gene, and therefore contain the knockout construct. The addition of Cre results in recombination between the loxP sites, removing the region of the gene containing exons 1, 2 and 3 and the *Neomycin* and *tk* genes. Gancyclovir kills cells expressing *tk*, and therefore selects cells in which recombination has occurred and the gene has been knocked out.

Unlike the conditional expression systems in transgenic mice, once Cre recombinase has been expressed, the change is irreversible, and there is evidence to suggest that Cre can be cytotoxic, perhaps due to recombination at pseudo-loxP sites (see Jonkers and Berns, 2002). In addition, the Cre-lox system cannot generate conditional point mutations, and this represents a significant limitation since point mutations and deletions do not always produce the same phenotype (Frese and Tuveson, 2007). However, the Cre-lox system has proved invaluable in creating models that would otherwise not arise or survive. For example, homozygous *Brca1* and *Brca2* knockouts die early in embryogenesis, and heterozygous mice are not tumour-prone, but mice harbouring a Cre-mediated deletion of *Brca1* (Xu *et al.*, 1999) or *Brca2* and *Trp53* (Jonkers *et al.*, 2001) in the adult mammary gland do develop mammary tumours. Likewise, *Trp53* mutations have been identified in many types of human cancer, but if *Trp53* is mutated in all cells, the mouse is most likely to develop lymphomas or sarcomas. Conditional *Trp53* mutations can be used to create models for human cancers that are driven by *TP53* mutation in other tissues (Jonkers and Berns, 2002). The Flp/FRT system from *Saccharomyces cerevisiae* is an alternative to Cre-lox that works in a similar way.

1.4.1.3 Mouse models in drug discovery

Mouse models that faithfully recapitulate human cancers are important for developing and testing therapeutic drugs. Studies on a mouse model for acute promyelocytic leukaemia (APL) have resulted in the development of an effective, retinoic-acid-based treatment for the disease (Lallemand-Breitenbach *et al.*, 1999; Soignet and Maslak, 2004). Mouse models can also be used to identify predictive markers of disease response and progression, and to understand drug toxicity and resistance. They have proved particularly useful in the study of oncogene addiction, which is an important consideration in drug target validation (see Section 1.2.7). Mouse models have demonstrated the requirement for persistent expression of *Hras*, *Myc*, *Bcr-Abl*, *Erbb2* and *Fgf7* in the maintenance of melanoma (Chin *et al.*, 1999), haematopoietic tumours (Felsher and Bishop, 1999), B-cell lymphoma and leukaemia (Huettner *et al.*, 2000), breast cancer (Xie *et al.*, 1999), and lung cancer (Tichelaar *et al.*, 2000), respectively.

1.4.1.4 Mouse models in cancer gene discovery

The methods described in Section 1.3 can also be applied to the identification of candidate cancer genes in the mouse. For example, array CGH has been used to identify regions of copy number change in mouse models of malignant melanoma (O'Hagan *et al.*, 2003) and pancreatic islet carcinomas (Hodgson *et al.*, 2001). However, as with human cancers, by the time the cancer has presented, it is difficult to distinguish the important driver mutations from the background of passenger mutations.

The genetically engineered mouse models discussed thus far are useful for studying the function of a particular gene or for representing a specific human cancer, but the tumours in these models do not evolve naturally. In general, the initiating event, i.e. the engineered mutation, is present throughout a tissue, whereas in natural tumourigenesis, the tumour develops from one mutated cell (see Section 1.2.3). Likewise, in mouse models used to study the combined action of multiple genes in cancer, the genes of interest are usually simultaneously mutated, whereas “natural” tumours progress through a multi-step process, where mutations are gradually acquired. Finally, many mouse models are designed to show high penetrance and short latency to keep costs down, but as a result they may not possess many of the co-operating oncogenic events that would eventually be acquired by a naturally evolving tumour (for review, see Frese and Tuveson, 2007; Sharpless and Depinho, 2006).

It is important that the mutations in mouse models used to identify novel cancer genes reflect the mutations found in human cancers, and this requires more accurate modelling of the natural evolution of tumours.

1.4.2 Forward genetic screens in the mouse

Forward genetic screens using somatic mutagens are a powerful approach for cancer gene discovery in which tumours undergo a process of evolution that mirrors that of human tumour formation. They allow for relatively unbiased, genome-wide identification of both novel cancer genes and collaborations between genes involved in cancer. Chemical mutagenesis is highly efficient but mutations are very difficult to identify. Insertional mutagenesis by retrovirus or transposon is an effective alternative approach in which the mutagen acts as a molecular tag for easy identification of the mutated allele.

1.4.2.1 Retroviral insertional mutagenesis

1.4.2.1.1 Mechanisms of mutagenesis

The slow transforming retroviruses murine leukaemia virus (MuLV) and mouse mammary tumour virus (MMTV) have been widely used for insertional mutagenesis in the mouse. Unlike acute transforming retroviruses, which induce tumours by expression of a viral oncogene, slow transforming retroviruses do not carry an oncogene, and tumours are induced by mutations caused by insertion of the retrovirus into the host genome. Consequently, tumours develop with a longer latency of 3-12 months, compared with 2-3 months for acute transforming retroviruses (Uren *et al.*, 2005). MMTV was identified as a causative agent in several strains of mice that were prone to mammary tumours, while MuLV was identified as a causative agent in the lymphoma-prone AKR mice (see Weiss, 2006). The principal dataset used in this thesis was generated using MuLV, and this mutagen is therefore the main focus of the background provided herein.

Retroviruses infect host cells by binding of the viral envelope proteins to cell surface receptors. Once the retrovirus has inserted into the host genome, forming a provirus, it will produce viral envelope proteins that occupy the cell surface receptors and prevent reinfection of the same cell. However, recombination with endogenous viral sequences results in the production of envelope proteins that bind to other receptors. This, combined with the fact that many proviruses have defective envelope coding sequences, enables retroviruses to reinfect the same cell, resulting in the accumulation of mutations. Mutations that confer a growth advantage on the cell co-operate in tumour formation, and the process therefore recapitulates the multi-step progression of human tumours (for review, see Mikkers and Berns, 2003; Uren *et al.*, 2005, see also Section 1.2.3).

The MuLV provirus consists of viral genes flanked by two long terminal repeats (LTRs), which are composed of three parts: U3, R and U5 (see Uren *et al.*, 2005; Figure 1.8). Elements within the LTRs drive expression of the viral genes but can also disrupt host genes. U3 contains enhancer and promoter sequences, while R contains transcription start and termination sites. High levels of viral transcription and, therefore, host gene disruption, will only occur in cells containing transcription factors that bind to U3. The propensity of MuLV to induce T- and B-cell lymphomas can be attributed to its dependence upon T- and B-cell-specific transcription factors, including *Runx*, *Ets* and *Myb* (see Neil and Cameron, 2002).

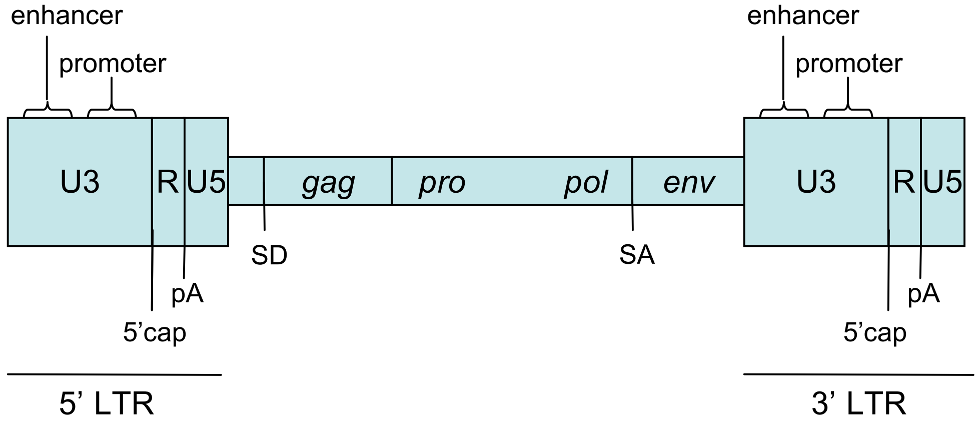

Figure 1.8. Structure of a retroviral provirus. The provirus contains two long terminal repeats (LTRs) flanking the genes required for viral assembly. Elements within the LTRs drive transcription of the viral genes but can also induce mutation of nearby cellular genes. Splicing of a viral splice donor (SD) or cryptic splice donor (not shown) to a splice acceptor or cryptic splice acceptor in the first intron or 5' UTR of a cellular gene results in the formation of a chimeric transcript, in which the cellular gene is coupled to the viral promoter. Splicing of a splice donor or cryptic splice donor in a cellular gene to a viral splice acceptor (SA) or cryptic splice acceptor (not shown) can cause premature termination of gene transcription owing to the presence of polyadenylation signals (pA) and cryptic polyadenylation signals (not shown) in the LTR. Adapted from Figure 1 of Uren *et al*. (2005). Figure is not to scale.

Retroviruses can mutate host genes in a number of different ways. The most common mechanism is enhancer mutation, where one of the U3 enhancers upregulates expression of host genes, which may be some distance away from the retroviral insertion (Figure 1.9A). Most proviruses causing enhancer mutations are found upstream of the mutated gene in the antisense orientation or downstream in the sense orientation. Several possible explanations for the directionality of the enhancer are that upregulation of the host gene may be impeded if the viral promoter intercepts the viral enhancer and host gene, or that viral enhancers may only be functional if they are not transcribed (Clausse *et al.*, 1993; see Uren *et al.*, 2005). *Myc* and *Gfi1* are frequent targets of enhancer mutation in retroviral insertional mutagenesis (Akagi *et al.*, 2004; Corcoran *et al.*, 1984; Selten *et al.*, 1984). *Myc* is mutated in many types of human cancer. It encodes a transcription factor that is thought to regulate the expression of 15% of all genes, including genes involved in cell division, cell growth and apoptosis (see Gearhart *et al.*, 2007). Gfi1 is a zinc finger transcriptional repressor that is involved in cell fate determination and differentiation, including in T- and B-cells (Rathinam and Klein, 2007; Yucel *et al.*, 2003).

An alternative mechanism of mutagenesis is promoter mutation, where the retrovirus inserts in the sense orientation into the promoter region of a host gene (Figure 1.9B). This uncouples the host gene from its own promoter and places it under the control of the viral promoters, resulting in the production of elevated levels of the wildtype protein from chimeric transcripts comprising part of the viral sequence and the complete coding region of the host gene (Mikkers *et al.*, 2002). Promoter mutations led to identification of *Evi1* as a potential oncogene (Copeland and Jenkins, 1990; Mucenski *et al.*, 1988a; Mucenski *et al.*, 1988b). *EVI1* encodes a zinc finger transcription factor that is frequently overexpressed in human myeloid malignancies. It is involved in several recurrent rearrangements, including 2 translocations that result in the fusion transcripts *AML1/MDS1/EVI1* and *ETV6/MDS1/EVI1*, where *MDS1* and *EVI1* are also expressed as a readthrough transcript in normal tissues (for review, see Wieser, 2007).

The retrovirus contains a polyadenylation signal within the R region of the LTR and a cryptic polyadenylation signal in the antisense orientation. Therefore, intragenic retroviral insertions in both orientations can cause premature termination of gene transcription. Insertions within the 3' UTR that truncate a transcript such that mRNA-destabilising motifs are removed will give rise to a more stable transcript and, as a result, increased levels of the wildtype protein (see Uren *et al.*, 2005; Figure 1.9C). Oncogenes

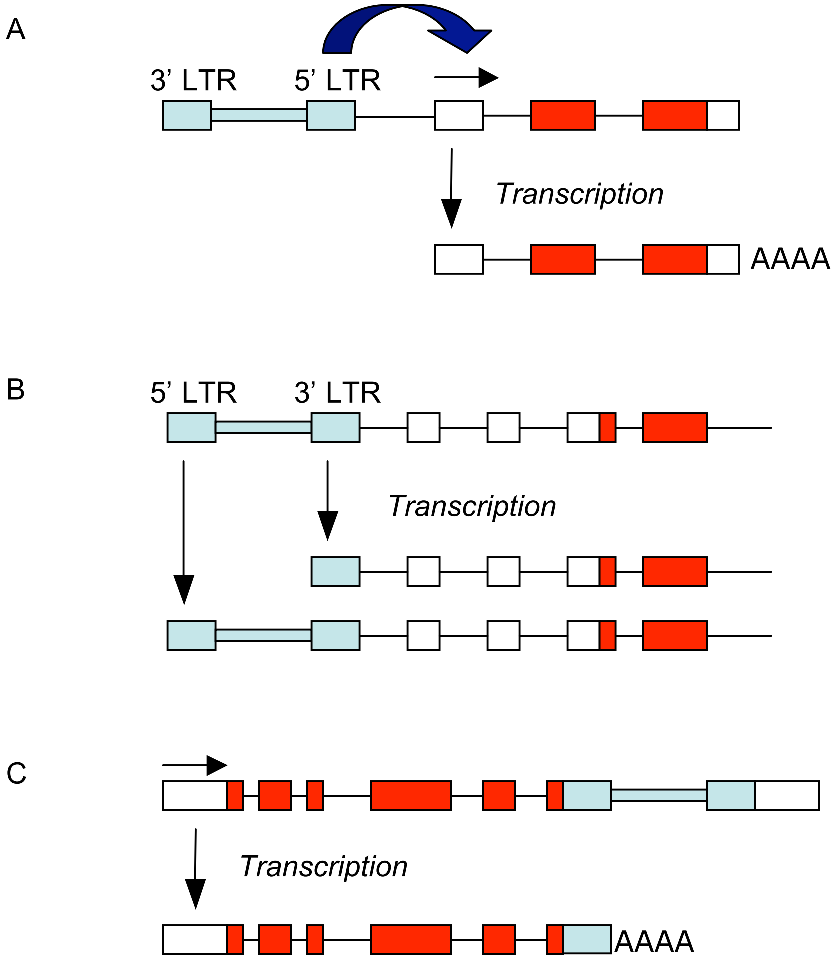

Figure 1.9. The mechanisms of mutagenesis of murine leukaemia virus include enhancer mutation (A), promoter mutation (B) and premature termination of gene transcription (C). The provirus is shown in blue; coding and non-coding exons are shown in red and white, respectively. **A.** An enhancer element in the 5' LTR of murine leukaemia virus (MuLV) can cause upregulation of nearby cellular genes. Oncogenic insertions of this type are most frequently found upstream and in the antisense orientation with respect to the cellular gene(s) that they are mutating. **B.** Insertion of MuLV into the promoter region of a cellular gene results in chimeric transcripts that are produced at higher levels than the endogenous gene transcript. **C.** Intragenic MuLV insertions can cause premature termination of gene transcription, resulting in either gene upregulation or gene inactivation. The figure shows an insertion within the 3' UTR region, which may remove mRNA-destabilising motifs, thereby stabilising the gene transcript. Adapted from figures in Uren *et al*. (2005).

Pim1 and *Mycn* are frequently mutated in this way (Cuypers *et al.*, 1984; Selten *et al.*, 1985; van Lohuizen *et al.*, 1989). *PIM1* encodes a serine/threonine kinase that is frequently overexpressed in human prostate cancer (Dhanasekaran *et al.*, 2001), while *MYCN* encodes a transcription factor related to *MYC* that is amplified in a variety of human tumours, most notably neuroblastomas (Brodeur *et al.*, 1984, 1985).

Intragenic insertions can also activate a gene by causing C-terminal or N-terminal truncation of the encoded protein. Insertions in oncogenes *Myb* and *Notch1* cause both N-terminal and C-terminal truncations (Rosson *et al.*, 1987; Uren *et al.*, 2005). C-terminally truncated Notch1 lacks the destabilising PEST domain and is therefore produced at increased levels, while N-terminal truncations remove the extracellular domain, resulting in a constitutively active intracellular domain expressed from the viral promoter or from a cryptic promoter in *Notch1* (Hoemann *et al.*, 2000). Activating mutations within the extracellular and PEST domains of NOTCH1 have been observed in human T-cell acute lymphoblastic leukaemia (Weng *et al.*, 2004), in which NOTCH1 plays an important role (see Section 1.2.5.1.4 for further details). Analysis of the distribution of insertions within an oncogene may therefore help to explain how the gene is mutated in human cancer.

Intragenic insertions may also cause gene inactivation, either through premature termination of transcription or by disrupting gene splicing (see Uren *et al.*, 2005). It is therefore possible to identify tumour suppressor genes by retroviral insertional mutagenesis, although they are found much less frequently than oncogenes because both copies of the gene must be inactivated. Mutation at the *Nf1* locus is observed in acute myeloid leukaemias in BXH2 mice (Largaespada *et al.*, 1996), which contain MuLV insertions (Bedigian *et al.*, 1984), while in an insertional mutagenesis screen of *Blm*-deficient mice, 11 genes met the criteria for tumour suppressor genes, including *Rbl1* and *Rbl2*, which are paralogues of *Rb1* (Suzuki *et al.*, 2006). *Blm*-deficient mice have a mutation in the RecQ protein-like-3 helicase gene (Ellis *et al.*, 1995) and show a predisposition to cancer due to increased frequencies of mitotic recombination (Luo *et al.*, 2000). There is an increased likelihood of finding tumour suppressor genes in these mice because they have a higher probability of a normal allele being lost so that only one insertion is required to inactivate the gene (Luo *et al.*, 2000). However, candidate tumour suppressor genes still only accounted for 5% of all genes identified in the screen by Suzuki *et al* (2006). In theory, insertional mutagenesis screens should have a better chance of finding haploinsufficient tumour suppressor genes, but none have yet been unambiguously identified (Uren *et al.*, 2005).

Insertional bias could also account for the paucity of tumour suppressor genes identified in retroviral screens. MuLV shows a strong preference for integration near to the transcription start sites of actively transcribed genes (Wu *et al.*, 2003) and is therefore less likely to disrupt a gene by intragenic insertion. However, it is possible that promoter mutations could also cause gene inactivation, as CpG islands in the retroviral LTRs are methylation targets, and DNA methylation could "spread" to CpG islands in the host gene, resulting in gene silencing (see Touw and Erkeland, 2007). Retroviruses prefer to insert into open chromatin (Muller and Varmus, 1994; Pryciak and Varmus, 1992), but different retroviruses show different target site preferences, suggesting that virus-specific interactions are involved (Mitchell *et al.*, 2004). DNA sequence does not seem to influence target site selection (Bushman *et al.*, 2005). The tendency for MuLV to insert into the promoter region indicates that the retrovirus interacts with cellular proteins bound near start sites (Mitchell *et al.*, 2004; Wu *et al.*, 2003).

1.4.2.1.2 Identifying candidate cancer genes

The retroviral insertions act as tags for identifying the mouse genes that are mutated by insertional mutagenesis, and sequencing of the mouse genome and the development of high-throughput genomic techniques have made it possible to identify hundreds or thousands of insertions in a single screen. Insertion sites were initially identified using methods that involved Southern blot analysis and genomic library screening, followed by genome walking to find the mutated gene (see Neil and Cameron, 2002; Uren *et al.*, 2005). However, these have been replaced by PCR-based methods, in which mouse genomic DNA flanking the insertion sites is amplified and is then mapped back to the genome. One such method, known as viral insertion site amplification (VISA) involves using a PCR primer designed to bind to the MuLV LTR and a degenerate, restriction-site-specific primer that enables amplification of the DNA between the insertion and a nearby restriction site (Hansen *et al.*, 2000; Weiser *et al.*, 2007). In inverse PCR and linker-mediated PCR-based methods, the genomic DNA is restriction-digested prior to PCR amplification.

In inverse PCR (Figure 1.10A), the digested genomic DNA is allowed to ligate to itself to form a circular template. PCR primers bind to the retroviral DNA and point out towards the genomic sequence, resulting in amplification of genomic DNA directly flanking the retrovirus (Ochman *et al.*, 1988; Triglia *et al.*, 1988). Only DNA fragments that are a suitable length for efficient circularisation and for PCR amplification will be detected (Uren *et al.*, 2005).

In linker-mediated PCR, rather than the digested DNA ligating to itself, it is ligated to a linker, and this enables shorter insertions to be identified. One primer is designed to bind to the linker, and the other binds to the retroviral sequence. A number of methods have been developed, each with a different approach for avoiding amplification of DNA that has linkers at both ends but contains no retroviral DNA. Vectorette PCR involves the use of a double-stranded linker with a cohesive end, designed for ligation to restricted DNA, and a central region with a mismatch (Riley *et al.*, 1990). The primer is the same sequence as the mismatched part of the upper strand, and this prevents initiation of priming from the linker until the complementary strand has been synthesised by priming from within the retroviral insertion. However, this method suffers from non-specific annealing of the primers and ‘end-repair’ priming, in which the ends of unligated linkers initiate priming and enable PCR amplification without involving the retroviral-specific primer (see Devon *et al.*, 1995). Any errors that cause amplification of DNA that is not flanking an insertion will lead to the false identification of insertion sites.

An improved method uses splinkerettes, which incorporate a hairpin structure on the bottom strand, rather than a mismatch sequence (Devon *et al.*, 1995; Figure 1.10B). The primer has the same sequence as the upper strand and, as with vectorette PCR, cannot anneal until the complementary strand has been synthesised. The stable hairpin does not enable end-repair priming and only the upper strand can act as a non-specific primer. In all the PCR-based methods, insertions are only identified if target sites for the chosen restriction endonuclease are close enough to the insertion for the intervening region to be amplified. Coverage can be improved by using multiple restriction endonucleases (Uren *et al.*, 2005).

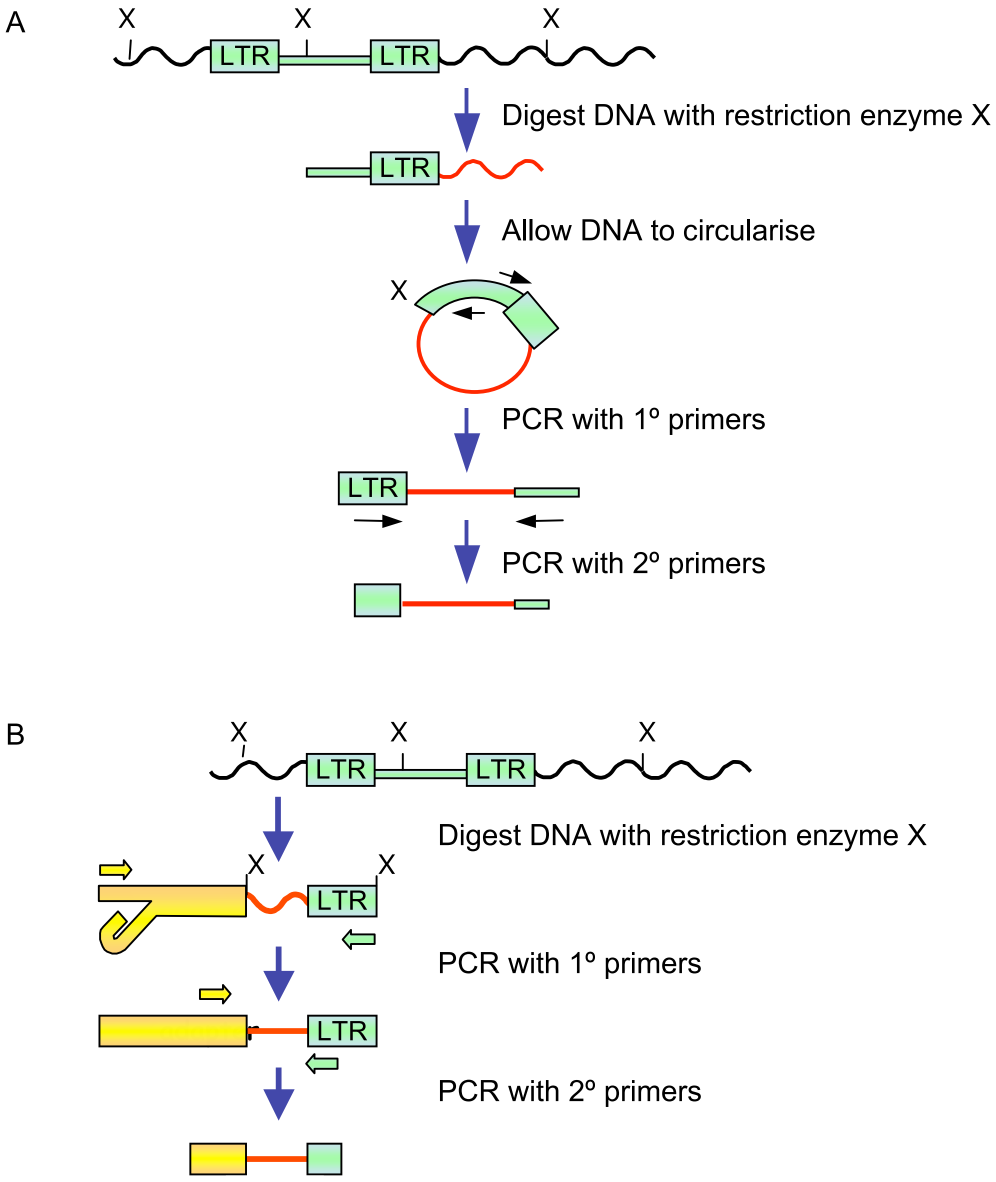

Figure 1.10 Isolation of retroviral insertion sites by inverse PCR (A) and splinkerette PCR (B). In inverse PCR, tumour DNA is digested using restriction enzyme X and the restricted DNA is allowed to circularise. Genomic DNA flanking retroviral insertions are amplified using PCR primers that bind within the insertion and point out towards the genomic DNA. A second round of PCR is performed using nested primers. The amplified DNA is sequenced and mapped to the mouse reference genome. Splinkerette PCR follows a similar procedure, except that instead of circularising the digested DNA, a splinkerette adapter (shown in yellow) is ligated to digested tumour DNA and genomic DNA flanking the retroviral insertions is amplified using PCR primers that bind to the adapter and the retroviral LTR.

Once the insertion-flanking genomic DNA has been amplified, the PCR products must be separated for sequencing. In the past, products were separated using agarose or polyacrylamide gels, but rare insertions are likely to be missed, and gel extraction is painstaking and subjective. An alternative method is to subclone the PCR products directly into a vector. By shotgun cloning the total mixture, it is possible to maintain the relative proportions of insertions from the starting material. However, it also means that more sequencing will be required to capture the rare insertions (see Uren *et al.*, 2005). The VISA approach sequences PCR products directly, without subcloning, which reduces the risk of sequencing contaminating products (Weiser *et al.*, 2007). The latest method uses massively parallel sequencing technology from 454 Life Sciences (http://www.454.com), in which fragmented genomic DNA is ligated to short adapters that are used for purification, amplification and sequencing. The DNA is denatured and immobilised onto beads, where PCR amplification and sequencing occur. This approach is extremely high-throughput, does not rely on cloning and is capable of detecting rare insertions. However, it can encounter problems when dealing with repetitive regions and long runs of a single nucleotide.

The next step is to map the sequenced DNA to the genome using a DNA alignment algorithm. For large screens, it is an advantage to be able to find high quality alignments quickly (Uren *et al.*, 2005). The Sequence Search and Alignment by Hashing Algorithm (SSAHA2, Ning *et al.*, 2001) converts the genome into a hash table, which can then be rapidly searched for matches. Sequences in the database (the mouse genome) are preprocessed into consecutive *k*-tuples of *k* contiguous bases and the hash table stores the position of each occurrence of each *k*-tuple. The query sequence (sequenced DNA) is also split into *k*-tuples and the locations of all occurrences of these sequences in the database, i.e. the "hits", are extracted from the hash table. The list of hits is sorted, and the algorithm searches for runs of hits in the database that match those in the query sequence. Having identified regions of high similarity, sequences are fully aligned using cross_match (Green, unpublished), which is based on the Smith-Waterman-Gotoh alignment algorithm (Gotoh, 1982; Smith and Waterman, 1981). Because the database is hashed, search time in SSAHA2 is independent of database size, provided *k* is not too small. SSAHA2 is therefore three to four orders of magnitude faster than the BLAST alignment algorithm (Altschul *et al.*, 1990), which scans the database and therefore performs at a speed that is directly related to database size (Ning *et al.*, 2001).

As the PCR mixture is shotgun cloned and preferably sequenced to a high depth, an insertion site may be represented by more than 1 sequence read. Reads from a single tumour that map to the same genomic region must therefore be clustered into single insertion sites. Like the mutations in human cancer, tumour DNA will contain both insertions that drive oncogenesis (oncogenic insertions) and insertions that are passengers (background insertions). In theory, most identified insertions should be oncogenic because these, and particularly the earliest events in tumourigenesis, should be present in most, if not all, tumour cells, whereas background insertions should be present in a smaller proportion of cells. However, background insertions that occur early in tumour development in a cell containing oncogenic insertions could also be highly represented in the final tumour (see de Ridder *et al.*, 2006).

Clustering of insertions from different tumours into common insertion sites (CISs) helps to distinguish oncogenic and background insertions. In theory, background insertions should be randomly distributed across the genome. Therefore, for small-scale screens, a gene in the vicinity of a cluster of insertion sites in different tumours is a strong candidate for a role in cancer. Methods for identifying statistically significant CISs, i.e. regions that are mutated by insertions in significantly more tumours than expected by chance, have involved generating a random distribution of insertions across the genome and obtaining an estimate of the number of false CISs in windows of fixed size using Monte Carlo simulation (Suzuki *et al.*, 2002) or the Poisson distribution (Mikkers *et al.*, 2002). These methods can be used to define the maximum window size in which insertions must fall to be considered non-randomly distributed. However, for larger scale screens, the window must be decreased to a size that is smaller than the spread of insertions within a single CIS so that many CIS are missed (de Ridder *et al.*, 2006). In addition, the above methods assume that insertions are randomly distributed and take no account of insertional biases, as mentioned in Section 1.4.2.1.1 (Wu *et al.*, 2006).

A more recent approach for CIS detection overcomes these problems by using a kernel convolution (KC)-based framework, which calculates a smoothed density distribution of inserts across the genome (de Ridder *et al.*, 2006). The scale (kernel size) can be varied so that CISs of varying widths can be identified. Decreasing the kernel size may identify separate CISs affecting the same gene, while increasing the kernel size will identify CISs where insertions are widely distributed in or around a gene. The method can be used for large-scale studies because it keeps control of the probability of detecting false CISs. The threshold for significant CISs is based on the alpha-level defined by the user and on a null-distribution of insertion densities obtained by performing random permutations. A background distribution, such as the location of transcription start sites, can be provided to correct for insertional biases. See de Ridder *et al.* (2006).

The final step is to identify the genes that are being mutated by insertions within CISs, which are known in this thesis as “CIS genes”. This may be relatively straightforward for intragenic insertions, but for enhancer mutations, which may have long distance effects, it is often difficult to identify the mutated gene unequivocally. Measuring the expression and transcript size of candidate genes in insertion-containing tumours can shed some light, but animal models and analysis of the orthologues in human cancer data are required for more conclusive evidence (Uren *et al.*, 2005).

A number of screens have been performed in recent years that have each identified hundreds of insertion sites (Hwang *et al.*, 2002; Johansson *et al.*, 2004; Li *et al.*, 1999; Lund *et al.*, 2002; Mikkers *et al.*, 2002; Slape *et al.*, 2007; Stewart *et al.*, 2007; Suzuki *et al.*, 2006; Suzuki *et al.*, 2002; Theodorou *et al.*, 2007; Uren *et al.*, 2008; Weiser *et al.*, 2007). The results of many screens have been collated and stored in the Retroviral Tagged Cancer Gene Database (RTCGD; http://rtcgd.abcc.ncifcrf.gov/) (Akagi *et al.*, 2004). At the time of writing, the database contains 503 CISs from 29 screens (database accessed May 2008). Users can search for individual genes of interest, or for CISs identified using particular mouse models and/or in particular tumour types. Genes with the most CISs are *Gfi1* and *Myc*, with 82 and 77 insertions across all screens, respectively.

1.4.2.1.3 Identifying co-operating cancer genes

Retroviral insertional mutagenesis is a powerful tool for identifying genes that collaborate in tumour development. Collaborations can be identified by analysing the co-occurrence of CISs in individual tumours. For example, proviral activation of *Meis1* and *Hoxa7* or *Hoxa9* is strongly correlated in myeloid leukaemias from BXH2 mice (Bedigian *et al.*, 1984; Nakamura *et al.*, 1996). *Meis1* and *Hoxa9* are targets of translocation in human pre-B leukaemia (Kamps *et al.*, 1990) and acute myeloid leukaemia (AML) (Calvo *et al.*, 2002), respectively, and they are frequently co-expressed in human AML (Lawrence *et al.*, 1999). Both genes encode homeodomain transcription factors that bind to Pbx, and

Meis1-Pbx and Hox-Pbx complexes have been shown to co-occupy the promoters of leukaemia-associated genes, such as *Flt3* (Wang *et al.*, 2006a).

A two-dimensional Gaussian Kernel Convolution method has recently been developed for identifying cooperating mutations in insertional mutagenesis data (de Ridder *et al.*, 2007). It is based on the kernel convolution framework used for identifying CISs (discussed in Section 1.4.2.1.2). The method has been applied to the data in RTCGD and, as well as finding previously characterised interactions, such as *Meis1* and *Hoxa9/Hoxa7*, it also finds novel interactions, such as *Rasgrp1* and *Cebpb*, which are both known to play a role in *Ras*-induced oncogenesis (de Ridder *et al.*, 2007).

As retroviral-induced tumours are oligoclonal, it is difficult to prove that tagged genes are in the same cell, and therefore that they collaborate (Largaespada, 2000). In an alternative approach, retroviral screens are performed on transgenic mice overexpressing known oncogenes, and knockout mice harbouring inactivated tumour suppressor genes, to identify genes that collaborate with the overexpression of oncogenes, and loss of tumour suppressor genes, respectively. For example, 35% of B-cell lymphomas generated in MuLV-infected *EμMyc* transgenic mice, in which *Myc* is overexpressed in B-cell progenitors under the control of the immunoglobulin heavy chain enhancer, have an insertion in *Pim1* or the polycomb group protein *Bmi1* (van Lohuizen *et al.*, 1991). Bmi1 collaborates with Myc by inhibiting *Cdkn2a* (*Ink4a/Arf*), and therefore inhibiting Myc-induced apoptosis (Jacobs *et al.*, 1999). In concurrence with these findings, *Myc* insertions were identified in 20% of tumours from MuLV-infected *Cdkn2a*-deficient mice, but none contained insertions in *Bmi1* (Lund *et al.*, 2002). Insertional mutagenesis also identifies genes that can functionally complement one another in tumour development. For example, in MuLV-infected *EμMyc* mice, activation of *Pim2* increases from 15% to 80% in compound mutant mice lacking *Pim1* expression (van der Lugt *et al.*, 1995), while *Pim3* is selectively activated in mice lacking *Pim1* and *Pim2* expression (Mikkers *et al.*, 2002). Pim1 is a coactivator of Myc that is required for expression of around 20% of all Myc target genes (Zippo *et al.*, 2007). Pim kinases also appear to suppress Myc-induced apoptosis, but it is not clear whether this mechanism or Myc coactivation is responsible for the co-occurrence of *Pim1* and *Myc* mutations observed in lymphomagenesis (for review, see Naud and Eilers, 2007)). *Pim1* also collaborates with *Myc* in human prostate cancers (Ellwood-Yen *et al.*, 2003).

Retroviral screening of a mouse model for human myeloid leukaemia has identified 6 CIS genes, including *Plagl* and *Plagl2*, which co-operate with the oncogenic fusion gene *CBFB-MYH11* (Castilla *et al.*, 2004). This screen used a replication-defective retrovirus, cloned amphotropic virus 4070A, to limit the number of mutations and therefore to show that mutation of only one or a few genes was sufficient to induce tumorigenesis. Other studies using replication-competent viruses report 3-6 insertions in a single tumour (Mikkers *et al.*, 2002; Suzuki *et al.*, 2002) but, as mentioned above, retroviral-induced tumours are oligoclonal and it is therefore difficult to make a reliable estimate of the number of insertions in a tumour clone (see Neil and Cameron, 2002).

1.4.2.1.4 Generating tumours of different types

As discussed in Section 1.4.2.1.1, the dependence of retroviruses on cell-type-specific transcription factors limits the range of tumours that they can induce. There have been some successful attempts to alter the propensity of MuLV for T-cell lymphomas by using an *EμMyc* transgenic mouse, which results in predominantly B-cell lymphomas (van Lohuizen *et al.*, 1991), and by expressing platelet derived growth factor B-chain (*PDGFβ*) from an MuLV-based retrovirus to generate mice with glioblastomas, which require activation of PDGF receptors for tumourigenesis (Johansson *et al.*, 2004). Mutations in the retroviral LTR may also lead to a change in tumour type, but manipulated viruses have a tendency to revert to wildtype (Uren *et al.*, 2005). In addition, MuLV and other retroviruses cannot infect nondividing cells, and infection is inefficient in slowly replicating cells and in tissues that have a basement membrane or mucin layer (Wang *et al.*, 2002a; Yamashita and Emerman, 2006). Transposon-mediated insertional mutagenesis is an alternative method that provides the possibility of generating a wider spectrum of tumours.

1.4.2.2 Transposon-mediated insertional mutagenesis

Like retroviruses, transposons are genetic elements that can mobilise within the genome. They are classified according to their mechanism of transposition. DNA transposons move by a “cut and paste” mechanism, in which they are excised from one site in the genome and integrated into another. Retrotransposons transpose via an RNA intermediate and are classified into LTR retrotransposons, which encode reverse transcriptase and transpose in a similar manner to retroviruses, and non-LTR retrotransposons, which are transcribed by host RNA polymerases and may or may not encode reverse transcriptase (Kapitonov and Jurka, 2008).

1.4.2.2.1 Sleeping Beauty

While DNA transposons are actively mobile in plants and invertebrates, all of the elements that have been so far identified in vertebrates are non-functional (Uren *et al.*, 2005). However, they can be mobilised in the mouse by using an invertebrate DNA transposon or by reconstructing a degenerate vertebrate transposon. *Sleeping Beauty* (SB) is a synthetic transposon derived from dormant DNA transposons of the Tc1/Mariner family in the genomes of salmonid fish. An active transposon, named SB10, was synthesised by directed mutagenesis on the basis of a consensus sequence obtained by aligning 12 degenerate transposon sequences from 8 species (Ivics *et al.*, 1997). SB consists of two inverted repeat/direct repeat (IR/DR) elements of ~230 bp each, flanking a cargo sequence (Collier *et al.*, 2005; Figure 1.11). Transposition occurs via binding of a transposase enzyme to two sites in each IR/DR (Izsvak *et al.*, 2000). All four binding sites are required for transposition and, in general, the closer the IR/DRs, the higher the transposition efficiency (Izsvak *et al.*, 2000). Higher levels of transposition have been achieved by introducing point mutations into the transposase, producing, for example, the SB11 (Geurts *et al.*, 2003) and SB12 (Zayed *et al.*, 2004) transposases.

The utility of SB for oncogenic insertional mutagenesis was first demonstrated in two studies published in 2005 (Collier *et al.*, 2005; Dupuy *et al.*, 2005). In both studies, transposons were introduced into mice by pronuclear injection of a linear plasmid containing one copy of the transposon, which forms a multicopy concatemer of variable length at a single site in the mouse genome. SB was mobilised by crossing these mice to mice expressing a transposase from a ubiquitous promoter. Collier and coworkers (2005) used a transgene containing the SB10 transposase under the control of the CAGGS promoter to mobilise around 25 T2/Onc transposons (Figure 1.11), while Dupuy *et al.* (2005) used the more active SB11 version knocked into the endogenous *Rosa26* locus to mobilise 150-350 copies of the T2/Onc2 transposon.

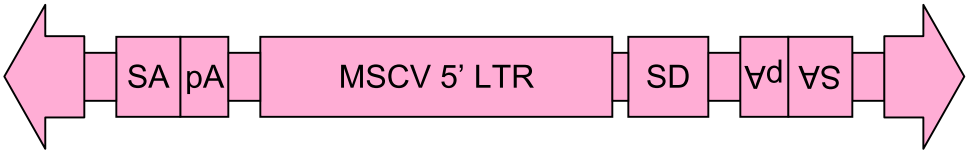

Figure 1.11. Structure of the *Sleeping Beauty* transposon. The presence of splice acceptors (SA) and polyadenylation signals (pA) in both orientations enables premature termination of gene transcription from intragenic insertions in both orientations. The transposon also contains the murine stem cell virus (MSCV) 5' LTR and a splice donor (SD) site that can induce promoter mutations in cellular genes. Elements for mutagenesis are flanked by 2 IR/DR elements, shown as arrows, which are required for transposon mobilisation. Redrawn and adapted from Figure 1a of Collier *et al*. (2005).

T2/Onc and T2/Onc2 were engineered to contain elements for mutagenesis much like those in retroviruses. The cargo of both transposons contains the 5' LTR of the murine stem cell virus (MSCV) followed by a splice donor, as well as splice acceptors followed by polyadenylation sites in both orientations. The transposons are therefore capable of disrupting genes by promoter mutation, N-terminal and C-terminal truncation and gene inactivation but, unlike retroviruses, they show low enhancer activity (Dupuy *et al.*, 2005). T2/Onc and T2/Onc2 are essentially the same, except that T2/Onc2 contains a larger fragment of the *Engrailed* splice acceptor and the IR/DRs have been optimised for transposase binding (Dupuy *et al.*, 2005). In the study by Dupuy and coworkers (2005), there was a high rate of embryonic lethality and, of the 24 T2/Onc2;Rosa26SB11 mice that survived to weaning, all developed cancer, most commonly T-cell lymphomas but also other haematopoietic malignancies plus a few cases of medulloblastomas and intestinal and pituitary neoplasias. Some mice had 2 or 3 types of cancer and all died within 17 weeks. In contrast, in the study by Collier *et al.* (2005), mice on a wildtype background did not develop tumours, but those on an *Arf*-null background developed sarcomas at an accelerated rate. The difference between the two studies most likely reflects the differences in transposon copy number and in transposase expression and activity (Collier and Largaespada, 2007). Transposase expression in CAGGS-SB10 mice has since been shown to be low and variegated in most tissues, probably due to epigenetic silencing of the transgene, while transposase expression is high in nearly all cell types in Rosa26SB11 mice (Collier and Largaespada, 2007). However, transposase is expressed in the testes of CAGGS-SB10 mice, which show high rates of transposition in the male germline (Collier and Largaespada, 2007; Dupuy *et al.*, 2001).

Transposons, like retroviruses, can be used to identify co-operating cancer genes. For example, *Braf* was frequently mutated in *Arf*-null mice, suggesting that these genes co-operate in tumour formation (Collier *et al.*, 2005), while of the six T-cell tumours containing *Notch1* mutations, three also contained insertions mutating *Rasgrp1*, and 2 of these contains *Sox8* mutations, suggesting that these three genes also co-operate (Dupuy *et al.*, 2005).

While a number of the genes identified in the haematopoietic malignancies of T2/Onc2;Rosa26SB11 mice had been previously identified in retroviral mutagenesis, other genes had not (Dupuy *et al.*, 2005). This indicates that transposon-mediated mutagenesis is a complementary approach for cancer gene discovery, and may reflect differences in insertional bias. While MuLV shows a strong preference for inserting near transcription start sites (Wu *et al.*, 2003), SB shows a less pronounced preference and shows no preference for actively transcribed genes (Yant *et al.*, 2005). SB inserts at TA dinucleotides and therefore shows a bias towards AT-rich sites, particularly those with the consensus sequence ANNTANNT (Carlson *et al.*, 2003; Vigdal *et al.*, 2002). However, most significant is the strong tendency of SB to transpose to sites close to the concatemer. This phenomenon, known as "local hopping", results in a non-random distribution of insertions that hampers CIS detection. Another potential hindrance to cancer gene identification is the ability of transposons to excise themselves and reinsert multiple times. SB leaves a small footprint upon excision, and it is possible that, at least in exons, this could continue to cause gene disruption that would not be identifiable (Collier and Largaespada, 2007). Likewise, the excision in some cells of transposons that had been critical for tumour development could result in a more heterogeneous tumour in which cancer gene identification would be more complicated. However, it is possible that such an event would be deleterious and that the cell would be eliminated (Collier and Largaespada, 2007) and, as SB transposition efficiency is higher for methylated (Yusa *et al.*, 2004) and heterochromatic (Ikeda *et al.*, 2007) transposons, excision of transposons involved in gene disruption may be relatively rare. A further drawback of SB, and possibly other DNA transposons, is that transposition induces genomic rearrangements, including deletions and inversions near to the transposon concatemer, and tumourigenesis could therefore be initiated by genes disrupted by these rearrangements rather than by mobilised transposons (Geurts *et al.*, 2006).

One of the key benefits of using a transposon such as SB for insertional mutagenesis is that the mutagenic elements can be modified to control the types of mutation that occur. For example, modifying the cargo to enable only truncating mutations could increase the likelihood of identifying tumour suppressor genes (Collier and Largaespada, 2007). Tissue-specific promoters can be integrated as cargo, making transposons an attractive mutagen for cancer gene discovery in specific cancer types (Dupuy *et al.*, 2006). Spatial and temporal transposition could also be achieved by introducing a lox-stop-lox cassette between the SB transposase promoter and cDNA, such that transposition is induced upon the addition of Cre (Dupuy *et al.*, 2006).

Identification of cancer genes in SB mutagenesis follows much the same procedure as for retroviruses. Largaespada and Collier (2008) have developed a technique that uses linker-mediated PCR, as described in Section 1.4.2.1.2, but that enables PCR amplification of DNA flanking both sides of the transposon to maximise coverage. Primers were designed to bind to the IR/DR sites and to synthetic adapters. Unlike in retroviral mutagenesis, tumour cells contain a concatemer of non-transposed elements. To avoid repeated cloning of the junctions between these elements, "blocking" primers can be used that bind to the plasmid DNA flanking each transposon in the concatemer but that have blocked 3' ends to prevent polymerase extension. Alternatively, after linker ligation, the DNA can be redigested with an endonuclease that cuts within the flanking plasmid DNA so that the primer binding sites are separated onto different molecules. (See Largaespada and Collier, 2008).

1.4.2.2.2 Alternative mutagens for transposon insertional mutagenesis

The active invertebrate transposons *piggyBac* and *Minos* are the only other DNA transposons that have so far been mobilised in the mouse (Collier and Largaespada, 2007). The *piggyBac* transposon, isolated from the cabbage looper moth, mobilises in mouse somatic cells and in the germline, and it can carry a larger cargo than SB (Ding *et al.*, 2005). The coding sequence of *piggyBac* has been codon-optimised to enable higher levels of transposition in the mouse, and inducible versions have been generated by fusing the transposon to the ERt^2 oestrogen receptor ligand-binding domain (Cadinanos and Bradley, 2007). Unlike SB, it shows a strong preference for inserting into genes in the mouse (Ding *et al.*, 2005) and in human cell lines (Wilson *et al.*, 2007). The *Minos* transposon, from *Drosophila hydei*, has attracted interest because it shows a low insertional bias and high transposition efficiency in a range of animals (for review, see Pavlopoulos *et al.*, 2007). However, it has so far shown only weak *in vivo* activity in the mouse (Drabek *et al.*, 2003; Zagoraiou *et al.*, 2001).

Retrotransposons are also gaining attention as potential insertional mutagens. Long interspersed nuclear elements (LINEs) are non-LTR retrotransposons that are transcribed into mRNA by RNA polymerase II and encode two proteins that are essential for transposition (Moran *et al.*, 1996): a protein that binds to single-stranded RNA (Hohjoh and Singer, 1997) and a protein with reverse transcriptase and endonuclease activity (Feng *et al.*, 1996; Mathias *et al.*, 1991). 17% of the human genome is composed of LINE-1 (L1) elements (Lander *et al.*, 2001). Transcription of endogenous L1 elements is generally inefficient but there are a small number of highly active "hot L1s", which were used to generate a transgenic mouse model of L1 retrotransposition that showed a higher frequency of de novo somatic L1 insertions (Babushok *et al.*, 2006). A 200-fold increase in transposition in the mouse germline has also been achieved by codon optimisation of the human L1 coding region (Han and Boeke, 2004). L1 mobilises by a "copy and paste" mechanism. It is therefore an attractive mutagen for forward genetic screens because, unlike DNA transposons, it is capable of self-expansion and the original insertion remains intact, aiding identification of mutated genes (Bestor, 2005; Collier and Largaespada, 2007). In addition, it appears to show no preference (An *et al.*, 2006), or only a slight preference (Babushok *et al.*, 2006), for inserting into genes and there is no local hopping because the RNA intermediate must exit and re-enter the nucleus before inserting into the genome. However, most L1 insertions are truncated at the 5' end (Babushok *et al.*, 2006), potentially resulting in the loss of promoters, splice acceptors and polyadenylation signals required for mutagenesis (Collier and Largaespada, 2007). Controlled insertional mutagenesis using L1 derivatives has not yet been reported and *Sleeping Beauty* remains the preferred transposon for cancer gene discovery.

1.5 Cross-species comparative analysis for cancer gene discovery

Important biological sequences, such as gene coding regions and regulatory elements, are conserved in evolution. Cross-species comparative sequence analysis may therefore potentially help in the characterisation of known cancer genes. Comparison of intronic sequences in human and mouse *BRCA1* led to the identification of two evolutionarily conserved regulatory elements in the second intron that, when mutated, had opposite effects on gene expression (Wardrop and Brown, 2005). However, cross-species comparative analysis also provides an extremely powerful approach for identifying novel genes and gene collaborations involved in cancer formation. As discussed in Section 1.3, the human cancer genome is highly complex. Many genes and pathways have been implicated in tumourigenesis, and most human cancers exhibit genomic instability, leading to the acquisition of genetic alterations that drive tumourigenesis but also many passenger mutations that do not contribute to the tumour phenotype. Distinguishing driver and passenger mutations is a major challenge. However, the molecular mechanisms that govern important biological processes are conserved in evolution, and cancer-associated mutation data from other species can therefore be used as a filter for identifying genes that represent strong candidates for a role in human cancer.

Genome-wide expression data for human tumours can be difficult to interpret, and a number of studies have therefore used cross-species comparative analysis to identify conserved expression signatures that are important in tumourigenesis. Expression profiles of intestinal polyps from patients with a germline mutation in *APC* were compared to those from *Apc*-deficient mice and the conserved signature showed an over-representation of genes involved in cell proliferation and activation of the Wnt/β-catenin signalling pathway (Gaspar *et al.*, 2008). Likewise, comparison of expression profiles for human lung adenocarcinoma and a mouse model of *Kras2*-mediated lung cancer led to the identification of a *KRAS2* expression signature that was not identified by analysing *KRAS2*-mutated human tumours alone (Sweet-Cordero *et al.*, 2005). More recently, a mutated *Kras*-specific signature that can be used to classify human and mouse lung tumours on the basis of their *KRAS* mutation status has been identified by comparing *KRAS*-mutated human cancer cells to mouse somatic cells containing knocked-in mutant *Kras* (Arena *et al.*, 2007).

Mouse prostate cancers induced by human *MYC* have an expression signature that defines a set of "*Myc*-like" human prostate tumours and includes overexpression of the oncogene *Pim1* (Ellwood-Yen *et al.*, 2003). Rat prostate tumours also have a similar expression profile to human prostate tumours, and have been used to identify conserved genes that are differentially expressed in both species in response to treatment with the chemopreventive agent Selenium (Schlicht *et al.*, 2004). The mouse is therefore not the only cancer model that has been used for cross-species comparison. The greater the evolutionary distance between the species, the greater the likelihood that conserved changes in gene expression contribute to the cancer phenotype. An expression signature in zebrafish liver tumours is more consistently associated with human liver tumours than with other human tumour types and, since human and zebrafish are distantly related, genes in the conserved signature are strong candidates for a role in cancer development (Lam *et al.*, 2006).

Another approach for cross-species analysis involves comparing the CGH profiles of human tumours to the CGH profiles of tumours generated from a mouse model of the corresponding human cancer. Such studies take advantage of the conserved synteny between the human and mouse genomes (Waterston *et al.*, 2002). Comparison of CGH profiles for human neuroblastomas with profiles for tumours and cell lines from a *MYCN* transgenic mouse model of neuroblastoma have shown that many genetic aberrations are conserved between species (Cheng *et al.*, 2007; Hackett *et al.*, 2003). Likewise, 80% of aberrations detected by array CGH in tumour cells of the mouse model for epithelial ovarian cancer are conserved in human epithelial ovarian cancer (Urzua *et al.*, 2005), and epithelial carcinomas in mice with telomere dysfunction show numerous copy number changes in regions syntenic to those in human cancers (O'Hagan *et al.*, 2002). Zender and coworkers (2006) used array CGH to identify regions of copy number change in the tumours of a mouse model for hepatocellular carcinoma. The CGH profiles were compared to array CGH data for human hepatocellular carcinomas to identify minimally conserved amplicons, and genes that showed increased expression in both species were chosen as candidate cancer genes. The authors identified 2 oncogenes, *cIAP1* and *Yap*, that act synergistically in a focal amplicon on mouse chromosome 9qA1, which is syntenic to an 11q22 amplicon in human tumours. Kim *et al.* (2006b) used a comparable approach to identify *Nedd9* as a candidate for a role in promoting melanoma metastasis. A focal amplicon comprising 8 genes, including *Nedd9*, was identified on chromosome 13 in 2 metastatic cell lines derived from a *Ras* mouse model of nonmetastic melanoma. 36% of metastatic melanomas contained a much larger amplicon in a syntenic region on human chromosome 6p25-24, and 35-52% of metastatic melanomas showed significant overexpression of *NEDD9*, with more advanced tumours showing higher levels.

Comparison of human cancers with mouse models of cancer relies on the use of mouse models that accurately recapitulate the human cancer (Tomlins and Chinnaiyan, 2006). While *cIAP1* and *Yap* overexpression was found to be important in *p53*$^{-/-}$;*Myc*-induced hepatoblasts in the study by Zender *et al.* (2006), neither gene contributed to tumourigenesis in *p53*$^{-/-}$;*Akt* or *Ras* hepatoblasts. Likewise, *Nedd9* did not contribute to melanoma metastasis in the absence of *Ras* or *Raf* activation (Kim *et al.*, 2006b). Cross-species comparison of genomic profiles for a particular cancer may therefore require some prior knowledge of the genetic events that drive tumourigenesis in that cancer so that an appropriate mouse model can be generated. However, cross-species analysis can also facilitate the selection of a suitable mouse model. Lee and coworkers (2004) used unsupervised hierarchical clustering of expression data from human and mouse hepatocellular carcinomas to identify the mouse models that provided the best fit for human cancers. Mouse and human tumours that clustered together due to similar expression profiles also shared phenotypic characteristics, such as proliferation rate and prognosis (Lee *et al.*, 2004). Most genetically engineered mouse models do not show the high levels of chromosome instability associated with human cancers. Mice that are engineered with telomere dysfunction, or defects in DNA damage checkpoints or DNA repair, may therefore represent better models for comparative oncogenomics (Maser *et al.*, 2007). Comparative analysis of copy number alterations in chromosomally unstable murine T-cell lymphomas and human solid tumours identified recurrent aberrations in the mouse that are conserved in human T-cell acute lymphoblastic leukaemias but also in other human tumour types (Maser *et al.*, 2007).

Candidate cancer genes can also be identified by comparing expression and CGH profiles for human tumours with mouse insertional mutagenesis screens. Genes in expression signatures associated with distinct subclasses of human acute myeloid leukaemia were significantly correlated with genes nearest to insertion sites in a Graffi 1.4 MuLV mouse model and with candidate leukaemia genes in BXH2 and AKXD mouse models (Erkeland *et al.*, 2006). There was little overlap between the candidates identified by Graffi 1.4 and BXH2/AKXD, demonstrating that retroviral screens involving multiple models and viruses may be required for a more effective cross-species comparison (Touw and Erkeland, 2007). Amplified regions in human pancreatic cancer have also been shown to contain more CIS in retrovirus-induced murine lymphomas and leukaemias than expected by chance (Aguirre *et al.*, 2004). As discussed in Section 1.4, insertional mutagenesis "tags" the mutated gene, therefore facilitating cancer gene identification. In contrast, copy number alterations in human cancer can be very large, encompassing many genes, and no systematic approach currently exists for identifying the critical genes within these regions (Degenhardt *et al.*, 2008). Thus comparative analysis of oncogenic insertions in mouse tumours and CGH data for human tumours is potentially a very powerful approach for narrowing down the candidates in regions of copy number change.

1.6 Aims of this thesis

The elucidation of the human genome sequence and the advent of high-throughput technologies for characterising cancer genomes have led to the discovery that the cancer genome is far more complex than previously thought. Genome-wide, cancer-associated mutation datasets can be generated at increasing speed and diminishing cost, yet identifying the mutations that contribute to the cancer phenotype remains a challenge. Integrative analyses, particularly cross-species comparisons, provide a means of distinguishing likely driver mutations from the background of passenger mutations that arise in unstable cancer genomes. The identification of cancer genes in regions of copy number change is especially problematic since such regions are often large and encompass many potential candidates. Forward genetic screens are purported to be a powerful tool for cancer gene discovery in the mouse, but how relevant are they to human cancer?

This thesis describes work undertaken to compare large-scale datasets generated by mouse insertional mutagenesis and CGH analysis of human cancer cell lines. The main aims of this project are to narrow down the candidate cancer genes in regions of copy number change in human cancers and, in so doing, demonstrate the utility of forward genetic screens in the mouse for the identification of human cancer genes. Chapter 2 describes the steps taken to identify mouse candidate cancer genes from a retroviral insertional mutagenesis dataset generated from 1,005 mouse tumours and a smaller transposon-mediated insertional mutagenesis dataset generated from 73 mouse tumours. Chapter 3 describes detailed analyses of the mouse candidate genes, including comparisons with numerous human and mouse cancer-associated mutation datasets, as well as an analysis of the types of mutations occurring in each candidate and the identification of collaborating cancer genes. Chapter 4 describes the work undertaken to identify regions of copy number change in Affymetrix 10K SNP array CGH data for 713 human cancer cell lines, and then to identify candidate cancer genes within these regions by comparison with mouse candidates from the retroviral screen. In Chapter 5, higher resolution Affymetrix SNP 6.0 CGH data generated from a subset of the same cell lines is used, again to identify putative cancer genes, but also for comparison with the lower resolution data to demonstrate the superiority of the high-resolution data for cancer gene discovery. Analyses that attempt to identify genes that co-occur, and therefore potentially co-operate, in both human and mouse cancers are also described. Finally, conclusions drawn from the analyses are presented in Chapter 6.

Chapter 2 Identifying insertion sites and candidate cancer genes by insertional mutagenesis in the mouse

2.1 Introduction

When a retroviral or transposon insertional mutagen inserts into the mouse genome, it acts as a molecular tag that facilitates the identification of genes that it disrupts. As discussed in Section 1.4.2.1.2, the elucidation of the mouse genome sequence and the development of high-throughput, PCR-based technologies for insertion site identification have allowed for larger scale mutagenesis screens that can identify a higher proportion of insertions across larger numbers of tumours. 1,005 mouse tumours were generated in a retroviral insertional mutagenesis screen performed by the Netherlands Cancer Institute (NKI). Murine leukaemia virus (MuLV) was used as the insertional mutagen, and insertions into the mouse genome were identified using splinkerette PCR (see Section 1.4.2.1.2). In a separate study at the University of Minnesota, 73 mouse tumours were generated by insertional mutagenesis using the *Sleeping Beauty* T2/Onc transposon (see Section 1.4.2.2.1). Genomic DNA flanking the retroviral and transposon insertion sites was sequenced at the Wellcome Trust Sanger Institute. This chapter begins with a description of the retroviral and transposon insertional mutagenesis datasets. While I did not contribute to the generation of tumours or sequence reads, all statistics are the result of my own analyses. A dataset of known cancer genes, compiled by the Sanger Institute Cancer Genome Project, is also described. This is followed by an account of the work undertaken to process the sequence reads into insertion sites, to filter out erroneous reads and insertion sites, and to measure the coverage of the screen. A relatively high proportion of reads could not be mapped, and the nature of non-mapping reads was therefore investigated. The remainder of the chapter focuses on the methods used to identify candidate cancer genes in the vicinity of mapped insertions. The identification of genes that are being mutated by retroviral insertions is complicated by the presence of enhancer mutations that may act at long range (see Section 1.4.2.1.2). Insertions were assigned to genes by defining rules based on an analysis of the distribution of insertions around mouse genes. Statistically significant common insertion sites (CISs) were defined using Monte Carlo simulations (Suzuki *et al.*, 2002) and a kernel convolution-based framework (de Ridder *et al.*, 2006), and CIS genes identified by the two approaches were compared. Data from the retroviral screen forms the principal mouse dataset used in this thesis, and is therefore discussed in greater detail than data from the transposon screen. The main steps involved in identifying candidate cancer genes from retroviral sequence reads are summarised in Figure 2.1. Unless otherwise stated, *P*-values provided in this chapter were generated using the Chi-squared test for independence.

2.2 Description of the datasets

2.2.1 The retroviral dataset

Mice of the *FVB* strain were engineered with a range of genetic backgrounds in order to identify cancer genes that collaborate with the loss of tumour suppressor genes (see Section 1.4.2.1.3). 1,005 tumours were generated, of which 22.7%, 12.5% and 23.0% were on a $p19^{ARF-/-}$ ($Cdkn2a^{-/-}$), $p53^{-/-}$ or wildtype genetic background, respectively. The remaining tumours were generated on a background deficient in *p15*, *p16*, *p21* or *p27*, or a combination of these (Table 2.1A). Equal numbers of males and females were used (500 each of males and females, 1 hermaphrodite and 4 unknown). The vast majority (at least 90.9%) of tumours originated in the spleen, thymus or lymph nodes (Table 2.1B). The 1-tailed Fisher Exact Test was performed to determine whether genetic background or gender was associated with particular tumour types. Wildtype and $p19^{-/-}$ genetic backgrounds were over-represented in tumours of the thymus (P=4.67x10^{-6} and P=9.87x10^{-5}, respectively), while among tumours of the spleen, there was an over-representation of $p53^{-/-}$ (P=5.05x10^{-5}) as well as wildtype and $p19^{-/-}$ genetic backgrounds (P=0.0240, and P=1.84x10^{-4}, respectively). Lymph node tumours were over-represented in $p16^{-/-}p19^{-/-}$ mice (P=1.10x10^{-5}) and in mice with a deficiency in *p21* or *p21* and *p27* ($p21^{-/-}$, P=2.32x10^{-13}; $p21^{-/-}p27^{+/-}$, P=4.34x10^{-4}; $p21^{-/-}p27^{-/-}$, P=2.17x10^{-4}; $p21^{+/-}p27^{+/-}$, P=0.0128). It is possible that these results represent a subjective bias in the selection of tumours. Alternatively, they may indicate that different genetic backgrounds are predisposed to different tumour types. Most striking was the over-representation of the $p16^{-/-}p19^{-/-}$ genotype among tumours in the liver (P=1.04x10^{-32}). At least 24 of the 33 liver tumours have been identified as tumours of the liver nodule. These are commonly observed in $p16^{-/-}p19^{-/-}$ mice infected with MuLV and may be lymphomas that have spread to the liver or they may be histiocytic sarcomas, which are a poorly-defined class of haematopoietic neoplasm (Lund *et al.*, 2002). There was no significant difference

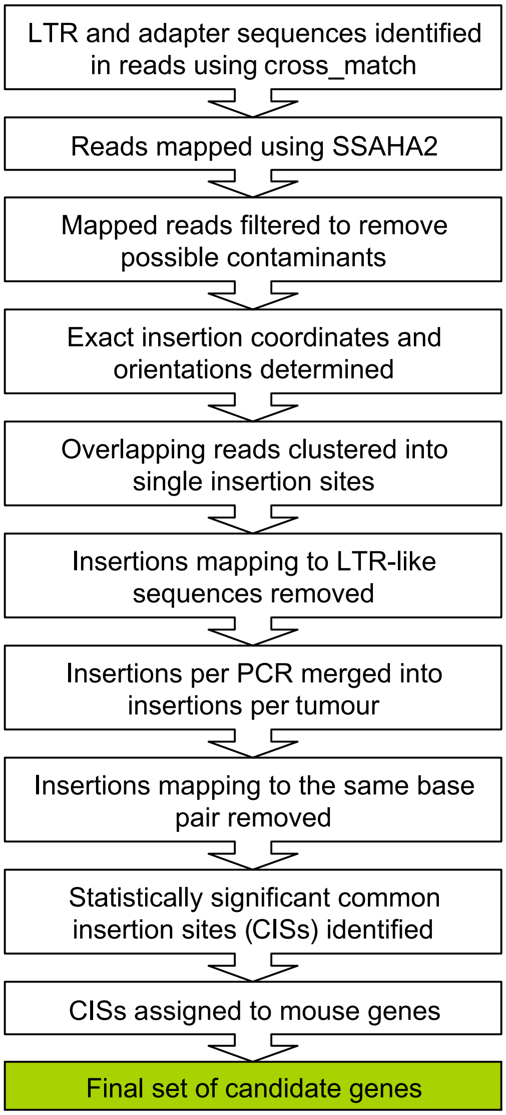
LTR and adapter sequences identified in reads using cross_match
Reads mapped using SSAHA2
Mapped reads filtered to remove possible contaminants
Exact insertion coordinates and orientations determined
Overlapping reads clustered into single insertion sites
Insertions mapping to LTR-like sequences removed
Insertions per PCR merged into insertions per tumour
Insertions mapping to the same base pair removed
Statistically significant common insertion sites (CISs) identified
CISs assigned to mouse genes
Final set of candidate genes

Figure 2.1. Workflow for identifying mouse candidate cancer genes from sequencing reads generated in a retroviral insertional mutagenesis screen.

A

Genotype	Number of tumours
wildtype	231
p19-/-	228
p53-/-	126
p16-/-, p19-/-	91
p15-/-	55
p21-/-, p27+/-	54
p21-/-	43
p27+/-	38
p21-/-, p27-/-	36
p27-/-	36
p16+/-, p19+/-	26
p21+/-, p27+/-	17
p15-/-, p21-/-	15
p21+/-, p27-/-	5
p53+/-	2
p21+/-	2
Total	1005

B

Tissue	Number of tumours
spleen	468
thymus	227
lymph node	125
spleen; lymph node	71
unknown	52
liver	33
thymus; spleen	15
spleen nodule	4
spleen; liver	3
kidney nodule	2
scapular tumour	1
uterine tract	1
uterine tumour	1
fascial lymphoma	1
uterine tumour; lymph node	1
Total	1005

C

Tissue	Number of tumours
spleen	38
thymus	22
lymph node	10
brain tumour	2
unknown	1
Total	73

Table 2.1. Characterisation of the insertional mutagenesis datasets. (A) The number of tumours from mice with different genetic backgrounds in the MuLV screen. (B) The number of tumours of each tissue type in the MuLV screen. (C) The number of tumours of each tissue type in the *Sleeping Beauty* T2/Onc screen.

between the number of males and females with tumours from different tissues or genetic backgrounds.

Following the isolation of tumour DNA, most samples were subjected to two separate splinkerette PCRs using different restriction enzymes, *Sau3AI* and *Tsp509I*, in order to increase the number of insertions that could be identified in the screen (see Section 1.4.2.1.2). The PCR products were shotgun cloned, and 96 reads were sequenced per PCR. Everything described from this point onwards is the result of my own work. The reads were converted to a CAF (Common Assembly Format) file, which contains the DNA sequence, base quality, and the coordinates of sequencing and cloning vector sequences within the read. The CAF file was then converted to FASTA format, in which the vector sequences were masked. The resulting dataset comprised 159,303 sequence reads from 2,060 PCRs. 14,767 reads from 199 PCRs were discarded because they were of unknown identity or had been flagged as invalid due to possible sample mix-up, no obvious tumour when killed, or contaminated or low quality PCR. The remaining 144,536 reads included 134,985 that were generated from 1,734 PCRs performed on 1,005 mouse tumours. For 62% of tumours, the dataset contained reads obtained from 2 PCR experiments, i.e. using both restriction enzymes, while for 33% of tumours, reads were only available for a single experiment. The remaining 5% of tumours were subjected to 3 or 4 PCRs, in which additional reactions using *Sau3AI* and/or *Tsp509I* were performed. The number of reads per tumour is shown in Figure 2.2. To facilitate the identification of PCR artefacts, 1,180 reads were also generated from 24 PCRs performed on uninfected mice. Finally, 8,371 reads were generated from 103 PCRs performed on samples that were harvested from mice 5 or 10 days post-MuLV infection. There has been limited time for cell re-infection, and thus for tumour initiation and progression, in these "short infection time" mice. A high proportion of insertions in samples from these mice are therefore expected to map to sites in the genome where the virus prefers to insert ("hotspots") and that may not contribute to tumourigenesis.

Cross_match (Green, unpublished) was used to identify and mask the retroviral LTR (5'-GCTAGCTTGCCAAACCTACAGGTGGGGTCTTTCA-3') and splinkerette adapter (5'-CCACTAGTGTCGACACCAGTCTCATTCAGCCAC-3') in order to prevent erroneous mapping of reads to regions of the mouse genome that resemble these sequences. The minimum length of the perfectly matching sequence (minmatch) and the minimum alignment score (minscore) were each set to 10. These parameters were used

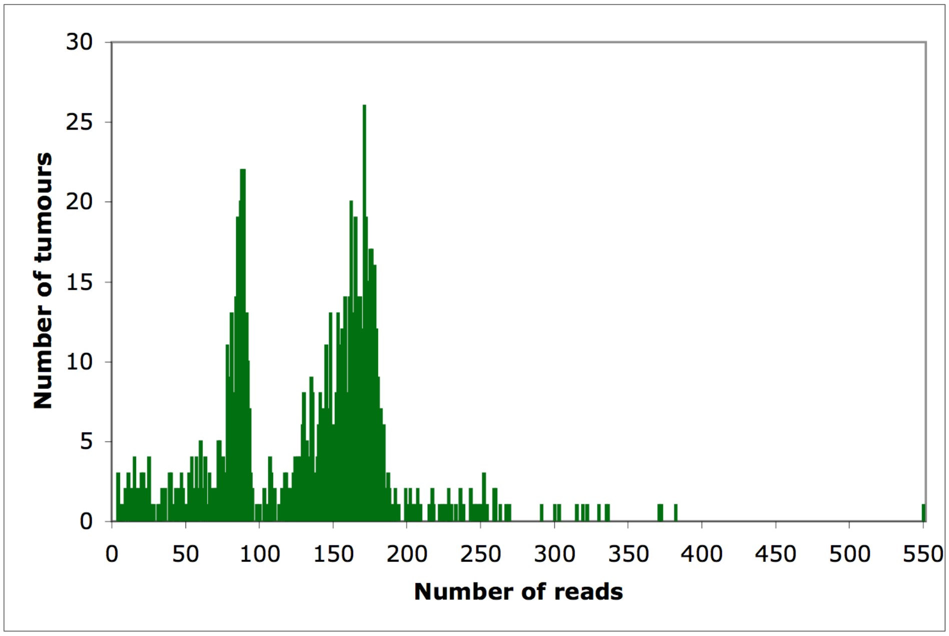

Figure 2.2. The number of sequence reads per tumour before mapping. Up to 96 reads were sequenced for each PCR. The bimodal distribution reflects the fact that 62% of tumours were subjected to 2 linker-mediated PCRs, while for 33% and 5% of tumours, 1 PCR or more than 2 PCRs, respectively, were performed.

for all cross_match runs, unless otherwise specified. Among the reads from tumour DNA, 110,318 (81.7%) contained LTR and adapter sequences, 9,534 (7.1%) contained an LTR but no adapter, 12,592 (9.3%) contained an adapter but no LTR, and 2,541 (1.9%) contained neither.

2.2.2 The Sleeping Beauty dataset

This smaller screen comprised 73 tumours, of which 60 were from wildtype mice on a mixed *C57BL/6J/FVB* background, and 13 were from *Bloom* (*Blm*)-deficient mice from the same strain. *Blm*-deficient tumours may be more likely to harbour mutations that inactivate tumour suppressor genes (see Section 1.4.2.1.1). These tumours, and 31 wildtype tumours, were generated from a transposon array (LC76) located on chromosome 1. The remaining wildtype tumours were generated from an array (LC68) on chromosome 15. As in the MuLV screen, tumours developed almost exclusively in the spleen, thymus and lymph node (Table 2.1C) since mice have a propensity for these tumour types.

Insertions were cloned using linker-mediated PCR in which genomic DNA flanking both sides of the insertion was amplified to maximise insertion site identification (see Section 1.4.2.2.1). The restriction enzymes *BfaI* and *NlaIII* were used to clone DNA flanking the 5’ and 3’ IR/DRs, respectively. As in the retroviral screen, PCR products were shotgun cloned and 96 reads were sequenced. All work described hereafter is my own. The initial dataset comprised 16,674 sequences. Although steps were taken to minimise the amplification of transposons within the concatemer (see Section 1.4.2.2.1), the sequence data inevitably contain some reads that map to the concatemer. Transposons in the concatemer are flanked by the sequence 5’-TATAGGGATCC-3’ and therefore any reads containing this sequence are likely to represent transposons that have not mobilised. 89 concatemer sequences were removed using cross_match (Green, unpublished).

The presence of the transposon IR/DR provides evidence that the genomic DNA is directly flanking an insertion. Using cross_match, IR/DR elements were identified and masked in 15,630 reads (94.2% of the total), and the rest were discarded. The linker, which was identified in 12,209 (78.1%) of the remaining reads, and extra vector sequence from the PROMEGA pGEM-T easy vector T7 promoter-multiple cloning site-SP6 promoter were also screened out with cross_match. 3,716 reads (23.8%) contained fewer than 25 bp of unmasked sequence after screening, and as these would be too short for mapping, they were removed from the dataset. Tumour details were not available for a further 1,123 reads, and so these were also removed. The final dataset comprised 10,791 reads generated from 138 PCRs. This included 60 tumours for which genomic DNA flanking both sides had been amplified and sequenced, and 11 tumours for which only one side had been amplified. For the remaining 2 tumours, both sides had been amplified, and PCR had been performed twice on one or both sides.

2.2.3 Known cancer genes in the Cancer Gene Census

The Cancer Gene Census is a list of genes for which there is strong evidence of a role in cancer (Futreal *et al.*, 2004; see Section 1.2.5.2). The complete working list dated 13/02/2007 was downloaded from http://www.sanger.ac.uk/genetics/CGP/Census/. The Ensembl (Hubbard *et al.*, 2007) Perl Application Programming Interface (API) was used to extract the Ensembl identifiers for each gene in the list from Ensembl version 48. Ensembl provides annotation on a selection of eukaryotic genomes, and it has been used throughout this project to obtain information about the mouse and human genomes. The API provides standardised methods for accessing data in the Ensembl MySQL databases through Perl scripts and it insulates developers from changes at the database level. From the 363 genes in the Cancer Gene Census, 354 human Ensembl genes were identified. 352 mouse Ensembl genes have a human orthologue in the Cancer Gene Census. 314 mouse genes have an orthologue with somatic mutations in cancer and 67 have an orthologue with germline mutations, including 32 that have an orthologue with both mutation types. The orthologues of 285 mouse genes bear mutations that are dominant at the cellular level, 66 bear recessive mutations, of which 2 are X-linked, and 1 has both dominant and recessive mutations. 205 have been implicated in leukaemia and/or lymphoma, 102 have been implicated in epithelial tumourigenesis and 84 have been implicated in mesenchymal tumourigenesis. The most common type of mutation is translocation, which affects the orthologues of 263 mouse genes. A list of the human cancer genes with mouse orthologues is provided in Appendix A.

2.3 Mapping the sequence reads using SSAHA2

As discussed in Section 1.4.2.1.2, SSAHA2 (Ning *et al.*, 2001) is a fast DNA alignment algorithm that is suited to mapping large numbers of insertions to the mouse genome.

The parameters of SSAHA2 were adjusted to maximise the number of mapped reads, and therefore to identify as many insertions as possible. A test set of 25,000 reads from the retroviral screen was mapped to the NCBI m34 mouse genome assembly. SSAHA2 preprocesses the query sequence (the read) and the subject (sequences in the NCBI m34 database) into consecutive *k*-tuples of *k* contiguous bases, called the word size or *k*-mer. Lowering the *k*-mer increases the sensitivity, and therefore yields more hits, but it also increases CPU time, and a *k*-mer of 13 or 14 is generally recommended for large databases, such as genome assemblies. The default *k*-mer of 12 was used for all runs of SSAHA2, since this offers a small gain in sensitivity without impacting too heavily on the speed. The "seeds" parameter defines the number of exact words that must match in the subject. Lowering the seeds increases the sensitivity, resulting in a higher proportion of low (<95%) identity and ambiguous mappings, but also more high identity unambiguous mappings (Table 2.2A). Initially, seeds 3 was chosen because seeds 2 yielded only 8 additional high identity unambiguous mappings and required more CPU time. By default, sequences are processed into consecutive *k*-mers with no overlap. Reducing the parameter "skip" increases the overlap between *k*-mers and should provide greater sensitivity. For seeds 3, decreasing skip to 4 (8 base overlap) and 6 (6 base overlap) did not increase numbers of high identity, single mapping reads. For higher seeds, numbers did increase but were lower than for seeds 3 alone (Table 2.2B). SSAHA2 with seeds 3 yielded more mappings than NCBI BLASTN (Altschul *et al.*, 1990; Table 2.2A) and was significantly faster. BLASTN parameters were set for moderately sized (~500 bp) genomic DNA (-G 1, -E 3, -W 30, -F 'm D', –U, -e 1e-20).

The full set of 144,536 retroviral reads was mapped to the NCBI m36 mouse build using SSAHA2 with seeds 3 and default values for all other parameters. Alignments with low identity were not segregated in this larger analysis because they may simply represent sequencing reads of poor quality and, if they are erroneous, they should be picked up in the filtering process (see Section 2.5). 86,290 reads (59.7%) mapped to a single location, 28,484 (19.7%) mapped to multiple locations, and 29,762 (20.6%) did not map at all. Further runs of SSAHA2 were performed with lower seeds to map as many of the unmapped reads as possible. 3,866 (13.0%) of unmapped reads could be mapped using seeds 2, and the same results were obtained with seeds 1. This is surprising, since the difference between seeds 3 and 2 was minimal when the 25,000-read test dataset was used. In the test set, analysis with seeds 2 did increase the number of alignments with <95% identity (Table 2.2A), and it is therefore likely that a proportion of the additional

A

		SSAHA2 seeds			
Mapping	**5**	**4**	**3**	**2**	**BLAST**
Single	13470	13894	14158	14164	14010
None	7060	6002	4971	3870	5960
Low	1837	2110	2414	3253	1365
Multiple	2633	2994	3457	3713	3665
Total	25000	25000	25000	25000	25000

B

	seeds 3			seeds 5		
Mapping	**default**	**skip 4**	**skip 6**	**default**	**skip 4**	**skip 6**
Single	14158	13699	13854	13470	13875	14004
None	4971	4187	4044	7060	3959	5096
Low	2414	3301	3281	1837	3301	2413
Multiple	3457	3813	3821	2633	3865	3487
Total	25000	25000	25000	25000	25000	25000

Table 2.2. The number of MuLV reads mapped using SSAHA2, with varying values for parameters seeds and skip, and BLASTN. (A) Lowering the number of seeds increases the number of reads mapped by SSAHA2. (B) Increasing the overlap between *k*-mers decreases the number of reads mapped using seeds 3 but increases the number mapped using seeds 5. Mapping types are Single (read maps to a single location in the genome), None (read unmapped), Low (read maps with an identity lower than 95%) and Multiple (read maps to multiple locations in the genome).

unambiguous mappings obtained using SSAHA2 with seeds 2 on the entire dataset have a low identity. The difference may also reflect developments in the algorithm in the time between the two analyses. The default minimum Smith-Waterman score is 30, and reducing this to 20 further increased the number of unmapped reads that could be mapped to a single location using seeds 2 to 4,382 (14.7%). The final set of mappings comprised 90,672 reads (62.7%) that mapped unambiguously and 29,769 (20.6%) that mapped to multiple locations. 24,095 (16.7%) remained unmapped.

Based on the observations for the retroviral dataset, the 10,791 reads of the *Sleeping Beauty* dataset were mapped to NCBI m36 using SSAHA2 with default parameters plus seeds 2 and score 20. 5,470 (50.7%) mapped to a single genomic location, 1,859 (17.2%) mapped to multiple locations, and 3,462 (32.1%) did not map at all.

2.4 Accounting for unmapped reads

Even after maximising the number of reads that could be mapped using SSAHA2, there was still a high proportion of unmapped reads in both the retroviral and *Sleeping Beauty* datasets. The lengths of the 96,072 single-mapping, and 24,095 non-mapping, retroviral reads are shown in Figures 2.3A and 2.3B, respectively. Since it is not known which part of the read, if any, is genomic DNA, all bases that were not masked as vector, LTR or linker were counted. 2,143 (8.9%) of the unmapped reads were exactly 132 base pairs in length and a high proportion of these shared an identical sequence flanked by LTR and splinkerette sequences. One read of length 132 bp was submitted to SSAHA2 and BLASTN on the Ensembl website (http://www.ensembl.org/). As expected, there were no matches to NCBI m36 using SSAHA2 with near exact or no optimisation. Using BLASTN optimised for near exact matches (–E 10 –B 100 –filter dust –RepeatMasker –W 15 –M 1 –N -3 –Q 3 –R 3), there were 96 hits, all of which were low scoring. The hit with the lowest E-value and *P*-value (both $4.2\text{x}10^{-7}$) was an alignment of 50 bp with a score of 22 and 86% identity to chromosome 8:126312491-126312540. The sequence was also submitted to the Ensembl Trace Server (http://trace.ensembl.org), which contains millions of single-pass DNA sequencing reads from over 1,000 different species. The full length of the read matched with 100% identity to 6 clones from the free-living nematode species *Pristionchus pacificus*. Since it was unclear how DNA from this organism would have become incorporated into the screen, a 132 bp read was also submitted to NCBI VecScreen (http://www.ncbi.nlm.nih.gov/VecScreen/), which

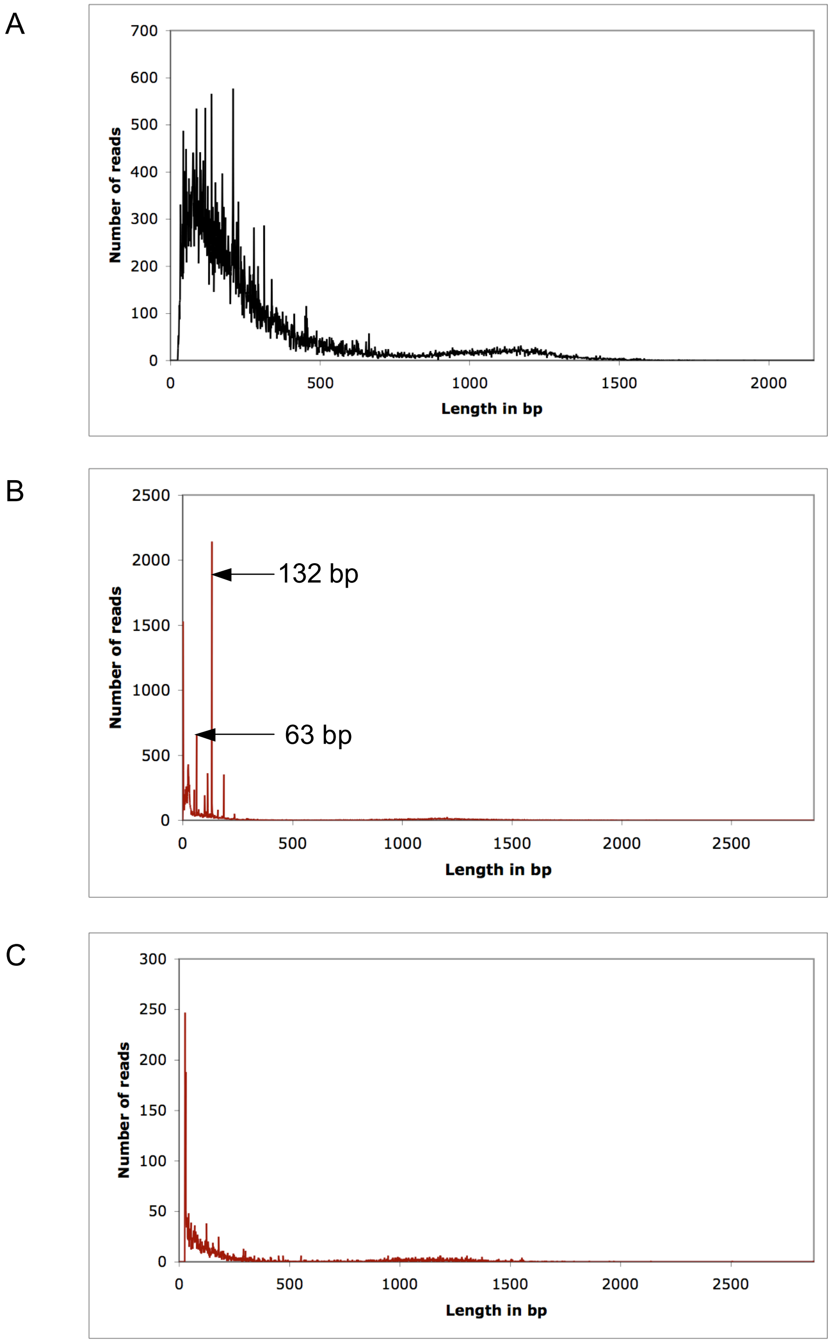

Figure 2.3. The lengths of retroviral reads that are unambiguously mapped (A), unmapped (B), and unmapped and uncharacterised (C). The reads of length 63 bp and 132 bp, which underwent further investigation, are shown.

searches for vector contamination in nucleic acid sequences by using BLAST to query the UniVec database. The entire sequence aligned to the MuLV retroviral vector pLNL6 with 100% identity and no gaps. Therefore, it appears that the 132 bp sequences are composed entirely of retroviral sequence, to which an adapter has been ligated.

There were 657 sequences of length 63 bp. One such sequence was submitted to BLASTN optimised for near exact matches, and one hit – an alignment of 18 bp with 100% identity to chromosome 15:90360616-90473373 – was obtained. The highest scoring hit obtained in a search against the Trace Server was just 75.9% identity, to a sequence from an unknown source. A VecScreen search revealed a 100% identity match along the entire length of the unmasked sequence to the cloning vector pBR322. Since these reads contain an adapter sequence, it is likely that they represent contamination during linker-mediated PCR.

Other reads containing the pLNL6 and pBR322 vector sequences were identified using cross_match. 19.4% of unmapped reads contained the pLNL6 sequence, while a further 4.8% contained the pBR322 sequence. In contrast, only 0.32% and 0.08% of reads mapping to a single location had matches to pLNL6 and pBR322, respectively. RepeatMasker (Smit *et al.*, 1996-2004) was used to identify repeat regions within the remaining unmapped reads. 13.6% contained low-complexity regions. Such regions are difficult to sequence, and these reads may have failed to map because they were not correctly sequenced. Alternatively, low-complexity regions may be the result of polymerase stuttering, where the polymerase transcribes the same nucleotide multiple times during PCR amplification, and the read may therefore no longer bear a close enough resemblance to the corresponding region of the mouse genome. The proportion of low-complexity regions was significantly lower in reads that mapped unambiguously (6.2%, P=0).

A further 26.6% of unmapped retroviral reads were below the minimum length (25 bp) that could be mapped using SSAHA2 with the chosen parameter values. Of the remaining sequences, 0.73% comprised more than 50% Ns (i.e. unknown nucleotides), and 1.2% contained other types of repeat element identified by RepeatMasker. This compared with 0.11% and 0.18%, respectively, for reads mapping to a single location (P=0 for both). The 2-tailed Fisher Exact test was used to determine whether there was any significant difference between the numbers of each type of repeat identified by

RepeatMasker in the mapped and unmapped reads. All types of low-complexity region were over-represented in the unmapped reads, as were simple repeats and a selection of retrovirus-related repeat elements (HAL1, GSAT_MM, L1MEd, LTR/ERV1, MuLV-int, MuRRS-int, RLTR4_MM-int and RLTR6-int, see Table 2.3A). This supports the theory that many of the reads could not be mapped either because they contain low complexity regions, or because they contain retroviral sequence and may not contain any genomic DNA. There were also numerous under-represented repeat elements among the unmapped reads (Table 2.3B). These included elements that one would expect to find in genomic DNA, such as 4.5SRNA and LINEs and SINEs, and elements that are specific to the genomes of rodents, such as the endogenous LTR MTE2a, and to the mouse in particular, such as the SINE B2_Mm2.

In summary, 15,996 (66.4%) of unmapped reads contained vector sequences or sequences of low complexity, low quality or short length (Table 2.4). The remaining 8,099 reads were searched against the Ensembl Trace Server using SSAHA2 with seeds 5. 5.1% had matches in the archive, of which 90.5% had matches to sequences of mouse origin. All of the non-mouse matches had an identity of less than 91%, except one, which matched with 100% identity to 2 sequences, with trace names rtn1ut06.g and rtn1yp83.g, from *Rattus norvegicus*. As rat and mouse are closely related, it is possible that this read does contain DNA from the mouse genome, but that it does not align to the mouse genome because of a genome assembly error. 245 reads mapped to mouse sequences in the Ensembl Trace Server with greater than 90% identity. The 8,099 uncharacterised reads were also searched against NCBI m36 using NCBI BLASTN. 2,901 (35.8%) had BLAST hits, but most had very low scores, just above the score threshold (half had a score of less than 33, 90% had a score of less than 59; Figure 2.4). The mapping algorithms of BLASTN and SSAHA2 therefore show small differences in output that may not significantly affect the final set of reliable mappings. Of the reads with BLAST hits, 76 also had hits to mouse sequences in the Ensembl Trace Server. However, there were also 295 reads that had hits to mouse sequences in the Trace Server but no BLAST hits. Again, these potentially represent sequences that have been incorrectly omitted from the mouse build.

Of the 5,198 (20.9%) remaining non-mapping reads, 4,363 were from tumours, 62 were from non-infected mice and 773 were from short infection time mice. Reads from control samples were highly over-represented (P=6.62x10^{-172}). There was also a highly

A

Repeat Element	*P*-value
A-rich	0
AT_rich	0
C-rich	7.64E-241
CT-rich	1.94E-09
G-rich	0
GA-rich	0
GC_rich	0
GSAT_MM	3.76E-04
HAL1	2.76E-05
L1MEd	2.02E-03
LTR/ERV1	9.60E-21
MuLV-int	1.87E-26
MuRRS-int	2.23E-08
RLTR4_MM-int	5.83E-45
RLTR6-int	8.18E-03
Simple_repeat	0
T-rich	8.15E-301
polypurine	4.87E-47
polypyrimidine	4.47E-07

B

Repeat Element	*P*-value	Repeat Element	*P*-value
4.5SRNA	7.90E-04	LTR/ERVL	1.21E-10
B1F	3.74E-22	LTR/MaLR	1.21E-82
B1F1	6.41E-08	Lx8	2.67E-13
B1F2	1.03E-14	Lx9	9.04E-08
B1_Mur1	1.07E-11	MIR	3.08E-20
B1_Mur2	1.05E-18	MIR3	8.14E-06
B1_Mur3	5.32E-05	MIRb	8.48E-23
B1_Mur4	1.09E-07	MTD	4.17E-13
B1_Mus1	5.17E-05	MTE-int	2.38E-08
B1_Mus2	5.55E-09	MTE2a	2.02E-05
B2_Mm2	4.07E-03	MTE2b	1.30E-06
B3	5.49E-49	MTEa	8.20E-07
B3A	3.50E-25	ORR1D2	3.79E-10
B4	1.93E-20	ORR1E	1.21E-04
B4A	7.06E-45	Other	2.10E-03
BC1_Mm	8.00E-03	PB1	2.48E-12
DNA/MER1_type	1.44E-27	PB1D10	1.99E-31
DNA/MER2_type	5.23E-04	PB1D9	5.99E-09
ID	8.22E-03	RMER15	2.11E-04
ID4	3.79E-10	RMER30	1.26E-04
ID4_	2.25E-10	RSINE1	7.00E-60
ID_B1	5.29E-69	SINE/Alu	9.88E-176
L1M	4.16E-04	SINE/B2	3.31E-73
L1M2	4.90E-03	SINE/B4	1.45E-182
L1MC3	8.00E-03	SINE/ID	9.77E-23
L1_Rod	3.19E-05	SINE/MIR	3.24E-47
L2	4.73E-07	THER1_MD	3.16E-03
LINE/L1	1.16E-65	URR1A	5.15E-03
LINE/L2	5.59E-13	URR1B	1.09E-07
LTR/ERVK	6.24E-07	scRNA	1.99E-04

Table 2.3. Repeat elements that are over-represented (A) and under-represented (B) among unmapped reads compared with unambiguously mapped reads. Over-represented elements include low-complexity regions and retrovirus-related elements, while under-represented elements include many that are frequently found in mouse genomic DNA. *P*-values were calculated using the 2-tailed Fisher Exact Test.

	Unmapped reads (%)	Unambiguous mappings (%)
MMLV vector sequence	19.37	0.32
pBR322	4.84	0.08
low complexity	13.63	6.19
<=25 bp in length	26.62	0
>50% Ns	0.73	0.11
Other repeats	1.19	0.18
Total	66.38	6.88

Table 2.4. Summary of the proportions of unmapped and unambiguously mapping reads that contain vector sequences, or sequences of low complexity, low quality or short length. ">50% Ns" refers to sequences where the identity of more than 50% of bases is unknown. "Other repeats" refers to sequences containing repeat regions other than low complexity regions that were identified using RepeatMasker.

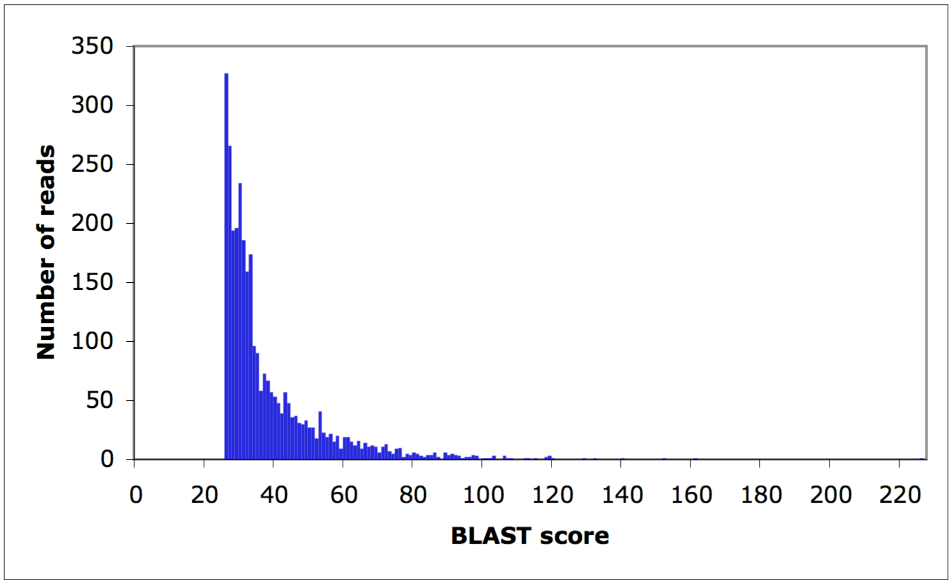

Figure 2.4. BLAST scores for uncharacterised unmapped reads. The majority of sequences that do not map with SSAHA2 but map with BLASTN have a low BLAST score.

significant under-representation (*P*=0) of reads containing both an LTR and an adapter sequence (1,972 reads) compared with those containing no LTR (1,065 reads) or no adapter (1,031 reads). These findings suggest a high presence of erroneous, contaminating reads. In addition, many reads were very short (Figure 2.3C) and may have failed to map due to the presence of a small number of differences from the reference genome sequence. Such differences may correspond to polymorphisms between the mouse strain *FVB*, from which the reads are derived, and strain *C57BL/6J*, upon which the mouse reference genome is based. 17.7% of reads were greater than 800 bp in length. The quality of reads rapidly deteriorates after ~700-900 bases of sequencing, which suggests that these are mostly of very poor quality or are chimeric sequences (discussed in Section 2.5).

There was also a highly significant over-representation of non-mapping reads without linker sequences (P=1.61x10^{-96}) in the *Sleeping Beauty* dataset. Most of the non-mapping sequences flanked by an IR/DR and linker were short, with 50.2% being shorter than the 25 bp threshold for SSAHA2. As with the retroviral reads, there was a higher proportion of low-complexity sequences among unmapped *Sleeping Beauty* reads greater than 25 bp in length (3.1%) than among those that mapped unambiguously (2.4%). There was also a significant over-representation of GC-rich elements (P=1.44x10^{-4}), and an under-representation of the LINE L1M2 (*P*=0.00265) and the rodent-specific LTR MTD (P=2.85x10^{-4}) and SINEs B3 (*P*=0.00348), B3A (*P*=0.00265), PBID10 (P=2.65x10^{-3}) and RSINE1 (P=2.74x10^{-4}).

2.5 Filtering the mapped reads

During PCR amplification, unrelated sequences can hybridise to one another, resulting in clones comprising chimeric sequences. It is important that retroviral reads contain the LTR sequence since, if the part of the read that maps to the genome is directly adjacent to the LTR, the location of the mapped DNA is likely to be the true location of the retroviral insertion. For reads that contain an LTR and an adapter, these sequences should directly flank the genomic DNA. Therefore, for each read, the coordinates of the LTR and adapter sequences identified by cross_match were compared to the coordinates of the region that mapped to the mouse genome using SSAHA2. If the gap between these regions was within 5 bp, the read was accepted. Since the junction between the LTR and the genomic DNA is most important, reads were also accepted if the DNA that mapped to the genome was within 5 bp of the LTR but there was a gap between the genomic DNA and the adapter, or if the read did not contain an adapter sequence. Base miscalling in low quality reads may result in a SSAHA2 alignment that does not extend right up to the LTR sequence even though the LTR and genomic DNA are directly adjacent. Therefore, up to a distance of 30 bp, reads were accepted if the sequence between the LTR and the aligning genomic DNA did not contain any restriction sites for *Tsp509I* (i.e. 5'-AATT-3') or *Sau3AI* (i.e. 5'-GATC-3'), depending on which had been used in the PCR. If a restriction site intercepts the LTR and genomic DNA, it is possible that the genomic DNA that immediately flanks the LTR, and represents the true location of the virus in the genome, may not have been mapped because it is too small or of poor quality but that it has ligated to a contaminating DNA fragment that has been mapped.

The components within the read should be in the configuration LTR-genome-adapter or adapter-genome-LTR. Therefore, any reads that had a different configuration were discarded. For example, the configuration LTR-adapter-genome suggests that a contaminating fragment of genomic DNA has ligated to the end of the adapter, and that the true flanking region of the LTR could not be mapped because it is too short or of poor quality. Reads containing multiple LTR or adapter sequences were subjected to the same filtering criteria, whereby reads were discarded if the sequence for one LTR did not directly abut the genomic sequence or the adapters intercepted the LTR and genomic sequence. Reads with no LTR were rejected unless an LTR identified by reducing the minimum score for cross_match to 5 followed the rules outlined above for stronger LTR matches.

81,846 reads (90.3%) were retained after filtering. Both accepted and rejected reads with gaps of greater than 5 bp were subjected to further analysis. If the average quality (Phred) score of the gap region was less than 30, the read was accepted as the gap may contain miscalled bases, causing SSAHA2 to prematurely terminate extension of the alignment across the full length of the genomic DNA within the read. Reads were also accepted if they mapped to the same location as other reads from the same tumour that did not contain a gap. The final set of accepted reads totalled 81,910 (90.3%). The filtering procedure is summarised in Figure 2.5. There were significantly more reads of greater than 800 bp in length among removed reads (39.5%) than retained reads (6.7%, *P*=0) and removed reads mapped to the genome with a lower percentage identity (92.6% ± 5.9) than retained reads (99.0% ± 2.3).

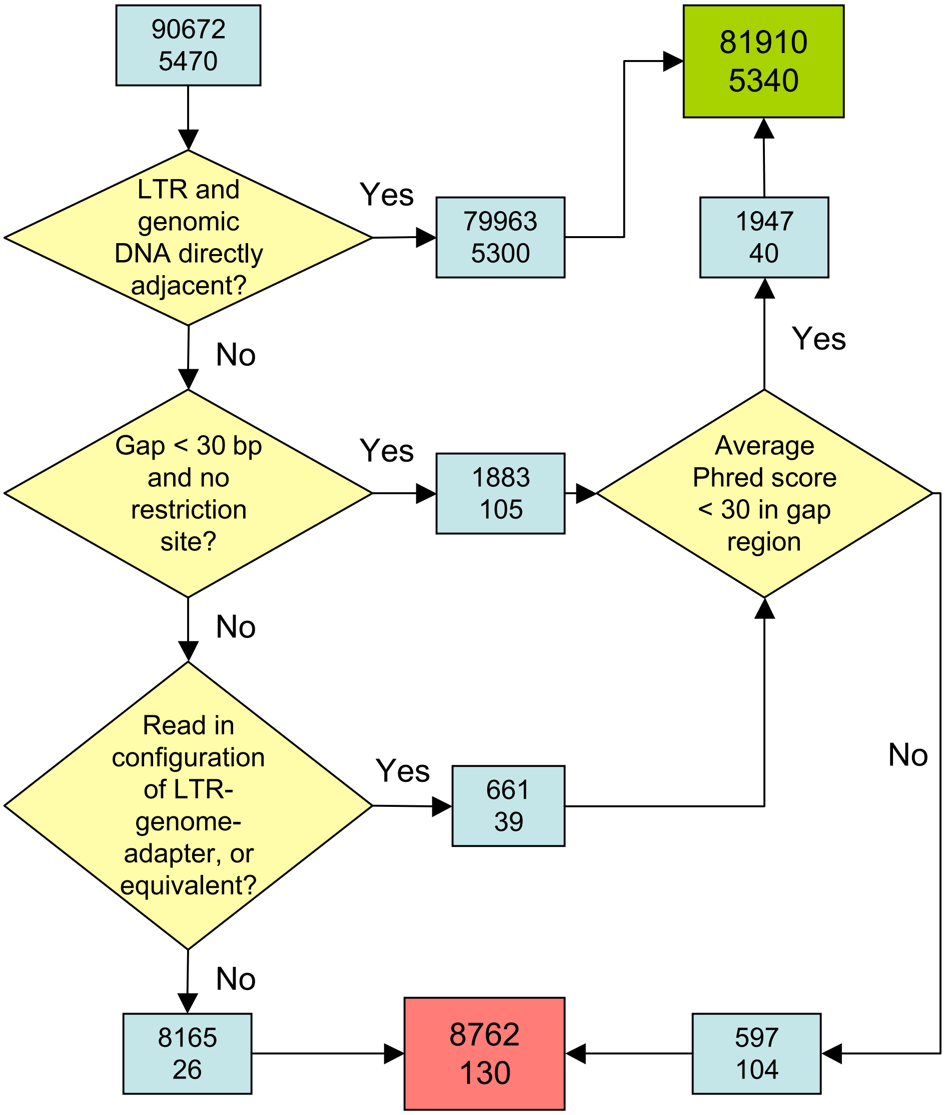
90672
5470
LTR and genomic DNA directly adjacent?
Yes
79963
5300
81910
5340
1947
40
Yes
No
Gap < 30 bp and no restriction site?
Yes
1883
105
Average Phred score < 30 in gap region
No
Read in configuration of LTR-genome-adapter, or equivalent?
Yes
661
39
No
No
8165
26
8762
130
597
104

Figure 2.5. The filtering process for single mapping reads. Blue boxes contain the counts for accepted or rejected reads at each stage, where the top number in each box refers to the count for retroviral reads and the bottom number refers to the count for transposon reads. Final counts for the accepted and rejected reads are shown in the green and red box, respectively.

29,769 reads mapped to multiple locations in the mouse genome. These may be chimeric or low quality reads, or they may represent retroviruses that have inserted into duplicated or repetitive regions of the genome. 15,036 reads were identified where at least one of the mappings matched the criteria used for filtering single mapping reads. For 8,429 of these reads, only one mapping matched the criteria, and this was retained in the dataset while other mappings were discarded. Among the remaining 6,607 reads, there were 465 where only one mapping had an alignment of 100% identity. These mappings were retained and all others were discarded. In total, 8,894 (29.9%) of reads that mapped ambiguously were retained. As with the unambiguous mappings, there was a significant over-representation of reads greater than 800 bp in length in the removed reads (13.4%) compared to the retained reads (8.9%, P=3.40x10^{-28}). The retained reads were pooled together with the retained single mapping reads, giving a total of 90,804 reads.

Transposon insertions were filtered using the same criteria, except that gaps between IR/DRs and genomic DNA were scanned for *NlaIII* (5'-CATG-3') or *BfaI* (5'-CTAG-3') restriction sites, depending on whether the IR/DR was from the left or right end of the transposon. 5,340 (97.6%) of reads mapping to a single genomic location and 941 (50.6%) of those mapping ambiguously were accepted. The filtering of reads that mapped unambiguously is summarised in Figure 2.5.

2.6 Identification and filtering of insertion sites

As 96 reads were sequenced for each PCR, there may be multiple reads that correspond to the same insertion site. The exact genomic coordinates and orientation of the retroviral or transposon insertion represented by each read were determined using the coordinates and orientation of the genomic DNA, resolved by SSAHA2. The methods are summarised in Figure 2.6. Reads from a single PCR mapping to within 2 kb were then clustered into a single insertion site, resulting in 29,553 retroviral insertion sites and 2,821 transposon insertion sites across all PCRs.

It is possible that endogenous LTR sequences within the mouse genome could be the target of non-specific PCR amplification in the retroviral screen. NCBI BLASTN, adjusted to search for short sequences (Word size 7, E value 10,000, filter OFF), was therefore used to identify sequences in NCBI m36 that resembled the MuLV LTR. In a preliminary analysis on NCBI m34, all 15 bp fragments of the LTR sequence

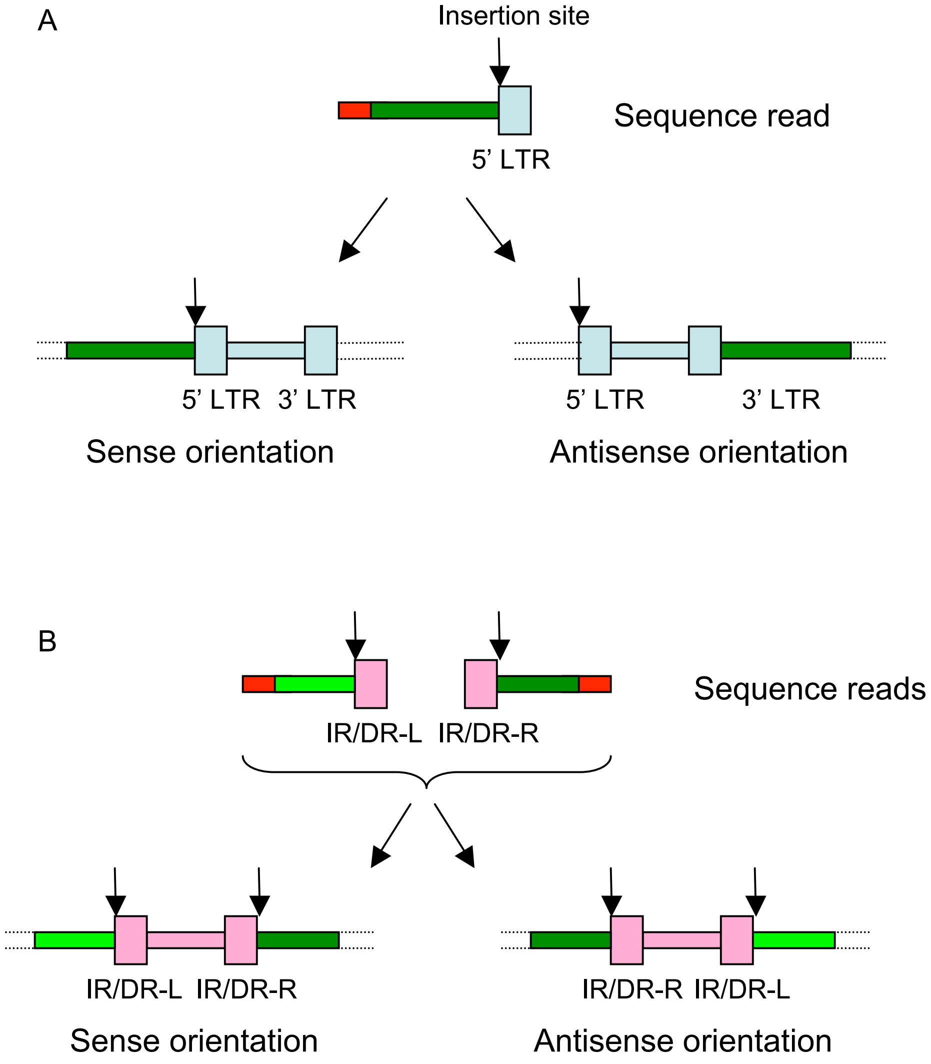

Figure 2.6. Determining the exact insertion site and orientation of retroviral (A) and transposon (B) insertions in the mouse genome. Adapter sequences are shown in red; genomic DNA is shown in green. **A.** The point of insertion is the genomic nucleotide adjacent to the 5' LTR of the MuLV retrovirus (shown in blue) in the sequence read. Alignment to the forward strand of the mouse genome indicates that the retrovirus has inserted in the 5'-3' orientation and the insertion site corresponds to the last nucleotide in the reported alignment. Alignment to the reverse strand indicates that the retrovirus has inserted in the 3'-5' orientation and the insertion site corresponds to the first nucleotide in the alignment. **B.** As for retroviral insertions, except that there are two sets of reads, containing a left or right IR/DR sequence. The T2/Onc transposon is shown in pink.

5'-GCTAGCTTGCCAAACCTACAGGTGGGGTCTTC-3' were used as query sequences, but 99% of the insertions near LTR-like sequences in short infection time mice were identified using LTR fragments 5'-GCTTGCCAAACCTAC-3' and 5'-CTTGCCAAACTACA-3', and therefore only these fragments were used in the current analysis. All of the apparent insertions in the uninfected control samples should be PCR artefacts, while short infection time DNA is expected to contain a higher proportion of PCR artefacts than tumour DNA. Among the 1,399 reads mapping to LTR-like sites, there were significantly more from uninfected samples and from short infection time samples than expected by chance (P=3.17x10^{-26} and P=0, respectively). These findings support the theory that reads mapping to sites that resemble the retroviral LTR are the result of non-specific PCR amplification and do not represent real insertion sites. For example, 174 samples contain an insertion in the aminoadipate-semialdehyde synthase (*Aass*) gene, but the insertions are adjacent to a 14 bp sequence that precisely matches the MuLV LTR and are therefore likely to be false positives. Figure 2.7 shows these insertions displayed in Ensembl. The Distributed Annotation System (DAS) server ProServer was used to display both the retroviral and the transposon insertion sites in the context of the mouse genome in Ensembl contigview. Ensembl is a DAS client that can integrate genome annotation information from multiple servers, enabling users to view and compare annotations from multiple sources in a single display. All 1,399 reads at 675 LTR-like sites were removed from the dataset.

Apparent insertions in non-infection and short infection time samples were removed from the dataset, but a decision was made not to remove tumour insertions that mapped to the same locations. A preliminary analysis, in which the reads were mapped to mouse build NCBI m34, showed that many of the reads from non-infection and short infection time samples mapped to cancer genes that are known targets of retroviral insertional mutagenesis. Insertions within 5 kb of *Myc* were identified in 41.7% of non-infection samples, 26.2% of short infection time samples and 30.4% of tumour samples (see Figure 2.8). Similarly, the proportions of insertions from non-infection and short infection time samples in and around *Mycn* were 12.5% and 35.9%, respectively, but just 8.9% in tumour samples. Findings for the short infection time dataset could indicate that *Myc* and *Mycn* are insertion hotspots, or that selection for *Myc* and *Mycn* insertions occurs at an early time point. However, these explanations do not justify the presence of such insertions in non-infection samples. As all non-infection insertions map to only 142 distinct coordinates, it seems an unlikely coincidence that *Myc* and *Mycn* are targeted by

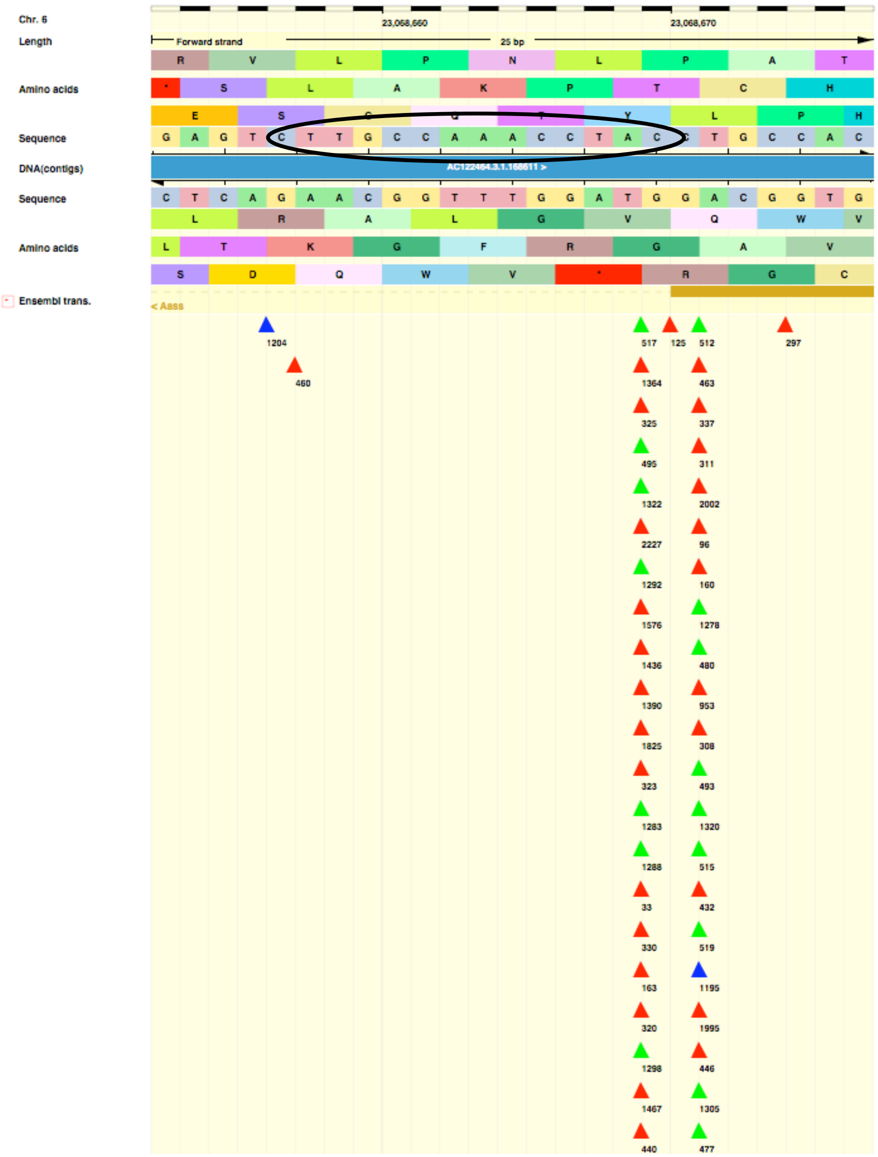

Figure 2.7. Insertions in the mouse aminoadipate-semialdehyde synthase (*Aass*) gene are PCR artefacts that map to an LTR-like sequence in the mouse genome. 174 samples contain an insertion in this region (46 are shown here as triangles). Insertions from tumours, short infection time samples and uninfected samples are shown as red, green and blue triangles, respectively. The LTR-like sequence is circled.

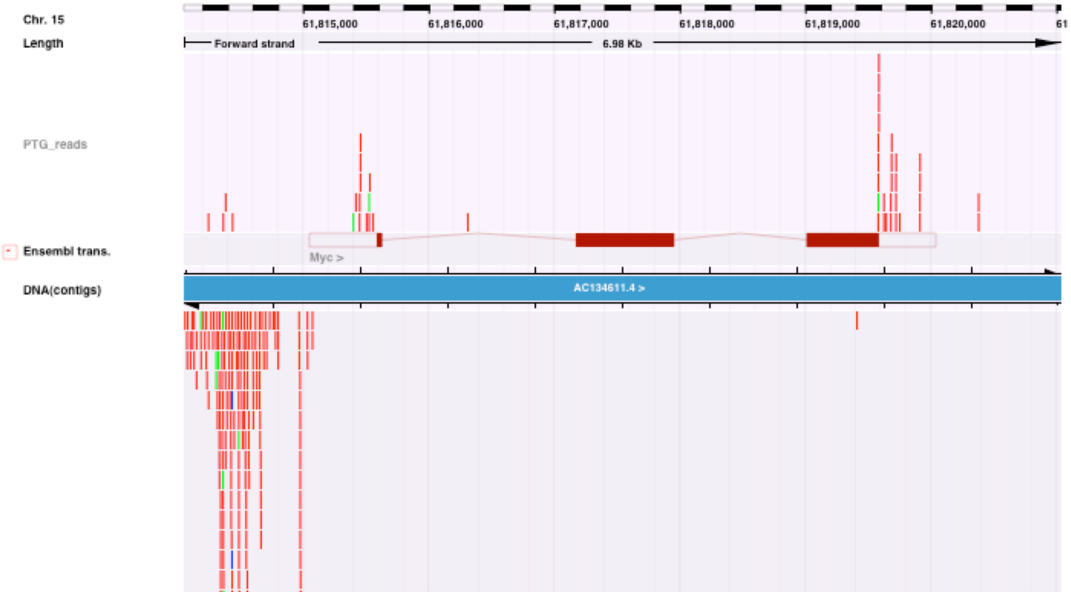

Figure 2.8. A high proportion of insertions in control samples map to the *Myc* gene. This figure shows some of the insertions in and around the *Myc* gene. Insertions from tumours, short infection time samples and uninfected samples are shown as red, green and blue rectangles, respectively.

non-specific primer binding, and there are no LTR-like sequences near these genes. The insertions may result from contamination during PCR or, even more worryingly, unintended infection of mice in the animal facility. The control samples are useful for picking out possible contaminants, like those described above that map to LTR-like sequences, but discarding all tumour insertions that map to the same sites as control insertions would most likely result in the removal of a considerable number of real insertions.

The insertion sites identified in individual PCRs were clustered into 22,579 retroviral insertion sites from 997 tumours. The average number of inserts per tumour was 23.49 ± 11.42 (Figure 2.9A). There were, on average, 3.72 ± 6.21 reads per insert (Figure 2.9B). The 2,821 transposon insertion sites identified in individual PCRs were clustered into 2,643 insertion sites from 73 tumours. There was an average of 36.21 ± 18.55 inserts per tumour, and 2.38 ± 4.08 reads per insert.

2.7 Estimating the coverage of the mutagenesis screens

Measuring the overlap of insertion sites between PCRs for an individual tumour gives some indication of the proportion of insertions that were identified in the screens. There were 616 tumours for which retroviral insertions had been identified from one PCR using *Sau3A1* and one using *Tsp509I.* These contained 10,733 and 8,580 insertions identified using *Sau3A1* and *Tsp509I*, respectively, of which 2,968 were identified using both enzymes. The overlap between PCR experiments was therefore 18.2%, rising to 32.9% if insertions represented by a single read were omitted. More than one enzyme is required because individual enzymes do not cut the genomic DNA sufficiently close to all insertions to enable PCR amplification of the intervening sequences. Since the overlap between PCRs is low, it seems likely that even two enzymes do not give sufficient coverage. However, the difference between the 2 PCRs may also result from insufficient sequencing, such that genomic DNA flanking an insertion is amplified but is not sequenced. This may explain why a high proportion of insertion sites represented by a single read are not identified by both PCRs, since they are more likely to be rare insertions that have a low representation in the PCR mixture and are less likely to be sequenced.

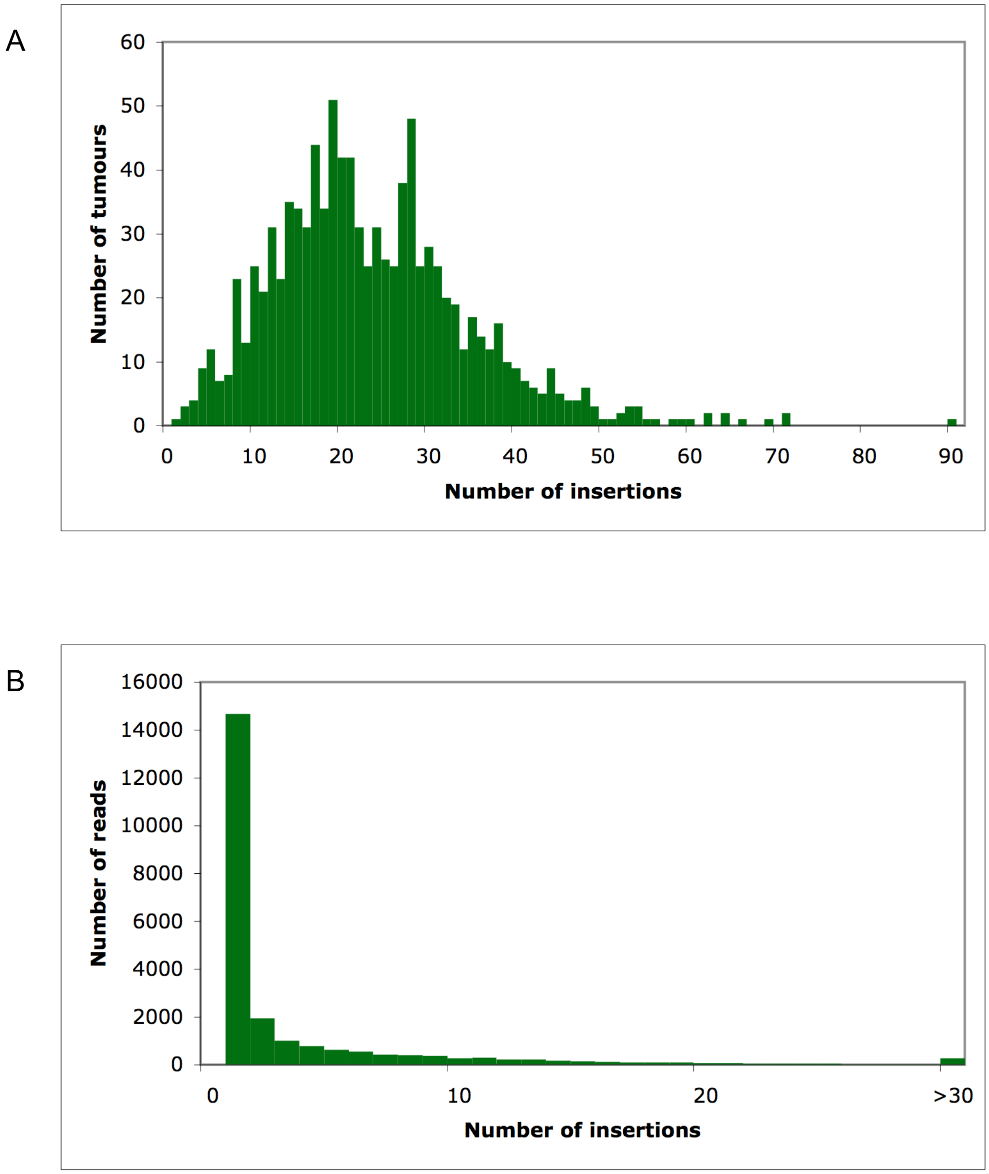

Figure 2.9. The number of insertions per tumour (A) and reads per insertion (B).

For 3 tumours, genomic DNA was also digested using *BstYI*, and 384 reads were sequenced. Reads were mapped to NCBI m34 and were compared to *Sau3A1* and *Tsp509I* reads from the same tumours, also mapped to NCBI m34. There was a 39.0% overlap between insertion sites identified from PCRs using *Sau3A1* and *BstY1*, and a 23.4% overlap between those identified from PCRs using *Tsp509I* and *BstYI*. A higher overlap is expected between *Sau3A1* and *BstYI* because the *BstYI* target site (5'-RGATCY-3') contains the target sequence for *Sau3A1*. *BstYI* cuts less frequently than *Sau3A1* and *Tsp509I*. For reads generated using *Sau3A1* or *Tsp509I*, the average distance between the LTR and the restriction site at which the DNA was cut was 308.41 bp, but for reads generated using *BstYI*, the average distance was 386.52 bp. It is therefore difficult to directly compare the PCRs because fragments of *BstYI*-digested DNA will be longer, on average, and there is likely to be a higher proportion that cannot be amplified by PCR. For insertion sites that were identified using *Sau3A1* or *Tsp509I* but not using *BstYI*, the genomic DNA within the corresponding reads was scanned for *BstYI* target sites. Likewise, for insertion sites that were uniquely identified using *BstYI*, the genomic DNA was scanned for *Sau3A1* and *Tsp509I* target sites. If the sequencing depth of 96 reads was sufficient, insertion sites should only be uniquely identified using *BstYI* if there are no *Sau3A1* and *Tsp509I* target sites close enough to the insertion site for successful PCR. A *BstYI* target site was identified at a distance equal to, or closer than, the *Sau3A1* or *Tsp509I* site in reads corresponding to 2 out of 15 unique *Sau3A1* insertion sites and 4 out of 20 unique *Tsp509I* insertion sites. However, for *Sau3A1* and *Tsp509I*, a target site was identified at a distance equal to, or closer than, the *BstYI* site for 21/21 and 14/29 unique *BstYI* insertions, respectively. This suggests that more insertion sites could be obtained by increasing the sequencing depth to 384 reads per PCR, and that an even greater depth may be required to saturate the screen. However, as only 3 tumours were used in this analysis, and different enzymes were used to generate the digested DNA for 96-read and 384-read sequencing, it is difficult to reach any firm conclusions about the number of enzymes and the sequencing depth required for maximum coverage.

For the Sleeping Beauty screen, there were 60 tumours for which 2 PCRs were performed using restriction enzymes *BfaI* and *NlaIII*. Only 159 insertions (6.9%) were shared from 1,161 insertion sites identified using *BfaI* and 1,310 identified using *NlaIII*.

2.8 Analysis of the distribution of insertions around mouse genes

The long-range effects of MuLV enhancer mutations can complicate the identification of mutated genes. Analysing the distribution of insertions around mouse genes, and in particular, around the mouse orthologues of known cancer genes, can help to define rules for predicting which gene is being mutated by an insertion. The genomic coordinates and orientation of all mouse protein-coding and miRNA genes were extracted from Ensembl using the Perl API version 45_36f, and insertions were counted in 100 bp intervals up to 20 kb upstream and downstream of each gene. The gene orientation was used to determine the orientation of insertions with respect to each gene. Figures 2.10A-D show the number of genes that contain insertions in each 100 bp interval upstream and downstream in the sense and antisense orientation with respect to each gene. In the full set of genes, the number of sense and antisense insertions peak at around 500-600 bp upstream, and a similar pattern is observed around the mouse orthologues of known cancer genes. These sense and antisense insertions are likely to represent promoter and enhancer mutations, respectively (see Section 1.4.2.1.1), with the peak representing the optimal distance for mutation. Downstream insertions show a relatively uniform distribution with similar proportions of insertions in the sense and antisense orientation. This may indicate that most are randomly occurring non-oncogenic insertions, or that there is no optimum distance for an enhancer mutation that acts downstream of a gene. It is also likely that some of these insertions are affecting adjacent genes, and variation in the distance between genes may contribute to the observed distribution. There is also no obvious pattern in the downstream counts of cancer genes with insertions. The plots in Figures 2.10A-D show the counts of genes with insertions up to 20 kb upstream or downstream, regardless of whether adjacent genes intercept the 20 kb region. However, counting only as far as the adjacent gene gives a similar distribution, with peaks at 500-600 bp upstream in both orientations, and an essentially uniform distribution downstream. Counting actual insertions, rather than the number of genes containing insertions, skews the distribution towards genes containing larger numbers of insertions. For example, *Myc* contains many enhancer mutations, and the highest peak might represent the optimal distance for an enhancer mutation of *Myc*, rather than for all genes. However, once again the highest peak is at 500-600 bp upstream. A similar distribution is also obtained by counting only the genes that contain insertions represented by more than one read.

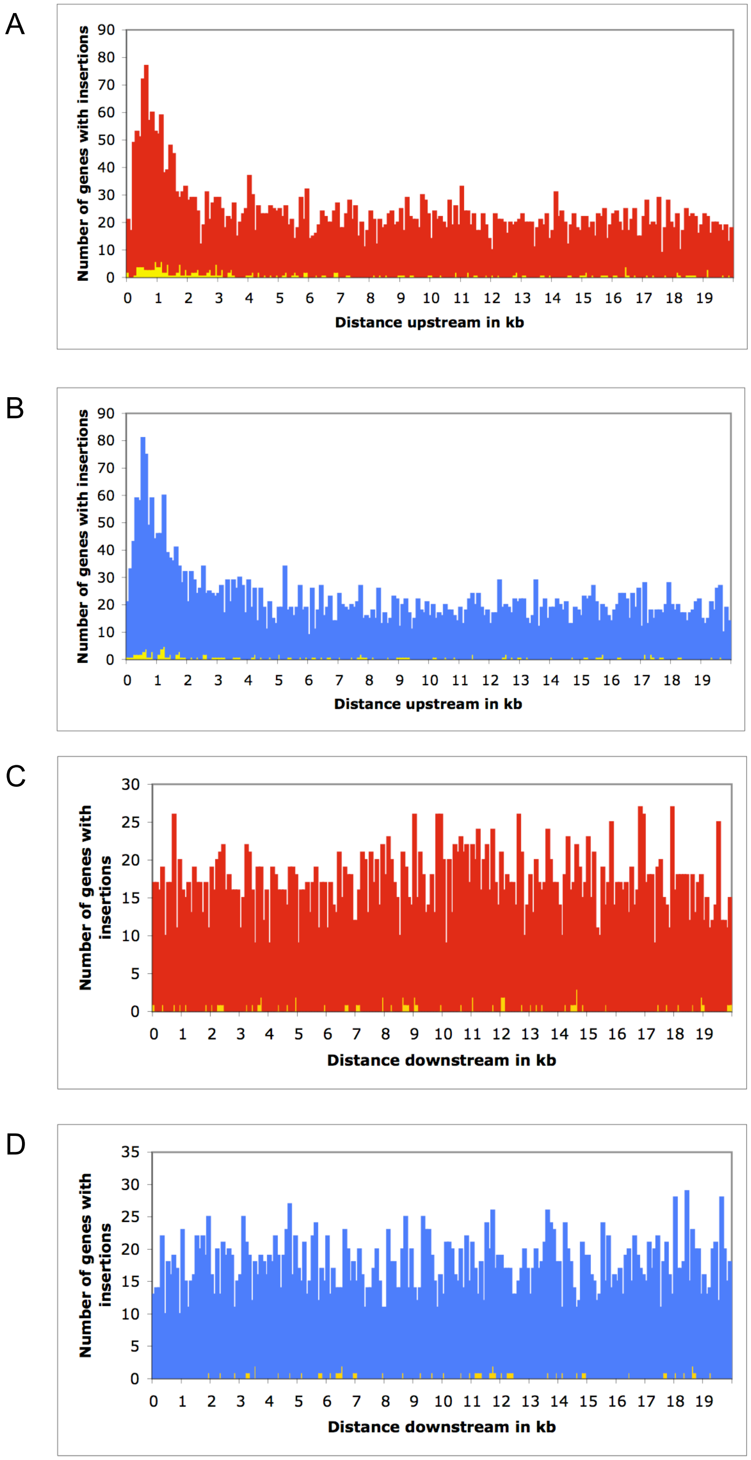

Figure 2.10. The number of genes with insertions in 100 bp intervals up to 20 kb upstream in the sense (A) and antisense (B) orientation and downstream in the sense (C) and antisense (D) orientation with respect to the gene. Counts of cancer genes with insertions in each interval are shown in yellow.

Known oncogenes and tumour suppressor genes with intergenic insertions up to 20 kb upstream and/or downstream are shown in Tables 2.5A and 2.5B, respectively. Tumour suppressor genes are expected to contain intragenic insertions that result in truncated, inactivated, transcripts. Of the 12 tumour suppressor genes flanked by intergenic insertions, only 1 has insertions represented by more than one read. This suggests that "singleton" insertions, i.e. insertions represented by a single read, are less likely to contribute to oncogenesis. They may be rare insertions that are not in the dominant tumour lineage or have integrated into a single lineage late on in tumour development, or they may be PCR artefacts. 8 known oncogenes had insertions within 2 kb upstream in the sense orientation, and 22 had insertions within 20 kb. These numbers fell to 5 and 9, respectively, if singleton insertions were removed. Likewise, there were 9 oncogenes with antisense insertions within 2 kb upstream, and 29 with insertions within 20 kb, but only 5 and 13, respectively, without singletons. As well as representing rare insertions, singleton insertions may result from limitations in PCR and sequencing depth. Therefore, in order to maximise the number of candidates that could be identified, singleton insertions were retained in the analysis since, if they are not important in tumourigenesis, they should not form statistically significant CISs (see Section 2.10).

Insertions around the *Pim1* oncogene (Figure 2.11A) suggest that downstream sense and antisense insertions can contribute to tumourigenesis. Downstream sense insertions also appear to affect the *Kit* oncogene (Figure 2.11B). However, there are fewer oncogenes with downstream sense insertions than upstream insertions, and even fewer with downstream antisense insertions. For some of the genes, e.g. *Gata1* (Figure 2.11C), it does appear that the downstream insertions are in fact mutating an adjacent gene. These observations concur with prior work, in suggesting that upstream antisense and sense insertions, corresponding to enhancer and promoter mutations, respectively, are the most common forms of mutation. Downstream insertions, while less common, appear to be more frequent in the sense orientation, which is the proposed orientation for downstream enhancer mutations (see Section 1.4.2.1.1).

Of the 22 oncogenes that had sense insertions within 20 kb upstream, 20 were still identified when the upstream limit was set to the 3' end of the upstream gene. All 9 genes without singleton insertions were similarly identified. Likewise, 23 out of 29 genes, including all 13 genes without singletons, that had antisense insertions within 20 kb upstream were still identified. 9 out of 14 genes containing downstream sense

A

Insertion		Gene		20 kb no		2 kb no	Within
Orientation	Mouse Ensembl ID	Name	20 kb all	singletons	2 kb all	singletons	limits
Upstream sense	ENSMUSG00000031103	*Elf4*	39	18	33	13	39
	ENSMUSG00000018654	*Ikzf1*	22	14	0	0	22
	ENSMUSG00000062312	*Erbb2*	13	0	0	0	1
	ENSMUSG00000022346	*Myc*	12	9	11	8	12
	ENSMUSG00000006362	*Cbfa2t3*	11	4	2	0	11
	ENSMUSG00000026923	*Notch1*	8	4	0	0	8
	ENSMUSG00000000409	*Lck*	7	6	6	5	7
	ENSMUSG00000034342	*Cbl*	7	3	0	0	6
	ENSMUSG00000024014	*Pim1*	6	4	6	4	6
	ENSMUSG00000029204	*Rhoh*	5	0	0	0	5
	ENSMUSG00000032688	*Malt1*	3	0	0	0	3
	ENSMUSG00000036986	*Pml*	3	0	0	0	0
	ENSMUSG00000025408	*Ddit3*	2	0	2	0	2
	ENSMUSG00000037169	*Mycn*	2	0	2	0	2
	ENSMUSG00000000184	*Ccnd2*	2	2	2	2	2
	ENSMUSG00000059248	*Sept9*	2	0	0	0	2
	ENSMUSG00000020893	*Per1*	2	0	0	0	2
	ENSMUSG00000021377	*Dek*	2	0	0	0	2
	ENSMUSG00000021356	*Irf4*	2	0	0	0	2
	ENSMUSG00000027829	*Ccnl1*	2	0	0	0	2
	ENSMUSG00000066306	*Numa1*	2	0	0	0	2
	ENSMUSG00000003282	*Plag1*	2	0	0	0	0
Upstream antisense	ENSMUSG00000022346	*Myc*	388	303	383	299	388
	ENSMUSG00000026923	*Notch1*	19	10	0	0	19
	ENSMUSG00000024014	*Pim1*	16	11	15	10	16
	ENSMUSG00000070348	*Ccnd1*	14	9	5	3	14
	ENSMUSG00000018654	*Ikzf1*	14	9	0	0	14
	ENSMUSG00000000184	*Ccnd2*	13	7	4	2	13
	ENSMUSG00000006362	*Cbfa2t3*	13	8	0	0	13
	ENSMUSG00000006389	*Mpl*	10	0	4	0	6
	ENSMUSG00000022952	*Runx1*	8	6	0	0	8
	ENSMUSG00000003282	*Plag1*	8	0	0	0	0
	ENSMUSG00000059248	*Sept9*	6	2	0	0	6
	ENSMUSG00000042817	*Flt3*	5	2	4	2	5
	ENSMUSG00000031103	*Elf4*	4	0	2	0	4
	ENSMUSG00000043962	*Akt3*	3	0	3	0	3
	ENSMUSG00000048251	*Bcl11b*	3	0	2	0	0
	ENSMUSG00000030745	*Il21r*	3	0	0	0	3
	ENSMUSG00000034342	*Cbl*	3	0	0	0	3
	ENSMUSG00000025958	*Creb1*	2	2	0	0	2
	ENSMUSG00000020453	*Patz1*	2	0	0	0	2
	ENSMUSG00000021457	*Syk*	2	0	0	0	2
	ENSMUSG00000021356	*Irf4*	2	2	0	0	2
	ENSMUSG00000056234	*Ncoa4*	2	0	0	0	2
	ENSMUSG00000022797	*Tfrc*	2	0	0	0	2
	ENSMUSG00000032698	*Lmo2*	2	0	0	0	2
	ENSMUSG00000002028	*Mll1*	2	0	0	0	2
	ENSMUSG00000025408	*Ddit3*	2	2	0	0	0
	ENSMUSG00000041358	*Nut*	2	0	0	0	0
	ENSMUSG00000000409	*Lck*	2	0	0	0	0
	ENSMUSG00000029438	*Bcl7a*	2	0	0	0	0
Downstream sense	ENSMUSG00000024014	*Pim1*	17	9	2	0	17
	ENSMUSG00000038227	*Hoxa9*	6	2	0	0	3
	ENSMUSG00000022346	*Myc*	5	2	0	0	5
	ENSMUSG00000020325	*Fstl3*	4	2	0	0	0
	ENSMUSG00000010755	*Cars*	3	0	3	0	0
	ENSMUSG00000032097	*Ddx6*	3	0	0	0	3
	ENSMUSG00000034041	*Lyl1*	2	0	0	0	3
	ENSMUSG00000057329	*Bcl2*	2	0	0	0	2
	ENSMUSG00000069305	*Hist4h4*	2	0	0	0	2
	ENSMUSG00000029204	*Rhoh*	2	0	0	0	2
	ENSMUSG00000005672	*Kit*	2	0	0	0	2
	ENSMUSG00000034165	*Ccnd3*	2	0	0	0	0
	ENSMUSG00000004895	*Prcc*	2	2	0	0	0
	ENSMUSG00000028718	*Stil*	2	0	0	0	0
Downstream antisense	ENSMUSG00000031162	*Gata1*	9	8	0	0	5
	ENSMUSG00000024014	*Pim1*	5	2	0	0	5
	ENSMUSG00000030745	*Il21r*	4	2	0	0	4
	ENSMUSG00000069305	*Hist4h4*	3	0	0	0	3
	ENSMUSG00000020453	*Patz1*	3	0	0	0	0
	ENSMUSG00000068860	*Gm128*	3	0	0	0	0
	ENSMUSG00000070002	*Ell*	3	2	0	0	0
	ENSMUSG00000026656	*Fcgr2b*	2	0	0	0	2
	ENSMUSG00000020167	*Tcfe2a*	2	0	0	0	0
	ENSMUSG00000034041	*Lyl1*	2	0	0	0	0

B

Insertion		Gene		20 kb no		2 kb no	Within
orientation	Mouse Ensembl ID	name	20 kb all	singletons	2 kb all	singletons	limits
Upstream sense	ENSMUSG00000003068	*Stk11*	2	0	0	0	2
	ENSMUSG00000009863	*Sdhb*	2	0	0	0	0
	ENSMUSG00000036712	*Cyld*	2	0	0	0	0
Upstream antisense	ENSMUSG00000009863	*Sdhb*	6	0	0	0	0
	ENSMUSG00000003068	*Stk11*	2	0	0	0	2
	ENSMUSG00000013663	*Pten*	2	0	0	0	2
	ENSMUSG00000026526	*Fh1*	2	0	0	0	0
	ENSMUSG00000028687	*Mutyh*	2	0	0	0	0
	ENSMUSG00000034023	*Fancd2*	2	0	0	0	0
Downstream sense	ENSMUSG00000030528	*Blm*	4	2	0	0	4
	ENSMUSG00000024947	*Men1*	2	0	0	0	0
	ENSMUSG00000025231	*Sufu*	2	0	0	0	0
	ENSMUSG00000044702	*Palb2*	2	0	0	0	0
Downstream antisense	ENSMUSG00000040084	*Bub1b*	3	0	0	0	3
	ENSMUSG00000024947	*Men1*	2	0	0	0	0

Table 2.5. Number of intergenic insertions up to 20 kb upstream and downstream of known oncogenes (A) and tumour suppressor genes (B) from the Cancer Gene Census. "20 kb all" and "2 kb all" give the total number of insertions up to 20 kb and 2 kb upstream/downstream. "2 kb no singletons" and "20 kb no singletons" give the number of insertions represented by more than 1 read. "Within limits" gives the number of insertions up to the adjacent upstream or downstream gene.

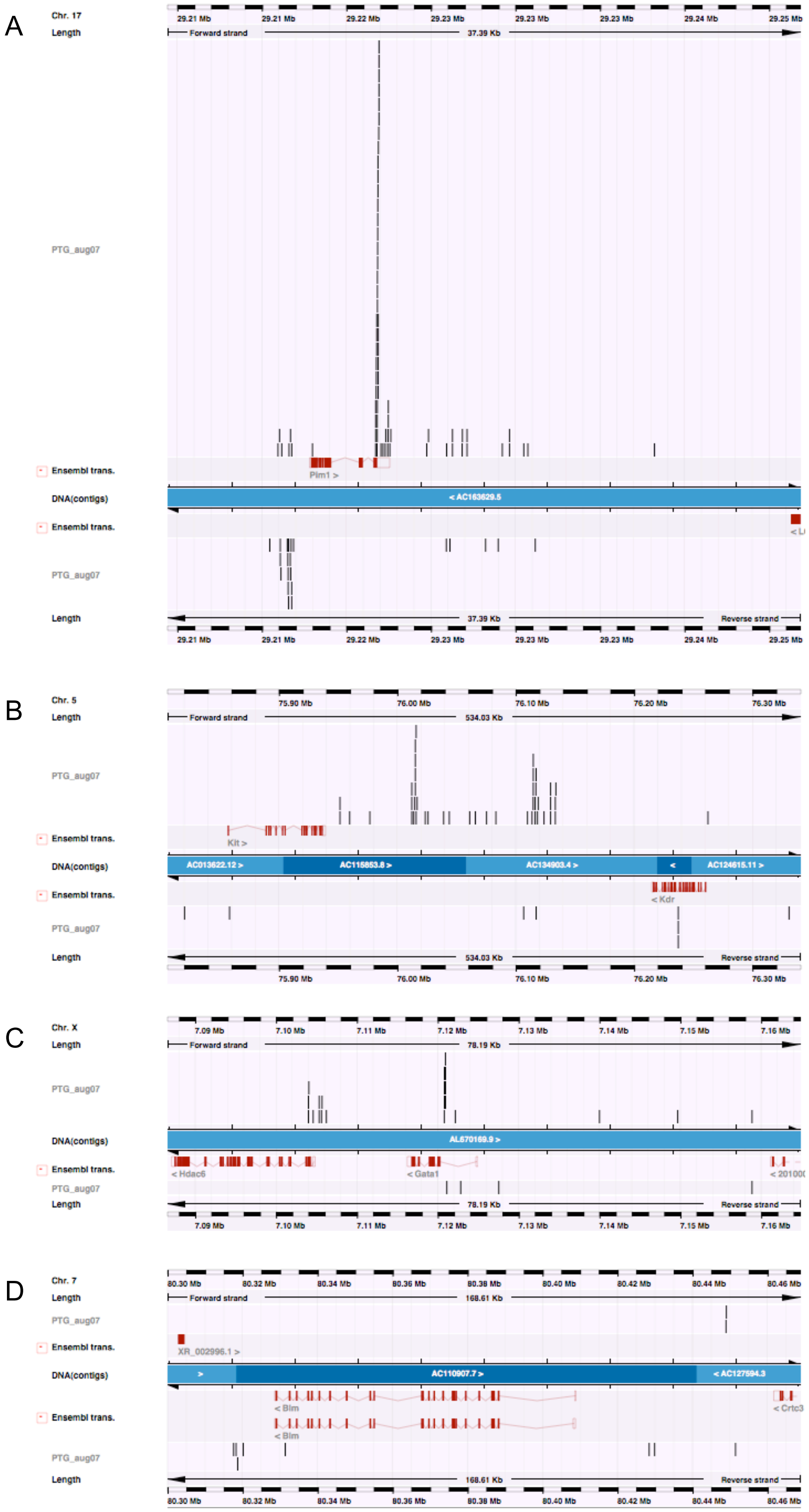

Figure 2.11. Insertions around known cancer genes *Pim1* (A), *Kit* (B), *Gata1* (C) and *Blm* (D). Insertions are shown as black bars in the context of Ensembl genes, shown in red. Insertions above and below the blue bar are in the sense and antisense orientation, respectively.

insertions, and 5 out of 10 genes containing downstream antisense insertions, were still identified when the downstream limit was set to the 5' end of downstream gene. The lower proportion for genes with downstream sense insertions, and lower still for downstream antisense insertions, most likely reflect the fact that these insertions are less likely to contribute to oncogenesis. The same applies to tumour suppressor genes, of which only 4 out of 12 were still identified when the upstream and downstream limits were set to adjacent genes. Therefore, based on these results, it seems reasonable to assign insertions to a gene only if they are within the limits of adjacent genes.

As indicated by the high proportion of singleton insertions and the presence of insertions beyond the boundaries of adjacent genes, it is likely that most of the tumour suppressor genes listed in Table 2.5 are not mutational targets. However, the *Blm* gene contains an intragenic insertion, as well downstream sense insertions (Figure 2.11D). It is possible that the intergenic insertions are not oncogenic, or that they are affecting a nearby gene, or there may be an error in the *Blm* gene prediction in Ensembl, such that the insertions appear to be intergenic but are in fact intragenic. Alternatively, the insertions could be disrupting a downstream regulatory element, resulting in reduced transcription or gene inactivation.

There is no obvious pattern in the distribution of transposon insertions upstream or downstream of genes. This is not surprising for upstream antisense and downstream insertions, since the *Sleeping Beauty* transposon T2/Onc has low enhancer activity. However, insertions in the upstream sense orientation might be expected to follow a similar distribution to those in the retroviral screen. The T2/Onc promoter is perhaps not as strong as the MuLV promoter and mostly mutates by producing truncated transcripts, rather than by increasing levels of the wildtype protein. Alternatively, some of the apparent promoter mutations in the retroviral screen may in fact be enhancer mutations, or a high background of non-oncogenic T2/Onc insertions may be masking the true pattern of oncogenic mutations. There is only one oncogene (*Irf4*) and no tumour suppressor genes with sense or antisense insertions up to 20 kb upstream or downstream. While this may in part reflect the smaller size of the dataset, it also suggests that oncogenic T2/Onc insertions are usually intragenic.

2.9 Assigning insertions to genes

The coordinates and orientation of the longest transcript of all protein-coding and miRNA genes in the mouse genome were extracted from Ensembl version 45_36f using the API. Genes nestled within other genes were removed from the analysis, since these complicate the specification of gene boundaries for assigning intergenic insertions to genes. Intragenic retroviral insertions were assigned to the genes within which they resided. For intergenic insertions, the flanking genes were identified. If an insertion was upstream of the first gene or downstream of the last gene on a chromosome, it was assigned to the first or last gene, respectively. If only one of the flanking genes was within 100 kb of the insertion, that gene was assigned the insertion. If one of the flanking genes contained intragenic insertions, the intergenic insertions were also assigned to that gene. Based on the observations of insertions around known cancer genes outlined in Section 2.8, if an insertion was in the downstream antisense orientation relative to one gene, but in a different orientation relative to the other gene, it was assigned to the other gene, and other intergenic insertions were also assigned to that gene. Finally, for the remaining unassigned intergenic insertions, the nearest insertion to each gene was identified, and all insertions were assigned to the gene that had the nearest insertion. *Sleeping Beauty* T2/Onc insertions were processed in a similar way, except that if an intergenic insertion was in the upstream sense orientation with respect to one gene, but in a different orientation with respect to another gene, it was assigned to the former gene.

2.10 Identifying statistically significant common insertion sites

Oncogenic insertions must be distinguished from a background of non-oncogenic insertions. Insertions from different tumours that reside in the same genomic region, defined as common insertion sites (CISs), are more likely to contribute to tumourigenesis, but statistical approaches are required to determine whether a CIS is significantly different to the random, background distribution of insertions. Monte Carlo simulations, and a more recent method that uses a kernel convolution-based statistical framework, have been applied to the retroviral and *Sleeping Beauty* datasets, and the results compared.

2.10.1 Monte Carlo simulations

This method is based on the procedures described in Suzuki *et al.* (2002) and Mikkers *et al.* (2002). The 26,144 retroviral insertions were randomised across the mouse genome (golden path length 2,661,205,088 bp, mouse build NCBI m36). A wide range of window sizes were used, and the number of windows containing at least *M* insertions were counted, where *M* was a number of 2 or more (up to 14 for large window sizes). The randomised insertions were ordered across the genome ($X_{[1]}$ to $X_{[26,144]}$), and windows were taken as the interval from $X_{[i]}$ to $X_{[i+M\text{-}1]}$ (see Suzuki *et al.*, 2002). If the distance between an insertion and the next *M*-1 insertions on the chromosome was less than the window size ($\mathrm{X}_{[i+M\text{-}1]} - \mathrm{X}_{[i]}$), it was counted as a CIS. The next window was positioned at *i*+*M*. 100,000 iterations were performed, and mean counts and the 0.99 upper quantile were calculated for each number (*M*) of insertions. This gives the number of CISs of *M* insertions that one would expect to find by chance in each window size, and the maximum number for *P*=0.01. As in Mikkers *et al.* (2002), fractions (represented as *Efr*) of 0.001, 0.005 and 0.01 of the total number of insertion sites expected to be random CIS clusters were calculated. These are 26.144, 130.72 and 261.44, respectively, for retroviral insertions, and 2.64, 13.22 and 26.43, respectively, for transposon insertions. Maximum window sizes for significant CISs for varying values of *M* can then be calculated by finding the window size at which the upper quantile of the random distribution is less than the expected number of false CISs (Table 2.6).

For each gene to which insertions had been assigned, the number of insertions was counted and the distance between insertions was calculated. If any of the insertions fell within a window size that met the criteria for a significant CIS, the gene was accepted as a candidate cancer gene. For an *Efr* of 0.001, 0.005 and 0.01, the number of identified candidates in the retroviral screen was 1,404, 1,677 and 1,829, respectively. For the *Sleeping Beauty* screen, the number of candidates was 62, 91 and 115, respectively. This approach differs from the method in Suzuki *et al.* (2002) in that insertions were considered in the context of each gene, and a consistent approach was used to identify all candidates. In Suzuki *et al.* (2002), CISs were identified independently of genes, and then assigned to genes, but further genes were selected as candidates if they contained multiple insertions that were not in significant CISs. In addition, the method in Suzuki *et al.* (2002) uses 3 fixed window sizes to define CISs, which, particularly in a screen of this

A

	Number of insertions												
Efr	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**
0.001	0.1	5.0	19.5	45.0	80.0	120.0	168.0	220.0	275.0	333.0	391.5	455.0	521.0
0.005	0.5	10.0	35.0	75.0	120.0	175.0	235.0	299.5	366.0	437.5	510.0	586.0	663.0
0.01	1.0	14.4	45.7	95.0	150.0	210.0	280.0	351.5	425.5	505.0	587.5	671.0	757.0

B

	Number of insertions						
Efr	**2**	**3**	**4**	**5**	**6**	**7**	**8**
0.001	1	45	193	450	800	1200	1650
0.005	5	101	345	750	1200	1750	N/A
0.01	10	142	452	950	1500	2100	N/A

Table 2.6. Maximum window sizes in kb for significant CISs for varying numbers of insertions in the retroviral (A) and *Sleeping Beauty* (B) screens. Window sizes are given for *Efr* (fraction of the number of insertion sites expected to be random CIS clusters) of 0.001, 0.005 and 0.01, for which the corresponding numbers of false CISs are 26.144, 130.72 and 261.44 for retroviral insertions, and 2.64, 13.22 and 26.43 for transposon insertions. N/A is given where the window size is larger than any gene plus 100 kb of flanking sequence, and is therefore not relevant to the analysis.

A

Method	**Number of cancer genes**	**Number of non-cancer genes**	**Accuracy**	**Coverage**	**MCC**
KC	42	487	0.0794	0.1193	0.1433
MC *Efr*=0.001	66	1144	0.0545	0.1875	0.1010
MC *Efr*=0.01	80	1500	0.0506	0.2273	0.0944
All	175	5483	0.0309	0.4972	0.0562

B

Method	**Number of cancer genes**	**Number of non-cancer genes**	**Accuracy**	**Coverage**	**MCC**
KC	6	21	0.2222	0.0170	0.3115
MC *Efr*=0.001	10	45	0.1818	0.0284	0.2708
MC *Efr*=0.01	11	90	0.1089	0.0313	0.1836
All	59	1279	0.0441	0.1676	0.0767

Table 2.7. Comparison of the methods used to generate candidate cancer genes lists from the retroviral (A) and *Sleeping Beauty* (B) screens. The accuracy, coverage and Matthew's correlation coefficient (MCC) are based on the number of known cancer genes in the candidate gene lists. KC = kernel convolution-based framework, MC *Efr*=0.001 and MC *Efr*=0.01 refer to Monte Carlo simulations using *Efr* (fraction of the number of insertion sites expected to be random CIS clusters) of 0.001 and 0.01, All = all genes to which insertions were assigned, regardless of whether they were statistically significant.

size, could result in some CISs being missed. Therefore, this method uses the approach in Mikkers *et al.* (2002) to define maximum window sizes for all values of *M*.

2.10.2 Kernel convolution

As discussed in Section 1.4.2.1.2, the Monte Carlo (MC) method may not be suitable for very large datasets. Significant CISs were therefore also identified using the kernel convolution (KC)-based statistical framework (de Ridder *et al.*, 2006). A list of insertions was supplied to the Netherlands Cancer Institute, where Jeroen de Ridder produced and returned a list of genomic coordinates corresponding to CISs generated using the KC method. In this method, a kernel function is placed at every insertion in the dataset and the number of insertions at any genomic position can be estimated by summing all the kernel functions. Insertions in close proximity to one another will produce a higher peak in the estimated number of insertions (de Ridder *et al.*, 2006, also discussed in Section 1.4.2.1.2).

867 retroviral cross-scale CISs were identified using the KC-based framework with *P*=0.05. These are all the CISs identified using a range of kernel widths (0.05, 0.1, 0.25, 0.5, 1, 2.5, 5, 10, 30, 50, 100 and 150 kb). The kernel width controls the smoothness of the estimated number of insertions (de Ridder *et al.*, 2006). In other words, it controls the size of the genomic region in which neighbouring insertions affect the estimate of the number of insertions at the observed insertion. For each CIS, the flanking genes were identified using the Ensembl API version 45_36f and were compared to the genes identified using MC simulations. Among the 867 KC CISs, there were 765 where the nearest or further gene was represented in the *Efr*=0.01 MC list of candidates. Genes flanking the remaining 102 KC CISs may be missing from the MC list because insertions have been misassigned or because of differences between the two statistical approaches. As described in Section 1.4.2.1.2, for large screens, the statistically significant window size in the MC method may be so small that it is less than the width of biologically relevant CISs, causing these to be missed. Many of the CISs unique to the MC analysis are likely to be false positives since, at an *Efr* of 0.01, 261.44 randomly occurring CISs are expected.

652 CISs identified using a kernel width of 30 kb (*P*=0.05) were chosen for further analysis since this width, which was also used in Uren *et al.* (2008), should capture a high

proportion of biologically relevant CISs without splitting independent CISs or merging CISs that represent different types of mutation within a gene. For example, in genes that are mutated by multiple mechanisms, upstream enhancer mutations may form one CIS, while intragenic or downstream enhancer mutations may form another. For intragenic CISs, the gene containing the CIS was defined as the candidate cancer gene. For intergenic CISs, the flanking genes were compared to the list of candidates generated using Monte Carlo (MC) simulations. Where one of the flanking genes was within the MC list, this was chosen as the candidate gene. Where both nearest genes were within the MC list, both were initially included in the KC list because it is possible that a CIS could be mutating multiple nearby genes. Where neither nearest gene was in the MC list, the nearest genes were compared to a list of all genes to which insertions had been assigned, rather than just those to which significant CISs had been assigned using MC simulations. Genes could not be identified for 26 CISs, and these were assigned to genes manually, by observing insertions in the context of genes using the Ensembl DAS track (see Section 2.6). 102 CISs were assigned to more than 1 gene, and these were also assessed manually to determine whether one gene could be removed from the list. 14 CISs were removed where all insertions mapped to the same genomic coordinates, as these are likely to be artefacts. The final dataset comprised 630 CISs assigned to 608 genes. 30 CISs were associated with more than 1 gene, and 37 genes contained more than 1 CIS.

The lists of genes generated by the KC and MC methods were compared to the list of mouse orthologues of known cancer genes (see Section 2.2.3) and the Matthew's correlation coefficient (MCC) was calculated.

$$\mathrm{MCC} = \frac{TP \text{ x } TN - FP \text{ x } FN}{\sqrt{\ }((\mathrm{TP} + \mathrm{FP})(\mathrm{TP} + \mathrm{FN})(\mathrm{TN} + \mathrm{FP})(\mathrm{TN} + \mathrm{FN}))}$$

TP is the number of cancer genes in the candidate cancer gene list (true positives), FP is the number of non-cancer genes in the list (false positives), TN is the number of non-cancer genes not in the list (true negatives), and FN is the number of cancer genes not in the list (false negatives). Genes that were not in the list were calculated as all 18,017 mouse genes with human orthologues, as identified in Ensembl version 48, minus those in the list. MCC is used in machine learning as a measure of the quality of a prediction and it takes into account the counts of true and false positives and negatives to generate a single number that can be compared across predictions. The candidate cancer gene lists generated from the retroviral screen are expected to contain known cancer genes, and these can therefore be used as a measure of the quality of the list. MCC is a more useful measure than accuracy or coverage alone, especially when comparing lists of different lengths. For example, a short list may have high accuracy but low coverage, and for a longer list, the reverse may be true. MCC returns a value between -1 and +1, where +1 is a perfect prediction (i.e. in this case, the list contains all known cancer genes and no non-cancer genes), 0 is a random prediction, and -1 is an inverse prediction. The MCC score, plus accuracy and coverage, for the KC list and each MC list are shown in Table 2.7A (page 98). All MCCs generated in this analysis are positive but are very small because many of the genes are not known cancer genes. The KC list had the lowest coverage but the highest accuracy, and achieved the highest MCC score. As expected, the MC list generated using a *Efr* of 0.001 achieved a higher MCC score than the *Efr*=0.01 list since there should be 10-fold reduction in the number of randomly occurring CISs, and the higher accuracy more than compensated for the lower coverage. The list containing all genes that were assigned to insertions, rather than just those with statistically significant insertions, achieved the highest coverage but performed worst overall. Despite the fact that the list of known cancer genes is incomplete, measurement of the MCC score enables direct comparison of the gene lists and is likely to be meaningful. In light of these findings, the KC list was judged to be most accurate and was chosen for the cross-species comparative analyses performed in Chapter 5.

In order to gain an impression of whether the correct genes had been chosen for the KC CISs, known oncogenes were identified within the list of genes flanking each CIS. 37 oncogenes had been chosen, while 16 had not. Of the unselected oncogenes, 3 were genes nearest to the CIS, and 13 were further away. The insertions around these genes were analysed in the context of the mouse genome in Ensembl contigview. Only one oncogene nearest the CIS and one further away appeared to have been wrongly assigned, and for one additional nearest oncogene, it appeared that both this gene, and the correctly assigned gene might be mutational targets. The list of CIS genes was modified to include these three genes (*Rhoh*, *Cbl* and *Ccnd2*), but the results of the MCC comparison and analysis of the distribution of insertions suggest that, by and large, the most likely candidate gene has been selected.

Of the 39 cross-scale *Sleeping Beauty* CISs identified by the KC method, 36 were also present in the *Efr*=0.01 MC list. The remaining 3 were a long way from the nearest gene, further than the 100 kb limit used in the MC simulations. As *Sleeping Beauty* has low enhancer activity, these are likely to be non-oncogenic insertions that have preferentially inserted into a particular genomic region, or are mutating a gene that has not been identified in the Ensembl gene build. 79 genes from the *Efr*=0.01 MC list were not identified by the KC method. 5 KC genes were missing from the MC list generated using an *Efr* of 0.001, and in all cases the CIS was greater than 100 kb from the gene, and 27 of the *Efr*=0.001 MC genes were not identified by KC. The KC method is designed primarily for large datasets and may therefore miss a significant proportion of biologically relevant CISs in the *Sleeping Beauty* dataset. However, the MCC score is highest for the candidate gene list generated using the KC method, and other lists follow the same pattern as the corresponding lists generated from the retroviral dataset (Table 2.7B, page 98).

21 *Sleeping Beauty* CISs were identified using the KC-based framework with a kernel width of 30 kb and *P*=0.05 (Appendix B1), but 5 were situated close to the transposon array on chromosome 1, and 4 were situated close to the array on chromosome 15. These were removed from the list because they are likely to result from "local hopping" of the transposon (see Section 1.4.2.2.1). The T2/Onc splice acceptor and splice donor sequences are derived from exon 2 of the *En2* gene and exon 1 of the *Foxf2* gene, respectively (Collier *et al.*, 2005). Statistically significant CISs were identified in both these genes, and the insertions were found to cluster around the splice junctions used to construct T2/Onc (Figure 2.12). These CISs most likely represent artefacts resulting from the mapping of T2/Onc sequences, rather than flanking genomic sequences, to the mouse genome, and they were removed from the dataset. This leaves just 10 CISs and so, for the purposes of comparison with the retroviral dataset, discussed in Chapter 3, the more inclusive MC lists of candidate cancer genes were also used.

2.10.3 Final set of candidate genes

Following a survey of the candidate genes identified from the retroviral screen by the kernel convolution-based method, it became clear that some insertions mapped to exactly the same coordinates. This is unlikely to occur by chance, except where mutation of a very localised region of a gene is required for oncogenesis. There were 26 animals from which 2 tumours had been collected, 70 from which 3 tumours had been collected, and 3 from which 4 tumours had been collected. Where a tumour has spread to a different site,

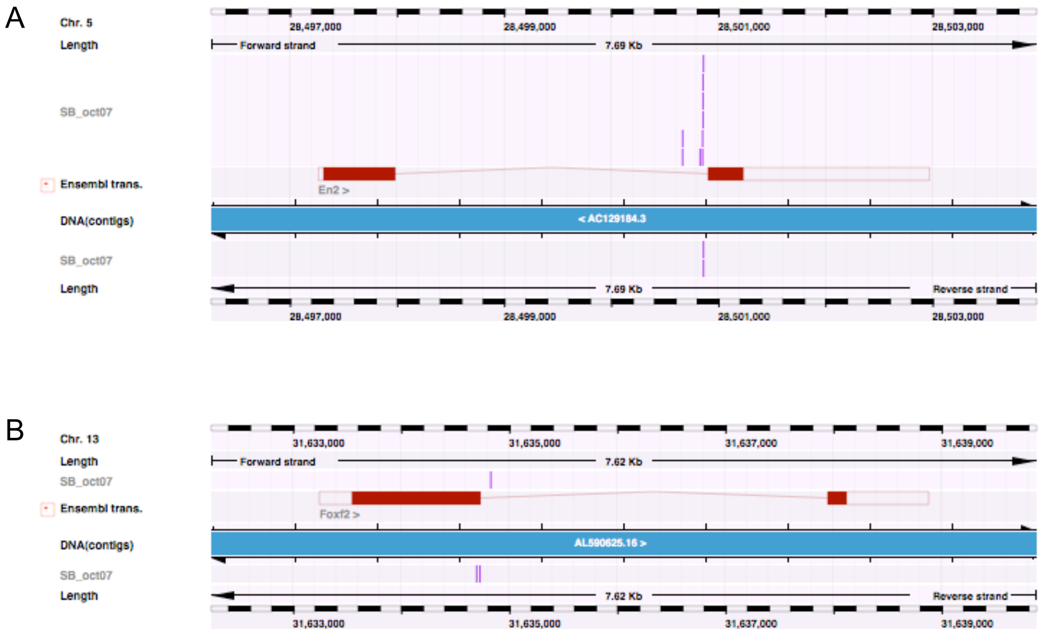

Figure 2.12. Insertions in *En2* (A) and *Foxf2* (B) are located at the splice junctions used to construct the T2/Onc transposon and are contaminating sequences. Insertions are shown as pink lines in the context of the Ensembl gene, shown in red. Insertions above and below the blue line are in the sense and antisense orientation, respectively.

a high proportion of insertions may be shared by both the original tumour and the secondary tumour, and this will influence the identification of significant CISs. Therefore, where 2 or more insertions from different tumours in the same animal occurred within 50 base pairs of one another, all but 1 of the insertions were removed from the dataset. A distance of 50 bp was chosen by counting the number of insertions that co-occurred within varying distances, and taking the distance at which the number levelled off. This reduced the dataset to 22,180 insertions. In addition, there is the possibility that insertions may map to the same position because of contamination during PCR. Therefore, if there were 2 or more sites in the genome where insertions from 2 tumours co-occurred within 10 bp, 1 of the co-occurring insertions was removed from the dataset at each location. A 10 bp window was used since it allows for a small amount of variation in the alignment of sequences using SSAHA2 (see Section 2.5), without significantly risking the removal of insertions that happen to fall into dense CISs. If the co-occurrence has resulted from aerosol contamination, it is assumed to be more likely that the insertion represented by the fewest number of reads is the contaminant and, therefore, in each case, this insertion was removed. The kernel convolution-based approach was applied to the final dataset of 20,114 insertions, and this resulted in 439 candidate cancer genes, of which 416 had a single CIS, 18 had 2 CISs, 2 had 3 CISs, 2 had 4 CISs and 1 had 5 CISs. The total number of CISs was 447, of which 24 were assigned to 2 genes. The CISs and associated genes are shown in Appendix B2.

2.11 Discussion

The aim of this chapter was to generate a reliable list of candidate cancer genes from insertional mutagenesis screens performed using the retrovirus MuLV and the *Sleeping Beauty* transposon T2/Onc. In order to maximise the number of insertions that could be identified within tumours, SSAHA2 was optimised to enable the mapping of as many reads as possible. The high number of unmapped reads was found to result from a high proportion of very short reads, especially in the *Sleeping Beauty* dataset, as well as reads containing genomic DNA of low complexity or low quality and reads that contained contaminating vector sequences. A small proportion may also result from errors in the mouse genome assembly. There did not seem to be any significant advantage in using BLASTN to map the reads, and as SSAHA2 is a faster algorithm, it is a good choice for mapping large numbers of reads. However, a possible alternative to SSAHA2 for future screens is the BLAST-Like Alignment Tool, BLAT (Kent, 2002). The UCSC Genome

Browser website (http://genome.ucsc.edu/, Kent *et al.*, 2002) uses BLAT to map users' sequences to the genome, and, because of its high speed and accuracy, BLAT has recently replaced BLAST as the default DNA search algorithm on the Ensembl website. Nevertheless, given the modest differences between SSAHA2 and BLAST (Altschul *et al.*, 1990), it is likely that BLAT would also perform similarly since the short, repetitive and low quality non-mapping reads can only be mapped at the expense of accuracy.

The reads were filtered to remove those in which the genomic DNA did not appear to represent the true location of the insertion. A gap between the genomic and retroviral DNA can result from low quality sequencing or the presence of unrelated DNA fragments within the clone, and efforts were made to retain low quality reads, whilst removing contaminating chimeric sequences. Comparisons between PCRs performed on the same tumours suggested that using more restriction enzymes and increasing the sequencing depth should increase the number of insertions that can be identified. Advances in sequencing technologies, such as 454 sequencing (see Section 1.4.2.1.2), will enable the use of more restriction enzymes and a greater depth of sequencing at a lower cost per read, thereby facilitating the identification of a higher proportion of insertions.

Identifying the genes that are most likely to have been mutated by insertions is hampered by the presence of enhancer insertions that can act at long range. Analysis of the distribution of insertions around mouse genes, and in particular, known cancer genes, suggested that the optimal distance is around 500-600 bp upstream, although the distance can be much greater, e.g. enhancer mutations can act as far as 270 kb downstream of the *Myc* promoter (Lazo *et al.*, 1990). It appears that downstream insertions are less likely to be oncogenic, although those in the sense orientation with respect to upstream genes may be more likely to contribute to oncogenesis. Enhancers can act over large distances via chromatin loop interactions, and they may therefore affect the activity of multiple genes (Uren *et al.*, 2005). However, analysis of the distribution of insertions around cancer genes suggests that, in general, enhancer insertions affect the promoters of the nearest, flanking genes.

Two approaches, Monte Carlo simulations (Suzuki *et al.*, 2002) and a kernel convolution-based statistical framework (de Ridder *et al.*, 2006), have been used to identify statistically significant CISs in the retroviral and transposon screens. Known cancer genes can be used as a partial set of true positives to evaluate candidate cancer genes in the vicinity of CISs, and Matthew's Correlation Coefficient was used to show that the kernel convolution-based framework gives the most reliable set of candidate cancer genes. The final set of candidates generated from the *Sleeping Beauty* screen comprises just 10 genes, reflecting the small size of the initial dataset and problems in mapping the reads. 439 candidates were identified from the MuLV screen, thereby supporting the theory that many genes contribute to tumourigenesis. The candidate cancer genes are analysed and characterised in Chapter 3.

Chapter 3 Analysis of mouse candidate cancer genes identified by insertional mutagenesis

3.1 Introduction

This chapter describes methods used to characterise the mouse candidate cancer genes identified by retroviral and transposon-mediated insertional mutagenesis. The integration of other cancer-associated datasets provides a means of filtering the genes to identify the strongest candidates for a role in tumourigenesis (see Section 1.5). Importantly, human cancer-associated datasets can be used to assess the relevance of insertional mutagenesis to human cancer. Analysis of Gene Ontology terms and gene pathways, as well as the identification of genes with binding sites for transcription factors relevant to cancer, can help to define the cancer pathways in which candidate genes may act. Comparative analyses between the mouse candidate genes and other cancer-related datasets are described in Section 3.2. The mutational profile varies between insertional mutagens, and is affected by insertional bias and the mechanisms by which the mutagen disrupts genes (see Section 1.4.2.1.1). Genes that are identified by multiple mutagens are strong candidates for a role in tumourigenesis. The candidate genes identified using MuLV and the *Sleeping Beauty* (SB) transposon T2/Onc are compared in Section 3.3.

The distribution of insertions in and around candidate cancer genes gives an indication of the likely mechanisms of mutagenesis (see Section 1.4.2.1.1) and therefore provides an insight into the structure and function of mutant oncoproteins. In Section 3.4.1, the distribution of intragenic insertions within candidates from the MuLV screen is explored, and genes are classified according to their predicted mutation type. The co-occurrence of both retroviral and transposon insertions within a localised region of a gene provides a strong indication that mutation within that region contributes to tumourigenesis. Therefore, in Section 3.4.2, the distribution of insertions in genes identified by both screens is used to predict the likely mechanisms of mutation. While it is clear that genes are frequently mutated by enhancer or promoter mutation or by premature termination of gene transcription, it is unclear whether the disruption of regulatory elements is a common mechanism of insertional mutagenesis. Therefore, Section 3.4.3 describes an analysis of insertions within regulatory features extracted from the Ensembl database.

Retroviral insertional mutagenesis identifies mainly oncogenes but it is also possible to identify tumour suppressor genes, and candidates are presented in Section 3.4.4. Finally, expression data for 18 MuLV-induced tumours is analysed in Section 3.4.5 in an attempt to confirm the deregulation of candidate genes.

Tumourigenesis is a multi-step process involving the co-operation of multiple cancer genes and pathways (see Section 1.2.3). Section 3.5 describes approaches for identifying co-operative cancer genes and presents a number of strong collaborations between genes identified in the retroviral screen. The work described in this chapter demonstrates the relevance of insertional mutagenesis to the study of human cancer, and identifies candidate cancer genes that warrant further investigation.

3.2 Comparative analyses between the insertional mutagenesis data and other cancer-related datasets

3.2.1 Description of the datasets

This section describes the datasets used for comparison with the candidate cancer genes identified by retroviral and transposon-mediated insertional mutagenesis. For all datasets where it was necessary to convert gene names to Ensembl identifiers, Ensembl BioMart (http://www.ensembl.org/biomart/index.html) was used. BioMart is a data mining tool that can be used to extract specific information from Ensembl for multiple genes simultaneously via a simple web interface. For all human datasets, mouse genes with human orthologues were also identified using Ensembl BioMart (version 48). The dataset of known cancer genes from the Cancer Gene Census (Futreal *et al.*, 2004) is described in Section 2.2.3.

3.2.1.1 The Retrovirus Tagged Cancer Gene Database (RTCGD)

As mentioned in Section 1.4.2.1.2, RTCGD (Akagi *et al.*, 2004; http://rtcgd.abcc.ncifcrf.gov/) is a database that manages data from retroviral and transposon-mediated insertional mutagenesis screens. All candidate cancer genes in the database were obtained from the website on 01/11/07. In total, the database contained 537 genes with unique MGI (Mouse Genome Informatics, http://www.informatics.jax.org/) symbols identified from 512 retroviral CISs (25 CISs had been assigned to 2 genes). The MGI symbols were used to identify 544 mouse Ensembl genes. 16 genes had 2 Ensembl identifiers (e.g. *Akap13*, which, according to Ensembl, is duplicated in 2 adjacent copies) and 9 could not be identified. 55 genes were identified from 52 transposon CISs (3 had been assigned to 2 genes) and all but 2 had Ensembl gene identifiers.

3.2.1.2 The Catalogue of Somatic Mutations in Cancer (COSMIC)

COSMIC stores and displays somatic mutation information relating to human cancers that has been curated from published scientific literature (Forbes *et al.*, 2006, see also Section 1.3.1). The complete set of mutations in COSMIC version 35 (dated 04/02/08) was downloaded from the website ftp://ftp.sanger.ac.uk/pub/CGP/cosmic/data_export. The list comprised 52,678 mutations in 1,550 genes from 45,743 tumours of 38 cancer types. Unique HGNC (HUGO Gene Nomenclature Committee; http://www.genenames.org/) human gene symbols were converted to human Ensembl gene identifiers. 1,521 mouse Emsembl genes were identified as having a human orthologue with somatic mutations in COSMIC. Individual genes were also searched against the COSMIC database via the website http://www.sanger.ac.uk/genetics/CGP/cosmic/.

3.2.1.3 Human candidate cancer genes from Sjöblom *et al.* (2006)

This dataset, which is described in Section 1.3.1, comprises 121 candidate breast cancer genes and 69 candidate colon cancer genes. HGNC symbols were used to extract Ensembl identifiers for all genes, and 181 mouse orthologues were identified.

3.2.1.4 Transcription factor binding sites

Mouse genes with Nanog and Oct4 binding sites were extracted from Loh *et al.* (2006), while human genes with p53 binding sites were extracted from Wei *et al.* (2006). Both datasets are described in Section 1.3.5. Ensembl gene IDs were identified from MGI symbols and/or Entrez Gene (http://www.ncbi.nlm.nih.gov/sites/entrez?db=gene) accession numbers for genes in the Nanog and Oct4 datasets, and from HGNC symbols and/or RefSeq (http://www.ncbi.nlm.nih.gov/RefSeq/) accession numbers for genes in the p53 dataset. Of the 3,006 Nanog binding sites, 2,408 were assigned to 1,923 mouse

Ensembl genes (i.e. some genes contained multiple binding sites), of which 1,889 encoded proteins or miRNAs. 817 mouse Ensembl genes, including 797 encoding proteins or miRNAs, were identified for 902 of the 1,083 Oct4 binding sites. 1,725 and 732 genes had human orthologues with Nanog and Oct4 binding sites, respectively. The p53 dataset contained 474 binding loci associated with human genes, of which 423 had Ensembl identifiers, resulting in 409 unique human Ensembl genes. 388 mouse Ensembl genes had a human orthologue with at least one binding site for p53.

3.2.1.5 Amplicons and deletions in paediatric acute lymphoblastic leukaemias

Regions of copy number change affecting more than one case of acute lymphoblastic leukaemia (ALL) were extracted from Mullighan *et al.* (2007; discussed in Section 1.3.3.3). In the publication, genomic coordinates were mapped to the human genome assembly NCBI 35, and these were therefore mapped across to NCBI 36 using the UCSC LiftOver tool (http://genome.ucsc.edu/cgi-bin/hgLiftOver). Overlapping amplicons and overlapping deletions were merged into single regions of copy number gain and loss, respectively. 1,905 mouse Ensembl genes contained a human orthologue within one of 8 non-overlapping amplicons, while 1,514 contained a human orthologue within one of 52 non-overlapping deletions.

3.2.1.6 Gene ontology (GO) terms

The GO project (http://www.geneontology.org) provides controlled vocabularies for describing genes and gene products in terms of their molecular function, their role in biological processes and their localisation to cellular components. The annotations assigned to genes and their products are known as GO terms. The public webserver g:Profiler (Reimand *et al.*, 2007, http://biit.cs.ut.ee/gprofiler/) was used to identify over-represented GO terms among the mouse candidate cancer genes, which were submitted as a list of mouse Ensembl gene identifiers. g:Profiler also identifies over-represented KEGG (http://www.genome.jp/kegg/) and REACTOME (http://www.reactome.org/) pathways, over-represented TRANSFAC (http://www.gene-regulation.com/) regulatory motifs, and miRNAs in miRBase (http://microrna.sanger.ac.uk/) for which target genes are over-represented among the candidate gene list.

3.2.2 Comparison with insertional mutagenesis data

The 1-tailed Fisher Exact Test was used to determine whether there was a significant overlap between candidate cancer genes identified by the MuLV and SB screens and those in other cancer-associated datasets. For comparison with the human datasets, the number of mouse orthologues of human genes was counted, whereas for comparison with the mouse Nanog and Oct4 binding site and RTCGD datasets, all mouse genes encoding proteins and miRNAs were counted. In total, there were 18,017 mouse Ensembl genes with human orthologues and 24,374 mouse genes that encoded proteins or miRNAs. All 439 of the candidate cancer genes identified in the MuLV screen using the kernel convolution (KC)-based approach (de Ridder *et al.*, 2006; see Section 2.10) encoded a protein or miRNA, and 384 had a human orthologue in Ensembl v48. All 10 of the *Sleeping Beauty* candidate genes were protein-coding, and 9 had human orthologues. For each significant CIS gene identified by the MuLV and SB screens, other cancer-associated datasets in which they also occurred are listed in Appendices C1 and C2, respectively.

Of the 439 CIS genes identified using MuLV, 118 (26.9%) were found among genes from other retroviral screens in the RTCGD database, and 6 (1.4%) were found among genes identified by transposon-mediated mutagenesis. This corresponds to a coverage of 21.7% and 11.3% of genes in the RTCGD database identified by retroviral and transposon-mediated mutagenesis, respectively. The larger overlap with candidates identified in retroviral screens suggests that retroviruses and transposons have different mutational profiles (see also Section 3.3) but may also reflect the fact that there is more retroviral data available. However, a high proportion of candidates are unique either to this dataset or to the RTCGD database. This may in part reflect limitations in insertion site identification, such as the number of restriction enzymes used in linker-mediated PCR and the depth of sequencing. Because of the variety of methods used to detect significant CISs, there may also be a high number of false detections in the RTCGD database, since 53% of CISs did not reach the significance threshold when the KC-based approach was applied to the data in RTCGD (de Ridder *et al.*, 2006). In addition, the RTCGD database contains genes that were identified by insertional mutagenesis on a range of genetic backgrounds and using a range of retroviral mutagens, and each mutagen and background may generate a different spectrum of candidates (see Section 3.5.1 for details on the identification of genotype-specific candidates). For example, *Sox4* and *Fgf3* are the 3rd and 6th most frequently mutated genes in RTCGD and yet they were not among the candidates identified in the MuLV screen. Almost all of the *Sox2* insertions were identified in mice with an AKxD or BXH2 strain background, while all of the *Fgf3* insertions were from a screen that used mouse mammary tumour virus (MMTV), rather than MuLV, as the retroviral mutagen.

Human orthologues of mouse candidate genes were enriched among oncogenes in the Cancer Gene Census (36 oncogenes, P=7.88x10^{-18}) and the COSMIC database (69 genes, P=1.36x10^{-9}). There were no recessive cancer genes among the candidates (P=1), demonstrating that the screen identifies predominantly oncogenes. Surprisingly, there were just 3 genes (*Lrrfip1*, *Nup214* and *Bcl11a*; P=0.74) that overlapped between the candidates of the retroviral screen and the candidates from Sjöblom *et al* (2006). This may reflect the fact that the Sjöblom dataset was an exon resequencing study of breast and bowel tumours exclusively, and it may be biased against genes mutated in lymphomas.

The 36 orthologues of mouse candidate cancer genes in the Cancer Gene Census were enriched for genes that are mutated in lymphoid tumours (31 genes, P=2.66x10^{-4}). This suggests that the retroviral screen mainly identifies genes that are important in the development of lymphoid malignancies. There was also a slight enrichment of genes that are mutated by chromosomal translocation, although this was not significant (31 genes, P=0.067). A more significant association might be expected because translocation is a common mechanism of mutation in lymphoid cancers, and a number of genes that are frequently targeted by insertional mutagenesis are involved in translocations in human tumours. In addition, MuLV mutagenesis may have a similar effect to translocations, since it often changes the regulatory environment of a gene and/or produces truncated oncoproteins. Chromosomal translocations, and leukaemias, lymphomas and mesenchymal tumours, all of which frequently harbour translocations, are over-represented in the Cancer Gene Census. This is partly because both translocation partners feature in the list of cancer genes, but also because, traditionally, cancer gene identification has been more frequently performed in these cancer types (Futreal, 2007). This may explain why the candidate cancer genes identified in the retroviral screen contain an over-representation of genes in the Cancer Gene Census, but that translocations are not over-represented among these candidates in the Census. Finally, there was an over-representation of known cancer genes that bear somatic mutations in human cancer (36 genes, *P*=0.0130). These findings demonstrate the efficacy of the MuLV retrovirus as a somatic mutagen that can be used to model the clonal evolution of human cancers, particularly those of lymphoid origin.

Human orthologues of mouse candidate cancer genes were significantly enriched among genes with p53 binding sites (14 genes, *P*=0.0394). The p53 pathway is important in tumourigenesis (see Section 1.2.6), and the identification of genes that act in this pathway provides further evidence that the screen has identified promising candidates for a role in cancer. It has been proposed that the CIS genes *Ptpre* and *Notch1* are upregulated by p53, while *Nedd4l* is downregulated (Wei *et al.*, 2006). *Ptpre* is required for p53-induced differentiation of IW32 erythroleukaemia cells (Tang and Wang, 2000), while upregulation of *Notch1* by p53 in human cancer cell lines contributes to cell fate determination (Alimirah *et al.*, 2007). *Nedd4l* is overexpressed in human prostate cancer cells (Qi *et al.*, 2003) and in the rare cutaneous T-cell lymphoma associated with Sézary Syndrome (Booken *et al.*, 2008), suggesting that suppression by p53 inhibits cancer growth. There was also a significant enrichment of mouse candidate cancer genes among genes with Nanog (53 genes, *P*=5.86x10^{-4}) and Oct4 (32 genes, *P*=1.64x10^{-5}) binding sites. Nanog and Oct4 regulate self-renewal, pluripotency and differentiation of ES cells (see Section 1.3.5). 9 CIS genes have binding sites for both Nanog and Oct4 and these include *Mycn*, *Il6st* and *Chd1*, which are upregulated in human ES cells, mesenchymal stem cells and haematopoietic stem/progenitor cells, respectively (Kim *et al.*, 2006a). *Il6st* (also known as *gp130*) is a key component of the signalling pathway required for the maintenance of embryonic stem cell pluripotency (Yoshida *et al.*, 1994) and mouse haematopoietic stem cell function (Audet *et al.*, 2001). These results suggest that a significant proportion of candidates may be involved in tumour cell self-renewal, therefore providing support for the cancer stem cell hypothesis, described in Section 1.2.3.2.

Mouse candidate genes with human orthologues were also over-represented in regions of copy number gain (54 genes, *P*=0.0180) and copy number loss (47 genes, *P*=5.82x10^{-3}) in paediatric ALL. CIS genes that were deleted in ALL included *Lef1*, *Ikzf1*, *Ikzf3*, *Etv6*, *Elf1* and *Erg*, while those that were amplified included *Runx1*, *Myb* and *Ahi1*. This suggests that genes that are mutated by insertional mutagenesis, and contribute to mouse tumourigenesis, may also be mutated by copy number changes in human cancers. However, the overlapping genes are implicated in B-cell development and differentiation, which are disrupted in human B-progenitor ALL and in MuLV-induced murine lymphomagenesis. It therefore remains to be seen whether CIS genes significantly overlap with regions of copy number change in other human cancers, and this is addressed in Chapters 4 and 5. The CIS genes may help to narrow down the candidates in regions of copy number change in the ALL dataset. For example, the deleted region on human chromosome 16q22.1 contains 11 genes, but the mouse orthologue of only 1 of these genes (*FAM65A*) is targeted by MuLV in insertional mutagenesis and therefore represents a putative target for deletion in ALL. Table 3.1 provides a list of the regions that are amplified and deleted in ALL and the CIS genes within these regions.

The candidate cancer genes were over-represented among genes in the KEGG pathways associated with acute and chronic myeloid leukaemia (*P*=2.14x10^{-13} and *P*=1.75x10^{-7}, respectively) and Jak-STAT signalling, and in the T cell receptor signalling KEGG and REACTOME pathways (*P*=1.35x10^{-5} and *P*=1.96x10^{-6}, respectively). This is encouraging, since the genes are candidates for a role in lymphomagenesis. However, genes were also over-represented in the endometrial cancer KEGG pathway (*P*=7.14x10^{-4}), demonstrating that some of the candidates (including *Pik3cd*, *Pik3r5*, *Akt1*, *Lef1*, *Myc*, *Ccnd1* and *Tcf7*) also contribute to other cancer types. Over-represented GO terms are listed in Table 3.2. These include terms related to the development, differentiation and proliferation of B- and T-cells, reflecting the lymphoid origin of the mouse tumours, and terms specifically related to cancer, such as cell proliferation, apoptosis, angiogenesis, cell motility and kinase activity.

Four transcription factor binding sites from the TRANSFAC database were also over-represented among the candidate genes. The most significant was the MAZ (Myc-associated zinc finger protein) binding matrix (TF:M00649, *P*=1.49x10^{-8}), which binds the MAZ transcription factor. MAZ interacts with MYC and histone deacetylases, and *MAZ* overexpression drives expression of the oncogene *PPARγ1* in human breast cancer cells (Wang *et al.*, 2008). It is also overexpressed in acute myeloid leukaemia (Greiner *et al.*, 2000) and in the terminal phase of chronic myeloid leukaemia (Daheron *et al.*, 1998). The second most significant binding matrix was TF:M01104 (*P*=2.51x10^{-6}), which binds the mouse Movo-b zinc finger protein. This protein is highly expressed in the mouse testis (Unezaki *et al.*, 2004), and has no known role in tumourigenesis, but has been shown to be involved in vascular angiogenesis in the developing embryo (Unezaki *et al.*, 2007). Finally, binding matrices for transcription factors LRF (leukaemia/lymphoma

A

Chromosome	Start (bp)	End (bp)	Comment in Mullighan et al.	CIS genes in region
1	127000000	247249719	719 genes telomeric of *PBX1*	*Mef2d, Nid1, Slamf6, Cd48, Anp32e, Lyst, Btg2, Ptprc, Mixl1, Ccdc19, Sell, Ppp2r5a, Zbtb7b, Rorc, Slamf7, Mcl1, Itpkb, AI848100, Rcsd1, 5730559C18Rik, Irf2bp2*
2	1	31987853	235 genes	*D12Ertd553e, Mycn*
6	1	26216000	190 genes	*Irf4, Exoc2, Rreb1, Sox19*
6	135556000	135714000	*MYB, MIRN548A2, AHI1*	*Myb, Ahi1*
9	60000000	140273252	155 genes telomeric of *ABL1*	*Nup214, Phyhd1, Sema4d, Gadd45g, Ccrk, Eng, Gfi1b, Ak1, Egfl7, Notch1, Coro2a, A2AN91_MOUSE, Akna, A130092J06Rik*
10	1	40290000	All 10p	*Cugbp2, Map3k8, Il2ra, Zfp438*
21	32896000	35199000	33 genes including *Runx1*	*Runx1, Ifnar1*
22	1	21888000	277 genes telomeric of *BCR*	*Bid, Tuba8, BC030863, Cecr5, Vpreb2*

B

Chromosome	Start (bp)	End (bp)	Comment in Mullighan et al.	CIS genes in region
2	232347739	242951149	124 genes	*Lrrfip1*
4	109254845	109303845	*LEF1*	*Lef1*
5	163535000	180857866	172 genes	*C330016O10Rik, Mgat1*
7	1	58058273	All 7p	*Stard3nl, Mafk, Ikzf1, Mad1l1, Lfng, Hibadh, Hoxa7, Sdk1, 3110082I17Rik, Jazf1*
9	1	50600000	All 9p	*Cd72, Anxa2, Dock8*
11	117882000	118379000	16 genes distal to *MLL*	*Treh, Bcl9l*
12	11694055	11939588	*ETV6*	*Etv6*
13	40453000	40484000	*ELF1*	*Elf1*
13	47885000	47968000	*RB1*	*Rcbtb2*
16	66116000	66423000	*FAM65A, CTCF, RLTPR, ACD, PARD6A, C16orf48, LOC388284, GFOD2, RANBP10, TSNAXIP1, CENPT*	*2310066E14Rik (FAM65A)*
17	1	18837000	383 genes	*Lgals9, AA536749, Prr6, Pik3r5, Ntn1, Slc43a2, Ovca2, Smg6, Rtn4rl1*
17	35185000	35230000	*IKZF3*	*Ikzf3*
19	229000	1531000	63 genes telomeric to *TCF3*	*Ptbp1, Arid3a, Midn*
20	27000000	62435964	All 20q	*Bcl2l1, Serinc3, Stk4, Ndrg3, Sla2, Ncoa3, Ppp1r16b, Prkcbp1, Zfp217*
21	38706000	38729000	*ERG*	*Erg*

Table 3.1. The human orthologues of mouse CIS genes can help to identify the critical gene(s) in regions of copy number change in acute lymphoblastic leukaemias (ALLs) from Mullighan *et al*. (2007). Recurrent amplifications and deletions in ALLs that contain CIS genes are shown in Tables A and B, respectively. The coordinates of each region in the NCBI 36 human assembly are shown. "Comment in Mullighan *et al*. (2007)" provides details of how the region was characterised in the publication. "CIS genes in region" provides a list of mouse genes that have human orthologues mapping to each region.

P-value	CIS genes	GO ID	Ontology	GO term
8.90E-14	**167**	**GO:0065007**	**BP**	**biological regulation**
4.89E-14	155	GO:0050789	BP	regulation of biological process
3.29E-14	64	GO:0048518	BP	positive regulation of biological process
3.98E-14	148	GO:0050794	BP	regulation of cellular process
2.07E-12	57	GO:0048522	BP	positive regulation of cellular process
1.01E-07	19	GO:0008284	BP	positive regulation of cell proliferation
1.48E-06	27	GO:0031325	BP	positive regulation of cellular metabolic process
2.18E-06	10	GO:0050867	BP	positive regulation of cell activation
1.51E-05	6	GO:0045787	BP	positive regulation of cell cycle
1.52E-05	79	GO:0019219	BP	regulation of nucleobase, nucleoside, nucleotide and nucleic acid metabolic process
5.73E-06	76	GO:0051252	BP	regulation of RNA metabolic process
3.83E-05	76	GO:0045449	BP	regulation of transcription
2.77E-05	73	GO:0006355	BP	regulation of transcription, DNA-dependent
4.32E-07	26	GO:0006357	BP	regulation of transcription from RNA polymerase II promoter
1.46E-11	30	GO:0051726	BP	regulation of cell cycle
1.29E-10	54	GO:0048523	BP	negative regulation of cellular process
1.60E-08	28	GO:0042127	BP	regulation of cell proliferation
2.47E-07	13	GO:0050865	BP	regulation of cell activation
4.84E-07	30	GO:0009966	BP	regulation of signal transduction
2.71E-06	89	GO:0031323	BP	regulation of cellular metabolic process
1.51E-05	9	GO:0051270	BP	regulation of cell motility
4.54E-05	80	GO:0010468	BP	regulation of gene expression
8.52E-11	57	GO:0048519	BP	negative regulation of biological process
8.80E-08	44	GO:0050793	BP	regulation of developmental process
3.09E-10	28	GO:0051094	BP	positive regulation of developmental process
5.83E-08	30	GO:0043067	BP	regulation of programmed cell death
2.86E-08	20	GO:0043068	BP	positive regulation of programmed cell death
2.47E-08	20	GO:0043065	BP	positive regulation of apoptosis
2.38E-06	15	GO:0006917	BP	induction of apoptosis
2.38E-06	15	GO:0012502	BP	induction of programmed cell death
4.87E-08	30	GO:0042981	BP	regulation of apoptosis
6.43E-06	21	GO:0051093	BP	negative regulation of developmental process
2.18E-07	17	GO:0002682	BP	regulation of immune system process
2.47E-07	13	GO:0002694	BP	regulation of leukocyte activation
1.26E-07	13	GO:0051249	BP	regulation of lymphocyte activation
6.55E-09	13	GO:0050863	BP	regulation of T cell activation
1.24E-06	5	GO:0050854	BP	regulation of antigen receptor-mediated signaling pathway
1.76E-05	4	GO:0050856	BP	regulation of T cell receptor signaling pathway
3.52E-06	8	GO:0002683	BP	negative regulation of immune system process
1.24E-05	13	GO:0002684	BP	positive regulation of immune system process
2.18E-06	10	GO:0002696	BP	positive regulation of leukocyte activation
1.69E-06	10	GO:0051251	BP	positive regulation of lymphocyte activation
1.31E-07	10	GO:0050870	BP	positive regulation of T cell activation
3.40E-05	6	GO:0042102	BP	positive regulation of T cell proliferation
2.43E-06	23	GO:0051239	BP	regulation of multicellular organismal process
9.53E-06	89	GO:0019222	BP	regulation of metabolic process
2.35E-06	27	GO:0009893	BP	positive regulation of metabolic process
2.86E-05	9	GO:0040012	BP	regulation of locomotion
1.29E-12	**27**	**GO:0001775**	**BP**	**cell activation**
6.04E-12	25	GO:0045321	BP	leukocyte activation
6.15E-12	24	GO:0046649	BP	lymphocyte activation
1.13E-11	19	GO:0042110	BP	T cell activation
3.85E-05	9	GO:0046651	BP	lymphocyte proliferation
2.29E-05	8	GO:0042098	BP	T cell proliferation
2.71E-11	**47**	**GO:0002376**	**BP**	**immune system process**
1.21E-05	6	GO:0001776	BP	leukocyte homeostasis
3.60E-05	5	GO:0002260	BP	lymphocyte homeostasis
2.89E-05	4	GO:0043029	BP	T cell homeostasis
8.96E-10	**41**	**GO:0007049**	**BP**	**cell cycle**
9.55E-10	**61**	**GO:0007242**	**BP**	**intracellular signaling cascade**
1.96E-09	**75**	**GO:0048869**	**BP**	**cellular developmental process**
1.96E-09	75	GO:0030154	BP	cell differentiation
1.15E-10	19	GO:0002521	BP	leukocyte differentiation
3.10E-09	15	GO:0030098	BP	lymphocyte differentiation
6.62E-06	9	GO:0030217	BP	T cell differentiation
5.95E-09	**103**	**GO:0032502**	**BP**	**developmental process**
3.25E-09	80	GO:0048856	BP	anatomical structure development
4.15E-08	39	GO:0016265	BP	death
2.83E-05	45	GO:0009653	BP	anatomical structure morphogenesis
6.25E-09	**84**	**GO:0007275**	**BP**	**multicellular organismal development**
2.88E-09	72	GO:0048731	BP	system development
1.80E-11	27	GO:0002520	BP	immune system development
4.63E-12	27	GO:0048534	BP	hemopoietic or lymphoid organ development
2.02E-11	25	GO:0030097	BP	hemopoiesis
5.95E-09	62	GO:0048513	BP	organ development
1.56E-08	**34**	**GO:0008283**	**BP**	**cell proliferation**
1.97E-05	7	GO:0050673	BP	epithelial cell proliferation
3.85E-05	9	GO:0032943	BP	mononuclear cell proliferation
1.59E-07	**53**	**GO:0048468**	**BP**	**cell development**
4.15E-08	39	GO:0008219	BP	cell death
1.67E-08	39	GO:0012501	BP	programmed cell death
1.24E-08	39	GO:0006915	BP	apoptosis
1.16E-05	**13**	**GO:0001525**	**BP**	**angiogenesis**
1.54E-05	**178**	**GO:0043170**	**BP**	**macromolecule metabolic process**
3.08E-09	153	GO:0043283	BP	biopolymer metabolic process
1.57E-05	63	GO:0043412	BP	biopolymer modification
7.15E-06	62	GO:0006464	BP	protein modification process
3.48E-05	54	GO:0043687	BP	post-translational protein modification
1.58E-05	**16**	**GO:0045944**	**BP**	**positive regulation of transcription from RNA polymerase II promoter**
2.30E-05	**7**	**GO:0030183**	**BP**	**B cell differentiation**
2.36E-05	**79**	**GO:0006350**	**BP**	**transcription**
2.44E-05	**86**	**GO:0016070**	**BP**	**RNA metabolic process**
1.71E-05	75	GO:0032774	BP	RNA biosynthetic process
1.63E-05	75	GO:0006351	BP	transcription, DNA-dependent
9.62E-06	25	GO:0006366	BP	transcription from RNA polymerase II promoter
2.81E-05	**6**	**GO:0050851**	**BP**	**antigen receptor-mediated signaling pathway**
2.45E-06	6	GO:0050852	BP	T cell receptor signaling pathway
3.38E-05	**11**	**GO:0001816**	**BP**	**cytokine production**
3.88E-05	**15**	**GO:0007265**	**BP**	**Ras protein signal transduction**
3.98E-05	**36**	**GO:0016310**	**BP**	**phosphorylation**
5.64E-06	35	GO:0006468	BP	protein amino acid phosphorylation
2.82E-07	**179**	**GO:0005515**	**MF**	**protein binding**
3.66E-06	**54**	**GO:0030528**	**MF**	**transcription regulator activity**
5.08E-06	41	GO:0003700	MF	transcription factor activity
3.95E-05	**39**	**GO:0016301**	**MF**	**kinase activity**
2.40E-05	**27**	**GO:0004674**	**MF**	**protein serine/threonine kinase activity**

Table 3.2. Over-represented GO terms among CIS genes identified using MuLV. “CIS genes” is the number of CIS genes annotated for each term. The ontologies shown are biological process (BP) and molecular function (MF). Terms are staggered to show GO term hierarchies, with terms of equivalent hierarchy being listed in order of decreasing significance. Terms associated with T- and B-cells are shown in blue.

related factor; TF:M01100) and VDR (vitamin D receptor; TF:M00444) were also significantly over-represented among candidate genes (*P*=1.38x10^{-5} and *P*=1.47x10-*5*, respectively). LRF is a master regulator of oncogenesis that directly represses transcription of the tumour suppressor gene *p19^{ARF}* (Maeda *et al.*, 2005) and plays an essential role in determining B- versus T-cell fate (Maeda *et al.*, 2007). VDR is also widely implicated in human tumourigenesis (for review, see Thorne and Campbell, 2008). Specific analysis of over-represented GO terms for the genes associated with each transcription factor showed that all were enriched for terms relating to the cell cycle and KEGG pathways for acute and chronic myeloid leukaemia. Only candidates associated with the Movo-b binding site were enriched for genes with protein serine/threonine kinase activity (*P*=2.21x10^{-5}), suggesting that Movo-b may play an important role in regulating protein kinases. Likewise, only genes with Lrf binding sites were enriched for terms associated with apoptosis, suggesting that Lrf may also repress other tumour suppressor genes, including *Wwox* and *Trp53inp1*.

The output from g:Profiler also showed that candidate genes were over-represented among the predicted targets of 3 miRNAs: mmu-miR-449b (*P*=4.31x10^{-5}), mmu-miR-449c (*P*=9,69x10^{-6}) and hsa-miR-565 (*P*=6.29x10^{-5}), which suggests that these miRNAs may play an important role in regulating genes involved in tumourigenesis. The list of candidates also included 5 genes that encode miRNAs: *mmu-miR-142*, *mmu-miR-17*, *mmu-miR-802*, *mmu-miR-181c* and *mmu-miR-23a*. miRNAs play an important role in haematopoiesis, and miRNA deregulation has been widely observed in leukaemias and lymphomas (for review, see Garzon and Croce, 2008). The human orthologues of *mmu-miR-142* and *mmu-miR-181* are both implicated in the regulation of mammalian haematopoiesis, since *hsa-miR-142* is at a translocation site within a case of aggressive B-cell leukaemia, while B-cell-specific *hsa-miR-181* promotes B cell differentiation (Chen and Lodish, 2005). Deregulated miRNAs also contribute to cancer in other cancer types. *hsa-miR-23a*, the human orthologue of *mmu-miR-23a*, is upregulated in human hepatocellular carcinomas (Kutay *et al.*, 2006), while the human orthologue of *mmu-miR-17* is implicated as an oncogene in a range of cancers, and is discussed further in Section 4.5.2.1.2.

Of the 10 candidates genes in the *Sleeping Beauty* dataset, 5 had been previously identified by retroviral insertional mutagenesis in RTCGD and were also identified in the MuLV screen described herein (see Section 3.3), while 3 had been previously identified by transposon-mediated mutagenesis. Therefore, a higher percentage (42.9%) of these candidates than those in the MuLV screen (1.4%) overlapped with other transposon screens, again highlighting the different mutational profiles of the two mutagens. The *Sleeping Beauty* dataset is small, and none of the genes had Nanog or Oct4 binding sites (*P*=1 for both tests), while 1 had a p53 binding site (*P*=0.178) and 1 overlapped with the Sjöblom *et al.* (2006) dataset (*P*=0.087). However, 6 candidates were identified in the Cancer Gene Census and 5 had mutations in COSMIC. This is significantly greater than the number expected by chance (*P*=4.26x10^{-9} and *P*=4.02x10^{-4}, respectively). 5 genes were dominant cancer genes in the Cancer Gene Census (*P*=1.16x10^{-7}), while 1 (*PTEN*) was recessive (*P*=0.032). Candidate genes were also enriched in regions of copy number loss (3 genes, *P*=0.0338) but not in regions of gain (2 genes, *P*=0.2450) in the Mullighan *et al.* (2007) dataset. There was an over-representation of genes (*AKT2* and *PTEN*) in the melanoma, endometrial cancer and glioma KEGG pathways (*P*=1.68x10^{-3}, *P*=9.26x10^{-4} and *P*=1.36x10^{-3}, respectively) and in the REACTOME pathway associated with negative regulation of the PI3K/AKT network (*P*=8.11x10^{-5}). Candidate genes were also over-represented in the B-cell receptor signalling pathway (*AKT2* and *PPP3CA; P*=1.40x10^{-3}). Only 2 GO terms, “regulation of biological process” (8 genes, *P*=6.10x10^{-5}) and “transcription factor activity” (*Notch1*, *Fli1*, *Myb*, *Ikzf1* and *Erg*; *P*=3.39x10^{-5}) were over-represented, but the test was limited by the small size of the dataset.

The enrichment of candidate genes from the MuLV and *Sleeping Beauty* screens within human cancer-associated datasets demonstrates the efficacy of insertional mutagenesis as a tool for discovering human cancer genes, as well as those in mice. In addition, overlaying other cancer-associated datasets on to the insertional mutagenesis data helps to characterise the candidate genes and facilitates the identification of novel cancer genes. However, the approach is biased towards the identification of genes involved in the development of cancers of lymphoid origin. Candidate genes were positively associated with the Mullighan *et al.* (2007) dataset, which was generated using ALLs, and the Cancer Gene Census and COSMIC database, in which genes implicated in haematopoietic and lymphoid tumourigenesis are over-represented. Many of the over-represented GO terms were also directly related to the differentiation and activation of B- and T-cells. Conversely, candidates showed no significant association with the Sjöblom *et al.* (2006) dataset of colon and breast cancer genes. This highlights the importance of developing insertional mutagenesis screens that can induce other types of tumour, e.g. by integrating tissue-specific promoters into transposons or by spatial and temporal regulation of transposase expression (see Section 1.4.2.2.1). The datasets discussed in this section are further referred to in the proceeding sections and chapters in relation to individual cancer gene candidates.

3.3 Comparison of candidate cancer genes in the MuLV and Sleeping Beauty datasets

There was a significant overlap between the lists of candidate cancer genes obtained using the retroviral and *Sleeping Beauty* (SB) screens. There was an overlap of 5 genes (P=9.64x10^{-31}) when both lists generated using the kernel convolution (KC)-based approach were compared. Comparing the KC list of candidates from the retroviral screen to the *Sleeping Beauty* candidates generated using Monte Carlo (MC) simulations (*Efr*=0.001) produced an overlap of 10 genes (P=1.95x10^{-8}). The KC lists therefore yielded the most significant overlap, which is consistent with the work described in Section 2.10.2, where the KC method was proposed to generate the most reliable set of candidate genes.

The distributions of MuLV and T2/Onc insertions across the mouse genome are shown in Figures 3.1A and 3.1B, respectively. The figures also show the location of candidate cancer genes that were identified by both screens, as well as all other *Sleeping Beauty* candidates identified using the KC method and a subset of the most frequently disrupted candidates from the MuLV screen. The most frequently mutated genes in MuLV-induced tumours were *Gfi1/Evi5*, *Myc/Pvt1* and *Ccnd3*. These genes had insertion densities of 427.28, 314.19 and 172.09, respectively, using the KC method with kernel width 30 kb. Remarkably, none of the SB-induced tumours contained insertions in or around these genes. While these genes are known to contribute to tumourigenesis, the frequency of insertions may reflect the bias of retroviruses to insert into particular sites in the genome (see Section 1.4.2.1.1). In addition, many of the MuLV insertions in these genes appear to be enhancer mutations, which do not feature in the *Sleeping Beauty* screen because T2/Onc has low enhancer activity. Therefore, the frequency of insertions may reflect the choice of mutagen and does not imply that a gene would contribute to a similar proportion of spontaneous tumours in the mouse.

Conversely, a significant CIS comprising insertions in 6 SB-induced tumours was identified in the tumour suppressor gene *Pten*, but none of the MuLV-induced tumours

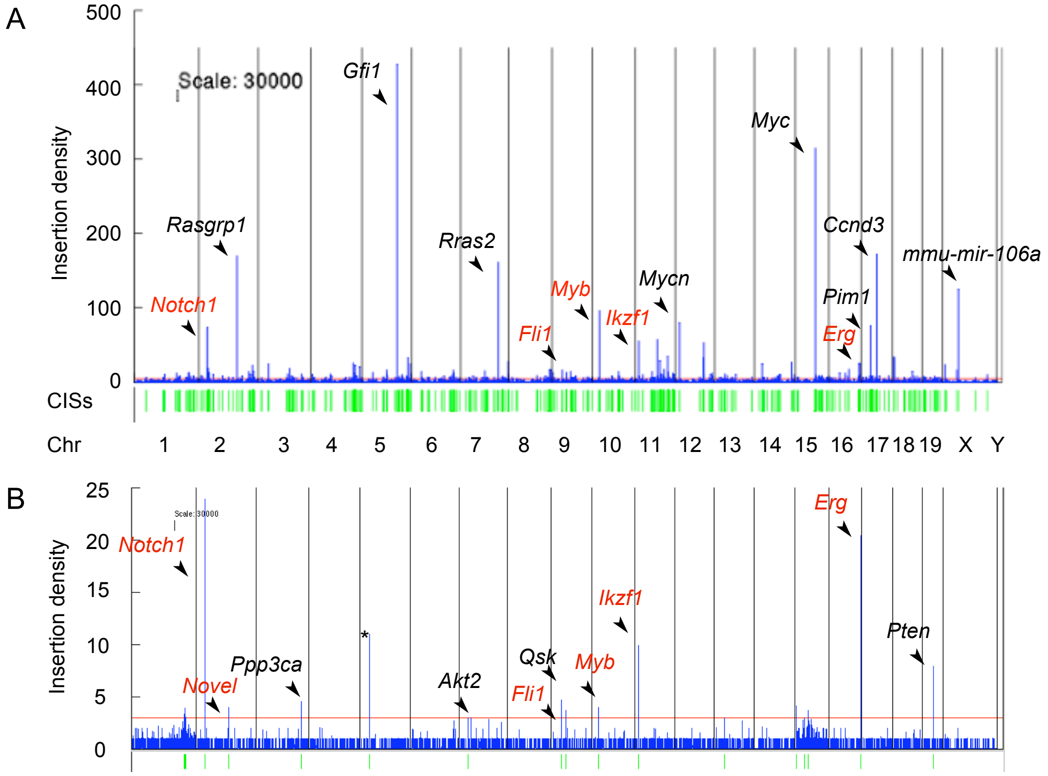

Figure 3.1. MuLV (A) and T2/Onc (B) insertions across the mouse genome. The plots show the density of insertions calculated using the kernel convolution-based method (de Ridder *et al*., 2006) with a kernel width of 30 kb. Common insertion sites (CISs) are shown in green. The red line represents the threshold above which insertions form significant CISs (*P*<0.05). Gene names shown in red contain significant CISs in both screens. Gene names shown in black contain significant common insertion sites that are unique to one screen. * marks artefacts in *En2.*

contained an insertion within this gene. None of the T2/Onc insertions were in *Bloom*-deficient tumours, which have an increased propensity for insertions within tumour suppressor genes (see Section 1.4.2.1.1). For 3 tumours, multiple insertions were identified, suggesting that the gene may be inactivated by insertions affecting both copies, rather than by one insertion in a region of loss of heterozygosity (LOH) (Figure 3.2A). The lack of MuLV insertions may reflect the fact that MuLV prefers to insert near to transcriptional start sites (Section 1.4.2.1.2) and is therefore more biased towards the identification of oncogenes than is the T2/Onc transposon. These observations suggest that MuLV and the T2/Onc transposon are unique mutagens with complementary mutagenic profiles, and that performing screens with both these mutagens can identify more candidate cancer genes than with either alone.

As well as the distinct differences between the mutagenic profiles of MuLV and T2/Onc, a number of known and implicated cancer genes (*Notch1*, *Erg*, *Ikzf1*, *Myb* and *Fli1*) contained significant CISs in both screens. The co-occurring MuLV and T2/Onc insertions within these genes are discussed in Section 3.4.2. After *Myc/PvtI*, *Gfi1/Evi5* and *Ccnd3*, the most highly mutated genes in the MuLV screen were *Rasgrp1* (insertion density 169.59) and *Rras2* (insertion density 161.49). Although significant *Sleeping Beauty* CISs were not identified in these genes using the KC method, *Rras2* did contain 1 T2/Onc insertion, and *Rasgrp1* contained 3 T2/Onc insertions, which was significant using the Monte Carlo method with *Efr*=0.005. Likewise, *AA536749, Zmiz1*, and known oncogenes *Irf4* and *Etv6*, contained significant MuLV CISs identified using the KC method and T2/Onc CISs that were significant using the MC method with *Efr*=0.001.

The human orthologue of *AA536749* is myosin phosphatase Rho-interacting protein (*p116Rip* or *M-RIP*). p116Rip is a filamentous actin-binding protein that is capable of disassembling the actomyosin-based cytoskeleton and acts downstream of RhoA (Mulder *et al.*, 2003). The actin cytoskeleton plays a role in many cancer-related functions such as cell motility, cell differentiation, cell survival and cell division. The LIM kinases (LIMK1 and LIMK2) are regulators of actin dynamics that also act downstream of Rho GTPase and play an important role in tumour invasion and metastasis (Scott and Olson, 2007). The identification of insertions in tumours generated by both mutagens suggests that *p116Rip* may also play an important role in tumourigenesis.

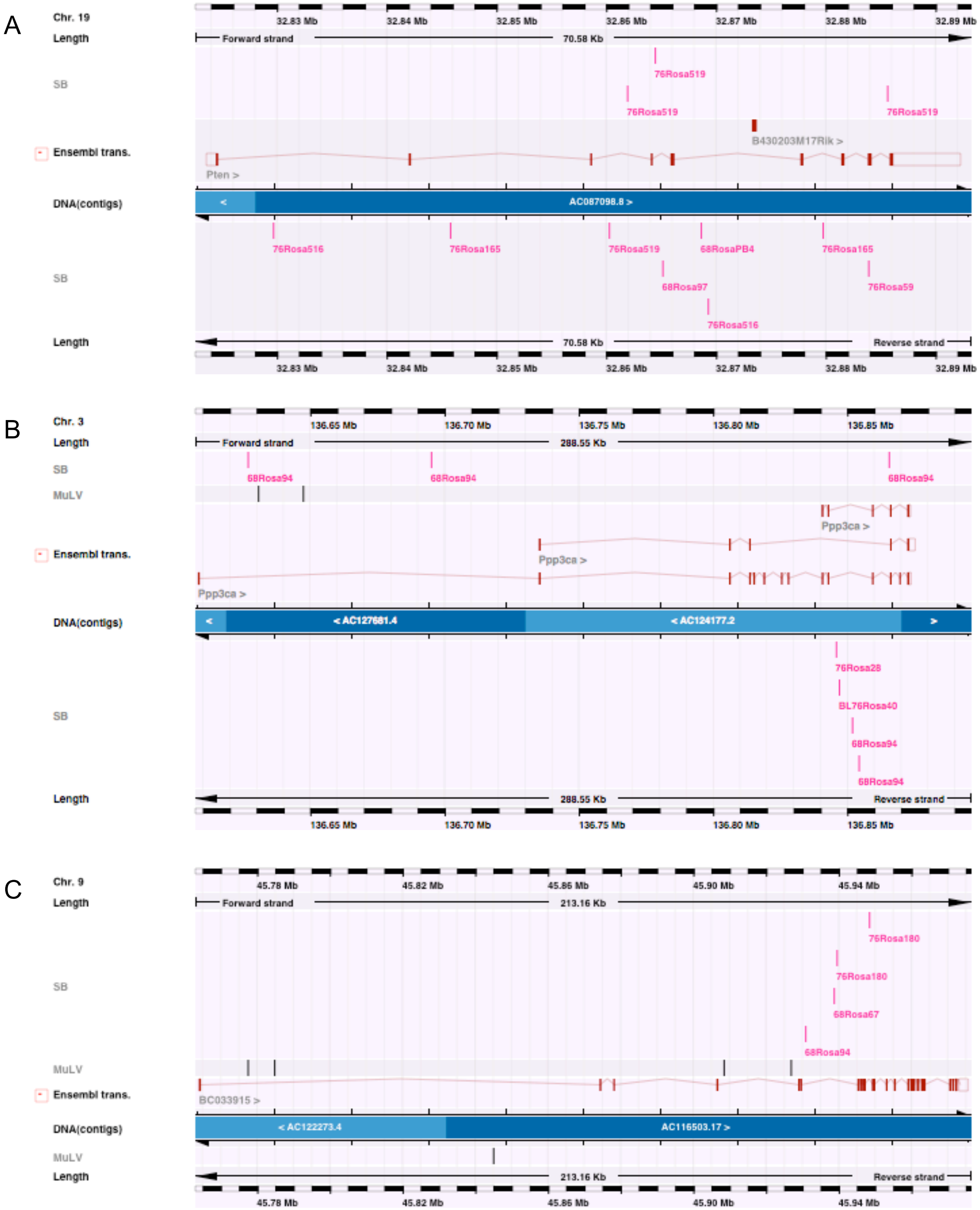

Figure 3.2. Known and putative tumour suppressor genes identified in the *Sleeping Beauty* (SB) screen. *Pten* (A) did not contain any MuLV insertions. *Ppp3ca* (B) and *BC033915* (C) contained MuLV insertions but not in a statistically significant CIS. For all genes, there was at least 1 SB tumour that contained more than 1 insertion, suggesting that inactivation of both genes may be required for tumourigenesis. The tumour in which each T2/Onc insertion was identified is provided as a label under the insertion, which is shown in pink. Ensembl genes are shown in red and, where applicable, MuLV insertions are shown as black vertical lines. Insertions above and below the blue line are in the forward and reverse orientation, respectively.

Zmiz1 enhances p53 (Lee *et al.*, 2007) and Smad transcriptional activity (Li *et al.*, 2006b), suggesting a tumour suppressive role. However, it is also required for vasculogenesis (Beliakoff *et al.*, 2008) and activates transcription of the androgen receptor (Beliakoff and Sun, 2006; Sharma *et al.*, 2003), which contributes to the formation and progression of human prostate cancer (for review, see Nieto *et al.*, 2007). In addition, a fusion between *ZMIZ1* and *ABL1* was recently identified in a human B-cell acute lymphoblastic leukaemia (Soler *et al.*, 2008). There are 4 other known fusion partners (*BCR*, *ETV6*, *NUP214* and *EML1*) for *ABL* in human haematological malignancies, and one putative partner, *RCSD1* (De Braekeleer *et al.*, 2007). Remarkably, *Zmiz1*, *Etv6*, *Nup214* and *Rcsd1* all contained statistically significant CISs in the retroviral screen and, although not significant, *Bcr* contained 2 MuLV insertions, while *Eml1* contained 1 MuLV and 1 T2/Onc insertion. This reflects the fact that mutagenesis by MuLV often resembles the effects of translocation, as mentioned in Section 3.3.2.

The remaining candidates containing significant *Sleeping Beauty* CISs identified using the KC method also contained retroviral insertions, although not significant CISs. The known oncogene *Akt2* and the serine/threonine protein phosphatase *Ppp3ca* contained 1 and 2 MuLV insertions, respectively. One of the 3 SB-induced tumours in which *Ppp3ca* was disrupted contained 4 insertion sites, suggesting that this gene encodes a tumour suppressor (Figure 3.2B). This is supported by research showing that *Ppp3ca* can dephosphorylate cyclin dependent kinases (CDKs), therefore potentially inhibiting cell cycle progression (Cheng *et al.*, 1999, see Section 1.2.6 for more on CDKs). In addition, *Ppp3ca* overexpression increases the levels of p53 and inhibits cell growth (Ofek *et al.*, 2003), and expression is reduced in androgen-independent prostate cancer cells (Singh *et al.*, 2008).

The gene encoding serine/threonine protein kinase BC033915 (known as QSK in humans) was also identified in both screens, and although the MuLV CIS was not significant, it did contain 5 retroviral insertions (Figure 3.2C). In the COSMIC database, 2 of the 296 human tumour samples that have been tested for mutations in the *QSK* gene contain missense mutations. One heterozygous S882C substitution was identified in the primary renal cell carcinoma PD1583a, which contains just one other small intragenic mutation in 519 genes examined, and a P836S substitution (zygosity unknown) was identified in the non-small cell lung cancer cell line NCI-H1770, which contains 201 small intragenic mutations in 4,688 genes examined. 1 silent mutation (heterozygous substitution R476R) was also identified in the malignant melanoma cell line MZ7-mel, but this line contains 428 mutations in 4,668 genes examined and therefore appears to have a hypermutable phenotype. Both missense mutations are located in a glutamine-rich region (Prosite profile PS50322). All of the retroviral and transposon insertions in *QSK* precede this region and may therefore produce truncated gene transcripts in which the region is missing. 2 heterozygous, missense mutations were also identified in the resequencing study by Sjöblom *et al* (2006). QSK and 11 other kinases related to AMP-activated protein kinase (AMPK) are known to be activated by the tumour suppressor kinase LKB1 (Lizcano *et al.*, 2004). Activation of one of these kinases (MARK1) by LKB1 has been shown to regulate microtubule dynamics by phosphorylating the microtubule-associated protein Tau, thereby reducing the affinity of Tau for microtubules and inhibiting tubulin polymerisation (Kojima *et al.*, 2007). QSK may play a similar role, since RNAi-mediated knockdown of the Drosophila orthologue of *QSK* resulted in spindle and chromosome alignment defects (Bettencourt-Dias *et al.*, 2004). One of the SB-induced lymphomas contained 2 insertion sites in *Qsk*, and this, coupled with the observations described above, suggests that *Qsk* may be a tumour suppressor gene.

Finally, a novel gene, *ENSMUSG00000075015*, contained a significant T2/Onc CIS and 1 MuLV insertion. 2 T2/Onc insertions were in the antisense orientation with respect to the gene, suggesting that the gene might encode a tumour suppressor, but functional analysis is required to determine the role of this gene in tumourigenesis.

The *Sleeping Beauty* dataset is relatively small and few candidate cancer genes have been identified. However, comparison with the MuLV screen demonstrates the potential benefits of using multiple mutagens to increase the spectrum of candidate cancer genes, but also to identify strong candidates that are independently mutated by both screens and are therefore unlikely to result solely from insertional bias. 3 of the 10 genes identified in the *Sleeping Beauty* screen are known or putative tumour suppressor genes, i.e. *Pten*, *Ppp3ca* and *Qsk*, suggesting that the T2/Onc mutagen is an effective tool for identifying recessive cancer genes. Scaling up the screen to identify further candidates would provide a valuable dataset to complement the retroviral insertional mutagenesis data.

3.4 Determining the mechanisms of MuLV insertional mutagenesis

3.4.1 Analysing the distribution of intragenic insertions

Analysis of the distribution of insertions within and around genes can help to determine the likely mechanisms of mutation. Oncogenic insertions in intergenic regions are likely to be promoter or enhancer mutations that result in increased levels of the wildtype protein. However, the effect of intragenic insertions, of which there are 8,447 (42.0% of the total), may be less obvious. The Ensembl API version 45_36f was used to identify the genomic coordinates of untranslated regions (UTRs), exons and introns in the longest transcript of each candidate cancer gene. From these, coding and non-coding exons, and introns within coding regions or UTRs, were distinguished. The "gene regions" were defined as 5' UTR, intron in 5' UTR, coding exon, intron flanked by coding exons, intron in 3' UTR, and 3' UTR. For each candidate, the number of insertions in each gene region was counted, and the orientation of each insertion with respect to the disrupted gene was determined. The total number of insertions in each gene region is shown in Figure 3.3A. The collective length of each gene region across all candidate genes was also calculated and, for each region, the insertion count was divided by the length in base pairs to give an indication of the proportion of insertions given the region size (Figure 3.3B). For insertions within introns or exons of the coding region, the identity, i.e. number, of the exon or intron containing the insertion was determined. This is helpful in determining the mechanism of mutation since, if a specific oncogenic gene product is formed, multiple insertions would be expected to localise to the same region of the gene.

Within genes, introns were the most common site of insertion, but for their size, they were the least commonly hit region. Intronic insertions may result in the formation of N- or C-terminal truncations. There are polyadenylation sites in both orientations of the retroviral provirus but promoters are only found in the forward orientation of the retroviral LTRs (see Section 1.4.2.1.1). Therefore, while antisense insertions can only form C-terminal truncations, sense insertions can form both N-terminal and C-terminal truncations. The distribution and orientation of intronic insertions will vary in different oncogenes depending on how an oncogenic mutant is created, and tumour suppressor genes can be inactivated by any distribution of insertions that results in non-functional N- or C-terminal truncations. There were therefore roughly equal numbers of sense and antisense insertions in introns. Genes with the highest numbers of intronic insertions were *Ikzf1* (56 insertions) and *Notch1* (52 insertions). The insertions in *Ikzf1*, a tumour

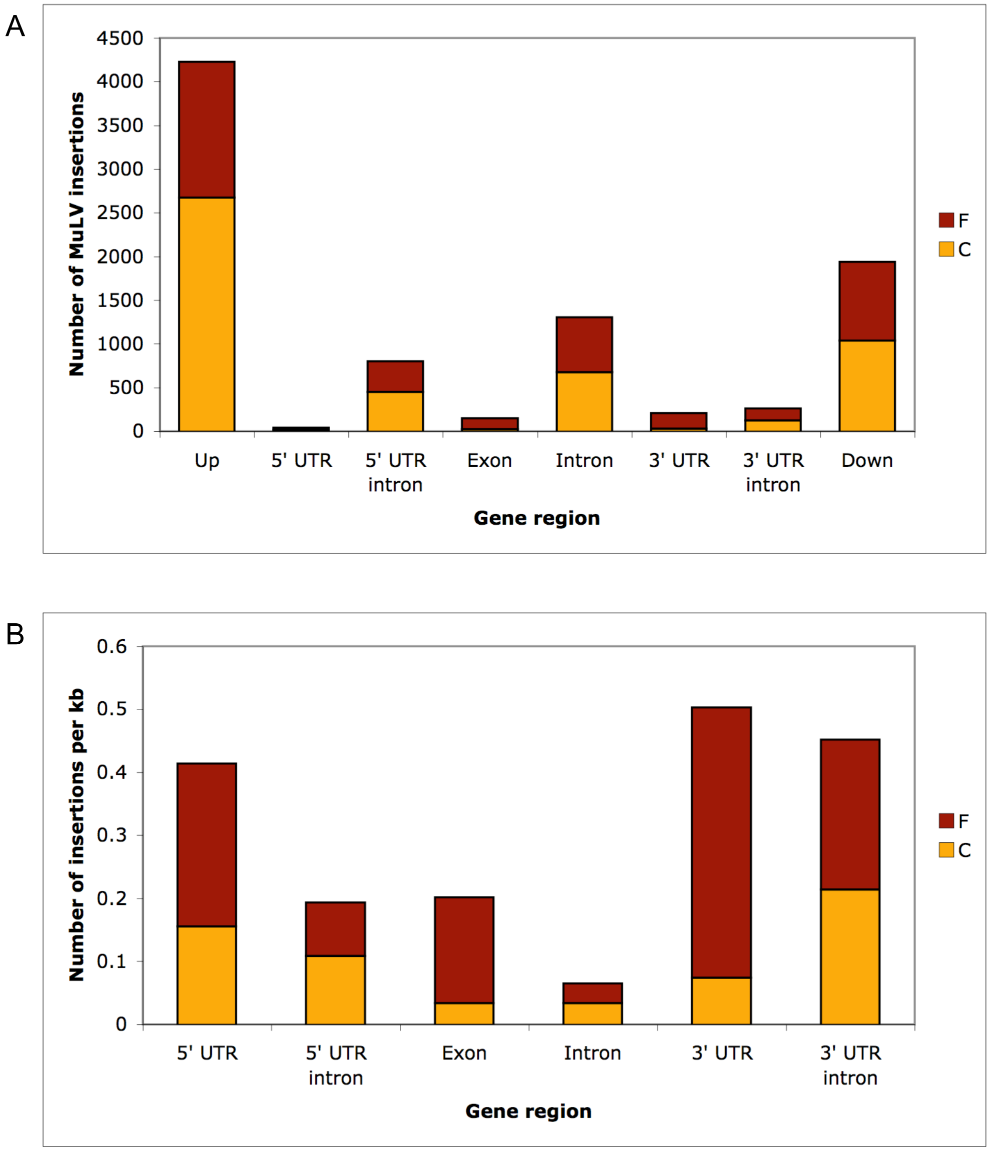

Figure 3.3. The distribution of MuLV insertions within candidate cancer genes. (A) The total number of insertions in each gene region. (B) The number of insertions as a proportion of the total length of the gene region across all candidate genes. The number of insertions in the sense orientation (F) with respect to genes is shown in red, while the number in the antisense orientation (C) is shown in yellow.

suppressor gene (see Section 3.4.4), are likely to cause premature termination of gene transcription, resulting in gene inactivation, while those in *Notch1* are likely to produce distinct, truncated mutant proteins that are implicated in tumourigenesis (see Section 3.4.2 for further details). *Flt3* contained 40 intronic insertions that were all in the sense orientation, suggesting that the gene is most likely disrupted by the formation of N-terminal truncations (Figure 3.4A). Most insertions were within intron 9 and are predicted to result in the production of proteins lacking the extracellular, ligand-binding, Ig-like domain. *FLT3* is mutated in around one third of human acute myeloid leukaemias, yet it is mutated by internal tandem duplications or by point mutations that produce a constitutively active protein (Small, 2006). 4,170 out of the 20,259 haematopoietic and lymphoid cancer samples tested in COSMIC have a mutation in *FLT3*, of which 185 have a missense mutation at amino acid 835 in the protein kinase core domain. Most of the remaining samples have internal tandem duplications that are represented in COSMIC by complex mutations and indels. This suggests that retroviral insertional mutagenesis may not always accurately recapitulate the mutations contributing to human cancers. Antisense insertions occurring in introns close to the 5' end of a gene could be inactivating mutations affecting tumour suppressor genes, or may result in the production of a truncated transcript from a cryptic transcription start site further downstream within the gene. It is also possible that they are acting as enhancer mutations.

The second most frequently hit regions were introns in the 5' UTR, i.e. introns that are flanked on each side by exons of the 5' UTR. Again, these collectively form a larger region than coding and non-coding exons. There were 28.8% more insertions in the antisense orientation than in the sense orientation. Sense insertions are most likely to be promoter mutations, which result in increased production of the full-length cellular protein. Antisense insertions may be prematurely terminating gene transcription, resulting in the complete absence of the gene product, as might be expected for tumour suppressor genes, or they may result in the production of a truncated transcript from a cryptic transcription start site. They could also be intragenic enhancer mutations or, as the longest gene transcript has been selected for this analysis, it is possible that some are enhancer mutations that are upstream of alternative gene transcripts, and are therefore producing full-length, wildtype proteins at increased levels. Cyclin D3 (*Ccnd3*) contained the highest number of insertions (204) within 5' UTR introns, with 85% occurring in the antisense orientation (Figure 3.4B). Since *Ccnd3* is an oncogene, and contains no known cryptic transcription start sites or alternative transcripts, it is likely that

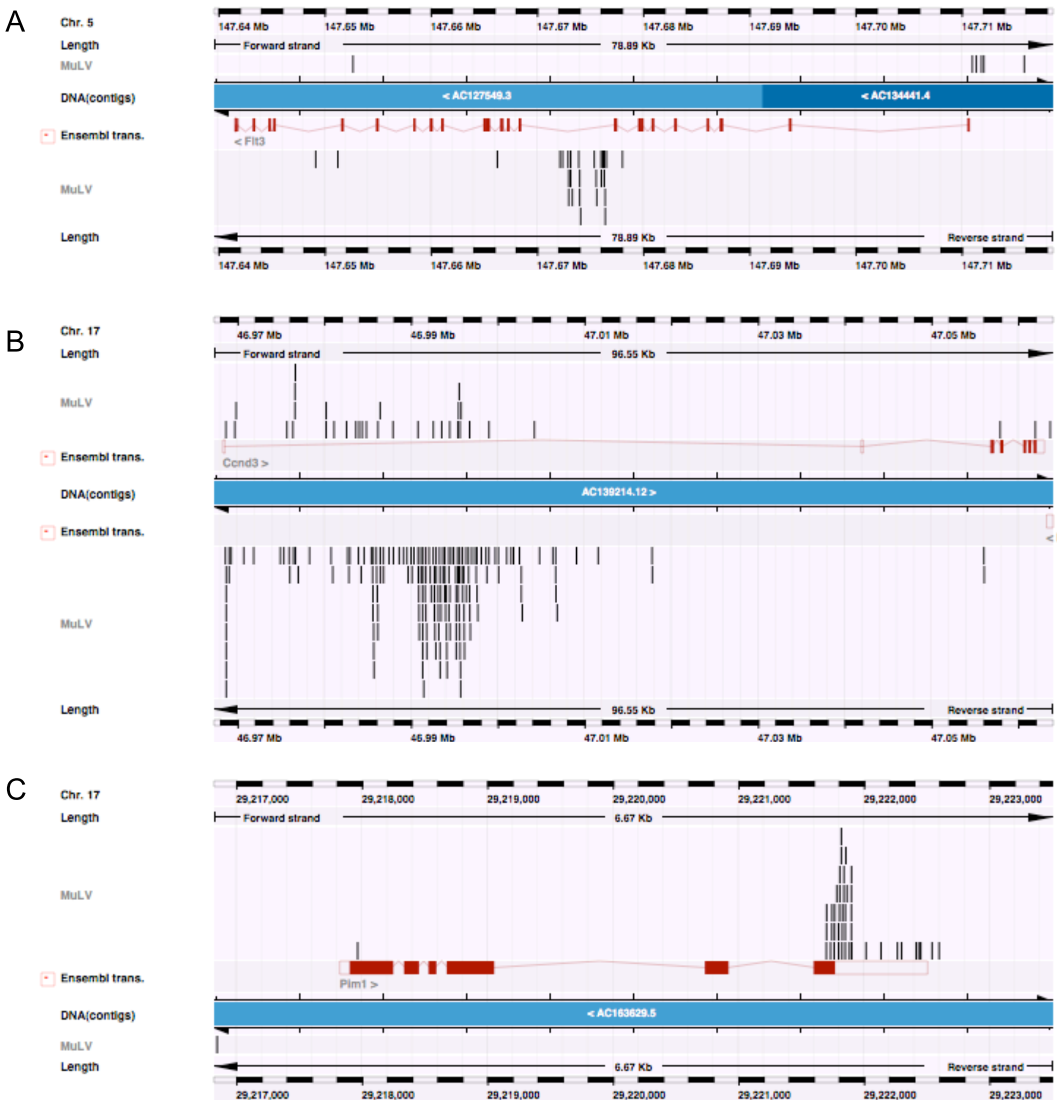

Figure 3.4. Intragenic MuLV insertions in candidate cancer genes. (A) Intronic insertions in *Flt3* are predicted to generate N-terminally truncated gene products. (B) Antisense insertions in the 5' UTR of *Ccnd3* are likely enhancer mutations. (C) Insertions in the final coding exon and 3' UTR of *Pim1* may cause premature termination of gene transcription that leads to a more stable gene product. Insertions are shown in black. Genes are shown in red. Insertions above and below the blue line are in the forward and reverse orientation, respectively.

the insertions are enhancer mutations. *Lck* contained 18 sense insertions but no antisense insertions in intronic regions of the 5' UTR, and a further 3 sense insertions in the 5' UTR and first coding exon, suggesting that all of the insertions are involved in the formation of chimeric transcripts in which the retroviral promoter drives increased expression of the cellular gene.

The 5' UTRs of candidate genes contained an over-representation of sense insertions, as expected for promoter insertions. *Myc* contained the most sense insertions, totalling eight. Antisense insertions in the 5' UTR may interfere with gene transcription, preventing protein production or resulting in transcription from a cryptic or alternative promoter.

Exons and 3' UTRs showed a strong bias towards insertions in the sense orientation. 80% of sense insertions in coding exons were in *Notch1*, *Mycn*, *Map3k8*, *Ccr7*, *Pim1* and *Jundm2*, and in all cases, insertions were at the 3' end of the gene, close to the 3' UTR. In the case of *Mycn* and *Pim1*, which also contained a large number of 3' UTR insertions, these insertions cause premature termination of gene transcription that result in the removal of mRNA-destabilising motifs and, therefore, the generation of a more stable gene transcript (Cuypers *et al.*, 1984; Selten *et al.*, 1985; van Lohuizen *et al.*, 1989). Insertions within *Pim1* are shown in Figure 3.4C. It is possible that *Ccr7* and *Jundm2* are disrupted by the same mechanism, since both contained sense insertions in the final coding exon and the 3' UTR. The near-exclusivity of sense insertions in these genes suggests that the polyadenylation site in the forward orientation of the retrovirus may have a stronger signal than the cryptic site in the reverse orientation. In summary, the density of insertions in exons and UTRs was higher than for introns, suggesting the importance of promoter insertions and "stabilising" insertions as mechanisms of mutagenesis.

For each candidate cancer gene, the distribution of insertions was used to predict the likely mechanisms of mutagenesis and, therefore, the likely structures of mutated gene products. Sense insertions that were upstream of the gene, within the 5' UTR or in an intron flanked by exons of the 5' UTR were classified as promoter mutations. Upstream insertions in the antisense orientation were classified as enhancer mutations. Insertions in the 3' UTR were classified as "stability" mutations, i.e. insertions that may result in the removal of mRNA-destabilising motifs, while sense and antisense insertions in exons or introns in the coding region were classified as C- or N-terminally and C-terminally truncating mutations, respectively. Finally, antisense insertions in introns within the 5' UTR remained unclassified, since these have a number of possible effects (see above). This yielded 360 genes with enhancer insertions, 309 with promoter insertions, 45 with stability mutations, 183 with C-terminally truncating insertions, 202 with C- or N-terminally truncating mutations, and 92 with antisense insertions in introns within the 5' UTR. Most genes are associated with multiple types of insertion (see Table 3.3). Genes were further classified according to the predicted protein generated by the mutations. Genes containing any combination of promoter, enhancer and stability mutations should generate the wildtype protein at increased levels compared with the endogenous gene. It was assumed that where both sense and antisense insertions occurred in the same exon and intron, the gene was C-terminally truncated, whereas if only sense insertions occurred, the gene was N-terminally truncated. Sense insertions in the last intron were classified as C-terminally truncating, as commonly observed, for example, in *Pim1* and *Mycn*. This generated 7 types of mutant – upregulated wildtype (201 genes), C-terminally truncated (30 genes), N-terminally truncated (2 genes), C- and N-terminally truncated (4 genes), and upregulated wildtype plus C-terminally truncated (122 genes), N-terminally truncated (56 genes) or C- and N-terminally truncated (24 genes). This suggests that a high proportion of genes contribute to tumourigenesis by increased production of the wildtype protein. C-terminally truncating mutations appear to be more common than N-terminally truncating mutations, but this may reflect the fact that insertions in the sense orientation were assumed to be C-terminally truncating if antisense insertions were also present, and therefore some may have been misclassified. The genes associated with each mutation type are presented in Table 3.3.

Elucidation of the mechanisms of mutagenesis is complicated by insertional bias and the ability of MuLV to disrupt a gene in multiple ways. Therefore, predictions must be experimentally validated, e.g. by measuring the length of transcripts generated by insertion-containing genes and by analysis of gene expression in MuLV-induced tumours (see Section 3.4.5). Reducing the number of ways in which an insertional mutagen can disrupt a gene would facilitate the analysis of insertions within genes. For example, by using a transposon engineered with a splice acceptor site and polyadenylation site on one strand only, it would be possible to distinguish C- and N-terminal truncations with a high degree of certainty.

Mutation type	Mechanisms of mutagenesis	Number of genes	Gene names
wildtype	P, E	100	*Cxcr4, Med13, mmu-mir-23a, Btg2, Art2b, Fgfr3, Fut8, Mknk2, Lyst, Fli1, Scyl1, Actr3, Tceb3, Kdr, Zfp438, 1700122O11Rik, Ier2, Acot11, Lta, Metrnl, Mpl, Cecr5, Pcgf5, Cd47, Haao, Rpl11, Mixl1, Hrbl, Q8BP09_MOUSE, Stmn1, C330024D12Rik, Ccnd2, Hspa9, Ttll10, Stat5a, Egfl7, Bex6, Irf2bp2, Cstad, Tpd52, Hibadh, Lfng, Lgals9, Psma1, ENSMUSG00000074256, Cd72, 1700019O17Rik, Anp32e, C130026L21Rik, Rcbtb2, Jph4, Ccrk, Brd2, AA536749, Frmd8, Mafk, Vpreb2, Hnrpf, A130050O07Rik, Lmo2, Ccdc19, AI848100, Aars2, Akt1, ENSMUSG00000073531, mmu-mir-181c, Rreb1, mmu-mir-17, Spn, 6430598A04Rik, Ppp2r5a, Ina, Slc39a13, mmu-mir-802, mmu-mir-142, Rcsd1, Lat, 2310016C08Rik, Trp53inp1, Chd2, Tcf7, Gadd45g, Rpl24, Cbl, Appl2, Ifnar1, Park7, Vdac1, ENSMUSG00000059894, ENSMUSG00000071320, ENSMUSG00000059313, ENSMUSG00000071576, ENSMUSG00000063435, ENSMUSG00000069082, ENSMUSG00000034596, ENSMUSG00000067988,*
wildtype	P, E, CT	41	*OTTMUSG00000012358, Rara, Bcl9l, Klf3, Mobkl2a, Tmem90a, Evi1, A530013C23Rik, Mgat1, Ldha, 4831426I19Rik, Tcof1, Slc38a1, 2310066E14Rik, Arhgdib, Plekha2, Gimap4, A630001O12Rik, Zfp217, Il6st, Rasgrp2, B3gnt2, A930002I21Rik, Fgr, Pag1, Sema4d, Ptp4a2, Rhoh, Zeb2, Mef2d, Ggta1, Stat5b, Pik3r5, Arid1a, Csk, Thra, Lcp1, Parvg, Hmga1, Gpr132,*
wildtype	E	27	*Ddr1, Zdhhc19, Chd7, Bid, Zfp36l2, Edg1, Dad1, Gfi1b, Mcl1, Fgd2, ENSMUSG00000074788, Rps14, Chst3, Trim47, Fgfr2, Aqp4, ENSMUSG00000074675, Frat2, Klhl25, Chchd7, Hhex, ENSMUSG00000072756, ENSMUSG00000061115, ENSMUSG00000046809, ENSMUSG00000072757, ENSMUSG00000052894*
wildtype	P, CT	10	*4932417H02Rik, Phyhd1, Cd53, Rhbdf2, Gpr56, Ptp4a3, Pitpnm2, Olfr56, 6330548G22Rik, Evi2b*
wildtype	P, E, S	9	*Hoxa7, 4632428N05Rik, BC027057, Sox19, OTTMUSG00000016805, Gadd45b, Ccnd1, 3930402G23Rik, Pim1*
wildtype	E, S	6	*Evi5, Gpr152, Clec2d, Scd1, BC030863, Scube1*
wildtype	P	2	*Ppp1r10, Pscd4*
wildtype	P, E, S, CT	2	*Pvt1, Rassf2*
wildtype	P, S	2	*Orai2, Mycn*
wildtype	E, CT	1	*Lrrc8c*
wildtype	P, S, CT	1	*Zbtb7b*
wildtype & C-trunc	P, E, C, C/N	30	*Nsmce1, Gse1, Ahi1, Prr6, Ski, Ptprc, Itpr2, Map3k1, Abcb9, Foxp1, Exoc2, Cd97, Srgn, Tmem131, Vamp8, AB041803, Etv6, Prkch, Ssbp3, Supt3h, Akna, Gm525, Rasgrp1, Zfp608, Tcfap4, Arid3a, Ntn1, 2010107G12Rik, Sema4b, Ets1*
wildtype & C-trunc	E, C, C/N	26	*Akap13, Cd2, Dym, Vps13d, Tmprss3, Pxn, Pygm, Tgfbr3, Zc3h12a, Bcl11b, Mad1l1, Spsb4, Gimap6, H2-D1, Nup214, Lims1, Znrf1, Kcnn4, Tmem49, Kctd2, Sept9, Slamf6, Dopey2, Itpkb,*
wildtype & C-trunc	P, E, C	16	*Exoc6, Colq, Hdac7a, Cybasc3, Lef1, Mylc2pl, Tmem173, ENSMUSG00000074787, Nfe2, B3gntl1, Treh, Slamf7, Aqp9, Cd27, 5730559C18Rik, EG433384*
wildtype & C-trunc	P, C, C/N	6	*Cldn10a, Spata13, Ahnak, Vil2, Stard10, Ubac2*
wildtype & C-trunc	E, S, C, C/N	6	*Cugbp2, Arhgap26, Arhgef3, 1110036O03Rik, Rab37, Psmb8*
wildtype & C-trunc	P, C, CT	5	*Asb2, Elf1, Adrbk1, Arpp21, Ptpre*
wildtype & C-trunc	P, E, C, CT	5	*Ncoa3, Emp3, Usp52, Il2ra, Wasf2*
wildtype & C-trunc	E, C	5	*A130092J06Rik, Cdkl3, Stra8, Gna15, Kit*
wildtype & C-trunc	P, E, C, C/N, CT	4	*Zmiz1, 4932422M17Rik, Ubxd5, Ikzf1*
wildtype & C-trunc	P, E, S, C, C/N	3	*Padi2, Ksr1, Pigv*
wildtype & C-trunc	P, C, C/N, CT	3	*1600014C10Rik, Anxa2, D18Ertd653e*
wildtype & C-trunc	P, C	2	*BC008155, Epha6*
wildtype & C-trunc	C, C/N, CT	2	*Bcl2l1, Ppp1r16b*
wildtype & C-trunc	E, C, C/N, CT	2	*Arhgef10l, Il21r*
wildtype & C-trunc	P, S, C, C/N	1	*Nfkb1*
wildtype & C-trunc	P, E, S, C, C/N, CT	1	*Jundm2*
wildtype & C-trunc	S, C, N	1	*Rnf43*
wildtype & C-trunc	S, C	1	*Ovca2*
wildtype & C-trunc	P, E, S, C	1	*Trpm1*
wildtype & C-trunc	E, S, C	1	*C330016O10Rik*
wildtype & N-trunc	P, E, C/N	31	*Runx1, Set, Cd3e, Cd48, 2410014A08Rik, Pctk2, Fchsd2, Paics, Thy1, Msh5, OTTMUSG00000005737, Jup, Sdk1, Prkcbp1, Mbd2, Arrdc5, Coro2a, Hipk1, Erg, D12Ertd553e, Rras2, Nedd4l, Myb, Eng, Plac8, Tspan2, 1190002H23Rik, Tap2, Ptbp1, Mns1, Ube1l, Chd1*
wildtype & N-trunc	E, C/N	8	*Slc36a3, Nfkbil1, Rtn4rl1, Bcl11a, Ak1, Ndrg3, NP_001074704.1, Map3k8*
wildtype & N-trunc	P, E, C/N, CT	6	*Tcte3, Sla2, Sla, Grap2, Pik3cd, Tbxa2r*
wildtype & N-trunc	P, C/N, CT	3	*Zfp710, Tspan14, Dnahc8*
wildtype & N-trunc	P, E, S, C/N	2	*Irf4, Midn*
wildtype & N-trunc	E, S, C/N	2	*Stard3nl, 1700081D17Rik*
wildtype & N-trunc	P, E, S, C/N, CT	1	*E230001N04Rik*
wildtype & N-trunc	P, E, N, CT	1	*Ccnd3*
wildtype & N-trunc	P, E, S, N, CT	1	*Myc*
C-trunc	C, C/N	29	*Nfix, Slc43a2, 2010106G01Rik, Nup210, Smg6, Hvcn1, Stk4, Lrrfip1, Dpp4, Rorc, Ramp1, Myo18a, Clec16a, Jazf1, Sh3bp5, Plxnd1, Prdm16, 3110082I17Rik, Serinc3, Fyb, B230120H23Rik, St6galnac5, Scotin, Il16, E2f2, Usp7, Recql5, Sirt2, Abcg1*
C-trunc	C	1	*Fmnl1*
N-trunc	C/N	2	*1300007F04Rik, Exosc5*
C-trunc & N-trunc	C, N	4	*Xrcc6, Rnf166, Sh3d19, Iqch*
wildtype & C-trunc & N-trunc	P, E, C, N	10	*Lck, Slc1a3, Dhx40, A2AN91_MOUSE, Rnf157, Katnal1, Mgat4a, Dock8, Cyb5, Cbfa2t3*
wildtype & C-trunc & N-trunc	E, C, N	6	*Wwox, Nid1, Capsl, Tcf25, Flt3, Tbc1d1*
wildtype & C-trunc & N-trunc	P, C, N	2	*Ubash3a, Pml*
wildtype & C-trunc & N-trunc	P, E, C, N, CT	2	*Runx3, Pecam1*
wildtype & C-trunc & N-trunc	P, C, N, CT	1	*Mrvi1*
wildtype & C-trunc & N-trunc	P, E, S, C, N	1	*Ccr7, Notch1, Gfi1, Ikzf3*

Table 3.3. The predicted mutation types and mechanisms of mutagenesis based on the distribution of MuLV insertions within and around candidate cancer genes. C-trunc = C-terminally truncated, N-trunc = N-terminally truncated, P = promoter insertion, E = enhancer insertion, C = antisense intragenic insertion, C/N = sense intragenic insertion, S = stabilising insertion, CT = antisense insertion 5' of first coding exon (i.e. truncating, leading to inactivation or use of cryptic transcription start site, or enhancer mutation).

3.4.2 Analysing co-occurring insertions in candidate genes disrupted by MuLV and T2/Onc

The co-occurrence of MuLV and T2/Onc insertions in distinct regions of genes provides strong evidence that the insertions do not result from insertional bias and that they play an important role in oncogenesis. Such insertions can provide important clues about the mechanism of mutation and, therefore, about the structure and function of genes and oncoproteins involved in cancer. In this section, the distributions of MuLV and T2/Onc insertions are compared within known and implicated cancer genes that overlap between the MuLV kernel convolution (KC)-based list of candidates and *Sleeping Beauty* (SB) candidates from the KC list, i.e. *Notch1*, *Myb*, *Fli1*, *Erg* and *Ikzf1*, and from the Monte Carlo (*Efr*=0.005) list, i.e. *Rasgrp1* and *Etv6*.

In *Notch1*, MuLV and T2/Onc insertions co-occurred in the same orientation in 3 distinct regions of the gene (Figure 3.5A). Antisense MuLV and T2/Onc insertions were identified in the second intron. The retroviral insertions could be assumed to be enhancer mutations, yet T2/Onc has low enhancer activity. Therefore, these are more likely to be truncating mutations, and this is consistent with the observation that radiation-induced deletions in the 5' region of *Notch1* result in truncated proteins that lead to the development of mouse thymic lymphomas (Tsuji *et al.*, 2003). The authors showed that deletion of, or MuLV insertion into, the juxtamembrane extracellular region encoded by exons 1 and 2 results in transcription from cryptic transcription start sites further downstream in *Notch1* and leads to the production of an active protein lacking most of the extracellular domain. Co-occurring sense insertions were also identified in the 28th and 29th introns. Based on their orientation, these insertions are expected to produce N-terminally truncated proteins containing only the intracellular domain of Notch1. This form of Notch1, called Notch1IC, is constitutively active and is associated with leukaemogenesis (for review, see Aster *et al.*, 2008). Finally, there were co-occurring insertions, again mostly in the sense orientation, within the final coding exon. These insertions were upstream of the PEST domain, which regulates turnover of Notch1IC (Aster *et al.*, 2008). Deletion of the PEST domain, by MuLV insertion in T-cell lymphomas and by radiation in the study by Tsuji *et al.* (2003), is believed to contribute to tumourigenesis in collaboration with other activated oncogenes (Feldman *et al.*, 2000; Hoemann *et al.*, 2000; Tsuji *et al.*, 2003). 184 human tumour samples out of 1,909 tested contain *NOTCH1* mutations in the COSMIC database. Of these, 180 are in

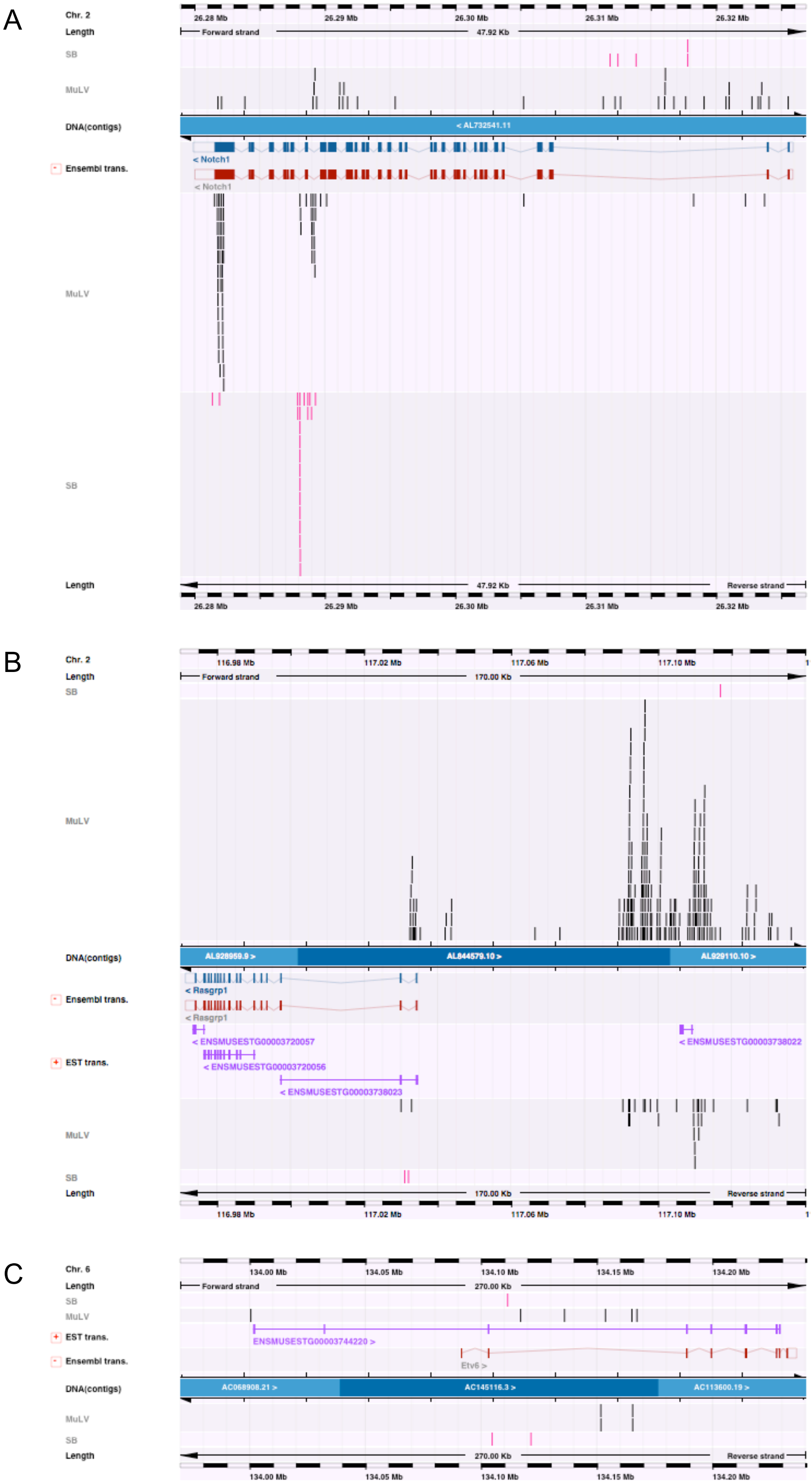

Figure 3.5. Co-occurring MuLV and T2/Onc insertions help to identify the mechanism of mutagenesis of genes *Notch1* (A), *Rasgrp1* (B) and *Etv6* (C). MuLV insertions are shown in black, T2/Onc insertions are shown in pink. Ensembl gene transcripts are shown in red and blue. ESTs are shown in purple. Insertions above and below the blue bar labelled DNA(contigs) are in the forward and reverse orientation, respectively.

haematopoietic and lymphoid tissue, which corresponds to 24% of all samples of this cancer type tested. *NOTCH1* is therefore specifically, and significantly, associated with cancers of this type.

Co-occurring MuLV and T2/Onc insertions were also identified in the first intron, preceding the first coding exon, of *Rasgrp1* (Figure 3.5B). These insertions, which are in the sense orientation with respect to the gene, are likely to be promoter mutations that result in overexpression of the full-length transcript. MuLV enhancer mutations were also found upstream in the antisense orientation but, unsurprisingly, these were not identified in the SB screen since T2/Onc has low enhancer activity. There were no intragenic insertions beyond the first intron, suggesting that only the full-length gene contributes to oncogenesis. This is supported by the observation that, among 273 tumour samples tested, there are none with somatic mutations in *RASGRP1* in the COSMIC database. Deregulated expression of full-length murine *Rasgrp1* contributes to the development of T lymphocytic leukaemias (Klinger *et al.*, 2005), and to the progression of skin carcinogenesis through the activation of the *Ras* oncogene (Luke *et al.*, 2007). Interestingly, in previous screens, insertions that are ~60-100 kb upstream have been assigned to *Rasgrp1* (Hansen *et al.*, 2000; Hwang *et al.*, 2002; Kim *et al.*, 2003a; Mikkers *et al.*, 2002; Stewart *et al.*, 2007; Suzuki *et al.*, 2006; Suzuki *et al.*, 2002). However, analysis of the insertions in the context of Ensembl shows that they are flanking an Ensembl EST gene for which there is no associated Ensembl gene transcript (Figure 3.5B). Expression analysis of *Rasgrp1* in the affected tumours is required to determine whether it is indeed disrupted by these insertions, but this observation suggests that the analysis of insertion sites in the context of the mouse genome could potentially help in the identification of "new" mouse transcripts.

All of the MuLV and T2/Onc insertions identified in gene *Etv6* were in the second intron, in both orientations (Figure 3.5C). Since sense insertions can form both N-terminal and C-terminal truncations but antisense insertions can form only C-terminal truncations, it is likely that where insertions occur in both orientations in the same intron of an oncogene, they are causing premature termination of gene transcription that results in C-terminally truncated gene products. In the case of *Etv6*, this would result in the production of polypeptide lacking both of its functional domains. Etv6 is a transcriptional repressor that is essential for haematopoietic stem cell function. The N-terminal sterile alpha motif/pointed (SAM_PNT) domain (IPR0003118) is responsible for hetero- and homodimerisation with other ETV6 and Ets (erythroblast transformation specific)-family proteins. Deletion of this domain decreases the inhibition of macrophage colony stimulating factor receptor (*MCSFR*) promoter activation by CBFA2B and C/EBPa, but does not completely abrogate it (Fears *et al.*, 1997). However, the SAM_PNT domain is necessary for interaction with, and inhibition of, the *FLI1* oncogene (Kwiatkowski *et al.*, 1998). The winged helix DNA-binding Ets domain (IPR000418) is essential for inhibiting the activation of *MCSFR* (Fears *et al.*, 1997). Therefore, deletion of both these domains in the mouse most likely produces a non-functional protein, resulting in the overexpression of Etv6 target genes, such as *Fli1*. As mentioned previously, *ETV6* forms a fusion with *ABL*, but also with many other genes, in human leukaemias (for review, see Bohlander, 2005). Interestingly, while most fusions with tyrosine kinase genes contain a breakpoint in intron 4 or 5 of *ETV6*, and, for example, fusions with *RUNX1* contain a breakpoint in intron 5, there are also fusions with unique or rare recurrent gene partners in which *ETV6* has a breakpoint in intron 2. It has been suggested that promoters in the latter *ETV6* truncation upregulate nearby oncogenes (Jalali *et al.*, 2008; Panagopoulos *et al.*, 2006). However, the distribution of MuLV and T2/Onc insertions within *ETV6* suggests that the nonfunctional truncation may also itself contribute to leukaemogenesis. The ETV6-RUNX1 fusion is consistently associated with deletion of the normal *ETV6* allele, suggesting that normal ETV6 represses ETV6-RUNX1 by interaction via the SAM_PNT domain (Hart and Foroni, 2002; Raynaud *et al.*, 1996). Homo- and heterodimerisation are believed to repress the activity of Ets proteins (Carrere *et al.*, 1998; see discussion below in relation to the *Erg* gene) and therefore, it is possible that by deleting one allele, fewer heterodimers will be formed with other Ets proteins, resulting in increased activity of those proteins. Incidentally, 1 sense and 1 antisense MuLV insertion were identified 91.05 kb and 147.04 kb, respectively, upstream of the *Etv6* gene. It could be assumed that these insertions are not oncogenic, since they are a considerable distance from the gene. However, the sense insertion is just 1.22 kb upstream of an Ensembl EST gene for which there is no associated Ensembl gene transcript, which suggests that there may be an unannotated alternative transcript of *Etv6* (see Figure 3.5C).

The rest of the genes that were disrupted by both MuLV and SB showed variation in the distribution of insertions. In some cases, this reflects differences in the mutational mechanisms of the two mutagens. For example, 1 retroviral insertion and 1 transposon insertion were found to co-occur just upstream of *Myb* (Figure 3.6A) in the sense orientation, where they are likely to be causing promoter mutation, but the vast majority of retroviral insertions were putative enhancer mutations, occurring upstream in the antisense orientation, or downstream, predominantly in the sense orientation with respect to *Myb*. The presence of MuLV sense and antisense insertions, and 1 SB sense insertion, just upstream of an Ensembl EST gene suggests that, as for *Etv6*, there is an additional *Myb* transcript that has not been annotated as an Ensembl gene transcript. The remaining T2/Onc insertions were intragenic. 3 were in the last (13th) intron in the antisense orientation with respect to *Myb*, while 1 was in the 10th intron in the sense orientation. A C-terminal truncation caused by the latter would truncate the C-terminal Myb domain (IPR015395), which is known to bind the inhibitor Cyp-40 (Leverson and Ness, 1998). Oncogenic *v-Myb* contains a mutated binding site that prevents binding of Cyp-40 and so prevents negative regulation (Leverson and Ness, 1998). The contribution of insertions within the last intron is unclear since a C-terminally truncated protein would contain an intact binding domain. It is possible that the last exon of *Myb* encodes a protein sequence with a hitherto uncharacterised role in oncogenesis.

Variation in the patterns of MuLV and T2/Onc insertions in *Fli1* also reflect differences in mutational mechanism. All of the MuLV insertions were upstream in the sense and antisense orientation, acting as promoter and enhancer mutations, respectively. Again, some of the upstream MuLV insertions were a considerable distance from *Fli1* but 3 sense insertions were within the first exon of an Ensembl EST gene for which there is no associated Ensembl gene transcript, suggesting the presence of an additional, unannotated *Fli1* gene transcript (Figure 3.6B). While 2 of the SB insertions were also upstream in the sense orientation, the remaining 3 were in the first intron in the sense orientation, most likely producing an overexpressed, N-terminally truncated transcript in which none of the functional domains are deleted.

Similarly, both mutagens were found upstream of *Erg* (Figure 3.6C) in the sense orientation, and MuLV insertions also occurred upstream in the antisense orientation. However, 19 intragenic T2/Onc insertions were found in the sense orientation within a 1,531 bp region in the 1st intron, while 3 MuLV sense insertions were found in the 2nd intron. Like *Etv6*, *Erg* encodes a SAM_PNT and an Ets domain and, assuming that the insertions are producing N-terminally truncated transcripts, the T2/Onc insertions may give rise to truncated proteins containing both domains, while the MuLV insertions would give rise to proteins with a disrupted SAM_PNT domain but full-length Ets domain. The

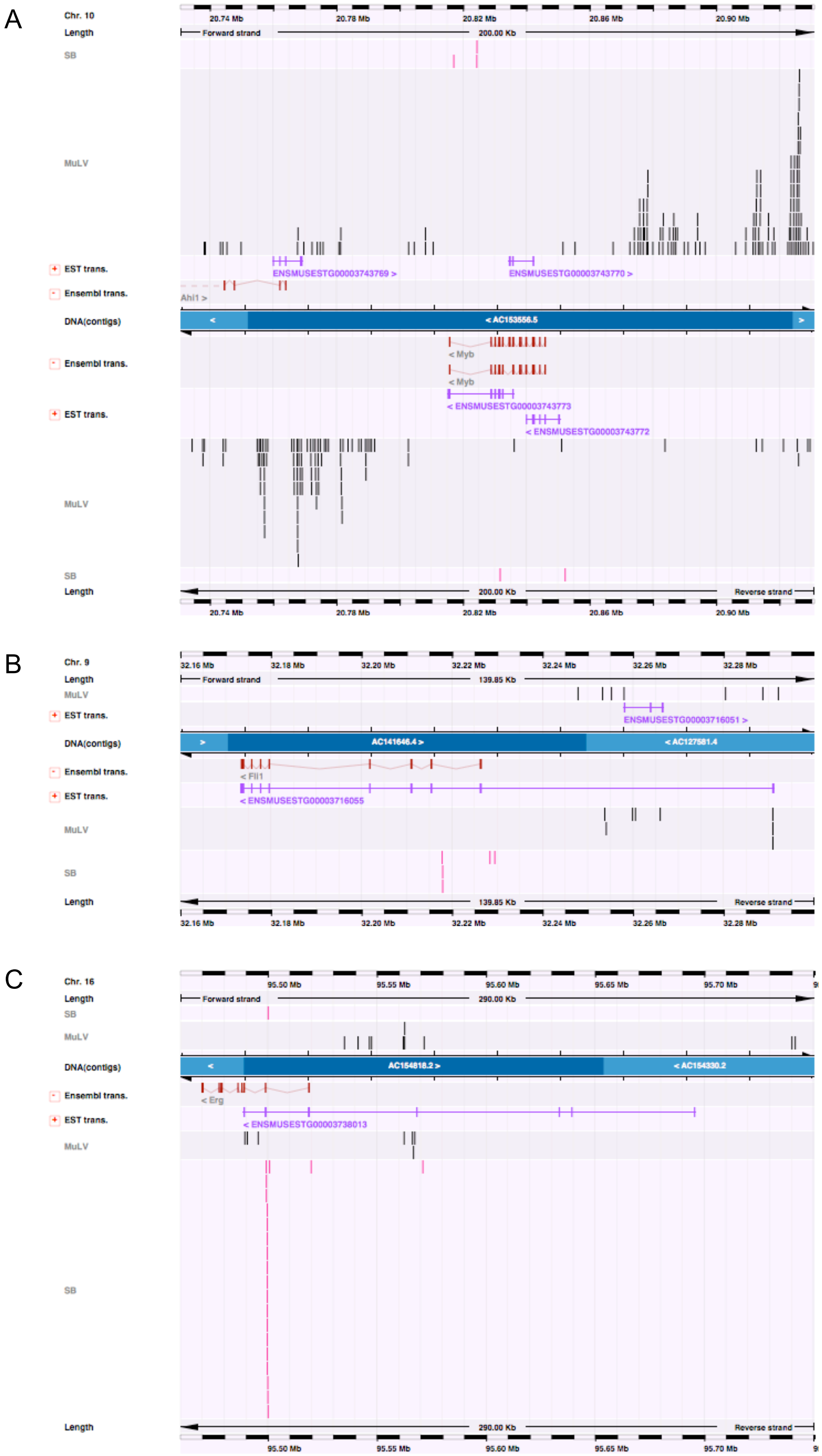

Figure 3.6. Variation in the distribution of MuLV and T2/Onc insertions in *Myb* (A), *Fli1* (B) and *Erg* (C) may reflect differences in the mechanisms of mutagenesis. MuLV insertions are shown in black, T2/Onc insertions are shown in pink. Ensembl gene transcripts are shown in red and blue. ESTs are shown in purple. Insertions above and below the blue bar labelled DNA(contigs) are in the forward and reverse orientation, respectively.

apparent presence of functional domains in the Erg truncations, but not in the Etv6 truncations, may reflect the fact that Erg is a transcriptional activator (Duterque-Coquillaud *et al.*, 1993), whereas Etv6 is a transcriptional repressor (see above). The closely aligned T2/Onc insertions in the first intron of *Erg* could be contaminants, but it is also possible that, in the absence of enhancer activity, the production of an overexpressed, N-terminally truncated protein is the most effective way to mutate the gene. The SAM_PNT domain is involved in the formation of heterodimers with other Ets proteins. Erg/Ets-2 dimer formation prevents Ets-2 from acting as a transcriptional activator of *Mmp3* (Basuyaux *et al.*, 1997; Buttice *et al.*, 1996) and dimerisation may prevent Ets proteins from binding to genomic DNA target sites (Carrere *et al.*, 1998). Therefore, it is possible that the 3 MuLV insertions in the 2nd intron that appear to disrupt the SAM_PNT domain prevent dimerisation and so cause an increase in the transcriptional activity of Erg and other Ets proteins that bind to Erg. The high proportion of MuLV promoter and enhancer insertions suggests that this may be a less efficient way of upregulating the gene, although it could also reflect the tendency of MuLV to insert close to transcription start sites. All of the MuLV insertions, and one of the T2/Onc insertions, that were identified upstream in the sense orientation were greater than 40 kb upstream of the *Erg* gene, which is a considerable distance for promoter mutation. However, the insertions resided within an Ensembl EST gene that overlaps with the *Erg* gene, suggesting that there may be an additional, unannotated, *Erg* transcript that is targeted by insertional mutagenesis.

In summary, differences in the distribution of MuLV and T2/Onc insertions may help to distinguish oncogenes and tumour suppressor genes. Intragenic insertions in oncogenes are more likely to be localised, since specific mutations, such as those described in *Notch1*, may be required for oncogenesis. However, it is more likely that tumour suppressor genes can be inactivated in multiple ways, and the distribution of insertions may be less defined, as demonstrated in *Ikzf1*, where MuLV and T2/Onc insertions were scattered throughout the gene (see also Section 3.4.3, below).

3.4.3 Identification of tumour suppressor genes inactivated by MuLV

Although retroviral insertional mutagenesis identifies predominantly oncogenes, tumour suppressor genes also featured in the list of candidate cancer genes. The most prevalent, with 93 insertions, was *Ikaros* (*Ikzf1*). *Ikaros* encodes a haematopoietic-specific zinc finger DNA-binding domain protein that regulates B- and T-cell differentiation (Georgopoulos *et al.*, 1997). Mice with reduced *Ikaros* expression develop leukaemias and lymphomas with complete penetrance (Winandy *et al.*, 1995). 31 insertions were also identified in *Aiolos* (*Ikzf3*), which is also a member of the Ikaros family. Ikaros and Aiolos appear to play dual roles in T cell development since they can regulate the activation or repression of lineage-specific genes through the formation of chromatin remodelling complexes in lymphocytes (Georgopoulos, 2002; Kim *et al.*, 1999). However, the importance of *Ikaros* and *Aiolos* as tumour suppressor genes is demonstrated by their frequent deletion in paediatric acute lymphoblastic leukaemia (ALL) (Mullighan *et al.*, 2007), and *Ikaros* is deleted in 83.7% of ALLs containing the *BCR-ABL* translocation (Mullighan *et al.*, 2008).

Other implicated tumour suppressor genes in the candidate list included *Wwox* (22 insertions), *E2f2* (17 insertions), *Mobkl2a* (*Mob1*; 16 insertions), *Xrcc6* (*Ku70*; 10 insertions), *Ovca2* (9 insertions) and *Adrbk1* (*Grk2*; 8 insertions). *Wwox* spans the human FRA16D fragile site and is frequently disrupted in human cancers (Bednarek *et al.*, 2000). *Wwox*$^{+/-}$ mice develop significantly more ethyl nitrosurea (ENU)-induced lung tumours and lymphomas than wildtype mice, suggesting that *Wwox* can act as a haploinsufficient tumour suppressor gene (Aqeilan *et al.*, 2007). Loss of *E2f2* accelerates *Myc*-induced lymphomagenesis in mice (Opavsky *et al.*, 2007), while *Xrcc6*-deficient mice develop thymic and disseminated T cell lymphomas (Li *et al.*, 1998). *MOB1* activates the tumour suppressor *LATS1*, which is inactivated in human sarcomas and ovarian and breast cancers (Hergovich *et al.*, 2006), and inactivating insertions in *Mob1* may therefore contribute to tumourigenesis by preventing the activation of *Lats1*. *OVCA2* is one of two adjacent genes that are frequently deleted in human ovarian, brain, breast and lung tumours (Schultz *et al.*, 1996). *GRK2* acts in a negative feedback loop to control TGFβ signal transduction, which is often dysregulated in cancer (Ho *et al.*, 2005), and was shown to significantly reduce proliferation of thyroid cancer cell lines (Metaye *et al.*, 2008).

There is no straightforward approach for identifying candidate tumour suppressor genes because, while they are likely to contain only intragenic insertions, some oncogenes, such as *Notch1* and *Pim1*, are also mutated predominantly by intragenic insertions. However, as mentioned in Section 3.4.2, insertions in oncogenes are more likely to form specific oncogenic mutants and may therefore be more localised within the gene. In addition, tumour suppressor genes are likely to contain multiple insertion sites within the same tumour, as described for *Pten* and *Qsk* in Section 3.3, because both copies of the gene must be inactivated. This does not hold for haploinsufficient tumour suppressor genes, which require only one inactivating insertion for tumourigenesis and are therefore more likely to be identified by insertional mutagenesis than genes requiring an insertion in both genes. 14 tumours contained multiple insertions within *Ikzf1*, while 1 contained multiple insertions in *Ikzf3*. However, none of the other genes so far discussed in this section were mutated by multiple insertions. This suggests either that they are haploinsufficient tumour suppressor genes, as demonstrated for *Wwox*, or that the coverage of the screen was too low, such that multiple insertions occurred but were not identified. To further complicate matters, oncogenes may also contain multiple insertion sites within the same tumour, either because the gene is a preferential target site for the virus, or because upregulation of both gene copies provides an even greater growth advantage to the cell. However, taken together, the distribution of insertions and the number of insertion sites within each tumour can help to identify potential tumour suppressor candidates.

Smg6 contained 53 insertions and was mutated by multiple insertions in 4 tumours. All insertions were intragenic and were distributed throughout the gene in both orientations, although many were clustered within a single intron (Figure 3.7A). The human orthologue, *EST1A/SMG6*, has been shown to interact with telomerase and the human telomerase reverse transcriptase (hTERT) (Redon *et al.*, 2007), and overexpression in kidney 293T cells leads to progressive telomere shortening (Snow *et al.*, 2003). Early in tumourigenesis, telomere shortening contributes to chromosomal destabilisation and therefore promotes genomic instability and cancer progression (see Sections 1.2.3.3 and 1.3.3.1). A telomere maintenance mechanism is subsequently activated and is required for tumour progression and immortality (Stewart, 2005). Therefore, *SMG6* could play an oncogenic or tumour suppressive role in this process. SMG6 is also an essential factor in the nonsense-mediated mRNA decay (NMD) pathway, which degrades mRNAs carrying premature stop codons and regulates the expression of naturally occurring transcripts, including those involved in cell cycle progression (Rehwinkel *et al.*, 2005). *SMG6* may therefore play a tumour suppressive role by negatively regulating oncogene expression via NMD. The presence of both sense and antisense insertions in the 9th intron suggests that they are involved in C-terminal truncation of the gene product. This would result in the removal of the PINc nucleotide binding domain (IPR006596), which is required for degradation of single-stranded RNA, and an inactivated domain has been shown to inhibit

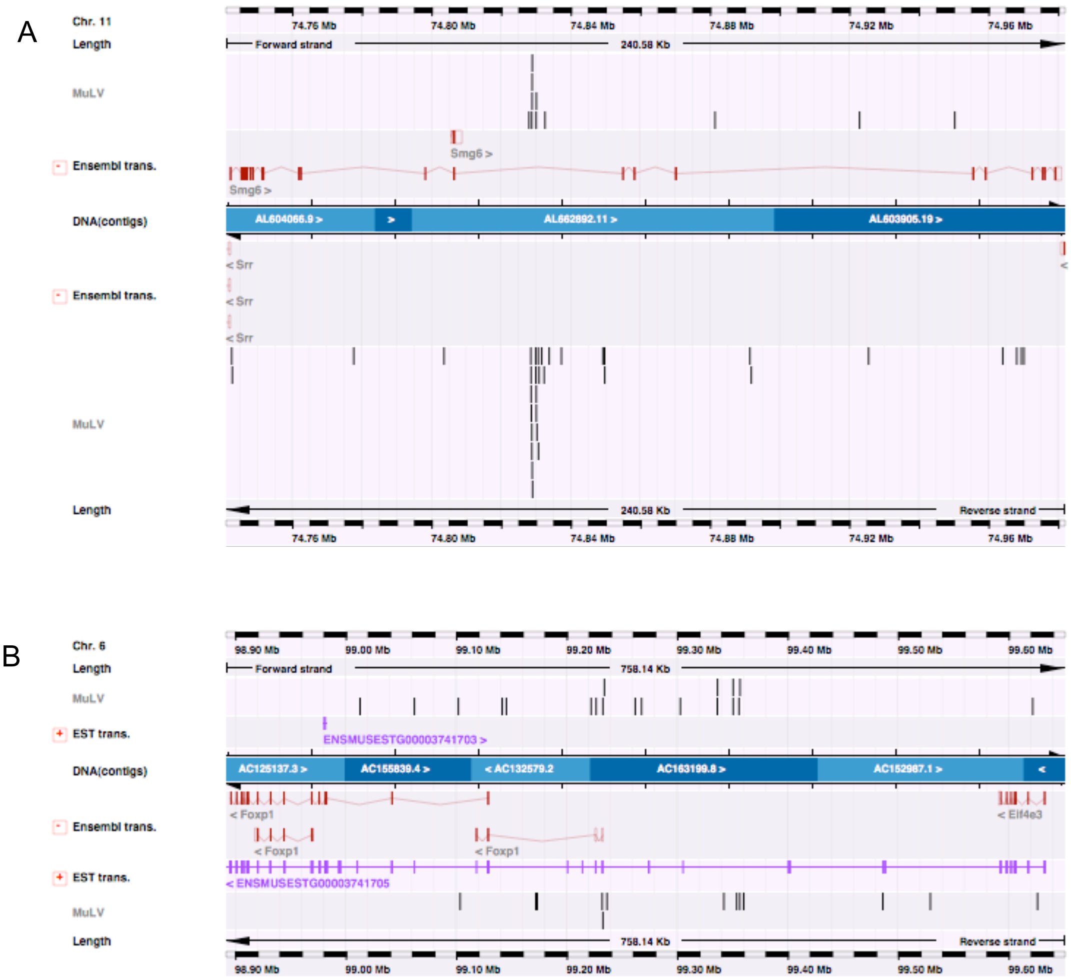

Figure 3.7. *Smg6* (A) and *Foxp1* (B) are putative tumour suppressor genes identified by MuLV insertional mutagenesis. MuLV insertions are shown in black. Ensembl gene transcripts are shown in red. ESTs are shown in purple. Insertions above and below the blue bar labelled DNA(contigs) are in the forward and reverse orientation, respectively.

NMD in *Drosophila* (Glavan *et al.*, 2006). This suggests that abrogation of NMD activity is the mechanism by which *Smg6* contributes to MuLV-induced lymphomagenesis. *SMG6* also resides within a deleted region containing 383 genes identified in 2.6% of human B-cell ALLs and 4.0% of T-cell ALLs in Mullighan *et al.* (2007).

Rassf2 contained 12 insertions, 2 of which were identified in a single tumour. Rassf2 is a negative regulator of Ras that is silenced by CpG island hypermethylation in a range of cancers, including gastric (Endoh *et al.*, 2005), liver (Nishida *et al.*, 2008), breast and lung (Cooper *et al.*, 2008; Kaira *et al.*, 2007). It has been shown to prevent cell transformation in primary colorectal cancers (Akino *et al.*, 2005).

Foxp1 contained 29 insertions, including 2 insertions in one tumour (Figure 3.7B). Overexpression of *Foxp1* is associated with poor prognosis in lymphomas (Banham *et al.*, 2005), but loss of *Foxp1* expression in breast cancer is also associated with poor prognosis (Fox *et al.*, 2004) and *Foxp1* maps to a region on chromosome 3 (p14.1) that frequently shows loss of heterozygosity in a range of human cancers (Banham *et al.*, 2001). This suggests that *Foxp1* can act as an oncogene or a tumour suppressor gene, depending on the tissue type (Koon *et al.*, 2007). The distribution of insertions in and around *Foxp1* suggests that many are upstream, and therefore that the gene is being upregulated, which is consistent with the oncogenic role of *Foxp1* in lymphomas. However, there is an Ensembl EST gene with no associated Ensembl gene transcript that spans the entire *Foxp1* CIS, suggesting that the insertions could in fact be intragenic.

This section suggests that the MuLV screen can be helpful in identifying candidate tumour suppressor genes. However, computational analysis of insertions in and around genes can only provide an indication of whether a candidate cancer gene is likely to be oncogenic or tumour suppressive, and analysis of gene expression in MuLV-induced tumours, followed by functional validation, is essential for further confirmation.

3.4.4 Identifying retroviral insertions in regulatory features

The orientation and distribution of insertions around genes helps to identify promoter and enhancer mutations and insertions that prematurely terminate gene transcription. However, it is also possible that insertions could disrupt a gene by inserting into regulatory elements, thereby preventing the binding of transcriptional activators or repressors. In Ensembl version 45, regulatory features were available for the human, but not the mouse, genome. Features were built using 3 genome-wide anchor datasets: DNaseI hypersensitivity sites identified by ChIP-seq analysis (Boyle *et al.*, 2008), CCCTC-binding factor (CTCF) binding sites identified by ChIP-Chip (Kim *et al.*, 2007b), and histone 3 lysine 4 tri-methylation (H3K4me3) also identified by ChIP-chip. ChIP-chip, ChIP-seq and DNaseI hypersensitivity (a marker of open chromatin) are discussed in Section 1.3.5. The DNaseI hypersensitivity sites were identified in CD4+ T cells, but most were also found in CD8+ T cells and B cells and around 10% were lymphocyte-specific. This dataset is therefore particularly relevant to the MuLV screen, which generated predominantly lymphomas (see Section 2.2.1). CTCF is an insulator protein that prevents the spread of heterochromatin and prevents enhancers from activating unrelated promoters. CTCF binding sites were identified in primary human fibroblasts but were largely conserved across cell types (Kim *et al.*, 2007b). The histone modification H3K4me3 is associated with transcription start sites of active genes. 5 supporting ChIP-Chip datasets of histone modifications (H4K20me3, H3K27me3, H3K36me3, H3K79me3 and H3K9me3) were also used. In the Ensembl regulatory build, overlapping elements identified in each analysis were merged into a single element, and each element was classified based on the datasets in which it was identified. Elements associated with DNaseI hypersensitivity and H3K36me3 were classified as promoter-associated elements, while elements associated with DNaseI and either H3K4me3 or H3K79me3, or DNaseI and H3K4me3 and either CTCF or H3K36me3, were classified as gene-associated elements. It is worth noting that elements in the regulatory build define regions that are much larger than individual transcription factor binding sites and only define regions that are likely to be involved in regulation.

Since the regulatory build was only available for the human genome, elements and their classifications were downloaded from ftp://anonymous@ftp.ensembl.org/pub/release-45/homo_sapiens_45_36g/data/reg_build/ and were mapped to the NCBI m36 mouse genome assembly using UCSC LiftOver (http://genome.ucsc.edu/cgi-bin/hgLiftOver). Out of 113,230 features, 77,446 (68.4%) were successfully mapped to the mouse build. Of those that failed to convert, 30 were split (i.e. they mapped to 2 locations in the mouse genome due to breaks in synteny), 466 were duplicated, 1,285 were partially deleted and 34,003 were completely deleted. It is not surprising that such a large number of elements did not map, since only a fraction of the human and mouse genomes align confidently and

LiftOver uses these alignments to find corresponding positions between human and mouse.

It has been estimated that 32-40% of human transcription factor binding sites are not functional in rodents, suggesting a high evolutionary turnover of sites (Dermitzakis and Clark, 2002). However, highly conserved elements have also been identified, and are particularly prominent around vertebrate developmental genes (Woolfe *et al.*, 2005). The parameters of LiftOver were set such that the minimum ratio of bases that must remap was 0.1, and therefore it is possible that some of the mapped elements are poorly conserved. However, while this could suggest that the elements are not functional in the mouse, there is evidence to suggest that regulatory function can be conserved without sequence similarity, e.g. in the case of the RET locus in humans and zebrafish (Fisher *et al.*, 2006). Therefore, a decision was made to be inclusive, rather than to reject elements of low sequence similarity.

2,036 (10.1%) of the 20,114 insertion sites mapped to regulatory features, of which 21 (0.02%) were promoter-associated and 1,394 (68.5%) were gene-associated. This compared with 743 (9.9%) of the 7,518 insertion sites associated with the 439 candidate cancer genes, of which 11 (0.02%) were promoter-associated and 443 (59.6%) were gene-associated. 228 candidate genes had insertions within regulatory features. There was no significant difference (P=0.38) between the number of insertions assigned to candidate cancer genes in regulatory features and the number of other insertions in regulatory features. However, insertions assigned to candidate cancer genes were under-represented in gene-associated elements (P=7.48x10^{-11}). This result was surprising since it might be assumed that oncogenic insertions would be more likely to be associated with regulatory elements.

Since many insertions may map to the same regulatory region, therefore skewing the results, the number of regions containing insertions was also counted. Insertions were identified in 1,483 regulatory features, of which 14 (0.94%) were promoter-associated and 971 (65.5%) were gene-associated. Insertions associated with candidate cancer genes were identified in 343 regulatory features, of which 5 (1.5%) were promoter-associated and 160 (46.6%) were gene-associated. Once again, gene-associated elements were under-represented among insertions associated with candidate cancer genes (P=6.00x10^{-17}). Counting the number of regulatory features of each type that contained insertions revealed an under-representation of features associated with H3K36me3 (P=2.04x10^{-3}), H3K4me3 (P=3.38x10^{-15}), H3K79me3 (P=1.65x10^{-4}) and DNaseI hypersensitivity sites (P=0.012). All significance tests were performed using the Chi-squared test for independence. Interestingly, DNaseI hypersensitivity and all of the histone modifications stated above are known to be associated with active genes, while those histone modifications that showed no significant difference (H3K27me3, H3K9me3 and H4K20me3) are associated with gene repression (Barski *et al.*, 2007). H3K4me3 and H3K27me3 have also been shown to be associated with active genes and silent genes, respectively, in human T cells (Roh *et al.*, 2006), which is of particular relevance to this MuLV dataset of lymphomas. Insertions that are not associated with candidate cancer genes are less likely to be oncogenic, and their over-representation in regulatory features associated with active genes may reflect the preference of MuLV for inserting within active genes. Since none of the regulatory features are over-represented among candidate genes, it appears that disruption of regulatory features may not be a common mechanism of mutagenesis of the MuLV retrovirus.

3.4.5 Expression analysis of MuLV-induced tumours

Computational approaches can be used to predict candidate cancer genes and the likely mechanisms of mutation, but these must be confirmed using experimental methods. Gene expression analysis is a useful tool towards validating candidates, since it is expected that genes that are disrupted by MuLV will be differentially expressed in insertion-containing tumours versus those that do not contain insertions. Although widespread expression analysis has not been performed on the MuLV-induced tumours, expression data was available for 18 tumours. The analysis was performed by David Adams using high density Nimblegen 5045 MM8 60mer expression arrays, where MM8 is the mouse build (the UCSC equivalent to NCBI m36) and 60mer is the length of the oligonucleotide probes on the array. The array covers 18,879 transcripts with unique RefSeq NM accession numbers, and 6,751 with RefSeq XM accession numbers. NM and XM refer to reported and predicted transcripts, respectively. Each NM transcript has three probes, while 1,861 XM transcripts have 3 probes and the rest have 2 probes. The normalised expression values across all probes in each transcript, as provided by Nimblegen, were used in this analysis.

81 candidate cancer genes from the MuLV screen contained MuLV insertions in at least 1 of the 18 tumours, and 20 contained insertions in at least 2 tumours. RefSeq accession numbers, Entrez Gene identifiers and MGI symbols were extracted from BioMart (version 49) for each of the 439 candidate cancer genes from the MuLV screen. All of the genes except *mmu-mir-17*, *ENSMUSG00000074675*, *Rnf157* and *Pvt1* were identified on the array. Genes directly flanking each candidate gene were identified using the coordinates of all genes in Ensembl version 45. For candidate genes with insertions in 2 or more tumours, a two-sided *t*-test was performed to determine whether the level of expression in tumours containing an insertion in the gene was significantly different to the level in tumours that did not contain an insertion in the gene. The results are shown in Table 3.4. The *t*-test was also performed on genes flanking the candidate cancer genes, in order to ascertain whether the insertions had been assigned to the correct gene.

Only 1 of the candidates, *Trpm1*, showed significant differential expression in tumours containing an insertion compared to those that did not, and the insertions appeared to cause a decrease in gene expression. Loss of *Trpm1*, also known as melastatin, correlates with metastatic potential in human and mouse melanoma cells (Deeds *et al.*, 2000). Interestingly, the insertions in tumours used in this analysis were 11.7 kb and 21.2 kb upstream of *Trpm1*, which suggests either that there is a longer transcript that is not annotated in Ensembl, or that the gene is disrupted by insertion into upstream regulatory elements, although the insertions did not overlap with regulatory features in the dataset described in Section 3.4.4. Although the difference was not significant, the mean expression level in insertion-containing tumours was at least 2-fold higher than in other tumours for genes *Notch1*, *Rasgrp1*, *Pik3r5*, *Jundm2*, *Pim1* and *Rras2*. None of the genes flanking the candidate genes showed significant differential expression, but *Spon1*, *Lrrc8b* and *Fos*, which flank genes *Rras2*, *Lrrc8c* and *Jundm2*, respectively, had a mean expression level that was at least 2-fold higher in insertion-containing tumours. Due to their enhancer activity, MuLV insertions can have long-range effects, and therefore it is possible that *Spon1* and *Fos* are also affected by insertions disrupting *Rras2* and *Jundm2*, respectively. On the other hand, *Lrrc8b* showed a greater difference in expression than did *Lrrc8c*, and it is possible that *Lrrc8b* is the true candidate cancer gene in this region.

The scale of this analysis was too small to provide any definitive evidence that the correct candidate cancer gene has been selected. The results suggest that there may not be a strong association between insertion-containing genes and higher expression levels. It is

Gene name	Number of tumours with insertions	Mean 1	SD 1	Mean 2	SD 2	*P*-value
Notch1	2	5815.239	6599.7549	1078.072	126.6176	0.1773
Mad1l1	2	4708.679	875.3291	5297.206	1253.8023	0.6712
Rasgrp1	2	14100.646	1935.705	7066.692	4240.0499	0.1195
Pik3r5	2	3813.572	2155.5715	800.739	2405.31807	0.1531
Jundm2	2	3250.289	3778.6357	579.06	571.0842	0.2024
Hnrpf	2	36521.059	7138.7373	37617.657	2704.6885	0.8494
Trpm1	2	153.399	27.9747	398.287	37.2045	0.0002
B3gnt2	2	2291.355	1024.3581	1947.636	857.7561	0.7632
Spn	2	1882.812	264.2865	1629.536	646.2697	0.7024
Hibadh	3	4449.752	1192.4941	5461.04	1742.7667	0.6032
Pim1	3	10856.253	13210.2491	1640.201	962.0191	0.4871
Myb	4	11458.869	5636.9595	8207.643	2707.9078	0.5177
Lrrc8c	4	1939.271	247.5314	1891.16	488.6951	0.927
Evi5	4	867.045	875.3993	797.081	706.3382	0.9485
Ccnd3	6	8402.866	4375.7706	7656.238	2977.1405	0.8832
Rras2	7	21102.81	4947.9028	6690.916	8678.9839	0.16
Myc	7	21190.356	7534.8977	19642.481	8734.1825	0.8881
Gfi1	8	7495.101	2592.9975	4365.898	2676.8333	0.3846

Table 3.4. Gene expression values for candidate cancer genes in insertion-containing tumours compared with tumours that do not contain insertions. Mean 1 and SD 1 are the mean and standard deviation of expression levels for genes in tumours containing insertions, and Mean 2 and SD 2 are the mean and standard deviation for genes in tumours that do not contain insertions. *P*-values were calculated using the *t*-test.

possible that, over time, insertions may lose their ability to disrupt cellular genes by promoter or enhancer mutation, e.g. because the retroviral LTRs are silenced by hypermethylation. Alternatively, since tumours are heterogeneous, and the tumour samples may also contain stromal cells, the effect of the insertion may be diluted by the presence of wildtype gene expression in contaminating cells. An analysis of gene expression across all tumours, with replicates, is required to substantiate these suggestions.

3.5 Identification of co-operating cancer genes in the MuLV dataset

As discussed in Section 1.4.2.1.3, there are two main approaches for identifying collaborating cancer genes using insertional mutagenesis. By conducting the screen in genetically engineered mice in which oncogenes are overexpressed or tumour suppressor genes are inactivated, it is possible to identify genes that collaborate with the gain or loss, respectively, of those cancer genes in oncogenesis. An alternative approach involves identifying co-occurring CIS genes in individual tumours. Both approaches have been employed to analyse the MuLV dataset of 439 statistically significant CIS genes.

3.5.1 Genotype-specific cancer genes

The retroviral screen described in this thesis was performed on mice deficient in a range of tumour suppressor genes (see Section 2.2.1). For each gene identified using the kernel convolution-based method for determining significant CISs, the number of insertions assigned to the gene and the number of insertions of each genotype were counted. See Section 2.9 for a description of the methods used to assign insertions to genes. The 2-tailed Fisher Exact Test was used to determine whether there was any significant difference between the number of insertions of a particular genotype within each gene and the number in the rest of the genome, and also between the number of insertions of a particular genotype and the number of wildtype insertions within a gene compared to the proportions in the rest of the genome:

a	*b*
c	*d*

a = Number of insertions of a given genotype assigned to the gene

b = Number of insertions of a given genotype not assigned to the gene
c = Number of insertions of other genotypes assigned to the gene, or number of wildtype insertions assigned to the gene
d = Number of insertions of other genotypes not assigned to the gene, or number of wildtype insertions not assigned to the gene

Significance tests were performed for each genotype, and also for groups of genotypes to increase the power of the analyses. For example, in order to test for a significant bias towards any *p21*-deficient background, all insertions on a *p21*-null homozygous or heterozygous background, and all insertions on a homozygous or heterozygous *p21*-null and *p27*-null double mutant background, were counted. In order to account for multiple testing, the R package QVALUE (Storey and Tibshirani, 2003) was used to generate a *q*-value for each test. The *q*-value is a measure of the minimum false discovery rate incurred if the test is called significant, where the false discovery rate is the number of false positives divided by the number of significant tests. This differs from the *P*-value, which is a measure of the minimum false positive rate incurred when the test is called significant, where the false positive rate is the number of false positives divided by the number of true null tests. Using a *P*-value of 0.05, 5% of tests will be called significant when they are in fact null, which would result in a very high number of false positives if a large number of tests were performed. On the other hand, using a *q*-value of 0.05, 5% of the tests that have been called significant will be false positives, which is more manageable for large numbers of tests. The QVALUE *bootstrap* method was used to estimate the overall proportion of true null hypotheses, since this method is deemed most appropriate for situations in which the distribution of null *P*-values is skewed towards a value of 1, as is the case in this analysis. The shape of the distribution also means that the calculated *q*-values are very conservative, but this method is still more inclusive than using, for example, the Bonferroni correction. The most significant tests, where the *q*-value is less than 0.05 for the comparison with insertions of all other genotypes and/or with only wildtype insertions, and that therefore suggest a bias towards, or away from, a particular tumour genotype, are presented in Tables 3.5A and 3.5B, respectively. Results are grouped by gene, with genes ordered according to the most significant association obtained from the comparison with insertions of all other genotypes. The two methods gave similar results, although the comparison of a given genotype versus wildtype gave fewer significant results. This may be because some genes always require the co-

A

					***vs.* insertions of all other genotypes**		***vs.* wildtype insertions**	
Gene	**Ensembl Gene ID**	**Total number of insertions**	**Genotype**	**Insertions of given genotype**	***P*-value**	***q*-value**	***P*-value**	***q*-value**
Evi5	ENSMUSG00000011831	466	*p27 all*	134	1.87E-29	8.19E-27	1.77E-15	4.50E-13
			p21 all	100	1.57E-10	3.45E-08	2.30E-06	5.06E-04
			p21, p27	71	7.83E-10	1.72E-07	7.17E-07	1.57E-04
			p27	66	2.93E-09	1.29E-06	7.91E-07	3.47E-04
			p27 ko	40	3.89E-08	1.71E-05	9.34E-07	4.10E-04
			p21 ko, p27 ko	35	1.38E-06	3.03E-04	8.50E-06	1.86E-03
Gfi1	ENSMUSG00000029275	458	*p27 all*	127	2.44E-26	5.35E-24	2.05E-15	4.50E-13
			p21, p27	77	1.01E-12	4.44E-10	1.40E-09	6.16E-07
			p21 all	102	1.19E-11	5.23E-09	7.63E-08	3.35E-05
			p21 ko, p27 ko	37	6.98E-08	3.07E-05	2.07E-07	9.10E-05
			p27 ko	35	4.21E-06	9.24E-04	8.24E-06	1.81E-03
			p27	55	1.27E-05	2.79E-03	3.97E-05	8.71E-03
			p21 ko, p27 het	35	8.59E-05	3.77E-02	8.13E-05	3.57E-02
Map3k8	ENSMUSG00000024235	37	*p16 ko, p19 ko*	27	2.77E-21	1.21E-18	5.59E-10	2.45E-07
			p16 all	28	1.67E-20	7.33E-18	5.33E-09	2.34E-06
			p16, p19	28	1.67E-20	7.33E-18	5.33E-09	2.34E-06
			p19 all	31	2.13E-09	7.57E-07	6.05E-03	5.67E-01
Myc	ENSMUSG00000022346	359	*p27 all*	74	6.14E-09	8.98E-07	3.56E-06	5.20E-04
Myb	ENSMUSG00000019982	247	*p27 all*	51	1.30E-06	1.35E-04	1.37E-03	1.00E-01
Art2b	ENSMUSG00000030651	11	*p27 all*	8	1.53E-06	1.35E-04	1.99E-03	1.22E-01
Pvt1	ENSMUSG00000072566	296	*p27 all*	55	1.72E-05	1.26E-03	2.23E-03	1.22E-01
A530013C23Rik	ENSMUSG00000006462	43	*p16 all*	15	1.75E-05	3.57E-03	7.46E-04	8.08E-02
			p16, p19	15	1.75E-05	3.57E-03	7.46E-04	8.08E-02
			p16 het, p19 het	6	1.22E-04	5.36E-02	8.57E-02	1.95E-02
Ccnd3	ENSMUSG00000034165	206	*p53 ko*	59	2.23E-05	9.79E-03	1.75E-06	7.68E-04
			p53	59	5.64E-05	2.47E-02	3.18E-06	7.68E-04
Zfp438	ENSMUSG00000050945	51	*p19 ko*	27	2.51E-05	3.67E-03	7.06E-03	8.13E-01
	ENSMUSG00000059894	170	*p27 all*	36	3.30E-05	2.07E-03	1.80E-02	6.64E-01
Pim1	ENSMUSG00000024014	118	*p21 all*	29	4.81E-05	7.04E-03	4.72E-03	4.14E-01
Rras2	ENSMUSG00000055723	224	*p27 all*	43	5.42E-05	2.97E-03	2.19E-04	2.40E-02
OTTMUSG00000012358	ENSMUSG00000052248	40	*p16 ko, p19 ko*	11	4.48E-04	3.93E-02	8.20E-03	5.14E-01
Zeb2	ENSMUSG00000026872	40	*p16 ko, p19 ko*	11	4.48E-04	3.93E-02	8.20E-03	5.14E-01
Mycn	ENSMUSG00000037169	81	*p27 all*	19	6.07E-04	2.96E-02	2.80E-04	2.46E-02
Ahi1	ENSMUSG00000019986	124	*p27 all*	25	9.15E-04	4.02E-02	5.17E-02	8.61E-01

B

					***vs.* insertions of all other genotypes**		***vs.* wildtype insertions**	
Gene	**Ensembl Gene ID**	**Total number of insertions**	**Genotype**	**Insertions of given genotype**	***P*-value**	***q*-value**	***P*-value**	***q*-value**
Evi5	ENSMUSG00000011831	466	*p19 ko*	71	2.43E-07	1.07E-04	7.20E-03	8.13E-01
Gfi1	ENSMUSG00000029275	458	*p19 all*	114	9.17E-07	1.34E-04	1.56E-01	1.00E+00
			p19 ko	76	1.51E-05	3.31E-03	9.18E-02	1.00E+00
Rasgrp1	ENSMUSG00000027347	237	*p16 all*	7	2.44E-05	3.57E-03	2.76E-05	6.06E-03
			p16, p19	7	2.44E-05	3.57E-03	2.76E-05	6.06E-03
			p16 ko, p19 ko	6	1.73E-04	2.53E-02	1.65E-04	2.41E-02
Ccnd3	ENSMUSG00000034165	206	*p16 ko, p19 ko*	4	9.19E-05	2.02E-02	1.25E-01	1.00E+00
			p16 all	7	2.27E-04	2.49E-02	3.15E-01	1.00E+00
			p16, p19	7	2.27E-04	2.49E-02	3.15E-01	1.00E+00
Ikzf1	ENSMUSG00000018654	93	*p53*	4	2.75E-04	5.30E-02	5.99E-05	8.76E-03
			p53 ko	2	4.14E-04	7.49E-02	6.54E-05	9.58E-03
Zmiz1	ENSMUSG00000007817	62	*p16 ko, p19 ko*	2	5.79E-03	2.06E-01	9.60E-05	2.11E-02
			p53	0	6.36E-03	1.99E-01	4.95E-05	8.76E-03
			p53 ko	2	9.62E-03	3.02E-01	5.37E-05	9.58E-03

Table 3.5. Genes containing an over-representation (A) or under-representation (B) of insertions on a given tumour background compared with all other backgrounds and compared with wild-type insertions only. All genes identified in tests with a *q*-value of less than 0.05 for one or both methods are shown. *p27* all = all genotypes that include a mutation in *p27* (homozygous or heterozygous, single or double mutant); *p27* = single homozygous or heterozygous mutation of *p27*; *p27* ko = single homozygous mutation of *p27*. All other genotypes follow the same rules. Genes are listed in order of decreasing significance (increasing *P*-value) with respect to the comparison with insertions from all other genotypes.

operation of other cancer genes and are therefore rarely or never mutated on a wildtype background.

The most significant result was the bias of insertions in *Evi5/Gfi1* towards *p27*-deficient genetic backgrounds. Interestingly, while insertions within this locus were identified in an MuLV screen performed on *p27*-deficient mice by Hwang and coworkers (2002), no significant difference was observed between the frequency of insertions in *p27*$^{-/-}$ and wildtype mice. The screen, which involved 50 tumours, was smaller than the screen described in this thesis, and therefore the difference may reflect the increased power of this larger dataset and supports the use of larger insertional mutagenesis screens for identifying co-operating oncogenes. In accordance with the observations of Hwang *et al.* (2002), insertions in *Myc* showed a significant bias towards *p27*-deficient genotypes. This is also supported by the finding that *p27*-deficient lymphomas show an increased frequency of *Myc* activation, and that *Myc*-induced tumourigenesis may be enhanced upon loss of *p27* (Martins and Berns, 2002). *Mycn*, which is structurally and functionally related to *Myc*, was also associated with *p27*-deficient tumours. *MYCN* amplification in human neuroblastomas is associated with poor prognosis (Seeger *et al.*, 1985). Low expression of *p27* is also correlated with poor prognosis in patients with neuroblastoma, yet *p27* expression and *MYCN* amplification are prognostic independent and are not significantly associated in neuroblastomas (Bergmann *et al.*, 2001). While this seems to suggest that disrupted *p27* and *MYCN* are not collaborating in neuroblastoma, it does indicate that the genes may act in different genetic pathways, as is generally expected for genes that collaborate in tumourigenesis.

Insertions in *Map3k8* showed the most significant bias to the *p16*$^{-/-}$*p19*$^{-/-}$ (or *Cdkn2a*$^{-/-}$) tumour genotype. *Map3k8* has previously been identified as *Cdkn2a*$^{-/-}$-specific in an MuLV screen performed on 115 mice (Lund *et al.*, 2002). Activation of Mek by Map3k8 in the mitogen-activated protein kinase (MAPK) signalling pathway (Salmeron *et al.*, 1996) induces p16 and p53, resulting in the permanent arrest of mouse fibroblasts (Lin *et al.*, 1998). However, in the absence of p16 or p53, the activation of the MAPK cascade causes cells to undergo uncontrolled mitogenesis and transformation (Lin *et al.*, 1998).

Insertions affecting the gene encoding zinc finger E-box-binding homeobox 2 (Zeb2 or Sip1) were also significantly associated with the *Cdkn2a*$^{-/-}$ genotype. *SIP1* plays a role in replicative senescence, which controls the number of cell divisions in human somatic tissues and so prevents the indefinite proliferation associated with tumour cells (Ozturk *et al.*, 2006). Inactivation of *SIP1* causes reactivation of the human telomerase reverse transcriptase (*hTERT*), resulting in the rescue of hepatocellular carcinoma cells from senescence arrest (Ozturk *et al.*, 2006). Replicative immortality also requires the inactivation of *Trp53* and *p16* (Ozturk *et al.*, 2006), therefore suggesting co-operation between *p16* and *SIP1* in tumourigenesis. Interestingly, all of the insertions within *Zeb2* are flanking an internal gene, and most are in the upstream sense orientation, suggesting that they are promoter insertions that upregulate the internal gene. This internal gene is a natural antisense transcript that, when overexpressed in epithelial cells, prevents splicing of the *Zeb2* 5' UTR (Beltran *et al.*, 2008). However, this is proposed to increase the levels of Zeb2 (Beltran *et al.*, 2008), which conflicts with the observations described above. Zeb2 also directly represses cyclin D1, resulting in initiation of the epithelial-mesenchymal transition (EMT), in which cells switch from a proliferative to an invasive state (Mejlvang *et al.*, 2007). *Cyclin D1* (*Ccnd1*) was also biased towards the *Cdkn2a*$^{-/-}$ tumour background, albeit with lower significance (*P*=2.16x10^{-3}, *q*=0.135). p16 binds to CDK4 and prevents it from forming a complex with cyclin D1, resulting in cell cycle arrest at the G1/S transition (Serrano *et al.*, 1993). An enhanced gene ratio of *CCND1:CDKN2A*, i.e. a high copy number of *CCND1* combined with deletion of *CDKN2A*, correlates with poor survival in patients with squamous cell carcinoma of the head and neck (Akervall *et al.*, 2003), while the combined loss of *p16* and overexpression of *cyclin D1* has been observed in 49% of gastric carcinomas (Kishimoto *et al.*, 2008). It therefore appears that *Ccnd1*, *Zeb2* and *Cdkn2a* may collaborate in tumourigenesis, where *Ccnd1* causes uncontrolled cell growth in the absence of *Cdnk2a*, and *Zeb2* represses *Ccnd1*, causing hyperproliferating cells to undergo EMT.

Insertions in the oncogene *Pim1* were associated with *p21*-deficient tumours. Phosphorylation of *p21* by Pim1 results in the cytoplasmic localisation (Wang *et al.*, 2002b) or stabilisation of p21 (Zhang *et al.*, 2007), and this is proposed to be a contributing factor in the tumourigenesis of cells overexpressing *Pim1* (Zhang *et al.*, 2007). However, the fact that *Pim1* mutagenesis is favoured in a *p21*-deficient background suggests that overexpression of *Pim1* alone cannot fully inactivate *p21* and the genes may have a more complex relationship that has not been elucidated. Insertions in *Runx1* were biased towards the *p53*$^{-/-}$ genetic background.

If a gene contains fewer insertions than expected in tumours bearing a particular inactivated tumour suppressor gene, this suggests that the CIS gene and the inactivated tumour suppressor gene may act in the same cancer pathway. Insertions in *Zmiz1* and *Ikaros* (*Ikzf1*) were under-represented in *p53*$^{-/-}$ tumours. Zmiz1 is a transcriptional co-activator of p53 (Lee *et al.*, 2007) and therefore, in the absence of p53, mutation of *Zmiz1* does not provide any additional growth advantage. The results for *Ikaros* are more surprising since, in chemically induced murine lymphomas, allelic loss of *Ikaros* was more frequently found in *p53*$^{-/-}$ lymphomas than in wildtype *p53* lymphomas, suggesting cooperation in lymphomagenesis (Okano *et al.*, 1999). Further functional evidence is required to validate this proposal. None of the genes identified as containing p53 binding sites in Section 3.2 were positively or negatively associated with the *p53*$^{-/-}$ tumour background. Some of the genes contained few insertions, and there may not be enough power to identify a significant association. For example, *Chd1* did not contain any insertions in *p53*$^{-/-}$ tumours but only contained 7 insertions overall. Alternatively, the relationship with p53 may not be relevant in the setting of MuLV-induced lymphomagenesis. For example, p53-mediated upregulation of *Notch1* contributes to cell fate determination (Alimirah *et al.*, 2007), but in MuLV insertional mutagenesis, *Notch1* is activated by truncating mutations and is therefore not dependent on p53 (P=3.73x10^{-4}, q=5.30x10^{-2}).

Further genes for which there was a strong bias towards or against a particular tumour genotype are listed in Table 3.5. Supporting evidence in the literature for the genes described above indicates that this may be a powerful method for identifying cooperating cancer genes.

3.5.2 Co-occurrence and mutual exclusivity of disrupted genes

The "genotype-specific" approach for identifying collaborating cancer genes only allows for the identification of collaborations with selected oncogenes or tumour suppressor genes. Identifying CIS genes that co-occur in tumours more often than expected by chance enables the identification of collaborations without any predetermined conditions. In Section 1.4.2.1.3, oligoclonality is cited as a potential disadvantage of this approach. However, since only significant CISs are utilised, insertions within these CISs are likely to be present, and to co-occur, in the dominant clone, rather than being rare insertions in less successful sublines of the tumour.

For each pair of CIS genes, the number of tumours that contained an insertion in both genes, or in one or other gene, was counted. The 2-tailed Fisher Exact Test was used to determine whether the number of tumours containing a co-occurrence of each gene pair was significantly different to the number expected by chance.

a	*b*
c	*d*

a = Number of tumours containing an insertion in both genes
b = Number of tumours containing an insertion in first gene
c = Number of tumours containing an insertion in second gene
d = Number of tumours containing an insertion in neither gene

To account for multiple testing, the R package QVALUE (Storey and Tibshirani, 2003), and specifically the *Bootstrap* method, was applied to all tests in which one or more co-occurrences were observed. Over- and under-represented co-occurrences with a *q*-value of less than 0.05 are shown in Tables 3.6A and 3.6B, respectively.

The most significant association was between genes *A530013C23Rik* and leukocyte-specific protein tyrosine kinase *Lck*. Lck initiates a tyrosine phosphorylation cascade in lymphocytes that results in T-cell antigen receptor signal transduction, and it is overexpressed in lymphomas, breast cancer and colon cancer (for review, see Palacios and Weiss, 2004). Interestingly, insertions in both *Lck* and *A530013C23Rik* were biased towards a *Cdkn2a*$^{-/-}$ genotype (*Lck*: *P*=9.12x10^{-4} and *q*=5.01x10^{-2}; *A530013C23Rik*: see Table 3.5A), suggesting that all 3 genes collaborate in tumourigenesis.

Co-occuring insertions were also identified in *Lck* and signal transducer and activation of transcription 5b (*Stat5b*). LCK has been shown to interact with STAT5b in cells, and induces tyrosine phosphorylation and DNA-binding of STAT5b (Shi *et al.*, 2006). Exogenous expression of wildtype *STAT5b* increases LCK-mediated cellular transformation (Shi *et al.*, 2006). This is consistent with the pattern of insertions in and around *Stat5b*, which suggests that the gene is upregulated by promoter and enhancer mutations that increase the levels of the wildtype protein. Finally, activation of *Lck* was also significantly associated with activation of the c-src tyrosine kinase gene *Csk*. Csk negatively regulates Lck by phosphorylation of a C-terminal tyrosine (Tyr-505) (Bergman *et al.*, 1992). The distribution of insertions in and around *Csk* suggests that the gene is

A

Gene name 1	Ensembl ID 1	Total tumours in which Gene 1 disrupted	Gene name 2	Ensembl ID 2	Total tumours in which Gene 2 disrupted	Number of tumours in which Gene 1 and Gene 2 disrupted	*P*-value	*q*-value
A530013C23Rik	ENSMUSG00000006462	43	*Lck*	ENSMUSG00000000409	26	10	2.47263E-08	8.59E-06
Ikzf1	ENSMUSG00000018654	93	*Notch1*	ENSMUSG00000026923	127	30	1.6337E-07	4.80E-05
Zfp438	ENSMUSG00000050945	51	*Ntn1*	ENSMUSG00000020902	59	14	3.41404E-07	9.66E-05
Pik3r5	ENSMUSG00000020901	64	*Zfp438*	ENSMUSG00000050945	51	14	1.02005E-06	2.52E-04
Epha6	ENSMUSG00000055540	20	*Pim1*	ENSMUSG00000024014	118	11	2.78357E-06	6.26E-04
Runx1	ENSMUSG00000022952	143	*Rasgrp1*	ENSMUSG00000027347	237	54	4.93807E-05	9.95E-03
Stat5b	ENSMUSG00000020919	18	*Lck*	ENSMUSG00000000409	26	5	5.55279E-05	1.06E-02
Nid1	ENSMUSG00000005397	12	*Cd3e*	ENSMUSG00000032093	8	3	7.27389E-05	1.32E-02
Lfng	ENSMUSG00000029570	33	*Notch1*	ENSMUSG00000026923	127	13	7.96187E-05	1.38E-02
Ppp2r5a	ENSMUSG00000026626	10	*Vps13d*	ENSMUSG00000020220	10	3	8.4762E-05	1.44E-02
Cd48	ENSMUSG00000015355	10	*Arhgap26*	ENSMUSG00000036452	11	3	1.16E-04	1.81E-02
Psma1	ENSMUSG00000030751	9	*mmu-mir-17*	ENSMUSG00000065508	33	4	1.13E-04	1.81E-02
Fgr	ENSMUSG00000028874	7	*Dad1*	ENSMUSG00000022174	16	3	1.15E-04	1.81E-02
Ubxd5	ENSMUSG00000012126	14	*Thra*	ENSMUSG00000058756	9	3	1.78E-04	2.56E-02
Rras2	ENSMUSG00000055723	224	*Rasgrp1*	ENSMUSG00000027347	237	75	1.75E-04	2.56E-02
Sdk1	ENSMUSG00000039683	13	*Mns1*	ENSMUSG00000032221	10	3	1.99E-04	2.77E-02
Zbtb7b	ENSMUSG00000028042	12	*Notch1*	ENSMUSG00000026923	127	7	2.15E-04	2.94E-02
Evi1	ENSMUSG00000027684	45	*AB041803*	ENSMUSG00000044471	14	5	2.23E-04	3.00E-02
Cd48	ENSMUSG00000015355	10	*Fgfr2*	ENSMUSG00000030849	14	3	2.52E-04	3.12E-02
Hvcn1	ENSMUSG00000064267	7	*Pygm*	ENSMUSG00000032648	4	2	2.53E-04	3.12E-02
D12Ertd553e	ENSMUSG00000020589	14	*Ptpre*	ENSMUSG00000041836	10	3	2.52E-04	3.12E-02
Eng	ENSMUSG00000026814	6	*Gse1*	ENSMUSG00000031822	25	3	2.67E-04	3.24E-02
Mylc2pl	ENSMUSG00000005474	11	*Bcl11a*	ENSMUSG00000000861	13	3	2.71E-04	3.24E-02
mmu-mir-802	ENSMUSG00000076457	143	*Rasgrp1*	ENSMUSG00000027347	237	52	2.79E-04	3.28E-02
Csk	ENSMUSG00000032312	14	*Lck*	ENSMUSG00000000409	26	4	3.08E-04	3.57E-02
Smg6	ENSMUSG00000038290	45	*Pik3r5*	ENSMUSG00000020901	64	10	3.18E-04	3.63E-02
Fgfr2	ENSMUSG00000030849	14	*Plac8*	ENSMUSG00000029322	11	3	3.43E-04	3.86E-02
A530013C23Rik	ENSMUSG00000006462	43	*Hhex*	ENSMUSG00000024986	16	5	3.66E-04	4.00E-02
A530013C23Rik	ENSMUSG00000006462	43	*Exoc6*	ENSMUSG00000053799	16	5	3.66E-04	4.00E-02
Spsb4	ENSMUSG00000046997	7	*6430598A04Rik*	ENSMUSG00000045348	5	2	4.20E-04	4.23E-02
Arid3a	ENSMUSG00000019564	6	*Rreb1*	ENSMUSG00000039087	29	3	4.21E-04	4.23E-02
Tcfap4	ENSMUSG00000005718	7	*Jph4*	ENSMUSG00000022208	5	2	4.20E-04	4.23E-02
Zfp608	ENSMUSG00000052713	24	*Olfr56*	ENSMUSG00000040328	7	3	4.06E-04	4.23E-02
Parvg	ENSMUSG00000022439	6	*Prr6*	ENSMUSG00000018509	6	2	4.50E-04	4.29E-02
Frmd8	ENSMUSG00000043488	12	*Ubxd5*	ENSMUSG00000012126	14	3	4.54E-04	4.29E-02
Frmd8	ENSMUSG00000043488	12	*AB041803*	ENSMUSG00000044471	14	3	4.54E-04	4.29E-02
Ubxd5	ENSMUSG00000012126	14	*Scyl1*	ENSMUSG00000024941	12	3	4.54E-04	4.29E-02
AB041803	ENSMUSG00000044471	14	*Scyl1*	ENSMUSG00000024941	12	3	4.54E-04	4.29E-02
Gimap6	ENSMUSG00000047867	3	*Bcl11a*	ENSMUSG00000000861	13	2	4.70E-04	4.38E-02
Zmiz1	ENSMUSG00000007817	62	*Notch1*	ENSMUSG00000026923	127	18	4.96E-04	4.58E-02
Cecr5	ENSMUSG00000058979	16	*Nfkb1*	ENSMUSG00000028163	11	3	5.22E-04	4.59E-02
B3gntl1	ENSMUSG00000046605	11	*Hhex*	ENSMUSG00000024986	16	3	5.22E-04	4.59E-02
B3gntl1	ENSMUSG00000046605	11	*Exoc6*	ENSMUSG00000053799	16	3	5.22E-04	4.59E-02
Irf2bp2	ENSMUSG00000051495	26	*Nsmce1*	ENSMUSG00000030750	7	3	5.18E-04	4.59E-02
Jundm2	ENSMUSG00000034271	105	*Runx1*	ENSMUSG00000022952	143	28	5.67E-04	4.87E-02
Myb	ENSMUSG00000019982	247	Rras2	ENSMUSG00000055723	224	76	5.79E-04	4.92E-02

B

Gene name 1	Ensembl ID 1	Total tumours in which Gene 1 disrupted	Gene name 2	Ensembl ID 2	Total tumours in which Gene 2 disrupted	Number of tumours in which Gene 1 and Gene 2 disrupted	*P*-value	*q*-value
Ikzf1	ENSMUSG00000018654	93	*Evi5*	ENSMUSG00000011831	466	11	6.30E-14	3.01E-11
Evi5	ENSMUSG00000011831	466	*Notch1*	ENSMUSG00000026923	127	25	1.79E-11	8.06E-09
Ikzf1	ENSMUSG00000018654	93	*Gfi1*	ENSMUSG00000029275	458	16	1.41E-09	5.68E-07
Evi5	ENSMUSG00000011831	466	*Rasgrp1*	ENSMUSG00000027347	237	71	1.92E-09	7.35E-07
Gfi1	ENSMUSG00000029275	458	*Rasgrp1*	ENSMUSG00000027347	237	72	2.73E-08	9.08E-06
Ikzf1	ENSMUSG00000018654	93	*Myc*	ENSMUSG00000022346	359	11	5.30E-08	1.69E-05
Gfi1	ENSMUSG00000029275	458	*Jundm2*	ENSMUSG00000034271	105	23	8.30E-08	2.54E-05
Jundm2	ENSMUSG00000034271	105	*Evi5*	ENSMUSG00000011831	466	25	4.60E-07	1.26E-04
Gfi1	ENSMUSG00000029275	458	*Notch1*	ENSMUSG00000026923	127	33	9.12E-07	2.41E-04
Myc	ENSMUSG00000022346	359	*Notch1*	ENSMUSG00000026923	127	22	9.88E-07	2.52E-04
Map3k8	ENSMUSG00000024235	37	*Myc*	ENSMUSG00000022346	359	1	1.68E-06	4.02E-04
Gfi1	ENSMUSG00000029275	458	*Lck*	ENSMUSG00000000409	26	1	2.60E-06	6.03E-04
Mycn	ENSMUSG00000037169	81	*Myc*	ENSMUSG00000022346	359	12	1.74E-05	3.70E-03
Evi5	ENSMUSG00000011831	466	*Lck*	ENSMUSG00000000409	26	2	2.57E-05	5.31E-03
Gfi1	ENSMUSG00000029275	458	*Zfp438*	ENSMUSG00000050945	51	10	7.48E-05	1.33E-02
A530013C23Rik	ENSMUSG00000006462	43	*Evi5*	ENSMUSG00000011831	466	8	1.25E-04	1.92E-02
Ccnd3	ENSMUSG00000034165	206	*Ikzf1*	ENSMUSG00000018654	93	6	1.34E-04	2.00E-02
Mycn	ENSMUSG00000037169	81	*Notch1*	ENSMUSG00000026923	127	1	1.99E-04	2.77E-02
Gfi1	ENSMUSG00000029275	458	*Rras2*	ENSMUSG00000055723	224	79	2.51E-04	3.12E-02
Ccr7	ENSMUSG00000037944	50	*Evi5*	ENSMUSG00000011831	466	11	2.42E-04	3.12E-02
Pvt1	ENSMUSG00000072566	296	*Mycn*	ENSMUSG00000037169	81	11	5.62E-04	4.87E-02

Table 3.6. Gene pairs in which insertions co-occur more often (A) or less often (B) than expected by chance. All tests with a *q*-value of less than 0.05 are shown.

upregulated by promoter and enhancer insertions, rather than being inactivated, as would be expected for co-operation in tumourigenesis. The distribution of *Lck*-associated insertions suggests that the full-length protein is produced (see Section 3.4.1), and can therefore be phosphorylated by Csk. Further experimental analysis of these genes is therefore required to understand their cooperative role.

A highly significant association was also identified between *Ikaros* (*Ikzf1*) and *Notch1*. An MuLV screen performed on transgenic mice expressing the oncogenic *Notch1* intracellular domain has previously identified the disruption of *Ikaros* as a co-operating event in lymphomagenesis (Beverly and Capobianco, 2003). Loss of heterozygosity of *Ikaros* and activation of *Notch1* have also been shown to co-occur in mouse thymic lymphomas induced by gamma-irradiation (Lopez-Nieva *et al.*, 2004; Ohi *et al.*, 2007). Activating insertions also co-occurred in *Notch1* and lunatic fringe (*Lfng*). *Lfng* encodes a glycosyltransferase that initiates elongation of *O*-linked fucose residues attached to the extracellular epidermal growth factor-like domain of Notch1 (Moloney *et al.*, 2000). This increases the sensitivity of Notch1 to Delta-like, rather than Jagged, Notch ligands and so promotes T cell, rather than B cell, development from haematopoietic progenitors (Besseyrias *et al.*, 2007; Haines and Irvine, 2003; Visan *et al.*, 2006). Upregulation of *Lfng* by insertional mutagenesis may therefore contribute to tumourigenesis by mediating an increase in the binding of oncogenic Notch1 to Delta-like ligands.

A significant co-occurrence was also identified between *Runx1* and *Rasgrp1*. The *Runx1* gene encodes the DNA binding alpha subunit of the Runt domain transcription factor PEBP2/CBF. *Runx1* translocations and point mutations are frequently implicated in human leukaemias and are often associated with activation of the Ras pathway (Goemans *et al.*, 2005). *Rasgrp1* is a Ras GTPase-specific guanine nucleotide exchange factor that activates Ras in lymphocytes (Roose *et al.*, 2007) and, in support of the observed co-occurrence, it was shown to be preferentially targeted by the endogenous retrovirus in BXH2-*Runx1*$^{+/-}$ mice (Yamashita *et al.*, 2005).

There were also numerous genes for which the number of co-occurrences was lower than expected. The lack of co-operation between *Myc* and *Mycn* reflects the fact that they are structurally and functionally related. Co-occurring insertions disrupting *Myc* and either *Ikaros* or *Notch1* were also under-represented. *Myc* is a transcriptional target of the Notch signalling pathway in T cell acute lymphoblastic leukaemia, and Notch1 is required to sustain the high levels of Myc that are required for continued growth and survival of the cancer (Sharma *et al.*, 2007). The mutual exclusivity of activated *Notch1* and *Myc* in mouse tumours suggests that *Notch1* activation may not provide a significant growth advantage when high levels of Myc are sustained by constitutive overexpression. *Mycn* and *Notch1* were also mutually exclusive, suggesting that Notch1 may play a similar role in the maintenance of *Mycn* expression during tumourigenesis. Co-occurring insertions that disrupt *Gfi1* and either *Rras2* or *Rasgrp1* were also under-represented. *Rasgrp1* and Ras-related *Rras2* were significantly associated, suggesting that Rasgrp1 activates *Rras2* and that overexpression of both genes contributes to tumourigenesis. The mutual exclusivity of disrupted *Gfi1* with both of these activated genes suggests that they may act in a common cancer pathway.

Many of the significant associations identified in this analysis are supported by observations in the literature, yet there are many more for which there is no evidence, in many cases because little is known about the genes involved. The list of co-occurring and mutually exclusive genes therefore provides a basis for future functional analyses, and demonstrates the potential of large scale insertional mutagenesis screens in the identification of cancer gene collaborations in mouse, and human, tumourigenesis.

3.6 Discussion

The purpose of the work described in this chapter was to characterise the candidate cancer genes identified by insertional mutagenesis, and to demonstrate their relevance to human tumourigenesis. The candidates showed a significant overlap with human mutation datasets associated with, or biased towards, cancers of haematopoietic and lymphoid tissue, but not with breast and colon candidate cancer genes. This suggests that the screen may only be effective in identifying novel candidates involved in the development of lymphomas and/or leukaemias. A number of the over-represented GO terms were also associated with the development, differentiation or carcinogenesis of T- and B-cells, but others were associated with general features of cancers, such as cell proliferation and apoptosis. An exciting observation to be followed up was the positive association between candidate genes and genes containing Nanog and Oct4 binding sites. This suggests that a significant proportion of the candidate genes may be involved in tumour cell self-renewal, which has not been previously reported in insertional mutagenesis screens. This chapter also presents evidence for an overlap between genes identified by insertional mutagenesis and regions of copy number change in human acute lymphoblastic leukaemias, therefore providing a justification for the cross-species comparative analyses performed in Chapters 4 and 5.

The comparison of candidate genes identified in the MuLV and *Sleeping Beauty* screens demonstrated differences in the mutational profiles of the two mutagens. This suggests that the use of different mutagens can increase the spectrum of candidate cancer genes, but the difference in profiles can only be fully appreciated by comparing fully saturated screens, since some CISs may be missing from one screen simply because of an insufficient number of PCRs or low sequencing depth. However, comparison of the screens does provide strong evidence that overlapping genes are involved in tumourigenesis, rather than resulting from insertional bias, which differs in MuLV and *Sleeping Beauty* (see Section 1.4.2.1.1). *Qsk* was flagged as a promising candidate following its identification in both screens. In light of this finding, Fanni Gergely at the Cambridge Research Institute performed RNAi-mediated knockdown of *QSK* in HeLa cells, which are an immortal cell line derived from human cervical cancer cells, and scored chromosome lagging in 40 late anaphase/early telophase cells in 2 separate experiments. Chromosome lagging was observed in 12.3 ± 2.2 control cells and 28.1 ± 4.0 cells with *QSK* knocked down by 95-100% (Figure 3.8). No other mitotic defects were observed. Chromosome lagging at anaphase can result in the failure of a chromosome or chromatid to become incorporated into one of the daughter nuclei following cell division. This causes aneuploidy and can therefore contribute to genomic instability and cancer formation. This study is ongoing, but suggests that *QSK* does play an important role in tumourigenesis. Likewise, *p116Rip*, *Zmiz1* and *ENSMUSG0000075015* contained both MuLV and T2/Onc insertions and are therefore promising candidates for which functional validation is required.

Co-occurring MuLV and T2/Onc insertions were also used in the prediction of the mechanisms of mutation of candidate cancer genes. While these may not always recapitulate the mutations observed in human cancer, as demonstrated for *Flt3*, in other cases, e.g. *Notch1*, similar mutations are observed. Identifying the structure and function of the mutant products of oncogenes is valuable in the development of therapeutic drugs that target those proteins. Experimental approaches are required to validate the predictions, although the efficacy of gene expression analysis appears to be variable, since in the limited analysis performed in Section 3.4.5, many CIS genes did not show

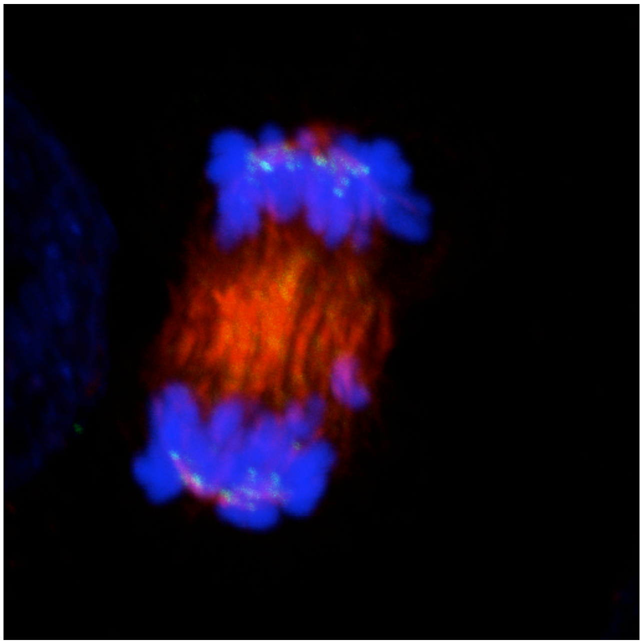

Figure 3.8. Knockdown of *QSK* in human HeLa cells is associated with increased chromosome lagging at anaphase. Figure shows a single cell at anaphase, with chromosomes stained blue and spindle fibres stained red. Image provided by Fanni Gergely at the Cambridge Research Institute.

significant differential expression in tumours containing insertions versus those without. Promoter and enhancer mutations appeared to be the most common mechanisms of mutation, with upregulation of the wildtype gene being the most common type of mutation overall. Initial comparison against a predicted set of regulatory features suggests that disruption of regulatory elements is not a common mechanism of mutation in insertional mutagenesis, although a more accurate analysis could be performed using a set of regulatory features specific for the mouse, rather than human. Analysis of the distribution of insertions within genes can also facilitate the identification of tumour suppressor genes, which are expected to contain only intragenic, truncating mutations, and may show a more random distribution of insertions that includes multiple insertions from the same tumour. For a number of the genes studied in this chapter (i.e. *Etv6*, *Myb*, *Fli1*, *Erg*, *Foxp1*), some of the insertions appeared to be associated with Ensembl EST genes for which there was no associated Ensembl gene transcript. Therefore, analysis of the distribution of insertions from insertional mutagenesis screens may also facilitate the identification of novel gene transcripts.

The identification of collaborating cancer genes is important for the development of targeted cancer therapies. As discussed in Section 1.2.7, cancers can develop resistance to targeted therapies but this may be alleviated by developing therapies that target multiple genes simultaneously. Collaborating cancer genes can also help in deciphering the complex landscape of cancer genomes and the events involved in the multi-step process of tumour evolution. The analyses described in Section 3.5 have identified a number of collaborations for which there is supporting evidence in the literature, as well as many novel collaborations.

In summary, this chapter demonstrates that insertional mutagenesis is a powerful tool for identifying both novel candidate cancer genes and collaborations between candidate cancer genes that are relevant to mouse and human tumourigenesis. In order to maximise the candidates and collaborations identified by this approach, the combined use of a variety of insertional mutagens and genetic backgrounds is recommended. In the future, the development of mutagens that can induce the formation of solid tumours should facilitate the identification of a larger repertoire of cancer gene candidates.

Chapter 4 Using mouse candidate cancer genes to narrow down the candidates in regions of copy number change in human cancers

4.1 Introduction

As discussed in Section 1.3.3, copy number changes are a common feature of cancer genomes, and can be identified using comparative genomic hybridisation (CGH)-based techniques. However, regions of copy number change are often large and encompass many genes, making it difficult to identify the "critical" genes that contribute to the tumourigenic process. Candidate cancer genes identified by insertional mutagenesis in the mouse can be used in a cross-species oncogenomics approach to narrow down the candidates within regions of copy number change in human tumours. The use of cross-species comparative analysis for cancer gene discovery is discussed in Section 1.5. In this chapter, mouse candidate cancer genes are used to identify orthologous candidates within regions of copy number change in 713 human cancer cell lines generated using SNP array CGH. The analyses were performed as part of a collaboration with the Netherlands Cancer Institute (NKI), published in Cell (Uren *et al.*, 2008), and therefore, rather than using the mouse candidate cancer genes generated from the work described in Chapters 2 and 3, lists of candidates were provided by the NKI.

The datasets are introduced in Section 4.2. This is followed, in Section 4.3, by a description of the methods used to process the copy number data into regions of copy number change, and gains and losses within the human cancer cell lines are characterised in Section 4.4. In Section 4.5.1, the mouse and human datasets are compared to determine whether retroviral insertional mutagenesis is relevant to the discovery of amplified and deleted cancer genes in humans. Promising cancer gene candidates that are both disrupted by insertional mutagenesis in the mouse and amplified or deleted in human cancers are presented in Section 4.5.2. A range of algorithms have been developed for identifying regions of copy number change within CGH data, and these are described and compared in Section 4.6. Finally, in Section 4.7, the mouse candidate cancer genes are combined with copy number variation (CNV) data from apparently healthy individuals to determine whether there is any overlap between candidates and regions of CNV.

Since the ploidy of the cell lines, and therefore the exact copy number of alterations, is difficult to establish, the terms “gain and “amplicon” are used interchangeably throughout this thesis to mean any gain of copy number, irrespective of the size or nature of the alteration.

4.2 Description of the datasets

As well as the datasets described below, the set of known cancer genes from the Cancer Gene Census (Futreal *et al.*, 2004) was also used. This is described in Section 2.2.3.

4.2.1 Mouse candidate cancer genes identified by retroviral insertional mutagenesis

As mentioned in the introduction, some of the work described in this chapter was undertaken as part of a collaboration with the NKI (Uren *et al.*, 2008, reprinted on p.365). The gene lists used in this chapter were therefore provided by the NKI but were generated from the analysis of insertion sites identified in the retroviral insertional mutagenesis screen described in Chapter 2. There were 6 lists of putative tumour suppressor genes. These included 3 lists comprising all genes in which there were insertions in the entire transcribed region, including UTRs and introns, only in the translated region (no UTRs) but including introns, and only in the coding region (no UTRs or introns). These lists are described throughout this thesis as genes in the transcribed region, translated region, and coding region, respectively. A further 3 lists contained genes with insertions in the same regions, but only where insertions comprised 2 or more sequence reads. Insertions represented by only 1 read are considered less likely to contribute to tumourigenesis (see Section 2.8) and are therefore predicted to have a reduced overlap with human deletions. 2 additional lists contained genes that were closest to CISs with *P*-values of less than 0.05 and 0.001, as determined using the kernel convolution (KC)-based statistical method (de Ridder *et al.*, 2006, see Sections 1.4.2.1.2 and 2.10.2). From these, lists were also generated for genes that were adjacent to CISs of *P*<0.05 and *P*<0.001 but were further away than the closest gene. For each gene list, the human orthologues and their genomic coordinates were extracted from Ensembl version 37 using Ensembl BioMart (see Section 3.2.1). Table 4.1 shows the number of mouse genes and human orthologues in each gene list. The *P*<0.001 and *P*<0.05 CISs and their associated nearest and further mouse genes

Gene List	Number of mouse genes	Number of mouse genes with human orthologues	Number of human orthologues in CIS gene list	% of human orthologues in CIS gene list
ORF only	266	240	41	17.1
ORF only (no singletons)	86	75	22	29.3
Translated region only	3024	2647	216	8.2
Translated region only (no singletons)	1331	1163	173	14.9
Transcribed region only	3773	3316	275	8.3
Transcribed region only (no singletons)	1706	1498	227	15.2
CIS nearest P<0.05	559	424	196	46.2
CIS nearest P<0.001	355	265	155	58.5
CIS further P<0.05	505	362	85	23.5
CIS further P<0.001	313	219	66	30.1

Table 4.1. Description of the lists of mouse candidate cancer genes used for comparison with human cancer copy number data. "[ORF, Translated region, Transcribed region] only" are lists of genes containing insertions only in the open reading frame, translated region (but including introns) or transcribed region, respectively. "no singletons" means that the list does not include genes that only contain insertions represented by a single read. "CIS nearest *P*<0.05" and "CIS nearest *P*<0.001" contain genes nearest to CISs identified by the kernel convolution (KC)-based method. "CIS further *P*<0.05" and "CIS further *P*<0.001" contain genes that flank CISs identified by the KC-based method but are not the nearest genes. The columns labelled "Number/% of human orthologues in CIS gene lists" show the overlap of each list with the list of candidate cancer genes generated and described in Chapters 2 and 3.

are listed in Appendix D. Due to their length, the lists of candidate tumour suppressor genes are not included, but are available on request.

Table 4.1 also shows the overlap of each gene list with the list of candidate cancer genes generated and described in Chapters 2 and 3 (shown in Appendix B2 and referred to here as the CIS gene list). The CIS gene list contains only genes that are associated with a significant CIS and this, together with the fact that the screen identifies mainly oncogenes, accounts for the small overlap with the tumour suppressor gene lists, in which genes may contain any number of insertions. The differences between the CIS gene list and the remaining lists may reflect differences in gene selection, i.e. a more sophisticated method was used to assign insertions to genes in the CIS gene list, and in read and insertion site processing, which were more conservative for the CIS gene list. Candidates from the CIS gene list are used in Chapter 5, where it is compared to higher resolution human CGH data (Section 5.3), as well as to the CGH data described in this chapter (Section 5.4).

4.2.2 Copy number data for human cancer cell lines

Comparative genomic hybridisation (CGH) data were generated by the Wellcome Trust Sanger Institute (WTSI) Cancer Genome Project for 713 human cancer cell lines from 29 tissues. A list of all cell lines and their tissue of origin is provided in Appendix E and is summarised in Table 4.2. None of the chosen cell lines had a common ancestor, according to cell line identity typing also performed by the WTSI Cancer Genome Project (http://www.sanger.ac.uk/genetics/CGP/Genotyping/synlinestable.shtml). This is important, since an amplicon or deletion might otherwise appear to be recurrent simply because it is within synonymous cell lines. CGH was performed using two Affymetrix GeneChip® Human Mapping Arrays. The 10K array, which comprises 11,555 SNPs, was used for 313 cell lines, while the 10K 2.0 array, comprising 10,204 SNPs, was used for the remaining 400 lines. 10,136 SNPs were shared between the two arrays, and both used the Affymetrix GeneChip® Mapping 10K assay, described in Section 1.3.3.2. The SNPs were mapped to the NCBI 35 human genome assembly. The mean distance between SNPs was 258.50 (±634.21) kb in the 10K array, and 292.82 (±683.49) kb in the 10K 2.0 array. The minimum distance was 2 bp and 11 bp for the 10K and 10K 2.0 arrays, respectively, and the maximum distance was 24.81 Mb for both arrays. 9.4% of human protein-coding genes in Ensembl v37 (extracted using Ensembl BioMart, see

Tissue of origin	**Number of cell lines**
Lung	131
Haematopoietic and lymphoid	117
Breast	43
Skin	42
Central nervous system	40
Unknown	39
Large intestine	38
Autonomic ganglia	29
Bone	23
Kidney	21
Soft tissue	20
Oesophagus	20
Stomach	19
Upper aerodigestive tract	19
Ovary	18
Pancreas	14
Urinary tract	13
Liver	11
Thyroid	11
Cervix	11
Endometrium	10
Biliary Tract	6
Pleura	5
Testis	3
Vulva	2
Prostate	2
Eye	2
Placenta	2
Adrenal gland	1
Small intestine	1
Total	713

Table 4.2. Tissues of origin of human cancer cell lines used in the 10K SNP array CGH analysis.

Section 3.2.1) contained at least one SNP in the 10K array, while 9.0% contained at least one SNP in the 10K 2.0 array. Genes were defined as the longest Ensembl gene transcript. The 10K and 10K 2.0 arrays contained an average of 0.176 (±0.735) and 0.157 (±0.648) SNPs per protein-coding gene, respectively. The interSNP distances and number of SNPs per gene are shown in Figures 4.1 and 4.2, respectively. The largest gaps between adjacent SNPs occur at the centromeres, while some gaps correspond to other regions of the genome that have not been assembled, e.g. due to highly repetitive sequences.

For each cell line. the raw intensity values were normalised internally. This involved calculating the value for each SNP as a total of all the SNPs on the array, and obtaining a copy number ratio for each SNP by dividing the SNP value by the value for the same SNP from a pool of reference normal samples. This is the point at which I received the data. The copy number data for all cell lines are available for download from ftp://ftp.sanger.ac.uk/pub/CGP/10kData. Data generated on the 10K and 10K 2.0 arrays is pooled in subsequent analyses and is collectively referred to as 10K data.

4.2.3 Copy number variants (CNVs)

CNVs are regions within the genome that vary in copy number. Germline CNV regions identified within 270 HapMap samples from Redon *et al.* (2006) were downloaded from http://www.sanger.ac.uk/humgen/cnv/data/cnv_data/. Merged CNVs identified using the Whole Genome Tilepath (WGTP) array and Affymetrix GeneChip Human Mapping 500K early access array (500K EA) were used. The WGTP array comprises 26,574 BAC clones, while the 500K EA array covers 474,642 SNPs. The WGTP and 500K EA platforms are complementary, since they are able to detect smaller and larger CNVs, respectively (Kehrer-Sawatzki, 2007). There are 1,447 merged CNVs that cover ~12% of the genome. 1,390 CNVs that mapped to autosomes in the NCBI 35 human build were used in this analysis.

4.3 Processing the copy number data

The copy number ratios at individual SNPs must be processed into regions of copy number change. As discussed in Section 1.3.3.2, a variety of methods have been

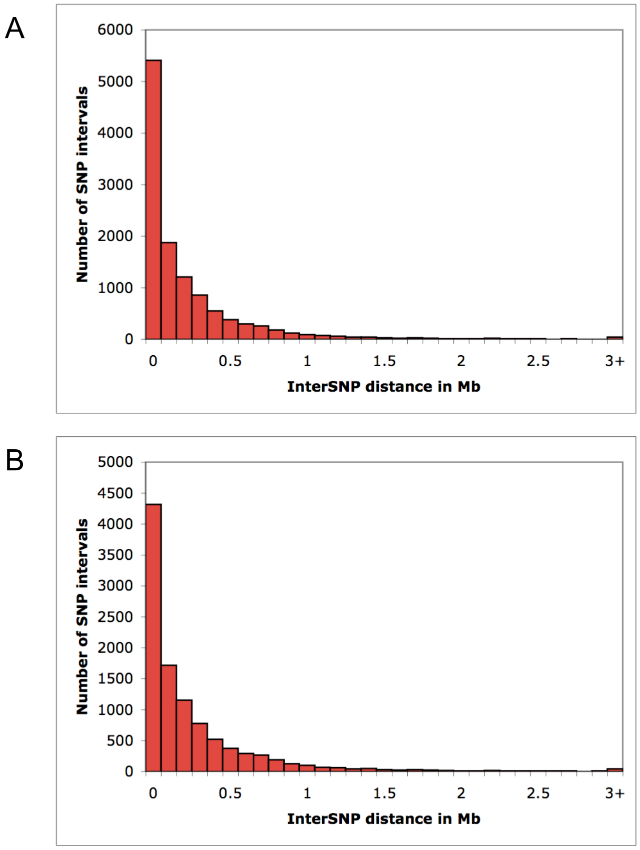

Figure 4.1. The distance between the genomic coordinates of adjacent SNPs on the 10K (A) and 10K 2.0 (B) SNP arrays.

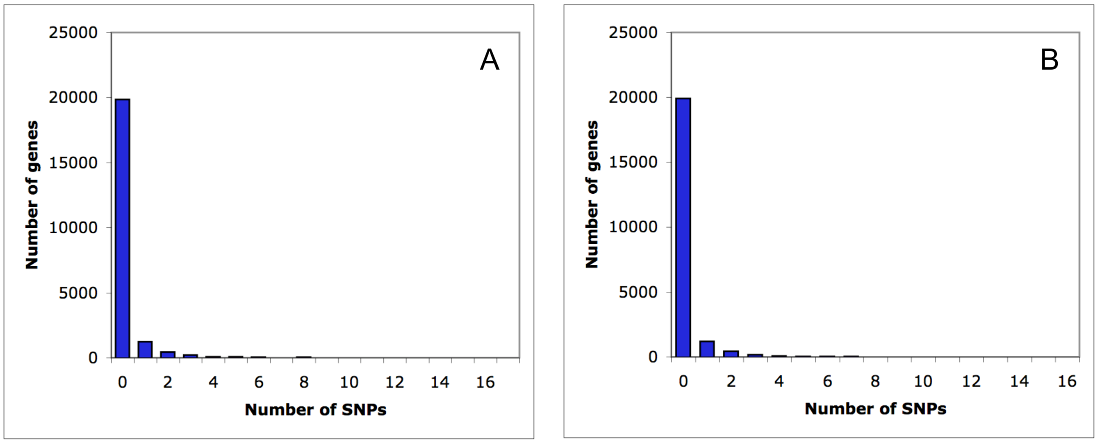

Figure 4.2. The number of SNPs per human protein-coding gene on the 10K (A) and 10K 2.0 (B) SNP arrays.

developed for this purpose. At the time of the analysis, most of the available algorithms had been developed primarily for conventional array CGH, i.e. using large genomic clones (see Section 1.3.3.2). In a comparison of 11 methods, DNAcopy (Olshen *et al.*, 2004) performed consistently well (Lai *et al.*, 2005), and a comparison of 3 segmentation methods by Willenbrock and Fridlyand (2005) demonstrated that DNAcopy performed better than GLAD (Hupe *et al.*, 2004) and HMM (Fridlyand *et al.*, 2004). A further benefit of DNAcopy is that it is freely available as an R package in BioConductor (http://www.bioconductor.org/). BioConductor is an open source software project that provides tools, mostly written in R, for analysing genomic data. DNAcopy (version 1.4.0) was therefore chosen as the method for detecting regions of copy number change in the 10K CGH data.

DNAcopy uses a method called circular binary segmentation (CBS) to identify change-points in CGH data, which is input as $\log_2$ intensity ratios at consecutive positions in the genome. The change-points correspond to positions in the genome where the DNA copy number has significantly changed. For each cell line, the copy number ratios for all SNPs were converted to $\log_2$-ratios and were smoothed, using a method within DNAcopy, to remove single point outliers before segmentation. Copy number ratios of 0 were given a $\log_2$-ratio of -6. Change-points may result from local trends in the data, and therefore all change-points that were less than 3 standard deviations apart were removed. Default parameters were used for the segmentation. Different values were tested for the parameter alpha but, upon visual inspection of the graphical outputs, the default value of alpha=0.01 appeared to be most suitable. Increasing alpha increases the sensitivity, resulting in more change-points but, potentially, more false positive change-points. Decreasing alpha results in fewer change-points, and regions of copy number change may therefore be missed. Increasing the number of standard deviations below which change-points were removed resulted in the loss of potentially important change-points. Figure 4.3 shows an example of how changing the parameters can affect the output of DNAcopy for chromosomes 1 and 6 of ovarian cancer cell line 41M-CISR. The removal of change-points less than 3 standard deviations apart results in the loss of a change-point in chromosome 1 (Figure 4.3B). However, the slight difference in copy number between the 2 arms of the chromosome may be due to trends in the data, and the difference in copy number is small. Increasing the number of standard deviations to 4 results in the loss of a change-point in chromosome 6, for which there is a clear step in copy number that does look real (Figure 4.3C). Increasing alpha from 0.01 to 0.05 results in the inclusion of

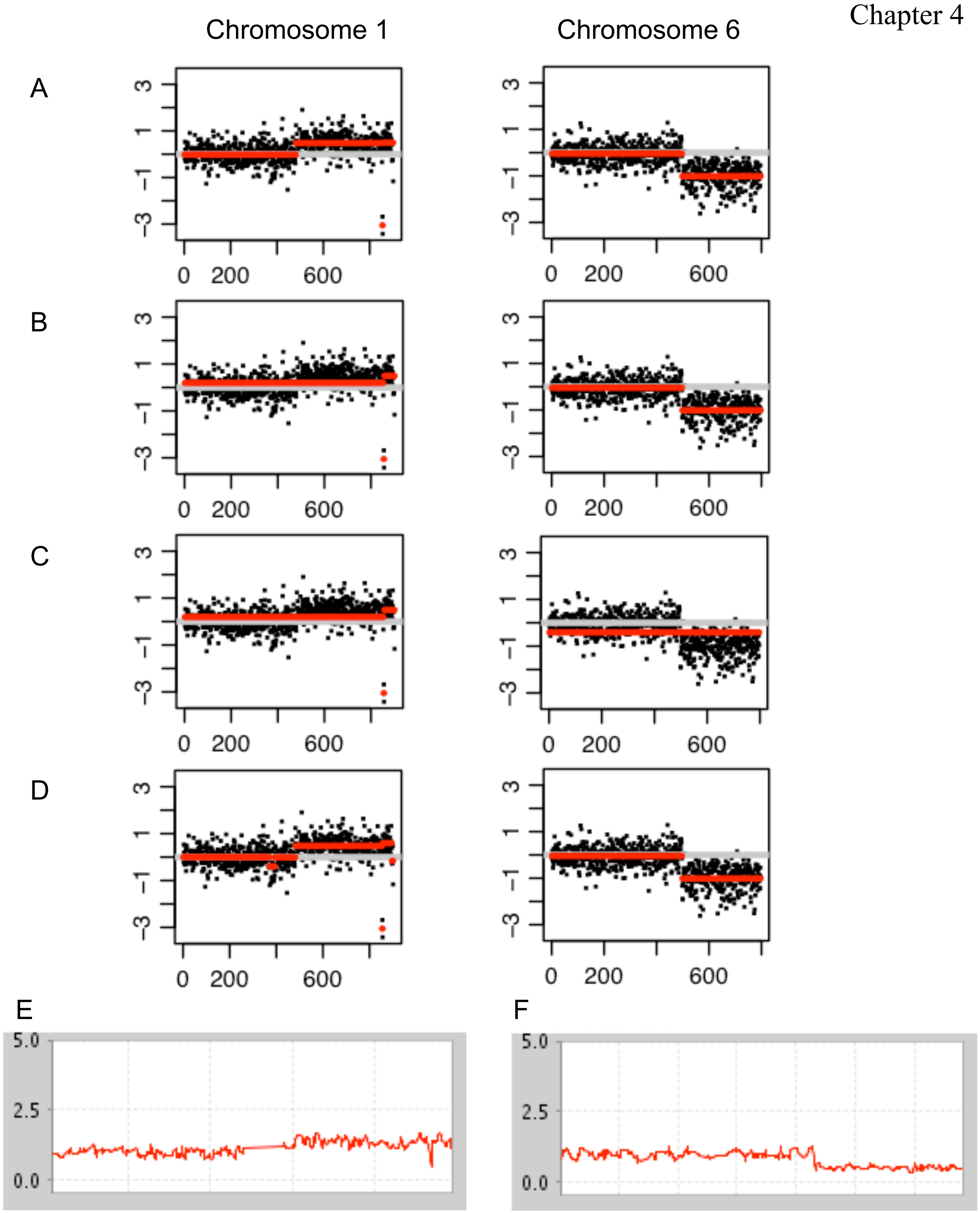

Figure 4.3. Altering the values for parameters in DNAcopy leads to differences in the regions of copy number change detected by the algorithm, as demonstrated for chromosomes 1 and 6 of ovarian cancer cell line 41M-CISR. (A) Default parameters. (B) Default parameters and removal of change-points less than 3 standard deviations apart. (C) Default parameters, smoothing and removal of change-points less than 4 standard deviations apart. (D) Alpha = 0.05 plus smoothing. (E) Copy number for chromosome 1, with values averaged across 3 consecutive SNPs. (F) Copy number for chromosome 6, with values averaged across 3 consecutive SNPs. Figures E and F are taken from the WTSI Cancer Genome Project website (http://www.sanger.ac.uk/genetics/CGP/) and give a clearer picture of the copy number across the chromosome. Figures A-D are extracted from the output of DNAcopy. Removing change-points that are close together results in fewer regions being detected, and the larger the number of standard deviations below which change-points are removed, the more regions are missed. Increasing alpha leads to the inclusion of additional change-points and, therefore, regions of copy number change.

additional change-points in chromosome 1 (Figure 4.3D). Since the data is relatively low resolution, it is highly possible that a region of copy number change may be represented by just 1 or 2 SNPs. However, it is also possible that such SNPs are anomalies and, to avoid the identification of false positives, this is the preferred assumption.

DNAcopy identifies changes in DNA copy number but does not indicate which regions are unchanged and which are gains or losses. It is therefore the responsibility of the user to set thresholds for calling gains and losses based on the mean $\log_2$-ratios of predicted segments. A disadvantage of DNAcopy is that it operates on individual chromosomes rather than the entire genome and the mean $\log_2$-ratios of segments representing no copy number change, or representing a gain or loss of a certain number of copies, will differ slightly across the genome. This makes it difficult to determine what is "normal" and therefore to call gains and losses, and the exact number of copies within a gain or loss cannot be clearly determined. Willenbrock and Fridlyand (2005) have developed an algorithm called MergeLevels that merges segments across the genome that are not significantly different from one another and so produces a more interpretable set of copy number levels. Combining DNAcopy and MergeLevels was found to be more effective than using DNAcopy alone (Willenbrock and Fridlyand, 2005). MergeLevels is freely available within an R/BioConductor package called aCGH. Therefore, for each cell line, the DNAcopy segmentation results were merged across all autosomes using MergeLevels with default parameters, which were considered appropriate upon inspection of the graphical outputs. Example outputs for the kidney cancer cell line 786-0 and endometrial cancer cell line AN3-CA are shown in Figure 4.4. The merged segments with a $\log_2$-ratio closest to 0 were defined as the level of no copy number change and, to enable comparison across cell lines, this $\log_2$-ratio was set to 0 and all other $\log_2$-ratios were normalised accordingly. Figure 4.5A shows the distribution of $\log_2$-ratios of the segments predicted by DNAcopy across all cell lines, while Figure 4.5B shows the distribution of $\log_2$-ratios of the merged segments. The $\log_2$-ratios of the merged segments show a series of peaks and troughs that may reflect distinct copy number levels. The large peak at -6 represents segments for which the copy number ratio of individual SNPs was 0. The $\log_2$-ratios of the merged segments were converted to copy number ratios (Figure 4.6), and troughs in the distribution were used to set thresholds for subsequent analyses (see Sections 4.4 and 4.5.1.1).

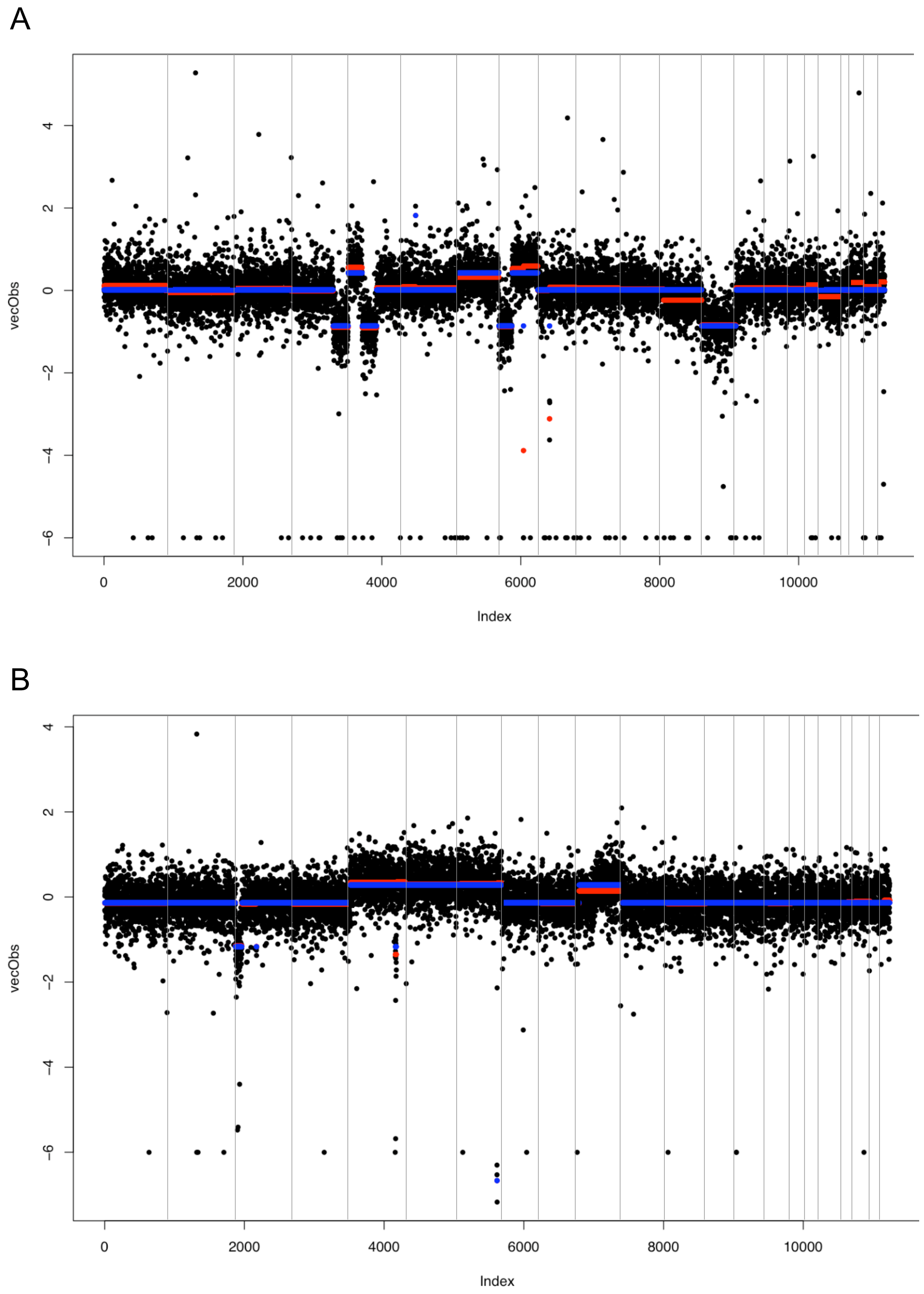

Figure 4.4. Graphical output from MergeLevels for human cancer cell lines 786-0 (A) and AN3-CA (B). Black dots represent the $\log_2$-ratios for individual SNPs ordered across the genome. The mean $\log_2$-ratios for segments identified by DNAcopy are shown in red. Merged segments generated by MergeLevels are shown in blue. Segments are merged across chromosomes into a set of copy number levels. The x-axis shows the position within the genome, while the y-axis measures the $\log_2$-ratio. Vertical lines represent the division of chromosomes.

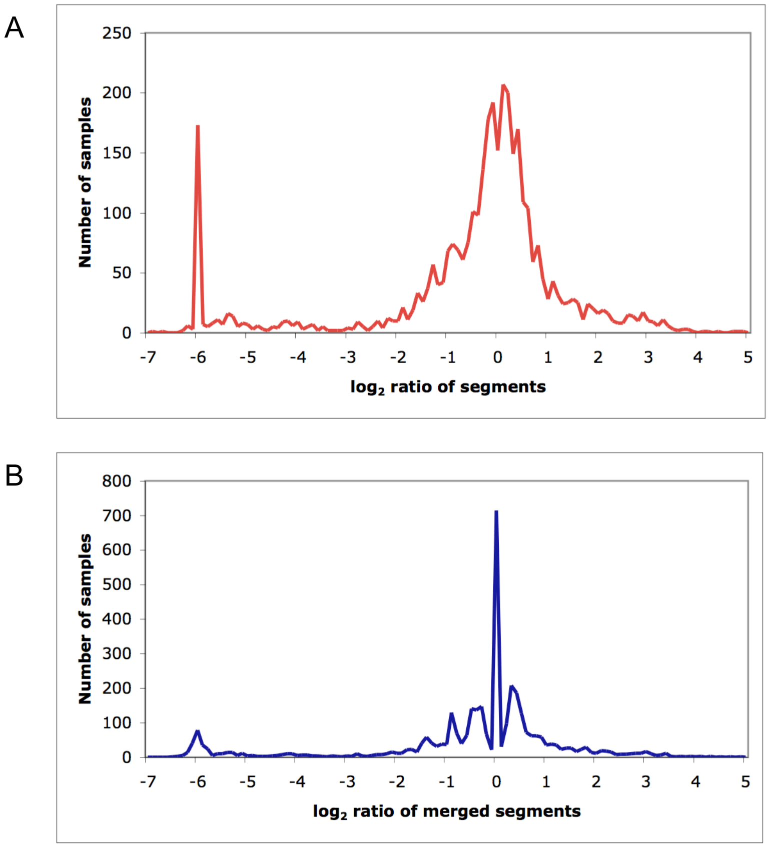

Figure 4.5. The number of human cancer cell lines with segments of varying log$_2$-ratio following processing with DNAcopy (A) and DNAcopy plus MergeLevels (B).

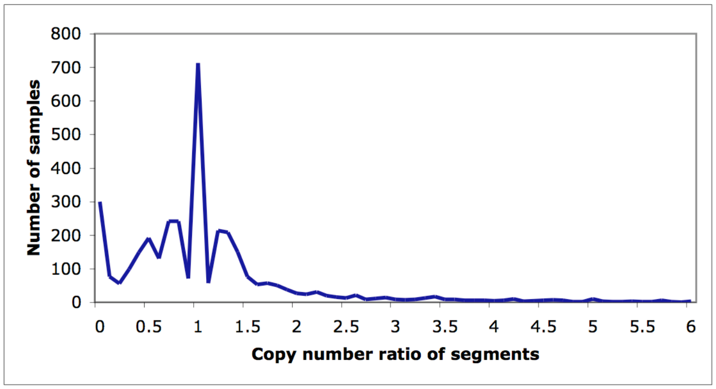

Figure 4.6. The number of human cancer cell lines with segments of varying copy number ratio following processing with DNAcopy plus MergeLevels. Troughs in the data were used to set thresholds for the analysis of gains and losses.

4.4 Characterising gains and losses in cancer genomes

All segments with a copy number of 1.6 or more were designated gains, and all segments with a copy number of 0.6 or less were designated losses. Segments with a copy number of 0.2 or less were designated homozygous deletions. Each of these thresholds was within a trough in the distribution of copy numbers for merged segments in Figure 4.6. The average number of gains per cancer cell line was 1.47 (±2.02), and the average size across all cell lines was 21.37 (±34.15) Mb. Amplicons contained an average of 213.68 (±341.45) genes. The average number of losses per cell line was 4.43 (±3.69) and the average size was 19.56 (±33.83) Mb, encompassing 195.62 (±338.32) genes. The average number of homozygous deletions was 1.53 (±1.99) per cell line. The average size was 0.98 (±2.68) Mb, encompassing 19.64 (±20.69) genes. Therefore, homozygous deletions were significantly smaller than amplicons and heterozygous deletions and contained fewer genes. Homozygous deletions have been previously shown to contain fewer genes than other regions of the genome (Cox *et al.*, 2005, see Section 1.3.3.3), and this analysis shows that, in general, the deletion of both copies of a gene is more likely to be deleterious to a cell than the loss of one copy or the gain of copies. The distributions of amplicon and deletion lengths are shown in Figure 4.7.

For each cancer type, the number of cell lines containing gains was counted, and the 2-tailed Fisher Exact Test was used to determine whether there was any significant difference between the observed and expected number of each cancer type. Cell lines derived from the oesophagus were over-represented (P=7.11x10^{-4}), suggesting that oesophageal tumours are particularly prone to genomic instability. Haematopoietic and lymphoid cancer cell lines were under-represented (P=5.32x10^{-5}), reflecting the fact that they often contain balanced translocations that do not show a change in DNA copy number (see Section 1.3.3.4) and that, in some cases, few genetic events are thought to be required for tumour development. For example, acute lymphoblastic leukaemias contain an average of 3.83 deletions and focal amplifications are rare (Mullighan *et al.*, 2007). Since most cell lines contained deletions, there was no significant difference between the observed and expected numbers of cancer types containing deletions.

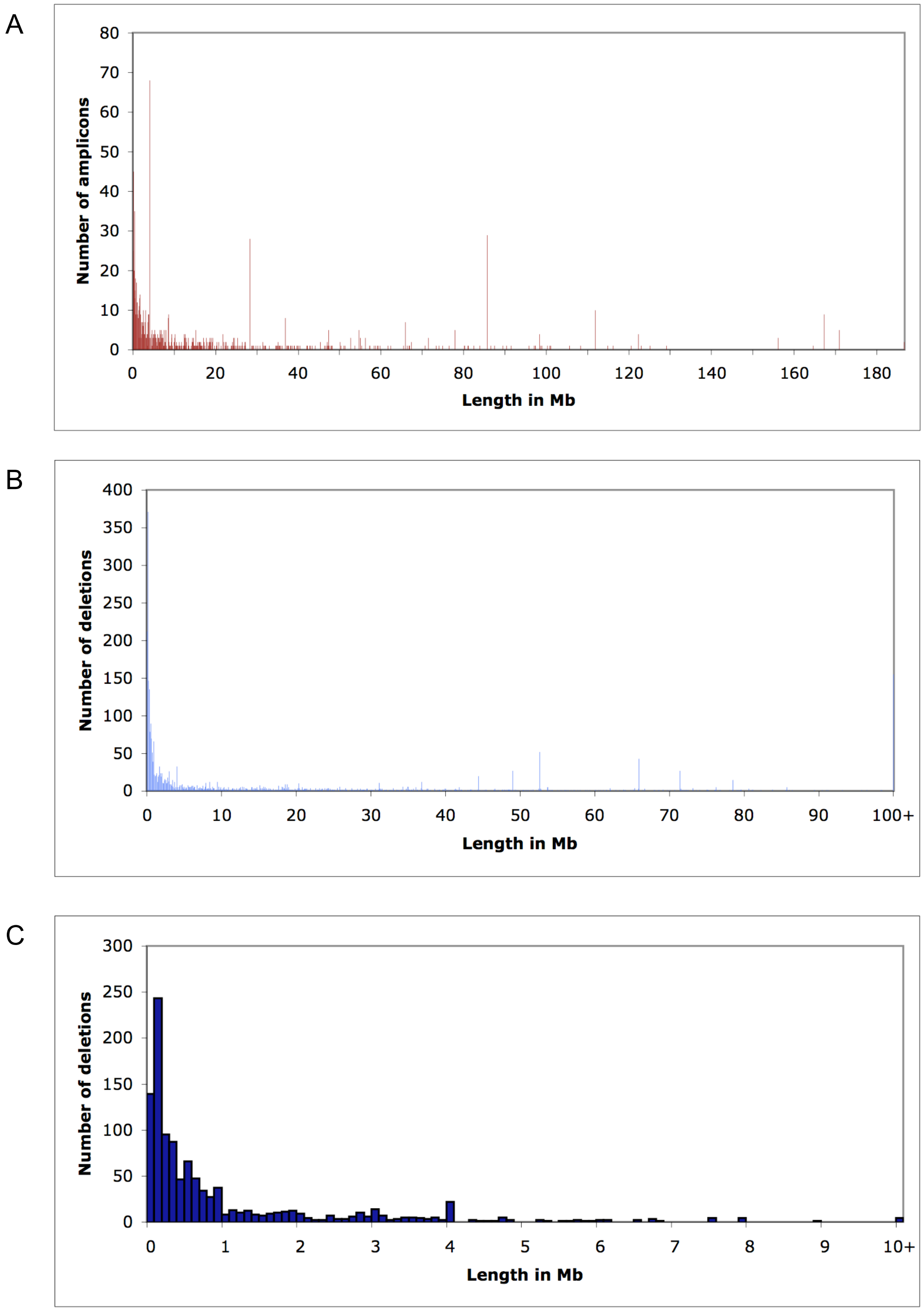

Figure 4.7. Distribution of the lengths of amplicons (A), deletions (B) and homozygous deletions (C) in 713 human cancer cell lines. Amplicons are defined as regions with a copy number greater than or equal to 1.6. Deletions and homozygous deletions are defined as regions with a copy number less than or equal to 0.6 and 0.2, respectively.

4.5 Comparative analysis of mouse candidate cancer genes and CGH data from human cancers

4.5.1 Global comparison

The purpose of the global comparison is to determine whether the human orthologues of candidate cancer genes identified by retroviral insertional mutagenesis by the Netherlands Cancer Institute are over-represented within regions of copy number change in the human cancer cell lines. Specifically, an over-representation of candidate oncogenes in human amplicons, and candidate tumour suppressor genes in human deletions, suggests that the retroviral insertional mutagenesis screen is relevant to human cancer, and may help to identify human cancer gene candidates within regions of copy number change.

4.5.1.1 Method

Rather than setting single copy number thresholds for gains and losses, a range of copy number thresholds were investigated. Thresholds were set as the centre-point of troughs in the graph shown in Figure 4.6, since these may represent transitions in the number of gene copies. The chosen thresholds were copy number ratios of less than or equal to 0.9, 0.6 and 0.2, and greater than or equal to 1.1, 1.6, 2.1, 2.5, 2.7, 3.1, 3.8, 4.3, 4.8, 5.2 and 5.9. The genomic coordinates of the human orthologues of all mouse genes were extracted from Ensembl version 37 using Ensembl BioMart. For each gene list described in Section 4.2.1, the number of mouse genes with human orthologues was counted. The same number of genes was selected randomly from among all mouse genes with human orthologues and this was repeated 1,000 times. For each of the 1,000 iterations, the number of human orthologues that resided within human cancer cell line segments with a mean copy number above or below a given threshold was counted. This produced a normal distribution of counts. The number of human orthologues of mouse candidate cancer genes in segments above or below each threshold was also counted. The Z-test was used to calculate the probability of obtaining a number greater than or equal to the observed count for the mouse candidates, based on the distribution of counts for the randomised genes. The procedure is summarised in Figure 4.8.

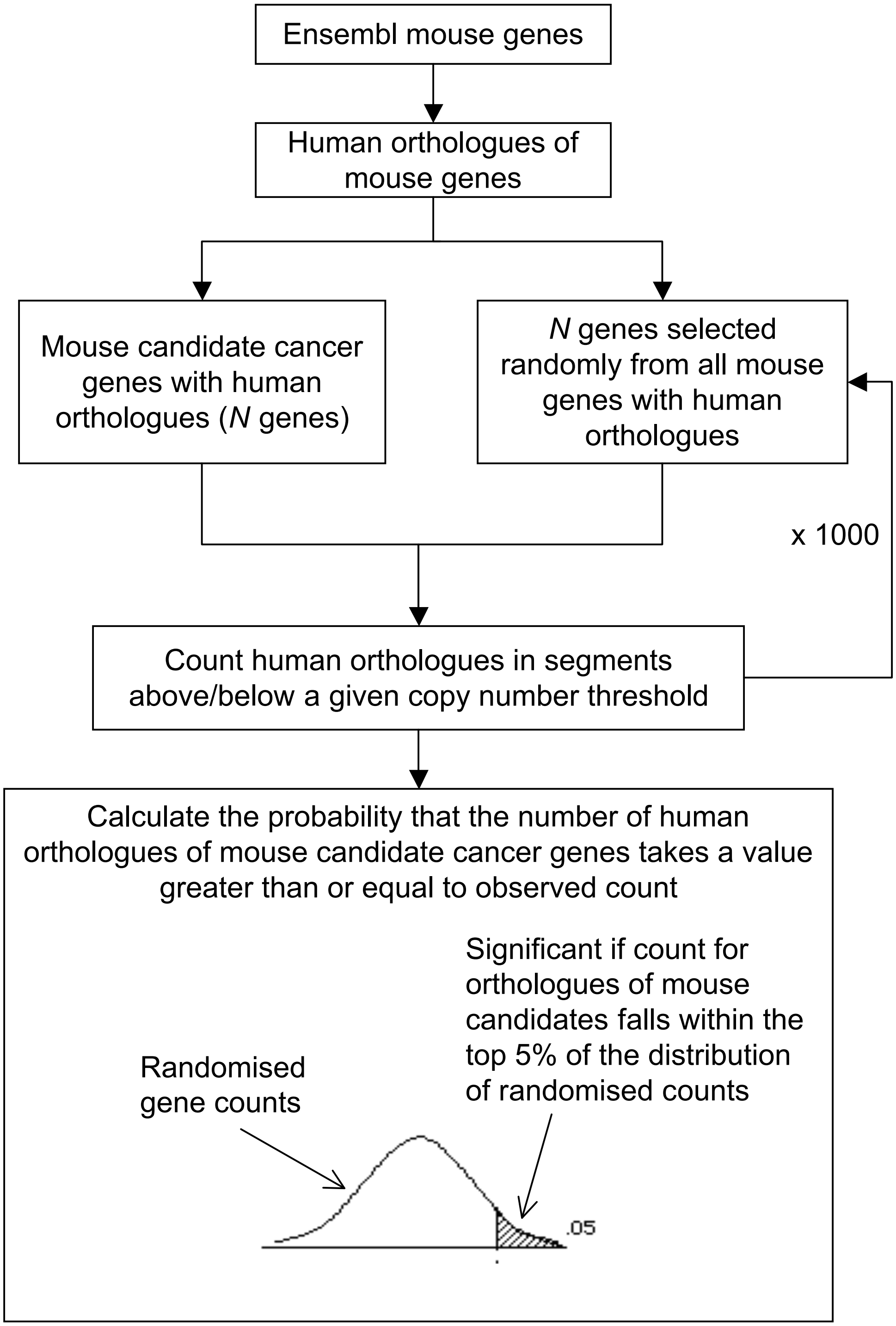
Ensembl mouse genes
Human orthologues of mouse genes
Mouse candidate cancer genes with human orthologues (*N* genes)
N genes selected randomly from all mouse genes with human orthologues
x 1000
Count human orthologues in segments above/below a given copy number threshold
Calculate the probability that the number of human orthologues of mouse candidate cancer genes takes a value greater than or equal to observed count
Randomised gene counts
Significant if count for orthologues of mouse candidates falls within the top 5% of the distribution of randomised counts
.05

Figure 4.8. Overview of the method for identifying over-representation of the human orthologues of mouse candidate cancer genes in regions of human copy number change.

4.5.1.2 Setting the boundaries of amplicons and deletions

The start and end coordinates of the copy number segments generated by DNAcopy and MergeLevels correspond to the first and last SNPs for which the $\log_2$-ratios are not significantly different to other SNPs in the segment. Therefore, the copy number can be determined for all coordinates between these positions. It is, however, impossible to determine the copy number for coordinates within the interval between the first SNP and the preceding SNP, which corresponds to the end coordinate of the preceding segment, and between the last SNP and the proceeding SNP, which corresponds to the start coordinate of the proceeding segment. As shown in Section 4.2.2, the distance between SNPs can be very large, especially across unassembled regions of the genome such as centromeres. Setting the boundaries of an amplicon or deletion as the end of the previous segment and start of the next segment, or even using half-way points, could therefore result in a very high number of false positives among genes predicted to be amplified or deleted.

In order to choose an appropriate distance for the boundaries of amplicons and deletions, the global comparison was performed using a range of distances. Assuming that CIS genes are more likely to be amplified or deleted in human cancers than are other genes, the most appropriate distance should be that which gives the highest over-representation of CIS genes. The list of genes nearest to CISs with $P<0.001$ was used in this analysis. Amplicon boundaries were extended beyond the first and last amplified SNP by a distance of 0 kb, 200 kb, 500 kb, 1 Mb, 3 Mb and 5 Mb, or as far as the adjacent SNP, whichever was closer. The results are shown in Figure 4.9. The association between CIS genes and amplicons was strongest when the boundaries were not extended at all. However, at lower copy numbers and in greater numbers of cell lines at higher copy number, the association was less significant than when the boundaries were extended to 500 kb. Extending the boundaries to 1 Mb and beyond resulted in a considerable decrease in the association between CIS genes and amplicons. Therefore, 500 kb was chosen as the most suitable distance.

Known oncogenes from among the CIS genes that were identified within full-length amplicons (i.e. where the amplicon was extended as far as the adjacent, non-amplified SNPs) were compared to those identified within amplicons with a 0 kb or 500 kb extension of the amplicon boundaries (Table 4.3). While the non-extended amplicons

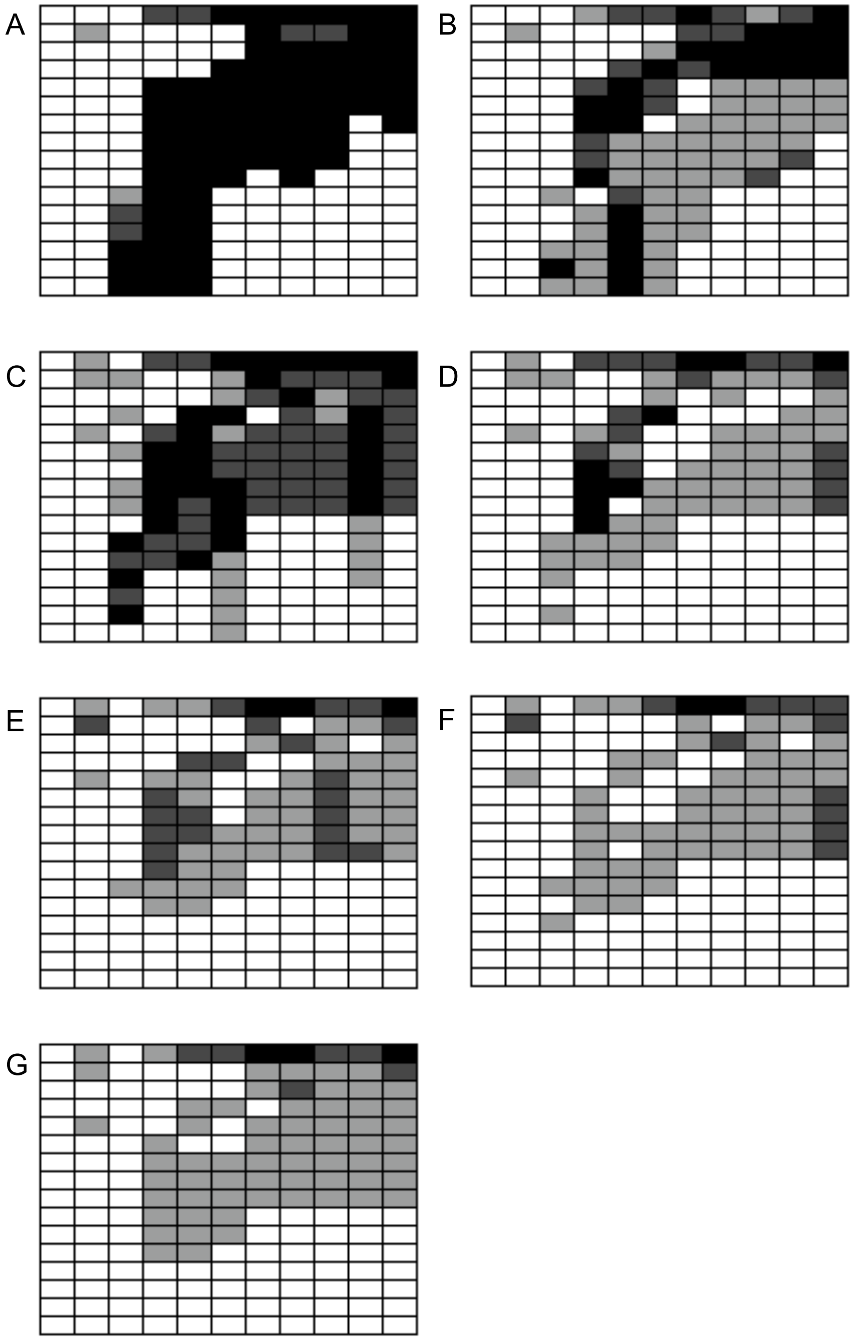

Figure 4.9. Over-representation of human orthologues of genes nearest to CISs in amplicons with boundaries extended beyond the first and last amplified SNP by a maximum distance of 0 kb (A), 200 kb (B), 500 kb (C), 1 Mb (D), 3 Mb (E), 5 Mb (F) and up to the adjacent, non-amplified SNPs (G). Each box represents the significance of the association between the selected genes and amplicons/deletions at a given copy number threshold and cell line number. $P<0.0001$, black; $P<0.001$, dark grey, $P<0.05$, light grey. Columns represent amplicons, with (from left to right) copy number thresholds of greater than 1.1, 1.6, 2.1, 2.5, 2.7, 3.1, 3.8, 4, 4.8, 5.2 and 5.9. Rows represent the number of cell lines, which increases in increments of 1, up to a cut-off of 16 cell lines. For example, the box in the bottom right-hand corner of each figure represents the *P*-value for the over-representation of CIS genes that occur in amplicons of copy number greater than or equal to 5.9 in at least 16 cancer cell lines.

Copy number	Gene	Full length	0 kb	500 kb
4.8	*MYC*	14	10	14
	MYCN	9	4	9
	CCND1	1	1	1
4.3	*MYC*	14	10	14
	MYCN	9	4	9
	CCND1	1	1	1
	ZFPN1A1	1	1	1
	LMO2	1	1	1
	CCND3	2	0	2
2.7	*MYC*	29	22	29
	MYCN	12	7	12
	CCND1	4	4	4
	ZFPN1A1	1	1	1
	LMO2	2	2	1
	CCND3	3	0	2
	CCND2	1	1	1
	PIM1	2	1	2
	EVI1	1	1	1
	IRF4	1	1	1

Table 4.3. The number of amplicons in which known cancer genes among genes nearest to CISs are identified when the amplicon boundaries are altered. "Full length" applies to amplicons extended to the next, non-amplified SNP. "0 kb" applies to amplicons where the start and end correspond to the first and last amplified SNP. "500 kb" applies to amplicons extended to a maximum of 500 kb. Copy number values are given as the minimum copy number of amplicons.

missed some of the occurrences of amplified *MYC* and *MYCN* that were identified in the full-length amplicons, all occurrences were identified in the amplicons extended by up to 500 kb. Likewise, occurrences of amplified *CCND3* and *PIM1* were identified in the 500 kb amplicons but not the non-extended amplicons.

To demonstrate that the observed association between candidate cancer genes and human amplicons was real, full-length amplicons were shuffled across the genome. The length and mean copy number of each amplicon were conserved, but the location was shuffled. The method from Section 4.5.1.1 was then performed on the shuffled amplicons. As shown in Figure 4.10, the association of candidates with regions of copy number gain was completely abolished.

4.5.1.3 Comparison with lists of candidate cancer genes

Having chosen 500 kb as the maximum distance for extending amplicon and deletion boundaries, the method of Section 4.5.1.1 was applied to all of the gene lists outlined in Section 4.2.1. The results are shown in Figure 4.11. The lists of genes nearest to CISs with *P*<0.001 or *P*<0.05 are lists of candidate oncogenes, with those nearest to CISs with *P*<0.001 being stronger candidates for a role in tumourigenesis. This is reflected in the results, since both lists showed an over-representation of candidates within regions of amplification, but the association was stronger for genes near to a CIS with *P*<0.001. For both gene lists, the association became significant at copy number 1.6 and above, but for low-level copy numbers, the association was generally strongest for genes that were amplified in higher numbers of cell lines. Figure 4.12A shows the over-representation of known oncogenes within regions of copy number gain. The pattern of association was very similar to that obtained using genes nearest to CISs with *P*<0.001, suggesting that this list contains oncogenes that are relevant to human cancer. Almost all of the mouse tumours generated in the retroviral screen were lymphomas, and therefore it could be assumed that the candidate cancer genes identified in the screen are only relevant to similar cancers within humans. Therefore, the human cancer cell lines were divided into haematopoietic and lymphoid cell lines and all other cell lines (from solid tumours) and the global comparison was performed on each subset using genes nearest to CISs with *P*<0.001. As shown in Figure 4.13, the association was much weaker when only haematopoietic and lymphoid cell lines were considered. This may be partly because amplification is not a common mechanism of mutation in these cell types (see Section

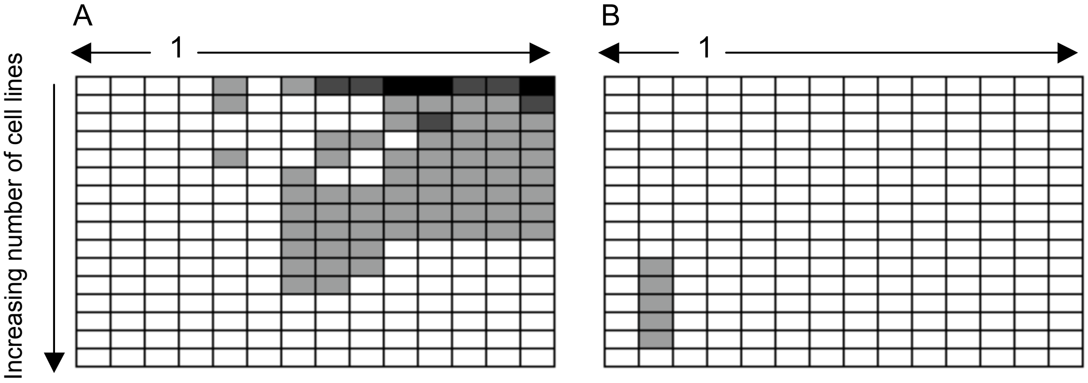

Figure 4.10. Over-representation of human orthologues of genes nearest to CISs in full-length human amplicons (A) and shuffled full-length amplicons (B). Each box represents the significance of the association between the selected genes and amplicons/deletions at a given copy number threshold and cell line number. *P*<0.0001, black; *P*<0.001, dark grey, *P*<0.05, light grey. Copy number thresholds below 1 represent deletions, with (from left to right) copy number thresholds of less than 0.2, 0.6 and 0.9. Copy number thresholds above 1 represent amplicons, with (from left to right) copy number thresholds of greater than 1.1, 1.6, 2.1, 2.5, 2.7, 3.1, 3.8, 4.3, 4.8, 5.2 and 5.9. The number of cell lines increases in increments of 1, up to a cut-off of 16 cell lines.

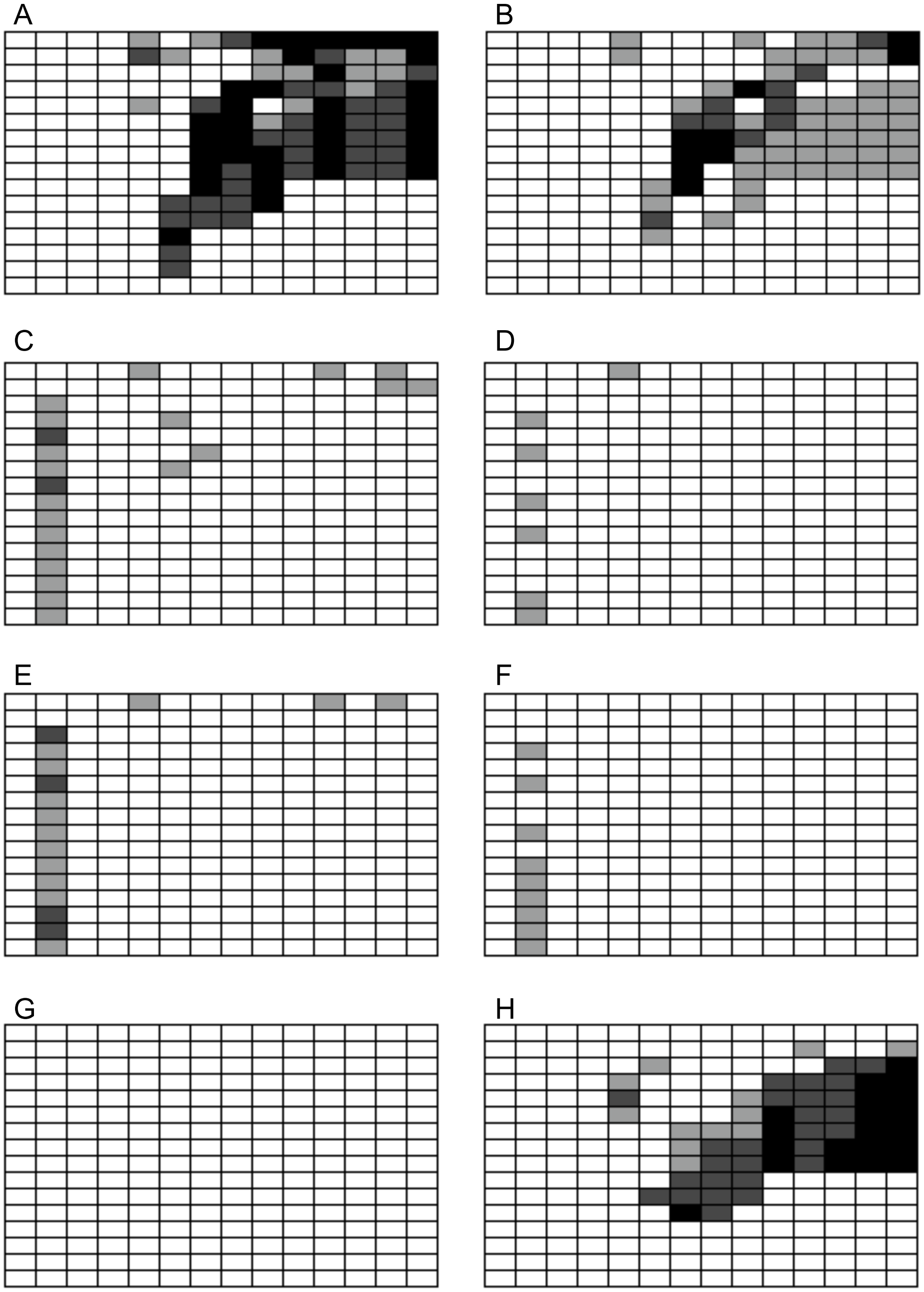

Figure 4.11. Over-representation of human orthologues of candidate cancer genes in regions of copy number change. (A) Genes nearest to CISs with *P*<0.001. (B) Genes nearest to CISs with *P*<0.05. (C) Genes with insertions within the transcribed region. (D) Genes with insertions but no singletons in the transcribed region. (E) Genes with insertions within the translated region. (F) Genes with insertions but no singletons in the translated region. (G) Genes with insertions in the coding region. (H) Genes with insertions but no singletons in the coding region. Each box represents the significance of the association between the selected genes and amplicons/deletions at a given copy number threshold and cell line number. *P*<0.0001, black; *P*<0.001, dark grey, *P*<0.05, light grey. Columns from left to right represent copy number thresholds of less than 0.2, 0.6 and 0.9 (deletions) and greater than 1.1, 1.6, 2.1, 2.5, 2.7, 3.1, 3.8, 4.3, 4.8, 5.2 and 5.9 (amplicons). The number of cell lines increases in increments of 1, up to a cut-off of 16 cell lines.

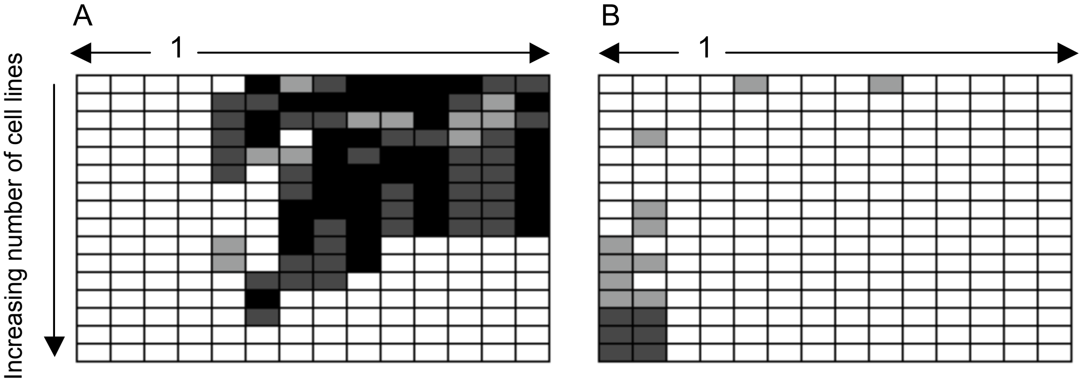

Figure 4.12. Over-representation of known oncogenes (A) and known tumour suppressor genes (B) in regions of copy number change in human cancer cell lines. Each box represents the significance of the association between the selected genes and amplicons/deletions at a given copy number threshold and cell line number. *P*<0.0001, black; *P*<0.001, dark grey, *P*<0.05, light grey. Copy number thresholds below 1 represent deletions, with (from left to right) copy number thresholds of less than 0.2, 0.6 and 0.9. Copy number thresholds above 1 represent amplicons, with (from left to right) copy number thresholds of greater than 1.1, 1.6, 2.1, 2.5, 2.7, 3.1, 3.8, 4.3, 4.8, 5.2 and 5.9. The number of cell lines increases in increments of 1, up to a cut-off of 16 cell lines.

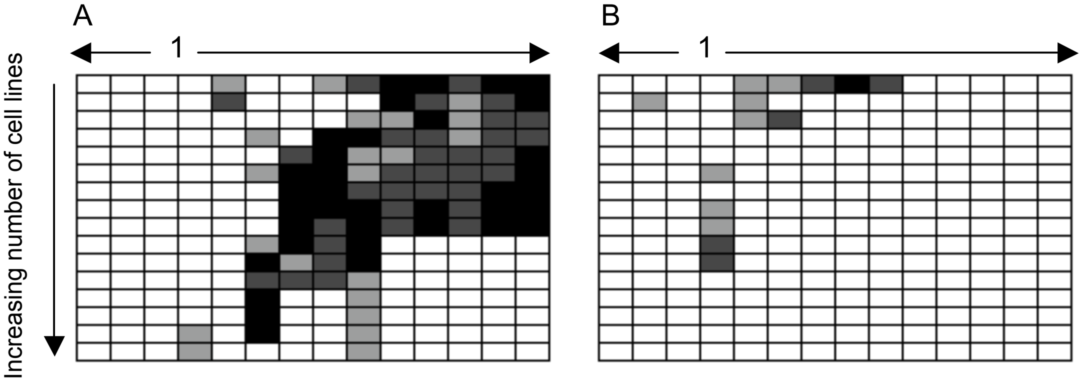

Figure 4.13. Over-representation of human orthologues of genes nearest to CISs with a *P*-value of <0.001 in regions of copy number change in human cancer cell lines derived from solid tumours (A) and haematopoietic and lymphoid cancers (B). See Figure 4.13 above for a description of how to interpret the figures.

4.4), but may also reflect the fact that the set of cell lines is smaller and therefore the comparison lacks power. Importantly, the pattern of association in solid tumours was similar to that for all cell lines and was highly significant. This demonstrates the relevance of retroviral insertional mutagenesis to the discovery of cancer genes in diverse human cancers, and shows that analysis of the full set of human cancer cell lines is warranted. Each cancer type provided in Table 4.2 was then separately tested for an association with the candidate cancer genes. Splitting the cancer cell lines into different types reduces the power of the analysis, and for most tumour types there was no clear association. However, cell lines derived from the autonomic ganglia, breast, upper aerodigestive tract, large intestine, oesophagus and stomach did show a significant overlap between mouse candidates and regions of copy number gain, although in the large intestine cell lines, there was also a significant overlap with regions of copy number loss (Figure 4.14).

The remaining lists are expected to contain candidate tumour suppressor genes. The results were similar for genes with insertions in transcribed and translated regions (Figure 4.11C-F). In both cases, including all genes containing insertions, rather than just those containing insertions represented by more than one read, generated a more significant association. This suggests that insertions represented by a single read ("singletons") in this retroviral screen are often important in tumourigenesis. As discussed in Section 2.7, the screen is not fully saturated due to the use of an insufficient number of enzymes in PCR and insufficient sequencing depth. Therefore, singleton insertions may result from these limitations, rather than because they are rare in the tumour mixture. However, for genes with insertions in the coding region, the reverse was observed, with a significant association only occurring when singleton insertions were omitted (Figure 4.11G-H, see below). As expected for tumour suppressor genes, the lists of genes with insertions in the transcribed and/or translated region were associated with deletions of copy number less than or equal to 0.6. However, the significance of the association was weak. When singleton insertions were included, there was also evidence of a weak association with regions of copy number gain. This is not surprising since the lists are likely to be contaminated with candidate oncogenes, as well as genes that do not play a role in tumourigenesis. The gene lists are long and yet tumour suppressor genes are less likely to be identified by insertional mutagenesis than are oncogenes and, as shown in Chapter 3, oncogenes are often disrupted by intragenic insertions. The association between known tumour suppressor genes and regions of copy number change is shown in Figure 4.12B.

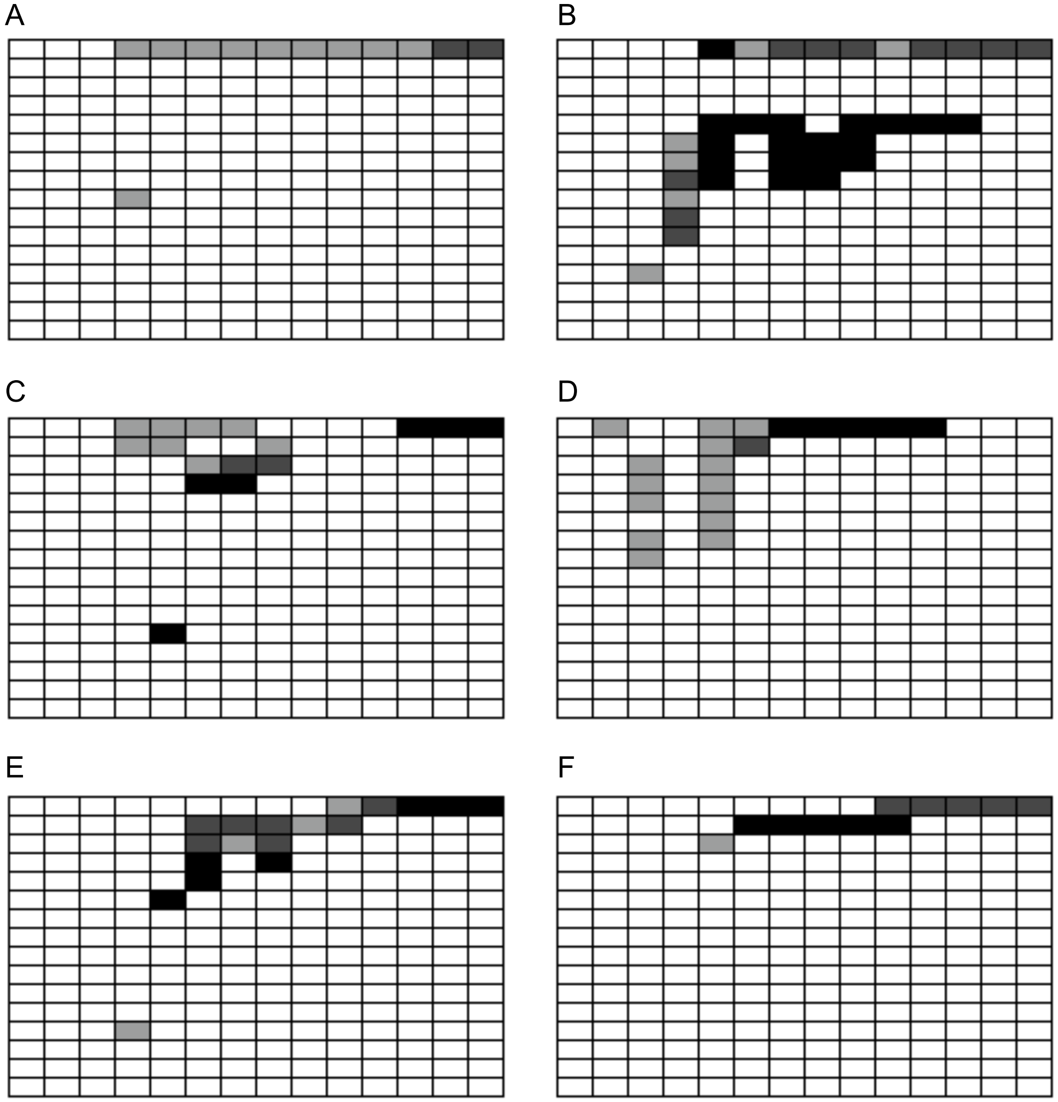

Figure 4.14. Over-representation of human orthologues of candidate cancer genes in regions of copy number change in cancer cell lines derived from the upper aerodigestive tract (A), autonomic ganglia (B), breast (C), large intestine (D), oesophagus (E) and stomach (F). Each box represents the significance of the association between the selected genes and amplicons/deletions at a given copy number threshold and cell line number. *P*<0.0001, black; *P*<0.001, dark grey, *P*<0.05, light grey. Columns from left to right represent copy number thresholds of less than 0.2, 0.6 and 0.9 (deletions) and greater than 1.1, 1.6, 2.1, 2.5, 2.7, 3.1, 3.8, 4.3, 4.8, 5.2 and 5.9 (amplicons). The number of cell lines increases in increments of 1, up to a cut-off of 16 cell lines.

There was a more significant over-representation of genes within deletions of copy number less than or equal to 0.6, but also in deletions of copy number less than or equal to 0.2. However, the fact that the pattern was broadly similar for genes with insertions in transcribed and translated regions is encouraging, and suggests that the lists do contain tumour suppressor genes that are relevant to human cancer.

The results for genes containing insertions within the coding region did not show the expected pattern for tumour suppressor genes (Figure 4.11G-H). As mentioned above, when singleton insertions were included, a significant association was not observed with either amplicons or deletions. Omission of singleton insertions resulted in a pattern of association representative of oncogenes, i.e. showing an over-representation of genes within amplicons. The identities of genes that reside within human amplicons and deletions are provided in Section 4.5.2.

4.5.1.4 Determining whether the nearest gene to a CIS is the most likely candidate cancer gene

As discussed in Chapter 2, it can be difficult to determine which gene is being mutated by insertions within a CIS, especially when the insertions are intergenic and disrupt genes by enhancer mutation. CISs are often assigned to the nearest gene. Therefore, to test whether this is a sensible assumption, the overlap of the human CGH data with candidate genes nearest to CISs was compared to that observed for the next nearest genes to CISs. The method was performed as described in Section 4.5.1.1, whereby the number of genes closest to CISs that occurred within amplicons or deletions was compared to the number of randomly occurring genes in amplicons or deletions. This was then repeated for genes adjacent to, but further from, CISs. Thresholds in this analysis were the same as for previous comparisons. The method was performed on CISs with a *P*-value of <0.05 and <0.001, and the results are shown in Figure 4.15. As previously shown, there was a more significant over-representation within human amplicons of genes nearest to CISs with *P*<0.001 than *P*<0.05. However, for both significance levels, the clear overlap between human amplicons and genes nearest to CISs was almost absent for genes further from CISs. This suggests that the nearest gene to a CIS is generally the disrupted gene. *Plekhf1* and *Ltap* (also known as *Vangl2*) were the only two genes in the set of genes that are further from the CIS for which the human orthologues were amplified to a copy number greater than or equal to 5.2. However, in both cases, the nearest gene to the CIS

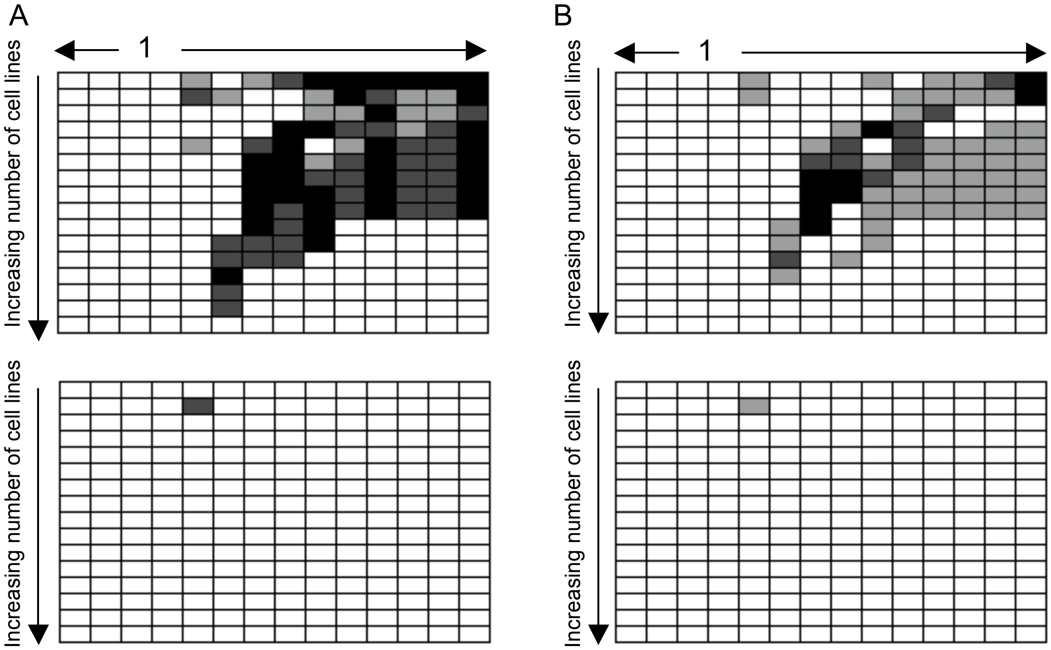

Figure 4.15. Over-representation of human orthologues of genes nearest to CISs (above) and genes further from CISs (below) in amplicons and deletions, where CISs have a *P*-value of <0.001 (A) and <0.05 (B). Each box represents the significance of the association between the selected genes and amplicons/deletions at a given copy number threshold and cell line number. *P*<0.0001, black; *P*<0.001, dark grey, *P*<0.05, light grey. Copy number thresholds below 1 represent deletions, with (from left to right) copy number thresholds of less than 0.2, 0.6 and 0.9. Copy number thresholds above 1 represent amplicons, with (from left to right) copy number thresholds of greater than 1.1, 1.6, 2.1, 2.5, 2.7, 3.1, 3.8, 4.3, 4.8, 5.2 and 5.9. The number of cell lines increases in increments of 1, up to a cut-off of 16 cell lines.

(*1600014C10Rik* and *Slamf6*, respectively) was also amplified, and analysis of the insertions around these genes suggests that the nearest gene is more likely to be disrupted by MuLV (Figure 4.16). *PLEKHF1* and *LTAP* may therefore be non-tumourigenic passengers within the amplified regions. *Tpcn2* and *Ccnd1* are neighbouring genes that both have nearby CISs, and both human orthologues were amplified, suggesting that both may be involved in tumourigenesis. For the remaining 9 genes nearest to CISs that were amplified to a copy number greater than or equal to 5.2, a human orthologue could not be found for the further gene. This explains why the lists of human orthologues of nearest genes are longer than the lists of human orthologues of further genes (see Table 4.1). Therefore, in some cases, the further gene may have an unidentified human orthologue that is also amplified in cancer. However, the fact that the nearest gene list contains a higher proportion of genes with human orthologues is itself significant, since cancer and cancer-related functions, such as cell growth, are well-studied and implicated genes may therefore be more likely to be characterised than genes with other functions, and such genes are also likely to be conserved between species.

To investigate whether the difference between comparisons of nearest and further genes is most likely to be due to the further gene not being amplified or not having a human orthologue, the 66 human orthologues nearest to CISs that were amplified to a copy number greater than or equal to 2.7 were analysed in greater detail. 10 of the amplified genes, including *TPCN2* and *CCND1*, were neighbouring genes for which both mouse orthologues had nearby CISs. For 27 genes, the human orthologue of the further gene could not be identified. For a further 27 genes, the further gene was also amplified and, in all cases, both the nearest and the further genes were amplified in the same number of cell lines. There were only 2 amplified nearest genes, *Slc9a8* and *Mafk*, for which the further gene, *B4galt5* and *1110015K06Rik* respectively, had a human orthologue that was not amplified, suggesting that the nearest genes are the likely candidate cancer genes. In both cases, the human and mouse regions containing these genes are syntenic, and thus the lack of amplification is not due to a break in synteny in the human genome.

The reciprocal analysis was also performed, whereby the 36 human orthologues further from CISs that were amplified to copy number 2.7 or above were also investigated. 29 had neighbouring, nearer genes that were also amplified. This number is higher than the reciprocal count of 27 genes because 2 of the genes were adjacent to nearer genes that had more than one adjacent gene because they contained multiple CISs. For the 7 remaining

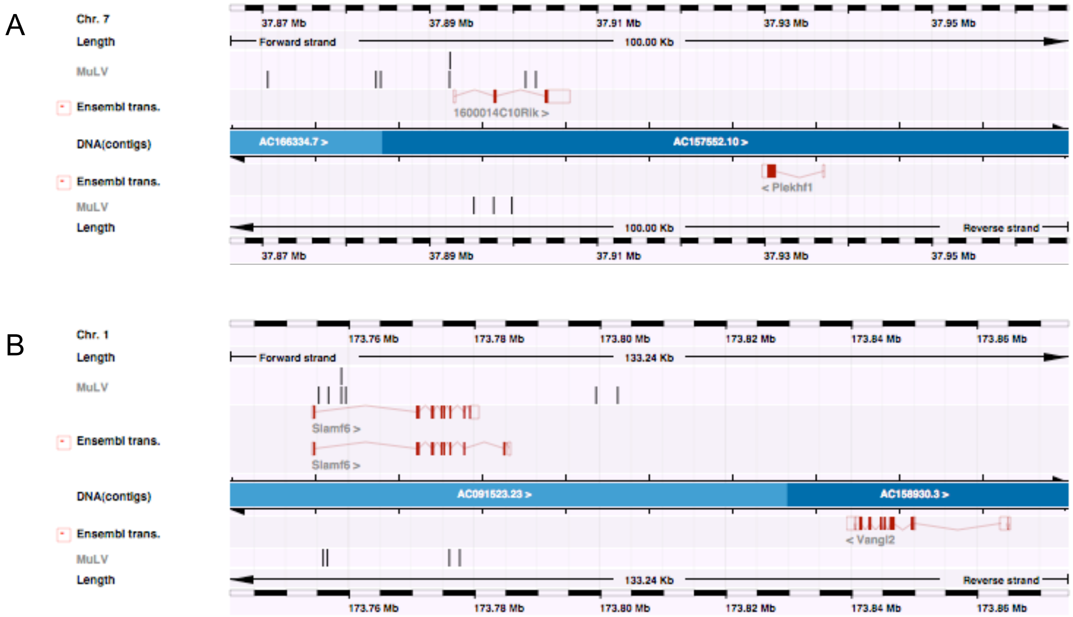

Figure 4.16. Insertions appear to be associated with the gene nearest to the CIS, i.e. *1600014C10Rik* (A) and *Slamf6* (B), even though adjacent genes are also amplified. Insertions are shown as black vertical lines. Those above the blue bar labelled DNA(contigs) are in the sense orientation, those below are in the antisense orientation. Ensembl genes are shown in red.

genes, there was no human orthologue for the nearest gene to the CIS. Therefore, it appears that the stronger overlap between the human amplicons and the human orthologues of genes nearest to CISs mainly reflects the inability to identify human orthologues for a higher percentage of the genes further from CISs. Since amplicons are generally large and encompass many genes (see Section 4.4), this is the more sensible explanation. However, several genes nearest to CISs were amplified in the absence of the further gene, while there were no examples of the reciprocal association. Choosing the nearest gene is the simplest method for assigning insertions to genes and will correctly identify genes that contain intragenic CISs. There is no evidence to suggest that genes nearest to CISs are preferentially amplified, but the fact that they are more likely to have a human orthologue does indicate that they may be the more likely candidates for a role in cancer. However, a more sophisticated method for assigning insertions to genes, such as the approach described in Sections 2.9 and 2.10, is likely to yield the most reliable list of cancer gene candidates. Comparative analyses involving these genes are discussed in Chapter 5.

4.5.2 Identification of individual candidates for a role in human cancer

For each gene list, the amplicons and deletions containing the human orthologues of candidate cancer genes were analysed in more detail to find the most promising candidates for a role in human cancer. The minimal amplified/deleted region was calculated by taking all of the amplicons/deletions in which the gene resided and finding the most 3’ start coordinate and the most 5’ end coordinate. The identities of other genes within the minimal region were established using the coordinates of human genes in Ensembl v37. For each gene within a minimum amplified or deleted region, the total number of cell lines in which that gene was amplified or deleted was calculated. In order to filter out the less likely candidates, genes in amplicons were discarded if they were co-amplified with known oncogenes or other mouse candidates from the gene list. Likewise, genes in deletions were discarded if they were co-deleted with known tumour suppressor genes or other mouse candidates from the gene list.

4.5.2.1 Candidate oncogenes among genes nearest to CISs

4.5.2.1.1 Protein-coding genes

The strongest candidates from the list of genes nearest to CISs with $P<0.001$ are shown in Table 4.4. Among 242 genes amplified to copy number 1.6 or above, 60 co-occurred with 1 or more known oncogenes and 128 co-occurred with other candidates in the list, of which 105 co-occurred with genes that were amplified in a greater number of cell lines. The filtered list of 54 candidates contained 14 known oncogenes, including *EVI1* and *FGFR2*, for which the murine insertions and human amplicons of less than 70 Mb are shown in Figure 4.17. 70 Mb is an arbitrary cut-off, but omits amplicons that are very large and for which there is therefore a low degree of certainty that the CIS genes are the targets of amplification. The kinase insert domain protein receptor gene (*KDR*) was amplified in 5 cell lines. Analysis of the insertions around *Kdr* in the mouse suggests that the adjacent gene, known oncogene *Kit*, may in fact be disrupted by the insertions assigned to *Kdr* (see Figure 2.11B, page 94). Likewise, the minimal amplified region containing *KDR* also contained *KIT*, which was amplified in an additional cell line (Figure 4.17) and is therefore the more likely target of amplification.

Further implicated oncogenes were also identified (Table 4.4). For example, the homeobox gene *MEIS1* is implicated in neuroblastoma. It was found to be amplified in the neuroblastoma cell line IMR-32 and was overexpressed in further neuroblastoma cell lines (Jones *et al.*, 2000). The single cell line in which it was amplified (to copy number 9.8) in this analysis was the neuroblastoma cell line GI-LI-N, which, according to the cell line typing analysis of the Cancer Genome Project (see Section 4.2.2), shares 96.0% identity with IMR-32, suggesting that they are derived from the same cancer. Even genes that are rarely amplified may therefore contribute to tumourigenesis. Likewise, the NF-κB transcription factor family member *NFKB1* was amplified to copy number 4.8 in one cell line (HH) derived from an adult T-cell lymphoma-leukaemia. Polymorphisms of *NFKB1* are associated with susceptibility to a number of cancers, including oral squamous cell carcinoma, myeloma, and cancers of the colon, liver and breast (for review, see Sun and Zhang, 2007). Interestingly, *NFKB1* maps to a region that is involved in translocations in certain types of acute lymphoblastic leukaemia (Liptay *et al.*, 1992).

Other implicated oncogenes that were amplified in human cancer and disrupted by retroviral insertions include matrix metalloproteinase-13 (*MMP13*) and mothers against decapentaplegic homolog 7 (*SMAD7*). *MMP13* shows recurrent amplification and overexpression in cervical cancer (Narayan *et al.*, 2007) and 2 of the 12 cell lines in this

CIS *P*-value	Gene name	Mouse Ensembl ID	Human Ensembl ID	Number of cell lines	Genes in minimal amplified region	Maximum copy number	Known oncogene?
0.001	*Myc*	ENSMUSG00000022346	ENSG00000136997	71	3	5.9+	Y
0.001	*Ccnd1*	ENSMUSG00000031071	ENSG00000110092	24	10	5.9+	Y
0.001	*Nmyc1*	ENSMUSG00000037169	ENSG00000134323	14	9	5.9+	Y
0.001	*Slamf6*	ENSMUSG00000015314	ENSG00000162739	14	21	5.9+	
0.001	*Smad7*	ENSMUSG00000025880	ENSG00000101665	6	27	5.9+	
0.001	*Fgfr2*	ENSMUSG00000030849	ENSG00000066468	5	7	5.9+	Y
0.001	*Kdr*	ENSMUSG00000062960	ENSG00000128052	5	15	5.9+	
0.001	*Tnfrsf7*	ENSMUSG00000030336	ENSG00000139193	5	26	5.9+	
0.001	*Meis1*	ENSMUSG00000020160	ENSG00000143995	1	2	5.9+	
0.001	*Mmp13*	ENSMUSG00000050578	ENSG00000137745	12	21	4.8	
0.001	*Nfkb1*	ENSMUSG00000028163	ENSG00000109320	1	7	4.8	
0.001	*Zfp217*	ENSMUSG00000052056	ENSG00000171940	43	15	4.3	
0.001	*Zfpn1a1*	ENSMUSG00000018654	ENSG00000185811	18	103	4.3	Y
0.001	*Ccnd3*	ENSMUSG00000034165	ENSG00000112576	8	37	4.3	Y
0.001	*Lmo2*	ENSMUSG00000032698	ENSG00000135363	4	43	4.3	Y
0.001	*Pim1*	ENSMUSG00000024014	ENSG00000137193	7	12	3.8	Y
0.001	*Ccnd2*	ENSMUSG00000000184	ENSG00000118971	6	15	3.8	Y
0.001	*Evi1*	ENSMUSG00000027684	ENSG00000085276	17	27	3.1	Y
0.001	*Btg2*	ENSMUSG00000020423	ENSG00000159388	10	35	3.1	
0.001	*Cd72*	ENSMUSG00000028459	ENSG00000137101	8	26	3.1	
0.001	*Rreb1*	ENSMUSG00000039087	ENSG00000124782	7	10	3.1	
0.001	*Aarsl*	ENSMUSG00000023938	ENSG00000124608	6	19	3.1	
0.001	*Taok3*	ENSMUSG00000061288	ENSG00000135090	3	9	3.1	
0.001	*Ntn1*	ENSMUSG00000020902	ENSG00000065320	2	31	3.1	
0.001	*Pik3r5*	ENSMUSG00000020901	ENSG00000141506	2	31	3.1	
0.001	*Eif4e3*	ENSMUSG00000030068	ENSG00000163412	2	33	3.1	
0.001	*Irf4*	ENSMUSG00000021356	ENSG00000137265	7	27	2.7	Y
0.001	*Ubb*	ENSMUSG00000019505	ENSG00000170315	5	72	2.5	
0.001	*Cd69*	ENSMUSG00000030156	ENSG00000110848	4	52	2.5	
0.001	*Lrrc5*	ENSMUSG00000046079	ENSG00000171492	2	25	2.5	
0.001	*Ptpn1*	ENSMUSG00000027540	ENSG00000196396	42	27	2.1	
0.001	*Sla2*	ENSMUSG00000027636	ENSG00000101082	41	84	2.1	
0.001	*E030003N15Rik*	ENSMUSG00000036661	ENSG00000105339	40	68	2.1	
0.001	*2310007D09Rik*	ENSMUSG00000027654	ENSG00000101447	38	61	2.1	
0.001	*Capsl*	ENSMUSG00000039676	ENSG00000152611	32	40	2.1	
0.001	*Cldn10*	ENSMUSG00000022132	ENSG00000134873	20	37	2.1	
0.001	*Ebi2*	ENSMUSG00000051212	ENSG00000169508	18	31	2.1	
0.001	*Flt3*	ENSMUSG00000042817	ENSG00000122025	11	122	2.1	Y
0.001	*Chc1l*	ENSMUSG00000022106	ENSG00000136161	10	53	2.1	
0.001	*Lcp1*	ENSMUSG00000021998	ENSG00000136167	10	128	2.1	Y
0.001	*4933403F05Rik*	ENSMUSG00000038121	ENSG00000177150	10	159	2.1	
0.001	*Dtl*	ENSMUSG00000037474	ENSG00000143476	8	70	2.1	
0.001	*2410129E14Rik*	ENSMUSG00000045136	ENSG00000137285	7	18	2.1	
0.001	*1110036O03Rik*	ENSMUSG00000006931	ENSG00000141696	6	43	2.1	
0.001	*Fmnl1*	ENSMUSG00000055805	ENSG00000184922	6	85	2.1	
0.001	*Ksr*	ENSMUSG00000018334	ENSG00000141068	5	34	2.1	
0.001	*Jundm2*	ENSMUSG00000034271	ENSG00000140044	13	42	1.6	
0.001	*Tomm20*	ENSMUSG00000058779	ENSG00000173726	8	252	1.6	
0.001	*Cyb5*	ENSMUSG00000024646	ENSG00000166347	6	20	1.6	
0.001	*Ldhd*	ENSMUSG00000031958	ENSG00000166816	6	74	1.6	
0.001	*Cbfa2t3h*	ENSMUSG00000006362	ENSG00000129993	5	133	1.6	Y
0.001	*Zfp608*	ENSMUSG00000052713	ENSG00000168916	3	30	1.6	
0.001	*2610307O08Rik*	ENSMUSG00000024349	ENSG00000184584	3	95	1.6	
0.001	*Hhex-rs2*	ENSMUSG00000024986	ENSG00000152804	2	40	1.6	
0.05	*D930036F22Rik*	ENSMUSG00000035181	ENSG00000129493	17	19	5.9+	
0.05	*Laptm5*	ENSMUSG00000028581	ENSG00000162511	1	11	2.7	
0.05	*Emp3*	ENSMUSG00000040212	ENSG00000142227	7	33	2.7	
0.05	*Rai1*	ENSMUSG00000062115	ENSG00000108557	5	72	2.1	

Table 4.4. Genes that are nearest to CISs in mouse lymphomas and are also promising candidates for targets of amplification in human cancer cell lines. "CIS *P*-value" is the minimum threshold for the significance of the CIS nearest to the given gene. "Number of cell lines" is the number of samples in which the gene is amplified to a copy number of greater than or equal to 1.6. "Genes in minimal amplified region" is the number of genes that co-occur with the CIS gene in the smallest region of amplification. "Maximum copy number" is the maximum copy number threshold above which the gene is identified as being amplified. "Known oncogene?" indicates whether the gene is a dominant cancer gene listed in the Cancer Gene Census.

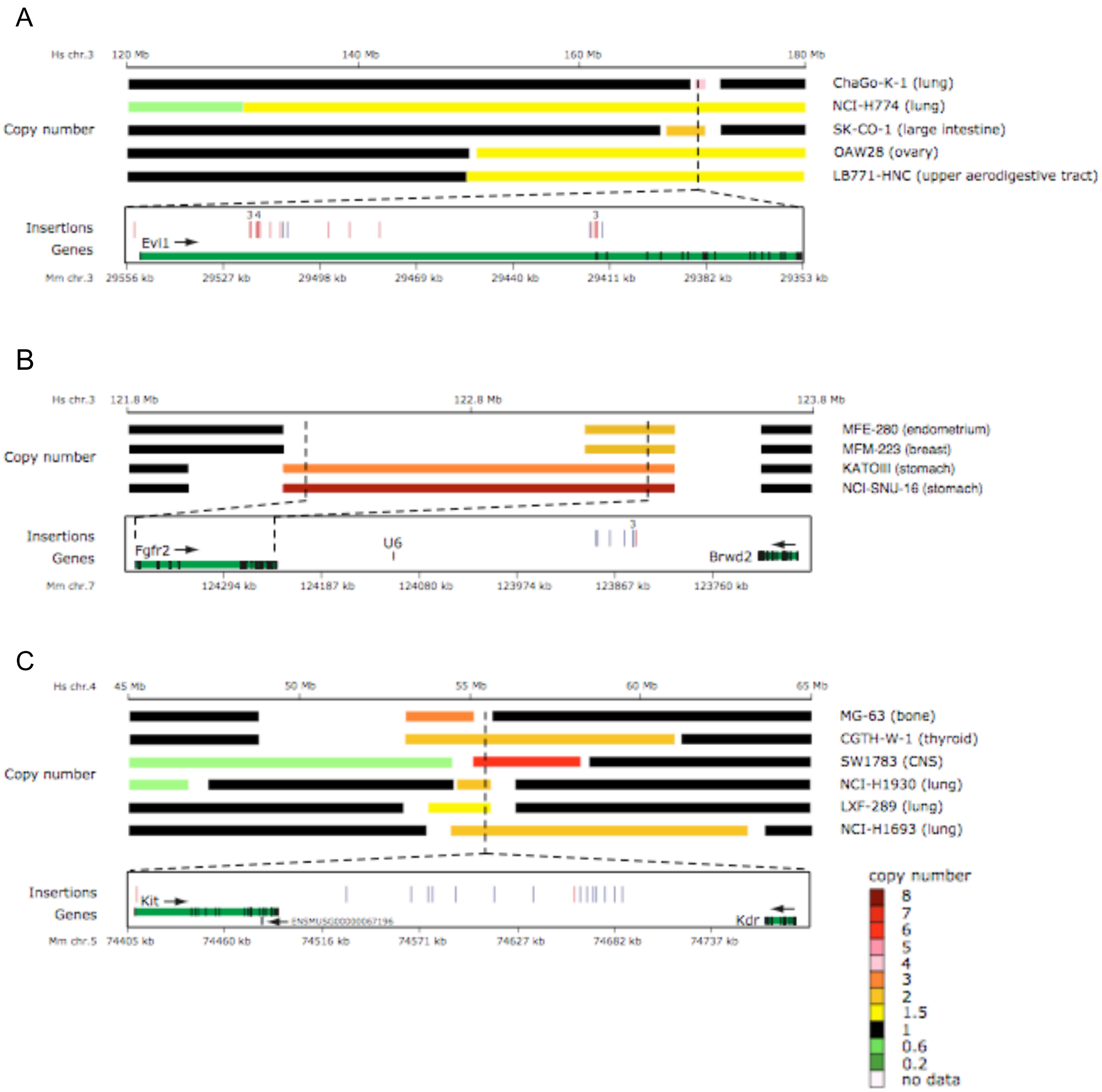

Figure 4.17. Known human oncogenes *EVI1* (A), *FGFR2* (B) and *KIT* (C) are amplified in human cancer cell lines and are disrupted by retroviral insertional mutagenesis in mouse lymphomas. The copy number of chromosomal regions in the human cell lines is depicted in colour. Names of human cell lines and tissue of origin are provided Only cell lines in which the amplicon containing the oncogene is less than 70 Mb are shown. The lower part of each figure shows insertions within mouse tumours, and was kindly provided by Jaap Kool and Jeroen de Ridder. Blue vertical lines represent insertions in the sense orientation, while red vertical lines represent antisense insertions. Genes are shown in green, with exons marked in black. Positions on the murine and human chromosomes are indicated on the black horizontal bars in kb and Mb, respectively. These figures can also be seen in Uren *et al*. (2008).

study in which *MMP13* was amplified were indeed derived from cervical cancers. *MMP13* has not been shown to be amplified in any other cancer types, but overexpression has been observed, e.g. in squamous cell carcinomas of the head and neck (Johansson *et al.*, 1997) and vulva (Johansson *et al.*, 1999). The results of this analysis suggest that *MMP13* is amplified in, and implicated in, a range of cancer types. The cell lines containing amplicons of less than 70 Mb that encompass *MMP13* are shown in Figure 4.18. Among these types are oesophageal, skin and breast cancers, in which *MMP13* overexpression has been observed (Freije *et al.*, 1994; Hu *et al.*, 2001; Kuivanen *et al.*, 2006). The minimal amplified region on chromosome 11 contains 21 genes, including a cluster of genes encoding matrix metalloproteinases, of which a number have been previously implicated in cancer. However, *MMP13* was the only gene disrupted by insertional mutagenesis.

SMAD7 duplication has been demonstrated in colorectal cancer (Boulay *et al.*, 2001) and the gene is overexpressed in a number of cancer types, including basal cell carcinoma (Gambichler *et al.*, 2007), endometrial cancer (Dowdy *et al.*, 2005) and thyroid follicular carcinoma cell lines (Cerutti *et al.*, 2003). The highest amplification of *SMAD7* was in the retinoblastoma cell line Y79. Interestingly, SMAD7 has been shown to suppress TGF-β1-mediated growth inhibition in pancreatic cancer cells through the inactivation of the retinoblastoma protein (Boyer Arnold and Korc, 2005) and it inhibits growth arrest and apoptosis in mouse B cells through the inactivation of retinoblastoma (Ishisaki *et al.*, 1998; Nakahara *et al.*, 2003). In addition, *SMAD7* is expressed in the eye, and suppresses TGF-β2-mediated inhibition of corneal endothelial cell proliferation, resulting in accelerated wound healing (Funaki *et al.*, 2003). *SMAD7* is therefore a promising target for amplification in the retinoblastoma cell line. Likewise, one of the amplicons encompassing *SMAD7* was identified in a Ewing's sarcoma cell line (EW-24) and, in osteogenesis, SMAD7 suppresses osteoblast differentiation and bone formation (Koinuma and Imamura, 2005) and inhibits Saos2 osteosarcoma cell differentiation (Eliseev *et al.*, 2006). *SMAD7* was also amplified in 2 haematopoietic and 2 lung cancer cell lines. SMAD7 promotes self-renewal of haematopoietic stem cells (Blank *et al.*, 2006) and is highly expressed in metastatic lung cancer cell lines (Shen *et al.*, 2003).

Other interesting candidates include SLAM family member 6 precursor (*SLAMF6*), serine/threonine-protein kinase TAO3 (*TAOK3*), RAS-responsive element-binding protein 1 (*RREB1*) and leucine-rich repeat-containing protein 8D (*LRRC5*). The minimal

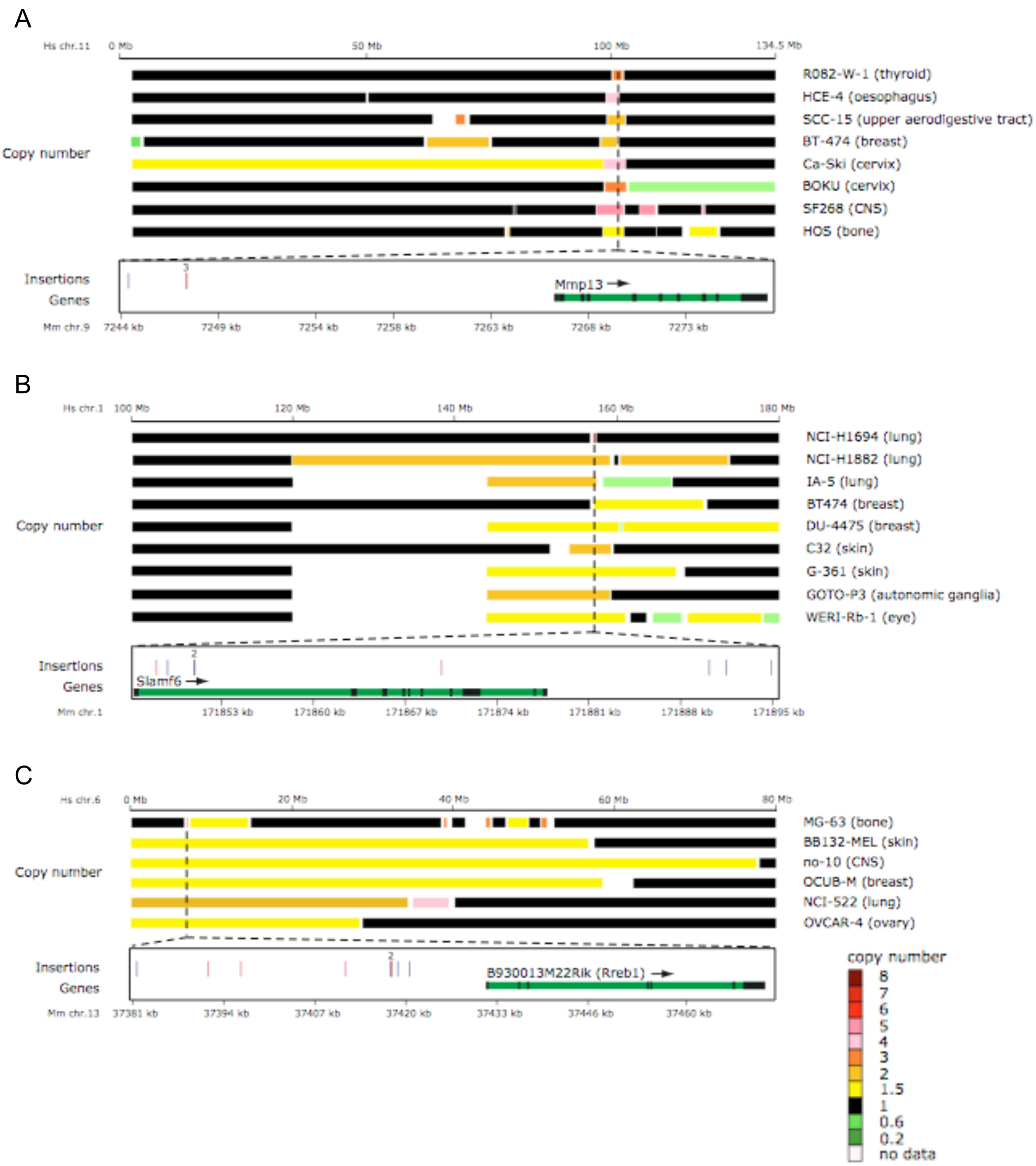

Figure 4.18. Candidate oncogenes *MMP13* (A), *SLAMF6* (B) and *RREB1* (C) are amplified in human cancer cell lines and are disrupted by retroviral insertional mutagenesis in mouse lymphomas. The copy number of chromosomal regions in the human cell lines is depicted in colour. Names of human cell lines and tissue of origin are provided Only cell lines in which the amplicon containing the oncogene is less than 70 Mb are shown. The lower part of each figure shows insertions within mouse tumours, and was kindly provided by Jaap Kool and Jeroen de Ridder. Blue vertical lines represent insertions in the sense orientation, while red vertical lines represent antisense insertions. Genes are shown in green, with exons marked in black. Positions on the murine and human chromosomes are indicated on the black horizontal bars in kb and Mb, respectively. These figures can also be seen in Uren *et al*. (2008).

amplified region encompassing *SLAMF6* comprised 21 genes and was recurrent across 14 cell lines. Cell lines containing an amplicon of less than 70 Mb are shown in Figure 4.18. The highest amplification of *SLAMF6* was within the lung cancer cell line NCI-H1694. However, it has been proposed that *SLAMF6*, also known as *Ly108*, is only expressed in lymphoid tissues (Peck and Ruley, 2000), where it regulates T cell development (Jordan *et al.*, 2007) and B cell tolerance (Kumar *et al.*, 2006). Polymorphisms within the gene are associated with systemic lupus erythematosus (Wandstrat *et al.*, 2004), which has been widely associated with an increased risk of developing a range of cancers, but most strongly with cancers arising from B lymphocytes (Bernatsky *et al.*, 2007). *TAOK3* was only amplified in 3 cell lines, but the minimum region contained just 9 genes. *TAOK3* is poorly characterised, but contains a somatic missense mutation in 2 lung cancers (small cell carcinoma cell line NCI-H28 and a primary adenocarcinoma) in the COSMIC database (Forbes *et al.*, 2006). Although a role in tumourigenesis has not been demonstrated for *TAOK3*, protein kinases are widely implicated in cancer (see Sections 1.2.5.2 and 1.3.1). *RREB1* was amplified in 7 cell lines within a minimal region of 10 genes. Cell lines containing an amplicon of less than 70 Mb are shown in Figure 4.18. Each cell line was derived from a different tissue, but *RREB1* has been shown to be ubiquitously expressed in human tissues apart from the adult brain (Thiagalingam *et al.*, 1997). Rreb1 binds to, and represses expression of, the $p16^{Ink4a}$ promoter, and the development of pristine-induced plasma cell tumours in Balb/C mice is attributable to a polymorphism in this Rreb1 binding site (Zhang *et al.*, 2003). In addition, *RREB1* is important in reducing cell-cell adhesion and collective migration of epithelial cells (Melani *et al.*, 2008), and it may therefore play a role in metastasis. RREB1 has also been identified as a transcriptional effector of RAL (Oxford *et al.*, 2007), and RALA is itself implicated in cancer cell migration, as well as other cancer-related functions (Oxford *et al.*, 2005). The most amplified occurrence of *RREB1* was in the osteosarcoma cell line MG-63, but no role for *RREB1* has previously been elucidated in bone tissue. *LRRC5* was amplified in just 2 cell lines, with a minimal amplified region of 25 genes. Although little is known about this gene, it is thought that it might be implicated in the proliferation and activation of lymphocytes and monocytes, suggesting a possible role in the oncogenesis of B cells which would account for the insertions disrupting *Lrrc5* in mouse lymphomas. However, *LRRC5* was amplified in human cancer cell lines derived from the ovary and upper aerodigestive tract.

Only 4 additional candidates (*D930036F22Rik*, *LAPTM5*, *EMP3* and *RAI1*) were identified using genes nearest to CISs with *P*<0.05 (see Table 4.4). *D930036F22Rik* is also known as HEAT repeat containing 5A (*HEATR5A*). The minimal amplified region also included Rho-GTPase-activating protein 5 (*p190-B*), which is known to be overexpressed in breast cancer (Chakravarty *et al.*, 2000), although only 1 breast cancer cell line contained an amplification of this region. Based on the distribution of insertions in the CIS, it is entirely possible that nearby genes *Hectd1* and/or *EG544864* were in fact the targets of MuLV mutagenesis (Figure 4.19). However, none of these genes have been previously implicated in tumourigenesis. Lysosomal-associated protein transmembrane 5 (*LAPTM5*) was amplified in a single cell line derived from an endometrial carcinoma (MFE-280). *LAPTM5* is inactivated by chromosomal rearrangement and DNA methylation in human multiple myeloma (Hayami *et al.*, 2003) but is overexpressed in malignant B lymphomas (Seimiya *et al.*, 2003) and is a predictor for early intrahepatic recurrence of hepatocellular carcinoma (Somura *et al.*, 2008). However, the amplified region in MFE-280 also contained the syndecan-3 gene, which is expressed in the human endometrium (Germeyer *et al.*, 2007) and is thought to play a role in uterine growth (Russo *et al.*, 2001). Epithelial membrane protein gene *EMP3* was proposed as a candidate tumour suppressor in glioma and neuroblastoma (Alaminos *et al.*, 2005), but it has since been shown to be overexpressed in oligodendroglial tumours (Li *et al.*, 2007a) and primary glioblastomas (Kunitz *et al.*, 2007). It is also overexpressed in invasive human mammary carcinoma cell lines (Evtimova *et al.*, 2003) and contains a polymorphism in prostate cancers (Burmester *et al.*, 2004). Retinoic acid induced 1 gene (*RAI1*) has not been previously implicated in cancer.

4.5.2.1.2 miRNA genes

The list of genes nearest to CISs with *P*<0.001 contained 6 miRNA genes, while the list nearest to CISs with *P*<0.05 contained 9. As mentioned in Section 3.2.2, deregulated miRNAs are implicated in promoting and suppressing tumourigenesis in a range of tissues. Currently, only protein-coding genes have human orthologues in Ensembl, and miRNA genes were therefore omitted from the global comparison and principal analysis of CIS genes within amplicons. However, it is possible to manually identify the human equivalents based on the miRNA name and the conserved synteny between the mouse and human genomes. Table 4.5 shows the name of the murine miRNA and the corresponding human miRNA for genes nearest to CISs, as well as lists of the miRNA genes within

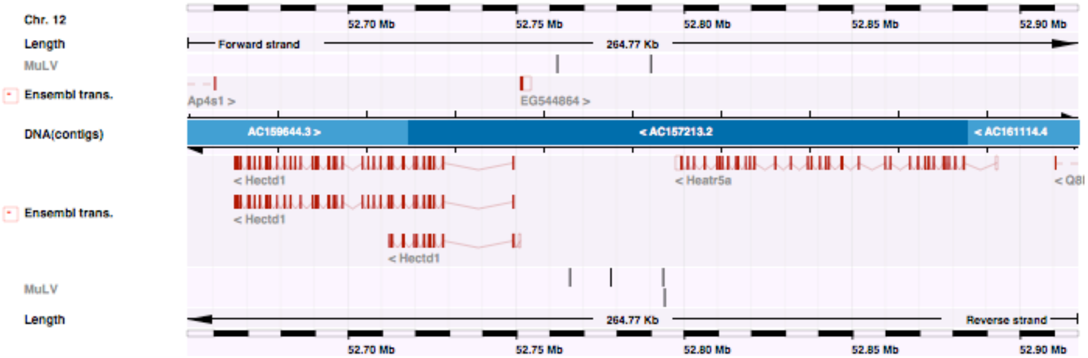

Figure 4.19. Insertions assigned to *Heatr5a* may be associated with *Hectd1* or *EG544864.* Insertions are shown as black vertical lines. Those above the blue bar labelled DNA(contigs) are in the sense orientation, those below are in the antisense orientation. Ensembl genes are shown in red.

A

P-value	Mouse miRNA	Human miRNA
0.001	rno-mir-128b	hsa-mir-128b
0.001	hsa-mir-142	hsa-mir-142
0.001	mmu-mir-21	hsa-mir-21
0.001	mmu-mir-23a	hsa-mir-23a
0.001	mmu-mir-17	hsa-mir-17
0.001	hsa-mir-106a	hsa-mir-106a
0.05	mmu-mir-26b	hsa-mir-26b
0.05	mmu-mir-22	hsa-mir-22
0.05	rno-mir-200b	hsa-mir-200b

B

Mouse miRNA gene	Mouse Ensembl ID	Human miRNA gene	Human Ensembl ID	Number of cell lines	Maximum copy number	Known oncogenes in minimal region?
mmu-mir-17	ENSMUSG00000065508	*hsa-mir-17*	ENSG00000198999	23	3.1	
hsa-mir-142	ENSMUSG00000065420	*hsa-mir-142*	ENSG00000199166	9	2.1	Y
mmu-mir-21	ENSMUSG00000065455	*hsa-mir-21*	ENSG00000199004	9	2.1	Y
mmu-mir-23a	ENSMUSG00000065611	*hsa-mir-23a*	ENSG00000199028	8	2.1	Y
rno-mir-128b	ENSMUSG00000065441	*hsa-mir-128b*	ENSG00000199105	1	2.1	

C

Mouse miRNA gene	Mouse Ensembl ID	Human miRNA gene	Human Ensembl ID	Number of cell lines	Minimum copy number
rno-mir-128b	ENSMUSG00000065441	*hsa-mir-128b*	ENSG00000199105	40(2)	0.2
mmu-mir-17	ENSMUSG00000065508	*hsa-mir-17*	ENSG00000198999	61	0.6
mmu-mir-22	ENSMUSG00000065529	*hsa-mir-22*	ENSG00000199060	17	0.6
mmu-mir-26b	ENSMUSG00000065468	*hsa-mir-26b*	ENSG00000199121	8	0.6
mmu-mir-21	ENSMUSG00000065455	*hsa-mir-21*	ENSG00000199004	6	0.6
hsa-mir-142	ENSMUSG00000065420	*hsa-mir-142*	ENSG00000199166	5	0.6
mmu-mir-23a	ENSMUSG00000065611	*hsa-mir-23a*	ENSG00000199028	3	0.6

Table 4.5. miRNA genes that are nearest to CISs in mouse lymphomas and are amplified and/or deleted in human cancer cell lines. (A) Names of the murine miRNAs and their human orthologues. (B) Amplified miRNA genes. (C) Deleted miRNA genes. "*P*-value" is the minimum threshold for the significance of the CIS nearest to the given gene. "Number of cell lines" is the number of samples in which the gene is amplified to a copy number of greater than or equal to 1.6 (B) or deleted to a copy number of less than or equal to 0.6, with the number for deletions of copy number 0.2 or below shown in brackets (C). "Maximum copy number" is the maximum copy number threshold above which the gene is identified as being amplified. "Minimum copy number" is the minimum copy number threshold below which the gene is identified as being deleted.

human amplicons and deletions. Genes encoding 5 of the miRNAs were amplified in human cancer cell lines. The minimal amplified regions for *hsa-miR-142* and *hsa-miR-23a* were very large and encompassed 4 and 3 known oncogenes, respectively, while *hsa-miR-21* was co-amplified with *hsa-miR-23a*, and *hsa-miR-128b* was amplified in just 1 cell line. The minimal amplified region of *hsa-miR-17* contained 14 genes, of which none were oncogenes and only 2 had a description in Ensembl. hsa-miR-17 is part of the miR-17-92 cluster of 6 miRNAs. All 3 of the cell lines in which *hsa-miR-17* was amplified to copy number 2.5 or above were derived from haematopoietic and lymphoid cancers, which is consistent with a role for miR-17-92 in both B-lymphocyte development and B-lymphoproliferative disorders (Garzon and Croce, 2008). The cluster is also overexpressed in other human cancers, including colorectal cancers (Monzo *et al.*, 2008), anaplastic thyroid cancer cells (Takakura *et al.*, 2008), neuroblastomas with *MYCN* amplification (Schulte *et al.*, 2008), bladder cancers (Gottardo *et al.*, 2007) and lung cancers (Hayashita *et al.*, 2005). Of the 23 cell lines in which *hsa-mir-17* was amplified to copy number 1.6 or above, 7 were from colon cancers, 7 were from haematopoietic and lymphoid cancers, 4 were from lung cancers, and 1 each were from cancers of the stomach, soft tissue, central nervous system, breast and eye. miR-17-92 is the likely target for amplification within this region, and demonstrates that the function of miRNAs may be conserved between species.

7 miRNA genes were identified within deletions of copy number 0.6 or below, but only *hsa-miR-17* and *hsa-miR-128b* were deleted in a large number of cell lines, and only *hsa-miR-128b* was within homozygous deletions. *hsa-miR-128b* was shown to be downregulated in classic Hodgkin lymphomas infected with Epstein-Barr virus (Navarro *et al.*, 2008). However, it is also upregulated in acute lymphoblastic leukaemia (Zanette *et al.*, 2007). The two homozygous deletions of *hsa-miR-128b* were within glioma and neuroblastoma cell lines, respectively. Interestingly, *hsa-miR-128* is highly expressed in the adult brain and preferentially in neurons, where it is thought to play a role in neural differentiation (Smirnova *et al.*, 2005). Downregulation of *hsa-miR-128* has been previously demonstrated in glioblastoma (Ciafre *et al.*, 2005), but deletion of the gene has not been previously demonstrated.

4.5.2.2 Candidate cancer genes among genes containing insertions within the coding region

Analyses involving the remaining lists were primarily designed to identify tumour suppressor genes, but as shown in Section 4.5.1.3, it is likely that the lists are contaminated with candidate oncogenes. The list for which the pattern of distribution was most similar to that of oncogenes was the list of genes containing insertions, not including those represented by a single read, within the coding region. The strongest candidates in regions of copy number gain, identified using the same filtering procedure as used for genes nearest to CISs, included 4 oncogenes (Table 4.6). As discussed in Section 3.4.1, insertions within the 3' UTR of *Mycn* and *Pim1* result in the formation of a more stable protein product, rather than gene inactivation. Some of the insertions were within the last exon of these genes, which explains their inclusion within the current list.

Candidate tumour suppressor genes were identified within regions of copy number loss. Among these was the gene encoding transmembrane protease, serine 2 precursor (*TMPRSS2*), which is a known oncogene that forms fusions with the ETS transcription factor genes *ERG* and *ETV1* in prostate cancer (Tomlins *et al.*, 2005) and is overexpressed in most prostate cancers (Vaarala *et al.*, 2001). A hemizygous microdeletion within the fusion has been observed on chromosome 21 between *ERG* and *TMPRSS2* (Yoshimoto *et al.*, 2006), but this does not explain the deletion of the entire gene, sometimes in both copies. In addition, the homozygous deletions, which were also the most focal deletions, were identified in cancer cell lines derived from the pancreas and upper aerodigestive tract, rather than the prostate. Most of the heterozygous deletions were very large, encompassing many genes, and the minimal deleted region contained 18 genes. Based on the known role of *TMPRSS2*, it seems unlikely that this is the target of deletion within this region. Likewise, although deleted in human cancers, *MAP3K8* and *IL6RA* are also more likely to act as oncogenes. *MAP3K8* is overexpressed in, for example, invasive endometrioid cancer (Aparecida Alves *et al.*, 2006), T-cell neoplasias (Christoforidou *et al.*, 2004) and breast cancer (Sourvinos *et al.*, 1999), while expression of the interleukin 6 receptor gene *IL6RA* is promoted by Epstein-Barr virus in immortalised B cells and Burkitt's lymphoma cells (Klein *et al.*, 1995). The minimal amplified region containing the gene encoding MAGUK p55 subfamily member 4 (MMP4) comprised just 4 genes. Interestingly, *MMP4* is a homologue of the Drosophila *Stardust* gene, which is involved in establishing and maintaining epithelial tissue polarity,

Gene name	Mouse Ensembl ID	Human Ensembl ID	Number of cell lines	Genes in minimal region	Copy number	Singletons only?	Known oncogene?
Capsl	ENSMUSG00000039676	ENSG00000152611	32	40	1.6+		
Bcl9	ENSMUSG00000038256	ENSG00000116128	17	62	1.6+		Y
Mycn	ENSMUSG00000037169	ENSG00000134323	14	9	1.6+		Y
Ccnd3	ENSMUSG00000034165	ENSG00000112576	8	37	1.6+		Y
Pim1	ENSMUSG00000024014	ENSG00000137193	7	12	1.6+		Y
NM_009283.2	ENSMUSG00000026104	ENSG00000115415	2	12	1.6+		
Mrps18b	ENSMUSG00000024436	ENSG00000137330	12	19	0.6		
Mpp4	ENSMUSG00000026024	ENSG00000003393	9	4	0.6		
Il6ra	ENSMUSG00000027947	ENSG00000160712	3	23	0.6		
Phgdhl1	ENSMUSG00000041765	ENSG00000134882	63(1)	9(21)	0.2		
Map3k8	ENSMUSG00000024235	ENSG00000107968	35(1)	13(20)	0.2		
Tmprss2	ENSMUSG00000000385	ENSG00000184012	14(2)	18(18)	0.2		Y
Tmem16f	ENSMUSG00000064210	ENSG00000177119	3	1	1.6+	Y	
Nfkb1	ENSMUSG00000028163	ENSG00000109320	1	7	1.6+	Y	
9030611O19Rik	ENSMUSG00000036136	ENSG00000184731	3	16	1.6+	Y	
Olfr1509	ENSMUSG00000035626	ENSG00000182735	12	26	1.6+	Y	
	ENSMUSG00000046186	ENSG00000156535	3	35	1.6+	Y	
Rasgrp4	ENSMUSG00000030589	ENSG00000171777	12	56	1.6+	Y	
Dsg1b	ENSMUSG00000061928	ENSG00000134760	76	89	0.6	Y	
XP_484397.2	ENSMUSG00000034731	ENSG00000102780	68	8	0.6	Y	
Riok3	ENSMUSG00000024404	ENSG00000101782	68	47	0.6	Y	
6330406I15Rik	ENSMUSG00000029659	ENSG00000102802	64	9	0.6	Y	
Il17rb	ENSMUSG00000015966	ENSG00000056736	46	14	0.6	Y	
Zmynd11	ENSMUSG00000021156	ENSG00000015171	46	32	0.6	Y	
Hmgb2	ENSMUSG00000054717	ENSG00000164104	43	13	0.6	Y	
Gtse1	ENSMUSG00000022385	ENSG00000075218	41	90	0.6	Y	
1700020C11Rik	ENSMUSG00000004748	ENSG00000100010	36	61	0.6	Y	
Slc37a2	ENSMUSG00000032122	ENSG00000134955	34	70	0.6	Y	
Man1a	ENSMUSG00000003746	ENSG00000111885	33	8	0.6	Y	
Ate1	ENSMUSG00000030850	ENSG00000107669	32	6	0.6	Y	
Nrap	ENSMUSG00000049134	ENSG00000197893	32	9	0.6	Y	
Q91VN2_MOUSE	ENSMUSG00000042293	ENSG00000180425	32	16	0.6	Y	
Snf1lk2	ENSMUSG00000037112	ENSG00000170145	32	40	0.6	Y	
Dnajc9	ENSMUSG00000021811	ENSG00000182180	31	26	0.6	Y	
3110003A17Rik	ENSMUSG00000019855	ENSG00000146386	30	11	0.6	Y	
Shb	ENSMUSG00000044813	ENSG00000107338	29	16	0.6	Y	
Hp1bp3	ENSMUSG00000028759	ENSG00000127483	24	84	0.6	Y	
1200009I06Rik	ENSMUSG00000021280	ENSG00000185215	23	94	0.6	Y	
8430406I07Rik	ENSMUSG00000027424	ENSG00000125871	21	20	0.6	Y	
Wdr5b	ENSMUSG00000034379	ENSG00000196981	17	14	0.6	Y	
Zdhhc23	ENSMUSG00000036304	ENSG00000184307	17	44	0.6	Y	
9030611O19Rik	ENSMUSG00000036136	ENSG00000184731	15	22	0.6	Y	
Gcnt2	ENSMUSG00000021360	ENSG00000111846	14	47	0.6	Y	
Itk	ENSMUSG00000020395	ENSG00000113263	13	31	0.6	Y	
	ENSMUSG00000039153	ENSG00000124813	13	97	0.6	Y	
Dok3	ENSMUSG00000035711	ENSG00000146094	13	183	0.6	Y	
Hivep3	ENSMUSG00000028634	ENSG00000127124	12	2	0.6	Y	
CSDE1_MOUSE	ENSMUSG00000068823	ENSG00000009307	11	25	0.6	Y	
8430438D04Rik	ENSMUSG00000036019	ENSG00000179104	10	4	0.6	Y	
Bcl10	ENSMUSG00000028191	ENSG00000142867	10	33	0.6	Y	
Rgs2	ENSMUSG00000026360	ENSG00000116741	7	16	0.6	Y	
Jmjd4	ENSMUSG00000036819	ENSG00000081692	7	93	0.6	Y	
Tssk6	ENSMUSG00000047654	ENSG00000178093	6	60	0.6	Y	
Tdrd5	ENSMUSG00000060985	ENSG00000162782	5	46	0.6	Y	
Sell	ENSMUSG00000026581	ENSG00000188404	4	51	0.6	Y	
Leprel1	ENSMUSG00000038168	ENSG00000090530	8(1)	29(33)	0.2	Y	
Olfr1509	ENSMUSG00000035626	ENSG00000182735	19(3)	19(27)	0.2	Y	
Dut	ENSMUSG00000027203	ENSG00000128951	14(1)	7(7)	0.2	Y	
3200002M19Rik	ENSMUSG00000030649	ENSG00000110200	11(1)	24(52)	0.2	Y	

Table 4.6. Mouse genes that contain retroviral insertions within the coding region and are also promising candidates for targets of amplification or deletion in human cancer cell lines. “Number of cell lines” is the number of samples in which the gene is amplified or deleted. “Copy number” is the maximum copy number threshold above which the gene is identified as being amplified, or the minimal threshold below which the gene is deleted. Where the copy number is 0.2, the number of cell lines and number of genes in the minimal deleted region are given for deletions of copy number <= 0.6, with numbers for copy number <= 0.2 being shown in brackets. “Genes in minimal region” is the number of genes that co-occur with the CIS gene in the smallest region of amplification/deletion. “Singletons only?” indicates whether the gene contains insertions other than those represented by a single read. “Known oncogene?” indicates whether the gene is a dominant cancer gene listed in the Cancer Gene Census.

which is disrupted in epithelial tumours. Although *MMP4* has not been implicated in cancer, expression of another family member, known as *MMP7*, has been demonstrated in tumours of the uterus and bladder, and in lymphomas (Katoh and Katoh, 2004).

Among the genes that contained insertions represented by a single read, 3 genes (*SHB*, *HIVEP3* and *BCL10*) stood out as potential tumour suppressor genes. Overexpression of the gene encoding the SHB adaptor protein causes increased activity of the pro-apoptotic kinase c-ABL, resulting in reduced tumour growth in PC3 prostate cancer cells (Davoodpour *et al.*, 2007). Therefore, it is possible that deletion of the gene may lead to tumour cell growth and proliferation. The human immunodeficiency virus type 1 enhancer binding protein 3 gene (*Hivep3*, also known as *Krc*), positively regulates transcription of the mouse metastasis-associated gene, *S100A4/mts1* (Hjelmsoe *et al.*, 2000). In addition, *KRC* was proposed as a potential tumour suppressor gene following the development of a teratoma from *KRC*-deficient embryonic stem cells introduced into an animal model (Allen *et al.*, 2002). Finally, the B-cell lymphoma/leukaemia 10 gene (*BCL10*) is a "hotspot" within the commonly deleted region 1p22.3 in mantle cell lymphomas. Interestingly, 5 of the 10 cell lines containing a deletion within this region were derived from tumours of the autonomic ganglia, but no role for *BCL10* has previously been demonstrated in these cancers (Balakrishnan *et al.*, 2006).

4.5.2.3 Candidate tumour suppressor genes among genes containing insertions within the translated or transcribed region

Candidates among the lists of genes containing insertions within the translated or transcribed region are combined in Table 4.7. The gene that was most frequently deleted below the copy number thresholds of both 0.6 and 0.2 was the known tumour suppressor gene *CDKN2A* (also known as the *INK4A/ARF* locus, and described in Section 1.2.6). This demonstrates the efficacy of the analysis, since homozygous and heterozygous deletions of *CDKN2A* are commonly observed in a wide range of cancers. The only other known tumour suppressor gene in the list, according to the Cancer Gene Census, was the gene encoding FAS, which is a member of the TNF receptor superfamily. Binding of the FAS ligand to the FAS receptor results in the formation of the death-inducing complex (DISC), which triggers apoptosis (for review, see Wajant, 2002). The implicated tumour suppressor gene *WWOX* was also frequently deleted. *WWOX* resides in a fragile site and therefore while it is frequently deleted in cancers, it is unclear whether it contributes to

Gene name	Mouse Ensembl ID	Human Ensembl ID	Number of cell lines	Genes in minimal region	Copy number	Insertions in translated region?	Singletons only?	Known TSG?
Cdkn2a	ENSMUSG00000044303	ENSG00000147889	207(145)	1(3)	0.2	Y		Y
Nfatc1	ENSMUSG00000033016	ENSG00000131196	112(2)	15(15)	0.2	Y		
Zfp532	ENSMUSG00000042439	ENSG00000074657	100(1)	14(14)	0.2			
Dock8	ENSMUSG00000052085	ENSG00000107099	88(5)	15(15)	0.2			
Rnf125	ENSMUSG00000033107	ENSG00000101695	78(1)	14(14)	0.2			
Sacs	ENSMUSG00000048279	ENSG00000151835	64(1)	4(5)	0.2	Y		
Arhgef3	ENSMUSG00000021895	ENSG00000163947	47(2)	10(10)	0.2	Y		
Rbms3	ENSMUSG00000039607	ENSG00000144642	43(4)	1(3)	0.2	Y		
Arpp21	ENSMUSG00000032503	ENSG00000172995	40(2)	2(2)	0.2			
Grm1	ENSMUSG00000019828	ENSG00000152822	37(6)	2(1)	0.2	Y		
Scye1	ENSMUSG00000028029	ENSG00000164022	36(1)	12(17)	0.2	Y		
Fas	ENSMUSG00000024778	ENSG00000026103	35(2)	11(13)	0.2	Y		Y
Mthfd1l	ENSMUSG00000040675	ENSG00000120254	35(1)	5(5)	0.2	Y		
Map3k8	ENSMUSG00000024235	ENSG00000107968	35(1)	13(20)	0.2	Y		
Wwox	ENSMUSG00000004637	ENSG00000186153	34(3)	1(2)	0.2	Y		
Esr1	ENSMUSG00000019768	ENSG00000091831	34(1)	5(5)	0.2	Y		
Prkg1	ENSMUSG00000052920	ENSG00000185532	33(1)	2(2)	0.2	Y		
Prep	ENSMUSG00000019849	ENSG00000085377	33(1)	10(19)	0.2	Y		
Utrn	ENSMUSG00000019820	ENSG00000152818	32(1)	2(1)	0.2	Y		
Cdc14b	ENSMUSG00000033102	ENSG00000081377	31(1)	19(19)	0.2			
Ank3	ENSMUSG00000069601	ENSG00000151150	29(1)	1(1)	0.2	Y		
XP_485387.1	ENSMUSG00000038578	ENSG00000106868	25(2)	5(5)	0.2	Y		
4831426I19Rik	ENSMUSG00000054150	ENSG00000176438	25(1)	2(7)	0.2			
A530016O06Rik	ENSMUSG00000050103	ENSG00000187546	24(7)	1(1)	0.2	Y		
Ches1	ENSMUSG00000033713	ENSG00000053254	24(1)	17(17)	0.2			
Auts2	ENSMUSG00000056924	ENSG00000158321	22(1)	1(14)	0.2	Y		
Rasgrp1	ENSMUSG00000027347	ENSG00000172575	21(6)	6(7)	0.2	Y		
Sec8l1	ENSMUSG00000029763	ENSG00000131558	20(1)	1(3)	0.2	Y		
Hars2	ENSMUSG00000027430	ENSG00000125821	20(1)	10(20)	0.2	Y		
Rad51l1	ENSMUSG00000059060	ENSG00000182185	19(1)	3(11)	0.2	Y		
Magi2	ENSMUSG00000040003	ENSG00000187391	18(3)	7(6)	0.2	Y		
Gys2	ENSMUSG00000030244	ENSG00000111713	17(1)	14(18)	0.2	Y		
Atg10	ENSMUSG00000021619	ENSG00000152348	16(4)	5(5)	0.2	Y		
Gnefr	ENSMUSG00000030839	ENSG00000129158	16(2)	2(4)	0.2	Y		
Dmxl1	ENSMUSG00000037416	ENSG00000172869	15(1)	23(23)	0.2	Y		
Frmd6	ENSMUSG00000048285	ENSG00000139926	14(1)	12(12)	0.2	Y		
1810060J02Rik	ENSMUSG00000030301	ENSG00000123106	14(1)	13(13)	0.2	Y		
Sipa1l2	ENSMUSG00000001995	ENSG00000116991	12(1)	7(9)	0.2	Y		
Zfp496	ENSMUSG00000020472	ENSG00000162714	11(2)	16(16)	0.2			
AI194318	ENSMUSG00000048058	ENSG00000179241	11(2)	3(3)	0.2	Y		
Eltd1	ENSMUSG00000039167	ENSG00000162618	11(1)	2(2)	0.2	Y		
Crim1	ENSMUSG00000024074	ENSG00000150938	9(1)	1(1)	0.2	Y		
Car2	ENSMUSG00000027562	ENSG00000104267	8(1)	12(12)	0.2	Y		
Ctnnd1	ENSMUSG00000034101	ENSG00000198561	8(1)	17(17)	0.2			
Evi1	ENSMUSG00000027684	ENSG00000085276	8(1)	19(26)	0.2			
Slc15a4	ENSMUSG00000029416	ENSG00000139370	7(1)	3(3)	0.2	Y		
Lpp	ENSMUSG00000033306	ENSG00000145012	7(1)	33(33)	0.2	Y		
Q8BG85_MOUSE	ENSMUSG00000028497	ENSG00000188921	105	10	0.6	Y		
Mbd2	ENSMUSG00000024513	ENSG00000134046	95	1	0.6	Y		
Glis3	ENSMUSG00000052942	ENSG00000107249	84	2	0.6	Y		
Diap3	ENSMUSG00000022021	ENSG00000139734	74	5	0.6			
Mtmr9	ENSMUSG00000035078	ENSG00000104643	70	26	0.6	Y		
Mobkl2b	ENSMUSG00000039945	ENSG00000120162	66	1	0.6			
2610206B13Rik	ENSMUSG00000022120	ENSG00000152193	66	4	0.6			
Lpin2	ENSMUSG00000024052	ENSG00000101577	62	17	0.6			
D18Ertd653e	ENSMUSG00000024544	ENSG00000168675	62	63	0.6	Y		
Elp3	ENSMUSG00000022031	ENSG00000134014	59	9	0.6	Y		
Lig4	ENSMUSG00000049717	ENSG00000174405	59	18	0.6			
Acsl1	ENSMUSG00000018796	ENSG00000151726	51	44	0.6	Y		
Frmd4b	ENSMUSG00000030064	ENSG00000114541	43	8	0.6	Y		
Pim3	ENSMUSG00000035828	ENSG00000198355	43	39	0.6			
Foxp1	ENSMUSG00000030067	ENSG00000114861	42	3	0.6	Y		
Pcaf	ENSMUSG00000000708	ENSG00000114166	42	4	0.6	Y		
Q8BKG9_MOUSE	ENSMUSG00000032035	ENSG00000134954	41	2	0.6	Y		
Fli1	ENSMUSG00000016087	ENSG00000151702	40	8	0.6	Y		
Cd38	ENSMUSG00000029084	ENSG00000004468	40	20	0.6	Y		
Prdm10	ENSMUSG00000042496	ENSG00000170325	39	10	0.6			
Dnmt2	ENSMUSG00000026723	ENSG00000107614	39	31	0.6	Y		
IGHA_MOUSE	ENSMUSG00000054328	ENSG00000177199	38	94	0.6			
Pde10a	ENSMUSG00000023868	ENSG00000112541	37	15	0.6	Y		
Myh9	ENSMUSG00000022443	ENSG00000100345	37	21	0.6			
Lef1	ENSMUSG00000027985	ENSG00000138795	36	6	0.6			
BC024806	ENSMUSG00000039048	ENSG00000110074	36	8	0.6			
Arhgap18	ENSMUSG00000039031	ENSG00000146376	35	4	0.6	Y		
Ptpre	ENSMUSG00000041836	ENSG00000132334	35	5	0.6	Y		
Tube1	ENSMUSG00000019845	ENSG00000074935	35	8	0.6	Y		
Centd1	ENSMUSG00000037999	ENSG00000047365	34	6	0.6			
Scfd2	ENSMUSG00000062110	ENSG00000184178	33	1	0.6	Y		
Kcnab2	ENSMUSG00000028931	ENSG00000069424	33	11	0.6			
Trim2	ENSMUSG00000027993	ENSG00000109654	33	32	0.6	Y		
TCA_MOUSE	ENSMUSG00000041018	ENSG00000166056	32	1	0.6	Y		
Nrap	ENSMUSG00000049134	ENSG00000197893	32	9	0.6	Y		
Sept11	ENSMUSG00000058013	ENSG00000138758	32	29	0.6			
Mcart1	ENSMUSG00000045973	ENSG00000122696	29	16	0.6			
Pip5k1a	ENSMUSG00000024867	ENSG00000107242	29	17	0.6			
Gpr56	ENSMUSG00000031785	ENSG00000159618	27	58	0.6			
Bcl11b	ENSMUSG00000048251	ENSG00000127152	24	17	0.6	Y		
4930402H24Rik	ENSMUSG00000027309	ENSG00000088854	23	16	0.6	Y		
5430432M24Rik	ENSMUSG00000027459	ENSG00000125898	23	40	0.6			
Ddx4	ENSMUSG00000021758	ENSG00000152670	20	15	0.6	Y		
Btla	ENSMUSG00000052013	ENSG00000186265	19	6	0.6	Y		
Trim30	ENSMUSG00000030921	ENSG00000132256	19	20	0.6			
6430601A21Rik	ENSMUSG00000040321	ENSG00000198146	18	17	0.6			
Man2a1	ENSMUSG00000024085	ENSG00000112893	17	9	0.6			
6330442E10Rik	ENSMUSG00000056219	ENSG00000198133	17	14	0.6			
Kif5c	ENSMUSG00000026764	ENSG00000168280	17	32	0.6			
Hivep1	ENSMUSG00000021366	ENSG00000095951	16	6	0.6	Y		
AI875199	ENSMUSG00000018995	ENSG00000137513	16	8	0.6	Y		
Slc30a5	ENSMUSG00000021629	ENSG00000145740	16	10	0.6	Y		
Pde3b	ENSMUSG00000030671	ENSG00000152270	16	11	0.6	Y		
Pnn	ENSMUSG00000020994	ENSG00000100941	16	12	0.6	Y		
Rffl	ENSMUSG00000020696	ENSG00000092871	16	26	0.6			
	ENSMUSG00000021171	ENSG00000117868	16	33	0.6	Y		
Ripk3	ENSMUSG00000022221	ENSG00000129465	16	35	0.6			
Tep1	ENSMUSG00000006281	ENSG00000129566	16	100	0.6			

continued on next page

continued from previous page

Gene name	Mouse Ensembl ID	Human Ensembl ID	Number of cell lines	Genes in minimal region	Copy number	Insertions in translated region?	Singletons only?	Known TSG?
Slco3a1	ENSMUSG00000025790	ENSG00000176463	15	1	0.6	Y		
NP_001019895.1	ENSMUSG00000033147	ENSG00000163393	15	4	0.6	Y		
St3gal6	ENSMUSG00000022747	ENSG00000064225	15	7	0.6	Y		
Slc36a3	ENSMUSG00000049491	ENSG00000186334	15	8	0.6	Y		
Itpr5	ENSMUSG00000030287	ENSG00000123104	15	10	0.6	Y		
	ENSMUSG00000042590	ENSG00000086200	15	36	0.6			
	ENSMUSG00000062252	ENSG00000197753	15	40	0.6			
NP_079558.1	ENSMUSG00000005583	ENSG00000081189	14	2	0.6	Y		
Phf14	ENSMUSG00000029629	ENSG00000106443	14	9	0.6	Y		
2810013C04Rik	ENSMUSG00000066411	ENSG00000173575	14	18	0.6			
Lyn	ENSMUSG00000042228	ENSG00000147507	14	20	0.6			
Cd53	ENSMUSG00000040747	ENSG00000143119	13	2	0.6			
St6galnac3	ENSMUSG00000052544	ENSG00000184005	13	2	0.6	Y		
Grik1	ENSMUSG00000022935	ENSG00000171189	13	6	0.6	Y		
Rab27a	ENSMUSG00000032202	ENSG00000069974	13	8	0.6	Y		
Zfhx1b	ENSMUSG00000026872	ENSG00000169554	13	10	0.6			
A130038L21Rik	ENSMUSG00000021703	ENSG00000164300	13	13	0.6	Y		
Dscr2	ENSMUSG00000022913	ENSG00000183527	13	14	0.6	Y		
1700001D09Rik	ENSMUSG00000010135	ENSG00000121933	13	26	0.6			
Sh3gl3	ENSMUSG00000030638	ENSG00000140600	13	27	0.6	Y		
Sdk1	ENSMUSG00000039683	ENSG00000146555	12	1	0.6	Y		
Hivep3	ENSMUSG00000028634	ENSG00000127124	12	2	0.6	Y		
	ENSMUSG00000021676	ENSG00000145703	12	3	0.6			
Mgat5	ENSMUSG00000036155	ENSG00000152127	12	17	0.6			
Cdc42se2	ENSMUSG00000052298	ENSG00000158985	12	54	0.6			
Wdfy1	ENSMUSG00000004377	ENSG00000085449	11	1	0.6	Y		
Bard1	ENSMUSG00000026196	ENSG00000138376	11	3	0.6			
D12Ertd553e	ENSMUSG00000020589	ENSG00000197872	10	1	0.6			
Nfia	ENSMUSG00000028565	ENSG00000162599	10	1	0.6	Y		
8430438D04Rik	ENSMUSG00000036019	ENSG00000179104	10	4	0.6	Y		
Acvr1	ENSMUSG00000026836	ENSG00000115170	10	4	0.6			
Mpp4	ENSMUSG00000026024	ENSG00000003393	9	4	0.6	Y		
Slc39a11	ENSMUSG00000041654	ENSG00000133195	9	12	0.6	Y		
	ENSMUSG00000053396	ENSG00000185676	8	32	0.6			
Dnmt3a	ENSMUSG00000020661	ENSG00000119772	7	22	0.6	Y		
1110014D18Rik	ENSMUSG00000059586	ENSG00000156831	7	29	0.6	Y		
Ccnl1	ENSMUSG00000027829	ENSG00000163660	7	29	0.6	Y		
Myc	ENSMUSG00000022346	ENSG00000136997	6	3	0.6			
1600014C10Rik	ENSMUSG00000054676	ENSG00000131943	6	12	0.6	Y		
Phf21a	ENSMUSG00000058318	ENSG00000135365	6	19	0.6	Y		
	ENSMUSG00000057788	ENSG00000105671	6	60	0.6	Y		
NM_011210.1	ENSMUSG00000026395	ENSG00000081237	5	3	0.6	Y		
Lrp12	ENSMUSG00000022305	ENSG00000147650	5	6	0.6	Y		
Stxbp4	ENSMUSG00000020546	ENSG00000166263	5	8	0.6	Y		
Galnt14	ENSMUSG00000024064	ENSG00000158089	5	16	0.6	Y		
Meis1	ENSMUSG00000020160	ENSG00000143995	4	3	0.6	Y		
Ccdc19	ENSMUSG00000026546	ENSG00000158710	4	40	0.6	Y		
Wdr7	ENSMUSG00000040560	ENSG00000091157	98(2)	2(3)	0.2	Y	Y	
Rfx3	ENSMUSG00000040929	ENSG00000080298	85(3)	3(3)	0.2		Y	
Nfib	ENSMUSG00000008575	ENSG00000147862	84(4)	2(2)	0.2	Y	Y	
	ENSMUSG00000064286	ENSG00000189076	75(3)	354(6)	0.2	Y	Y	
Htr2a	ENSMUSG00000034997	ENSG00000102468	72(1)	5(5)	0.2	Y	Y	
6430573F11Rik	ENSMUSG00000039620	ENSG00000170941	67(1)	3(11)	0.2		Y	
Gpc5	ENSMUSG00000022112	ENSG00000179399	65(2)	4(2)	0.2	Y	Y	
Flt1	ENSMUSG00000029648	ENSG00000102755	63(1)	6(6)	0.2	Y	Y	
Gas7	ENSMUSG00000033066	ENSG00000007237	57(2)	1(16)	0.2		Y	
Rac1	ENSMUSG00000001847	ENSG00000136238	45(1)	126(22)	0.2	Y	Y	
Park2	ENSMUSG00000023826	ENSG00000185345	42(5)	2(1)	0.2	Y	Y	
Robo1	ENSMUSG00000022883	ENSG00000169855	42(1)	4(6)	0.2	Y	Y	
Htr1f	ENSMUSG00000050783	ENSG00000179097	37(2)	9(9)	0.2		Y	
Il15	ENSMUSG00000031712	ENSG00000164136	34(1)	4(4)	0.2		Y	
Pank1	ENSMUSG00000033610	ENSG00000152782	33(2)	1(1)	0.2	Y	Y	
Slc44a1	ENSMUSG00000028412	ENSG00000070214	25(1)	20(20)	0.2	Y	Y	
D16Ertd472e	ENSMUSG00000022864	ENSG00000154642	24(4)	13(13)	0.2	Y	Y	
Ubxd3	ENSMUSG00000043621	ENSG00000162543	24(1)	5(5)	0.2		Y	
Arfrp2	ENSMUSG00000042348	ENSG00000185305	20(2)	1(7)	0.2		Y	
Col19a1	ENSMUSG00000026141	ENSG00000082293	20(1)	2(15)	0.2	Y	Y	
Accn1	ENSMUSG00000020704	ENSG00000108684	18(1)	1(1)	0.2	Y	Y	
Rhoj	ENSMUSG00000046768	ENSG00000126785	17(1)	14(14)	0.2	Y	Y	
Usp47	ENSMUSG00000059263	ENSG00000170242	16(2)	4(4)	0.2		Y	
4833446K15Rik	ENSMUSG00000058152	ENSG00000198108	16(1)	3(2)	0.2	Y	Y	
Dut	ENSMUSG00000027203	ENSG00000128951	14(1)	7(7)	0.2	Y	Y	
Klf7	ENSMUSG00000025959	ENSG00000118263	13(1)	1(1)	0.2	Y	Y	
Ifngr2	ENSMUSG00000022965	ENSG00000159128	12(1)	12(12)	0.2	Y	Y	
Tmem16f	ENSMUSG00000064210	ENSG00000177119	10(1)	21(21)	0.2	Y	Y	
Lrrk2	ENSMUSG00000036273	ENSG00000188906	10(1)	22(22)	0.2	Y	Y	
	ENSMUSG00000014781	ENSG00000164256	7(2)	2(7)	0.2	Y	Y	
Thada	ENSMUSG00000024251	ENSG00000115970	7(1)	5(5)	0.2	Y	Y	

Table 4.7. Mouse genes that contain retroviral insertions within the transcribed or translated region and are also promising candidates for targets of deletion in human cancer cell lines. "Number of cell lines" is the number of samples in which the gene is deleted. "Genes in minimal region" is the number of genes that co-occur with the CIS gene in the smallest region of deletion. "Copy number" is the minimal threshold below which the gene is deleted. Where the copy number is 0.2, the number of cell lines and number of genes in the minimal deleted region are given for deletions of copy number <= 0.6, with numbers for copy number <= 0.2 being shown in brackets. "Insertions in translated region?" indicates whether any of the insertions are within the translated region of the gene. "Singletons only?" indicates whether the gene contains insertions other than those represented by a single read. "Known TSG?" indicates whether the gene is a recessive cancer gene listed in the Cancer Gene Census.

tumourigenesis (see Section 1.3.3.3). The identification of insertions within the gene provides strong evidence that it does contribute to cancer (see also Section 3.4.3). Deletions of less than 70 Mb encompassing *WWOX* and insertions in *Wwox* are shown in Figure 4.20. In Section 3.4.3, *Foxp1* was proposed as a putative tumour suppressor gene. Deletion of *FOXP1* was observed in 42 cell lines, with a minimal amplified region of 3 genes, therefore providing additional evidence that this gene contributes to cancer and that it does so in both species. *Mobkl2a* was also presented as a putative tumour suppressor gene in Section 3.4.3, and while the human orthologue of this gene was not deleted in cancer, the human orthologue of paralogue *Mobkl2b* was deleted. Another implicated tumour suppressor gene identified in this analysis was *DOCK8*, which is deleted and under-expressed in human lung cancers (Takahashi *et al.*, 2006).

Known oncogenes *EVI1*, *MYC* and *FLI1* were also identified in the analysis, demonstrating that the results must be viewed with caution and that functional validation, as well as analysis of the distribution of insertions within the mouse candidate, is required to determine whether deletion of the identified genes is likely to contribute to tumourigenesis. Other candidates that have been implicated as oncogenes include *GRM1*, which plays an important role in the transformation of melanocytes in melanoma (Shin *et al.*, 2008), *RASGPR1*, which contributes to tumour progression in murine skin cancer (Luke *et al.*, 2007; Oki-Idouchi and Lorenzo, 2007) and, as mentioned in the previous sections, *MAP3K8*, *MPP4* and *MEIS1*. Likewise, amplification and overexpression of genes encoding cyclin L1 (*CCNL1*), low-density lipoprotein receptor-related protein 12 precursor (*LRP12*) and glypican-5 (*GPC5*) have been demonstrated in human head and neck squamous cell carcinomas (Muller *et al.*, 2006; Redon *et al.*, 2002), oral squamous cell carcinomas (Garnis *et al.*, 2004) and rhabdomyosarcomas (Williamson *et al.*, 2007), respectively. It is therefore likely that other genes are the targets of deletion in the regions containing these known and implicated oncogenes. *GRM1* was the only gene for which the minimal deleted region did not contain additional genes. However, this does not prove that *GRM1* must be the critical gene, since deletions affecting upstream and downstream genes may simply overlap at *GRM1*.

The list contains many genes for which there is limited evidence in the literature to suggest that they may act as tumour suppressor genes. The results of this analysis therefore lend further support to these findings. Some of these candidates (*RBMS3*, *PCAF*, *UTRN*, *ANK3*, *ACCN1*, *CDC14B*, *CHES1* and *PARK2*) are briefly discussed

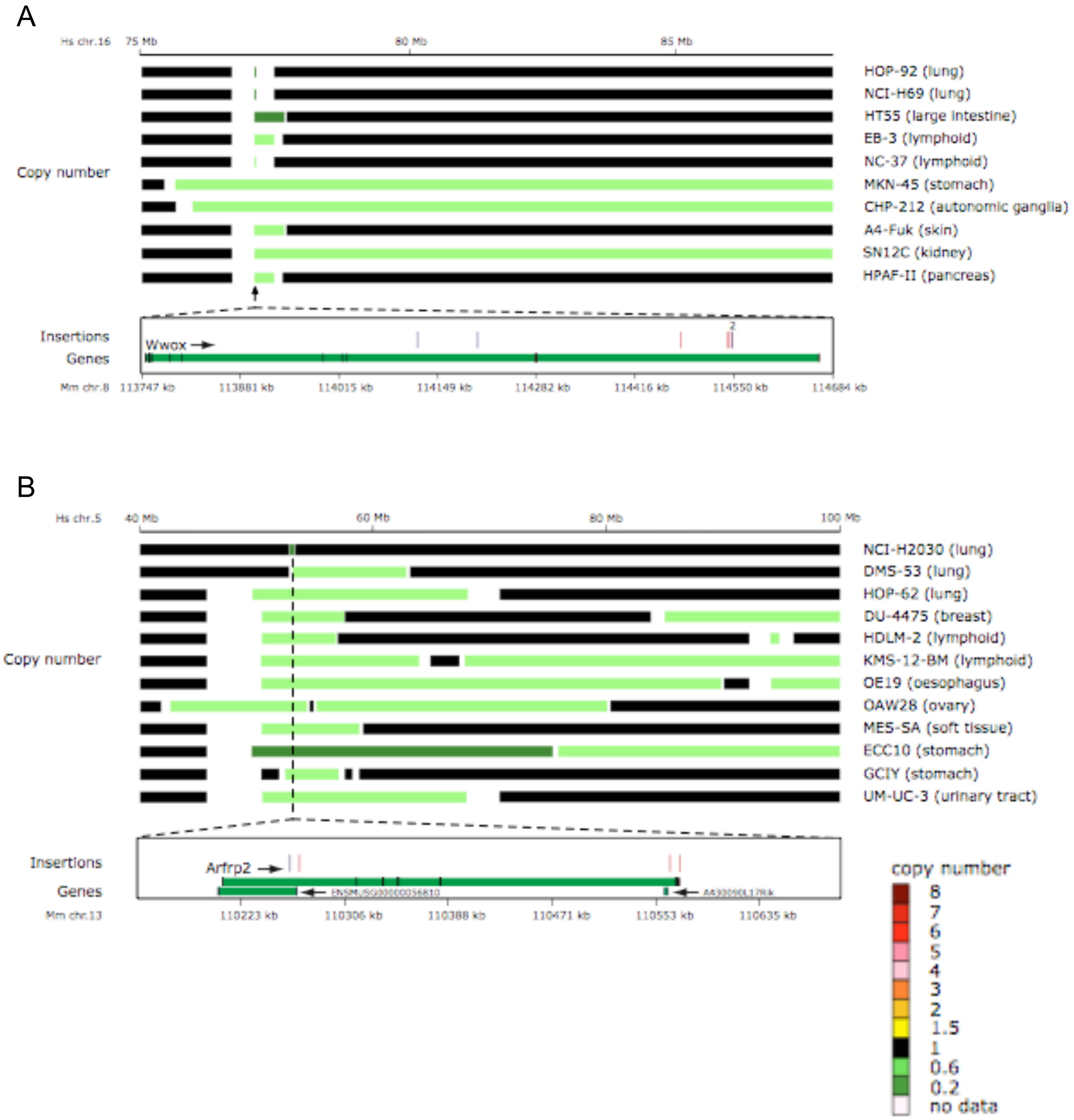

Figure 4.20. Candidate tumour suppressor genes *WWOX* (A) and *ARFRP2* (B) are deleted in human cancer cell lines and are disrupted by retroviral insertional mutagenesis in mouse lymphomas. The copy number of chromosomal regions in the human cell lines is depicted in colour. Names of human cell lines and tissue of origin are provided Only cell lines in which the deletion containing the gene is less than 70 Mb are shown. The lower part of each figure shows insertions within mouse tumours, and was kindly provided by Jaap Kool and Jeroen de Ridder. Blue vertical lines represent insertions in the sense orientation, while red vertical lines represent antisense insertions. Genes are shown in green, with exons marked in black. Positions on the murine and human chromosomes are indicated on the black horizontal bars in kb and Mb, respectively. These figures can also be seen in Uren *et al*. (2008).

below. *RBMS3* and *PCAF* reside in commonly deleted regions, and are down-regulated, in oesophageal squamous cell carcinomas (Qin *et al.*, 2008). One of the homozygous deletions that contained *RBMS3* was from an oesophageal cancer cell line (COLO-608N), but the remaining three were from the large intestine (NCI-H747), ovary (TYK-nu) and cervix (SKG-IIIa). The single homozygous deletion of *PCAF* was in a biliary tract cell line (EGI-1), and neither gene was deleted below copy number 0.6 in any additional oesophageal cancer cell lines. This suggests that the genes may also contribute to other cancers. The utrophin gene (*UTRN*) resides within a deletion of the long arm of chromosome 6 that is frequently observed in a range of tumours, and *UTRN* has been recently proposed as a putative tumour suppressor gene within this region (Li *et al.*, 2007b). Ankyrin-3 (*ANK3*) is a target of the transcription factor hepatocyte nuclear factor 4 alpha that down-regulates cell proliferation in kidney cells (Grigo *et al.*, 2008). None of the 29 deletions containing *ANK3* were within cell lines derived from kidney cancer, but the fact that this was the only gene in the minimal deleted region provides support for a role in tumour suppression. *ACCN1* was proposed as a putative glioma tumour suppressor gene following the observation that surface expression of one of the two isoforms reduces cell growth and migration (Vila-Carriles *et al.*, 2006), while the gene was also shown to be disrupted by a translocation within a neuroblastoma (Vandepoele *et al.*, 2008). Notably, the single homozygous deletion containing this gene was within a glioma cell line (8-MG-BA), while 3 of the remaining 17 deletions of copy number less than or equal to 0.6 were within neuroblastomas. The rest of the deletions were in a range of tumours, including 3 breast, 2 bone, 2 lung and 2 ovarian. *CDC14B* and *CHES1* are both involved in regulating cell cycle checkpoints related to DNA damage response (Bassermann *et al.*, 2008; Busygina *et al.*, 2006), and the deletion of these genes could therefore contribute to tumourigenesis by allowing damaged cells to enter mitosis. Like *WWOX*, the Parkin gene (*PARK2*) resides within a common fragile site (FRA6E) and, therefore, while the gene is frequently deleted in cancer, it is unclear whether it contributes to cancer development. However, deletions involving *PARK2* are associated with ovarian cancer (Denison *et al.*, 2003) and glioblastoma multiforme (Mulholland *et al.*, 2006), and promoter hypermethylation of *PARK2*, resulting in down-regulation of gene expression, has been observed in leukaemias (Agirre *et al.*, 2006). *PARK2* is a long gene, measuring 994.53 kb, and contains just 2 insertions that could have occurred by chance. Therefore, the presence of insertions within the gene does not provide convincing support for a role in tumourigenesis. Interestingly, a break in FRA6E was associated with poor outcome in breast carcinomas, but expression of *PARK2* was not associated, while the loss of *AF-6* gene, which is telomeric of *PARK2*, was associated, suggesting that this may be a tumour suppressor gene affected by the break (Letessier *et al.*, 2007). Other candidates for which there is evidence in the literature of a tumour suppressive role in cancer include *BARD1*, *DMXL1*, *GPR56*, *HIVEP1*, *KCNAB2*, *LEF1*, *LIG4*, *PHF14*, *RAD51L1* and *RIPK3*. Further candidates *SDK1*, *BCL11B* and *MBD2* are discussed in Section 5.3.2.2.

ARFRP2 is a novel candidate tumour suppressor gene for which there is currently no evidence in the literature for a role in cancer. ARFRP2, also known as ARL15, is a member of the ADP-ribosylation factor-like family. Another member of this family, *ARL11*, is a tumour suppressor gene for which truncating germline mutations and promoter methylation contribute to leukaemia, breast cancer, ovarian cancer and melanoma (Frank *et al.*, 2005; Petrocca *et al.*, 2006). Deletions of less than 70 Mb that encompass *ARFRP2* are shown in Figure 4.20. There is also no evidence in the literature to suggest that the sec1 family domain containing gene *SCFD2* is a tumour suppressor gene. However, *SCFD2* is a transcriptional target of p53 (Krieg *et al.*, 2006), and it is the only gene within the minimal deleted region of 33 cancer cell lines.

4.6 Comparison of methods for calling gains and losses

As discussed in Section 4.3, DNAcopy and MergeLevels were the algorithms chosen for detecting regions of copy number change because they had been shown to perform better than other methods, and were freely available. However, it is not known whether DNAcopy and MergeLevels out-perform other methods in processing copy number data generated on the 10K SNP array CGH platform, and a variety of methods were therefore compared. The methods tested were DNAcopy alone (Olshen *et al.*, 2004), DNAcopy and MergeLevels (Olshen *et al.*, 2004; Willenbrock and Fridlyand, 2005), FASeg (Yu *et al.*, 2007), BioHMM (Marioni *et al.*, 2006) and a selection of the methods included within ADaCGH (Diaz-Uriarte and Rueda, 2007), i.e. CGHseg (Picard *et al.*, 2005), HMM (Fridlyand *et al.*, 2004), Wavelets (Hsu *et al.*, 2005) and GLAD (Hupe *et al.*, 2004).

27 different runs of DNAcopy version 1.4.0 were performed, each time varying the parameters. Alpha values of 0.1, 0.05, 0.01, 0.005 and 0.001 were tested, change-points that differed by less than 1, 2, 3 or 4 standard deviations were removed or all change-points were retained, and the smoothing step was either performed or was omitted from

the process (see Section 4.3 for details of these parameters). A further 17 runs of DNAcopy plus MergeLevels were performed, with various combinations of values for the DNAcopy parameters and the Wilcoxon and Ansari-Bradley thresholds within MergeLevels. The Wilcoxon rank sum test is used to determine whether there is a significant difference (according to the Wilcoxon threshold) between the observed values for two copy number levels, or whether they should be merged. The Ansari-Bradley 2-sample test determines whether there is any significant difference between the distribution of merged values minus observed $\log_2$-ratios (i.e. the original ratios at individual SNPs) compared with the distribution of original segmented values minus observed $\log_2$-ratios. The optimal Ansari-Bradley threshold is the largest threshold where the distributions do not differ significantly (Willenbrock and Fridlyand, 2005).

BioHMM is available as part of the BioConductor/R package, snapCGH. It is the only method that takes into account the distance between clones (or in this case SNPs), rather than simply ordering the clones or SNPs along the chromosome. BioHMM uses a Hidden Markov Model to segment data into a finite number of hidden states, where all of the data-points within a state have an equivalent copy number (Marioni *et al.*, 2006). A single run of BioHMM version 1.2.0 was performed using default parameters.

ADaCGH (analysis of data from aCGH) is a web-based tool that provides a selection of the best-performing methods via a simple user interface. DNAcopy and MergeLevels are available within this tool, but it is only possible to use default parameters and the MergeLevels output has been post-processed into three states: -1 (loss), 0 (no change) and 1 (gain). Methods within ADaCGH were chosen because they have been shown to perform well in the comparisons by Lai *et al.* (2005) and Willenbrock and Fridlyand (2005) and/or because they help to present a cross-section of the types of algorithm available for detecting copy number changes. CGHseg models the CGH data as a random Gaussian process and segments the data at points where the mean $\log_2$-ratio changes abruptly. A threshold must be set for the adaptive penalisation, which is a threshold used to estimate the number of segments in the data. Picard *et al.* (2005) proposed a threshold of -0.05 as the default value, but Diaz-Uriarte and Rueda (2007) found that values around -0.005 were more appropriate but recommended experimenting with different values, which must be less than 0. For this analysis, 5 runs of CGHseg were performed, using thresholds of -0.005, -0.01, -0.05, -0.1 and -0.2. The smoothing approach of Hsu *et al.* (2005) uses wavelets to “denoise” the DNA copy number data and so to capture copy number changes while smoothing out the noise. HMM is another method in which Hidden Markov Models are fitted to the data to identify different states, or copy number levels (Fridlyand *et al.*, 2004). However, unlike BioHMM, it does not take account of distances between data-points. Finally, the detection of breakpoints in GLAD is based on the Adaptive Weights Smoothing (AWS) procedure. This method finds the maximal neighbourhood around each data-point in which the local constant assumption holds true. In other words, it finds regions within which the copy number does not differ significantly and the boundaries of these regions represent breakpoints where the copy number changes. Default parameters were used for GLAD, HMM and the wavelets approach. All runs were performed in December 2007 on the website http://adacgh.bioinfo.cnio.es/.

FASeg, or Forward-Backward Fragment-Annealing Segmentation, is available as an R package from http://www.sph.emory.edu/bios/FASeg/. It is proposed to be especially suitable for SNP array CGH, which has a higher probe density but lower signal-to-noise ratio than traditional array CGH. According to the developers, the performance of FASeg was superior to 6 R packages, including DNAcopy, GLAD, BioHMM and CGHseg, in the detection of small segments with a low signal-to-noise ratio, although GLAD and BioHMM also performed well when the signal-to-noise ratio was low and the segments flanking copy number changes were long. When the signal-to-noise ratio was high, most methods performed well, although the HMM-based methods were less effective when there were multiple copy number levels within a single chromosome. This is a significant drawback, since multiple states are common in unstable cancer genomes. FASeg breaks each chromosome into small segments in an over-sensitive edge (or breakpoint) detection step that involves LOESS smoothing. It then iteratively merges consecutive segments until all remaining edges pass a significance threshold, based on testing for equal means between the groups of copy number values for SNPs before and after the edge using the unpaired Student's *t*-test. 15 different runs of FASeg version 1.2 were performed, in which parameters were altered for the smoothing span, which is the number of SNPs used to calculate the weights around each probe in the LOESS smoother, and the *P*-value cut-off for defining the significance of each edge. (See Yu *et al.*, 2007)

In total, 69 different method and/or parameter combinations were compared. Each method was performed on the same 50 randomly selected cancer cell lines. The results were compared using Matthew's Correlation Coefficient (MCC), which is described in

Section 2.10.2. 280 Ensembl genes corresponding to known oncogenes involved in translocations or amplifications were extracted from the Cancer Gene Census. The number of known oncogenes and the number of other Ensembl genes within, and outside of, amplicons of copy number greater than or equal to 2.7 were counted. Oncogenes and other genes within amplicons were defined as true positives and false positives, respectively. Oncogenes and other genes that were not within amplicons were defined as false negatives and true negatives, respectively. The numbers of true and false positives and negatives in each cell line were then added together to give the number across all cell lines, and the MCC score was calculated. This analysis was performed individually on each method. It is possible that some of the known oncogenes that are involved in translocations are not amplified in human cancer, and of course there will be a proportion of non-oncogenes that are amplified in, and contribute to the development of, cancer. However, this analysis gives an indication of the performance of the method in comparison to other methods. The coverage was defined as the proportion of known oncogenes that were represented in amplicons, and the accuracy was defined as the proportion of genes in amplicons that were known oncogenes. The coverage, accuracy and MCC score for each method are shown in Table 4.8.

The wavelet, HMM and BioHMM algorithms all performed poorly. In the case of HMM and BioHMM, this may reflect the fact that there are often multiple copy number levels within a chromosome (see above). The low signal-to-noise ratio may account for the poor performance of the wavelet approach, since this method involves "denoising" the data but was developed for conventional array CGH data, which has a higher signal-to-noise ratio. Denoising the SNP CGH data may result in the removal of biologically relevant copy number changes. In addition, only the default parameters were used for this method. Changing the penalty constant in CGHseg made a considerable difference to the number of amplicons that were detected. This demonstrates the importance of choosing suitable parameter values. The closer the value was to 0, the greater the number of amplicons and the higher the coverage. However, the accuracy fell considerably. The default parameter value of -0.05 gave the best overall results, but this was lower than many of the results obtained using FASeg or DNAcopy. The value suggested in ADaCGH, i.e. -0.005, produced the highest coverage of all methods, but at the expense of a low accuracy. Although only the default parameters were used, GLAD performed reasonably well, obtaining similar results to the best-performing DNAcopy runs.

Method	Parameters	TP	FP	TN	FN	Coverage	Accuracy	MCC
FASeg	p=0.01, smooth=7	31	1027	883573	13969	0.00221	0.02930	0.00380
FASeg	p=0.001, smooth=7	25	830	883770	13975	0.00179	0.02924	0.00340
FASeg	p=0.001, smooth=5	22	757	883843	13978	0.00157	0.02824	0.00301
DNAcopy & MergeLevels	alpha=0.05, smooth, SD=1	25	907	883693	13975	0.00179	0.02682	0.00293
DNAcopy & MergeLevels	alpha=0.1, smooth, w=0.00001	28	1055	883545	13972	0.00200	0.02585	0.00288
DNAcopy	alpha=0.05, smooth, SD=3	17	560	884040	13983	0.00121	0.02946	0.00284
FASeg	p=0.0001, smooth=10	18	608	883992	13982	0.00129	0.02875	0.00281
GLAD		25	931	883669	13975	0.00179	0.02615	0.00279
DNAcopy	alpha=0.01, smooth, SD=2	17	571	884029	13983	0.00121	0.02891	0.00275
DNAcopy	alpha=0.01, smooth, SD=1	18	620	883980	13982	0.00129	0.02821	0.00272
DNAcopy & MergeLevels	alpha=0.1, smooth, SD=1 w=0.00001	28	1087	883513	13972	0.00200	0.02511	0.00271
DNAcopy & MergeLevels	alpha=0.1, smooth, SD=1 w=0.00001, ans=0.01	28	1087	883513	13972	0.00200	0.02511	0.00271
DNAcopy & MergeLevels	alpha=0.1, smooth, SD=1 w=0.00001, ans=0.1	28	1087	883513	13972	0.00200	0.02511	0.00271
DNAcopy	alpha=0.05, smooth, SD=1	23	854	883746	13977	0.00164	0.02623	0.00269
DNAcopy	alpha=0.1, smooth, SD=2	23	855	883745	13977	0.00164	0.02620	0.00268
DNAcopy & MergeLevels	alpha=0.1, smooth, SD=1	28	1103	883497	13972	0.00200	0.02476	0.00263
DNAcopy & MergeLevels	alpha=0.1, smooth, SD=1, ans=0.01	28	1103	883497	13972	0.00200	0.02476	0.00263
DNAcopy & MergeLevels	alpha=0.1, smooth, SD=1, ans=0.1	28	1103	883497	13972	0.00200	0.02476	0.00263
FASeg	p=0.01, smooth=10	24	911	883689	13976	0.00171	0.02567	0.00263
DNAcopy	alpha=0.1, smooth, SD=1	26	1007	883593	13974	0.00186	0.02517	0.00263
DNAcopy	alpha=0.05, smooth	23	865	883735	13977	0.00164	0.02590	0.00262
DNAcopy	alpha=0.05, smooth, SD=4	13	415	884185	13987	0.00093	0.03037	0.00261
DNAcopy	alpha=0.01, smooth	18	636	883964	13982	0.00129	0.02752	0.00260
DNAcopy & MergeLevels	alpha=0.1, smooth	27	1061	883539	13973	0.00193	0.02482	0.00260
DNAcopy	alpha=0.005, smooth, SD=2	16	558	884042	13984	0.00114	0.02787	0.00251
DNAcopy	alpha=0.1, smooth	26	1029	883571	13974	0.00186	0.02464	0.00251
DNAcopy	alpha=0.01, smooth, SD=4	12	383	884217	13988	0.00086	0.03038	0.00251
DNAcopy	alpha=0.05, smooth, SD=2	20	744	883856	13980	0.00143	0.02618	0.00250
DNAcopy	alpha=0.01, smooth, SD=3	14	471	884129	13986	0.00100	0.02887	0.00249
CGHseg	penalty=-0.05	16	561	884039	13984	0.00114	0.02773	0.00249
DNAcopy	alpha=0.001, smooth, SD=2	15	526	884074	13985	0.00107	0.02773	0.00241
FASeg	p=0.001, smooth=10	20	758	883842	13980	0.00143	0.02571	0.00241
DNAcopy	alpha=0.001, smooth	16	575	884025	13984	0.00114	0.02707	0.00238
DNAcopy	alpha=0.001, smooth, SD=4	12	396	884204	13988	0.00086	0.02941	0.00238
DNAcopy	alpha=0.005, smooth, SD=4	12	396	884204	13988	0.00086	0.02941	0.00238
DNAcopy & MergeLevels	alpha=0.01, smooth, w=0.00001	19	717	883883	13981	0.00136	0.02582	0.00237
DNAcopy & MergeLevels	alpha=0.1, smooth, SD=1 w=0.001	26	1059	883541	13974	0.00186	0.02396	0.00235
DNAcopy & MergeLevels	alpha=0.1, smooth, SD=1 w=0.001, ans=0.01	26	1059	883541	13974	0.00186	0.02396	0.00235
DNAcopy & MergeLevels	alpha=0.1, smooth, SD=1 w=0.001, ans=0.1	26	1059	883541	13974	0.00186	0.02396	0.00235
DNAcopy & MergeLevels	alpha=0.1, smooth, SD=2	21	816	883784	13979	0.00150	0.02509	0.00234
DNAcopy	alpha=0.001, smooth, SD=1	16	582	884018	13984	0.00114	0.02676	0.00233
DNAcopy	alpha=0.005, smooth, SD=1	16	587	884013	13984	0.00114	0.02653	0.00229
DNAcopy	alpha=0.001, smooth, SD=3	13	454	884146	13987	0.00093	0.02784	0.00226
DNAcopy & MergeLevels	alpha=0.01, smooth, SD=1	16	593	884007	13984	0.00114	0.02627	0.00225
DNAcopy	alpha=0.005, smooth, SD=3	13	458	884142	13987	0.00093	0.02760	0.00222
FASeg	p=0.0001, smooth=7	18	694	883906	13982	0.00129	0.02528	0.00221
DNAcopy	alpha=0.005, smooth	16	603	883997	13984	0.00114	0.02585	0.00218
DNAcopy & MergeLevels	alpha=0.005, smooth, w=0.00001	15	568	884032	13985	0.00107	0.02573	0.00209
FASeg	p=0.1, smooth=10	28	1261	883339	13972	0.00200	0.02172	0.00188
CGHseg	penalty=-0.01	37	1751	882849	13963	0.00264	0.02069	0.00184
DNAcopy & MergeLevels	alpha=0.05, smooth, w=0.00001	24	1119	883481	13976	0.00171	0.02100	0.00156
DNAcopy	alpha=0.01	21	961	883639	13979	0.00150	0.02138	0.00155
DNAcopy	alpha=0.001	16	703	883897	13984	0.00114	0.02225	0.00152
DNAcopy	alpha=0.005	19	891	883709	13981	0.00136	0.02088	0.00136
DNAcopy	alpha=0.05	24	1169	883431	13976	0.00171	0.02012	0.00134
CGHseg	penalty=-0.01	5	187	884413	13995	0.00036	0.02604	0.00123
FASeg	p=0.1, smooth=5	33	1710	882890	13967	0.00236	0.01893	0.00119
CGHseg	penalty=-0.2	2	72	884528	13998	0.00014	0.02703	0.00084
DNAcopy	alpha=0.1	23	1302	883298	13977	0.00164	0.01736	0.00055
BioHMM		15	849	883751	13985	0.00107	0.01736	0.00045
CGHseg	penalty=-0.005	42	2492	882108	13958	0.00300	0.01657	0.00043
FASeg	p=0.01, smooth=5	8	438	884162	13992	0.00057	0.01794	0.00042
FASeg	p=0.001	4	264	884336	13996	0.00029	0.01493	-0.00009
FASeg	p=0.0001	4	264	884336	13996	0.00029	0.01493	-0.00009
FASeg	p=0.000001	4	264	884336	13996	0.00029	0.01493	-0.00009
FASeg	p=0.0001, smooth=50	5	370	884230	13995	0.00036	0.01333	-0.00037
Wavelets		25	1769	882831	13975	0.00179	0.01394	-0.00059
HMM		9	717	883883	13991	0.00064	0.01240	-0.00073
FASeg	p=0.01	5	467	884133	13995	0.00036	0.01059	-0.00092

Table 4.8. Comparison of methods for detecting regions of copy number gain in 50 randomly selected cancer cell lines. Abbreviations for describing parameters are as follows: FASeg: p=significance threshold, smooth=smoothing range; DNAcopy: alpha=parameter alpha, smooth=outliers smoothed, SD=change-points differing by less than X standard deviations removed; MergeLevels: w=Wilcoxon threshold, ans=Ansari-Bradley threshold; CGHseg: penalty=penalty constant. Undefined parameters are default. TP=number of true positives (amplified oncogenes), FP=number of false positives (amplified non-oncogenes), TN=number of true negatives (non-amplified non-oncogenes), FN=number of false negatives (non-amplified oncogenes). Numbers are calculated across all cell lines. Coverage=TP/(TP+FN), Accuracy=TP/(TP+FP). MCC = Matthew's Correlation Coefficient.

Of the runs involving DNAcopy alone, those in which the data were not smoothed before segmentation performed worst. Higher values for the parameter alpha, which result in increased sensitivity, generally performed better due mainly to a higher coverage. For the purposes of the cross-species comparison, higher coverage, even at the expense of lower accuracy, is preferable since the mouse candidate cancer genes help to identify the targets of amplification in the human amplicons, and false positives are therefore likely to be ignored. Reducing the number of standard deviations below which change-points were removed resulted in a higher coverage of oncogenes. This may be because the highest peak of amplification, which often contains the critical cancer gene(s), is more likely to remain distinct from lower level copy number gains and the segment will have a higher mean copy number and will contain fewer amplified passengers. For higher values of alpha, merging the segments using default parameters also resulted in higher coverage. However, upon inspection of the results, it appeared that some oncogenes were lost upon merging, while some were gained. All of the oncogenes that were unique to the run without merging were still amplified in the run with merging, and vice versa, but they did not reach the copy number threshold of greater than or equal to 2.7. This is because merging increases the mean copy number of some segments and decreases the mean copy number of others, in line with the copy numbers of other segments in the genome. This demonstrates why it is useful to use a range of copy number thresholds in the comparative analysis. Changing the Ansari-Bradley threshold from 0.1 to 0.01 made no difference to the results, but lowering the value for the Wilcoxon threshold increased the MCC score. Using a lower value for the Wilcoxon threshold means that a higher proportion of segments will not be considered significantly different from one another and will therefore be merged. However, more detailed analysis suggests that lowering the value may not produce sensible results. For example, using a value of $1.0\text{x}10^{-5}$ rather than $1.0\text{x}10^{-4}$, the segment of copy number 3.00 that contains *CCND1* in the neck squamous cell carcinoma cell line SCC-15 is merged with a segment of copy number 1.78 to give an overall copy number of 1.85. While the oncogene is still amplified, merging of this kind removes the peaks in amplification, which are most likely to harbour the critical targets of amplification. Similarly, a segment of copy number 0.12 on chromosome 4 of the bone cancer cell line CAL-72 is merged with other segments to give a copy number of 0.47. This segment is likely to be a homozygous deletion but is merged with segments that are more likely to represent heterozygous deletions.

Comparison of the FASeg runs showed that using the default smoothing span of 25 rather than a lower value resulted in lower accuracy and coverage and, therefore, a lower MCC score. When a significance threshold of *P*=0.0001 was used, a smoothing span of 10 rather than 50 not only identified more oncogenes (18 rather than 4) but also had tighter amplicon boundaries that still retained the oncogene. For example, lung cancer cell line LC-2-ad and pancreatic cancer cell line HuP-T4 contained amplicons that encompassed the oncogenes *MYC* and *POU5F1*, respectively. Using smoothing spans of 50 and 10, the number of SNPs within the amplicon containing *MYC* was calculated as 17 and 15, respectively, while the number within the amplicon containing *POU5F1* was 102 and 94, respectively. Increasing the significance threshold generally decreased the MCC score. Using a smoothing span of 7, a significance threshold of *P*=0.001 yielded 64 amplicons, while a significance threshold of *P*=0.0001 yielded 43 amplicons. 35 amplicons were identical, while the rest were either missing from the latter run or were shared but spanned a larger region when the threshold was higher. For example, an amplicon in the lung cancer cell line ChaGo-K-1 spanned 1.39 Mb and had a mean copy number ratio of 3.97 using a threshold of *P*=0.0001, and 919.42 kb with a mean copy number ratio of 4.16 using a threshold of *P*=0.001. Likewise, the amplicon encompassing *MYC* in lung cancer cell line LC-2-ad was also larger (6.74 Mb rather than 4.77 Mb) using a threshold of *P*=0.0001 rather than *P*=0.001 and had a lower mean copy number (5.06 rather than 5.49). The amplicons that were missing from the run with a higher significance threshold may still be present, but the mean copy number may not reach the copy number threshold of 2.7 because a larger region, including less amplified or non-amplified SNPs, is defined as the region of copy number change and this dilutes the mean copy number ratio for the entire segment. Overall, using a significance threshold of *P*=0.01 and a smoothing span of 7 appeared to give the best results, with the highest MCC score and highest accuracy and coverage. It is worth noting, however, that the parameter value for the smoothing span is well below that recommended in Yu *et al.* (2007). The results obtained using the top scoring runs from FASeg and DNAcopy plus MergeLevels were compared. 20 cancer genes were identified by both algorithms. 11 were unique to the FASeg output, and 5 were unique to the DNAcopy and MergeLevels output. In most cases, the missing genes were still amplified, but were below the copy number threshold of 2.7. This analysis indicates that the choice of method and parameters can make a considerable difference to the output and involves finding a suitable balance between accuracy and coverage.

4.7 Global comparison of mouse candidate cancer genes and human CNVs

The global comparison method of Section 4.5.1.1 was applied to human CNVs (see Section 4.2.3) and the gene lists from Section 4.2.1. Rather than using copy number thresholds, CNVs were separated into deletions and duplications, which were specified in the original downloaded file. As with previous analyses, the number of deletions/duplications within which each gene resided was counted. For each number of deletions/duplications, the number of mouse candidates was compared to the distribution of randomised genes using the Z-test. The results are depicted in Figure 4.21. None of the gene lists showed over-representation within deletions or duplications. The only positive association was observed for 7 known oncogenes (namely *DDIT3*, *NSD1*, *IRF4*, 2 genes encoding Histone H4, *NUT* and *PDE4DIP*) that were within 32 or more duplications. The association increased as the number of duplications increased, to a maximum of P=2.07x10^{-4} for 5 known oncogenes in 93 or more duplications. This suggests that some oncogenes are amplified in the normal population, and these individuals may have a predisposition to cancer. However, in general, genes involved in cancer were not found within CNVs. In fact, genes nearest to CISs (P<0.001 and P<0.05) and genes with insertions in coding regions were slightly under-represented in deletions, and genes within translated and transcribed regions were highly under-represented in both duplications and deletions. Many of the genes that are involved in oncogenesis are also involved in other important cellular functions, and this may explain why candidate oncogenes are rarely deleted in healthy individuals. Duplication of tumour suppressor genes could also lead to oncogene repression, producing a similar outcome, while deletion of tumour suppressor genes could lead to tumourigenesis. The results show that cells do not tolerate changes in copy number in genes that are important in tumourigenesis.

For each gene list, the number of genes residing within CNV deletions and within regions of copy number loss (less than or equal to a ratio of 0.6) in human cancer cell lines was counted. A 2-tailed Fisher Exact Test was performed to determine whether there was any association between genes found in deletions in normal individuals and deletions in cancer cell lines. The same analysis was performed using CNV duplications and regions of copy number gain (greater than or equal to 2.7). The P-values are provided in Table 4.9. In accordance with the results obtained in the global analysis, there was an under-

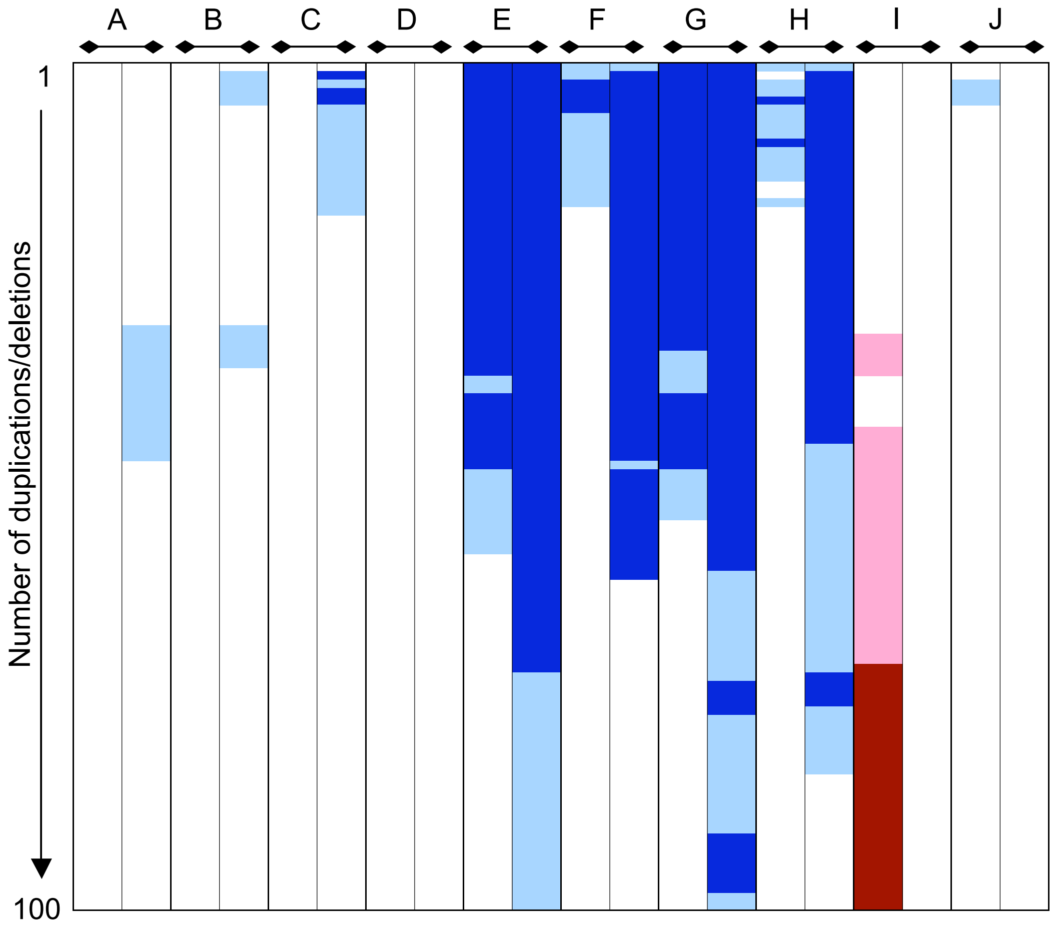

Figure 4.21. Under- and over-representation of human orthologues of candidate cancer genes in regions of copy number variation (CNV). (A) Genes nearest to CISs with *P*<0.001. (B) Genes nearest to CISs with *P*<0.05. (C) Genes with insertions within the coding region. (D) Genes with insertions but no singletons in the coding region. (E) Genes with insertions within the translated region. (F) Genes with insertions but no singletons in the translated region. (G) Genes with insertions in the transcribed region. (H) Genes with insertions but no singletons in the transcribed region. (I) Known oncogenes. (J) Known tumour suppressor genes. For each gene list, the left-hand column represents the significance of the association between the genes and CNV duplications, with rows representing the number of duplications, increasing in increments of 1 to a maximum of 100. Each box in the right-hand column represents the significance of the association between the genes and CNV deletions. *P*<0.01, dark blue for under-representation and dark red for over-representation; *P*<0.05, light blue for under-representation and pink for over-representation.

Gene list	Deletions	Amplicons
ORF only	3.76E-04	0.442
ORF only (no singletons)	6.12E-02	0.290
Translated region only	2.70E-10	0.780
Translated region only (no singletons)	7.02E-06	0.747
Transcribed region only	2.77E-14	0.082
Transcribed region only (no singletons)	6.68E-09	0.178
CIS nearest P<0.05	2.07E-04	5.12E-03
CIS nearest P<0001	2.09E-04	0.229

Table 4.9. *P*-values for the co-occurrence between genes from each gene list within CNVs and regions of copy number change in human cancer cell lines. "Deletions" gives the *P*-values for the co-occurrence of genes in CNV deletions and deletions of copy number less than or equal to 0.6 in human cancers, while "Amplicons" gives the *P*-values for the co-occurrence of genes in CNV duplications and amplicons of copy number greater than or equal to 2.7 in human cancers. *P*-values were calculated using a 2-tailed Fisher Exact Test. All significant *P*-values in "Deletions" represent an under-representation of genes in both CNVs and cancer deletions, while the significant *P*-value in "Amplicons" represents an over-representation of genes in CNVs and cancer amplicons.

representation in all lists of genes that co-occurred in both CNV deletions and deletions in human cancers. There was no association between genes in CNV duplications and copy number gains in human cancers, except for genes nearest to CISs with $P<0.05$, for which more genes than expected co-occurred in CNVs and amplicons. Again, this suggests that amplification of these genes in the general population may confer a predisposition to the development of cancer.

4.8 Discussion

The most significant finding from this chapter is that retroviral insertional mutagenesis is relevant to the discovery of cancer genes in regions of copy number change in human cancers. As anticipated, the overlap is stronger between candidate oncogenes and regions of copy number gain than between candidate tumour suppressor genes and regions of copy number loss. This partly reflects the fact that retroviral insertional mutagenesis predominantly identifies oncogenes due to the major mechanisms by which the retrovirus mutates genes and the requirement for both copies of a tumour suppressor gene to be mutated (see Section 3.4). It may, however, facilitate the identification of haploinsufficient tumour suppressor genes, for which the deletion of one gene copy can contribute to cancer. The other reason for the weaker association between tumour suppressor genes and deletions is that all genes that contained at least one insertion within the transcribed, translated or coding region were included in the analysis. Firstly, this can result in the inclusion of oncogenes that are activated by intragenic truncating mutations (see Section 3.4) and, secondly, many of the insertions may have occurred randomly and may not contribute to oncogenesis. However, the kernel convolution-based method for identifying CISs (de Ridder *et al.*, 2006, see Section 2.10.2) is biased towards oncogenes because insertions within many parts of a tumour suppressor gene may cause its inactivation and therefore insertions may not cluster into tight CISs. For this reason, including all genes provides a more comprehensive list of candidates for a role in tumour suppression.

Significantly, CIS genes were over-represented in amplicons from both haematopoietic and lymphoid cell lines and lines derived from solid tumours. This demonstrates that retroviral insertional mutagenesis is relevant to the discovery of cancer genes in cancers other than lymphomas. This is also proven in the identification of individual candidates, since many were amplified or deleted in a range of cancer types, and some, including

MEIS1, *MMP13* and *ACCN1*, were amplified or deleted in cancer types in which they had previously been implicated. While this study does not include any functional validation, the candidates include a considerable number of known and implicated cancer genes, demonstrating that the method is effective. In general, the discussion of individual genes has focussed on those for which there is some evidence, albeit sometimes limited, that gives cause for presenting the genes as potential oncogenes or tumour suppressor genes. However, the genes listed in Tables 4.4, 4.6 and 4.7 provide a large number of novel candidates that may be of interest to the cancer community. Interestingly, candidate cancer genes were under-represented in CNVs in apparently healthy individuals, further suggesting that amplification and/or deletion of these genes can have a detrimental effect on the cell and, in turn, on the individual.

Despite the promising results, there are a number of potential limitations associated with the analysis. Firstly, all of the human cancers were cell lines, rather than primary tumours. Cancer cells cultured *in vitro* lack the microenvironment of the tumour from which they are derived. While this means that they may not be fully representative of the original tumour, the homogeneity of cell lines can be an advantage since it prevents contamination by stromal cells and potential dilution of the copy number changes identified by CGH. It is, however, possible that the phenotype and genotype of cancer cell lines may differ from those of the original tumour due to genomic instability. Gene expression profiling of lung tumours and cell lines has demonstrated that, in culture, adenocarcinomas progress towards poorly differentiated phenotypes with expression profiles similar to those for squamous cell and small cell lung carcinomas (Virtanen *et al.*, 2002). However, comparisons of human breast and lung cancer cell lines and their corresponding tumours demonstrated an extremely high correlation for both genotype and phenotype, concluding that cell lines from both cancer types are suitable model systems for the original tumours (Wistuba *et al.*, 1998; Wistuba *et al.*, 1999). In addition, gene expression profiles for the NCI60 cell lines, which are the most commonly used cancer cell lines in cancer research and constitute a proportion of the cell lines used in this chapter, also showed that most were representative of their corresponding tumour types (Wang *et al.*, 2006b). Therefore, the use of cancer cell lines is warranted in this analysis, especially as the study is generally concerned with the number of copy number changes affecting a gene, rather than the tissue specificity.

A second potential drawback is that the ploidy of the cancer cell lines is not known. None of the methods used for detecting copy number changes within CGH data can determine the ploidy, and yet aneuploidy is a common characteristic of cancers. Attempts were made to determine the ploidy of cell lines based on the copy numbers of merged segments since, for example, a triploid cell line should only have copy number gains of 1.33, 1.67, 2.00, 2.33, 2.67, and so on, while a tetraploid should have copy number gains of 1.25, 1.5, 1.75, 2, 2.25, and so on. However, the mean copy number ratios for segments are not accurate enough to reliably assign cancers to a particular state. Irrespective of the ploidy, a copy number ratio of 3 indicates that there is a 3-fold increase in the number of copies. In this study, it is assumed that the balance of genes is more important than the actual number, i.e. a 3-fold increase in the number of copies of an oncogene is expected to have the same effect whether the baseline copy number is 2 or 4 genes. In addition, this study is concerned less with the exact copy number of genes, and more with whether genes are amplified or deleted, and the use of a set of copy number thresholds, rather than just one for amplification and one for deletion, ensures that as many candidates as possible are identified.

The analysis does not determine whether an amplified or deleted gene is significantly recurrent. However, genes that are only amplified or deleted in a single cell line may be biologically relevant, as demonstrated for *MEIS1*, and as many different cancer types were used in the analysis, tissue-specific amplicons and deletions may not be significantly recurrent across all cell lines. A gene for which there is no evidence of a role in cancer may not be a convincing candidate if it is amplified or deleted in a single cell line, but the presence of retroviral insertions within the mouse orthologue provides further support. For all candidates, the number of amplicons or deletions containing the gene and the number of additional genes in the minimal amplified or deleted region are provided to help in assessing the contribution of a gene to tumourigenesis. In Chapter 5, efforts are made to make it easier to identify the most promising candidates by ranking genes and assigning a *P*-value based on the number of samples in which they are amplified or deleted. In an attempt to filter out less promising candidates, any genes that were co-amplified with oncogenes or other mouse candidates were removed from the analysis, and yet co-amplified genes may co-operate in tumourigenesis (see Section 1.3.3.3). Nevertheless, given the number of candidates identified, it was considered more important to remove false positives, even at the expense of some “real” cancer genes.

As demonstrated in Section 4.5.1.4, some mouse candidates do not have human orthologues and are therefore excluded from the analysis. In some cases, the human orthologue may not have been identified, while in others, there may not be an orthologue in the human genome. However, the results of the analysis in Section 4.5.1.4 suggest that the proportion of human orthologues may be higher for "real" mouse candidates than for incorrectly assigned candidates. Any discrepancy in the number of mouse genes and the number of human orthologues does not affect the global comparison of Section 4.5.1, since the randomisation takes only mouse genes with human orthologues. This also prevents any introduction of bias resulting from the fact that only protein-coding genes have human orthologues, and that cancer genes are likely to be predominantly protein-coding. Another possible method for comparing the human and mouse data would be to map the insertion sites across to the human genome and then to assign the insertions to human genes. This could be achieved using the Ensembl Compara API, which enables the retrieval of genomic alignments between mouse and human. This would avoid the problem of lack of orthologues but there are many gaps in the alignment, which would prevent the precise mapping of a considerable proportion of insertions. To demonstrate, prior to mapping the retroviral insertions of Chapter 2 and 3 to the NCBI m36 mouse assembly, insertions were mapped to NCBI m34. Only 64.3% of insertions were successfully mapped across to the human genome (NCBI 35) using the Ensembl Compara API. A further drawback of mapping insertions could be that if there really is no human orthologue for a given mouse candidate gene, or there is a break in synteny between mouse and human, the insertions mapped to the human genome will be assigned to an incorrect gene.

The analysis is also limited by the resolution of the data. Efforts have been made to choose suitable boundaries for the ends of amplicons and deletions, but without increasing the density of the SNPs it is impossible to determine whether genes beyond the first or last amplified or deleted SNP are indeed amplified or deleted. It is also possible that small amplicons and deletions may be missed, while the high levels of noise in the data may also lead to regions of copy number change being missed or falsely identified. Encouragingly, the most successful methods for detecting changes produced similar outputs, and the fact that known and implicated oncogenes and tumour suppressor genes were identified, often in cancer types in which they have previously been shown to be disrupted, was also reassuring. However, in Chapter 5, a higher density SNP array is used, and is compared to the 10K array to determine whether it represents a significant improvement.

Chapter 5 Identifying human cancer genes in high-resolution copy number data

5.1 Introduction

Advances in comparative genomic hybridisation (CGH)-based technology have led to the development of higher resolution platforms that can identify smaller amplicons and deletions in cancers, and can more accurately define the breakpoints of regions of copy number change. Higher resolution SNP array CGH platforms have been generated by increasing the density of SNPs from across the genome that are represented on the array. This chapter describes comparative analyses between mouse candidate cancer genes identified by retroviral insertional mutagenesis and copy number data for 598 human cancer cell lines generated by the Wellcome Trust Sanger Institute (WTSI) Cancer Genome Project using high-resolution SNP array CGH. Section 5.2 describes the datasets used in these analyses, and the methods and results are described in Section 5.3. In Section 5.4, the results are compared to those obtained with the 10K CGH data described in Chapter 4 to determine whether there is a significant advantage in using the higher resolution data for integrative analyses. While the analyses in Chapter 4 involved lists of mouse candidate cancer genes that were generated by the Netherlands Cancer Institute, the analyses within this chapter involve candidates identified from the work described in Chapter 2 of this thesis, and different methods have been used to identify interesting candidates. Therefore, in order to directly compare both platforms, the analyses described in Section 5.3 are repeated using the 10K CGH data. Finally, Section 5.5 describes the identification of amplified and deleted human orthologues of mouse candidate cancer genes that co-occur with *TP53* and/or *CDKN2A* mutations, or co-occur with one another, in the human cancer cell lines. These results are then compared to co-occurring CIS genes identified in mouse lymphomas in Section 3.5.2 in an attempt to identify cross-species conservation of co-operation between cancer genes. This chapter represents the culmination of work to characterise the mouse candidate cancer genes described in Chapters 2 and 3, and to demonstrate their relevance to human tumourigenesis. It also provides a clear illustration of how integrative data analysis can facilitate the identification of human cancer gene candidates, which can then be functionally validated in the laboratory.

5.2 Description and processing of the datasets

5.2.1 High-resolution copy number data

Copy number data were generated by the Wellcome Trust Sanger Institute Cancer Genome Project for 598 human cancer cell lines from 29 different tissues (see Table 5.1 and, for more detail, Appendix Table 4.2) using the Affymetrix Genome-Wide Human SNP Array 6.0, which comprises 1.8 million genetic markers for measuring copy number change, of which more than 906,600 are SNPs and more than 946,000 are probes for detecting copy number variation. The genetic markers were mapped to the NCBI 36 human genome assembly. The intensity values were processed into copy number ratios using the method described for the 10K data in Section 4.2.2. This is the point at which I received the data. The analysis only considers autosomes, for which the total number of markers is 1,773,325. The number of copy number markers per chromosome and the distances between adjacent markers are shown in Table 5.2 and Figure 5.1, respectively. The average distance between markers is 1.65 (±41.69) kb, which compares very favourably with the mean distances of 258.50 (±634.21) kb and 292.82 (±683.49) kb for the two 10K arrays used in Chapter 4.

For each cell line, the R packages DNAcopy (Olshen *et al.*, 2004), with default parameters plus removal of change-points less than 2 standard deviations (SD) apart, and MergeLevels (Willenbrock and Fridlyand, 2005) with default parameters, were used to segment each autosome into regions of uniform copy number and to merge segments that were not significantly different across the genome, respectively. DNAcopy and MergeLevels are described in Section 4.3. In this analysis, change-points less than 2 SD, rather than 3 SD, apart were removed because the comparison of methods in Section 4.6 demonstrated that reducing the number of standard deviations should result in the identification of a higher proportion of critical cancer genes. The mean number of segments per cell line was 675.33 (±428.30), which is an average of 24.92 segments per chromosome. Given that DNAcopy and MergeLevels were originally developed for BAC array CGH and, therefore, for dealing with a considerably smaller set of copy number values than the current dataset, it is likely that the genomes are over-segmented. Decreasing the DNAcopy parameter α results in fewer change-points but requires an increased number of permutations, which is unfeasible for a dataset of this size, for which

Site of origin	**Number of cell lines**
Haematopoietic and lymphoid	103
Lung	100
Central nervous system	49
Large intestine	36
Skin	36
Breast	35
Autonomic ganglia	29
Bone	24
Stomach	19
Kidney	18
Ovary	18
Upper aerodigestive tract	18
Soft tissue	17
Oesophagus	15
Pancreas	12
Urinary tract	12
Cervix	11
Thyroid	9
Endometrium	8
Biliary tract	6
Liver	6
Pleura	5
Testis	3
Placenta	2
Prostate	2
Adrenal gland	1
Eye	1
Gastrointestinal tract	1
Small intestine	1
Vulva	1
Total	598

Table 5.1. Tissues of origin of human cancer cell lines used in high-resolution copy number analysis.

Chromosome	Number of markers
1	145591
2	152881
3	127049
4	119457
5	115131
6	112395
7	100581
8	97736
9	81856
10	93272
11	89214
12	86990
13	65757
14	56782
15	53389
16	53920
17	46469
18	51802
19	30236
20	43457
21	24984
22	24376
Total	1773325

Table 5.2. Number of copy number probes per human autosome.

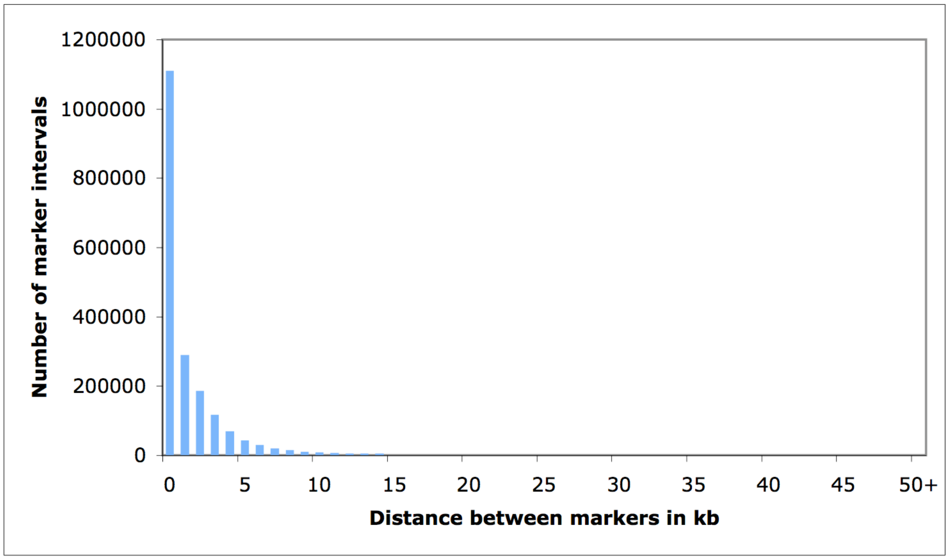

Figure 5.1. Distance between adjacent copy number probes across the human genome.

the CPU time is already very large. However, this analysis does not attempt to elucidate the events that have led to the observed changes in copy number across individual genomes, but focuses instead on individual genes that are amplified or deleted across a significant number of cell lines. Here, the use of DNAcopy and MergeLevels provides a means of filtering out anomalies at individual SNPs, rather than accurately defining entire regions of copy number change.

Since markers are closely spaced, the problems associated with defining amplicon and deletion boundaries that were described in Section 4.5.1.2 are unlikely to arise in this analysis. Therefore, boundaries were simply defined as the halfway point between the first/last amplified or deleted SNP and the preceding/proceeding SNP in the genome. None of the human cancer cell lines selected for use in this study had a shared common ancestor (see Section 4.2.2 for further details).

As in Chapter 4, the terms “gain” and “amplicon” are used interchangeably (see p.162). In the proceeding analyses, an amplification was defined as a gain of copy number greater than or equal to 1.7 and a deletion was defined as a loss of copy number less than or equal to 0.6, or less than or equal to 0.3. A threshold of 1.7 was chosen because this was the lowest copy number at which an over-representation of human orthologues of mouse candidate cancer genes from the insertional mutagenesis screen was observed (*P*=0.00846 for genes amplified in 2 or more cell lines). Likewise, copy number 0.6 was the highest threshold at which an over-representation of orthologues in deleted regions was observed (*P*=0.0289 for genes deleted in 10 or more cell lines). Copy number 0.3 was the highest threshold at which an over-representation of orthologues was observed in 1 or more cell lines (*P*=0.00467), and these may represent homozygous deletions. The method used to generate the above *P*-values is described in Section 5.3.1. The average number of gains of copy number greater than or equal to 1.7 was 34.03 (±36.57) per cell line. The average size of these amplicons was 299.10 (±1667.93) kb and an average of 2.99 (±14.50) genes was found in each amplicon. The average number of losses of copy number less than or equal to 0.6 per cell line was 204.10 (±194.36). These losses were on average 196.87 (±3058.58) kb in size, encompassing 2.61 (±32.98) genes. Deletions were therefore smaller on average than amplicons, suggesting that deletion of gene copies may have a more detrimental effect on a cell than an increase in gene copies. Figure 5.2 shows the distribution of the number of amplicons and deletions in this collection of cell lines and the distribution of the length of aberrations.

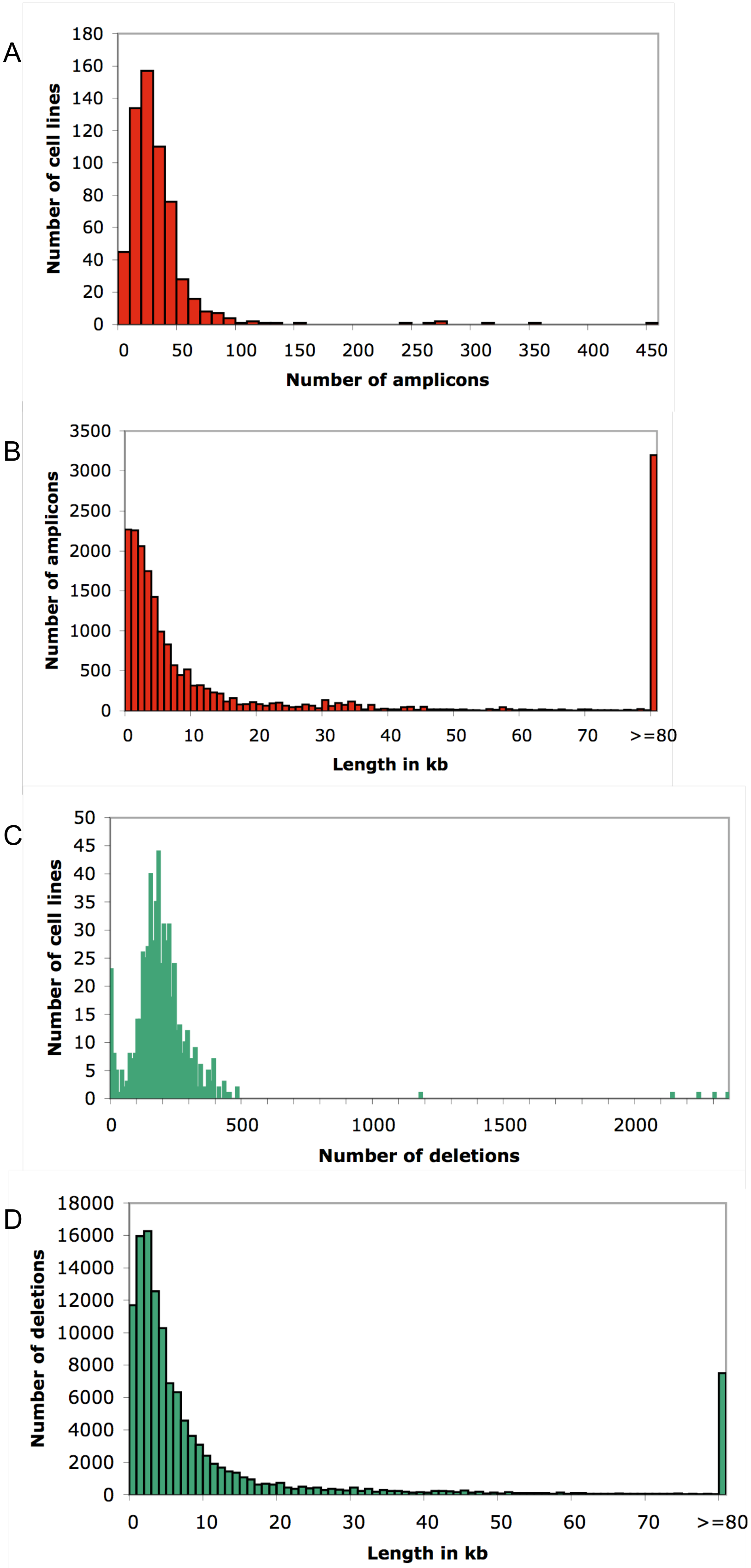

Figure 5.2. Characterisation of amplicons and deletions in 598 human cancer cell lines analysed using high-resolution SNP array CGH. (A) Number of amplicons per cell line. (B) Length of amplicons. (C) Number of deletions per cell line. (D) Length of deletions.

A 2-tailed Fisher Exact Test was used to identify types of cancer that were over- or under-represented among cell lines containing amplicons of copy number greater than or equal to 1.7 and deletions of copy number less than or equal to 0.6. 362 cell lines contained at least one amplicon, while 542 contained at least one deletion. Cell lines derived from cancers of the oesophagus were over-represented among those containing amplicons (*P*=6.79x10^{-4}) while cell lines derived from haematopoietic and lymphoid cancers were under-represented (*P*=3.87x10^{-3}). This is consistent with the results obtained in the 10K analysis described in Section 4.4. Most cell lines contained at least one deletion, and consequently there was no significant difference between the numbers of each cancer type containing deletions.

5.2.2 Additional datasets

The dataset of copy number variants (CNVs) from Redon *et al.* (2006) is described in Section 4.2.3. However, in this chapter, 1,390 CNVs mapping to autosomes on the NCBI 36 human genome assembly, rather than NCBI 35, were used. These are available for download from http://www.sanger.ac.uk/humgen/cnv/data/cnv_data/. Known cancer genes from the Cancer Gene Census are described in Section 2.2.3. The mouse candidate cancer genes used in this chapter were the 439 genes identified in the murine leukaemia virus (MuLV) insertional mutagenesis screen described in Chapter 2. 384 of the 439 candidate cancer genes had human orthologues in Ensembl v48 (see Section 3.2.2). Mouse candidate genes are referred to here as CIS genes. Other datasets referred to in this chapter are described in Section 3.2.1.

5.3 Comparative analysis of human high-resolution CGH data versus mouse insertional mutagenesis data

5.3.1 Global comparison

The number of CIS genes with human orthologues was counted within amplicons above copy number thresholds ranging from 1.1 to 5.0 with increments of 0.1, and within deletions below copy number thresholds ranging from 0.9 to 0.1 with increments of 0.1. The number of non-CIS genes with human orthologues in amplicons and deletions was also calculated, and a 2-tailed Fisher Exact Test was used to determine whether, for each copy number threshold, the number of CIS genes was significantly different to that expected by chance. The number of cell lines in which the genes were amplified or deleted above or below a given threshold was also counted. This global comparison is similar to that described in Section 4.5.1, except that the Fisher Exact Test, rather than the randomisation approach, was used to generate *P*-values. This gives an accurate comparison of the number of CIS genes to the exact number of non-CIS genes, rather than to an estimated number. Copy number increments of 0.1, rather than the smaller set of thresholds from Chapter 4, were used to maximise the amount of information provided by the comparison. Figures 5.3A-C show the pattern of over-represented CIS genes at varying thresholds of copy and cell line number for all cell lines, as well as for those derived specifically from haematopoietic and lymphoid tissues or from solid tumours.

For copy number thresholds of 1.7 and above, there was an over-representation of CIS genes, demonstrating that a significant proportion are amplified in human cancers and may play an important role in human tumourigenesis. The most significant result was a *P*-value of 3.23×10^{-7} for genes within 1 or more cell lines at a copy number greater than or equal to 3.4. In general, the significance increased with increasing copy number. This may reflect the fact that regions of higher copy number are more likely to be tumourigenic, and will therefore contain a higher proportion of candidate cancer genes than amplicons with a lower copy number. In addition, regions amplified to a higher copy number are likely to be more localised, containing fewer genes, and the highest peak in amplification is most likely to harbour the critical gene. The significance generally decreased with increasing numbers of cell lines. This is mainly because fewer CIS genes are amplified across larger numbers of cell lines, and thus there may not be enough power for a significant *P*-value. It may also indicate that regions that are amplified across a large number of cell lines from different types of cancer are less likely to be involved in tumourigenesis. As for the candidate genes in Section 4.5.1.3, the CIS genes were over-represented in amplicons in haematopoietic and lymphoid cancer cell lines but with less significance than for the whole set of samples, and in cell lines derived from solid tumours. The *P*-value for genes within 1 or more solid tumour cell lines at a copy number greater than or equal to 3.4 was 6.29×10^{-7}, which is roughly double that observed for the full set of cell lines, but is still highly significant.

An over-representation of CIS genes was also identified in deletions but with lower significance than for genes in amplicons. The most significant result was a *P*-value of 0.00467 for genes within 1 or more cell lines at a copy number less than or equal to 0.3.

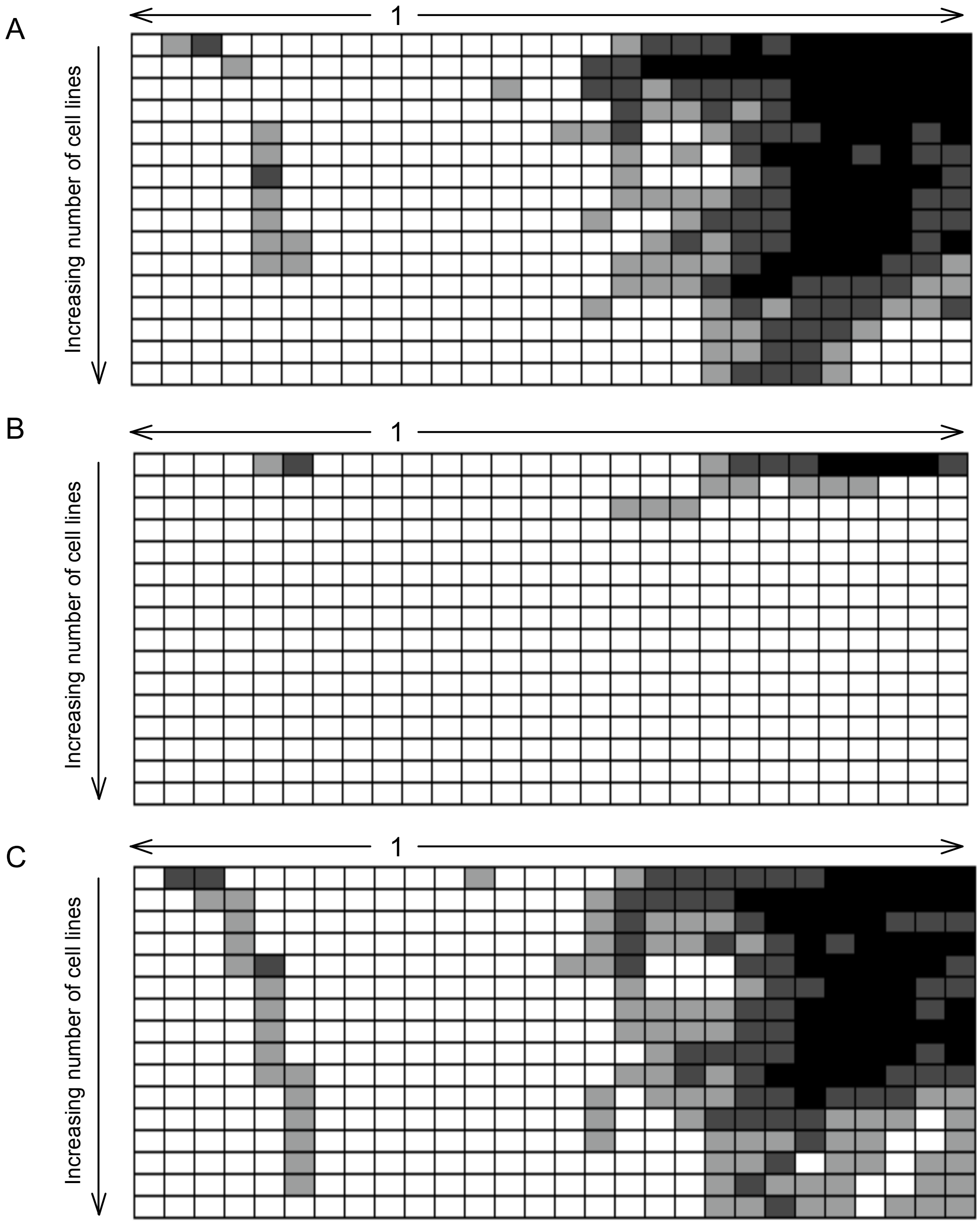

Figure 5.3. Over-representation of CIS genes in amplicons and deletions of varying copy number threshold and number of cell lines across all cell lines (A), haematopoietic and lymphoid cancer cell lines (B), and cell lines derived from solid tumours (C). Each box represents the significance of the association between CIS genes and amplicons/deletions at a given copy number threshold and cell line number. *P*<0.0001, black; *P*<0.001, dark grey, *P*<0.05, light grey. Copy number thresholds below 1 represent deletions, and range from 0.1 to 0.9 with 0.1 increments. Copy number thresholds above 1 represent amplicons, and range from 1.1 to 2.9 with 0.1 increments. The number of cell lines increases in increments of 1, up to a cut-off of 16 cell lines. For example, the box in the bottom right-hand corner represents the *P*-value for the over-representation of CIS genes that occur in amplicons of copy number greater than or equal to 2.9 in at least 16 cancer cell lines.

All other *P*-values were greater than 0.01. As mentioned in Section 4.5.1.3, it is expected that the overlap with deletions will be smaller than with amplicons because most of the CIS genes are likely to function as oncogenes. Deletions in haematopoietic and lymphoid cell lines were only slightly over-represented at copy numbers less than or equal to 0.6 and 0.5, while deletions in cell lines derived from solid tumours showed a similar pattern of over-representation to the full set of cell lines.

5.3.2 Identifying individual cancer gene candidates

5.3.2.1 Methods

The human orthologues of all mouse genes were extracted from Ensembl v48 and the number of amplicons containing each gene was calculated. Non-CIS genes were ranked according to the number of amplicons in which they resided, and a *P*-value was calculated for each CIS gene by counting the number of non-CIS genes within a higher number of amplicons and dividing it by the total number of non-CIS genes. The same procedure was used to calculate *P*-values for deleted CIS genes. These analyses were performed using the full set of cancer cell lines, as well as tissue-specific sets for cancers that were represented by 10 or more cell lines. Where genes contained regions of variable copy number within one cell line, the highest copy number was chosen to represent that gene, and the maximum copy number for that gene across all cell lines was determined. A *P*-value for the maximum peak of amplification of each CIS gene was then calculated by comparing the maximum copy number to that of non-CIS genes using the same method as used for the number of amplicons and deletions.

As for the 10K data in Section 4.5.2, minimal amplified and deleted regions were identified by calculating the coordinates of the smallest overlap of regions that contained the CIS gene. Other genes within the region were identified using the coordinates of human genes in Ensembl v45. For each gene within a minimal amplified region, the total number of cell lines in which that gene was amplified was calculated, and the maximum copy number across all lines was determined. A similar procedure was applied to genes within a minimal deleted region, except that the number of deletions was calculated, and the minimum copy number was determined. The minimal amplified or deleted region within which a CIS gene resides may not necessarily be a minimal amplified or deleted region from across the entire genome. The coordinates of minimal amplified and deleted regions across the genome were therefore determined and were compared to the coordinates of CIS genes to determine whether the CIS genes resided in these regions. To avoid confusion, these regions are known as MCRs (minimal common regions). 8,694 MCRs were identified among amplicons, while 35,213 were identified among deletions of copy number less than or equal to 0.6.

The position of a CIS relative to a CIS gene was calculated using the genomic coordinates and orientation of the mouse gene, extracted from Ensembl v45, and the coordinates of the CIS in the output from the kernel convolution-based method for CIS identification (de Ridder *et al.*, 2006; see Section 2.10.2). The coordinates of the CIS are given as the position of the highest peak in insertion density within a 30 kb kernel.

Global comparisons of the number of amplicons/deletions and maximum copy number in CIS genes versus non-CIS genes were calculated using the Mann Whitney U test, wherein all the values for CIS genes were compared to all the values for non-CIS genes to determine whether the values for CIS genes tended to be greater. A 2-tailed Fisher Exact Test was used to determine whether there was any significant difference between the number of cell lines of different tissue types containing amplicons and deletions. Unless otherwise stated, all other *P*-values were generated using a 1-sided Fisher Exact Test to determine whether genes were over-represented.

5.3.2.2 Candidate cancer genes within amplicons

There were 9,681 mouse genes with human orthologues within the amplicons of human cancer cell lines. Of these, 232 were CIS genes, which is more than expected by chance (P=0.00447). 27 CIS genes were found in statistically significant recurrent amplicons (P<0.05). This included a significant over-representation of genes that were designated dominant cancer genes in the Cancer Gene Census, compared with CIS genes that were not recurrently amplified ($P=2.85\text{x}10^{-4}$). 2,730 mouse genes with human orthologues were found within amplicons in haematopoietic and lymphoid cancer cell lines. CIS genes were over-represented among these genes (71 genes, P=0.0408) but, surprisingly, not as significantly as those amplified in all tumours. This further demonstrates that genes identified by insertional mutagenesis in the mouse may be relevant to the identification of cancer genes in a range of human tumours, not just in those originating in lymphoid tissue. This is at odds with the findings in Section 3.2.2, where the genes showed no association with candidate breast and colon cancer genes from Sjöblom *et al.* (2006). However, the breast and colon candidates were identified by exon resequencing and therefore represent genes mutated by point mutations and indels in cancer. The most common effects of insertional mutagenesis, i.e. gene upregulation by promoter or enhancer insertion, more closely resemble those resulting from changes in copy number. It is therefore more likely that the human orthologues of mouse oncogenes identified by insertional mutagenesis will be disrupted by copy number changes than by small intragenic substitutions and indels. Interestingly, there was no over-representation of genes with mutations in COSMIC among CIS genes that were recurrently amplified, suggesting that copy number changes and small intragenic mutations are not positively associated within candidate cancer genes. All genes in recurrent amplicons across all cell lines, and specifically in haematopoietic and lymphoid cell lines, with a *P*-value of less than 0.1 are shown in Table 5.3.

Of the 27 statistically significant genes, there were 16 where the minimal amplified region contained only that gene. Among the remaining genes, FYN binding protein gene (*FYB*), myeloid cell leukaemia sequence 1 (*MCL1*) and acidic nuclear phosphoprotein 32 family, member E (*ANP32e*) co-occurred with known oncogenes, while all other genes co-occurred with genes that were amplified in more cell lines and, in some cases, to a higher maximum copy number. *FYB* co-occurred with the oncogene leukaemia inhibitory factor, *LIFR*, and all 6 of the additional genes in the minimal amplified region were amplified in a higher number of cell lines than *FYB*, suggesting that *FYB* is not the likely target for amplification in this region. Likewise, *MCL1* and *ANP32e* co-occurred with oncogenes ALL1 fused gene from chromosome 1q (*AF1Q*) and aryl hydrocarbon receptor nuclear translocator (*ARNT*). *AF1Q* is a fusion partner of *MLL* that is involved in leukaemogenesis (Tse *et al.*, 1995), and high expression of *AF1Q* is associated with poor prognosis in paediatric acute leukaemia (Tse *et al.*, 2004). It has also been shown to be overexpressed in thyroid oncocytic tumours, which are a type of tumour characterised by the presence of abundant mitochondria (Jacques *et al.*, 2005), and in breast cancer cells, where it was associated with enhanced proliferation and metastastic potential (Chang *et al.*, 2008; Li *et al.*, 2006a). *ARNT* forms a fusion protein with *ETV6* in acute myeloblastic leukaemia (Salomon-Nguyen *et al.*, 2000) and encodes a component of the transcription factor Hypoxia-inducible factor 1 (HIF1), which is implicated in tumour growth and angiogenesis (for review, see Semenza, 2002). However, the antiapoptotic gene *MCL1*

A

CIS gene	Mouse Ensembl ID	Position of CIS relative to gene	Cancer gene	COSMIC	Mullighan	CNV	Nanog BS	Oct4 BS	p53 BS	Number of amplicons	*P*-value	Gene in MCR?	Number of genes in minimal region	Number of oncogenes in minimal region	Other genes amplified in more cell lines?	Other genes amplified to higher copy?
Wwox	ENSMUSG00000004637	Inside	-	-	-	CNV	Nanog	Oct4	-	69	0.0041	Y	0	0		
Etv6	ENSMUSG00000030199	Inside	Dominant	-	-	-	-	-	-	54	0.0043	Y	0	0		
Myc	ENSMUSG00000022346	Upstream	Dominant	-	-	-	Nanog	-	p53	55	0.0043	Y	0	0		
Mycn	ENSMUSG00000037169	Inside	Dominant	COSMIC	Mullighan	-	Nanog	Oct4	-	25	0.0101	Y	0	0		
Ccnd2	ENSMUSG00000000184	Upstream	Dominant	-	-	-	-	-	-	25	0.0101	Y	0	0		
Ccnd1	ENSMUSG00000070348	Upstream	Dominant	-	-	-	-	-	-	23	0.0111	Y	0	0		
Ikzf3	ENSMUSG00000018168	Inside	-	COSMIC	-	-	-	Oct4	-	22	0.0121	Y	0	0		
Sla	ENSMUSG00000022372	Inside	-	-	-	-	-	-	-	20	0.0132		1	0	Y	Y
Lgals9	ENSMUSG00000001123	Upstream	-	-	-	CNV	-	-	-	17	0.0154	Y	0	0		
Pml	ENSMUSG00000036986	Inside	Dominant	COSMIC	-	CNV	-	Oct4	-	14	0.0212	Y	0	0		
Itpr2	ENSMUSG00000030287	Upstream	-	COSMIC	-	-	-	-	-	14	0.0212	Y	0	0		
Fyb	ENSMUSG00000022148	Inside	-	-	-	-	Nanog	Oct4	-	14	0.0212		6	1	Y	
D12Ertd553e	ENSMUSG00000020589	Downstream	-	-	-	-	Nanog	-	-	14	0.0212	Y	0	0		
Slc1a3	ENSMUSG00000005360	Downstream	-	-	-	-	-	-	-	13	0.0248		9	0	Y	Y
Capsl	ENSMUSG00000039676	Downstream	-	-	-	-	-	-	-	13	0.0248		18	0	Y	Y
Cugbp2	ENSMUSG00000002107	Inside	-	-	Mullighan	-	-	Oct4	-	13	0.0248	Y	0	0		
Sdk1	ENSMUSG00000039683	Downstream	-	-	-	CNV	Nanog	Oct4	-	13	0.0248	Y	0	0		
Trp53inp1	ENSMUSG00000028211	Upstream	-	-	-	-	-	-	-	12	0.0295		6	0	Y	
Ptp4a3	ENSMUSG00000059895	Inside	-	-	-	-	-	-	-	10	0.0365		30	0	Y	Y
Mcl1	ENSMUSG00000038612	Downstream	-	-	Mullighan	-	-	-	-	10	0.0365		20	2	Y	
Erg	ENSMUSG00000040732	Upstream	Dominant	-	-	-	-	-	-	10	0.0365	Y	0	0		
1600014C10Rik	ENSMUSG00000054676	Inside	-	-	-	-	-	-	-	9	0.0447	Y	0	0		
Pag1	ENSMUSG00000027508	Upstream	-	-	-	-	-	-	-	9	0.0447		15	0	Y	Y
Anp32e	ENSMUSG00000015749	Upstream	-	-	Mullighan	-	-	-	-	9	0.0447		54	2	Y	Y
Tpd52	ENSMUSG00000027506	Upstream	-	-	-	-	Nanog	-	-	9	0.0447		15	0	Y	Y
Evi1	ENSMUSG00000027684	Inside	Dominant	COSMIC	-	-	-	-	-	9	0.0447	Y	0	0		
Flt3	ENSMUSG00000042817	Inside	Dominant	COSMIC	-	-	-	-	-	9	0.0447	Y	0	0		
Dock8	ENSMUSG00000052085	Upstream	-	-	-	CNV	-	-	-	8	0.0564	Y	0	0		
Ccnd3	ENSMUSG00000034165	Inside	Dominant	-	-	-	-	-	-	8	0.0564	Y	1	0		
Bcl11b	ENSMUSG00000048251	Inside	Dominant	COSMIC	-	-	-	-	-	7	0.0709	Y	0	0		
Supt3h	ENSMUSG00000038954	Upstream	-	-	-	CNV	-	-	-	7	0.0709	Y	0	0		
Cldn10a	ENSMUSG00000022132	Inside	-	-	-	-	-	-	-	7	0.0709	Y	0	0		
Rorc	ENSMUSG00000028150	Inside	-	COSMIC	Mullighan	-	-	-	-	7	0.0709		8	0	Y	Y
Zfp217	ENSMUSG00000052056	Upstream	-	COSMIC	-	-	Nanog	-	-	7	0.0709		4	0	Y	Y
Kit	ENSMUSG00000005672	Downstream	Dominant	COSMIC	-	-	-	-	-	6	0.0891	Y	0	0		
Ubac2	ENSMUSG00000041765	Inside	-	-	-	-	-	-	-	6	0.0891		8	0	Y	Y
Med13	ENSMUSG00000034297	Upstream	-	-	-	-	-	-	-	6	0.0891		16	2	Y	Y
Cd48	ENSMUSG00000015355	Upstream	-	-	Mullighan	-	-	-	-	6	0.0891	Y	0	0		
Thra	ENSMUSG00000058756	Inside	-	-	-	-	-	-	-	6	0.0891		3	0	Y	Y
Tmem49	ENSMUSG00000018171	Inside	-	-	-	-	-	-	-	6	0.0891		9	1	Y	Y
St6galnac5	ENSMUSG00000039037	Inside	-	-	-	-	-	-	-	6	0.0891	Y	0	0		
Myb	ENSMUSG00000019982	Downstream	-	-	Mullighan	-	-	-	-	6	0.0891	Y	0	0		
Mad1l1	ENSMUSG00000029554	Inside	-	-	-	-	-	Oct4	-	6	0.0891	Y	0	0		

B

CIS gene	Mouse Ensembl ID	Position of CIS relative to gene	Cancer gene	COSMIC	Mullighan	CNV	Nanog BS	Oct4 BS	p53 BS	Number of amplicons	*P*-value	Gene in MCR?	Number of genes in minimal region	Number of oncogenes in minimal region	Other genes amplified in more cell lines?	Other genes amplified to higher copy?
Wwox	ENSMUSG00000004637	Inside	-	-	-	CNV	Nanog	Oct4	-	9	0.0024	Y	0	0		
Myc	ENSMUSG00000022346	Upstream	Dominant	-	-	-	Nanog	-	p53	7	0.0025	Y	0	0		
Cugbp2	ENSMUSG00000002107	Inside	-	-	Mullighan	-	-	Oct4	-	5	0.0052	Y	0	0		
Ccnd2	ENSMUSG00000000184	Upstream	Dominant	-	-	-	-	-	-	5	0.0052	Y	0	0		
Pag1	ENSMUSG00000027508	Upstream	-	-	-	-	-	-	-	4	0.0196		25	0		
Trp53inp1	ENSMUSG00000028211	Upstream	-	-	-	-	-	-	-	4	0.0196		20	0	Y	Y
Tpd52	ENSMUSG00000027506	Upstream	-	-	-	-	Nanog	-	-	4	0.0196		25	0		
Nedd4l	ENSMUSG00000024589	Upstream	-	-	-	CNV	Nanog	-	p53	3	0.0314		9	1		
Ptp4a3	ENSMUSG00000059895	Inside	-	-	-	-	-	-	-	3	0.0314		30	0	Y	Y
Sla	ENSMUSG00000022372	Inside	-	-	-	-	-	-	-	3	0.0314		5	0	Y	
Pml	ENSMUSG00000036986	Inside	Dominant	COSMIC	-	CNV	-	Oct4	-	3	0.0314	Y	0	0		
Bcl11b	ENSMUSG00000048251	Inside	Dominant	COSMIC	-	-	-	-	-	3	0.0314	Y	0	0		
Mcl1	ENSMUSG00000038612	Downstream	-	-	Mullighan	-	-	-	-	3	0.0314		20	2		
Rorc	ENSMUSG00000028150	Inside	-	COSMIC	Mullighan	-	-	-	-	2	0.0478		8	0	Y	
Evi1	ENSMUSG00000027684	Inside	Dominant	COSMIC	-	-	-	-	-	2	0.0478	Y	0	0		
Mbd2	ENSMUSG00000024513	Upstream	-	-	-	-	-	-	-	2	0.0478		34	2	Y	Y
Anp32e	ENSMUSG00000015749	Upstream	-	-	Mullighan	-	-	-	-	2	0.0478		63	2	Y	Y
Ikzf3	ENSMUSG00000018168	Inside	-	COSMIC	-	-	-	Oct4	-	2	0.0478	Y	0	0		
Vpreb2	ENSMUSG00000059280	Upstream	-	-	Mullighan	-	-	-	-	2	0.0478		3	0		

Table 5.3. Lists of CIS genes that are in recurrent amplicons across all cell lines (A) and across haematopoietic and lymphoid cancer cell lines only (B). Cancer gene = known cancer gene in Cancer Gene Census; COSMIC = gene contains somatic mutations in COSMIC database; Mullighan = gene within amplicon in Mullighan *et al*. (2007) dataset of acute lymphoblastic leukaemias; CNV = gene within CNV identified in Redon *et al*. (2006); "[Nanog, Oct4, p53] BS" = gene contains binding site for Nanog, Oct4 and p53, respectively. "minimal region" = minimal amplified region containing CIS gene; "MCR" = minimal amplified region from across genome, not centred on CIS gene; "Number of genes in minimal region" = number of genes other than the CIS gene within the minimal amplified region.

was shown to be amplified and overexpressed in drug-resistant cancer cell lines (Yasui *et al.*, 2004) and in a resistant subline of BL41 Burkitt lymphoma cells (Vrana *et al.*, 2002), and it is overexpressed in a range of leukaemias (Nagy *et al.*, 2003). Therefore, the presence of known oncogenes within a minimal amplified region does not necessarily exclude other genes within the region as candidates for a role in tumourigenesis. As discussed in Section 1.3.3.3, genes within an amplicon may co-operate in tumour development. The protein phosphatase 2 inhibitor *ANP32e* is, however, a less convincing candidate, since high expression of the gene is associated with a better survival rate in patients with follicular lymphoma (Bjorck *et al.*, 2005) and it is downregulated in a progressive, compared to a regressive, murine fibrosarcoma cancer cell line (Hayashi *et al.*, 2005). A recurrent amplicon containing *MCL1* and *ANP32e* was identified in the Mullighan *et al.* (2007) dataset of acute lymphoblastic leukaemias (ALLs; see Table 3.1) and amplification of *MCL1* and *ANP32e* is significantly recurrent in the subset of haematopoietic and lymphoid cell lines. These observations are consistent with the theory that *LIFR*, *AF1Q*, *ARNT*, and possibly *MCL1*, are the critical cancer genes in the amplicon, since all are implicated in leukaemogenesis.

A role in tumourigenesis cannot be ruled out for the 7 amplified genes that co-occurred with genes that were amplified in a greater number of cell lines. For example, the Src-like adaptor gene *Sla*, known as *SLAP* in humans, is upregulated in *FLI1*-transformed erythroblasts (Lebigot *et al.*, 2003), while protein tyrosine phosphatase type IVA, member 3 (*PTP4A3*) is involved in promoting metastasis (for review, see Bessette *et al.*, 2007).

Of the 16 genes that were identified as the only gene in the minimal amplified region, WW domain-containing oxidoreductase (*WWOX*), Aiolos (*IKZF3*) and CUG triplet repeat, RNA binding protein 2 (*CUGBP2*) might be considered more likely to play a role in tumour suppression. Their presence within amplicons therefore suggests that they reside in unstable regions of the genome. It is possible that duplication within the gene may be a mechanism for disrupting the gene, or that the copy number values across a group of markers are erroneous, such that the gene appears to be the target of amplification when in fact it is not. However, as mentioned in Section 3.4.3, *IKZF3* can both activate and repress lineage-specific cells in lymphocytes and may therefore play a dual role in tumourigenesis, and it has been found to be upregulated in chronic lymphocytic leukaemia (Duhamel *et al.*, 2008). Overexpression of *CUGBP2* induces apoptosis of colon cancer cells exposed to radiation (Natarajan *et al.*, 2008) by inhibiting expression of *MCL1*, which is described above (Subramaniam *et al.*, 2008). All of the amplicons in the *CUGBP2* gene were localised to exon 2, suggesting either an error with the 3 markers in this region, or that duplication of this exon may lead to gene disruption (Figure 5.4A). Likewise, the minimal amplified region of *IKZF3* was focused on exon 3, but many amplicons also spanned the entire gene (Figure 5.4B). The presence of amplified *WWOX* may reflect the fact that it resides in the fragile site FRA16D, which is prone to rearrangement in human cancer (see Section 4.5.2.3). Almost all of the amplicons were clustered into 2 distinct regions, one of which contained amplicons spanning up to 19 markers and the other contained amplicons spanning 2 or 3 markers. Both of these regions were within introns, suggesting that they may not affect gene expression. In addition, recurrent amplification of *WWOX* was observed across all cell types, and was significant in 15 of the 19 cancer types for which there were more than 10 cell lines, which suggests that it may reflect some sort of global effect rather than tumourigenicity. As expected for tumour suppressor genes, *Ikzf3*, *Cugbp2* and *Wwox* are all disrupted by intragenic CISs.

As mentioned previously, a significant number of known oncogenes were also discovered in this analysis, demonstrating that the method does successfully identify candidate cancer genes. 10 known oncogenes were also among the 35 CIS genes for which the maximum copy number was significantly higher than for non-CIS genes (P<0.05). All genes for which the maximum copy number has a P-value of less than 0.05 are shown in Table 5.4. After accounting for known oncogenes, the remaining candidates that were recurrently amplified with a P-value of <0.05 included *ITPR2*, *SDK1*, *NP_001035167.2* and *FAM49A* (known as *D12Ertd553e* in the mouse), all of which were the only gene in the minimal amplified region. *FAM49A* co-occurred with *MYCN* in some cell lines, but since this gene was also mutated by insertional mutagenesis, a role in tumourigenesis for the amplified gene cannot be ruled out, and it is within the boundaries of a recurrent amplicon, also containing *MYCN*, that was identified in ALLs by Mullighan *et al.* (2007). *ITPR2* encodes an inositol 1,4,5-trisphosphate receptor that plays an essential role in calcium signalling. Although a role in tumourigenesis has not been confirmed, *ITPR2* is co-amplified with *KRAS2* and *KRAG* in a range of human tumours (Heighway *et al.*, 1996). In the current study, significant recurrent amplification of *ITPR2* was observed in cell lines derived from cancers of the pancreas, colon, ovary, cervix and endometrium. *KRAS2* has been implicated in the development of all of these cancer types. However, the

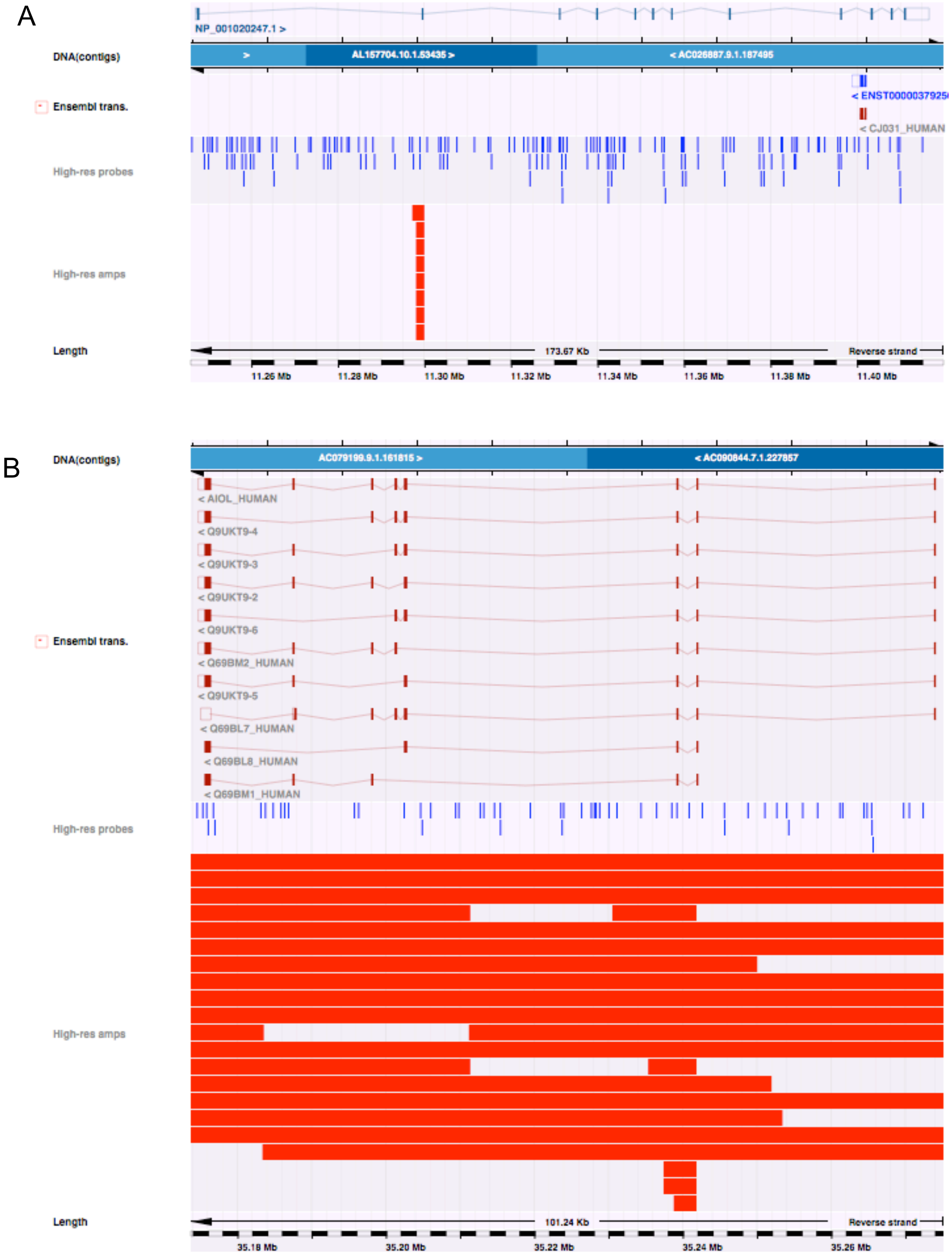

Figure 5.4. The minimal amplified regions within putative tumour suppressor genes *CUGBP2* (A) and *IKZF3* (B) are localised around specific exons. The Ensembl transcripts for *CUGBP2* and *IKZF3* are shown in blue and red, respectively. Copy number markers are shown as blue vertical lines (labelled High-res probes) and amplicons are shown as red rectangles (labelled High-res amps).

		Cancer	Number of	Maximum	
CIS gene	**Mouse Ensembl ID**	**gene**	**amplicons**	**copy number**	***P*-value**
Tgfbr3	ENSMUSG00000029287		5	30.23	0.0030
Fli1	ENSMUSG00000016087	Dominant	1	24.70	0.0054
Mad1l1	ENSMUSG00000029554		6	22.19	0.0069
Cldn10a	ENSMUSG00000022132		7	21.05	0.0091
Ppp1r16b	ENSMUSG00000037754		2	21.05	0.0091
Ptpre	ENSMUSG00000041836		1	14.73	0.0154
A2AN91_MOUSE	ENSMUSG00000038578		1	14.37	0.0161
Mgat4a	ENSMUSG00000026110		2	14.29	0.0161
Stard3nl	ENSMUSG00000003062		3	13.59	0.0169
Anxa2	ENSMUSG00000032231		1	12.12	0.0190
Dym	ENSMUSG00000035765		4	10.95	0.0216
C330024D12Rik	ENSMUSG00000030553		1	10.95	0.0216
Ccnd3	ENSMUSG00000034165	Dominant	8	10.95	0.0216
Ikzf3	ENSMUSG00000018168		22	8.01	0.0281
Dock8	ENSMUSG00000052085		8	7.42	0.0295
Fgfr2	ENSMUSG00000030849	Dominant	5	6.82	0.0308
4932417H02Rik	ENSMUSG00000025583		2	6.76	0.0308
Myc	ENSMUSG00000022346	Dominant	55	6.66	0.0309
Etv6	ENSMUSG00000030199	Dominant	54	6.60	0.0310
Cugbp2	ENSMUSG00000002107		13	6.28	0.0335
Fgd2	ENSMUSG00000024013		3	6.28	0.0335
Mycn	ENSMUSG00000037169	Dominant	25	6.24	0.0338
D12Ertd553e	ENSMUSG00000020589		14	6.23	0.0339
Rara	ENSMUSG00000037992	Dominant	5	5.88	0.0351
Flt3	ENSMUSG00000042817	Dominant	9	5.80	0.0352
Evi1	ENSMUSG00000027684	Dominant	9	5.37	0.0367
Erg	ENSMUSG00000040732	Dominant	10	5.24	0.0369
1110036O03Rik	ENSMUSG00000006931		3	5.10	0.0401
Med13	ENSMUSG00000034297		6	4.91	0.0408
Rcsd1	ENSMUSG00000040723		3	4.80	0.0414
Slamf6	ENSMUSG00000015314		5	4.70	0.0438
Recql5	ENSMUSG00000020752		4	4.62	0.0449
Thra	ENSMUSG00000058756		6	4.62	0.0449
Sla	ENSMUSG00000022372		20	4.58	0.0455
Cyb5	ENSMUSG00000024646		2	4.41	0.0475

Table 5.4. A list of CIS genes for which the maximum copy number across all cell lines is significantly higher than expected by chance.

identification of insertions within *Itpr2*, and a previous study showing that *Itpr2* is targeted by Hepatitis B virus insertional mutagenesis in hepatocellular carcinomas (Paterlini-Brechot *et al.*, 2003), suggests that amplification of *ITPR2* may also contribute to cancer development. In addition, 5 out of 136 human cancer samples tested have a somatic misssense mutation in *ITPR2* in the COSMIC database (Forbes *et al.*, 2006). SDK1, or sphingosine-dependent protein kinase 1, has the same amino acid sequence as the kinase domain of PKCδ (Hamaguchi *et al.*, 2003) and specifically phosphorylates certain isoforms of 14-3-3 that regulate signal transduction and have been implicated as potential oncogenes (Megidish *et al.*, 1998; for review on 14-3-3 proteins, see Tzivion *et al.*, 2006). However, activation of SDK1 leads to apoptosis (Suzuki *et al.*, 2004), suggesting that it has a tumour suppressive, rather than an oncogenic, role in cancer. 5 of the amplicons overlapping *SDK1* were very long, spanning a region of at least 4.4 Mb. The remaining amplicons were small and occurred in intronic regions, suggesting that they may not disrupt the gene. LGALS9 (galectin-9) appears to play dual roles in cancer, since it is associated with antimetastatic potential in breast cancer and oral squamous cell carcinoma cell lines (Irie *et al.*, 2005; Kasamatsu *et al.*, 2005), but it also stimulates phosphorylation of Tim-3, which is implicated in the survival of melanoma cells (Wiener *et al.*, 2007). The amplicons were identified in gene *NP_001035167.2*, which is a paralogue of *LGALS9*. A CNV locus from Redon *et al.* (2006) spans the entire *NP_001035167.2* gene and includes 61 gains in copy number and 5 losses from 270 HapMap individuals. The amplicons in *NP_001035167.2* roughly overlap with the CNV (Figure 5.5). Therefore, while mouse *Lgals9*, and potentially its human orthologue *LGALS9*, may contribute to tumourigenesis, the human paralogue *NP_001035167.2* may not, since it is commonly amplified or deleted in normal individuals as well as in human cancer cell lines. Incidentally, CIS genes are under-represented in CNVs (P=0.0217), which provides support for the fact that they contribute to cancer and are therefore unlikely to be disrupted in healthy individuals.

Of the 71 CIS genes in haematopoietic and lymphoid tumours, 19 were found in statistically recurrent amplicons (P<0.05). 5 of the genes with a P-value of less than 0.05 (i.e. *NEDD4L*, *BCL11B*, *RORC*, *MBD2* and *VPREB2*) were not significantly amplified in the full set of tumours, suggesting that they are specifically associated with cancers of haematopoietic and lymphoid tissue. However, *NEDD4L* and *MBD2* both co-occurred with known oncogene mucosa associated lymphoid tissue lymphoma translocation gene 1 (*MALT1*), while *MBD2* also co-occurred with the B-cell leukaemia/lymphoma 2 gene

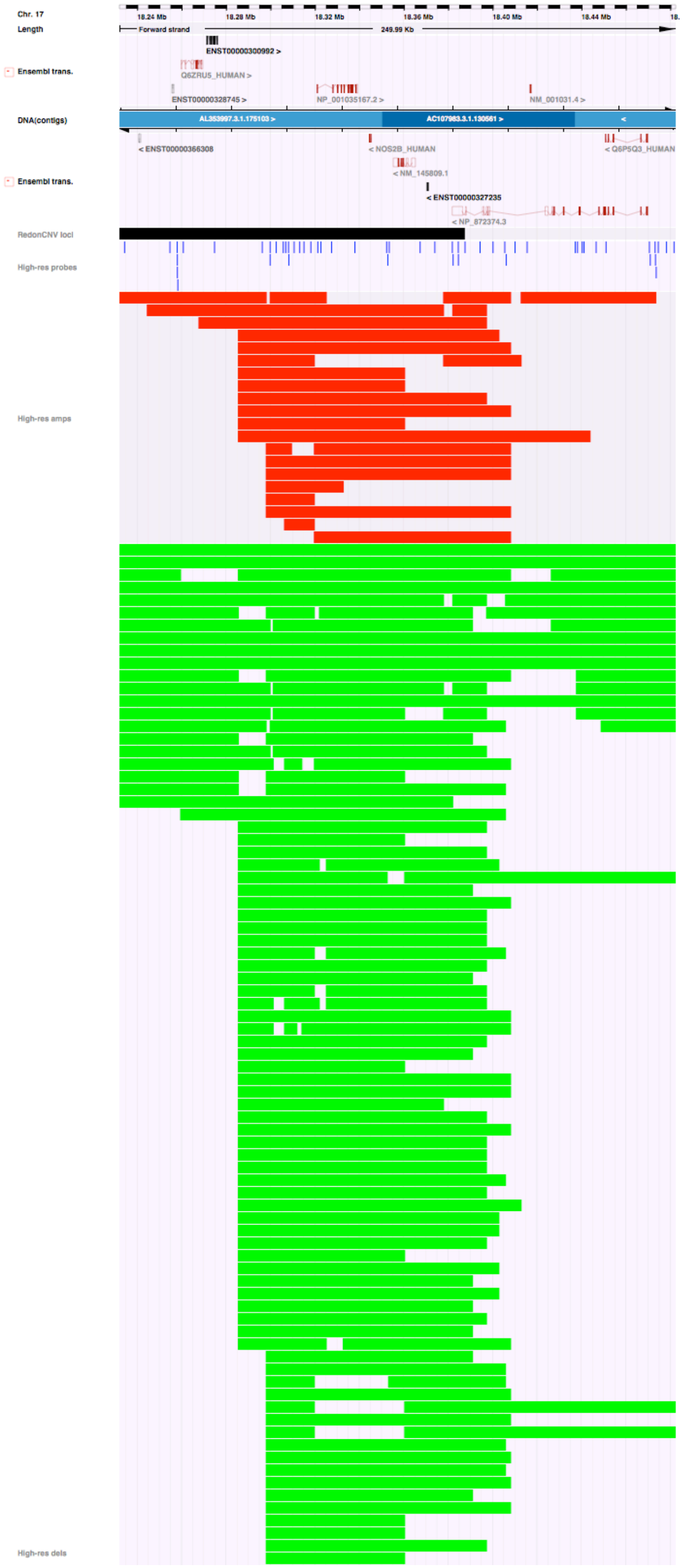

Figure 5.5. Amplicons and deletions in the *LGALS9* paralogue *NP_001035167.2* overlap with a copy number variant (CNV) from Redon *et al.* (2006). The CNV and amplicons and deletions are shown as black, red and green rectangles, respectively. All amplicons, but only around half of all deletions, are shown. Copy number markers are shown in blue.

BCL2. Both *MALT1* and *BCL2* are implicated in lymphomagenesis. The retinoic acid receptor-related orphan receptor C, *RORC*, and the immunoglobulin omega chain precursor, *VPREB2*, have not been previously implicated in cancer, but mutations in *Rorc* were associated with abnormalities in the development of lymphoid organs in immunodeficient mice (Seymour *et al.*, 2006), while *VPREB2* is selectively expressed in pre-B lymphocytes (Kudo and Melchers, 1987; Okabe *et al.*, 1992) and contributes to B-cell development (Dul *et al.*, 1996; Mundt *et al.*, 2001; Shimizu *et al.*, 2002). *RORC* was identified in the recurrent ALL amplicon of Mullighan *et al.* (2007) that also contained *ANP32e* and *MCL1*, while *VPREB2* resides in a recurrent ALL amplicon that is telomeric of *BCR*. However, in this analysis, neither resided within an MCR in haematopoietic and lymphoid cell lines. B-cell leukaemia/lymphoma 11B gene *BCL11B* does reside in an MCR, and is the only gene within it. *BCL11B* encodes a Kruppel-like zinc finger protein that is involved in thymopoiesis and is required for the survival of human T-cell leukaemia and lymphoma cell lines, suggesting an antiapoptotic role in these cancers (Grabarczyk *et al.*, 2007).

Interestingly, 9 of the 27 genes significantly amplified across all cancer types were not amplified in any haematopoietic and lymphoid cell lines, while a further 4 were amplified but not significantly. This may seem surprising since the genes play a role in lymphomagenesis in the mouse, and therefore might be expected to do the same in the human disease. However, it is possible that they are simply not disrupted by amplification in haematopoietic and lymphoid cell lines, which tend to show lower levels of copy number change than cell lines derived from solid tumours. Amplification of ets variant gene 6 (*ETV6*) was significantly under-represented in haematopoietic and lymphoid lines compared with other cancer types (P=3.59x10^{-5}). *ETV6* contributes to human leukaemia through the formation of gene fusions that are not associated with an increase in copy number. There is no evidence to suggest that overexpression of *ETV6* is implicated in tumourigenesis. It resides on, but is not believed to be a critical gene in, an amplicon that is frequently found in breast tumours and osteosarcomas (Gisselsson *et al.*, 2002; Yao *et al.*, 2006). 41 identical amplicons spanning 2 intronic markers were identified, suggesting that the copy number at these markers may be erroneous. *MYCN* is also under-represented in haematopoietic and lymphoid cell lines and is over-represented in cell lines derived from tumours of the autonomic ganglia (P=9.12x10^{-24}). This is consistent with the role of amplified *MYCN* in the development of neuroblastomas. It is possible that genes that are activated by insertional mutagenesis, and contribute to lymphomagenesis, in the mouse might not contribute to the formation of spontaneous mouse or human lymphomas because they are not normally expressed in lymphocytes. However, in the case of *MYCN*, activation by translocation has been demonstrated in non-Hodgkin's lymphoma (Finnegan *et al.*, 1995), and *MYCN* appears to co-operate with the ETS family gene *TEL2* in B-cell lymphomagenesis (Cardone *et al.*, 2005). As mentioned previously, it is also found within a recurrent amplicon in ALL (Mullighan *et al.*, 2007).

Other genes that were significantly over-represented in a particular tissue type were *CCND1* (oesophagus, P=6.51x10^{-6}), *TMEM49* (breast, P=1.37x10^{-4}), *NCOA3* (breast, P=1.85x10^{-4}) and *RCBTB2* (large intestine, P=2.01x10^{-4}). Amplification and overexpression of *CCND1*, or Cyclin D1, has been demonstrated in 32% of human oesophageal squamous cell carcinomas (Jiang *et al.*, 1993) and 64% of oesophageal adenocarcinomas (Arber *et al.*, 1996), while overexpression of nuclear receptor coactivator 3, *NCOA3*, is associated with poor prognosis in breast tumours (Zhao *et al.*, 2003). All 3 amplicons containing *NCOA3* were in breast cancer cell lines. This corresponded to a significant recurrence across breast cancers (P=0.0259) but not across all cell lines (P=0.249), suggesting that *NCOA3* may contribute to tumourigenesis, but only in the breast. However, of the 3 genes that co-occurred with *NCOA3*, 2 (*PRKCBP1* and *EYA2*) were amplified to a higher copy number, and *Prkcbp1* is in fact a CIS gene. 6 cell lines contained a *TMEM49* amplification, of which 4 were derived from breast tumours. Again, this corresponded to a significant recurrence across breast cancers (P=0.0117) but not across all cell lines (P=0.0891). *TMEM49* has not been previously implicated in cancer but it falls within a common region of amplification in breast cancers on chromosome 17q23 that is associated with poor prognosis. Other genes in the minimal amplified region, including *PPM1D*, *APPBP2*, *RPS6KB1* and *BCAS3*, have been implicated in breast cancer development (for review, see Sinclair *et al.*, 2003) and of these, *RPS6KB1* and *BCAS3* were amplified to a higher copy number than *TMEM49* in other cell lines. However, the identification of insertions within *Tmem49* suggests that this gene may also be important in tumourigenesis. All 3 amplicons containing *RCBTB2* were within colon cancer cell lines, which corresponded to a significant recurrence across colon cancers (P=0.0107) but not across all cell lines (P=0.249). *RCBTB2* co-occurs with known tumour suppressor gene *RB1*, and is in fact a candidate tumour suppressor gene in a region on chromosome 13q14 that frequently shows loss of heterozygosity in prostate cancer (Latil *et al.*, 2002). The insertions assigned to *Rcbtb2* are upstream in the sense and antisense orientation, suggesting that the gene plays an oncogenic role, but all insertions are within a longer Ensembl EST gene transcript that is not annotated as an Ensembl gene transcript and could therefore be intragenic, inactivating insertions (Figure 5.6, page 249). *RCBTB2* is not within an MCR, suggesting that it may not contribute to colon tumourigenesis.

Other genes were also amplified in a significant number of cell lines of a particular tissue type but not across all cell lines. Among these were genes encoding suppressor of Ty 3 homolog (*SUPT3H*), claudin-10 (*CLDN10*) and MAD1 mitotic arrest deficient-like 1 (*MAD1L1*), which were significantly amplified in soft tissue (*P*=0.00566), colon (*P*=0.0107) and lung (*P*=0.0267) cancer cell lines, respectively. All 3 genes were within MCRs. SUPT3H is a transcription factor that forms part of a multiprotein complex that mediates transcriptional activation (Brand *et al.*, 1999). Importantly, it is required for transcription and cell proliferation induced by the MYC oncoprotein (Liu *et al.*, 2008). Claudins are components of tight junctions and have been implicated in tumour progression (Kominsky, 2006). *CLDN10* is overexpressed in papillary thyroid carcinoma (Aldred *et al.*, 2004) and overexpression in hepatocellular carcinoma cells promotes cancer cell survival, motility and invasiveness, leading to malignancy (Ip *et al.*, 2007). Claudin-1 plays an important role in cellular transformation and metastasis in colon cancer (Dhawan *et al.*, 2005; Resnick *et al.*, 2005), and overexpression of claudin-7 and claudin-12 is also implicated (Darido *et al.*, 2008; Grone *et al.*, 2007). These findings suggest that amplification of claudin-10 may also contribute to colon cancer, which has not been previously shown. *MAD1L1* is a mitotic checkpoint gene that has also been shown to harbour somatic missense mutations in lung cancer cell lines (Nomoto *et al.*, 1999) and has been presented as a putative tumour suppressor gene (Tsukasaki *et al.*, 2001). However, consistent with the results of the analysis described herein, a region on chromosome 7p22.3 that is centred on *MAD1L1* was found to be the most frequently observed copy number change in small-cell lung cancer cell lines (Coe *et al.*, 2006), suggesting that it may play an oncogenic role that requires further investigation. The maximum copy numbers of *CLDN10* and *MAD1L1* were also significantly higher than for non-CIS genes (*P*<0.05, see Table 5.4).

The number of recurrent amplifications was significantly higher across CIS genes than across non-CIS genes (*P*=0.00396). The median number of amplifications was 1 for both samples, and the maximum number was larger for non-CIS genes, at 115 recurrent amplifications, compared with 69 for CIS genes (specifically *MYC*). However, the mean number was 2.77 for CIS genes compared with 2.30 for non-CIS genes. The maximum peak of amplification was also higher among CIS genes (*P*=0.00135). The minimum and maximum copy numbers were 1.253 (*UBE1L*) and 30.230 (*TGFBR3*) for CIS genes, and 1.227 and 67.960 for non-CIS genes. The median and mean peaks in amplification were 1.832 and 2.692 for CIS genes, and 1.721 and 2.405 for non-CIS genes. These results further suggest that a significant proportion of CIS genes may contribute to human tumourigenesis through the mechanism of amplification.

5.3.2.3 Candidate cancer genes within deletions

16,973 human genes with mouse orthologues, including all but one of the CIS genes, were identified in deletions of copy number less than or equal to 0.6 in human tumours. However, only 24 CIS genes were found in a significant number of deletions (*P*<0.05). Table 5.5A shows all CIS genes with a *P*-value of less than 0.1. Unlike those in amplicons, CIS genes in deletions were not significantly over-represented among known oncogenes (*P*=0.177) compared with other CIS genes. There was also no over-representation of CIS genes with mutations in COSMIC (*P*=0.951).

5 genes (*CCND2*, *ETV6*, *LGALS9*, *SDK1* and *WWOX*) were both significantly amplified and significantly deleted. This is a larger overlap than expected by chance (*P*=0.0196). As discussed in the previous section, deletions within *ETV6*, *WWOX* and *SDK1* are predicted to contribute to tumourigenesis, while the *LGALS9* paralogue *NP001035167.2* shows both gains and losses in copy number in the normal population. The deletions in *NP001035167.2* (Figure 5.5) and *ETV6* were in exactly the same location as the corresponding amplicons, further indicating that the copy number values may be erroneous or due to CNVs. This is also somewhat true of the deletions in *WWOX* and *SDK1*, but *WWOX* and, to a lesser extent, *SDK1* also contained many deletions that spanned other regions, or the entire gene. However, only 2 MuLV insertions were found within *Sdk1*, and the coordinates of the CIS are 161.05 kb downstream of the gene, therefore shedding doubt on a tumour suppressive role for the gene in mouse lymphomas. *CCND2* is a known oncogene that is amplified and overexpressed in a range of human cancers, including malignant gliomas (Buschges *et al.*, 1999), B-cell neoplasms (Werner *et al.*, 1997) and testicular germ cell tumours (Rodriguez *et al.*, 2003). However, its proximity to $p27^{KIP1}$ on the short arm of chromosome 12 means that it is frequently

A

CIS gene	Mouse Ensembl ID	Position of CIS relative to gene	Cancer gene	COSMIC	Mullighan	CNV	Nanog BS	Oct4 BS	p53 BS	Number of deletions	*P*-value	Gene in MCR?	Number of genes in minimal region	Number of TSGs in minimal region	Other genes deleted in more cell lines?	Other genes deleted to lower copy?
Wwox	ENSMUSG00000004637	Inside	-	-	-	CNV	Nanog	Oct4	-	237	0.0000	Y	0	0		
Lgals9	ENSMUSG00000001123	Upstream	-	-	Mullighan	CNV	-	-	-	146	0.0027	Y	0	0		
Etv6	ENSMUSG00000030199	Inside	Dominant	-	Mullighan	-	-	-	-	120	0.0049	Y	0	0		
Sdk1	ENSMUSG00000039683	Downstream	-	-	Mullighan	CNV	Nanog	Oct4	-	83	0.0084	Y	0	0		
Metrnl	ENSMUSG00000039208	Upstream	-	-	-	-	-	-	-	72	0.0101	Y	0	0		
Zfp438	ENSMUSG00000050945	Upstream	-	-	-	-	-	-	-	67	0.0125	Y	0	0		
Midn	ENSMUSG00000035621	Inside	-	-	Mullighan	CNV	-	-	-	66	0.0136	Y	2	0	Y	
Arid3a	ENSMUSG00000019564	Inside	-	-	Mullighan	CNV	-	-	-	65	0.0151		6	0	Y	
Ptbp1	ENSMUSG00000006498	Upstream	-	-	Mullighan	-	-	-	-	65	0.0151		8	0	Y	
Mknk2	ENSMUSG00000020190	Upstream	-	-	-	-	-	-	-	56	0.0197		8	0	Y	
Mobkl2a	ENSMUSG00000003348	Downstream	-	-	-	-	-	-	-	56	0.0197		2	0	Y	
Ets1	ENSMUSG00000032035	Upstream/Downstream	-	COSMIC	-	-	-	-	-	55	0.0202	Y	0	0		
Ccnd2	ENSMUSG00000000184	Upstream	Dominant	-	-	-	-	-	-	53	0.0219	Y	0	0		
Acot11	ENSMUSG00000034853	Upstream	-	-	-	-	-	-	-	52	0.0222	Y	0	0		
Gadd45b	ENSMUSG00000015312	Inside	-	-	-	-	-	-	-	52	0.0222		15	0	Y	
Dym	ENSMUSG00000035765	Inside	-	-	-	-	-	-	-	50	0.0240	Y	0	0		
Notch1	ENSMUSG00000026923	Inside	Dominant	COSMIC	-	CNV	-	-	p53	50	0.0240	Y	0	0		
Gna15	ENSMUSG00000034792	Inside	-	-	-	-	-	-	-	47	0.0281		10	0	Y	
Tbxa2r	ENSMUSG00000034881	Upstream	-	-	-	-	-	-	-	46	0.0293		22	0	Y	
Nedd4l	ENSMUSG00000024589	Upstream	-	-	-	CNV	Nanog	-	p53	43	0.0323	Y	0	0		
Cyb5	ENSMUSG00000024646	Downstream	-	-	-	-	Nanog	-	-	42	0.0351	Y	0	0		
Dock8	ENSMUSG00000052085	Upstream	-	-	Mullighan	CNV	-	-	-	42	0.0351	Y	0	0		
Ntn1	ENSMUSG00000020902	Downstream	-	-	Mullighan	-	Nanog	-	-	42	0.0351	Y	0	0		
Cbfa2t3	ENSMUSG00000006362	Upstream	Dominant	-	-	-	-	-	-	39	0.0455	Y	3	0		
Rtn4rl1	ENSMUSG00000045287	Downstream	-	-	Mullighan	-	-	-	-	37	0.0502		28	1	Y	Y
Slc43a2	ENSMUSG00000038178	Inside	-	-	Mullighan	-	-	-	-	37	0.0502		12	0	Y	Y
Smg6	ENSMUSG00000038290	Inside	-	COSMIC	Mullighan	-	-	-	-	37	0.0502		4	0	Y	
Ovca2	ENSMUSG00000038268	Downstream	-	-	Mullighan	-	-	-	-	36	0.0538		28	1	Y	Y
Ski	ENSMUSG00000029050	Upstream	-	-	-	-	-	-	-	36	0.0538	Y	0	0		
Tcf25	ENSMUSG00000001472	Downstream	-	-	-	CNV	-	-	-	36	0.0538		3	0	Y	
Ttll10	ENSMUSG00000029074	Upstream	-	-	-	-	-	-	-	36	0.0538	Y	27	0	Y	
Vps13d	ENSMUSG00000020220	Inside	-	-	-	-	-	-	-	36	0.0538	Y	0	0		
Arrdc5	ENSMUSG00000073380	Inside	-	-	-	-	-	-	-	35	0.0569		27	0	Y	Y
Rnf166	ENSMUSG00000014470	Inside	-	-	-	-	-	-	-	35	0.0569		18	0	Y	
Mbd2	ENSMUSG00000024513	Upstream	-	-	-	-	-	-	-	34	0.0601	Y	0	0		
Pik3cd	ENSMUSG00000039936	Upstream	-	-	-	-	-	Oct4	-	34	0.0601	Y	1	0	Y	
Prdm16	ENSMUSG00000039410	Inside	Dominant	COSMIC	-	CNV	-	-	-	33	0.0637	Y	0	0		
Foxp1	ENSMUSG00000030067	Inside	-	COSMIC	-	-	Nanog	-	-	32	0.0667	Y	0	0		
BC008155	ENSMUSG00000057411	Inside	-	-	-	CNV	-	-	-	29	0.0772	Y	4	0	Y	
1700081D17Rik	ENSMUSG00000022085	Downstream	-	-	-	CNV	-	-	-	28	0.0824	Y	0	0		
Kit	ENSMUSG00000005672	Downstream	Dominant	COSMIC	-	-	-	-	-	28	0.0824	Y	0	0		
Erg	ENSMUSG00000040732	Upstream	Dominant	-	Mullighan	-	-	-	-	26	0.0927	Y	0	0		
Park7	ENSMUSG00000028964	Upstream	-	-	-	-	-	-	-	25	0.0986		2	0	Y	

B

CIS gene	Mouse Ensembl ID	Position of CIS relative to gene	Cancer gene	COSMIC	Mullighan	CNV	Nanog BS	Oct4 BS	p53 BS	Number of deletions	*P*-value	Gene in MCR?	Number of genes in minimal region	Number of TSGs in minimal region	Other genes deleted in more cell lines?	Other genes deleted to lower copy?
Wwox	ENSMUSG00000004637	Inside	-	-	-	CNV	Nanog	Oct4	-	41	0.0006	Y	0	0		
Lgals9	ENSMUSG00000001123	Upstream	-	-	Mullighan	CNV	-	-	-	22	0.0045	Y	0	0		
Sdk1	ENSMUSG00000039683	Downstream	-	-	Mullighan	CNV	Nanog	Oct4	-	19	0.0088	Y	0	0		
Etv6	ENSMUSG00000030199	Inside	Dominant	-	Mullighan	-	-	-	-	18	0.0094	Y	0	0		
Ntn1	ENSMUSG00000020902	Downstream	-	-	Mullighan	-	Nanog	-	-	10	0.0227	Y	0	0		
Acot11	ENSMUSG00000034853	Upstream	-	-	-	-	-	-	-	9	0.0259	Y	0	0		
Zfp438	ENSMUSG00000050945	Upstream	-	-	-	-	-	-	-	7	0.0343	Y	0	0		
Ccnd2	ENSMUSG00000000184	Upstream	Dominant	-	-	-	-	-	-	7	0.0343	Y	0	0		
Kit	ENSMUSG00000005672	Downstream	Dominant	COSMIC	-	-	-	-	-	7	0.0343	Y	0	0		
Smg6	ENSMUSG00000038290	Inside	-	COSMIC	Mullighan	-	-	-	-	6	0.0518		36	1	Y	Y
Ovca2	ENSMUSG00000038268	Downstream	-	-	Mullighan	-	-	-	-	6	0.0518		36	1	Y	Y
Slc43a2	ENSMUSG00000038178	Inside	-	-	Mullighan	-	-	-	-	6	0.0518		36	1	Y	Y
Rtn4rl1	ENSMUSG00000045287	Downstream	-	-	Mullighan	-	-	-	-	6	0.0518		36	1	Y	Y
Foxp1	ENSMUSG00000030067	Inside	-	COSMIC	-	-	Nanog	-	-	6	0.0518	Y	0	0		
Tbc1d1	ENSMUSG00000029174	Downstream	-	-	-	-	Nanog	-	-	6	0.0518	Y	0	0		
Notch1	ENSMUSG00000026923	Inside	Dominant	COSMIC	-	CNV	-	-	p53	6	0.0518	Y	0	0		
Sema4d	ENSMUSG00000021451	Inside	-	-	-	-	-	-	-	5	0.0609	Y	0	0		
Pml	ENSMUSG00000036986	Inside	Dominant	COSMIC	-	CNV	-	Oct4	-	5	0.0609	Y	0	0		
Rcbtb2	ENSMUSG00000022106	Upstream	-	COSMIC	Mullighan	-	-	-	-	5	0.0609	Y	0	0		
Fut8	ENSMUSG00000021065	Upstream	-	-	-	-	-	-	-	5	0.0609	Y	0	0		
Lrrfip1	ENSMUSG00000026305	Inside	-	-	Mullighan	-	Nanog	-	-	4	0.0840	Y	1	0		
Pik3r5	ENSMUSG00000020901	Upstream	-	-	Mullighan	-	-	-	-	4	0.0840		10	0	Y	Y
2310016C08Rik	ENSMUSG00000043421	Downstream	-	-	-	-	-	-	-	4	0.0840		43	0	Y	Y
Gse1	ENSMUSG00000031822	Upstream/Inside	-	COSMIC	-	-	-	-	-	4	0.0840	Y	0	0		
Tspan14	ENSMUSG00000037824	Inside	-	-	-	-	-	-	-	4	0.0840	Y	0	0		
Ets1	ENSMUSG00000032035	Upstream/Downstream	-	COSMIC	-	-	-	-	-	4	0.0840	Y	0	0		
Dock8	ENSMUSG00000052085	Upstream	-	-	Mullighan	CNV	-	-	-	4	0.0840	Y	0	0		
Vps13d	ENSMUSG00000020220	Inside	-	-	-	-	-	-	-	4	0.0840	Y	0	0		
Gadd45g	ENSMUSG00000021453	Upstream	-	-	-	-	-	Oct4	-	4	0.0840		47	0	Y	
Ccrk	ENSMUSG00000021483	Upstream	-	-	-	-	-	-	-	4	0.0840		47	0	Y	
Metrnl	ENSMUSG00000039208	Upstream	-	-	-	-	-	-	-	4	0.0840	Y	0	0		
Mad1l1	ENSMUSG00000029554	Inside	-	-	Mullighan	-	-	Oct4	-	4	0.0840	Y	0	0		

Table 5.5. A list of CIS genes that are in recurrent deletions of copy number less than or equal to 0.6 across all cell lines (A) and across haematopoietic and lymphoid cancer cell lines (B). Cancer gene = known cancer gene in Cancer Gene Census; COSMIC = gene contains somatic mutations in COSMIC database; Mullighan = gene within deletion in Mullighan et al. (2007) dataset of acute lymphoblastic leukaemias; CNV = gene within CNV identified in Redon *et al*. (2006); "[Nanog, Oct4, p53] BS" = gene contains binding site for Nanog, Oct4 and p53, respectively. "minimal region" = minimal deleted region containing CIS gene; "MCR" = minimal deleted region from across genome, not centred on CIS gene; "Number of genes in minimal region" = number of genes other than the CIS gene within the minimal deleted region.

deleted along with *p27*^{KIP1} in childhood acute lymphoblastic leukaemia (Komuro *et al.*, 1999). Only 2 of the deletions spanning *CCND2* were long enough to include *p27*^{KIP1}. Most were very small (~500 bp) and 18 co-occurred in the same region as 23 focal amplicons, suggesting that they may represent errors in copy number measurement.

For a number of the remaining genes that were recurrently deleted, there is supporting evidence in the literature suggesting that they are tumour suppressor genes. These include *OVCA2* and *MOBKL2A*, which were discussed in Section 3.4.3, plus *DOCK8*, which is deleted and under-expressed in human lung cancers (Takahashi *et al.*, 2006), and *CBFA2T3*, which is a putative breast tumour suppressor gene (Kochetkova *et al.*, 2002; Powell *et al.*, 2002). Interestingly, the human orthologue of *Smg6*, which was proposed to be a putative mouse tumour suppressor gene in Section 3.4.3, was also recurrently deleted, although *SMG6* was located in a minimal deleted region that included 4 other genes that were deleted in a greater number of cell lines. Of these candidates, only *Smg6* has a CIS within the gene. However, all of the genes contain insertions and, especially in the cases of *Ovca2* and *Mobkl2a*, which contain 9 and 6 insertions respectively, it appears that 2 nearby CISs may have been merged, resulting in a CIS location that does not reflect the true location of either CIS. This is one of the limitations of the kernel convolution-based method for identifying CISs, since CISs vary in size and the chosen kernel width may not be appropriate for all CISs.

Although *METRNL* and *ZNF438* (or *Zfp438* in the mouse) were frequently deleted and *ZNF438* has been shown to act as a transcriptional repressor (Zhong *et al.*, 2007), the distribution of insertions around the mouse genes suggests that they are unlikely to act as tumour suppressor genes, at least not in MuLV-induced lymphomagenesis. More promising candidate tumour suppressor genes include cytochrome b5 (*CYB5*), which is frequently deleted in uterine leiomyosarcoma (Cho *et al.*, 2005), and netrin-1 (*NTN1*), the expression of which is reduced in prostate tumours (Latil *et al.*, 2003). Loss of function of dymeclin (*DYM*) is implicated in the rare autosomal recessive Dyggve-Melchior Clausen syndrome (El Ghouzzi *et al.*, 2003), which is associated with mental retardation. No role in tumourigenesis has previously been observed, but the deletion of *DYM* and the presence of intragenic insertions within *Dym* implicates the gene as a potential tumour suppressor. Most of the deletions in *CYB5*, *NTN1* and *DYM* span the entire gene. Once again, only *Dym* contains an internal CIS, but both *Cyb5* and *Ntn1* contain intragenic insertions (3 and 5 insertions, respectively).

The significantly deleted genes also included known and implicated oncogenes *NOTCH1* and *ETS1*. In the case of *NOTCH1*, the minimal deleted region was within intron 2, where deletions that result in the formation of an N-terminally truncated oncoprotein are commonly observed in cancer (see Section 3.4.2). While some of the deletions spanned the entire gene, many left the last 3-8 exons intact, which again may result in the production of the N-terminally truncated, intracellular oncogenic NOTCH-IC protein (Figure 5.7). *ETS1* is a member of the ets protein family, which also includes *ERG* and *ETV6*. The distribution of MuLV insertions in *Erg* and *Etv6*, which are described in Section 3.4.2, suggests that truncation and/or deletion of these genes is implicated in tumourigenesis. A search in the COSMIC database revealed that a nonsense mutation (replacing arginine at reside 211) that would result in the removal of the DNA-binding Ets domain while still retaining the SAM_PNT domain has been observed in the *ETS1* gene in pleural cancer cell line NCI-H2052. The minimal deleted region in *ETS1* spans the final 2 exons of the gene and would result in a similar protein product containing the SAM_PNT domain but no Ets domain (Figure 5.8).

A number of additional candidates were identified among genes that showed significantly recurrent deletion at copy numbers less than or equal to 0.3. All CIS genes occurring within recurrent deletions with a *P*-value of less than 0.1 are shown in Table 5.6A. These included vacuolar protein sorting 13D (*VPS13D*), juxtaposed with another zinc finger protein 1 (*JAZF1*), regulator of chromosome condensation and BTB domain containing protein 2 (*RCBTB2*) and phosphatidylinositol-4,5-bisphosphate 3-kinase catalytic subunit delta isoform (*PIK3CD*). *VPS13D* is poorly characterised and is not known to play a role in cancer, while *JAZF1* forms fusions with *JJAZ1* and *PHF1* in endometrial stromal tumours (Koontz *et al.*, 2001; Micci *et al.*, 2006) and deletion within *JAZF1* has not been implicated in tumourigenesis. In both genes, there is an intronic region of deletion that is exactly the same across each cell line, suggesting that it may not contribute to cancer. In the previous section, *RCBTB2* was shown to be significantly amplified but evidence in the literature suggested that it was a tumour suppressor gene. The identification of recurrent deletions within *RCBTB2* lends further support to this theory. *PIK3CD* is overexpressed in, and contributes to the survival and proliferation of, blast cells in patients with acute myeloid leukaemia (Sujobert *et al.*, 2005), implicating it as an oncogene. However, *PIK3CD* also resides within a region on chromosome 1p36 that is frequently deleted in neuroblastomas and was identified as the most interesting candidate for further study (Caren *et al.*, 2007). Importantly, deletions with a copy number of 0.6 or

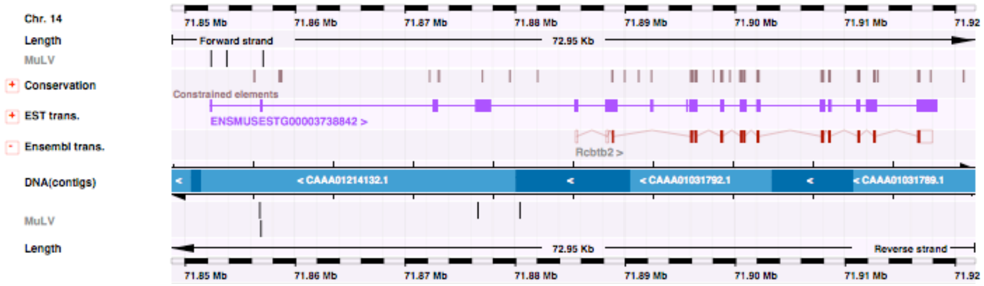

Figure 5.6. All of the MuLV insertions assigned to the *Rcbtb2* gene are within a larger, unannotated, EST transcript. Insertions are shown as black vertical lines. The Ensembl gene and EST are shown in red and purple, respectively.

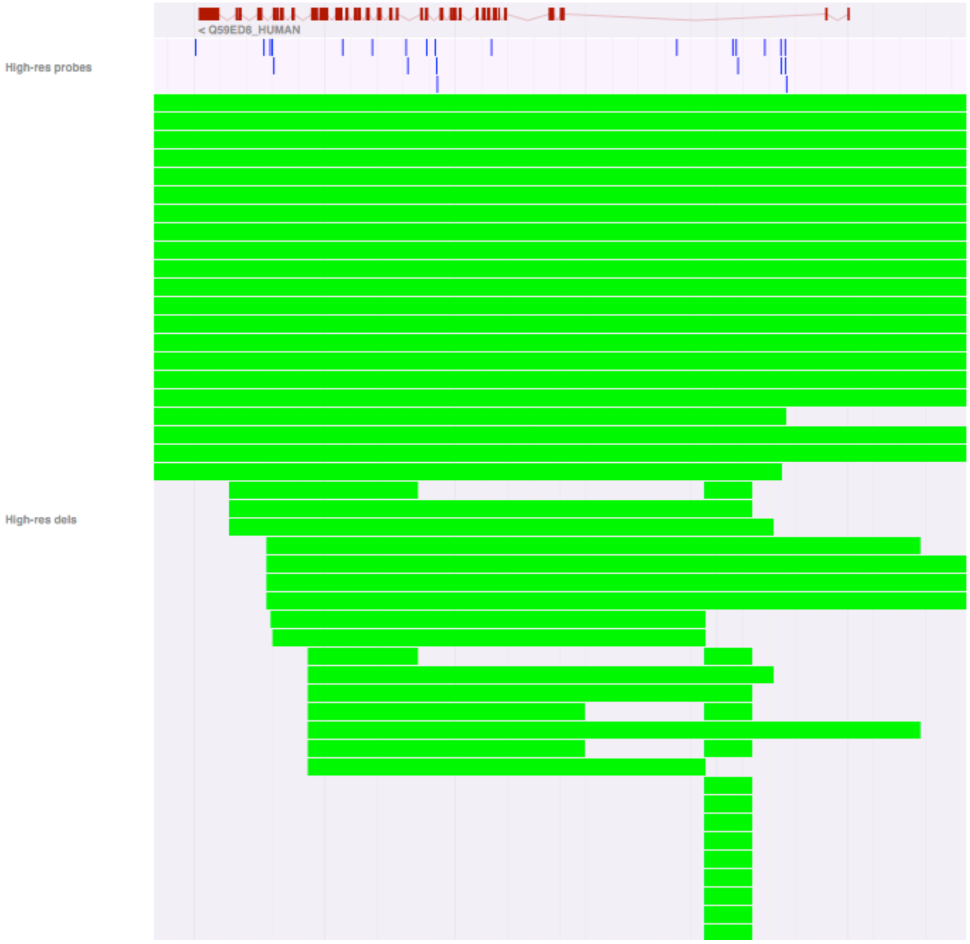

Figure 5.7. Intragenic deletions within *NOTCH1* result in the formation of the oncogenic NOTCH-IC protein. Although some deletions span the entire gene, some may result in a C-terminal truncation containing only the intracellular part of the protein. The *NOTCH1* gene is shown in red. Copy number markers are shown in blue. Deletions are shown in green.

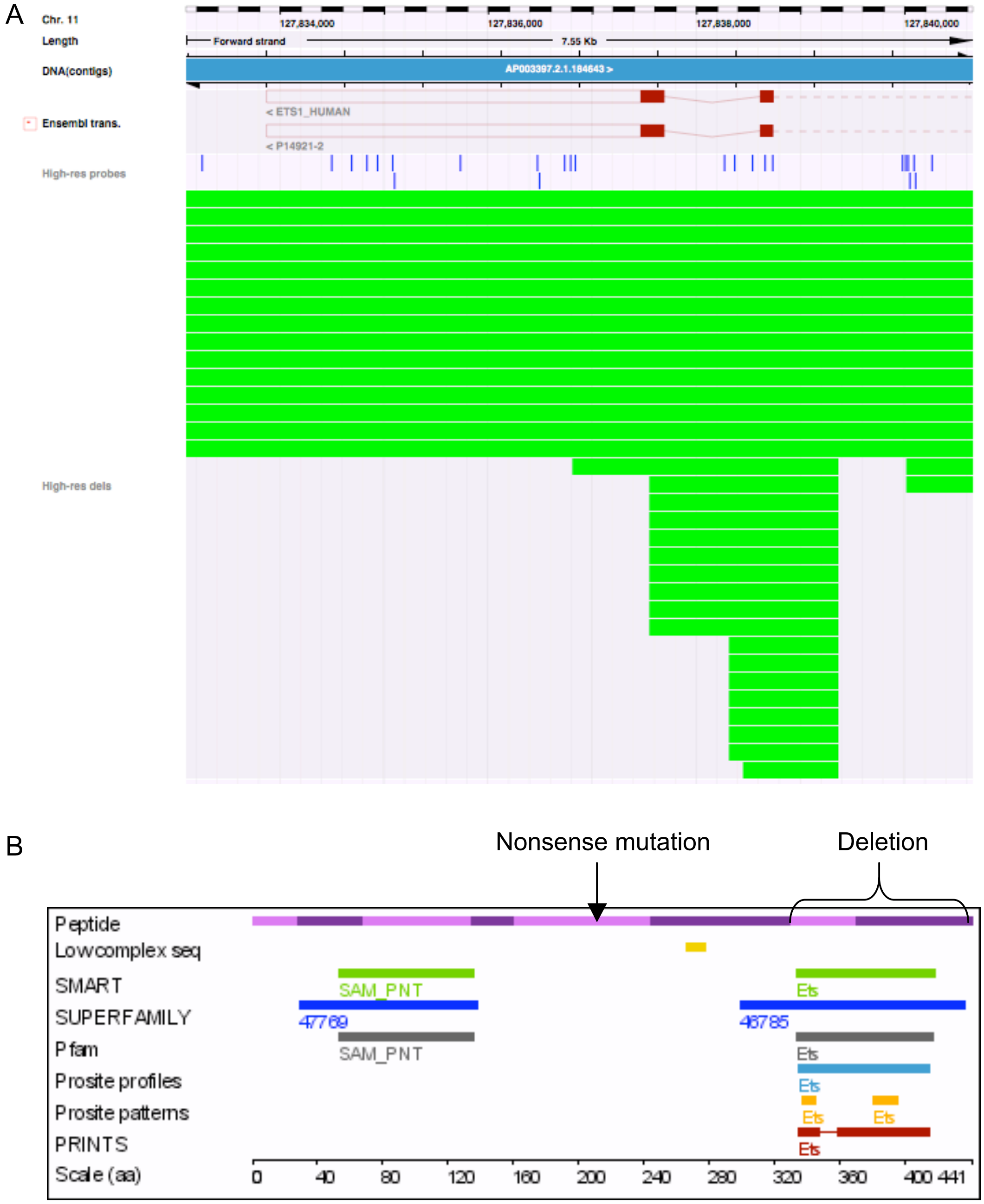

Figure 5.8. Mutations in the *ETS1* gene result in removal of the Ets domain. (A) Deletions within *ETS1* in human cancer cell lines. (B) The location of mutations in the context of the ETS1 protein. In Figure A, *ETS1* gene transcripts are shown in red, copy number markers are shown in blue and deletions are shown in green. Figure B shows the location of the SAM_PNT and Ets domains in the ETS1 protein (extracted from Ensembl geneview), and the position of the nonsense mutation in COSMIC and focal deletions in the human cancer cell lines.

A

CIS gene	Mouse Ensembl ID	Position of CIS relative to gene	Cancer gene	COSMIC	Mullighan	CNV	Nanog BS	Oct4 BS	p53 BS	Number of deletions	*P*-value	Gene in MCR?	Number of genes in minimal region	Number of TSGs in minimal region	Other genes deleted in more cell lines?
Wwox	ENSMUSG00000004637	Inside	-	-	-	CNV	Nanog	Oct4	-	87	0.0011	Y	0	0	
Etv6	ENSMUSG00000030199	Inside	Dominant	-	Mullighan	-	-	-	-	60	0.0018	Y	0	0	
Zfp438	ENSMUSG00000050945	Upstream	-	-	-	-	-	-	-	33	0.0077	Y	0	0	
Lgals9	ENSMUSG00000001123	Upstream	-	-	Mullighan	CNV	-	-	-	31	0.0083		0	0	
Dym	ENSMUSG00000035765	Inside	-	-	-	-	-	-	-	19	0.0138	Y	0	0	
Sdk1	ENSMUSG00000039683	Downstream	-	-	Mullighan	CNV	Nanog	Oct4	-	19	0.0138	Y	0	0	
Acot11	ENSMUSG00000034853	Upstream	-	-	-	-	-	-	-	15	0.0176	Y	0	0	
Notch1	ENSMUSG00000026923	Inside	Dominant	COSMIC	-	CNV	-	-	p53	12	0.0237	Y	0	0	
Dock8	ENSMUSG00000052085	Upstream	-	-	Mullighan	CNV	-	-	-	11	0.0265	Y	0	0	
Ntn1	ENSMUSG00000020902	Downstream	-	-	Mullighan	-	Nanog	-	-	10	0.0282	Y	0	0	
Vps13d	ENSMUSG00000020220	Inside	-	-	-	-	-	-	-	10	0.0282	Y	0	0	
Pik3cd	ENSMUSG00000039936	Upstream	-	-	-	-	-	Oct4	-	8	0.0382	Y	1	0	Y
Ccnd2	ENSMUSG00000000184	Upstream	Dominant	-	-	-	-	-	-	7	0.0422	Y	0	0	
Jazf1	ENSMUSG00000063568	Inside	Dominant	-	Mullighan	CNV	-	-	-	7	0.0422	Y	0	0	
Rcbtb2	ENSMUSG00000022106	Upstream	-	COSMIC	Mullighan	-	-	-	-	7	0.0422	Y	0	0	
Zfp608	ENSMUSG00000052713	Upstream	-	COSMIC	-	-	Nanog	-	-	7	0.0422	Y	0	0	
Zmiz1	ENSMUSG00000007817	Upstream/Inside	-	-	-	-	-	Oct4	-	6	0.0500	Y	0	0	
Ets1	ENSMUSG00000032035	Upstream/Downstream	-	COSMIC	-	-	-	-	-	6	0.0500	Y	0	0	
Nedd4l	ENSMUSG00000024589	Upstream	-	-	-	CNV	Nanog	-	p53	5	0.0569	Y	0	0	
Lrrfip1	ENSMUSG00000026305	Inside	-	-	Mullighan	-	Nanog	-	-	5	0.0569	Y	1	0	Y
Cldn10a	ENSMUSG00000022132	Inside	-	-	-	-	-	-	-	5	0.0569	Y	0	0	
Foxp1	ENSMUSG00000030067	Inside	-	COSMIC	-	-	Nanog	-	-	5	0.0569	Y	0	0	
Arhgef3	ENSMUSG00000021895	Inside	-	-	-	-	Nanog	-	-	5	0.0569	Y	0	0	
Bcl11b	ENSMUSG00000048251	Inside	Dominant	COSMIC	-	-	-	-	-	5	0.0569	Y	0	0	
Fgfr2	ENSMUSG00000030849	Downstream	Dominant	COSMIC	-	-	-	-	-	4	0.0714	Y	0	0	
Tspan2	ENSMUSG00000027858	Upstream	-	-	-	CNV	-	-	-	4	0.0714	Y	0	0	
Tbc1d1	ENSMUSG00000029174	Downstream	-	-	-	-	Nanog	-	-	4	0.0714	Y	0	0	
Flt3	ENSMUSG00000042817	Inside	Dominant	COSMIC	-	-	-	-	-	4	0.0714	Y	0	0	
D12Ertd55.	ENSMUSG00000020589	Downstream	-	-	-	-	Nanog	-	-	4	0.0714	Y	0	0	
Pml	ENSMUSG00000036986	Inside	Dominant	COSMIC	-	CNV	-	Oct4	-	4	0.0714	Y	0	0	
Evi2b	ENSMUSG00000070354	Inside	-	-	-	-	-	-	-	4	0.0714		4	1	Y
Fut8	ENSMUSG00000021065	Upstream	-	-	-	-	-	-	-	4	0.0714	Y	0	0	
Vpreb2	ENSMUSG00000059280	Upstream	-	-	-	-	-	-	-	4	0.0714		1	0	
Ksr1	ENSMUSG00000018334	Downstream	-	COSMIC	-	CNV	-	-	-	3	0.0940	Y	0	0	
Tgfbr3	ENSMUSG00000029287	Inside	-	-	-	-	-	Oct4	-	3	0.0940	Y	0	0	
Sema4d	ENSMUSG00000021451	Inside	-	-	-	-	-	-	-	3	0.0940	Y	0	0	
Cugbp2	ENSMUSG00000002107	Inside	-	-	-	-	-	Oct4	-	3	0.0940	Y	0	0	
Lef1	ENSMUSG00000027985	Upstream	-	-	Mullighan	-	Nanog	-	-	3	0.0940	Y	0	0	
Runx1	ENSMUSG00000022952	Upstream	Dominant	COSMIC	-	-	-	-	-	3	0.0940	Y	0	0	
Kit	ENSMUSG00000005672	Downstream	Dominant	COSMIC	-	-	-	-	-	3	0.0940	Y	0	0	

B

CIS gene	Mouse Ensembl ID	Position of CIS relative to gene	Cancer gene	COSMIC	Mullighan	CNV	Nanog BS	Oct4 BS	p53 BS	Number of deletions	*P*-value	Gene in MCR?	Number of genes in minimal region	Number of TSGs in minimal region	Other genes deleted in more cell lines?
Wwox	ENSMUSG00000004637	Inside	-	-	-	CNV	Nanog	Oct4	-	16	0.0014	Y	0	0	
Etv6	ENSMUSG00000030199	Inside	Dominant	-	Mullighan	-	-	-	-	11	0.0067	Y	0	0	
Lgals9	ENSMUSG00000001123	Upstream	-	-	Mullighan	CNV	-	-	-	5	0.0194		0	0	
Rcbtb2	ENSMUSG00000022106	Upstream	-	COSMIC	Mullighan	-	-	-	-	4	0.0229	Y	0	0	
Zfp438	ENSMUSG00000050945	Upstream	-	-	-	-	-	-	-	3	0.0280	Y	0	0	
Ccnd2	ENSMUSG00000000184	Upstream	Dominant	-	-	-	-	-	-	3	0.0280	Y	0	0	
Notch1	ENSMUSG00000026923	Inside	Dominant	COSMIC	-	CNV	-	-	p53	3	0.0280	Y	0	0	
Dym	ENSMUSG00000035765	Inside	-	-	-	-	-	-	-	2	0.0379	Y	0	0	
Zmiz1	ENSMUSG00000007817	Upstream/Inside	-	-	-	-	-	Oct4	-	2	0.0379	Y	0	0	
Foxp1	ENSMUSG00000030067	Inside	-	COSMIC	-	-	Nanog	-	-	2	0.0379	Y	0	0	
Tbc1d1	ENSMUSG00000029174	Downstream	-	-	-	-	Nanog	-	-	2	0.0379	Y	0	0	
Vps13d	ENSMUSG00000020220	Inside	-	-	-	-	-	-	-	2	0.0379	Y	0	0	
Sdk1	ENSMUSG00000039683	Downstream	-	-	Mullighan	CNV	Nanog	Oct4	-	2	0.0379	Y	0	0	
Ntn1	ENSMUSG00000020902	Downstream	-	-	Mullighan	-	Nanog	-	-	2	0.0379	Y	0	0	
Fut8	ENSMUSG00000021065	Upstream	-	-	-	-	-	-	-	2	0.0379	Y	0	0	
Vpreb2	ENSMUSG00000059280	Upstream	-	-	-	-	-	-	-	2	0.0379		1	0	

Table 5.6. A list of CIS genes that are in recurrent deletions of copy number 0.3 or less across all cell lines (A) and across haematopoietic and lymphoid cancer cell lines only (B). Cancer gene = known cancer gene in Cancer Gene Census; COSMIC = gene contains somatic mutations in COSMIC database; Mullighan = gene within deletion in Mullighan et al. (2007) dataset of acute lymphoblastic leukaemias; CNV = gene within CNV identified in Redon *et al.* (2006); "[Nanog, Oct4, p53] BS" = gene contains binding site for Nanog, Oct4 and p53, respectively. "minimal region" = minimal deleted region containing CIS gene; "MCR" = minimal deleted region from across genome, not centred on CIS gene; "Number of genes in minimal region" = number of genes other than the CIS gene within the minimal deleted region.

less within *PIK3CD* were significantly over-represented among cancer cell lines derived from the autonomic ganglia in this study (*P*=6.37x10^{-8}). The human orthologues of 11 other CIS genes on chromosome 1 were also over-represented among these cell lines, but *PIK3CD* was deleted in the highest number of cell lines (11), which lends support to the conclusions of Caren *et al.* (2007). However, while some of the insertions disrupting *Pik3cd* were within the gene, most were upstream in the antisense orientation, suggesting that *Pik3cd* is upregulated in the MuLV-induced mouse lymphomas. Deletions within *WWOX* that had a copy number of 0.3 or less were over-represented in lung cancer cell lines (*P*=2.91x10^{-4}). Loss of *WWOX* is strongly associated with tumour histology and aggressiveness of non-small cell lung cancers (Donati *et al.*, 2007).

All of the haematopoietic and lymphoid tissue-specific candidates within deletions of copy number less than or equal to 0.6 were also found among candidates identified across all cell lines (Table 5.5B). Of the candidates with *P*-values of less than 0.05 in deletions of copy number 0.3 or less (Table 5.6B), 11 out of 16 haematopoietic and lymphoid candidates also had *P*-values of less than 0.05 across all cell lines. The remaining 5 had *P*-values of less than 0.1 and included *FOXP1*, which was also identified in the 10K analysis (Section 4.5.2.3) and was presented as a putative tumour suppressor gene in Section 3.4.3, and *ZMIZ1*, which may play both tumour suppressive and oncogenic roles (see Section 3.3).

A number of interesting candidates were also identified among those genes occurring in significantly recurrent deletions in specific tissue types. These included *FLI1*, *MYO18A*, *TGFBR3* and *SKI*, all of which were in MCRs across all cell lines. *FLI1* and *MYO18A* contained recurrent deletions in bone cancer cell lines (*P*=0.0371 for both genes). Interestingly, a translocation that leads to the formation of an EWS-FLI1 fusion protein is present in 95% of Ewing's sarcomas, which are tumours of the bone and soft tissue (Delattre *et al.*, 1994). However, *FLI1* is an oncogene, and all but three of the deletions spanning this gene also encompassed other genes (including *ETS1*). *MYO18A* has not been implicated in cancer, but has been shown to be expressed in bone marrow stromal cells, where it may play a role in the maintenance of cell architecture (Furusawa *et al.*, 2000). TGFβ receptor type III precursor *TGFBR3* is implicated as a tumour suppressor gene in a range of cancer types, including non-small cell lung cancer (Finger *et al.*, 2008), pancreatic cancer (Gordon *et al.*, 2008), prostate cancer (Turley *et al.*, 2007) and breast cancer (Dong *et al.*, 2007), but the identification of recurrent deletions in soft tissue cancer cell lines (*P*=0.0270) suggests it is also important in these tumours. *SKI* plays an oncogenic role in some cancers, such as human melanomas (for review, see Reed *et al.*, 2005), but *Ski*-deficient heterozygous mice show an increased susceptibility to tumourigenesis (Shinagawa *et al.*, 2001) and reduced *SKI* expression in breast and lung cancer cells enhances tumour metastasis (Le Scolan *et al.*, 2008). The *P*-value for the number of deletions of *SKI* across all cell lines was 0.0538, but was lower for deletions in cancer cell lines of the kidney (*P*=0.0351), central nervous system (*P*=0.0166) and autonomic ganglia (*P*=0.0130).

The number of deletions was slightly higher across CIS genes than non-CIS genes for deletions with a copy number of 0.6 or less (*P*=0.0306), and significantly higher for deletions with a copy number of 0.3 or less (*P*=0.00340). For CIS genes in deletions of copy number less than or equal to 0.6, the median and mean number of deletions were 8 and 13.54, while for non-CIS genes these were 7 and 11.85, respectively. The maximum number of deletions among CIS genes was 237, occurring within *WWOX*. For CIS genes in deletions of copy number less than or equal to 0.3, the median and mean number were 0 and 1.449, respectively, compared with 0 and 1.303 for non-CIS genes. The maximum number among CIS genes was 87, again occurring within *WWOX*.

The work presented in this section demonstrates the overlap between mouse CIS genes and regions of copy number change in human cancer, and provides a selection of candidates that warrant further investigation.

5.4 Comparison between high-resolution and 10K CGH data

Figure 5.9 provides an illustration of the difference in resolution between the two SNP array CGH platforms. A number of deletions that are clearly visible in the high-resolution data are completely missed by the 10K data for B-cell lymphoma cell line DOHH-2 (Figures 5.9A and B). The breast cancer cell line HCC1143 shows considerable variation in copy number across the genome, but the large distances between adjacent SNPs, and therefore between segments of differing copy number, in the 10K data make it impossible to determine the copy numbers of genes in the intervening regions (Figure 5.9C). This problem is alleviated in the high-resolution data, where most genes contain, or are very close to, a copy number marker (Figure 5.9D).

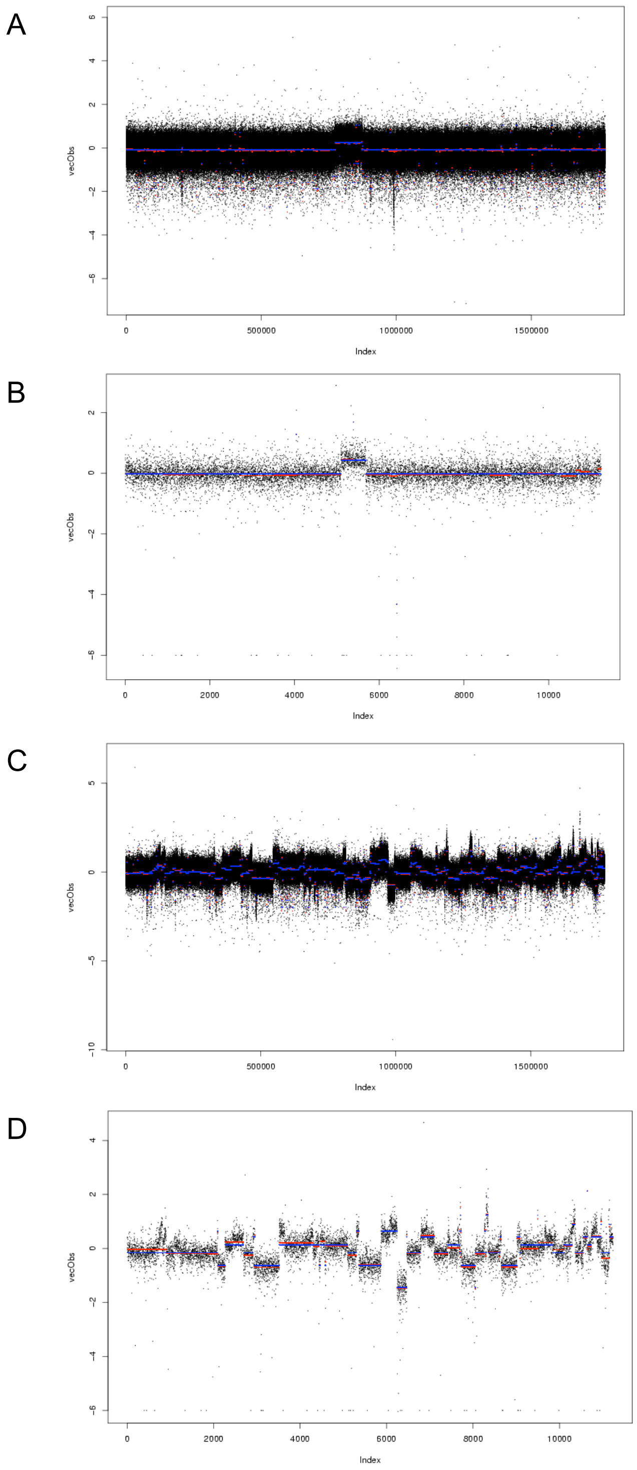

Figure 5.9. High-resolution and 10K SNP array CGH data for the entire genome of B-cell lymphoma cell line DOHH-2 (A and B, respectively) and breast cancer cell line HCC1143 (C and D, respectively). Black points are copy number values for individual SNPs, red lines are mean copy numbers for DNAcopy segments, blue lines are copy number values for merged segments. Markers are positioned according to their order in the genome rather than their exact coordinates. Copy numbers are provided as $\log_2$-ratios.

In order to directly compare the use of high-resolution and 10K CGH data for integrative analyses with the mouse CIS genes, the procedures described in Section 5.3 were applied to the 598 cell lines within the 10K dataset that were also within the high-resolution dataset. Figure 5.10 shows the distribution of the number of amplicons and deletions in these cell lines and the distribution of the lengths of aberrations as determined using the 10K SNP array. To ensure that the dataset was treated identically to the higher resolution data, the start and end coordinates of an amplicon or deletion were taken as the halfway point between the first or last amplified or deleted SNP in a segment and the nearest SNP in the adjacent segment, and amplicons and deletions were defined as regions with a copy number of 1.7 or more, and 0.6 or less, respectively. The average number of amplicons was 1.34 (±1.92) per cell line. The amplicons were on average 17.0 (±28.31) Mb in size and contained 173.14 (±293.84) genes. The average number of deletions was 4.45 (±3.79). These deletions were on average 20.97 (±36.04) Mb in size, encompassing 198.05 (±336.49) genes. Amplicons and deletions were therefore considerably longer in the lower resolution dataset. As mentioned in Section 5.2.1, the higher resolution data are over-segmented, but the lower resolution data may be missing small regions of copy number change and some genes may be incorrectly assigned to amplicons and deletions because of the large distances between probes. The number of minimal common regions (MCRs) for amplicons and deletions were 300 and 741, respectively. These are ~30-fold and 50-fold less than the numbers obtained using the high-resolution data, again reflecting possible over-segmentation of the high-resolution data but also the increased ability to detect small regions of copy number change.

The number of amplicons in the 10K data that contained each CIS gene were counted and compared, using the Mann Whitney U test, to the values for the high-resolution data generated in Section 5.3.2. For the 10K data, the median and mean numbers of amplicons were 2 and 3.82, compared with 1 and 2.768 for the high-resolution data. Therefore, the values across all CIS genes were significantly lower in the high-resolution data (P=6.77x10^{-8}). Likewise, the number of deletions of copy number less than or equal to 0.6 was significantly lower (P=5.12x10^{-5}), with the median and mean measuring 11 and 16.54, respectively, in the 10K data, compared with 8 and 13.54 in the high-resolution data. However, a global analysis of CIS genes in amplicons and deletions in the 10K data, based on the analysis described in Section 5.3.1, demonstrated that the significance of the overlap between CIS genes and amplicons was much lower in the 10K data than in the high-resolution data, and there was no over-representation of CIS genes within

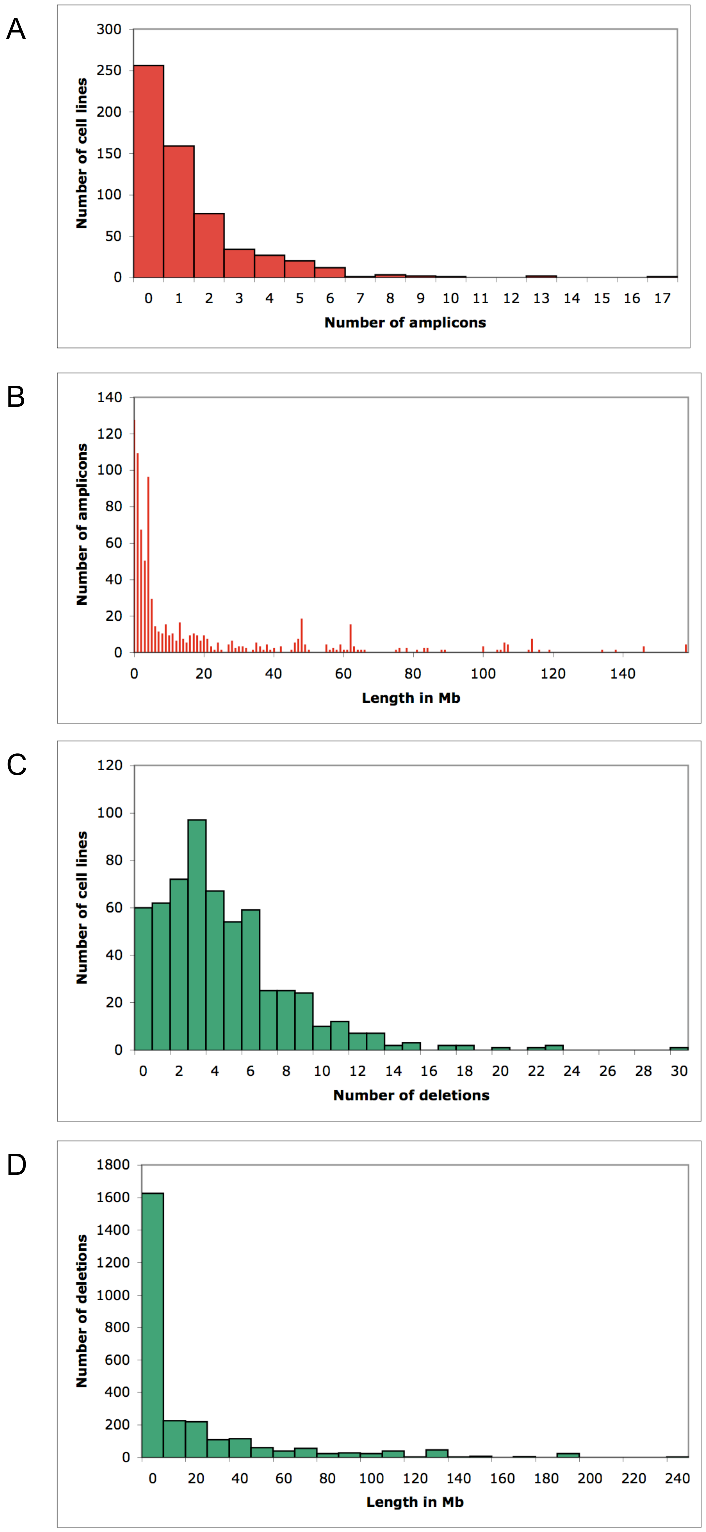

Figure 5.10. Characterisation of amplicons and deletions in 598 human cancer cell lines analysed using 10K SNP array CGH. (A) Number of amplicons per cell line. (B) Length of amplicons. (C) Number of deletions per cell line. (D) Length of deletions.

deletions (Figure 5.11). Therefore, while CIS genes were amplified and deleted in more cell lines, on average, in the 10K dataset than in the high-resolution dataset, non-CIS genes were also amplified and deleted in more cell lines. It is possible that due to the sparseness of the data, regions of copy number change have been incorrectly called or have been extended beyond their true boundaries.

The median and mean number of deletions of copy number 0.3 or less that contained each CIS gene was 0 and 0.444, respectively, for the 10K data, compared with 0 and 1.449 for the high-resolution data. Consequently, the values across all CIS genes were significantly higher in the high-resolution dataset (*P*=0.00144), which is in contrast to the results for amplicons and higher copy deletions. This may reflect the fact that homozygous deletions are likely to be small and are therefore missed by the lower resolution analysis. CIS genes were over-represented in deletions of this copy number in the high-resolution, but not the lower resolution, dataset.

The superiority of the high-resolution dataset was demonstrated by applying the Matthew's Correlation Coefficient (MCC), which is described in Section 2.10.2 and also used in Section 4.6. In each cell line, the number of CIS genes in amplicons was counted and defined as the number of true positives. Non-CIS genes in amplicons were defined as false positives. CIS genes that did not occur in amplicons were defined as false negatives and non-CIS outside of amplicons were defined as true negatives. The number of true and false positives and negatives in each cell line were then added together to give the number across all cell lines. This prevents a CIS gene that is amplified in just one cell line from having the same weight as one that is amplified in multiple cell lines. The procedure was repeated using deletions of copy number less than 0.6 and 0.3. It was also repeated using known cancer genes from the Cancer Gene Census in place of CIS genes, with oncogenes being counted in amplicons, and tumour suppressor genes being counted in deletions. The results are presented in Table 5.7. Although it is not expected that all CIS genes or known cancer genes contribute to human cancer through a change in copy number, these tests do give an indication of the comparative reliability of the two datasets. In concurrence with the results above, the coverage of CIS genes in amplicons and deletions of copy number 0.6 or below was higher in the 10K data but the accuracy was higher in the high-resolution data, while both the coverage and accuracy of CIS genes in deletions of copy number 0.3 or less were higher in the high-resolution data. In all cases, a higher MCC score was obtained using the high-resolution dataset. The same

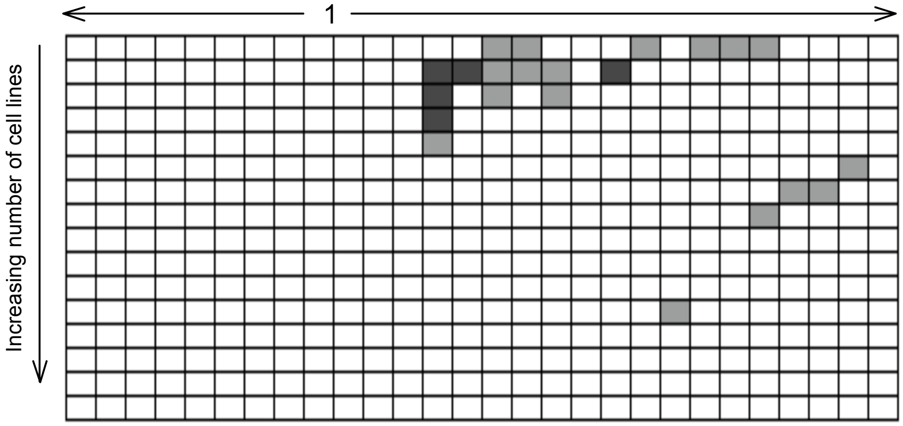

Figure 5.11. Over-representation of CIS genes in amplicons of varying copy number threshold and number of cell lines in the 10K dataset. Each box represents the significance of the association between CIS genes and amplicons/deletions at a given copy number threshold and cell line number. *P*<0.0001, black; *P*<0.001, dark grey, *P*<0.05, light grey. Copy number thresholds below 1 represent deletions, and range from 0.1 to 0.9 with 0.1 increments. Copy number thresholds above 1 represent amplicons, and range from 1.1 to 2.9 with 0.1 increments. The number of cell lines increases in increments of 1, up to a cut-off of 16 cell lines.

A

Region	Resolution	TP	TN	FP	FN	Accuracy	Coverage	MCC
Amplicons	High	1060	9954616	39034	227974	0.026438	0.00463	0.001710
	Low	1704	9922569	71081	227330	0.023411	0.00744	0.000576
Deletions <= 0.6	High	5184	9792549	201101	223850	0.025130	0.02263	0.002643
	Low	7939	9636527	357123	221095	0.021747	0.03466	-0.000855
Deletions <= 0.3	High	555	9971540	22110	228479	0.024487	0.00242	0.000663
	Low	210	9981105	12545	228824	0.016464	0.00092	-0.001419

B

Region	Resolution	TP	TN	FP	FN	Accuracy	Coverage	MCC
Amplicons	High	1017	10032313	39077	150277	0.025365	0.006722	0.005491
	Low	1199	9999804	71586	150095	0.016473	0.007925	0.001174
Deletions <= 0.6	High	1147	9979872	205138	36527	0.005560	0.030445	0.004440
	Low	1919	9821867	363143	35755	0.005257	0.050937	0.004990
Deletions <= 0.3	High	289	10162634	22376	37385	0.012751	0.007671	0.007052
	Low	193	10172448	12562	37481	0.015131	0.005123	0.006676

Table 5.7. Comparison of the high- and low-resolution datasets based on the proportion of CIS genes (A) and known cancer genes (B) that are amplified and deleted. TP = number of CIS/cancer genes in amplicons/deletions, TN = number of non-CIS/non-cancer genes that are not in amplicons/deletions, FP = number of non-CIS/non-cancer genes in amplicons/deletions, FN = number of CIS/cancer genes that are not in amplicons/deletions. Accuracy is given by TP/(TP+FP); Coverage is given by TP/(TP+FN). MCC = Matthew's Correlation Coefficient. Deletions <= 0.6/03 = deletions with a copy number of less than or equal to 0.6/0.3. Amplicons are regions with a copy number of 2.7 or above.

pattern was observed for known cancer genes, except for copy number 0.6 or less, where the MCC score was very slightly higher in the 10K dataset. These results demonstrate that at higher resolution, regions of copy number change are likely to be more defined, making it easier to identify the critical gene(s) that contribute to tumourigenesis. In addition, it is possible to identify smaller changes, particularly deletions, that are missed by lower resolution CGH. Importantly, it appears that the possible over-segmentation of the high-resolution data has not been detrimental to the analysis.

Finally, individual candidates identified in the 10K analysis were compared to those identified in the high-resolution analysis. The 10K dataset contained 20 genes in statistically significant recurrent amplicons (Table 5.8A). Only 2 of these genes were known oncogenes, compared with 9 in the high-resolution dataset, and both were identified in fewer cell lines in the 10K dataset. All of the remaining significantly recurring genes in the 10K dataset were amplified in a higher number of cell lines than those in the high-resolution dataset. However, 11 of these genes were not within an MCR in the 10K data, while a further 2 (*ADRBK1* and *GPR152*) were co-amplified with *CCND1* in 21 cell lines. In the high-resolution dataset, *ADRBK1* and *GPR152* were found in 5 and 4 cell lines, respectively, while *CCND1* was found in 23 lines. Therefore, in this case, it is only possible to discern the critical cancer gene at higher resolution. *NDRG3* and *SLA2* were amplified in 24 and 23 cell lines, respectively, in the 10K dataset but just 1 and 0 cell lines, respectively, in the high-resolution dataset. Visual inspection of the SNPs from the 10K dataset in the context of Ensembl showed that there were none overlapping *NDRG3* and *SLA2*, and therefore the copy number for these genes could not be accurately determined (Figure 5.12A). In order to directly compare the positions of copy number markers in the 10K and high-resolution datasets, the coordinates of the 10K SNPs, which were originally mapped to the NCBI 35 human genome assembly, were converted to NCBI 36 using the UCSC LiftOver tool (http://genome.ucsc.edu/cgi-bin/hgLiftOver). Copy number markers for both datasets were displayed in Ensembl contigview using a DAS track (see Section 2.6 for further details regarding DAS). Interestingly, the regions surrounding the closest 10K SNPs to *NDRG3* and *SLA2* were not amplified in the high-resolution dataset, suggesting that the sparseness of probes can lead to the miscalling of copy number changes across large regions. *CAPSL* and *ZNF217* (known as *Zfp217* in the mouse) were also amplified in a greater number of cell lines in the 10K dataset, but these too did not contain any 10K SNPs. In the case of *ZNF217*, the number of high-resolution amplicons overlapping the adjacent gene, *BCAS1*, was slightly

A

Gene	Position of CIS relative to gene	Number of amplicons	*P*-value	Gene in MCR?	Number of additional genes in MCR	Number of known oncogenes in MCR	Other genes in MCR	Amplified in more cell lines than high-resolution?
Myc	Upstream	39	0.0008	Y	0	0		
Capsl	Downstream	25	0.0055	Y	0	0		Y
Slc1a3	Downstream	24	0.0089	N	0	0		Y
Fyb	Inside	24	0.0089	N	0	0		Y
Sla	Inside	21	0.0114	N	0	0		Y
Zfp217	Upstream	21	0.0114	Y	0	0		Y
Ptp4a3	Inside	20	0.0178	N	0	0		Y
Ndrg3	Downstream	20	0.0178	Y	1	0	*Sla2*	Y
Sla2	Inside	19	0.0227	Y	1	0	*Ndrg3*	Y
Bcl2l1	Inside	18	0.0298	N	0	0		Y
Ncoa3	Upstream	18	0.0298	N	0	0		Y
Prkcbp1	Upstream	17	0.0375	N	0	0		Y
Ccnd1	Upstream	17	0.0375	Y	2	0	*Gpr152, Adrbk1*	
Serinc3	Inside	17	0.0375	N	0	0		Y
Ppp1r16b	Inside	17	0.0375	N	0	0		Y
Gpr152	Downstream	17	0.0375	Y	2	1	*Adrbk1, Ccnd1*	Y
Stk4	Inside	17	0.0375	N	0	0		Y
Adrbk1	Inside	17	0.0375	Y	2	1	*Gpr152, Ccnd1*	Y
Cldn10a	Inside	16	0.0394	N	0	0		Y
Ubac2	Inside	14	0.0458	Y	0	0		Y

B

Gene	Position of CIS relative to gene	Number of amplicons	*P*-value	Gene in MCR?	Number of additional genes in MCR	Amplified in more cell lines than high-resolution?
Cyb5	Downstream	89	0.0030	N	0	Y
Nedd4l	Upstream	84	0.0050	N	0	Y
Mbd2	Upstream	80	0.0060	N	0	Y
Dym	Inside	76	0.0082	N	0	Y
Dock8	Upstream	76	0.0082	Y	0	Y
Rcbtb2	Upstream	71	0.0124	Y	0	Y
Lcp1	Inside	66	0.0177	N	0	Y
1190002H23Rik	Downstream	64	0.0227	N	0	Y
Aqp4	Downstream	64	0.0227	Y	0	Y
Elf1	Inside	63	0.0241	N	0	Y
1700081D17Rik	Downstream	61	0.0291	N	0	Y
Katnal1	Upstream	60	0.0311	Y	0	Y
Spata13	Inside	59	0.0350	N	0	Y
Ubac2	Inside	59	0.0350	N	0	Y
Flt3	Inside	58	0.0370	N	0	Y
Cldn10a	Inside	55	0.0404	N	0	Y
D18Ertd653e	Inside	52	0.0440	N	0	Y

C

Gene	Position of CIS relative to gene	Number of amplicons	*P*-value	Gene in MCR?	Number of additional genes in MCR	Number of known oncogenes in MCR	Other genes in MCR	Amplified in more cell lines than high-resolution?
Cyb5	Downstream	9	0.0052	N	0	0		Y
Mbd2	Upstream	8	0.0061	N	0	0		Y
Nedd4l	Upstream	6	0.0093	N	0	0		Y
Rasgrp1	Upstream	6	0.0093	N	0	0		Y
Rcbtb2	Upstream	4	0.0170	Y	0	0		
Dock8	Upstream	4	0.0170	N	0	0		Y
Bid	Upstream	3	0.0405	Y	4	0	*Vpreb2, BC030863, Cecr5, Tuba8*	Y
Cecr5	Upstream	3	0.0405	Y	4	0	*Vpreb2, BC030863, Bid, Tuba8*	Y
Wwox	Inside	3	0.0405	Y	0	0		
BC030863	Downstream	3	0.0405	Y	4	0	*Vpreb2, Cecr5, Bid, Tuba8*	Y
Ubac2	Inside	3	0.0405	N	0	0		Y
Vpreb2	Upstream	3	0.0405	Y	4	0	*BC030863, Cecr5, Bid, Tuba8*	Y
Tuba8	Downstream	3	0.0405	Y	4	0	*BC030863, Cecr5, Bid, Vpreb2*	Y

Table 5.8. A list of CIS genes that are in recurrent amplicons (A), recurrent deletions of copy number 0.6 or less (B) and recurrent deletions of copy number 0.3 or less (C) in the 10K CGH dataset. “MCR” = minimal amplified or deleted region.

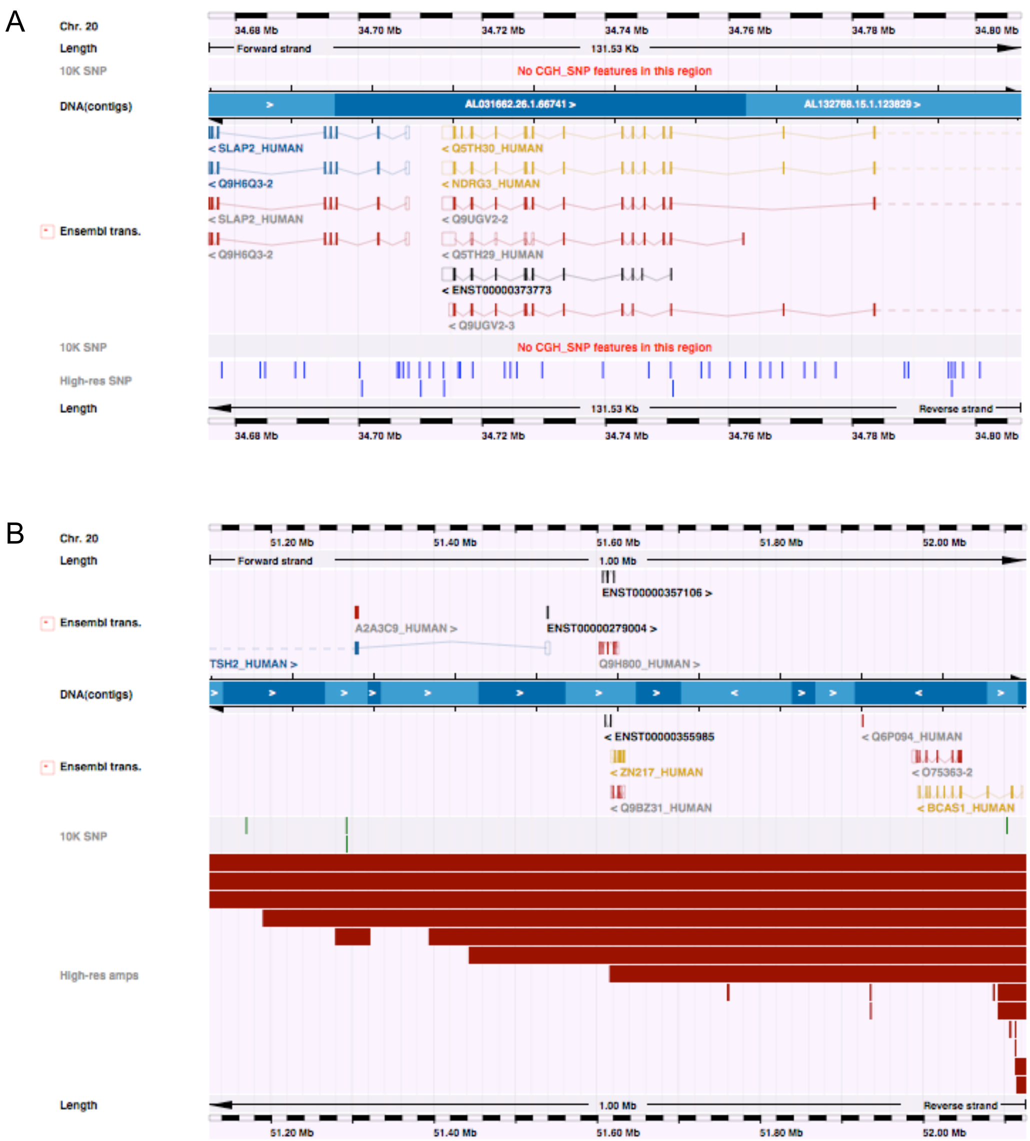

Figure 5.12. *SLA2* and *NDRG3* (A) and *ZNF217* (B) are amplified in a greater number of cell lines in the 10K dataset than in the high-resolution dataset but do not contain SNPs in the 10K dataset. In Figure A, copy number markers in the high-resolution dataset are shown in blue. In Figure B, copy number SNPs in the 10K dataset are shown in dark green. Amplicons identified in the high-resolution analysis are shown as red rectangles.

higher and may contribute to the amplicons incorrectly assigned to *ZNF217* in the 10K dataset (Figure 5.12B).

Among deletions of copy number less than or equal to 0.6, all of the 17 significantly recurrent genes in the 10K dataset (Table 5.8B) were deleted in a greater number of cell lines than in the high-resolution dataset. However, 13 of these genes were not within an MCR. Of the 4 that were, 3 (*RCBTB2*, *AQP4* and *KATNAL1*) did not contain any 10K SNPs, and therefore the copy numbers of these genes in the 10K dataset are not accurate. In addition, neither *Rcbtb2* nor *Aqp4* contained any intragenic MuLV insertions in mouse lymphomas, while *Katnal1* contained just 2 insertions. Only *DOCK8* contained SNPs, and it is not clear why the number of deletions was so much greater at lower resolution. Just 4 of the 24 significantly recurrent deleted genes in the high-resolution dataset were significantly recurrent in the 10K dataset. A number of promising candidates from the high-resolution analysis, including *WWOX*, *SDK1* and *CBFA2T3*, were deleted in fewer cell lines, and were not significant, in the 10K dataset.

For deletions of copy number 0.3 or less, 11 genes were deleted in a higher number of cell lines in the 10K dataset. 6 were not within an MCR and the remaining 5 resided in the same MCR (Table 5.8C). *WWOX* was deleted in fewer cell lines than in the high-resolution data, and *RCBTB2* was deleted in the same number of lines. *WWOX* and *RCBTB2* were the only significantly recurrent deleted genes from the high-resolution dataset that were represented among significant genes from the 10K dataset.

The lists of significantly recurrent amplified and deleted CIS genes differ considerably between the 10K and high-resolution datasets. These differences reflect the low density of copy number markers in the 10K analysis, which results in small deletions being missed and, therefore, putative tumour suppressor genes going undetected. They also result from genes being incorrectly flagged as promising candidates when copy number regions are miscalled or are extended beyond their true boundaries. However, as discussed in Sections 5.3.2.2 and 5.3.2.3, in the high-resolution analysis, some of the significantly amplified and deleted genes, including *ETV6*, *CUGBP2* and *CCND2*, contained small, repetitive, regions of copy number change across multiple cell lines that did not look likely to contribute to tumourigenesis. These tiny changes were not identified in the lower resolution analysis and therefore amplifications and deletions within these genes were not significantly recurrent. Therefore, for amplicons, which are often large and encompass multiple genes, the most convincing candidate oncogenes may be those that are identified using both platforms. For deletions, which are often smaller, the identification of candidate tumour suppressor genes is more reliant on the high-resolution dataset.

Having established that candidates may be incorrectly assigned to amplicons or deletions because of the low resolution of the 10K data, it is important to revisit the candidates identified in Chapter 4 to determine whether they are still valid. Among mouse candidates near to CISs with a *P*-value of less than 0.001, 17 of those presented as putative oncogenes in Table 4.4 were not included within the more conservative dataset used in this chapter and have therefore not been studied in relation to the high-resolution data. These included *Meis1*, *Mmp13*, *Smad7*, *Lrrc5* and *Taok3*, all of which were discussed in Section 4.5.2.1.1. Of the remaining 37 candidates, all were found within at least one human amplicon, but only 11 had a *P*-value of less than 0.1 in the recurrence analysis. Apart from *Zfp217*, all of these genes were known oncogenes from the Cancer Gene Census. The *P*-values for *Nfkb1*, *Slamf6* and *Rreb1*, which were presented as candidates in Section 4.5.2.1.1, were greater than 0.1. However, *Nfkb1* was amplified in a single cell line in both datasets and, while *Slamf6* and *Rreb1* were amplified in fewer cell lines in the high-resolution data (5 and 2 cell lines, respectively), the minimal amplified regions contained just 3 and 2 genes, respectively, and included no known oncogenes. Hence there is no convincing evidence from the higher resolution data to suggest that these genes are not amplified in cancer.

Of the candidates that were identified in deletions in the 10K data, 39 were discussed in Section 4.5.2.3. 27 of these genes were not within the list of mouse candidate cancer genes used in this chapter. This is because the lists used in Chapter 4 included all genes containing insertions, whether the insertions clustered into a CIS or not, whereas only CIS genes were considered in the list generated in Chapter 2 and used in this chapter. Among the remaining 12 genes, 5 were recurrently deleted to copy number 0.6 and below and/or 0.2 and below with a *P*-value of less than 0.1. These were *Wwox*, *Foxp1*, *Sdk1*, *Bcl11b*, *Lef1* and *Mbd2*. The only implicated tumour suppressor gene that was no longer identified in the high-resolution analysis was *Gpr56*. Known and implicated oncogenes *Evi1*, *Myc*, *Fli1*, *Rasgrp1* and *Map3k8* were deleted, but not significantly, therefore providing further evidence that these deleted genes are most likely passengers, rather than causative genes, in human cancers.

5.5 *Identification of co-operating cancer genes*

5.5.1 Genotype-specific cancer genes

Most of the cancer cell lines used in this study are part of the Cancer Cell Line Project undertaken by the Wellcome Trust Sanger Institute Cancer Genome Project (http://www.sanger.ac.uk/genetics/CGP/CellLines/). The aim of the project is to systematically sequence all known cancer genes in all of the selected cell lines. 595 of the 598 cell lines used in this study have been analysed for somatic mutations in *TP53* and *CDKN2A*. 311 have mutations in *TP53*, while 160 have mutations in *CDKN2A*. For each amplified or deleted CIS gene, a 2-tailed Fisher Exact Test was used to determine whether there was any significant difference between the number of cell lines with *TP53* or *CDKN2A* mutations that contained the amplified or deleted gene compared with the number with *TP53* or *CDKN2A* mutations that did not contain the amplified or deleted gene. A positive association between an amplified or deleted gene and cell lines bearing a *TP53* or *CDKN2A* mutation suggests that the disrupted CIS gene may co-operate with inactivation of *TP53* or *CDKN2A* in human tumourigenesis. Likewise, a negative association suggests that the genes never co-operate, possibly because they act in the same cancer pathway. This analysis is equivalent to the genotype-specific analysis described in Section 3.5.1 that was performed on tumours generated in mice deficient in *p53*, or *p19* and/or *p16* (*p16* and *p19* are collectively known as *Cdkn2a*). The identification of genes that co-operate in both species provides strong evidence that the co-operation is real, and that cancer pathways are conserved between species.

This analysis was performed on CIS genes amplified to copy number 1.7 or above and deleted to copy number 0.6, or 0.3, or below. While a small number of tests had *P*-values of less than 0.05, none of the tests were below the significance level adjusted to account for multiple testing (either with *q*-values, determined using the R package QVALUE (see Section 3.5.1), or using the Bonferroni correction, where the significance level of 0.05 was divided by the number of amplified or deleted genes). Nevertheless, associations with a *P*-value of less than 0.05, which are presented in Table 5.9, were investigated to determine whether there was any evidence in the literature to support the findings. *CCND1* amplification co-occurred with *TP53* mutation with a *P*-value of 0.0173. Overexpression of *CCND1* and mutant *TP53* have been shown to co-occur in human uterine endometrial carcinomas (Nikaido *et al.*, 1996). However, insertional mutagenesis of *Ccnd1* was associated with the loss of *Cdkn2a*, and not with *Trp53*, in mouse

Region	Genotype	CIS gene	Mouse Ensembl ID	***P*-value**
Amplicons	*TP53* up	*Ccnd1*	ENSMUSG00000070348	0.0173
		Sdk1	ENSMUSG00000039683	0.0228
Deletions <= 0.6	*TP53* up	*BC008155*	ENSMUSG00000057411	0.0009
		Dock8	ENSMUSG00000052085	0.0012
		B3gntl1	ENSMUSG00000046605	0.0016
		1300007F04Rik	ENSMUSG00000000686	0.0035
		Zfp608	ENSMUSG00000052713	0.0075
		Mobkl2a	ENSMUSG00000003348	0.0104
		Mknk2	ENSMUSG00000020190	0.0104
		Metrnl	ENSMUSG00000039208	0.0115
		Arrdc5	ENSMUSG00000073380	0.0126
		Ubac2	ENSMUSG00000041765	0.0130
		Cldn10a	ENSMUSG00000022132	0.0173
		Gadd45b	ENSMUSG00000015312	0.0184
		Abcg1	ENSMUSG00000024030	0.0217
		Arid3a	ENSMUSG00000019564	0.0247
		Ptbp1	ENSMUSG00000006498	0.0247
		Mbd2	ENSMUSG00000024513	0.0332
		Il6st	ENSMUSG00000021756	0.0356
		Midn	ENSMUSG00000035621	0.0360
		Gna15	ENSMUSG00000034792	0.0448
	TP53 down	*Ski*	ENSMUSG00000029050	0.0091
		Park7	ENSMUSG00000028964	0.0143
		Pml	ENSMUSG00000036986	0.0160
		Prdm16	ENSMUSG00000039410	0.0308
	CDKN2A up	*Kdr*	ENSMUSG00000062960	0.0109
		Acot11	ENSMUSG00000034853	0.0469
Deletions <= 0.3	*CDKN2A* up	*Vpreb2*	ENSMUSG00000059280	0.0051

Table 5.9. A list of amplified and deleted CIS genes that are over- or under-represented in cell lines that contain a mutation in *TP53* or *CDKN2A*. "Amplicons" represents regions with a copy number greater than or equal to 1.7; "Deletions <= 0.6" and "Deletions <= 0.3" represent regions with a copy number less than or equal to 0.6 and 0.3, respectively. "*TP53* up" and "*TP53* down" represent CIS genes that are over-represented and under-represented, respectively, in *TP53*-mutated cell lines; "*CDKN2A* up" represents CIS genes that are over-represented in *CDKN2A*-mutated cell lines. *P*-values were generated using the 2-tailed Fisher Exact Test.

lymphomas. A positive association was observed between deletion of *KDR* and mutation in *CDKN2A*, and mutagenesis of *Kdr* was also weakly associated with loss of *Cdkn2a* in mouse tumours (*P*=0.015). KDR is a receptor for vascular endothelial growth factor (VEGF) that plays a role in tumour angiogenesis and would generally be expected to be amplified, rather than deleted, in cancer cell lines, and none of the MuLV insertions in *Kdr* were intragenic. However, non-endothelial *KDR* expression is associated with increased survival of patients with urothelial bladder carcinomas (Gakiopoulou-Givalou *et al.*, 2003).

Deletion within DJ-1 (*PARK7*) was negatively associated with *TP53* mutation. However, *PARK7* is a putative oncogene, and it does not reside within an MCR. Deletions within *PARK7* are implicated in Parkinson's disease rather than cancer (Abou-Sleiman *et al.*, 2004). Also negatively associated with *TP53* mutation were deletions within the promyelocytic leukaemia gene *PML*. PML forms an oncogenic fusion protein with RARα in acute promyelocytic leukaemia, but alone, *PML* functions as a tumour suppressor gene. PML and p53 were identified as independent prognostic markers in gallbladder carcinomas, suggesting that they do not co-operate in tumourigenesis (Chang *et al.*, 2007). *PML* is a direct target of p53 (de Stanchina *et al.*, 2004), which suggests that there is no selective advantage in mutating both genes and provides support for the observed association.

Among the deleted genes that were positively associated with *TP53* mutations were AT-rich interactive domain 3A (*ARID3A*) and growth arrest and DNA-damage-inducible protein beta (*GADD45β*). *ARID3A* contains a putative binding site for p53, and has been proposed to play a role in growth suppression mediated by p53 (Ma *et al.*, 2003). The fact that *ARID3A* deletions co-occur with *TP53* mutation suggests that other mechanisms may also contribute to the activation of *ARID3A* since inactivation of *TP53* would otherwise be sufficient to inactivate *ARID3A*. Down-regulation of *GADD45β* was found to be associated with human hepatocellular carcinoma but was inversely correlated with the presence of mutant p53 (Qiu *et al.*, 2003). In addition, p53 is believed to be involved in regulating the expression of murine *Gadd45β* (Balliet *et al.*, 2001), suggesting that inactivation of *GADD45β* and *TP53* should be negatively, rather than positively, associated. However, p53 knockdown in a human glioma cell line showed that *GADD45α* could be induced by a p53-independent mechanism (Heminger *et al.*, 2006), and it is therefore possible that the same may be true of *GADD45β*, in which case inactivation of both genes may provide an additional growth advantage on the tumour. Neither *ARID3A* nor *GADD45β* resides within an MCR, which casts doubt on the observed associations.

There was no evidence in the literature to directly support the remaining associations. This analysis has unfortunately not yielded any strong candidates for co-operation with inactivated tumour suppressor genes *TP53* and *CDKN2A*, and for those that showed a slight association, there was very little correlation with the co-operating candidates identified in the insertional mutagenesis screen. A possible explanation for the lack of association is that for some of the cell lines in which a gene is amplified or deleted, that gene may not be a critical gene and may not be contributing to tumourigenesis. Thus any association in cell lines where the CIS gene is the critical gene could be concealed. Another confounding factor is that some of the cell lines have not been screened to completion and therefore, in some cases, a cell line that is labelled as not containing a mutation may in fact do so.

5.5.2 Co-occurrence of amplified and deleted candidate cancer genes

The aim of this analysis was to identify amplified and deleted CIS genes that co-occur across human cancer cell lines. Co-occurring cancer genes may co-operate in tumourigenesis and therefore represent attractive targets for combined therapies (see Section 1.2.7). This analysis is equivalent to the analysis performed on mouse tumours in Section 3.5.2. Any candidates that are disrupted in common in both mouse and human cancers are particularly strong candidates for a co-operating role in tumourigenesis, and co-occurring genes were therefore compared across species.

Co-occurring genes were identified by taking each pair of CIS genes and counting the number of cell lines in which they were both amplified, and the number in which they were uniquely amplified. A 2-tailed Fisher Exact Test was used to determine whether there was any significant difference between the observed number of co-amplifications, and the expected number.

a	*b*
c	*d*

a = Number of tumours in which both genes were amplified
b = Number of tumours in which the first gene was uniquely amplified
c = Number of tumours in which the second gene was uniquely amplified
d = Number of tumours in which neither gene was amplified

The test was repeated to identify co-occurring deleted CIS genes, and co-occurring amplified and deleted CIS genes. Deleted genes were defined as genes with a copy number of 0.3 or less, since regions of copy number 0.6 or less can be very large and all but one of the CIS genes were identified in at least one of these deletions (see Section 5.3.2.3). To account for multiple testing, *q*-values were calculated using the R package QVALUE (see Section 3.5.1). Most of the co-occurrences of amplified genes were on the same chromosome and the genes may therefore co-occur simply because they are located close to one another. Significant co-occurrences between CIS genes on different chromosomes are shown in Table 5.10. None of these co-occurrences were identified among the mouse lymphomas in Section 3.5.2.

There were 3 significant co-occurrences of amplified genes, of which 2 included sodium-coupled neutral amino acid transporter 1 (*SLC38A1*). The minimal amplified region containing *SLC38A1* also contained 33 additional genes, of which 29 were amplified in a greater number of cell lines and 14 of these were amplified to higher copy number. This suggests that *SLC38A1* is not the critical gene with which *KDR* and/or *KIT* co-operate. Interestingly, both genes in the remaining gene pair of *TMEM49* and *NCOA3* were significantly over-represented in breast cancer cell lines (see Section 5.3.2.2) and all 3 of the cell lines in which they co-occurred were derived from breast cancers. However, *TMEM49* and *NCOA3* were co-amplified with other putative breast cancer genes on chromosomes 17q23 and 20q, respectively. While it is therefore not possible to conclude that *TMEM49* and *NCOA3* co-operate in breast cancer, it does appear that genes within the two amplicons do co-operate.

There were 8 significant co-occurrences of deleted genes in the human cancer cell lines. There is no evidence in the literature to support any of these associations, and some of the genes seem unlikely candidates for a role in tumour suppression. *LRRFIP1* is over-expressed in Burkitt's Lymphoma and other cancer cell lines (Rikiyama *et al.*, 2003) and represses expression of tumour necrosis factor alpha, which is involved in apoptosis (Suriano *et al.*, 2005). Likewise, *BCL9L* is overexpressed in acute lymphoblastic

A

Gene 1	Gene 2	Gene 1 human coordinates	Gene 2 human coordinates	Shared	Gene 1 unique	Gene 2 unique	Shared cell lines	*P*-value	q-value
Slc38a1	*Kdr*	12:44863110-44949475	4:55639416-55686519	2	0	3	NCI-H1930 (lung), NCI-H1693 (lung)	5.60E-05	0.0082
Slc38a1	*Kit*	12:44863110-44949475	4:55218863-55301636	2	0	4	NCI-H1930 (lung), NCI-H1693 (lung)	8.40E-05	0.0116
Tmem49	*Ncoa3*	17:55139811-55273235	20:45564053-45719023	2	4	1	HCC2218 (breast), MCF7 (breast), BT474 (breast)	2.51E-04	0.0278

B

Gene 1	Gene 2	Gene 1 human coordinates	Gene 2 human coordinates	Shared	Gene 1 unique	Gene 2 unique	Shared cell lines	*P*-value	q-value
D12Ertd553e	*Tspan2*	2:16594890-16710580	1:115392155-115433638	4	0	0	no-11 (glioma), ATN-1 (leukaemia), COLO-205 (large intestine), NCI-H322M (lung)	1.90E-10	1.54E-06
Dock8	*Tspan2*	9:263048-455255	1:115392155-115433638	4	7	0	no-11 (glioma), ATN-1 (leukaemia), COLO-205 (large intestine), NCI-H322M (lung)	6.26E-08	1.70E-04
Dock8	*D12Ertd553e*	9:263048-455255	2:16594890-16710580	4	7	0	no-11 (glioma), ATN-1 (leukaemia), COLO-205 (large intestine), NCI-H322M (lung)	6.26E-08	1.70E-04
Bcl9l	*D12Ertd553e*	11:118272059-118286823	2:16594890-16710580	2	0	2	ATN-1 (leukaemia), COLO-205 (large intestine)	3.36E-05	0.0391
Bcl9l	*Tspan2*	11:118272059-118286823	1:115392155-115433638	2	0	2	ATN-1 (leukaemia), COLO-205 (large intestine)	3.36E-05	0.0391
Pml	*Mad1l1*	15:72074067-72127204	7:1821956-2239109	2	2	0	NCI-H1417 (lung), NCI-H322M (lung)	3.36E-05	0.0391
Bcl9l	*Tbc1d1*	11:118272059-118286823	4:37569115-37817189	2	0	2	ATN-1 (leukaemia), COLO-205 (large intestine)	3.36E-05	0.0391
Dock8	*Lrrfip1*	9:263048-455255	2:238200958-238353697	3	8	2	no-11 (glioma), ATN-1 (leukaemia), COLO-205 (large intestine), NCI-H322M (lung)	4.56E-05	0.0464

Table 5.10. A list of CIS genes that are co-amplified (A) or co-deleted (B) across a significant number of human cancer cell lines. The column "Shared" shows the number of cell lines in which the genes are co-amplified/co-deleted. "Gene 1 unique" and "Gene 2 unique" show the number of cell lines in which gene 1 and gene 2 are amplified/deleted without amplification/deletion of the other gene in the pair. *P*-values were calculated using the 2-tailed Fisher Exact Test, *q*-values were calculated using R package QVALUE.

leukaemias and a range of other tumour types (Katoh and Katoh, 2003), and *TBC1D1* has been implicated as an oncogene in gastric cancer (Yang *et al.*, 2007). *FAM49A* (or *D12Ertd553e*) is frequently co-amplified with *MYCN* in neuroblastomas and acute lymphoblastic leukaemias and so is also more likely to play an oncogenic role in cancer (see Section 5.3.2.2). However, *DOCK8* and *PML* are putative tumour suppressor genes (see Sections 5.3.2.3 and 5.5.1, respectively) and *TSPAN2* is methylated in breast cancer, suggesting that it may play a tumour suppressive role (Miyamoto *et al.*, 2005), although it is also a potential marker of metastasis in tongue tumours (Carinci *et al.*, 2005). 7 of the co-occurrences involved the leukaemia cell line ATN-1 and colon cancer cell line COLO-205, and 3 of these also involved glioma cell line no-11 and lung cancer cell line NCI-H322M. Each of these cell lines contains a very high number of deletions, i.e. 1228, 1431, 667 and 1284 in cell lines ATN-1, COLO-205, no-11 and NCI-H322M, respectively. They therefore appear to be extremely unstable and the observed co-occurrences may not be functionally important. This is supported by the diversity of cancer types. More interesting is the co-occurrence between deletions involving *PML* and *MAD1L1* (*MAD1*). Both cell lines are from lung cancers. Lack of expression of *PML* has been demonstrated in human lung cancers (Gurrieri *et al.*, 2004; Zhang *et al.*, 2000; Zhao *et al.*, 2006). In contrast, amplification of a region containing *MAD1L1* is the most commonly observed event in small cell lung cancer cell lines (Coe *et al.*, 2006) and amplified *MAD1L1* was significantly over-represented in lung cancer cell lines in Section 5.3.2.2. However, an oncogenic role has not been proven, and *MAD1L1* is more widely implicated as a tumour suppressor gene, e.g. in human stomach cancer (Osaki *et al.*, 2007). Interestingly, PML promotes MAD-mediated transcriptional repression (Khan *et al.*, 2001), which may contribute to tumour suppression through the repression of *MYC* (Grandori *et al.*, 2000). However, as the genes are co-deleted, it appears that inactivation of *PML* may not be sufficient to completely abrogate the tumour suppressive functions of *MAD*. There were no significant co-occurrences of gene pairs in which one gene was amplified and the other was deleted.

All co-occurrences with a *P*-value of less than 0.05 were then compared to co-occurrences with a *P*-value of less than 0.05 within mouse lymphomas to identify candidates that might be conserved across species. It is possible that genes that co-occur within the same amplicon may also co-operate in tumourigenesis (see Section 1.3.3.3) and these were therefore included in the analysis. Gene pairs that co-occurred in both mouse and human are shown in Table 5.11.

A

		Human cancer cell lines								Mouse lymphomas				
Gene 1	Gene 2	Gene 1 coordinates	Gene 2 coordinates	Shared	Gene 1 unique	Gene 2 unique	*P*-value	q-value	Shared cell lines	Shared	Gene 1 unique	Gene 2 unique	*P*-value	q-value
Cldn10a	*Rcbtb2*	13:94883859-95029907	13:47961104-48005317	3	4	0	9.87E-07	2.91E-04	NCI-H630 (large intestine), NCI-H508 (large intestine), COLO-205 (large intestine)	1	7	7	0.0484	0.1514
Rhbdf2	*Rnf157*	17:71978571-72009103	17:71651451-71747985	2	0	0	5.60E-06	1.35E-03	CP66MEL (skin), MDA-MB-361 (breast)	1	5	9	0.0445	0.1514
Brd2	*Rreb1*	6:33044415-33057260	6:7052829-7197212	1	0	1	3.34E-03	0.254	MG-63 (bone)	2	6	29	0.0114	0.1514
Jup	*Ccr7*	17:37164382-37196476	17:35963550-35975250	1	2	2	0.0150	0.561	NCI-N87 (stomach)	3	9	50	0.0081	0.1368
Dym	*Fgd2*	18:44824172-45241077	6:37081400-37489422	1	3	2	0.0200	0.655	TE-206-1 (unknown)	3	13	58	0.0354	0.1514
Pecam1	*Ncoa3*	17:59752964-59794505	20:45564053-45719023	1	3	2	0.0200	0.655	MCF7 (breast)	2	16	17	0.0287	0.1514
Prkcbp1	*Pecam1*	20:45271266-45418881	17:59752964-59794505	1	3	3	0.0266	0.781	MCF7 (breast)	2	18	16	0.0319	0.1514
Ccnd1	*Ppp1r10*	11:69165054-69178422	6:30676162-30692999	1	22	0	0.0385	1	TE-6 (oesophagus)	2	69	3	0.0136	0.1514

B

		Human cancer cell lines								Mouse lymphomas				
Gene 1	Gene 2	Gene 1 coordinates	Gene 2 coordinates	Shared	Gene 1 unique	Gene 2 unique	*P*-value	q-value	Shared cell lines	Shared	Gene 1 unique	Gene 2 unique	*P*-value	q-value
B230120H23Rik	*Ikzf3*	2:173648811-173840979	17:35174724-35273967	1	0	0	0.0017	0.0774	NCI-H1417 (lung)	2	10	31	0.0362	0.1514
Tcte3	*Lef1*	6:169882140-169893563	4:109188150-109309027	1	0	2	0.0050	0.1241	NCI-H1417 (lung)	3	11	27	0.0025	0.0987
Tbc1d1	*Ubash3a*	4:37569115-37817189	21:42697088-42740843	1	3	0	0.0067	0.1241	ATN-1 (leukaemia)	2	26	6	0.0092	0.1455
Flt3	*Ncoa3*	13:27475411-27572729	20:45564053-45719023	1	3	0	0.0067	0.1241	NCI-H1417 (lung)	3	32	17	0.0151	0.1514
Ubac2	*Lef1*	13:98651109-98836682	4:109188150-109309027	1	1	2	0.0100	0.1550	NCI-H1417 (lung)	4	27	27	0.0049	0.116
Lef1	*Mad1l1*	4:109188150-109309027	7:1821956-2239109	1	2	1	0.0100	0.1550	NCI-H1417 (lung)	2	27	8	0.0180	0.1514

C

		Human cancer cell lines								Mouse lymphomas				
Gene 1	Gene 2	Gene 1 coordinates	Gene 2 coordinates	Shared	Gene 1 unique	Gene 2 unique	*P*-value	q-value	Shared cell lines	Shared	Gene 1 unique	Gene 2 unique	*P*-value	q-value
2010106G01Rik	*Tgfbr3*	15:48787031-48845202	1:91918488-92144147	1	0	2	0.0050	1	MCF7 (breast)	1	6	6	0.0357	0.1514
Spata13	*D12Ertd553e*	13:23632887-23779204	2:16594890-16710580	1	0	3	0.006689	1	COLO-205 (large intestine)	2	9	14	0.00626	0.1237
Mylc2pl	*Arid1a*	7:101043326-101059296	1:26895109-26981188	1	3	1	0.0133	1	NCI-SNU-5 (stomach)	1	2	11	0.0189	0.1514
Flt3	*Chst3*	13:27475411-27572729	10:73394126-73443318	1	8	0	0.0151	1	COLO-205 (large intestine)	1	2	32	0.0313	0.1514

Table 5.11. A list of genes that are co-amplified (A), co-deleted (B) or amplified and deleted (C) across human cancer cell lines and are also co-disrupted by MuLV in mouse lymphomas. The columns labelled "Shared" shows the number of cell lines in which the genes are co-amplified/co-deleted and the number of mouse lymphomas in which both genes are disrupted by MuLV. "Gene 1 unique" and "Gene 2 unique" show the number of cell lines in which gene 1 and gene 2 are amplified/deleted without amplification/deletion of the other gene in the pair, and the number of mouse lymphomas in which only one or other gene is disrupted by MuLV. In Figure C, Gene 1 is amplified and Gene 2 is deleted. *P*-values were calculated using the 2-tailed Fisher Exact Test, *q*-values were calculated using R package QVALUE.

8 co-amplified genes in human cancer cell lines were also co-mutated by MuLV in mouse lymphomas. Claudin-10 (*CLDN10*) and *RCBTB2* were co-amplified in 3 cancer cell lines derived from the large intestine but, as discussed in Section 5.3, *RCBTB2* is not a likely target of amplification. The same applies to *RHBDF2* and *RNF157*, which were co-amplified in 2 carcinoma cell lines derived from breast and skin. The minimal amplified region for both genes encompassed 196 genes, of which 4 were known oncogenes, 189 genes were amplified in a greater number of cell lines and 130 were amplified to a higher copy number.

The remaining co-amplifications occurred in a single human cancer cell line, but since they were also co-mutated in MuLV mutagenesis, they are worthy of further investigation. Ras-responsive element binding protein 1 (*RREB1*) and bromodomain containing 2 (*BRD2*) were co-amplified in a bone cancer cell line. Both have been previously implicated in cancer but 5 of the 23 genes within the minimal amplified region containing *BRD2* were amplified in more cell lines and to greater copy number, suggesting that *BRD2* is not a critical target gene. The co-occurrence of genes junction plakoglobin (*JUP*) and C-C chemokine receptor type 7 precursor (*CCR7*) is potentially more interesting since, while co-operation between these genes has not been previously observed, both have been implicated in stomach cancer, which is the origin of the cancer cell line in which a co-occurrence was identified. *JUP* is overexpressed in an amplicon on chromosome 17q that is frequently found in gastric cancer (Varis *et al.*, 2002), while high *CCR7* expression correlates with poorer surgical outcomes in gastric cancer patients (Ishigami *et al.*, 2007). These genes showed the highest significance for MuLV co-occurrence but it is worth noting that the *q*-value was still low, at 0.137. Two of the remaining co-occurrences were in breast cancer cell line MCF7 and involved *PECAM1* plus *NCOA3* or *PRKCBP1*. Like *TMEM49*, mentioned above, *PECAM1* resides on the breast-cancer-specific amplicon mapping to chromosome 17q23, while *NCOA3* and *PRKCBP1* reside on the 20q breast cancer amplicon. These observations therefore lend further support to the theory that the 17q23 and 20q amplicons co-operate in breast tumourigenesis. The fact that MuLV insertions in *PECAM1* also co-occurred with insertions in *NCOA3* and *PRKCBP1* suggests that *PECAM1*, rather than *TMEM49*, may be the critical cancer gene involved in a co-operation with both *NCOA3* and *PRKCBP1*. However, caution must be applied since, while the *P*-values are less than 0.05, the *q*-values for both tests are considerably higher. Finally, *CCND1* and *PPP1R10* co-occurred in an oesophageal cancer cell line. PPP1R10 inhibits the catalytic subunit of protein phosphatase 1 alpha (PPP1CA) (Kim *et al.*, 2003b). Along with *CCND1*, *PPP1CA* is overexpressed in an amplicon on chromosome 11q13 that is frequently observed in oral squamous cell carcinomas (Hsu *et al.*, 2006) but also in oesophageal squamous cell carcinomas (Huang *et al.*, 2006b). PPP1CA has recently been shown to contribute to oncogenic Ras- and p53-induced cell cycle arrest (Castro *et al.*, 2008). Therefore, it is possible that the amplification of *PPP1R10* counteracts the tumour suppressive effects of *PPP1CA*, which is amplified and overexpressed due to its proximity to *CCND1* and, potentially, other amplified cancer genes. However, this hypothesis clearly needs to be verified experimentally, especially since the overlap between genes is small.

Of the 6 pairs of co-deleted CIS genes, 5 were within the lung cancer cell line NCI-H1417 and 1 was within the leukaemia cell line ATN-1. As mentioned earlier, both cell lines contain a large number of deletions, which reduces the reliability of these associations. However, deletions within at least 3 of the genes have been observed in cancer. For example, heterozygous deletions in *MAD1L1* increase the incidence of tumours in mice (Iwanaga *et al.*, 2007), while recurrent deletions in *LEF1* and *IKZF3* were detected in acute lymphoblastic leukaemias in the study by Mullighan *et al.* (2007). The 4 co-occurrences of amplified and deleted gene pairs also included 2 that were from a highly unstable cell line, COLO-205. In addition, *FLT3* was the only amplified gene for which the minimal amplified region did not contain other genes that were amplified in a greater number of cell lines and, for all associations, the *q*-value was 1. None of the associations between co-deleted or amplified and deleted genes have been previously observed in the literature, and in all cases, there was only 1 shared human cancer cell line. As all of the observed associations have a low significance, it has not been possible to identify any strong collaborations between cancer genes. However, the co-occurrence of the associations in both human and mouse provides additional evidence, and the gene pairs are therefore presented here as potential candidates for a co-operating role in tumourigenesis.

5.6 Discussion

The purpose of the work described in this chapter was to identify CIS genes that were significantly amplified or deleted in human cancer cell lines. As discussed in Chapter 4, mouse candidate cancer genes identified by retroviral insertional mutagenesis can help to narrow down the candidates in regions of copy number change in the human cancers. In this chapter, the CIS genes identified in Chapter 2 were used. These are a more reliable set of candidates than those used in Chapter 4. The other major difference between the comparative analysis described in Chapters 4 and 5 was the resolution of the copy number data used. The high-resolution SNP 6.0 data has only recently become available. Due to time constraints and the lack of published methods for dealing with data of this size, the data were subjected to the same segmentation and merging algorithms as the lower resolution, 10K dataset. This may not be entirely appropriate but for the purposes of this study, it was deemed to be sufficient since only the copy numbers at CIS genes were relevant to the analysis, and only those CIS genes that were recurrently amplified or deleted were investigated. The efficacy of the analysis is demonstrated by the over-representation of known oncogenes in amplicons, and the identification of implicated tumour suppressor genes in deletions. The high-resolution analysis identified a higher proportion of CIS genes and known cancer genes than the 10K analysis, demonstrating its superiority. It was also shown that some of the most significant candidates identified in the 10K analysis did not appear to be within the true boundaries of regions of copy number change. However, the 10K data may be helpful in filtering out genes that have been incorrectly selected due to errors in the measurement of copy number at clusters of probes or in the calling of gains and losses in the high-resolution data. These problems could also be avoided by using the raw data, rather than the segmentation and merging algorithms, and averaging the copy number values across all markers within a gene. However, this could result in the loss of relevant data in genes where small deletions and amplifications are biologically important. For example, oncogenes *NOTCH1* and *ETS1* contained intragenic deletions that were helpful in defining the mechanisms of mutation.

Known oncogenes were over-represented among significantly amplified CIS genes, demonstrating that the method used to rank genes was a successful approach for finding promising candidates. Some genes may not be amplified or deleted in a large number of cell lines across all cancer types, but may be heavily implicated in tumourigenesis in a subset of cancers. To identify such tissue-specific candidates, it is essential to perform the analysis independently on each type of cancer. The chosen method ignores genes that are amplified or deleted in few cell lines. However, it is difficult to endorse such candidates unless evidence to support a role in tumourigenesis has been previously demonstrated, as it has for *MEIS1* and *NFKB1*, which were identified in Section 4.5.2.1.1. Nevertheless, a significance threshold of $P<0.05$ is potentially too stringent. The P-value for each candidate is based on the number of amplifications compared with the distribution of the number of amplifications for non-CIS genes. Since amplicons are often large and encompass multiple genes, it is highly probable that a number of non-CIS genes will be amplified for every CIS gene, even when the CIS gene is the target of amplification and other genes are passengers. It is therefore unwise to implement a strict cut-off when selecting promising candidates, but the method does provide a way to rank the candidates so that they can be compared against one another. Ranking the candidates according to the maximum copy number, rather than the number of amplicons, provides a method for identifying alternative candidates that, like *MEIS1* and *SLAMF6*, are amplified to high copy number but not across a large number of cell lines.

The search for collaborating cancer genes was somewhat disappointing. Most genes were amplified or deleted in only a small number of cell lines and the power of the analysis was therefore insufficient. In addition, the CIS gene may not be the main target of amplification in all of the cell lines containing a copy number change in that region, which would dilute the association. The lack of an overlap between observed associations and collaborations identified in mouse lymphomas may reflect the fact that pathways and collaborations can differ between tumour types, and associations may therefore be concealed in an analysis of co-occurrences across all human cancer cell lines. However, the numbers, and therefore the power of the analysis, would be too low if each tumour type was independently analysed. The results demonstrate the complexity of human cancer genomes, where different genes are mutated in different tumours and by different mechanisms. This analysis only considers CIS genes in regions of copy number change. However, genes can also be mutated by epigenetic changes, balanced translocations, point mutations and other small, intragenic mutations. Therefore, the analysis is perhaps too restrictive and requires the use of copy number data for a larger set of human cancers and/or datasets representing other types of cancer-associated mutations.

Importantly, CIS genes were over-represented in amplicons and deletions in human cancer cell lines of diverse origin, therefore demonstrating that retroviral insertional mutagenesis in the mouse is relevant to the identification of candidate genes in human cancers derived from a range of tissues. The CIS genes presented in this chapter are stronger candidates than those presented in the gene lists of Chapter 4, and, combined with the high-resolution copy number data for human cancers, they provide a resource of promising candidates for a role in both human and mouse cancer.

Chapter 6 Summary and Conclusions

In light of recent developments in high-throughput genome analysis, genome-wide mutation datasets can be generated for large numbers of cancer genomes at increasing speed and resolution and diminishing cost. However, extracting meaningful information from these datasets can be challenging, particularly as the human and mouse cancer genomes are highly complex, with hundreds of genes being implicated in cancer, and different cancers showing high variability in the spectrum of mutated cancer genes, and in the mechanisms of mutation. The main purpose of this project was to use genome-wide, cancer-associated mutation datasets from mouse and human to facilitate the identification of human genes involved in cancer development.

The principal mouse dataset used in this project was generated by insertional mutagenesis using the retrovirus murine leukaemia virus (MuLV). Retroviral insertional mutagenesis is an established approach for cancer gene discovery in the mouse, but the elucidation of the mouse genome sequence and advances in PCR-based methods for identifying insertion sites have greatly increased the efficiency and the size of the screens that can be performed. The main aim of Chapter 2 was to identify insertion sites and candidate cancer genes within 1,005 MuLV-induced mouse tumours. Following mapping with SSAHA2, the reads, and the insertion sites into which they were clustered, were filtered to remove contaminants. While the filtering procedure is rather conservative and may result in the removal of some true positives, it is essential in large-scale analyses where it would be impractical to manually check each read independently.

An important consideration when planning an insertional mutagenesis screen is how to maximise the coverage, i.e. the proportion of insertion sites successfully identified, whilst keeping control of costs. A comparison of the insertion sites identified in two separate PCR reactions performed on the same tumours demonstrated that the screen is not fully saturated. Without data for greater numbers of enzymes, and for the same enzymes with greater sequencing depth, it was not possible to accurately determine the conditions that would maximise coverage, but it is likely that both the number of enzymes and the depth of sequencing are important factors. Future studies can benefit from next-generation sequencing technologies, such as 454 sequencing (http://www.454.com), which, compared with the traditional Sanger chain-termination methods, can sequence larger quantities of DNA at lower cost, thereby allowing for an increase in both the number of PCRs and the sequencing depth.

Insertions were assigned to genes by analysing the distribution of insertions around known cancer genes, which are assumed to be targets of mutation. The identification of common insertion sites was performed using two methods: Monte Carlo (MC) simulations (Suzuki *et al.*, 2002) and a kernel convolution (KC)-based framework (de Ridder *et al.*, 2006), and the candidate gene list generated using the KC method was chosen for subsequent analyses following a comparison based on the number of known cancer genes within the MC and KC gene lists. Some cancer-associated genes will almost certainly be missed by both methods. For example, some genes, particularly tumour suppressor genes, can be mutated in a variety of ways, and the insertions may not cluster into a sufficiently tight region to be detected as a single CIS. It is also possible that genes for which insertions are identified in just one or two tumours, and are therefore below the threshold for a significant CIS, do contribute to tumourigenesis. The paucity of insertions may reflect problems in mapping some of the insertions (see above) or may indicate that the gene is not frequently disrupted, e.g. because a specific set of co-operating cancer genes must also be mutated for the gene to contribute to tumourigenesis. However, for the purposes of this study, in which the mouse candidates were used to identify likely candidates within the human dataset, a smaller set of strong candidates is preferable to a larger set containing a high proportion of false positives. A smaller dataset of 73 mouse tumours generated using the *Sleeping Beauty* (SB) transposon was processed in a similar way to the MuLV dataset.

The main aim of Chapter 3 was to characterise the mouse candidate cancer genes identified in Chapter 2, both as a complete list and as individual candidates. A significant overlap between the mouse candidates and human cancer-associated genes within the Cancer Gene Census (Futreal *et al.*, 2004) and COSMIC database (Forbes *et al.*, 2006) demonstrated the relevance of insertional mutagenesis to human tumourigenesis. This is important on two counts, since it shows that genes disrupted by MuLV mutagenesis contribute to spontaneous cancer development, and that mouse cancer gene candidates are involved in human cancer. Interestingly, candidates were also over-represented among genes with Nanog and Oct4 binding sites, suggesting that a significant proportion may be involved in tumour cell self-renewal, which is consistent with the cancer stem cell hypothesis discussed in Section 1.2.3.2. The overlap with regions of copy number change in human paediatric acute lymphoblastic leukaemias (Mullighan *et al.*, 2007) provided the first indication that a significant proportion of the mouse candidates may be amplified or deleted in human cancer, and may help to narrow down the candidates in regions of copy number change in human cancers. The analysis also identified miRNA genes among the mouse candidates, and uncovered an over-representation of targets for 3 miRNAs (mmu-miR-449b, mmu-miR-449c and hsa-miR-565) that have not been previously implicated in tumourigenesis.

A comparison of the candidates generated by retroviral and transposon-mediated insertional mutagenesis suggested that while some genes are frequently disrupted by both mutagens, others are unique to one screen. This most likely reflects differences in the insertional bias and mechanisms of mutation of the two mutagens. It suggests that the screens are complementary and that the use of multiple mutagens should increase the yield of candidate cancer genes. The implementation of a larger *Sleeping Beauty* screen is therefore highly recommended, while the use of tissue-specific transposons would further increase the repertoire of candidates. It is, however, worth noting that some of the difference between the mutagens may reflect the incomplete saturation of the screens, which results in some insertion sites going undetected. Genes that are mutated by both MuLV and SB are very strong candidates for a role in tumourigenesis since they are unlikely to result from insertional bias, which differs between the two mutagens. The analysis revealed several MuLV- and SB-disrupted genes that warrant further investigation, including *p116Rip*, *Zmiz1*, *ENSMUSG00000075105* and *Qsk*. Preliminary work has already commenced into the functional validation of *QSK*, and results have demonstrated that knockdown of the gene in human HeLa cells causes chromosome lagging, which can lead to aneuploidy and cancer formation.

The chapter also focuses on the analysis of the distribution of insertions in and around cancer genes, and this was used to predict the likely mechanism of mutation of candidate genes. For example, *Ccr7* and *Jundm2* were predicted to be mutated by the same mechanism as *Mycn* and *Pim1*, wherein insertions are known to cause premature termination of gene transcription that results in the removal of mRNA-destabilising motifs and, therefore, a more stable gene transcript. This analysis is complicated by the number of ways in which the mutagen can disrupt a gene, and would be aided by using a transposon with a more limited repertoire of mutational mechanisms, e.g. the ability to induce C-terminal truncations in one orientation only. For genes that are mutated by both MuLV and SB, determination of the mechanism of mutation is facilitated by the co-analysis of MuLV and SB insertions. For example, *Notch1* contains both MuLV and SB insertions in 3 distinct locations of the gene that represent different types of oncogenic mutation, and the resulting gene products have been observed in human cancers. Elucidation of the mechanism of mutation may provide an insight into the role of a gene in cancer, and may facilitate the development of therapies targeted against specific mutants. The distribution of insertions can also help to distinguish oncogenes, in which insertions often form distinct clusters, from tumour suppressor genes, in which insertions are more likely to be scattered throughout the gene and may be more likely to include multiple insertions from the same tumour, potentially representing inactivation of both gene copies. As well as a number of known and implicated tumour suppressor genes, this analysis identified *Qsk*, *Smg6* and *Foxp1* as potential tumour suppressor genes. Interestingly, compared with random insertions, those associated with candidate cancer genes were under-represented in a predicted set of gene-associated regulatory regions and, specifically, within predicted regulatory features associated with active genes. This supports the notion that these insertions are oncogenic and do not simply result from a bias towards insertion into the 5' ends of transcriptionally active genes. It also shows that disruption of regulatory regions is not a common mechanism of mutagenesis for MuLV.

The final part of this chapter was concerned with identifying genes that co-operate in cancer development. The aim of the first analysis was to identify genes that contain an over-representation of insertions in tumours from mice deficient in a particular tumour suppressor gene, therefore suggesting that the genes may collaborate with the loss of that tumour suppressor gene. A number of the significant associations were supported by evidence in the literature, suggesting that many of the unsupported associations may also be real. The same applies to the second analysis, in which pairs of genes that were co-mutated in a significant number of tumours were identified. Collaborating cancer genes are of particular interest because they can help to elucidate cancer pathways and may represent suitable targets for combined therapies, which may have a lower rate of resistance than therapies targeted to a single gene. Interesting collaborations included positive associations between the loss of *Cdkn2a* and MuLV-disrupted *Zeb2* and *Ccnd1*, and between MuLV-disrupted *Lck* and *A530013C23Rik*, *Stat5b* and *Csk*. The analyses also identified a number of mutually exclusive genes that co-occur less frequently than expected by chance. Such genes may act in the same pathway, and are therefore helpful in elucidating cancer pathways. Understanding the pathway in which a gene acts may help in the development of cancer therapies since, in some cases, it may be more effective to target a gene that acts downstream of the mutated gene, rather than targeting the mutated gene itself.

In the remaining chapters, mouse candidate cancer genes identified by MuLV insertional mutagenesis were used in cross-species comparative analyses with copy number data from human cancer cell lines. The main aims of the analyses were to demonstrate that retroviral insertional mutagenesis is relevant to the discovery of cancer genes that are amplified or deleted in human cancer, and to help narrow down the candidates within regions of copy number change, which are often large and encompass many genes. In the course of this project, there have been rapid developments in genome-wide copy number analysis and therefore two datasets were used, the second being of much higher resolution and representing the current state of the art. However, even using the lower resolution dataset, it was clear that mouse candidate oncogenes were over-represented in regions of copy number gain and tumour suppressor genes, although less likely to be identified by insertional mutagenesis, were slightly over-represented in regions of copy number loss. The association was observed in cell lines derived from both haematopoietic and lymphoid cancers and solid tumours. Therefore, despite the fact that MuLV insertional mutagenesis generates mouse lymphomas and the resulting candidate cancer genes showed an over-representation of GO terms related to the activation and differentiation of B- and T-cells, the study is relevant to the identification of cancer genes in a diverse range of human cancers. Lymphomas are the most common cancer in mice. As mentioned in Section 1.4.1.2.2, *Trp53* is mutated in many types of human cancer, but mutation in the mouse leads to the development of lymphomas or sarcomas (Jonkers and Berns, 2002). Therefore, simply because mutation in a certain gene leads to lymphomagenesis in the mouse, it does not necessarily follow that mutation of the orthologous human gene would result in the same type of cancer. In addition, insertional mutagenesis may result in genes being switched on in cells that would never normally express the gene, even through translocation. Comparison with human cancer-associated mutation data helps to demonstrate how, and whether, the MuLV-disrupted genes contribute to the formation of spontaneous human tumours.

The CGH data from both the 10K and SNP 6.0 SNP arrays were processed into regions of copy number change before being compared to the mouse dataset. Most of the available methods were designed for processing BAC array CGH data, but the chosen method of DNAcopy (Olshen *et al.*, 2004) plus MergeLevels (Willenbrock and Fridlyand, 2005), was found to be suitable for the 10K SNP array data. The limiting factor was the resolution of the data, since regions of copy number change may be missed or may be falsely predicted. For the higher resolution data, the analysis was limited by the lack of available methods specifically developed for high-density SNP arrays. However, comparative analyses using both datasets uncovered a significant number of known oncogenes in regions of copy number gain, therefore demonstrating the efficacy of the methods. In the lower resolution analysis, interesting candidate oncogenes that were disrupted in the mouse and amplified in human cancer included *SLAMF6*, *MMP13*, *RREB1* and *TAOK3*, while candidate tumour suppressor genes included *ARFRP2*, *SCFD2*, *RBMS3*, *UTRN*, *ANK3* and *ACCN1*. Further candidates are provided in Section 4.5.2. Recurrent amplification of *hsa-miR-23a* and deletion of *hsa-miR-128b* were also demonstrated. In the high-resolution analysis, candidate oncogenes included *ITPR2*, *FAM49A*, *BCL11B*, *TMEM49*, *SUPT3H* and *CLDN10*. Candidate tumour suppressor genes included *DYM*, *CYB5*, *NTN1* and *SKI*. As anticipated, the high-resolution data appeared to provide a better representation of the cancer genome than did the lower resolution data.

The development of high-resolution platforms for copy number analysis has helped to more accurately define regions of copy number change, but new or modified algorithms are required to deal with the data. The Wellcome Trust Sanger Institute Cancer Genome Project has very recently unveiled a Hidden Markov Model (HMM)-based method, in which data are segmented and are assigned to a state representing the copy number level, which is based on both SNP intensity values and allele ratios (Greenman *et al.*, unpublished). This method analyses each sample independently, and takes account of the genotype at each SNP, making it possible to identify complex, yet fairly common, changes such as loss of heterozygosity without decrease in copy number, and decrease in copy number without loss of heterozygosity. Also in development is a modification of Christiaan Klijn's KC-SMART method (Klijn *et al.*, 2008), in which non-discretised CGH data are input into the program and regions that are significantly aberrant across all tumours are detected, enabling the detection of recurrent amplicons and deletions. Both methods allow for deviation from the 3 states (gain, loss or no change) upon which many previous methods are dependent, therefore permitting more accurate analysis of heterogeneous and polyploid samples.

Finally, analyses were performed to identify co-amplified and co-deleted candidate cancer genes, and candidate genes that were preferentially amplified or deleted in cell lines containing a somatic mutation in *CDKN2A* or *TP53*. There was evidence for co-operation between the breast-cancer-specific amplicons on chromosomes 17q23 and 20q. Other possible associations were also identified, e.g. deletion of *PML* was found to be both negatively associated with mutation of *TP53* and positively associated with deletion of *MAD1L1*, while *JUP* and *CCR7* were co-amplified. However, the *P*-values and *q*-values for all associations were relatively high, and there was no clear overlap with co-disrupted genes in the retroviral screen. The most likely explanation for the lack of significant associations identified by these analyses is that the dataset is simply not large enough. Although 598 cancer cell lines were used, many of these do not have highly unstable genomes, and therefore contain few regions of copy number change. In addition, the cell lines are derived from a variety of tissues and many amplicons and deletions appear to be cell-type-specific, but considering each cancer type independently further reduces the power of the analysis. Finally, there are a number of different mechanisms by which a gene may be disrupted in cancer, and therefore some of the genes that are co-disrupted in the MuLV screen may be mutated by another mechanism, e.g. point mutation or hypermethylation, in human cancers.

Copy number changes are only one feature of cancer genomes and, therefore, a greater understanding of cancers and the pathways involved in cancer development could be achieved by integrating additional, complementary, cancer-associated mutation datasets. For example, a more in-depth comparison could be performed between the insertional mutagenesis data and mutations within the COSMIC database (Forbes *et al.*, 2006). In addition, some of the candidates identified in the insertional mutagenesis screen are known to be aberrantly methylated in cancer, e.g. *RASSF2*, *LAPTM5*, *PARK2* and *TSPAN2*, and a comparison with human epigenetic data may reveal further candidates. The human copy number data could also be compared to copy number data generated in the mouse, to identify regions that overlap in both species. In future, when datasets are available for large-scale, tissue-specific insertional mutagenesis screens, it will be possible to compare these with specific types of human cancer to identify tissue-specific cancer genes.

While informatics has taken on an increasingly important role in cancer genomics, it is still essential that candidate cancer genes identified *in silico* are experimentally validated. This thesis provides a number of strong candidates but their role is not yet proven. To demonstrate a possible involvement in tumourigenesis, the gene of interest can be knocked out or overexpressed in an animal model, such as zebrafish or mouse. Co-operation between cancer genes can be verified by knocking out or overexpressing both genes and comparing the phenotype to single mutants. The creation of a transgenic or endogenous mouse model is a lengthy procedure, and a less protracted method for validating oncogenes involves directly studying the effect on transgenic or chimeric founder mice. However, the use of animal models does not prove that the gene is important in human cancer. As discussed for *QSK*, the candidate gene can be knocked down in human cells using RNAi (RNA interference) and the effect on cell division and cell growth observed. For candidates that are amplified or deleted in human cancers, expression of those genes can be measured to determine whether there is any significant difference in the expression levels of cancers bearing the copy number change versus those that do not. Candidate genes can also be resequenced across multiple cancers to determine whether they contain somatic mutations in cancer.

The work described in this thesis provides a detailed analysis and characterisation of a large-scale retroviral insertional mutagenesis screen, and demonstrates the relevance of insertional mutagenesis to the discovery of human cancer genes. A selection of strong candidates for a role in tumourigenesis are presented, including mouse candidate cancer genes identified by both MuLV and SB insertional mutagenesis, co-operating mouse cancer genes, and human candidate cancer genes that are both disrupted by MuLV in the mouse and amplified or deleted in human cancers.

References

Abou-Sleiman, P. M., Healy, D. G., and Wood, N. W. (2004). Causes of Parkinson's disease: genetics of DJ-1. *Cell Tissue Res* **318**, 185-188.

Adams, J. M., Harris, A. W., Pinkert, C. A., Corcoran, L. M., Alexander, W. S., Cory, S., Palmiter, R. D., and Brinster, R. L. (1985). The c-myc oncogene driven by immunoglobulin enhancers induces lymphoid malignancy in transgenic mice. *Nature* **318**, 533-538.

Agirre, X., Roman-Gomez, J., Vazquez, I., Jimenez-Velasco, A., Garate, L., Montiel-Duarte, C., Artieda, P., Cordeu, L., Lahortiga, I., Calasanz, M. J.*, et al.* (2006). Abnormal methylation of the common PARK2 and PACRG promoter is associated with downregulation of gene expression in acute lymphoblastic leukemia and chronic myeloid leukemia. *Int J Cancer* **118**, 1945-1953.

Aguirre, A. J., Brennan, C., Bailey, G., Sinha, R., Feng, B., Leo, C., Zhang, Y., Zhang, J., Gans, J. D., Bardeesy, N.*, et al.* (2004). High-resolution characterization of the pancreatic adenocarcinoma genome. *Proc Natl Acad Sci U S A* **101**, 9067-9072.

Akagi, K., Suzuki, T., Stephens, R. M., Jenkins, N. A., and Copeland, N. G. (2004). RTCGD: retroviral tagged cancer gene database. *Nucleic Acids Res* **32**, D523-527.

Akervall, J., Bockmuhl, U., Petersen, I., Yang, K., Carey, T. E., and Kurnit, D. M. (2003). The gene ratios c-MYC:cyclin-dependent kinase (CDK)N2A and CCND1:CDKN2A correlate with poor prognosis in squamous cell carcinoma of the head and neck. *Clin Cancer Res* **9**, 1750-1755.

Akino, K., Toyota, M., Suzuki, H., Mita, H., Sasaki, Y., Ohe-Toyota, M., Issa, J. P., Hinoda, Y., Imai, K., and Tokino, T. (2005). The Ras effector RASSF2 is a novel tumor-suppressor gene in human colorectal cancer. *Gastroenterology* **129**, 156-169.

Alaminos, M., Davalos, V., Ropero, S., Setien, F., Paz, M. F., Herranz, M., Fraga, M. F., Mora, J., Cheung, N. K., Gerald, W. L., and Esteller, M. (2005). EMP3, a myelin-related gene located in the critical 19q13.3 region, is epigenetically silenced and exhibits features of a candidate tumor suppressor in glioma and neuroblastoma. *Cancer Res* **65**, 2565-2571.

Albertson, D. G., Collins, C., McCormick, F., and Gray, J. W. (2003). Chromosome aberrations in solid tumors. *Nat Genet* **34**, 369-376.

Albertson, D. G., and Pinkel, D. (2003). Genomic microarrays in human genetic disease and cancer. *Hum Mol Genet* **12 Spec No 2**, R145-152.

Albertson, D. G., Ylstra, B., Segraves, R., Collins, C., Dairkee, S. H., Kowbel, D., Kuo, W. L., Gray, J. W., and Pinkel, D. (2000). Quantitative mapping of amplicon structure by array CGH identifies CYP24 as a candidate oncogene. *Nat Genet* **25**, 144-146.

Aldred, M. A., Huang, Y., Liyanarachchi, S., Pellegata, N. S., Gimm, O., Jhiang, S., Davuluri, R. V., de la Chapelle, A., and Eng, C. (2004). Papillary and follicular thyroid carcinomas show distinctly different microarray expression profiles and can be distinguished by a minimum of five genes. *J Clin Oncol* **22**, 3531-3539.

Alimirah, F., Panchanathan, R., Davis, F. J., Chen, J., and Choubey, D. (2007). Restoration of p53 expression in human cancer cell lines upregulates the expression of Notch1: implications for cancer cell fate determination after genotoxic stress. *Neoplasia* **9**, 427-434.

Allen, C. E., Muthusamy, N., Weisbrode, S. E., Hong, J. W., and Wu, L. C. (2002). Developmental anomalies and neoplasia in animals and cells deficient in the large zinc finger protein KRC. *Genes Chromosomes Cancer* **35**, 287-298.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. *J Mol Biol* **215**, 403-410.

An, W., Han, J. S., Wheelan, S. J., Davis, E. S., Coombes, C. E., Ye, P., Triplett, C., and Boeke, J. D. (2006). Active retrotransposition by a synthetic L1 element in mice. *Proc Natl Acad Sci U S A* **103**, 18662-18667.

Aparecida Alves, C., Silva, I. D., Villanova, F. E., Nicolau, S. M., Custodio, M. A., Bortoletto, C., and Goncalves, W. J. (2006). Differential gene expression profile reveals overexpression of MAP3K8 in invasive endometrioid carcinoma. *Eur J Gynaecol Oncol* **27**, 589-593.

Apperley, J. F., Gardembas, M., Melo, J. V., Russell-Jones, R., Bain, B. J., Baxter, E. J., Chase, A., Chessells, J. M., Colombat, M., Dearden, C. E*., et al.* (2002). Response to imatinib mesylate in patients with chronic myeloproliferative diseases with rearrangements of the platelet-derived growth factor receptor beta. *N Engl J Med* **347**, 481-487.

Aqeilan, R. I., Trapasso, F., Hussain, S., Costinean, S., Marshall, D., Pekarsky, Y., Hagan, J. P., Zanesi, N., Kaou, M., Stein, G. S*., et al.* (2007). Targeted deletion of Wwox reveals a tumor suppressor function. *Proc Natl Acad Sci U S A* **104**, 3949-3954.

Arber, N., Lightdale, C., Rotterdam, H., Han, K. H., Sgambato, A., Yap, E., Ahsan, H., Finegold, J., Stevens, P. D., Green, P. H*., et al.* (1996). Increased expression of the cyclin D1 gene in Barrett's esophagus. *Cancer Epidemiol Biomarkers Prev* **5**, 457-459.

Arena, S., Isella, C., Martini, M., de Marco, A., Medico, E., and Bardelli, A. (2007). Knock-in of oncogenic Kras does not transform mouse somatic cells but triggers a transcriptional response that classifies human cancers. *Cancer Res* **67**, 8468-8476.

Armitage, P., and Doll, R. (1954). The age distribution of cancer and a multi-stage theory of carcinogenesis. *Br J Cancer* **8**, 1-12.

Artandi, S. E., Chang, S., Lee, S. L., Alson, S., Gottlieb, G. J., Chin, L., and DePinho, R. A. (2000). Telomere dysfunction promotes non-reciprocal translocations and epithelial cancers in mice. *Nature* **406**, 641-645.

Ashley, D. J. (1969). The two "hit" and multiple "hit" theories of carcinogenesis. *Br J Cancer* **23**, 313-328.

Aster, J. C., Pear, W. S., and Blacklow, S. C. (2008). Notch signaling in leukemia. *Annu Rev Pathol* **3**, 587-613.

Audet, J., Miller, C. L., Rose-John, S., Piret, J. M., and Eaves, C. J. (2001). Distinct role of gp130 activation in promoting self-renewal divisions by mitogenically stimulated murine hematopoietic stem cells. *Proc Natl Acad Sci U S A* **98**, 1757-1762.

Babushok, D. V., Ostertag, E. M., Courtney, C. E., Choi, J. M., and Kazazian, H. H., Jr. (2006). L1 integration in a transgenic mouse model. *Genome Res* **16**, 240-250.

Bachman, K. E., Argani, P., Samuels, Y., Silliman, N., Ptak, J., Szabo, S., Konishi, H., Karakas, B., Blair, B. G., Lin, C*., et al.* (2004). The PIK3CA gene is mutated with high frequency in human breast cancers. *Cancer Biol Ther* **3**, 772-775.

Balak, M. N., Gong, Y., Riely, G. J., Somwar, R., Li, A. R., Zakowski, M. F., Chiang, A., Yang, G., Ouerfelli, O., Kris, M. G*., et al.* (2006). Novel D761Y and common secondary T790M mutations in epidermal growth factor receptor-mutant lung adenocarcinomas with acquired resistance to kinase inhibitors. *Clin Cancer Res* **12**, 6494-6501.

Balakrishnan, A., von Neuhoff, N., Rudolph, C., Kamphues, K., Schraders, M., Groenen, P., van Krieken, J. H., Callet-Bauchu, E., Schlegelberger, B., and Steinemann, D. (2006). Quantitative microsatellite analysis to delineate the commonly deleted region 1p22.3 in mantle cell lymphomas. *Genes Chromosomes Cancer* **45**, 883-892.

Balliet, A. G., Hatton, K. S., Hoffman, B., and Liebermann, D. A. (2001). Comparative analysis of the genetic structure and chromosomal location of the murine MyD118 (Gadd45beta) gene. *DNA Cell Biol* **20**, 239-247.

Banham, A. H., Beasley, N., Campo, E., Fernandez, P. L., Fidler, C., Gatter, K., Jones, M., Mason, D. Y., Prime, J. E., Trougouboff, P*., et al.* (2001). The FOXP1 winged helix transcription factor is a novel candidate tumor suppressor gene on chromosome 3p. *Cancer Res* **61**, 8820-8829.

Banham, A. H., Connors, J. M., Brown, P. J., Cordell, J. L., Ott, G., Sreenivasan, G., Farinha, P., Horsman, D. E., and Gascoyne, R. D. (2005). Expression of the FOXP1 transcription factor is strongly associated with inferior survival in patients with diffuse large B-cell lymphoma. *Clin Cancer Res* **11**, 1065-1072.

Bardelli, A., Parsons, D. W., Silliman, N., Ptak, J., Szabo, S., Saha, S., Markowitz, S., Willson, J. K., Parmigiani, G., Kinzler, K. W*., et al.* (2003). Mutational analysis of the tyrosine kinome in colorectal cancers. *Science* **300**, 949.

Baron, U., and Bujard, H. (2000). Tet repressor-based system for regulated gene expression in eukaryotic cells: principles and advances. *Methods Enzymol* **327**, 401-421.

Baross, A., Delaney, A. D., Li, H. I., Nayar, T., Flibotte, S., Qian, H., Chan, S. Y., Asano, J., Ally, A., Cao, M*., et al.* (2007). Assessment of algorithms for high throughput detection of genomic copy number variation in oligonucleotide microarray data. *BMC Bioinformatics* **8**, 368.

Barski, A., Cuddapah, S., Cui, K., Roh, T. Y., Schones, D. E., Wang, Z., Wei, G., Chepelev, I., and Zhao, K. (2007). High-resolution profiling of histone methylations in the human genome. *Cell* **129**, 823-837.

Baselga, J. (2006). Targeting tyrosine kinases in cancer: the second wave. *Science* **312**, 1175-1178.

Bashir, A., Volik, S., Collins, C., Bafna, V., and Raphael, B. J. (2008). Evaluation of paired-end sequencing strategies for detection of genome rearrangements in cancer. *PLoS Comput Biol* **4**, e1000051.

Bassermann, F., Frescas, D., Guardavaccaro, D., Busino, L., Peschiaroli, A., and Pagano, M. (2008). The Cdc14B-Cdh1-Plk1 axis controls the G2 DNA-damage-response checkpoint. *Cell* **134**, 256-267.

Basuyaux, J. P., Ferreira, E., Stehelin, D., and Buttice, G. (1997). The Ets transcription factors interact with each other and with the c-Fos/c-Jun complex via distinct protein domains in a DNA-dependent and -independent manner. *J Biol Chem* **272**, 26188-26195.

Bedigian, H. G., Johnson, D. A., Jenkins, N. A., Copeland, N. G., and Evans, R. (1984). Spontaneous and induced leukemias of myeloid origin in recombinant inbred BXH mice. *J Virol* **51**, 586-594.

Bednarek, A. K., Laflin, K. J., Daniel, R. L., Liao, Q., Hawkins, K. A., and Aldaz, C. M. (2000). WWOX, a novel WW domain-containing protein mapping to human chromosome 16q23.3-24.1, a region frequently affected in breast cancer. *Cancer Res* **60**, 2140-2145.

Beer, D. G., Kardia, S. L., Huang, C. C., Giordano, T. J., Levin, A. M., Misek, D. E., Lin, L., Chen, G., Gharib, T. G., Thomas, D. G*., et al.* (2002). Gene-expression profiles predict survival of patients with lung adenocarcinoma. *Nat Med* **8**, 816-824.

Beliakoff, J., Lee, J., Ueno, H., Aiyer, A., Weissman, I. L., Barsh, G. S., Cardiff, R. D., and Sun, Z. (2008). The PIAS-like protein Zimp10 is essential for embryonic viability and proper vascular development. *Mol Cell Biol* **28**, 282-292.

Beliakoff, J., and Sun, Z. (2006). Zimp7 and Zimp10, two novel PIAS-like proteins, function as androgen receptor coregulators. *Nucl Recept Signal* **4**, e017.

Beltran, M., Puig, I., Pena, C., Garcia, J. M., Alvarez, A. B., Pena, R., Bonilla, F., and de Herreros, A. G. (2008). A natural antisense transcript regulates Zeb2/Sip1 gene expression during Snail1-induced epithelial-mesenchymal transition. *Genes Dev* **22**, 756-769.

Ben-Porath, I., Thomson, M. W., Carey, V. J., Ge, R., Bell, G. W., Regev, A., and Weinberg, R. A. (2008). An embryonic stem cell-like gene expression signature in poorly differentiated aggressive human tumors. *Nat Genet* **40**, 499-507.

Berenblum, I., and Shubik, P. (1949). An experimental study of the initiating state of carcinogenesis, and a re-examination of the somatic cell mutation theory of cancer. *Br J Cancer* **3**, 109-118.

Bergman, M., Mustelin, T., Oetken, C., Partanen, J., Flint, N. A., Amrein, K. E., Autero, M., Burn, P., and Alitalo, K. (1992). The human p50csk tyrosine kinase phosphorylates p56lck at Tyr-505 and down regulates its catalytic activity. *Embo J* **11**, 2919-2924.

Bergmann, E., Wanzel, M., Weber, A., Shin, I., Christiansen, H., and Eilers, M. (2001). Expression of P27(KIP1) is prognostic and independent of MYCN amplification in human neuroblastoma. *Int J Cancer* **95**, 176-183.

Bernardini, M., Lee, C. H., Beheshti, B., Prasad, M., Albert, M., Marrano, P., Begley, H., Shaw, P., Covens, A., Murphy, J.*, et al.* (2005). High-resolution mapping of genomic imbalance and identification of gene expression profiles associated with differential chemotherapy response in serous epithelial ovarian cancer. *Neoplasia* **7**, 603-613.

Bernatsky, S., Ramsey-Goldman, R., Isenberg, D., Rahman, A., Dooley, M. A., Sibley, J., Boivin, J. F., Joseph, L., Armitage, J., Zoma, A., and Clarke, A. (2007). Hodgkin's lymphoma in systemic lupus erythematosus. *Rheumatology (Oxford)* **46**, 830-832.

Berns, K., Horlings, H. M., Hennessy, B. T., Madiredjo, M., Hijmans, E. M., Beelen, K., Linn, S. C., Gonzalez-Angulo, A. M., Stemke-Hale, K., Hauptmann, M.*, et al.* (2007). A functional genetic approach identifies the PI3K pathway as a major determinant of trastuzumab resistance in breast cancer. *Cancer Cell* **12**, 395-402.

Bessette, D. C., Wong, P. C., and Pallen, C. J. (2007). PRL-3: a metastasis-associated phosphatase in search of a function. *Cells Tissues Organs* **185**, 232-236.

Besseyrias, V., Fiorini, E., Strobl, L. J., Zimber-Strobl, U., Dumortier, A., Koch, U., Arcangeli, M. L., Ezine, S., Macdonald, H. R., and Radtke, F. (2007). Hierarchy of Notch-Delta interactions promoting T cell lineage commitment and maturation. *J Exp Med* **204**, 331-343.

Bestor, T. H. (2005). Transposons reanimated in mice. *Cell* **122**, 322-325.

Bettencourt-Dias, M., Giet, R., Sinka, R., Mazumdar, A., Lock, W. G., Balloux, F., Zafiropoulos, P. J., Yamaguchi, S., Winter, S., Carthew, R. W.*, et al.* (2004). Genome-wide survey of protein kinases required for cell cycle progression. *Nature* **432**, 980-987.

Beverly, L. J., and Capobianco, A. J. (2003). Perturbation of Ikaros isoform selection by MLV integration is a cooperative event in Notch(IC)-induced T cell leukemogenesis. *Cancer Cell* **3**, 551-564.

Beverly, L. J., Felsher, D. W., and Capobianco, A. J. (2005). Suppression of p53 by Notch in lymphomagenesis: implications for initiation and regression. *Cancer Res* **65**, 7159-7168.

Bhattacharjee, A., Richards, W. G., Staunton, J., Li, C., Monti, S., Vasa, P., Ladd, C., Beheshti, J., Bueno, R., Gillette, M.*, et al.* (2001). Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses. *Proc Natl Acad Sci U S A* **98**, 13790-13795.

Bignell, G. R., Huang, J., Greshock, J., Watt, S., Butler, A., West, S., Grigorova, M., Jones, K. W., Wei, W., Stratton, M. R.*, et al.* (2004). High-resolution analysis of DNA copy number using oligonucleotide microarrays. *Genome Res* **14**, 287-295.

Bignell, G. R., Santarius, T., Pole, J. C., Butler, A. P., Perry, J., Pleasance, E., Greenman, C., Menzies, A., Taylor, S., Edkins, S.*, et al.* (2007). Architectures of somatic genomic rearrangement in human cancer amplicons at sequence-level resolution. *Genome Res* **17**, 1296-1303.

Bjorck, E., Ek, S., Landgren, O., Jerkeman, M., Ehinger, M., Bjorkholm, M., Borrebaeck, C. A., Porwit-MacDonald, A., and Nordenskjold, M. (2005). High expression of cyclin B1 predicts a favorable outcome in patients with follicular lymphoma. *Blood* **105**, 2908-2915.

Blank, U., Karlsson, G., Moody, J. L., Utsugisawa, T., Magnusson, M., Singbrant, S., Larsson, J., and Karlsson, S. (2006). Smad7 promotes self-renewal of hematopoietic stem cells. *Blood* **108**, 4246-4254.

Bohlander, S. K. (2005). ETV6: a versatile player in leukemogenesis. *Semin Cancer Biol* **15**, 162-174.

Booken, N., Gratchev, A., Utikal, J., Weiss, C., Yu, X., Qadoumi, M., Schmuth, M., Sepp, N., Nashan, D., Rass, K.*, et al.* (2008). Sezary syndrome is a unique cutaneous T-cell lymphoma as identified by an expanded gene signature including diagnostic marker molecules CDO1 and DNM3. *Leukemia* **22**, 393-399.

Bottinger, E. P., Jakubczak, J. L., Haines, D. C., Bagnall, K., and Wakefield, L. M. (1997). Transgenic mice overexpressing a dominant-negative mutant type II transforming growth factor beta receptor show enhanced tumorigenesis in the mammary gland and lung in response to the carcinogen 7,12-dimethylbenz-[a]-anthracene. *Cancer Res* **57**, 5564-5570.

Boulay, J. L., Mild, G., Reuter, J., Lagrange, M., Terracciano, L., Lowy, A., Laffer, U., Orth, B., Metzger, U., Stamm, B.*, et al.* (2001). Combined copy status of 18q21 genes in colorectal cancer shows frequent retention of SMAD7. *Genes Chromosomes Cancer* **31**, 240-247.

Boveri, T. (1926, 1914). *The Origin of Malignant Tumours* (Baltimore, MD: Williams & Wilkins).

Boyer Arnold, N., and Korc, M. (2005). Smad7 abrogates transforming growth factor-beta1-mediated growth inhibition in COLO-357 cells through functional inactivation of the retinoblastoma protein. *J Biol Chem* **280**, 21858-21866.

Boyer, L. A., Lee, T. I., Cole, M. F., Johnstone, S. E., Levine, S. S., Zucker, J. P., Guenther, M. G., Kumar, R. M., Murray, H. L., Jenner, R. G.*, et al.* (2005). Core transcriptional regulatory circuitry in human embryonic stem cells. *Cell* **122**, 947-956.

Boyle, A. P., Davis, S., Shulha, H. P., Meltzer, P., Margulies, E. H., Weng, Z., Furey, T. S., and Crawford, G. E. (2008). High-resolution mapping and characterization of open chromatin across the genome. Cell 132, 311-322.

Brand, M., Yamamoto, K., Staub, A., and Tora, L. (1999). Identification of TATA-binding protein-free TAFII-containing complex subunits suggests a role in nucleosome acetylation and signal transduction. *J Biol Chem* **274**, 18285-18289.

Brodeur, G. M., Seeger, R. C., Schwab, M., Varmus, H. E., and Bishop, J. M. (1984). Amplification of N-myc in untreated human neuroblastomas correlates with advanced disease stage. *Science* **224**, 1121-1124.

Brodeur, G. M., Seeger, R. C., Schwab, M., Varmus, H. E., and Bishop, J. M. (1985). Amplification of N-myc sequences in primary human neuroblastomas: correlation with advanced disease stage. *Prog Clin Biol Res* **175**, 105-113.

Brugge, J., Hung, M. C., and Mills, G. B. (2007). A new mutational AKTivation in the PI3K pathway. *Cancer Cell* **12**, 104-107.

Buchdunger, E., Zimmermann, J., Mett, H., Meyer, T., Muller, M., Druker, B. J., and Lydon, N. B. (1996). Inhibition of the Abl protein-tyrosine kinase in vitro and in vivo by a 2-phenylaminopyrimidine derivative. *Cancer Res* **56**, 100-104.

Burmester, J. K., Suarez, B. K., Lin, J. H., Jin, C. H., Miller, R. D., Zhang, K. Q., Salzman, S. A., Reding, D. J., and Catalona, W. J. (2004). Analysis of candidate genes for prostate cancer. *Hum Hered* **57**, 172-178.

Buschges, R., Weber, R. G., Actor, B., Lichter, P., Collins, V. P., and Reifenberger, G. (1999). Amplification and expression of cyclin D genes (CCND1, CCND2 and CCND3) in human malignant gliomas. *Brain Pathol* **9**, 435-442; discussion 432-433.

Bushman, F., Lewinski, M., Ciuffi, A., Barr, S., Leipzig, J., Hannenhalli, S., and Hoffmann, C. (2005). Genome-wide analysis of retroviral DNA integration. *Nat Rev Microbiol* **3**, 848-858.

Busygina, V., Kottemann, M. C., Scott, K. L., Plon, S. E., and Bale, A. E. (2006). Multiple endocrine neoplasia type 1 interacts with forkhead transcription factor CHES1 in DNA damage response. *Cancer Res* **66**, 8397-8403.

Buttice, G., Duterque-Coquillaud, M., Basuyaux, J. P., Carrere, S., Kurkinen, M., and Stehelin, D. (1996). Erg, an Ets-family member, differentially regulates human collagenase1 (MMP1) and stromelysin1 (MMP3) gene expression by physically interacting with the Fos/Jun complex. *Oncogene* **13**, 2297-2306.

Cadinanos, J., and Bradley, A. (2007). Generation of an inducible and optimized piggyBac transposon system. *Nucleic Acids Res* **35**, e87.

Calhoun, E. S., Gallmeier, E., Cunningham, S. C., Eshleman, J. R., Hruban, R. H., and Kern, S. E. (2006). Copy-number methods dramatically underestimate loss of heterozygosity in cancer. *Genes Chromosomes Cancer* **45**, 1070-1071.

Callinan, P. A., and Feinberg, A. P. (2006). The emerging science of epigenomics. *Hum Mol Genet* **15 Spec No 1**, R95-101.

Calvo, K. R., Sykes, D. B., Pasillas, M. P., and Kamps, M. P. (2002). Nup98-HoxA9 immortalizes myeloid progenitors, enforces expression of Hoxa9, Hoxa7 and Meis1, and alters cytokine-specific responses in a manner similar to that induced by retroviral co-expression of Hoxa9 and Meis1. *Oncogene* **21**, 4247-4256.

Campbell, I. G., Russell, S. E., Choong, D. Y., Montgomery, K. G., Ciavarella, M. L., Hooi, C. S., Cristiano, B. E., Pearson, R. B., and Phillips, W. A. (2004). Mutation of the PIK3CA gene in ovarian and breast cancer. *Cancer Res* **64**, 7678-7681.

Campbell, P. J., Stephens, P. J., Pleasance, E. D., O'Meara, S., Li, H., Santarius, T., Stebbings, L. A., Leroy, C., Edkins, S., Hardy, C., *et al.* (2008). Identification of somatically acquired rearrangements in cancer using genome-wide massively parallel paired-end sequencing. *Nat Genet* **40**, 722-729.

Cardone, M., Kandilci, A., Carella, C., Nilsson, J. A., Brennan, J. A., Sirma, S., Ozbek, U., Boyd, K., Cleveland, J. L., and Grosveld, G. C. (2005). The novel ETS factor TEL2 cooperates with Myc in B lymphomagenesis. *Mol Cell Biol* **25**, 2395-2405.

Caren, H., Fransson, S., Ejeskar, K., Kogner, P., and Martinsson, T. (2007). Genetic and epigenetic changes in the common 1p36 deletion in neuroblastoma tumours. *Br J Cancer* **97**, 1416-1424.

Carinci, F., Lo Muzio, L., Piattelli, A., Rubini, C., Chiesa, F., Ionna, F., Palmieri, A., Maiorano, E., Pastore, A., Laino, G*., et al.* (2005). Potential markers of tongue tumor progression selected by cDNA microarray. *Int J Immunopathol Pharmacol* **18**, 513-524.

Carlson, C. M., Dupuy, A. J., Fritz, S., Roberg-Perez, K. J., Fletcher, C. F., and Largaespada, D. A. (2003). Transposon mutagenesis of the mouse germline. *Genetics* **165**, 243-256.

Carrere, S., Verger, A., Flourens, A., Stehelin, D., and Duterque-Coquillaud, M. (1998). Erg proteins, transcription factors of the Ets family, form homo, heterodimers and ternary complexes via two distinct domains. *Oncogene* **16**, 3261-3268.

Castilla, L. H., Perrat, P., Martinez, N. J., Landrette, S. F., Keys, R., Oikemus, S., Flanegan, J., Heilman, S., Garrett, L., Dutra, A*., et al.* (2004). Identification of genes that synergize with Cbfb-MYH11 in the pathogenesis of acute myeloid leukemia. *Proc Natl Acad Sci U S A* **101**, 4924-4929.

Castro, M. E., Ferrer, I., Cascon, A., Guijarro, M. V., Lleonart, M., Cajal, S. R., Leal, J. F., Robledo, M., and Carnero, A. (2008). PPP1CA contributes to the senescence program induced by oncogenic Ras. *Carcinogenesis* **29**, 491-499.

Cawley, S., Bekiranov, S., Ng, H. H., Kapranov, P., Sekinger, E. A., Kampa, D., Piccolboni, A., Sementchenko, V., Cheng, J., Williams, A. J*., et al.* (2004). Unbiased mapping of transcription factor binding sites along human chromosomes 21 and 22 points to widespread regulation of noncoding RNAs. *Cell* **116**, 499-509.

Cerutti, J. M., Ebina, K. N., Matsuo, S. E., Martins, L., Maciel, R. M., and Kimura, E. T. (2003). Expression of Smad4 and Smad7 in human thyroid follicular carcinoma cell lines. *J Endocrinol Invest* **26**, 516-521.

Chakravarty, G., Roy, D., Gonzales, M., Gay, J., Contreras, A., and Rosen, J. M. (2000). P190-B, a Rho-GTPase-activating protein, is differentially expressed in terminal end buds and breast cancer. *Cell Growth Differ* **11**, 343-354.

Chang, H. J., Yoo, B. C., Kim, S. W., Lee, B. L., and Kim, W. H. (2007). Significance of PML and p53 protein as molecular prognostic markers of gallbladder carcinomas. *Pathol Oncol Res* **13**, 326-335.

Chang, X. Z., Li, D. Q., Hou, Y. F., Wu, J., Lu, J. S., Di, G. H., Jin, W., Ou, Z. L., Shen, Z. Z., and Shao, Z. M. (2008). Identification of the functional role of AF1Q in the progression of breast cancer. *Breast Cancer Res Treat* **111**, 65-78.

Chen, C. Z., and Lodish, H. F. (2005). MicroRNAs as regulators of mammalian hematopoiesis. *Semin Immunol* **17**, 155-165.

Chen, J., and Sadowski, I. (2005). Identification of the mismatch repair genes PMS2 and MLH1 as p53 target genes by using serial analysis of binding elements. *Proc Natl Acad Sci U S A* **102**, 4813-4818.

Cheng, A., Ross, K. E., Kaldis, P., and Solomon, M. J. (1999). Dephosphorylation of cyclin-dependent kinases by type 2C protein phosphatases. *Genes Dev* **13**, 2946-2957.

Cheng, A. J., Cheng, N. C., Ford, J., Smith, J., Murray, J. E., Flemming, C., Lastowska, M., Jackson, M. S., Hackett, C. S., Weiss, W. A.*, et al.* (2007). Cell lines from MYCN transgenic murine tumours reflect the molecular and biological characteristics of human neuroblastoma. *Eur J Cancer* **43**, 1467-1475.

Chin, K., de Solorzano, C. O., Knowles, D., Jones, A., Chou, W., Rodriguez, E. G., Kuo, W. L., Ljung, B. M., Chew, K., Myambo, K.*, et al.* (2004). In situ analyses of genome instability in breast cancer. *Nat Genet* **36**, 984-988.

Chin, K., DeVries, S., Fridlyand, J., Spellman, P. T., Roydasgupta, R., Kuo, W. L., Lapuk, A., Neve, R. M., Qian, Z., Ryder, T.*, et al.* (2006). Genomic and transcriptional aberrations linked to breast cancer pathophysiologies. *Cancer Cell* **10**, 529-541.

Chin, L., Tam, A., Pomerantz, J., Wong, M., Holash, J., Bardeesy, N., Shen, Q., O'Hagan, R., Pantginis, J., Zhou, H.*, et al.* (1999). Essential role for oncogenic Ras in tumour maintenance. *Nature* **400**, 468-472.

Cho, Y. L., Bae, S., Koo, M. S., Kim, K. M., Chun, H. J., Kim, C. K., Ro, D. Y., Kim, J. H., Lee, C. H., Kim, Y. W., and Ahn, W. S. (2005). Array comparative genomic hybridization analysis of uterine leiomyosarcoma. *Gynecol Oncol* **99**, 545-551.

Christoforidou, A. V., Papadaki, H. A., Margioris, A. N., Eliopoulos, G. D., and Tsatsanis, C. (2004). Expression of the Tpl2/Cot oncogene in human T-cell neoplasias. *Mol Cancer* **3**, 34.

Ciafre, S. A., Galardi, S., Mangiola, A., Ferracin, M., Liu, C. G., Sabatino, G., Negrini, M., Maira, G., Croce, C. M., and Farace, M. G. (2005). Extensive modulation of a set of microRNAs in primary glioblastoma. *Biochem Biophys Res Commun* **334**, 1351-1358.

Clausse, N., Baines, D., Moore, R., Brookes, S., Dickson, C., and Peters, G. (1993). Activation of both Wnt-1 and Fgf-3 by insertion of mouse mammary tumor virus downstream in the reverse orientation: a reappraisal of the enhancer insertion model. *Virology* **194**, 157-165.

Cobleigh, M. A., Vogel, C. L., Tripathy, D., Robert, N. J., Scholl, S., Fehrenbacher, L., Wolter, J. M., Paton, V., Shak, S., Lieberman, G., and Slamon, D. J. (1999). Multinational study of the efficacy and safety of humanized anti-HER2 monoclonal antibody in women who have HER2-overexpressing metastatic breast cancer that has progressed after chemotherapy for metastatic disease. *J Clin Oncol* **17**, 2639-2648.

Coe, B. P., Lee, E. H., Chi, B., Girard, L., Minna, J. D., Gazdar, A. F., Lam, S., MacAulay, C., and Lam, W. L. (2006). Gain of a region on 7p22.3, containing MAD1L1, is the most frequent event in small-cell lung cancer cell lines. *Genes Chromosomes Cancer* **45**, 11-19.

Cokus, S. J., Feng, S., Zhang, X., Chen, Z., Merriman, B., Haudenschild, C. D., Pradhan, S., Nelson, S. F., Pellegrini, M., and Jacobsen, S. E. (2008). Shotgun bisulphite sequencing of the Arabidopsis genome reveals DNA methylation patterning. *Nature* **452**, 215-219.

Collier, L. S., Carlson, C. M., Ravimohan, S., Dupuy, A. J., and Largaespada, D. A. (2005). Cancer gene discovery in solid tumours using transposon-based somatic mutagenesis in the mouse. *Nature* **436**, 272-276.

Collier, L. S., and Largaespada, D. A. (2007). Transposons for cancer gene discovery: Sleeping Beauty and beyond. *Genome Biol* **8 Suppl 1**, S15.

Cooper, W. N., Dickinson, R. E., Dallol, A., Grigorieva, E. V., Pavlova, T. V., Hesson, L. B., Bieche, I., Broggini, M., Maher, E. R., Zabarovsky, E. R*., et al.* (2008). Epigenetic regulation of the ras effector/tumour suppressor RASSF2 in breast and lung cancer. *Oncogene* **27**, 1805-1811.

Copeland, N. G., and Jenkins, N. A. (1990). Retroviral integration in murine myeloid tumors to identify Evi-1, a novel locus encoding a zinc-finger protein. *Adv Cancer Res* **54**, 141-157.

Corcoran, L. M., Adams, J. M., Dunn, A. R., and Cory, S. (1984). Murine T lymphomas in which the cellular myc oncogene has been activated by retroviral insertion. *Cell* **37**, 113-122.

Cox, C., Bignell, G., Greenman, C., Stabenau, A., Warren, W., Stephens, P., Davies, H., Watt, S., Teague, J., Edkins, S*., et al.* (2005). A survey of homozygous deletions in human cancer genomes. *Proc Natl Acad Sci U S A* **102**, 4542-4547.

Crawford, G. E., Holt, I. E., Whittle, J., Webb, B. D., Tai, D., Davis, S., Margulies, E. H., Chen, Y., Bernat, J. A., Ginsburg, D*., et al.* (2006). Genome-wide mapping of DNase hypersensitive sites using massively parallel signature sequencing (MPSS). *Genome Res* **16**, 123-131.

CRUK (2008). CancerStats (http://info.cancerresearchuk.org/cancerstats/).

Cuypers, H. T., Selten, G., Quint, W., Zijlstra, M., Maandag, E. R., Boelens, W., van Wezenbeek, P., Melief, C., and Berns, A. (1984). Murine leukemia virus-induced T-cell lymphomagenesis: integration of proviruses in a distinct chromosomal region. *Cell* **37**, 141-150.

Daheron, L., Salmeron, S., Patri, S., Brizard, A., Guilhot, F., Chomel, J. C., and Kitzis, A. (1998). Identification of several genes differentially expressed during progression of chronic myelogenous leukemia. *Leukemia* **12**, 326-332.

Danaei, G., Vander Hoorn, S., Lopez, A. D., Murray, C. J., and Ezzati, M. (2005). Causes of cancer in the world: comparative risk assessment of nine behavioural and environmental risk factors. *Lancet* **366**, 1784-1793.

Darido, C., Buchert, M., Pannequin, J., Bastide, P., Zalzali, H., Mantamadiotis, T., Bourgaux, J. F., Garambois, V., Jay, P., Blache, P.*, et al.* (2008). Defective claudin-7 regulation by Tcf-4 and Sox-9 disrupts the polarity and increases the tumorigenicity of colorectal cancer cells. *Cancer Res* **68**, 4258-4268.

Davies, H., Bignell, G. R., Cox, C., Stephens, P., Edkins, S., Clegg, S., Teague, J., Woffendin, H., Garnett, M. J., Bottomley, W.*, et al.* (2002). Mutations of the BRAF gene in human cancer. *Nature* **417**, 949-954.

Davies, H., Hunter, C., Smith, R., Stephens, P., Greenman, C., Bignell, G., Teague, J., Butler, A., Edkins, S., Stevens, C.*, et al.* (2005). Somatic mutations of the protein kinase gene family in human lung cancer. *Cancer Res* **65**, 7591-7595.

Davoodpour, P., Landstrom, M., and Welsh, M. (2007). Reduced tumor growth in vivo and increased c-Abl activity in PC3 prostate cancer cells overexpressing the Shb adapter protein. *BMC Cancer* **7**, 161.

de Alboran, I. M., O'Hagan, R. C., Gartner, F., Malynn, B., Davidson, L., Rickert, R., Rajewsky, K., DePinho, R. A., and Alt, F. W. (2001). Analysis of C-MYC function in normal cells via conditional gene-targeted mutation. *Immunity* **14**, 45-55.

De Braekeleer, E., Douet-Guilbert, N., Le Bris, M. J., Berthou, C., Morel, F., and De Braekeleer, M. (2007). A new partner gene fused to ABL1 in a t(1;9)(q24;q34)-associated B-cell acute lymphoblastic leukemia. *Leukemia* **21**, 2220-2221.

de Ridder, J., Kool, J., Uren, A., Bot, J., Wessels, L., and Reinders, M. (2007). Co-occurrence analysis of insertional mutagenesis data reveals cooperating oncogenes. *Bioinformatics* **23**, i133-141.

de Ridder, J., Uren, A., Kool, J., Reinders, M., and Wessels, L. (2006). Detecting statistically significant common insertion sites in retroviral insertional mutagenesis screens. *PLoS Comput Biol* **2**, e166.

de Stanchina, E., Querido, E., Narita, M., Davuluri, R. V., Pandolfi, P. P., Ferbeyre, G., and Lowe, S. W. (2004). PML is a direct p53 target that modulates p53 effector functions. *Mol Cell* **13**, 523-535.

Deeds, J., Cronin, F., and Duncan, L. M. (2000). Patterns of melastatin mRNA expression in melanocytic tumors. *Hum Pathol* **31**, 1346-1356.

Degenhardt, Y. Y., Wooster, R., McCombie, R. W., Lucito, R., and Powers, S. (2008). High-content analysis of cancer genome DNA alterations. *Curr Opin Genet Dev* **18**, 68-72.

Deininger, M. W., and Druker, B. J. (2003). Specific targeted therapy of chronic myelogenous leukemia with imatinib. *Pharmacol Rev* **55**, 401-423.

Delattre, O., Zucman, J., Melot, T., Garau, X. S., Zucker, J. M., Lenoir, G. M., Ambros, P. F., Sheer, D., Turc-Carel, C., Triche, T. J., and et al. (1994). The Ewing family of tumors--a subgroup of small-round-cell tumors defined by specific chimeric transcripts. *N Engl J Med* **331**, 294-299.

Denison, S. R., Callahan, G., Becker, N. A., Phillips, L. A., and Smith, D. I. (2003). Characterization of FRA6E and its potential role in autosomal recessive juvenile parkinsonism and ovarian cancer. *Genes Chromosomes Cancer* **38**, 40-52.

Dermitzakis, E. T., and Clark, A. G. (2002). Evolution of transcription factor binding sites in Mammalian gene regulatory regions: conservation and turnover. *Mol Biol Evol* **19**, 1114-1121.

Devon, R. S., Porteous, D. J., and Brookes, A. J. (1995). Splinkerettes--improved vectorettes for greater efficiency in PCR walking. *Nucleic Acids Res* **23**, 1644-1645.

Dhanasekaran, S. M., Barrette, T. R., Ghosh, D., Shah, R., Varambally, S., Kurachi, K., Pienta, K. J., Rubin, M. A., and Chinnaiyan, A. M. (2001). Delineation of prognostic biomarkers in prostate cancer. *Nature* **412**, 822-826.

Dhawan, P., Singh, A. B., Deane, N. G., No, Y., Shiou, S. R., Schmidt, C., Neff, J., Washington, M. K., and Beauchamp, R. D. (2005). Claudin-1 regulates cellular transformation and metastatic behavior in colon cancer. *J Clin Invest* **115**, 1765-1776.

Diaz-Uriarte, R., and Rueda, O. M. (2007). ADaCGH: A parallelized web-based application and R package for the analysis of aCGH data. *PLoS ONE* **2**, e737.

Ding, S., Wu, X., Li, G., Han, M., Zhuang, Y., and Xu, T. (2005). Efficient transposition of the piggyBac (PB) transposon in mammalian cells and mice. *Cell* **122**, 473-483.

Diskin, S. J., Eck, T., Greshock, J., Mosse, Y. P., Naylor, T., Stoeckert, C. J., Jr., Weber, B. L., Maris, J. M., and Grant, G. R. (2006). STAC: A method for testing the significance of DNA copy number aberrations across multiple array-CGH experiments. *Genome Res* **16**, 1149-1158.

Donati, V., Fontanini, G., Dell'Omodarme, M., Prati, M. C., Nuti, S., Lucchi, M., Mussi, A., Fabbri, M., Basolo, F., Croce, C. M., and Aqeilan, R. I. (2007). WWOX expression in different histologic types and subtypes of non-small cell lung cancer. *Clin Cancer Res* **13**, 884-891.

Dong, M., How, T., Kirkbride, K. C., Gordon, K. J., Lee, J. D., Hempel, N., Kelly, P., Moeller, B. J., Marks, J. R., and Blobe, G. C. (2007). The type III TGF-beta receptor suppresses breast cancer progression. *J Clin Invest* **117**, 206-217.

Dowdy, S. C., Mariani, A., Reinholz, M. M., Keeney, G. L., Spelsberg, T. C., Podratz, K. C., and Janknecht, R. (2005). Overexpression of the TGF-beta antagonist Smad7 in endometrial cancer. *Gynecol Oncol* **96**, 368-373.

Down, T. A., Rakyan, V. K., Turner, D. J., Flicek, P., Li, H., Kulesha, E., Graf, S., Johnson, N., Herrero, J., Tomazou, E. M., *et al.* (2008). A Bayesian deconvolution strategy for immunoprecipitation-based DNA methylome analysis. *Nat Biotechnol* **26**, 779-785.

Drabek, D., Zagoraiou, L., deWit, T., Langeveld, A., Roumpaki, C., Mamalaki, C., Savakis, C., and Grosveld, F. (2003). Transposition of the Drosophila hydei Minos transposon in the mouse germ line. *Genomics* **81**, 108-111.

Druker, B. J., Guilhot, F., O'Brien, S. G., Gathmann, I., Kantarjian, H., Gattermann, N., Deininger, M. W., Silver, R. T., Goldman, J. M., Stone, R. M., *et al.* (2006). Five-year follow-up of patients receiving imatinib for chronic myeloid leukemia. *N Engl J Med* **355**, 2408-2417.

Druker, B. J., Tamura, S., Buchdunger, E., Ohno, S., Segal, G. M., Fanning, S., Zimmermann, J., and Lydon, N. B. (1996). Effects of a selective inhibitor of the Abl tyrosine kinase on the growth of Bcr-Abl positive cells. *Nat Med* **2**, 561-566.

Duhamel, M., Arrouss, I., Merle-Beral, H., and Rebollo, A. (2008). The Aiolos transcription factor is up-regulated in chronic lymphocytic leukemia. *Blood* **111**, 3225-3228.

Dul, J. L., Argon, Y., Winkler, T., ten Boekel, E., Melchers, F., and Martensson, I. L. (1996). The murine VpreB1 and VpreB2 genes both encode a protein of the surrogate light chain and are co-expressed during B cell development. *Eur J Immunol* **26**, 906-913.

Dupuy, A. J., Akagi, K., Largaespada, D. A., Copeland, N. G., and Jenkins, N. A. (2005). Mammalian mutagenesis using a highly mobile somatic Sleeping Beauty transposon system. *Nature* **436**, 221-226.

Dupuy, A. J., Fritz, S., and Largaespada, D. A. (2001). Transposition and gene disruption in the male germline of the mouse. *Genesis* **30**, 82-88.

Dupuy, A. J., Jenkins, N. A., and Copeland, N. G. (2006). Sleeping beauty: a novel cancer gene discovery tool. *Hum Mol Genet* **15 Spec No 1**, R75-79.

Duterque-Coquillaud, M., Niel, C., Plaza, S., and Stehelin, D. (1993). New human erg isoforms generated by alternative splicing are transcriptional activators. *Oncogene* **8**, 1865-1873.

Eden, P., Ritz, C., Rose, C., Ferno, M., and Peterson, C. (2004). "Good Old" clinical markers have similar power in breast cancer prognosis as microarray gene expression profilers. *Eur J Cancer* **40**, 1837-1841.

Eilers, M., Picard, D., Yamamoto, K. R., and Bishop, J. M. (1989). Chimaeras of myc oncoprotein and steroid receptors cause hormone-dependent transformation of cells. *Nature* **340**, 66-68.

Eischen, C. M., Weber, J. D., Roussel, M. F., Sherr, C. J., and Cleveland, J. L. (1999). Disruption of the ARF-Mdm2-p53 tumor suppressor pathway in Myc-induced lymphomagenesis. *Genes Dev* **13**, 2658-2669.

El Ghouzzi, V., Dagoneau, N., Kinning, E., Thauvin-Robinet, C., Chemaitilly, W., Prost-Squarcioni, C., Al-Gazali, L. I., Verloes, A., Le Merrer, M., Munnich, A., *et al.* (2003). Mutations in a novel gene Dymeclin (FLJ20071) are responsible for Dyggve-Melchior-Clausen syndrome. *Hum Mol Genet* **12**, 357-364.

Eliseev, R. A., Schwarz, E. M., Zuscik, M. J., O'Keefe, R. J., Drissi, H., and Rosier, R. N. (2006). Smad7 mediates inhibition of Saos2 osteosarcoma cell differentiation by NFkappaB. *Exp Cell Res* **312**, 40-50.

Ellis, N. A., Groden, J., Ye, T. Z., Straughen, J., Lennon, D. J., Ciocci, S., Proytcheva, M., and German, J. (1995). The Bloom's syndrome gene product is homologous to RecQ helicases. *Cell* **83**, 655-666.

Ellisen, L. W., Bird, J., West, D. C., Soreng, A. L., Reynolds, T. C., Smith, S. D., and Sklar, J. (1991). TAN-1, the human homolog of the Drosophila notch gene, is broken by chromosomal translocations in T lymphoblastic neoplasms. *Cell* **66**, 649-661.

Ellwood-Yen, K., Graeber, T. G., Wongvipat, J., Iruela-Arispe, M. L., Zhang, J., Matusik, R., Thomas, G. V., and Sawyers, C. L. (2003). Myc-driven murine prostate cancer shares molecular features with human prostate tumors. *Cancer Cell* **4**, 223-238.

Endoh, M., Tamura, G., Honda, T., Homma, N., Terashima, M., Nishizuka, S., and Motoyama, T. (2005). RASSF2, a potential tumour suppressor, is silenced by CpG island hypermethylation in gastric cancer. *Br J Cancer* **93**, 1395-1399.

Engelman, J. A., Zejnullahu, K., Mitsudomi, T., Song, Y., Hyland, C., Park, J. O., Lindeman, N., Gale, C. M., Zhao, X., Christensen, J.*, et al.* (2007). MET amplification leads to gefitinib resistance in lung cancer by activating ERBB3 signaling. *Science* **316**, 1039-1043.

Engler, D. A., Mohapatra, G., Louis, D. N., and Betensky, R. A. (2006). A pseudolikelihood approach for simultaneous analysis of array comparative genomic hybridizations. *Biostatistics* **7**, 399-421.

Erkeland, S. J., Verhaak, R. G., Valk, P. J., Delwel, R., Lowenberg, B., and Touw, I. P. (2006). Significance of murine retroviral mutagenesis for identification of disease genes in human acute myeloid leukemia. *Cancer Res* **66**, 622-626.

Esteller, M., Silva, J. M., Dominguez, G., Bonilla, F., Matias-Guiu, X., Lerma, E., Bussaglia, E., Prat, J., Harkes, I. C., Repasky, E. A.*, et al.* (2000). Promoter hypermethylation and BRCA1 inactivation in sporadic breast and ovarian tumors. *J Natl Cancer Inst* **92**, 564-569.

Euskirchen, G. M., Rozowsky, J. S., Wei, C. L., Lee, W. H., Zhang, Z. D., Hartman, S., Emanuelsson, O., Stolc, V., Weissman, S., Gerstein, M. B.*, et al.* (2007). Mapping of transcription factor binding regions in mammalian cells by ChIP: comparison of array- and sequencing-based technologies. *Genome Res* **17**, 898-909.

Evtimova, V., Zeillinger, R., and Weidle, U. H. (2003). Identification of genes associated with the invasive status of human mammary carcinoma cell lines by transcriptional profiling. *Tumour Biol* **24**, 189-198.

Faderl, S., Talpaz, M., Estrov, Z., O'Brien, S., Kurzrock, R., and Kantarjian, H. M. (1999). The biology of chronic myeloid leukemia. *N Engl J Med* **341**, 164-172.

Fearon, E. R., and Vogelstein, B. (1990). A genetic model for colorectal tumorigenesis. *Cell* **61**, 759-767.

Fears, S., Gavin, M., Zhang, D. E., Hetherington, C., Ben-David, Y., Rowley, J. D., and Nucifora, G. (1997). Functional characterization of ETV6 and ETV6/CBFA2 in the regulation of the MCSFR proximal promoter. *Proc Natl Acad Sci U S A* **94**, 1949-1954.

Feldman, B. J., Hampton, T., and Cleary, M. L. (2000). A carboxy-terminal deletion mutant of Notch1 accelerates lymphoid oncogenesis in E2A-PBX1 transgenic mice. *Blood* **96**, 1906-1913.

Felsher, D. W., and Bishop, J. M. (1999). Reversible tumorigenesis by MYC in hematopoietic lineages. *Mol Cell* **4**, 199-207.

Feng, Q., Moran, J. V., Kazazian, H. H., Jr., and Boeke, J. D. (1996). Human L1 retrotransposon encodes a conserved endonuclease required for retrotransposition. *Cell* **87**, 905-916.

Ferlay, J., Bray, F., Pisani, P., and Parkin, D. M. (2004). *GLOBOCAN 2002: Cancer Incidence. Mortality and Prevalence Worldwide.* (Lyon: IARC Press 2004).

Feuk, L., Carson, A. R., and Scherer, S. W. (2006). Structural variation in the human genome. *Nat Rev Genet* **7**, 85-97.

Fiegler, H., Geigl, J. B., Langer, S., Rigler, D., Porter, K., Unger, K., Carter, N. P., and Speicher, M. R. (2007). High resolution array-CGH analysis of single cells. *Nucleic Acids Res* **35**, e15.

Finger, E. C., Turley, R. S., Dong, M., How, T., Fields, T. A., and Blobe, G. C. (2008). TbetaRIII suppresses non-small cell lung cancer invasiveness and tumorigenicity. *Carcinogenesis* **29**, 528-535.

Finnegan, M. C., Hammond, D. W., Hancock, B. W., and Goyns, M. H. (1995). Activation of MYCN in a case of non-Hodgkin's lymphoma. *Leuk Lymphoma* **18**, 511-514.

Fisher, S., Grice, E. A., Vinton, R. M., Bessling, S. L., and McCallion, A. S. (2006). Conservation of RET regulatory function from human to zebrafish without sequence similarity. *Science* **312**, 276-279.

Forbes, S., Clements, J., Dawson, E., Bamford, S., Webb, T., Dogan, A., Flanagan, A., Teague, J., Wooster, R., Futreal, P. A., and Stratton, M. R. (2006). Cosmic 2005. *Br J Cancer* **94**, 318-322.

Forster, A., Pannell, R., Drynan, L. F., McCormack, M., Collins, E. C., Daser, A., and Rabbitts, T. H. (2003). Engineering de novo reciprocal chromosomal translocations associated with Mll to replicate primary events of human cancer. *Cancer Cell* **3**, 449-458.

Fox, S. B., Brown, P., Han, C., Ashe, S., Leek, R. D., Harris, A. L., and Banham, A. H. (2004). Expression of the forkhead transcription factor FOXP1 is associated with estrogen receptor alpha and improved survival in primary human breast carcinomas. *Clin Cancer Res* **10**, 3521-3527.

Frank, B., Klaes, R., and Burwinkel, B. (2005). Familial cancer and ARLTS1. *N Engl J Med* **353**, 313-314; author reply 313-314.

Freije, J. M., Diez-Itza, I., Balbin, M., Sanchez, L. M., Blasco, R., Tolivia, J., and Lopez-Otin, C. (1994). Molecular cloning and expression of collagenase-3, a novel human matrix metalloproteinase produced by breast carcinomas. *J Biol Chem* **269**, 16766-16773.

Frese, K. K., and Tuveson, D. A. (2007). Maximizing mouse cancer models. *Nat Rev Cancer* **7**, 645-658.

Fridlyand, J., Snijders, A. M., Pinkel, D., and Albertson, D. (2004). Hidden markov models approach to the analysis of array cgh data. *Journal of Multivariate Analysis* **90**, 132-153.

Fridlyand, J., Snijders, A. M., Ylstra, B., Li, H., Olshen, A., Segraves, R., Dairkee, S., Tokuyasu, T., Ljung, B. M., Jain, A. N.*, et al.* (2006). Breast tumor copy number aberration phenotypes and genomic instability. *BMC Cancer* **6**, 96.

Frieboes, H. B., and Brody, J. P. (2005). Age-incidence data support the telomere crisis hypothesis of epithelial carcinogenesis. *Unpublished*.

Frigola, J., Song, J., Stirzaker, C., Hinshelwood, R. A., Peinado, M. A., and Clark, S. J. (2006). Epigenetic remodeling in colorectal cancer results in coordinate gene suppression across an entire chromosome band. *Nat Genet* **38**, 540-549.

Frohling, S., Scholl, C., Levine, R. L., Loriaux, M., Boggon, T. J., Bernard, O. A., Berger, R., Dohner, H., Dohner, K., Ebert, B. L.*, et al.* (2007). Identification of driver and passenger mutations of FLT3 by high-throughput DNA sequence analysis and functional assessment of candidate alleles. *Cancer Cell* **12**, 501-513.

Fukuoka, M., Yano, S., Giaccone, G., Tamura, T., Nakagawa, K., Douillard, J. Y., Nishiwaki, Y., Vansteenkiste, J., Kudoh, S., Rischin, D.*, et al.* (2003). Multi-institutional randomized phase II trial of gefitinib for previously treated patients with advanced non-small-cell lung cancer (The IDEAL 1 Trial) [corrected]. *J Clin Oncol* **21**, 2237-2246.

Fulci, G., Labuhn, M., Maier, D., Lachat, Y., Hausmann, O., Hegi, M. E., Janzer, R. C., Merlo, A., and Van Meir, E. G. (2000). p53 gene mutation and ink4a-arf deletion appear to be two mutually exclusive events in human glioblastoma. *Oncogene* **19**, 3816-3822.

Funaki, T., Nakao, A., Ebihara, N., Setoguchi, Y., Fukuchi, Y., Okumura, K., Ra, C., Ogawa, H., and Kanai, A. (2003). Smad7 suppresses the inhibitory effect of TGF-beta2 on corneal endothelial cell proliferation and accelerates corneal endothelial wound closure in vitro. *Cornea* **22**, 153-159.

Furusawa, T., Ikawa, S., Yanai, N., and Obinata, M. (2000). Isolation of a novel PDZ-containing myosin from hematopoietic supportive bone marrow stromal cell lines. *Biochem Biophys Res Commun* **270**, 67-75.

Futreal, P. A. (2007). Backseat drivers take the wheel. *Cancer Cell* **12**, 493-494.

Futreal, P. A., Coin, L., Marshall, M., Down, T., Hubbard, T., Wooster, R., Rahman, N., and Stratton, M. R. (2004). A census of human cancer genes. *Nat Rev Cancer* **4**, 177-183.

Gakiopoulou-Givalou, H., Nakopoulou, L., Panayotopoulou, E. G., Zervas, A., Mavrommatis, J., and Giannopoulos, A. (2003). Non-endothelial KDR/flk-1 expression is associated with increased survival of patients with urothelial bladder carcinomas. *Histopathology* **43**, 272-279.

Gambichler, T., Skrygan, M., Kaczmarczyk, J. M., Hyun, J., Tomi, N. S., Sommer, A., Bechara, F. G., Boms, S., Brockmeyer, N. H., Altmeyer, P., and Kreuter, A. (2007).

Increased expression of TGF-beta/Smad proteins in basal cell carcinoma. *Eur J Med Res* **12**, 509-514.

Garber, M. E., Troyanskaya, O. G., Schluens, K., Petersen, S., Thaesler, Z., Pacyna-Gengelbach, M., van de Rijn, M., Rosen, G. D., Perou, C. M., Whyte, R. I*., et al.* (2001). Diversity of gene expression in adenocarcinoma of the lung. *Proc Natl Acad Sci U S A* **98**, 13784-13789.

Garnis, C., Coe, B. P., Zhang, L., Rosin, M. P., and Lam, W. L. (2004). Overexpression of LRP12, a gene contained within an 8q22 amplicon identified by high-resolution array CGH analysis of oral squamous cell carcinomas. *Oncogene* **23**, 2582-2586.

Garraway, L. A., and Sellers, W. R. (2006). From integrated genomics to tumor lineage dependency. *Cancer Res* **66**, 2506-2508.

Garraway, L. A., Widlund, H. R., Rubin, M. A., Getz, G., Berger, A. J., Ramaswamy, S., Beroukhim, R., Milner, D. A., Granter, S. R., Du, J*., et al.* (2005). Integrative genomic analyses identify MITF as a lineage survival oncogene amplified in malignant melanoma. *Nature* **436**, 117-122.

Garzon, R., and Croce, C. M. (2008). MicroRNAs in normal and malignant hematopoiesis. *Curr Opin Hematol* **15**, 352-358.

Gaspar, C., Cardoso, J., Franken, P., Molenaar, L., Morreau, H., Moslein, G., Sampson, J., Boer, J. M., de Menezes, R. X., and Fodde, R. (2008). Cross-species comparison of human and mouse intestinal polyps reveals conserved mechanisms in adenomatous polyposis coli (APC)-driven tumorigenesis. *Am J Pathol* **172**, 1363-1380.

Gearhart, J., Pashos, E. E., and Prasad, M. K. (2007). Pluripotency redux--advances in stem-cell research. *N Engl J Med* **357**, 1469-1472.

Georgopoulos, K. (2002). Haematopoietic cell-fate decisions, chromatin regulation and ikaros. *Nat Rev Immunol* **2**, 162-174.

Georgopoulos, K., Winandy, S., and Avitahl, N. (1997). The role of the Ikaros gene in lymphocyte development and homeostasis. *Annu Rev Immunol* **15**, 155-176.

Germeyer, A., Klinkert, M. S., Huppertz, A. G., Clausmeyer, S., Popovici, R. M., Strowitzki, T., and von Wolff, M. (2007). Expression of syndecans, cell-cell interaction regulating heparan sulfate proteoglycans, within the human endometrium and their regulation throughout the menstrual cycle. *Fertil Steril* **87**, 657-663.

Getz, G., Hofling, H., Mesirov, J. P., Golub, T. R., Meyerson, M., Tibshirani, R., and Lander, E. S. (2007). Comment on "The consensus coding sequences of human breast and colorectal cancers". *Science* **317**, 1500.

Geurts, A. M., Collier, L. S., Geurts, J. L., Oseth, L. L., Bell, M. L., Mu, D., Lucito, R., Godbout, S. A., Green, L. E., Lowe, S. W*., et al.* (2006). Gene mutations and genomic rearrangements in the mouse as a result of transposon mobilization from chromosomal concatemers. *PLoS Genet* **2**, e156.

Geurts, A. M., Yang, Y., Clark, K. J., Liu, G., Cui, Z., Dupuy, A. J., Bell, J. B., Largaespada, D. A., and Hackett, P. B. (2003). Gene transfer into genomes of human cells by the sleeping beauty transposon system. *Mol Ther* **8**, 108-117.

Gisselsson, D., Palsson, E., Hoglund, M., Domanski, H., Mertens, F., Pandis, N., Sciot, R., Dal Cin, P., Bridge, J. A., and Mandahl, N. (2002). Differentially amplified chromosome 12 sequences in low- and high-grade osteosarcoma. *Genes Chromosomes Cancer* **33**, 133-140.

Glavan, F., Behm-Ansmant, I., Izaurralde, E., and Conti, E. (2006). Structures of the PIN domains of SMG6 and SMG5 reveal a nuclease within the mRNA surveillance complex. *Embo J* **25**, 5117-5125.

Goemans, B. F., Zwaan, C. M., Miller, M., Zimmermann, M., Harlow, A., Meshinchi, S., Loonen, A. H., Hahlen, K., Reinhardt, D., Creutzig, U*., et al.* (2005). Mutations in KIT and RAS are frequent events in pediatric core-binding factor acute myeloid leukemia. *Leukemia* **19**, 1536-1542.

Gordon, K. J., Dong, M., Chislock, E. M., Fields, T. A., and Blobe, G. C. (2008). Loss of type III transforming growth factor beta receptor expression increases motility and invasiveness associated with epithelial to mesenchymal transition during pancreatic cancer progression. *Carcinogenesis* **29**, 252-262.

Gorre, M. E., Mohammed, M., Ellwood, K., Hsu, N., Paquette, R., Rao, P. N., and Sawyers, C. L. (2001). Clinical resistance to STI-571 cancer therapy caused by BCR-ABL gene mutation or amplification. *Science* **293**, 876-880.

Gotoh, O. (1982). An improved algorithm for matching biological sequences. *J Mol Biol* **162**, 705-708.

Gottardo, F., Liu, C. G., Ferracin, M., Calin, G. A., Fassan, M., Bassi, P., Sevignani, C., Byrne, D., Negrini, M., Pagano, F*., et al.* (2007). Micro-RNA profiling in kidney and bladder cancers. *Urol Oncol* **25**, 387-392.

Grabarczyk, P., Przybylski, G. K., Depke, M., Volker, U., Bahr, J., Assmus, K., Broker, B. M., Walther, R., and Schmidt, C. A. (2007). Inhibition of BCL11B expression leads to apoptosis of malignant but not normal mature T cells. *Oncogene* **26**, 3797-3810.

Grandori, C., Cowley, S. M., James, L. P., and Eisenman, R. N. (2000). The Myc/Max/Mad network and the transcriptional control of cell behavior. *Annu Rev Cell Dev Biol* **16**, 653-699.

Green, P. (unpublished). *http://wwwphraporg/*.

Greenman, C., Stephens, P., Smith, R., Dalgliesh, G. L., Hunter, C., Bignell, G., Davies, H., Teague, J., Butler, A., Stevens, C*., et al.* (2007). Patterns of somatic mutation in human cancer genomes. *Nature* **446**, 153-158.

Greiner, J., Ringhoffer, M., Simikopinko, O., Szmaragowska, A., Huebsch, S., Maurer, U., Bergmann, L., and Schmitt, M. (2000). Simultaneous expression of different immunogenic antigens in acute myeloid leukemia. *Exp Hematol* **28**, 1413-1422.

Grigo, K., Wirsing, A., Lucas, B., Klein-Hitpass, L., and Ryffel, G. U. (2008). HNF4 alpha orchestrates a set of 14 genes to down-regulate cell proliferation in kidney cells. *Biol Chem* **389**, 179-187.

Grone, J., Weber, B., Staub, E., Heinze, M., Klaman, I., Pilarsky, C., Hermann, K., Castanos-Velez, E., Ropcke, S., Mann, B*., et al.* (2007). Differential expression of genes encoding tight junction proteins in colorectal cancer: frequent dysregulation of claudin-1, -8 and -12. *Int J Colorectal Dis* **22**, 651-659.

Gunther, T., Schneider-Stock, R., Hackel, C., Kasper, H. U., Pross, M., Hackelsberger, A., Lippert, H., and Roessner, A. (2000). Mdm2 gene amplification in gastric cancer correlation with expression of Mdm2 protein and p53 alterations. *Mod Pathol* **13**, 621-626.

Gurrieri, C., Capodieci, P., Bernardi, R., Scaglioni, P. P., Nafa, K., Rush, L. J., Verbel, D. A., Cordon-Cardo, C., and Pandolfi, P. P. (2004). Loss of the tumor suppressor PML in human cancers of multiple histologic origins. *J Natl Cancer Inst* **96**, 269-279.

Hackett, C. S., Hodgson, J. G., Law, M. E., Fridlyand, J., Osoegawa, K., de Jong, P. J., Nowak, N. J., Pinkel, D., Albertson, D. G., Jain, A*., et al.* (2003). Genome-wide array CGH analysis of murine neuroblastoma reveals distinct genomic aberrations which parallel those in human tumors. *Cancer Res* **63**, 5266-5273.

Haines, N., and Irvine, K. D. (2003). Glycosylation regulates Notch signalling. *Nat Rev Mol Cell Biol* **4**, 786-797.

Hamaguchi, A., Suzuki, E., Murayama, K., Fujimura, T., Hikita, T., Iwabuchi, K., Handa, K., Withers, D. A., Masters, S. C., Fu, H., and Hakomori, S. (2003). Sphingosine-dependent protein kinase-1, directed to 14-3-3, is identified as the kinase domain of protein kinase C delta. *J Biol Chem* **278**, 41557-41565.

Han, J. S., and Boeke, J. D. (2004). A highly active synthetic mammalian retrotransposon. *Nature* **429**, 314-318.

Hanahan, D., and Weinberg, R. A. (2000). The hallmarks of cancer. *Cell* **100**, 57-70.

Hansen, G. M., Skapura, D., and Justice, M. J. (2000). Genetic profile of insertion mutations in mouse leukemias and lymphomas. *Genome Res* **10**, 237-243.

Hart, S. M., and Foroni, L. (2002). Core binding factor genes and human leukemia. *Haematologica* **87**, 1307-1323.

Hayami, Y., Iida, S., Nakazawa, N., Hanamura, I., Kato, M., Komatsu, H., Miura, I., Dave, B. J., Sanger, W. G., Lim, B*., et al.* (2003). Inactivation of the E3/LAPTm5 gene by chromosomal rearrangement and DNA methylation in human multiple myeloma. *Leukemia* **17**, 1650-1657.

Hayashi, E., Kuramitsu, Y., Okada, F., Fujimoto, M., Zhang, X., Kobayashi, M., Iizuka, N., Ueyama, Y., and Nakamura, K. (2005). Proteomic profiling for cancer progression: Differential display analysis for the expression of intracellular proteins between regressive and progressive cancer cell lines. *Proteomics* **5**, 1024-1032.

Hayashita, Y., Osada, H., Tatematsu, Y., Yamada, H., Yanagisawa, K., Tomida, S., Yatabe, Y., Kawahara, K., Sekido, Y., and Takahashi, T. (2005). A polycistronic microRNA cluster, miR-17-92, is overexpressed in human lung cancers and enhances cell proliferation. *Cancer Res* **65**, 9628-9632.

Heighway, J., Betticher, D. C., Hoban, P. R., Altermatt, H. J., and Cowen, R. (1996). Coamplification in tumors of KRAS2, type 2 inositol 1,4,5 triphosphate receptor gene, and a novel human gene, KRAG. *Genomics* **35**, 207-214.

Heminger, K., Jain, V., Kadakia, M., Dwarakanath, B., and Berberich, S. J. (2006). Altered gene expression induced by ionizing radiation and glycolytic inhibitor 2-deoxy-glucose in a human glioma cell line: implications for radio sensitization. *Cancer Biol Ther* **5**, 815-823.

Hergovich, A., Schmitz, D., and Hemmings, B. A. (2006). The human tumour suppressor LATS1 is activated by human MOB1 at the membrane. *Biochem Biophys Res Commun* **345**, 50-58.

Herman, J. G., Latif, F., Weng, Y., Lerman, M. I., Zbar, B., Liu, S., Samid, D., Duan, D. S., Gnarra, J. R., Linehan, W. M., and et al. (1994). Silencing of the VHL tumor-suppressor gene by DNA methylation in renal carcinoma. *Proc Natl Acad Sci U S A* **91**, 9700-9704.

Herman, J. G., Merlo, A., Mao, L., Lapidus, R. G., Issa, J. P., Davidson, N. E., Sidransky, D., and Baylin, S. B. (1995). Inactivation of the CDKN2/p16/MTS1 gene is frequently associated with aberrant DNA methylation in all common human cancers. *Cancer Res* **55**, 4525-4530.

Hernando, E., Nahle, Z., Juan, G., Diaz-Rodriguez, E., Alaminos, M., Hemann, M., Michel, L., Mittal, V., Gerald, W., Benezra, R*., et al.* (2004). Rb inactivation promotes genomic instability by uncoupling cell cycle progression from mitotic control. *Nature* **430**, 797-802.

Hicks, J., Krasnitz, A., Lakshmi, B., Navin, N. E., Riggs, M., Leibu, E., Esposito, D., Alexander, J., Troge, J., Grubor, V*., et al.* (2006). Novel patterns of genome rearrangement and their association with survival in breast cancer. *Genome Res* **16**, 1465-1479.

Hjelmsoe, I., Allen, C. E., Cohn, M. A., Tulchinsky, E. M., and Wu, L. C. (2000). The kappaB and V(D)J recombination signal sequence binding protein KRC regulates transcription of the mouse metastasis-associated gene S100A4/mts1. *J Biol Chem* **275**, 913-920.

Ho, J., Cocolakis, E., Dumas, V. M., Posner, B. I., Laporte, S. A., and Lebrun, J. J. (2005). The G protein-coupled receptor kinase-2 is a TGFbeta-inducible antagonist of TGFbeta signal transduction. *Embo J* **24**, 3247-3258.

Hodgson, G., Hager, J. H., Volik, S., Hariono, S., Wernick, M., Moore, D., Nowak, N., Albertson, D. G., Pinkel, D., Collins, C*., et al.* (2001). Genome scanning with array CGH delineates regional alterations in mouse islet carcinomas. *Nat Genet* **29**, 459-464.

Hoemann, C. D., Beaulieu, N., Girard, L., Rebai, N., and Jolicoeur, P. (2000). Two distinct Notch1 mutant alleles are involved in the induction of T-cell leukemia in c-myc transgenic mice. *Mol Cell Biol* **20**, 3831-3842.

Hohjoh, H., and Singer, M. F. (1997). Sequence-specific single-strand RNA binding protein encoded by the human LINE-1 retrotransposon. *Embo J* **16**, 6034-6043.

Horak, C. E., Mahajan, M. C., Luscombe, N. M., Gerstein, M., Weissman, S. M., and Snyder, M. (2002). GATA-1 binding sites mapped in the beta-globin locus by using mammalian chIp-chip analysis. *Proc Natl Acad Sci U S A* **99**, 2924-2929.

Howarth, K. D., Blood, K. A., Ng, B. L., Beavis, J. C., Chua, Y., Cooke, S. L., Raby, S., Ichimura, K., Collins, V. P., Carter, N. P., and Edwards, P. A. (2008). Array painting reveals a high frequency of balanced translocations in breast cancer cell lines that break in cancer-relevant genes. *Oncogene* **27**, 3345-3359.

Hsu, L., Self, S. G., Grove, D., Randolph, T., Wang, K., Delrow, J. J., Loo, L., and Porter, P. (2005). Denoising array-based comparative genomic hybridization data using wavelets. *Biostatistics* **6**, 211-226.

Hsu, L. C., Huang, X., Seasholtz, S., Potter, D. M., and Gollin, S. M. (2006). Gene amplification and overexpression of protein phosphatase 1alpha in oral squamous cell carcinoma cell lines. *Oncogene* **25**, 5517-5526.

Hu, Y. C., Lam, K. Y., Law, S., Wong, J., and Srivastava, G. (2001). Profiling of differentially expressed cancer-related genes in esophageal squamous cell carcinoma (ESCC) using human cancer cDNA arrays: overexpression of oncogene MET correlates with tumor differentiation in ESCC. *Clin Cancer Res* **7**, 3519-3525.

Huang, J., Gusnanto, A., O'Sullivan, K., Staaf, J., Borg, A., and Pawitan, Y. (2007). Robust smooth segmentation approach for array CGH data analysis. *Bioinformatics* **23**, 2463-2469.

Huang, J., Wei, W., Chen, J., Zhang, J., Liu, G., Di, X., Mei, R., Ishikawa, S., Aburatani, H., Jones, K. W., and Shapero, M. H. (2006a). CARAT: a novel method for allelic detection of DNA copy number changes using high density oligonucleotide arrays. *BMC Bioinformatics* **7**, 83.

Huang, J., Wei, W., Zhang, J., Liu, G., Bignell, G. R., Stratton, M. R., Futreal, P. A., Wooster, R., Jones, K. W., and Shapero, M. H. (2004). Whole genome DNA copy number changes identified by high density oligonucleotide arrays. *Hum Genomics* **1**, 287-299.

Huang, X., Godfrey, T. E., Gooding, W. E., McCarty, K. S., Jr., and Gollin, S. M. (2006b). Comprehensive genome and transcriptome analysis of the 11q13 amplicon in human oral cancer and synteny to the 7F5 amplicon in murine oral carcinoma. *Genes Chromosomes Cancer* **45**, 1058-1069.

Hubbard, T. J., Aken, B. L., Beal, K., Ballester, B., Caccamo, M., Chen, Y., Clarke, L., Coates, G., Cunningham, F., Cutts, T*., et al.* (2007). Ensembl 2007. *Nucleic Acids Res* **35**, D610-617.

Huettner, C. S., Zhang, P., Van Etten, R. A., and Tenen, D. G. (2000). Reversibility of acute B-cell leukaemia induced by BCR-ABL1. *Nat Genet* **24**, 57-60.

Hug, B. A., Ahmed, N., Robbins, J. A., and Lazar, M. A. (2004). A chromatin immunoprecipitation screen reveals protein kinase Cbeta as a direct RUNX1 target gene. *J Biol Chem* **279**, 825-830.

Hupe, P., Stransky, N., Thiery, J. P., Radvanyi, F., and Barillot, E. (2004). Analysis of array CGH data: from signal ratio to gain and loss of DNA regions. *Bioinformatics* **20**, 3413-3422.

Hwang, H. C., Martins, C. P., Bronkhorst, Y., Randel, E., Berns, A., Fero, M., and Clurman, B. E. (2002). Identification of oncogenes collaborating with p27Kip1 loss by insertional mutagenesis and high-throughput insertion site analysis. *Proc Natl Acad Sci U S A* **99**, 11293-11298.

Hyman, E., Kauraniemi, P., Hautaniemi, S., Wolf, M., Mousses, S., Rozenblum, E., Ringner, M., Sauter, G., Monni, O., Elkahloun, A*., et al.* (2002). Impact of DNA amplification on gene expression patterns in breast cancer. *Cancer Res* **62**, 6240-6245.

IARC (1994). Schistosomes, Liver Flukes and helicobacter pylori, In IARC Monograph on the Evaluation of Carcinogenic Risks to Humans (Lyon: International Agency for Research on Cancer).

Ikeda, R., Kokubu, C., Yusa, K., Keng, V. W., Horie, K., and Takeda, J. (2007). Sleeping beauty transposase has an affinity for heterochromatin conformation. *Mol Cell Biol* **27**, 1665-1676.

Impey, S., McCorkle, S. R., Cha-Molstad, H., Dwyer, J. M., Yochum, G. S., Boss, J. M., McWeeney, S., Dunn, J. J., Mandel, G., and Goodman, R. H. (2004). Defining the CREB regulon: a genome-wide analysis of transcription factor regulatory regions. *Cell* **119**, 1041-1054.

Ip, Y. C., Cheung, S. T., Lee, Y. T., Ho, J. C., and Fan, S. T. (2007). Inhibition of hepatocellular carcinoma invasion by suppression of claudin-10 in HLE cells. *Mol Cancer Ther* **6**, 2858-2867.

Irie, A., Yamauchi, A., Kontani, K., Kihara, M., Liu, D., Shirato, Y., Seki, M., Nishi, N., Nakamura, T., Yokomise, H., and Hirashima, M. (2005). Galectin-9 as a prognostic factor with antimetastatic potential in breast cancer. *Clin Cancer Res* **11**, 2962-2968.

Ishigami, S., Natsugoe, S., Nakajo, A., Tokuda, K., Uenosono, Y., Arigami, T., Matsumoto, M., Okumura, H., Hokita, S., and Aikou, T. (2007). Prognostic value of CCR7 expression in gastric cancer. *Hepatogastroenterology* **54**, 1025-1028.

Ishisaki, A., Yamato, K., Nakao, A., Nonaka, K., Ohguchi, M., ten Dijke, P., and Nishihara, T. (1998). Smad7 is an activin-inducible inhibitor of activin-induced growth arrest and apoptosis in mouse B cells. *J Biol Chem* **273**, 24293-24296.

Ivics, Z., Hackett, P. B., Plasterk, R. H., and Izsvak, Z. (1997). Molecular reconstruction of Sleeping Beauty, a Tc1-like transposon from fish, and its transposition in human cells. *Cell* **91**, 501-510.

Iwanaga, Y., Chi, Y. H., Miyazato, A., Sheleg, S., Haller, K., Peloponese, J. M., Jr., Li, Y., Ward, J. M., Benezra, R., and Jeang, K. T. (2007). Heterozygous deletion of mitotic arrest-deficient protein 1 (MAD1) increases the incidence of tumors in mice. *Cancer Res* **67**, 160-166.

Izsvak, Z., Ivics, Z., and Plasterk, R. H. (2000). Sleeping Beauty, a wide host-range transposon vector for genetic transformation in vertebrates. *J Mol Biol* **302**, 93-102.

Jackson, E. L., Willis, N., Mercer, K., Bronson, R. T., Crowley, D., Montoya, R., Jacks, T., and Tuveson, D. A. (2001). Analysis of lung tumor initiation and progression using conditional expression of oncogenic K-ras. *Genes Dev* **15**, 3243-3248.

Jacobs, J. J., Scheijen, B., Voncken, J. W., Kieboom, K., Berns, A., and van Lohuizen, M. (1999). Bmi-1 collaborates with c-Myc in tumorigenesis by inhibiting c-Myc-induced apoptosis via INK4a/ARF. *Genes Dev* **13**, 2678-2690.

Jacques, C., Baris, O., Prunier-Mirebeau, D., Savagner, F., Rodien, P., Rohmer, V., Franc, B., Guyetant, S., Malthiery, Y., and Reynier, P. (2005). Two-step differential expression analysis reveals a new set of genes involved in thyroid oncocytic tumors. *J Clin Endocrinol Metab* **90**, 2314-2320.

Jalali, G. R., An, Q., Konn, Z. J., Worley, H., Wright, S. L., Harrison, C. J., Strefford, J. C., and Martineau, M. (2008). Disruption of ETV6 in intron 2 results in upregulatory and insertional events in childhood acute lymphoblastic leukaemia. *Leukemia* **22**, 114-123.

Jiang, W., Zhang, Y. J., Kahn, S. M., Hollstein, M. C., Santella, R. M., Lu, S. H., Harris, C. C., Montesano, R., and Weinstein, I. B. (1993). Altered expression of the cyclin D1 and retinoblastoma genes in human esophageal cancer. *Proc Natl Acad Sci U S A* **90**, 9026-9030.

Joensuu, H., Roberts, P. J., Sarlomo-Rikala, M., Andersson, L. C., Tervahartiala, P., Tuveson, D., Silberman, S., Capdeville, R., Dimitrijevic, S., Druker, B., and Demetri, G. D. (2001). Effect of the tyrosine kinase inhibitor STI571 in a patient with a metastatic gastrointestinal stromal tumor. *N Engl J Med* **344**, 1052-1056.

Johansson, F. K., Brodd, J., Eklof, C., Ferletta, M., Hesselager, G., Tiger, C. F., Uhrbom, L., and Westermark, B. (2004). Identification of candidate cancer-causing genes in mouse brain tumors by retroviral tagging. *Proc Natl Acad Sci U S A* **101**, 11334-11337.

Johansson, N., Airola, K., Grenman, R., Kariniemi, A. L., Saarialho-Kere, U., and Kahari, V. M. (1997). Expression of collagenase-3 (matrix metalloproteinase-13) in squamous cell carcinomas of the head and neck. *Am J Pathol* **151**, 499-508.

Johansson, N., Vaalamo, M., Grenman, S., Hietanen, S., Klemi, P., Saarialho-Kere, U., and Kahari, V. M. (1999). Collagenase-3 (MMP-13) is expressed by tumor cells in invasive vulvar squamous cell carcinomas. *Am J Pathol* **154**, 469-480.

Jones, P. A., and Baylin, S. B. (2002). The fundamental role of epigenetic events in cancer. *Nat Rev Genet* **3**, 415-428.

Jones, P. A., and Baylin, S. B. (2007). The epigenomics of cancer. *Cell* **128**, 683-692.

Jones, T. A., Flomen, R. H., Senger, G., Nizetic, D., and Sheer, D. (2000). The homeobox gene MEIS1 is amplified in IMR-32 and highly expressed in other neuroblastoma cell lines. *Eur J Cancer* **36**, 2368-2374.

Jonkers, J., and Berns, A. (2002). Conditional mouse models of sporadic cancer. *Nat Rev Cancer* **2**, 251-265.

Jonkers, J., Meuwissen, R., van der Gulden, H., Peterse, H., van der Valk, M., and Berns, A. (2001). Synergistic tumor suppressor activity of BRCA2 and p53 in a conditional mouse model for breast cancer. *Nat Genet* **29**, 418-425.

Jordan, M. A., Fletcher, J. M., Pellicci, D., and Baxter, A. G. (2007). Slamf1, the NKT cell control gene Nkt1. *J Immunol* **178**, 1618-1627.

Kaira, K., Sunaga, N., Tomizawa, Y., Yanagitani, N., Ishizuka, T., Saito, R., Nakajima, T., and Mori, M. (2007). Epigenetic inactivation of the RAS-effector gene RASSF2 in lung cancers. *Int J Oncol* **31**, 169-173.

Kaiser, J. (2006). Cancer. First pass at cancer genome reveals complex landscape. *Science* **313**, 1370.

Kallioniemi, A., Kallioniemi, O. P., Sudar, D., Rutovitz, D., Gray, J. W., Waldman, F., and Pinkel, D. (1992). Comparative genomic hybridization for molecular cytogenetic analysis of solid tumors. *Science* **258**, 818-821.

Kamps, M. P., Murre, C., Sun, X. H., and Baltimore, D. (1990). A new homeobox gene contributes the DNA binding domain of the t(1;19) translocation protein in pre-B ALL. *Cell* **60**, 547-555.

Kapitonov, V. V., and Jurka, J. (2008). A universal classification of eukaryotic transposable elements implemented in Repbase. *Nat Rev Genet* **9**, 411-412; author reply 414.

Kasamatsu, A., Uzawa, K., Nakashima, D., Koike, H., Shiiba, M., Bukawa, H., Yokoe, H., and Tanzawa, H. (2005). Galectin-9 as a regulator of cellular adhesion in human oral squamous cell carcinoma cell lines. *Int J Mol Med* **16**, 269-273.

Katoh, M., and Katoh, M. (2003). Identification and characterization of human BCL9L gene and mouse Bcl9l gene in silico. *Int J Mol Med* **12**, 643-649.

Katoh, M., and Katoh, M. (2004). Identification and characterization of human MPP7 gene and mouse Mpp7 gene in silico. *Int J Mol Med* **13**, 333-338.

Kehrer-Sawatzki, H. (2007). What a difference copy number variation makes. *Bioessays* **29**, 311-313.

Kendall, J., Liu, Q., Bakleh, A., Krasnitz, A., Nguyen, K. C., Lakshmi, B., Gerald, W. L., Powers, S., and Mu, D. (2007). Oncogenic cooperation and coamplification of developmental transcription factor genes in lung cancer. *Proc Natl Acad Sci U S A* **104**, 16663-16668.

Kennedy, G. C., Matsuzaki, H., Dong, S., Liu, W. M., Huang, J., Liu, G., Su, X., Cao, M., Chen, W., Zhang, J*., et al.* (2003). Large-scale genotyping of complex DNA. *Nat Biotechnol* **21**, 1233-1237.

Kent, W. J. (2002). BLAT--the BLAST-like alignment tool. *Genome Res* **12**, 656-664.

Kent, W. J., Sugnet, C. W., Furey, T. S., Roskin, K. M., Pringle, T. H., Zahler, A. M., and Haussler, D. (2002). The human genome browser at UCSC. *Genome Res* **12**, 996-1006.

Kersemaekers, A. M., Kenter, G. G., Hermans, J., Fleuren, G. J., and van de Vijver, M. J. (1998). Allelic loss and prognosis in carcinoma of the uterine cervix. *Int J Cancer* **79**, 411-417.

Khan, M. M., Nomura, T., Kim, H., Kaul, S. C., Wadhwa, R., Shinagawa, T., Ichikawa-Iwata, E., Zhong, S., Pandolfi, P. P., and Ishii, S. (2001). Role of PML and PML-RARalpha in Mad-mediated transcriptional repression. *Mol Cell* **7**, 1233-1243.

Khojasteh, M., Lam, W. L., Ward, R. K., and MacAulay, C. (2005). A stepwise framework for the normalization of array CGH data. *BMC Bioinformatics* **6**, 274.

Kilbey, A., Blyth, K., Wotton, S., Terry, A., Jenkins, A., Bell, M., Hanlon, L., Cameron, E. R., and Neil, J. C. (2007). Runx2 disruption promotes immortalization and confers resistance to oncogene-induced senescence in primary murine fibroblasts. *Cancer Res* **67**, 11263-11271.

Kim, C. G., Lee, J. J., Jung, D. Y., Jeon, J., Heo, H. S., Kang, H. C., Shin, J. H., Cho, Y. S., Cha, K. J., Kim, C. G.*, et al.* (2006a). Profiling of differentially expressed genes in human stem cells by cDNA microarray. *Mol Cells* **21**, 343-355.

Kim, J., Bhinge, A. A., Morgan, X. C., and Iyer, V. R. (2005a). Mapping DNA-protein interactions in large genomes by sequence tag analysis of genomic enrichment. *Nat Methods* **2**, 47-53.

Kim, J., Sif, S., Jones, B., Jackson, A., Koipally, J., Heller, E., Winandy, S., Viel, A., Sawyer, A., Ikeda, T.*, et al.* (1999). Ikaros DNA-binding proteins direct formation of chromatin remodeling complexes in lymphocytes. *Immunity* **10**, 345-355.

Kim, M., Gans, J. D., Nogueira, C., Wang, A., Paik, J. H., Feng, B., Brennan, C., Hahn, W. C., Cordon-Cardo, C., Wagner, S. N.*, et al.* (2006b). Comparative oncogenomics identifies NEDD9 as a melanoma metastasis gene. *Cell* **125**, 1269-1281.

Kim, R., Trubetskoy, A., Suzuki, T., Jenkins, N. A., Copeland, N. G., and Lenz, J. (2003a). Genome-based identification of cancer genes by proviral tagging in mouse retrovirus-induced T-cell lymphomas. *J Virol* **77**, 2056-2062.

Kim, S. W., Kim, J. W., Kim, Y. T., Kim, J. H., Kim, S., Yoon, B. S., Nam, E. J., and Kim, H. Y. (2007a). Analysis of chromosomal changes in serous ovarian carcinoma using high-resolution array comparative genomic hybridization: Potential predictive markers of chemoresistant disease. *Genes Chromosomes Cancer* **46**, 1-9.

Kim, T. H., Abdullaev, Z. K., Smith, A. D., Ching, K. A., Loukinov, D. I., Green, R. D., Zhang, M. Q., Lobanenkov, V. V., and Ren, B. (2007b). Analysis of the vertebrate insulator protein CTCF-binding sites in the human genome. *Cell* **128**, 1231-1245.

Kim, T. H., Barrera, L. O., Zheng, M., Qu, C., Singer, M. A., Richmond, T. A., Wu, Y., Green, R. D., and Ren, B. (2005b). A high-resolution map of active promoters in the human genome. *Nature* **436**, 876-880.

Kim, Y. M., Watanabe, T., Allen, P. B., Kim, Y. M., Lee, S. J., Greengard, P., Nairn, A. C., and Kwon, Y. G. (2003b). PNUTS, a protein phosphatase 1 (PP1) nuclear targeting subunit. Characterization of its PP1- and RNA-binding domains and regulation by phosphorylation. *J Biol Chem* **278**, 13819-13828.

Kishimoto, I., Mitomi, H., Ohkura, Y., Kanazawa, H., Fukui, N., and Watanabe, M. (2008). Abnormal Expression of p16(INK4a), Cyclin D1, Cyclin-Dependent Kinase 4 and Retinoblastoma Protein in Gastric Carcinomas. *J Surg Oncol* **98**, 60-66.

Klein, S. C., Jucker, M., Abts, H., and Tesch, H. (1995). IL6 and IL6 receptor expression in Burkitt's lymphoma and lymphoblastoid cell lines: promotion of IL6 receptor expression by EBV. *Hematol Oncol* **13**, 121-130.

Klijn, C., Holstege, H., de Ridder, J., Liu, X., Reinders, M., Jonkers, J., and Wessels, L. (2008). Identification of cancer genes using a statistical framework for multiexperiment analysis of nondiscretized array CGH data. *Nucleic Acids Res* **36**, e13.

Klinger, M. B., Guilbault, B., Goulding, R. E., and Kay, R. J. (2005). Deregulated expression of RasGRP1 initiates thymic lymphomagenesis independently of T-cell receptors. *Oncogene* **24**, 2695-2704.

Kloth, J. N., Oosting, J., van Wezel, T., Szuhai, K., Knijnenburg, J., Gorter, A., Kenter, G. G., Fleuren, G. J., and Jordanova, E. S. (2007). Combined array-comparative genomic hybridization and single-nucleotide polymorphism-loss of heterozygosity analysis reveals complex genetic alterations in cervical cancer. *BMC Genomics* **8**, 53.

Kmita, M., Kondo, T., and Duboule, D. (2000). Targeted inversion of a polar silencer within the HoxD complex re-allocates domains of enhancer sharing. *Nat Genet* **26**, 451-454.

Knudson, A. G. (2001). Two genetic hits (more or less) to cancer. *Nat Rev Cancer* **1**, 157-162.

Knudson, A. G., Jr. (1971). Mutation and cancer: statistical study of retinoblastoma. *Proc Natl Acad Sci U S A* **68**, 820-823.

Knutsen, T., Gobu, V., Knaus, R., Padilla-Nash, H., Augustus, M., Strausberg, R. L., Kirsch, I. R., Sirotkin, K., and Ried, T. (2005). The interactive online SKY/M-FISH & CGH database and the Entrez cancer chromosomes search database: linkage of chromosomal aberrations with the genome sequence. *Genes Chromosomes Cancer* **44**, 52-64.

Kobayashi, S., Boggon, T. J., Dayaram, T., Janne, P. A., Kocher, O., Meyerson, M., Johnson, B. E., Eck, M. J., Tenen, D. G., and Halmos, B. (2005). EGFR mutation and resistance of non-small-cell lung cancer to gefitinib. *N Engl J Med* **352**, 786-792.

Kochetkova, M., McKenzie, O. L., Bais, A. J., Martin, J. M., Secker, G. A., Seshadri, R., Powell, J. A., Hinze, S. J., Gardner, A. E., Spendlove, H. E*., et al.* (2002). CBFA2T3 (MTG16) is a putative breast tumor suppressor gene from the breast cancer loss of heterozygosity region at 16q24.3. *Cancer Res* **62**, 4599-4604.

Koinuma, D., and Imamura, T. (2005). [Bone formation and inflammation]. *Nippon Rinsho* **63**, 1523-1528.

Kojima, Y., Miyoshi, H., Clevers, H. C., Oshima, M., Aoki, M., and Taketo, M. M. (2007). Suppression of tubulin polymerization by the LKB1-microtubule-associated protein/microtubule affinity-regulating kinase signaling. *J Biol Chem* **282**, 23532-23540.

Kominsky, S. L. (2006). Claudins: emerging targets for cancer therapy. *Expert Rev Mol Med* **8**, 1-11.

Komuro, H., Valentine, M. B., Rubnitz, J. E., Saito, M., Raimondi, S. C., Carroll, A. J., Yi, T., Sherr, C. J., and Look, A. T. (1999). p27KIP1 deletions in childhood acute lymphoblastic leukemia. *Neoplasia* **1**, 253-261.

Koon, H. B., Ippolito, G. C., Banham, A. H., and Tucker, P. W. (2007). FOXP1: a potential therapeutic target in cancer. *Expert Opin Ther Targets* **11**, 955-965.

Koontz, J. I., Soreng, A. L., Nucci, M., Kuo, F. C., Pauwels, P., van Den Berghe, H., Cin, P. D., Fletcher, J. A., and Sklar, J. (2001). Frequent fusion of the JAZF1 and JJAZ1 genes in endometrial stromal tumors. *Proc Natl Acad Sci U S A* **98**, 6348-6353.

Krieg, A. J., Hammond, E. M., and Giaccia, A. J. (2006). Functional analysis of p53 binding under differential stresses. *Mol Cell Biol* **26**, 7030-7045.

Kudo, A., and Melchers, F. (1987). A second gene, VpreB in the lambda 5 locus of the mouse, which appears to be selectively expressed in pre-B lymphocytes. *Embo J* **6**, 2267-2272.

Kuivanen, T. T., Jeskanen, L., Kyllonen, L., Impola, U., and Saarialho-Kere, U. K. (2006). Transformation-specific matrix metalloproteinases, MMP-7 and MMP-13, are present in epithelial cells of keratoacanthomas. *Mod Pathol* **19**, 1203-1212.

Kumar, K. R., Li, L., Yan, M., Bhaskarabhatla, M., Mobley, A. B., Nguyen, C., Mooney, J. M., Schatzle, J. D., Wakeland, E. K., and Mohan, C. (2006). Regulation of B cell tolerance by the lupus susceptibility gene Ly108. *Science* **312**, 1665-1669.

Kunitz, A., Wolter, M., van den Boom, J., Felsberg, J., Tews, B., Hahn, M., Benner, A., Sabel, M., Lichter, P., Reifenberger, G., *et al.* (2007). DNA hypermethylation and aberrant expression of the EMP3 gene at 19q13.3 in Human Gliomas. *Brain Pathol* **17**, 363-370.

Kutay, H., Bai, S., Datta, J., Motiwala, T., Pogribny, I., Frankel, W., Jacob, S. T., and Ghoshal, K. (2006). Downregulation of miR-122 in the rodent and human hepatocellular carcinomas. *J Cell Biochem* **99**, 671-678.

Kwiatkowski, B. A., Bastian, L. S., Bauer, T. R., Jr., Tsai, S., Zielinska-Kwiatkowska, A. G., and Hickstein, D. D. (1998). The ets family member Tel binds to the Fli-1 oncoprotein and inhibits its transcriptional activity. *J Biol Chem* **273**, 17525-17530.

LaFramboise, T., Weir, B. A., Zhao, X., Beroukhim, R., Li, C., Harrington, D., Sellers, W. R., and Meyerson, M. (2005). Allele-specific amplification in cancer revealed by SNP array analysis. *PLoS Comput Biol* **1**, e65.

Lai, W., Choudhary, V., and Park, P. J. (2008). CGHweb: a tool for comparing DNA copy number segmentations from multiple algorithms. *Bioinformatics* **24**, 1014-1015.

Lai, W. R., Johnson, M. D., Kucherlapati, R., and Park, P. J. (2005). Comparative analysis of algorithms for identifying amplifications and deletions in array CGH data. *Bioinformatics* **21**, 3763-3770.

Lakso, M., Sauer, B., Mosinger, B., Jr., Lee, E. J., Manning, R. W., Yu, S. H., Mulder, K. L., and Westphal, H. (1992). Targeted oncogene activation by site-specific recombination in transgenic mice. *Proc Natl Acad Sci U S A* **89**, 6232-6236.

Lallemand-Breitenbach, V., Guillemin, M. C., Janin, A., Daniel, M. T., Degos, L., Kogan, S. C., Bishop, J. M., and de The, H. (1999). Retinoic acid and arsenic synergize to eradicate leukemic cells in a mouse model of acute promyelocytic leukemia. *J Exp Med* **189**, 1043-1052.

Lam, S. H., Wu, Y. L., Vega, V. B., Miller, L. D., Spitsbergen, J., Tong, Y., Zhan, H., Govindarajan, K. R., Lee, S., Mathavan, S.*, et al.* (2006). Conservation of gene expression signatures between zebrafish and human liver tumors and tumor progression. *Nat Biotechnol* **24**, 73-75.

Lamy, P., Andersen, C. L., Dyrskjot, L., Torring, N., and Wiuf, C. (2007). A Hidden Markov Model to estimate population mixture and allelic copy-numbers in cancers using Affymetrix SNP arrays. *BMC Bioinformatics* **8**, 434.

Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W.*, et al.* (2001). Initial sequencing and analysis of the human genome. *Nature* **409**, 860-921.

Largaespada, D. A. (2000). Genetic heterogeneity in acute myeloid leukemia: maximizing information flow from MuLV mutagenesis studies. *Leukemia* **14**, 1174-1184.

Largaespada, D. A., Brannan, C. I., Shaughnessy, J. D., Jenkins, N. A., and Copeland, N. G. (1996). The neurofibromatosis type 1 (NF1) tumor suppressor gene and myeloid leukemia. *Curr Top Microbiol Immunol* **211**, 233-239.

Largaespada, D. A., and Collier, L. S. (2008). Transposon-mediated mutagenesis in somatic cells: identification of transposon-genomic DNA junctions. *Methods Mol Biol* **435**, 95-108.

Latil, A., Chene, L., Cochant-Priollet, B., Mangin, P., Fournier, G., Berthon, P., and Cussenot, O. (2003). Quantification of expression of netrins, slits and their receptors in human prostate tumors. *Int J Cancer* **103**, 306-315.

Latil, A., Morant, P., Fournier, G., Mangin, P., Berthon, P., and Cussenot, O. (2002). CHC1-L, a candidate gene for prostate carcinogenesis at 13q14.2, is frequently affected by loss of heterozygosity and underexpressed in human prostate cancer. *Int J Cancer* **99**, 689-696.

Lawrence, H. J., Rozenfeld, S., Cruz, C., Matsukuma, K., Kwong, A., Komuves, L., Buchberg, A. M., and Largman, C. (1999). Frequent co-expression of the HOXA9 and MEIS1 homeobox genes in human myeloid leukemias. *Leukemia* **13**, 1993-1999.

Lazo, P. A., Lee, J. S., and Tsichlis, P. N. (1990). Long-distance activation of the Myc protooncogene by provirus insertion in Mlvi-1 or Mlvi-4 in rat T-cell lymphomas. *Proc Natl Acad Sci U S A* **87**, 170-173.

Le Scolan, E., Zhu, Q., Wang, L., Bandyopadhyay, A., Javelaud, D., Mauviel, A., Sun, L., and Luo, K. (2008). Transforming growth factor-beta suppresses the ability of Ski to inhibit tumor metastasis by inducing its degradation. *Cancer Res* **68**, 3277-3285.

Lebigot, I., Gardellin, P., Lefebvre, L., Beug, H., Ghysdael, J., and Quang, C. T. (2003). Up-regulation of SLAP in FLI-1-transformed erythroblasts interferes with EpoR signaling. *Blood* **102**, 4555-4562.

Lee, J., Beliakoff, J., and Sun, Z. (2007). The novel PIAS-like protein hZimp10 is a transcriptional co-activator of the p53 tumor suppressor. *Nucleic Acids Res* **35**, 4523-4534.

Lee, J. S., Chu, I. S., Mikaelyan, A., Calvisi, D. F., Heo, J., Reddy, J. K., and Thorgeirsson, S. S. (2004). Application of comparative functional genomics to identify best-fit mouse models to study human cancer. *Nat Genet* **36**, 1306-1311.

Lee, T. I., Jenner, R. G., Boyer, L. A., Guenther, M. G., Levine, S. S., Kumar, R. M., Chevalier, B., Johnstone, S. E., Cole, M. F., Isono, K*., et al.* (2006). Control of developmental regulators by Polycomb in human embryonic stem cells. *Cell* **125**, 301-313.

Lengauer, C., Kinzler, K. W., and Vogelstein, B. (1998). Genetic instabilities in human cancers. *Nature* **396**, 643-649.

Letessier, A., Garrido-Urbani, S., Ginestier, C., Fournier, G., Esterni, B., Monville, F., Adelaide, J., Geneix, J., Xerri, L., Dubreuil, P*., et al.* (2007). Correlated break at PARK2/FRA6E and loss of AF-6/Afadin protein expression are associated with poor outcome in breast cancer. *Oncogene* **26**, 298-307.

Leverson, J. D., and Ness, S. A. (1998). Point mutations in v-Myb disrupt a cyclophilin-catalyzed negative regulatory mechanism. *Mol Cell* **1**, 203-211.

Levine, D. A., Bogomolniy, F., Yee, C. J., Lash, A., Barakat, R. R., Borgen, P. I., and Boyd, J. (2005). Frequent mutation of the PIK3CA gene in ovarian and breast cancers. *Clin Cancer Res* **11**, 2875-2878.

Li, D. Q., Hou, Y. F., Wu, J., Chen, Y., Lu, J. S., Di, G. H., Ou, Z. L., Shen, Z. Z., Ding, J., and Shao, Z. M. (2006a). Gene expression profile analysis of an isogenic tumour metastasis model reveals a functional role for oncogene AF1Q in breast cancer metastasis. *Eur J Cancer* **42**, 3274-3286.

Li, G. C., Ouyang, H., Li, X., Nagasawa, H., Little, J. B., Chen, D. J., Ling, C. C., Fuks, Z., and Cordon-Cardo, C. (1998). Ku70: a candidate tumor suppressor gene for murine T cell lymphoma. *Mol Cell* **2**, 1-8.

Li, J., Shen, H., Himmel, K. L., Dupuy, A. J., Largaespada, D. A., Nakamura, T., Shaughnessy, J. D., Jr., Jenkins, N. A., and Copeland, N. G. (1999). Leukaemia disease genes: large-scale cloning and pathway predictions. *Nat Genet* **23**, 348-353.

Li, J., Yen, C., Liaw, D., Podsypanina, K., Bose, S., Wang, S. I., Puc, J., Miliaresis, C., Rodgers, L., McCombie, R*., et al.* (1997). PTEN, a putative protein tyrosine phosphatase gene mutated in human brain, breast, and prostate cancer. *Science* **275**, 1943-1947.

Li, K. K., Pang, J. C., Chung, N. Y., Ng, Y. L., Chan, N. H., Zhou, L., Poon, W. S., and Ng, H. K. (2007a). EMP3 overexpression is associated with oligodendroglial tumors retaining chromosome arms 1p and 19q. *Int J Cancer* **120**, 947-950.

Li, X., Thyssen, G., Beliakoff, J., and Sun, Z. (2006b). The novel PIAS-like protein hZimp10 enhances Smad transcriptional activity. *J Biol Chem* **281**, 23748-23756.

Li, Y., Huang, J., Zhao, Y. L., He, J., Wang, W., Davies, K. E., Nose, V., and Xiao, S. (2007b). UTRN on chromosome 6q24 is mutated in multiple tumors. *Oncogene* **26**, 6220-6228.

Liggett, W. H., Jr., and Sidransky, D. (1998). Role of the p16 tumor suppressor gene in cancer. *J Clin Oncol* **16**, 1197-1206.

Lilljebjorn, H., Heidenblad, M., Nilsson, B., Lassen, C., Horvat, A., Heldrup, J., Behrendtz, M., Johansson, B., Andersson, A., and Fioretos, T. (2007). Combined high-resolution array-based comparative genomic hybridization and expression profiling of ETV6/RUNX1-positive acute lymphoblastic leukemias reveal a high incidence of cryptic Xq duplications and identify several putative target genes within the commonly gained region. *Leukemia* **21**, 2137-2144.

Lin, A. W., Barradas, M., Stone, J. C., van Aelst, L., Serrano, M., and Lowe, S. W. (1998). Premature senescence involving p53 and p16 is activated in response to constitutive MEK/MAPK mitogenic signaling. *Genes Dev* **12**, 3008-3019.

Lin, J., Gan, C. M., Zhang, X., Jones, S., Sjoblom, T., Wood, L. D., Parsons, D. W., Papadopoulos, N., Kinzler, K. W., Vogelstein, B., *et al.* (2007). A multidimensional analysis of genes mutated in breast and colorectal cancers. *Genome Res* **17**, 1304-1318.

Liptay, S., Schmid, R. M., Perkins, N. D., Meltzer, P., Altherr, M. R., McPherson, J. D., Wasmuth, J. J., and Nabel, G. J. (1992). Related subunits of NF-kappa B map to two distinct loci associated with translocations in leukemia, NFKB1 and NFKB2. *Genomics* **13**, 287-292.

Liu, X., Vorontchikhina, M., Wang, Y. L., Faiola, F., and Martinez, E. (2008). STAGA recruits Mediator to the MYC oncoprotein to stimulate transcription and cell proliferation. *Mol Cell Biol* **28**, 108-121.

Lizcano, J. M., Goransson, O., Toth, R., Deak, M., Morrice, N. A., Boudeau, J., Hawley, S. A., Udd, L., Makela, T. P., Hardie, D. G., and Alessi, D. R. (2004). LKB1 is a master kinase that activates 13 kinases of the AMPK subfamily, including MARK/PAR-1. *Embo J* **23**, 833-843.

Loeb, L. A., Loeb, K. R., and Anderson, J. P. (2003). Multiple mutations and cancer. *Proc Natl Acad Sci U S A* **100**, 776-781.

Loh, Y. H., Wu, Q., Chew, J. L., Vega, V. B., Zhang, W., Chen, X., Bourque, G., George, J., Leong, B., Liu, J., *et al.* (2006). The Oct4 and Nanog transcription network regulates pluripotency in mouse embryonic stem cells. *Nat Genet* **38**, 431-440.

Lopez-Nieva, P., Santos, J., and Fernandez-Piqueras, J. (2004). Defective expression of Notch1 and Notch2 in connection to alterations of c-Myc and Ikaros in gamma-radiation-induced mouse thymic lymphomas. *Carcinogenesis* **25**, 1299-1304.

Luke, C. T., Oki-Idouchi, C. E., Cline, J. M., and Lorenzo, P. S. (2007). RasGRP1 overexpression in the epidermis of transgenic mice contributes to tumor progression during multistage skin carcinogenesis. *Cancer Res* **67**, 10190-10197.

Lund, A. H., Turner, G., Trubetskoy, A., Verhoeven, E., Wientjens, E., Hulsman, D., Russell, R., DePinho, R. A., Lenz, J., and van Lohuizen, M. (2002). Genome-wide retroviral insertional tagging of genes involved in cancer in Cdkn2a-deficient mice. *Nat Genet* **32**, 160-165.

Luo, G., Santoro, I. M., McDaniel, L. D., Nishijima, I., Mills, M., Youssoufian, H., Vogel, H., Schultz, R. A., and Bradley, A. (2000). Cancer predisposition caused by elevated mitotic recombination in Bloom mice. *Nat Genet* **26**, 424-429.

Ma, K., Araki, K., Ichwan, S. J., Suganuma, T., Tamamori-Adachi, M., and Ikeda, M. A. (2003). E2FBP1/DRIL1, an AT-rich interaction domain-family transcription factor, is regulated by p53. *Mol Cancer Res* **1**, 438-444.

Maeda, T., Hobbs, R. M., Merghoub, T., Guernah, I., Zelent, A., Cordon-Cardo, C., Teruya-Feldstein, J., and Pandolfi, P. P. (2005). Role of the proto-oncogene Pokemon in cellular transformation and ARF repression. *Nature* **433**, 278-285.

Maeda, T., Merghoub, T., Hobbs, R. M., Dong, L., Maeda, M., Zakrzewski, J., van den Brink, M. R., Zelent, A., Shigematsu, H., Akashi, K., *et al.* (2007). Regulation of B versus T lymphoid lineage fate decision by the proto-oncogene LRF. *Science* **316**, 860-866.

Maher, E. A., Brennan, C., Wen, P. Y., Durso, L., Ligon, K. L., Richardson, A., Khatry, D., Feng, B., Sinha, R., Louis, D. N., *et al.* (2006). Marked genomic differences characterize primary and secondary glioblastoma subtypes and identify two distinct molecular and clinical secondary glioblastoma entities. *Cancer Res* **66**, 11502-11513.

Mani, S. A., Guo, W., Liao, M. J., Eaton, E. N., Ayyanan, A., Zhou, A. Y., Brooks, M., Reinhard, F., Zhang, C. C., Shipitsin, M., *et al.* (2008). The epithelial-mesenchymal transition generates cells with properties of stem cells. *Cell* **133**, 704-715.

Marchetti, A., Buttitta, F., Girlando, S., Dalla Palma, P., Pellegrini, S., Fina, P., Doglioni, C., Bevilacqua, G., and Barbareschi, M. (1995a). mdm2 gene alterations and mdm2 protein expression in breast carcinomas. *J Pathol* **175**, 31-38.

Marchetti, A., Buttitta, F., Pellegrini, S., Merlo, G., Chella, A., Angeletti, C. A., and Bevilacqua, G. (1995b). mdm2 gene amplification and overexpression in non-small cell lung carcinomas with accumulation of the p53 protein in the absence of p53 gene mutations. *Diagn Mol Pathol* **4**, 93-97.

Marioni, J. C., Thorne, N. P., and Tavare, S. (2006). BioHMM: a heterogeneous hidden Markov model for segmenting array CGH data. *Bioinformatics* **22**, 1144-1146.

Marioni, J. C., Thorne, N. P., Valsesia, A., Fitzgerald, T., Redon, R., Fiegler, H., Andrews, T. D., Stranger, B. E., Lynch, A. G., Dermitzakis, E. T., *et al.* (2007). Breaking the waves: improved detection of copy number variation from microarray-based comparative genomic hybridization. *Genome Biol* **8**, R228.

Martins, C. P., and Berns, A. (2002). Loss of p27(Kip1) but not p21(Cip1) decreases survival and synergizes with MYC in murine lymphomagenesis. *Embo J* **21**, 3739-3748.

Maser, R. S., Choudhury, B., Campbell, P. J., Feng, B., Wong, K. K., Protopopov, A., O'Neil, J., Gutierrez, A., Ivanova, E., Perna, I.*, et al.* (2007). Chromosomally unstable mouse tumours have genomic alterations similar to diverse human cancers. *Nature* **447**, 966-971.

Maser, R. S., and DePinho, R. A. (2002). Connecting chromosomes, crisis, and cancer. *Science* **297**, 565-569.

Mathias, S. L., Scott, A. F., Kazazian, H. H., Jr., Boeke, J. D., and Gabriel, A. (1991). Reverse transcriptase encoded by a human transposable element. *Science* **254**, 1808-1810.

Matsuzaki, H., Dong, S., Loi, H., Di, X., Liu, G., Hubbell, E., Law, J., Berntsen, T., Chadha, M., Hui, H.*, et al.* (2004). Genotyping over 100,000 SNPs on a pair of oligonucleotide arrays. *Nat Methods* **1**, 109-111.

McClintock, B. (1941). The Stability of Broken Ends of Chromosomes in Zea Mays. *Genetics* **26**, 234-282.

Megidish, T., Cooper, J., Zhang, L., Fu, H., and Hakomori, S. (1998). A novel sphingosine-dependent protein kinase (SDK1) specifically phosphorylates certain isoforms of 14-3-3 protein. *J Biol Chem* **273**, 21834-21845.

Mejlvang, J., Kriajevska, M., Vandewalle, C., Chernova, T., Sayan, A. E., Berx, G., Mellon, J. K., and Tulchinsky, E. (2007). Direct repression of cyclin D1 by SIP1 attenuates cell cycle progression in cells undergoing an epithelial mesenchymal transition. *Mol Biol Cell* **18**, 4615-4624.

Melani, M., Simpson, K. J., Brugge, J. S., and Montell, D. (2008). Regulation of cell adhesion and collective cell migration by hindsight and its human homolog RREB1. *Curr Biol* **18**, 532-537.

Merlo, A., Herman, J. G., Mao, L., Lee, D. J., Gabrielson, E., Burger, P. C., Baylin, S. B., and Sidransky, D. (1995). 5' CpG island methylation is associated with transcriptional silencing of the tumour suppressor p16/CDKN2/MTS1 in human cancers. *Nat Med* **1**, 686-692.

Metaye, T., Levillain, P., Kraimps, J. L., and Perdrisot, R. (2008). Immunohistochemical detection, regulation and antiproliferative function of G-protein-coupled receptor kinase 2 in thyroid carcinomas. *J Endocrinol* **198**, 101-110.

Meza-Zepeda, L. A., Kresse, S. H., Barragan-Polania, A. H., Bjerkehagen, B., Ohnstad, H. O., Namlos, H. M., Wang, J., Kristiansen, B. E., and Myklebost, O. (2006). Array comparative genomic hybridization reveals distinct DNA copy number differences between gastrointestinal stromal tumors and leiomyosarcomas. *Cancer Res* **66**, 8984-8993.

Micci, F., Panagopoulos, I., Bjerkehagen, B., and Heim, S. (2006). Consistent rearrangement of chromosomal band 6p21 with generation of fusion genes JAZF1/PHF1 and EPC1/PHF1 in endometrial stromal sarcoma. *Cancer Res* **66**, 107-112.

Mikkers, H., Allen, J., Knipscheer, P., Romeijn, L., Hart, A., Vink, E., and Berns, A. (2002). High-throughput retroviral tagging to identify components of specific signaling pathways in cancer. *Nat Genet* **32**, 153-159.

Mikkers, H., and Berns, A. (2003). Retroviral insertional mutagenesis: tagging cancer pathways. *Adv Cancer Res* **88**, 53-99.

Misra, J. S., Das, K., and Chandrawati (1998). Results of clinically downstaging cervical cancer in a cytological screening programme. *Diagn Cytopathol* **19**, 344-348.

Mitchell, R. S., Beitzel, B. F., Schroder, A. R., Shinn, P., Chen, H., Berry, C. C., Ecker, J. R., and Bushman, F. D. (2004). Retroviral DNA integration: ASLV, HIV, and MLV show distinct target site preferences. *PLoS Biol* **2**, E234.

Mitelman, F., Johansson, B., and Mertens, F. (2004). Fusion genes and rearranged genes as a linear function of chromosome aberrations in cancer. *Nat Genet* **36**, 331-334.

Miyamoto, K., Fukutomi, T., Akashi-Tanaka, S., Hasegawa, T., Asahara, T., Sugimura, T., and Ushijima, T. (2005). Identification of 20 genes aberrantly methylated in human breast cancers. *Int J Cancer* **116**, 407-414.

Moloney, D. J., Panin, V. M., Johnston, S. H., Chen, J., Shao, L., Wilson, R., Wang, Y., Stanley, P., Irvine, K. D., Haltiwanger, R. S., and Vogt, T. F. (2000). Fringe is a glycosyltransferase that modifies Notch. *Nature* **406**, 369-375.

Momand, J., Wu, H. H., and Dasgupta, G. (2000). MDM2--master regulator of the p53 tumor suppressor protein. *Gene* **242**, 15-29.

Monzo, M., Navarro, A., Bandres, E., Artells, R., Moreno, I., Gel, B., Ibeas, R., Moreno, J., Martinez, F., Diaz, T.*, et al.* (2008). Overlapping expression of microRNAs in human embryonic colon and colorectal cancer. *Cell Res* **18**, 823-833.

Moran, J. V., Holmes, S. E., Naas, T. P., DeBerardinis, R. J., Boeke, J. D., and Kazazian, H. H., Jr. (1996). High frequency retrotransposition in cultured mammalian cells. *Cell* **87**, 917-927.

Mucenski, M. L., Taylor, B. A., Copeland, N. G., and Jenkins, N. A. (1988a). Chromosomal location of Evi-1, a common site of ecotropic viral integration in AKXD murine myeloid tumors. *Oncogene Res* **2**, 219-233.

Mucenski, M. L., Taylor, B. A., Ihle, J. N., Hartley, J. W., Morse, H. C., 3rd, Jenkins, N. A., and Copeland, N. G. (1988b). Identification of a common ecotropic viral integration site, Evi-1, in the DNA of AKXD murine myeloid tumors. *Mol Cell Biol* **8**, 301-308.

Mulder, J., Poland, M., Gebbink, M. F., Calafat, J., Moolenaar, W. H., and Kranenburg, O. (2003). p116Rip is a novel filamentous actin-binding protein. *J Biol Chem* **278**, 27216-27223.

Mulholland, P. J., Fiegler, H., Mazzanti, C., Gorman, P., Sasieni, P., Adams, J., Jones, T. A., Babbage, J. W., Vatcheva, R., Ichimura, K.*, et al.* (2006). Genomic profiling identifies discrete deletions associated with translocations in glioblastoma multiforme. *Cell Cycle* **5**, 783-791.

Muller, D., Millon, R., Theobald, S., Hussenet, T., Wasylyk, B., du Manoir, S., and Abecassis, J. (2006). Cyclin L1 (CCNL1) gene alterations in human head and neck squamous cell carcinoma. *Br J Cancer* **94**, 1041-1044.

Muller, H. P., and Varmus, H. E. (1994). DNA bending creates favored sites for retroviral integration: an explanation for preferred insertion sites in nucleosomes. *Embo J* **13**, 4704-4714.

Muller, U. (1999). Ten years of gene targeting: targeted mouse mutants, from vector design to phenotype analysis. *Mech Dev* **82**, 3-21.

Mullighan, C. G., Goorha, S., Radtke, I., Miller, C. B., Coustan-Smith, E., Dalton, J. D., Girtman, K., Mathew, S., Ma, J., Pounds, S. B*., et al.* (2007). Genome-wide analysis of genetic alterations in acute lymphoblastic leukaemia. *Nature* **446**, 758-764.

Mullighan, C. G., Miller, C. B., Radtke, I., Phillips, L. A., Dalton, J., Ma, J., White, D., Hughes, T. P., Le Beau, M. M., Pui, C. H*., et al.* (2008). BCR-ABL1 lymphoblastic leukaemia is characterized by the deletion of Ikaros. *Nature* **453**, 110-114.

Mundt, C., Licence, S., Shimizu, T., Melchers, F., and Martensson, I. L. (2001). Loss of precursor B cell expansion but not allelic exclusion in VpreB1/VpreB2 double-deficient mice. *J Exp Med* **193**, 435-445.

Nagar, B., Bornmann, W. G., Pellicena, P., Schindler, T., Veach, D. R., Miller, W. T., Clarkson, B., and Kuriyan, J. (2002). Crystal structures of the kinase domain of c-Abl in complex with the small molecule inhibitors PD173955 and imatinib (STI-571). *Cancer Res* **62**, 4236-4243.

Nagata, Y., Lan, K. H., Zhou, X., Tan, M., Esteva, F. J., Sahin, A. A., Klos, K. S., Li, P., Monia, B. P., Nguyen, N. T*., et al.* (2004). PTEN activation contributes to tumor inhibition by trastuzumab, and loss of PTEN predicts trastuzumab resistance in patients. *Cancer Cell* **6**, 117-127.

Nagy, B., Lundan, T., Larramendy, M. L., Aalto, Y., Zhu, Y., Niini, T., Edgren, H., Ferrer, A., Vilpo, J., Elonen, E*., et al.* (2003). Abnormal expression of apoptosis-related genes in haematological malignancies: overexpression of MYC is poor prognostic sign in mantle cell lymphoma. *Br J Haematol* **120**, 434-441.

Nakahara, T., Tominaga, K., Koseki, T., Yamamoto, M., Yamato, K., Fukuda, J., and Nishihara, T. (2003). Growth/differentiation factor-5 induces growth arrest and apoptosis in mouse B lineage cells with modulation by Smad. *Cell Signal* **15**, 181-187.

Nakamura, T., Largaespada, D. A., Shaughnessy, J. D., Jr., Jenkins, N. A., and Copeland, N. G. (1996). Cooperative activation of Hoxa and Pbx1-related genes in murine myeloid leukaemias. *Nat Genet* **12**, 149-153.

Nannya, Y., Sanada, M., Nakazaki, K., Hosoya, N., Wang, L., Hangaishi, A., Kurokawa, M., Chiba, S., Bailey, D. K., Kennedy, G. C., and Ogawa, S. (2005). A robust algorithm for copy number detection using high-density oligonucleotide single nucleotide polymorphism genotyping arrays. *Cancer Res* **65**, 6071-6079.

Narayan, G., Bourdon, V., Chaganti, S., Arias-Pulido, H., Nandula, S. V., Rao, P. H., Gissmann, L., Durst, M., Schneider, A., Pothuri, B*., et al.* (2007). Gene dosage alterations revealed by cDNA microarray analysis in cervical cancer: identification of candidate amplified and overexpressed genes. *Genes Chromosomes Cancer* **46**, 373-384.

Natarajan, G., Ramalingam, S., Ramachandran, I., May, R., Queimado, L., Houchen, C. W., and Anant, S. (2008). CUGBP2 downregulation by prostaglandin E2 protects colon cancer cells from radiation-induced mitotic catastrophe. *Am J Physiol Gastrointest Liver Physiol* **294**, G1235-1244.

Naud, J. F., and Eilers, M. (2007). PIM1 and MYC: a changing relationship? *Nat Cell Biol* **9**, 873-875.

Navarro, A., Gaya, A., Martinez, A., Urbano-Ispizua, A., Pons, A., Balague, O., Gel, B., Abrisqueta, P., Lopez-Guillermo, A., Artells, R*., et al.* (2008). MicroRNA expression profiling in classic Hodgkin lymphoma. *Blood* **111**, 2825-2832.

Neil, J. C., and Cameron, E. R. (2002). Retroviral insertion sites and cancer: fountain of all knowledge? *Cancer Cell* **2**, 253-255.

Neuvial, P., Hupe, P., Brito, I., Liva, S., Manie, E., Brennetot, C., Radvanyi, F., Aurias, A., and Barillot, E. (2006). Spatial normalization of array-CGH data. *BMC Bioinformatics* **7**, 264.

Nieto, M., Finn, S., Loda, M., and Hahn, W. C. (2007). Prostate cancer: Re-focusing on androgen receptor signaling. *Int J Biochem Cell Biol* **39**, 1562-1568.

Nikaido, T., Li, S. F., Shiozawa, T., and Fujii, S. (1996). Coabnormal expression of cyclin D1 and p53 protein in human uterine endometrial carcinomas. *Cancer* **78**, 1248-1253.

Ning, Z., Cox, A. J., and Mullikin, J. C. (2001). SSAHA: a fast search method for large DNA databases. *Genome Res* **11**, 1725-1729.

Nishida, N., Nagasaka, T., Nishimura, T., Ikai, I., Boland, C. R., and Goel, A. (2008). Aberrant methylation of multiple tumor suppressor genes in aging liver, chronic hepatitis, and hepatocellular carcinoma. *Hepatology* **47**, 908-918.

Nomoto, S., Haruki, N., Takahashi, T., Masuda, A., Koshikawa, T., Takahashi, T., Fujii, Y., Osada, H., and Takahashi, T. (1999). Search for in vivo somatic mutations in the mitotic checkpoint gene, hMAD1, in human lung cancers. *Oncogene* **18**, 7180-7183.

Nowell, P. C. (1976). The clonal evolution of tumor cell populations. *Science* **194**, 23-28.

O'Hagan, R. C., Brennan, C. W., Strahs, A., Zhang, X., Kannan, K., Donovan, M., Cauwels, C., Sharpless, N. E., Wong, W. H., and Chin, L. (2003). Array comparative genome hybridization for tumor classification and gene discovery in mouse models of malignant melanoma. *Cancer Res* **63**, 5352-5356.

O'Hagan, R. C., Chang, S., Maser, R. S., Mohan, R., Artandi, S. E., Chin, L., and DePinho, R. A. (2002). Telomere dysfunction provokes regional amplification and deletion in cancer genomes. *Cancer Cell* **2**, 149-155.

Ochman, H., Gerber, A. S., and Hartl, D. L. (1988). Genetic applications of an inverse polymerase chain reaction. *Genetics* **120**, 621-623.

Ofek, P., Ben-Meir, D., and Lavi, S. (2003). An inducible system to study the growth arrest properties of protein phosphatase 2C. *Methods Enzymol* **366**, 338-347.

Ohi, H., Mishima, Y., Kamimura, K., Maruyama, M., Sasai, K., and Kominami, R. (2007). Multi-step lymphomagenesis deduced from DNA changes in thymic lymphomas and atrophic thymuses at various times after gamma-irradiation. *Oncogene* **26**, 5280-5289.

Okabe, T., Bauer, S. R., and Kudo, A. (1992). Pre-B lymphocyte-specific transcriptional control of the mouse VpreB gene. *Eur J Immunol* **22**, 31-36.

Okano, H., Saito, Y., Miyazawa, T., Shinbo, T., Chou, D., Kosugi, S., Takahashi, Y., Odani, S., Niwa, O., and Kominami, R. (1999). Homozygous deletions and point mutations of the Ikaros gene in gamma-ray-induced mouse thymic lymphomas. *Oncogene* **18**, 6677-6683.

Oki-Idouchi, C. E., and Lorenzo, P. S. (2007). Transgenic overexpression of RasGRP1 in mouse epidermis results in spontaneous tumors of the skin. *Cancer Res* **67**, 276-280.

Okuda, K., Weisberg, E., Gilliland, D. G., and Griffin, J. D. (2001). ARG tyrosine kinase activity is inhibited by STI571. *Blood* **97**, 2440-2448.

Olshen, A. B., Venkatraman, E. S., Lucito, R., and Wigler, M. (2004). Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics* **5**, 557-572.

Opavsky, R., Tsai, S. Y., Guimond, M., Arora, A., Opavska, J., Becknell, B., Kaufmann, M., Walton, N. A., Stephens, J. A., Fernandez, S. A.*, et al.* (2007). Specific tumor suppressor function for E2F2 in Myc-induced T cell lymphomagenesis. *Proc Natl Acad Sci U S A* **104**, 15400-15405.

Orlow, I., Lacombe, L., Hannon, G. J., Serrano, M., Pellicer, I., Dalbagni, G., Reuter, V. E., Zhang, Z. F., Beach, D., and Cordon-Cardo, C. (1995). Deletion of the p16 and p15 genes in human bladder tumors. *J Natl Cancer Inst* **87**, 1524-1529.

Osaki, M., Inoue, T., Yamaguchi, S., Inaba, A., Tokuyasu, N., Jeang, K. T., Oshimura, M., and Ito, H. (2007). MAD1 (mitotic arrest deficiency 1) is a candidate for a tumor suppressor gene in human stomach. *Virchows Arch* **451**, 771-779.

Oxford, G., Owens, C. R., Titus, B. J., Foreman, T. L., Herlevsen, M. C., Smith, S. C., and Theodorescu, D. (2005). RalA and RalB: antagonistic relatives in cancer cell migration. *Cancer Res* **65**, 7111-7120.

Oxford, G., Smith, S. C., Hampton, G., and Theodorescu, D. (2007). Expression profiling of Ral-depleted bladder cancer cells identifies RREB-1 as a novel transcriptional Ral effector. *Oncogene* **26**, 7143-7152.

Ozturk, N., Erdal, E., Mumcuoglu, M., Akcali, K. C., Yalcin, O., Senturk, S., Arslan-Ergul, A., Gur, B., Yulug, I., Cetin-Atalay, R.*, et al.* (2006). Reprogramming of replicative senescence in hepatocellular carcinoma-derived cells. *Proc Natl Acad Sci U S A* **103**, 2178-2183.

Palacios, E. H., and Weiss, A. (2004). Function of the Src-family kinases, Lck and Fyn, in T-cell development and activation. *Oncogene* **23**, 7990-8000.

Panagopoulos, I., Strombeck, B., Isaksson, M., Heldrup, J., Olofsson, T., and Johansson, B. (2006). Fusion of ETV6 with an intronic sequence of the BAZ2A gene in a paediatric pre-B acute lymphoblastic leukaemia with a cryptic chromosome 12 rearrangement. *Br J Haematol* **133**, 270-275.

Pao, W., Wang, T. Y., Riely, G. J., Miller, V. A., Pan, Q., Ladanyi, M., Zakowski, M. F., Heelan, R. T., Kris, M. G., and Varmus, H. E. (2005). KRAS mutations and primary resistance of lung adenocarcinomas to gefitinib or erlotinib. *PLoS Med* **2**, e17.

Parkin, D. M., Bray, F., Ferlay, J., and Pisani, P. (2005). Global cancer statistics, 2002. *CA Cancer J Clin* **55**, 74-108.

Parsons, D. W., Wang, T. L., Samuels, Y., Bardelli, A., Cummins, J. M., DeLong, L., Silliman, N., Ptak, J., Szabo, S., Willson, J. K., *et al.* (2005). Colorectal cancer: mutations in a signalling pathway. *Nature* **436**, 792.

Paterlini-Brechot, P., Saigo, K., Murakami, Y., Chami, M., Gozuacik, D., Mugnier, C., Lagorce, D., and Brechot, C. (2003). Hepatitis B virus-related insertional mutagenesis occurs frequently in human liver cancers and recurrently targets human telomerase gene. *Oncogene* **22**, 3911-3916.

Pavlopoulos, A., Oehler, S., Kapetanaki, M. G., and Savakis, C. (2007). The DNA transposon Minos as a tool for transgenesis and functional genomic analysis in vertebrates and invertebrates. *Genome Biol* **8 Suppl 1**, S2.

Peck, S. R., and Ruley, H. E. (2000). Ly108: a new member of the mouse CD2 family of cell surface proteins. *Immunogenetics* **52**, 63-72.

Pelengaris, S., and Khan, M. (2006). *The Molecular Biology of Cancer*, 2 edn: Blackwell Publishing Ltd).

Pelengaris, S., Khan, M., and Evan, G. I. (2002). Suppression of Myc-induced apoptosis in beta cells exposes multiple oncogenic properties of Myc and triggers carcinogenic progression. *Cell* **109**, 321-334.

Peto, J. (2001). Cancer epidemiology in the last century and the next decade. *Nature* **411**, 390-395.

Petrocca, F., Iliopoulos, D., Qin, H. R., Nicoloso, M. S., Yendamuri, S., Wojcik, S. E., Shimizu, M., Di Leva, G., Vecchione, A., Trapasso, F., *et al.* (2006). Alterations of the tumor suppressor gene ARLTS1 in ovarian cancer. *Cancer Res* **66**, 10287-10291.

Picard, F., Robin, S., Lavielle, M., Vaisse, C., and Daudin, J. J. (2005). A statistical approach for array CGH data analysis. *BMC Bioinformatics* **6**, 27.

Pinkel, D., and Albertson, D. G. (2005). Comparative genomic hybridization. *Annu Rev Genomics Hum Genet* **6**, 331-354.

Pinkel, D., Segraves, R., Sudar, D., Clark, S., Poole, I., Kowbel, D., Collins, C., Kuo, W. L., Chen, C., Zhai, Y., *et al.* (1998). High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays. *Nat Genet* **20**, 207-211.

Pollack, J. R., Perou, C. M., Alizadeh, A. A., Eisen, M. B., Pergamenschikov, A., Williams, C. F., Jeffrey, S. S., Botstein, D., and Brown, P. O. (1999). Genome-wide analysis of DNA copy-number changes using cDNA microarrays. *Nat Genet* **23**, 41-46.

Pollack, J. R., Sorlie, T., Perou, C. M., Rees, C. A., Jeffrey, S. S., Lonning, P. E., Tibshirani, R., Botstein, D., Borresen-Dale, A. L., and Brown, P. O. (2002). Microarray analysis reveals a major direct role of DNA copy number alteration in the transcriptional program of human breast tumors. *Proc Natl Acad Sci U S A* **99**, 12963-12968.

Powell, J. A., Gardner, A. E., Bais, A. J., Hinze, S. J., Baker, E., Whitmore, S., Crawford, J., Kochetkova, M., Spendlove, H. E., Doggett, N. A*., et al.* (2002). Sequencing, transcript identification, and quantitative gene expression profiling in the breast cancer loss of heterozygosity region 16q24.3 reveal three potential tumor-suppressor genes. *Genomics* **80**, 303-310.

Pruitt, K. D., Tatusova, T., and Maglott, D. R. (2007). NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res* **35**, D61-65.

Pryciak, P. M., and Varmus, H. E. (1992). Nucleosomes, DNA-binding proteins, and DNA sequence modulate retroviral integration target site selection. *Cell* **69**, 769-780.

Qi, H., Grenier, J., Fournier, A., and Labrie, C. (2003). Androgens differentially regulate the expression of NEDD4L transcripts in LNCaP human prostate cancer cells. *Mol Cell Endocrinol* **210**, 51-62.

Qin, Y. R., Fu, L., Sham, P. C., Kwong, D. L., Zhu, C. L., Chu, K. K., Li, Y., and Guan, X. Y. (2008). Single-nucleotide polymorphism-mass array reveals commonly deleted regions at 3p22 and 3p14.2 associate with poor clinical outcome in esophageal squamous cell carcinoma. *Int J Cancer* **123**, 826-830.

Qiu, W., David, D., Zhou, B., Chu, P. G., Zhang, B., Wu, M., Xiao, J., Han, T., Zhu, Z., Wang, T*., et al.* (2003). Down-regulation of growth arrest DNA damage-inducible gene 45beta expression is associated with human hepatocellular carcinoma. *Am J Pathol* **162**, 1961-1974.

Quackenbush, J. (2002). Microarray data normalization and transformation. *Nat Genet* **32 Suppl**, 496-501.

Quelle, D. E., Zindy, F., Ashmun, R. A., and Sherr, C. J. (1995). Alternative reading frames of the INK4a tumor suppressor gene encode two unrelated proteins capable of inducing cell cycle arrest. *Cell* **83**, 993-1000.

Radtke, F., and Raj, K. (2003). The role of Notch in tumorigenesis: oncogene or tumour suppressor? *Nat Rev Cancer* **3**, 756-767.

Radtke, F., Wilson, A., Stark, G., Bauer, M., van Meerwijk, J., MacDonald, H. R., and Aguet, M. (1999). Deficient T cell fate specification in mice with an induced inactivation of Notch1. *Immunity* **10**, 547-558.

Raghavan, M., Lillington, D. M., Skoulakis, S., Debernardi, S., Chaplin, T., Foot, N. J., Lister, T. A., and Young, B. D. (2005). Genome-wide single nucleotide polymorphism analysis reveals frequent partial uniparental disomy due to somatic recombination in acute myeloid leukemias. *Cancer Res* **65**, 375-378.

Raphael, B. J., Volik, S., Yu, P., Wu, C., Huang, G., Linardopoulou, E. V., Trask, B. J., Waldman, F., Costello, J., Pienta, K. J*., et al.* (2008). A sequence-based survey of the complex structural organization of tumor genomes. *Genome Biol* **9**, R59.

Rathinam, C., and Klein, C. (2007). Transcriptional repressor Gfi1 integrates cytokine-receptor signals controlling B-cell differentiation. *PLoS ONE* **2**, e306.

Raynaud, S., Cave, H., Baens, M., Bastard, C., Cacheux, V., Grosgeorge, J., Guidal-Giroux, C., Guo, C., Vilmer, E., Marynen, P., and Grandchamp, B. (1996). The 12;21 translocation involving TEL and deletion of the other TEL allele: two frequently associated alterations found in childhood acute lymphoblastic leukemia. *Blood* **87**, 2891-2899.

Redon, R., Hussenet, T., Bour, G., Caulee, K., Jost, B., Muller, D., Abecassis, J., and du Manoir, S. (2002). Amplicon mapping and transcriptional analysis pinpoint cyclin L as a candidate oncogene in head and neck cancer. *Cancer Res* **62**, 6211-6217.

Redon, R., Ishikawa, S., Fitch, K. R., Feuk, L., Perry, G. H., Andrews, T. D., Fiegler, H., Shapero, M. H., Carson, A. R., Chen, W*., et al.* (2006). Global variation in copy number in the human genome. *Nature* **444**, 444-454.

Redon, S., Reichenbach, P., and Lingner, J. (2007). Protein RNA and protein protein interactions mediate association of human EST1A/SMG6 with telomerase. *Nucleic Acids Res* **35**, 7011-7022.

Reed, J. A., Lin, Q., Chen, D., Mian, I. S., and Medrano, E. E. (2005). SKI pathways inducing progression of human melanoma. *Cancer Metastasis Rev* **24**, 265-272.

Rehwinkel, J., Letunic, I., Raes, J., Bork, P., and Izaurralde, E. (2005). Nonsense-mediated mRNA decay factors act in concert to regulate common mRNA targets. *Rna* **11**, 1530-1544.

Reimand, J., Kull, M., Peterson, H., Hansen, J., and Vilo, J. (2007). g:Profiler--a web-based toolset for functional profiling of gene lists from large-scale experiments. *Nucleic Acids Res* **35**, W193-200.

Ren, B., Robert, F., Wyrick, J. J., Aparicio, O., Jennings, E. G., Simon, I., Zeitlinger, J., Schreiber, J., Hannett, N., Kanin, E*., et al.* (2000). Genome-wide location and function of DNA binding proteins. *Science* **290**, 2306-2309.

Resnick, M. B., Konkin, T., Routhier, J., Sabo, E., and Pricolo, V. E. (2005). Claudin-1 is a strong prognostic indicator in stage II colonic cancer: a tissue microarray study. *Mod Pathol* **18**, 511-518.

Rikiyama, T., Curtis, J., Oikawa, M., Zimonjic, D. B., Popescu, N., Murphy, B. A., Wilson, M. A., and Johnson, A. C. (2003). GCF2: expression and molecular analysis of repression. *Biochim Biophys Acta* **1629**, 15-25.

Riley, J., Butler, R., Ogilvie, D., Finniear, R., Jenner, D., Powell, S., Anand, R., Smith, J. C., and Markham, A. F. (1990). A novel, rapid method for the isolation of terminal sequences from yeast artificial chromosome (YAC) clones. *Nucleic Acids Res* **18**, 2887-2890.

Robertson, E., Bradley, A., Kuehn, M., and Evans, M. (1986). Germ-line transmission of genes introduced into cultured pluripotential cells by retroviral vector. *Nature* **323**, 445-448.

Robertson, G., Hirst, M., Bainbridge, M., Bilenky, M., Zhao, Y., Zeng, T., Euskirchen, G., Bernier, B., Varhol, R., Delaney, A.*, et al.* (2007). Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. *Nat Methods* **4**, 651-657.

Rodriguez, S., Jafer, O., Goker, H., Summersgill, B. M., Zafarana, G., Gillis, A. J., van Gurp, R. J., Oosterhuis, J. W., Lu, Y. J., Huddart, R.*, et al.* (2003). Expression profile of genes from 12p in testicular germ cell tumors of adolescents and adults associated with i(12p) and amplification at 12p11.2-p12.1. *Oncogene* **22**, 1880-1891.

Roh, T. Y., Cuddapah, S., Cui, K., and Zhao, K. (2006). The genomic landscape of histone modifications in human T cells. *Proc Natl Acad Sci U S A* **103**, 15782-15787.

Roh, T. Y., Cuddapah, S., and Zhao, K. (2005). Active chromatin domains are defined by acetylation islands revealed by genome-wide mapping. *Genes Dev* **19**, 542-552.

Roose, J. P., Mollenauer, M., Ho, M., Kurosaki, T., and Weiss, A. (2007). Unusual interplay of two types of Ras activators, RasGRP and SOS, establishes sensitive and robust Ras activation in lymphocytes. *Mol Cell Biol* **27**, 2732-2745.

Rosson, D., Dugan, D., and Reddy, E. P. (1987). Aberrant splicing events that are induced by proviral integration: implications for myb oncogene activation. *Proc Natl Acad Sci U S A* **84**, 3171-3175.

Rouveirol, C., Stransky, N., Hupe, P., Rosa, P. L., Viara, E., Barillot, E., and Radvanyi, F. (2006). Computation of recurrent minimal genomic alterations from array-CGH data. *Bioinformatics* **22**, 849-856.

Rueda, O. M., and Diaz-Uriarte, R. (2007). Flexible and accurate detection of genomic copy-number changes from aCGH. *PLoS Comput Biol* **3**, e122.

Russo, L. A., Calabro, S. P., Filler, T. A., Carey, D. J., and Gardner, R. M. (2001). In vivo regulation of syndecan-3 expression in the rat uterus by 17 beta-estradiol. *J Biol Chem* **276**, 686-692.

Salmeron, A., Ahmad, T. B., Carlile, G. W., Pappin, D., Narsimhan, R. P., and Ley, S. C. (1996). Activation of MEK-1 and SEK-1 by Tpl-2 proto-oncoprotein, a novel MAP kinase kinase kinase. *Embo J* **15**, 817-826.

Salomon-Nguyen, F., Della-Valle, V., Mauchauffe, M., Busson-Le Coniat, M., Ghysdael, J., Berger, R., and Bernard, O. A. (2000). The t(1;12)(q21;p13) translocation of human acute myeloblastic leukemia results in a TEL-ARNT fusion. *Proc Natl Acad Sci U S A* **97**, 6757-6762.

Samuels, Y., and Velculescu, V. E. (2004). Oncogenic mutations of PIK3CA in human cancers. *Cell Cycle* **3**, 1221-1224.

Sauer, B., and Henderson, N. (1988). Site-specific DNA recombination in mammalian cells by the Cre recombinase of bacteriophage P1. *Proc Natl Acad Sci U S A* **85**, 5166-5170.

Sawyers, C. L., Hochhaus, A., Feldman, E., Goldman, J. M., Miller, C. B., Ottmann, O. G., Schiffer, C. A., Talpaz, M., Guilhot, F., Deininger, M. W.*, et al.* (2002). Imatinib induces hematologic and cytogenetic responses in patients with chronic myelogenous leukemia in myeloid blast crisis: results of a phase II study. *Blood* **99**, 3530-3539.

Schlicht, M., Matysiak, B., Brodzeller, T., Wen, X., Liu, H., Zhou, G., Dhir, R., Hessner, M. J., Tonellato, P., Suckow, M.*, et al.* (2004). Cross-species global and subset gene expression profiling identifies genes involved in prostate cancer response to selenium. *BMC Genomics* **5**, 58.

Schulte, J. H., Horn, S., Otto, T., Samans, B., Heukamp, L. C., Eilers, U. C., Krause, M., Astrahantseff, K., Klein-Hitpass, L., Buettner, R.*, et al.* (2008). MYCN regulates oncogenic MicroRNAs in neuroblastoma. *Int J Cancer* **122**, 699-704.

Schultz, D. C., Vanderveer, L., Berman, D. B., Hamilton, T. C., Wong, A. J., and Godwin, A. K. (1996). Identification of two candidate tumor suppressor genes on chromosome 17p13.3. *Cancer Res* **56**, 1997-2002.

Schwab, M. (1999). Oncogene amplification in solid tumors. *Semin Cancer Biol* **9**, 319-325.

Scott, R. W., and Olson, M. F. (2007). LIM kinases: function, regulation and association with human disease. *J Mol Med* **85**, 555-568.

Seeger, R. C., Brodeur, G. M., Sather, H., Dalton, A., Siegel, S. E., Wong, K. Y., and Hammond, D. (1985). Association of multiple copies of the N-myc oncogene with rapid progression of neuroblastomas. *N Engl J Med* **313**, 1111-1116.

Seimiya, M., J, O. W., Bahar, R., Kawamura, K., Wang, Y., Saisho, H., and Tagawa, M. (2003). Stage-specific expression of Clast6/E3/LAPTM5 during B cell differentiation: elevated expression in human B lymphomas. *Int J Oncol* **22**, 301-304.

Selten, G., Cuypers, H. T., and Berns, A. (1985). Proviral activation of the putative oncogene Pim-1 in MuLV induced T-cell lymphomas. *Embo J* **4**, 1793-1798.

Selten, G., Cuypers, H. T., Zijlstra, M., Melief, C., and Berns, A. (1984). Involvement of c-myc in MuLV-induced T cell lymphomas in mice: frequency and mechanisms of activation. *Embo J* **3**, 3215-3222.

Semenza, G. L. (2002). Involvement of hypoxia-inducible factor 1 in human cancer. *Intern Med* **41**, 79-83.

Serrano, M., Hannon, G. J., and Beach, D. (1993). A new regulatory motif in cell-cycle control causing specific inhibition of cyclin D/CDK4. *Nature* **366**, 704-707.

Seymour, R., Sundberg, J. P., and Hogenesch, H. (2006). Abnormal lymphoid organ development in immunodeficient mutant mice. *Vet Pathol* **43**, 401-423.

Shah, S. P., Xuan, X., DeLeeuw, R. J., Khojasteh, M., Lam, W. L., Ng, R., and Murphy, K. P. (2006). Integrating copy number polymorphisms into array CGH analysis using a robust HMM. *Bioinformatics* **22**, e431-439.

Shapiro, S. (1997). Periodic screening for breast cancer: the HIP Randomized Controlled Trial. Health Insurance Plan. *J Natl Cancer Inst Monogr*, 27-30.

Sharma, M., Li, X., Wang, Y., Zarnegar, M., Huang, C. Y., Palvimo, J. J., Lim, B., and Sun, Z. (2003). hZimp10 is an androgen receptor co-activator and forms a complex with SUMO-1 at replication foci. *Embo J* **22**, 6101-6114.

Sharma, V. M., Draheim, K. M., and Kelliher, M. A. (2007). The Notch1/c-Myc pathway in T cell leukemia. *Cell Cycle* **6**, 927-930.

Sharpless, N. E., and Depinho, R. A. (2006). The mighty mouse: genetically engineered mouse models in cancer drug development. *Nat Rev Drug Discov* **5**, 741-754.

Shattuck, D. L., Miller, J. K., Carraway, K. L., 3rd, and Sweeney, C. (2008). Met receptor contributes to trastuzumab resistance of Her2-overexpressing breast cancer cells. *Cancer Res* **68**, 1471-1477.

Shen, J. L., Yan, C. H., Liu, Y., Yan, X. Q., Zhang, X. L., Jin, Y., Zhang, K. F., Sang, Z. F., Zhang, G. Y., Li, P., and Fu, S. B. (2003). [Studies of TGF-beta/Smads expression in lung cancer]. *Yi Chuan Xue Bao* **30**, 681-686.

Sherr, C. J. (2001). The INK4a/ARF network in tumour suppression. *Nat Rev Mol Cell Biol* **2**, 731-737.

Shi, M., Cooper, J. C., and Yu, C. L. (2006). A constitutively active Lck kinase promotes cell proliferation and resistance to apoptosis through signal transducer and activator of transcription 5b activation. *Mol Cancer Res* **4**, 39-45.

Shimizu, T., Mundt, C., Licence, S., Melchers, F., and Martensson, I. L. (2002). VpreB1/VpreB2/lambda 5 triple-deficient mice show impaired B cell development but functional allelic exclusion of the IgH locus. *J Immunol* **168**, 6286-6293.

Shin, S. S., Namkoong, J., Wall, B. A., Gleason, R., Lee, H. J., and Chen, S. (2008). Oncogenic activities of metabotropic glutamate receptor 1 (Grm1) in melanocyte transformation. *Pigment Cell Melanoma Res* **21**, 368-378.

Shinagawa, T., Nomura, T., Colmenares, C., Ohira, M., Nakagawara, A., and Ishii, S. (2001). Increased susceptibility to tumorigenesis of ski-deficient heterozygous mice. *Oncogene* **20**, 8100-8108.

Shipitsin, M., and Polyak, K. (2008). The cancer stem cell hypothesis: in search of definitions, markers, and relevance. *Lab Invest* **88**, 459-463.

Sinclair, C. S., Rowley, M., Naderi, A., and Couch, F. J. (2003). The 17q23 amplicon and breast cancer. *Breast Cancer Res Treat* **78**, 313-322.

Singh, A. P., Bafna, S., Chaudhary, K., Venkatraman, G., Smith, L., Eudy, J. D., Johansson, S. L., Lin, M. F., and Batra, S. K. (2008). Genome-wide expression profiling reveals transcriptomic variation and perturbed gene networks in androgen-dependent and androgen-independent prostate cancer cells. *Cancer Lett* **259**, 28-38.

Sjoblom, T., Jones, S., Wood, L. D., Parsons, D. W., Lin, J., Barber, T. D., Mandelker, D., Leary, R. J., Ptak, J., Silliman, N.*, et al.* (2006). The consensus coding sequences of human breast and colorectal cancers. *Science* **314**, 268-274.

Slape, C., Hartung, H., Lin, Y. W., Bies, J., Wolff, L., and Aplan, P. D. (2007). Retroviral insertional mutagenesis identifies genes that collaborate with NUP98-HOXD13 during leukemic transformation. *Cancer Res* **67**, 5148-5155.

Small, D. (2006). FLT3 mutations: biology and treatment. *Hematology Am Soc Hematol Educ Program*, 178-184.

Smirnova, L., Grafe, A., Seiler, A., Schumacher, S., Nitsch, R., and Wulczyn, F. G. (2005). Regulation of miRNA expression during neural cell specification. *Eur J Neurosci* **21**, 1469-1477.

Smit, A. F. A., Hubley, R., and Green, P. (1996-2004). RepeatMasker Open-3.0. *http://wwwrepeatmaskerorg*.

Smith, A. J., Xian, J., Richardson, M., Johnstone, K. A., and Rabbitts, P. H. (2002). Cre-loxP chromosome engineering of a targeted deletion in the mouse corresponding to the 3p21.3 region of homozygous loss in human tumours. *Oncogene* **21**, 4521-4529.

Smith, T. F., and Waterman, M. S. (1981). Identification of common molecular subsequences. *J Mol Biol* **147**, 195-197.

Snow, B. E., Erdmann, N., Cruickshank, J., Goldman, H., Gill, R. M., Robinson, M. O., and Harrington, L. (2003). Functional conservation of the telomerase protein Est1p in humans. *Curr Biol* **13**, 698-704.

Soda, M., Choi, Y. L., Enomoto, M., Takada, S., Yamashita, Y., Ishikawa, S., Fujiwara, S., Watanabe, H., Kurashina, K., Hatanaka, H.*, et al.* (2007). Identification of the transforming EML4-ALK fusion gene in non-small-cell lung cancer. *Nature* **448**, 561-566.

Soignet, S., and Maslak, P. (2004). Therapy of acute promyelocytic leukemia. *Adv Pharmacol* **51**, 35-58.

Soler, G., Radford-Weiss, I., Ben-Abdelali, R., Mahlaoui, N., Ponceau, J. F., Macintyre, E. A., Vekemans, M., Bernard, O. A., and Romana, S. P. (2008). Fusion of ZMIZ1 to ABL1 in a B-cell acute lymphoblastic leukaemia with a t(9;10)(q34;q22.3) translocation. *Leukemia* **22**, 1278-1280.

Somura, H., Iizuka, N., Tamesa, T., Sakamoto, K., Hamaguchi, T., Tsunedomi, R., Yamada-Okabe, H., Sawamura, M., Eramoto, M., Miyamoto, T.*, et al.* (2008). A three-gene predictor for early intrahepatic recurrence of hepatocellular carcinoma after curative hepatectomy. *Oncol Rep* **19**, 489-495.

Soto, A. M., and Sonnenschein, C. (2004). The somatic mutation theory of cancer: growing problems with the paradigm? *Bioessays* **26**, 1097-1107.

Sourvinos, G., Tsatsanis, C., and Spandidos, D. A. (1999). Overexpression of the Tpl-2/Cot oncogene in human breast cancer. *Oncogene* **18**, 4968-4973.

Staaf, J., Jonsson, G., Ringner, M., and Vallon-Christersson, J. (2007). Normalization of array-CGH data: influence of copy number imbalances. *BMC Genomics* **8**, 382.

Stephens, P., Edkins, S., Davies, H., Greenman, C., Cox, C., Hunter, C., Bignell, G., Teague, J., Smith, R., Stevens, C*., et al.* (2005). A screen of the complete protein kinase gene family identifies diverse patterns of somatic mutations in human breast cancer. *Nat Genet* **37**, 590-592.

Stewart, M., Mackay, N., Hanlon, L., Blyth, K., Scobie, L., Cameron, E., and Neil, J. C. (2007). Insertional mutagenesis reveals progression genes and checkpoints in MYC/Runx2 lymphomas. *Cancer Res* **67**, 5126-5133.

Stewart, S. A. (2005). Telomere maintenance and tumorigenesis: an "ALT"ernative road. *Curr Mol Med* **5**, 253-257.

Stewart, T. A., Pattengale, P. K., and Leder, P. (1984). Spontaneous mammary adenocarcinomas in transgenic mice that carry and express MTV/myc fusion genes. *Cell* **38**, 627-637.

Stjernqvist, S., Ryden, T., Skold, M., and Staaf, J. (2007). Continuous-index hidden Markov modelling of array CGH copy number data. *Bioinformatics* **23**, 1006-1014.

Storey, J. D., and Tibshirani, R. (2003). Statistical significance for genomewide studies. *Proc Natl Acad Sci U S A* **100**, 9440-9445.

Stranger, B. E., Forrest, M. S., Dunning, M., Ingle, C. E., Beazley, C., Thorne, N., Redon, R., Bird, C. P., de Grassi, A., Lee, C*., et al.* (2007). Relative impact of nucleotide and copy number variation on gene expression phenotypes. *Science* **315**, 848-853.

Subramaniam, D., Natarajan, G., Ramalingam, S., Ramachandran, I., May, R., Queimado, L., Houchen, C. W., and Anant, S. (2008). Translation inhibition during cell cycle arrest and apoptosis: Mcl-1 is a novel target for RNA binding protein CUGBP2. *Am J Physiol Gastrointest Liver Physiol* **294**, G1025-1032.

Sujobert, P., Bardet, V., Cornillet-Lefebvre, P., Hayflick, J. S., Prie, N., Verdier, F., Vanhaesebroeck, B., Muller, O., Pesce, F., Ifrah, N*., et al.* (2005). Essential role for the p110delta isoform in phosphoinositide 3-kinase activation and cell proliferation in acute myeloid leukemia. *Blood* **106**, 1063-1066.

Sun, X. F., and Zhang, H. (2007). NFKB and NFKBI polymorphisms in relation to susceptibility of tumour and other diseases. *Histol Histopathol* **22**, 1387-1398.

Suriano, A. R., Sanford, A. N., Kim, N., Oh, M., Kennedy, S., Henderson, M. J., Dietzmann, K., and Sullivan, K. E. (2005). GCF2/LRRFIP1 represses tumor necrosis factor alpha expression. *Mol Cell Biol* **25**, 9073-9081.

Suzuki, E., Handa, K., Toledo, M. S., and Hakomori, S. (2004). Sphingosine-dependent apoptosis: a unified concept based on multiple mechanisms operating in concert. *Proc Natl Acad Sci U S A* **101**, 14788-14793.

Suzuki, T., Minehata, K., Akagi, K., Jenkins, N. A., and Copeland, N. G. (2006). Tumor suppressor gene identification using retroviral insertional mutagenesis in Blm-deficient mice. *Embo J* **25**, 3422-3431.

Suzuki, T., Shen, H., Akagi, K., Morse, H. C., Malley, J. D., Naiman, D. Q., Jenkins, N. A., and Copeland, N. G. (2002). New genes involved in cancer identified by retroviral tagging. *Nat Genet* **32**, 166-174.

Sweet-Cordero, A., Mukherjee, S., Subramanian, A., You, H., Roix, J. J., Ladd-Acosta, C., Mesirov, J., Golub, T. R., and Jacks, T. (2005). An oncogenic KRAS2 expression signature identified by cross-species gene-expression analysis. *Nat Genet* **37**, 48-55.

Takahashi, K., Kohno, T., Ajima, R., Sasaki, H., Minna, J. D., Fujiwara, T., Tanaka, N., and Yokota, J. (2006). Homozygous deletion and reduced expression of the DOCK8 gene in human lung cancer. *Int J Oncol* **28**, 321-328.

Takakura, S., Mitsutake, N., Nakashima, M., Namba, H., Saenko, V. A., Rogounovitch, T. I., Nakazawa, Y., Hayashi, T., Ohtsuru, A., and Yamashita, S. (2008). Oncogenic role of miR-17-92 cluster in anaplastic thyroid cancer cells. *Cancer Sci* **99**, 1147-1154.

Tang, P. P., and Wang, F. F. (2000). Induction of IW32 erythroleukemia cell differentiation by p53 is dependent on protein tyrosine phosphatase. *Leukemia* **14**, 1292-1300.

Theodorou, V., Kimm, M. A., Boer, M., Wessels, L., Theelen, W., Jonkers, J., and Hilkens, J. (2007). MMTV insertional mutagenesis identifies genes, gene families and pathways involved in mammary cancer. *Nat Genet* **39**, 759-769.

Thiagalingam, A., Lengauer, C., Baylin, S. B., and Nelkin, B. D. (1997). RREB1, a ras responsive element binding protein, maps to human chromosome 6p25. *Genomics* **45**, 630-632.

Thomas, R. K., Weir, B., and Meyerson, M. (2006). Genomic approaches to lung cancer. *Clin Cancer Res* **12**, 4384s-4391s.

Thorne, J., and Campbell, M. J. (2008). The vitamin D receptor in cancer. *Proc Nutr Soc* **67**, 115-127.

Tichelaar, J. W., Lu, W., and Whitsett, J. A. (2000). Conditional expression of fibroblast growth factor-7 in the developing and mature lung. *J Biol Chem* **275**, 11858-11864.

Tomlins, S. A., and Chinnaiyan, A. M. (2006). Of mice and men: cancer gene discovery using comparative oncogenomics. *Cancer Cell* **10**, 2-4.

Tomlins, S. A., Rhodes, D. R., Perner, S., Dhanasekaran, S. M., Mehra, R., Sun, X. W., Varambally, S., Cao, X., Tchinda, J., Kuefer, R*., et al.* (2005). Recurrent fusion of TMPRSS2 and ETS transcription factor genes in prostate cancer. *Science* **310**, 644-648.

Touw, I. P., and Erkeland, S. J. (2007). Retroviral insertion mutagenesis in mice as a comparative oncogenomics tool to identify disease genes in human leukemia. *Mol Ther* **15**, 13-19.

Toyota, M., Ahuja, N., Ohe-Toyota, M., Herman, J. G., Baylin, S. B., and Issa, J. P. (1999). CpG island methylator phenotype in colorectal cancer. *Proc Natl Acad Sci U S A* **96**, 8681-8686.

Triglia, T., Peterson, M. G., and Kemp, D. J. (1988). A procedure for in vitro amplification of DNA segments that lie outside the boundaries of known sequences. *Nucleic Acids Res* **16**, 8186.

Trotman, L. C., Niki, M., Dotan, Z. A., Koutcher, J. A., Di Cristofano, A., Xiao, A., Khoo, A. S., Roy-Burman, P., Greenberg, N. M., Van Dyke, T.*, et al.* (2003). Pten dose dictates cancer progression in the prostate. *PLoS Biol* **1**, E59.

Tsai, J., Lee, J. T., Wang, W., Zhang, J., Cho, H., Mamo, S., Bremer, R., Gillette, S., Kong, J., Haass, N. K.*, et al.* (2008). Discovery of a selective inhibitor of oncogenic B-Raf kinase with potent antimelanoma activity. *Proc Natl Acad Sci U S A* **105**, 3041-3046.

Tse, W., Meshinchi, S., Alonzo, T. A., Stirewalt, D. L., Gerbing, R. B., Woods, W. G., Appelbaum, F. R., and Radich, J. P. (2004). Elevated expression of the AF1q gene, an MLL fusion partner, is an independent adverse prognostic factor in pediatric acute myeloid leukemia. *Blood* **104**, 3058-3063.

Tse, W., Zhu, W., Chen, H. S., and Cohen, A. (1995). A novel gene, AF1q, fused to MLL in t(1;11) (q21;q23), is specifically expressed in leukemic and immature hematopoietic cells. *Blood* **85**, 650-656.

Tsugane, S. (2005). Salt, salted food intake, and risk of gastric cancer: epidemiologic evidence. *Cancer Sci* **96**, 1-6.

Tsuji, H., Ishii-Ohba, H., Ukai, H., Katsube, T., and Ogiu, T. (2003). Radiation-induced deletions in the 5' end region of Notch1 lead to the formation of truncated proteins and are involved in the development of mouse thymic lymphomas. *Carcinogenesis* **24**, 1257-1268.

Tsukasaki, K., Miller, C. W., Greenspun, E., Eshaghian, S., Kawabata, H., Fujimoto, T., Tomonaga, M., Sawyers, C., Said, J. W., and Koeffler, H. P. (2001). Mutations in the mitotic check point gene, MAD1L1, in human cancers. *Oncogene* **20**, 3301-3305.

Turley, R. S., Finger, E. C., Hempel, N., How, T., Fields, T. A., and Blobe, G. C. (2007). The type III transforming growth factor-beta receptor as a novel tumor suppressor gene in prostate cancer. *Cancer Res* **67**, 1090-1098.

Tzivion, G., Gupta, V. S., Kaplun, L., and Balan, V. (2006). 14-3-3 proteins as potential oncogenes. *Semin Cancer Biol* **16**, 203-213.

Unezaki, S., Horai, R., Sudo, K., Iwakura, Y., and Ito, S. (2007). Ovol2/Movo, a homologue of Drosophila ovo, is required for angiogenesis, heart formation and placental development in mice. *Genes Cells* **12**, 773-785.

Unezaki, S., Nishizawa, M., Okuda-Ashitaka, E., Masu, Y., Mukai, M., Kobayashi, S., Sawamoto, K., Okano, H., and Ito, S. (2004). Characterization of the isoforms of MOVO zinc finger protein, a mouse homologue of Drosophila Ovo, as transcription factors. *Gene* **336**, 47-58.

Uren, A. G., Kool, J., Berns, A., and van Lohuizen, M. (2005). Retroviral insertional mutagenesis: past, present and future. *Oncogene* **24**, 7656-7672.

Uren, A. G., Kool, J., Matentzoglu, K., de Ridder, J., Mattison, J., van Uitert, M., Lagcher, W., Sie, D., Tanger, E., Cox, T.*, et al.* (2008). Large-scale mutagenesis in p19(ARF)- and p53-deficient mice identifies cancer genes and their collaborative networks. *Cell* **133**, 727-741.

Urzua, U., Frankenberger, C., Gangi, L., Mayer, S., Burkett, S., and Munroe, D. J. (2005). Microarray comparative genomic hybridization profile of a murine model for epithelial ovarian cancer reveals genomic imbalances resembling human ovarian carcinomas. *Tumour Biol* **26**, 236-244.

Vaarala, M. H., Porvari, K., Kyllonen, A., Lukkarinen, O., and Vihko, P. (2001). The TMPRSS2 gene encoding transmembrane serine protease is overexpressed in a majority of prostate cancer patients: detection of mutated TMPRSS2 form in a case of aggressive disease. *Int J Cancer* **94**, 705-710.

van der Lugt, N. M., Domen, J., Verhoeven, E., Linders, K., van der Gulden, H., Allen, J., and Berns, A. (1995). Proviral tagging in E mu-myc transgenic mice lacking the Pim-1 proto-oncogene leads to compensatory activation of Pim-2. *Embo J* **14**, 2536-2544.

van Lohuizen, M., Verbeek, S., Krimpenfort, P., Domen, J., Saris, C., Radaszkiewicz, T., and Berns, A. (1989). Predisposition to lymphomagenesis in pim-1 transgenic mice: cooperation with c-myc and N-myc in murine leukemia virus-induced tumors. *Cell* **56**, 673-682.

van Lohuizen, M., Verbeek, S., Scheijen, B., Wientjens, E., van der Gulden, H., and Berns, A. (1991). Identification of cooperating oncogenes in E mu-myc transgenic mice by provirus tagging. *Cell* **65**, 737-752.

van Oosterom, A. T., Judson, I., Verweij, J., Stroobants, S., Donato di Paola, E., Dimitrijevic, S., Martens, M., Webb, A., Sciot, R., Van Glabbeke, M.*, et al.* (2001). Safety and efficacy of imatinib (STI571) in metastatic gastrointestinal stromal tumours: a phase I study. *Lancet* **358**, 1421-1423.

Vandepoele, K., Andries, V., Van Roy, N., Staes, K., Vandesompele, J., Laureys, G., De Smet, E., Berx, G., Speleman, F., and van Roy, F. (2008). A constitutional translocation t(1;17)(p36.2;q11.2) in a neuroblastoma patient disrupts the human NBPF1 and ACCN1 genes. *PLoS ONE* **3**, e2207.

Varis, A., Wolf, M., Monni, O., Vakkari, M. L., Kokkola, A., Moskaluk, C., Frierson, H., Jr., Powell, S. M., Knuutila, S., Kallioniemi, A., and El-Rifai, W. (2002). Targets of gene amplification and overexpression at 17q in gastric cancer. *Cancer Res* **62**, 2625-2629.

Veltman, J. A., Fridlyand, J., Pejavar, S., Olshen, A. B., Korkola, J. E., DeVries, S., Carroll, P., Kuo, W. L., Pinkel, D., Albertson, D.*, et al.* (2003). Array-based comparative genomic hybridization for genome-wide screening of DNA copy number in bladder tumors. *Cancer Res* **63**, 2872-2880.

Venkatraman, E. S., and Olshen, A. B. (2007). A faster circular binary segmentation algorithm for the analysis of array CGH data. *Bioinformatics* **23**, 657-663.

Vigdal, T. J., Kaufman, C. D., Izsvak, Z., Voytas, D. F., and Ivics, Z. (2002). Common physical properties of DNA affecting target site selection of sleeping beauty and other Tc1/mariner transposable elements. *J Mol Biol* **323**, 441-452.

Vila-Carriles, W. H., Kovacs, G. G., Jovov, B., Zhou, Z. H., Pahwa, A. K., Colby, G., Esimai, O., Gillespie, G. Y., Mapstone, T. B., Markert, J. M*., et al.* (2006). Surface expression of ASIC2 inhibits the amiloride-sensitive current and migration of glioma cells. *J Biol Chem* **281**, 19220-19232.

Virtanen, C., Ishikawa, Y., Honjoh, D., Kimura, M., Shimane, M., Miyoshi, T., Nomura, H., and Jones, M. H. (2002). Integrated classification of lung tumors and cell lines by expression profiling. *Proc Natl Acad Sci U S A* **99**, 12357-12362.

Visan, I., Tan, J. B., Yuan, J. S., Harper, J. A., Koch, U., and Guidos, C. J. (2006). Regulation of T lymphopoiesis by Notch1 and Lunatic fringe-mediated competition for intrathymic niches. *Nat Immunol* **7**, 634-643.

Vogelstein, B., and Kinzler, K. W. (2004). Cancer genes and the pathways they control. *Nat Med* **10**, 789-799.

Vogelstein, B., Lane, D., and Levine, A. J. (2000). Surfing the p53 network. *Nature* **408**, 307-310.

Volik, S., Raphael, B. J., Huang, G., Stratton, M. R., Bignel, G., Murnane, J., Brebner, J. H., Bajsarowicz, K., Paris, P. L., Tao, Q*., et al.* (2006). Decoding the fine-scale structure of a breast cancer genome and transcriptome. *Genome Res* **16**, 394-404.

Volik, S., Zhao, S., Chin, K., Brebner, J. H., Herndon, D. R., Tao, Q., Kowbel, D., Huang, G., Lapuk, A., Kuo, W. L*., et al.* (2003). End-sequence profiling: sequence-based analysis of aberrant genomes. *Proc Natl Acad Sci U S A* **100**, 7696-7701.

von Hansemann, D. (1890). Uber assymetrische Zellteilung in Epithekrebsen und deren biologische Bedeutung. *Virchow's Arch Path Anat* **119**, 299-326.

Vrana, J. A., Bieszczad, C. K., Cleaveland, E. S., Ma, Y., Park, J. P., Mohandas, T. K., and Craig, R. W. (2002). An MCL1-overexpressing Burkitt lymphoma subline exhibits enhanced survival on exposure to serum deprivation, topoisomerase inhibitors, or staurosporine but remains sensitive to 1-beta-D-arabinofuranosylcytosine. *Cancer Res* **62**, 892-900.

Wajant, H. (2002). The Fas signaling pathway: more than a paradigm. *Science* **296**, 1635-1636.

Wandstrat, A. E., Nguyen, C., Limaye, N., Chan, A. Y., Subramanian, S., Tian, X. H., Yim, Y. S., Pertsemlidis, A., Garner, H. R., Jr., Morel, L., and Wakeland, E. K. (2004). Association of extensive polymorphisms in the SLAM/CD2 gene cluster with murine lupus. *Immunity* **21**, 769-780.

Wang, G., Williams, G., Xia, H., Hickey, M., Shao, J., Davidson, B. L., and McCray, P. B. (2002a). Apical barriers to airway epithelial cell gene transfer with amphotropic retroviral vectors. *Gene Ther* **9**, 922-931.

Wang, G. G., Pasillas, M. P., and Kamps, M. P. (2006a). Persistent transactivation by meis1 replaces hox function in myeloid leukemogenesis models: evidence for co-occupancy of meis1-pbx and hox-pbx complexes on promoters of leukemia-associated genes. *Mol Cell Biol* **26**, 3902-3916.

Wang, H., Huang, S., Shou, J., Su, E. W., Onyia, J. E., Liao, B., and Li, S. (2006b). Comparative analysis and integrative classification of NCI60 cell lines and primary tumors using gene expression profiling data. *BMC Genomics* **7**, 166.

Wang, P., Kim, Y., Pollack, J., Narasimhan, B., and Tibshirani, R. (2005). A method for calling gains and losses in array CGH data. *Biostatistics* **6**, 45-58.

Wang, X., Southard, R. C., Allred, C. D., Talbert, D. R., Wilson, M. E., and Kilgore, M. W. (2008). MAZ drives tumor-specific expression of PPAR gamma 1 in breast cancer cells. *Breast Cancer Res Treat* **111**, 103-111.

Wang, Y., and Armstrong, S. A. (2007). Genome-wide SNP analysis in cancer: leukemia shows the way. *Cancer Cell* **11**, 308-309.

Wang, Z., Bhattacharya, N., Mixter, P. F., Wei, W., Sedivy, J., and Magnuson, N. S. (2002b). Phosphorylation of the cell cycle inhibitor p21Cip1/WAF1 by Pim-1 kinase. *Biochim Biophys Acta* **1593**, 45-55.

Wang, Z., Shen, D., Parsons, D. W., Bardelli, A., Sager, J., Szabo, S., Ptak, J., Silliman, N., Peters, B. A., van der Heijden, M. S*., et al.* (2004). Mutational analysis of the tyrosine phosphatome in colorectal cancers. *Science* **304**, 1164-1166.

Wardrop, S. L., and Brown, M. A. (2005). Identification of two evolutionarily conserved and functional regulatory elements in intron 2 of the human BRCA1 gene. *Genomics* **86**, 316-328.

Waterston, R. H., Lindblad-Toh, K., Birney, E., Rogers, J., Abril, J. F., Agarwal, P., Agarwala, R., Ainscough, R., Alexandersson, M., An, P*., et al.* (2002). Initial sequencing and comparative analysis of the mouse genome. *Nature* **420**, 520-562.

Watson, S. K., deLeeuw, R. J., Horsman, D. E., Squire, J. A., and Lam, W. L. (2007). Cytogenetically balanced translocations are associated with focal copy number alterations. *Hum Genet* **120**, 795-805.

Wei, C. L., Wu, Q., Vega, V. B., Chiu, K. P., Ng, P., Zhang, T., Shahab, A., Yong, H. C., Fu, Y., Weng, Z*., et al.* (2006). A global map of p53 transcription-factor binding sites in the human genome. *Cell* **124**, 207-219.

Weinberg, R. A. (1990). The retinoblastoma gene and cell growth control. *Trends Biochem Sci* **15**, 199-202.

Weinmann, A. S., Bartley, S. M., Zhang, T., Zhang, M. Q., and Farnham, P. J. (2001). Use of chromatin immunoprecipitation to clone novel E2F target promoters. *Mol Cell Biol* **21**, 6820-6832.

Weinmann, A. S., Yan, P. S., Oberley, M. J., Huang, T. H., and Farnham, P. J. (2002). Isolating human transcription factor targets by coupling chromatin immunoprecipitation and CpG island microarray analysis. *Genes Dev* **16**, 235-244.

Weinstein, I. B. (2002). Cancer. Addiction to oncogenes--the Achilles heal of cancer. *Science* **297**, 63-64.

Weir, B. A., Woo, M. S., Getz, G., Perner, S., Ding, L., Beroukhim, R., Lin, W. M., Province, M. A., Kraja, A., Johnson, L. A.*, et al.* (2007). Characterizing the cancer genome in lung adenocarcinoma. *Nature* **450**, 893-898.

Weir, H. K., Thun, M. J., Hankey, B. F., Ries, L. A., Howe, H. L., Wingo, P. A., Jemal, A., Ward, E., Anderson, R. N., and Edwards, B. K. (2003). Annual report to the nation on the status of cancer, 1975-2000, featuring the uses of surveillance data for cancer prevention and control. *J Natl Cancer Inst* **95**, 1276-1299.

Weisenberger, D. J., Siegmund, K. D., Campan, M., Young, J., Long, T. I., Faasse, M. A., Kang, G. H., Widschwendter, M., Weener, D., Buchanan, D.*, et al.* (2006). CpG island methylator phenotype underlies sporadic microsatellite instability and is tightly associated with BRAF mutation in colorectal cancer. *Nat Genet* **38**, 787-793.

Weiser, K. C., Liu, B., Hansen, G. M., Skapura, D., Hentges, K. E., Yarlagadda, S., Morse Iii, H. C., and Justice, M. J. (2007). Retroviral insertions in the VISION database identify molecular pathways in mouse lymphoid leukemia and lymphoma. *Mamm Genome* **18**, 709-722.

Weiss, R. A. (2006). The discovery of endogenous retroviruses. *Retrovirology* **3**, 67.

Weng, A. P., Ferrando, A. A., Lee, W., Morris, J. P. t., Silverman, L. B., Sanchez-Irizarry, C., Blacklow, S. C., Look, A. T., and Aster, J. C. (2004). Activating mutations of NOTCH1 in human T cell acute lymphoblastic leukemia. *Science* **306**, 269-271.

Werner, C. A., Dohner, H., Joos, S., Trumper, L. H., Baudis, M., Barth, T. F., Ott, G., Moller, P., Lichter, P., and Bentz, M. (1997). High-level DNA amplifications are common genetic aberrations in B-cell neoplasms. *Am J Pathol* **151**, 335-342.

WHO (2008). World Health Organization. http://www.who.int/topics/cancer/en/.

Widschwendter, M., Fiegl, H., Egle, D., Mueller-Holzner, E., Spizzo, G., Marth, C., Weisenberger, D. J., Campan, M., Young, J., Jacobs, I., and Laird, P. W. (2007). Epigenetic stem cell signature in cancer. *Nat Genet* **39**, 157-158.

Wiener, Z., Kohalmi, B., Pocza, P., Jeager, J., Tolgyesi, G., Toth, S., Gorbe, E., Papp, Z., and Falus, A. (2007). TIM-3 is expressed in melanoma cells and is upregulated in TGF-beta stimulated mast cells. *J Invest Dermatol* **127**, 906-914.

Wieser, R. (2007). The oncogene and developmental regulator EVI1: expression, biochemical properties, and biological functions. *Gene* **396**, 346-357.

Willenbrock, H., and Fridlyand, J. (2005). A comparison study: applying segmentation to array CGH data for downstream analyses. *Bioinformatics* **21**, 4084-4091.

Williamson, D., Selfe, J., Gordon, T., Lu, Y. J., Pritchard-Jones, K., Murai, K., Jones, P., Workman, P., and Shipley, J. (2007). Role for amplification and expression of glypican-5 in rhabdomyosarcoma. *Cancer Res* **67**, 57-65.

Wilson, M. H., Coates, C. J., and George, A. L., Jr. (2007). PiggyBac transposon-mediated gene transfer in human cells. *Mol Ther* **15**, 139-145.

Winandy, S., Wu, P., and Georgopoulos, K. (1995). A dominant mutation in the Ikaros gene leads to rapid development of leukemia and lymphoma. *Cell* **83**, 289-299.

Wistuba, II, Behrens, C., Milchgrub, S., Syed, S., Ahmadian, M., Virmani, A. K., Kurvari, V., Cunningham, T. H., Ashfaq, R., Minna, J. D., and Gazdar, A. F. (1998). Comparison of features of human breast cancer cell lines and their corresponding tumors. *Clin Cancer Res* **4**, 2931-2938.

Wistuba, II, Bryant, D., Behrens, C., Milchgrub, S., Virmani, A. K., Ashfaq, R., Minna, J. D., and Gazdar, A. F. (1999). Comparison of features of human lung cancer cell lines and their corresponding tumors. *Clin Cancer Res* **5**, 991-1000.

Wong, D. J., Liu, H., Ridky, T. W., Cassarino, D., Segal, E., and Chang, H. Y. (2008). Module map of stem cell genes guides creation of epithelial cancer stem cells. *Cell Stem Cell* **2**, 333-344.

Wood, L. D., Parsons, D. W., Jones, S., Lin, J., Sjoblom, T., Leary, R. J., Shen, D., Boca, S. M., Barber, T., Ptak, J.*, et al.* (2007). The genomic landscapes of human breast and colorectal cancers. *Science* **318**, 1108-1113.

Woolfe, A., Goodson, M., Goode, D. K., Snell, P., McEwen, G. K., Vavouri, T., Smith, S. F., North, P., Callaway, H., Kelly, K.*, et al.* (2005). Highly conserved non-coding sequences are associated with vertebrate development. *PLoS Biol* **3**, e7.

Wotton, S. F., Blyth, K., Kilbey, A., Jenkins, A., Terry, A., Bernardin-Fried, F., Friedman, A. D., Baxter, E. W., Neil, J. C., and Cameron, E. R. (2004). RUNX1 transformation of primary embryonic fibroblasts is revealed in the absence of p53. *Oncogene* **23**, 5476-5486.

Wu, X., Li, Y., Crise, B., and Burgess, S. M. (2003). Transcription start regions in the human genome are favored targets for MLV integration. *Science* **300**, 1749-1751.

Wu, X., Luke, B. T., and Burgess, S. M. (2006). Redefining the common insertion site. *Virology* **344**, 292-295.

Xiao, Y., Segal, M. R., Yang, Y. H., and Yeh, R. F. (2007). A multi-array multi-SNP genotyping algorithm for Affymetrix SNP microarrays. *Bioinformatics* **23**, 1459-1467.

Xie, W., Chow, L. T., Paterson, A. J., Chin, E., and Kudlow, J. E. (1999). Conditional expression of the ErbB2 oncogene elicits reversible hyperplasia in stratified epithelia and up-regulation of TGFalpha expression in transgenic mice. *Oncogene* **18**, 3593-3607.

Xu, X., Wagner, K. U., Larson, D., Weaver, Z., Li, C., Ried, T., Hennighausen, L., Wynshaw-Boris, A., and Deng, C. X. (1999). Conditional mutation of Brca1 in mammary epithelial cells results in blunted ductal morphogenesis and tumour formation. *Nat Genet* **22**, 37-43.

Yamashita, M., and Emerman, M. (2006). Retroviral infection of non-dividing cells: old and new perspectives. *Virology* **344**, 88-93.

Yamashita, N., Osato, M., Huang, L., Yanagida, M., Kogan, S. C., Iwasaki, M., Nakamura, T., Shigesada, K., Asou, N., and Ito, Y. (2005). Haploinsufficiency of Runx1/AML1 promotes myeloid features and leukaemogenesis in BXH2 mice. *Br J Haematol* **131**, 495-507.

Yang, S., Jeung, H. C., Jeong, H. J., Choi, Y. H., Kim, J. E., Jung, J. J., Rha, S. Y., Yang, W. I., and Chung, H. C. (2007). Identification of genes with correlated patterns of variations in DNA copy number and gene expression level in gastric cancer. *Genomics* **89**, 451-459.

Yant, S. R., Wu, X., Huang, Y., Garrison, B., Burgess, S. M., and Kay, M. A. (2005). High-resolution genome-wide mapping of transposon integration in mammals. *Mol Cell Biol* **25**, 2085-2094.

Yao, J., Weremowicz, S., Feng, B., Gentleman, R. C., Marks, J. R., Gelman, R., Brennan, C., and Polyak, K. (2006). Combined cDNA array comparative genomic hybridization and serial analysis of gene expression analysis of breast tumor progression. *Cancer Res* **66**, 4065-4078.

Yasui, K., Mihara, S., Zhao, C., Okamoto, H., Saito-Ohara, F., Tomida, A., Funato, T., Yokomizo, A., Naito, S., Imoto, I.*, et al.* (2004). Alteration in copy numbers of genes as a mechanism for acquired drug resistance. *Cancer Res* **64**, 1403-1410.

Yoshida, K., Chambers, I., Nichols, J., Smith, A., Saito, M., Yasukawa, K., Shoyab, M., Taga, T., and Kishimoto, T. (1994). Maintenance of the pluripotential phenotype of embryonic stem cells through direct activation of gp130 signalling pathways. *Mech Dev* **45**, 163-171.

Yoshimoto, M., Joshua, A. M., Chilton-Macneill, S., Bayani, J., Selvarajah, S., Evans, A. J., Zielenska, M., and Squire, J. A. (2006). Three-color FISH analysis of TMPRSS2/ERG fusions in prostate cancer indicates that genomic microdeletion of chromosome 21 is associated with rearrangement. *Neoplasia* **8**, 465-469.

Yu, T., Ye, H., Sun, W., Li, K. C., Chen, Z., Jacobs, S., Bailey, D. K., Wong, D. T., and Zhou, X. (2007). A forward-backward fragment assembling algorithm for the identification of genomic amplification and deletion breakpoints using high-density single nucleotide polymorphism (SNP) array. *BMC Bioinformatics* **8**, 145.

Yucel, R., Karsunky, H., Klein-Hitpass, L., and Moroy, T. (2003). The transcriptional repressor Gfi1 affects development of early, uncommitted c-Kit+ T cell progenitors and CD4/CD8 lineage decision in the thymus. *J Exp Med* **197**, 831-844.

Yusa, K., Takeda, J., and Horie, K. (2004). Enhancement of Sleeping Beauty transposition by CpG methylation: possible role of heterochromatin formation. *Mol Cell Biol* **24**, 4004-4018.

Zagoraiou, L., Drabek, D., Alexaki, S., Guy, J. A., Klinakis, A. G., Langeveld, A., Skavdis, G., Mamalaki, C., Grosveld, F., and Savakis, C. (2001). In vivo transposition of Minos, a Drosophila mobile element, in mammalian tissues. *Proc Natl Acad Sci U S A* **98**, 11474-11478.

Zanette, D. L., Rivadavia, F., Molfetta, G. A., Barbuzano, F. G., Proto-Siqueira, R., Silva-Jr, W. A., Falcao, R. P., and Zago, M. A. (2007). miRNA expression profiles in chronic lymphocytic and acute lymphocytic leukemia. *Braz J Med Biol Res* **40**, 1435-1440.

Zayed, H., Izsvak, Z., Walisko, O., and Ivics, Z. (2004). Development of hyperactive sleeping beauty transposon vectors by mutational analysis. *Mol Ther* **9**, 292-304.

Zeller, K. I., Zhao, X., Lee, C. W., Chiu, K. P., Yao, F., Yustein, J. T., Ooi, H. S., Orlov, Y. L., Shahab, A., Yong, H. C.*, et al.* (2006). Global mapping of c-Myc binding sites and target gene networks in human B cells. *Proc Natl Acad Sci U S A* **103**, 17834-17839.

Zender, L., Spector, M. S., Xue, W., Flemming, P., Cordon-Cardo, C., Silke, J., Fan, S. T., Luk, J. M., Wigler, M., Hannon, G. J.*, et al.* (2006). Identification and validation of oncogenes in liver cancer using an integrative oncogenomic approach. *Cell* **125**, 1253-1267.

Zhang, P., Chin, W., Chow, L. T., Chan, A. S., Yim, A. P., Leung, S. F., Mok, T. S., Chang, K. S., Johnson, P. J., and Chan, J. Y. (2000). Lack of expression for the suppressor PML in human small cell lung carcinoma. *Int J Cancer* **85**, 599-605.

Zhang, S., Qian, X., Redman, C., Bliskovski, V., Ramsay, E. S., Lowy, D. R., and Mock, B. A. (2003). p16 INK4a gene promoter variation and differential binding of a repressor, the ras-responsive zinc-finger transcription factor, RREB. *Oncogene* **22**, 2285-2295.

Zhang, Y., Wang, Z., and Magnuson, N. S. (2007). Pim-1 kinase-dependent phosphorylation of p21Cip1/WAF1 regulates its stability and cellular localization in H1299 cells. *Mol Cancer Res* **5**, 909-922.

Zhao, C., Yasui, K., Lee, C. J., Kurioka, H., Hosokawa, Y., Oka, T., and Inazawa, J. (2003). Elevated expression levels of NCOA3, TOP1, and TFAP2C in breast tumors as predictors of poor prognosis. *Cancer* **98**, 18-23.

Zhao, X., Li, C., Paez, J. G., Chin, K., Janne, P. A., Chen, T. H., Girard, L., Minna, J., Christiani, D., Leo, C.*, et al.* (2004). An integrated view of copy number and allelic alterations in the cancer genome using single nucleotide polymorphism arrays. *Cancer Res* **64**, 3060-3071.

Zhao, Z. L., Huang, Q. Y., Xu, S., Zhang, L., and Zhao, H. R. (2006). [Expression of promyelocytic leukaemia protein in lung carcinomas and clinical significance thereof]. *Zhonghua Yi Xue Za Zhi* **86**, 3362-3366.

Zhong, Z., Wan, B., Qiu, Y., Ni, J., Tang, W., Chen, X., Yang, Y., Shen, S., Wang, Y., Bai, M.*, et al.* (2007). Identification of a novel human zinc finger gene, ZNF438, with transcription inhibition activity. *J Biochem Mol Biol* **40**, 517-524.

Zippo, A., De Robertis, A., Serafini, R., and Oliviero, S. (2007). PIM1-dependent phosphorylation of histone H3 at serine 10 is required for MYC-dependent transcriptional activation and oncogenic transformation. *Nat Cell Biol* **9**, 932-944.

Appendix A. Human Ensembl genes and their mouse orthologues for known cancer genes in the Cancer Gene Census.

Abbreviations:
Tissue type: E=epithelial, L=leukaemia/lymphoma, M=mesenchymal, O=other
Dom/Rec: Dom=dominant, Rec=recessive
Mutation type: A=amplification, D=large deletion, F=frameshift, Mis=missense, N=nonsense, S=splice site, T=translocation

Human Ensembl ID	Mouse Ensembl ID	Gene name	Somatic mutation	Germline mutation	Tissue type	Dom/ Rec	Mutation type
ENSG00000136754	ENSMUSG00000058835	*ABI1*	yes		L	Dom	T
ENSG00000097007	ENSMUSG00000026842	*ABL1*	yes		L	Dom	T, Mis
ENSG00000143322	ENSMUSG00000026596	*ABL2*	yes		L	Dom	T
ENSG00000164398	ENSMUSG00000020333	*ACSL6*	yes		L	Dom	T
ENSG00000143443	ENSMUSG00000068860	*AF1q*	yes		L	Dom	T
ENSG00000172493	ENSMUSG00000029313	*AFF1*	yes		L	Dom	T
ENSG00000144218	ENSMUSG00000037138	*AFF3*	yes		L	Dom	T
ENSG00000072364	ENSMUSG00000049470	*AFF4*	yes		L	Dom	T
ENSG00000127914	ENSMUSG00000040407	*AKAP9*	yes		E	Dom	T
ENSG00000105221	ENSMUSG00000004056	*AKT2*	yes		E	Dom	A
ENSG00000105221	ENSMUSG00000073134	*AKT2*	yes		E	Dom	A
ENSG00000171094	ENSMUSG00000055471	*ALK*	yes		L	Dom	T
ENSG00000134982	ENSMUSG00000005871	*APC*	yes	yes	E, M, O	Rec	D, Mis, N, F, S
ENSG00000145819	ENSMUSG00000036452	*ARHGAP26*	yes		L	Dom	T, F, S
ENSG00000196914	ENSMUSG00000059495	*ARHGEF12*	yes		L	Dom	T
ENSG00000143437	ENSMUSG00000015522	*ARNT*	yes		L	Dom	T
ENSG00000169696	ENSMUSG00000025142	*ASPSCR1*	yes		M	Dom	T
ENSG00000123268	ENSMUSG00000055574	*ATF1*	yes		E, M	Dom	T
ENSG00000123268	ENSMUSG00000023027	*ATF1*	yes		E, M	Dom	T
ENSG00000138363	ENSMUSG00000026192	*ATIC*	yes		L	Dom	T
ENSG00000149311	ENSMUSG00000034218	*ATM*	yes	yes	L, O	Rec	D, Mis, N, F, S
ENSG00000142867	ENSMUSG00000028191	*BCL10*	yes		L	Dom	T
ENSG00000119866	ENSMUSG00000000861	*BCL11A*	yes		L	Dom	T
ENSG00000127152	ENSMUSG00000048251	*BCL11B*	yes		L	Dom	T
ENSG00000171791	ENSMUSG00000057329	*BCL2*	yes		L	Dom	T
ENSG00000069399	ENSMUSG00000053175	*BCL3*	yes		L	Dom	T
ENSG00000113916	ENSMUSG00000022508	*BCL6*	yes		L	Dom	T, Mis
ENSG00000110987	ENSMUSG00000029438	*BCL7A*	yes		L	Dom	T
ENSG00000116128	ENSMUSG00000038256	*BCL9*	yes		L	Dom	T
ENSG00000186716	ENSMUSG00000009681	*BCR*	yes		L	Dom	T
ENSG00000023445	ENSMUSG00000032000	*BIRC3*	yes		L	Dom	T
ENSG00000197299	ENSMUSG00000030528	*BLM*		yes	L, E	Rec	Mis, N, F
ENSG00000157764	ENSMUSG00000002413	*BRAF*	yes		E	Dom	Mis, T
ENSG00000012048	ENSMUSG00000017146	*BRCA1*	yes	yes	E	Rec	D, Mis, N, F, S
ENSG00000139618	ENSMUSG00000041147	*BRCA2*	yes	yes	L, E	Rec	D, Mis, N, F, S
ENSG00000141867	ENSMUSG00000024002	*BRD4*	yes		E	Dom	T
ENSG00000136492	ENSMUSG00000034329	*BRIP1*		yes	L, E	Rec	F, N, Mis
ENSG00000133639	ENSMUSG00000036478	*BTG1*	yes		L	Dom	T
ENSG00000156970	ENSMUSG00000040084	*BUB1B*		yes	M	Rec	Mis, N, F, S
ENSG00000110619	ENSMUSG00000010755	*CARS*	yes		L	Dom	T
ENSG00000137812	ENSMUSG00000027326	*CASC5*	yes		L	Dom	T
ENSG00000129993	ENSMUSG00000006362	*CBFA2T3*	yes		L	Dom	T
ENSG00000067955	ENSMUSG00000031885	*CBFB*	yes		L	Dom	T
ENSG00000110395	ENSMUSG00000034342	*CBL*	yes		L	Dom	T
ENSG00000100814	ENSMUSG00000071470	*CCNB1IP1*	yes		M	Dom	T
ENSG00000110092	ENSMUSG00000070348	*CCND1*	yes		L, E	Dom	T
ENSG00000118971	ENSMUSG00000000184	*CCND2*	yes		L	Dom	T
ENSG00000112576	ENSMUSG00000034165	*CCND3*	yes		L	Dom	T
ENSG00000163660	ENSMUSG00000027829	*CCNL1*	yes		E	Dom	T
ENSG00000128283	ENSMUSG00000049521	*CDC42EP1*	yes		L	Dom	T
ENSG00000134371	ENSMUSG00000026361	*CDC73*	yes	yes	E, M	Rec	Mis, N, F
ENSG00000039068	ENSMUSG00000000303	*CDH1*	yes	yes	E	Rec	Mis, N, F, S
ENSG00000140937	ENSMUSG00000031673	*CDH11*	yes		M	Dom	T
ENSG00000135446	ENSMUSG00000006728	*CDK4*		yes	E	Dom	Mis
ENSG00000105810	ENSMUSG00000040274	*CDK6*	yes		L	Dom	T
ENSG00000147889	ENSMUSG00000044303	*CDKN2A*	yes	yes	L, E, M, O	Rec	D, S
ENSG00000165556	ENSMUSG00000029646	*CDX2*	yes		L	Dom	T
ENSG00000183765	ENSMUSG00000029521	*CHEK2*		yes	E	Rec	F
ENSG00000109220	ENSMUSG00000029229	*CHIC2*	yes		L	Dom	T
ENSG00000128656	ENSMUSG00000056486	*CHN1*	yes		M	Dom	T
ENSG00000179583	ENSMUSG00000022504	*CIITA*	yes		L	Dom	T
ENSG00000172409	ENSMUSG00000027079	*CLP1*	yes		L	Dom	T
ENSG00000141367	ENSMUSG00000047126	*CLTC*	yes		L	Dom	T
ENSG00000169714	ENSMUSG00000030057	*CNBP*	yes		M	Dom	T
ENSG00000108821	ENSMUSG00000001506	*COL1A1*	yes		M	Dom	T
ENSG00000164919	ENSMUSG00000066491	*COX6C*	yes		M	Dom	T
ENSG00000164919	ENSMUSG00000014313	*COX6C*	yes		M	Dom	T
ENSG00000164919	ENSMUSG00000069096	*COX6C*	yes		M	Dom	T
ENSG00000118260	ENSMUSG00000025958	*CREB1*	yes		M	Dom	T
ENSG00000005339	ENSMUSG00000022521	*CREBBP*	yes		L	Dom	T
ENSG00000105662	ENSMUSG00000003575	*CRTC1*	yes		E	Dom	T
ENSG00000168036	ENSMUSG00000006932	*CTNNB1*	yes		E, M, O	Dom	H, Mis
ENSG00000144476	ENSMUSG00000044337	*CXCR7*	yes		M	Dom	T
ENSG00000138336	ENSMUSG00000075011	*CXXC6*	yes		L	Dom	T
ENSG00000083799	ENSMUSG00000036712	*CYLD*	yes	yes	E	Rec	Mis, N, F, S
ENSG00000108091	ENSMUSG00000048701	*D10S170*	yes		E	Dom	T
ENSG00000134574	ENSMUSG00000002109	*DDB2*		yes	E	Rec	Mis, N
ENSG00000175197	ENSMUSG00000025408	*DDIT3*	yes		M	Dom	T

ENSG00000178105	ENSMUSG00000053289	*DDX10*	yes		L	Dom	T
ENSG00000110367	ENSMUSG00000059707	*DDX6*	yes		L	Dom	T
ENSG00000110367	ENSMUSG00000032097	*DDX6*	yes		L	Dom	T
ENSG00000124795	ENSMUSG00000021377	*DEK*	yes		L	Dom	T
ENSG00000146648	ENSMUSG00000020122	*EGFR*	yes		E, O	Dom	A, O, Mis
ENSG00000156976	ENSMUSG00000022884	*EIF4A2*	yes		L	Dom	T
ENSG00000156976	ENSMUSG00000022884	*EIF4A2*	yes		L	Dom	T
ENSG00000102034	ENSMUSG00000031103	*ELF4*	yes		E	Dom	T
ENSG00000105656	ENSMUSG00000070002	*ELL*	yes		L	Dom	T
ENSG00000100393	ENSMUSG00000055024	*EP300*	yes		L, E	Rec	T
ENSG00000085832	ENSMUSG00000028552	*EPS15*	yes		L	Dom	T
ENSG00000141736	ENSMUSG00000062312	*ERBB2*	yes		E	Dom	A, Mis, O
ENSG00000104884	ENSMUSG00000030400	*ERCC2*		yes	E	Rec	Mis, N, F, S
ENSG00000163161	ENSMUSG00000024382	*ERCC3*		yes	E	Rec	Mis, S
ENSG00000175595	ENSMUSG00000022545	*ERCC4*		yes	E	Rec	Mis, N, F
ENSG00000134899	ENSMUSG00000026048	*ERCC5*		yes	E	Rec	Mis, N, F
ENSG00000157554	ENSMUSG00000040732	*ERG*	yes		M, E, L	Dom	T
ENSG00000006468	ENSMUSG00000047643	*ETV1*	yes		M, E	Dom	T
ENSG00000006468	ENSMUSG00000004151	*ETV1*	yes		M, E	Dom	T
ENSG00000175832	ENSMUSG00000017724	*ETV4*	yes		M	Dom	T
ENSG00000139083	ENSMUSG00000030199	*ETV6*	yes		L, E, M	Dom	T
ENSG00000085276	ENSMUSG00000027684	*EVI1*	yes		L	Dom	T
ENSG00000182944	ENSMUSG00000038649	*EWSR1*	yes		L, M	Dom	T
ENSG00000182944	ENSMUSG00000070829	*EWSR1*	yes		L, M	Dom	T
ENSG00000182944	ENSMUSG00000009079	*EWSR1*	yes		L, M	Dom	T
ENSG00000182197	ENSMUSG00000061731	*EXT1*		yes	M	Rec	Mis, N, F, S
ENSG00000151348	ENSMUSG00000027198	*EXT2*		yes	M	Rec	Mis, N, F, S
ENSG00000187741	ENSMUSG00000032815	*FANCA*		yes	L	Rec	D, Mis, N, F, S
ENSG00000158169	ENSMUSG00000021461	*FANCC*		yes	L	Rec	D, Mis, N, F, S
ENSG00000144554	ENSMUSG00000034023	*FANCD2*		yes	L	Rec	D, Mis, N, F
ENSG00000112039	ENSMUSG00000007570	*FANCE*		yes	L	Rec	N, F, S
ENSG00000183161	ENSMUSG00000043480	*FANCF*		yes	L	Rec	N, F
ENSG00000165281	ENSMUSG00000028453	*FANCG*		yes	L	Rec	Mis, N, F, S
ENSG00000026103	ENSMUSG00000024778	*FAS*	yes		L, E, O	Rec	Mis
ENSG00000109670	ENSMUSG00000028086	*FBXW7*	yes		E	Dom	Mis, N
ENSG00000072694	ENSMUSG00000026656	*FCGR2B*	yes		L	Dom	T
ENSG00000072694	ENSMUSG00000059498	*FCGR2B*	yes		L	Dom	T
ENSG00000163497	ENSMUSG00000055197	*FEV*	yes		M	Dom	T
ENSG00000077782	ENSMUSG00000031565	*FGFR1*	yes		L	Dom	T
ENSG00000112486	ENSMUSG00000069135	*FGFR1OP*	yes		L	Dom	T
ENSG00000066468	ENSMUSG00000030849	*FGFR2*	yes		E	Dom	Mis
ENSG00000068078	ENSMUSG00000054252	*FGFR3*	yes		L, E	Dom	Mis, T
ENSG00000091483	ENSMUSG00000026526	*FH*		yes	E, M	Rec	Mis, N, F
ENSG00000091483	ENSMUSG00000037498	*FH*		yes	E, M	Rec	Mis, N, F
ENSG00000145216	ENSMUSG00000029227	*FIP1L1*	yes		L	Dom	T
ENSG00000154803	ENSMUSG00000032633	*FLCN*		yes	E, M	Rec?	Mis. N, F
ENSG00000122025	ENSMUSG00000042817	*FLT3*	yes		L	Dom	Mis, O
ENSG00000187239	ENSMUSG00000075415	*FNBP1*	yes		L	Dom	T
ENSG00000150907	ENSMUSG00000044167	*FOXO1A*	yes		M	Dom	T
ENSG00000118689	ENSMUSG00000048756	*FOXO3A*	yes		L	Dom	T
ENSG00000070404	ENSMUSG00000020325	*FSTL3*	yes		L	Dom	T
ENSG00000089280	ENSMUSG00000066554	*FUS*	yes		M, L	Dom	T
ENSG00000089280	ENSMUSG00000030795	*FUS*	yes		M, L	Dom	T
ENSG00000119537	ENSMUSG00000009905	*FVT1*	yes		L	Dom	T
ENSG00000007237	ENSMUSG00000033066	*GAS7*	yes		L	Dom	T
ENSG00000102145	ENSMUSG00000031162	*GATA1*	yes		L	Dom	Mis, F
ENSG00000163655	ENSMUSG00000027823	*GMPS*	yes		L	Dom	T
ENSG00000087460	ENSMUSG00000027523	*GNAS*	yes		E	Dom	Mis
ENSG00000066455	ENSMUSG00000021192	*GOLGA5*	yes		E	Dom	T
ENSG00000047932	ENSMUSG00000019861	*GOPC*	yes		O	Dom	O
ENSG00000147257	ENSMUSG00000055653	*GPC3*		yes	O	X-linked Rec	T, D, Mis, N, F, S
ENSG00000171723	ENSMUSG00000047454	*GPHN*	yes		L	Dom	T
ENSG00000127946	ENSMUSG00000039959	*HIP1*	yes		L	Dom	T
ENSG00000198339	ENSMUSG00000069305	*HIST1H4I*	yes		L	Dom	T
ENSG00000108924	ENSMUSG00000003949	*HLF*	yes		L	Dom	T
ENSG00000130675	ENSMUSG00000001566	*HLXB9*	yes		L	Dom	T
ENSG00000149948	ENSMUSG00000056758	*HMGA2*	yes		M	Dom	T
ENSG00000005073	ENSMUSG00000038210	*HOXA11*	yes		L	Dom	T
ENSG00000106031	ENSMUSG00000038203	*HOXA13*	yes		L	Dom	T
ENSG00000078399	ENSMUSG00000038227	*HOXA9*	yes		L	Dom	T
ENSG00000123388	ENSMUSG00000001656	*HOXC11*	yes		L	Dom	T
ENSG00000123364	ENSMUSG00000001655	*HOXC13*	yes		L	Dom	T
ENSG00000128713	ENSMUSG00000042499	*HOXD11*	yes		L	Dom	T
ENSG00000128714	ENSMUSG00000001819	*HOXD13*	yes		L	Dom	T
ENSG00000174775	ENSMUSG00000025499	*HRAS*	yes	yes	E, L, M	Dom	Mis
ENSG00000185811	ENSMUSG00000018654	*IKZF1*	yes		L	Dom	T
ENSG00000109471	ENSMUSG00000027720	*IL2*	yes		L	Dom	T
ENSG00000103522	ENSMUSG00000030745	*IL21R*	yes		L	Dom	T
ENSG00000137265	ENSMUSG00000021356	*IRF4*	yes		L	Dom	T
ENSG00000113263	ENSMUSG00000020395	*ITK*	yes		L	Dom	T
ENSG00000096968	ENSMUSG00000024789	*JAK2*	yes		L	Dom	T, Mis, O
ENSG00000153814	ENSMUSG00000063568	*JAZF1*	yes		M	Dom	T
ENSG00000157404	ENSMUSG00000005672	*KIT*	yes	yes	L, M, O	Dom	Mis, O
ENSG00000067082	ENSMUSG00000000078	*KLF6*	yes		E, O	Rec	Mis, N
ENSG00000133703	ENSMUSG00000030265	*KRAS*	yes		L, E, M, O	Dom	Mis
ENSG00000126777	ENSMUSG00000021843	*KTN1*	yes		E	Dom	T
ENSG00000002834	ENSMUSG00000038366	*LASP1*	yes		L	Dom	T
ENSG00000002834	ENSMUSG00000037792	*LASP1*	yes		L	Dom	T
ENSG00000182866	ENSMUSG00000000409	*LCK*	yes		L	Dom	T
ENSG00000136167	ENSMUSG00000021998	*LCP1*	yes		L	Dom	T
ENSG00000183722	ENSMUSG00000048332	*LHFP*	yes		M	Dom	T
ENSG00000113594	ENSMUSG00000054263	*LIFR*	yes		E	Dom	T
ENSG00000166407	ENSMUSG00000036111	*LMO1*	yes		L	Dom	T
ENSG00000135363	ENSMUSG00000032698	*LMO2*	yes		L	Dom	T

ENSG00000145012	ENSMUSG00000033306	*LPP*	yes		L, M	Dom	T
ENSG00000104903	ENSMUSG00000034041	*LYL1*	yes		L	Dom	T
ENSG00000178573	ENSMUSG00000055435	*MAF*	yes		L	Dom	T
ENSG00000204103	ENSMUSG00000074622	*MAFB*	yes		L	Dom	T
ENSG00000172175	ENSMUSG00000032688	*MALT1*	yes		L	Dom	T
ENSG00000184384	ENSMUSG00000031925	*MAML2*	yes		E	Dom	T
ENSG00000065559	ENSMUSG00000033352	*MAP2K4*	yes		E	Rec	D, Mis, N
ENSG00000206115	ENSMUSG00000051636	*MDS1*	yes		L	Dom	T
ENSG00000133895	ENSMUSG00000024947	*MEN1*	yes	yes	E	Rec	D, Mis, N, F, S
ENSG00000105976	ENSMUSG00000009376	*MET*	yes		E	Dom	Mis
ENSG00000162775	ENSMUSG00000048109	*MKL1*	yes		L	Dom	T
ENSG00000178053	ENSMUSG00000048416	*MLF1*	yes		L	Dom	T
ENSG00000076242	ENSMUSG00000032498	*MLH1*	yes	yes	E, O	Rec	D, Mis, N, F, S
ENSG00000118058	ENSMUSG00000002028	*MLL*	yes		L	Dom	T, O
ENSG00000130382	ENSMUSG00000024212	*MLLT1*	yes		L	Dom	T
ENSG00000078403	ENSMUSG00000026743	*MLLT10*	yes		L	Dom	T
ENSG00000171843	ENSMUSG00000028496	*MLLT3*	yes		L	Dom	T
ENSG00000130396	ENSMUSG00000068036	*MLLT4*	yes		L	Dom	T
ENSG00000108292	ENSMUSG00000038437	*MLLT6*	yes		L	Dom	T
ENSG00000184481	ENSMUSG00000042903	*MLLT7*	yes		L	Dom	T
ENSG00000169184	ENSMUSG00000070576	*MN1*	yes		L, O	Dom	T
ENSG00000117400	ENSMUSG00000006389	*MPL*	yes	yes	L	Dom	Mis
ENSG00000095002	ENSMUSG00000024151	*MSH2*	yes	yes	E	Rec	D, Mis, N, F, S
ENSG00000116062	ENSMUSG00000005370	*MSH6*	yes	yes	E	Rec	Mis, N, F, S
ENSG00000153944	ENSMUSG00000069769	*MSI2*	yes		L	Dom	T
ENSG00000147065	ENSMUSG00000031207	*MSN*	yes		L	Dom	T
ENSG00000182712	ENSMUSG00000031200	*MTCP1*	yes		L	Dom	T
ENSG00000185499	ENSMUSG00000042784	*MUC1*	yes		L	Dom	T
ENSG00000132781	ENSMUSG00000028687	*MUTYH*		yes	E	Rec	Mis
ENSG00000136997	ENSMUSG00000022346	*MYC*	yes		L, E	Dom	A, T
ENSG00000116990	ENSMUSG00000028654	*MYCL1*	yes		E	Dom	A
ENSG00000134323	ENSMUSG00000037169	*MYCN*	yes		O	Dom	A
ENSG00000133392	ENSMUSG00000018830	*MYH11*	yes		L	Dom	T
ENSG00000100345	ENSMUSG00000022443	*MYH9*	yes		L	Dom	T
ENSG00000083168	ENSMUSG00000031540	*MYST3*	yes		L	Dom	T
ENSG00000156650	ENSMUSG00000021767	*MYST4*	yes		L	Dom	T
ENSG00000104320	ENSMUSG00000028224	*NBN*		yes	L, E, M, O	Rec	Mis, N, F
ENSG00000140396	ENSMUSG00000005886	*NCOA2*	yes		L	Dom	T
ENSG00000138293	ENSMUSG00000021908	*NCOA4*	yes		E	Dom	T
ENSG00000138293	ENSMUSG00000056234	*NCOA4*	yes		E	Dom	T
ENSG00000196712	ENSMUSG00000020716	*NF1*	yes	yes	O	Rec	D, Mis, N, F, S, O
ENSG00000186575	ENSMUSG00000009073	*NF2*	yes	yes	O	Rec	D, Mis, N, F, S, O
ENSG00000077150	ENSMUSG00000025225	*NFKB2*	yes		L	Dom	T
ENSG00000100503	ENSMUSG00000021068	*NIN*	yes		L	Dom	T
ENSG00000147140	ENSMUSG00000067514	*NONO*	yes		E	Dom	T
ENSG00000147140	ENSMUSG00000031311	*NONO*	yes		E	Dom	T
ENSG00000148400	ENSMUSG00000026923	*NOTCH1*	yes		L	Dom	T, Mis, O
ENSG00000119508	ENSMUSG00000028341	*NR4A3*	yes		M	Dom	T
ENSG00000009307	ENSMUSG00000072420	*NRAS*	yes		L, E	Dom	Mis
ENSG00000009307	ENSMUSG00000068823	*NRAS*	yes		L, E	Dom	Mis
ENSG00000009307	ENSMUSG00000047588	*NRAS*	yes		L, E	Dom	Mis
ENSG00000165671	ENSMUSG00000021488	*NSD1*	yes		L	Dom	T
ENSG00000198400	ENSMUSG00000028072	*NTRK1*	yes		E	Dom	T
ENSG00000140538	ENSMUSG00000059146	*NTRK3*	yes		E, M	Dom	T
ENSG00000137497	ENSMUSG00000066306	*NUMA1*	yes		L	Dom	T
ENSG00000110713	ENSMUSG00000063550	*NUP98*	yes		L	Dom	T
ENSG00000184507	ENSMUSG00000041358	*NUT*	yes		E	Dom	T
ENSG00000205927	ENSMUSG00000039830	*OLIG2*	yes		L	Dom	T
ENSG00000127083	ENSMUSG00000048368	*OMD*	yes		M	Dom	T
ENSG00000168092	ENSMUSG00000003131	*PAFAH1B2*	yes		L	Dom	T
ENSG00000083093	ENSMUSG00000044702	*PALB2*		yes	L, O, E	Rec	F, N, Mis
ENSG00000100105	ENSMUSG00000020453	*PATZ1*	yes		M	Dom	T
ENSG00000135903	ENSMUSG00000004872	*PAX3*	yes		M	Dom	T
ENSG00000196092	ENSMUSG00000014030	*PAX5*	yes		L	Dom	T
ENSG00000009709	ENSMUSG00000028736	*PAX7*	yes		M	Dom	T
ENSG00000125618	ENSMUSG00000026976	*PAX8*	yes		E	Dom	T
ENSG00000185630	ENSMUSG00000052534	*PBX1*	yes		L	Dom	T
ENSG00000078674	ENSMUSG00000031592	*PCM1*	yes		E	Dom	T
ENSG00000160613	ENSMUSG00000035382	*PCSK7*	yes		L	Dom	T
ENSG00000178104	ENSMUSG00000038170	*PDE4DIP*	yes		L	Dom	T
ENSG00000100311	ENSMUSG00000000489	*PDGFB*	yes		M	Dom	T
ENSG00000134853	ENSMUSG00000029231	*PDGFRA*	yes		L, M, O	Dom	Mis, O, T
ENSG00000113721	ENSMUSG00000024620	*PDGFRB*	yes		L	Dom	T
ENSG00000179094	ENSMUSG00000020893	*PER1*	yes		L	Dom	T
ENSG00000109132	ENSMUSG00000012520	*PHOX2B*	yes	yes	O	Rec	Mis, F
ENSG00000073921	ENSMUSG00000039361	*PICALM*	yes		L	Dom	T
ENSG00000121879	ENSMUSG00000027665	*PIK3CA*	yes		E, O	Dom	Mis
ENSG00000137193	ENSMUSG00000024014	*PIM1*	yes		L	Dom	T
ENSG00000181690	ENSMUSG00000003282	*PLAG1*	yes		E	Dom	T
ENSG00000140464	ENSMUSG00000036986	*PML*	yes		L	Dom	T
ENSG00000064933	ENSMUSG00000026098	*PMS1*		yes	E	Rec	Mis, N
ENSG00000110777	ENSMUSG00000032053	*POU2AF1*	yes		L	Dom	T
ENSG00000204531	ENSMUSG00000024406	*POU5F1*	yes		M	Dom	T
ENSG00000132170	ENSMUSG00000000440	*PPARG*	yes		E	Dom	T
ENSG00000143294	ENSMUSG00000004895	*PRCC*	yes		E	Dom	T
ENSG00000142611	ENSMUSG00000039410	*PRDM16*	yes		L	Dom	T
ENSG00000108946	ENSMUSG00000020612	*PRKAR1A*	yes	yes	E, M	Dom, Rec	T, Mis, N, F, S
ENSG00000116132	ENSMUSG00000026586	*PRRX1*	yes		L	Dom	T
ENSG00000164985	ENSMUSG00000034033	*PSIP1*	yes		L	Dom	T
ENSG00000164985	ENSMUSG00000028484	*PSIP1*	yes		L	Dom	T
ENSG00000185920	ENSMUSG00000021466	*PTCH1*	yes	yes	E, M	Rec	Mis, N, F, S
ENSG00000171862	ENSMUSG00000013663	*PTEN*	yes	yes	L, E, M, O	Rec	D, Mis, N, F, S

ENSG00000179295	ENSMUSG00000043733	*PTPN11*	yes		L	Dom	Mis
ENSG00000029725	ENSMUSG00000020817	*RABEP1*	yes		L	Dom	T
ENSG00000182185	ENSMUSG00000059060	*RAD51L1*	yes		M	Dom	T
ENSG00000204764	ENSMUSG00000040594	*RANBP17*	yes		L	Dom	T
ENSG00000138698	ENSMUSG00000028149	*RAP1GDS1*	yes		L	Dom	T
ENSG00000131759	ENSMUSG00000037992	*RARA*	yes		L	Dom	T
ENSG00000139687	ENSMUSG00000022105	*RB1*	yes	yes	L, E, M, O	Rec	D, Mis, N, F, S
ENSG00000162775	ENSMUSG00000048109	*RBM15*	yes		L	Dom	T
ENSG00000160957	ENSMUSG00000033762	*RECQL4*		yes	M	Rec	N, F, S
ENSG00000162924	ENSMUSG00000020275	*REL*	yes		L	Dom	A
ENSG00000165731	ENSMUSG00000030110	*RET*	yes	yes	E, O	Dom	T, Mis, N, F
ENSG00000168421	ENSMUSG00000029204	*RHOH*	yes		L	Dom	T
ENSG00000047936	ENSMUSG00000019893	*ROS1*	yes		O	Dom	T
ENSG00000116251	ENSMUSG00000028936	*RPL22*	yes		L	Dom	T
ENSG00000163902	ENSMUSG00000030062	*RPN1*	yes		L	Dom	T
ENSG00000159216	ENSMUSG00000022952	*RUNX1*	yes		L	Dom	T
ENSG00000079102	ENSMUSG00000006586	*RUNX1T1*	yes		L	Dom	T
ENSG00000126524	ENSMUSG00000025337	*SBDS*		yes	L	Rec	Gene Conversion
ENSG00000117118	ENSMUSG00000009863	*SDHB*		yes	O	Rec	Mis, N, F
ENSG00000143252	ENSMUSG00000058076	*SDHC*		yes	O	Rec	Mis, N, F
ENSG00000204370	ENSMUSG00000000171	*SDHD*		yes	O	Rec	Mis, N, F, S
ENSG00000184702	ENSMUSG00000072214	*SEPT5*	yes		L	Dom	T
ENSG00000125354	ENSMUSG00000050379	*SEPT6*	yes		L	Dom	T
ENSG00000184640	ENSMUSG00000059248	*SEPT9*	yes		L	Dom	T
ENSG00000116560	ENSMUSG00000028820	*SFPQ*	yes		E	Dom	T
ENSG00000141985	ENSMUSG00000003200	*SH3GL1*	yes		L	Dom	T
ENSG00000008300	ENSMUSG00000023473	*SLC26A6*	yes		L	Dom	T
ENSG00000141646	ENSMUSG00000024515	*SMAD4*	yes	yes	E	Rec	D, Mis, N, F
ENSG00000099956	ENSMUSG00000000902	*SMARCB1*	yes	yes	M	Rec	D, N, F, S
ENSG00000128602	ENSMUSG00000001761	*SMO*	yes		E	Dom	Mis
ENSG00000185338	ENSMUSG00000038037	*SOCS1*	yes		L	Rec	F, O
ENSG00000128487	ENSMUSG00000042331	*SPECC1*	yes		L	Dom	T
ENSG00000141380	ENSMUSG00000037013	*SS18*	yes		M	Dom	T
ENSG00000184402	ENSMUSG00000039086	*SS18L1*	yes		M	Dom	T
ENSG00000126752	ENSMUSG00000035371	*SSX1*	yes		M	Dom	T
ENSG00000126752	ENSMUSG00000071816	*SSX1*	yes		M	Dom	T
ENSG00000126752	ENSMUSG00000023165	*SSX1*	yes		M	Dom	T
ENSG00000126752	ENSMUSG00000068218	*SSX1*	yes		M	Dom	T
ENSG00000126752	ENSMUSG00000062814	*SSX1*	yes		M	Dom	T
ENSG00000126752	ENSMUSG00000068219	*SSX1*	yes		M	Dom	T
ENSG00000187754	ENSMUSG00000035371	*SSX2*	yes		M	Dom	T
ENSG00000187754	ENSMUSG00000071816	*SSX2*	yes		M	Dom	T
ENSG00000187754	ENSMUSG00000023165	*SSX2*	yes		M	Dom	T
ENSG00000187754	ENSMUSG00000068218	*SSX2*	yes		M	Dom	T
ENSG00000187754	ENSMUSG00000062814	*SSX2*	yes		M	Dom	T
ENSG00000187754	ENSMUSG00000068219	*SSX2*	yes		M	Dom	T
ENSG00000204645	ENSMUSG00000035371	*SSX4*	yes		M	Dom	T
ENSG00000204645	ENSMUSG00000071816	*SSX4*	yes		M	Dom	T
ENSG00000204645	ENSMUSG00000023165	*SSX4*	yes		M	Dom	T
ENSG00000204645	ENSMUSG00000068218	*SSX4*	yes		M	Dom	T
ENSG00000204645	ENSMUSG00000062814	*SSX4*	yes		M	Dom	T
ENSG00000204645	ENSMUSG00000068219	*SSX4*	yes		M	Dom	T
ENSG00000123473	ENSMUSG00000028718	*STIL*	yes		L	Dom	T
ENSG00000118046	ENSMUSG00000003068	*STK11*	yes	yes	E, M, O	Rec	D, Mis, N, F, S
ENSG00000107882	ENSMUSG00000025231	*SUFU*	yes	yes	O	Rec	D, F, S
ENSG00000178691	ENSMUSG00000017548	*SUZ12*	yes		M	Dom	T
ENSG00000165025	ENSMUSG00000021457	*SYK*	yes		L	Dom	T
ENSG00000172660	ENSMUSG00000020680	*TAF15*	yes		L, M	Dom	T
ENSG00000162367	ENSMUSG00000028717	*TAL1*	yes		L	Dom	T
ENSG00000186051	ENSMUSG00000028417	*TAL2*	yes		L	Dom	T
ENSG00000187735	ENSMUSG00000033813	*TCEA1*	yes		E	Dom	T
ENSG00000135100	ENSMUSG00000029556	*TCF1*	yes	yes	E	Rec	Mis, F
ENSG00000140262	ENSMUSG00000032228	*TCF12*	yes		M	Dom	T
ENSG00000071564	ENSMUSG00000020167	*TCF3*	yes		L	Dom	T
ENSG00000100721	ENSMUSG00000041359	*TCL1A*	yes		L	Dom	T
ENSG00000135605	ENSMUSG00000029217	*TEC*	yes		M	Dom	T
ENSG00000068323	ENSMUSG00000000134	*TFE3*	yes		E	Dom	T
ENSG00000112561	ENSMUSG00000023990	*TFEB*	yes		E,M	Dom	T
ENSG00000114354	ENSMUSG00000022757	*TFG*	yes		E, L	Dom	T
ENSG00000105619	ENSMUSG00000006335	*TFPT*	yes		L	Dom	T
ENSG00000072274	ENSMUSG00000022797	*TFRC*	yes		L	Dom	T
ENSG00000054118	ENSMUSG00000072862	*THRAP3*	yes		M	Dom	T
ENSG00000054118	ENSMUSG00000043962	*THRAP3*	yes		M	Dom	T
ENSG00000107807	ENSMUSG00000025215	*TLX1*	yes		L	Dom	T
ENSG00000164438	ENSMUSG00000040610	*TLX3*	yes		L	Dom	T
ENSG00000184012	ENSMUSG00000000385	*TMPRSS2*	yes		E	Dom	T
ENSG00000048462	ENSMUSG00000022496	*TNFRSF17*	yes		L	Dom	T
ENSG00000198900	ENSMUSG00000070544	*TOP1*	yes		L	Dom	T
ENSG00000141510	ENSMUSG00000059552	*TP53*	yes	yes	L, E, M, O	Rec	Mis, N, F
ENSG00000167460	ENSMUSG00000031799	*TPM4*	yes		L	Dom	T
ENSG00000047410	ENSMUSG00000006005	*TPR*	yes		E	Dom	T
ENSG00000122779	ENSMUSG00000029833	*TRIM24*	yes		L	Dom	T
ENSG00000197323	ENSMUSG00000033014	*TRIM33*	yes		E	Dom	T
ENSG00000100815	ENSMUSG00000021188	*TRIP11*	yes		L	Dom	T
ENSG00000165699	ENSMUSG00000026812	*TSC1*		yes	E, O	Rec	D, Mis, N, F, S
ENSG00000103197	ENSMUSG00000002496	*TSC2*		yes	E, O	Rec	D, Mis, N, F, S
ENSG00000165409	ENSMUSG00000020963	*TSHR*	yes	yes	E	Dom	Mis
ENSG00000114999	ENSMUSG00000027394	*TTL*	yes		L	Dom	T
ENSG00000129204	ENSMUSG00000000804	*USP6*	yes		M	Dom	T
ENSG00000134086	ENSMUSG00000033933	*VHL*	yes	yes	E, M, O	Rec	D, Mis, N, F, S
ENSG00000015285	ENSMUSG00000031165	*WAS*			L	X-linked Rec	Mis, N, F, S
ENSG00000109685	ENSMUSG00000057406	*WHSC1*	yes		L	Dom	T
ENSG00000147548	ENSMUSG00000054823	*WHSC1L1*	yes		L	Dom	T

ENSG00000165392	ENSMUSG00000031583	*WRN*		yes	L, E, M, O	Rec	Mis, N, F, S
ENSG00000184937	ENSMUSG00000016458	*WT1*	yes	yes	O	Rec	D, Mis, N, F, S
ENSG00000136936	ENSMUSG00000028329	*XPA*		yes	E	Rec	Mis, N, F, S
ENSG00000154767	ENSMUSG00000030094	*XPC*		yes	E	Rec	Mis, N, F, S
ENSG00000109906	ENSMUSG00000066687	*ZBTB16*	yes		L	Dom	T
ENSG00000121741	ENSMUSG00000021945	*ZMYM2*	yes		L	Dom	T
ENSG00000126746	ENSMUSG00000038346	*ZNF384*	yes		L	Dom	T

Appendix B1. *Sleeping Beauty* CISs and predicted CIS genes obtained using the kernel convolution-based framework with a kernel width of 30 kb.

"None" is used where there is no obvious candidate gene associated with the CIS.
"CIS chr" and "CIS position" are the chromosome and base pair coordinates of the CIS, which correspond to the centre of the peak in insertion density identified by the kernel convolution-based method.
"Biotype" indicates whether the gene encodes a protein or miRNA.
"Real" indicates whether the gene is believed to be a real cancer gene candidate.

CIS chr	CIS position	Gene name	Ensembl ID	Biotype	Real?
1	161802810	*Tnr*	ENSMUSG00000015829	protein_coding	N
1	162750720	None	N/A	N/A	N
1	162810060	*Rc3h1*	ENSMUSG00000040423	protein_coding	N
1	162918420	*Cenpl*	ENSMUSG00000026708	protein_coding	N
1	164494053	*Q3UWS6_mouse*	ENSMUSG00000073516	protein_coding	N
2	26289038	*Notch1*	ENSMUSG00000026923	protein_coding	Y
2	98464591		ENSMUSG00000075015	protein_coding	Y
3	136852080	*Ppp3ca*	ENSMUSG00000028161	protein_coding	Y
5	28500825	*En2*	ENSMUSG00000039095	protein_coding	N
7	27307506	*Akt2*	ENSMUSG00000004056	protein_coding	Y
9	32222013	*Fli1*	ENSMUSG00000016087	protein_coding	Y
9	45939227	*BC033915*	ENSMUSG00000034135	protein_coding	Y
10	20825984	*Myb*	ENSMUSG00000019982	protein_coding	Y
11	11653286	*Ikzf1*	ENSMUSG00000018654	protein_coding	Y
13	31634753	*Foxf2*	ENSMUSG00000038402	protein_coding	N
15	3434345	*Ghr*	ENSMUSG00000055737	protein_coding	N
15	3536237	*Ghr*	ENSMUSG00000055737	protein_coding	N
15	28104183	*ENSMUSESTG00003714390*	N/A	EST	N
15	39369050	*Rims2*	ENSMUSG00000037386	protein_coding	N
16	95500025	*Erg*	ENSMUSG00000040732	protein_coding	Y
19	32868157	*Pten*	ENSMUSG00000013663	protein_coding	Y

Appendix B2. MuLV CISs and predicted CIS genes obtained using the kernel convolution-based framework with a kernel width of 30 kb.

"CIS chr" and "CIS" position are the chromosome and base pair coordinates of the CIS, which correspond to the centre of the peak in insertion density identified by the kernel convolution-based method.
"Biotype" indicates whether the gene encodes a protein or miRNA.

CIS chr	CIS position	Gene name	Ensembl ID	Biotype
1	36867188	*Tmem131*	ENSMUSG00000026116	protein_coding
1	37429920	*Mgat4a*	ENSMUSG00000026110	protein_coding
1	88266570	*1700019O17Rik*	ENSMUSG00000036574	protein_coding
1	92886988	*Lrrfip1*	ENSMUSG00000026305	protein_coding
1	93017760	*Ramp1*	ENSMUSG00000034353	protein_coding
1	127266450	*Actr3*	ENSMUSG00000026341	protein_coding
1	130440120	*Cxcr4*	ENSMUSG00000045382	protein_coding
1	135882360	*Btg2*	ENSMUSG00000020423	protein_coding
1	138068640	*5730559C18Rik*	ENSMUSG00000041605	protein_coding
1	139624096	*A130050O07Rik*	ENSMUSG00000051480	protein_coding
1	139988370	*Ptprc*	ENSMUSG00000026395	protein_coding
1	158862420	*ENSMUSG00000073531*	ENSMUSG00000073531	protein_coding
1	163733040	*AI848100*	ENSMUSG00000040297	protein_coding
1	165910570	*Sell*	ENSMUSG00000026581	protein_coding
1	167543910	*Rcsd1*	ENSMUSG00000040723	protein_coding
1	173513310	*Cd48*	ENSMUSG00000015355	protein_coding
1	173513310	*Slamf7*	ENSMUSG00000038179	protein_coding
1	173759850	*Slamf6*	ENSMUSG00000015314	protein_coding
1	174351491	*Ccdc19*	ENSMUSG00000026546	protein_coding
1	182195685	*Itpkb*	ENSMUSG00000038855	protein_coding
1	182545770	*Mixl1*	ENSMUSG00000026497	protein_coding
1	183688021	*EG433384*	ENSMUSG00000056699	protein_coding
1	193218780	*Ppp2r5a*	ENSMUSG00000026626	protein_coding
2	6625263	*Cugbp2*	ENSMUSG00000002107	protein_coding
2	11544848	*Il2ra*	ENSMUSG00000026770	protein_coding
2	18526199		ENSMUSG00000046809	protein_coding
2	26286578	*Notch1*	ENSMUSG00000026923	protein_coding
2	26391458	*Egfl7*	ENSMUSG00000026921	protein_coding
2	28422542	*Gfi1b*	ENSMUSG00000026815	protein_coding
2	29890628	*Set*	ENSMUSG00000054766	protein_coding
2	30074651	*Phyhd1*	ENSMUSG00000007476	protein_coding
2	30478556	*Cstad*	ENSMUSG00000047363	protein_coding
2	30515798	*OTTMUSG00000016805*	ENSMUSG00000075416	protein_coding
2	31883557	*A130092J06Rik*	ENSMUSG00000050592	protein_coding
2	31883557	*Nup214*	ENSMUSG00000001855	protein_coding

2	32463278	*Eng*	ENSMUSG00000026814	protein_coding
2	32463278	*Ak1*	ENSMUSG00000026817	protein_coding
2	35322278	*Ggta1*	ENSMUSG00000035778	protein_coding
2	44747181	*Zeb2*	ENSMUSG00000026872	protein_coding
2	44933888	*OTTMUSG00000012358*	ENSMUSG00000052248	protein_coding
2	45088748	*OTTMUSG00000012358*	ENSMUSG00000052248	protein_coding
2	62199128	*Dpp4*	ENSMUSG00000035000	protein_coding
2	72205228	*B230120H23Rik*	ENSMUSG00000004085	protein_coding
2	90880286	*Slc39a13*	ENSMUSG00000002105	protein_coding
2	103740068	*Lmo2*	ENSMUSG00000032698	protein_coding
2	117035648	*Rasgrp1*	ENSMUSG00000027347	protein_coding
2	117102556	*Rasgrp1*	ENSMUSG00000027347	protein_coding
2	126612068	*2010106G01Rik*	ENSMUSG00000027366	protein_coding
2	131712974	*Rassf2*	ENSMUSG00000027339	protein_coding
2	132201420	*ENSMUSG00000074787*	ENSMUSG00000074787	protein_coding
2	152476268	*Bcl2l1*	ENSMUSG00000007659	protein_coding
2	156577123	*Sla2*	ENSMUSG00000027636	protein_coding
2	156894698	*Ndrg3*	ENSMUSG00000027634	protein_coding
2	158413178	*Ppp1r16b*	ENSMUSG00000037754	protein_coding
2	163335883	*Serinc3*	ENSMUSG00000017707	protein_coding
2	163772315	*Stk4*	ENSMUSG00000018209	protein_coding
2	165576218	*Prkcbp1*	ENSMUSG00000039671	protein_coding
2	165645467	*Ncoa3*	ENSMUSG00000027678	protein_coding
2	167395179	*A530013C23Rik*	ENSMUSG00000006462	protein_coding
2	167496748	*A530013C23Rik*	ENSMUSG00000006462	protein_coding
2	167541311	*A530013C23Rik*	ENSMUSG00000006462	protein_coding
2	169859468	*Zfp217*	ENSMUSG00000052056	protein_coding
3	9136653	*Tpd52*	ENSMUSG00000027506	protein_coding
3	9865266	*Pag1*	ENSMUSG00000027508	protein_coding
3	30204130	*Evi1*	ENSMUSG00000027684	protein_coding
3	30293163	*Evi1*	ENSMUSG00000027684	protein_coding
3	86192646	*Sh3d19*	ENSMUSG00000028082	protein_coding
3	88228899	*Mef2d*	ENSMUSG00000001419	protein_coding
3	89472216	*Zbtb7b*	ENSMUSG00000028042	protein_coding
3	94464096	*Rorc*	ENSMUSG00000028150	protein_coding
3	95758692	*Mcl1*	ENSMUSG00000038612	protein_coding
3	96052686	*Anp32e*	ENSMUSG00000015749	protein_coding
3	101406666	*Cd2*	ENSMUSG00000027863	protein_coding
3	102859296	*Tspan2*	ENSMUSG00000027858	protein_coding
3	103913316	*Hipk1*	ENSMUSG00000008730	protein_coding
3	106915709	*Cd53*	ENSMUSG00000040747	protein_coding
3	107215144	*A930002I21Rik*	ENSMUSG00000050179	protein_coding
3	115671696	*Edg1*	ENSMUSG00000045092	protein_coding
3	130929576	*Lef1*	ENSMUSG00000027985	protein_coding
3	130965711	*Lef1*	ENSMUSG00000027985	protein_coding
3	135539739	*Nfkb1*	ENSMUSG00000028163	protein_coding
3	152867598	*St6galnac5*	ENSMUSG00000039037	protein_coding
4	3880368	*Chchd7*	ENSMUSG00000042198	protein_coding
4	8849854	*Chd7*	ENSMUSG00000041235	protein_coding
4	11076294	*Trp53inp1*	ENSMUSG00000028211	protein_coding
4	43483494	*Cd72*	ENSMUSG00000028459	protein_coding
4	46597074	*Coro2a*	ENSMUSG00000028337	protein_coding
4	59536494	*A2AN91_MOUSE*	ENSMUSG00000038578	protein_coding
4	62886131	*Akna*	ENSMUSG00000039158	protein_coding
4	106332684	*Acot11*	ENSMUSG00000034853	protein_coding
4	106408884	*Ssbp3*	ENSMUSG00000061887	protein_coding
4	117956382	*Mpl*	ENSMUSG00000006389	protein_coding
4	124626818	*Zc3h12a*	ENSMUSG00000042677	protein_coding
4	129061985	*Lck*	ENSMUSG00000000409	protein_coding
4	129313443	*Ptp4a2*	ENSMUSG00000028788	protein_coding
4	132248912	*Fgr*	ENSMUSG00000028874	protein_coding
4	132364674	*Wasf2*	ENSMUSG00000028868	protein_coding
4	132943044	*Pigv*	ENSMUSG00000043257	protein_coding
4	133027869	*Arid1a*	ENSMUSG00000007880	protein_coding
4	133088622	*Arid1a*	ENSMUSG00000007880	protein_coding
4	133384494	*Ubxd5*	ENSMUSG00000012126	protein_coding
4	133674693	*Stmn1*	ENSMUSG00000028832	protein_coding
4	134350074	*Runx3*	ENSMUSG00000070691	protein_coding
4	134479185	*Runx3*	ENSMUSG00000070691	protein_coding
4	135302163	*Tceb3*	ENSMUSG00000028668	protein_coding
4	135361554	*Rpl11*	ENSMUSG00000059291	protein_coding
4	135449154	*E2f2*	ENSMUSG00000018983	protein_coding
4	139870464	*Arhgef10l*	ENSMUSG00000040964	protein_coding
4	140215854	*Padi2*	ENSMUSG00000028927	protein_coding
4	144278324	*Vps13d*	ENSMUSG00000020220	protein_coding
4	148545129	*Pik3cd*	ENSMUSG00000039936	protein_coding
4	149758314	*Park7*	ENSMUSG00000028964	protein_coding
4	153385044	*Prdm16*	ENSMUSG00000039410	protein_coding
4	154073934	*Ski*	ENSMUSG00000029050	protein_coding
4	154914733	*Ttll10*	ENSMUSG00000029074	protein_coding
5	34003443	*Fgfr3*	ENSMUSG00000054252	protein_coding
5	64884363	*Tbc1d1*	ENSMUSG00000029174	protein_coding
5	65091206	*Klf3*	ENSMUSG00000029178	protein_coding
5	66139051	*Rhoh*	ENSMUSG00000029204	protein_coding
5	66198603	*Rhoh*	ENSMUSG00000029204	protein_coding
5	76017843	*Kit*	ENSMUSG00000005672	protein_coding
5	76117974	*Kdr*	ENSMUSG00000062960	protein_coding
5	78014613	*Paics*	ENSMUSG00000029247	protein_coding
5	100814883	*Plac8*	ENSMUSG00000029322	protein_coding
5	105886383	*Lrrc8c*	ENSMUSG00000054720	protein_coding
5	107894133	*Tgfbr3*	ENSMUSG00000029287	protein_coding
5	107974475	*Evi5*	ENSMUSG00000011831	protein_coding
5	107974475	*Gfi1*	ENSMUSG00000029275	protein_coding
5	111746943	*C130026L21Rik*	ENSMUSG00000052848	protein_coding
5	111746943	*NP_001074704.1*	ENSMUSG00000070576	protein_coding

5	115425603	*2410014A08Rik*	ENSMUSG00000048578	protein_coding
5	115786225	*Pxn*	ENSMUSG00000029528	protein_coding
5	122492313	*Hvcn1*	ENSMUSG00000064267	protein_coding
5	123399652	*4932422M17Rik*	ENSMUSG00000062946	protein_coding
5	124332723	*Abcb9*	ENSMUSG00000029408	protein_coding
5	124482875	*Pitpnm2*	ENSMUSG00000029406	protein_coding
5	124740393	*6330548G22Rik*	ENSMUSG00000029402	protein_coding
5	136456199	*Orai2*	ENSMUSG00000039747	protein_coding
5	136908642	*Mylc2pl*	ENSMUSG00000005474	protein_coding
5	137925903	*Hrbl*	ENSMUSG00000029722	protein_coding
5	137967993	*6430598A04Rik*	ENSMUSG00000045348	protein_coding
5	139635153	*3110082I17Rik*	ENSMUSG00000053553	protein_coding
5	140035023	*Mafk*	ENSMUSG00000018143	protein_coding
5	140464293	*Mad1l1*	ENSMUSG00000029554	protein_coding
5	140850741	*Lfng*	ENSMUSG00000029570	protein_coding
5	142625943	*Sdk1*	ENSMUSG00000039683	protein_coding
5	147673695	*Flt3*	ENSMUSG00000042817	protein_coding
5	149285433	*Katnal1*	ENSMUSG00000041298	protein_coding
6	29247822	*2310016C08Rik*	ENSMUSG00000043421	protein_coding
6	31094170		ENSMUSG00000052894	protein_coding
6	31149400	*AB041803*	ENSMUSG00000044471	protein_coding
6	34876806	*Stra8*	ENSMUSG00000029848	protein_coding
6	34876806	*2010107G12Rik*	ENSMUSG00000029847	protein_coding
6	48618880	*Gimap4*	ENSMUSG00000054435	protein_coding
6	48618880	*Gimap6*	ENSMUSG00000047867	protein_coding
6	52145472	*Hoxa7*	ENSMUSG00000038236	protein_coding
6	52574318	*Hibadh*	ENSMUSG00000029776	protein_coding
6	52836602	*Jazf1*	ENSMUSG00000063568	protein_coding
6	72313232	*Vamp8*	ENSMUSG00000050732	protein_coding
6	91077280	*Nup210*	ENSMUSG00000030091	protein_coding
6	99231728	*Foxp1*	ENSMUSG00000030067	protein_coding
6	99350416	*Foxp1*	ENSMUSG00000030067	protein_coding
6	115936008	*Plxnd1*	ENSMUSG00000030123	protein_coding
6	117854680	*Hnrpf*	ENSMUSG00000042079	protein_coding
6	120536200	*Cecr5*	ENSMUSG00000058979	protein_coding
6	120896025	*Bid*	ENSMUSG00000004446	protein_coding
6	120896025	*BC030863*	ENSMUSG00000051586	protein_coding
6	121193790	*Tuba8*	ENSMUSG00000030137	protein_coding
6	125199490	*Cd27*	ENSMUSG00000030336	protein_coding
6	127120398	*Ccnd2*	ENSMUSG00000000184	protein_coding
6	127120398		ENSMUSG00000067988	protein_coding
6	127218280		ENSMUSG00000072757	protein_coding
6	127299190	*ENSMUSG00000072756*	ENSMUSG00000072756	protein_coding
6	129195516	*Clec2d*	ENSMUSG00000030157	protein_coding
6	134158510	*Etv6*	ENSMUSG00000030199	protein_coding
6	134894031		ENSMUSG00000072697	protein_coding
6	136905370	*Arhgdib*	ENSMUSG00000030220	protein_coding
6	146471350	*Itpr2*	ENSMUSG00000030287	protein_coding
7	24088656	*Kcnn4*	ENSMUSG00000054342	protein_coding
7	25365205	*Exosc5*	ENSMUSG00000061286	protein_coding
7	28499665	*Sirt2*	ENSMUSG00000015149	protein_coding
7	37893655	*1600014C10Rik*	ENSMUSG00000054676	protein_coding
7	45787645	*Emp3*	ENSMUSG00000040212	protein_coding
7	46707317	*Ldha*	ENSMUSG00000063229	protein_coding
7	63884467	*Trpm1*	ENSMUSG00000030523	protein_coding
7	68149465	*C330024D12Rik*	ENSMUSG00000030553	protein_coding
7	73477495	*Chd2*	ENSMUSG00000025788	protein_coding
7	75456595	*Akap13*	ENSMUSG00000066406	protein_coding
7	75669385	*Klhl25*	ENSMUSG00000055652	protein_coding
7	79917073	*Zfp710*	ENSMUSG00000048897	protein_coding
7	80018807	*Sema4b*	ENSMUSG00000030539	protein_coding
7	80078065	*Sema4b*	ENSMUSG00000030539	protein_coding
7	83533892	*Il16*	ENSMUSG00000001741	protein_coding
7	100958305	*Fchsd2*	ENSMUSG00000030691	protein_coding
7	101212495	*Stard10*	ENSMUSG00000030688	protein_coding
7	101438455	*Art2b*	ENSMUSG00000030651	protein_coding
7	110744965	*Mrvi1*	ENSMUSG00000005611	protein_coding
7	113952958	*Rras2*	ENSMUSG00000055723	protein_coding
7	114089008	*Psma1*	ENSMUSG00000030751	protein_coding
7	125282161	*Nsmce1*	ENSMUSG00000030750	protein_coding
7	125421045	*Il21r*	ENSMUSG00000030745	protein_coding
7	126145721	*Lat*	ENSMUSG00000030742	protein_coding
7	126935011	*Spn*	ENSMUSG00000051457	protein_coding
7	126975488	*Spn*	ENSMUSG00000051457	protein_coding
7	129572365	*Fgfr2*	ENSMUSG00000030849	protein_coding
7	135434206	*Ptpre*	ENSMUSG00000041836	protein_coding
7	144752546	*Ccnd1*	ENSMUSG00000070348	protein_coding
7	144785845	*Ccnd1*	ENSMUSG00000070348	protein_coding
7	144855464	*Ccnd1*	ENSMUSG00000070348	protein_coding
7	144917305	*Ccnd1*	ENSMUSG00000070348	protein_coding
8	10928751	*3930402G23Rik*	ENSMUSG00000038917	protein_coding
8	26567421	*Plekha2*	ENSMUSG00000031557	protein_coding
8	86631381	*Cd97*	ENSMUSG00000002885	protein_coding
8	87090501	*mmu-mir-181c*	ENSMUSG00000065483	miRNA
8	87090501	*mmu-mir-23a*	ENSMUSG00000065611	miRNA
8	87550590	*Ier2*	ENSMUSG00000053560	protein_coding
8	87615283	*Nfix*	ENSMUSG00000001911	protein_coding
8	97863021	*Gpr56*	ENSMUSG00000031785	protein_coding
8	108503751	*2310066E14Rik*	ENSMUSG00000038604	protein_coding
8	114480441	*Znrf1*	ENSMUSG00000033545	protein_coding
8	118116064	*Wwox*	ENSMUSG00000004637	protein_coding
8	123373041	*Gse1*	ENSMUSG00000031822	protein_coding
8	123424581	*Gse1*	ENSMUSG00000031822	protein_coding
8	125353476	*Rnf166*	ENSMUSG00000014470	protein_coding
8	125592411	*Cbfa2t3*	ENSMUSG00000006362	protein_coding

8	126293180	*Tcf25*	ENSMUSG00000001472	protein_coding
8	129550574	*Irf2bp2*	ENSMUSG00000051495	protein_coding
8	129725139	*A630001O12Rik*	ENSMUSG00000074025	protein_coding
9	3015310		ENSMUSG00000074565	protein_coding
9	7253103		ENSMUSG00000076378	miRNA
9	14649123		ENSMUSG00000059658	protein_coding
9	32260042	*Fli1*	ENSMUSG00000016087	protein_coding
9	32361388	*Ets1*	ENSMUSG00000032035	protein_coding
9	32677124	*Ets1*	ENSMUSG00000032035	protein_coding
9	43776933	*Thy1*	ENSMUSG00000032011	protein_coding
9	43989290	*Cbl*	ENSMUSG00000034342	protein_coding
9	44252133	*Bcl9l*	ENSMUSG00000063382	protein_coding
9	44402613	*Treh*	ENSMUSG00000032098	protein_coding
9	44741223	*Cd3e*	ENSMUSG00000032093	protein_coding
9	57440949	*Csk*	ENSMUSG00000032312	protein_coding
9	58049343	*Pml*	ENSMUSG00000036986	protein_coding
9	60795243	*ENSMUSG00000074256*	ENSMUSG00000074256	protein_coding
9	60925102	*ENSMUSG00000074256*	ENSMUSG00000074256	protein_coding
9	63303429	*Iqch*	ENSMUSG00000037801	protein_coding
9	65342463		ENSMUSG00000066510	protein_coding
9	69262560	*Anxa2*	ENSMUSG00000032231	protein_coding
9	70960083	*Aqp9*	ENSMUSG00000032204	protein_coding
9	72192475	*Mns1*	ENSMUSG00000032221	protein_coding
9	96761643	*Spsb4*	ENSMUSG00000046997	protein_coding
9	107835483	*Ube1l*	ENSMUSG00000032596	protein_coding
9	108912277	*Scotin*	ENSMUSG00000025647	protein_coding
9	112063923	*Arpp21*	ENSMUSG00000032503	protein_coding
10	20683844	*Ahi1*	ENSMUSG00000019986	protein_coding
10	20769944	*Myb*	ENSMUSG00000019982	protein_coding
10	20926098		ENSMUSG00000059894	protein_coding
10	43510904		ENSMUSG00000071320	protein_coding
10	57764624	*Lims1*	ENSMUSG00000019920	protein_coding
10	59562434	*Chst3*	ENSMUSG00000057337	protein_coding
10	59771804	*4632428N05Rik*	ENSMUSG00000020101	protein_coding
10	61900544	*Srgn*	ENSMUSG00000020077	protein_coding
10	79255214	*Ptbp1*	ENSMUSG00000006498	protein_coding
10	79331066	*Arid3a*	ENSMUSG00000019564	protein_coding
10	79559572	*Midn*	ENSMUSG00000035621	protein_coding
10	80085434	*Mknk2*	ENSMUSG00000020190	protein_coding
10	80085434	*Mobkl2a*	ENSMUSG00000003348	protein_coding
10	80333647	*Gadd45b*	ENSMUSG00000015312	protein_coding
10	80729264	*Tbxa2r*	ENSMUSG00000034881	protein_coding
10	80906834	*Gna15*	ENSMUSG00000034792	protein_coding
10	83148554	*Appl2*	ENSMUSG00000020263	protein_coding
10	92559224	*Pctk2*	ENSMUSG00000020015	protein_coding
10	120086745	*ENSMUSG00000074675*	ENSMUSG00000074675	protein_coding
10	127706760	*Usp52*	ENSMUSG00000005682	protein_coding
11	11613558	*Ikzf1*	ENSMUSG00000018654	protein_coding
11	22713536	*B3gnt2*	ENSMUSG00000051650	protein_coding
11	24147350	*Bcl11a*	ENSMUSG00000000861	protein_coding
11	48925646	*Olfr56*	ENSMUSG00000040328	protein_coding
11	49090466	*Mgat1*	ENSMUSG00000020346	protein_coding
11	51485800	*C330016O10Rik*	ENSMUSG00000001053	protein_coding
11	51918686	*Cdkl3*	ENSMUSG00000020389	protein_coding
11	52162416	*Vdac1*	ENSMUSG00000020402	protein_coding
11	52162416	*Tcf7*	ENSMUSG00000000782	protein_coding
11	54965779	*Slc36a3*	ENSMUSG00000049491	protein_coding
11	58249556	*OTTMUSG00000005737*	ENSMUSG00000058287	protein_coding
11	59445765	*AA536749*	ENSMUSG00000005417	protein_coding
11	62365226	*Prr6*	ENSMUSG00000018509	protein_coding
11	68016822	*Ntn1*	ENSMUSG00000020902	protein_coding
11	68246923	*Pik3r5*	ENSMUSG00000020901	protein_coding
11	74830706	*Smg6*	ENSMUSG00000038290	protein_coding
11	74991866	*Ovca2*	ENSMUSG00000038268	protein_coding
11	75094676	*Rtn4rl1*	ENSMUSG00000045287	protein_coding
11	75352973	*Slc43a2*	ENSMUSG00000038178	protein_coding
11	77334143	*1300007F04Rik*	ENSMUSG00000000686	protein_coding
11	77603558	*Myo18a*	ENSMUSG00000000631	protein_coding
11	78820796	*Lgals9*	ENSMUSG00000001123	protein_coding
11	78820796	*Ksr1*	ENSMUSG00000018334	protein_coding
11	79340216	*Evi2b*	ENSMUSG00000070354	protein_coding
11	86223896	*Med13*	ENSMUSG00000034297	protein_coding
11	86402486	*Tmem49*	ENSMUSG00000018171	protein_coding
11	86709049	*Dhx40*	ENSMUSG00000018425	protein_coding
11	87560521	*mmu-mir-142*	ENSMUSG00000065420	miRNA
11	87560521	*Rnf43*	ENSMUSG00000034177	protein_coding
11	88890550	*Gm525*	ENSMUSG00000072553	protein_coding
11	98292206	*Ikzf3*	ENSMUSG00000018168	protein_coding
11	98561306	*Thra*	ENSMUSG00000058756	protein_coding
11	98760020	*Rara*	ENSMUSG00000037992	protein_coding
11	98971556	*Ccr7*	ENSMUSG00000037944	protein_coding
11	100220216	*Jup*	ENSMUSG00000001552	protein_coding
11	100220216	*1110036O03Rik*	ENSMUSG00000006931	protein_coding
11	100665059	*Stat5b*	ENSMUSG00000020919	protein_coding
11	100665059	*Stat5a*	ENSMUSG00000004043	protein_coding
11	102990952	*Fmnl1*	ENSMUSG00000055805	protein_coding
11	106488626	*Pecam1*	ENSMUSG00000020717	protein_coding
11	106564886	*Pecam1*	ENSMUSG00000020717	protein_coding
11	114976084	*Rab37*	ENSMUSG00000020732	protein_coding
11	115250456	*Kctd2*	ENSMUSG00000016940	protein_coding
11	115719645	*Recql5*	ENSMUSG00000020752	protein_coding
11	115907726	*Trim47*	ENSMUSG00000020773	protein_coding
11	116175931	*Rnf157*	ENSMUSG00000052949	protein_coding
11	116421746	*Rhbdf2*	ENSMUSG00000020806	protein_coding
11	117141136	*Sept9*	ENSMUSG00000059248	protein_coding

11	119660186	*4932417H02Rik*	ENSMUSG00000025583	protein_coding
11	121502319	*Metrnl*	ENSMUSG00000039208	protein_coding
11	121502319	*B3gntl1*	ENSMUSG00000046605	protein_coding
12	12591194	*D12Ertd553e*	ENSMUSG00000020589	protein_coding
12	12963050	*Mycn*	ENSMUSG00000037169	protein_coding
12	74496944	*Prkch*	ENSMUSG00000021108	protein_coding
12	78104114	*Fut8*	ENSMUSG00000021065	protein_coding
12	85570394	*Tmem90a*	ENSMUSG00000071234	protein_coding
12	86433840	*Jundm2*	ENSMUSG00000034271	protein_coding
12	86521836	*Jundm2*	ENSMUSG00000034271	protein_coding
12	87718814		ENSMUSG00000061115	protein_coding
12	103748676	*Asb2*	ENSMUSG00000021200	protein_coding
12	105408674	*4831426I19Rik*	ENSMUSG00000054150	protein_coding
12	107534684		ENSMUSG00000059313	protein_coding
12	108374384	*Bcl11b*	ENSMUSG00000048251	protein_coding
12	113124164	*Akt1*	ENSMUSG00000001729	protein_coding
12	113308480	*Gpr132*	ENSMUSG00000021298	protein_coding
13	13322617	*Lyst*	ENSMUSG00000019726	protein_coding
13	13322617	*Nid1*	ENSMUSG00000005397	protein_coding
13	19354435	*Stard3nl*	ENSMUSG00000003062	protein_coding
13	28955545	*Sox19*	ENSMUSG00000076431	protein_coding
13	30759835	*Irf4*	ENSMUSG00000021356	protein_coding
13	30836965	*Exoc2*	ENSMUSG00000021357	protein_coding
13	37869265	*Rreb1*	ENSMUSG00000039087	protein_coding
13	51756303	*Sema4d*	ENSMUSG00000021451	protein_coding
13	52096246	*Gadd45g*	ENSMUSG00000021453	protein_coding
13	56170108	*BC027057*	ENSMUSG00000049625	protein_coding
13	64424644	*Ccrk*	ENSMUSG00000021483	protein_coding
13	113122455	*Map3k1*	ENSMUSG00000021754	protein_coding
13	113544085	*Il6st*	ENSMUSG00000021756	protein_coding
14	14966028		ENSMUSG00000071576	protein_coding
14	24133166	*Zmiz1*	ENSMUSG00000007817	protein_coding
14	24278597	*Zmiz1*	ENSMUSG00000007817	protein_coding
14	24350627	*Zmiz1*	ENSMUSG00000007817	protein_coding
14	26187647	*Arhgef3*	ENSMUSG00000021895	protein_coding
14	30246581	*Sh3bp5*	ENSMUSG00000021892	protein_coding
14	30385342	*Colq*	ENSMUSG00000057606	protein_coding
14	39863417	*Tspan14*	ENSMUSG00000037824	protein_coding
14	53134547	*Dad1*	ENSMUSG00000022174	protein_coding
14	54145686	*Jph4*	ENSMUSG00000022208	protein_coding
14	59679497	*Spata13*	ENSMUSG00000021990	protein_coding
14	68797904	*1700081D17Rik*	ENSMUSG00000022085	protein_coding
14	71858837	*Rcbtb2*	ENSMUSG00000022106	protein_coding
14	73916687	*Lcp1*	ENSMUSG00000021998	protein_coding
14	78008237	*1190002H23Rik*	ENSMUSG00000022018	protein_coding
14	78257370	*Elf1*	ENSMUSG00000036461	protein_coding
14	113921387	*mmu-mir-17*	ENSMUSG00000065508	miRNA
14	117778812	*Cldn10a*	ENSMUSG00000022132	protein_coding
14	121096162	*Ubac2*	ENSMUSG00000041765	protein_coding
14	121142627	*Ubac2*	ENSMUSG00000041765	protein_coding
15	6538277	*Fyb*	ENSMUSG00000022148	protein_coding
15	8506397	*Slc1a3*	ENSMUSG00000005360	protein_coding
15	9413447	*Capsl*	ENSMUSG00000039676	protein_coding
15	61814465	*Myc*	ENSMUSG00000022346	protein_coding
15	61866994	*Pvt1*	ENSMUSG00000072566	protein_coding
15	61924727	*Pvt1*	ENSMUSG00000072566	protein_coding
15	62002017	*Pvt1*	ENSMUSG00000072566	protein_coding
15	62287787	*Pvt1*	ENSMUSG00000072566	protein_coding
15	62562732	*Pvt1*	ENSMUSG00000072566	protein_coding
15	63088217		ENSMUSG00000063435	protein_coding
15	63481547		ENSMUSG00000069082	protein_coding
15	66641961	*Sla*	ENSMUSG00000022372	protein_coding
15	73562505	*Ptp4a3*	ENSMUSG00000059895	protein_coding
15	74753147		ENSMUSG00000034596	protein_coding
15	78420675	*Pscd4*	ENSMUSG00000018008	protein_coding
15	80394347	*Grap2*	ENSMUSG00000042351	protein_coding
15	81855512	*Xrcc6*	ENSMUSG00000022471	protein_coding
15	83428377	*Scube1*	ENSMUSG00000016763	protein_coding
15	84144702	*Parvg*	ENSMUSG00000022439	protein_coding
15	96478907	*Slc38a1*	ENSMUSG00000023169	protein_coding
15	97669457	*Hdac7a*	ENSMUSG00000022475	protein_coding
15	103086193	*Nfe2*	ENSMUSG00000058794	protein_coding
16	4472427	*Tcfap4*	ENSMUSG00000005718	protein_coding
16	8659080	*Usp7*	ENSMUSG00000022710	protein_coding
16	10587300	*Clec16a*	ENSMUSG00000068663	protein_coding
16	17885456	*Vpreb2*	ENSMUSG00000059280	protein_coding
16	32108280	*Bex6*	ENSMUSG00000075269	protein_coding
16	32439060	*Zdhhc19*	ENSMUSG00000052363	protein_coding
16	49720413	*Cd47*	ENSMUSG00000055447	protein_coding
16	55872427	*Rpl24*	ENSMUSG00000022601	protein_coding
16	60578594	*Epha6*	ENSMUSG00000055540	protein_coding
16	91341300	*Ifnar1*	ENSMUSG00000022967	protein_coding
16	92638260	*Runx1*	ENSMUSG00000022952	protein_coding
16	92692779	*Runx1*	ENSMUSG00000022952	protein_coding
16	92755807	*Runx1*	ENSMUSG00000022952	protein_coding
16	92815635	*Runx1*	ENSMUSG00000022952	protein_coding
16	93041760	*mmu-mir-802*	ENSMUSG00000076457	miRNA
16	93080556	*mmu-mir-802*	ENSMUSG00000076457	miRNA
16	93705556	*Dopey2*	ENSMUSG00000022946	protein_coding
16	95560470	*Erg*	ENSMUSG00000040732	protein_coding
17	6423331	*Vil2*	ENSMUSG00000052397	protein_coding
17	14738010	*Tcte3*	ENSMUSG00000036648	protein_coding
17	15382728	*Chd1*	ENSMUSG00000023852	protein_coding
17	25521271	*BC008155*	ENSMUSG00000057411	protein_coding
17	27282631	*Hmga1*	ENSMUSG00000046711	protein_coding

17	28241401	*E230001N04Rik*	ENSMUSG00000066170	protein_coding
17	29125068	*Fgd2*	ENSMUSG00000024013	protein_coding
17	29221740	*Pim1*	ENSMUSG00000024014	protein_coding
17	30355057	*Dnahc8*	ENSMUSG00000033826	protein_coding
17	30796931	*Abcg1*	ENSMUSG00000024030	protein_coding
17	30927241	*Ubash3a*	ENSMUSG00000042345	protein_coding
17	30927241	*Tmprss3*	ENSMUSG00000024034	protein_coding
17	33731492	*Brd2*	ENSMUSG00000024335	protein_coding
17	33809236	*Tap2*	ENSMUSG00000024339	protein_coding
17	33809236	*Psmb8*	ENSMUSG00000024338	protein_coding
17	34656136	*Msh5*	ENSMUSG00000007035	protein_coding
17	34814198	*Lta*	ENSMUSG00000024402	protein_coding
17	34814198	*Nfkbil1*	ENSMUSG00000042419	protein_coding
17	34960681	*H2-D1*	ENSMUSG00000073411	protein_coding
17	35524109	*Ddr1*	ENSMUSG00000003534	protein_coding
17	35630144	*Ppp1r10*	ENSMUSG00000039220	protein_coding
17	44245111	*Supt3h*	ENSMUSG00000038954	protein_coding
17	44956111	*Aars2*	ENSMUSG00000023938	protein_coding
17	46991279	*Ccnd3*	ENSMUSG00000034165	protein_coding
17	47060071	*Ccnd3*	ENSMUSG00000034165	protein_coding
17	47504543	*1700122O11Rik*	ENSMUSG00000042494	protein_coding
17	55928551	*Arrdc5*	ENSMUSG00000073380	protein_coding
17	71390787	*Q8BP09_MOUSE*	ENSMUSG00000041913	protein_coding
17	83869302	*Haao*	ENSMUSG00000000673	protein_coding
17	84056736	*Zfp36l2*	ENSMUSG00000045817	protein_coding
18	4333434	*Map3k8*	ENSMUSG00000024235	protein_coding
18	5352149	*Zfp438*	ENSMUSG00000050945	protein_coding
18	15349211	*Aqp4*	ENSMUSG00000024411	protein_coding
18	35081394	*Hspa9*	ENSMUSG00000024359	protein_coding
18	35868041	*Tmem173*	ENSMUSG00000024349	protein_coding
18	39149231	*Arhgap26*	ENSMUSG00000036452	protein_coding
18	55115767	*Zfp608*	ENSMUSG00000052713	protein_coding
18	60880541	*Rps14*	ENSMUSG00000024608	protein_coding
18	61000211	*Tcof1*	ENSMUSG00000024613	protein_coding
18	65035691	*Nedd4l*	ENSMUSG00000024589	protein_coding
18	68349821	*D18Ertd653e*	ENSMUSG00000024544	protein_coding
18	70687614	*Mbd2*	ENSMUSG00000024513	protein_coding
18	75397751	*Dym*	ENSMUSG00000035765	protein_coding
18	85017491	*Cyb5*	ENSMUSG00000024646	protein_coding
19	4144414	*Gpr152*	ENSMUSG00000044724	protein_coding
19	4299940	*Adrbk1*	ENSMUSG00000024858	protein_coding
19	5810614	*Frmd8*	ENSMUSG00000043488	protein_coding
19	5810614	*Scyl1*	ENSMUSG00000024941	protein_coding
19	5844306	*Frmd8*	ENSMUSG00000024816	protein_coding
19	6399309	*Rasgrp2*	ENSMUSG00000032946	protein_coding
19	6399309	*Pygm*	ENSMUSG00000032648	protein_coding
19	9110974	*Ahnak*	ENSMUSG00000069833	protein_coding
19	10649744	*Cybasc3*	ENSMUSG00000034445	protein_coding
19	25054444	*Dock8*	ENSMUSG00000052085	protein_coding
19	36321364	*Pcgf5*	ENSMUSG00000024805	protein_coding
19	37556224	*Hhex*	ENSMUSG00000024986	protein_coding
19	37556224	*Exoc6*	ENSMUSG00000053799	protein_coding
19	41890234	*Frat2*	ENSMUSG00000047604	protein_coding
19	44437354	*Scd1*	ENSMUSG00000037071	protein_coding
19	47024957	*Ina*	ENSMUSG00000034336	protein_coding
19	53487874	*ENSMUSG00000074788*	ENSMUSG00000074788	protein_coding

Appendix C1. List showing other cancer-associated datasets in which the MuLV CIS genes appear.

RTCGD (retro) = genes identified by retroviral insertional mutagenesis in the RTCGD database (RETRO = CIS gene in RTCGD dataset)
RTCGD (SB) = genes identified by transposon-mediated insertional mutagenesis in the RTCGD database (SB = CIS gene in RTCGD dataset)
CGC = Cancer Gene Census (Dom = CIS gene is a dominant cancer gene)
COSMIC = genes mutated in COSMIC database (COSMIC = CIS gene mutated in COSMIC)
Sjoblom = candidate cancer genes from Sjöblom *et al*. (2006) (Sjoblom = CIS gene in Sjoblom dataset)
Nanog BS = genes with Nanog binding site (Nanog = CIS gene in Nanog dataset)
Oct4 BS = genes with Oct4 binding site (OCT4 = CIS gene in Oct4 dataset)
p53 BS = genes with p53 binding site (p53 (dn) = CIS gene downregulated by p53; p53 (up) = CIS gene upregulated by p53; p53 = effect of p53 binding unknown)
Mullighan Amps/Dels = genes in recurrent amplicons and deletions in acute lymphoblastic leukaemia dataset of Mullighan *et al*. (2007) (AMP = CIS gene in amplicon; DEL = CIS gene in deletion)
n/a is used where the gene has no known human orthologue and cannot be compared to human cancer-associated datasets.

		RTCGD	RTCGD				Nanog	Oct4	p53	Mullighan	
Gene name	Ensembl ID	(retro)	(SB)	CGC	COSMIC	Sjoblom	BS	BS	BS	Amps /	Dels
1110036O03Rik	ENSMUSG00000006931	-	-	-	-	-	-	-	-	-	-
1190002H23Rik	ENSMUSG00000022018	-	-	-	COSMIC	-	-	-	-	-	-
1300007F04Rik	ENSMUSG00000000686	-	-	-	-	-	-	-	-	-	-
1600014C10Rik	ENSMUSG00000054676	-	-	-	-	-	-	-	-	-	-
1700019O17Rik	ENSMUSG00000036574	-	-	-	-	-	Nanog	-	-	-	-
1700081D17Rik	ENSMUSG00000022085	-	-	-	-	-	-	-	-	-	-
1700122O11Rik	ENSMUSG00000042494	-	-	n/a	n/a	n/a	-	-	n/a	n/a	n/a
2010106G01Rik	ENSMUSG00000027366	-	-	-	-	-	-	-	-	-	-
2010107G12Rik	ENSMUSG00000029847	-	-	n/a	n/a	n/a	-	-	n/a	n/a	n/a
2310016C08Rik	ENSMUSG00000043421	-	-	-	-	-	-	-	-	-	-
2310066E14Rik	ENSMUSG00000038604	-	-	-	-	-	-	-	-	-	DEL
2410014A08Rik	ENSMUSG00000048578	-	-	-	-	-	Nanog	-	-	-	-
3110082I17Rik	ENSMUSG00000053553	-	-	-	-	-	-	-	-	-	DEL
3930402G23Rik	ENSMUSG00000038917	RETRO	-	n/a	n/a	n/a	-	-	n/a	n/a	n/a
4632428N05Rik	ENSMUSG00000020101	-	-	-	-	-	-	-	-	-	-
4831426I19Rik	ENSMUSG00000054150	RETRO	-	-	-	-	Nanog	-	-	-	-
4932417H02Rik	ENSMUSG00000025583	-	-	-	COSMIC	-	-	OCT4	-	-	-
4932422M17Rik	ENSMUSG00000062946	-	-	n/a	n/a	n/a	-	-	n/a	n/a	n/a
5730559C18Rik	ENSMUSG00000041605	-	-	-	-	-	-	-	-	AMP	-
6330548G22Rik	ENSMUSG00000029402	-	-	-	-	-	-	-	-	-	-
6430598A04Rik	ENSMUSG00000045348	-	-	-	-	-	-	-	-	-	-
A130050O07Rik	ENSMUSG00000051480	-	-	n/a	n/a	n/a	-	-	n/a	n/a	n/a
A130092J06Rik	ENSMUSG00000050592	-	-	-	-	-	-	-	-	AMP	-
A2AN91_MOUSE	ENSMUSG00000038578	-	-	-	-	-	-	-	p53	AMP	-
A530013C23Rik	ENSMUSG00000006462	-	-	n/a	n/a	n/a	-	-	n/a	n/a	n/a
A630001O12Rik	ENSMUSG00000074025	-	-	n/a	n/a	n/a	-	-	n/a	n/a	n/a
A930002I21Rik	ENSMUSG00000050179	-	-	n/a	n/a	n/a	-	-	n/a	n/a	n/a
AA536749	ENSMUSG00000005417	-	-	-	-	-	-	-	-	-	DEL
Aars2	ENSMUSG00000023938	-	-	-	-	-	-	-	-	-	-
AB041803	ENSMUSG00000044471	RETRO	-	n/a	n/a	n/a	Nanog	-	n/a	n/a	n/a
Abcb9	ENSMUSG00000029408	-	-	-	-	-	-	-	p53	-	-
Abcg1	ENSMUSG00000024030	-	-	-	-	-	-	-	-	-	-
Acot11	ENSMUSG00000034853	-	-	-	-	-	-	-	-	-	-
Actr3	ENSMUSG00000026341	-	-	-	-	-	-	-	-	-	-
Adrbk1	ENSMUSG00000024858	-	-	-	COSMIC	-	-	-	-	-	-
Ahi1	ENSMUSG00000019986	RETRO	-	-	COSMIC	-	Nanog	-	-	AMP	-
Ahnak	ENSMUSG00000069833	-	-	-	-	-	-	-	-	-	-
AI848100	ENSMUSG00000040297	-	-	-	-	-	-	-	-	AMP	-
Ak1	ENSMUSG00000026817	-	-	-	-	-	-	-	-	AMP	-
Akap13	ENSMUSG00000066406	RETRO	-	n/a	n/a	n/a	-	-	n/a	n/a	n/a
Akna	ENSMUSG00000039158	-	-	-	-	-	-	-	-	AMP	-
Akt1	ENSMUSG00000001729	RETRO	-	-	-	-	-	-	-	-	-
Anp32e	ENSMUSG00000015749	-	-	-	-	-	-	-	-	AMP	-
Anxa2	ENSMUSG00000032231	-	-	-	-	-	-	-	-	-	DEL
Appl2	ENSMUSG00000020263	-	-	-	-	-	-	-	-	-	-
Aqp4	ENSMUSG00000024411	-	-	-	-	-	-	-	-	-	-
Aqp9	ENSMUSG00000032204	-	-	-	-	-	-	-	-	-	-
Arhgap26	ENSMUSG00000036452	-	-	Dom	-	-	-	OCT4	-	-	-
Arhgdib	ENSMUSG00000030220	-	-	-	-	-	-	-	-	-	-
Arhgef10l	ENSMUSG00000040964	-	-	-	-	-	-	OCT4	-	-	-
Arhgef3	ENSMUSG00000021895	-	-	-	-	-	Nanog	-	-	-	-
Arid1a	ENSMUSG00000007880	RETRO	-	-	COSMIC	-	-	-	-	-	-
Arid3a	ENSMUSG00000019564	-	-	-	-	-	-	-	-	-	DEL
Arpp21	ENSMUSG00000032503	-	-	-	-	-	-	-	-	-	-
Arrdc5	ENSMUSG00000073380	-	-	-	-	-	-	-	-	-	-
Art2b	ENSMUSG00000030651	-	-	n/a	n/a	n/a	-	-	n/a	n/a	n/a
Asb2	ENSMUSG00000021200	-	-	-	-	-	-	-	-	-	-
B230120H23Rik	ENSMUSG00000004085	-	-	-	COSMIC	-	-	-	-	-	-
B3gnt2	ENSMUSG00000051650	-	-	-	-	-	-	-	-	-	-
B3gntl1	ENSMUSG00000046605	-	-	-	-	-	-	-	-	-	-
BC008155	ENSMUSG00000057411	-	-	-	-	-	-	-	-	-	-

BC027057	ENSMUSG00000049625	-	-	-	-	-	-	-	-	-	-
BC030863	ENSMUSG00000051586	-	-	-	-	-	-	-	-	AMP	-
Bcl11a	ENSMUSG00000000861	RETRO	-	Dom	COSMIC	Sjoblom	-	-	-	-	-
Bcl11b	ENSMUSG00000048251	-	-	Dom	COSMIC	-	-	-	-	-	-
Bcl2l1	ENSMUSG00000007659	RETRO	-	-	-	-	Nanog	-	-	-	DEL
Bcl9l	ENSMUSG00000063382	-	-	-	-	-	-	-	-	-	DEL
Bex6	ENSMUSG00000075269	-	-	-	COSMIC	-	-	-	-	-	-
Bid	ENSMUSG00000004446	RETRO	-	-	-	-	-	-	-	AMP	-
Brd2	ENSMUSG00000024335	RETRO	-	-	COSMIC	-	-	OCT4	-	-	-
Btg2	ENSMUSG00000020423	RETRO	-	-	-	-	-	-	-	AMP	-
C130026L21Rik	ENSMUSG00000052848	-	-	n/a	n/a	n/a	-	-	n/a	n/a	n/a
C330016O10Rik	ENSMUSG00000001053	RETRO	-	-	-	-	-	-	-	-	DEL
C330024D12Rik	ENSMUSG00000030553	-	-	-	-	-	-	-	-	-	-
Capsl	ENSMUSG00000039676	-	-	-	-	-	-	-	-	-	-
Cbfa2t3	ENSMUSG00000006362	RETRO	-	Dom	-	-	-	-	-	-	-
Cbl	ENSMUSG00000034342	-	-	Dom	-	-	-	-	-	-	-
Ccdc19	ENSMUSG00000026546	-	-	-	-	-	-	-	-	AMP	-
Ccnd1	ENSMUSG00000070348	RETRO	-	Dom	-	-	-	-	-	-	-
Ccnd2	ENSMUSG00000000184	RETRO	-	Dom	-	-	-	-	-	-	-
Ccnd3	ENSMUSG00000034165	RETRO	-	Dom	-	-	-	-	-	-	-
Ccr7	ENSMUSG00000037944	RETRO	-	-	-	-	-	-	-	-	-
Ccrk	ENSMUSG00000021483	RETRO	-	-	-	-	-	-	-	AMP	-
Cd2	ENSMUSG00000027863	-	-	-	-	-	-	-	-	-	-
Cd27	ENSMUSG00000030336	-	-	-	-	-	-	-	-	-	-
Cd3e	ENSMUSG00000032093	-	-	-	-	-	-	-	-	-	-
Cd47	ENSMUSG00000055447	-	-	-	-	-	-	-	-	-	-
Cd48	ENSMUSG00000015355	-	-	-	-	-	-	-	-	AMP	-
Cd53	ENSMUSG00000040747	-	-	-	-	-	-	-	-	-	-
Cd72	ENSMUSG00000028459	-	-	-	-	-	-	-	-	-	DEL
Cd97	ENSMUSG00000002885	-	-	-	-	-	-	OCT4	-	-	-
Cdkl3	ENSMUSG00000020389	-	-	-	COSMIC	-	-	-	-	-	-
Cecr5	ENSMUSG00000058979	RETRO	-	-	-	-	Nanog	-	-	AMP	-
Chchd7	ENSMUSG00000042198	-	-	-	-	-	-	-	p53	-	-
Chd1	ENSMUSG00000023852	-	-	-	COSMIC	-	Nanog	OCT4	p53	-	-
Chd2	ENSMUSG00000025788	RETRO	-	n/a	n/a	n/a	-	-	n/a	n/a	n/a
Chd7	ENSMUSG00000041235	-	-	-	-	-	Nanog	-	-	-	-
Chst3	ENSMUSG00000057337	-	-	-	-	-	Nanog	OCT4	-	-	-
Cldn10a	ENSMUSG00000022132	-	-	-	-	-	-	-	-	-	-
Clec16a	ENSMUSG00000068663	-	-	-	-	-	-	-	-	-	-
Clec2d	ENSMUSG00000030157	-	-	-	-	-	-	-	-	-	-
Colq	ENSMUSG00000057606	-	-	-	-	-	-	-	-	-	-
Coro2a	ENSMUSG00000028337	RETRO	-	-	-	-	-	-	-	AMP	-
Csk	ENSMUSG00000032312	-	-	-	COSMIC	-	-	-	-	-	-
Cstad	ENSMUSG00000047363	RETRO	-	n/a	n/a	n/a	-	-	n/a	n/a	n/a
Cugbp2	ENSMUSG00000002107	-	-	-	-	-	-	OCT4	-	AMP	-
Cxcr4	ENSMUSG00000045382	-	-	-	COSMIC	-	-	-	-	-	-
Cyb5	ENSMUSG00000024646	-	-	-	-	-	Nanog	-	-	-	-
Cybasc3	ENSMUSG00000034445	-	-	-	-	-	-	-	-	-	-
D12Ertd553e	ENSMUSG00000020589	RETRO	-	-	-	-	Nanog	-	-	AMP	-
D18Ertd653e	ENSMUSG00000024544	-	-	-	-	-	-	-	-	-	-
Dad1	ENSMUSG00000022174	-	-	-	-	-	-	-	-	-	-
Ddr1	ENSMUSG00000003534	RETRO	-	-	COSMIC	-	-	-	-	-	-
Dhx40	ENSMUSG00000018425	-	-	-	-	-	-	-	-	-	-
Dnahc8	ENSMUSG00000033826	-	-	-	COSMIC	-	Nanog	OCT4	-	-	-
Dock8	ENSMUSG00000052085	-	-	-	-	-	-	-	-	-	DEL
Dopey2	ENSMUSG00000022946	RETRO	-	-	-	-	Nanog	-	-	-	-
Dpp4	ENSMUSG00000035000	-	-	-	-	-	-	-	-	-	-
Dym	ENSMUSG00000035765	RETRO	-	-	-	-	-	-	-	-	-
E230001N04Rik	ENSMUSG00000066170	-	-	n/a	n/a	n/a	-	-	n/a	n/a	n/a
E2f2	ENSMUSG00000018983	RETRO	-	-	-	-	-	-	-	-	-
Edg1	ENSMUSG00000045092	-	-	-	-	-	Nanog	-	-	-	-
EG433384	ENSMUSG00000056699	-	-	n/a	n/a	n/a	-	-	n/a	n/a	n/a
Egfl7	ENSMUSG00000026921	-	-	-	-	-	-	-	-	AMP	-
Elf1	ENSMUSG00000036461	-	-	-	-	-	-	-	-	-	DEL
Emp3	ENSMUSG00000040212	-	-	-	-	-	-	-	-	-	-
Eng	ENSMUSG00000026814	-	-	-	-	-	-	-	-	AMP	-
Epha6	ENSMUSG00000055540	-	-	n/a	n/a	n/a	-	-	n/a	n/a	n/a
Erg	ENSMUSG00000040732	RETRO	SB	Dom	-	-	-	-	-	-	DEL
Ets1	ENSMUSG00000032035	RETRO	SB	-	COSMIC	-	-	-	-	-	-
Etv6	ENSMUSG00000030199	-	-	Dom	-	-	-	-	-	-	DEL
Evi1	ENSMUSG00000027684	RETRO	-	Dom	COSMIC	-	-	-	-	-	-
Evi2b	ENSMUSG00000070354	-	-	-	-	-	-	-	-	-	-
Evi5	ENSMUSG00000011831	-	-	-	-	-	-	-	-	-	-
Exoc2	ENSMUSG00000021357	-	-	-	COSMIC	-	-	-	-	AMP	-
Exoc6	ENSMUSG00000053799	-	-	-	-	-	-	-	-	-	-
Exosc5	ENSMUSG00000061286	-	-	-	-	-	-	-	-	-	-
Fchsd2	ENSMUSG00000030691	-	-	-	-	-	-	-	-	-	-
Fgd2	ENSMUSG00000024013	-	-	-	COSMIC	-	Nanog	-	-	-	-
Fgfr2	ENSMUSG00000030849	-	-	Dom	COSMIC	-	-	-	-	-	-
Fgfr3	ENSMUSG00000054252	RETRO	-	Dom	COSMIC	-	-	-	-	-	-
Fgr	ENSMUSG00000028874	RETRO	-	-	COSMIC	-	Nanog	-	-	-	-
Fli1	ENSMUSG00000016087	RETRO	SB	Dom	COSMIC	-	-	-	-	-	-
Flt3	ENSMUSG00000042817	RETRO	-	Dom	COSMIC	-	-	-	-	-	-
Fmnl1	ENSMUSG00000055805	-	-	-	-	-	Nanog	-	-	-	-
Foxp1	ENSMUSG00000030067	-	-	-	COSMIC	-	Nanog	-	-	-	-
Frat2	ENSMUSG00000047604	-	-	-	-	-	-	-	-	-	-
Frmd8	ENSMUSG00000024816	RETRO	-	-	COSMIC	-	-	-	-	-	-
Frmd8	ENSMUSG00000043488	RETRO	-	n/a	n/a	n/a	-	-	n/a	n/a	n/a
Fut8	ENSMUSG00000021065	-	-	-	-	-	-	-	-	-	-
Fyb	ENSMUSG00000022148	-	-	-	-	-	Nanog	OCT4	-	-	-
Gadd45b	ENSMUSG00000015312	-	-	-	-	-	-	-	-	-	-
Gadd45g	ENSMUSG00000021453	RETRO	-	-	-	-	-	OCT4	-	AMP	-
Gfi1	ENSMUSG00000029275	RETRO	-	-	-	-	-	-	-	-	-
Gfi1b	ENSMUSG00000026815	RETRO	-	-	-	-	-	-	-	AMP	-
Ggta1	ENSMUSG00000035778	RETRO	-	n/a	n/a	n/a	-	-	n/a	n/a	n/a

Gimap4	ENSMUSG00000054435	-	-	-	-	-	-	-	-	-	-
Gimap6	ENSMUSG00000047867	-	-	-	-	-	-	-	-	-	-
Gm525	ENSMUSG00000072553	-	-	-	-	-	-	-	-	-	-
Gna15	ENSMUSG00000034792	-	-	-	-	-	-	-	-	-	-
Gpr132	ENSMUSG00000021298	-	-	-	-	-	-	-	-	-	-
Gpr152	ENSMUSG00000044724	-	-	-	-	-	-	-	-	-	-
Gpr56	ENSMUSG00000031785	-	-	-	-	-	-	-	-	-	-
Grap2	ENSMUSG00000042351	-	-	-	-	-	Nanog	OCT4	-	-	-
Gse1	ENSMUSG00000031822	RETRO	-	-	COSMIC	-	-	-	-	-	-
H2-D1	ENSMUSG00000073411	-	-	n/a	n/a	n/a	-	-	n/a	n/a	n/a
Haao	ENSMUSG00000000673	-	-	-	-	-	-	-	p53	-	-
Hdac7a	ENSMUSG00000022475	-	-	-	COSMIC	-	-	-	-	-	-
Hhex	ENSMUSG00000024986	RETRO	-	-	-	-	-	OCT4	-	-	-
Hibadh	ENSMUSG00000029776	-	-	-	-	-	Nanog	-	-	-	DEL
Hipk1	ENSMUSG00000008730	-	-	-	COSMIC	-	-	-	-	-	-
Hmga1	ENSMUSG00000046711	RETRO	-	Dom	-	-	-	-	-	-	-
Hnrpf	ENSMUSG00000042079	-	-	-	-	-	-	-	-	-	-
Hoxa7	ENSMUSG00000038236	RETRO	-	-	-	-	-	-	-	-	DEL
Hrbl	ENSMUSG00000029722	-	-	-	-	-	-	-	-	-	-
Hspa9	ENSMUSG00000024359	RETRO	-	-	-	-	-	-	-	-	-
Hvcn1	ENSMUSG00000064267	-	-	-	-	-	-	-	-	-	-
Ier2	ENSMUSG00000053560	RETRO	-	-	-	-	-	-	-	-	-
Ifnar1	ENSMUSG00000022967	RETRO	-	-	-	-	Nanog	-	p53	AMP	-
Ikzf1	ENSMUSG00000018654	RETRO	-	Dom	-	-	-	-	-	-	DEL
Ikzf3	ENSMUSG00000018168	RETRO	-	-	COSMIC	-	-	OCT4	-	-	DEL
Il16	ENSMUSG00000001741	RETRO	-	-	COSMIC	-	-	-	-	-	-
Il21r	ENSMUSG00000030745	-	-	Dom	-	-	-	-	-	-	-
Il2ra	ENSMUSG00000026770	RETRO	-	-	-	-	-	-	-	AMP	-
Il6st	ENSMUSG00000021756	RETRO	-	-	-	-	Nanog	OCT4	-	-	-
Ina	ENSMUSG00000034336	-	-	-	-	-	Nanog	-	-	-	-
Iqch	ENSMUSG00000037801	-	-	-	-	-	-	-	-	-	-
Irf2bp2	ENSMUSG00000051495	RETRO	-	-	-	-	-	-	p53	AMP	-
Irf4	ENSMUSG00000021356	-	-	Dom	-	-	-	-	-	AMP	-
Itpkb	ENSMUSG00000038855	-	-	-	-	-	-	-	-	AMP	-
Itpr2	ENSMUSG00000030287	-	-	-	COSMIC	-	-	-	-	-	-
Jazf1	ENSMUSG00000063568	RETRO	-	Dom	-	-	-	-	-	-	DEL
Jph4	ENSMUSG00000022208	-	-	-	-	-	-	-	-	-	-
Jundm2	ENSMUSG00000034271	RETRO	-	-	-	-	-	-	-	-	-
Jup	ENSMUSG00000001552	-	-	-	-	-	-	-	-	-	-
Katnal1	ENSMUSG00000041298	-	-	-	-	-	-	-	-	-	-
Kcnn4	ENSMUSG00000054342	-	-	-	-	-	-	-	-	-	-
Kctd2	ENSMUSG00000016940	RETRO	-	-	-	-	-	OCT4	-	-	-
Kdr	ENSMUSG00000062960	RETRO	-	-	COSMIC	-	-	-	-	-	-
Kit	ENSMUSG00000005672	-	-	Dom	COSMIC	-	-	-	-	-	-
Klf3	ENSMUSG00000029178	RETRO	-	-	COSMIC	-	-	-	-	-	-
Klhl25	ENSMUSG00000055652	-	-	-	-	-	-	-	-	-	-
Ksr1	ENSMUSG00000018334	-	-	-	COSMIC	-	-	-	-	-	-
Lat	ENSMUSG00000030742	-	-	-	-	-	-	-	-	-	-
Lck	ENSMUSG00000000409	-	-	Dom	COSMIC	-	-	-	-	-	-
Lcp1	ENSMUSG00000021998	RETRO	-	Dom	-	-	-	-	-	-	-
Ldha	ENSMUSG00000063229	-	-	-	-	-	-	-	-	-	-
Lef1	ENSMUSG00000027985	RETRO	-	-	-	-	Nanog	-	-	-	DEL
Lfng	ENSMUSG00000029570	RETRO	-	-	-	-	-	-	-	-	DEL
Lgals9	ENSMUSG00000001123	-	-	-	-	-	-	-	-	-	DEL
Lims1	ENSMUSG00000019920	-	-	n/a	n/a	n/a	-	-	n/a	n/a	n/a
Lmo2	ENSMUSG00000032698	RETRO	-	n/a	n/a	n/a	-	-	n/a	n/a	n/a
Lrrc8c	ENSMUSG00000054720	-	-	-	-	-	-	-	-	-	-
Lrrfip1	ENSMUSG00000026305	-	-	-	-	Sjoblom	Nanog	-	-	-	DEL
Lta	ENSMUSG00000024402	-	-	-	-	-	-	-	-	-	-
Lyst	ENSMUSG00000019726	-	-	-	-	-	Nanog	-	p53	AMP	-
Mad1l1	ENSMUSG00000029554	RETRO	-	-	-	-	-	OCT4	-	-	DEL
Mafk	ENSMUSG00000018143	RETRO	-	-	-	-	-	-	-	-	DEL
Map3k1	ENSMUSG00000021754	-	-	-	COSMIC	-	Nanog	-	-	-	-
Map3k8	ENSMUSG00000024235	RETRO	-	-	-	-	Nanog	-	-	AMP	-
Mbd2	ENSMUSG00000024513	-	-	-	-	-	-	-	-	-	-
Mcl1	ENSMUSG00000038612	RETRO	-	-	-	-	-	-	-	AMP	-
Med13	ENSMUSG00000034297	-	-	-	-	-	-	-	-	-	-
Mef2d	ENSMUSG00000001419	RETRO	-	-	-	-	-	-	-	AMP	-
Metrnl	ENSMUSG00000039208	-	-	-	-	-	-	-	-	-	-
Mgat1	ENSMUSG00000020346	-	-	-	-	-	-	-	-	-	DEL
Mgat4a	ENSMUSG00000026110	-	-	-	-	-	-	-	-	-	-
Midn	ENSMUSG00000035621	-	-	-	-	-	-	-	-	-	DEL
Mixl1	ENSMUSG00000026497	-	-	-	-	-	-	-	-	AMP	-
Mknk2	ENSMUSG00000020190	-	-	-	-	-	-	-	-	-	-
mmu-mir-142	ENSMUSG00000065420	-	-	n/a	n/a	n/a	-	-	n/a	n/a	n/a
mmu-mir-17	ENSMUSG00000065508	-	-	n/a	n/a	n/a	-	-	n/a	n/a	n/a
mmu-mir-181c	ENSMUSG00000065483	-	-	n/a	n/a	n/a	-	-	n/a	n/a	n/a
mmu-mir-23a	ENSMUSG00000065611	-	-	n/a	n/a	n/a	-	-	n/a	n/a	n/a
mmu-mir-802	ENSMUSG00000076457	-	-	n/a	n/a	n/a	-	-	n/a	n/a	n/a
Mns1	ENSMUSG00000032221	-	-	-	-	-	-	-	-	-	-
Mobkl2a	ENSMUSG00000003348	-	-	-	-	-	-	-	-	-	-
Mpl	ENSMUSG00000006389	-	-	Dom	COSMIC	-	-	-	-	-	-
Mrvi1	ENSMUSG00000005611	RETRO	-	-	-	-	-	-	-	-	-
Msh5	ENSMUSG00000007035	-	-	-	-	-	-	OCT4	-	-	-
Myb	ENSMUSG00000019982	RETRO	-	-	-	-	-	-	-	AMP	-
Myc	ENSMUSG00000022346	RETRO	-	Dom	-	-	Nanog	-	p53	-	-
Mycn	ENSMUSG00000037169	RETRO	-	Dom	COSMIC	-	Nanog	OCT4	-	AMP	-
Mylc2pl	ENSMUSG00000005474	-	-	-	-	-	-	-	-	-	-
Myo18a	ENSMUSG00000000631	-	-	-	-	-	-	OCT4	-	-	-
Ncoa3	ENSMUSG00000027678	-	-	-	COSMIC	-	Nanog	-	-	-	DEL
Ndrg3	ENSMUSG00000027634	-	-	-	-	-	-	-	-	-	DEL
Nedd4l	ENSMUSG00000024589	-	-	-	-	-	Nanog	-	p53 (dn)	-	-
Nfe2	ENSMUSG00000058794	-	-	-	-	-	-	-	-	-	-
Nfix	ENSMUSG00000001911	RETRO	-	-	-	-	-	OCT4	-	-	-

Nfkb1	ENSMUSG00000028163	RETRO	-	-	COSMIC	-	-	-	-	-	-
Nfkbil1	ENSMUSG00000042419	-	-	-	-	-	-	-	-	-	-
Nid1	ENSMUSG00000005397	-	-	-	-	-	-	-	-	AMP	-
Notch1	ENSMUSG00000026923	RETRO	SB	Dom	COSMIC	-	-	-	p53 (up)	AMP	-
NP_001074704.1	ENSMUSG00000070576	-	-	Dom	-	-	-	-	-	-	-
Nsmce1	ENSMUSG00000030750	-	-	-	-	-	-	-	-	-	-
Ntn1	ENSMUSG00000020902	-	-	-	-	-	Nanog	-	-	-	DEL
Nup210	ENSMUSG00000030091	-	-	-	COSMIC	-	-	-	-	-	-
Nup214	ENSMUSG00000001855	-	-	Dom	COSMIC	Sjoblom	-	-	-	AMP	-
Olfr56	ENSMUSG00000040328	-	-	n/a	n/a	n/a	-	-	n/a	n/a	n/a
Orai2	ENSMUSG00000039747	RETRO	-	-	-	-	-	-	-	-	-
	ENSMUSG00000058287	-	-	n/a	n/a	n/a	-	-	n/a	n/a	n/a
	ENSMUSG00000052248	-	-	n/a	n/a	n/a	-	-	n/a	n/a	n/a
	ENSMUSG00000075416	-	-	n/a	n/a	n/a	-	-	n/a	n/a	n/a
Ovca2	ENSMUSG00000038268	-	-	-	-	-	-	-	-	-	DEL
Padi2	ENSMUSG00000028927	RETRO	-	-	-	-	-	-	-	-	-
Pag1	ENSMUSG00000027508	-	-	-	-	-	-	-	-	-	-
Paics	ENSMUSG00000029247	-	-	-	-	-	-	-	-	-	-
Park7	ENSMUSG00000028964	-	-	-	-	-	-	-	-	-	-
Parvg	ENSMUSG00000022439	-	-	-	-	-	-	-	-	-	-
Pcgf5	ENSMUSG00000024805	-	-	-	COSMIC	-	-	-	p53	-	-
Pctk2	ENSMUSG00000020015	-	-	-	COSMIC	-	-	-	-	-	-
Pecam1	ENSMUSG00000020717	RETRO	-	-	-	-	Nanog	-	-	-	-
Phyhd1	ENSMUSG00000007476	-	-	-	-	-	-	-	-	AMP	-
Pigv	ENSMUSG00000043257	-	-	-	-	-	Nanog	-	-	-	-
Pik3cd	ENSMUSG00000039936	-	-	-	-	-	-	OCT4	-	-	-
Pik3r5	ENSMUSG00000020901	RETRO	-	-	-	-	-	-	-	-	DEL
Pim1	ENSMUSG00000024014	RETRO	-	Dom	COSMIC	-	Nanog	-	-	-	-
Pitpnm2	ENSMUSG00000029406	-	-	-	-	-	-	-	-	-	-
Plac8	ENSMUSG00000029322	-	-	-	-	-	-	-	-	-	-
Plekha2	ENSMUSG00000031557	RETRO	-	-	-	-	-	-	-	-	-
Plxnd1	ENSMUSG00000030123	-	-	-	-	-	-	-	-	-	-
Pml	ENSMUSG00000036986	-	-	Dom	COSMIC	-	-	OCT4	-	-	-
Ppp1r10	ENSMUSG00000039220	-	-	-	COSMIC	-	-	-	-	-	-
Ppp1r16b	ENSMUSG00000037754	RETRO	-	-	-	-	-	-	-	-	DEL
Ppp2r5a	ENSMUSG00000026626	-	-	-	-	-	-	-	-	AMP	-
Prdm16	ENSMUSG00000039410	RETRO	-	Dom	COSMIC	-	-	-	-	-	-
Prkcbp1	ENSMUSG00000039671	RETRO	-	-	COSMIC	-	-	OCT4	-	-	DEL
Prkch	ENSMUSG00000021108	-	-	-	COSMIC	-	-	-	-	-	-
Prr6	ENSMUSG00000018509	-	-	-	-	-	-	-	-	-	DEL
Pscd4	ENSMUSG00000018008	-	-	-	-	-	-	-	-	-	-
Psma1	ENSMUSG00000030751	-	-	-	-	-	-	-	-	-	-
Psmb8	ENSMUSG00000024338	-	-	-	-	-	-	-	-	-	-
Ptbp1	ENSMUSG00000006498	RETRO	-	-	-	-	-	-	-	-	DEL
Ptp4a2	ENSMUSG00000028788	-	-	-	-	-	-	-	-	-	-
Ptp4a3	ENSMUSG00000059895	RETRO	-	-	-	-	-	-	-	-	-
Ptprc	ENSMUSG00000026395	-	-	-	COSMIC	-	-	-	-	AMP	-
Ptpre	ENSMUSG00000041836	RETRO	-	-	-	-	-	-	p53 (up)	-	-
Pvt1	ENSMUSG00000072566	-	-	n/a	n/a	n/a	-	-	n/a	n/a	n/a
Pxn	ENSMUSG00000029528	RETRO	-	-	-	-	-	-	-	-	-
Pygm	ENSMUSG00000032648	-	-	-	-	-	-	-	-	-	-
Q8BP09_MOUSE	ENSMUSG00000041913	-	-	-	-	-	-	-	-	-	-
Rab37	ENSMUSG00000020732	-	-	-	-	-	Nanog	-	-	-	-
Ramp1	ENSMUSG00000034353	RETRO	-	n/a	n/a	n/a	-	-	n/a	n/a	n/a
Rara	ENSMUSG00000037992	-	-	Dom	COSMIC	-	-	OCT4	-	-	-
Rasgrp1	ENSMUSG00000027347	RETRO	SB	-	-	-	-	-	-	-	-
Rasgrp2	ENSMUSG00000032946	RETRO	-	-	-	-	-	-	-	-	-
Rassf2	ENSMUSG00000027339	-	-	-	-	-	-	-	-	-	-
Rcbtb2	ENSMUSG00000022106	-	-	-	COSMIC	-	-	-	-	-	DEL
Rcsd1	ENSMUSG00000040723	-	-	-	-	-	-	-	-	AMP	-
Recql5	ENSMUSG00000020752	-	-	-	-	-	-	-	-	-	-
Rhbdf2	ENSMUSG00000020806	-	-	-	-	-	-	-	-	-	-
Rhoh	ENSMUSG00000029204	-	-	Dom	-	-	-	-	-	-	-
Rnf157	ENSMUSG00000052949	RETRO	-	-	-	-	-	-	-	-	-
Rnf166	ENSMUSG00000014470	-	-	-	-	-	-	-	-	-	-
Rnf43	ENSMUSG00000034177	-	-	-	-	-	-	-	-	-	-
Rorc	ENSMUSG00000028150	RETRO	-	-	COSMIC	-	-	-	-	AMP	-
Rpl11	ENSMUSG00000059291	-	-	n/a	n/a	n/a	Nanog	-	n/a	n/a	n/a
Rpl24	ENSMUSG00000022601	-	-	-	-	-	-	-	-	-	-
Rps14	ENSMUSG00000024608	-	-	-	-	-	-	-	-	-	-
Rras2	ENSMUSG00000055723	RETRO	-	-	COSMIC	-	-	-	-	-	-
Rreb1	ENSMUSG00000039087	RETRO	-	-	-	-	-	-	-	AMP	-
Rtn4rl1	ENSMUSG00000045287	-	-	-	-	-	-	-	-	-	DEL
Runx1	ENSMUSG00000022952	RETRO	-	Dom	COSMIC	-	-	-	-	AMP	-
Runx3	ENSMUSG00000070691	RETRO	-	-	-	-	-	-	-	-	-
Scd1	ENSMUSG00000037071	-	-	-	-	-	Nanog	-	-	-	-
Scotin	ENSMUSG00000025647	-	-	-	-	-	-	-	-	-	-
Scube1	ENSMUSG00000016763	-	-	-	-	-	-	-	-	-	-
Scyl1	ENSMUSG00000024941	RETRO	-	-	COSMIC	-	-	-	-	-	-
Sdk1	ENSMUSG00000039683	-	-	-	-	-	Nanog	OCT4	-	-	DEL
Sell	ENSMUSG00000026581	-	-	-	-	-	-	-	-	AMP	-
Sema4b	ENSMUSG00000030539	RETRO	-	-	-	-	Nanog	-	-	-	-
Sema4d	ENSMUSG00000021451	-	-	-	-	-	-	-	-	AMP	-
Sept9	ENSMUSG00000059248	RETRO	-	Dom	-	-	-	-	-	-	-
Serinc3	ENSMUSG00000017707	RETRO	-	-	-	-	-	-	-	-	DEL
Set	ENSMUSG00000054766	RETRO	-	n/a	n/a	n/a	-	-	n/a	n/a	n/a
Sh3bp5	ENSMUSG00000021892	-	-	-	-	-	-	-	-	-	-
Sh3d19	ENSMUSG00000028082	-	-	-	-	-	-	-	-	-	-
Sirt2	ENSMUSG00000015149	-	-	-	-	-	-	-	-	-	-
Ski	ENSMUSG00000029050	-	-	-	-	-	-	-	-	-	-
Sla	ENSMUSG00000022372	-	-	-	-	-	-	-	-	-	-
Sla2	ENSMUSG00000027636	-	-	-	-	-	-	-	-	-	DEL
Slamf6	ENSMUSG00000015314	RETRO	-	-	-	-	-	-	-	AMP	-

Slamf7	ENSMUSG00000038179	-	-	-	-	-	-	-	-	AMP	-
Slc1a3	ENSMUSG00000005360	-	-	-	-	-	-	-	-	-	-
Slc36a3	ENSMUSG00000049491	-	-	-	-	-	-	-	-	-	-
Slc38a1	ENSMUSG00000023169	RETRO	-	-	-	-	-	-	-	-	-
Slc39a13	ENSMUSG00000002105	-	-	-	-	-	-	-	-	-	-
Slc43a2	ENSMUSG00000038178	-	-	-	-	-	-	-	-	-	DEL
Smg6	ENSMUSG00000038290	RETRO	-	-	COSMIC	-	-	-	-	-	DEL
Sox19	ENSMUSG00000076431	RETRO	-	-	-	-	-	-	-	AMP	-
Spata13	ENSMUSG00000021990	-	-	-	COSMIC	-	-	-	-	-	-
Spn	ENSMUSG00000051457	-	-	-	-	-	-	-	-	-	-
Spsb4	ENSMUSG00000046997	-	-	-	-	-	-	-	-	-	-
Srgn	ENSMUSG00000020077	-	-	-	-	-	-	-	-	-	-
Ssbp3	ENSMUSG00000061887	RETRO	-	-	-	-	-	-	-	-	-
St6galnac5	ENSMUSG00000039037	-	-	-	-	-	-	-	-	-	-
Stard10	ENSMUSG00000030688	-	-	-	-	-	-	-	-	-	-
Stard3nl	ENSMUSG00000003062	-	-	-	-	-	-	-	-	-	DEL
Stat5a	ENSMUSG00000004043	RETRO	-	-	-	-	-	-	-	-	-
Stat5b	ENSMUSG00000020919	RETRO	-	-	-	-	-	-	-	-	-
Stk4	ENSMUSG00000018209	-	-	-	-	-	-	-	-	-	DEL
Stmn1	ENSMUSG00000028832	-	-	-	-	-	-	-	-	-	-
Stra8	ENSMUSG00000029848	-	-	-	-	-	-	-	-	-	-
Supt3h	ENSMUSG00000038954	-	-	-	-	-	-	-	-	-	-
Tap2	ENSMUSG00000024339	RETRO	-	-	-	-	-	-	-	-	-
Tbc1d1	ENSMUSG00000029174	-	-	-	-	-	Nanog	-	-	-	-
Tbxa2r	ENSMUSG00000034881	-	-	-	-	-	-	-	-	-	-
Tceb3	ENSMUSG00000028668	-	-	-	-	-	-	-	-	-	-
Tcf25	ENSMUSG00000001472	-	-	-	-	-	-	-	-	-	-
Tcf7	ENSMUSG00000000782	-	-	-	-	-	-	OCT4	-	-	-
Tcfap4	ENSMUSG00000005718	RETRO	-	-	-	-	-	-	-	-	-
Tcof1	ENSMUSG00000024613	-	-	-	-	-	-	-	-	-	-
Tcte3	ENSMUSG00000036648	-	-	-	-	-	-	-	-	-	-
Tgfbr3	ENSMUSG00000029287	-	-	-	-	-	-	OCT4	-	-	-
Thra	ENSMUSG00000058756	-	-	-	-	-	-	-	-	-	-
Thy1	ENSMUSG00000032011	-	-	-	-	-	-	-	-	-	-
Tmem131	ENSMUSG00000026116	-	-	-	-	-	-	OCT4	-	-	-
Tmem173	ENSMUSG00000024349	-	-	-	-	-	-	-	-	-	-
Tmem49	ENSMUSG00000018171	-	-	-	-	-	-	-	-	-	-
Tmem90a	ENSMUSG00000071234	-	-	-	-	-	-	-	-	-	-
Tmprss3	ENSMUSG00000024034	-	-	-	-	-	-	-	-	-	-
Tpd52	ENSMUSG00000027506	-	-	-	-	-	Nanog	-	-	-	-
Treh	ENSMUSG00000032098	-	-	-	-	-	-	-	-	-	DEL
Trim47	ENSMUSG00000020773	-	-	-	COSMIC	-	-	-	-	-	-
Trp53inp1	ENSMUSG00000028211	-	-	-	-	-	-	-	-	-	-
Trpm1	ENSMUSG00000030523	-	-	-	-	-	-	-	p53 (dn)	-	-
Tspan14	ENSMUSG00000037824	-	-	-	-	-	-	-	-	-	-
Tspan2	ENSMUSG00000027858	-	-	-	-	-	-	-	-	-	-
Ttll10	ENSMUSG00000029074	-	-	-	-	-	-	-	-	-	-
Tuba8	ENSMUSG00000030137	-	-	-	-	-	-	-	-	AMP	-
Ubac2	ENSMUSG00000041765	-	-	-	-	-	-	-	-	-	-
Ubash3a	ENSMUSG00000042345	-	-	-	-	-	-	-	-	-	-
Ube1l	ENSMUSG00000032596	-	-	-	-	-	-	-	-	-	-
Ubxd5	ENSMUSG00000012126	-	-	-	-	-	-	-	-	-	-
Usp52	ENSMUSG00000005682	-	-	-	COSMIC	-	-	-	-	-	-
Usp7	ENSMUSG00000022710	-	-	-	-	-	-	OCT4	-	-	-
Vamp8	ENSMUSG00000050732	RETRO	-	-	-	-	-	-	-	-	-
Vdac1	ENSMUSG00000020402	-	-	-	-	-	-	-	-	-	-
Vil2	ENSMUSG00000052397	-	-	-	-	-	-	-	-	-	-
Vpreb2	ENSMUSG00000059280	-	-	-	-	-	-	-	-	AMP	-
Vps13d	ENSMUSG00000020220	-	-	-	-	-	-	-	-	-	-
Wasf2	ENSMUSG00000028868	RETRO	-	-	COSMIC	-	Nanog	-	-	-	-
Wwox	ENSMUSG00000004637	-	-	-	-	-	Nanog	OCT4	-	-	-
Xrcc6	ENSMUSG00000022471	-	-	-	COSMIC	-	-	-	-	-	-
Zbtb7b	ENSMUSG00000028042	-	-	-	-	-	-	-	-	AMP	-
Zc3h12a	ENSMUSG00000042677	-	-	-	-	-	Nanog	-	-	-	-
Zdhhc19	ENSMUSG00000052363	-	-	-	-	-	-	-	-	-	-
Zeb2	ENSMUSG00000026872	RETRO	-	-	-	-	Nanog	-	-	-	-
Zfp217	ENSMUSG00000052056	RETRO	-	-	COSMIC	-	Nanog	-	-	-	DEL
Zfp36l2	ENSMUSG00000045817	RETRO	-	-	-	-	-	-	-	-	DEL
Zfp438	ENSMUSG00000050945	-	-	-	-	-	-	-	-	AMP	-
Zfp608	ENSMUSG00000052713	RETRO	-	-	COSMIC	-	Nanog	-	-	-	-
Zfp710	ENSMUSG00000048897	-	-	-	-	-	-	-	-	-	-
Zmiz1	ENSMUSG00000007817	RETRO	SB	-	-	-	-	OCT4	-	-	-
Znrf1	ENSMUSG00000033545	-	-	-	-	-	Nanog	-	-	-	-
	ENSMUSG00000072756	-	-	n/a	n/a	n/a	-	-	n/a	n/a	n/a
	ENSMUSG00000073531	-	-	n/a	n/a	n/a	-	-	n/a	n/a	n/a
	ENSMUSG00000074256	-	-	n/a	n/a	n/a	-	-	n/a	n/a	n/a
	ENSMUSG00000074675	-	-	n/a	n/a	n/a	-	-	n/a	n/a	n/a
	ENSMUSG00000074787	-	-	n/a	n/a	n/a	-	-	n/a	n/a	n/a
	ENSMUSG00000074788	-	-	n/a	n/a	n/a	-	-	n/a	n/a	n/a
	ENSMUSG00000034596	-	-	n/a	n/a	n/a	-	-	n/a	n/a	n/a
	ENSMUSG00000046809	-	-	n/a	n/a	n/a	-	-	n/a	n/a	n/a
	ENSMUSG00000052894	-	-	n/a	n/a	n/a	-	-	n/a	n/a	n/a
	ENSMUSG00000059313	-	-	n/a	n/a	n/a	-	-	n/a	n/a	n/a
	ENSMUSG00000059658	-	-	-	-	-	-	-	-	-	-
	ENSMUSG00000059894	-	-	n/a	n/a	n/a	-	-	n/a	n/a	n/a
	ENSMUSG00000061115	-	-	n/a	n/a	n/a	-	-	n/a	n/a	n/a
	ENSMUSG00000063435	-	-	n/a	n/a	n/a	-	-	n/a	n/a	n/a
	ENSMUSG00000066510	-	-	-	-	-	-	-	-	-	-
	ENSMUSG00000067988	-	-	n/a	n/a	n/a	-	-	n/a	n/a	n/a
	ENSMUSG00000069082	-	-	n/a	n/a	n/a	-	-	n/a	n/a	n/a
	ENSMUSG00000071320	-	-	n/a	n/a	n/a	-	-	n/a	n/a	n/a
	ENSMUSG00000071576	-	-	-	-	-	-	-	-	-	-
	ENSMUSG00000072697	-	-	n/a	n/a	n/a	-	-	n/a	n/a	n/a
	ENSMUSG00000072757	-	-	n/a	n/a	n/a	-	-	n/a	n/a	n/a

	ENSMUSG00000074565	-	-	n/a	n/a	n/a	-	-	n/a	n/a	n/a
	ENSMUSG00000076378	-	-	n/a	n/a	n/a	-	-	n/a	n/a	n/a

Appendix C2. List showing other cancer-associated datasets in which the *Sleeping Beauty* CIS genes appear.

See Appendix C1 for an explanation of each column.

Gene		RTCGD	RTCGD				Nanog	OCT4	p53	Mullighan	
name	Ensembl ID	(retro)	(SB)	CGC	COSMIC	Sjoblom	BS	BS	BS	Amps/	Dels
Akt2	ENSMUSG00000004056	-	-	Dom	-	-	-	-	-	-	-
BC033915	ENSMUSG00000034135	-	-	-	COSMIC	Sjoblom	-	-	-	-	-
Erg	ENSMUSG00000040732	RETRO	SB	Dom	-	-	-	-	-	-	-
Fli1	ENSMUSG00000016087	RETRO	SB	Dom	COSMIC	-	-	-	-	-	-
Ikzf1	ENSMUSG00000018654	RETRO	-	Dom	-	-	-	-	-	-	-
Myb	ENSMUSG00000019982	RETRO	-	-	-	-	-	-	-	AMP	DEL
Notch1	ENSMUSG00000026923	RETRO	SB	Dom	COSMIC	-	-	-	p53 (up)	AMP	DEL
Ppp3ca	ENSMUSG00000028161	-	-	-	COSMIC	-	-	-	-	-	-
Pten	ENSMUSG00000013663	-	-	Rec	COSMIC	-	-	-	-	-	DEL
	ENSMUSG00000075015	-	-	n/a	n/a	n/a	-	-	n/a	n/a	n/a

Appendix D. Nearest and further genes from CISs with a *P*-value of <0.001 or <0.05 in lists supplied by the Netherlands Cancer Institute.

"CIS chr" and "CIS position" are the chromosome and base pair coordinates of the CIS, which correspond to the centre of the peak in insertion density identified by the kernel convolution-based method.

CIS *P*-value	CIS chr	CIS position	Nearest gene name	Nearest gene Ensembl ID	Further gene name	Further gene Ensembl ID
0.001	1	37217235		ENSMUSG00000061123	*Tmem131*	ENSMUSG00000026116
0.001	1	37780290	*Mgat4a*	ENSMUSG00000026110	*2010300C02Rik*	ENSMUSG00000026090
0.001	1	86239852	*1700019O17Rik*	ENSMUSG00000036574		ENSMUSG00000067125
0.001	1	86295450		ENSMUSG00000067125	*OTTMUSG00000007209*	ENSMUSG00000026238
0.001	1	90879243	*Lrrfip1*	ENSMUSG00000026305		ENSMUSG00000048068
0.001	1	91009545	*Ramp1*	ENSMUSG00000034353		ENSMUSG00000047217
0.001	1	125266838	*Actr3*	ENSMUSG00000026341	*Slc35f5*	ENSMUSG00000026342
0.001	1	128455260	*Cxcr4*	ENSMUSG00000045382		ENSMUSG00000064737
0.001	1	133917960	*Btg2*	ENSMUSG00000020423	*Fmod*	ENSMUSG00000041559
0.001	1	136100550	*Camsap1l1*	ENSMUSG00000041570	*GPR25_MOUSE*	ENSMUSG00000052759
0.001	1	137658090	*A130050O07Rik*	ENSMUSG00000051480		ENSMUSG00000065205
0.001	1	156936780	*1700057K13Rik*	ENSMUSG00000026592	*Angptl1*	ENSMUSG00000033544
0.001	1	163991159	*Sell*	ENSMUSG00000026581	*Selp*	ENSMUSG00000026580
0.001	1	165628770	*Creg1*	ENSMUSG00000040713	*Rcsd1*	ENSMUSG00000040723
0.001	1	171607260	*Cd48*	ENSMUSG00000015355	*Slamf7*	ENSMUSG00000038179
0.001	1	171851676	*Slamf6*	ENSMUSG00000015314	*Vangl2*	ENSMUSG00000026556
0.001	1	172443511	*Ccdc19*	ENSMUSG00000026546		ENSMUSG00000066683
0.001	1	180313439	*Itpkb*	ENSMUSG00000038855	*ENSMUSG00000056615*	ENSMUSG00000056615
0.001	1	180660300	*Mixl1*	ENSMUSG00000026497	*Acbd3*	ENSMUSG00000026499
0.001	1	180803760	*BC031781*	ENSMUSG00000038806		ENSMUSG00000066654
0.001	1	191254946	*Dtl*	ENSMUSG00000037474	*Ppp2r5a*	ENSMUSG00000026626
0.001	2	6626911	*Cugbp2*	ENSMUSG00000002107		ENSMUSG00000069177
0.001	2	11543491		ENSMUSG00000069172	*Rbm17*	ENSMUSG00000037197
0.001	2	26397160	*Notch1*	ENSMUSG00000026923	*Egfl7*	ENSMUSG00000026921
0.001	2	26493080	*Egfl7*	ENSMUSG00000026921	*Notch1*	ENSMUSG00000026923
0.001	2	28534226	*Gfi1b*	ENSMUSG00000026815	*Gtf3c5*	ENSMUSG00000026816
0.001	2	29999738	*Set*	ENSMUSG00000054766	*Pkn3*	ENSMUSG00000026785
0.001	2	30588260	*Cstad*	ENSMUSG00000047363	*1700001O22Rik*	ENSMUSG00000044320
0.001	2	30652628	*1700001O22Rik*	ENSMUSG00000044320	*Cstad*	ENSMUSG00000047363
0.001	2	31994682	*A130092J06Rik*	ENSMUSG00000050592	*Nup214*	ENSMUSG00000001855
0.001	2	32573666	*Eng*	ENSMUSG00000026814	*Ak1*	ENSMUSG00000026817
0.001	2	45044311	*OTTMUSG00000012358*	ENSMUSG00000052248		ENSMUSG00000057684
0.001	2	62252693	*Dpp4*	ENSMUSG00000035000	*Gcg*	ENSMUSG00000000394
0.001	2	65193998	*9330158F14Rik*	ENSMUSG00000061171		ENSMUSG00000060270
0.001	2	72102343		ENSMUSG00000068863	*B230120H23Rik*	ENSMUSG00000004085
0.001	2	90786925	*Sfpi1*	ENSMUSG00000002111	*Slc39a13*	ENSMUSG00000002105
0.001	2	103644128	*Lmo2*	ENSMUSG00000032698		ENSMUSG00000056848
0.001	2	116858239	*Rasgrp1*	ENSMUSG00000027347	*Thbs1*	ENSMUSG00000040152
0.001	2	116924160	*Rasgrp1*	ENSMUSG00000027347	*Thbs1*	ENSMUSG00000040152
0.001	2	126843788	*Dusp2*	ENSMUSG00000027368	*Stard7*	ENSMUSG00000027367
0.001	2	131535747	*Rassf2*	ENSMUSG00000027339	*Slc23a2*	ENSMUSG00000027340
0.001	2	132024599	*Prei4*	ENSMUSG00000027346		ENSMUSG00000068247
0.001	2	152242778	*Bcl2l1*	ENSMUSG00000007659	*Tpx2*	ENSMUSG00000027469
0.001	2	156342081	*Sla2*	ENSMUSG00000027636	*Ndrg3*	ENSMUSG00000027634
0.001	2	158180458	*2310007D09Rik*	ENSMUSG00000027654	*Ppp1r16b*	ENSMUSG00000037754
0.001	2	165342117		ENSMUSG00000051789	*Ncoa3*	ENSMUSG00000027678
0.001	2	165410755	*Ncoa3*	ENSMUSG00000027678		ENSMUSG00000051789
0.001	2	166824814	*Slc9a8*	ENSMUSG00000039463	*B4galt5*	ENSMUSG00000017929
0.001	2	167210199	*A530013C23Rik*	ENSMUSG00000006462	*Ptpn1*	ENSMUSG00000027540
0.001	2	167357666	*Ptpn1*	ENSMUSG00000027540	*A530013C23Rik*	ENSMUSG00000006462
0.001	2	169676198	*Zfp217*	ENSMUSG00000052056		ENSMUSG00000055544
0.001	2	180161851	*Slco4a1*	ENSMUSG00000038963	*mmu-mir-133a-2*	ENSMUSG00000065460
0.001	3	29414978	*Evi1*	ENSMUSG00000027684	*Mds1*	ENSMUSG00000051636
0.001	3	29514295	*Mds1*	ENSMUSG00000051636	*Evi1*	ENSMUSG00000027684
0.001	3	51372385		ENSMUSG00000069019	*Maml3*	ENSMUSG00000061143
0.001	3	87886915	*Mef2d*	ENSMUSG00000001419		ENSMUSG00000068928
0.001	3	88367220	*Arhgef2*	ENSMUSG00000028059	*Rxfp4*	ENSMUSG00000049741
0.001	3	89142895	*Zbtb7b*	ENSMUSG00000028042	*Flad1*	ENSMUSG00000042642
0.001	3	93865375	*Rorc*	ENSMUSG00000028150	*Lingo4*	ENSMUSG00000044505
0.001	3	95157949	*Adamtsl4*	ENSMUSG00000015850	*Mcl1*	ENSMUSG00000038612
0.001	3	95452585		ENSMUSG00000057781		ENSMUSG00000058910
0.001	3	100703412	*Cd2*	ENSMUSG00000027863		ENSMUSG00000050461
0.001	3	102156745	*Tspan2*	ENSMUSG00000027858	*Ngfb*	ENSMUSG00000027859
0.001	3	103210135	*Dclre1b*	ENSMUSG00000027845		ENSMUSG00000068800
0.001	3	106883096	*A930002I21Rik*	ENSMUSG00000050179	*AI504432*	ENSMUSG00000056145
0.001	3	114451945	*Edg1*	ENSMUSG00000045092	*Olfm3*	ENSMUSG00000027965
0.001	3	129871645		ENSMUSG00000058642		ENSMUSG00000049622
0.001	3	129907165		ENSMUSG00000058642		ENSMUSG00000049622
0.001	3	134486207	*Nfkb1*	ENSMUSG00000028163	*ENSMUSG00000045520*	ENSMUSG00000045520
0.001	4	3880430	*Chchd7*	ENSMUSG00000042198		ENSMUSG00000061390
0.001	4	11075624	*Trp53inp1*	ENSMUSG00000028211		ENSMUSG00000058164
0.001	4	43376874	*Cd72*	ENSMUSG00000028459		ENSMUSG00000066198
0.001	4	46494350		ENSMUSG00000066194	*Tbc1d2*	ENSMUSG00000039813
0.001	4	59383145	*Rod1*	ENSMUSG00000028382	*A2AN91_MOUSE*	ENSMUSG00000038578
0.001	4	62489630	*Whrn*	ENSMUSG00000039137		ENSMUSG00000039165
0.001	4	105793040	*Ssbp3*	ENSMUSG00000061887	*Acot11*	ENSMUSG00000034853
0.001	4	117418370	*Tie1*	ENSMUSG00000033191	*Mpl*	ENSMUSG00000006389
0.001	4	128586798		ENSMUSG00000066049	*BC030183*	ENSMUSG00000050493
0.001	4	128794580		ENSMUSG00000049089	*Ptp4a2*	ENSMUSG00000028788
0.001	4	131936973	*Fgr*	ENSMUSG00000028874	*Ahdc1*	ENSMUSG00000037692

0.001	4	132051188	*ENSMUSG00000066041*	ENSMUSG00000066041		ENSMUSG00000066040
0.001	4	132629392	*Pigv*	ENSMUSG00000043257	*Arid1a*	ENSMUSG00000007880
0.001	4	132779960		ENSMUSG00000064705	*Rps6ka1*	ENSMUSG00000003644
0.001	4	133074230	*Ubxd5*	ENSMUSG00000012126	*Sh3bgrl3*	ENSMUSG00000028843
0.001	4	134785370	*Cnr2*	ENSMUSG00000062585	*Fuca1*	ENSMUSG00000028673
0.001	4	134910709	*Tceb3*	ENSMUSG00000028668	*Id3*	ENSMUSG00000007872
0.001	4	134971837	*Id3*	ENSMUSG00000007872	*Tceb3*	ENSMUSG00000028668
0.001	4	135058700	*E2f2*	ENSMUSG00000018983	*Ddefl1*	ENSMUSG00000036995
0.001	4	139825430	*Sdhb*	ENSMUSG00000009863		ENSMUSG00000064443
0.001	4	148194053	*D4Ertd429e*	ENSMUSG00000044700	*Pik3cd*	ENSMUSG00000039936
0.001	5	32173336	*Tacc3*	ENSMUSG00000037313	*Fgfr3*	ENSMUSG00000054252
0.001	5	63408106	*Klf3*	ENSMUSG00000029178	*Tbc1d1*	ENSMUSG00000029174
0.001	5	64659750	*Rhoh*	ENSMUSG00000029204		ENSMUSG00000067246
0.001	5	64720625	*Chrna9*	ENSMUSG00000029205	*Rhoh*	ENSMUSG00000029204
0.001	5	74573986		ENSMUSG00000067196	*Kdr*	ENSMUSG00000062960
0.001	5	74669557	*Kdr*	ENSMUSG00000062960		ENSMUSG00000067196
0.001	5	99594826		ENSMUSG00000067084	*Plac8*	ENSMUSG00000029322
0.001	5	104706836	*Lrrc8d*	ENSMUSG00000046079	*Lrrc8c*	ENSMUSG00000054720
0.001	5	106723992	*Rpap2*	ENSMUSG00000033773	*Gfi1*	ENSMUSG00000029275
0.001	5	106805480	*Evi5*	ENSMUSG00000011831	*Gfi1*	ENSMUSG00000029275
0.001	5	110555672	*C130026L21Rik*	ENSMUSG00000052848	*Pitpnb*	ENSMUSG00000050017
0.001	5	112943262	*Coro1c*	ENSMUSG00000004530	*Selplg*	ENSMUSG00000048163
0.001	5	114275290	*Cabp1*	ENSMUSG00000029544	*2410014A08Rik*	ENSMUSG00000048578
0.001	5	114642496	*Pxn*	ENSMUSG00000029528		ENSMUSG00000064401
0.001	5	116270566	*Taok3*	ENSMUSG00000061288	*Pebp1*	ENSMUSG00000032959
0.001	5	122293554	*4932422M17Rik*	ENSMUSG00000062946		ENSMUSG00000066795
0.001	5	122492536	*Bcl7a*	ENSMUSG00000029438	*Wdr66*	ENSMUSG00000029442
0.001	5	123399436		ENSMUSG00000029398	*Pitpnm2*	ENSMUSG00000029406
0.001	5	135184914	*Prkrip1*	ENSMUSG00000039737	*Orai2*	ENSMUSG00000039747
0.001	5	135637883	*Mylc2pl*	ENSMUSG00000005474	*4731417B20Rik*	ENSMUSG00000046548
0.001	5	136645576		ENSMUSG00000066690		ENSMUSG00000007324
0.001	5	137199166		ENSMUSG00000064676	*0910001L09Rik*	ENSMUSG00000050552
0.001	5	138379186	*Gpr146*	ENSMUSG00000044197	*C130050O18Rik*	ENSMUSG00000044092
0.001	5	138777046	*Mafk*	ENSMUSG00000018143	*Ints1*	ENSMUSG00000029547
0.001	5	139597070	*Lfng*	ENSMUSG00000029570	*Grifin*	ENSMUSG00000036586
0.001	5	146254722	*Flt3*	ENSMUSG00000042817		ENSMUSG00000057157
0.001	5	147866815		ENSMUSG00000066552	*Hmgb1*	ENSMUSG00000066551
0.001	5	148368619	*4930588N13Rik*	ENSMUSG00000029660	*6330406I15Rik*	ENSMUSG00000029659
0.001	6	29344053	*BC048651*	ENSMUSG00000039742	*2310016C08Rik*	ENSMUSG00000043421
0.001	6	31202455	*AB041803*	ENSMUSG00000044471		ENSMUSG00000052894
0.001	6	35041709	*2010107G12Rik*	ENSMUSG00000029847	*Cnot4*	ENSMUSG00000038784
0.001	6	48825985	*Gimap6*	ENSMUSG00000047867	*Gimap4*	ENSMUSG00000054435
0.001	6	52360435	*Hoxa7*	ENSMUSG00000038236	*Hoxa9*	ENSMUSG00000038227
0.001	6	52787635		ENSMUSG00000064792	*Tax1bp1*	ENSMUSG00000004535
0.001	6	53055274	*Jazf1*	ENSMUSG00000063568		ENSMUSG00000038157
0.001	6	99854785	*Eif4e3*	ENSMUSG00000030068		ENSMUSG00000024343
0.001	6	121003391	*Cecr5*	ENSMUSG00000058979		ENSMUSG00000040893
0.001	6	121362051	*BC030863*	ENSMUSG00000051586	*Bid*	ENSMUSG00000004446
0.001	6	125891365	*Cd27*	ENSMUSG00000030336		ENSMUSG00000064520
0.001	6	127827619	*Ccnd2*	ENSMUSG00000000184		ENSMUSG00000064191
0.001	6	127922935		ENSMUSG00000064191		ENSMUSG00000057412
0.001	6	128009665		ENSMUSG00000058306		ENSMUSG00000067666
0.001	6	129919277	*Cd69*	ENSMUSG00000030156	*Clec2d*	ENSMUSG00000030157
0.001	6	134950855		ENSMUSG00000020939	*Bcl2l14*	ENSMUSG00000030200
0.001	6	147434665	*4933424B01Rik*	ENSMUSG00000040250		ENSMUSG00000067323
0.001	7	19555460	*Kcnn4*	ENSMUSG00000054342	*1500002O20Rik*	ENSMUSG00000002210
0.001	7	24194070		ENSMUSG00000047345	*Sirt2*	ENSMUSG00000015149
0.001	7	33353130	*1600014C10Rik*	ENSMUSG00000054676		ENSMUSG00000051425
0.001	7	40923804		ENSMUSG00000039955	*Ldha*	ENSMUSG00000063229
0.001	7	62158275	*C330024D12Rik*	ENSMUSG00000030553	*EG384639*	ENSMUSG00000042668
0.001	7	67496820		ENSMUSG00000065729		ENSMUSG00000066411
0.001	7	69487127		ENSMUSG00000025750	*Akap13*	ENSMUSG00000066406
0.001	7	73844970	*Zfp710*	ENSMUSG00000048897	*2610034B18Rik*	ENSMUSG00000039043
0.001	7	73943257	*Sema4b*	ENSMUSG00000030539	*Idh2*	ENSMUSG00000030541
0.001	7	74000850	*Sema4b*	ENSMUSG00000030539	*Cib1*	ENSMUSG00000030538
0.001	7	74632350	*Zscan2*	ENSMUSG00000038797	*Iqgap1*	ENSMUSG00000030536
0.001	7	95190900	*Fchsd2*	ENSMUSG00000030691	*P2ry2*	ENSMUSG00000032860
0.001	7	95447100	*Stard10*	ENSMUSG00000030688	*Centd2*	ENSMUSG00000032812
0.001	7	95667465		ENSMUSG00000066310	*mmu-mir-139*	ENSMUSG00000065446
0.001	7	108014978	*Copb1*	ENSMUSG00000030754	*Rras2*	ENSMUSG00000055723
0.001	7	108155370	*Psma1*	ENSMUSG00000030751	*Pde3b*	ENSMUSG00000030671
0.001	7	112137030	*4930583K01Rik*	ENSMUSG00000055159	*Syt17*	ENSMUSG00000058420
0.001	7	119538513	*Nsmce1*	ENSMUSG00000030750	*Il4ra*	ENSMUSG00000030748
0.001	7	119676960	*Gtf3c1*	ENSMUSG00000032777	*Il21r*	ENSMUSG00000030745
0.001	7	120402570	*Lat*	ENSMUSG00000030742		ENSMUSG00000066180
0.001	7	120757800	*Coro1a*	ENSMUSG00000030707	*Mapk3*	ENSMUSG00000063065
0.001	7	121141669	*AI467606*	ENSMUSG00000045165	*Qprt*	ENSMUSG00000030674
0.001	7	121233233	*Cd2bp2*	ENSMUSG00000042502	*Spn*	ENSMUSG00000051457
0.001	7	121331984		ENSMUSG00000061513	*Sephs2*	ENSMUSG00000049091
0.001	7	123850980	*Brwd2*	ENSMUSG00000042055		ENSMUSG00000065763
0.001	7	129963457	*Ptpre*	ENSMUSG00000041836	*Mki67*	ENSMUSG00000031004
0.001	7	134601900	*Olfr524*	ENSMUSG00000050366	*Cd163l1*	ENSMUSG00000025461
0.001	7	135393702	*Ifitm3*	ENSMUSG00000025492	*Ifitm1*	ENSMUSG00000025491
0.001	7	139355962		ENSMUSG00000031071	*Tpcn2*	ENSMUSG00000048677
0.001	7	139459278		ENSMUSG00000031071	*Tpcn2*	ENSMUSG00000048677
0.001	7	139520040	*Tpcn2*	ENSMUSG00000048677		ENSMUSG00000031071
0.001	8	10298663	*3930402G23Rik*	ENSMUSG00000038917	*Irs2*	ENSMUSG00000038894
0.001	8	83003110		ENSMUSG00000064787	*Cd97*	ENSMUSG00000002885
0.001	8	83464133	*mmu-mir-23a*	ENSMUSG00000065611	*mmu-mir-181c*	ENSMUSG00000065483
0.001	8	83925867	*Ier2*	ENSMUSG00000053560	*Cacna1a*	ENSMUSG00000034656
0.001	8	83990843	*Nfix*	ENSMUSG00000001911	*Dand5*	ENSMUSG00000053226
0.001	8	104916233	*2310066E14Rik*	ENSMUSG00000038604	*Ctcf*	ENSMUSG00000005698
0.001	8	110892610	*Ldhd*	ENSMUSG00000031958	*Znrf1*	ENSMUSG00000033545
0.001	8	116868923	*Plcg2*	ENSMUSG00000034330	*4632417N05Rik*	ENSMUSG00000034308
0.001	8	119895988	*1700016A09Rik*	ENSMUSG00000042840	*Gins2*	ENSMUSG00000031821

0.001	8	121849515	*Rnf166*	ENSMUSG00000014470	*2310061F22Rik*	ENSMUSG00000049482
0.001	8	122086021	*Cbfa2t3*	ENSMUSG00000006362	*BC021611*	ENSMUSG00000015016
0.001	8	122807753		ENSMUSG00000057475	*Def8*	ENSMUSG00000001482
0.001	8	126128303	*Tomm20*	ENSMUSG00000058779		ENSMUSG00000069852
0.001	9	7246771	*Mmp13*	ENSMUSG00000050578	*Dcun1d5*	ENSMUSG00000032002
0.001	9	32432540		ENSMUSG00000064492	*Ets1*	ENSMUSG00000032035
0.001	9	32533657	*Ets1*	ENSMUSG00000032035		ENSMUSG00000064492
0.001	9	44013881	*Thy1*	ENSMUSG00000032011		ENSMUSG00000059979
0.001	9	44227331	*Pdzd3*	ENSMUSG00000032105	*Cbl*	ENSMUSG00000034342
0.001	9	44494631	*Blr1*	ENSMUSG00000047880	*Bcl9l*	ENSMUSG00000063382
0.001	9	44979071	*Cd3d*	ENSMUSG00000032094	*Cd3e*	ENSMUSG00000032093
0.001	9	57755625	*Cyp1a1*	ENSMUSG00000032315	*Csk*	ENSMUSG00000032312
0.001	9	69594821	*Anxa2*	ENSMUSG00000032231	*Foxb1*	ENSMUSG00000059246
0.001	9	71304551	*Aqp9*	ENSMUSG00000032204	*Aldh1a2*	ENSMUSG00000013584
0.001	9	72524010	*Mns1*	ENSMUSG00000032221	*Zfp280d*	ENSMUSG00000038535
0.001	9	108045941	*Ube1l*	ENSMUSG00000032596	*6230427J02Rik*	ENSMUSG00000042106
0.001	9	109100985	*Trex1*	ENSMUSG00000049734	*Scotin*	ENSMUSG00000025647
0.001	9	110949851	*Als2cl*	ENSMUSG00000044037	*Tdgf1*	ENSMUSG00000032494
0.001	9	112261091	*mmu-mir-128b*	ENSMUSG00000065441	*Pdcd6ip*	ENSMUSG00000032504
0.001	9	114407511	*Glb1*	ENSMUSG00000045594	*Ccr4*	ENSMUSG00000047898
0.001	9	123784927	*Ccr9*	ENSMUSG00000029530	*Lztfl1*	ENSMUSG00000025245
0.001	10	19508682	*Ifngr1*	ENSMUSG00000020009	*Olig3*	ENSMUSG00000045591
0.001	10	20924120	*Ahi1*	ENSMUSG00000019986	*Myb*	ENSMUSG00000019982
0.001	10	21009938	*Myb*	ENSMUSG00000019982	*Ahi1*	ENSMUSG00000019986
0.001	10	21125838	*Myb*	ENSMUSG00000019982		ENSMUSG00000059894
0.001	10	21171964		ENSMUSG00000059894	*Myb*	ENSMUSG00000019982
0.001	10	40527010	*Slc22a16*	ENSMUSG00000019834	*Cdc2l6*	ENSMUSG00000038481
0.001	10	60125292	*Chst3*	ENSMUSG00000057337	*Spock2*	ENSMUSG00000058297
0.001	10	62471532	*Srgn*	ENSMUSG00000020077	*2510003E04Rik*	ENSMUSG00000036955
0.001	10	80050543	*Arid3a*	ENSMUSG00000019564	*Wdr18*	ENSMUSG00000035754
0.001	10	80284284	*Cirbp*	ENSMUSG00000045193	*Midn*	ENSMUSG00000035621
0.001	10	80817672	*Mobkl2a*	ENSMUSG00000003348	*NP_082105.3*	ENSMUSG00000055862
0.001	10	81464292	*Tbxa2r*	ENSMUSG00000034881	*Gipc3*	ENSMUSG00000034872
0.001	10	81637110	*Gna15*	ENSMUSG00000034792	*Edg6*	ENSMUSG00000044199
0.001	10	93100182	*Pctk2*	ENSMUSG00000020015	*4930485B16Rik*	ENSMUSG00000020014
0.001	10	120266886	*4921513I03Rik*	ENSMUSG00000044544	*1700006J14Rik*	ENSMUSG00000034764
0.001	10	128040691	*Usp52*	ENSMUSG00000005682		ENSMUSG00000064948
0.001	11	5403337	*Xbp1*	ENSMUSG00000020484	*Znrf3*	ENSMUSG00000041961
0.001	11	11610839	*Ikzf1*	ENSMUSG00000018654	*Fignl1*	ENSMUSG00000035455
0.001	11	18916027	*Meis1*	ENSMUSG00000020160		ENSMUSG00000069956
0.001	11	22705312		ENSMUSG00000062899	*B3gnt2*	ENSMUSG00000051650
0.001	11	48835439		ENSMUSG00000040335	*Olfr1396*	ENSMUSG00000047511
0.001	11	52072054	*Vdac1*	ENSMUSG00000020402	*Tcf7*	ENSMUSG00000000782
0.001	11	54874286		ENSMUSG00000044751	*Slc36a3*	ENSMUSG00000049491
0.001	11	58159185	*OTTMUSG00000005737*	ENSMUSG00000058287	*1810065E05Rik*	ENSMUSG00000013653
0.001	11	62272627	*Ubb*	ENSMUSG00000019505	*Prr6*	ENSMUSG00000018509
0.001	11	67925874	*Ntn1*	ENSMUSG00000020902	*Stx8*	ENSMUSG00000020903
0.001	11	68156194	*Pik3r5*	ENSMUSG00000020901	*Ntn1*	ENSMUSG00000020902
0.001	11	72682830	*Atp2a3*	ENSMUSG00000020788		ENSMUSG00000065298
0.001	11	74740528		ENSMUSG00000061832	*Smg6*	ENSMUSG00000038290
0.001	11	74900917	*Ovca2*	ENSMUSG00000038268	*mmu-mir-132*	ENSMUSG00000065537
0.001	11	75004843	*Rtn4rl1*	ENSMUSG00000045287	*Rpa1*	ENSMUSG00000000751
0.001	11	77243999	*1300007F04Rik*	ENSMUSG00000000686	*Taok1*	ENSMUSG00000017291
0.001	11	78730665	*Ksr1*	ENSMUSG00000018334	*Lgals9*	ENSMUSG00000001123
0.001	11	79250571		ENSMUSG00000046628	*Rab11fip4*	ENSMUSG00000017639
0.001	11	82817405		ENSMUSG00000035163	*Ap2b1*	ENSMUSG00000035152
0.001	11	86133847		ENSMUSG00000064526	*Med13*	ENSMUSG00000034297
0.001	11	86318401	*mmu-mir-21*	ENSMUSG00000065455		ENSMUSG00000069780
0.001	11	86624827	*Ypel2*	ENSMUSG00000018427	*Dhx40*	ENSMUSG00000018425
0.001	11	87475267	*mmu-mir-142*	ENSMUSG00000065420	*Supt4h1*	ENSMUSG00000020485
0.001	11	88896397	*A930013B10Rik*	ENSMUSG00000063109	*Nog*	ENSMUSG00000048616
0.001	11	98296507		ENSMUSG00000065891	*Med24*	ENSMUSG00000017210
0.001	11	98568907	*Thra*	ENSMUSG00000058756	*Nr1d1*	ENSMUSG00000020889
0.001	11	98766517		ENSMUSG00000064491	*Rara*	ENSMUSG00000037992
0.001	11	98977357	*Ccr7*	ENSMUSG00000037944	*Smarce1*	ENSMUSG00000037935
0.001	11	100225567	*1110036O03Rik*	ENSMUSG00000006931	*Jup*	ENSMUSG00000001552
0.001	11	100668367	*Stat5a*	ENSMUSG00000004043	*Stat5b*	ENSMUSG00000020919
0.001	11	100709155	*Stat3*	ENSMUSG00000004040	*Ptrf*	ENSMUSG00000004044
0.001	11	102995551	*Fmnl1*	ENSMUSG00000055805	*1700023F06Rik*	ENSMUSG00000020940
0.001	11	106494037	*Pecam1*	ENSMUSG00000020717	*Gm885*	ENSMUSG00000040528
0.001	11	114981951	*Slc9a3r1*	ENSMUSG00000020733	*Cd300lf*	ENSMUSG00000047798
0.001	11	115250587	*Kctd2*	ENSMUSG00000016940	*Armc7*	ENSMUSG00000057219
0.001	11	115723675	*2210020M01Rik*	ENSMUSG00000048442	*Recql5*	ENSMUSG00000020752
0.001	11	116181215		ENSMUSG00000052791		ENSMUSG00000020786
0.001	11	116238127	*Rnf157*	ENSMUSG00000052949	*1110014K08Rik*	ENSMUSG00000050628
0.001	11	116424757	*Rhbdf2*	ENSMUSG00000020806	*Cygb*	ENSMUSG00000020810
0.001	11	117148567	*Sept9*	ENSMUSG00000059248		ENSMUSG00000021618
0.001	11	117605527	*Tmc8*	ENSMUSG00000050106	*6030468B19Rik*	ENSMUSG00000025573
0.001	11	119663437	*Chmp6*	ENSMUSG00000025371		ENSMUSG00000056397
0.001	12	12295330	*Mycn*	ENSMUSG00000037169		ENSMUSG00000048957
0.001	12	52103161		ENSMUSG00000019260	*1110008L16Rik*	ENSMUSG00000021023
0.001	12	70425031	*Prkch*	ENSMUSG00000021108	*Tmem30b*	ENSMUSG00000034435
0.001	12	82414330	*Jundm2*	ENSMUSG00000034271	*Fos*	ENSMUSG00000021250
0.001	12	82502328	*Jundm2*	ENSMUSG00000034271	*Batf*	ENSMUSG00000034266
0.001	12	83703031		ENSMUSG00000061115		ENSMUSG00000059114
0.001	12	102594114		ENSMUSG00000057566		ENSMUSG00000064099
0.001	12	103397701	*Bcl11b*	ENSMUSG00000048251	*Setd3*	ENSMUSG00000056770
0.001	12	106174472	*Hsp90aa1*	ENSMUSG00000021270	*Wdr20a*	ENSMUSG00000037957
0.001	12	108155761	*EG382639*	ENSMUSG00000037638	*Akt1*	ENSMUSG00000001729
0.001	12	108336933	*Gpr132*	ENSMUSG00000021298	*Jag2*	ENSMUSG00000002799
0.001	13	18833698	*Stard3nl*	ENSMUSG00000003062		ENSMUSG00000003061
0.001	13	23035361	*Hist1h1d*	ENSMUSG00000052565	*Hist2h3c2*	ENSMUSG00000069273
0.001	13	28426378		ENSMUSG00000043626		ENSMUSG00000069256
0.001	13	30230008	*Irf4*	ENSMUSG00000021356	*Exoc2*	ENSMUSG00000021357
0.001	13	30308308	*Exoc2*	ENSMUSG00000021357		ENSMUSG00000069254

0.001	13	33651514	*Tubb2b*	ENSMUSG00000045136	*Tubb2a*	ENSMUSG00000058672
0.001	13	37400278	*Rreb1*	ENSMUSG00000039087		ENSMUSG00000052366
0.001	13	50295208	*Sema4d*	ENSMUSG00000021451	*Gadd45g*	ENSMUSG00000021453
0.001	13	50519968	*Gadd45g*	ENSMUSG00000021453	*Diras2*	ENSMUSG00000047842
0.001	13	50636638	*Gadd45g*	ENSMUSG00000021453	*Diras2*	ENSMUSG00000047842
0.001	13	54778354	*BC027057*	ENSMUSG00000049625	*H2afy*	ENSMUSG00000015937
0.001	13	61798581	*Ccrk*	ENSMUSG00000021483	*Ctsl*	ENSMUSG00000021477
0.001	13	98887438	*Cd180*	ENSMUSG00000021624		ENSMUSG00000069077
0.001	13	108815959	*Il6st*	ENSMUSG00000021756		ENSMUSG00000069060
0.001	14	23671933		ENSMUSG00000068669		ENSMUSG00000068670
0.001	14	23821465		ENSMUSG00000068668	*Zmiz1*	ENSMUSG00000007817
0.001	14	23892372	*Zmiz1*	ENSMUSG00000007817	*Ppif*	ENSMUSG00000021868
0.001	14	25493395	*D14Abb1e*	ENSMUSG00000040651	*Arhgef3*	ENSMUSG00000021895
0.001	14	39100453	*5730469M10Rik*	ENSMUSG00000021792	*Tspan14*	ENSMUSG00000037824
0.001	14	55259395	*Spata13*	ENSMUSG00000021990		ENSMUSG00000021992
0.001	14	64378994	*Egr3*	ENSMUSG00000033730		ENSMUSG00000068098
0.001	14	67472035	*Rcbtb2*	ENSMUSG00000022106	*Cysltr2*	ENSMUSG00000033470
0.001	14	69524875	*Lcp1*	ENSMUSG00000021998	*Cpb2*	ENSMUSG00000021999
0.001	14	73628125	*1190002H23Rik*	ENSMUSG00000022018		ENSMUSG00000045353
0.001	14	73878564		ENSMUSG00000067917	*Sugt1*	ENSMUSG00000022024
0.001	14	109596475	*mmu-mir-17*	ENSMUSG00000065508		ENSMUSG00000050988
0.001	14	113352627	*Cldn10a*	ENSMUSG00000022132	*Dzip1*	ENSMUSG00000042156
0.001	14	116528750	*Ebi2*	ENSMUSG00000051212	*Timm8a1*	ENSMUSG00000045455
0.001	14	116576305	*Timm8a1*	ENSMUSG00000045455	*Ebi2*	ENSMUSG00000051212
0.001	15	8345129		ENSMUSG00000068679	*Slc1a3*	ENSMUSG00000005360
0.001	15	9283820	*Capsl*	ENSMUSG00000039676	*Il7r*	ENSMUSG00000003882
0.001	15	61996917	*Myc*	ENSMUSG00000022346		ENSMUSG00000065812
0.001	15	62053677	*Myc*	ENSMUSG00000022346		ENSMUSG00000061375
0.001	15	62111705		ENSMUSG00000061375	*Myc*	ENSMUSG00000022346
0.001	15	62188686		ENSMUSG00000061375	*Myc*	ENSMUSG00000022346
0.001	15	62472724		ENSMUSG00000056590		ENSMUSG00000056864
0.001	15	62749453		ENSMUSG00000056864		ENSMUSG00000056590
0.001	15	63277845		ENSMUSG00000064826		ENSMUSG00000056864
0.001	15	63669185		ENSMUSG00000068466		ENSMUSG00000043083
0.001	15	73551396	*Dennd3*	ENSMUSG00000036661	*Slc45a4*	ENSMUSG00000036649
0.001	15	73763915		ENSMUSG00000064808		ENSMUSG00000058416
0.001	15	74957525		ENSMUSG00000034596	*Ly6e*	ENSMUSG00000022587
0.001	15	80620823	*Grap2*	ENSMUSG00000042351		ENSMUSG00000068166
0.001	15	82079568		ENSMUSG00000068119	*Xrcc6*	ENSMUSG00000022471
0.001	15	83649146	*Scube1*	ENSMUSG00000016763	*Ttll12*	ENSMUSG00000016757
0.001	15	103321602	*Copz1*	ENSMUSG00000060992	*mmu-mir-148b*	ENSMUSG00000065560
0.001	16	16743887	*Vpreb2*	ENSMUSG00000059280	*Slc25a1*	ENSMUSG00000003528
0.001	16	28771453	*Hes1*	ENSMUSG00000022528	*9530020O07Rik*	ENSMUSG00000056344
0.001	16	30992563	*Lrrc33*	ENSMUSG00000052384	*Fbxo45*	ENSMUSG00000035764
0.001	16	31329160	*Osta*	ENSMUSG00000035699		ENSMUSG00000068350
0.001	16	48643063		ENSMUSG00000060444	*Cd47*	ENSMUSG00000055447
0.001	16	54845060		ENSMUSG00000051061	*Rpl24*	ENSMUSG00000022601
0.001	16	90606834	*Ifnar1*	ENSMUSG00000022967	*Il10rb*	ENSMUSG00000022969
0.001	16	91914433	*Runx1*	ENSMUSG00000022952	*Setd4*	ENSMUSG00000022948
0.001	16	91968509	*Runx1*	ENSMUSG00000022952	*Setd4*	ENSMUSG00000022948
0.001	16	92019606	*Runx1*	ENSMUSG00000022952	*Setd4*	ENSMUSG00000022948
0.001	16	92092059	*Runx1*	ENSMUSG00000022952	*Setd4*	ENSMUSG00000022948
0.001	16	92318566	*Setd4*	ENSMUSG00000022948	*Runx1*	ENSMUSG00000022952
0.001	16	92987190	*Morc3*	ENSMUSG00000039456	*Dopey2*	ENSMUSG00000022946
0.001	16	94858873	*Erg*	ENSMUSG00000040732	*Ets2*	ENSMUSG00000022895
0.001	16	95302243	*Dscr2*	ENSMUSG00000022913	*Ets2*	ENSMUSG00000022895
0.001	17	13030875	*9030025P20Rik*	ENSMUSG00000036552	*Dll1*	ENSMUSG00000014773
0.001	17	25350593	*Hmga1*	ENSMUSG00000046711		ENSMUSG00000060209
0.001	17	26317381	*4930511I11Rik*	ENSMUSG00000024223	*Tmhs*	ENSMUSG00000062252
0.001	17	27208140	*Fgd2*	ENSMUSG00000024013	*Pim1*	ENSMUSG00000024014
0.001	17	27304540	*Pim1*	ENSMUSG00000024014	*Tbc1d22b*	ENSMUSG00000042203
0.001	17	31905906	*Psmb8*	ENSMUSG00000024338	*Tap2*	ENSMUSG00000024339
0.001	17	32908903	*Lta*	ENSMUSG00000024402	*Nfkbil1*	ENSMUSG00000042419
0.001	17	33029595		ENSMUSG00000064174		ENSMUSG00000057294
0.001	17	33737865		ENSMUSG00000057787	*H2-T24*	ENSMUSG00000053835
0.001	17	43003035	*Aars2*	ENSMUSG00000023938	*Spats1*	ENSMUSG00000023935
0.001	17	45047008	*Ccnd3*	ENSMUSG00000034165		ENSMUSG00000054271
0.001	17	45115665	*Bysl*	ENSMUSG00000023988		ENSMUSG00000054271
0.001	17	53932665	*Uhrf1*	ENSMUSG00000001228	*M6prbp1*	ENSMUSG00000024197
0.001	17	81795735	*Haao*	ENSMUSG00000000673	*Zfp36l2*	ENSMUSG00000045817
0.001	17	81983535	*Zfp36l2*	ENSMUSG00000045817	*Haao*	ENSMUSG00000000673
0.001	18	4338405	*Map3k8*	ENSMUSG00000024235	*Papd1*	ENSMUSG00000024234
0.001	18	5357605	*Zfp438*	ENSMUSG00000050945	*ENSMUSG00000063087*	ENSMUSG00000063087
0.001	18	15396409		ENSMUSG00000069445		ENSMUSG00000042886
0.001	18	35179674		ENSMUSG00000065904	*Ctnna1*	ENSMUSG00000037815
0.001	18	35975425	*Tmem173*	ENSMUSG00000024349		ENSMUSG00000060500
0.001	18	55208965	*Zfp608*	ENSMUSG00000052713		ENSMUSG00000064942
0.001	18	61099162	*Arsi*	ENSMUSG00000036412	*Tcof1*	ENSMUSG00000024613
0.001	18	62394033	*Adrb2*	ENSMUSG00000045730	*Sh3tc2*	ENSMUSG00000045629
0.001	18	68452825	*4933403F05Rik*	ENSMUSG00000038121	*D18Ertd653e*	ENSMUSG00000024544
0.001	18	70795365	*Mbd2*	ENSMUSG00000024513		ENSMUSG00000058475
0.001	18	75510805	*Smad7*	ENSMUSG00000025880		ENSMUSG00000069347
0.001	18	85054024	*Cyb5*	ENSMUSG00000024646	*1700034H14Rik*	ENSMUSG00000024645
0.001	19	4087019	*Adrbk1*	ENSMUSG00000024858	*Fbxl11*	ENSMUSG00000054611
0.001	19	5598539	*Frmd8*	ENSMUSG00000043488	*Scyl1*	ENSMUSG00000024941
0.001	19	6189172	*Rasgrp2*	ENSMUSG00000032946	*Nrxn2*	ENSMUSG00000033768
0.001	19	28738605	*C030046E11Rik*	ENSMUSG00000038658	*Pdcd1lg2*	ENSMUSG00000016498
0.001	19	36835733	*Hhex*	ENSMUSG00000024986	*Exoc6*	ENSMUSG00000053799
0.001	19	40571759	*Dntt*	ENSMUSG00000025014	*Blnk*	ENSMUSG00000061132
0.001	19	41384622	*Frat2*	ENSMUSG00000047604		ENSMUSG00000067199
0.001	20	6105829		ENSMUSG00000050227	*Pim2*	ENSMUSG00000031155
0.001	20	6193879	*Gata1*	ENSMUSG00000031162	*Hdac6*	ENSMUSG00000031161
0.001	20	42949129	*Elf4*	ENSMUSG00000031103	*Q8C3U4_MOUSE*	ENSMUSG00000053512
0.001	20	47271754	*mmu-mir-106a*	ENSMUSG00000065456		ENSMUSG00000046863
0.001	20	98219522	*Cnbp2*	ENSMUSG00000031330		ENSMUSG00000060041

0.001	20	135168160	*Irs4*	ENSMUSG00000054667		ENSMUSG00000061044
0.001	20	135305959		ENSMUSG00000061044	*Irs4*	ENSMUSG00000054667
0.05	1	51885139	*Obfc2a*	ENSMUSG00000026107	*Myo1b*	ENSMUSG00000018417
0.05	1	59019360	*ENSMUSG00000050533*	ENSMUSG00000026031	*Casp8*	ENSMUSG00000026029
0.05	1	59476356	*Als2*	ENSMUSG00000026024	*Als2cr7*	ENSMUSG00000026023
0.05	1	74696520	*mmu-mir-26b*	ENSMUSG00000065468	*Ctdsp1*	ENSMUSG00000026176
0.05	1	133102528	*Plekha6*	ENSMUSG00000041757	*Ppp1r15b*	ENSMUSG00000046062
0.05	1	138026353	*Atp6v1g3*	ENSMUSG00000026394	*Ptprc*	ENSMUSG00000026395
0.05	1	162567874	*Bat2d*	ENSMUSG00000040225	*Myoc*	ENSMUSG00000026697
0.05	1	181811700	*EG433384*	ENSMUSG00000056699	*Enah*	ENSMUSG00000022995
0.05	2	11128538	*Prkcq*	ENSMUSG00000026778		ENSMUSG00000062973
0.05	2	13999812	*Stam*	ENSMUSG00000026718	*Ptpla*	ENSMUSG00000063275
0.05	2	18650614		ENSMUSG00000046809	*Dnajc1*	ENSMUSG00000026740
0.05	2	27054140	*C630035N08Rik*	ENSMUSG00000009216	*Dbh*	ENSMUSG00000000889
0.05	2	35430442		ENSMUSG00000068954	*Dab2ip*	ENSMUSG00000026883
0.05	2	44857908	*Zeb2*	ENSMUSG00000026872	*Gtdc1*	ENSMUSG00000036890
0.05	2	91340558	*F2*	ENSMUSG00000027249	*Arhgap1*	ENSMUSG00000027247
0.05	2	126434377		ENSMUSG00000009337	*Ap4e1*	ENSMUSG00000001998
0.05	2	163113278	*Pkig*	ENSMUSG00000035268	*0610039K10Rik*	ENSMUSG00000058812
0.05	2	163541060	*Stk4*	ENSMUSG00000018209	*Kcns1*	ENSMUSG00000040164
0.05	2	166100558		ENSMUSG00000065871	*BC067047*	ENSMUSG00000039621
0.05	2	166203470	*Trp53rk*	ENSMUSG00000042854		ENSMUSG00000065871
0.05	3	9883817		ENSMUSG00000069113		ENSMUSG00000065364
0.05	3	85851115		ENSMUSG00000065847	*Rps3a*	ENSMUSG00000028081
0.05	3	95419375	*Anp32e*	ENSMUSG00000015749		ENSMUSG00000057781
0.05	3	96046585	*Txnip*	ENSMUSG00000038393	*Polr3gl*	ENSMUSG00000028104
0.05	3	106583483	*Cd53*	ENSMUSG00000040747	*Olfr266*	ENSMUSG00000043529
0.05	3	151932715	*St6galnac5*	ENSMUSG00000039037	*St6galnac3*	ENSMUSG00000052544
0.05	4	8846480	*Chd7*	ENSMUSG00000041235		ENSMUSG00000064716
0.05	4	101529157	*Sgip1*	ENSMUSG00000028524	*Pde4b*	ENSMUSG00000028525
0.05	4	105875450	*Ssbp3*	ENSMUSG00000061887	*Mrpl37*	ENSMUSG00000028622
0.05	4	124151905	*Zc3h12a*	ENSMUSG00000042677	*Grik3*	ENSMUSG00000001985
0.05	4	128836280	*Ptp4a2*	ENSMUSG00000028788		ENSMUSG00000049089
0.05	4	129704690	*Laptm5*	ENSMUSG00000028581	*Sdc3*	ENSMUSG00000025743
0.05	4	131763693	*BC013712*	ENSMUSG00000037731	*Ppp1r8*	ENSMUSG00000028882
0.05	4	132590465		ENSMUSG00000059194	*Zdhhc18*	ENSMUSG00000037553
0.05	4	132717213		ENSMUSG00000064705	*Arid1a*	ENSMUSG00000007880
0.05	4	133877778	*D4Wsu53e*	ENSMUSG00000037266	*Tmem50a*	ENSMUSG00000028822
0.05	4	134087857	*Clic4*	ENSMUSG00000037242	*Syf2*	ENSMUSG00000028821
0.05	4	143924846		ENSMUSG00000066025		ENSMUSG00000066024
0.05	4	149406416	*Tnfrsf9*	ENSMUSG00000028965	*Park7*	ENSMUSG00000028964
0.05	4	150672620	*Acot7*	ENSMUSG00000028937	*Hes2*	ENSMUSG00000028940
0.05	4	153034815		ENSMUSG00000065993	*Actrt2*	ENSMUSG00000051276
0.05	4	153728683		ENSMUSG00000058333	*Ski*	ENSMUSG00000029050
0.05	4	154563626	*mmu-mir-200b*	ENSMUSG00000065549	*9430015G10Rik*	ENSMUSG00000059939
0.05	5	63607486	*Klf3*	ENSMUSG00000029178	*Tlr1*	ENSMUSG00000044827
0.05	5	71995966	*Ociad1*	ENSMUSG00000029152		ENSMUSG00000051990
0.05	5	99658186	*Coq2*	ENSMUSG00000029319		ENSMUSG00000067084
0.05	5	107022286	*2900024C23Rik*	ENSMUSG00000029270		ENSMUSG00000051165
0.05	5	111352516	*Tpst2*	ENSMUSG00000029344	*Tfip11*	ENSMUSG00000029345
0.05	5	121289205	*Ppp1cc*	ENSMUSG00000004455	*Ccdc63*	ENSMUSG00000043036
0.05	5	121383706	*TECT1_MOUSE*	ENSMUSG00000038593	*Hvcn1*	ENSMUSG00000064267
0.05	5	123243706	*Ogfod2*	ENSMUSG00000023707	*Hip1r*	ENSMUSG00000000915
0.05	5	123653743	*6330548G22Rik*	ENSMUSG00000029402	*BC003324*	ENSMUSG00000029401
0.05	5	133950556	*Wbscr27*	ENSMUSG00000040557	*Cldn4*	ENSMUSG00000047501
0.05	5	136689166	*Tsc22d4*	ENSMUSG00000029723	*6430598A04Rik*	ENSMUSG00000045348
0.05	5	139207846	*Ftsj2*	ENSMUSG00000029557	*Mad1l1*	ENSMUSG00000029554
0.05	5	141379487	*Foxk1*	ENSMUSG00000056493	*C330006K01Rik*	ENSMUSG00000039623
0.05	5	147914211	*Hmgb1*	ENSMUSG00000066551		ENSMUSG00000066552
0.05	6	31167314		ENSMUSG00000052894	*mmu-mir-29b-1*	ENSMUSG00000065604
0.05	6	31257583		ENSMUSG00000052507	*Mkln1*	ENSMUSG00000025609
0.05	6	72717805	*Vamp8*	ENSMUSG00000050732	*Ggcx*	ENSMUSG00000053460
0.05	6	91554600	*Hdac11*	ENSMUSG00000034245	*Nup210*	ENSMUSG00000030091
0.05	6	99735040		ENSMUSG00000024343	*Eif4e3*	ENSMUSG00000030068
0.05	6	121662055	*Usp18*	ENSMUSG00000030107	*Tuba8*	ENSMUSG00000030137
0.05	6	125559038	*Lag3*	ENSMUSG00000030124	*Cd4*	ENSMUSG00000023274
0.05	6	135683425	*Cdkn1b*	ENSMUSG00000003031		ENSMUSG00000043984
0.05	6	145942855	*Lrmp*	ENSMUSG00000030263	*Bcat1*	ENSMUSG00000030268
0.05	7	20823368	*Exosc5*	ENSMUSG00000061286	*Bckdha*	ENSMUSG00000060376
0.05	7	25834009	*Hcst*	ENSMUSG00000064109	*AY078069*	ENSMUSG00000036931
0.05	7	26241588	*Ffar2*	ENSMUSG00000051314	*Ffar3*	ENSMUSG00000019429
0.05	7	39652369	*Bcat2*	ENSMUSG00000030826		ENSMUSG00000062241
0.05	7	40007490	*Emp3*	ENSMUSG00000040212	*Ccdc114*	ENSMUSG00000040189
0.05	7	57779910	*Klf13*	ENSMUSG00000052040	*Trpm1*	ENSMUSG00000030523
0.05	7	72517740	*Mrpl46*	ENSMUSG00000030612	*ENSMUSG00000052629*	ENSMUSG00000052629
0.05	7	74247357	*Blm*	ENSMUSG00000030528		ENSMUSG00000030529
0.05	7	77513550		ENSMUSG00000066374		ENSMUSG00000053138
0.05	7	93910346	*Neu3*	ENSMUSG00000035239	*Spcs2*	ENSMUSG00000035227
0.05	7	121192906	*Spn*	ENSMUSG00000051457	*Cd2bp2*	ENSMUSG00000042502
0.05	7	123882300	*Brwd2*	ENSMUSG00000042055		ENSMUSG00000065763
0.05	7	136753950	*6330512M04Rik*	ENSMUSG00000045777		ENSMUSG00000066099
0.05	7	137972370	*Cars*	ENSMUSG00000010755	*Tnfrsf26*	ENSMUSG00000045362
0.05	8	21562896	*1700041G16Rik*	ENSMUSG00000054822	*Myst3*	ENSMUSG00000031540
0.05	8	23824673	*Tacc1*	ENSMUSG00000065954	*Plekha2*	ENSMUSG00000031557
0.05	8	32948963	*Tmem66*	ENSMUSG00000031532	*Leprotl1*	ENSMUSG00000031513
0.05	8	94258013	*Gpr56*	ENSMUSG00000031785	*Gpr114*	ENSMUSG00000061577
0.05	8	105403341	*Rbm35b*	ENSMUSG00000033824	*Nfatc3*	ENSMUSG00000031902
0.05	8	106461053	*Cyb5b*	ENSMUSG00000031924		ENSMUSG00000057657
0.05	8	114543956		ENSMUSG00000055043		ENSMUSG00000065090
0.05	8	119848093	*Gse1*	ENSMUSG00000031822	*1700120B06Rik*	ENSMUSG00000042269
0.05	8	120081812	*Irf8*	ENSMUSG00000041515	*Cox4i1*	ENSMUSG00000031818
0.05	8	120106043	*Irf8*	ENSMUSG00000041515	*Foxf1a*	ENSMUSG00000042812
0.05	8	125951548		ENSMUSG00000069852	*Irf2bp2*	ENSMUSG00000051495
0.05	9	20507368	*Fbxl12*	ENSMUSG00000066892	*5730577I03Rik*	ENSMUSG00000062470
0.05	9	21207041	*Slc44a2*	ENSMUSG00000057193	*Ap1m2*	ENSMUSG00000003309

0.05	9	44636891	*Treh*	ENSMUSG00000032098	*Ddx6*	ENSMUSG00000032097
0.05	9	45297401	*Tmprss13*	ENSMUSG00000037129	*Il10ra*	ENSMUSG00000032089
0.05	9	64914275	*F730015K02Rik*	ENSMUSG00000053641		ENSMUSG00000066519
0.05	9	65666861	*Plekhq1*	ENSMUSG00000050721		ENSMUSG00000041270
0.05	9	96852131	*Spsb4*	ENSMUSG00000046997	*Slc25a36*	ENSMUSG00000032449
0.05	10	28620462		ENSMUSG00000069691	*E430004N04Rik*	ENSMUSG00000049109
0.05	10	42998622	*Scml4*	ENSMUSG00000044770		ENSMUSG00000056871
0.05	10	43934922		ENSMUSG00000064118	*Qrsl1*	ENSMUSG00000019863
0.05	10	60332832	*4632428N05Rik*	ENSMUSG00000020101	*Slc29a3*	ENSMUSG00000020100
0.05	10	79916922	*Palm*	ENSMUSG00000035863	*BC005764*	ENSMUSG00000035835
0.05	10	79975739	*Ptbp1*	ENSMUSG00000006498	*9130017N09Rik*	ENSMUSG00000035852
0.05	10	80531388	*Uqcr*	ENSMUSG00000020163	*Mbd3*	ENSMUSG00000035478
0.05	10	81060296	*Gadd45b*	ENSMUSG00000015312		ENSMUSG00000035103
0.05	10	89422752	*Nr1h4*	ENSMUSG00000047638		ENSMUSG00000019928
0.05	10	91154799	*Tmpo*	ENSMUSG00000019961		ENSMUSG00000019985
0.05	10	93279846	*Elk3*	ENSMUSG00000008398	*Lta4h*	ENSMUSG00000015889
0.05	10	117780612		ENSMUSG00000069509	*Rap1b*	ENSMUSG00000052681
0.05	10	118556982	*4932442E05Rik*	ENSMUSG00000050709	*Cand1*	ENSMUSG00000020114
0.05	10	128150682	*Rnf41*	ENSMUSG00000025373	*Smarcc2*	ENSMUSG00000025369
0.05	11	3200617	*OTTMUSG00000007639*	ENSMUSG00000053263	*Pik3ip1*	ENSMUSG00000034614
0.05	11	6424357		ENSMUSG00000064513		ENSMUSG00000065783
0.05	11	6450157	*Ccm2*	ENSMUSG00000000378	*Nacad*	ENSMUSG00000041073
0.05	11	18767407	*Meis1*	ENSMUSG00000020160		ENSMUSG00000069957
0.05	11	33905838	*Lcp2*	ENSMUSG00000002699	*Kcnmb1*	ENSMUSG00000020155
0.05	11	44369918	*Ebf1*	ENSMUSG00000057098	*Rnf145*	ENSMUSG00000019189
0.05	11	49001235	*Mgat1*	ENSMUSG00000020346	*Olfr1393*	ENSMUSG00000059864
0.05	11	51394259	*C330016O10Rik*	ENSMUSG00000001053	*Rmnd5b*	ENSMUSG00000001054
0.05	11	57827011	*Cnot8*	ENSMUSG00000020515	*Q3TQ57_MOUSE*	ENSMUSG00000069875
0.05	11	59355197	*Olfr225*	ENSMUSG00000044061	*AA536749*	ENSMUSG00000005417
0.05	11	59833387	*Rai1*	ENSMUSG00000062115	*4930412M03Rik*	ENSMUSG00000051008
0.05	11	67995067	*Ntn1*	ENSMUSG00000020902	*Pik3r5*	ENSMUSG00000020901
0.05	11	68747377	*1500010J02Rik*	ENSMUSG00000020898	*Aurkb*	ENSMUSG00000020897
0.05	11	69307020	*Trp53*	ENSMUSG00000059552	*Atp1b2*	ENSMUSG00000041329
0.05	11	72206047	*Spns3*	ENSMUSG00000020798	*Spns2*	ENSMUSG00000040447
0.05	11	75185176	*mmu-mir-22*	ENSMUSG00000065529	*Wdr81*	ENSMUSG00000045374
0.05	11	75372697	*Myo1c*	ENSMUSG00000017774	*Pps*	ENSMUSG00000006127
0.05	11	100549426	*NP_001074663.1*	ENSMUSG00000035355	*Rab5c*	ENSMUSG00000019173
0.05	11	106575247	*Gm885*	ENSMUSG00000040528	*Polg2*	ENSMUSG00000020718
0.05	11	107239162	*Psmd12*	ENSMUSG00000020720	*Pitpnc1*	ENSMUSG00000040430
0.05	11	115911097	*Wbp2*	ENSMUSG00000034341	*Trim47*	ENSMUSG00000020773
0.05	11	117776339	*Socs3*	ENSMUSG00000053113	*Tha1*	ENSMUSG00000017713
0.05	11	118287487	*D230014K01Rik*	ENSMUSG00000033857		ENSMUSG00000055046
0.05	11	119500262	*4932417H02Rik*	ENSMUSG00000025583		ENSMUSG00000056397
0.05	11	120450487	*Mafg*	ENSMUSG00000051510	*Pycr1*	ENSMUSG00000025140
0.05	11	121373610		ENSMUSG00000069562	*B3gntl1*	ENSMUSG00000046605
0.05	11	121505573	*Metrnl*	ENSMUSG00000039208	*B3gntl1*	ENSMUSG00000046605
0.05	12	48611011	*Heatr5a*	ENSMUSG00000035181	*EG544864*	ENSMUSG00000059605
0.05	12	66379019	*Gm71*	ENSMUSG00000049882	*Arf6*	ENSMUSG00000044147
0.05	12	73221091	*Zbtb1*	ENSMUSG00000033454	*Zbtb25*	ENSMUSG00000056459
0.05	12	74041861	*Fut8*	ENSMUSG00000021065	*Max*	ENSMUSG00000059436
0.05	12	81551551	*Npc2*	ENSMUSG00000021242	*Abcd4*	ENSMUSG00000021240
0.05	12	83765431	*6430527G18Rik*	ENSMUSG00000034168	*2310044G17Rik*	ENSMUSG00000034157
0.05	12	100447651		ENSMUSG00000064849	*4831426I19Rik*	ENSMUSG00000054150
0.05	12	106910208	*Tnfaip2*	ENSMUSG00000021281	*Gm266*	ENSMUSG00000010529
0.05	12	108902461		ENSMUSG00000061907	*4930523C11Rik*	ENSMUSG00000051804
0.05	12	110690701		ENSMUSG00000066297		ENSMUSG00000066299
0.05	13	12949185		ENSMUSG00000069331	*Lyst*	ENSMUSG00000019726
0.05	13	28237845		ENSMUSG00000062217		ENSMUSG00000069257
0.05	13	51121798	*Syk*	ENSMUSG00000021457	*Diras2*	ENSMUSG00000047842
0.05	13	60341458	*2010111I01Rik*	ENSMUSG00000021458		ENSMUSG00000064997
0.05	13	95179819		ENSMUSG00000069108		ENSMUSG00000060323
0.05	13	97865338	*Pik3r1*	ENSMUSG00000041417		ENSMUSG00000069079
0.05	14	14668912		ENSMUSG00000068752	*Oxsm*	ENSMUSG00000021786
0.05	14	25535035	*D14Abb1e*	ENSMUSG00000040651	*Arhgef3*	ENSMUSG00000021895
0.05	14	28775305		ENSMUSG00000042565	*Prkcd*	ENSMUSG00000021948
0.05	14	29548276	*Sh3bp5*	ENSMUSG00000021892	*Mettl6*	ENSMUSG00000021891
0.05	14	29691175	*Colq*	ENSMUSG00000057606	*Hacl1*	ENSMUSG00000021884
0.05	14	29803018	*Btd*	ENSMUSG00000021900	*Ankrd28*	ENSMUSG00000014496
0.05	14	67752804	*Itm2b*	ENSMUSG00000022108	*Med4*	ENSMUSG00000022109
0.05	15	36452585	*Ankrd46*	ENSMUSG00000048307		ENSMUSG00000068579
0.05	15	62364378		ENSMUSG00000056590		ENSMUSG00000056864
0.05	15	66834725		ENSMUSG00000068418	*Wisp1*	ENSMUSG00000005124
0.05	15	78644705	*Pscd4*	ENSMUSG00000018008	*Rac2*	ENSMUSG00000033220
0.05	15	79263425	*Pick1*	ENSMUSG00000068206	*Sox10*	ENSMUSG00000033006
0.05	15	81568145		ENSMUSG00000068142	*Ep300*	ENSMUSG00000055024
0.05	15	84366995	*Parvg*	ENSMUSG00000022439	*Parvb*	ENSMUSG00000022438
0.05	15	97449954	*5830453K13Rik*	ENSMUSG00000044325	*BC038822*	ENSMUSG00000044250
0.05	15	97542935	*5830453K13Rik*	ENSMUSG00000044325	*Rpap3*	ENSMUSG00000022466
0.05	15	97916315	*Vdr*	ENSMUSG00000022479	*Hdac7a*	ENSMUSG00000022475
0.05	16	4228813	*Tcfap4*	ENSMUSG00000005718	*Glis2*	ENSMUSG00000014303
0.05	16	8415673		ENSMUSG00000065902	*Usp7*	ENSMUSG00000022710
0.05	16	29200453	*Atp13a3*	ENSMUSG00000022533	*Tmem44*	ENSMUSG00000022537
0.05	16	54969715	*Senp7*	ENSMUSG00000052917		ENSMUSG00000064994
0.05	16	75626803		ENSMUSG00000050299		ENSMUSG00000053608
0.05	16	93648349	*Ripply3*	ENSMUSG00000022941	*Hlcs*	ENSMUSG00000040820
0.05	16	93919303	*Dscr3*	ENSMUSG00000022898	*Dyrk1a*	ENSMUSG00000022897
0.05	17	13683105	*Chd1*	ENSMUSG00000023852	*Prdm9*	ENSMUSG00000051977
0.05	17	23599995	*Metrn*	ENSMUSG00000002274	*BC008155*	ENSMUSG00000057411
0.05	17	23795145	*Rab11fip3*	ENSMUSG00000037098	*Decr2*	ENSMUSG00000036775
0.05	17	28883660	*Abcg1*	ENSMUSG00000024030	*Tff3*	ENSMUSG00000024029
0.05	17	32751890		ENSMUSG00000065139	*Msh5*	ENSMUSG00000007035
0.05	17	33636415	*Ppp1r10*	ENSMUSG00000039220		ENSMUSG00000049832
0.05	17	42287145	*Supt3h*	ENSMUSG00000038954	*4930564C03Rik*	ENSMUSG00000048820
0.05	17	44277195	*2310039H08Rik*	ENSMUSG00000062619	*NP_035325.1*	ENSMUSG00000036858
0.05	17	45563715	*1700122O11Rik*	ENSMUSG00000042494	*1700067P10Rik*	ENSMUSG00000021545

0.05	17	49419135		ENSMUSG00000058380		ENSMUSG00000053770
0.05	17	51109845	*Pcaf*	ENSMUSG00000000708		ENSMUSG00000063839
0.05	17	63594898	*Twsg1*	ENSMUSG00000024098	*Ralbp1*	ENSMUSG00000024096
0.05	17	69257535	*Ndc80*	ENSMUSG00000024056	*Q8BP09_MOUSE*	ENSMUSG00000041913
0.05	18	39246982	*Arhgap26*	ENSMUSG00000036452	*Nr3c1*	ENSMUSG00000024431
0.05	18	50244433	*Tnfaip8*	ENSMUSG00000062210		ENSMUSG00000069381
0.05	18	60980215	*Rps14*	ENSMUSG00000024608	*Ndst1*	ENSMUSG00000054008
0.05	18	65140045	*Nedd4l*	ENSMUSG00000024589	*A330084C13Rik*	ENSMUSG00000055362
0.05	18	80828071	*Atp9b*	ENSMUSG00000024566	*Nfatc1*	ENSMUSG00000033016
0.05	19	3938519	*Coro1b*	ENSMUSG00000024835	*Rps6kb2*	ENSMUSG00000024830
0.05	19	4011389	*Clcf1*	ENSMUSG00000040663	*Pold4*	ENSMUSG00000024854
0.05	19	5629379	*Frmd8*	ENSMUSG00000043488	*Scyl1*	ENSMUSG00000024941
0.05	19	7953175	*Nxf1*	ENSMUSG00000010097	*Stx5a*	ENSMUSG00000010110
0.05	19	8242954	*Scgb1a1*	ENSMUSG00000024653		ENSMUSG00000059508
0.05	19	9780269	*Cybasc3*	ENSMUSG00000034445	*Dak*	ENSMUSG00000034371
0.05	19	24225959	*Dock8*	ENSMUSG00000052085		ENSMUSG00000064896
0.05	19	31586856	*2700046G09Rik*	ENSMUSG00000024893	*Sgms1*	ENSMUSG00000040451
0.05	19	32424629		ENSMUSG00000052009	*Pten*	ENSMUSG00000013663
0.05	19	35598899	*Pcgf5*	ENSMUSG00000024805	*Ankrd1*	ENSMUSG00000024803
0.05	19	43928459	*Scd1*	ENSMUSG00000037071	*Scd4*	ENSMUSG00000050195
0.05	19	45077552	*Poll*	ENSMUSG00000025218		ENSMUSG00000062336
0.05	19	46522796	*Ina*	ENSMUSG00000034336	*Q99JM3_MOUSE*	ENSMUSG00000025039
0.05	19	46999894	*Obfc1*	ENSMUSG00000042694	*Sh3pxd2a*	ENSMUSG00000053617
0.05	19	53470079	*Pdcd4*	ENSMUSG00000024975	*Rbm20*	ENSMUSG00000043639
0.05	20	11481710		ENSMUSG00000058791	*Ddx3x*	ENSMUSG00000000787
0.05	20	42661369	*1200013B08Rik*	ENSMUSG00000031101	*Xpnpep2*	ENSMUSG00000037005
0.05	20	128113669	*Gla*	ENSMUSG00000031266	*Btk*	ENSMUSG00000031264

Appendix E. Human cancer cell lines used in the 10K and SNP 6.0 CGH analyses.

SNP6.0? indicates whether the cell line was used in the higher resolution SNP 6.0 CGH analysis. Abbreviations: auto_ganglia = autonomic ganglia; CNS = central nervous system; haem_and_lymph = haematopoietic and lymphoid tissue.
unknown(…) is used where the cell line was classified as unknown in the 10K analysis, but was later identified for the SNP 6.0 analysis.

Cell line	Tissue	SNP6.0?	Cell line	Tissue	SNP6.0?	Cell line	Tissue	SNP6.0?
SW13	adrenal_gland	Y	MOLT-4	haem_and_lymph	Y	SCLC-21H	lung	Y
GI-CA-N	auto_ganglia		MONO-MAC-6	haem_and_lymph	Y	SHP-77	lung	Y
KP-N-YS	auto_ganglia	Y	Mo-T	haem_and_lymph	Y	SK-LU-1	lung	Y
NB-10	auto_ganglia	Y	MUTZ-1	haem_and_lymph	Y	SK-MES-1	lung	Y
NB-13	auto_ganglia	Y	MV-4-11	haem_and_lymph		SW1573	lung	Y
NB-14	auto_ganglia	Y	NALM-1	haem_and_lymph	Y	SW900	lung	Y
NB-17	auto_ganglia	Y	NALM-6	haem_and_lymph	Y	VMRC-LCP	lung	Y
NB-5	auto_ganglia	Y	NC-37	haem_and_lymph		COLO-680N	oesophagus	Y
NB-6	auto_ganglia	Y	NKM-1	haem_and_lymph	Y	EC-GI-10	oesophagus	Y
NB-7	auto_ganglia	Y	NOMO-1	haem_and_lymph	Y	HCE-4	oesophagus	Y
ACN	auto_ganglia	Y	OCI-AML2	haem_and_lymph	Y	KYSE-140	oesophagus	
BE2-C	auto_ganglia		OPM-2	haem_and_lymph	Y	KYSE-150	oesophagus	
CHP-126	auto_ganglia	Y	P12-ICHIKAWA	haem_and_lymph	Y	KYSE-180	oesophagus	
CHP-212	auto_ganglia	Y	P30-OHK	haem_and_lymph	Y	KYSE-30	oesophagus	
GI-LI-N	auto_ganglia		P31-FUJ	haem_and_lymph	Y	KYSE-520	oesophagus	Y
GOTO-P3	auto_ganglia	Y	PF-382	haem_and_lymph		KYSE-70	oesophagus	
KP-N-RT-BM-1	auto_ganglia	Y	QIMR-WIL	haem_and_lymph	Y	OE19	oesophagus	Y
KP-N-S19s	auto_ganglia	Y	Raji	haem_and_lymph	Y	OE33	oesophagus	Y
LAN-5	auto_ganglia	Y	Ramos-2G6-4C10	haem_and_lymph	Y	TE-1	oesophagus	Y
LAN-6	auto_ganglia	Y	REH	haem_and_lymph	Y	TE-10	oesophagus	Y
MC-IXC	auto_ganglia	Y	RL	haem_and_lymph	Y	TE-11	oesophagus	Y
MHH-NB-11	auto_ganglia	Y	RPMI-6666	haem_and_lymph	Y	TE-12	oesophagus	Y
NB16-RIKEN	auto_ganglia	Y	RPMI-8226	haem_and_lymph	Y	TE-15	oesophagus	Y
NH-12	auto_ganglia	Y	RPMI-8402	haem_and_lymph	Y	TE-5	oesophagus	Y
NH-6	auto_ganglia	Y	RPMI-8866	haem_and_lymph	Y	TE-6	oesophagus	Y
SCCH-26	auto_ganglia	Y	RS4-11	haem_and_lymph	Y	TE-8	oesophagus	Y
SIMA	auto_ganglia	Y	SCC-3	haem_and_lymph	Y	TE-9	oesophagus	Y
SK-N-AS	auto_ganglia	Y	SIG-M5	haem_and_lymph	Y	41MCISR	ovary	
SK-N-DZ	auto_ganglia	Y	SKM-1	haem_and_lymph	Y	Caov-3	ovary	Y
TGW	auto_ganglia	Y	SK-MM-2	haem_and_lymph	Y	EFO-21	ovary	Y
EGI-1	biliary_tract	Y	SKW	haem_and_lymph		EFO-27	ovary	Y
ETK-1	biliary_tract	Y	SKW-3	haem_and_lymph		IGROV-1	ovary	Y
HuCCT1	biliary_tract	Y	ST486	haem_and_lymph	Y	KGN	ovary	Y
HuH-28	biliary_tract	Y	SU-DHL-1	haem_and_lymph	Y	KURAMOCHI	ovary	Y
TGBC1TKB	biliary_tract	Y	SUP-B8	haem_and_lymph	Y	OAW28	ovary	Y
TGBC24TKB	biliary_tract	Y	SUP-T1	haem_and_lymph	Y	OAW-42	ovary	Y
CADO-ES1	bone	Y	TALL-1	haem_and_lymph	Y	OC-314	ovary	Y
CAL-72	bone	Y	THP-1	haem_and_lymph		OVCAR3	ovary	Y
ES-1	bone	Y	TUR	haem_and_lymph	Y	OVCAR-4	ovary	Y
ES-4	bone	Y	U-266	haem_and_lymph	Y	OVCAR-5	ovary	Y
EW-1	bone	Y	U-698-M	haem_and_lymph	Y	PA-1	ovary	Y
EW-12	bone	Y	WSU-NHL	haem_and_lymph	Y	RMG-I	ovary	Y
EW-13	bone	Y	YT	haem_and_lymph	Y	RTSG	ovary	Y
EW-16	bone	Y	NB4	haem_and_lymph		SK-OV-3	ovary	Y
EW-18	bone		769-P	kidney	Y	TYK-nu	ovary	Y
EW-22	bone	Y	786-0	kidney	Y	AsPC-1	pancreas	Y
EW-24	bone	Y	A498	kidney	Y	BxPC-3	pancreas	Y
EW-3	bone	Y	ACHN	kidney	Y	Capan-1	pancreas	
EW-7	bone	Y	BB65-RCC	kidney	Y	Capan-2	pancreas	Y
H-EMC-SS	bone	Y	BFTC-909	kidney	Y	CFPAC-1	pancreas	Y
HOS	bone	Y	CAKI-1	kidney	Y	HPAF-II	pancreas	Y
HuO9	bone	Y	CAL-54	kidney	Y	HuP-T3	pancreas	Y
MG-63	bone	Y	HA7-RCC	kidney	Y	HuP-T4	pancreas	Y
NOS-1	bone	Y	LB1047-RCC	kidney	Y	MIA-PaCa-2	pancreas	
NY	bone	Y	LB2241-RCC	kidney	Y	MZ1-PC	pancreas	
Saos-2	bone	Y	LB996-RCC	kidney	Y	PANC-03-27	pancreas	Y
SJSA-1	bone	Y	OS-RC-2	kidney		PANC-10-05	pancreas	Y
SK-PN-DW	bone	Y	RCC10RGB	kidney		SW1990	pancreas	Y
U-2-OS	bone	Y	RXF393	kidney	Y	YAPC	pancreas	Y
BT-20	breast		SK-NEP-1	kidney	Y	JAR	placenta	Y
BT-474	breast	Y	SN12C	kidney	Y	JEG-3	placenta	Y
BT-549	breast	Y	SW156	kidney		IST-MES1	pleura	Y
CAL-120	breast	Y	TK10	kidney	Y	MPP-89	pleura	Y
CAL-148	breast	Y	U031	kidney	Y	MSTO-211H	pleura	Y
CAL-51	breast	Y	VMRC-RCZ	kidney	Y	NCI-H2452	pleura	Y
CAL-85-1	breast	Y	VMRC-MELG	large intestine	Y	NCI-H28	pleura	Y
CAMA-1	breast	Y	293	large_intestine		BPH-1	prostate	Y
COLO-824	breast	Y	C2BBe1	large_intestine	Y	DU-145	prostate	Y
DU-4475	breast	Y	CA46	large_intestine	Y	A101D	skin	Y
EVSA-T	breast	Y	COLO-205	large_intestine	Y	A375	skin	Y
HCC1143	breast	Y	COLO-320	large_intestine	Y	A4-Fuk	skin	Y

HCC1187	breast	Y
HCC1395	breast	Y
HCC1419	breast	
HCC1569	breast	Y
HCC1599	breast	Y
HCC1806	breast	Y
HCC1937	breast	Y
HCC1954	breast	Y
HCC2157	breast	Y
HCC2218	breast	Y
HCC38	breast	Y
HCC70	breast	Y
Hs-578-T	breast	Y
MCF7	breast	Y
MDA-MB-134-VI	breast	Y
MDA-MB-157	breast	Y
MDA-MB-175-VII	breast	Y
MDA-MB-231	breast	Y
MDA-MB-361	breast	Y
MDA-MB-415	breast	Y
MDA-MB-453	breast	Y
MDA-MB-468	breast	
MFM-223	breast	
MRK-nu-1	breast	
NCI-ADR-RES	breast	
OCUB-M	breast	Y
SK-BR-3	breast	
SW527	breast	
T47D	breast	Y
UACC-812	breast	Y
UACC-893	breast	Y
8-MG-BA	CNS	Y
A172	CNS	Y
AM-38	CNS	Y
Becker	CNS	Y
CAS-1	CNS	Y
CCF-STTG1	CNS	Y
D-247-MG	CNS	Y
D-283-MED	CNS	Y
D-542-MG	CNS	Y
DBTRG-05MG	CNS	Y
DK-MG	CNS	Y
GAMG	CNS	Y
GB-1	CNS	Y
GI-1	CNS	Y
GMS-10	CNS	Y
GOS-3	CNS	
KALS-1	CNS	Y
KNS-42	CNS	Y
KNS-81	CNS	
KS-1	CNS	Y
LN-405	CNS	Y
MOG-G-CCM	CNS	Y
MOG-G-UVW	CNS	Y
NMC-G1	CNS	Y
no-10	CNS	Y
no-11	CNS	Y
ONS-76	CNS	Y
PFSK	CNS	Y
SF126	CNS	Y
SF268	CNS	Y
SF539	CNS	Y
SK-MG-1	CNS	Y
SNB75	CNS	Y
SW1783	CNS	Y
T98G	CNS	Y
U-118-MG	CNS	Y
U251	CNS	Y
U-87-MG	CNS	Y
YH-13	CNS	Y
YKG-1	CNS	
BOKU	cervix	Y
C-33-A	cervix	Y
C-4-II	cervix	Y
Ca-Ski	cervix	Y
DoTc2-4510	cervix	Y
HT-3	cervix	Y
ME-180	cervix	Y
OMC-1	cervix	Y
SiHa	cervix	Y
SKG-IIIa	cervix	Y
SW756	cervix	Y
COLO-678	large_intestine	Y
COLO-741	large_intestine	Y
CW-2	large_intestine	Y
GP5d	large_intestine	Y
H4	large_intestine	Y
HCC2998	large_intestine	Y
HCT-116	large_intestine	Y
HT-115	large_intestine	
HT-29	large_intestine	Y
HT55	large_intestine	Y
LoVo	large_intestine	Y
LS-123	large_intestine	Y
LS-174T	large_intestine	Y
LS-411N	large_intestine	Y
LS-513	large_intestine	Y
NCI-H508	large_intestine	Y
NCI-H630	large_intestine	Y
NCI-H716	large_intestine	Y
NCI-H747	large_intestine	Y
RCM-1	large_intestine	Y
RKO	large_intestine	Y
SK-CO-1	large_intestine	Y
SNU-C1	large_intestine	Y
SNU-C2B	large_intestine	Y
SW1116	large_intestine	Y
SW1417	large_intestine	Y
SW403	large_intestine	Y
SW48	large_intestine	Y
SW620	large_intestine	Y
SW837	large_intestine	Y
SW948	large_intestine	Y
T84	large_intestine	Y
C3A	liver	Y
HLE	liver	
HuH-6-clone5	liver	
HuH-7	liver	
SK-HEP-1	liver	Y
SNU-387	liver	Y
SNU-398	liver	
SNU-423	liver	
SNU-449	liver	Y
SNU-475	liver	Y
T24	liver	
A549	lung	Y
ABC-1	lung	Y
BEN	lung	Y
CAL-12T	lung	Y
Calu-1	lung	Y
Calu-3	lung	Y
Calu-6	lung	Y
ChaGo-K-1	lung	
COLO-668	lung	Y
COR-L105	lung	Y
COR-L23	lung	
COR-L279	lung	Y
COR-L47	lung	
COR-L88	lung	Y
DMS-114	lung	
DMS-153	lung	
DMS-273	lung	
DMS-53	lung	Y
DMS-79	lung	Y
DV-90	lung	Y
EBC-1	lung	
EKVX	lung	Y
EPLC-272H	lung	Y
HOP-62	lung	Y
HOP-92	lung	Y
IST-SL1	lung	Y
IST-SL2	lung	Y
KNS-62	lung	Y
LB647-SCLC	lung	Y
LC-2-ad	lung	
LCLC-103H	lung	
LK-2	lung	Y
LU-134-A	lung	Y
LU-135	lung	Y
LU-139	lung	Y
LU-165	lung	Y
LU-65	lung	Y
LU-99A	lung	
LUDLU-1	lung	
AKI	skin	
BB132-MEL	skin	
BSCC-93	skin	
C32	skin	Y
CAL-39	skin	Y
CHL-1	skin	Y
COLO-679	skin	Y
COLO-792	skin	Y
COLO-800	skin	Y
COLO-853	skin	
CP50-MEL-B	skin	Y
CP66-MEL	skin	Y
G-361	skin	Y
GR-M	skin	
HMV-II	skin	Y
IGR-1	skin	Y
IPC-298	skin	Y
IST-MEL1	skin	Y
LB2518-MEL	skin	Y
LB2531-MEL-Z	skin	
LB33-MEL-A	skin	
LB373-MEL-D	skin	Y
Malme-3M	skin	
MEL-HO	skin	
MEL-JUSO	skin	
MLMA	skin	Y
MMAC-SF	skin	Y
MZ2-mel	skin	
RVH-421	skin	Y
SK-MEL-2	skin	Y
SK-MEL-24	skin	Y
SK-MEL-28	skin	Y
SK-MEL-3	skin	Y
SK-MEL-30	skin	Y
SR	skin	Y
T107	skin	
UACC-257	skin	Y
UACC-62	skin	Y
WM-115	skin	Y
HUTU-80	small_intestine	Y
A204	soft_tissue	Y
A673	soft_tissue	Y
G-401	soft_tissue	Y
G-402	soft_tissue	Y
GCT	soft_tissue	Y
HT-1080	soft_tissue	Y
MES-SA	soft_tissue	Y
RH18	soft_tissue	Y
RMS	soft_tissue	
S-117	soft_tissue	Y
SJRH30	soft_tissue	Y
SK-LMS-1	soft_tissue	Y
SK-UT-1	soft_tissue	Y
SW684	soft_tissue	Y
SW872	soft_tissue	Y
SW982	soft_tissue	Y
T-174	soft_tissue	
TE-159-T	soft_tissue	
TE-441-T	soft_tissue	Y
VA-ES-BJ	soft_tissue	Y
23132-87	stomach	Y
A3-KAW	stomach	Y
AGS	stomach	Y
ECC10	stomach	Y
ECC12	stomach	Y
GCIY	stomach	Y
GT3TKB	stomach	Y
KATOIII	stomach	Y
MKN1	stomach	Y
MKN28	stomach	Y
MKN45	stomach	Y
MKN7	stomach	Y
NCI-N87	stomach	Y
NCI-SNU-1	stomach	Y
NCI-SNU-16	stomach	Y
NCI-SNU-5	stomach	Y
NUGC-3	stomach	Y
SCH	stomach	Y
TGBC11TKB	stomach	Y
ITO-II	testis	Y
NEC8	testis	Y
NTERA-S-cl-	testis	Y

D1		
AN3-CA	endometrium	Y
COLO-684	endometrium	Y
EFE-184	endometrium	Y
ESS-1	endometrium	Y
HEC-1	endometrium	Y
KLE	endometrium	Y
MFE-280	endometrium	
MFE-296	endometrium	
RL95-2	endometrium	Y
SNG-M	endometrium	Y
WERI-Rb-1	eye	Y
Y79	eye	
380	haem_and_lymph	
697	haem_and_lymph	Y
ALL-PO	haem_and_lymph	Y
ARH-77	haem_and_lymph	Y
ATN-1	haem_and_lymph	Y
BALL-1	haem_and_lymph	Y
BC-1	haem_and_lymph	Y
BC-3	haem_and_lymph	
BE-13	haem_and_lymph	Y
BL-41	haem_and_lymph	Y
BL-70	haem_and_lymph	Y
BONNA-12	haem_and_lymph	Y
BV-173	haem_and_lymph	
C8166	haem_and_lymph	Y
CESS	haem_and_lymph	Y
CMK	haem_and_lymph	Y
CML-T1	haem_and_lymph	Y
CRO-AP5	haem_and_lymph	Y
CTV-1	haem_and_lymph	Y
Daudi	haem_and_lymph	Y
DB	haem_and_lymph	
DEL	haem_and_lymph	Y
DG-75	haem_and_lymph	Y
DOHH-2	haem_and_lymph	Y
EB-2	haem_and_lymph	Y
EB-3	haem_and_lymph	Y
EHEB	haem_and_lymph	Y
EM-2	haem_and_lymph	Y
EoL-1-cell	haem_and_lymph	Y
GA-10-Clone-20	haem_and_lymph	Y
GA-10-Clone-4	haem_and_lymph	Y
GDM-1	haem_and_lymph	Y
GR-ST	haem_and_lymph	Y
H9	haem_and_lymph	
HC-1	haem_and_lymph	Y
HDLM-2	haem_and_lymph	Y
HD-MY-Z	haem_and_lymph	Y
HEL-92-1-7	haem_and_lymph	
HH	haem_and_lymph	Y
HL-60	haem_and_lymph	Y
HT	haem_and_lymph	Y
IM-9	haem_and_lymph	Y
J45-01	haem_and_lymph	
JiyoyeP-2003	haem_and_lymph	Y
JVM-2	haem_and_lymph	Y
JVM-3	haem_and_lymph	Y
K052	haem_and_lymph	Y
K-562	haem_and_lymph	Y
KARPAS-299	haem_and_lymph	Y
KARPAS-422	haem_and_lymph	Y
KARPAS-45	haem_and_lymph	Y
KASUMI-1	haem_and_lymph	Y
KE-37	haem_and_lymph	Y
KM-H2	haem_and_lymph	Y
KMOE-2	haem_and_lymph	Y
KMS-12-BM	haem_and_lymph	Y
KU812	haem_and_lymph	Y
KY821	haem_and_lymph	Y
L-363	haem_and_lymph	Y
L-540	haem_and_lymph	Y
LAMA-84	haem_and_lymph	Y
LC4-1	haem_and_lymph	Y
LOUCY	haem_and_lymph	Y
LP-1	haem_and_lymph	Y
MC-1010	haem_and_lymph	Y
MEG-01	haem_and_lymph	Y
MHH-CALL-2	haem_and_lymph	Y
MHH-CALL-4	haem_and_lymph	
LXF-289	lung	
NCI-H1048	lung	
NCI-H1062	lung	
NCI-H1092	lung	
NCI-H1173	lung	
NCI-H1184	lung	
NCI-H128	lung	
NCI-H1284	lung	
NCI-H1299	lung	
NCI-H1304	lung	Y
NCI-H1355	lung	
NCI-H1417	lung	Y
NCI-H146	lung	Y
NCI-H1563	lung	Y
NCI-H157	lung	Y
NCI-H1573	lung	Y
NCI-H1581	lung	Y
NCI-H1618	lung	Y
NCI-H1623	lung	Y
NCI-H1648	lung	Y
NCI-H1650	lung	Y
NCI-H1651	lung	Y
NCI-H1666	lung	Y
NCI-H1693	lung	Y
NCI-H1694	lung	Y
NCI-H1703	lung	Y
NCI-H1734	lung	Y
NCI-H1755	lung	Y
NCI-H1770	lung	Y
NCI-H1792	lung	Y
NCI-H1793	lung	Y
NCI-H1838	lung	Y
NCI-H187	lung	Y
NCI-H1882	lung	Y
NCI-H1926	lung	Y
NCI-H1930	lung	Y
NCI-H1975	lung	Y
NCI-H2009	lung	Y
NCI-H2029	lung	Y
NCI-H2030	lung	
NCI-H2052	lung	Y
NCI-H2073	lung	
NCI-H2085	lung	
NCI-H2107	lung	Y
NCI-H2141	lung	
NCI-H2170	lung	
NCI-H2195	lung	
NCI-H2227	lung	Y
NCI-H2228	lung	Y
NCI-H226	lung	Y
NCI-H2291	lung	Y
NCI-H23	lung	Y
NCI-H2330	lung	Y
NCI-H2342	lung	
NCI-H2347	lung	Y
NCI-H250	lung	Y
NCI-H292	lung	
NCI-H322M	lung	Y
NCI-H345	lung	Y
NCI-H358	lung	Y
NCI-H378	lung	
NCI-H441	lung	Y
NCI-H446	lung	Y
NCI-H460	lung	Y
NCI-H510A	lung	Y
NCI-H522	lung	Y
NCI-H596	lung	Y
NCI-H64	lung	Y
NCI-H650	lung	Y
NCI-H661	lung	Y
NCI-H69	lung	Y
NCI-H719	lung	Y
NCI-H720	lung	Y
NCI-H727	lung	Y
NCI-H748	lung	
NCI-H774	lung	Y
NCI-H810	lung	Y
NCI-H82	lung	Y
NCI-H835	lung	Y
NCI-H889	lung	Y
8305C	thyroid	Y
8505C	thyroid	
BCPAP	thyroid	Y
BHT-101	thyroid	Y
CAL-62	thyroid	Y
CGTH-W-1	thyroid	Y
FTC133	thyroid	
K5	thyroid	Y
RO82-W-1	thyroid	Y
TCO-1	thyroid	
TT	thyroid	Y
22-RVI	unknown	
CAR-1	unknown	
COR-L24	unknown	
HTCC3	unknown	
IA5	unknown	
LC-1-SQ-SF	unknown	
LCLC-97TM1	unknown	
MFN-INO	unknown	
NB-I-IFO	unknown	
NCI-H417	unknown	
SW1315	unknown	
TCYIK	unknown	
TE-206-T	unknown	Y
TuHR-14-TKB	unknown	
NB-12	unknown (auto_ganglia)	Y
NB-SuS-SR	unknown (auto_ganglia)	Y
SK-N-FI	unknown (auto_ganglia)	Y
D-245-MG	unknown (CNS)	Y
D-263-MG	unknown (CNS)	Y
D-336-MG	unknown (CNS)	Y
D-384-MED	unknown (CNS)	Y
D-392-MG	unknown (CNS)	Y
D-397-MG	unknown (CNS)	Y
D-423-MG	unknown (CNS)	Y
D-458-MED	unknown (CNS)	Y
D-502-MG	unknown (CNS)	Y
D-538-MG	unknown (CNS)	Y
D-556-MED	unknown (CNS)	Y
D-566-MG	unknown (CNS)	Y
ECC4	unknown (gastro_tract)	Y
PLCPRF5	unknown (liver)	Y
SW626	unknown (ovary)	Y
PANC-08-13	unknown (pancreas)	Y
A2058	unknown (skin)	Y
A388	unknown (skin)	Y
A431	unknown (skin)	Y
GAK	unknown (skin)	Y
RPMI-7951	unknown (skin)	Y
SK-MEL-5	unknown (skin)	Y
A253	upper_aero_tract	
BB30-HNC	upper_aero_tract	Y
BB49-HNC	upper_aero_tract	Y
BHY	upper_aero_tract	Y
Ca9-22	upper_aero_tract	Y
CAL-27	upper_aero_tract	Y
CAL-33	upper_aero_tract	Y
Detroit562	upper_aero_tract	Y
DOK	upper_aero_tract	Y
HCE-T	upper_aero_tract	Y
HN	upper_aero_tract	Y
HO-1-N-1	upper_aero_tract	Y
KOSC-2	upper_aero_tract	Y
LB771-HNC	upper_aero_tract	Y
RPMI-2650	upper_aero_tract	Y
SAS	upper_aero_tract	Y
SCC-15	upper_aero_tract	Y
SCC-25	upper_aero_tract	Y
SCC-9	upper_aero_tract	Y
5637	urinary_tract	Y
639-V	urinary_tract	Y
647-V	urinary_tract	Y
BFTC-905	urinary_tract	Y
DSH1	urinary_tract	Y
HT-1197	urinary_tract	Y
HT-1376	urinary_tract	Y
KU-19-19	urinary_tract	
RT-112	urinary_tract	Y
RT4	urinary_tract	Y
SW1710	urinary_tract	

MJ	haem_and_lymph	Y	PC-14	lung		SW780	urinary_tract	Y
ML-2	haem_and_lymph	Y	PC-3	lung	Y	UM-UC-3	urinary_tract	Y
MN-60	haem_and_lymph	Y	RERF-LC-FM	lung	Y	SW954	vulva	Y
MOLT-13	haem_and_lymph	Y	SBC-1	lung	Y	SW962	vulva	
MOLT-16	haem_and_lymph	Y	SBC-5	lung	Y			

Publication

The following publication is reprinted from Cell 133(4): 727-41. Uren A.G., Kool, J., Matentzoglu, K., de Ridder, J., Mattison, J., van Uitert, M., Lagcher, W., Sie, D., Tanger, E., Cox, T., Reinders, M., Hubbard, T.J., Rogers, J., Jonkers, J., Wessels, L., Adams, D.J., van Lohuizen, M., and Berns, A. Large-scale mutagenesis in p19(ARF)- and p53-deficient mice identifies cancer genes and their collaborative networks. Copyright © 2008, with permission from Elsevier.