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).

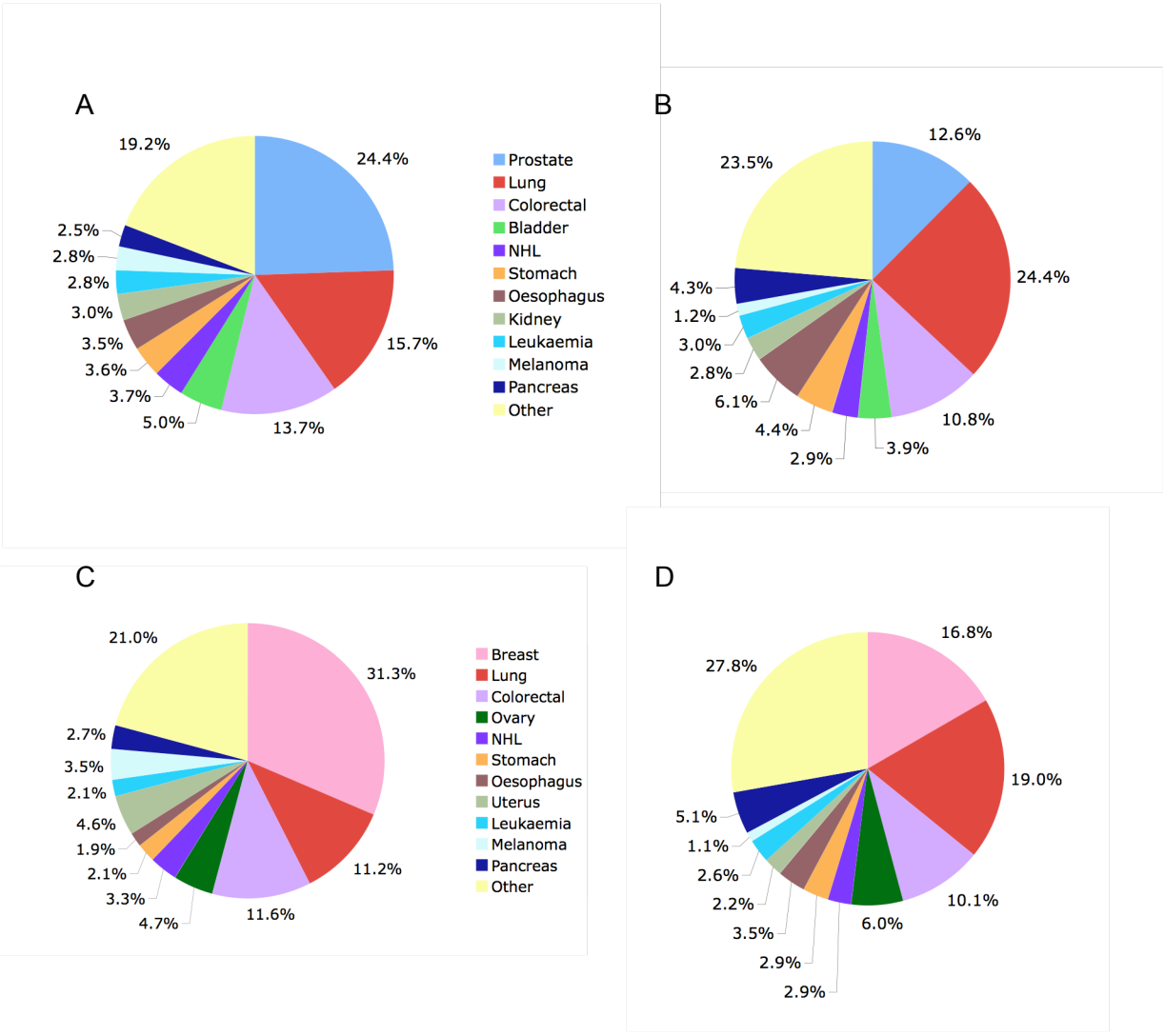

Figure 1.1. Summary of cancer incidence in 2004 and deaths from cancer in 2005 for the most common sites of cancer in males and females in the UK. Cancer incidence and mortality among males are shown in Figures A and B, respectively. Cancer incidence and mortality among females are shown in Figures C 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.

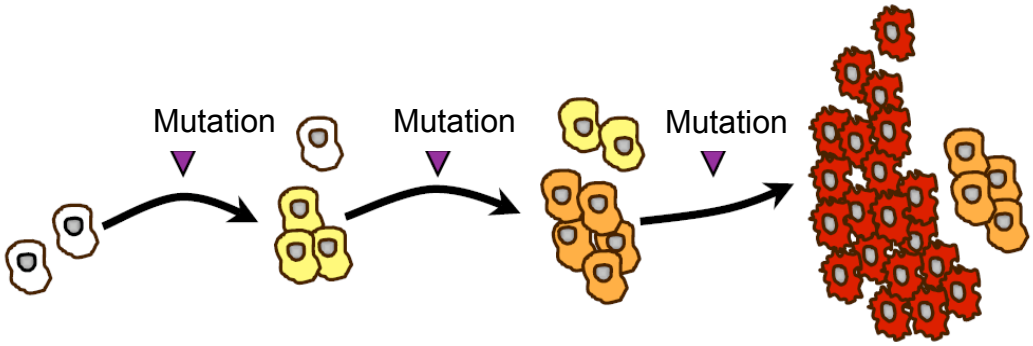

Figure 1.2. The clonal evolution of cancer. Tumours evolve from a single cell through a series of stepwise genetic changes within the original clone. Cells containing mutations that confer the greatest growth advantage are selected and become the dominant clone. Adapted from a figure supplied by D.J.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

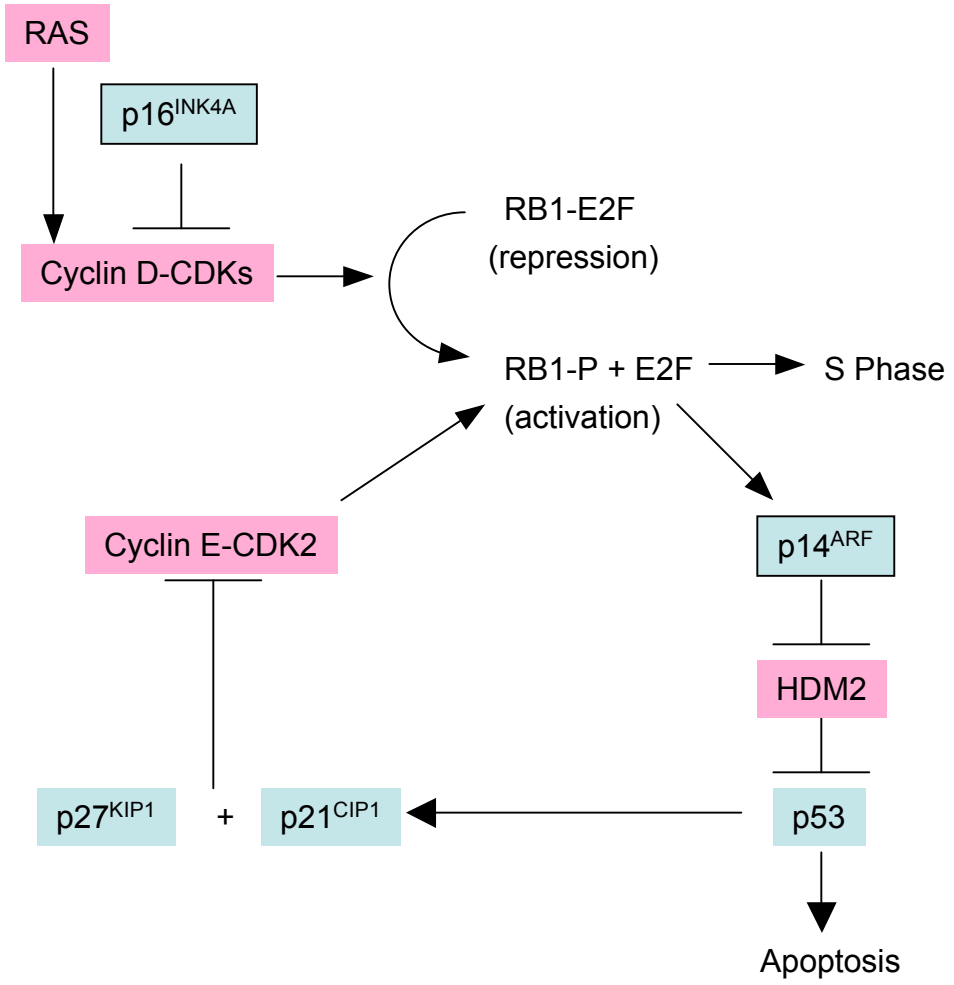

Figure 1.3. Mutations in different genes in the same pathway can have an equivalent effect. The figure shows a simple representation of the p53 and RB1 genetic pathways. The pathways are coupled through the *INK4A/ARF* locus, which encodes $p16^{INK4A}$ and $p14^{ARF}$, shown here in black boxes, and 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: $p16^{INK4A}$, described above, and $p14^{ARF}$ (also known as $p19^{ARF}$ 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).

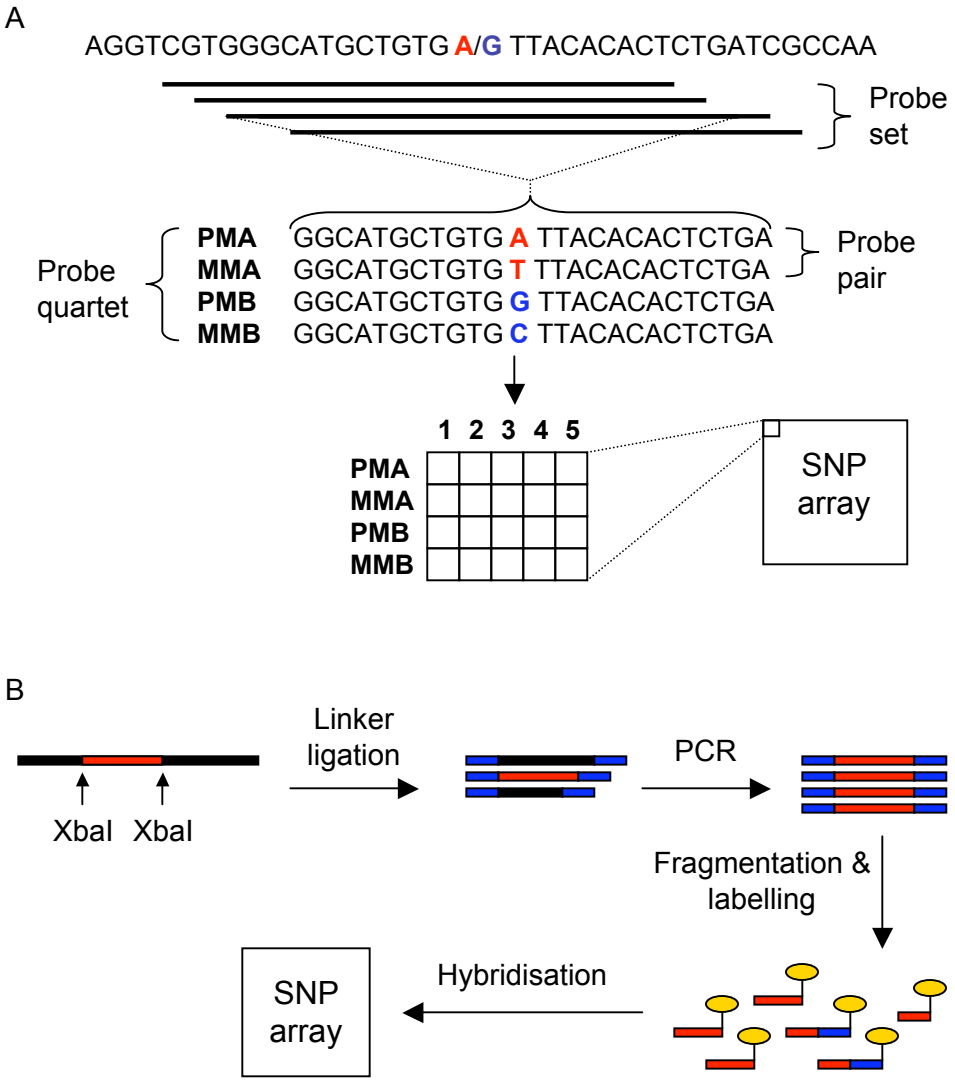

Figure 1.4. Array design (A) and whole-genome sampling assay (B) for the Affymetrix SNP array. A. A SNP in the DNA sequence is shown in red/blue. The SNP is represented in the array by a probe set, which comprises multiple probe quartets that differ from one another in the position of the polymorphic 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.

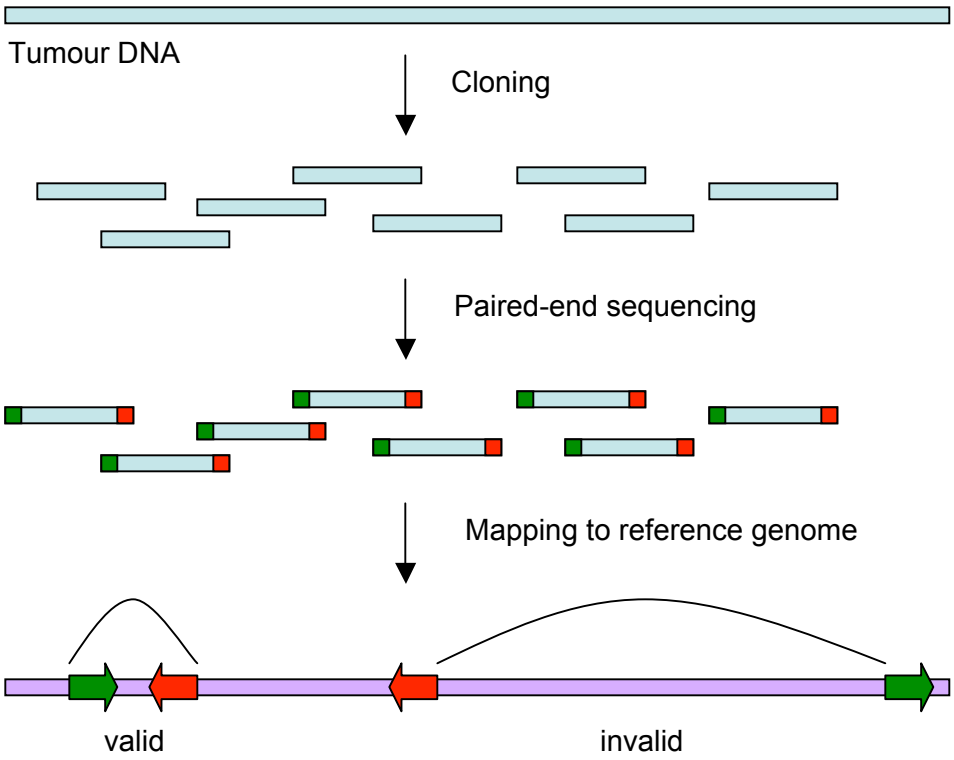

Figure 1.5. End sequence profiling of tumour DNA. 100-250 kb regions of the tumour genome are cloned and a 500 bp region at the end of each clone is sequenced. The ends are mapped to the human reference genome. Ends that are an abnormal distance apart or in an abnormal orientation, shown here as "invalid", 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

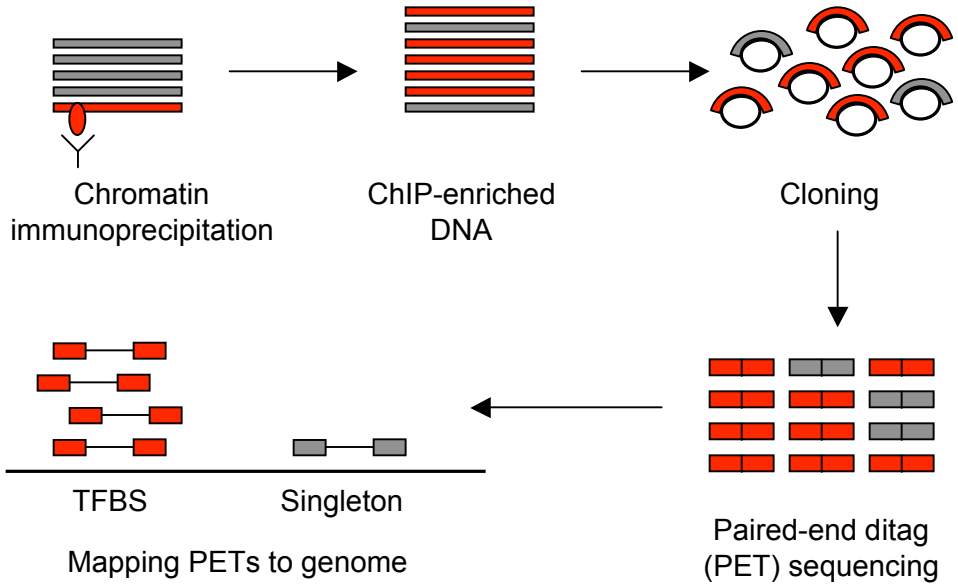

Figure 1.6. Overview of ChIP-PET for mapping transcription factor binding sites. In chromatin precipitation (ChIP), the chromatin is fragmented by sonication and an antibody against the transcription factor of interest is used to precipitate the transcription factor bound to DNA. The ChIP-enriched DNA 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).

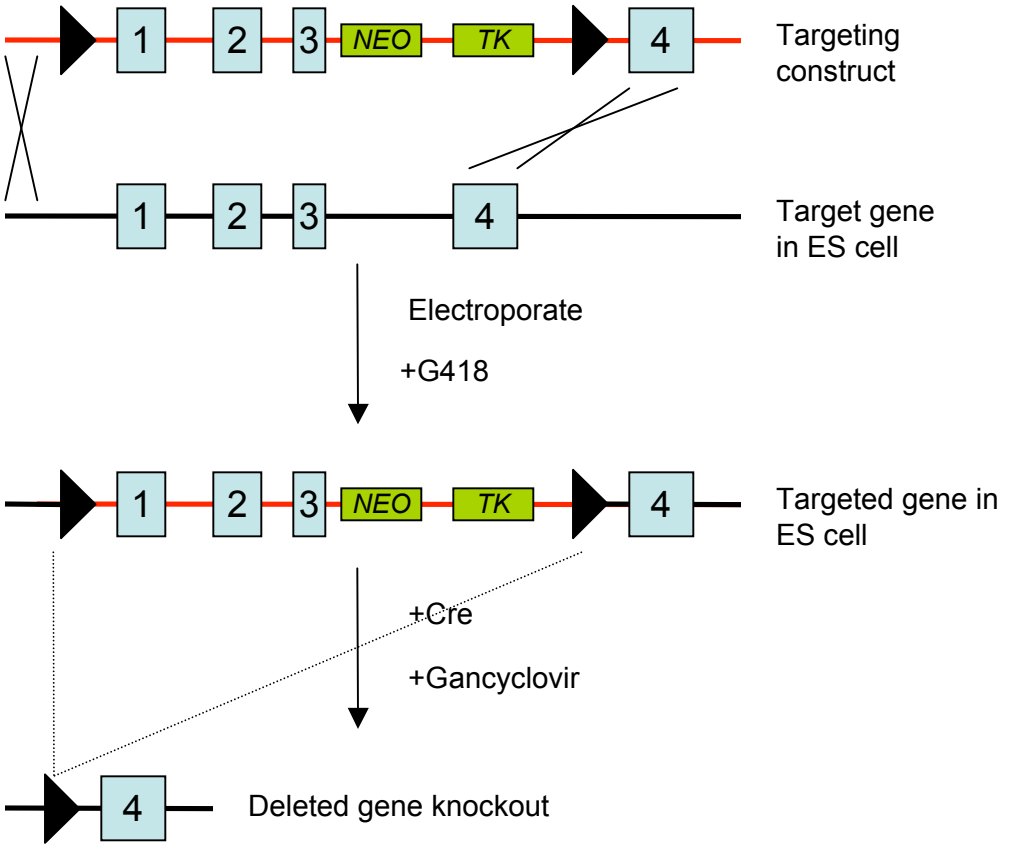

Figure 1.7. Generation of a conditional knockout allele in ES cells. A targeted gene construct is designed that contains loxP sites flanking the region of the gene to be deleted as well as genes for selection. Upon introduction into ES cells, DNA in the construct replaces endogenous DNA in the target 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).

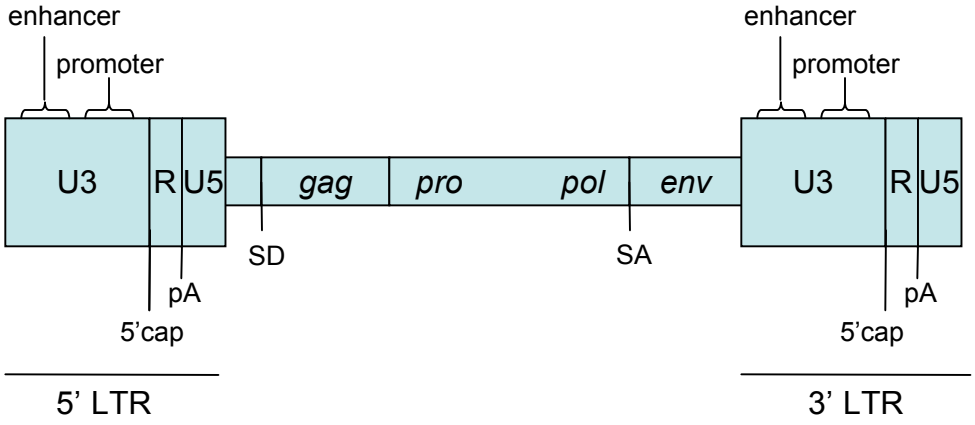

Figure 1.8. Structure of a retroviral provirus. The provirus contains two long terminal repeats (LTRs) flanking the genes required for viral assembly. Elements within the LTRs drive transcription of the viral genes but can also induce mutation of nearby cellular genes. Splicing of a viral splice donor (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

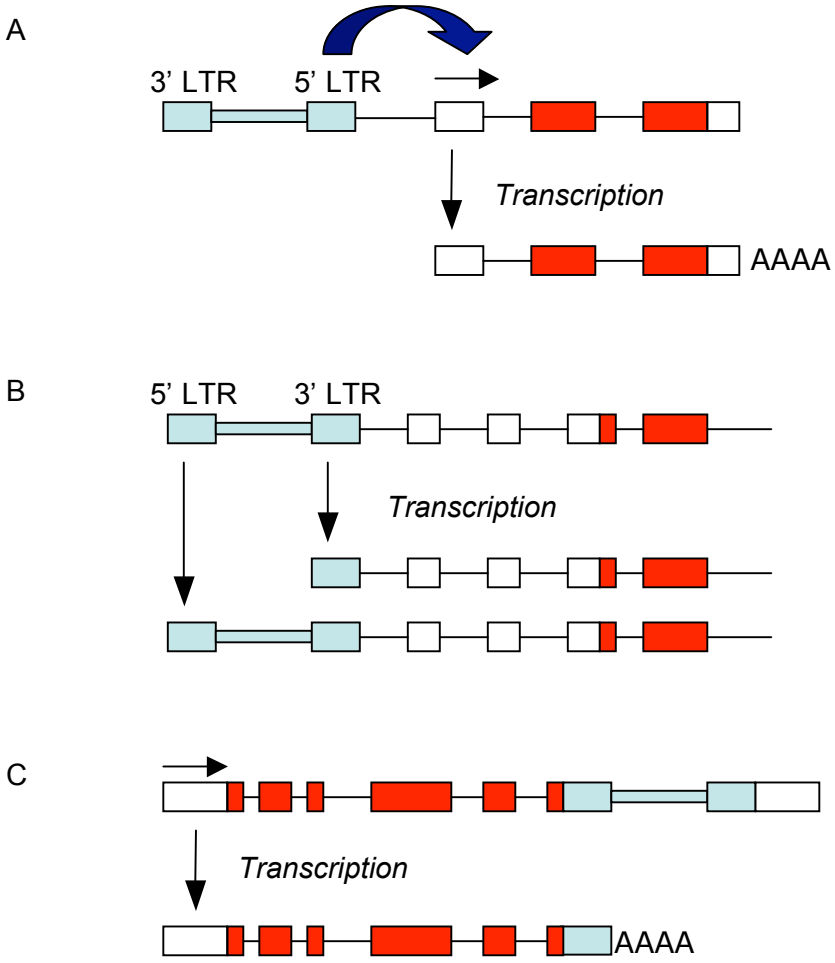

Figure 1.9. The mechanisms of mutagenesis of murine leukaemia virus include enhancer mutation (A), promoter mutation (B) and premature termination of gene transcription (C). The provirus is shown in blue; coding and non-coding exons are shown in red and white, respectively. **A.** An enhancer element 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).

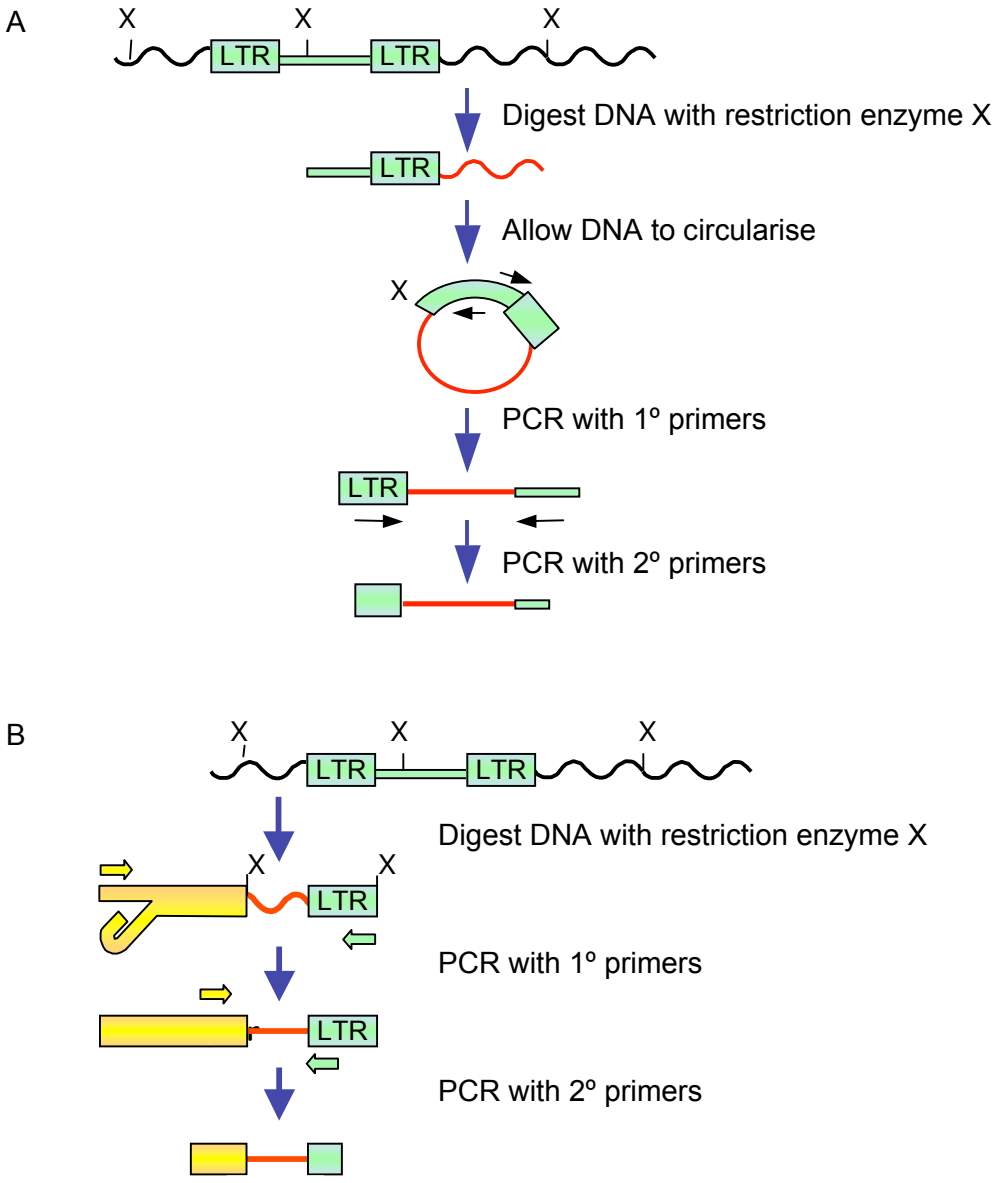

Figure 1.10 Isolation of retroviral insertion sites by inverse PCR (A) and splinkerette PCR (B). In inverse PCR, tumour DNA is digested using restriction enzyme X and the restricted DNA is allowed to circularise. Genomic DNA flanking retroviral insertions are amplified using PCR primers that bind within 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 *Plag1* 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 *Eμ*Myc 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.

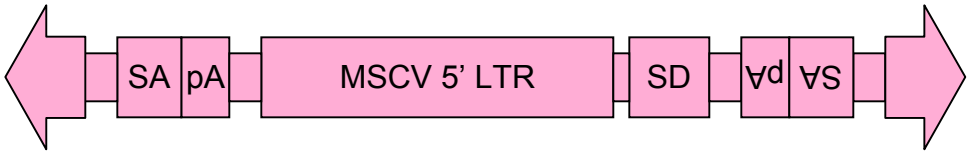

Figure 1.11. Structure of the *Sleeping Beauty* transposon. The presence of splice acceptors (SA) and polyadenylation signals (pA) in both orientations enables premature termination of gene transcription from intragenic insertions in both orientations. The transposon also contains the murine stem cell 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.