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

3.1 Introduction

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

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

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

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

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

3.2.1 Description of the datasets

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

3.2.1.1 The Retrovirus Tagged Cancer Gene Database (RTCGD)

As mentioned in Section 1.4.2.1.2, RTCGD (Akagi *et al.*, 2004; http://rtcgd.abcc.ncifcrf.gov/) is a database that manages data from retroviral and transposon-mediated insertional mutagenesis screens. All candidate cancer genes in the database were obtained from the website on 01/11/07. In total, the database contained 537 genes with unique MGI (Mouse Genome Informatics, http://www.informatics.jax.org/) symbols identified from 512 retroviral CISs (25 CISs

had been assigned to 2 genes). The MGI symbols were used to identify 544 mouse Ensembl genes. 16 genes had 2 Ensembl identifiers (e.g. *Akap13*, which, according to Ensembl, is duplicated in 2 adjacent copies) and 9 could not be identified. 55 genes were identified from 52 transposon CISs (3 had been assigned to 2 genes) and all but 2 had Ensembl gene identifiers.

3.2.1.2 The Catalogue of Somatic Mutations in Cancer (COSMIC)

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

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

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

3.2.1.4 Transcription factor binding sites

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

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

3.2.1.5 Amplicons and deletions in paediatric acute lymphoblastic leukaemias

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

3.2.1.6 Gene ontology (GO) terms

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

3.2.2 Comparison with insertional mutagenesis data

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

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

and 6th most frequently mutated genes in RTCGD and yet they were not among the candidates identified in the MuLV screen. Almost all of the *Sox2* insertions were identified in mice with an AKxD or BXH2 strain background, while all of the *Fgf3* insertions were from a screen that used mouse mammary tumour virus (MMTV), rather than MuLV, as the retroviral mutagen.

Human orthologues of mouse candidate genes were enriched among oncogenes in the Cancer Gene Census (36 oncogenes, $P=7.88\text{x}10^{-18}$) and the COSMIC database (69 genes, $P=1.36\text{x}10^{-9}$). There were no recessive cancer genes among the candidates ($P=1$), demonstrating that the screen identifies predominantly oncogenes. Surprisingly, there were just 3 genes (*Lrrfip1*, *Nup214* and *Bcl11a*; $P=0.74$) that overlapped between the candidates of the retroviral screen and the candidates from Sjöblom *et al* (2006). This may reflect the fact that the Sjöblom dataset was an exon resequencing study of breast and bowel tumours exclusively, and it may be biased against genes mutated in lymphomas.

The 36 orthologues of mouse candidate cancer genes in the Cancer Gene Census were enriched for genes that are mutated in lymphoid tumours (31 genes, $P=2.66\text{x}10^{-4}$). This suggests that the retroviral screen mainly identifies genes that are important in the development of lymphoid malignancies. There was also a slight enrichment of genes that are mutated by chromosomal translocation, although this was not significant (31 genes, $P=0.067$). A more significant association might be expected because translocation is a common mechanism of mutation in lymphoid cancers, and a number of genes that are frequently targeted by insertional mutagenesis are involved in translocations in human tumours. In addition, MuLV mutagenesis may have a similar effect to translocations, since it often changes the regulatory environment of a gene and/or produces truncated oncoproteins. Chromosomal translocations, and leukaemias, lymphomas and mesenchymal tumours, all of which frequently harbour translocations, are over-represented in the Cancer Gene Census. This is partly because both translocation partners feature in the list of cancer genes, but also because, traditionally, cancer gene identification has been more frequently performed in these cancer types (Futreal, 2007). This may explain why the candidate cancer genes identified in the retroviral screen contain an over-representation of genes in the Cancer Gene Census, but that translocations are not over-represented among these candidates in the Census. Finally, there was an over-representation of known cancer genes that bear somatic mutations in

human cancer (36 genes, P=0.0130). These findings demonstrate the efficacy of the MuLV retrovirus as a somatic mutagen that can be used to model the clonal evolution of human cancers, particularly those of lymphoid origin.

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

Mouse candidate genes with human orthologues were also over-represented in regions of copy number gain (54 genes, P=0.0180) and copy number loss (47 genes, P=5.82x10^{-3}) in paediatric ALL. CIS genes that were deleted in ALL included *Lef1*, *Ikzf1*, *Ikzf3*, *Etv6*, *Elf1* and *Erg*, while those that were amplified included *Runx1*, *Myb* and *Ahi1*. This suggests that genes that are mutated by insertional mutagenesis, and contribute to mouse tumourigenesis, may also be mutated by copy number changes in human cancers. However, the overlapping genes are implicated in B-cell development and differentiation,

which are disrupted in human B-progenitor ALL and in MuLV-induced murine lymphomagenesis. It therefore remains to be seen whether CIS genes significantly overlap with regions of copy number change in other human cancers, and this is addressed in Chapters 4 and 5. The CIS genes may help to narrow down the candidates in regions of copy number change in the ALL dataset. For example, the deleted region on human chromosome 16q22.1 contains 11 genes, but the mouse orthologue of only 1 of these genes (*FAM65A*) is targeted by MuLV in insertional mutagenesis and therefore represents a putative target for deletion in ALL. Table 3.1 provides a list of the regions that are amplified and deleted in ALL and the CIS genes within these regions.

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

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

A

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

B

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

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

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

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

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

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

Of the 10 candidates genes in the *Sleeping Beauty* dataset, 5 had been previously identified by retroviral insertional mutagenesis in RTCGD and were also identified in the MuLV screen described herein (see Section 3.3), while 3 had been previously identified

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

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

regulation of transposase expression (see Section 1.4.2.2.1). The datasets discussed in this section are further referred to in the proceeding sections and chapters in relation to individual cancer gene candidates.

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

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

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

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

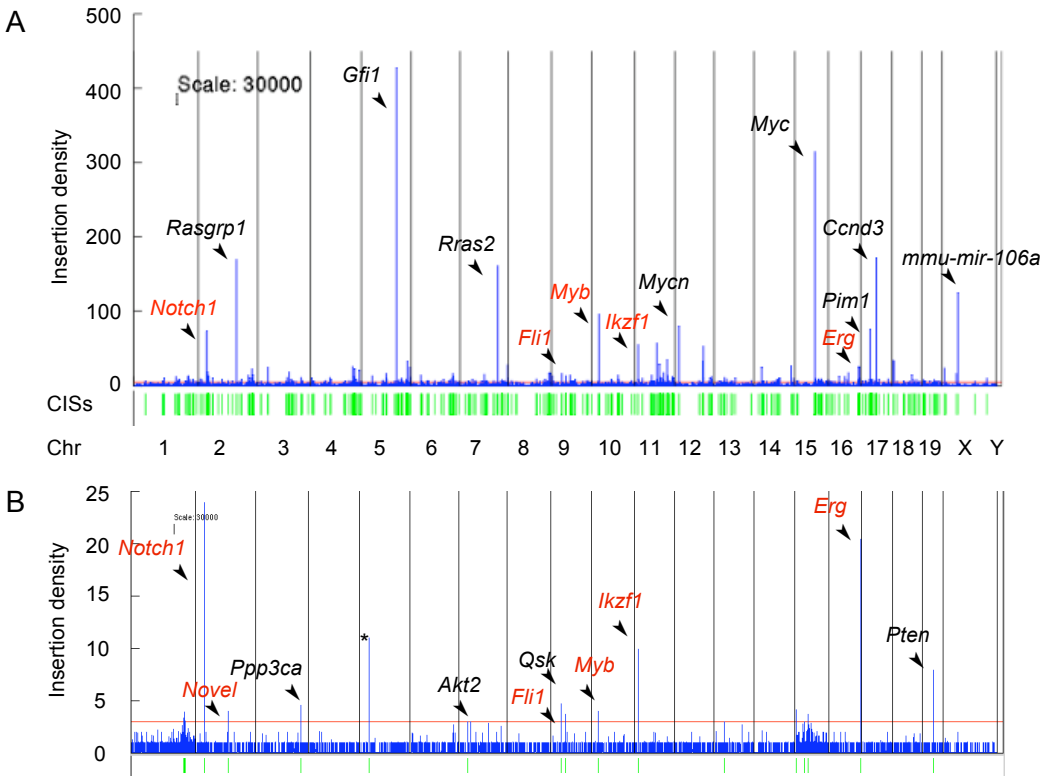

Figure 3.1. MuLV (A) and T2/Onc (B) insertions across the mouse genome. The plots show the density of insertions calculated using the kernel convolution-based method (de Ridder *et al*., 2006) with a kernel width of 30 kb. Common insertion sites (CISs) are shown in green. The red line represents the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

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

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

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

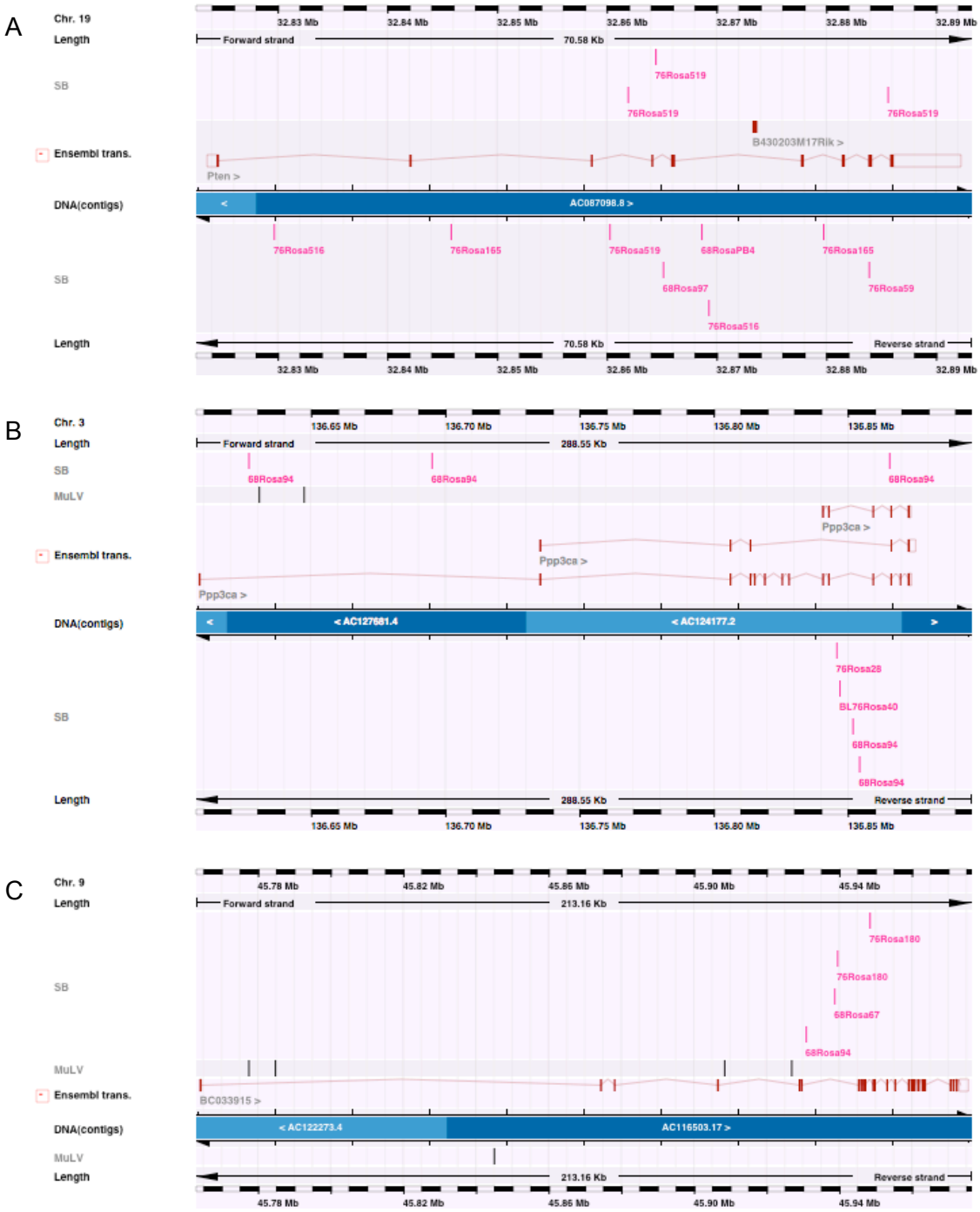

Figure 3.2. Known and putative tumour suppressor genes identified in the *Sleeping Beauty* (SB) screen. *Pten* (A) did not contain any MuLV insertions. *Ppp3ca* (B) and *BC033915* (C) contained MuLV insertions but not in a statistically significant CIS. For all genes, there was at least 1 SB tumour that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

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

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

The gene encoding serine/threonine protein kinase BC033915 (known as QSK in humans) was also identified in both screens, and although the MuLV CIS was not significant, it did contain 5 retroviral insertions (Figure 3.2C). In the COSMIC database, 2 of the 296 human tumour samples that have been tested for mutations in the *QSK* gene contain missense mutations. One heterozygous S882C substitution was identified in the primary renal cell carcinoma PD1583a, which contains just one other small intragenic mutation in 519 genes examined, and a P836S substitution (zygosity unknown) was identified in the non-small cell lung cancer cell line NCI-H1770, which contains 201 small intragenic mutations in 4,688 genes examined. 1 silent mutation (heterozygous

substitution R476R) was also identified in the malignant melanoma cell line MZ7-mel, but this line contains 428 mutations in 4,668 genes examined and therefore appears to have a hypermutable phenotype. Both missense mutations are located in a glutamine-rich region (Prosite profile PS50322). All of the retroviral and transposon insertions in *QSK* precede this region and may therefore produce truncated gene transcripts in which the region is missing. 2 heterozygous, missense mutations were also identified in the resequencing study by Sjöblom *et al* (2006). QSK and 11 other kinases related to AMP-activated protein kinase (AMPK) are known to be activated by the tumour suppressor kinase LKB1 (Lizcano *et al.*, 2004). Activation of one of these kinases (MARK1) by LKB1 has been shown to regulate microtubule dynamics by phosphorylating the microtubule-associated protein Tau, thereby reducing the affinity of Tau for microtubules and inhibiting tubulin polymerisation (Kojima *et al.*, 2007). QSK may play a similar role, since RNAi-mediated knockdown of the Drosophila orthologue of *QSK* resulted in spindle and chromosome alignment defects (Bettencourt-Dias *et al.*, 2004). One of the SB-induced lymphomas contained 2 insertion sites in *Qsk*, and this, coupled with the observations described above, suggests that *Qsk* may be a tumour suppressor gene.

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

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

3.4 Determining the mechanisms of MuLV insertional mutagenesis

3.4.1 Analysing the distribution of intragenic insertions

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

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

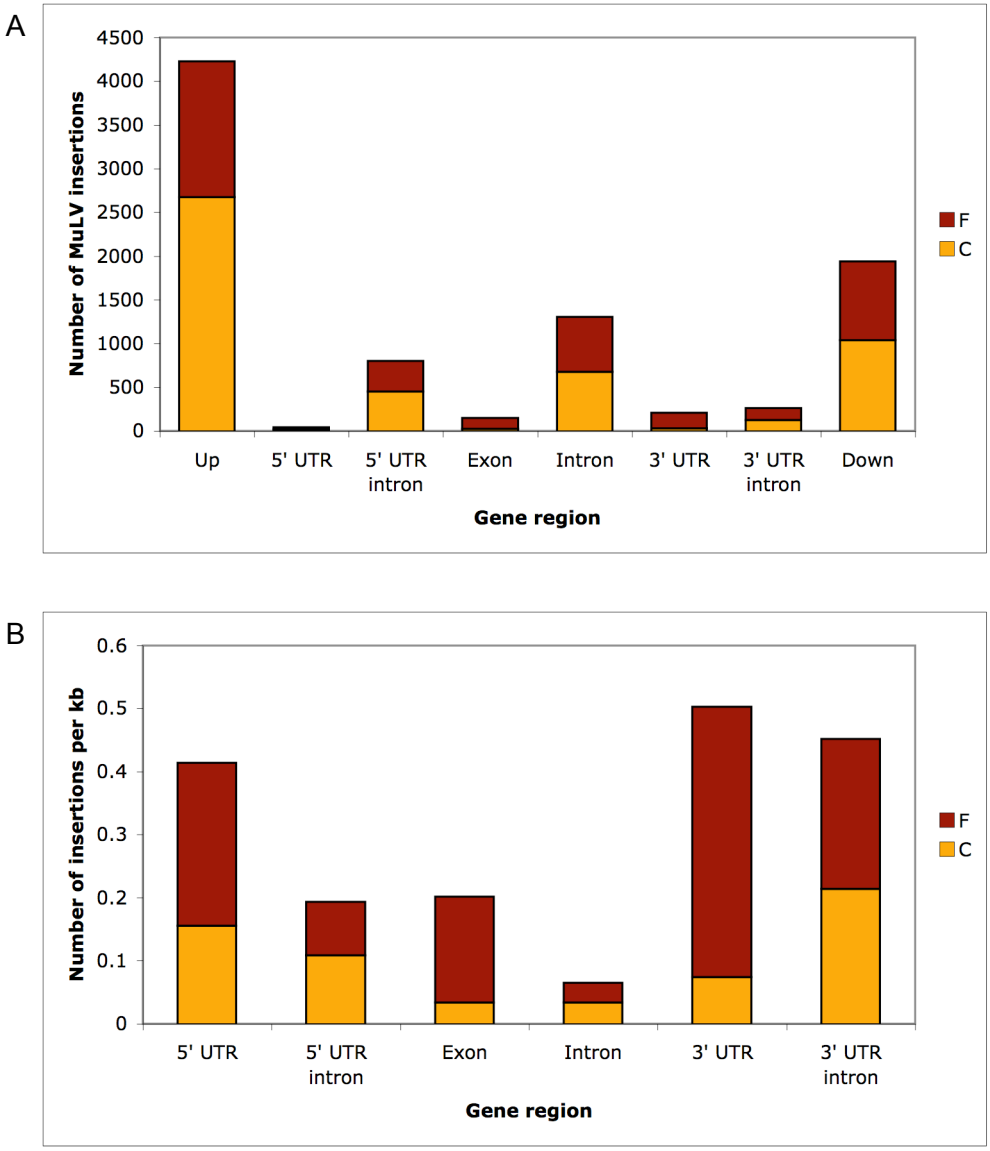

Figure 3.3. The distribution of MuLV insertions within candidate cancer genes. (A) The total number of insertions in each gene region. (B) The number of insertions as a proportion of the total length of the gene region across all candidate genes. The number of insertions in the sense orientation (F)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

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

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

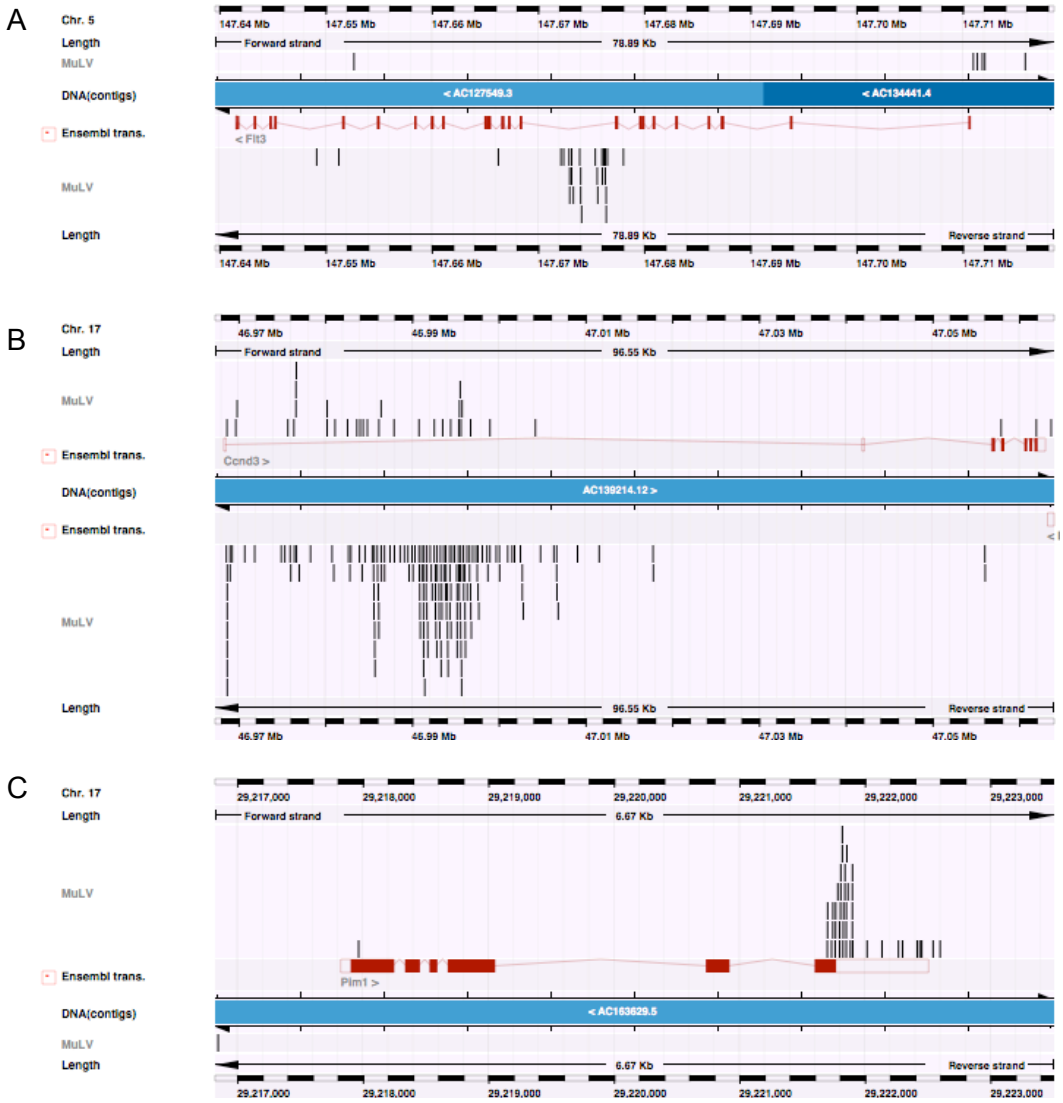

Figure 3.4. Intragenic MuLV insertions in candidate cancer genes. (A) Intronic insertions in *Flt3* are predicted to generate N-terminally truncated gene products. (B) Antisense insertions in the 5' UTR of *Ccnd3* are likely enhancer mutations. (C) Insertions in the final coding exon and 3' UTR of *Pim1*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

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

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

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

For each candidate cancer gene, the distribution of insertions was used to predict the likely mechanisms of mutagenesis and, therefore, the likely structures of mutated gene products. Sense insertions that were upstream of the gene, within the 5' UTR or in an intron flanked by exons of the 5' UTR were classified as promoter mutations. Upstream insertions in the antisense orientation were classified as enhancer mutations. Insertions in the 3' UTR were classified as "stability" mutations, i.e. insertions that may result in the removal of mRNA-destabilising motifs, while sense and antisense insertions in exons or

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

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

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

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

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

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

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

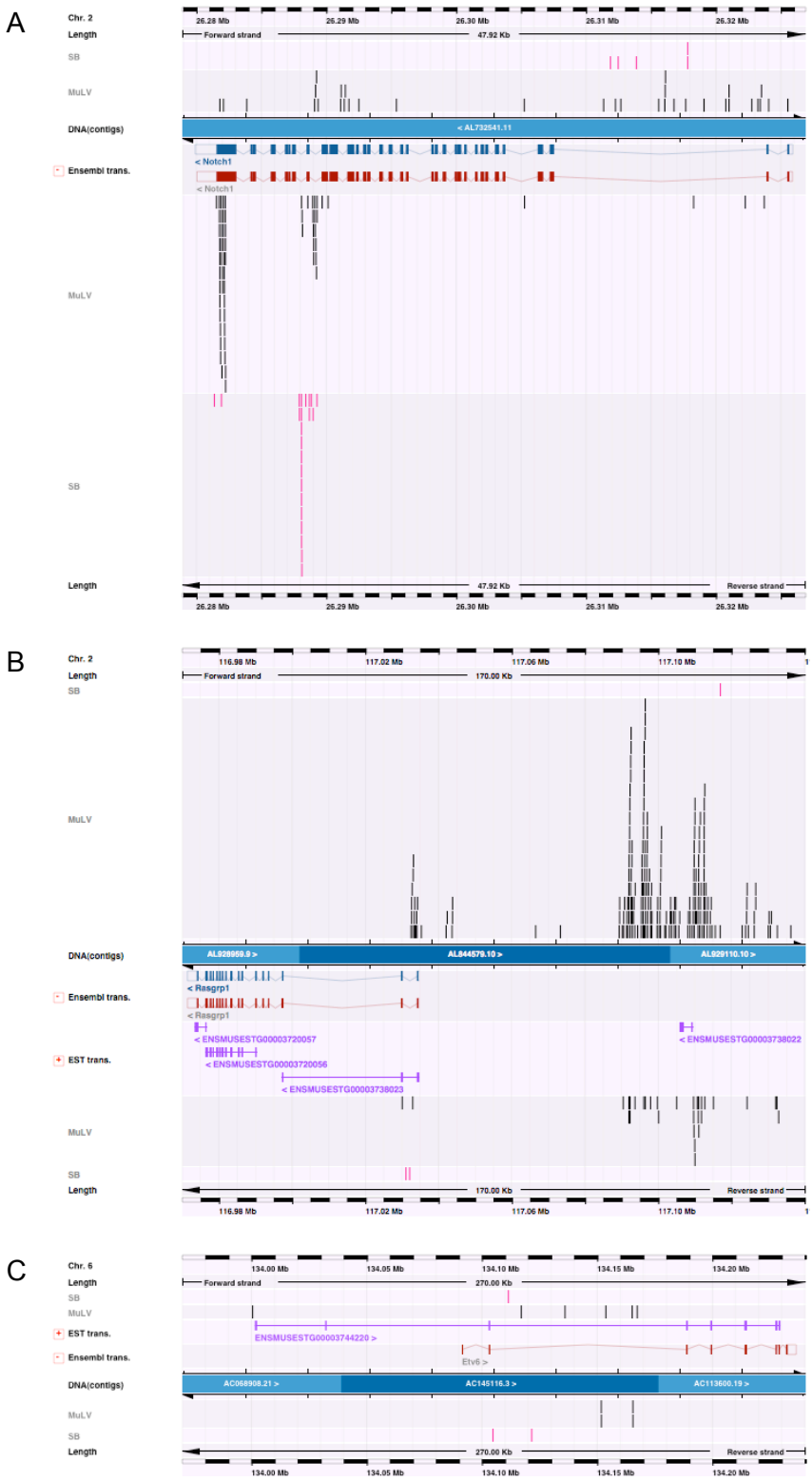

Figure 3.5. Co-occurring MuLV and T2/Onc insertions help to identify the mechanism of mutagenesis of genes *Notch1* (A), *Rasgrp1* (B) and *Etv6* (C). MuLV insertions are shown in black, T2/Onc insertions are shown in pink. Ensembl gene transcripts are shown in red and blue. ESTs are shown in purple.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

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

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

All of the MuLV and T2/Onc insertions identified in gene *Etv6* were in the second intron, in both orientations (Figure 3.5C). Since sense insertions can form both N-terminal and C-terminal truncations but antisense insertions can form only C-terminal truncations, it is likely that where insertions occur in both orientations in the same intron of an oncogene, they are causing premature termination of gene transcription that results in C-terminally truncated gene products. In the case of *Etv6*, this would result in the production of polypeptide lacking both of its functional domains. Etv6 is a transcriptional repressor that is essential for haematopoietic stem cell function. The N-terminal sterile alpha motif/pointed (SAM_PNT) domain (IPR0003118) is responsible for hetero- and

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

The rest of the genes that were disrupted by both MuLV and SB showed variation in the distribution of insertions. In some cases, this reflects differences in the mutational mechanisms of the two mutagens. For example, 1 retroviral insertion and 1 transposon insertion were found to co-occur just upstream of *Myb* (Figure 3.6A) in the sense orientation, where they are likely to be causing promoter mutation, but the vast majority

of retroviral insertions were putative enhancer mutations, occurring upstream in the antisense orientation, or downstream, predominantly in the sense orientation with respect to *Myb*. The presence of MuLV sense and antisense insertions, and 1 SB sense insertion, just upstream of an Ensembl EST gene suggests that, as for *Etv6*, there is an additional *Myb* transcript that has not been annotated as an Ensembl gene transcript. The remaining T2/Onc insertions were intragenic. 3 were in the last (13^{th}) intron in the antisense orientation with respect to *Myb*, while 1 was in the 10^{th} intron in the sense orientation. A C-terminal truncation caused by the latter would truncate the C-terminal Myb domain (IPR015395), which is known to bind the inhibitor Cyp-40 (Leverson and Ness, 1998). Oncogenic *v-Myb* contains a mutated binding site that prevents binding of Cyp-40 and so prevents negative regulation (Leverson and Ness, 1998). The contribution of insertions within the last intron is unclear since a C-terminally truncated protein would contain an intact binding domain. It is possible that the last exon of *Myb* encodes a protein sequence with a hitherto uncharacterised role in oncogenesis.

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

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

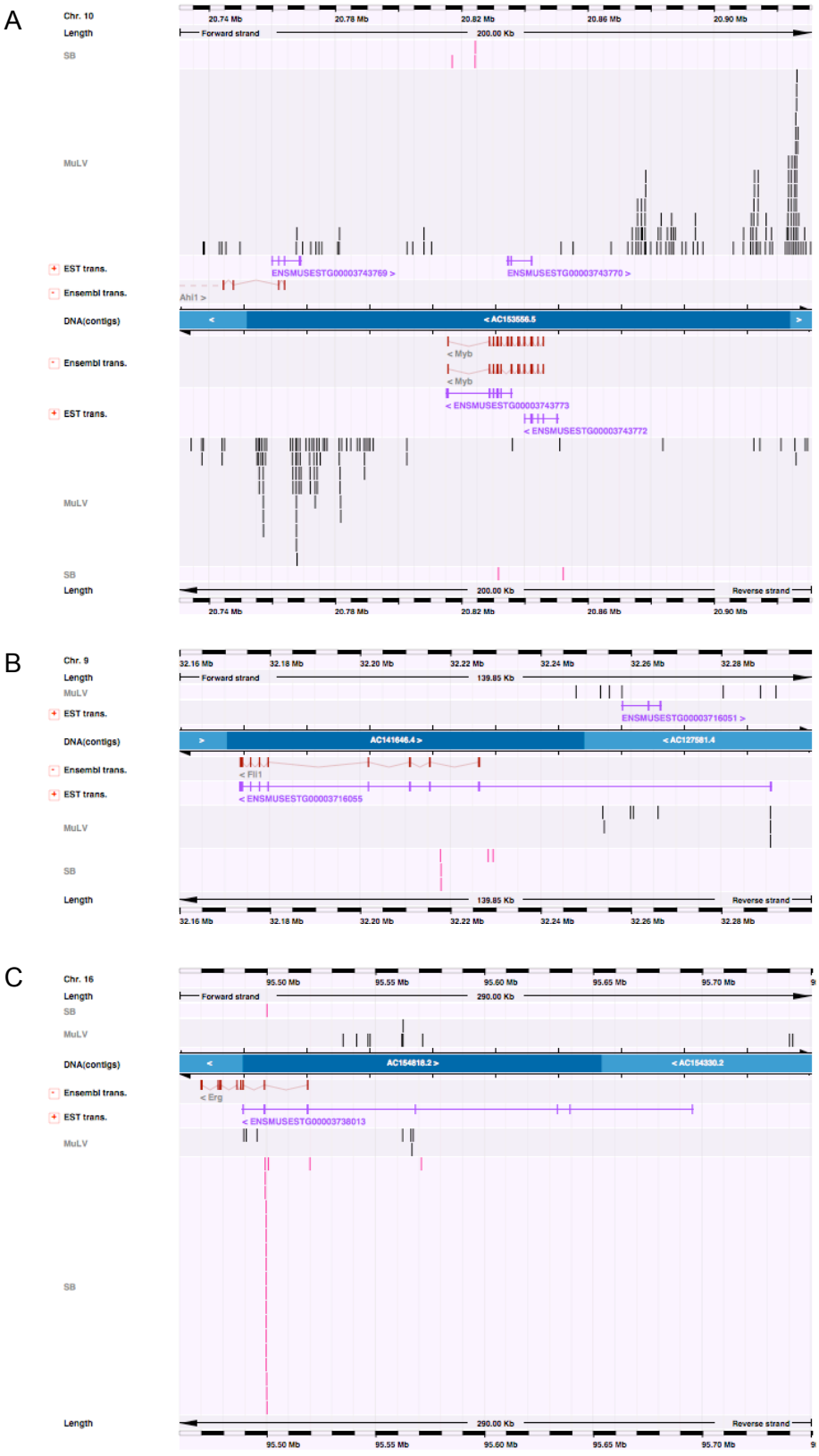

Figure 3.6. Variation in the distribution of MuLV and T2/Onc insertions in *Myb* (A), *Fli1* (B) and *Erg* (C) may reflect differences in the mechanisms of mutagenesis. MuLV insertions are shown in black, T2/Onc insertions are shown in pink. Ensembl gene transcripts are shown in red and blue. ESTs are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

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

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

3.4.3 Identification of tumour suppressor genes inactivated by MuLV

Although retroviral insertional mutagenesis identifies predominantly oncogenes, tumour suppressor genes also featured in the list of candidate cancer genes. The most prevalent, with 93 insertions, was *Ikaros* (*Ikzf1*). *Ikaros* encodes a haematopoietic-specific zinc

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

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

There is no straightforward approach for identifying candidate tumour suppressor genes because, while they are likely to contain only intragenic insertions, some oncogenes, such as *Notch1* and *Pim1*, are also mutated predominantly by intragenic insertions. However, as mentioned in Section 3.4.2, insertions in oncogenes are more likely to form specific oncogenic mutants and may therefore be more localised within the gene. In addition,

tumour suppressor genes are likely to contain multiple insertion sites within the same tumour, as described for *Pten* and *Qsk* in Section 3.3, because both copies of the gene must be inactivated. This does not hold for haploinsufficient tumour suppressor genes, which require only one inactivating insertion for tumourigenesis and are therefore more likely to be identified by insertional mutagenesis than genes requiring an insertion in both genes. 14 tumours contained multiple insertions within *Ikzf1*, while 1 contained multiple insertions in *Ikzf3*. However, none of the other genes so far discussed in this section were mutated by multiple insertions. This suggests either that they are haploinsufficient tumour suppressor genes, as demonstrated for *Wwox*, or that the coverage of the screen was too low, such that multiple insertions occurred but were not identified. To further complicate matters, oncogenes may also contain multiple insertion sites within the same tumour, either because the gene is a preferential target site for the virus, or because upregulation of both gene copies provides an even greater growth advantage to the cell. However, taken together, the distribution of insertions and the number of insertion sites within each tumour can help to identify potential tumour suppressor candidates.

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

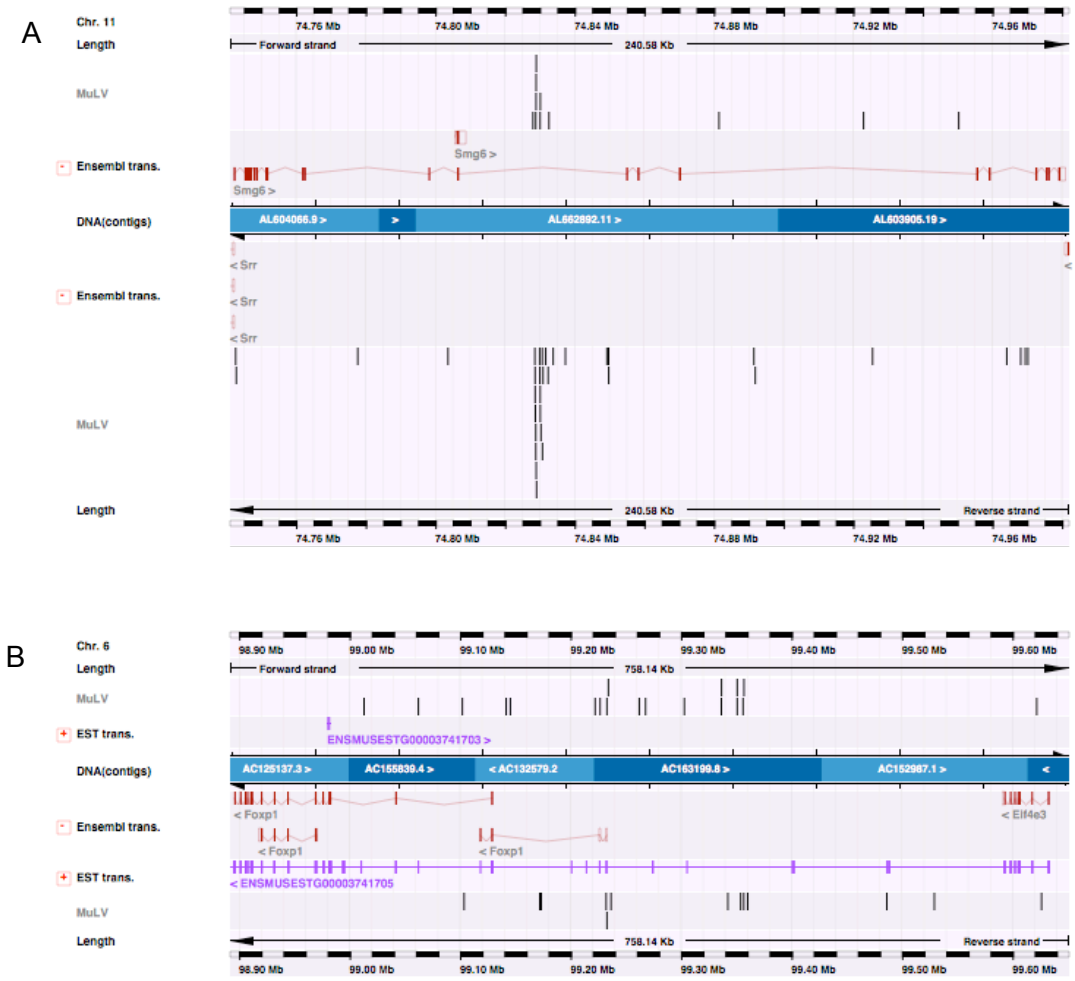

Figure 3.7. ***Smg6*** **(A) and** ***Foxp1*** **(B) are putative tumour suppressor genes identified by MuLV insertional mutagenesis.** MuLV insertions are shown in black. Ensembl gene transcripts are shown in red. ESTs are shown in purple. Insertions above and below the blue bar labelled DNA(contigs)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

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

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

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

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

3.4.4 Identifying retroviral insertions in regulatory features

The orientation and distribution of insertions around genes helps to identify promoter and enhancer mutations and insertions that prematurely terminate gene transcription. However, it is also possible that insertions could disrupt a gene by inserting into regulatory elements, thereby preventing the binding of transcriptional activators or

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

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

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

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

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

Since many insertions may map to the same regulatory region, therefore skewing the results, the number of regions containing insertions was also counted. Insertions were identified in 1,483 regulatory features, of which 14 (0.94%) were promoter-associated and 971 (65.5%) were gene-associated. Insertions associated with candidate cancer genes were identified in 343 regulatory features, of which 5 (1.5%) were promoter-associated and 160 (46.6%) were gene-associated. Once again, gene-associated elements were under-represented among insertions associated with candidate cancer genes (P=6.00x10^{-17}). Counting the number of regulatory features of each type that contained insertions

revealed an under-representation of features associated with H3K36me3 (P=2.04x10^{-3}), H3K4me3 (P=3.38x10^{-15}), H3K79me3 (P=1.65x10^{-4}) and DNaseI hypersensitivity sites (P=0.012). All significance tests were performed using the Chi-squared test for independence. Interestingly, DNaseI hypersensitivity and all of the histone modifications stated above are known to be associated with active genes, while those histone modifications that showed no significant difference (H3K27me3, H3K9me3 and H4K20me3) are associated with gene repression (Barski *et al.*, 2007). H3K4me3 and H3K27me3 have also been shown to be associated with active genes and silent genes, respectively, in human T cells (Roh *et al.*, 2006), which is of particular relevance to this MuLV dataset of lymphomas. Insertions that are not associated with candidate cancer genes are less likely to be oncogenic, and their over-representation in regulatory features associated with active genes may reflect the preference of MuLV for inserting within active genes. Since none of the regulatory features are over-represented among candidate genes, it appears that disruption of regulatory features may not be a common mechanism of mutagenesis of the MuLV retrovirus.

3.4.5 Expression analysis of MuLV-induced tumours

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

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

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

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

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

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

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

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

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

3.5.1 Genotype-specific cancer genes

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

a	*b*
c	*d*

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

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

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

A

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

B

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

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

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

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

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

Insertions affecting the gene encoding zinc finger E-box-binding homeobox 2 (Zeb2 or Sip1) were also significantly associated with the $Cdkn2a^{-/-}$ genotype. *SIP1* plays a role in replicative senescence, which controls the number of cell divisions in human somatic

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

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

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

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

3.5.2 Co-occurrence and mutual exclusivity of disrupted genes

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

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

a	*b*
c	*d*

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

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

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

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

A

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

B

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

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

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

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

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

There were also numerous genes for which the number of co-occurrences was lower than expected. The lack of co-operation between *Myc* and *Mycn* reflects the fact that they are structurally and functionally related. Co-occurring insertions disrupting *Myc* and either *Ikaros* or *Notch1* were also under-represented. *Myc* is a transcriptional target of the Notch signalling pathway in T cell acute lymphoblastic leukaemia, and Notch1 is

required to sustain the high levels of Myc that are required for continued growth and survival of the cancer (Sharma *et al.*, 2007). The mutual exclusivity of activated *Notch1* and *Myc* in mouse tumours suggests that *Notch1* activation may not provide a significant growth advantage when high levels of Myc are sustained by constitutive overexpression. *Mycn* and *Notch1* were also mutually exclusive, suggesting that Notch1 may play a similar role in the maintenance of *Mycn* expression during tumourigenesis. Co-occurring insertions that disrupt *Gfi1* and either *Rras2* or *Rasgrp1* were also under-represented. *Rasgrp1* and Ras-related *Rras2* were significantly associated, suggesting that Rasgrp1 activates *Rras2* and that overexpression of both genes contributes to tumourigenesis. The mutual exclusivity of disrupted *Gfi1* with both of these activated genes suggests that they may act in a common cancer pathway.

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

3.6 Discussion

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

insertional mutagenesis and regions of copy number change in human acute lymphoblastic leukaemias, therefore providing a justification for the cross-species comparative analyses performed in Chapters 4 and 5.

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

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

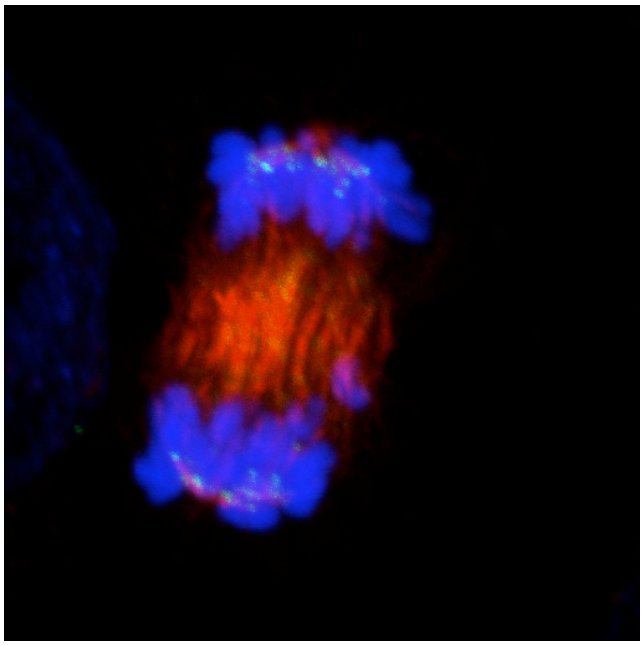

Figure 3.8. Knockdown of *QSK* in human HeLa cells is associated with increased chromosome lagging at anaphase. Figure shows a single cell at anaphase, with chromosomes stained blue and spindle fibres stained red. Image provided by Fanni Gergely at the Cambridge Research Institute.

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

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

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