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

2.1 Introduction

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

framework (de Ridder *et al.*, 2006), and CIS genes identified by the two approaches were compared. Data from the retroviral screen forms the principal mouse dataset used in this thesis, and is therefore discussed in greater detail than data from the transposon screen. The main steps involved in identifying candidate cancer genes from retroviral sequence reads are summarised in Figure 2.1. Unless otherwise stated, *P*-values provided in this chapter were generated using the Chi-squared test for independence.

2.2 Description of the datasets

2.2.1 The retroviral dataset

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

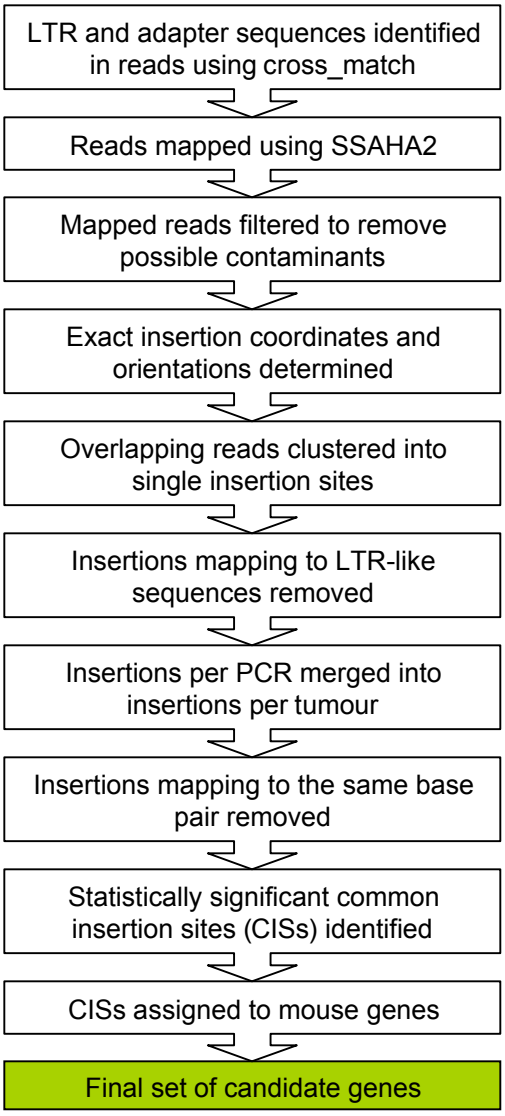

Figure 2.1. Workflow for identifying mouse candidate cancer genes from sequencing reads generated in a retroviral insertional mutagenesis screen.

A

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

B

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

C

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

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

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

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

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

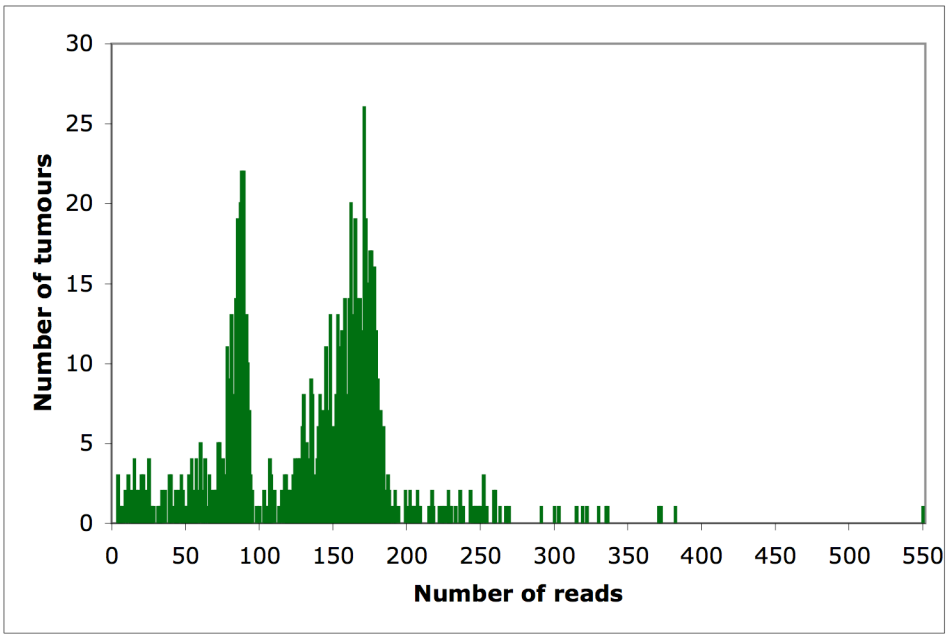

Figure 2.2. The number of sequence reads per tumour before mapping. Up to 96 reads were sequenced for each PCR. The bimodal distribution reflects the fact that 62% of tumours were subjected to 2 linker-mediated PCRs, while for 33% and 5% of tumours, 1 PCR or more than 2 PCRs, respectively, were performed.

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

2.2.2 The Sleeping Beauty dataset

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

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

The presence of the transposon IR/DR provides evidence that the genomic DNA is directly flanking an insertion. Using cross_match, IR/DR elements were identified and masked in 15,630 reads (94.2% of the total), and the rest were discarded. The linker, which was identified in 12,209 (78.1%) of the remaining reads, and extra vector sequence from the PROMEGA pGEM-T easy vector T7 promoter-multiple cloning site-SP6 promoter were also screened out with cross_match. 3,716 reads (23.8%) contained fewer

than 25 bp of unmasked sequence after screening, and as these would be too short for mapping, they were removed from the dataset. Tumour details were not available for a further 1,123 reads, and so these were also removed. The final dataset comprised 10,791 reads generated from 138 PCRs. This included 60 tumours for which genomic DNA flanking both sides had been amplified and sequenced, and 11 tumours for which only one side had been amplified. For the remaining 2 tumours, both sides had been amplified, and PCR had been performed twice on one or both sides.

2.2.3 Known cancer genes in the Cancer Gene Census

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

2.3 Mapping the sequence reads using SSAHA2

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

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

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

A

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

B

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

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

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

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

2.4 Accounting for unmapped reads

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

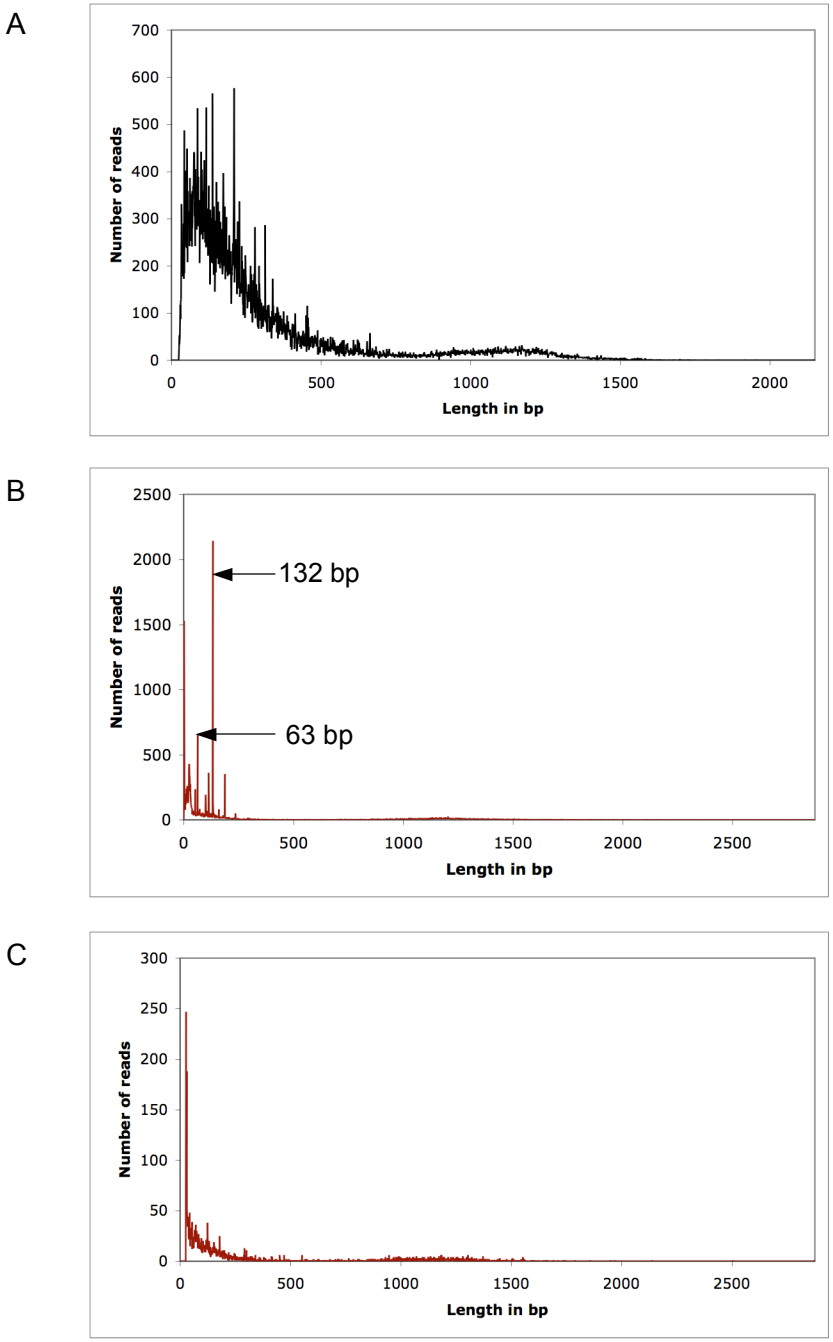

Figure 2.3. The lengths of retroviral reads that are unambiguously mapped (A), unmapped (B), and unmapped and uncharacterised (C). The reads of length 63 bp and 132 bp, which underwent further investigation, are shown.

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

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

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

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

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

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

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

A

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

B

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

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

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

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

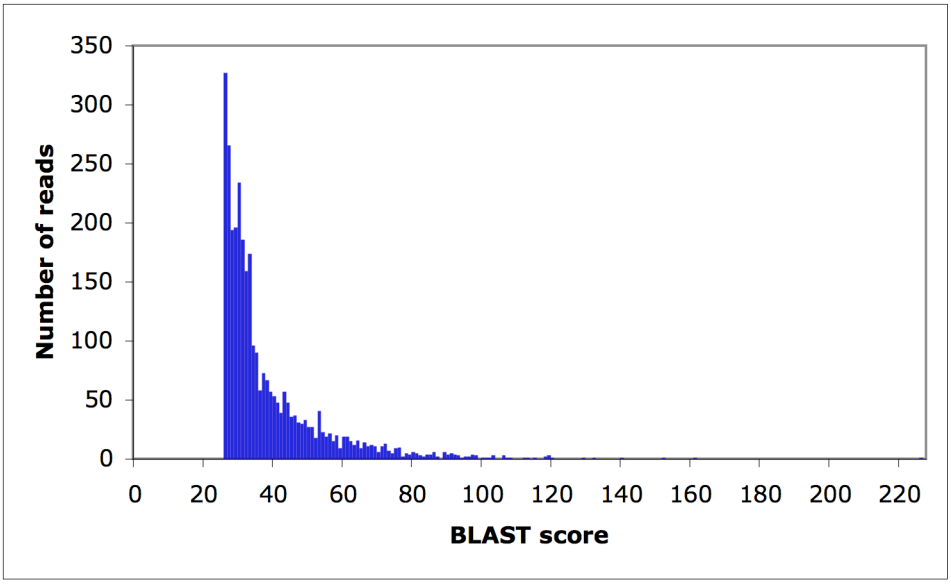

Figure 2.4. BLAST scores for uncharacterised unmapped reads. The majority of sequences that do not map with SSAHA2 but map with BLASTN have a low BLAST score.

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

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

2.5 Filtering the mapped reads

During PCR amplification, unrelated sequences can hybridise to one another, resulting in clones comprising chimeric sequences. It is important that retroviral reads contain the LTR sequence since, if the part of the read that maps to the genome is directly adjacent to the LTR, the location of the mapped DNA is likely to be the true location of the retroviral insertion. For reads that contain an LTR and an adapter, these sequences should directly flank the genomic DNA. Therefore, for each read, the coordinates of the LTR and adapter sequences identified by cross_match were compared to the coordinates of the region that mapped to the mouse genome using SSAHA2. If the gap between these regions was within 5 bp, the read was accepted. Since the junction between the LTR and the genomic DNA is most important, reads were also accepted if the DNA that mapped to

the genome was within 5 bp of the LTR but there was a gap between the genomic DNA and the adapter, or if the read did not contain an adapter sequence. Base miscalling in low quality reads may result in a SSAHA2 alignment that does not extend right up to the LTR sequence even though the LTR and genomic DNA are directly adjacent. Therefore, up to a distance of 30 bp, reads were accepted if the sequence between the LTR and the aligning genomic DNA did not contain any restriction sites for *Tsp509I* (i.e. 5'-AATT-3') or *Sau3AI* (i.e. 5'-GATC-3'), depending on which had been used in the PCR. If a restriction site intercepts the LTR and genomic DNA, it is possible that the genomic DNA that immediately flanks the LTR, and represents the true location of the virus in the genome, may not have been mapped because it is too small or of poor quality but that it has ligated to a contaminating DNA fragment that has been mapped.

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

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

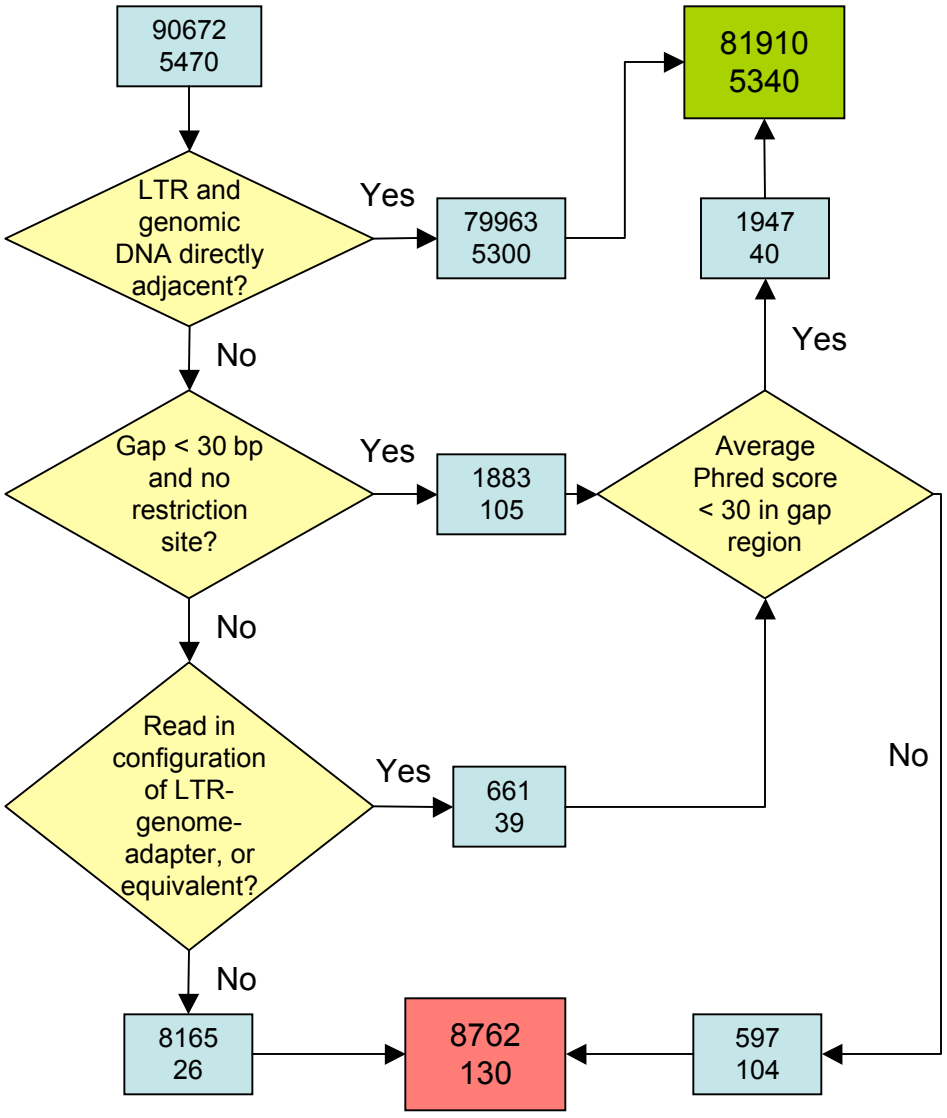

Figure 2.5. The filtering process for single mapping reads. Blue boxes contain the counts for accepted or rejected reads at each stage, where the top number in each box refers to the count for retroviral reads and the bottom number refers to the count for transposon reads. Final counts for the accepted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

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

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

2.6 Identification and filtering of insertion sites

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

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

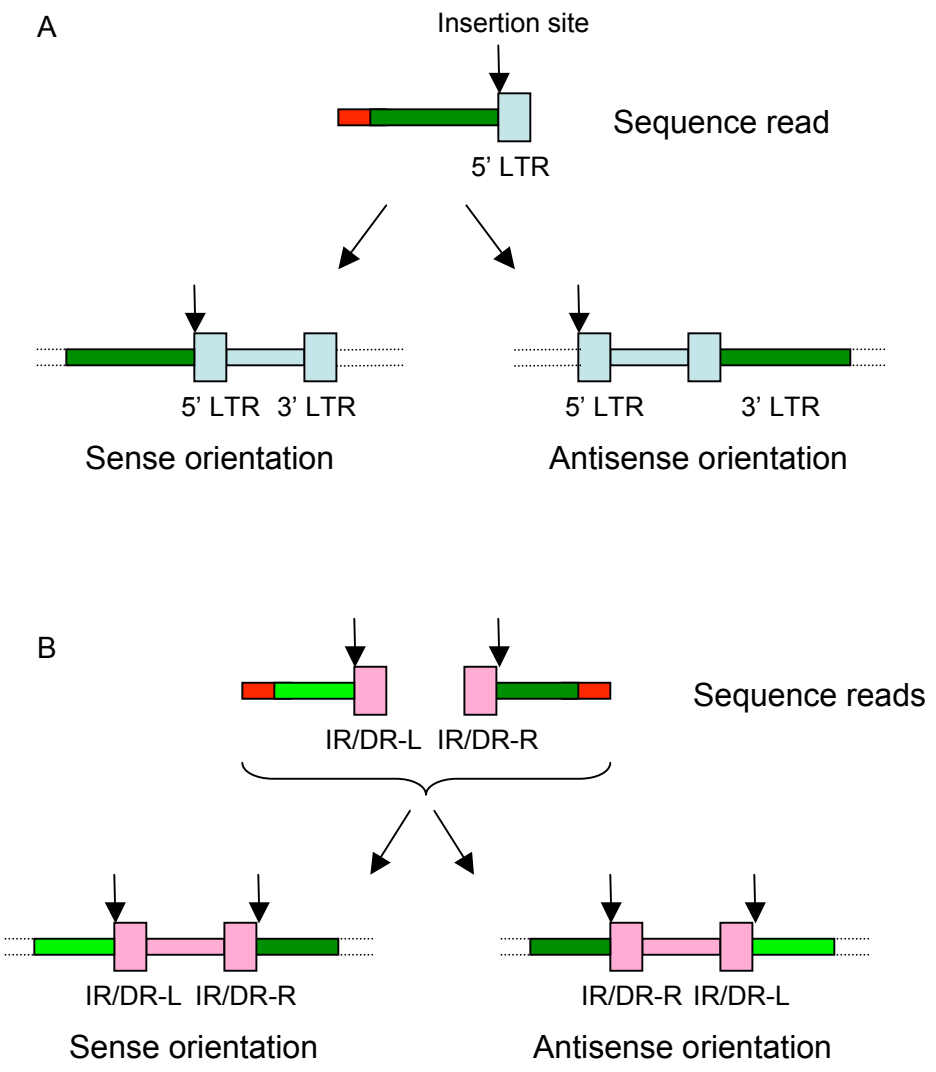

Figure 2.6. Determining the exact insertion site and orientation of retroviral (A) and transposon (B) insertions in the mouse genome. Adapter sequences are shown in red; genomic DNA is shown in green. **A.** The point of insertion is the genomic nucleotide adjacent to the 5' LTR of the MuLV retrovirus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

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

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

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

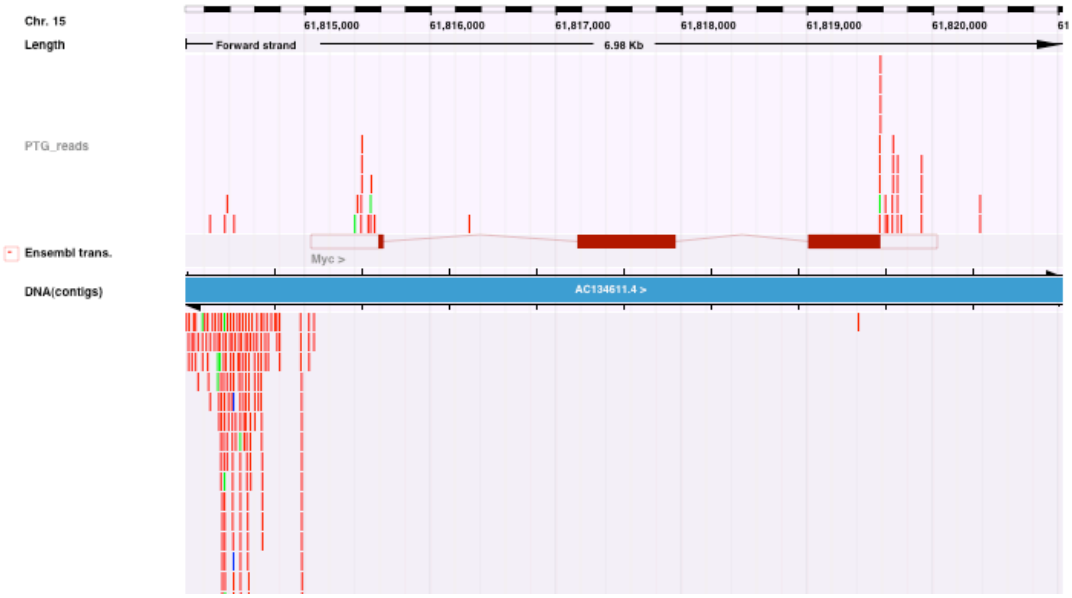

Figure 2.8. A high proportion of insertions in control samples map to the *Myc* gene. This figure shows some of the insertions in and around the *Myc* gene. Insertions from tumours, short infection time samples and uninfected samples are shown as red, green and blue rectangles, respectively.

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

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

2.7 Estimating the coverage of the mutagenesis screens

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

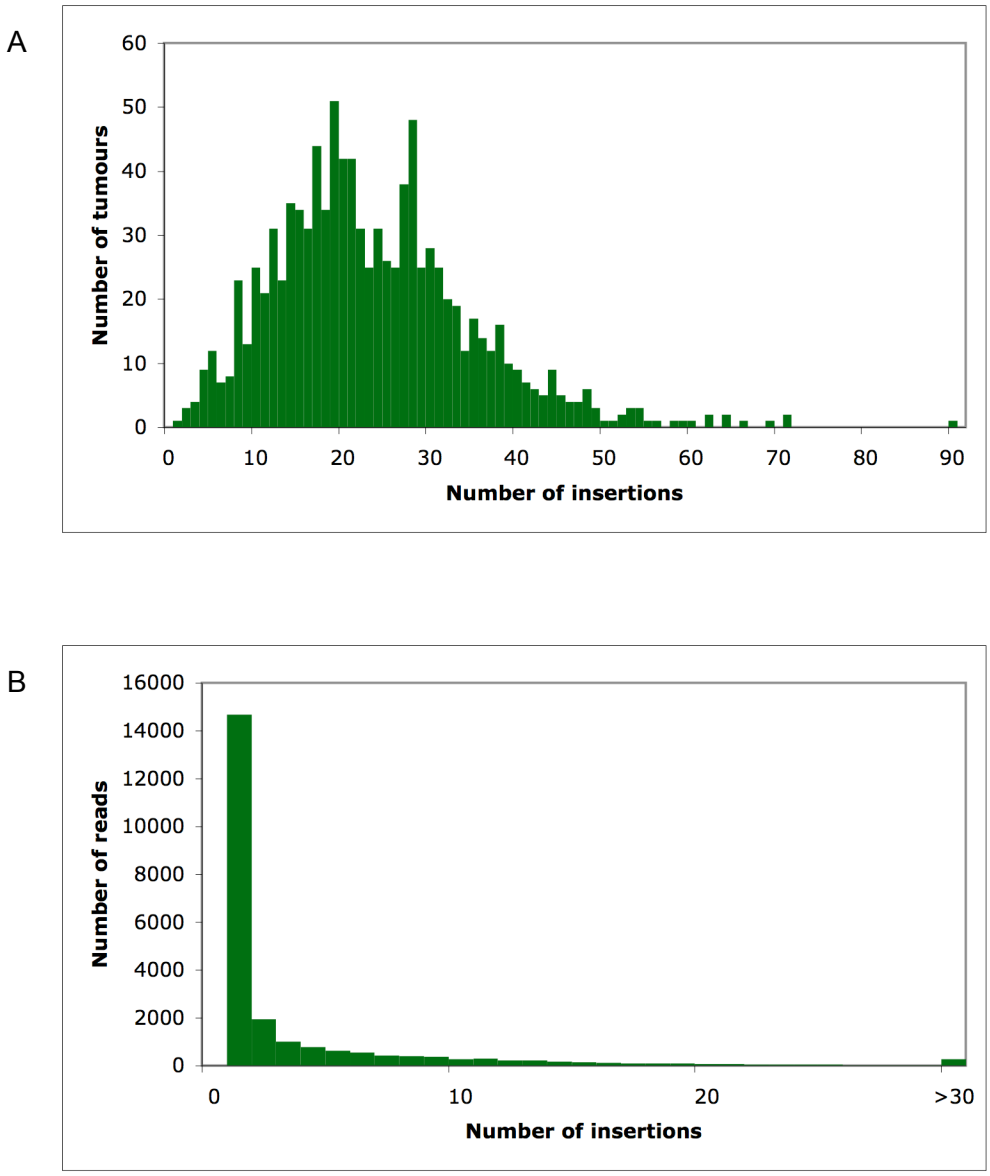

Figure 2.9. The number of insertions per tumour (A) and reads per insertion (B).

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

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

2.8 Analysis of the distribution of insertions around mouse genes

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

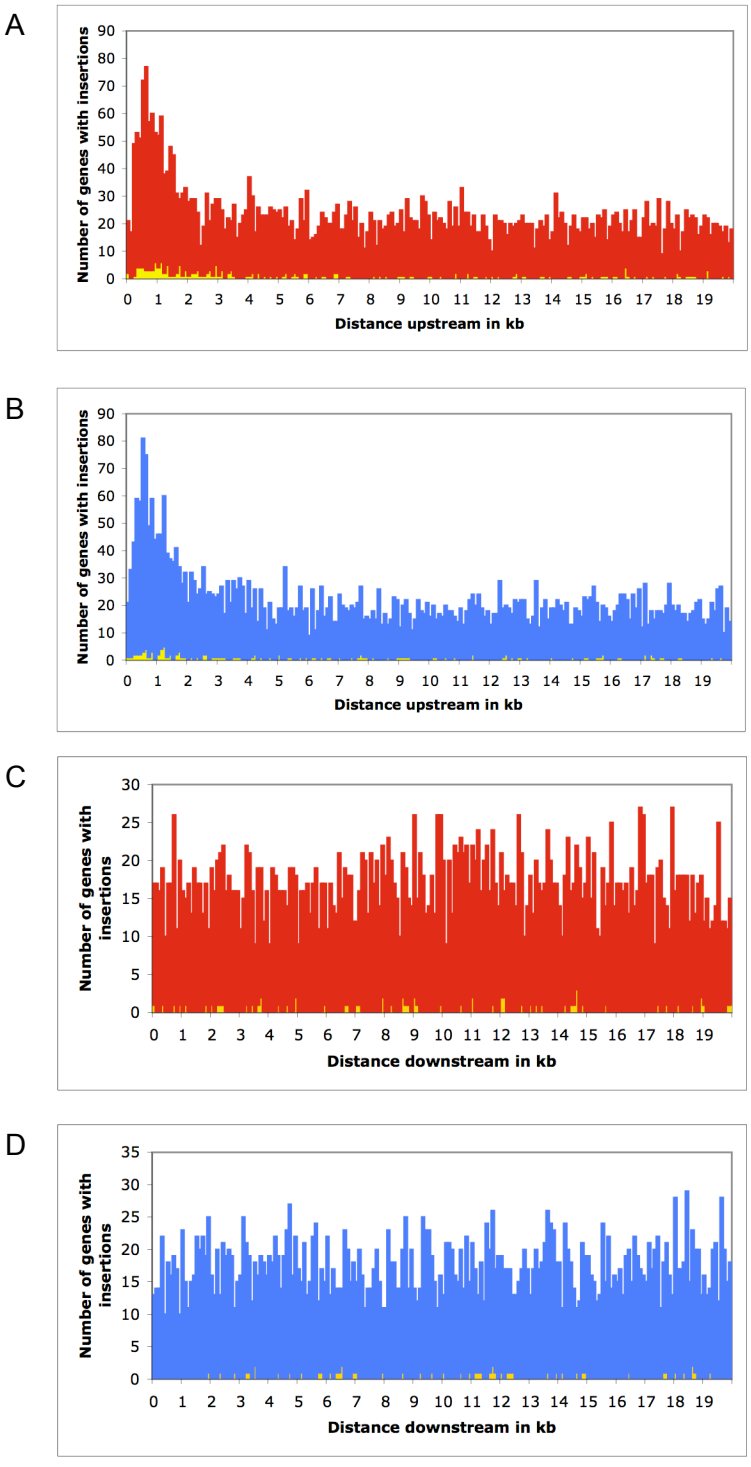

Figure 2.10. The number of genes with insertions in 100 bp intervals up to 20 kb upstream in the sense (A) and antisense (B) orientation and downstream in the sense (C) and antisense (D) orientation with respect to the gene. Counts of cancer genes with insertions in each interval are shown in yellow.

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

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

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

A

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

B

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

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

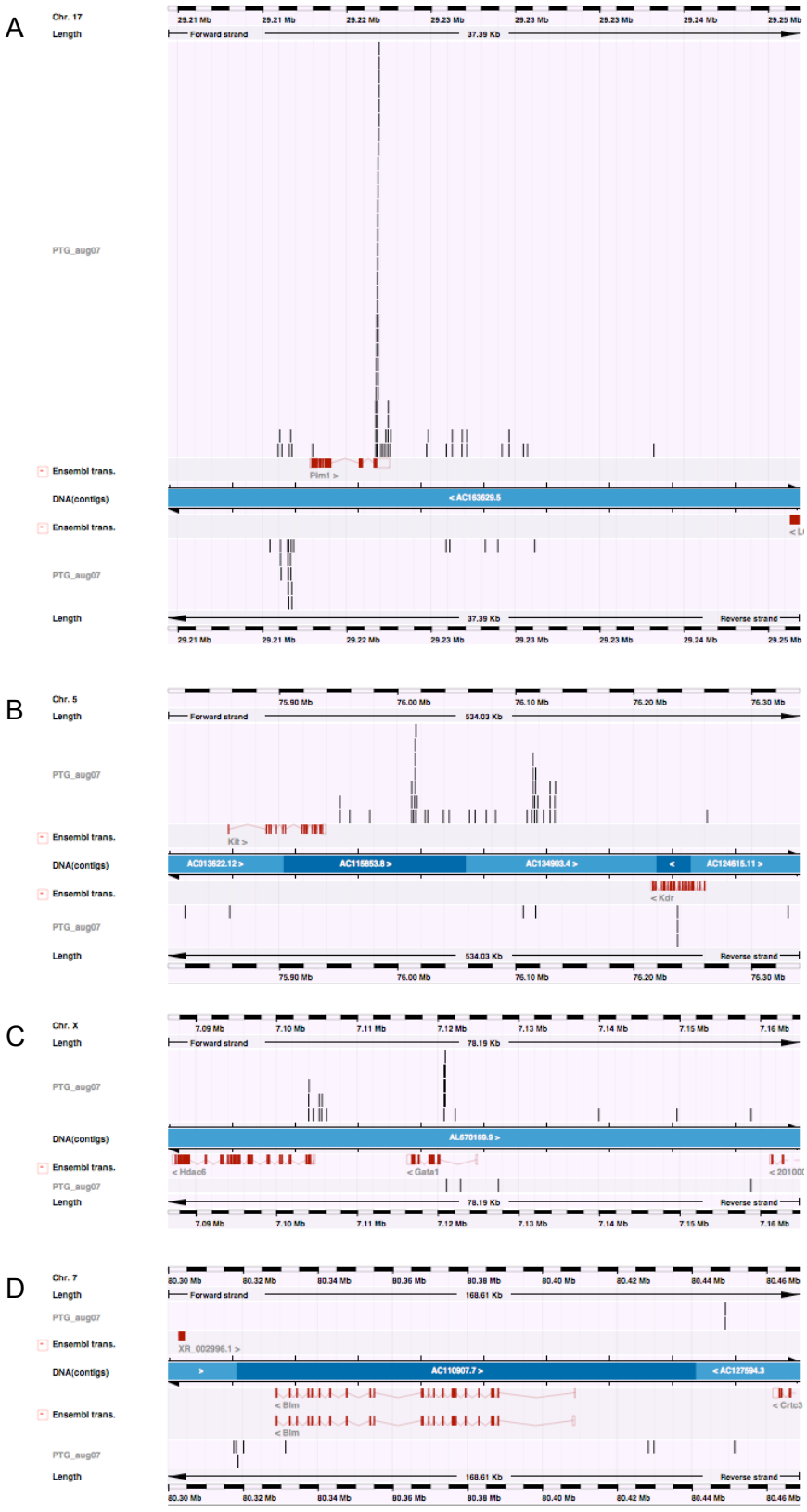

Figure 2.11. Insertions around known cancer genes *Pim1* (A), *Kit* (B), *Gata1* (C) and *Blm* (D). Insertions are shown as black bars in the context of Ensembl genes, shown in red. Insertions above and below the blue bar are in the sense and antisense orientation, respectively.

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

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

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

2.9 Assigning insertions to genes

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

2.10 Identifying statistically significant common insertion sites

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

2.10.1 Monte Carlo simulations

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

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

A

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

B

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

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

A

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

B

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

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

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

2.10.2 Kernel convolution

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

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

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

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

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

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

TP is the number of cancer genes in the candidate cancer gene list (true positives), FP is the number of non-cancer genes in the list (false positives), TN is the number of non-cancer genes not in the list (true negatives), and FN is the number of cancer genes not in the list (false negatives). Genes that were not in the list were calculated as all 18,017 mouse genes with human orthologues, as identified in Ensembl version 48, minus those in the list. MCC is used in machine learning as a measure of the quality of a prediction and it takes into account the counts of true and false positives and negatives to generate a single number that can be compared across predictions. The candidate cancer gene lists

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

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

Of the 39 cross-scale *Sleeping Beauty* CISs identified by the KC method, 36 were also present in the *Efr*=0.01 MC list. The remaining 3 were a long way from the nearest gene,

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

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

2.10.3 Final set of candidate genes

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

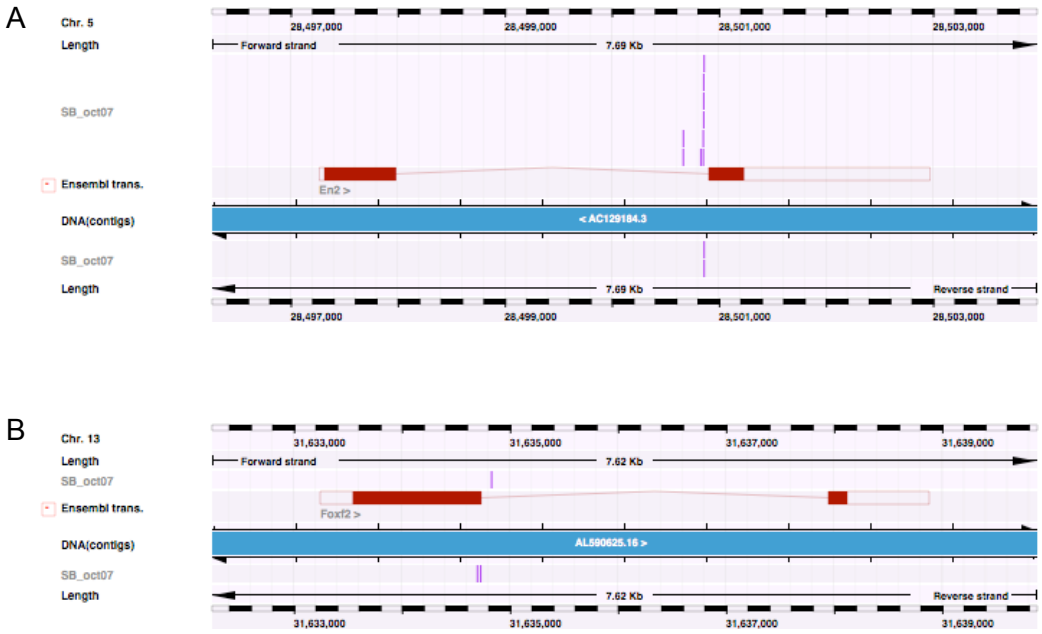

Figure 2.12. Insertions in *En2* (A) and *Foxf2* (B) are located at the splice junctions used to construct the T2/Onc transposon and are contaminating sequences. Insertions are shown as pink lines in the context of the Ensembl gene, shown in red. Insertions above and below the blue line are in the sense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

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

2.11 Discussion

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

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

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

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

Two approaches, Monte Carlo simulations (Suzuki *et al.*, 2002) and a kernel convolution-based statistical framework (de Ridder *et al.*, 2006), have been used to identify statistically significant CISs in the retroviral and transposon screens. Known cancer genes can be used as a partial set of true positives to evaluate candidate cancer genes in

the vicinity of CISs, and Matthew's Correlation Coefficient was used to show that the kernel convolution-based framework gives the most reliable set of candidate cancer genes. The final set of candidates generated from the *Sleeping Beauty* screen comprises just 10 genes, reflecting the small size of the initial dataset and problems in mapping the reads. 439 candidates were identified from the MuLV screen, thereby supporting the theory that many genes contribute to tumourigenesis. The candidate cancer genes are analysed and characterised in Chapter 3.