Chapter 6 Summary and Conclusions

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

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

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

compared with the traditional Sanger chain-termination methods, can sequence larger quantities of DNA at lower cost, thereby allowing for an increase in both the number of PCRs and the sequencing depth.

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

The main aim of Chapter 3 was to characterise the mouse candidate cancer genes identified in Chapter 2, both as a complete list and as individual candidates. A significant overlap between the mouse candidates and human cancer-associated genes within the Cancer Gene Census (Futreal *et al.*, 2004) and COSMIC database (Forbes *et al.*, 2006) demonstrated the relevance of insertional mutagenesis to human tumourigenesis. This is important on two counts, since it shows that genes disrupted by MuLV mutagenesis contribute to spontaneous cancer development, and that mouse cancer gene candidates are involved in human cancer. Interestingly, candidates were also over-represented among genes with Nanog and Oct4 binding sites, suggesting that a significant proportion may be involved in tumour cell self-renewal, which is consistent with the cancer stem cell

hypothesis discussed in Section 1.2.3.2. The overlap with regions of copy number change in human paediatric acute lymphoblastic leukaemias (Mullighan *et al.*, 2007) provided the first indication that a significant proportion of the mouse candidates may be amplified or deleted in human cancer, and may help to narrow down the candidates in regions of copy number change in human cancers. The analysis also identified miRNA genes among the mouse candidates, and uncovered an over-representation of targets for 3 miRNAs (mmu-miR-449b, mmu-miR-449c and hsa-miR-565) that have not been previously implicated in tumourigenesis.

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

The chapter also focuses on the analysis of the distribution of insertions in and around cancer genes, and this was used to predict the likely mechanism of mutation of candidate genes. For example, *Ccr7* and *Jundm2* were predicted to be mutated by the same mechanism as *Mycn* and *Pim1*, wherein insertions are known to cause premature termination of gene transcription that results in the removal of mRNA-destabilising motifs and, therefore, a more stable gene transcript. This analysis is complicated by the number of ways in which the mutagen can disrupt a gene, and would be aided by using a transposon with a more limited repertoire of mutational mechanisms, e.g. the ability to

induce C-terminal truncations in one orientation only. For genes that are mutated by both MuLV and SB, determination of the mechanism of mutation is facilitated by the co-analysis of MuLV and SB insertions. For example, *Notch1* contains both MuLV and SB insertions in 3 distinct locations of the gene that represent different types of oncogenic mutation, and the resulting gene products have been observed in human cancers. Elucidation of the mechanism of mutation may provide an insight into the role of a gene in cancer, and may facilitate the development of therapies targeted against specific mutants. The distribution of insertions can also help to distinguish oncogenes, in which insertions often form distinct clusters, from tumour suppressor genes, in which insertions are more likely to be scattered throughout the gene and may be more likely to include multiple insertions from the same tumour, potentially representing inactivation of both gene copies. As well as a number of known and implicated tumour suppressor genes, this analysis identified *Qsk*, *Smg6* and *Foxp1* as potential tumour suppressor genes. Interestingly, compared with random insertions, those associated with candidate cancer genes were under-represented in a predicted set of gene-associated regulatory regions and, specifically, within predicted regulatory features associated with active genes. This supports the notion that these insertions are oncogenic and do not simply result from a bias towards insertion into the 5' ends of transcriptionally active genes. It also shows that disruption of regulatory regions is not a common mechanism of mutagenesis for MuLV.

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

in elucidating cancer pathways. Understanding the pathway in which a gene acts may help in the development of cancer therapies since, in some cases, it may be more effective to target a gene that acts downstream of the mutated gene, rather than targeting the mutated gene itself.

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

The CGH data from both the 10K and SNP 6.0 SNP arrays were processed into regions of copy number change before being compared to the mouse dataset. Most of the available

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

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

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

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

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

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