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

5.1 Introduction

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

5.2 Description and processing of the datasets

5.2.1 High-resolution copy number data

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

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

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

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

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

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

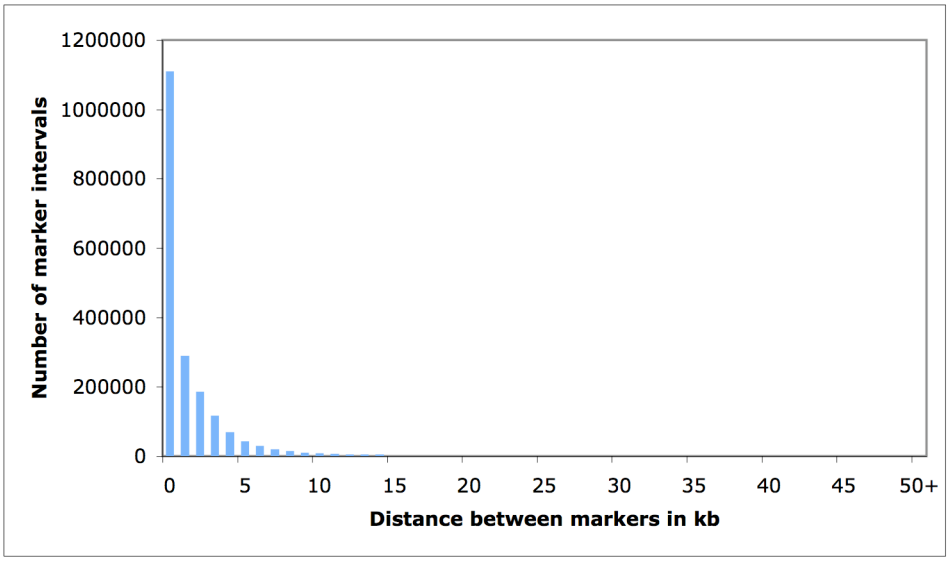

Figure 5.1. Distance between adjacent copy number probes across the human genome.

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

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

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

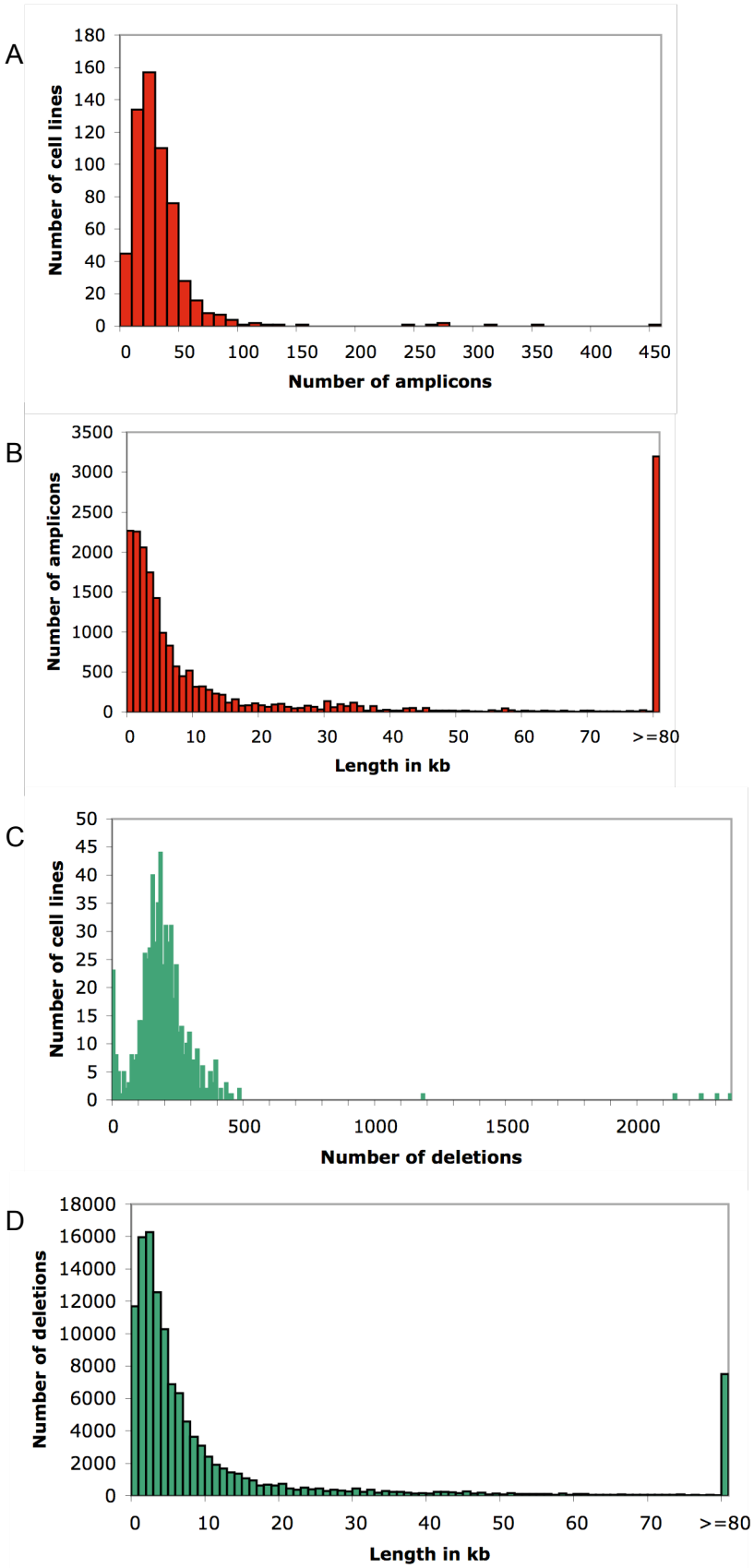

Figure 5.2. Characterisation of amplicons and deletions in 598 human cancer cell lines analysed using high-resolution SNP array CGH. (A) Number of amplicons per cell line. (B) Length of amplicons. (C) Number of deletions per cell line. (D) Length of deletions.

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

5.2.2 Additional datasets

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

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

5.3.1 Global comparison

The number of CIS genes with human orthologues was counted within amplicons above copy number thresholds ranging from 1.1 to 5.0 with increments of 0.1, and within deletions below copy number thresholds ranging from 0.9 to 0.1 with increments of 0.1. The number of non-CIS genes with human orthologues in amplicons and deletions was also calculated, and a 2-tailed Fisher Exact Test was used to determine whether, for each copy number threshold, the number of CIS genes was significantly different to that

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

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

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

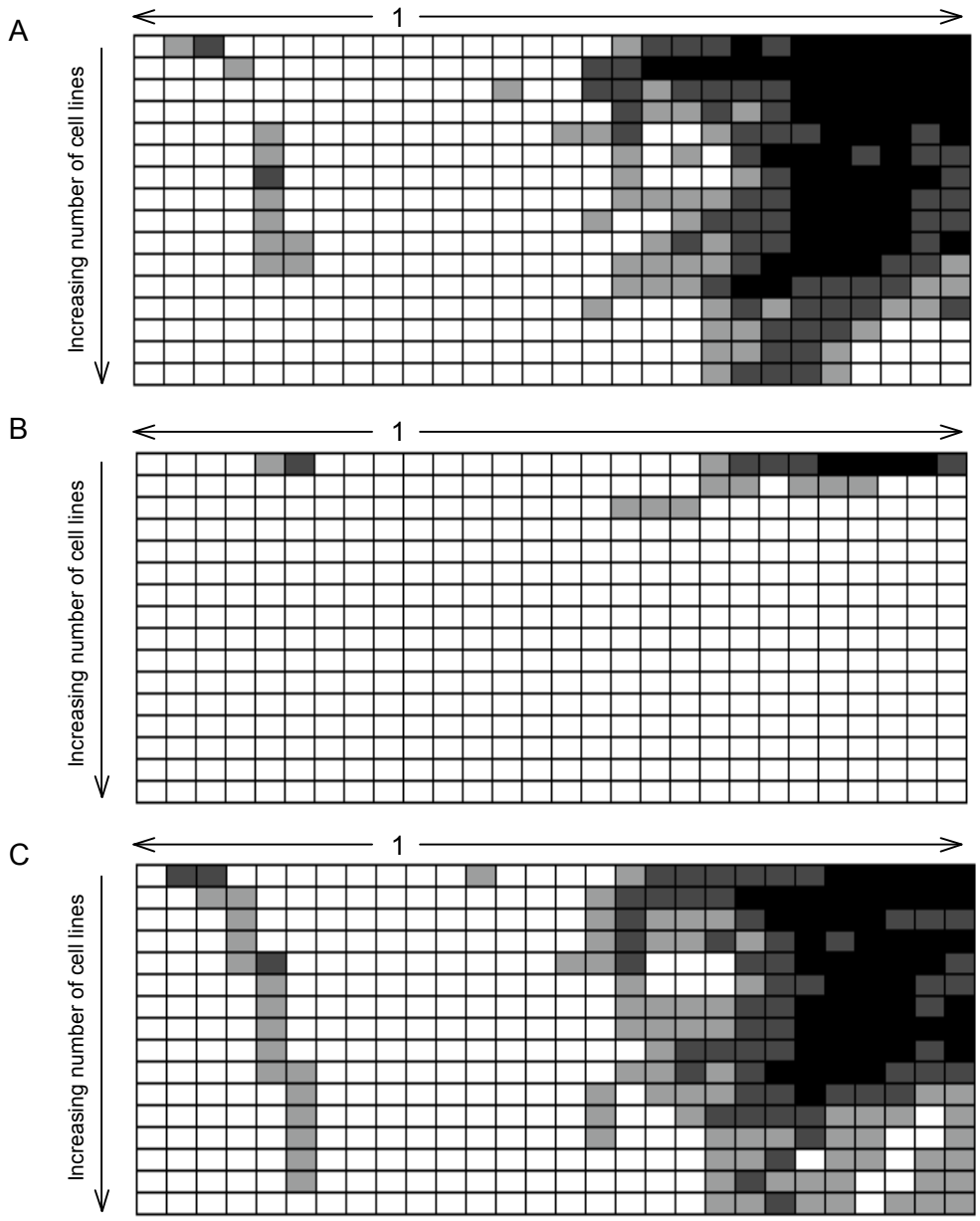

Figure 5.3. Over-representation of CIS genes in amplicons and deletions of varying copy number threshold and number of cell lines across all cell lines (A), haematopoietic and lymphoid cancer cell lines (B), and cell lines derived from solid tumours (C). Each box represents the significance of the association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

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

5.3.2 Identifying individual cancer gene candidates

5.3.2.1 Methods

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

As for the 10K data in Section 4.5.2, minimal amplified and deleted regions were identified by calculating the coordinates of the smallest overlap of regions that contained the CIS gene. Other genes within the region were identified using the coordinates of human genes in Ensembl v45. For each gene within a minimal amplified region, the total number of cell lines in which that gene was amplified was calculated, and the maximum copy number across all lines was determined. A similar procedure was applied to genes within a minimal deleted region, except that the number of deletions was calculated, and the minimum copy number was determined. The minimal amplified or deleted region within which a CIS gene resides may not necessarily be a minimal amplified or deleted region from across the entire genome. The coordinates of minimal amplified and deleted

regions across the genome were therefore determined and were compared to the coordinates of CIS genes to determine whether the CIS genes resided in these regions. To avoid confusion, these regions are known as MCRs (minimal common regions). 8,694 MCRs were identified among amplicons, while 35,213 were identified among deletions of copy number less than or equal to 0.6.

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

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

5.3.2.2 Candidate cancer genes within amplicons

There were 9,681 mouse genes with human orthologues within the amplicons of human cancer cell lines. Of these, 232 were CIS genes, which is more than expected by chance (P=0.00447). 27 CIS genes were found in statistically significant recurrent amplicons (P<0.05). This included a significant over-representation of genes that were designated dominant cancer genes in the Cancer Gene Census, compared with CIS genes that were not recurrently amplified (P=2.85x10^{-4}). 2,730 mouse genes with human orthologues were found within amplicons in haematopoietic and lymphoid cancer cell lines. CIS genes were over-represented among these genes (71 genes, P=0.0408) but, surprisingly, not as significantly as those amplified in all tumours. This further demonstrates that genes identified by insertional mutagenesis in the mouse may be relevant to the identification of cancer genes in a range of human tumours, not just in those originating in lymphoid tissue. This is at odds with the findings in Section 3.2.2, where the genes

showed no association with candidate breast and colon cancer genes from Sjöblom *et al.* (2006). However, the breast and colon candidates were identified by exon resequencing and therefore represent genes mutated by point mutations and indels in cancer. The most common effects of insertional mutagenesis, i.e. gene upregulation by promoter or enhancer insertion, more closely resemble those resulting from changes in copy number. It is therefore more likely that the human orthologues of mouse oncogenes identified by insertional mutagenesis will be disrupted by copy number changes than by small intragenic substitutions and indels. Interestingly, there was no over-representation of genes with mutations in COSMIC among CIS genes that were recurrently amplified, suggesting that copy number changes and small intragenic mutations are not positively associated within candidate cancer genes. All genes in recurrent amplicons across all cell lines, and specifically in haematopoietic and lymphoid cell lines, with a *P*-value of less than 0.1 are shown in Table 5.3.

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

A

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

B

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

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

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

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

Of the 16 genes that were identified as the only gene in the minimal amplified region, WW domain-containing oxidoreductase (*WWOX*), Aiolos (*IKZF3*) and CUG triplet repeat, RNA binding protein 2 (*CUGBP2*) might be considered more likely to play a role in tumour suppression. Their presence within amplicons therefore suggests that they reside in unstable regions of the genome. It is possible that duplication within the gene may be a mechanism for disrupting the gene, or that the copy number values across a group of markers are erroneous, such that the gene appears to be the target of amplification when in fact it is not. However, as mentioned in Section 3.4.3, *IKZF3* can both activate and repress lineage-specific cells in lymphocytes and may therefore play a dual role in tumourigenesis, and it has been found to be upregulated in chronic lymphocytic leukaemia (Duhamel *et al.*, 2008). Overexpression of *CUGBP2* induces

apoptosis of colon cancer cells exposed to radiation (Natarajan *et al.*, 2008) by inhibiting expression of *MCL1*, which is described above (Subramaniam *et al.*, 2008). All of the amplicons in the *CUGBP2* gene were localised to exon 2, suggesting either an error with the 3 markers in this region, or that duplication of this exon may lead to gene disruption (Figure 5.4A). Likewise, the minimal amplified region of *IKZF3* was focused on exon 3, but many amplicons also spanned the entire gene (Figure 5.4B). The presence of amplified *WWOX* may reflect the fact that it resides in the fragile site FRA16D, which is prone to rearrangement in human cancer (see Section 4.5.2.3). Almost all of the amplicons were clustered into 2 distinct regions, one of which contained amplicons spanning up to 19 markers and the other contained amplicons spanning 2 or 3 markers. Both of these regions were within introns, suggesting that they may not affect gene expression. In addition, recurrent amplification of *WWOX* was observed across all cell types, and was significant in 15 of the 19 cancer types for which there were more than 10 cell lines, which suggests that it may reflect some sort of global effect rather than tumourigenicity. As expected for tumour suppressor genes, *Ikzf3*, *Cugbp2* and *Wwox* are all disrupted by intragenic CISs.

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

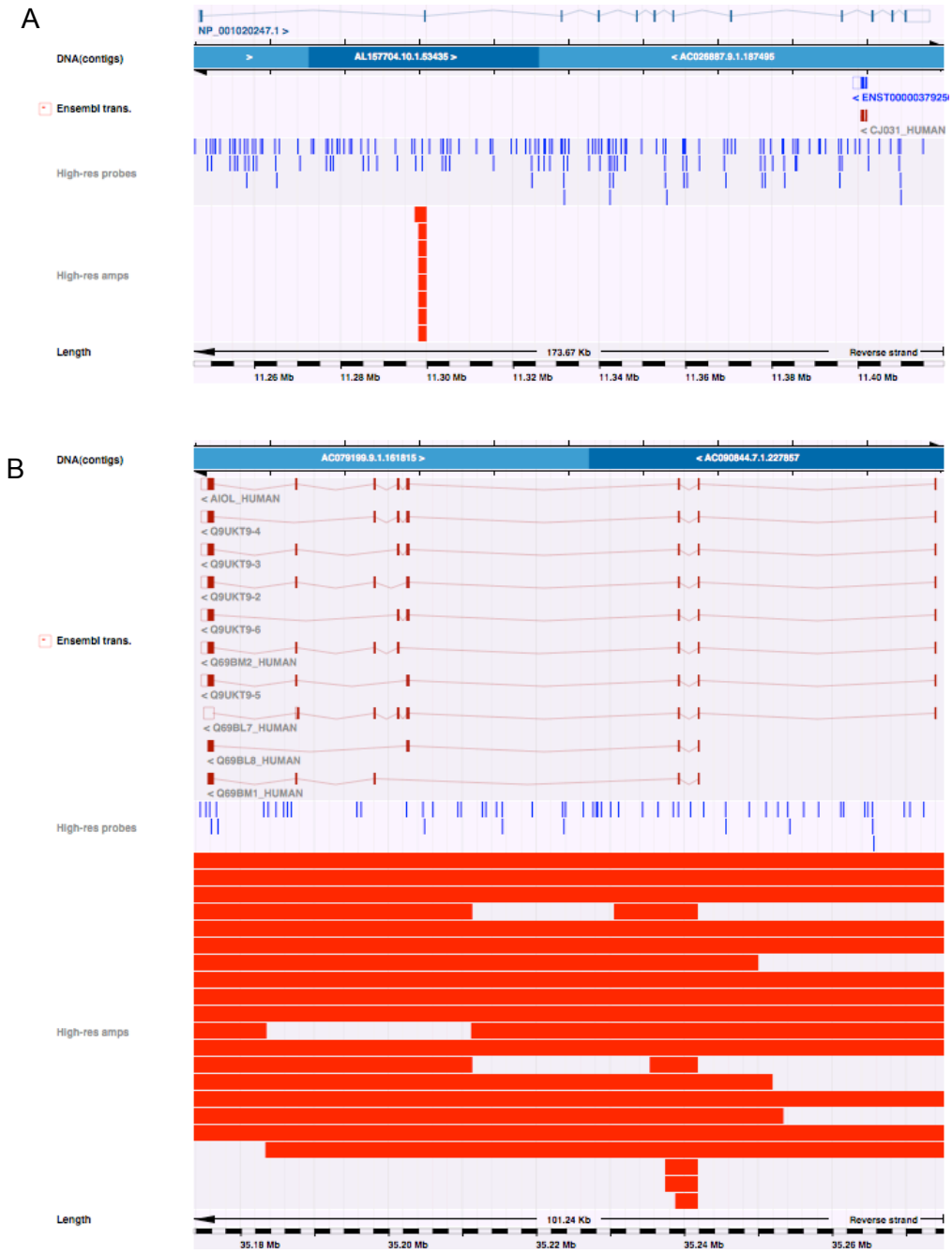

Figure 5.4. The minimal amplified regions within putative tumour suppressor genes *CUGBP2* (A) and *IKZF3* (B) are localised around specific exons. The Ensembl transcripts for *CUGBP2* and *IKZF3* are shown in blue and red, respectively. Copy number markers are shown as blue vertical lines (labelled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

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

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

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

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

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

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

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

lymphomagenesis, in the mouse might not contribute to the formation of spontaneous mouse or human lymphomas because they are not normally expressed in lymphocytes. However, in the case of *MYCN*, activation by translocation has been demonstrated in non-Hodgkin's lymphoma (Finnegan *et al.*, 1995), and *MYCN* appears to co-operate with the ETS family gene *TEL2* in B-cell lymphomagenesis (Cardone *et al.*, 2005). As mentioned previously, it is also found within a recurrent amplicon in ALL (Mullighan *et al.*, 2007).

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

insertions are within a longer Ensembl EST gene transcript that is not annotated as an Ensembl gene transcript and could therefore be intragenic, inactivating insertions (Figure 5.6, page 249). *RCBTB2* is not within an MCR, suggesting that it may not contribute to colon tumourigenesis.

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

The number of recurrent amplifications was significantly higher across CIS genes than across non-CIS genes (P=0.00396). The median number of amplifications was 1 for both samples, and the maximum number was larger for non-CIS genes, at 115 recurrent amplifications, compared with 69 for CIS genes (specifically *MYC*). However, the mean

number was 2.77 for CIS genes compared with 2.30 for non-CIS genes. The maximum peak of amplification was also higher among CIS genes (P=0.00135). The minimum and maximum copy numbers were 1.253 (*UBE1L*) and 30.230 (*TGFBR3*) for CIS genes, and 1.227 and 67.960 for non-CIS genes. The median and mean peaks in amplification were 1.832 and 2.692 for CIS genes, and 1.721 and 2.405 for non-CIS genes. These results further suggest that a significant proportion of CIS genes may contribute to human tumourigenesis through the mechanism of amplification.

5.3.2.3 Candidate cancer genes within deletions

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

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

A

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

B

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

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

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

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

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

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

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

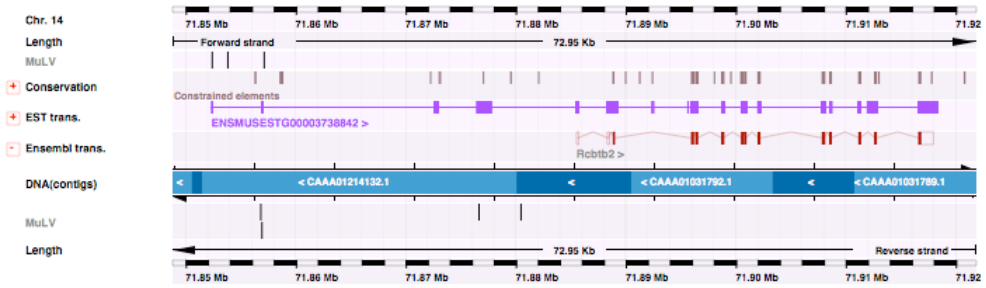

Figure 5.6. All of the MuLV insertions assigned to the *Rcbtb2* gene are within a larger, unannotated, EST transcript. Insertions are shown as black vertical lines. The Ensembl gene and EST are shown in red and purple, respectively.

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

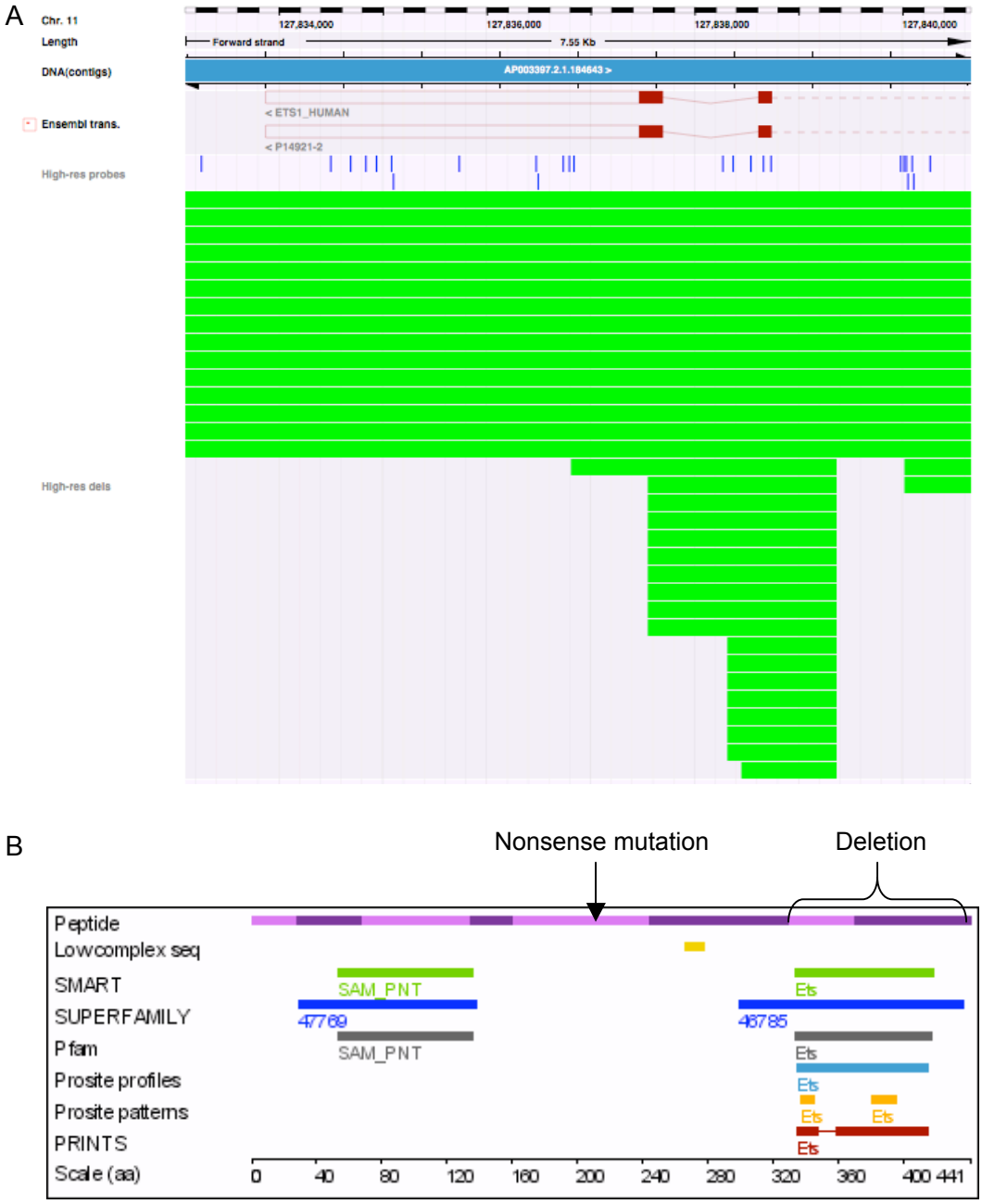

Figure 5.8. Mutations in the *ETS1* gene result in removal of the Ets domain. (A) Deletions within *ETS1* in human cancer cell lines. (B) The location of mutations in the context of the ETS1 protein. In Figure A, *ETS1* gene transcripts are shown in red, copy number markers are shown in blue and deletions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

A

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

B

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

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

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

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

A number of interesting candidates were also identified among those genes occurring in significantly recurrent deletions in specific tissue types. These included *FLI1*, *MYO18A*, *TGFBR3* and *SKI*, all of which were in MCRs across all cell lines. *FLI1* and *MYO18A* contained recurrent deletions in bone cancer cell lines (P=0.0371 for both genes). Interestingly, a translocation that leads to the formation of an EWS-FLI1 fusion protein is present in 95% of Ewing's sarcomas, which are tumours of the bone and soft tissue (Delattre *et al.*, 1994). However, *FLI1* is an oncogene, and all but three of the deletions spanning this gene also encompassed other genes (including *ETS1*). *MYO18A* has not been implicated in cancer, but has been shown to be expressed in bone marrow stromal cells, where it may play a role in the maintenance of cell architecture (Furusawa *et al.*, 2000). TGFβ receptor type III precursor *TGFBR3* is implicated as a tumour suppressor gene in a range of cancer types, including non-small cell lung cancer (Finger *et al.*, 2008), pancreatic cancer (Gordon *et al.*, 2008), prostate cancer (Turley *et al.*, 2007) and breast cancer (Dong *et al.*, 2007), but the identification of recurrent deletions in soft tissue

cancer cell lines (*P*=0.0270) suggests it is also important in these tumours. *SKI* plays an oncogenic role in some cancers, such as human melanomas (for review, see Reed *et al.*, 2005), but *Ski*-deficient heterozygous mice show an increased susceptibility to tumourigenesis (Shinagawa *et al.*, 2001) and reduced *SKI* expression in breast and lung cancer cells enhances tumour metastasis (Le Scolan *et al.*, 2008). The *P*-value for the number of deletions of *SKI* across all cell lines was 0.0538, but was lower for deletions in cancer cell lines of the kidney (*P*=0.0351), central nervous system (*P*=0.0166) and autonomic ganglia (*P*=0.0130).

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

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

5.4 Comparison between high-resolution and 10K CGH data

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

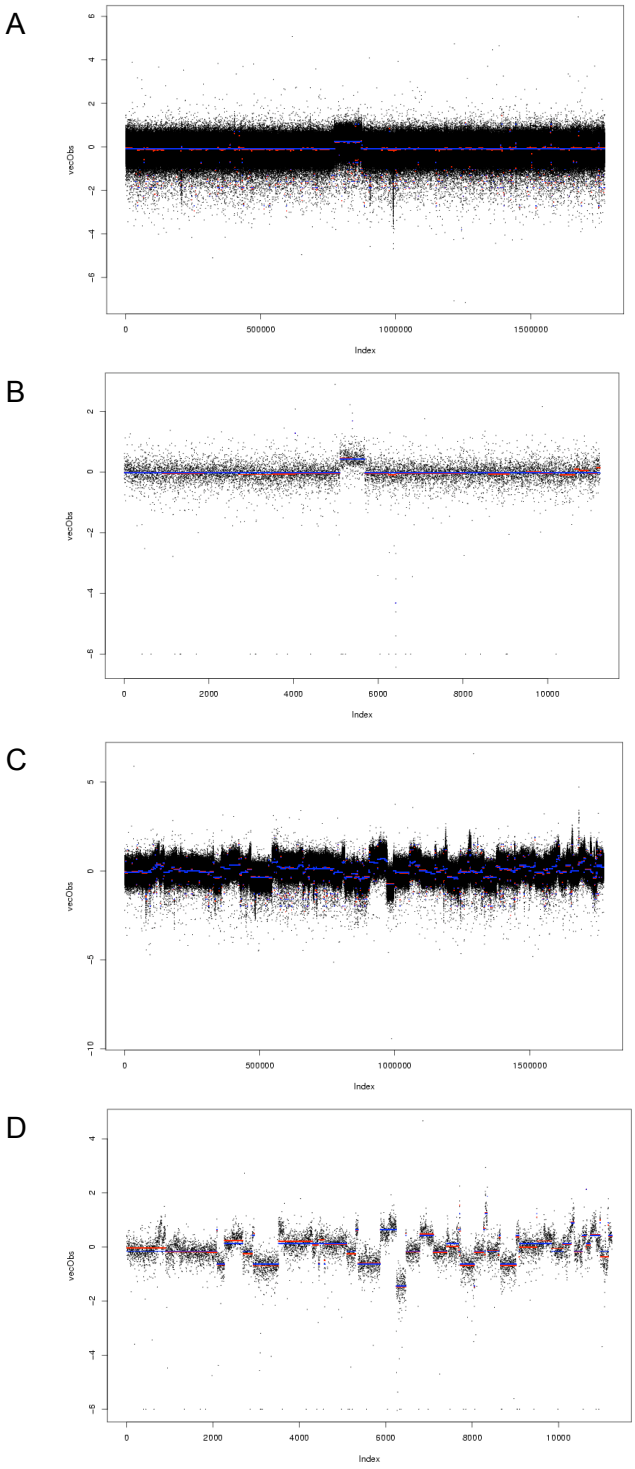

Figure 5.9. High-resolution and 10K SNP array CGH data for the entire genome of B-cell lymphoma cell line DOHH-2 (A and B, respectively) and breast cancer cell line HCC1143 (C and D, respectively). Black points are copy number values for individual SNPs, red lines are mean copy numbers for DNAcopy segments,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

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

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

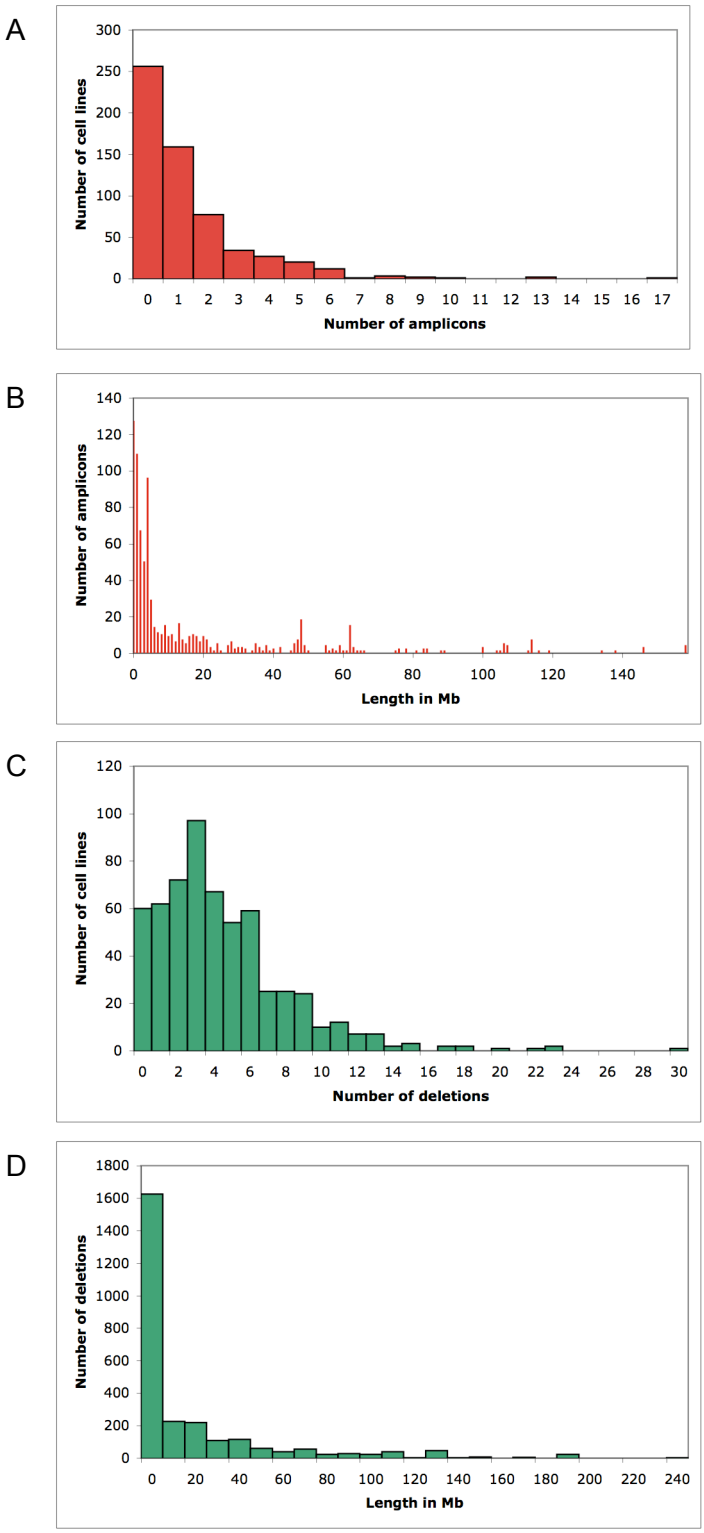

Figure 5.10. Characterisation of amplicons and deletions in 598 human cancer cell lines analysed using 10K SNP array CGH. (A) Number of amplicons per cell line. (B) Length of amplicons. (C) Number of deletions per cell line. (D) Length of deletions.

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

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

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

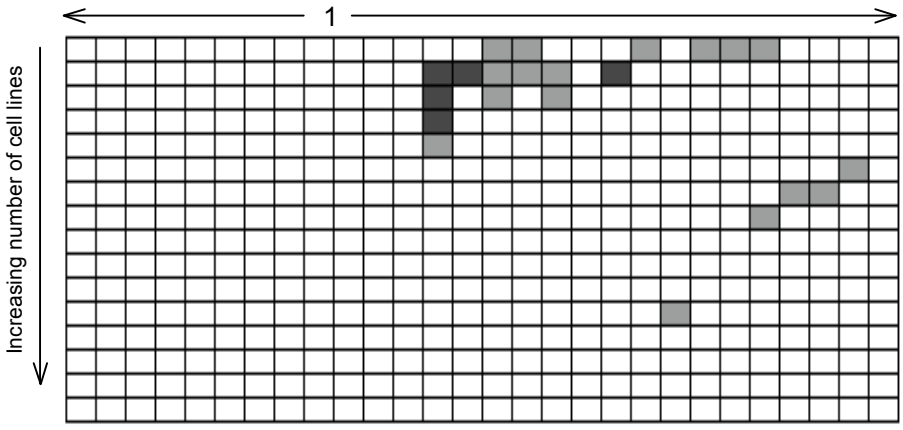

Figure 5.11. Over-representation of CIS genes in amplicons of varying copy number threshold and number of cell lines in the 10K dataset. Each box represents the significance of the association between CIS genes and amplicons/deletions at a given copy number threshold and cell line number. $P<0.0001$,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

A

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

B

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

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

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

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

A

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

B

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

C

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

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

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

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

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

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

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

often large and encompass multiple genes, the most convincing candidate oncogenes may be those that are identified using both platforms. For deletions, which are often smaller, the identification of candidate tumour suppressor genes is more reliant on the high-resolution dataset.

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

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

5.5 Identification of co-operating cancer genes

5.5.1 Genotype-specific cancer genes

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

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

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

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

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

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

Among the deleted genes that were positively associated with *TP53* mutations were AT-rich interactive domain 3A (*ARID3A*) and growth arrest and DNA-damage-inducible protein beta (*GADD45β*). *ARID3A* contains a putative binding site for p53, and has been proposed to play a role in growth suppression mediated by p53 (Ma *et al.*, 2003). The fact that *ARID3A* deletions co-occur with *TP53* mutation suggests that other mechanisms may also contribute to the activation of *ARID3A* since inactivation of *TP53* would otherwise be sufficient to inactivate *ARID3A*. Down-regulation of *GADD45β* was found to be associated with human hepatocellular carcinoma but was inversely correlated with the presence of mutant p53 (Qiu *et al.*, 2003). In addition, p53 is believed to be involved in regulating the expression of murine *Gadd45β* (Balliet *et al.*, 2001), suggesting that inactivation of *GADD45β* and *TP53* should be negatively, rather than positively, associated. However, p53 knockdown in a human glioma cell line showed that *GADD45α* could be induced by a p53-independent mechanism (Heminger *et al.*, 2006), and it is therefore possible that the same may be true of *GADD45β*, in which case

inactivation of both genes may provide an additional growth advantage on the tumour. Neither *ARID3A* nor *GADD45β* resides within an MCR, which casts doubt on the observed associations.

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

5.5.2 Co-occurrence of amplified and deleted candidate cancer genes

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

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

a	*b*
c	*d*

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

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

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

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

A

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

B

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

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

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

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

A

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

B

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

C

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

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

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

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

phosphatase 1 alpha (PPP1CA) (Kim *et al.*, 2003b). Along with *CCND1*, *PPP1CA* is overexpressed in an amplicon on chromosome 11q13 that is frequently observed in oral squamous cell carcinomas (Hsu *et al.*, 2006) but also in oesophageal squamous cell carcinomas (Huang *et al.*, 2006b). PPP1CA has recently been shown to contribute to oncogenic Ras- and p53-induced cell cycle arrest (Castro *et al.*, 2008). Therefore, it is possible that the amplification of *PPP1R10* counteracts the tumour suppressive effects of *PPP1CA*, which is amplified and overexpressed due to its proximity to *CCND1* and, potentially, other amplified cancer genes. However, this hypothesis clearly needs to be verified experimentally, especially since the overlap between genes is small.

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

5.6 Discussion

The purpose of the work described in this chapter was to identify CIS genes that were significantly amplified or deleted in human cancer cell lines. As discussed in Chapter 4, mouse candidate cancer genes identified by retroviral insertional mutagenesis can help to narrow down the candidates in regions of copy number change in the human cancers. In

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

Known oncogenes were over-represented among significantly amplified CIS genes, demonstrating that the method used to rank genes was a successful approach for finding promising candidates. Some genes may not be amplified or deleted in a large number of cell lines across all cancer types, but may be heavily implicated in tumourigenesis in a subset of cancers. To identify such tissue-specific candidates, it is essential to perform the analysis independently on each type of cancer. The chosen method ignores genes that are amplified or deleted in few cell lines. However, it is difficult to endorse such candidates unless evidence to support a role in tumourigenesis has been previously demonstrated, as it has for *MEIS1* and *NFKB1*, which were identified in Section 4.5.2.1.1. Nevertheless, a significance threshold of $P<0.05$ is potentially too stringent. The *P*-value for each candidate is based on the number of amplifications compared with the

distribution of the number of amplifications for non-CIS genes. Since amplicons are often large and encompass multiple genes, it is highly probable that a number of non-CIS genes will be amplified for every CIS gene, even when the CIS gene is the target of amplification and other genes are passengers. It is therefore unwise to implement a strict cut-off when selecting promising candidates, but the method does provide a way to rank the candidates so that they can be compared against one another. Ranking the candidates according to the maximum copy number, rather than the number of amplicons, provides a method for identifying alternative candidates that, like *MEIS1* and *SLAMF6*, are amplified to high copy number but not across a large number of cell lines.

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

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