Chapter 4 Using mouse candidate cancer genes to narrow down the candidates in regions of copy number change in human cancers

4.1 Introduction

As discussed in Section 1.3.3, copy number changes are a common feature of cancer genomes, and can be identified using comparative genomic hybridisation (CGH)-based techniques. However, regions of copy number change are often large and encompass many genes, making it difficult to identify the "critical" genes that contribute to the tumourigenic process. Candidate cancer genes identified by insertional mutagenesis in the mouse can be used in a cross-species oncogenomics approach to narrow down the candidates within regions of copy number change in human tumours. The use of cross-species comparative analysis for cancer gene discovery is discussed in Section 1.5. In this chapter, mouse candidate cancer genes are used to identify orthologous candidates within regions of copy number change in 713 human cancer cell lines generated using SNP array CGH. The analyses were performed as part of a collaboration with the Netherlands Cancer Institute (NKI), published in Cell (Uren *et al*., 2008), and therefore, rather than using the mouse candidate cancer genes generated from the work described in Chapters 2 and 3, lists of candidates were provided by the NKI.

The datasets are introduced in Section 4.2. This is followed, in Section 4.3, by a description of the methods used to process the copy number data into regions of copy number change, and gains and losses within the human cancer cell lines are characterised in Section 4.4. In Section 4.5.1, the mouse and human datasets are compared to determine whether retroviral insertional mutagenesis is relevant to the discovery of amplified and deleted cancer genes in humans. Promising cancer gene candidates that are both disrupted by insertional mutagenesis in the mouse and amplified or deleted in human cancers are presented in Section 4.5.2. A range of algorithms have been developed for identifying regions of copy number change within CGH data, and these are described and compared in Section 4.6. Finally, in Section 4.7, the mouse candidate cancer genes are combined with copy number variation (CNV) data from apparently healthy individuals to determine whether there is any overlap between candidates and regions of CNV.

Since the ploidy of the cell lines, and therefore the exact copy number of alterations, is difficult to establish, the terms "gain and "amplicon" are used interchangeably throughout this thesis to mean any gain of copy number, irrespective of the size or nature of the alteration.

4.2 Description of the datasets

As well as the datasets described below, the set of known cancer genes from the Cancer Gene Census (Futreal *et al.*, 2004) was also used. This is described in Section 2.2.3.

4.2.1 Mouse candidate cancer genes identified by retroviral insertional mutagenesis

As mentioned in the introduction, some of the work described in this chapter was undertaken as part of a collaboration with the NKI (Uren *et al*., 2008, reprinted on p.365). The gene lists used in this chapter were therefore provided by the NKI but were generated from the analysis of insertion sites identified in the retroviral insertional mutagenesis screen described in Chapter 2. There were 6 lists of putative tumour suppressor genes. These included 3 lists comprising all genes in which there were insertions in the entire transcribed region, including UTRs and introns, only in the translated region (no UTRs) but including introns, and only in the coding region (no UTRs or introns). These lists are described throughout this thesis as genes in the transcribed region, translated region, and coding region, respectively. A further 3 lists contained genes with insertions in the same regions, but only where insertions comprised 2 or more sequence reads. Insertions represented by only 1 read are considered less likely to contribute to tumourigenesis (see Section 2.8) and are therefore predicted to have a reduced overlap with human deletions. 2 additional lists contained genes that were closest to CISs with *P*-values of less than 0.05 and 0.001, as determined using the kernel convolution (KC)-based statistical method (de Ridder *et al.*, 2006, see Sections 1.4.2.1.2 and 2.10.2). From these, lists were also generated for genes that were adjacent to CISs of $P<0.05$ and $P<0.001$ but were further away than the closest gene. For each gene list, the human orthologues and their genomic coordinates were extracted from Ensembl version 37 using Ensembl BioMart (see Section 3.2.1). Table 4.1 shows the number of mouse genes and human orthologues in each gene list. The $P<0.001$ and $P<0.05$ CISs and their associated nearest and further mouse genes

Gene List	Number of mouse genes	Number of mouse genes with human orthologues	Number of human orthologues in CIS gene list	% of human orthologues in CIS gene list
ORF only	266	240	41	17.1
ORF only (no singletons)	86	75	22	29.3
Translated region only	3024	2647	216	8.2
Translated region only (no singletons)	1331	1163	173	14.9
Transcribed region only	3773	3316	275	8.3
Transcribed region only (no singletons)	1706	1498	227	15.2
CIS nearest P<0.05	559	424	196	46.2
CIS nearest P<0.001	355	265	155	58.5
CIS further P<0.05	505	362	85	23.5
CIS further P<0.001	313	219	66	30.1

Table 4.1. Description of the lists of mouse candidate cancer genes used for comparison with human cancer copy number data. "[ORF, Translated region, Transcribed region] only" are lists of genes containing insertions only in the open reading frame, translated region (but including introns) or transcribed region, respectively. "no singletons" means that the list does not include genes that only contain insertions represented by a single read. "CIS nearest $P<0.05$" and "CIS nearest $P<0.001$" contain genes nearest to CISs identified by the kernel convolution (KC)-based method. "CIS further $P<0.05$" and "CIS further $P<0.001$" contain genes that flank CISs identified by the KC-based method but are not the nearest genes. The columns labelled "Number/% of human orthologues in CIS gene lists" show the overlap of each list with the list of candidate cancer genes generated and described in Chapters 2 and 3.

are listed in Appendix D. Due to their length, the lists of candidate tumour suppressor genes are not included, but are available on request.

Table 4.1 also shows the overlap of each gene list with the list of candidate cancer genes generated and described in Chapters 2 and 3 (shown in Appendix B2 and referred to here as the CIS gene list). The CIS gene list contains only genes that are associated with a significant CIS and this, together with the fact that the screen identifies mainly oncogenes, accounts for the small overlap with the tumour suppressor gene lists, in which genes may contain any number of insertions. The differences between the CIS gene list and the remaining lists may reflect differences in gene selection, i.e. a more sophisticated method was used to assign insertions to genes in the CIS gene list, and in read and insertion site processing, which were more conservative for the CIS gene list. Candidates from the CIS gene list are used in Chapter 5, where it is compared to higher resolution human CGH data (Section 5.3), as well as to the CGH data described in this chapter (Section 5.4).

4.2.2 Copy number data for human cancer cell lines

Comparative genomic hybridisation (CGH) data were generated by the Wellcome Trust Sanger Institute (WTSI) Cancer Genome Project for 713 human cancer cell lines from 29 tissues. A list of all cell lines and their tissue of origin is provided in Appendix E and is summarised in Table 4.2. None of the chosen cell lines had a common ancestor, according to cell line identity typing also performed by the WTSI Cancer Genome Project (http://www.sanger.ac.uk/genetics/CGP/Genotyping/synlinestable.shtml). This is important, since an amplicon or deletion might otherwise appear to be recurrent simply because it is within synonymous cell lines. CGH was performed using two Affymetrix GeneChip® Human Mapping Arrays. The 10K array, which comprises 11,555 SNPs, was used for 313 cell lines, while the 10K 2.0 array, comprising 10,204 SNPs, was used for the remaining 400 lines. 10,136 SNPs were shared between the two arrays, and both used the Affymetrix GeneChip® Mapping 10K assay, described in Section 1.3.3.2. The SNPs were mapped to the NCBI 35 human genome assembly. The mean distance between SNPs was 258.50 (±634.21) kb in the 10K array, and 292.82 (±683.49) kb in the 10K 2.0 array. The minimum distance was 2 bp and 11 bp for the 10K and 10K 2.0 arrays, respectively, and the maximum distance was 24.81 Mb for both arrays. 9.4% of human protein-coding genes in Ensembl v37 (extracted using Ensembl BioMart, see

Tissue of origin	Number of cell lines
Lung	131
Haematopoietic and lymphoid	117
Breast	43
Skin	42
Central nervous system	40
Unknown	39
Large intestine	38
Autonomic ganglia	29
Bone	23
Kidney	21
Soft tissue	20
Oesophagus	20
Stomach	19
Upper aerodigestive tract	19
Ovary	18
Pancreas	14
Urinary tract	13
Liver	11
Thyroid	11
Cervix	11
Endometrium	10
Biliary Tract	6
Pleura	5
Testis	3
Vulva	2
Prostate	2
Eye	2
Placenta	2
Adrenal gland	1
Small intestine	1
Total	713

Table 4.2. Tissues of origin of human cancer cell lines used in the 10K SNP array CGH analysis.

Section 3.2.1) contained at least one SNP in the 10K array, while 9.0% contained at least one SNP in the 10K 2.0 array. Genes were defined as the longest Ensembl gene transcript. The 10K and 10K 2.0 arrays contained an average of 0.176 (±0.735) and 0.157 (±0.648) SNPs per protein-coding gene, respectively. The interSNP distances and number of SNPs per gene are shown in Figures 4.1 and 4.2, respectively. The largest gaps between adjacent SNPs occur at the centromeres, while some gaps correspond to other regions of the genome that have not been assembled, e.g. due to highly repetitive sequences.

For each cell line. the raw intensity values were normalised internally. This involved calculating the value for each SNP as a total of all the SNPs on the array, and obtaining a copy number ratio for each SNP by dividing the SNP value by the value for the same SNP from a pool of reference normal samples. This is the point at which I received the data. The copy number data for all cell lines are available for download from ftp://ftp.sanger.ac.uk/pub/CGP/10kData. Data generated on the 10K and 10K 2.0 arrays is pooled in subsequent analyses and is collectively referred to as 10K data.

4.2.3 Copy number variants (CNVs)

CNVs are regions within the genome that vary in copy number. Germline CNV regions identified within 270 HapMap samples from Redon *et al.* (2006) were downloaded from http://www.sanger.ac.uk/humgen/cnv/data/cnv_data/. Merged CNVs identified using the Whole Genome Tilepath (WGTP) array and Affymetrix GeneChip Human Mapping 500K early access array (500K EA) were used. The WGTP array comprises 26,574 BAC clones, while the 500K EA array covers 474,642 SNPs. The WGTP and 500K EA platforms are complementary, since they are able to detect smaller and larger CNVs, respectively (Kehrer-Sawatzki, 2007). There are 1,447 merged CNVs that cover ~12% of the genome. 1,390 CNVs that mapped to autosomes in the NCBI 35 human build were used in this analysis.

4.3 Processing the copy number data

The copy number ratios at individual SNPs must be processed into regions of copy number change. As discussed in Section 1.3.3.2, a variety of methods have been

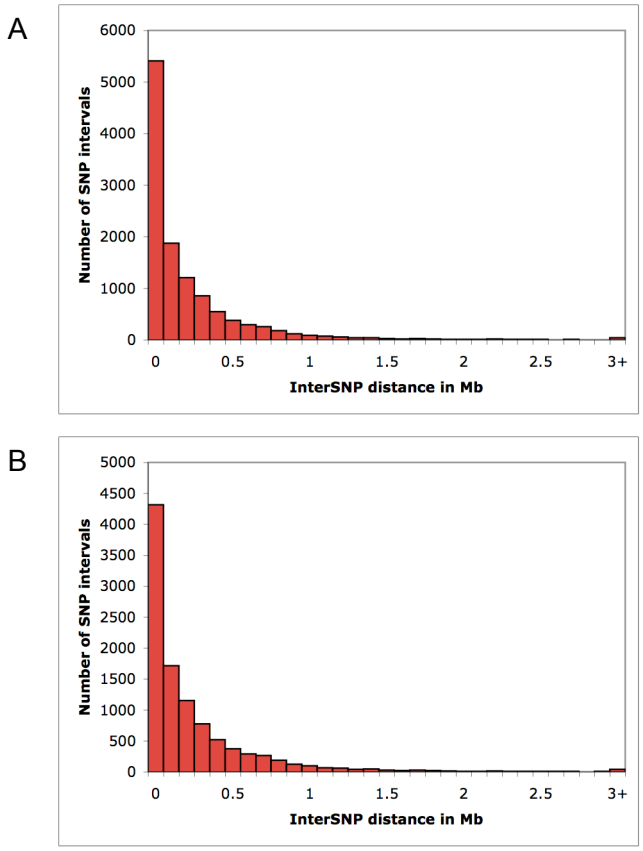

Figure 4.1. The distance between the genomic coordinates of adjacent SNPs on the 10K (A) and 10K 2.0 (B) SNP arrays.

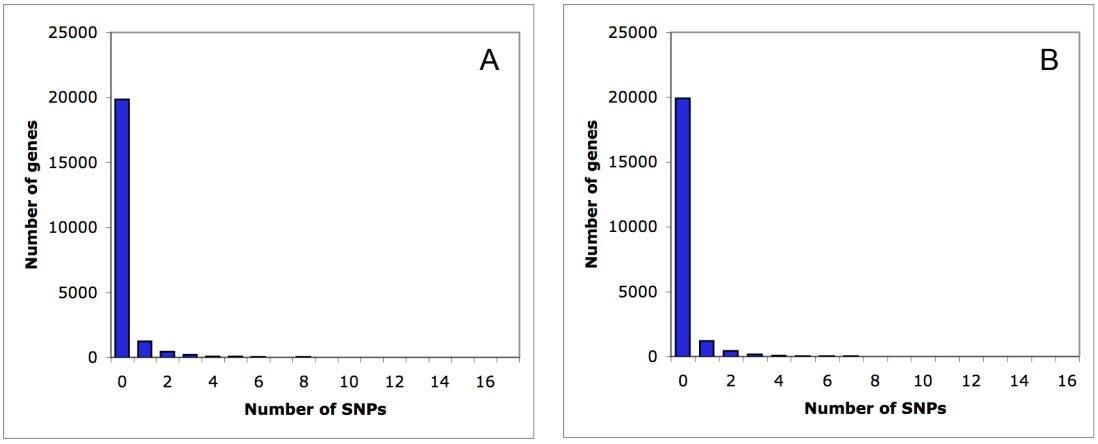

Figure 4.2. The number of SNPs per human protein-coding gene on the 10K (A) and 10K 2.0 (B) SNP arrays.

developed for this purpose. At the time of the analysis, most of the available algorithms had been developed primarily for conventional array CGH, i.e. using large genomic clones (see Section 1.3.3.2). In a comparison of 11 methods, DNAcopy (Olshen *et al.*, 2004) performed consistently well (Lai *et al.*, 2005), and a comparison of 3 segmentation methods by Willenbrock and Fridlyand (2005) demonstrated that DNAcopy performed better than GLAD (Hupe *et al.*, 2004) and HMM (Fridlyand *et al.*, 2004). A further benefit of DNAcopy is that it is freely available as an R package in BioConductor (http://www.bioconductor.org/). BioConductor is an open source software project that provides tools, mostly written in R, for analysing genomic data. DNAcopy (version 1.4.0) was therefore chosen as the method for detecting regions of copy number change in the 10K CGH data.

DNAcopy uses a method called circular binary segmentation (CBS) to identify change-points in CGH data, which is input as $\log_2$ intensity ratios at consecutive positions in the genome. The change-points correspond to positions in the genome where the DNA copy number has significantly changed. For each cell line, the copy number ratios for all SNPs were converted to $\log_2$-ratios and were smoothed, using a method within DNAcopy, to remove single point outliers before segmentation. Copy number ratios of 0 were given a $\log_2$-ratio of -6. Change-points may result from local trends in the data, and therefore all change-points that were less than 3 standard deviations apart were removed. Default parameters were used for the segmentation. Different values were tested for the parameter alpha but, upon visual inspection of the graphical outputs, the default value of alpha=0.01 appeared to be most suitable. Increasing alpha increases the sensitivity, resulting in more change-points but, potentially, more false positive change-points. Decreasing alpha results in fewer change-points, and regions of copy number change may therefore be missed. Increasing the number of standard deviations below which change-points were removed resulted in the loss of potentially important change-points. Figure 4.3 shows an example of how changing the parameters can affect the output of DNAcopy for chromosomes 1 and 6 of ovarian cancer cell line 41M-CISR. The removal of change-points less than 3 standard deviations apart results in the loss of a change-point in chromosome 1 (Figure 4.3B). However, the slight difference in copy number between the 2 arms of the chromosome may be due to trends in the data, and the difference in copy number is small. Increasing the number of standard deviations to 4 results in the loss of a change-point in chromosome 6, for which there is a clear step in copy number that does look real (Figure 4.3C). Increasing alpha from 0.01 to 0.05 results in the inclusion of

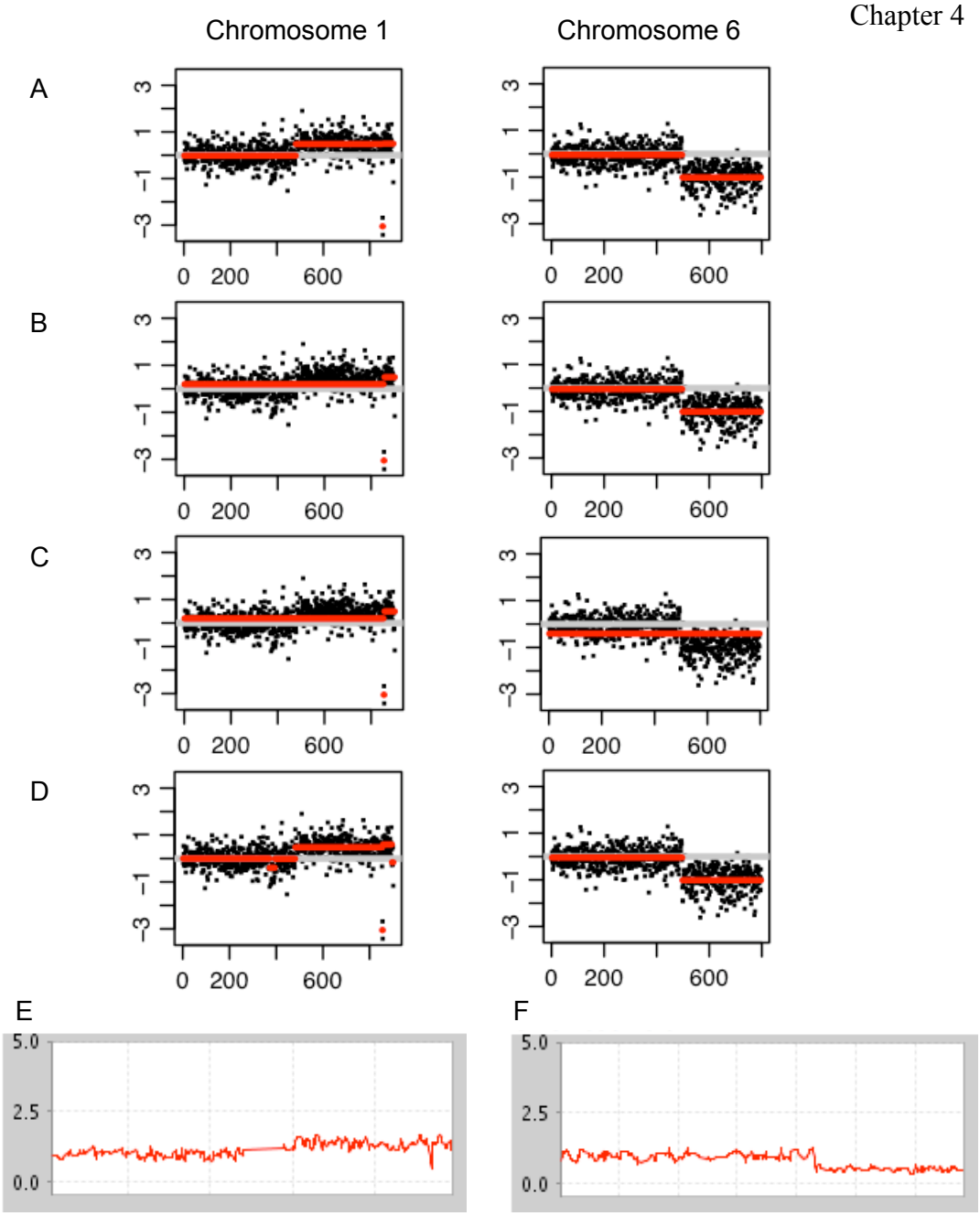

Figure 4.3. Altering the values for parameters in DNAcopy leads to differences in the regions of copy number change detected by the algorithm, as demonstrated for chromosomes 1 and 6 of ovarian cancer cell line 41M-CISR. (A) Default parameters. (B) Default parameters and removal of change-points less than 3 standard deviations apart. (C) Default parameters, smoothing and removal of change-points less than 4 standard deviations apart. (D) Alpha = 0.05 plus smoothing. (E) Copy number for chromosome 1, with values averaged across 3 consecutive SNPs. (F) Copy number for chromosome 6, with values averaged across 3 consecutive SNPs. Figures E and F are taken from the WTSI Cancer Genome Project website (http://www.sanger.ac.uk/genetics/CGP/) and give a clearer picture of the copy number across the chromosome. Figures A-D are extracted from the output of DNAcopy. Removing change-points that are close together results in fewer regions being detected, and the larger the number of standard deviations below which change-points are removed, the more regions are missed. Increasing alpha leads to the inclusion of additional change-points and, therefore, regions of copy number change.

additional change-points in chromosome 1 (Figure 4.3D). Since the data is relatively low resolution, it is highly possible that a region of copy number change may be represented by just 1 or 2 SNPs. However, it is also possible that such SNPs are anomalies and, to avoid the identification of false positives, this is the preferred assumption.

DNAcopy identifies changes in DNA copy number but does not indicate which regions are unchanged and which are gains or losses. It is therefore the responsibility of the user to set thresholds for calling gains and losses based on the mean $\log_2$-ratios of predicted segments. A disadvantage of DNAcopy is that it operates on individual chromosomes rather than the entire genome and the mean $\log_2$-ratios of segments representing no copy number change, or representing a gain or loss of a certain number of copies, will differ slightly across the genome. This makes it difficult to determine what is "normal" and therefore to call gains and losses, and the exact number of copies within a gain or loss cannot be clearly determined. Willenbrock and Fridlyand (2005) have developed an algorithm called MergeLevels that merges segments across the genome that are not significantly different from one another and so produces a more interpretable set of copy number levels. Combining DNAcopy and MergeLevels was found to be more effective than using DNAcopy alone (Willenbrock and Fridlyand, 2005). MergeLevels is freely available within an R/BioConductor package called aCGH. Therefore, for each cell line, the DNAcopy segmentation results were merged across all autosomes using MergeLevels with default parameters, which were considered appropriate upon inspection of the graphical outputs. Example outputs for the kidney cancer cell line 786-0 and endometrial cancer cell line AN3-CA are shown in Figure 4.4. The merged segments with a $\log_2$-ratio closest to 0 were defined as the level of no copy number change and, to enable comparison across cell lines, this $\log_2$-ratio was set to 0 and all other $\log_2$-ratios were normalised accordingly. Figure 4.5A shows the distribution of $\log_2$-ratios of the segments predicted by DNAcopy across all cell lines, while Figure 4.5B shows the distribution of $\log_2$-ratios of the merged segments. The $\log_2$-ratios of the merged segments show a series of peaks and troughs that may reflect distinct copy number levels. The large peak at -6 represents segments for which the copy number ratio of individual SNPs was 0. The $\log_2$-ratios of the merged segments were converted to copy number ratios (Figure 4.6), and troughs in the distribution were used to set thresholds for subsequent analyses (see Sections 4.4 and 4.5.1.1).

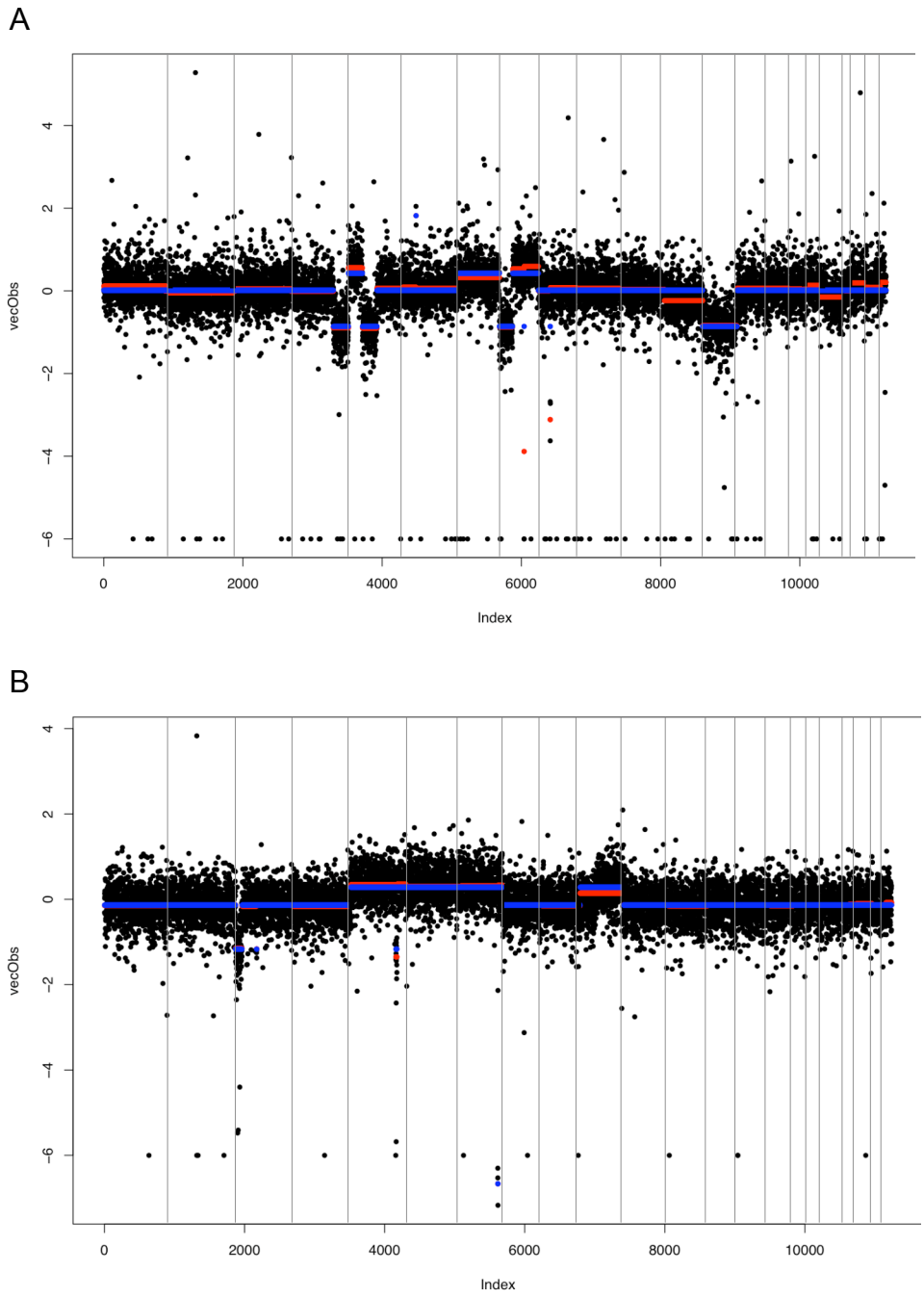

Figure 4.4. Graphical output from MergeLevels for human cancer cell lines 786-0 (A) and AN3-CA (B). Black dots represent the $\log_2$-ratios for individual SNPs ordered across the genome. The mean $\log_2$-ratios for segments identified by DNAcopy are shown in red. Merged segments generated by MergeLevels are shown in blue. Segments are merged across chromosomes into a set of copy number levels. The x-axis shows the position within the genome, while the y-axis measures the $\log_2$-ratio. Vertical lines represent the division of chromosomes.

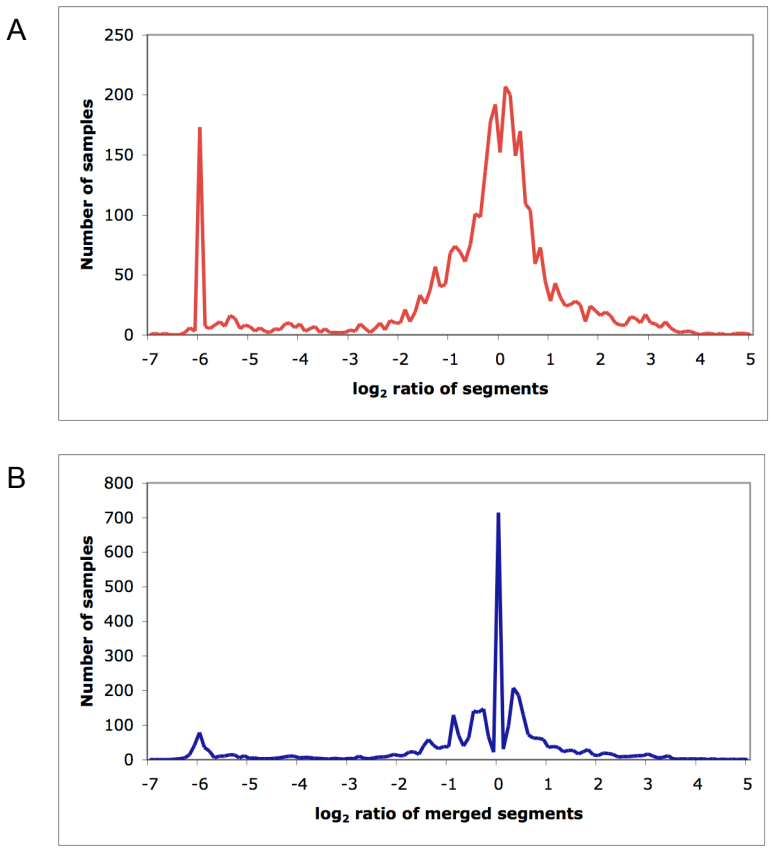

Figure 4.5. The number of human cancer cell lines with segments of varying $\log_2$-ratio following processing with DNAcopy (A) and DNAcopy plus MergeLevels (B).

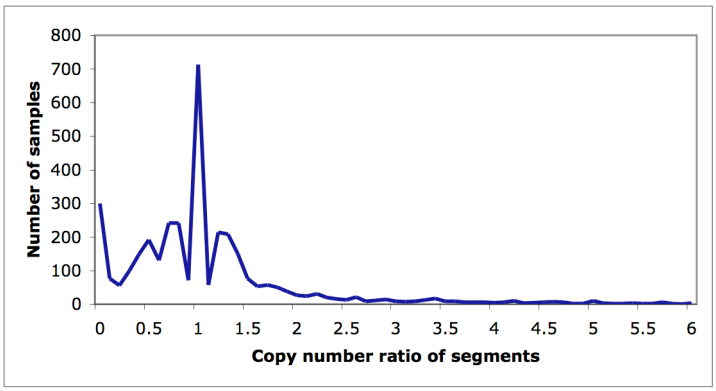

Figure 4.6. The number of human cancer cell lines with segments of varying copy number ratio following processing with DNAcopy plus MergeLevels. Troughs in the data were used to set thresholds for the analysis of gains and losses.

4.4 Characterising gains and losses in cancer genomes

All segments with a copy number of 1.6 or more were designated gains, and all segments with a copy number of 0.6 or less were designated losses. Segments with a copy number of 0.2 or less were designated homozygous deletions. Each of these thresholds was within a trough in the distribution of copy numbers for merged segments in Figure 4.6. The average number of gains per cancer cell line was 1.47 (±2.02), and the average size across all cell lines was 21.37 (±34.15) Mb. Amplicons contained an average of 213.68 (±341.45) genes. The average number of losses per cell line was 4.43 (±3.69) and the average size was 19.56 (±33.83) Mb, encompassing 195.62 (±338.32) genes. The average number of homozygous deletions was 1.53 (±1.99) per cell line. The average size was 0.98 (±2.68) Mb, encompassing 19.64 (±20.69) genes. Therefore, homozygous deletions were significantly smaller than amplicons and heterozygous deletions and contained fewer genes. Homozygous deletions have been previously shown to contain fewer genes than other regions of the genome (Cox *et al.*, 2005, see Section 1.3.3.3), and this analysis shows that, in general, the deletion of both copies of a gene is more likely to be deleterious to a cell than the loss of one copy or the gain of copies. The distributions of amplicon and deletion lengths are shown in Figure 4.7.

For each cancer type, the number of cell lines containing gains was counted, and the 2-tailed Fisher Exact Test was used to determine whether there was any significant difference between the observed and expected number of each cancer type. Cell lines derived from the oesophagus were over-represented (P=7.11x10^{-4}), suggesting that oesophageal tumours are particularly prone to genomic instability. Haematopoietic and lymphoid cancer cell lines were under-represented (P=5.32x10^{-5}), reflecting the fact that they often contain balanced translocations that do not show a change in DNA copy number (see Section 1.3.3.4) and that, in some cases, few genetic events are thought to be required for tumour development. For example, acute lymphoblastic leukaemias contain an average of 3.83 deletions and focal amplifications are rare (Mullighan *et al.*, 2007). Since most cell lines contained deletions, there was no significant difference between the observed and expected numbers of cancer types containing deletions.

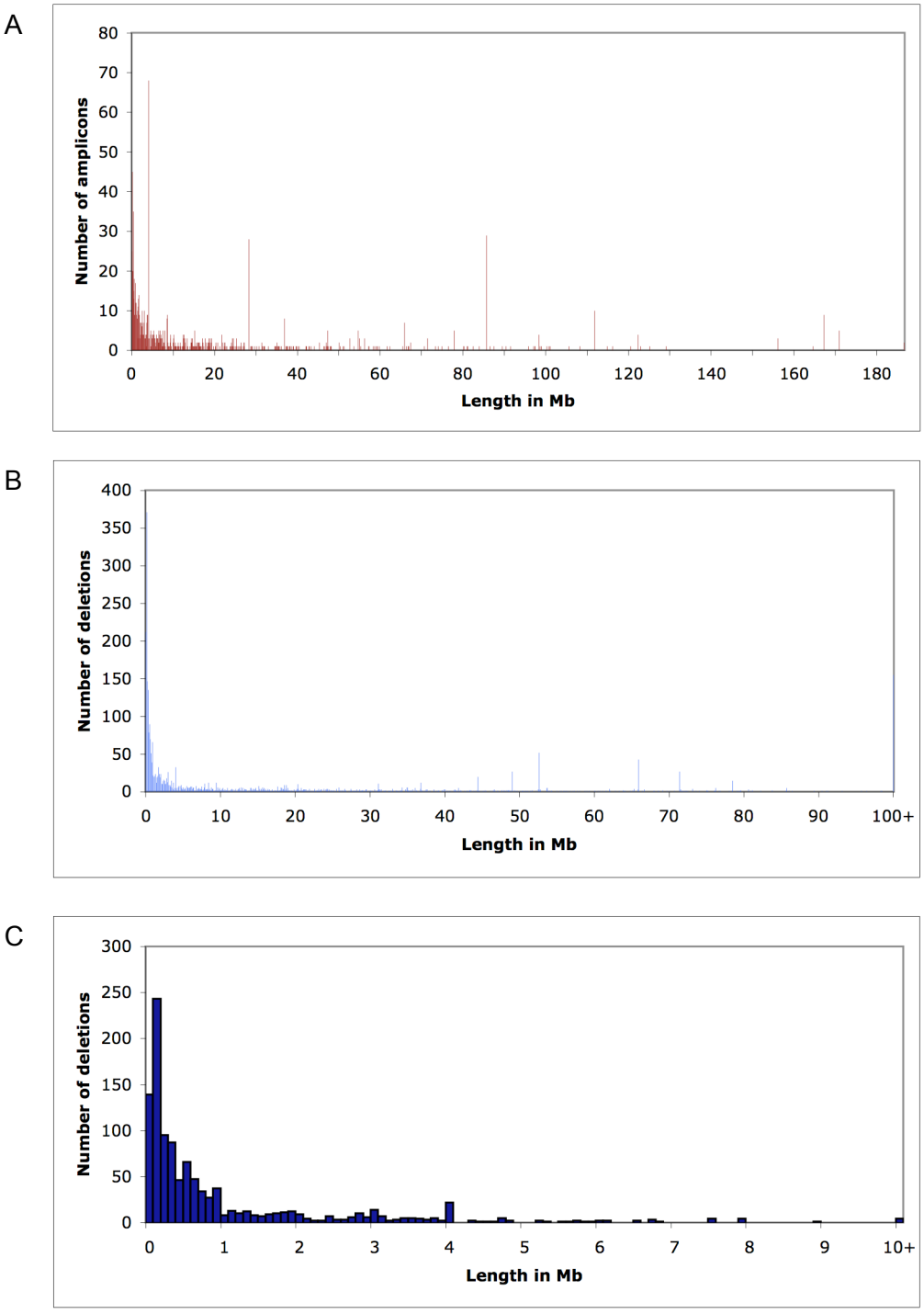

Figure 4.7. Distribution of the lengths of amplicons (A), deletions (B) and homozygous deletions (C) in 713 human cancer cell lines. Amplicons are defined as regions with a copy number greater than or equal to 1.6. Deletions and homozygous deletions are defined as regions with a copy number less than or equal to 0.6 and 0.2, respectively.

4.5 Comparative analysis of mouse candidate cancer genes and CGH data from human cancers

4.5.1 Global comparison

The purpose of the global comparison is to determine whether the human orthologues of candidate cancer genes identified by retroviral insertional mutagenesis by the Netherlands Cancer Institute are over-represented within regions of copy number change in the human cancer cell lines. Specifically, an over-representation of candidate oncogenes in human amplicons, and candidate tumour suppressor genes in human deletions, suggests that the retroviral insertional mutagenesis screen is relevant to human cancer, and may help to identify human cancer gene candidates within regions of copy number change.

4.5.1.1 Method

Rather than setting single copy number thresholds for gains and losses, a range of copy number thresholds were investigated. Thresholds were set as the centre-point of troughs in the graph shown in Figure 4.6, since these may represent transitions in the number of gene copies. The chosen thresholds were copy number ratios of less than or equal to 0.9, 0.6 and 0.2, and greater than or equal to 1.1, 1.6, 2.1, 2.5, 2.7, 3.1, 3.8, 4.3, 4.8, 5.2 and 5.9. The genomic coordinates of the human orthologues of all mouse genes were extracted from Ensembl version 37 using Ensembl BioMart. For each gene list described in Section 4.2.1, the number of mouse genes with human orthologues was counted. The same number of genes was selected randomly from among all mouse genes with human orthologues and this was repeated 1,000 times. For each of the 1,000 iterations, the number of human orthologues that resided within human cancer cell line segments with a mean copy number above or below a given threshold was counted. This produced a normal distribution of counts. The number of human orthologues of mouse candidate cancer genes in segments above or below each threshold was also counted. The Z-test was used to calculate the probability of obtaining a number greater than or equal to the observed count for the mouse candidates, based on the distribution of counts for the randomised genes. The procedure is summarised in Figure 4.8.

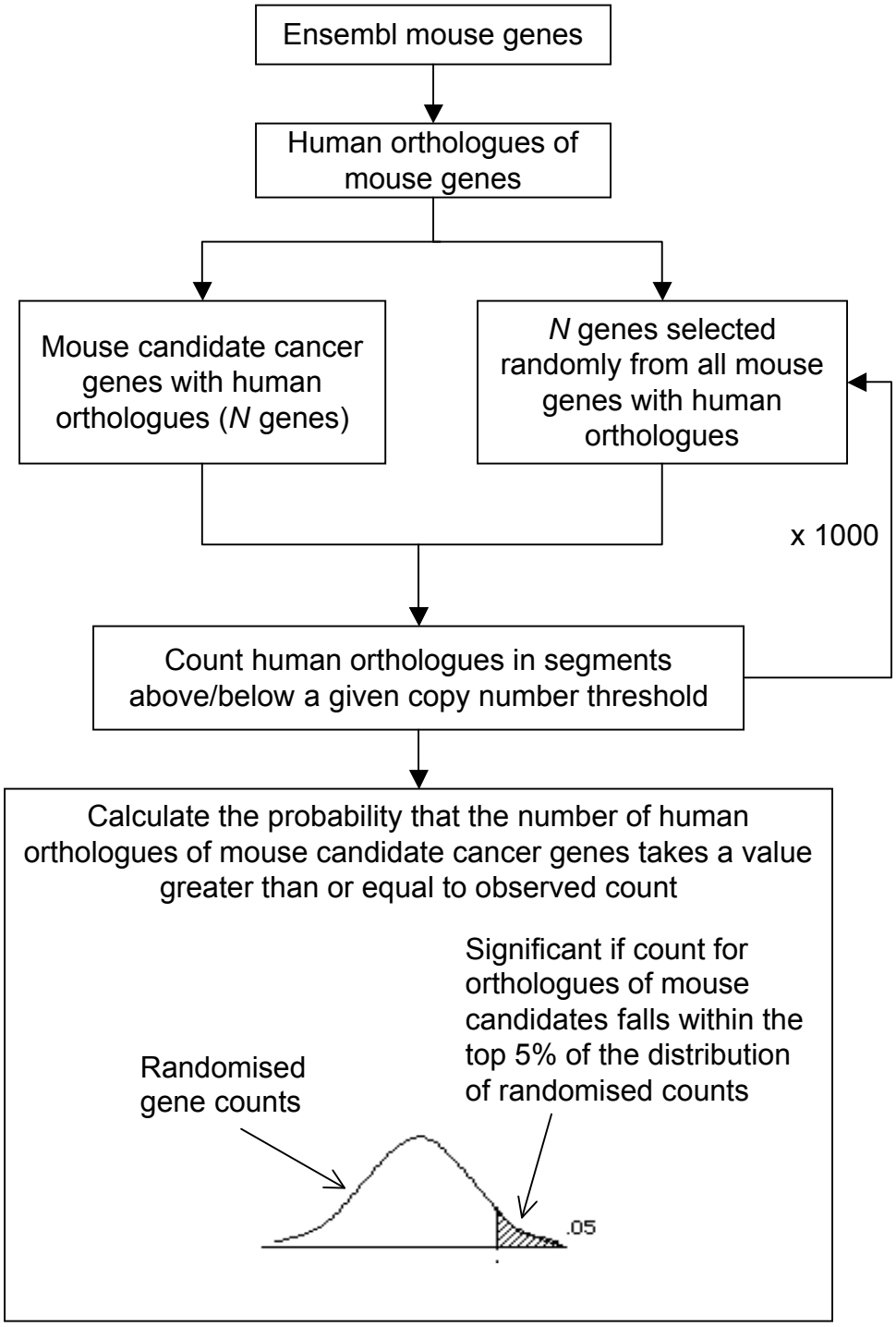

Figure 4.8. Overview of the method for identifying over-representation of the human orthologues of mouse candidate cancer genes in regions of human copy number change.

4.5.1.2 Setting the boundaries of amplicons and deletions

The start and end coordinates of the copy number segments generated by DNAcopy and MergeLevels correspond to the first and last SNPs for which the $\log_2$-ratios are not significantly different to other SNPs in the segment. Therefore, the copy number can be determined for all coordinates between these positions. It is, however, impossible to determine the copy number for coordinates within the interval between the first SNP and the preceding SNP, which corresponds to the end coordinate of the preceding segment, and between the last SNP and the proceeding SNP, which corresponds to the start coordinate of the proceeding segment. As shown in Section 4.2.2, the distance between SNPs can be very large, especially across unassembled regions of the genome such as centromeres. Setting the boundaries of an amplicon or deletion as the end of the previous segment and start of the next segment, or even using half-way points, could therefore result in a very high number of false positives among genes predicted to be amplified or deleted.

In order to choose an appropriate distance for the boundaries of amplicons and deletions, the global comparison was performed using a range of distances. Assuming that CIS genes are more likely to be amplified or deleted in human cancers than are other genes, the most appropriate distance should be that which gives the highest over-representation of CIS genes. The list of genes nearest to CISs with $P<0.001$ was used in this analysis. Amplicon boundaries were extended beyond the first and last amplified SNP by a distance of 0 kb, 200 kb, 500 kb, 1 Mb, 3 Mb and 5 Mb, or as far as the adjacent SNP, whichever was closer. The results are shown in Figure 4.9. The association between CIS genes and amplicons was strongest when the boundaries were not extended at all. However, at lower copy numbers and in greater numbers of cell lines at higher copy number, the association was less significant than when the boundaries were extended to 500 kb. Extending the boundaries to 1 Mb and beyond resulted in a considerable decrease in the association between CIS genes and amplicons. Therefore, 500 kb was chosen as the most suitable distance.

Known oncogenes from among the CIS genes that were identified within full-length amplicons (i.e. where the amplicon was extended as far as the adjacent, non-amplified SNPs) were compared to those identified within amplicons with a 0 kb or 500 kb extension of the amplicon boundaries (Table 4.3). While the non-extended amplicons

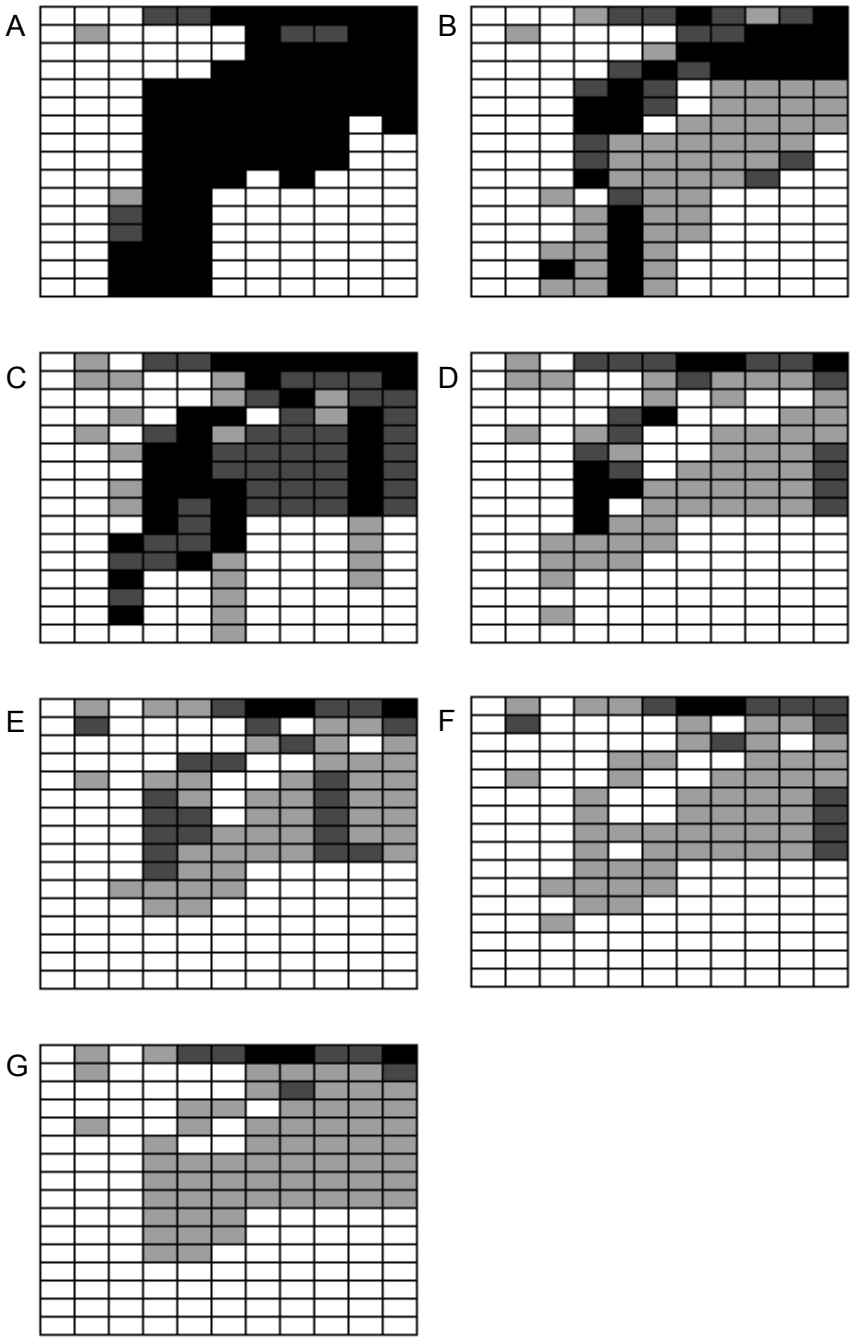

Figure 4.9. Over-representation of human orthologues of genes nearest to CISs in amplicons with boundaries extended beyond the first and last amplified SNP by a maximum distance of 0 kb (A), 200 kb (B), 500 kb (C), 1 Mb (D), 3 Mb (E), 5 Mb (F) and up to the adjacent, non-amplified SNPs (G). Each box represents the significance of the association between the selected 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. Columns represent amplicons, with (from left to right) copy number thresholds of greater than 1.1, 1.6, 2.1, 2.5, 2.7, 3.1, 3.8, 4, 4.8, 5.2 and 5.9. Rows represent the number of cell lines, which increases in increments of 1, up to a cut-off of 16 cell lines. For example, the box in the bottom right-hand corner of each figure represents the *P*-value for the over-representation of CIS genes that occur in amplicons of copy number greater than or equal to 5.9 in at least 16 cancer cell lines.

Copy number	Gene	Full length	0 kb	500 kb
4.8	*MYC*	14	10	14
	MYCN	9	4	9
	CCND1	1	1	1
4.3	*MYC*	14	10	14
	MYCN	9	4	9
	CCND1	1	1	1
	ZFPN1A1	1	1	1
	LMO2	1	1	1
	CCND3	2	0	2
2.7	*MYC*	29	22	29
	MYCN	12	7	12
	CCND1	4	4	4
	ZFPN1A1	1	1	1
	LMO2	2	2	1
	CCND3	3	0	2
	CCND2	1	1	1
	PIM1	2	1	2
	EVI1	1	1	1
	IRF4	1	1	1

Table 4.3. The number of amplicons in which known cancer genes among genes nearest to CISs are identified when the amplicon boundaries are altered. "Full length" applies to amplicons extended to the next, non-amplified SNP. "0 kb" applies to amplicons where the start and end correspond to the first and last amplified SNP. "500 kb" applies to amplicons extended to a maximum of 500 kb. Copy number values are given as the minimum copy number of amplicons.

missed some of the occurrences of amplified *MYC* and *MYCN* that were identified in the full-length amplicons, all occurrences were identified in the amplicons extended by up to 500 kb. Likewise, occurrences of amplified *CCND3* and *PIM1* were identified in the 500 kb amplicons but not the non-extended amplicons.

To demonstrate that the observed association between candidate cancer genes and human amplicons was real, full-length amplicons were shuffled across the genome. The length and mean copy number of each amplicon were conserved, but the location was shuffled. The method from Section 4.5.1.1 was then performed on the shuffled amplicons. As shown in Figure 4.10, the association of candidates with regions of copy number gain was completely abolished.

4.5.1.3 Comparison with lists of candidate cancer genes

Having chosen 500 kb as the maximum distance for extending amplicon and deletion boundaries, the method of Section 4.5.1.1 was applied to all of the gene lists outlined in Section 4.2.1. The results are shown in Figure 4.11. The lists of genes nearest to CISs with $P<0.001$ or $P<0.05$ are lists of candidate oncogenes, with those nearest to CISs with $P<0.001$ being stronger candidates for a role in tumourigenesis. This is reflected in the results, since both lists showed an over-representation of candidates within regions of amplification, but the association was stronger for genes near to a CIS with $P<0.001$. For both gene lists, the association became significant at copy number 1.6 and above, but for low-level copy numbers, the association was generally strongest for genes that were amplified in higher numbers of cell lines. Figure 4.12A shows the over-representation of known oncogenes within regions of copy number gain. The pattern of association was very similar to that obtained using genes nearest to CISs with $P<0.001$, suggesting that this list contains oncogenes that are relevant to human cancer. Almost all of the mouse tumours generated in the retroviral screen were lymphomas, and therefore it could be assumed that the candidate cancer genes identified in the screen are only relevant to similar cancers within humans. Therefore, the human cancer cell lines were divided into haematopoietic and lymphoid cell lines and all other cell lines (from solid tumours) and the global comparison was performed on each subset using genes nearest to CISs with $P<0.001$. As shown in Figure 4.13, the association was much weaker when only haematopoietic and lymphoid cell lines were considered. This may be partly because amplification is not a common mechanism of mutation in these cell types (see Section

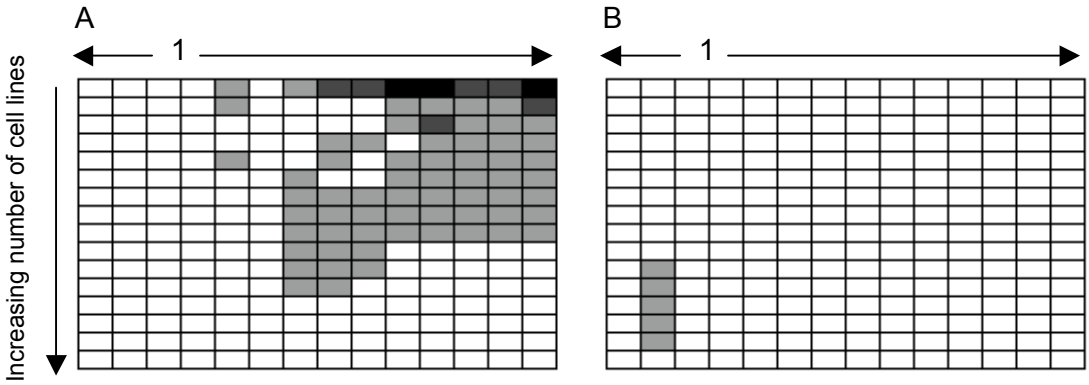

Figure 4.10. Over-representation of human orthologues of genes nearest to CISs in full-length human amplicons (A) and shuffled full-length amplicons (B). Each box represents the significance of the association between the selected genes and amplicons/deletions at a given copy number threshold and cell line number. $P<0.0001$, black; $P<0.001$, dark grey, $P<0.05$, light grey. Copy number thresholds below 1 represent deletions, with (from left to right) copy number thresholds of less than 0.2, 0.6 and 0.9. Copy number thresholds above 1 represent amplicons, with (from left to right) copy number thresholds of greater than 1.1, 1.6, 2.1, 2.5, 2.7, 3.1, 3.8, 4.3, 4.8, 5.2 and 5.9. The number of cell lines increases in increments of 1, up to a cut-off of 16 cell lines.

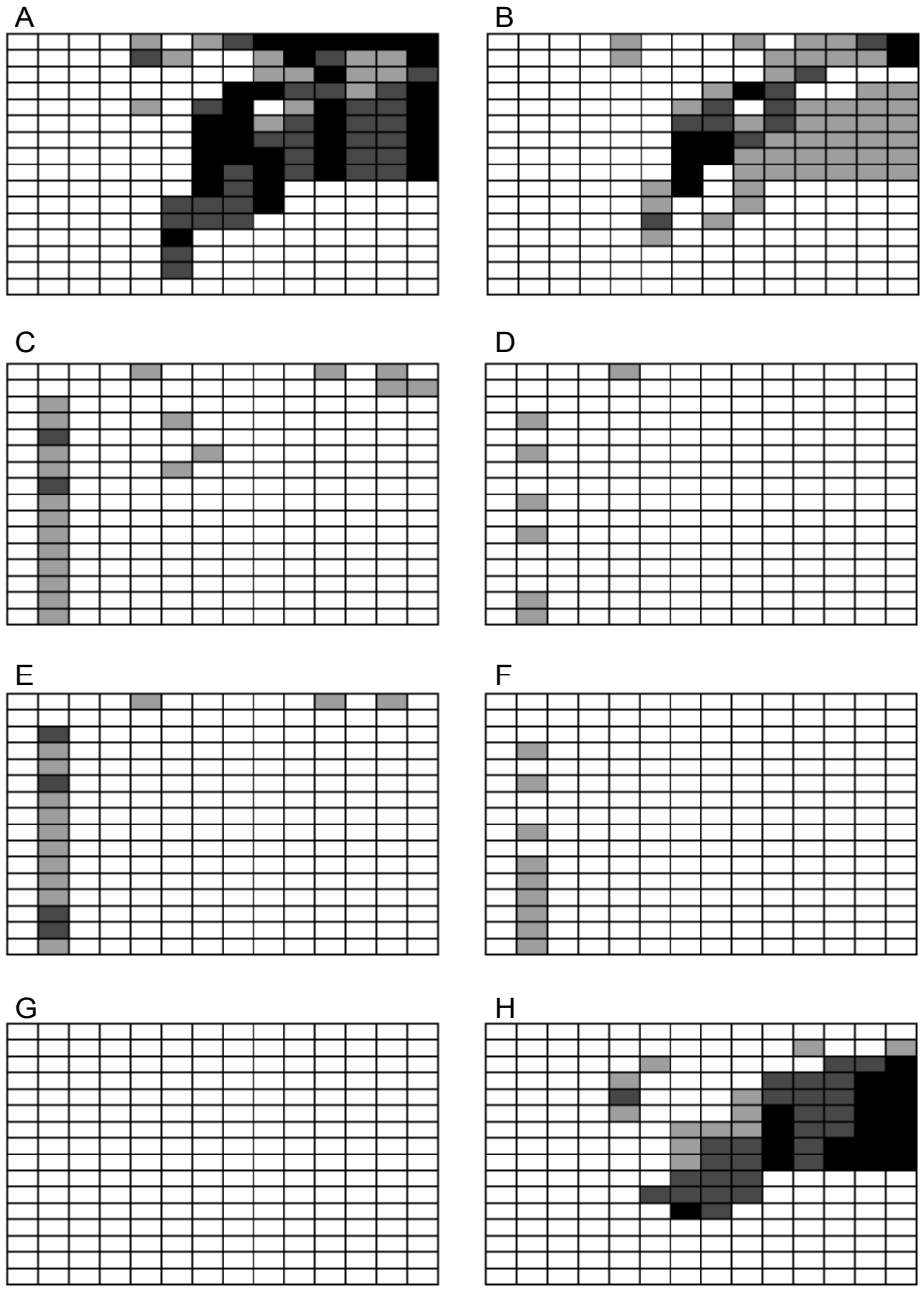

Figure 4.11. Over-representation of human orthologues of candidate cancer genes in regions of copy number change. (A) Genes nearest to CISs with *P*<0.001. (B) Genes nearest to CISs with *P*<0.05. (C) Genes with insertions within the transcribed region. (D) Genes with insertions but no singletons in the transcribed region. (E) Genes with insertions within the translated region. (F) Genes with insertions but no singletons in the translated region. (G) Genes with insertions in the coding region. (H) Genes with insertions but no singletons in the coding region. Each box represents the significance of the association between the selected 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. Columns from left to right represent copy number thresholds of less than 0.2, 0.6 and 0.9 (deletions) and greater than 1.1, 1.6, 2.1, 2.5, 2.7, 3.1, 3.8, 4.3, 4.8, 5.2 and 5.9 (amplicons). The number of cell lines increases in increments of 1, up to a cut-off of 16 cell lines.

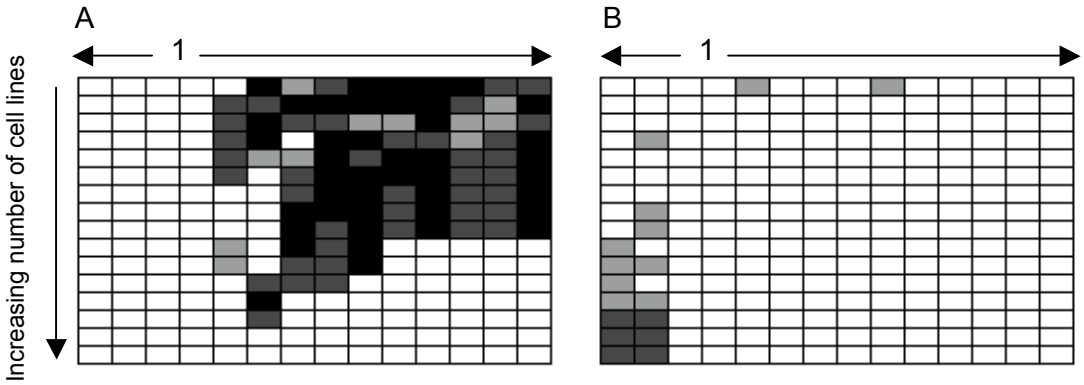

Figure 4.12. Over-representation of known oncogenes (A) and known tumour suppressor genes (B) in regions of copy number change in human cancer cell lines. Each box represents the significance of the association between the selected genes and amplicons/deletions at a given copy number threshold and cell line number. $P<0.0001$, black; $P<0.001$, dark grey, $P<0.05$, light grey. Copy number thresholds below 1 represent deletions, with (from left to right) copy number thresholds of less than 0.2, 0.6 and 0.9. Copy number thresholds above 1 represent amplicons, with (from left to right) copy number thresholds of greater than 1.1, 1.6, 2.1, 2.5, 2.7, 3.1, 3.8, 4.3, 4.8, 5.2 and 5.9. The number of cell lines increases in increments of 1, up to a cut-off of 16 cell lines.

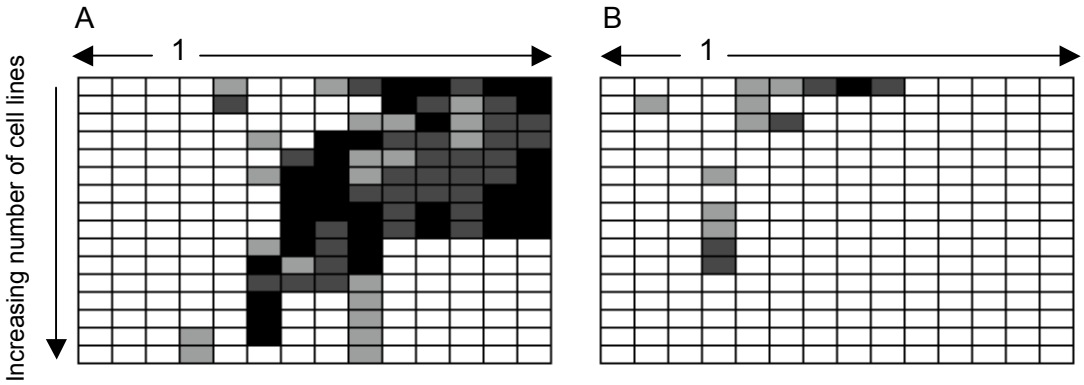

Figure 4.13. Over-representation of human orthologues of genes nearest to CISs with a *P*-value of <0.001 in regions of copy number change in human cancer cell lines derived from solid tumours (A) and haematopoietic and lymphoid cancers (B). See Figure 4.13 above for a description of how to interpret the figures.

4.4), but may also reflect the fact that the set of cell lines is smaller and therefore the comparison lacks power. Importantly, the pattern of association in solid tumours was similar to that for all cell lines and was highly significant. This demonstrates the relevance of retroviral insertional mutagenesis to the discovery of cancer genes in diverse human cancers, and shows that analysis of the full set of human cancer cell lines is warranted. Each cancer type provided in Table 4.2 was then separately tested for an association with the candidate cancer genes. Splitting the cancer cell lines into different types reduces the power of the analysis, and for most tumour types there was no clear association. However, cell lines derived from the autonomic ganglia, breast, upper aerodigestive tract, large intestine, oesophagus and stomach did show a significant overlap between mouse candidates and regions of copy number gain, although in the large intestine cell lines, there was also a significant overlap with regions of copy number loss (Figure 4.14).

The remaining lists are expected to contain candidate tumour suppressor genes. The results were similar for genes with insertions in transcribed and translated regions (Figure 4.11C-F). In both cases, including all genes containing insertions, rather than just those containing insertions represented by more than one read, generated a more significant association. This suggests that insertions represented by a single read ("singletons") in this retroviral screen are often important in tumourigenesis. As discussed in Section 2.7, the screen is not fully saturated due to the use of an insufficient number of enzymes in PCR and insufficient sequencing depth. Therefore, singleton insertions may result from these limitations, rather than because they are rare in the tumour mixture. However, for genes with insertions in the coding region, the reverse was observed, with a significant association only occurring when singleton insertions were omitted (Figure 4.11G-H, see below). As expected for tumour suppressor genes, the lists of genes with insertions in the transcribed and/or translated region were associated with deletions of copy number less than or equal to 0.6. However, the significance of the association was weak. When singleton insertions were included, there was also evidence of a weak association with regions of copy number gain. This is not surprising since the lists are likely to be contaminated with candidate oncogenes, as well as genes that do not play a role in tumourigenesis. The gene lists are long and yet tumour suppressor genes are less likely to be identified by insertional mutagenesis than are oncogenes and, as shown in Chapter 3, oncogenes are often disrupted by intragenic insertions. The association between known tumour suppressor genes and regions of copy number change is shown in Figure 4.12B.

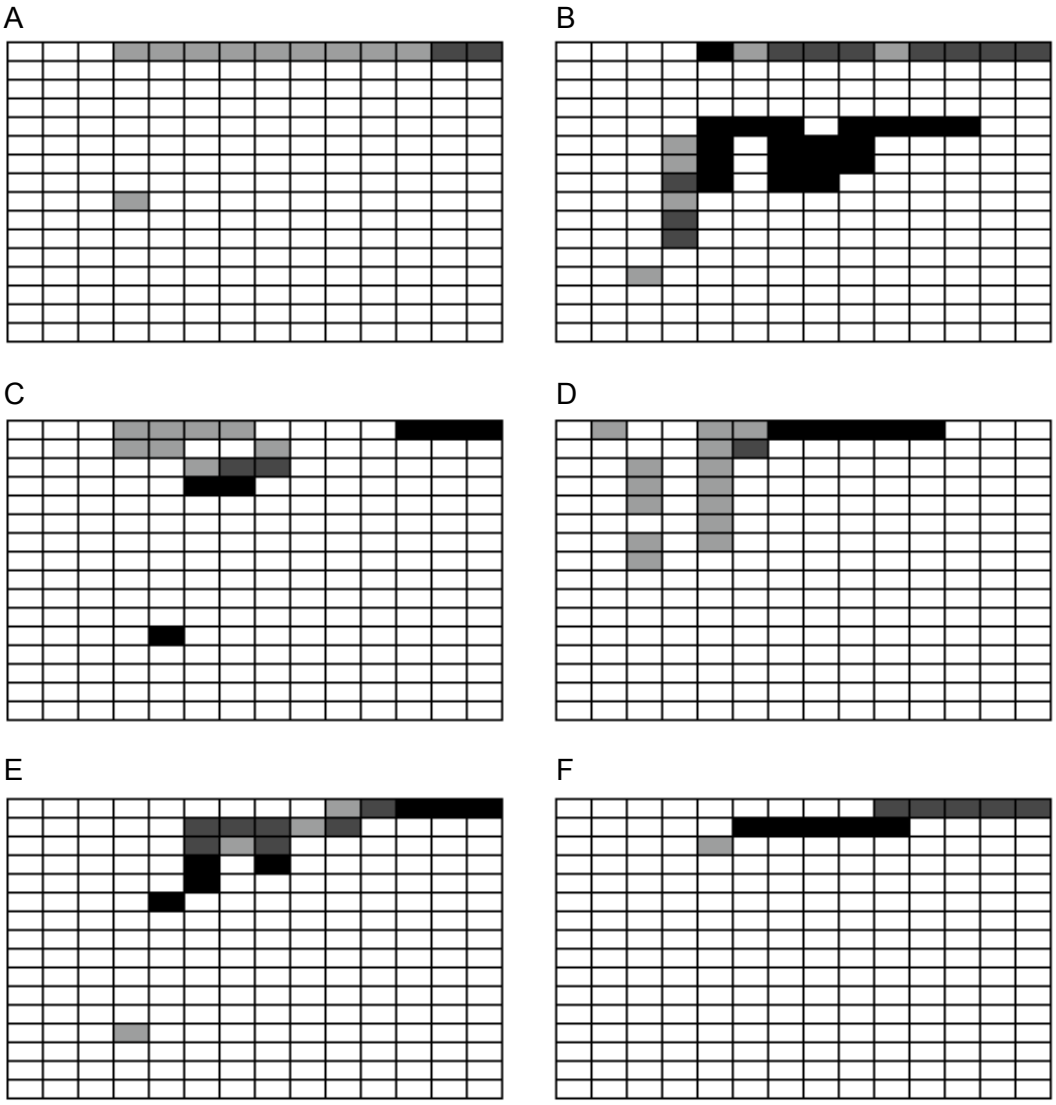

Figure 4.14. Over-representation of human orthologues of candidate cancer genes in regions of copy number change in cancer cell lines derived from the upper aerodigestive tract (A), autonomic ganglia (B), breast (C), large intestine (D), oesophagus (E) and stomach (F). Each box represents the significance of the association between the selected 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. Columns from left to right represent copy number thresholds of less than 0.2, 0.6 and 0.9 (deletions) and greater than 1.1, 1.6, 2.1, 2.5, 2.7, 3.1, 3.8, 4.3, 4.8, 5.2 and 5.9 (amplicons). The number of cell lines increases in increments of 1, up to a cut-off of 16 cell lines.

There was a more significant over-representation of genes within deletions of copy number less than or equal to 0.6, but also in deletions of copy number less than or equal to 0.2. However, the fact that the pattern was broadly similar for genes with insertions in transcribed and translated regions is encouraging, and suggests that the lists do contain tumour suppressor genes that are relevant to human cancer.

The results for genes containing insertions within the coding region did not show the expected pattern for tumour suppressor genes (Figure 4.11G-H). As mentioned above, when singleton insertions were included, a significant association was not observed with either amplicons or deletions. Omission of singleton insertions resulted in a pattern of association representative of oncogenes, i.e. showing an over-representation of genes within amplicons. The identities of genes that reside within human amplicons and deletions are provided in Section 4.5.2.

4.5.1.4 Determining whether the nearest gene to a CIS is the most likely candidate cancer gene

As discussed in Chapter 2, it can be difficult to determine which gene is being mutated by insertions within a CIS, especially when the insertions are intergenic and disrupt genes by enhancer mutation. CISs are often assigned to the nearest gene. Therefore, to test whether this is a sensible assumption, the overlap of the human CGH data with candidate genes nearest to CISs was compared to that observed for the next nearest genes to CISs. The method was performed as described in Section 4.5.1.1, whereby the number of genes closest to CISs that occurred within amplicons or deletions was compared to the number of randomly occurring genes in amplicons or deletions. This was then repeated for genes adjacent to, but further from, CISs. Thresholds in this analysis were the same as for previous comparisons. The method was performed on CISs with a *P*-value of <0.05 and <0.001, and the results are shown in Figure 4.15. As previously shown, there was a more significant over-representation within human amplicons of genes nearest to CISs with $P<0.001$ than $P<0.05$. However, for both significance levels, the clear overlap between human amplicons and genes nearest to CISs was almost absent for genes further from CISs. This suggests that the nearest gene to a CIS is generally the disrupted gene. *Plekhf1* and *Ltap* (also known as *Vangl2*) were the only two genes in the set of genes that are further from the CIS for which the human orthologues were amplified to a copy number greater than or equal to 5.2. However, in both cases, the nearest gene to the CIS

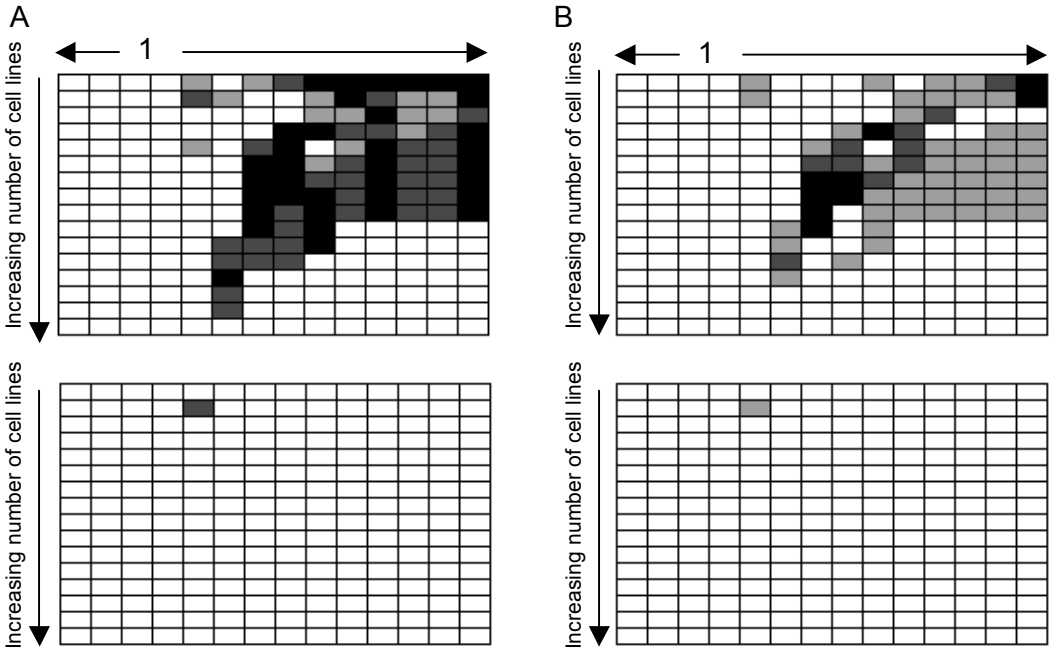

Figure 4.15. Over-representation of human orthologues of genes nearest to CISs (above) and genes further from CISs (below) in amplicons and deletions, where CISs have a *P*-value of <0.001 (A) and <0.05 (B). Each box represents the significance of the association between the selected genes and amplicons/deletions 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, with (from left to right) copy number thresholds of less than 0.2, 0.6 and 0.9. Copy number thresholds above 1 represent amplicons, with (from left to right) copy number thresholds of greater than 1.1, 1.6, 2.1, 2.5, 2.7, 3.1, 3.8, 4.3, 4.8, 5.2 and 5.9. The number of cell lines increases in increments of 1, up to a cut-off of 16 cell lines.

(*1600014C10Rik* and *Slamf6*, respectively) was also amplified, and analysis of the insertions around these genes suggests that the nearest gene is more likely to be disrupted by MuLV (Figure 4.16). *PLEKHF1* and *LTAP* may therefore be non-tumourigenic passengers within the amplified regions. *Tpcn2* and *Ccnd1* are neighbouring genes that both have nearby CISs, and both human orthologues were amplified, suggesting that both may be involved in tumourigenesis. For the remaining 9 genes nearest to CISs that were amplified to a copy number greater than or equal to 5.2, a human orthologue could not be found for the further gene. This explains why the lists of human orthologues of nearest genes are longer than the lists of human orthologues of further genes (see Table 4.1). Therefore, in some cases, the further gene may have an unidentified human orthologue that is also amplified in cancer. However, the fact that the nearest gene list contains a higher proportion of genes with human orthologues is itself significant, since cancer and cancer-related functions, such as cell growth, are well-studied and implicated genes may therefore be more likely to be characterised than genes with other functions, and such genes are also likely to be conserved between species.

To investigate whether the difference between comparisons of nearest and further genes is most likely to be due to the further gene not being amplified or not having a human orthologue, the 66 human orthologues nearest to CISs that were amplified to a copy number greater than or equal to 2.7 were analysed in greater detail. 10 of the amplified genes, including *TPCN2* and *CCND1*, were neighbouring genes for which both mouse orthologues had nearby CISs. For 27 genes, the human orthologue of the further gene could not be identified. For a further 27 genes, the further gene was also amplified and, in all cases, both the nearest and the further genes were amplified in the same number of cell lines. There were only 2 amplified nearest genes, *Slc9a8* and *Mafk*, for which the further gene, *B4galt5* and *1110015K06Rik* respectively, had a human orthologue that was not amplified, suggesting that the nearest genes are the likely candidate cancer genes. In both cases, the human and mouse regions containing these genes are syntenic, and thus the lack of amplification is not due to a break in synteny in the human genome.

The reciprocal analysis was also performed, whereby the 36 human orthologues further from CISs that were amplified to copy number 2.7 or above were also investigated. 29 had neighbouring, nearer genes that were also amplified. This number is higher than the reciprocal count of 27 genes because 2 of the genes were adjacent to nearer genes that had more than one adjacent gene because they contained multiple CISs. For the 7 remaining

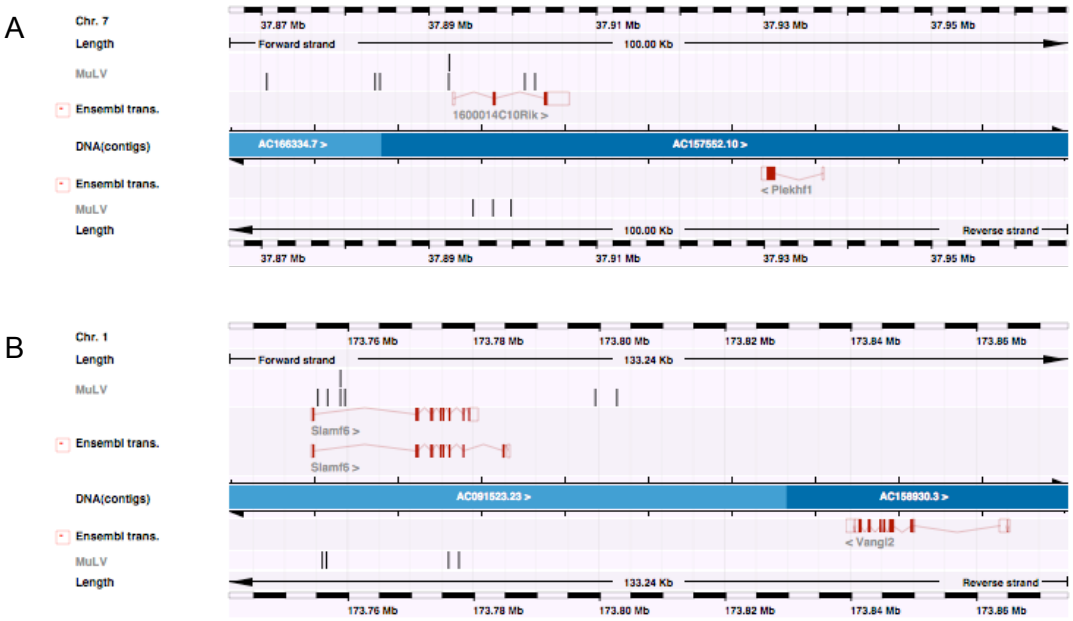

Figure 4.16. Insertions appear to be associated with the gene nearest to the CIS, i.e. *1600014C10Rik* (A) and *Slamf6* (B), even though adjacent genes are also amplified. Insertions are shown as black vertical lines. Those above the blue bar labelled DNA(contigs) are in the sense orientation, those below are in the antisense orientation. Ensembl genes are shown in red.

genes, there was no human orthologue for the nearest gene to the CIS. Therefore, it appears that the stronger overlap between the human amplicons and the human orthologues of genes nearest to CISs mainly reflects the inability to identify human orthologues for a higher percentage of the genes further from CISs. Since amplicons are generally large and encompass many genes (see Section 4.4), this is the more sensible explanation. However, several genes nearest to CISs were amplified in the absence of the further gene, while there were no examples of the reciprocal association. Choosing the nearest gene is the simplest method for assigning insertions to genes and will correctly identify genes that contain intragenic CISs. There is no evidence to suggest that genes nearest to CISs are preferentially amplified, but the fact that they are more likely to have a human orthologue does indicate that they may be the more likely candidates for a role in cancer. However, a more sophisticated method for assigning insertions to genes, such as the approach described in Sections 2.9 and 2.10, is likely to yield the most reliable list of cancer gene candidates. Comparative analyses involving these genes are discussed in Chapter 5.

4.5.2 Identification of individual candidates for a role in human cancer

For each gene list, the amplicons and deletions containing the human orthologues of candidate cancer genes were analysed in more detail to find the most promising candidates for a role in human cancer. The minimal amplified/deleted region was calculated by taking all of the amplicons/deletions in which the gene resided and finding the most 3' start coordinate and the most 5' end coordinate. The identities of other genes within the minimal region were established using the coordinates of human genes in Ensembl v37. For each gene within a minimum amplified or deleted region, the total number of cell lines in which that gene was amplified or deleted was calculated. In order to filter out the less likely candidates, genes in amplicons were discarded if they were co-amplified with known oncogenes or other mouse candidates from the gene list. Likewise, genes in deletions were discarded if they were co-deleted with known tumour suppressor genes or other mouse candidates from the gene list.

4.5.2.1 Candidate oncogenes among genes nearest to CISs

4.5.2.1.1 Protein-coding genes

The strongest candidates from the list of genes nearest to CISs with $P<0.001$ are shown in Table 4.4. Among 242 genes amplified to copy number 1.6 or above, 60 co-occurred with 1 or more known oncogenes and 128 co-occurred with other candidates in the list, of which 105 co-occurred with genes that were amplified in a greater number of cell lines. The filtered list of 54 candidates contained 14 known oncogenes, including *EVI1* and *FGFR2*, for which the murine insertions and human amplicons of less than 70 Mb are shown in Figure 4.17. 70 Mb is an arbitrary cut-off, but omits amplicons that are very large and for which there is therefore a low degree of certainty that the CIS genes are the targets of amplification. The kinase insert domain protein receptor gene (*KDR*) was amplified in 5 cell lines. Analysis of the insertions around *Kdr* in the mouse suggests that the adjacent gene, known oncogene *Kit*, may in fact be disrupted by the insertions assigned to *Kdr* (see Figure 2.11B, page 94). Likewise, the minimal amplified region containing *KDR* also contained *KIT*, which was amplified in an additional cell line (Figure 4.17) and is therefore the more likely target of amplification.

Further implicated oncogenes were also identified (Table 4.4). For example, the homeobox gene *MEIS1* is implicated in neuroblastoma. It was found to be amplified in the neuroblastoma cell line IMR-32 and was overexpressed in further neuroblastoma cell lines (Jones *et al.*, 2000). The single cell line in which it was amplified (to copy number 9.8) in this analysis was the neuroblastoma cell line GI-LI-N, which, according to the cell line typing analysis of the Cancer Genome Project (see Section 4.2.2), shares 96.0% identity with IMR-32, suggesting that they are derived from the same cancer. Even genes that are rarely amplified may therefore contribute to tumourigenesis. Likewise, the NF-κB transcription factor family member *NFKB1* was amplified to copy number 4.8 in one cell line (HH) derived from an adult T-cell lymphoma-leukaemia. Polymorphisms of *NFKB1* are associated with susceptibility to a number of cancers, including oral squamous cell carcinoma, myeloma, and cancers of the colon, liver and breast (for review, see Sun and Zhang, 2007). Interestingly, *NFKB1* maps to a region that is involved in translocations in certain types of acute lymphoblastic leukaemia (Liptay *et al.*, 1992).

Other implicated oncogenes that were amplified in human cancer and disrupted by retroviral insertions include matrix metalloproteinase-13 (*MMP13*) and mothers against decapentaplegic homolog 7 (*SMAD7*). *MMP13* shows recurrent amplification and overexpression in cervical cancer (Narayan *et al.*, 2007) and 2 of the 12 cell lines in this

CIS *P*-value	Gene name	Mouse Ensembl ID	Human Ensembl ID	Number of cell lines	Genes in minimal amplified region	Maximum copy number	Known oncogene?
0.001	*Myc*	ENSMUSG00000022346	ENSG00000136997	71	3	5.9+	Y
0.001	*Ccnd1*	ENSMUSG00000031071	ENSG00000110092	24	10	5.9+	Y
0.001	*Nmyc1*	ENSMUSG00000037169	ENSG00000134323	14	9	5.9+	Y
0.001	*Slamf6*	ENSMUSG00000015314	ENSG00000162739	14	21	5.9+	
0.001	*Smad7*	ENSMUSG00000025880	ENSG00000101665	6	27	5.9+	
0.001	*Fgfr2*	ENSMUSG00000030849	ENSG00000066468	5	7	5.9+	Y
0.001	*Kdr*	ENSMUSG00000062960	ENSG00000128052	5	15	5.9+	
0.001	*Tnfrsf7*	ENSMUSG00000030336	ENSG00000139193	5	26	5.9+	
0.001	*Meis1*	ENSMUSG00000020160	ENSG00000143995	1	2	5.9+	
0.001	*Mmp13*	ENSMUSG00000050578	ENSG00000137745	12	21	4.8	
0.001	*Nfkb1*	ENSMUSG00000028163	ENSG00000109320	1	7	4.8	
0.001	*Zfp217*	ENSMUSG00000052056	ENSG00000171940	43	15	4.3	
0.001	*Zfpn1a1*	ENSMUSG00000018654	ENSG00000185811	18	103	4.3	Y
0.001	*Ccnd3*	ENSMUSG00000034165	ENSG00000112576	8	37	4.3	Y
0.001	*Lmo2*	ENSMUSG00000032698	ENSG00000135363	4	43	4.3	Y
0.001	*Pim1*	ENSMUSG00000024014	ENSG00000137193	7	12	3.8	Y
0.001	*Ccnd2*	ENSMUSG00000000184	ENSG00000118971	6	15	3.8	Y
0.001	*Evi1*	ENSMUSG00000027684	ENSG00000085276	17	27	3.1	Y
0.001	*Btg2*	ENSMUSG00000020423	ENSG00000159388	10	35	3.1	
0.001	*Cd72*	ENSMUSG00000028459	ENSG00000137101	8	26	3.1	
0.001	*Rreb1*	ENSMUSG00000039087	ENSG00000124782	7	10	3.1	
0.001	*Aarsl*	ENSMUSG00000023938	ENSG00000124608	6	19	3.1	
0.001	*Taok3*	ENSMUSG00000061288	ENSG00000135090	3	9	3.1	
0.001	*Ntn1*	ENSMUSG00000020902	ENSG00000065320	2	31	3.1	
0.001	*Pik3r5*	ENSMUSG00000020901	ENSG00000141506	2	31	3.1	
0.001	*Eif4e3*	ENSMUSG00000030068	ENSG00000163412	2	33	3.1	
0.001	*Irf4*	ENSMUSG00000021356	ENSG00000137265	7	27	2.7	Y
0.001	*Ubb*	ENSMUSG00000019505	ENSG00000170315	5	72	2.5	
0.001	*Cd69*	ENSMUSG00000030156	ENSG00000110848	4	52	2.5	
0.001	*Lrrc5*	ENSMUSG00000046079	ENSG00000171492	2	25	2.5	
0.001	*Ptpn1*	ENSMUSG00000027540	ENSG00000196396	42	27	2.1	
0.001	*Sla2*	ENSMUSG00000027636	ENSG00000101082	41	84	2.1	
0.001	*E030003N15Rik*	ENSMUSG00000036661	ENSG00000105339	40	68	2.1	
0.001	*2310007D09Rik*	ENSMUSG00000027654	ENSG00000101447	38	61	2.1	
0.001	*Capsl*	ENSMUSG00000039676	ENSG00000152611	32	40	2.1	
0.001	*Cldn10*	ENSMUSG00000022132	ENSG00000134873	20	37	2.1	
0.001	*Ebi2*	ENSMUSG00000051212	ENSG00000169508	18	31	2.1	
0.001	*Flt3*	ENSMUSG00000042817	ENSG00000122025	11	122	2.1	Y
0.001	*Chc1l*	ENSMUSG00000022106	ENSG00000136161	10	53	2.1	
0.001	*Lcp1*	ENSMUSG00000021998	ENSG00000136167	10	128	2.1	Y
0.001	*4933403F05Rik*	ENSMUSG00000038121	ENSG00000177150	10	159	2.1	
0.001	*Dtl*	ENSMUSG00000037474	ENSG00000143476	8	70	2.1	
0.001	*2410129E14Rik*	ENSMUSG00000045136	ENSG00000137285	7	18	2.1	
0.001	*1110036O03Rik*	ENSMUSG00000006931	ENSG00000141696	6	43	2.1	
0.001	*Fmnl1*	ENSMUSG00000055805	ENSG00000184922	6	85	2.1	
0.001	*Ksr*	ENSMUSG00000018334	ENSG00000141068	5	34	2.1	
0.001	*Jundm2*	ENSMUSG00000034271	ENSG00000140044	13	42	1.6	
0.001	*Tomm20*	ENSMUSG00000058779	ENSG00000173726	8	252	1.6	
0.001	*Cyb5*	ENSMUSG00000024646	ENSG00000166347	6	20	1.6	
0.001	*Ldhd*	ENSMUSG00000031958	ENSG00000166816	6	74	1.6	
0.001	*Cbfa2t3h*	ENSMUSG00000006362	ENSG00000129993	5	133	1.6	Y
0.001	*Zfp608*	ENSMUSG00000052713	ENSG00000168916	3	30	1.6	
0.001	*2610307O08Rik*	ENSMUSG00000024349	ENSG00000184584	3	95	1.6	
0.001	*Hhex-rs2*	ENSMUSG00000024986	ENSG00000152804	2	40	1.6	
0.05	*D930036F22Rik*	ENSMUSG00000035181	ENSG00000129493	17	19	5.9+	
0.05	*Laptm5*	ENSMUSG00000028581	ENSG00000162511	1	11	2.7	
0.05	*Emp3*	ENSMUSG00000040212	ENSG00000142227	7	33	2.7	
0.05	*Rai1*	ENSMUSG00000062115	ENSG00000108557	5	72	2.1	

Table 4.4. Genes that are nearest to CISs in mouse lymphomas and are also promising candidates for targets of amplification in human cancer cell lines. "CIS *P*-value" is the minimum threshold for the significance of the CIS nearest to the given gene. "Number of cell lines" is the number of samples in which the gene is amplified to a copy number of greater than or equal to 1.6. "Genes in minimal amplified region" is the number of genes that co-occur with the CIS gene in the smallest region of amplification. "Maximum copy number" is the maximum copy number threshold above which the gene is identified as being amplified. "Known oncogene?" indicates whether the gene is a dominant cancer gene listed in the Cancer Gene Census.

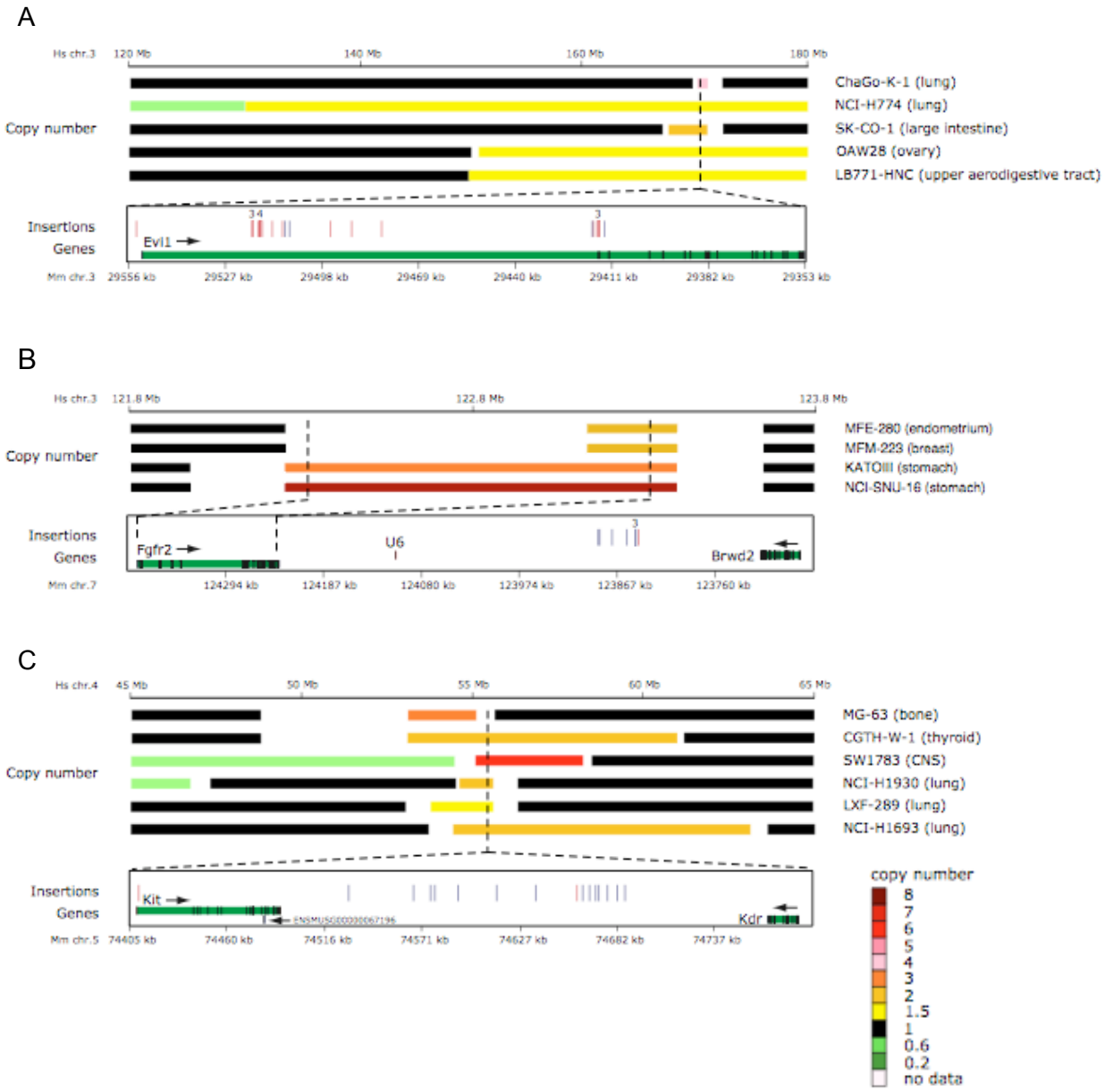

Figure 4.17. Known human oncogenes *EVI1* (A), *FGFR2* (B) and *KIT* (C) are amplified in human cancer cell lines and are disrupted by retroviral insertional mutagenesis in mouse lymphomas. The copy number of chromosomal regions in the human cell lines is depicted in colour. Names of human cell lines and tissue of origin are provided Only cell lines in which the amplicon containing the oncogene is less than 70 Mb are shown. The lower part of each figure shows insertions within mouse tumours, and was kindly provided by Jaap Kool and Jeroen de Ridder. Blue vertical lines represent insertions in the sense orientation, while red vertical lines represent antisense insertions. Genes are shown in green, with exons marked in black. Positions on the murine and human chromosomes are indicated on the black horizontal bars in kb and Mb, respectively. These figures can also be seen in Uren *et al*. (2008).

study in which *MMP13* was amplified were indeed derived from cervical cancers. *MMP13* has not been shown to be amplified in any other cancer types, but overexpression has been observed, e.g. in squamous cell carcinomas of the head and neck (Johansson *et al.*, 1997) and vulva (Johansson *et al.*, 1999). The results of this analysis suggest that *MMP13* is amplified in, and implicated in, a range of cancer types. The cell lines containing amplicons of less than 70 Mb that encompass *MMP13* are shown in Figure 4.18. Among these types are oesophageal, skin and breast cancers, in which *MMP13* overexpression has been observed (Freije *et al.*, 1994; Hu *et al.*, 2001; Kuivanen *et al.*, 2006). The minimal amplified region on chromosome 11 contains 21 genes, including a cluster of genes encoding matrix metalloproteinases, of which a number have been previously implicated in cancer. However, *MMP13* was the only gene disrupted by insertional mutagenesis.

SMAD7 duplication has been demonstrated in colorectal cancer (Boulay *et al.*, 2001) and the gene is overexpressed in a number of cancer types, including basal cell carcinoma (Gambichler *et al.*, 2007), endometrial cancer (Dowdy *et al.*, 2005) and thyroid follicular carcinoma cell lines (Cerutti *et al.*, 2003). The highest amplification of *SMAD7* was in the retinoblastoma cell line Y79. Interestingly, SMAD7 has been shown to suppress TGF-β1-mediated growth inhibition in pancreatic cancer cells through the inactivation of the retinoblastoma protein (Boyer Arnold and Korc, 2005) and it inhibits growth arrest and apoptosis in mouse B cells through the inactivation of retinoblastoma (Ishisaki *et al.*, 1998; Nakahara *et al.*, 2003). In addition, *SMAD7* is expressed in the eye, and suppresses TGF-β2-mediated inhibition of corneal endothelial cell proliferation, resulting in accelerated wound healing (Funaki *et al.*, 2003). *SMAD7* is therefore a promising target for amplification in the retinoblastoma cell line. Likewise, one of the amplicons encompassing *SMAD7* was identified in a Ewing's sarcoma cell line (EW-24) and, in osteogenesis, SMAD7 suppresses osteoblast differentiation and bone formation (Koinuma and Imamura, 2005) and inhibits Saos2 osteosarcoma cell differentiation (Eliseev *et al.*, 2006). *SMAD7* was also amplified in 2 haematopoietic and 2 lung cancer cell lines. SMAD7 promotes self-renewal of haematopoietic stem cells (Blank *et al.*, 2006) and is highly expressed in metastatic lung cancer cell lines (Shen *et al.*, 2003).

Other interesting candidates include SLAM family member 6 precursor (*SLAMF6*), serine/threonine-protein kinase TAO3 (*TAOK3*), RAS-responsive element-binding protein 1 (*RREB1*) and leucine-rich repeat-containing protein 8D (*LRRC5*). The minimal

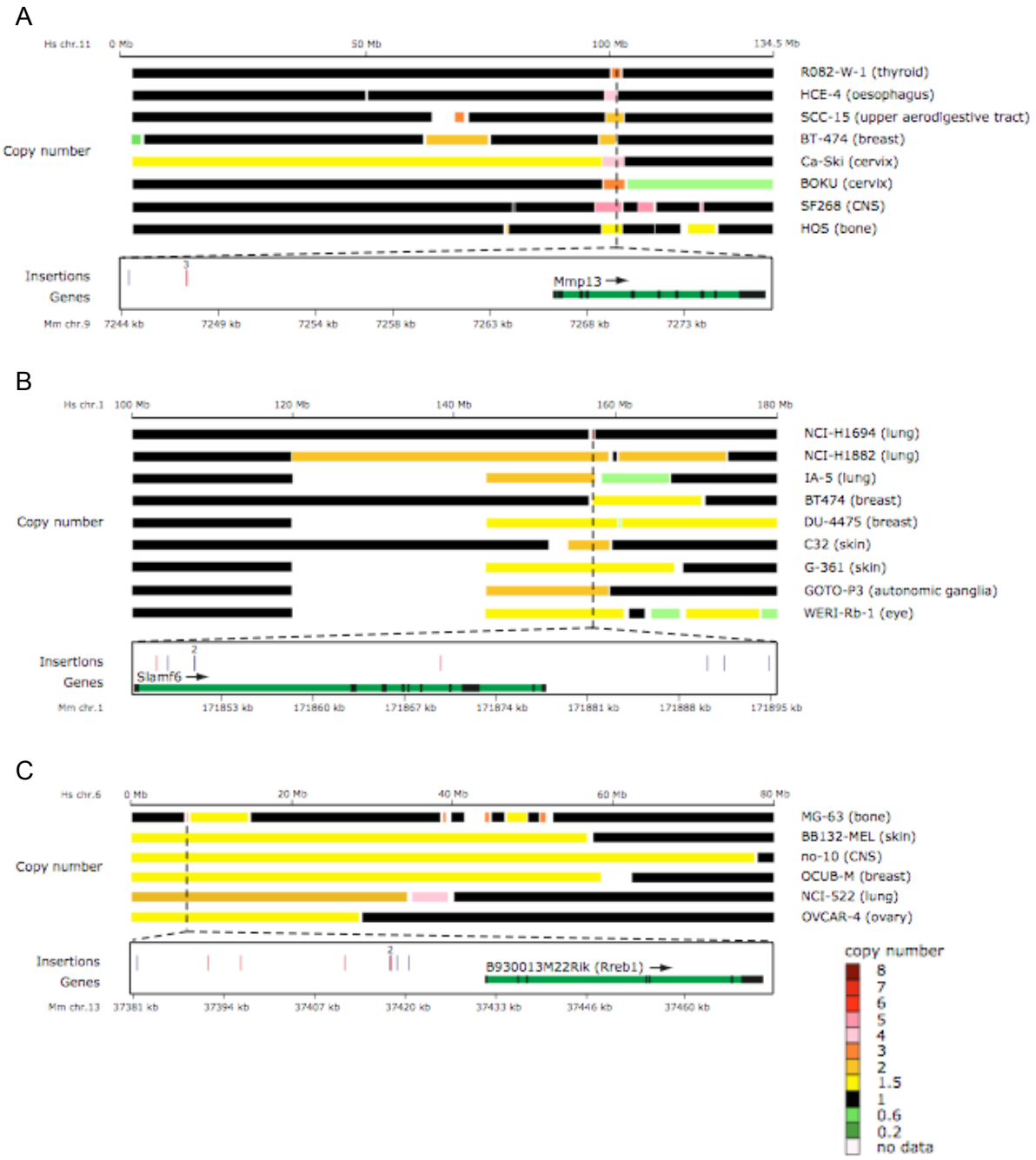

Figure 4.18. Candidate oncogenes *MMP13* (A), *SLAMF6* (B) and *RREB1* (C) are amplified in human cancer cell lines and are disrupted by retroviral insertional mutagenesis in mouse lymphomas. The copy number of chromosomal regions in the human cell lines is depicted in colour. Names of human cell lines and tissue of origin are provided Only cell lines in which the amplicon containing the oncogene is less than 70 Mb are shown. The lower part of each figure shows insertions within mouse tumours, and was kindly provided by Jaap Kool and Jeroen de Ridder. Blue vertical lines represent insertions in the sense orientation, while red vertical lines represent antisense insertions. Genes are shown in green, with exons marked in black. Positions on the murine and human chromosomes are indicated on the black horizontal bars in kb and Mb, respectively. These figures can also be seen in Uren *et al*. (2008).

amplified region encompassing *SLAMF6* comprised 21 genes and was recurrent across 14 cell lines. Cell lines containing an amplicon of less than 70 Mb are shown in Figure 4.18. The highest amplification of *SLAMF6* was within the lung cancer cell line NCI-H1694. However, it has been proposed that *SLAMF6*, also known as *Ly108*, is only expressed in lymphoid tissues (Peck and Ruley, 2000), where it regulates T cell development (Jordan *et al.*, 2007) and B cell tolerance (Kumar *et al.*, 2006). Polymorphisms within the gene are associated with systemic lupus erythematosus (Wandstrat *et al.*, 2004), which has been widely associated with an increased risk of developing a range of cancers, but most strongly with cancers arising from B lymphocytes (Bernatsky *et al.*, 2007). *TAOK3* was only amplified in 3 cell lines, but the minimum region contained just 9 genes. *TAOK3* is poorly characterised, but contains a somatic missense mutation in 2 lung cancers (small cell carcinoma cell line NCI-H28 and a primary adenocarcinoma) in the COSMIC database (Forbes *et al.*, 2006). Although a role in tumourigenesis has not been demonstrated for *TAOK3*, protein kinases are widely implicated in cancer (see Sections 1.2.5.2 and 1.3.1). *RREB1* was amplified in 7 cell lines within a minimal region of 10 genes. Cell lines containing an amplicon of less than 70 Mb are shown in Figure 4.18. Each cell line was derived from a different tissue, but *RREB1* has been shown to be ubiquitously expressed in human tissues apart from the adult brain (Thiagalingam *et al.*, 1997). Rreb1 binds to, and represses expression of, the $p16^{Ink4a}$ promoter, and the development of pristine-induced plasma cell tumours in Balb/C mice is attributable to a polymorphism in this Rreb1 binding site (Zhang *et al.*, 2003). In addition, *RREB1* is important in reducing cell-cell adhesion and collective migration of epithelial cells (Melani *et al.*, 2008), and it may therefore play a role in metastasis. RREB1 has also been identified as a transcriptional effector of RAL (Oxford *et al.*, 2007), and RALA is itself implicated in cancer cell migration, as well as other cancer-related functions (Oxford *et al.*, 2005). The most amplified occurrence of *RREB1* was in the osteosarcoma cell line MG-63, but no role for *RREB1* has previously been elucidated in bone tissue. *LRRC5* was amplified in just 2 cell lines, with a minimal amplified region of 25 genes. Although little is known about this gene, it is thought that it might be implicated in the proliferation and activation of lymphocytes and monocytes, suggesting a possible role in the oncogenesis of B cells which would account for the insertions disrupting *Lrrc5* in mouse lymphomas. However, *LRRC5* was amplified in human cancer cell lines derived from the ovary and upper aerodigestive tract.

Only 4 additional candidates (*D930036F22Rik*, *LAPTM5*, *EMP3* and *RAI1*) were identified using genes nearest to CISs with $P<0.05$ (see Table 4.4). *D930036F22Rik* is also known as HEAT repeat containing 5A (*HEATR5A*). The minimal amplified region also included Rho-GTPase-activating protein 5 (*p190-B*), which is known to be overexpressed in breast cancer (Chakravarty *et al.*, 2000), although only 1 breast cancer cell line contained an amplification of this region. Based on the distribution of insertions in the CIS, it is entirely possible that nearby genes *Hectd1* and/or *EG544864* were in fact the targets of MuLV mutagenesis (Figure 4.19). However, none of these genes have been previously implicated in tumourigenesis. Lysosomal-associated protein transmembrane 5 (*LAPTM5*) was amplified in a single cell line derived from an endometrial carcinoma (MFE-280). *LAPTM5* is inactivated by chromosomal rearrangement and DNA methylation in human multiple myeloma (Hayami *et al.*, 2003) but is overexpressed in malignant B lymphomas (Seimiya *et al.*, 2003) and is a predictor for early intrahepatic recurrence of hepatocellular carcinoma (Somura *et al.*, 2008). However, the amplified region in MFE-280 also contained the syndecan-3 gene, which is expressed in the human endometrium (Germeyer *et al.*, 2007) and is thought to play a role in uterine growth (Russo *et al.*, 2001). Epithelial membrane protein gene *EMP3* was proposed as a candidate tumour suppressor in glioma and neuroblastoma (Alaminos *et al.*, 2005), but it has since been shown to be overexpressed in oligodendroglial tumours (Li *et al.*, 2007a) and primary glioblastomas (Kunitz *et al.*, 2007). It is also overexpressed in invasive human mammary carcinoma cell lines (Evtimova *et al.*, 2003) and contains a polymorphism in prostate cancers (Burmester *et al.*, 2004). Retinoic acid induced 1 gene (*RAI1*) has not been previously implicated in cancer.

4.5.2.1.2 miRNA genes

The list of genes nearest to CISs with $P<0.001$ contained 6 miRNA genes, while the list nearest to CISs with $P<0.05$ contained 9. As mentioned in Section 3.2.2, deregulated miRNAs are implicated in promoting and suppressing tumourigenesis in a range of tissues. Currently, only protein-coding genes have human orthologues in Ensembl, and miRNA genes were therefore omitted from the global comparison and principal analysis of CIS genes within amplicons. However, it is possible to manually identify the human equivalents based on the miRNA name and the conserved synteny between the mouse and human genomes. Table 4.5 shows the name of the murine miRNA and the corresponding human miRNA for genes nearest to CISs, as well as lists of the miRNA genes within

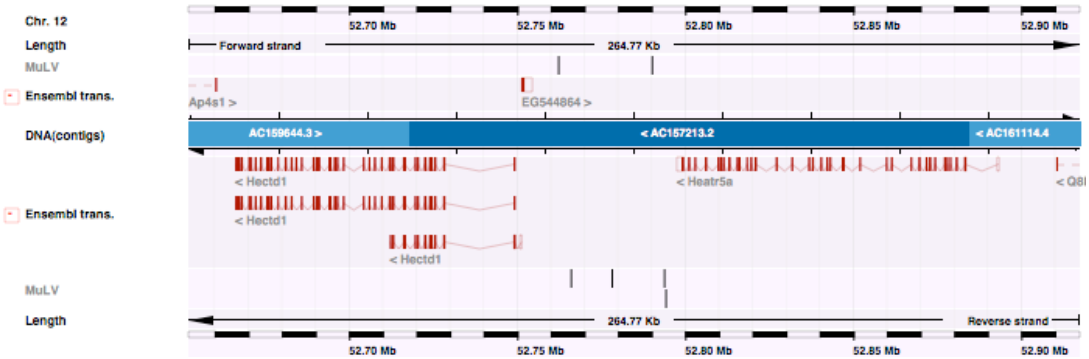

Figure 4.19. Insertions assigned to *Heatr5a* may be associated with *Hectd1* or *EG544864*. Insertions are shown as black vertical lines. Those above the blue bar labelled DNA(contigs) are in the sense orientation, those below are in the antisense orientation. Ensembl genes are shown in red.

A

P-value	Mouse miRNA	Human miRNA
0.001	rno-mir-128b	hsa-mir-128b
0.001	hsa-mir-142	hsa-mir-142
0.001	mmu-mir-21	hsa-mir-21
0.001	mmu-mir-23a	hsa-mir-23a
0.001	mmu-mir-17	hsa-mir-17
0.001	hsa-mir-106a	hsa-mir-106a
0.05	mmu-mir-26b	hsa-mir-26b
0.05	mmu-mir-22	hsa-mir-22
0.05	rno-mir-200b	hsa-mir-200b

B

Mouse miRNA gene	Mouse Ensembl ID	Human miRNA gene	Human Ensembl ID	Number of cell lines	Maximum copy number	Known oncogenes in minimal region?
mmu-mir-17	ENSMUSG00000065508	*hsa-mir-17*	ENSG00000198999	23	3.1	
hsa-mir-142	ENSMUSG00000065420	*hsa-mir-142*	ENSG00000199166	9	2.1	Y
mmu-mir-21	ENSMUSG00000065455	*hsa-mir-21*	ENSG00000199004	9	2.1	Y
mmu-mir-23a	ENSMUSG00000065611	*hsa-mir-23a*	ENSG00000199028	8	2.1	Y
rno-mir-128b	ENSMUSG00000065441	*hsa-mir-128b*	ENSG00000199105	1	2.1	

C

Mouse miRNA gene	Mouse Ensembl ID	Human miRNA gene	Human Ensembl ID	Number of cell lines	Minimum copy number
rno-mir-128b	ENSMUSG00000065441	*hsa-mir-128b*	ENSG00000199105	40(2)	0.2
mmu-mir-17	ENSMUSG00000065508	*hsa-mir-17*	ENSG00000198999	61	0.6
mmu-mir-22	ENSMUSG00000065529	*hsa-mir-22*	ENSG00000199060	17	0.6
mmu-mir-26b	ENSMUSG00000065468	*hsa-mir-26b*	ENSG00000199121	8	0.6
mmu-mir-21	ENSMUSG00000065455	*hsa-mir-21*	ENSG00000199004	6	0.6
hsa-mir-142	ENSMUSG00000065420	*hsa-mir-142*	ENSG00000199166	5	0.6
mmu-mir-23a	ENSMUSG00000065611	*hsa-mir-23a*	ENSG00000199028	3	0.6

Table 4.5. miRNA genes that are nearest to CISs in mouse lymphomas and are amplified and/or deleted in human cancer cell lines. (A) Names of the murine miRNAs and their human orthologues. (B) Amplified miRNA genes. (C) Deleted miRNA genes. "*P*-value" is the minimum threshold for the significance of the CIS nearest to the given gene. "Number of cell lines" is the number of samples in which the gene is amplified to a copy number of greater than or equal to 1.6 (B) or deleted to a copy number of less than or equal to 0.6, with the number for deletions of copy number 0.2 or below shown in brackets (C). "Maximum copy number" is the maximum copy number threshold above which the gene is identified as being amplified. "Minimum copy number" is the minimum copy number threshold below which the gene is identified as being deleted.

human amplicons and deletions. Genes encoding 5 of the miRNAs were amplified in human cancer cell lines. The minimal amplified regions for *hsa-miR-142* and *hsa-miR-23a* were very large and encompassed 4 and 3 known oncogenes, respectively, while *hsa-miR-21* was co-amplified with *hsa-miR-23a*, and *hsa-miR-128b* was amplified in just 1 cell line. The minimal amplified region of *hsa-miR-17* contained 14 genes, of which none were oncogenes and only 2 had a description in Ensembl. hsa-miR-17 is part of the miR-17-92 cluster of 6 miRNAs. All 3 of the cell lines in which *hsa-miR-17* was amplified to copy number 2.5 or above were derived from haematopoietic and lymphoid cancers, which is consistent with a role for miR-17-92 in both B-lymphocyte development and B-lymphoproliferative disorders (Garzon and Croce, 2008). The cluster is also overexpressed in other human cancers, including colorectal cancers (Monzo *et al.*, 2008), anaplastic thyroid cancer cells (Takakura *et al.*, 2008), neuroblastomas with *MYCN* amplification (Schulte *et al.*, 2008), bladder cancers (Gottardo *et al.*, 2007) and lung cancers (Hayashita *et al.*, 2005). Of the 23 cell lines in which *hsa-mir-17* was amplified to copy number 1.6 or above, 7 were from colon cancers, 7 were from haematopoietic and lymphoid cancers, 4 were from lung cancers, and 1 each were from cancers of the stomach, soft tissue, central nervous system, breast and eye. miR-17-92 is the likely target for amplification within this region, and demonstrates that the function of miRNAs may be conserved between species.

7 miRNA genes were identified within deletions of copy number 0.6 or below, but only *hsa-miR-17* and *hsa-miR-128b* were deleted in a large number of cell lines, and only *hsa-miR-128b* was within homozygous deletions. *hsa-miR-128b* was shown to be downregulated in classic Hodgkin lymphomas infected with Epstein-Barr virus (Navarro *et al.*, 2008). However, it is also upregulated in acute lymphoblastic leukaemia (Zanette *et al.*, 2007). The two homozygous deletions of *hsa-miR-128b* were within glioma and neuroblastoma cell lines, respectively. Interestingly, *hsa-miR-128* is highly expressed in the adult brain and preferentially in neurons, where it is thought to play a role in neural differentiation (Smirnova *et al.*, 2005). Downregulation of *hsa-miR-128* has been previously demonstrated in glioblastoma (Ciafre *et al.*, 2005), but deletion of the gene has not been previously demonstrated.

4.5.2.2 Candidate cancer genes among genes containing insertions within the coding region

Analyses involving the remaining lists were primarily designed to identify tumour suppressor genes, but as shown in Section 4.5.1.3, it is likely that the lists are contaminated with candidate oncogenes. The list for which the pattern of distribution was most similar to that of oncogenes was the list of genes containing insertions, not including those represented by a single read, within the coding region. The strongest candidates in regions of copy number gain, identified using the same filtering procedure as used for genes nearest to CISs, included 4 oncogenes (Table 4.6). As discussed in Section 3.4.1, insertions within the 3' UTR of *Mycn* and *Pim1* result in the formation of a more stable protein product, rather than gene inactivation. Some of the insertions were within the last exon of these genes, which explains their inclusion within the current list.

Candidate tumour suppressor genes were identified within regions of copy number loss. Among these was the gene encoding transmembrane protease, serine 2 precursor (*TMPRSS2*), which is a known oncogene that forms fusions with the ETS transcription factor genes *ERG* and *ETV1* in prostate cancer (Tomlins *et al.*, 2005) and is overexpressed in most prostate cancers (Vaarala *et al.*, 2001). A hemizygous microdeletion within the fusion has been observed on chromosome 21 between *ERG* and *TMPRSS2* (Yoshimoto *et al.*, 2006), but this does not explain the deletion of the entire gene, sometimes in both copies. In addition, the homozygous deletions, which were also the most focal deletions, were identified in cancer cell lines derived from the pancreas and upper aerodigestive tract, rather than the prostate. Most of the heterozygous deletions were very large, encompassing many genes, and the minimal deleted region contained 18 genes. Based on the known role of *TMPRSS2*, it seems unlikely that this is the target of deletion within this region. Likewise, although deleted in human cancers, *MAP3K8* and *IL6RA* are also more likely to act as oncogenes. *MAP3K8* is overexpressed in, for example, invasive endometrioid cancer (Aparecida Alves *et al.*, 2006), T-cell neoplasias (Christoforidou *et al.*, 2004) and breast cancer (Sourvinos *et al.*, 1999), while expression of the interleukin 6 receptor gene *IL6RA* is promoted by Epstein-Barr virus in immortalised B cells and Burkitt's lymphoma cells (Klein *et al.*, 1995). The minimal amplified region containing the gene encoding MAGUK p55 subfamily member 4 (MMP4) comprised just 4 genes. Interestingly, *MMP4* is a homologue of the Drosophila *Stardust* gene, which is involved in establishing and maintaining epithelial tissue polarity,

Gene name	Mouse Ensembl ID	Human Ensembl ID	Number of cell lines	Genes in minimal region	Copy number	Singletons only?	Known oncogene?
Capsl	ENSMUSG00000039676	ENSG00000152611	32	40	1.6+		
Bcl9	ENSMUSG00000038256	ENSG00000116128	17	62	1.6+		Y
Mycn	ENSMUSG00000037169	ENSG00000134323	14	9	1.6+		Y
Ccnd3	ENSMUSG00000034165	ENSG00000112576	8	37	1.6+		Y
Pim1	ENSMUSG00000024014	ENSG00000137193	7	12	1.6+		Y
NM_009283.2	ENSMUSG00000026104	ENSG00000115415	2	12	1.6+		
Mrps18b	ENSMUSG00000024436	ENSG00000137330	12	19	0.6		
Mpp4	ENSMUSG00000026024	ENSG00000003393	9	4	0.6		
Il6ra	ENSMUSG00000027947	ENSG00000160712	3	23	0.6		
Phgdhl1	ENSMUSG00000041765	ENSG00000134882	63(1)	9(21)	0.2		
Map3k8	ENSMUSG00000024235	ENSG00000107968	35(1)	13(20)	0.2		
Tmprss2	ENSMUSG00000000385	ENSG00000184012	14(2)	18(18)	0.2		Y
Tmem16f	ENSMUSG00000064210	ENSG00000177119	3	1	1.6+	Y	
Nfkb1	ENSMUSG00000028163	ENSG00000109320	1	7	1.6+	Y	
9030611O19Rik	ENSMUSG00000036136	ENSG00000184731	3	16	1.6+	Y	
Olfr1509	ENSMUSG00000035626	ENSG00000182735	12	26	1.6+	Y	
	ENSMUSG00000046186	ENSG00000156535	3	35	1.6+	Y	
Rasgrp4	ENSMUSG00000030589	ENSG00000171777	12	56	1.6+	Y	
Dsg1b	ENSMUSG00000061928	ENSG00000134760	76	89	0.6	Y	
XP_484397.2	ENSMUSG00000034731	ENSG00000102780	68	8	0.6	Y	
Riok3	ENSMUSG00000024404	ENSG00000101782	68	47	0.6	Y	
6330406I15Rik	ENSMUSG00000029659	ENSG00000102802	64	9	0.6	Y	
Il17rb	ENSMUSG00000015966	ENSG00000056736	46	14	0.6	Y	
Zmynd11	ENSMUSG00000021156	ENSG00000015171	46	32	0.6	Y	
Hmgb2	ENSMUSG00000054717	ENSG00000164104	43	13	0.6	Y	
Gtse1	ENSMUSG00000022385	ENSG00000075218	41	90	0.6	Y	
1700020C11Rik	ENSMUSG00000004748	ENSG00000100010	36	61	0.6	Y	
Slc37a2	ENSMUSG00000032122	ENSG00000134955	34	70	0.6	Y	
Man1a	ENSMUSG00000003746	ENSG00000111885	33	8	0.6	Y	
Ate1	ENSMUSG00000030850	ENSG00000107669	32	6	0.6	Y	
Nrap	ENSMUSG00000049134	ENSG00000197893	32	9	0.6	Y	
Q91VN2_MOUSE	ENSMUSG00000042293	ENSG00000180425	32	16	0.6	Y	
Snf1lk2	ENSMUSG00000037112	ENSG00000170145	32	40	0.6	Y	
Dnajc9	ENSMUSG00000021811	ENSG00000182180	31	26	0.6	Y	
3110003A17Rik	ENSMUSG00000019855	ENSG00000146386	30	11	0.6	Y	
Shb	ENSMUSG00000044813	ENSG00000107338	29	16	0.6	Y	
Hp1bp3	ENSMUSG00000028759	ENSG00000127483	24	84	0.6	Y	
1200009I06Rik	ENSMUSG00000021280	ENSG00000185215	23	94	0.6	Y	
8430406I07Rik	ENSMUSG00000027424	ENSG00000125871	21	20	0.6	Y	
Wdr5b	ENSMUSG00000034379	ENSG00000196981	17	14	0.6	Y	
Zdhhc23	ENSMUSG00000036304	ENSG00000184307	17	44	0.6	Y	
9030611O19Rik	ENSMUSG00000036136	ENSG00000184731	15	22	0.6	Y	
Gcnt2	ENSMUSG00000021360	ENSG00000111846	14	47	0.6	Y	
Itk	ENSMUSG00000020395	ENSG00000113263	13	31	0.6	Y	
	ENSMUSG00000039153	ENSG00000124813	13	97	0.6	Y	
Dok3	ENSMUSG00000035711	ENSG00000146094	13	183	0.6	Y	
Hivep3	ENSMUSG00000028634	ENSG00000127124	12	2	0.6	Y	
CSDE1_MOUSE	ENSMUSG00000068823	ENSG00000009307	11	25	0.6	Y	
8430438D04Rik	ENSMUSG00000036019	ENSG00000179104	10	4	0.6	Y	
Bcl10	ENSMUSG00000028191	ENSG00000142867	10	33	0.6	Y	
Rgs2	ENSMUSG00000026360	ENSG00000116741	7	16	0.6	Y	
Jmjd4	ENSMUSG00000036819	ENSG00000081692	7	93	0.6	Y	
Tssk6	ENSMUSG00000047654	ENSG00000178093	6	60	0.6	Y	
Tdrd5	ENSMUSG00000060985	ENSG00000162782	5	46	0.6	Y	
Sell	ENSMUSG00000026581	ENSG00000188404	4	51	0.6	Y	
Leprel1	ENSMUSG00000038168	ENSG00000090530	8(1)	29(33)	0.2	Y	
Olfr1509	ENSMUSG00000035626	ENSG00000182735	19(3)	19(27)	0.2	Y	
Dut	ENSMUSG00000027203	ENSG00000128951	14(1)	7(7)	0.2	Y	
3200002M19Rik	ENSMUSG00000030649	ENSG00000110200	11(1)	24(52)	0.2	Y	

Table 4.6. Mouse genes that contain retroviral insertions within the coding region and are also promising candidates for targets of amplification or deletion in human cancer cell lines. "Number of cell lines" is the number of samples in which the gene is amplified or deleted. "Copy number" is the maximum copy number threshold above which the gene is identified as being amplified, or the minimal threshold below which the gene is deleted. Where the copy number is 0.2, the number of cell lines and number of genes in the minimal deleted region are given for deletions of copy number <= 0.6, with numbers for copy number <= 0.2 being shown in brackets. "Genes in minimal region" is the number of genes that co-occur with the CIS gene in the smallest region of amplification/deletion. "Singletons only?" indicates whether the gene contains insertions other than those represented by a single read. "Known oncogene?" indicates whether the gene is a dominant cancer gene listed in the Cancer Gene Census.

which is disrupted in epithelial tumours. Although *MMP4* has not been implicated in cancer, expression of another family member, known as *MMP7*, has been demonstrated in tumours of the uterus and bladder, and in lymphomas (Katoh and Katoh, 2004).

Among the genes that contained insertions represented by a single read, 3 genes (*SHB*, *HIVEP3* and *BCL10*) stood out as potential tumour suppressor genes. Overexpression of the gene encoding the SHB adaptor protein causes increased activity of the pro-apoptotic kinase c-ABL, resulting in reduced tumour growth in PC3 prostate cancer cells (Davoodpour *et al.*, 2007). Therefore, it is possible that deletion of the gene may lead to tumour cell growth and proliferation. The human immunodeficiency virus type 1 enhancer binding protein 3 gene (*Hivep3*, also known as *Krc*), positively regulates transcription of the mouse metastasis-associated gene, *S100A4/mts1* (Hjelmsoe *et al.*, 2000). In addition, *KRC* was proposed as a potential tumour suppressor gene following the development of a teratoma from *KRC*-deficient embryonic stem cells introduced into an animal model (Allen *et al.*, 2002). Finally, the B-cell lymphoma/leukaemia 10 gene (*BCL10*) is a "hotspot" within the commonly deleted region 1p22.3 in mantle cell lymphomas. Interestingly, 5 of the 10 cell lines containing a deletion within this region were derived from tumours of the autonomic ganglia, but no role for *BCL10* has previously been demonstrated in these cancers (Balakrishnan *et al.*, 2006).

4.5.2.3 Candidate tumour suppressor genes among genes containing insertions within the translated or transcribed region

Candidates among the lists of genes containing insertions within the translated or transcribed region are combined in Table 4.7. The gene that was most frequently deleted below the copy number thresholds of both 0.6 and 0.2 was the known tumour suppressor gene *CDKN2A* (also known as the *INK4A/ARF* locus, and described in Section 1.2.6). This demonstrates the efficacy of the analysis, since homozygous and heterozygous deletions of *CDKN2A* are commonly observed in a wide range of cancers. The only other known tumour suppressor gene in the list, according to the Cancer Gene Census, was the gene encoding FAS, which is a member of the TNF receptor superfamily. Binding of the FAS ligand to the FAS receptor results in the formation of the death-inducing complex (DISC), which triggers apoptosis (for review, see Wajant, 2002). The implicated tumour suppressor gene *WWOX* was also frequently deleted. *WWOX* resides in a fragile site and therefore while it is frequently deleted in cancers, it is unclear whether it contributes to

Gene name	Mouse Ensembl ID	Human Ensembl ID	Number of cell lines	Genes in minimal region	Copy number	Insertions in translated region?	Singletons only?	Known TSG?
Cdkn2a	ENSMUSG00000044303	ENSG00000147889	207(145)	1(3)	0.2	Y		Y
Nfatc1	ENSMUSG00000033016	ENSG00000131196	112(2)	15(15)	0.2	Y		
Zfp532	ENSMUSG00000042439	ENSG00000074657	100(1)	14(14)	0.2			
Dock8	ENSMUSG00000052085	ENSG00000107099	88(5)	15(15)	0.2			
Rnf125	ENSMUSG00000033107	ENSG00000101695	78(1)	14(14)	0.2			
Sacs	ENSMUSG00000048279	ENSG00000151835	64(1)	4(5)	0.2	Y		
Arhgef3	ENSMUSG00000021895	ENSG00000163947	47(2)	10(10)	0.2	Y		
Rbms3	ENSMUSG00000039607	ENSG00000144642	43(4)	1(3)	0.2	Y		
Arpp21	ENSMUSG00000032503	ENSG00000172995	40(2)	2(2)	0.2			
Grm1	ENSMUSG00000019828	ENSG00000152822	37(6)	2(1)	0.2	Y		
Scye1	ENSMUSG00000028029	ENSG00000164022	36(1)	12(17)	0.2	Y		
Fas	ENSMUSG00000024778	ENSG00000026103	35(2)	11(13)	0.2	Y		Y
Mthfd1l	ENSMUSG00000040675	ENSG00000120254	35(1)	5(5)	0.2	Y		
Map3k8	ENSMUSG00000024235	ENSG00000107968	35(1)	13(20)	0.2	Y		
Wwox	ENSMUSG00000004637	ENSG00000186153	34(3)	1(2)	0.2	Y		
Esr1	ENSMUSG00000019768	ENSG00000091831	34(1)	5(5)	0.2	Y		
Prkg1	ENSMUSG00000052920	ENSG00000185532	33(1)	2(2)	0.2	Y		
Prep	ENSMUSG00000019849	ENSG00000085377	33(1)	10(19)	0.2	Y		
Utrn	ENSMUSG00000019820	ENSG00000152818	32(1)	2(1)	0.2	Y		
Cdc14b	ENSMUSG00000033102	ENSG00000081377	31(1)	19(19)	0.2			
Ank3	ENSMUSG00000069601	ENSG00000151150	29(1)	1(1)	0.2	Y		
XP_485387.1	ENSMUSG00000038578	ENSG00000106868	25(2)	5(5)	0.2	Y		
4831426I19Rik	ENSMUSG00000054150	ENSG00000176438	25(1)	2(7)	0.2			
A530016O06Rik	ENSMUSG00000050103	ENSG00000187546	24(7)	1(1)	0.2	Y		
Ches1	ENSMUSG00000033713	ENSG00000053254	24(1)	17(17)	0.2			
Auts2	ENSMUSG00000056924	ENSG00000158321	22(1)	1(14)	0.2	Y		
Rasgrp1	ENSMUSG00000027347	ENSG00000172575	21(6)	6(7)	0.2	Y		
Sec8l1	ENSMUSG00000029763	ENSG00000131558	20(1)	1(3)	0.2	Y		
Hars2	ENSMUSG00000027430	ENSG00000125821	20(1)	10(20)	0.2	Y		
Rad51l1	ENSMUSG00000059060	ENSG00000182185	19(1)	3(11)	0.2	Y		
Magi2	ENSMUSG00000040003	ENSG00000187391	18(3)	7(6)	0.2	Y		
Gys2	ENSMUSG00000030244	ENSG00000111713	17(1)	14(18)	0.2	Y		
Atg10	ENSMUSG00000021619	ENSG00000152348	16(4)	5(5)	0.2	Y		
Gnefr	ENSMUSG00000030839	ENSG00000129158	16(2)	2(4)	0.2	Y		
Dmxl1	ENSMUSG00000037416	ENSG00000172869	15(1)	23(23)	0.2	Y		
Frmd6	ENSMUSG00000048285	ENSG00000139926	14(1)	12(12)	0.2	Y		
1810060J02Rik	ENSMUSG00000030301	ENSG00000123106	14(1)	13(13)	0.2	Y		
Sipa1l2	ENSMUSG00000001995	ENSG00000116991	12(1)	7(9)	0.2	Y		
Zfp496	ENSMUSG00000020472	ENSG00000162714	11(2)	16(16)	0.2			
AI194318	ENSMUSG00000048058	ENSG00000179241	11(2)	3(3)	0.2	Y		
Eltd1	ENSMUSG00000039167	ENSG00000162618	11(1)	2(2)	0.2	Y		
Crim1	ENSMUSG00000024074	ENSG00000150938	9(1)	1(1)	0.2	Y		
Car2	ENSMUSG00000027562	ENSG00000104267	8(1)	12(12)	0.2	Y		
Ctnnd1	ENSMUSG00000034101	ENSG00000198561	8(1)	17(17)	0.2			
Evi1	ENSMUSG00000027684	ENSG00000085276	8(1)	19(26)	0.2			
Slc15a4	ENSMUSG00000029416	ENSG00000139370	7(1)	3(3)	0.2	Y		
Lpp	ENSMUSG00000033306	ENSG00000145012	7(1)	33(33)	0.2	Y		
Q8BG85_MOUSE	ENSMUSG00000028497	ENSG00000188921	105	10	0.6	Y		
Mbd2	ENSMUSG00000024513	ENSG00000134046	95	1	0.6	Y		
Glis3	ENSMUSG00000052942	ENSG00000107249	84	2	0.6	Y		
Diap3	ENSMUSG00000022021	ENSG00000139734	74	5	0.6			
Mtmr9	ENSMUSG00000035078	ENSG00000104643	70	26	0.6	Y		
Mobkl2b	ENSMUSG00000039945	ENSG00000120162	66	1	0.6			
2610206B13Rik	ENSMUSG00000022120	ENSG00000152193	66	4	0.6			
Lpin2	ENSMUSG00000024052	ENSG00000101577	62	17	0.6			
D18Ertd653e	ENSMUSG00000024544	ENSG00000168675	62	63	0.6	Y		
Elp3	ENSMUSG00000022031	ENSG00000134014	59	9	0.6	Y		
Lig4	ENSMUSG00000049717	ENSG00000174405	59	18	0.6			
Acsl1	ENSMUSG00000018796	ENSG00000151726	51	44	0.6	Y		
Frmd4b	ENSMUSG00000030064	ENSG00000114541	43	8	0.6	Y		
Pim3	ENSMUSG00000035828	ENSG00000198355	43	39	0.6			
Foxp1	ENSMUSG00000030067	ENSG00000114861	42	3	0.6	Y		
Pcaf	ENSMUSG00000000708	ENSG00000114166	42	4	0.6	Y		
Q8BKG9_MOUSE	ENSMUSG00000032035	ENSG00000134954	41	2	0.6	Y		
Fli1	ENSMUSG00000016087	ENSG00000151702	40	8	0.6	Y		
Cd38	ENSMUSG00000029084	ENSG00000004468	40	20	0.6	Y		
Prdm10	ENSMUSG00000042496	ENSG00000170325	39	10	0.6			
Dnmt2	ENSMUSG00000026723	ENSG00000107614	39	31	0.6	Y		
IGHA_MOUSE	ENSMUSG00000054328	ENSG00000177199	38	94	0.6			
Pde10a	ENSMUSG00000023868	ENSG00000112541	37	15	0.6	Y		
Myh9	ENSMUSG00000022443	ENSG00000100345	37	21	0.6			
Lef1	ENSMUSG00000027985	ENSG00000138795	36	6	0.6			
BC024806	ENSMUSG00000039048	ENSG00000110074	36	8	0.6			
Arhgap18	ENSMUSG00000039031	ENSG00000146376	35	4	0.6	Y		
Ptpre	ENSMUSG00000041836	ENSG00000132334	35	5	0.6	Y		
Tube1	ENSMUSG00000019845	ENSG00000074935	35	8	0.6	Y		
Centd1	ENSMUSG00000037999	ENSG00000047365	34	6	0.6			
Scfd2	ENSMUSG00000062110	ENSG00000184178	33	1	0.6	Y		
Kcnab2	ENSMUSG00000028931	ENSG00000069424	33	11	0.6			
Trim2	ENSMUSG00000027993	ENSG00000109654	33	32	0.6	Y		
TCA_MOUSE	ENSMUSG00000041018	ENSG00000166056	32	1	0.6	Y		
Nrap	ENSMUSG00000049134	ENSG00000197893	32	9	0.6	Y		
Sept11	ENSMUSG00000058013	ENSG00000138758	32	29	0.6			
Mcart1	ENSMUSG00000045973	ENSG00000122696	29	16	0.6			
Pip5k1a	ENSMUSG00000024867	ENSG00000107242	29	17	0.6			
Gpr56	ENSMUSG00000031785	ENSG00000159618	27	58	0.6			
Bcl11b	ENSMUSG00000048251	ENSG00000127152	24	17	0.6	Y		
4930402H24Rik	ENSMUSG00000027309	ENSG00000088854	23	16	0.6	Y		
5430432M24Rik	ENSMUSG00000027459	ENSG00000125898	23	40	0.6			
Ddx4	ENSMUSG00000021758	ENSG00000152670	20	15	0.6	Y		
Btla	ENSMUSG00000052013	ENSG00000186265	19	6	0.6	Y		
Trim30	ENSMUSG00000030921	ENSG00000132256	19	20	0.6			
6430601A21Rik	ENSMUSG00000040321	ENSG00000198146	18	17	0.6			
Man2a1	ENSMUSG00000024085	ENSG00000112893	17	9	0.6			
6330442E10Rik	ENSMUSG00000056219	ENSG00000198133	17	14	0.6			
Kif5c	ENSMUSG00000026764	ENSG00000168280	17	32	0.6			
Hivep1	ENSMUSG00000021366	ENSG00000095951	16	6	0.6	Y		
AI875199	ENSMUSG00000018995	ENSG00000137513	16	8	0.6	Y		
Slc30a5	ENSMUSG00000021629	ENSG00000145740	16	10	0.6	Y		
Pde3b	ENSMUSG00000030671	ENSG00000152270	16	11	0.6	Y		
Pnn	ENSMUSG00000020994	ENSG00000100941	16	12	0.6	Y		
Rffl	ENSMUSG00000020696	ENSG00000092871	16	26	0.6			
	ENSMUSG00000021171	ENSG00000117868	16	33	0.6	Y		
Ripk3	ENSMUSG00000022221	ENSG00000129465	16	35	0.6			
Tep1	ENSMUSG00000006281	ENSG00000129566	16	100	0.6			

continued on next page

continued from previous page

Gene name	Mouse Ensembl ID	Human Ensembl ID	Number of cell lines	Genes in minimal region	Copy number	Insertions in translated region?	Singletons only?	Known TSG?
Slco3a1	ENSMUSG00000025790	ENSG00000176463	15	1	0.6	Y		
NP_001019895.1	ENSMUSG00000033147	ENSG00000163393	15	4	0.6	Y		
St3gal6	ENSMUSG00000022747	ENSG00000064225	15	7	0.6	Y		
Slc36a3	ENSMUSG00000049491	ENSG00000186334	15	8	0.6	Y		
Itpr5	ENSMUSG00000030287	ENSG00000123104	15	10	0.6	Y		
	ENSMUSG00000042590	ENSG00000086200	15	36	0.6			
	ENSMUSG00000062252	ENSG00000197753	15	40	0.6			
NP_079558.1	ENSMUSG00000005583	ENSG00000081189	14	2	0.6	Y		
Phf14	ENSMUSG00000029629	ENSG00000106443	14	9	0.6	Y		
2810013C04Rik	ENSMUSG00000066411	ENSG00000173575	14	18	0.6			
Lyn	ENSMUSG00000042228	ENSG00000147507	14	20	0.6			
Cd53	ENSMUSG00000040747	ENSG00000143119	13	2	0.6			
St6galnac3	ENSMUSG00000052544	ENSG00000184005	13	2	0.6	Y		
Grik1	ENSMUSG00000022935	ENSG00000171189	13	6	0.6	Y		
Rab27a	ENSMUSG00000032202	ENSG00000069974	13	8	0.6	Y		
Zfhx1b	ENSMUSG00000026872	ENSG00000169554	13	10	0.6			
A130038L21Rik	ENSMUSG00000021703	ENSG00000164300	13	13	0.6	Y		
Dscr2	ENSMUSG00000022913	ENSG00000183527	13	14	0.6	Y		
1700001D09Rik	ENSMUSG00000010135	ENSG00000121933	13	26	0.6			
Sh3gl3	ENSMUSG00000030638	ENSG00000140600	13	27	0.6	Y		
Sdk1	ENSMUSG00000039683	ENSG00000146555	12	1	0.6	Y		
Hivep3	ENSMUSG00000028634	ENSG00000127124	12	2	0.6	Y		
	ENSMUSG00000021676	ENSG00000145703	12	3	0.6			
Mgat5	ENSMUSG00000036155	ENSG00000152127	12	17	0.6			
Cdc42se2	ENSMUSG00000052298	ENSG00000158985	12	54	0.6			
Wdfy1	ENSMUSG00000004377	ENSG00000085449	11	1	0.6	Y		
Bard1	ENSMUSG00000026196	ENSG00000138376	11	3	0.6			
D12Ertd553e	ENSMUSG00000020589	ENSG00000197872	10	1	0.6			
Nfia	ENSMUSG00000028565	ENSG00000162599	10	1	0.6	Y		
8430438D04Rik	ENSMUSG00000036019	ENSG00000179104	10	4	0.6	Y		
Acvr1	ENSMUSG00000026836	ENSG00000115170	10	4	0.6			
Mpp4	ENSMUSG00000026024	ENSG00000003393	9	4	0.6	Y		
Slc39a11	ENSMUSG00000041654	ENSG00000133195	9	12	0.6	Y		
	ENSMUSG00000053396	ENSG00000185676	8	32	0.6			
Dnmt3a	ENSMUSG00000020661	ENSG00000119772	7	22	0.6	Y		
1110014D18Rik	ENSMUSG00000059586	ENSG00000156831	7	29	0.6	Y		
Ccnl1	ENSMUSG00000027829	ENSG00000163660	7	29	0.6	Y		
Myc	ENSMUSG00000022346	ENSG00000136997	6	3	0.6			
1600014C10Rik	ENSMUSG00000054676	ENSG00000131943	6	12	0.6	Y		
Phf21a	ENSMUSG00000058318	ENSG00000135365	6	19	0.6	Y		
	ENSMUSG00000057788	ENSG00000105671	6	60	0.6	Y		
NM_011210.1	ENSMUSG00000026395	ENSG00000081237	5	3	0.6	Y		
Lrp12	ENSMUSG00000022305	ENSG00000147650	5	6	0.6	Y		
Stxbp4	ENSMUSG00000020546	ENSG00000166263	5	8	0.6	Y		
Galnt14	ENSMUSG00000024064	ENSG00000158089	5	16	0.6	Y		
Meis1	ENSMUSG00000020160	ENSG00000143995	4	3	0.6	Y		
Ccdc19	ENSMUSG00000026546	ENSG00000158710	4	40	0.6	Y		
Wdr7	ENSMUSG00000040560	ENSG00000091157	98(2)	2(3)	0.2	Y	Y	
Rfx3	ENSMUSG00000040929	ENSG00000080298	85(3)	3(3)	0.2		Y	
Nfib	ENSMUSG00000008575	ENSG00000147862	84(4)	2(2)	0.2	Y	Y	
	ENSMUSG00000064286	ENSG00000189076	75(3)	354(6)	0.2	Y	Y	
Htr2a	ENSMUSG00000034997	ENSG00000102468	72(1)	5(5)	0.2	Y	Y	
6430573F11Rik	ENSMUSG00000039620	ENSG00000170941	67(1)	3(11)	0.2		Y	
Gpc5	ENSMUSG00000022112	ENSG00000179399	65(2)	4(2)	0.2	Y	Y	
Flt1	ENSMUSG00000029648	ENSG00000102755	63(1)	6(6)	0.2	Y	Y	
Gas7	ENSMUSG00000033066	ENSG00000007237	57(2)	1(16)	0.2		Y	
Rac1	ENSMUSG00000001847	ENSG00000136238	45(1)	126(22)	0.2	Y	Y	
Park2	ENSMUSG00000023826	ENSG00000185345	42(5)	2(1)	0.2	Y	Y	
Robo1	ENSMUSG00000022883	ENSG00000169855	42(1)	4(6)	0.2	Y	Y	
Htr1f	ENSMUSG00000050783	ENSG00000179097	37(2)	9(9)	0.2		Y	
Il15	ENSMUSG00000031712	ENSG00000164136	34(1)	4(4)	0.2		Y	
Pank1	ENSMUSG00000033610	ENSG00000152782	33(2)	1(1)	0.2	Y	Y	
Slc44a1	ENSMUSG00000028412	ENSG00000070214	25(1)	20(20)	0.2	Y	Y	
D16Ertd472e	ENSMUSG00000022864	ENSG00000154642	24(4)	13(13)	0.2	Y	Y	
Ubxd3	ENSMUSG00000043621	ENSG00000162543	24(1)	5(5)	0.2		Y	
Arfrp2	ENSMUSG00000042348	ENSG00000185305	20(2)	1(7)	0.2		Y	
Col19a1	ENSMUSG00000026141	ENSG00000082293	20(1)	2(15)	0.2	Y	Y	
Accn1	ENSMUSG00000020704	ENSG00000108684	18(1)	1(1)	0.2	Y	Y	
Rhoj	ENSMUSG00000046768	ENSG00000126785	17(1)	14(14)	0.2	Y	Y	
Usp47	ENSMUSG00000059263	ENSG00000170242	16(2)	4(4)	0.2		Y	
4833446K15Rik	ENSMUSG00000058152	ENSG00000198108	16(1)	3(2)	0.2	Y	Y	
Dut	ENSMUSG00000027203	ENSG00000128951	14(1)	7(7)	0.2	Y	Y	
Klf7	ENSMUSG00000025959	ENSG00000118263	13(1)	1(1)	0.2	Y	Y	
Ifngr2	ENSMUSG00000022965	ENSG00000159128	12(1)	12(12)	0.2	Y	Y	
Tmem16f	ENSMUSG00000064210	ENSG00000177119	10(1)	21(21)	0.2	Y	Y	
Lrrk2	ENSMUSG00000036273	ENSG00000188906	10(1)	22(22)	0.2	Y	Y	
	ENSMUSG00000014781	ENSG00000164256	7(2)	2(7)	0.2	Y	Y	
Thada	ENSMUSG00000024251	ENSG00000115970	7(1)	5(5)	0.2	Y	Y	

Table 4.7. Mouse genes that contain retroviral insertions within the transcribed or translated region and are also promising candidates for targets of deletion in human cancer cell lines. "Number of cell lines" is the number of samples in which the gene is deleted. "Genes in minimal region" is the number of genes that co-occur with the CIS gene in the smallest region of deletion. "Copy number" is the minimal threshold below which the gene is deleted. Where the copy number is 0.2, the number of cell lines and number of genes in the minimal deleted region are given for deletions of copy number <= 0.6, with numbers for copy number <= 0.2 being shown in brackets. "Insertions in translated region?" indicates whether any of the insertions are within the translated region of the gene. "Singletons only?" indicates whether the gene contains insertions other than those represented by a single read. "Known TSG?" indicates whether the gene is a recessive cancer gene listed in the Cancer Gene Census.

tumourigenesis (see Section 1.3.3.3). The identification of insertions within the gene provides strong evidence that it does contribute to cancer (see also Section 3.4.3). Deletions of less than 70 Mb encompassing *WWOX* and insertions in *Wwox* are shown in Figure 4.20. In Section 3.4.3, *Foxp1* was proposed as a putative tumour suppressor gene. Deletion of *FOXP1* was observed in 42 cell lines, with a minimal amplified region of 3 genes, therefore providing additional evidence that this gene contributes to cancer and that it does so in both species. *Mobkl2a* was also presented as a putative tumour suppressor gene in Section 3.4.3, and while the human orthologue of this gene was not deleted in cancer, the human orthologue of paralogue *Mobkl2b* was deleted. Another implicated tumour suppressor gene identified in this analysis was *DOCK8*, which is deleted and under-expressed in human lung cancers (Takahashi *et al.*, 2006).

Known oncogenes *EVI1*, *MYC* and *FLI1* were also identified in the analysis, demonstrating that the results must be viewed with caution and that functional validation, as well as analysis of the distribution of insertions within the mouse candidate, is required to determine whether deletion of the identified genes is likely to contribute to tumourigenesis. Other candidates that have been implicated as oncogenes include *GRM1*, which plays an important role in the transformation of melanocytes in melanoma (Shin *et al.*, 2008), *RASGPR1*, which contributes to tumour progression in murine skin cancer (Luke *et al.*, 2007; Oki-Idouchi and Lorenzo, 2007) and, as mentioned in the previous sections, *MAP3K8*, *MPP4* and *MEIS1*. Likewise, amplification and overexpression of genes encoding cyclin L1 (*CCNL1*), low-density lipoprotein receptor-related protein 12 precursor (*LRP12*) and glypican-5 (*GPC5*) have been demonstrated in human head and neck squamous cell carcinomas (Muller *et al.*, 2006; Redon *et al.*, 2002), oral squamous cell carcinomas (Garnis *et al.*, 2004) and rhabdomyosarcomas (Williamson *et al.*, 2007), respectively. It is therefore likely that other genes are the targets of deletion in the regions containing these known and implicated oncogenes. *GRM1* was the only gene for which the minimal deleted region did not contain additional genes. However, this does not prove that *GRM1* must be the critical gene, since deletions affecting upstream and downstream genes may simply overlap at *GRM1*.

The list contains many genes for which there is limited evidence in the literature to suggest that they may act as tumour suppressor genes. The results of this analysis therefore lend further support to these findings. Some of these candidates (*RBMS3*, *PCAF*, *UTRN*, *ANK3*, *ACCN1*, *CDC14B*, *CHES1* and *PARK2*) are briefly discussed

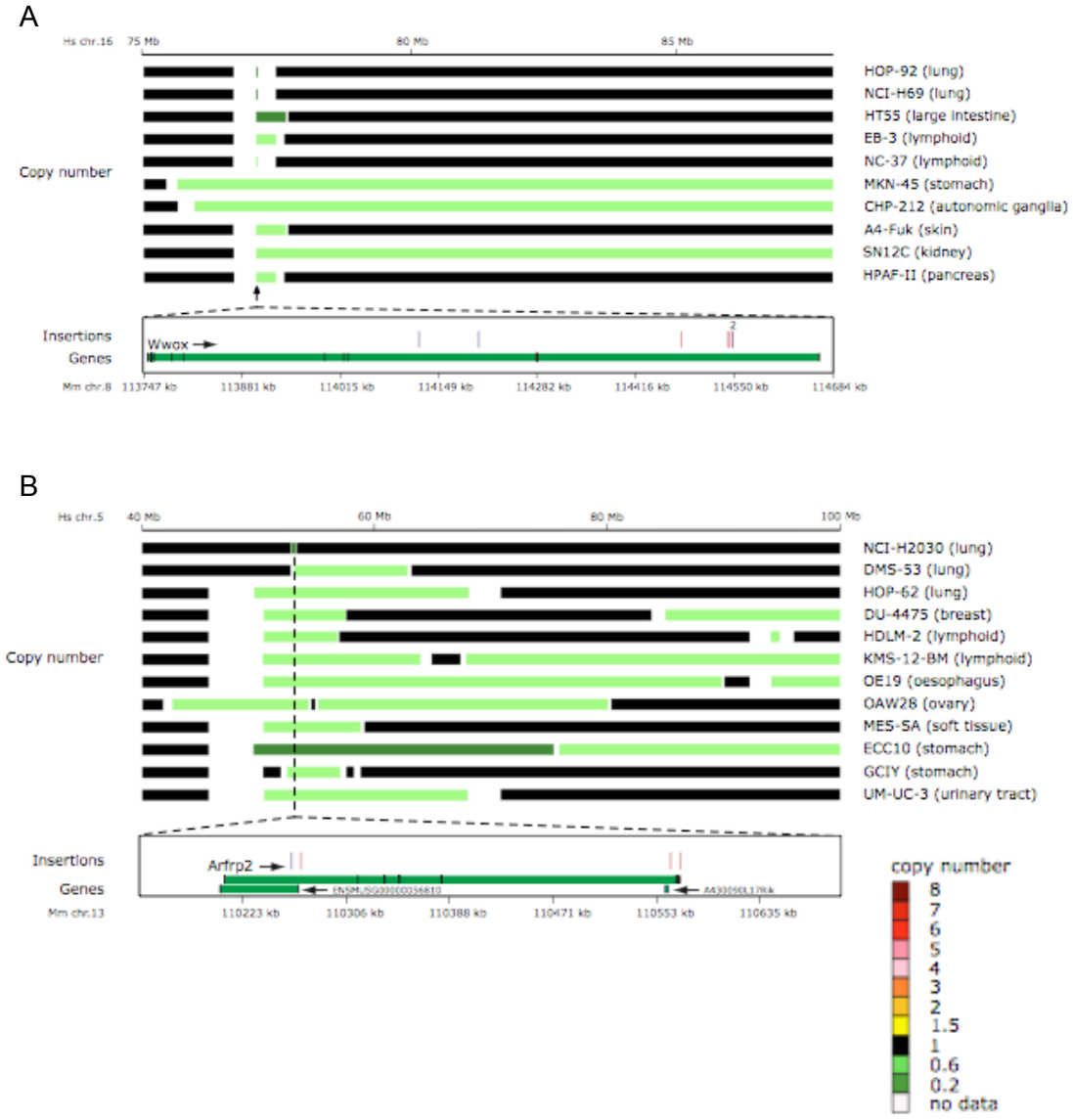

Figure 4.20. Candidate tumour suppressor genes *WWOX* (A) and *ARFRP2* (B) are deleted in human cancer cell lines and are disrupted by retroviral insertional mutagenesis in mouse lymphomas. The copy number of chromosomal regions in the human cell lines is depicted in colour. Names of human cell lines and tissue of origin are provided Only cell lines in which the deletion containing the gene is less than 70 Mb are shown. The lower part of each figure shows insertions within mouse tumours, and was kindly provided by Jaap Kool and Jeroen de Ridder. Blue vertical lines represent insertions in the sense orientation, while red vertical lines represent antisense insertions. Genes are shown in green, with exons marked in black. Positions on the murine and human chromosomes are indicated on the black horizontal bars in kb and Mb, respectively. These figures can also be seen in Uren *et al*. (2008).

below. *RBMS3* and *PCAF* reside in commonly deleted regions, and are down-regulated, in oesophageal squamous cell carcinomas (Qin *et al.*, 2008). One of the homozygous deletions that contained *RBMS3* was from an oesophageal cancer cell line (COLO-608N), but the remaining three were from the large intestine (NCI-H747), ovary (TYK-nu) and cervix (SKG-IIIa). The single homozygous deletion of *PCAF* was in a biliary tract cell line (EGI-1), and neither gene was deleted below copy number 0.6 in any additional oesophageal cancer cell lines. This suggests that the genes may also contribute to other cancers. The utrophin gene (*UTRN*) resides within a deletion of the long arm of chromosome 6 that is frequently observed in a range of tumours, and *UTRN* has been recently proposed as a putative tumour suppressor gene within this region (Li *et al.*, 2007b). Ankyrin-3 (*ANK3*) is a target of the transcription factor hepatocyte nuclear factor 4 alpha that down-regulates cell proliferation in kidney cells (Grigo *et al.*, 2008). None of the 29 deletions containing *ANK3* were within cell lines derived from kidney cancer, but the fact that this was the only gene in the minimal deleted region provides support for a role in tumour suppression. *ACCN1* was proposed as a putative glioma tumour suppressor gene following the observation that surface expression of one of the two isoforms reduces cell growth and migration (Vila-Carriles *et al.*, 2006), while the gene was also shown to be disrupted by a translocation within a neuroblastoma (Vandepoele *et al.*, 2008). Notably, the single homozygous deletion containing this gene was within a glioma cell line (8-MG-BA), while 3 of the remaining 17 deletions of copy number less than or equal to 0.6 were within neuroblastomas. The rest of the deletions were in a range of tumours, including 3 breast, 2 bone, 2 lung and 2 ovarian. *CDC14B* and *CHES1* are both involved in regulating cell cycle checkpoints related to DNA damage response (Bassermann *et al.*, 2008; Busygina *et al.*, 2006), and the deletion of these genes could therefore contribute to tumourigenesis by allowing damaged cells to enter mitosis. Like *WWOX*, the Parkin gene (*PARK2*) resides within a common fragile site (FRA6E) and, therefore, while the gene is frequently deleted in cancer, it is unclear whether it contributes to cancer development. However, deletions involving *PARK2* are associated with ovarian cancer (Denison *et al.*, 2003) and glioblastoma multiforme (Mulholland *et al.*, 2006), and promoter hypermethylation of *PARK2*, resulting in down-regulation of gene expression, has been observed in leukaemias (Agirre *et al.*, 2006). *PARK2* is a long gene, measuring 994.53 kb, and contains just 2 insertions that could have occurred by chance. Therefore, the presence of insertions within the gene does not provide convincing support for a role in tumourigenesis. Interestingly, a break in FRA6E was associated with poor outcome in breast carcinomas, but expression of *PARK2* was not

associated, while the loss of *AF-6* gene, which is telomeric of *PARK2*, was associated, suggesting that this may be a tumour suppressor gene affected by the break (Letessier *et al.*, 2007). Other candidates for which there is evidence in the literature of a tumour suppressive role in cancer include *BARD1*, *DMXL1*, *GPR56*, *HIVEP1*, *KCNAB2*, *LEF1*, *LIG4*, *PHF14*, *RAD51L1* and *RIPK3*. Further candidates *SDK1*, *BCL11B* and *MBD2* are discussed in Section 5.3.2.2.

ARFRP2 is a novel candidate tumour suppressor gene for which there is currently no evidence in the literature for a role in cancer. ARFRP2, also known as ARL15, is a member of the ADP-ribosylation factor-like family. Another member of this family, *ARL11*, is a tumour suppressor gene for which truncating germline mutations and promoter methylation contribute to leukaemia, breast cancer, ovarian cancer and melanoma (Frank *et al.*, 2005; Petrocca *et al.*, 2006). Deletions of less than 70 Mb that encompass *ARFRP2* are shown in Figure 4.20. There is also no evidence in the literature to suggest that the sec1 family domain containing gene *SCFD2* is a tumour suppressor gene. However, *SCFD2* is a transcriptional target of p53 (Krieg *et al.*, 2006), and it is the only gene within the minimal deleted region of 33 cancer cell lines.

4.6 Comparison of methods for calling gains and losses

As discussed in Section 4.3, DNAcopy and MergeLevels were the algorithms chosen for detecting regions of copy number change because they had been shown to perform better than other methods, and were freely available. However, it is not known whether DNAcopy and MergeLevels out-perform other methods in processing copy number data generated on the 10K SNP array CGH platform, and a variety of methods were therefore compared. The methods tested were DNAcopy alone (Olshen *et al.*, 2004), DNAcopy and MergeLevels (Olshen *et al.*, 2004; Willenbrock and Fridlyand, 2005), FASeg (Yu *et al.*, 2007), BioHMM (Marioni *et al.*, 2006) and a selection of the methods included within ADaCGH (Diaz-Uriarte and Rueda, 2007), i.e. CGHseg (Picard *et al.*, 2005), HMM (Fridlyand *et al.*, 2004), Wavelets (Hsu *et al.*, 2005) and GLAD (Hupe *et al.*, 2004).

27 different runs of DNAcopy version 1.4.0 were performed, each time varying the parameters. Alpha values of 0.1, 0.05, 0.01, 0.005 and 0.001 were tested, change-points that differed by less than 1, 2, 3 or 4 standard deviations were removed or all change-points were retained, and the smoothing step was either performed or was omitted from

the process (see Section 4.3 for details of these parameters). A further 17 runs of DNAcopy plus MergeLevels were performed, with various combinations of values for the DNAcopy parameters and the Wilcoxon and Ansari-Bradley thresholds within MergeLevels. The Wilcoxon rank sum test is used to determine whether there is a significant difference (according to the Wilcoxon threshold) between the observed values for two copy number levels, or whether they should be merged. The Ansari-Bradley 2-sample test determines whether there is any significant difference between the distribution of merged values minus observed $\log_2$-ratios (i.e. the original ratios at individual SNPs) compared with the distribution of original segmented values minus observed $\log_2$-ratios. The optimal Ansari-Bradley threshold is the largest threshold where the distributions do not differ significantly (Willenbrock and Fridlyand, 2005).

BioHMM is available as part of the BioConductor/R package, snapCGH. It is the only method that takes into account the distance between clones (or in this case SNPs), rather than simply ordering the clones or SNPs along the chromosome. BioHMM uses a Hidden Markov Model to segment data into a finite number of hidden states, where all of the data-points within a state have an equivalent copy number (Marioni *et al.*, 2006). A single run of BioHMM version 1.2.0 was performed using default parameters.

ADaCGH (analysis of data from aCGH) is a web-based tool that provides a selection of the best-performing methods via a simple user interface. DNAcopy and MergeLevels are available within this tool, but it is only possible to use default parameters and the MergeLevels output has been post-processed into three states: -1 (loss), 0 (no change) and 1 (gain). Methods within ADaCGH were chosen because they have been shown to perform well in the comparisons by Lai *et al.* (2005) and Willenbrock and Fridlyand (2005) and/or because they help to present a cross-section of the types of algorithm available for detecting copy number changes. CGHseg models the CGH data as a random Gaussian process and segments the data at points where the mean $\log_2$-ratio changes abruptly. A threshold must be set for the adaptive penalisation, which is a threshold used to estimate the number of segments in the data. Picard *et al.* (2005) proposed a threshold of -0.05 as the default value, but Diaz-Uriarte and Rueda (2007) found that values around -0.005 were more appropriate but recommended experimenting with different values, which must be less than 0. For this analysis, 5 runs of CGHseg were performed, using thresholds of -0.005, -0.01, -0.05, -0.1 and -0.2. The smoothing approach of Hsu *et al.* (2005) uses wavelets to "denoise" the DNA copy number data and so to capture copy

number changes while smoothing out the noise. HMM is another method in which Hidden Markov Models are fitted to the data to identify different states, or copy number levels (Fridlyand *et al.*, 2004). However, unlike BioHMM, it does not take account of distances between data-points. Finally, the detection of breakpoints in GLAD is based on the Adaptive Weights Smoothing (AWS) procedure. This method finds the maximal neighbourhood around each data-point in which the local constant assumption holds true. In other words, it finds regions within which the copy number does not differ significantly and the boundaries of these regions represent breakpoints where the copy number changes. Default parameters were used for GLAD, HMM and the wavelets approach. All runs were performed in December 2007 on the website http://adacgh.bioinfo.cnio.es/.

FASeg, or Forward-Backward Fragment-Annealing Segmentation, is available as an R package from http://www.sph.emory.edu/bios/FASeg/. It is proposed to be especially suitable for SNP array CGH, which has a higher probe density but lower signal-to-noise ratio than traditional array CGH. According to the developers, the performance of FASeg was superior to 6 R packages, including DNAcopy, GLAD, BioHMM and CGHseg, in the detection of small segments with a low signal-to-noise ratio, although GLAD and BioHMM also performed well when the signal-to-noise ratio was low and the segments flanking copy number changes were long. When the signal-to-noise ratio was high, most methods performed well, although the HMM-based methods were less effective when there were multiple copy number levels within a single chromosome. This is a significant drawback, since multiple states are common in unstable cancer genomes. FASeg breaks each chromosome into small segments in an over-sensitive edge (or breakpoint) detection step that involves LOESS smoothing. It then iteratively merges consecutive segments until all remaining edges pass a significance threshold, based on testing for equal means between the groups of copy number values for SNPs before and after the edge using the unpaired Student's *t*-test. 15 different runs of FASeg version 1.2 were performed, in which parameters were altered for the smoothing span, which is the number of SNPs used to calculate the weights around each probe in the LOESS smoother, and the *P*-value cut-off for defining the significance of each edge. (See Yu *et al.*, 2007)

In total, 69 different method and/or parameter combinations were compared. Each method was performed on the same 50 randomly selected cancer cell lines. The results were compared using Matthew's Correlation Coefficient (MCC), which is described in

Section 2.10.2. 280 Ensembl genes corresponding to known oncogenes involved in translocations or amplifications were extracted from the Cancer Gene Census. The number of known oncogenes and the number of other Ensembl genes within, and outside of, amplicons of copy number greater than or equal to 2.7 were counted. Oncogenes and other genes within amplicons were defined as true positives and false positives, respectively. Oncogenes and other genes that were not within amplicons were defined as false negatives and true negatives, respectively. The numbers of true and false positives and negatives in each cell line were then added together to give the number across all cell lines, and the MCC score was calculated. This analysis was performed individually on each method. It is possible that some of the known oncogenes that are involved in translocations are not amplified in human cancer, and of course there will be a proportion of non-oncogenes that are amplified in, and contribute to the development of, cancer. However, this analysis gives an indication of the performance of the method in comparison to other methods. The coverage was defined as the proportion of known oncogenes that were represented in amplicons, and the accuracy was defined as the proportion of genes in amplicons that were known oncogenes. The coverage, accuracy and MCC score for each method are shown in Table 4.8.

The wavelet, HMM and BioHMM algorithms all performed poorly. In the case of HMM and BioHMM, this may reflect the fact that there are often multiple copy number levels within a chromosome (see above). The low signal-to-noise ratio may account for the poor performance of the wavelet approach, since this method involves "denoising" the data but was developed for conventional array CGH data, which has a higher signal-to-noise ratio. Denoising the SNP CGH data may result in the removal of biologically relevant copy number changes. In addition, only the default parameters were used for this method. Changing the penalty constant in CGHseg made a considerable difference to the number of amplicons that were detected. This demonstrates the importance of choosing suitable parameter values. The closer the value was to 0, the greater the number of amplicons and the higher the coverage. However, the accuracy fell considerably. The default parameter value of -0.05 gave the best overall results, but this was lower than many of the results obtained using FASeg or DNAcopy. The value suggested in ADaCGH, i.e. -0.005, produced the highest coverage of all methods, but at the expense of a low accuracy. Although only the default parameters were used, GLAD performed reasonably well, obtaining similar results to the best-performing DNAcopy runs.

Method	Parameters	TP	FP	TN	FN	Coverage	Accuracy	MCC
FASeg	p=0.01, smooth=7	31	1027	883573	13969	0.00221	0.02930	0.00380
FASeg	p=0.001, smooth=7	25	830	883770	13975	0.00179	0.02924	0.00340
FASeg	p=0.001, smooth=5	22	757	883843	13978	0.00157	0.02824	0.00301
DNAcopy & MergeLevels	alpha=0.05, smooth, SD=1	25	907	883693	13975	0.00179	0.02682	0.00293
DNAcopy & MergeLevels	alpha=0.1, smooth, w=0.00001	28	1055	883545	13972	0.00200	0.02585	0.00288
DNAcopy	alpha=0.05, smooth, SD=3	17	560	884040	13983	0.00121	0.02946	0.00284
FASeg	p=0.0001, smooth=10	18	608	883992	13982	0.00129	0.02875	0.00281
GLAD		25	931	883669	13975	0.00179	0.02615	0.00279
DNAcopy	alpha=0.01, smooth, SD=2	17	571	884029	13983	0.00121	0.02891	0.00275
DNAcopy	alpha=0.01, smooth, SD=1	18	620	883980	13982	0.00129	0.02821	0.00272
DNAcopy & MergeLevels	alpha=0.1, smooth, SD=1 w=0.00001	28	1087	883513	13972	0.00200	0.02511	0.00271
DNAcopy & MergeLevels	alpha=0.1, smooth, SD=1 w=0.00001, ans=0.01	28	1087	883513	13972	0.00200	0.02511	0.00271
DNAcopy & MergeLevels	alpha=0.1, smooth, SD=1 w=0.00001, ans=0.1	28	1087	883513	13972	0.00200	0.02511	0.00271
DNAcopy	alpha=0.05, smooth, SD=1	23	854	883746	13977	0.00164	0.02623	0.00269
DNAcopy	alpha=0.1, smooth, SD=2	23	855	883745	13977	0.00164	0.02620	0.00268
DNAcopy & MergeLevels	alpha=0.1, smooth, SD=1	28	1103	883497	13972	0.00200	0.02476	0.00263
DNAcopy & MergeLevels	alpha=0.1, smooth, SD=1, ans=0.01	28	1103	883497	13972	0.00200	0.02476	0.00263
DNAcopy & MergeLevels	alpha=0.1, smooth, SD=1, ans=0.1	28	1103	883497	13972	0.00200	0.02476	0.00263
FASeg	p=0.01, smooth=10	24	911	883689	13976	0.00171	0.02567	0.00263
DNAcopy	alpha=0.1, smooth, SD=1	26	1007	883593	13974	0.00186	0.02517	0.00263
DNAcopy	alpha=0.05, smooth	23	865	883735	13977	0.00164	0.02590	0.00262
DNAcopy	alpha=0.05, smooth, SD=4	13	415	884185	13987	0.00093	0.03037	0.00261
DNAcopy	alpha=0.01, smooth	18	636	883964	13982	0.00129	0.02752	0.00260
DNAcopy & MergeLevels	alpha=0.1, smooth	27	1061	883539	13973	0.00193	0.02482	0.00260
DNAcopy	alpha=0.005, smooth, SD=2	16	558	884042	13984	0.00114	0.02787	0.00251
DNAcopy	alpha=0.1, smooth	26	1029	883571	13974	0.00186	0.02464	0.00251
DNAcopy	alpha=0.01, smooth, SD=4	12	383	884217	13988	0.00086	0.03038	0.00251
DNAcopy	alpha=0.05, smooth, SD=2	20	744	883856	13980	0.00143	0.02618	0.00250
DNAcopy	alpha=0.01, smooth, SD=3	14	471	884129	13986	0.00100	0.02887	0.00249
CGHseg	penalty=-0.05	16	561	884039	13984	0.00114	0.02773	0.00249
DNAcopy	alpha=0.001, smooth, SD=2	15	526	884074	13985	0.00107	0.02773	0.00241
FASeg	p=0.001, smooth=10	20	758	883842	13980	0.00143	0.02571	0.00241
DNAcopy	alpha=0.001, smooth	16	575	884025	13984	0.00114	0.02707	0.00238
DNAcopy	alpha=0.001, smooth, SD=4	12	396	884204	13988	0.00086	0.02941	0.00238
DNAcopy	alpha=0.005, smooth, SD=4	12	396	884204	13988	0.00086	0.02941	0.00238
DNAcopy & MergeLevels	alpha=0.01, smooth, w=0.00001	19	717	883883	13981	0.00136	0.02582	0.00237
DNAcopy & MergeLevels	alpha=0.1, smooth, SD=1 w=0.001	26	1059	883541	13974	0.00186	0.02396	0.00235
DNAcopy & MergeLevels	alpha=0.1, smooth, SD=1 w=0.001, ans=0.01	26	1059	883541	13974	0.00186	0.02396	0.00235
DNAcopy & MergeLevels	alpha=0.1, smooth, SD=1 w=0.001, ans=0.1	26	1059	883541	13974	0.00186	0.02396	0.00235
DNAcopy & MergeLevels	alpha=0.1, smooth, SD=2	21	816	883784	13979	0.00150	0.02509	0.00234
DNAcopy	alpha=0.001, smooth, SD=1	16	582	884018	13984	0.00114	0.02676	0.00233
DNAcopy	alpha=0.005, smooth, SD=1	16	587	884013	13984	0.00114	0.02653	0.00229
DNAcopy	alpha=0.001, smooth, SD=3	13	454	884146	13987	0.00093	0.02784	0.00226
DNAcopy & MergeLevels	alpha=0.01, smooth, SD=1	16	593	884007	13984	0.00114	0.02627	0.00225
DNAcopy	alpha=0.005, smooth, SD=3	13	458	884142	13987	0.00093	0.02760	0.00222
FASeg	p=0.0001, smooth=7	18	694	883906	13982	0.00129	0.02528	0.00221
DNAcopy	alpha=0.005, smooth	16	603	883997	13984	0.00114	0.02585	0.00218
DNAcopy & MergeLevels	alpha=0.005, smooth, w=0.00001	15	568	884032	13985	0.00107	0.02573	0.00209
FASeg	p=0.1, smooth=10	28	1261	883339	13972	0.00200	0.02172	0.00188
CGHseg	penalty=-0.01	37	1751	882849	13963	0.00264	0.02069	0.00184
DNAcopy & MergeLevels	alpha=0.05, smooth, w=0.00001	24	1119	883481	13976	0.00171	0.02100	0.00156
DNAcopy	alpha=0.01	21	961	883639	13979	0.00150	0.02138	0.00155
DNAcopy	alpha=0.001	16	703	883897	13984	0.00114	0.02225	0.00152
DNAcopy	alpha=0.005	19	891	883709	13981	0.00136	0.02088	0.00136
DNAcopy	alpha=0.05	24	1169	883431	13976	0.00171	0.02012	0.00134
CGHseg	penalty=-0.01	5	187	884413	13995	0.00036	0.02604	0.00123
FASeg	p=0.1, smooth=5	33	1710	882890	13967	0.00236	0.01893	0.00119
CGHseg	penalty=-0.2	2	72	884528	13998	0.00014	0.02703	0.00084
DNAcopy	alpha=0.1	23	1302	883298	13977	0.00164	0.01736	0.00055
BioHMM		15	849	883751	13985	0.00107	0.01736	0.00045
CGHseg	penalty=-0.005	42	2492	882108	13958	0.00300	0.01657	0.00043
FASeg	p=0.01, smooth=5	8	438	884162	13992	0.00057	0.01794	0.00042
FASeg	p=0.001	4	264	884336	13996	0.00029	0.01493	-0.00009
FASeg	p=0.0001	4	264	884336	13996	0.00029	0.01493	-0.00009
FASeg	p=0.000001	4	264	884336	13996	0.00029	0.01493	-0.00009
FASeg	p=0.0001, smooth=50	5	370	884230	13995	0.00036	0.01333	-0.00037
Wavelets		25	1769	882831	13975	0.00179	0.01394	-0.00059
HMM		9	717	883883	13991	0.00064	0.01240	-0.00073
FASeg	p=0.01	5	467	884133	13995	0.00036	0.01059	-0.00092

Table 4.8. Comparison of methods for detecting regions of copy number gain in 50 randomly selected cancer cell lines. Abbreviations for describing parameters are as follows: FASeg: p=significance threshold, smooth=smoothing range; DNAcopy: alpha=parameter alpha, smooth=outliers smoothed, SD=change-points differing by less than X standard deviations removed; MergeLevels: w=Wilcoxon threshold, ans=Ansari-Bradley threshold; CGHseg: penalty=penalty constant. Undefined parameters are default. TP=number of true positives (amplified oncogenes), FP=number of false positives (amplified non-oncogenes), TN=number of true negatives (non-amplified non-oncogenes), FN=number of false negatives (non-amplified oncogenes). Numbers are calculated across all cell lines. Coverage=TP/(TP+FN), Accuracy=TP/(TP+FP). MCC = Matthew's Correlation Coefficient.

Of the runs involving DNAcopy alone, those in which the data were not smoothed before segmentation performed worst. Higher values for the parameter alpha, which result in increased sensitivity, generally performed better due mainly to a higher coverage. For the purposes of the cross-species comparison, higher coverage, even at the expense of lower accuracy, is preferable since the mouse candidate cancer genes help to identify the targets of amplification in the human amplicons, and false positives are therefore likely to be ignored. Reducing the number of standard deviations below which change-points were removed resulted in a higher coverage of oncogenes. This may be because the highest peak of amplification, which often contains the critical cancer gene(s), is more likely to remain distinct from lower level copy number gains and the segment will have a higher mean copy number and will contain fewer amplified passengers. For higher values of alpha, merging the segments using default parameters also resulted in higher coverage. However, upon inspection of the results, it appeared that some oncogenes were lost upon merging, while some were gained. All of the oncogenes that were unique to the run without merging were still amplified in the run with merging, and vice versa, but they did not reach the copy number threshold of greater than or equal to 2.7. This is because merging increases the mean copy number of some segments and decreases the mean copy number of others, in line with the copy numbers of other segments in the genome. This demonstrates why it is useful to use a range of copy number thresholds in the comparative analysis. Changing the Ansari-Bradley threshold from 0.1 to 0.01 made no difference to the results, but lowering the value for the Wilcoxon threshold increased the MCC score. Using a lower value for the Wilcoxon threshold means that a higher proportion of segments will not be considered significantly different from one another and will therefore be merged. However, more detailed analysis suggests that lowering the value may not produce sensible results. For example, using a value of $1.0x10^{-5}$ rather than $1.0x10^{-4}$, the segment of copy number 3.00 that contains *CCND1* in the neck squamous cell carcinoma cell line SCC-15 is merged with a segment of copy number 1.78 to give an overall copy number of 1.85. While the oncogene is still amplified, merging of this kind removes the peaks in amplification, which are most likely to harbour the critical targets of amplification. Similarly, a segment of copy number 0.12 on chromosome 4 of the bone cancer cell line CAL-72 is merged with other segments to give a copy number of 0.47. This segment is likely to be a homozygous deletion but is merged with segments that are more likely to represent heterozygous deletions.

Comparison of the FASeg runs showed that using the default smoothing span of 25 rather than a lower value resulted in lower accuracy and coverage and, therefore, a lower MCC score. When a significance threshold of *P*=0.0001 was used, a smoothing span of 10 rather than 50 not only identified more oncogenes (18 rather than 4) but also had tighter amplicon boundaries that still retained the oncogene. For example, lung cancer cell line LC-2-ad and pancreatic cancer cell line HuP-T4 contained amplicons that encompassed the oncogenes *MYC* and *POU5F1*, respectively. Using smoothing spans of 50 and 10, the number of SNPs within the amplicon containing *MYC* was calculated as 17 and 15, respectively, while the number within the amplicon containing *POU5F1* was 102 and 94, respectively. Increasing the significance threshold generally decreased the MCC score. Using a smoothing span of 7, a significance threshold of *P*=0.001 yielded 64 amplicons, while a significance threshold of *P*=0.0001 yielded 43 amplicons. 35 amplicons were identical, while the rest were either missing from the latter run or were shared but spanned a larger region when the threshold was higher. For example, an amplicon in the lung cancer cell line ChaGo-K-1 spanned 1.39 Mb and had a mean copy number ratio of 3.97 using a threshold of *P*=0.0001, and 919.42 kb with a mean copy number ratio of 4.16 using a threshold of *P*=0.001. Likewise, the amplicon encompassing *MYC* in lung cancer cell line LC-2-ad was also larger (6.74 Mb rather than 4.77 Mb) using a threshold of *P*=0.0001 rather than *P*=0.001 and had a lower mean copy number (5.06 rather than 5.49). The amplicons that were missing from the run with a higher significance threshold may still be present, but the mean copy number may not reach the copy number threshold of 2.7 because a larger region, including less amplified or non-amplified SNPs, is defined as the region of copy number change and this dilutes the mean copy number ratio for the entire segment. Overall, using a significance threshold of *P*=0.01 and a smoothing span of 7 appeared to give the best results, with the highest MCC score and highest accuracy and coverage. It is worth noting, however, that the parameter value for the smoothing span is well below that recommended in Yu *et al.* (2007). The results obtained using the top scoring runs from FASeg and DNAcopy plus MergeLevels were compared. 20 cancer genes were identified by both algorithms. 11 were unique to the FASeg output, and 5 were unique to the DNAcopy and MergeLevels output. In most cases, the missing genes were still amplified, but were below the copy number threshold of 2.7. This analysis indicates that the choice of method and parameters can make a considerable difference to the output and involves finding a suitable balance between accuracy and coverage.

4.7 Global comparison of mouse candidate cancer genes and human CNVs

The global comparison method of Section 4.5.1.1 was applied to human CNVs (see Section 4.2.3) and the gene lists from Section 4.2.1. Rather than using copy number thresholds, CNVs were separated into deletions and duplications, which were specified in the original downloaded file. As with previous analyses, the number of deletions/duplications within which each gene resided was counted. For each number of deletions/duplications, the number of mouse candidates was compared to the distribution of randomised genes using the Z-test. The results are depicted in Figure 4.21. None of the gene lists showed over-representation within deletions or duplications. The only positive association was observed for 7 known oncogenes (namely *DDIT3*, *NSD1*, *IRF4*, 2 genes encoding Histone H4, *NUT* and *PDE4DIP*) that were within 32 or more duplications. The association increased as the number of duplications increased, to a maximum of $P=2.07\text{x}10^{-4}$ for 5 known oncogenes in 93 or more duplications. This suggests that some oncogenes are amplified in the normal population, and these individuals may have a predisposition to cancer. However, in general, genes involved in cancer were not found within CNVs. In fact, genes nearest to CISs ($P<0.001$ and $P<0.05$) and genes with insertions in coding regions were slightly under-represented in deletions, and genes within translated and transcribed regions were highly under-represented in both duplications and deletions. Many of the genes that are involved in oncogenesis are also involved in other important cellular functions, and this may explain why candidate oncogenes are rarely deleted in healthy individuals. Duplication of tumour suppressor genes could also lead to oncogene repression, producing a similar outcome, while deletion of tumour suppressor genes could lead to tumourigenesis. The results show that cells do not tolerate changes in copy number in genes that are important in tumourigenesis.

For each gene list, the number of genes residing within CNV deletions and within regions of copy number loss (less than or equal to a ratio of 0.6) in human cancer cell lines was counted. A 2-tailed Fisher Exact Test was performed to determine whether there was any association between genes found in deletions in normal individuals and deletions in cancer cell lines. The same analysis was performed using CNV duplications and regions of copy number gain (greater than or equal to 2.7). The *P*-values are provided in Table 4.9. In accordance with the results obtained in the global analysis, there was an under-

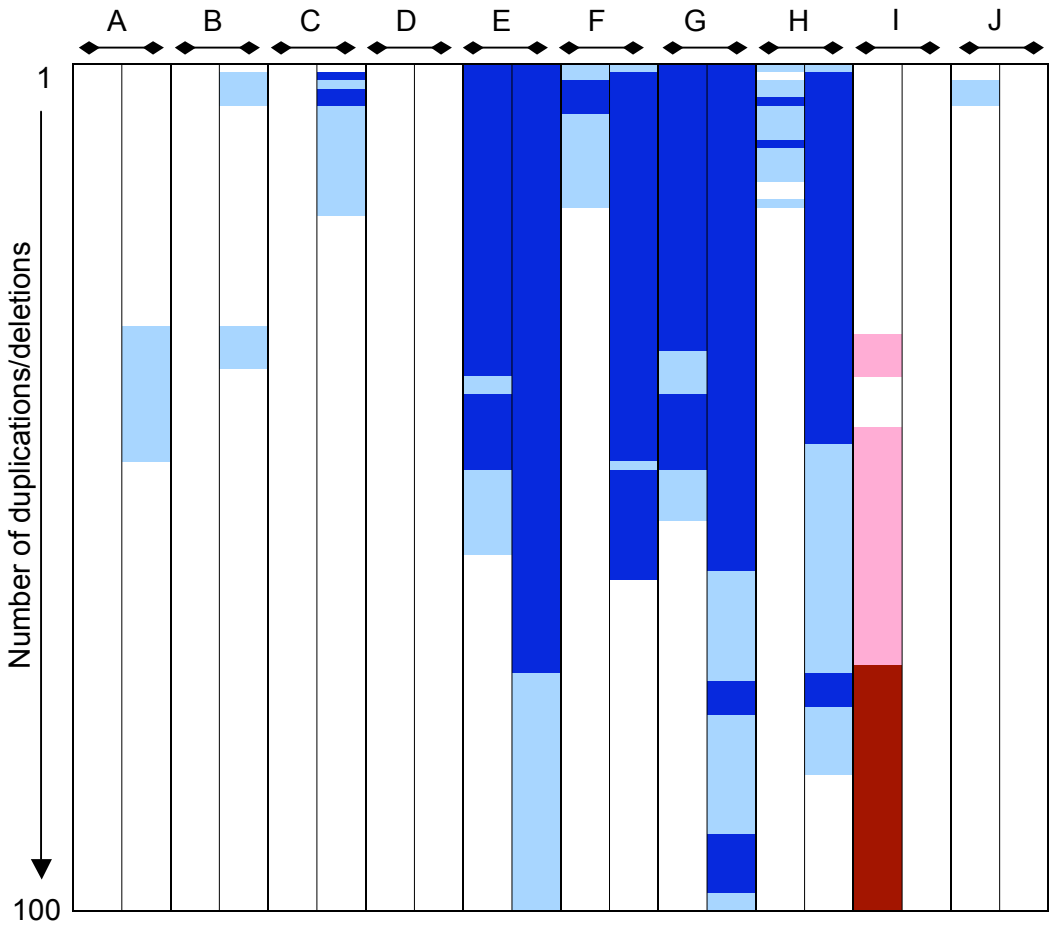

Figure 4.21. Under- and over-representation of human orthologues of candidate cancer genes in regions of copy number variation (CNV). (A) Genes nearest to CISs with *P*<0.001. (B) Genes nearest to CISs with *P*<0.05. (C) Genes with insertions within the coding region. (D) Genes with insertions but no singletons in the coding region. (E) Genes with insertions within the translated region. (F) Genes with insertions but no singletons in the translated region. (G) Genes with insertions in the transcribed region. (H) Genes with insertions but no singletons in the transcribed region. (I) Known oncogenes. (J) Known tumour suppressor genes. For each gene list, the left-hand column represents the significance of the association between the genes and CNV duplications, with rows representing the number of duplications, increasing in increments of 1 to a maximum of 100. Each box in the right-hand column represents the significance of the association between the genes and CNV deletions. *P*<0.01, dark blue for under-representation and dark red for over-representation; *P*<0.05, light blue for under-representation and pink for over-representation.

Gene list	Deletions	Amplicons
ORF only	3.76E-04	0.442
ORF only (no singletons)	6.12E-02	0.290
Translated region only	2.70E-10	0.780
Translated region only (no singletons)	7.02E-06	0.747
Transcribed region only	2.77E-14	0.082
Transcribed region only (no singletons)	6.68E-09	0.178
CIS nearest P<0.05	2.07E-04	5.12E-03
CIS nearest P<0001	2.09E-04	0.229

Table 4.9. *P*-values for the co-occurrence between genes from each gene list within CNVs and regions of copy number change in human cancer cell lines. "Deletions" gives the *P*-values for the co-occurrence of genes in CNV deletions and deletions of copy number less than or equal to 0.6 in human cancers, while "Amplicons" gives the *P*-values for the co-occurrence of genes in CNV duplications and amplicons of copy number greater than or equal to 2.7 in human cancers. *P*-values were calculated using a 2-tailed Fisher Exact Test. All significant *P*-values in "Deletions" represent an under-representation of genes in both CNVs and cancer deletions, while the significant *P*-value in "Amplicons" represents an over-representation of genes in CNVs and cancer amplicons.

representation in all lists of genes that co-occurred in both CNV deletions and deletions in human cancers. There was no association between genes in CNV duplications and copy number gains in human cancers, except for genes nearest to CISs with $P<0.05$, for which more genes than expected co-occurred in CNVs and amplicons. Again, this suggests that amplification of these genes in the general population may confer a predisposition to the development of cancer.

4.8 Discussion

The most significant finding from this chapter is that retroviral insertional mutagenesis is relevant to the discovery of cancer genes in regions of copy number change in human cancers. As anticipated, the overlap is stronger between candidate oncogenes and regions of copy number gain than between candidate tumour suppressor genes and regions of copy number loss. This partly reflects the fact that retroviral insertional mutagenesis predominantly identifies oncogenes due to the major mechanisms by which the retrovirus mutates genes and the requirement for both copies of a tumour suppressor gene to be mutated (see Section 3.4). It may, however, facilitate the identification of haploinsufficient tumour suppressor genes, for which the deletion of one gene copy can contribute to cancer. The other reason for the weaker association between tumour suppressor genes and deletions is that all genes that contained at least one insertion within the transcribed, translated or coding region were included in the analysis. Firstly, this can result in the inclusion of oncogenes that are activated by intragenic truncating mutations (see Section 3.4) and, secondly, many of the insertions may have occurred randomly and may not contribute to oncogenesis. However, the kernel convolution-based method for identifying CISs (de Ridder *et al.*, 2006, see Section 2.10.2) is biased towards oncogenes because insertions within many parts of a tumour suppressor gene may cause its inactivation and therefore insertions may not cluster into tight CISs. For this reason, including all genes provides a more comprehensive list of candidates for a role in tumour suppression.

Significantly, CIS genes were over-represented in amplicons from both haematopoietic and lymphoid cell lines and lines derived from solid tumours. This demonstrates that retroviral insertional mutagenesis is relevant to the discovery of cancer genes in cancers other than lymphomas. This is also proven in the identification of individual candidates, since many were amplified or deleted in a range of cancer types, and some, including

MEIS1, *MMP13* and *ACCN1*, were amplified or deleted in cancer types in which they had previously been implicated. While this study does not include any functional validation, the candidates include a considerable number of known and implicated cancer genes, demonstrating that the method is effective. In general, the discussion of individual genes has focussed on those for which there is some evidence, albeit sometimes limited, that gives cause for presenting the genes as potential oncogenes or tumour suppressor genes. However, the genes listed in Tables 4.4, 4.6 and 4.7 provide a large number of novel candidates that may be of interest to the cancer community. Interestingly, candidate cancer genes were under-represented in CNVs in apparently healthy individuals, further suggesting that amplification and/or deletion of these genes can have a detrimental effect on the cell and, in turn, on the individual.

Despite the promising results, there are a number of potential limitations associated with the analysis. Firstly, all of the human cancers were cell lines, rather than primary tumours. Cancer cells cultured *in vitro* lack the microenvironment of the tumour from which they are derived. While this means that they may not be fully representative of the original tumour, the homogeneity of cell lines can be an advantage since it prevents contamination by stromal cells and potential dilution of the copy number changes identified by CGH. It is, however, possible that the phenotype and genotype of cancer cell lines may differ from those of the original tumour due to genomic instability. Gene expression profiling of lung tumours and cell lines has demonstrated that, in culture, adenocarcinomas progress towards poorly differentiated phenotypes with expression profiles similar to those for squamous cell and small cell lung carcinomas (Virtanen *et al.*, 2002). However, comparisons of human breast and lung cancer cell lines and their corresponding tumours demonstrated an extremely high correlation for both genotype and phenotype, concluding that cell lines from both cancer types are suitable model systems for the original tumours (Wistuba *et al.*, 1998; Wistuba *et al.*, 1999). In addition, gene expression profiles for the NCI60 cell lines, which are the most commonly used cancer cell lines in cancer research and constitute a proportion of the cell lines used in this chapter, also showed that most were representative of their corresponding tumour types (Wang *et al.*, 2006b). Therefore, the use of cancer cell lines is warranted in this analysis, especially as the study is generally concerned with the number of copy number changes affecting a gene, rather than the tissue specificity.

A second potential drawback is that the ploidy of the cancer cell lines is not known. None of the methods used for detecting copy number changes within CGH data can determine the ploidy, and yet aneuploidy is a common characteristic of cancers. Attempts were made to determine the ploidy of cell lines based on the copy numbers of merged segments since, for example, a triploid cell line should only have copy number gains of 1.33, 1.67, 2.00, 2.33, 2.67, and so on, while a tetraploid should have copy number gains of 1.25, 1.5, 1.75, 2, 2.25, and so on. However, the mean copy number ratios for segments are not accurate enough to reliably assign cancers to a particular state. Irrespective of the ploidy, a copy number ratio of 3 indicates that there is a 3-fold increase in the number of copies. In this study, it is assumed that the balance of genes is more important than the actual number, i.e. a 3-fold increase in the number of copies of an oncogene is expected to have the same effect whether the baseline copy number is 2 or 4 genes. In addition, this study is concerned less with the exact copy number of genes, and more with whether genes are amplified or deleted, and the use of a set of copy number thresholds, rather than just one for amplification and one for deletion, ensures that as many candidates as possible are identified.

The analysis does not determine whether an amplified or deleted gene is significantly recurrent. However, genes that are only amplified or deleted in a single cell line may be biologically relevant, as demonstrated for *MEIS1*, and as many different cancer types were used in the analysis, tissue-specific amplicons and deletions may not be significantly recurrent across all cell lines. A gene for which there is no evidence of a role in cancer may not be a convincing candidate if it is amplified or deleted in a single cell line, but the presence of retroviral insertions within the mouse orthologue provides further support. For all candidates, the number of amplicons or deletions containing the gene and the number of additional genes in the minimal amplified or deleted region are provided to help in assessing the contribution of a gene to tumourigenesis. In Chapter 5, efforts are made to make it easier to identify the most promising candidates by ranking genes and assigning a *P*-value based on the number of samples in which they are amplified or deleted. In an attempt to filter out less promising candidates, any genes that were co-amplified with oncogenes or other mouse candidates were removed from the analysis, and yet co-amplified genes may co-operate in tumourigenesis (see Section 1.3.3.3). Nevertheless, given the number of candidates identified, it was considered more important to remove false positives, even at the expense of some "real" cancer genes.

As demonstrated in Section 4.5.1.4, some mouse candidates do not have human orthologues and are therefore excluded from the analysis. In some cases, the human orthologue may not have been identified, while in others, there may not be an orthologue in the human genome. However, the results of the analysis in Section 4.5.1.4 suggest that the proportion of human orthologues may be higher for "real" mouse candidates than for incorrectly assigned candidates. Any discrepancy in the number of mouse genes and the number of human orthologues does not affect the global comparison of Section 4.5.1, since the randomisation takes only mouse genes with human orthologues. This also prevents any introduction of bias resulting from the fact that only protein-coding genes have human orthologues, and that cancer genes are likely to be predominantly protein-coding. Another possible method for comparing the human and mouse data would be to map the insertion sites across to the human genome and then to assign the insertions to human genes. This could be achieved using the Ensembl Compara API, which enables the retrieval of genomic alignments between mouse and human. This would avoid the problem of lack of orthologues but there are many gaps in the alignment, which would prevent the precise mapping of a considerable proportion of insertions. To demonstrate, prior to mapping the retroviral insertions of Chapter 2 and 3 to the NCBI m36 mouse assembly, insertions were mapped to NCBI m34. Only 64.3% of insertions were successfully mapped across to the human genome (NCBI 35) using the Ensembl Compara API. A further drawback of mapping insertions could be that if there really is no human orthologue for a given mouse candidate gene, or there is a break in synteny between mouse and human, the insertions mapped to the human genome will be assigned to an incorrect gene.

The analysis is also limited by the resolution of the data. Efforts have been made to choose suitable boundaries for the ends of amplicons and deletions, but without increasing the density of the SNPs it is impossible to determine whether genes beyond the first or last amplified or deleted SNP are indeed amplified or deleted. It is also possible that small amplicons and deletions may be missed, while the high levels of noise in the data may also lead to regions of copy number change being missed or falsely identified. Encouragingly, the most successful methods for detecting changes produced similar outputs, and the fact that known and implicated oncogenes and tumour suppressor genes were identified, often in cancer types in which they have previously been shown to be disrupted, was also reassuring. However, in Chapter 5, a higher density SNP array is

used, and is compared to the 10K array to determine whether it represents a significant improvement.