SUPPLEMENTARY RESULTS

Substitution detection

We have previously analysed the coding exons of ~4000 genes (5Mb) from NCI-H209 by PCR and capillary sequencing (http://www.sanger.ac.uk/perl/genetics/CGP/cosmic?action=sample&id=688013). This identified 29 known single-base substitutions (supplementary table 6). The substitution algorithm recaptured 22 of these, for a *sensitivity* of 76% for known coding variants. Of the 7 that were missed, 3 were not identified because there were no reads containing the variant mapped to the given genomic position. This would be either because there were no reads from the variant allele by play of chance or, more likely, because the reads covering the variant allele mapped incorrectly elsewhere in the genome (so-called reference bias). A further 2 known mutations were missed because, although there were reads spanning the variant allele present in the data set, these were not of sufficient numbers to meet the pre-determined thresholds for the overall level of coverage at that base. One mutation was excluded by the algorithm for the reason that there was insufficient coverage of the normal genome at that position to determine whether the variant found in the tumour reads was somatic or germline. The final known substitution that was excluded was one where there was a read reporting the variant allele found in the sequencing data from the normal genome: whether this represents contamination of the normal DNA library by tumour DNA or two adjacent sequencing errors in colour space is unclear.

To assess the *specificity* of the algorithm for identifying somatic point mutations, we assessed by capillary sequencing a set of 79 coding substitutions and 354 randomly chosen genome-wide substitutions predicted by the algorithm (supplementary table 7). A total of 77 coding mutations and 333 genome-wide variant calls were confirmed as somatic substitutions of the type predicted. Thus, the true positive rate of the algorithm is 98% in coding regions and 94% in non-coding regions. Of the false positives, 6 were due to germline SNPs being miscalled as somatic variants because there were no reads from the polymorphic allele in the sequencing data from the normal genome. This is likely to be due to the stochastic nature of allele sampling in shotgun sequencing and fits with the simulations done for the power calculations (figure 1A). Of the

other mis-calls, neighbouring indels and other variants accounted for 9, and a further 8 had no clear explanation for the discrepancy between the SOLiD and capillary sequencing data sets.

Insertion and deletion detection

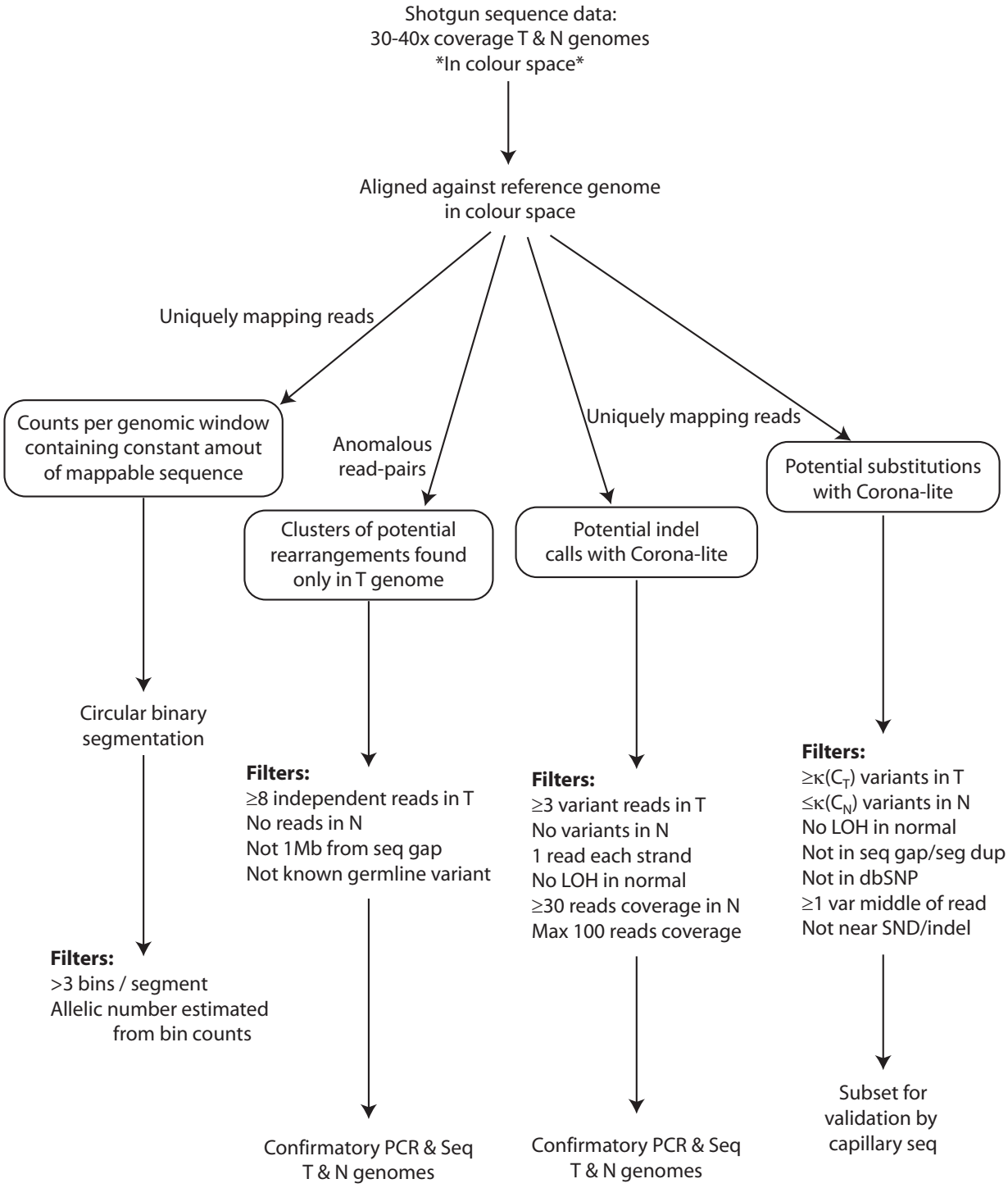
Small insertions (up to 3 bp) and deletions (up to 11 bp) were called using corona-lite (version 0.4). Indels found in the tumour and not in the normal were further filtered to require (i) minimum three supporting tumour reads, (ii) minimum one read on each strand, (iii) no LOH in the normal, (iv) maximum 100X coverage (to remove regions of misalignment) and (v) minimum 30X normal coverage (to reduce the number of germline indels in the set). This is a fairly stringent algorithm, designed to minimise the number of false positive calls. The reason for such stringency is that mapping reads spanning indel variants from paired 25bp DNA fragments presents major difficulties to the current generation of short-read aligners. Thus, there will be considerable reference bias, manifesting as a high reference:variant ratio for a heterozygous indel. Since it is likely that germline indels outnumber somatic indels by several orders of magnitude, extensive reference bias could potentially result in a large number of false positive calls of somatic variants due to failure to identify the variant-containing reads in the sequence data from the normal genome.

From the ~5Mb of the NCI-H209 genome sequenced by exon PCR and capillary sequencing, we had previously identified 2 coding indels (http://www.sanger.ac.uk/perl/genetics/CGP/cosmic?action=sample&id=688013). Neither of these was identified by the algorithm described above, confirming that our sensitivity for detecting somatic indels is low.

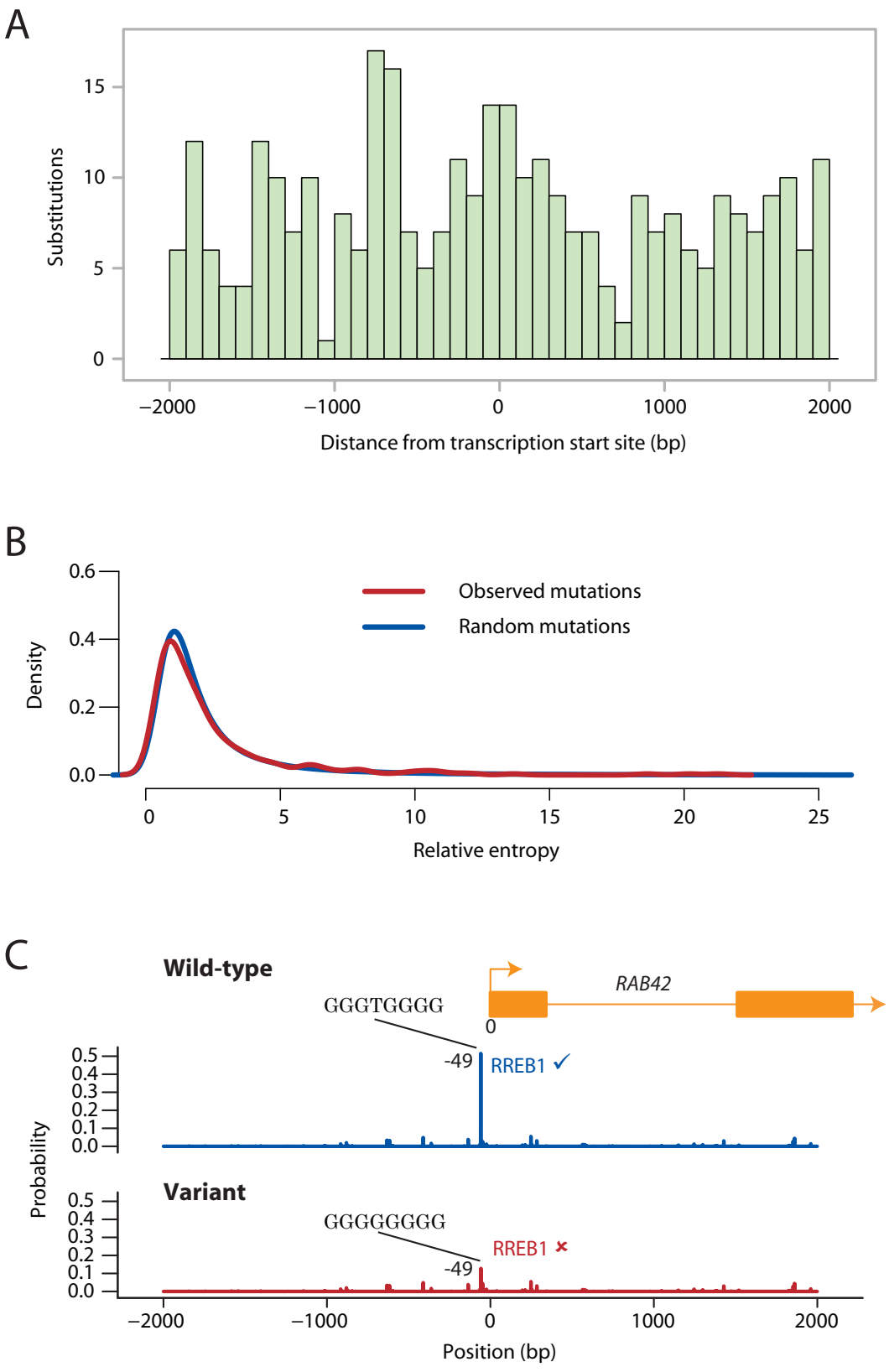
We took a set of 262 putatively somatic indels called by the algorithm for confirmatory capillary sequencing (supplementary table 8). Of these, 65 were confirmed as genuine somatic variants, with a resultant true positive rate of 25%. This suggests that, as expected, it is very difficult to reliably call indels from short-read data. Of the false positives, 113 (43%) were wild-type on capillary sequencing. Many of these were called at long (.5-6bp) tracts of identical nucleotides or microsatellite repeats, which might result either from polymerase slippage during library production or mis-ligation during sequencing. A total of 84 (32%) miscalls were due to germline

indels being called as somatic. This underscores the difficulty described above of identifying genuine somatic indels in situations where there is extensive reference bias coupled with a large excess of germline polymorphisms in the genome.

Since the algorithm for identifying indels had such a low true positive rate, only those indels confirmed as genuine and somatically acquired by capillary sequencing are reported in this paper (table 1; supplementary table 2).

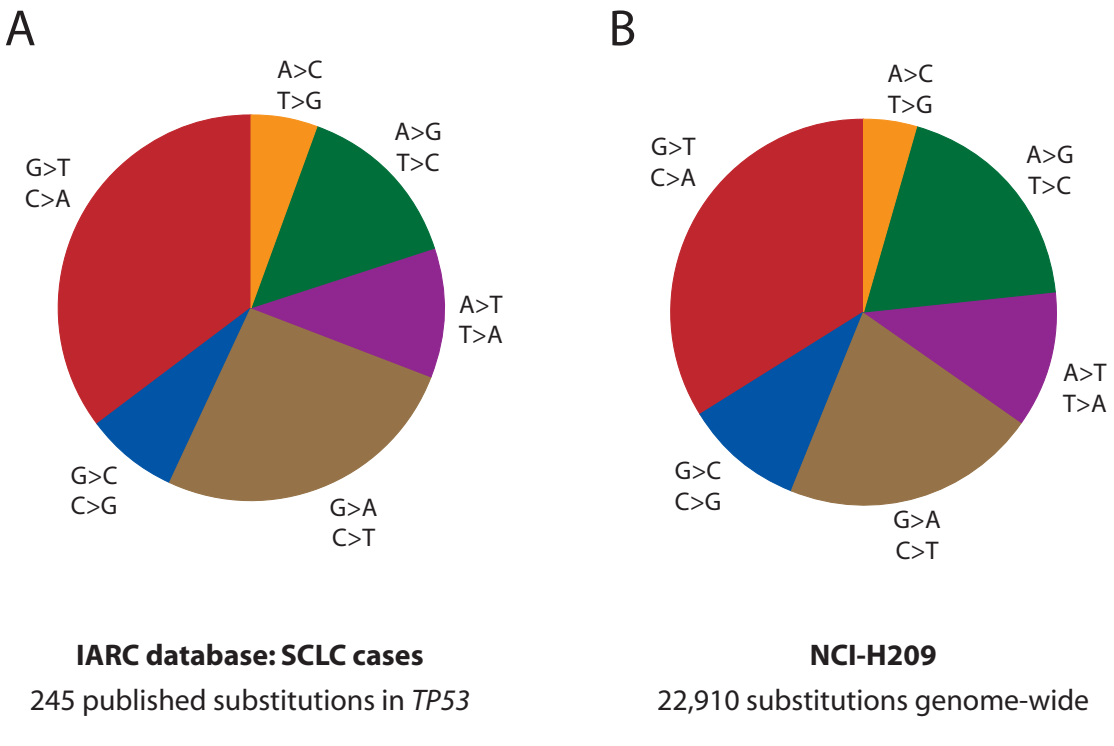


Supplementary figure 1. Outline of bioinformatic algorithms for identification of somatically acquired genetic variants of all classes. T, tumour genome; N, normal genome; seq, sequencing; var, variant; $\kappa(C_T)$, threshold for number of reads containing variant allele required in tumour genome sequence data, for given level of coverage at that base; $\kappa(C_N)$, maximum number of reads containing variant allele in normal genome sequence data at that base.



Supplementary figure 2. Analysis of somatic substitutions falling in regions predicted to contain the promoters of protein-coding genes in the Ensembl database. (A) Number of substitutions falling in 100bp bins for the 2kb either side of gene transcription start sites in the Ensembl database. (B) Comparison of the predicted effects on transcription factor

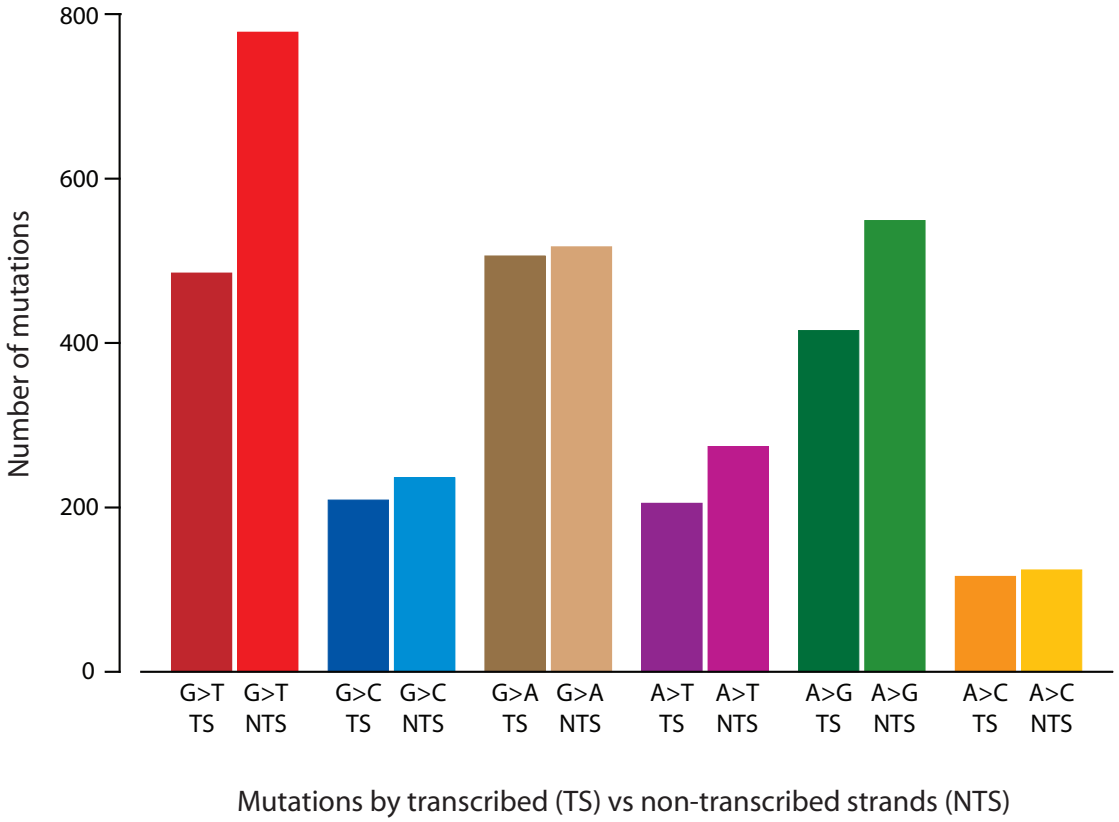
binding sites of observed mutations (red) and random sets of mutations (blue). Relative entropy denotes the magnitude of the predicted effect of the variant on the likelihood of transcription factor binding. (C) Example of a mutation predicted to obliterate a binding site for the RAS-responsive transcription factor, RREB1. The consensus binding sequence for RREB1 is (CCCCACCA / TGGTGGGG), and a putative element is seen 49bp downstream of the transcription start site of the gene *RAB42*. A T>G somatic substitution in NCI-H209 is predicted to reduce the potential for RREB1 to bind to the site with high confidence ($p=3x10^{-98}$).



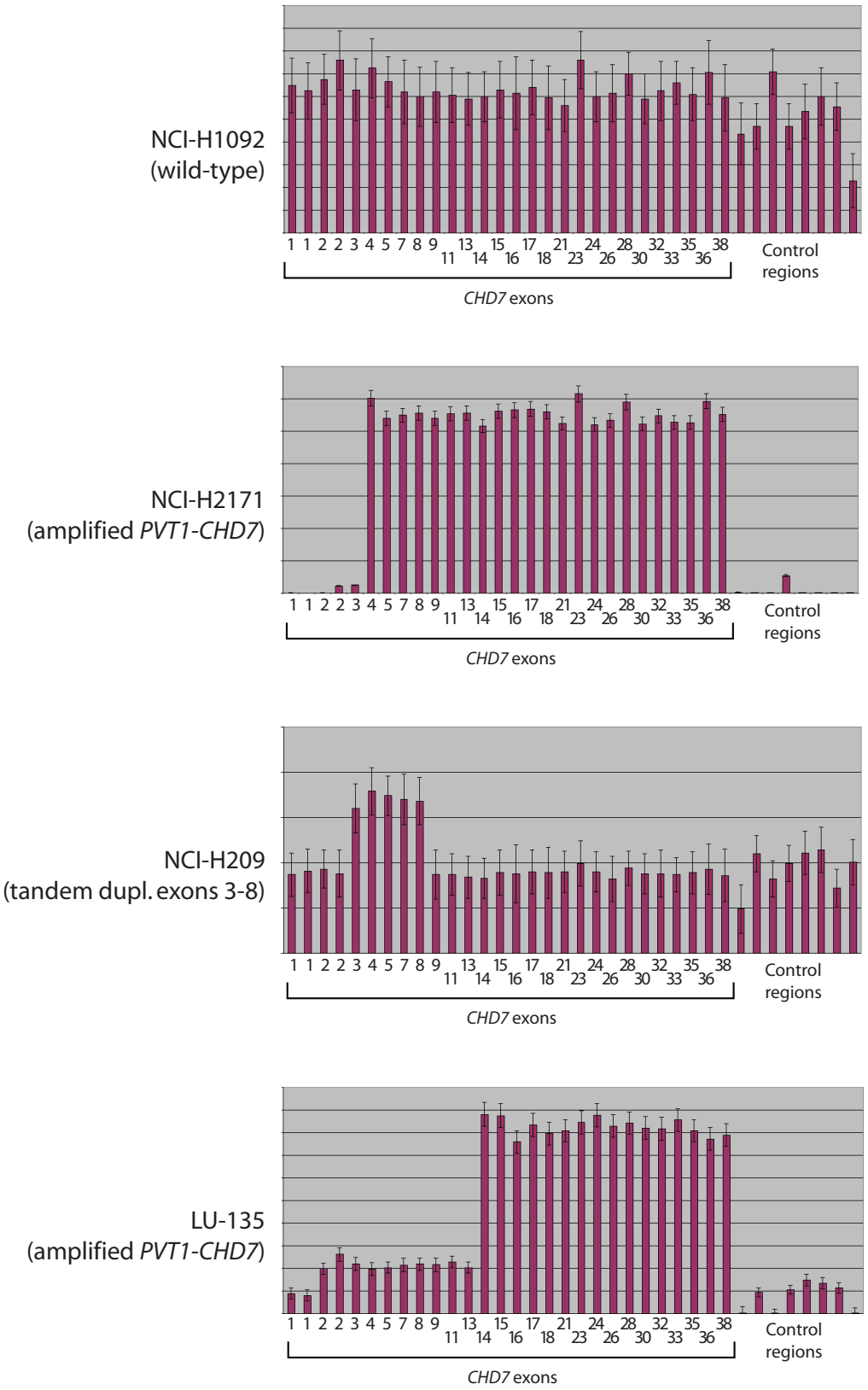
IARC database: SCLC cases
245 published substitutions in *TP53*

NCI-H209
22,910 substitutions genome-wide

Supplementary figure 3. Comparison of the mutation spectrum between (A) 245 published substitutions in *TP53* in SCLC cases recorded by the IARC database (http://www-p53.iarc.fr/) and (B) 22,910 genome-wide somatic substitutions in NCI-H209.



Supplementary figure 4. Strand bias in mutation prevalence across transcribed regions of the NCI-H209 genome. The numbers of mutations of each of the 6 classes of guanine and adenine mutations are shown, split by whether they occur on the transcribed (TS) or non-transcribed strand (NTS), as defined by the Ensembl transcript database.



Supplementary figure 5. Multiplex ligation-dependent probe amplification (MLPA) of *CHD7* in SCLC cell lines. Each histogram shows the normalised signal intensity for probes in exon order across the gene, followed by probes from control genomic regions. The first histogram is a representative example from a cell line without evidence for exon copy number variation. The following three plots represent the three cell lines in which rearrangements of *CHD7* were identified.

SUPPLEMENTARY TABLE LEGENDS

Supplementary table 1. Genome-wide list of 22,910 somatic substitutions identified by bioinformatic analysis of sequencing data from NCI-H209. Chromosome and base-pair are shown (by the NCBI36 genome build), together with the reference base at that position, the predicted variant, the zygosity and the sequence context (10bp either side of and including the variant base).

Supplementary table 2. Genome-wide list of 65 somatically acquired small insertions and deletions identified by bioinformatic analysis of short-read sequencing data from NCI-H209, all subsequently confirmed as genuine, somatic changes by capillary sequencing. Chromosome and base-pair (in a range where necessary) are shown according to the NCBI36 genome build, together with the zygosity, the predicted variant, status and the sequence context (10bp either side of and including the variant base).

Supplementary table 3. List of 58 somatically acquired genomic rearrangements in NCI-H209. All structural variants have been confirmed by PCR across the breakpoint, with bidirectional sequencing confirming the segments involved. All but two have had the breakpoint annotated to base-pair resolution: for the other two, we provide a range of genomic positions encompassing the breakpoints. Length and sequence of either microhomology or non-templated sequence at the junction are shown.

Supplementary table 4. List of copy number segments across the genome for NCI-H209. The copy number changes associated with identified somatically acquired genomic rearrangements are marked.

Supplementary table 5. Table of somatically acquired substitutions and indels in coding sequence. The mutations, genomic positions and genes are shown. Predicted protein consequences are shown based on the Ensembl transcript quoted. The validation status refers to whether the variant has been confirmed by capillary sequencing or whether no confirmatory sequencing was attempted.

Supplementary table 6. Sensitivity of bioinformatic algorithm for identifying known somatic single-base substitutions in NCI-H209. From capillary sequencing of ~4000 genes (5Mb), we previously identified 29 somatic single-base substitutions. Of these, 22 (76%) were correctly identified independently from the SOLiD sequencing data.

Supplementary table 7. Specificity of bioinformatic algorithm for identifying somatic single-base substitutions in NCI-H209. A set of 79 novel coding substitutions and 354 randomly chosen genome-wide variants called by the algorithm were tested by capillary sequencing to ascertain veracity. Of these, 77 (97%) of the coding substitutions and 333 (94%) of the random variants were confirmed as genuine somatic mutations. The true and false positives are shown, together with possible explanations, where identified, for the false positives.

Supplementary table 8. Specificity of bioinformatic algorithm for identifying somatic indels in NCI-H209. A set of 262 novel indel calls made by the algorithm were assessed by capillary sequencing, of which 65 (25%) were confirmed as genuine somatic indels.