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

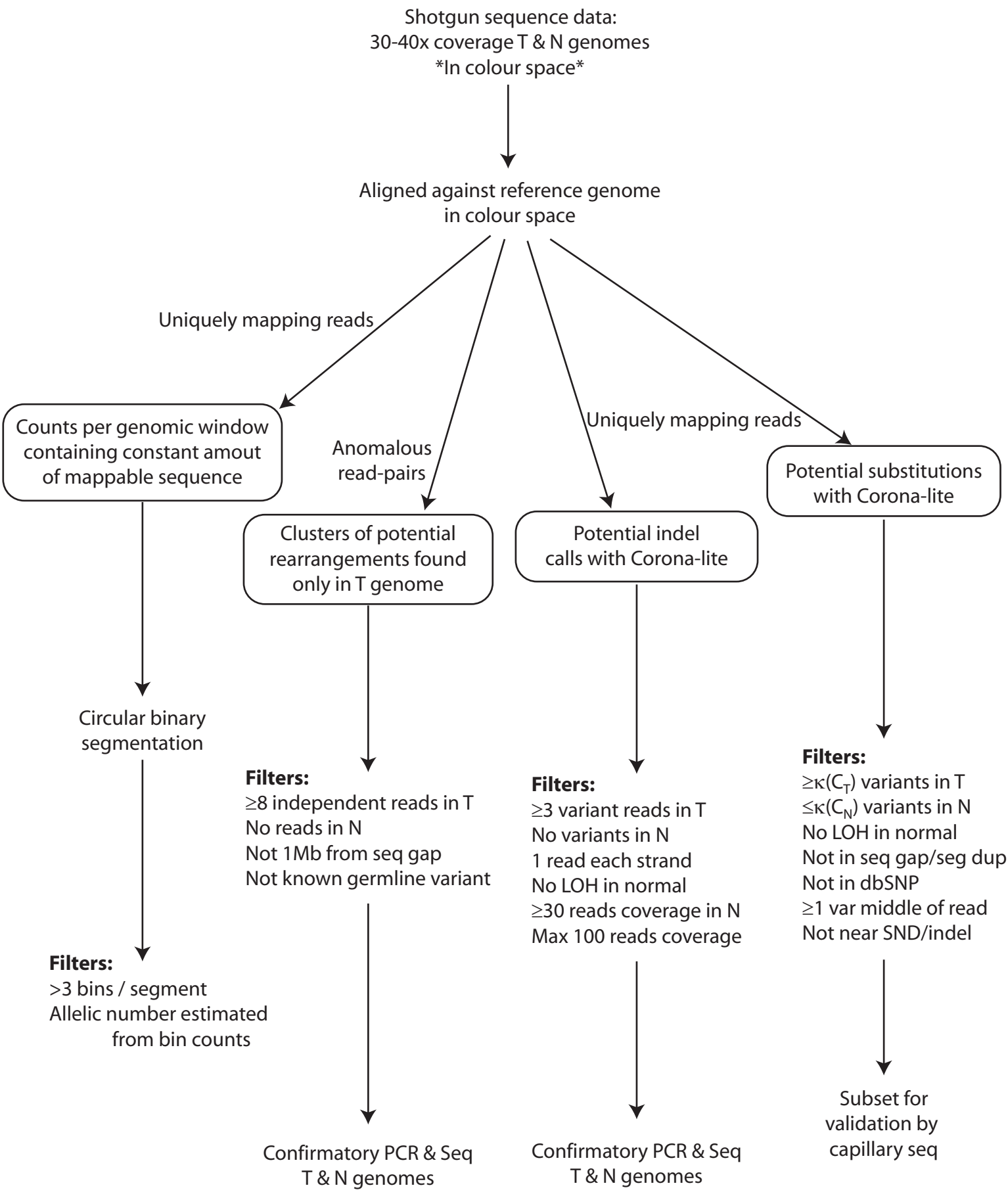
Small insertions (up to 3 bp) and deletions (up to 11 bp) were called using corona-lite (version 0.4). Indels found in the tumour and not in the normal were further filtered to require (i) minimum three supporting tumour reads, (ii) minimum one read on each strand, (iii) no LOH in the normal, (iv) maximum 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.

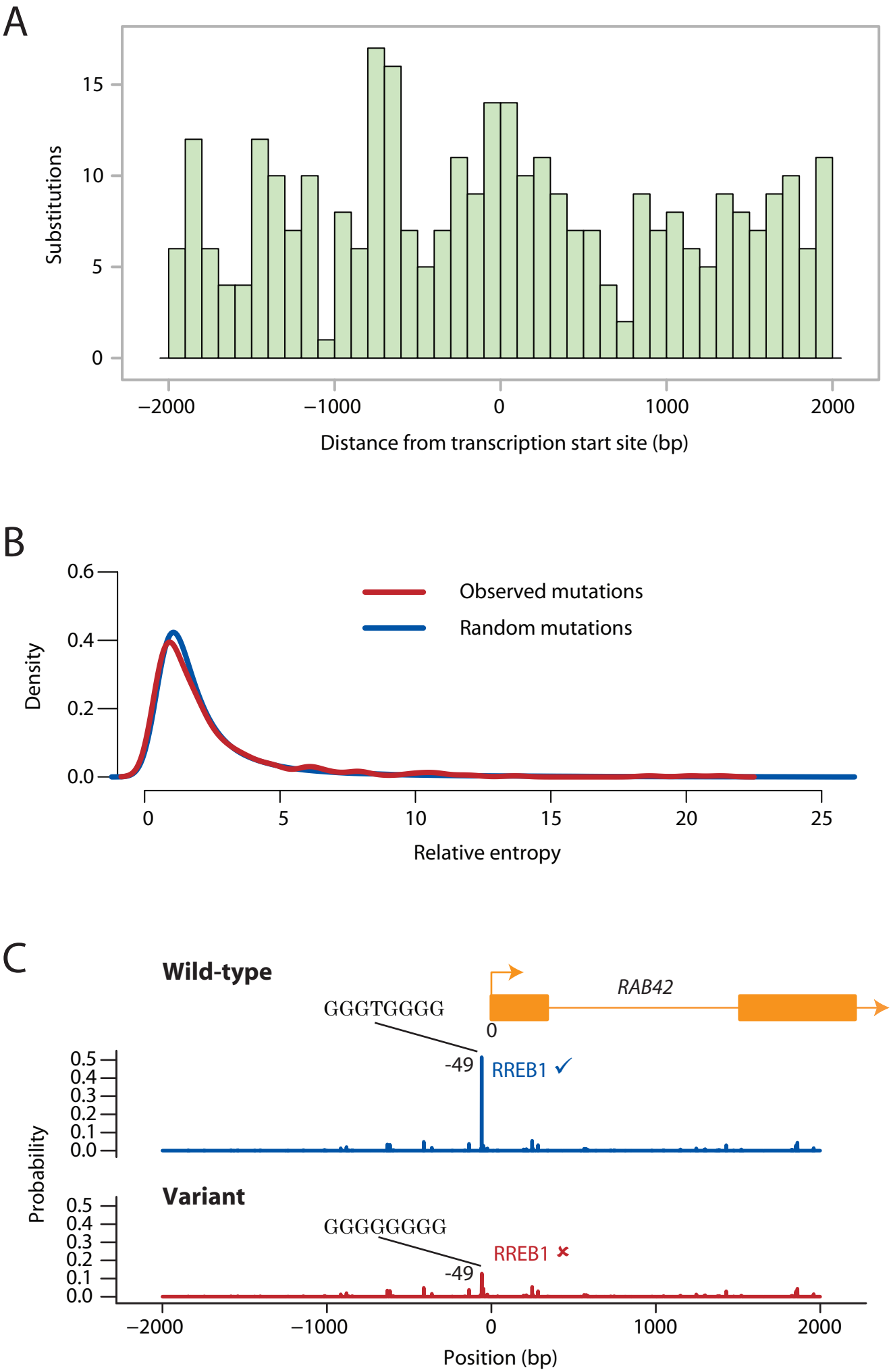
We took a set of 262 putatively somatic indels called by the algorithm for confirmatory capillary sequencing (supplementary table 8). Of these, 65 were confirmed as genuine somatic variants, with a resultant true positive rate of 25%. This suggests that, as expected, it is very difficult to reliably 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).

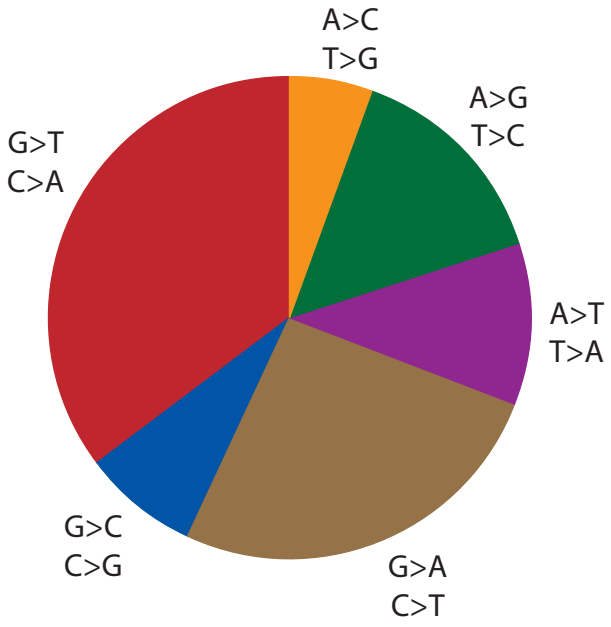


Supplementary figure 1



Supplementary figure 2

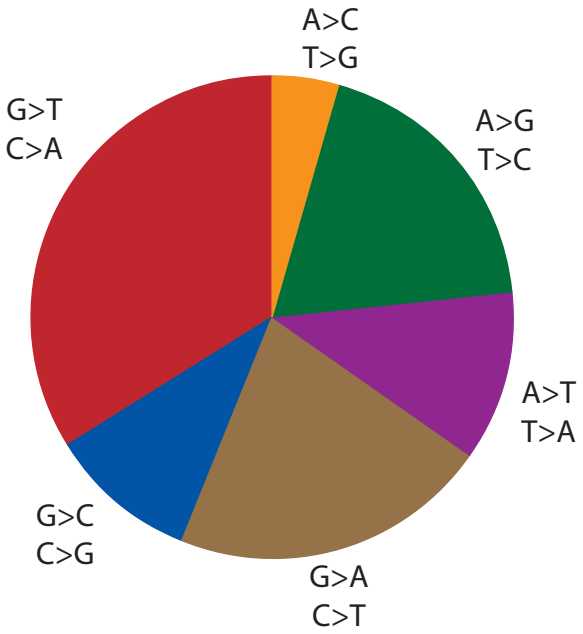
A



IARC database: SCLC cases

245 published substitutions in *TP53*

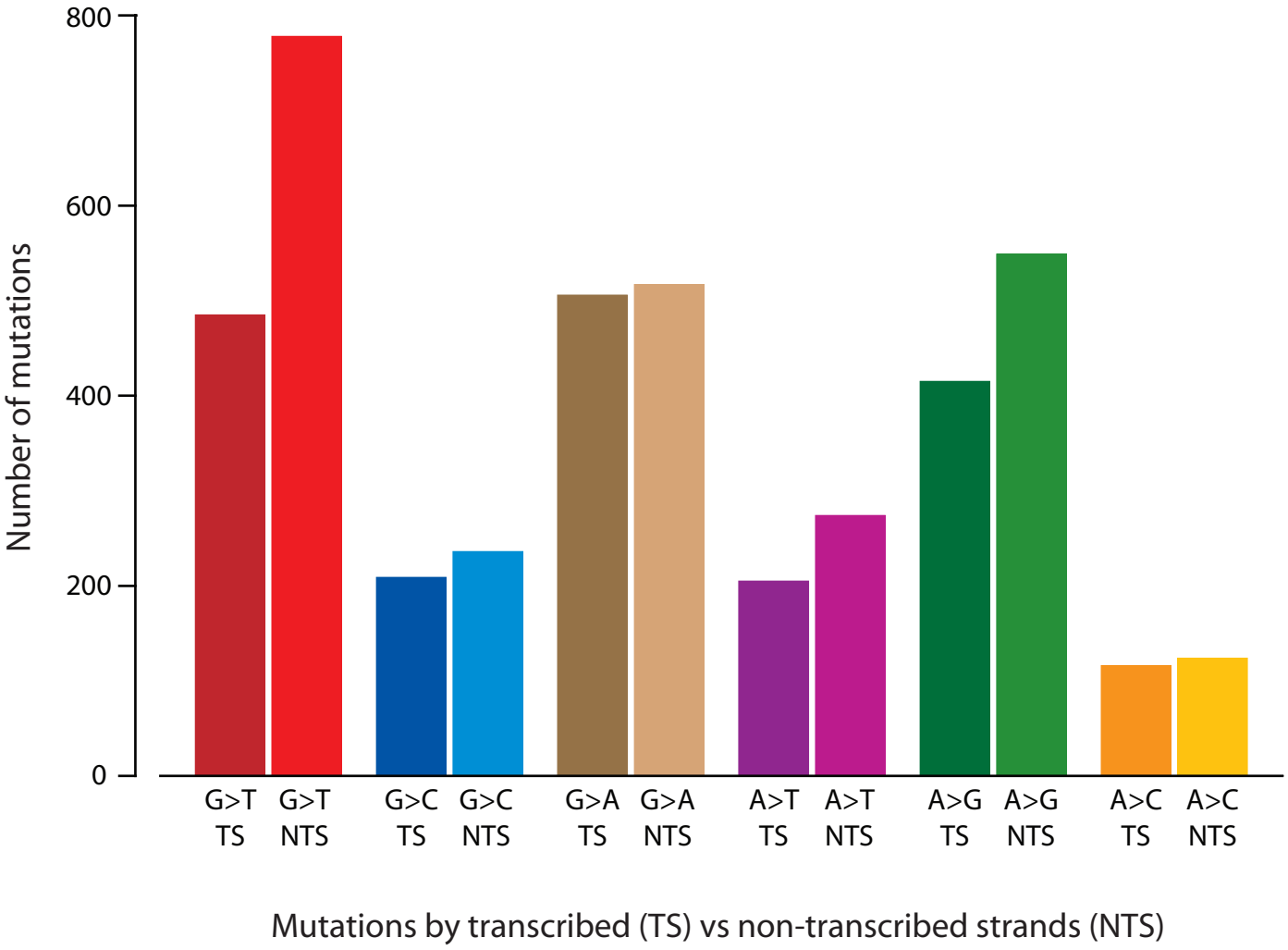
B



NCI-H209

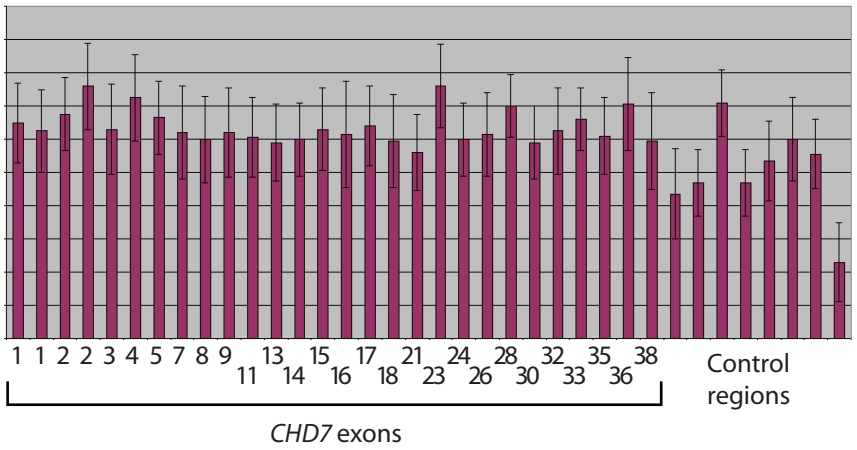
22,910 substitutions genome-wide

Supplementary figure 3

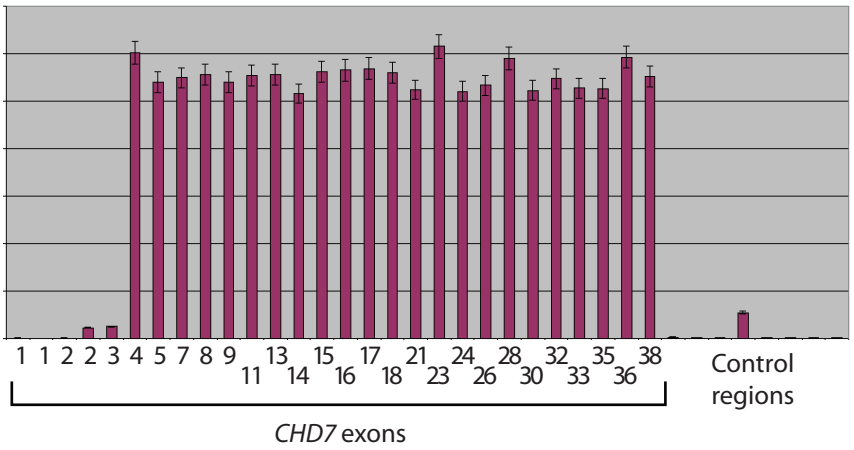


Supplementary figure 4

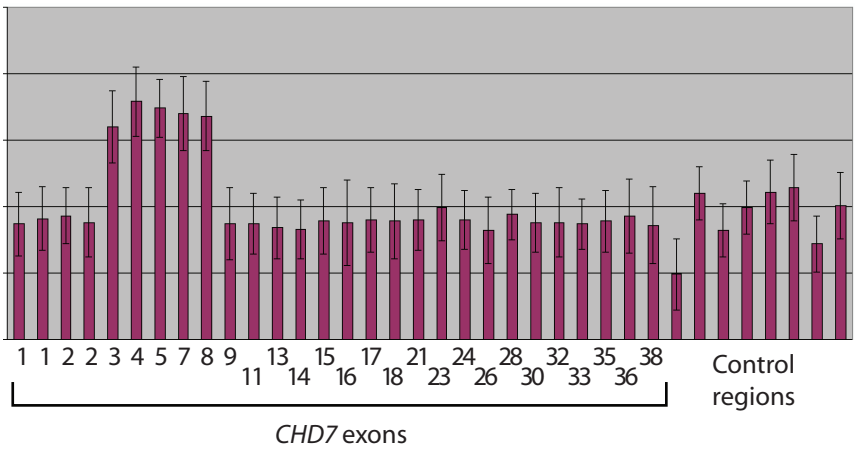
NCI-H1092
(wild-type)



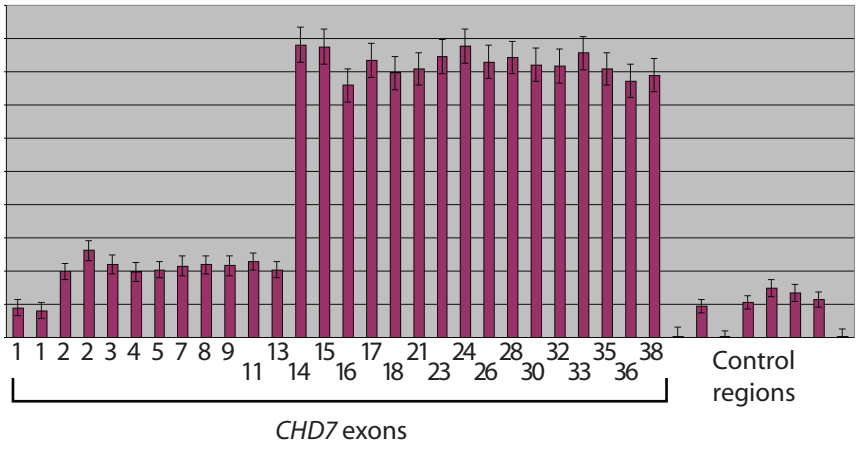
NCI-H2171
(amplified *PVT1-CHD7*)



NCI-H209
(tandem dupl. exons 3-8)



LU-135
(amplified *PVT1-CHD7*)



Supplementary figure 5