Supplementary Note

1. Germline *APOBEC3A* and *APOBEC3B* deletion allele

The germline deletion allele is believed to involve a segment of sequence flanked by 350bp of sequence homology between APOBEC3A (coordinates 22:39358281-39358630) and APOBEC3B (coordinates 22:39388217-39388566). The 350bp of sequence homology includes fifteen base pairs constituting the five terminal amino acids of the final exon in both proteins and continues to 130bp into the 3'UTR of both *APOBEC3A* and *APOBEC3B*. In the deletion allele, only a single copy of the 350bp sequence remains with the 29,500bp intervening sequence between them removed. It is therefore impossible to be precise regarding the absolute breakpoints of this deletion allele within this 350bp window of sequence homology. Hence, some uncertainty remains regarding what constitutes the final transcript: whether it is the terminal exon of *APOBEC3B* and all its 3'UTR or whether it is the terminal exon of *APOBEC3A* and part of the 3'UTR of *APOBEC3B* that makes up the protein. Either way, at least part of the *APOBEC3B* 3'UTR is conveyed to genomic *APOBEC3A* and may come under its regulation. Nevertheless, the predicted protein resembles the APOBEC3A protein.

2. Detection of germline *APOBEC3A/3B* deletion polymorphism

2.1 Principle of detection of the large germline deletion polymorphism at the *APOBEC3A/3B* locus

The critical region of the germline deletion polymorphism was reported to be in the region of chr22:39,363,619-39,375,307 (hg19) based on 24 probes on the Affymetrix SNP6.0 array[5,7]. This approximately 29kb deletion is at the limit of resolution for detection of copy number polymorphisms (CNPs) using genomic arrays. It had been previously highlighted through analysis of data from the International HapMap Project that no single existing tag single nucleotide polymorphism (SNP) served as an effective surrogate for the deletion variant[5], emphasizing the need for direct genotyping of the CNP. However, based on 1000 Genomes Project Asian data, the *APOBEC3A/3B* deletion allele was shown to be in strong LD with SNP rs12628403 (r^2 =0.91)[6]. A WTCCC study seeking CNP associations with breast cancer as well as many other diseases was also unable to define a tagging SNP[6] in relation to this deletion polymorphism. Nevertheless, in order to detect this deletion reliably in next-

generation sequencing data, sequence coverage at thirty loci within the critical region and thirty loci immediately flanking the critical region were sampled from the matched normal BAM files resulting in 90 sampled loci in total (Supplementary Fig 2A, Supplementary Table 2A) for whole-genome sequenced (WGS) samples. For exome-sequenced samples (WES), 45 loci were sampled within and flanking the CNP to allow for variation in capture efficiency. Consistent failure of capture occurred at some loci which were excluded, resulting in total sampling of 35 loci within the CNP and 80 loci flanking it (Supplementary Table 2A).

If a matched normal BAM file was unavailable, a tumor BAM file was used. However, matched normal BAM files were favoured for calling this germline deletion allele (the CNP) because of the possibility of changes of ploidy that can occur in tumors. If normal and tumor BAM files were unavailable, suppressed, redacted or corrupted, then the sample was excluded from further analysis. The source of CNP calling for each sample included in this study is provided in Supplementary Table 1B (column 12: "tumor" sourced from tumor sample, "normal" sourced from normal sample). In total, the *APOBEC3A/3B* polymorphism detection was sourced from 561 tumors (99 BLCA, 117 BRCA, 1 CESC, 19 HNSC, 2 KIRP, 303 LUAD, 12 STAD, 2 THCA and 6 UCEC) and 2158 normals.

The mathematical method of calling the allelic status has been described in the Online Methods section.

2.2 Robustness of method used for detection of the deletion polymorphism

In order to examine the reproducibility of our method, we sought the concordance of copy number polymorphism detection between tumor and normal samples from the same individuals. Tumor and normal DNA libraries from the same patient are independently prepared and are therefore distinct experiments. Concordant detection of the germline polymorphism in these separate preparations provides independent verification of the carrier status in the patient. We checked the concordance of the detection of the polymorphism when obtained from tumor and normal BAM files from each of 123 whole-genome sequenced (WGS) patients and 166 exome-sequenced (WES) patients. 100% concordance was seen in polymorphism detection between tumor and normal BAM files of WGS samples. These WGS libraries were made using the Illumina no-PCR protocol and therefore do not suffer amplification-related complications. 97.6% concordance was seen between tumor and normal BAM files of exome-sequenced samples (4 out of 166 were

discordant calls). Exome library preparation involves an amplification step, which may diminish the difference between copy number statuses (Supplementary Table 2C).

We next looked at samples, which were both WGS and independently exome sequenced. The sequencing reads generated in WGS and WES are different in their parameters. WGS insert sizes tend to be longer (~400-700bp) with sequencing reads of 100bp in length. In contrast, WES insert sizes are shorter (~300-450bp) with 75bp sequenced reads. Mapping parameters will therefore be different between these BAM files. There were only ten such samples. However, concordance was 100% in detection of the germline *APOBEC3A/3B* CNV for samples that had been WGS and WES (Supplementary Table 2C).

As mentioned previously, sampling of normal BAM files was preferred and sought above sampling tumor BAM files because of the possibility of changes of ploidy that occurs in tumors. In theory, true discordance between CNV detection in the tumor and normal could occur if loss of heterozygosity of chromosome 22 arose in the tumor of the wild-type parental allele in a patient who is heterozygous for the deletion allele. This would be apparent as a reduction in copy number from being a heterozygous carrier of the germline deletion allele to being homozygous in the tumor. This occurred in one out of the 166 samples (PD5873a/PD5873b) in this dataset (Supplementary Table 2C). This also occurred in a cell line, HCC38a/HCC38b that was analyzed concurrently (but not included in this study formally because it was not primary cancer). This discordance was observed in BAM files, detected by the informatics method and confirmed by PCR in the tumor and normal (Supplementary Figure 2C).

Furthermore, in an analysis of expression levels of various members of the APOBEC family of enzymes, we find that patients who are homozygotes for the deletion polymorphism lack APOBEC3B expression and heterozygotes have reduced expression of APOBEC3B (Supplementary Figure 5A). This provides independent, supporting evidence of the copy number polymorphism status.

3. The relationship between the *APOBEC3A/3B* germline deletion allele and somatic mutational signatures in cancer

The germline *APOBEC3A/3B* deletion polymorphism locus was previously reported to be a modest breast cancer susceptibility risk allele. Given the speculation that the APOBEC family of enzymes may be involved in generating Signatures 2/13, we sought a relationship between the germline deletion allele and the somatic mutation. In this section, we describe the intriguing findings.

3.1 Carriers of at least one copy of the germline deletion allele show a higher mutation rate of Signatures 2/13 in breast cancer patients

By grouping breast cancer patients according to whether they were carrying at least one copy of the germline deletion allele or not, we found that breast cancers derived from people with at least one copy of the germline deletion allele had a *higher* mutation burden of Signatures 2/13. The dataset comprised genome-sequenced (123) as well as exome-sequenced (800) cancers. In order to perform the analyses, the rate of mutation was calculated for each cancer (rate of Signature 2/13 per Mb), which effectively corrects for whether the samples had been genome- or exome-sequenced.

Because the rates of Signatures 2/13 were not normally distributed (Supplementary Figure 5A: QQ plots), a one-sided Wilcoxon rank-sum test was performed to see whether carrying one copy of the deletion allele had an overall effect on the mutation rate of Signatures 2/13. We found that breast cancer patients carrying at least one copy of the germline deletion allele had higher rates of Signatures 2/13 ($p=2.70e^{-3}$) in their cancers (Supplementary Figure 5B).

3.2 In contrast, Signatures 1A/1B do not show an association with the *APOBEC3A/3B* deletion allele in breast cancer patients

If it is true that there is a biological relationship between the germline *APOBEC3A/3B* deletion allele and Signatures 2/13, then the deletion allele should not show any association with other mutational signatures. Pervasive signatures seen across many different cancers are Signatures 1A/1B. One-sided Wilcoxon rank sum tests were carried out on the set of 923 breast cancer patients, to test whether samples with at least one copy of the deletion allele had a significantly higher rate of signature 2/13 mutations than those samples without the deletion allele. We found no correlation between the germline *APOBEC3A/3B* deletion allele status and the rate of Signatures 1A/1B (p=0.9354).

3.3 Increasing the power of the analysis: Including more cancers

In order to increase the power of the analysis, we sought to include more cancer samples. There were no further available breast cancer samples with BAM files ready for download, hence we sought inclusion of other cancer types that had previously been analyzed[9,18]. However, there were two factors that we took into consideration. First, that the distribution of rates of Signatures 2/13 varied considerably between cancer-types (Supplementary Figure 3C). Second, there were clear outliers in all the cancer types skewing the distribution of mutation rates (Supplementary Figure 3D).

3.4 Average rates of signatures 2/13 varied considerably between cancers and outliers existed within each cancer-type

It was observed that the distribution of mutation rates of Signatures 2/13 was different between cancers. Overall, cancers such as bladder, cervical, head and neck, lung adeno, lung squamous, stomach and uterine carcinomas had higher average rates of Signatures 2/13. In contrast, cancers such as breast cancer had an overall lower rate of Signatures 2/13 mutagenesis (Supplementary Figure 3C). It is envisaged that other factors are likely to be influencing the occurrence of Signature 2/13 mutagenesis in different tissue-types. Pooling the datasets would therefore not be biologically appropriate and could simply dilute what would be only a modest signal.

In addition, all cancer types demonstrated striking outliers that skewed the distribution of mutation rates. The non-normally distributed nature of Signature 2/13 rates for each cancer subtype (Supplementary Figure 3A) are reflected in mean rates which were almost always significantly higher than median rates of Signatures 2/13. Furthermore, we observed that the deletion allele was enriched amongst cancers that had a significantly higher fraction of mutations associated with Signatures 2/13 (Supplementary Figure 3E), or outliers. We therefore sought a method of identifying outliers in a more formal manner.

3.5 Identification of hypermutators (outliers)

Some cancers were observed to have a strikingly high proportion of total mutations associated with Signatures 2/13 and/or have higher rates of mutagenesis associated with this signature (Supplementary Figure 1C-1O, Supplementary Figure 3C). Using the rate of Signatures 2/13 mutagenesis, outliers were identified as patients with cancers that had a mutation rate exceeding 1.5 times the length of the interquartile range from the 75th percentile for each type of cancer[19]. These outliers will hitherto be referred to as “hypermutators” although we do not suggest that there is an on-going biological process

attached to this name. Given the considerable variation of the mutation rates for different cancer tissue-types (Supplementary Figure 3A-B), each cancer type was analyzed separately. A summary of the hypermutators versus non-hypermutators is provided in the Supplementary Table 3A.

3.6 An enrichment of the deletion allele is seen in hypermutators of Signatures 2/13 in breast cancers

We next performed an analysis comparing the prevalence of the deletion allele amongst hypermutators and non-hypermutators in the 923 breast cancer patients. We found an enrichment of the deletion allele amongst patients who had breast cancers with the hypermutator phenotype (Supplementary Figure 3E) and a trend was observed with the number of copies of the deletion allele (Cochrane-Armitage test, $p=6.251e^{-6}$, OR 2.37 CI 1.64-3.46). We extended this analysis across many individual cancer tissues. Although the sample sizes are much smaller in all the other cancer types, the trend was observed in ALL ($p=2.51e^{-5}$) and BLCA ($p=0.038$). We did not find a similar degree of enrichment for many other cancers (Supplementary Figure 3D). When analyzed in aggregate across 2,719 cancer samples, the trend remains (OR 1.65 CI 1.18-2.20, $p=9.505e^{-4}$) although this is likely to be driven by specific cancer types, chiefly BRCA which contributes the largest number of samples to the dataset.

3.7 The enrichment of the deletion allele is not seen in hypermutators of Signature 1A/1B, another common mutational signature present in human cancers

A similar "hypermutator" phenomenon is not observed for Signatures 1A/1B (Supplementary Figure 1C-1O) at least for the cancers analyzed so far. Nevertheless, we attempted to identify outliers using the same method as applied for Signatures 2/13 and identified a total of 110 cancers that met the criteria of falling above 1.5-fold the interquartile distance from the 75^{th} percentile point in each cancer type. Because there were cancer types which showed no outliers of Signature 1A/1B (ALL), only one outlier (CESC, HNSC) or no Signatures 1A/1B (KIRP, LUSC, MM and THCA), it was not possible to perform the analyses in individual cancer types. However, aggregating all the cancers together, we found that there was no enrichment of the deletion allele amongst outliers for Signatures 1A/1B ($p=0.767$).

Although comparisons have not been done against other signatures, Signatures 1A/1B and Signatures 2/13 are the most ubiquitous signatures in human cancers, transcending many

cancer types, making them excellent subjects in comparison to other less prevalent and therefore less comparable mutational signatures.

3.8 Variation in population frequencies of the *APOBEC3A/3B* CNV

The race reported for the breast cancer patients were sought in order to explore whether the association was possibly being driven solely by population stratification of the *APOBEC3A/3B* germline deletion allele given the variation in allele frequencies across the globe[5]. Removing non-Caucasian patients from the analysis would reduce the power of the analysis. Hence this was not a favoured approach. Due to limitations of reporting and of ethical restrictions on genome-wide genotyping of the germline, we compared the different reported races to see if elevated rates of Signatures 2/13 were restricted to particular racial groups. We found that this was not the case and that hypermutators characterized by an excess of Signatures 2/13 were present in all racial groups (Supplementary Table 3B). Despite the reported variation in allele frequencies across the world, carriers of at least one copy of the deletion allele are enriched amongst the hypermutators regardless of race.

4: Additional characteristics of Signatures 2 and 13 which resemble APOBEC-induced mutations

4.1 Strand-coordinated mutagenesis

It was previously highlighted[1,18] that mutations associated with Signatures 2 and 13 were observed to occur more frequently on the same strand (e.g. C>T..C>T..C>G or G>T..G>A..G>A) than would be expected by chance, an observation referred to as "strand-coordinated mutagenesis".

Neighboring mutations could arise on either of two strands of a double-helix (Supplementary Figure 4A) particularly if they had arisen as independent events during different cycles of cell division. If more mutations are observed to occur on the same strand than expected by chance (Supplementary Figure 4B), this would imply one of two scenarios: Either those neighboring mutations arose over different rounds of cell division with preferential targeting of one strand over another or they arose during a single round of cell division and potentially occurred in the same instance.

APOBECs have been implicated in Signatures 2/13 on the basis of the predilection of cytosine mutations at a TpC context. APOBECs also require single-stranded DNA for

deamination. If stretches of single-stranded DNA become available to APOBECs for deamination during the development of a cancer, then this could manifest as strand-coordinated mutagenesis. Furthermore, these mutations may frequently be closer together than would be expected. Documenting strand-coordinated mutagenesis in Signatures 2/13 would lend support to the speculation that APOBECs are involved in the generating the Signatures 2/13.

We therefore sought to first, formally document that neighboring mutations are occurring on the same strand more often than expected in some whole-genome sequenced cancers. Second, show that this pattern of strand-coordinate mutagenesis is a feature of Signatures 2/13 in particular. Third, demonstrate additional features, which could support the latter model of having arisen in a single moment of hypermutability by APOBECs during a single cell cycle event.

4.2 Demonstrating genome-wide strand coordination

It has been shown previously[3] that series of closely-spaced mutations, known as *kataegis* preferentially occur on the same strand. Kataegis shares similarity of sequence-specificity to Signatures 2/13 but shows additional features of dense localisation and marked co-occurrence with somatic rearrangements. Recent experimental evidence suggests that clusters of mutations are found surrounding induced double-strand breaks[25], in other words, that DNA double-strand breaks instigate the occurrence of kataegis. This would suggest that kataegis may arise by an alternative mechanism to the global genome-wide nature of mutagenesis of Signatures 2/13. In order to remove the effects of kataegis, mutations within regions of *kataegis* were identified as previously described[18] and were removed prior to analysis for strand-coordination in order to reduce bias.

In order to demonstrate genome-wide strand coordination, analysis was carried out on all whole-genome sequence data for which BAM files were available (Supplementary Table 4A). The principle of this analysis was described in Online Methods section 4.

Because same-strand mutations were ascertained in an unbiased way from any mutation type (not restricted to just cytosine mutations at TpCs), to see whether strand-coordinated mutations were a particular feature of Signatures 2/13, we sought a relationship between the degree of "strand-coordination", given by the OR of strand-coordination, and the fractional burden of Signatures 2/13 in each cancer (Supplementary Figure 4D). Here, we see a direct relationship between the degree of strand coordination and the burden of

Signatures 2/13 (r=0.74, p=1.14e^{-21}) supporting the notion that strand-coordination was an inherent feature of Signatures 2/13.

4.3 Strand-coordinated mutations are closer in proximity to each other

It was observed from interrogation of BAM files, that strand-coordinated mutations tended to be in close proximity to each other. This was not an insignificant observation. If it is true that Signatures 2/13 mutations have arisen through APOBEC enzyme activity, which requires single-stranded DNA to deaminate cytosines, then it may be more likely for same-strand mutations to be closer to each other.

To test whether strand-coordinated mutations were indeed closer to one another than expected, all successive mutation pairs were ordered by intermutation distance. Fisher exact tests were then performed using a sliding window of bins containing 100 mutation pairs, starting with the 100 mutation pairs with the shortest intermutation distance and extending out to an intermutation distance of 10 KB (Supplementary Figure 4D).

The trumpet plots in Supplementary Figure 4E show the *expected* proportion of same-strand pairs of mutations (dark blue line) and different-strand pairs of mutations (pale blue line) corrected for the overall number of mutations for each cancer as well as correcting for the overall pattern of mutations (the distribution of C>A, C>G, C>T, T>A, T>C and T>G) for individual cancers. The *observed* proportion of same-strand pairs (black dots if significantly different from expected, dark blue dots otherwise) and different-strand pairs of mutations (red dots if significantly different from expected, light blue dots otherwise) are also plotted. Each bin of observed mutations will therefore contain two dots reporting the proportion of same and different strand pairs of mutations (if added together would amount to 1). LOESS fitted lines are also presented, showing the smoothed proportion of 'same' (black line) and 'diff' (red line) mutation pairs.

For each window, whether the pairs of mutations show a significant enrichment for 'same' pairs is assessed as follows. The probability that a pair of successive mutations would affect the same base given randomly positioned mutations is given by $p_A^2 + p_C^2 + p_G^2 + p_T^2$, where p_X is the fraction of mutations that occur at nucleotide *X*. The odds ratio, *OR*, is therefore calculated as

$$OR = \mathrm{n}_{\mathrm{diff}} \times (p_A^2 + p_C^2 + p_G^2 + p_T^2) \,/\, n_{same} \times (1 - p_A^2 + p_C^2 + p_G^2 + p_T^2)$$

where n_{diff} is the number of pairs of successive mutations that occur at different nucleotides and n_{same} is the number of pairs of successive mutations that occur at the same nucleotide (see Supplementary Figure 4A). An enrichment of mutations within a processive series is indicated by a greater than expected value of n_{same}. The probability of observing such a value by chance is the p-value obtained from a Fisher's exact test of the contingency table for expected vs observed n_{same} and n_{diff}. Windows assigned a p-value less than or equal to 0.05 are considered to have a significant enrichment of same strand pairs of mutations.

In samples that do not have much strand-coordinated mutagenesis, the observed (red/black) Loess lines approach the expected for all bins of intermutation distances (e.g. PD7321a, PD7216a, Supplementary Figure 4D). In contrast, cancers with a lot of strand-coordinated mutagenesis show not just an excess of same strand mutations, but also that the same-strand mutations are over-represented particularly at shorter intermutation distances (e.g. PD4120a & PD4224a, black Loess line for observed same strand mutations deviates away from expected at short distances between successive mutations, Supplementary Figure 4E. Intermediary samples: PD6042a and PD4958a). This suggests that strand-coordinated mutations are frequently closer in proximity to each other.

4.4 Strand-coordinated mutations are in *cis*

Although we have demonstrated strand-coordinated mutagenesis in Signatures 2/13 and shown how pairs of strand-coordinated mutations are usually in close proximity to each other, mutations can only truly be strand-coordinated if it is possible to demonstrate that they arise on the same parental haplotype i.e. are in *cis* with each other.

Whole-genome sequenced libraries of the cancers in this analysis had NGS insert sizes of approximately 400-700bp. Each insert will have had 100bp sequenced at either end. These 100bp paired-reads are referred to as read-mates of each other. A single NGS insert (and its pair of read-mates) represents DNA obtained from a single DNA molecule. Mutations present on the same read or on read-mates therefore are in *cis* with respect to each other, come from the same DNA molecule and are present on the same parental allele (Supplementary Figure 4C). We set out to show that informative same-strand mutations (close enough to be on the same insert i.e. within 700bp of each other) were indeed in *cis*.

As with the identification of strand-coordinated mutations, mutations within *kataegis* regions were not included in this analysis. Pairs of successive mutations within 700 bp were identified (Supplementary Table 4A, columns F-M) and BAM files were searched to identify

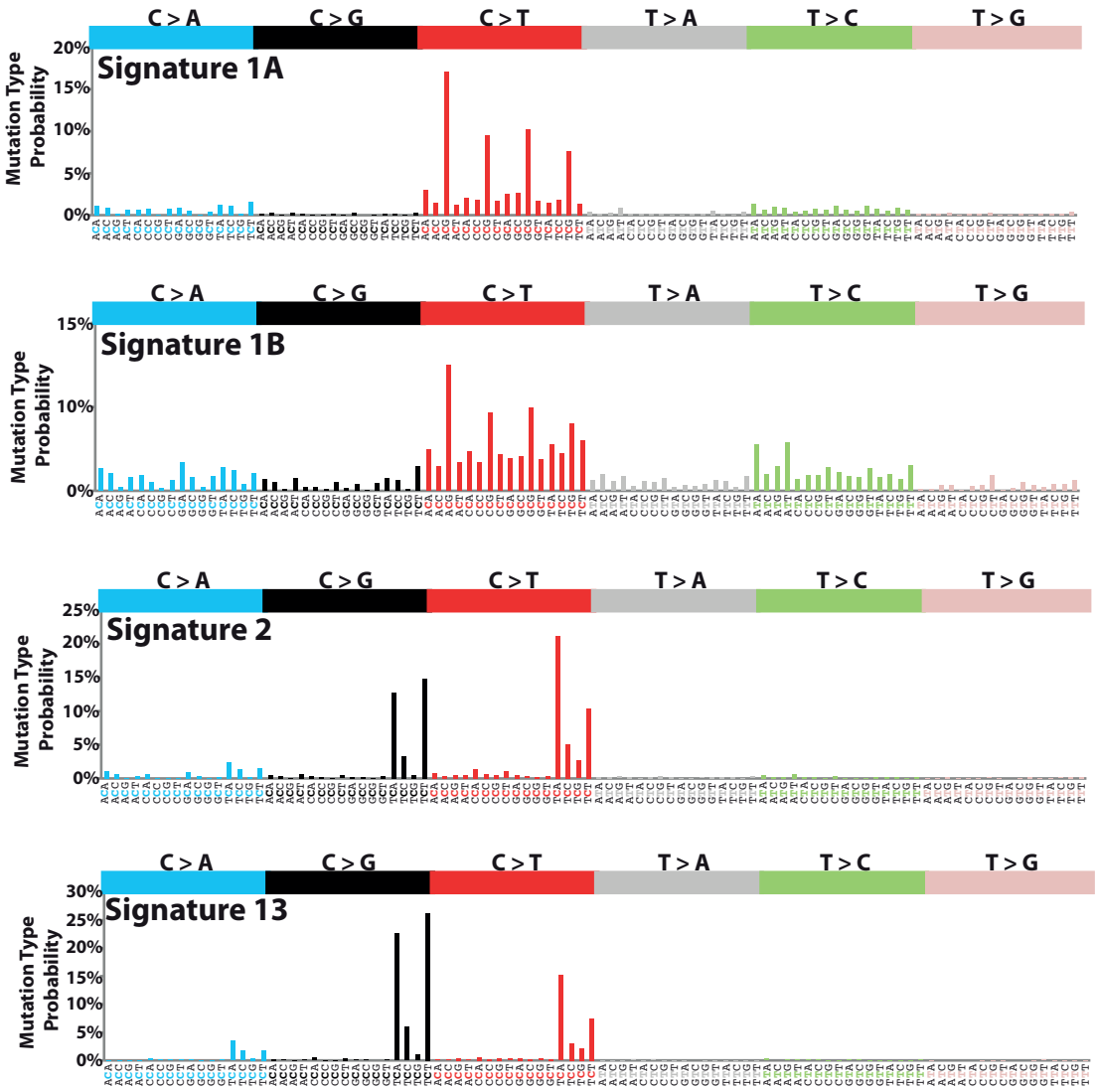
all read pairs that covered the positions of both loci. For each mutation position, a read may be wild type (WT) or mutated (MUT). All reads were classified as WT-WT, WT-MUT, MUT-WT or MUT-MUT. Mutation pairs in *trans* have all reads classified into 3 categories - WT-WT, WT-MUT or MUT-WT – and have no MUT-MUT pairs. Mutation pairs in *cis*, on the other hand, have reads classified as WT-WT and MUT-MUT. For the purposes of this analysis, we concentrated on demonstrating an enrichment of strand-coordinated variants in *cis* than in *trans* (Supplementary Table 4A). We find many such examples particularly in cancers with a lot of strand-coordination. The excess of mutations in *cis* is demonstrated using an OR of mutations in cis, although, in some instances the numbers of mutations were too few to be informative.

5: Relationship between expression of APOBEC family members and rates of mutation of Signatures 2/13

RNA-seq derived expression data was obtained from the https://browser.cghub.ucsc.edu/ for 1691 patients. Expression levels for each APOBEC family member were standardized relative to the levels of *TBP* (TATA-binding protein).

It was previously suggested that APOBEC3B may be the main instigator of APOBEC-related mutagenesis[27-29]. We explored the relationship between the APOBEC3B expression levels and germline deletion allele status in these cancers (Supplementary Figure 5A). Intriguingly, we find a very clear relationship between the copies of deletion allele and the expression levels of APOBEC3B. In particular, homozygous carriers of the deletion allele, which effectively deletes genomic APOBEC3B, show very little (if any) APOBEC3B expression (Supplementary Figure 5A), yet still show apobec-related mutagenesis (arrows Supplementary Figure 5B). At least in these cancers, factors other than APOBEC3B must be contributing to the increased rate of Signature 2/13 mutagenesis. This was not seen for other APOBEC family members (Supplementary Figure 5A, Supplementary Table 5).

A

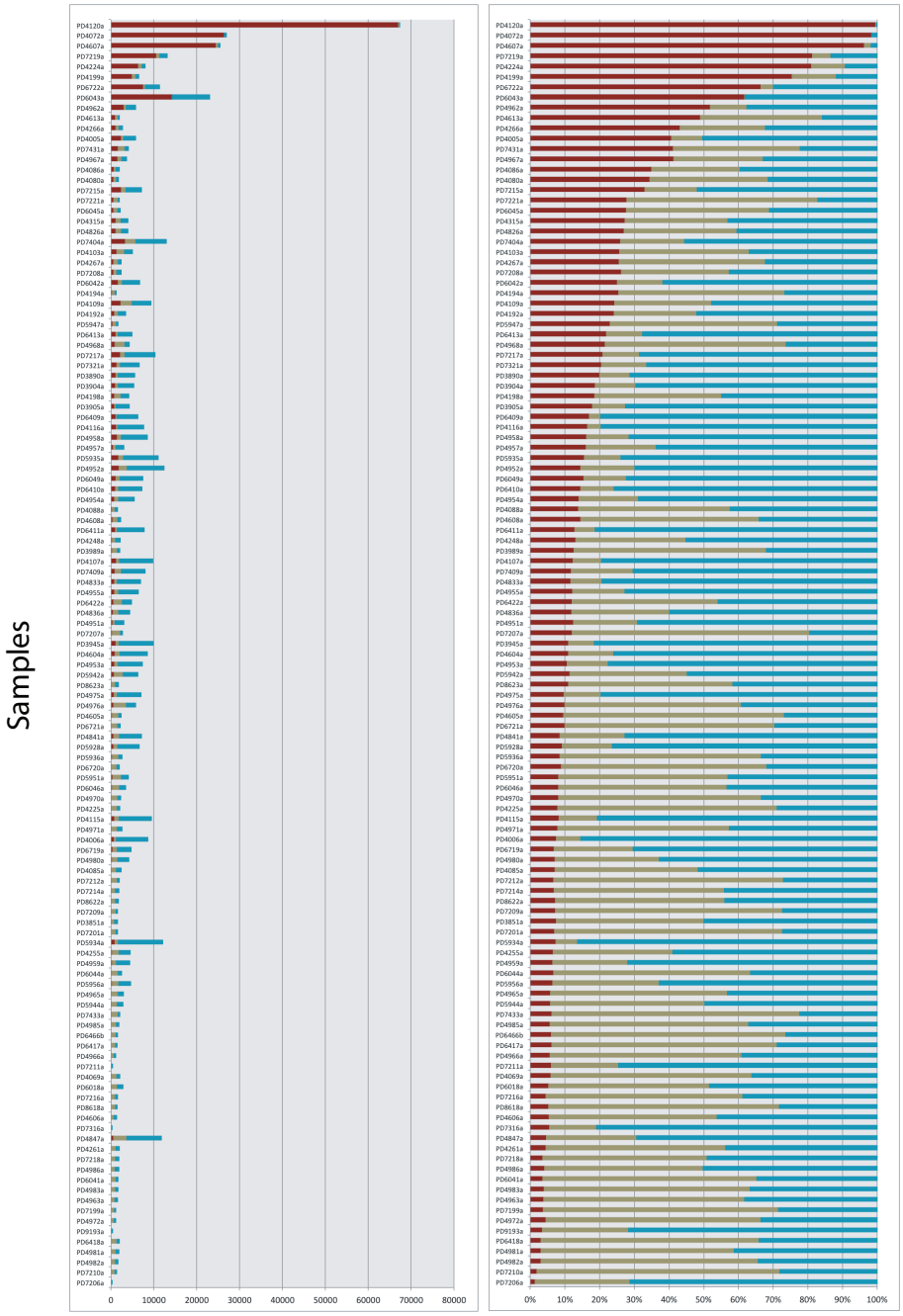


B

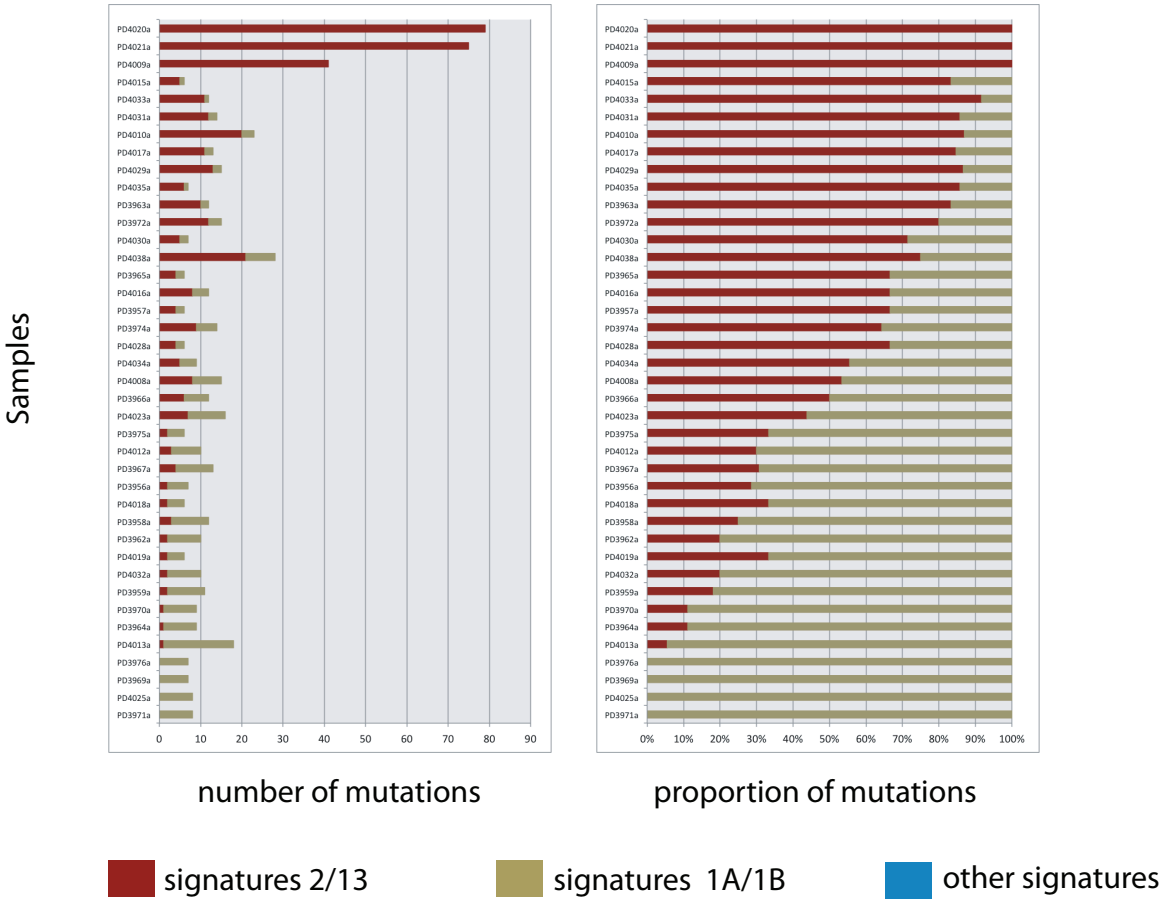
	ALL	AML	Bladder	Breast	Cervix	CLL	Colorectum	Esophagus	Glioblastoma	Glioma Low Grade	Head and Neck	Kidney Chromophobe	Kidney Clear Cell	Kidney Papillary	Liver	Lung Adeno	Lung Small Cell	Lung Squamous	Lymphoma B-cell	Medulloblastoma	Melanoma	Myeloma	Neuroblastoma	Ovary	Pancreas	Pilocytic astrocytoma	Prostate	Stomach	Thyroid	Uterus
Signature 1A		■	■						■	■		■								■							■	■		
Siganture 1B	■	■	■	■	■	■	■	■			■		■		■	■			■		■		■	■	■	■				■
Signature 2	■		■	■	■	■		■			■			■		■		■	■			■			■			■	■	■
Signature 13			■	■																										

c

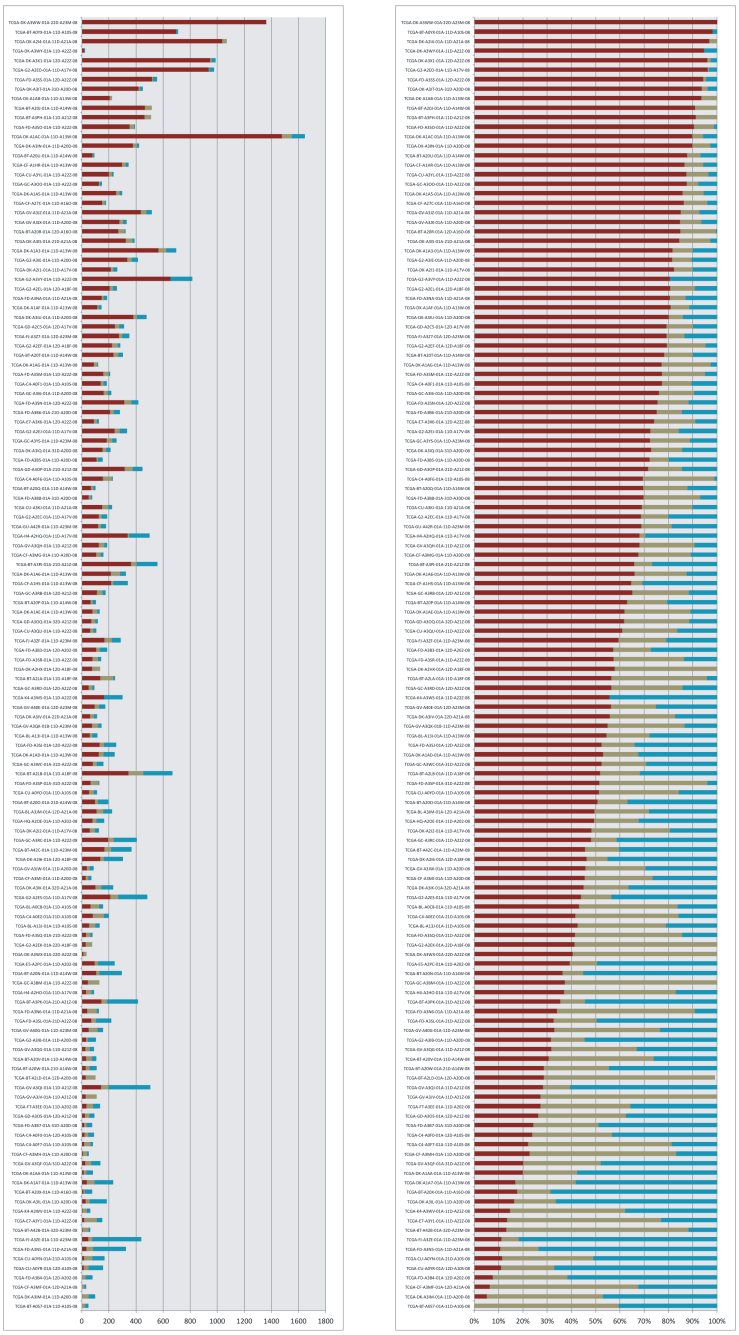
Whole-genome sequenced breast cancer (BRCA)



D Exome sequenced acute lymphoblastic leukaemia (ALL)



E Exome sequenced bladder cancer (BLCA)



number of mutations

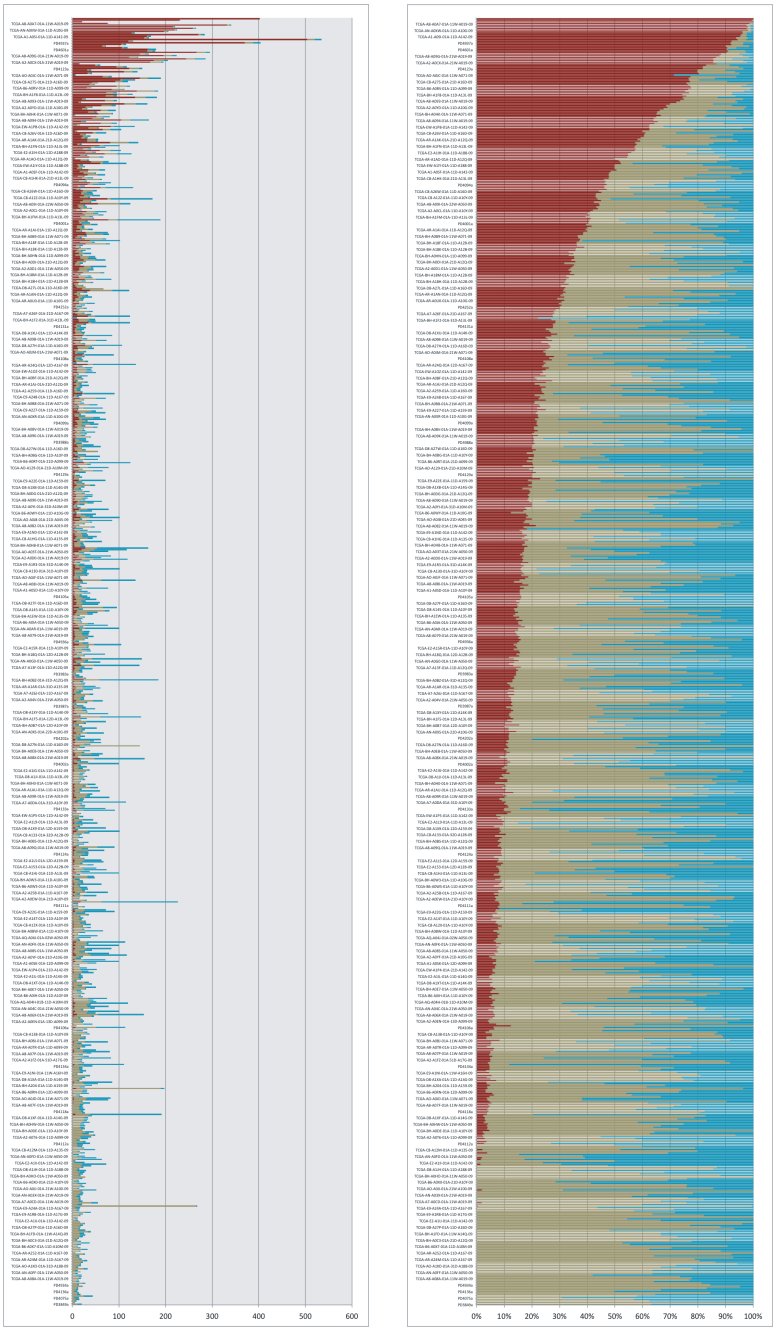
proportion of mutations

signatures 2/13

signatures 1A/1B

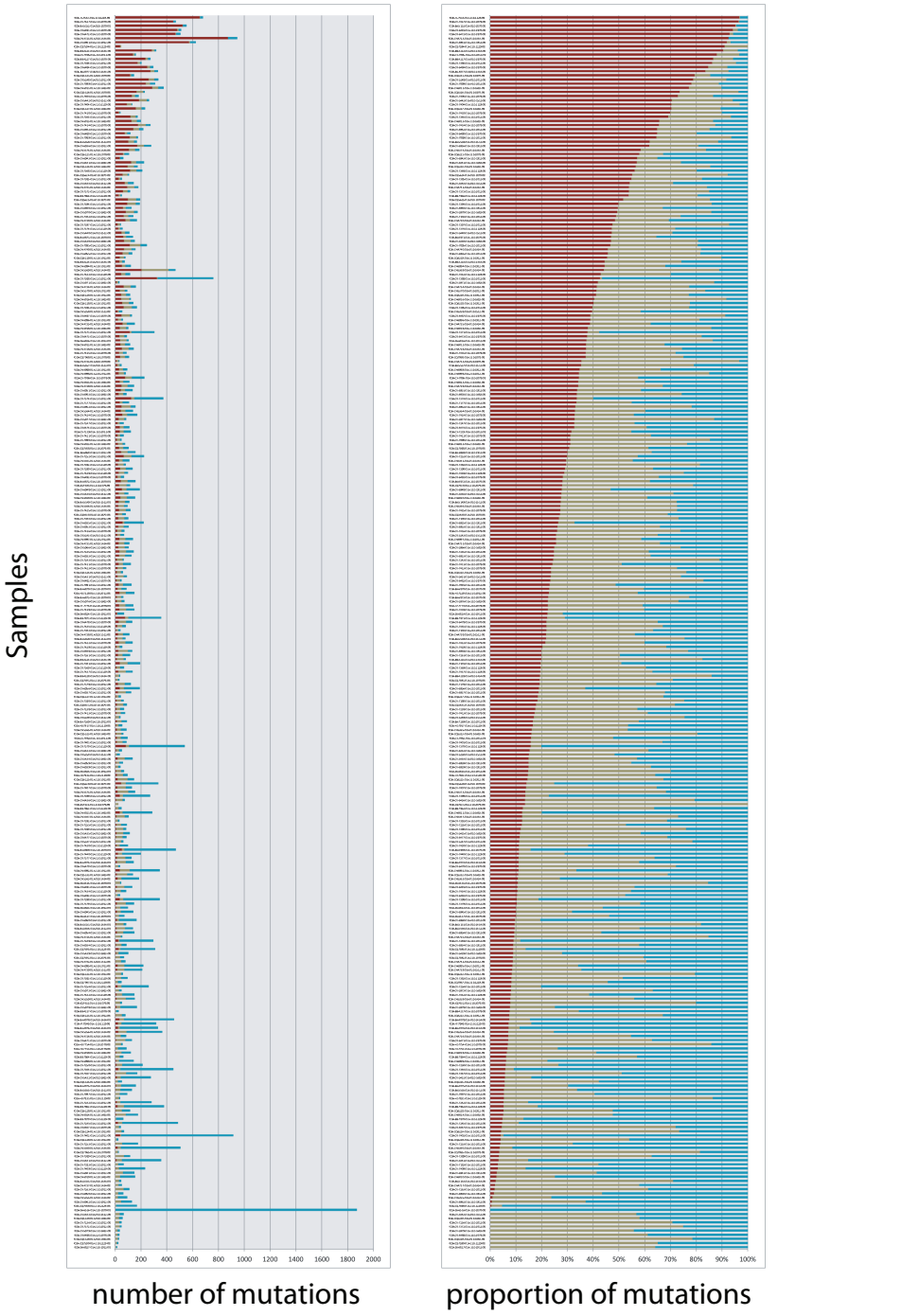
other signatures

F Exome sequenced breast cancer (BRCA)

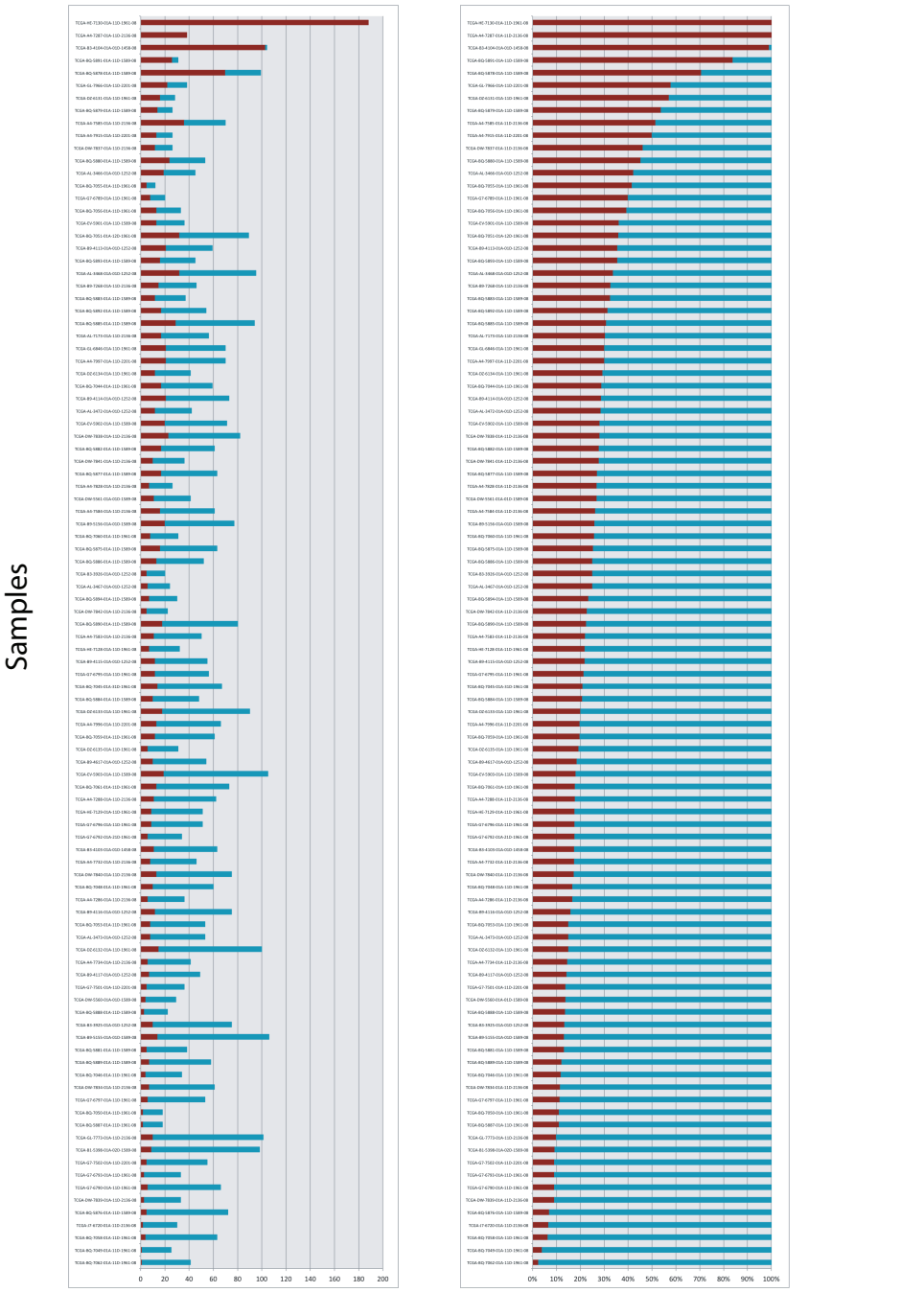


G Exome sequenced cervical cancer (CESC)

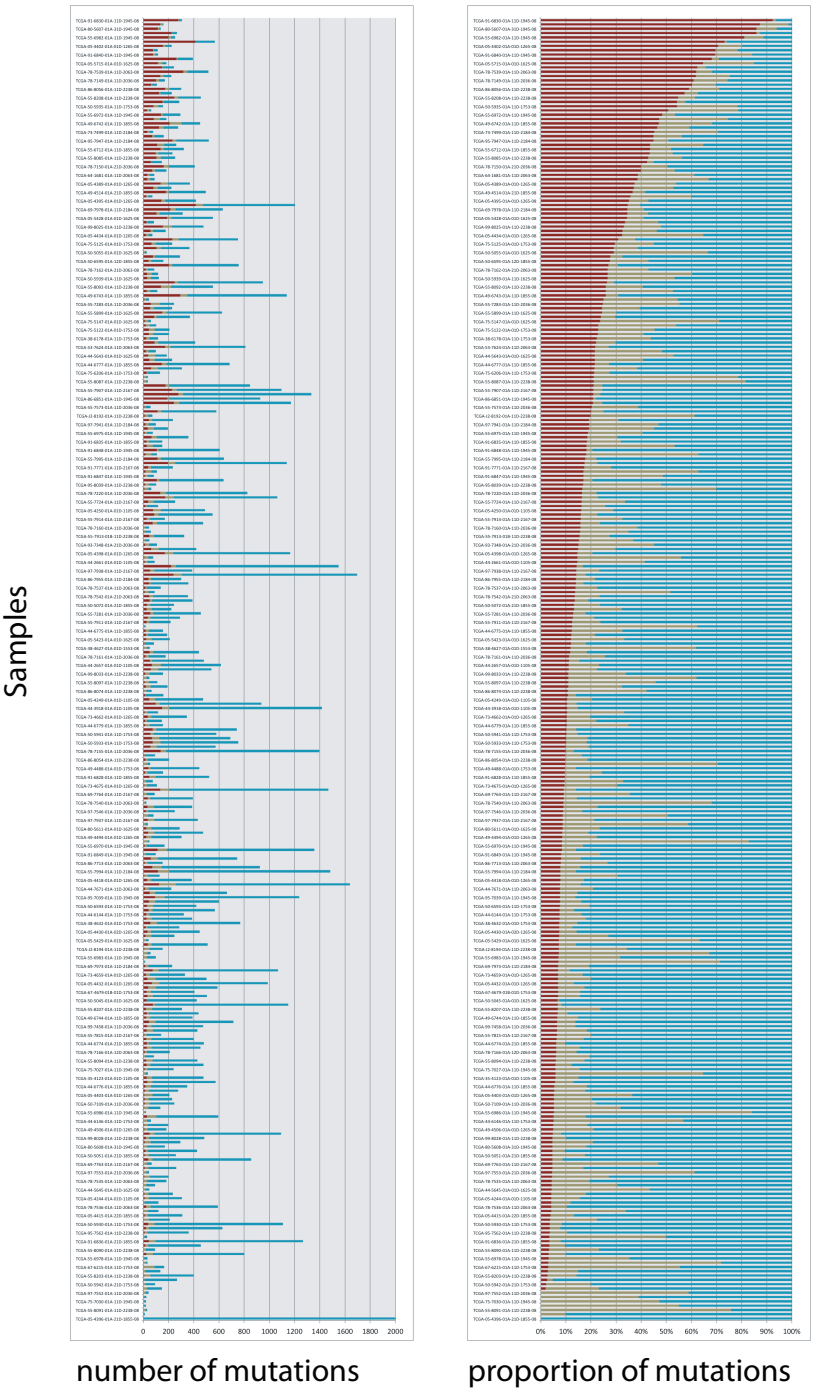
H Exome sequenced head and neck cancers (HNSC)



I Exome sequenced renal papillary cancers (KIRP)



J Exome sequenced lung adenocarcinoma (LUAD)

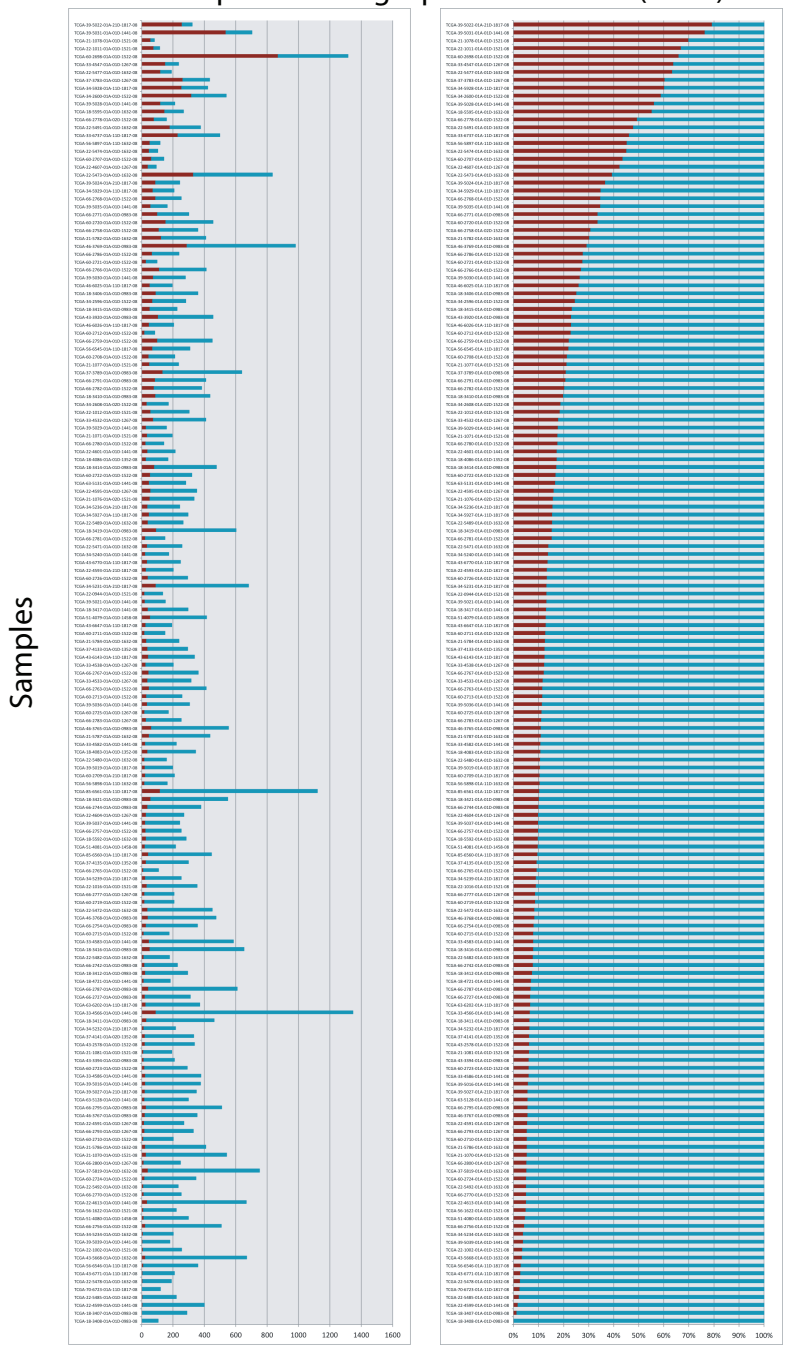


number of mutations

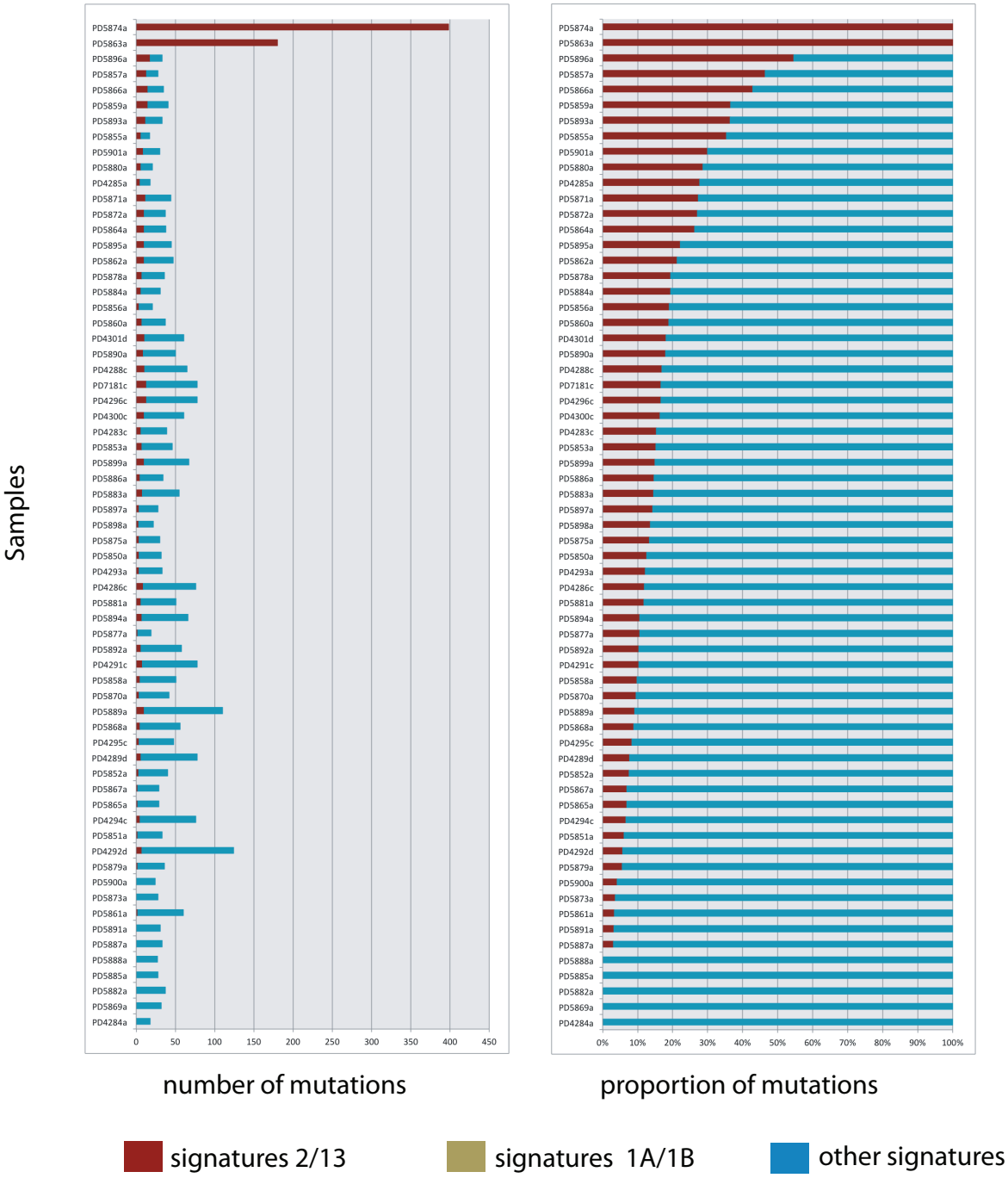
proportion of mutations

signatures 2/13 signatures 1A/1B other signatures

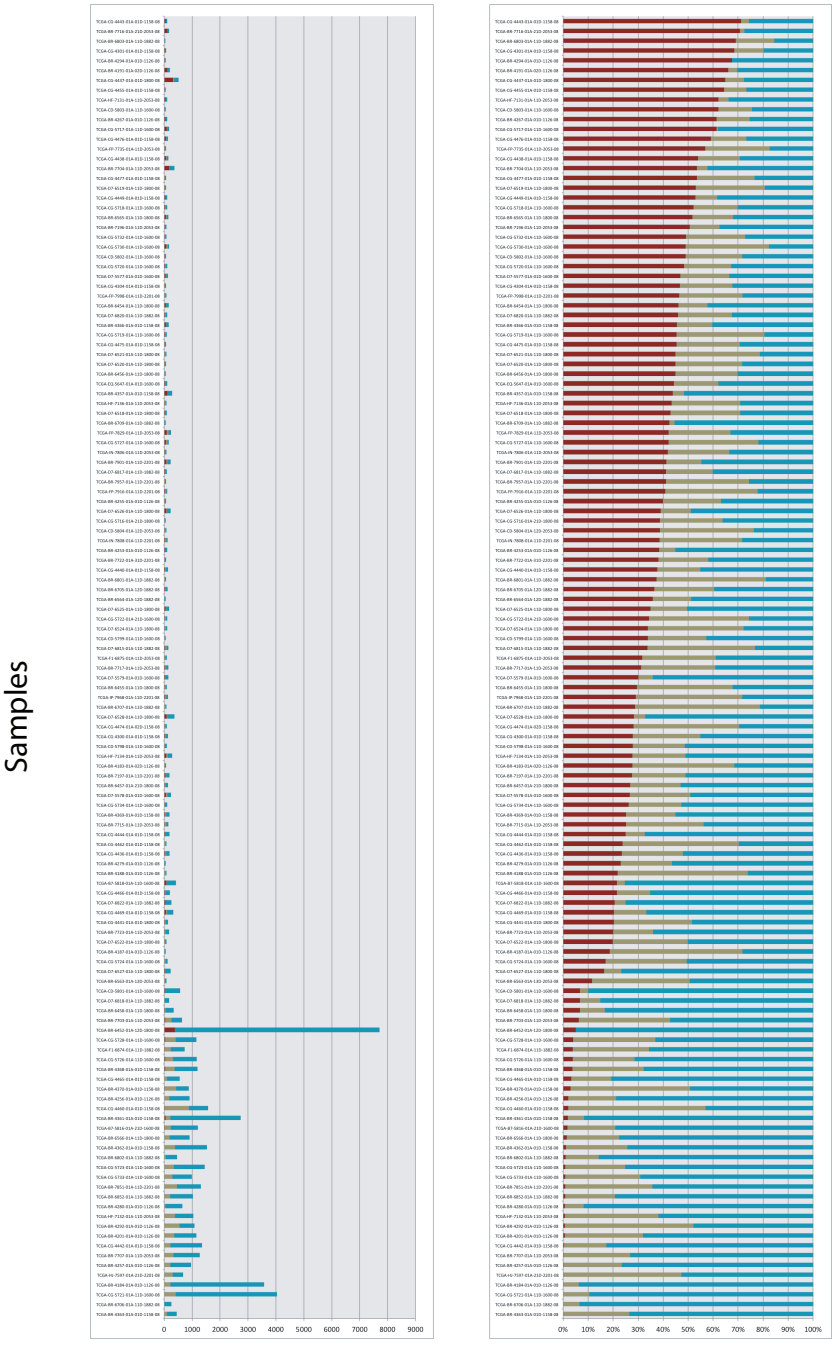
K Exome sequenced lung squamous cancer (LUSC)



L Exome sequenced multiple myoloma (MM)



M Exome sequenced stomach adenocarcinoma (STAD)



number of mutations

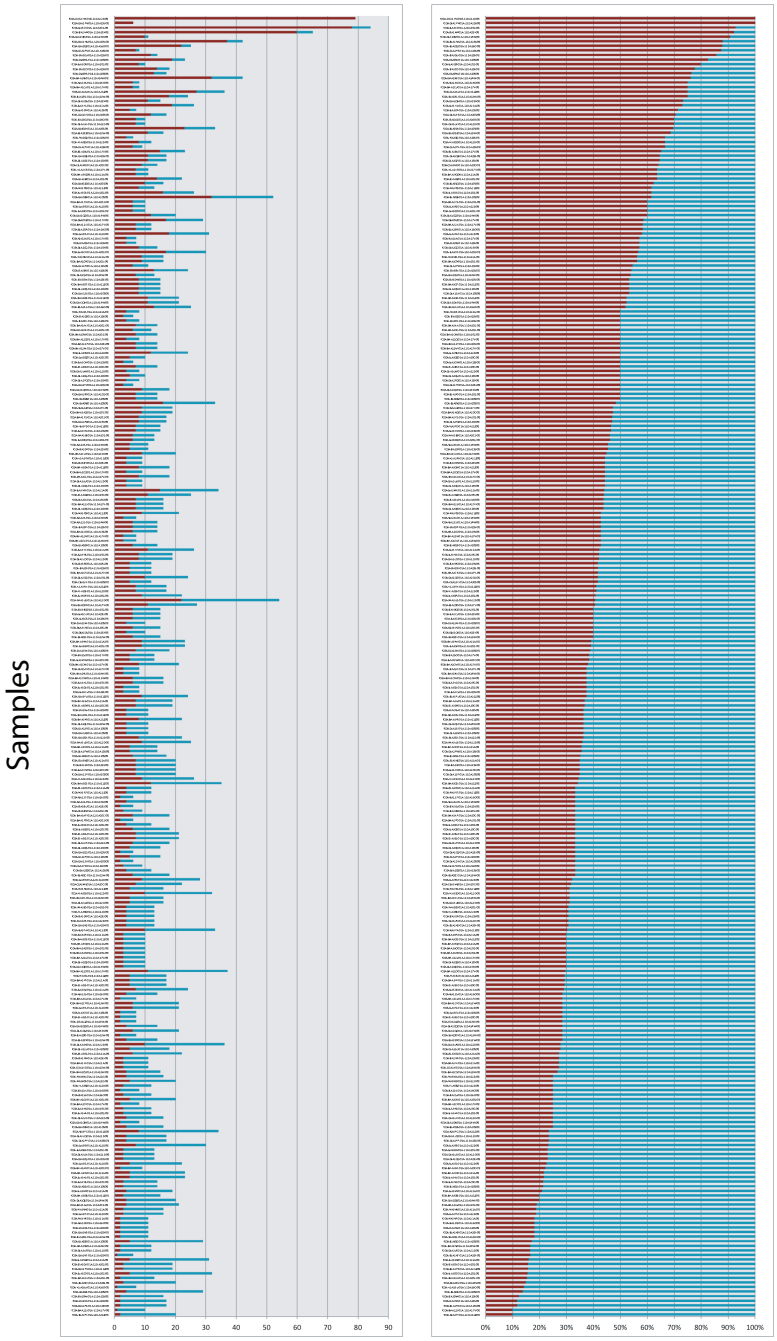
proportion of mutations

signatures 2/13

signatures 1A/1B

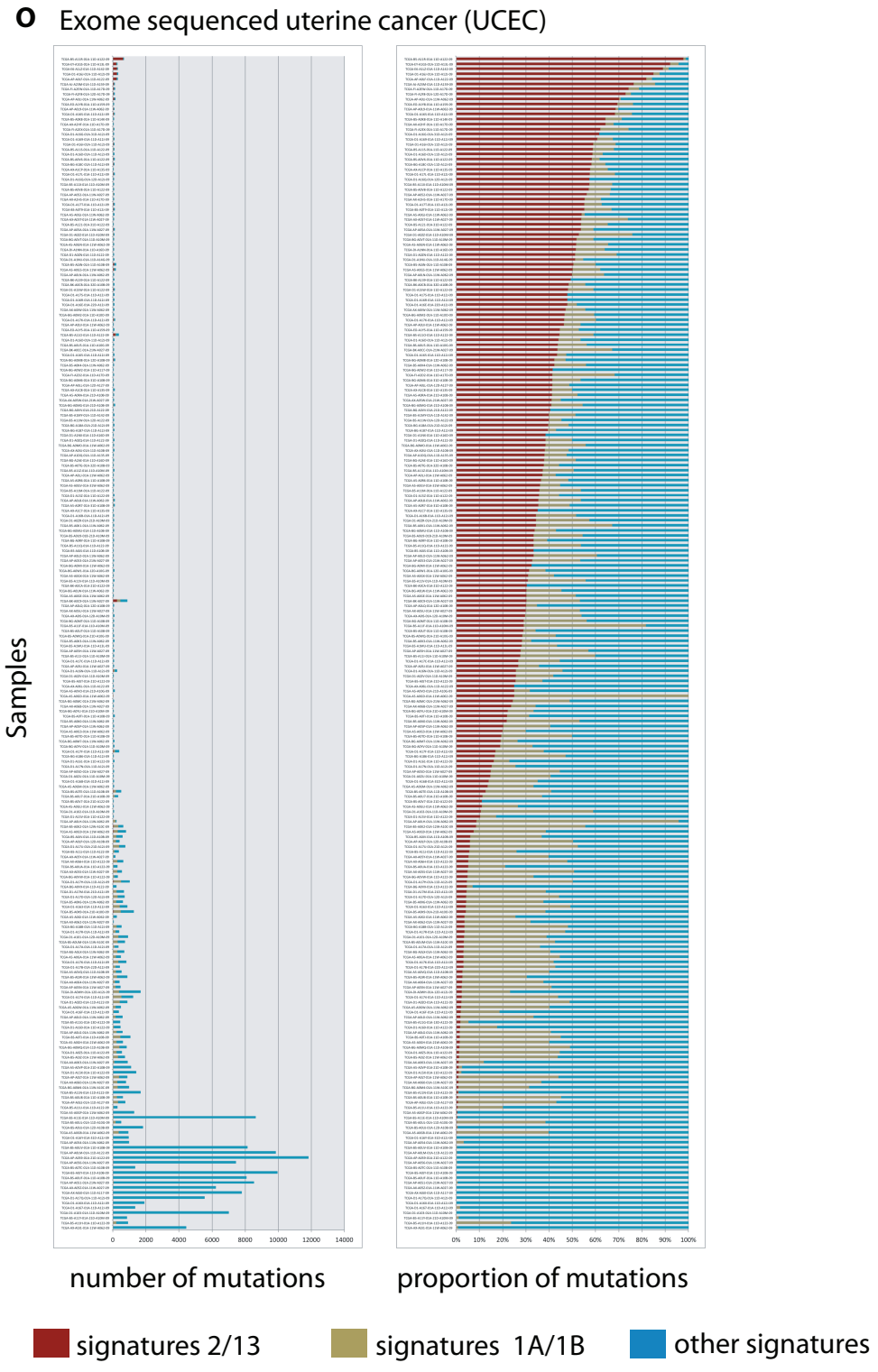
other signatures

N Exome sequenced thyroid adenocarcinoma (THCA)



number of mutations proportion of mutations

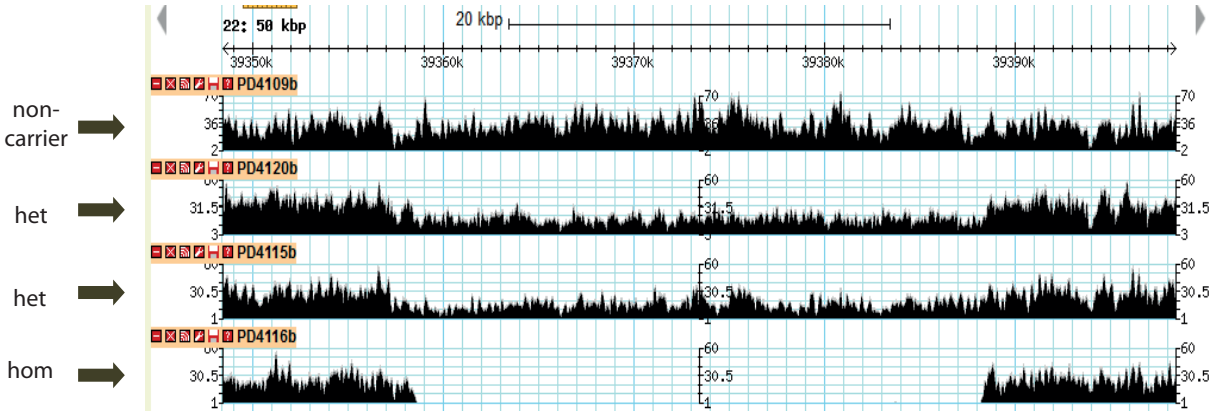
signatures 2/13 signatures 1A/1B other signatures



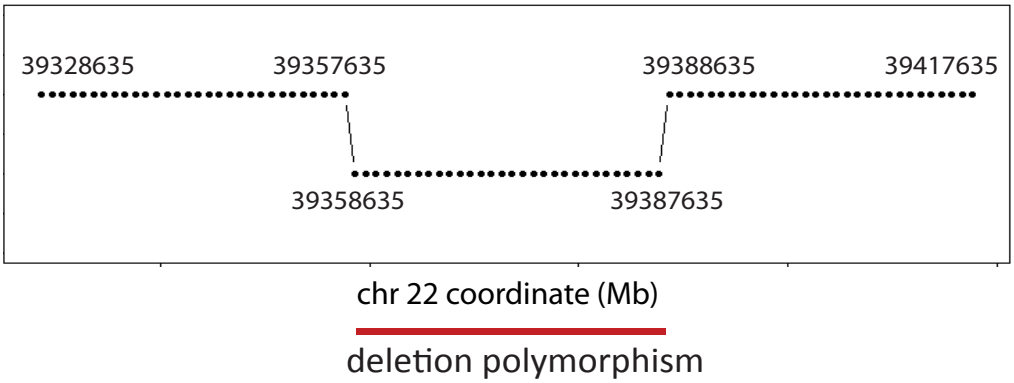
Supplementary Figure 1: Sequence features of Signature 1A/1B and Signature 2/13 as extracted by NNMF. (A) Signatures 1A and 1B are subtly different to each other. Both are dominated by C>T at a XpCpG context but the peaks are more dominant in Signature 1A.There is additionally a T>C component to Signature 1B. Signature 2 is dominated by C>T transitions whilst signature 13 is dominated by C>G transversions at the same TpCpX context. (B) Presence of these signatures in different cancer types. These signatures were almost ubiquitously seen across all cancer types. Green square denotes the cancer-type in which each signature is seen. This data has been taken from Alexandrov et al 2013. (C) Graphical display showing absolute numbers of Signatures 2/13 (dark red), Signatures 1A/1B (khaki) and other signatures (blue), as well as proportional contributions of these signatures in (C) whole-genome sequenced breast cancers and exome -sequenced (D) ALL, (E) BLCA, (F) BRCA, (G) CESC, (H) HNSC, (I) KIRP, (J) LUAD, (K) LUSC, (L) MM, (M) STAD, (N) THCA and (O) UCEC (see Supplementary Table 3A for description of abbreviations.

A

Visualisation of BAM files

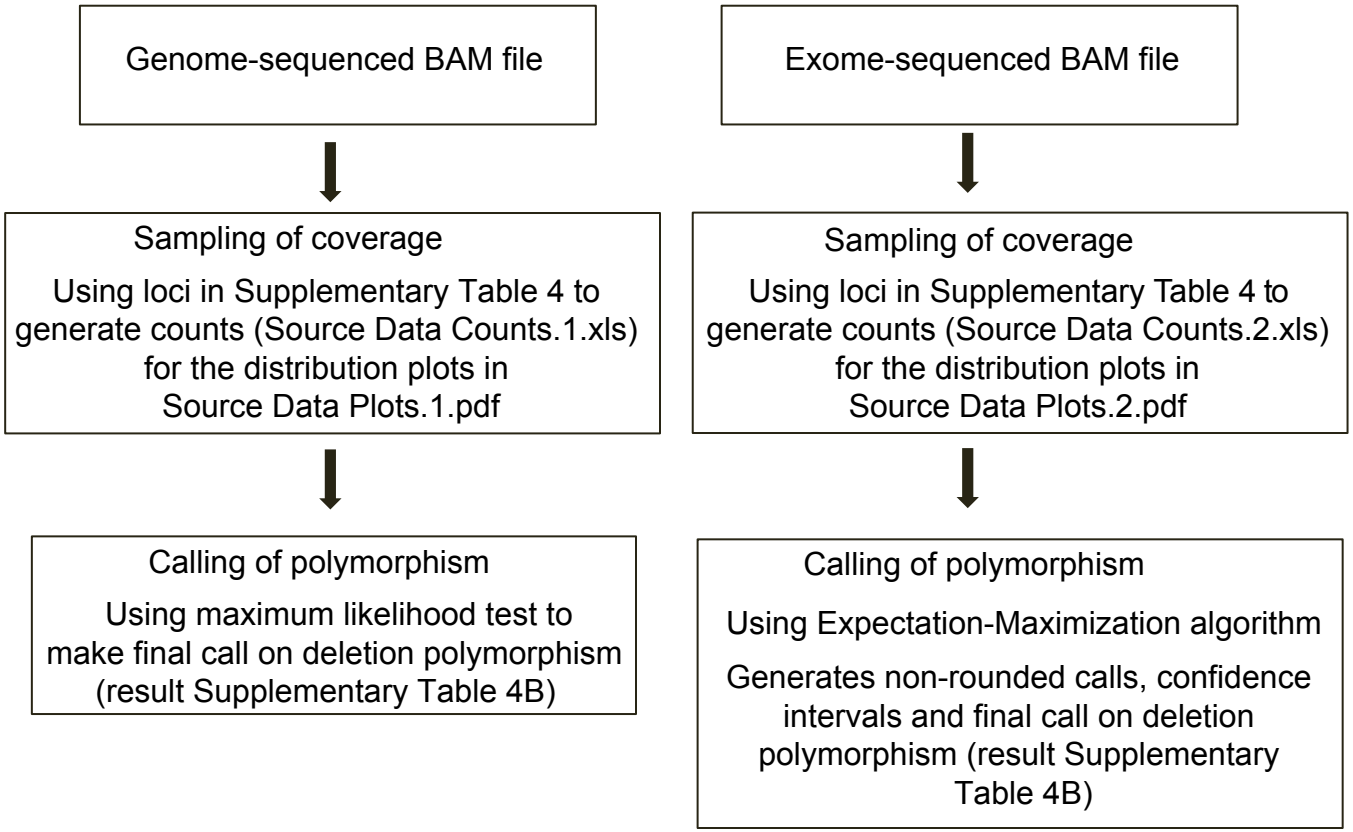


Sampling of BAM files

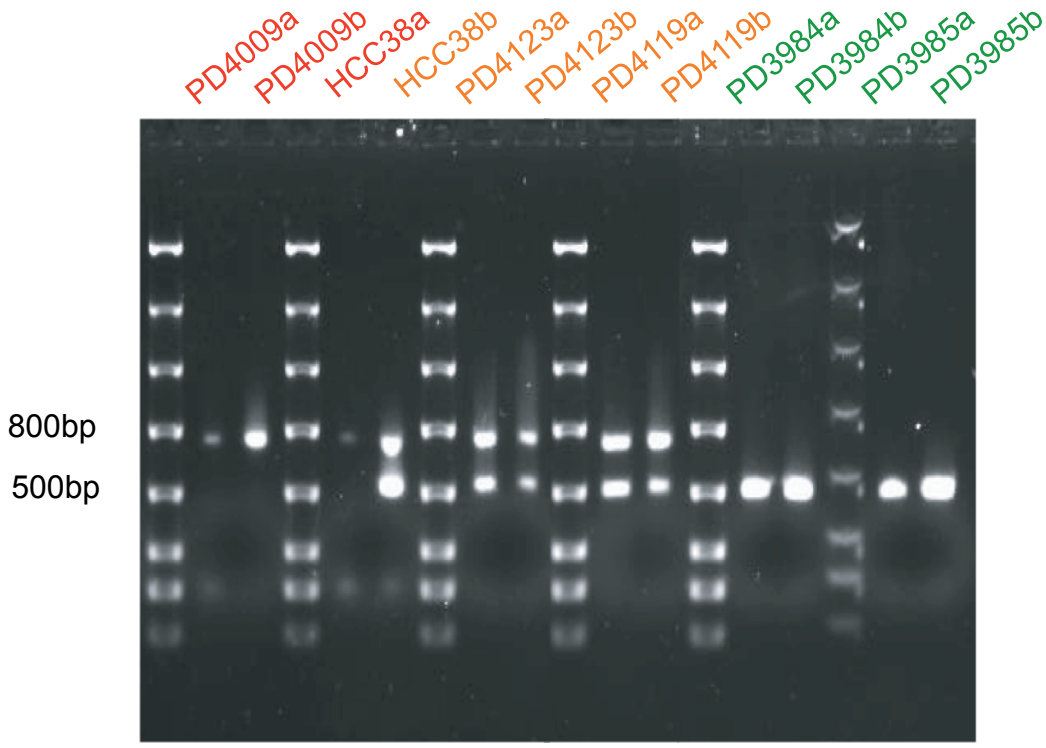


Supplementary Figure 2: (A) Germline copy number polymorphism detection. Image of next-generation sequencing BAM files. 50kb region including the APOBEC3A/3B germline deletion allele locus in 4 different individuals is presented. Non-carrier = patient not carrying the deletion allele, het = heterozygous carrier, hom = homozygous for the APOBEC3A/3B deletion polymorphism. Principle of sampling BAM files for germline APOBEC3A/3B deletion polymorphism also depicted.

B

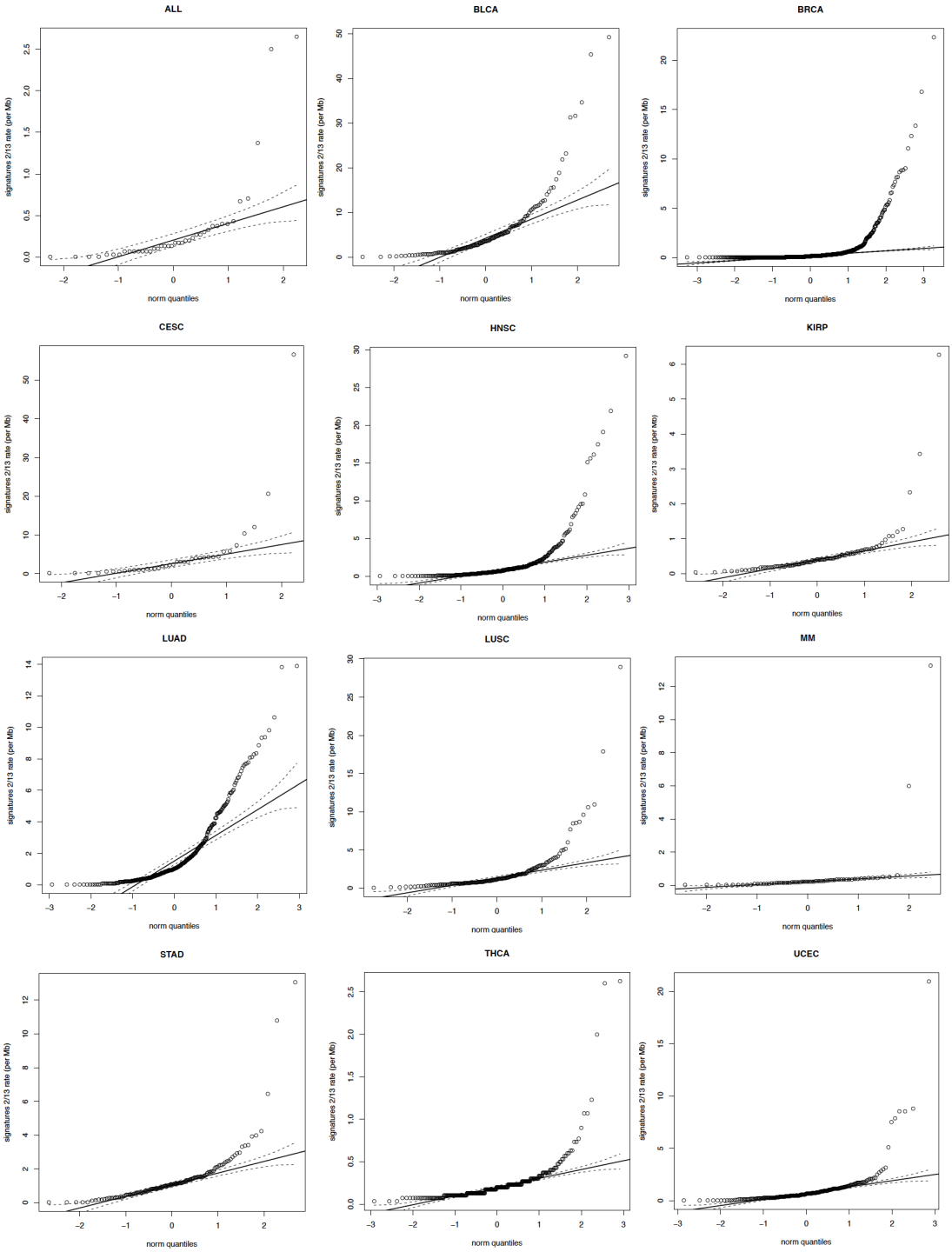


Supplementary Figure 2: (B) Workflow of detection of germline *APOBEC3A* and *APOBEC3B* deletion polymorphism (Supplementary Dataset 1 and Supplementary Dataset 2).

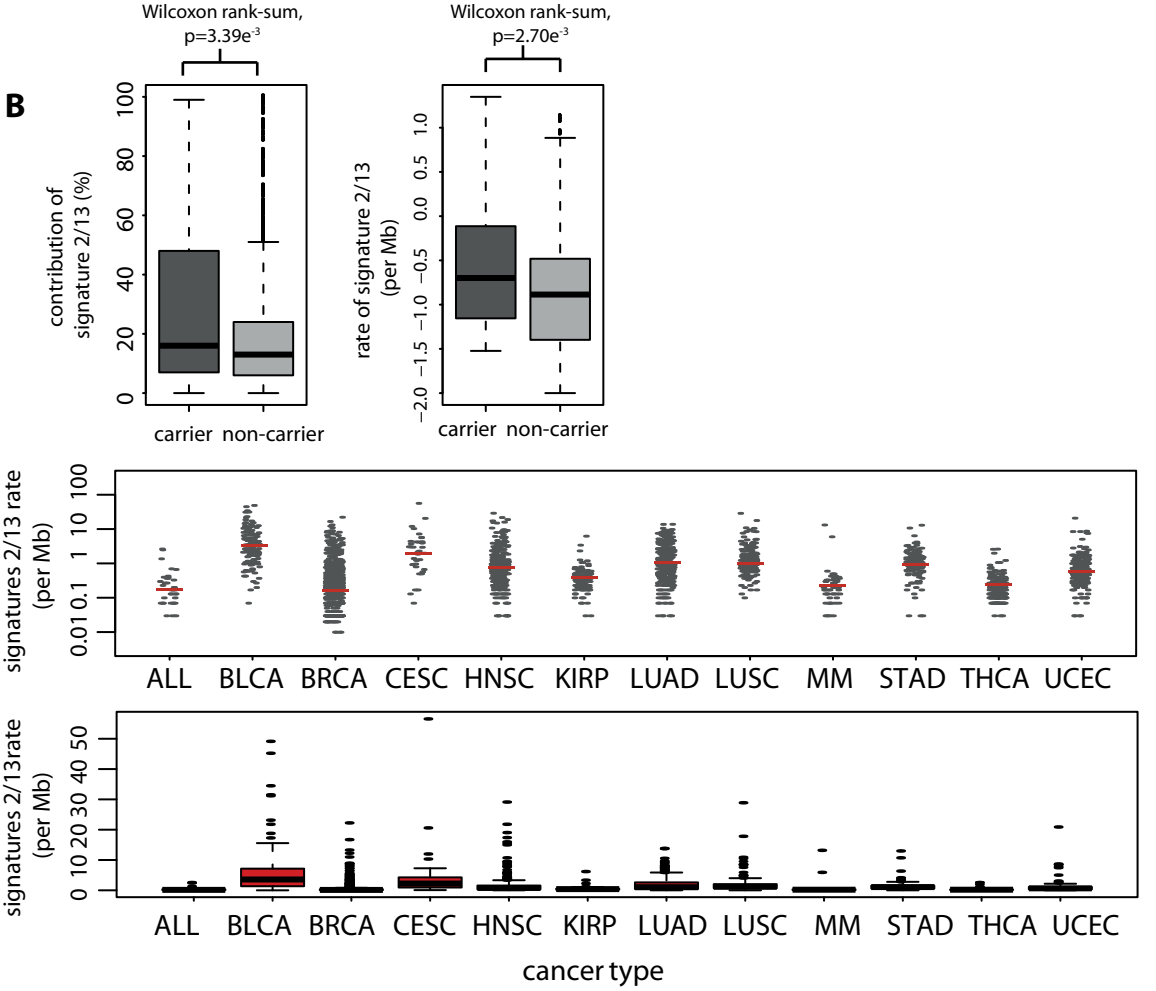


Primer name	5' sequence 3'
APOBEC3A/B_Deletion_F	TAGGTGCCACCCCGAT
APOBEC3A/B_ Deletion_R	TTGAGCATAATCTTACTCTTGTAC
APOBEC3A/B_ Wildtype_F	TTGGTGCTGCCCCCTC
APOBEC3A/B_ Wildtype_R	TAGAGACTGAGGCCCAT

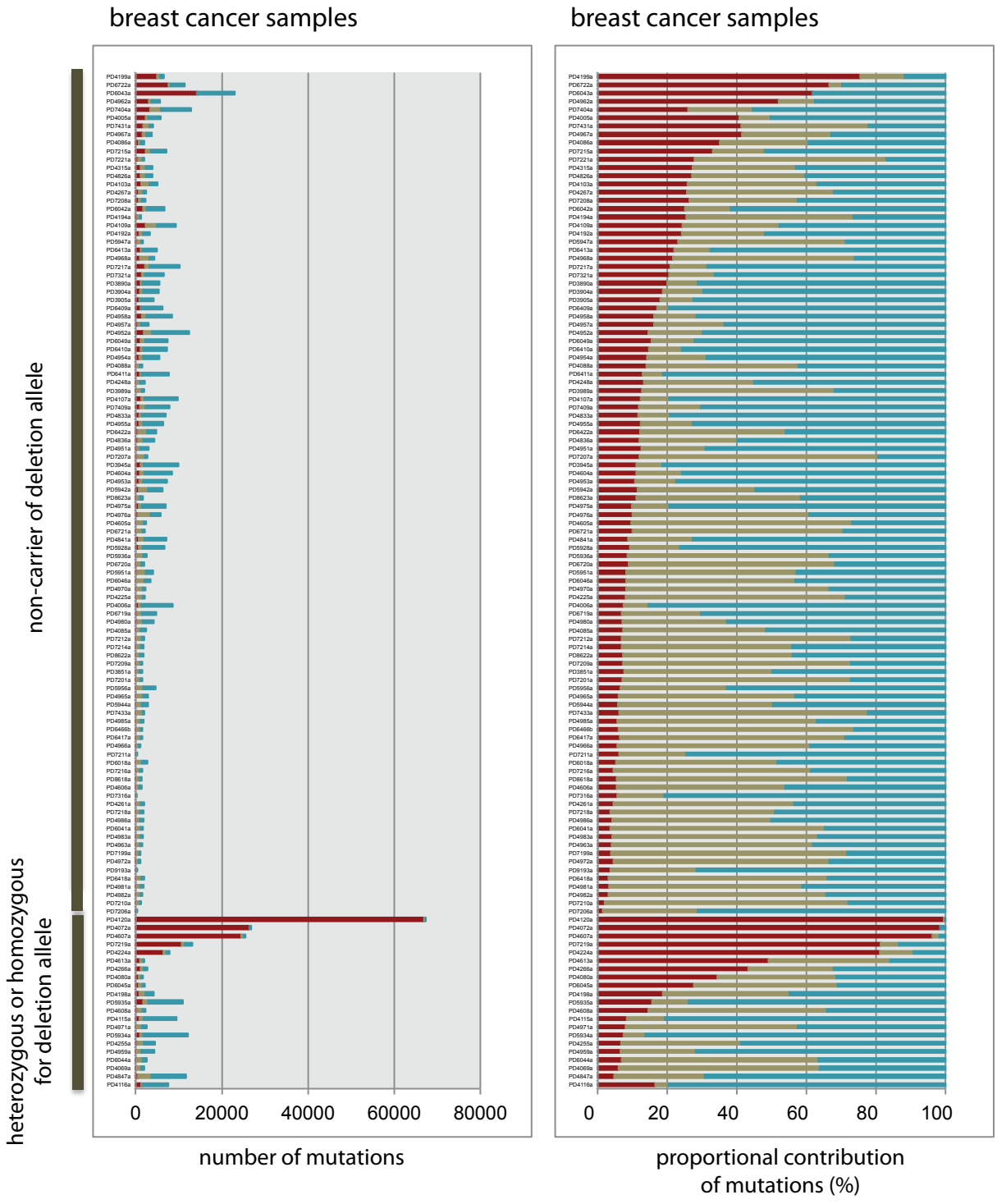
Supplementary Figure 2: (C) Genomic PCR demonstrating concordance with informatic method of deriving polymorphism status in a subset of samples, as well as true discordance between tumor and normal samples of HCC38a/b (PDXXXXa = tumour, PDXXXXb = normal). Red = homozygous for deletion allele, orange = heterozygous for deletion allele, green = non-carrier. Expected size of PCR products: wild-type 490bp, deletion allele 700bp.



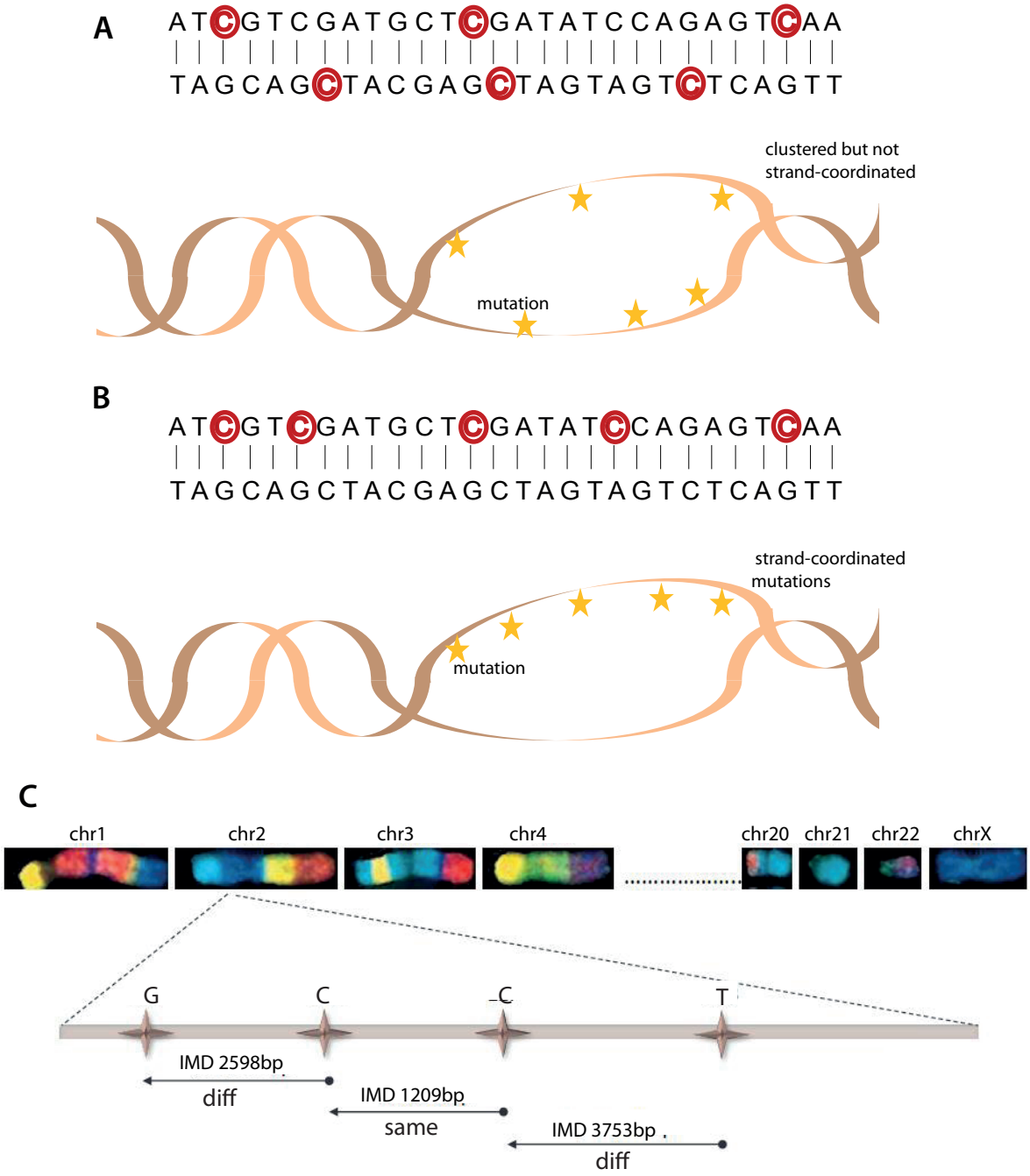
Supplementary Figure 3: (A) QQ plots of all the cancer types used in this analysis.



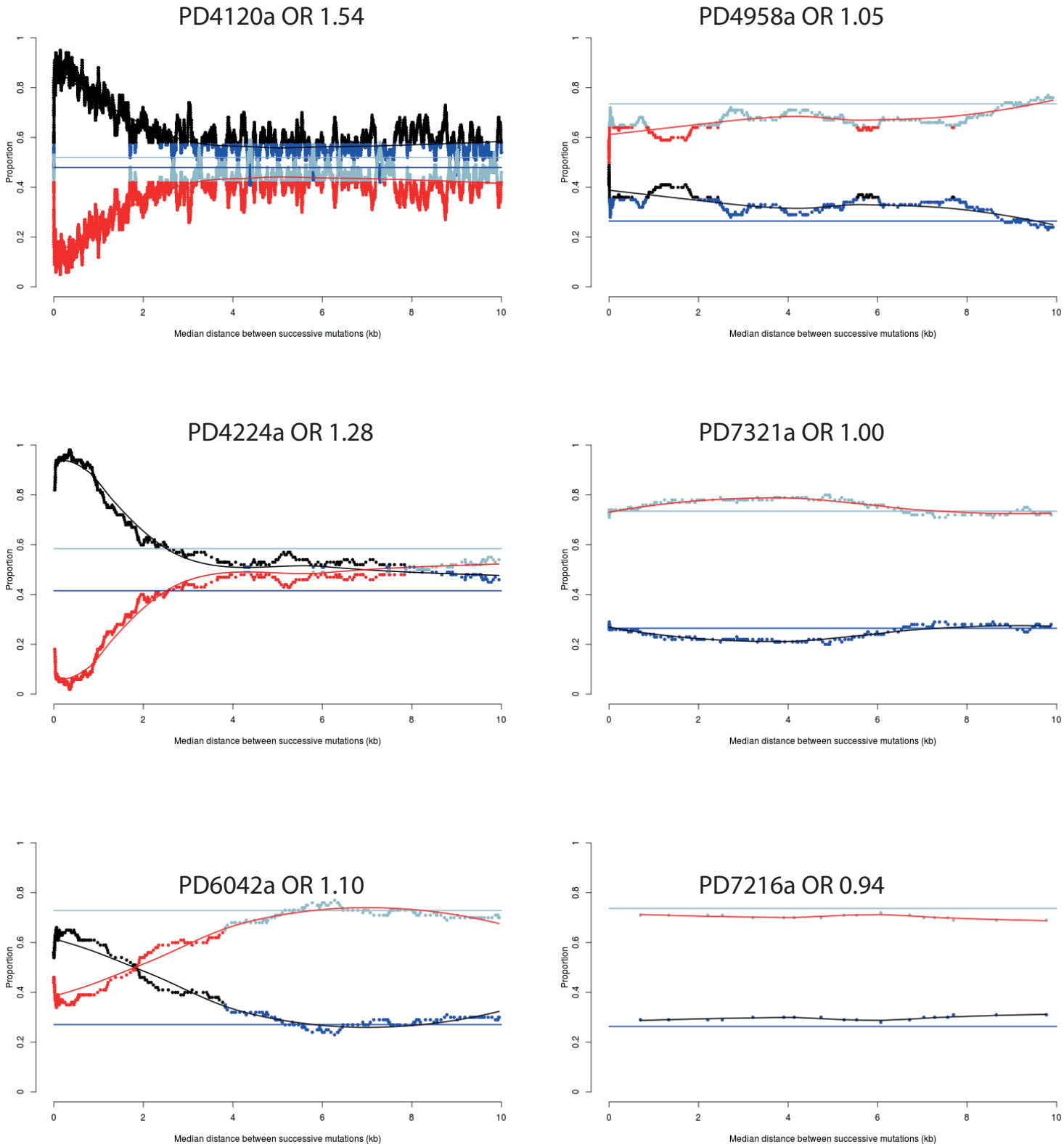
Supplementary Figure 3: (B) Comparison of rates of mutation and proportional contribution of Signatures 2/13 between carriers of at least one copy of the deletion allele and non-carriers in breast cancers (C) The rate of mutagenesis attributed to Signatures 2/13 varied between tumour types with cancers such as bladder, cervical, head and neck, lung cancers and stomach cancers showing a higher mutation rate than the other tumour types. (D) All tumor types demonstrated a tail of outliers which we have subsequently defined as "hypermutators". Standard box-and-whisker plots, coefficient of 1.5.



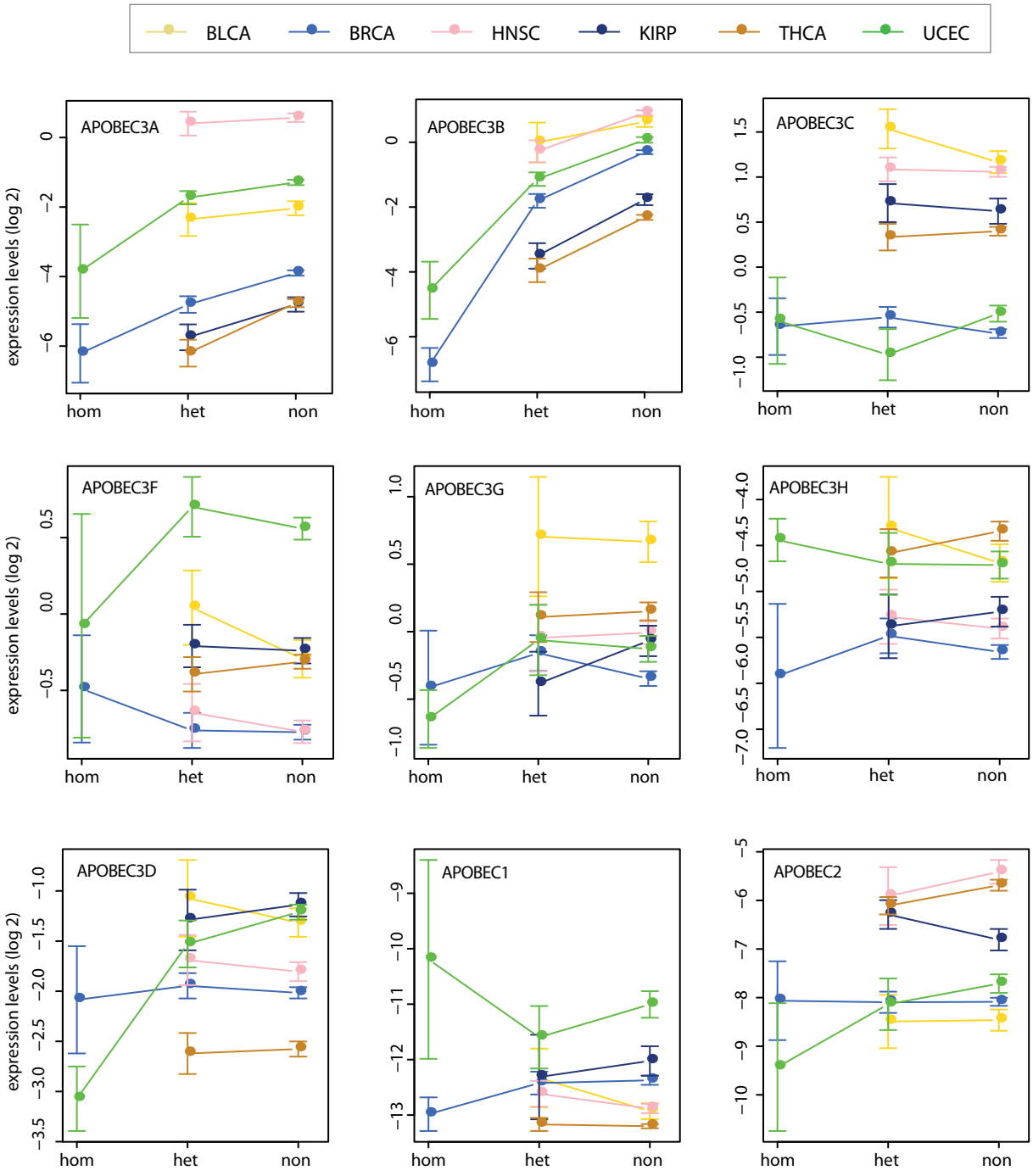
Supplementary Figure 3E: Distribution of number of mutations and contribution of mutations by carrier status of the deletion allele in123 whole-genome sequenced breast cancers. Signatures 2/13 (dark red), Signatures 1A/1B (khaki) and other signatures (blue).



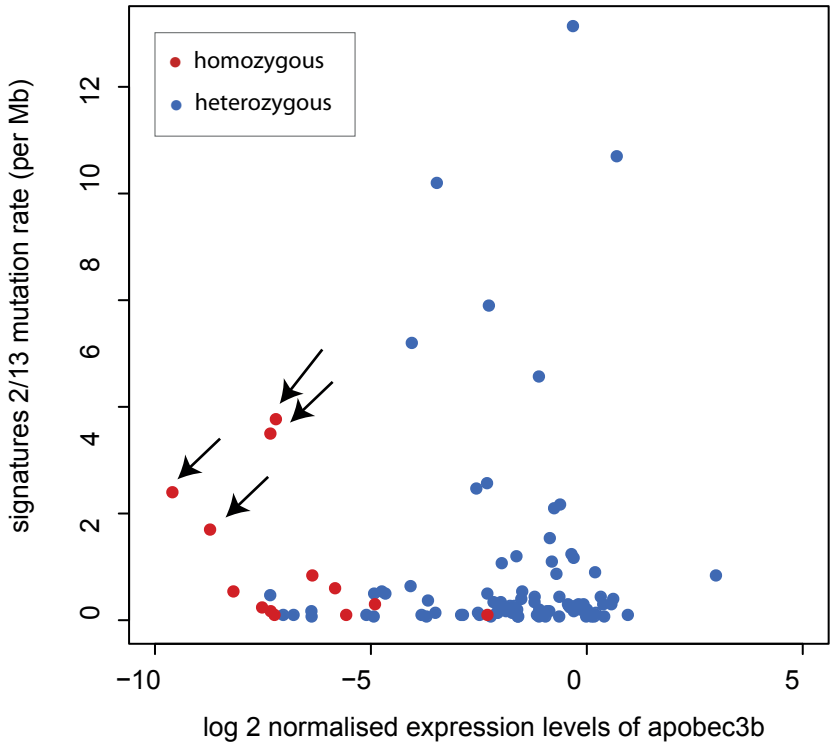
Supplementary Figure 4: Additional characteristics of Signatures 2 and 13 which resemble APOBEC-induced mutations. (A) Mutations could arise on both strands of one parental chromosome. (B) In contrast, strand-coordinated mutations arise on the same strand of a parental chromosome. (C) Principle of classification of successive pairs of variants for the purpose of demonstrating genome-wide strand coordination. Successive pairs of variants are classified as 'same' if both mutations are of the same originating base and 'diff' if not. The distance between successive pairs or intermutation distance is also calculated (IMD = intermutation distance).



Supplementary Figure 4D: Trumpet plots were generated by using a sliding window taking intermutation distances of 100 successive pairs of mutations per bin and presenting the proportion of same and different strand mutations per bin. Each bin should therefore add up to 1. The expected number of same strand (dark blue line) or different strand (light blue line) mutations was obtained for each cancer corrected for the mutation spectrum and mutation rate for each cancer. In each bin, the proportion of observed same strand mutations and different strand mutations remained light blue and dark blue respectively if there was no significant deviation from expected. If significantly more same strand mutations were observed in each bin than expected, the proportions of same strand and different strand mutations were plotted in black and red respectively. Strand-coordinated variants were more often closer together than not, suggesting that they had arisen at the same time on the same stretch of single-stranded DNA.



Supplementary Figure 5A: Demonstrating the relationship between expression levels of several different APOBECs with *APOBEC3A/3B* deletion polymorphism status in 1,691 patients with different cancers types. Hom = homozygous deletion carrier, het = heterozygous deletion carrier, non= not carrying *APOBEC3A/3B* deletion allele. Source data: Expression levels.



Supplementary Figure 5B: Relationship between expression level and deletion allele carrier status in individual breast cancers. Arrows to show that patients who are homozygous for the deletion allele have very little expression of APOBEC3B but still have a high rate of mutation of Signatures 2/13. For clarity, only patients who are heterozygous or homoyzgous are shown in this figure. Source data: Expression levels.

Supplementary Table 3A: Summary of hypermutators for each cancer type.

cancer type	hypermutators	non-hypermutators	total samples
ALL	3	37	40
BLCA	9	127	136
BRCA	106	817	923
CESC	4	34	38
HNSC	36	262	298
KIRP	5	95	100
LUAD	25	278	303
LUSC	15	150	165
MM	2	63	65
STAD	11	122	133
THCA	20	264	284
UCEC	12	222	234
TOTAL	248	2471	2719

Supplementary Table 3B: The trend for enrichment is not restricted to a specific racial group

	hypermutator			non-hypermutator				enrichment	
cancer type	carrier	non	total	carrier	non	total	total	OR	CI
Asian	5	3	8	19	14	33	41	1.23	0.3-6.0
White	11	45	56	50	436	486	542	2.13	1.0-4.4
Not Available	16	18	34	35	217	252	286	5.51	2.6-11.8