

## Chapter 3

# Genome properties and the rate of rearrangement

In Chapter 2, I introduced the PCAWG dataset of classified structural variants in over 2500 cancer samples. Having previously described the properties of SV class, size, and junction homology, I now turn to their specific location in the genome. Somatic rearrangements in clinically-detectable cancer samples reflect the distribution of events at generation, filtered by the forces of positive and negative selection. In this way, the total observed SV catalogue reveals biases about the dynamics of DNA breakage and repair, and highlights particular cancer-associated loci which recurrently drive oncogenesis through altered gene dosage, disruption, fusion, or regulation.

When considering the distribution of PCAWG SV events along the genome (Figure 3.1), a few dozen ‘hotspots’ immediately emerge at fragile sites, immune loci, and certain cancer genes under positive selection for rearrangement.<sup>a</sup> Outside these anomalous genome regions, variation in the rearrangement rate is more modest, and associates with a variety of genome properties such as replication timing and chromatin state. In this chapter, I describe a library of quantitative metrics to measure more than 30 properties across the genome (Section 3.1); show the pattern of association between these properties and the different SV classes described in Chapter 2 (Section 3.2); examine their utility for modelling the rate of rearrangement (Section 3.3); define and analyse fragile sites in the PCAWG dataset (Section 3.4); and, finally, explore the different SV patterns observed around cancer genes (Section 3.5).

---

<sup>a</sup>Note that somatic retrotransposition events were excluded from this study at the outset; some ‘hot’ L1 elements have a comparable rate of somatic activity and would also be marked in Figure 3.1 if retrotransposition was included.

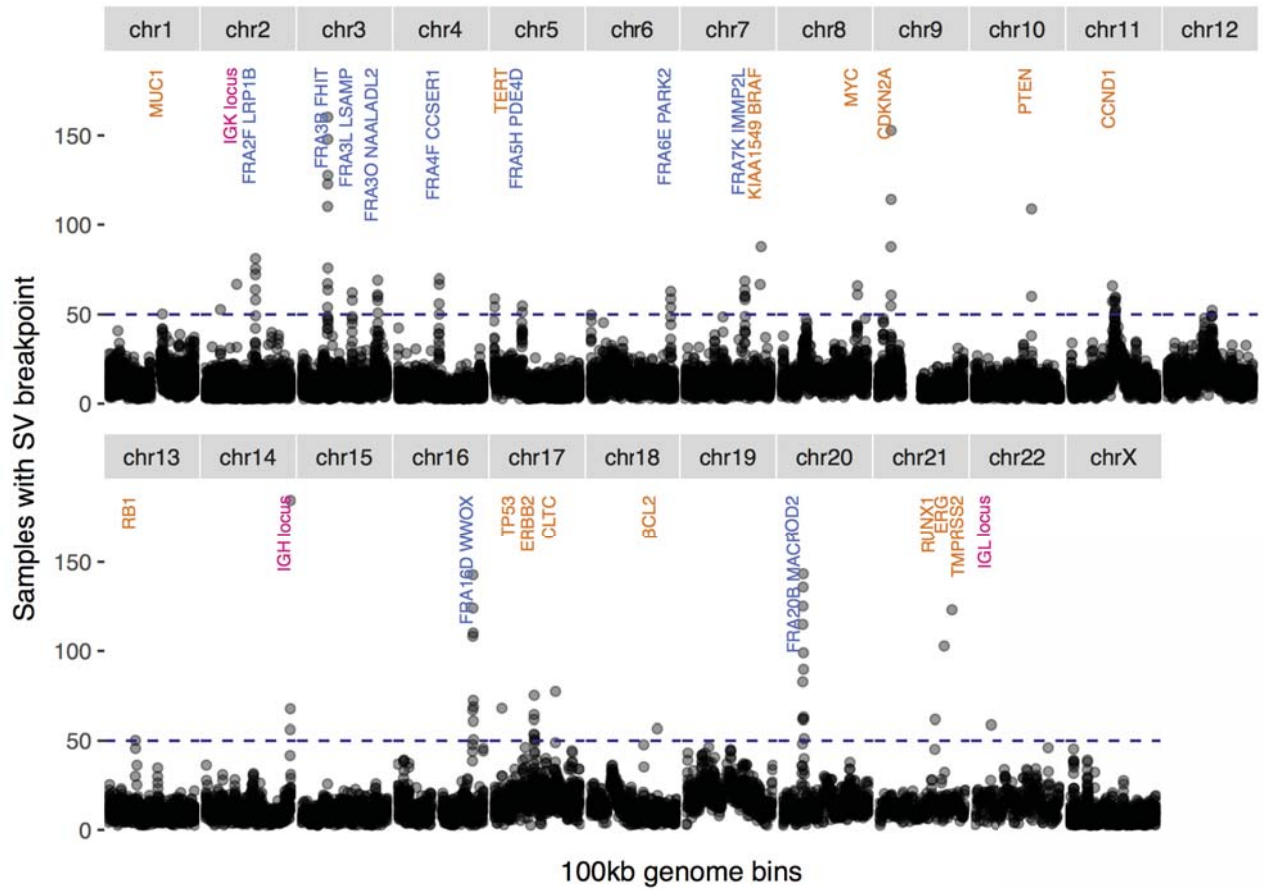


Figure 3.1: The genome-wide distribution of somatic rearrangements across 2559 PCAWG samples. Each dot records the number of samples containing a somatic SV breakpoint in a 100kb bin. Bins with breakpoints in fewer than three samples are excluded. A selection of peak regions with more than 50 rearranged samples are labelled for the presence of cancer genes (orange), fragile sites (blue), and immune loci (pink). The equal chromosome facet width means the horizontal scale is not constant across chromosomes.

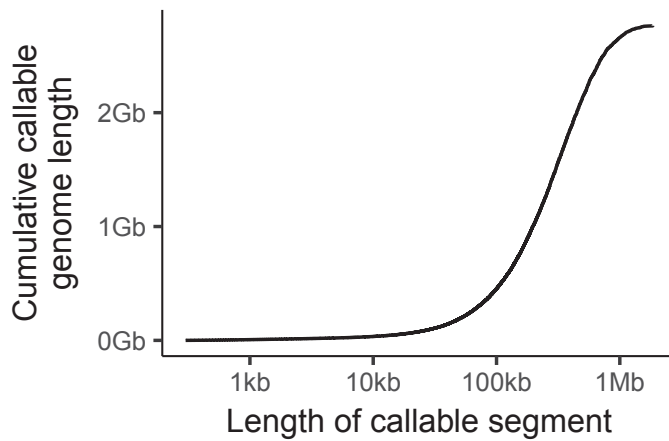


Figure 3.2: Cumulative length of callable genome regions, sorted by size.

## 3.1 A library of genome properties

### 3.1.1 Defining the callable genome

Before characterising the rate of rearrangement, I first defined the ‘callable’ subset of the hg19 reference genome to account for unmappable regions in which variants are unable to be detected.

To estimate these boundaries, I ran a random collection of 200 BAM files from PCAWG normal samples through the GATK CallableLoci tool (McKenna et al., 2010)<sup>b</sup>. Summarising results across these 200 normals, I defined the callable genome space to be positions callable in  $\geq 40\%$  of samples, such that non-callable tracts must be at least 100 bp in length, and callable regions at least 300 bp.

The resulting callable genome covers 95.3% of non-N bases in hg19 (2.76 Gb, Figure 3.2). Of the non-callable fraction, the vast majority is excluded due to consistently poor mapping quality, less than a fifth because of low coverage, and less than a thousandth because of excessive coverage.

Of 551,872 total breakpoint positions in the PCAWG cohort, only 1102 (0.20%) are outside this callable genome definition. As these 1102 positions are spread across 883 different loci in 609 samples<sup>c</sup>, I consider this a negligible discrepancy

<sup>b</sup>GATK CallableLoci v3.3-0 run with options `maxFractionOfReadsWithLowMAPQ=0.25`, `maxDepth=1000`, and otherwise default settings.

<sup>c</sup>Grouping breakpoints within 20 kb of each other, no locus contains more than 8 breakpoints in non-callable regions (worst cases: 8 breakpoints in 8 samples around *IGH* on chr14; 7 breakpoints in 6 samples in chr17:58061250–58088813; and 6 breakpoints in 5 samples in a 78 bp stretch on chr7:107410599–107410676 containing poly-T tracts). Only 15 samples have more than five breakpoints outside the callable genome.

with no strong systematic bias to affect downstream analyses, and do not filter out these calls nor do I extend the callable genome definition to encompass them. Strikingly, 63% of breakpoints outside the callable genome are returned by BRASS and just one other caller, a combination matching 13.5% of breakpoints in general. This suggests the BRASS SV calling algorithm is most vulnerable to dubious calls in regions of consistently poor mapping quality.

### 3.1.2 Defining pixel metrics

I divide the hg19 (GRCh37) human reference genome (autosomes and chromosome X) into 3,036,315 pixels of 1 kb, and calculate a suite of metrics per-pixel to summarise a variety of genome properties with potential relevance to the rate of rearrangement. The metric definitions aim to optimise three desirable, and often competing, properties: clarity of interpretation and communication; a genome-wide distribution that is (where possible) symmetric, uni-modal, and without extreme zero-inflation; and a preference for measuring local sequence effects operating at short-range.

#### Basic sequence features

The following properties are with respect to the hg19 reference genome sequence.

**GC sequence content** The calculated metric is  $(g + c)/w$  where  $g$  and  $c$  are the number of guanine and cytosine bases in the pixel, and  $w$  is the number of known (non-N) bases in the pixel. Pixels with 50% or more unknown bases ( $w < 500$ ) are disregarded.

**Sequence complexity/simplicity** The calculated metric is  $(\sum_i x_i^2)/w^2$  where  $x_i$  is the number of trinucleotide motifs of identity  $i$  in the pixel, for all possible trinucleotide motifs.

**Centromeres and telomeres** The calculated metric is  $\log_{10}(d_M + 1)$  where  $d_M$  is the distance in megabases to the feature (centromere or telomere). Centromere and telomere positions are taken from the UCSC Genome Table Browser ‘Gap’ track (Karolchik et al., 2014).



**CpG islands** The calculated metric is  $\log_{10}(d_k + 1)$  where  $d_k$  is the distance in kilobases to the nearest CpG island (zero for pixels containing one). CpG island positions are taken from the UCSC Genome Table Browser (Karolchik et al., 2014). In brief, islands are defined as segments at least 200 bp long, with GC content above 50% and more CpG dinucleotides than expected given the GC content. In total, CpG islands make up 21 Mb of genome (0.7%), with median width 562 bp and median gap between islands of 27 kb.

### Repeat sequences

The calculated metric for each of the following repeat types is  $\log_{10}(d_k + 1)$  where  $d_k$  is the distance in kilobases to the nearest annotated repeat (zero for pixels containing a repeat). Repeat sequence annotations are from Repeatmasker (repeat library version 20140131, hg19 genome build (*RepeatMasker Open-4.0*)).

**LTR retrotransposons** Long terminal repeat (LTR) transposable elements (TE) are autonomous retrotransposons with characteristic direct repeats at either end. The canonical active versions are about 5–7 kb in full, but the annotated LTR repeats are typically much shorter (< 1 kb) remnants of historic transposition activity. In total, the LTR family makes up 266 Mb of genome (9%), with median width 329 bp and median gap between repeats of 1.2 kb.

**L1 and L2** L1 and L2 TES (LINES) are autonomous non-LTR retrotransposons about 5–7 kb in their active form, although the annotated repeats are typically much shorter remnants. The median annotation width is 287 bp for L1 and 146 bp for L2, in total covering 510 Mb (17%) and 111 Mb (4%) of the genome respectively. The median gap between L1s is 470 bp and between L2s is 2 kb.

**Alu and MIR** Alu and MIR TES (SINES) are non-autonomous non-LTR retrotransposons. Alu elements have median width of 295 bp, totalling 304 Mb of genome (10%) with a median gap of 850 bp. MIR elements have median width 142 bp, covering 85 Mb (3%) with a median gap of 2 kb.

**DNA transposons** DNA transposons have a ‘cut-and-paste’ mechanism acting directly via DNA as opposed to the ‘copy-and-paste’ retrotransposon mechanism with an RNA intermediate. The canonical active versions are about 1–5 kb in full, but the annotated DNA TE repeats are typically much shorter

remnants. In total, the DNA transposon family makes up 109 Mb of genome (4%), with median width 156 bp and median gap between repeats of 2.5 kb.

**Simple repeats** Simple repeats are runs of identical motifs (mostly 1–6 bp), including single or di- nucleotide tracts. In total, they cover 35 Mb of genome (1%), with median width 36 bp and median gap between repeats of 2.7 kb.

### Non-B DNA forming motifs

Unless otherwise specified, the calculated metric for each of the following motif types is  $\log_{10}(d_k + 1)$  where  $d_k$  is the distance in kilobases to the nearest annotated motif in the non-B DNA database (version 2.0 (Cer et al., 2013); see review by Bacolla and Wells (2009)).

**Direct repeats** Direct repeats are sequences of 10–300 bp repeated directly one or more times 0–10 bp away, with the potential to form loop structures by misalignment. Their median length is 28 bp, median gap between annotations is 1.5 kb, and total in the genome is 52 Mb (2%).

**G-quadruplex forming motifs** G-quadruplex forming motifs are four runs of three G (or three C) bases, with 1–4 bp between each run (a subset of those in the non-B DNA database, guided by results in Piazza et al. (2015)). Their median length is 22 bp, median gap is 7.5 kb, and total in the genome is 4.6 Mb (0.15%).

**Triplex-forming mirror repeats** Triplex-forming mirror repeats are sequences of 10 or more bases with 90% pyrimidine (C or T) content on one strand, repeated as a mirror up to 8 bp away. Their median length is 24 bp, median gap is 4.5 kb, and total in the genome is 11 Mb (0.4%).

**Z-DNA forming motifs** Z-DNA forming motifs are alternating purine-pyrimidine tracts of 10 or more bases, excluding AT dinucleotide repeats. Their median length is 12 bp, median gap is 3.7 kb, and total in the genome is 7 Mb (0.2%).

**Cruciform-forming inverted repeats** Cruciform-forming inverted repeats are sequences of six or more bases repeated inversely up to 4 bp away. Their median length is 15 bp, median gap is 365 bp, and total in the genome is 83 Mb (3%). The calculated metric is the proportion of bases belonging to a cruciform inverted repeat in a 3 kb sliding window (i.e. considering one pixel either side).

**Short tandem repeats** Short tandem repeats are sequences of 1–9 bp repeated perfectly three or more times with no bases between. Their median length is 13 bp, median gap is 600 bp, and total in the genome is 46 Mb (1.5%). The calculated metric is the proportion of bases belonging to a short tandem repeat in a 3 kb sliding window (i.e. considering one pixel either side).

## ROADMAP Epigenomics

I derive the following properties from imputed signal tracks (Ernst and Kellis, 2015) from the Roadmap Epigenomics Consortium et al. (2015). Table E.1 details the match between each tissue type in the PCAWG cohort and one or more cell lines in the ROADMAP database, with the average taken as a tissue-matched metric. The tissue-matched definition is unique to the ROADMAP properties; all properties derived from other data are defined once, with no tissue-type information considered.

**DNase hypersensitivity** The calculated metric is the average imputed negative log  $p$ -value in the pixel from DNase-seq experiments, with high values indicating high chromatin accessibility (as required for binding of regulatory proteins etc.).

**RNA expression level** The calculated metric is the average logRPKM value in the pixel from RNA-seq experiments. RPKM denotes reads per kilobase of transcript per million mapped reads, so high values indicate high expression in the tissue type.

**DNA methylation** The calculated metric is the average fractional methylation value in the pixel from DNAMethylSBS experiments. High values indicate an increased tendency for CpG methylation at that locus in the tissue.

Table 3.1: Histone mark interpretations adapted from ENCODE Project Consortium (2012)

H2A.Z	regulatory elements with dynamic chromatin
H3K4me1	enhancers, and downstream of transcription starts
H3K4me2	promoters and enhancers
H3K4me3	promoters and transcription starts
H3K9ac	active regulatory elements, including promoters
H3K9me3	repressive mark, heterochromatin, repeats
H3K27ac	active regulatory elements, promoters and enhancers
H3K27me3	repressive mark, Polycomb repression
H3K36me3	transcribed genes, especially after first intron
H3K79me2	transcribed genes, especially at 5' end
H4K20me1	5' end of genes

**Histone marks** I chose a subset of 11 (out of 31) available ChIP-seq tracks to represent the landscape of histone modifications, as listed in Table 3.1. These 11 tracks were used for the 25-state chromatin segmentation analysis reported by the Roadmap Epigenomics Consortium et al. (2015). For each, the calculated metric is the average imputed negative log  $p$ -value in the pixel.

### Genome organisation

**Topologically associating domains** The calculated metric is  $\log_{10}(d_k + 1)$  where  $d_k$  is the distance in kilobases to the nearest TAD boundary taken from a Hi-C experiment in the IMR90 cell line of normal human embryonic lung fibroblasts (Dixon et al., 2012), lifted over to hg19 coordinates.

**Lamina associated domains** The calculated metric is the proportion of bases in a lamina associated domain in a 1.001 Mb sliding window (i.e. considering 500 pixels either side). LADs are taken from a DamID experiment by Guelen et al. (2008) in the Tig3 cell line of normal human embryonic lung fibroblasts, lifted over to hg19 coordinates.

**Nucleosome occupancy** Nucleosome occupancy is the only property for which the metric is not calculated per-pixel. Instead, for any given genome position, the raw value is taken at base-pair resolution using nucleosome occupancy data from a MNase-seq experiment by the ENCODE Project Consortium (2012) in the K562 cell line of myelogenous leukaemia lymphoblasts. High signal values

indicate core DNA wrapped around a nucleosome, and low signal indicates linker DNA between nucleosomes.

### Other properties

**DNA replication timing** For replication timing, I calculated the per-pixel average of three wavelet-smoothed signal tracks from the ENCODE Project Consortium (2012) summarizing Repli-seq experiments in three different cell lines: NHEK (normal skin, ectoderm), GM12878 (normal blood, mesoderm), and IMR90 (normal lung, endoderm). All three original tracks had a Pearson correlation of 0.93 or higher with the average track. High values indicate early replicating DNA, and low values indicate late replicating DNA.

**Germline recombination rate** The calculated metric is the germline recombination rate of the nearest SNP, using data from the HapMap consortium (Frazer et al., 2007)<sup>d</sup>.

**Protein-coding genes** The calculated metric is the proportion of bases in a protein-coding gene in a 1.001 Mb sliding window (i.e. considering 500 pixels either side). Protein-coding gene positions are taken from GENCODE v19 (Harrow et al., 2012).

### 3.1.3 Correlation between genome properties

As shown in Figure 3.3, there is a complex correlation structure between the different genome properties. The nine histone marks associated with active genes have strong positive correlations amongst themselves, and with high DNase hypersensitivity and high RNA expression. The two histone marks associated with repressive regions have a strong positive correlation with each other, and, curiously, a mild positive correlation with the histone marks for active genes. High gene density correlates with early replication timing, high GC content, low density of lamina-associated domains, and close proximity to CpG islands and TAD boundaries.

---

<sup>d</sup>[ftp://ftp.ncbi.nlm.nih.gov/hapmap/recombination/2011-01\\_phaseII\\_B37/](ftp://ftp.ncbi.nlm.nih.gov/hapmap/recombination/2011-01_phaseII_B37/)

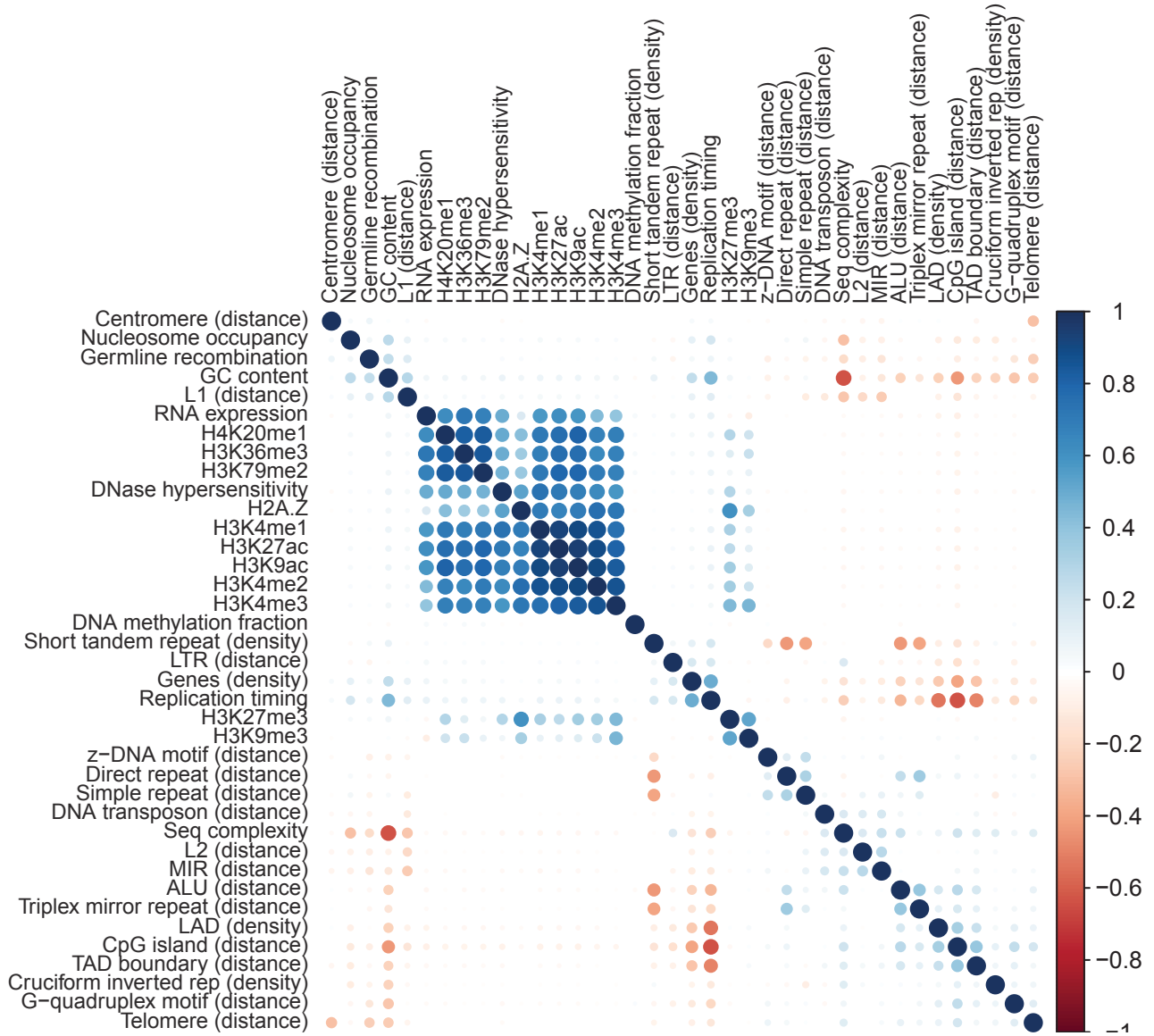


Figure 3.3: Spearman correlation between 38 genome properties at 100,000 random uniform positions in the callable genome space. Circle size is proportional to the magnitude of correlation.

## 3.2 SV classes associate with genome properties

### 3.2.1 Property quantile skew at SV breakpoints

To test for association between SV event classes and the library of genome properties described in Section 3.1.2, I compared genome property metrics between real SV positions and one million uniform random positions from the callable genome space. To compare the tissue-specific ROADMAP properties, each simulated random position was assigned a random tissue type, drawing from the observed tissue type distribution in the SV call set. To reduce dependence between observations, I only included one side of each BPJ, ensuring that the side chosen was:

- random for BPJ classified as complex, deletion, tandem duplication, unbalanced inversion, foldback, or unbalanced translocation;
- one side per motif for reciprocal translocation, templated insertions, chromoplexy, or translocation with tandem duplication (i.e. pick one side per BPJ with the stipulation that they must be in different loci);
- the outermost side for each BPJ in a reciprocal inversion, dup–inv–dup, or dup–trp–dup structure;
- the opposite side for each BPJ in a loss–inv–dup structure; and
- the distal translocation side for a BPJ in translocation with foldback or translocation with inverted insertion, and the side closest to the translocation for the partner intrachromosomal BPJ.

For each genome property and each event class (separately), I pool the real observations amongst the million random values, then rank transform and normalise on a scale from zero to one to calculate quantiles. Under the null hypothesis of no event–property association, the quantiles of the real observations would follow a uniform distribution. In each case, I assess departure from uniformity with a Kolmogorov-Smirnov test, and apply a Benjamini-Yekutieli correction for false discovery rate across the entire suite of tests, setting the reporting threshold at 0.01 FDR. In this analysis, I flip the distance-type metrics so that positions close to the feature of interest score higher than positions far away, and thus higher values correspond to signal enrichment (similar to density metrics).

Figure 3.4 presents the results for 13 of the genome properties considered, with the other 25 properties shown in Figure D.8.

Both small and large deletions (separating the groups at 10 kb) are enriched in late-replicating, AT-rich DNA, with breakpoints preferentially occurring in linker DNA between nucleosomes. Small deletions are the only SV class significantly associated with low gene density, whereas a small proportion of larger deletions skew massively towards genic regions—mostly in large gene related common fragile sites (analysed in Section 3.4). Reciprocal inversions also have a mild skew towards late-replicating AT-rich regions with breakpoints in linking DNA between nucleosomes.

Small and large tandem duplications (separating at 50 kb), templated insertion events, and unbalanced translocations are all enriched in early-replicating, gene- and GC-rich DNA, with breakpoints preferentially occurring close to ALU elements, short tandem repeats, and mirror repeats. The skew towards early-replicating DNA is particularly strong for larger tandem duplications; indeed, for every 10-point increase in the replication timing metric (roughly equivalent to a quantile position 0.1 higher/earlier), the average size of a tandem duplication at that location increases by 8%<sup>e</sup>.

Unbalanced translocations are more likely to occur close to centromeres, and also, to a lesser extent, close to telomeres. Proximity to centromeres is the only significant association observed for unbalanced inversions, and is also a very strong characteristic of foldback rearrangements. Reciprocal translocations are strongly enriched close to telomeres, and, like most SV classes, are enriched in early replicating regions.

With the exception of complex BPJ, most SV classes are positively associated with histone marks at active genes, with H3K4me3 shown in Figure 3.4 and the other histone marks shown in Figure D.8.

The general tendency of SV breakpoints (except deletions and reciprocal inversions) to occur in early-replicating, active, genic DNA has the flipside of breakpoint depletion in lamina-associated domains and L1 and LTR repeats.

Given the correlation structure between genome properties (Figure 3.3), all univariate associations must be interpreted with caution in the context of competing biological explanations, including properties not measured here.

---

<sup>e</sup> $p$ -value  $< 10^{-15}$ , linear regression of  $\log_{10}$  tandem dup size vs replication timing, converting back to the ratio interpretation on a base-pair (non-log) scale.



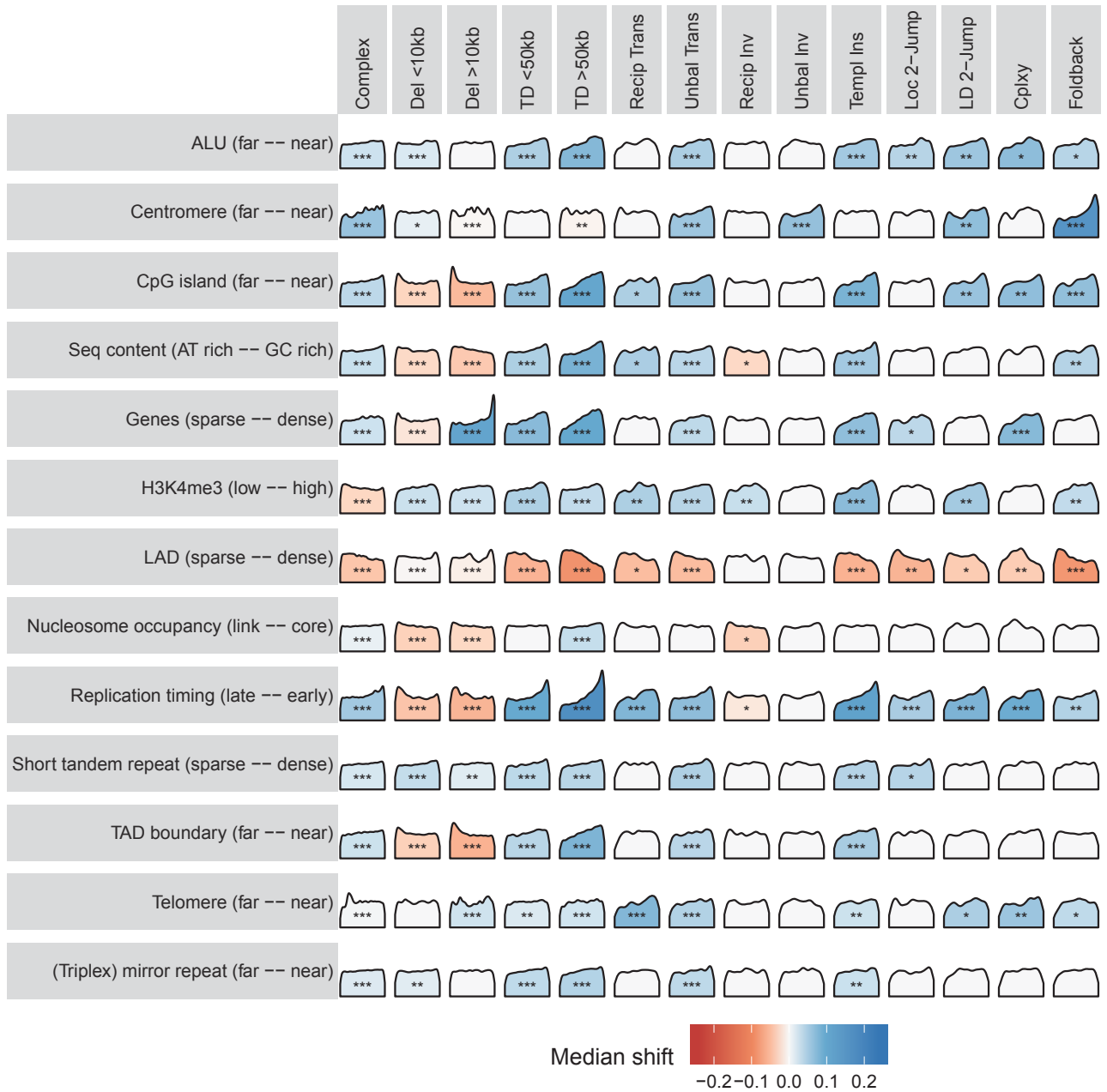


Figure 3.4: Associations between genome properties (rows) and sv classes (columns). Each density curve represents the quantile distribution of the genome property metrics at observed breakpoints compared to random genome positions, with stars indicating significant departure from uniform quantiles:  $\text{FDR} < 0.01$  \*,  $< 0.001$  \*\*, and  $< 10^{-6}$  \*\*\*. Significant property associations are shaded by the magnitude of the shift of the median observed quantile above (blue) or below (red) 0.5. The interpretation of each property metric from left to right is indicated in parentheses.

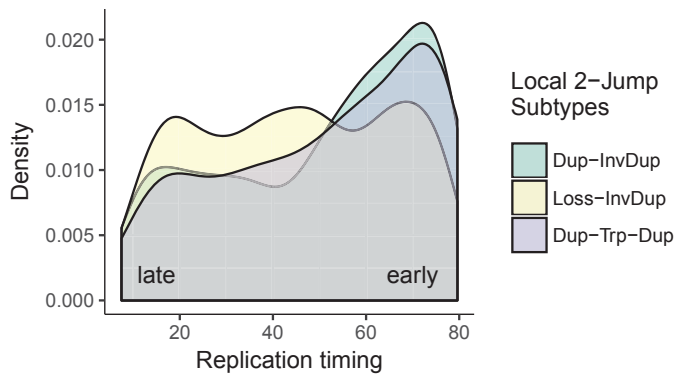


Figure 3.5: Replication timing distribution for the three sub-groups of local 2-jump SV classes. ANOVA  $p$  value for a difference between groups is  $7 \times 10^{-9}$ .

This analysis does not attempt to quantify differences between more specific breakpoint classifications—such as templated insertion bridges compared to chains or cycles—and may be averaging over subtle distinctions. For example, a comparison of the replication timing distribution of breakpoints in the three sub-groups of local 2-jump (Figure 3.5) reveals that the loss-inv-dup structure does not share the same strong preference for early replicating DNA as the dup-inv-dup and dup-trp-dup structures. Interestingly, this places the loss-inv-dup structure combining copy gain and copy loss in a middle zone between the copy gain event types with a preference for early replicating regions (tandem dup, templated insertion, and dup-inv-dup/dup-trp-dup) and the copy loss events (deletion) with a preference for late replicating regions.

### 3.2.2 Breakpoints in close proximity with short repeats

Using the property metric library, description of the positive association between SV classes and small sequence repeats is limited by the 1 kb pixel resolution, and may reflect broad correlation with other genome properties rather than specific localisation of breakpoints within repeats. To check whether these associations hold at a shorter range, I tallied the proportion of breakpoints (using one side per BPJ as described in Section 3.2.1) within a short radius around each class of SINE and non-B DNA motif. Comparing against the proportion of random uniform positions in the callable genome that also sit within these repeat radii, I checked for significant enrichment/depletion with a binomial proportion test followed by Benjamini-Hochberg FDR correction within each repeat class.

The results in Figure 3.6 confirm a significant enrichment for several SV classes around ALU elements, as well as around direct repeats, short tandem repeats, and triplex-forming mirror repeats. In most cases, any significant enrichment only accounts for an extra 1–2% of breakpoints above expectation under a

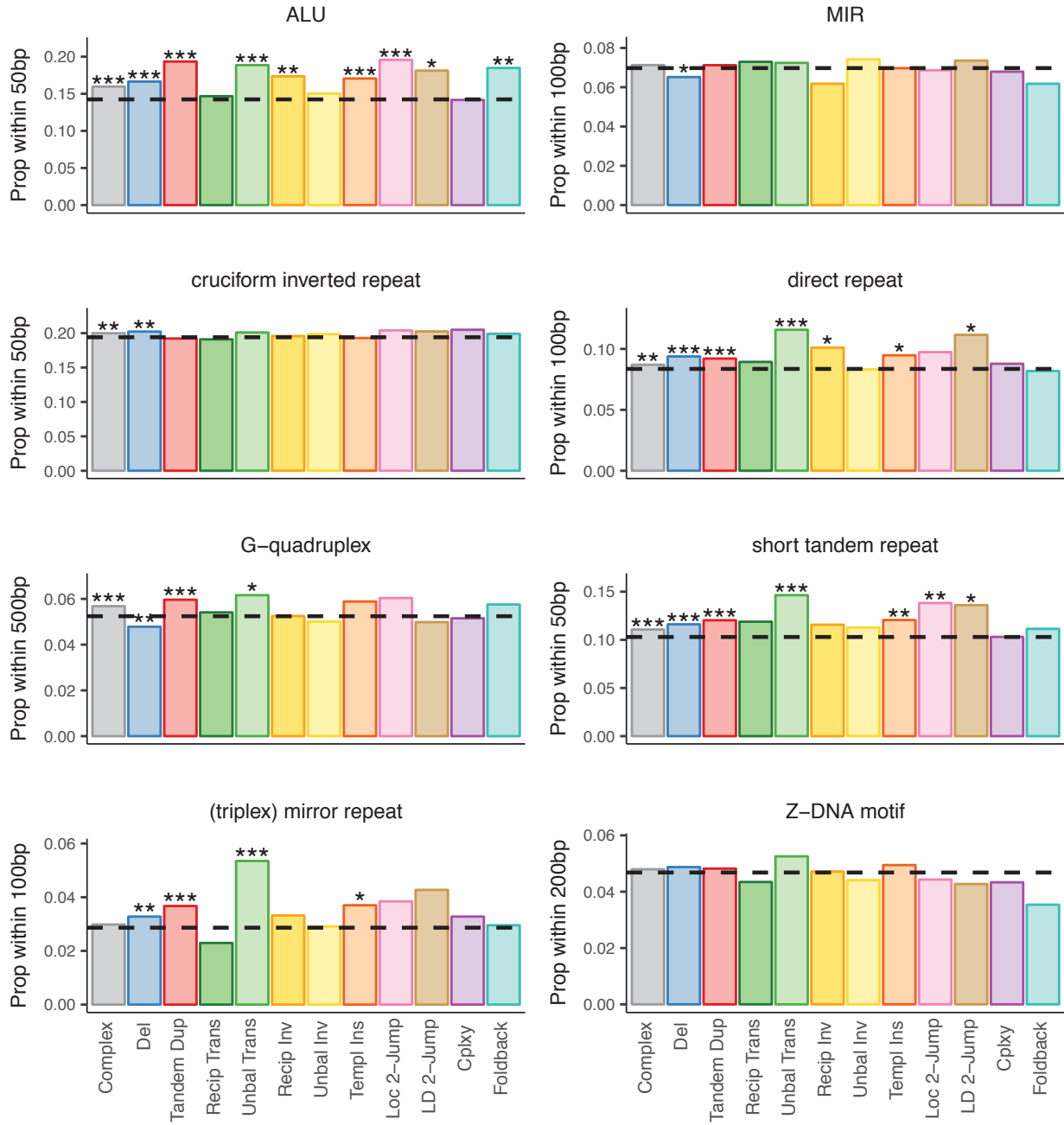


Figure 3.6: The proportion of SV breakpoints (one side per BPJ) within a short radius of each SINE and non-B DNA motif class. The proportion expected under a uniform null over the callable genome is indicated with a black dashed line. Significant departure from the uniform expectation is assessed with a binomial proportion test, marked at BH-corrected FDR:  $< 0.01$  \*,  $< 10^{-4}$  \*\*,  $< 10^{-6}$  \*\*\*

uniform null. However, the enrichment is greater for unbalanced translocation, with 11.6% of breaks within 100 bp of a direct repeat (8.4% expected), 14.6% within 50 bp of a short tandem repeat (10.4% expected) and 5.3% within 100 bp of a triplex-forming mirror repeat (2.8% expected). The ALU association is strongest for tandem duplication, with 19.3% of breaks within 50 bp of an ALU element compared to 14.3% expected. These univariate tests do not account for other correlated property associations.

### 3.2.3 Replication timing at hypermutator breakpoints

Sections 3.2.1 and 3.2.2 consider property associations of SV breakpoints grouped by classification, pooling observations across all samples and histology types. Any potential differences between samples and/or cancer types are averaged out, with results skewing towards those groups with large sample size and high rearrangement burden.

In general, I choose to avoid direct quantitative comparison of property associations between cancer types because differences in metric accuracy for each tissue would confound any biological variation in rearrangement rate. Furthermore, tissue-specific SV driver events promoted by natural selection would exacerbate biases in observed location properties if separated by histology.

To circumvent these problems with bulk histology comparison, I instead tested for variation in SV–property associations by comparing hypermutator samples with the general cohort of the same cancer type. Although many relevant genome properties could be considered, I limited this exploration to replication timing—a strong correlate of the rearrangement rate as shown in Section 3.2.1 and a reasonable proxy for other correlated properties such as GC content and gene content as shown in Figure 3.3.

For each of six cancer types<sup>f</sup> with large sample size and high rearrangement burden, I considered three SV classes: deletion, tandem duplication, and unbalanced translocation. For each SV class in each cancer type, I defined hypermutator samples to be the subset with over three times as many events as the upper quartile (0.75 quantile). Then, I modelled event replication timing as a linear regression with two predictors: hypermutator status (each hypermutator represented by one dummy variable, with the non-hypermutator samples pooled together as the baseline level); and  $\log_{10}$  event size (for deletion and tandem

---

<sup>f</sup>Breast, esophagus, liver, pancreatic (adenocarcinoma), prostate, and skin (melanoma).

duplication only, size is irrelevant for translocation). As in Section 3.2.1, dependence between observations was reduced by only including one side per BPJ. The replication timing outcome variable was taken to be the quantile value when pooled with one million uniform random positions from the callable genome. As shown in Figure 3.7, I only report those hypermutator samples whose (absolute) regression coefficient is at least 0.07<sup>g</sup> with  $p$ -value  $< 0.01$ .

Although deletions, on average, skew towards late replicating regions, some hypermutator samples have deletions significantly skewing towards earlier replicating DNA, including one breast, two liver, and four pancreatic cancer samples. In contrast, seven deletion hypermutators in the prostate group have a stronger predilection towards late replicating regions than the pool average.

Tandem duplications generally skew towards early replication, and the extent of this bias is even greater in many hypermutators, including one breast, six esophageal, eight liver, five pancreatic, two prostate, and two melanoma samples. Some hypermutators have tandem duplications in later replicating DNA than the group average, including two breast, two liver, and two melanoma samples. Note that these results for deletion and tandem duplication account for event size, which is known to vary between samples (Section 2.4).

Unbalanced translocations generally skew towards early replicating regions, with two hypermutators displaying an even stronger association with early regions (one esophageal, one pancreatic) and one translocation hypermutator skewing late (melanoma).

Figure 3.7 also lays out histology-specific replication timing for the pool of non-hypermutator samples. As discussed above, caution should be applied to general property comparisons across histology groups because the metric may not be accurate for some tissues. Although replication timing is known to vary across cell types and individuals (Hansen et al., 2010; Koren et al., 2014), it may be consistent enough to warrant modest consideration (the three contributing tracks—each from a different germ layer—had high correlation; Section 3.1.2). Of the six cancer types explored here: the late-replicating deletion bias is less pronounced in liver, and may be absent altogether in breast; the early-replicating tandem duplication bias may be absent in prostate and esophagus (aside from the hypermutators); and the early-replicating translocation bias may be absent in pancreas (aside from one hypermutator).

---

<sup>g</sup>A coefficient of 0.07 means that, on average, events in this hypermutator sample had a replication timing quantile 0.07 away from the average in non-hypermutator samples of the same cancer type.

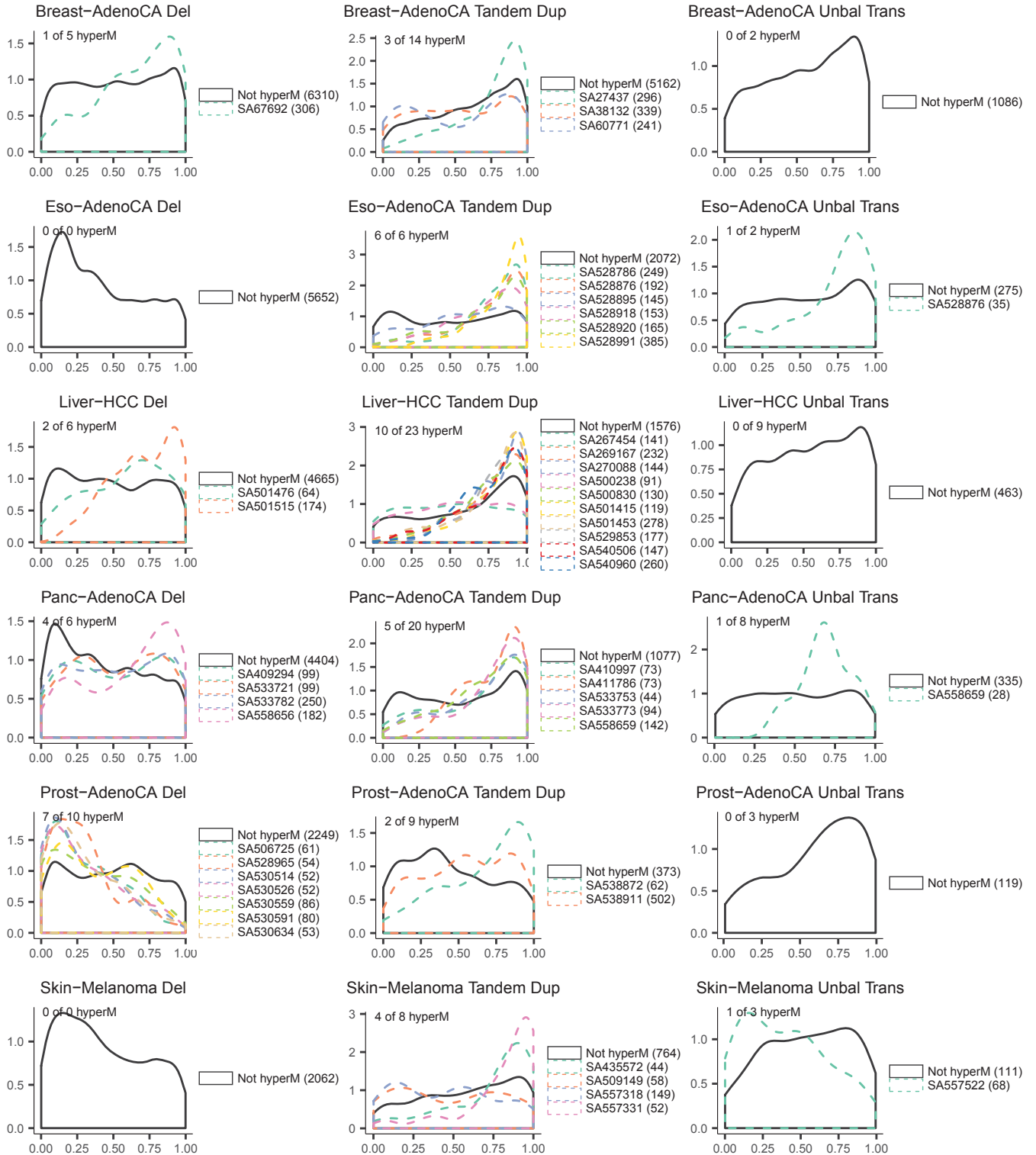


Figure 3.7: Density skew of replication timing quantiles for hypermutators compared to the pool of non-hypermutators for deletions, tandem duplications, and unbalanced translocations in six cancer types. The number of events in the sample (or pool of non-hypermutators) is indicated in the legend. Only those samples with a significant absolute average difference  $> 0.07$  are plotted, with the top-left annotation indicating how many hypermutators were considered. Low quantiles are late replicating; high quantiles are early replicating.

Assuming that characteristic patterns in hypermutator samples are signatures left by specific over-active mechanisms of breakage and/or repair, this analysis suggests that subtypes of the simple SV classes have different biases in genome location as measured by replication timing (in addition to subtypes by size, introduced in Section 2.4).

### 3.2.4 Property correlation at the junction

Sections 3.2.1–3.2.3 consider the property associations of individual breakpoint positions, selecting one side to represent each BPJ. The additional complexity of two genome positions joining in a breakpoint junction adds another dimension in which genome properties may influence the rate of rearrangement. In a companion paper analysing the same dataset, Wala et al. (2017a) found significant enrichment of BPJ within the same TAD, and significant enrichment of BPJ between repeat elements of the same class for LTRs, SINES, and LINES—partly driven by microhomology.

To extend our understanding of correlation at breakpoint junctions beyond intrachromosomal TAD structures and repeat-driven microhomology, I first considered the role of replication timing at interchromosomal BPJ.

For SV events classified as templated insertion, chromplexy, or unbalanced or reciprocal translocation, I collected the set of interchromosomal BPJ (ignoring any intrachromosomal) and took the absolute difference between replication timing estimates at either side of the junction. To compare against a null expectation that preserves the class-specific marginal distribution, I shuffled the footprint IDs within each SV class group such that the two breakpoints in a  $[+ -]$  or  $[- +]$  motif adopted the replication timing of another such motif, and any singleton breakpoints adopted the replication timing of another single break. Over ten iterations of footprint shuffling, I compared the difference in replication timing across the simulated and observed junctions.

The results in Figure 3.8 show a modest significant increase in the proportion of interchromosomal BPJ with similar replication timing. Given that replication timing correlates with physical proximity in broad nuclear compartments (Rhind and Gilbert, 2013), and, as shown in Figure 3.7, some samples have a particularly different replication timing bias, this result is somewhat expected and does not necessarily indicate a mechanistic role for rearrangements generated during replication.



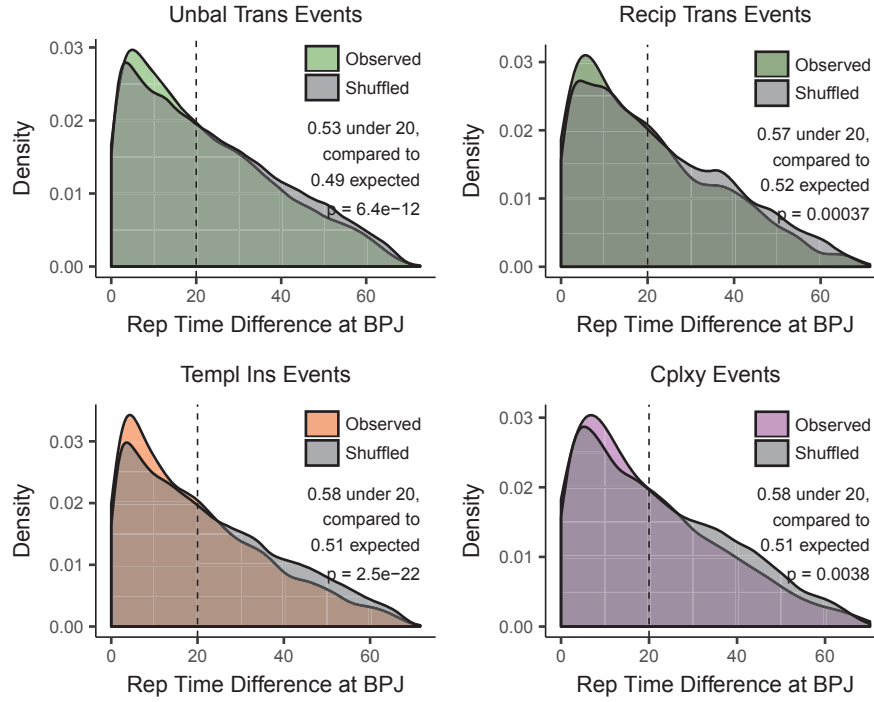


Figure 3.8: Difference in replication timing estimates across interchromosomal BPJ, compared to the expected distribution at randomly shuffled junctions. The proportion of junctions with a replication timing difference less than 20 is compared with a binomial proportion test, annotated middle right.

Nonetheless, this motivated a hypothesis that there may be a significant association between the direction of leading or lagging strand replication and the orientation of interchromosomal BPJ. Using annotations generated by Haradhvala et al. (2016) that mark about 40% of the callable genome as either predominantly ‘right’ or ‘left’ leading, I considered all BPJ with both sides in annotated regions for the same SV classes tested in Figure 3.8. About 15% of BPJ have known replication direction at both sides. Annotating + orientation breakpoints in right replicating regions and – orientation breakpoints in left replicating regions as “type 1”, and the reverse cases as “type 2”, I tested the null hypothesis that 25% of junctions are both type 1, 25% are both type 2, and 50% are type 1 and 2. Using a  $\chi^2$  goodness-of-fit test, I found no significant associations between the replication strand direction and BPJ orientation for translocations or templated insertions or chromoplexy.

For any future analysis quantifying correlations between junction sides, it may indeed be sufficient to consider only physical proximity (including TAD structure) and homology, as demonstrated by Wala et al. (2017a).



## 3.3 Modelling the rate of rearrangement

In addition to the biological insight about factors affecting genome alteration, the other major reason for characterising genome property associations is the need for appropriate mutation rate models to underpin recurrence-based driver discovery<sup>h</sup>. To explore the utility of my genome property library (Section 3.1) for predicting rearrangement rate along the genome, I aimed to fit multivariate logistic regression models to distinguish real SV breakpoints from a background of randomly distributed positions. This exercise also serves to test the strength of property associations (Section 3.2) in a multivariate setting.

### 3.3.1 Methods

#### Outcome variable

Each logistic regression model considered the set of observed breakpoints for a given SV class (one side per BPJ) against one million uniform random positions in the callable genome space. The six SV classes were: small and large deletion (split at 10 kb); small and large tandem duplication (split at 50 kb); unbalanced translocation; and foldback.

#### Predictor variables

To reduce multicollinearity among the predictors, I followed guidelines by James et al. (2013) to remove three (of 38) property library metrics with variance inflation factor above five. The three discarded variables with high correlation to other predictors were the histone marks H3K9ac, H3K4me2, and H3K4me3.

The remaining 35 predictors (Section 3.1) were scaled to have mean zero and variance one. All random genome positions were assigned tissue-specific ROADMAP property metrics according to an empirically matched tissue distribution. No interaction or histology model terms were included.

---

<sup>h</sup>Driver discovery methods aim to distinguish positively-selected cancer loci from predisposed mutational hotspots with negligible fitness effect, and require background mutation rate models to account for bias in the formation distribution.

### GLM models with lasso regularisation

Lasso regularisation on a generalised linear model (GLM) performs variable selection by restricting the absolute coefficient sum to a total budget, naturally forcing coefficients to zero as the budget shrinks. To find the optimal lasso tuning parameter (budget constraint) for logistic GLM with each SV class, I ran five-fold cross validation in a two-thirds training set to find the model with minimal classification error. Using this optimal model from the training set, I recorded model predictions for the separate testing third, and then finally report coefficients fitted to the whole dataset.

GLM lasso models were fitted with the `glmnet` (v2.0-13) R package by Friedman et al. (2010). Coefficient confidence intervals and significance were calculated with the `selectiveInference` (v1.2.3) package which accounts for the lasso selection procedure (Lee et al., 2016; Taylor and Tibshirani, 2017).

### GAM models with lasso-type regularisation

Generalised additive models (GAM) allow predictors to have a non-linear effect, typically via a spline function. Extending the concept of lasso regularisation to the GAM case, the `gamse1` (v1.8-0) package by Chouldechova and Hastie (2015) restricts the (adjusted) coefficient sum in a similar way, such that increasing the budget constraint reduces spline terms to linear terms and linear terms to zero (predictor removal). To find the optimal lasso-type tuning parameters for logistic GAM with each SV class, I ran five-fold cross validation in a two-thirds training set to find the model with minimal classification error. As above, I used this optimal training model to record predictions for the separate testing third, and then finally report coefficients fitted to the whole dataset. Spline functions were constructed from at most ten orthonormal basis functions of degree five.

### 3.3.2 Results

Figure 3.9 illustrates the coefficient paths in each GLM as the lasso tuning parameter reduces the total budget from unlimited to zero. For the optimal model choice with the best cross-validation performance, the coefficients and their confidence intervals are shown in Figure 3.10. In contrast to the strictly linear effects allowed in the GLM model, the GAM regressions permit non-linear

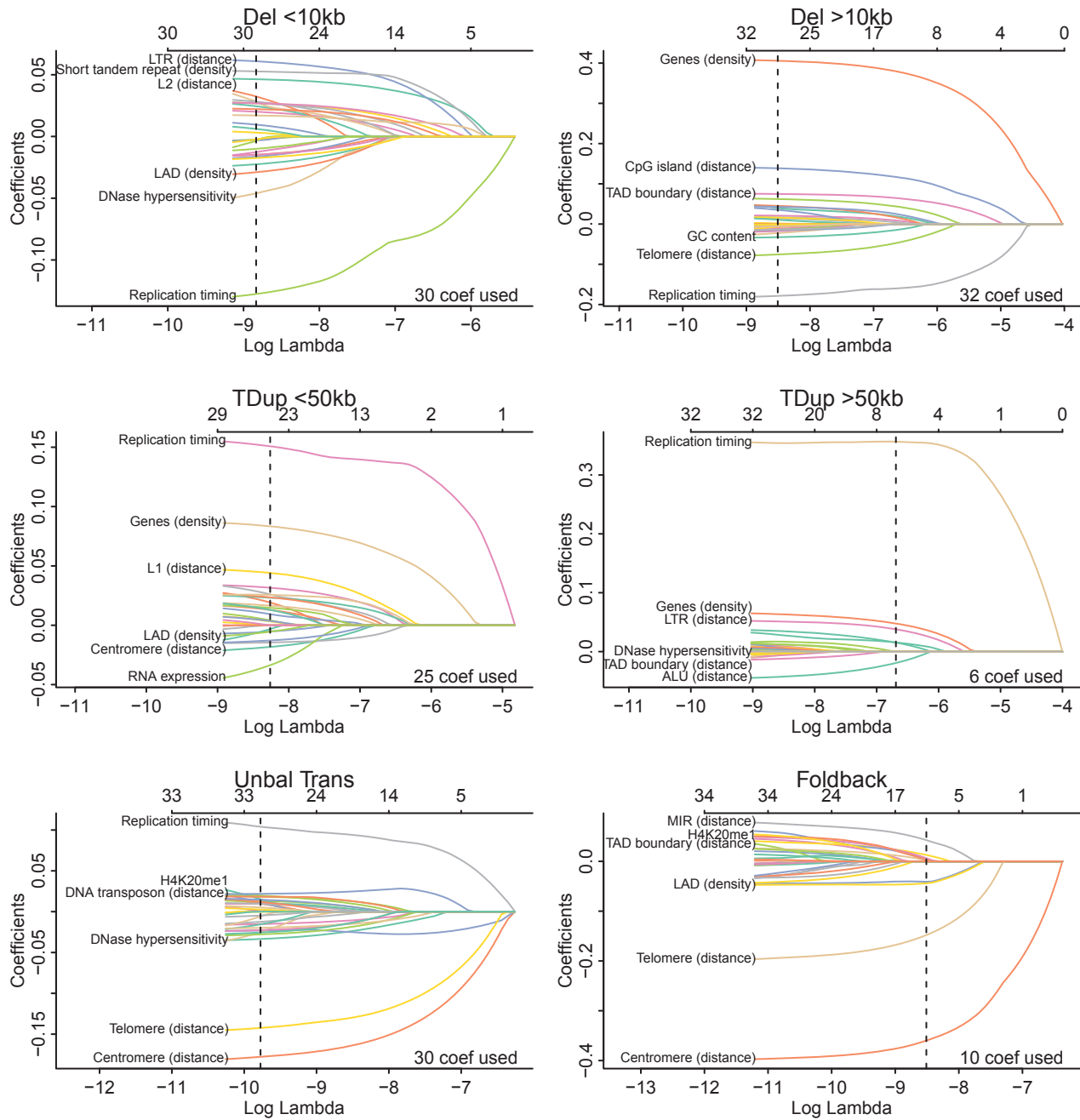


Figure 3.9: The lasso tuning parameter controls coefficient paths and number of selected predictors (annotated top) for logistic GLMs classifying real and random breakpoint positions for six SV classes. The optimal tuning parameter (best cross-validation performance) is marked with a vertical dashed line, and the number of included predictors is annotated bottom right. A subset of the most predictive coefficients are labelled on the left.

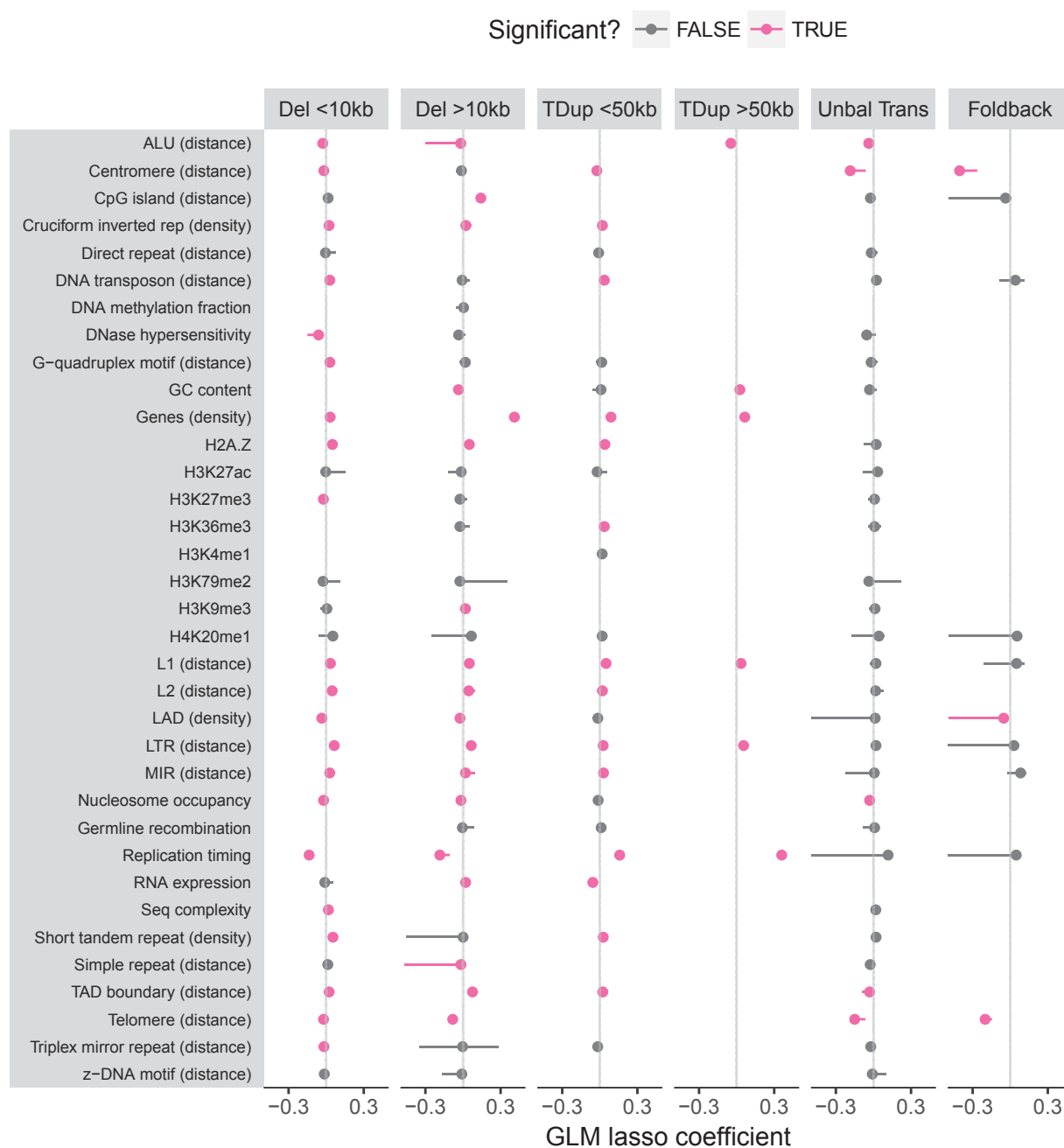


Figure 3.10: Fitted coefficient values (dots) and their confidence intervals (horizontal lines) for predictors in the optimal lasso GLM for each SV class, coloured by lasso-adjusted significance below a 0.05 type 1 threshold. Vertical guide lines mark zero.

spline effects. The optimal GAM fit for small tandem duplications is shown in Figure 3.11, with models for the other SV classes shown in Figures D.9–D.13.

To interpret the direction of predictor effects on the log odds of a position being a real breakpoint, recall that high replication time values are early, and that (unlike the reversed distances used in Section 3.2.1) high distance metrics are *far* from the feature while high density metrics are *close* to the feature.

Different SV class models select different subsets of the 35 available predictors to achieve optimal classification performance. For the GLMs, only six predictors are included for large tandem duplication, whereas the large deletion model uses 32 predictors. For the GAMs, just one predictor (centromere proximity) is included for foldback, whereas the small deletion model uses 31 predictors.

The major findings from Section 3.2.1 are recapitulated in the multivariate setting, with replication timing a strong predictor of deletion (late) and tandem duplication (early). High gene density stands out as predictor for large deletion, whereas centromere and telomere proximity are the most important predictors of translocation and foldback. Interestingly, although gene density skews low for small deletion in a univariate dimension (Figure 3.4), when conditioning on other properties in the multivariate model, small deletions have a significant association with high gene density in both GLM and GAM models.

The non-linear GAM terms offer more detailed insight into the domain of a predictor’s effect. For translocation, small deletion, and both tandem duplication sizes, the GAM models suggest that replication timing effects are specifically limited to the earliest few deciles. Other non-linear associations include small tandem duplications with mid-range values of the active histone mark H3K36me3, and small deletions with mid-range values of the repressive histone mark H3K9me3. Despite these hints at non-linear effects, when the predictive performance of the GLM and GAM models is compared on a held-out test set, the difference between them is minimal (Figure 3.12). The similar area-under-the-curve (AUC) performance metrics of the two approaches suggests that linear terms are generally adequate for rearrangement rate estimation with this property library.

To illustrate the predicted rearrangement rate with the GLM model, Figure 3.13 plots the average prediction in 10 kb bins for each SV class along two chromosomes, normalising the rates to have the same total sum. As the ROADMAP predictors are tissue-specific, the illustration is chosen for breast tissue properties. Notable features include: the predicted increase in foldback rate around

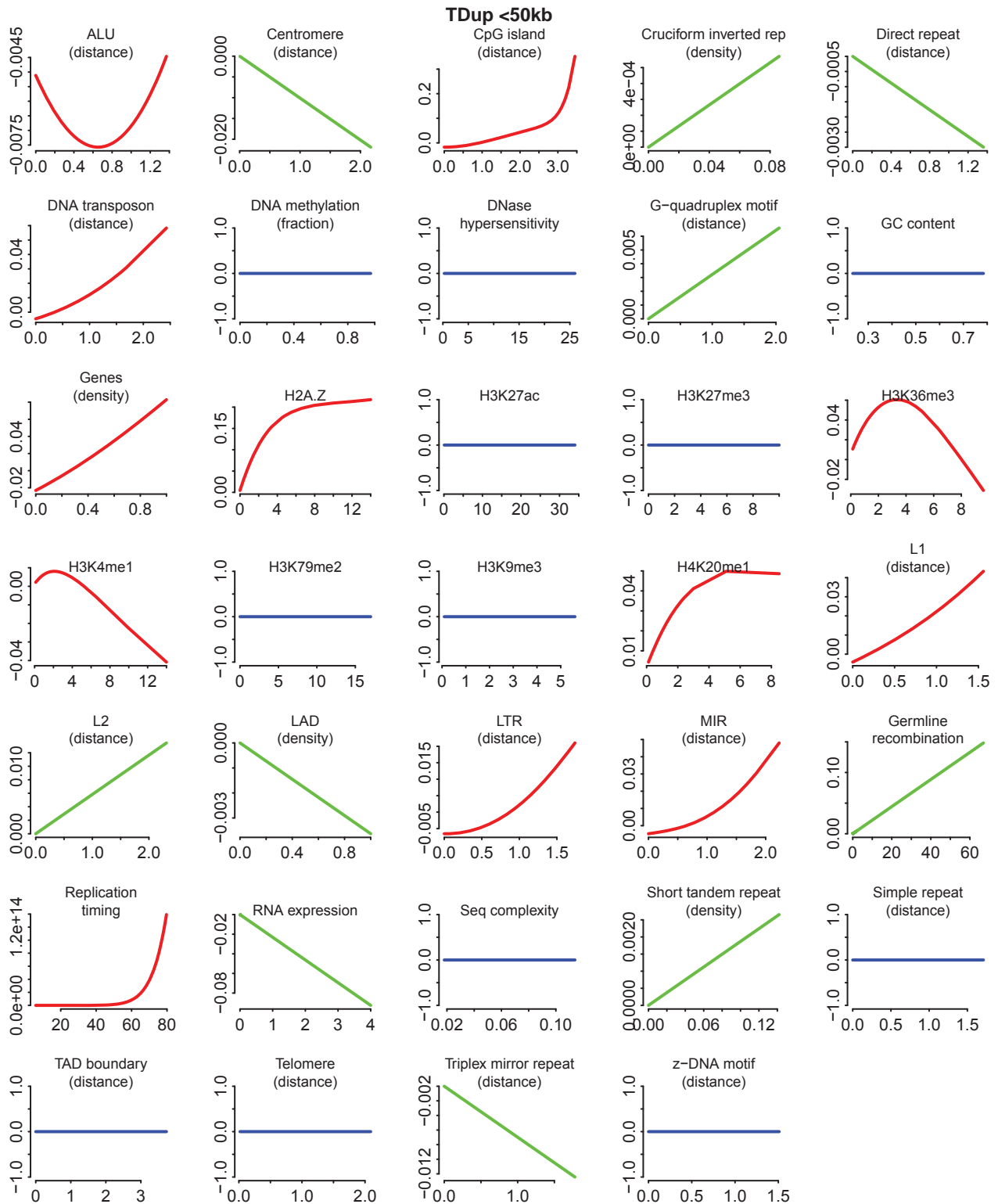


Figure 3.11: The optimal (best cross-validation performance) logistic GAM with lasso-type regularisation for small tandem duplications. The effect on the log odds of a position being a genuine breakpoint is shown as a function over each predictor's domain (back-transformed from scaled model predictors), in red for splines, green for linear terms, and blue for removed predictors.

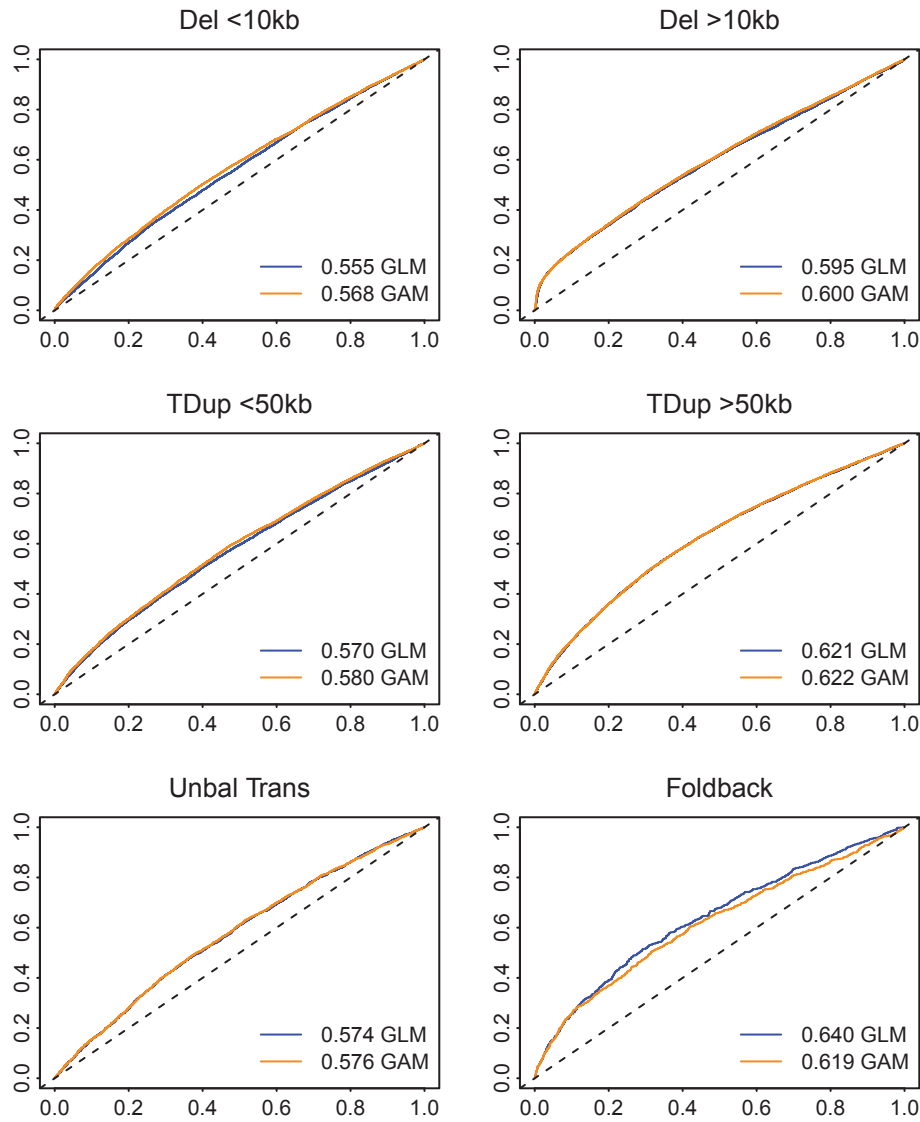


Figure 3.12: The ROC curves for true positive rate (vertical axis) and true negative rate (horizontal axis) in the testing subset for GLM and GAM models classifying real and random breakpoint positions for six SV classes. The total area under the curve is annotated bottom right, quantifying the degree of improvement compared to random guessing with area 0.5 (dashed line).

each centromere<sup>i</sup> and—to a lesser extent—telomere; the predicted increase in large deletion rate around loci with high gene density, including two fragile site genes (Section 3.4) annotated on chromosome 16; and the general tendency for large tandem duplications to have greater rate fluctuations (but in the same direction) as their smaller counterparts.

### 3.3.3 Discussion

In this section, I explored the utility of GLM and GAM logistic regression for distinguishing genuine breakpoints from uniform random genome positions. As shown in Figure 3.12, these modelling strategies achieve AUC performance ranging from 0.56 for small deletion to 0.64 for foldback. As the two outcome categories have substantial physical overlap, the AUC metric does not hold its standard interpretation as a value between 0.5 (no predictive power) and 1.0 (perfect predictive power). Rather, the upper bound is an unknown value less than one, which depends on the true breakpoint distribution’s departure from uniformity. It is unclear whether the observed performance around AUC 0.6 reflects a genuine upper bound on achievable classification, or that the model predictors do not adequately describe all factors influencing rearrangement rate. Quantifying the fraction of unexplained variance is beyond the scope of this work, as the standard  $R^2$  statistics are not applicable to logistic regression.

My exploratory attempt at rearrangement rate modelling did not consider interaction terms, histology differences, or finer SV class distinctions, any of which might improve the model fit. In particular, the illustration of predicted rearrangement rate in Figure 3.13 shows the massive rate hikes predicted for large deletion in certain loci with extremely high gene density. This gene density metric encompasses fragile sites in large genes, and causes the predicted rate to skyrocket in any similar region, fragile site or no. As discussed in the following Section 3.4, most fragile sites are characterised not only by long genes, but also by late replication time. To more accurately predict the deletion rate without dummy variables for known or suspected fragile loci, it would be advisable to include an interaction term between gene density and replication timing. As it stands, of the predicted deletion peaks shown in Figure 3.13, only two correspond to real fragile sites (Section 3.4), while the others correspond to large genes in earlier replicating regions without such a high deletion rate,

---

<sup>i</sup>The *q*-arm side of the chromosome 16 centromere is missing because that region is not included in the callable genome definition from Section 3.1.1.



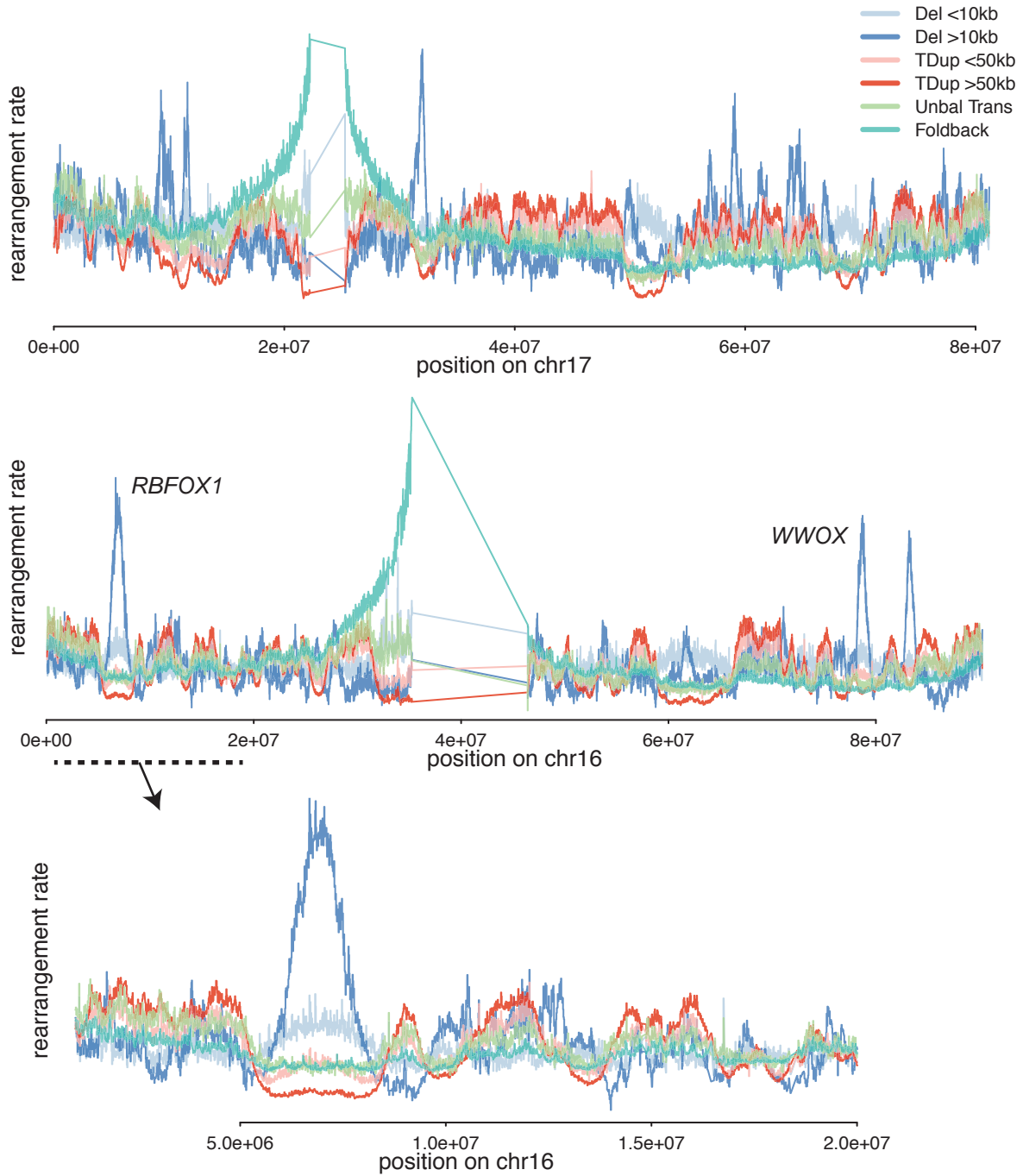


Figure 3.13: Predicted GLM rearrangement rate for six SV classes in 10kb bins along chromosomes 16 and 17, using the breast tissue-specific ROADMAP predictors. Two peaks in the predicted large deletion rate around fragile site genes *RBFOX1* and *WWOX* are annotated on chr16.

but the current model is unable to accommodate this distinction without an interaction term.

Overall, these models demonstrate important differences in the rearrangement rate of different SV classes, and suggest that SV breaks should not be modelled as one generic process. Future work could develop more sophisticated models including interaction terms, with SV classes divided by specific signatures of size, sample, histology, and property association.

### 3.4 Fragile sites and other anomalous genome regions

Within the human genome, there are several regions (besides centromeres and telomeres) with unusual properties and roles. For example, the short *p*-arms of acrocentric chromosomes contain large clusters of ribosomal RNA genes termed nucleolar organising regions. Due to their highly repetitive nature, these regions are missing from the human reference genome and their possible contribution to the cancer rearrangement landscape is largely unknown (McStay, 2016). Other anomalous regions include: the mitochondrial genome; immune loci encoding hyper-variable immunoglobulin products following V(D)J recombination; and the sex chromosomes with different gender dosage and random X inactivation in female cells.

In this section, I focus mainly on particular regions termed common<sup>j</sup> fragile sites (FS), reviewed by Sarni and Kerem (2016) and Glover et al. (2017). Cytogenetic studies first characterised FS bands by their innate propensity to develop gaps and breaks under replication stress<sup>k</sup>. Wilson et al. (2015) proposed a transcription-dependent double fork failure model to account for the cell-type-specific FS locations within unusually long, late-replicating genes. As FS genes have a paucity of dormant replication origins, contain difficult-to-replicate sequences, and suffer replication interference from transcription bubbles, conditions of replication stress may cause un-replicated regions between two stalled forks to persist into M phase. These lesions often resolve as deletion SVs, and cause a high rate of focal deletion at fragile sites in cancer genomes (Le Tallec et al., 2013).

---

<sup>j</sup>‘Common’ FS because they are common to all individuals, as opposed to rare FS which express fragility only in certain polymorphic forms.

<sup>k</sup>In vitro replication stress typically induced by the DNA polymerase inhibitor aphidicolon.

### 3.4.1 Defining fragile sites in the PCAWG cohort

To define the set of fragile sites with appreciable activity in the PCAWG dataset, I split the genome into 500 kb tiles sliding every 50 kb and calculated the density of deletion breakpoints, normalising by the length of callable regions (Section 3.1.1) within each tile. As an initial set of fragile candidates, 56 contiguous regions had deletion breakpoint density above 100 breaks/Mb<sup>l</sup> for at least 500 kb, and an absolute deletion break count over 100. Fragile sites are characterised not only by high deletion density, but also by the predominance of deletion events above all other SV classes. Considering the proportion of breaks classified as deletion in each candidate region, I set thresholds at > 42% for candidates overlapping known CFS<sup>m</sup> and > 50% otherwise. After removing some regions overlapping known cancer census genes (*ERBB4* and *GPHN*) and the *IGK* locus on chr2, 27 candidate fragile sites remained, including 22 overlapping known CFS. Of these 27 fragile regions (listed in Table E.2), 21 are located at long protein-coding genes and are used in downstream analyses. Three fragile genes have no overlap with a known CFS: *CSMD1* on chr8; *PTPRD* on chr9; and *RBFOX1* on chr16. The six fragile regions without an explanatory transcript are not carried forward in the rest of this section.

### 3.4.2 Fragile site activity

Figure 3.14 illustrates the nine most active FS, sorted by the number of samples affected (see Figure 3.15A for ranking, and Figure D.14 for the other twelve FS). Deletions are particularly enriched in fragile sites, accounting for 64% of all breakpoints in the nine major FS, and 54% of all breakpoints in the other twelve FS. Indeed, 9.7% of all deletions have both ends inside these FS regions which span only 1.4% of the callable genome. Tandem duplications and reciprocal inversions are also mildly enriched in fragile sites, with 2.3% of each event class within the bounds of a FS.

Fragile site tandem duplications tend to occur in the same samples as FS deletions, suggesting a similar aetiology. Outside the FS, most deletions and tandem duplications in the cohort are observed in breast, ovary, and liver cancer samples. However, inside the FS, esophageal cancers contribute the

<sup>l</sup>per total cohort, not per sample!

<sup>m</sup>Using 109 CFS defined in the Supplementary Materials from Bignell et al. (2010) and Le Tallec et al. (2013), lifting over to hg19 coordinates and using the UCSC Genome Browser to find coordinates of cytogenic bands where necessary.

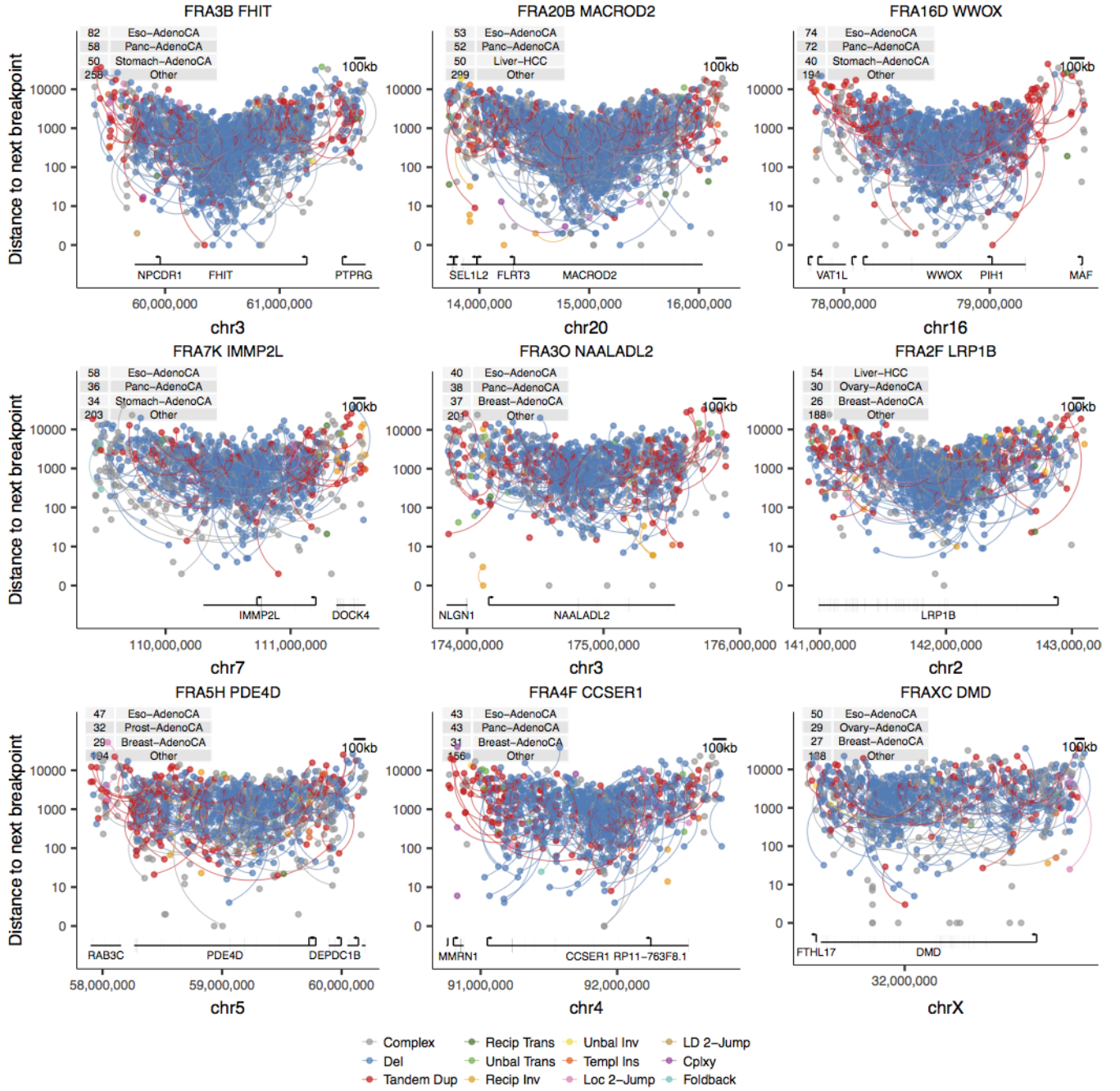


Figure 3.14: All sv breakpoint positions in nine major fragile sites, sorted by number of affected samples. Breakpoint positions are coloured by classification, and vertically spaced by the distance to the next breakpoint in the cohort. If the two sides of a BPJ are contained within the plotting window, they are joined with a curved line. The number of samples with a breakpoint in the plotting window is annotated top left.

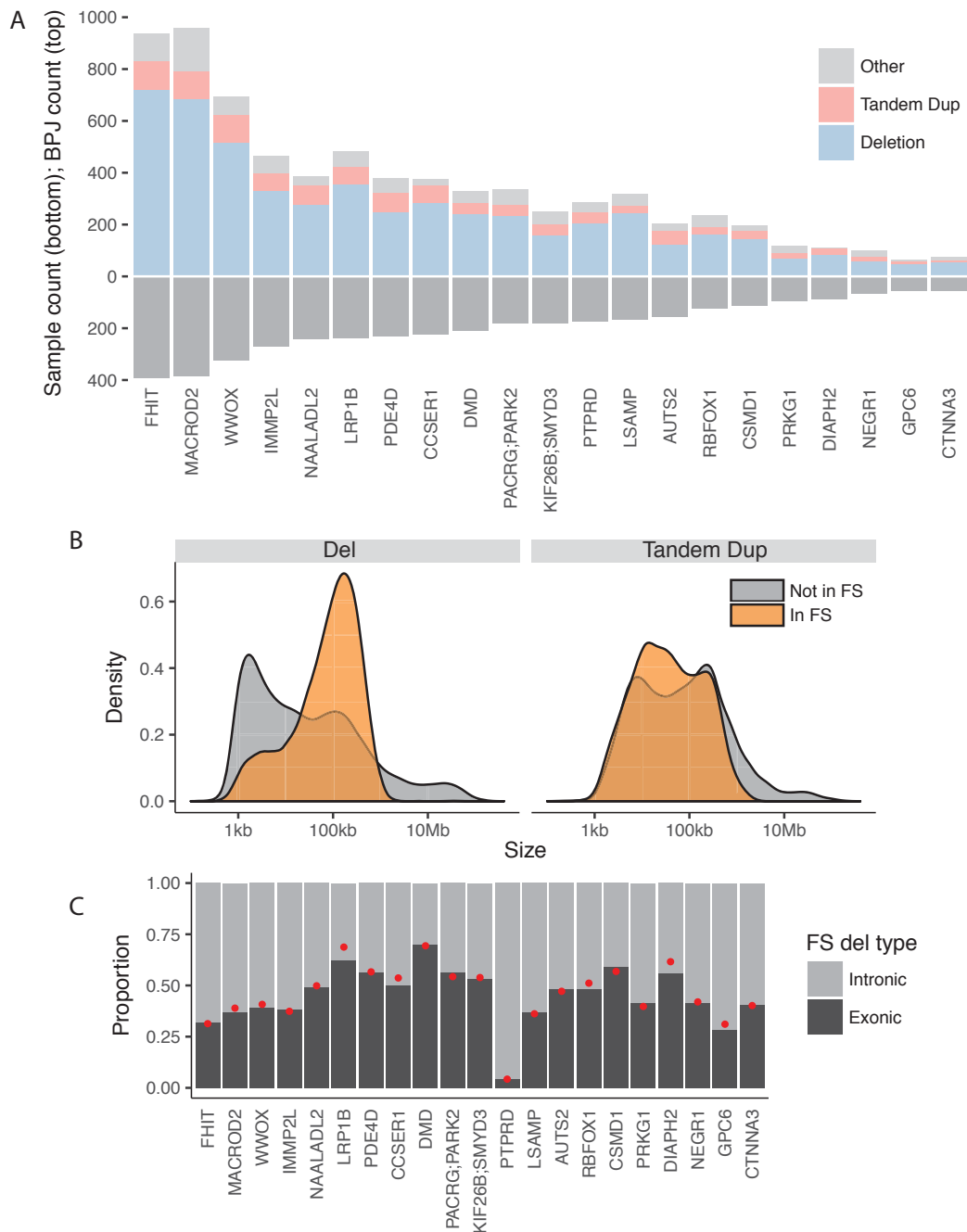


Figure 3.15: (a) number of deletions, tandem duplications and other BPJ within each of the 21 fragile sites considered (upper), sorted by the number of affected samples (lower); (b) size distribution of deletions and tandem duplications in fragile sites compared to the rest of the genome; (c) proportion of FS deletions intersecting an exon (plus 5 bp flanks), with the red dot indicating the proportion expected by random chance.

most deletions *and* tandem duplications. The Pearson correlation between the number of FS deletions and FS tandem duplications in a sample is 0.52<sup>n</sup>.

The size distribution of FS rearrangements differs from the general genome-wide distribution (Figure 3.15), with fragile deletions skewed towards larger events above 100 kb (on average, 2.2–2.4 times larger than a non-FS deletion<sup>o</sup>) and fragile tandem duplications skewed towards smaller events below 100 kb (on average, 1.7–2.1 times smaller than a non-FS tandem duplication<sup>o</sup>).

Figures 3.16 and D.15 show how closely the fragile site definitions correspond to dramatic local peaks in deletion density. Most FS have a symmetric ‘bell-shaped’ deletion distribution, with notable exceptions including: *DMD* with a peak in the 3’ gene end and a long tail stretching across to the TSS<sup>p</sup>; and *FRA1I* with a peak over the *SMYD3* gene and a tail stretching over the adjacent *KIF26B* gene (see Figure D.14 for gene positions). Some of the less active fragile sites may be imprecisely defined, with areas of elevated deletion density flanking the *GPC6* and *PRKG1* regions. As expected, these FS definitions correlate with late replication, and sometimes co-locate almost perfectly with a local timing dip between two early loci (presumably between replication origins). *FHIT*, *WWOX*, *PACRG*; *PARK2*, *LSAMP*, *RBFOX1*, *PRKG1*, and *DIAPH2* are all good examples of fragility demarcated by protein-coding genes in a local replication timing dip. Reassuringly, all three fragile genes without a known CFS overlap have very late replication, supporting genuine fragility over positive selection. These plots also illustrate slightly elevated rates of tandem duplication at some fragile sites, and suggest a possible enrichment in the edge regions—as previously reported by Wilson et al. (2015)—where replication forks may tend to stall. This duplication effect is most noticeable in the weaker fragile sites like *AUTS2*, *PRKG1*, and *DIAPH2* whose vertical scales (Figure D.15) do not compress the duplication track, but is also hinted at for some of the more common sites like *PDE4D*, *IMMP2L*, and *NAALADL2*.

### 3.4.3 Fragile site deletions are mostly intronic

With FS genes accounting for about half the recurrent deletion foci in cancer genomes (Le Tallec et al., 2013), the question of whether these events drive the cancer phenotype has received ongoing attention. Genuine tumour suppres-

<sup>n</sup>*p*-value testing null hypothesis of zero correlation is  $< 10^{-15}$ .

<sup>o</sup>95% confidence interval for mean difference between FS and non-FS events, using a *t*-test on the log<sub>10</sub>-scale and then converting back to the ratio on a base-pair scale.

<sup>p</sup>*DMD* is on the – strand.



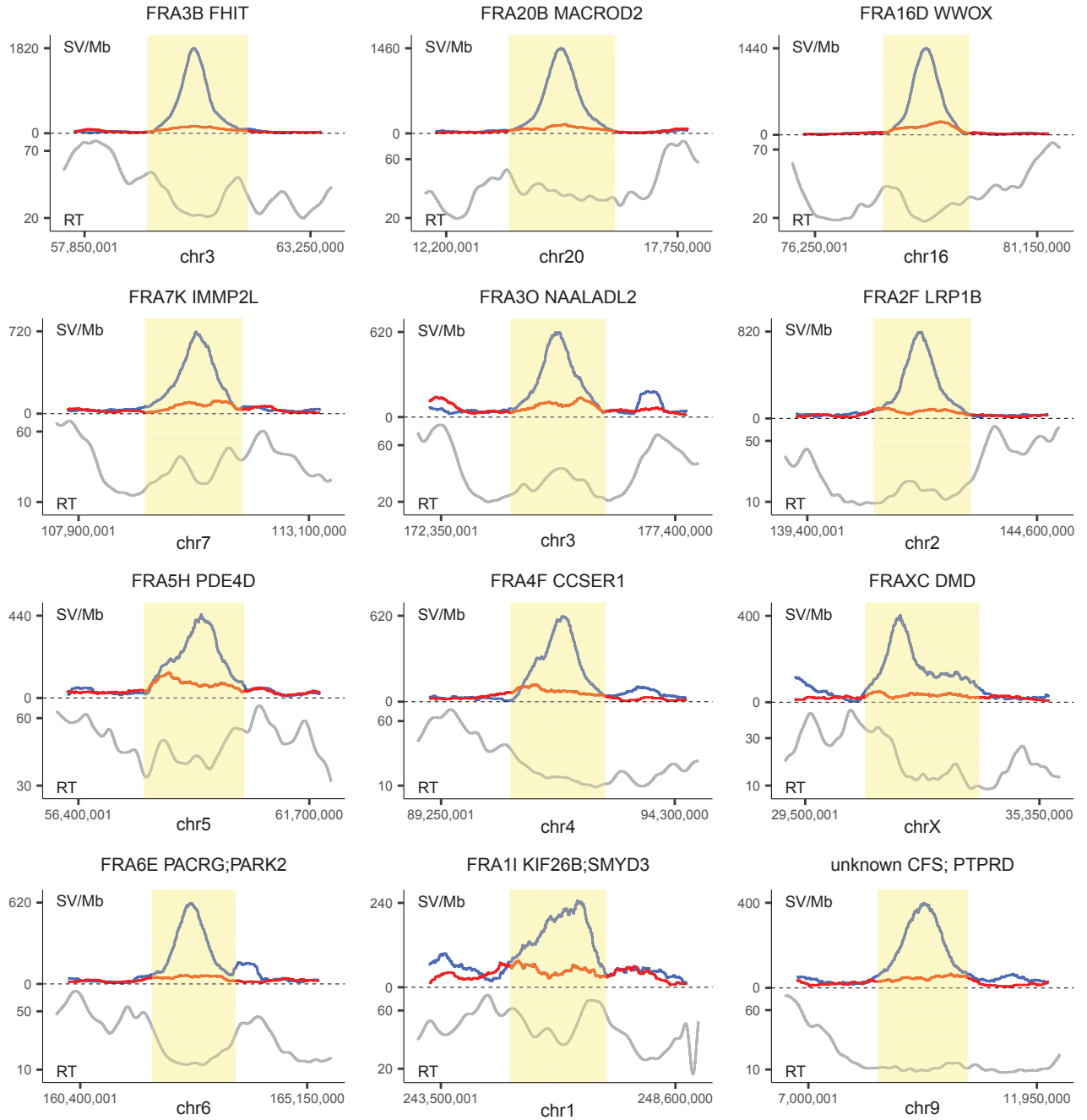


Figure 3.16: The upper plot shows the density of deletion (blue) and tandem duplication (red) breakpoints in 500 kb windows sliding every 10 kb for the 12 major FS marked in yellow, with 2 Mb flanks either side. The lower plot shows the replication timing track, with high values for early and low for late.

sor genes whose disruption is subject to positive selection typically have an enrichment of inactivating point mutations and/or homozygous loss, neither of which are observed for FS genes (Bignell et al., 2010; Lawrence et al., 2014). On the other hand, some functional studies support a tumour suppressing role for *FHIT*, *WWOX*, and *PARK2* (Gong et al., 2014; Karras et al., 2017; Glover et al., 2017), and it remains entirely plausible that a subset of FS deletion confers a modest selective advantage.

To capitalise on the precise breakpoint resolution of this WGS dataset, I compared the observed frequency of exon disruption by FS deletion with the rate expected by random chance. I marked any event crossing within 5 bp of a FS gene exon as having an exonic effect, regarding all other deletions as purely intronic events unlikely to change cell fitness. To estimate the expected rate in the absence of selection at each FS, I considered the specific distribution of exon placement, deletion size, and deletion position—aiming to roughly account for the bell-shaped concentration patterns shown in Figure 3.16. Within each FS region, I binned the deletion sizes on a  $\log_{10}$  scale divided every 0.25 units, and found the median event size within each bin. For that particular size, I simulated  $\sim 500,000$  deletions within the FS window, centred in accordance to a lowess-smoothed empirical distribution function capturing the observed mid positions of all deletions in the locus. Finally, the overall expected proportion of exonic-vs-intronic deletions was taken as the sum of each simulated fraction weighted by the proportion of deletions in that size bin.

As shown in Figure 3.15C, most FS deletions are purely intronic, and never exceed the expected rate of exon-disruption by any notable margin (granted that this cursory analysis did not extend to a formal statistical test). The variation across FS loci is almost entirely due to exon placement within the gene. For example, deletions within *PTPRD* are almost exclusively intronic because the exons are concentrated in the shoulder region with a much lower deletion rate. The only two loci with a noticeable departure from the estimated background rate are *LRP1B* and *DIAPH2* with slightly less exon disruption than expected.

The absence of protein-disrupting enrichment supports the view that FS deletions are mostly passenger events with recurrence stemming from inherent fragility, and are not under strong positive selection for their possible phenotypic effects.



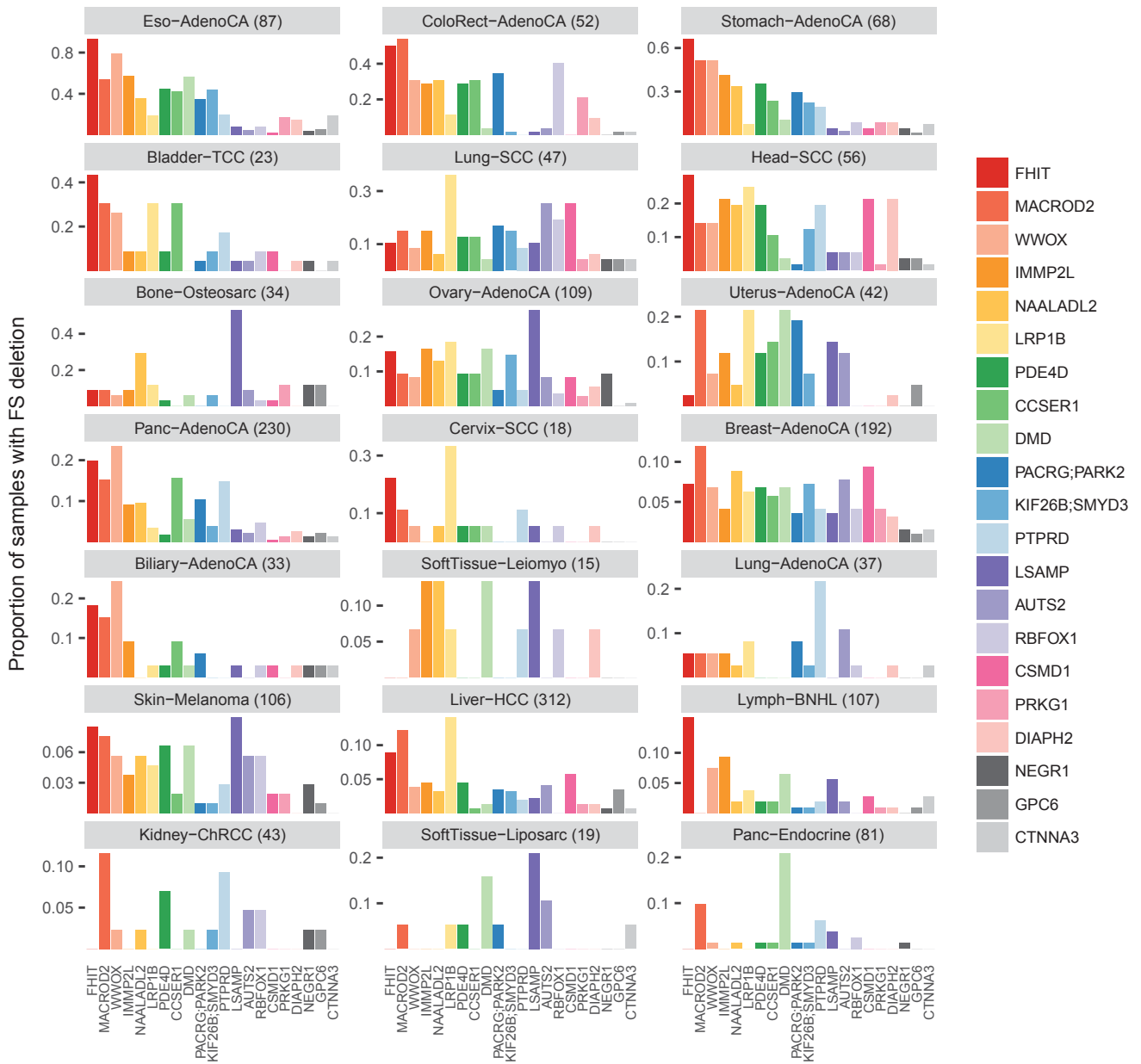


Figure 3.17: Fragile site preference by cancer histology group as measured by the proportion of samples with a deletion in each of the 21 fragile sites considered here. The number of samples is indicated in parentheses.

### 3.4.4 Tissue specificity of fragile sites

This pan-cancer dataset also offers a rare opportunity to compare fragile site activity across many different tissues. In Figure 3.17, I compare the proportion of samples with deletion in each FS region across histology groups.

Gastrointestinal cancers are the most affected by FS deletion, with esophageal, colorectal, and stomach cancers all commonly expressing fragility in *FHIT*,

*MACROD2*, *WVOX*, *IMMP2L*, *NAALADL2*, and others. There are some tissue-specific differences even within this group, with *DMD* deletions in 56% of esophagus samples but only 4% and 10% of colorectal and stomach, and *RBFOX1* deletions in 40% of colorectal samples but only 8% and 9% of esophagus and stomach.

*LSAMP* is the dominant FS in osteosarcoma (53%), ovarian adenocarcinoma (28%), and liposarcoma (21%). For squamous cell carcinomas, *LRP1B* is the dominant FS in the lung (36%), cervix (33%), and—to a lesser extent—head (25%). Other unusual tissues where one fragile site is affected more than the others are lung adenocarcinoma with 22% of samples having a *PTPRD* deletion, and pancreatic endocrine cancer with 21% of samples having a *DMD* deletion.

The cell type differences in FS fragility may be partly explained by different transcriptional programs (Wilson et al., 2015), replication timing variance (Letessier et al., 2011), and other unknown factors including oncogene-specific effects described by Miron et al. (2015). Aside from the site-specificity, the overall differences in FS deletion frequency likely reflect the incidence of replication stress triggers, with gastrointestinal cancers particularly vulnerable.

### 3.4.5 Complex SV in fragile sites

The extent of fragile site deletion is slightly underestimated due to misclassification of deletion clusters as complex events. It is common to have many deletions at the same FS within one sample, and in some samples where they overlap too much (both with each other and with different BPJ), the SV classification method groups the FS deletions into one complex unexplained cluster. In total, 83 complex clusters have at least half their BPJ within a fragile site, as summarised in Figure 3.18A. Some of these events are genuine FS deletion clusters (for example, Figure 3.18B–D at the *MACROD2* gene in esophageal and colorectal cancers), while others are different SV events. Figure 3.18E shows a complex cluster of mostly inversion-orientation BPJ at the *DMD* FS gene causing amplification of the promoter region in a pancreatic adenocarcinoma. Figure 3.18F shows a complex cluster with all BPJ orientations within the *PARK2* FS gene, causing a copy loss pattern reminiscent of chromothripsis, but unusually restricted to a ~100 kb region.

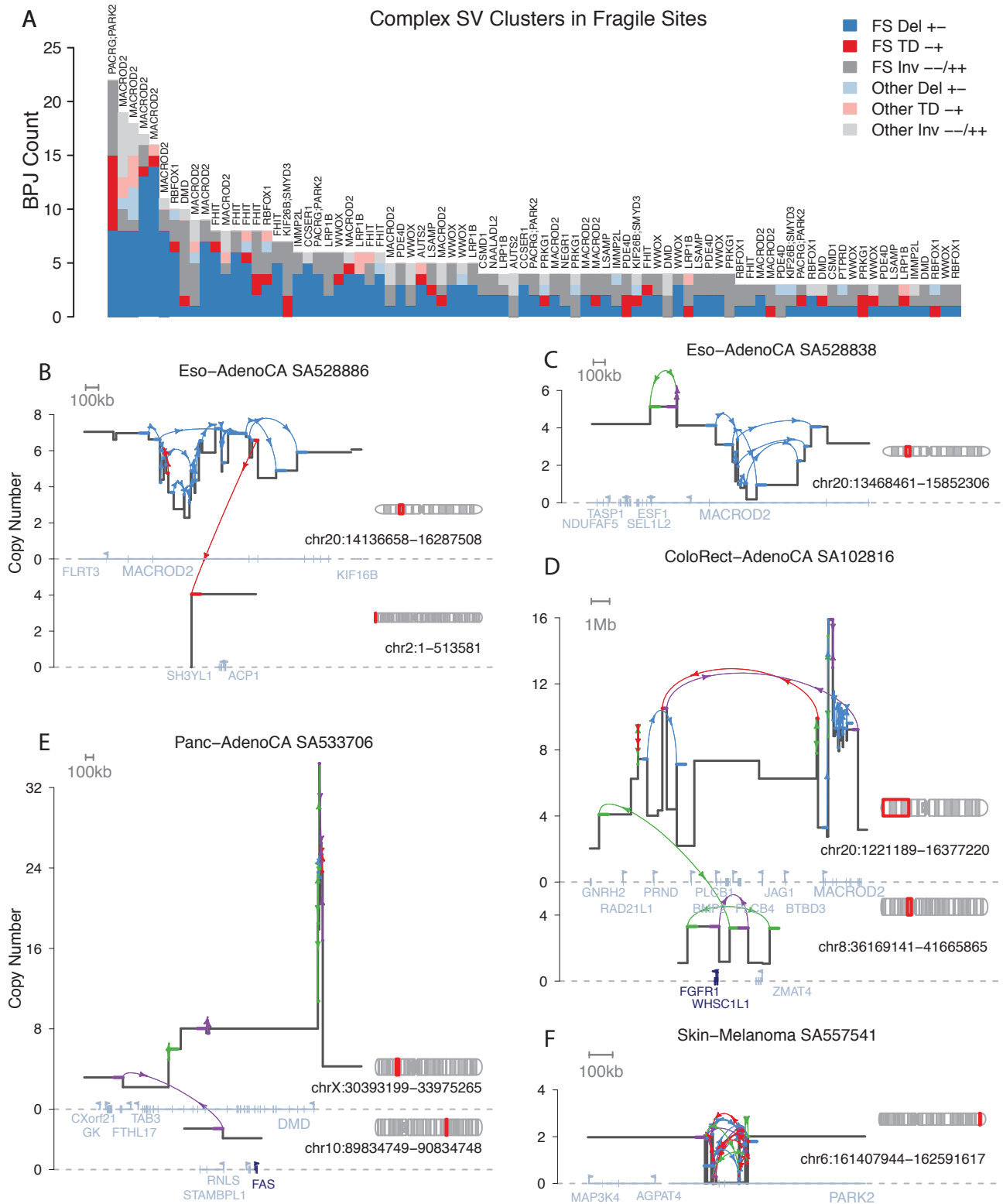


Figure 3.18: (a) 83 complex SV clusters have  $\geq 50\%$  BPJ within a fragile site; (b-f) five examples of complex clusters overlapping fragile sites.

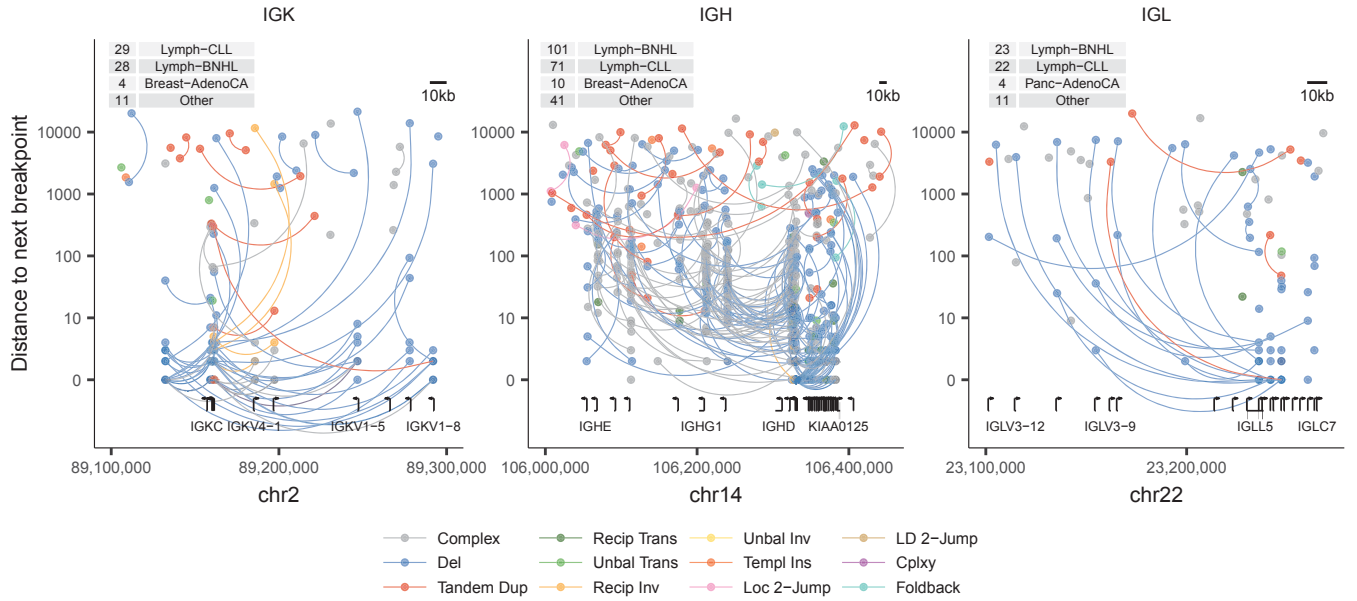


Figure 3.19: All SV breakpoint positions in three immunoglobulin loci: *IGK*, *IGH*, and *IGL*. If the two sides of a BPJ are contained within the plotting window, they are joined with a curved line.

### 3.4.6 Other anomalous genome regions

Of the SV in other anomalous genome regions, some—mitochondrial insertions, L1 retrotranspositions, and telomere length—were excluded from this dataset and analysed separately by other PCAWG members.

The three immunoglobulin loci (Figure 3.19) are notable outliers in any somatic rearrangement catalogue involving lymphocytes, with a high rate of programmed deletion for V(D)J recombination. As a result of their enzymatic DSB generation and highly active enhancer/promoter regions, these immune loci are also prone to forming recurrent oncogenic fusion translocations with spatially proximal genes including *MYC* and *BCL2* (Roix et al., 2003).

Chromosomes X and Y are another special case, with chrY excluded from this dataset and chrX present at half the dosage in male cells. Interestingly, the male PCAWG samples have SV events on chromosome X at about 60% the rate of female samples<sup>a</sup>, closer than the approximately 50% expected by CN difference alone. A likely explanation is that heterochromatin inactivation of one female X copy goes some way to protecting it from rearrangements biased towards active, open chromatin (Section 3.2).

<sup>a</sup>Considering the average number of separate SV events (clusters) on chrX in male or female samples, pooling only those histologies with at least a 30:70 gender balance (either way) to reduce cancer type confounding.

### 3.5 Structural variation affecting cancer genes

As shown in Figure 3.1, recurrent SV loci are usually explained by the presence of inherently breakable fragile sites, or cancer genes under positive selection for disruption (at tumour suppressors) or up-regulation (at oncogenes). Attempts to quantify the selection pressures conferred by rearrangement and discover novel cancer SV drivers are beyond the scope of this thesis (although can be found in a companion paper by Wala et al. (2017a) for the same dataset). In lieu of a formal SV driver analysis, I present a brief overview of different SV class patterns around several canonical cancer genes. To guide this exploration, Table 3.2 ranks COSMIC census cancer genes by the event density of various SV classes.

Table 3.2: COSMIC cancer census genes ranked by number of samples with a classified SV breakpoint in the region (gene plus 70 kb flanks), normalised by the region length and requiring at least five samples with the classification.

Gene	All SV	Complex	Del	Tandem Dup	Recip Trans	Unbal Trans	Recip Inv	Unbal Inv	Templ Ins	Cplx
CDKN2A	1	-	1	-	-	-	1	-	-	-
TMPRSS2	2	3	3	-	-	-	-	-	-	1
PTEN	3	-	2	-	-	-	-	-	-	-
MYC	4	8	-	5	1	4	-	-	-	-
CCND1	5	1	-	-	-	-	-	-	-	-
TERT	6	4	-	-	-	1	-	-	1	-
ERBB2	7	2	-	-	-	-	-	-	-	-
TP53	8	-	7	-	-	-	-	-	-	-
RARA	9	9	-	-	-	6	-	-	-	-
CDK12	-	5	-	-	-	-	-	-	-	-
CCNE1	-	6	-	-	-	-	-	-	-	-
CDK4	-	7	-	-	-	-	-	-	-	-
FHIT	-	-	4	-	-	-	-	-	-	-
SMAD4	-	-	5	-	-	-	-	-	-	-
BRD4	-	-	6	-	-	-	-	-	-	-
RB1	-	-	8	-	3	-	-	-	5	-
CDKN2C	-	-	9	-	-	-	-	-	-	-
KIAA1549	-	-	-	1	-	-	-	-	-	-
BRAF	-	-	-	2	-	-	-	-	-	-

Continued on next page

**Table 3.2 – continued from previous page**

Gene	All SV	Complex	Del	Tandem Dup	Recip Trans	Unbal Trans	Recip Inv	Unbal Inv	Templ Ins	Cplx
FGFR3	-	-	-	3	-	-	-	-	-	-
MUC1	-	-	-	4	-	-	-	-	-	-
H3F3B	-	-	-	6	-	-	-	-	-	-
CALR	-	-	-	7	-	-	-	-	-	-
STK11	-	-	-	8	-	-	-	-	-	-
LMNA	-	-	-	9	-	-	-	-	-	-
BCL2	-	-	-	-	2	-	-	-	-	-
RUNX1	-	-	-	-	4	-	-	-	-	-
ELK4	-	-	-	-	-	2	-	-	-	-
PCSK7	-	-	-	-	-	3	-	-	-	-
SLC45A3	-	-	-	-	-	5	-	-	-	-
CNTRL	-	-	-	-	-	7	-	-	-	-
CRTC3	-	-	-	-	-	8	-	-	-	-
SETD2	-	-	-	-	-	9	-	-	-	-
NCOA4	-	-	-	-	-	-	-	1	-	-
ERBB3	-	-	-	-	-	-	-	-	2	-
MPL	-	-	-	-	-	-	-	-	3	-
ACSL3	-	-	-	-	-	-	-	-	4	-
TCF12	-	-	-	-	-	-	-	-	6	-
KMT2C	-	-	-	-	-	-	-	-	7	-
RAD51B	-	-	-	-	-	-	-	-	8	-
CAMTA1	-	-	-	-	-	-	-	-	9	-
ERG	-	-	-	-	-	-	-	-	-	2

### 3.5.1 Cancer genes are affected by different SV classes

Figure 3.20 shows all SV breakpoints in the PCAWG cohort around eight example cancer genes with different rearrangement profiles.

Some tumour suppressors—like *CDKN2A* and *SMAD4*—are mostly lost through simple deletion. Others—like *PTEN* and *TP53*—are commonly disrupted by deletion or complex SV events. In contrast, the homologous recombination repair gene *RAD51B* is commonly disrupted by internal tandem duplications

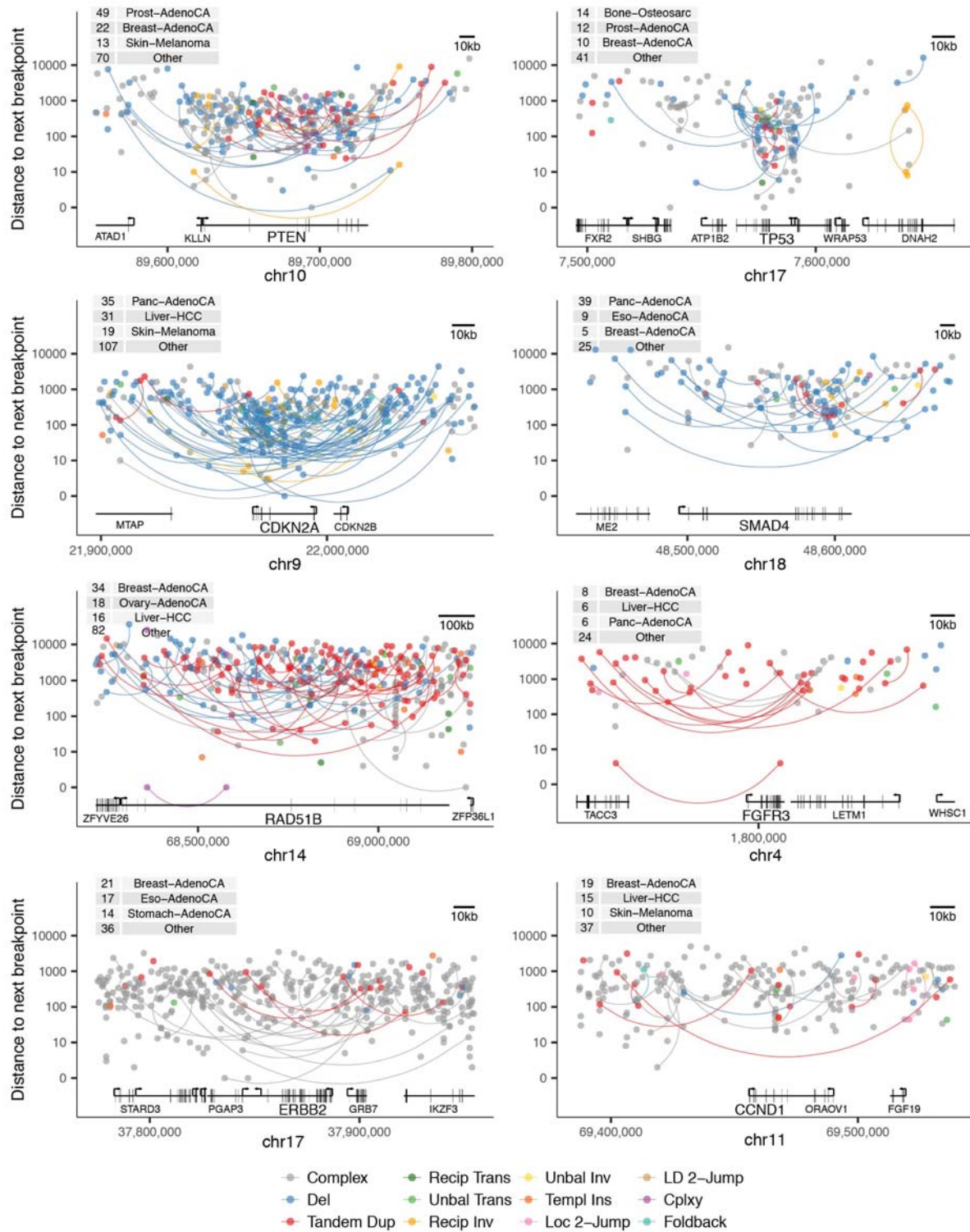


Figure 3.20: Sv breakpoint positions around known cancer genes (plus 70 kb flanks). Breakpoints are coloured by sv class, and vertically spaced by distance to the next breakpoint in the cohort. If the two sides of a BPJ are contained within the plotting window, they are joined with a curved line. The number of samples with a breakpoint in the plotting window is annotated top left.



and templated insertions, both of which are frequent events in the breast and ovary tissues observed to have *RAD51B* rearrangement. Tandem duplication within a gene causes loss-of-function by duplicating exons to disrupt the open reading frame, and are also observed within *PTEN* and *TP53*. Around an oncogene like *FGFR3*, most tandem duplications span—rather than interrupt—the transcript, and presumably up-regulate gene expression through increased dosage. Unlike *FGFR3* with its propensity for simple local duplication, other oncogenes like *ERBB2* and *CCND1* are the focus of complex SV clusters forming local amplicon structures (not shown).

### 3.5.2 Fusion drivers are formed by different SV classes

Figure 3.21 illustrates six genes involved in recurrent fusion events, with several breakpoints of the same SV class and cancer type stacking in a tightly defined cluster (usually between particular exons).

In pilocytic astrocytoma, the *KIAA1549-BRAF* driver is caused by a distinctive tandem duplication event spanning 1.9 Mb. In lymphoma, most recurrent fusion drivers are formed via translocation with an immunoglobulin locus, activating oncogenes such as *MYC* and *BCL2*. Another example of a recurrent translocation fusion is the ‘*RUNX1* translocation partner’ gene (*RUNX1T1*) frequently fused with *RUNX1* in acute myeloid leukaemia. Other SV classes generating fusion drivers include reciprocal inversion at the *RET* gene in thyroid cancer, and deletion and chromoplexy at the *TMPRSS2* gene in prostate cancer.

Figure 3.22 illustrates breakpoints in the prostate fusion partners *TMPRSS2* and *ERG*. Approximately 40% of these fusions arise through simple deletion events spanning almost 3 Mb, with the remainder resulting from chromoplexy type events involving reciprocal exchange across multiple loci. Using the stringent definition outlined in Section 2.1.3, eight (out of 199) prostate cancer samples have a clear chromoplexy-mediated *TMPRSS2-ERG* fusion. However, a further 49 samples have a complex unexplained cluster intersecting both genes, with manual inspection revealing the vast majority to be chromoplexy-type events with a complex character that is currently inaccessible to our automated classification algorithm. Other prostate fusion events involving different *ETS* family transcription factors are also mediated by chromoplexy-type events, mostly involving convoluted BPJ structures consigned to the complex unexplained bin (not shown).



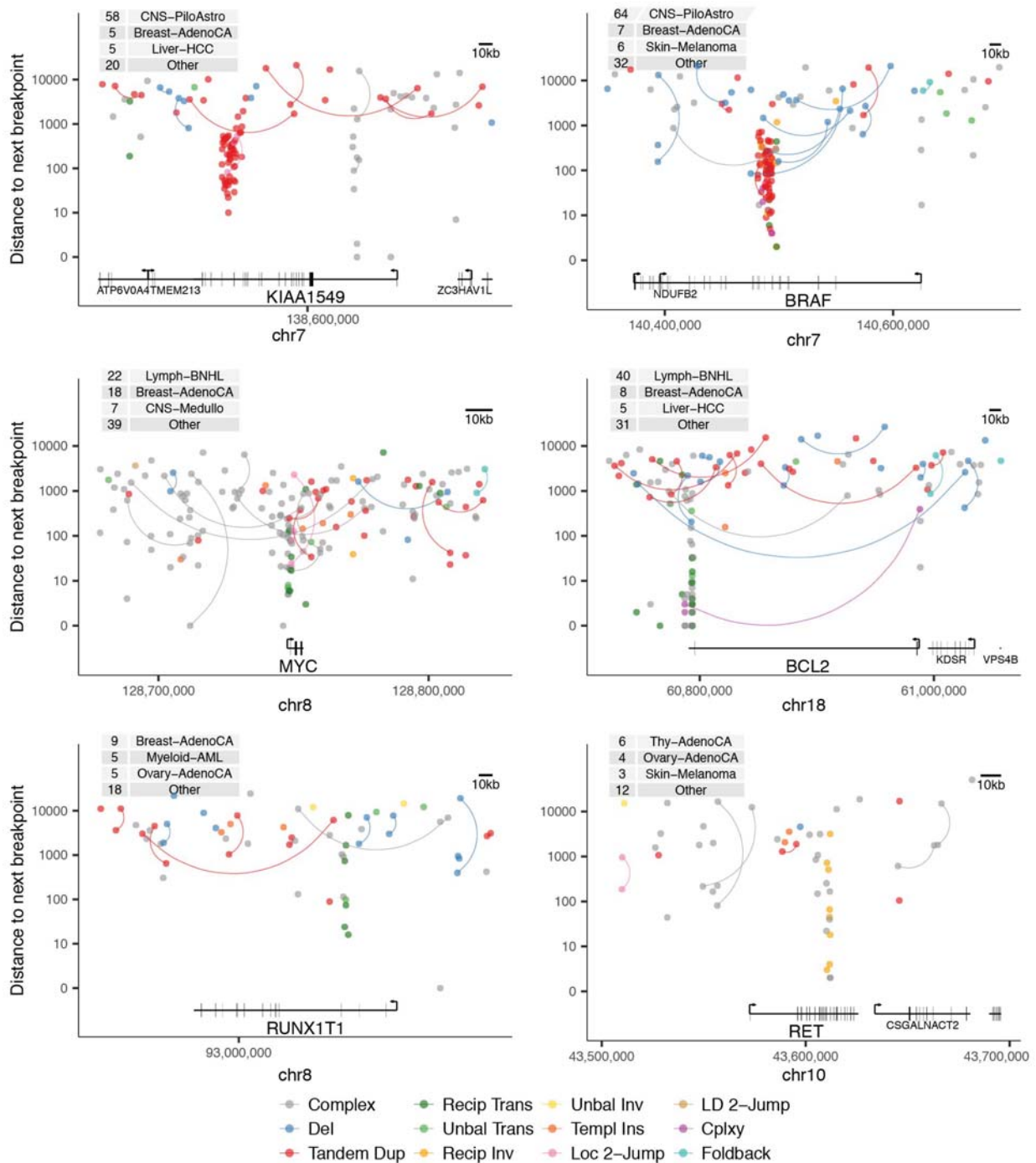


Figure 3.21: Sv breakpoints around six genes (plus 70 kb flanks) with recurrent fusion drivers. The *KIAA1549* and *BRAF* plots illustrate two sides of the same fusion event in pilocytic astrocytoma.

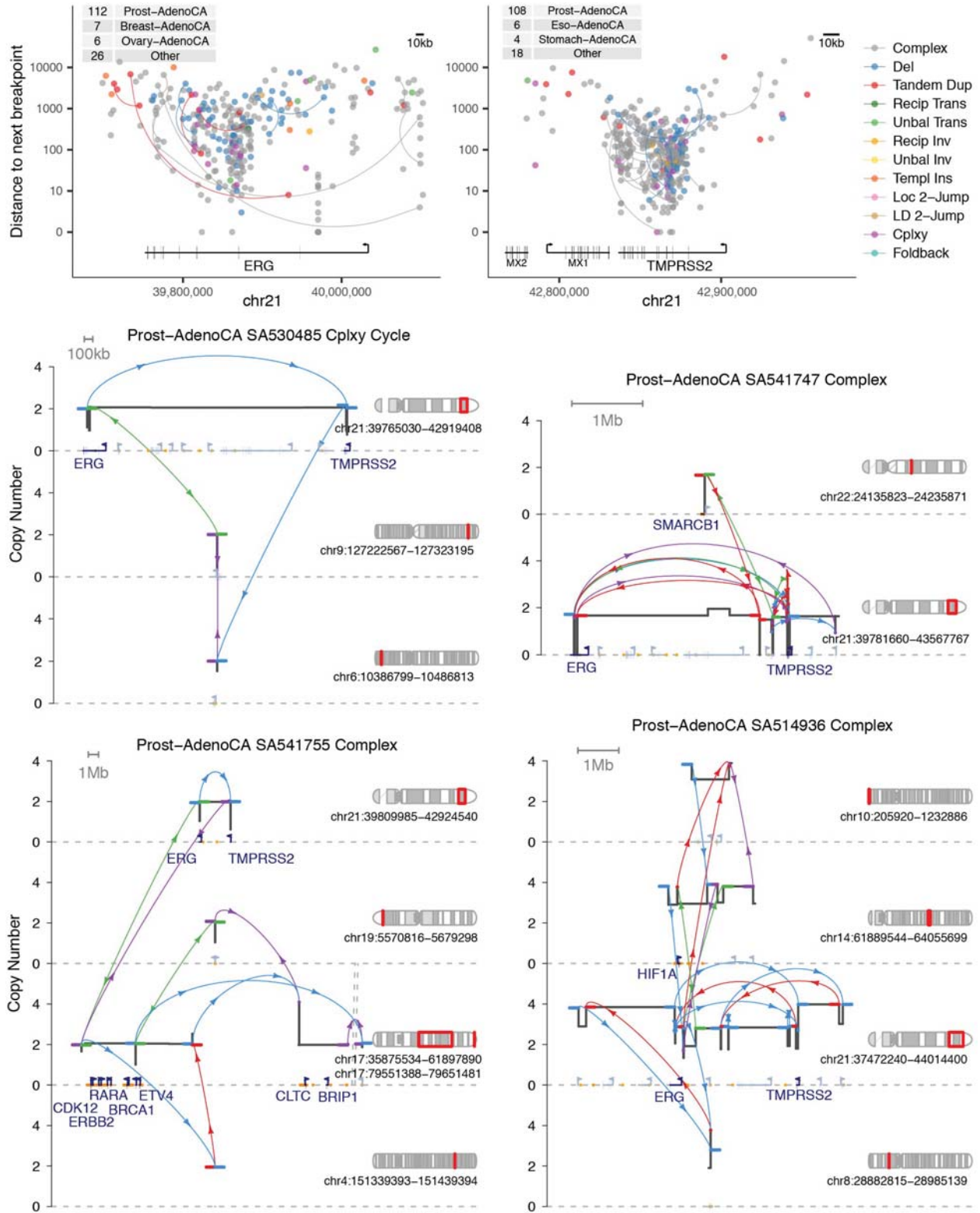


Figure 3.22: Sv breakpoints around *TMPRSS2* and *ERG*, plus four example fusion events in prostate cancer: one simple chromoplexy cycle, and three complex clusters with chromoplexy features. Annotations mark: known cancer census genes in navy; other protein-coding genes in light grey-blue (without names); and enhancer sites in orange.

### 3.5.3 Rearrangement structures around MYC

As previously indicated in Figure 3.21, the *MYC* oncogene is rearranged through many different SV forms, including translocation, tandem duplication, templated insertion, and a range of complex structures. To further illustrate this variety, Figure 3.23 shows ten SV examples affecting *MYC* in different cancer types—a small subset of the total.

Although the canonical chr8;chr14 translocation generating the *IGH-MYC* lymphoma fusion is typically a simple event, the pattern in some samples is more complex. For example, sample SA321030 has a translocation with foldback structure, and in sample SA320830 the canonical reciprocal translocation sits within a chromoplexy-type SV cluster.

In other cancer types, *MYC* is more commonly up-regulated by amplification rather than fusion. In two uterus examples (SA514439; SA460859), *MYC* is amplified through templated insertion<sup>r</sup>. The breast sample SA6128 amplifies *MYC* with a similar structure to the dup–trp–dup local 2-jump, confounded with an additional duplication-type BPJ. Another breast sample (SA77461) appears to achieve amplification via a nested series of simple tandem duplications. Three of the examples—SA411786 (pancreas), SA517281 (medulloblastoma), SA466124 (uterus)—have extremely high CN estimates indicative of double minute (DM) amplification. In the pancreas example, the outermost BPJ ( $\langle - + \rangle$  type) demarcates the circularised fragment, with other BPJs from some internal DM rearrangement. In the medulloblastoma example, the CN profile suggests a highly rearranged DM containing five distinct fragments from the same original neighbourhood. In the uterus example, the interchromosomal BPJ appear to demarcate a circularised DM formed from two distant fragments, again spanning some internal rearrangement. Finally, the *MYC* amplification in the ovary sample SA505563 is not obviously consistent with either a DM structure (expect a discrete and extreme CN profile) or with the successive overlap of simple SV structures (expect graduated CN and few BPJ). Instead, I conjecture that the complex sequence of low to mid level copy gains is indicative of a chromoanasythesis mechanism with multiple MMBIR template switches.

---

<sup>r</sup>Figure 3.23 shows one simple classified templated insertion example, and one complex unexplained cluster with features approximating templated insertion.

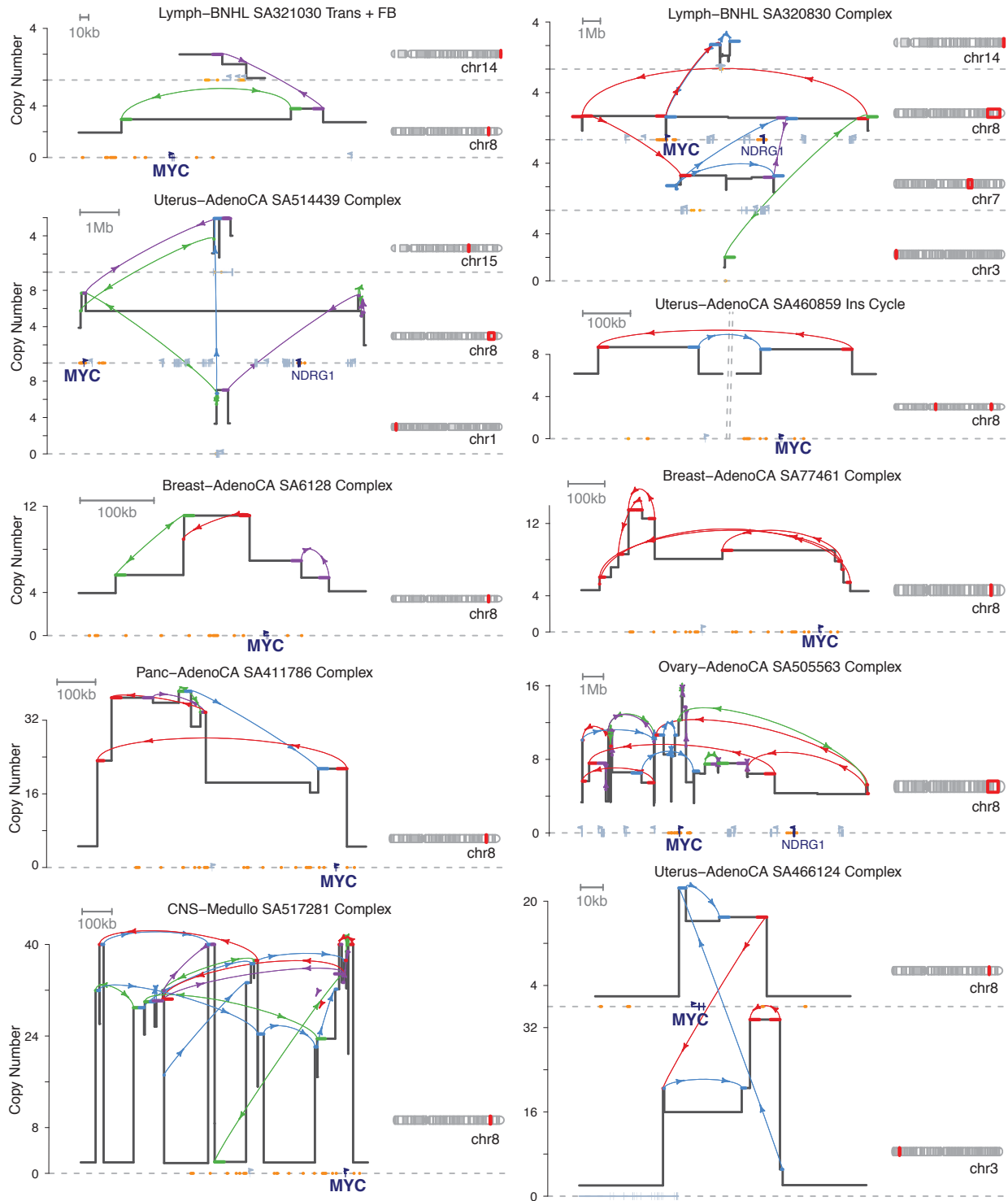


Figure 3.23: Example SV events around the *MYC* oncogene, with annotations to mark: known cancer census genes in navy; other protein-coding genes in light grey-blue (without names); and enhancer sites in orange.

### 3.5.4 Templated insertion effects

To highlight the importance of templated insertion, Figures 3.24 and 3.25 illustrate how this novel SV event can activate oncogenes (*TERT*) and disrupt tumour suppressors (*RB1*).

*TERT* encodes the catalytic subunit of telomerase, and is over-expressed in most cancers to preserve telomere length over many cell divisions (Bell et al., 2016). In addition to common promoter SNV drivers, *TERT* can also be up-regulated by enhancer-hijacking genome rearrangements (Davis et al., 2014; Peifer et al., 2015; Alaei-Mahabadi et al., 2016; Fujimoto et al., 2016; Weischenfeldt et al., 2017; Barthel et al., 2017). This observation is confirmed once again in the PCAWG cohort, with 64 samples having a SV breakpoint within 20 kb upstream (or 500 bp downstream) of the TSS (Figure 3.24A). Templated insertion is a frequent contributor to the *TERT* SV profile, with ten events in the liver cohort (of 312 samples) and four in other cancers (biliary, medulloblastoma, head squamous cell)<sup>s</sup>. Considering the 100 liver cancers with available RNA data, both templated insertion and other SV correlate with high *TERT* expression (Figure 3.24B)<sup>t</sup>. The second highest *TERT* RPKM in a liver cancer is observed in SA270088 with a three-BPJ insertion cycle shown in Figure 3.24C. Templated insertion cycles may up-regulate an oncogene by both increasing gene dosage and introducing the gene to new regulatory elements.

*RB1* encodes an inhibitor of cell cycle progression, and is inactivated in many cancers (Dyson, 2016). As shown in Figure 3.25, many SV classes intersect and disrupt *RB1*, including deletion, tandem duplication, translocation, chromoplexy, and local 2-jumps. Templated insertion also acts to disrupt this tumour suppressor, with six events observed in both breast and ovarian cancer cohorts<sup>u</sup>. Although insertion cycles could theoretically leave the *RB1* locus undisturbed (host chromosome unknown), RNA data in the breast cohort (and ovary, not shown) suggests *RB1* expression is significantly reduced in the templated insertion samples (Figure 3.25B), perhaps due to nonsense mediated decay of the rearranged open reading frame.

---

<sup>s</sup>These 14 purported templated insertion events in the *TERT* locus (gene plus 70 kb flanks) include ten classified events and four complex unexplained clusters manually curated as having strong resemblance to templated insertion.

<sup>t</sup>Despite previous reports (Totoki et al., 2014; Fujimoto et al., 2016) that the majority of liver cancers contain the canonical *TERT* promoter SNV (also expected to drive high expression in the ‘None’ SV status category in Figure 3.24B), only twelve (two with RNA) PCAWG liver samples are annotated with this mutation—a possible false negative result.

<sup>u</sup>These 12 templated insertions in *RB1* include manual curation of three complex clusters.

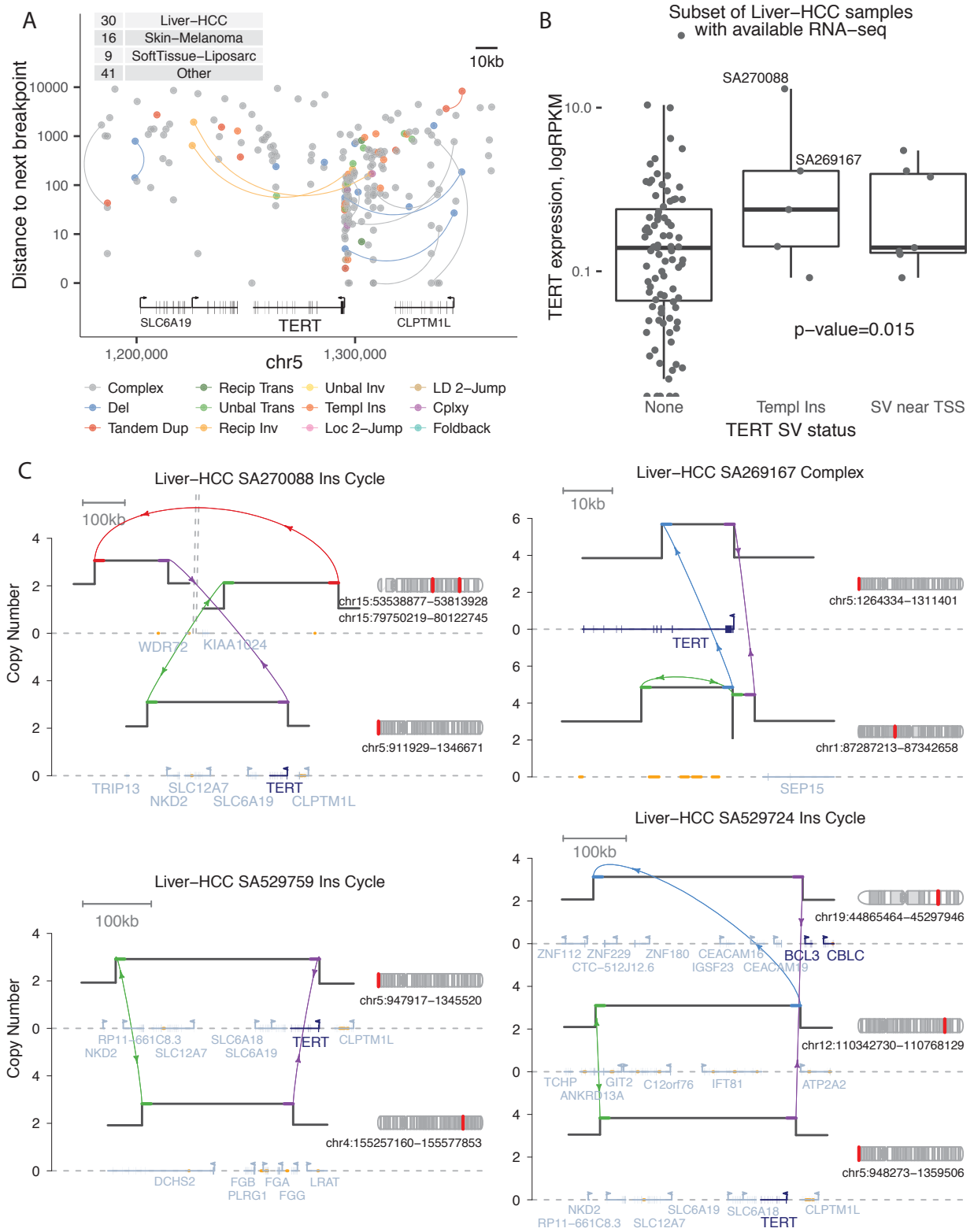


Figure 3.24: (a) Sv breakpoints around *TERT*; (b) RNA expression in 100 liver cancer samples, with the  $p$ -value from a one-sided Wilcoxon test for higher expression in samples with either a local templated insertion or another SV within 20 kb upstream of the TSS; (c) example templated insertion events in liver cancer, with annotations to mark: known cancer census genes in navy; other protein-coding genes in light grey-blue; and enhancer sites in orange.



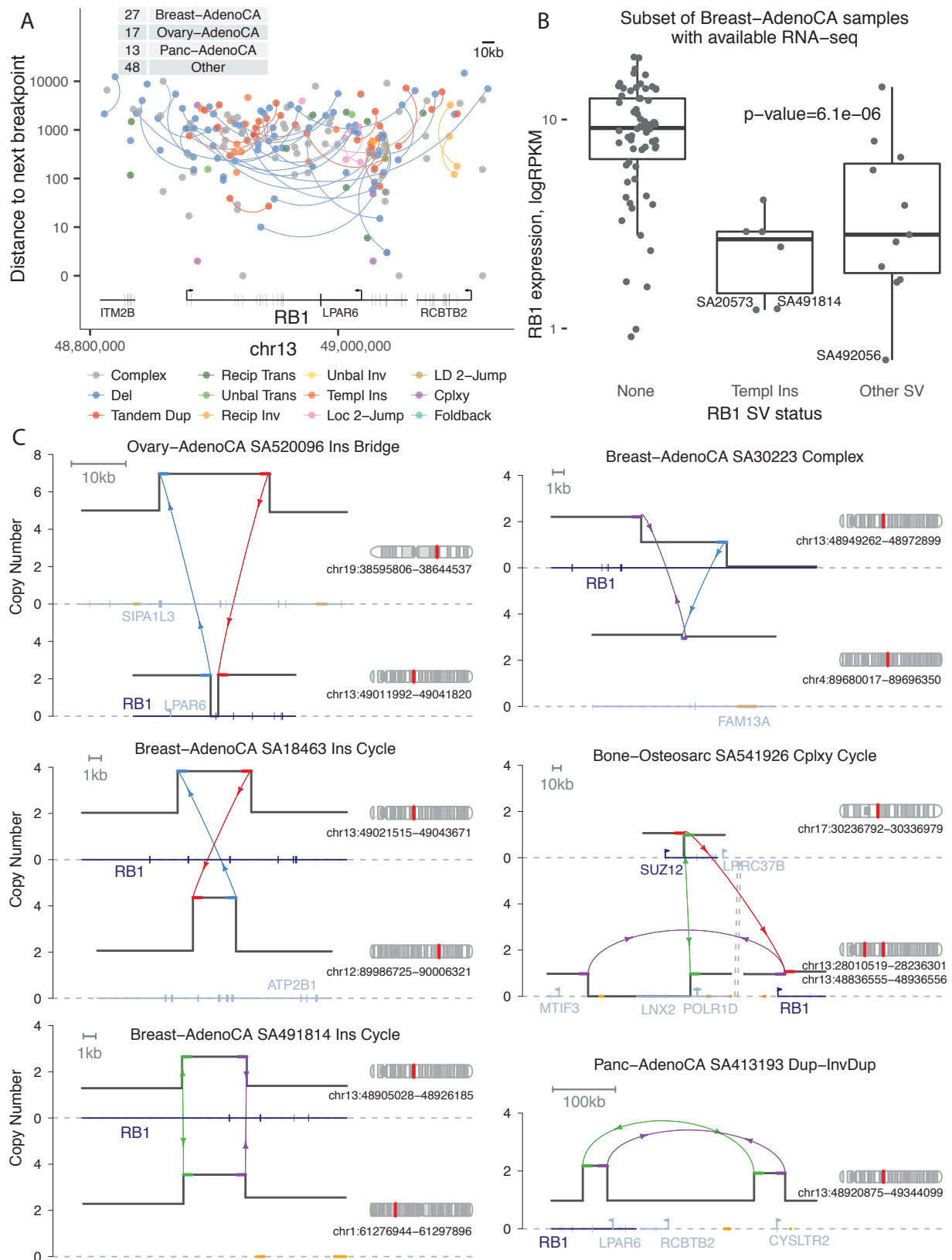


Figure 3.25: (a) Sv breakpoints around *RB1*; (b) RNA expression in 83 breast cancer samples, with the  $p$ -value from a one-sided Wilcoxon test for lower expression in samples with either a local templated insertion or another SV break in the region; (c) example events disrupting *RB1*.

## 3.6 Discussion

In this chapter, I analysed the distribution of SV classes across the genome, and attributed variance in the observed rearrangement rate to a combination of genome property correlations (Sections 3.1–3.3) and particular hotspot loci with inherent fragility (Section 3.4) or relevance to the cancer phenotype (Section 3.5).

Much of this work depends on the library of quantitative property metrics described in Section 3.1, and further improvements in accuracy and detail could be made by refining this suite of properties. For example, instead of using predicted G-quadruplex motifs based solely on sequence composition, it may be preferable to use experimentally determined G-quadruplex locations in ChIP-seq data from Hänsel-Hertsch et al. (2016). Likewise, instead of using TAD boundary estimates from just one cell line, it may be more accurate to define a consistent boundary set across multiple cell lines as reported by Akdemir et al. (2017). For the tissue-specific ROADMAP epigenome data, one major limitation was that some PCAWG cancer types—biliary, bladder, prostate, uterus—had no close cell type available, and were instead matched to a generic average over many epithelial cell lines (Table E.1). This discrepancy could already be mitigated for RNA expression, with more tissues—including prostate and uterus—now available in the GTEx atlas (GTEx Consortium, 2017). Ideally, replication timing—which is known to correlate with the plastic topology of chromatin domains (Hansen et al., 2010; Rhind and Gilbert, 2013)—should also be upgraded to a tissue-specific variable as the data becomes available. The chosen pixel size of 1 kb causes: zero-inflation of some metrics (such as distance to the nearest L1); a slightly arbitrary series of edge effects; and obfuscation of highly local effects from non-B DNA motifs. In theory, the property library is calculable for pixels of any length, with file size the only practical limitation. In future, a compact property library at single base resolution could feasibly be constructed from rounded values with run-length encoding.

Attempts to describe SV-property associations at event generation are somewhat confounded by the fact that observed rearrangements in cancer cohorts are disproportionately skewed towards recurrent events conferring positive selection. However, if we assume that: (a) most observed SVs are passenger events largely impervious to selection forces; and that (b) positively selected loci are situated in different topographical genome features; then it is reasonable to suppose that biased associations average out across the driver regions, particularly in the



heterogeneous pan-cancer setting of this study. To further reduce the influence of events under positive selection, one possible approach would be to ignore samples with a low SV burden in which each individual event is more likely to confer a relevant driver effect (for example, nearly all SV in the quiet pilocytic astrocytoma genomes are specific tandem duplications causing the driver gene fusion). Another interesting caveat is that large-scale genome rearrangements are likely to reduce the congruence between genome properties in the reference library and the derivative chromosomes present in the cancer sample. However, this only effects a subset of events occurring after major rearrangement, and any inaccurate property annotations may again be assumed to average out across a large sample size.

Overall, I found that different SV classes have different correlations with replication timing, gene density, open chromatin marks, telomere/centromere proximity, and repeat features. Within the same SV class and cancer type, some hypermutator samples have remarkably distinctive property associations, indicating that separate mutational processes may have unique effect topologies, depending on the pathways of DNA breakage and repair. For example, I hypothesise that the location of SV events following replication fork stalling will depend on the underlying cause, be it nucleotide pool depletion or collision with transcription bubbles, DNA adducts, and/or DSBs.

In somatic SNV studies, a widely adopted paradigm is for mutational processes—with differential activity across samples—to be characterised by their signature distributions of alteration class and genome topography (Alexandrov et al., 2013b; Helleday et al., 2014; Haradhvala et al., 2016; Morganella et al., 2016). Crucially, this variation across samples, mutation classes, and genome regions has important consequences for selection analysis and driver detection (Lawrence et al., 2013; Martincorena et al., 2017). In following a similar logic for structural variants, a truly comprehensive set of rearrangement rate models may need to account for the mutational process<sup>v</sup>, in addition to tissue-specific properties and two-dimensional correlations such as TAD structure and homology. However, unlike simple point mutations, about half the rearrangement burden is currently intractable to automatic classification, and therefore cannot have a sensible background rate estimation with even the simplest strategy. These difficulties pose a serious challenge for SV driver analysis, as further discussed in Chapter 6.

---

<sup>v</sup>See Chapter 4 for a signatures decomposition of somatic SV events.

