

Chapter 2

Census of the rearrangement landscape in 2500 human cancer genomes

Over the last century, the fundamental connection between cancer and chromosomal aberration has been described in increasingly forensic detail as technologies evolved from crude cytogenetic visualisation to copy number arrays and whole genome sequencing. These studies have established that the vast majority of cancer genomes carry some degree of somatic rearrangement, with massive variation in form and frequency across cancer types and samples. Efforts to provide a truly comprehensive survey of the rearrangement landscape accessible by WGS have so far been limited to simplistic classification schemes of four to eight structural categories in a few hundred samples (Yang et al., 2013; Alaei-Mahabadi et al., 2016). In an unprecedented opportunity to extend the breadth and depth of structural cancer genome analysis, the ICGC PCAWG consortium now presents a uniform callset of somatic SVs in more than 2500 samples from 30 common cancer types.

In this chapter, I describe the SV dataset assembled by various PCAWG working groups (Section 2.1), and explain the output of my colleague’s SV classification algorithm using illustrations from my own novel plotting method (Section 2.2). By identifying the precise structure of individual rearrangements and separating out complex clusters, this classification scheme allows for detailed feature exploration of all simple rearrangements. After first presenting an overall SV census (Section 2.3), I focus on the structural properties of size (Section 2.4), breakpoint homology (Section 2.5), and accompanying kataegis (Section 2.6).

2.1 Pan-Cancer Analysis of Whole Genomes structural variation dataset

The ICGC project on the Pan-Cancer Analysis of Whole Genomes (PCAWG) was a coordinated international endeavour over 2013–2017 to analyse more than 2500 matched cancer-normal samples with a uniform bioinformatics pipeline for read mapping, variant calling, and quality control (Campbell et al., 2017b). Including more than 30 common cancer types, the PCAWG dataset is by far the largest single collection of cancer whole genomes yet analysed.

2.1.1 Sample set

All matched cancer-normal samples were originally sequenced as part of tissue-specific TCGA or ICGC studies using the Illumina Hi-Seq platform to $\geq 30\times$ whole genome coverage ($\geq 25\times$ in normal) using paired-end 100–150 bp reads with insert sizes of 200–1000 bp. To ensure comparable results across cancer types, the PCAWG technical working group re-aligned all raw sequencing reads to the hg19 reference genome using BWA-MEM (Yung et al., 2017).

After extensive quality control to remove unreliable samples and, where necessary, to identify just one representative sample per donor, the PCAWG consortium agreed upon a high quality ‘white-list’ of 2583 samples (Whalley et al., 2017). Twenty-four failed to complete SV calling, and so the dataset presented in this thesis consists of 2559 samples from 37 histology groups, as tallied in Table 2.1. Six histology groups had fewer than 15 samples, and are not considered in histology-specific analyses. The largest histology classes are liver hepatocellular carcinoma (312 samples), pancreatic adenocarcinoma (230 samples) and prostate adenocarcinoma (199 samples).

For one prostate cancer donor with multiple samples (DO52513), the consortium-selected representative sample did not pass SV calling and was missing from the SV dataset. To represent this donor, I instead selected sample SA541762 because it had the highest purity as estimated by the working group on evolution and heterogeneity (Dentro et al., 2017).

Table 2.1: Sample counts by histology group in the PCAWG dataset. The geographic origin of samples is denoted by standard ICGC abbreviations. The values shown for donor age and mean sequencing coverage in the tumour (T) and normal (N) samples are the median, minimum, and maximum.

Histology	Samp	Origin	Age	T SeqCov	N SeqCov
Biliary-AdenoCA	33	SG, JP	63 (37-84)	46 (31-72)	36 (28-76)
Bladder-TCC	23	US	65 (34-84)	37 (31-60)	37 (32-45)
Bone-Benign	16	UK	unknown	44 (39-49)	32 (30-38)
Bone-Epith	10	UK	unknown	44 (42-51)	34 (28-69)
Bone-Osteosarc	34	UK	unknown	43 (39-74)	34 (29-55)
Breast-AdenoCA	192	EU,UK,US	56 (30-89)	51 (29-76)	38 (28-124)
Breast-DCIS	3	EU, UK	55 (40-61)	53 (38-54)	36 (34-36)
Breast-LobularCA	13	EU,UK,US	52 (40-76)	50 (32-84)	35 (30-39)
Cervix-AdenoCA	2	US	39 (32-46)	58 (56-59)	34 (33-34)
Cervix-SCC	18	US	39 (21-58)	58 (38-63)	35 (27-38)
CNS-GBM	38	US	59 (21-76)	41 (34-76)	40 (28-65)
CNS-Medullo	141	DE	9 (1-49)	38 (29-61)	37 (28-58)
CNS-Oligo	18	US	40 (17-62)	37 (31-68)	36 (31-57)
CNS-PiloAstro	89	DE	8 (1-50)	39 (31-51)	36 (28-54)
ColoRect-AdenoCA	52	US	68 (31-89)	47 (29-78)	35 (29-44)
Eso-AdenoCA	87	UK	70 (47-87)	67 (52-91)	40 (31-74)
Head-SCC	56	US, IN	53 (19-76)	64 (35-82)	38 (30-50)
Kidney-ChRCC	43	US	47 (17-86)	64 (54-78)	37 (30-43)
Kidney-RCC	143	US, EU	60 (38-84)	58 (29-92)	46 (23-116)
Liver-HCC	312	FR,US,JP	67 (23-89)	39 (27-126)	34 (24-108)
Lung-AdenoCA	37	US	66 (41-81)	44 (33-87)	42 (35-73)
Lung-SCC	47	US	68 (47-83)	65 (40-92)	43 (31-81)
Lymph-BNHL	107	US, DE	57 (4-85)	37 (30-77)	36 (27-58)
Lymph-CLL	90	ES	61 (40-86)	33 (24-79)	32 (25-47)
Myeloid-AML	13	UK, KR	50 (35-75)	35 (29-48)	31 (24-42)
Myeloid-MDS	2	UK	76 (74-77)	40 (40-40)	33 (32-34)
Myeloid-MPN	23	UK	54 (27-85)	44 (39-49)	34 (30-43)
Ovary-AdenoCA	109	AU, US	60 (39-81)	55 (34-78)	40 (26-77)
Panc-AdenoCA	230	AU, CA	67 (34-90)	66 (36-122)	45 (27-178)
Panc-Endocrine	81	AU, IT	59 (17-81)	66 (40-82)	41 (27-54)
Prost-AdenoCA	199	DE,CA,UK,US	59 (38-80)	62 (30-107)	41 (28-85)
Skin-Melanoma	106	AU, US	58 (16-87)	59 (33-145)	40 (21-138)
SoftTissue-Leiomyo	15	US	unknown	53 (46-60)	33 (31-37)
SoftTissue-Liposarc	19	US	unknown	54 (49-64)	33 (30-37)

Continued on next page

Table 2.1 – continued from previous page

Histology	Samp	Origin	Age	T SeqCov	N SeqCov
Stomach-AdenoCA	68	CN, US	65 (36-90)	40 (30-83)	37 (30-78)
Thy-AdenoCA	48	US	50 (17-85)	71 (32-87)	42 (30-57)
Uterus-AdenoCA	42	US	70 (38-90)	58 (35-63)	36 (26-40)

2.1.2 Calling breakpoint junctions and copy number

Consensus SV breakpoint junctions

The technical working group called SV breakpoint junctions in 2559 samples using four algorithms (Yung et al., 2017; Wala et al., 2017a). They were: BRASS from the Wellcome Trust Sanger Institute (Cancer Genome Project, 2017); DELLY from DKFZ (Rausch et al., 2012); and SvABA (Wala et al., 2017b) and dRanger (Drier et al., 2013) both from the Broad Institute.

BPJ calls consist of two genomic base locations (the breakpoint positions), each with one of two possible orientations: + for a read group leading into the break 5' to 3' on the reference strand; and – for a read group leading into the break 3' to 5' on the reference, as illustrated in Figure 2.1. SV calling algorithms also estimate the extent of possible microhomology (MH), where a run of homologous bases obscures the specific break position within the junction.

The PCAWG structural variation working group, in this task led by Joachim Weischenfeldt, defined a final consensus SV dataset after matching up estimated breakpoint positions and retaining all BPJ calls returned by two or more algorithms (autosomes and chrX only) (Wala et al., 2017a). Consensus MH was taken to be the longest estimate reported. Any BPJ attributed to somatic retrotransposition was excluded from this dataset and analysed separately by Rodriguez-Martin et al. (2017). In this thesis, I use breakpoint positions adjusted for soft-clipping evidence as described in Li et al. (2017); these adjusted positions deviate as much as 200 bp from the original consensus breakpoints.

The consensus SV call set contains 275,936 BPJ (551,872 breakpoints) in 2429 samples, with 130 samples containing no identifiable BPJ. Figure 2.2 shows the overlap between each calling algorithm in the consensus dataset, with 46% of consensus BPJ agreed upon by all four callers, and a further 34% agreed upon by three.

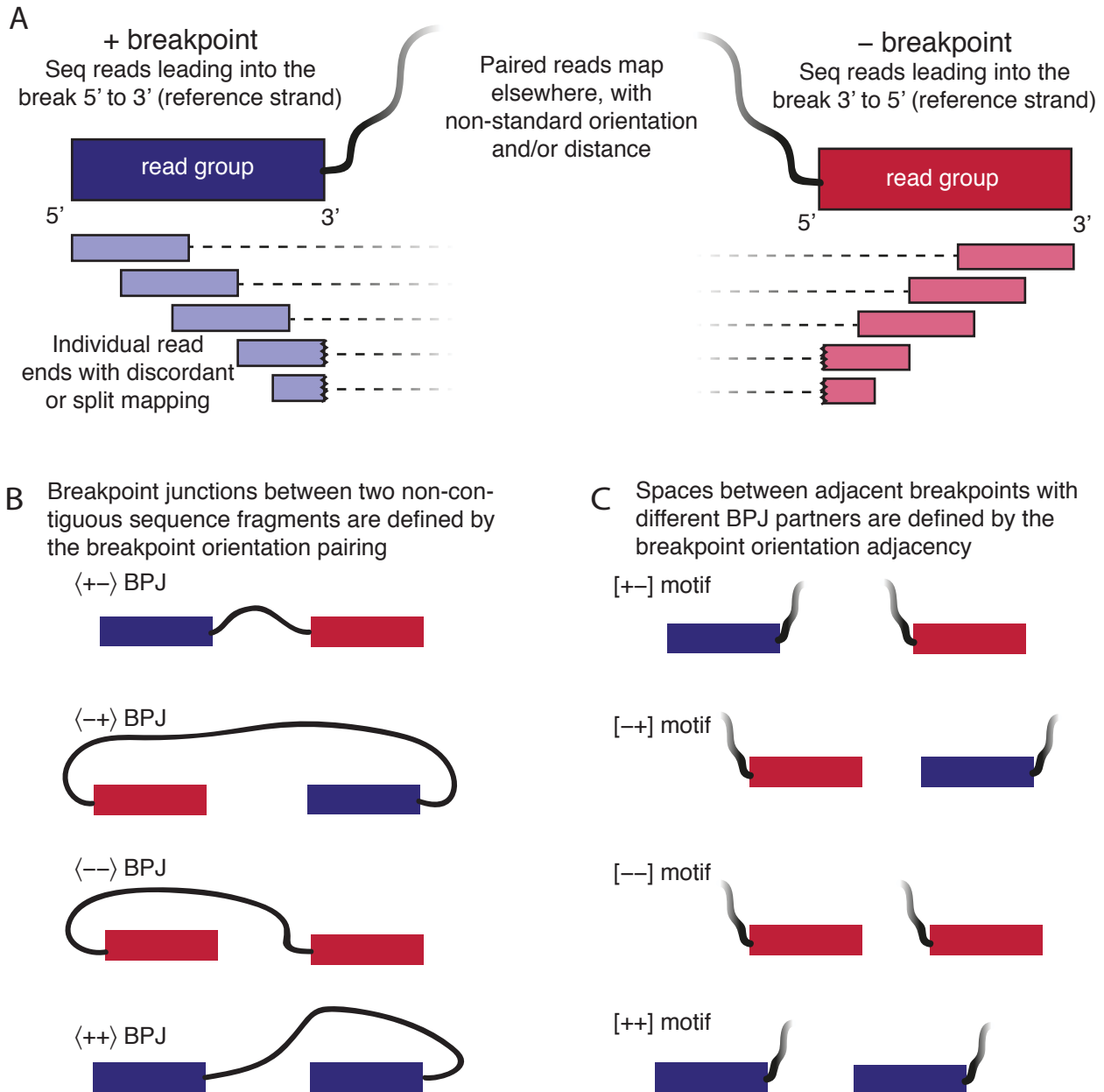


Figure 2.1: (A) Breakpoints of structural variation may have either a + or – orientation, yielding (B) four possible orientations for (intra-chromosomal) breakpoint junctions connecting two non-contiguous sequence fragments in the sv event, and (C) four possible motifs between adjacent breakpoints belonging to different junctions.

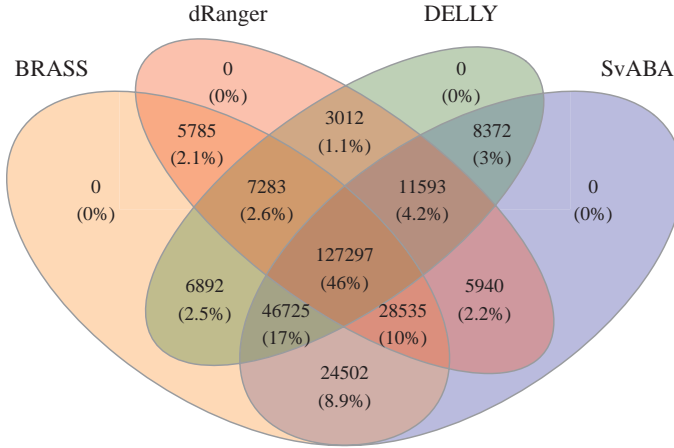


Figure 2.2: Overlap between consensus breakpoint junctions returned by the four SV calling algorithms in the PCAWG dataset.

CN segmentation

SV events are nearly always accompanied by some degree of copy number (CN) change, as even so-called “balanced” rearrangements often lose a small segment between adjacent breakpoints.

Unless stated otherwise (as in parts of Chapter 5), the CN segmentation estimates in this thesis were generated by Yilong Li with a custom algorithm described in Li et al. (2017), and henceforth referred to as YL CN calls. To briefly summarise the YL method, the tumour-normal read depth ratio in 500 bp windows was first normalised by GC content, density of fold-back read pairs, and sample purity and ploidy, and then segmented into CN estimates using known BPJ positions and additional change-points estimated with a piecewise constant regression fit. These CN estimates are non-integer, allowing for subclonal CN change and flexible fitting to noisy or complex regions, but they are occasionally unreliable over small segments and in a few problematic samples (Section 5.3).

For complex SV clusters, I sometimes switch to CN segmentation estimates provided by the evolution and heterogeneity working group, and henceforth referred to as P11 CN. Section 5.3 describes the conditions for triggering a switch to the P11 CN calls for complex SV in a particular sample. The P11 CN estimates are a consensus result from six CN calling algorithms, restricted to integer values (Dentro et al., 2017). In comparison to the non-integer YL CN estimates, these are relatively conservative and unable to capture subclonal change levels.

2.1.3 Classifying rearrangement event types

Robust methods for: (a) separating BPJ into independent clusters, and (b) classifying their structural forms; are a critical prerequisite to distinguishing the various simple and complex SV events generated by different underlying mechanisms. Without careful BPJ classification, any subsequent analysis of properties and prevalence may be strongly confounded by heterogeneous phenomena. However, meaningful classification is a difficult goal, compounded by overlapping and adjacent SV events, missing data, noisy CN estimation, lack of phasing information, tumour heterogeneity, and the germline SV background. To illustrate the problem, Figure D.1 plots intrachromosomal BPJ configurations on the *p*-arm of chromosome 17 in ten different cancer samples, each with a unique combination of rearrangements to codify appropriately.

In Chapters 2–4 of this thesis, I use SV clustering and classification provided by Yilong Li, described in detail in the supplementary methods of Li et al. (2017).

Table 2.2 summarises the SV classification scheme, employing a notation of angle brackets for an intrachromosomal breakpoint pair comprising the two halves of one breakpoint junction (e.g. $\langle + - \rangle$ for a deletion-type BPJ, all combinations illustrated in Figure 2.1B), as distinct from square brackets denoting a pair of adjacent breakpoint positions belonging to two separate BPJ (e.g. the $[+ -]$ motif indicates the left- and right-most segments lead into different BPJ with a gap in-between, all combinations illustrated in Figure 2.1C).

In brief, the BPJ clustering procedure within each sample was:

1. for every given pair of BPJ, estimate the expected number of BPJ that would be closer to either of these by chance, given the sample-specific frequency distribution of BPJ distances and types (interchromosomal, or three intrachromosomal types: $\langle + - \rangle$, $\langle - + \rangle$, and $\langle ++ \rangle / \langle -- \rangle$);
2. using the expected number of closer BPJ as a distance metric, group BPJ using agglomerative hierarchical clustering with single linkage;
3. define the first set of clusters with cut-off distance of 0.01 expected BPJ;
4. repeat steps 1 and 2 excluding the newly clustered BPJ;
5. finally, define the second round of clusters using a cut-off distance of 0.05 expected BPJ.

Following this initial clustering procedure, BPJ within each cluster were divided into local genome footprints on the assumption that distances between break

Table 2.2: Classification of simple structural variants in PCAWG cohort

SV class	Sub-group	Definition	BPJ
Complex	-	unexplained clusters	151212
Deletion	-	local $\langle + - \rangle$ BPJ	54311
Tandem Dup	-	local $\langle - + \rangle$ BPJ	45669
Recip Trans	-	distant BPJ pair, $[+ -]$ motifs	1220
Unbal Trans	-	distant BPJ	6394
Recip Inv	-	interlocked $\langle ++ \rangle / \langle -- \rangle$ BPJ pair	2800
Unbal Inv	-	$\langle ++ \rangle$ or $\langle -- \rangle$ BPJ	1995
Foldback	-	close local $\langle ++ \rangle$ or $\langle -- \rangle$ BPJ	1894
Replicative Local 2-Jump	Dup-InvDup	interlocked $\langle -- \rangle / \langle ++ \rangle$ BPJ pair	968
	Loss-InvDup	nested $\langle ++ \rangle / \langle -- \rangle$ BPJ pair	846
	Dup-Trp-Dup	disjoint $\langle -- \rangle / \langle ++ \rangle$ BPJ pair	240
Local+	Trans w/	distant BPJ adjoining $\langle ++ \rangle$ or $\langle -- \rangle$	580
Distant	Foldback	BPJ w/ $[- +]$ motif	
2-Jump	Trans w/	distant BPJ intersecting $\langle ++ \rangle$ or $\langle -- \rangle$	508
	InvIns	BPJ w/ $[- +]$ motif	
	Trans w/ TandemDup	distant BPJ pair w/ $[- +]$ motifs & unbalanced CN	176
Templated Insertion	Ins Cycle	loop of $[- +]$ motifs	3052
	Ins Bridge	loop of $[- +]$ motif/s into $[+ -]$ motif	2601
	Ins Chain	chain of $[- +]$ motif/s	616
Chromoplexy	Cplx Cycle	loop of $[+ -]$ motifs	326
	Cplx Chain	chain of $[+ -]$ motif/s	366
	Cplx Cycle w/ Ins	loop of $[+ -]$ and $[- +]$ motifs	162

Dup = duplication; Trp = triplication; Trans = translocation; Recip = reciprocal; Unbal = unbalanced; Inv = inversion; Ins = insertion; Cplx = chromoplexy

positions within a footprint should fit an exponential distribution (and that distances between footprints will be larger than this). The footprinting step and further heuristic adjustments separated out peripheral deletions or tandem duplications, and identified isolated $[- +]$ or $[+ -]$ motifs for the definition of templated insertion and chromoplexy events respectively.

Finally, clusters of one or two BPJ and clusters of isolated $[- +]$ or $[+ -]$ footprints were classified by the relative orientation of the BPJ as summarised and tallied in Table 2.2. In addition, overlaps of a few simple BPJ were separated into their constituent events by comparison against a library of all possible overlap structures and selection of the parsimonious solution.

In total, this method classified 45% of the BPJ calls, leaving the remaining 55% (151,212 BPJ) in unexplained complex clusters. Section 2.2.2 provides additional description of the different SV classes alongside visualisation of example events.

2.1.4 Additional sample information

Additional PCAWG sample information used in this thesis includes: whole genome duplication estimates from the evolution and heterogeneity group (Dentro et al., 2017); driver annotation of individual SNV and indel events from the drivers and functional impact group (Sabarinathan et al., 2017); gene expression estimates from the transcriptome group (Fonseca et al., 2017); and microsatellite instability typing from the mutational signatures group (personal communication with Akihiro Fujimoto).

2.2 Visualising structural variants

The somatic SV set in the PCAWG cohort includes a diverse range of rearrangement phenomena involving multiple genome loci in many varied combinations. WGS over these rearrangements yields two types of informative data: breakpoint junctions at base-pair resolution, and copy number segmentation estimates. Given the complexity of the underlying biology and resulting data, visualisation is absolutely paramount for understanding and communicating SV analysis.

2.2.1 A robust plotting method for structural variation

To visualise any SV structure (or group of structures) ranging from the simplest deletion to the largest chromothripsis event spanning multiple chromosomes, I developed a scalable plotting method to present CN estimates with BPJ calls. As WGS data does not afford the additional benefit of phasing information, all data is shown relative to the reference genome rather than the physical derivative chromosomes present in the sample. Without phasing information, the precise order of BPJ on the derivative chromosome cannot generally be reconstructed, nor can the possibility of independent events on different homologous chromosome copies be ruled out.

To arrange the data, I divide the plotting window into columns of variable-height rectangles (one per reference chromosome) with linear reference space on the horizontal axis and copy number on the vertical.

First, chromosomes order themselves in the grid to minimise the sum of squares of the plotting distance traversed by interchromosomal BPJ, with a double penalty for horizontally adjacent chromosomes compared to vertically adjacent.

For context, the ideogram of major Giemsa bands lies on the outer edge of each chromosome’s plotting area.

Second, I define the local genome footprints^a to plot by flanking each breakpoint by some set flank size (variable, usually many kb), leaving no gaps smaller than some minimum distance (variable, usually 10 Mb). For chromosomes with more than one constituent footprint, the horizontal plotting coordinates break into two disjoint windows if there is a gap between footprints spanning over 40% of the total window (axis break indicated by parallel dashed lines). Red highlights on each ideogram indicate the genome region/s represented.

Third, the vertical height of each chromosome’s plotting area is set to include the maximum CN estimate in the footprint region/s.

Having established the layout and scale, mapping functions convert genome positions and CN values into their equivalent plot coordinates. A step function outlines the CN segmentation in each footprint, and curved lines mark the BPJ connections, with arrows pointing away from the break for + orientation and towards the break for – orientation. The default option is to colour BPJ blue for $\langle + - \rangle$, red for $\langle - + \rangle$, purple for $\langle ++ \rangle$, and green for $\langle -- \rangle$. To further assist the visual distinction between + and – ends, the segment leading into the break is coloured to match. An alternative option is to colour BPJ by any other categorical factor, used in Chapter 5 to distinguish BPJ in separate clusters.

Finally, annotation of genes and other functional elements is an optional addition along the lower edge of each chromosome’s plotting area.

2.2.2 Visual examples of all SV classes

To supplement the SV class definitions outlined in Table 2.2, here I include some example events for illustration. For the complex SV clusters left unexplained by the current classification scheme, I refer the reader to Chapter 5.

The simplest SV classes comprise just one BPJ, as illustrated in Figure 2.3. They are: deletion, tandem duplication, foldback, unbalanced inversion, and unbalanced translocation. As foldback and unbalanced inversion are both defined by one lone $\langle ++ \rangle$ or $\langle -- \rangle$ orientation BPJ, their only distinguishing feature is the distance between breakpoints, although the specific threshold is somewhat arbitrary. Foldback refers to a highly local one-sided inversion (the sequence almost literally ‘folds back’ on itself, median distance 4 kb), whereas

^aPlotting footprints are different to the classification footprints described in Section 2.1.3

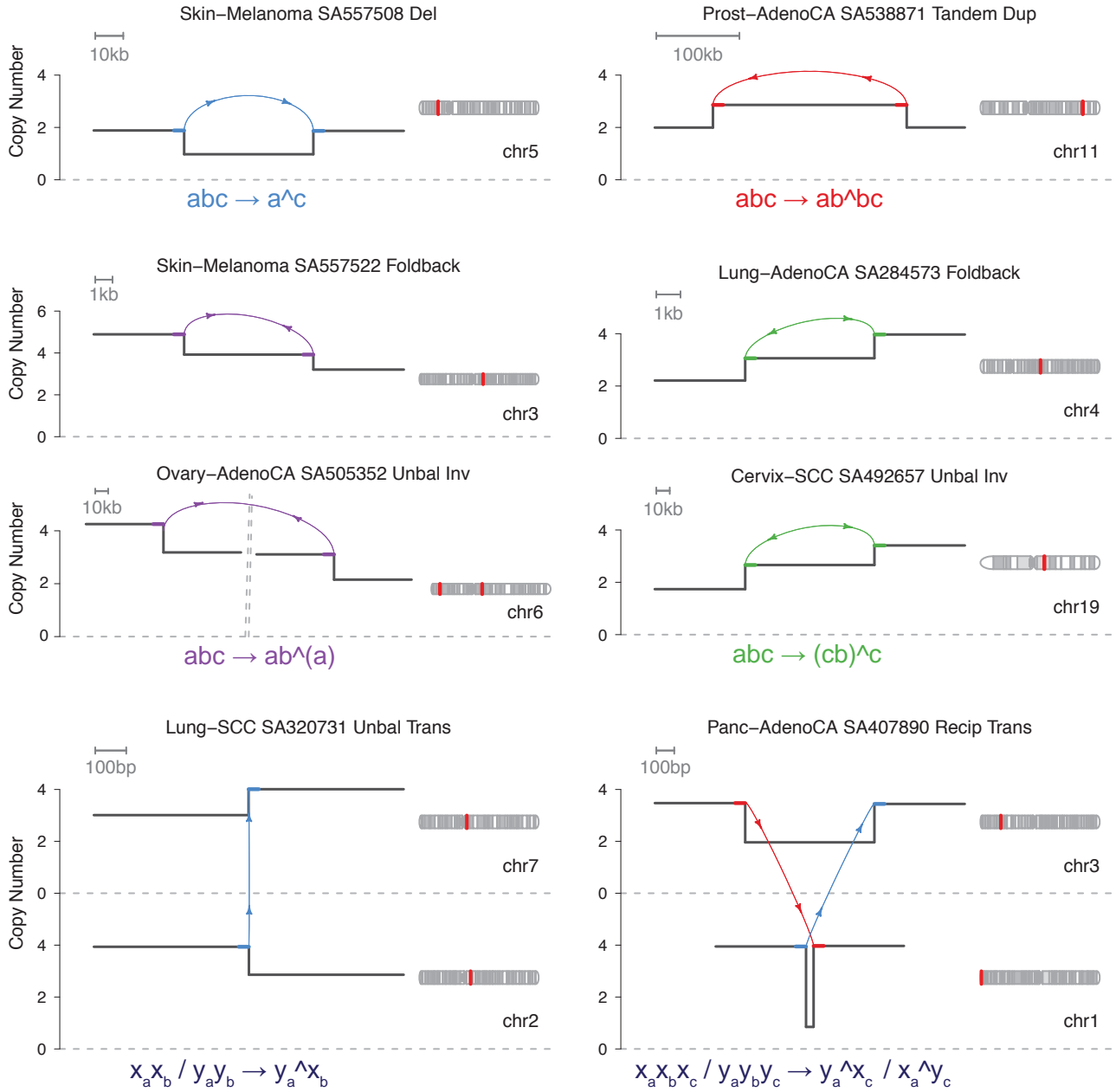


Figure 2.3: Example plots of the simple SV event classes: deletion, tandem duplication, foldback, unbalanced inversion, unbalanced translocation, and reciprocal translocation. The transformation between germline segment order and the somatic rearrangement is annotated below, with carets to denote breakpoint junctions between non-contiguous reference segments, parentheses to indicate inverted segments, and a forward-slash to separate different chromosomes.

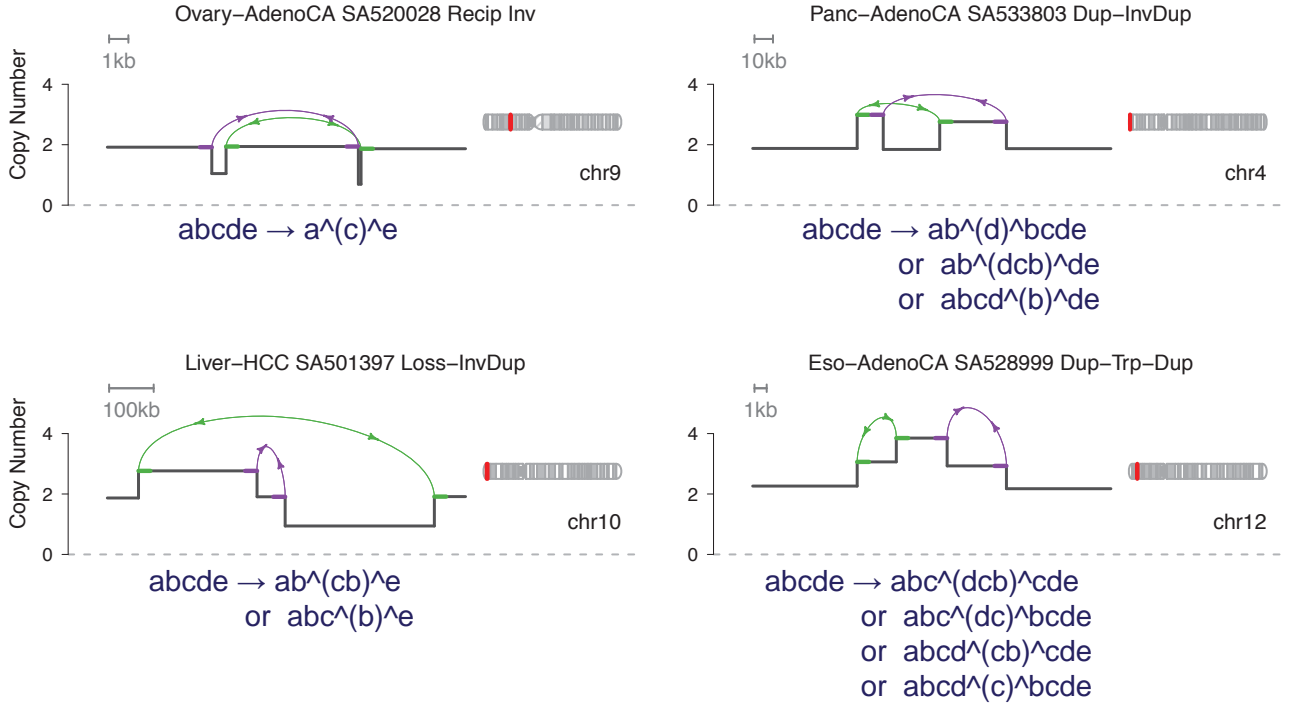


Figure 2.4: Example plots of reciprocal inversion and the three types of local 2-jump: duplication-inverted duplication, loss-inverted duplication, and duplication-inverted triplication-duplication. The transformation between germline segment order and the possible rearranged derivative structures is annotated below, with carets to denote breakpoint junctions between non-contiguous reference segments, and parentheses to indicate inverted segments.

unbalanced inversion refers to a BPJ between more distant loci (median distance 8 Mb). An intrachromosomal BPJ could, in theory, be a translocation between two homologous chromosomes. However, given the low frequency of reciprocal translocations detected on homologous chromosomes, I estimate that approximately 0.4% of single intrachromosomal BPJ might actually be unbalanced translocation^b—a negligible fraction for subsequent analyses. Figure 2.3 also includes an example of reciprocal translocation—a pair of interchromosomal BPJ with characteristic $[+-]$ motifs demarcating a small region of copy loss between breakpoints.

Figure 2.4 illustrates the SV classes involving two opposite inverting BPJ. The

^bConsidering all inter-chrom translocations, unbalanced events outnumber reciprocal at a ratio of 11:1. We detect 21 reciprocal translocations of type $\langle + - \rangle / \langle - + \rangle$ on homologous chromosomes. Assuming this is approximately half the true total ($\langle + + \rangle / \langle - - \rangle$ classified as reciprocal inversion), the total number of unbalanced translocations between homologous chromosomes might be estimated in the ballpark of $11 \times 21 \times 2 = 462$. There are 103,869 total deletions, tandem dups, foldbacks and unbalanced inversions in the cohort, so if approximately 460 are actually translocations, then this is an error rate of $\approx 0.4\%$.

reciprocal inversion has a $\langle ++ \rangle$ BPJ interlocking with a $\langle -- \rangle$ BPJ^c, leaving $[-+]$ motifs with accompanying copy number loss either side of the middle segment that now sits inverted in the derivative chromosome. The other interlocking pattern of $\langle -- \rangle$ followed by $\langle ++ \rangle^c$ forms the dup-inv-dup structure, imparting $[-+]$ motifs with accompanying copy number gain. Similar regions of local copy gain are found in the loss-inv-dup with nested inverting BPJ (either $\langle -- \rangle$ within $\langle ++ \rangle$ or $\langle ++ \rangle$ within $\langle -- \rangle$), and the dup-trp-dup structure of disjoint BPJ in the order $\langle -- \rangle$ then $\langle ++ \rangle^c$. These last three structures cannot be generated by any plausible combination of ‘break and ligate’ mechanisms^d, and thus the group name ‘local 2-jump’ refers to the purported ‘template and replicate’ mechanism with two rounds of strand invasion. Small template switch events have previously been described in germline developmental disorders (Lee et al., 2007; Carvalho et al., 2011), but this is the first analysis to formally identify them in somatic cancer genomes.

Extending the concept of local 2-jump structures, Figure 2.5 illustrates three types of local plus distant 2-jump. One structure results in an unbalanced translocation with sequence foldback close to the breakpoint on one side. Given that the distal side of the unbalanced translocation is preserved, it seems likely these events are precipitated by foldback and end in translocation. The segment of copy number gain implicates a possible role for replication-based polymerase jumping. Another structure of unbalanced translocation with a local segment inserted in inverted orientation could plausibly result from polymerase jumping as well, although the absence of copy gain means simple breakage and ligation is also a possible route. Less intuitive is the structure generated by an unbalanced translocation followed by tandem duplication spanning the break. In the bottom left example of Figure 2.5, the blue BPJ marks an initial translocation between chr11 and chr2, with a subsequent tandem duplication in red on the derivative chromosome—duplicating the segment containing the original translocation BPJ. Although the two $[-+]$ motifs match the pattern generated by templated insertion cycles shown in Figure 2.7, the unbalanced CN either side identifies this as tandem duplication after translocation. Likewise, the bottom right example in Figure 2.5 illustrates an initial translocation in green, followed by tandem duplication on the derivative chromosome in purple.

Templated insertion events come in three varieties, all characterised by $[-+]$ motifs with accompanying copy number gain indicative of replication-based SV

^cIn the order moving 5’ to 3’ along the reference strand, left to right in plotting space.

^dComparing against the library of possible overlap patterns generated by Yilong Li; more details in Li et al. (2017).

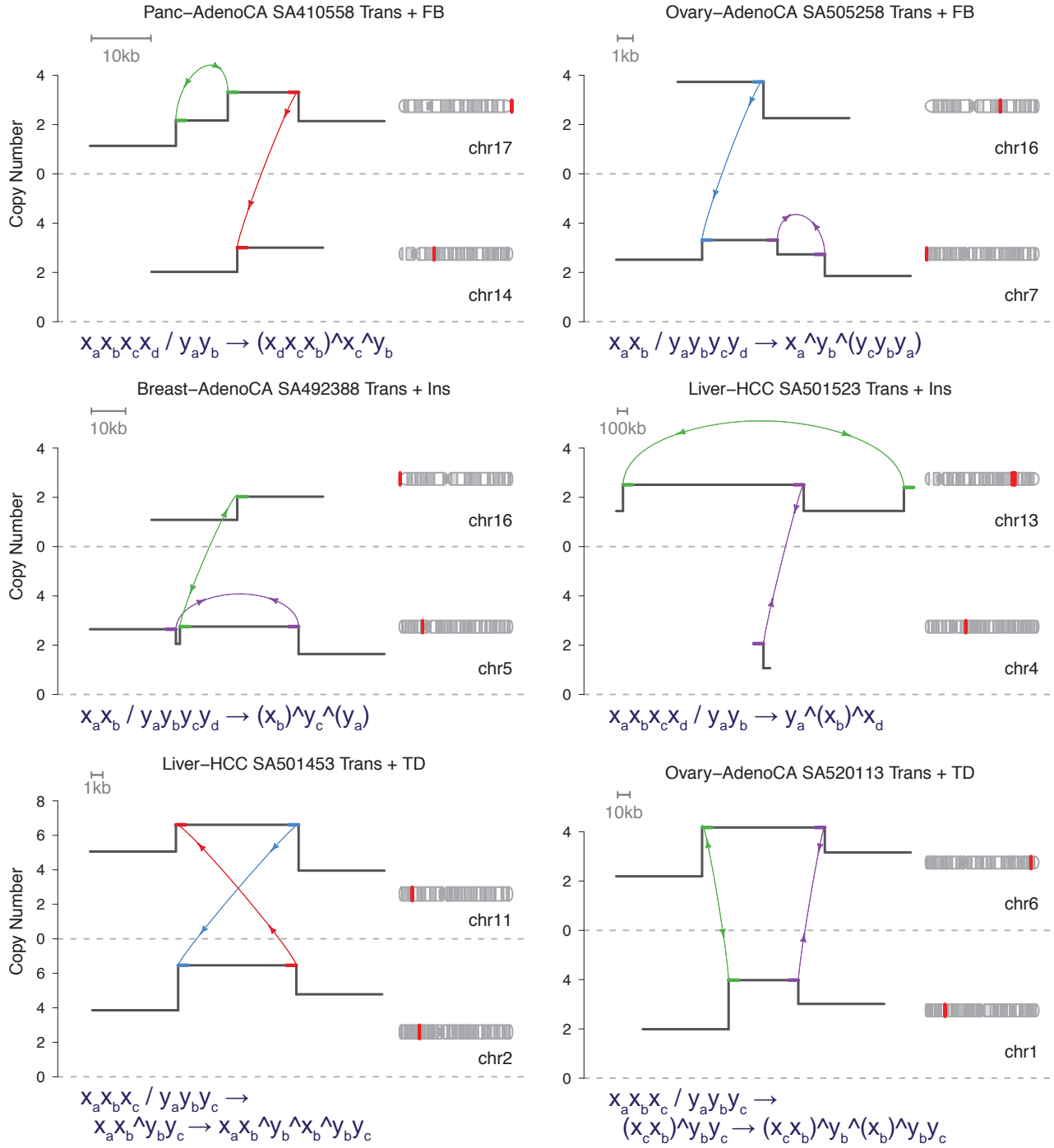


Figure 2.5: Example plots for three classes of local + distant 2-jump: translocation with foldback; translocation with inverted insertion; and translocation with overlapping tandem duplication. The transformation between germline segment order and the somatic rearrangement is annotated below, with carets to denote breakpoint junctions between non-contiguous reference segments, parentheses to indicate inverted segments, and a forward-slash to separate different chromosomes. The intermediate structure is included for translocation plus tandem duplication.

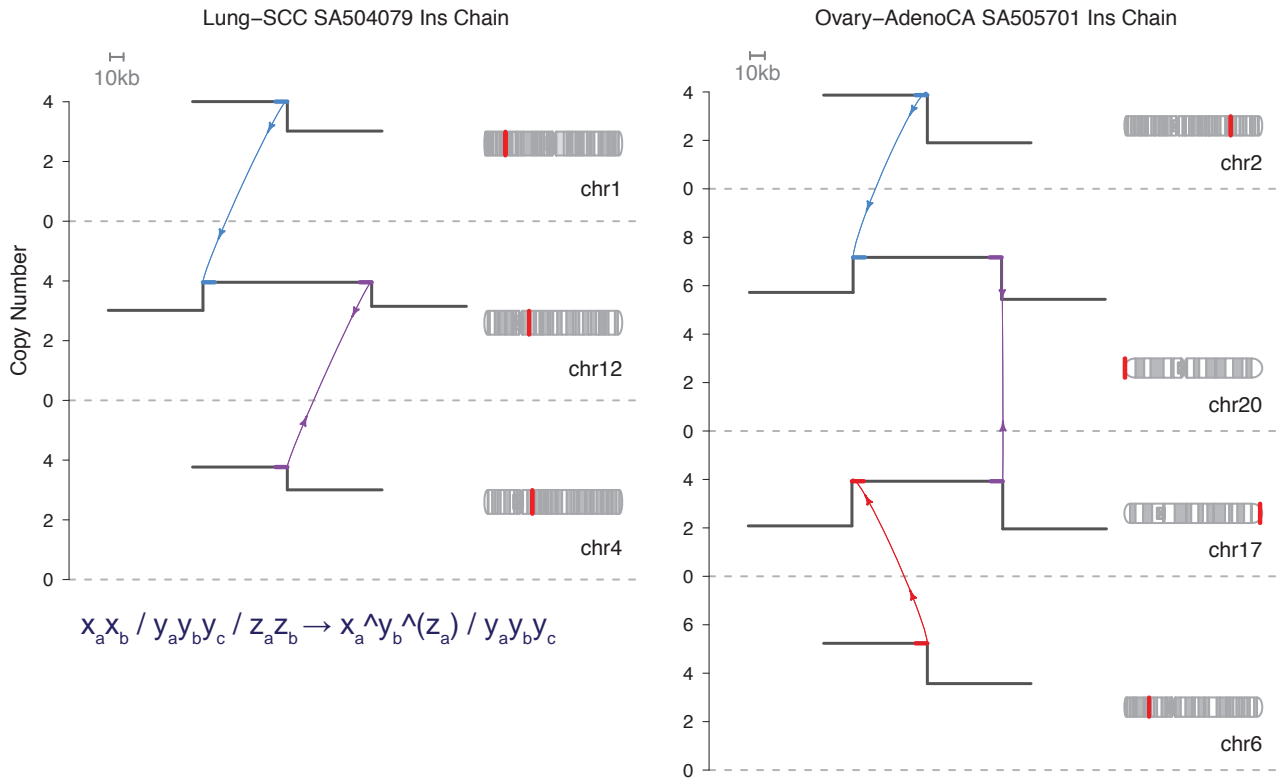


Figure 2.6: Example plots for chains of templated insertion, where two distant loci are joined by one or more templated inserts ($[-+]$ motif). The transformation between germline segment order and the somatic rearrangement is annotated for the simplest example, with carets to denote breakpoint junctions between non-contiguous reference segments, parentheses to indicate inverted segments, and a forward-slash to separate different chromosomes. Note that the original locus of the insert segment (y_b) remains intact.

formation. Insertion chains, shown in Figure 2.6, link two distant loci through a path of one or more templated inserts. The overall derivative structure is an unbalanced translocation, with a chain of distant segment/s copied into the join. Insertion bridges and cycles, shown in Figure 2.7, both loop back to the original locus. In a bridge event, the point of return is after the point of departure—leaving a deletion on the host chromosome with a chain of distant segment/s copied into the gap. In a cycle event, the point of return is behind the point of departure, thus re-replicating a segment on the host to generate a tandem duplication with a chain of distant segment/s copied in-between. The symmetry of BPJ and CN generated by templated insertion cycles means the identity of the host chromosome cannot be determined by WGS. For all templated insertion events, the original loci of the insert segments remain intact. This specific definition of templated insertion events is the first of its kind in either somatic or germline genome studies.

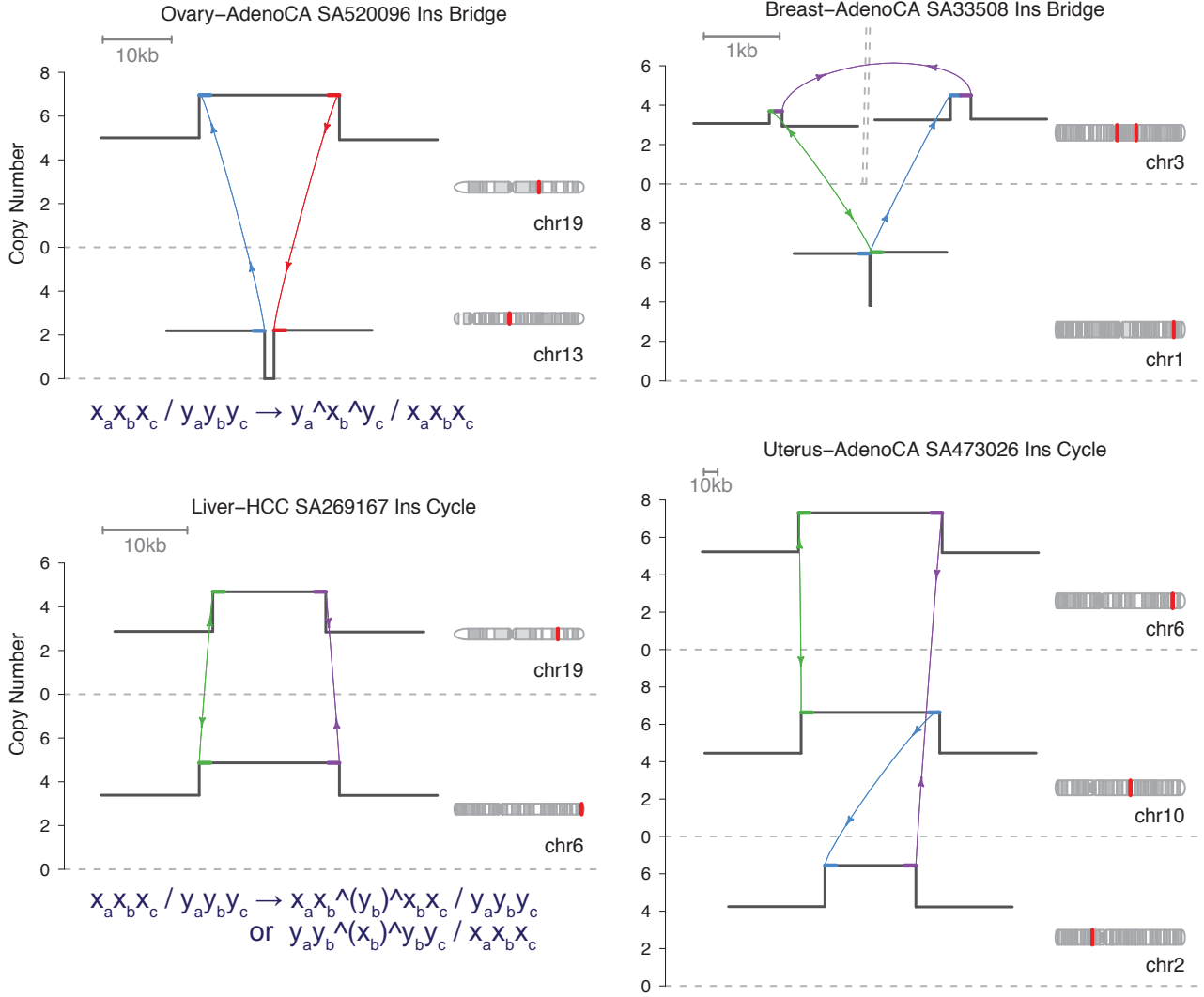


Figure 2.7: Example plots for bridges and cycles of templated insertion ($[-+]$ motif). The ‘bridge’ events insert one or more templated inserts into a gap ($[+-]$ motif) on the host chromosome. The ‘cycle’ events insert one or more templated inserts between a local duplication on the (unknown) host chromosome. The transformation between germline segment order and the somatic rearrangement is annotated for the simplest examples (detail in previous figure legends).

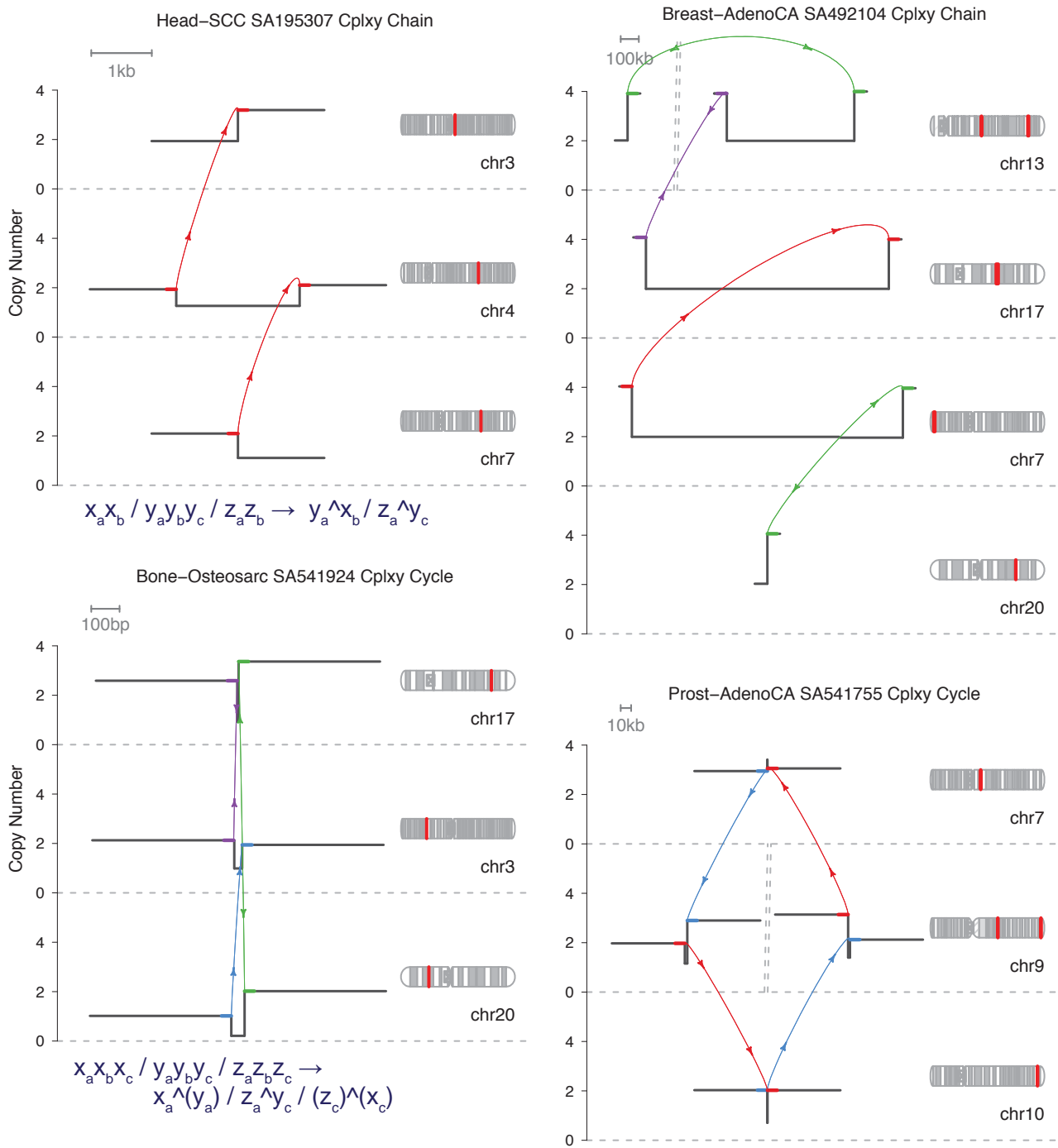


Figure 2.8: Example plots for chains and cycles of chromoplexy. The ‘chain’ events involve one or more footprints of balanced translocation ([+–] motif) that start and end in isolated breakpoints (unbalanced translocation). The ‘cycle’ events involve three or more footprints of balanced translocation ([+–] motif) in a closed loop (all derivatives are balanced). The transformation between germline segment order and the somatic rearrangement is annotated for the simplest examples (detail in previous figure legends).

Finally, Figure 2.8 shows chromoplexy events characterised by $[+ -]$ motifs with accompanying copy number loss—extending the simple balanced structure of reciprocal translocation to three or more loci. Chromoplexy chains start and end in unbalanced translocation, connected to partners in balanced translocation motifs. Chromoplexy cycles are a complete loop of three or more balanced translocation motifs, with all breakpoints finding a ligation partner within the closed set. As discussed in the supplementary methods of Li et al. (2017), repair at balanced translocation breakpoints can sometimes result in short $[- +]$ motifs instead of the canonical $[+ -]$ pattern, and these can only be distinguished from short templated insertions by the presence of reads extending through the other break position.

2.3 Initial census of SV events

The detailed classification of SV structures in 2559 PCAWG samples allows for a comprehensive census of SV prevalence across individual cancers and different histology groups.

2.3.1 SV prevalence by histology

Figure 2.9 presents an overview of all major SV class frequencies in cancer samples grouped by histology. Overall, liposarcoma has the greatest SV burden with a median of 825 BPJ per sample (IQR 549–1195), followed by ovarian adenocarcinoma and osteosarcoma with per-sample BPJ medians of 231 (IQR 157–317) and 195 (IQR 110–390) respectively. At the other extreme, myelo-proliferative neoplasms have the lowest SV burden with a median of 0 BPJ per sample (IQR 0–0.5), followed by pilocytic astrocytoma and benign bone cancers^e with per-sample medians of 1 (IQR 1–2) and 2 (IQR 0–6) BPJ respectively.

In most histology groups, over 40% of all BPJ occur in complex unexplained clusters, with particularly high rates in liposarcoma (96%), glioblastoma multiforme (85%), osteosarcoma (80%), and melanoma (77%). Cancer types with low rearrangement burden are the major exception to this general preponderance of complex SV. For example, the CLL cohort (median 5 BPJ per sample) has a relatively high proportion of simple deletions (50% of all BPJ, compared to 34%

^eThe benign bone cancers include cartilaginous neoplasm, osteblastoma, and osteofibrous dysplasia.

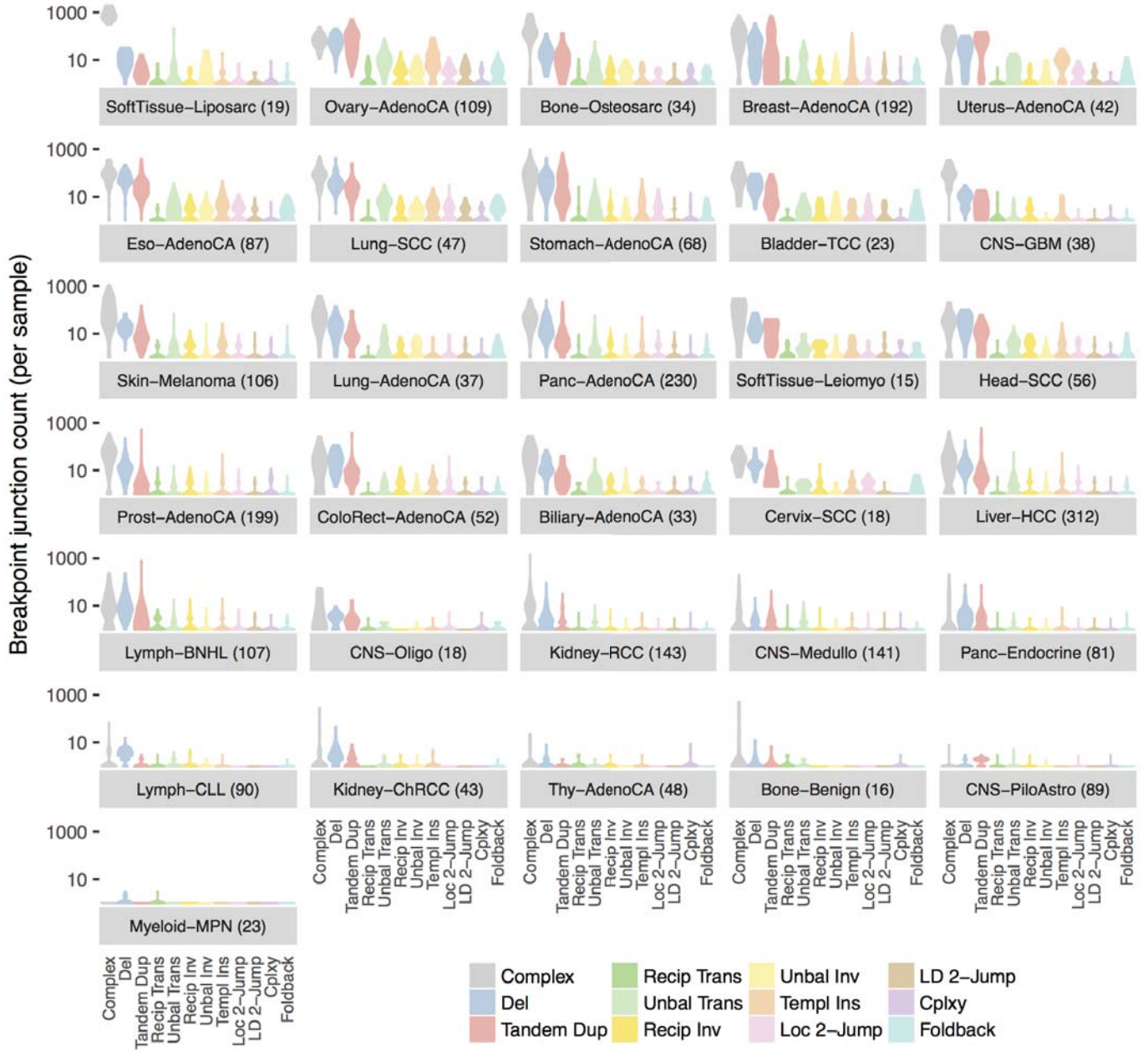


Figure 2.9: The number of classified breakpoint junctions across samples grouped by cancer histology, with the number of samples indicated in parentheses. Histology groups are sorted by the median number of BPJ per sample.

complex). Strikingly, 53% of BPJ in the pilocytic astrocytoma cohort (median 1 BPJ) are tandem duplications, compared to just 8% complex; upon inspection, the vast majority of the tandem duplications generate the characteristic *KIAA1549-BRAF* fusion driver.

Deletions explain the greatest fraction of classified BPJ, and make up a particularly high proportion of all BPJ in colorectal adenocarcinoma (36%), head squamous cell carcinoma (35%), and B-cell non-Hodgkin lymphoma (33%). Just below deletion in overall frequency, tandem duplications are most enriched in adenocarcinomas of the female reproductive tissues—ovary (32% of all BPJ), uterus (32%), and breast (23%)—as well as stomach (26%). Similarly, the three histology groups with the highest overall proportion of templated insertion BPJ are ovary (5.8%), uterus (4.1%), and breast (3.4%).

Overall, only 8.5% of translocation events are reciprocal (rather than unbalanced), although the reciprocal fraction is significantly greater in thyroid (6 out of 9), glioblastoma (18 out of 47), lymphoma (43 out of 124), and prostate (75 out of 228)^f. In contrast, liver cancer is significantly skewed towards unbalanced events, with only 22 reciprocal translocations observed from 755 total.

The preference for reciprocal translocation in the relatively quiet thyroid genome extends to reciprocal exchange at several loci in chromoplexy events. Astonishingly, 14% of BPJ in thyroid adenocarcinoma are attributed to chromoplexy, although the small sample size and low SV burden mean this amounts to only 23 total BPJ across five (out of 48) samples. Nevertheless, this represents an enormous enrichment for balanced chromoplexy, with the next highest proportions of BPJ classified as chromoplexy in pilocytic astrocytoma (1.6%), oligodendroglioma (1.5%) and prostate adenocarcinoma (1.1%)^g.

2.3.2 SV prevalence by sample

The number of BPJ across samples within the same histology class often varies by more than two orders of magnitude, illustrated in Figure 2.10 and Figure D.2. For example, in the osteosarcoma cohort, the two *least* rearranged samples have fewer than 10 identifiable BPJ, whereas, at the other extreme, the two *most* rearranged samples have more than 850 BPJ.

^fTwo-sided binomial test against 0.085 null hypothesis, reporting significant results below 0.001 Benjamini–Hochberg-corrected FDR.

^gAlthough only 1.1% of prostate cancer BPJ are classified as chromoplexy under the stringent definition used in this section, many of the complex unexplained clusters in prostate probably derive from a chromoplexy-type origin, as discussed in Chapter 5.

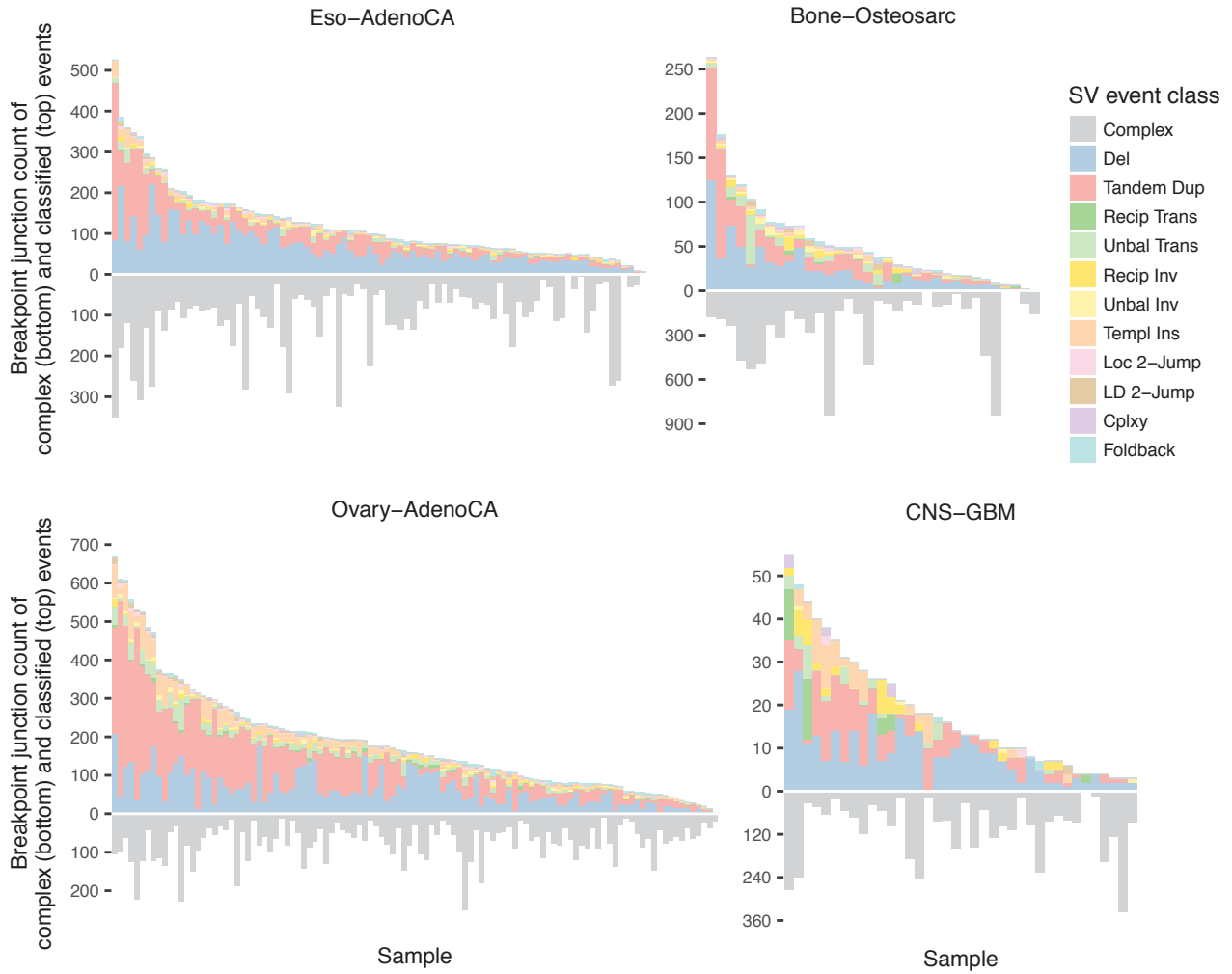


Figure 2.10: Per-sample counts of complex (lower) and classified (upper) breakpoint junctions for esophageal adenocarcinoma, osteosarcoma, ovarian adenocarcinoma, and glioblastoma multiforme. The lower plot for complex BPJ is on a different scale to the upper plot for classified BPJ.

In some histology groups, the number of complex BPJ mildly correlates with the number of classified BPJ—for example, prostate, uterus, and stomach all have Spearman rank correlations above 0.65 (Figure D.3). However, this correlation is weak or non-existent in most cancer groups, and many samples with a high burden of complex BPJ have very few SV classified with simple structure.

Some samples are particularly biased towards one SV class. For example, of the 646 samples with more than 50 classified (not complex) BPJ, 55 samples have more than 80% of their classified junctions assigned to the same type (36 to deletion, 17 to tandem duplication, and 2 to unbalanced translocation).

To find sample covariates associated with SV burden, I considered the 13 histology groups with 40 or more samples and a median BPJ count above ten (1607 samples total). For each separate histology group, I fitted a quasi-Poisson linear regression between a set of covariates and the number of classified or complex BPJ (two separate regressions per histology). The covariates were donor age, mean WGS coverage of the tumour, driver status at genes of interest^h, presence of microsatellite instability (MSI), and whole genome duplication. Each categorical variable was only included in the histology-specific model if present in at least five samples. Any outlying samples with Cook’s distance greater than one were excluded from the model fit. Finally, the p -values of the regression coefficients were adjusted for multiple testing across all histologies, and reported as FDR-adjusted q -values (Benjamini–Hochberg method).

Table 2.3 presents the histology-specific covariate–SV associations below a 10% FDR cut-off. Age is a positive predictor of simple rearrangement burden in prostate cancer, but does not emerge as a significant factor in any other group. MSI does not significantly relate to SV burden in any of the five histology groups with sufficient MSI samples to test. As expected, higher rates of rearrangement are associated with biallelic *BRCA* loss, *TP53* mutations, and whole genome duplication in several tissues. Driver mutations in the *NEAT1* long non-coding RNA are associated with higher rates of complex SV in esophagus and simple SV in prostate and liver. Promoter mutations at the *WDR74* gene have a particularly strong correlation with complex BPJ in B-cell non-Hodgkin lymphoma. The prospect of a significant link between rearrangement burden and non-coding disruptions in RNA genes or promoter regions exemplifies the novel findings made possible by WGS data.

^hThe gene set considered was the top 40 most commonly annotated drivers (only considering SNV and indel mutations) from the PCAWG driver catalogue described by Sabarinathan et al. (2017). An additional variable registered biallelic loss of *BRCA1* or *BRCA2* in germline and/or soma. Genes were only included in histology strata with five or more affected samples.

In the liver cancer cohort, the depth of sequencing coverage positively correlates with the number of simple and complex BPJ identified, perhaps indicating a tendency towards false negatives in lower coverage samples and/or false positives in higher coverage samples. On the other hand, coverage may simply be a proxy for some hidden variable/s unevenly distributed across the constituent projects, as the sub-cohorts of liver cancer from France and the Riken center in Japan have lower coverage (range 31–49 \times) than the sub-cohorts from the USA (range 55–80 \times) or the Japanese National Cancer Centre (range 33–126 \times).

Table 2.3: Significant associations between sample covariates and the number of classified or complex BPJ in a histology group. The effect size (ES, interpreted as linear effect on the natural logarithm of the mean) is estimated by quasi-Poisson multivariate linear regression, stratified by histology and printing only those associations with Benjamini–Hochberg corrected q -value (Q) below 0.1 (121 other rows not shown). The number of samples with each categorical variable is indicated in parentheses.

Histology	Variable	Classified BPJ			Complex BPJ		
		ES	Q		ES	Q	
Prost-AdenoCA(199)	Age	0.05	0.000	***	0.02	0.195	
Panc-AdenoCA(230)	BRCA.bi(13)	1.16	0.000	***	-0.52	0.471	
Breast-AdenoCA(192)	BRCA.bi(18)	0.80	0.078		-0.29	0.651	
Ovary-AdenoCA(109)	BRCA.bi(22)	0.50	0.042	*	-0.13	0.730	
Prost-AdenoCA(199)	BRCA.bi(5)	1.15	0.013	*	0.53	0.439	
Liver-HCC(312)	CTNNB1(80)	-0.47	0.117		-0.84	0.006	**
Eso-AdenoCA(87)	NEAT1(12)	0.43	0.237		0.55	0.088	
Prost-AdenoCA(199)	NEAT1(12)	0.92	0.000	***	0.41	0.374	
Liver-HCC(312)	NEAT1(91)	0.55	0.004	**	0.11	0.764	
Skin-Melanoma(106)	NRAS(25)	-0.36	0.295		-1.13	0.026	*
Prost-AdenoCA(199)	PTEN(10)	0.59	0.089		0.33	0.541	
Liver-HCC(312)	SeqCover	0.02	0.003	**	0.02	0.001	***
Panc-AdenoCA(230)	SF3B1(6)	0.73	0.078		-0.34	0.764	
Skin-Melanoma(106)	TERT(53)	-0.31	0.247		-0.65	0.099	
Breast-AdenoCA(192)	TP53(100)	1.14	0.000	***	-0.08	0.859	
Panc-AdenoCA(230)	TP53(172)	0.29	0.295		0.76	0.005	**
Uterus-AdenoCA(42)	TP53(30)	1.37	0.087		0.99	0.201	
Liver-HCC(312)	TP53(99)	-0.01	0.982		0.47	0.039	*
Lymph-BNHL(107)	WDR74(13)	-0.11	0.894		1.36	0.001	***
Stomach-AdenoCA(68)	WGD(29)	1.28	0.008	**	0.97	0.082	
Skin-Melanoma(106)	WGD(58)	0.56	0.013	*	0.49	0.203	
Ovary-AdenoCA(109)	WGD(66)	0.39	0.117		0.50	0.014	*
Liver-HCC(312)	WGD(77)	0.34	0.194		0.51	0.024	*
Panc-AdenoCA(230)	WGD(90)	0.38	0.078		0.27	0.192	
Breast-AdenoCA(192)	WGD(95)	-0.03	0.938		0.63	0.006	**

BRCA.bi is biallelic *BRCA1* or *BRCA2* loss including germline status. SeqCover is the tumour sample mean coverage. WGD is whole genome duplication.

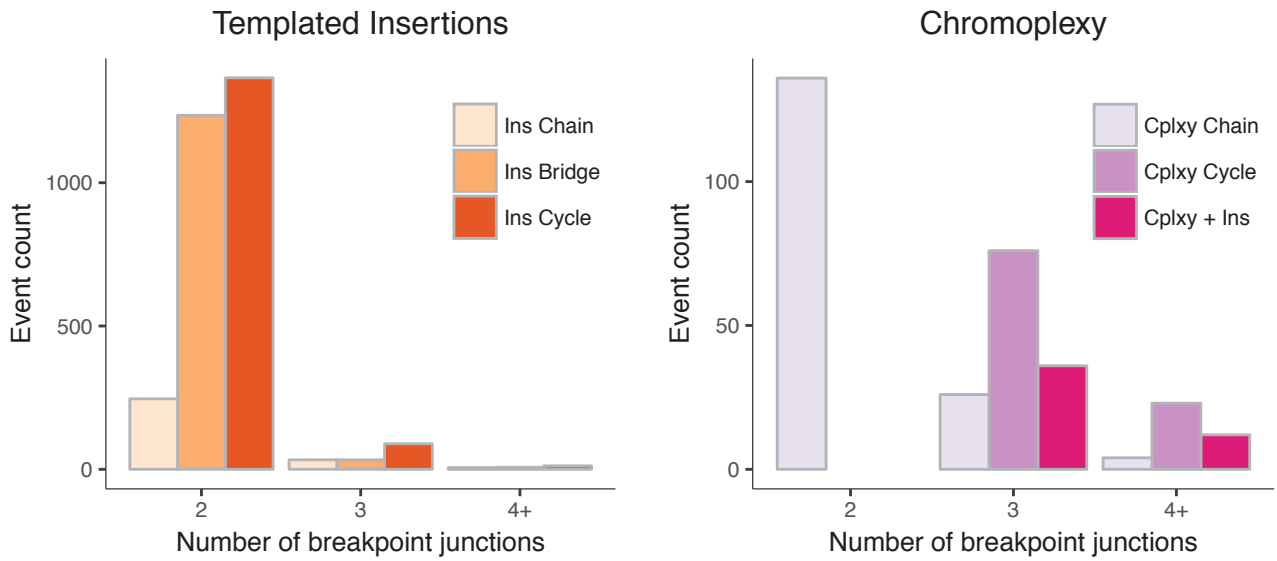


Figure 2.11: Number of BPJ in all templated insertion and chromoplexy events. Only chromoplexy *chains* have two BPJ because the minimal chromoplexy cycle is classified as reciprocal translocation and the minimal chromoplexy plus insertion cycle is classified as insertion bridge.

2.3.3 Length of templated insertions and chromoplexy

Setting aside the complex unexplained clusters, most simple SV classifications outlined in Table 2.2 refer to a specific configuration of one or two BPJ. The exceptions to this are the templated insertion and chromoplexy classifications which involve two or more BPJ, as tallied in Figure 2.11.

The shortest events in the chromoplexy group are chains of two BPJ forming one $[+ -]$ motif and two singleton ends. This minimal case may be a poor representation of the chromoplexy term, originally defined for several DSB positions repaired through balanced exchange. Instead, it may be preferable for future classification schemes to regard such events as another translocation variant (perhaps ‘split’ translocation), where the two sides of one DSB ligate to different partners, without the chromoplexy hallmark of multi-locus reciprocity.

For templated insertion, the longest observed events are a bridge of eight BPJ (Figure 2.12, in cervix), two cycles of seven and six BPJ (Figure 2.13, in uterus and pancreas), and one chain of five BPJ (not shown, in uterus)ⁱ. The pancreatic and both uterus samples also have three to five additional and independent templated insertion events in other genome regions.

ⁱNone of the four samples with these long templated insertions are annotated with any germline or somatic mutations or copy loss affecting *BRCA1* or *BRCA2*.

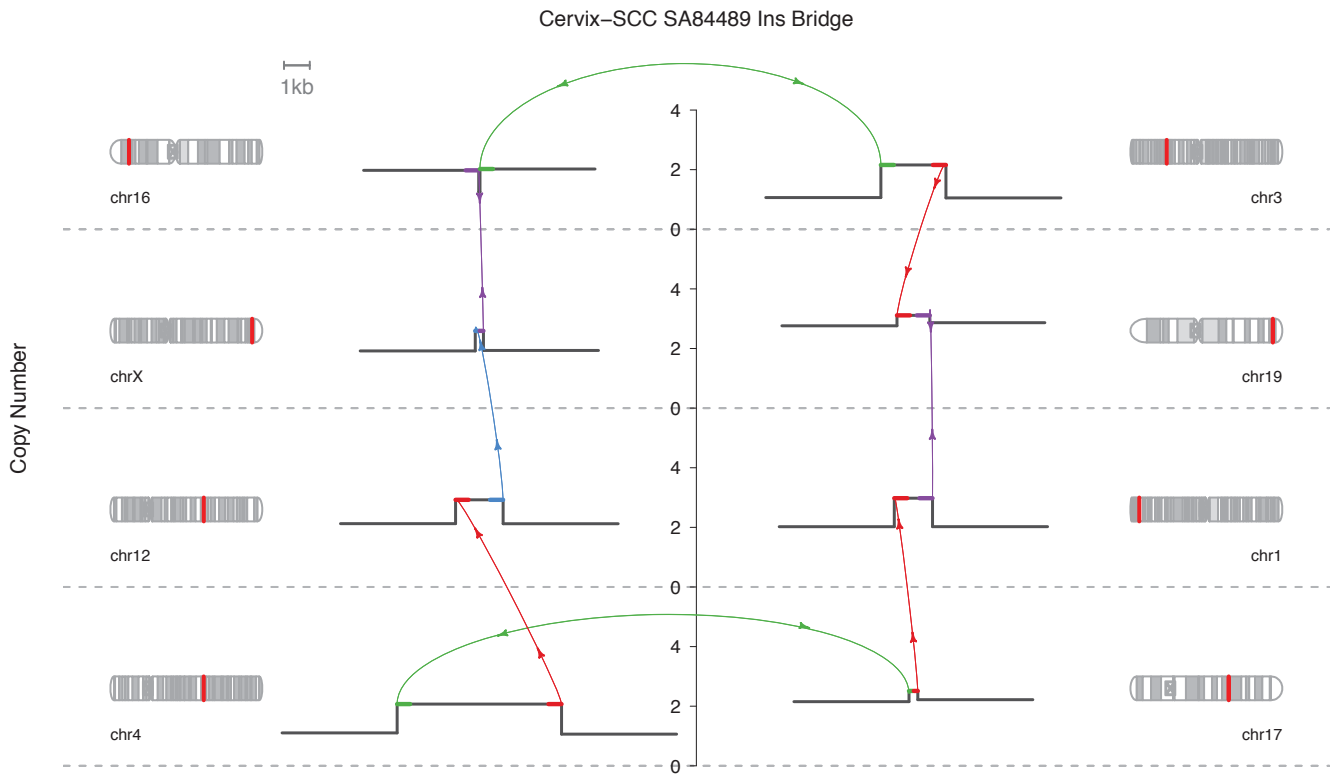


Figure 2.12: Longest templated insertion bridge event

The insertion bridge in Figure 2.12 copies seven distant genome segments into a break on chr16 in a cervical squamous cell carcinoma, coinciding with the *CLEC16A* gene^j. Interestingly, the longest insertion ‘cycle’ (Figure 2.13) has unbalanced CN estimates either side of the event on chr3 and chr21. If these CN estimates are correct, then a more logical mechanistic explanation is a long templated insertion *chain* forming an unbalanced translocation between chr3 and chr21, with a subsequent tandem duplication spanning the entire set of inserted fragments (the BPJ in purple would be the tandem duplication, similar to the translocation and tandem duplication example in Figure 2.5).

For chromoplexy, the longest observed events are one cycle of six BPJ and ten cycles of five BPJ (mix of pure chromoplexy as in Figure D.4 and chromoplexy with insertion as in Figure D.5), whereas the four longest chains comprise four BPJ (one illustrated in Figure 2.8). Some chromoplexy classifications involve multiple adjacent $[+ -]$ motifs on the same chromosome, and may involve local breakage in addition to the balanced exchange between distant loci on different chromosomes or arms.

^j*CLEC16A* is annotated by the COSMIC database (Forbes et al., 2015) as having over-expression in 6% and under-expression in 2% of cervical cancers. *CLEC16A* polymorphisms are associated with multiple sclerosis and type 1 diabetes (Soleimanpour et al., 2014).

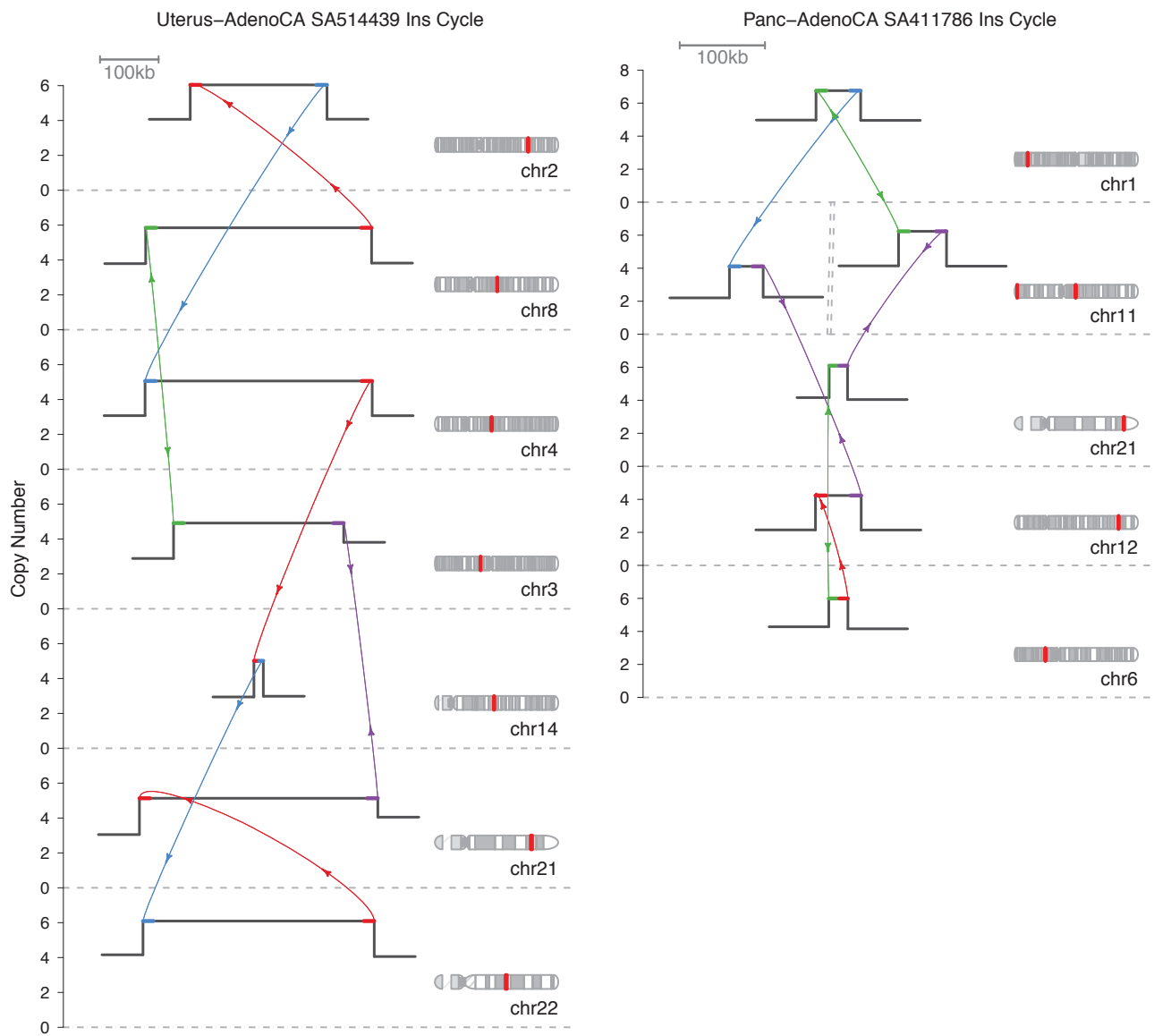


Figure 2.13: Longest templated insertion cycle events

As the current SV classification scheme for chromoplexy and templated insertion requires all $[+-]$ and $[-+]$ motifs to be isolated in separate footprint divisions, some similar SV patterns obfuscated by additional local break sets are consigned to the complex unexplained BPJ set. Larger events with profiles reminiscent of chromoplexy or templated insertion are presented in Chapter 5.

2.4 Size distribution of SV classes

Event size is the simplest structural property, yet a historic lack of appropriate SV classification methods for WGS data have prohibited structurally-aware size analysis across a pan-cancer cohort. The existing literature on CNA size registers the aggregate effect of many heterogeneous and complex rearrangement mechanisms, and offers little insight into the underlying event properties. Tandem duplication and deletion size is a known correlate of *BRCA* status in breast cancer (Nik-Zainal et al., 2016), indicating that event size distribution is a characteristic readout of the mutational mechanism.

2.4.1 Deletion and tandem duplication

The overall size distributions for deletion and tandem duplication are multi-modal, with recurrent peak positions shared across different histology groups, even as their relative contribution varies (Figure 2.14). For example, deletion size peaks around 2 kb and 160 kb in most cancer types, and is dominated by the small peak in lung squamous cell carcinoma, by the large peak in colorectal adenocarcinoma, and is quite evenly apportioned in liver and stomach cancers. Peak duplication sizes are not as consistent across all cancer types, with the striking exception of shared modes around 8 kb and 300 kb in breast, ovary, and prostate. The tandem duplication pattern is not so bi-modal in other tissues, but varies between large events over 100 kb in uterus and pancreatic endocrine cancers, and smaller events below 50 kb in cervical and colorectal cancers.

To assess whether these cohort-level patterns emerge from a consistent multi-modal distribution preserved within individual samples or the summation of sample-specific size preferences, I set out to cluster the constituent samples. Running separate analyses for deletion and tandem duplication, I considered only those samples with 30 or more events, in histology groups with at least

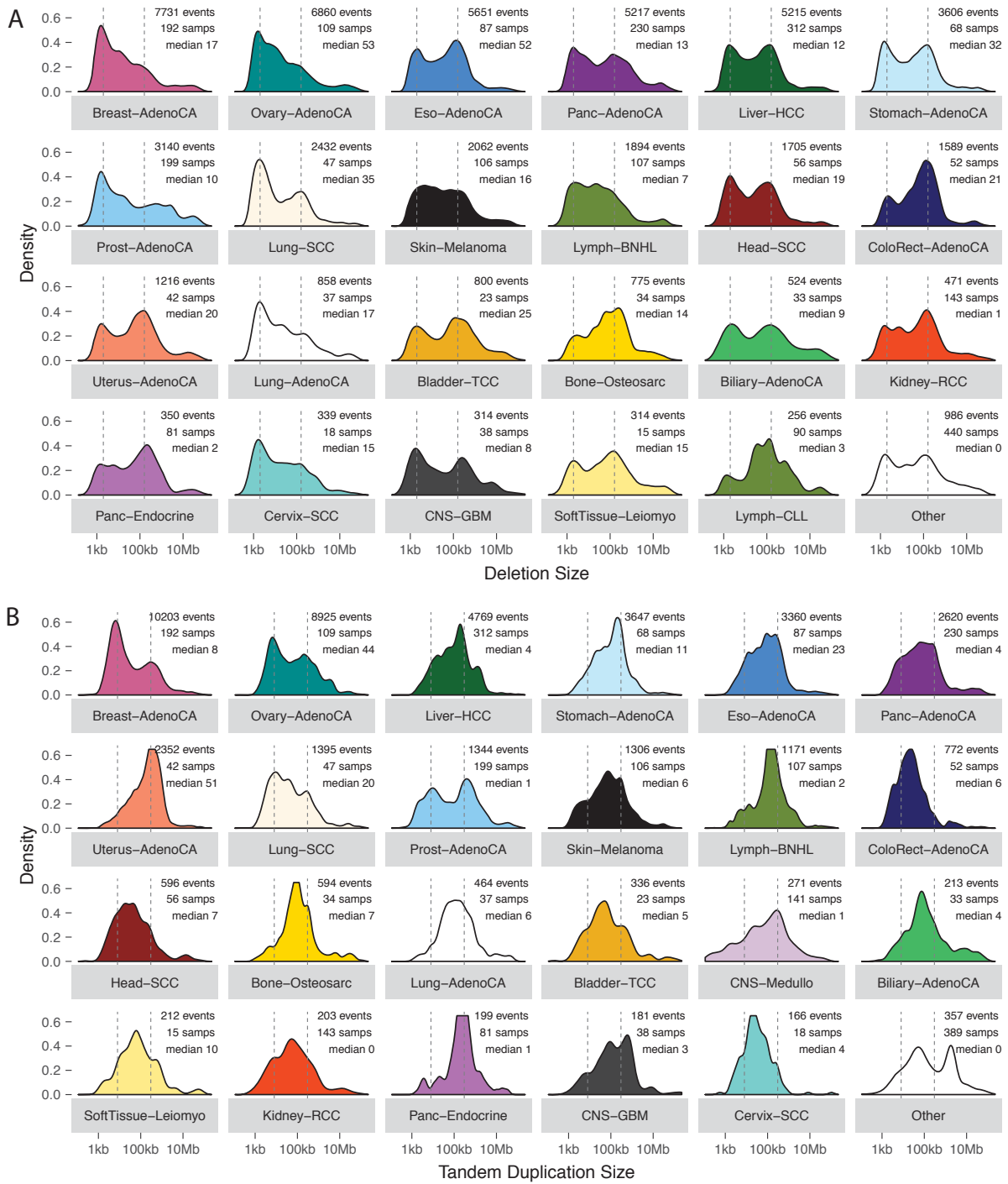


Figure 2.14: Deletion and tandem duplication size distributions over a \log_{10} scale. Histology groups are sorted by total number of events in the cohort. The median number of events per sample is annotated in the top right for each group. Guide lines are marked at 2 kb and 160 kb for deletion (A), and 8 kb and 300 kb for tandem duplication (B).

five such samples^k. Then, I performed hierarchical agglomerative clustering using the earth mover’s distance^l between samples’ size distributions, cutting clusters at the complete linkage threshold of 0.8^m.

Figures 2.15 and 2.16 illustrate the deletion and tandem duplication size distributions in a subset of individual samples randomly chosen to represent each cluster. Events within a sample are predominantly drawn from a uni-modal size range, often with narrow variance. Of the 14 samples in deletion ‘cluster 7’ with extremely large events spanning hundreds of kilobases, 13 are pancreatic cancers, revealing a specific large deletion phenotype almost unique to that tissue. The largest tandem duplications are found in eight samples assigned to ‘cluster 5’, with some degree of bimodality and an average event frequency well above the norm. For example, one unusual liver sample (SA269323) has 574 tandem duplications with an inter-quartile size range of 609–1710 kb (subset plotted in Figure D.6). Upon inspection, these events are: evenly distributed across the genome; mostly (70%) agreed upon by all four calling algorithms; have accompanying CN support (99% logical); and therefore appear to be real events. At the other extreme of small events, two outlying prostate cancers have deletions (SA530428; SA506736) and tandem duplications (SA530428) almost exclusively smaller than 2 kb. Upon inspection, these events are: evenly distributed across the genome; mostly (> 80%) returned by only two callers (predominantly BRASS+SvABA); have somewhat unreliable CN support (~ 60% logical); and are possible false positives (although CN calling is inherently difficult in small segments and may be inaccurate even in real SV).

Confirming the pattern in breast cancer (Nik-Zainal et al., 2016), 21 of 24 samples with biallelic *BRCA1* loss in the tandem duplication analysis belong to the small size ‘cluster 2’ group, while all 34 samples with biallelic *BRCA2* loss in the deletion analysis are assigned to the small size ‘cluster 3’ or ‘cluster 4’ⁿ.

These results suggest that multiple mechanisms generate deletions or tandem duplications, with individual samples predominately affected by one pathway acting over a tell-tale size distribution.

^k538 samples included for deletion; 288 samples included for tandem duplication.

^lThe earth mover’s distance measures the minimal work (mass \times distance) to transform between two probability distributions. Here, I use bins at 0.25 intervals along a \log_{10} scale.

^mIn context, the 0.8 earth mover’s distance means that *any* two samples in the same cluster must be similar enough that if 60% of their size distribution is the same, then the remaining 40% must be within a factor of 100. Equally, if 20% of their size distribution is the same, then the remaining 80% must be within a factor of 10.

ⁿFor tandem dup, only one of four *BRCA2* samples is assigned to cluster 2 (small dup). For deletion, 14 of 21 *BRCA1* samples are assigned to clusters 3/4 (small deletion).

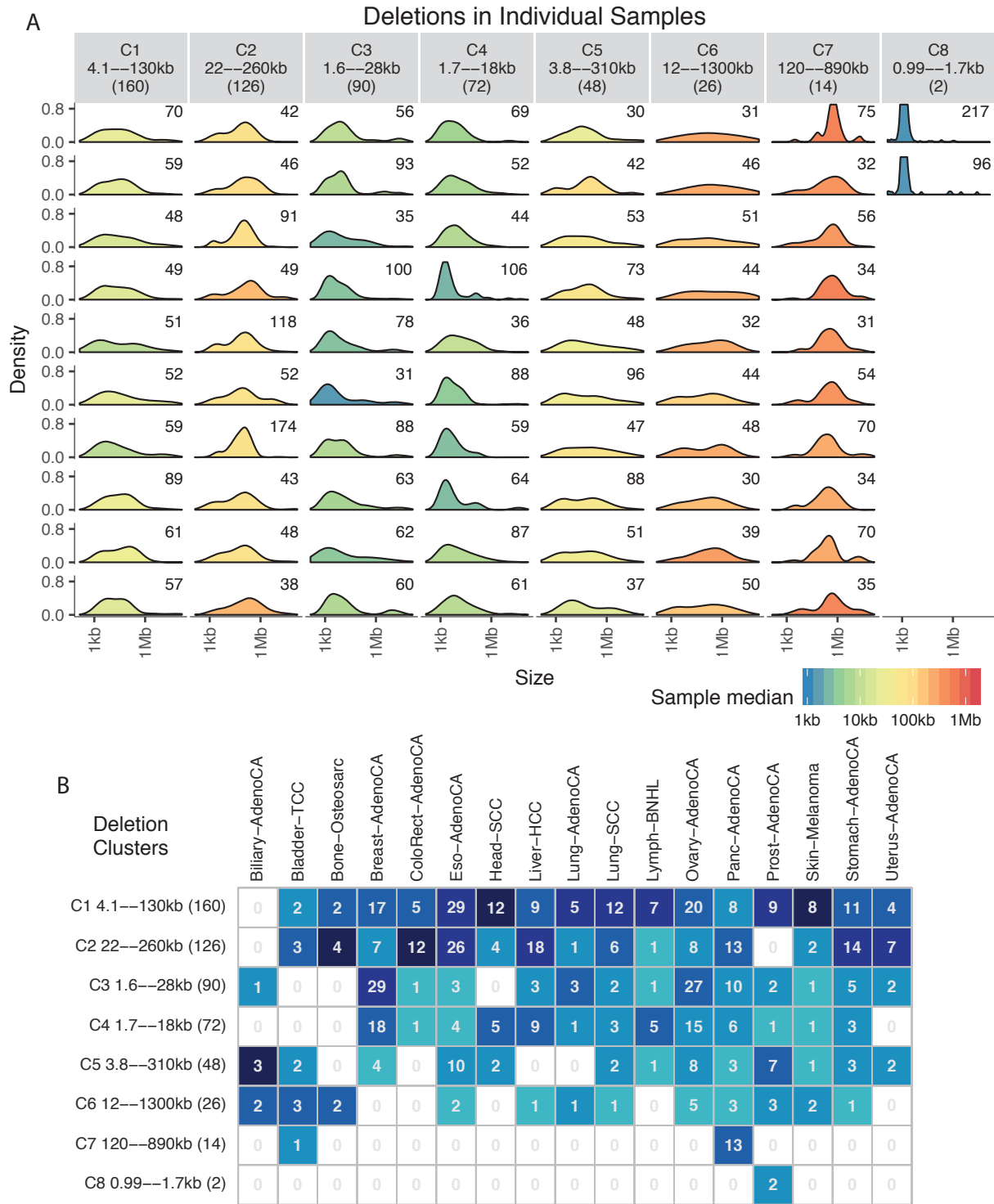


Figure 2.15: Samples with 30 or more deletions, clustered by their size distribution. Clusters are labelled with the inter-quartile deletion size range (pooling samples in the cluster), and the number of samples in parentheses. (a) Deletion size distributions of randomly chosen individual samples from each cluster, coloured by the median size, with number of deletions annotated top-right. (b) The number of samples allocated to each cluster, shaded by the proportion of samples in each histology group.

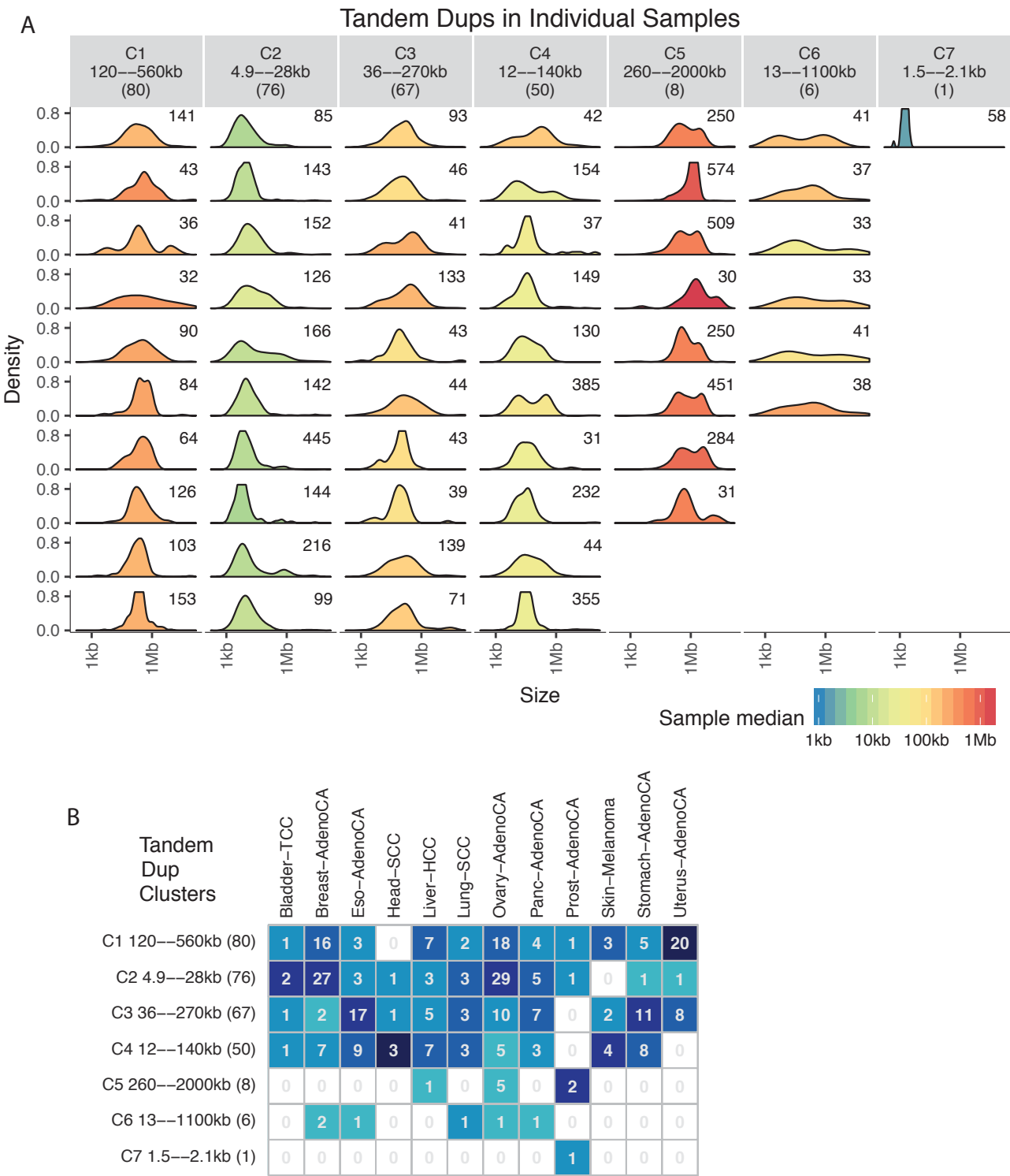


Figure 2.16: Samples with 30 or more tandem duplications, clustered by their size distribution. Clusters are labelled with the inter-quartile duplication size range (pooling samples in the cluster), and the number of samples in parentheses. (a) Tandem dup size distributions of randomly chosen individual samples from each cluster, coloured by the median size, with number of duplications annotated top right. (b) The number of samples allocated to each cluster, shaded by the proportion of samples in each histology group.

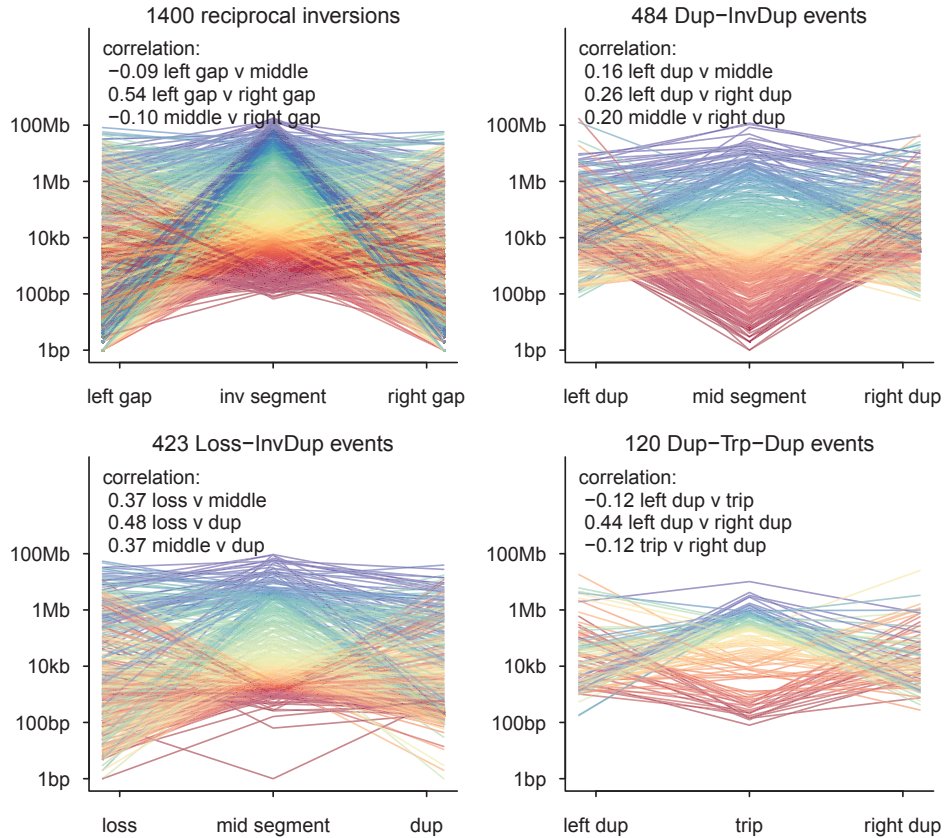


Figure 2.17: Segment size distribution for reciprocal inversion and local 2-jumps, shaded by the size of the middle segment. Pearson correlation coefficients between segment lengths on a log₁₀ scale are annotated top left.

2.4.2 Reciprocal inversion and local 2-jumps

The SV structures defined by specific configurations of two inverting BPJ are the reciprocal inversion, and three sub-classes of ‘local 2-jump’. In each case, the event size is comprised of three distinct segments between adjacent breakpoints, as summarised in Figure 2.17. In all four structures, the two outermost segments (such as the gaps bordering a reciprocal inversion or the duplications in the dup-inv-dup or dup-trp-dup) are modestly correlated in size, presumably reflecting some mechanistic symmetry, such as the length a MMBIR D-loop travels before dissociating and triggering another round of strand invasion. Although the correlations suggest some internal consistency *within* each event, the overall size range varies massively, from about 1 kb to over 100 Mb. Some reciprocal inversion classifications consist of a tiny (< 1 kb) inverted segment captured in a much larger deletion spanning several megabases, and, from a copy number standpoint, might alternatively be considered a variant of canonical deletion rather than a true reciprocal inversion as classically imagined.

2.4.3 Templated insertion

Regarding templated insertion SV, the insert fragments ($[-+]$ motifs) are remarkably bi- or tri-modal in every histology group, with recurrent peaks around 200 bp, 8 kb, and 300 kb (Figure 2.18A). Intriguingly, these larger two peak positions match those in the tandem duplication analysis, and implicate common underlying ‘template and replicate’ mechanisms which have previously been characterised in the bimodal context of short and long tract gene conversions (Nagaraju et al., 2009; Yim et al., 2014).

In general, inserts in cycle events tend to draw from the larger sizes, whereas inserts in bridge events are predominantly under 1 kb. The pattern varies across cancer types, with cycles of small inserts being relatively common in ovary, breast, and prostate, but quite rare in uterus, glioblastoma, and esophagus.

Insertion bridge events are also characterised by deletion size on the host chromosome ($[+-]$ motif), with the insert fragment/s slotting in the gap (Figure 2.18B). This gap is typically smaller than 1 kb, with little variation across cancer types. If the mechanism of formation involves template switching, it seems the event most often resolves with polymerase re-start just after the point of departure, causing minimal sequence loss.

Events involving two or more insert fragments (all cycles, plus bridges and chains with ≥ 3 BPJ) fall into two distinct clusters (Figure 2.18C): those with highly correlated insert sizes, and those with at least one small (< 1 kb) and one arbitrarily-sized insert. I found no obvious associations between these two clusters and either *BRCA* status, sub-class (chain, cycle, or bridge), or histology.

As shown in Figure 2.18D, most events with three or more large (> 1 kb) inserts have extremely consistent internal size, even as the mean size varies between events. There are also many events with a mix of small and large insert sizes.

While the distinctive copy gain patterns imply that most templated insertion events are generated by a replication-based mechanism, some *intra*-chromosomal insertion chains of two BPJ (not shown) are also consistent with DSB-mediated deletion with a small intervening fragment rescued in its native orientation in the junction (similar to the deletion-type reciprocal inversion cases discussed in the previous section). Future classification methods may wish to separate this special case.

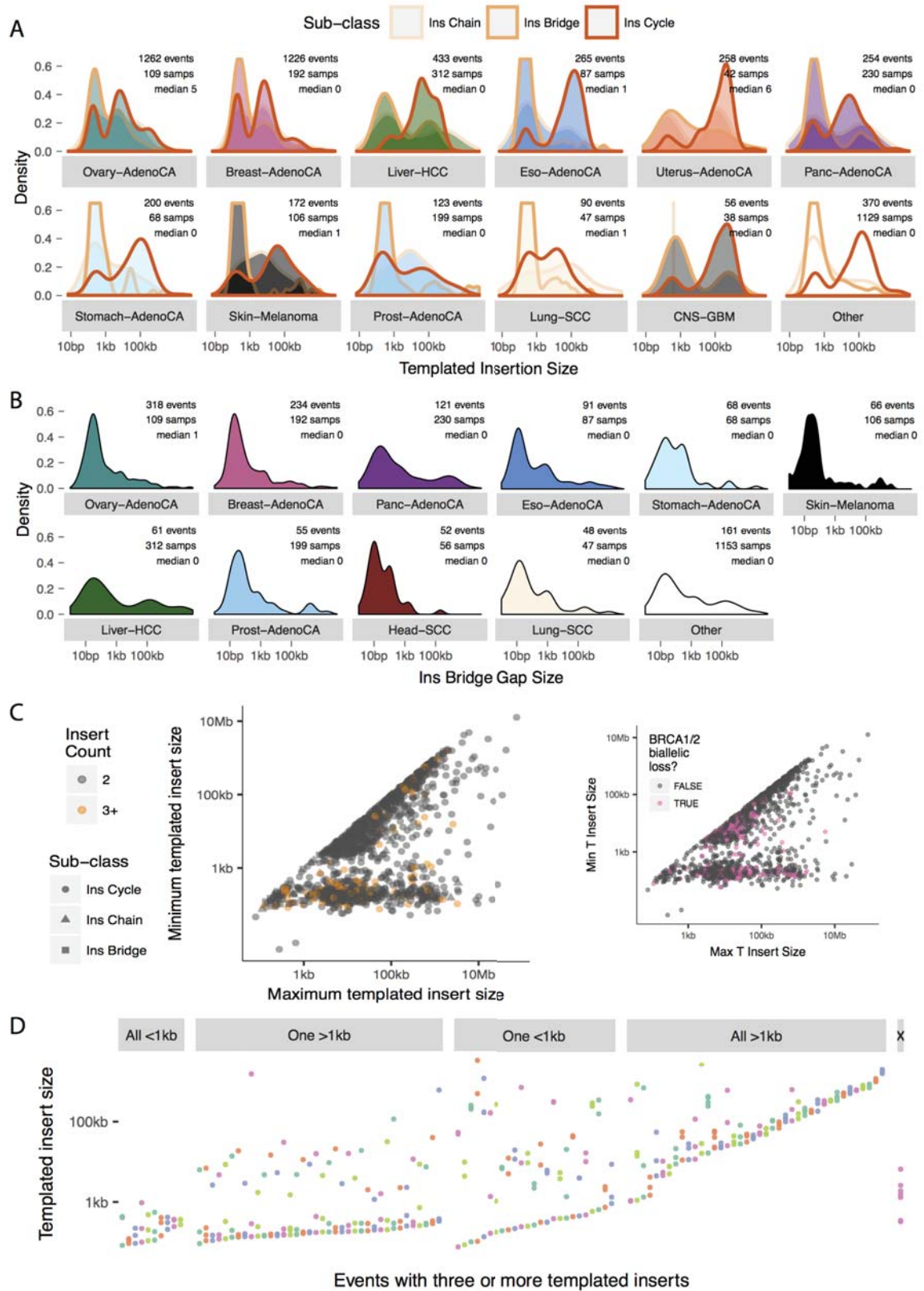


Figure 2.18: (a) Size distribution of templated inserts ($[-+]$ motifs) by sub-class. (b) Size distribution of insertion bridge gaps ($[+-]$ motifs). (c) Correlation between the smallest and largest insert in the same event (no chains/bridges of only two BPJ). (d) Events with three or more inserts, sorted by size composition.

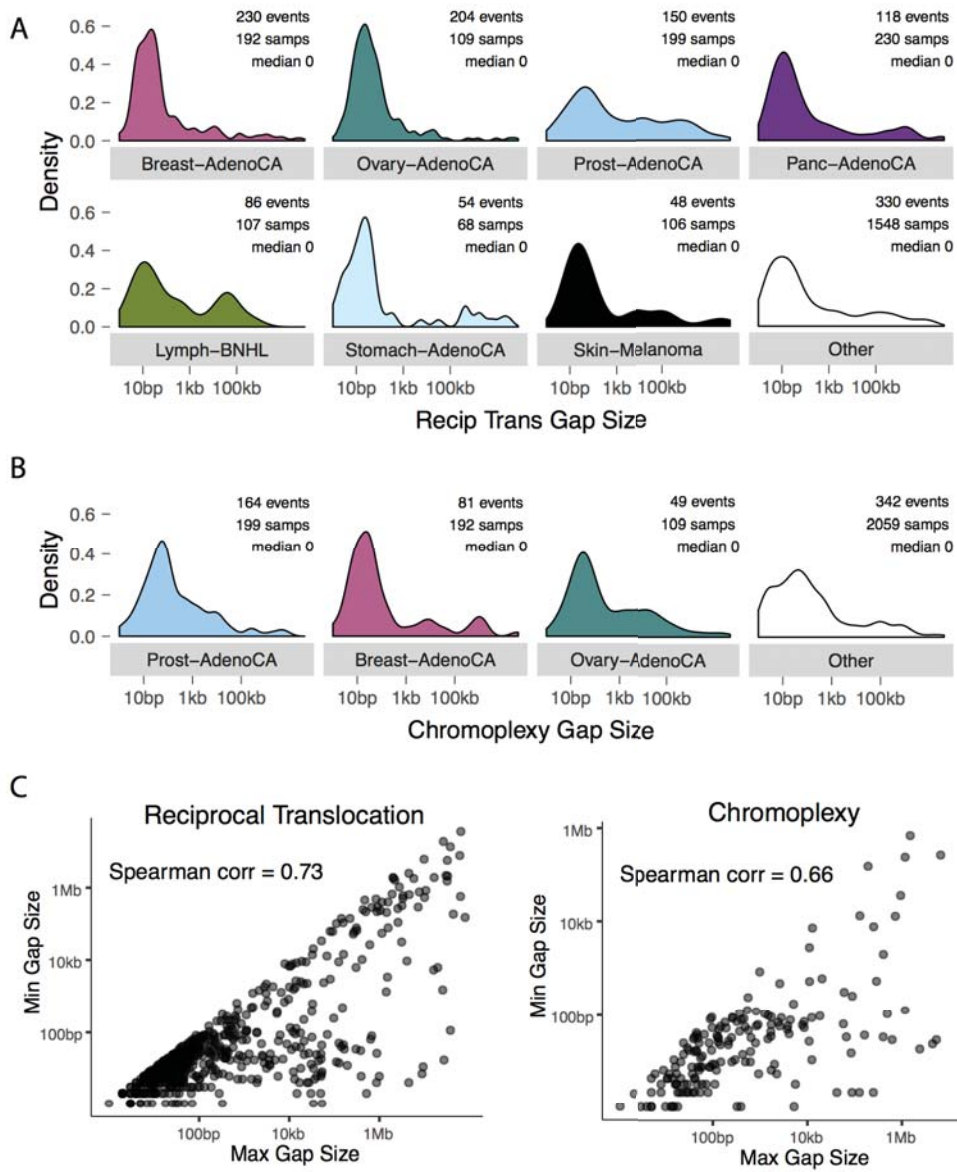


Figure 2.19: Gap sizes of $[+−]$ motifs in (a) reciprocal translocation, (b) chromoplexy, and (c) correlation within the same event.

2.4.4 Gaps in reciprocal translocation and chromoplexy

The gap size ($[+−]$ motif) in reciprocal translocation and chromoplexy is typically smaller than 1 kb, but occasionally stretches beyond 100 kb in this classification scheme (Figure 2.19). Translocations with larger stretches of lost sequence are particularly prevalent in prostate cancer and lymphoma, and possibly arise from ligation repair across two sets of two correlated break positions rather than extreme resection at a pair of individual DSBs.

Within individual events, the gap size at distant loci is modestly correlated.

Gap size correlation may result from the underlying biology—such as nuclease activity levels eroding free DNA ends—or bias imposed by the BPJ clustering method which only groups breaks within a sample-specific threshold by orientation type.

2.5 Homology at the breakpoint junction

With the exception of NHEJ, most DSB repair pathways rely on some degree of sequence homology to facilitate annealing or strand invasion. MMEJ and MMBIR only require a few bases of homology, whereas SSA, BIR, and HR require much longer matching (details unclear, see Renkawitz et al. (2014) and Anand et al. (2017)). WGS data provides enough sequence detail at each breakpoint junction to detect short runs of homology, although this is somewhat muddled in the PCAWG dataset where consensus BPJ calls are merged from four different callers. Despite some slight confounding from different SV calling algorithms, the consensus estimates are sufficient to indicate the relative degree of MH enrichment across samples and SV classes. Longer tracts of potentially imperfect homology are not reported with the BPJ calls, but could be estimated in future research by comparing the reference genome sequence either side.

2.5.1 Microhomology by SV class and histology

To analyse the extent of microhomology enrichment in the PCAWG cohort, I modelled MH as an ordinal variable from zero to four-plus bases using proportional odds (cumulative logit) regression with histology group as the sole predictor in separate strata for each SV class (excluding complex unexplained BPJ). In each model fit, the baseline MH level was set by 100,000 dummy observations from the background of random position pairs in the callable genome space^o. For this analysis, I pooled all histology groups with fewer than one thousand classified BPJ into a mixed ‘Other’ category, and only included histologies with at least 30 BPJ in the SV class stratum. To correct for multiple testing and *p*-value inflation from the dummy sample size, I ran a conservative

^oIn the callable genome space (see Section 3.1.1), the empirical MH distribution at random position pairs is $\text{Pr}(0) = 0.743$, $\text{Pr}(1) = 0.187$, $\text{Pr}(2) = 0.050$, $\text{Pr}(3) = 0.014$, $\text{Pr}(\geq 4) = 0.006$. Curiously, this empirical distribution has slightly more one-length MH than the theoretical proportion in completely random sequence. This possibly emerges because of microsatellite depletion in callable genome areas, and overall GC bias.

Bonferroni adjustment over the coefficient p -values from all model fits, and report significant MH enrichment at a 0.01 FWER threshold.

Figure 2.20 shows the MH distribution for each SV class and cancer type. Randomly matched junctions have one or more MH bases about 25% of the time, so any significantly larger proportion indicates activation of non-NHEJ repair. Overall, ovarian cancer has the greatest degree of MH enrichment, while prostate cancer has the least. Many distributions peak at two bases, indicating a mechanistic role for very short MH. Reciprocal translocation is the only SV class with no significant MH, suggesting that NHEJ is perhaps the only major mechanism of reciprocal translocation. All other SV classes have some degree of MH enrichment, from low levels observed in reciprocal inversion and unbalanced translocation to high levels observed in tandem duplication, foldback, and many other structural forms. Chromoplexy—as the multi-locus extension of the no-MH reciprocal translocation class—does have more MH than random expectation, perhaps indicating a greater time delay between DSB formation and repair, during which time strand resection triggers a switch to MMEJ mechanisms. Surprisingly, of the SV classes hypothesised to result from BIR/MMBIR—that is templated insertion, local 2-jumps, and some fraction of tandem duplications—about half of these events have no discernible MH. This may reflect: the ability of low-fidelity translesion polymerases to create small *de novo* MH (Sakofsky et al., 2015; Ceccaldi et al., 2016); failure to report homology interspersed with mismatches; and/or the insertion of non-templated bases which some SV callers treat as a mutually exclusive feature to MH.

2.5.2 Microhomology by sample

To roughly gauge MH variation across samples, I considered deletion and tandem duplication in four cancer types (esophagus, ovary, pancreas, and prostate) and compared the samples with the most events against the pool of all other samples in the same histology group using the proportional odds model described above (without the dummy background observations).

As shown in Figure D.7, most samples have reasonably consistent MH distributions, with a few notable exceptions. One pancreatic and five prostate samples have considerably greater MH in their tandem duplications, a signature of sample-specific repair preferences. The underlying reason is unclear, although two of the high-MH prostate examples are known to have biallelic *BRCA* loss. Interestingly, some samples have considerably less MH than the pool of other

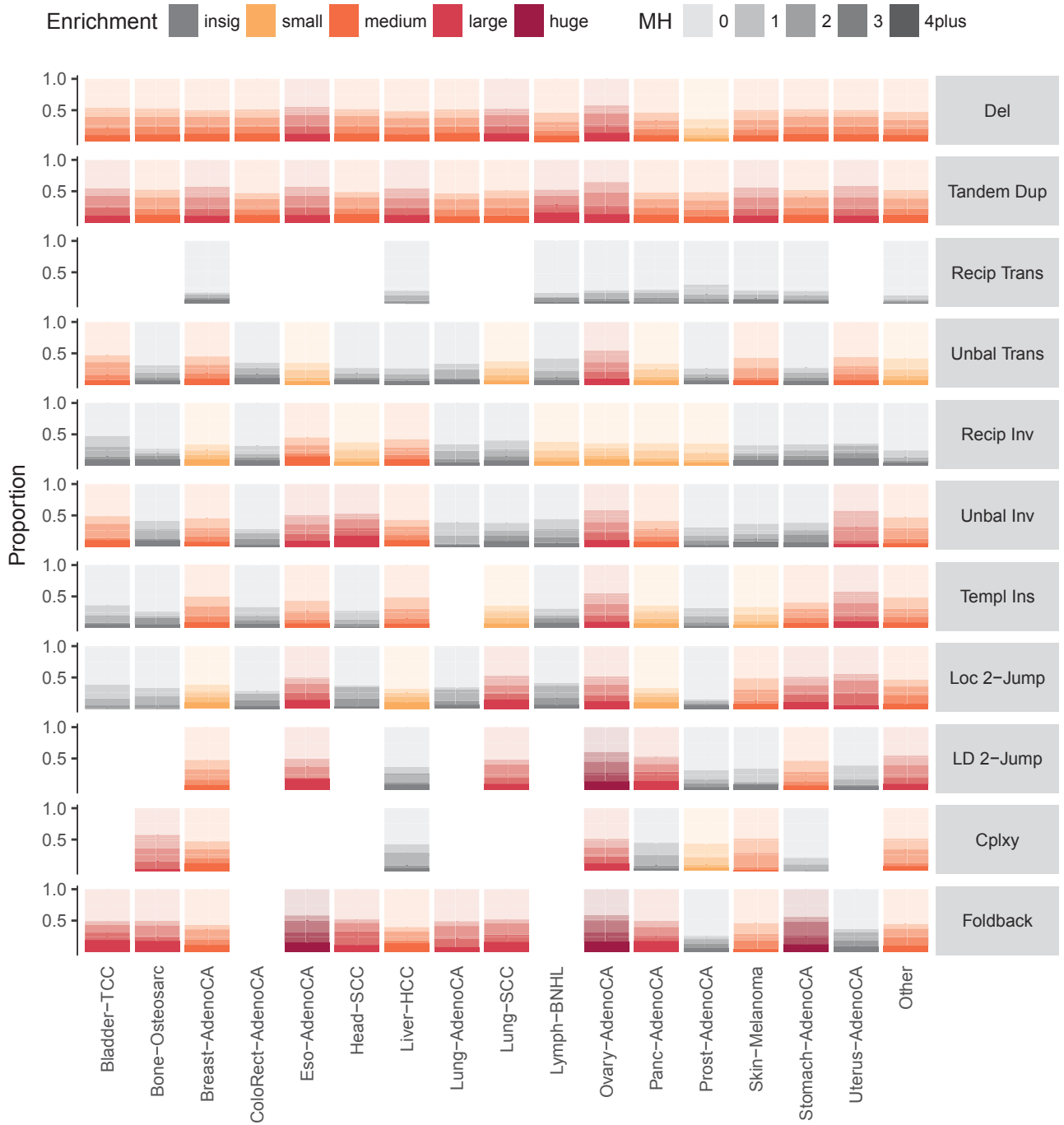


Figure 2.20: Distribution of microhomology at the breakpoint junction for different SV classes, separated by cancer histology. The magnitude of significant enrichment (compared to random background expectation) is coloured by the proportional odds regression coefficient, split into small (0.5–1.0], medium (1.0–1.5], large (1.5–2.0], and huge (2.0– ∞) effect sizes. Non-significant categories (at Bonferroni-adjusted 0.01 threshold) are shaded grey. Categories with fewer than 30 BPJ are excluded from consideration and left blank.

samples. Although this could be interpreted (up to a point) as a preference for NHEJ, about a quarter of random junctions should have at least one base of homology. For the samples with significantly less MH, the difference may be attributed to the variable treatment of non-templated base insertions by the different SV callers. Many SV events insert a few random nucleotides into the junction, which may be considered part of the potential MH sequence by some algorithms, and mutually exclusive to MH by others. In the four examples with significant MH depletion (deletion in one esophagus and two prostate samples; tandem duplication in one esophagus), the vast majority of events are returned by only two SV callers (in a variety of combinations), perhaps indicating some systematic problem with breakpoint reconstruction, or loss of MH information due to different modelling approaches and the consensus reporting method.

2.6 Kataegis and SV classes

Kataegis regions are dense hypermutation clusters of several SNV in far closer proximity than chance expectation (Nik-Zainal et al., 2012). Most clusters are attributed to APOBEC cytidine deaminase activity targeting single stranded DNA (Taylor et al., 2013), accounting for the observed signature of strand-coordinated C>N SNV in a TpC context with frequent proximity to SV breakpoints. Nik-Zainal et al. (2016) recently described a non-APOBEC signature in just 1% of all breast cancer kataegis foci, mostly consisting of T>G and T>C mutations with a pattern reminiscent of translesion polymerase η activity. This finding was further investigated by Supek and Lehner (2017), who propose that polymerase η participates in error-prone mismatch repair following carcinogen exposure.

Although kataegis clusters have long been associated with rearrangement breakpoints, a lack of appropriate SV classification has prevented structurally-aware analysis of hypermutation frequency around different SV classes.

2.6.1 Defining kataegis regions

To correlate kataegis events with SV in the PCAWG cohort, I searched for hypermutation clusters by fitting a piecewise constant model^P to the sequence of inter-SNV distances on a \log_{10} scale, one chromosome at a time. All segments

^PPiecewise constant fit assuming Gaussian noise with constant standard deviation, using the narrowest-over-threshold method from Baranowski et al. (2016).

with at least five SNV and a mean inter-SNV distance less than 1 kb^q were defined as kataegis, with any gaps over 10 kb dividing separate clusters. Each cluster was associated with the closest SV breakpoint up to a maximum distance of 50 kb, and labelled as APOBEC type if more than 70% of the SNV were C>N or G>N. To avoid false positive clusters and recurrently mutated immune loci, I excluded 39 samples^r with extremely high mutational burdens (more than 150,000 SNV) as well as the entire lymphoma and CLL cohorts.

2.6.2 Analysing kataegis in the PCAWG cohort

In total, 9149 kataegis foci are spread genome-wide (no recurrent hotspots) over 1281 samples, with a median of four foci per sample (range 1–124) and a median of eight SNV per cluster (range 5–169). Figure 2.21 illustrates example kataegis events in fifteen samples. The vast majority of clusters have the distinctive APOBEC signature (91.4%, in 1175 samples), while just 790 clusters (8.6%, in 334 samples) have an alternative signature shown in Figure 2.22A. As previously observed, this non-APOBEC kataegis signature bears some resemblance to the polymerase η pattern (Alexandrov et al., 2013b; Nik-Zainal et al., 2016), but is by no means an exact recapitulation and may instead derive from one or more processes yet to be determined. Further investigation would need to apply signature decomposition methods (discussed in Chapter 4) to obtain detailed kataegis subdivisions by mutational process, following a similar logic to Supek and Lehner (2017).

The distribution of kataegis classes in each major histology group is shown in Figure 2.22B. Bladder transitional cell carcinomas have the highest average kataegis count per sample by a wide margin, strongly biased towards APOBEC clusters *without* a nearby SV breakpoint. Squamous cell carcinomas (SCC) from all tissues show a similar predilection for high APOBEC kataegis independent of SVs. In contrast, sarcomas, which also have a particularly high APOBEC cluster rate, have a very strong connection between kataegis and SV breakpoint positions. Of the twelve samples with more than 50 kataegis foci, six are bladder cancers, three are SCC (two head, one lung), two are liposarcomas, and one is a breast cancer. Kataegis foci with the other (non-APOBEC) signature are mostly found in stomach, esophageal, and liver cancers.

^qIn rare cases where the median inter-SNV distance m on the chromosome was under 15 kb, the kataegis threshold was lowered from 1 kb to $\frac{m}{15}$, down to a lower bound of 100 bp.

^rExcluded hypermutator samples included 25 melanoma, 8 colorectal, 2 lung, and 4 other cancers.

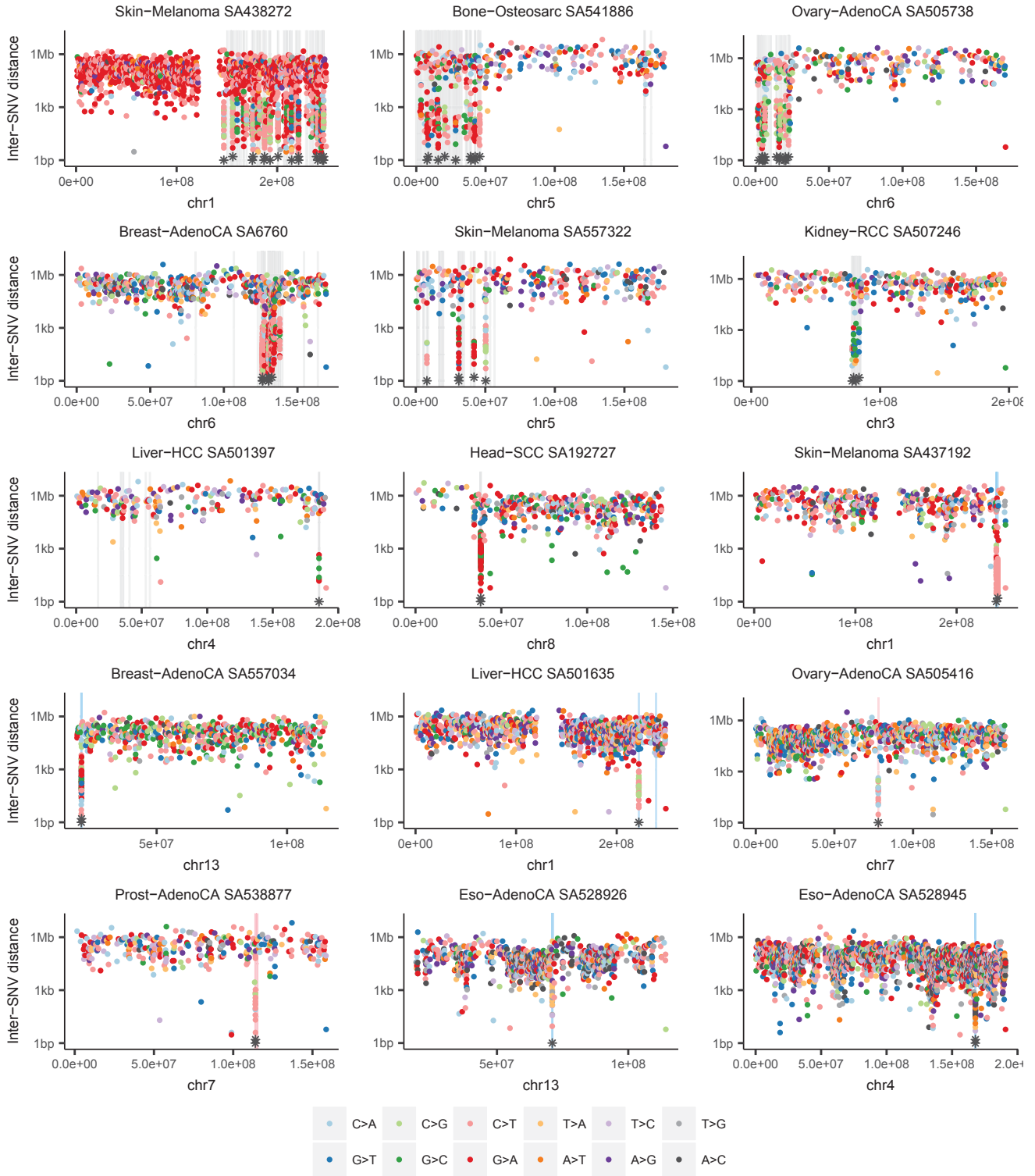


Figure 2.21: Fifteen chromosomes with identified kataegis regions, marked by dark gray stars along the lower edge. Rearrangement BPJ in associated events are marked by vertical lines in gray (complex sv), blue (deletion sv) or pink (other sv class; reciprocal inversion in the prostate example and unbalanced translocation in the ovary example).

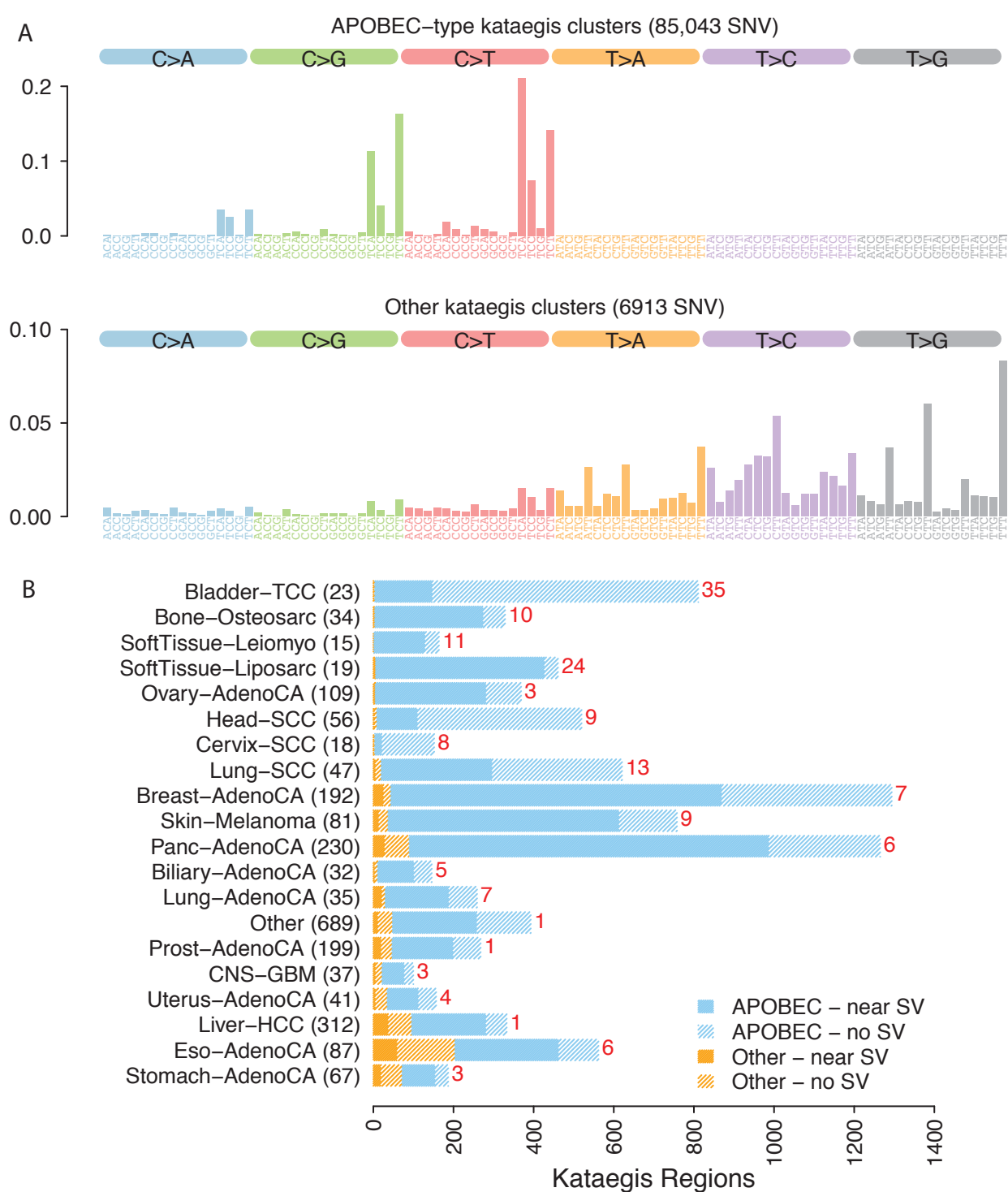


Figure 2.22: Kataegis distributions in the PCAWG cohort. (a) Somatic SNV distribution in a trinucleotide context around the pyrimidine reference base for two types of kataegis cluster: APOBEC type (mostly C>N in TCN), and other. (b) Number of kataegis regions in each histology group, shaded by SNV signature and proximity to SV breakpoint (within 50 kb or not), and sorted by proportion of APOBEC type clusters. The number of considered samples is indicated in parentheses (accounting for hypermutator exclusion), and the mean kataegis count per sample is annotated in red.

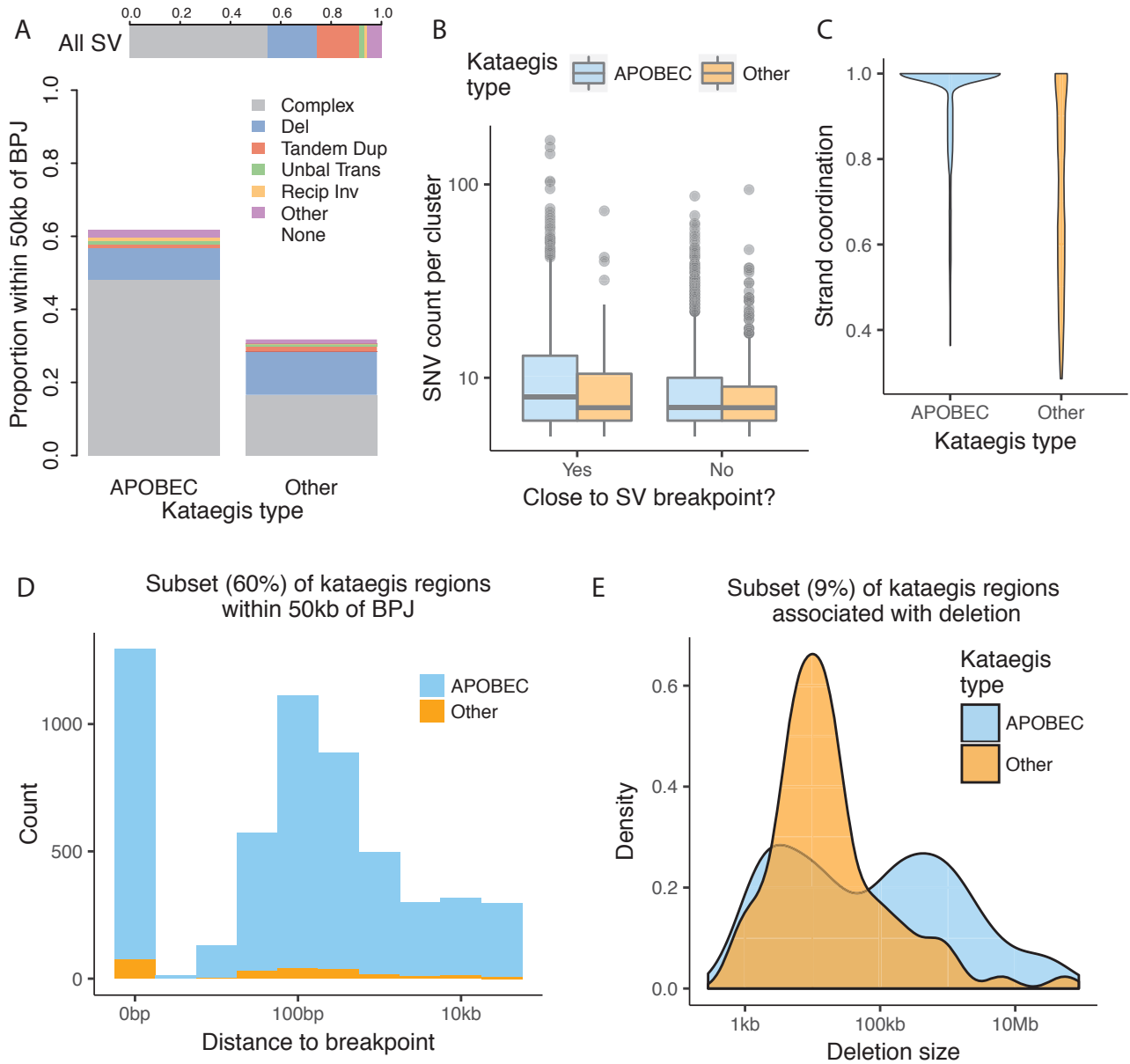


Figure 2.23: Kataegis properties in the PCAWG cohort. (a) Proportion of kataegis regions within 50kb of a SV breakpoint, shaded by SV class. The background distribution of all SV classes is indicated above. (b) SNV counts per kataegis cluster (\log_{10} scale). (c) Extent of strand coordination within kataegis clusters, measuring the maximum proportion of SNV from the same reference base. (d) Distance from kataegis region to SV breakpoint. (e) Size of deletions with associated kataegis.

Within a cut-off distance of 50 kb, 62% of APOBEC and 32% of other clusters are close to an identified SV breakpoint. The vast majority of these associations are very close indeed, usually well within 1 kb (Figure 2.23D). Most SV-associated APOBEC clusters are found around complex SV events (78%) or deletions of any size (14%), with a marked depletion around tandem duplications^s as shown in Figure 2.23A. Presumably, APOBEC enzymes mutate single stranded DNA exposed by resection at the DSB. The non-APOBEC kataegis regions are also found at complex SV events (52%), and have a specific bias towards small (< 100 kb) deletions (38%) (Figure 2.23A,E). These clusters are also set apart by their lack of strand-coordination (Figure 2.23C), indicating that single stranded DNA is not the major substrate for this alternative process. Supek and Lehner (2017) attribute most of these clusters to mismatch repair error, but that does not necessarily account for their frequent SV association. I conjecture that the small deletion preference may point to translesion polymerase restart of stalled replication forks, possibly coupled with error-prone mismatch repair.

Kataegis is notably absent from most tandem duplication, local 2-jump, and templated insertion events, despite the hypothesised role for MMBIR generating mechanisms known to expose single stranded DNA with a vulnerability to APOBEC mutagenesis (Sakofsky et al., 2014). Perhaps single strand protection (by RPA binding) is particularly efficient in these contexts, although complex template switching events consigned to the unexplained SV bin may yet be found to have a kataegis association.

For those thousands of kataegis foci with no associated SV event, the mutation clusters may mark sites of competent break repair, or APOBEC targeting of transcribed or lagging strand DNA, both of which are general—but not necessarily kataegis—APOBEC biases described by Haradhvala et al. (2016) and Morganella et al. (2016).

Visual inspection of SNV plots like those in Figure 2.21 reveals that my current method of kataegis calling occasionally misses some adjacent clusters, and so the analysis presented here slightly underestimates the kataegis burden, particularly around complex SV.

^sAlthough tandem duplications make up almost 17% of the total BPJ set, only 1.7% of SV-associated APOBEC clusters are near a tandem dup.

2.7 Discussion

In this chapter, I explored a novel SV classification scheme in a pan-cancer WGS dataset of 2559 samples, and presented a census of somatic rearrangement classes and their structural properties.

Although the PCAWG consortium strived to ensure the reliable quality of all sequencing data and variant calling, no orthogonal validation could be meaningfully applied to the somatic SVs. Consequently, the sensitivity and specificity of the BPJ callset is unclear. Data visualisation and CN concordance suggest the data is optimised for high specificity; however, it is practical to assume a small fraction are false positives from germline polymorphism or mapping/sequencing artefacts. For example, two prostate samples had unusually small deletion calls with atypically low evidentiary support (Section 2.4.1), and their inconsistency with the dataset at large suggest possible false positive contamination. It is also reasonable to assume a false negative rate of at least 5%, as short read WGS data cannot reliably map to approximately that fraction of the genome, even without counting centromeres and telomeres (Section 3.1.1). Although all samples were processed with the same bioinformatics pipeline, the underlying differences in sequencing centre, platform iteration, depth, and library insert size will inevitably impart some variant detection bias across the sub-cohorts by cancer type. All results should be interpreted in the context of these potential data quality caveats.

The task of BPJ clustering and classification fell chiefly to my colleague, Yilong Li. In collaboration, we developed the scheme outlined in Sections 2.1.3 and 2.2.2, and classified about 45% of all BPJ in the cohort. Alongside the traditional classes of deletion, tandem duplication, inversion, and translocation, we formalised a variety of medium-complexity SV structures in the cancer genome for the first time, including local 2-jumps and templated insertions. BPJ classification in highly convoluted cancer genomes is a difficult task, confounded by complex and overlapping SV events with ambiguous phasing. Even clean BPJ calls can have more than one plausible interpretation. For example, the relatively simple SV pattern in the osteosarcoma shown in Figure D.1 (second row, first column) was classified as a reciprocal inversion overlapping a prior tandem duplication, but is equally consistent with a templated insertion bridge on one chromosome. To present this diverse array of somatic SV events, I developed a novel plotting method in use throughout this thesis. Arguably, the modular layout and leveraging of clean CN segments provides a more

interpretable visualisation of complex structures than most existing approaches, particularly the ubiquitous ‘circos’ plot.

Most BPJ clusters were too large and/or cryptic to be interpreted against a library of simple SV overlaps, and a preliminary exploration of these complex unexplained clusters is deferred to Chapter 5. Other rearrangement phenomena excluded from this census were aneuploidy, SV on chrY, retrotransposition (analysed separately by Rodriguez-Martin et al. (2017)), mitochondrial insertions (Yuan et al. (2017)), and telomere length (Sieverling et al. (2017)).

Careful SV classification facilitated downstream analysis of properties and prevalence, without confounding from heterogeneous structures. Deletion and tandem duplication were by far the most common simple SVs, together accounting for about 80% of all classified BPJ in the cohort. Among the other SV classes, the extent of templated insertion was a revelatory finding, accounting for just over 5% of all classified BPJ across the three variant structures of chain, cycle, and bridge that re-route the genome through as many as eight distant loci, possibly via a MMBIR template switching mechanism.

The multi-modality of SV size distributions presumably reflects structural attributes about TAD size, resection rates, replication fork dynamics, strand invasion search, D-loop migration, and other unknown factors. The tendency of individual samples to incur events within the same characteristic size range suggests distinct underlying mechanisms have differential activity across samples and tissues, depending on the nature of DNA injury and subsequent repair.

Microhomology analysis implicated some level of MH-mediated repair in all SV classes except reciprocal translocation, with (mostly minor) variation between samples and cancer types. Even in the SV classes with the most MH-enrichment, about 40–50% of BPJ had no reported homology. This may indicate that repair mechanisms in cancer are less reliant on MH matching than previously expected, or reflect the failure of SV callers to estimate junction homology in the presence of non-templated base insertions and/or a few mismatching bases. Unfortunately, base insertion estimates could not be consolidated across the four SV callers, and the only reported values (from SvABA) were often inconsistent with the consensus break set, and ultimately too difficult to include.

The connection between APOBEC kataegis clusters and rearrangement breakpoints was confirmed for deletion and complex events, but was rarely observed for any other SV class. I also described a non-APOBEC kataegis signature with a striking preference for small deletion in stomach, esophagus, and liver.

Significant scope remains for further structural analysis of genome rearrangement in the PCAWG cohort. Besides extending and refining the BPJ classification procedure, further work could: quantify and improve the concordance between BPJ calls and CN estimates; identify regions of longer and imperfect junction homology; and explore SV-connected LOH as previously described for germline local 2-jumps (Carvalho et al., 2015).

