Chapter 5

Complex rearrangement events

Complex SV events spanning tens to hundreds of BPJ are a common feature in the cancer rearrangement landscape. The various complex phenomena—reviewed in Section 1.4.2—include chromothripsis (Stephens et al., 2009), chromoplexy (Berger et al., 2011; Baca et al., 2013), extrachromosomal double minutes (Cox et al., 1965; Turner et al., 2017), breakage-fusion-bridge cycles (McClintock, 1941; Greenman et al., 2016), and chromoanasynthesis (Liu et al., 2011; Meier et al., 2014). As described in Section 2.1.3, Yilong Li's classification of SV in the PCAWG dataset focused on (relatively) simple rearrangement structures involving a small handful of BPJ at most. This classification scheme left 151,212 BPJ from 1889 samples in complex unexplained clusters. In this chapter, I embark on a preliminary attempt to meaningfully partition and describe these complex rearrangements, and propose strategies for further investigation in future projects.

5.1 Clustering complex unexplained breakpoint junctions

All BPJ in the PCAWG dataset were previously clustered by the original SV classification pipeline described by Li et al. (2017). However, these existing BPJ clusters are a poor starting point for comprehensive analysis of the complex SV landscape for several reasons. First, the original BPJ clustering method was optimized to extract and explain the non-complex structures, and was never refined to generate distinct and classifiable complex clusters. Second, the original method demarcated cluster boundaries solely based on the immediate adjacency distance between breakpoints on the same chromosome, and did not consider additional information about breakpoint groups neighbouring at multiple distant loci. Third, two complex SV structures would be joined in the same cluster with as little as one BPJ spanning between them, even if each side was a large interconnected "hairball" of dozens of BPJ with no other external connections. Finally, one known oversight of the original algorithm left some BPJ together in the same cluster even after the linking BPJ that joined them were siphoned out as classifiable sub-structures.

Given these problems with the existing cluster breakdown of the complex unexplained BPJ, I set out to develop a new clustering algorithm as follows.

5.1.1 New BPJ clustering method

For each sample in the PCAWG cohort, I considered the set of 'complex' BPJ left unexplained by the original SV classification scheme. Then, I grouped the breakpoints into primary local footprints by placing a partition between adjacent (on same reference chromosome) breakpoints if the distance between them was greater than some sample-specific threshold (and requiring double the threshold before separating any pair of adjacent breakpoints belonging to the same BPJ).

To choose the sample-specific footprint partition threshold, I fitted a mixture of two gamma distributions to the collection of inter-break distances on a $\log_{10}$ scale, and calculated the 0.95 quantile of the lower gamma component, subject to the following caveats:

- the footprint cut-point was constrained to a minimum of 40 kb and a maximum of 4 Mb, and
- if the sample had fewer than 20 inter-break distances, the cut-point defaulted to 1 Mb.

By fitting a mixture of two gamma distributions, I aimed to quantify the expected inter-break distances between positions which are and are not mechanistically linked, with a cut-point chosen to keep related positions in the same footprint 95% of the time. Figure 5.1 illustrates the gamma fit and cut-point choice for 64 randomly chosen samples. The variation across samples suggests that this approach will work better for some samples than for others, and will not pick the ideal initial footprint grouping in all cases.

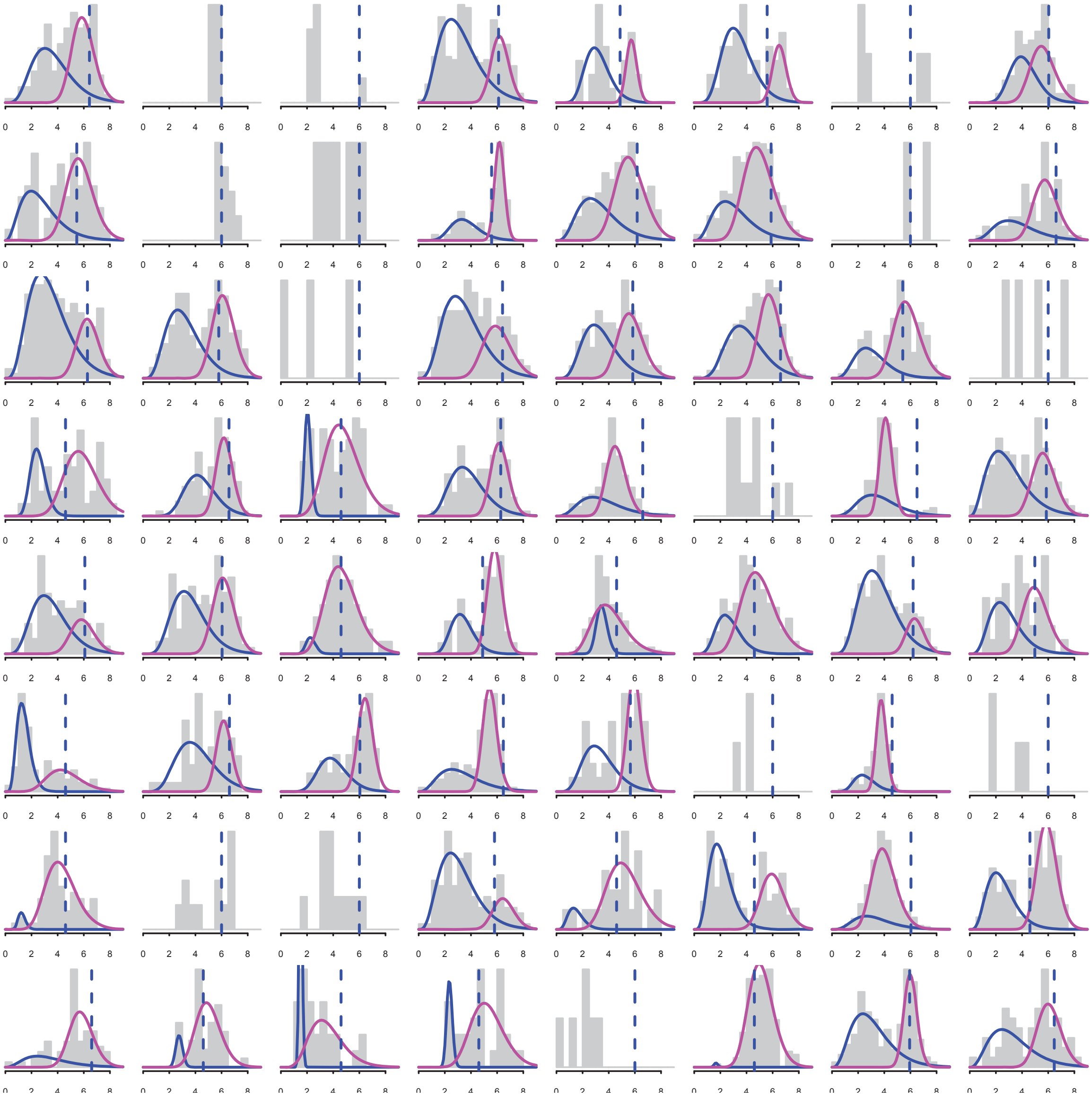


Figure 5.1: The distribution (shown in grey histogram on a $\log_{10}$ scale) of inter-break distances between adjacent (on same reference chromosome) positions of complex unexplained BPJ in 64 randomly chosen PCAWG samples. For samples with 20 or more inter-break distances, the primary footprint partition cut-point (blue dashed line) is placed at the 0.95 quantile of the lower component in a two gamma mixture (constrained to minimum 40 kb and maximum 4 Mb). For samples with few inter-break distances, the cut-point is fixed at 1 Mb.

As a final refinement to the primary footprint definition, any footprint larger than 1 Mb with at least two breakpoints on either side of a gap spanning $> 70\%$ of the footprint region was then split apart in the gap.

I then proceeded to represent the complex SV network in a sample with a weighted, undirected, node-edge graph. Each node is a primary footprint region with a size attribute representing the number of contained breakpoints. Each edge represents the BPJ with a side in each node, with a weight attribute representing the number of connecting BPJ. The disjoint (unconnected) components in the node-edge graph provide the initial candidates for a BPJ cluster division.

Next, I aimed to reduce under-clustering by grouping graph components with several nodes adjacent in genome space. Two candidate BPJ clusters were merged if:

- any two "foldback" type footprints were within 5 Mb of each other[a],
- four unique footprints were within 8 Mb of a footprint from the other cluster (either $2 * (1 \leftrightarrow 1)$, $(1 \leftrightarrow 3)$ or $(2 \leftrightarrow 2)$ arrangement),
- five unique footprints were within 12 Mb of a footprint from the other cluster (either $(1 \leftrightarrow 2)/(1 \leftrightarrow 1)$, $(1 \leftrightarrow 4)$ or $(2 \leftrightarrow 3)$ arrangement),
- six unique footprints were within 16 Mb of a footprint from the other cluster (either $3{*}(1 \leftrightarrow 1)$, $2{*}(1 \leftrightarrow 2)$, $(1 \leftrightarrow 3)/(1 \leftrightarrow 1)$, $(2 \leftrightarrow 2)/(1 \leftrightarrow 1)$, $(1 \leftrightarrow 5)$, $(2 \leftrightarrow 4)$ or $(3 \leftrightarrow 3)$ arrangement), and
- if *and only if* a cluster had just one or two nodes, three footprints were within 4 Mb of a footprint from the other cluster ($(1 \leftrightarrow 2)$ arrangement).

After every merge, the resulting cluster was compared against the sample's current BPJ cluster set to check for subsequent merges now meeting the criteria.

One final part of the cluster merging stage aimed to capture cycles of multiple graph components that cannot be captured through simple pairwise cluster comparison. To look for cycles, I considered any small BPJ clusters of 2–4 footprint nodes, and merged any maximal subset of these clusters for which:

- there were at least two footprints in each cluster within 15 Mb of another cluster in the subset, and
- each cluster was within 15 Mb of at least two footprints from another cluster in the subset (subtle distinction from the first criterion).

[a] Foldback-type footprints defined as those solely comprised of one or two (non-overlapping) foldback-type BPJ, i.e. $\langle ++ \rangle$ or $\langle -- \rangle$.

Figure 5.2 illustrates the graph component and merging steps for four samples.

As the last step in the BPJ clustering algorithm, I aimed to reduce over-clustering by separating out distinct graph communities within large candidate clusters. For any candidate cluster involving 15 or more BPJ (≥ 30 breakpoints), I first defined larger secondary footprints to construct a new node-edge graph representation. For a cluster with b breakpoints, local footprints were partitioned in gaps larger than some threshold t_M in megabase units such that

$$t_M = 10 - 6 \times \frac{\min(b, 1500) - 30}{1500 - 30}\,.$$

This set the partition threshold on a sliding scale between 4 Mb for clusters involving ≥ 1500 breakpoints and 10 Mb for clusters involving 30 breakpoints.

Using these new footprint definitions to define the nodes, and using a double weighting on any intrachromosomal BPJ edges between nodes, I identified candidate sub-clusters using the "walktrap" community detection algorithm with s steps where

$$s = 7 + \left\lfloor 14 \times \frac{\min(b, 1500) - 30}{1500 - 30} \right\rceil\,.$$

This walktrap algorithm (Pons and Latapy, 2006) finds sub-graph community structures using short random walks along graph edges (accounting for edge weights) to measure the distance between nodes. Considering this community division, I separated a sub-graph into a new BPJ cluster if:

- it had at least eight breakpoints; and
- less than 12.5% of breakpoints (up to a maximum of six) were connected to a BPJ leading outside the sub-graph (double-counting any intrachromosomal BPJ).

Figures 5.3–5.5 illustrate several examples, with full event plots in Figure D.19.

If the walktrap algorithm returned more than four candidate sub-graphs and less than a quarter of these met the criteria for separation, I then tried to agglomerate the sub-graphs and reassess the separation criteria (example Figure 5.6). I also checked whether sub-graph removal isolated any other sub-graph into its own disjoint component. Finally, any BPJ spanning two separated clusters was assigned to the smaller of the two.

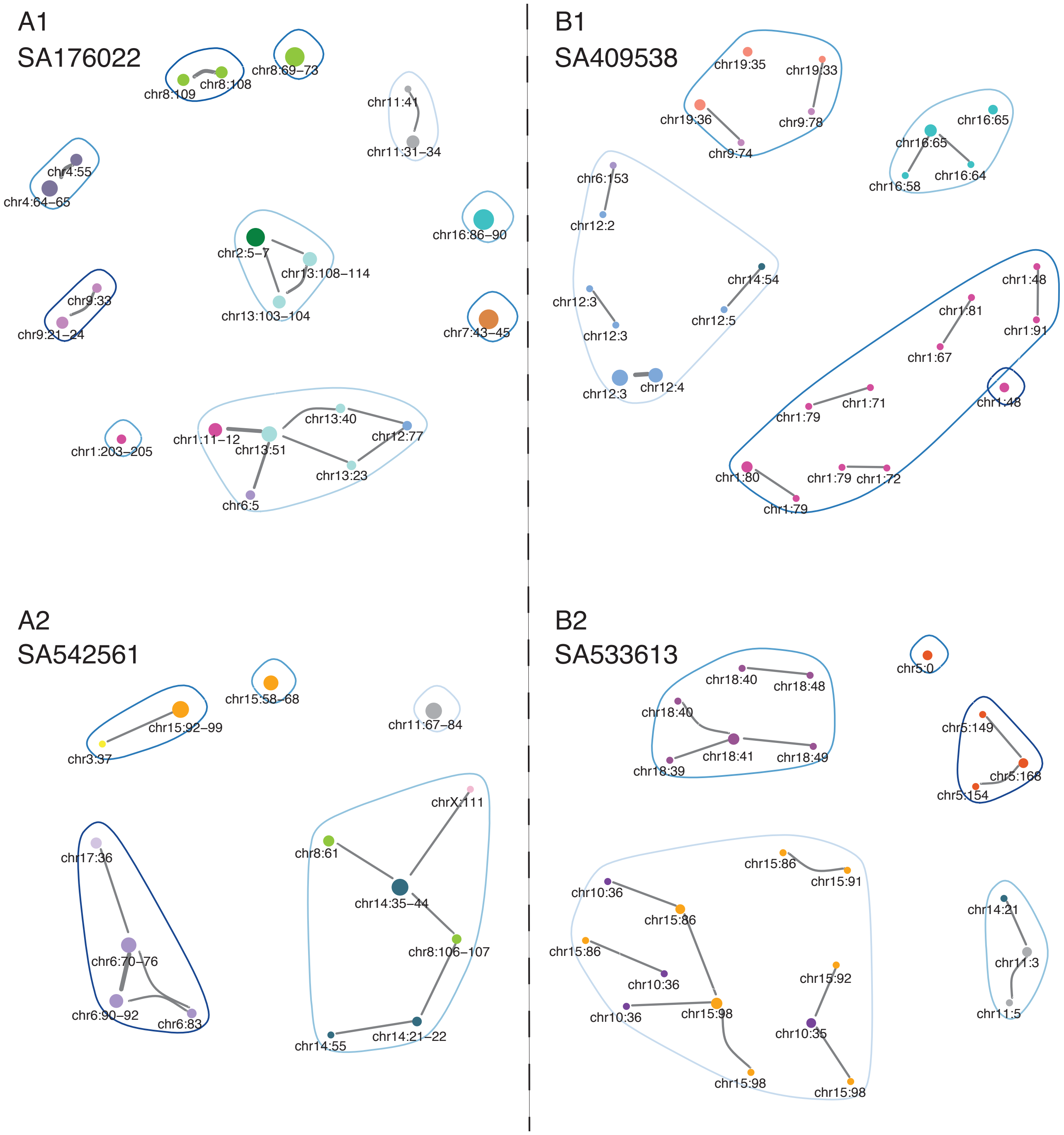


Figure 5.2: Node-edge graph representation of the complex unexplained BPJ in four PCAWG samples. Each node is a genome footprint, coloured by reference chromosome with size corresponding to breakpoint count. Node labels indicate the chromosome position in megabase units. Each edge indicates breakpoint junctions between footprints, with edge weights corresponding to the number of linking BPJ. In side (A), none of the initial disjoint graph components are merged any further. In side (B), the blue circles indicate graph components merged into the same final BPJ cluster.

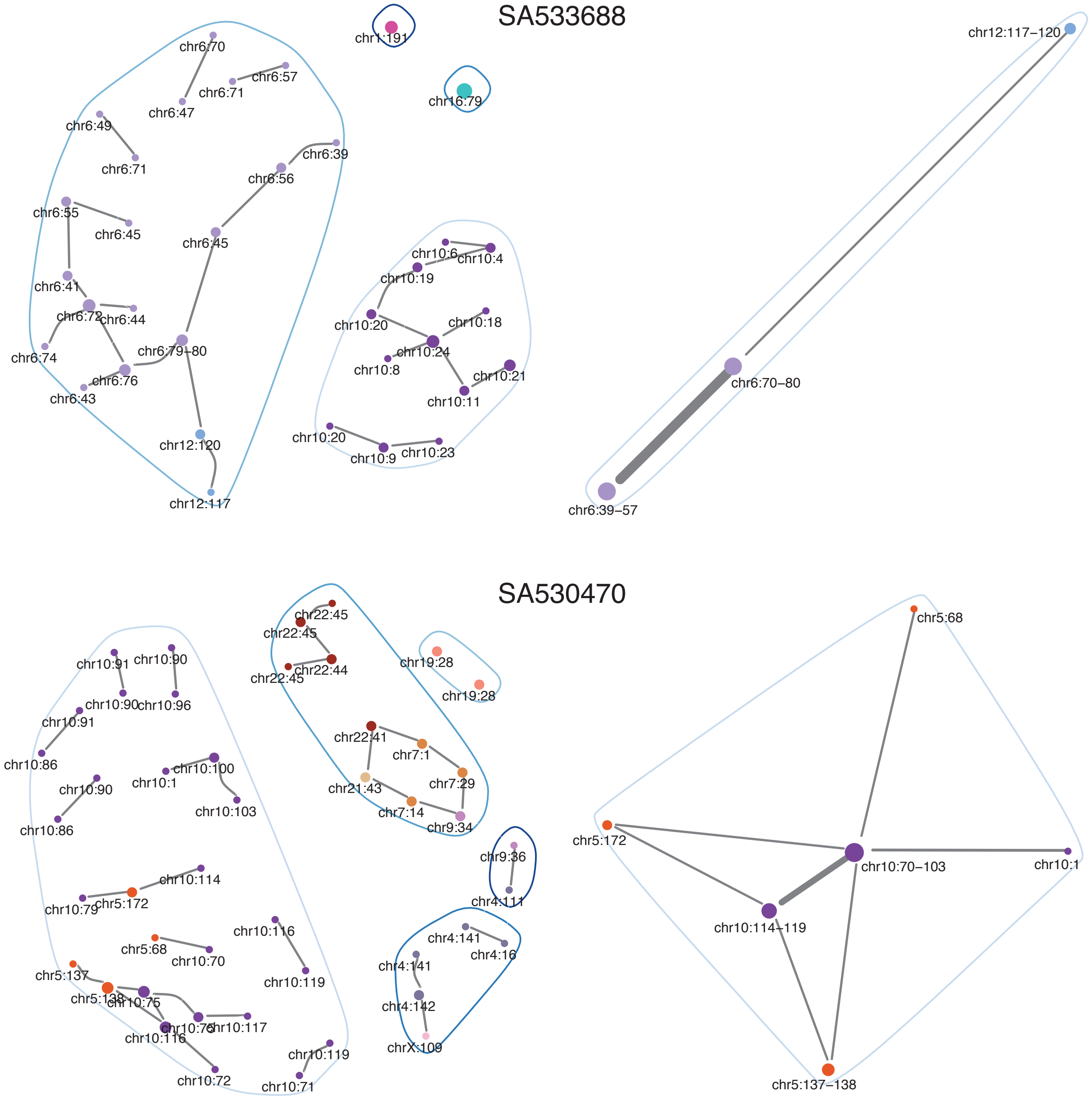


Figure 5.3: Two samples containing large BPJ clusters with no separable sub-graphs. The left side graphs show all complex BPJ in each sample. The right side graphs show the secondary footprint partition of the large candidate cluster, with blue circles indicating that the walktrap community detection algorithm finds no significant sub-graph structures. The large candidate groups are accepted as the final BPJ clusters.

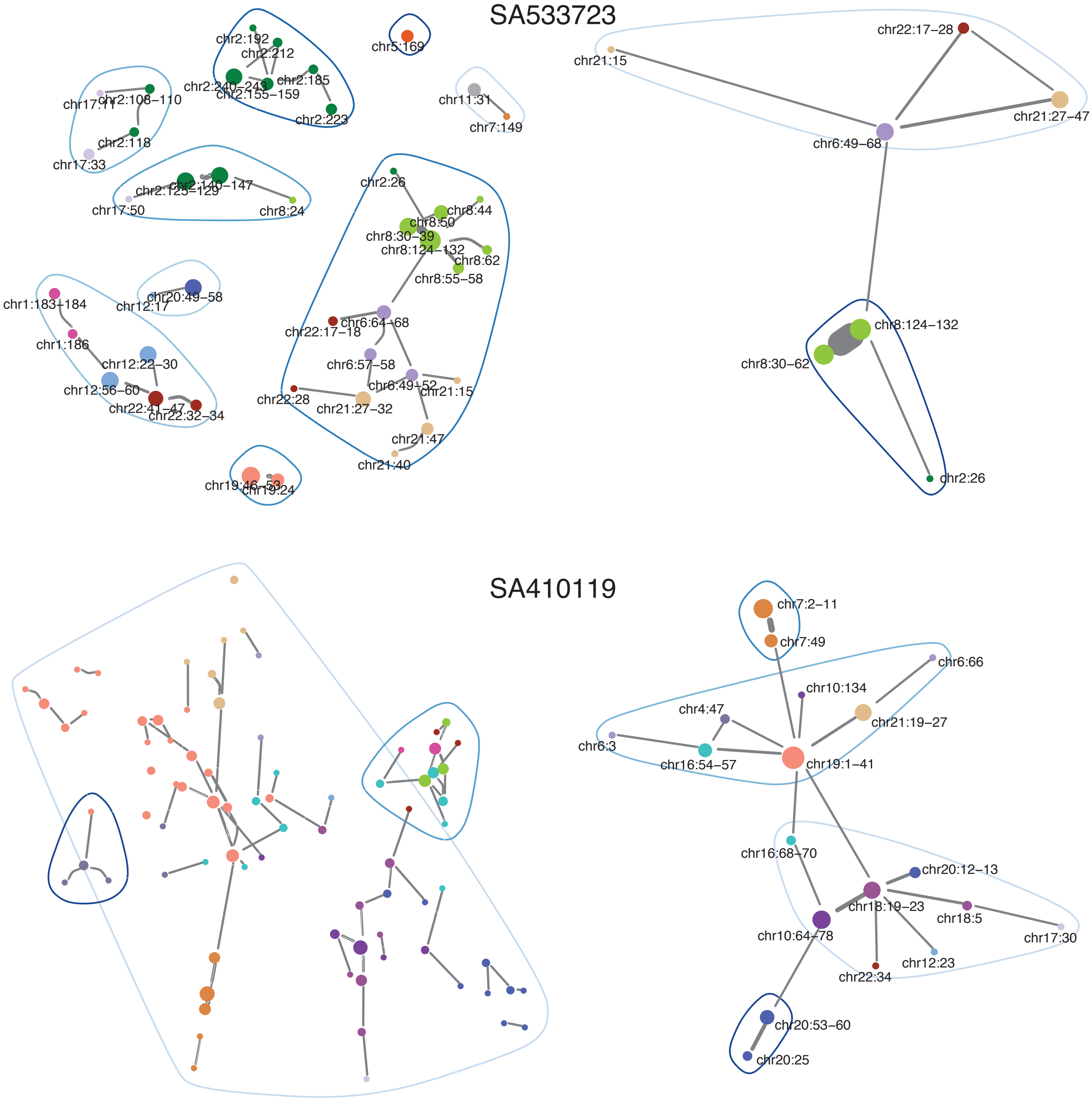


Figure 5.4: Two samples containing large BPJ candidate clusters with fully separable sub-graphs. The left side graphs show all complex BPJ in each sample. The right side graphs show the secondary footprint partition of the large candidate cluster, with blue circles indicating sub-graphs found by walktrap community detection. All sub-graphs meet the criteria for separation into different BPJ clusters. Full BPJ plots are available in Figure D.19.

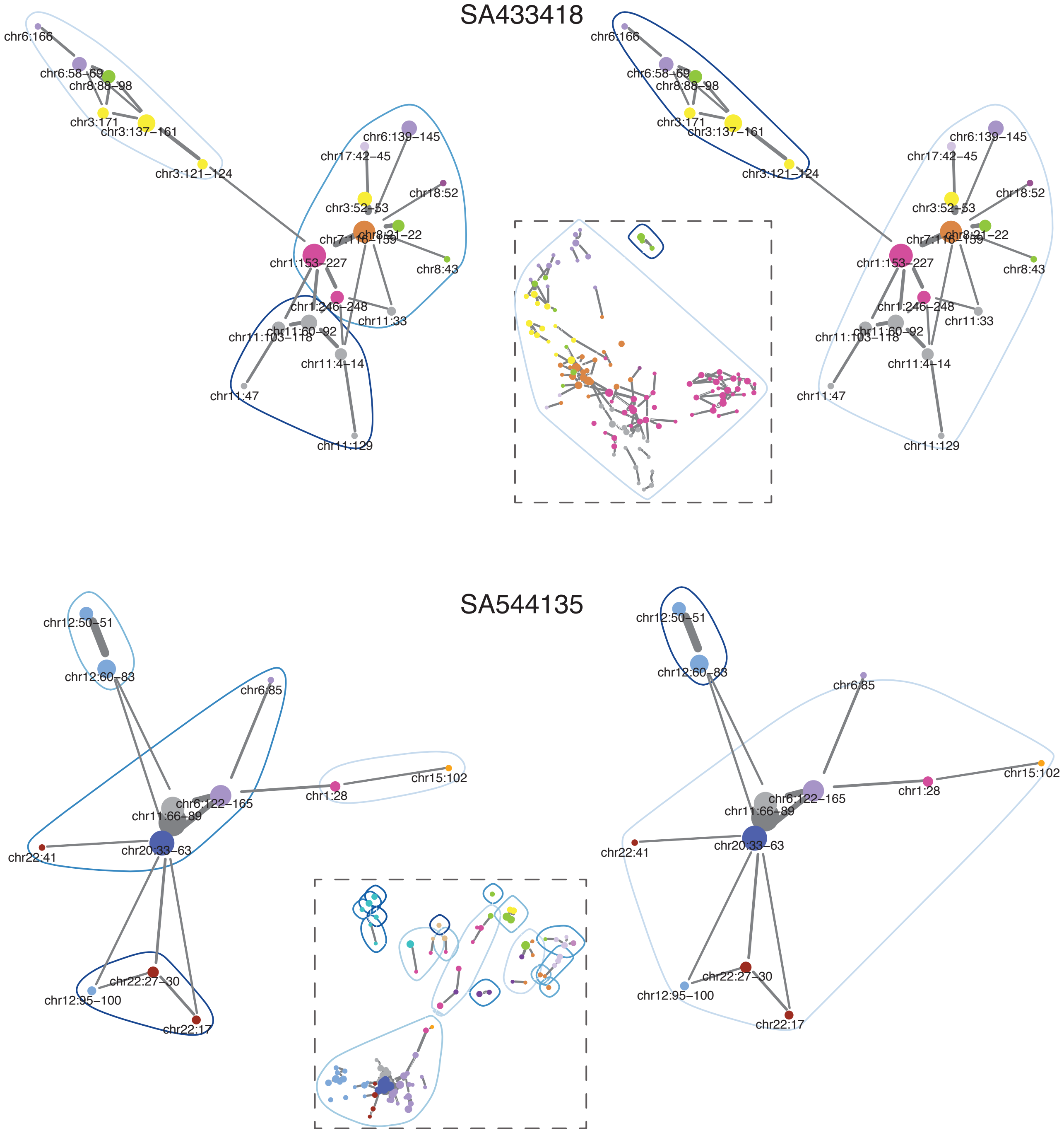


Figure 5.5: Two samples containing large BPJ candidate clusters with partially separable sub-graphs. The left side graphs show the secondary footprint partition of the candidate cluster, with blue circles indicating sub-graphs found by walktrap community detection. In each case, only one sub-graph meets the criteria for separation into a different BPJ cluster, with the final cluster allocation indicated in the right side graphs. The inset boxes show all complex BPJ in the sample for context. Full BPJ plots are available in Figure D.19.

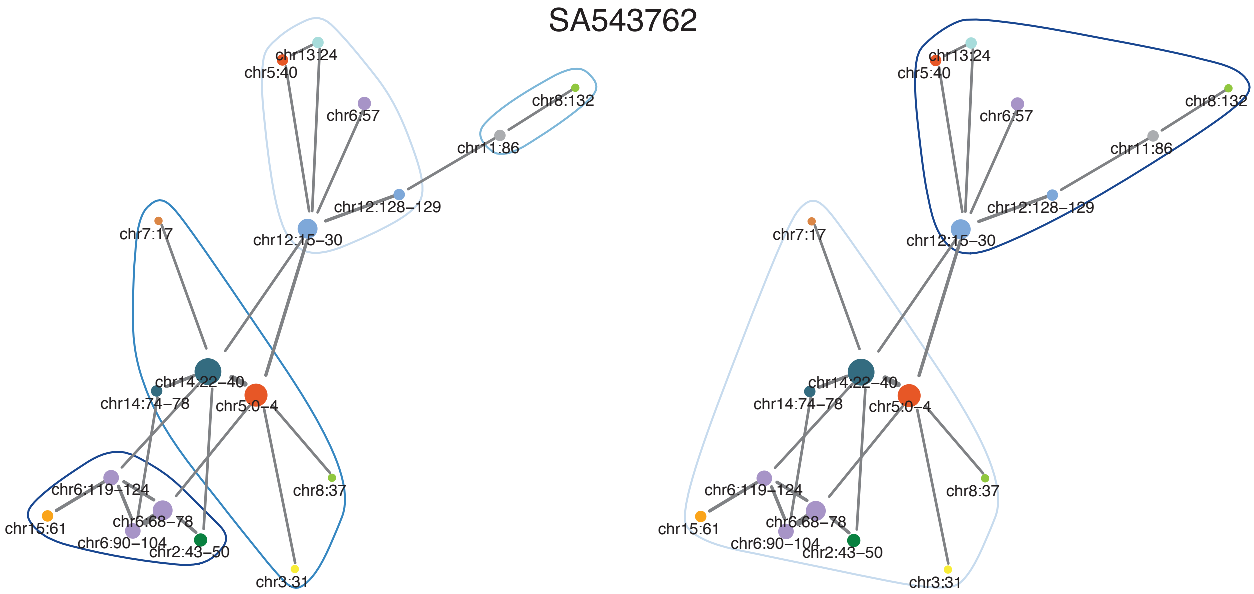


Figure 5.6: A large candidate BPJ cluster with separable sub-graphs following extra agglomeration. The left node-edge graph shows the secondary footprint partition, with blue circles indicating community sub-graphs. In this case, none of the four initial sub-graphs meet the separation criteria. Following extra agglomeration into two sub-graphs shown in the right side plot, the separation criteria are now met and the final allocation divides the SV into two clusters. The full BPJ plot is available in Figure D.19.

5.1.2 Comparison between old and new BPJ clustering

Of the 1889 PCAWG samples with complex unexplained BPJ, 78 samples have all BPJ assigned to tiny clusters of one or two BPJ in the new clustering scheme (summarised in Section 5.2). Of the remaining samples, 582 have exactly the same cluster breakdown as the old method, and a further 455 have the same cluster breakdown if BPJ now allocated to tiny clusters are disregarded. This leaves 774 samples with a different cluster breakdown by the old and new methods (Figure 5.7), including 555 samples with *more* clusters in the new scheme and 219 samples with *fewer* clusters in the new scheme. As summarised in Table 5.1, the samples with different cluster divisions tend to be those with greater rearrangement burdens.

Figures D.20–D.26 illustrate the new and old cluster divisions in a range of samples with either a greater or lesser degree of cluster separation with my novel method outlined in Section 5.1.1. In particular, the extreme outlying melanoma sample with more than 60 clusters in the old scheme and fewer than 10 clusters in the new scheme is included in Figure D.26. Although the old partition appears to over-split these melanoma rearrangements, the massive

Table 5.1: Number of samples (n) with the same or different complex BPJ cluster divisions by the old and new methods. The total number of complex BPJ and new-scheme clusters per sample are summarised by the median, minimum and maximum. The samples with the most junctions (J) and clusters (C) are listed for each group. Samples with 'nearly' the same cluster breakdown differ only by the separation of tiny clusters of one or two BPJ.

	n	total BPJ	total clust.	max BPJ	max clust.
all 'complex' BPJ in tiny clusters	78	2 (2–8)	1 (1–5)	SA515309, 8J in 5C	see left
exactly the same cluster breakdown	582	14 (3–1183)	2 (1–27)	SA554721, 1183J in 7C	SA54378, 242J in 27C
nearly the same cluster breakdown	455	26 (3–1387)	3 (1–21)	SA236844, 1387J in 2C	SA541880, 168J in 21C
different cluster breakdown	774	80 (8–1954)	6 (1–32)	SA554739, 1954J in 6C (11C before)	SA440859, 949J in 32C (26C before)

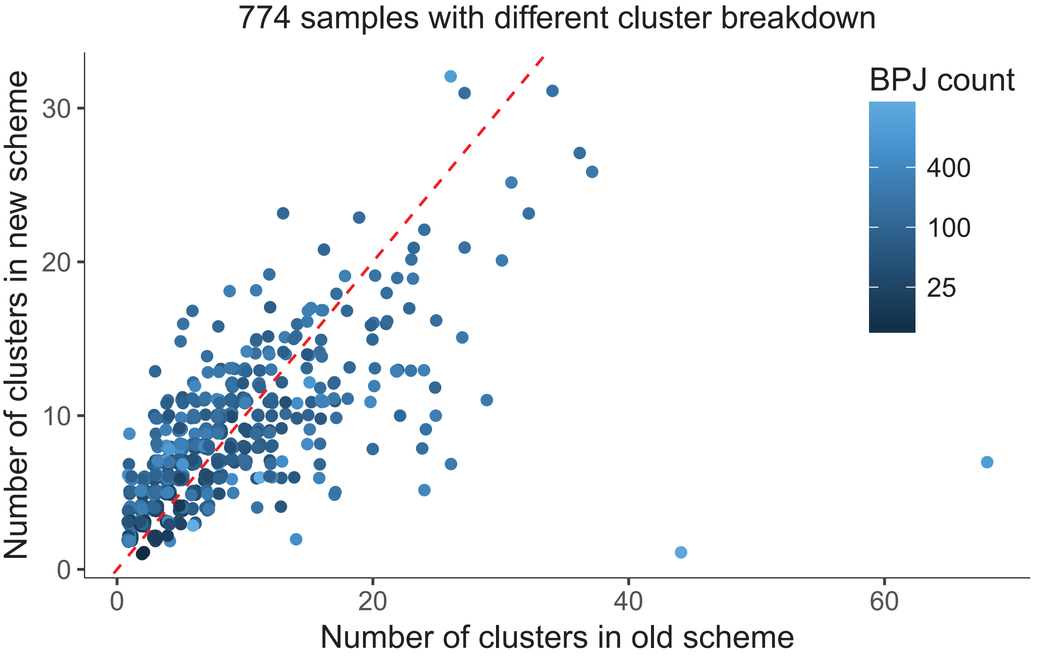


Figure 5.7: Discrepancy in complex unexplained BPJ cluster counts between new and old schemes for 774 PCAWG samples. Red dashed line separates samples with more clusters in new scheme (top left) from those with fewer clusters in new scheme (bottom right).

cluster from my new scheme may be failing to separate distinct sub-structures. In future work, it would be helpful to define objective summary statistics to quantify the fit of different BPJ cluster partitions. From manual inspection of these examples (and dozens more not shown), I conclude that my current partitions are a more logical division of the BPJ terrain than the pre-existing clusters. In many cases, this improvement is due to known oversights in the previous algorithm which left BPJ in the same cluster even after their connecting SVs were separated out. Despite this progress, many samples may yet have poor clustering results, and substantial opportunities remain for further development of BPJ clustering algorithms, ideally accompanied by more formal statistics for performance comparison.

5.2 Tiny unexplained BPJ clusters

Of the 151,212 complex unexplained BPJ, 6964 (4.6%) are separated into tiny clusters of one or two BPJ by the method described in Section 5.1.1. Some of the two-BPJ clusters are the same as those generated by Yilong Li (Section 2.1.3), in combinations unaccounted for by the existing classification scheme.

As summarised in Table 5.2, these BPJ are configured in a variety of known and unknown structural forms. The majority of single BPJs newly separated from larger complex clusters are unbalanced translocations (978) and foldback SVs (869). Of the recovered BPJ pairs with familiar structures, 270 junctions are in reciprocal inversions, 78 in reciprocal translocations, 544 in local 2-jumps, and 232 in templated insertion chains, cycles or bridges. Additionally, I identified a new SV class and termed it templated insertion mediated foldback (198 observations). This novel structure is characterised by the 'insertion' fragment ([−+] motif) mediating an overall rearrangement of foldback in another locus ([++] or [−−] motif). For the BPJ pairs with other, unclassified configurations, the majority involve foldback-type BPJ intersecting or adjoining another junction with uncertain derivative structure (possibly involving chance proximity of unphased events on separate homologous chromosomes). The remaining small proportion of unexplained pairs are simple overlaps of deletion, tandem duplication and/or translocation.

Table 5.2: Isolated BPJs (singles and pairs) unexplained by initial classification.

SV class	Sub-group	Definition	BPJ
Deletion	-	local $\langle +- \rangle$ BPJ	236
Tandem Dup	-	local $\langle -+ \rangle$ BPJ	179
Foldback	-	$\langle ++ \rangle$ or $\langle -- \rangle$ BPJ	869
Unbal Trans	-	distant BPJ	978
Recip Inv	-	interlocked $\langle ++ \rangle / \langle -- \rangle$ BPJ pair	270
Recip Trans	-	distant BPJ pair, $[+-]$ motifs	78
Foldback Pair	-	adjacent inverting BPJ, same orientation	180
Local 2-Jump	Dup-InvDup	interlocked $\langle -- \rangle / \langle ++ \rangle$ BPJ pair	182
	Loss-InvDup	nested $\langle ++ \rangle / \langle -- \rangle$ BPJ pair	232
	Dup-Trp-Dup	disjoint $\langle -- \rangle / \langle ++ \rangle$ BPJ pair	130
Local+ Distant 2-Jump	Trans w/ Foldback	distant BPJ adjoining $\langle ++ \rangle$ or $\langle -- \rangle$ BPJ w/ $[-+]$ motif	136
	Trans w/ InvIns	distant BPJ intersecting $\langle ++ \rangle$ or $\langle -- \rangle$ BPJ w/ $[-+]$ motif	138
Templated Insertion	Cycle	two $[-+]$ motifs	78
	Bridge	$[-+]$ and $[+-]$ motif	84
	Chain	$[-+]$ motif and two single breaks	70
	Foldback	$[-+]$ and $[++]$ *or* $[--]$ motif	198
Chromoplexy	Chain	$[+-]$ motif and two single breakpoints	60
Other Complex	Local	two other BPJ in local configuration	2124
	Distant	distant BPJ intersecting or adjoining other local BPJ	530
	Unphased	distant BPJ pair with $[++]$ or $[--]$ motifs	160
	Other	rare configurations	54

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

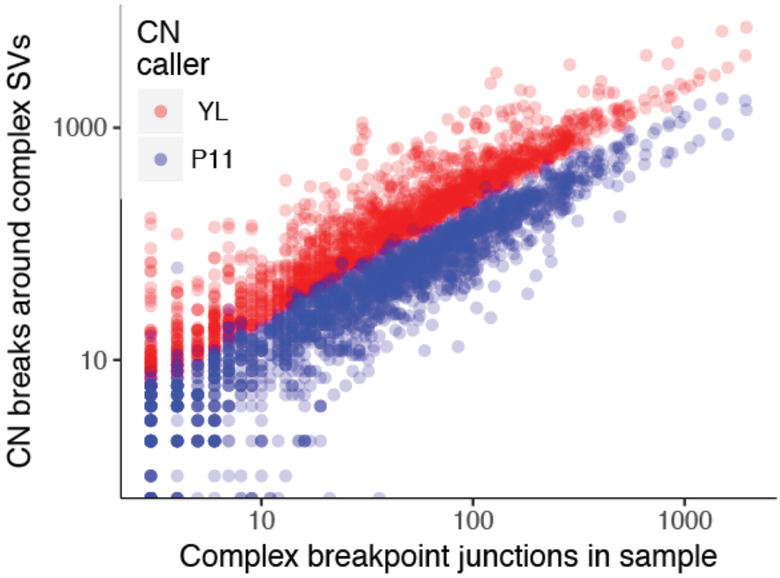
5.3 Matching complex SV with CN estimates

To describe complex SV clusters with more than two BPJ, the breakpoint calls must be considered in conjunction with the CN profile calculated from WGS read depth. As described in Section 2.1.2, most CN estimates used in this thesis are the YL calls with non-integer (sub-clonal) segmentation values. Upon inspection, these YL CN calls are unreliable in a minority of samples. Fortunately, Dentro et al. (2017) generated another set of CN estimates (the P11 calls) for the PCAWG cohort by calculating a consensus segmentation from several algorithms constrained by the simplifying assumption of integer (clonal) values. Prior to the characterisation and visualisation of the remaining complex unexplained SVs, I set out to determine which samples had sufficiently poor YL CN estimates as to necessitate a switch to the more conservative P11 estimates.

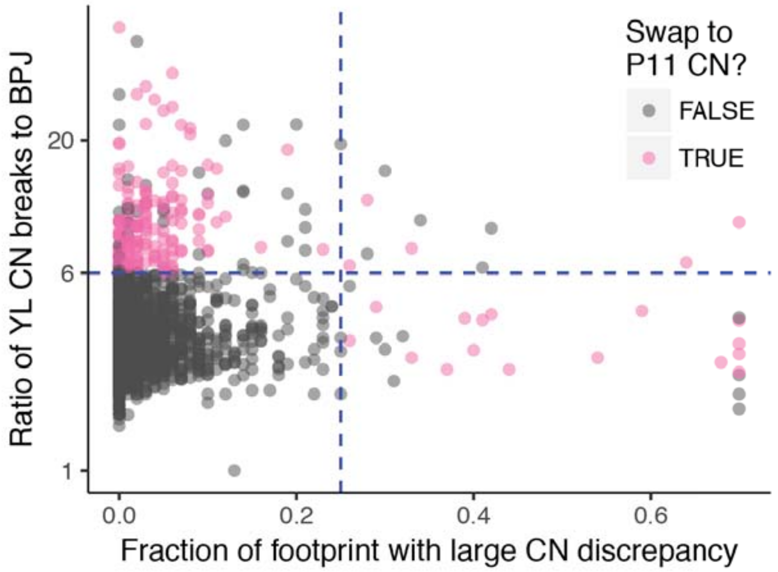
For the 1811 samples with complex unexplained BPJ (excluding tiny clusters from Section 5.2), I consider the CN profiles returned by YL and P11 in 1 Mb flanks around each breakpoint, leaving no gaps smaller than 5 Mb. I also round the non-integer YL calls to 0.05 intervals to disregard any minor change-points between very similar adjacent segments. As shown in Figure 5.8A, the YL CN segmentation around complex BPJ consistently involves many more change-points than the P11 calls. My criteria for switching a sample to P11 CN estimates are that:

- the YL CN has 6-fold more change-points than there are BPJ in the footprint of interest; *or*
- at least 25% of the footprint has a major CN discrepancy, defined as any region where $(Y + 0.4)/(P + 0.4)$ is either > 2.5 or < 0.4—that is, the two CN callers differ by more than 2.5-fold after adding a dummy value to disregard differences in the 0–1 CN range; *except*
- the CN estimates are *not* switched in samples where the number of P11 CN change-points is fewer than half the number of BPJ in the footprint *or* in cases where the P11 CN contains more than double the length of `NA` values over at least 10% of the total footprint.

With these criteria, I switched 174 samples (9.6%) to the integer P11 CN estimates for the remaining analyses in this chapter (Figure 5.8B). Figures 5.9 and D.28 provide a side-by-side comparison of the two CN call sets in five of these qualifying samples.



(a) Number of change-points in the CN calls around complex BPJ for each sample.



(b) Approximately 9% of samples are switched to P11 CN calls, in cases with excessive change-points in the YL set (vertical axis) or a large discrepancy in overall copy estimation (horizontal axis), barring a few exceptions as detailed in the text.

Figure 5.8: Comparison of YL and P11 copy number estimates around all complex unexplained BPJ in 1811 PCAWG samples (considering CN in 1 Mb flanks around each breakpoint, leaving no gaps smaller than 5 Mb).

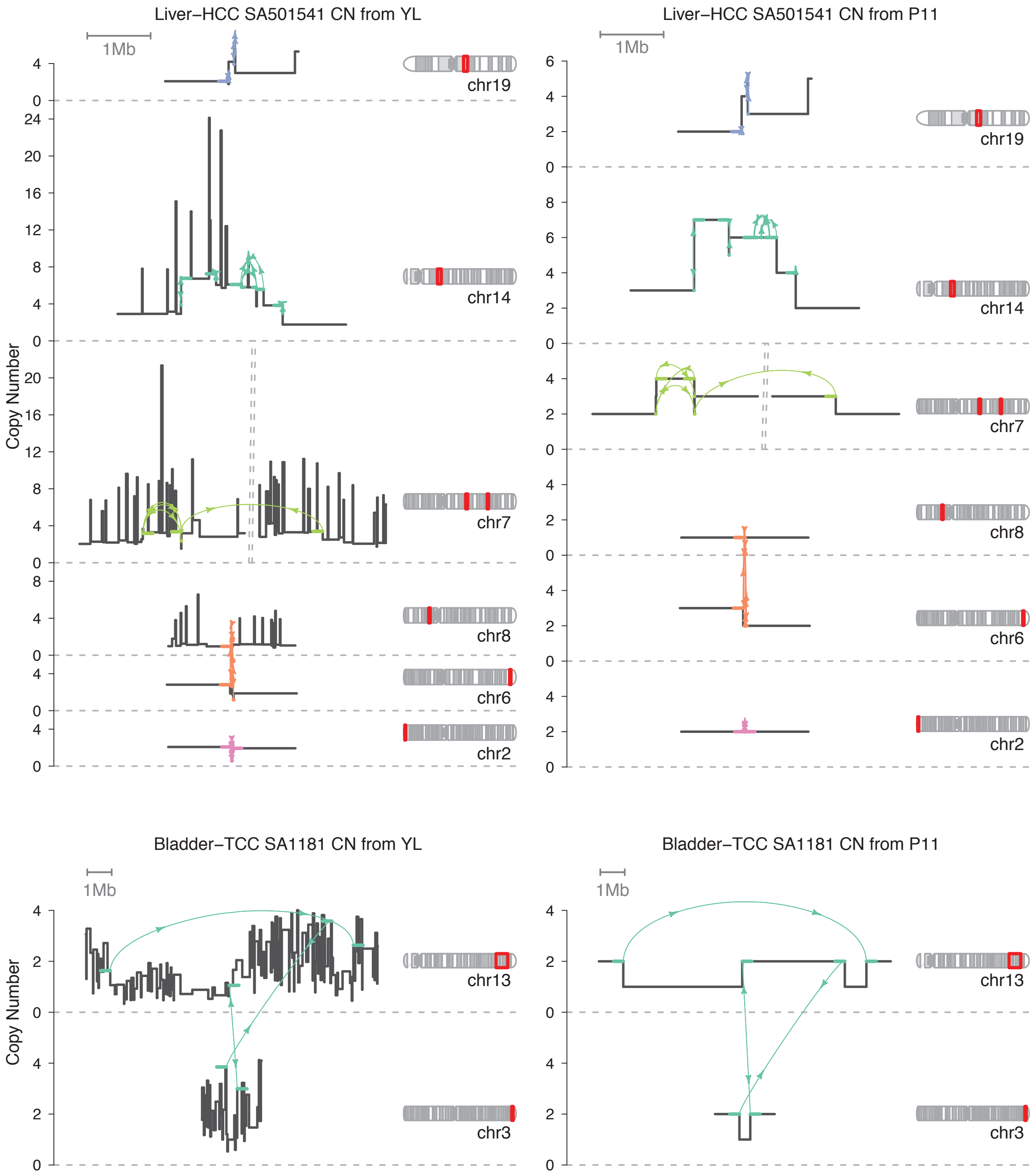


Figure 5.9: Copy number profiles returned by YL (left) and P11 (right) around complex unexplained BPJ in samples qualifying for a switch to P11 CN. Breakpoint junctions are coloured by cluster assignment within the sample.

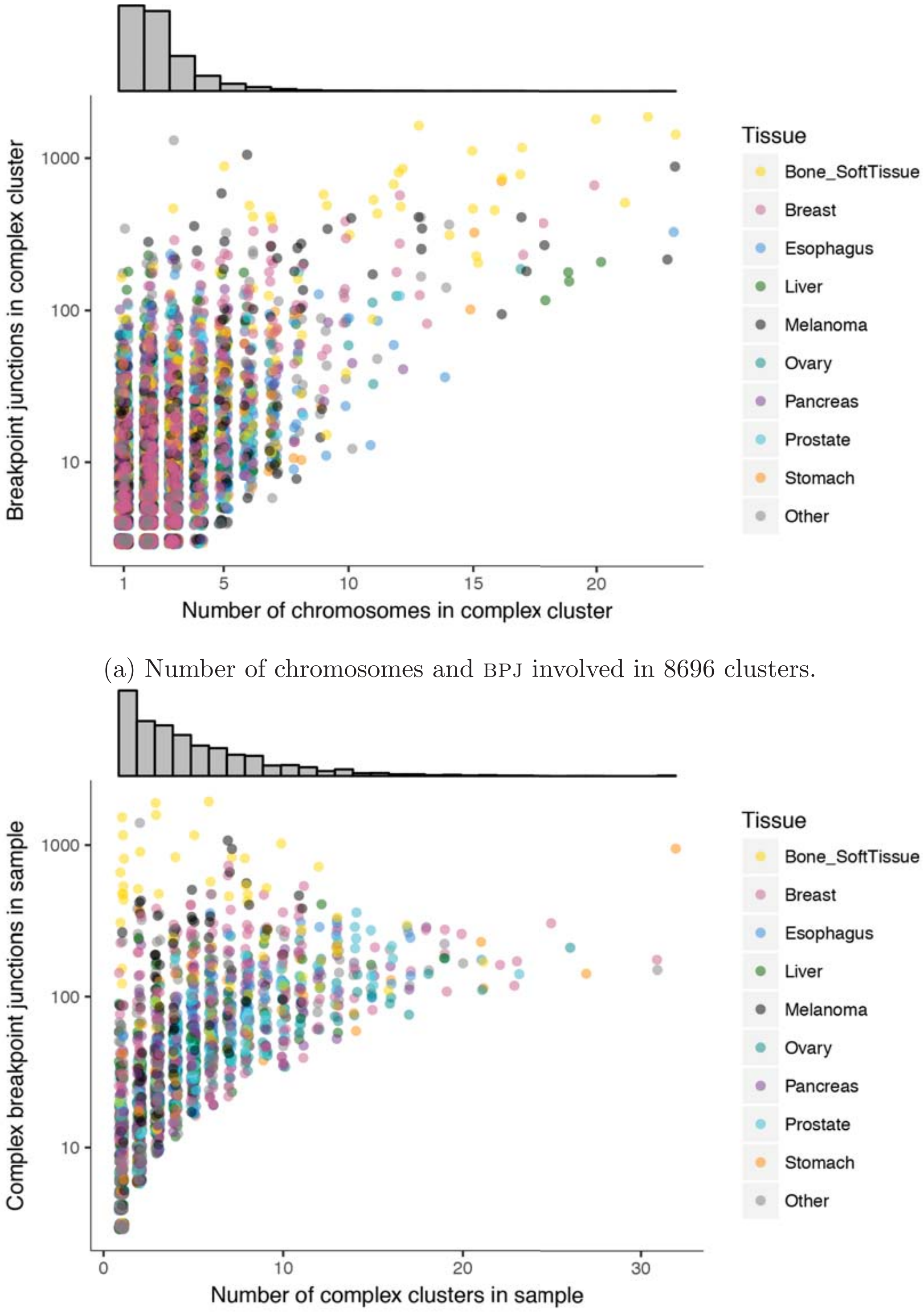
5.4 Outlying clusters and samples

Having excluded the set of 6964 BPJ in tiny clusters of one or two junctions (Section 5.2), 144,248 BPJ remain in 8696 unexplained clusters of three or more BPJ, spread across 1811 samples. As illustrated in Figure 5.10, the vast majority of samples contain fewer than 100 unexplained BPJ spread across a small handful of clusters, with most events containing fewer than 10 BPJ within one or two chromosomes. Indeed, just under 40% of these unexplained clusters involve only three or four BPJ. However, each of these distributions has a long tail, with many outlying clusters and samples.

One outlying event involving more than 1000 BPJ distributed over just three chromosomes—and primarily two chromosomes upon inspection—is the kidney renal cell cancer rearrangement shown in Figure 5.11. This event has the characteristic hallmarks of chromothripsis, with short fragments along two distinct chromosome arms randomly shuffled together to generate an oscillating CN profile. The number of breaks is unusually high (even for chromothripsis), particularly within this relatively contained region spanning 128 Mb on chr21 and chrX (15 kb median gap between adjacent breaks).

In contrast, Figure 5.12 shows two outlying events with relatively few BPJ spanning a large number of chromosomes in esophageal cancer. The distinctive 'star' pattern of multiple translocations emanating from one confined source locus is a hallmark of retrotransposition from an active L1 element. Although the PCAWG structural variation working group endeavoured to separate all retrotransposition events for independent analysis by Rodriguez-Martin et al. (2017), some complex clusters appear to have slipped through this filter, presumably because the activity stems from a secondary (somatically transposed) element. The two samples presented in Figure 5.12 are both known to have high retrotransposition activity more generally, with Rodriguez-Martin et al. (2017) reporting 427 transpositions in SA528901 and 125 transpositions in SA130917.

Another set of outlying SV clusters are massive rearrangements involving hundreds to thousands of BPJ spanning more than a dozen reference chromosomes. Four BPJ clusters even extend to the entire complement of 23 reference chromosomes. The twenty BPJ clusters spanning 17 or more chromosomes are represented as node-edge graphs in Figure 5.13, including six sarcomas, five melanomas, four liver cancers, and three breast cancers. To demonstrate the level of detail underlying each simplified graph representation, Figures 5.14 and 5.15 present the full BPJ plot for two examples: a liver sample with relatively



(a) Number of chromosomes and BPJ involved in 8696 clusters.

(b) Number of clusters and BPJ within 1811 samples.

Figure 5.10: Complex SV events of three or more BPJ in the PCAWG cohort.

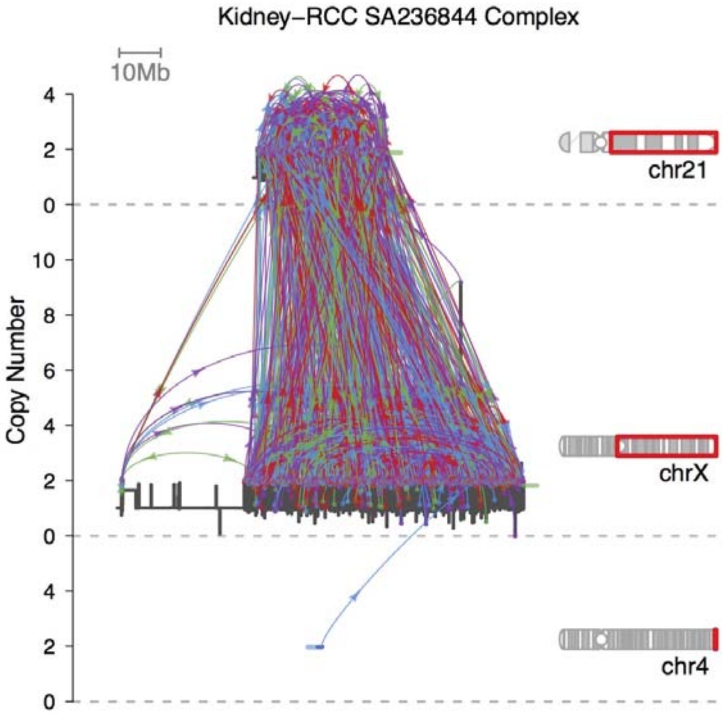


Figure 5.11: Unusual chromothripsis event with 1365 BPJ spanning two chromosome arms in a kidney renal cell cancer.

sparse connectivity, and a liposarcoma sample with high connectivity between most nodes. In the liver example (Figure 5.14), the small local copy gains implicate a dominant role for template and replicate repair, whereas the sharp copy spikes over a low oscillating SV background in the sarcoma example (Figure 5.15) are consistent with a break and ligate model of chromothripsis with subsequent DM amplification and integration. In all examples, the complex network structures were unable to be subdivided with the current methodology into smaller, more local, clusters. It remains unclear whether these giant clusters amass through the chance proximity of independent events on separate homologous chromosomes and/or in separate subclonal populations, or are genuinely connected on the same derivative chromosomes through one or more rounds of punctuated genome evolution. In future work, samples with mass SV overlap may require specialised analytic approaches to divide and describe their relevant features via simplifying assumptions that are generally unnecessary in samples with more isolated rearrangement.

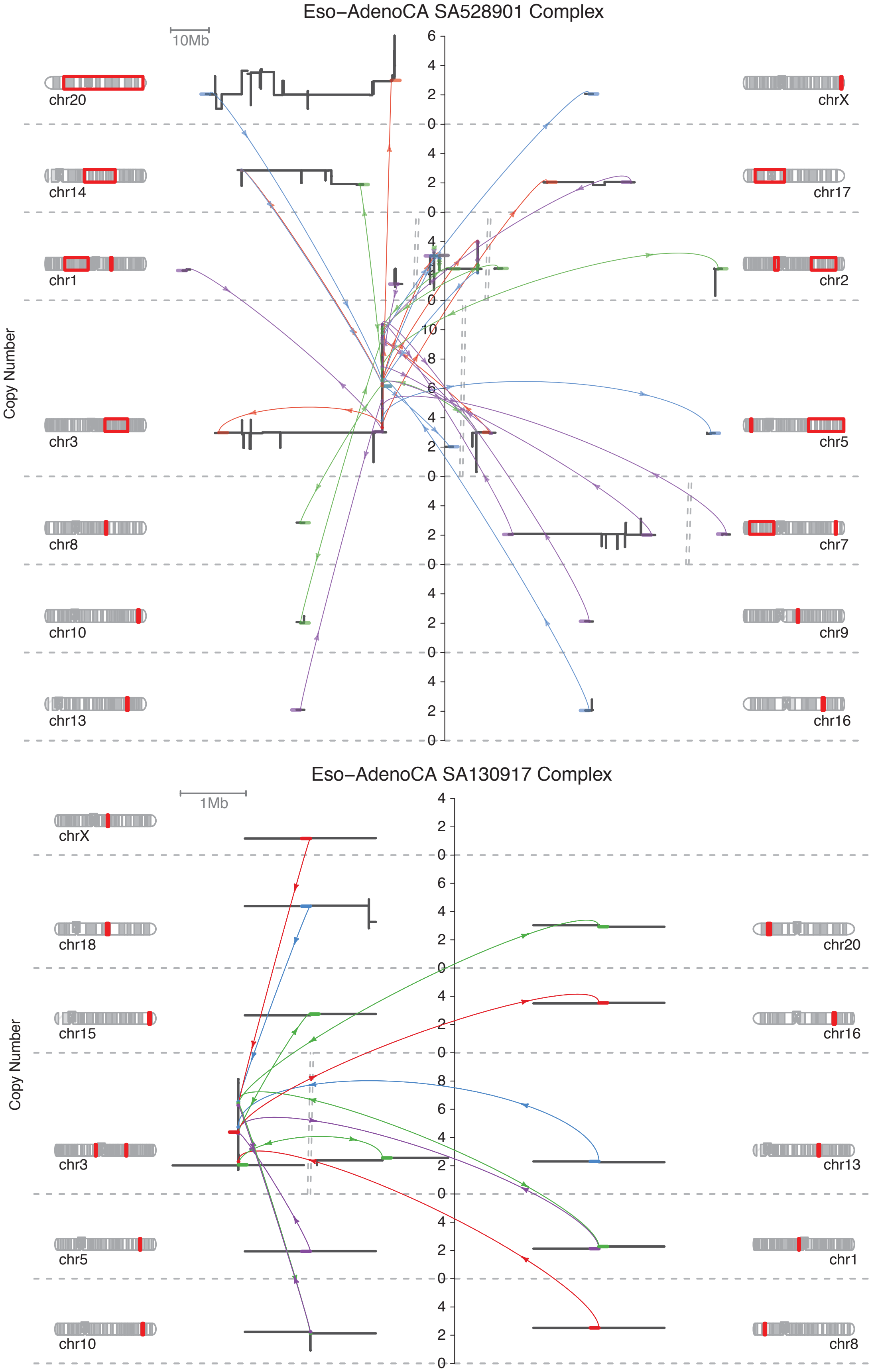


Figure 5.12: Somatic retrotransposition clusters spanning many chromosomes with relatively few BPJ.

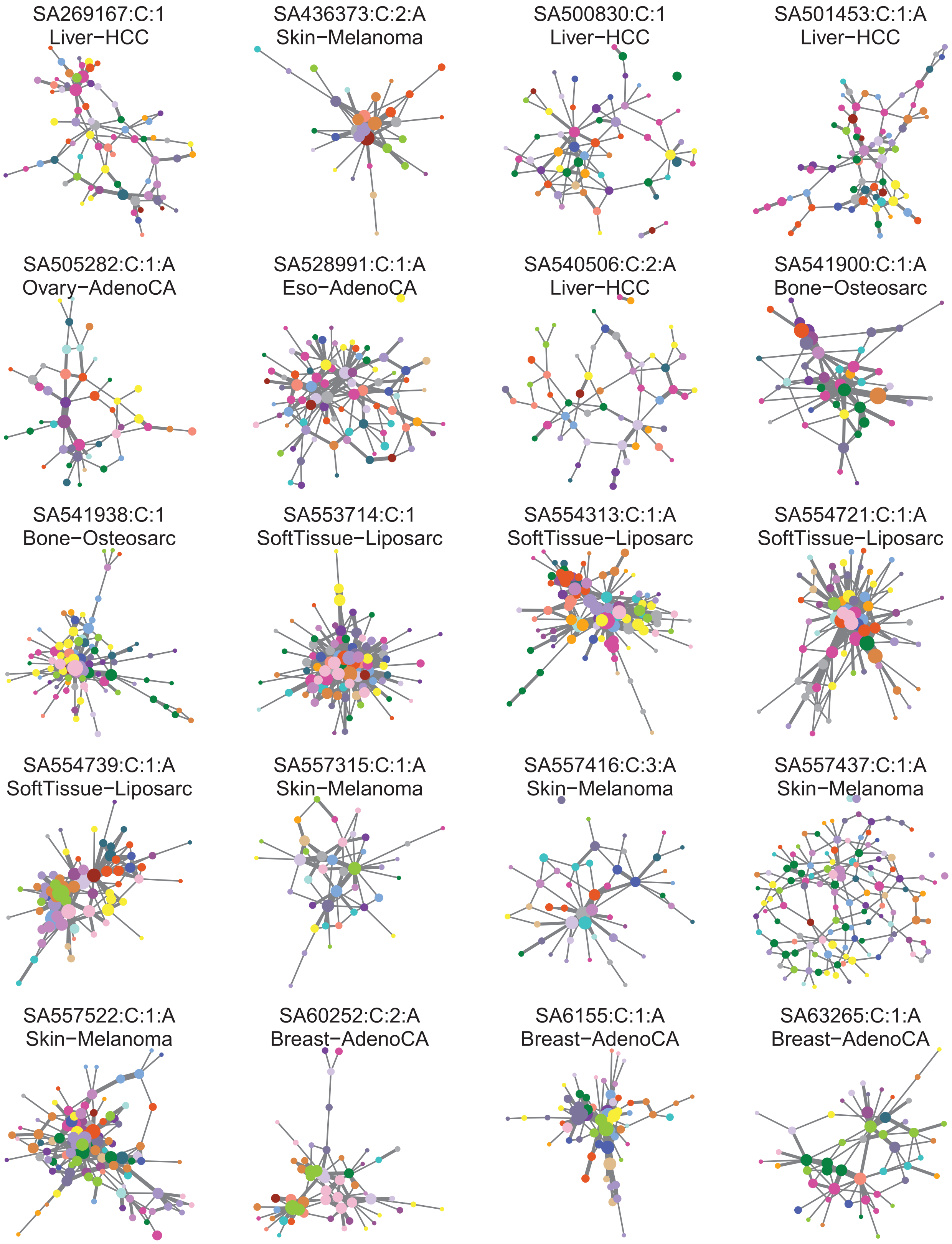


Figure 5.13: Graph representation of all BPJ clusters spanning 17 or more chromosomes. The footprint nodes partition adjacency gaps greater than 5 Mb.

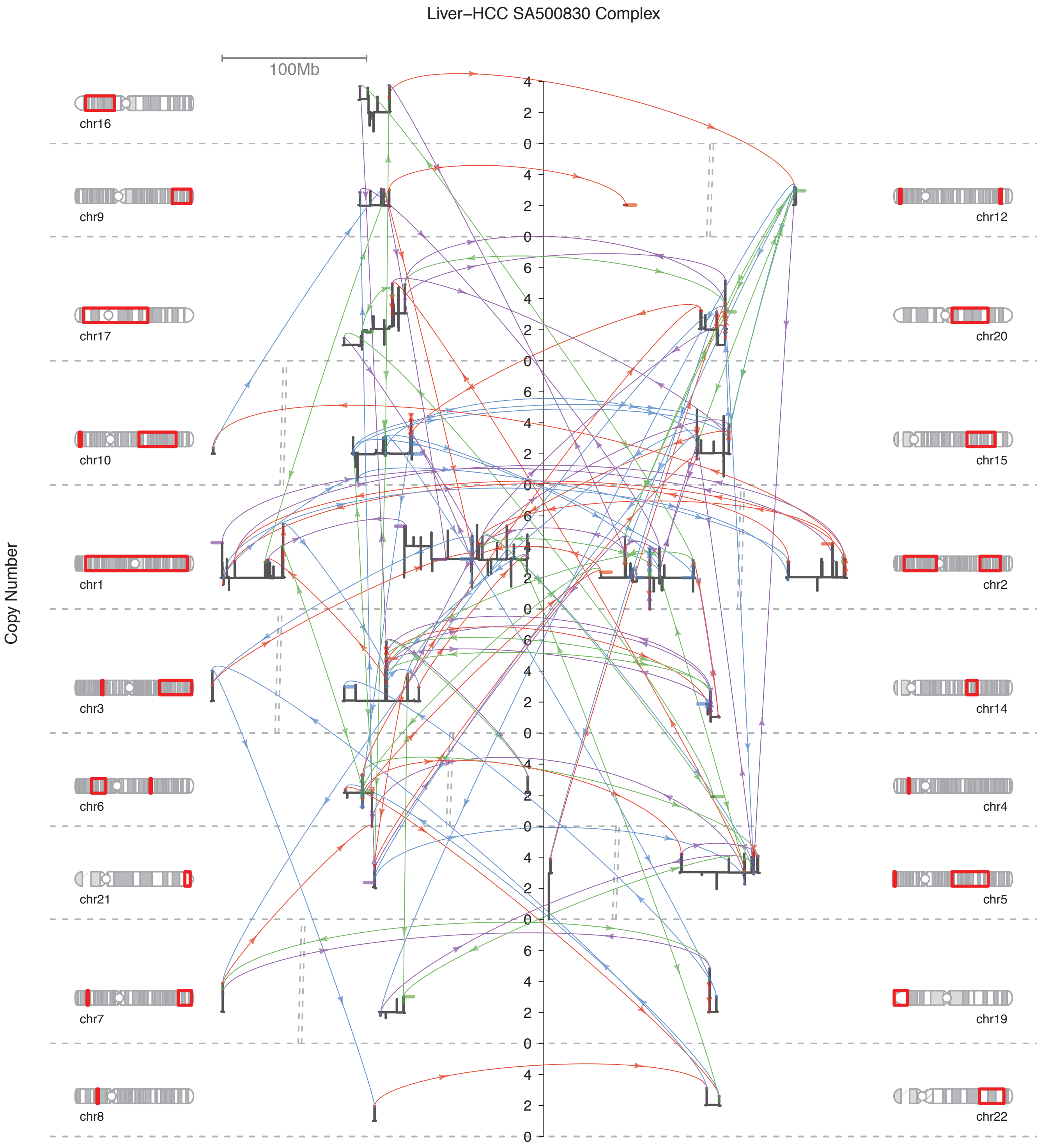


Figure 5.14: Complex SV cluster in a liver cancer sample spanning 19 chromosomes with 155 BPJ.

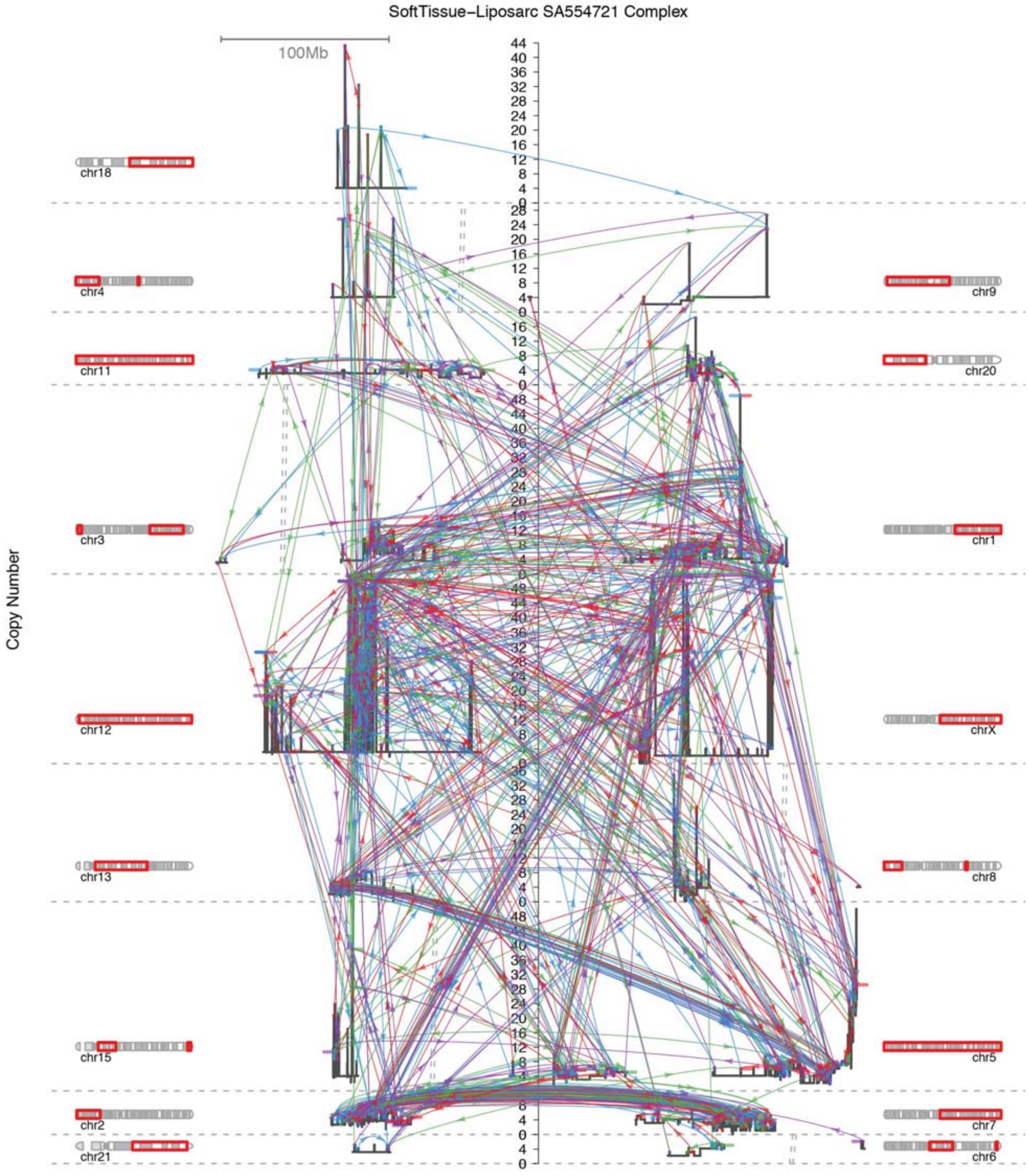


Figure 5.15: Complex cluster in a liposarcoma sample spanning 17 chromosomes with 1122 BPJ. The vertical copy number scale is limited to a maximum of 50.

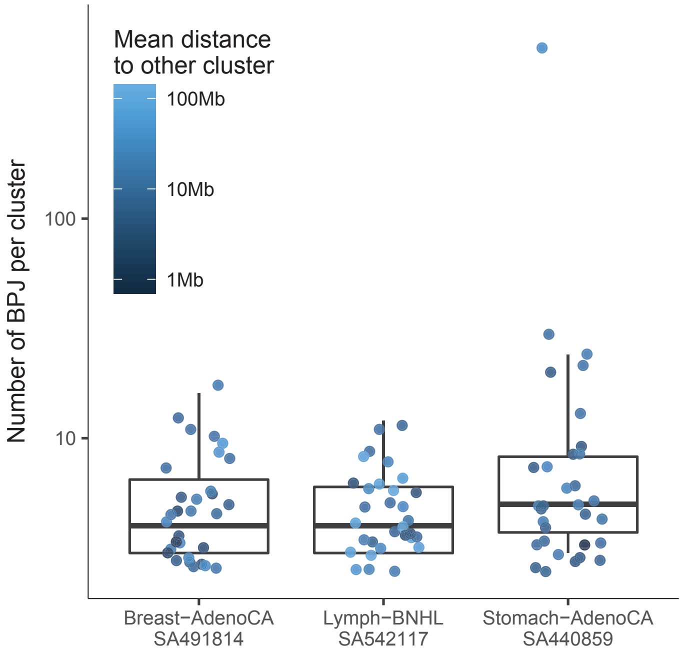


Figure 5.16: Number of BPJ per cluster in three outlying samples with more than 30 separate complex clusters. Each dot is shaded by the average distance between breakpoints within that cluster and the next closest breakpoint in a different complex cluster.

As shown in Figure 5.10B, the PCAWG cohort includes three outlying samples—a breast, lymphoma, and stomach cancer—each containing over 30 separate complex SV clusters. In each case, the vast majority of clusters are small to medium events (fewer than ~20 BPJ) separated by several megabases (Figure 5.16). Manual inspection revealed that most events in these recurrently affected samples have characteristic hallmarks of template and replicate repair, including small local copy gains and many [−+] insertion motifs. A selection of these events are shown in Figure D.27, including one interesting example in the breast sample (third row, first column) of a templated insertion cycle crossing back on itself to re-replicate and insert the same locus (at different lengths) twice over. These examples are testament to the sample-specific activity of particular rearrangement mechanisms, in this instance generating multiple complex configurations with broadly similar features.

5.5 Small unexplained BPJ clusters

Of the 8696 complex clusters, 3435 involve only three or four BPJ (total of 11,537 BPJ). For future method development, I propose that categorisation of these medium-complexity SV events may best be achieved as a separate task, as strategies optimised for success on large clusters of dozens of BPJ are unlikely to extend to these (relatively) small configurations. Here, I present a diverse—but not exhaustive—selection of the major SV patterns found in these small unexplained BPJ clusters. In lieu of a systematic taxonomy, I aim to provide a summary of the dominant features to expect and account for in further studies. Of the small rearrangements *not* summarised in this section, the most common structures are simple DM circles presenting with highly amplified copy number, and groups of adjacent foldback BPJ indicative of BFB cycles.

5.5.1 Break and ligate SV

The hallmarks of break and ligate DNA repair are small copy loss regions demarcated by [+−] gap motifs with junction reciprocity across local or distant loci.

Figure 5.17 illustrates small SV clusters consistent with three or four DSBs along one locus, with subsequent ligation repair to reorder and/or reorient the internal segments after some degree of copy loss at each break. For example, three local breaks may transmute a reference sequence of `abcd` segments into various derivatives harbouring junctions of non-contiguous sequence, including `acbd`, `ac(b)d`, `a(c)bd` or `a(b)(c)d`[b]. These events occupy a middle ground between simple reciprocal inversion and larger break and ligate events across multiple loci (chromoplexy) or dozens of breaks (chromothripsis). As such, these small clusters may warrant a novel classification term of "k-break" (for small $k = 3, 4, \dots$).

Figure 5.18 illustrates small break and ligate clusters spanning two chromosomes. The upper two rows show events where the middle fragment in a deletion SV is rescued and inserted into a distant break. In an unusual variation, the lymphoma example (second row, first column) is consistent with fragmentation of the deleted chr13 segment, with *two* small fragments ligated into a break on chrX. These events share similar features to chromoplexy, but instead of reciprocal exchange between loci, the lost fragment from one side is captured as a simple insertion in the other side. In the third row, these unbalanced translocation events share similar features to the translocation plus inverted insertion 2-jumps first illustrated in Figure 2.5. The complex extensions shown here involve multiple fragments on one or both sides of the translocation. In the fourth row, the prostate and lymphoma examples show reciprocal translocation overall, with the added complexity of intervening fragment capture in one of

[b] Parentheses denote inverted segments.

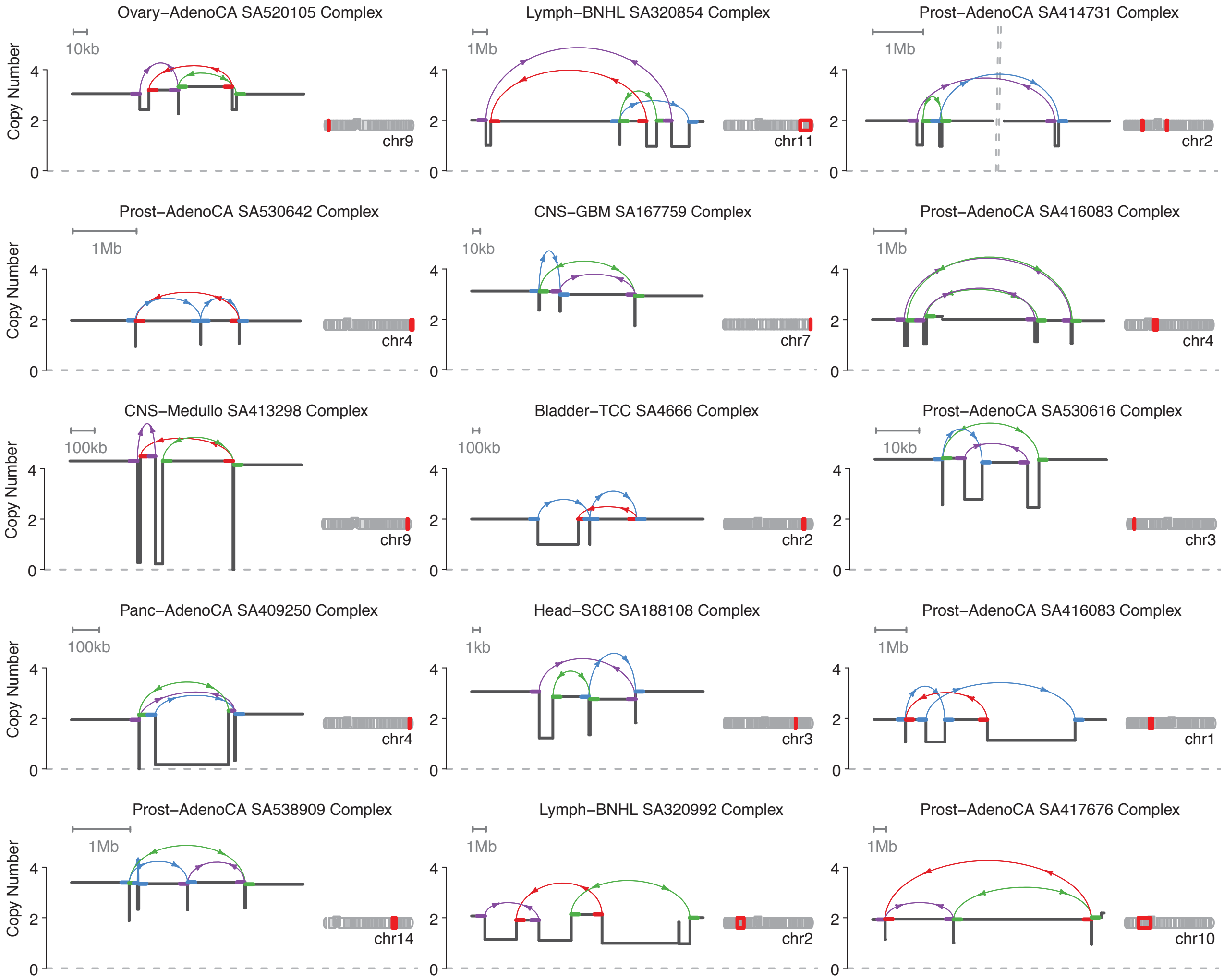


Figure 5.17: Small break and ligate clusters on one chromosome

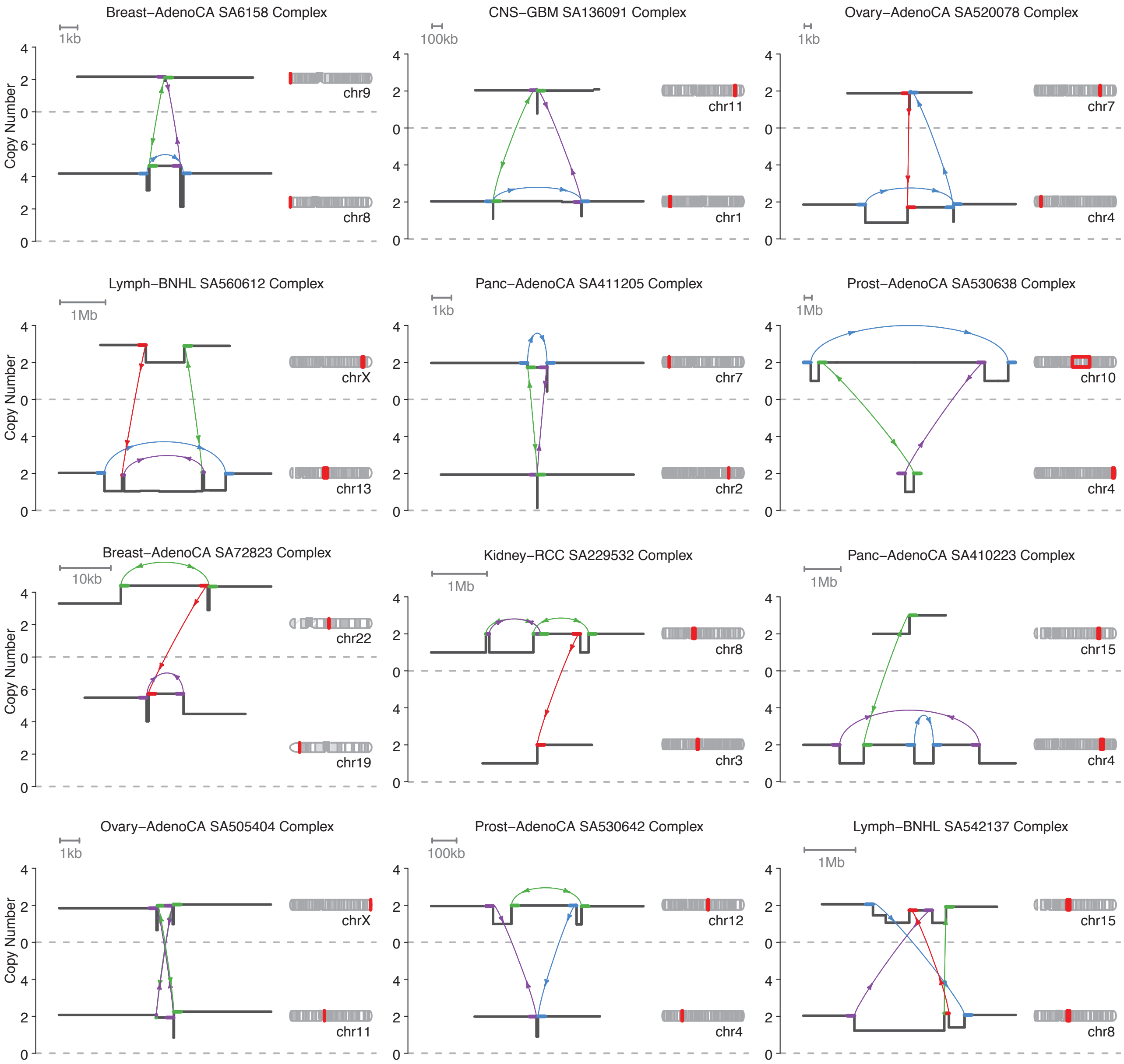


Figure 5.18: Small break and ligate clusters on two chromosomes

the translocation derivatives. Finally, the ovary example (bottom left) is an unusual event of double reciprocal translocation consistent with non-crossover recombination whereby small fragments (about 1 kb) on chrX and chr11 are mutually exchanged. Although this rare configuration presents with hallmark break and ligate features, this structure is likely to result from a rare somatic double Holliday junction resolution following non-allelic HR.

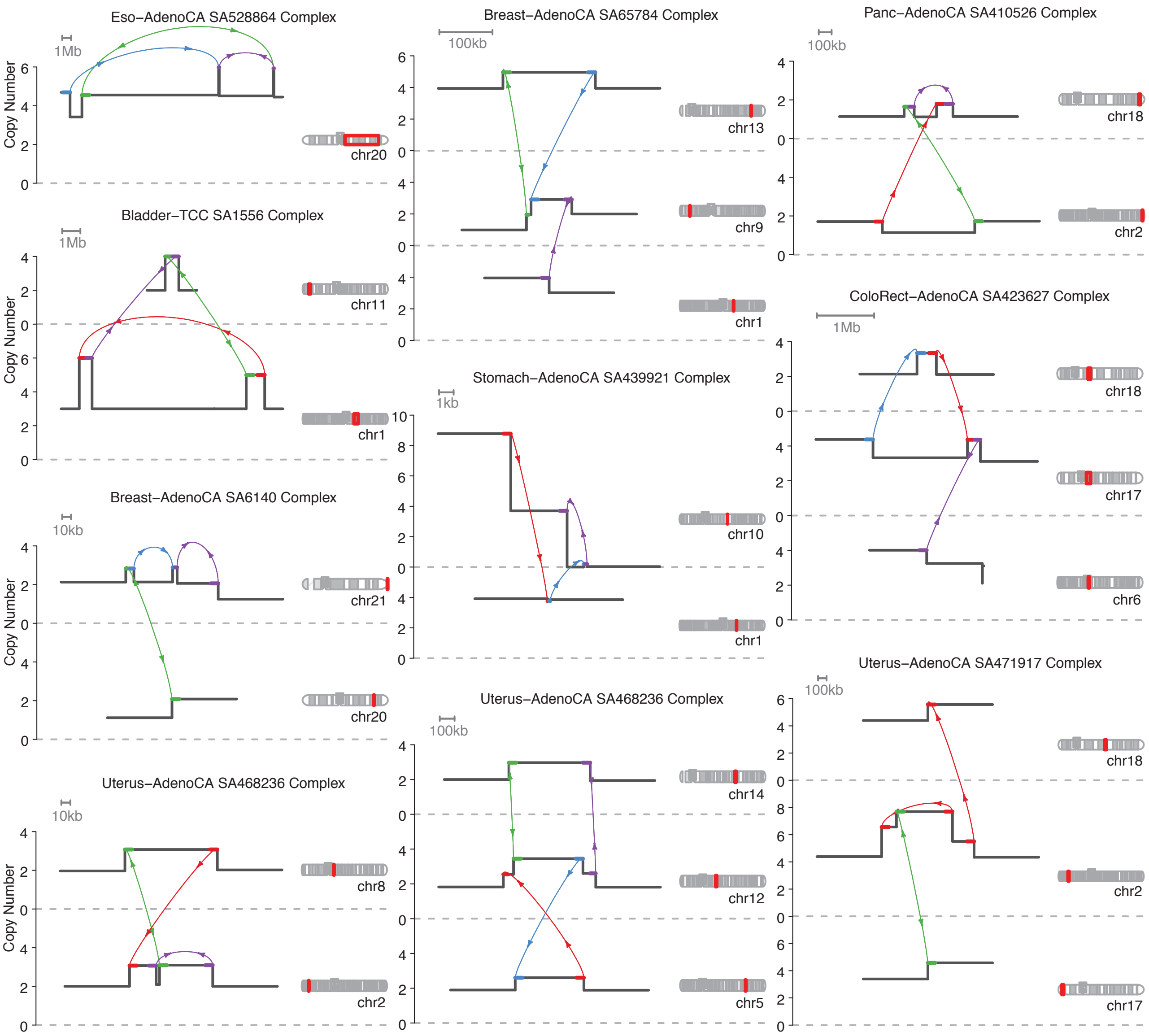


Figure 5.19: Small complex templated insertion events with adjacent or overlapping footprints.

5.5.2 Template and replicate SV

The hallmarks of template and replicate DNA repair are small copy gain regions demarcated by [−+] insertion motifs or overlapping intrachromosomal BPJ.

Figure 5.19 illustrates a subset of the many templated insertion events that were missed in the initial classification scheme (Section 2.1.3) because the footprints were either adjacent or overlapping, and therefore not detected as completely isolated [−+] motifs. These overlooked templated insertions include bridges, chains, cycles, and at least one insertion-mediated foldback shown for a stomach cancer sample. In future projects, the definition of templated

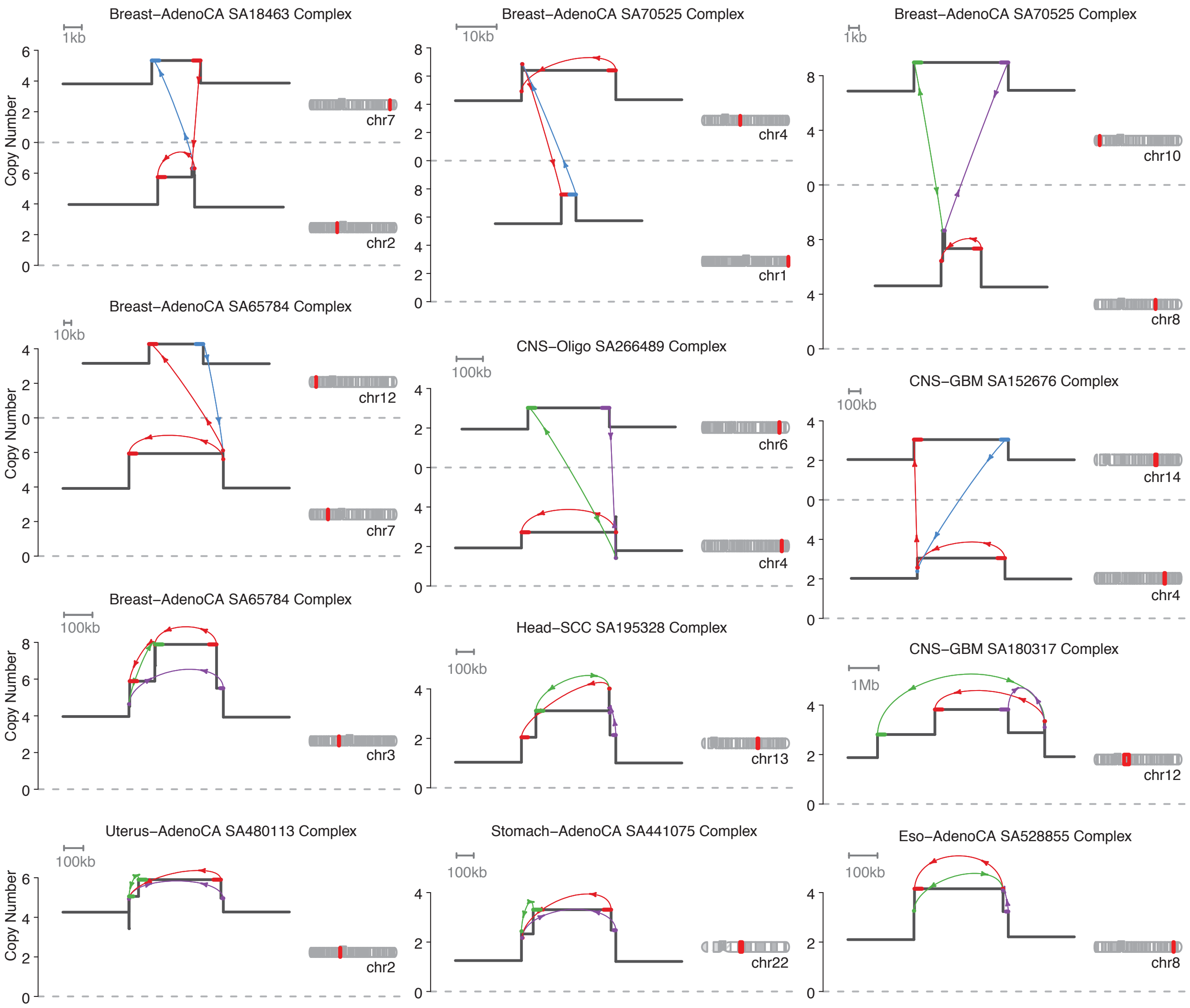


Figure 5.20: Small template and replicate clusters with three BPJ converging at one recurrent break position.

insertion should ideally account for these additional possibilities.

Figure 5.20 illustrates a very common pattern consistent with local or distant polymerase template switching where three or more BPJ all converge at (or emanate from) the same recurrent break locus. I hypothesise that these events are precipitated by a persistent DNA lesion—such as an inter-strand crosslink (Meier et al., 2014)—triggering multiple template switches at the same position.

5.5.3 Combination SV

Occasionally, small BPJ clusters present with unexpected configurations (and no obvious false negative or false positive calls) that are inconsistent with

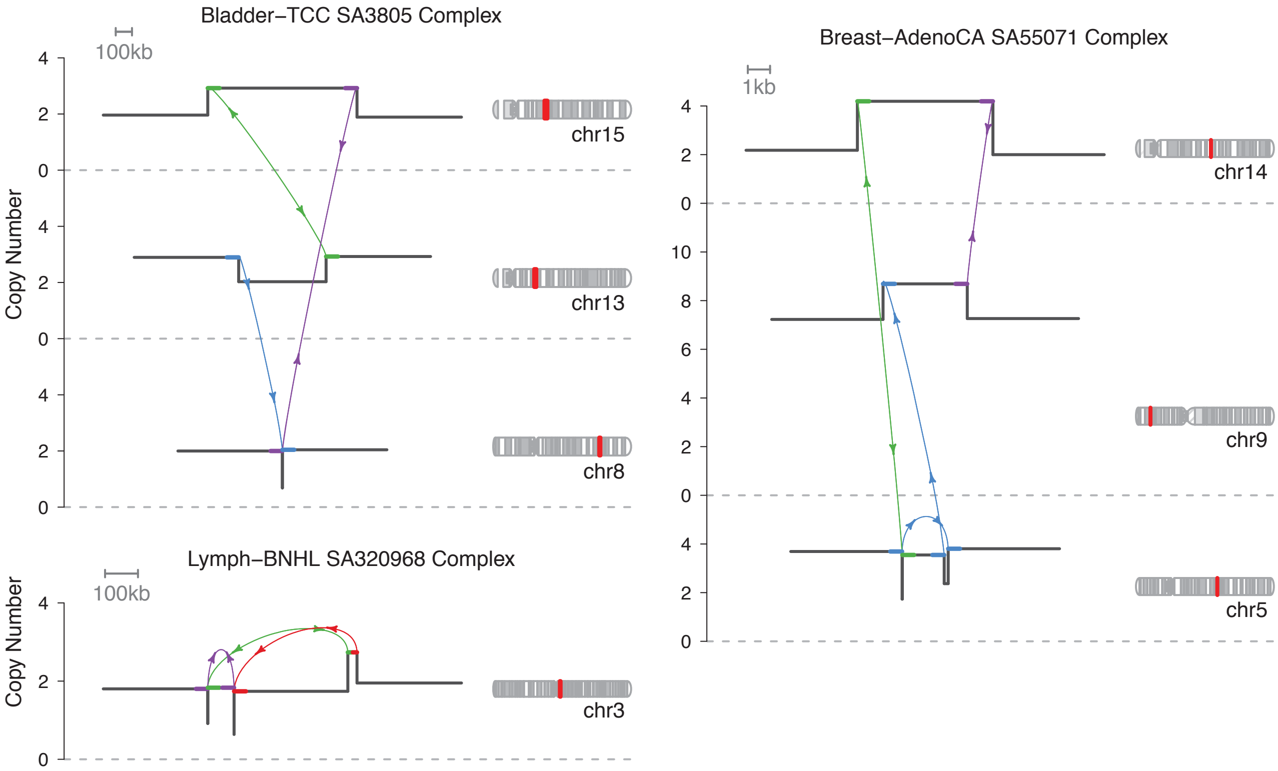


Figure 5.21: Combination SV clusters with hallmarks of both break and ligate *and* template and replicate repair mechanisms.

either repair mechanism acting in isolation. Three such examples are shown in Figure 5.21. In the bladder sample cluster of three BPJ, the data suggest an overall effect of reciprocal translocation, combined with the added complexity of a templated insertion from a distant locus within one derivative chromosome. The breast sample cluster of four BPJ appears to be a small templated insertion cycle, additionally capturing a fragment lost through deletion on another chromosome (as previously introduced in Figure 5.18). Finally, the lymphoma sample cluster of three BPJ appears to generate a reciprocal inversion with a templated insertion copied into one of the breaks. These observations are somewhat incongruous with our current understanding of rearrangement mechanisms, hinting at unexplored subtleties in the repertoire of DNA repair.

To complete this overview of the major patterns generated by three or four BPJ, Figure 5.22 illustrates a range of clusters involving overlapping BPJ that may or may not result from chance proximity of independent events. For example, in the top row, the breast and head SCC examples are possibly consistent with a dup–inv-dup local 2-jump following by subsequent tandem duplication or deletion, *or* may possibly result from three polymerase template switches. Likewise, the pancreas example is consistent with overlapping deletion and

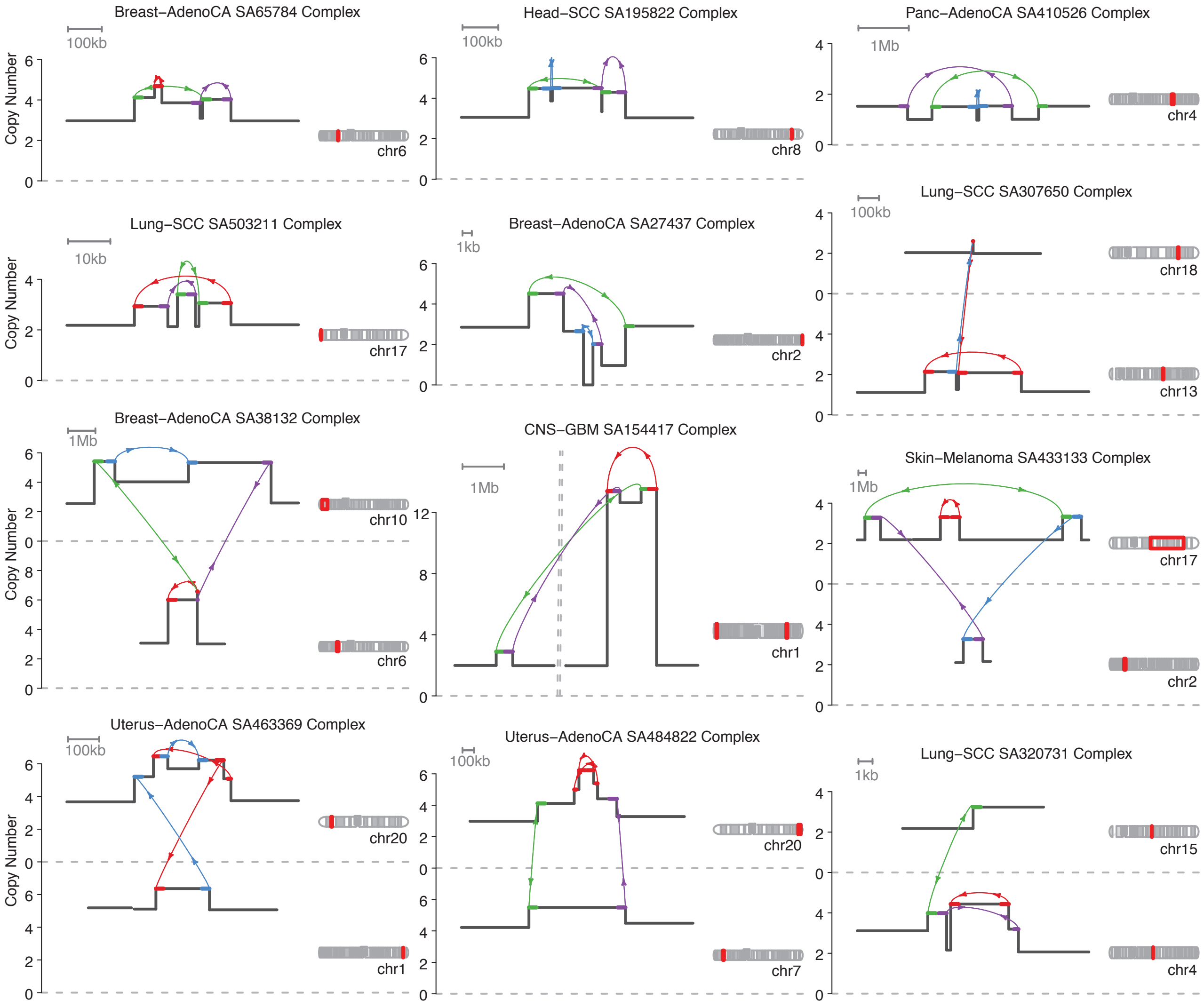


Figure 5.22: Overlapping or adjacent SV clusters

reciprocal inversion events independently acquired, *or* with a local 3-break (as in Section 5.5.1) repaired in the order `a(c)(b)d`. Future BPJ classification projects will ideally address the complexity and ambiguity generated by overlapping clusters of few BPJ, perhaps by conditioning on the sample-specific frequencies and sizes of the various isolated SV classes.

5.6 Heuristic classification of complex SV

To complete a *tour d'horizon* of the complex SV landscape, this section explores the remaining tranche of unclassified SV with an approximate parametrisation of various rearrangement phenomena. This survey is preliminary in nature, aiming to furnish future endeavours with a base appreciation of the challenges involved.

After filtering out 30 clusters of fragile site deletion and 16 clusters of immune loci recombination, there remain 5215 complex SV of five or more BPJ. To initially assess the character and scope of these unexplained clusters, I defined a suite of heuristic classification rules to mark each event as a 'first tier' or 'second tier' candidate example of different SV categories (detailed in Appendix C). These pilot classifications are *not* enforced to be mutually exclusive, so one SV cluster may match the provisional criteria for several groups.

For the six categories currently implemented—breakage fusion bridge, complex chromoplexy, chromothripsis *without* double minutes, complex amplification (possibly chromoanasynthesis), isolated double minutes *without* chromothripsis, and retrotransposition hotspots—1051 SV clusters (20%) meet first tier criteria for at least one class. The overlap at first tier is minimal for most categories (Figure 5.23), with the exception of chromothripsis and complex chromoplexy which manifest on a spectrum of break and ligate repair, sometimes with ambiguous origin. I estimated the specificity of the first tier classifications by manually curating fifty randomly chosen examples in each category (or the maximum possible for retrotransposition), counting half a point for uncertain candidates. The specificity estimates ranged from 95% or higher for retrotransposition and double minutes, to just above 70% for chromothripsis (Table 5.3).

Double minute candidates are often found in glioblastoma samples, and involve one or more reference fragments in highly amplified extrachromosomal circles (Figures 5.24 and 5.25). Breakage-fusion-bridge candidates are enriched in esophageal, pancreatic, and many other cancer types (including SCC in lung and head), causing step-wise copy gain profiles (Figures 5.24 and 5.26). Complex amplifying events are enriched in cancers of female reproductive tissues, recapitulating the tissue preference of small template and replicate events like tandem duplication and templated insertion (Figures 5.24 and 5.27). I hypothesise that many of these amplifications are caused by multiple polymerase template switches, and could possibly be termed 'chromoanasynthesis'.

Table 5.3: Complex SV clusters (five or more BPJ) meeting the first tier criteria for preliminary classification as defined in Appendix C. The specificity of each category was estimated by manual curation of fifty randomly chosen examples.

Group	Clusters	Specificity	Median BPJ	Total BPJ
Break-Fus-Bridge	168	0.80	9	1688
C-plexy	515	0.90	8	5904
C-thripsis (noDM)	228	0.72	16	5025
Complex Amplify	130	0.88	19	4396
Double Minute	52	0.95	23	2735
Retrotrans	14	1.00	7	119
Unexplained	4164	NA	10	109491

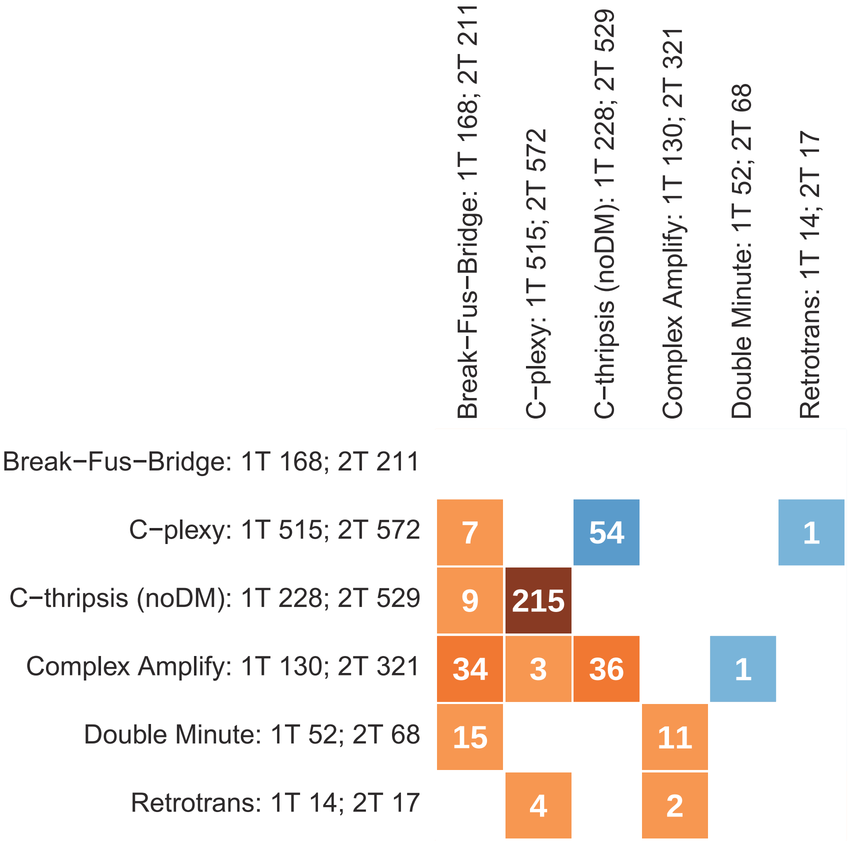


Figure 5.23: Overlap between the pilot classification groupings for the first tier (upper right, in blue) and second *or* first tier (lower left, in orange) complex SV events.

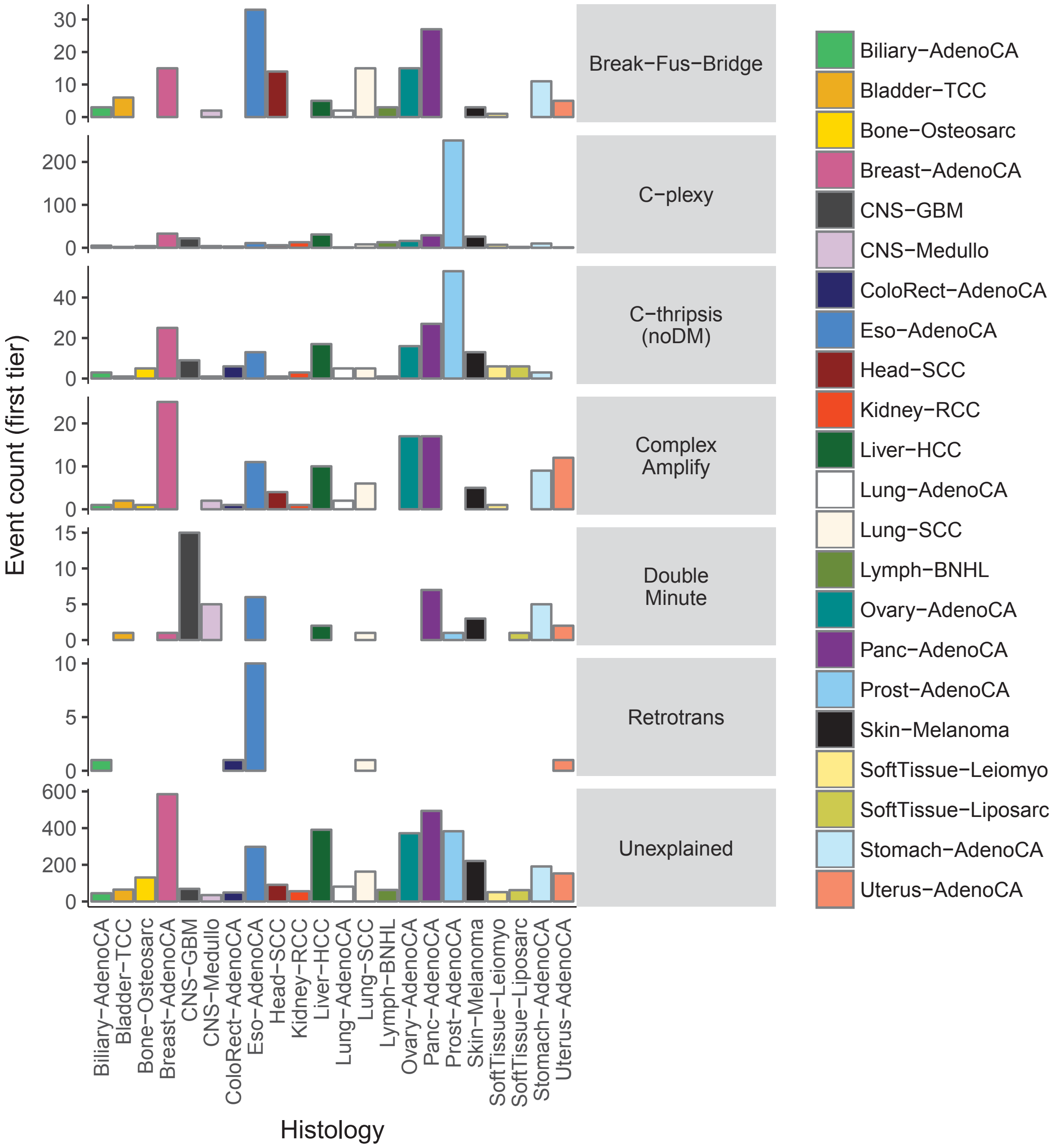


Figure 5.24: Histology distribution of first tier classifications for complex SV, without normalising by sample size or BPJ count.

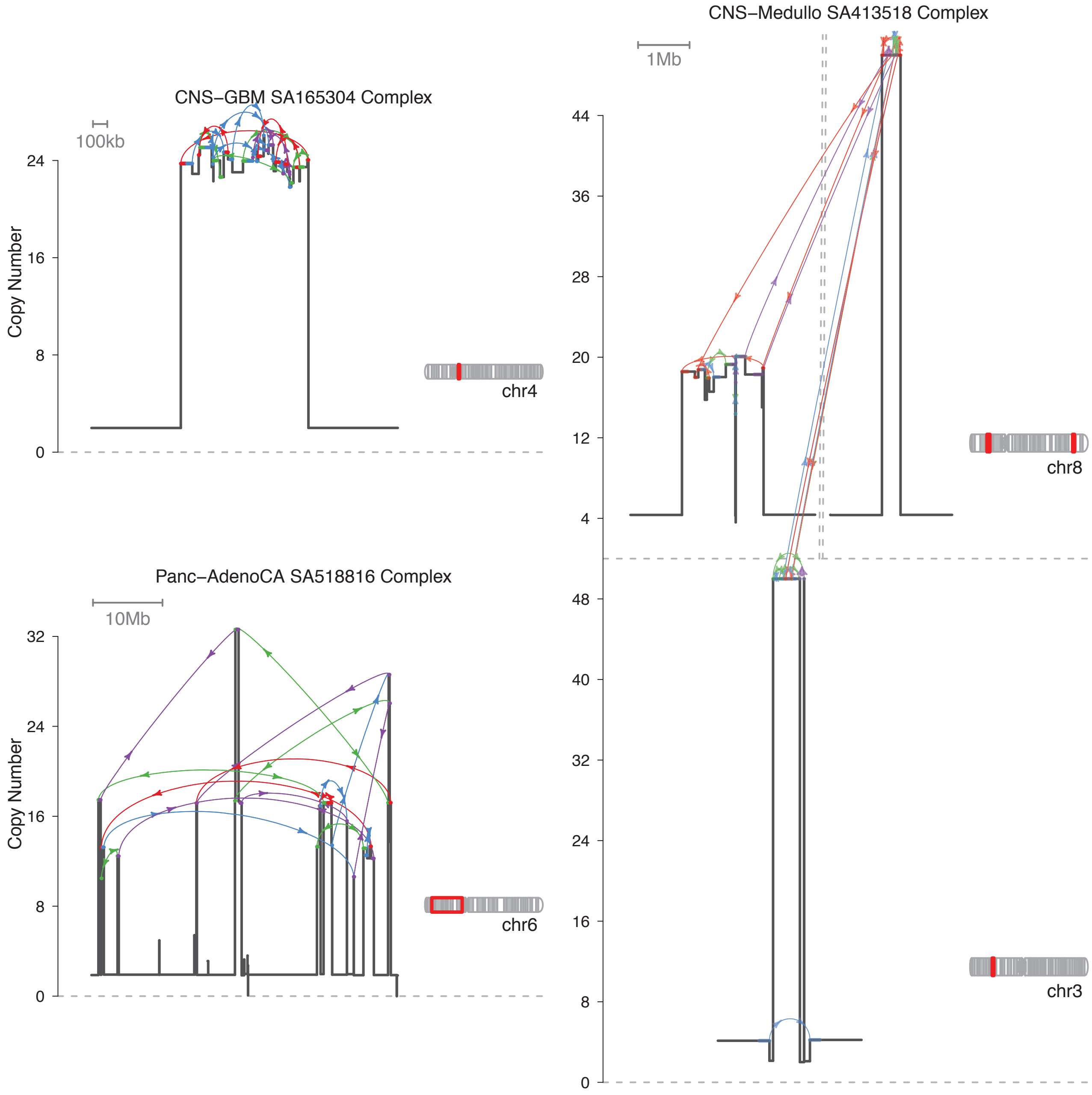


Figure 5.25: Example double minute events (first tier). The vertical CN scale is capped at 50.

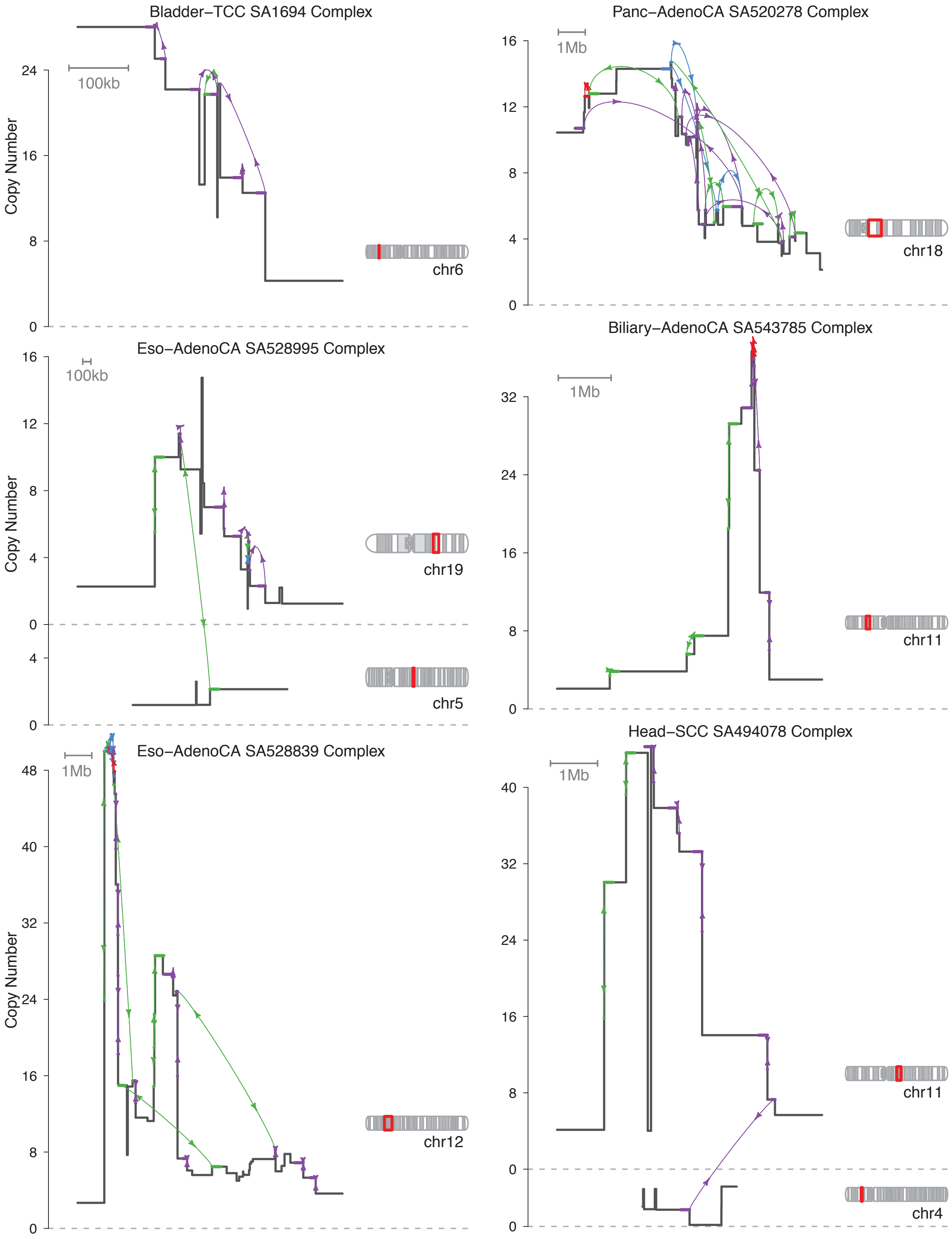


Figure 5.26: Example breakage-fusion-bridge events (first tier)

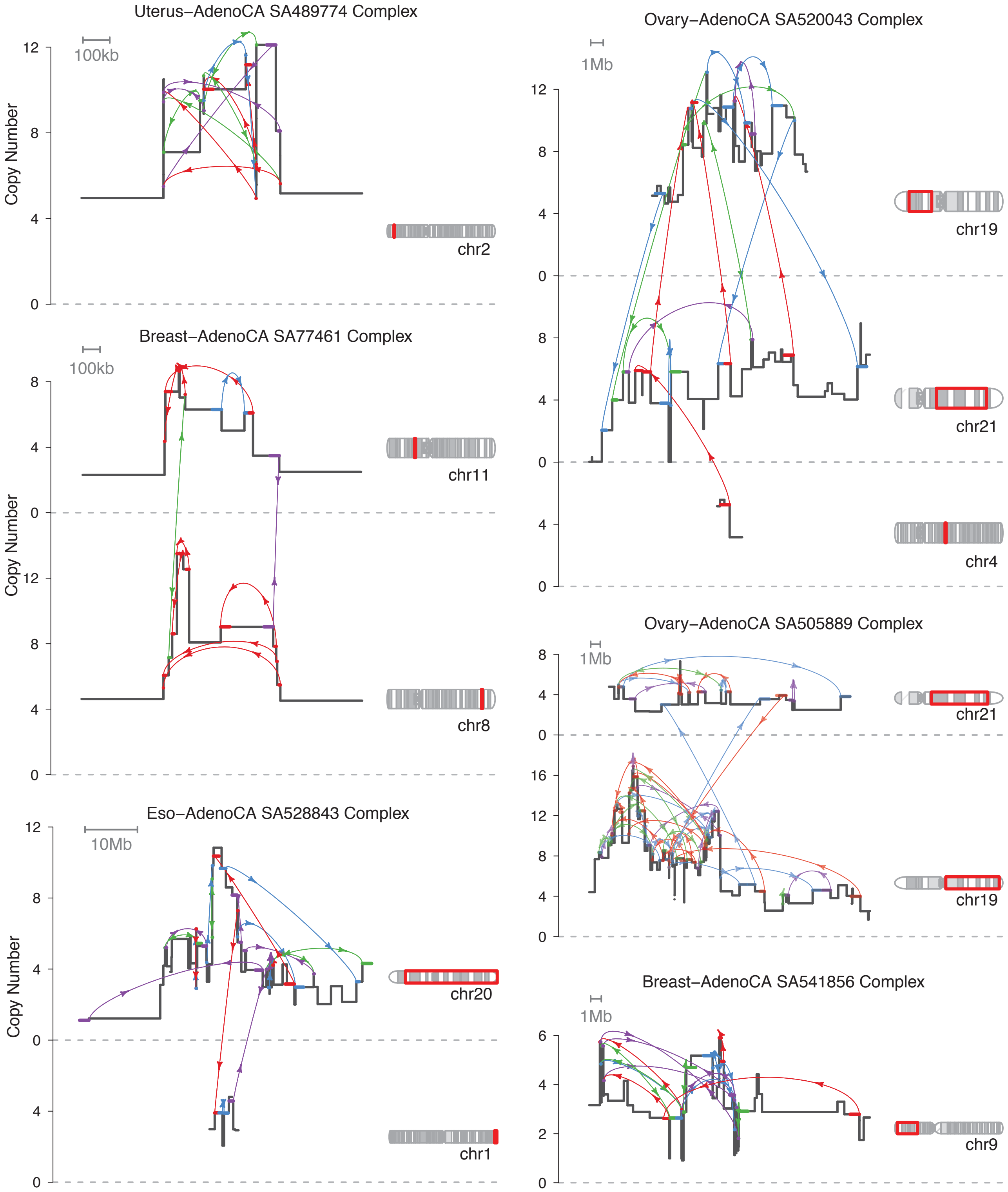


Figure 5.27: Example complex amplification events (first tier), possibly mediated by chromoanasynthesis.

Over 50% of all complex chromoplexy candidates are found in the prostate cancer cohort, often involving micro-fragmentation at each ‘macro’ break locus (Figures 5.24 and 5.28). Chromothripsis events are found in many different cancer types, but are often difficult to distinguish from one end of the chromoplexy spectrum (Figures 5.24 and 5.29). Some chromothripsis candidates span entire arms or chromosomes in a manner consistent with micronucleus capture of lagging DNA (Zhang et al., 2015), whereas other localised events span just a few megabases, and potentially reflect the alternative trigger of chromatin bridge shattering following telomere crisis (Maciejowski et al., 2015).

My heuristic classification rules for preliminary description of the complex SV remain a work in progress, and currently miss chromothripsis events associated with double minute amplification, as well as a range of medium-complexity templated insertions, and other novel patterns yet to be described.

5.7 Discussion

In this chapter, I outlined an exploratory sketch of the structural content within the 55% of PCAWG BPJ left unexplained by the simple SV classifications presented in previous chapters.

As the pre-existing BPJ cluster divisions were not optimised for the meaningful separation of complex events, I developed an alternative BPJ clustering procedure (Section 5.1) using a novel node-edge graph description of connectivity across variably sized footprints. By inspection only, these new cluster partitions appear to be a more logical division of the complex SV landscape, with the ability to merge SV groups connected via multiple distant loci, and separate out distinct sub-graphs with negligible external connection. In its current implementation, the major shortcomings of my alternative clustering procedure relate to the over-reliance on fixed threshold decision points for footprint definition, merging, and separation, without a statistical justification accounting for the sample-specific rearrangement landscape.

The BPJ cluster divisions are assumed to demarcate a set of independent (or at least punctuated) SV events, with hallmark features indicative of the underlying generating mechanism. Clusters of 2–4 BPJ (Sections 5.2 and 5.5) manifest in a huge variety of possible configurations, usually—but not always—consistent with the activity of ‘break and ligate’ *or* ‘template and replicate’ repair across one or two loci. Despite the relatively small number of constituent breaks,

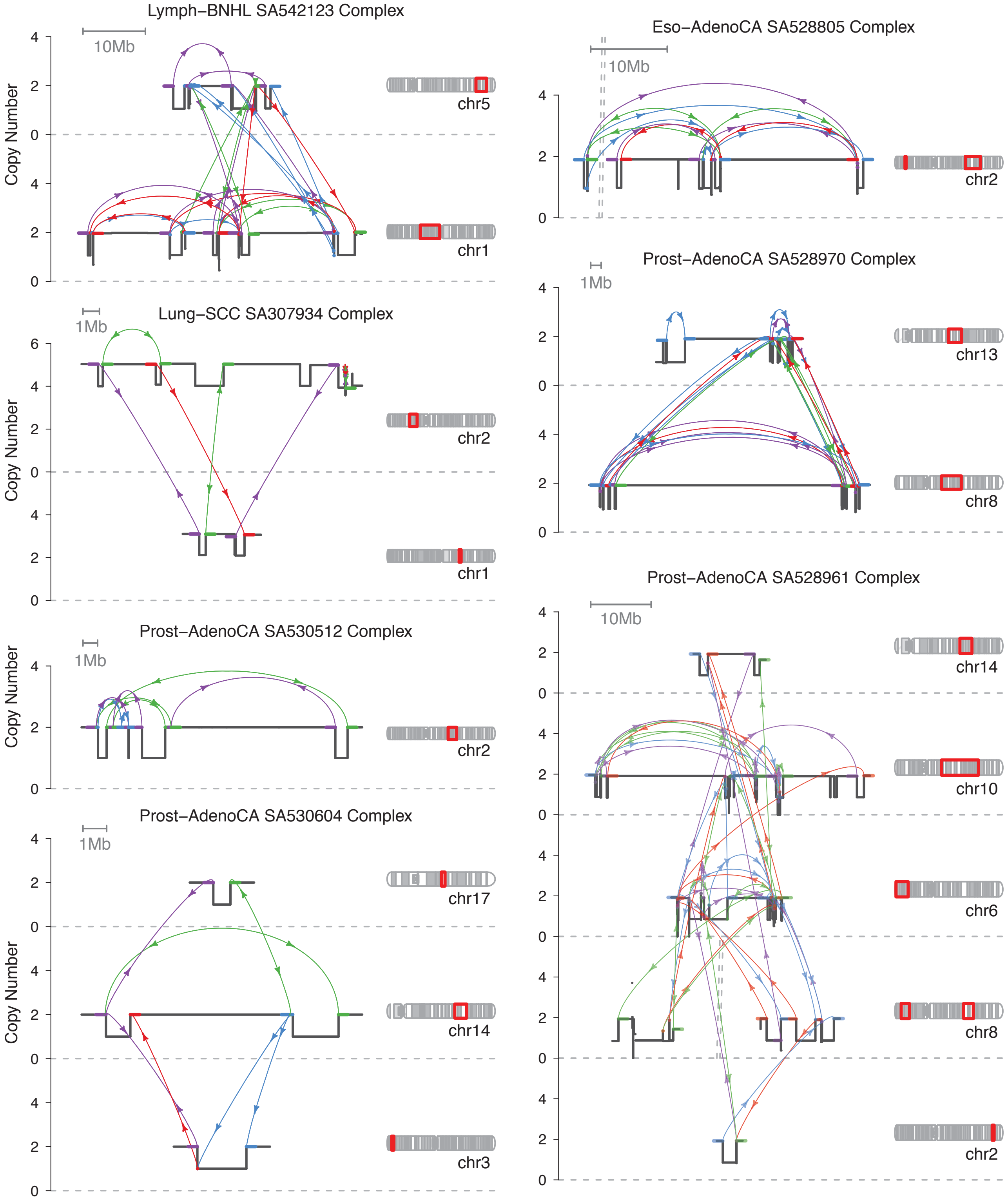


Figure 5.28: Example complex chromoplexy events (first tier)

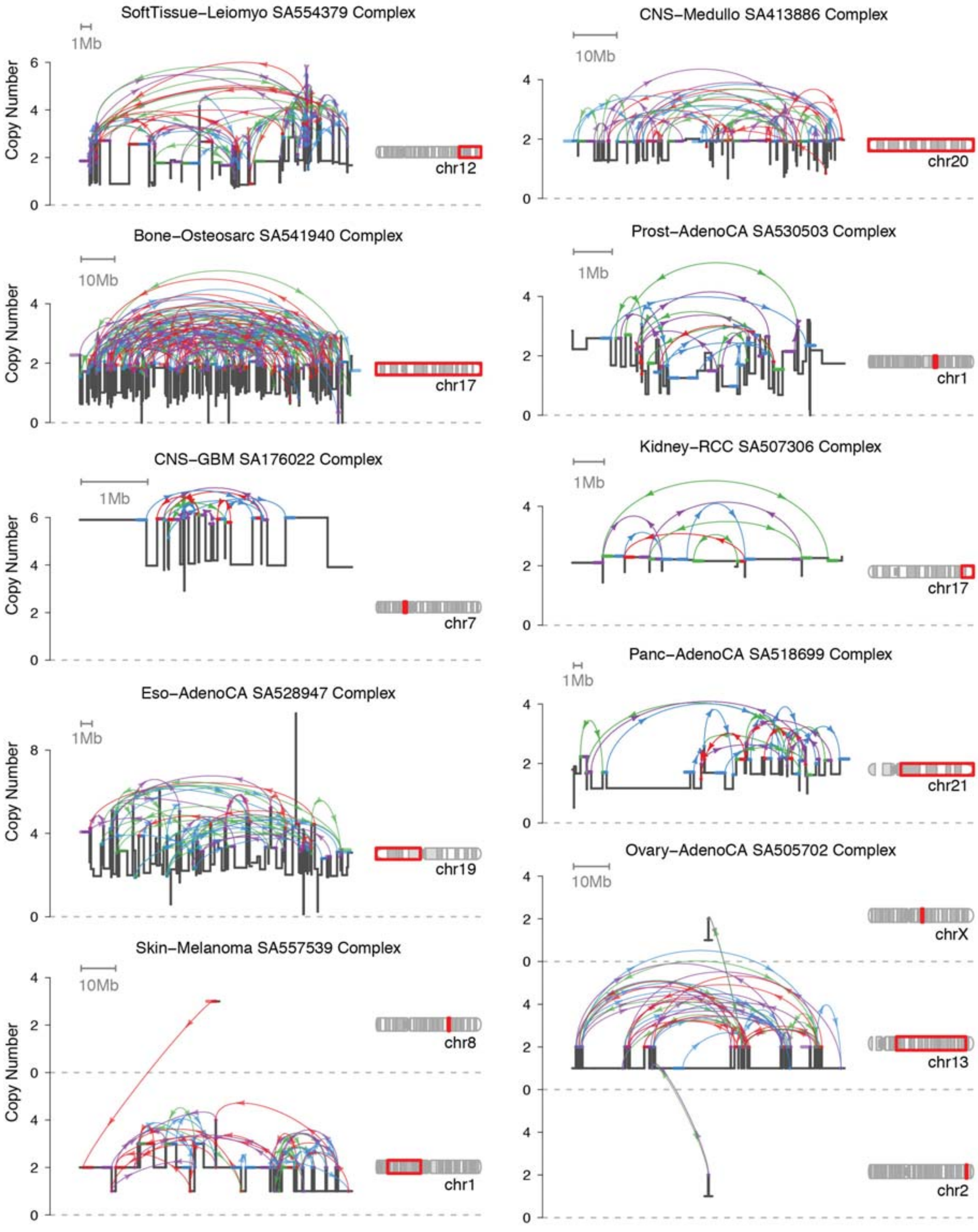


Figure 5.29: Example chromothripsis events (first tier)

these rearrangements are difficult to systematically catalogue. Even for just three BPJ, there are hundreds of possible unique configurations which vary by order, orientation, and connection across loci. I anticipate that these medium-complexity rearrangements will require an intermediate classification strategy between the two extremes of exact motif recognition allowing no variation (as for simple SV) and top-down characterisation of the overall feature distribution (as for large SV clusters).

Large rearrangements of five or more BPJ are highly variable, with some outlying clusters involving more than a thousand BPJ and/or more than a dozen chromosomes (Section 5.4). In a pilot survey, about 20% of complex clusters were approximately compatible with a canonical rearrangement phenomenon (Section 5.6). Of the 80% of clusters with no putative explanation, some fraction may be described by missing categories such as chromothripsis *with* double minutes, others may be retrieved with improved BPJ clustering methods, and some may be confounded by overlapping events, false positive or negative BPJ calls, and/or poor CN segmentation (which is occasionally unreliable, even after the mitigation described in Section 5.3).

The results presented in this chapter describe the major contours of the complex SV landscape, but do not represent a definitive solution to the ongoing challenge of systematic complex rearrangement classification. Strategies for improving the separation and interpretation of complex SV are discussed in Chapter 6.