Patterns of somatic genome rearrangement in human cancer

Nicola Diane Roberts
Trinity College
University of Cambridge

January, 2018

Dissertation submitted for the degree of Doctor of Philosophy

Declaration

This dissertation is the result of my own work and includes nothing which is the outcome of work done in collaboration except as specified in the text. It is not substantially the same as any that I have submitted, or is being concurrently submitted, for a degree or other qualification at the University of Cambridge or any other university. It does not exceed the prescribed limit of 60,000 words.

Summary

Cancer development is driven by somatic genome alterations, ranging from single point mutations to larger structural variants (SV) affecting kilobases to megabases of one or more chromosomes. Studies of somatic rearrangement have previously been limited by a paucity of whole genome sequencing data, and a lack of methods for comprehensive structural classification and downstream analysis. The ICGC project on the Pan-Cancer Analysis of Whole Genomes provides an unprecedented opportunity to analyse somatic SVs at base-pair resolution in more than 2500 samples from 30 common cancer types.

In this thesis, I build on a recently developed SV classification pipeline to present a census of rearrangement across the pan-cancer cohort, including chromoplexy, replicative two-jumps, and templated insertions connecting as many as eight distant loci. By identifying the precise structure of individual breakpoint junctions and separating out complex clusters, the classification scheme empowers detailed exploration of all simple SV properties and signatures.

After illustrating the various SV classes and their frequency across cancer types and samples, Chapter 2 focuses on structural properties including event size and breakpoint homology. Then, in Chapter 3, I consider the SV distribution across the genome, and show patterns of association with various genome properties. Upon examination of rearrangement hotspot loci, I describe tissue-specific fragile site deletion patterns, and a variety of SV profiles around known cancer genes, including recurrent templated insertion cycles affecting *TERT* and *RB1*.

Turning to co-occurring alteration patterns, Chapter 4 introduces the Hierarchical Dirichlet Process as a non-parametric Bayesian model of mutational signatures. After developing methods for consensus signature extraction, I detour to the domain of single nucleotide variants to test the HDP method on real and simulated data, and to illustrate its utility for simultaneous signature discovery and matching. Finally, I return to the PCAWG SV dataset, and extract SV signatures delineated by structural class, size, and replication timing.

In Chapter 5, I move on to the complex SV clusters (largely set aside throughout Chapters 2–4), and develop an improved breakpoint clustering method to subdivide the complex rearrangement landscape. I propose a raft of summary metrics for groups of five or more breakpoint junctions, and explore their utility for preliminary classification of chromothripsis and other complex phenomena.

This comprehensive study of somatic genome rearrangement provides detailed insight into SV patterns and properties across event classes, genome regions, samples, and cancer types. To extrapolate from the progress made in this thesis, Chapter 6 suggests future strategies for addressing unanswered questions about complex SV mechanisms, annotation of functional consequences, and selection analysis to discover novel drivers of the cancer phenotype.

Acknowledgements

With sincere gratitude, I thank my doctoral supervisor Dr Peter Campbell for his steadfast patience, support, and advice over four transformative years. Plumbing the depths and idiosyncrasies of cancer genome rearrangement proved to be both a rewarding and exasperating endeavour, with Peter's insight and good humour reliably tilting the balance in favour of scientific inspiration.

I was fortunate to benefit from many talented colleagues in the Wellcome Sanger Institute's Cancer Genome Project, absorbing illuminating discussion as well as practical guidance in all matters biological, computational, statistical, and philosophical. In particular, I thank my office companions who helped kick-start my research in the early stages, including Dr Moritz Gerstung, Dr Inigo Martincorena, Dr David Wedge, Dr Kevin Dawson, and Dr Yilong Li. Yilong was especially instrumental as the original architect of the SV classifications underpinning much of this thesis, and I offer particular gratitude for his unfailing generosity in explaining the minutiae of genome rearrangement.

The ICGC PCAWG dataset analysed in this thesis was the result of many years of work by hundreds of researchers across several continents, and I thank the donors and consortium members—especially the structural variation working group—for the privilege of accessing this vast collection of high-quality cancer genome data.

On a personal note, this thesis would not have been possible without the love and understanding of my friends and family in England and Australia. Special thanks to Katie, Ellese, Mariel, and Alice, and to my parents—Barb and Tony—for their tireless support and encouragement, especially in the final six months!

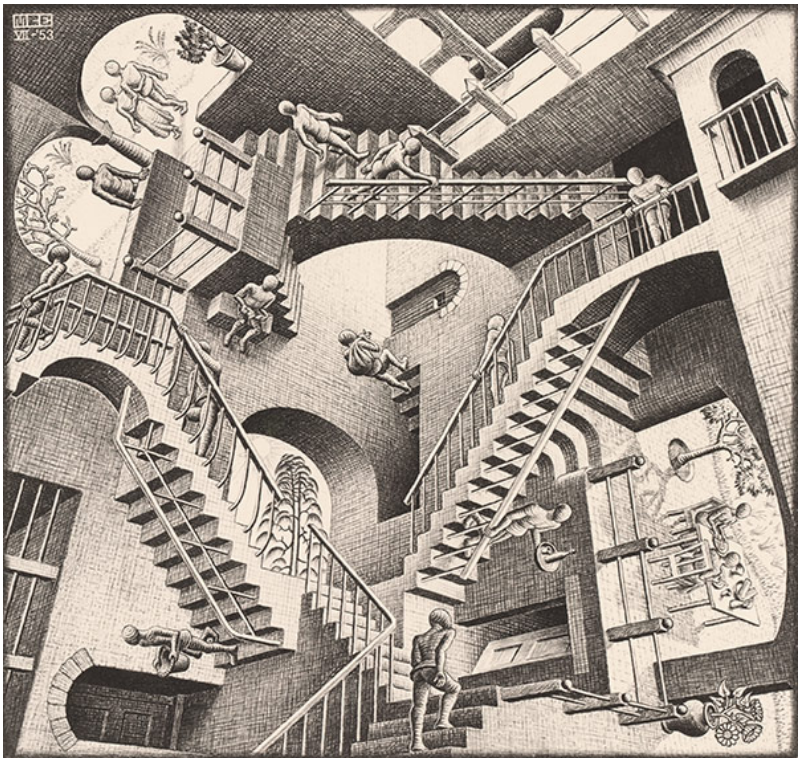
I also thank the Wellcome Trust for their financial support, and Trinity College for providing the social nexus of my studentship in Cambridge.

Clouds are not spheres, mountains are not cones, coastlines are not circles, and bark is not smooth, nor does lightning travel in a straight line... Nature exhibits not simply a higher degree but an altogether different level of complexity.

— *Benoit Mandelbrot*

I am progressing very slowly, for nature reveals herself to me in very complex forms; and the progress needed is incessant.

— *Paul Cézanne*



Contents

Chapter 1

Introduction to the cancer genome

The journey of an individual human genome begins with its formation in the fertilised egg—a chance meeting between maternal and paternal chromosomes in a totally unique combination, never to be repeated. After the normal germline genome is first established in the zygote, it faces the immediate prospect of copying itself into two daughter cells as faithfully as possible. Indeed, in each mitotic cell division through embryogenesis, infancy, and adulthood, a volley of biochemical activity operates to replicate and disseminate the six gigabases of inherited genome many millions of times over. Inevitably, occasional errors in DNA repair, replication, and segregation accrue with each cell division, and somatic genomes gradually diverge from their common ancestor in the zygote. A subset of these somatic mutations confer a selective advantage to the cell lineage, sometimes culminating in pathological unchecked cell growth broadly classified as cancer. With advances in whole genome DNA sequencing technology, somatic mutation in cancer samples can now be identified at base-pair resolution on any scale from single base substitution to rearrangement of kilobases, megabases, and whole chromosomes. In this thesis I analyse somatic rearrangement observed in more than 2500 cancer genomes from common cancer types all over the human body. The diverse structural patterns which emerge are testament to the complex bio-molecular challenges a genome may encounter in the course of its somatic evolution. By charting the landscape of possible genome configurations in the soma, we begin to understand the repertoire of genetic manoeuvres available to a cancer, and can better appreciate the underlying reasons for cancer's heterogeneous clinical presentation.

1.1 The somatic genome in mitosis and cancer

Throughout the mitotic cell cycle, the information content and structural integrity of the nuclear genome must be preserved and carefully promulgated to maintain regulated programs of cell behaviour and function. To this end, breaks or lesions in the DNA are repaired where possible, or may trigger cell death. The DNA content is replicated in S phase to produce sister chromatid pairs, which then condense and separate into opposite daughter cells during M phase. Errors in the dynamic orchestration of genome state generate mutations which transmit through the descendent cell lineage. Such genome alterations include single nucleotide variants (SNV), small insertions or deletions (indels), and a diverse range of larger structural variation (SV). Although most mutations have negligible fitness effects, some may confer a selective advantage driving clonal expansion into oncogenesis. (Stratton et al., 2009; Martincorena and Campbell, 2015; Tubbs and Nussenzweig, 2017)

1.1.1 DNA damage response

DNA lesions arise from endogenous and exogenous sources, including UV radiation, ionizing radiation, reactive oxygen species, chemical mutagens, and the inherent instability of biochemical molecules in a reactive environment. Different lesion types signal specialised DNA damage response pathways. For example, abasic sites and spontaneous deamination of 5-methylcytosine are repaired by base excision repair; pyrimidine dimers and bulky adducts by nucleotide excision repair; and incorrect DNA base-pairing by mismatch repair. Double-stranded DNA breaks may signal a variety of repair pathways, including non-homologous end-joining and homologous recombination, discussed further in Section 1.4. If the DNA injury is beyond repair, then the p53 pathway may trigger senescence or apoptosis to remove the cell from the population. When a DNA lesion is replicated without repair, or repaired incorrectly, mutations fix into the cell lineage. Some cancers have loss-of-function mutations in the genes controlling DNA repair, and develop a hypermutator phenotype as a result of compromised repair capacity. (Jackson and Bartek, 2009; Helleday et al., 2014; Tubbs and Nussenzweig, 2017)

1.1.2 DNA replication

DNA replication begins at many thousands of licensed origins[a], which 'fire' at different time points during S phase to recruit the replisome complex at two bi-directional replication forks. The replisome includes: helicase for separating parental duplex DNA into single stranded templates; topoisomerase for cutting the DNA backbone to release super-coil tension ahead of the fork and precatenane[b] structures behind the fork; polymerases for synthesising new DNA strands; and the DNA clamp PCNA for tethering the polymerase to the template strand. Different DNA polymerases have specialised roles in priming DNA synthesis and elongating nascent DNA along the leading and lagging strands[c]. The polymerases completing the bulk of replication have an inbuilt proof-reading domain and an estimated error rate of 10^{-7} mismatches per base. In contrast, the specialised translesion polymerases for replicating past DNA damage have lower fidelity, and are prone to incorporating small indels and SNVs. (Loeb and Monnat, 2008; Branzei and Foiani, 2010; Gaillard et al., 2015)

In addition to the small mutations caused by polymerase error, DNA replication can also generate larger structural variation through aberrant origin licensing, topoisomerase errors, and replication fork stalling and collapse. For example, inefficient origin licensing leads to incomplete replication and breaks in late-replicating regions, whereas unscheduled origin firing can lead to re-replication and fork collisions. Fork progression is also impeded by nucleotide pool depletion or physical obstacles such as DNA lesions or breaks, non-B DNA structures, or transcription bubbles. S phase checkpoint pathways respond to stalled forks and try to complete replication via translesion polymerases, template switching to the sister chromatid, or licensing of dormant origins. Failure to do so gives rise to double-stranded DNA breaks and subsequent error-prone repair. (Branzei and Foiani, 2010; Gaillard et al., 2015; Cortez, 2015)

[a] A 'licensed' replication origin is bound by helicases and the origin recognition complex during G1 phase, in preparation for active replication 'firing' during S phase.

[b] A precatenane is formed by sister DNA duplexes intertwining after synthesis.

[c] As DNA polymerases add new nucleotides to the free 3' hydroxyl group on the sugar-phosphate backbone, synthesis must proceed in a 5' to 3' direction. At the replication fork, leading strand synthesis is able to proceed continuously as it travels in the same direction as the opening fork. On the lagging strand, DNA is synthesised in discontinuous fragments building away from the replication fork and later joined through ligation.

1.1.3 Chromosome segregation

During interphase, the nuclear DNA spreads out to occupy large chromosomal territories with looping domain structures to regulate gene expression (Gibcus and Dekker, 2013). In preparation for mitotic cell division, the nuclear membrane breaks down as the chromosomes condense into their compact form, with sister chromatids initially still linked together via cohesin complexes (prophase). To achieve equal chromosome segregation, each chromatid in a sister pair must attach to kinetochore microtubules emanating from opposite spindle poles (metaphase). As the mitotic checkpoint proteins decay to signal successful spindle attachments, the cohesin disbands and sister chromatids are pulled to opposite poles (anaphase). In the final stages of telophase and cytokinesis, nuclear membranes reform around the two separated DNA masses, and the cellular membrane cleaves the cytoplasm to produce two daughter cells with equal chromosome content. (Hirano, 2015; Funk et al., 2016)

Errors in mitotic division can change the overall ploidy, and even be a root cause of DNA breaks and rearrangement. If cytokinesis fails to divide the replicated DNA into separate daughter cells, then the doubled genome content can persist in tetraploid state. If successful cytokinesis follows uneven chromosome segregation, then the abnormal chromosome count can persist in aneuploid state. Causes of chromosome missegregation include: mitotic checkpoint failure permitting premature entry into anaphase; cohesin defects causing sister chromatids to prematurely decouple or remain linked during anaphase; and aberrant kinetochore attachments (syntelic or merotelic) or centromere content (dicentric or acentric). These errors may pull both sister chromatids into the same daughter cell, or may result in DNA being caught between poles (either an entire lagging chromosome, or a smaller DNA section caught in a 'bridge'). Lagging or bridge DNA can be a substrate for large-scale rearrangement, as discussed further in Section 1.4. (Orr et al., 2015; Funk et al., 2016)

1.1.4 Genome and chromosome instability

In some cancers, the normal programs of DNA repair, replication, and mitotic segregation become so disordered that the cells develop persistent genomic and/or chromosomal instability. Genomic instability (GIN) refers to the continual generation of structural rearrangements *within* chromosomes, whereas chromosomal instability (CIN) refers specifically to unstable aneuploidy and a consistently high rate of chromosome missegregation.

Both instability phenotypes are associated with ongoing replication stress as a result of excessive DNA damage, excessive oncogenic transcriptional programs, or loss-of-function mutations in relevant genes (Burrell et al., 2013; Macheret and Halazonetis, 2015). Under these stress conditions, slow, stalled, or collapsed replication forks give rise to SVs and missegregating acentric or dicentric chromosomes. CIN is also possible in a competent replication background with compromised mitotic function. Although high rates of CIN are associated with cell death and tumour suppression, low rates of CIN are thought to be weakly tumour promoting, and provide a gradually diversifying genetic pool to facilitate adaptation. In the Mitelman cytogenetic database, 44% of solid tumours and 14% of blood cancers show evidence of CIN, while a further 42% (solid) and 58% (blood) have stable aneuploidies. (Zasadil et al., 2013; Funk et al., 2016)

1.1.5 Somatic mutations give rise to cancer

The hallmark properties of cancer include: sustained proliferative signalling and replicative immortality; evasion of growth suppression and cell death; and acquisition of invasive and metastatic abilities. These abnormal cellular properties are acquired via driver genome alterations, and thus somatic genome instability and mutation are considered an 'enabling' cancer hallmark (Hanahan and Weinberg, 2011).

Oncogenesis requires a small accumulation of driver events, with between two and ten currently identifiable in most cancer genomes (Vogelstein et al., 2013; Tomasetti et al., 2015; Martincorena et al., 2017; Sabarinathan et al., 2017). In general, oncogenes promoting cell growth are up-regulated by gain-of-function mutations, and tumour suppressor genes providing normal control and repair functions are down-regulated by loss-of-function mutations. Although most driver mutations are acquired in the soma, some may be inherited in the germline and increase the lifetime cancer risk (for example, *BRCA1* and *BRCA2* polymorphisms). Active mutagenic processes also generate a vast number of 'passenger' somatic alterations with no fitness benefit, thus confounding the search for genuine drivers in cancer genome sequencing studies (Pon and Marra, 2015).

1.2 Cancer genome sequencing projects

In a prescient opinion piece, Dulbecco (1986) predicted that an undertaking to sequence the human genome would yield invaluable insight into cancer biology. Despite being a stretch of blue-sky thinking at the time, his initial vision—to interrogate any gene of interest with probes designed off the reference—has long since been surpassed. The advent of affordable high-throughput DNA sequencing technologies ushered in a new field of cancer genomics research, with the first samples sequenced in their entirety by Ley et al. (2008) and Pleasance et al. (2010). Following this success, large collaborations within the International Cancer Genome Consortium (ICGC), The Cancer Genome Atlas (TCGA), and other local projects, set out to systematically catalogue genetic mutations in most common cancer types (International Cancer Genome Consortium et al., 2010; Cancer Genome Atlas Research Network et al., 2013; Wheeler and Wang, 2013). To date, research publications have summarised the genome landscape in dozens of patient cohorts, from the earliest reports characterising hundreds of *exomes* in ovarian and colorectal cancer (Cancer Genome Atlas Research Network, 2011; Cancer Genome Atlas Research Network, 2012) to more recent work analysing hundreds of *whole genomes* in breast cancer (Nik-Zainal et al., 2016) and medulloblastoma (Northcott et al., 2017), to cite just a few examples.

1.2.1 Study design

The classical study design for a cancer genome project is to sequence the bulk DNA of matched cancer–normal samples from a cohort of donors with the same or similar disease pathology (Mwenifumbo and Marra, 2013). Matching each cancer sample with normal DNA from the same individual[d] is critical for distinguishing somatic mutations specific to the cancer lineage from germline polymorphisms present in all tissues of the body.

To date, the vast majority of cancer genome projects have used the Illumina DNA sequencing platform. This technology sequences the last 100–150 bases of billions of DNA fragments by detecting the stepwise addition of fluorescently-labelled, reversibly-terminating nucleotides (Reuter et al., 2015). Sophisticated bioinformatics pipelines map these short reads (usually paired ends from a

[d]Normal DNA is usually taken from blood, or nearby non-cancerous tissue surgically extracted at the same time as the tumour. For blood cancers, the normal sample must be taken from an isolate of non-cancerous cell type/s (or another tissue if available).

fragment $< 1\,\text{kb}$ long) to their most likely origin in the reference genome, and identify variants which differ from the reference sequence.

So far, TCGA studies have primarily focussed on whole exome capture sequencing (WES), limiting high resolution findings to protein-coding regions covering less than 2% of the total genome. Studies by the ICGC and other groups are now turning to the more expensive whole genome sequencing (WGS) methods, which allow variation to be detected in non-coding regions and in the form of structural rearrangement. In addition to DNA sequencing, most cancer genome projects include complementary assays such as SNP arrays to detect copy number variation (CNV) and RNA-seq to quantify gene expression levels. (Mwenifumbo and Marra, 2013)

Moving beyond this traditional template of bulk DNA sequencing in matched cancer–normal pairs, other approaches to cancer genome interrogation include multi-sample, multi-region, and single-cell designs, combined with a burgeoning variety of new long-read and single-molecule sequencing technologies.

1.2.2 Insight from somatic SNVs

As a core output of both WES and WGS data with relatively simple properties to identify and analyse, the SNV has been the most intensively studied class of somatic genome alteration in the modern sequencing era. Analysis of somatic SNVs has yielded substantial insight into their underlying generative mechanisms (Alexandrov et al., 2013b; Helleday et al., 2014) and functional implications as driver events within genes (Kandoth et al., 2013; Lawrence et al., 2014) and, to a lesser extent, non-coding regions (Khurana et al., 2016). Patterns of SNV allele fraction have shed light on the sub-clonal phylogenetic evolution of tumours, and the relationships between primary and metastatic sites (Macintyre et al., 2016a; Schwartz and Schäffer, 2017). In concert with other -omics assays, SNV data has also been instrumental in describing the molecular subtypes of different cancer histologies (Hoadley et al., 2014; Bailey et al., 2016). Comprehensive studies of structural variation have been slower to emerge, partly because of the paucity of WGS relative to exome data, and partly because the complexity and variety of rearrangement events pose considerable analytical challenges. Section 1.4 outlines our current understanding of the somatic rearrangement landscape in human cancer.

1.2.3 Clinical translation

Efforts to characterise cancer genomes are motivated partly by the insight into molecular biology, and partly by the promise of clinical translation and improved patient outcomes. Findings from cancer genome studies are already proving their clinical worth, with at least eleven genetic alterations specifically targeted by FDA-approved therapies in ten different cancer types (as of early 2017), and dozens more genes on track for targeted drug development (Hyman et al., 2017). As diagnosis moves to incorporate molecular and genetic markers, new 'basket' clinical trials are beginning to test therapies by gene target in addition—or even in preference—to histology and tissue of origin. For example, drugs approved to target BRAF V600 mutations in melanoma may be used to treat other cancers with the same driver mutation (Hyman et al., 2015). Beyond precision therapies, detailed genomic profiling also improves prognostic accuracy (Ng et al., 2016; Gerstung et al., 2017), and has led to novel technologies for personalised medicine such as relapse monitoring of circulating cell-free DNA (Wan et al., 2017; Siravegna et al., 2017). Personalised, genome-driven oncology may soon be a routine addition to patient care, with the Genomics England initiative currently in progress to sequence whole genomes of 25,000 cancer patients in a clinical setting (Peplow, 2016; Genomics England, 2017).

1.3 Discovering rearrangements in the cancer genome

In addition to SNVs and small indels, somatic genomes also develop larger structural variants wherein kilobases, megabases, or whole chromosomes are deleted, amplified, or otherwise rearranged from the germline state. In this thesis I use the terms genome rearrangement and structural variation (SV) interchangeably. With the first deluge of cancer sequencing data over 2010–2015, publication of SNV analyses far outpaced those on SVs. However, long before high-throughput DNA sequencing and the focus on point mutations, cancer genomes were historically described in terms of large cytogenetic aberrations. As the cancer genomics field matures and the task of gleaning new insight from SNVs becomes harder, the time is right to refocus attention on somatic rearrangements, capitalising on the improved power and resolution afforded by WGS technology.

1.3.1 History of SV discovery in cancer

Advances in biotechnology have revealed several types of genome rearrangement. In the late 19th and early 20th centuries, David Paul von Hansemann and Theodor Boveri proposed the first chromosomal theories on the origins of cancer after observing abnormal chromosome content and asymmetric mitoses in tumour cells (contributions reviewed by Bignold et al. (2006)). As cytogenetic techniques improved, researchers visualised whole chromosome gains and losses (Spriggs et al., 1962), double minutes (Cox et al., 1965), translocations (Rowley, 1973), breakage-fusion-bridge cycles (Gisselsson et al., 2000), and megabase-scale deletions, insertions, and inversions (Sandberg, 1991).

One of the earliest successes from the cytogenetic era was the characterisation of the chr9;chr22 translocation causing the *BCR–ABL* oncogenic fusion gene in chronic myeloid leukaemia (Rowley, 1973) (Nowell (2007) recounts the history of its discovery). With the consequent development of targeted tyrosine kinase inhibitors, the life expectancy of CML patients is now comparable to the general population (Bower et al., 2016).

Moving beyond cytogenetic visualisation of M-phase chromosomes, the detection resolution for copy number alterations (CNA) was refined to a sub-megabase scale with the development of aCGH (Pinkel et al., 1998) and SNP arrays (Zhao et al., 2004; Bignell et al., 2004). CN array methods quantify the degree of copy loss or gain along the reference genome to a resolution of several kilobases, and are still commonly used to supplement WES studies (Zack et al., 2013). However, array technology cannot pinpoint the underlying events actually causing copy number change, and are powerless to detect copy-neutral rearrangement (with the exception of loss-of-heterozygosity (LOH) detectable by SNP array).

1.3.2 Somatic SVs in WGS data

Whole genome sequencing allows all rearrangement classes at any size[e] to be identified at base-pair breakpoint resolution (Korbel et al., 2007; Campbell et al., 2008). In addition to the many chromosome abnormalities identified in the cytogenetic era, sequencing data has revealed novel rearrangement patterns including chromothripsis (Stephens et al., 2011), chromoplexy (Berger et al., 2011; Baca et al., 2013), and chromoanasynthesis (Liu et al., 2011), described further in Section 1.4.2.

[e] SV detection below ~ 1 kb is poor if the read-group orientation is normal (deletion-type).

First generation SV calling algorithms (reviewed by Liu et al. (2015)) use reference-mapped paired-end reads to find groups of split[f] and/or discordantly mapping[g] read pairs which demarcate breakpoint junction positions. The range of possible SV detection methods continues to expand, with more than 20 published algorithms for short-read WGS data available as of late 2017. Some of the more recent contributions concentrate on:

- incorporating depth of coverage (copy number) (for example, SV-Bay finds likely breakpoints under a Bayesian model linking discordant read positions with concordant read depth (Iakovishina et al., 2016); COSMOS prioritises SV calls using strand-specific coverage (Yamagata et al., 2016));
- local assembly around purported breakpoints (for example, novoBreak assembles reads containing the same cancer-specific k-mers (Chong et al., 2016); SvABA assembles abnormal reads mapping to the same reference loci (Wala et al., 2017b)); and
- different ways of comparing matched cancer-normal samples to account for the germline SV background (for example, SMUFIN performs reference-free raw read comparison (Moncunill et al., 2014); PSSV estimates the joint probability of specific hidden genotype states (Chen et al., 2016)).

Regardless of the method, all algorithms are bound by the intrinsic limitation of read lengths being shorter than some repeat sequences, and have low power to detect SVs in ambiguous regions around telomeres and centromeres.

The core output from a standard SV caller is a set of breakpoint junctions (BPJ), each identifying two reference positions juxtaposed in a specified orientation. In addition, the nucleotide sequence detail can detect microhomology or small non-templated base insertions at each junction. WGS data also facilitates genome-wide CN estimation by segmentation of normalised read depth (reviewed by Liu et al. (2013)).

Ideally, a bioinformatics pipeline would also classify SV events by their broader structural context to distinguish simple events of one or two BPJ from medium complexity events of ∼3–9 BPJ or highly complex clusters of ∼10–1000 BPJ. So far, systematic SV classification in cancer WGS data has been largely confined to the basic orientation pattern of individual junctions (Yang et al., 2013; Zhuang and Weng, 2015; Alaei-Mahabadi et al., 2016). Some studies have

[f] A split read has a portion mapping to the reference location, with the remaining portion soft-clipped.

[g] A discordantly mapping read pair has non-standard orientations and/or a mapping distance inconsistent with the library insert size.

augmented this with one or two additional caveats by copy number, cluster separation, and/or broad classification of chromothriptic patterns (Patch et al., 2015; Nik-Zainal et al., 2016; Fraser et al., 2017).

1.4 Patterns of structural variation

The breadth of rearrangement observed in cancer sequencing data reflects the diverse range of DNA alteration that is not only possible, but evidently both consistent with and beneficial to cellular survival, even to the point of continuous pathological growth. Somatic SV catalogues are a window into the dynamics of genome upkeep, and hint at where and when different structural changes arise, whether in specific genome loci, cell types, genotype background, stage of tumour evolution and so on. However, the underlying mechanisms actually generating these rearrangements are not always obvious, and we rely on characteristic fingerprints such as microhomology and copy number profile to implicate known and undiscovered pathways of DNA damage and repair.

1.4.1 Mechanisms of repair and rearrangement at a DNA break

Genome rearrangements are generated by a variety of mechanisms, with many details still unknown. In general, they form during repair of double-strand breaks (DSB) caused by DNA damage, replication fork collapse, telomere attrition, or enzymatic activity. Free DNA ends are substrates for several possible processes, including resection, annealing, ligation, strand invasion, polymerisation, and telomere capture (Kasparek and Humphrey, 2011). DNA repair pathways employ these steps in varying combinations to secure ongoing genome integrity, even at the expense of some local rearrangement.

DSB repair pathways fall into two broad camps: 'break and ligate' mechanisms where two free DNA ends are pasted together; and 'template and replicate' mechanisms where one free end is extended through DNA polymerisation against a template sequence. For detailed reviews, see Willis et al. (2015), Ceccaldi et al. (2016), and Rodgers and McVey (2016).

In brief, the classic 'break and ligate' pathway of non-homologous end-joining (NHEJ) operates throughout the cell cycle (especially in G0/G1) to ligate blunt DNA ends. An alternative mechanism termed microhomology-mediated

end-joining (MMEJ) ligates slightly resected DNA ends[h] with a few bases of overlapping microhomology (MH). If heavily resected DNA ends share long (> 20 bp) homology, then single-stranded annealing (SSA) can stabilise their connection in new duplex DNA, and ligate the backbones after 3' flap digestion and DNA synthesis to fill in the gaps.

The classic 'template and replicate' pathway of homologous recombination (HR) operates during S and G2 phases of the cell cycle, and starts with strand invasion of a 3' single strand overhang to a template sequence with shared homology—preferably finding the sister chromatid for exact sequence preservation. Following strand invasion, DNA synthesis extends the nascent strand along the template, leaving the other strand displaced in a 'D-loop'. Somatic cells primarily resolve HR with synthesis-dependent strand annealing, in which the nascent strand is free to anneal to homologous sequence as it detaches from the template, and ideally finds its duplex partner on the opposing side of the original DSB to mediate error-free repair. An alternative form termed break-induced replication (BIR) continues synthesis of the invading strand in a migrating D-loop for many kilobases, proceeding until the D-loop destabilises or encounters the next obstacle (e.g. replication fork, transcription bubble, chromosome end). The stretch of newly synthesised single stranded DNA trailing from the D-loop is vulnerable to mutation, and is a probable substrate for APOBEC-mediated kataegis clusters[i]. In contrast to the established BIR model which relies on RAD51 homology search to initiate strand invasion, a RAD51-independent pathway termed microhomology-mediated break-induced replication (MMBIR)[j] appears to act in similar fashion, with the relaxed requirement of short MH between the invading and template strands. Indeed, the low-fidelity action of translesion polymerases may even facilitate MMBIR strand invasion in the absence of any pre-existing MH (Sakofsky et al., 2015).

DNA break repair mechanisms have a propensity to generate rearrangement structures through ligation of non-contiguous sequences, or inappropriate template choice and template switching. For example, stalled replication forks may trigger tandem duplication, either by end-joining of staggered breaks in two sister chromatids or re-replication bubble (break and ligate), or by strand invasion to the sister behind the original break locus (template and replicate)

[h]Enzymatic resection at the DSB leaves 3' overhanging single stranded DNA.

[i]Kataegis is a dense hypermutation cluster of ∼5–100 SNV. APOBEC is a family of cytidine deaminases which act on single stranded nucleic acid, with an important role in mutational disarmament of invading viral sequence.

[j]The MMBIR mechanism is also described in the literature as fork-stalling and template switching (FOSTES, Lee et al. (2007)).

(Finn and Li, 2013; Costantino et al., 2014; Willis et al., 2015). Likewise, deletions and translocations may be caused by aberrant end-joining of two DSB positions, or by strand invasion to a distant locus (Roukos and Misteli, 2014; Sakofsky and Malkova, 2017).

The repercussions of structural DNA repair and remodelling extend well beyond one or two break positions, and occasional bursts of genomic upheaval generate complex SV spanning tens or hundreds of breakpoint junctions.

1.4.2 Complex rearrangements

SV clusters arise from special cases of DNA breakage, and are not typically the mere overlap of simple events independently acquired.

Stephens et al. (2011) first described chromothripsis, characterised by dozens of BPJ shuffled together over one or more reference chromosomes with an oscillating copy number profile (Korbel and Campbell, 2013). This complex configuration results from a catastrophic shattering event, such as befalls lagging DNA caught in a micronucleus (Zhang et al., 2015) or chromatin bridge (Maciejowski et al., 2015) after aberrant mitosis. Subsequent ligation of a random combination of disjoint fragments generates a highly disordered derivative chromosome, with several fragments lost altogether.

Another 'break and ligate' pattern termed chromoplexy was first described in prostate cancer as a largely copy-neutral cycle of reciprocal exchange at multiple loci (Berger et al., 2011; Baca et al., 2013). The observed balancing of translocation partners across many chromosomes is hypothesised to result from correlated DSBs in spatio-temporal proximity, presumably mediated by androgen receptor activity in prostate.

Extrachromosomal DNA fragments generated by chromothripsis-type shattering events (or other means) often circularise to form double minutes (DM). These acentric DNA circles are free to segregate asymmetrically during mitosis, and are an efficient vehicle for oncogene amplification. DM copies can also re-integrate into the linear chromosome complement, forming intrachromosomal amplicon structures (also known as homogeneously staining regions). Internal DM composition may combine non-templated sequence insertions with small and large segments from several reference chromosomes, evolving through multiple rounds of integration and recombination. (Sanborn et al., 2013; L'Abbate et al., 2014; Vogt et al., 2014; Turner et al., 2017)

A different route to intrachromosomal sequence amplification is through successive breakage-fusion-bridge (BFB) cycles. In the classic model proposed by McClintock (1941), fusion of two atelomeric sister chromatids forms a dicentric chromosome which gets pulled apart during anaphase, passing a foldback SV (one-sided inversion) to one daughter cell. If multiple cell divisions repeat this cycle before the derivative is stabilised via telomere acquisition, then BFB imparts a characteristic foldback SV cluster with a step-like CN profile (Kinsella and Bafna, 2012; Greenman et al., 2016).

Break and ligate events—such as BFB, DM formation and chromothripsis—sometimes overlap to generate highly convoluted derivatives with little resemblance to their germline chromosome antecedents (Garsed et al., 2014; Li et al., 2014; Notta et al., 2016). Presumably, the inherent instability of some aberrant structures means that one large rearrangement may beget another, thus accounting for the prevalence of complex overlap observed in several cancers.

Replication mechanisms also generate complex SV via serial template switching, with distinctive patterns of copy gain, MH enrichment, and small, locally-templated insertions in the junctions between more distal BPJ (Lee et al., 2007; Zhang et al., 2009). These events have primarily been described in germline developmental disorders, and range from medium complexity SV like the duplication–inverted triplication–duplication (Carvalho et al., 2011), to high complexity events involving five or more BPJ termed chromoanasynthesis (Liu et al., 2011), possibly triggered by interstrand crosslinks or other persistent DNA lesions (Meier et al., 2014). Experimental studies support a MMBIR mechanism (Sakofsky et al., 2015; Hartlerode et al., 2016) with low-fidelity polymerases also generating nearby SNVs and indels (Carvalho et al., 2013).

1.4.3 Prevalence and distribution across the genome

The character and extent of somatic rearrangement is highly variable, depending on the fidelity of replication, rate of DNA breakage, choice of repair pathway, and subsequent effectiveness of that repair. WGS data indicate that most cancer samples have tens to hundreds of detectable BPJ, with the burden varying by an order of magnitude both across and within cancer types, from highly rearranged breast and ovarian genomes, to relatively stable genomes in kidney and thyroid cancer (Yang et al., 2013; Alaei-Mahabadi et al., 2016). Some cancers present with a strong tandem duplicator phenotype, especially those breast and ovarian cancers with both *BRCA1* and *TP53* mutations (Menghi

et al., 2016). Moreover, bone and soft-tissue cancers are particularly prone to chromothripsis (Stephens et al., 2011; Cai et al., 2014), while prostate cancer is notable for the prevalence of chromoplexy (Baca et al., 2013). The observation that most somatic BPJ have no or micro (1–5 bp) junction homology suggests that NHEJ, MMEJ, and MMBIR are the major pathways to cancer rearrangement, while non-allelic HR is largely confined to germline disorders (Drier et al., 2013; Malhotra et al., 2013; Yang et al., 2013; Carvalho and Lupski, 2016).

The variable forces of DNA breakage and repair not only dictate the *number* of BPJ per sample, but also their *location* in the genome. In B cells, deliberate enzymatic DSB generation renders immune loci particularly prone to translocation, often contributing to oncogenic fusions (Vaandrager et al., 2000; Alt et al., 2013). In prostate, androgen receptor signalling leads to topoisomerase DSBs in specific regulatory locations, often triggering the *TMPRSS2–ERG* fusion driver (Lin et al., 2009; Haffner et al., 2010). Retrotransposons are another source of recurrent SV, with particular L1 hotspots generating dozens of somatic insertion/transduction events in some cancers (Lee et al., 2012; Tubio et al., 2014; Helman et al., 2014). Common fragile sites are recurrent foci of deletion in many cancer types, associated with low density of replication forks, late replication time, large genes, and active transcription (Ozeri-Galai et al., 2012; Sarni and Kerem, 2016; Glover et al., 2017). Aside from these rearrangement hotspots, BPJ also correlate more generally with: spatial proximity inside the nucleus (Fudenberg et al., 2011; Hakim et al., 2012; Zhang et al., 2012); replication timing[k] (De and Michor, 2011; Pedersen and De, 2013); simple repeats (Bacolla et al., 2016); chromatin modifications (Black et al., 2013; Burman et al., 2015); and show sample-specific association patterns (Drier et al., 2013).

1.5 Functional consequences of rearrangement

Rearrangement landscapes observed in clinically-detectable cancer samples reflect the distribution of events at generation, moulded by selection on the functional consequences. Events which substantially reduce cell fitness are subject to purifying selection, and are not typically observed. Conventional theories posit that most somatic mutations are passenger events with negligible fitness effect, and that only a handful of positively-selected drivers are responsible for clonal expansion of the cancer lineage. A high passenger-to-driver ratio is well substantiated for point mutations (Tomasetti et al., 2015; Martincorena et al.,

[k]Replication timing tends to be late for copy loss, and early for both copy gain and LOH.

2017; Sabarinathan et al., 2017), and presumably extends to most SV classes as well. As a probable exception to this general paradigm, those complex SV events that restructure hundreds of megabases effect such a drastic departure from the normal diploid genome that passenger status seems unlikely.

Rearrangements drive the cancer phenotype through various means, including production of oncogenic fusion genes, amplification of oncogenes, deletion or disruption of tumour suppressors, and repurposing of regulatory regions. These alterations play a major role in cancer development, with COSMIC curating 73% of 547 census cancer genes as being affected by translocation or CNA (v71, (Forbes et al., 2015)). Even with the additional insight provided by RNA-seq data, it remains extremely challenging to distinguish the driver SVs from the passengers, and to discern which of the many changes to genes and/or regulatory elements meaningfully contribute to oncogenesis.

1.5.1 Fusion genes

Some rearrangements create fusion genes by placing one gene (or part thereof) downstream of a different promoter region (with or without the 5' end of the promoter's native open reading frame). Fusion genes drive cancer by placing a proto-oncogene under the control of a highly active promoter, or by the translation of a chimeric protein product with novel oncogenic properties. (Mertens et al., 2015)

Any SV class is capable of generating a fusion gene via the juxtaposition of non-contiguous sequences. For example, the *BCR–ABL* fusion driving chronic myeloid leukaemia is generated by translocation (Salesse and Verfaillie, 2002; Nowell, 2007); *KIAA1549–BRAF* in pilocytic astrocytoma is generated by tandem duplication (Jones et al., 2008); whereas *TMPRSS2–ERG* in prostate cancer is fused through deletion or chromoplexy (St John et al., 2012; Baca et al., 2013).

1.5.2 Gene dosage

A gene's transcriptional output is roughly correlated with its copy number in the genome (Fehrmann et al., 2015), and thus SV events generating regions of copy gain or loss may drive cancer by oncogene over-expression or tumour suppressor haploinsufficiency or two-hit loss. Roughly 40 peak regions of recurrent CNA span a known cancer gene (Beroukhim et al., 2010; Zack et al., 2013), such as

the *MYC* oncogene amplified in 13–17% of all breast and ovarian cancers and the *CDKN2A* tumour suppressor lost in 33% of brain cancers[l].

Regions of copy alteration often span multiple genes, and may drive cancer through the combined fitness effect of their synchronous dosage change. In the maximal case, whole chromosome or arm-level aneuploidy simultaneously alters the copy level for hundreds of genes. Some arms are strongly biased towards gain (e.g. 7p, 8q, 20q) or loss (e.g. 9p, 13q, 17p), reflecting the uneven distribution of tumour promoting or suppressing regions (Beroukhim et al., 2010; Kim et al., 2013; Davoli et al., 2013). Considering a smaller scale of several megabases, Liu et al. (2016) reported that the selective advantage of *TP53* tumour suppressor loss is boosted by co-deletion of neighbouring genes. Likewise, Hagerstrand et al. (2013) described the joint amplification in 3q26 of two oncogenes promoting cell growth and invasion. Beyond the single-copy gains proffered by aneuploidy or tandem duplication, the most efficient route to high-magnitude amplification is via extrachromosomal DMs, frequently boosting oncogenes like *MYC* and *EGFR* to CN levels above ten (Turner et al., 2017).

Amplifying enhancer[m] dosage is another route to oncogene over-expression, without necessarily changing the copy level of the gene itself (Zhang et al., 2016; Glodzik et al., 2017).

1.5.3 Altered regulation

Interphase chromosomes are organised in a looping architecture of topologically associating domains (TAD) which divide the linear sequence into self-interacting blocks (typically hundreds of kilobases) with coordinated gene expression and replication timing. TADs are physically separated from their neighbours by insulating boundary regions held together by CTCF and cohesin. Within a TAD, DNA looping allows enhancer elements to recruit transcription factors for genes up to a megabase away. DNA looping also ensures that enhancers are typically restricted from accessing and regulating genes in any separate TAD. Although TAD boundaries are conserved across cell types (and even species), the TADs themselves are dynamic units, localising in either active or repressive nuclear compartments to regulate tissue-specific gene expression programs. (Pombo and Dillon, 2015; Ruiz-Velasco and Zaugg, 2017)

[l]CNA statistics from the COSMIC database (Forbes et al., 2015); other cancer types not specified are also commonly affected by CNA at *MYC* and *CDKN2A*.

[m]Enhancer elements are *cis*-acting regulatory regions which recruit transcription factors to promote expression of genes brought in to proximity by DNA looping.

Mouse models show that SV events which duplicate or delete TAD boundaries result in merged or neo-TAD structures. Such alterations place genes in a novel regulatory context, drastically changing their expression levels with potentially serious phenotypic consequences. (Lupiáñez et al., 2015; Franke et al., 2016)

Chromatin topology remodelling and ectopic enhancer activity has also been observed in cancer, with the capacity to activate oncogenes and down-regulate tumour suppressors (Valton and Dekker, 2016). Early findings highlighted recurrent 'enhancer-hijacking' rearrangements up-regulating *EVI1* (alias *MECOM*) in acute myeloid leukaemia (Gröschel et al., 2014) and *GFI1A/B* in medulloblastoma (Northcott et al., 2014). Weischenfeldt et al. (2017) surveyed over 7000 cancer samples to find more than a dozen oncogenes likely to be activated in this manner, including the *IGF2* gene recurrently involved in a boundary-spanning tandem duplication in colorectal cancer. This simple SV event generates a neo-TAD structure linking *IGF2* with an active super-enhancer from the neighbouring region (usually insulated from each other by the boundary), causing an oncogene expression increase of more than 250-fold.

Given the immense influence of enhancer contact on gene regulation, TAD-disrupting SV events can drastically affect genes as far as a megabase from the breakpoint, irrespective of any fusion or dosage changes. The ability of rearrangements to transmute the chromatin organisation domains so faithfully preserved across tissues and species is now emerging as an under-appreciated pathway to the cancer phenotype.

1.6 Overview of this work

In this thesis I analyse somatic genome rearrangements within 2559 samples from the ICGC Pan-Cancer Analysis of Whole Genomes dataset, focussing on structural classes and properties (Chapter 2), the genome-wide distribution pattern (Chapter 3), co-occurrence signatures of underlying process (Chapter 4), and complex SV intractable to simple classification (Chapter 5).

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.

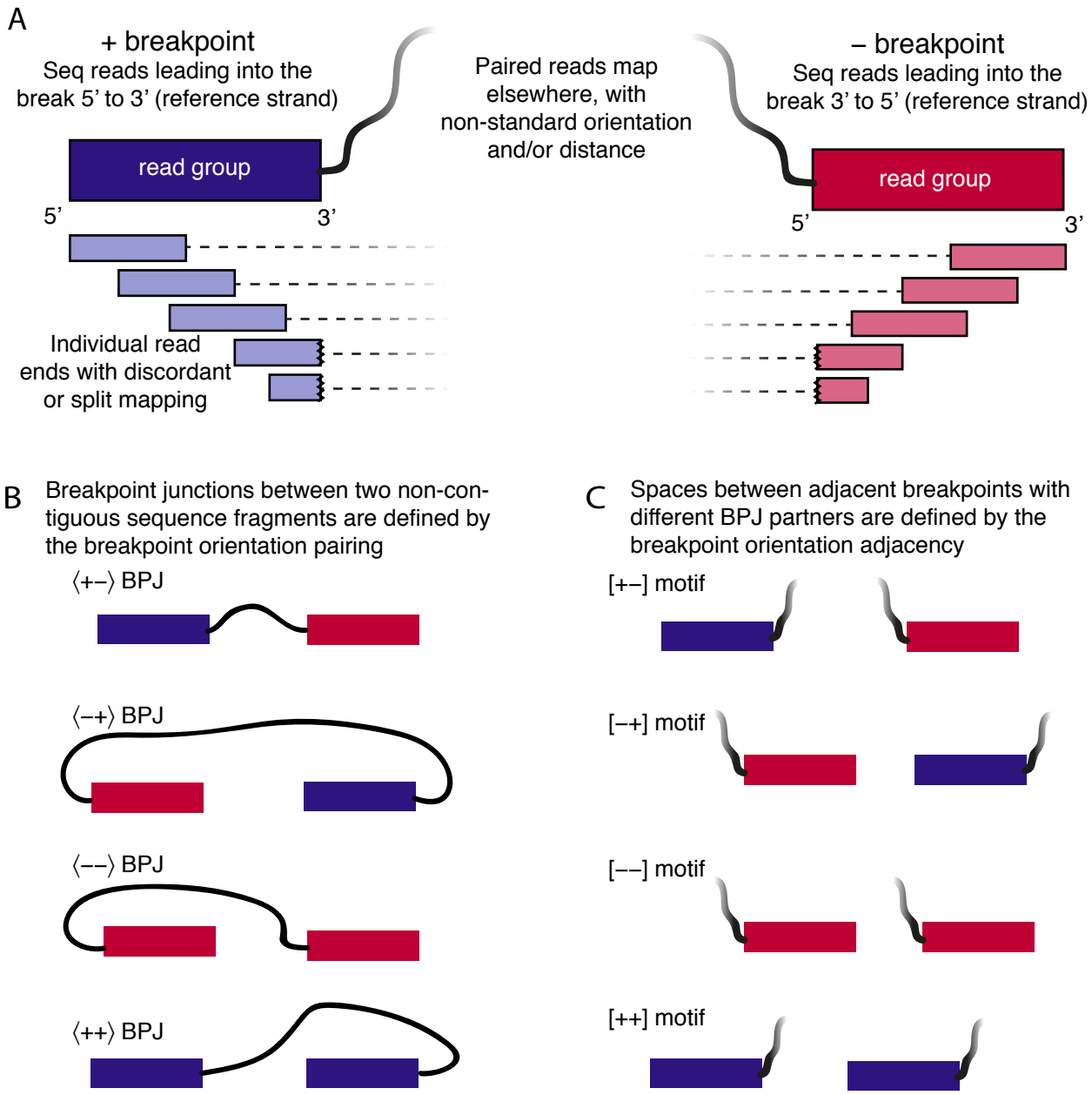


Figure 2.1: (A) Breakpoints of structural variation may have either a + or − orientation, yielding (B) four possible orientations for (intra-chromosomal) breakpoint junctions connecting two non-contiguous sequence fragments in the SV event, and (C) four possible motifs between adjacent breakpoints belonging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

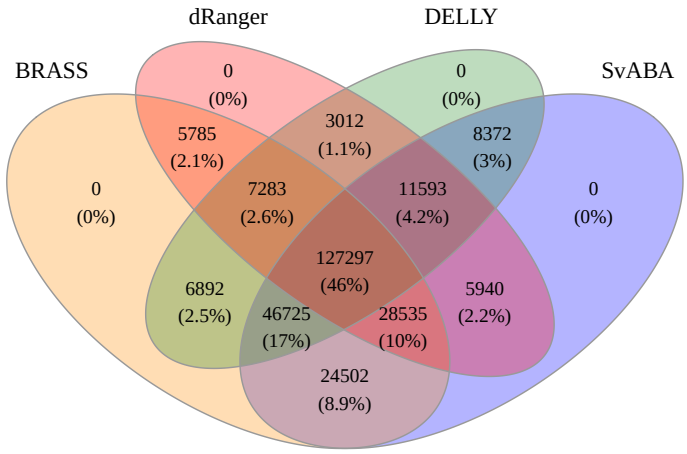


Figure 2.2: Overlap between consensus breakpoint junctions returned by the four SV calling algorithms in the PCAWG dataset.

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+ Distant 2-Jump	Trans w/ Foldback	distant BPJ adjoining $\langle ++ \rangle$ or $\langle -- \rangle$ BPJ w/ $[-+]$ motif	580
	Trans w/ InvIns	distant BPJ intersecting $\langle ++ \rangle$ or $\langle -- \rangle$ BPJ w/ $[-+]$ motif	508
	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	Cplxy Cycle	loop of $[+-]$ motifs	326
	Cplxy Chain	chain of $[+-]$ motif/s	366
	Cplxy Cycle w/ Ins	loop of $[+-]$ and $[-+]$ motifs	162

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

[a]Plotting footprints are different to the classification footprints described in Section 2.1.3

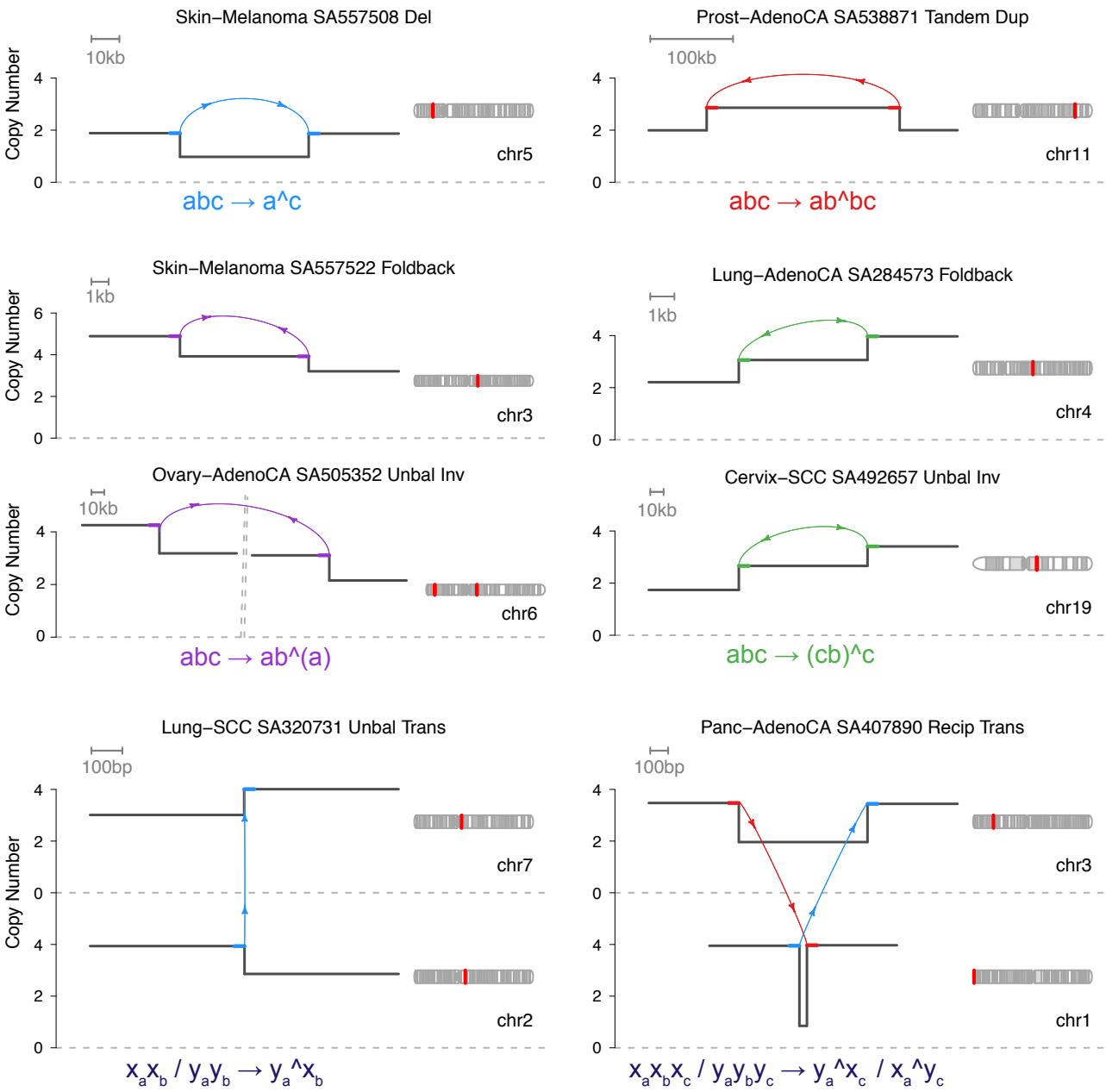


Figure 2.3: Example plots of the simple SV event classes: deletion, tandem duplication, foldback, unbalanced inversion, unbalanced translocation, and reciprocal translocation. The transformation between germline segment order and the somatic rearrangement is annotated below, with carets to denote breakpoint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

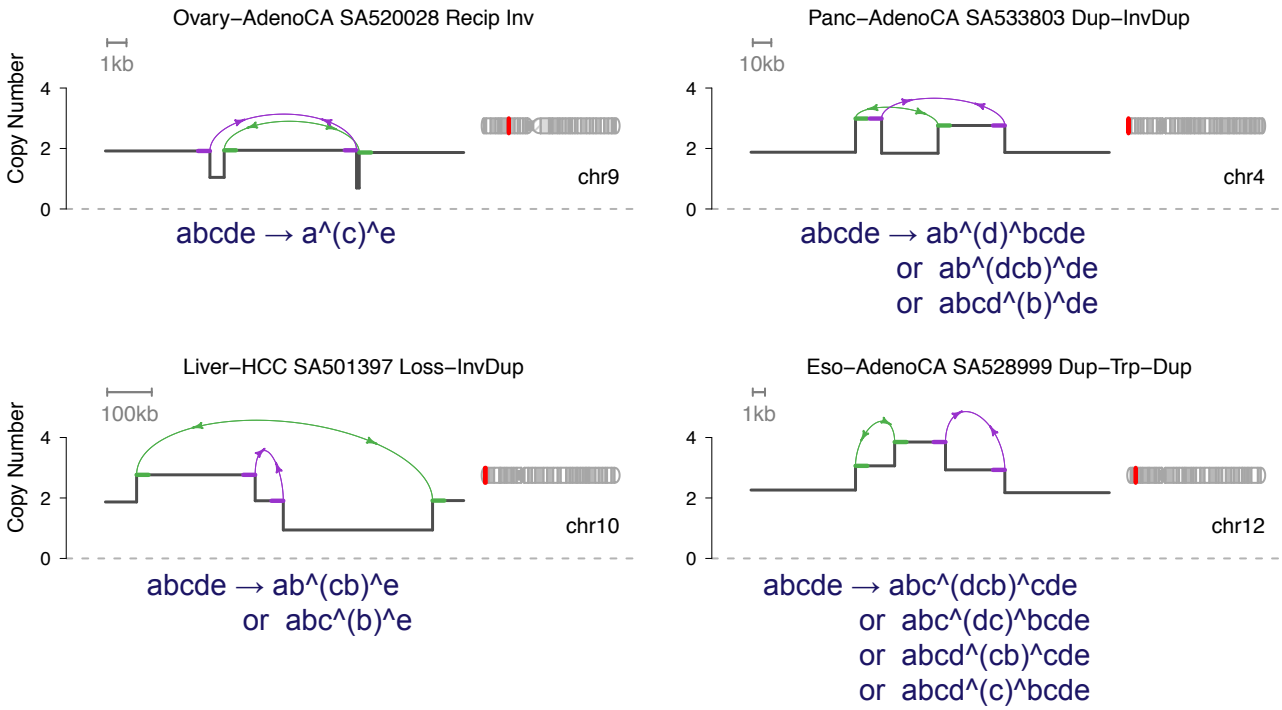


Figure 2.4: Example plots of reciprocal inversion and the three types of local 2-jump: duplication–inverted duplication, loss–inverted duplication, and duplication–inverted triplication–duplication. The transformation between germline segment order and the possible rearranged derivative structures is 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

[b]Considering 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

[c]In the order moving 5' to 3' along the reference strand, left to right in plotting space.

[d]Comparing against the library of possible overlap patterns generated by Yilong Li; more details in Li et al. (2017).

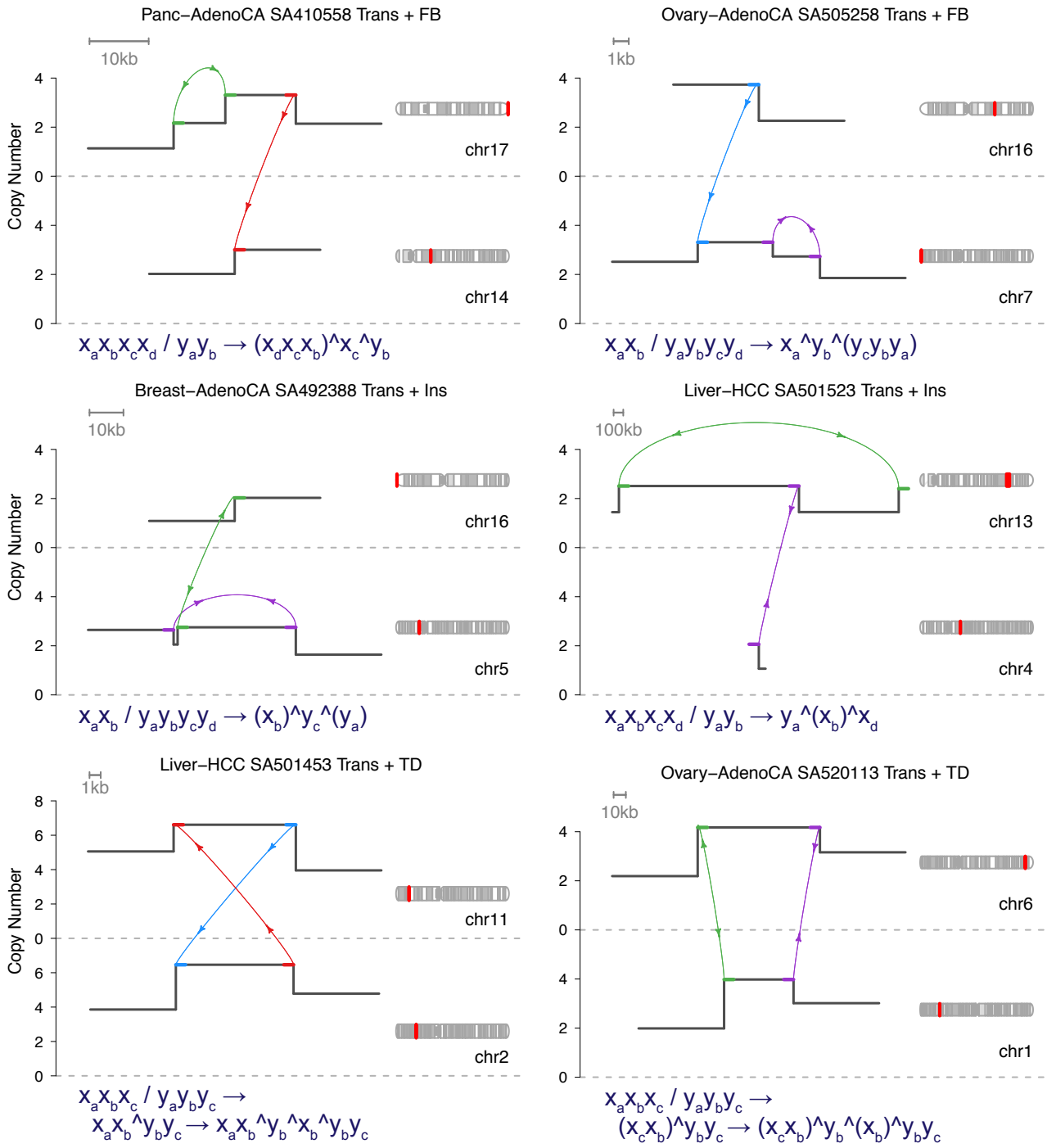


Figure 2.5: Example plots for three classes of local + distant 2-jump: translocation with foldback; translocation with inverted insertion; and translocation with overlapping tandem duplication. The transformation between germline segment order and the somatic rearrangement is annotated below, with carets 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.

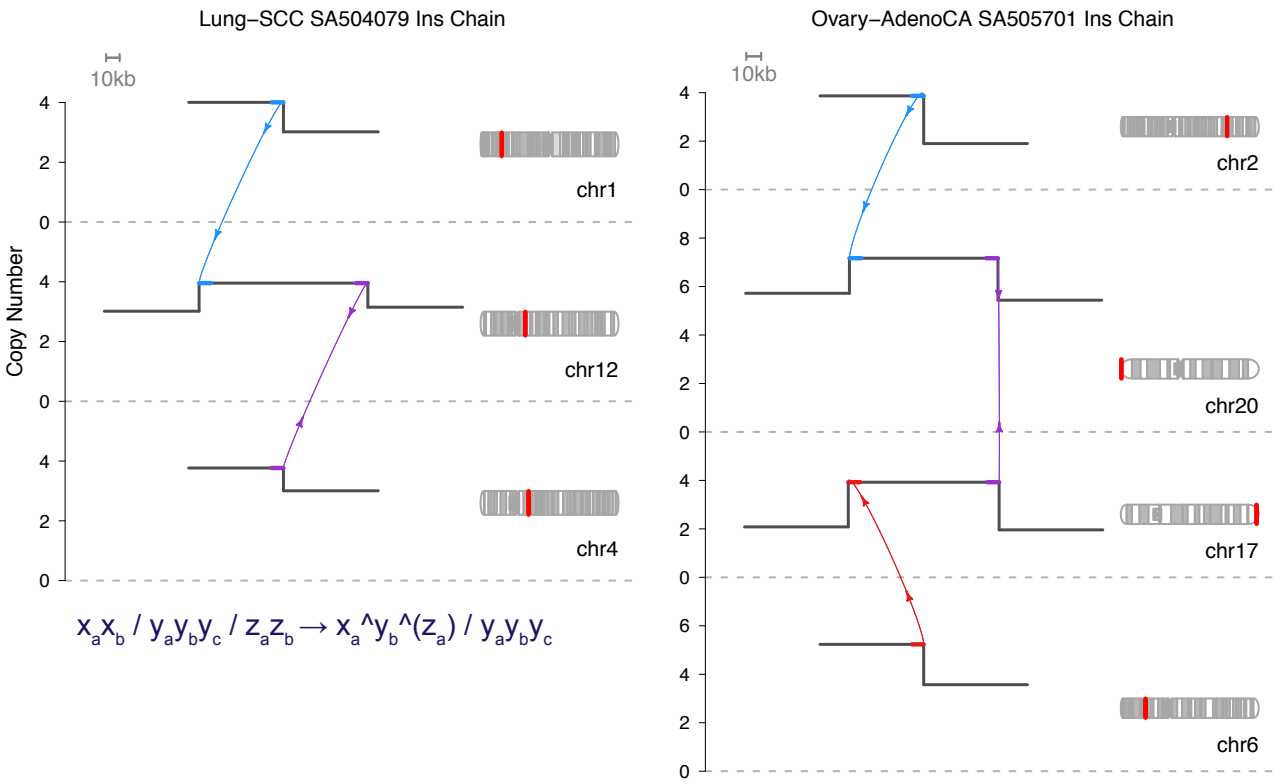


Figure 2.6: Example plots for chains of templated insertion, where two distant loci are joined by one or more templated inserts ([−+] motif). The transformation between germline segment order and the somatic rearrangement is annotated for the simplest example, with carets to denote breakpoint junctions 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.

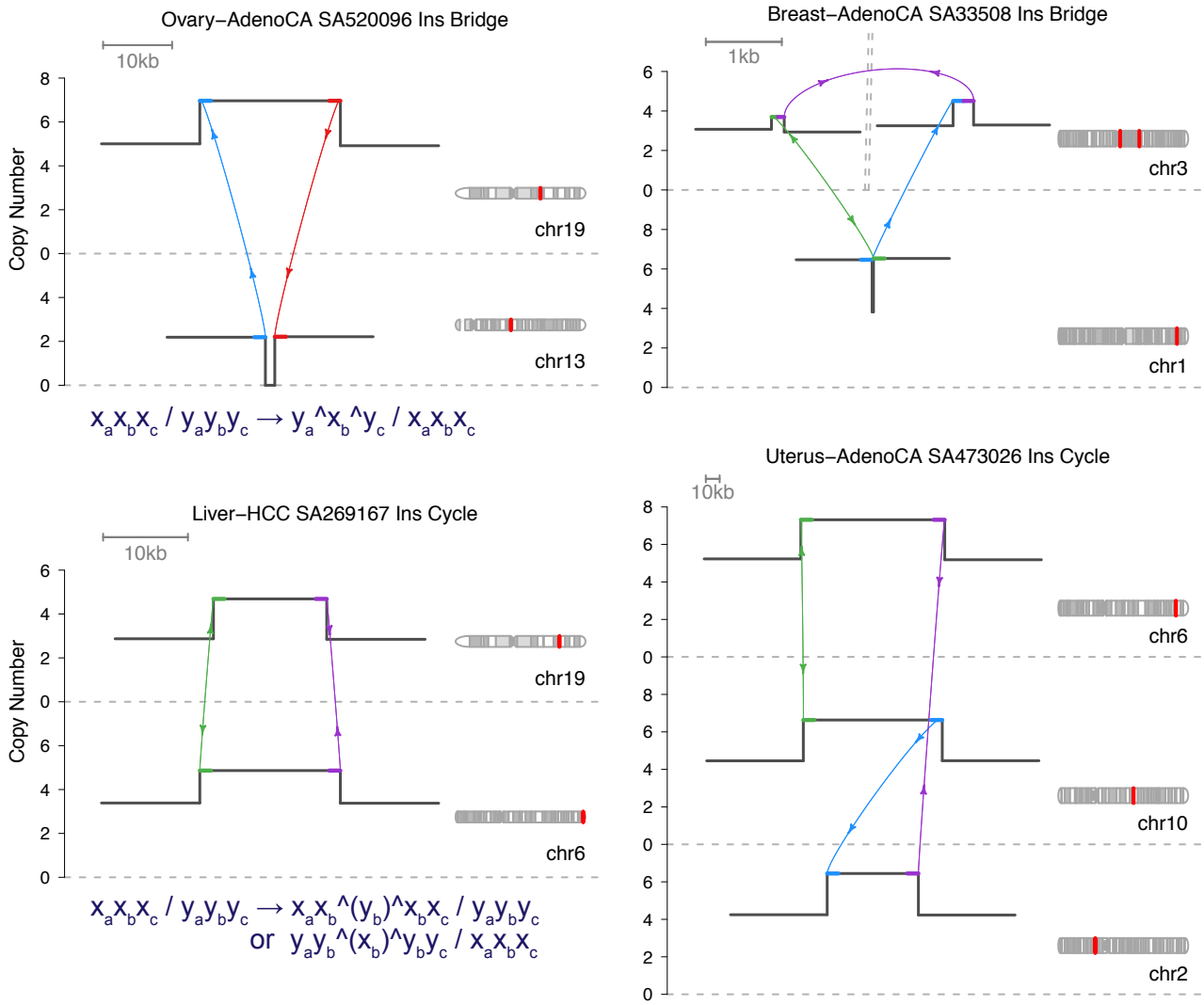


Figure 2.7: Example plots for bridges and cycles of templated insertion ([−+] motif). The 'bridge' events insert one or more templated inserts into a gap ([+−] motif) on the host chromosome. The 'cycle' events insert one or more templated inserts between a local duplication on the (unknown) host chromosome.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

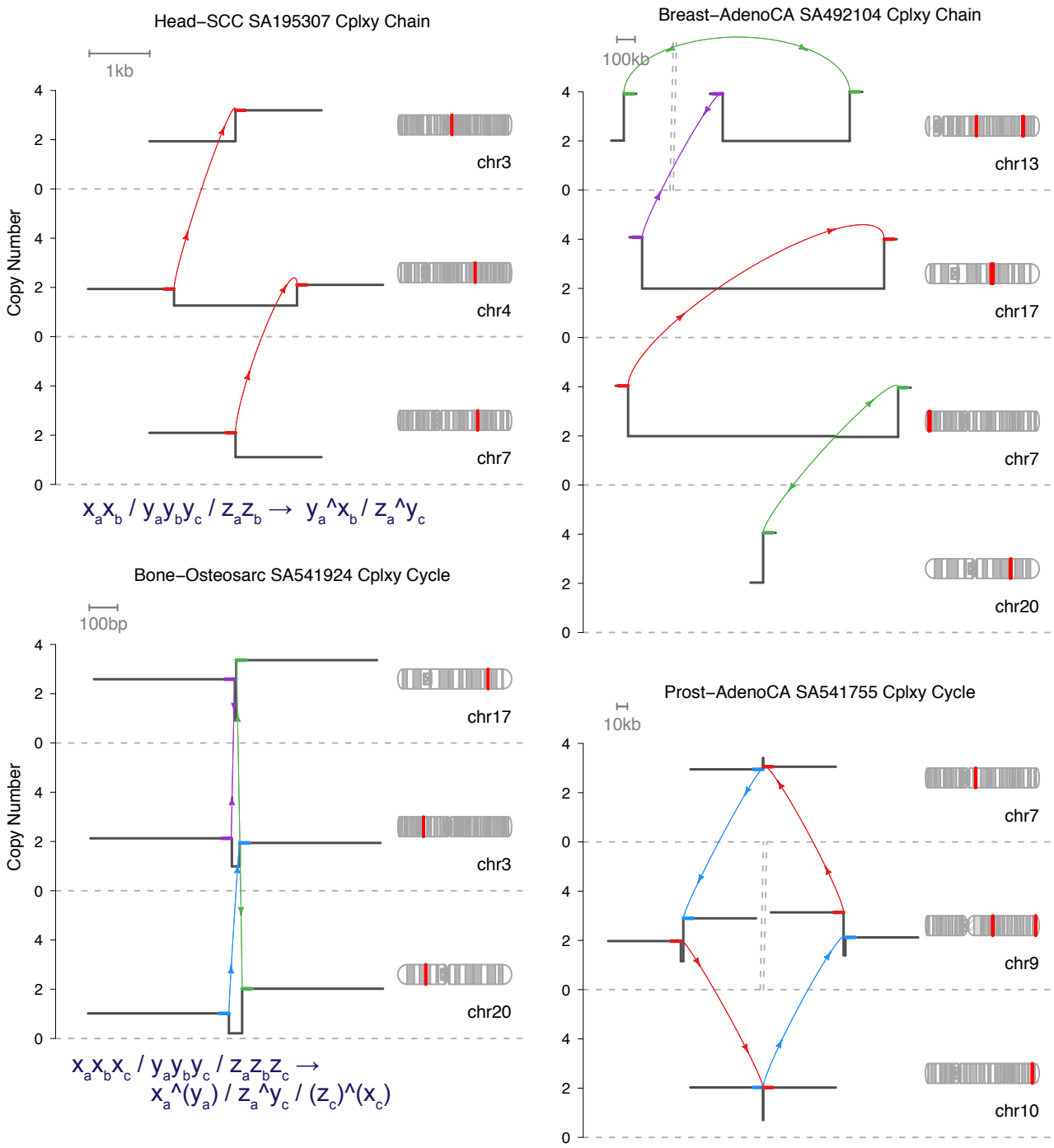


Figure 2.8: Example plots for chains and cycles of chromoplexy. The 'chain' events involve one or more footprints of balanced translocation ([+−] motif) that start and end in isolated breakpoints (unbalanced translocation). The 'cycle' events involve three or more footprints of balanced translocation ([+−] 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 of 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, myeloproliferative 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%

[e]The benign bone cancers include cartilaginous neoplasm, osteoblastoma, and osteofibrous dysplasia.

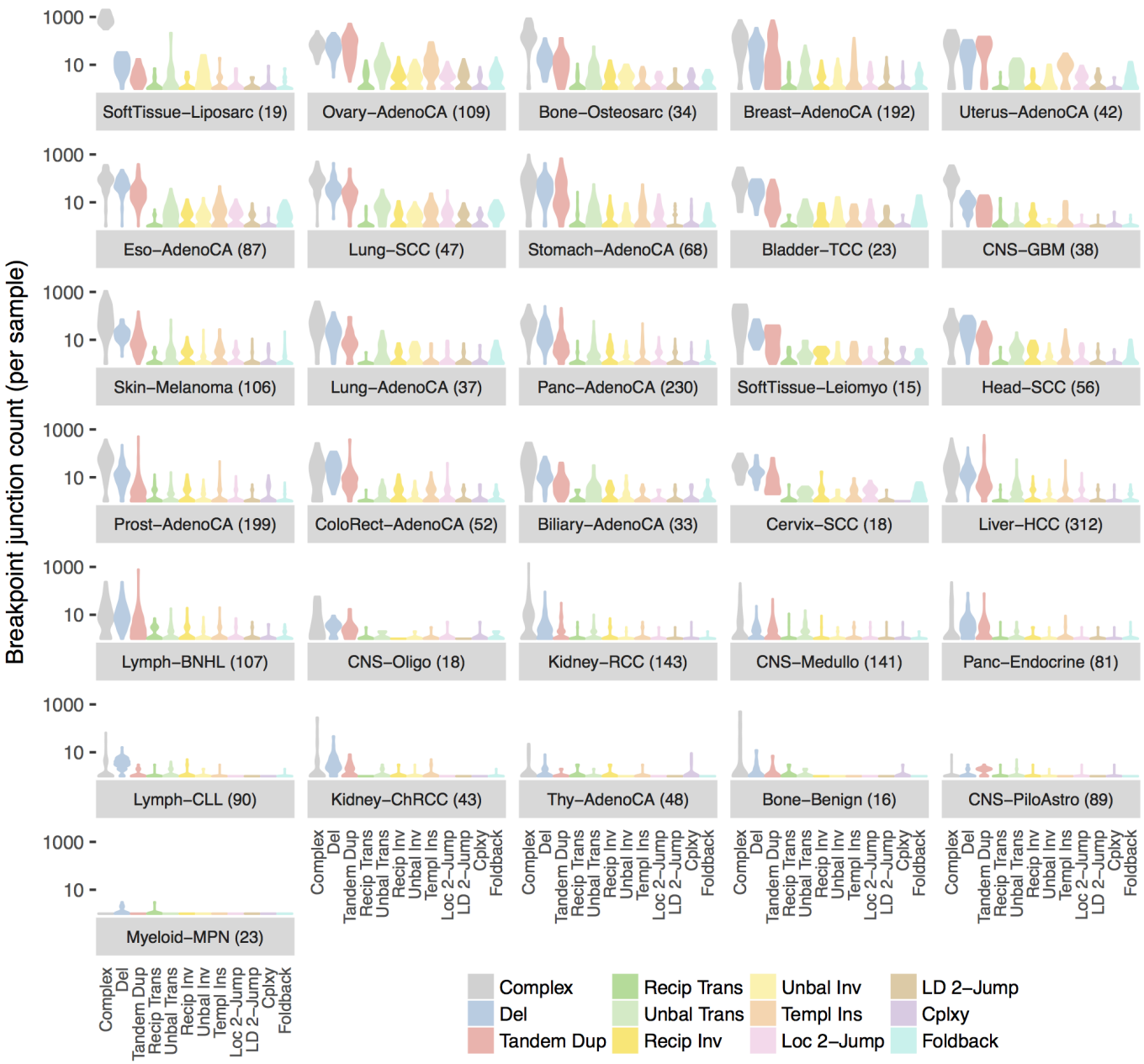


Figure 2.9: The number of classified breakpoint junctions across samples grouped by cancer histology, with the number of samples indicated in parentheses. Histology groups are sorted by the median number of BPJ per sample.

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.

[f]Two-sided binomial test against 0.085 null hypothesis, reporting significant results below 0.001 Benjamini–Hochberg-corrected FDR.

[g]Although 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.

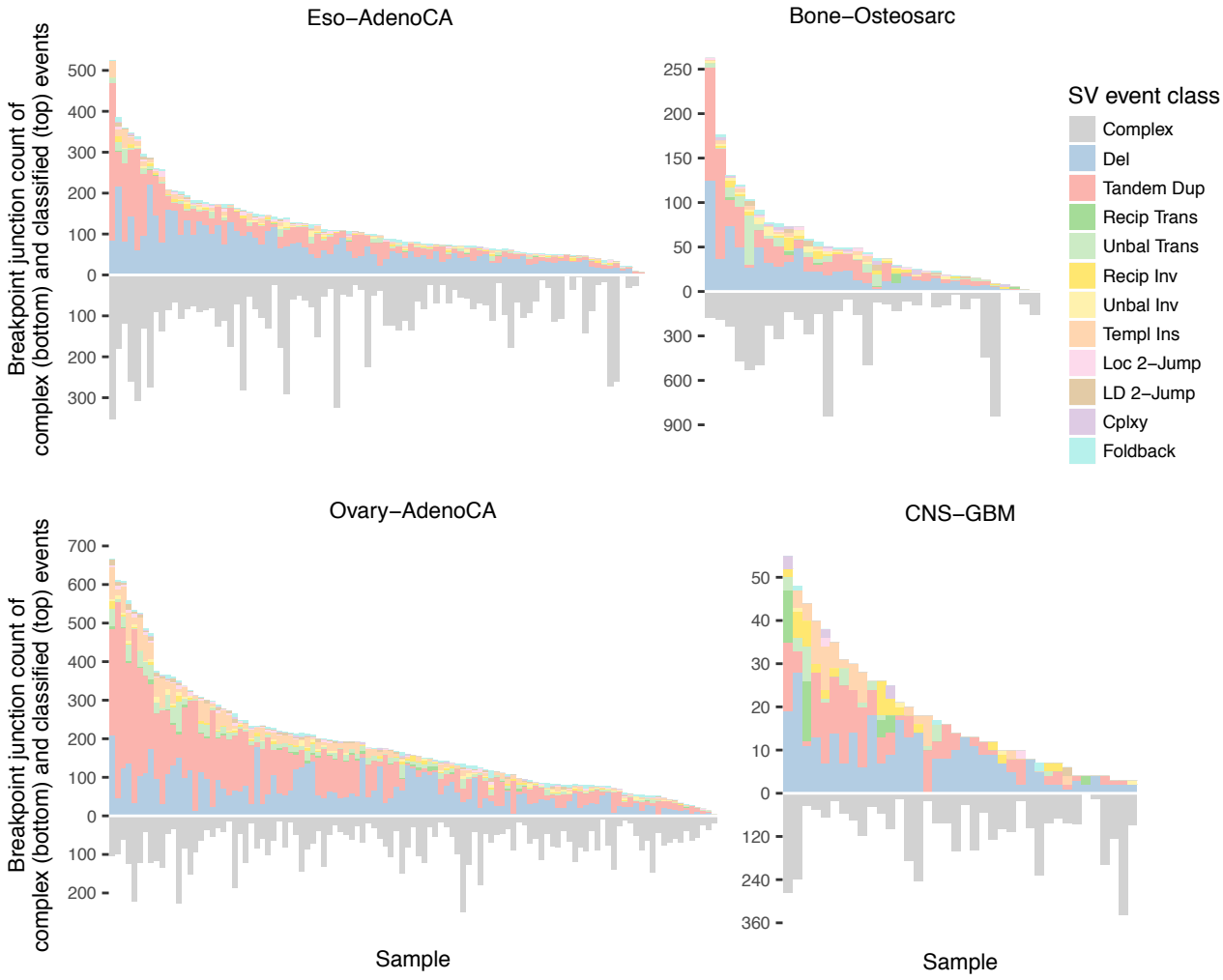


Figure 2.10: Per-sample counts of complex (lower) and classified (upper) breakpoint junctions for esophageal adenocarcinoma, osteosarcoma, ovarian adenocarcinoma, and glioblastoma multiforme. The lower plot for complex BPJ is on a different scale to the upper plot for classified BPJ.

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.

[h]The 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×) than the sub-cohorts from the USA (range 55–80×) or the Japanese National Cancer Centre (range 33–126×).

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.

		Classified BPJ			Complex BPJ		
Histology	Variable	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.

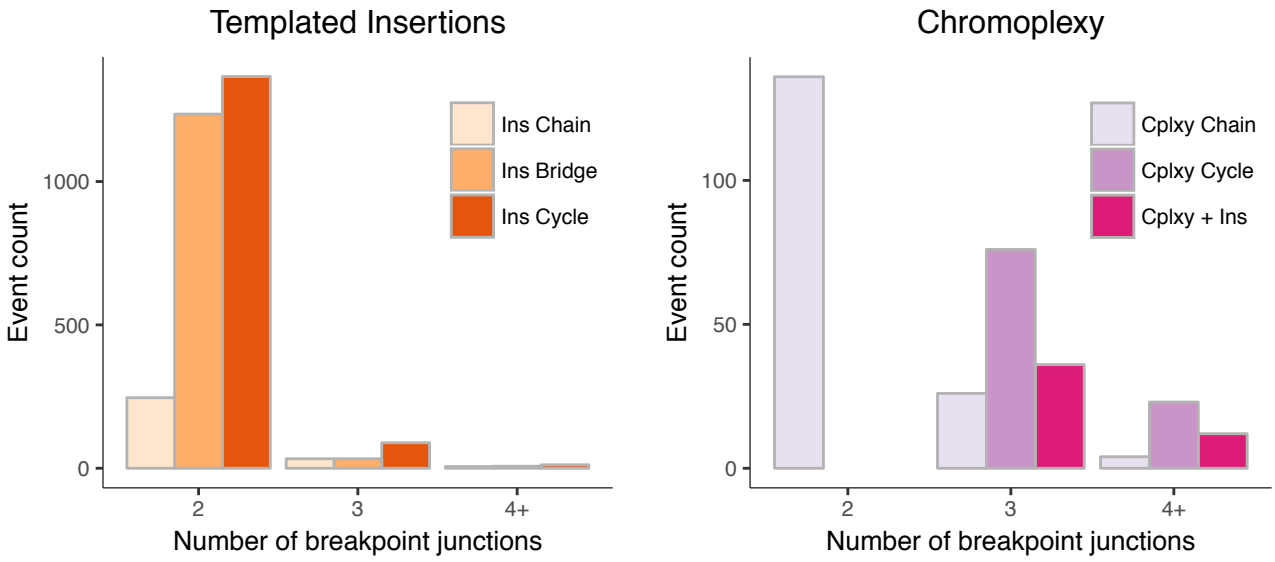


Figure 2.11: Number of BPJ in all templated insertion and chromoplexy events. Only chromoplexy *chains* have two BPJ because the minimal chromoplexy cycle is classified as reciprocal translocation and the minimal chromoplexy plus insertion cycle is classified as insertion bridge.

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.

[i]None of the four samples with these long templated insertions are annotated with any germline or somatic mutations or copy loss affecting *BRCA1* or *BRCA2*.

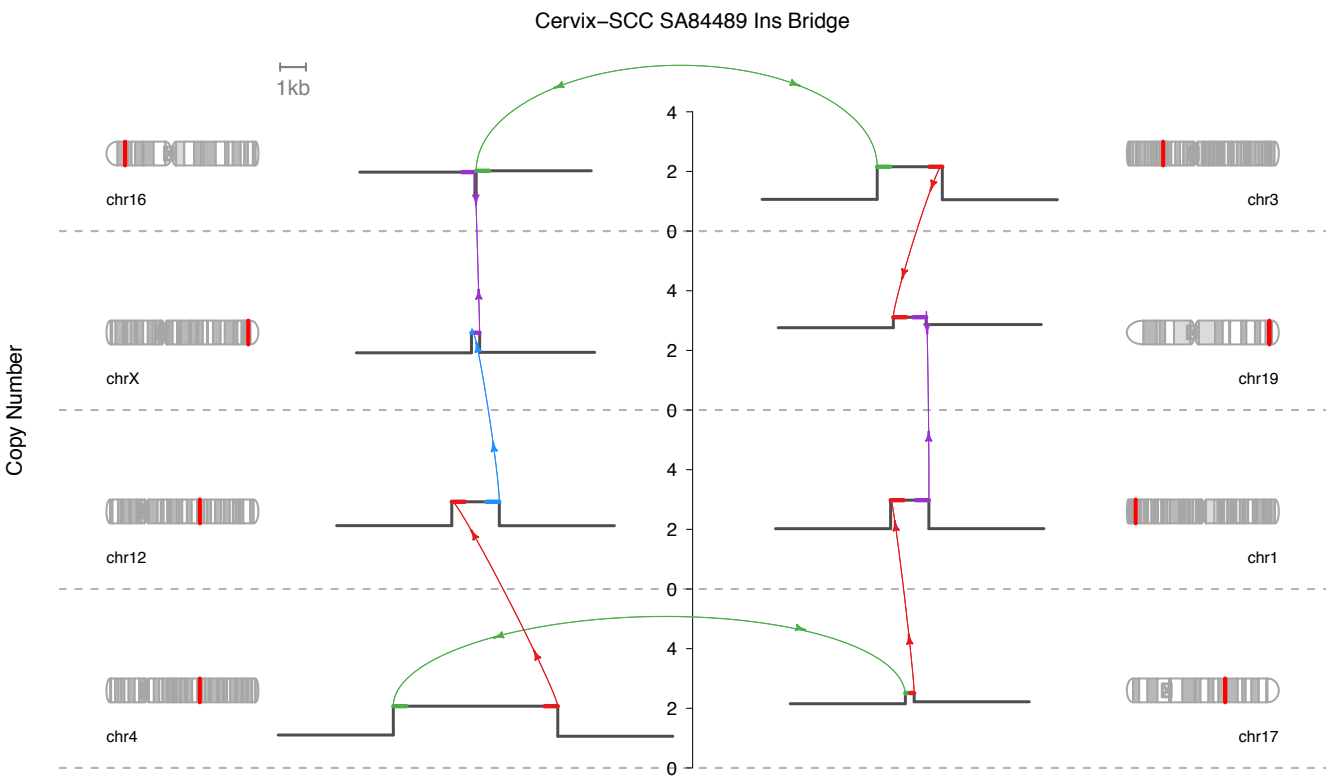


Figure 2.12: Longest templated insertion bridge event

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).

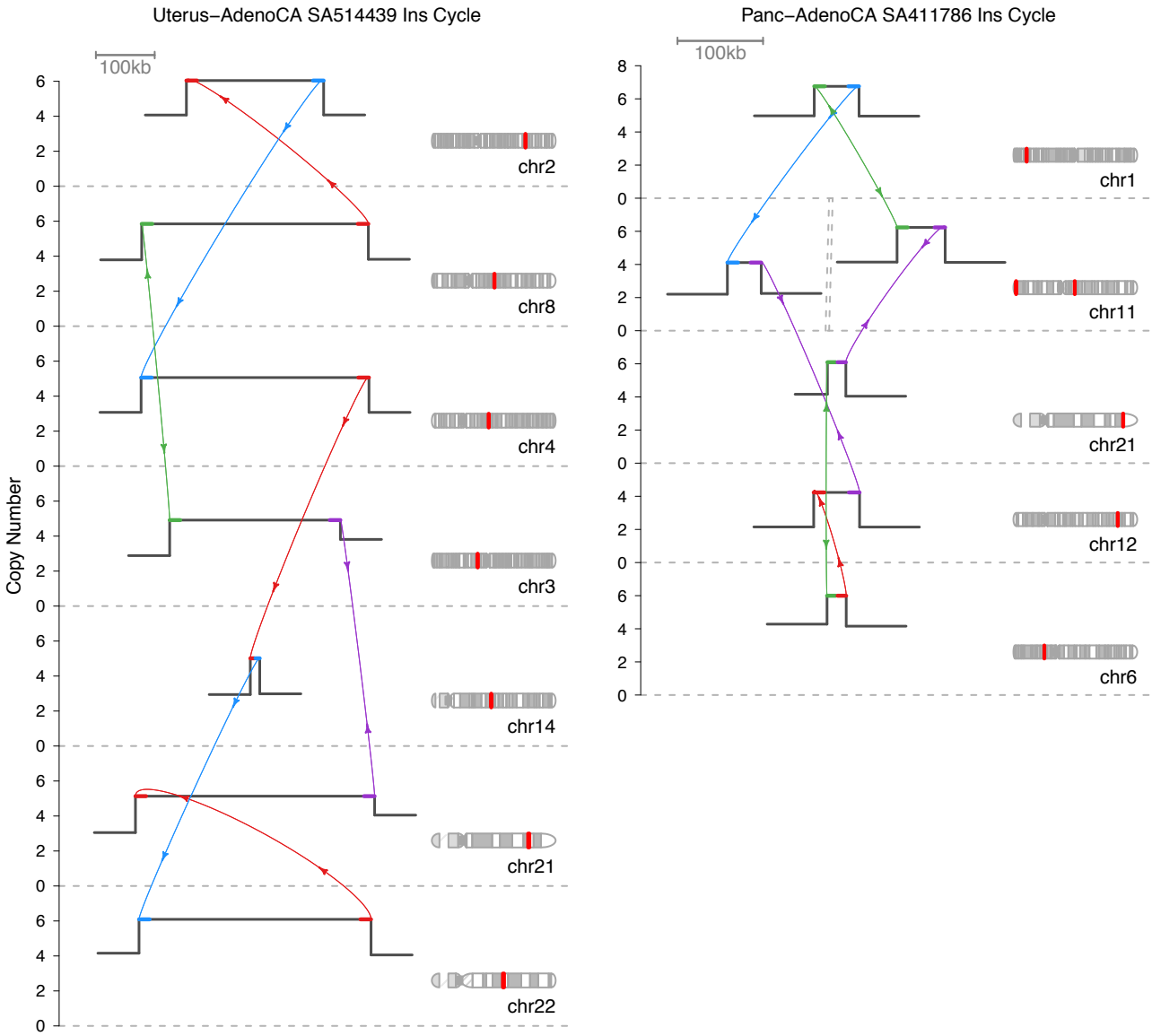


Figure 2.13: Longest templated insertion cycle events

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

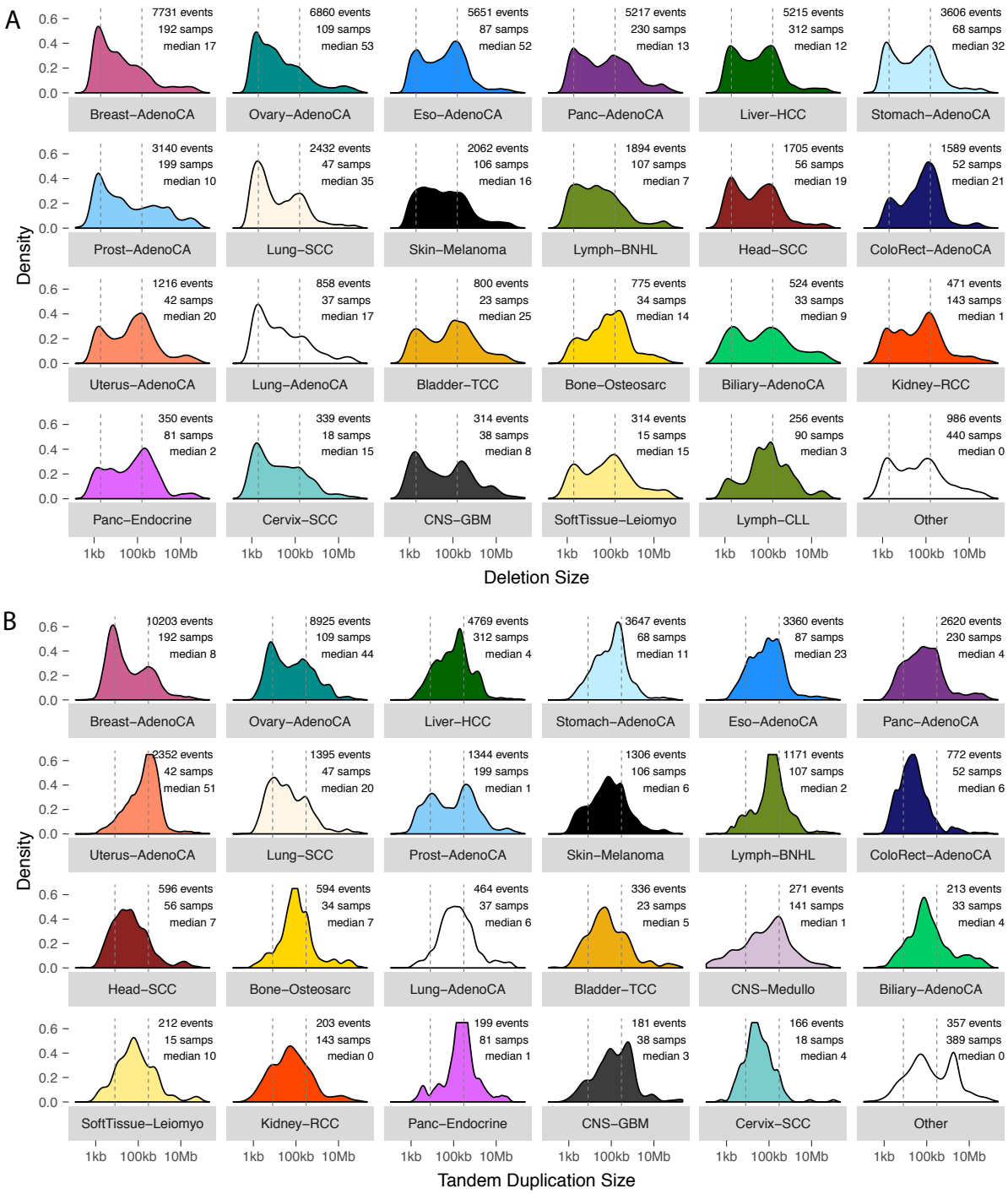


Figure 2.14: Deletion and tandem duplication size distributions over a $\log_{10}$ scale. Histology groups are sorted by total number of events in the cohort. The median number of events per sample is annotated in the top right for each group. Guide lines are marked at 2 kb and 160 kb for deletion (A), 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 unimodal 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 ($\sim 60\%$ 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.

[k] 538 samples included for deletion; 288 samples included for tandem duplication.

[l] The 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.

[m] In 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.

[n] For 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).

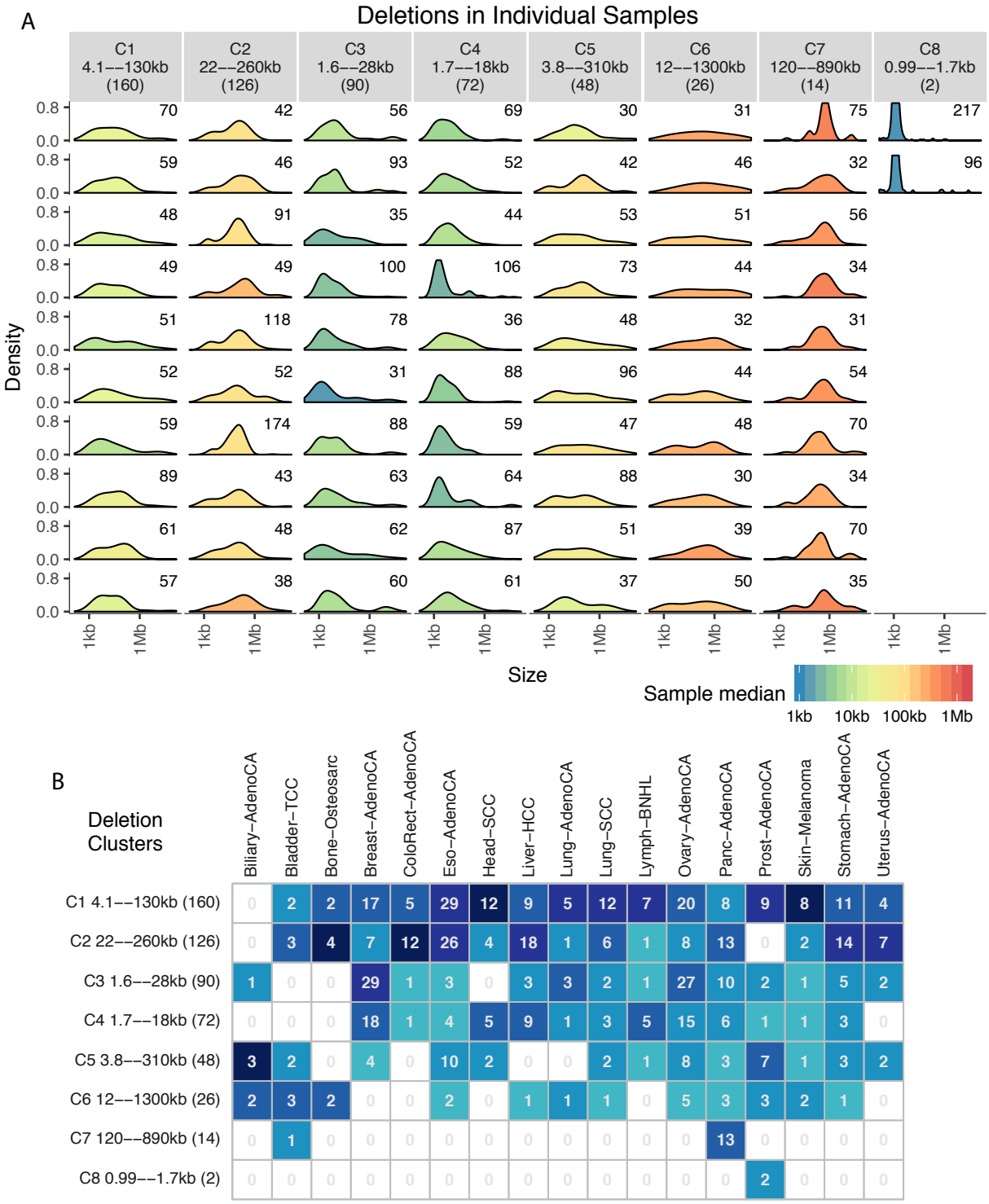
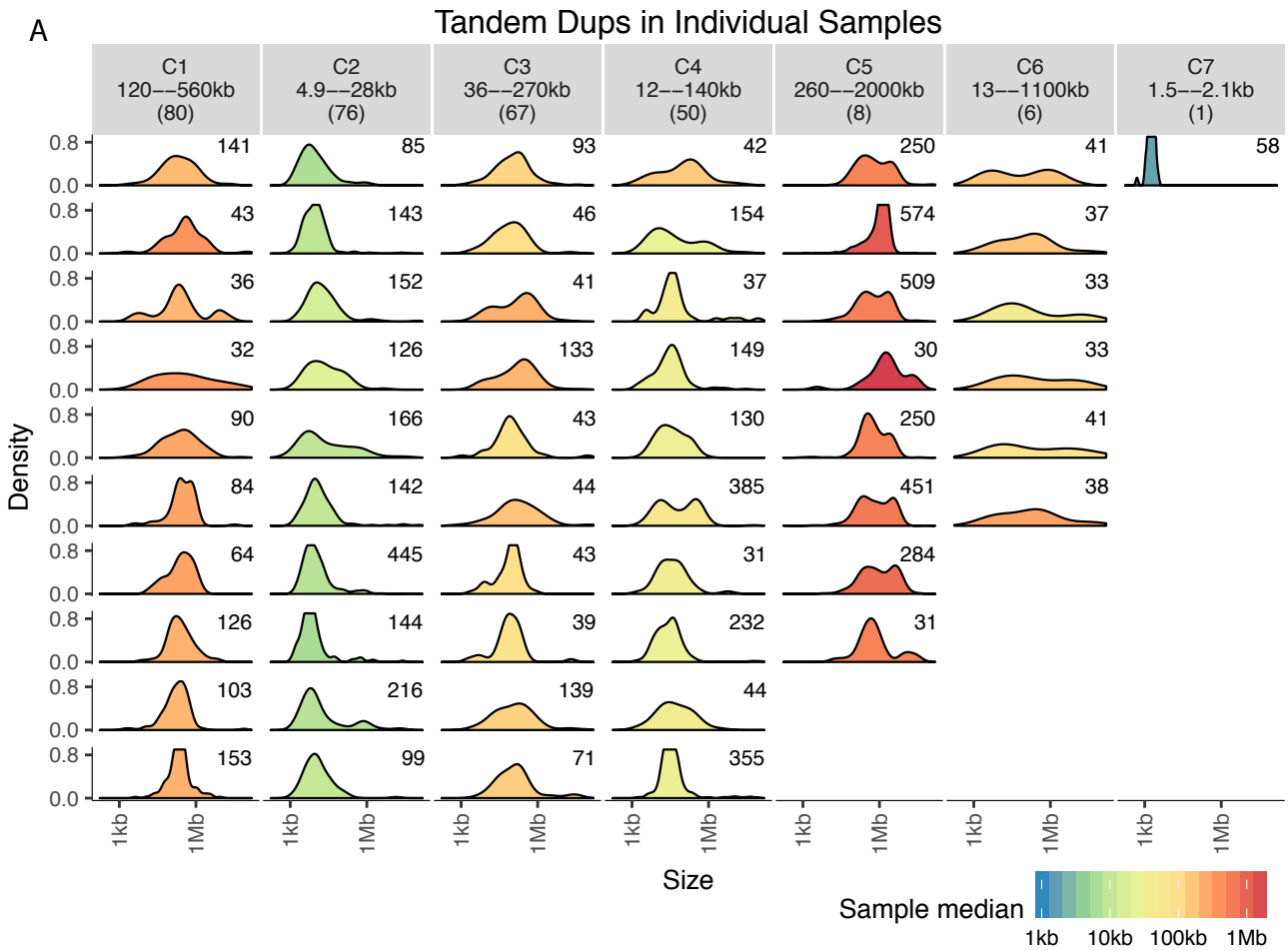


Figure 2.15: Samples with 30 or more deletions, clustered by their size distribution. Clusters are labelled with the inter-quartile deletion size range (pooling samples in the cluster), and the number of samples in parentheses. (a) Deletion size distributions of randomly chosen individual samples from 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.



B

Tandem Dup Clusters	Bladder–TCC	Breast–AdenoCA	Eso–AdenoCA	Head–SCC	Liver–HCC	Lung–SCC	Ovary–AdenoCA	Panc–AdenoCA	Prost–AdenoCA	Skin–Melanoma	Stomach–AdenoCA	Uterus–AdenoCA
C1 120––560kb (80)	1	16	3	0	7	2	18	4	1	3	5	20
C2 4.9––28kb (76)	2	27	3	1	3	3	29	5	1	0	1	1
C3 36––270kb (67)	1	2	17	1	5	3	10	7	0	2	11	8
C4 12––140kb (50)	1	7	9	3	7	3	5	3	0	4	8	0
C5 260––2000kb (8)	0	0	0	0	1	0	5	0	2	0	0	0
C6 13––1100kb (6)	0	2	1	0	0	1	1	1	0	0	0	0
C7 1.5––2.1kb (1)	0	0	0	0	0	0	0	0	1	0	0	0

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.

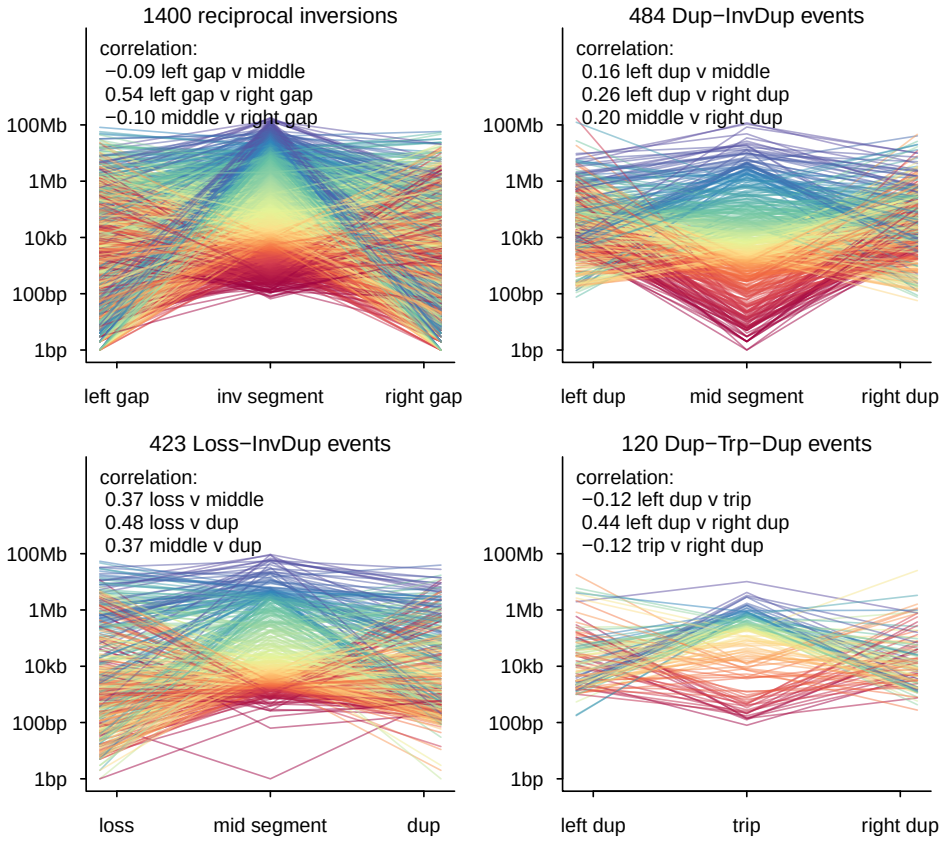


Figure 2.17: Segment size distribution for reciprocal inversion and local 2-jumps, shaded by the size of the middle segment. Pearson correlation coefficients between segment lengths on a $\log_{10}$ scale are annotated top left.

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.

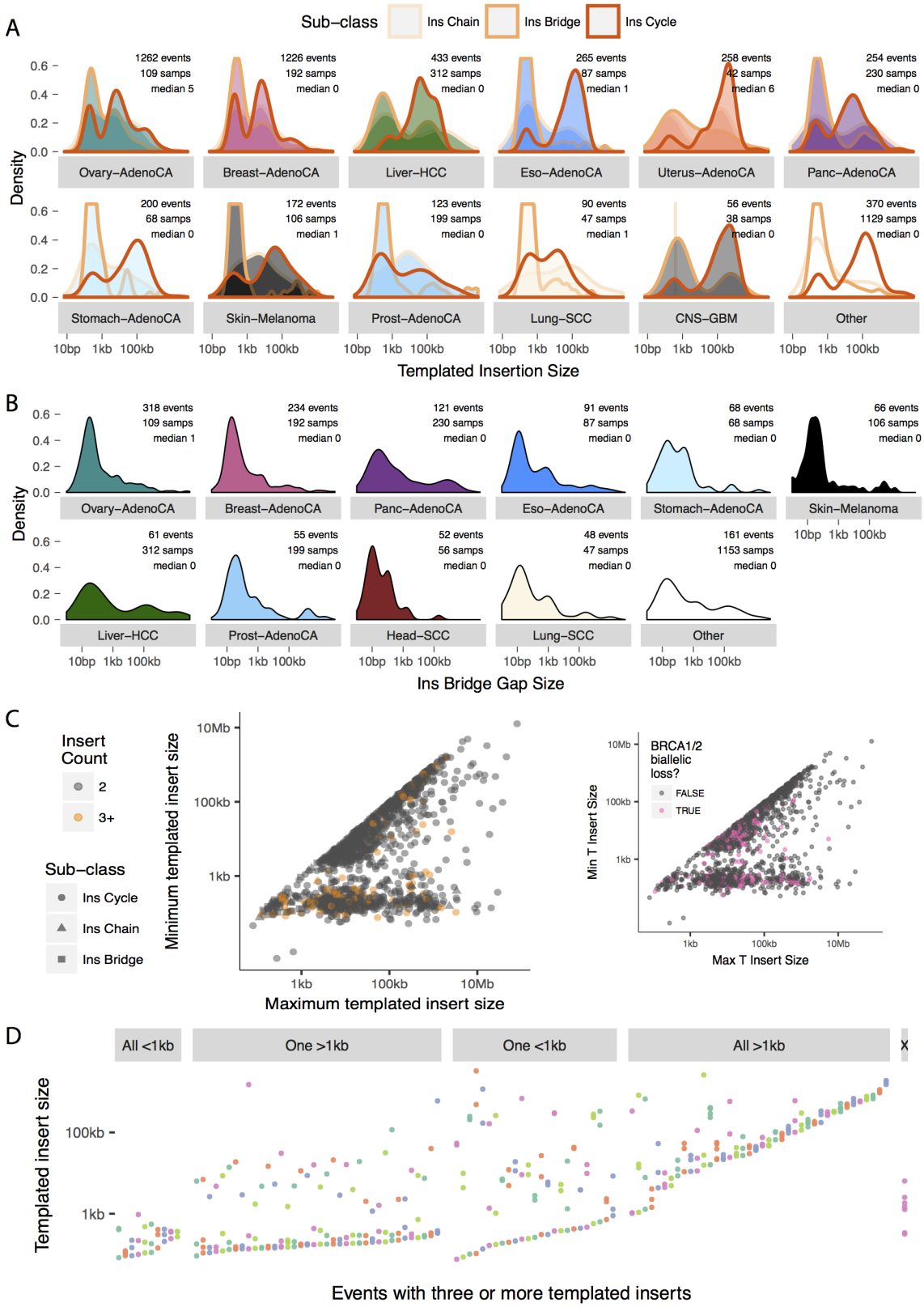


Figure 2.18: (a) Size distribution of templated inserts ([−+] motifs) by sub-class. (b) Size distribution of insertion bridge gaps ([+−] motifs). (c) Correlation between the smallest and largest insert in the same event (no chains/bridges of only two BPJ). (d) Events with three or more inserts, sorted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

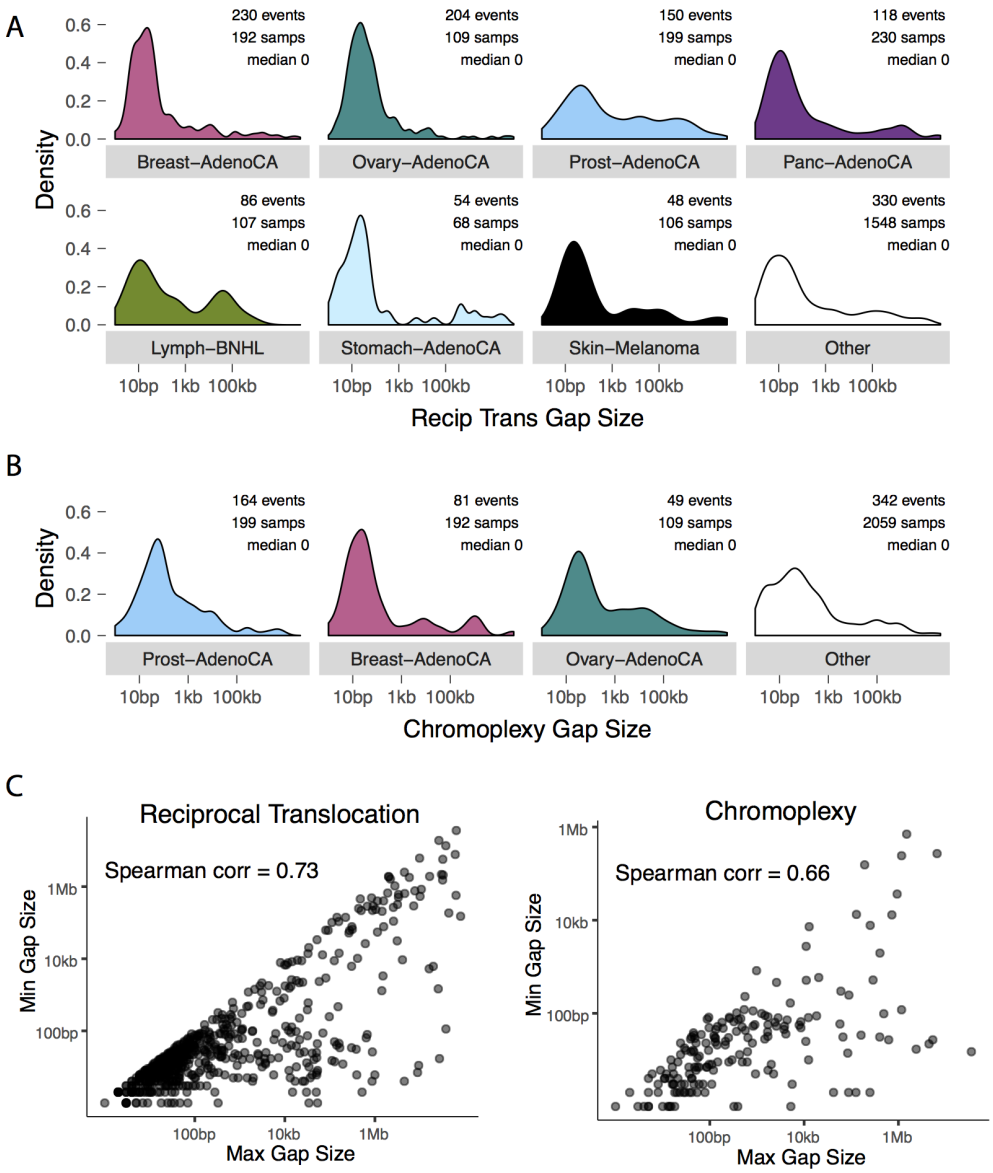


Figure 2.19: Gap sizes of [+−] motifs in (a) reciprocal translocation, (b) chromoplexy, and (c) correlation within the same event.

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 muddied 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

[o]In the callable genome space (see Section 3.1.1), the empirical MH distribution at random position pairs is $\Pr(0) = 0.743$, $\Pr(1) = 0.187$, $\Pr(2) = 0.050$, $\Pr(3) = 0.014$, $\Pr(\geq 4) = 0.006$. 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

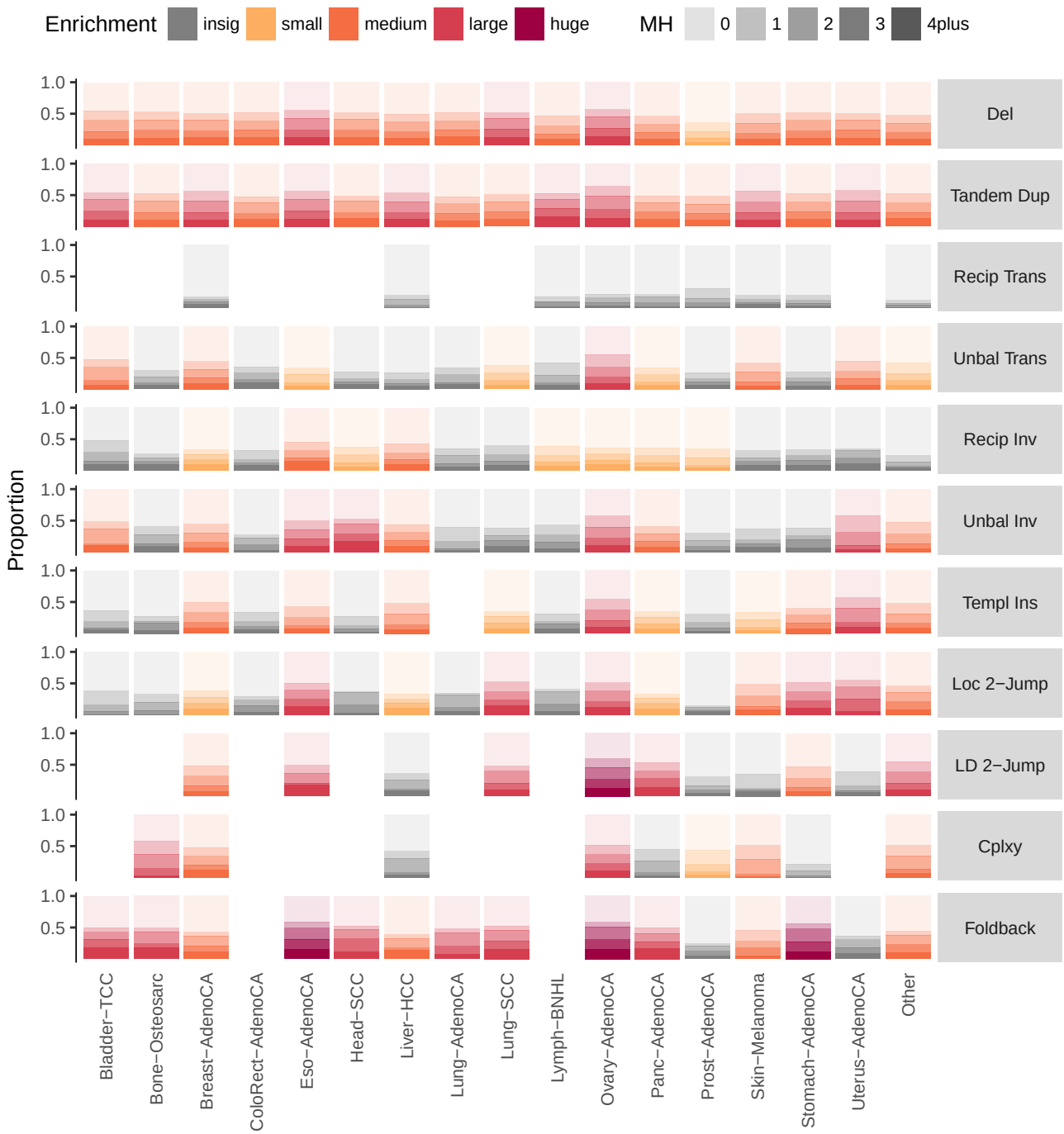


Figure 2.20: Distribution of microhomology at the breakpoint junction for different SV classes, separated by cancer histology. The magnitude of significant enrichment (compared to random background expectation) is coloured by the proportional odds regression coefficient, split into small (0.5–1.0], medium (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

[p]Piecewise 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.

[q]In 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.

[r]Excluded hypermutator samples included 25 melanoma, 8 colorectal, 2 lung, and 4 other cancers.

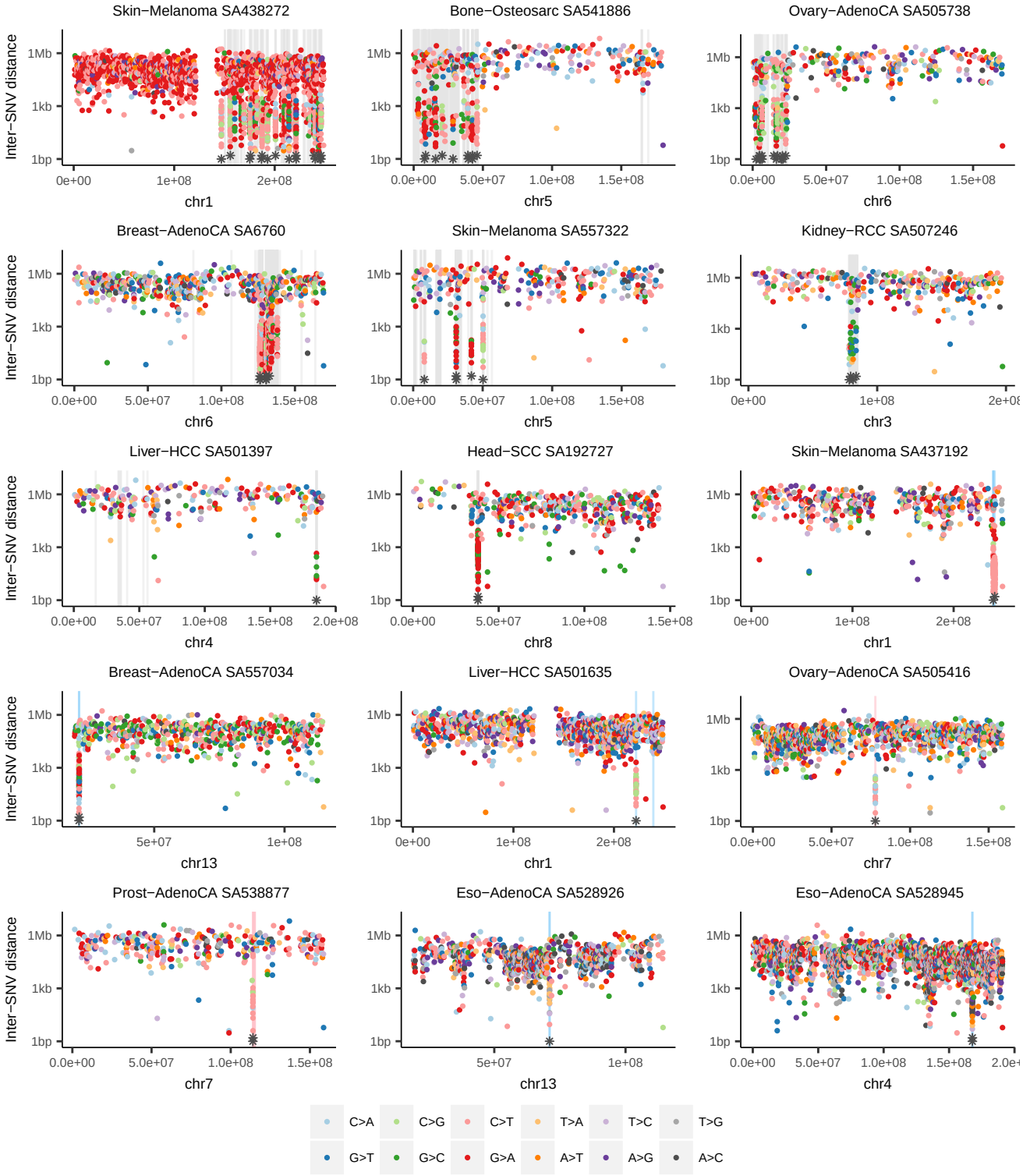


Figure 2.21: Fifteen chromosomes with identified kataegis regions, marked by dark gray stars along the lower edge. Rearrangement BPJ in associated events are marked by vertical lines in gray (complex SV), blue (deletion SV) or pink (other SV class; reciprocal inversion in the prostate example and unbalanced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

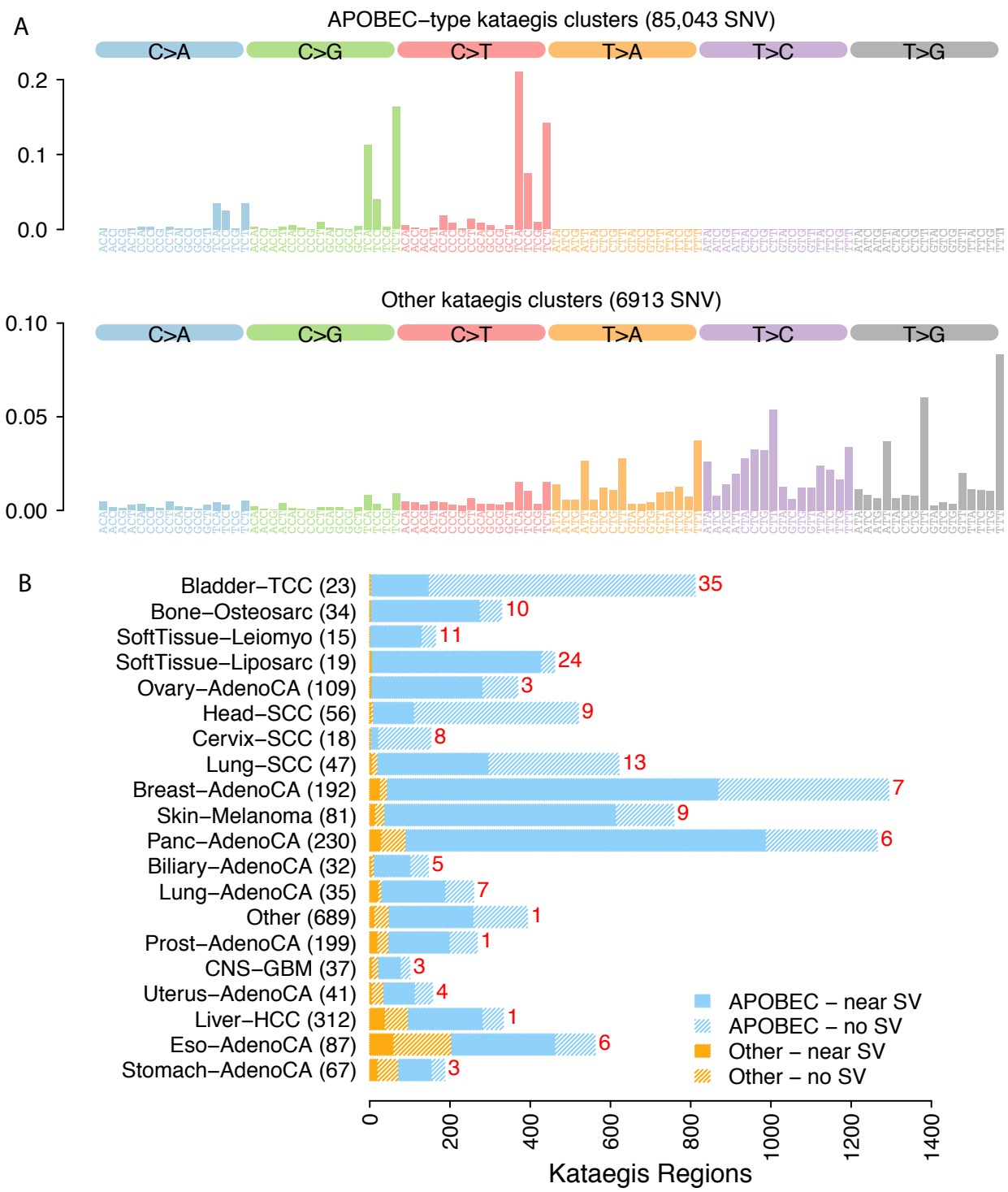


Figure 2.22: Kataegis distributions in the PCAWG cohort. (a) Somatic SNV distribution in a trinucleotide context around the pyrimidine reference base for two types of kataegis cluster: APOBEC type (mostly `C>N` in `TCN`), and other. (b) Number of kataegis regions in each histology group, shaded by SNV 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.

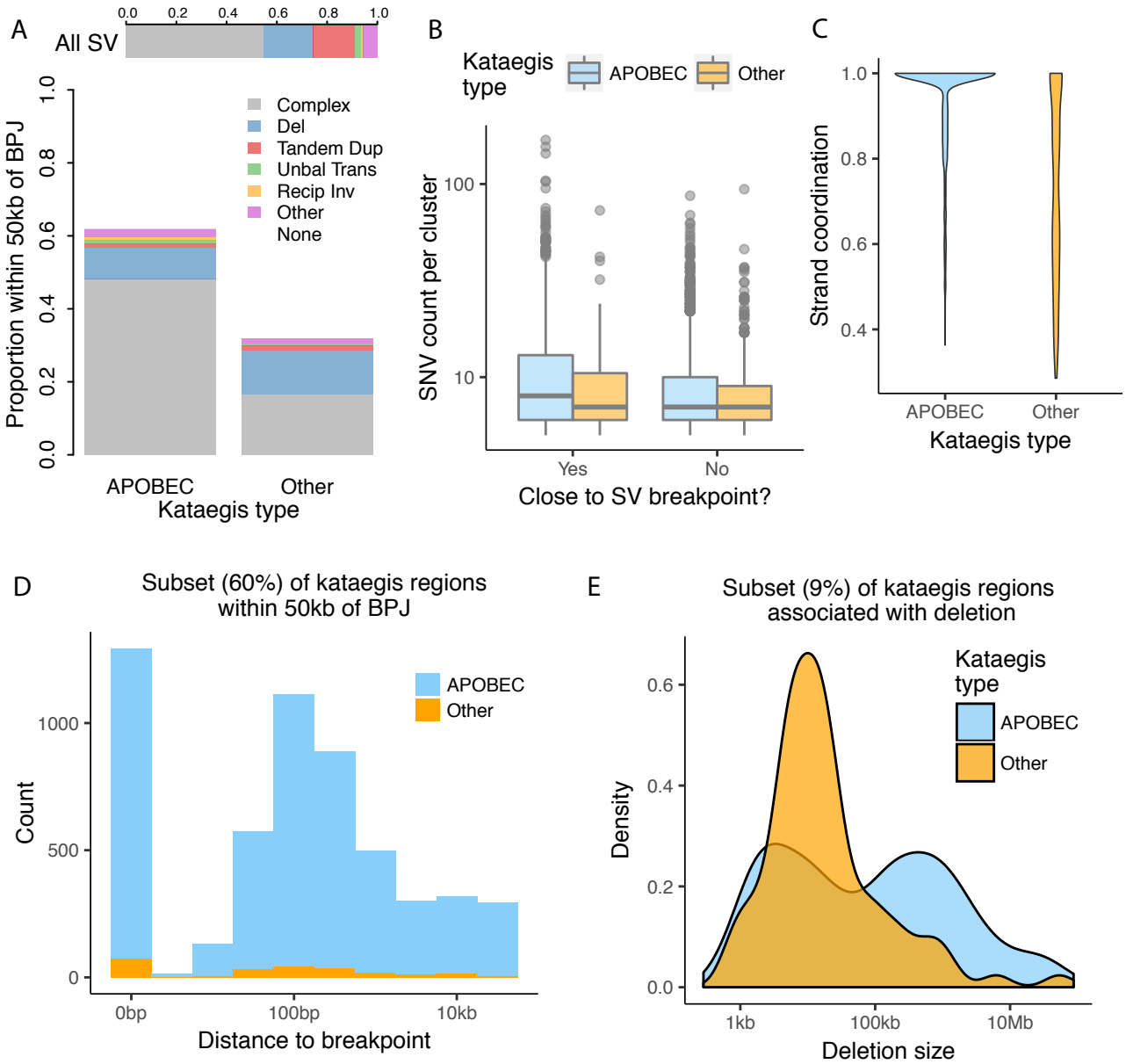


Figure 2.23: Kataegis properties in the PCAWG cohort. (a) Proportion of kataegis regions within 50 kb of a SV breakpoint, shaded by SV class. The background distribution of all SV classes is indicated above. (b) SNV counts per kataegis cluster ($\log_{10}$ scale). (c) Extent of strand coordination within 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.

[s] Although 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).

Chapter 3

Genome properties and the rate of rearrangement

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

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

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

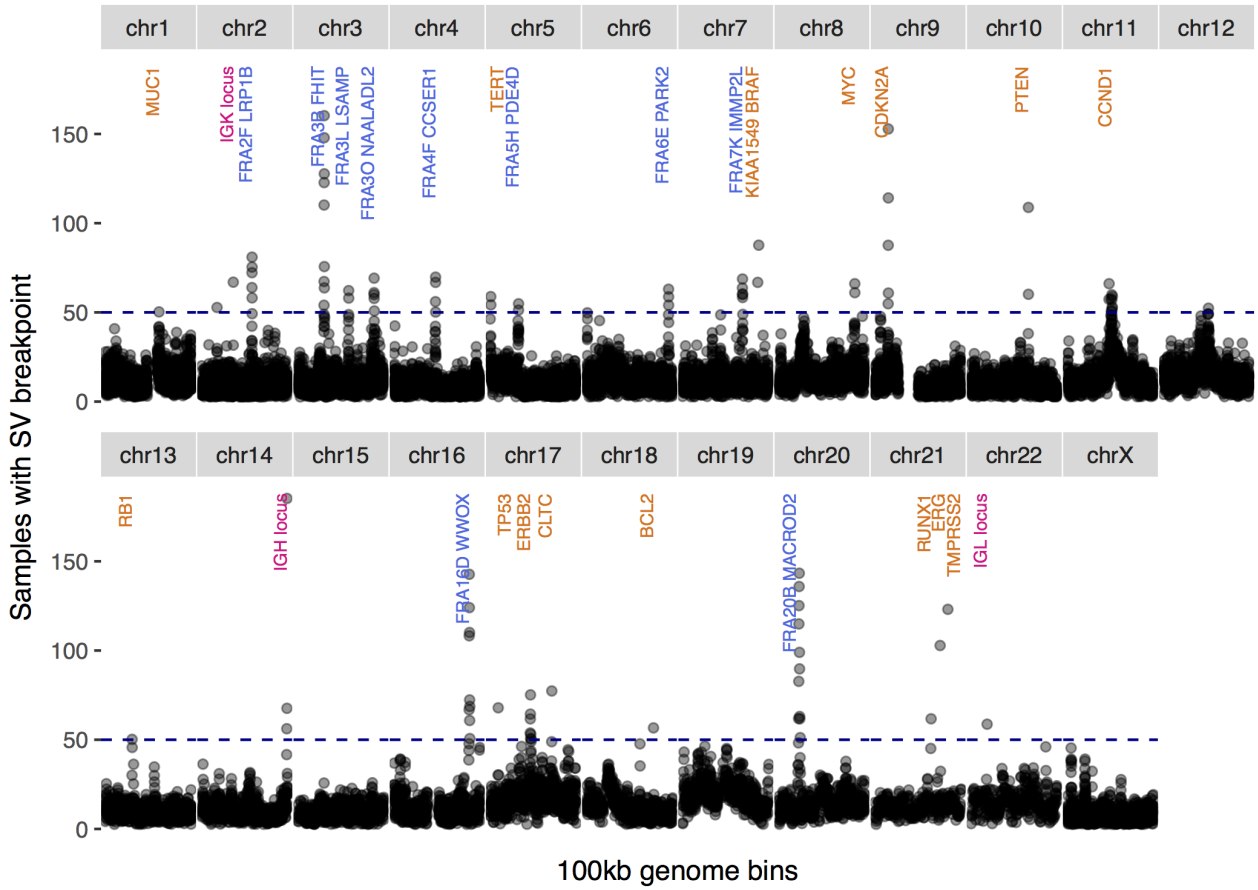


Figure 3.1: The genome-wide distribution of somatic rearrangements across 2559 PCAWG samples. Each dot records the number of samples containing a somatic SV breakpoint in a 100 kb bin. Bins with breakpoints in fewer than three samples are excluded. A selection of peak regions with more than 50 rearranged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

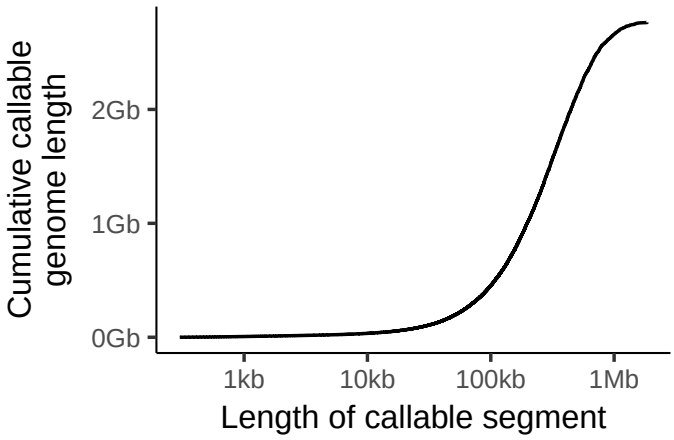


Figure 3.2: Cumulative length of callable genome regions, sorted by size.

3.1 A library of genome properties

3.1.1 Defining the callable genome

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

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

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

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

[b]GATK CallableLoci v3.3-0 run with options `maxFractionOfReadsWithLowMAPQ=0.25`, `maxDepth=1000`, and otherwise default settings.

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

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

3.1.2 Defining pixel metrics

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

Basic sequence features

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

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

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

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

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

Repeat sequences

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

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

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

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

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

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

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

Non-B DNA forming motifs

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

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

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

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

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

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

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

ROADMAP Epigenomics

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

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

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

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

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

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

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

Genome organisation

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

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

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

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

Other properties

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

Germline recombination rate The calculated metric is the germline recombination rate of the nearest SNP, using data from the HapMap consortium (Frazer et al., 2007)[d].

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

3.1.3 Correlation between genome properties

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

[d] `ftp://ftp.ncbi.nlm.nih.gov/hapmap/recombination/2011-01_phaseII_B37/`

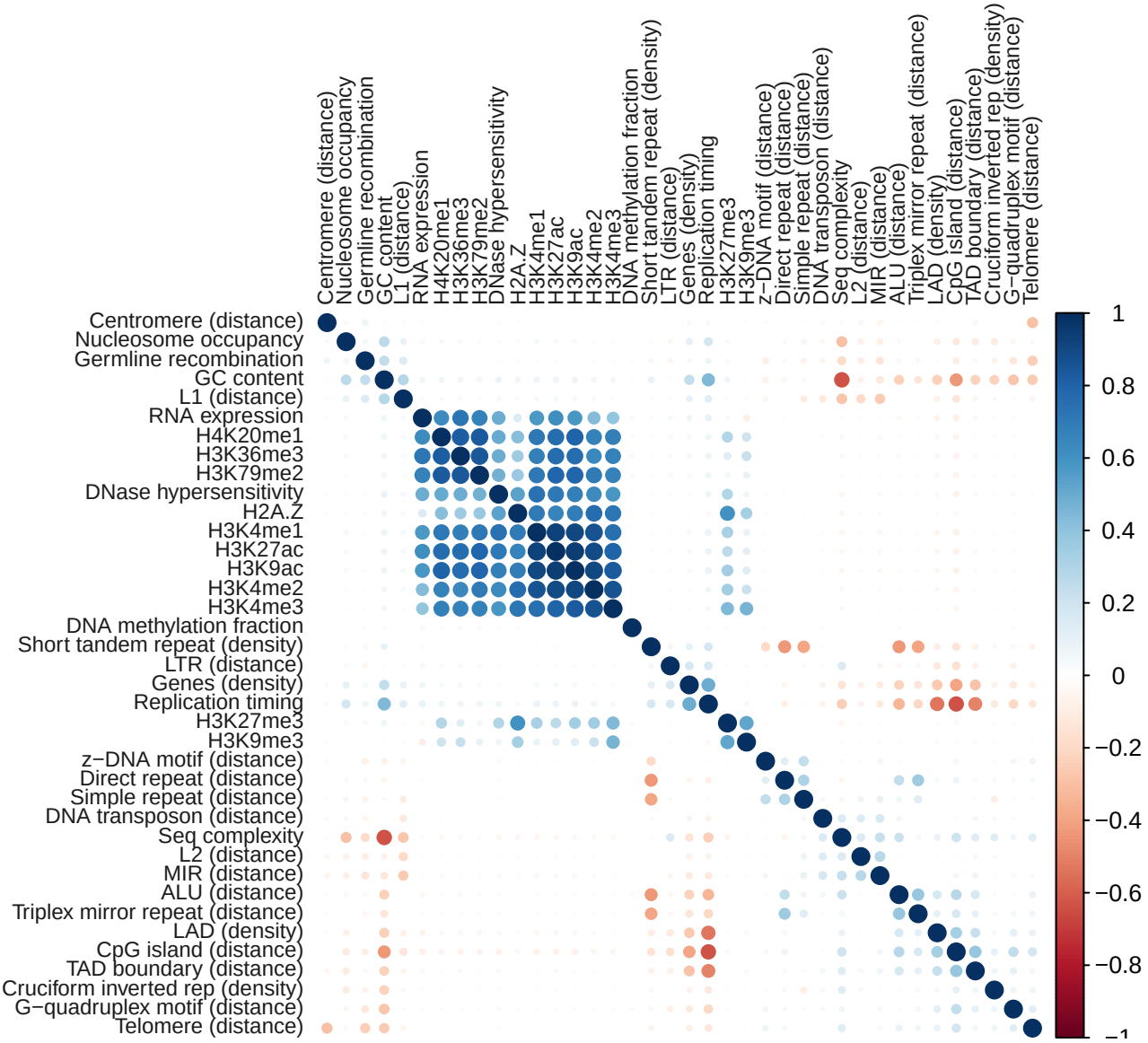


Figure 3.3: Spearman correlation between 38 genome properties at 100,000 random uniform positions in the callable genome space. Circle size is proportional to the magnitude of correlation.

3.2 SV classes associate with genome properties

3.2.1 Property quantile skew at SV breakpoints

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

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

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

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

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

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

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

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

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

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

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

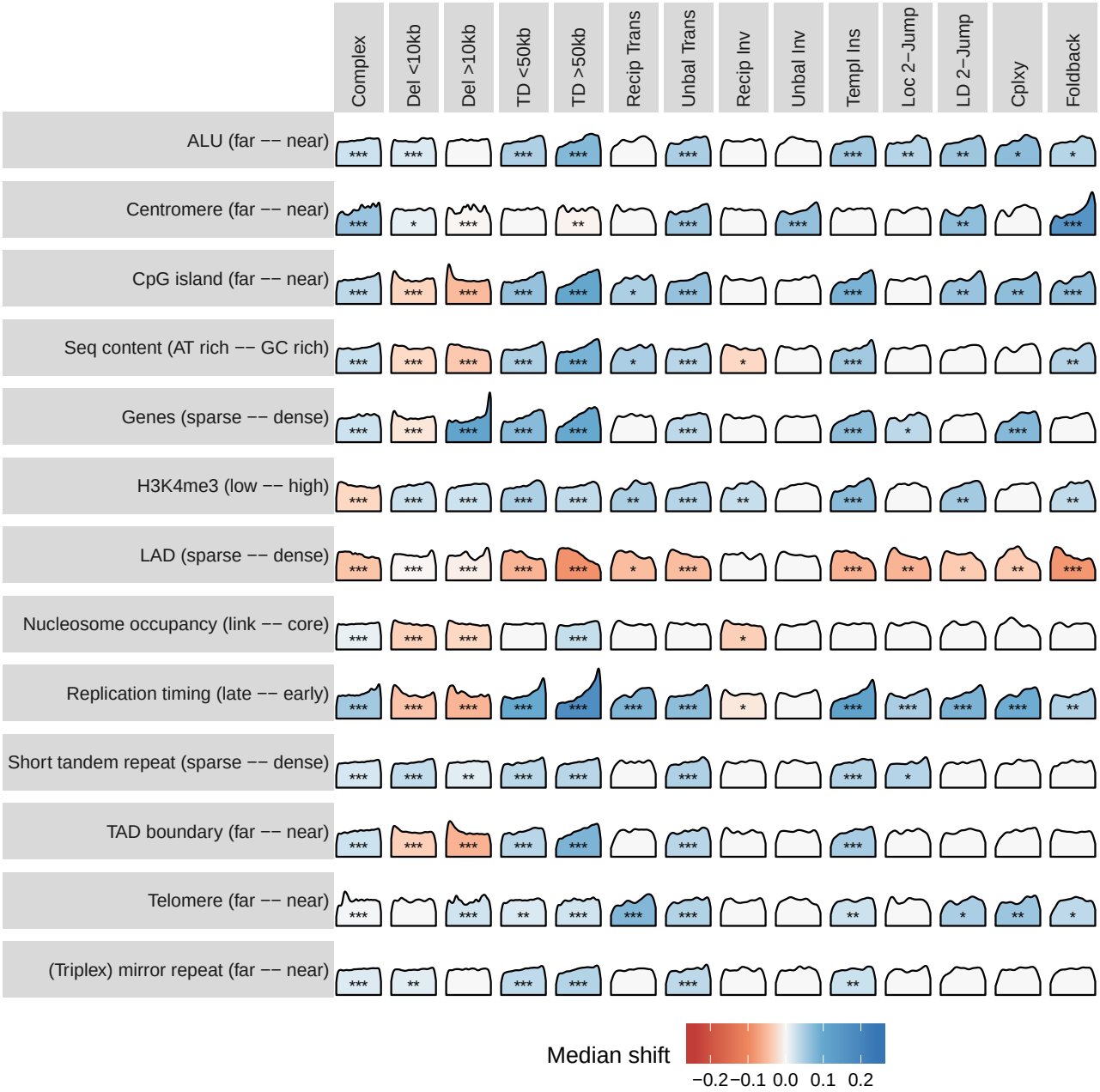


Figure 3.4: Associations between genome properties (rows) and SV classes (columns). Each density curve represents the quantile distribution of the genome property metrics at observed breakpoints compared to random genome positions, with stars indicating significant departure from uniform quantiles: FDR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

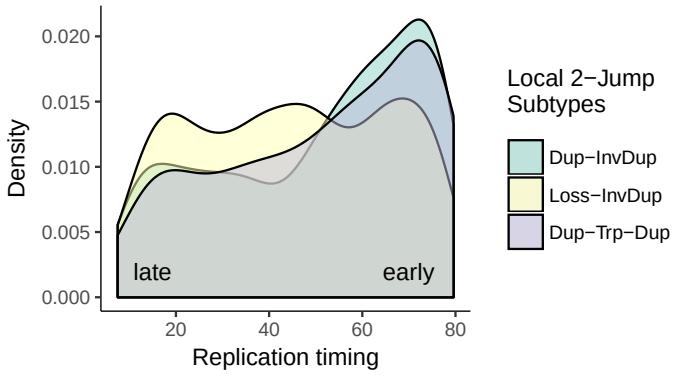


Figure 3.5: Replication timing distribution for the three sub-groups of local 2-jump SV classes. ANOVA p value for a difference between groups is 7×10^{-9}.

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

3.2.2 Breakpoints in close proximity with short repeats

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

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

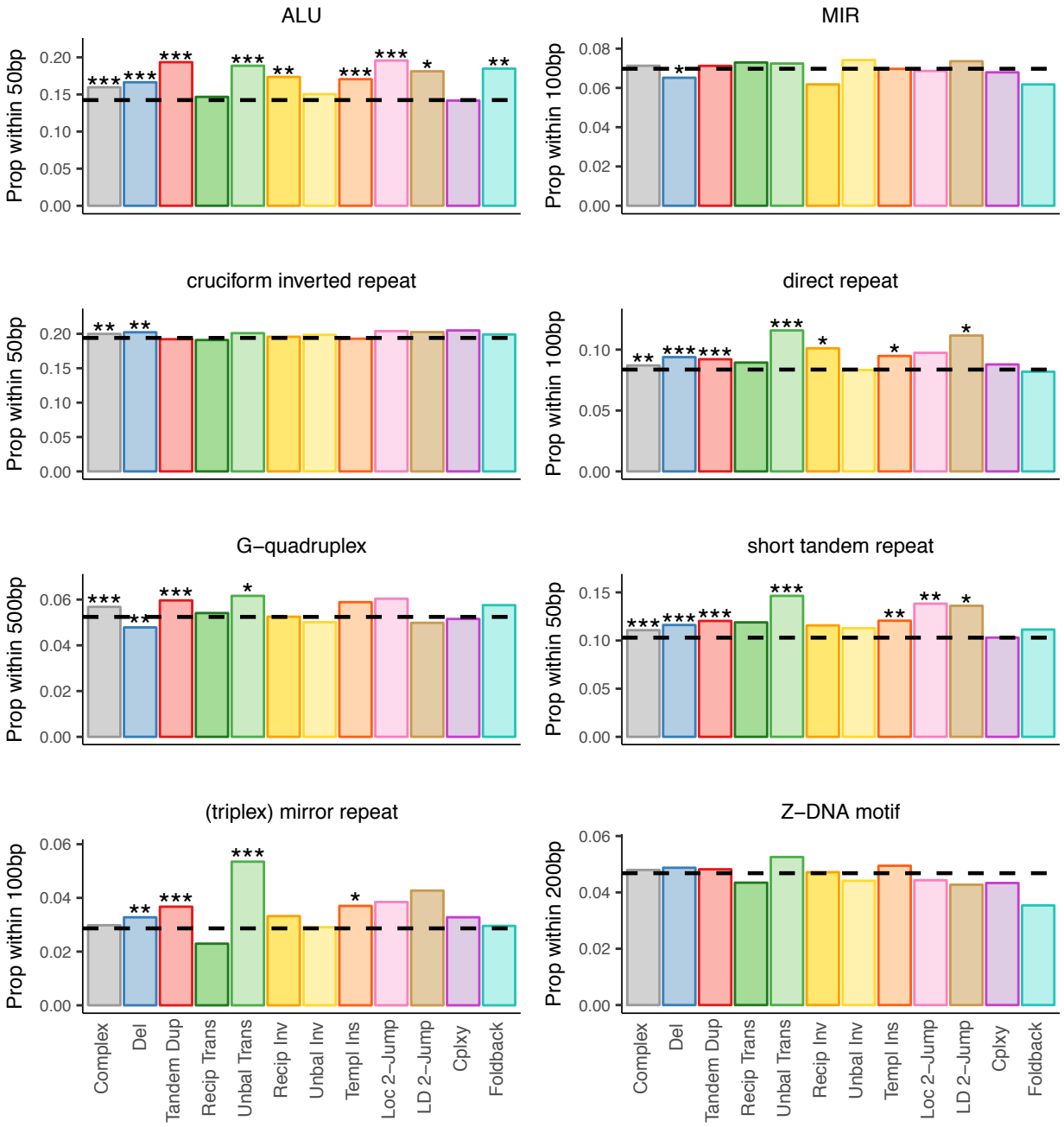


Figure 3.6: The proportion of SV breakpoints (one side per BPJ) within a short radius of each SINE and non-B DNA motif class. The proportion expected under a uniform null over the callable genome is indicated with a black dashed line. Significant departure from the uniform expectation is assessed with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

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

3.2.3 Replication timing at hypermutator breakpoints

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

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

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

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

[f]Breast, esophagus, liver, pancreatic (adenocarcinoma), prostate, and skin (melanoma).

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

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

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

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

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

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

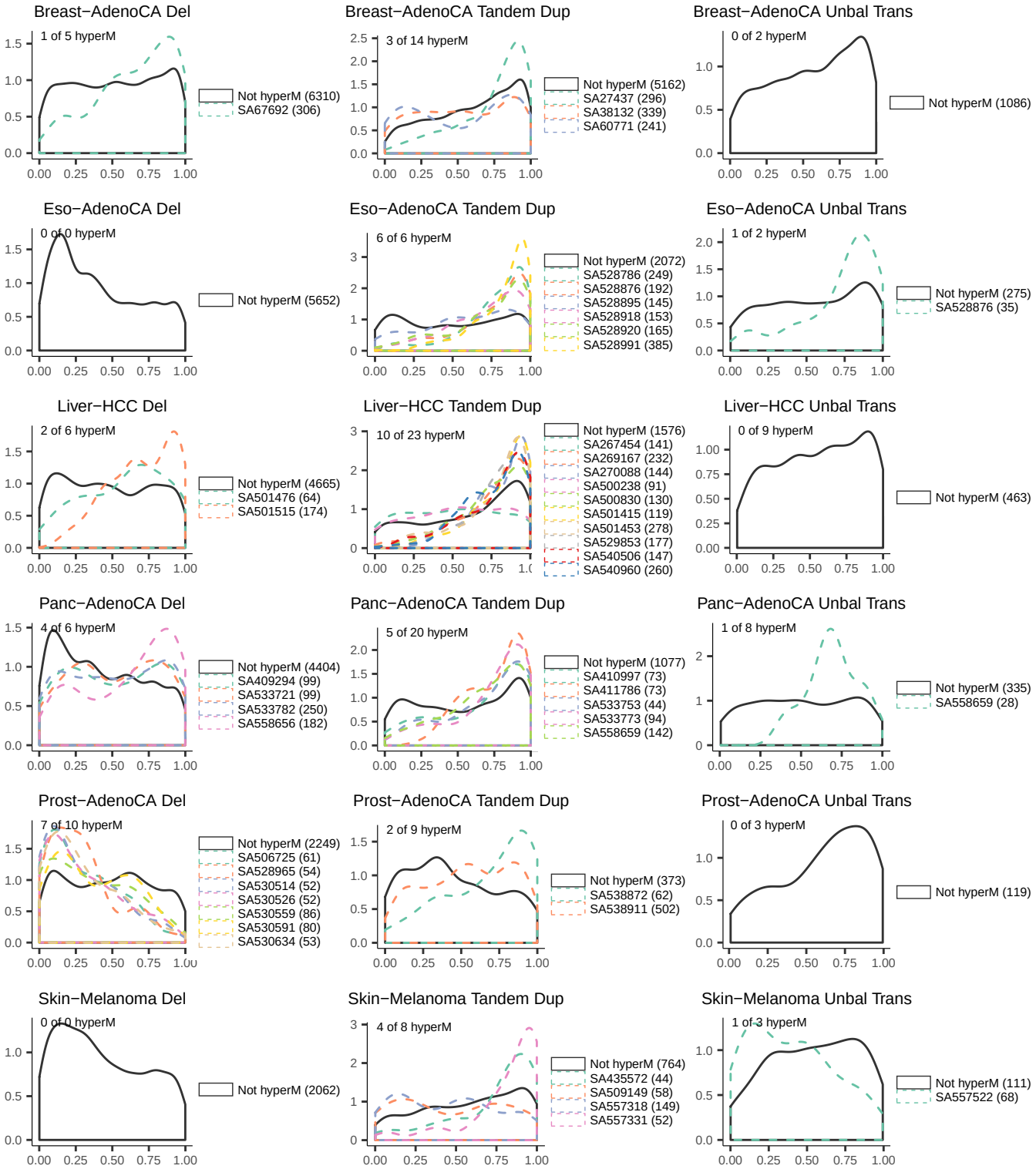


Figure 3.7: Density skew of replication timing quantiles for hypermutators compared to the pool of non-hypermutators for deletions, tandem duplications, and unbalanced translocations in six cancer types. The number of events in the sample (or pool of non-hypermutators) is indicated in the legend. Only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

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

3.2.4 Property correlation at the junction

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

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

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

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

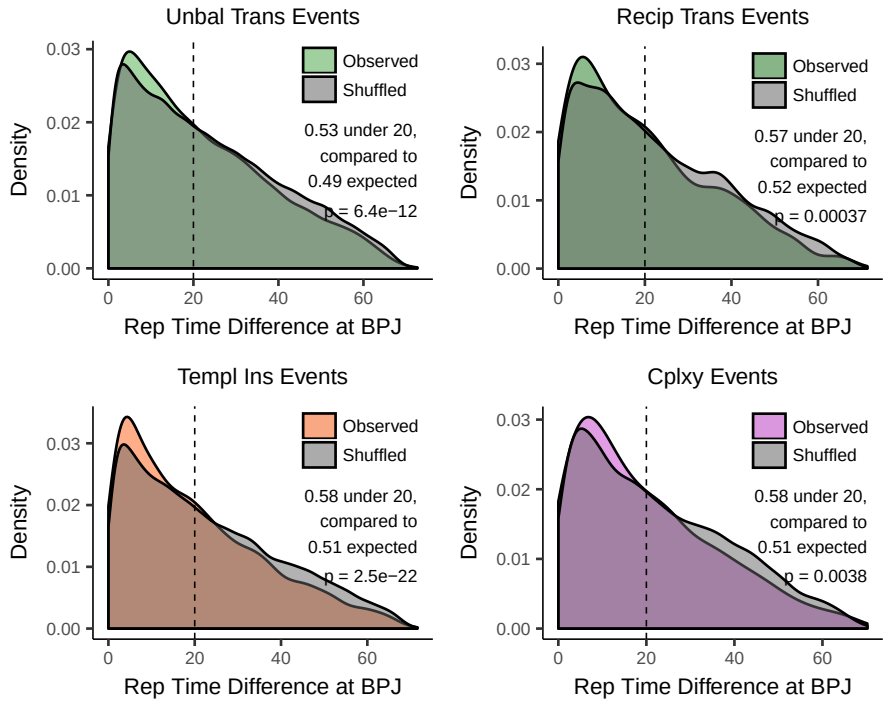


Figure 3.8: Difference in replication timing estimates across interchromosomal BPJ, compared to the expected distribution at randomly shuffled junctions. The proportion of junctions with a replication timing difference less than 20 is compared with a binomial proportion test, annotated middle right.

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

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

3.3 Modelling the rate of rearrangement

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

3.3.1 Methods

Outcome variable

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

Predictor variables

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

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

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

GLM models with lasso regularisation

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

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

GAM models with lasso-type regularisation

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

3.3.2 Results

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

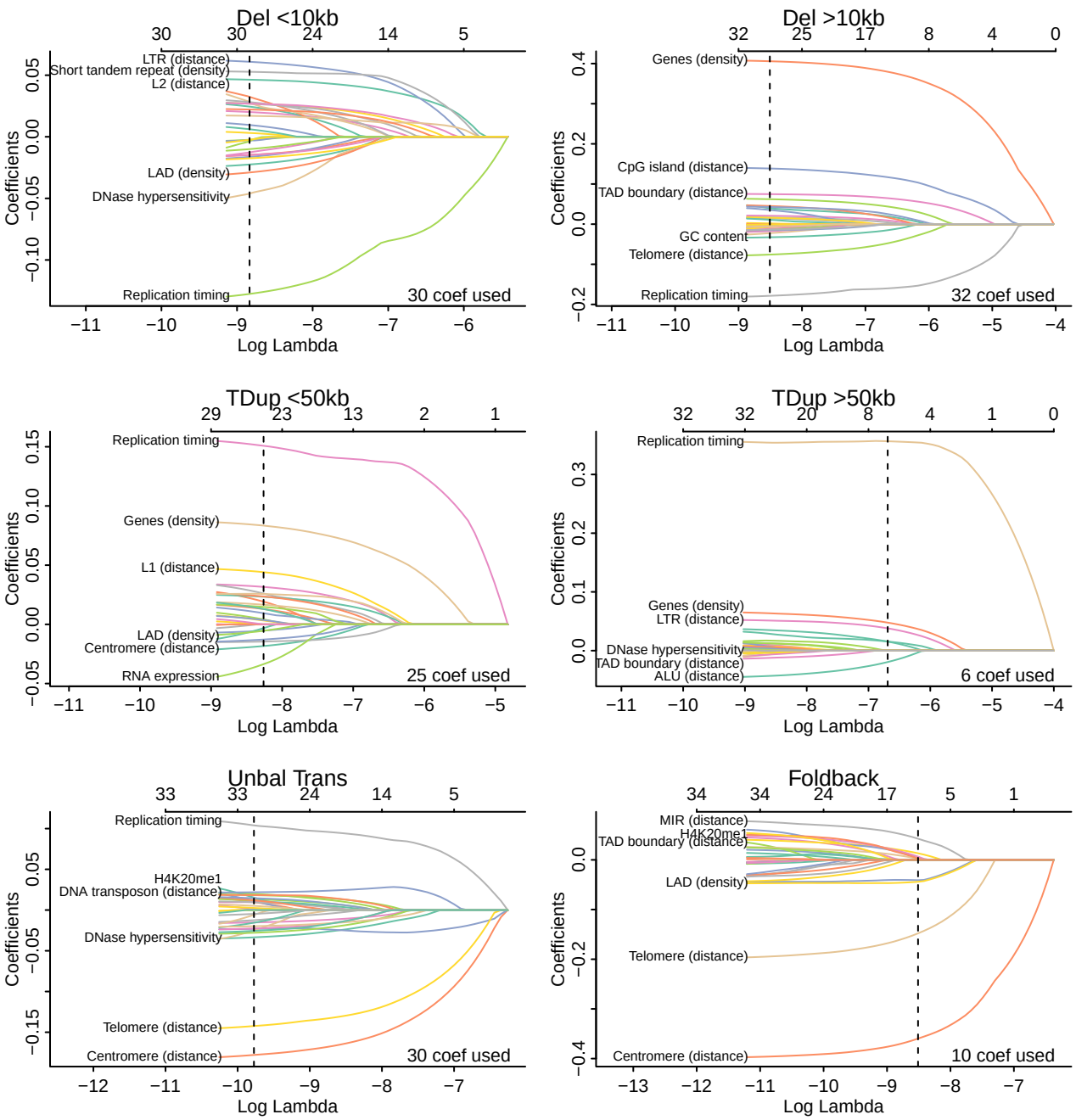


Figure 3.9: The lasso tuning parameter controls coefficient paths and number of selected predictors (annotated top) for logistic GLMs classifying real and random breakpoint positions for six SV classes. The optimal tuning parameter (best cross-validation performance) is marked with a vertical dashed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

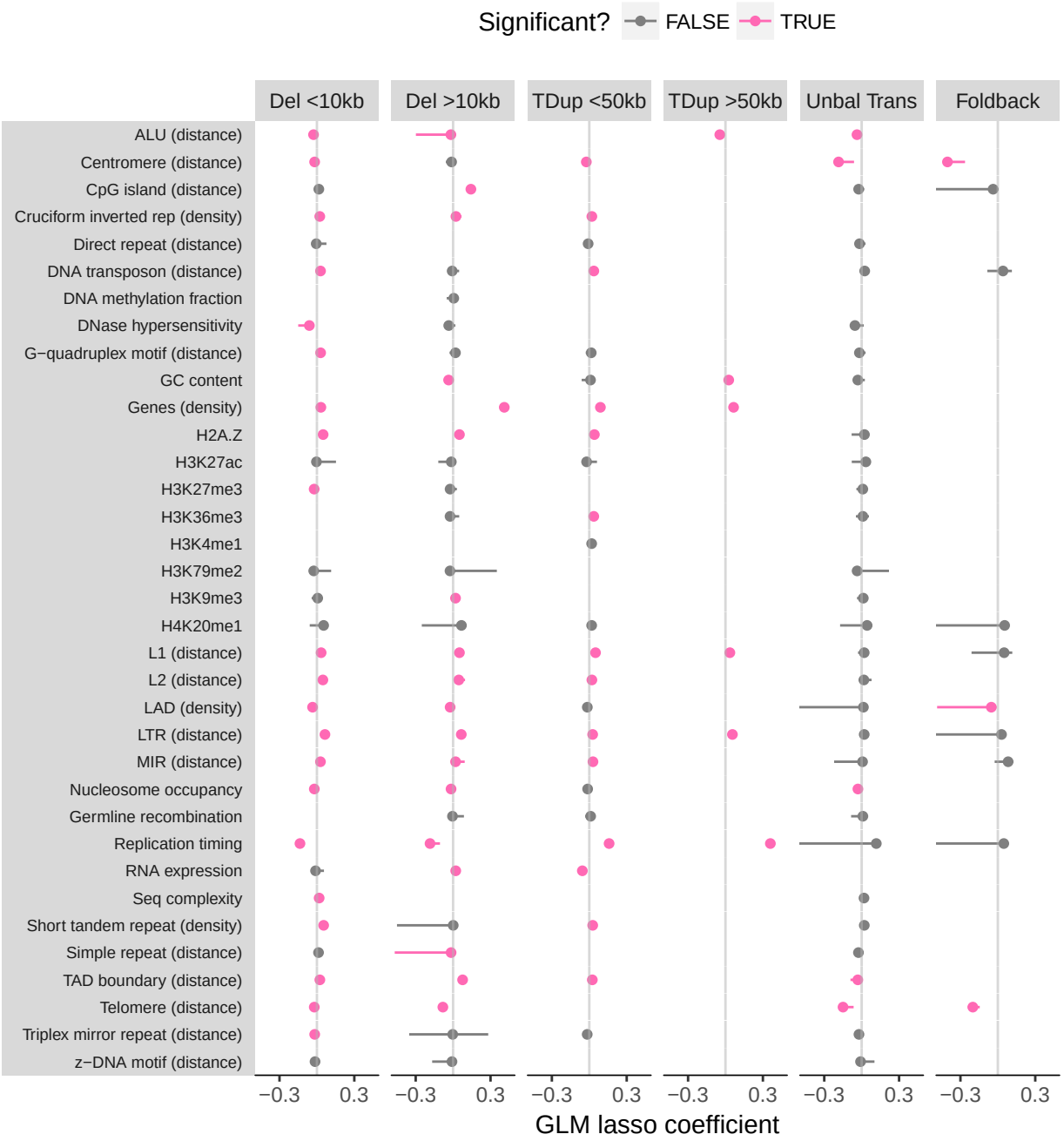


Figure 3.10: Fitted coefficient values (dots) and their confidence intervals (horizontal lines) for predictors in the optimal lasso GLM for each SV class, coloured by lasso-adjusted significance below a 0.05 type 1 threshold. Vertical guide lines mark zero.

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

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

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

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

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

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

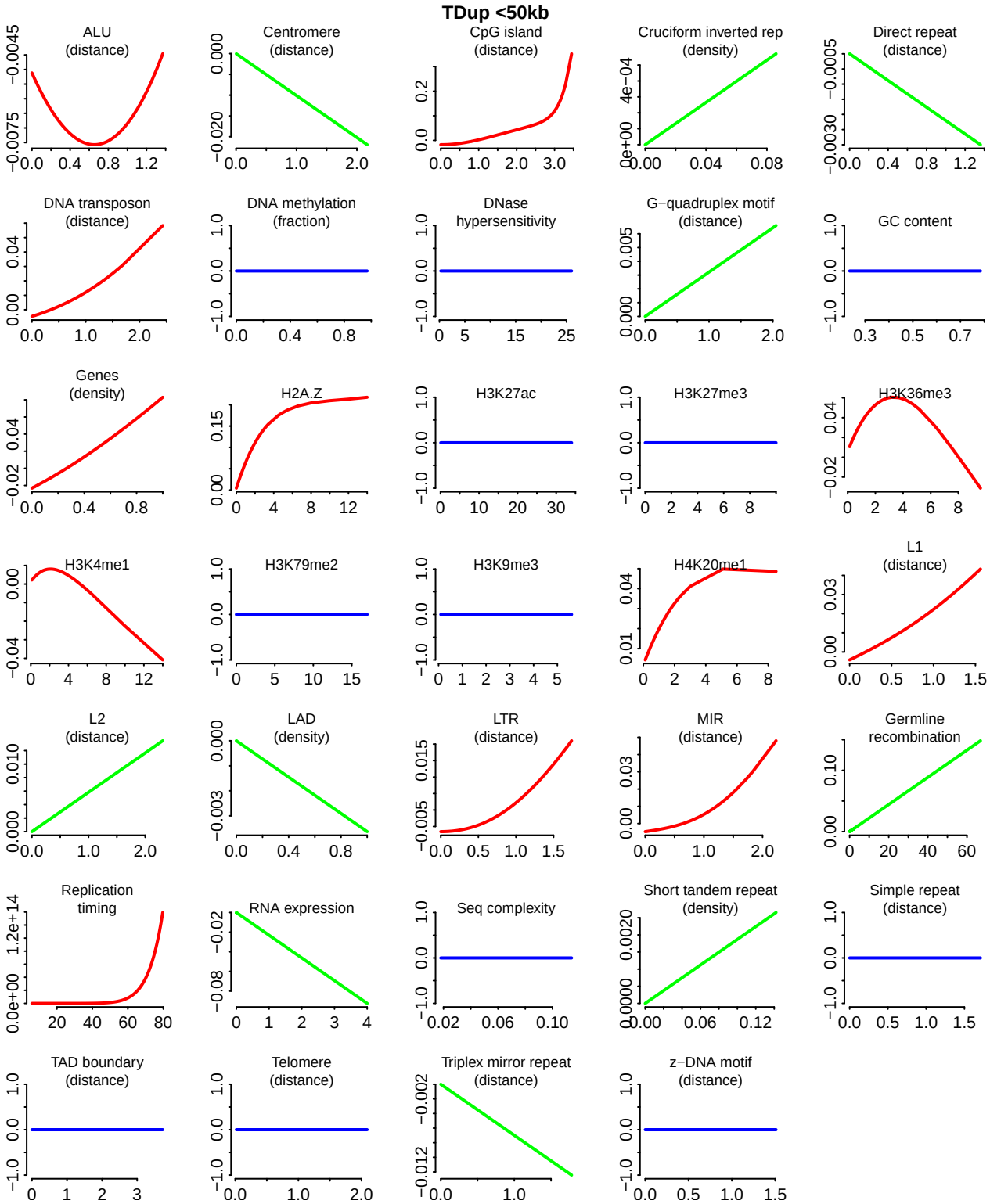


Figure 3.11: The optimal (best cross-validation performance) logistic GAM with lasso-type regularisation for small tandem duplications. The effect on the log odds of a position being a genuine breakpoint is shown as a function over each predictor's domain (back-transformed from scaled model predictors),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

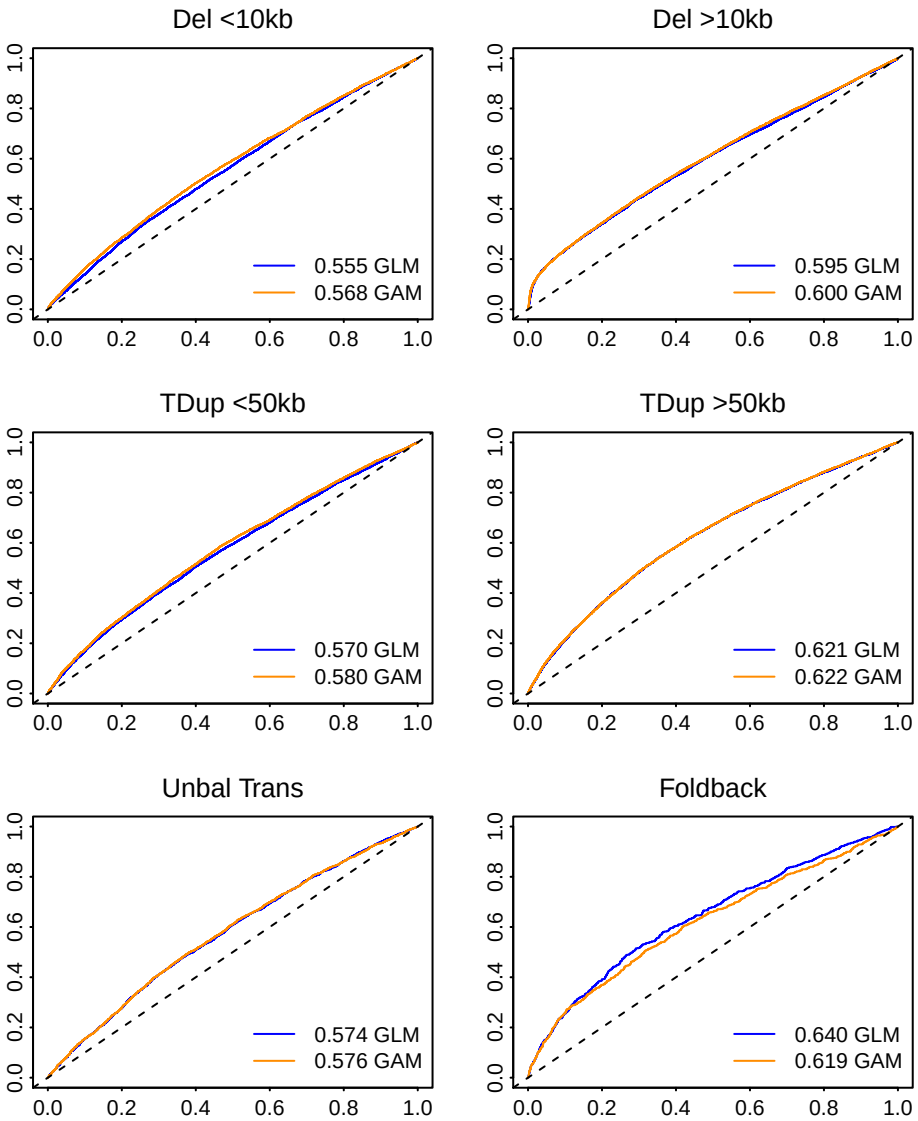


Figure 3.12: The ROC curves for true positive rate (vertical axis) and true negative rate (horizontal axis) in the testing subset for GLM and GAM models classifying real and random breakpoint positions for six SV classes. The total area under the curve is annotated bottom right, quantifying the degree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

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

3.3.3 Discussion

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

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

[i]The q-arm side of the chromosome 16 centromere is missing because that region is not included in the callable genome definition from Section 3.1.1.

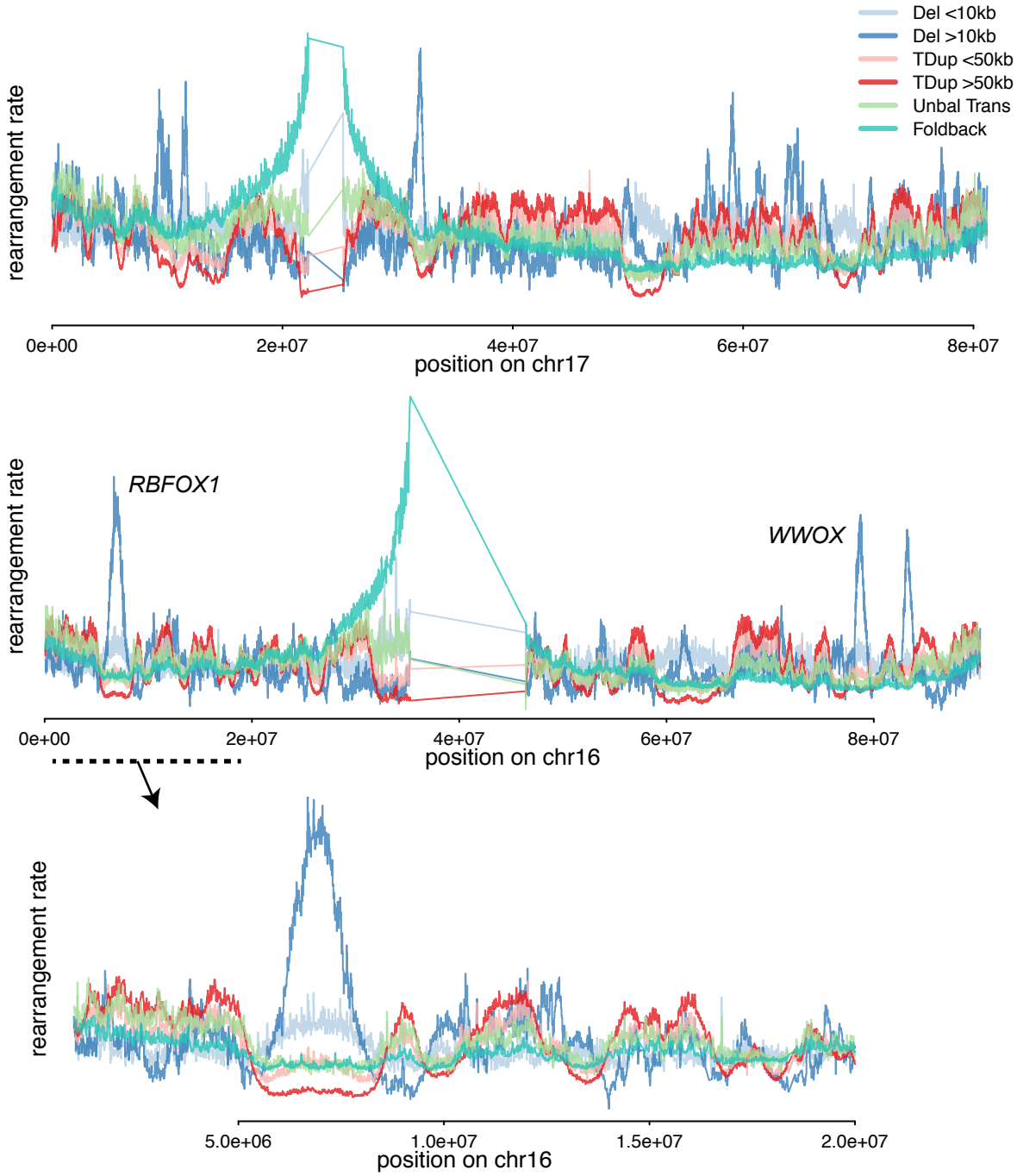


Figure 3.13: Predicted GLM rearrangement rate for six SV classes in 10 kb bins along chromosomes 16 and 17, using the breast tissue-specific ROADMAP predictors. Two peaks in the predicted large deletion rate around fragile site genes *RBFOX1* and *WWOX* are annotated on chr16.

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

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

3.4 Fragile sites and other anomalous genome regions

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

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

[j] 'Common' FS because they are common to all individuals, as opposed to rare FS which express fragility only in certain polymorphic forms.

[k] In vitro replication stress typically induced by the DNA polymerase inhibitor aphidicolon.

3.4.1 Defining fragile sites in the PCAWG cohort

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

3.4.2 Fragile site activity

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

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

[l] per total cohort, not per sample!

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

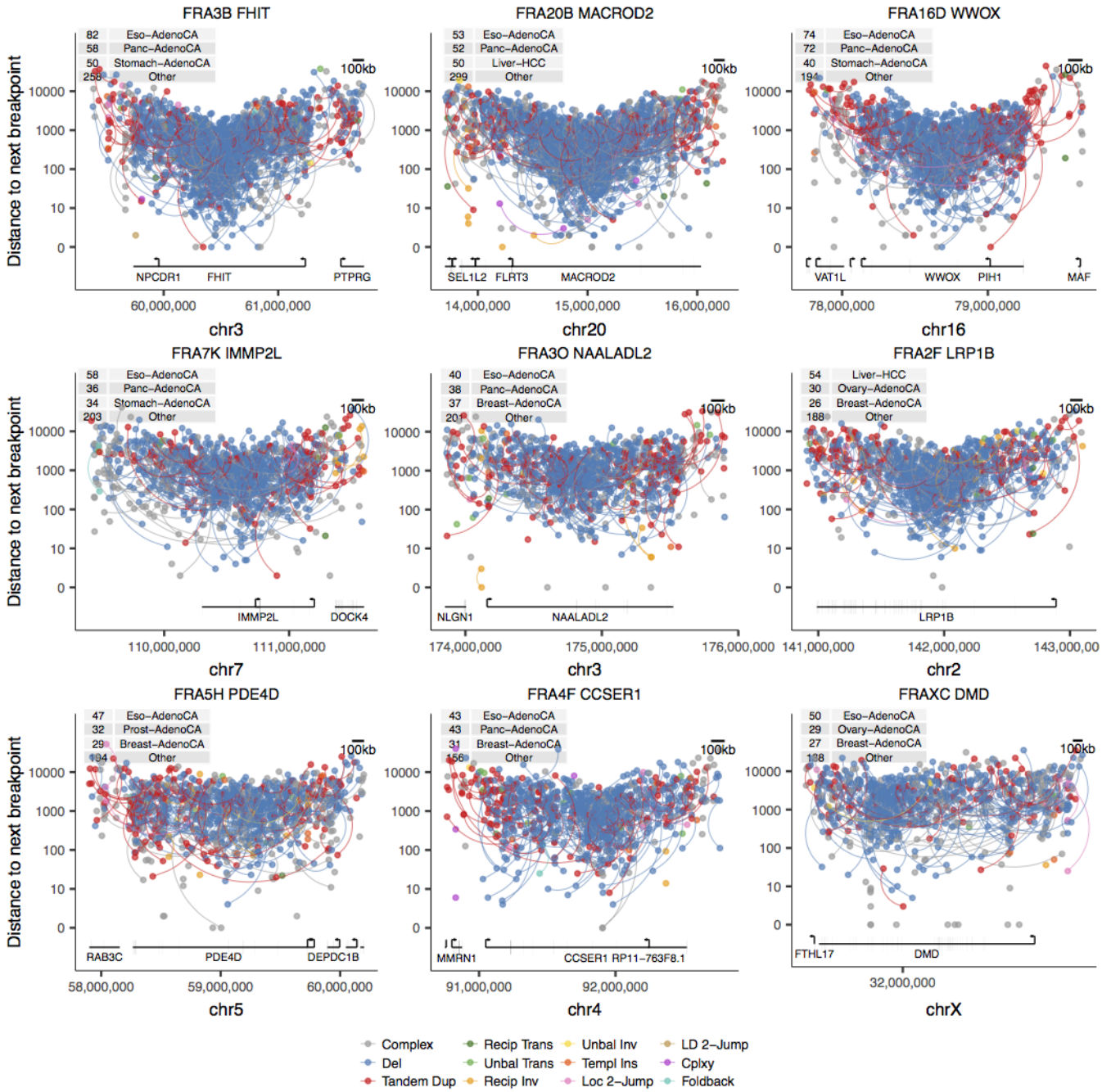


Figure 3.14: All SV breakpoint positions in nine major fragile sites, sorted by number of affected samples. Breakpoint positions are coloured by classification, and vertically spaced by the distance to the next breakpoint in the cohort. If the two sides of a BPJ are contained within the plotting window,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

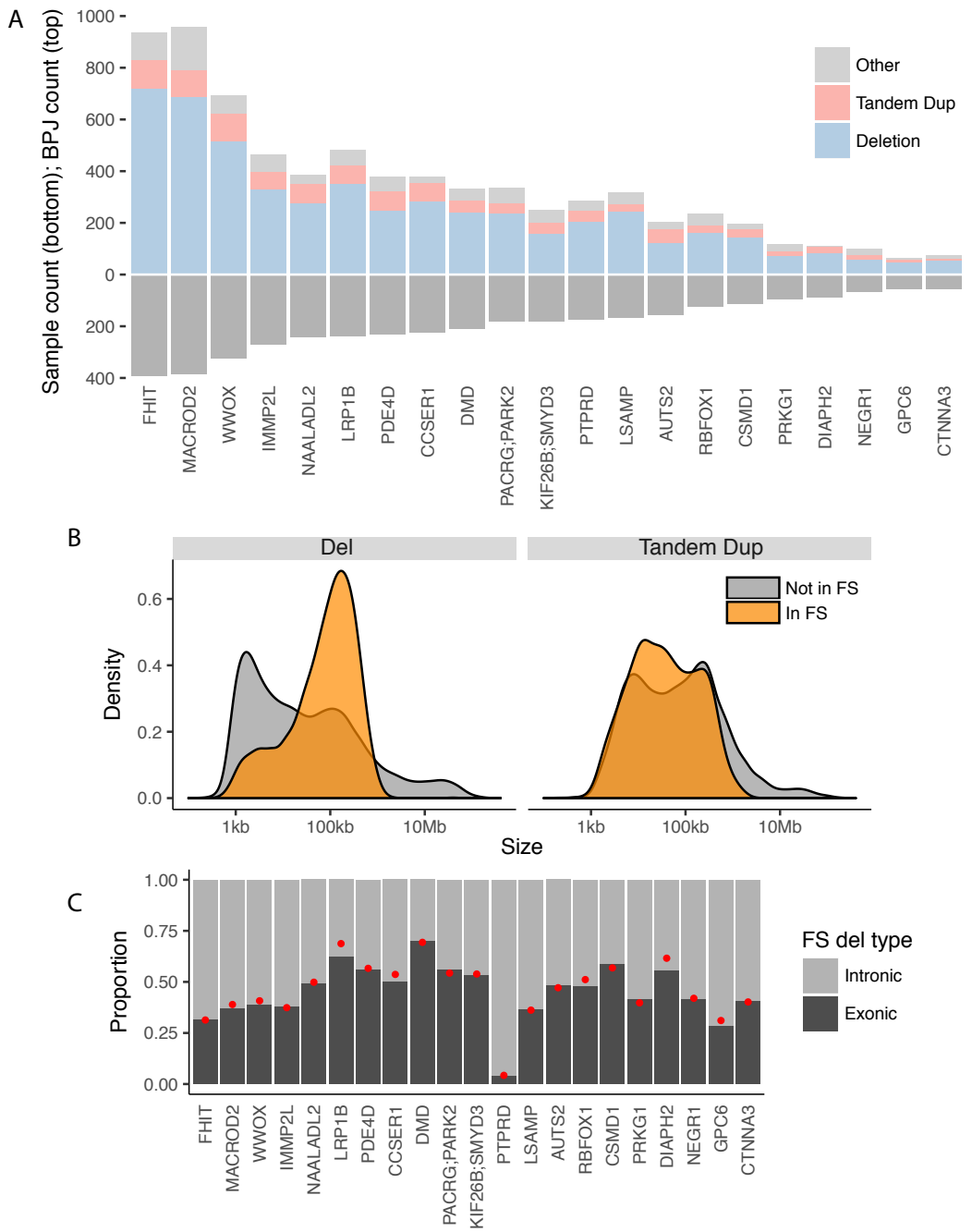


Figure 3.15: (a) number of deletions, tandem duplications and other BPJ within each of the 21 fragile sites considered (upper), sorted by the number of affected samples (lower); (b) size distribution of deletions and tandem duplications in fragile sites compared to the rest of the genome; (c) proportion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

most deletions *and* tandem duplications. The Pearson correlation between the number of FS deletions and FS tandem duplications in a sample is 0.52[n].

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

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

3.4.3 Fragile site deletions are mostly intronic

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

[n] p-value testing null hypothesis of zero correlation is $< 10^{-15}$.

[o] 95% confidence interval for mean difference between FS and non-FS events, using a t-test on the $\log_{10}$-scale and then converting back to the ratio on a base-pair scale.

[p] *DMD* is on the − strand.

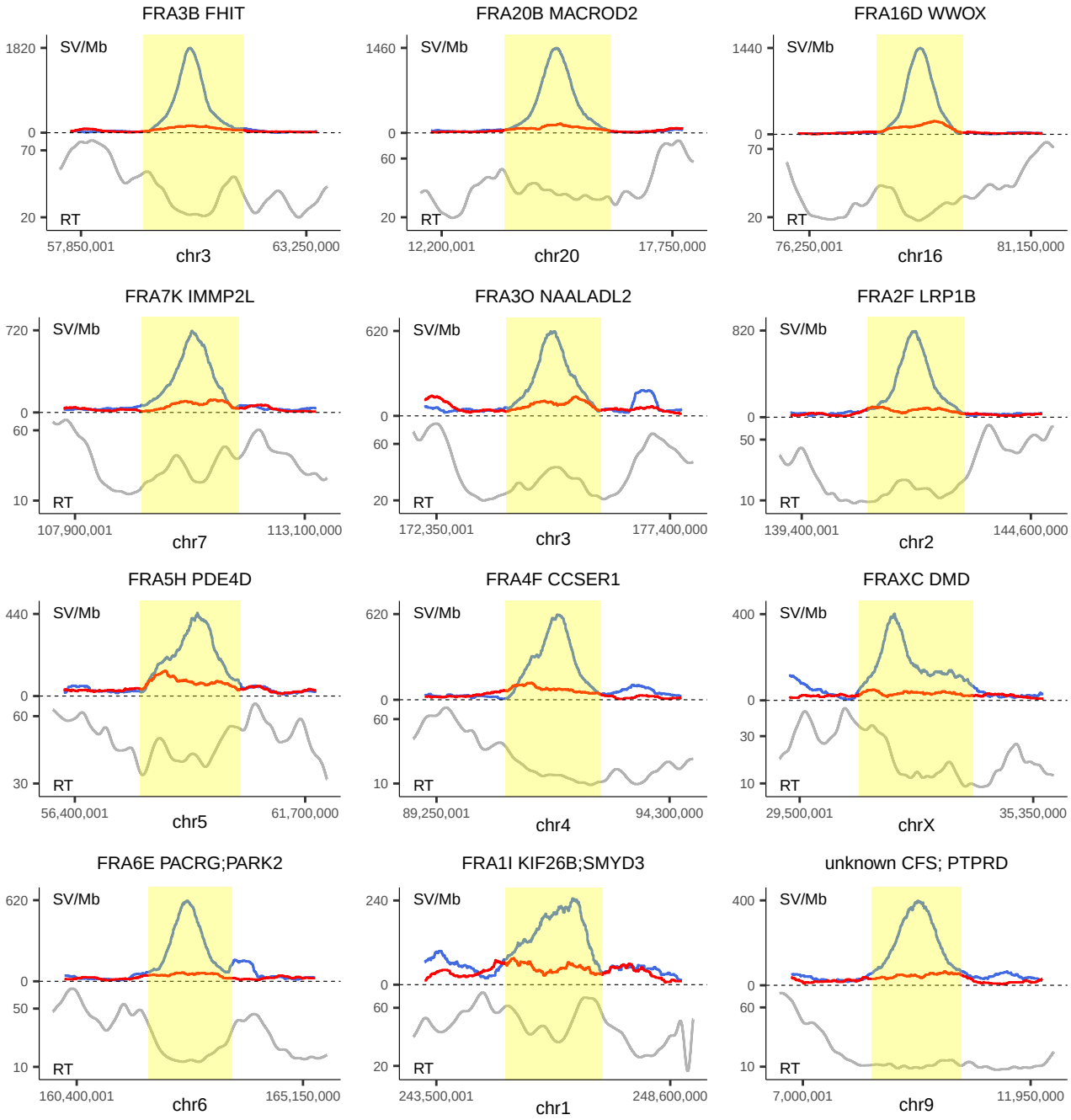


Figure 3.16: The upper plot shows the density of deletion (blue) and tandem duplication (red) breakpoints in 500 kb windows sliding every 10 kb for the 12 major FS marked in yellow, with 2 Mb flanks either side. The lower plot shows the replication timing track, with high values for early and low for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

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

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

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

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

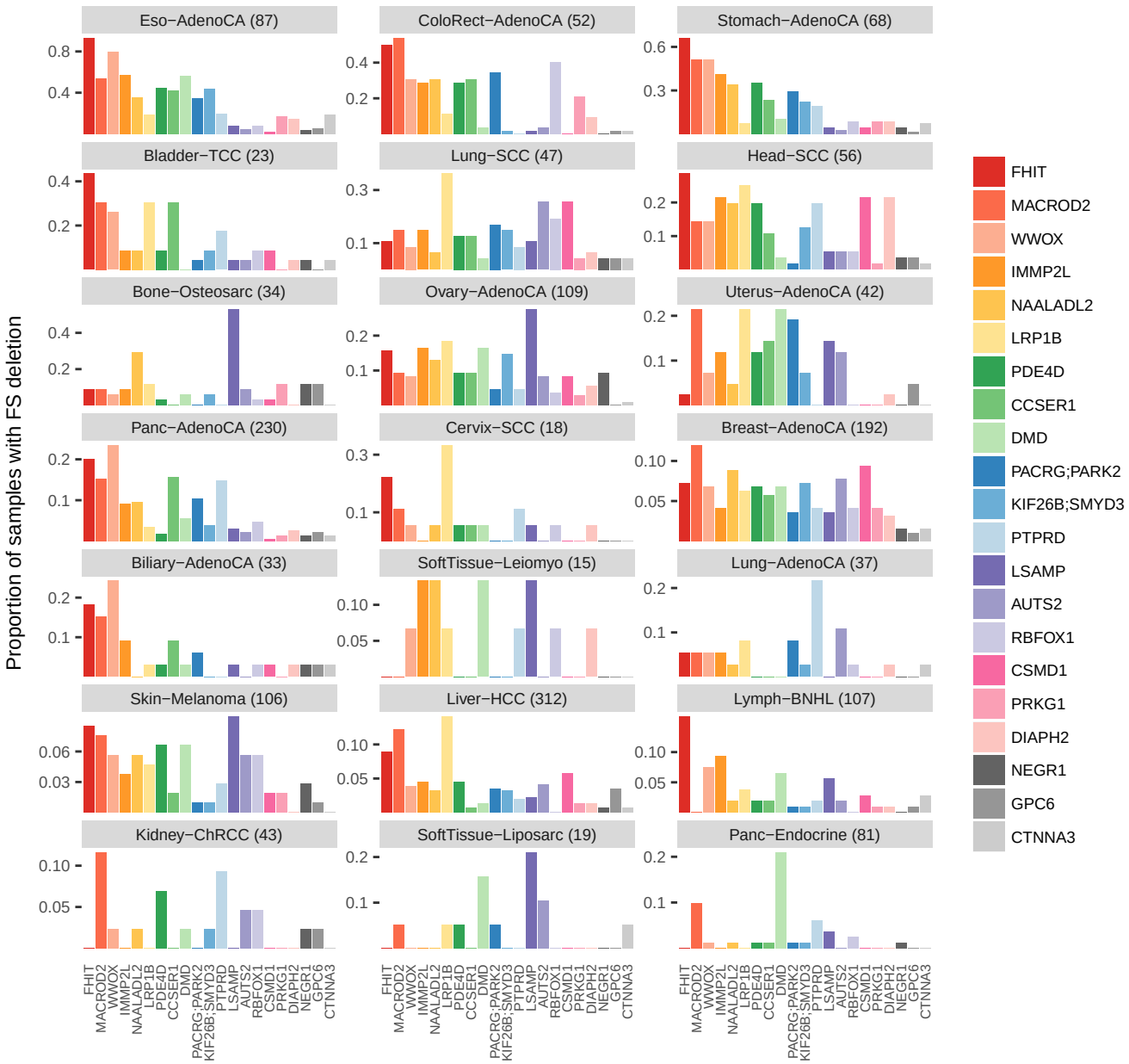


Figure 3.17: Fragile site preference by cancer histology group as measured by the proportion of samples with a deletion in each of the 21 fragile sites considered here. The number of samples is indicated in parentheses.

3.4.4 Tissue specificity of fragile sites

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

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

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

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

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

3.4.5 Complex SV in fragile sites

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

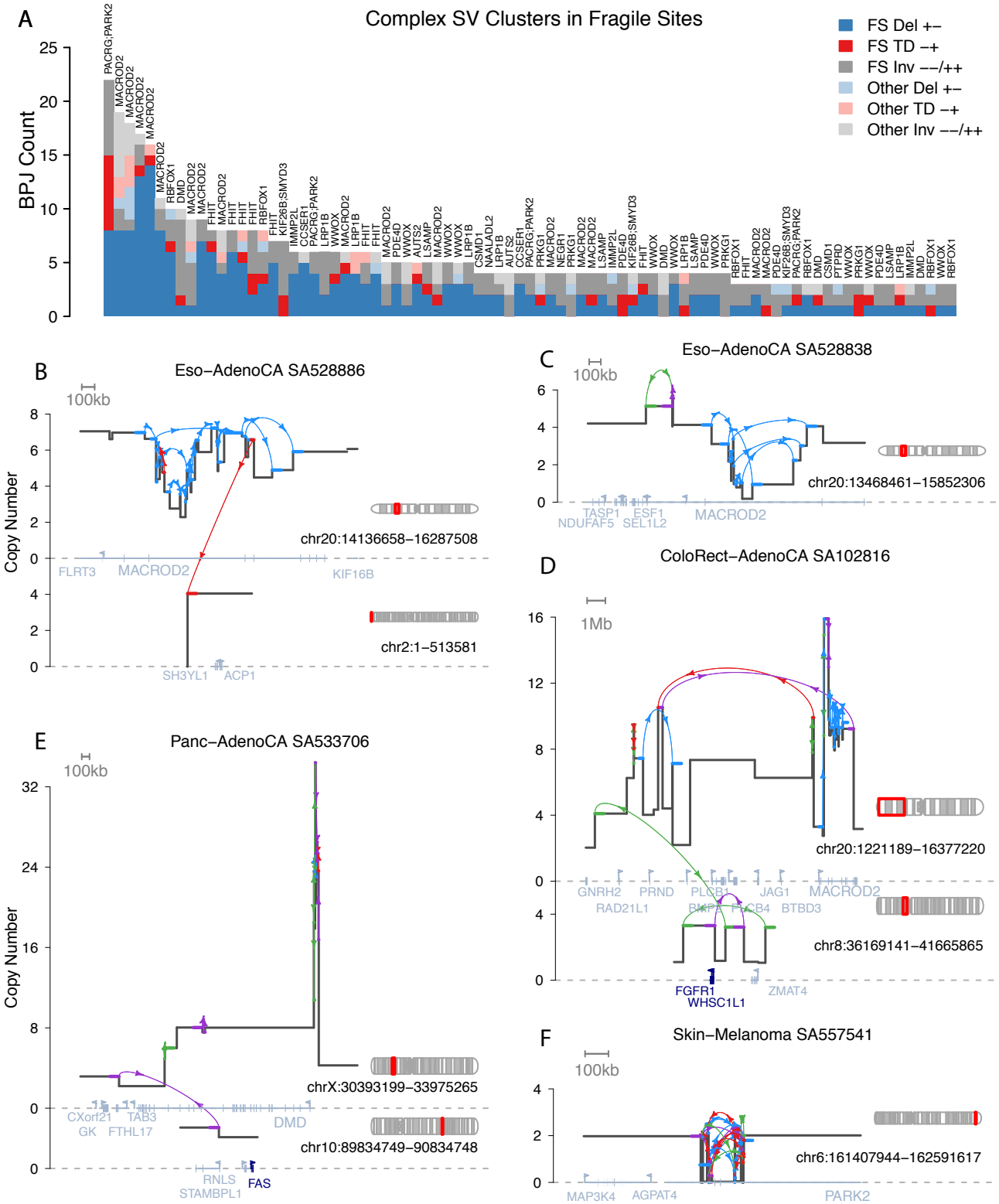


Figure 3.18: (a) 83 complex SV clusters have $\geq 50\%$ BPJ within a fragile site; (b–f) five examples of complex clusters overlapping fragile sites.

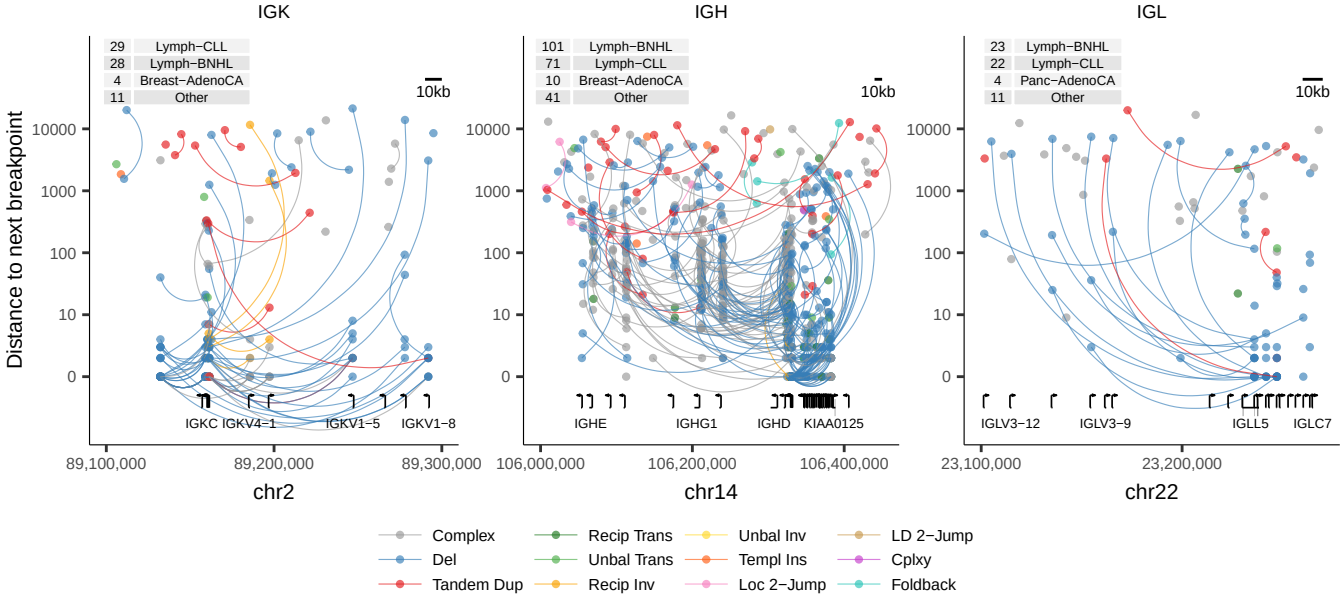


Figure 3.19: All SV breakpoint positions in three immunoglobulin loci: *IGK*, *IGH*, and *IGL*. If the two sides of a BPJ are contained within the plotting window, they are joined with a curved line.

3.4.6 Other anomalous genome regions

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

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

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

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

3.5 Structural variation affecting cancer genes

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

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

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

Continued on next page

Table 3.2 – continued from previous page

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

3.5.1 Cancer genes are affected by different SV classes

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

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

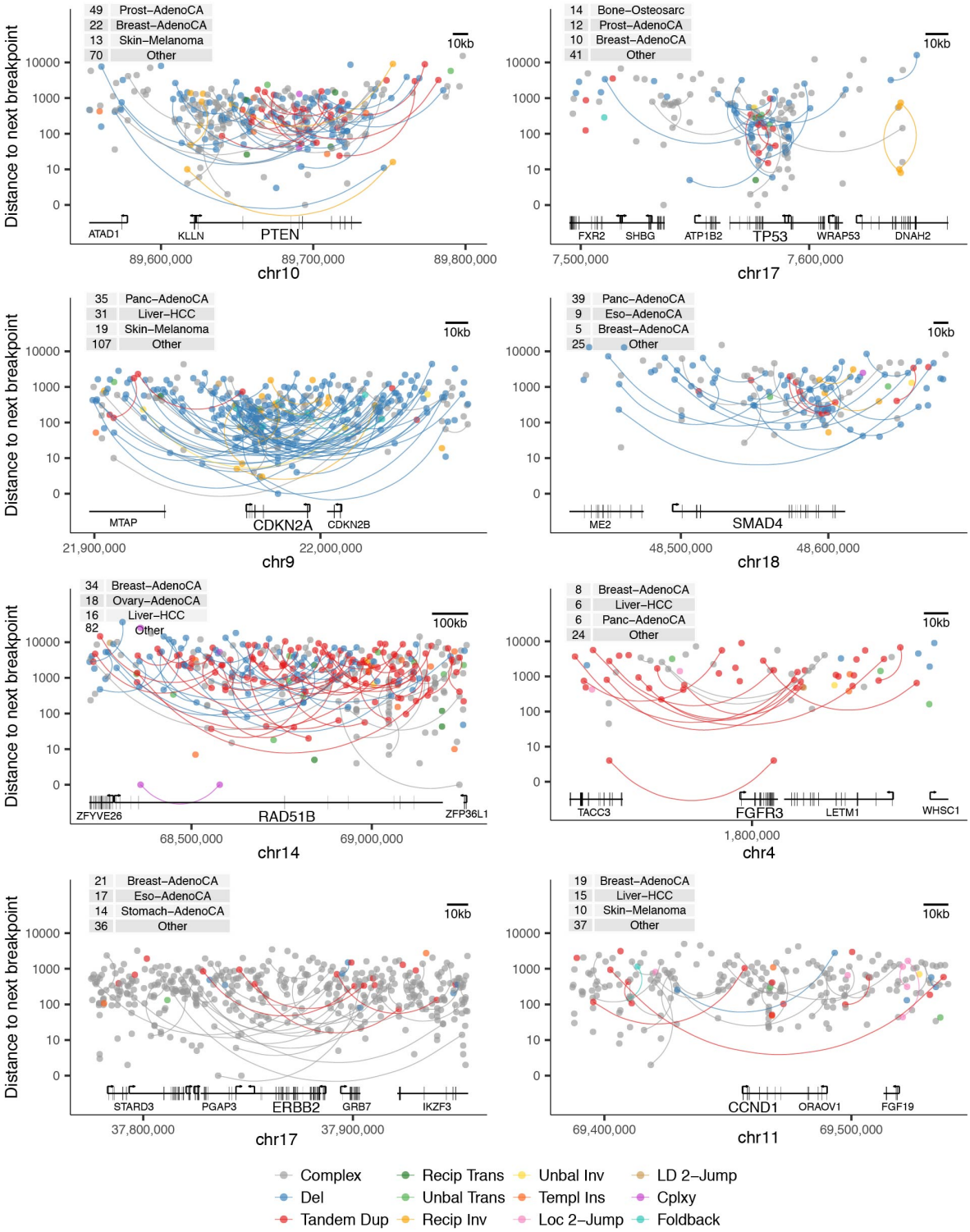


Figure 3.20: SV breakpoint positions around known cancer genes (plus 70 kb flanks). Breakpoints are coloured by SV class, and vertically spaced by distance to the next breakpoint in the cohort. If the two sides of a BPJ are contained within the plotting window, they are joined with a curved line. The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

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

3.5.2 Fusion drivers are formed by different SV classes

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

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

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

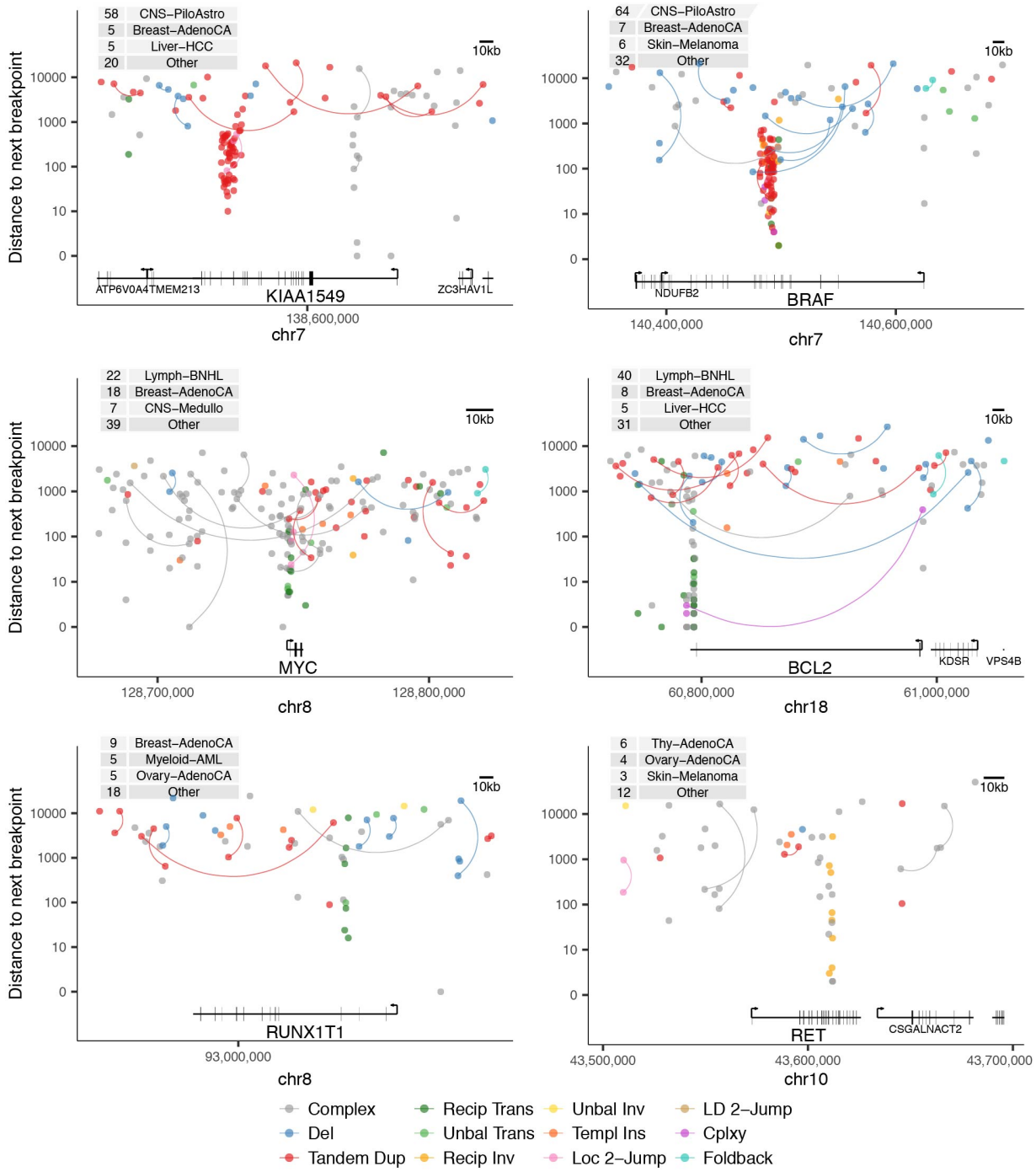


Figure 3.21: Sv breakpoints around six genes (plus 70 kb flanks) with recurrent fusion drivers. The *KIAA1549* and *BRAF* plots illustrate two sides of the same fusion event in pilocytic astrocytoma.

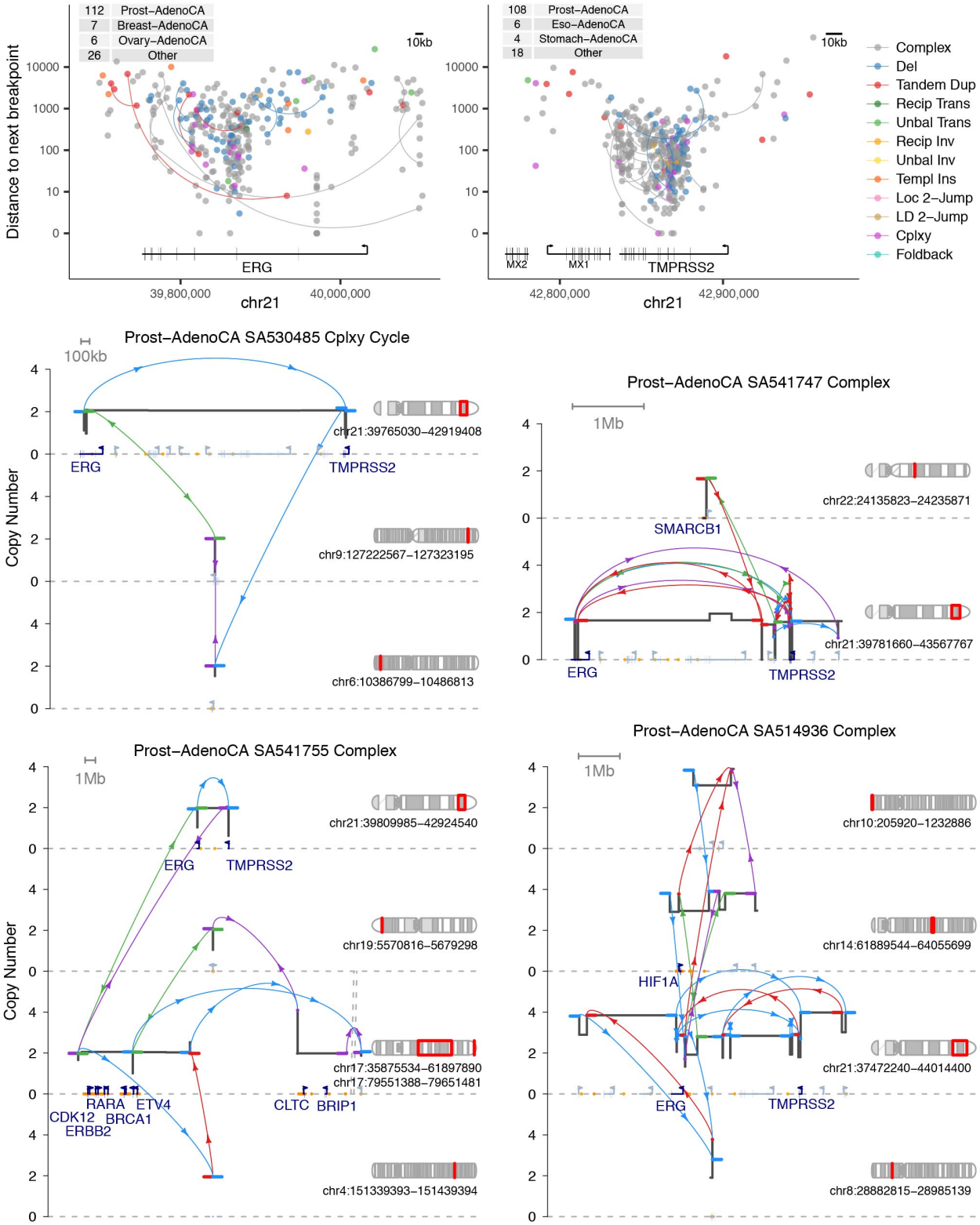


Figure 3.22: Sv breakpoints around *TMPRSS2* and *ERG*, plus four example fusion events in prostate cancer: one simple chromoplexy cycle, and three complex clusters with chromoplexy features. Annotations mark: known cancer census genes in navy; other protein-coding genes in light grey-blue (without names);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

3.5.3 Rearrangement structures around MYC

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

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

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

[r]Figure 3.23 shows one simple classified templated insertion example, and one complex unexplained cluster with features approximating templated insertion.

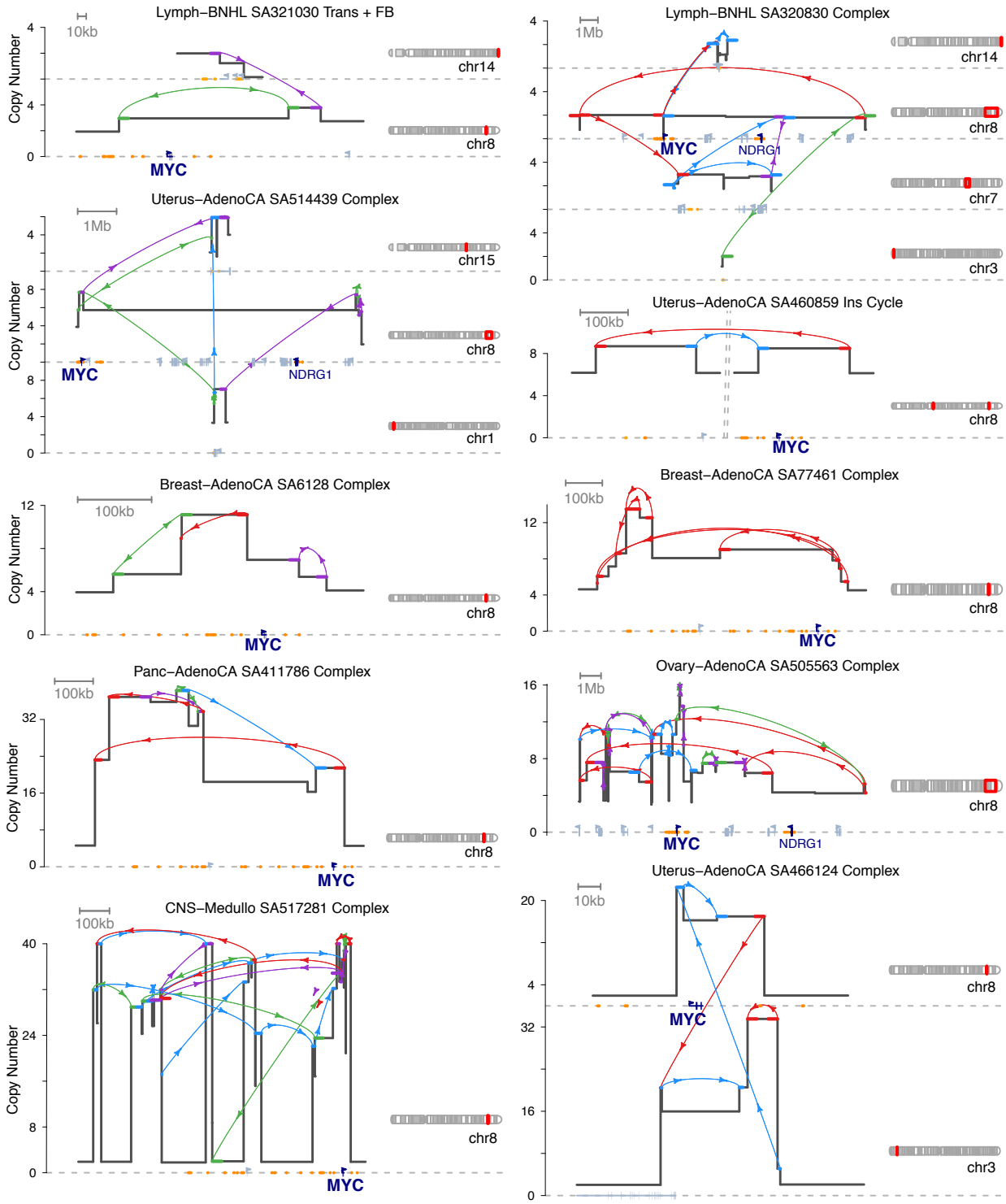


Figure 3.23: Example SV events around the *MYC* oncogene, with annotations to mark: known cancer census genes in navy; other protein-coding genes in light grey-blue (without names); and enhancer sites in orange.

3.5.4 Templated insertion effects

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

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

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

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

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

[u]These 12 templated insertions in *RB1* include manual curation of three complex clusters.

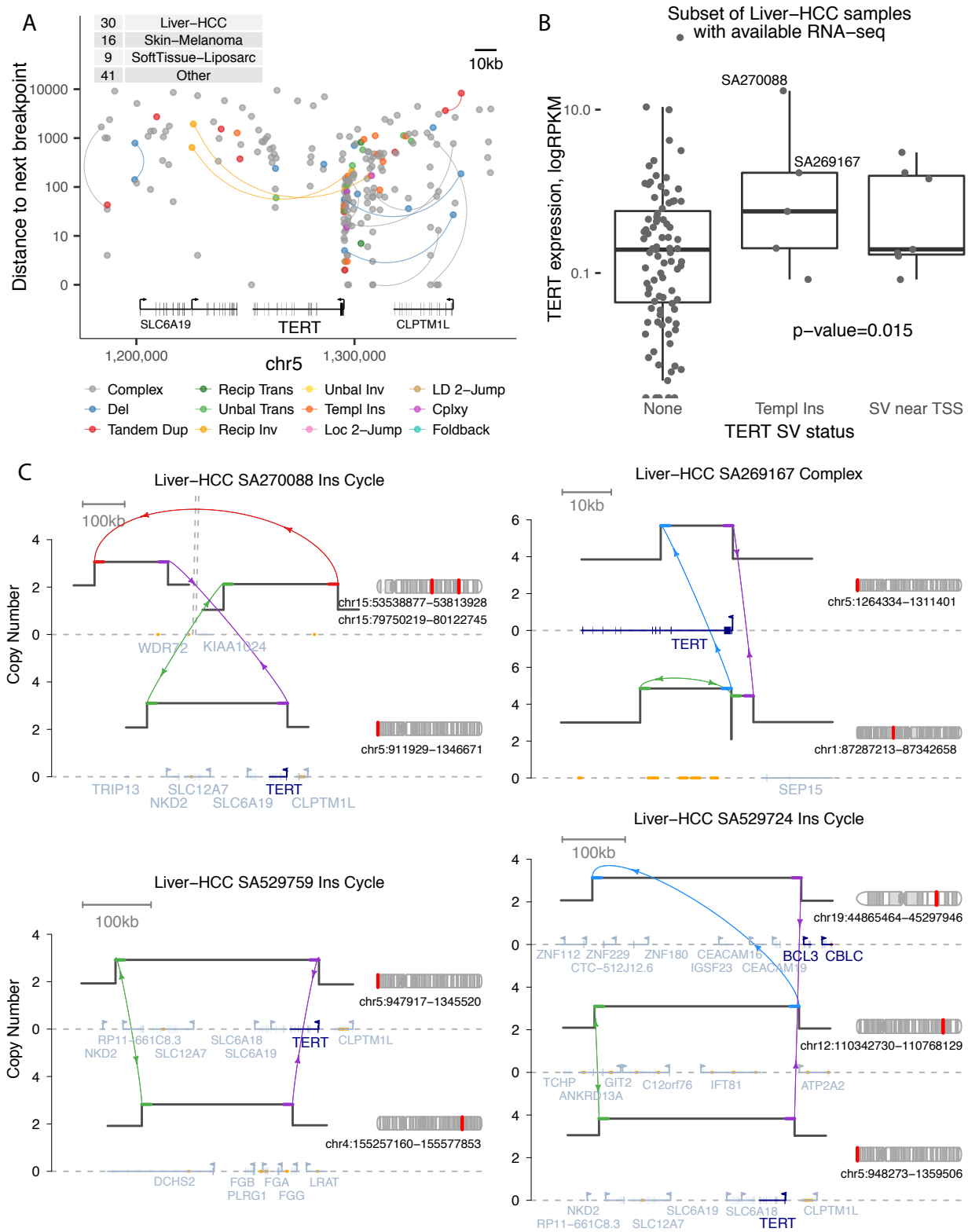


Figure 3.24: (a) SV breakpoints around *TERT*; (b) RNA expression in 100 liver cancer samples, with the *p*-value from a one-sided Wilcoxon test for higher expression in samples with either a local templated insertion or another SV within 20 kb upstream of the TSS; (c) example templated insertion events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

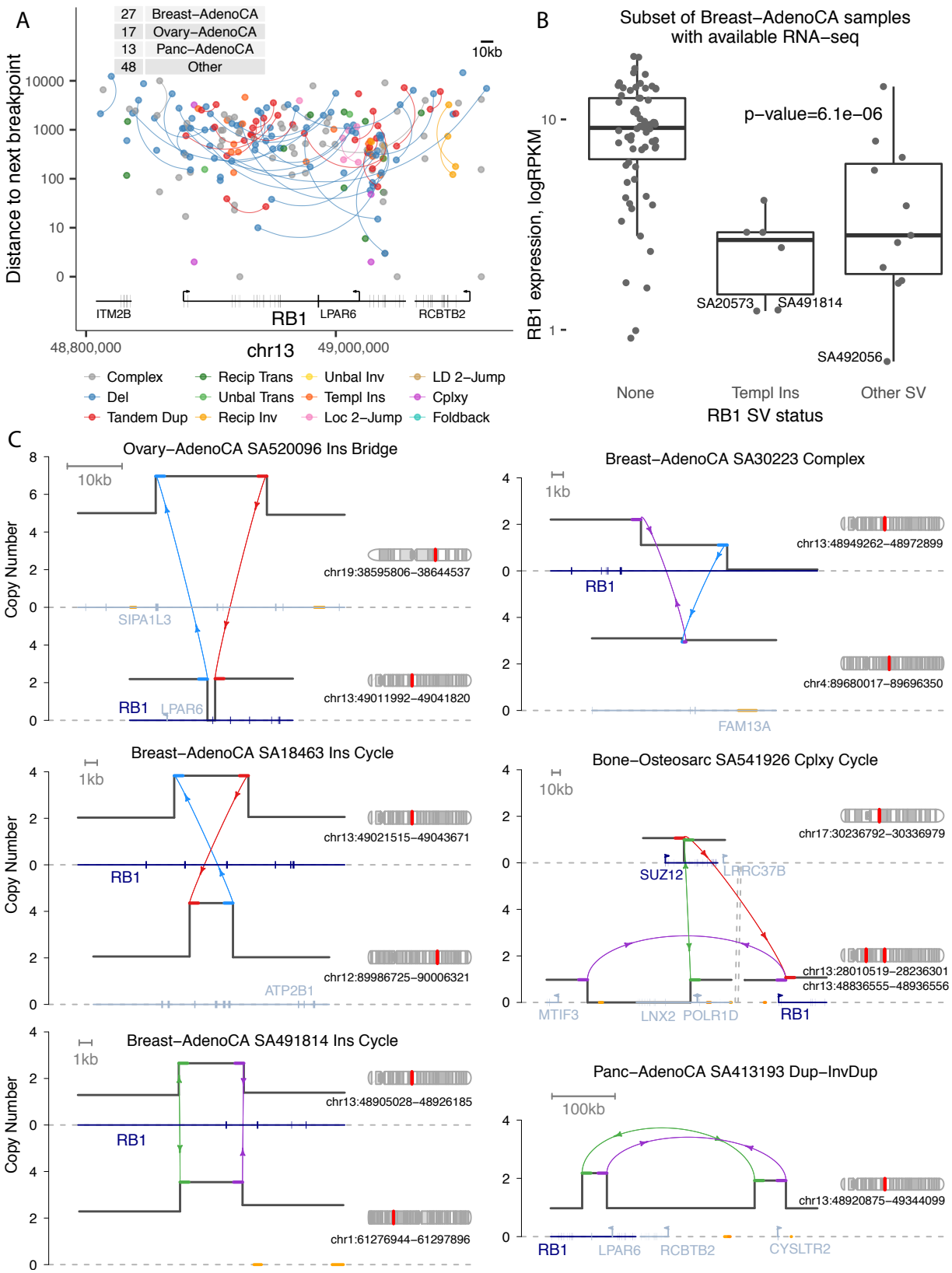


Figure 3.25: (a) SV breakpoints around *RB1*; (b) RNA expression in 83 breast cancer samples, with the *p*-value from a one-sided Wilcoxon test for lower expression in samples with either a local templated insertion or another SV break in the region; (c) example events disrupting *RB1*.

3.6 Discussion

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

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

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

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

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

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

[v]See Chapter 4 for a signatures decomposition of somatic SV events.

Chapter 4

Mutational process signatures: an application for the Hierarchical Dirichlet Process

Somatic genome alterations stem from a variety of underlying processes, including mutagen exposure, replication error, and defective repair. As each different process generates a characteristic distribution or 'signature' of alteration classes, somatic mutation catalogues serve as useful records of historic and ongoing mutagenic activity. To decipher the constituent signatures, observed mutations in a sample cohort are fractionated by their co-occurrence patterns (Nik-Zainal et al., 2012). For SNVs, a subset of about twenty derived signatures have a proposed aetiology as genuine mutational processes or known sequencing artefacts (Alexandrov et al., 2013b; Helleday et al., 2014), often confirmed through experimental data (Segovia et al., 2015; Drost et al., 2017). Cancer sample characterisation by signature exposure has important applications such as: quantifying the effect of environmental carcinogens like aristolochic acid (Poon et al., 2013; Poon et al., 2015), and revealing druggable opportunities like HR-deficiency (with or without *BRCA* loss) (Alexandrov et al., 2015a; Davies et al., 2017; Polak et al., 2017).

In this chapter, I briefly review published methods for mutational signature decomposition (Section 4.1), and describe a different statistical approach using the hierarchical Dirichlet process (Section 4.2). I illustrate the performance of this HDP method on simulated (Section 4.3) and real (Section 4.4) SNV catalogues, and then examine its ability to match new data to an existing signature library while *simultaneously* discovering novel signatures (Section 4.5).

Finally, I return to the PCAWG SV dataset and find fifteen novel rearrangement signatures defined by SV class, size, and replication timing (Section 4.6).

4.1 Existing methods for mutational signature analysis

In the first formal analysis of this kind, Nik-Zainal et al. (2012) used non-negative matrix factorization (NMF) to find five underlying signatures in 21 breast cancer genomes. Alexandrov et al. (2013a) further expounded the details of this NMF application, developing the most widespread signature analysis framework to date. This NMF method assumes each signature is a discrete probability distribution over a finite set of unordered mutation classes. For the $p \times n$ count matrix M which tallies p mutation classes in a cohort of n samples, standard NMF algorithms approximate $M \approx S \times E$, where S is the $p \times k$ matrix of k signatures (constrained to have non-negative columns sum to one), and E is the $k \times n$ sample exposure matrix recording the non-negative burden of each signature in each sample. That is, the sample *exposure* to a signature is the estimated number of mutations generated by that signature in that particular sample. Most SNV studies define mutation classes by the trinucleotide context, yielding $p = 96$[a].

To determine the number of signatures (k unknown), Alexandrov et al. (2013a) calculate NMF solutions at a range of plausible k values for a series of bootstrap-resampled M matrices, returning the consensus solution with stability across bootstraps and minimal reconstruction error by the Frobenius norm. NMF was most notably used to extract 21 validated SNV signatures from 7000 cancer samples analysed by Alexandrov et al. (2013b). However, NMF is not a formal statistical model with probabilistic interpretation, and the choice of Frobenius norm does not account for the integer nature of the input matrix.

Three alternative methods—EMu (Fischer et al., 2013), signeR (Rosales et al., 2016), and SignatureAnalyzer (Kim et al., 2016)—make similar assumptions as the NMF approach, namely that mutations derive from sample-specific mixtures of shared underlying signatures, where each signature is a discrete probability distribution over unordered mutation classes. Crucially, all three methods assume the observed mutation counts follow a Poisson distribution. Their

[a]There are six possible single base substitutions (reported from the pyrimidine side), with four possible flanking bases either side, so $4 \times 6 \times 4 = 96$ SNV classes.

major distinctions lie in the method of estimation and signature number choice. EMu uses expectation-maximisation (EM) iterations to fit signatures and sample exposures until convergence, whereas signeR models the signature probabilities and sample exposures with gamma priors in a Bayesian framework. Both EMu and signeR use the Bayesian information criterion to select a model with high likelihood and few parameters (penalising too many signatures). SignatureAnalyzer aims to minimise the Kullback-Leibler divergence between the NMF solution and input matrix (rather than Frobenius norm), which is equivalent to maximising the Poisson likelihood. To select the number of signatures, SignatureAnalyzer adopts a Bayesian skrinkage methodology (Tan and Févotte, 2013) which automatically determines the relevant components by driving some signature weights to a small lower bound (effectively zero).

Taking a different approach, the 'pmsignature' method (Shiraishi et al., 2015) does *not* consider a mutational signature to be one discrete probability distribution over a large set of mutation classes. Instead of assigning each mutation to just one classification, Shiraishi et al. (2015) model each mutation as having a *set* of observed categorical variables such as substitution type, flanking base identity, and transcriptional strand (in genic regions). In this paradigm, signatures are defined by a collection of distributions over each separate variable, under the simplifying assumption of independence between all mutation features. With this strategy, many relevant features beyond simple trinucleotide context are included within a relatively small parameter space. Pmsignature uses EM iterations to calculate all signature and sample exposure parameters, and selects the number of signatures that yields high likelihood without splitting into multiple components with very similar distributions.

In this chapter, I propose that the hierarchical Dirichlet process (HDP) (Teh et al., 2006) is well-suited to the problem of mutational signature decomposition, particularly in the context of multiple sample groups and/or prior signature information. The signatures defined by the HDP model match the 'traditional' paradigm of one discrete probability distribution per signature, as previously established by NMF and most other methods (with the notable exception of 'pmsignature'). With HDP, a flexible hierarchical model borrows information across samples and groups to identify shared signatures, while also quantifying differences between samples and groups. Under the nonparametric Bayes assumption of infinitely many generating processes, HDP automatically determines the underlying signature number, and can discover novel patterns while simultaneously matching against a prior library of known signatures.

4.2 HDP method for mutational signatures

Teh et al. (2006) first developed the hierarchical Dirichlet process mixture model for the problem of topic modelling in corpora (collections of written text; concept reviewed by Blei (2012)). The HDP is a non-parametric Bayesian clustering method, and infers the number of clusters directly from the complexity of the dataset by assuming the data is drawn from some finite subset of infinitely many generative processes. In Appendix B, I describe the HDP for the novel use case of signature patterns within somatic mutation catalogues.

4.2.1 HDP overview

An overview of the HDP model is shown in Figure 4.1, as designed for multiple groups of samples. Other designs with different hierarchical levels are also possible; for example, an additional child node layer could capture multiple samples from the same individual.

In brief, mutations observed in each sample are tallied into discrete, unordered categories. I assume these mutation counts are randomly drawn from a sample-specific mixture of an infinite number of multinomial distributions (the signatures) over the set of possible mutation classes. Under the HDP model, the sample-specific signature distribution is a Dirichlet process (DP) draw from the group-specific signature distribution. A DP can be intuitively understood as taking in one probability density function, and outputting a sparser, more discretised probability function defined on the same domain[b]. That is, the signature distribution in a sample is based on the parent distribution of its group, but with the probability density further concentrated at particular values/signatures. Moving up one hierarchical level, the group-specific signature distribution is itself a DP draw from the overall distribution of signatures in the dataset. At the top level, the dataset-specific signature distribution is a DP draw from the uniform probability over the infinite set of all possible signatures.

In practice, we observe the mutation catalogues at the bottom of the tree, and specify the uniform Dirichlet prior at the top of the tree, but must estimate the signatures (their identity and prevalence) at each node in-between. To

[b]To use the stick-breaking analogy, a DP draw is built from an infinite random sample from the input distribution, weighted by an infinite series of successive weights randomly broken off an imaginary 'stick' of unit length. A concentration parameter controlling the proportion of 'stick' broken off each time (rate at which the weights attenuate) controls the degree of sparsity in the output.

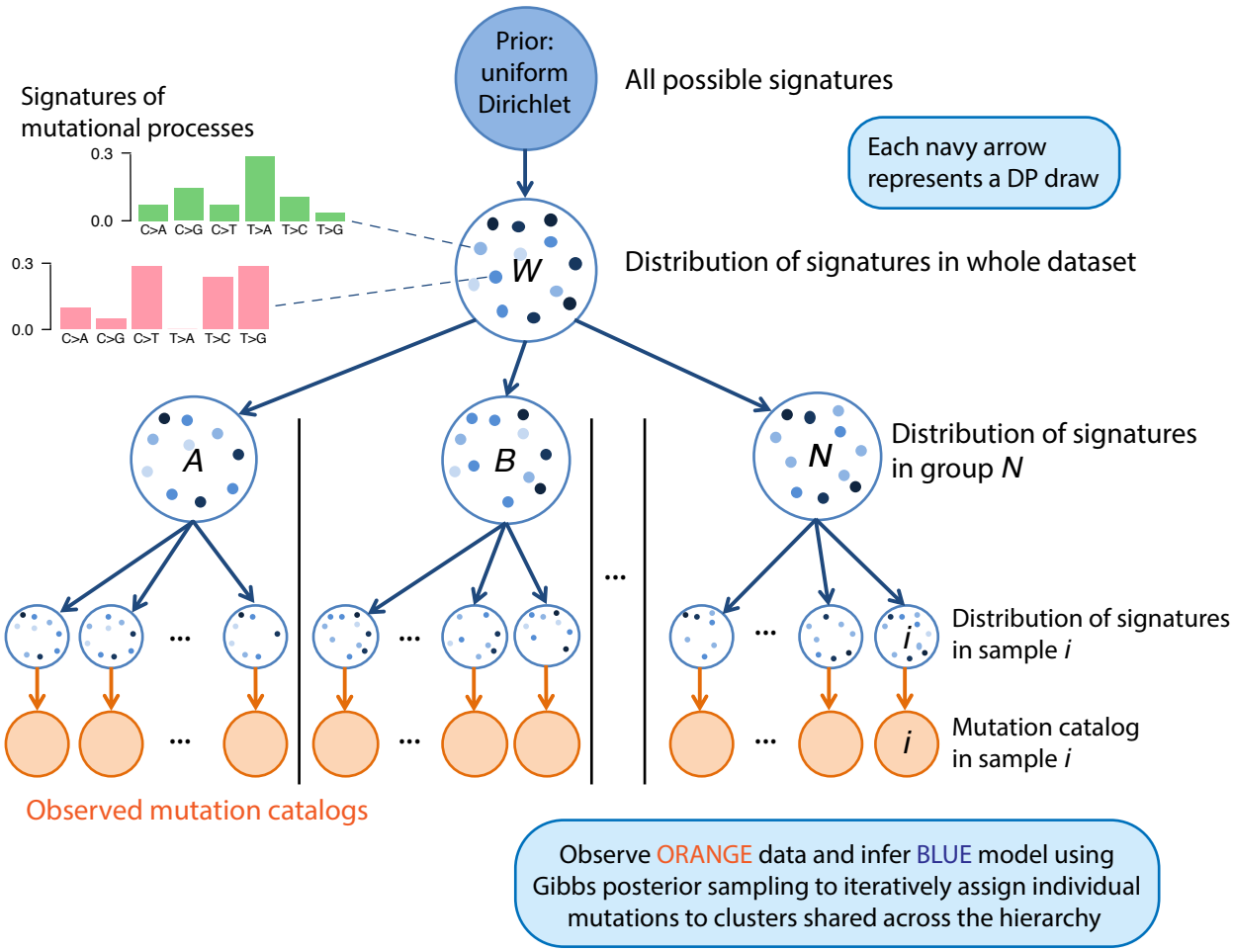


Figure 4.1: Schematic overview of the HDP model for multiple sample groups. Each blue node represents a distribution over the infinite set of all possible signatures, and is a Dirichlet process draw from its parent node. At the top of the tree, the prior distribution is uniform over all possible signatures. Each successive child node concentrates the probability density at particular signature values. The small blue dots inside each node represent particular signatures (discrete probability distributions over the mutation classes), with different shades representing the probability of that signature in the node. The two example signatures over six mutation classes are illustrative only; in practice, mutations are classified into more specific groups (e.g. 96 SNV classes in a trinucleotide context). At the bottom of the tree, the observed data are per-sample mutation catalogues (tallies of mutation classes), assumed to be drawn from sample-specific multinomial mixtures.

perform this posterior inference with the Gibbs sampling method by Teh et al. (2006), all observed mutations are initialised with a random cluster allocation (clusters of mutations define the estimated signatures, and are shared across nodes/samples). Then, the Gibbs procedure cycles through each mutation in turn and assigns an updated cluster allocation, most likely moving to a cluster with: (a) a high proportion of that same mutation class (across all samples); and/or (b) a high proportion of mutations in that sample and/or parent DP (across all mutation classes). At any iteration, there is also a small chance (controlled by the concentration parameter) that a mutation gets assigned to a brand new cluster by itself. In this way, the number of clusters fluctuates throughout the MCMC sampling chain, and there is no need to specify how many clusters should be found. Concentration parameters for each DP are sampled from a Gamma hyper-prior as one of the Gibbs iterations. More details are available in Appendix B. After a burn-in period, posterior samples taken at regular intervals provide a snapshot of possible cluster allocations that defines the space of probable signatures and their prevalence at each node.

Originally, Teh et al. (2006) implemented this Gibbs scheme for the HDP as a suite of functions written in MATLAB and C[c]. To encourage the adoption of HDP in the bioinformatics community, I developed the open-source R package `hdp` as a practical front-end to the original C engine for MCMC inference (R Core Team, 2017; Roberts, 2015). In addition to providing a user-friendly package with detailed documentation and examples, I also developed a suite of post-processing functions for practical reporting across MCMC chains, and a convenient method for setting up pseudo-counts in frozen nodes to condition on prior knowledge. Although this work was motivated by mutational signatures analysis, the utility of my `hdp` package extends to any similar problem involving categorical count data, and was used by Papaemmanuil et al. (2016) to cluster co-occurring driver alterations in acute myeloid leukaemia. Given the range of applications, the package documentation refers to the generic nomenclature of 'components' rather than mutational signatures.

4.2.2 Extracting consensus signatures

Each posterior sample collected off an MCMC chain consists of per-mutation cluster allocations. This output is not immediately amenable to direct reporting because:

[c] `http://www.stats.ox.ac.uk/~teh/research/npbayes/npbayes-r21.tgz` available as of December 2017

- the number of raw clusters varies across posterior samples;
- many raw clusters are very small, with only a few mutations assigned (because HDP assumes infinitely many underlying signatures); and
- multiple clusters can have the same data distribution (strong signatures sometimes found multiple times).

To extract meaningful output from a collection of posterior samples (ideally from multiple independent MCMC chains), I developed a new post-processing method[d] to hone in on the stable set of consistently returned clusters while consolidating the smaller, transitory raw clusters into an additional component capturing noise and uncertainty. While useful for accessible interpretation, this method loses the variable posterior distribution over the number of signatures. However, as the number of raw clusters in a non-parametric Bayesian model is known to scale logarithmically with the number of observed data items (Teh and Jordan, 2009), I conjecture that the raw clusters do not provide the best biological insight by themselves, and instead propose the following approach.

Let S be the number of posterior samples collected, and $K^{[s]}$ the number of raw clusters in posterior sample s for $s \in 1, \ldots, S$. Each posterior sample s assigns each individual mutation to a raw cluster $k^{[s]} \in 1, 2, \ldots, K^{[s]}$. Let the maximum number of raw clusters be denoted K^m. For p mutation classes, let $\boldsymbol{r}_k^{[s]}$ be a p-length count vector of mutations assigned to raw cluster k in posterior sample s, and $\mathcal{R}^{[s]}$ denote the $p \times K^{[s]}$ count matrix of mutation classes in all raw clusters from that posterior sample.

My method for extracting consensus signatures is as follows.

1. Append $K^m - K^{[s]}$ zero vectors to each $\mathcal{R}^{[s]}$ so all count matrices have dimension $p \times K^m$. That is, $\mathcal{R}^{[s]\prime} = \begin{bmatrix} \mathcal{R}^{[s]} & 0_{p \times \left(K^m - K^{[s]}\right)} \end{bmatrix}$.
2. Match up raw clusters across posterior samples by K^m-centroid clustering of all $\boldsymbol{r}_k^{[s]\prime}$, minimising the Manhattan distance to the median and imposing a cannot-link constraint on raw clusters from the same posterior sample. This enforces a result of K^m super-clusters (components), each with S members all from different posterior samples. Let $\mathcal{C}_\ell = \begin{bmatrix} \boldsymbol{r}_\ell^{[1]\prime}, \boldsymbol{r}_\ell^{[2]\prime}, \cdots, \boldsymbol{r}_\ell^{[S]\prime} \end{bmatrix}$ be the $p \times S$ matrix of all raw clusters (and possibly some zero vectors) assigned to component ℓ for $\ell \in 1, \ldots, K^m$.
3. Merge components with very similar mutation class distributions. Let

[d] Available in the `hdp_extract_components` function within the `hdp` package.

the average mutation class distribution for $\mathcal{C}_\ell$ be

$$\bar{\boldsymbol{c}}_\ell = \left[\sum_{s=1}^{S} \left(\boldsymbol{r}_\ell^{[s]\prime} \Big/ \left\| \boldsymbol{r}_\ell^{[s]\prime} \right\|_1 \right) \right] \Big/ S \,.$$

If $\text{cosine.similarity}(\bar{\boldsymbol{c}}_a, \bar{\boldsymbol{c}}_b) \geq 0.9$, then $\mathcal{C}_{\text{new}} = \mathcal{C}_a + \mathcal{C}_b$ [e].

4. Assign components with no significantly non-zero mutation classes to 'component zero' to capture the fraction of noise/uncertainty. Let $\text{HPD}_{0.95}(\boldsymbol{y})$ return the highest posterior density interval containing 95% of $\boldsymbol{y}$ values, so an indicator for the absence of significant mutation classes is

$$z_\ell = \begin{cases} 1 & \text{if } 0 \in \text{HPD}_{0.95}(\mathcal{C}_{\ell,i,:}) \text{ for all rows } i = 1, \ldots, p \,, \\ 0 & \text{otherwise.} \end{cases}$$

Initialise the zero component as

$$\mathcal{C}_{\text{zero\_init}} = \sum_{\{\ell | z_\ell = 1\}} \mathcal{C}_\ell \,,$$

removing non-significant components (with $z_\ell = 1$) from the main set.

5. Assign components with no significantly non-zero sample exposures to 'component zero'. Where previously we have pooled samples and looked at the distribution across mutation classes (p rows), now pool mutation classes and consider the distribution across samples. For n samples (leaf nodes), let $\mathcal{C}_\ell^*$ be the $n \times S$ count matrix of mutations assigned to component ℓ for each sample (row) in each posterior sample (column). An indicator for the absence of significant sample exposures is

$$z_\ell^* = \begin{cases} 1 & \text{if } 0 \in \text{HPD}_{0.95}(\mathcal{C}_{\ell,i,:}^*) \text{ for all rows } i = 1, \ldots, n \,, \\ 0 & \text{otherwise.} \end{cases}$$

Add to the zero component, such that

$$\mathcal{C}_{\text{zero}} = \mathcal{C}_{\text{zero\_init}} + \sum_{\{\ell | z_\ell^* = 1\}} \mathcal{C}_\ell \,,$$

removing non-significant components (with $z_\ell^* = 1$) from the main set[f].

[e] 0.9 is the default similarity threshold for merging components, but can be changed.

[f] An optional variation is to require non-zero sample exposure in two (or more) samples, changing the z_ℓ^* indicator to one if all but one (or more) rows have credibility intervals including zero.

6. Finally, the remaining components are ranked by their prevalence (total number of mutations assigned, averaged over posterior samples) and reported as the set of consensus signatures.

This method returns a set of robust signatures with significant exposure in at least one sample and significant presence of at least one mutation class. The number of signatures is empirically determined without resorting to separate model fitting for every plausible number. A fraction of mutations are assigned to component zero, and reflect the extent of noise and uncertainty in the signature estimation method. Credibility intervals for the mutation classes in each signature, and for the level of signature exposure in each sample and group, are simply constructed as highest posterior density intervals from the set of posterior samples.

4.2.3 Conditioning on prior knowledge

Given the availability of known SNV signatures extracted from large datasets (Alexandrov et al., 2013b; Alexandrov et al., 2015b), it will often be desirable to match a new mutation catalogue to existing signatures, rather than performing *de novo* signature discovery every time. Conditioning on prior knowledge not only saves computational time and effort, but also improves accuracy for small datasets, and leverages existing signature aetiology explanations.

Matching a new dataset to an existing library of mutational signatures is already possible with several methods. For example, with NMF, any mutation tally matrix can be factored into a fixed matrix of known signatures and an unknown sample exposure matrix to be estimated. Alternatively, the 'deconstructSigs' R package by Rosenthal et al. (2016) matches new mutation data to existing signatures with brute-force iterations to minimise the reconstruction error. However, both these approaches are restricted to the pre-defined signature set, and will find a poor solution if the new dataset contains previously unreported signatures, either from cohort-specific mutational mechanisms or a specific profile of artefactual variant calls. Artefacts vary with DNA library preparation, sequencing platform, and bioinformatics calling pipelines, so may appear as novel signatures even in well-studied cancer types.

With its non-parametric Bayes assumption of infinitely many generating signatures, the HDP model is uniquely suited to address this problem, and can *simultaneously* match data to known signatures *and* allow for potential discovery of new signatures.

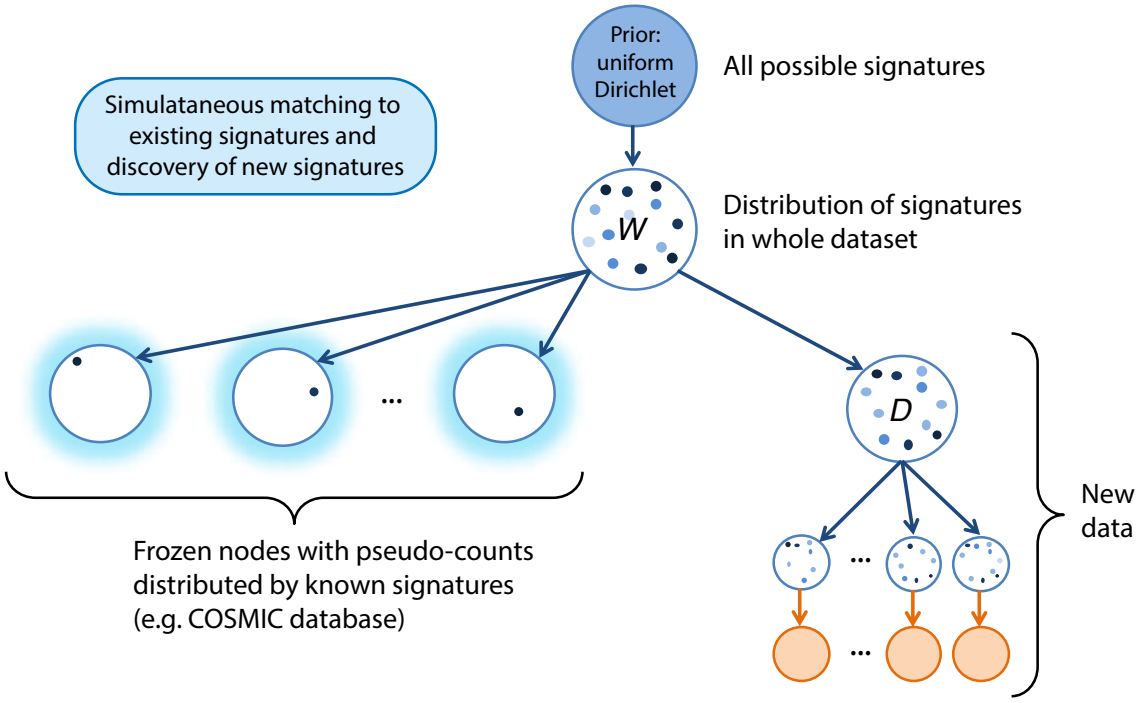


Figure 4.2: Overview of the HDP model conditioning on prior knowledge about known mutational signatures. For each known signature, a number of pseudo-counts following the expected mutation class distribution are assigned to a frozen node and allocated to one fixed cluster. The 'frozen' node status means the cluster allocation of these pseudo-counts is fixed throughout the MCMC posterior sampling. This forces their parent node describing the distribution of signatures in the whole dataset to always apportion some probability to these known signatures. The other nodes behave as in Figure 4.1, and observed mutations in the new dataset are free to cluster either with the fixed pseudo-counts of prior signatures, or in separate clusters describing novel signatures.

The diagram in Figure 4.2 overviews the pseudo-count strategy for conditioning on prior knowledge. When initialising the HDP structure describing the generative model for a new dataset, each known signature is assigned to a 'frozen' child node (DP draw) as shown in Figure 4.2. For each prior signature, the characteristic distribution over mutation classes is instantiated as a set of pseudo-counts fixed to one cluster throughout the posterior sampling process. While the pseudo-counts are held frozen in their cluster allocation, the mutations observed in the new dataset are free to cluster with either the fixed pseudo-counts of a prior signature, or in novel clusters solely composed of new data observations. Following the collection of posterior samples and extraction of consensus signatures (Section 4.2.2), the signatures are labelled by their match in the prior set or with a new label for novel discoveries.

4.3 HDP performance on simulated data

4.3.1 Simulated mutation catalogues

To assess the performance of the HDP method for mutational signatures analysis, I generated a collection of simulated SNV mutation catalogues and compared the HDP reconstruction with the known underlying signatures and sample exposures under a range of conditions.

In total, I simulated 240 separate datasets (parameters in Table 4.1) by varying the number of samples, number of underlying signatures, similarity of sample exposure patterns, number of distinct sample sub-groups, and whether or not the total mutational burdens are consistent with WES or WGS data.

Table 4.1: Parameter combinations for simulated SNV catalogues. Five independent datasets were simulated with every possible combination of parameters within each column, generating 240 simulated datasets in total.

	base combinations	different exposure	three sub-groups
samples	50, 100, 200	50, 100	50,100
generating signatures	5, 10, 15, 20	5, 10	5, 10
seq tech / burden	WES, WGS	WES, WGS	WES, WGS
exposure similarity	medium	low, high	medium
number of groups	1	1	3
replicates	5	5	5
total datasets	120	80	40

Each simulated dataset randomly sampled K underlying signatures from a set of 30 published by the COSMIC database (v74, Forbes et al. (2015)[g]) after NMF analysis of 10,952 exomes and 1,048 whole genomes (Alexandrov et al., 2013b; Alexandrov et al., 2015b). Each COSMIC signature $\boldsymbol{\theta}_k$ is a discrete probability distribution over 96 mutation classes (SNVs in trinucleotide context).

The number of mutations in sample j was taken to be $n_j = \min(\lfloor 10^{x_j} \rceil, 20000)$ for $X \sim \text{Gamma}(\alpha, \beta)$, with shape and rate parameters specific to either WES or WGS data[h].

Next, the signature exposure vector $\boldsymbol{\phi}_j$ for sample j (probability distribution over the set of K signatures) was sampled from $\boldsymbol{\phi} \sim \text{Dirichlet}_K(\boldsymbol{\tau} \times \eta)$, where

[g] `http://cancer.sanger.ac.uk/cosmic/signatures` available as of December 2017

[h] By fitting gamma distributions to the $\log_{10}$-transformed per-sample mutation counts in exome and genome datasets described in Section 4.4, I obtain shape $\alpha_E = 8.23$ and rate $\beta_E = 4.55$ for WES data, and $\alpha_G = 10.02$ and $\beta_G = 3.15$ for WGS data.

the concentration parameters $\boldsymbol{\tau} \sim \text{Dirichlet}_K(\mathbf{1})$ were newly sampled for each simulated cohort, and the exposure similarity weight η was set to 10 for 'medium' similarity across samples, 1 for 'low', and 20 for 'high'. In simulations with three sample sub-groups ($g \in [1, 2, 3]$), the sample exposures were drawn from group-specific distributions with $\boldsymbol{\tau}_g \sim \text{Dirichlet}_K(\mathbf{1} \times 0.5)$. As an additional constraint on the exposure profile, each of the K generating signatures was forced to contribute at least 2% (for $K \in 5, 10$) or 1% (for $K \in 15, 20$) of the total mutations in the cohort.

Finally, the n_j mutations in sample j were drawn from a sample-specific distribution over the 96 mutation classes as defined by $[\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K] \times \boldsymbol{\phi}_j$ (signatures mixed by sample-specific exposure proportions).

4.3.2 Posterior inference settings

For the 240 separate simulated datasets outlined in Table 4.1, I attempted to reconstruct the generating signatures and sample exposures using the HDP model with a variety of settings. Unless otherwise specified, the default HDP design is for one shared concentration parameter across all nodes, with one top parent node modelling the dataset distribution of signatures, and one child node per sample descended from the same shared parent.

The base setting was to collect 500 posterior samples (50 iterations apart) from four independent MCMC chains after 5000 burn-in iterations (2000 posterior samples total). Under the base setting, I initialised all models with 10 clusters, and set the gamma hyper-parameters for the shared concentration parameter at shape = 1 and rate = 1. All 240 datasets were put through HDP clustering with these base settings, and some were also run with additional combinations. As the generating signatures from the COSMIC set include one pair with cosine similarity just below 0.92, I set 0.92 as the similarity threshold for component merging during signature extraction.

To assess the influence of initial clustering, 60 datasets were also run with initial cluster counts of 5 or 15, holding the other settings constant[i].

To assess the influence of the concentration parameter, 40 datasets were also run with a shape hyper-parameter of 0.1 or 10, holding the other settings

[i]For datasets with 50 or 100 samples; WES or WGS burden; 5, 10, or 15 underlying signatures; medium sample exposure similarity; and one shared group.

constant[j].

Finally, to assess the influence of specifying a sub-group structure in cases where it does and does not exist, 80 datasets were also run with a three-group node hierarchy and group-specific concentration parameters, holding the other settings constant[k].

4.3.3 HDP performance on simulated data

Metrics To assess performance of the HDP method, Figures 4.3–4.8 compare the following four metrics:

- number of signatures returned (compare against number of generating signatures, indicated by colour);
- proportion of mutations explained by the fit (proportion of mutations *not* assigned to component zero, averaged over posterior samples);
- cosine similarity of returned signatures with underlying signatures; and
- cosine similarity of estimated sample exposures with the true underlying exposure vectors.

Above each plot is a p-value for the independent variable in question (either a posterior sampling setting, or a property of the simulated dataset) and its relation to the performance metric, controlling for all other variables with a Poisson regression for number of signatures returned, or a beta regression for the other three metrics (defined on a 0–1 scale).

Posterior sampling settings Overall, HDP solutions are robust to the posterior sampling settings, and do not change significantly as the number of initial clusters varies from five to fifteen (Figure 4.4), nor as the mean of the hyper-prior for the concentration parameter varies by a factor of ten (Figure 4.5). Increasing the number of independent MCMC chains from two to eight has little impact (Figure 4.3), indicating that the sampling procedure is mixing around the posterior distribution in a reasonably representative manner even within one chain.

[j]For datasets with 50 or 100 samples; WES or WGS burden; 5 or 10 underlying signatures; medium exposure similarity; and one shared group.

[k]For datasets with 50 or 100 samples; WES or WGS burden; 5 or 10 underlying signatures; medium exposure similarity; and one or three shared groups.

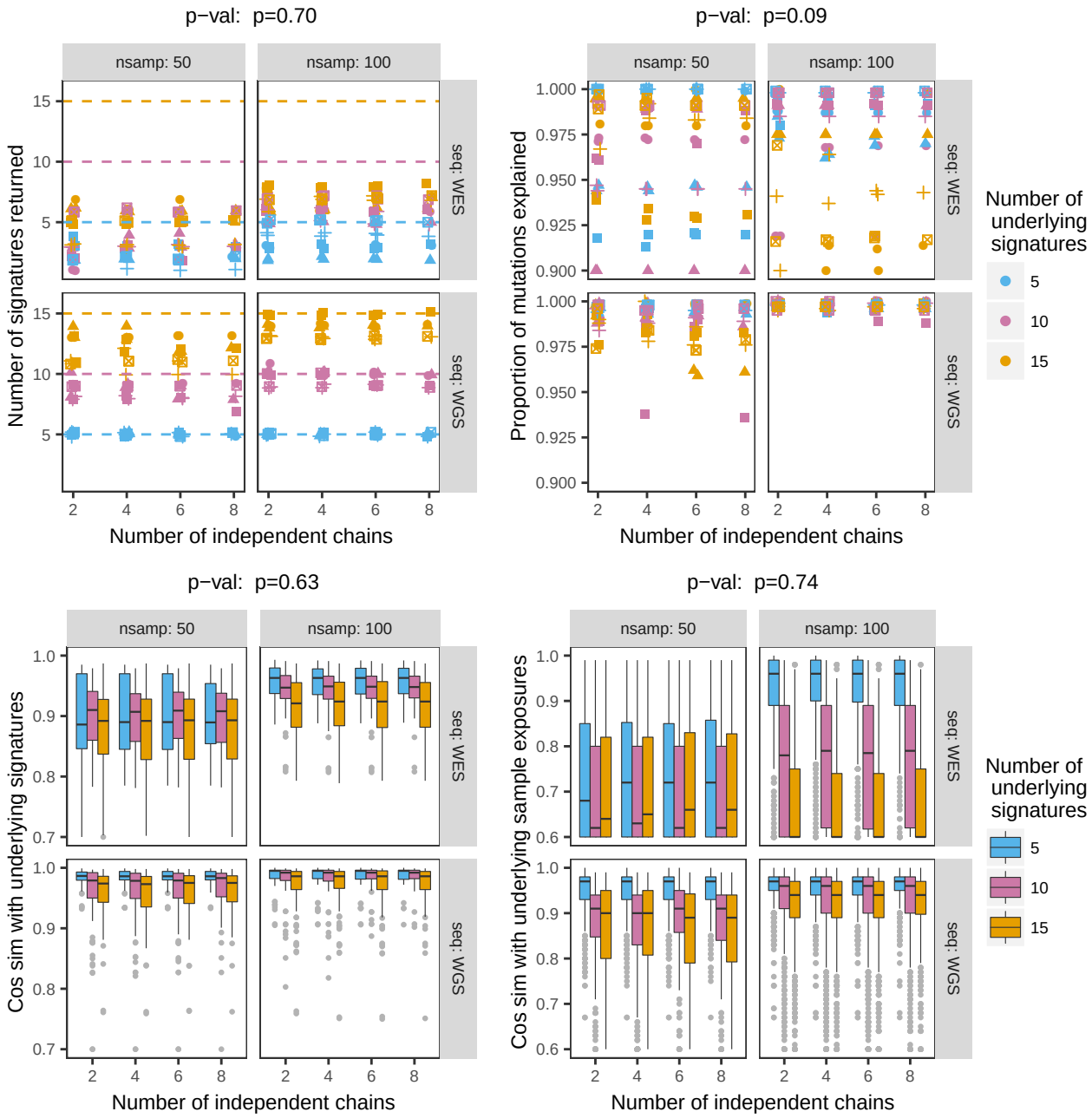


Figure 4.3: HDP performance as the number of independent MCMC chains varies. P-values above each plot are for the number of chains as a quantitative predictor of each vertical axis metric, controlling for number of samples, number of underlying signatures, and sequencing type.

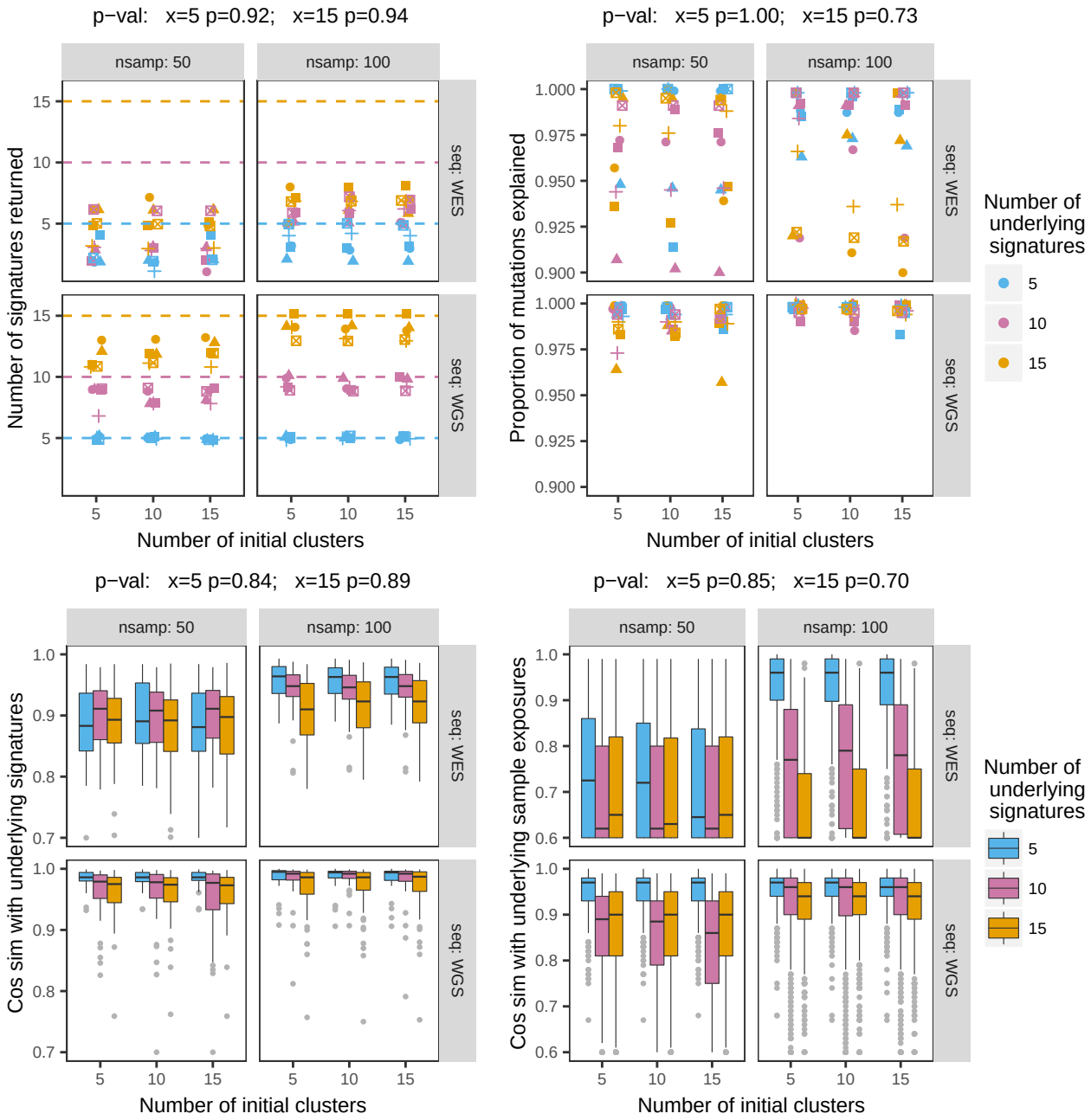


Figure 4.4: HDP performance as the number of initial clusters varies. P-values above each plot are for the number of initial clusters compared to a baseline of 10, controlling for number of samples, number of underlying signatures, and sequencing type.

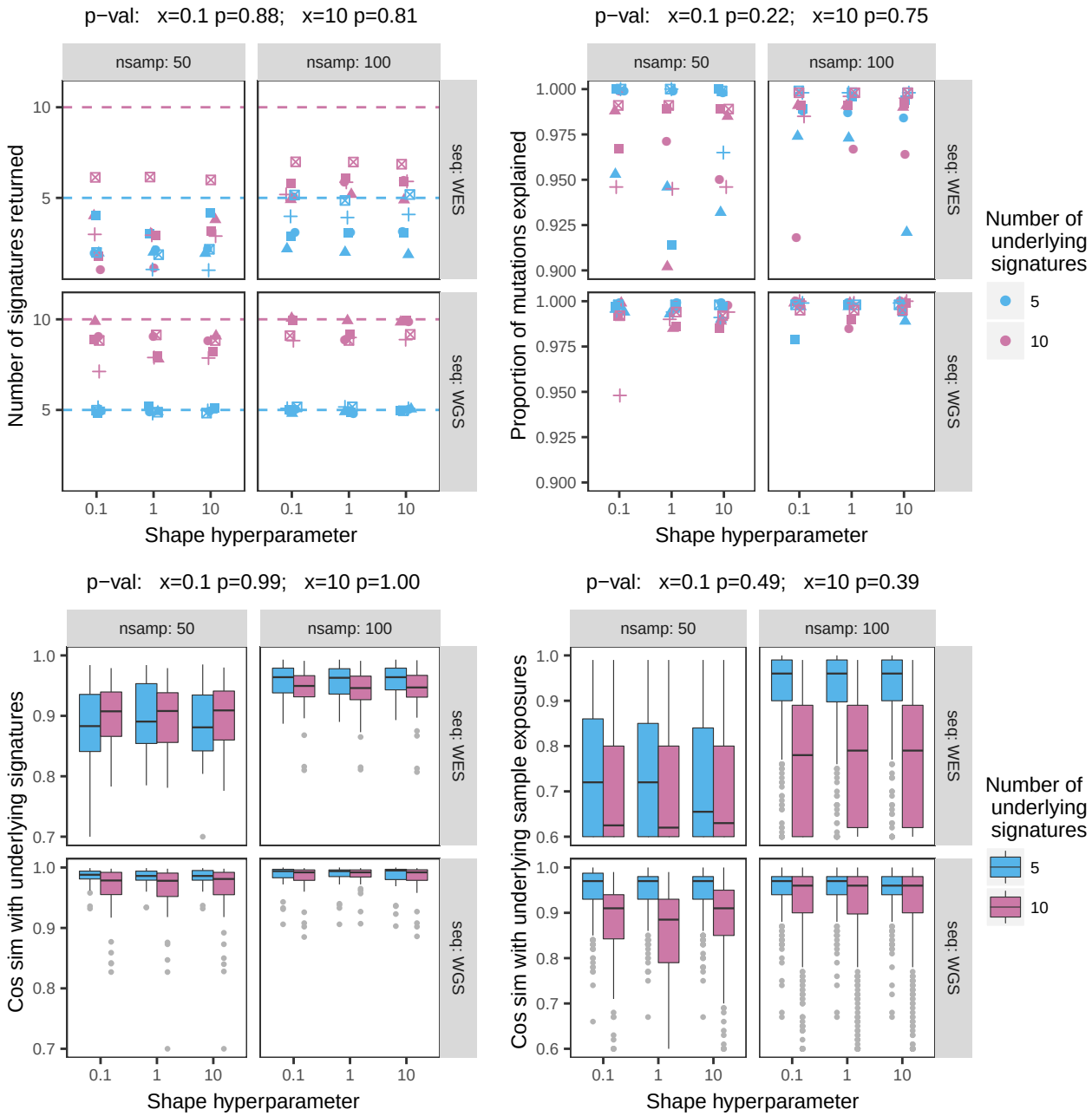


Figure 4.5: HDP performance as the shape hyper-parameter for the DP concentration parameter varies. P-values above each plot are for the shape hyper-parameter levels compared to a baseline of 1, controlling for number of samples, number of underlying signatures, and sequencing type.

Dataset properties The greatest determinants of accurate signature reconstruction are dataset size and sample exposure similarity. As expected, HDP performance improves in larger datasets with more samples and more observed mutations (WGS better than WES, Figure 4.6). Under the simulation conditions established here, five mutational signatures are reliably reconstructed with about 200 exomes or 50 whole genomes. For datasets generated with 15 or 20 underlying signatures, 200 exomes will only reconstruct about 10 of these, whereas 200 whole genomes can accurately return all 15, and almost all 20 signatures. However, these guidelines are heavily dependent on the sample exposure patterns. As shown in Figure 4.7, accurate signature reconstruction requires much less data when samples have very different signature exposures, as the co-occurrence profile is more distinct for variably assorting signatures.

Sub-group structure Finally, for these simulations, the HDP results are broadly similar whether or not the samples' sub-group structure is accounted for (Figure 4.8). Although modelling the sub-group structure has no discernible influence on signature estimation, it does significantly improve the sample exposure estimates when there *is* a genuine underlying difference, and is of no detriment when the sub-group division is erroneous.

Factors influencing accuracy In all the HDP model fits on simulated data, some signatures and sample exposures are more reliably reconstructed than others. To investigate factors influencing reconstruction accuracy, I considered the subset of base setting simulations with 50 or 100 WGS samples with 5 or 10 underlying signatures and medium exposure similarity. Pooling observations across simulated cohorts, I fitted a beta regression for the outcome variable of cosine similarity between the estimation and underlying truth, with predictor variables as indicated in Figures 4.9 and 4.10. This exercise shows that sample exposure recall (Figure 4.9) is more likely to be inaccurate if the number of extracted signatures is incorrect, and/or the sample has: similar contributions from most signatures (low standard deviation); low mutation count; or a higher proportion of mutations in rare signatures. For the signatures (but not the exposures), I subset to the models returning the *correct* number of underlying signatures (eliminating poor reconstruction due to incorrect number). Signature recall (Figure 4.10) is more likely to be inaccurate if the dataset is small, or if the signature in question is: rare in the cohort (contributes a low proportion of total mutations); close to uniform across mutation classes (low standard deviation); or roughly similar to another generating signature in the cohort.

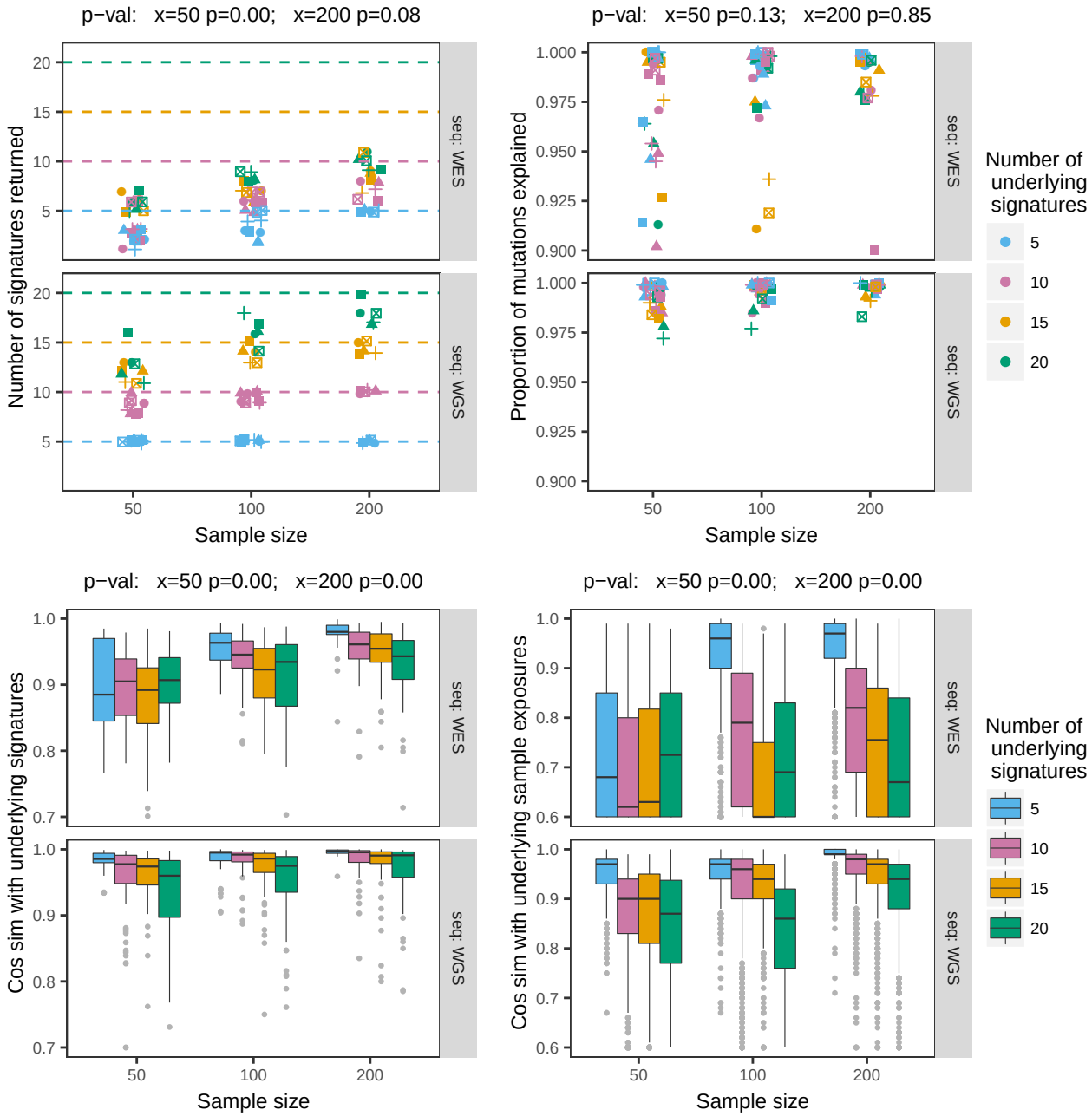


Figure 4.6: HDP performance as the number of samples varies. P-values above each plot are for the number of samples compared to a baseline of 100, controlling for number of underlying signatures, and sequencing type.

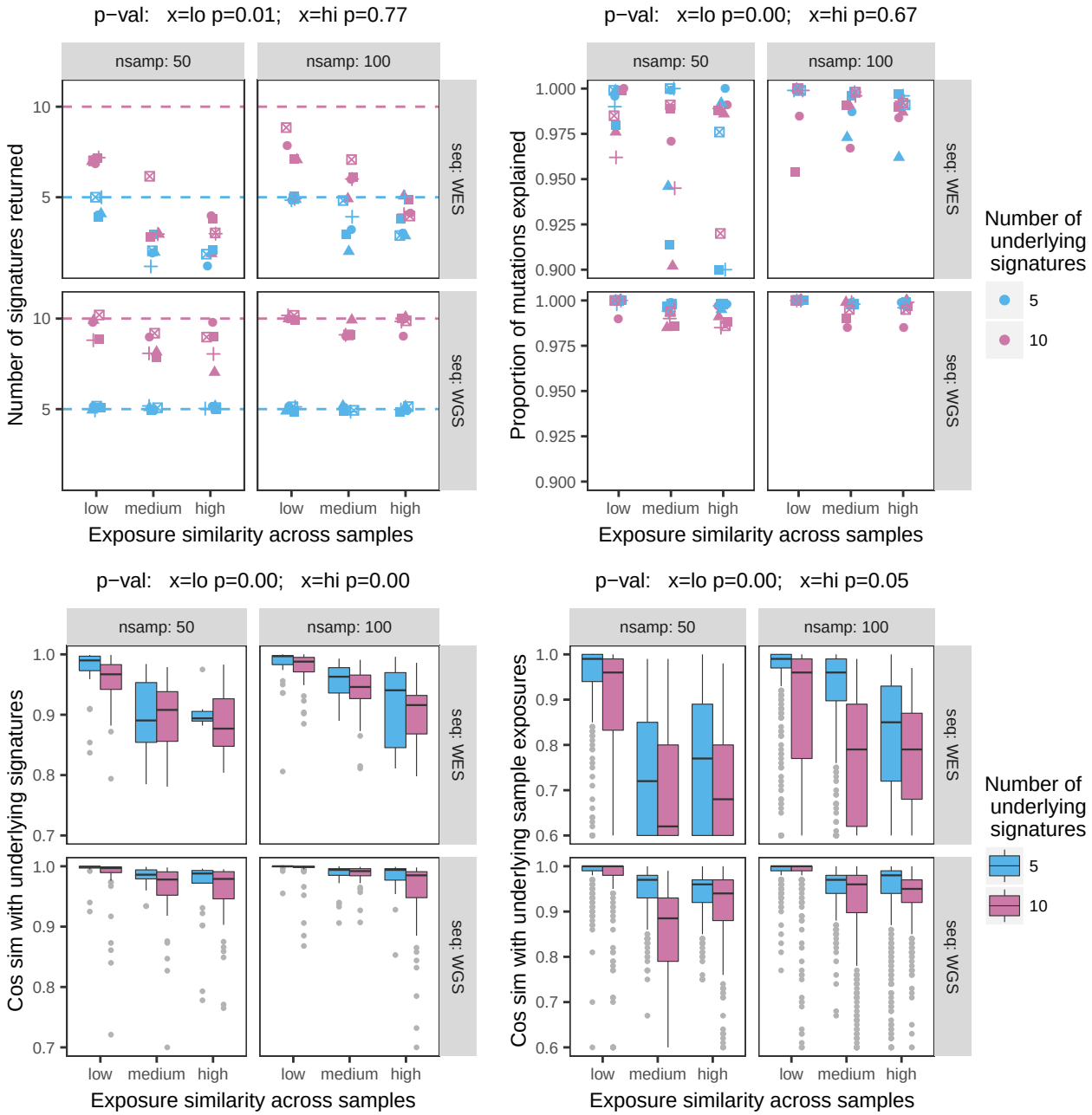


Figure 4.7: HDP performance as the level of signature exposure similarity across samples varies. P-values above each plot are for the level of exposure similarity compared to a 'medium' baseline, controlling for number of samples, number of underlying signatures, and sequencing type. Exposure similarity was controlled by a weight on Dirichlet concentration parameters for the sample exposure vectors when generating the simulated data (Section 4.3.1).

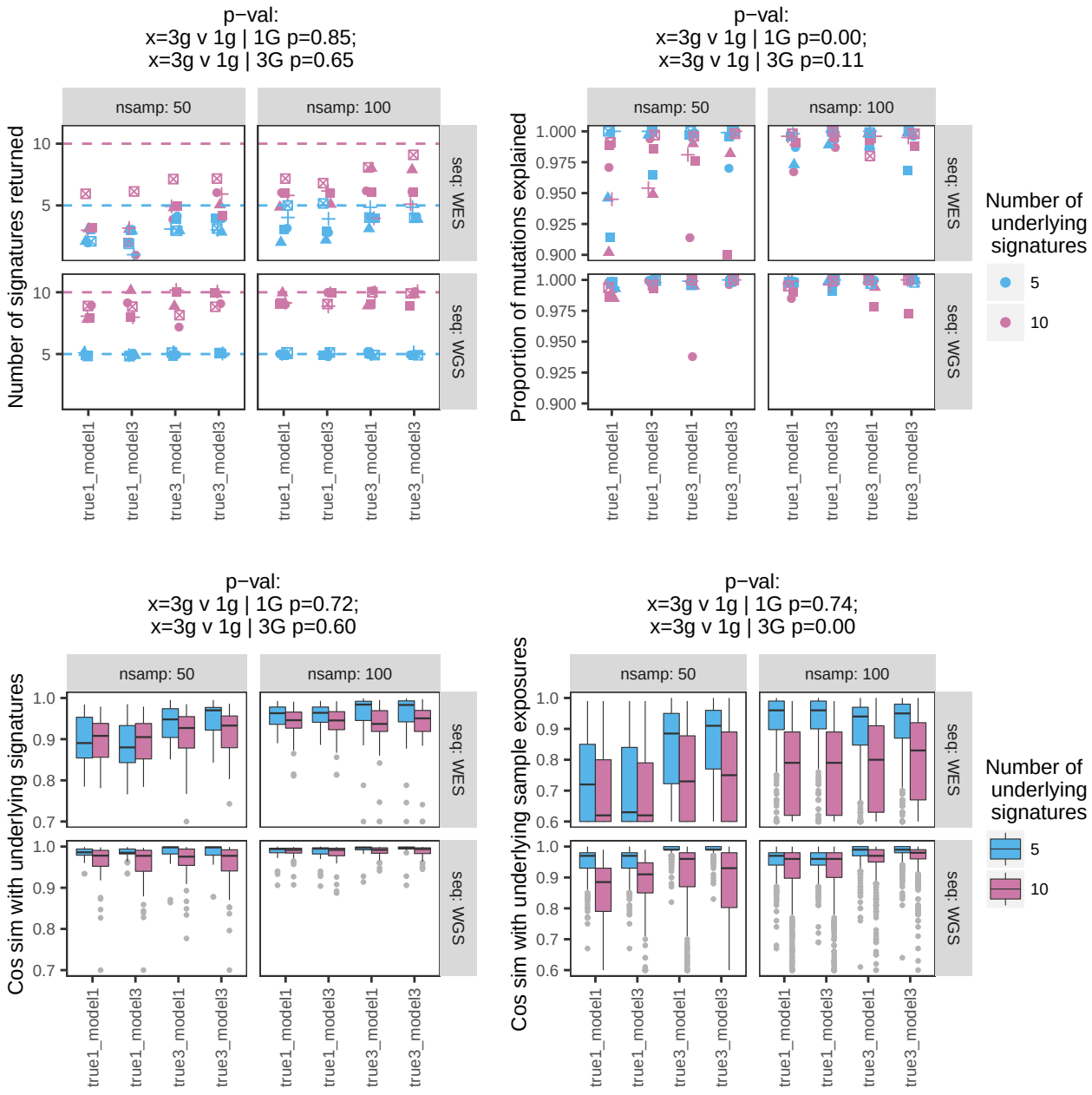


Figure 4.8: HDP performance when modelling the sub-group structure of samples, in cases where this sub-group structure genuinely existed (true3) and in cases where it did not (true1). P-values above each plot compare the 3-group model with the 1-group model, given that the dataset was simulated from one true group or from three true groups. Regression tests controlled for number of samples, number of underlying signatures, and sequencing type.

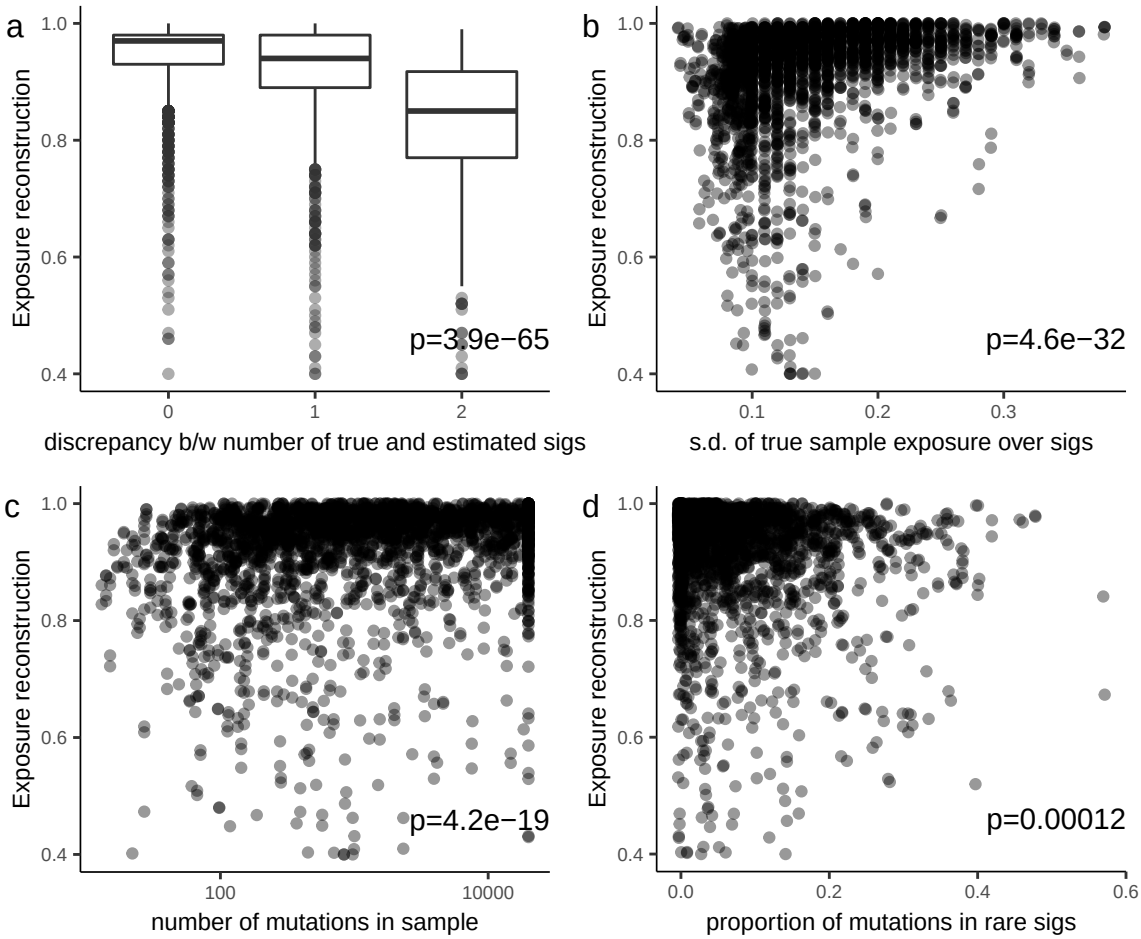


Figure 4.9: Factors influencing cosine similarity between true and estimated sample exposures (vertical axis). P-values shown are from a multivariate beta regression fit on the four predictor variables: (a) discrepancy between the number of underlying signatures and the number of signatures returned by HDP for the cohort; (b) standard deviation of the true exposure values for a sample; (c) number of mutations in a sample; and (d) proportion of mutations in a sample from 'rare' signatures (defined as the maximal subset which cumulatively contribute less than 10% of total cohort mutations).

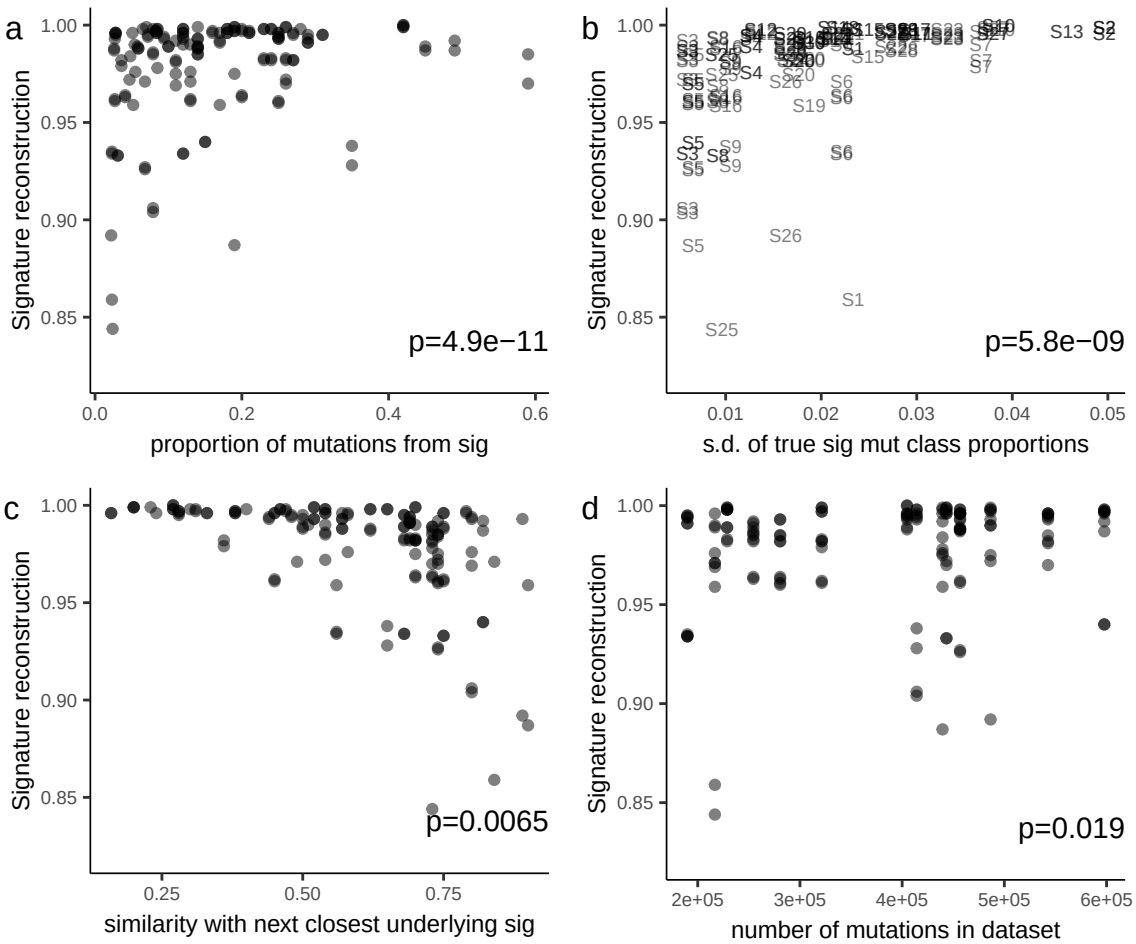


Figure 4.10: Factors influencing cosine similarity between true and estimated mutational signatures (vertical axis). P-values shown are from a multivariate beta regression fit on the four predictor variables: (a) proportion of mutations in the cohort from that signature; (b) standard deviation of the true mutation class probabilities in that signature; (c) cosine similarity with the most similar generating signature in the cohort; and (d) total number of mutations in the cohort. For panel (b), the underlying signatures are marked with their identifier in the COSMIC database (horizontal position for underlying signature s.d. is constant across datasets).

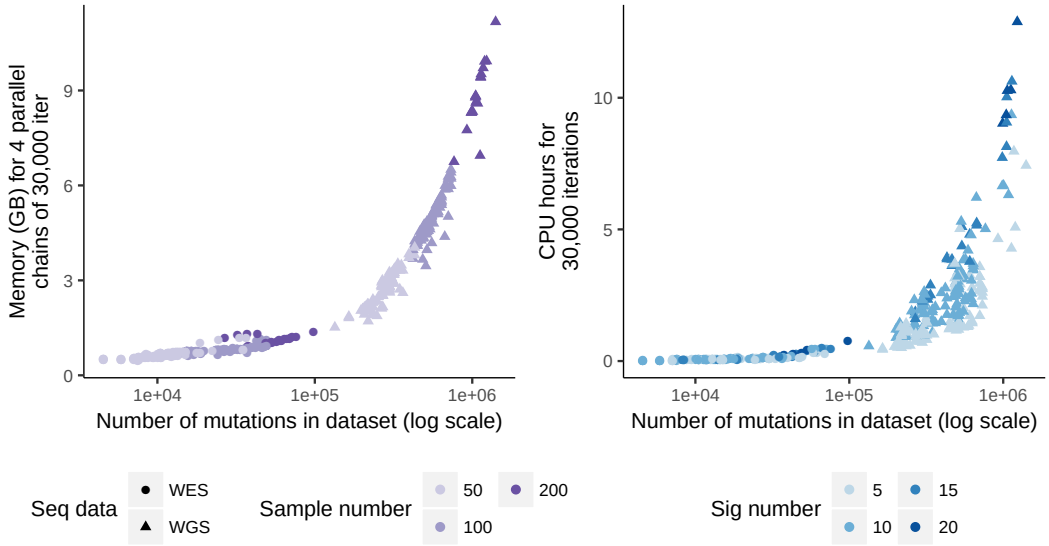


Figure 4.11: Computational resources for HDP posterior inference (MCMC approach) on simulated datasets.

Computational resources The computational time and memory required for HDP inference with MCMC (Figure 4.11) scales with the number of mutations that must be iterated over and tracked through successive cluster allocations. The CPU time also increases with the complexity of the data (number of underlying signatures), as the volume of calculations at each step relates to the number of clusters. The easiest way to reduce computational cost is to sub-sample the mutation set in hypermutators, thereby reducing the number of data items. The memory requirements are also reduced by collecting fewer posterior samples, and time in human hours (rather than CPU hours) is reduced by running more chains in parallel, particularly after the burn-in period.

4.4 Application to SNVs in original signature discovery dataset

In the first major effort to describe mutational signatures in a large pan-cancer somatic SNV dataset, Alexandrov et al. (2013b) applied NMF to almost 5 million mutations from over 7000 samples (mostly exomes) representing 30 different cancer types. By focusing on 96 SNV classes in a trinucleotide context, the original report presented 27 consensus signatures, including: 22 which validated (including two versions of the `CpG` deamination 'signature 1'); 3 confirmed artefacts; and 2 unable to be validated. The COSMIC database (Forbes et al.,

2015) has since released an updated set of 30 signatures (numbers 22–30 not reported in the 2013 paper) extracted with the same methods from an updated set of more than 10,000 samples (Alexandrov et al., 2015b). In this section, I return to the original signature discovery cohort of ∼7000 samples (summarised in Table E.3), and compare HDP results in a practical real-world dataset.

4.4.1 Model design, combining exomes and genomes

One obstacle to combining exome and genome data in signatures analysis is the difference in background trinucleotide frequency (Figure D.16). To take the extreme examples, `ATA` has a trinucleotide frequency of 4.1% in the whole genome but 2.7% in the exome (ratio 1.5)[l], while `GCG` has frequency 0.47% in the genome but 1.3% in the exome (ratio 0.36)[l]. The upshot of this discrepancy is that the same underlying mutational process will present with different mutation class proportions in exome or genome data. In their original signatures analysis paper, Alexandrov et al. (2013b) ran NMF in separate sample groups divided by cancer type and sequencing type (exome or genome), then matched the signatures post hoc, adjusting for exome biases on the signature distributions at this stage[m].

In contrast, I choose to pool the exome and genome data, and fit the HDP signatures model to all samples simultaneously, grouping cancer types by parent nodes as illustrated in Figure 4.1. This approach empowers the clustering method to share information across cancer type boundaries, while upholding the prior expectation of significant differences between groups. However, mutation class tallies in the exome samples require adjustment to reflect mutational signatures on a comparable background.

For an exome sample j with observed 96-length mutation class count vector $\boldsymbol{\mu}_j$ and total SNV count of $\|\boldsymbol{\mu}_j\|_1 = m_j$, the adjusted mutation class counts are

$$\boldsymbol{\mu}'_j = \left\lfloor \left\{ \left(\boldsymbol{\gamma} \odot \boldsymbol{\mu}_j m_j^{-1}\right) \Big/ \left\|\boldsymbol{\gamma} \odot \boldsymbol{\mu}_j m_j^{-1}\right\|_1 \right\} \times m_j \right\rceil ,$$

[l]For trinucleotide frequency in the whole genome, I only include the callable genome regions defined in Section 3.1.1. For the exome, I include all protein-coding exons plus 100 bp flanks as variant calls are often made in flanks and off-target regions.

[m]I follow Alexandrov et al. (2013b) in reporting mutation class signature probabilities as their expected relative frequency in a (human) genome-wide landscape, without normalising by background trinucleotide frequency. That is, the reported probabilities inherently account for how rare (e.g. `ACG` or `TCG`) or common (e.g. `TTT`) the context is. If the genome composition was adjusted for (maybe useful to generate species-agnostic signatures), the signature probabilities would increase for the rare contexts, and decrease for the common contexts.

where $\odot$ denotes element-wise multiplication, $\boldsymbol{\gamma}$ is the genome-to-exome ratio of the trinucleotide context for each mutation class, and $\lfloor \ldots \rceil$ is shorthand for integer rounding using a modified procedure guaranteed to preserve m_j total SNV count for sample j.

Using these adjusted mutation tallies for any exome sample, and down-sampling hypermutator samples to a maximum of 20,000 SNVs each[n], I allocated each sample to a leaf node using the HDP design for multiple cancer type groups as in Figure 4.1.

In the first instance, I ran four independent burn-in chains for 15,000 iterations, each separately initialised with 30 random clusters. Picking up from the end of each initial chain, I started another four independent MCMC chains for a further 10,000 burn-in iterations and then collected 50 posterior samples at intervals of 300 iterations (800 total samples from 16 separate chains).

4.4.2 Sampling chain diagnostics

Theoretically, an infinitely long MCMC chain would sample all possible cluster allocations in proportion to their likelihood. In practice, we aim to have a finite posterior sample set that approximates the true random sampling space without strong biases imparted by the initialisation state or by slow mixing between successive iterations. The diagnostic plots in Figure 4.12 show no strong trends in the likelihood or number of raw clusters across the MCMC chains which might indicate poor sampling. In future method development, it would be beneficial to include more formal convergence diagnostics.

4.4.3 HDP signature and exposure estimates

Using the method outlined in Section 4.2.2[o], the HDP model returned 54 consensus mutational signatures and assigned 19.8% of mutations (on average) to the zero component for noise and uncertainty.

For each HDP-estimated mutational signature ('HSig'; all presented in Figure D.17), I matched the mean mutation class distribution with its closest

[n] 30 hypermutator samples downsampled.

[o] I set the similarity threshold for signature merging to 0.92 as the COSMIC set includes one pair with this level of similarity. Also, I required every reported signature to have significant exposure (95% credibility interval above zero) in at least two samples.

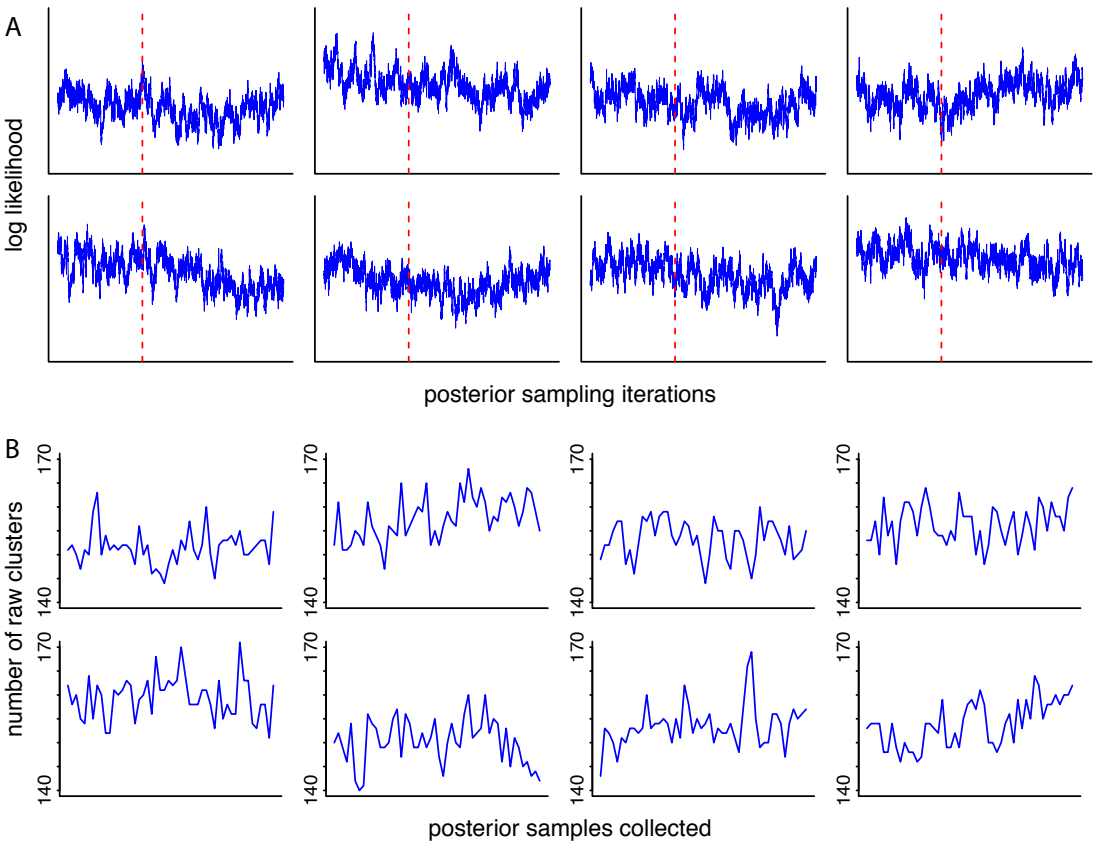


Figure 4.12: Diagnostic plots to assess HDP sampling chains for the signature discovery dataset (8 of 16 chains shown). (a) Log-likelihood of HDP state at each iteration, showing the end of the burn-in period up to the red dashed line, with posterior samples collected at regular intervals thereafter. (b) Number of raw data clusters at each collected posterior sample (50 per chain).

match in the current COSMIC set ('CSig'; including the artefact and un-validated signatures from Alexandrov et al. (2013b)), down to 0.9 cosine similarity.

Of the artefactual and un-validated signatures described by Alexandrov et al. (2013b), HDP only recovered the three artefacts R1–R3. Of the 21 validated signatures (CSig1–CSig21), HDP recovered all but four. The four missing signatures were CSig3, CSig5, CSig6, and CSig19. COSMIC annotates CSig3 as a HR-deficiency signature, and CSig5 as a 'clock-like' process associated with age (Alexandrov et al., 2015b). CSig3 and CSig5 have relatively uniform mutation class profiles which HDP may struggle to differentiate in exome data, presumably apportioning many of these mutations to the zero component with uncertain allocation. Part of the HR-deficiency signature is probably captured by HSig7, with a 0.88 similarity to CSig3 and frequent contribution to the breast cancer cohort. CSig6 is annotated as defective DNA mismatch repair, often co-occurring with the other mismatch repair signatures CSig15 and CSig20. Of the HDP-estimated signatures, HSig30 matches CSig15 extremely closely (0.98 similarity) while HSig10 matches CSig20 quite roughly (0.91 similarity). Further investigation reveals that HSig10 is a much closer match to a blend of CSig20 and CSig6[P], so it seems that HDP does not distinguish between these aspects of defective mismatch repair. The missing CSig19 is solely identified in pilocytic astrocytoma (Alexandrov et al., 2013b), and is not apparent in the HDP output. Considering the signature-tissue overview presented in Figure 4.13, the mutations originally attributed to CSig19 are presumably subsumed by the zero component.

Of the nine validated signatures subsequently added to the COSMIC database after analysis of more data (Alexandrov et al., 2015b), HDP recovered CSig22 (aristolochic acid) and CSig28 (mostly `T>G` in `NTT`) without requiring the extra samples. This suggests the HDP method may have greater sensitivity for detecting some genuine signatures.

The NMF analysis of this dataset recovered two versions of the common CSig1 `CpG` deamination signature (Alexandrov et al., 2013b). Similarly, HDP outputs three signatures resembling CSig1: HSig17 as a very pure distribution of `C>T` in `NCG` (even stronger peaks than the current COSMIC estimate); and HSig13 and HSig9 as relatively 'muddled' versions (Figure D.17) with particular prominence in esophageal and breast cancers respectively. It seems likely that these latter estimates mix different underlying processes. Two other COSMIC signatures are also represented multiple times in the HDP output. The CSig7 UV radiation

[P] 0.97 cosine similarity between HSig10 and a 60:40 mixture of CSig20 and CSig6.

	ALL:140	AML:154	Bladder:136	Breast:963	Cervix:38	CLL:131	Colorectum:559	Esophageal:146	Glioblastoma:98	Glioma_Low_Grade:217	Head_and_Neck:380	Kidney_Chromophobe:65	Kidney_Clear_Cell:325	Kidney_Papillary:100	Liver:88	Lung_Adeno:660	Lung_Small_Cell:70	Lung_Squamous:176	Lymphoma_B-cell:48	Medulloblastoma:100	Melanoma:396	Myeloma:69	Neuroblastoma:210	Ovary:471	Pancreas:113	Pilocytic_Astrocytoma:101	Prostate:330	Stomach:212	Thyroid:304	Uterus:241
HSig0 matches (0.91) CSig5	11	104	101	295	11	50	168	132	41	202	309	22	322	100	88	514	46	146	38	38	110	65	127	451	19	69	324	112	251	143
HSig1 matches (0.97) CSig4																347	22	14												
HSig2 matches (0.98) CSig7																		1			372									
HSig3 matches (0.96) CSig12															88	1			1		1					4				1
HSig4 matches (0.90) CSig13	10		127	86	36	8	2	8			207			4	1	195	7	88				1		4	2		2	7	6	19
HSig5 has no match				9											5				5	46					14					
HSig6 has no match															88															
HSig7 has no match				41											9	2									5					
HSig8 matches (0.97) CSig8				26											18	1			2	39					1					
HSig9 matches (0.90) CSig1				50												1														
HSig10 matches (0.91) CSig20				5			83	2		1	2	1	1			1		1					1				3	39		70
HSig11 has no match																3														
HSig12 has no match																18														
HSig13 matches (0.90) CSig1		1		10			39	104				3			1										2		2	14		1
HSig14 matches (1.00) CSig2			27	58	6		2				13					21		1	1			3			1			4	4	3
HSig15 matches (0.94) CSig10				1			13												1											14
HSig16 matches (0.91) CSig16															68															
HSig17 matches (0.97) CSig1		1		8			35	1		1		1			1	3			1	90					19	68	2	32		16
HSig18 matches (0.92) CSig7																					82									
HSig19 has no match											3				54															
HSig20 matches (0.98) CSig10							8																							10
HSig21 matches (0.98) CSigR3							253			2																				
HSig22 matches (0.90) CSig14							4			1																				8
HSig23 matches (0.96) CSig9						16													15		1									
HSig24 matches (0.96) CSig18				1																			5							
HSig25 matches (0.99) CSig17							7	120								14			10									71		
HSig26 has no match																				71										
HSig27 matches (0.97) CSig21				1			13	1							1						1							17		7

Figure 4.13: Number of samples with significant exposure to HDP-extracted signatures (95% credibility interval above zero). HDP signatures ('HSig') are labelled with their closest match in the COSMIC signature library ('CSig'), with known artefacts prefixed 'R'. 'HSig0' denotes the zero component for noise and uncertainty. The number of samples considered is indicated in the column label for each cancer type. Figure continues on the next page.

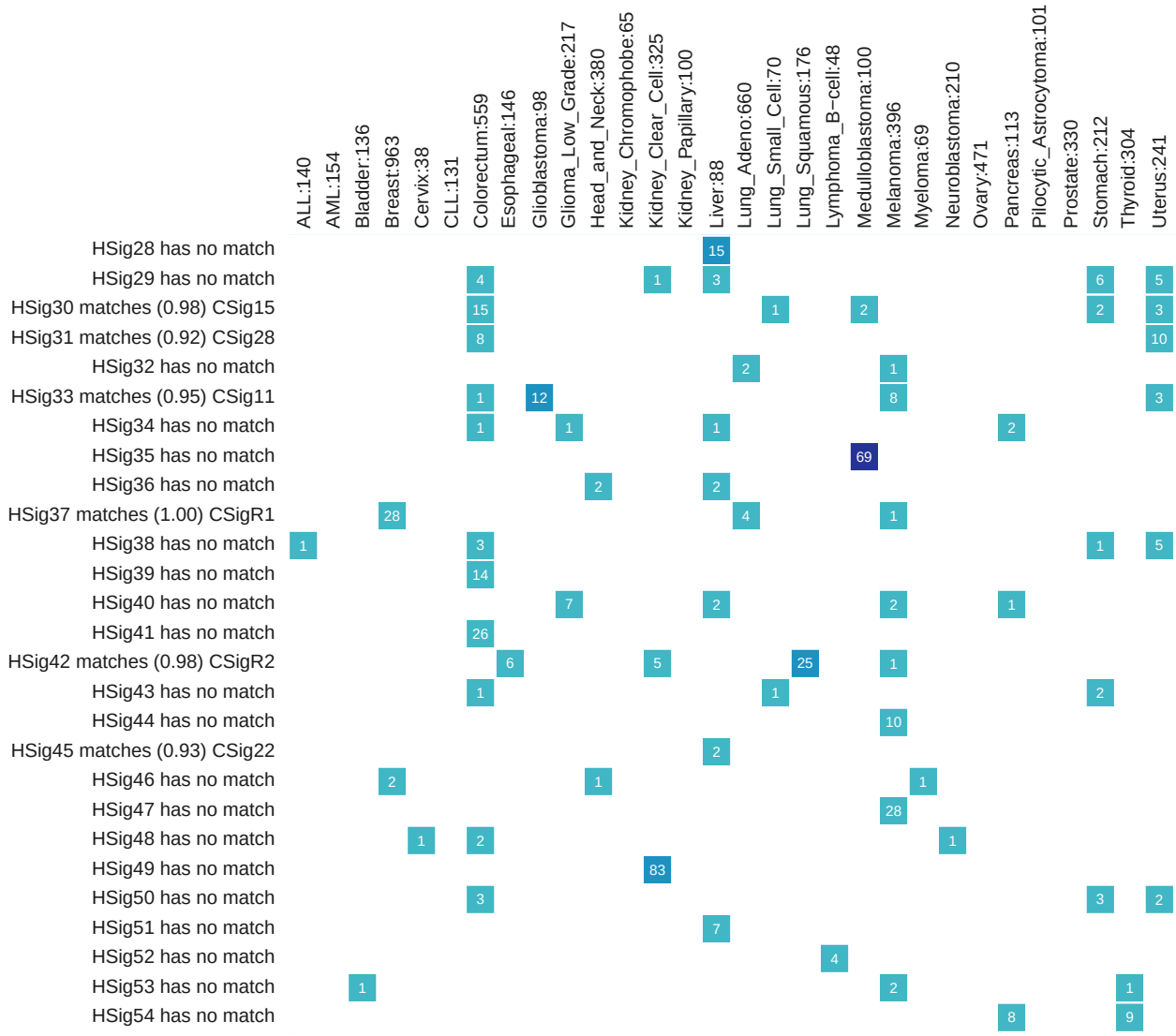


Figure 4.13: Number of samples with significant exposure to HDP-extracted signatures (95% credibility interval above zero)—continued from previous page.

signature is matched by both HSig2 and HSig18, both with strong exposure in the melanoma cohort (Figure 4.13). It remains unclear whether this is a genuine biological difference in the UV profile affecting some samples, or a false positive signature split caused by over-fitting. Finally, the CSig10 signature of mutant POLE activity is matched by both HSig15 and HSig20, differentiated by the relative probabilities of `C>A` in `TCT` and `C>T` in `TCG` (Figure D.17). This signature split may plausibly reflect biological variation in the POLE effect of different mutant protein residues (Rayner et al., 2016; Campbell et al., 2017a).

Of the 28 HDP signature estimates with no match in the COSMIC database, some are close to uniform with similar patterning to other known signatures (centre of Figure 4.14), whereas others have unique, distinctive peaks (edges of Figure 4.14). Some novel signatures with particularly clear patterns (see all in Figure D.17) include:

- HSig19 in 54 liver cancers (`T>C` in `ATN`, possibly a cleaner extraction of CSig16/HSig16);
- HSig28 in 15 liver cancers (`C>T` in `TCT`);
- HSig29 in 19 various samples including stomach, uterus (`T>C` in `NTG`);
- HSig40 in 12 various samples, including 7 gliomas (`C>G` in `WCW`);
- HSig44 in 10 melanomas (`T>C` in `GTT`);
- HSig47 in 28 melanomas (`T>N` in `TTT`);
- HSig49 in 83 kidney clear cell cancers (`T>C` in `NTY`);
- HSig51 in 7 liver cancers (`T>G` in `GTN`); and
- HSig53 in rare bladder, melanoma and thyroid samples (`C>G` in `CCN`).

The tendency for many novel signature estimates to group similar mutation classes supports their biological validity, as the model regards all 96 SNV categories as equally independent. Furthermore, the novel melanoma signatures in a `TpT` context are consistent with the known modality of UV radiation causing thymine dimer lesions. However, validating these signature estimates as sequencing artefacts, genuine mutational processes, or over-fitting to random data correlations, is beyond the scope of this work.

In contrast to the simulated data (Section 4.3) for which HDP typically consigned one or two percent to the zero component, this real-world example grouped almost 20% of total mutations in the zero component for noise and uncertainty.

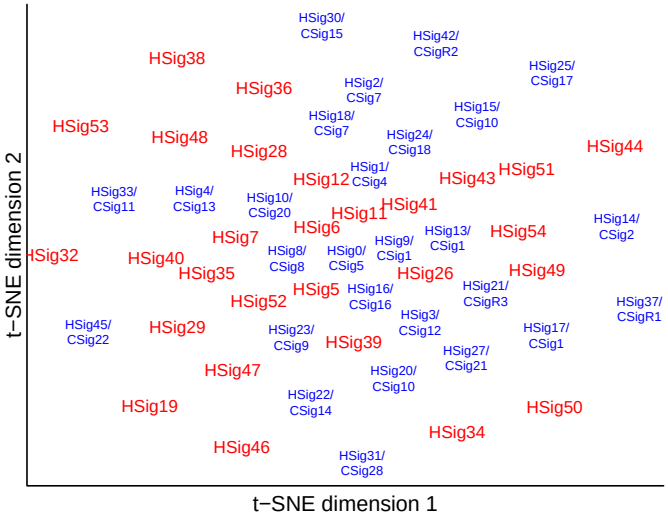


Figure 4.14: Map of HDP-extracted signatures (blue for COSMIC match) using t-distributed stochastic neighbour embedding (two dimensions mapping similarity between signature distributions).

With thousands of small exome samples, my consensus signature extraction method (Section 4.2.2) failed to reproduce the close-to-uniform CSig3 and CSig5. Further method development may improve the sensitivity to uniform patterns, possibly by upgrading the raw cluster matching step (across posterior samples) to consider sample exposure similarity in addition to signature similarity. It is also possible that my approach for normalising exome data to a genome-wide background (Section 4.4.1) introduced unintended biases that confounded the clustering procedure. Overall, the zero component is a similar match to the CSig5 signature found in all cancer types (Alexandrov et al., 2013b). To extract CSig5 as its own confident signature, the HDP method may require cleaner genome data and/or a more robust post-processing procedure.

In addition to the mutational signatures, HDP also estimates the exposure weights within each sample (leaf node) and group (parent node). By taking 95% highest posterior density credibility intervals, HDP formalises the significance of each signature's activity. As shown in Figure 4.15, this leads to my novel approach of plotting sample exposure with a fraction left unexplained. By plotting the mean exposure estimate for significant findings only, the blank proportion of sample exposure represents the proportion of mutations with uncertain cluster allocation (to the same, or different, signatures). This strategy is important for communicating uncertainty, and emphasizes our reduced confidence in the signature exposures of low-burden samples. To highlight some examples in Figure 4.15: bladder samples are dominated by the APOBEC signatures (HSig4/CSig13, HSig14/CSig2); lung cancers by tobacco (HSig1/CSig4); melanoma by UV radiation (HSig2/CSig7, HSig18/CSig7); and high-burden stomach cancers by defective mismatch repair (HSig10/CSig20).

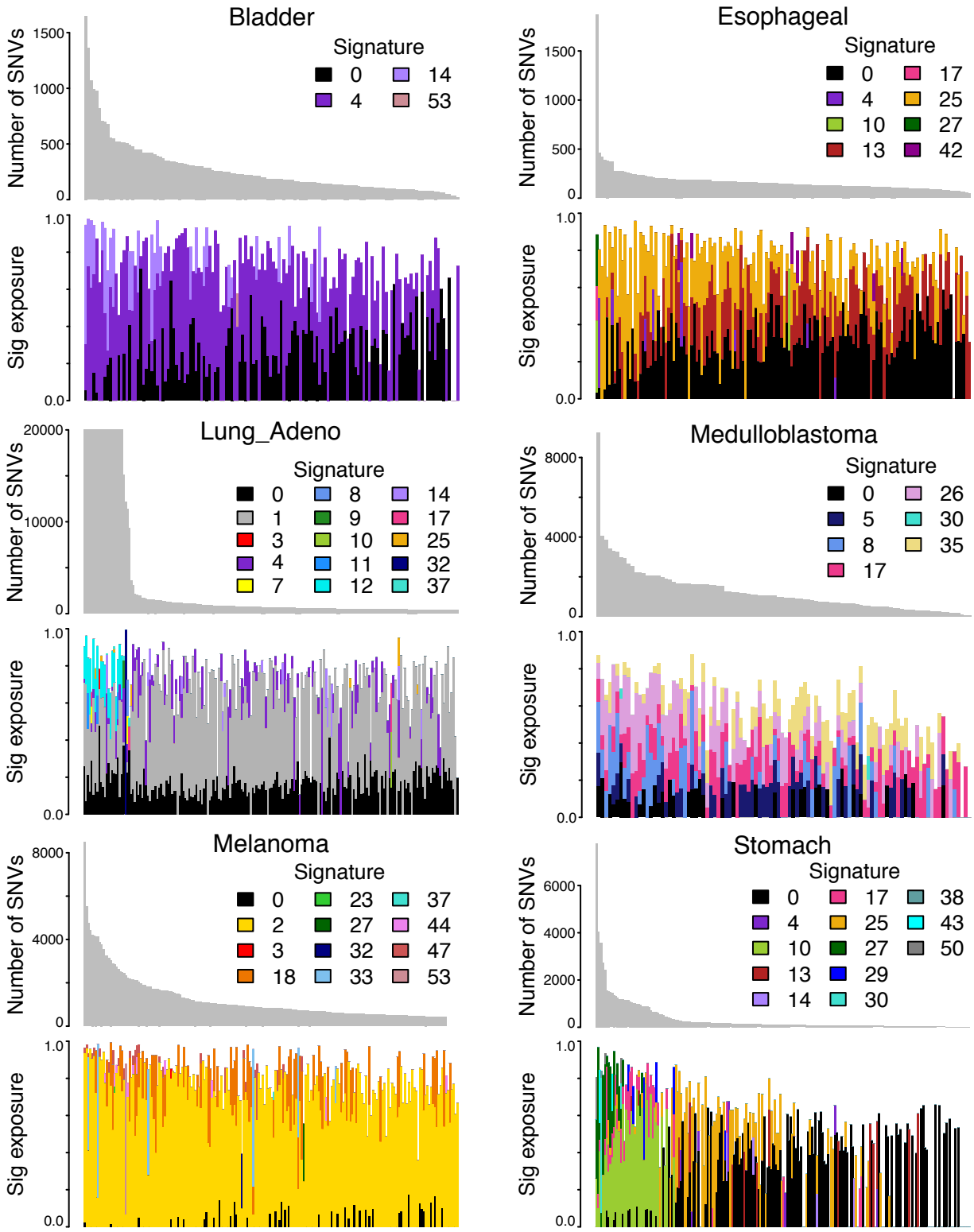


Figure 4.15: Average estimated sample exposures to HDP-extracted signatures for six cancer types with samples sorted by observed mutational burden, capped at 20,000 SNV. Large cohorts are subset to a maximum of 200 samples for presentation. For each sample, mean estimated exposures are only plotted for significant signatures (95% credibility interval above zero), leaving a blank proportion with uncertain signature allocation.

4.5 Simultaneous signature matching and discovery

One of the key theoretical advantages of the HDP method for mutational signatures analysis is the ability (outlined in Section 4.2.3) to match a new dataset to an existing library of known signatures while simultaneously empowering any novel mutational signatures in the dataset to emerge as separate clusters.

Data and methods

To briefly illustrate this approach on real data from the PCAWG cohort, I selected somatic SNV calls from the pancreatic endocrine cancer group (81 samples; 252,930 SNVs) and prostate adenocarcinoma group (198 samples; 635,688 SNVs). Following the HDP design illustrated in Figure 4.2, the 30 known mutational signatures in the current COSMIC database[q] were each represented by 500 pseudo-count mutations in a frozen node. Analysing each cohort separately, I randomly initialised the real mutations into 35 clusters—30 linked with a prior signature, and five others solely comprised of observed mutations from the new dataset. After running four burn-in chains for 10,000 iterations, two chains bifurcated from the end of each burn-in and ran a further 2000 burn-in iterations before collecting 125 posterior samples separated by 100 iterations (1000 total posterior samples from eight independent MCMC chains).

Results

For the pancreatic endocrine cohort, HDP identifies four additional signatures (shown in Figure 4.16) while simultaneously matching to the set of known signatures. The newly extracted signature N1 is characterised by a consistent distribution of `C>A` mutations, with high-confidence peaks in `TCT` and `TCA` contexts. Scarpa et al. (2017) recently attributed this pancreatic neuroendocrine signature to *MUTYH* loss and consequent deficiency in base excision repair. Seven samples have significant exposure to this *MUTYH* signature, including three with high overall burden (> 7500 SNV) and $> 92\%$ of mutations attributed to signature N1 (Figure 4.17). Signatures N2 and N3 have greater uncertainty about their mutation class distribution, and have significant exposure in three and six samples respectively, with estimated exposure proportions as high as

[q] `http://cancer.sanger.ac.uk/cosmic/signatures` available as of December 2017

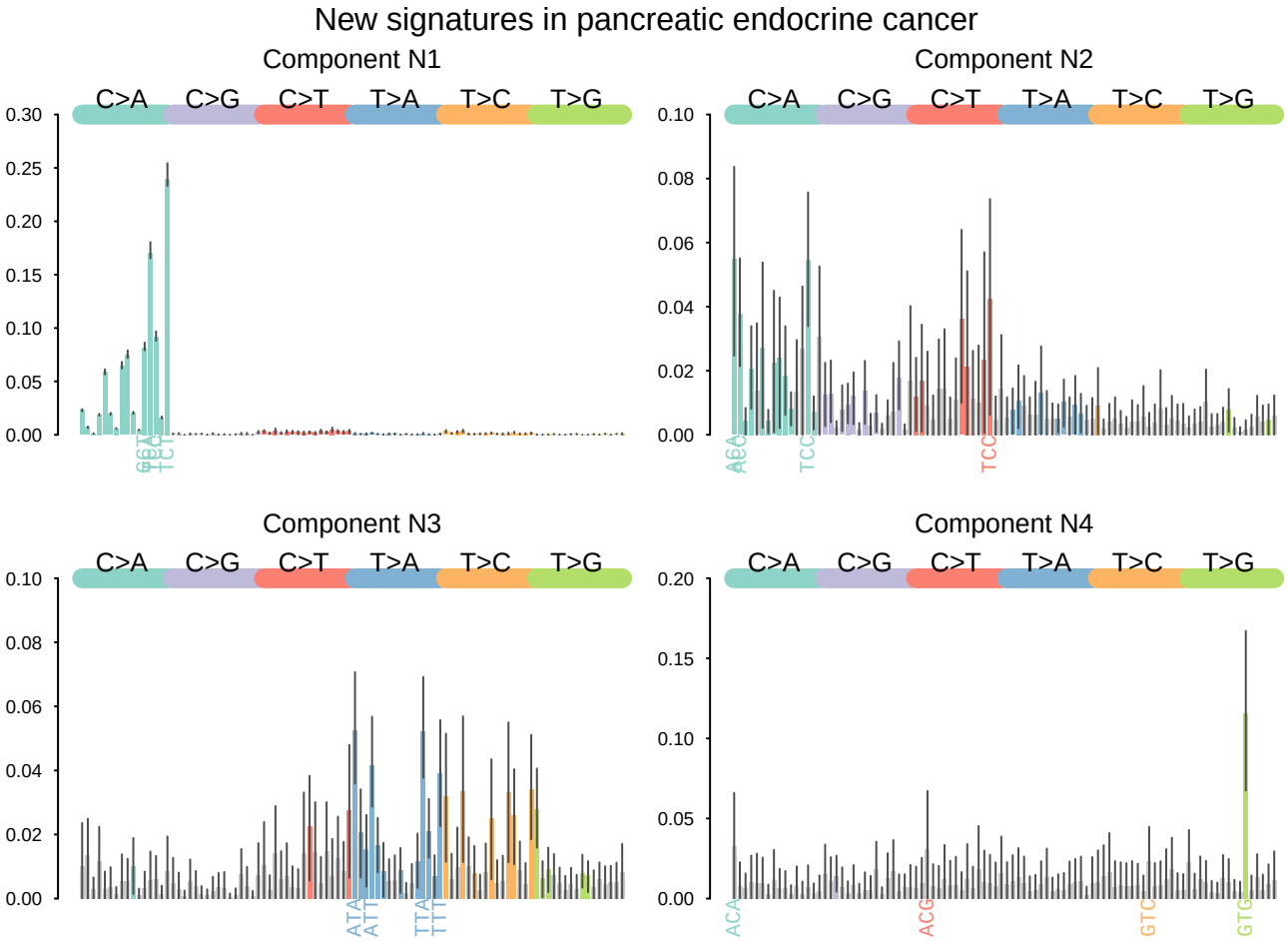


Figure 4.16: Newly discovered mutational signatures in pancreatic endocrine cancer WGS cohort (mean and 95% credibility interval from HDP posterior samples, with non-significant mutation classes in grey and four major peaks labelled with trinucleotide context). Component N4 is similar to the known artefact signature R1.

24% and 27%. Signature N4 has one dominant peak of `T>G` in `GTG` and probably corresponds to the known artefact signature 'R1' (Alexandrov et al. (2013b), artefact signatures not included as priors). Across the cohort, samples also have significant exposure to many prior COSMIC signatures, including age-related signatures 1 and 5 (Alexandrov et al., 2015b), APOBEC signatures 2 and 13, and signature 8 (low `C>A` peaks, with `CC>AA` double nucleotide substitutions). One unusual sample has 66% of 1463 SNVs attributed to signature 12 (peaks in `T>C`), previously described in liver cancer only (Alexandrov et al., 2013b).

For the prostate cohort, HDP identifies six novel signatures shown in Figure 4.18. Signature N1 is quite common in the cohort (Figure 4.19), with significant exposure in 30 samples including several with over 3000 SNVs attributed to this novel signature. In contrast, signature N6 has significant exposure in

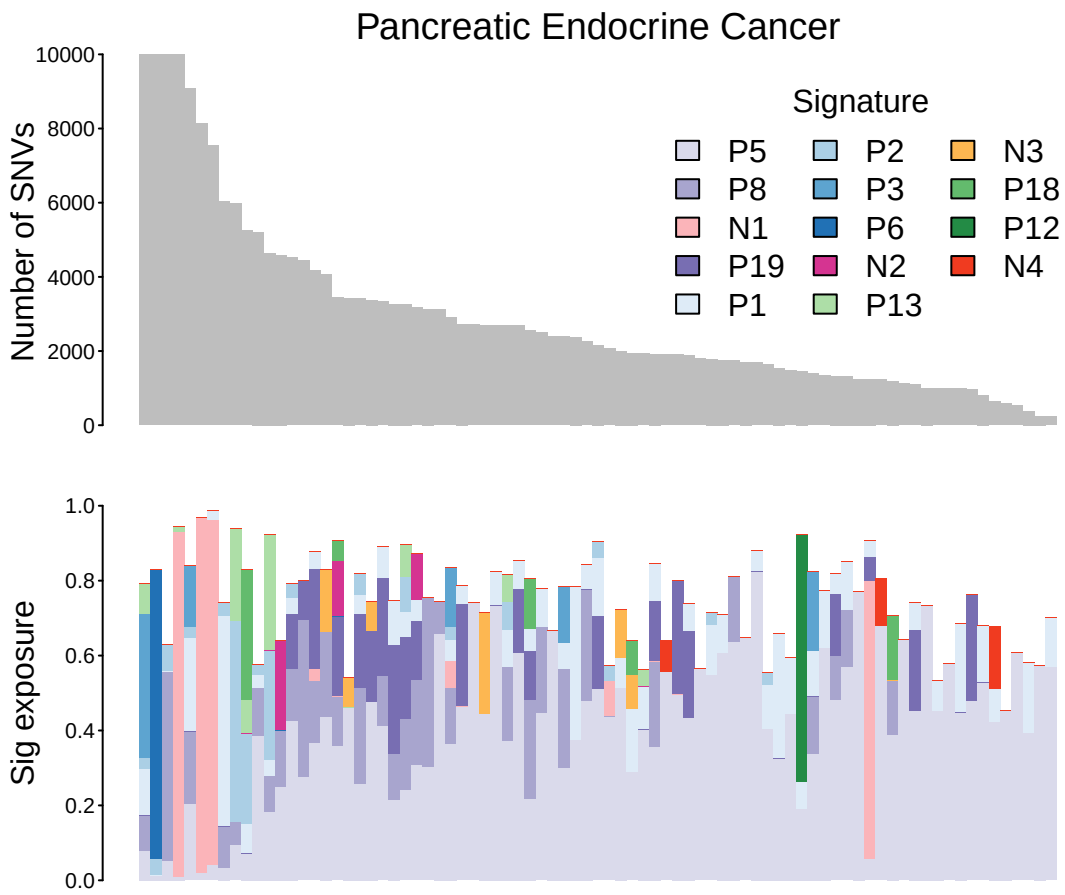


Figure 4.17: Pancreatic endocrine cancer sample exposures (average from HDP posterior samples) to a library of known signatures (labelled 'P' for prior) and newly discovered signatures (labelled 'N' for new) with samples sorted by observed mutational burden, capped at 10,000 SNVs. For each sample, mean estimated exposures are only plotted for significant signatures (95% credibility interval above zero), leaving a blank proportion with uncertain signature allocation.

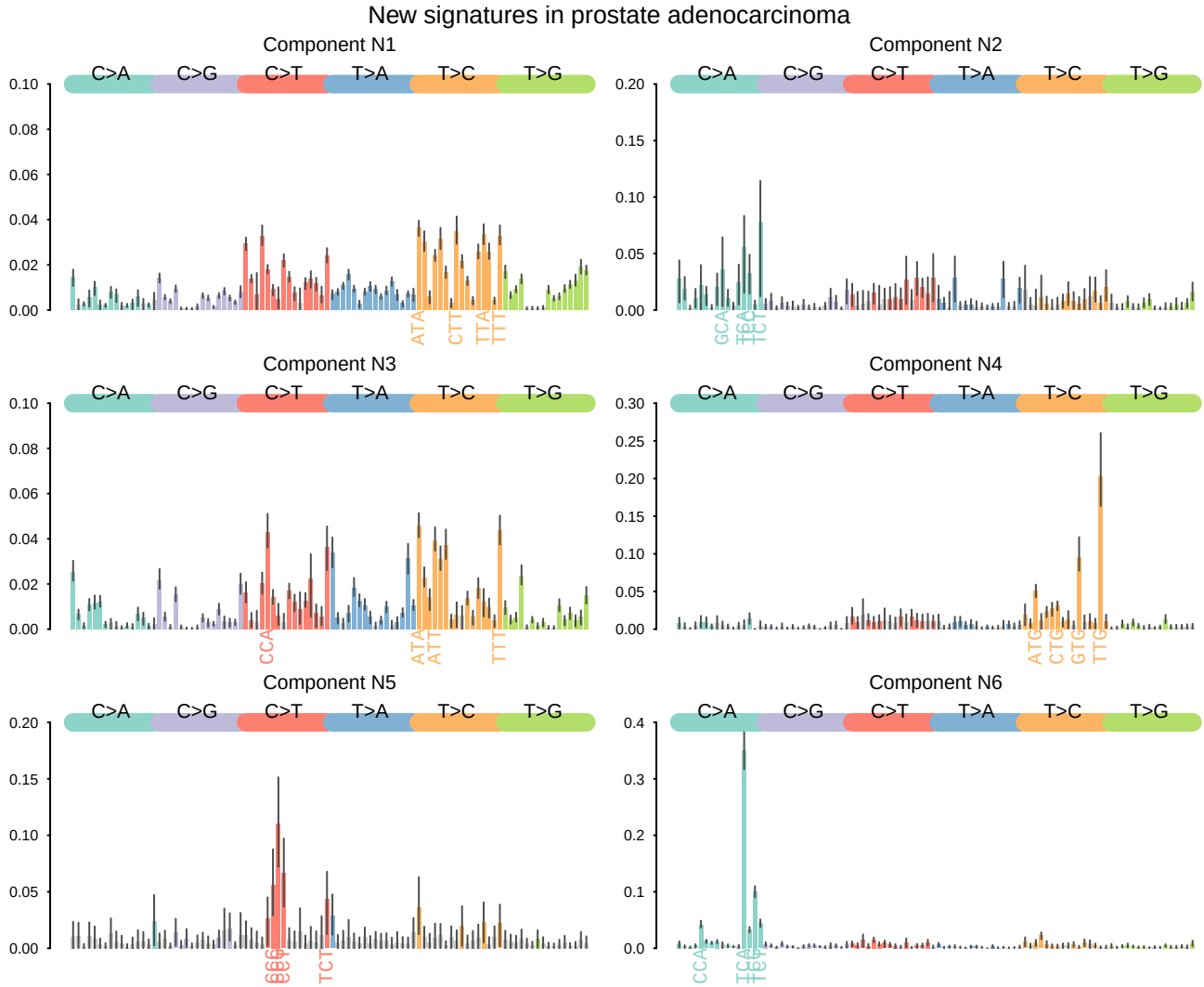


Figure 4.18: Newly discovered mutational signatures in prostate adenocarcinoma WGS cohort (mean and 95% credibility interval from HDP posterior samples, with non-significant mutation classes in grey and four major peaks labelled with trinucleotide context).

just one sample, contributing an estimated 70% of its 2775 SNV calls, with a huge peak of `C>A` mutations in a `TCA` context. Prostate samples also have significant exposure to some prior COSMIC signatures, including 1, 5, and 8. Interestingly, one sample in the prostate cancer cohort has almost 1000 SNVs attributed to COSMIC signature 9, thought to be the mark of polymerase η activity and previously identified in CLL and B-cell lymphomas only (cells with AID hypermutation) (Alexandrov et al., 2013b). This HDP finding implicates rare polymerase η activity under other conditions in prostate tissue.

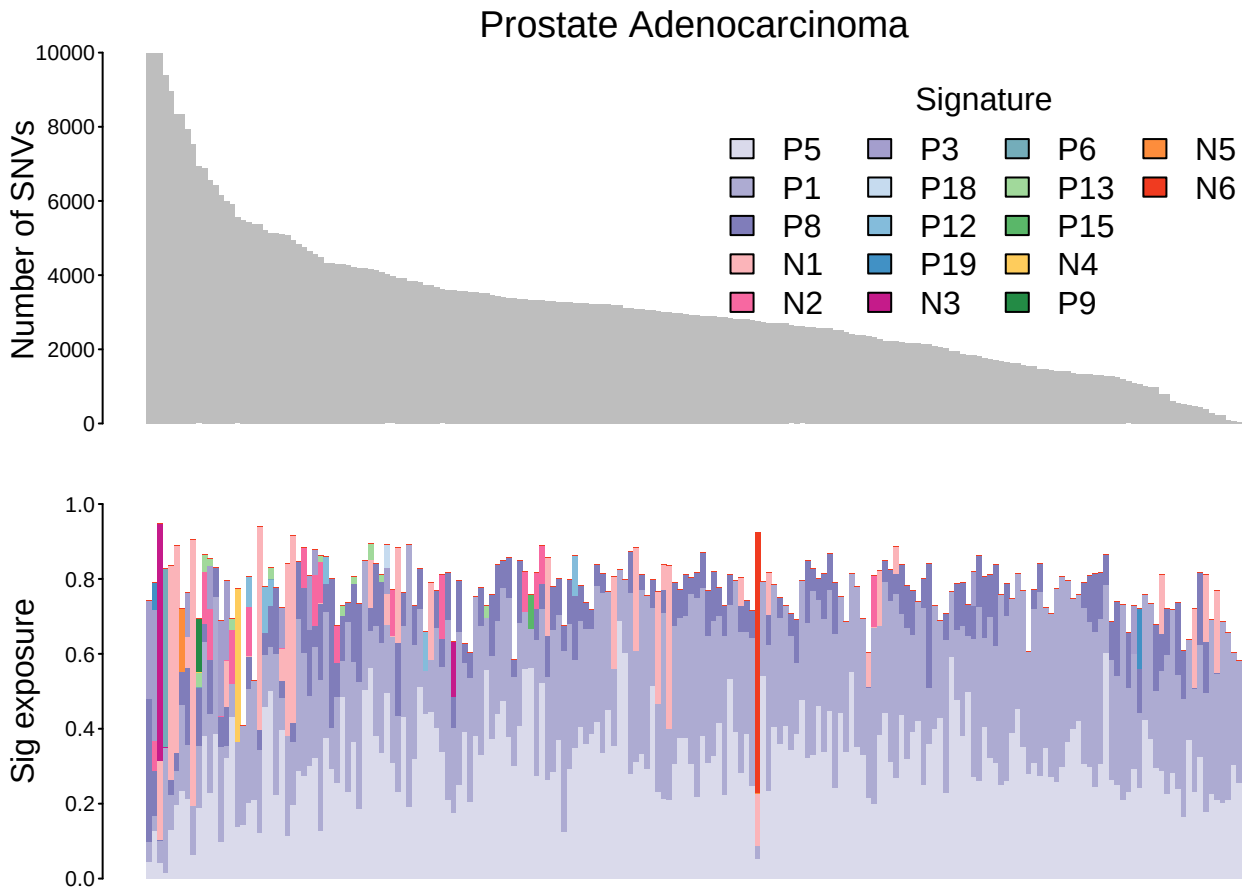


Figure 4.19: Prostate cancer sample exposures (average from HDP posterior samples) to a library of known signatures (labelled 'P' for prior) and newly discovered signatures (labelled 'N' for new) with samples sorted by observed mutational burden, capped at 10,000 SNVs. For each sample, mean estimated exposures are only plotted for significant signatures (95% credibility interval above zero), leaving a blank proportion with uncertain signature allocation.

Discussion

Any newly discovered pattern of co-occurring mutation classes may: reflect a genuine mutational process previously missed by signature analysis; be a variant form of a previously described signature (genuine biological variant, or different form due to calling bias); or be a specific profile of artefact calls from contamination, sequencing errors, etc. Validating the new signatures described in pancreas and prostate as artefact or genuine is beyond the scope of this thesis.

While these results demonstrate the efficacy of this HDP approach for discovering novel mutational signatures and quantifying the uncertainty in their distribution with credibility intervals, interpretation of the prior signature matching has possible pitfalls.

First, when observed mutations from the new dataset cluster with fixed pseudo-counts corresponding to a prior signature, the distribution over mutation classes in that cluster can sometimes deviate substantially from the prior signature as primed by the pseudo-counts. The signature extraction method attempts to resolve this problem when reporting results, and will de-couple a cluster from the prior identity of its pseudo-counts if the overall pattern has diverged from the original intended signature. With the current implementation, the final estimated signature will still be labelled with its closest match in the prior set down to a threshold of 0.85 cosine similarity. As a result, reported exposure to a prior signature may sometimes indicate a rough match only. For example, the cosine similarity between the reference version of COSMIC signature 19 and the version reported in the pancreatic endocrine cancers is 0.92, and it remains unclear whether or not this represents the same underlying process.

Second, it may be the case that more uniform signatures (low variation over mutation class proportions) are particularly difficult to distinguish when conditioning on prior knowledge. Whereas the roughly uniform COSMIC signature 5 has previously been reported in all cancer types, another roughly uniform signature 8 has only been reported in breast and medulloblastoma (Alexandrov et al., 2013b) and yet was estimated to have a significant presence in most of the pancreatic and prostate samples analysed here. It seems plausible that this signature 8 exposure stems from mis-clustering of genuine signature 5 mutations, primed by pseudo-counts with similar spread over all mutation classes. The COSMIC database reports that signature 8 has a weak transcription strand bias for `C>A` mutations and a tendency for double nucleotide `CC>AA`

substitutions, so it would be interesting to check in future research whether or not the purported signature 8 exposures in pancreas and prostate have similar distinguishing properties.

Finally, conditioning on a large number of prior signatures may increase the likelihood of a small subset of mutations consistently clustering with a known prior by chance, introducing small false positive signature exposures not adequately sampled away by the finite MCMC sample collection. For example, previous analysis with NMF reported only three mutational signatures with significant exposure in 520 prostate cancer exomes (COSMIC signatures 1, 5 and 6, Alexandrov et al. (2015b)), whereas my HDP analysis of 198 prostate genomes found significant exposure to eleven known COSMIC signatures and a further six novel signatures. This discrepancy could partially result from greater detection power in genomes rather than exomes; greater sensitivity of the HDP approach (particularly when conditioning on prior signatures); and possible false positive matching to prior signatures.

To assess how the inclusion of prior signature information impacts results, it would be informative to compare against *de novo* HDP signature extraction on these same cohorts, and see whether a pattern like COSMIC signature 8 emerges separately to COSMIC signature 5, and whether the low frequency signature exposures are still reported. The performance of HDP matching to prior signatures could also be investigated with simulation studies under a range of conditions, with particular interest in close-to-uniform signatures, and possible false-positive clustering with fixed pseudo-counts.

In practice, including *all* previously reported signatures as equally weighted priors may be naive and possibly confounding, particularly as the number of known mutational signatures will continue to grow. I conjecture that a better approach would be to weight known priors by their reported prevalence in related cancer types, even to the point of excluding priors only described in completely unrelated cancer types.

If future studies come to model mutational signatures in more detail than a simple categorical distribution over 96 SNV classes—for example, including strand bias, double nucleotide substitutions, replication timing bias—then methods to match new data to previously reported signatures will have more diverse evidence to draw from, likely improving results.

4.6 Signatures of genome rearrangement

In contrast to the detailed analyses of somatic SNV signatures (Alexandrov et al., 2013b; Helleday et al., 2014; Alexandrov et al., 2015b), few publications have attempted a similar decomposition of somatic rearrangement signatures. In 560 breast cancer genomes, Nik-Zainal et al. (2016) applied NMF to a set of BPJ calls classified by their orientation, size, and presence in clustered or isolated SVs. This yielded six rearrangement signatures, broadly defined by: large or small tandem duplications (two separate signatures); small deletions; unbalanced translocations with large deletions and inversions; and intra- or inter-chromosomal complex rearrangements (two separate signatures). With essentially the same SV classification and NMF signature pipeline, Hillman et al. (2018) extracted five rearrangement signatures from 80 ovarian cancers, finding similar results with basic separation by SV class and size for isolated BPJ.

In this section, I return to the SV dataset of approximately 2500 PCAWG samples (introduced in Chapter 2) and leverage our detailed BPJ classifications (Section 2.1.3) to calculate signatures of co-occurring rearrangement patterns.

4.6.1 Generating the SV tally matrix

Using the BPJ classification in Table 2.2 as a starting point, I defined 76 SV categories to use an input alteration classes for a HDP signature analysis.

For deletion and tandem duplication, I first separated the fragile site events with both breakpoints inside one of the 21 FS regions defined in Table E.2 (Section 3.4). Then, I classified the remaining deletions and tandem duplications by both size (breaks at 50 kb, 500 kb, and 5 Mb) and replication timing (at the event mid-point; early > 60, late < 30, or in-between, using the definition in Section 3.1.2). Events larger than 5 Mb were not sub-categorised by replication timing.

For classified SV like translocation (unbalanced and reciprocal), 2-jumps (local and distant), chromoplexy (cycles and chains), and templated insertion (cycles, chains and bridges), I tallied the counts per-event rather than per-BPJ. Local 2-jumps and reciprocal inversions (two BPJ per event count) were additionally separated by size categories split at 100 kb (measuring the total event span). Templated insertions were divided by the size of the insert fragment (split at 5 kb), taking the median insert size for multi-insert events were necessary.

Finally, of the ∼150,000 BPJ in complex unexplained clusters, I included categories for the subset of individual local footprints with a recurrent BPJ pattern present at least one hundred times in the cohort (tallied once per-footprint, not per-BPJ). In the category labels shown in Figure 4.20, the complex footprint patterns are annotated with an alphabetical segment notation, using + for the 3' end, − for the 5' end, carets for BPJ joins, and a forward slash for adjacent breakpoints in separate BPJ. The many complex unexplained BPJ in rarer, more convoluted local footprints were excluded from the signatures analysis.

After removing samples with less than three counted SV events, the final matrix tallied 147,508 SV events across 2050 samples. Of the 76 SV categories, the most common were: a single translocation breakpoint within a complex cluster (12,753) and deletions smaller than 50 kb with mid-range replication timing (11,289). The least common categories were: dup-trp-dup local 2-jumps smaller than 100 kb (31) and local+distant 2-jumps of translocation with subsequent tandem duplication (88).

4.6.2 HDP model for SV signatures

Following the HDP design for multiple cancer type groups as in Figure 4.1, I allocated each sample to a leaf node, using a separate concentration parameter for each group of child nodes and the set of all parent nodes, using gamma hyper-priors with shape = 1 and rate = 1.

I ran eight independent burn-in chains—each separately initialised with ten clusters—for 40,000 iterations, and then collected 125 posterior samples at intervals of 300 iterations (1000 total samples from 8 separate chains).

4.6.3 Estimated SV signatures

Using the method outlined in Section 4.2.2, the HDP model returned 15 consensus rearrangement signatures and assigned just 0.3% of SV events (on average) to the zero component for noise and uncertainty—a far lower proportion of uncertain clustering than for the SNV signatures in Section 4.4.

Figure 4.20 presents the fifteen PCAWG rearrangement signatures (sorted by structure, not frequency) with an inverse normalisation such that event class proportions (across signatures) sum to one. This is a different interpretation to

the standard plot showing individual signatures as proper probability distributions integrating to one. Given the extreme differences in SV class frequency, this inverted visualisation allows rare SV classes to be seen alongside common structures. However, the values shown *within* each signature need careful interpretation, as the rearrangement process will generate common SV classes far more frequently than rare SV classes at the same plotted height.

The complex SV footprints are mostly split across two signatures: one generic 'Complex' group, and one 'Complex+Chromoplexy' group co-occurring with chromoplexy cycle events. Fragile site deletions almost exclusively assort to their own 'Fragile Site' signature, which also includes about half the FS tandem duplications, and a range of other deletions enriched in late-replicating regions. Other deletions separate into four different signatures:

- 'Small Deletion', co-occurring with several other classes including: small reciprocal inversion, small insertion bridge, and reciprocal translocation;
- 'Mid Deletion' with few other SV classes;
- 'Large Deletion', co-occurring with large reciprocal inversions and reciprocal inversions within complex clusters; and
- 'Late Deletion' of late-replicating events at any size, also co-occurring with a small fraction of reciprocal inversion events at any size.

Tandem duplications mostly assort over five signatures:

- 'Early Small TD', co-occurring with templated insertions (particularly small insertion cycles) and translocation plus tandem duplication events;
- 'Late Small TD', co-occurring with small dup–inv-dup 2-jumps;
- 'Early Mid TD', co-occurring with large insertion cycles and chains;
- 'Late Mid TD', co-occurring with large dup–inv-dup 2 jumps; and
- 'Large TD' with few other SV classes.

'Unbalanced Translocation' forms a largely separate signature, co-occurring with a small fraction of chromoplexy chains. The 'Reciprocal SV' signature pairs reciprocal translocations with other balanced events like chromoplexy cycles and some reciprocal inversions. Finally, the miscellaneous 'Break+Ligate' signature groups foldback rearrangements with extremely large deletions and duplications, as well as local+distant 2-jumps, chromoplexy chains, and some complex SV footprints involving foldback BPJs.

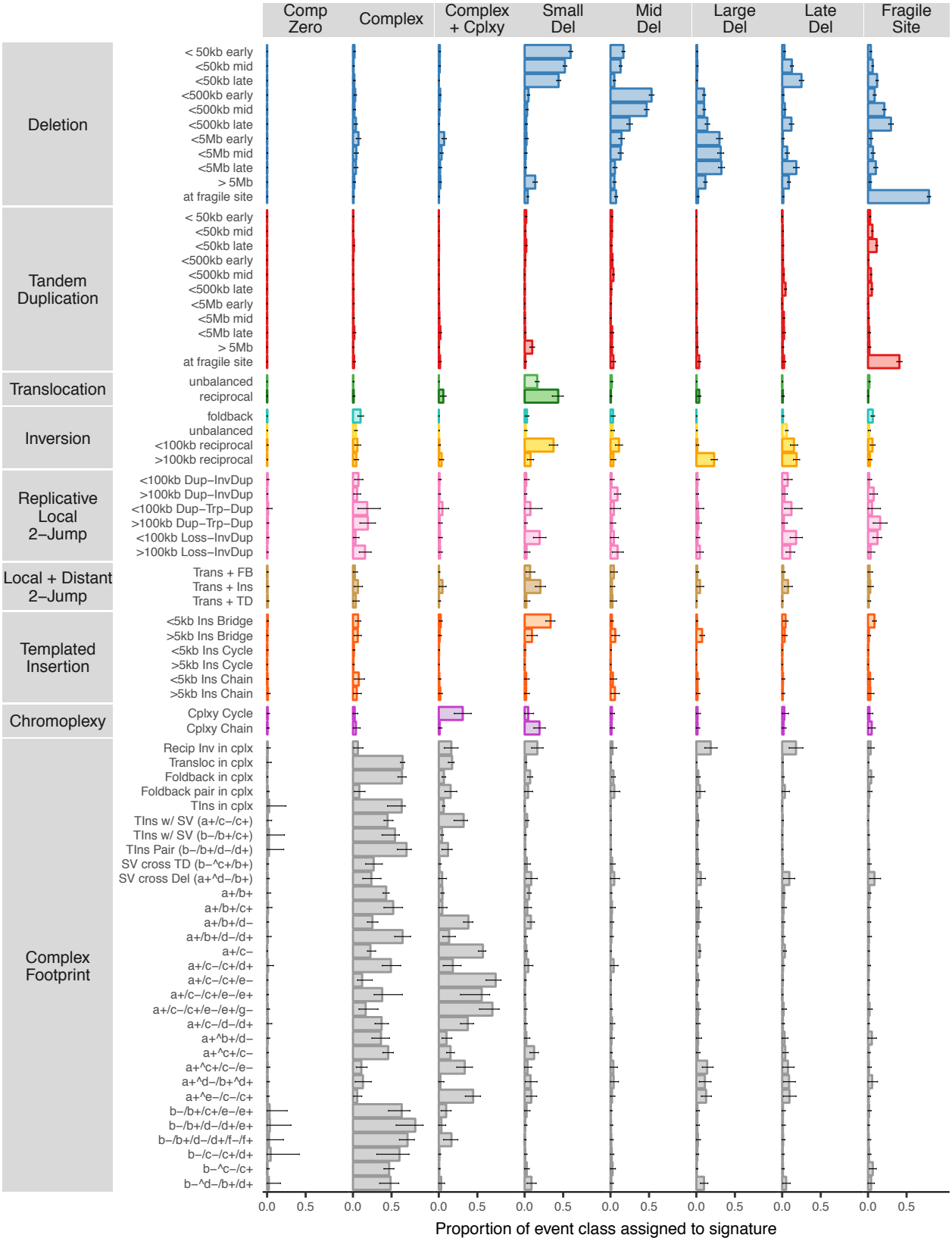


Figure 4.20: SV signatures and 95% credibility intervals, normalised by event class fraction (rows—not columns—sum to one, including the figure continuation on the next page).

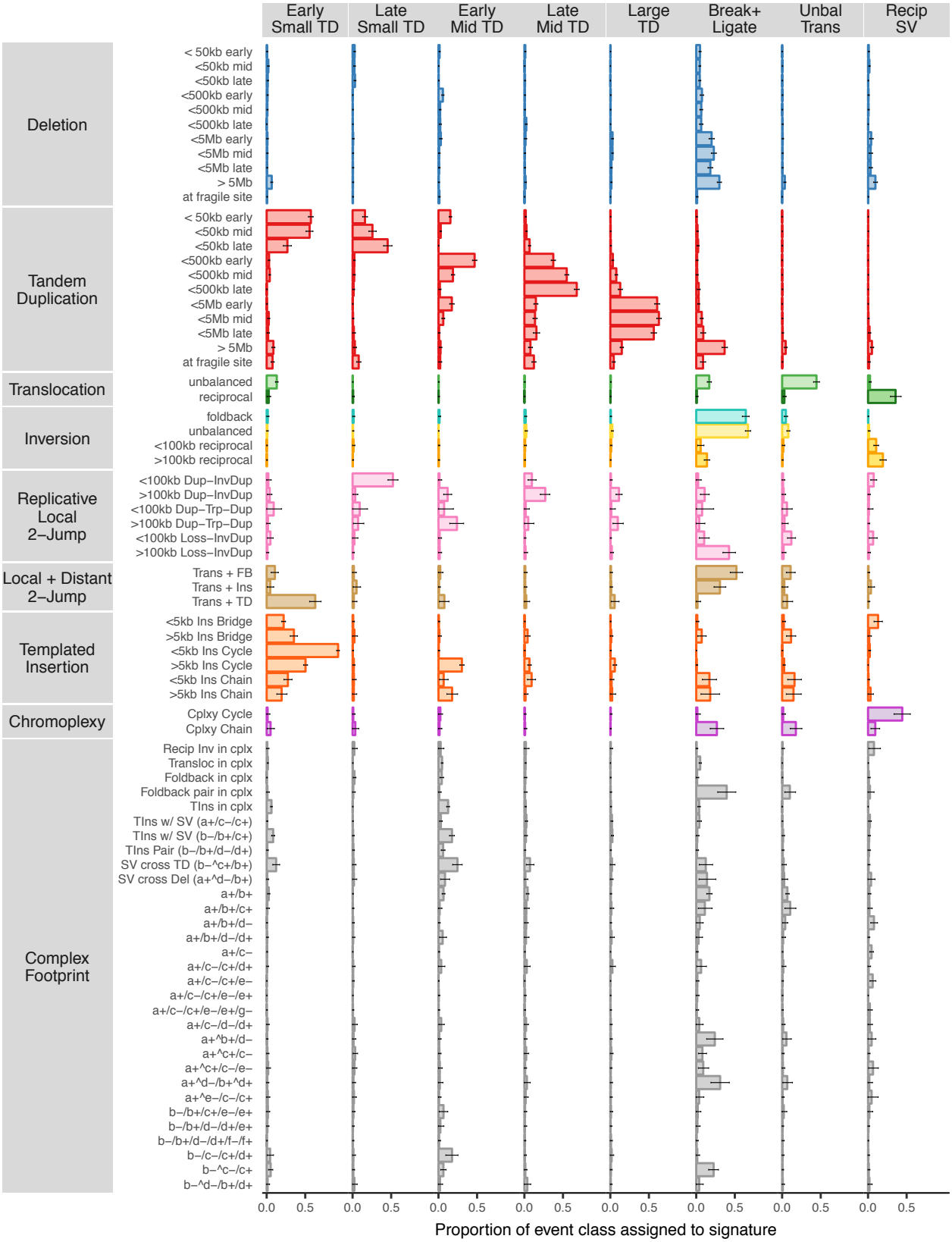


Figure 4.20: Sv signatures and 95% credibility intervals, normalised by event class fraction (rows—not columns—sum to one, including the figure continuation on the previous page).

Figures 4.21 and D.18 show a subset of the estimated sample exposures, recapitulating the basic BPJ census presented in Figures 2.10 and D.2 with the enhanced signature context of size, replication timing, and SV group. Prostate cancer is particularly enriched for the late-replicating deletion and complex chromoplexy signatures; this latter exposure indicates that many BPJ in chromoplexy-associated events are found in the complex unexplained bin. Other cancer types with particularly high exposure to certain rearrangement signatures include: bladder cancer with large deletion and 'break+ligate' SVs; osteosarcoma with complex SVs; medulloblastoma with the unbalanced translocation signature; and colorectal cancer with the fragile site signature. Not all SV events in the fragile site signature are confined to the annotated FS regions, as the other deletions in the signature are more common than their relative values in Figure 4.20 suggest (because of the inverted normalisation to visualise common and rare SV classes concurrently). The exposure patterns in breast, liver[r] and uterus highlight the different replication timing skews of tandem duplication by sample, extending the results of Section 3.2.3.

4.6.4 Discussion

The fifteen rearrangement signatures presented in this section are an apotheosis of the results presented in Chapters 2 and 3, combining SV class, size, and location (as represented by replication timing) into one decomposition of underlying rearrangement processes with characteristic structural readouts and varying activity levels across samples and groups.

The co-occurrence patterns indicate the same underlying condition may generate different structural forms with similar properties of size and/or location. For example, deletions coincide with reciprocal inversions of a similar size range, presumably mediated by break and ligate repair of DSBs at consistent intervals. Similarly, tandem duplications in late-replicating regions coincide with dup–inv-dup 2-jumps of a similar size range, presumably mediated by template and replicate repair of invading strands with consistent processivity (mechanisms reviewed in Section 1.4.1). The conditions generating fragile site deletions also foster FS tandem duplications and a range of other late-replicating deletions, possibly present in un-annotated FS regions.

Compared to previous reports of five or six relatively simplistic SV signatures

[r]For liver cancer, the outlying high-burden sample with large duplications was previously illustrated in Figure D.6.

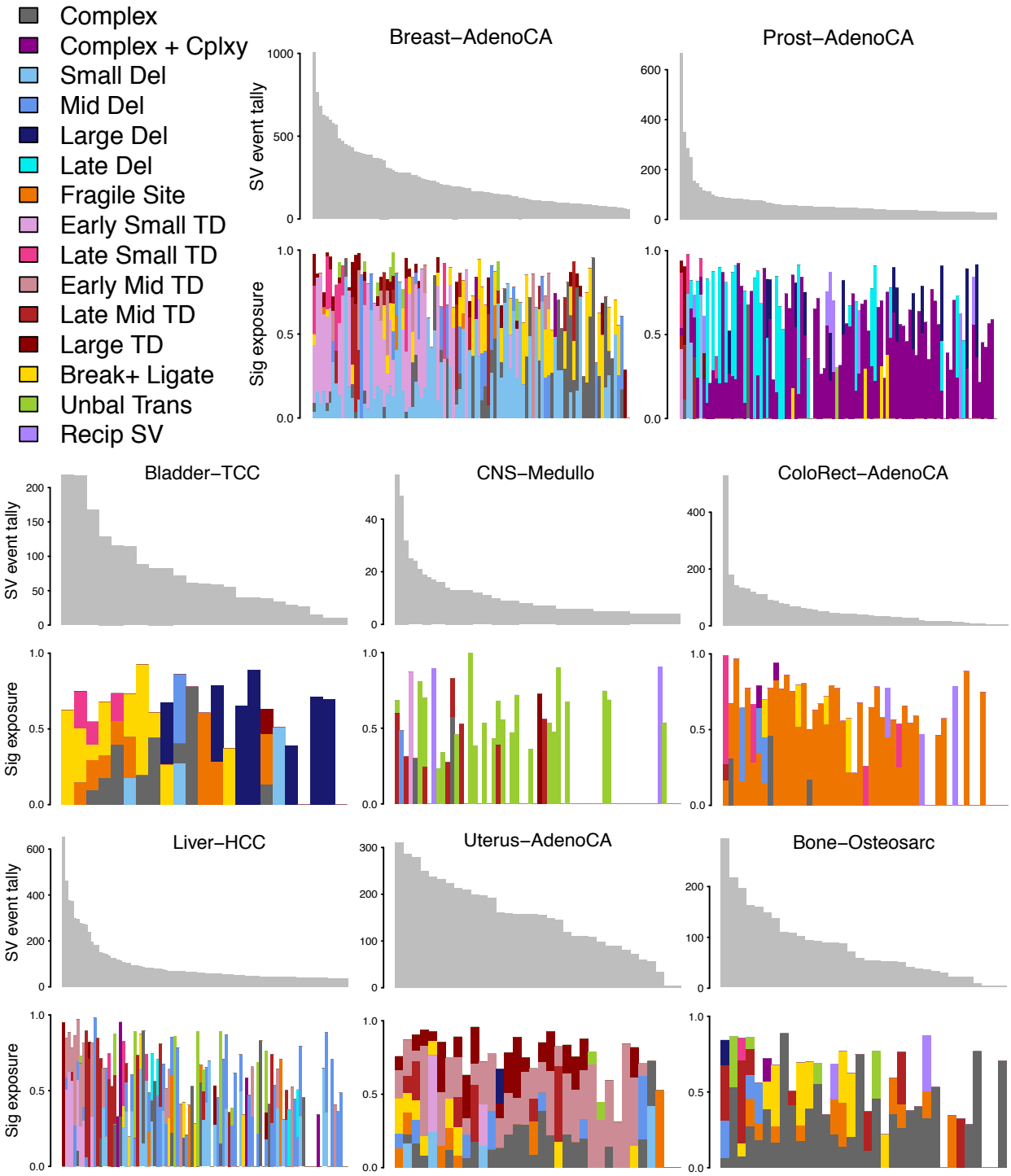


Figure 4.21: Average estimated sample exposures to HDP-extracted SV signatures (Figure 4.20) for eight of the PCAWG cancer types. Large cohorts are subset to a maximum of 100 samples for presentation. For each sample, mean estimated exposures are only plotted for significant signatures (95% credibility interval above zero), leaving a blank proportion with uncertain allocation to the same or different signatures.

(Nik-Zainal et al., 2016; Hillman et al., 2018), this analysis of 76 SV categories—including many novel structures—tallied across 2050 samples in a pan-cancer cohort is the most detailed and comprehensive summary of rearrangement signatures yet produced. However, the scope for further improvement is vast. My HDP method currently requires genome alterations to be tallied in discrete, unordered categories, as do the NMF-based methods reviewed in Section 4.1. When classifying SV events, a large swathe of the complex rearrangement landscape remains intractable to simple categorisation, and is largely precluded from signature analysis. Even if BPJ in complex events like chromothripsis or BFB were to be classified, it remains unclear how these large-scale phenomena should be tallied for meaningful comparison against a simple deletion count, for example. For the SV events that do have classifiable structures, their other pertinent features of size and replication timing are best measured as quantitative variables. My current categorisation of size and timing is a crude substitute for the real value, causing edge effect bias and violating the assumed independence of separate alteration classes. Ideally, future signature analysis methodologies will handle quantitative event features (perhaps using a similar approach to Shiraishi et al. (2015)), and SV signatures may extend to additional features such as microhomology and chromatin state.

4.7 Discussion

In this chapter, I introduced the hierarchical Dirichlet process as a novel strategy for mutational signature decomposition, and derived a set of fifteen somatic rearrangement signatures with unprecedented detail and scale.

The HDP model was first developed by Teh et al. (2006) for topic modelling in corpora, but it is also well-suited to mixed-membership cluster problems in biology, such as the mutational signature decompositions explored in this thesis. The flexible tree of hierarchical DP nodes provides a natural framework for grouping samples by any number of pertinent factors, such as cancer type, germline genotype, mutagen exposure, or patient of origin (if multiple metastases or subclones are available from the same individual). This consideration of sample relatedness empowers the clustering procedure to borrow information across disparate groups, while upholding the prior expectation of differences between groups. In contrast, most other methods perform siloed signature extraction in separate cancer types, with a post hoc consolidation to match results across groups. As the MCMC posterior sampling method naturally gener-

ates credibility intervals for every signature and sample exposure estimate, HDP quantifies significant differences between samples and groups with a justifiable comparison often lacking in other methods. Two further advantages of the HDP approach stem from its nonparametric Bayesian assumption of infinitely many generative processes. First, this enables the number of underlying signatures to be automatically determined from the complexity of the data itself so that—unlike many other methods—HDP does not need to separately assess all plausible signature numbers to find the optimal fit. Second, HDP easily conditions on a prior set of known signatures while simultaneously finding novel clusters in a new dataset. This property is particularly important for small and/or heterogeneous cancer cohorts which might be underpowered for completely *de novo* signature extraction, but which nevertheless contain some number of previously undescribed signatures (particularly artefacts).

One of the major downsides to my current `hdp` R package implementation is that runtime and memory both scale at a roughly linear rate with the number of observed mutations. For the dataset of about 5 million SNVs analysed in Section 4.4, every 1000 MCMC iterations required approximately 3 CPU hours. Although this speed was sufficient to complete analysis in under a week (human time) with parallel computing, the computational expense is prohibitive for larger datasets such as the entire PCAWG SNV catalogue of almost 44 million mutations. For a collection of 10–100 million items, MCMC inference could still be made in separate silos of relevant cancer type groups (as is done for NMF and other methods), *or* an alternative variational inference method (reviewed by Blei et al. (2017)) could approximate the optimal solution in far less time. Several variational inference methods have been proposed for the HDP model, with two available Python packages to support multinomial data (Wang et al., 2011; Hughes et al., 2015).

Another limitation of my current HDP package is the multinomial distribution definition for mutational signatures. By modelling genome alterations as discrete, unordered categories, any quantitative features are forced into crude bins, relationships between similar alteration types are ignored, and the parameter space multiplies with each subdivision for additional features. The requirement for a modest number of separate mutation categories (perhaps less than one thousand) is the major reason most SNV analyses are restricted to 96 classes defined by trinucleotide context, despite the relevance of other signature features such as pentanucleotide context, replication timing, chromatin state, and transcriptional and replication strand bias (Shiraishi et al., 2015; Haradhvala

et al., 2016; Morganella et al., 2016). With the simplifying assumption of independence between features, Shiraishi et al. (2015) proposed a novel approach where each signature is modelled as a collection of distributions over different mutational features. Although currently implemented for categorical features only, the same principle should extend to quantitative variables, and offers an appealing solution for SV alterations defined by many disparate properties such as form, size, microhomology, complexity, and genome topography. In future work, I propose that the nonparametric Bayesian HDP framework—with its many advantages for modelling sample relatedness, conditioning on prior knowledge, quantifying uncertainty, and learning the signature number—could be extended to signature clustering based on sets of independent multinomial and Gaussian distributions, and thus combine the best aspects of both strategies.

Other directions for future improvement include: adoption of formal convergence diagnostics to assess MCMC chains; refinement of my post-processing signature extraction procedure (Section 4.2.2); and incorporation of other topic modelling developments to explicitly account for correlation between topics/signatures (Blei and Lafferty, 2007; Kim and Sudderth, 2011) and/or impose sparsity constraints (Wang and Blei, 2009; Williamson et al., 2010).

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.

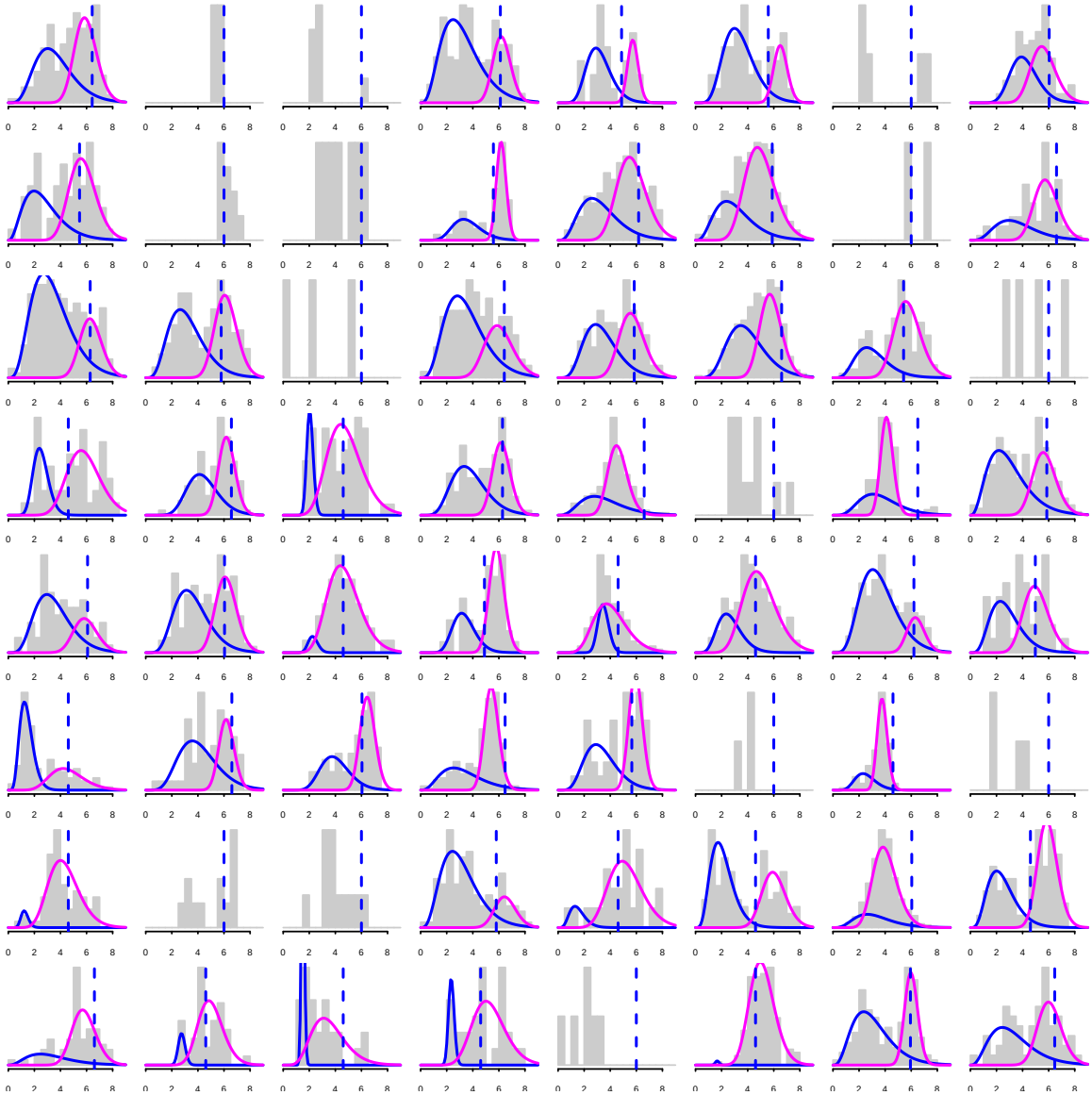


Figure 5.1: The distribution (shown in grey histogram on a $\log_{10}$ scale) of inter-break distances between adjacent (on same reference chromosome) positions of complex unexplained BPJ in 64 randomly chosen PCAWG samples. For samples with 20 or more inter-break distances, the primary footprint partition 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.

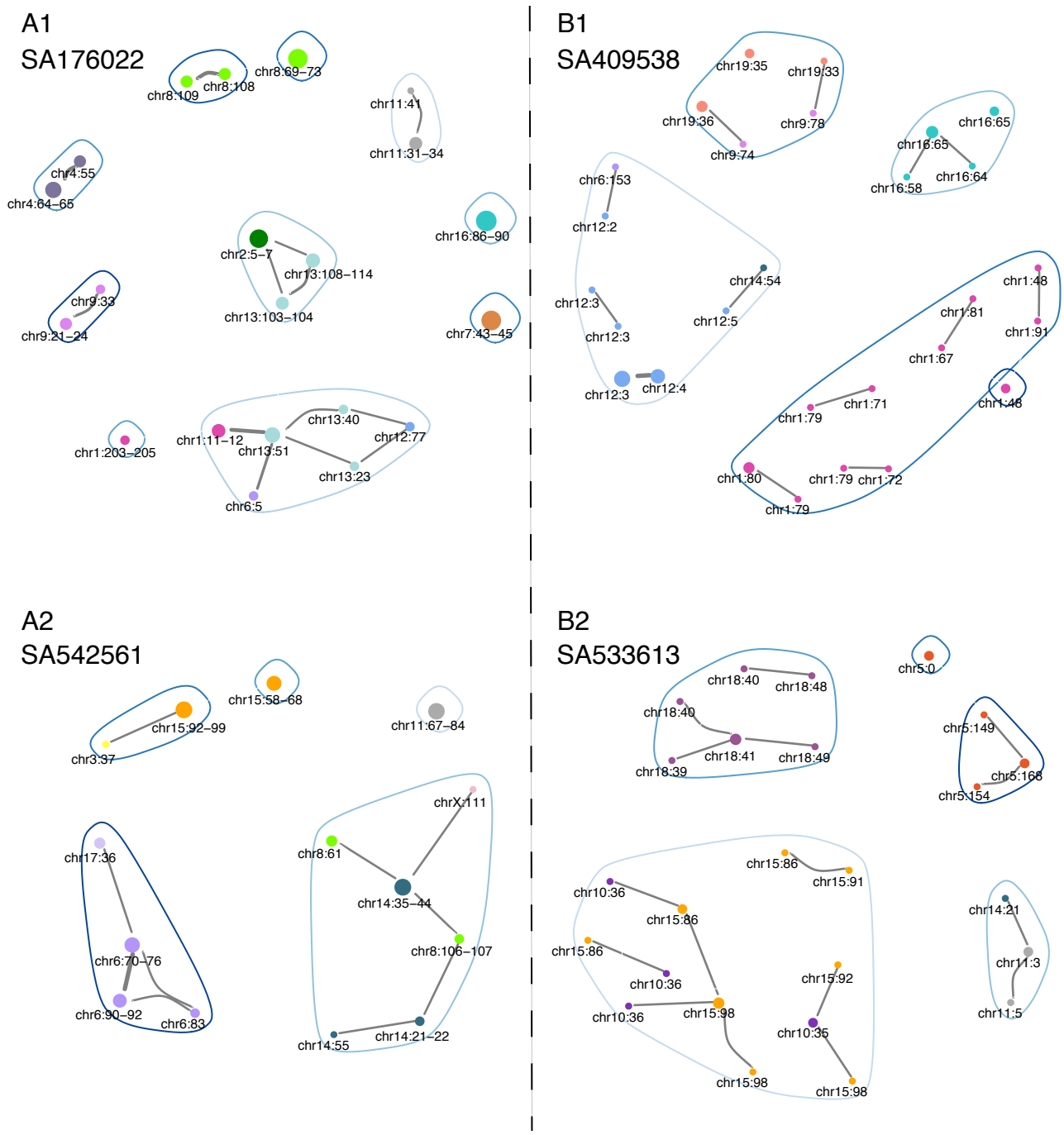


Figure 5.2: Node-edge graph representation of the complex unexplained BPJ in four PCAWG samples. Each node is a genome footprint, coloured by reference chromosome with size corresponding to breakpoint count. Node labels indicate the chromosome position in megabase units. Each edge indicates breakpoint 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.

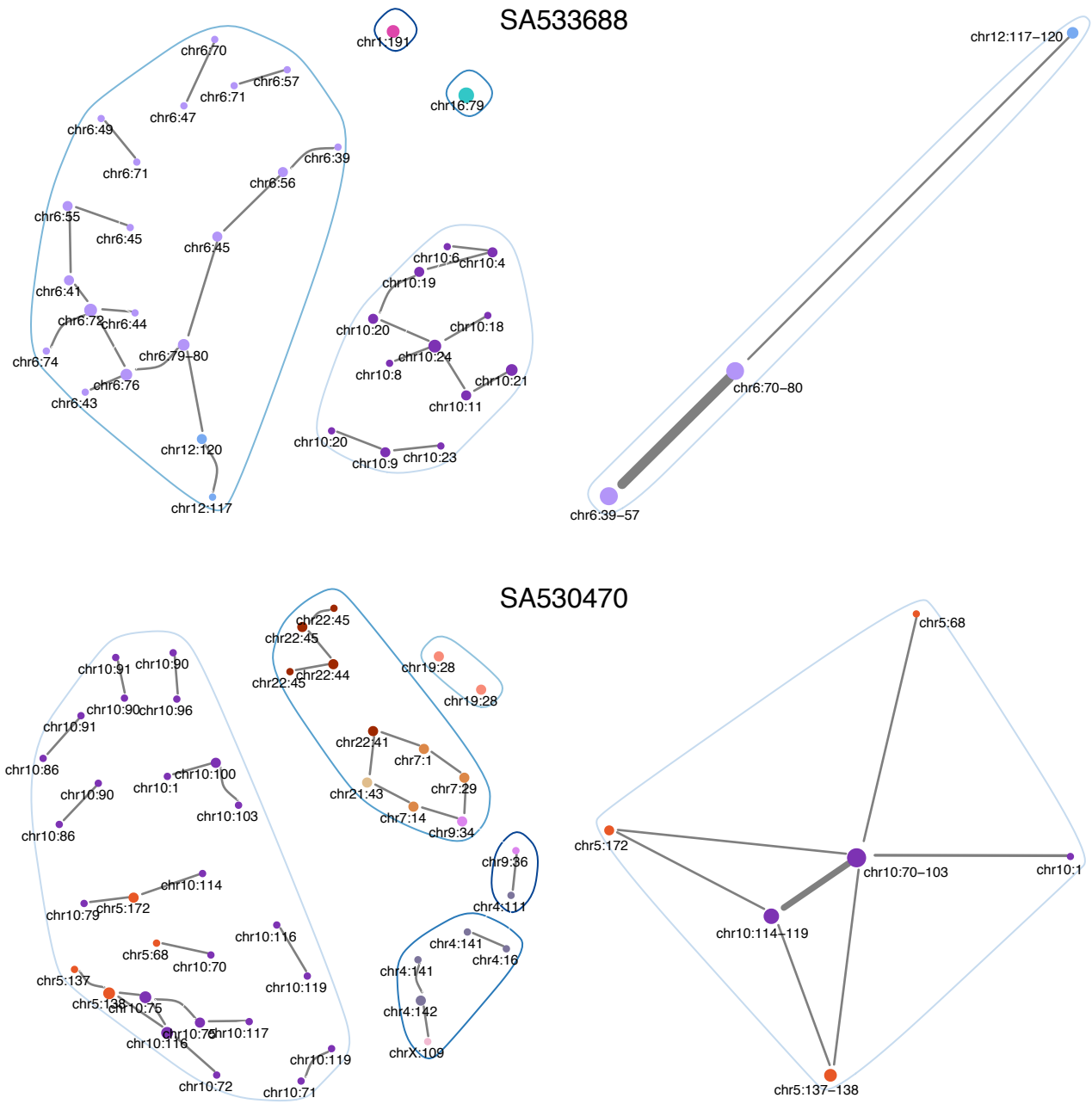


Figure 5.3: Two samples containing large BPJ clusters with no separable sub-graphs. The left side graphs show all complex BPJ in each sample. The right side graphs show the secondary footprint partition of the large candidate cluster, with blue circles indicating that the walktrap community detection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

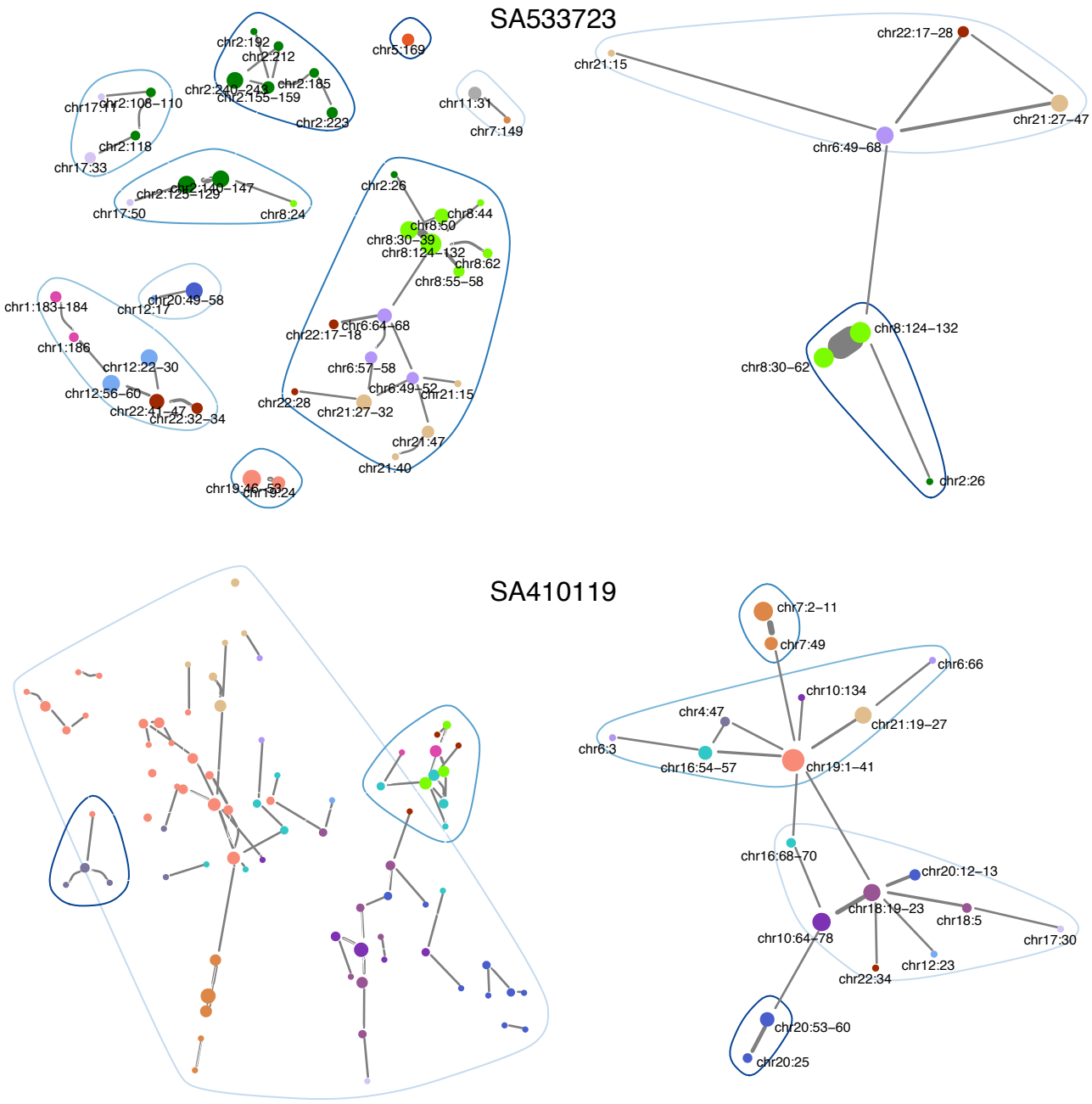


Figure 5.4: Two samples containing large BPJ candidate clusters with fully separable sub-graphs. The left side graphs show all complex BPJ in each sample. The right side graphs show the secondary footprint partition of the large candidate cluster, with blue circles indicating sub-graphs found by walktrap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

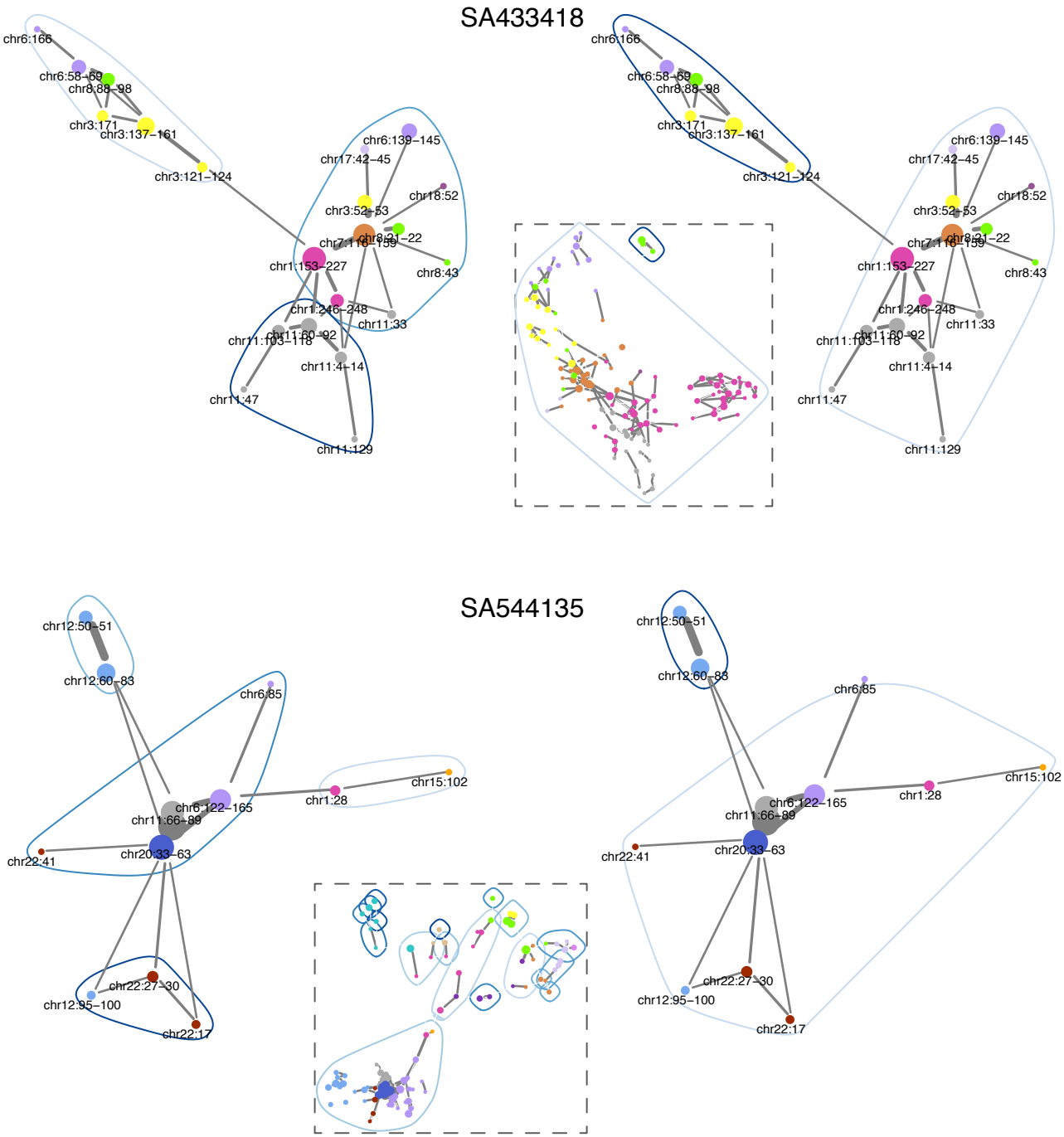


Figure 5.5: Two samples containing large BPJ candidate clusters with partially separable sub-graphs. The left side graphs show the secondary footprint partition of the candidate cluster, with blue circles indicating sub-graphs found by walktrap community detection. In each case, only one sub-graph meets 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.

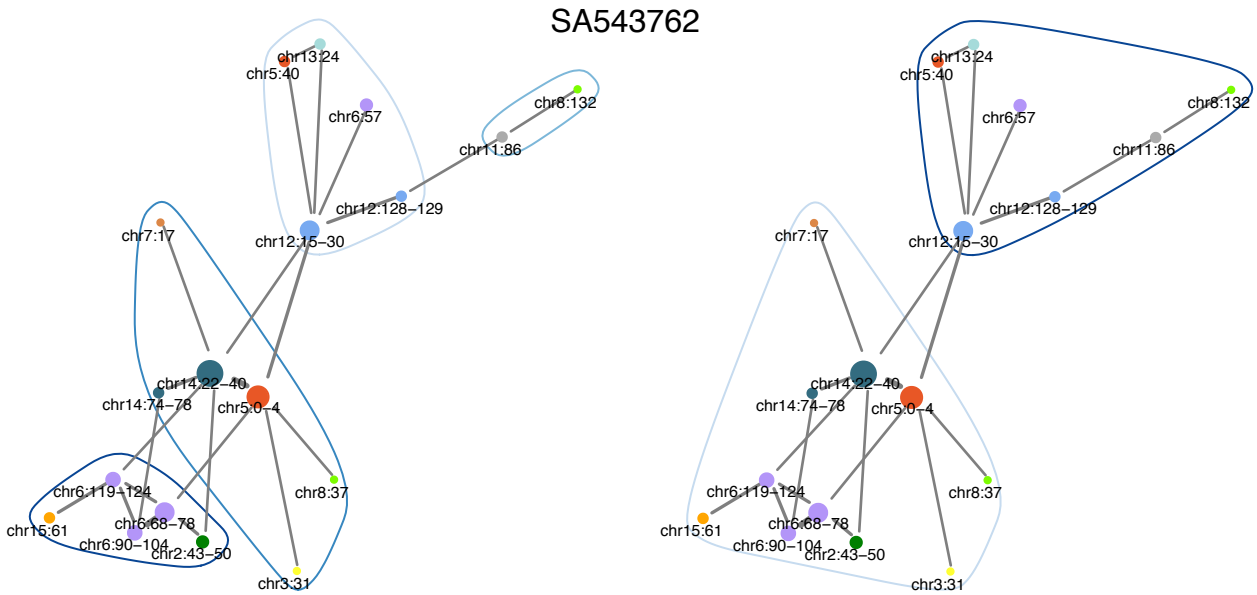


Figure 5.6: A large candidate BPJ cluster with separable sub-graphs following extra agglomeration. The left node-edge graph shows the secondary footprint partition, with blue circles indicating community sub-graphs. In this case, none of the four initial sub-graphs meet the separation criteria. Following 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)

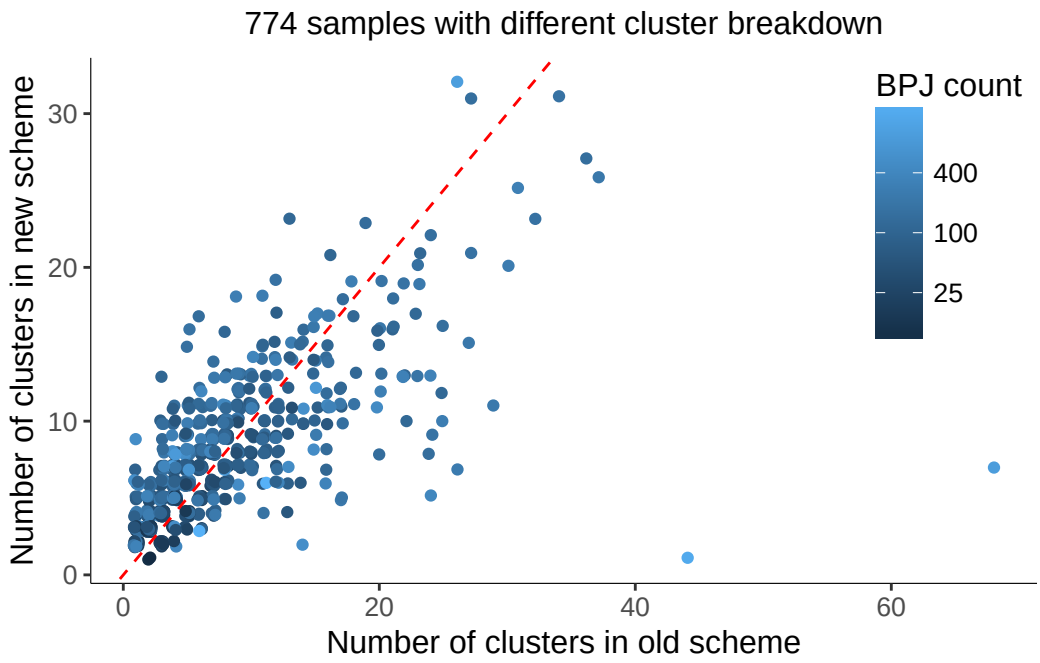


Figure 5.7: Discrepancy in complex unexplained BPJ cluster counts between new and old schemes for 774 PCAWG samples. Red dashed line separates samples with more clusters in new scheme (top left) from those with fewer clusters in new scheme (bottom right).

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

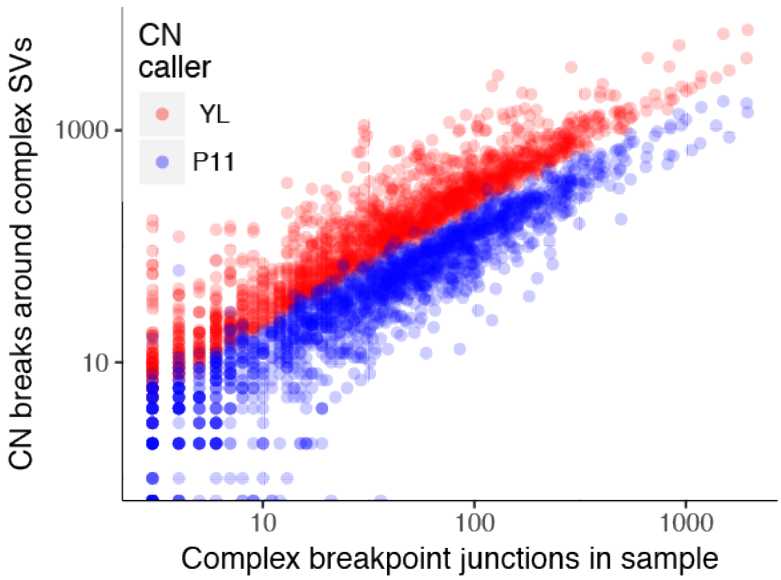
5.3 Matching complex SV with CN estimates

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.

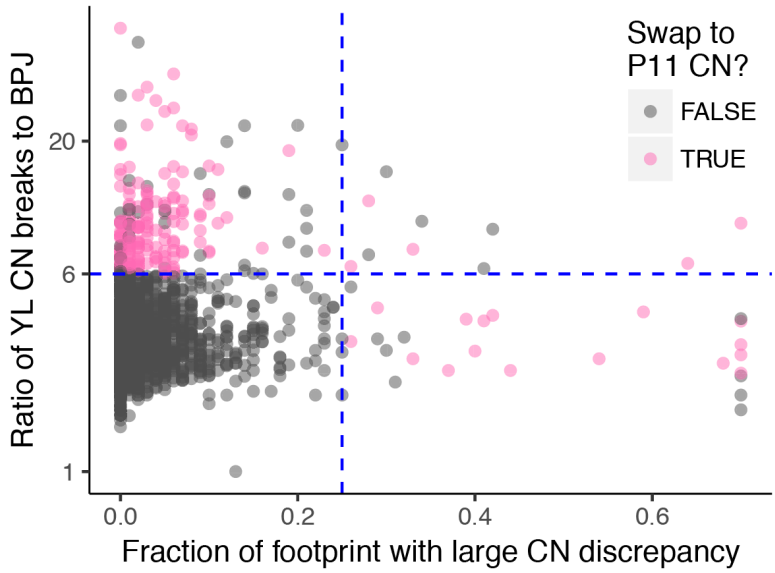
For the 1811 samples with complex unexplained BPJ (excluding tiny clusters from Section 5.2), I consider the CN profiles returned by YL and P11 in 1 Mb flanks around each breakpoint, leaving no gaps smaller than 5 Mb. I also round the non-integer YL calls to 0.05 intervals to disregard any minor change-points 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).

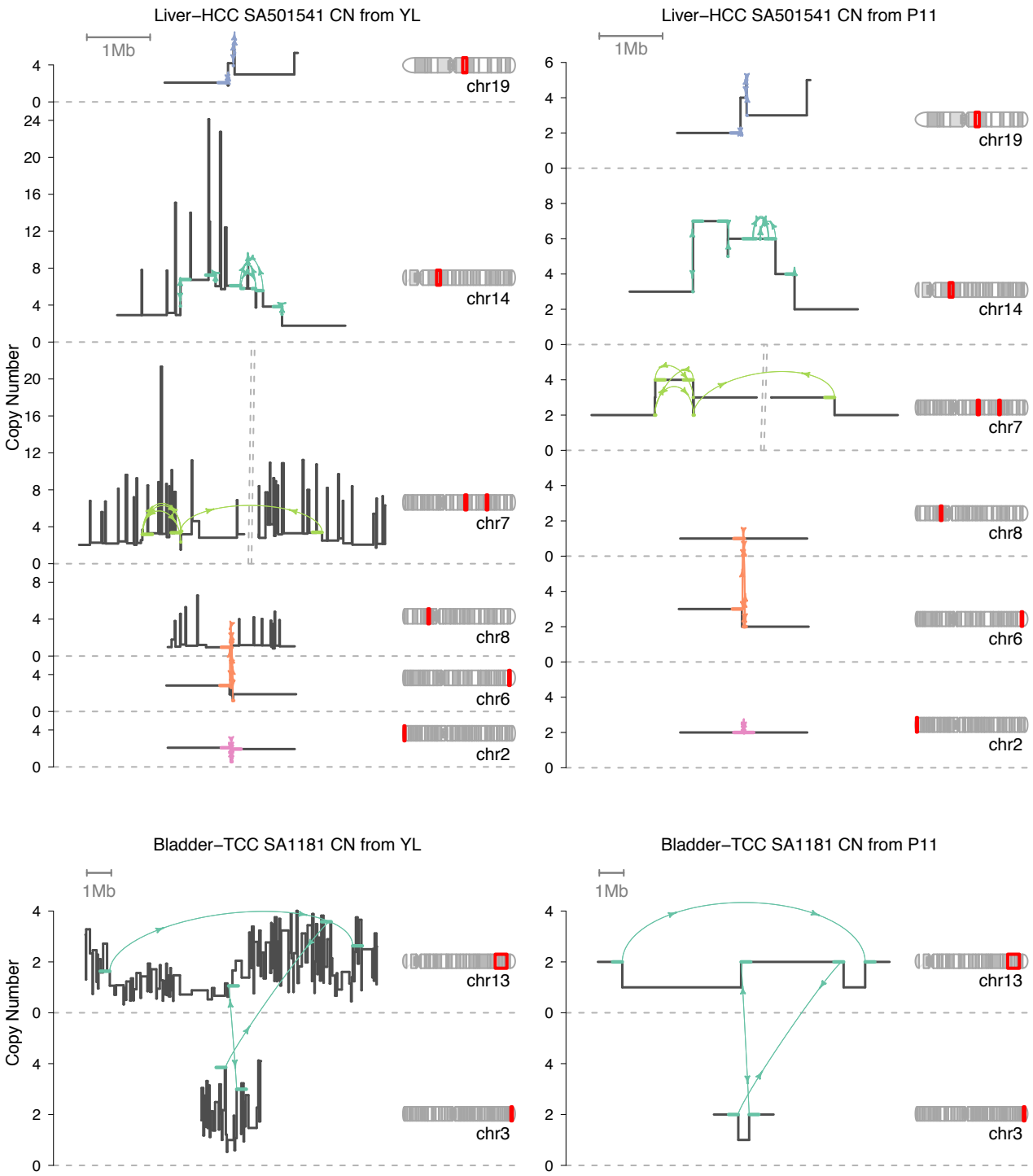


Figure 5.9: Copy number profiles returned by YL (left) and P11 (right) around complex unexplained BPJ in samples qualifying for a switch to P11 CN. Breakpoint junctions are coloured by cluster assignment within the sample.

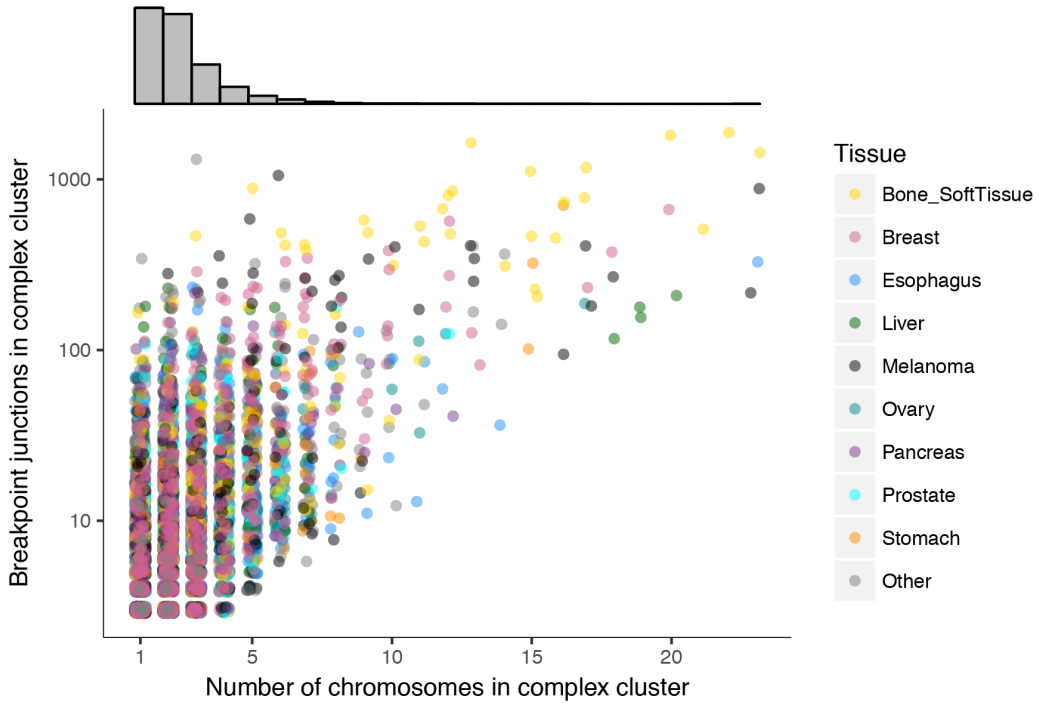
5.4 Outlying clusters and samples

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.

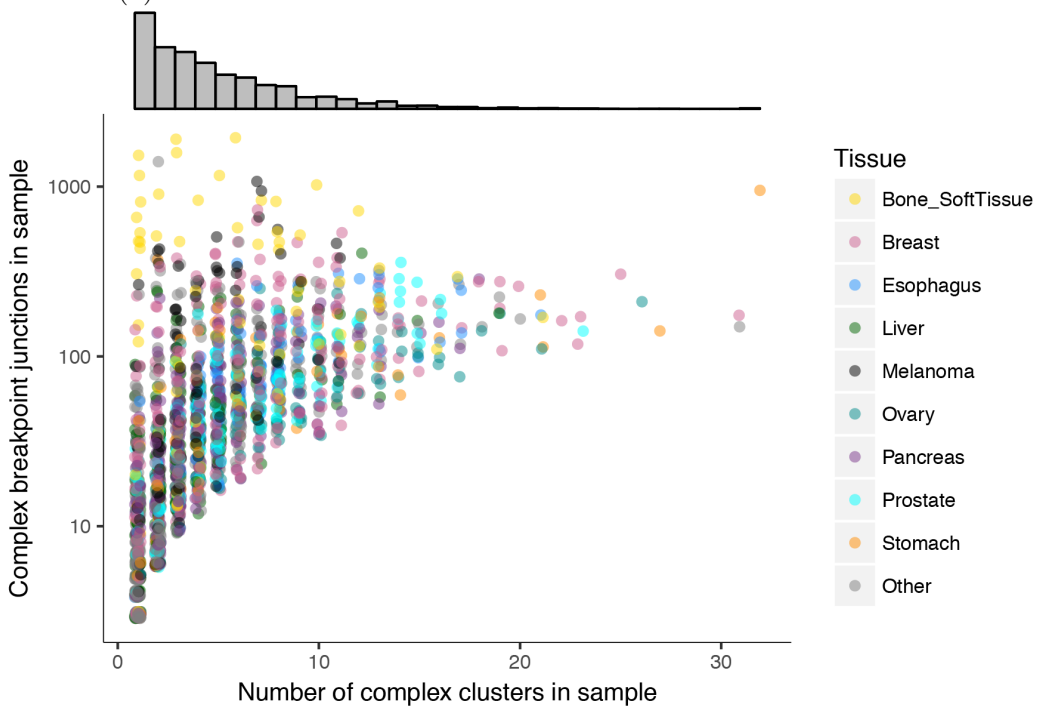
One outlying event involving more than 1000 BPJ distributed over just three chromosomes—and primarily two chromosomes upon inspection—is the kidney renal cell cancer rearrangement shown in Figure 5.11. This event has the characteristic hallmarks of chromothripsis, with short fragments along two distinct 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.

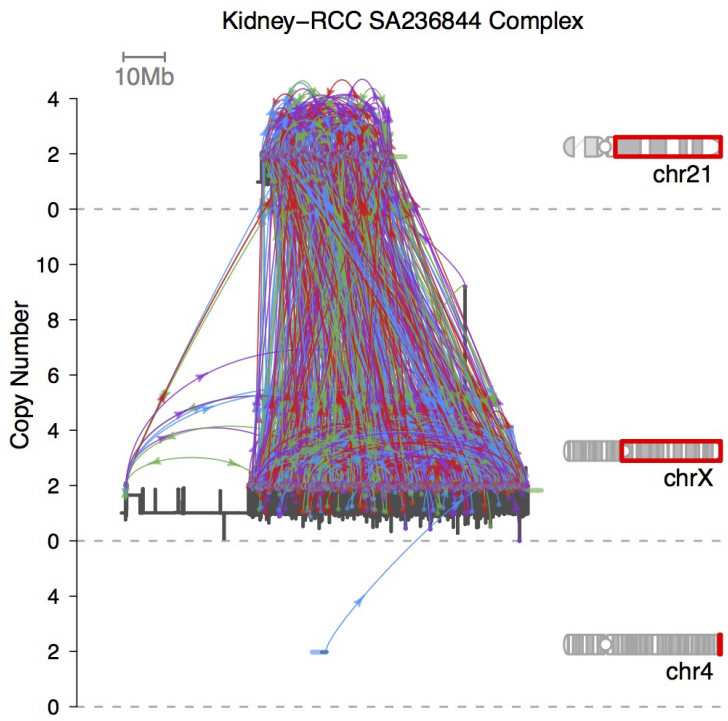


Figure 5.11: Unusual chromothripsis event with 1365 BPJ spanning two chromosome arms in a kidney renal cell cancer.

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.

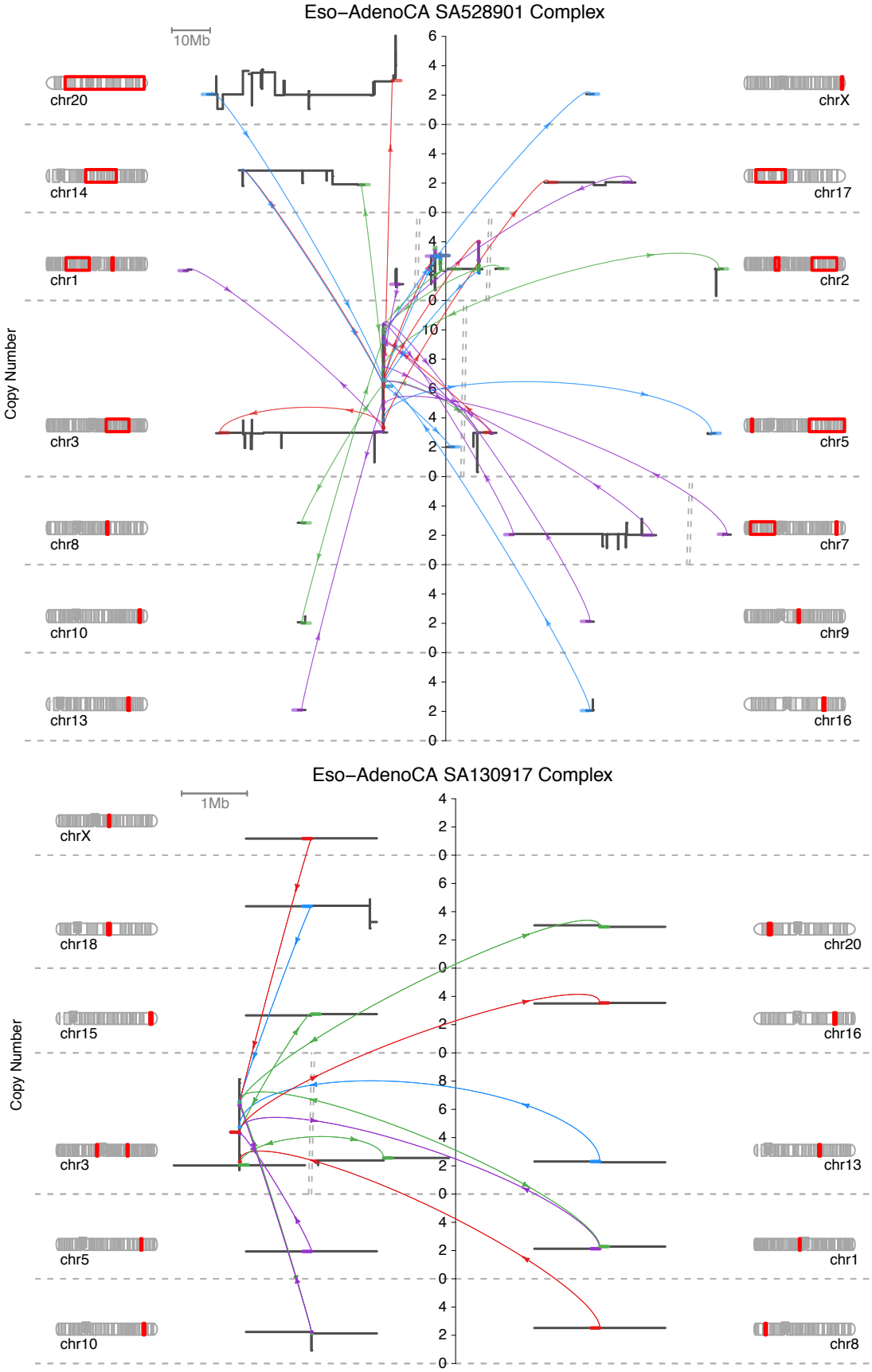


Figure 5.12: Somatic retrotransposition clusters spanning many chromosomes with relatively few BPJ.

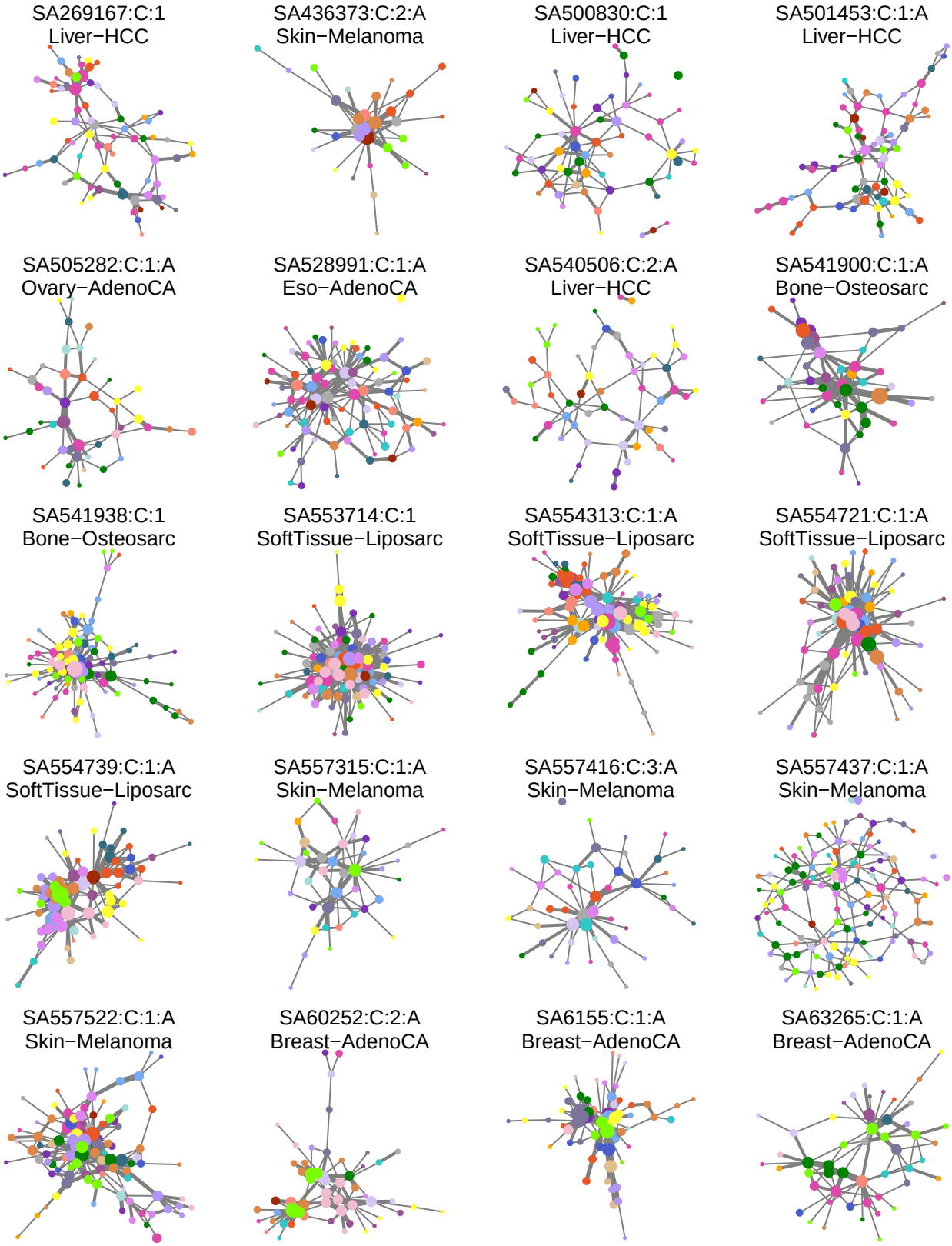


Figure 5.13: Graph representation of all BPJ clusters spanning 17 or more chromosomes. The footprint nodes partition adjacency gaps greater than 5 Mb.

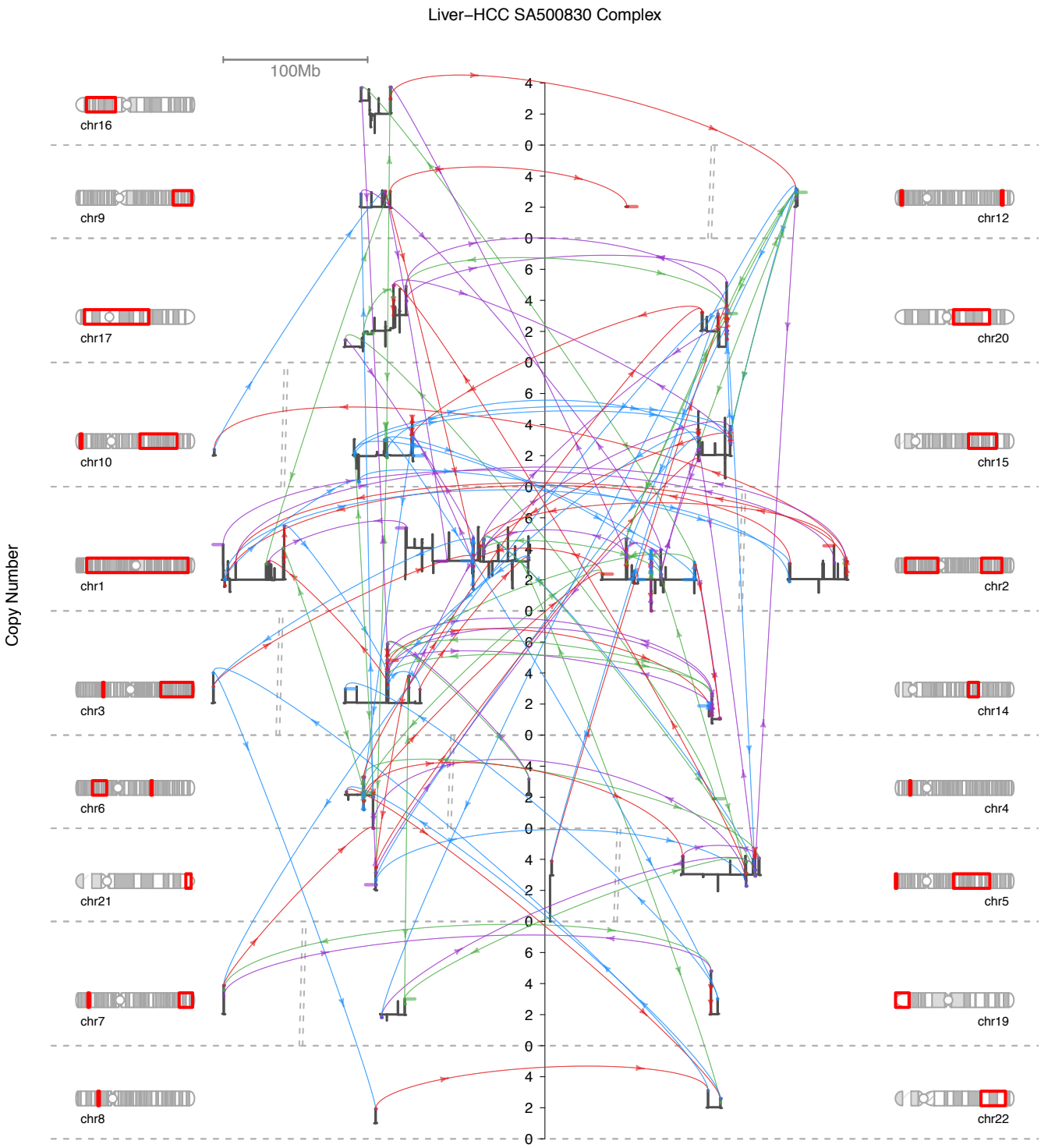


Figure 5.14: Complex SV cluster in a liver cancer sample spanning 19 chromosomes with 155 BPJ.

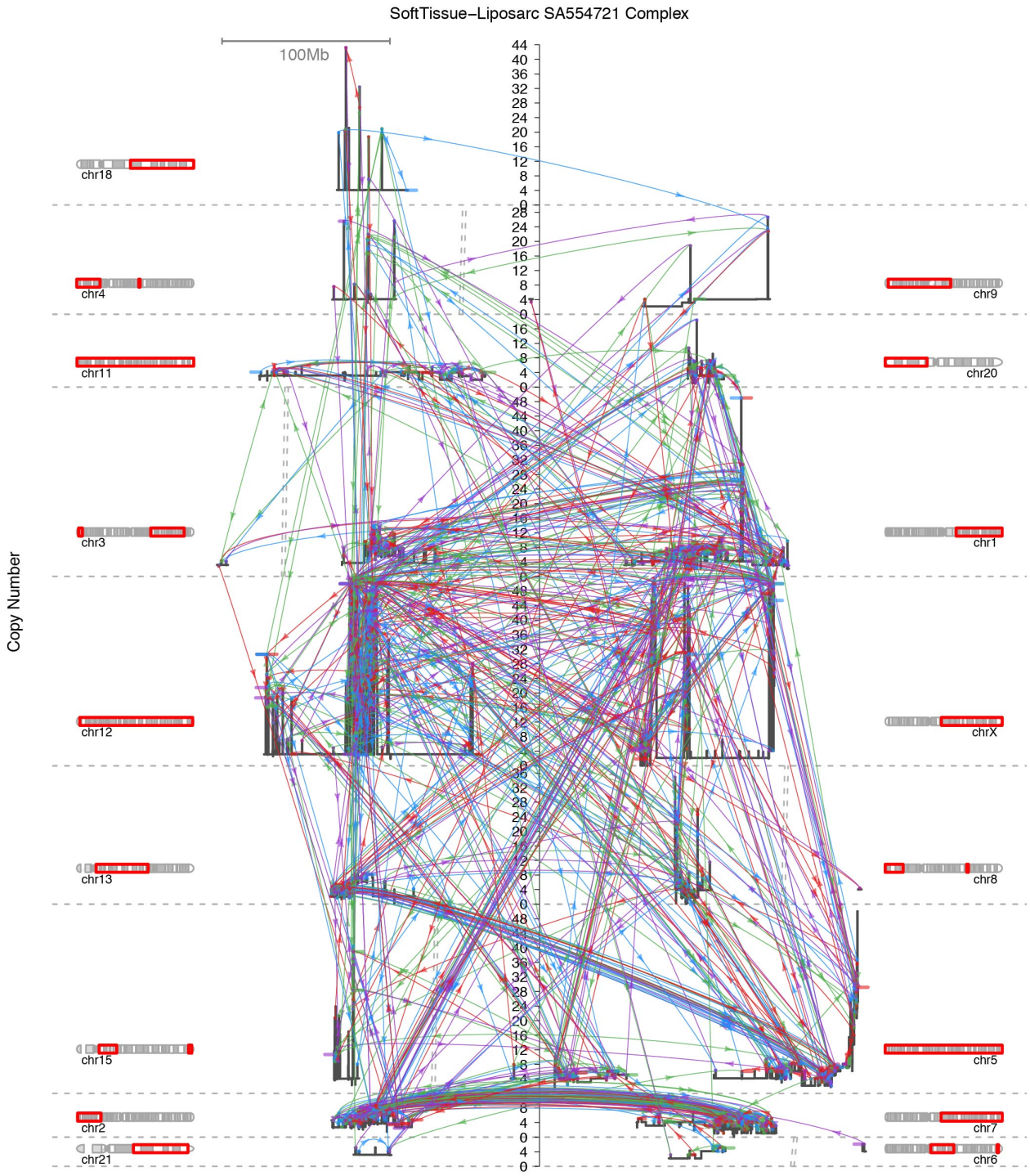


Figure 5.15: Complex cluster in a liposarcoma sample spanning 17 chromosomes with 1122 BPJ. The vertical copy number scale is limited to a maximum of 50.

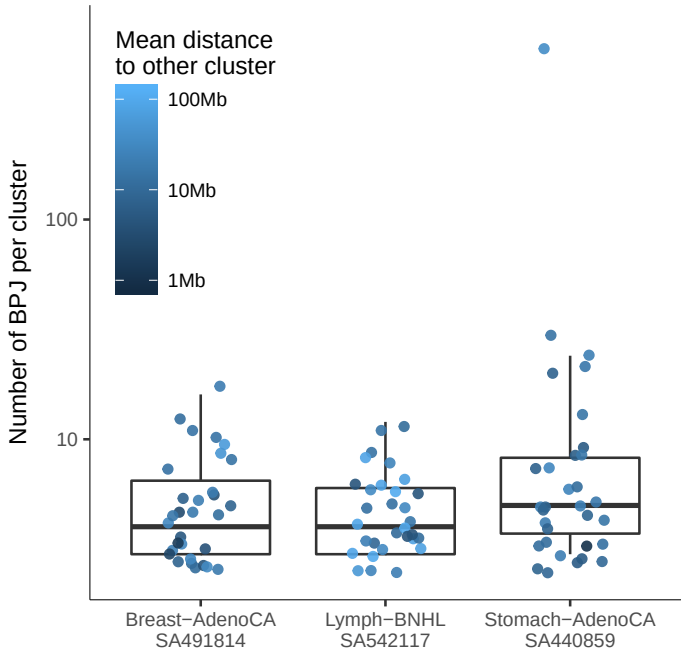


Figure 5.16: Number of BPJ per cluster in three outlying samples with more than 30 separate complex clusters. Each dot is shaded by the average distance between breakpoints within that cluster and the next closest breakpoint in a different complex cluster.

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 $k = 3, 4, \ldots$).

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.

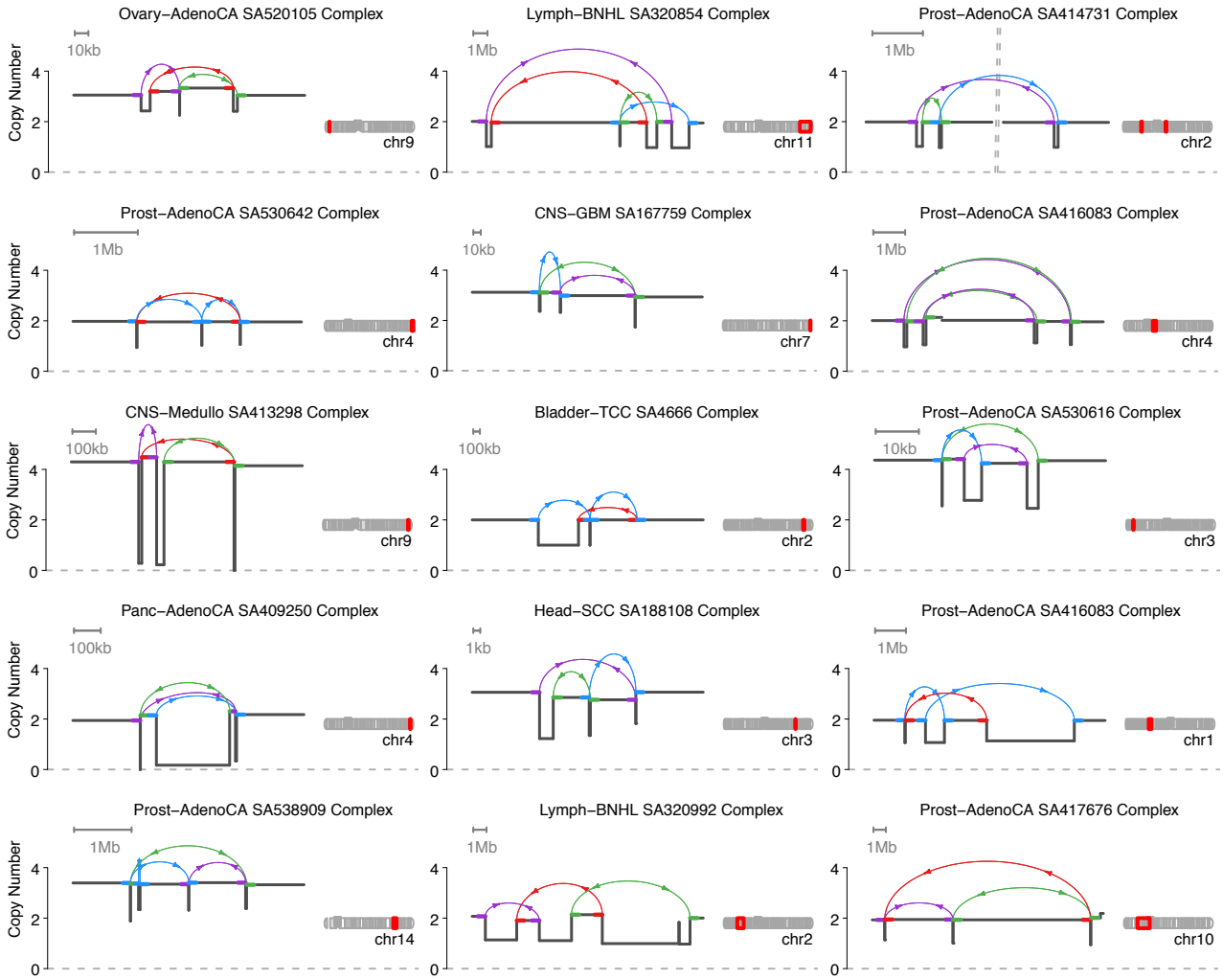


Figure 5.17: Small break and ligate clusters on one chromosome

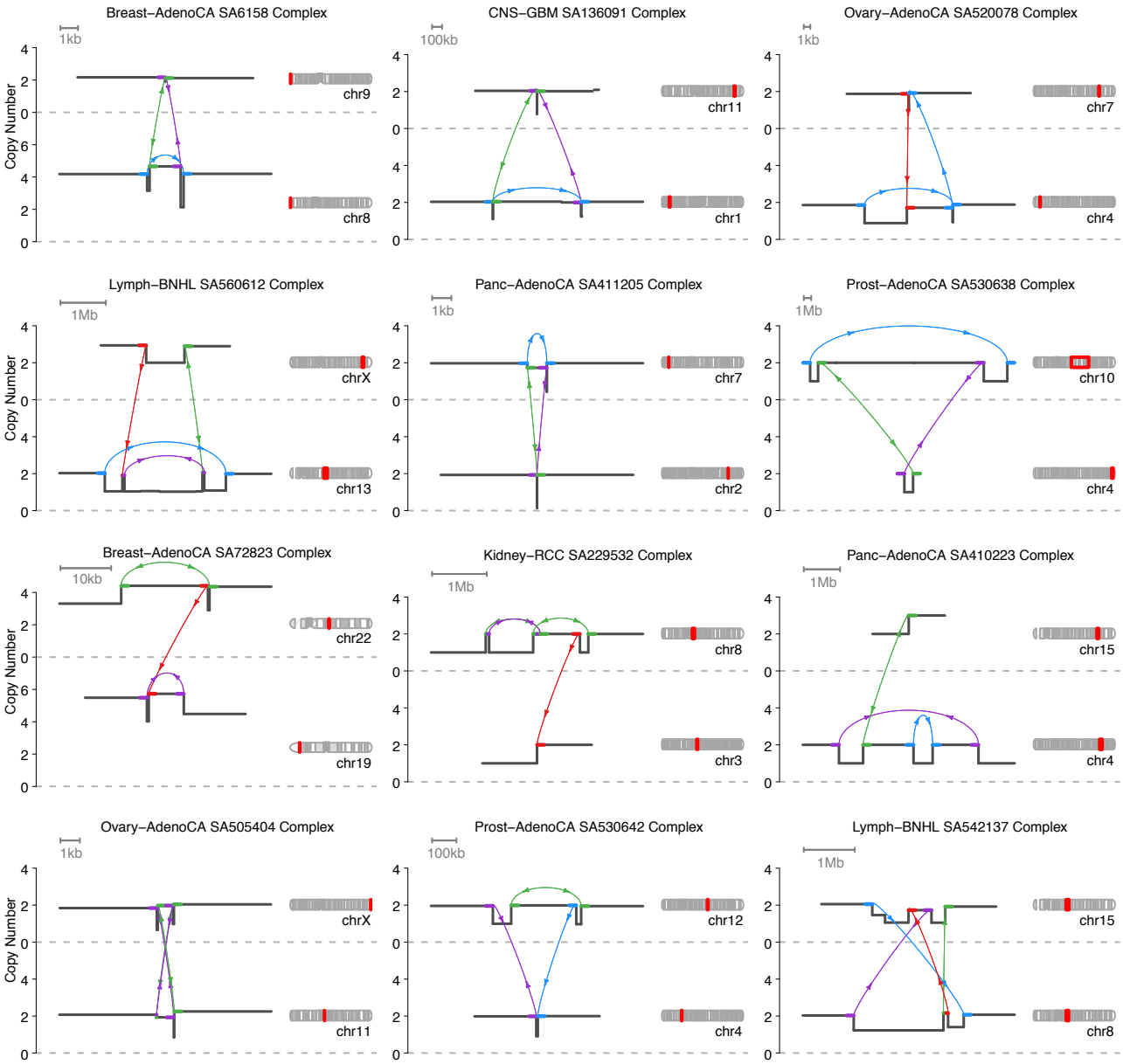


Figure 5.18: Small break and ligate clusters on two chromosomes

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.

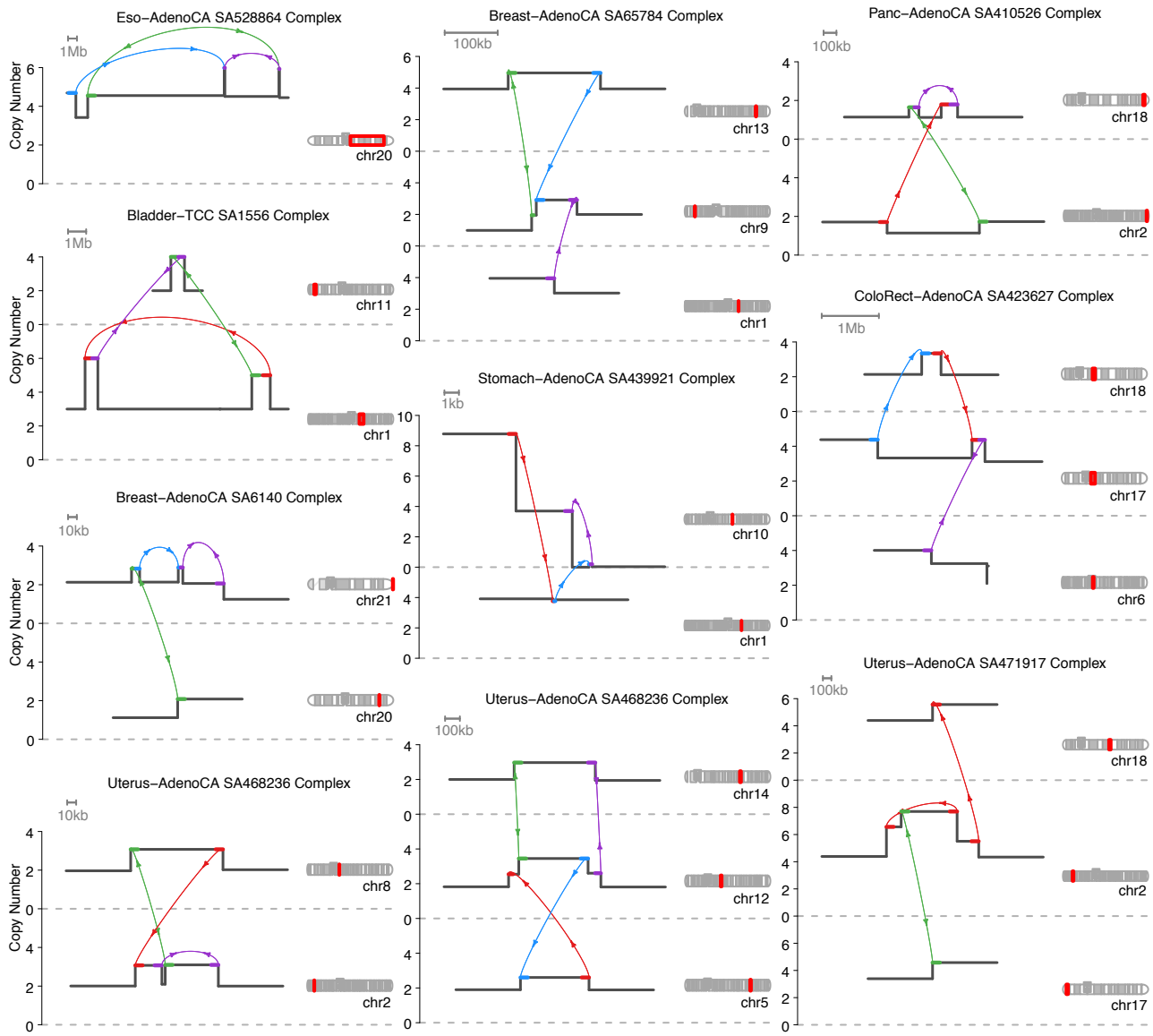


Figure 5.19: Small complex templated insertion events with adjacent or overlapping footprints.

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

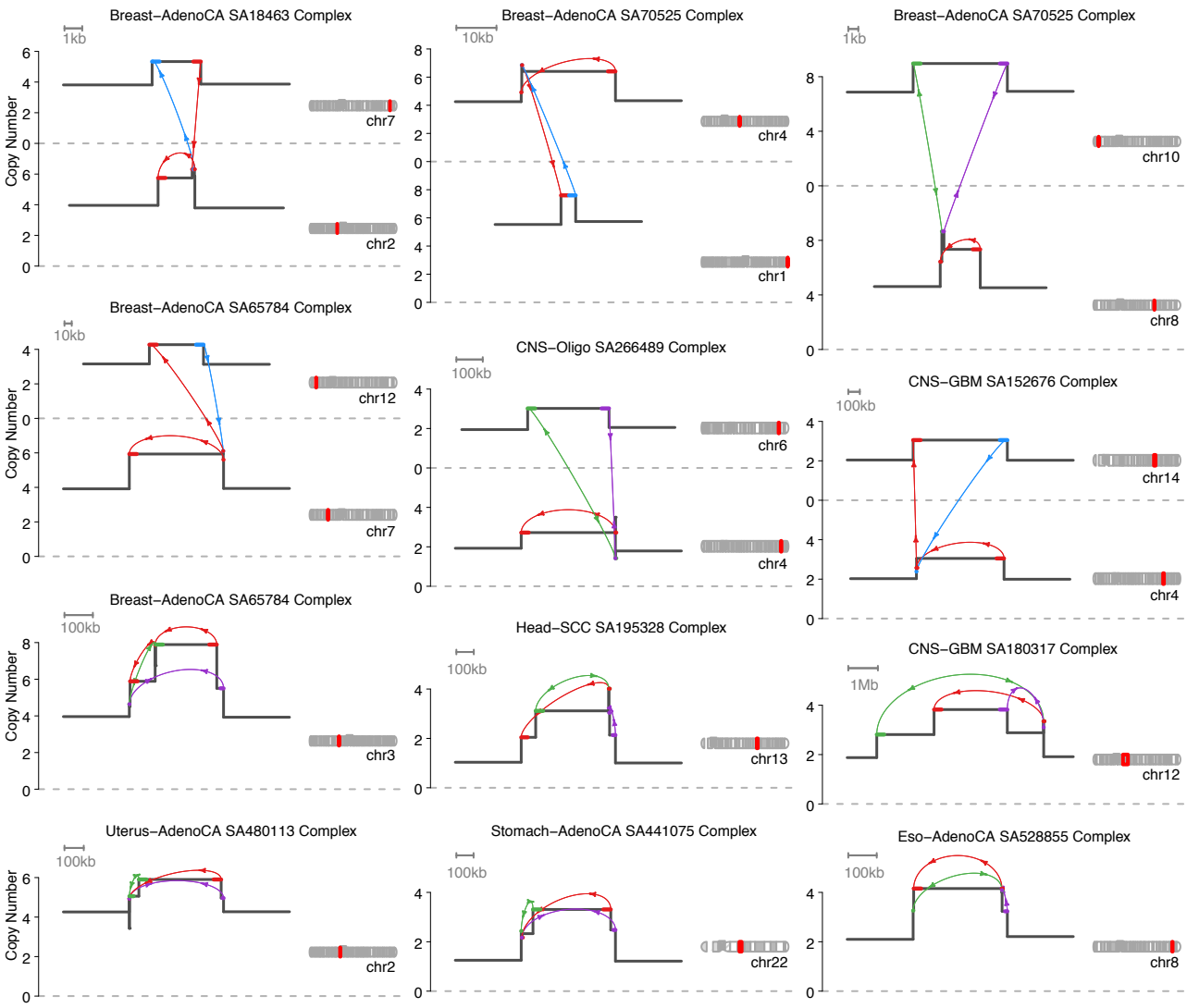


Figure 5.20: Small template and replicate clusters with three BPJ converging at one recurrent break position.

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

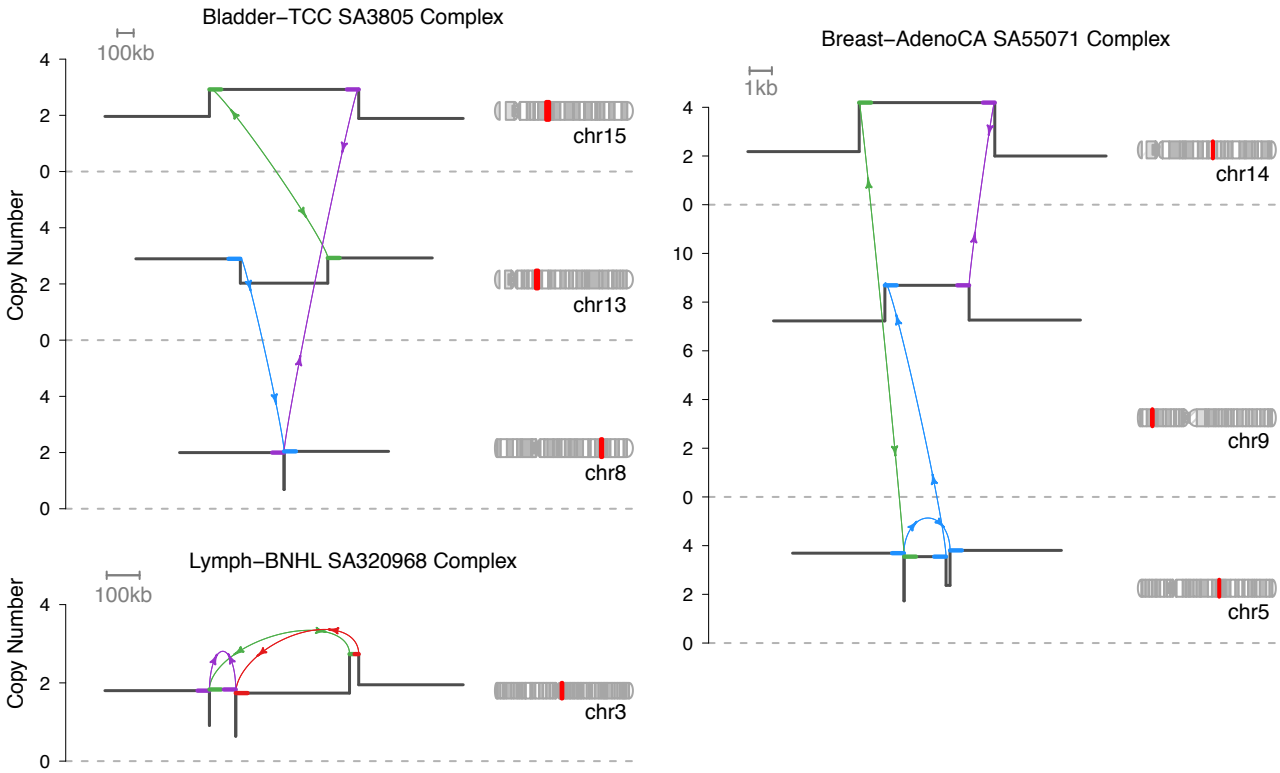


Figure 5.21: Combination SV clusters with hallmarks of both break and ligate *and* template and replicate repair mechanisms.

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

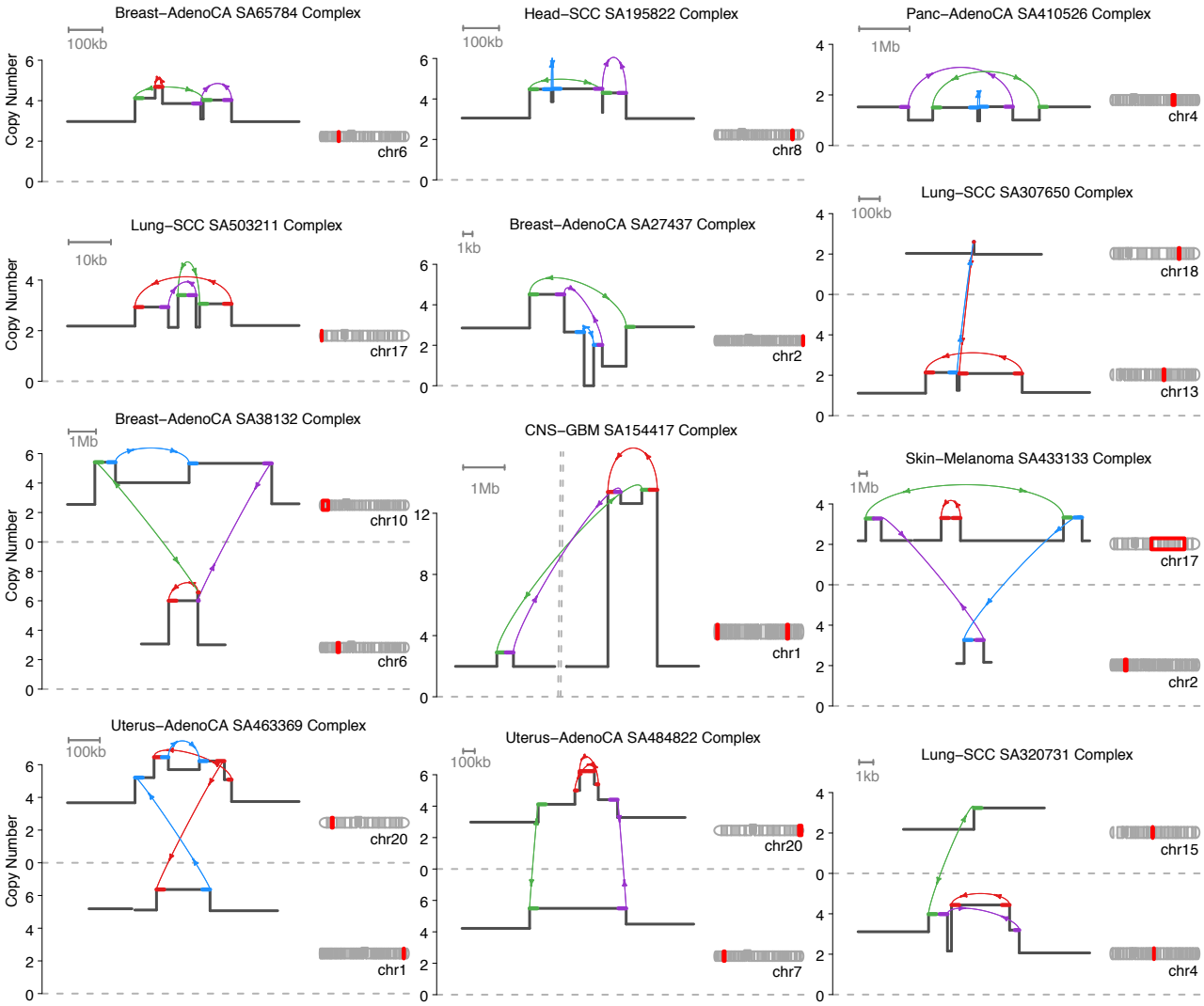


Figure 5.22: Overlapping or adjacent SV clusters

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

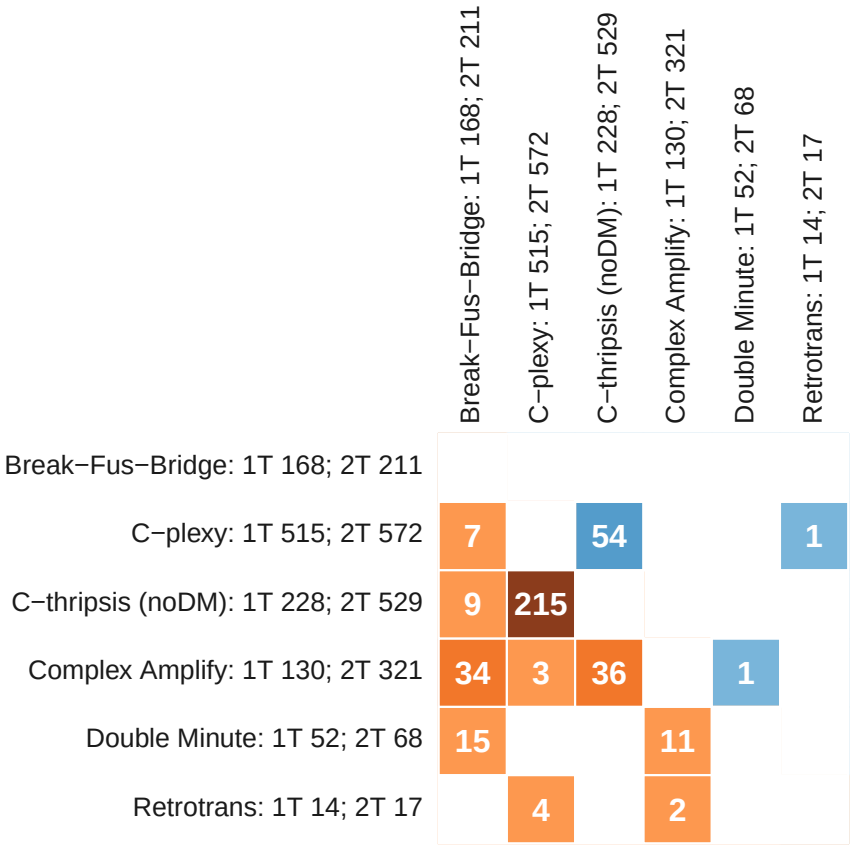


Figure 5.23: Overlap between the pilot classification groupings for the first tier (upper right, in blue) and second *or* first tier (lower left, in orange) complex SV events.

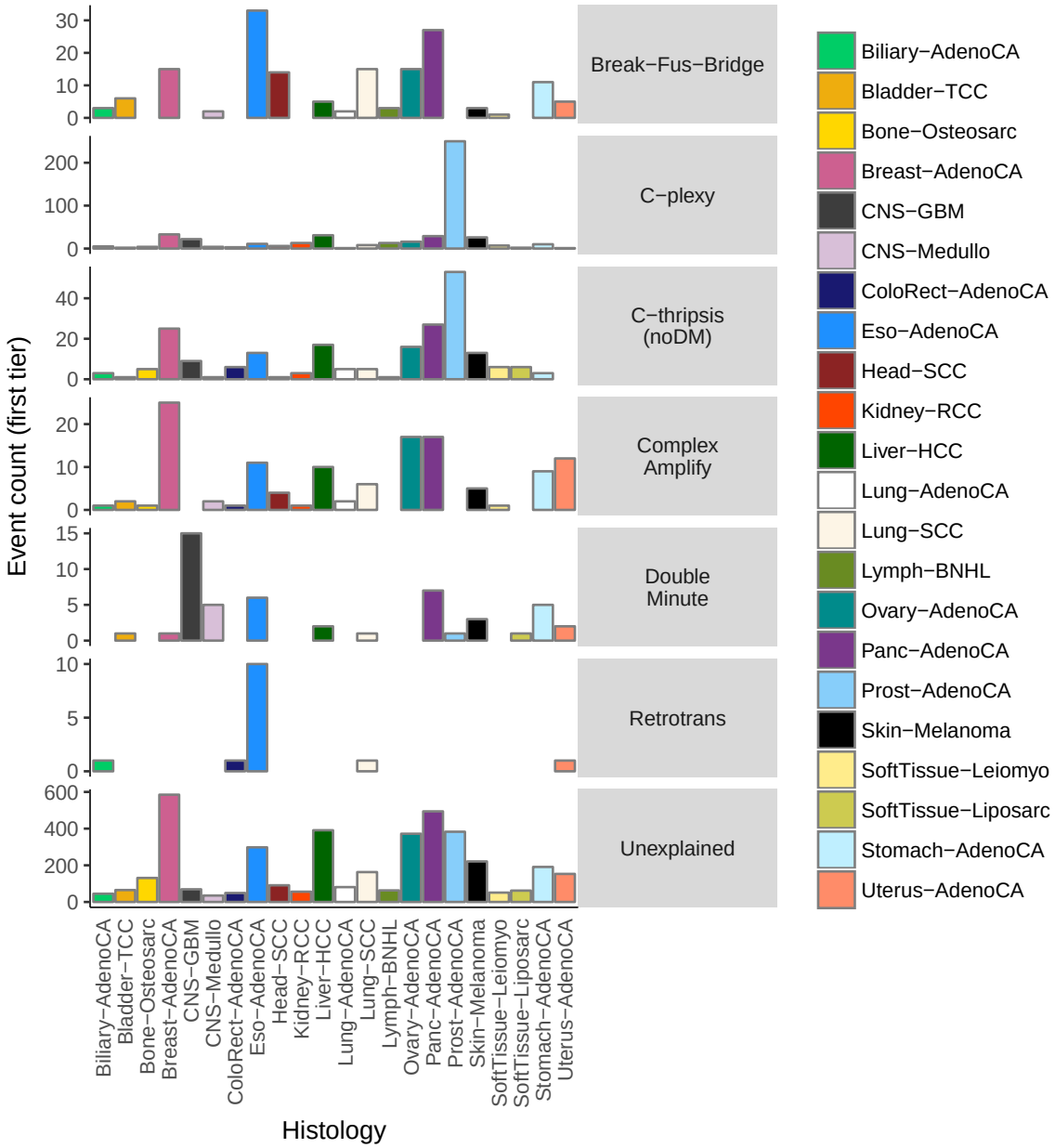


Figure 5.24: Histology distribution of first tier classifications for complex SV, without normalising by sample size or BPJ count.

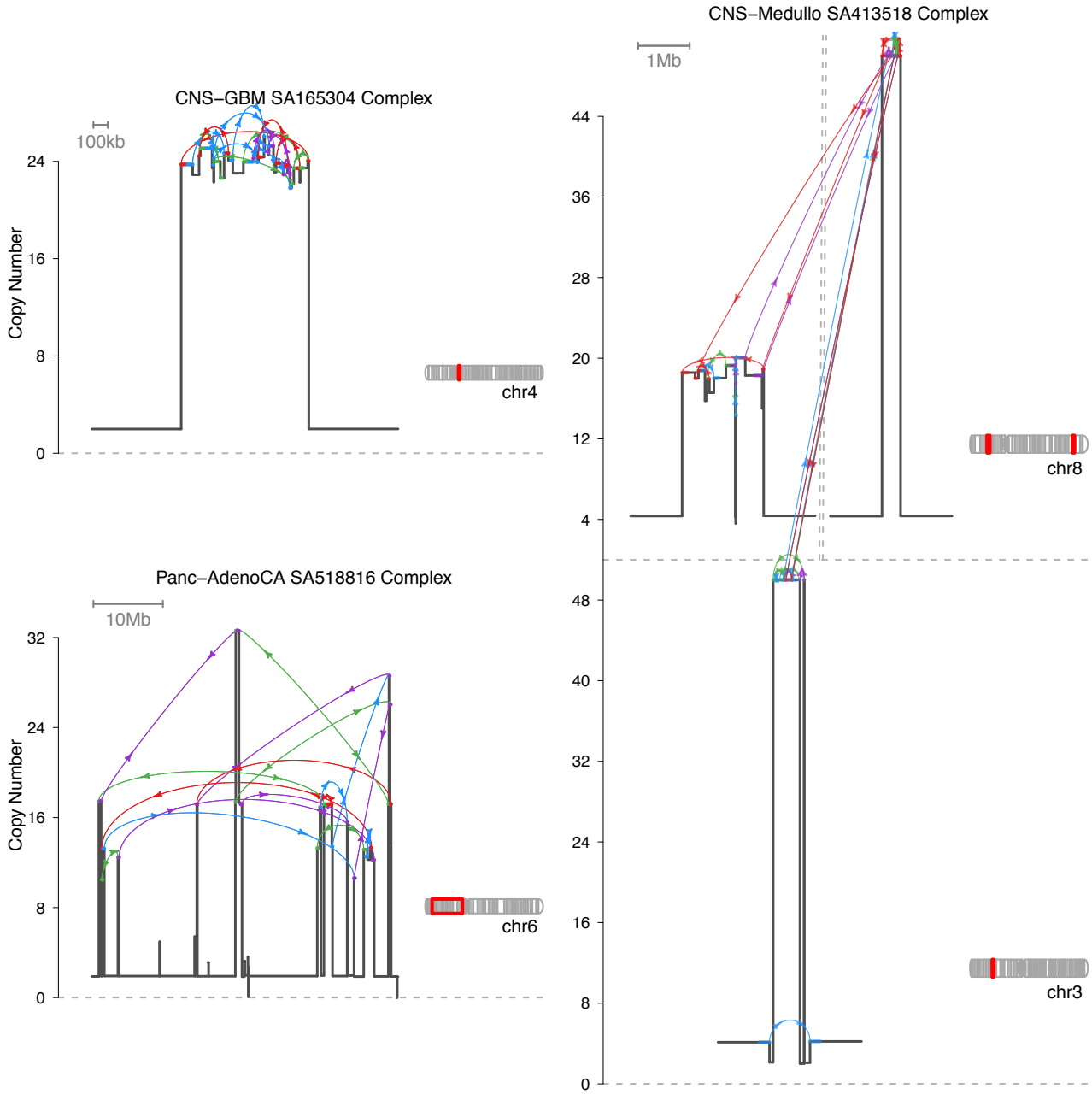


Figure 5.25: Example double minute events (first tier). The vertical CN scale is capped at 50.

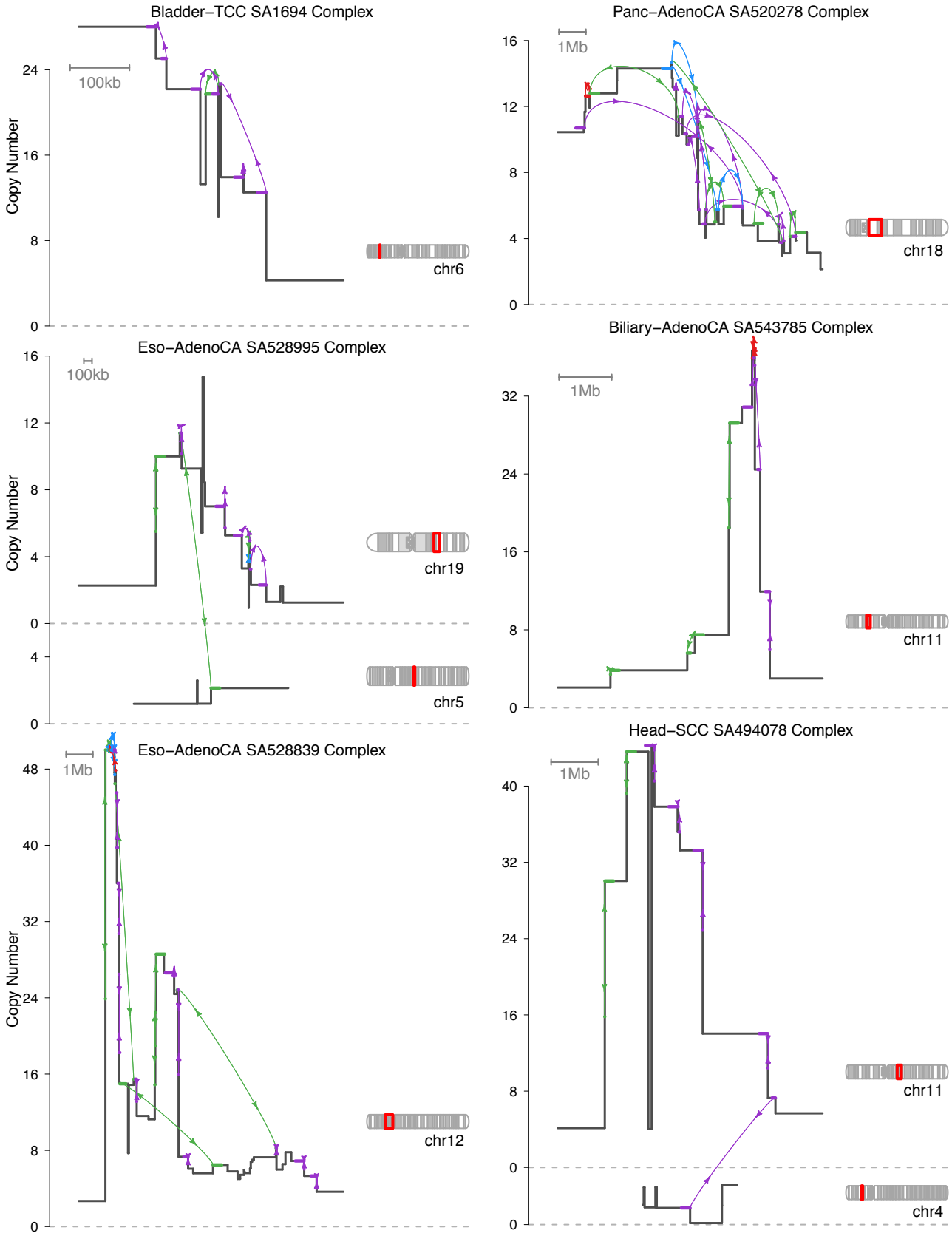


Figure 5.26: Example breakage-fusion-bridge events (first tier)

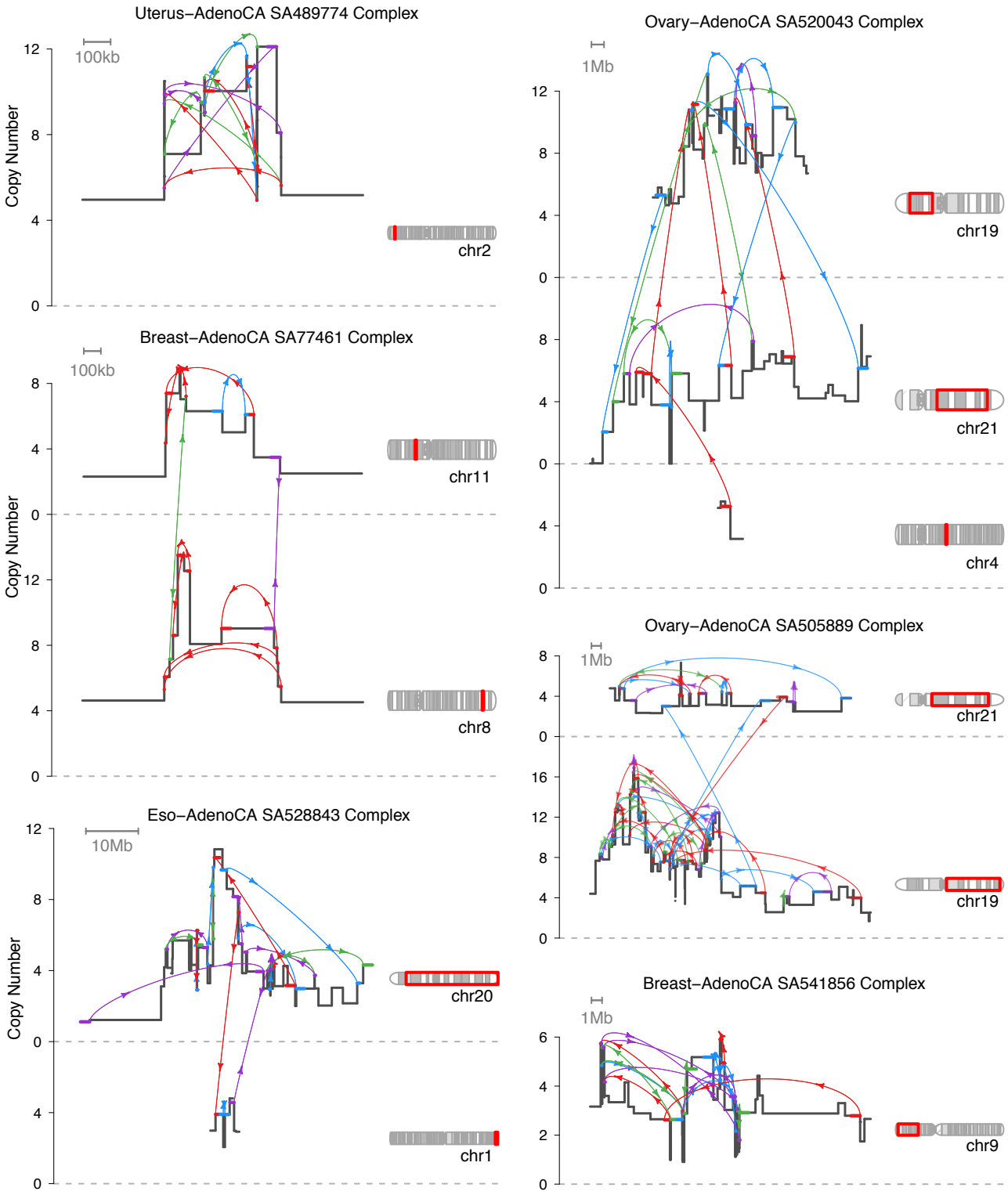


Figure 5.27: Example complex amplification events (first tier), possibly mediated by chromoanasynthesis.

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,

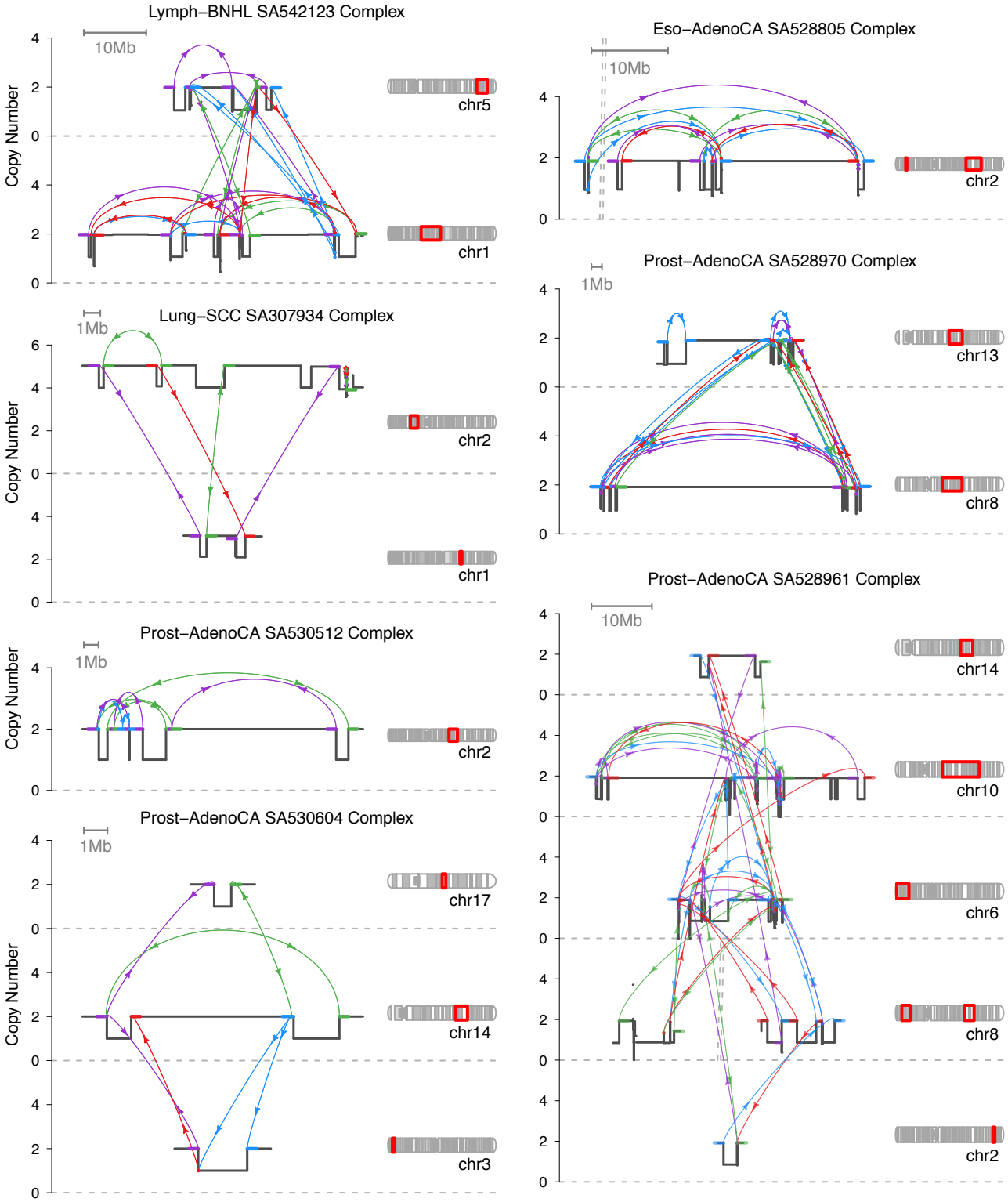


Figure 5.28: Example complex chromoplexy events (first tier)

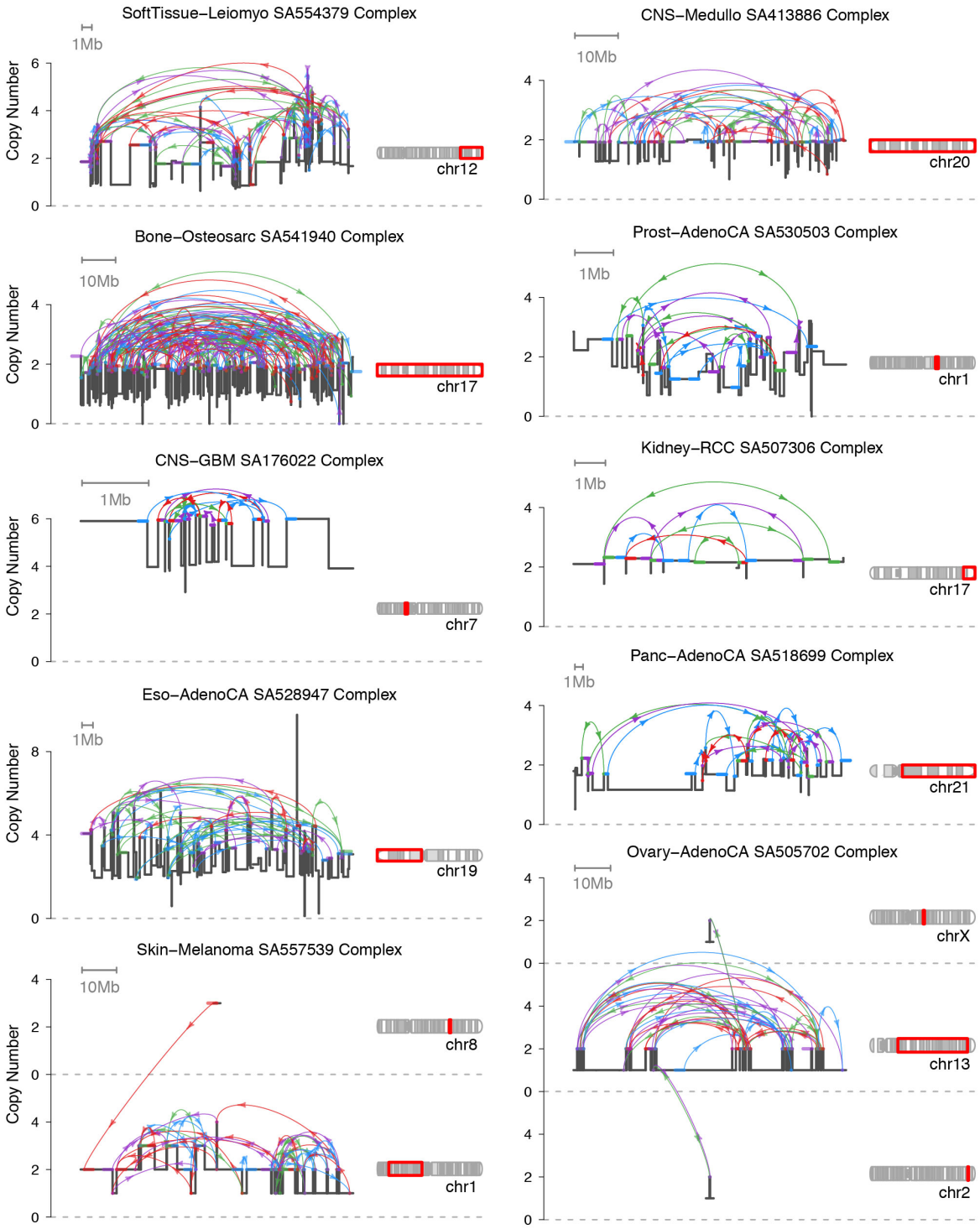


Figure 5.29: Example chromothripsis events (first tier)

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.

Chapter 6

Future perspectives

With the advent of high-throughput DNA sequencing technology, somatic alterations in cancer genomes are now identified at base-pair resolution in ever-expanding patient cohorts across a wide variety of histological subtypes. In contrast to the well-studied catalogues of single nucleotide variants, comprehensive studies of somatic rearrangement have lagged in development, impeded by the intrinsic complexity of their irregular and multifaceted structural forms. Consequently, the cancer genomics field lacks a robust and well-founded methodology for systematic SV specification, visualisation, and annotation.

The main aims of this thesis were to capitalise on a newly collated WGS dataset of somatic SV calls in 2559 cancer samples in order to: survey the diverse panorama of cancer rearrangement in different cell types; analyse signatures of SV form, location, and prevalence; and define a consistent framework for understanding and reporting genome rearrangement to advance the capabilities of future projects. Building on a recently developed classification scheme to identify the precise structure of individual breakpoint junctions and separate out complex clusters, I described the pan-cancer SV landscape of structural features (Chapter 2), genome property associations (Chapter 3), co-occurrence patterns (Chapter 4), and complex events (Chapter 5). To conclude, I highlight opportunities for further research and development with a focus on: algorithms and technology for SV detection (Section 6.1); the need for complete SV classification tools (Section 6.2); open questions regarding SV signature analysis (Section 6.3); and, finally, discovery and annotation of key functional consequences, with the ultimate goal of pinpointing relevant SV drivers of the cancer phenotype in a clinical setting (Section 6.4).

6.1 Identifying somatic genome rearrangement

In the PCAWG dataset used in this thesis, structural variants were identified by discordant and split paired-end sequences from the short-read Illumina Hi-Seq platform. Taking the intersection of SV calls from four different algorithms, the PCAWG consortium aimed to report high-confidence somatic events with congruent copy number support from read depth evidence. Without orthogonal technologies for SV detection, the specificity and sensitivity of this dataset is unknown. In future projects analysing cancer rearrangement, other biotechnology platforms may supplement or supersede the current Illumina pipelines to: validate SV calls with independent data; capture previously unmapped SVs in longer repeat regions; find variation in the 50 bp–1 kb range mostly overlooked by short-read sequencing; and phase BPJ to the same or different derivative chromosomes. New technologies with established benefits for SV detection (germline or somatic) include linked-read sequencing (Greer et al., 2017; Xia et al., 2017), long-read sequencing (Nattestad et al., 2017; Merker et al., 2018), optical mapping (Chan et al., 2017; Jaratlerdsiri et al., 2017), and Hi-C chromosome conformation assays (Harewood et al., 2017). A combination of approaches will yield the richest portrait of rearrangement (Chaisson et al., 2017), subject to comprehensive algorithmic development for integrating disparate lines of evidence within multi-platform datasets.

Although technological developments may eventually render short-read sequencing obsolete, the short to medium term prospects for SV analysis in large patient cohorts is still largely dominated by Illumina data in the legacy repositories of TCGA and ICGC, as well as ongoing sequencing projects by Genomics England (2017) and other initiatives. As such, there remains considerable value in continuing to improve variant calling pipelines for short-read WGS data. Ideally, SV events and CN segmentation would be jointly estimated by one inclusive algorithm aiming to uphold the logical expectation of higher CN states on the read-group side of every genuine breakpoint. The sub-clonal CN calls used in this thesis were obviously inaccurate in a sizeable minority of cases (Section 5.3), leaving an unmet demand for reliable estimation of non-integer CN states in complex and heterogeneous cancer cell populations. SV detection and CN estimation in the soma is further confounded by germline polymorphism. CN segmentation may benefit from explicit modelling of the germline SV states found in the matched normal sample, possibly using catalogues of common, population-matched SV inheritance (Sudmant et al., 2015) to better separate the germline events from the cancer-specific alterations.

6.2 Classifying breakpoint junctions

The major insights contributed by this thesis were facilitated by a novel BPJ classification scheme described in Sections 2.1.3 and 2.2.2. Where cancer rearrangement studies were previously limited to a handful of basic SV classes defined by BPJ orientation with one or two additional caveats, I was instead able to leverage twenty well-reasoned SV classes, including local 2-jump subtypes and long looping events of chromoplexy or templated insertion. All downstream investigation benefited from this detailed codification of individual breakpoints, empowering meaningful stratification within every SV property analysis to avoid massive confounding from heterogeneous phenomena. Notwithstanding this advancement, BPJ clustering and taxonomy remain deeply challenging tasks, with over half of all PCAWG breakpoints unexplained by the current system.

In the existing pipeline, BPJ are first clustered into groups with closer than expected proximity given sample-specific SV rates, and then adjacent breakpoints are partitioned into footprints labelled by their break orientation pattern. The final event classification depends on these footprint motifs and the connecting BPJ, shelving all cryptic configurations to a complex unexplained bin. The output depends heavily on the initial BPJ clustering, and it this clustering step which provides the first opportunity for improvement.

The limitations of the current approach (Section 2.1.3) include: the failure to account for BPJ interrelation in loops across multiple loci; the inability to separate distinct clusters connected by an unrelated BPJ; and the dependence on BPJ orientation frequencies oblivious to the broader structural context. I attempted to overcome some of these limitations with an alternative clustering method on the complex unexplained fraction using node-edge graph models (Section 5.1.1). However, in its present implementation, my graph method is also compromised by a reliance on arbitrary thresholds for node partitioning and component merging/separation. Ideally, the next generation of BPJ clustering methods will address these shortcomings in a formal probabilistic framework conditioning on the sample-specific SV composition. I propose that an iterative approach may offer the optimal solution—clustering and classifying by turns until the updates converge on a final stable solution. For example, if a sample has 250 foldback-type BPJ ($\langle ++ \rangle$ or $\langle -- \rangle$), and two such junctions fall within one or two megabases of each other, an initial cluster partition might separate these BPJ into independent events given the high overall rate of this junction class. However, if the subsequent classification step estimates that 220 of these

BPJ are actually explained by one chromothripsis event, the purported rate of isolated foldback would drastically reduce, causing the next cluster iteration to group the two inverting BPJ in one related event such as a dup-trp-dup or small BFB cluster (depending on the orientation). In an iterative framework, the clustering procedure could even account for the estimated location distribution of different event classifications, such as independent tandem duplications enriched in early replicating DNA, and independent deletions enriched in late replicating DNA (especially fragile sites). Sub-clonality provides another line of evidence informing cluster estimation (Cmero et al., 2017), assuming that high-confidence sub-clonal BPJ should only cluster with SV in the same approximate cell fraction. As 'third' (and 'fourth') generation sequencing becomes more ubiquitous, additional phasing information may greatly reduce the ambiguity of SV patterns along homologous chromosomes and/or in different cell fractions.

Given a particular partition of a sample's BPJ terrain, the next logical step is classification of the separated clusters, assuming they are generated by independent (or at least punctuated) rearrangement events. The current classification scheme (Section 2.1.3) is limited to isolated footprints in simple combinations, augmented with a library of possible overlaps to dissect a fraction of those convoluted clusters up to three or four BPJ. From my exploratory analysis of the complex unexplained SV in Chapter 5, I established that a different BPJ clustering scheme may recover additional SV events conforming to simple definitions, and that many more templated insertion and chromoplexy events would be recovered by extending the classification scheme to adjacent and overlapping footprint motifs. Furthermore, the current library of theoretical overlap structures does not account for templated insertion or local 2-jump events, and so upgrading this reference library may readily yield automatic classifications for another tranche of SV clusters.

These avenues for refining the current classification procedure are ultimately limited to small and relatively simple events, as larger clusters rapidly approach a unique parameter space that cannot possibly be afforded individual categories by specific BPJ configurations. At some point, SV classification strategies must transition from bottom-up to top-down, such that complex SV clusters are characterised by their overall feature profile, linking the total formation—where possible—to compatible underlying mechanisms such as chromothripsis, chromoplexy, chromoanasynthesis, and so forth. A top-down view is also more robust to false negative and/or false positive contamination; problems not accommodated by the simple SV classifier assuming complete BPJ information.

An outstanding question is how best to summarise the characteristic attributes of complex SV events in order to generate taxonomical divisions with proven correspondence to the underlying rearrangement mechanism. Thus, future research could consider: *de novo* event clustering using distance measures between independent SV; fixed classification rules trained on clear examples of canonical mechanisms; and computer simulations of genome rearrangement under a range of mechanistic models applied in varying combinations. In any case, the distinctions between different phenomena must be measured via a raft of summary statistics to capture relevant aspects of copy number, orientation, and connectivity. Experimental systems which generate complex SV events via known pathways of breakage and repair may provide additional validation and guidance in optimising these efforts (Meier et al., 2014; Maciejowski et al., 2015; Mardin et al., 2015; Zhang et al., 2015).

Given the importance of somatic rearrangement in cancer biology—and the role of similarly complex germline SVs in developmental disorders (Heesch et al., 2014; Collins et al., 2017)—complete SV specification tools are in high demand for research and clinical use, and must be regarded a major priority of bioinformatic development in the next few years.

6.3 Signatures of mutational process

As a valuable window into cancer aetiology and DNA dynamics, the mutational signatures imparted by different underlying processes are estimated by co-occurrence pattern matching across cancer sample cohorts. Throughout Chapter 4, I discussed my future proposals for extending the current signature paradigm, with particular attention to the hierarchical Dirichlet process. Here, I briefly highlight some open questions in relation to SV signatures in particular. In the abiding signature model, genome alteration classes are tallied as independent events in discrete, unordered categories. This may be a partially false premise in the structural variant realm, with events spanning a wide spectrum of size and complexity without neatly dividing into independent categories of comparable scale. It remains unclear how large, rare events like chromothripsis and chromoplexy should be compared against small, common SV like deletion and tandem duplication. Furthermore, the relevant features of size, microhomology, and replication timing skew, more naturally suit a signature framework of distributions over separate variables rather than discrete categorical observations. Regardless of the model, another frontier for signature

research is the integration of SNV, indel, and SV alteration classes within one overarching analysis, possibly using hierarchical models to share information across disparate data types reflecting a shared underlying condition such as HR deficiency or UV radiation.

6.4 Functional consequences of rearrangement

The investigations in this thesis focused mainly on the patterns and properties of somatic rearrangement, irrespective of their functional import as passenger or driver genome alterations. In this section, I discuss the prospects for annotation and selection analysis of functional consequences, as informed by the SV landscape surveyed throughout this work.

6.4.1 Annotation

As reviewed in Section 1.5, one rearrangement event may impart several gene-altering effects, including gene disruption or fusion across breakpoint junctions, gene dosage changes within the span of a SV footprint, and ectopic gene-enhancer regulation within merged or neo-TAD structures. At present, there are no available tools to annotate the full consequence spectrum of BPJ clusters in varying configurations.

For simple SV between two genome positions, it would be feasible to construct a complete atlas of gene-level consequences in the two-dimensional space of all possible events. Figure 6.1 outlines a proposed design for partitioning the space of all possible deletions or tandem duplications along a chromosome into functional consequences for one particular gene of interest. In theory, a simple rule set could construct a similar map for every gene, with any observed event easily annotated by position look-ups across the atlas of relevant maps. Although this construct may seem more convoluted than on-the-fly calculations for individually observed events, the annotation atlas has useful implications for recurrence-based driver analysis, as discussed in the following section.

For more complex SV events spanning multiple genome loci, it is impractical to calculate a full atlas of functional consequences for all possible structures. Instead, the individually observed BPJ and CN profiles could be parsed for likely fusions and dosage change, with ectopic enhancer contacts predicted from TAD boundary placement along likely derivatives. As adjacent BPJ may have

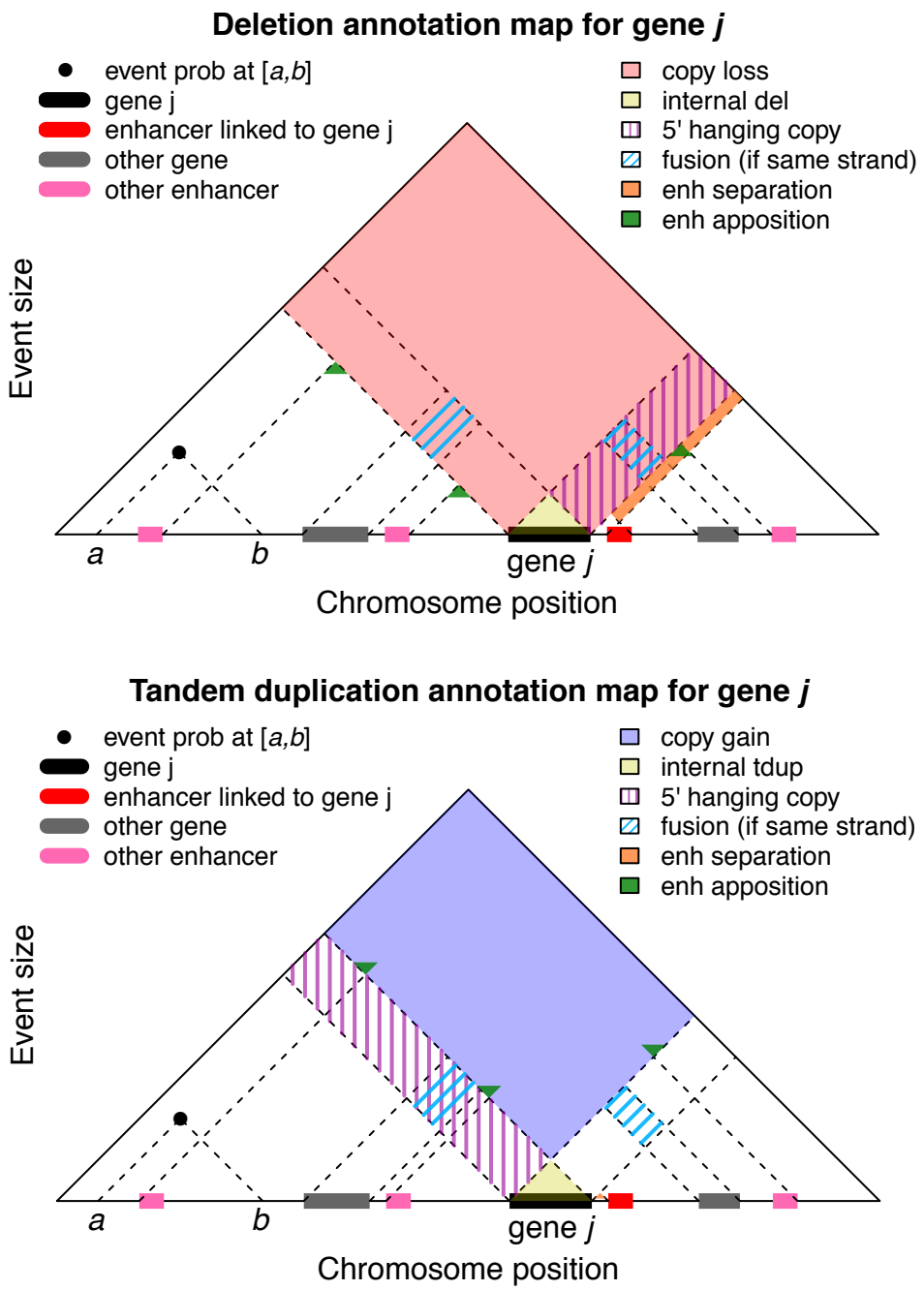


Figure 6.1: Schematic annotation maps of the functional consequences at gene j (assuming a + strand gene) imparted by all possible deletions (upper) or tandem duplications (lower) along the chromosome, with every point in the triangle representing a possible BPJ between two perpendicular points (like a and b).

uncertain phasing, some annotations (particularly enhancer apposition) may best be reported as probabilistic possibilities given the sample-specific likelihood of adjacent breakpoints occurring by chance on different chromosomes.

Existing knowledge banks of established cancer genes can be utilised to highlight putative driver SV events. For example, Notta et al. (2016) annotated pancreatic cancer rearrangements with simultaneous knockout of several canonical cancer genes. Aside from highlighting alterations to known oncogenes and tumour suppressors, it would also be beneficial to assess which of the many other functional consequences have relevance to cancer progression. Although hundreds of genes have already been labelled with known cancer effects, many more may yet be found, with one recent estimate suggesting half of all coding SNV drivers occur outside known cancer genes (Martincorena et al., 2017).

6.4.2 Driver discovery

So far, SV driver discovery efforts have focussed on: foci of recurrent copy gain or copy loss (Beroukhim et al., 2010; Mermel et al., 2011); enhancer-hijacking (Weischenfeldt et al., 2017); and one or two dimensional breakpoint recurrence, agnostic to BPJ classification (Wala et al., 2017a). As discussed in Chapter 3, different SV classes have markedly different formation rates across the genome, and, therefore, recurrence-based driver discovery should ideally account for structure-specific (and tissue-specific) background distribution estimates (before selection), in concert with sample-specific SV class exposures. Additionally, it would be preferable to integrate multiple effects—dosage, disruption, fusion, and regulation—to maximise available evidence for positive selection at the level of individual genes.

To this end, I return to the annotation map concept illustrated in Figure 6.1. For a given set of observed annotations, the question arises: which of these functional effects has occurred significantly more or less often than expected in the absence of selection? If we could determine the background probability of every possible SV event—that is, the probability at each point in the annotation map—then the expected rate of each annotated consequence before selection is the summation of event probabilities within the relevant partition. In this way, effects can be integrated across disparate SV classes while upholding the class-specific genome distributions and sample exposures to quantify the selection coefficients (neutral, positive, or negative) acting on functional up-regulation or inactivation for different genes. This approach is limited to simple SV classes—

such as translocation, foldback, reciprocal inversion, as well as deletion and tandem duplication shown in Figure 6.1—subject to appropriate estimation of the tissue-specific rearrangement rate at every position in the class-specific 2D annotation map.

To estimate the SV probability at every point (or pixelated square for reduced computation) in the 2D map (triangle for intra-chromosomal events; rectangle for inter-chromosomal events), recall that $\Pr(A \cap B) = \Pr(A) \Pr(B \mid A)$. In this context, the probability of a BPJ between positions (or pixels) A and B is the marginal breakpoint probability at A, multiplied by the conditional probability of a partner break at B. The first factor is easily obtained via class-specific logistic regression models explored in Section 3.3. The second factor is harder to obtain, and depends on the class (or signature) size distribution, sequence homology, physical proximity imposed by TAD structure and neighbouring chromosome territories, and the marginal breakpoint likelihood of B for this SV class. If this proves intractable to estimate, another possibility is to eschew 1D breakpoint likelihood models (such as logistic regression) in favour of 2D spatial point process models for the event locations directly observed within the space of possible junctions. For the spatial point process, the predictor variables at each 2D location could include size, homology, proximity (from Hi-C data), and a range of properties along the 1D genome that somehow require translation to the 2D junction space. In either scenario, properties such as chromatin state and gene expression should ideally be regarded as tissue-specific predictors. With a spatial point process, it may even be beneficial to regard tissue type as a third dimension, along which some tissue-agnostic properties are held constant, and the tissue-specific properties allowed to vary by identity pixels sorted by relatedness of tissue development and/or chromatin correlation.

One important caveat to using observed cancer variation datasets as the basis for background rearrangement rate models is the disproportionate bias towards positively-selected driver events, as previously discussed in Section 3.6. A more critical limitation is that background rate models do not readily extend to complex structures involving several genome loci in convoluted configurations. If the annotated consequences of templated insertion, chromoplexy, chromothripsis, and other structures, cannot be modelled as probabilistic distributions in the absence of selection, it is difficult to conceive how recurrence-based driver analysis will be possible without massive simplification. In the short to medium term, the prospects for driver discovery with complex SV may be limited to existing approaches on a reduced profile—such as copy number (Mermel et al.,

2011) or junction enrichment (Wala et al., 2017a)—*or* depend on functional assessment of related alterations to transcription, translation, and/or chromatin conformation, where complementary data (such as RNA-seq) are available to elaborate on the SV effect. Given the many challenges in interpreting SV structures and consequences, experimental validation of putative drivers is especially pertinent, perhaps using CRISPR technology to recreate specific rearrangement structures with a predicted functional consequence (Maddalo et al., 2014).

6.4.3 Clinical translation

Method development for somatic SV specification and annotation has important clinical ramifications (Macintyre et al., 2016b), with driver alterations *and* signatures of underlying repair deficiency illuminating diagnosis, prognosis, therapeutic opportunities, and the dynamics of ongoing genome instability which facilitates adaptation and acquired drug resistance.

6.5 Concluding remarks

Through errors of DNA repair, replication, and segregation, somatic genomes gradually diverge from their common ancestor in the zygote, occasionally evolving into cancerous cell populations with unregulated growth. Genome alterations at any scale may contribute to oncogenic transformation, with this thesis focussing on structural variation (typically larger than 1 kb) detected through whole genome sequencing of 2559 PCAWG samples. In addition to previously recognised SV phenomena involving isolated junctions of non-contiguous sequence or, at the other extreme, mass rearrangement under catastrophic stress, the PCAWG dataset revealed a vast intervening continuum of medium-complexity structure with hallmarks of both 'break and ligate' and 'template and replicate' repair modalities. By methodically surveying this panorama of SV structures and properties, the available repertoire of genetic manoeuvres is revealed with unprecedented breadth and resolution across dozens of common cancer types. The tissue and sample specificity of SV form, size, location, and complexity are testament to the many diverse rearrangement mechanisms driving somatic genomes towards pathological cancer phenotypes.

Appendix A

List of abbreviations

aCGH	array comparative genomic hybridisation
BFB	breakage fusion bridge
BIR	break-induced replication
bp	base pairs
BPJ	breakpoint junction
CFS	common fragile site
chr	chromosome
CN	copy number
CNA	copy number alteration
CNV	copy number variation
COSMIC	catalogue of somatic mutation in cancer
DNA	deoxyribose nucleic acid
DP	Dirichlet process
DSB	double-stranded break (in DNA)
FDR	false discovery rate
FS	fragile site
FWER	family-wise error rate
GAM	generalised additive model
GLM	generalised linear model
HDP	hierarchical dirichlet process
HR	homologous recombination
ICGC	international cancer genome consortium
IQR	inter-quartile range
kb	kilobase
LAD	lamina associated domain
LOH	loss of heterozygosity
LTR	long terminal repeat
Mb	megabase

MCMC	Markov chain Monte Carlo
MH	microhomology
MMBIR	microhomology-mediate break-induced replication
MMEJ	microhomology-mediated end-joining
MSI	microsatellite instability
NHEJ	non-homologous end-joining
NMF	non-negative matrix factorization
PCAWG	pan-cancer analysis of whole genomes
RNA	ribose nucleic acid
RPKM	reads per kilobase of transcript per million mapped reads
SNP	single nucleotide polymorphism
SNV	single nucleotide variant
SSA	single stranded annealing
SV	structural variation or structural variant
TAD	topologically associating domain
TCGA	the cancer genome atlas
TE	transposable element
TSS	transcription start site
WES	whole exome sequencing
WGD	whole genome duplication
WGS	whole genome sequencing

Appendix B

A description of the Hierarchical Dirichlet Process in the context of mutational signatures

In the following, I assume a mutational process is characterized by a discrete probability distribution over V mutation classes, hereafter termed its 'signature'.

Model description for one group of cancer samples

For each of N cancer samples, observe M_j total mutations across V mutation classes ($j = 1, \ldots, N$).

Let G_0 be a distribution over some countably infinite set of V-length probability vectors, describing the set of signatures found across the group of samples. G_0 is drawn from a Dirichlet process (DP) with prior H and concentration parameter γ_0, such that

$$\begin{aligned} \gamma_0 \mid \alpha_0, \beta_0 &\sim \text{Gamma}(\alpha_0, \beta_0)\,, \\ G_0 \mid \gamma_0, H &\sim \text{DP}(\gamma_0, H)\,. \end{aligned}$$

Let G_j be a distribution over the same set of probability vectors, describing the (sub)set of signatures found in sample j. G_j is drawn from a DP with prior G_0 and concentration parameter γ_j, such that

$$\begin{aligned} \gamma_j \mid \alpha_j, \beta_j &\sim \text{Gamma}(\alpha_j, \beta_j)\,, \\ G_j \mid \gamma_j, G_0 &\sim \text{DP}(\gamma_j, G_0) \quad \text{for } j = 1, \ldots, N\,. \end{aligned}$$

Define θ_{ji} to be the signature that causes the i-th mutation x_{ji} in cancer sample j. Each θ_{ji} is a probability vector over the V classes, and each x_{ji} is one categorical draw from that distribution, such that

$$\begin{aligned} \theta_{ji} \mid G_j &\sim G_j\,, \\ x_{ji} \mid \theta_{ji} &\sim \text{Categorical}(\theta_{ji}) \quad \text{for } i = 1, \ldots, M_j\,. \end{aligned}$$

Figure B.1 illustrates this HDP for one group.

Model description for multiple groups of cancer samples

Assume P groups of cancer samples with N_g samples in each group ($g = 1, \ldots, P$). For each cancer sample, observe M_{gj} mutations across V mutation classes ($j = 1, \ldots, N_g$). Let G_0 be defined as above.

Let G_g be a distribution over the set of probability vectors, describing the (sub)set of signatures found in group g. G_g is drawn from a DP with prior G_0 and concentration parameter γ_g, such that

$$\begin{aligned} \gamma_g \mid \alpha_g, \beta_g &\sim \text{Gamma}(\alpha_g, \beta_g)\,, \\ G_g \mid \gamma_g, G_0 &\sim \text{DP}(\gamma_g, G_0) \quad \text{for } g = 1, \ldots, P\,. \end{aligned}$$

Similarly, let G_{gj} be a distribution over probability vectors, describing the (sub)set of signatures found in cancer sample j from group g. G_{gj} is drawn from a DP with prior G_g and concentration parameter γ_{gj}, such that

$$\begin{aligned} \gamma_{gj} \mid \alpha_{gj}, \beta_{gj} &\sim \text{Gamma}(\alpha_{gj}, \beta_{gj})\,, \\ G_{gj} \mid \gamma_{gj}, G_g &\sim \text{DP}(\gamma_{gj}, G_g) \quad \text{for } j = 1, \ldots, N_g\,. \end{aligned}$$

Define θ_{gji} to be the signature that causes the i-th mutation x_{gji} in sample j from group g. Each θ_{gji} is a probability vector over the V classes, and each x_{gji} is one categorical draw from that distribution, such that

$$\begin{aligned} \theta_{gji} \mid G_{gj} &\sim G_{gj}\,, \\ x_{gji} \mid \theta_{gji} &\sim \text{Categoical}(\theta_{gji}) \quad \text{for } i = 1, \ldots, M_{gj}\,. \end{aligned}$$

Figure B.2 illustrates this HDP for $P = 2$

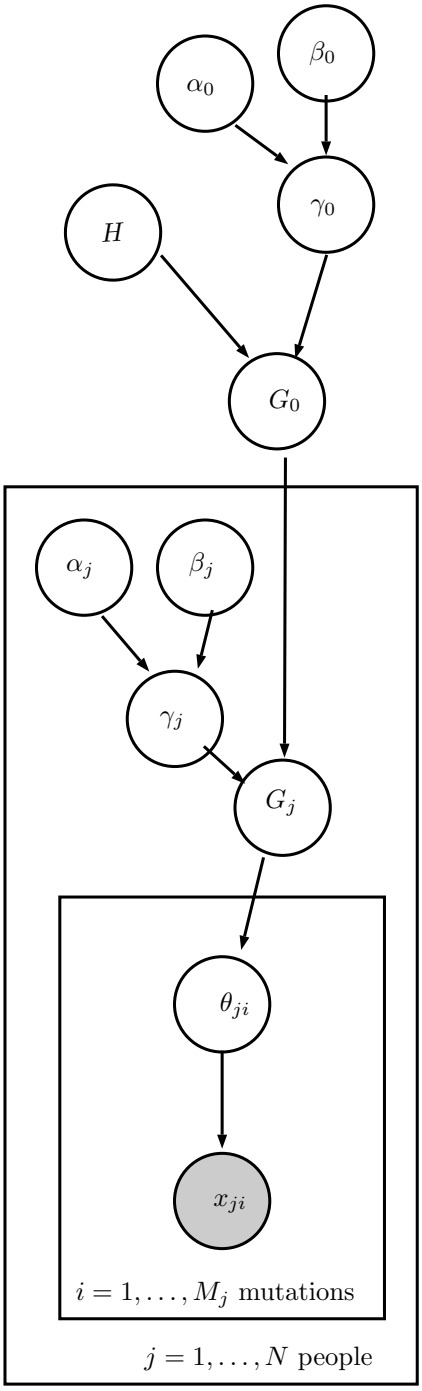


Figure B.1: The hierarchical Dirichlet process mixture model for one group of cancer samples.

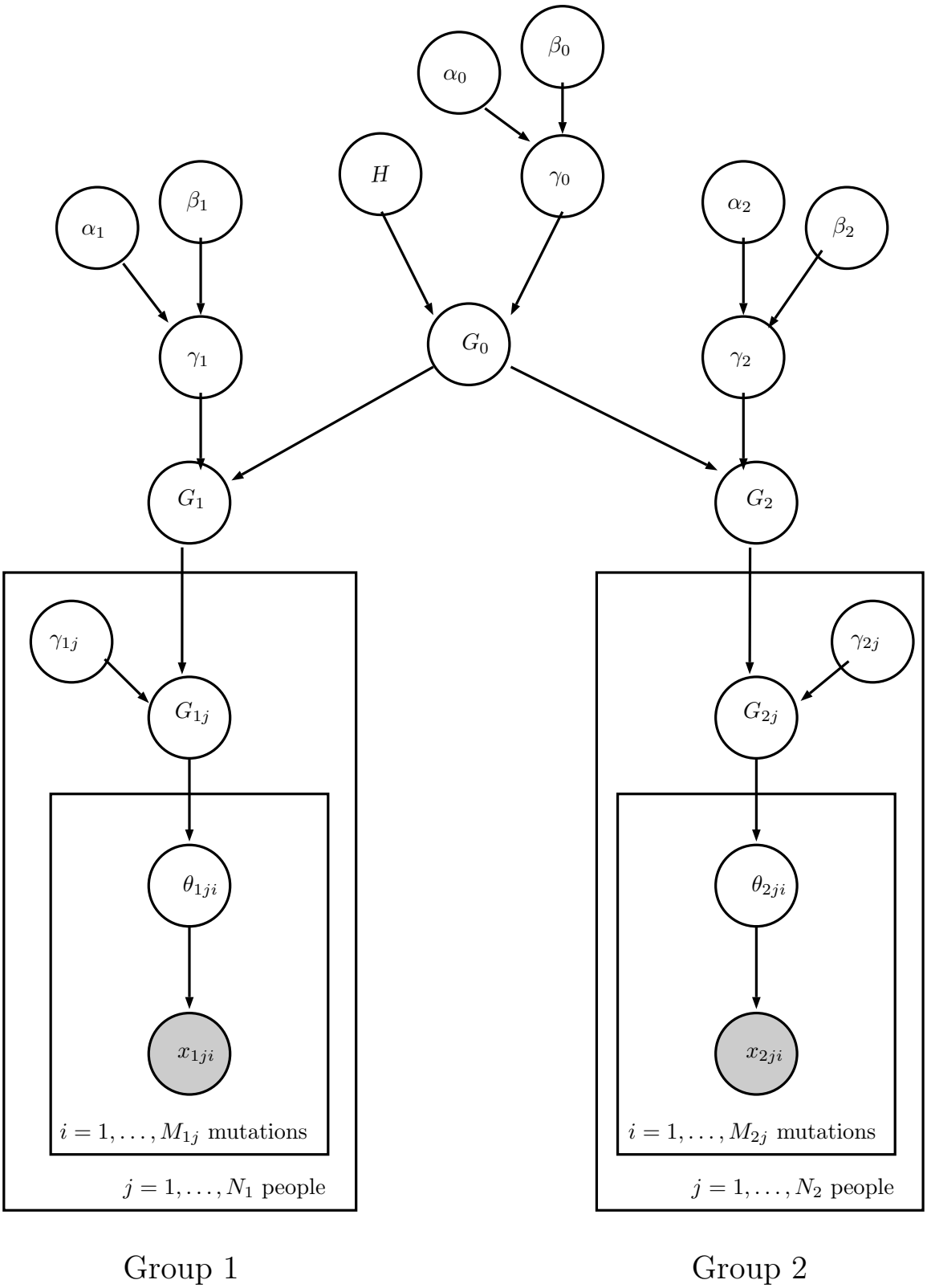


Figure B.2: The hierarchical Dirichlet process mixture model for two groups. Gamma priors for γ_{1j} and γ_{2j} not shown for convenience.

Posterior sampling in the Chinese Restaurant Franchise

For any such HDP, we observe the values of x (the mutations) and specify the prior distribution H and the hyperparameters α and β, but must estimate all other variables to make inference. Teh et al. (2006) described Gibbs sampling schemes in the general case for any data distribution. Here, I derive the equations of the 'Chinese Restaurant Franchise' Gibbs sampling scheme for categorical data in the context of mutational process signatures. This scheme fits the 'one group' HDP as shown in Figure B.1.

Assume a franchise of restaurants (cancer samples), each containing an unlimited number of tables. Each table is associated with one dish (mutational process), characterised by a probability distribution over V categories (mutation classes). Customers (mutations) are assigned to tables within the restaurant (sample), and take values from the probability distribution assigned to that table. Note that more than one table in the restaurant (sample) can be generating customer values (mutations) from the same dish (mutational process/signature).

Let t_{ji} be the index of the table in sample j that mutation i belongs to. Let $k_{jt_{ij}}$ be the index of the mutational process at the table in sample j with the i-th mutation. Let n_{jtkc} be the number of mutations in sample j at table t assigned to process k equal to class c. Let m_{jk} be the number of tables in sample j associated with process k. Any of these variables can be summed over (denoted with a bullet) to represent the marginal counts for n or m.

Let H be a Dirichlet distribution with concentration parameters $\boldsymbol{\tau}$. Each mutational signature ϕ is a draw from H, such that

$$\begin{aligned} \phi &\sim H(\boldsymbol{\tau}), \\ h(\phi \mid \boldsymbol{\tau}) &= \frac{1}{B(\boldsymbol{\tau})} \prod_{v=1}^{V} \phi_v^{\tau_v - 1}. \end{aligned}$$

The probability of mutation x_{ji} being equal to its observed mutation class c, given that x_{ji} originates from a particular process k, given all other mutations currently assigned to process k and integrating over all possible values for the signature ϕ_k, is

$$p_k^{-x_{ji}}(x_{ji} = c) = \frac{n_{..kc}^{-x_{ji}} + \tau_c}{n_{..k\cdot}^{-x_{ji}} + \sum_{v=1}^{V} \tau_v}.$$

where $n_{..kc}^{-x_{ji}}$ is the number of mutations (across all samples and tables) currently assigned to signature k and class c (excluding x_{ji}), $n_{..k.}^{-x_{ji}}$ is the total number of mutations (across all samples, tables and classes) currently assigned to signature k (excluding x_{ji}), and τ_c is the concentration parameter for class c from the prior H (like a pseudocount).

The probability of mutation x_{ji} being equal to its observed mutation class c, given that x_{ji} originates from a new process k^{new}, integrating over all possible values for the signature ϕ_k, is

$$p_{k^{\text{new}}}^{-x_{ji}}(x_{ji} = c) \quad = \quad \frac{\tau_c}{\sum_{i=1}^{V} \tau_i} \, .$$

The probability of mutation x_{ji} being equal to its observed mutation class c, given that x_{ji} belongs to a new table t^{new} in sample j, given all other table assignments in all other samples and given the current set of mutational processes, is

$$p(x_{ji} = c \mid \boldsymbol{t}^{-ji}, t_{ji} = t^{\text{new}}, \boldsymbol{k}) = \sum_{k=1}^{K} \frac{m_{.k}}{m_{..} + \gamma_0} p_k^{-x_{ji}}(x_{ji} = c) + \frac{\gamma_0}{m_{..} + \gamma_0} p_{k^{\text{new}}}^{-x_{ji}}(x_{ji} = c)$$

where $m_{.k}$ is the number of tables associated with process k (across all samples), $m_{..}$ is the total number of tables across all samples, and γ_0 is the concentration parameter of the Dirichlet process prior for G_0.

The probability of the set of table t mutations $\boldsymbol{x}_{jt}$ given they originate from a particular process k, given all other mutations currently assigned to process k and integrating over all possible values for the signature ϕ_k, is

$$p_k^{-\boldsymbol{x}_{jt}}(\boldsymbol{x}_{jt}) = \frac{\Gamma(n_{..k.}^{-\boldsymbol{x}_{jt}} + \sum_{v=1}^{V} \tau_v)}{\Gamma(n_{jtk.} + n_{..k.}^{-\boldsymbol{x}_{jt}} + \sum_{v=1}^{V} \tau_v)} \prod_{v=1}^{V} \frac{\Gamma(n_{jtkv} + n_{..kv}^{-\boldsymbol{x}_{jt}} + \tau_v)}{\Gamma(n_{..kv}^{-\boldsymbol{x}_{jt}} + \tau_v)}$$

where $n_{..k.}^{-\boldsymbol{x}_{jt}}$ is the number of mutations (across all samples, tables and classes) assigned to process k (excluding the mutations at table t in sample j), $n_{jtk.}$ is the number of mutations (across all classes) at table t in sample j, n_{jtkv} is the number of mutations in sample j at table t equal to class v, and $n_{..kv}^{-\boldsymbol{x}_{jt}}$ is the number of mutations (across all samples, tables) assigned to process k and class v (excluding the mutations at table t in sample j).

The probability of the set of table t mutations $\boldsymbol{x}_{jt}$ given they originate from a

new process k^{new}, integrating over all possible values for the signature ϕ_k, is

$$p_{k^{\text{new}}}^{-\boldsymbol{x}_{jt}}(\boldsymbol{x}_{jt}) = \frac{\Gamma(\sum_{v=1}^{V} \tau_v)}{\Gamma(n_{jtk\cdot} + \sum_{v=1}^{V} \tau_v)} \prod_{v=1}^{V} \frac{\Gamma(n_{jtkv} + \tau_v)}{\Gamma(\tau_v)} .$$

Gibbs sampling scheme

The sampling scheme is initialised with some total number of mutational processes (K) and some number of tables in each cancer sample ($m_{j\cdot}$ for $j = 1, \ldots, N$). Each table is assigned a mutational process (initialise each k_{jt} - the index of the process associated with table t in sample j) and each mutation is assigned to a particular table (initialise each t_{ji} - the index of the table in sample j with mutation i).

Iterate steps:

1. For each mutation, sample a new value for t_{ji}.
2. For each table, sample a new value for k_{jt}.
3. For each concentration parameter, sample a new value given the current cluster allocations.

After removing the burn-in period, the Gibbs sampling scheme thus generates a posterior sample to estimate K, and all $m_{j\cdot}$, t_{ji} and k_{jt}.

Sampling $\boldsymbol{t}$

The probability that the i-th mutation in sample j belongs to a particular table t, given all other table assignments in all other samples and given the current set of mutational processes, is:

$$\Pr(t_{ji} = t \mid \boldsymbol{t}^{-ji}, \boldsymbol{k}) \propto \begin{cases} n_{jt\cdot\cdot}^{-ji} p_{k_{jt}}^{-x_{ji}}(x_{ji}) & \text{if } t \text{ previously used}, \\ \alpha_0 p(x_{ji} \mid \boldsymbol{t}^{-ji}, t_{ji} = t^{\text{new}}, \boldsymbol{k}) & \text{if } t = t^{\text{new}}, \end{cases}$$

where n_{jt}^{-ji} is the number of mutations in sample j already at table t (excluding x_{ji}), and α_0 is the concentration parameter for the Dirichlet process prior on G_j (the distribution of mutational signatures in sample j).

If the sampled value of t_{ji} is t^{new}, then a mutational process must be assigned to the new table by sampling a value for $k_{jt^{\text{new}}}$ from

$$\Pr(k_{jt^{\text{new}}} = k \mid \boldsymbol{t}, \boldsymbol{k}^{jt^{\text{new}}}) \propto \begin{cases} m_{\cdot k} p_k^{-x_{ji}}(x_{ji}) & \text{if } k \text{ previously used}\,, \\ \gamma p_{k^{\text{new}}}^{-x_{ji}}(x_{ji}) & \text{if } k = k^{\text{new}}\,. \end{cases}$$

Sampling $\boldsymbol{k}$

Changing the mutational process assigned to a particular table (updating k_{jt}) changes the mutational process assigned to all mutations at that table. Therefore

$$\Pr(k_{jt} = k \mid \boldsymbol{t}, \boldsymbol{k}^{-jt}) \propto \begin{cases} m_{\cdot k}^{-jt} p_k^{-\boldsymbol{x}_{jt}}(\boldsymbol{x}_{jt}) & \text{if } k \text{ previously used}\,, \\ \gamma p_{k^{\text{new}}}^{-\boldsymbol{x}_{jt}}(\boldsymbol{x}_{jt}) & \text{if } k = k^{\text{new}}\,. \end{cases}$$

Sampling concentration parameters

See Appendix in Teh et al. (2006).

Appendix C

Heuristic classification rules for complex SV

In the following list of pilot classification criteria, second tier thresholds are given in parentheses following the first tier threshold values. The estimated specificity (by manual curation) of the first tier preliminary classification is reported in Table 5.3.

The heuristics for breakage-fusion-bridge are:

- the proportion of breaks on one chromosome is at least 0.75 (0.65);
- the proportion of intra-chromosomal BPJ with foldback-type orientations is greater than or equal to 0.7 (0.6), and also outnumbers the frequency of inter-chromosomal BPJ; and
- in the footprint with the highest number of breaks, the flanking CN states differ by more than 6 (4).

The heuristics for retrotransposition hotspots are:

- at least 4 (3) involved chromosomes;
- one (and only one) footprint contains 6 or more breaks, and there are at least four other footprints containing no more than 2 (3) breaks;
- the footprint with the most breaks spans less than 100 kb (1 Mb); and
- the proportion of inter-chromosomal BPJ at the footprint with the most breaks is at least 0.6 (0.5).

The heuristics for isolated double minutes (with no chromothripsis) are:

- at least 75% (70%) of breaks have one CN side higher than 12 (9);
- the CN at least one side of the footprint with the most breaks is less than 6 (8);

- no more than 3 (6) chromosomes have three or more breaks;
- at least two copy jumps are larger than 10 (6);
- at least two copy jumps larger than 6 are at least 50 kb (10 kb) apart; and
- on the chromosome with the largest copy area, less than 60% (70%) of the junctions are foldback-type.

The heuristics for a diverse range of complex graduated amplifications possibly involving chromoanasynthesis mechanisms are:

- at least one footprint has 10 (8) breakpoints;
- the proportion of intra-chromosomal BPJ with foldback-type orientations is less than or equal to 0.6 (0.67);
- at least one chromosome has over 95% of its involved CN profile spread over at least 4 (3) rough CN states above 2 (using integer rounding);
- every footprint with 5 or more breakpoints has an internal CN average 1 (0.5) copy higher than at least one flanking side; and
- the 0.9 quantile of absolute CN jump magnitude is less than 4 (6).

The heuristics for chromoplexy are:

- no footprint has more than 50 (75) breaks;
- no chromosome contains more than 8 (12) separate footprints;
- at least 35% (30%) of the inter-break motifs are [+−] gaps smaller than 1 Mb (3 Mb);
- the geometric mean [+−] gap motif is less than 0.5 (1.0) times the geometric mean [−+] motif (disregarded if both are <10 kb);
- at least 50% (33%) of footprints start with a + break orientation and end with a − break orientation;
- no one orientation type contributes more than 50% (60%) of all intra-chromosomal junctions;
- all copy jumps are smaller than 3 (5);
- every chromosome has over 85% (75%) of its involved CN profile spread over at most 2 rough CN states (using integer rounding);
- every footprint with 5 or more breakpoints has an internal CN average not more than 0.75 (1.1) copies higher than either flanking side;
- any chromosome with more than 15 breaks should have a non-uniform break distribution, with a Kolmogorov-Smirnov test significant at 0.05 (0.1); and
- if the event is restricted to one chromosome, it must span more than 100 kb.

The heuristics for chromothripsis (with no double minute amplification) are:

- at least one chromosome has 15 (10) breakpoints;
- the proportion of breaks in footprints containing three or fewer breaks is less than or equal to 0.25 (0.4);
- every intra-chromosomal BPJ orientation is observed at least once, with no one orientation type contributing more than 0.45 (0.55) of all intra-chromosomal junctions;
- the median span of intra-chromosomal junction types varies by less than 50-fold (500-fold) across the four possible BPJ orientations;
- inter-break motifs are at least 0.33 (0.25) [+−] and 0.33 (0.25) [−+];
- the 0.95 quantile of absolute CN jump magnitude is smaller than 3 (4) times the median CN jump, up to a maximum of 4 (6);
- every chromosome has over 85% (75%) of its involved CN profile spread over at most 3 rough CN states (using integer rounding);
- every footprint with 5 or more breakpoints has an internal CN average not more than 1 (2) copies higher than either flanking side;
- if there is more than one footprint, at least one footprint is larger than 500 kb (100 kb); and,
- *to attempt differentiation from chromoplexy*, if there are four or more breaks on two different chromosomes:
  - the median size of a [+−] gap motif is not smaller than 1 kb if the median size of a [−+] retained motif is larger than 10 kb;
  - a Kolmogorov-Smirnov test for uniform breakpoint positioning in footprints with 12 or more breaks is non-significant at a 10^{-3} (10^{-6}) threshold, *or* has a test statistic smaller than 0.25.

Appendix D

Supplementary Figures

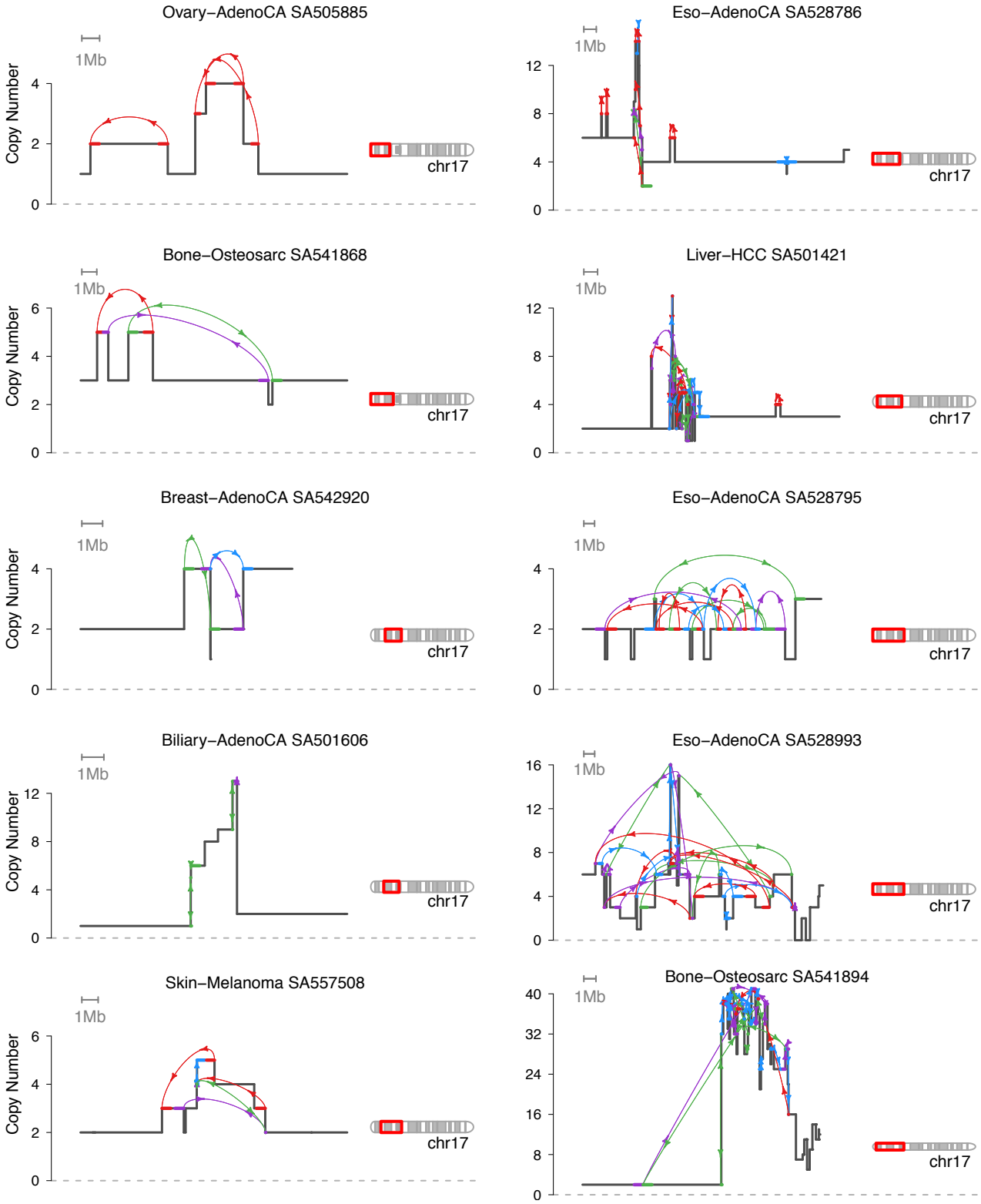


Figure D.1: All intrachromosomal BPJ on the p-arm of chromosome 17 in ten different samples, coloured by orientation. Blue denotes deletion type $\langle +- \rangle$, red is tandem duplication type $\langle -+ \rangle$, and purple and green indicate inversion type $\langle ++ \rangle$ or $\langle -- \rangle$.

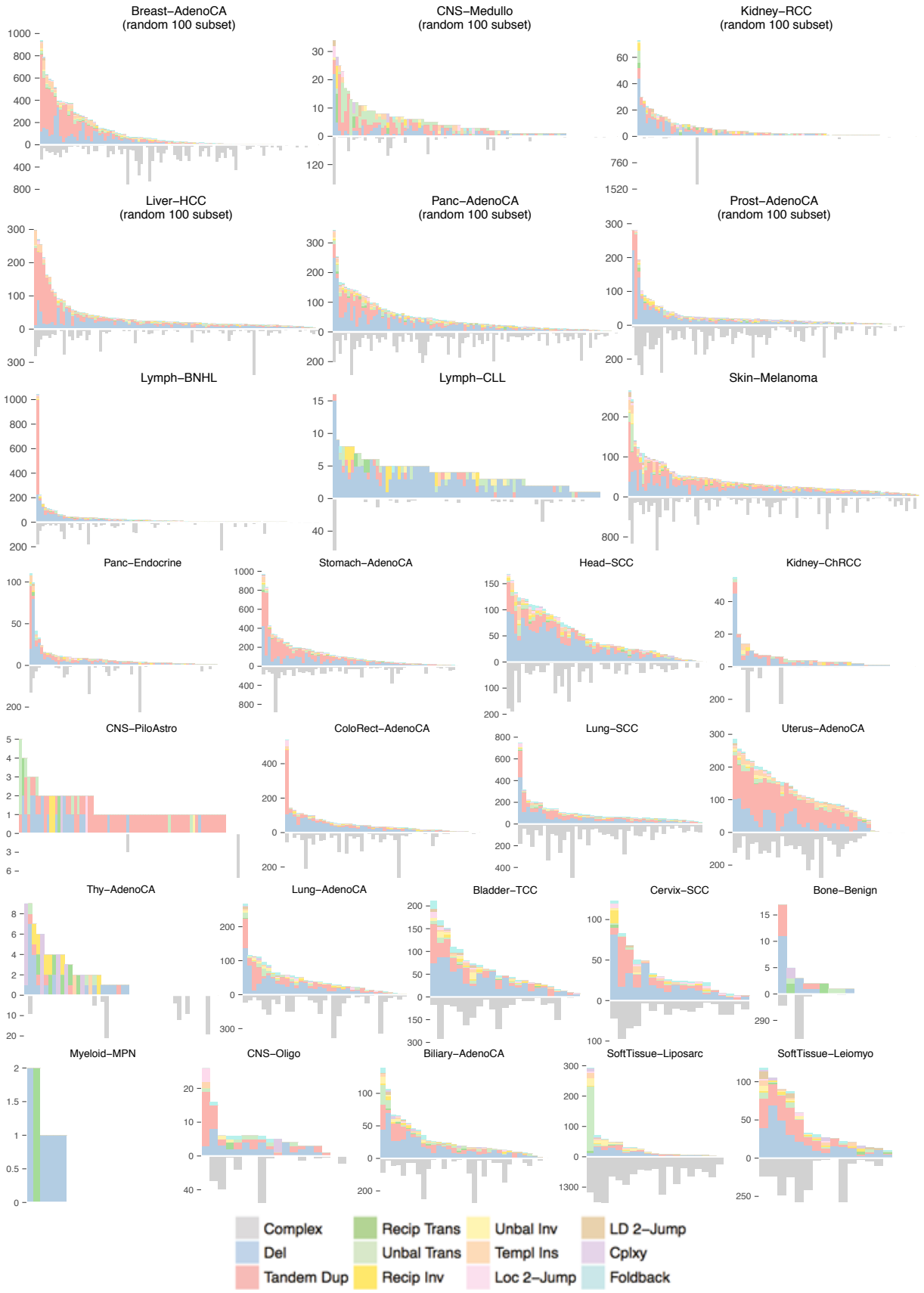


Figure D.2: Per-sample counts of complex (lower) and classified (upper) breakpoint junctions.

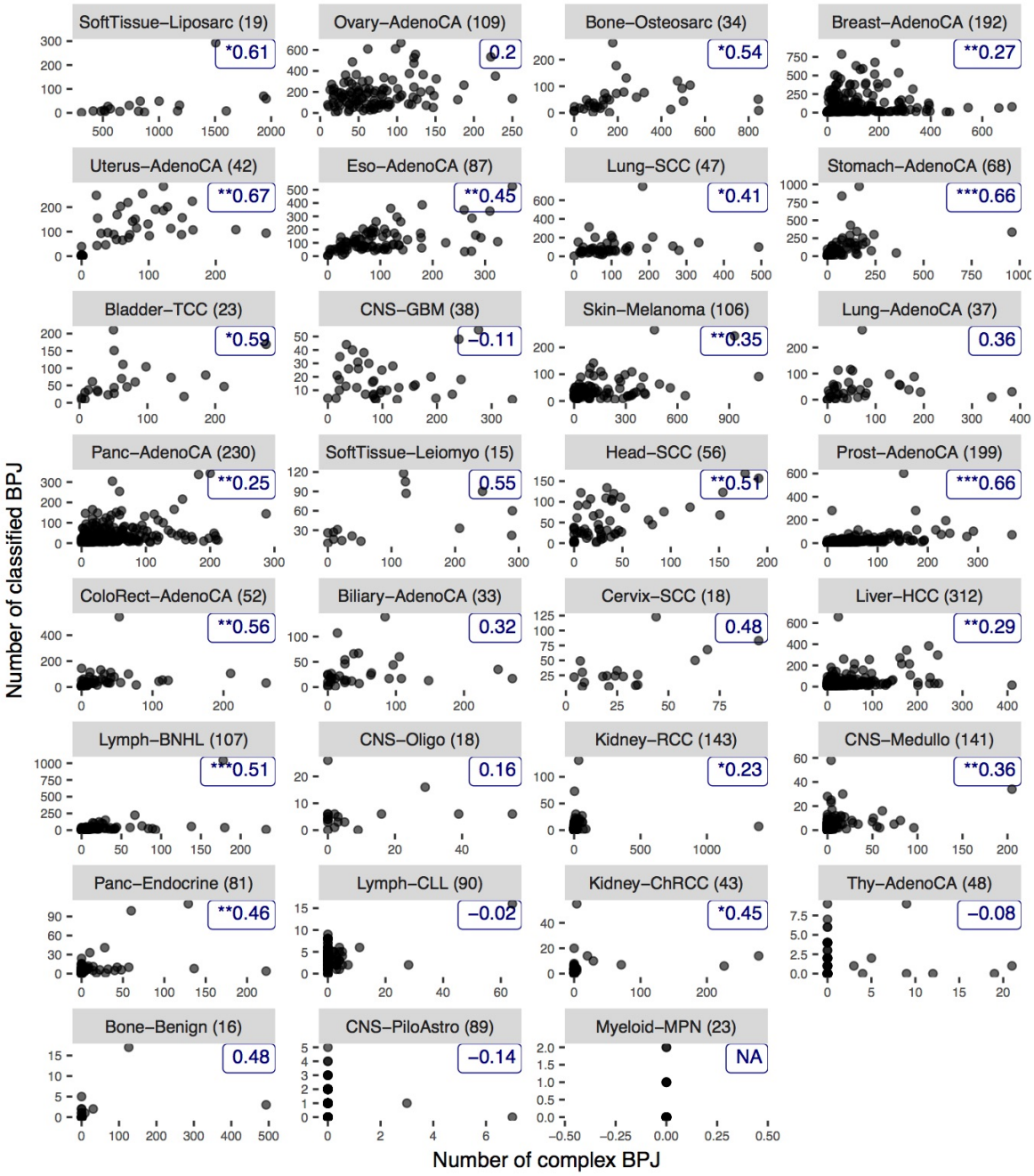


Figure D.3: Spearman's rank correlation coefficient between complex (horizontal) and classified (vertical) BPJ counts in samples grouped by histology. Benjamini–Hochberg-corrected FDR for the null hypothesis of zero correlation is indicated at levels: * < 0.01, ** 0.001, and *** $< 10^{-6}$.

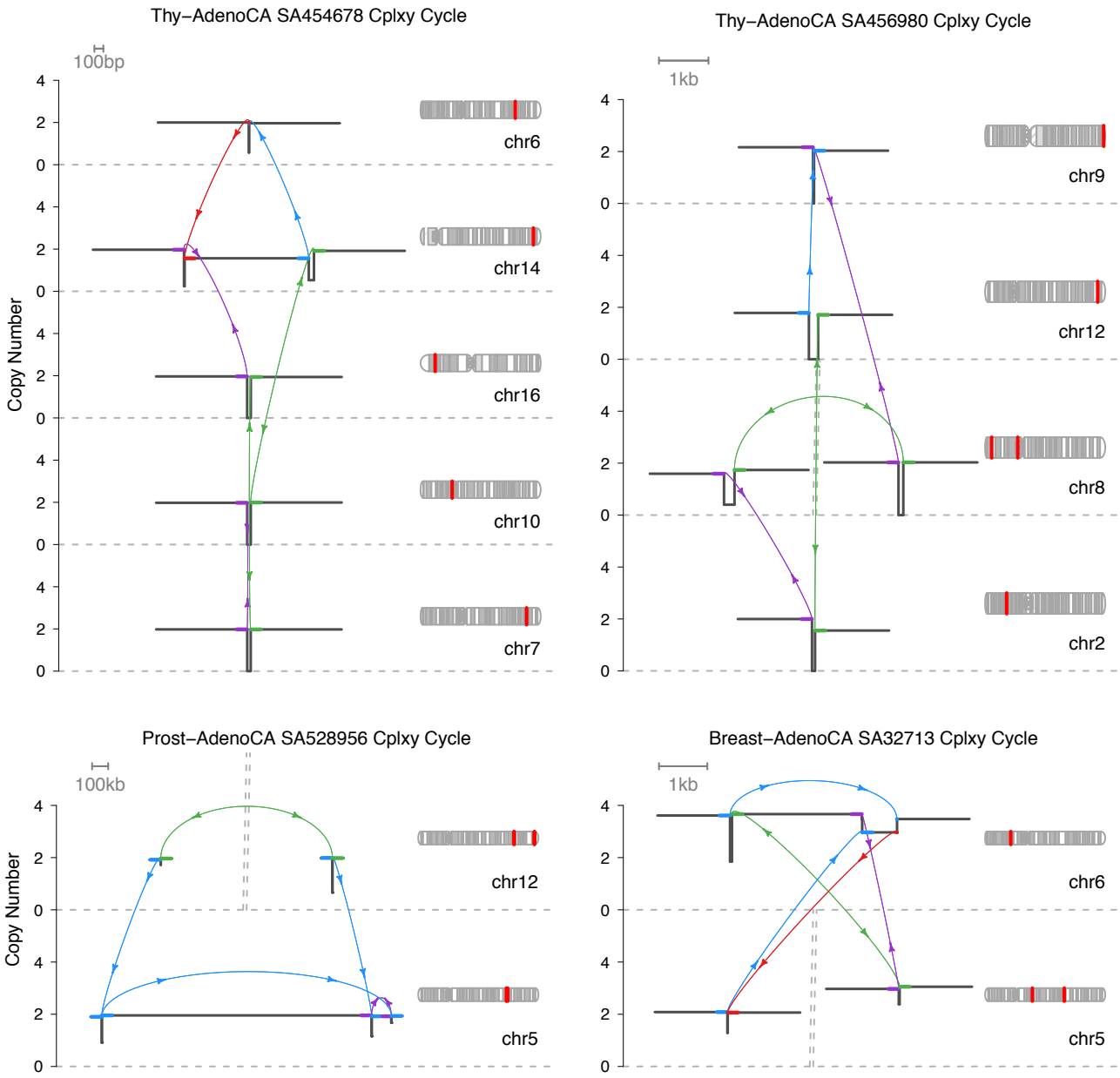


Figure D.4: Four of the longest chromoplexy events

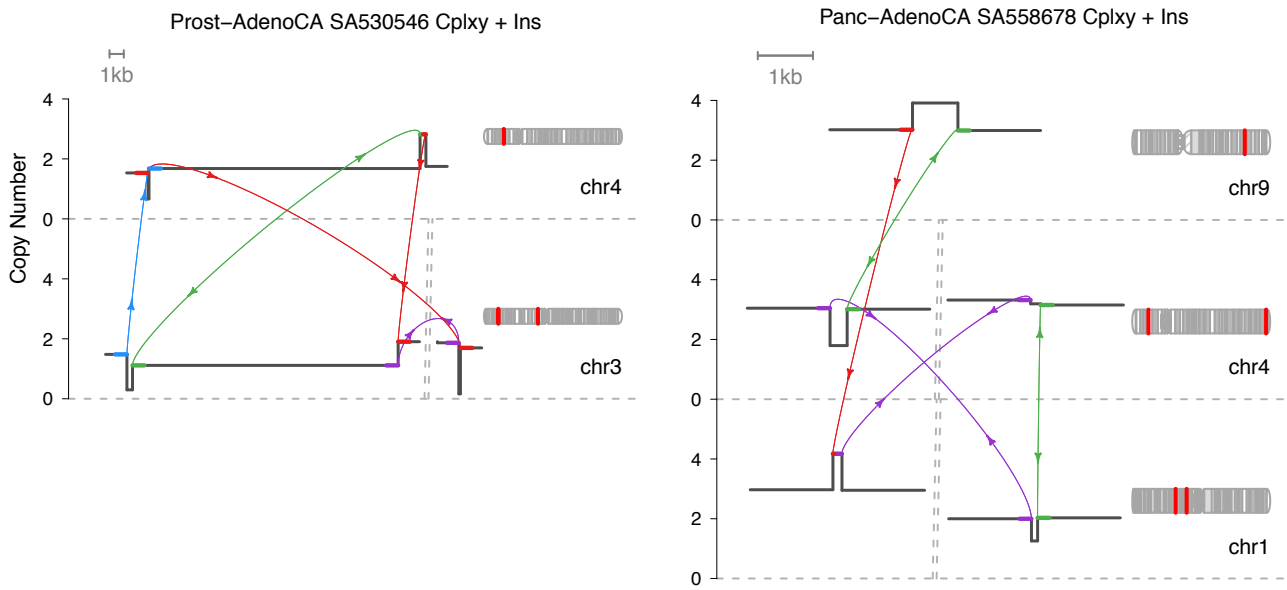


Figure D.5: Two of the longest chromoplexy with insertion events. Note that copy number estimates are unreliable in short segments $< 1\,\text{kb}$.

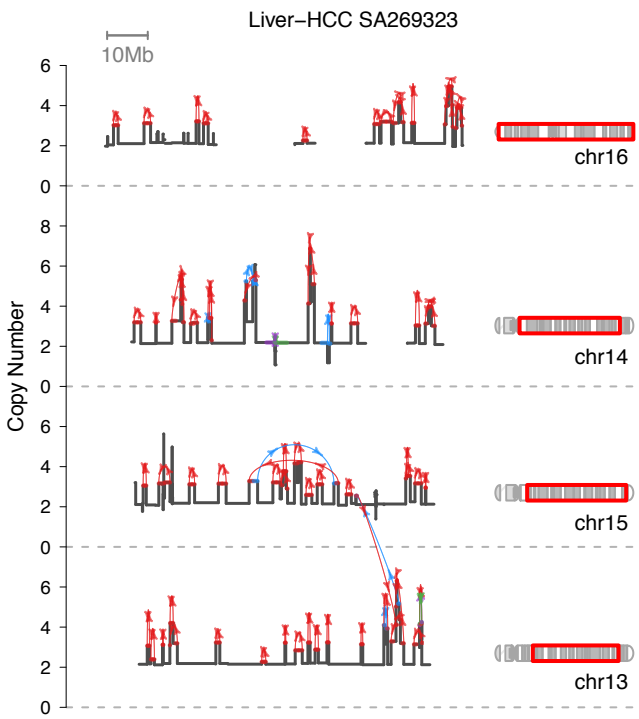


Figure D.6: All SV along four representative chromosomes from an unusual liver cancer sample with a high frequency of extremely large tandem duplications.

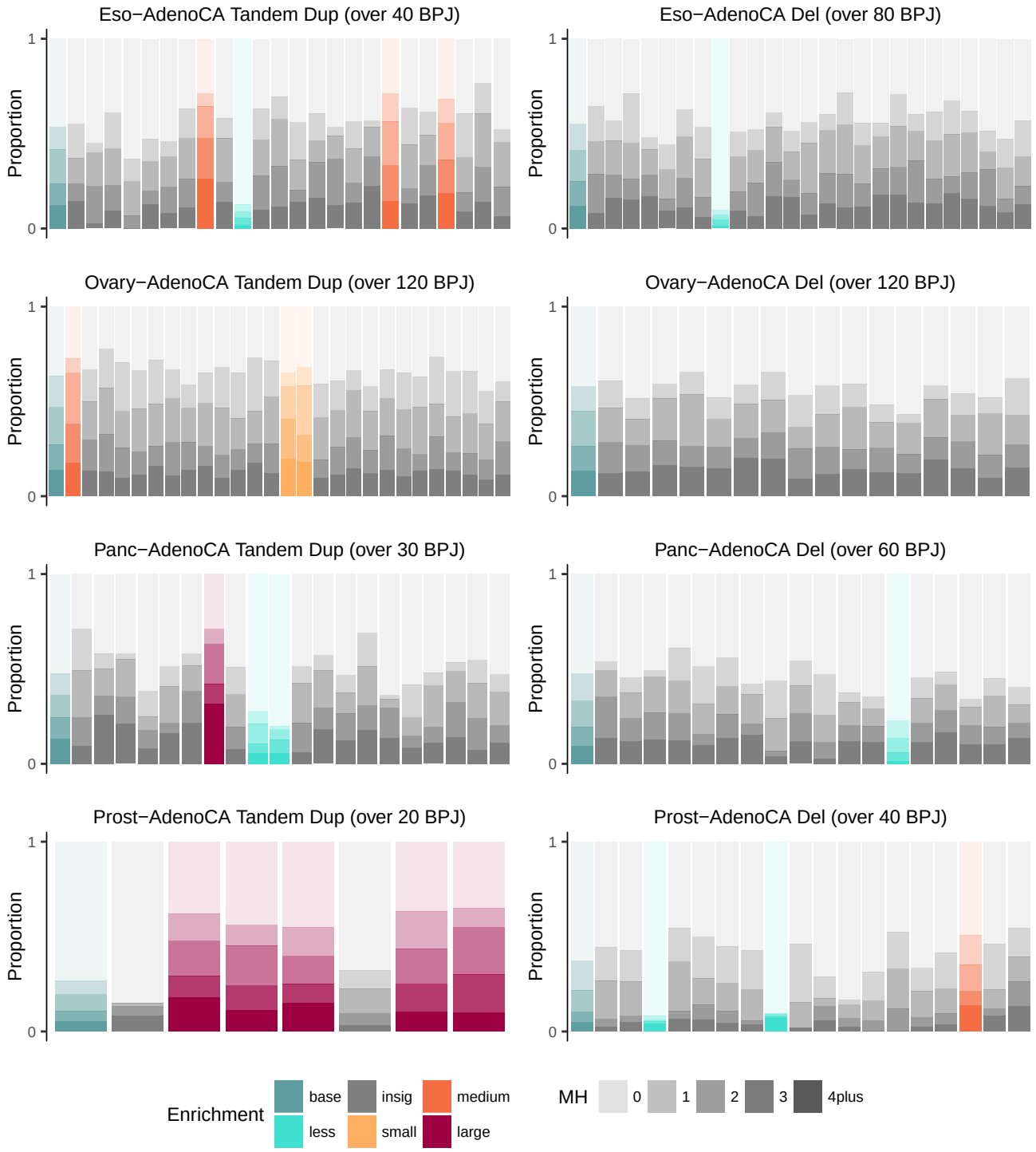


Figure D.7: Distribution of microhomology at the breakpoint junction for deletion and tandem duplication in individual samples with event counts above the indicated threshold. The magnitude of significant enrichment (compared to pool of other samples in the same histology class, shown in leftmost bar) is coloured by the proportional odds regression coefficient, split into less ($-\infty$–-0.25], small (0.25–0.50], medium (0.5–1.0], and large (1.0–∞) effect sizes. Non-significant samples (at Benjamini-Hochberg 0.01 FDR threshold) are shaded grey.

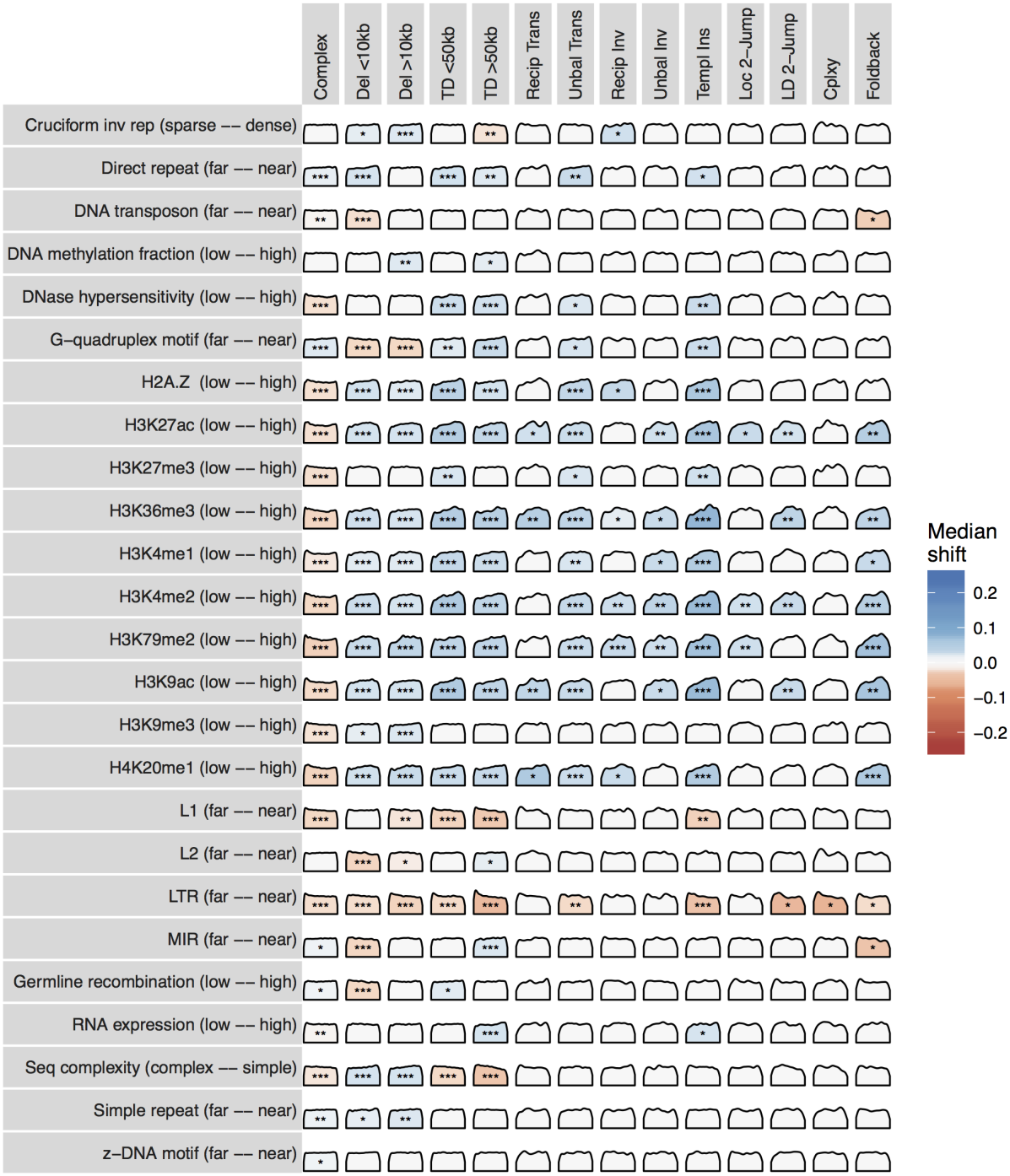


Figure D.8: For each SV class, the quantile distribution of the genomic property metrics at observed breakpoints compared to random positions, with significant departure from uniform quantiles marked by: FDR < 0.01 *, < 0.001 **, and $< 10^{-6}$ ***; shading the magnitude of the shift of the median observed quantile above (blue) or below (red) 0.5.

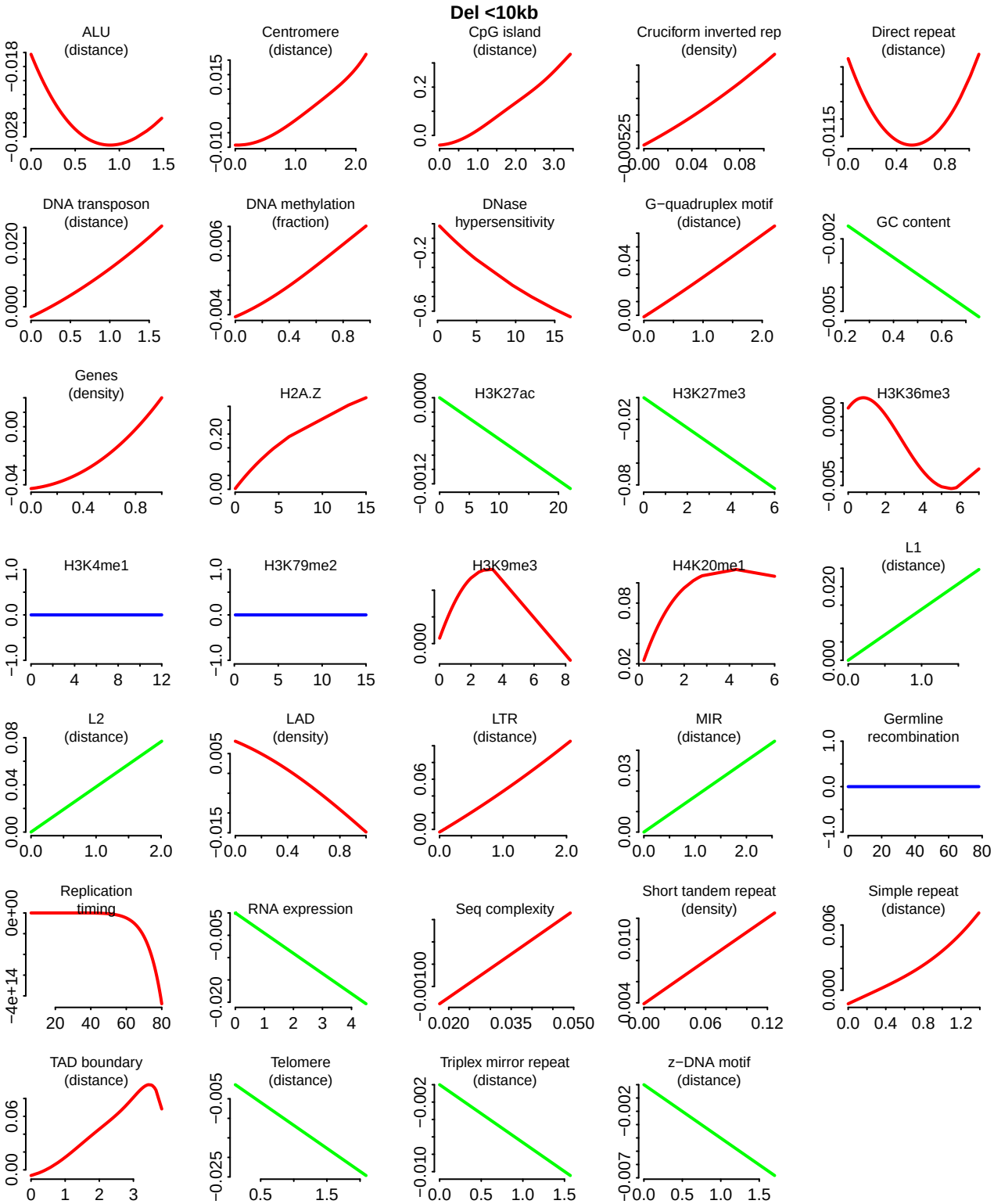


Figure D.9: The optimal lasso GAM for small deletions, with predictor effects on the log odds of a real breakpoint in red for splines, green for linear terms, and blue for removed predictors.

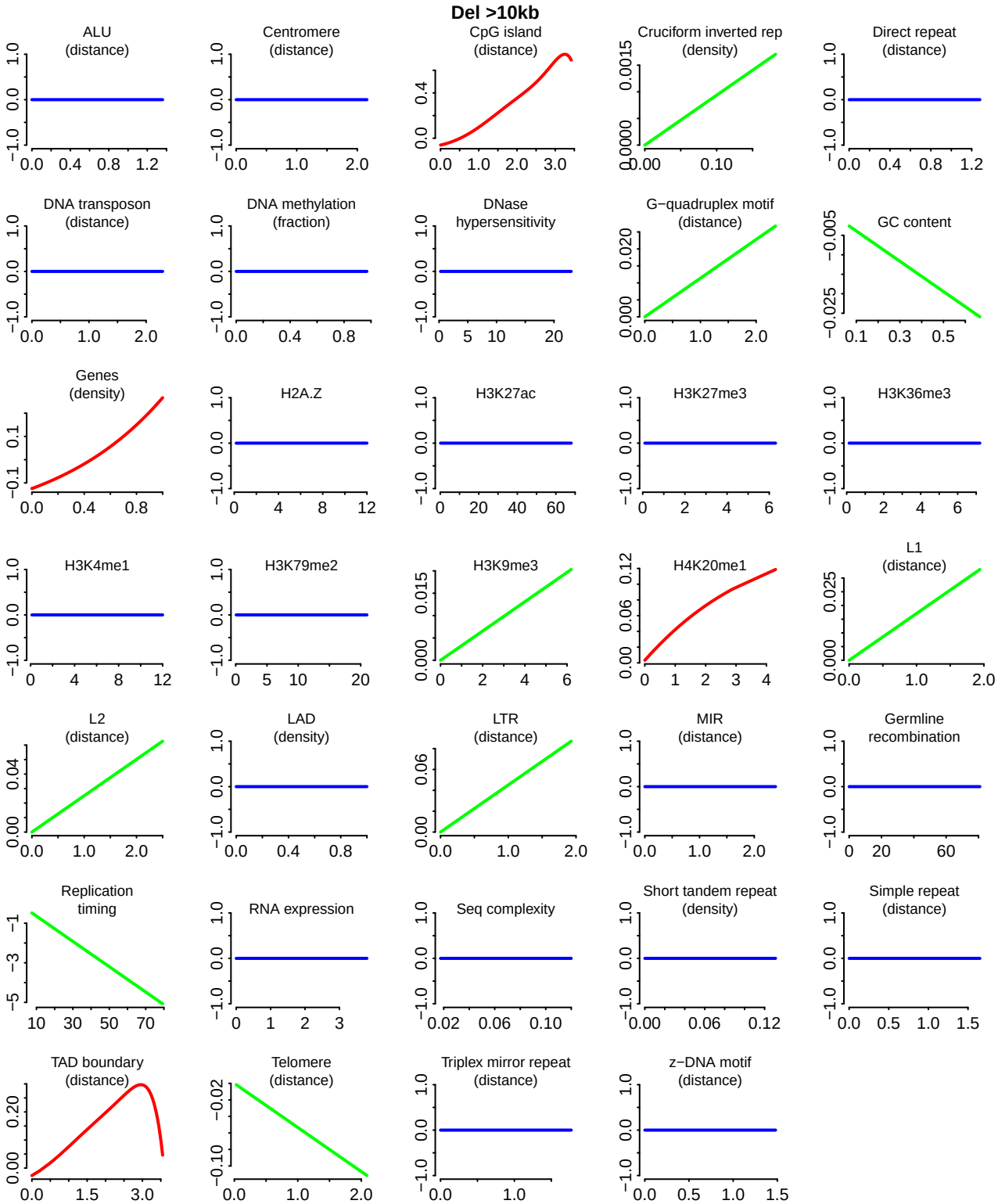


Figure D.10: The optimal lasso GAM for large deletions, with predictor effects on the log odds of a real breakpoint in red for splines, green for linear terms, and blue for removed predictors.

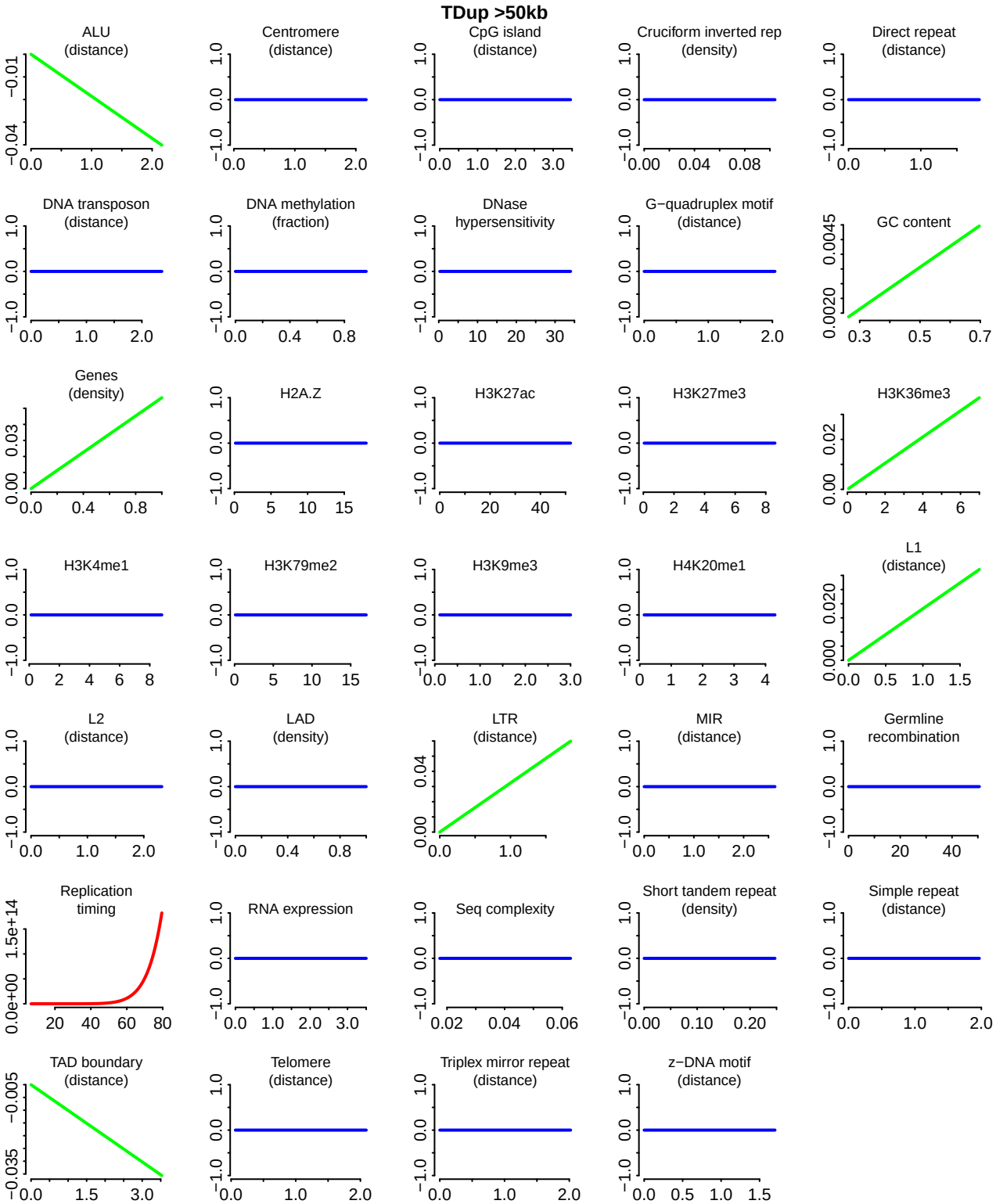


Figure D.11: The optimal lasso GAM for large tandem duplication, with predictor effects on the log odds of a real breakpoint in red for splines, green for linear terms, and blue for removed predictors.

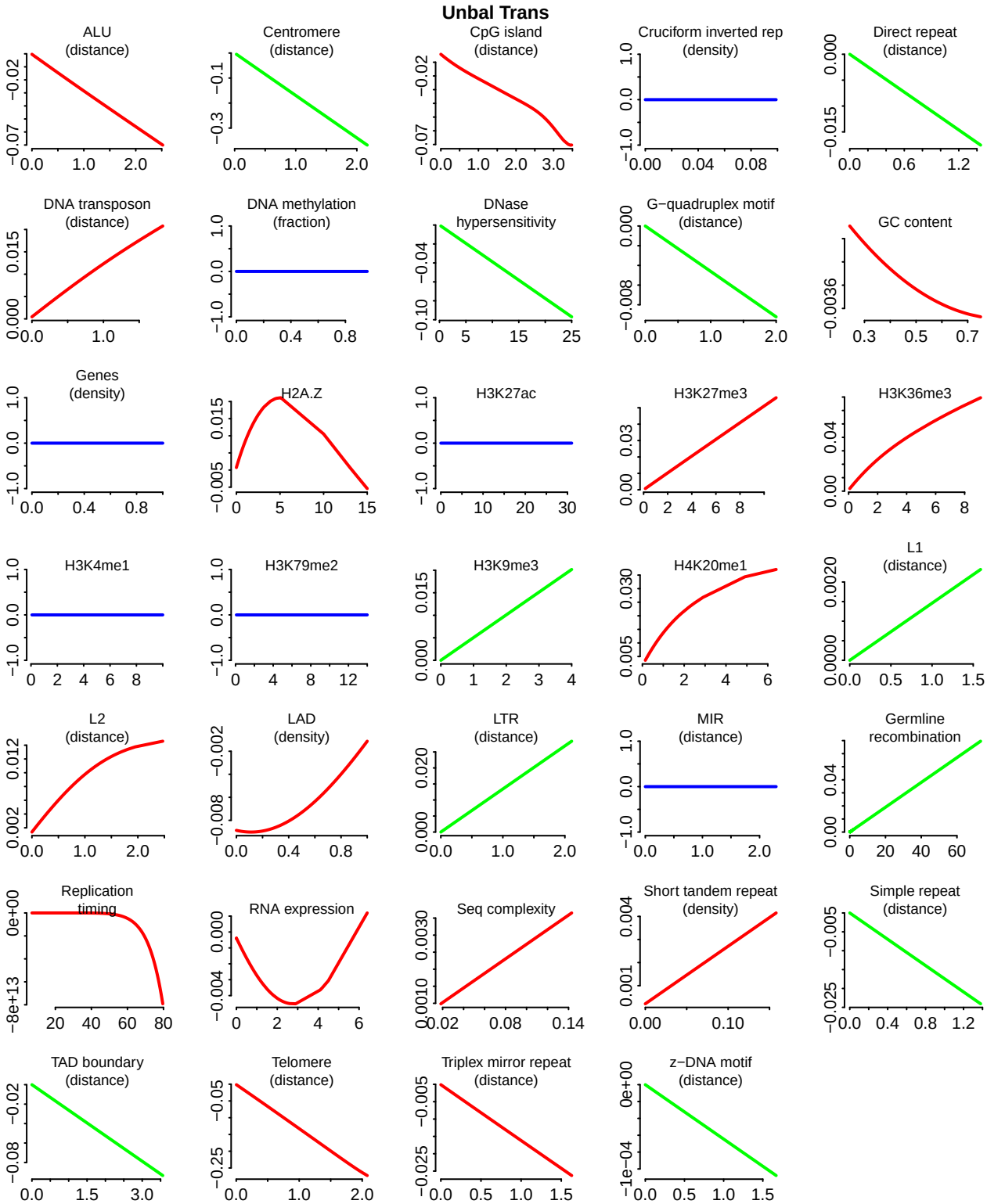


Figure D.12: The optimal lasso GAM for unbalanced translocation, with predictor effects on the log odds of a real breakpoint in red for splines, green for linear terms, and blue for removed predictors.

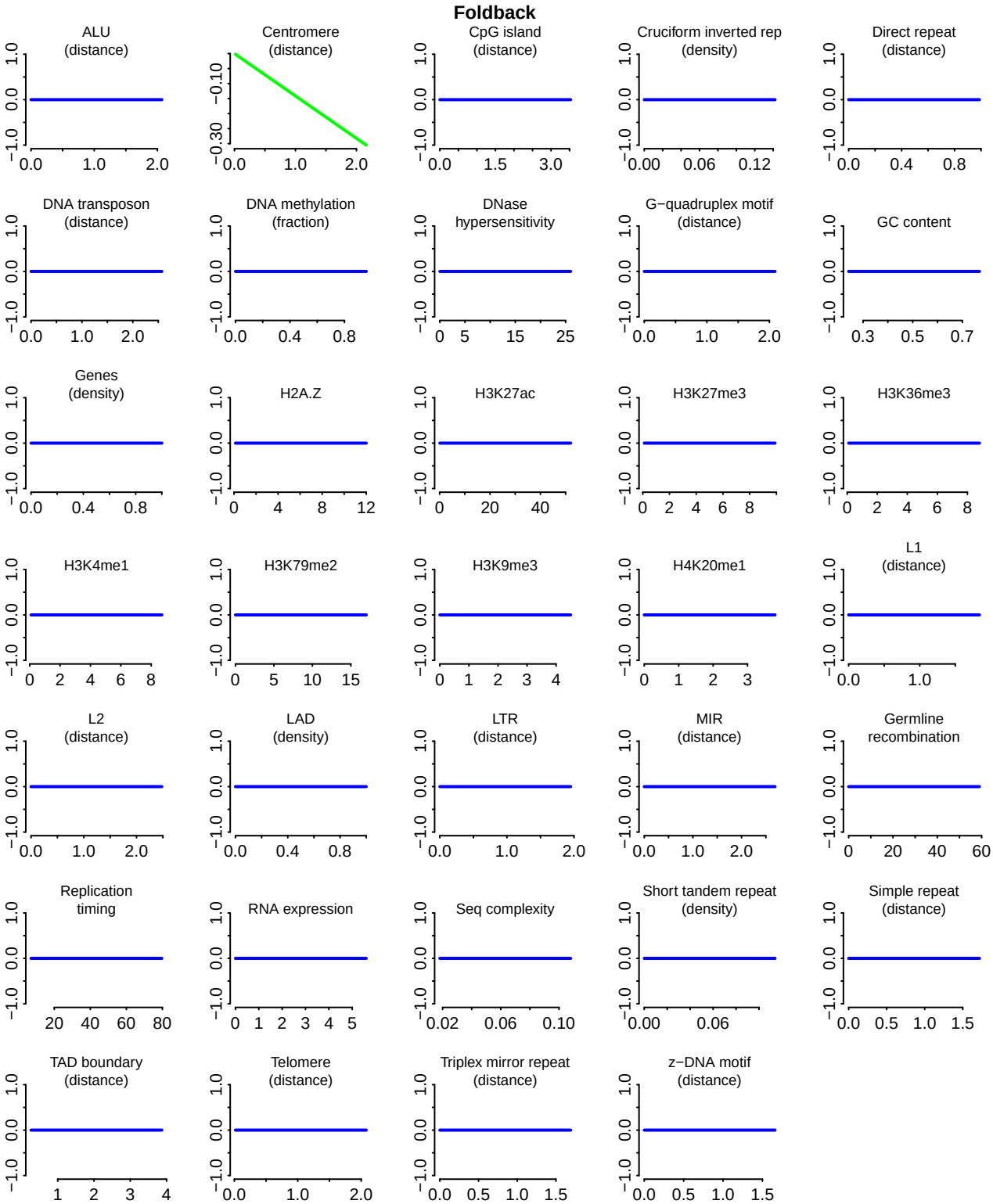


Figure D.13: The optimal lasso GAM for foldback, with predictor effects on the log odds of a real breakpoint in red for splines, green for linear terms, and blue for removed predictors.

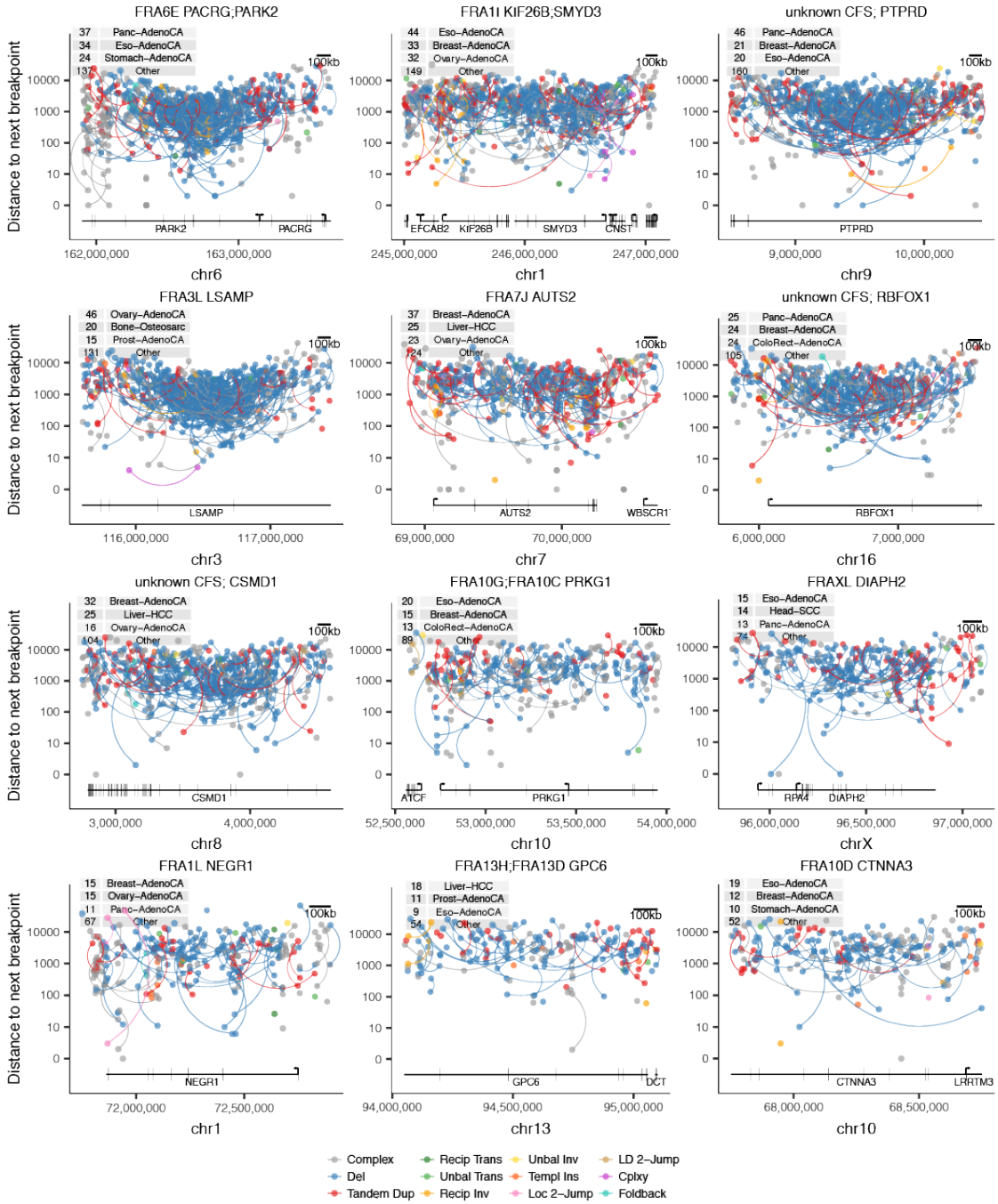


Figure D.14: All SV breakpoint positions in the 12 minor fragile sites. If the two sides of a BPJ are contained within the plotting window, they are joined with a curved line. The number of samples with a breakpoint in the plotting window is annotated top left.

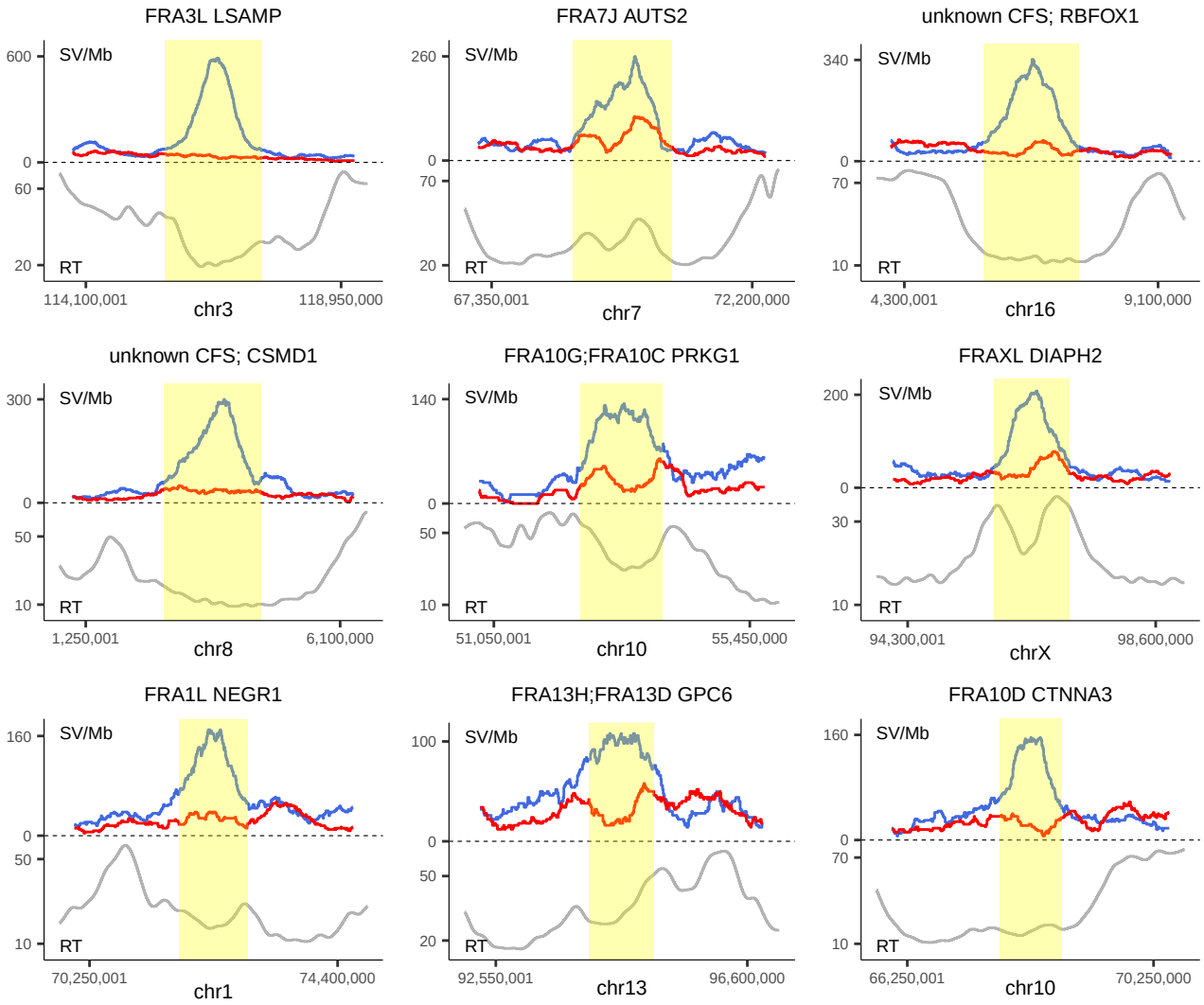


Figure D.15: Extending to 2 Mb flanks either side of nine minor FS marked in yellow, the upper plot shows the density of deletion (blue) and tandem dup (red) breakpoints in 500 kb windows sliding every 10 kb. The lower plot shows the replication timing track, with high values for early and low for late.

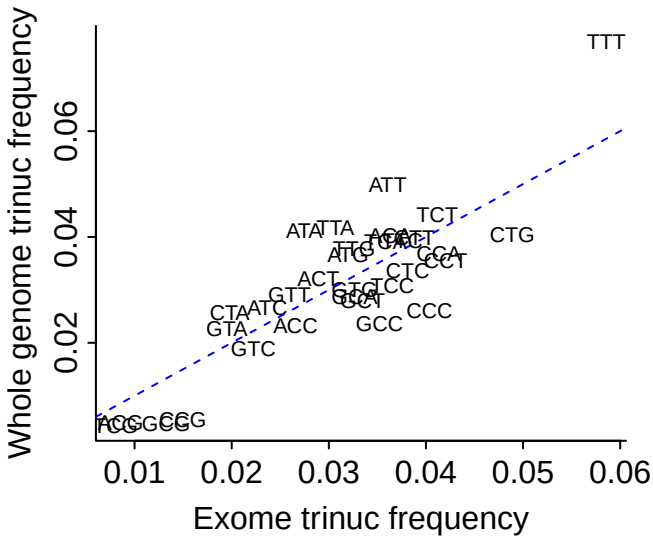


Figure D.16: Trinucleotide frequency in human exome (plus 100 bp flanks) and whole genome (callable regions), with blue line denoting equal values. Reverse complements are consolidated, and reported with the middle base as the pyrimidine (`C` or `T`).

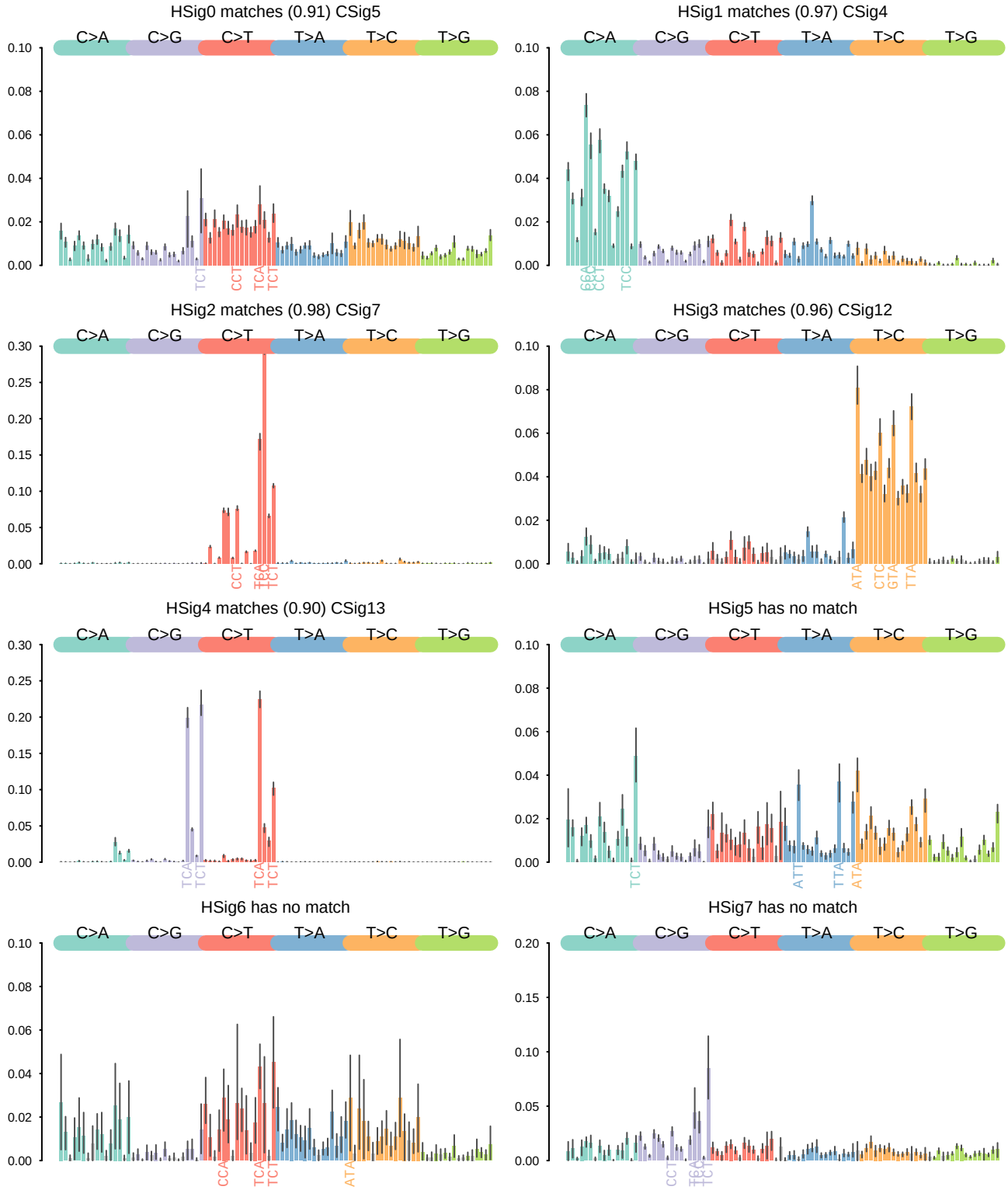


Figure D.17: HDP mutational signatures in discovery dataset (mean and 95% credibility interval from MCMC posterior samples, with non-significant mutation classes in grey and four major peaks labelled with trinucleotide context).

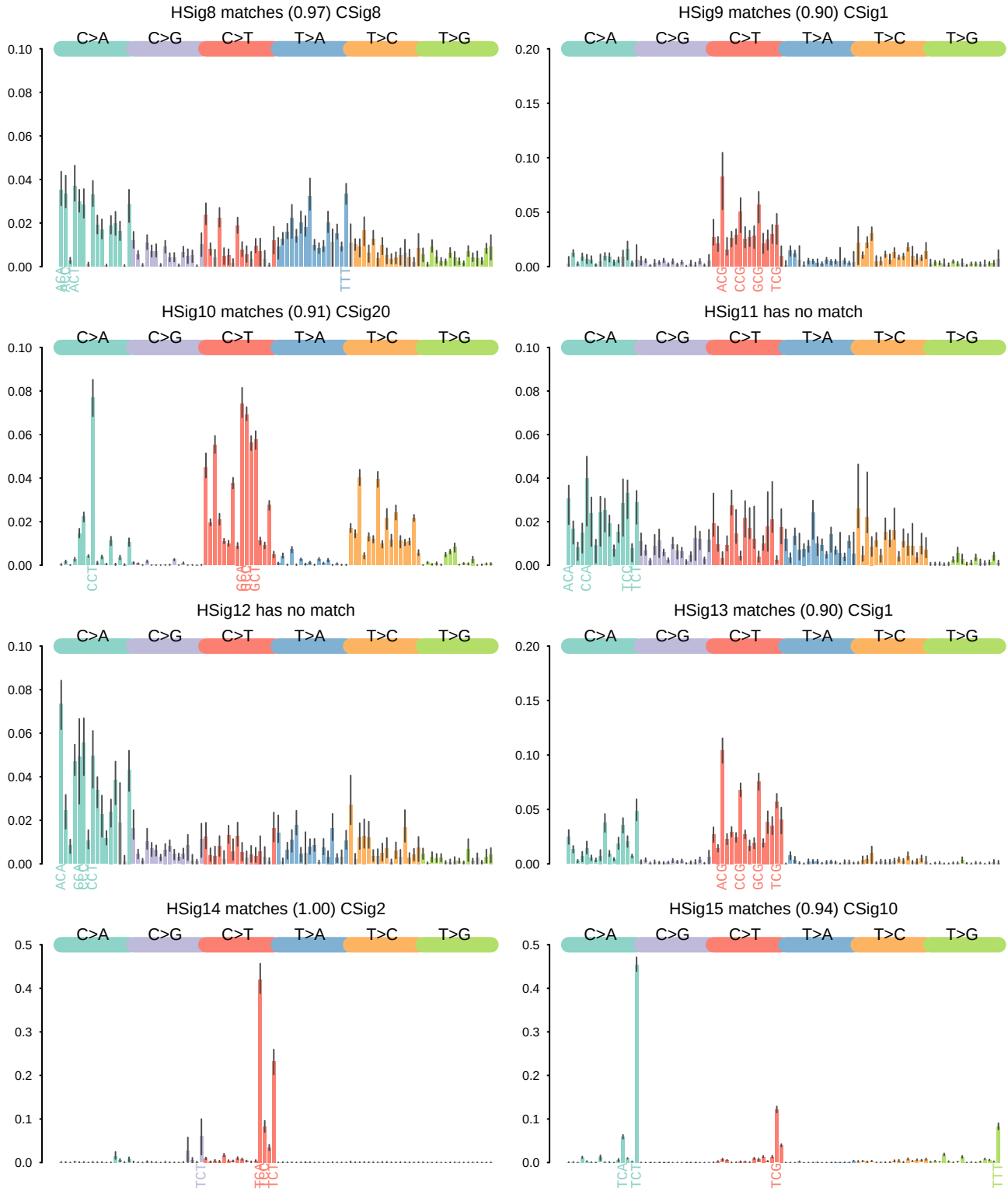


Figure D.17: HDP mutational signatures in discovery dataset (continued from previous)

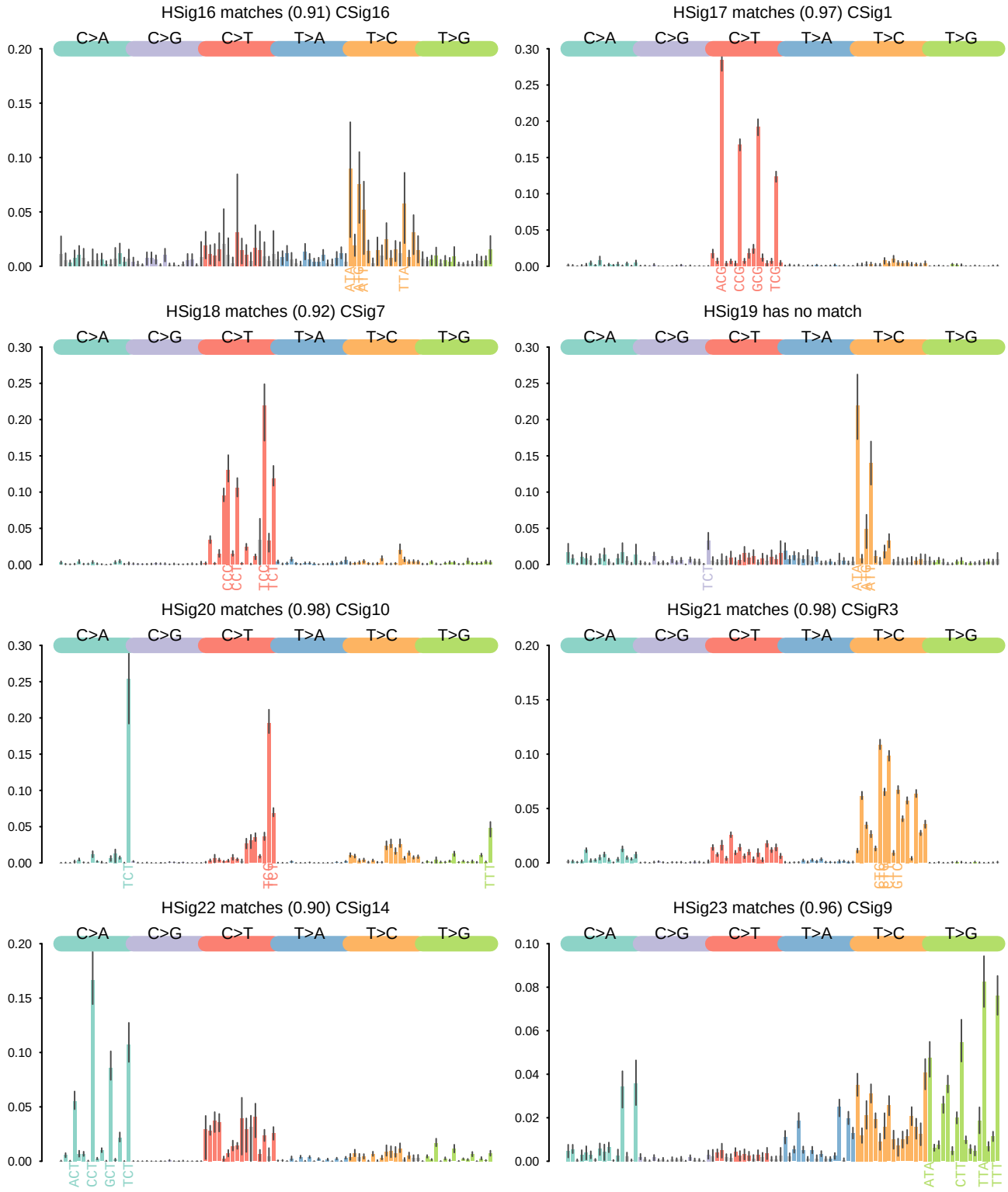


Figure D.17: HDP mutational signatures in discovery dataset (continued from previous)

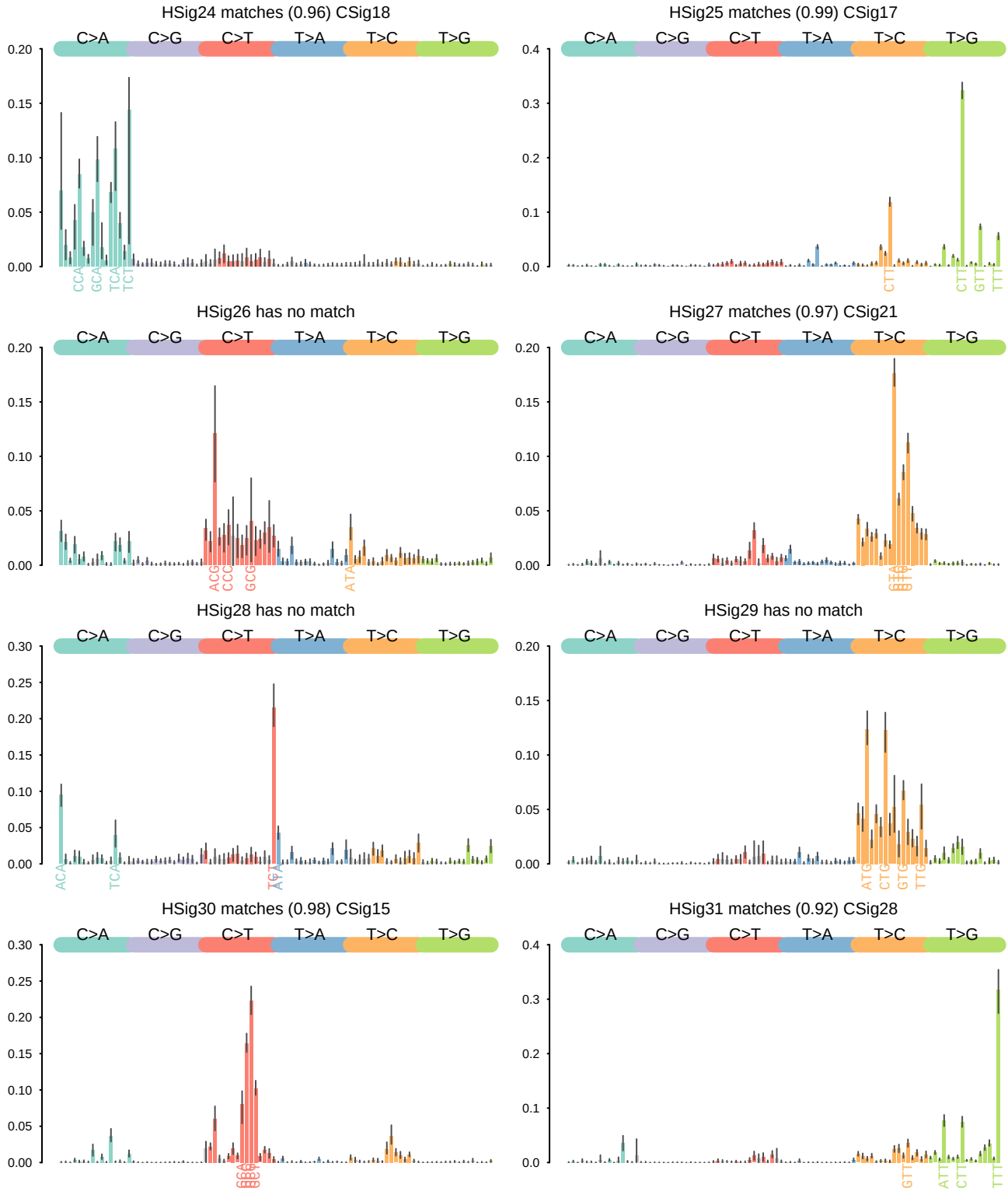


Figure D.17: HDP mutational signatures in discovery dataset (continued from previous)

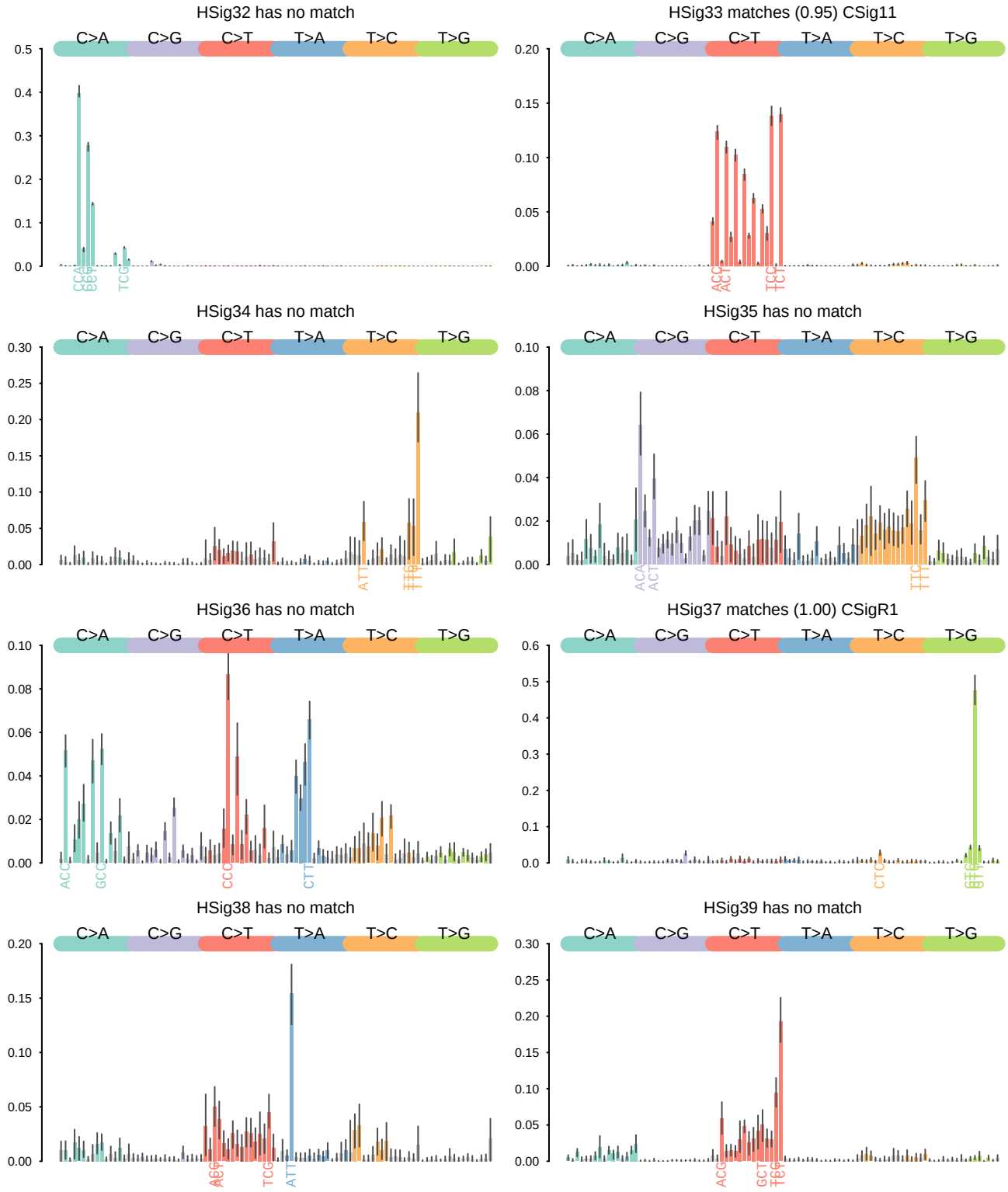


Figure D.17: HDP mutational signatures in discovery dataset (continued from previous)

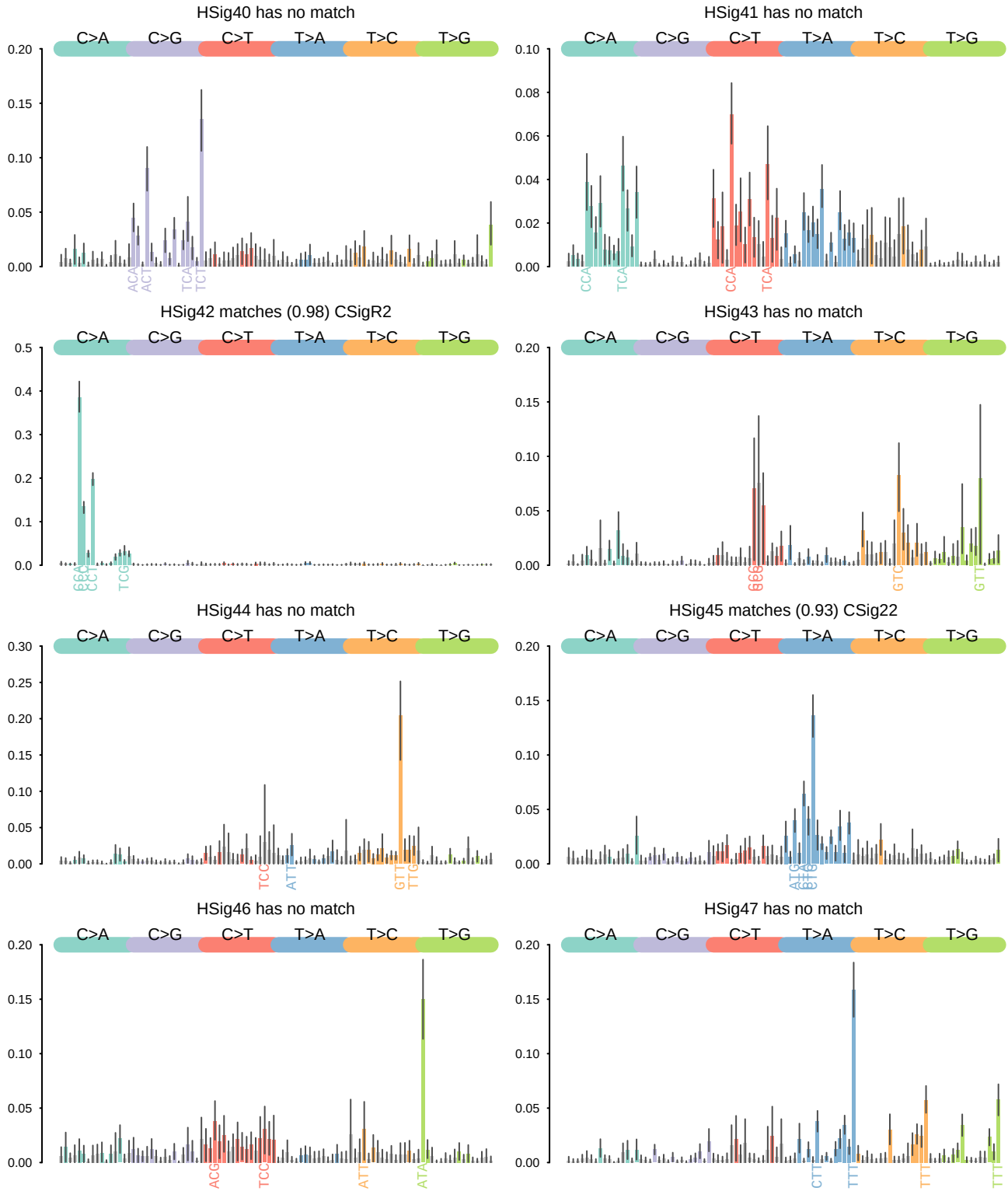


Figure D.17: HDP mutational signatures in discovery dataset (continued from previous)

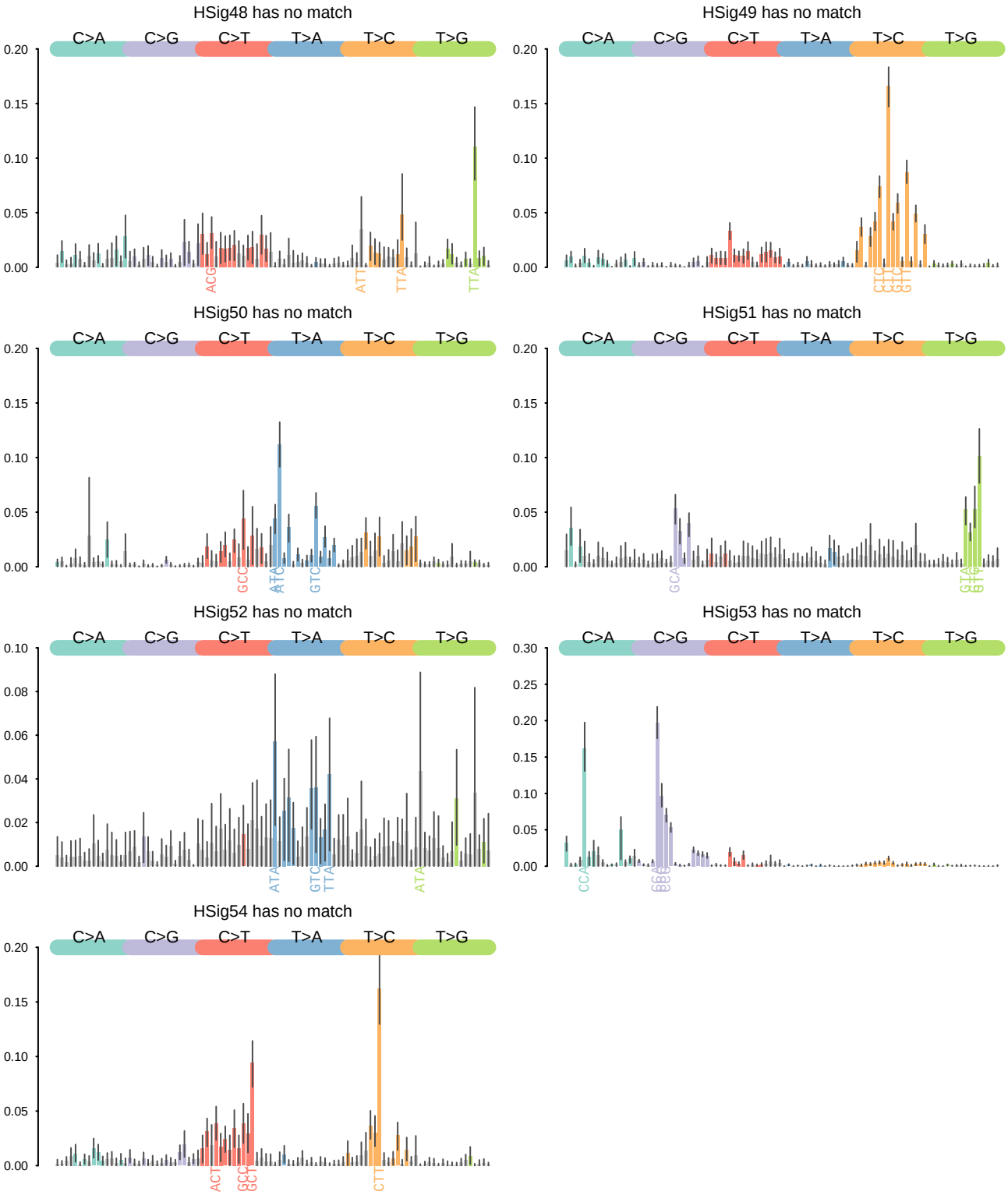


Figure D.17: HDP mutational signatures in discovery dataset (continued from previous)

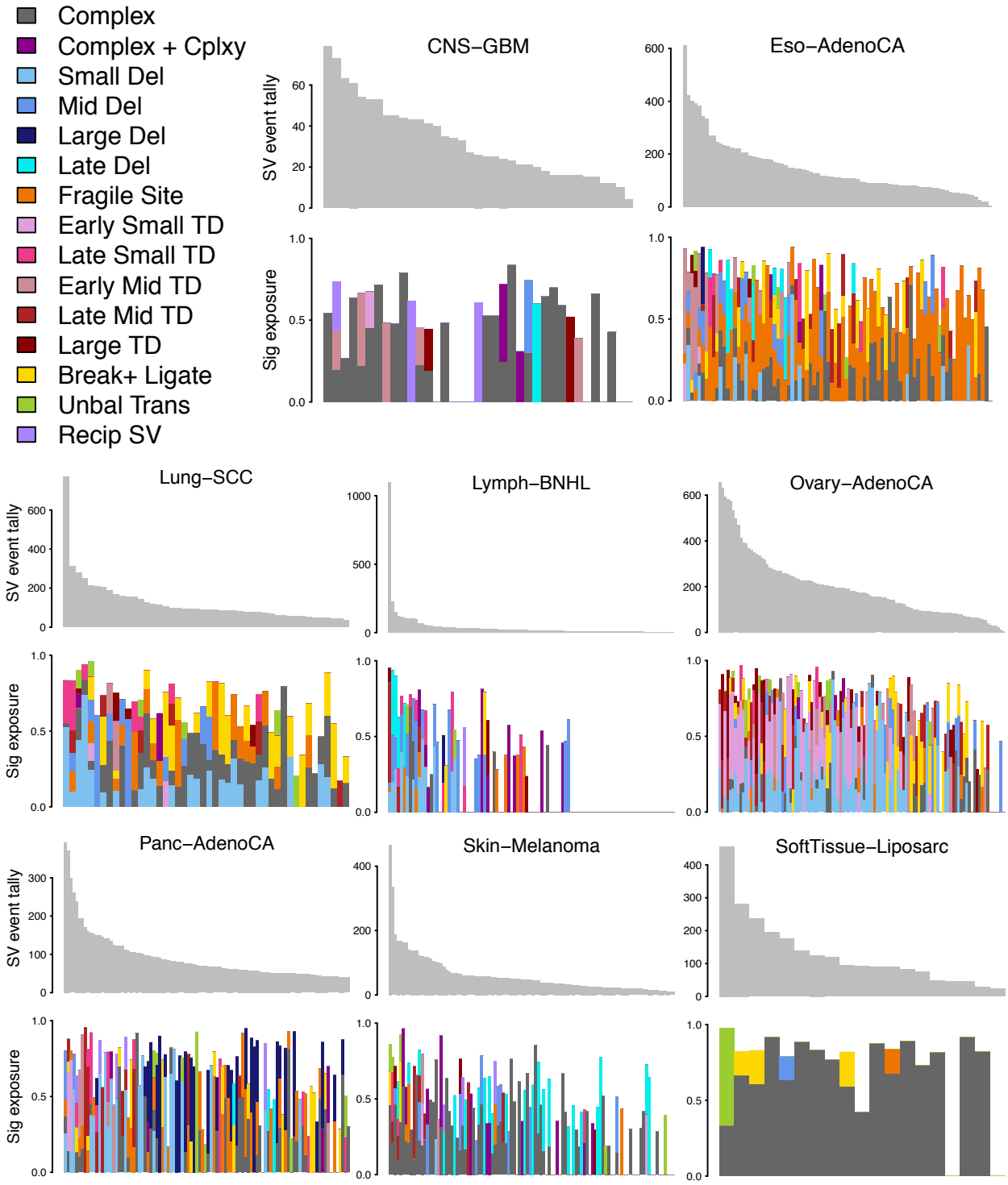


Figure D.18: Average estimated sample exposures to HDP-extracted SV signatures (Figure 4.20) for eight of the PCAWG cancer types. Large cohorts are subset to a maximum of 100 samples for presentation. For each sample, mean estimated exposures are only plotted for significant signatures (95% credibility interval above zero), leaving a blank proportion with uncertain allocation to the same or different signatures.

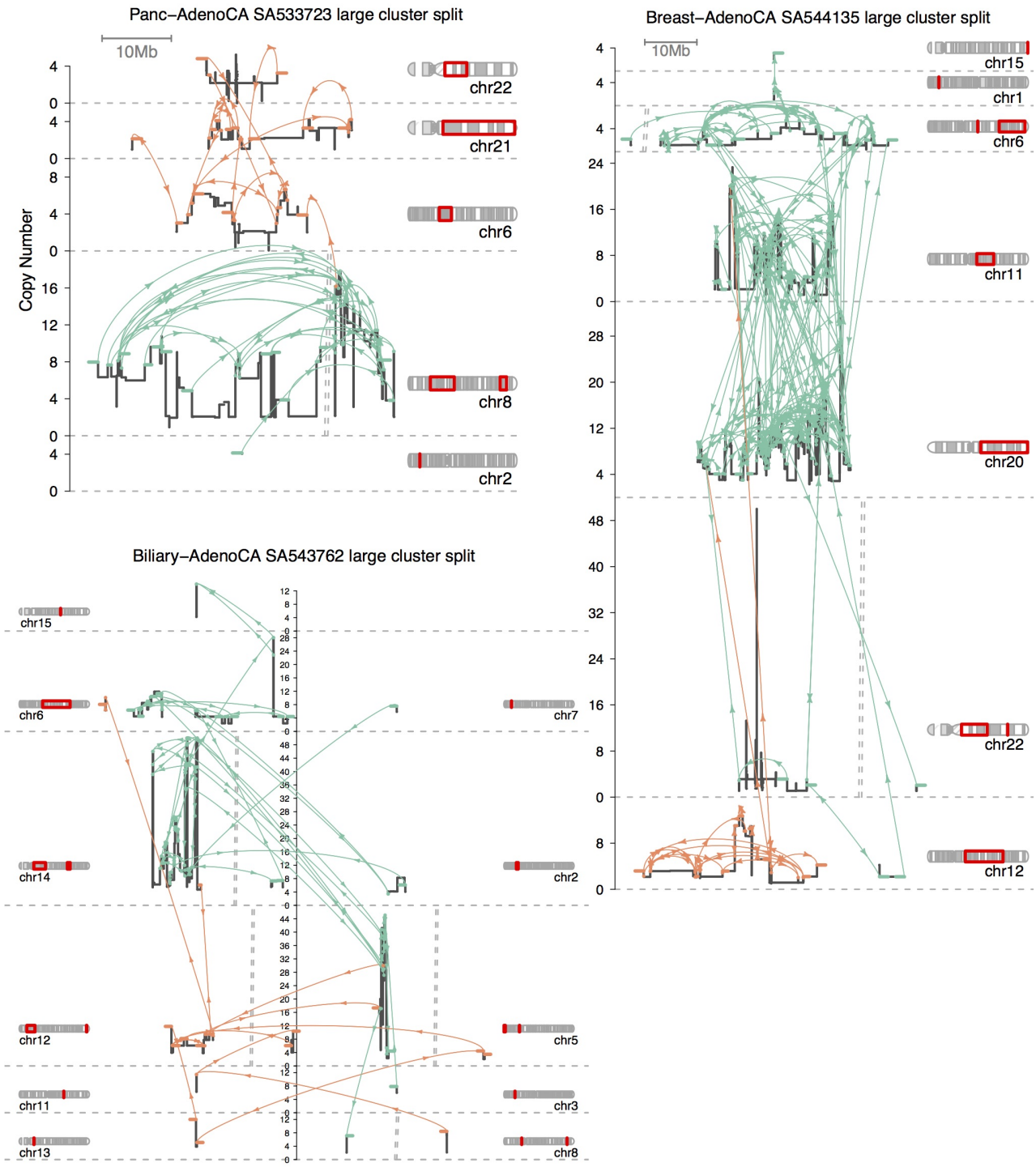
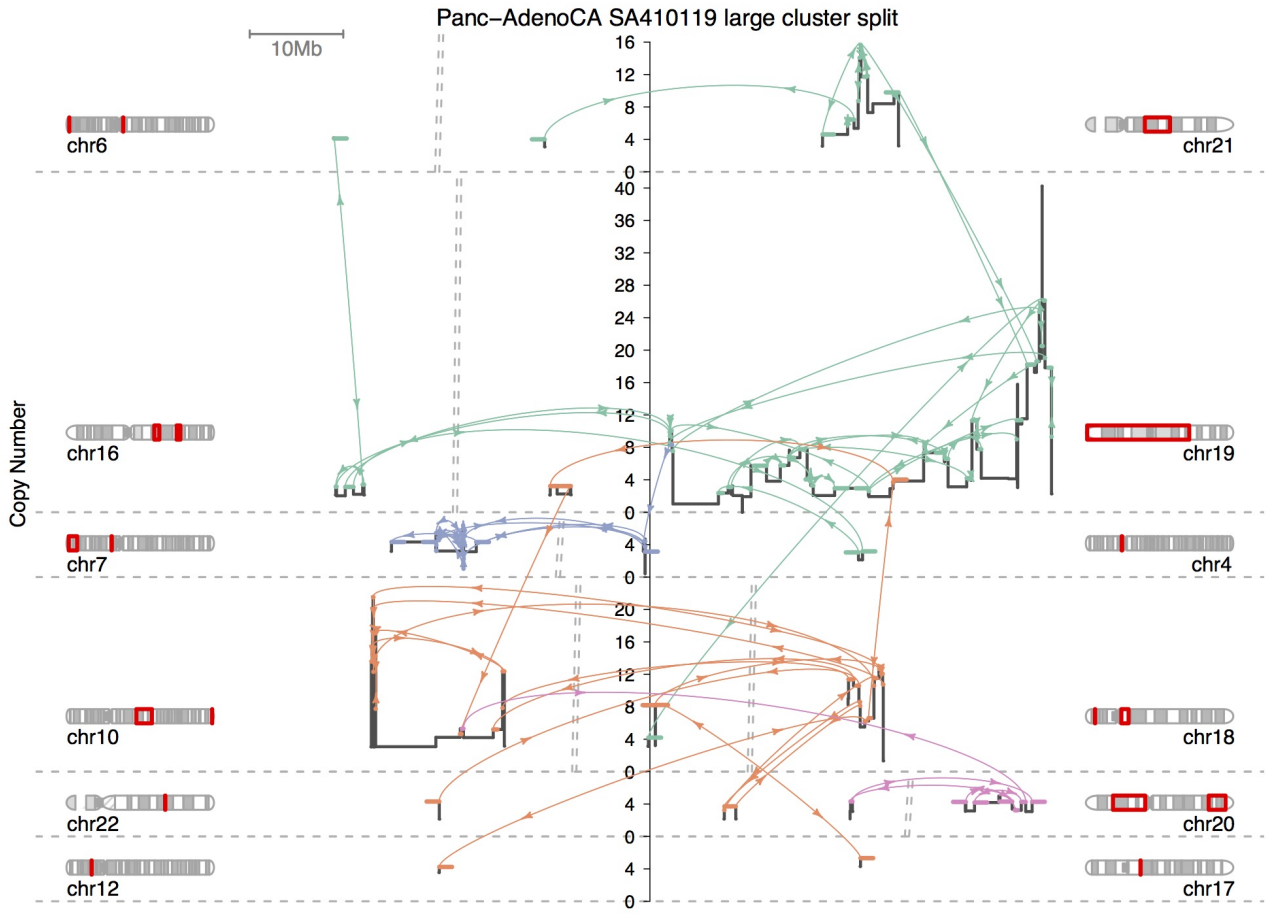


Figure D.19: Large BPJ candidate clusters, split by walktrap graph community detection as described in Section 5.1.1. Node-edge graph visualisations are available in Figures 5.4–5.6. Figure continues on the next page.



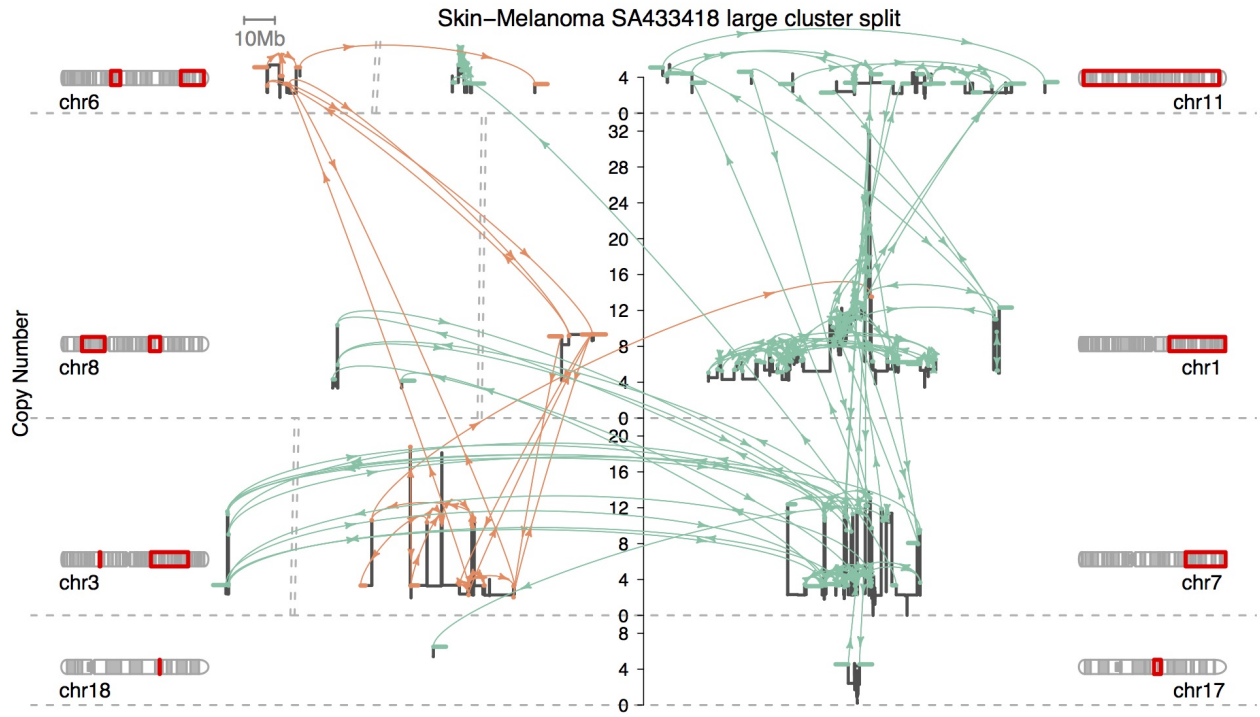


Figure D.19: Large BPJ candidate clusters, split by walktrap graph community detection as described in Section 5.1.1. Node-edge graph visualisations are available in Figures 5.4–5.6. Figure continued from the previous page.

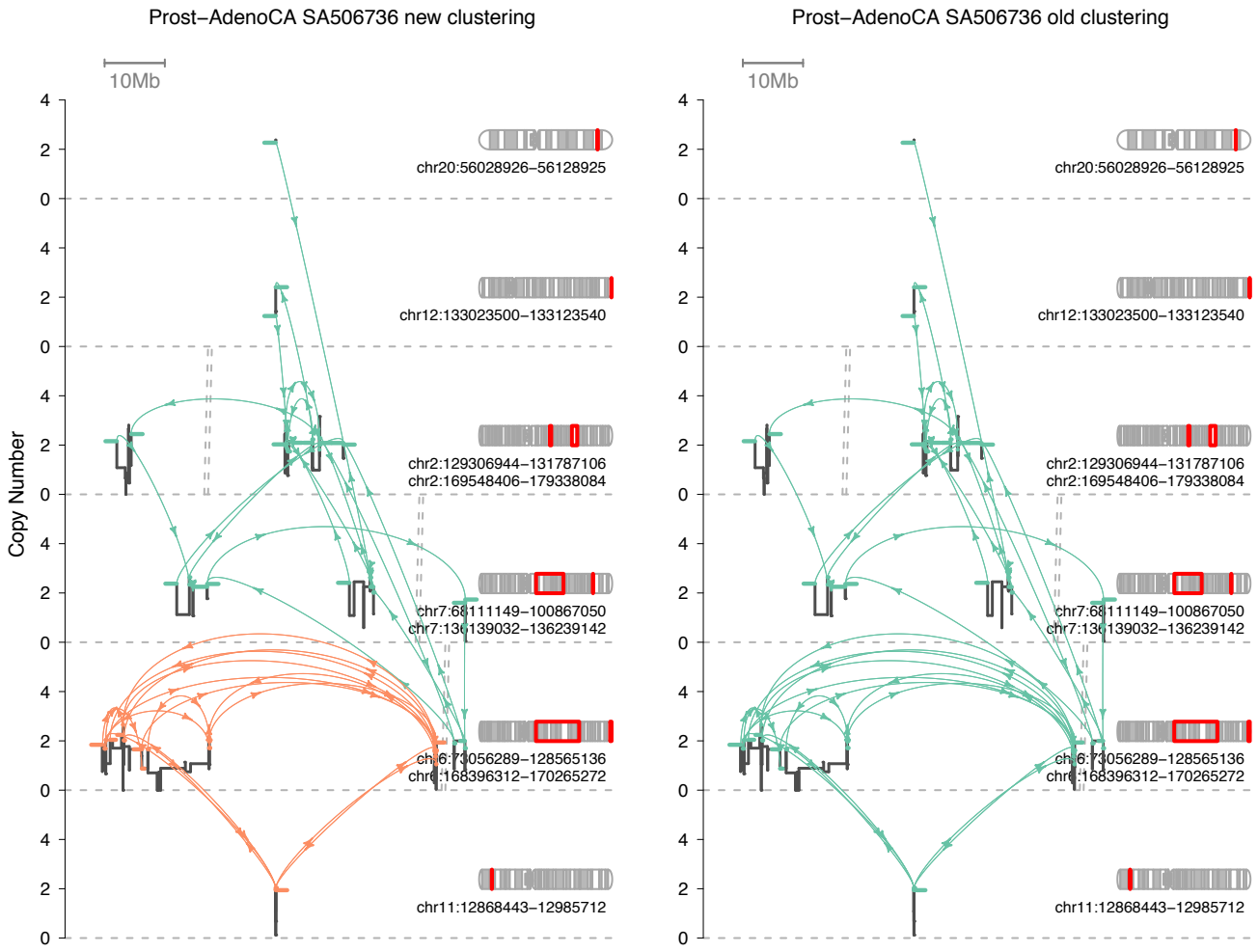


Figure D.20: New and old clustering schemes (more clusters in new) for complex unexplained BPJ in prostate cancer sample SA506736, only showing clusters with disagreement.

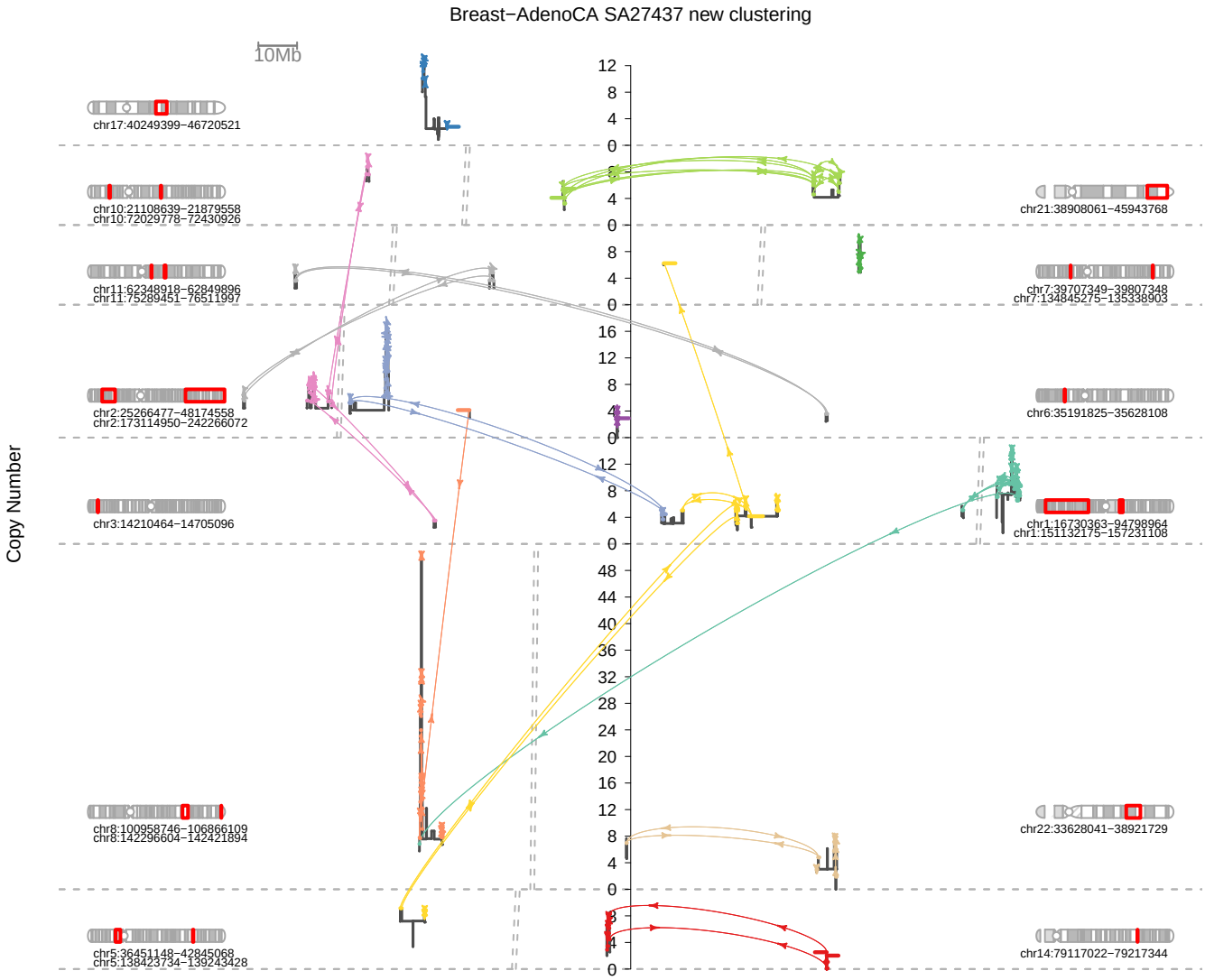


Figure D.21: New and old clustering schemes (more clusters in new) for complex unexplained BPJ in breast cancer sample SA27437, only showing clusters with disagreement. Figure continues on the next page.

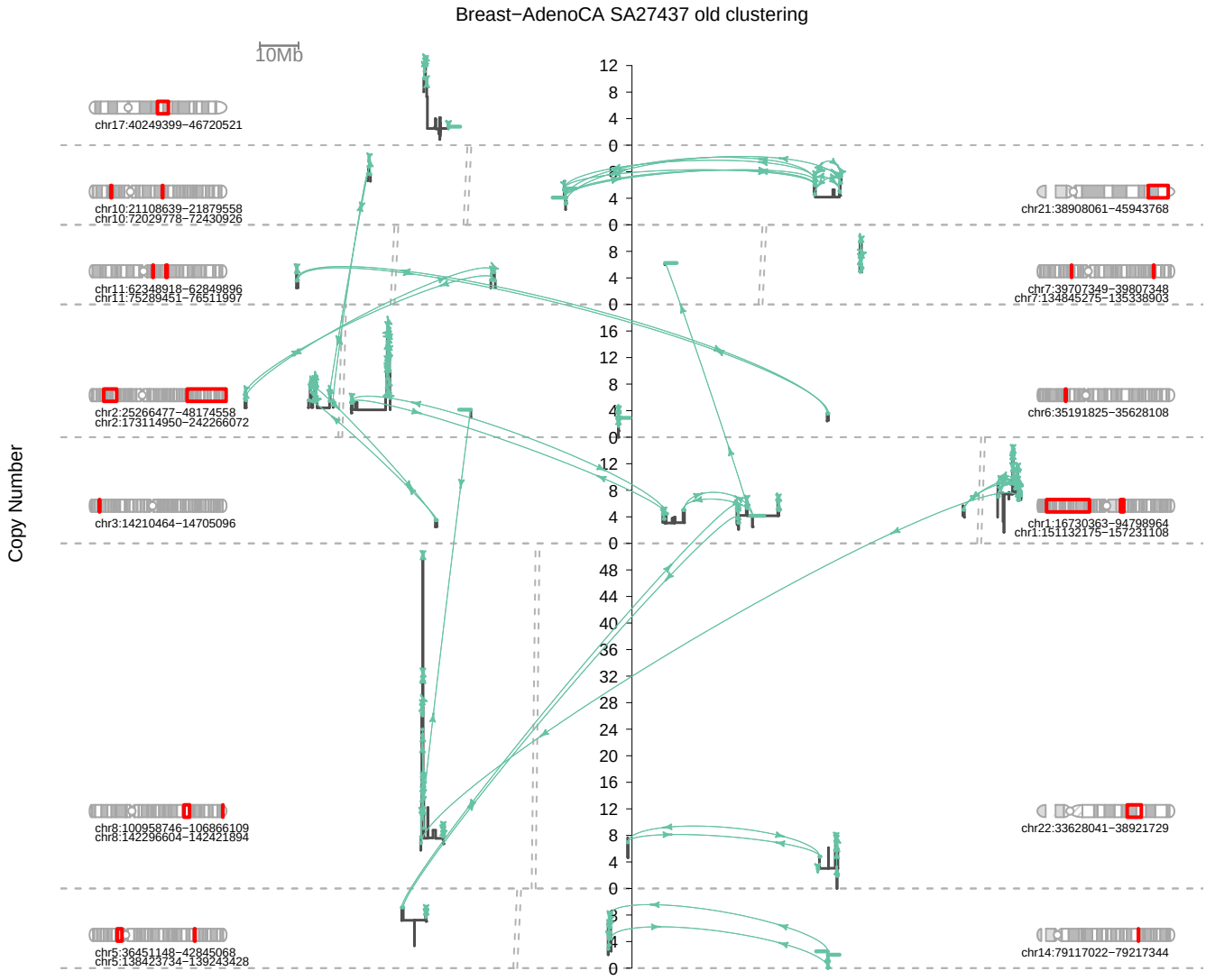


Figure D.21: New and old clustering schemes (more clusters in new) for complex unexplained BPJ in breast cancer sample SA27437, only showing clusters with disagreement. Figure is continued from the previous page.

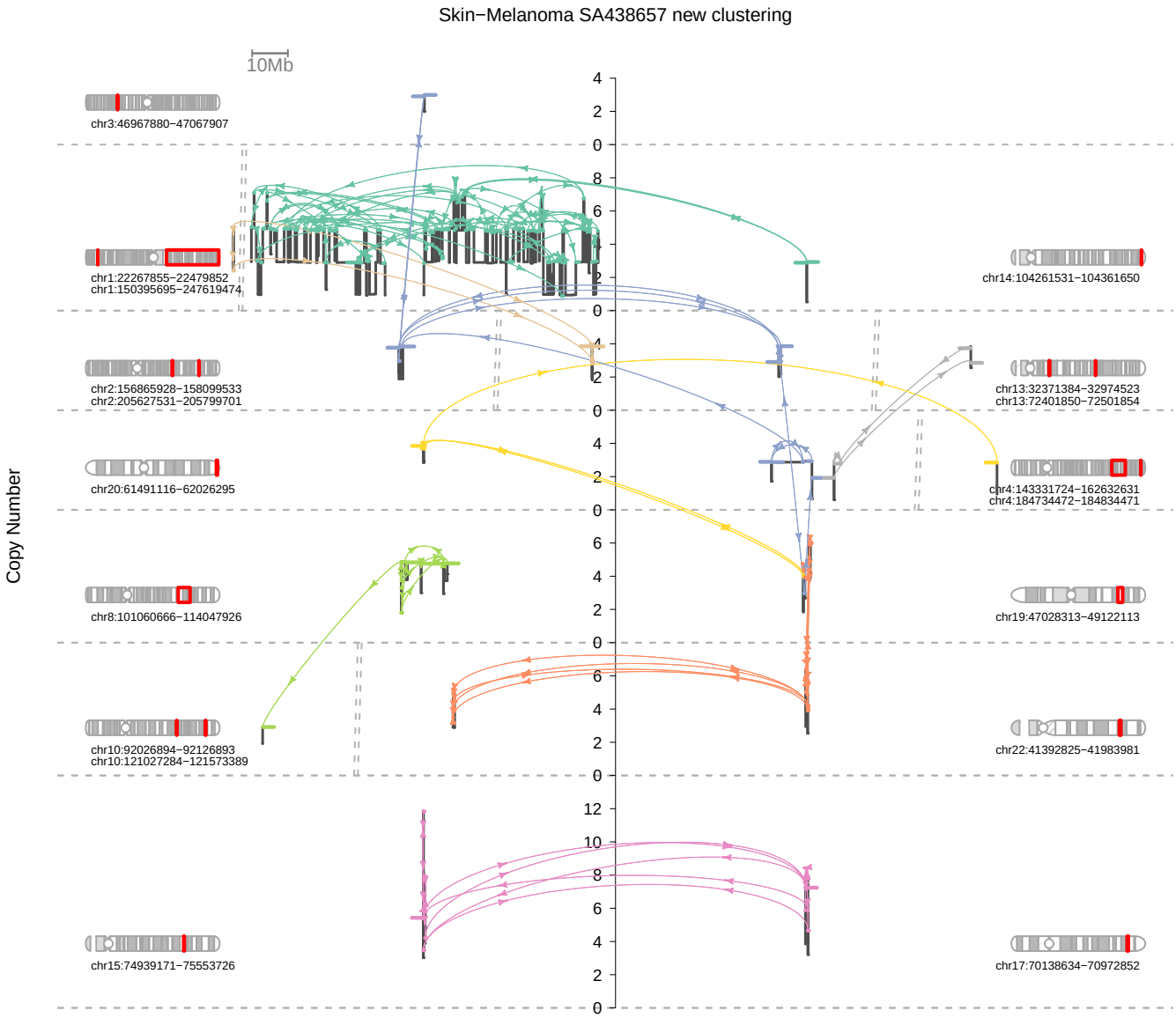


Figure D.22: New and old clustering schemes (more clusters in new) for complex unexplained BPJ in melanoma sample SA438657, only showing clusters with disagreement. Figure continues on the next page.

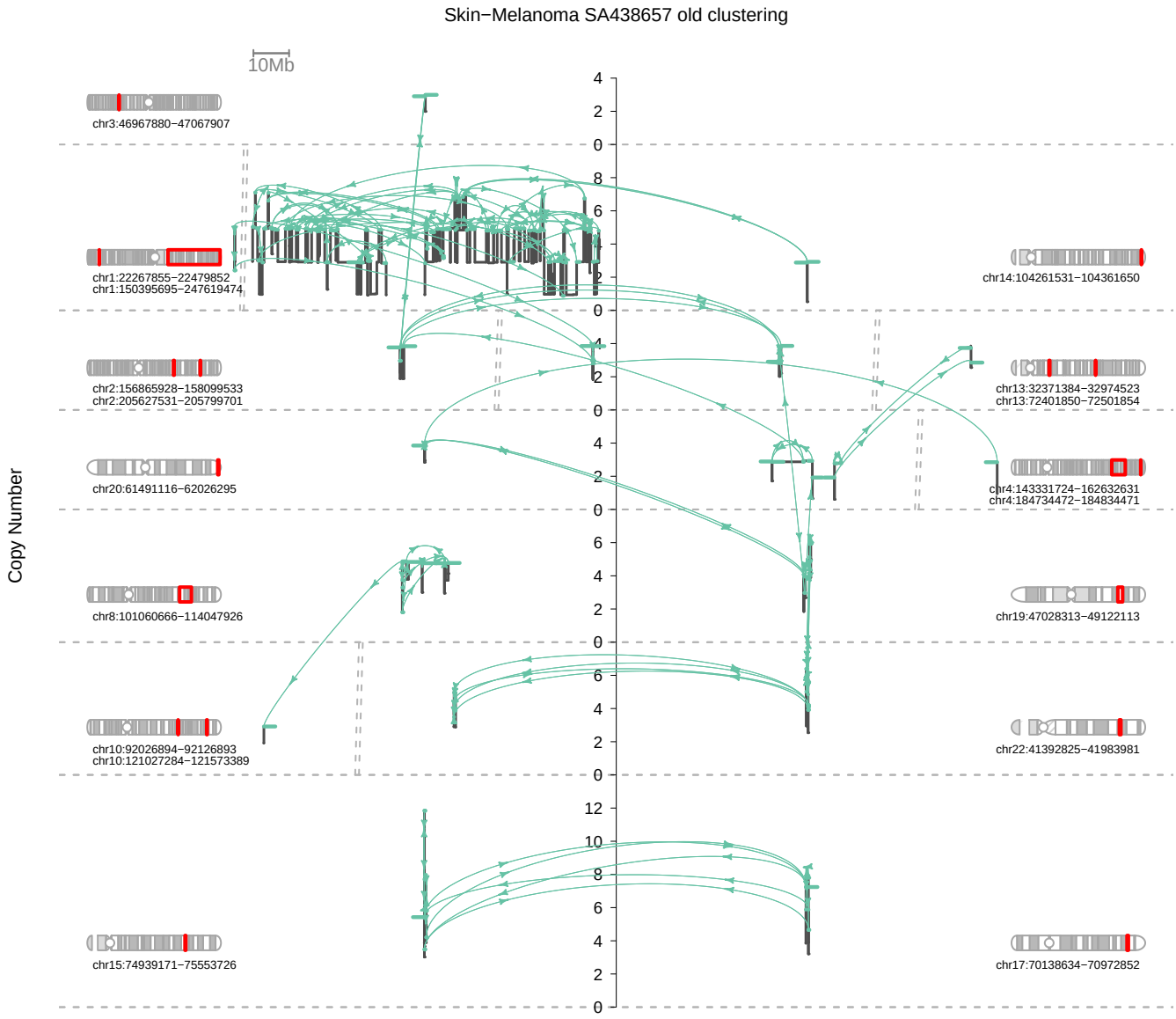


Figure D.22: New and old clustering schemes (more clusters in new) for complex unexplained BPJ in melanoma sample SA438657, only showing clusters with disagreement. Figure is continued from the previous page.

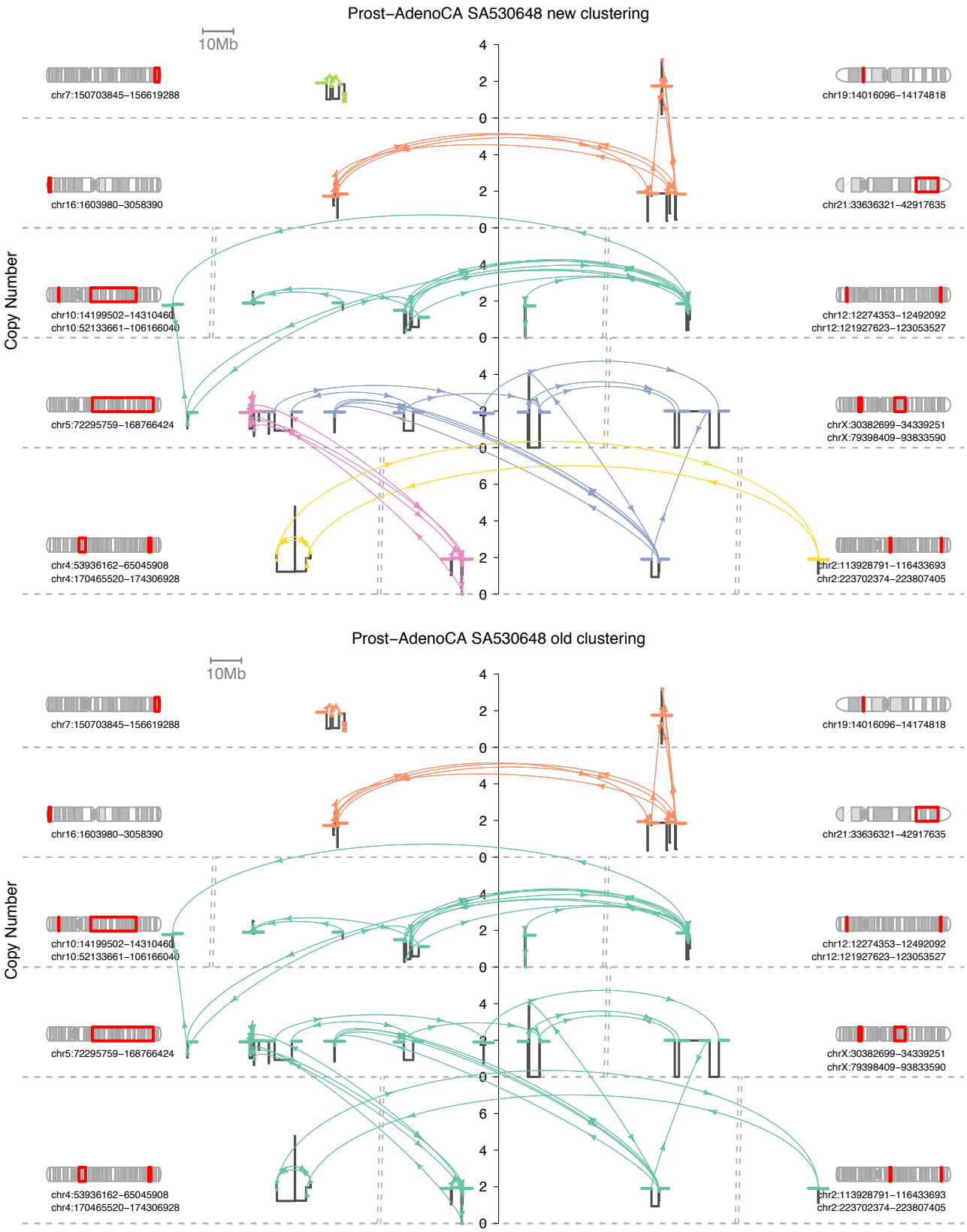


Figure D.23: New and old clustering schemes (more clusters in new) for complex unexplained BPJ in prostate cancer sample SA530648, only showing clusters with disagreement.

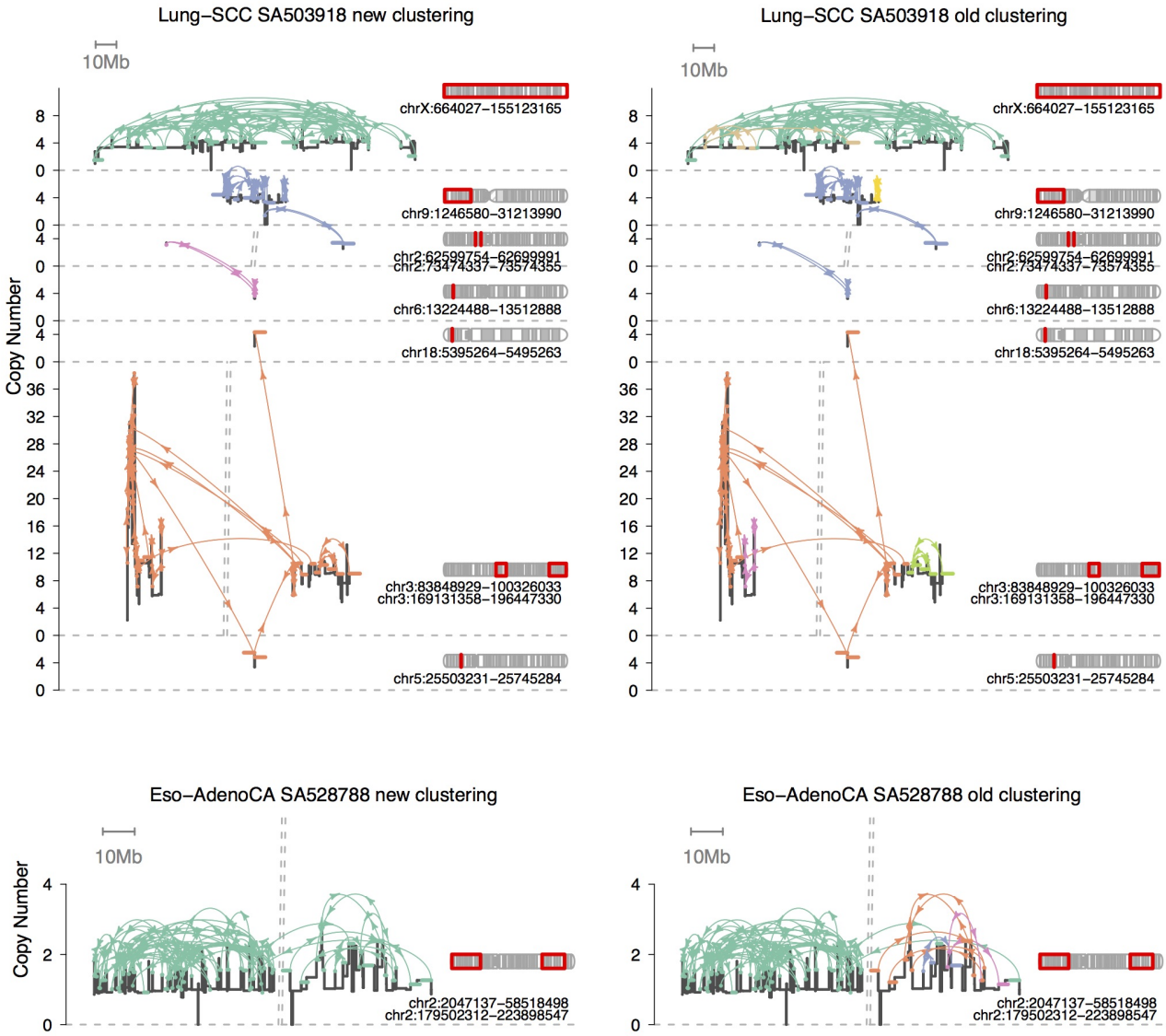


Figure D.24: New and old clustering schemes (fewer clusters in new) for complex unexplained BPJ in lung squamous cell cancer sample SA503918 and esophageal cancer sample SA528788, only showing clusters with disagreement.

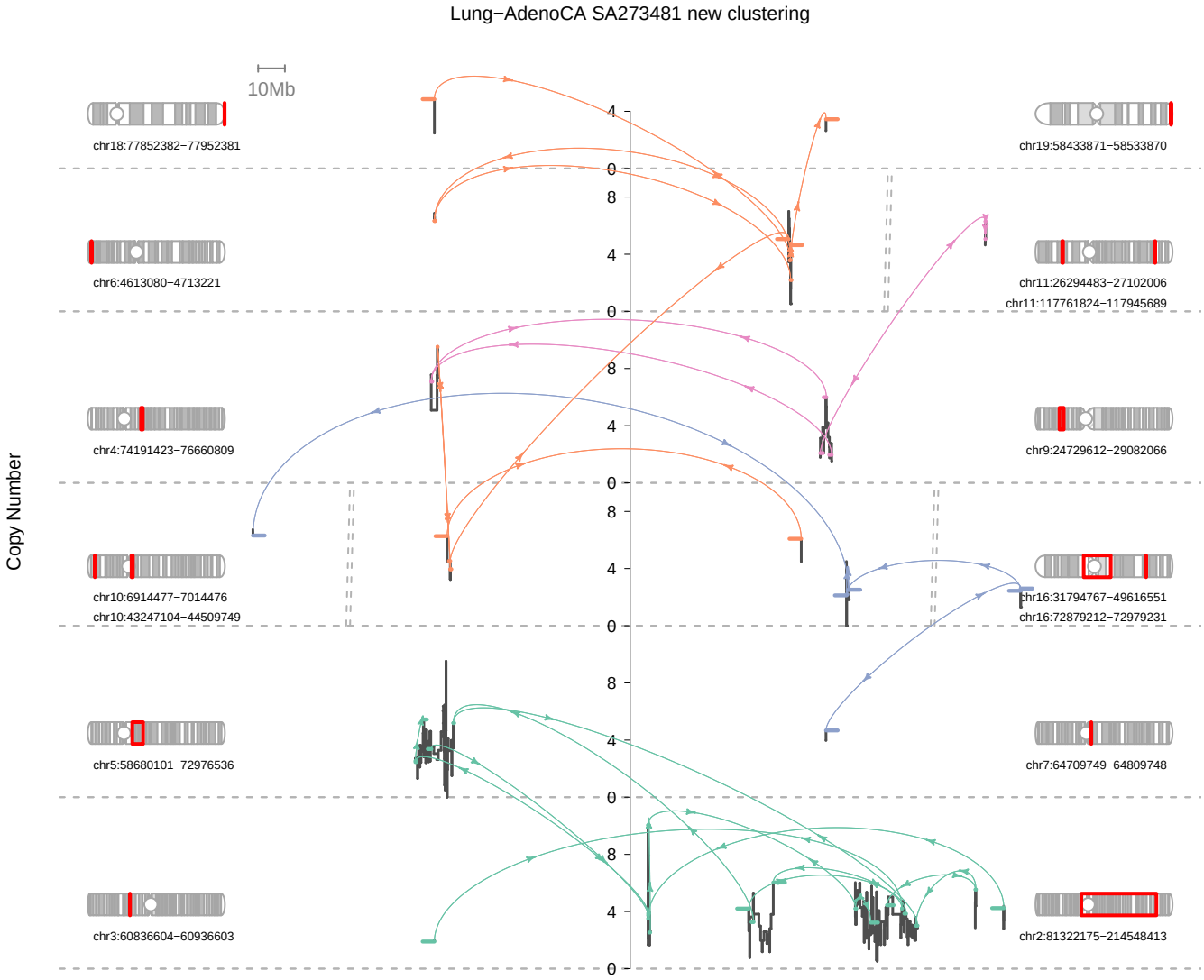


Figure D.25: New and old clustering schemes (fewer clusters in new) for complex unexplained BPJ in lung cancer sample SA273481, only showing clusters with disagreement. Figure continues on the next page.

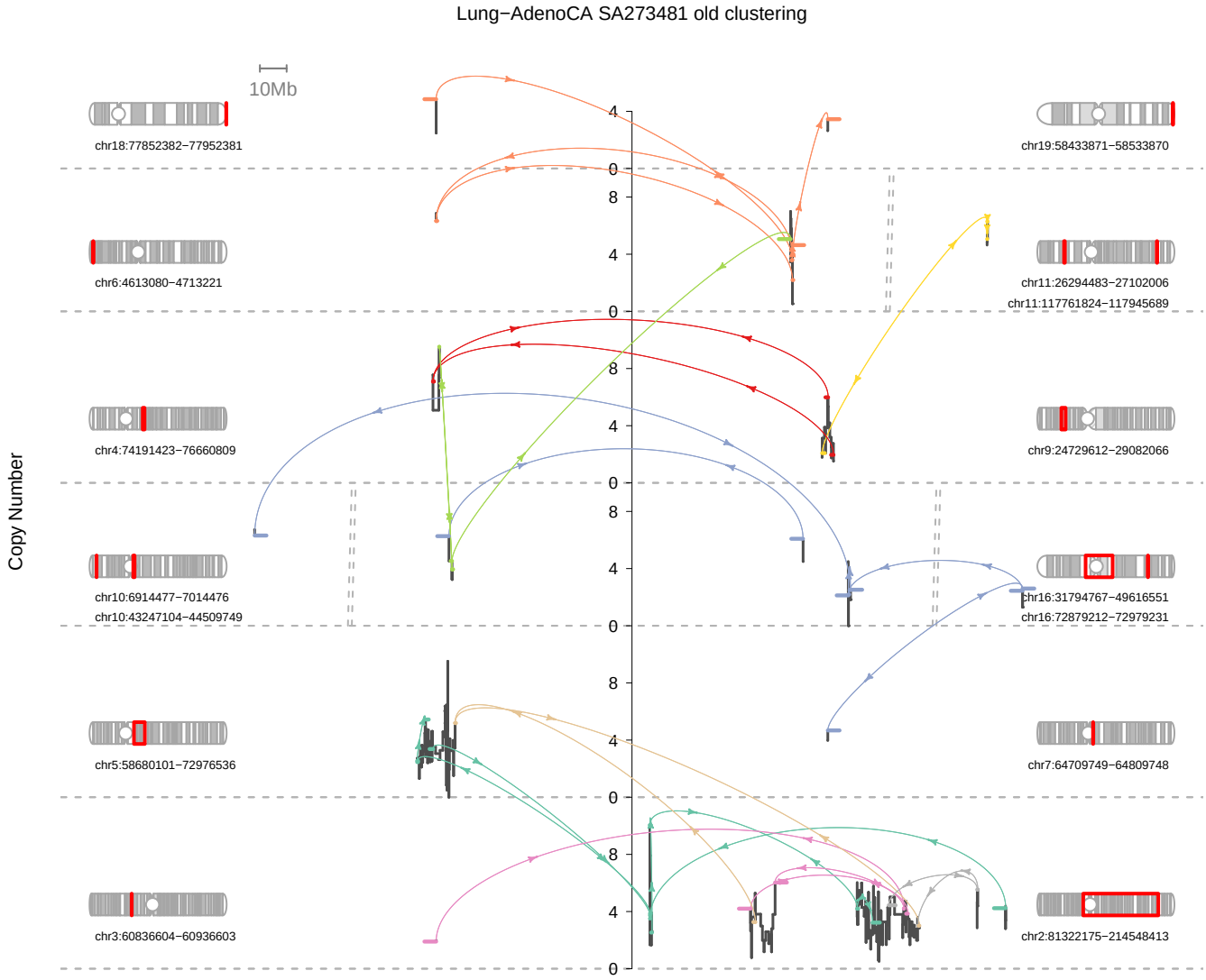


Figure D.25: New and old clustering schemes (fewer clusters in new) for complex unexplained BPJ in lung cancer sample SA273481, only showing clusters with disagreement. Figure is continued from the previous page.

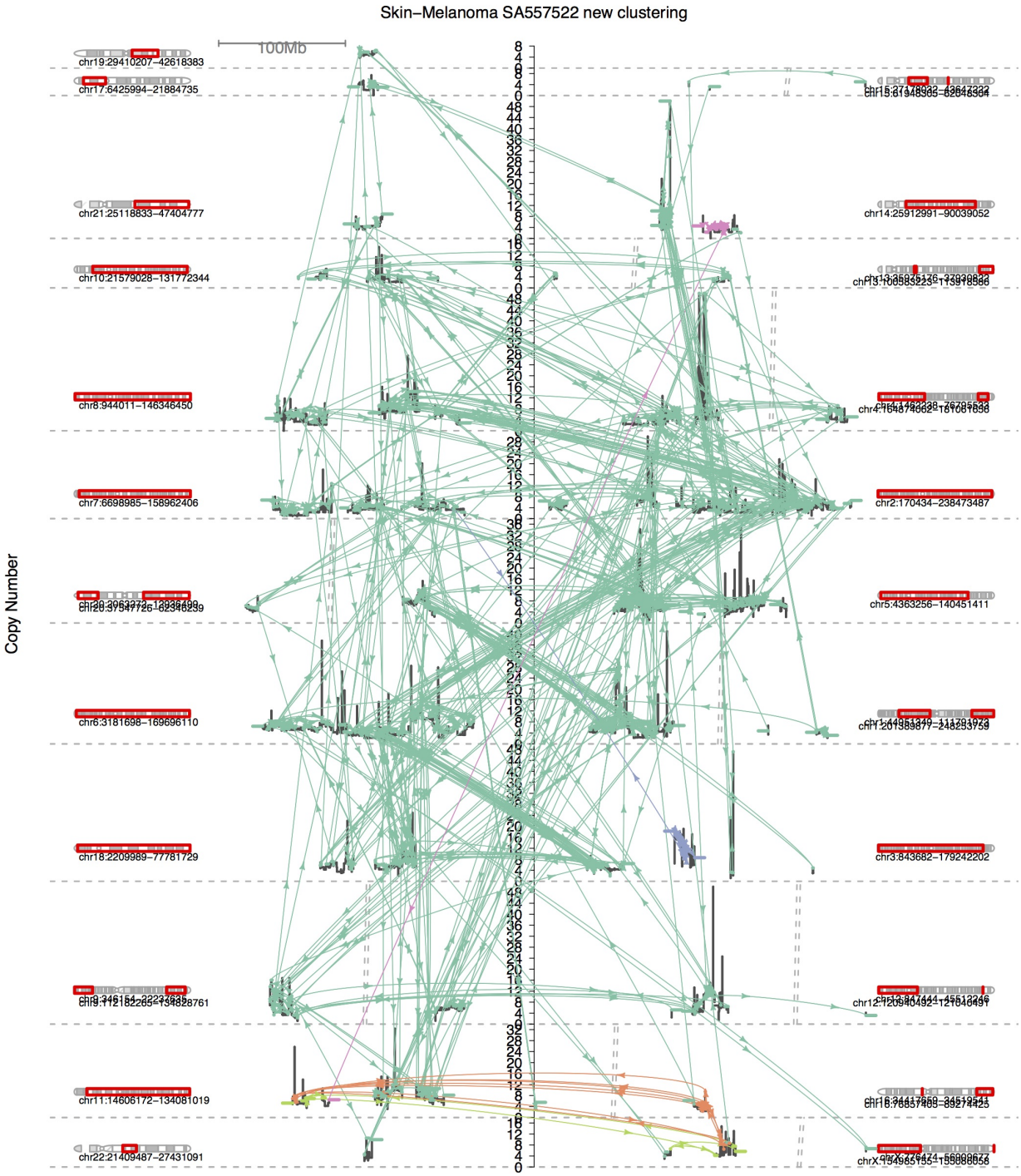


Figure D.26: New and old clustering schemes (fewer clusters in new) for complex unexplained BPJ in melanoma sample SA557522, only showing clusters with disagreement. Figure continues on the next page.

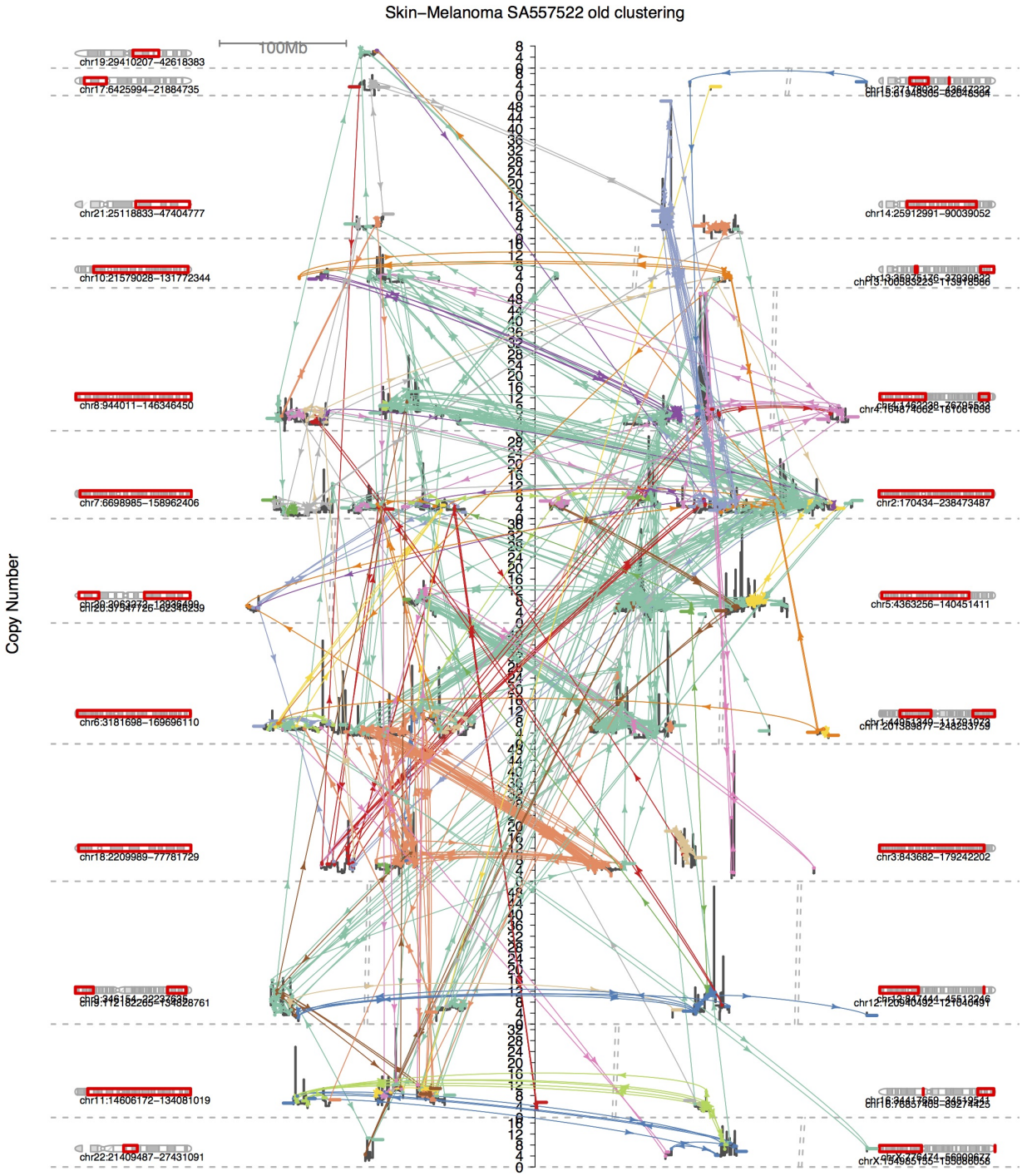


Figure D.26: New and old clustering schemes (fewer clusters in new) for complex unexplained BPJ in melanoma sample SA557522, only showing clusters with disagreement. Figure is continued from the previous page.

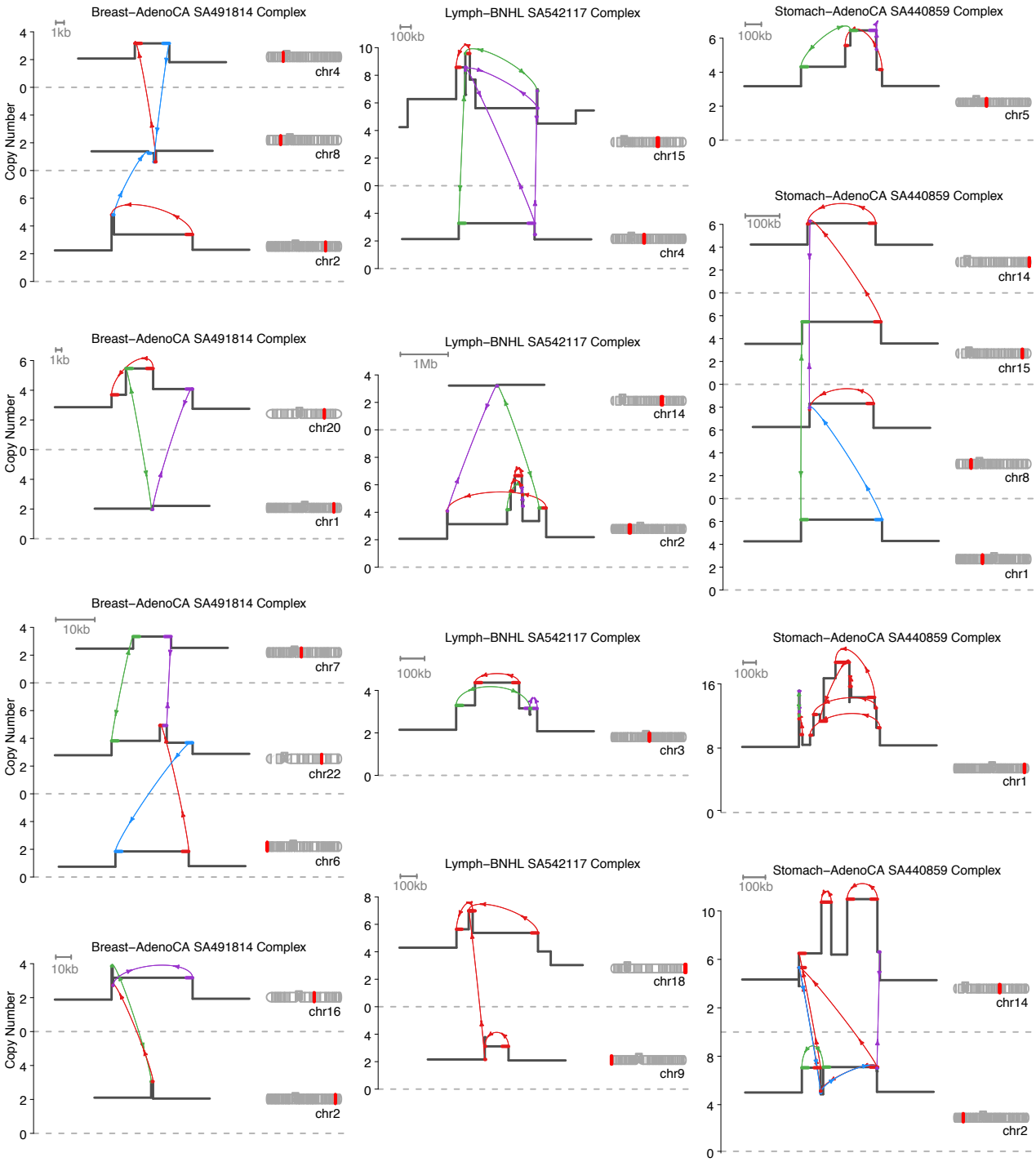


Figure D.27: Four example events from each of three outlying samples containing more than 30 separate complex clusters. These samples are further summarised in Figure 5.16.

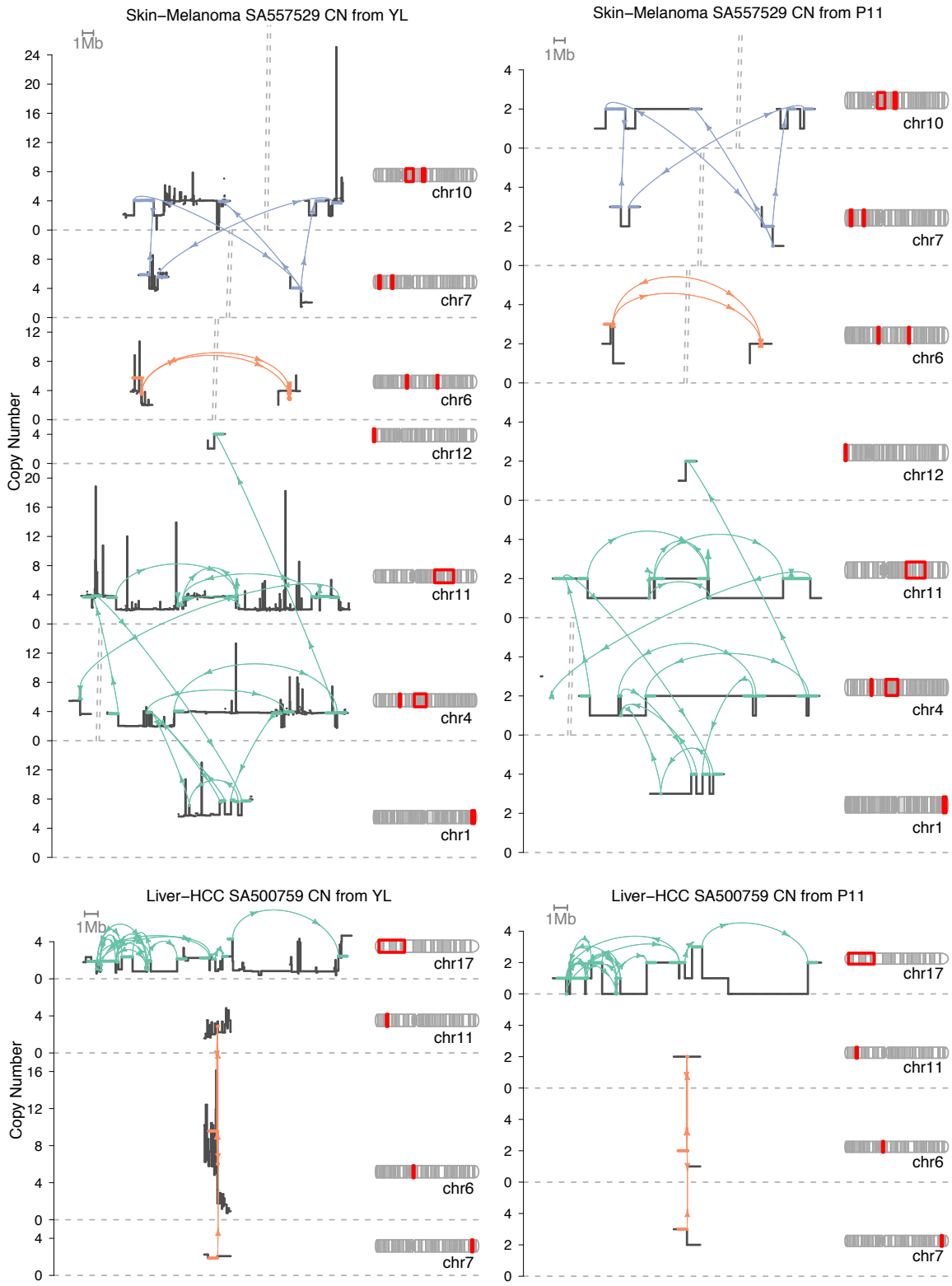


Figure D.28: Comparison of CN calls returned by YL (left) and P11 (right) around complex unexplained BPJ in samples qualifying for a switch to P11 CN. Breakpoint junctions are coloured by cluster assignment within the sample. Figure continues on the next page.

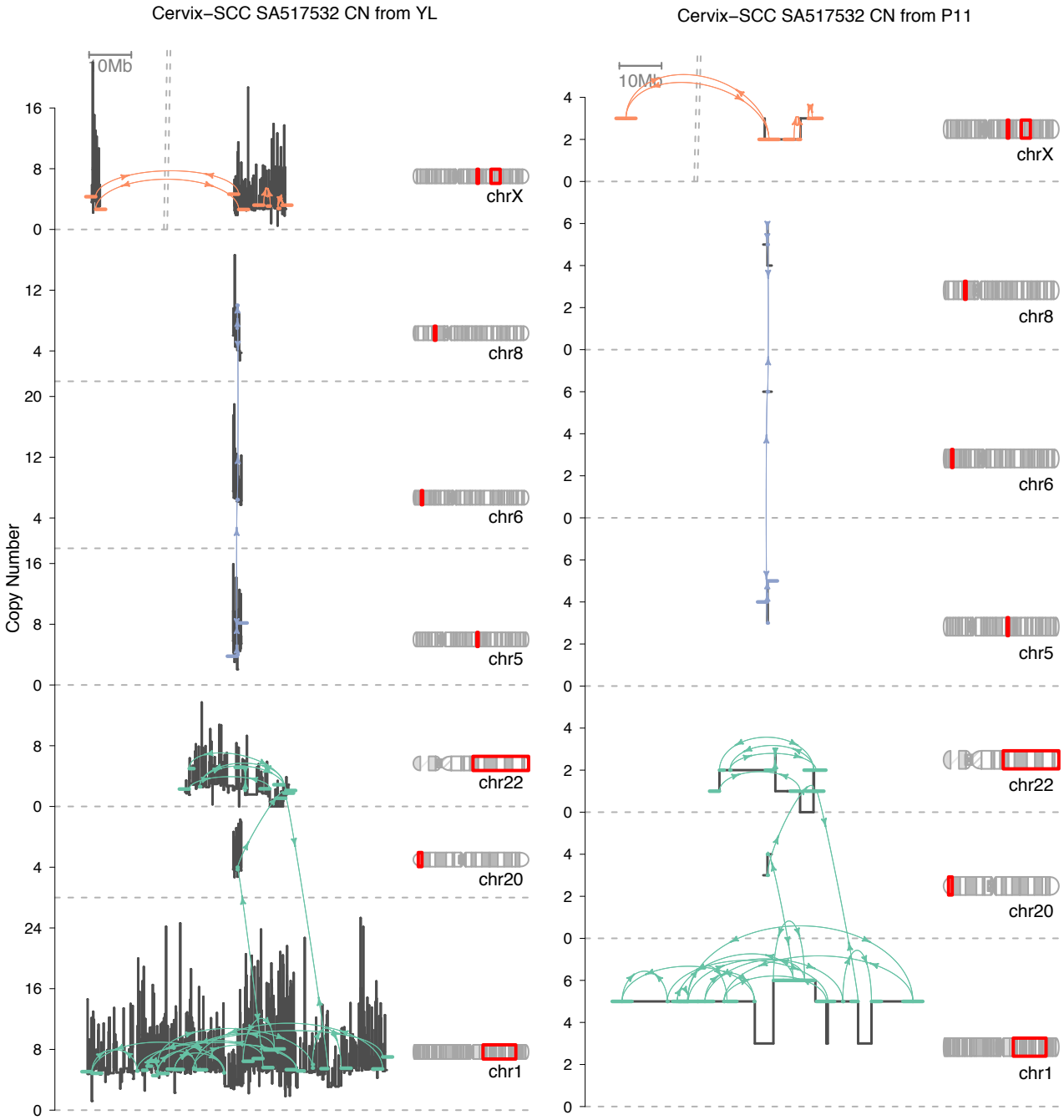


Figure D.28: Comparison of CN calls returned by YL (left) and P11 (right) around complex unexplained BPJ in samples qualifying for a switch to P11 CN. Breakpoint junctions are coloured by cluster assignment within the sample. Figure continued from the previous page.

Appendix E

Supplementary Tables

Table E.1: ROADMAP cell lines chosen to estimate tissue-specific epigenomic properties for PCAWG tissues. Tissues without a close match in ROADMAP are instead matched to the average over many epithelial cell types. Details of the cell lines are available in Roadmap Epigenomics Consortium et al. (2015).

PCAWG tissue group	Matching ROADMAP cell line IDs
Biliary	E028, E065, E076, E079, E094, E096, E098, E109, E126, E127
Bladder	E028, E065, E076, E079, E094, E096, E098, E109, E126, E127
BoneSoftTissue	E025, E107, E108, E129
Breast	E027, E028, E119
Cervix	E117
CNS	E067, E068, E069, E070, E071, E072, E073, E074
ColonRectum	E075, E076, E102, E103
Esophagus	E079
HeadNeck	E079
Kidney	E086
Liver	E066
Lung	E088, E096, E128
Lymphoid	E032, E034
Myeloid	E029, E030
Ovary	E097
Pancreas	E087, E098
Prostate	E028, E065, E076, E079, E094, E096, E098, E109, E126, E127
Skin	E059, E061, E126, E127
Stomach	E094, E110, E111
Thyroid	E080
Uterus	E028, E065, E076, E079, E094, E096, E098, E109, E126, E127

Table E.2: Fragile site definitions for the PCAWG cohort

Chr	Start	End	CFS	Gene	Note
chr1	71750001	72900000	FRA1L	NEGR1	
chr1	245000001	247100000	FRA1I	KIF26B;SMYD3	
chr2	140900001	143100000	FRA2F	LRP1B	
chr3	59350001	61750000	FRA3B	FHIT	
chr3	115600001	117450000	FRA3L	LSAMP	
chr3	173850001	175900000	FRA3O	NAALADL2	
chr4	90750001	92800000	FRA4F	CCSER1	
chr5	57900001	60200000	FRA5H	PDE4D	
chr6	161900001	163650000	FRA6E	PACRG;PARK2	
chr7	68850001	70700000	FRA7J	AUTS2	
chr7	109400001	111600000	FRA7K	IMMP2L	
chr8	2750001	4600000	no CFS name	CSMD1	
chr9	8500001	10450000	no CFS name	PTPRD	
chr10	52550001	53950000	FRA10G;FRA10C	PRKG1	
chr10	67750001	68750000	FRA10D	CTNNA3	
chr13	94050001	95100000	FRA13H;FRA13D	GPC6	
chr16	5800001	7600000	no CFS name	RBFOX1	
chr16	77750001	79650000	FRA16D	WWOX	
chr20	13700001	16250000	FRA20B	MACROD2	
chrX	31000001	33850000	FRAXC	DMD	
chrX	95800001	97100000	FRAXL	DIAPH2	
chr2	77350001	78350000	no CFS name	no long gene	excluded
chr2	186500001	188000000	FRA2H	no long gene	excluded
chr4	19050001	20100000	FRA4D	no long gene	excluded
chr4	181000001	183100000	no CFS name	no long gene	excluded
chr18	36600001	37600000	FRA18A	no long gene	excluded
chrX	6400001	8250000	FRAXB	no long gene	excluded

Table E.3: Sample counts of somatic SNV dataset from original mutational signatures discovery project by Alexandrov et al. (2013b).

Cancer	Exomes	Genomes	Total SNV
ALL	140	0	1562
AML	147	7	4903
Bladder	136	0	36390
Breast	844	119	687514
Cervix	38	0	7563
CLL	103	28	53513
Colorectum	559	0	204630
Esophageal	146	0	24861
Glioblastoma	98	0	3508
Glioma Low Grade	217	0	20601
Head and Neck	380	0	56078
Kidney Chromophobe	65	0	1287
Kidney Clear Cell	325	0	24999
Kidney Papillary	100	0	5489
Liver	0	88	850734
Lung Adeno	636	24	1658098
Lung Small Cell	70	0	13950
Lung Squamous	176	0	62412
Lymphoma B-cell	24	24	128212
Medulloblastoma	0	100	124941
Melanoma	396	0	280918
Myeloma	69	0	3467
Neuroblastoma	210	0	4508
Ovary	471	0	22307
Pancreas	98	15	115645
Pilocytic Astrocytoma	0	101	10577
Prostate	330	0	15176
Stomach	212	0	77345
Thyroid	304	0	4910
Uterus	241	0	163742
Total	6535	506	4669840

Bibliography

Akdemir, Kadir C, Yilong Li, Roel G Verhaak, Rameen Beroukhim, Peter Cambell, Lynda Chin, and Andrew Futreal (2017). "Spatial Genome Organization as a Framework for Somatic Alterations in Human Cancer". *bioRxiv*. DOI: 10.1101/179176.

Alaei-Mahabadi, Babak, Joydeep Bhadury, Joakim W Karlsson, Jonas A Nilsson, and Erik Larsson (2016). "Global analysis of somatic structural genomic alterations and their impact on gene expression in diverse human cancers". *Proc. Natl. Acad. Sci. U. S. A.* DOI: 10.1073/pnas.1606220113.

Alexandrov, Ludmil B, Serena Nik-Zainal, David C Wedge, Peter J Campbell, and Michael R Stratton (2013a). "Deciphering signatures of mutational processes operative in human cancer". *Cell Rep.* 3.1, pp. 246–259. DOI: 10.1016/j.celrep.2012.12.008.

Alexandrov, Ludmil B, Serena Nik-Zainal, David C Wedge, Samuel A J R Aparicio, Sam Behjati, Andrew V Biankin, Graham R Bignell, Niccolò Bolli, Ake Borg, et al. (2013b). "Signatures of mutational processes in human cancer". *Nature* 500.7463, pp. 415–421. DOI: 10.1038/nature12477.

Alexandrov, Ludmil B, Serena Nik-Zainal, Hoi Cheong Siu, Suet Yi Leung, and Michael R Stratton (2015a). "A mutational signature in gastric cancer suggests therapeutic strategies". *Nat. Commun.* 6, p. 8683. DOI: 10.1038/ncomms9683.

Alexandrov, Ludmil B, Philip H Jones, David C Wedge, Julian E Sale, Peter J Campbell, Serena Nik-Zainal, and Michael R Stratton (2015b). "Clock-like mutational processes in human somatic cells". *Nat. Genet.* 47.12, pp. 1402–1407. DOI: 10.1038/ng.3441.

Alt, Frederick W, Yu Zhang, Fei-Long Meng, Chunguang Guo, and Bjoern Schwer (2013). "Mechanisms of programmed DNA lesions and genomic instability in the immune system". *Cell* 152.3, pp. 417–429. DOI: 10.1016/j.cell.2013.01.007.

Anand, Ranjith, Annette Beach, Kevin Li, and James Haber (2017). "Rad51-mediated double-strand break repair and mismatch correction of divergent substrates". *Nature* 544.7650, pp. 377–380. DOI: 10.1038/nature22046.

Baca, Sylvan C, Davide Prandi, Michael S Lawrence, Juan Miguel Mosquera, Alessandro Romanel, Yotam Drier, Kyung Park, Naoki Kitabayashi, Theresa Y MacDonald, et al. (2013). "Punctuated evolution of prostate

cancer genomes". *Cell* 153.3, pp. 666–677. DOI: 10.1016/j.cell.2013.03.021.

Bacolla, Albino and Robert D Wells (2009). "Non-B DNA conformations as determinants of mutagenesis and human disease". *Mol. Carcinog.* 48.4, pp. 273–285. DOI: 10.1002/mc.20507.

Bacolla, Albino, John A Tainer, Karen M Vasquez, and David N Cooper (2016). "Translocation and deletion breakpoints in cancer genomes are associated with potential non-B DNA-forming sequences". *Nucleic Acids Res.* 44.12, pp. 5673–5688. DOI: 10.1093/nar/gkw261.

Bailey, Peter, David K Chang, Katia Nones, Amber L Johns, Ann-Marie Patch, Marie-Claude Gingras, David K Miller, Angelika N Christ, Tim J C Bruxner, et al. (2016). "Genomic analyses identify molecular subtypes of pancreatic cancer". *Nature* 531.7592, pp. 47–52. DOI: 10.1038/nature16965.

Baranowski, Rafal, Yining Chen, and Piotr Fryzlewicz (2016). "Narrowest-Over-Threshold Detection of Multiple Change-points and Change-point-like Features". arXiv: 1609.00293 [stat.ME].

Barthel, Floris P, Wei Wei, Ming Tang, Emmanuel Martinez-Ledesma, Xin Hu, Samirkumar B Amin, Kadir C Akdemir, Sahil Seth, Xingzhi Song, et al. (2017). "Systematic analysis of telomere length and somatic alterations in 31 cancer types". *Nat. Genet.* 49.3, pp. 349–357. DOI: 10.1038/ng.3781.

Bell, Robert J A, H Tomas Rube, Ana Xavier-Magalhães, Bruno M Costa, Andrew Mancini, Jun S Song, and Joseph F Costello (2016). "Understanding TERT Promoter Mutations: A Common Path to Immortality". *Mol. Cancer Res.* 14.4, pp. 315–323. DOI: 10.1158/1541-7786.MCR-16-0003.

Berger, Michael F, Michael S Lawrence, Francesca Demichelis, Yotam Drier, Kristian Cibulskis, Andrey Y Sivachenko, Andrea Sboner, Raquel Esgueva, Dorothee Pflueger, et al. (2011). "The genomic complexity of primary human prostate cancer". *Nature* 470.7333, pp. 214–220. DOI: 10.1038/nature09744.

Beroukhim, Rameen, Craig H Mermel, Dale Porter, Guo Wei, Soumya Raychaudhuri, Jerry Donovan, Jordi Barretina, Jesse S Boehm, Jennifer Dobson, et al. (2010). "The landscape of somatic copy-number alteration across human cancers". *Nature* 463.7283, pp. 899–905. DOI: 10.1038/nature08822.

Bignell, Graham R, Jing Huang, Joel Greshock, Stephen Watt, Adam Butler, Sofie West, Mira Grigorova, Keith W Jones, Wen Wei, et al. (2004). "High-resolution analysis of DNA copy number using oligonucleotide microarrays". *Genome Res.* 14.2, pp. 287–295. DOI: 10.1101/gr.2012304.

Bignell, Graham R, Chris D Greenman, Helen Davies, Adam P Butler, Sarah Edkins, Jenny M Andrews, Gemma Buck, Lina Chen, David Beare, et al. (2010). "Signatures of mutation and selection in the cancer genome". *Nature* 463.7283, pp. 893–898. DOI: 10.1038/nature08768.

Bignold, Leon P, Brian L D Coghlan, and Hubertus P A Jersmann (2006). "Hansemann, Boveri, chromosomes and the gametogenesis-related theories of tumours". *Cell Biol. Int.* 30.7, pp. 640–644. DOI: 10.1016/j.cellbi.2006.04.002.

Black, Joshua C, Amity L Manning, Capucine Van Rechem, Jaegil Kim, Brendon Ladd, Juok Cho, Cristiana M Pineda, Nancy Murphy, Danette L Daniels, et al. (2013). "KDM4A Lysine Demethylase Induces Site-Specific Copy Gain and Rereplication of Regions Amplified in Tumors". *Cell* 154.3, pp. 541–555. DOI: 10.1016/j.cell.2013.06.051.

Blei, David M and John D Lafferty (2007). "A correlated topic model of Science". *Ann. Appl. Stat.* 1.1, pp. 17–35. DOI: 10.1214/07-AOAS114.

Blei, David M, Alp Kucukelbir, and Jon D McAuliffe (2017). "Variational Inference: A Review for Statisticians". *J. Am. Stat. Assoc.* 112.518, pp. 859–877. DOI: 10.1080/01621459.2017.1285773.

Blei, David (2012). "Probabilistic topic models". *Commun. ACM* 55.4, pp. 77–84. DOI: 10.1109/MSP.2010.938079.

Bower, Hannah, Magnus Björkholm, Paul W Dickman, Martin Höglund, Paul C Lambert, and Therese M-L Andersson (2016). "Life Expectancy of Patients With Chronic Myeloid Leukemia Approaches the Life Expectancy of the General Population". *J. Clin. Oncol.* 34.24, pp. 2851–2857. DOI: 10.1200/JCO.2015.66.2866.

Branzei, Dana and Marco Foiani (2010). "Maintaining genome stability at the replication fork". *Nat. Rev. Mol. Cell Biol.* 11.3, pp. 208–219. DOI: 10.1038/nrm2852.

Burman, Bharat, Zhuzhu Z Zhang, Gianluca Pegoraro, Jason D Lieb, and Tom Misteli (2015). "Histone modifications predispose genome regions to breakage and translocation". *Genes Dev.* 29.13, pp. 1393–1402. DOI: 10.1101/gad.262170.115.

Burrell, Rebecca A, Sarah E McClelland, David Endesfelder, Petra Groth, Marie-Christine Weller, Nadeem Shaikh, Enric Domingo, Nnennaya Kanu, Sally M Dewhurst, et al. (2013). "Replication stress links structural and numerical cancer chromosomal instability". *Nature* 494.7438, pp. 492–496. DOI: 10.1038/nature11935.

Cai, Haoyang, Nitin Kumar, Homayoun C Bagheri, Christian von Mering, Mark D Robinson, and Michael Baudis (2014). "Chromothripsis-like patterns are recurring but heterogeneously distributed features in a survey of 22,347 cancer genome screens". *BMC Genomics* 15, p. 82. DOI: 10.1186/1471-2164-15-82.

Campbell, Brittany B, Nicholas Light, David Fabrizio, Matthew Zatzman, Fabio Fuligni, Richard de Borja, Scott Davidson, Melissa Edwards, Julia A Elvin, et al. (2017a). "Comprehensive Analysis of Hypermutation in Human Cancer". *Cell.* DOI: 10.1016/j.cell.2017.09.048.

Campbell, Peter J, Philip J Stephens, Erin D Pleasance, Sarah O'Meara, Heng Li, Thomas Santarius, Lucy A Stebbings, Catherine Leroy, Sarah Edkins, et al. (2008). "Identification of somatically acquired rearrangements in cancer using genome-wide massively parallel paired-end sequencing". *Nat. Genet.* 40.6, pp. 722–729. DOI: 10.1038/ng.128.

Campbell, Peter J, Gad Getz, Joshua M Stuart, Jan O Korbel, Lincoln D Stein, and ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Network (2017b). "Pan-cancer analysis of whole genomes". *bioRxiv.* DOI: 10.1101/162784.

Cancer Genome Atlas Research Network (2011). "Integrated genomic analyses of ovarian carcinoma". *Nature* 474.7353, pp. 609–615. DOI: 10.1038/nature10166.

– (2012). "Comprehensive molecular characterization of human colon and rectal cancer". *Nature* 487.7407, pp. 330–337. DOI: 10.1038/nature11252.

Cancer Genome Atlas Research Network, John N Weinstein, Eric A Collisson, Gordon B Mills, Kenna R Mills Shaw, Brad A Ozenberger, Kyle Ellrott, Ilya Shmulevich, Chris Sander, and Joshua M Stuart (2013). "The Cancer Genome Atlas Pan-Cancer analysis project". *Nat. Genet.* 45.10, pp. 1113–1120. DOI: 10.1038/ng.2764.

Cancer Genome Project (2017). *BRASS*. https://github.com/cancerit/BRASS. Accessed: 2017-11-1.

Carvalho, Claudia M B and James R Lupski (2016). "Mechanisms underlying structural variant formation in genomic disorders". *Nat. Rev. Genet.* 17.4, pp. 224–238. DOI: 10.1038/nrg.2015.25.

Carvalho, Claudia M B, Melissa B Ramocki, Davut Pehlivan, Luis M Franco, Claudia Gonzaga-Jauregui, Ping Fang, Alanna McCall, Eniko Karman Pivnick, Stacy Hines-Dowell, et al. (2011). "Inverted genomic segments and complex triplication rearrangements are mediated by inverted repeats in the human genome". *Nat. Genet.* 43.11, pp. 1074–1081. DOI: 10.1038/ng.944.

Carvalho, Claudia M B, Davut Pehlivan, Melissa B Ramocki, Ping Fang, Benjamin Alleva, Luis M Franco, John W Belmont, P J Hastings, and James R Lupski (2013). "Replicative mechanisms for CNV formation are error prone". *Nat. Genet.* 45.11, pp. 1319–1326. DOI: 10.1038/ng.2768.

Carvalho, Claudia M B, Rolph Pfundt, Daniel A King, Sarah J Lindsay, Luciana W Zuccherato, Merryn V E Macville, Pengfei Liu, Diana Johnson, Pawel Stankiewicz, et al. (2015). "Absence of heterozygosity due to template switching during replicative rearrangements". *Am. J. Hum. Genet.* 96.4, pp. 555–564. DOI: 10.1016/j.ajhg.2015.01.021.

Ceccaldi, Raphael, Beatrice Rondinelli, and Alan D D'Andrea (2016). "Repair Pathway Choices and Consequences at the Double-Strand Break". *Trends Cell Biol.* 26.1, pp. 52–64. DOI: 10.1016/j.tcb.2015.07.009.

Cer, Regina Z, Duncan E Donohue, Uma S Mudunuri, Nuri A Temiz, Michael A Loss, Nathan J Starner, Goran N Halusa, Natalia Volfovsky, Ming Yi, et al. (2013). "Non-B DB v2.0: a database of predicted non-B DNA-forming motifs and its associated tools". *Nucleic Acids Res.* 41.Database issue, pp. D94–D100. DOI: 10.1093/nar/gks955.

Chaisson, Mark J P, Ashley D Sanders, Xuefang Zhao, Ankit Malhotra, David Porubsky, Tobias Rausch, Eugene J Gardner, Oscar Rodriguez, Li Guo, et al. (2017). "Multi-platform discovery of haplotype-resolved structural variation in human genomes". *bioRxiv*. DOI: 10.1101/193144.

Chan, Eva K F, Desiree C Petersen, Ruth J Lyons, Benedetta F Baldi, Anthony T Papenfuss, David M Thomas, and Vanessa M Hayes (2017). "Whole genome optical mapping reveals multiple fusion events chained by large novel sequences in cancer". *bioRxiv*. DOI: 10.1101/166173.

Chen, Xi, Xu Shi, Leena Hilakivi-Clarke, Ayesha N Shajahan-Haq, Robert Clarke, and Jianhua Xuan (2016). "PSSV: A novel pattern-based probabilistic approach for somatic structural variation identification". *Bioinformatics*. DOI: 10.1093/bioinformatics/btw605.

Chong, Zechen, Jue Ruan, Min Gao, Wanding Zhou, Tenghui Chen, Xian Fan, Li Ding, Anna Y Lee, Paul Boutros, et al. (2016). "novoBreak: local assembly for breakpoint detection in cancer genomes". *Nat. Methods*. DOI: 10.1038/nmeth.4084.

Chouldechova, Alexandra and Trevor Hastie (2015). "Generalized Additive Model Selection". *arXiv [stat.ML]*. arXiv: 1506.03850 [stat.ML].

Cmero, Marek, Cheng Soon Ong, Ke Yuan, Jan Schröder, Kangbo Mo, PCAWG Evolution and Heterogeneity Working Group, Niall M Corcoran, Anthony T Papenfuss, Christopher M Hovens, et al. (2017). "SVclone: inferring structural variant cancer cell fraction". *bioRxiv*. DOI: 10.1101/172486.

Collins, Ryan L, Harrison Brand, Claire E Redin, Carrie Hanscom, Caroline Antolik, Matthew R Stone, Joseph T Glessner, Tamara Mason, Giulia Pregno, et al. (2017). "Defining the diverse spectrum of inversions, complex structural variation, and chromothripsis in the morbid human genome". *Genome Biol.* 18.1, p. 36. DOI: 10.1186/s13059-017-1158-6.

Cortez, David (2015). "Preventing replication fork collapse to maintain genome integrity". *DNA Repair* 32, pp. 149–157. DOI: 10.1016/j.dnarep.2015.04.026.

Costantino, Lorenzo, Sotirios K Sotiriou, Juha K Rantala, Simon Magin, Emil Mladenov, Thomas Helleday, James E Haber, George Iliakis, Olli P Kallioniemi, and Thanos D Halazonetis (2014). "Break-induced replication repair of damaged forks induces genomic duplications in human cells". *Science* 343.6166, pp. 88–91. DOI: 10.1126/science.1243211.

Cox, D, C Yuncken, and A I Spriggs (1965). "Minute Chromatin Bodies in Malignant Tumours of Childhood". *Lancet* 1.7402, pp. 55–58.

Davies, Helen, Dominik Glodzik, Sandro Morganella, Lucy R Yates, Johan Staaf, Xueqing Zou, Manasa Ramakrishna, Sancha Martin, Sandrine Boyault, et al. (2017). "HRDetect is a predictor of BRCA1 and BRCA2 deficiency based on mutational signatures". *Nat. Med.* DOI: 10.1038/nm.4292.

Davis, Caleb F, Christopher J Ricketts, Min Wang, Lixing Yang, Andrew D Cherniack, Hui Shen, Christian Buhay, Hyojin Kang, Sang Cheol Kim, et al. (2014). "The somatic genomic landscape of chromophobe renal cell carcinoma". *Cancer Cell* 26.3, pp. 319–330. DOI: 10.1016/j.ccr.2014.07.014.

Davoli, Teresa, Andrew Wei Xu, Kristen E Mengwasser, Laura M Sack, John C Yoon, Peter J Park, and Stephen J Elledge (2013). "Cumulative haploinsufficiency and triplosensitivity drive aneuploidy patterns and shape the cancer genome". *Cell* 155.4, pp. 948–962. DOI: 10.1016/j.cell.2013.10.011.

De, Subhajyoti and Franziska Michor (2011). "DNA replication timing and long-range DNA interactions predict mutational landscapes of cancer genomes". *Nat. Biotechnol.* 29.12, pp. 1103–1108. DOI: 10.1038/nbt.2030.

Dentro, Stefan C, Ignaty Leshchiner, Kerstin Haase, Jeff Wintersinger, Amit G Deshwar, Maxime Tarabichi, Yulia Rubanova, Kaixian Yu, Ignacio Vázquez-García, et al. (2017). "Pervasive intra-tumour heterogeneity and subclonal selection across cancer types". *Manuscript in preparation.*

Dixon, Jesse R, Siddarth Selvaraj, Feng Yue, Audrey Kim, Yan Li, Yin Shen, Ming Hu, Jun S Liu, and Bing Ren (2012). "Topological domains in mammalian genomes identified by analysis of chromatin interactions". *Nature* 485.7398, pp. 376–380. DOI: 10.1038/nature11082.

Drier, Yotam, Michael S Lawrence, Scott L Carter, Chip Stewart, Stacey B Gabriel, Eric S Lander, Matthew Meyerson, Rameen Beroukhim, and Gad Getz (2013). "Somatic rearrangements across cancer reveal classes of samples with distinct patterns of DNA breakage and rearrangement-induced hypermutability". *Genome Res.* 23.2, pp. 228–235. DOI: 10.1101/gr.141382.112.

Drost, Jarno, Ruben van Boxtel, Francis Blokzijl, Tomohiro Mizutani, Nobuo Sasaki, Valentina Sasselli, Joep de Ligt, Sam Behjati, Judith E Grolleman, et al. (2017). "Use of CRISPR-modified human stem cell organoids to study the origin of mutational signatures in cancer". *Science.* DOI: 10.1126/science.aao3130.

Dulbecco, R (1986). "A turning point in cancer research: sequencing the human genome". *Science* 231.4742, pp. 1055–1056.

Dyson, Nicholas J (2016). "RB1: a prototype tumor suppressor and an enigma". *Genes Dev.* 30.13, pp. 1492–1502. DOI: 10.1101/gad.282145.116.

ENCODE Project Consortium (2012). "An integrated encyclopedia of DNA elements in the human genome". *Nature* 489.7414, pp. 57–74. DOI: 10.1038/nature11247.

Ernst, Jason and Manolis Kellis (2015). "Large-scale imputation of epigenomic datasets for systematic annotation of diverse human tissues". *Nat. Biotechnol.* 33.4, pp. 364–376. DOI: 10.1038/nbt.3157.

Fehrmann, Rudolf S N, Juha M Karjalainen, Małgorzata Krajewska, Harm-Jan Westra, David Maloney, Anton Simeonov, Tune H Pers, Joel N Hirschhorn, Ritsert C Jansen, et al. (2015). "Gene expression analysis identifies global gene dosage sensitivity in cancer". *Nat. Genet.* 47.2, pp. 115–125. DOI: 10.1038/ng.3173.

Finn, Kenneth J and Joachim J Li (2013). "Single-stranded annealing induced by re-initiation of replication origins provides a novel and efficient mechanism for generating copy number expansion via non-allelic homologous recombination". *PLoS Genet.* 9.1, e1003192. DOI: 10.1371/journal.pgen.1003192.

Fischer, Andrej, Christopher J R Illingworth, Peter J Campbell, and Ville Mustonen (2013). "EMu: probabilistic inference of mutational processes and their localization in the cancer genome". *Genome Biol.* 14.4, R39. DOI: 10.1186/gb-2013-14-4-r39.

Fonseca, Nuno A, Andre Kahles, Kjong-Van Lehmann, Claudia Calabrese, Aurelien Chateigner, Natalie R Davidson, Deniz Demircioğlu, Yao He, Fabien C Lamaze, et al. (2017). "Pan-cancer study of heterogeneous RNA aberrations". *bioRxiv*. DOI: 10.1101/183889.

Forbes, Simon A, David Beare, Prasad Gunasekaran, Kenric Leung, Nidhi Bindal, Harry Boutselakis, Minjie Ding, Sally Bamford, Charlotte Cole, et al. (2015). "COSMIC: exploring the world's knowledge of somatic mutations in human cancer". *Nucleic Acids Res.* 43.Database issue, pp. D805–11. DOI: 10.1093/nar/gku1075.

Franke, Martin, Daniel M Ibrahim, Guillaume Andrey, Wibke Schwarzer, Verena Heinrich, Robert Schöpflin, Katerina Kraft, Rieke Kempfer, Ivana Jerković, et al. (2016). "Formation of new chromatin domains determines pathogenicity of genomic duplications". *Nature*. DOI: 10.1038/nature19800.

Fraser, Michael, Veronica Y Sabelnykova, Takafumi N Yamaguchi, Lawrence E Heisler, Julie Livingstone, Vincent Huang, Yu-Jia Shiah, Fouad Yousif, Xihui Lin, et al. (2017). "Genomic hallmarks of localized, non-indolent prostate cancer". *Nature*. DOI: 10.1038/nature20788.

Frazer, Kelly A, Dennis G Ballinger, David R Cox, David A Hinds, Laura L Stuve, Richard A Gibbs, John W Belmont, Andrew Boudreau, Paul Hardenbol, et al. (2007). "A second generation human haplotype map of over 3.1 million SNPs". *Nature* 449.7164, pp. 851–861. DOI: 10.1038/nature06258.

Friedman, Jerome, Trevor Hastie, and Rob Tibshirani (2010). "Regularization Paths for Generalized Linear Models via Coordinate Descent". *J. Stat. Softw.* 33.1, pp. 1–22. DOI: 10.18637/jss.v033.i01.

Fudenberg, Geoff, Gad Getz, Matthew Meyerson, and Leonid A Mirny (2011). "High order chromatin architecture shapes the landscape of chromosomal alterations in cancer". *Nat. Biotechnol.* 29.12, pp. 1109–1113. DOI: 10.1038/nbt.2049.

Fujimoto, Akihiro, Mayuko Furuta, Yasushi Totoki, Tatsuhiko Tsunoda, Mamoru Kato, Yuichi Shiraishi, Hiroko Tanaka, Hiroaki Taniguchi, Yoshiiku Kawakami, et al. (2016). "Whole-genome mutational landscape and characterization of noncoding and structural mutations in liver cancer". *Nat. Genet.* 48.5, pp. 500–509. DOI: 10.1038/ng.3547.

Funk, Laura C, Lauren M Zasadil, and Beth A Weaver (2016). "Living in CIN: Mitotic Infidelity and Its Consequences for Tumor Promotion and Suppression". *Dev. Cell* 39.6, pp. 638–652. DOI: 10.1016/j.devcel.2016.10.023.

GTEx Consortium (2017). "Genetic effects on gene expression across human tissues". *Nature* 550.7675, pp. 204–213. DOI: 10.1038/nature24277.

Gaillard, Hélène, Tatiana García-Muse, and Andrés Aguilera (2015). "Replication stress and cancer". *Nat. Rev. Cancer* 15.5, pp. 276–289. DOI: 10.1038/nrc3916.

Garsed, Dale W, Owen J Marshall, Vincent D A Corbin, Arthur Hsu, Leon Di Stefano, Jan Schröder, Jason Li, Zhi-Ping Feng, Bo W Kim, et al.

(2014). "The architecture and evolution of cancer neochromosomes". *Cancer Cell* 26.5, pp. 653–667. DOI: 10.1016/j.ccell.2014.09.010.

Genomics England (2017). *The 100,000 Genomes Project Protocol v3*. DOI: 10.6084/m9.figshare.4530893.v2.

Gerstung, Moritz, Elli Papaemmanuil, Inigo Martincorena, Lars Bullinger, Verena I Gaidzik, Peter Paschka, Michael Heuser, Felicitas Thol, Niccolo Bolli, et al. (2017). "Precision oncology for acute myeloid leukemia using a knowledge bank approach". *Nat. Genet.* DOI: 10.1038/ng.3756.

Gibcus, Johan H and Job Dekker (2013). "The hierarchy of the 3D genome". *Mol. Cell* 49.5, pp. 773–782. DOI: 10.1016/j.molcel.2013.02.011.

Gisselsson, D, L Pettersson, M Höglund, M Heidenblad, L Gorunova, J Wiegant, F Mertens, P Dal Cin, F Mitelman, and N Mandahl (2000). "Chromosomal breakage-fusion-bridge events cause genetic intratumor heterogeneity". *Proc. Natl. Acad. Sci. U. S. A.* 97.10, pp. 5357–5362. DOI: 10.1073/pnas.090013497.

Glodzik, Dominik, Sandro Morganella, Helen Davies, Peter T Simpson, Yilong Li, Xueqing Zou, Javier Diez-Perez, Johan Staaf, Ludmil B Alexandrov, et al. (2017). "A somatic-mutational process recurrently duplicates germline susceptibility loci and tissue-specific super-enhancers in breast cancers". *Nat. Genet.* DOI: 10.1038/ng.3771.

Glover, Thomas W, Thomas E Wilson, and Martin F Arlt (2017). "Fragile sites in cancer: more than meets the eye". *Nat. Rev. Cancer* 17.8, pp. 489–501. DOI: 10.1038/nrc.2017.52.

Gong, Yongxing, Travis Ian Zack, Luc G T Morris, Kan Lin, Ellen Hukkelhoven, Radhika Raheja, I-Li Tan, Sevin Turcan, Selvaraju Veeriah, et al. (2014). "Pan-cancer genetic analysis identifies PARK2 as a master regulator of G1/S cyclins". *Nat. Genet.* 46.6, pp. 588–594. DOI: 10.1038/ng.2981.

Greenman, C D, S L Cooke, J Marshall, M R Stratton, and P J Campbell (2016). "Modeling the evolution space of breakage fusion bridge cycles with a stochastic folding process". *J. Math. Biol.* 72.1-2, pp. 47–86. DOI: 10.1007/s00285-015-0875-2.

Greer, Stephanie U, Lincoln D Nadauld, Billy T Lau, Jiamin Chen, Christina Wood-Bouwens, James M Ford, Calvin J Kuo, and Hanlee P Ji (2017). "Linked read sequencing resolves complex genomic rearrangements in gastric cancer metastases". *Genome Med.* 9.1, p. 57. DOI: 10.1186/s13073-017-0447-8.

Gröschel, Stefan, Mathijs A Sanders, Remco Hoogenboezem, Elzo de Wit, Britta A M Bouwman, Claudia Erpelinck, Vincent H J van der Velden, Marije Havermans, Roberto Avellino, et al. (2014). "A single oncogenic enhancer rearrangement causes concomitant EVI1 and GATA2 deregulation in leukemia". *Cell* 157.2, pp. 369–381. DOI: 10.1016/j.cell.2014.02.019.

Guelen, Lars, Ludo Pagie, Emilie Brasset, Wouter Meuleman, Marius B Faza, Wendy Talhout, Bert H Eussen, Annelies de Klein, Lodewyk Wessels, et al. (2008). "Domain organization of human chromosomes revealed by mapping of nuclear lamina interactions". *Nature* 453.7197, pp. 948–951. DOI: 10.1038/nature06947.

Haffner, Michael C, Martin J Aryee, Antoun Toubaji, David M Esopi, Roula Albadine, Bora Gurel, William B Isaacs, G Steven Bova, Wennuan Liu, et al. (2010). "Androgen-induced TOP2B-mediated double-strand breaks and prostate cancer gene rearrangements". *Nat. Genet.* 42.8, pp. 668–675. DOI: 10.1038/ng.613.

Hagerstrand, Daniel, Alexander Tong, Steven E Schumacher, Nina Ilic, Rhine R Shen, Hiu Wing Cheung, Francisca Vazquez, Yashaswi Shrestha, So Young Kim, et al. (2013). "Systematic interrogation of 3q26 identifies TLOC1 and SKIL as cancer drivers". *Cancer Discov.* 3.9, pp. 1044–1057. DOI: 10.1158/2159-8290.CD-12-0592.

Hakim, Ofir, Wolfgang Resch, Arito Yamane, Isaac Klein, Kyong-Rim Kieffer-Kwon, Mila Jankovic, Thiago Oliveira, Anne Bothmer, Ty C Voss, et al. (2012). "DNA damage defines sites of recurrent chromosomal translocations in B lymphocytes". *Nature* 484.7392, pp. 69–74. DOI: 10.1038/nature10909.

Hanahan, Douglas and Robert A Weinberg (2011). "Hallmarks of cancer: the next generation". *Cell* 144.5, pp. 646–674. DOI: 10.1016/j.cell.2011.02.013.

Hänsel-Hertsch, Robert, Dario Beraldi, Stefanie V Lensing, Giovanni Marsico, Katherine Zyner, Aled Parry, Marco Di Antonio, Jeremy Pike, Hiroshi Kimura, et al. (2016). "G-quadruplex structures mark human regulatory chromatin". *Nat. Genet.* DOI: 10.1038/ng.3662.

Hansen, R Scott, Sean Thomas, Richard Sandstrom, Theresa K Canfield, Robert E Thurman, Molly Weaver, Michael O Dorschner, Stanley M Gartler, and John A Stamatoyannopoulos (2010). "Sequencing newly replicated DNA reveals widespread plasticity in human replication timing". *Proc. Natl. Acad. Sci. U. S. A.* 107.1, pp. 139–144. DOI: 10.1073/pnas.0912402107.

Haradhvala, Nicholas J, Paz Polak, Petar Stojanov, Kyle R Covington, Eve Shinbrot, Julian M Hess, Esther Rheinbay, Jaegil Kim, Yosef E Maruvka, et al. (2016). "Mutational Strand Asymmetries in Cancer Genomes Reveal Mechanisms of DNA Damage and Repair". *Cell* 164.3, pp. 538–549. DOI: 10.1016/j.cell.2015.12.050.

Harewood, Louise, Kamal Kishore, Matthew D Eldridge, Steven Wingett, Danita Pearson, Stefan Schoenfelder, V Peter Collins, and Peter Fraser (2017). "Hi-C as a tool for precise detection and characterisation of chromosomal rearrangements and copy number variation in human tumours". *Genome Biol.* 18.1, p. 125. DOI: 10.1186/s13059-017-1253-8.

Harrow, Jennifer, Adam Frankish, Jose M Gonzalez, Electra Tapanari, Mark Diekhans, Felix Kokocinski, Bronwen L Aken, Daniel Barrell, Amonida Zadissa, et al. (2012). "GENCODE: the reference human genome annotation for The ENCODE Project". *Genome Res.* 22.9, pp. 1760–1774. DOI: 10.1101/gr.135350.111.

Hartlerode, Andrea J, Nicholas A Willis, Anbazhagan Rajendran, John P Manis, and Ralph Scully (2016). "Complex Breakpoints and Template Switching Associated with Non-canonical Termination of Homologous Recombination in Mammalian Cells". *PLoS Genet.* 12.11, e1006410. DOI: 10.1371/journal.pgen.1006410.

Heesch, Sebastiaan van, Marieke Simonis, Markus J van Roosmalen, Vamsee Pillalamarri, Harrison Brand, Ewart W Kuijk, Kim L de Luca, Nico Lansu, A Koen Braat, et al. (2014). "Genomic and functional overlap between somatic and germline chromosomal rearrangements". *Cell Rep.* 9.6, pp. 2001–2010. DOI: 10.1016/j.celrep.2014.11.022.

Helleday, Thomas, Saeed Eshtad, and Serena Nik-Zainal (2014). "Mechanisms underlying mutational signatures in human cancers". *Nat. Rev. Genet.* 15.9, pp. 585–598. DOI: 10.1038/nrg3729.

Helman, Elena, Michael S Lawrence, Chip Stewart, Carrie Sougnez, Gad Getz, and Matthew Meyerson (2014). "Somatic retrotransposition in human cancer revealed by whole-genome and exome sequencing". *Genome Res.* 24.7, pp. 1053–1063. DOI: 10.1101/gr.163659.113.

Hillman, R Tyler, Gary B Chisholm, Karen H Lu, and P Andrew Futreal (2018). "Genomic Rearrangement Signatures and Clinical Outcomes in High-Grade Serous Ovarian Cancer". *J. Natl. Cancer Inst.* 110.3. DOI: 10.1093/jnci/djx176.

Hirano, Tatsuya (2015). "Chromosome Dynamics during Mitosis". *Cold Spring Harb. Perspect. Biol.* 7.6. DOI: 10.1101/cshperspect.a015792.

Hoadley, Katherine A, Christina Yau, Denise M Wolf, Andrew D Cherniack, David Tamborero, Sam Ng, Max D M Leiserson, Beifang Niu, Michael D McLellan, et al. (2014). "Multiplatform analysis of 12 cancer types reveals molecular classification within and across tissues of origin". *Cell* 158.4, pp. 929–944. DOI: 10.1016/j.cell.2014.06.049.

Hughes, Michael, Dae Il Kim, and Erik Sudderth (2015). "Reliable and Scalable Variational Inference for the Hierarchical Dirichlet Process". *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics*, pp. 370–378.

Hyman, David M, Igor Puzanov, Vivek Subbiah, Jason E Faris, Ian Chau, Jean-Yves Blay, Jürgen Wolf, Noopur S Raje, Eli L Diamond, et al. (2015). "Vemurafenib in Multiple Nonmelanoma Cancers with BRAF V600 Mutations". *N. Engl. J. Med.* 373.8, pp. 726–736. DOI: 10.1056/NEJMoa1502309.

Hyman, David M, Barry S Taylor, and José Baselga (2017). "Implementing Genome-Driven Oncology". *Cell* 168.4, pp. 584–599. DOI: 10.1016/j.cell.2016.12.015.

Iakovishina, Daria, Isabelle Janoueix-Lerosey, Emmanuel Barillot, Mireille Regnier, and Valentina Boeva (2016). "SV-Bay: structural variant detection in cancer genomes using a Bayesian approach with correction for GC-content and read mappability". *Bioinformatics* 32.7, pp. 984–992. DOI: 10.1093/bioinformatics/btv751.

International Cancer Genome Consortium, Thomas J Hudson, Warwick Anderson, Axel Artez, Anna D Barker, Cindy Bell, Rosa R Bernabé, M K Bhan, Fabien Calvo, et al. (2010). "International network of cancer genome projects". *Nature* 464.7291, pp. 993–998. DOI: 10.1038/nature08987.

Jackson, Stephen P and Jiri Bartek (2009). "The DNA-damage response in human biology and disease". *Nature* 461.7267, pp. 1071–1078. DOI: 10.1038/nature08467.

James, Gareth, Daniela Witten, Trevor Hastie, and Robert Tibshirani (2013). *An Introduction to Statistical Learning*. Springer Texts in Statistics.

Jaratlerdsiri, Weerachai, Eva K F Chan, Desiree C Petersen, Claire Yang, Peter I Croucher, M S Riana Bornman, Palak Sheth, and Vanessa M Hayes (2017). "Next generation mapping reveals novel large genomic rearrangements in prostate cancer". *Oncotarget* 8.14, pp. 23588–23602. DOI: 10.18632/oncotarget.15802.

Jones, David T W, Sylvia Kocialkowski, Lu Liu, Danita M Pearson, L Magnus Bäcklund, Koichi Ichimura, and V Peter Collins (2008). "Tandem duplication producing a novel oncogenic BRAF fusion gene defines the majority of pilocytic astrocytomas". *Cancer Res.* 68.21, pp. 8673–8677. DOI: 10.1158/0008-5472.CAN-08-2097.

Kandoth, Cyriac, Michael D McLellan, Fabio Vandin, Kai Ye, Beifang Niu, Charles Lu, Mingchao Xie, Qunyuan Zhang, Joshua F McMichael, et al. (2013). "Mutational landscape and significance across 12 major cancer types". *Nature* 502.7471, pp. 333–339. DOI: 10.1038/nature12634.

Karolchik, Donna, Galt P Barber, Jonathan Casper, Hiram Clawson, Melissa S Cline, Mark Diekhans, Timothy R Dreszer, Pauline A Fujita, Luvina Guruvadoo, et al. (2014). "The UCSC Genome Browser database: 2014 update". *Nucleic Acids Res.* 42.Database issue, pp. D764–70. DOI: 10.1093/nar/gkt1168.

Karras, Jenna R, Morgan S Schrock, Bahadir Batar, and Kay Huebner (2017). "Fragile Genes That Are Frequently Altered in Cancer: Players Not Passengers". *Cytogenet. Genome Res.* DOI: 10.1159/000455753.

Kasparek, Torben R and Timothy C Humphrey (2011). "DNA double-strand break repair pathways, chromosomal rearrangements and cancer". *Semin. Cell Dev. Biol.* 22.8, pp. 886–897. DOI: 10.1016/j.semcdb.2011.10.007.

Khurana, Ekta, Yao Fu, Dimple Chakravarty, Francesca Demichelis, Mark A Rubin, and Mark Gerstein (2016). "Role of non-coding sequence variants in cancer". *Nat. Rev. Genet.* 17.2, pp. 93–108. DOI: 10.1038/nrg.2015.17.

Kim, Dae I and Erik B Sudderth (2011). "The Doubly Correlated Nonparametric Topic Model". *Advances in Neural Information Processing Systems 24*, pp. 1980–1988.

Kim, Jaegil, Kent W Mouw, Paz Polak, Lior Z Braunstein, Atanas Kamburov, Grace Tiao, David J Kwiatkowski, Jonathan E Rosenberg, Eliezer M Van Allen, et al. (2016). "Somatic ERCC2 mutations are associated with a distinct genomic signature in urothelial tumors". *Nat. Genet.* 48.6, pp. 600–606. DOI: 10.1038/ng.3557.

Kim, Tae-Min, Ruibin Xi, Lovelace J Luquette, Richard W Park, Mark D Johnson, and Peter J Park (2013). "Functional genomic analysis of chromosomal aberrations in a compendium of 8000 cancer genomes". *Genome Res.* 23.2, pp. 217–227. DOI: 10.1101/gr.140301.112.

Kinsella, Marcus and Vineet Bafna (2012). "Combinatorics of the Breakage-Fusion-Bridge Mechanism". *J. Comput. Biol.* 19.6, pp. 662–678. DOI: 10.1089/cmb.2012.0020.

Korbel, Jan O and Peter J Campbell (2013). "Criteria for inference of chromothripsis in cancer genomes". *Cell* 152.6, pp. 1226–1236. DOI: 10.1016/j.cell.2013.02.023.

Korbel, Jan O, Alexander Eckehart Urban, Jason P Affourtit, Brian Godwin, Fabian Grubert, Jan Fredrik Simons, Philip M Kim, Dean Palejev, Nicholas J Carriero, et al. (2007). "Paired-end mapping reveals extensive structural variation in the human genome". *Science* 318.5849, pp. 420–426. DOI: 10.1126/science.1149504.

Koren, Amnon, Robert E Handsaker, Nolan Kamitaki, Rosa Karlić, Sulagna Ghosh, Paz Polak, Kevin Eggan, and Steven A McCarroll (2014). "Genetic variation in human DNA replication timing". *Cell* 159.5, pp. 1015–1026. DOI: 10.1016/j.cell.2014.10.025.

L'Abbate, Alberto, Gemma Macchia, Pietro D'Addabbo, Angelo Lonoce, Doron Tolomeo, Domenico Trombetta, Klaas Kok, Christoph Bartenhagen, Christopher W Whelan, et al. (2014). "Genomic organization and evolution of double minutes/homogeneously staining regions with MYC amplification in human cancer". *Nucleic Acids Res.* 42.14, pp. 9131–9145. DOI: 10.1093/nar/gku590.

Lawrence, Michael S, Petar Stojanov, Paz Polak, Gregory V Kryukov, Kristian Cibulskis, Andrey Sivachenko, Scott L Carter, Chip Stewart, Craig H Mermel, et al. (2013). "Mutational heterogeneity in cancer and the search for new cancer-associated genes". *Nature* 499.7457, pp. 214–218. DOI: 10.1038/nature12213.

Lawrence, Michael S, Petar Stojanov, Craig H Mermel, James T Robinson, Levi A Garraway, Todd R Golub, Matthew Meyerson, Stacey B Gabriel, Eric S Lander, and Gad Getz (2014). "Discovery and saturation analysis of cancer genes across 21 tumour types". *Nature* 505.7484, pp. 495–501. DOI: 10.1038/nature12912.

Le Tallec, Benoît, Gaël Armel Millot, Marion Esther Blin, Olivier Brison, Bernard Dutrillaux, and Michelle Debatisse (2013). "Common fragile site profiling in epithelial and erythroid cells reveals that most recurrent cancer deletions lie in fragile sites hosting large genes". *Cell Rep.* 4.3, pp. 420–428. DOI: 10.1016/j.celrep.2013.07.003.

Lee, Eunjung, Rebecca Iskow, Lixing Yang, Omer Gokcumen, Psalm Haseley, Lovelace J Luquette 3rd, Jens G Lohr, Christopher C Harris, Li Ding, et al. (2012). "Landscape of somatic retrotransposition in human cancers". *Science* 337.6097, pp. 967–971. DOI: 10.1126/science.1222077.

Lee, Jason D, Dennis L Sun, Yuekai Sun, and Jonathan E Taylor (2016). "Exact post-selection inference, with application to the lasso". *Ann. Stat.* 44.3, pp. 907–927. DOI: 10.1214/15-AOS1371.

Lee, Jennifer A, Claudia M B Carvalho, and James R Lupski (2007). "A DNA replication mechanism for generating nonrecurrent rearrangements associated with genomic disorders". *Cell* 131.7, pp. 1235–1247. DOI: 10.1016/j.cell.2007.11.037.

Letessier, Anne, Gaël A Millot, Stéphane Koundrioukoff, Anne-Marie Lachagès, Nicolas Vogt, R Scott Hansen, Bernard Malfoy, Olivier Brison, and Michelle Debatisse (2011). "Cell-type-specific replication initiation programs set fragility of the FRA3B fragile site". *Nature* 470.7332, pp. 120–123. DOI: 10.1038/nature09745.

Ley, Timothy J, Elaine R Mardis, Li Ding, Bob Fulton, Michael D McLellan, Ken Chen, David Dooling, Brian H Dunford-Shore, Sean McGrath, et al. (2008). "DNA sequencing of a cytogenetically normal acute myeloid leukaemia genome". *Nature* 456.7218, pp. 66–72. DOI: 10.1038/nature07485.

Li, Yilong, Claire Schwab, Sarra L Ryan, Elli Papaemmanuil, Hazel M Robinson, Patricia Jacobs, Anthony V Moorman, Sara Dyer, Julian Borrow, et al. (2014). "Constitutional and somatic rearrangement of chromosome 21 in acute lymphoblastic leukaemia". *Nature* 508.7494, pp. 98–102. DOI: 10.1038/nature13115.

Li, Yilong, Nicola Roberts, Joachim Weischenfeldt, Jeremiah Anthony Wala, Ofer Shapira, Steven Schumacher, Ekta Khurana, Jan O Korbel, Marcin Imielinski, et al. (2017). "Patterns of structural variation in human cancer". *bioRxiv*. DOI: 10.1101/181339.

Lin, Chunru, Liuqing Yang, Bogdan Tanasa, Kasey Hutt, Bong-Gun Ju, Kenny Ohgi, Jie Zhang, David W Rose, Xiang-Dong Fu, et al. (2009). "Nuclear receptor-induced chromosomal proximity and DNA breaks underlie specific translocations in cancer". *Cell* 139.6, pp. 1069–1083. DOI: 10.1016/j.cell.2009.11.030.

Liu, Biao, Carl D Morrison, Candace S Johnson, Donald L Trump, Maochun Qin, Jeffrey C Conroy, Jianmin Wang, and Song Liu (2013). "Computational methods for detecting copy number variations in cancer genome using next generation sequencing: principles and challenges". *Oncotarget* 4.11, pp. 1868–1881. DOI: 10.18632/oncotarget.1537.

Liu, Biao, Jeffrey M Conroy, Carl D Morrison, Adekunle O Odunsi, Maochun Qin, Lei Wei, Donald L Trump, Candace S Johnson, Song Liu, and Jianmin Wang (2015). "Structural variation discovery in the cancer genome using next generation sequencing: computational solutions and perspectives". *Oncotarget* 6.8, pp. 5477–5489. DOI: 10.18632/oncotarget.3491.

Liu, Pengfei, Ayelet Erez, Sandesh C Sreenath Nagamani, Shweta U Dhar, Katarzyna E Kołodziejska, Avinash V Dharmadhikari, M Lance Cooper, Joanna Wiszniewska, Feng Zhang, et al. (2011). "Chromosome catastrophes involve replication mechanisms generating complex genomic rearrangements". *Cell* 146.6, pp. 889–903. DOI: 10.1016/j.cell.2011.07.042.

Liu, Yu, Chong Chen, Zhengmin Xu, Claudio Scuoppo, Cory D Rillahan, Jianjiong Gao, Barbara Spitzer, Benedikt Bosbach, Edward R Kastenhuber, et al. (2016). "Deletions linked to TP53 loss drive cancer through p53-independent mechanisms". *Nature* 531.7595, pp. 471–475. DOI: 10.1038/nature17157.

Loeb, Lawrence A and Raymond J Monnat (2008). "DNA polymerases and human disease". *Nat. Rev. Genet.* 9.8, pp. 594–604. DOI: 10.1038/nrg2345.

Lupiáñez, Darío G, Katerina Kraft, Verena Heinrich, Peter Krawitz, Francesco Brancati, Eva Klopocki, Denise Horn, Hülya Kayserili, John M Opitz, et al. (2015). "Disruptions of topological chromatin domains cause pathogenic rewiring of gene-enhancer interactions". *Cell* 161.5, pp. 1012–1025. DOI: 10.1016/j.cell.2015.04.004.

Macheret, Morgane and Thanos D Halazonetis (2015). "DNA replication stress as a hallmark of cancer". *Annu. Rev. Pathol.* 10, pp. 425–448. DOI: 10.1146/annurev-pathol-012414-040424.

Maciejowski, John, Yilong Li, Nazario Bosco, Peter J Campbell, and Titia de Lange (2015). "Chromothripsis and Kataegis Induced by Telomere Crisis". *Cell* 163.7, pp. 1641–1654. DOI: 10.1016/j.cell.2015.11.054.

Macintyre, Geoff, Peter Van Loo, Niall M Corcoran, David C Wedge, Florian Markowetz, and Christopher M Hovens (2016a). "How subclonal modelling is changing the metastatic paradigm". *Clin. Cancer Res.* DOI: 10.1158/1078-0432.CCR-16-0234.

Macintyre, Geoff, Bauke Ylstra, and James D Brenton (2016b). "Sequencing Structural Variants in Cancer for Precision Therapeutics". *Trends Genet.* 32.9, pp. 530–542. DOI: 10.1016/j.tig.2016.07.002.

Maddalo, Danilo, Eusebio Manchado, Carla P Concepcion, Ciro Bonetti, Joana A Vidigal, Yoon-Chi Han, Paul Ogrodowski, Alessandra Crippa, Natasha Rekhtman, et al. (2014). "In vivo engineering of oncogenic chromosomal rearrangements with the CRISPR/Cas9 system". *Nature* 516.7531, pp. 423–427. DOI: 10.1038/nature13902.

Malhotra, Ankit, Michael Lindberg, Gregory G Faust, Mitchell L Leibowitz, Royden A Clark, Ryan M Layer, Aaron R Quinlan, and Ira M Hall (2013). "Breakpoint profiling of 64 cancer genomes reveals numerous complex rearrangements spawned by homology-independent mechanisms". *Genome Res.* 23.5, pp. 762–776. DOI: 10.1101/gr.143677.112.

Mardin, Balca R, Alexandros P Drainas, Sebastian M Waszak, Joachim Weischenfeldt, Mayumi Isokane, Adrian M Stütz, Benjamin Raeder, Theocharis Efthymiopoulos, Christopher Buccitelli, et al. (2015). "A cell-based model system links chromothripsis with hyperploidy". *Mol. Syst. Biol.* 11.9, p. 828. DOI: 10.15252/msb.20156505.

Martincorena, Iñigo and Peter J Campbell (2015). "Somatic mutation in cancer and normal cells". *Science* 349.6255, pp. 1483–1489. DOI: 10.1126/science.aab4082.

Martincorena, Iñigo, Keiran M Raine, Moritz Gerstung, Kevin J Dawson, Kerstin Haase, Peter Van Loo, Helen Davies, Michael R Stratton, and Peter J Campbell (2017). "Universal Patterns of Selection in Cancer and Somatic Tissues". *Cell.* DOI: 10.1016/j.cell.2017.09.042.

McClintock, B (1941). "The Stability of Broken Ends of Chromosomes in Zea Mays". *Genetics* 26.2, pp. 234–282.

McKenna, Aaron, Matthew Hanna, Eric Banks, Andrey Sivachenko, Kristian Cibulskis, Andrew Kernytsky, Kiran Garimella, David Altshuler, Stacey Gabriel, et al. (2010). "The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data". *Genome Res.* 20.9, pp. 1297–1303. DOI: 10.1101/gr.107524.110.

McStay, Brian (2016). "Nucleolar organizer regions: genomic 'dark matter' requiring illumination". *Genes Dev.* 30.14, pp. 1598–1610. DOI: 10.1101/gad.283838.116.

Meier, Bettina, Susanna L Cooke, Joerg Weiss, Aymeric P Bailly, Ludmil B Alexandrov, John Marshall, Keiran Raine, Mark Maddison, Elizabeth Anderson, et al. (2014). "C. elegans whole-genome sequencing reveals mutational signatures related to carcinogens and DNA repair deficiency". *Genome Res.* 24.10, pp. 1624–1636. DOI: 10.1101/gr.175547.114.

Menghi, Francesca, Koichiro Inaki, Xingyi Woo, Pooja A Kumar, Krzysztof R Grzeda, Ankit Malhotra, Vinod Yadav, Hyunsoo Kim, Eladio J Marquez, et al. (2016). "The tandem duplicator phenotype as a distinct genomic configuration in cancer". *Proc. Natl. Acad. Sci. U. S. A.* 113.17, E2373–82. DOI: 10.1073/pnas.1520010113.

Merker, Jason D, Aaron M Wenger, Tam Sneddon, Megan Grove, Zachary Zappala, Laure Fresard, Daryl Waggott, Sowmi Utiramerur, Yanli Hou, et al. (2018). "Long-read genome sequencing identifies causal structural variation in a Mendelian disease". *Genet. Med.* 20.1, pp. 159–163. DOI: 10.1038/gim.2017.86.

Mermel, Craig H, Steven E Schumacher, Barbara Hill, Matthew L Meyerson, Rameen Beroukhim, and Gad Getz (2011). "GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers". *Genome Biol.* 12.4, R41. DOI: 10.1186/gb-2011-12-4-r41.

Mertens, Fredrik, Bertil Johansson, Thoas Fioretos, and Felix Mitelman (2015). "The emerging complexity of gene fusions in cancer". *Nat. Rev. Cancer* 15.6, pp. 371–381. DOI: 10.1038/nrc3947.

Miron, Karin, Tamar Golan-Lev, Raz Dvir, Eyal Ben-David, and Batsheva Kerem (2015). "Oncogenes create a unique landscape of fragile sites". *Nat. Commun.* 6, p. 7094. DOI: 10.1038/ncomms8094.

Moncunill, Valentí, Santi Gonzalez, Sílvia Beà, Lise O Andrieux, Itziar Salaverria, Cristina Royo, Laura Martinez, Montserrat Puiggròs, Maia Segura-Wang, et al. (2014). "Comprehensive characterization of complex structural variations in cancer by directly comparing genome sequence reads". *Nat. Biotechnol.* 32.11, pp. 1106–1112. DOI: 10.1038/nbt.3027.

Morganella, Sandro, Ludmil B Alexandrov, Dominik Glodzik, Xueqing Zou, Helen Davies, Johan Staaf, Anieta M Sieuwerts, Arie B Brinkman, Sancha Martin, et al. (2016). "The topography of mutational processes in breast cancer genomes". *Nat. Commun.* 7.May, p. 11383. DOI: 10.1038/ncomms11383.

Mwenifumbo, Jill C and Marco A Marra (2013). "Cancer genome-sequencing study design". *Nat. Rev. Genet.* 14.5, pp. 321–332. DOI: 10.1038/nrg3445.

Nagaraju, Ganesh, Andrea Hartlerode, Amy Kwok, Gurushankar Chandramouly, and Ralph Scully (2009). "XRCC2 and XRCC3 regulate the balance between short- and long-tract gene conversions

between sister chromatids". *Mol. Cell. Biol.* 29.15, pp. 4283–4294. DOI: 10.1128/MCB.01406-08.

Nattestad, Maria, Sara Goodwin, Karen Ng, Timour Baslan, Fritz J Sedlazeck, Philipp Rescheneder, Tyler Garvin, Han Fang, James Gurtowski, et al. (2017). "Complex rearrangements and oncogene amplifications revealed by long-read DNA and RNA sequencing of a breast cancer cell line". *bioRxiv*. DOI: 10.1101/174938.

Ng, Stanley W K, Amanda Mitchell, James A Kennedy, Weihsu C Chen, Jessica McLeod, Narmin Ibrahimova, Andrea Arruda, Andreea Popescu, Vikas Gupta, et al. (2016). "A 17-gene stemness score for rapid determination of risk in acute leukaemia". *Nature* 540.7633, pp. 433–437. DOI: 10.1038/nature20598.

Nik-Zainal, Serena, Ludmil B Alexandrov, David C Wedge, Peter Van Loo, Christopher D Greenman, Keiran Raine, David Jones, Jonathan Hinton, John Marshall, et al. (2012). "Mutational processes molding the genomes of 21 breast cancers". *Cell* 149.5, pp. 979–993. DOI: 10.1016/j.cell.2012.04.024.

Nik-Zainal, Serena, Helen Davies, Johan Staaf, Manasa Ramakrishna, Dominik Glodzik, Xueqing Zou, Inigo Martincorena, Ludmil B Alexandrov, Sancha Martin, et al. (2016). "Landscape of somatic mutations in 560 breast cancer whole-genome sequences". *Nature* 534.7605, pp. 47–54. DOI: 10.1038/nature17676.

Northcott, Paul A, Catherine Lee, Thomas Zichner, Adrian M Stütz, Serap Erkek, Daisuke Kawauchi, David J H Shih, Volker Hovestadt, Marc Zapatka, et al. (2014). "Enhancer hijacking activates GFI1 family oncogenes in medulloblastoma". *Nature* 511.7510, pp. 428–434. DOI: 10.1038/nature13379.

Northcott, Paul A, Ivo Buchhalter, A Sorana Morrissy, Volker Hovestadt, Joachim Weischenfeldt, Tobias Ehrenberger, Susanne Gröbner, Maia Segura-Wang, Thomas Zichner, et al. (2017). "The whole-genome landscape of medulloblastoma subtypes". *Nature* 547.7663, pp. 311–317. DOI: 10.1038/nature22973.

Notta, Faiyaz, Michelle Chan-Seng-Yue, Mathieu Lemire, Yilong Li, Gavin W Wilson, Ashton A Connor, Robert E Denroche, Sheng-Ben Liang, Andrew M K Brown, et al. (2016). "A renewed model of pancreatic cancer evolution based on genomic rearrangement patterns". *Nature*. DOI: 10.1038/nature19823.

Nowell, Peter C (2007). "Discovery of the Philadelphia chromosome: a personal perspective". *J. Clin. Invest.* 117.8, pp. 2033–2035. DOI: 10.1172/JCI31771.

Orr, Bernardo, Kristina M Godek, and Duane Compton (2015). "Aneuploidy". *Curr. Biol.* 25.13, R538–42. DOI: 10.1016/j.cub.2015.05.010.

Ozeri-Galai, Efrat, Assaf C Bester, and Batsheva Kerem (2012). "The complex basis underlying common fragile site instability in cancer". *Trends Genet.* 28.6, pp. 295–302. DOI: 10.1016/j.tig.2012.02.006.

Papaemmanuil, Elli, Moritz Gerstung, Lars Bullinger, Verena I Gaidzik, Peter Paschka, Nicola D Roberts, Nicola E Potter, Michael Heuser, Felicitas Thol, et al. (2016). "Genomic Classification and Prognosis in Acute

Myeloid Leukemia". *N. Engl. J. Med.* 374.23, pp. 2209–2221. DOI: 10.1056/NEJMoa1516192.

Patch, Ann-Marie, Elizabeth L Christie, Dariush Etemadmoghadam, Dale W Garsed, Joshy George, Sian Fereday, Katia Nones, Prue Cowin, Kathryn Alsop, et al. (2015). "Whole-genome characterization of chemoresistant ovarian cancer". *Nature* 521.7553, pp. 489–494. DOI: 10.1038/nature14410.

Pedersen, Brent S and Subhajyoti De (2013). "Loss of heterozygosity preferentially occurs in early replicating regions in cancer genomes". *Nucleic Acids Res.* 41.16, pp. 7615–7624. DOI: 10.1093/nar/gkt552.

Peifer, Martin, Falk Hertwig, Frederik Roels, Daniel Dreidax, Moritz Gartlgruber, Roopika Menon, Andrea Krämer, Justin L Roncaioli, Frederik Sand, et al. (2015). "Telomerase activation by genomic rearrangements in high-risk neuroblastoma". *Nature* 526.7575, pp. 700–704. DOI: 10.1038/nature14980.

Peplow, Mark (2016). "The 100,000 Genomes Project". *BMJ* 353, p. i1757. DOI: 10.1136/bmj.i1757.

Piazza, Aurèle, Michael Adrian, Frédéric Samazan, Brahim Heddi, Florian Hamon, Alexandre Serero, Judith Lopes, Marie-Paule Teulade-Fichou, Anh Tuân Phan, and Alain Nicolas (2015). "Short loop length and high thermal stability determine genomic instability induced by G-quadruplex-forming minisatellites". *EMBO J.* 34.12, pp. 1718–1734. DOI: 10.15252/embj.201490702.

Pinkel, D, R Segraves, D Sudar, S Clark, I Poole, D Kowbel, C Collins, W L Kuo, C Chen, et al. (1998). "High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays". *Nat. Genet.* 20.2, pp. 207–211. DOI: 10.1038/2524.

Pleasance, Erin D, R Keira Cheetham, Philip J Stephens, David J McBride, Sean J Humphray, Chris D Greenman, Ignacio Varela, Meng-Lay Lin, Gonzalo R Ordóñez, et al. (2010). "A comprehensive catalogue of somatic mutations from a human cancer genome". *Nature* 463.7278, pp. 191–196. DOI: 10.1038/nature08658.

Polak, Paz, Jaegil Kim, Lior Z Braunstein, Rosa Karlic, Nicholas J Haradhavala, Grace Tiao, Daniel Rosebrock, Dimitri Livitz, Kirsten Kübler, et al. (2017). "A mutational signature reveals alterations underlying deficient homologous recombination repair in breast cancer". *Nat. Genet.* DOI: 10.1038/ng.3934.

Pombo, Ana and Niall Dillon (2015). "Three-dimensional genome architecture: players and mechanisms". *Nat. Rev. Mol. Cell Biol.* 16.4, pp. 245–257. DOI: 10.1038/nrm3965.

Pon, Julia R and Marco A Marra (2015). "Driver and passenger mutations in cancer". *Annu. Rev. Pathol.* 10, pp. 25–50. DOI: 10.1146/annurev-pathol-012414-040312.

Pons, Pascal and Matthieu Latapy (2006). "Computing Communities in Large Networks Using Random Walks". *Journal of Graph Algorithms and Applications* 10.2, pp. 191–218. DOI: 10.7155/jgaa.00124.

Poon, Song Ling, See-Tong Pang, John R McPherson, Willie Yu, Kie Kyon Huang, Peiyong Guan, Wen-Hui Weng, Ee Yan Siew, Yujing Liu, et al. (2013). "Genome-wide mutational signatures of aristolochic acid and its application as a screening tool". *Sci. Transl. Med.* 5.197, 197ra101. DOI: 10.1126/scitranslmed.3006086.

Poon, Song Ling, Mi Ni Huang, Yang Choo, John R McPherson, Willie Yu, Hong Lee Heng, Anna Gan, Swe Swe Myint, Ee Yan Siew, et al. (2015). "Mutation signatures implicate aristolochic acid in bladder cancer development". *Genome Med.* 7.1, p. 38. DOI: 10.1186/s13073-015-0161-3.

R Core Team (2017). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rausch, Tobias, Thomas Zichner, Andreas Schlattl, Adrian M Stütz, Vladimir Benes, and Jan O Korbel (2012). "DELLY: structural variant discovery by integrated paired-end and split-read analysis". *Bioinformatics* 28.18, pp. i333–i339. DOI: 10.1093/bioinformatics/bts378.

Rayner, Emily, Inge C van Gool, Claire Palles, Stephen E Kearsey, Tjalling Bosse, Ian Tomlinson, and David N Church (2016). "A panoply of errors: polymerase proofreading domain mutations in cancer". *Nat. Rev. Cancer* 16.2, pp. 71–81. DOI: 10.1038/nrc.2015.12.

Renkawitz, Jörg, Claudio A Lademann, and Stefan Jentsch (2014). "Mechanisms and principles of homology search during recombination". *Nat. Rev. Mol. Cell Biol.* 15.6, pp. 369–383. DOI: 10.1038/nrm3805.

Reuter, Jason A, Damek V Spacek, and Michael P Snyder (2015). "High-throughput sequencing technologies". *Mol. Cell* 58.4, pp. 586–597. DOI: 10.1016/j.molcel.2015.05.004.

Rhind, Nicholas and David M Gilbert (2013). "DNA replication timing". *Cold Spring Harb. Perspect. Biol.* 5.8, a010132. DOI: 10.1101/cshperspect.a010132.

Roadmap Epigenomics Consortium, Anshul Kundaje, Wouter Meuleman, Jason Ernst, Misha Bilenky, Angela Yen, Alireza Heravi-Moussavi, Pouya Kheradpour, Zhizhuo Zhang, et al. (2015). "Integrative analysis of 111 reference human epigenomes". *Nature* 518.7539, pp. 317–330. DOI: 10.1038/nature14248.

Roberts, Nicola D (2015). *R pkg for Hierarchical Dirichlet Process*. https://github.com/nicolaroberts/hdp. Accessed: 2017-10-4.

Rodgers, Kasey and Mitch McVey (2016). "Error-Prone Repair of DNA Double-Strand Breaks". *J. Cell. Physiol.* 231.1, pp. 15–24.

Rodriguez-Martin, Bernardo, Eva G Alvarez, Adrian Baez-Ortega, Jonas Demeulemeester, Young Seok Ju, Jorge Zamora, Harald Detering, Yilong Li, Gianmarco Contino, et al. (2017). "Pan-cancer analysis of whole genomes reveals driver rearrangements promoted by LINE-1 retrotransposition in human tumours". *bioRxiv*. DOI: 10.1101/179705.

Roix, Jeffrey J, Philip G McQueen, Peter J Munson, Luis A Parada, and Tom Misteli (2003). "Spatial proximity of translocation-prone gene loci in human lymphomas". *Nat. Genet.* 34, p. 287. DOI: 10.1038/ng1177.

Rosales, Rafael A, Rodrigo D Drummond, Renan Valieris, Emmanuel Dias-Neto, and Israel T da Silva (2016). "signeR: An empirical

Bayesian approach to mutational signature discovery". *Bioinformatics*. DOI: 10.1093/bioinformatics/btw572.

Rosenthal, Rachel, Nicholas McGranahan, Javier Herrero, Barry S Taylor, and Charles Swanton (2016). "DeconstructSigs: delineating mutational processes in single tumors distinguishes DNA repair deficiencies and patterns of carcinoma evolution". *Genome Biol.* 17.1, p. 31. DOI: 10.1186/s13059-016-0893-4.

Roukos, Vassilis and Tom Misteli (2014). "The biogenesis of chromosome translocations". *Nat. Cell Biol.* 16.4, pp. 293–300. DOI: 10.1038/ncb2941.

Rowley, J D (1973). "A new consistent chromosomal abnormality in chronic myelogenous leukaemia identified by quinacrine fluorescence and Giemsa staining". *Nature* 243.5405, pp. 290–293. DOI: 10.1038/243290a0.

Ruiz-Velasco, Mariana and Judith B Zaugg (2017). "Structure meets function: How chromatin organisation conveys functionality". *Current Opinion in Systems Biology* 1.Supplement C, pp. 129–136. DOI: 10.1016/j.coisb.2017.01.003.

Sabarinathan, Radhakrishnan, Oriol Pich, Inigo Martincorena, Carlota Rubio-Perez, Malene Juul, Jeremiah Wala, Steven Schumacher, Ofer Shapira, Nikos Sidiropoulos, et al. (2017). "The whole-genome panorama of cancer drivers". *bioRxiv*. DOI: 10.1101/190330.

Sakofsky, Cynthia J and Anna Malkova (2017). "Break induced replication in eukaryotes: mechanisms, functions, and consequences". *Crit. Rev. Biochem. Mol. Biol.* 52.4, pp. 395–413. DOI: 10.1080/10409238.2017.1314444.

Sakofsky, Cynthia J, Steven A Roberts, Ewa Malc, Piotr A Mieczkowski, Michael A Resnick, Dmitry A Gordenin, and Anna Malkova (2014). "Break-induced replication is a source of mutation clusters underlying kataegis". *Cell Rep.* 7.5, pp. 1640–1648. DOI: 10.1016/j.celrep.2014.04.053.

Sakofsky, Cynthia J, Sandeep Ayyar, Angela K Deem, Woo-Hyun Chung, Grzegorz Ira, and Anna Malkova (2015). "Translesion Polymerases Drive Microhomology-Mediated Break-Induced Replication Leading to Complex Chromosomal Rearrangements". *Mol. Cell* 60.6, pp. 860–872. DOI: 10.1016/j.molcel.2015.10.041.

Salesse, Stephanie and Catherine M Verfaillie (2002). "BCR/ABL: from molecular mechanisms of leukemia induction to treatment of chronic myelogenous leukemia". *Oncogene* 21.56, pp. 8547–8559. DOI: 10.1038/sj.onc.1206082.

Sanborn, J Zachary, Sofie R Salama, Mia Grifford, Cameron W Brennan, Tom Mikkelsen, Suresh Jhanwar, Sol Katzman, Lynda Chin, and David Haussler (2013). "Double minute chromosomes in glioblastoma multiforme are revealed by precise reconstruction of oncogenic amplicons". *Cancer Res.* 73.19, pp. 6036–6045. DOI: 10.1158/0008-5472.CAN-13-0186.

Sandberg, A A (1991). "Chromosome abnormalities in human cancer and leukemia". *Mutat. Res.* 247.2, pp. 231–240. DOI: 10.1016/0027-5107(91)90019-K.

Sarni, Dan and Batsheva Kerem (2016). "The complex nature of fragile site plasticity and its importance in cancer". *Curr. Opin. Cell Biol.* 40, pp. 131–136. DOI: 10.1016/j.ceb.2016.03.017.

Scarpa, Aldo, David K Chang, Katia Nones, Vincenzo Corbo, Ann-Marie Patch, Peter Bailey, Rita T Lawlor, Amber L Johns, David K Miller, et al. (2017). "Whole-genome landscape of pancreatic neuroendocrine tumours". *Nature*. DOI: 10.1038/nature21063.

Schwartz, Russell and Alejandro A Schäffer (2017). "The evolution of tumour phylogenetics: principles and practice". *Nat. Rev. Genet.* DOI: 10.1038/nrg.2016.170.

Segovia, Romulo, Annie S Tam, and Peter C Stirling (2015). "Dissecting genetic and environmental mutation signatures with model organisms". *Trends Genet.* 31.8, pp. 465–474. DOI: 10.1016/j.tig.2015.04.001.

Shiraishi, Yuichi, Georg Tremmel, Satoru Miyano, and Matthew Stephens (2015). "A Simple Model-Based Approach to Inferring and Visualizing Cancer Mutation Signatures". *PLoS Genet.* 11.12, e1005657. DOI: 10.1371/journal.pgen.1005657.

Sieverling, Lina, Chen Hong, Sandra D Koser, Philip Ginsbach, Kortine Kleinheinz, Barbara Hutter, Delia M Braun, Isidro Cortes-Ciriano, Ruibin Xi, et al. (2017). "Genomic footprints of activated telomere maintenance mechanisms in cancer". *bioRxiv*, p. 157560. DOI: 10.1101/157560.

Siravegna, Giulia, Silvia Marsoni, Salvatore Siena, and Alberto Bardelli (2017). "Integrating liquid biopsies into the management of cancer". *Nat. Rev. Clin. Oncol.* 14.9, pp. 531–548. DOI: 10.1038/nrclinonc.2017.14.

Smit, A F A, R Hubley, and P Green. *RepeatMasker Open-4.0*. http://www.repeatmasker.org. Accessed: 2015-1-21.

Soleimanpour, Scott A, Aditi Gupta, Marina Bakay, Alana M Ferrari, David N Groff, João Fadista, Lynn A Spruce, Jake A Kushner, Leif Groop, et al. (2014). "The diabetes susceptibility gene Clec16a regulates mitophagy". *Cell* 157.7, pp. 1577–1590. DOI: 10.1016/j.cell.2014.05.016.

Spriggs, A I, M M Boddington, and C M Clarke (1962). "Chromosomes of human cancer cells". *Br. Med. J.* 2.5317, pp. 1431–1435.

St John, Jason, Katelyn Powell, M Katie Conley-Lacomb, and Sreenivasa R Chinni (2012). "TMPRSS2-ERG Fusion Gene Expression in Prostate Tumor Cells and Its Clinical and Biological Significance in Prostate Cancer Progression". *J. Cancer Sci. Ther.* 4.4, pp. 94–101. DOI: 10.4172/1948-5956.1000119.

Stephens, Philip J, David J McBride, Meng-Lay Lin, Ignacio Varela, Erin D Pleasance, Jared T Simpson, Lucy A Stebbings, Catherine Leroy, Sarah Edkins, et al. (2009). "Complex landscapes of somatic rearrangement in human breast cancer genomes". *Nature* 462.7276, pp. 1005–1010. DOI: 10.1038/nature08645.

Stephens, Philip J, Chris D Greenman, Beiyuan Fu, Fengtang Yang, Graham R Bignell, Laura J Mudie, Erin D Pleasance, King Wai Lau, David Beare, et al. (2011). "Massive genomic rearrangement acquired in a

single catastrophic event during cancer development". *Cell* 144.1, pp. 27–40. DOI: 10.1016/j.cell.2010.11.055.

Stratton, Michael R, Peter J Campbell, and P Andrew Futreal (2009). "The cancer genome". *Nature* 458.7239, pp. 719–724. DOI: 10.1038/nature07943.

Sudmant, Peter H, Tobias Rausch, Eugene J Gardner, Robert E Handsaker, Alexej Abyzov, John Huddleston, Yan Zhang, Kai Ye, Goo Jun, et al. (2015). "An integrated map of structural variation in 2,504 human genomes". *Nature* 526.7571, pp. 75–81. DOI: 10.1038/nature15394.

Supek, Fran and Ben Lehner (2017). "Clustered Mutation Signatures Reveal that Error-Prone DNA Repair Targets Mutations to Active Genes". *Cell* 170.3, 534–547.e23. DOI: 10.1016/j.cell.2017.07.003.

Tan, Vincent Y F and Cédric Févotte (2013). "Automatic relevance determination in nonnegative matrix factorization with the β-divergence". *IEEE Trans. Pattern Anal. Mach. Intell.* 35.7, pp. 1592–1605. DOI: 10.1109/TPAMI.2012.240.

Taylor, Benjamin Jm, Serena Nik-Zainal, Yee Ling Wu, Lucy A Stebbings, Keiran Raine, Peter J Campbell, Cristina Rada, Michael R Stratton, and Michael S Neuberger (2013). "DNA deaminases induce break-associated mutation showers with implication of APOBEC3B and 3A in breast cancer kataegis". *Elife* 2, e00534. DOI: 10.7554/eLife.00534.

Taylor, Jonathan and Robert Tibshirani (2017). "Post-selection inference for 1-penalized likelihood models". *Can. J. Stat.* DOI: 10.1002/cjs.11313.

Teh, Y and Michael I Jordan (2009). "Hierarchical Bayesian nonparametric models with applications". *Bayesian Nonparametrics*, pp. 1–48.

Teh, Yee Whye, Michael I Jordan, Matthew J Beal, and David M Blei (2006). "Hierarchical Dirichlet Processes". *J. Am. Stat. Assoc.* 101.476, pp. 1566–1581. DOI: 10.1198/016214506000000302.

Tomasetti, Cristian, Luigi Marchionni, Martin A Nowak, Giovanni Parmigiani, and Bert Vogelstein (2015). "Only three driver gene mutations are required for the development of lung and colorectal cancers". *Proc. Natl. Acad. Sci. U. S. A.* 112.1, pp. 118–123. DOI: 10.1073/pnas.1421839112.

Totoki, Yasushi, Kenji Tatsuno, Kyle R Covington, Hiroki Ueda, Chad J Creighton, Mamoru Kato, Shingo Tsuji, Lawrence A Donehower, Betty L Slagle, et al. (2014). "Trans-ancestry mutational landscape of hepatocellular carcinoma genomes". *Nat. Genet.* 46.12, pp. 1267–1273. DOI: 10.1038/ng.3126.

Tubbs, Anthony and André Nussenzweig (2017). "Endogenous DNA Damage as a Source of Genomic Instability in Cancer". *Cell* 168.4, pp. 644–656. DOI: 10.1016/j.cell.2017.01.002.

Tubio, Jose M C, Yilong Li, Young Seok Ju, Inigo Martincorena, Susanna L Cooke, Marta Tojo, Gunes Gundem, Christodoulos P Pipinikas, Jorge Zamora, et al. (2014). "Mobile DNA in cancer. Extensive transduction of nonrepetitive DNA mediated by L1 retrotransposition in cancer genomes". *Science* 345.6196, p. 1251343. DOI: 10.1126/science.1251343.

Turner, Kristen M, Viraj Deshpande, Doruk Beyter, Tomoyuki Koga, Jessica Rusert, Catherine Lee, Bin Li, Karen Arden, Bing Ren, et al. (2017).

"Extrachromosomal oncogene amplification drives tumour evolution and genetic heterogeneity". *Nature*. DOI: 10.1038/nature21356.

Vaandrager, J W, E Schuuring, K Philippo, and P M Kluin (2000). "V(D)J recombinase-mediated transposition of the BCL2 gene to the IGH locus in follicular lymphoma". *Blood* 96.5, pp. 1947–1952.

Valton, Anne-Laure and Job Dekker (2016). "TAD disruption as oncogenic driver". *Curr. Opin. Genet. Dev.* 36, pp. 34–40. DOI: 10.1016/j.gde.2016.03.008.

Vogelstein, Bert, Nickolas Papadopoulos, Victor E Velculescu, Shibin Zhou, Luis A Diaz Jr, and Kenneth W Kinzler (2013). "Cancer genome landscapes". *Science* 339.6127, pp. 1546–1558. DOI: 10.1126/science.1235122.

Vogt, Nicolas, Anne Gibaud, Frédéric Lemoine, Pierre de la Grange, Michelle Debatisse, and Bernard Malfoy (2014). "Amplicon rearrangements during the extrachromosomal and intrachromosomal amplification process in a glioma". *Nucleic Acids Res.* 42.21, pp. 13194–13205. DOI: 10.1093/nar/gku1101.

Wala, Jeremiah Anthony, Ofer Shapira, Yilong Li, David Craft, Steven E Schumacher, Marcin Imielinski, James E Haber, Nicola Roberts, Xiaotong Yao, et al. (2017a). "Selective and mechanistic sources of recurrent rearrangements across the cancer genome". *bioRxiv*. DOI: 10.1101/187609.

Wala, Jeremiah, Pratiti Bandopadhayay, Noah Greenwald, Ryan O'Rourke, Ted Sharpe, Chip Stewart, Steven E Schumacher, Yilong Li, Joachim Weischenfeldt, et al. (2017b). "Genome-wide detection of structural variants and indels by local assembly". *bioRxiv*. DOI: 10.1101/105080.

Wan, Jonathan C M, Charles Massie, Javier Garcia-Corbacho, Florent Mouliere, James D Brenton, Carlos Caldas, Simon Pacey, Richard Baird, and Nitzan Rosenfeld (2017). "Liquid biopsies come of age: towards implementation of circulating tumour DNA". *Nat. Rev. Cancer* 17.4, pp. 223–238. DOI: 10.1038/nrc.2017.7.

Wang, Chong and David M Blei (2009). "Decoupling Sparsity and Smoothness in the Discrete Hierarchical Dirichlet Process". *Advances in Neural Information Processing Systems 22*, pp. 1982–1989.

Wang, Chong, John Paisley, and David M Blei (2011). "Online Variational Inference for the Hierarchical Dirichlet Process". *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics* 15, pp. 752–760.

Weischenfeldt, Joachim, Taronish Dubash, Alexandros P Drainas, Balca R Mardin, Yuanyuan Chen, Adrian M Stütz, Sebastian M Waszak, Graziella Bosco, Ann Rita Halvorsen, et al. (2017). "Pan-cancer analysis of somatic copy-number alterations implicates IRS4 and IGF2 in enhancer hijacking". *Nat. Genet.* 49.1, pp. 65–74. DOI: 10.1038/ng.3722.

Whalley, Justin P, Ivo Buchhalter, Esther Rheinbay, Keiran M Raine, Kortine Kleinheinz, Miranda D Stobbe, Johannes Werner, Sergi Beltran, Marta Gut, et al. (2017). "Framework For Quality Assessment Of Whole Genome, Cancer Sequences". *bioRxiv*. DOI: 10.1101/140921.

Wheeler, David A and Linghua Wang (2013). "From human genome to cancer genome: the first decade". *Genome Res.* 23.7, pp. 1054–1062. DOI: 10.1101/gr.157602.113.

Williamson, Sinead, Chong Wang, Katherine A Heller, and David M Blei (2010). "The IBP Compound Dirichlet Process and Its Application to Focused Topic Modeling". *Proceedings of the 27th International Conference on Machine Learning*, pp. 1151–1158.

Willis, Nicholas A, Emilie Rass, and Ralph Scully (2015). "Deciphering the Code of the Cancer Genome: Mechanisms of Chromosome Rearrangement". *Trends Cancer Res.* 1.4, pp. 217–230. DOI: 10.1016/j.trecan.2015.10.007.

Wilson, Thomas E, Martin F Arlt, So Hae Park, Sountharia Rajendran, Michelle Paulsen, Mats Ljungman, and Thomas W Glover (2015). "Large transcription units unify copy number variants and common fragile sites arising under replication stress". *Genome Res.* 25.2, pp. 189–200. DOI: 10.1101/gr.177121.114.

Xia, Li C, John M Bell, Christina Wood-Bouwens, Jiamin J Chen, Nancy R Zhang, and Hanlee P Ji (2017). "Identification of large rearrangements in cancer genomes with barcode linked reads". *Nucleic Acids Res.* DOI: 10.1093/nar/gkx1193.

Yamagata, Koichi, Ayako Yamanishi, Chikara Kokubu, Junji Takeda, and Jun Sese (2016). "COSMOS: accurate detection of somatic structural variations through asymmetric comparison between tumor and normal samples". *Nucleic Acids Res.* 44.8, e78. DOI: 10.1093/nar/gkw026.

Yang, Lixing, Lovelace J Luquette, Nils Gehlenborg, Ruibin Xi, Psalm S Haseley, Chih-Heng Hsieh, Chengsheng Zhang, Xiaojia Ren, Alexei Protopopov, et al. (2013). "Diverse mechanisms of somatic structural variations in human cancer genomes". *Cell* 153.4, pp. 919–929. DOI: 10.1016/j.cell.2013.04.010.

Yim, Eunice, Karen E O'Connell, Jordan St Charles, and Thomas D Petes (2014). "High-resolution mapping of two types of spontaneous mitotic gene conversion events in Saccharomyces cerevisiae". *Genetics* 198.1, pp. 181–192. DOI: 10.1534/genetics.114.167395.

Yuan, Yuan, Young Seok Ju, Youngwook Kim, Jun Li, Yumeng Wang, Yang Yang, Inigo Martincorena, Chad Creighton, John N Weinstein, et al. (2017). "Comprehensive Molecular Characterization of Mitochondrial Genomes in Human Cancers". *bioRxiv.* DOI: 10.1101/161356.

Yung, Christina K, Brian D O'Connor, Sergei Yakneen, Junjun Zhang, Kyle Ellrott, Kortine Kleinheinz, Naoki Miyoshi, Keiran M Raine, Romina Royo, et al. (2017). "Large-Scale Uniform Analysis of Cancer Whole Genomes in Multiple Computing Environments". *bioRxiv.* DOI: 10.1101/161638.

Zack, Travis I, Stephen E Schumacher, Scott L Carter, Andre D Cherniack, Gordon Saksena, Barbara Tabak, Michael S Lawrence, Cheng-Zhong Zhsng, Jeremiah Wala, et al. (2013). "Pan-cancer patterns of somatic copy number alteration". *Nat. Genet.* 45.10, pp. 1134–1140. DOI: 10.1038/ng.2760.

Zasadil, Lauren M, Eric M C Britigan, and Beth A Weaver (2013). "2n or not 2n: Aneuploidy, polyploidy and chromosomal instability in primary and tumor cells". *Semin. Cell Dev. Biol.* 24.4, pp. 370–379. DOI: 10.1016/j.semcdb.2013.02.001.

Zhang, Cheng-Zhong, Alexander Spektor, Hauke Cornils, Joshua M Francis, Emily K Jackson, Shiwei Liu, Matthew Meyerson, and David Pellman (2015). "Chromothripsis from DNA damage in micronuclei". *Nature* 522.7555, pp. 179–184. DOI: 10.1038/nature14493.

Zhang, Feng, Claudia M B Carvalho, and James R Lupski (2009). "Complex human chromosomal and genomic rearrangements". *Trends Genet.* 25.7, pp. 298–307. DOI: 10.1016/j.tig.2009.05.005.

Zhang, Xiaoyang, Peter S Choi, Joshua M Francis, Marcin Imielinski, Hideo Watanabe, Andrew D Cherniack, and Matthew Meyerson (2016). "Identification of focally amplified lineage-specific super-enhancers in human epithelial cancers". *Nat. Genet.* 48.2, pp. 176–182. DOI: 10.1038/ng.3470.

Zhang, Yu, Rachel Patton McCord, Yu-Jui Ho, Bryan R Lajoie, Dominic G Hildebrand, Aline C Simon, Michael S Becker, Frederick W Alt, and Job Dekker (2012). "Spatial organization of the mouse genome and its role in recurrent chromosomal translocations". *Cell* 148.5, pp. 908–921. DOI: 10.1016/j.cell.2012.02.002.

Zhao, Xiaojun, Cheng Li, J Guillermo Paez, Koei Chin, Pasi A Jänne, Tzu-Hsiu Chen, Luc Girard, John Minna, David Christiani, et al. (2004). "An integrated view of copy number and allelic alterations in the cancer genome using single nucleotide polymorphism arrays". *Cancer Res.* 64.9, pp. 3060–3071.

Zhuang, Jiali and Zhiping Weng (2015). "Local sequence assembly reveals a high-resolution profile of somatic structural variations in 97 cancer genomes". *Nucleic Acids Res.* 43.17, pp. 8146–8156. DOI: 10.1093/nar/gkv831.

List of Tables

List of Figures