

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 ($j = 1, \dots, N$).

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

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

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

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

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 ($g = 1, \dots, P$). For each cancer sample, observe M_{gj} mutations across V mutation classes ($j = 1, \dots, N_g$). 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

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

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

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

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

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

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

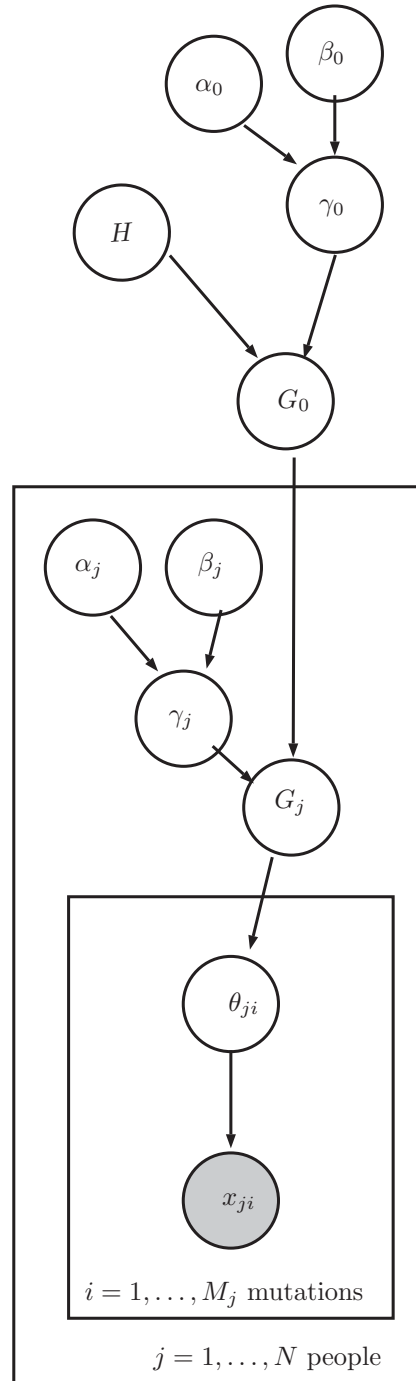


Figure B.1: The hierarchical Dirichlet process mixture model for one group of cancer samples.

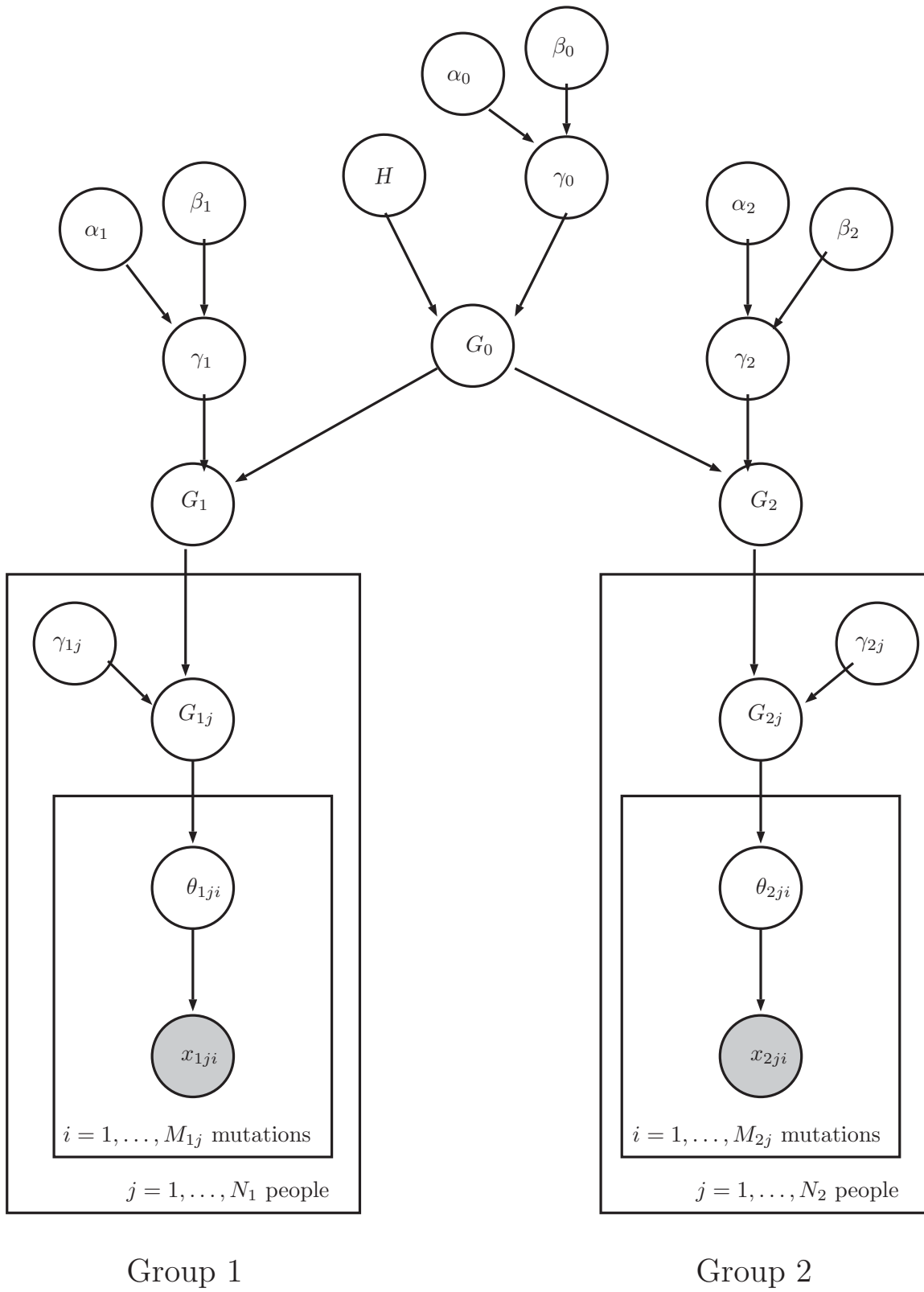


Figure B.2: The hierarchical Dirichlet process mixture model for two groups. Gamma priors for γ_{1j} and γ_{2j} not shown for convenience.

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

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

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

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

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

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

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 \mathbf{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

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

where $n_{..k}^{-\mathbf{x}_{jt}}$ 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 $n_{..kv}^{-\mathbf{x}_{jt}}$ 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 \mathbf{x}_{jt} given they originate from a

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

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

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 for $j = 1, \dots, N$). 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 , t_{ji} and k_{jt} .

Sampling 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:

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

where $n_{jt\cdot}^{-ji}$ 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

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

Sampling 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

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

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 chromoanasyntesis 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

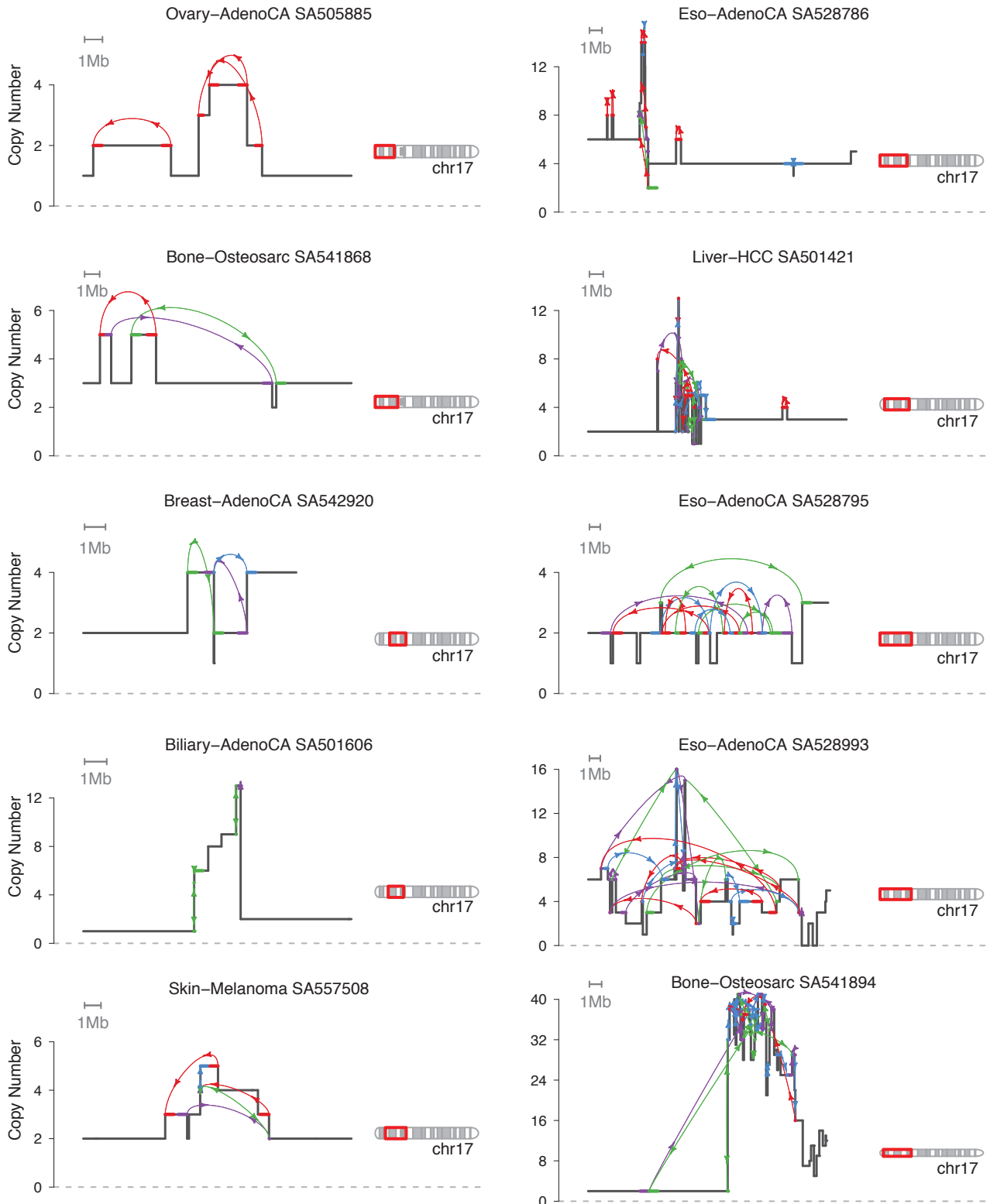


Figure D.1: All intrachromosomal BPTT on the *p*-arm of chromosome 17 in ten different samples, coloured by orientation. Blue denotes deletion type $\langle + - \rangle$, red is tandem duplication type $\langle - + \rangle$, and purple and green indicate inversion type $\langle ++ \rangle$ or $\langle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

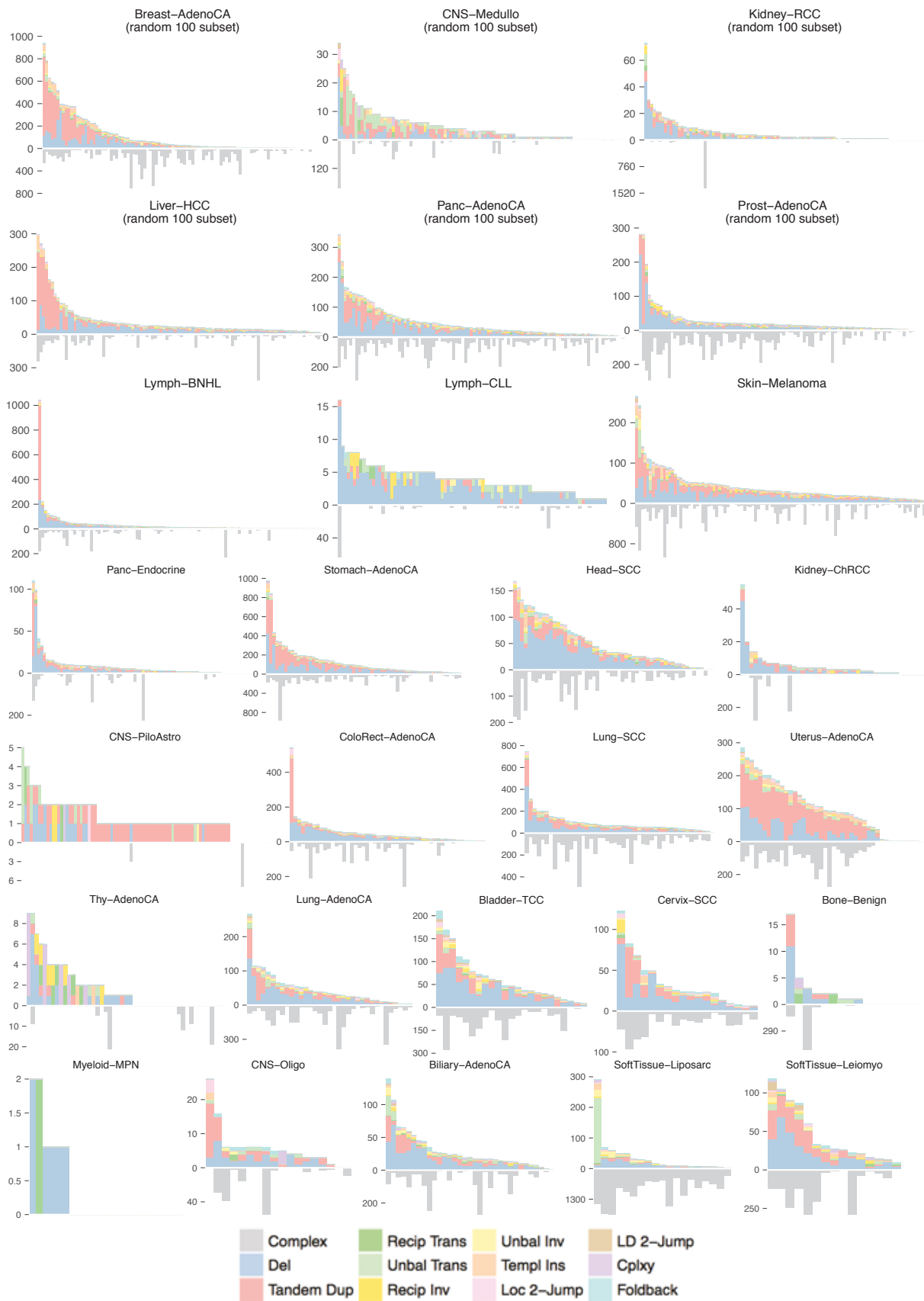


Figure D.2: Per-sample counts of complex (lower) and classified (upper) break-point junctions.

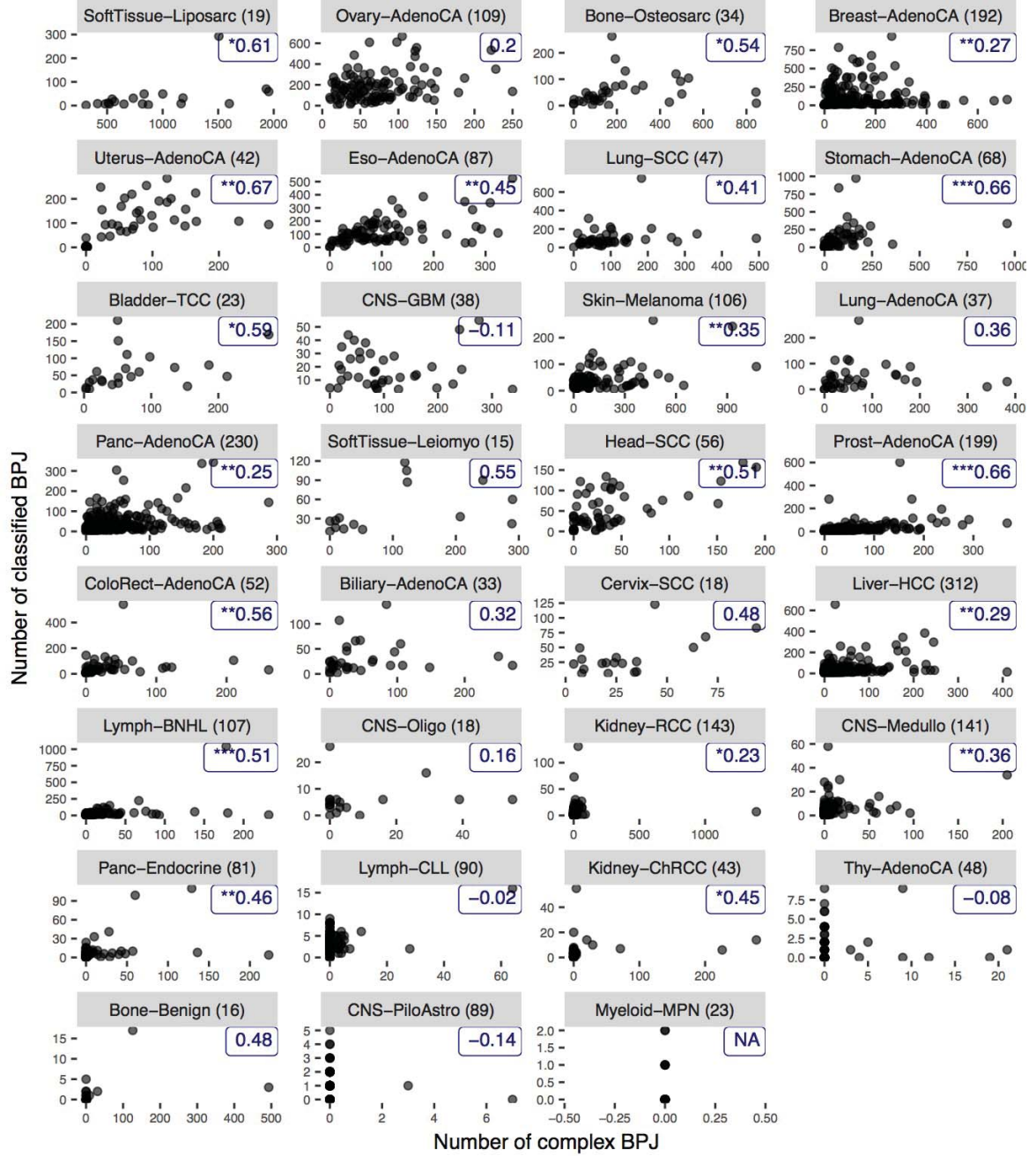


Figure D.3: Spearman's rank correlation coefficient between complex (horizontal) and classified (vertical) BPJ counts in samples grouped by histology. Benjamini-Hochberg-corrected FDR for the null hypothesis of zero correlation is indicated at levels: * < 0.01 , ** < 0.001 , and *** $< 10^{-6}$.

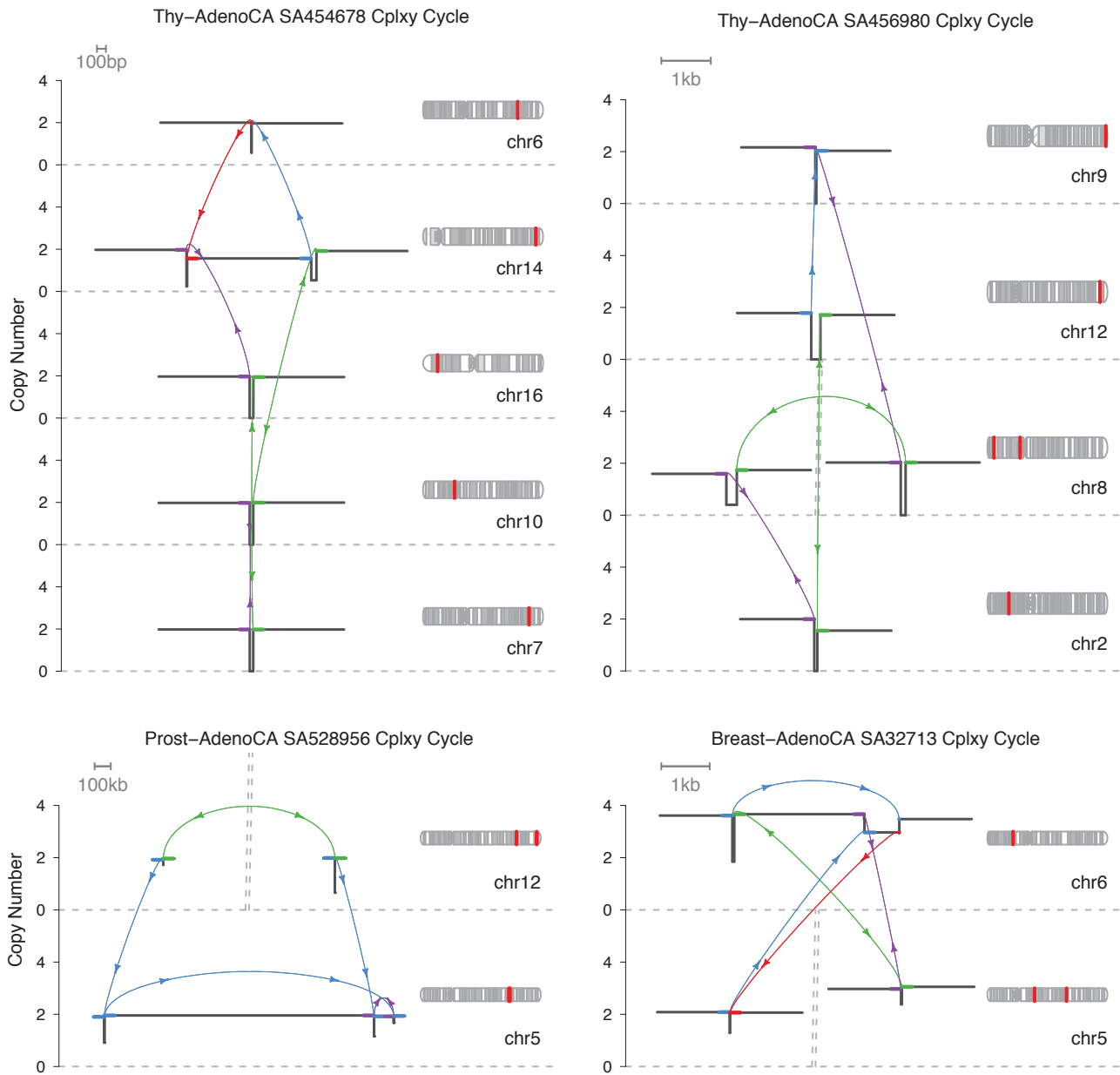


Figure D.4: Four of the longest chromoplexy events

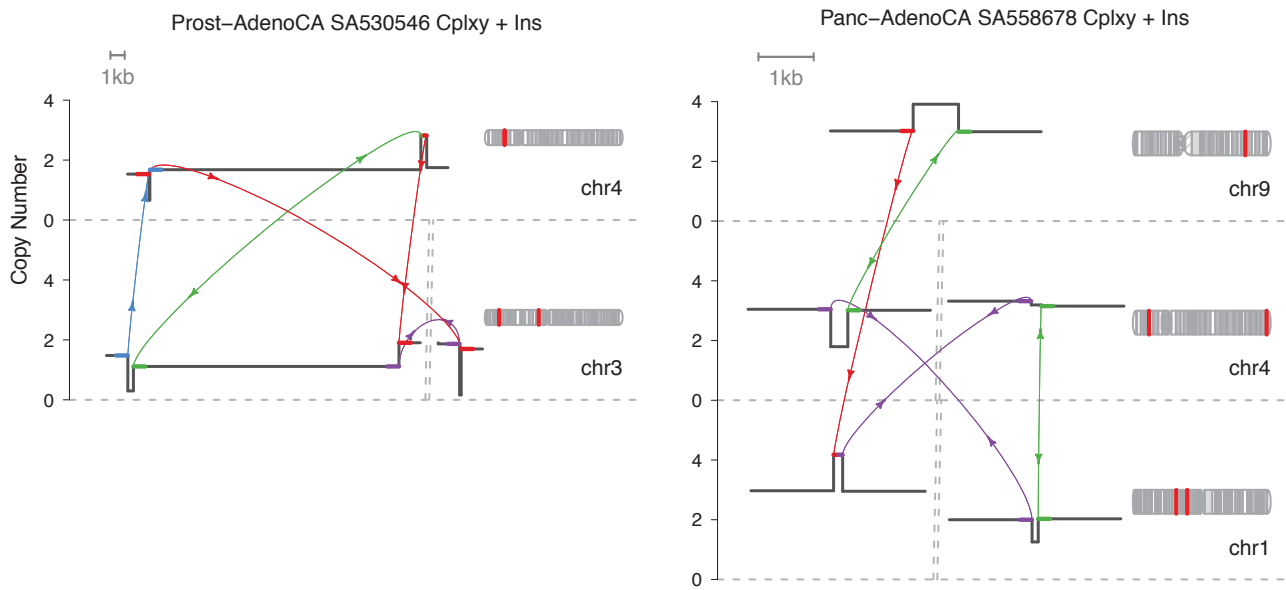


Figure D.5: Two of the longest chromoplexy with insertion events. Note that copy number estimates are unreliable in short segments < 1 kb.

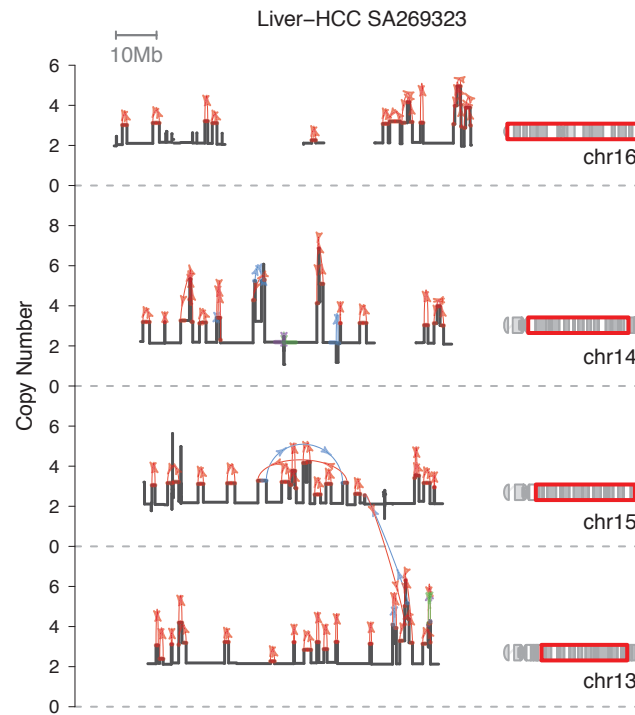


Figure D.6: All SV along four representative chromosomes from an unusual liver cancer sample with a high frequency of extremely large tandem duplications.

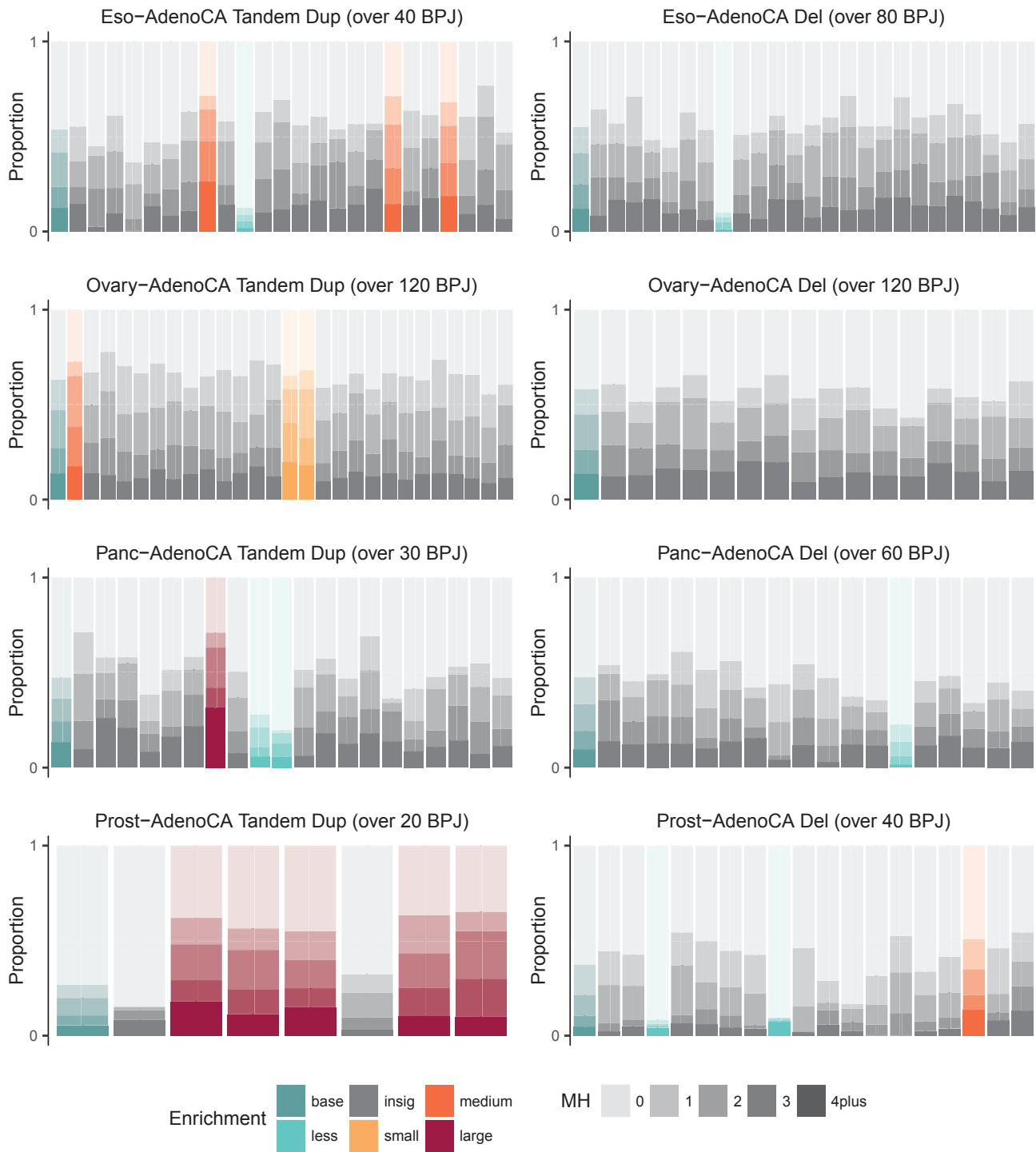


Figure D.7: Distribution of microhomology at the breakpoint junction for deletion and tandem duplication in individual samples with event counts above the indicated threshold. The magnitude of significant enrichment (compared to pool of other samples in the same histology class, shown in leftmost bar) 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.

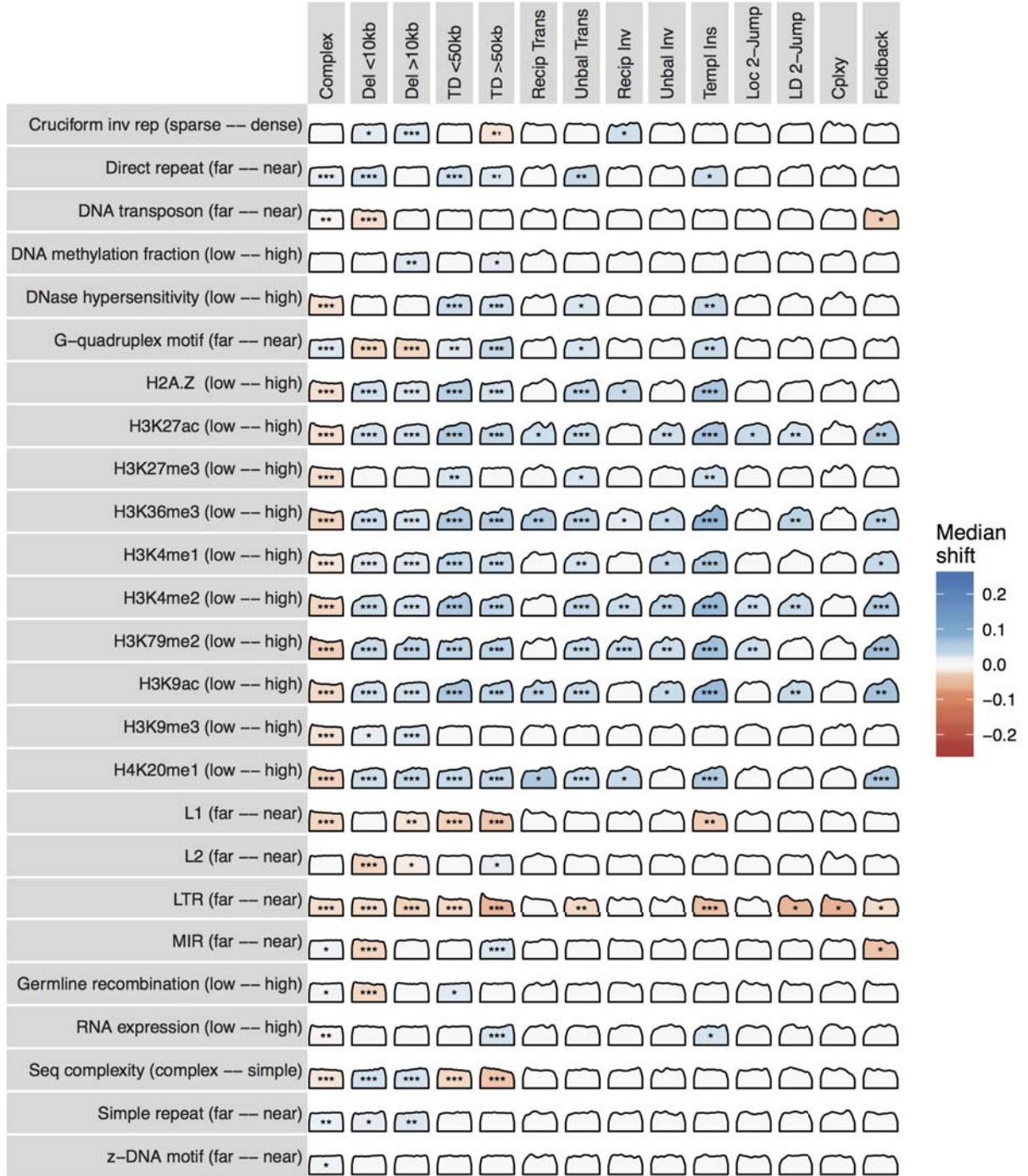


Figure D.8: For each SV class, the quantile distribution of the genomic property metrics at observed breakpoints compared to random positions, with significant departure from uniform quantiles marked by: FDR < 0.01 *, < 0.001 **, and < 10⁻⁶ ***; shading the magnitude of the shift of the median observed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

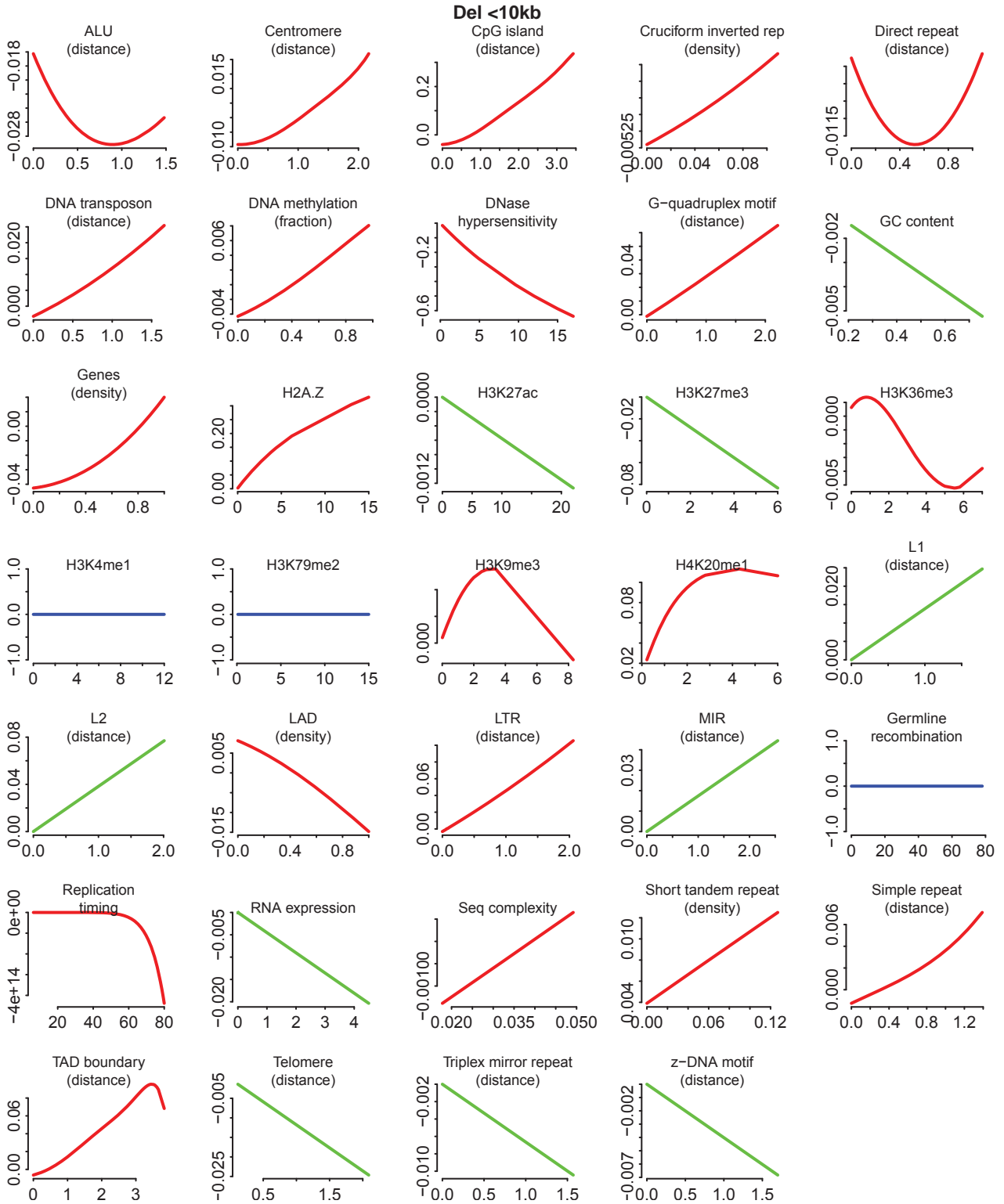


Figure D.9: The optimal lasso GAM for small deletions, with predictor effects on the log odds of a real breakpoint in red for splines, green for linear terms, and blue for removed predictors.

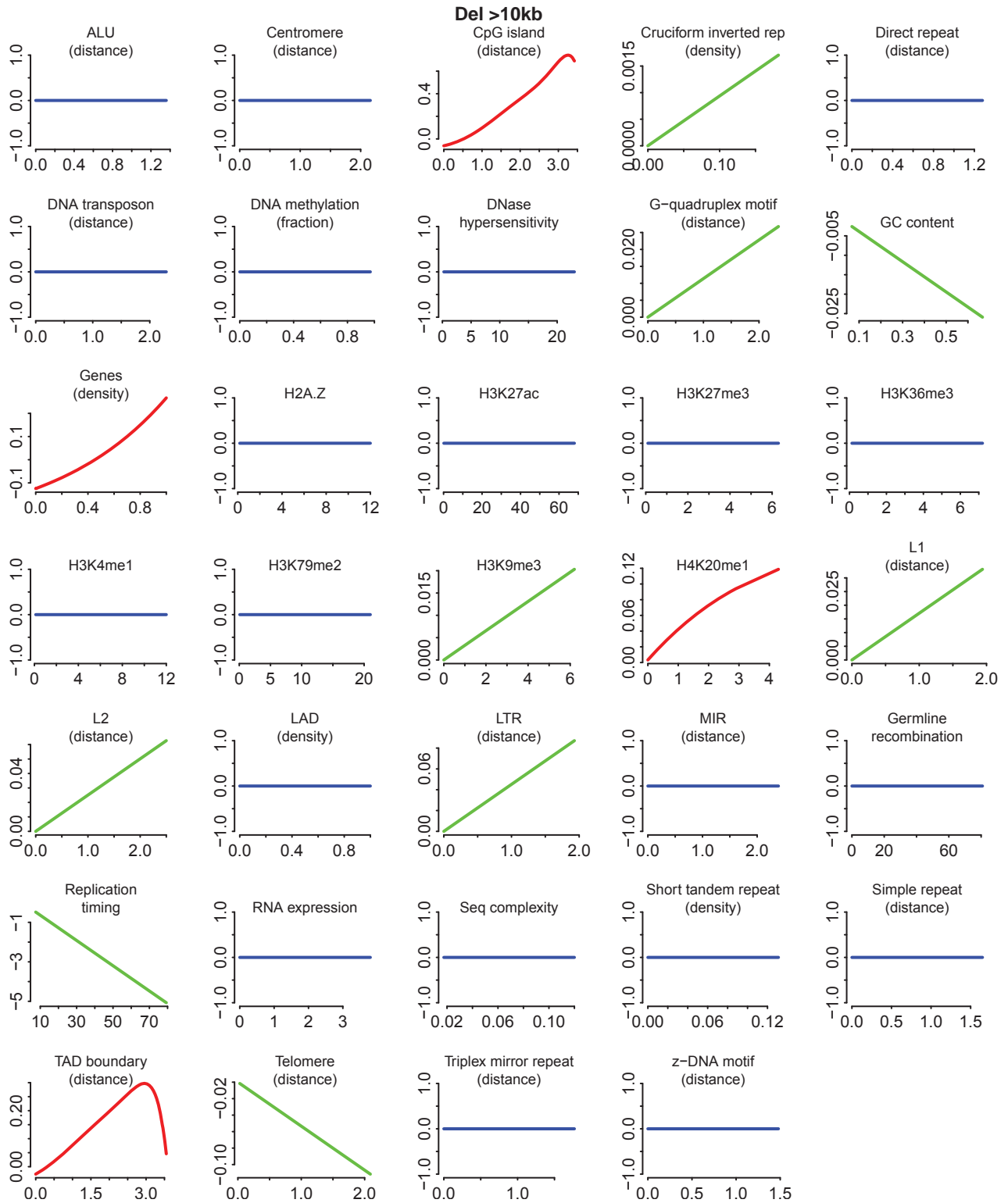


Figure D.10: The optimal lasso GAM for large deletions, with predictor effects on the log odds of a real breakpoint in red for splines, green for linear terms, and blue for removed predictors.

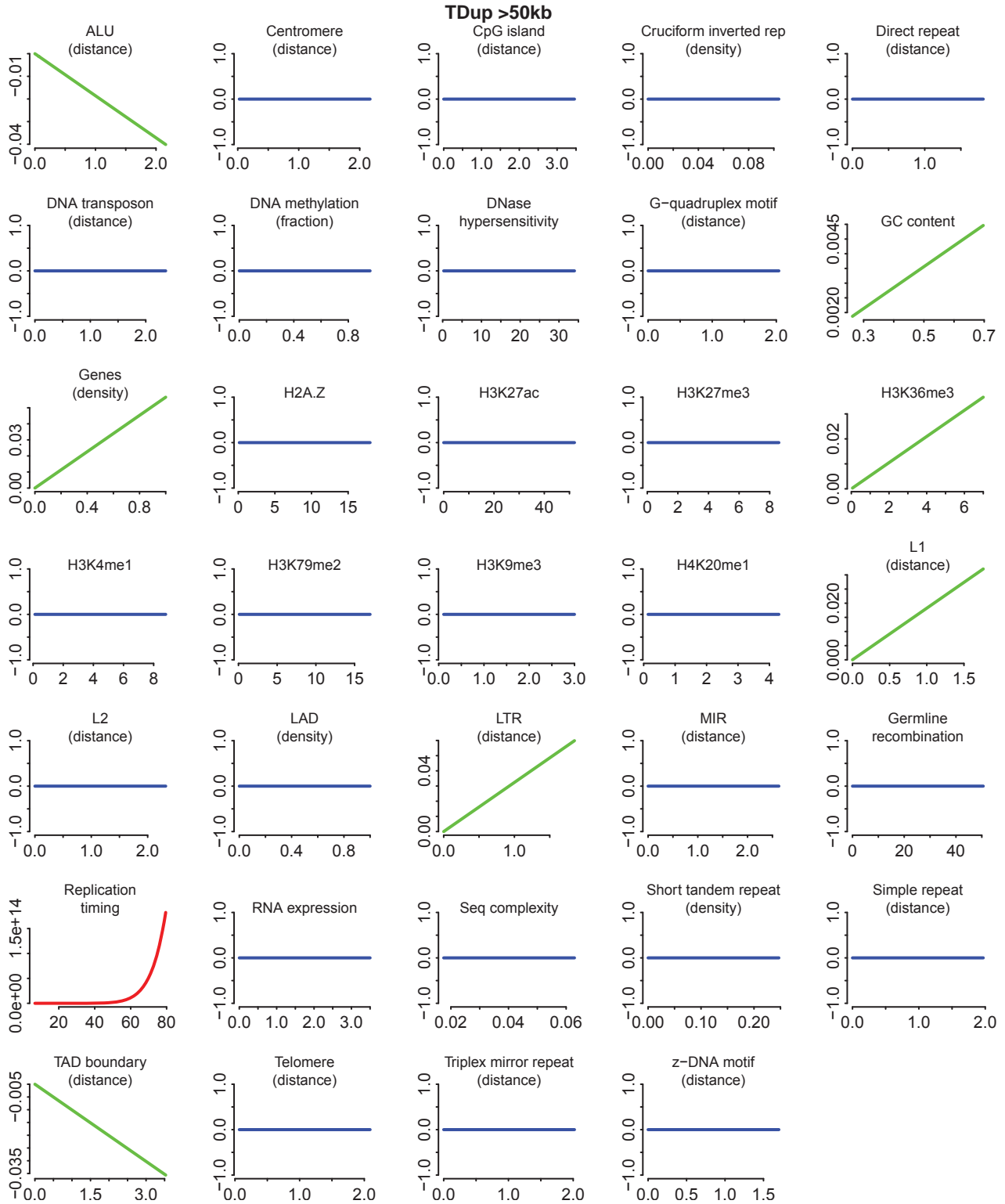


Figure D.11: The optimal lasso GAM for large tandem duplication, with predictor effects on the log odds of a real breakpoint in red for splines, green for linear terms, and blue for removed predictors.

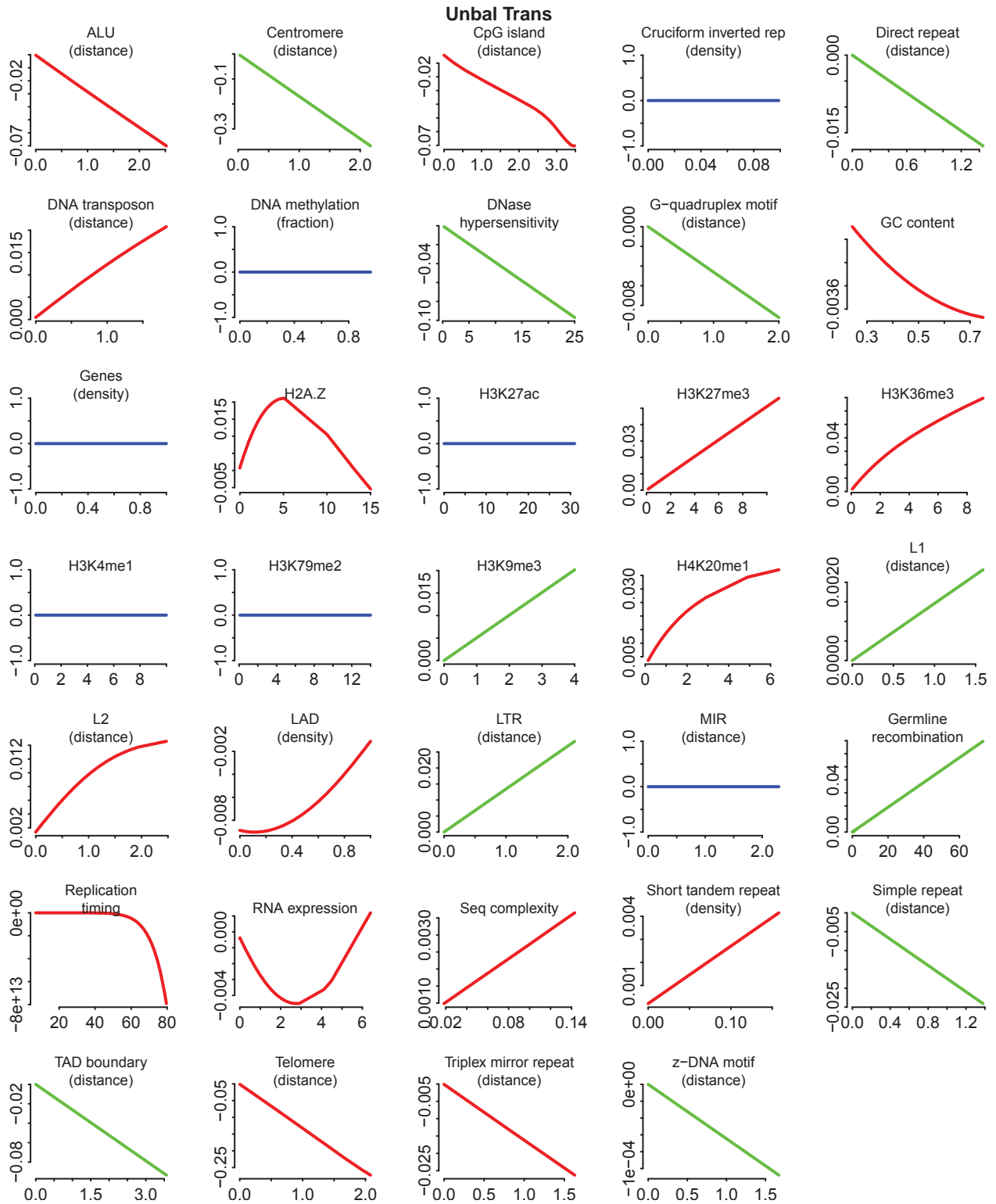


Figure D.12: The optimal lasso GAM for unbalanced translocation, with predictor effects on the log odds of a real breakpoint in red for splines, green for linear terms, and blue for removed predictors.

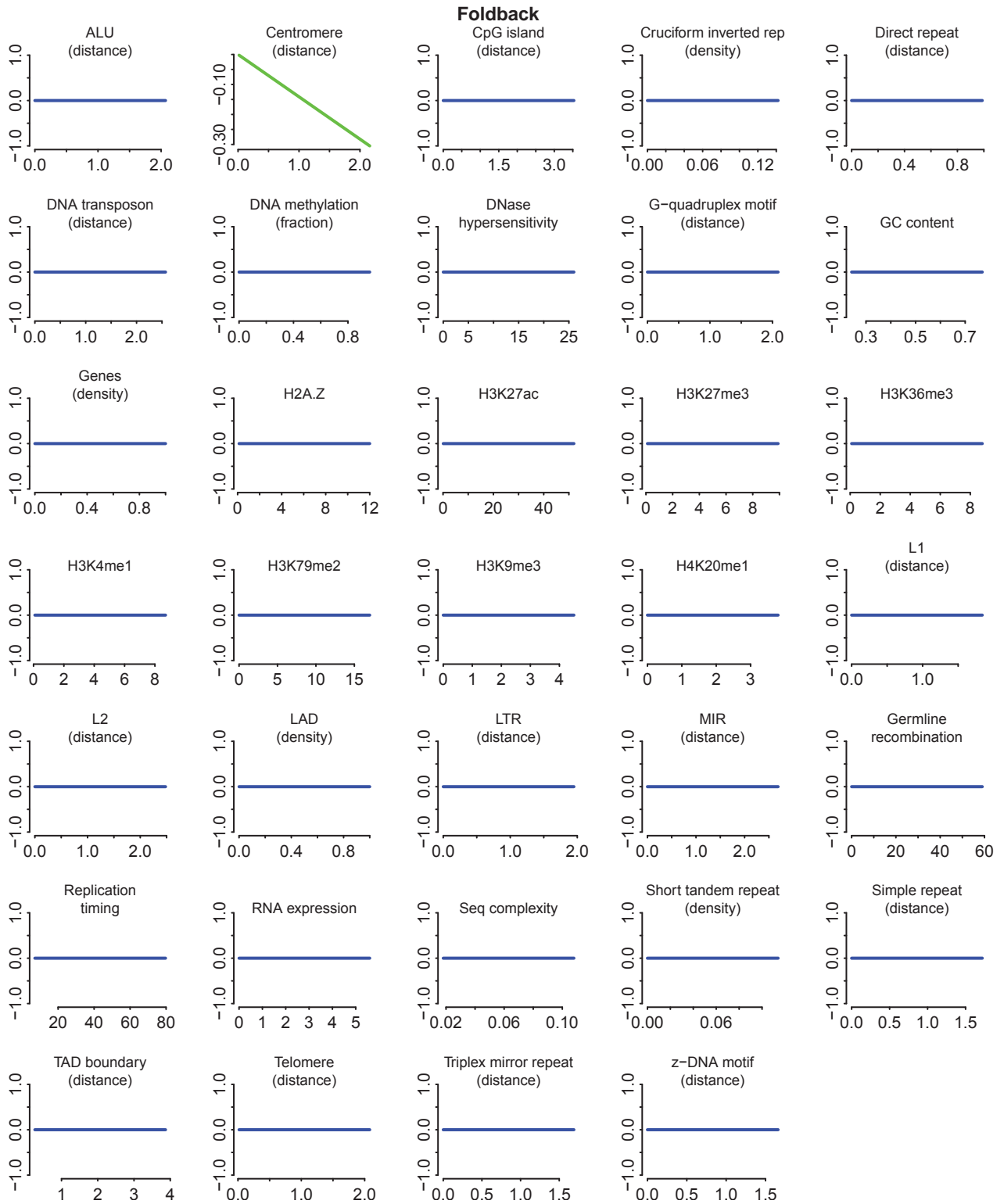


Figure D.13: The optimal lasso GAM for foldback, with predictor effects on the log odds of a real breakpoint in red for splines, green for linear terms, and blue for removed predictors.

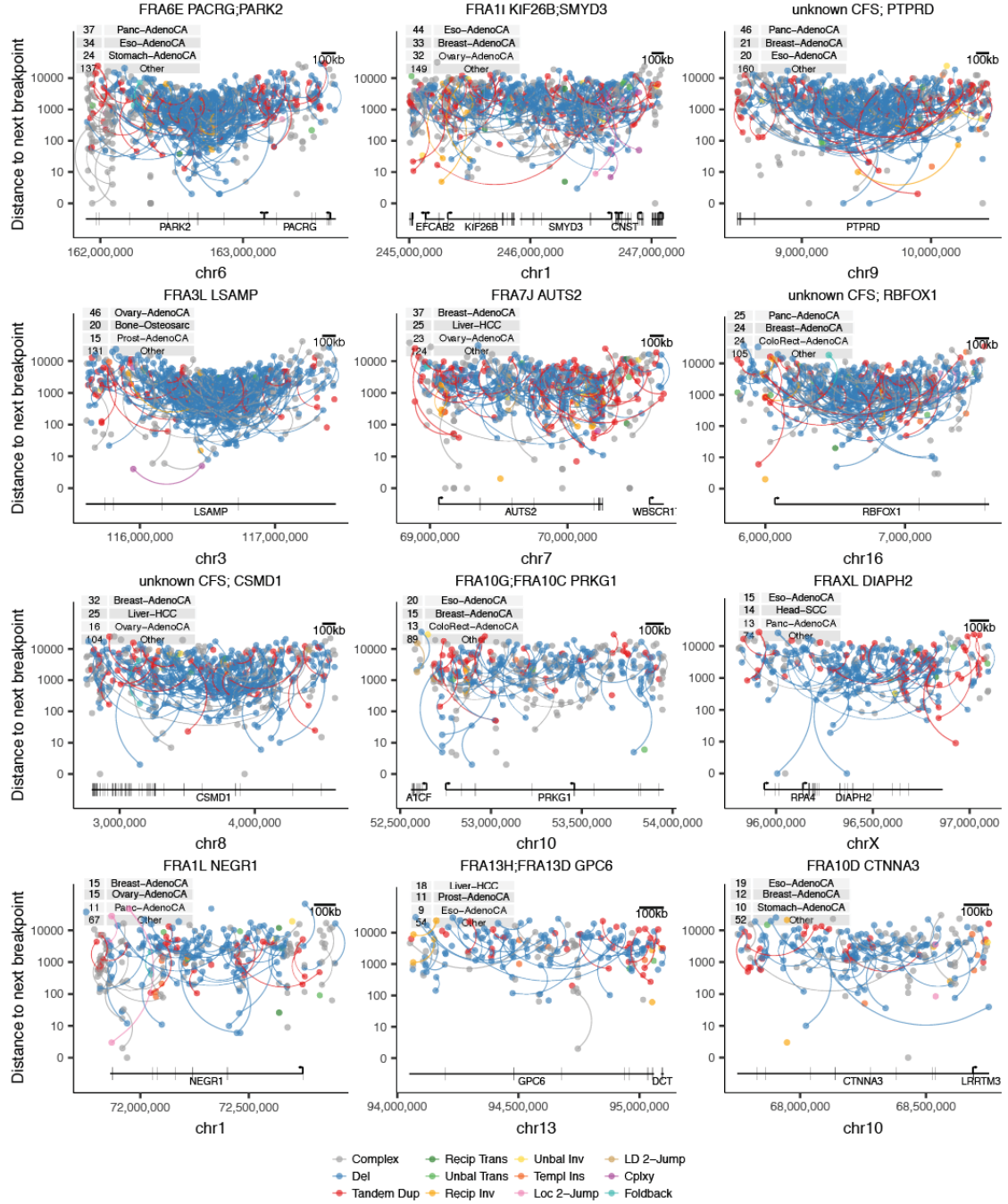


Figure D.14: All sv breakpoint positions in the 12 minor fragile sites. If the two sides of a BPJ are contained within the plotting window, they are joined with a curved line. The number of samples with a breakpoint in the plotting window is annotated top left.

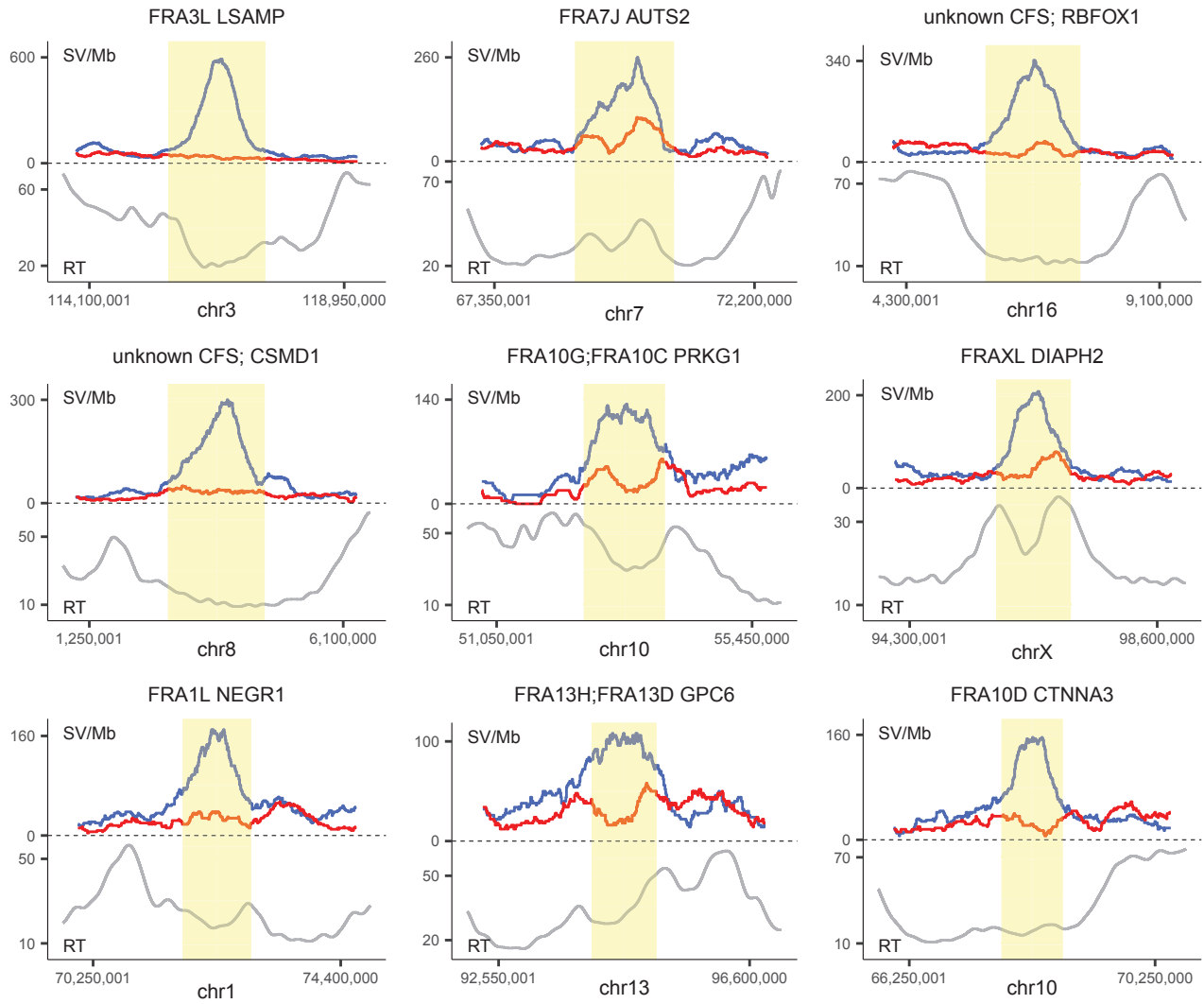


Figure D.15: Extending to 2 Mb flanks either side of nine minor FS marked in yellow, the upper plot shows the density of deletion (blue) and tandem dup (red) breakpoints in 500 kb windows sliding every 10 kb. The lower plot shows the replication timing track, with high values for early and low for late.

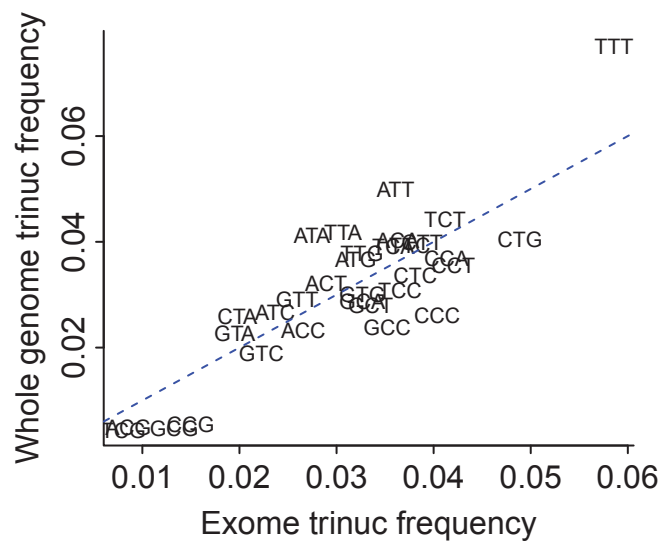


Figure D.16: Trinucleotide frequency in human exome (plus 100 bp flanks) and whole genome (callable regions), with blue line denoting equal values. Reverse complements are consolidated, and reported with the middle base as the pyrimidine (C or T).

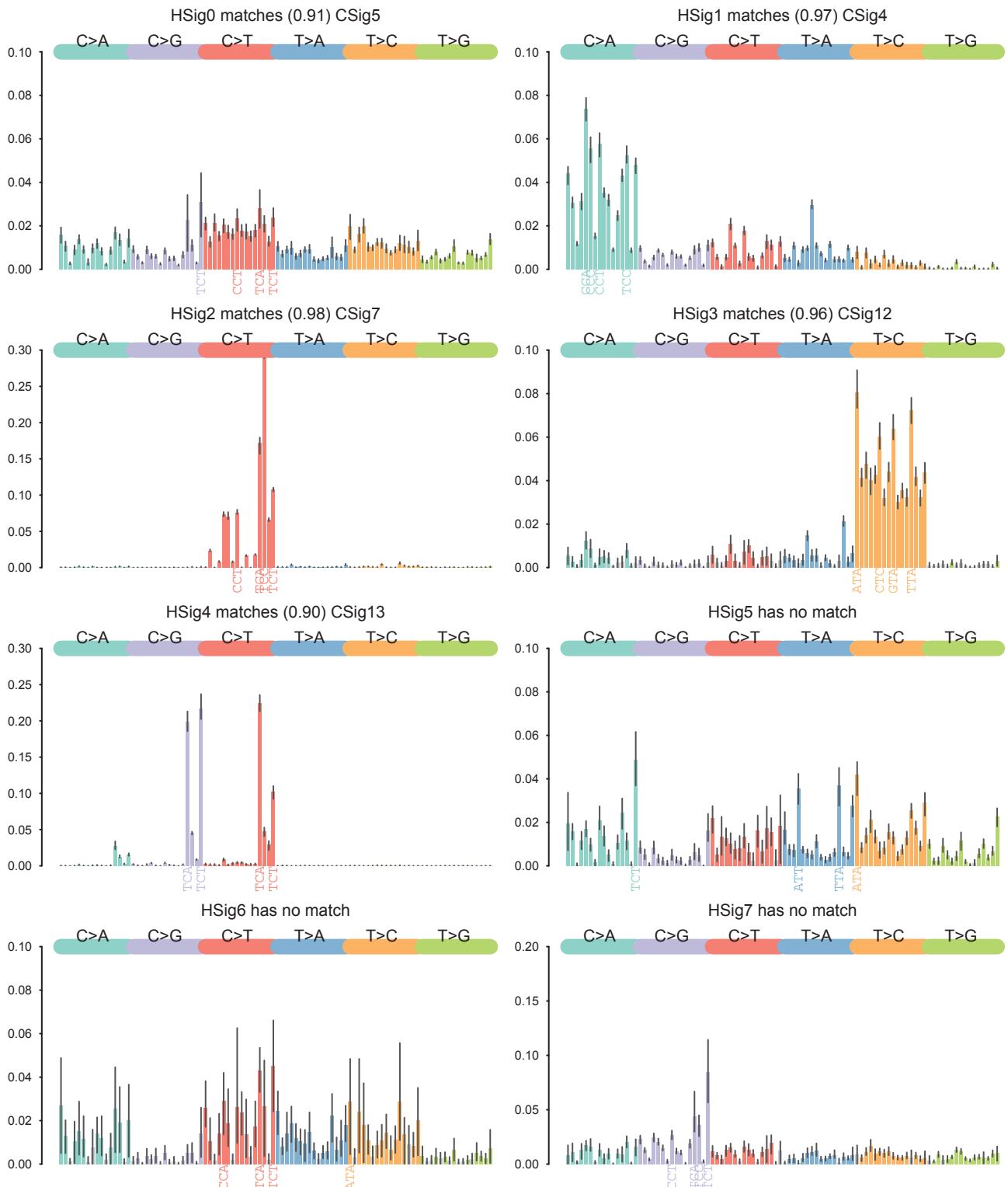


Figure D.17: HDP mutational signatures in discovery dataset (mean and 95% credibility interval from MCMC posterior samples, with non-significant mutation classes in grey and four major peaks labelled with trinucleotide context).

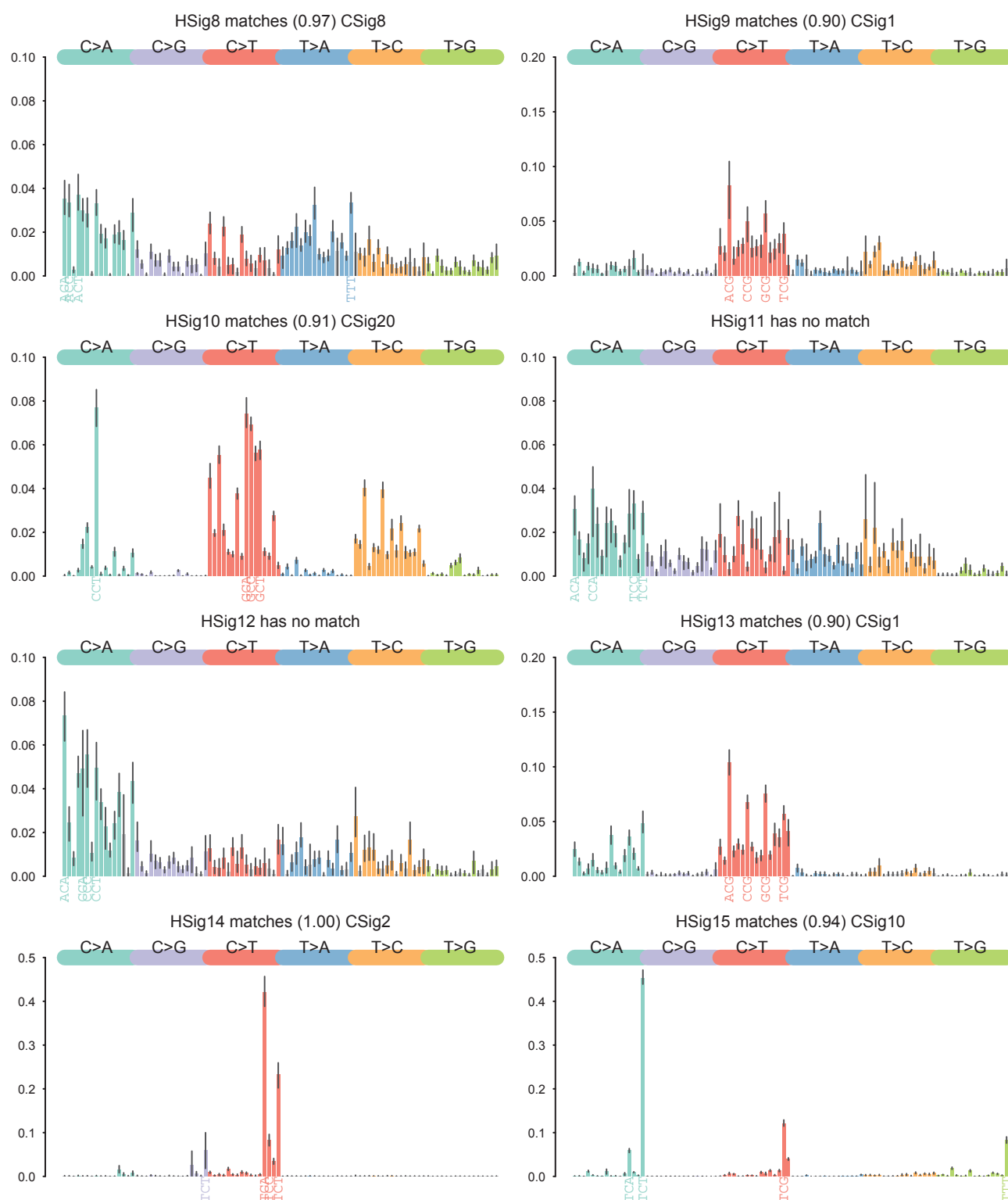


Figure D.17: HDP mutational signatures in discovery dataset (continued from previous)

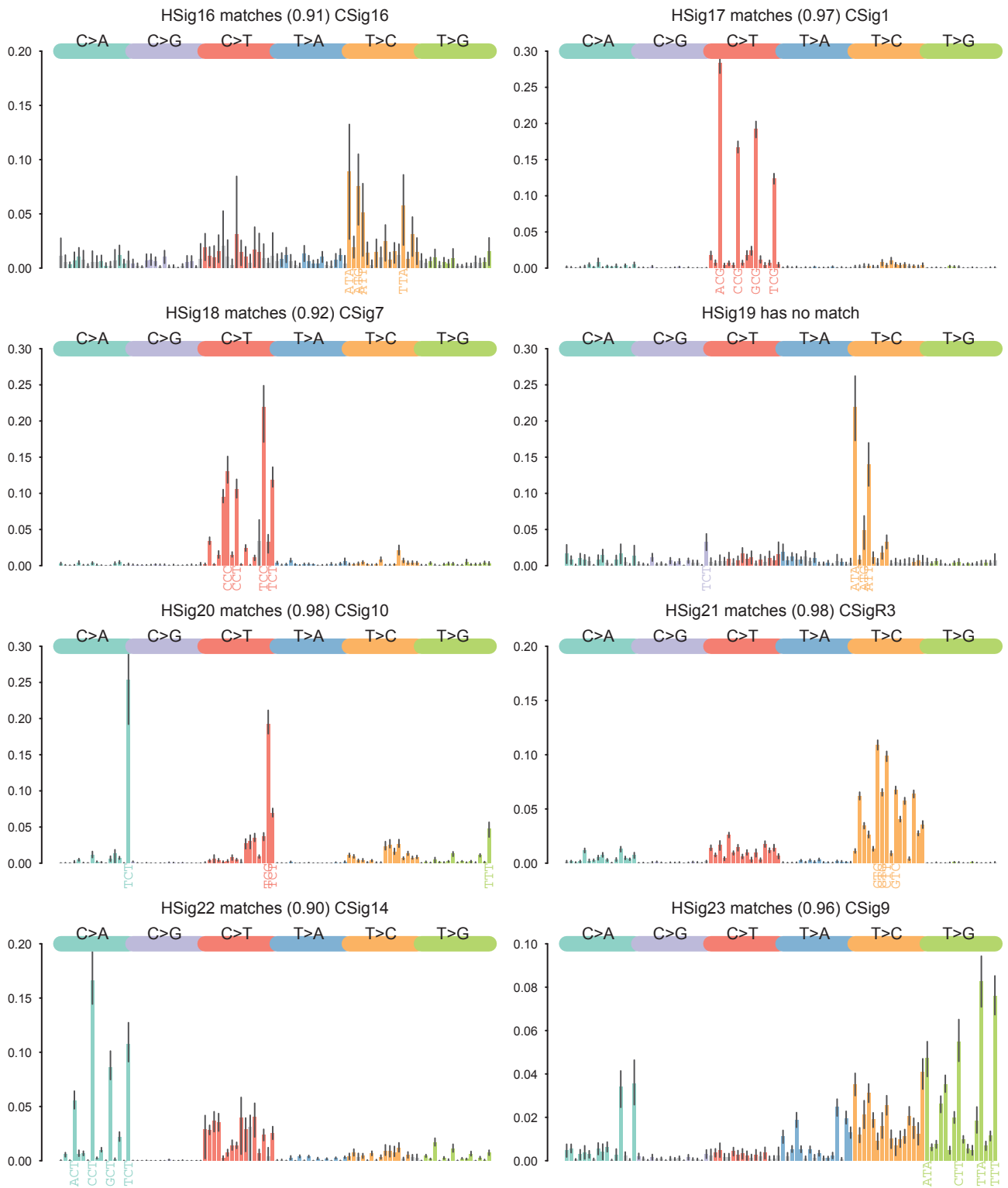


Figure D.17: HDP mutational signatures in discovery dataset (continued from previous)

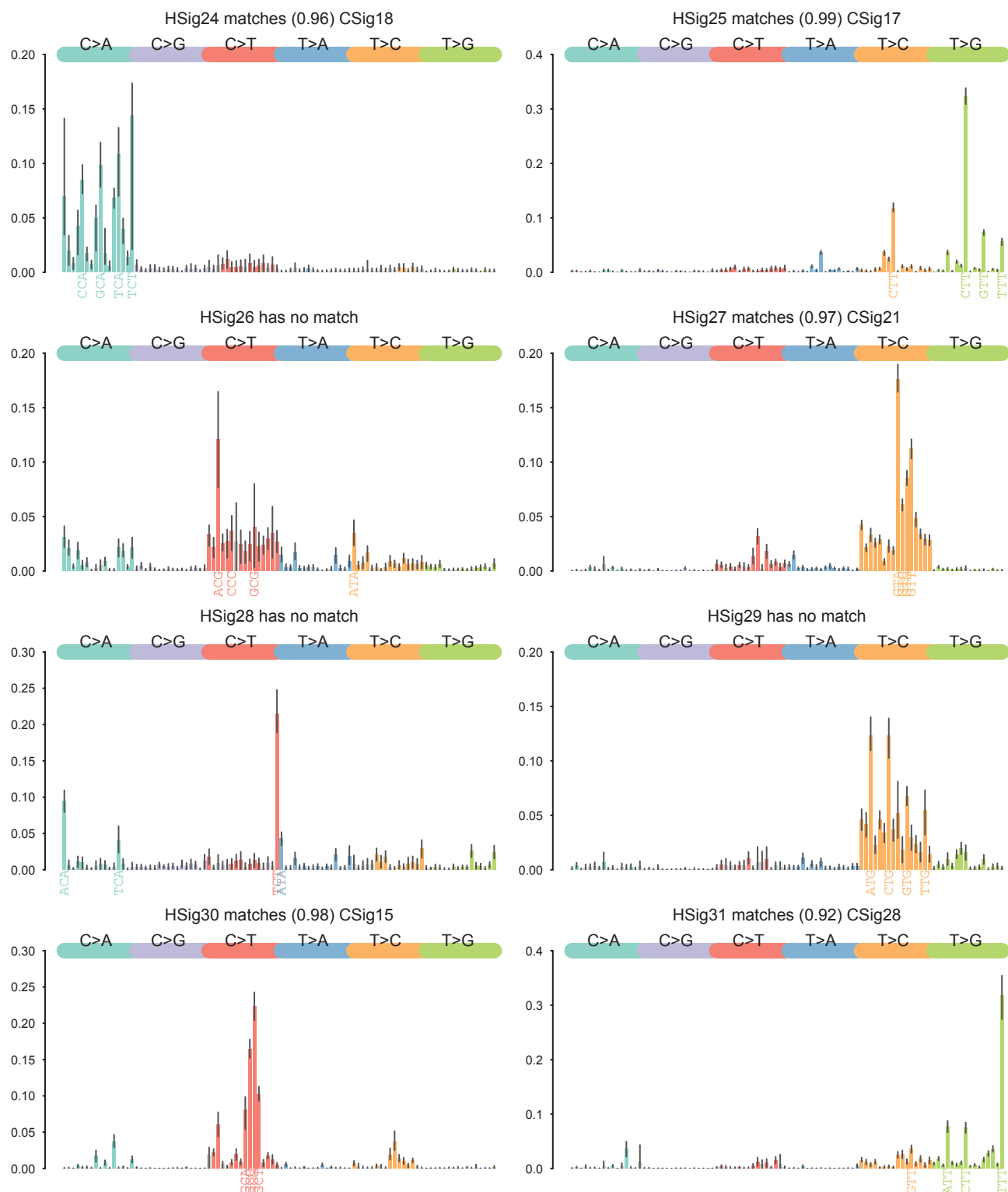


Figure D.17: HDP mutational signatures in discovery dataset (continued from previous)

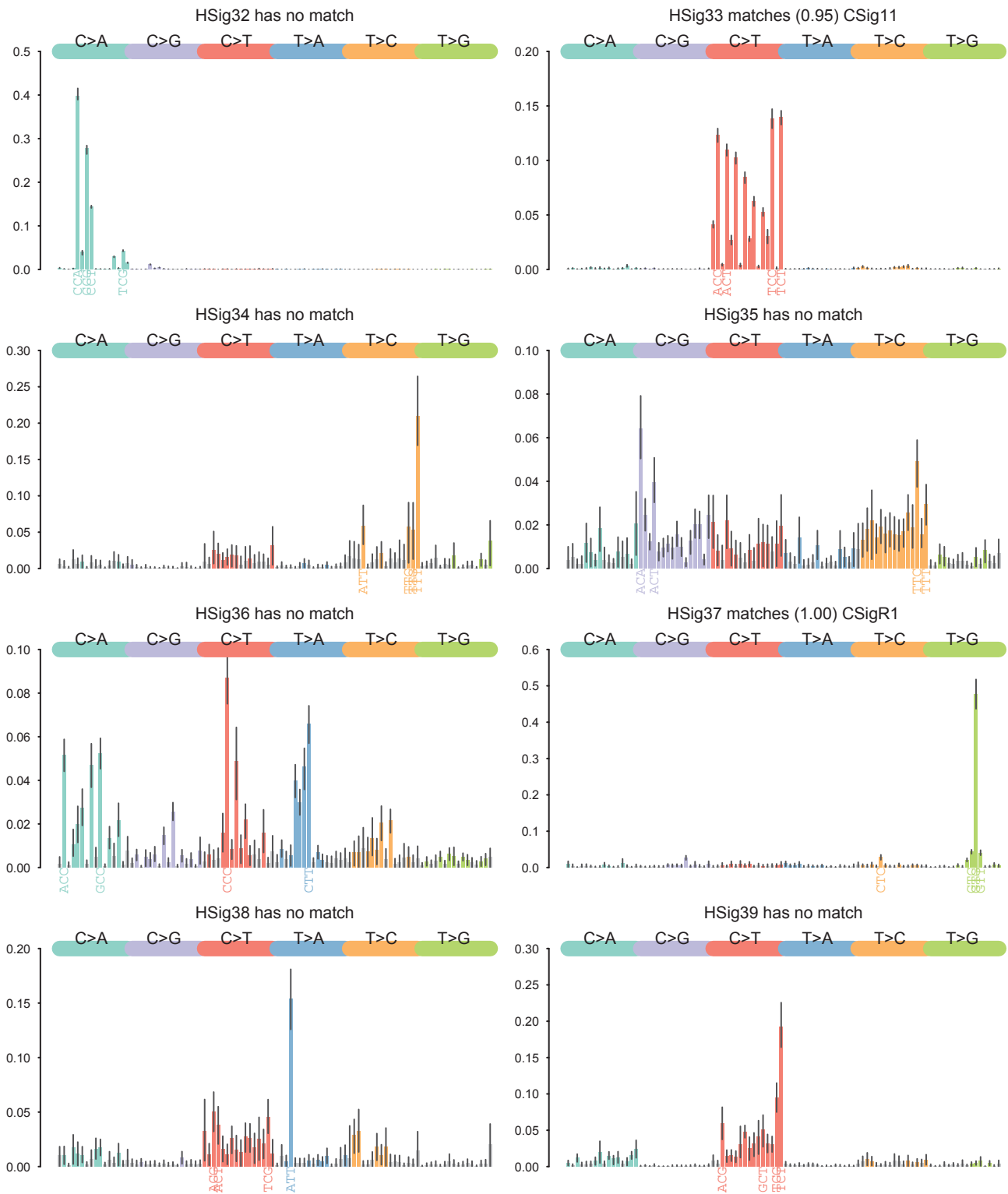


Figure D.17: HDP mutational signatures in discovery dataset (continued from previous)

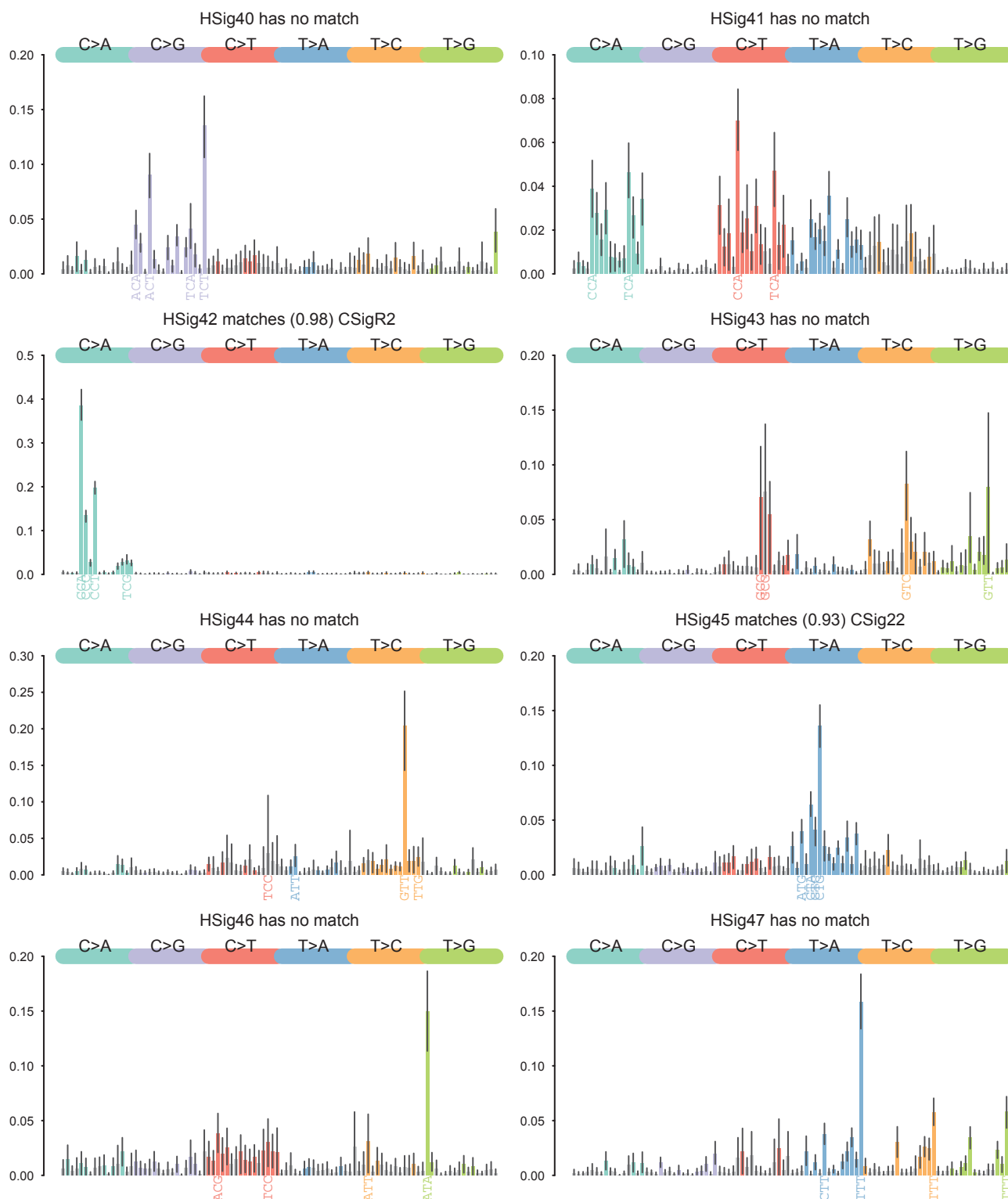


Figure D.17: HDP mutational signatures in discovery dataset (continued from previous)

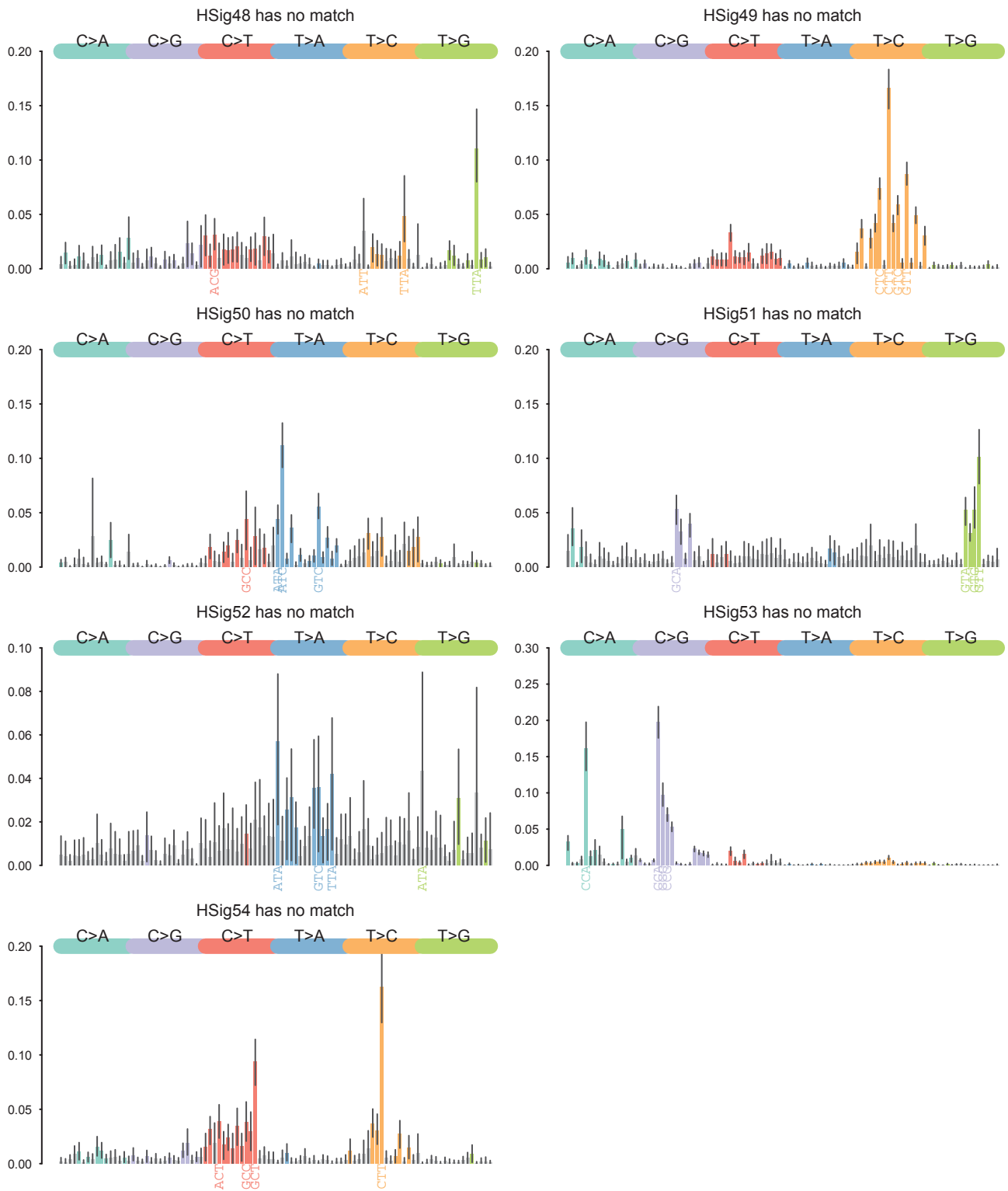


Figure D.17: HDP mutational signatures in discovery dataset (continued from previous)

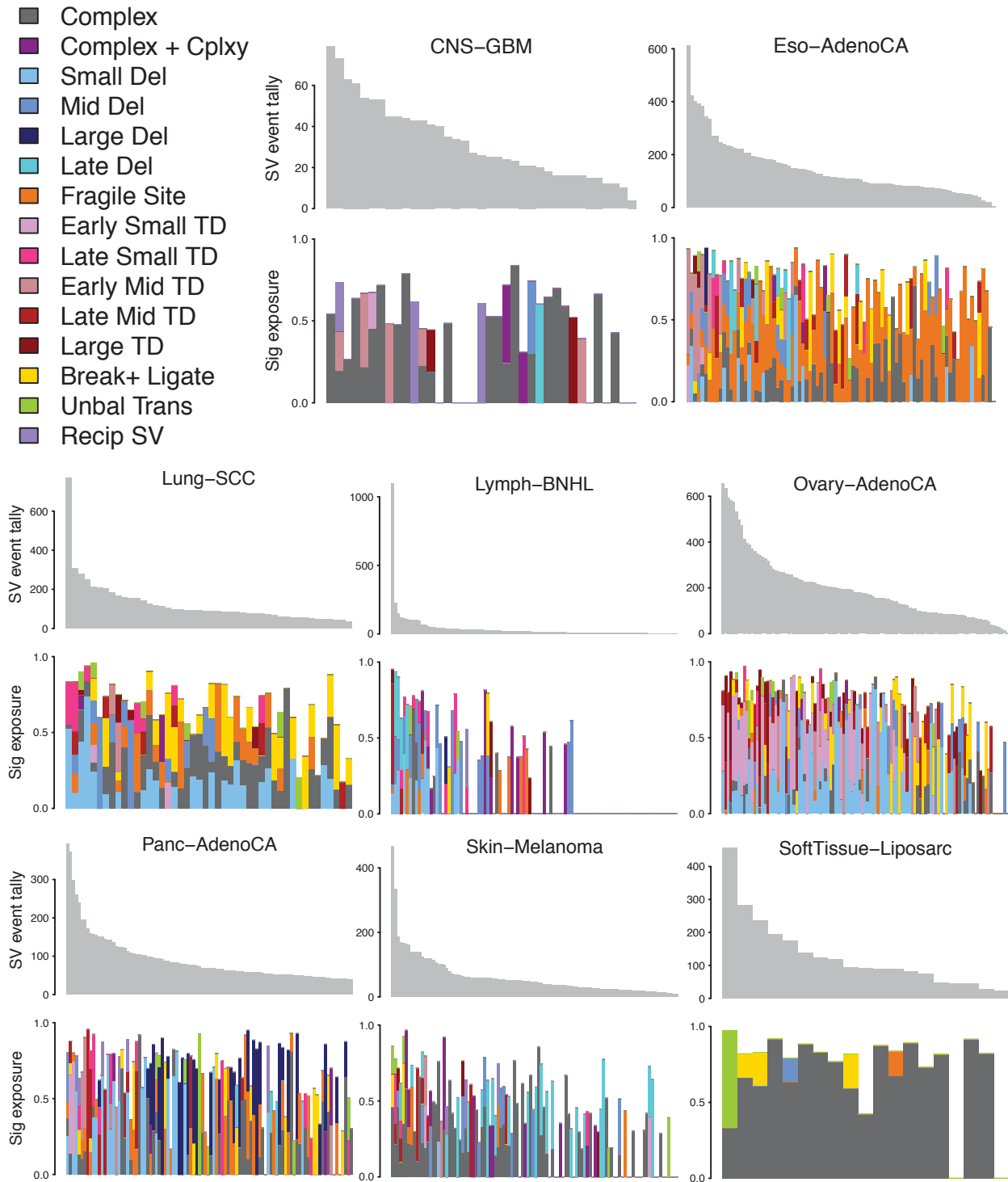


Figure D.18: Average estimated sample exposures to HDP-extracted SV signatures (Figure 4.20) for eight of the PCAWG cancer types. Large cohorts are subset to a maximum of 100 samples for presentation. For each sample, mean estimated exposures are only plotted for significant signatures (95% credibility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

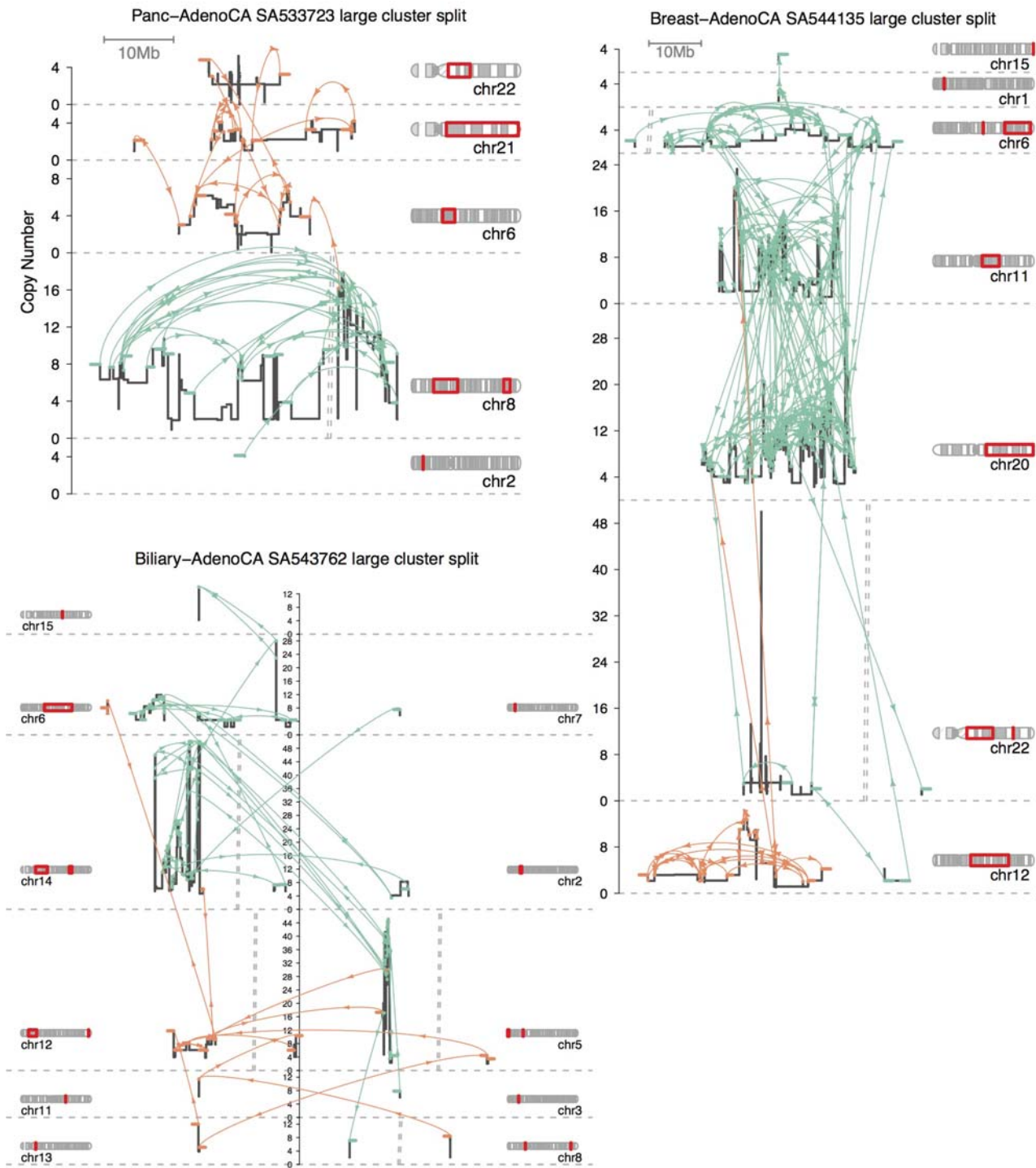


Figure D.19: Large BPJ candidate clusters, split by walktrap graph community detection as described in Section 5.1.1. Node-edge graph visualisations are available in Figures 5.4–5.6. Figure continues on the next page.

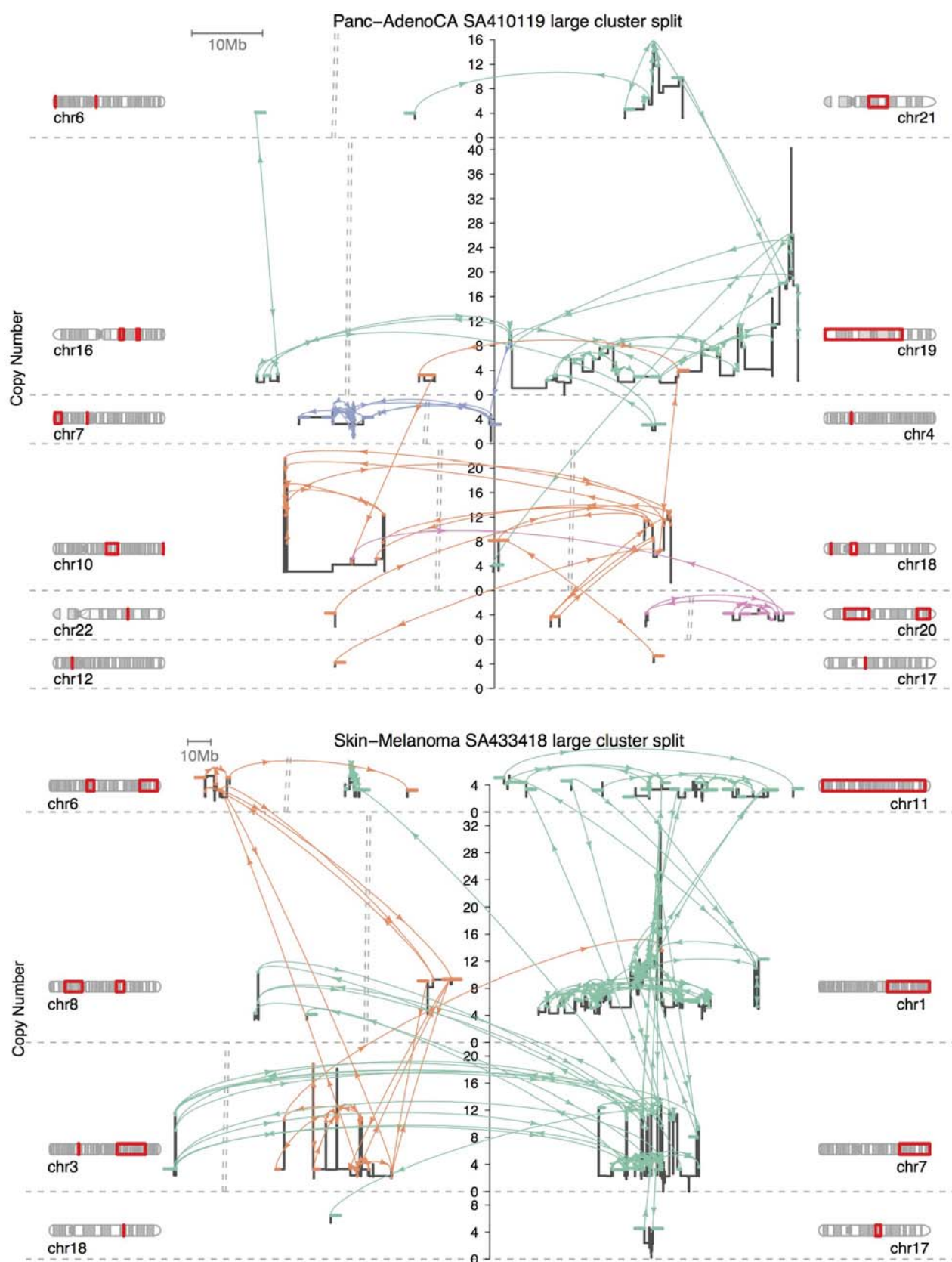


Figure D.19: Large BPJ candidate clusters, split by walktrap graph community detection as described in Section 5.1.1. Node-edge graph visualisations are available in Figures 5.4–5.6. Figure continued from the previous page.

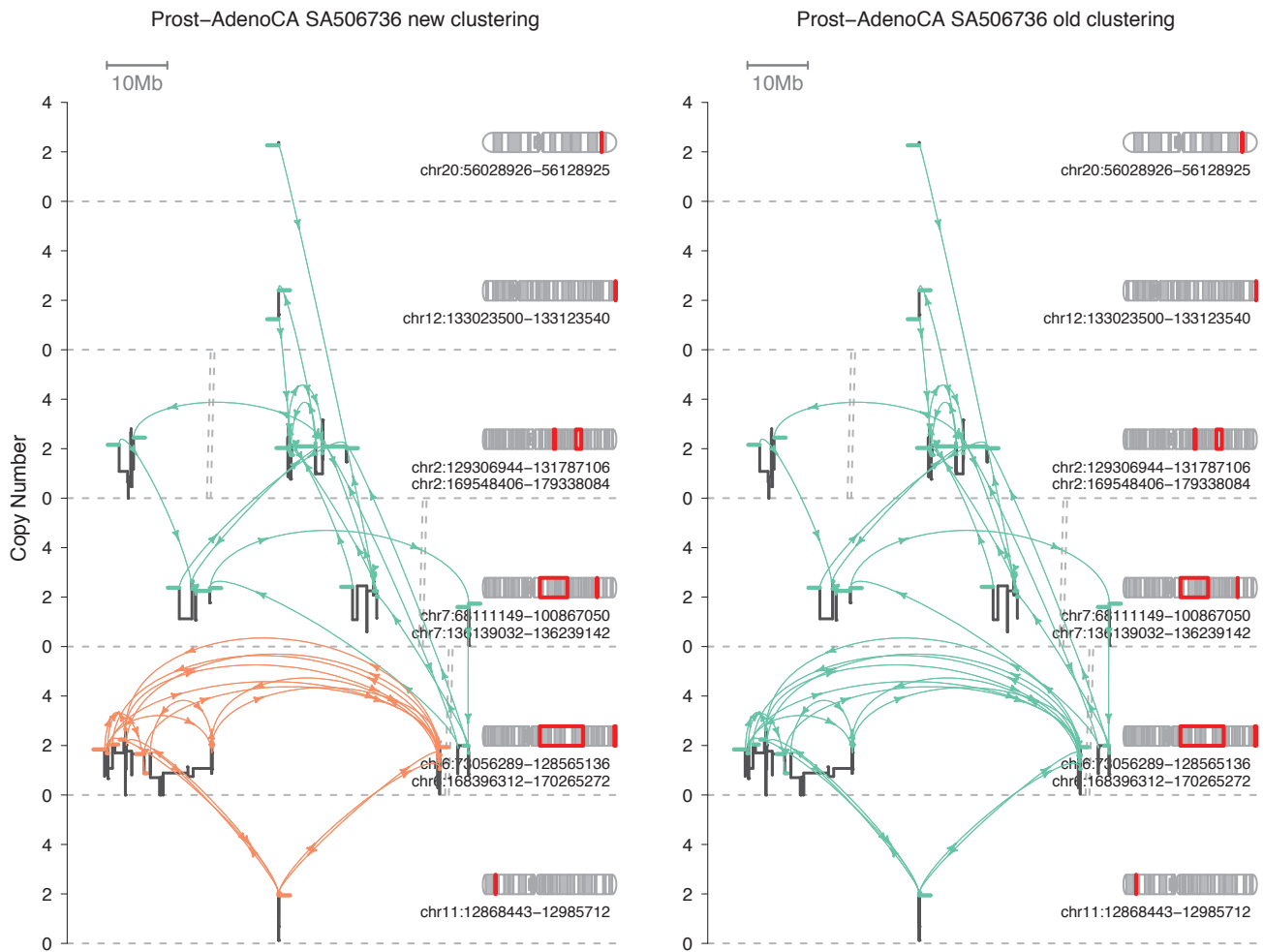


Figure D.20: New and old clustering schemes (more clusters in new) for complex unexplained BPJ in prostate cancer sample SA506736, only showing clusters with disagreement.

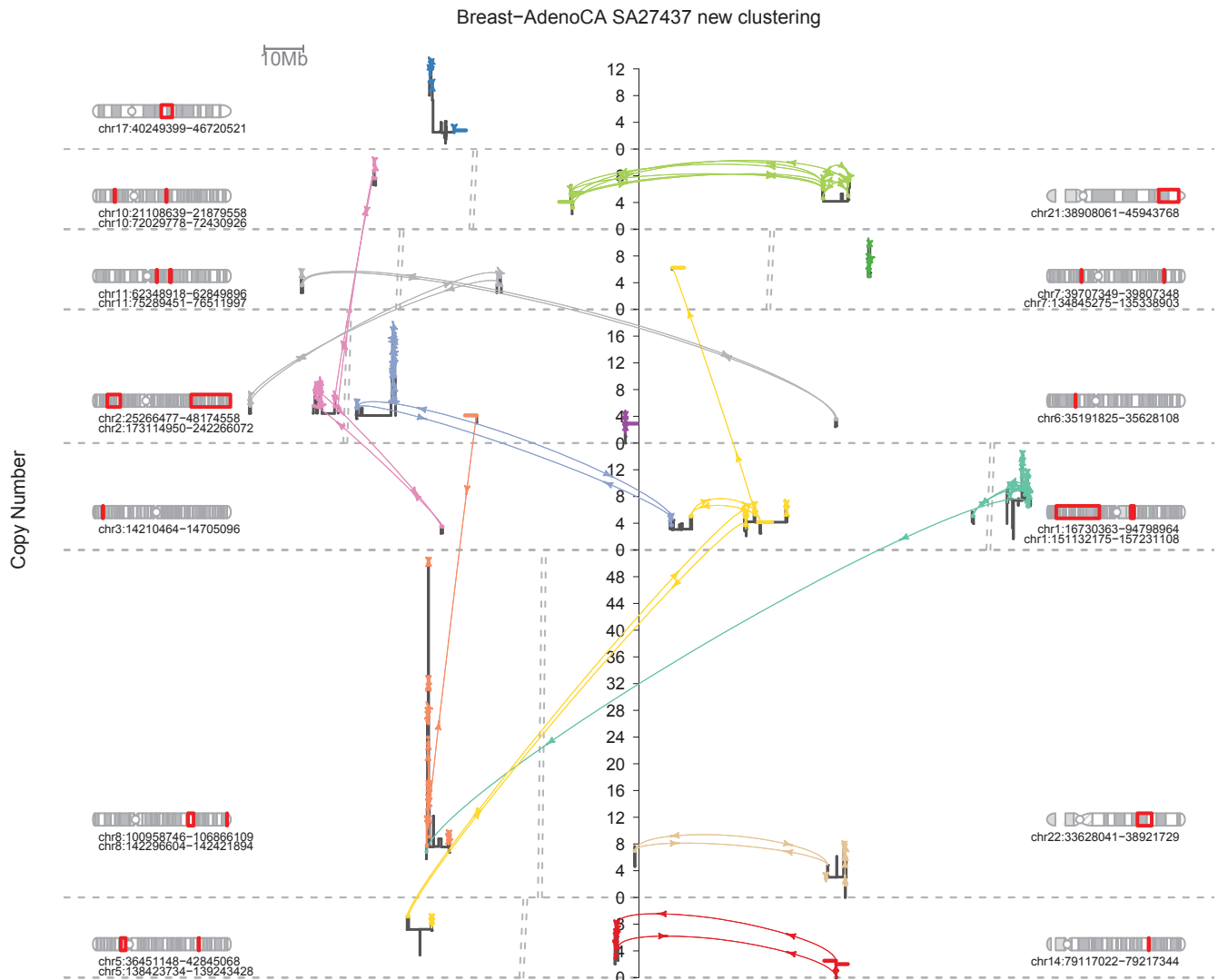


Figure D.21: New and old clustering schemes (more clusters in new) for complex unexplained BPJ in breast cancer sample SA27437, only showing clusters with disagreement. Figure continues on the next page.

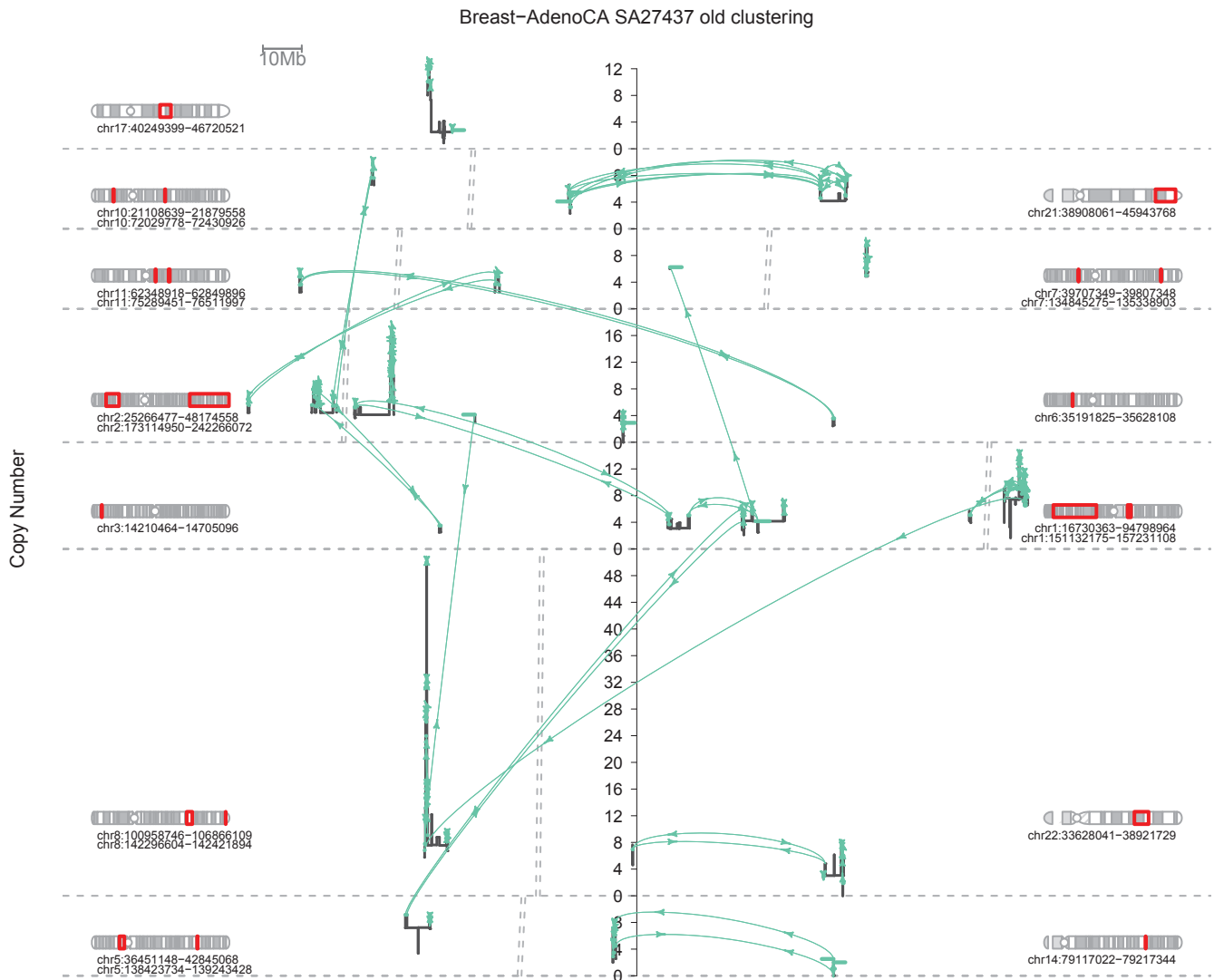


Figure D.21: New and old clustering schemes (more clusters in new) for complex unexplained BPJ in breast cancer sample SA27437, only showing clusters with disagreement. Figure is continued from the previous page.

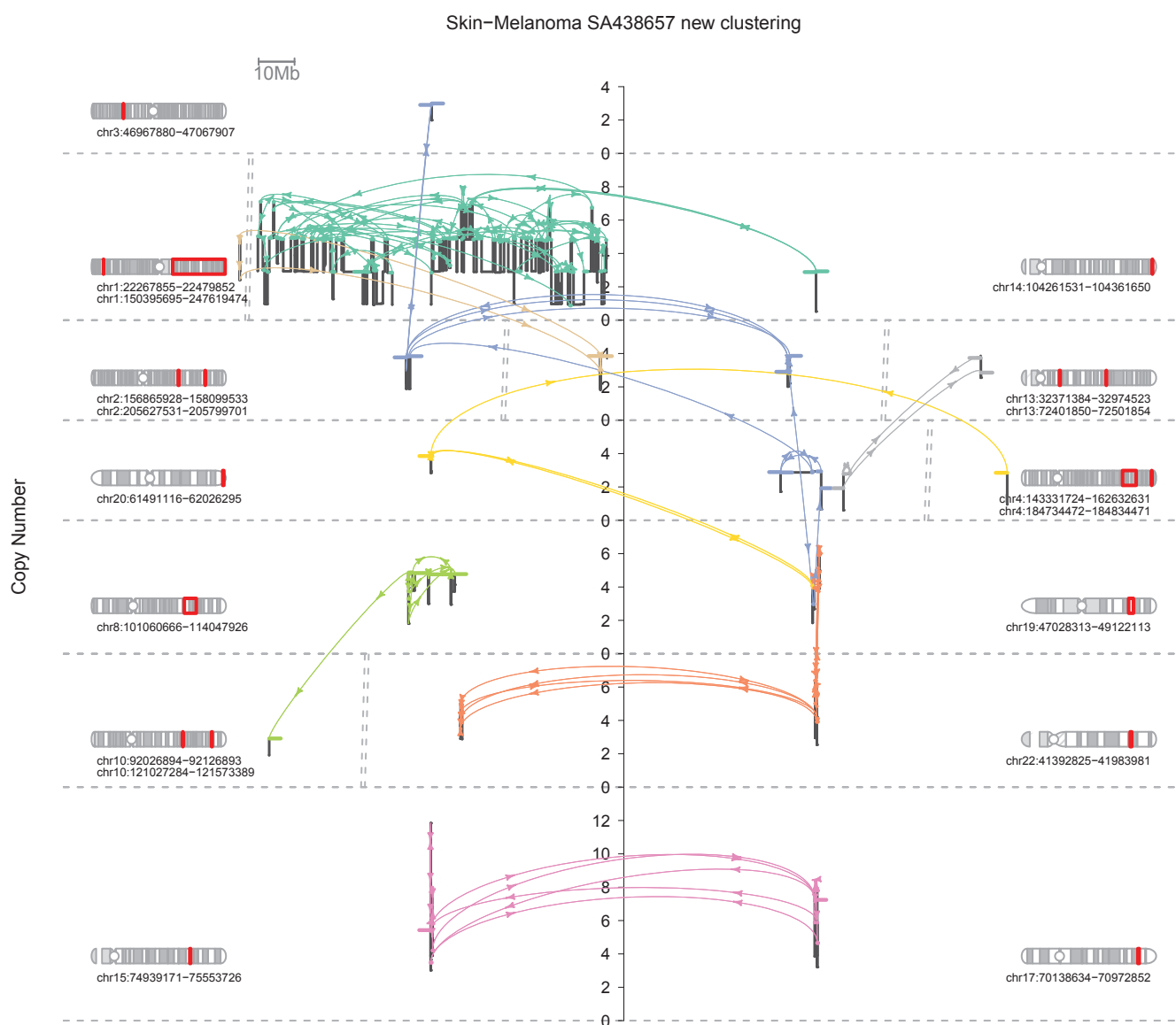


Figure D.22: New and old clustering schemes (more clusters in new) for complex unexplained BPJ in melanoma sample SA438657, only showing clusters with disagreement. Figure continues on the next page.

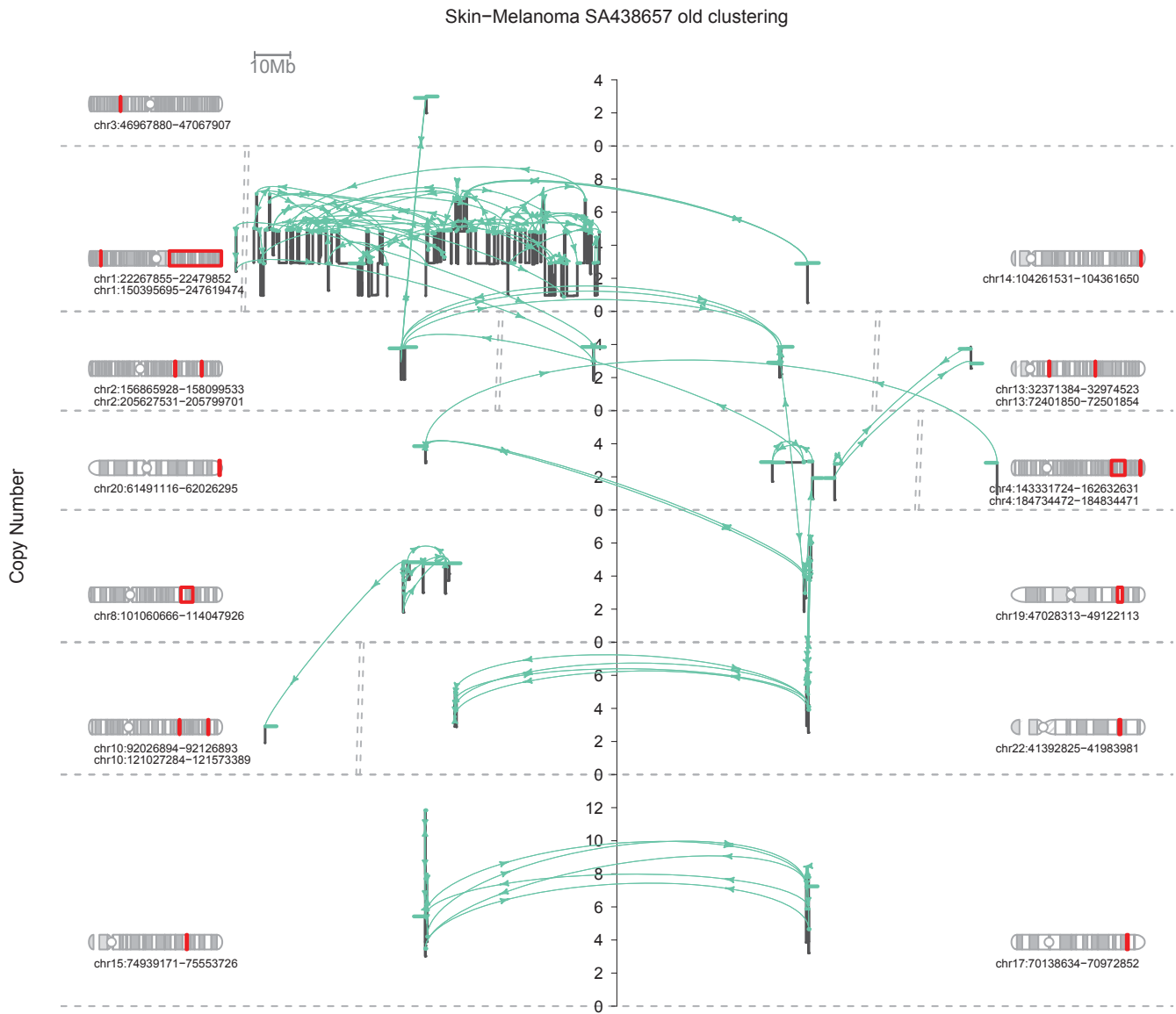


Figure D.22: New and old clustering schemes (more clusters in new) for complex unexplained BPJ in melanoma sample SA438657, only showing clusters with disagreement. Figure is continued from the previous page.

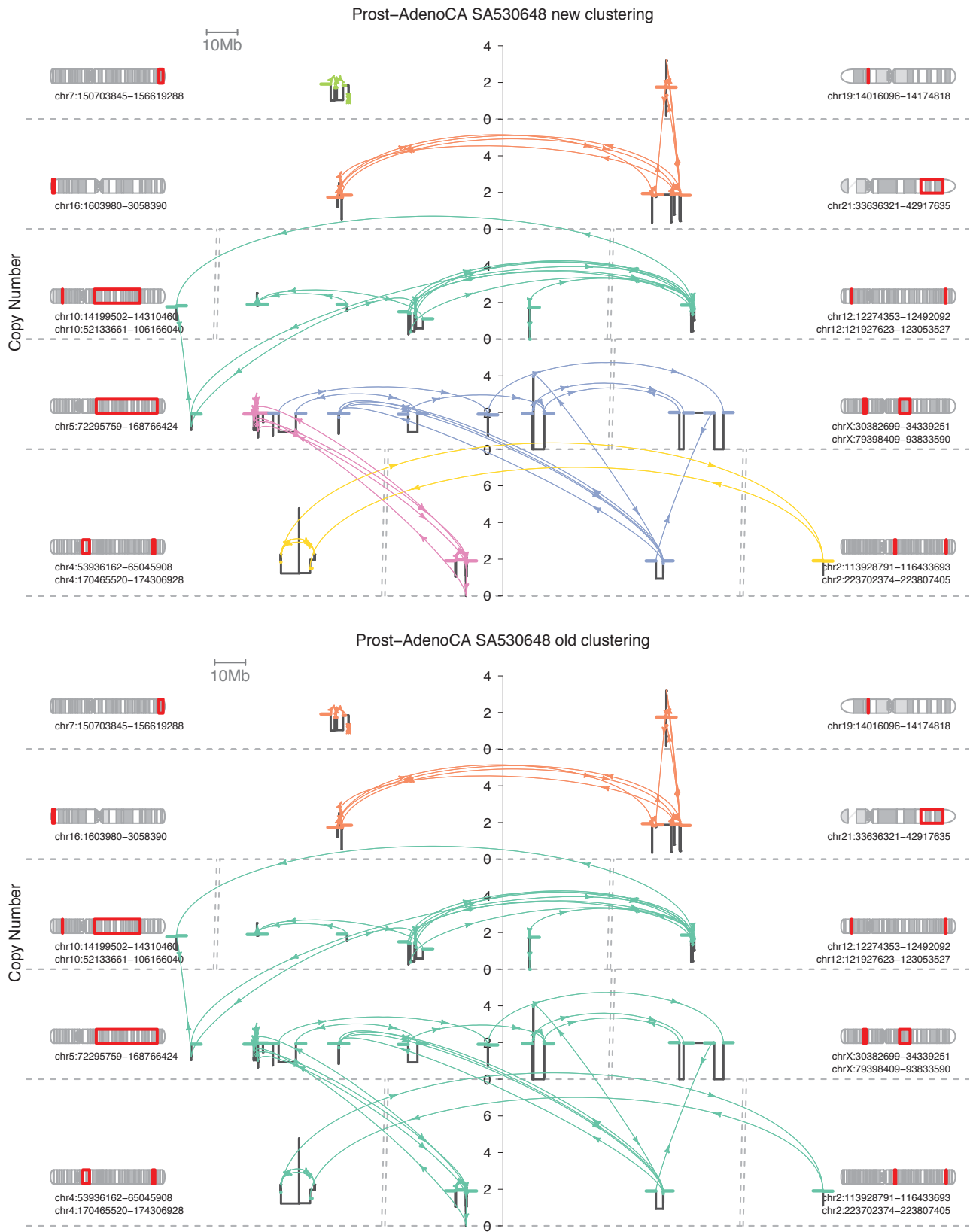


Figure D.23: New and old clustering schemes (more clusters in new) for complex unexplained BPJ in prostate cancer sample SA530648, only showing clusters with disagreement.

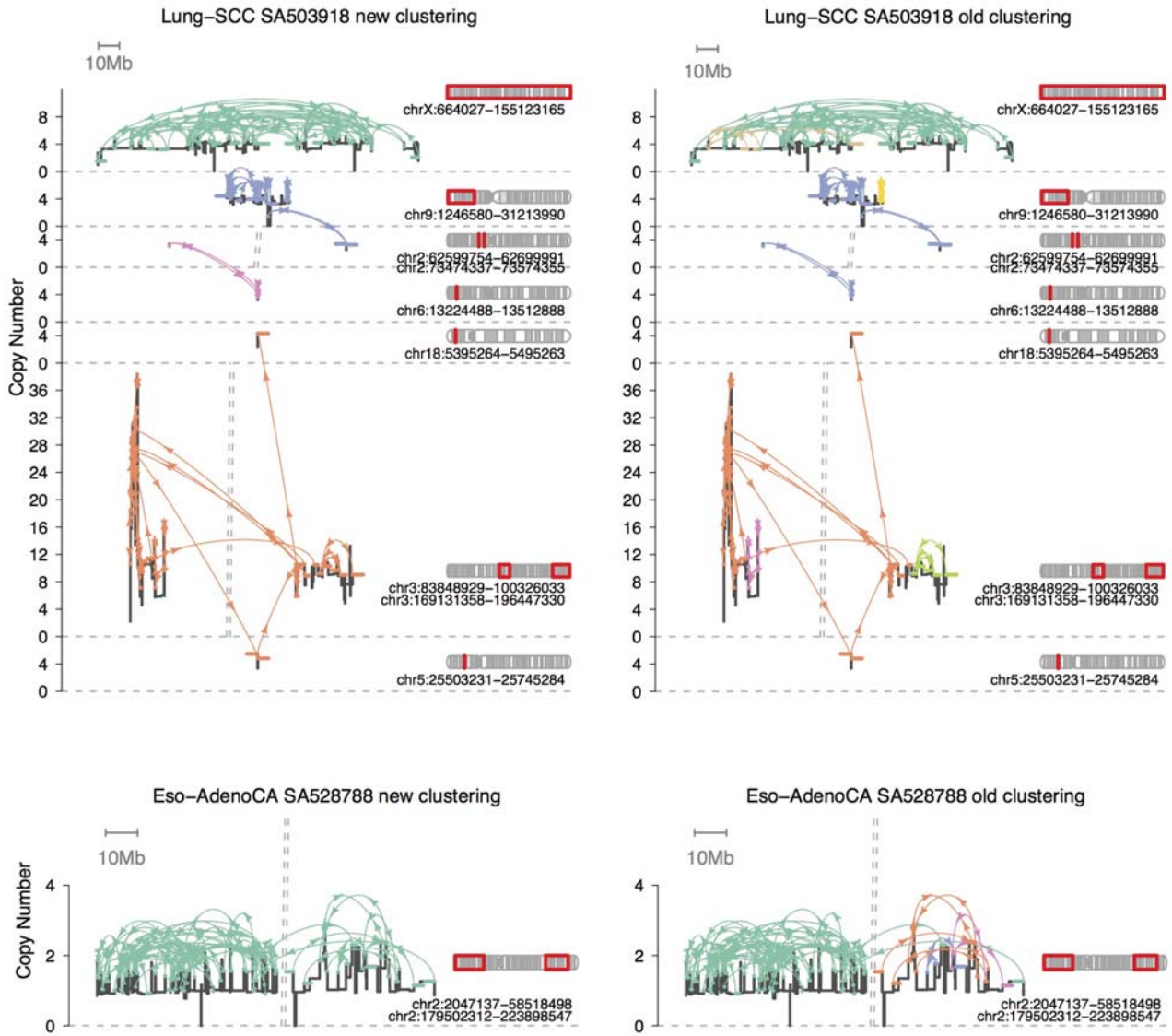


Figure D.24: New and old clustering schemes (fewer clusters in new) for complex unexplained BPJ in lung squamous cell cancer sample SA503918 and esophageal cancer sample SA528788, only showing clusters with disagreement.

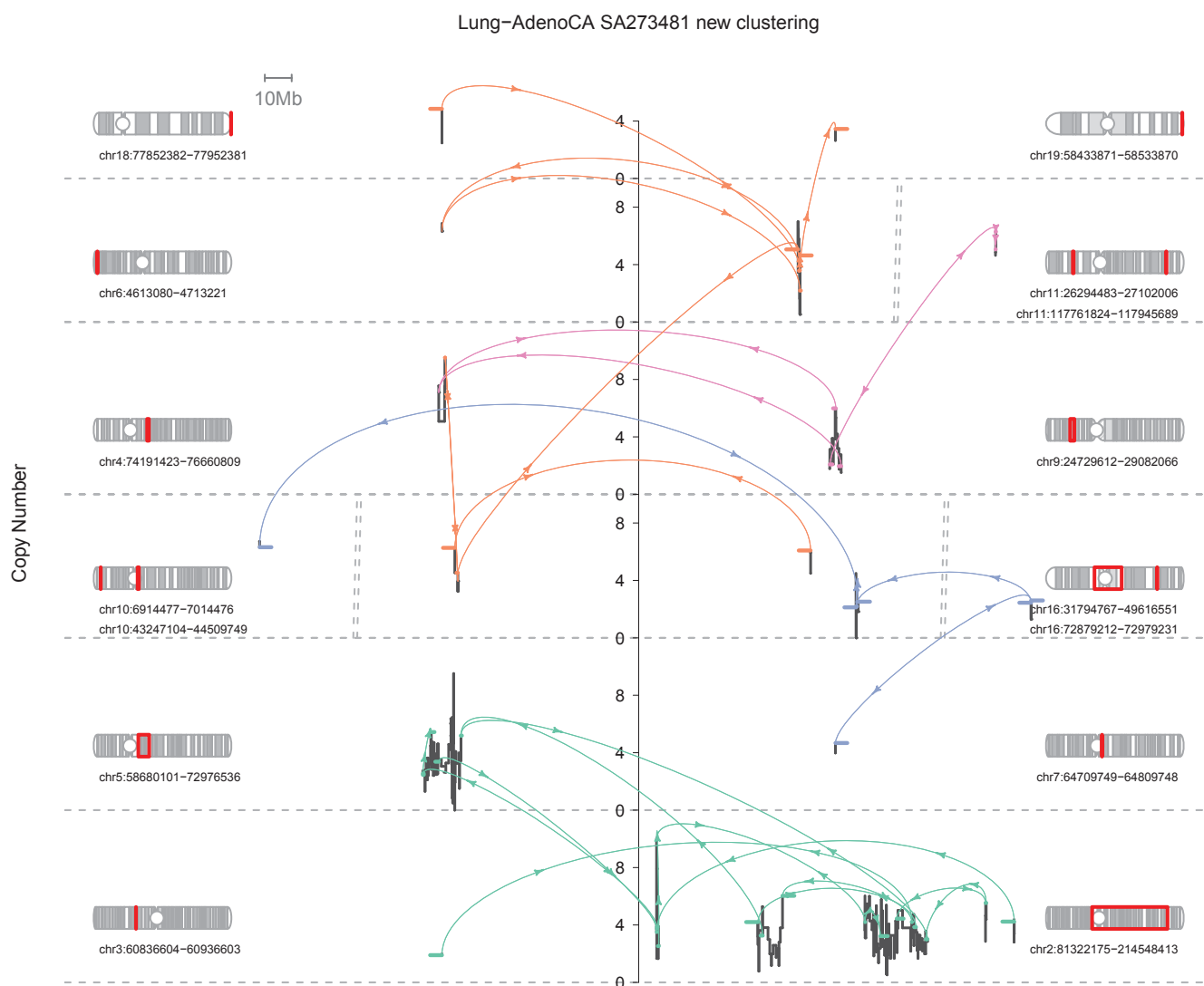


Figure D.25: New and old clustering schemes (fewer clusters in new) for complex unexplained BPJ in lung cancer sample SA273481, only showing clusters with disagreement. Figure continues on the next page.

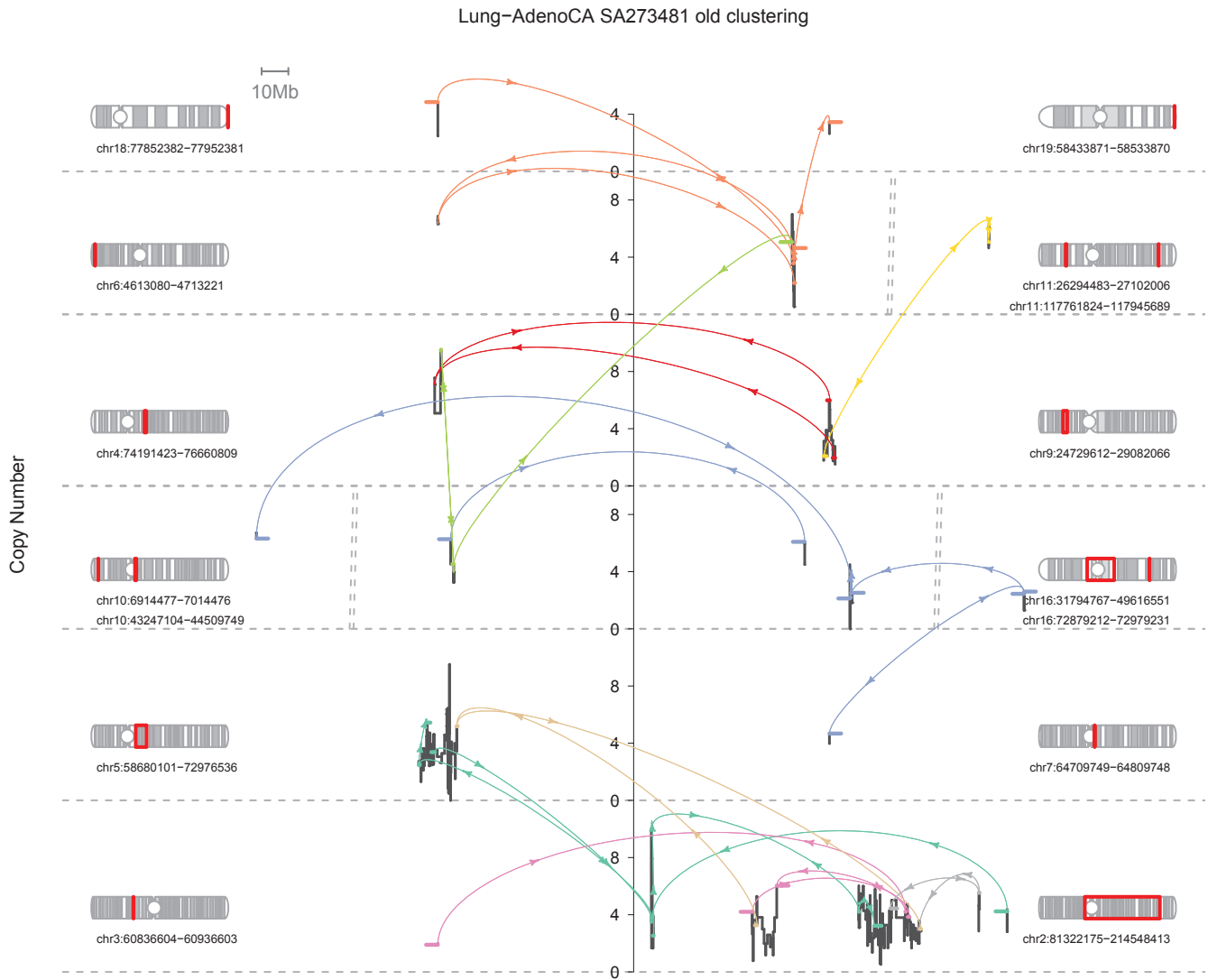


Figure D.25: New and old clustering schemes (fewer clusters in new) for complex unexplained BPJ in lung cancer sample SA273481, only showing clusters with disagreement. Figure is continued from the previous page.

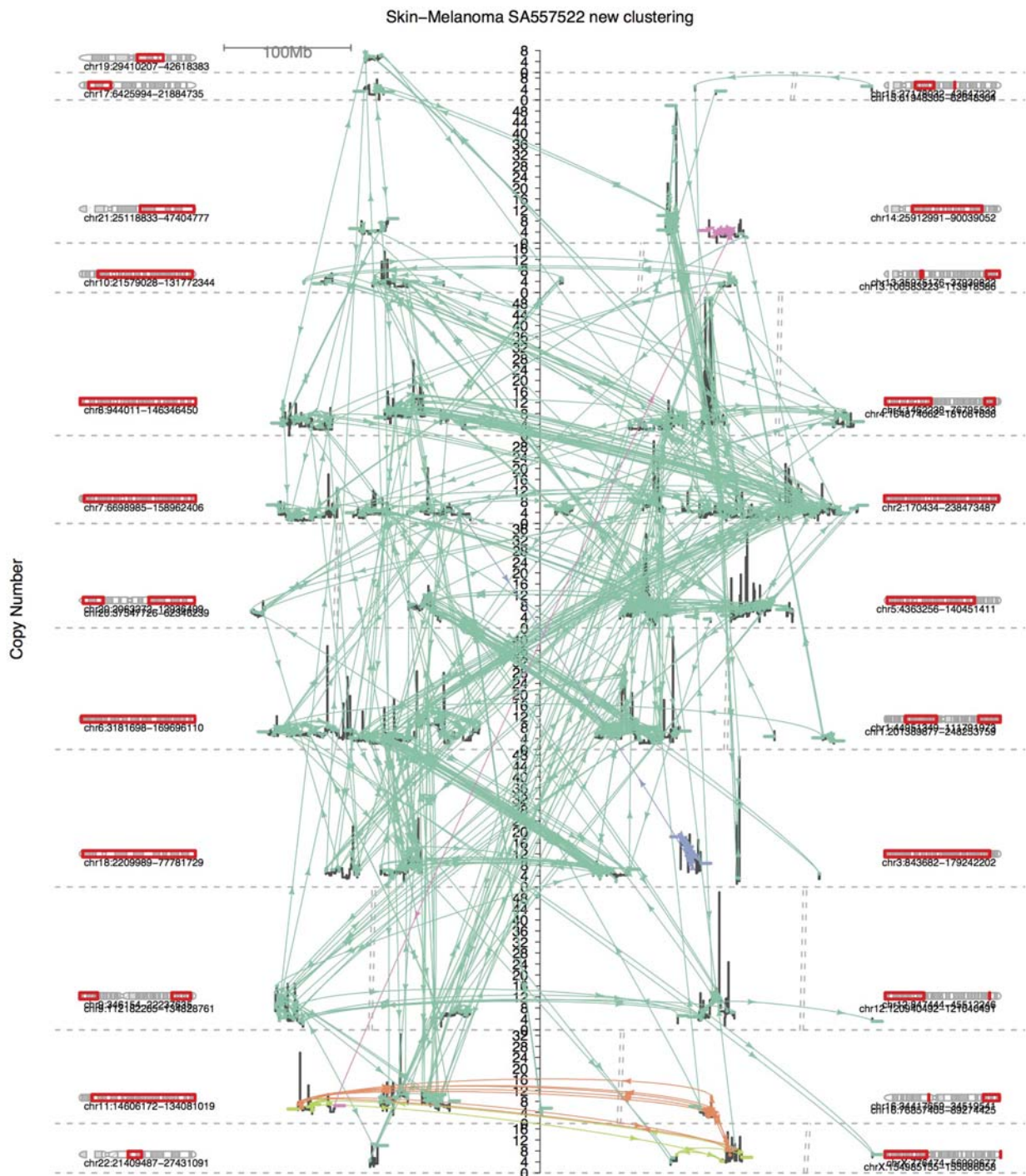


Figure D.26: New and old clustering schemes (fewer clusters in new) for complex unexplained BPJ in melanoma sample SA557522, only showing clusters with disagreement. Figure continues on the next page.

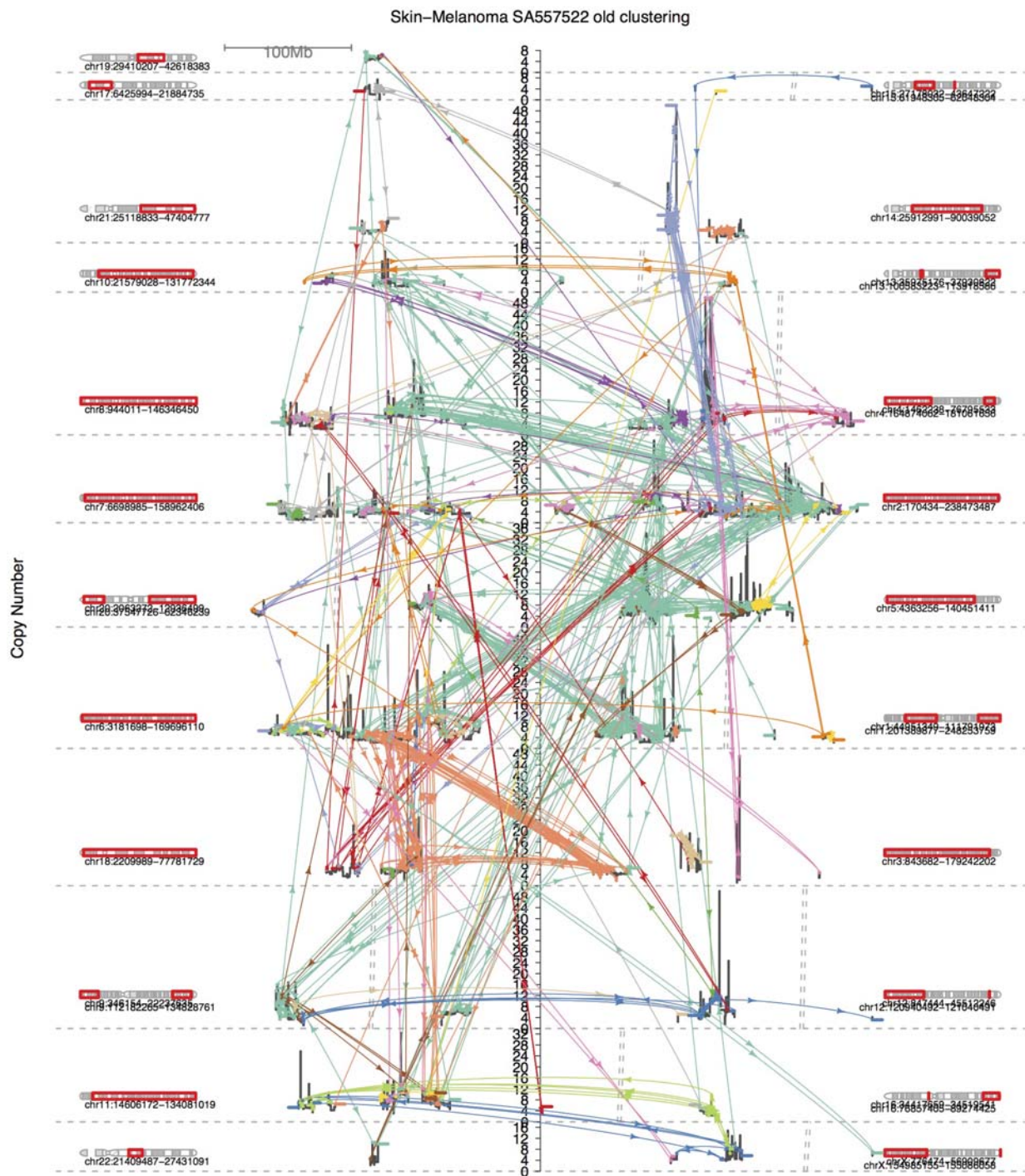


Figure D.26: New and old clustering schemes (fewer clusters in new) for complex unexplained BPJ in melanoma sample SA557522, only showing clusters with disagreement. Figure is continued from the previous page.

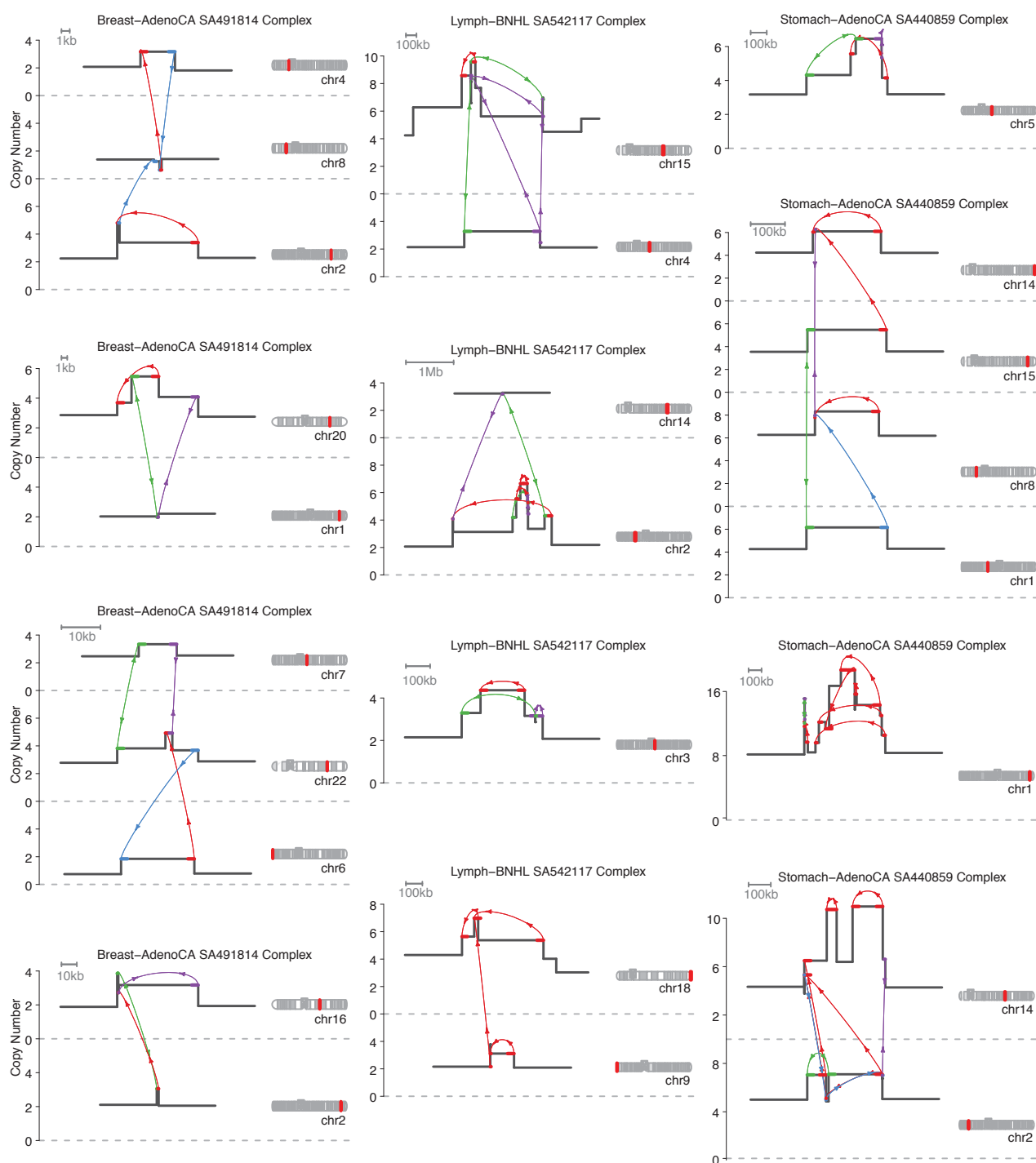


Figure D.27: Four example events from each of three outlying samples containing more than 30 separate complex clusters. These samples are further summarised in Figure 5.16.

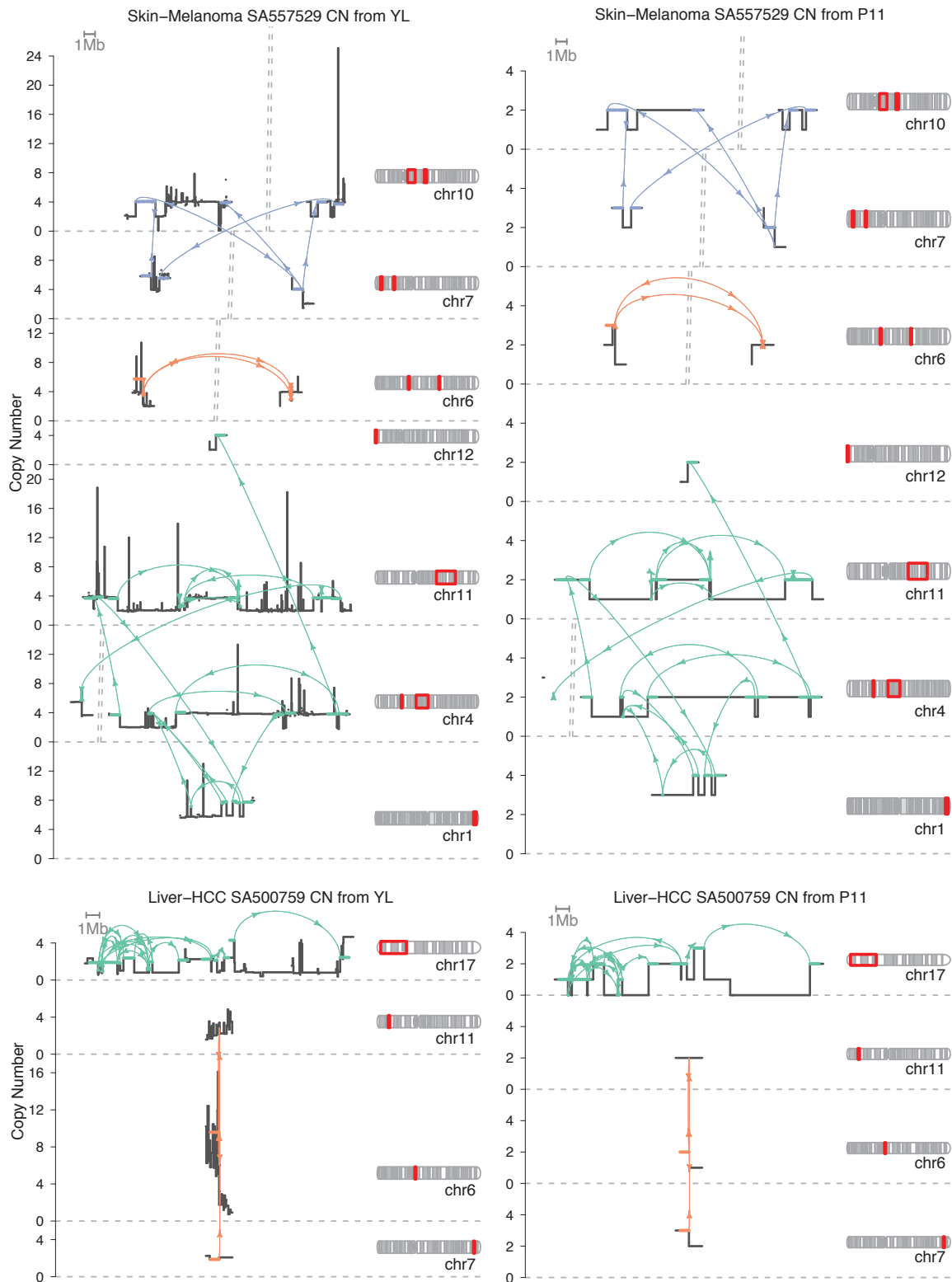


Figure D.28: Comparison of CN calls returned by YL (left) and P11 (right) around complex unexplained BPJ in samples qualifying for a switch to P11 CN. Breakpoint junctions are coloured by cluster assignment within the sample. Figure continues on the next page.

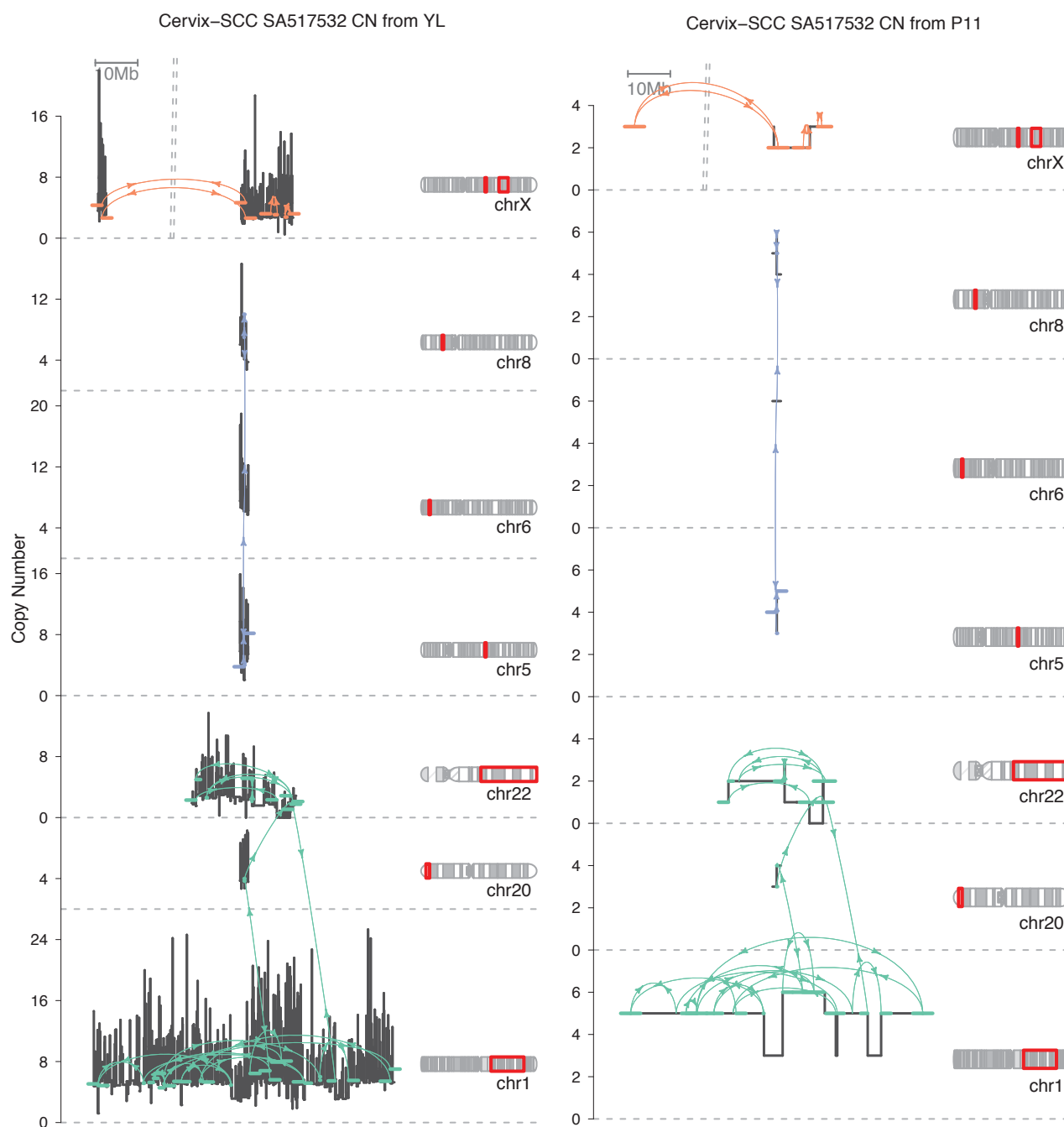


Figure D.28: Comparison of CN calls returned by YL (left) and P11 (right) around complex unexplained BPJ in samples qualifying for a switch to P11 CN. Breakpoint junctions are coloured by cluster assignment within the sample. Figure continued from the previous page.

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

List of Tables

2.1	Sample counts by histology group in PCAWG	21
2.2	Classification of simple structural variants in PCAWG cohort . .	26
2.3	Sample variables associated with SV burden	41
3.1	Histone mark interpretations	74
3.2	Cancer census genes ranked by SV classes	107
4.1	Parameters for simulated SNV catalogues	131
5.1	Comparison of old and new complex BPJ clustering	181
5.2	Tiny unexplained BPJ clusters	183
5.3	Summary of first tier complex classes (pilot)	203
E.1	ROADMAP cell lines matched to PCAWG tissue types	278
E.2	Fragile site definitions for the PCAWG cohort	279
E.3	Sample counts of Alexandrov somatic SNV dataset	280

List of Figures

2.1	BPJ orientations and motifs	23
2.2	Overlap between four SV calling algorithms	24
2.3	Example plots for simple SV classes	29
2.4	Example plots for local 2-jump SV	30
2.5	Example plots for local + distant 2-jump SV	32
2.6	Example plots for templated insertion chain	33
2.7	Example plots for templated insertion bridge and cycle	34
2.8	Example plots for chromoplexy	35
2.9	SV class distributions across cancer histology groups	37
2.10	SV class distribution across samples in four histology groups	39
2.11	BPJ count in templated insertion and chromoplexy	42
2.12	Longest templated insertion bridge event	43
2.13	Longest templated insertion cycle events	44
2.14	Deletion and tandem dup size distributions	46
2.15	Samples clustered by deletion size distribution	48
2.16	Samples clustered by tandem dup size distribution	49
2.17	Size distribution of local 2-jumps	50
2.18	Templated insertion size distribution	52
2.19	Gap size distribution for recip trans and chromoplexy	53
2.20	Microhomology enrichment by histology and SV class	56
2.21	Example kataegis regions with SV	59
2.22	Kataegis signatures and tally	60
2.23	Kataegis association with SV	61
3.1	Genome-wide distribution of rearrangements	68
3.2	Cumulative length of callable genome regions	69
3.3	Spearman correlation between genome properties	76
3.4	SV class associations with 13 genome properties	79
3.5	Replication timing for three sub-groups of local 2-jump	80
3.6	Proportion of SV breaks near short repeats	81
3.7	SV vs replication timing in hypermutators	84
3.8	Replication timing across interchromosomal BPJ	86
3.9	Lasso paths for GLM logistic regression	89
3.10	Coefficients for GLM logistic regression	90
3.11	GAM fit for short tandem dups	92
3.12	ROC curves for GLM and GAM logistic regression	93
3.13	Predicted rearrangement rate on chr16 and chr17	95

3.14	Sv breakpoints within nine major fragile sites	98
3.15	Fragile site ranking and size distribution	99
3.16	Del density and replication timing around 12 FS	101
3.17	Fragile site preference by histology group	103
3.18	Complex sv in fragile sites	105
3.19	Sv breakpoints within three immune loci	106
3.20	Sv breakpoints around eight example cancer genes	109
3.21	Sv breakpoints around six genes with recurrent fusion drivers .	111
3.22	Sv events generating the <i>TPR2-ERG</i> fusion driver	112
3.23	Sv events around <i>MYC</i>	114
3.24	Sv events around <i>TERT</i>	116
3.25	Sv events around <i>RB1</i>	117
4.1	Overview of HDP model for multiple sample groups	125
4.2	Overview of HDP model conditioning on prior knowledge	130
4.3	HDP performance by number of MCMC chains	134
4.4	HDP performance by number of initial clusters	135
4.5	HDP performance by shape prior for concentration parameter .	136
4.6	HDP performance by sample size	138
4.7	HDP performance by sample exposure similarity	139
4.8	HDP performance by sub-group modelling	140
4.9	Factors influencing sample exposure reconstruction	141
4.10	Factors influencing mutational signature reconstruction	142
4.11	Computational resources for HDP on simulated data	143
4.12	HDP diagnostic plots for signature discovery dataset	146
4.13	HDP signature exposures in discovery samples	148
4.13	HDP exposures in discovery samples (cont.)	149
4.14	t-SNE view of HDP-extracted signatures	151
4.15	Sample exposures to HDP signatures for six cancer types	152
4.16	New mutational signatures in pancreas	154
4.17	Pancreatic endocrine cancer mutational signature exposures . .	155
4.18	New mutational signatures in prostate	156
4.19	Prostate cancer mutational signature exposures	157
4.20	Sv signatures	163
4.20	Sv signatures (continued)	164
4.21	Sv signature exposures, incl breast and prostate	166
5.1	Gamma mixture fit to inter-break distances in 64 samples	173
5.2	Node-edge graphs of complex BPJ in simple samples	176
5.3	Large BPJ clusters with no separable sub-graphs	177
5.4	Fully separable sub-graphs in large BPJ clusters	178
5.5	Partially separable sub-graphs in large BPJ clusters	179
5.6	Separable sub-graphs after extra agglomeration	180
5.7	BPJ clustering discrepancies	181
5.8	Comparison of two CN callers	185
5.9	CN comparison (YL vs P11) around sv in two samples	186
5.10	Complex sv overview	188
5.11	Unusual c-thripsis with over 1000 BPJ on two chrom	189

5.12	Somatic retrotransposition clusters	190
5.13	BPJ clusters spanning 17 or more chromosomes	191
5.14	Complex cluster spanning 19 chromosomes w/ 155 BPJ	192
5.15	Complex cluster spanning 17 chromosomes w/ 1122 BPJ	193
5.16	Three samples with many complex clusters	194
5.17	Small break and ligate clusters on one chrom	196
5.18	Small break and ligate clusters on two chrom	197
5.19	Small complex templated insertions	198
5.20	Small template and replicate clusters with shared break	199
5.21	Combination SV	200
5.22	Overlap SV clusters	201
5.23	Overlap between complex class annotations (pilot)	203
5.24	Histology distribution for complex SV	204
5.25	Example double minute events	205
5.26	Example BFB events	206
5.27	Example complex amplification events	207
5.28	Example complex chromoplexy events	209
5.29	Example chromothripsis events	210
6.1	Annotation map theory	219
B.1	HDP for one group	227
B.2	HDP for two groups	228
D.1	Example SV configurations on chr17 <i>p</i> -arm	238
D.2	SV class distribution across samples in 27 histology groups	239
D.3	Correlation between complex and classified BPJ counts	240
D.4	Longest chromoplexy events	241
D.5	Longest chromoplexy with insertion events	242
D.6	Liver cancer with large tandem dup	242
D.7	Microhomology variation across samples	243
D.8	SV class associations with 25 genome properties	244
D.9	GAM fit for short deletions	245
D.10	GAM fit for large deletions	246
D.11	GAM fit for large tandem dups	247
D.12	GAM fit for unbalanced translocations	248
D.13	GAM fit for foldbacks	249
D.14	SV breakpoints within 12 minor fragile sites	250
D.15	Del density and replication timing around nine minor FS	251
D.16	Trinucleotide frequency in exome vs genome	252
D.17	HDP signatures in discovery dataset	253
D.17	HDP signatures in discovery dataset (cont.)	254
D.17	HDP signatures in discovery dataset (cont.)	255
D.17	HDP signatures in discovery dataset (cont.)	256
D.17	HDP signatures in discovery dataset (cont.)	257
D.17	HDP signatures in discovery dataset (cont.)	258
D.17	HDP signatures in discovery dataset (cont.)	259
D.18	SV signature exposures, incl esophagus and ovary	260

D.19 Large BPJ cluster splits	261
D.19 Large BPJ cluster splits (continued)	262
D.20 BPJ clustering example 1	263
D.21 BPJ clustering example 2	264
D.21 BPJ clustering example 2	265
D.22 BPJ clustering example 3	266
D.22 BPJ clustering example 3	267
D.23 BPJ clustering example 4	268
D.24 BPJ clustering example 5	269
D.25 BPJ clustering example 6	270
D.25 BPJ clustering example 6	271
D.26 BPJ clustering example 7	272
D.26 BPJ clustering example 7	273
D.27 Example SV from samples with many complex clusters	274
D.28 CN comparison (YL vs P11) around SV in same samples	275
D.28 CN comparison (YL vs P11) around SV in same samples (cont.)	276