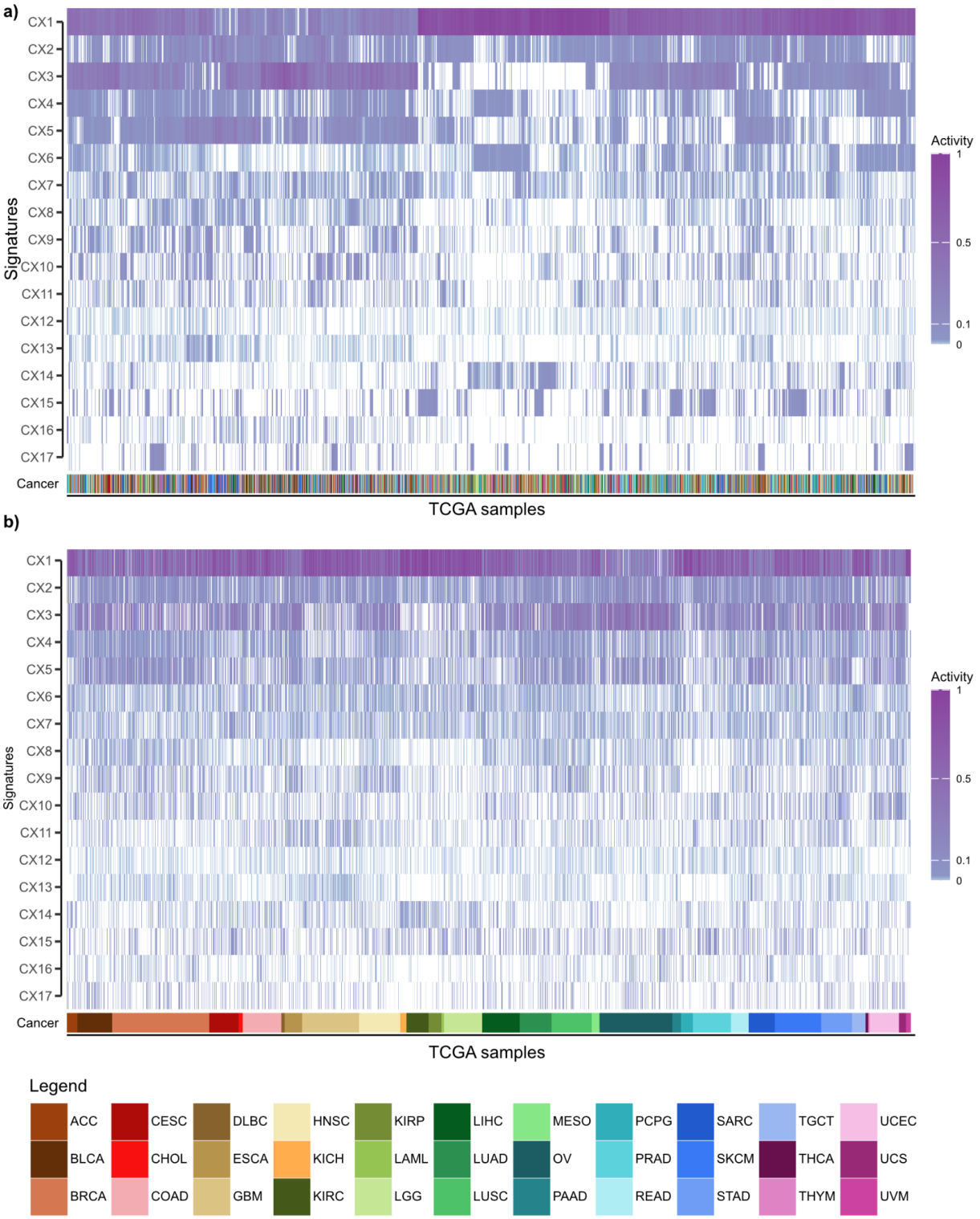
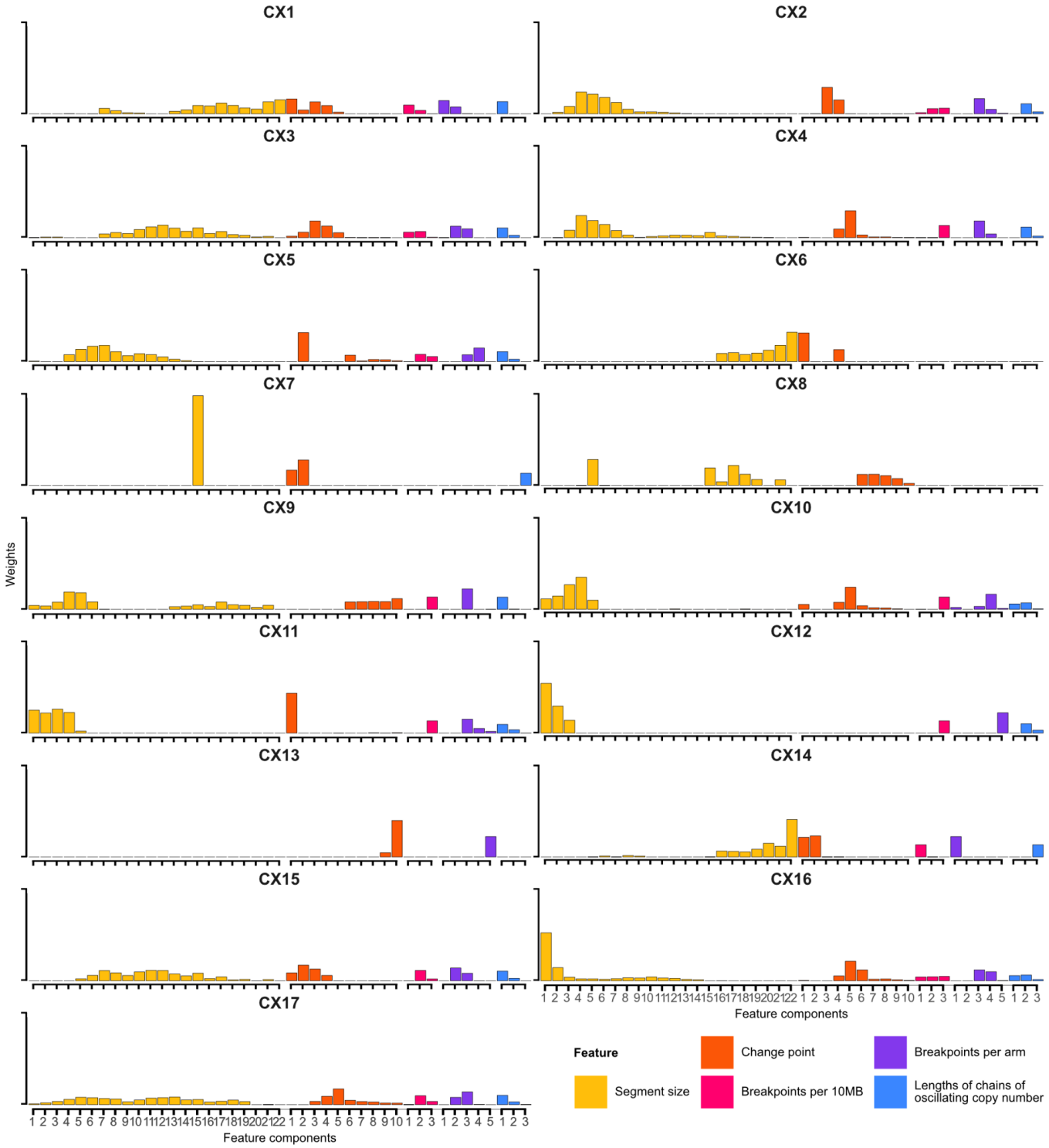
Supplementary figures



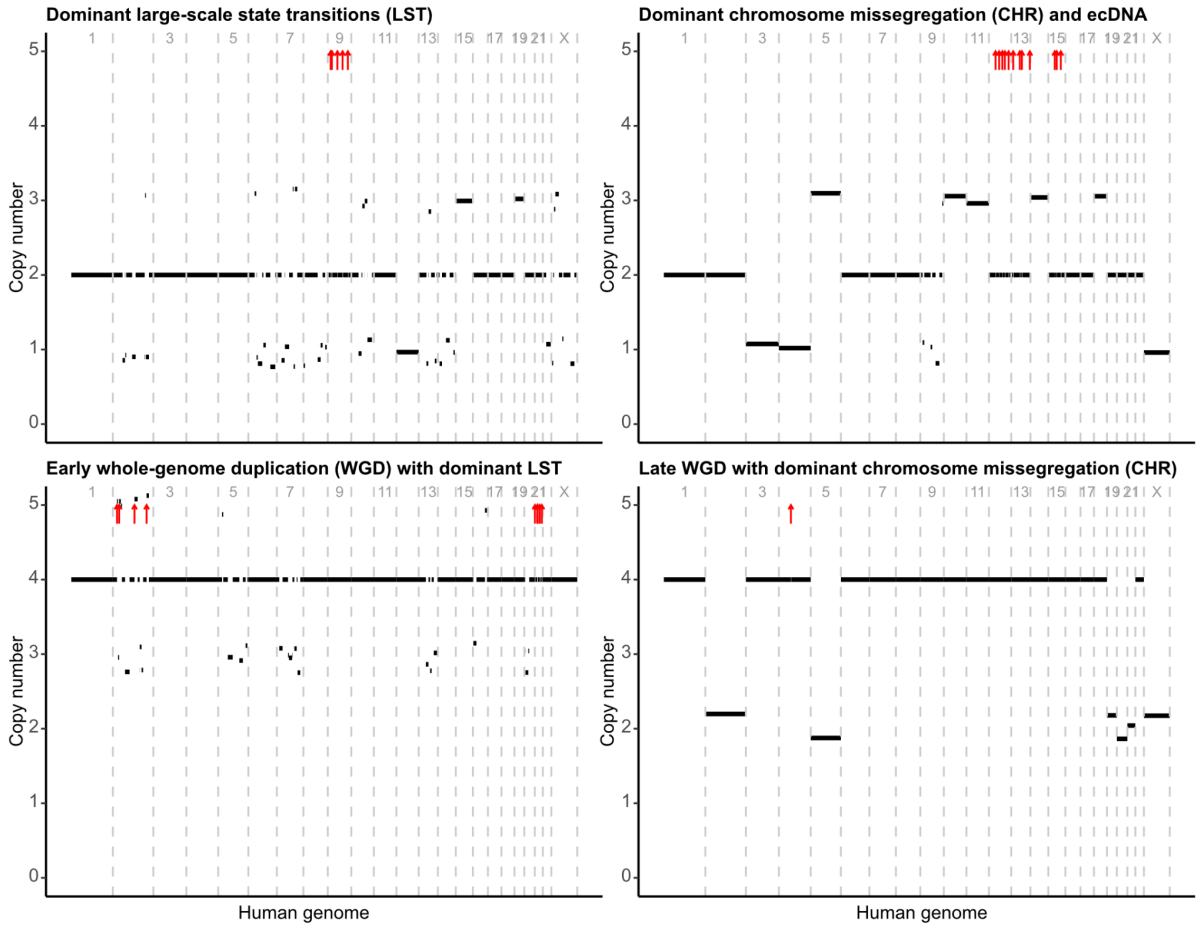
Supplementary Fig. 1: Signature activities of 6,335 TCGA samples

a: Heatmap of normalised signature activities of 6,335 TCGA samples sorted by hierarchical clustering. **b:** Samples of (a) sorted by cancer type and decreasing CX1 activity.



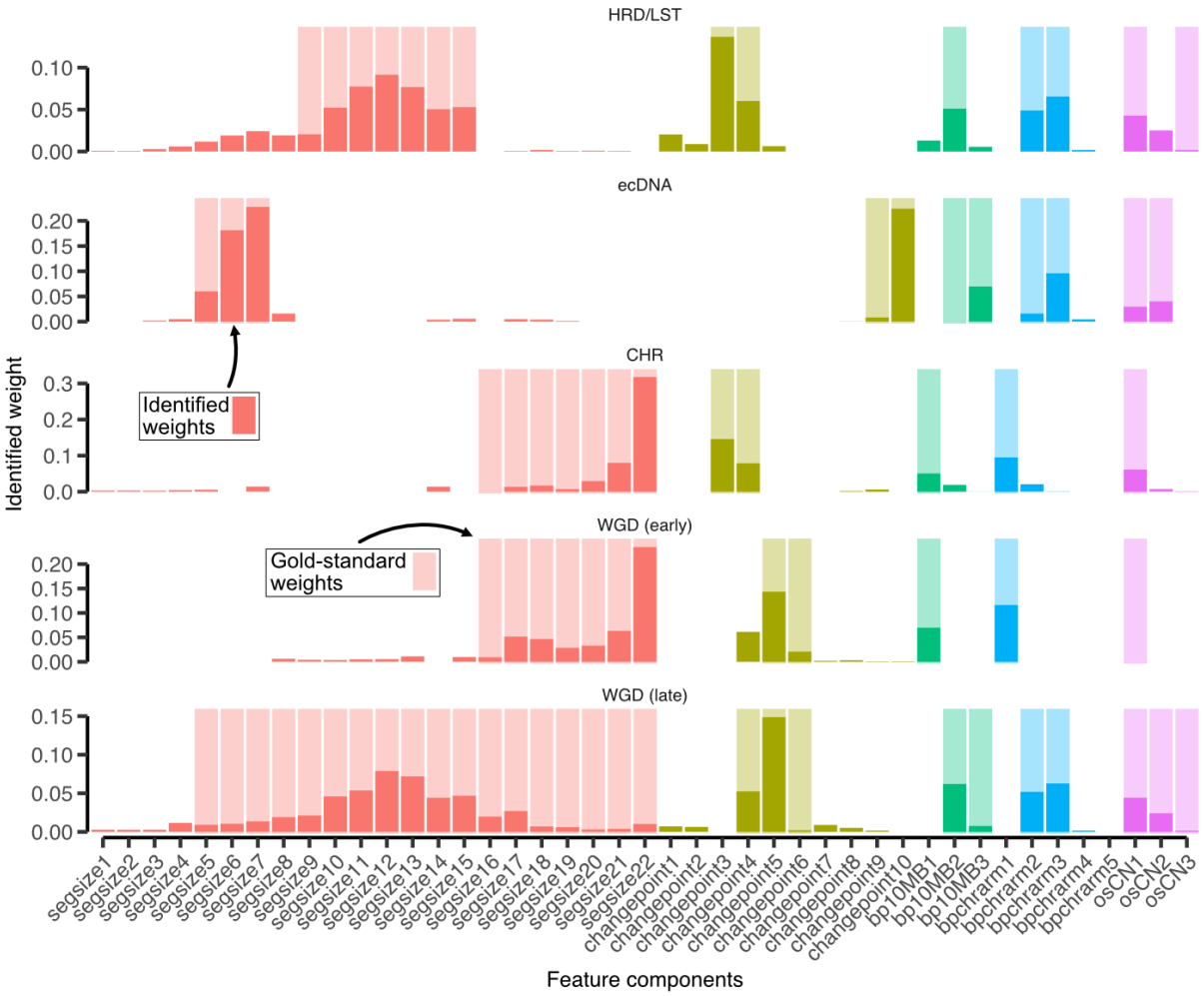
Supplementary Fig. 2: Barplot of signature definitions

Displayed on the x-axis are the 43 components of the five features. Components go from small to large means from left to right (see Extended Data Fig. 4 or Supplementary Material 2, Supplementary Table 2 for component means). The heatmap bar height shows the adjusted weights for each signature. Weights were normalised per signature and adjusted for each feature by the number of components.



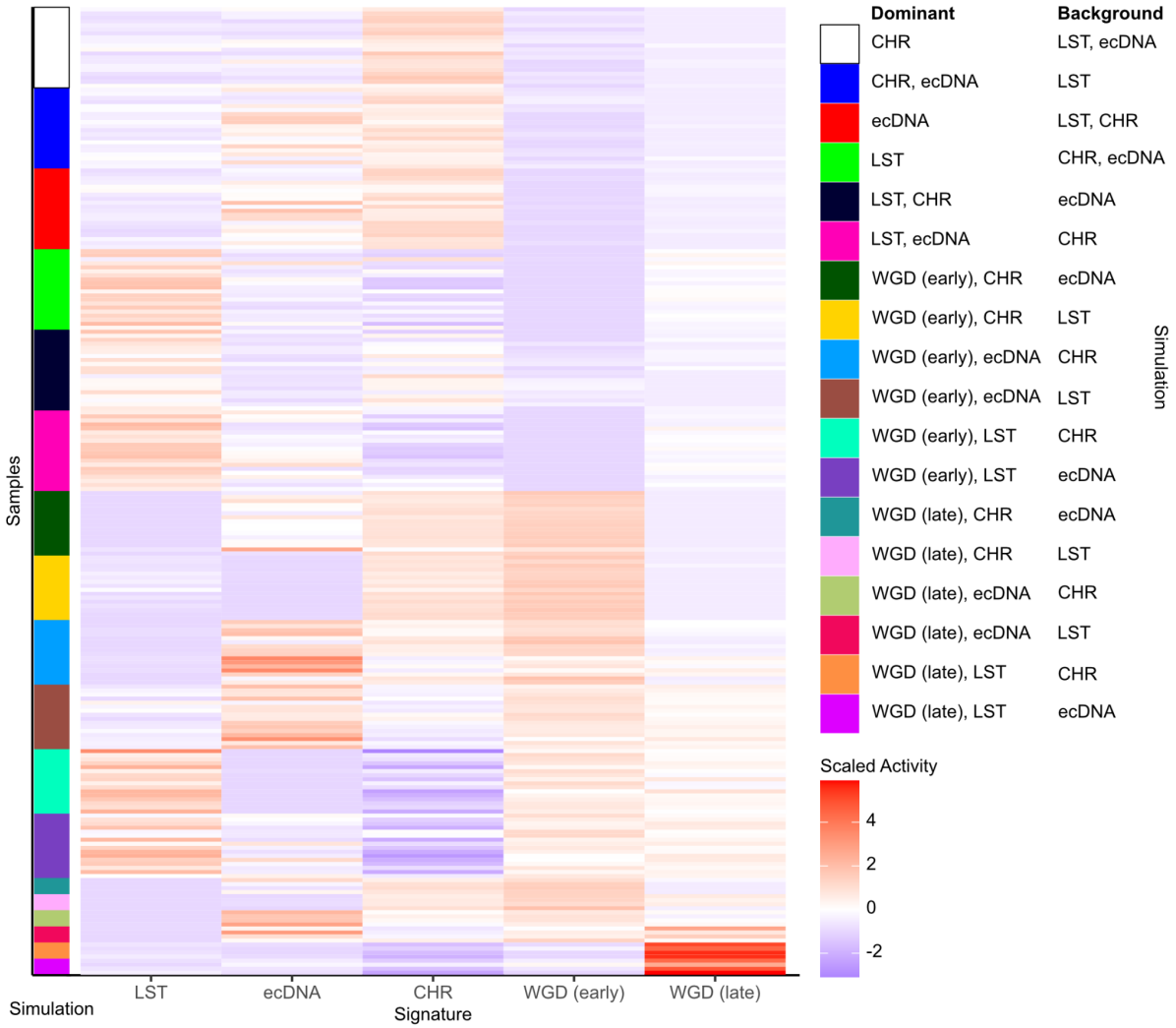
Supplementary Fig. 3: Example copy number profiles of the simulation producing genomes exposed to mutational processes

The titles of each plot state the dominant mutational processes that acted on this sample. The x-axis denotes the length of the human genome, with vertical, grey dashed lines showing chromosome boundaries. Y-axis shows the absolute copy number. Red arrows denote copy number aberrations with a copy number higher than 5.

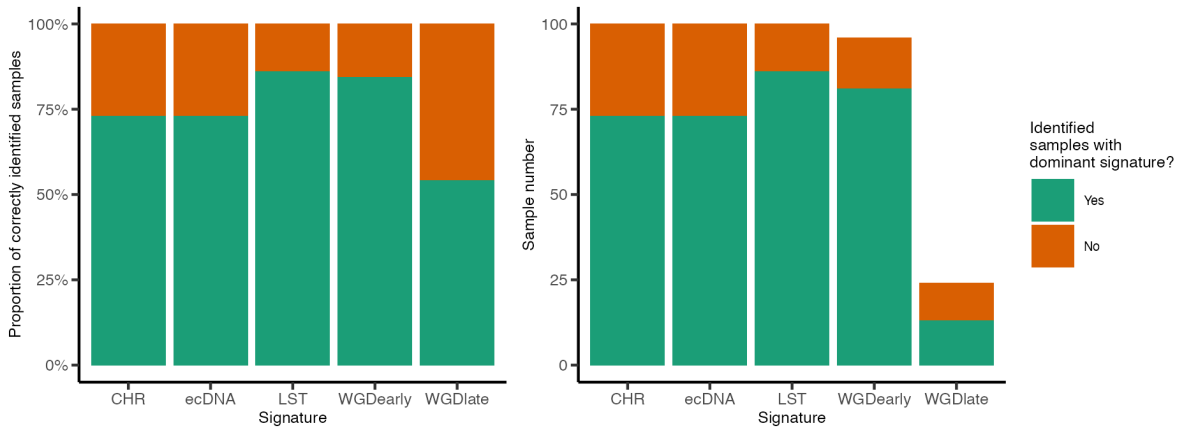


Supplementary Fig. 4: Gold-standard and identified copy numbers signatures of five mutational processes modelled for the simulation of genomes

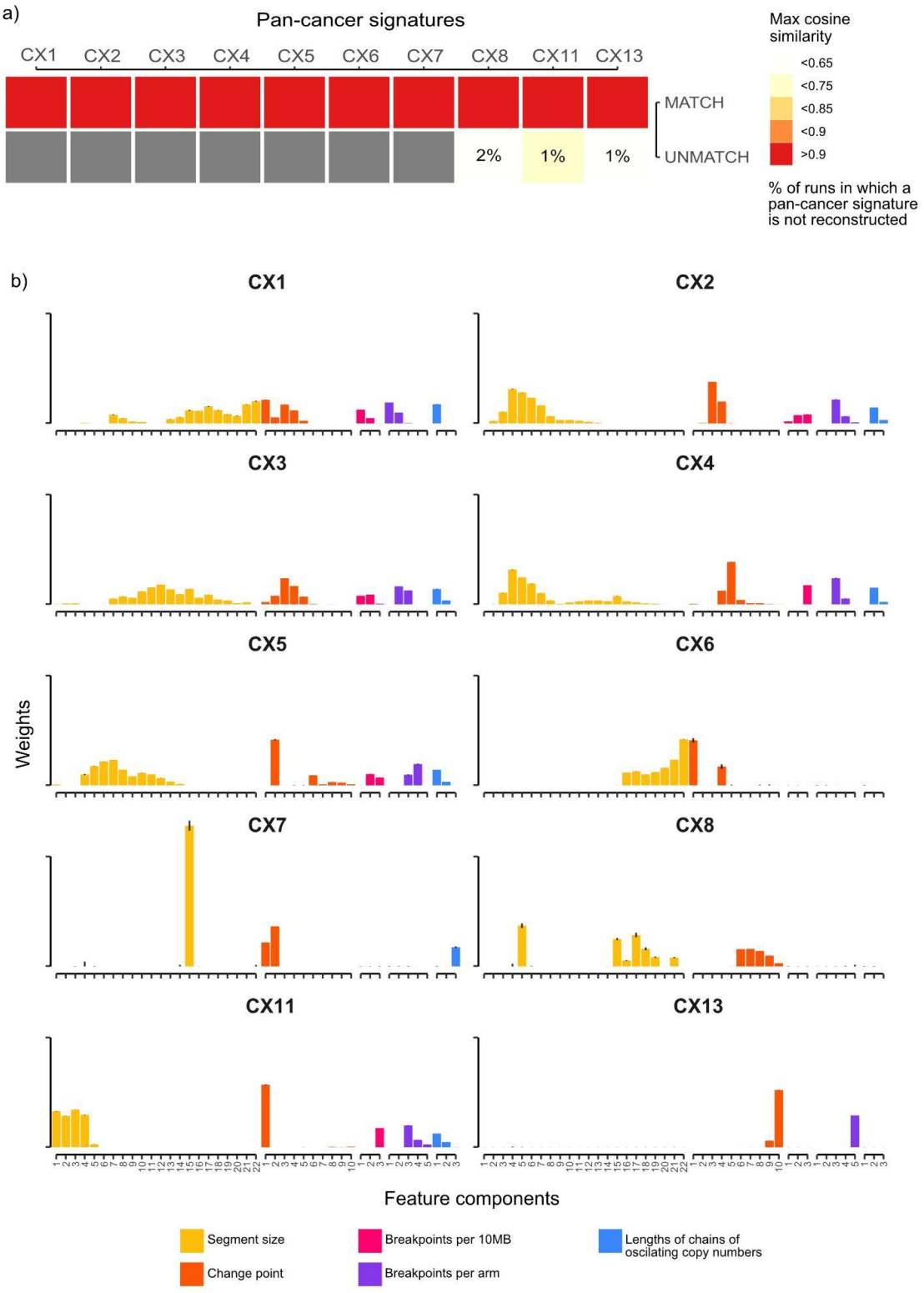
Shaded bars: Components received weight depending on the importance of this component to reflect the mutational process. Fully coloured bars: Identified signatures weights.



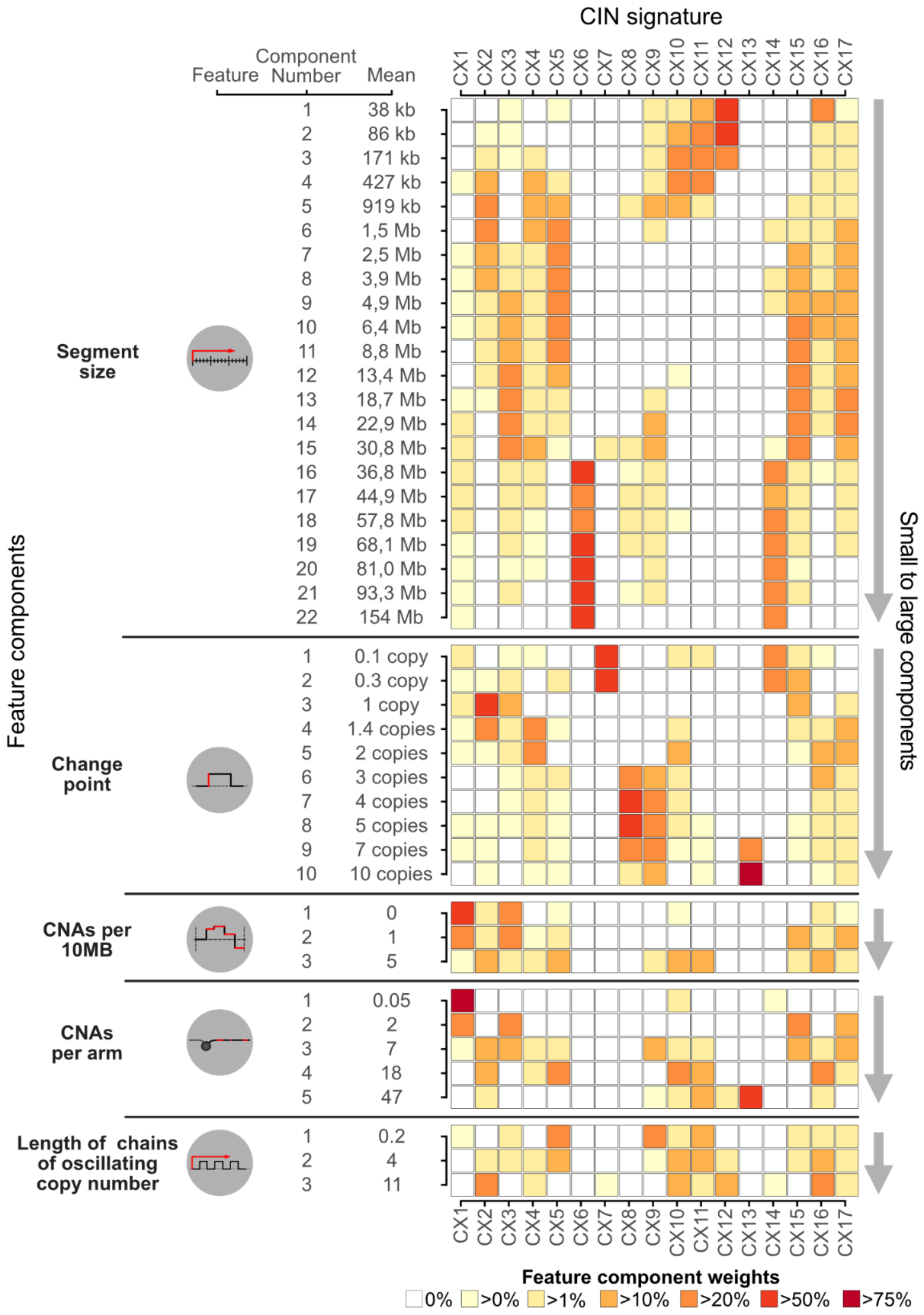
Supplementary Fig. 5: Heatmap of scaled signature activities from simulated samples
Heatmap of 240 samples and their scaled signature activities to the 5 identified copy number signatures from the simulation of genomes. The coloured bar on the left shows the partition of the simulated samples.



Supplementary Fig. 6: Barplot of the accuracy of signature activities identifying dominant samples as produced by simulation parameters
Left: Proportion of correctly identified samples to the five mutational processes. **Right:** Absolute number of samples that have been correctly identified to have a dominant mutational process.

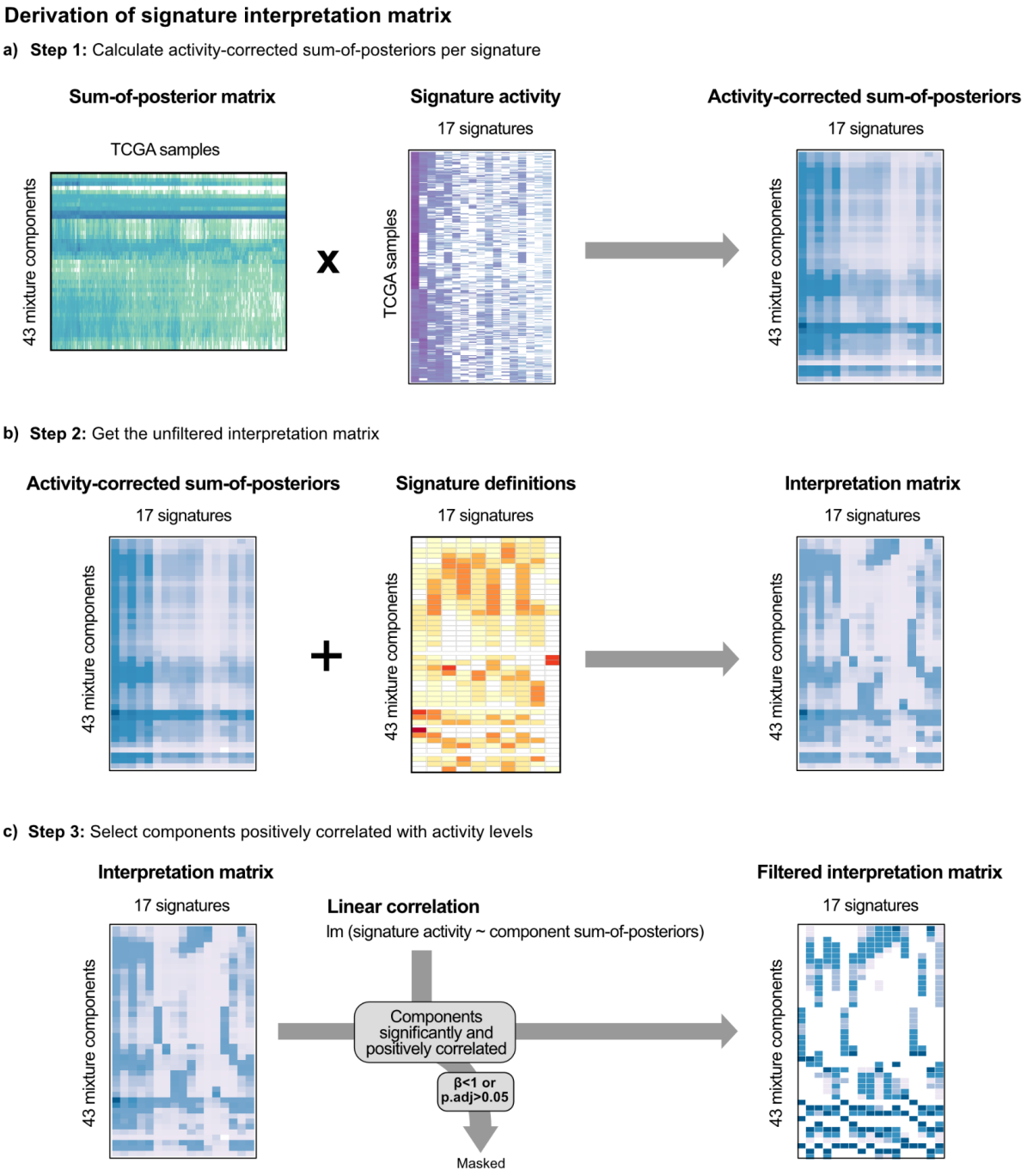


Supplementary Fig. 7: Monte Carlo simulation for determining variability in signature definitions
a: Heatmap of signature definition robustness. Grey denotes that all 100 noisy matrices matched with cosine similarity > 0.85). Three signatures had mismatches of up to 1-2% of cases. **b:** Barplots showing the variability of signature definitions with added noise. Displayed on the x-axis are the 43 components of the five features. The bar height shows the adjusted weights for each signature. Weights were normalised per signature. Error bars represent the standard error of weights for signatures estimated from 100 noisy simulations.



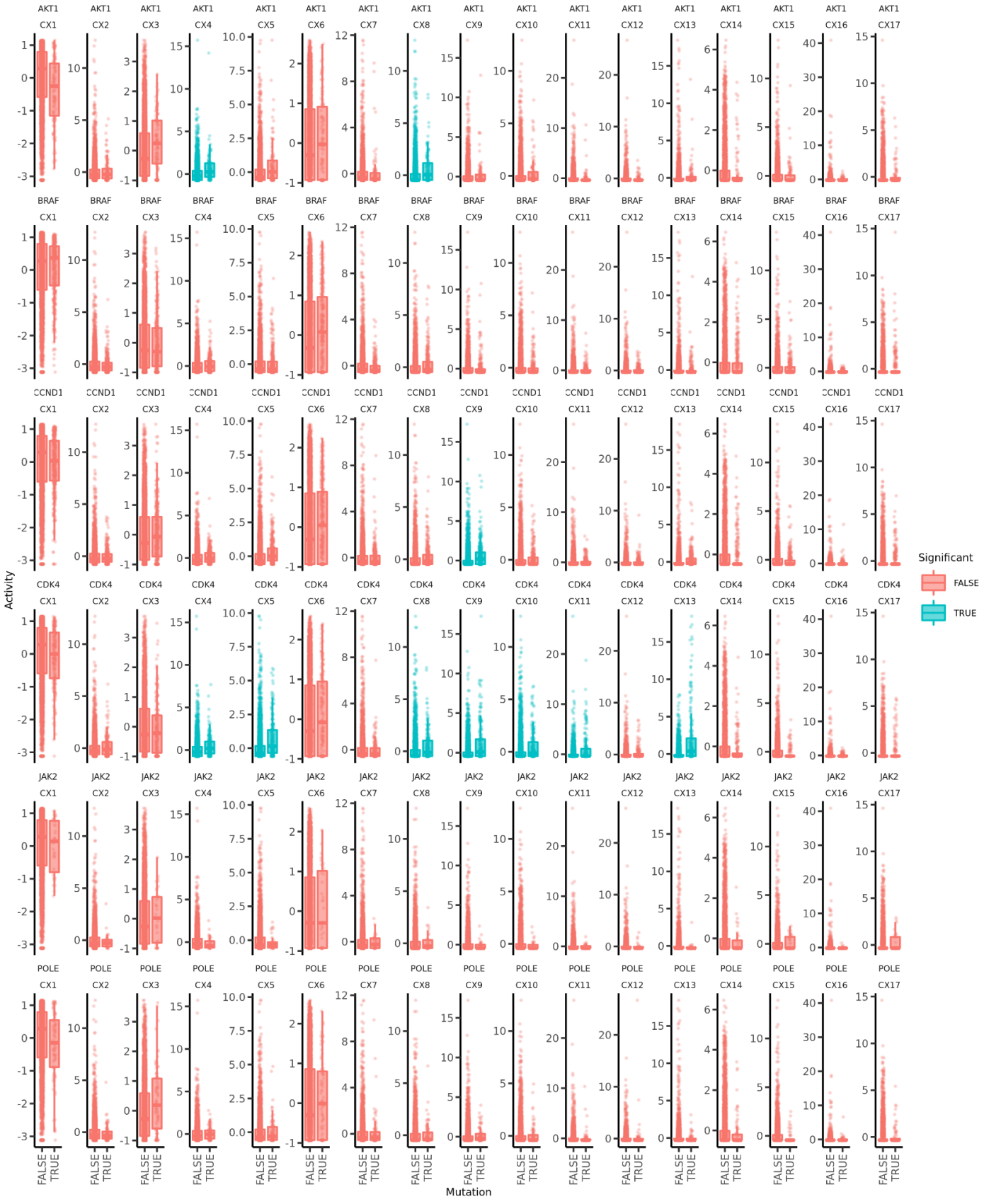
Supplementary Fig. 8: Signature definitions

Displayed on the left are the five features, their mixture components and component means. The heatmap on the right shows weights for each signature. Weights are normalised per row (component) for ease in comparison across signatures for each component.

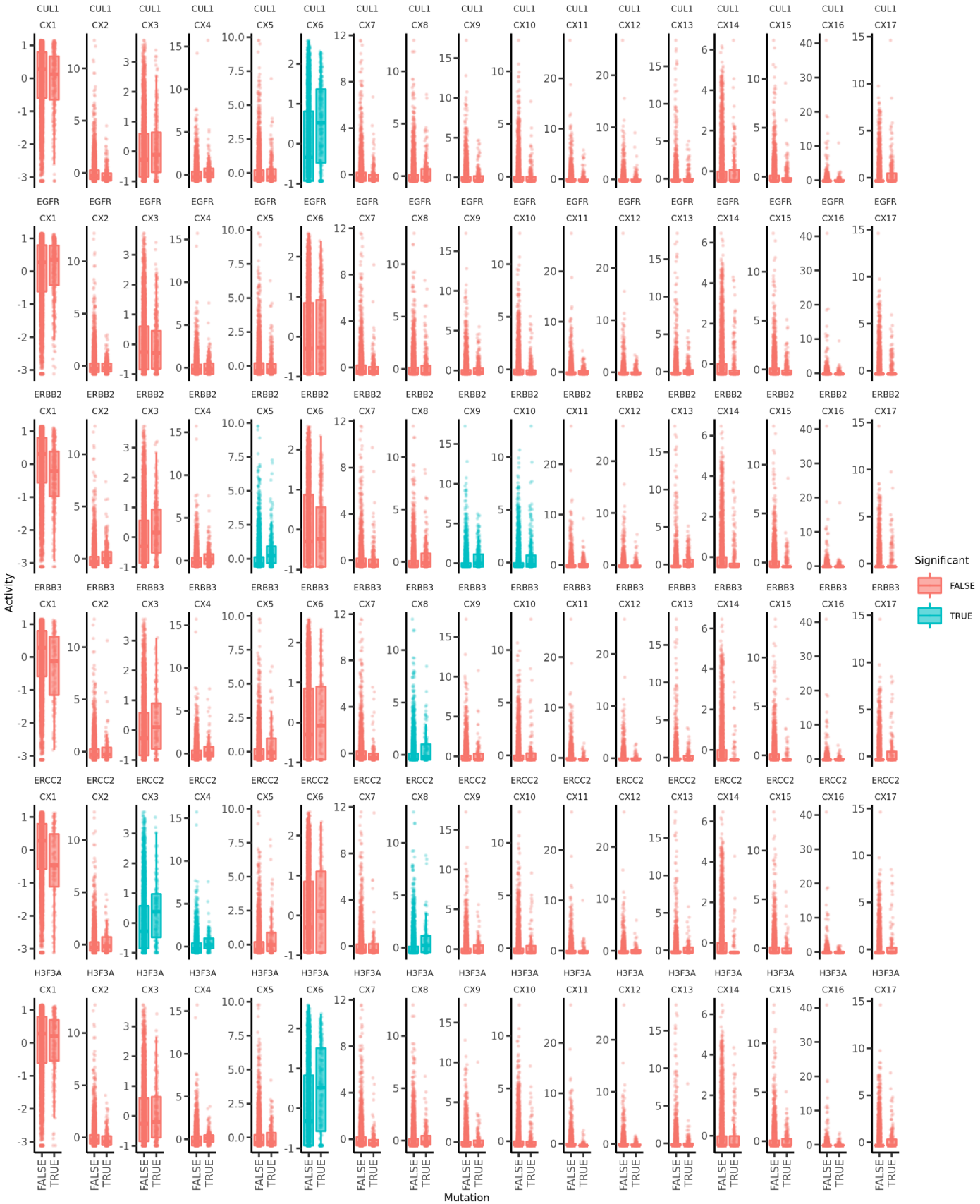


Supplementary Fig. 9: Overview of the signature interpretation matrix derivation

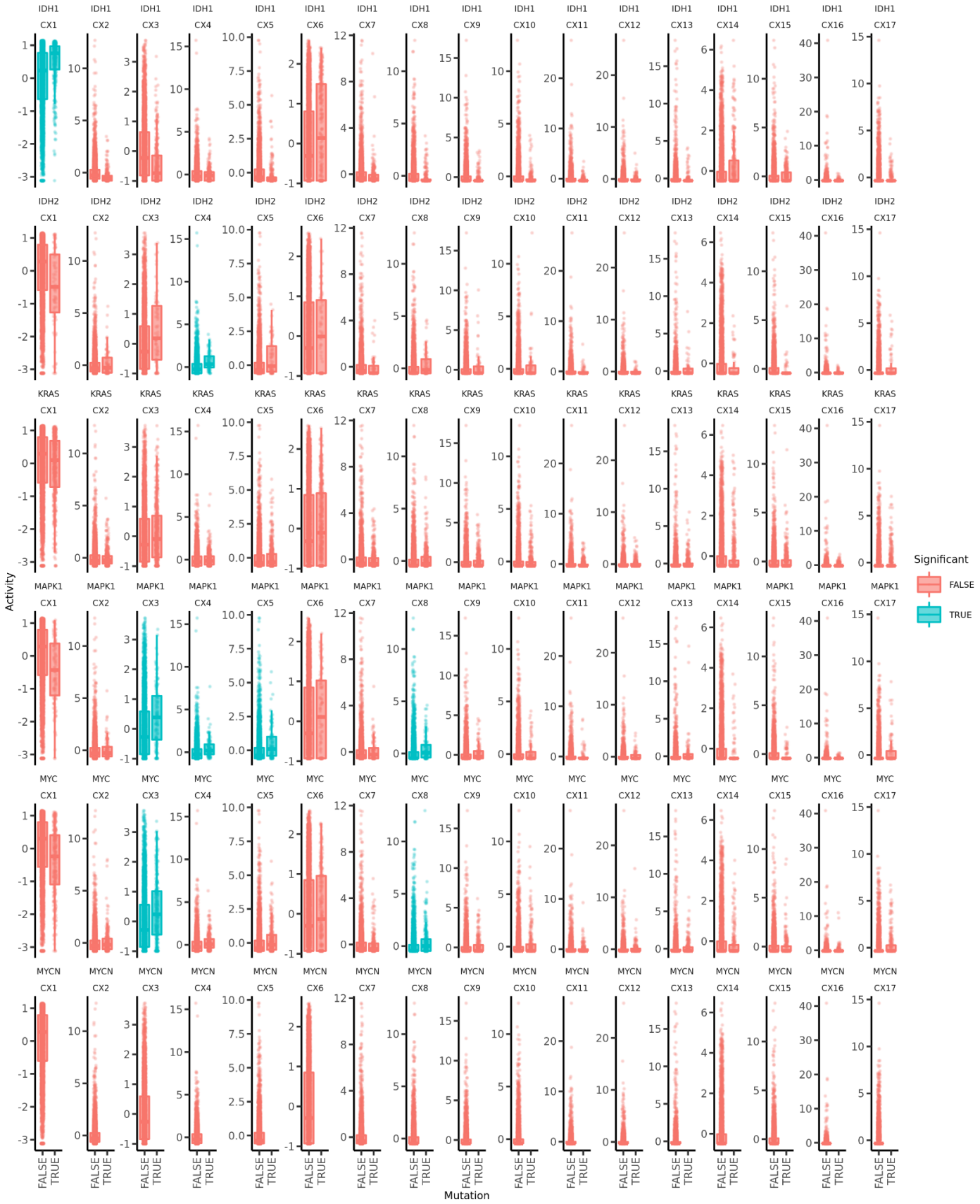
a: The sum-of-posterior input matrix is multiplied by the signature activity matrix to get an activity-corrected sum-of-posteriors matrix. b: Both the activity-corrected sum-of-posteriors and signature definition matrix are normalised to sum to 1, guaranteeing they have equal contributions, and are then summed to generate the interpretation matrix. c: To remove any signal from the sum-of-posterior contribution that does not positively correlate with signature activity, we performed a linear correlation between signature activity and component sum-of-posteriors and only retained weights which showed significant, positive correlation ($\beta>0$ and Benjamini–Hochberg corrected *P*-value < 0.05) generating the final filtered interpretation matrix.



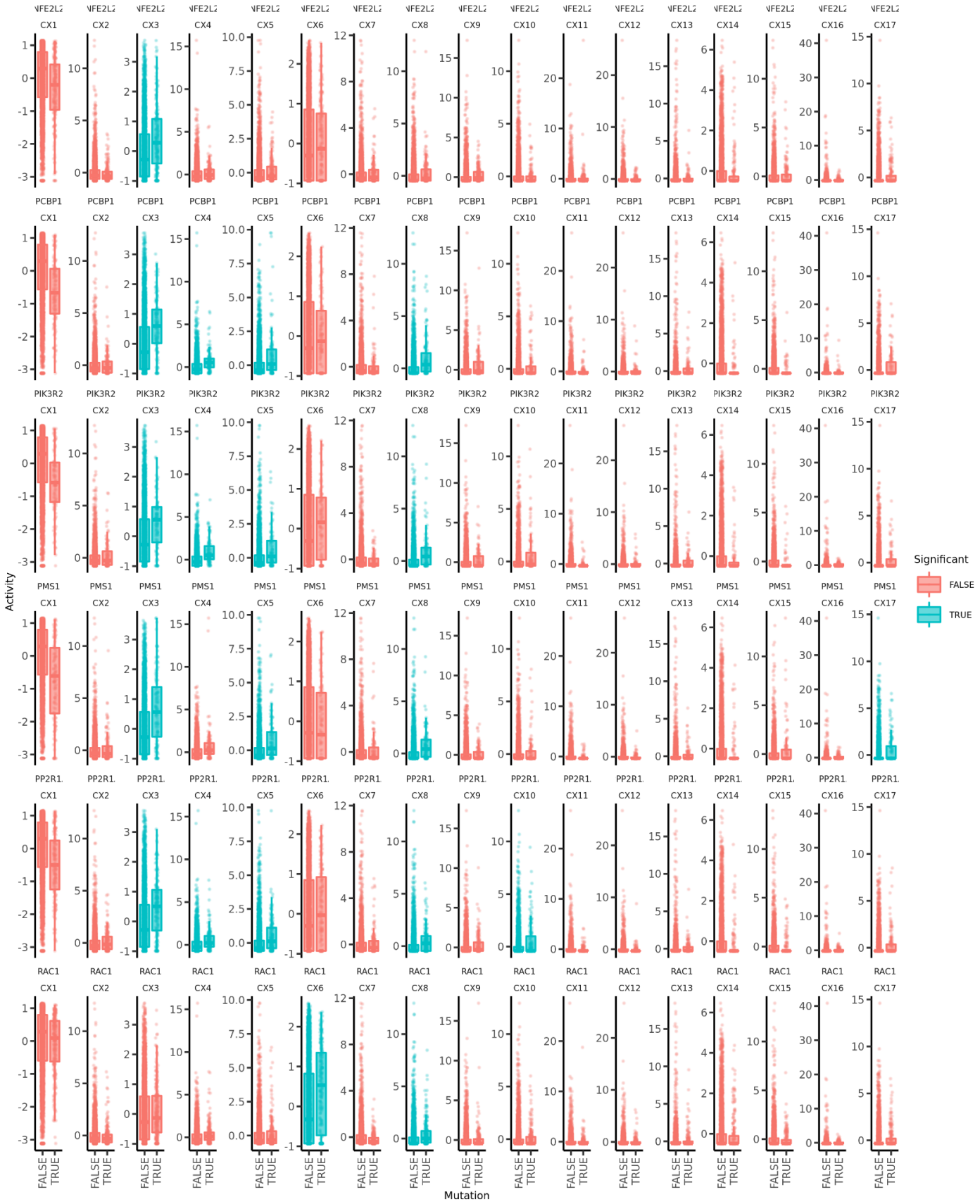
Supplementary Fig. 10: Change in signature activities for mutations in high-confidence genes (Part 1)
Boxplots comparing the signature activities of samples split by the presence of a mutated gene. If a statistical comparison with a two-sided Welch's t-test was significant and directionality was positive (mutants have higher activity), then the boxplots were coloured blue, otherwise red. Boxes represent the interquartile range (IQR) with the median as a bolded line. The whiskers extend to the largest/smallest value no further than 1.5 * IQR from the hinge. Outliers beyond the end of the whiskers are marked individually as points.



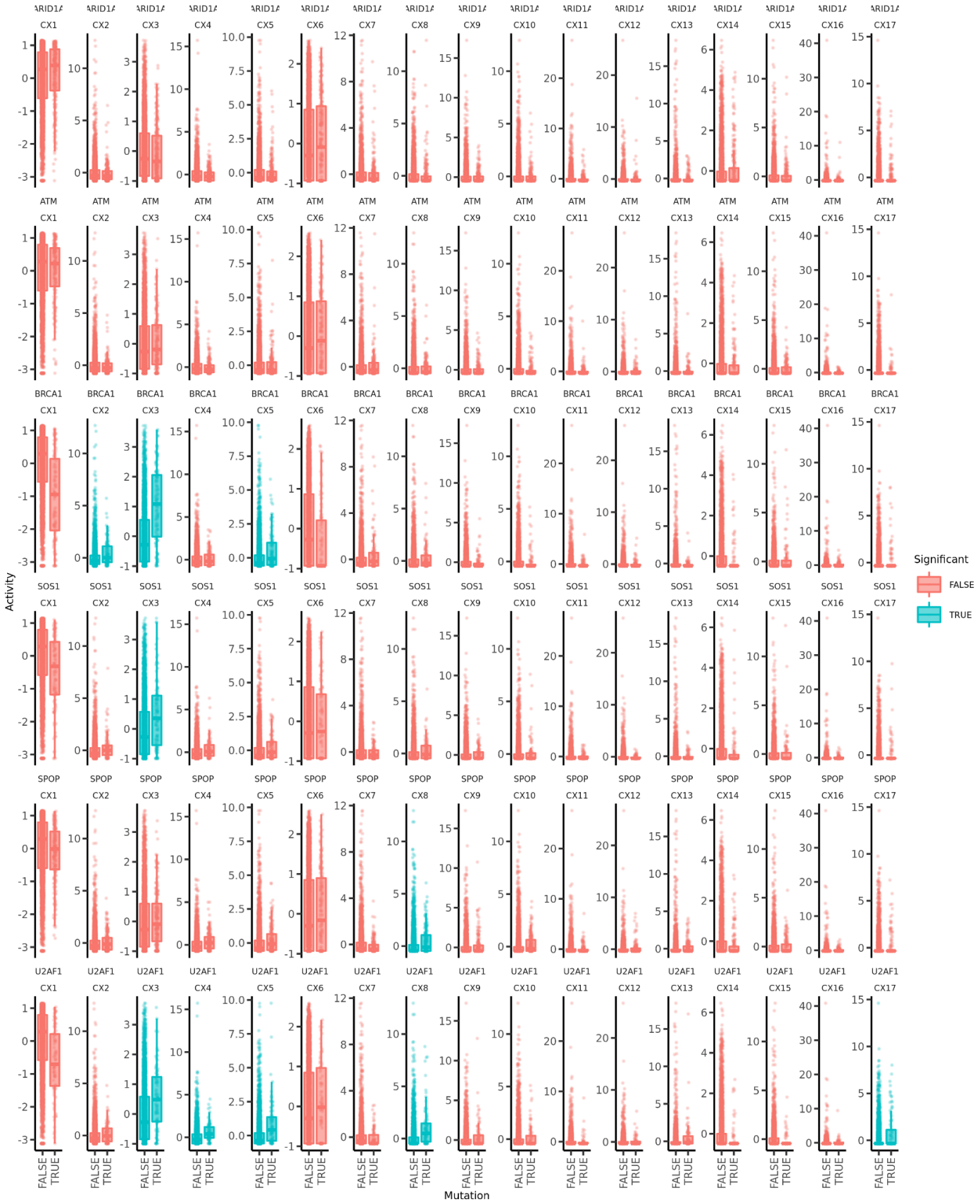
Supplementary Fig. 11: Change in signature activities for mutations in high-confidence genes (Part 2)
Boxplots comparing the signature activities of samples split by the presence of a mutated gene. If a statistical comparison with a two-sided Welch's t-test was significant and directionality was positive (mutants have higher activity), then the boxplots were coloured blue, otherwise red. Boxes represent the interquartile range (IQR) with the median as a bolded line. The whiskers extend to the largest/smallest value no further than 1.5 * IQR from the hinge. Outliers beyond the end of the whiskers are marked individually as points.



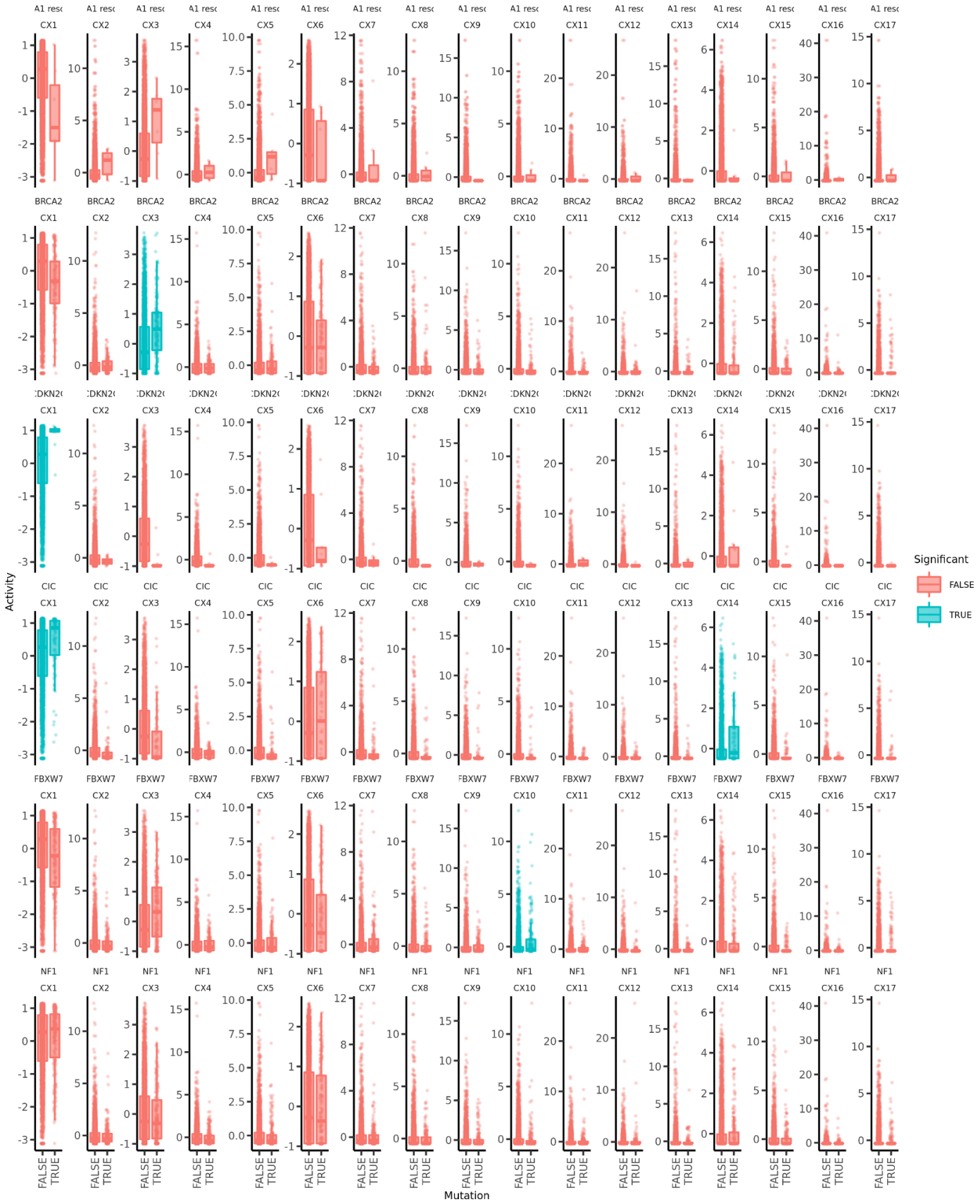
Supplementary Fig. 12: Change in signature activities for mutations in high-confidence genes (Part 3)
Boxplots comparing the signature activities of samples split by the presence of a mutated gene. If a statistical comparison with a two-sided Welch's t-test was significant and directionality was positive (mutants have higher activity), then the boxplots were coloured blue, otherwise red. Boxes represent the interquartile range (IQR) with the median as a bolded line. The whiskers extend to the largest/smallest value no further than 1.5 * IQR from the hinge. Outliers beyond the end of the whiskers are marked individually as points.



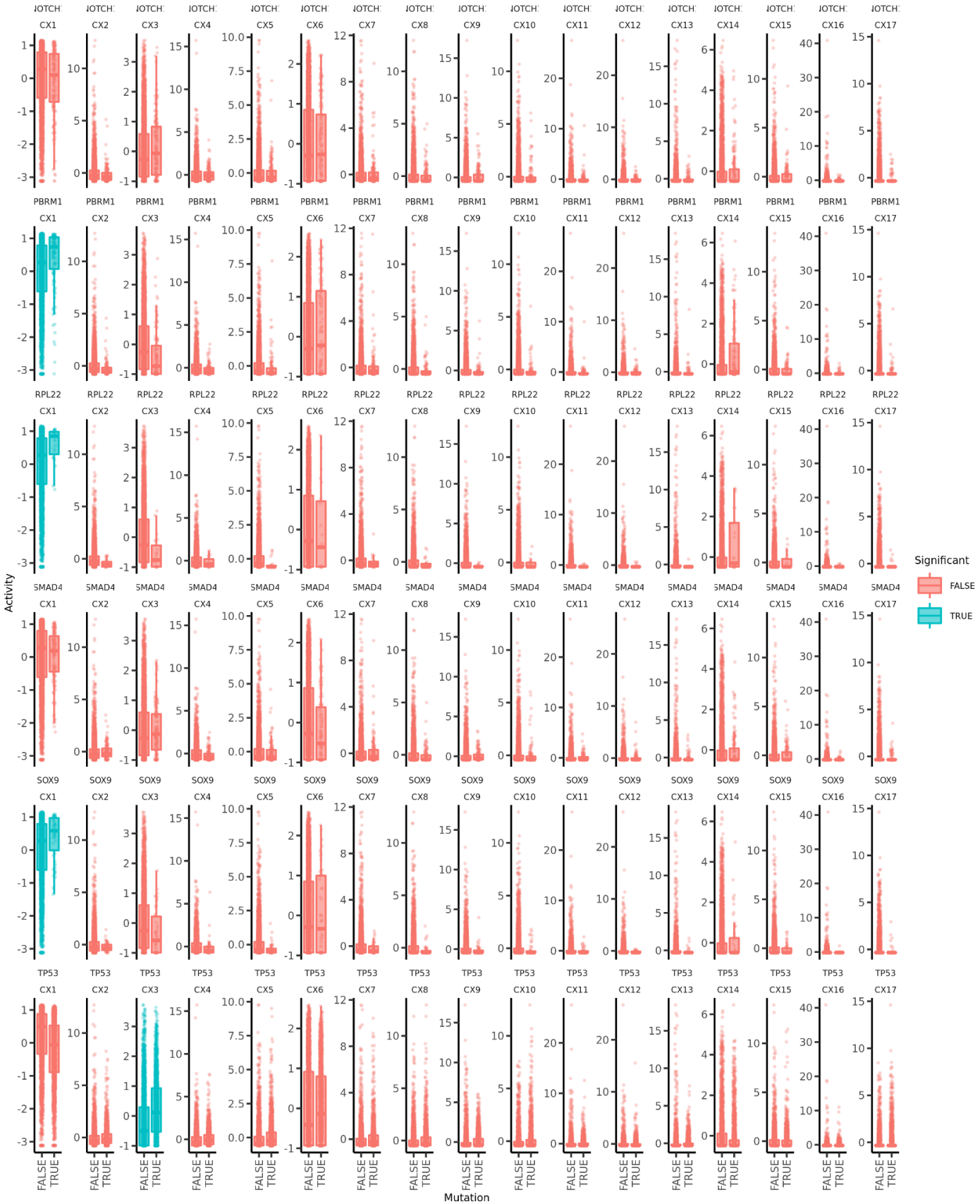
Supplementary Fig. 13: Change in signature activities for mutations in high-confidence genes (Part 4)
Boxplots comparing the signature activities of samples split by the presence of a mutated gene. If a statistical comparison with a two-sided Welch's t-test was significant and directionality was positive (mutants have higher activity), then the boxplots were coloured blue, otherwise red. Boxes represent the interquartile range (IQR) with the median as a bolded line. The whiskers extend to the largest/smallest value no further than 1.5 * IQR from the hinge. Outliers beyond the end of the whiskers are marked individually as points.



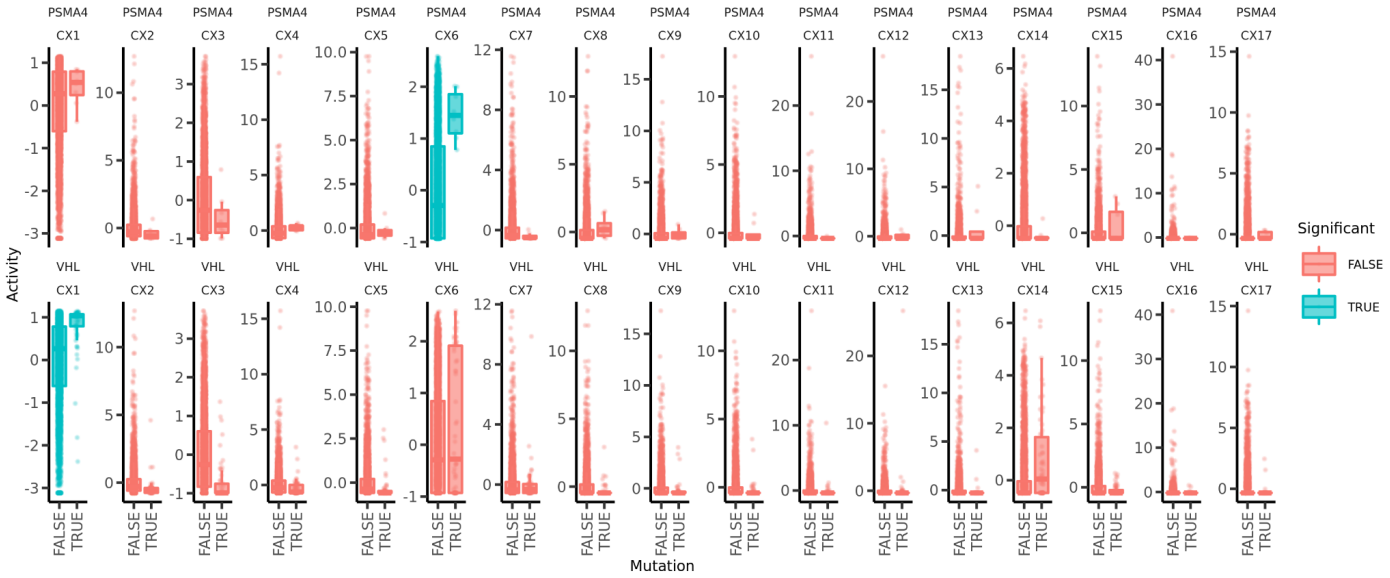
Supplementary Fig. 14: Change in signature activities for mutations in high-confidence genes (Part 5)
Boxplots comparing the signature activities of samples split by the presence of a mutated gene. If a statistical comparison with a two-sided Welch's t-test was significant and directionality was positive (mutants have higher activity), then the boxplots were coloured blue, otherwise red. Boxes represent the interquartile range (IQR) with the median as a bolded line. The whiskers extend to the largest/smallest value no further than 1.5 * IQR from the hinge. Outliers beyond the end of the whiskers are marked individually as points.



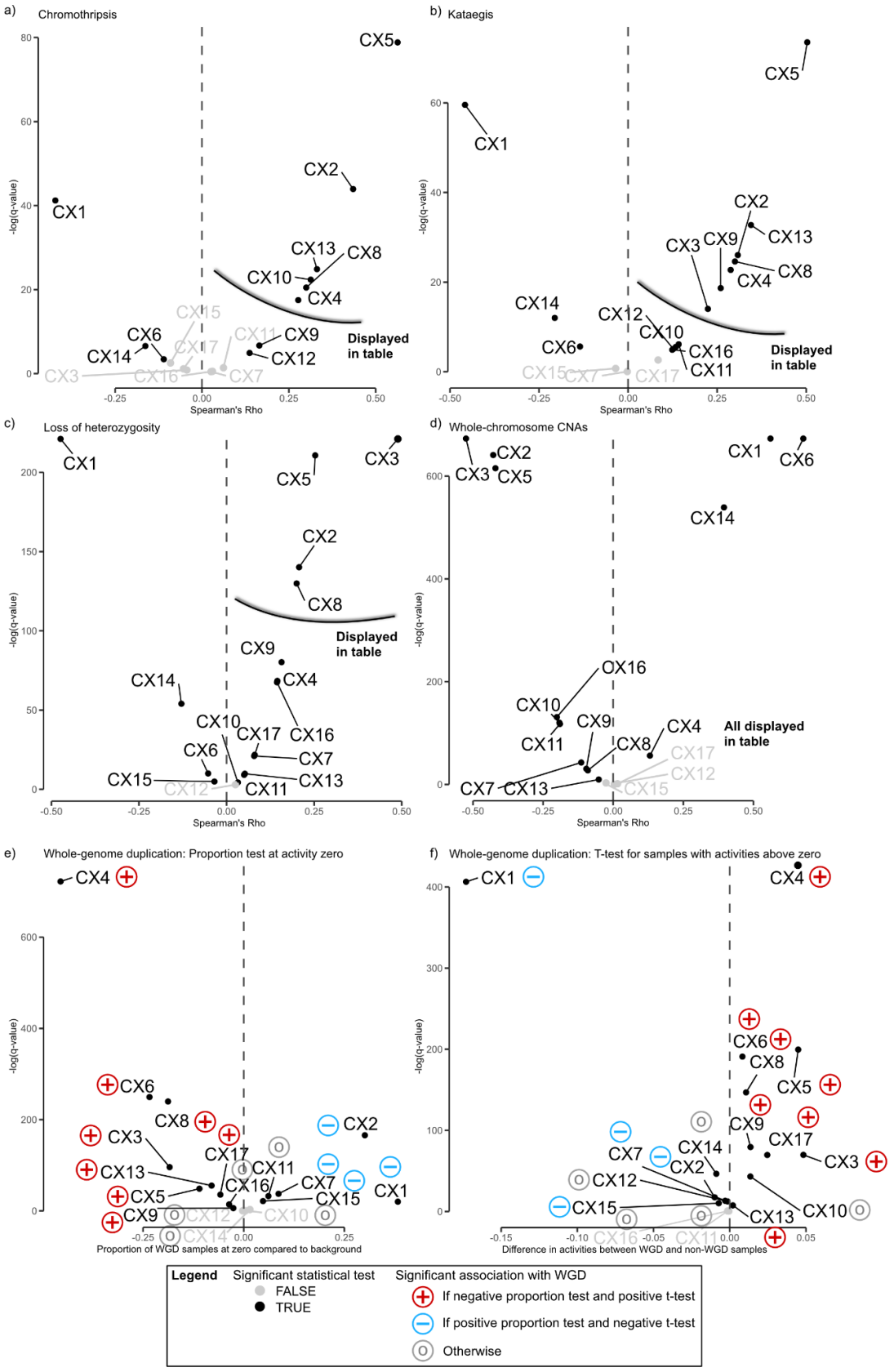
Supplementary Fig. 15: Change in signature activities for mutations in high-confidence genes (Part 6)
Boxplots comparing the signature activities of samples split by the presence of a mutated gene. If a statistical comparison with a two-sided Welch's t-test was significant and directionality was positive (mutants have higher activity), then the boxplots were coloured blue, otherwise red. Boxes represent the interquartile range (IQR) with the median as a bolded line. The whiskers extend to the largest/smallest value no further than 1.5 * IQR from the hinge. Outliers beyond the end of the whiskers are marked individually as points.



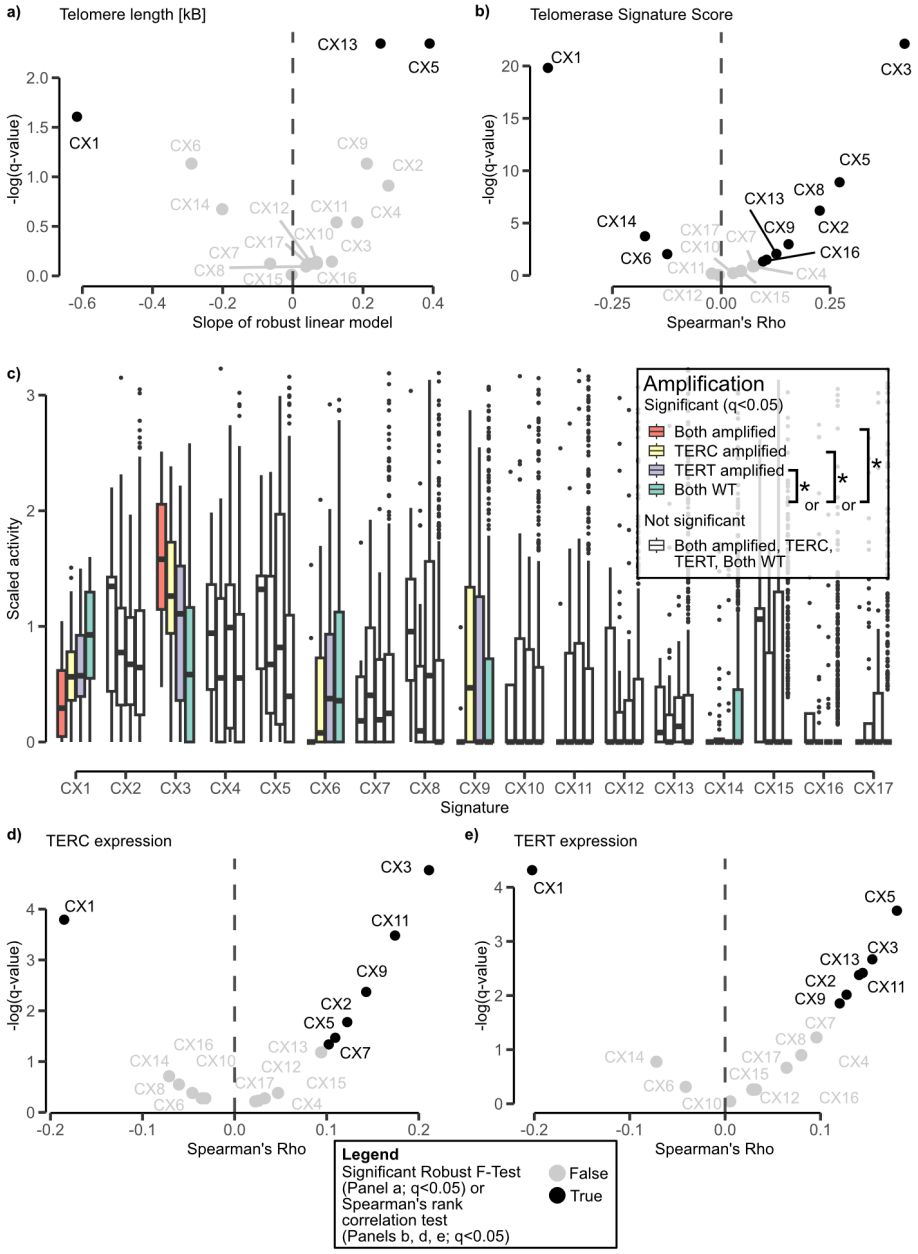
Supplementary Fig. 16: Change in signature activities for mutations in high-confidence genes (Part 7)
Boxplots comparing the signature activities of samples split by the presence of a mutated gene. If a statistical comparison with a two-sided Welch's t-test was significant and directionality was positive (mutants have higher activity), then the boxplots were coloured blue, otherwise red. Boxes represent the interquartile range (IQR) with the median as a bolded line. The whiskers extend to the largest/smallest value no further than 1.5 * IQR from the hinge. Outliers beyond the end of the whiskers are marked individually as points.



Supplementary Fig. 17: Change in signature activities for mutations in high-confidence genes (Part 8)
Boxplots comparing the signature activities of samples split by the presence of a mutated gene. If a statistical comparison with a two-sided Welch's t-test was significant and directionality was positive (mutants have higher activity), then the boxplots were coloured blue, otherwise red. Boxes represent the interquartile range (IQR) with the median as a bolded line. The whiskers extend to the largest/smallest value no further than 1.5 * IQR from the hinge. Outliers beyond the end of the whiskers are marked individually as points.

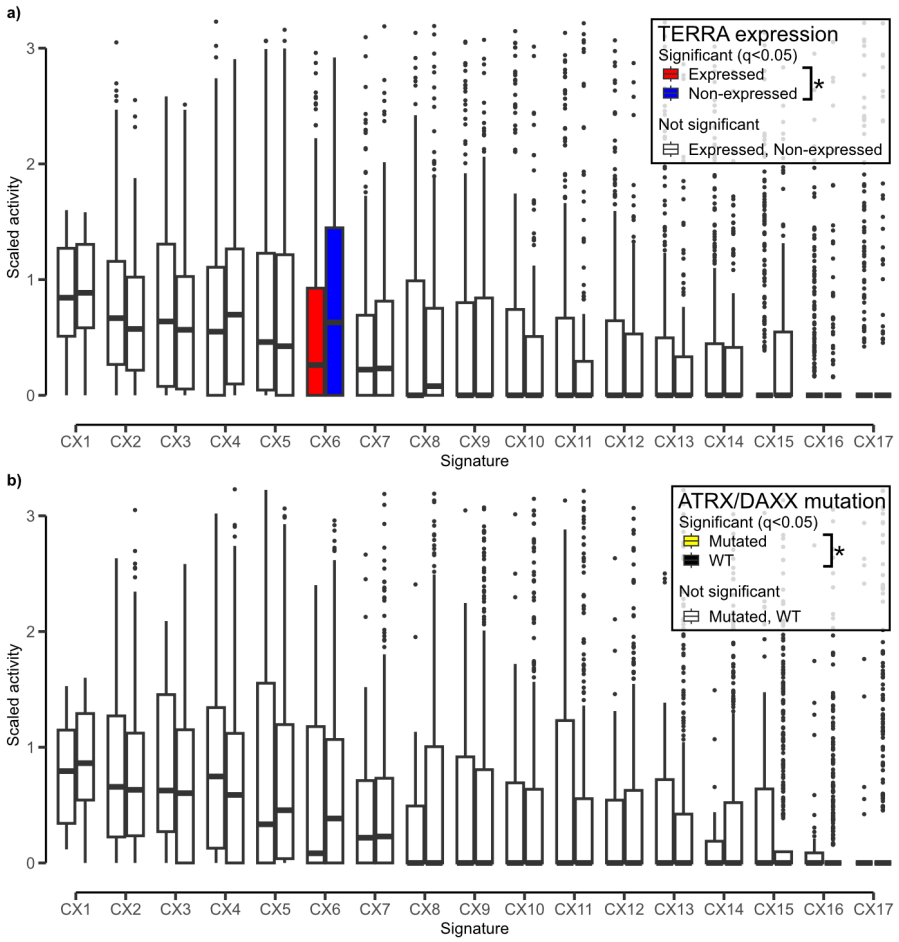


Supplementary Fig. 18: Associations of signature activities with five important biological covariates
Scatterplots of the Spearman correlation between signature activities and chromothriptic events (**a**), kataegis events (**b**), proportion of the genome altered by loss of heterozygosity (**c**) and number of whole-chromosome CNAs (**d**). Summary table is displayed in Extended Data Fig 8b. Scatterplot of two statistical tests studying the relationship of signatures with whole-genome duplication: results of proportion tests for samples with zero activity (**e**) and results of Welch's t-tests for samples with activities above zero (**f**).



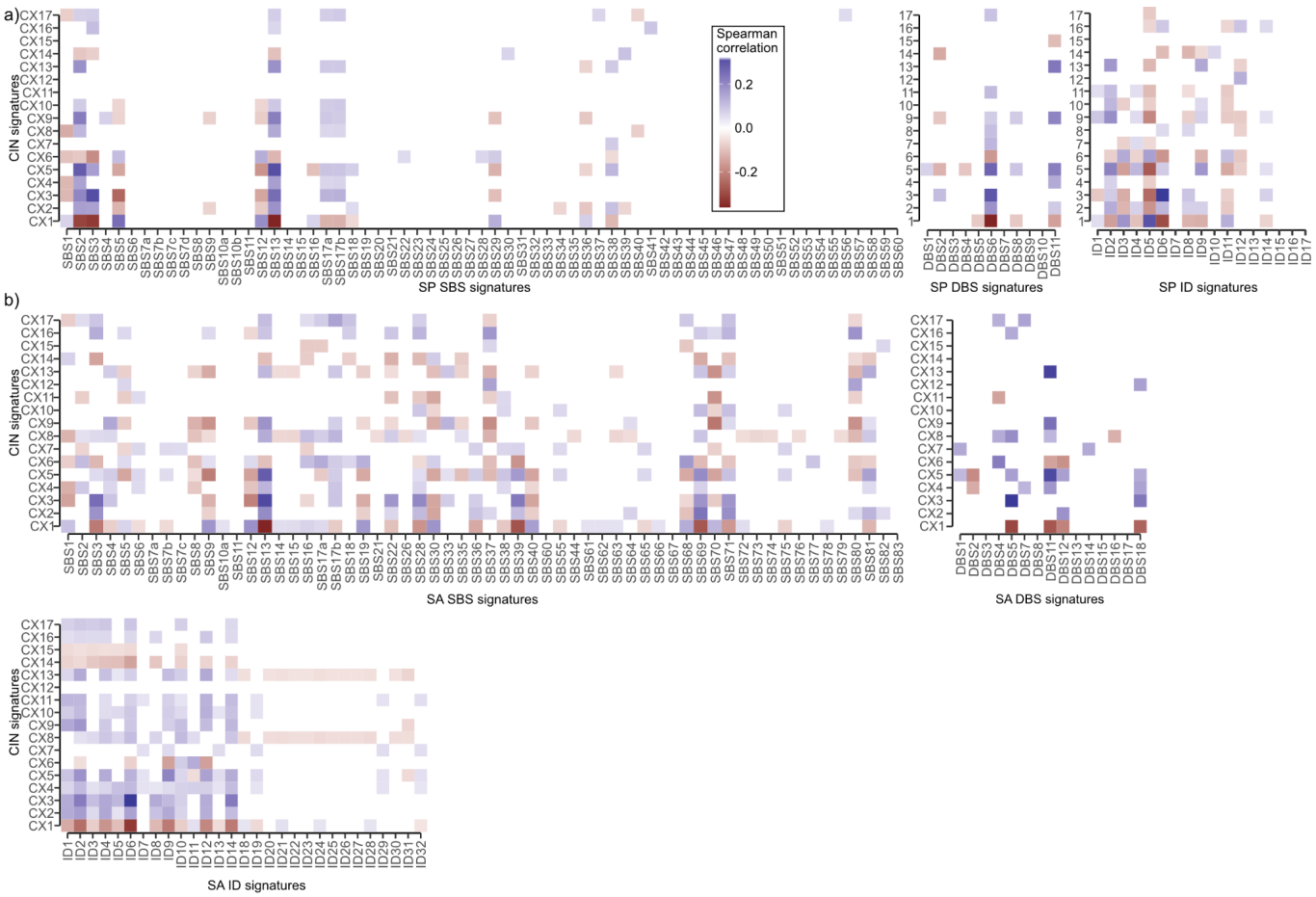
Supplementary Fig. 19: Associations of signature activities with telomere length and telomere elongation machinery (Part 1)

a: Scatterplot of the slopes of robust linear models fitted to signature activities and telomere length. **b:** Scatterplots of the Spearman correlation between signature activities and Telomerase Signature Score. **c:** Boxplots comparing signature activities of samples split by the genomic amplification status of TERT and TERC, crucial parts of the telomerase elongation machinery. **d-e:** Scatterplots of the Spearman correlation between signature activities and TERC (d) or TERC (e) gene expression. **For all boxplots:** Boxes represent the interquartile range (IQR) with the median as a bolded line. The whiskers extend to the largest/smallest value no further than 1.5 * IQR from the hinge. Outliers beyond the end of the whiskers are marked individually as points.



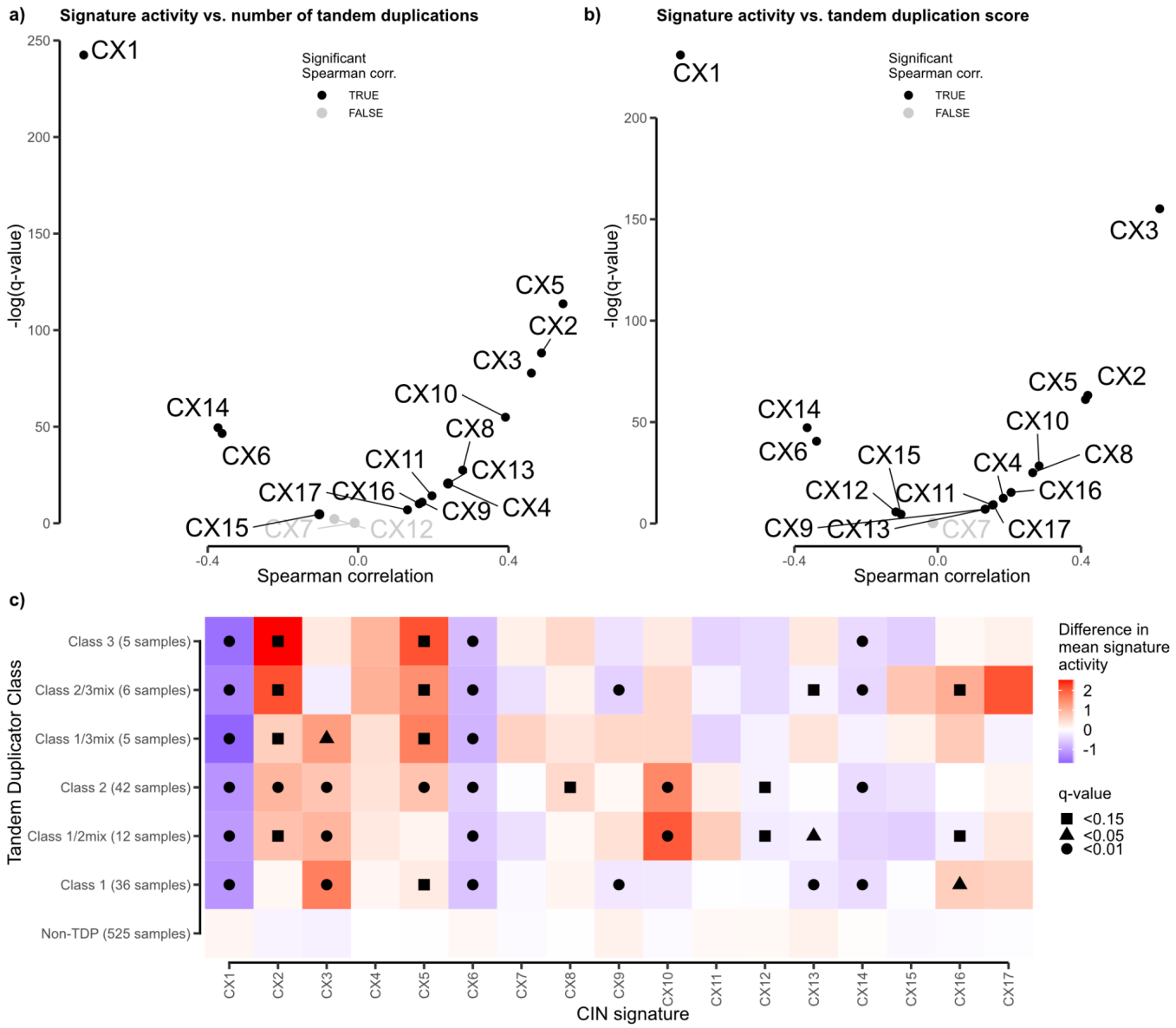
Supplementary Fig. 20: Associations of signature activities with telomere length and telomere elongation machinery (Part 2)

a: Boxplots comparing signature activities of samples split by gene expression status of lncRNA TERRA. **b:** Boxplots comparing signature activities of samples split by the mutation status of ATRX and DAXX, associated with the activation of alternative telomere lengthening (ALT). **For all boxplots:** Boxes represent the interquartile range (IQR) with the median as a bolded line. The whiskers extend to the largest/smallest value no further than 1.5 * IQR from the hinge. Outliers beyond the end of the whiskers are marked individually as points.



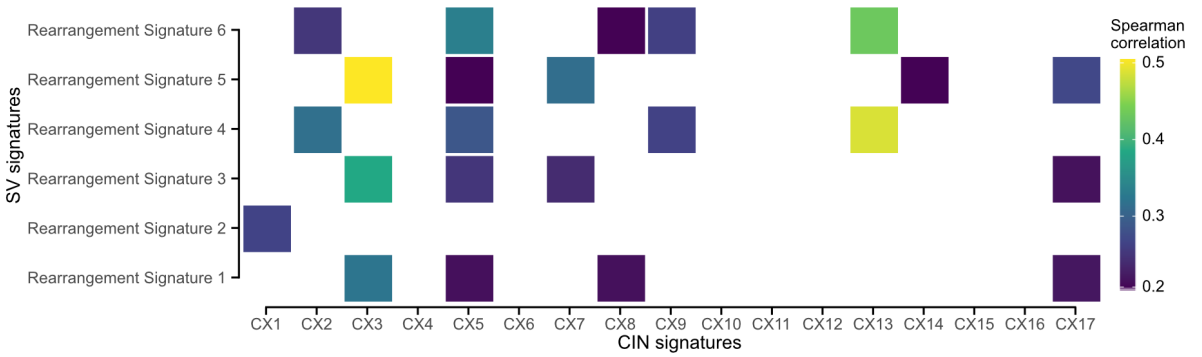
Supplementary Fig. 21: Associations of signature activities with other mutational signatures (SBS, DBS, ID)

a: Heatmaps of significant Spearman correlations between copy number signature activities of the PCAWG cohort and single base substitutions (SBS), doublet base substitutions (DBS) and small insertion and deletion (ID) signatures extracted by Signature Profiler (SP). **b:** Heatmaps of the same comparisons extracted by Signature Analyzer (SA).

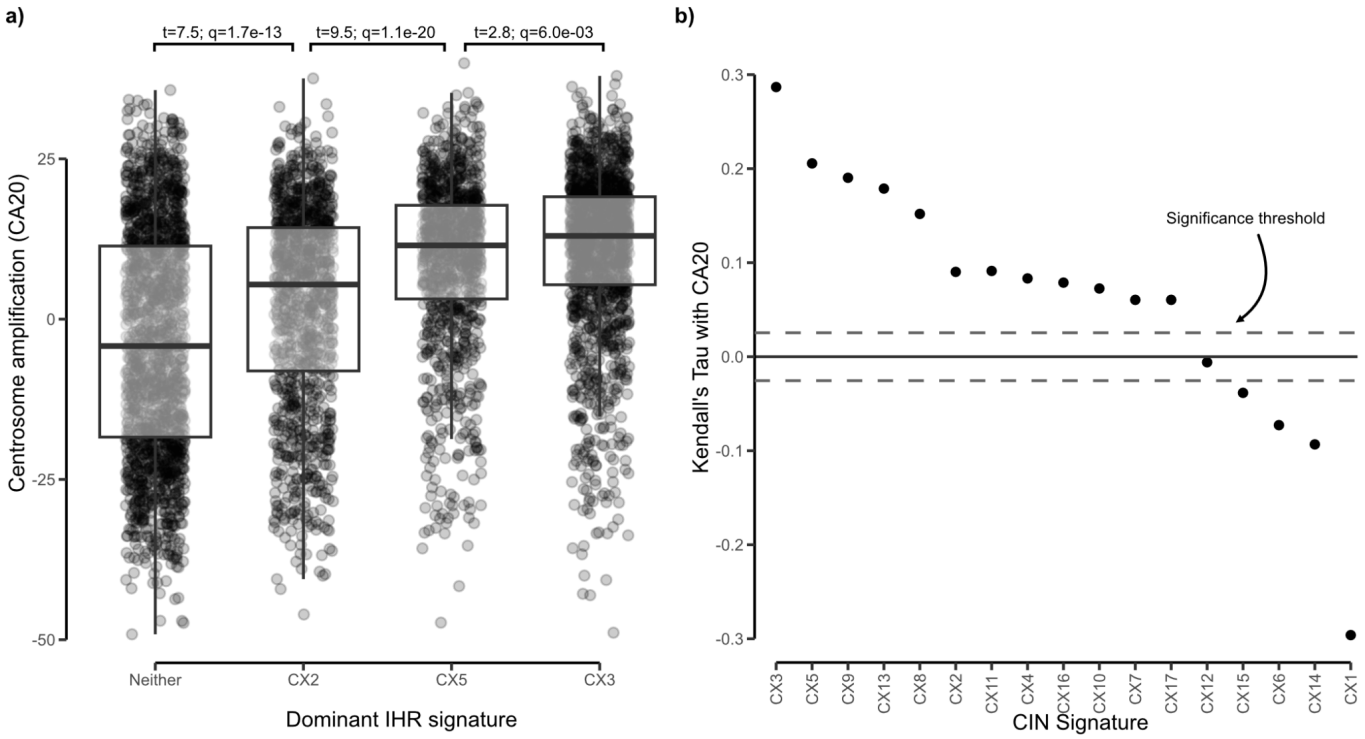


Supplementary Fig. 22: Associations of signature activities with tandem duplication events, scores and classes

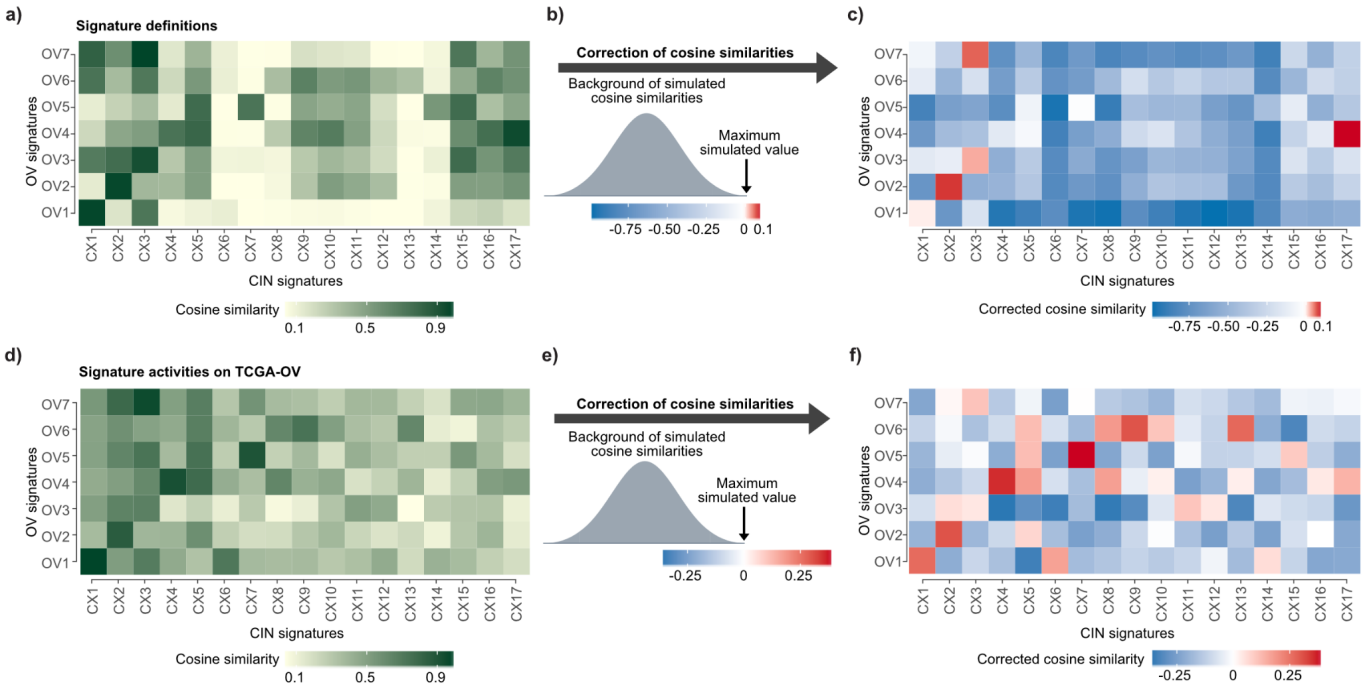
a-b: Scatterplots of the Spearman correlation between signature activities and number of tandem duplications (a) and tandem duplication score (b). c: Heatmap of significant differences in signature activities of TCGA samples with no tandem duplications (non-TDP) and samples categorised into six tandem duplication classes tested by two-sided Welch's t-test.



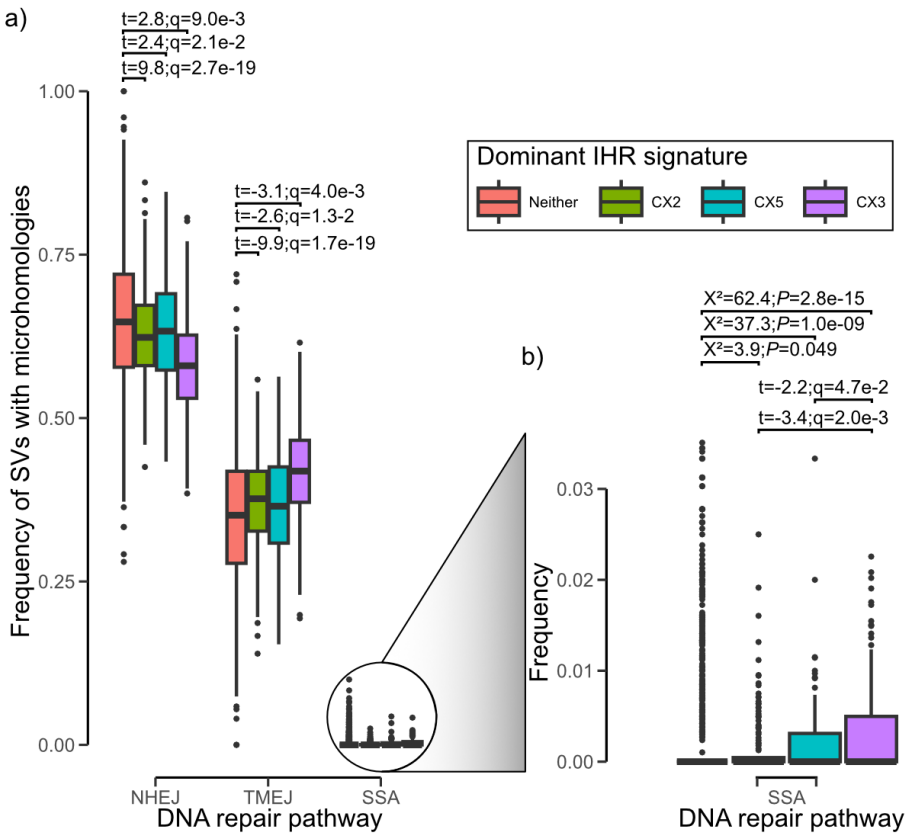
Supplementary Fig. 23: Associations of signature activities with rearrangement signatures
Heatmap of significant Spearman correlations between copy number signature activities and Rearrangement Signature exposures of the ICGC breast cancer cohort.



Supplementary Fig. 24: Associations of signature activities with centrosome amplifications (CA20)
a: Boxplots comparing centrosome amplification score (CA20) of samples split by dominant impaired HR signature. Statistical significance was tested by two-sided Welch's t-test corrected for multiple testing using Benjamini-Hochberg. Stars were assigned when q-value was less than 0.05. Boxes represent the interquartile range (IQR) with the median as a bolded line. The whiskers extend to the largest/smallest value no further than 1.5 * IQR from the hinge. Outliers beyond the end of the whiskers are marked individually as points. **b:** Scatterplot of Kendall correlation between copy number signature activity and CA20 score. Dashed line shows significance threshold for q-values (*P*-values after multiple testing correction using Benjamini-Hochberg method).

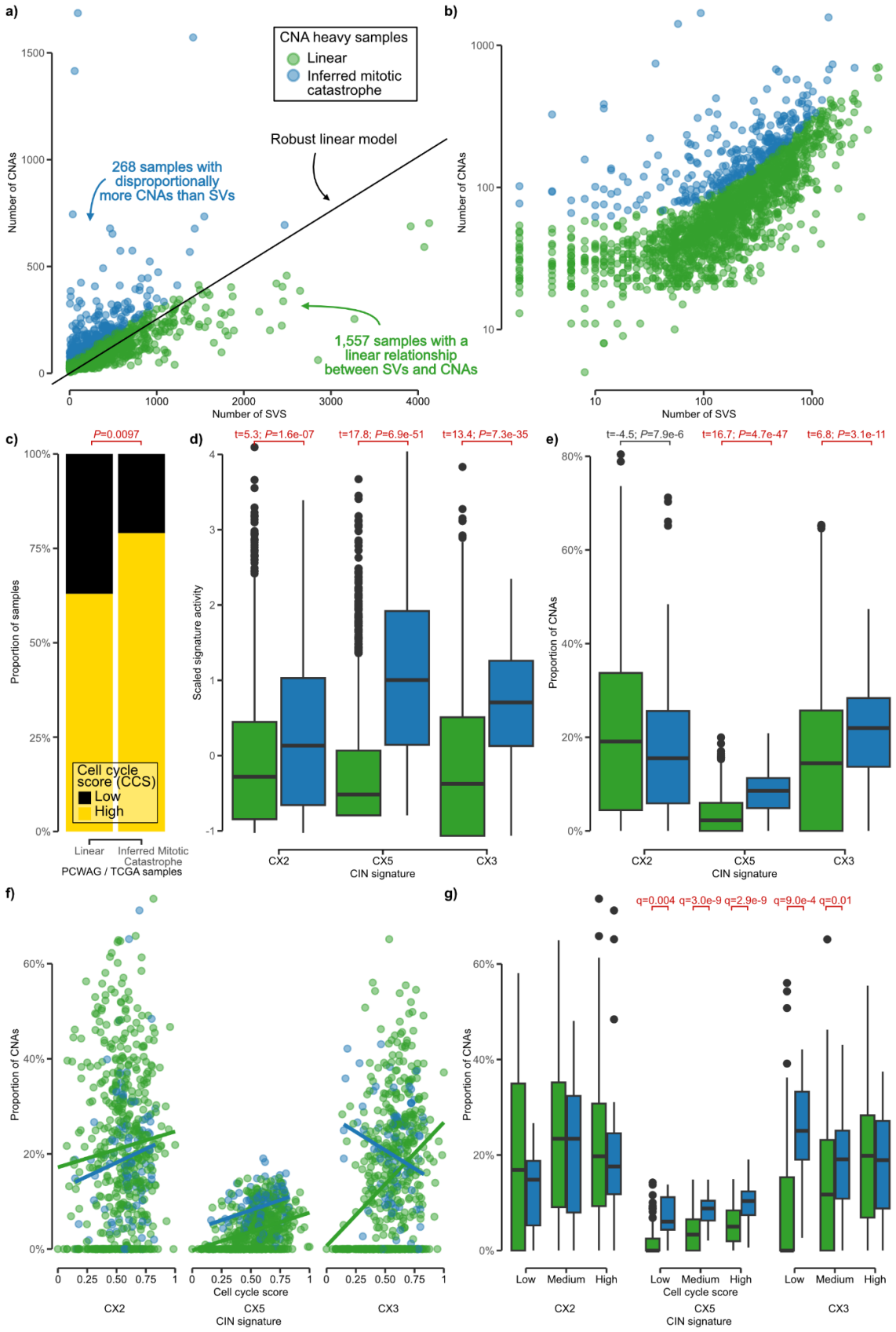


Supplementary Fig. 25: Comparison of 17 CIN signatures with 7 ovarian copy number signatures
a: Cosine similarities between the definitions of the 17 CIN signatures (x-axis) and ovarian copy number signatures (y-axis). **b:** Schematic of cosine similarity correction by maximum of simulated background distribution for each signature pair. **c:** Corrected cosine similarities between signature definitions. Positive values (white to red colour) are significant by default according to a permutation test as they indicate that no simulated value from 10,000 simulations was as large as the observed value (*P*=0.0001). **d:** Cosine similarities between the activities of the 17 CIN signatures (x-axis) and ovarian copy number signatures (y-axis) on the samples of the TCGA ovarian cancer cohort. **e:** Schematic of cosine similarity correction by maximum of simulated background distribution for each signature pair. **f:** Corrected cosine similarities between signature definitions. Positive values (white to red colour) are significant by default according to a permutation test as they indicate that no simulated value from 10,000 simulations was as large as the observed value (*P*=0.0001).



Supplementary Fig. 26: Associations of samples grouped by their dominant IHR signature activities with frequency of microhomologies of SVs

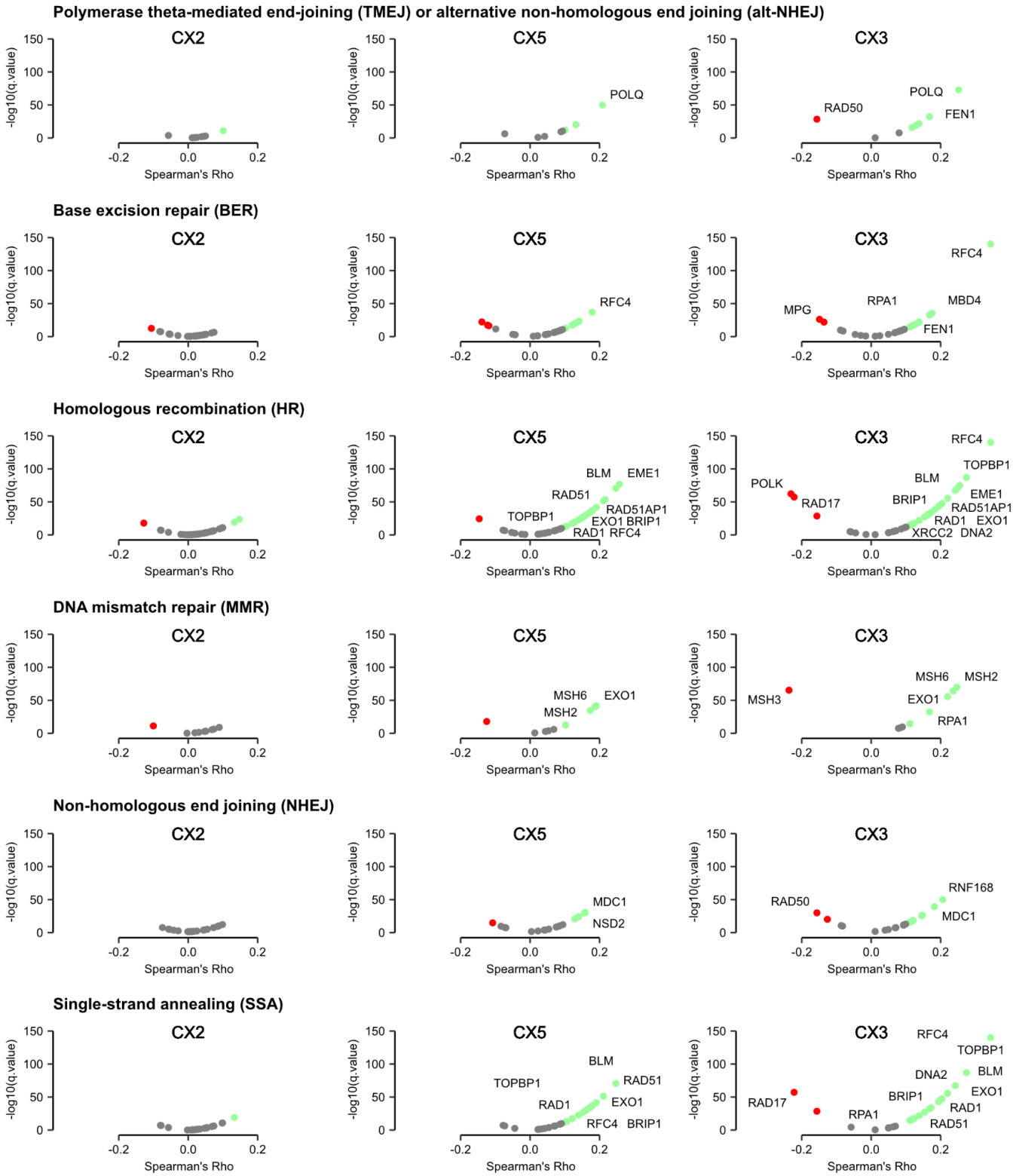
a: Boxplots comparing frequency of structural variants (SVs) with microhomologies split by the repair pathway that most likely created them. PCAWG Samples were split by dominant impaired HR signature. Statistical significance was tested by two-sided Welch's t-test with multiple testing correction using Benjamini-Hochberg method. **b:** Zoomed in boxplot for single strand annealing (SSA) repair pathway. Because of outliers driving the Neither group, statistical significance to the Neither group was tested by a proportion test of samples having no evidence of SSA activity. Statistical significance between the IHR signatures was tested by a two-sided Welch's t-test with multiple testing correction using Benjamini-Hochberg method. **For boxplots:** Boxes represent the interquartile range (IQR) with the median as a bolded line. The whiskers extend to the largest/smallest value no further than 1.5 * IQR from the hinge. Outliers beyond the end of the whiskers are marked individually as points.



Supplementary Fig. 27: Samples with more CNAs experience a mitotic catastrophe phenotype with major contributions by CX2, CX3 and CX5

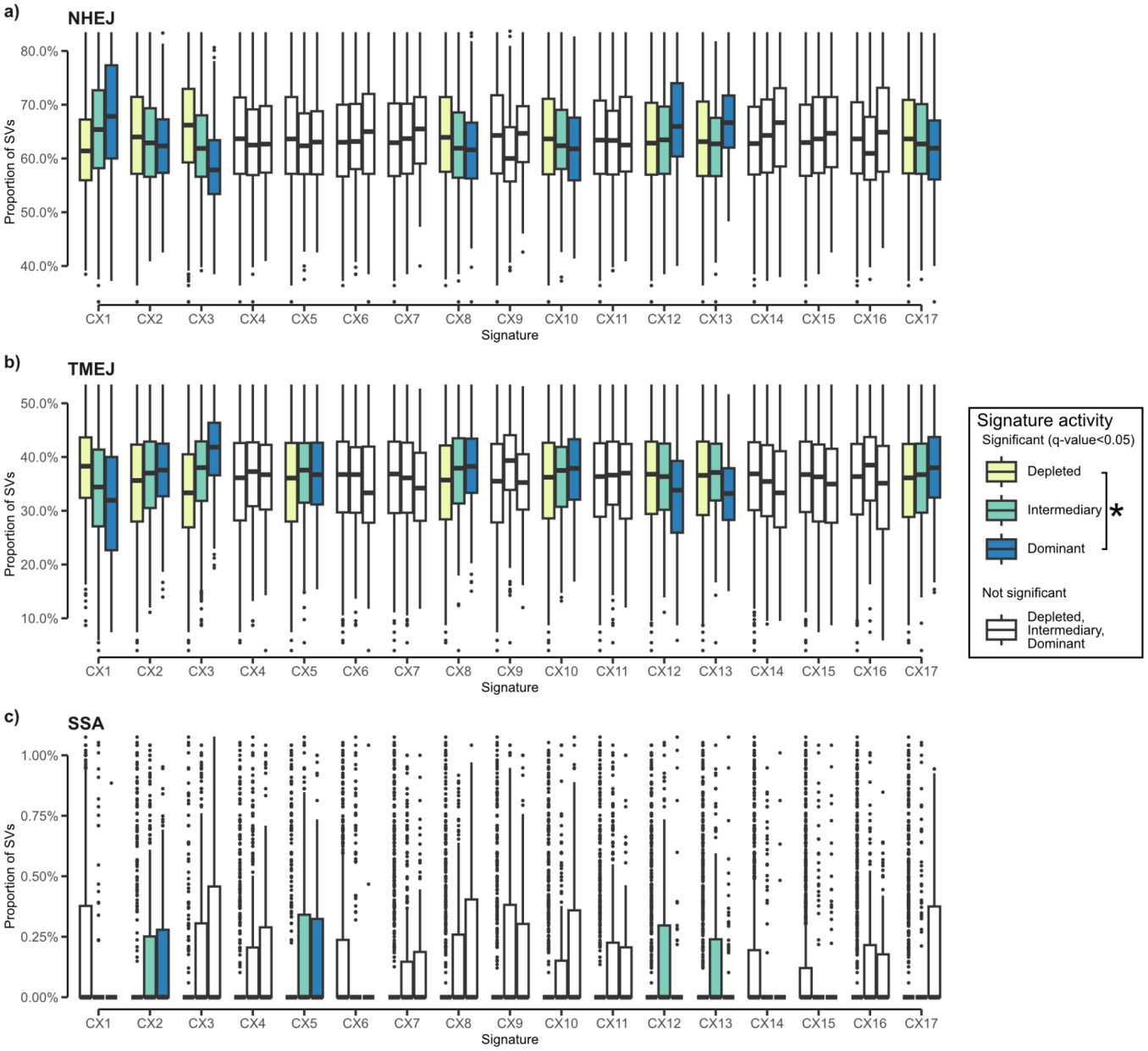
a: Scatterplot of the number of structural variants and the number of copy number aberrations for the PCAWG cohort. Samples with low weights assigned in a robust linear model are samples where we predict an accumulation of CNAs through mitotic catastrophe (blue samples). **b:** Scatterplot of (a) with log10 axis. **c:** Barplot comparing the proportion of high and low Cell cycle scores for samples with predicted mitotic catastrophe (see blue samples in panel a) and samples without (see green samples in panel a). Statistical test was Fisher's exact

test. **d:** Boxplots comparing signature activities for samples with (blue) and without (green) inferred mitotic catastrophe. Comparisons made with a two-sided Welch's t-test. **e:** Boxplots comparing the frequency of estimated number of CNAs caused by impaired HR signatures for samples with (blue) and without (green) inferred mitotic catastrophe. Comparisons made with a two-sided Welch's t-test. **f:** Scatterplot comparing Cell cycle scores and proportion of estimated CNAs for samples with (blue) and without (green) inferred mitotic catastrophe. Lines are fitted linear models. **g:** Boxplots comparing the frequency of estimated number of CNAs caused by impaired HR signatures for samples with (blue) and without (green) inferred mitotic catastrophe and split by Cell cycle score category (low, medium, high). Comparisons made with a two-sided Welch's t-test. **For boxplots:** Boxes represent the interquartile range (IQR) with the median as a bolded line. The whiskers extend to the largest/smallest value no further than 1.5 * IQR from the hinge. Outliers beyond the end of the whiskers are marked individually as points.

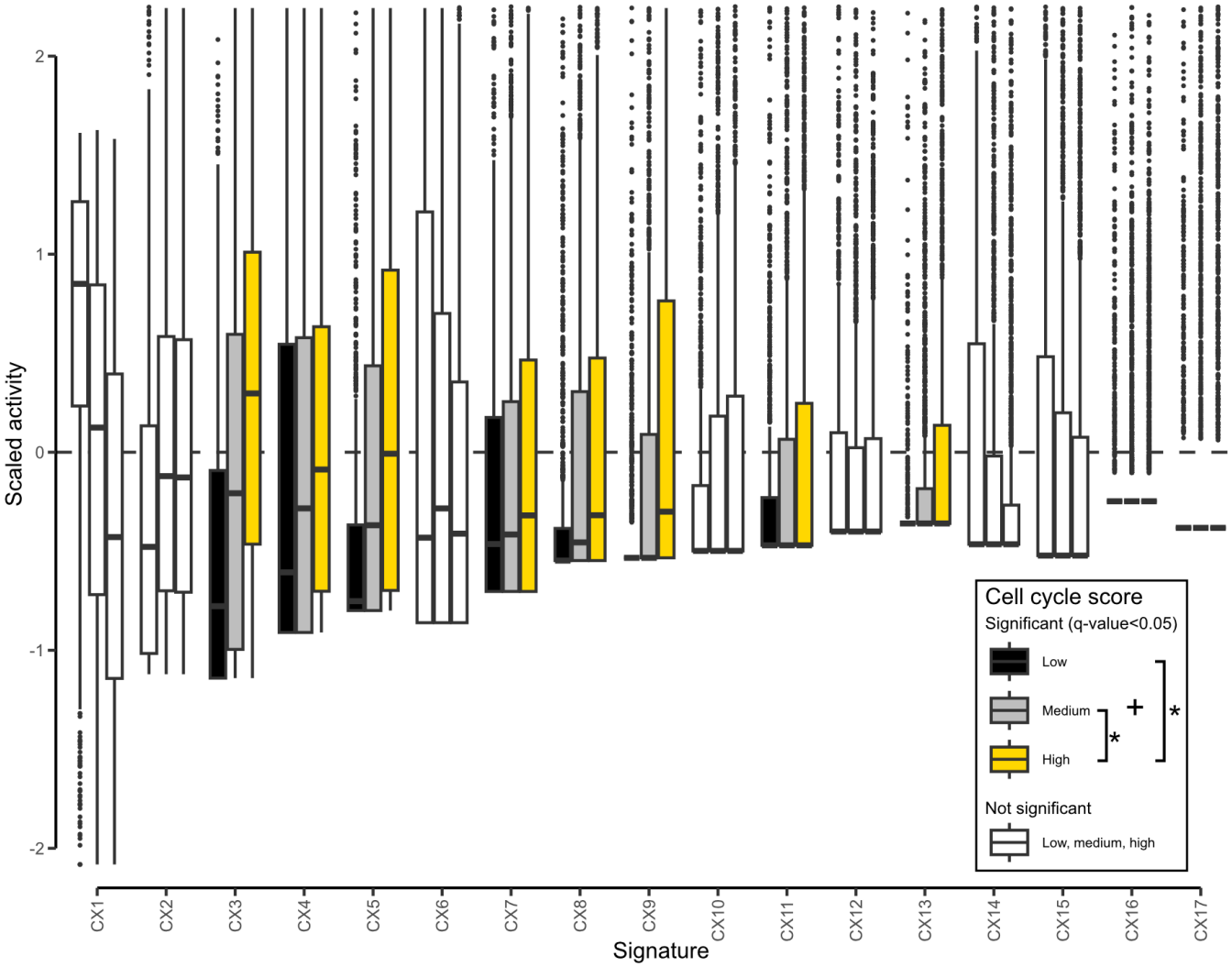


Supplementary Fig. 28: Correlation of gene expression of known DNA repair pathway genes with the activities of the impaired HR signatures

Scatterplots of the Spearman correlations between signature activities of the three impaired HR signatures ordered by levels of perturbation (CX2, CX5, CX3) and gene expression of selected genes shown to be of importance to six DNA repair pathways. Positively associated genes are marked green and negatively associated genes red.

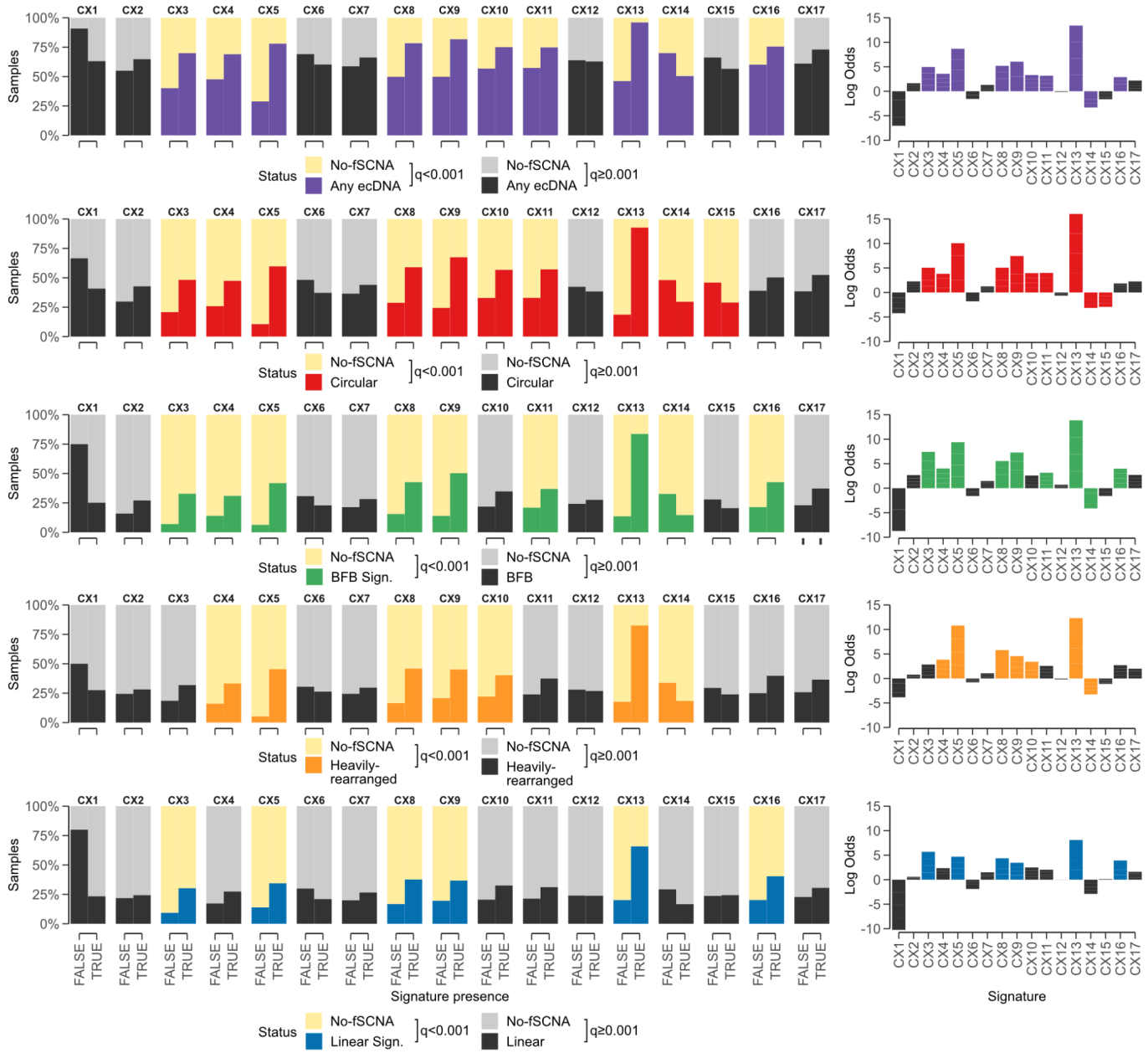


Supplementary Fig. 29: Associations of signature activities with frequency of microhomologies of SVs
a-c: Boxplots comparing frequency of structural variants (SVs) with microhomologies split by the repair pathway that most likely created them: non-homologous end joining (NHEJ) (a), polymerase theta-mediated end joining (TMEJ) (b) and single strand annealing (SSA) (c). PCAWG Samples were split by dominant impaired HR signature. Statistical significance was tested by two-sided Welch's t-test and coloured when there was a significant difference between depleted samples and samples with a dominant signature. Boxes represent the interquartile range (IQR) with the median as a bolded line. The whiskers extend to the largest/smallest value no further than 1.5 * IQR from the hinge. Outliers beyond the end of the whiskers are marked individually as points.

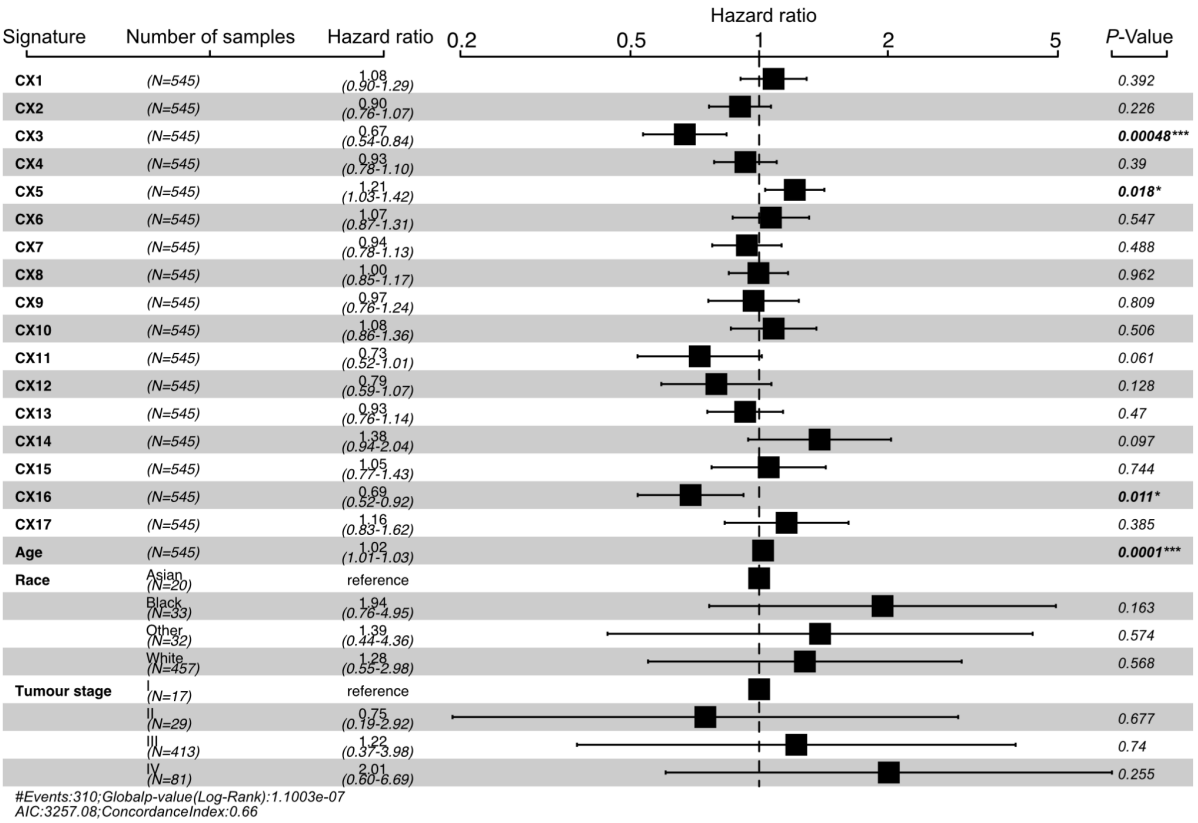


Supplementary Fig. 30: Associations of signature activities with cell cycle score

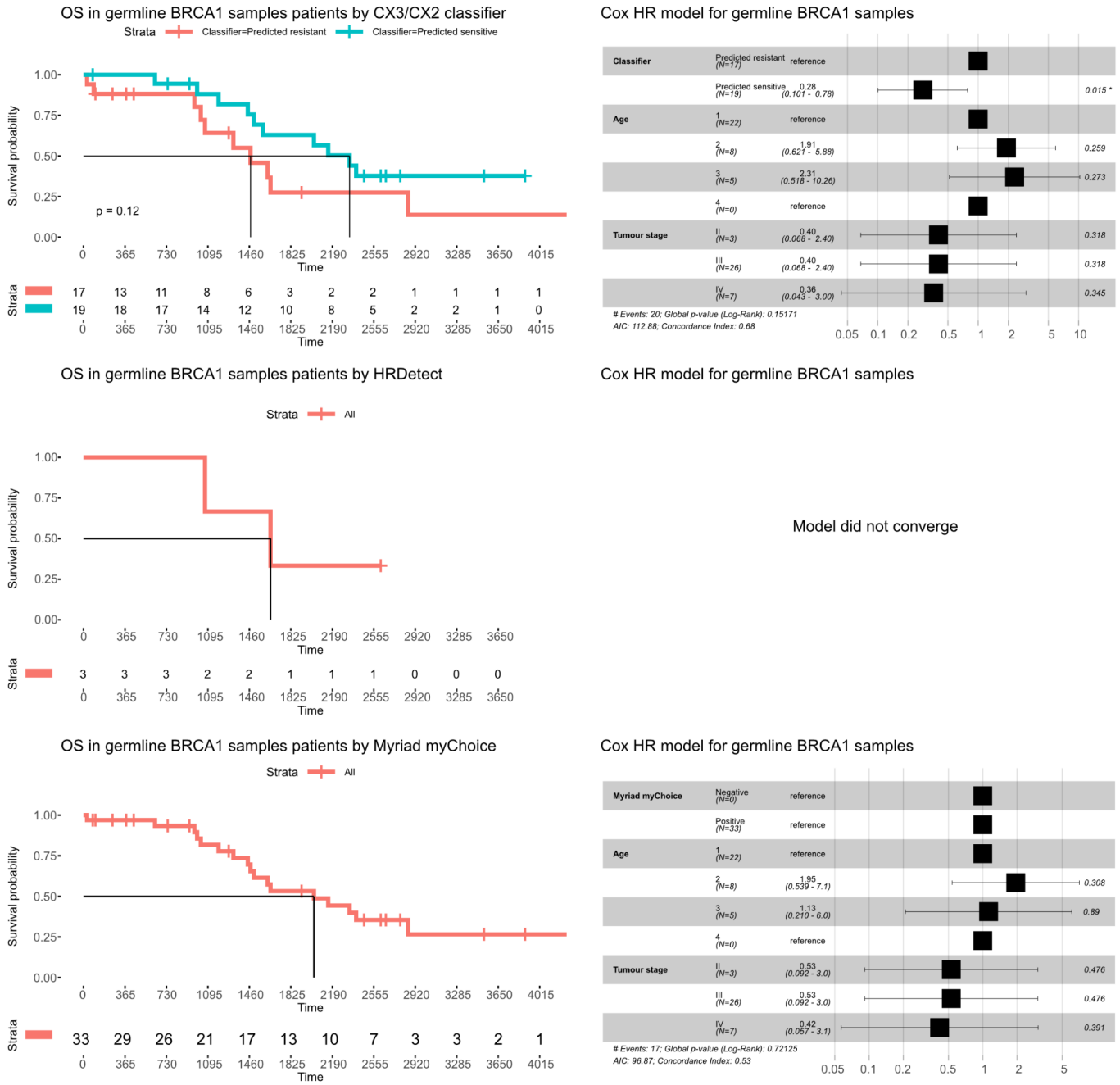
Boxplots comparing signature activity between groups of samples split by their Cell cycle score (CCS). Statistical significance was tested by two-sided Welch's t-test and coloured when there were significant differences between the low and high cell cycle group and the medium and high cell cycle group. Boxes represent the interquartile range (IQR) with the median as a bolded line. The whiskers extend to the largest/smallest value no further than 1.5 * IQR from the hinge. Outliers beyond the end of the whiskers are marked individually as points.



Supplementary Fig. 31: Associations of signature activities with extrachromosomal DNA (ecDNA)
Barplots comparing the presence of ecDNA amplicons with the presence of the 17 CIN signatures for TCGA samples. Statistical significance was tested with Fisher's exact test. Log odds are displayed in the barplot on the right. If found significant, grey and black were switched to brighter colours: beige for no focal SCNA (No-fSCNA), and purple/red/green/yellow/blue for specific types of ecDNA amplicons.

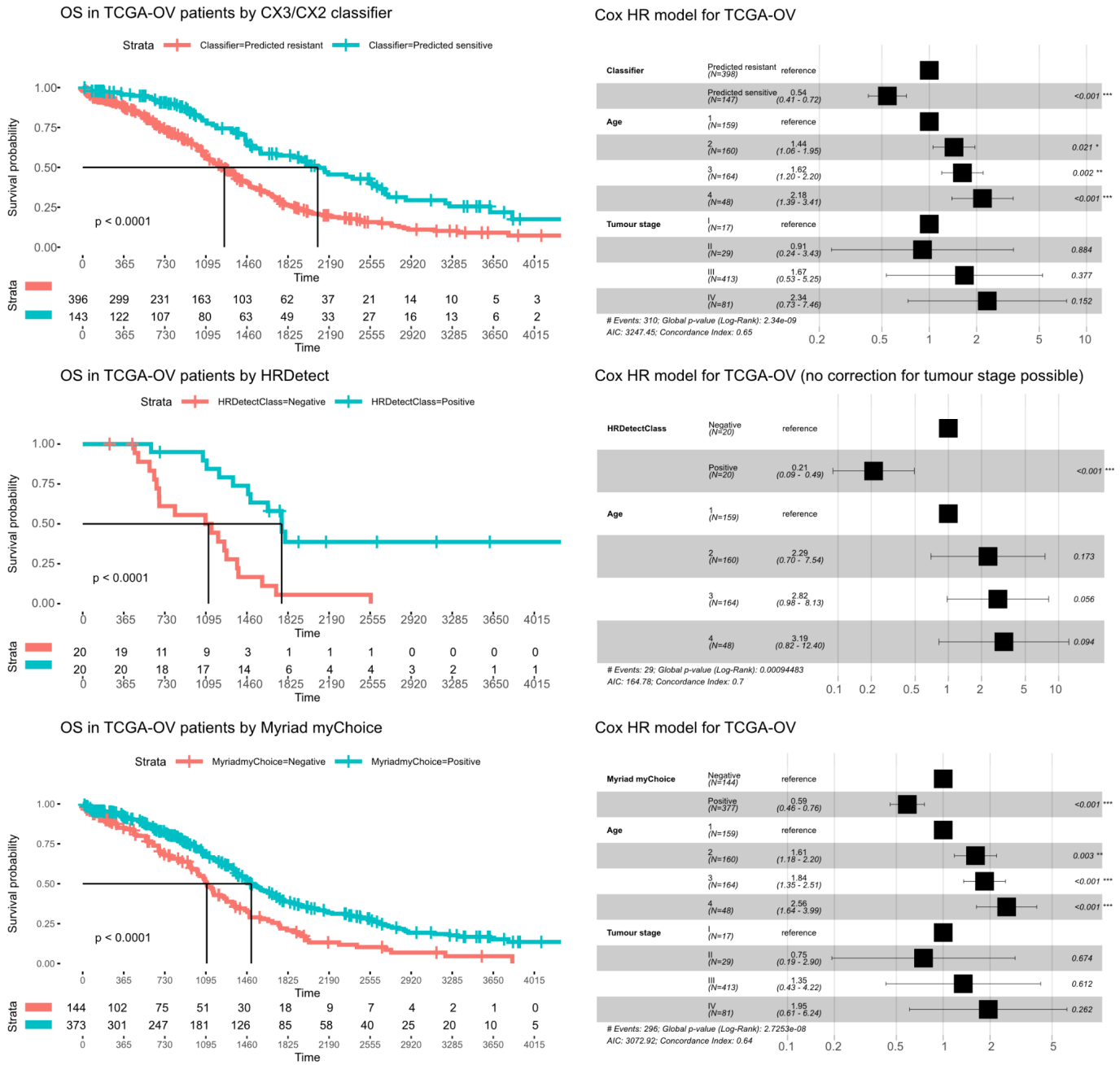


Supplementary Fig. 32: Associations of signature activities with overall survival of TCGA-OV cancers
Hazard ratios and their 95% confidence interval obtained from a Cox proportional hazards model on the overall survival of the TCGA ovarian cancer cohort. Covariates are the activity of the 17 signatures, age, race and tumour stage of the 545 samples.



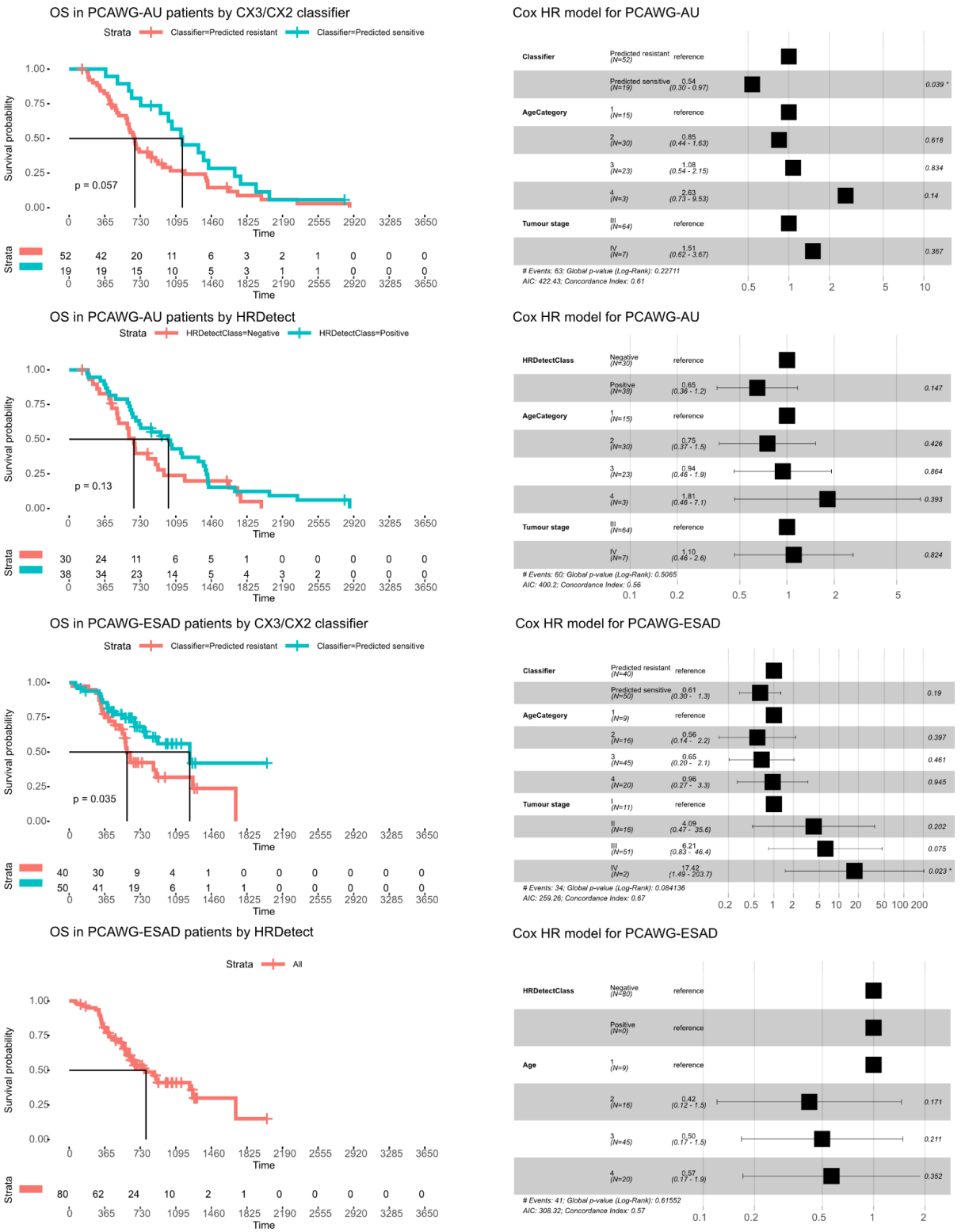
Supplementary Fig. 33: Kaplan-Meier curves and Cox proportional hazards model for germline BRCA1 samples of the TCGA-OV cohort

Left column: Univariate survival modelling by a Kaplan-Meier estimate for germline BRCA1 mutant samples of the TCGA-OV cohort. **Right column:** Hazard ratios and their 95% confidence interval obtained from a Cox proportional hazards model correcting for age and tumour stage (multivariate survival modelling). **Top row:** Clinical classification predicting sensitivity to platinum-based chemotherapy done by our CX2/CX3 classifier. **Middle row:** Clinical classification performed by HRDetect. **Bottom row:** Clinical classification performed by HRD score (Myriad myChoice).



Supplementary Fig. 34: Kaplan-Meier curves and Cox proportional hazards model for all TCGA-OV samples

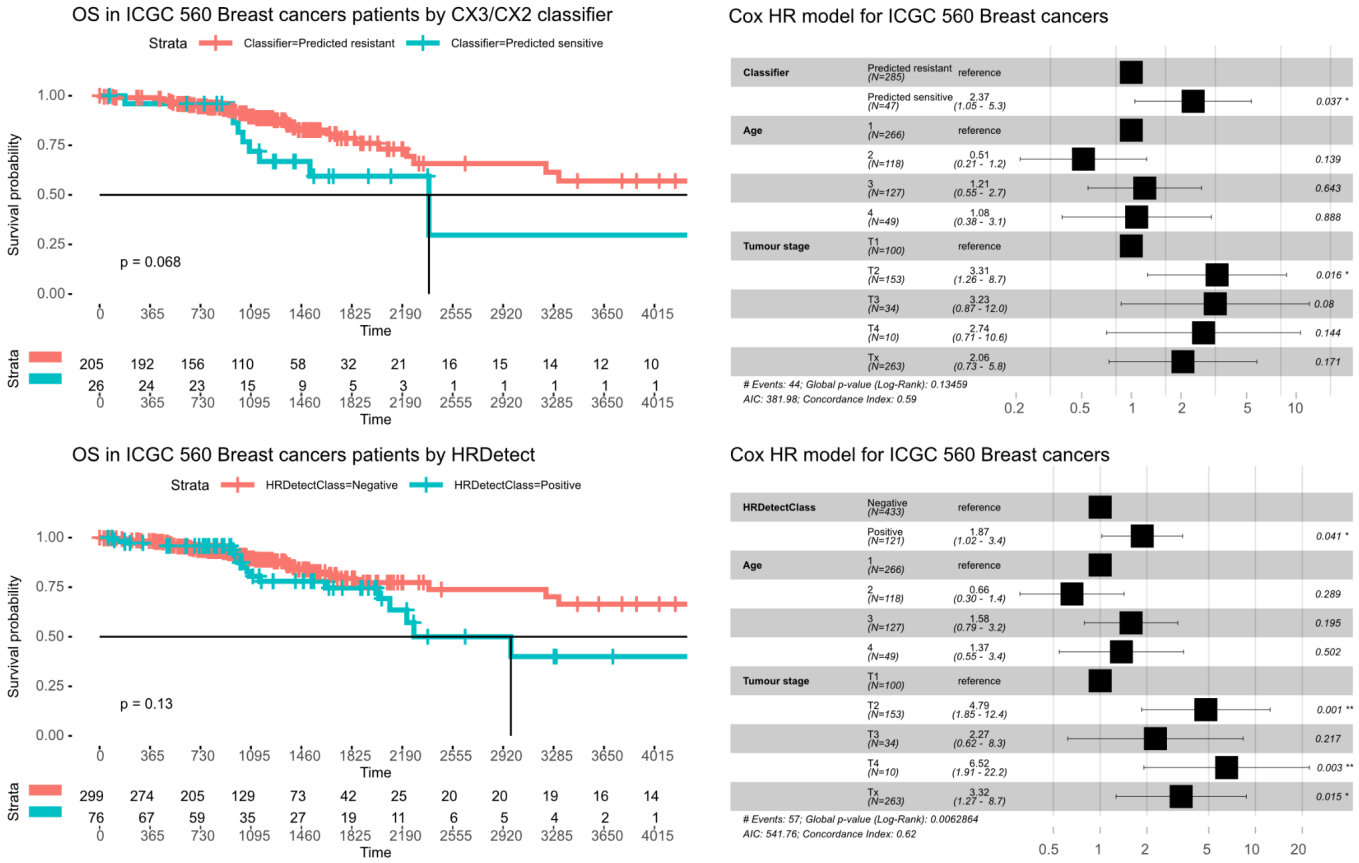
Left column: Univariate survival modelling by a Kaplan-Meier estimate for the TCGA-OV cohort. **Right column:** Hazard ratios and their 95% confidence interval obtained from a Cox proportional hazards model correcting for age and tumour stage (multivariate survival modelling). Cox model for HRDetect excluded tumour stage. **Top row:** Clinical classification predicting sensitivity to platinum-based chemotherapy done by our CX2/CX3 classifier. **Middle row:** Clinical classification performed by HRDetect. **Bottom row:** Clinical classification performed by HRD score (Myriad myChoice).



Supplementary Fig. 35: Kaplan-Meier curves and Cox proportional hazards model for two PCAWG cohorts

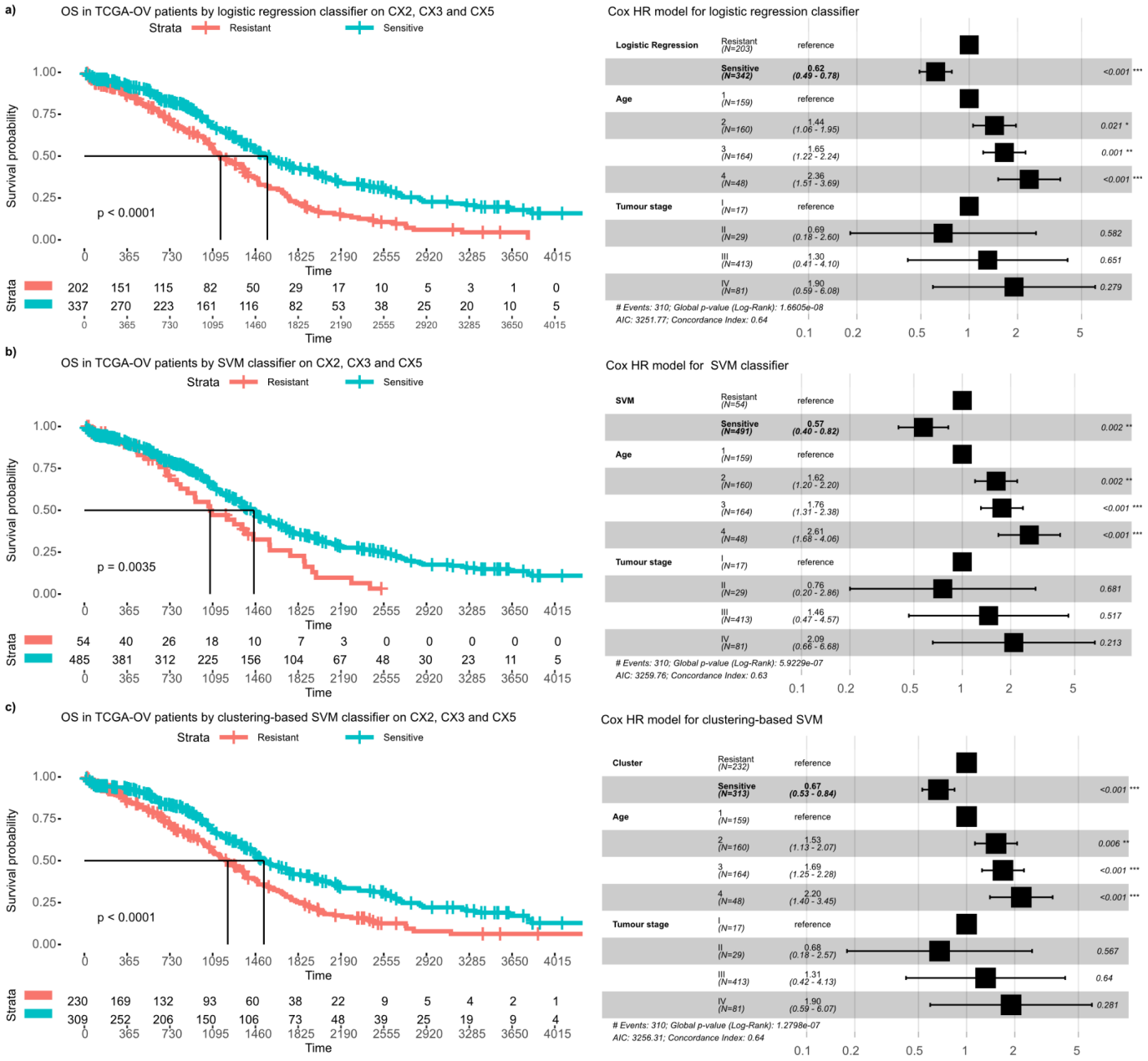
Left column: Univariate survival modelling by a Kaplan-Meier estimate for two PCAWG cohorts: Australian ovarian cancer cohort (OV-AU) and British oesophageal cancer cohort (ESAD-UK). **Right column:** Hazard ratios and their 95% confidence interval obtained from a Cox proportional hazards model correcting for age and tumour stage (multivariate survival modelling). One Cox model for HRDetect excluded tumour stage. **Top row:** Clinical classification predicting sensitivity to platinum-based chemotherapy done by our CX2/CX3 classifier. **Middle top row:** Clinical classification performed by HRDetect. **Middle bottom row:** Clinical classification predicting

sensitivity to platinum-based chemotherapy done by our CX2/CX3 classifier. **Bottom row:** Clinical classification performed by HRDetect.



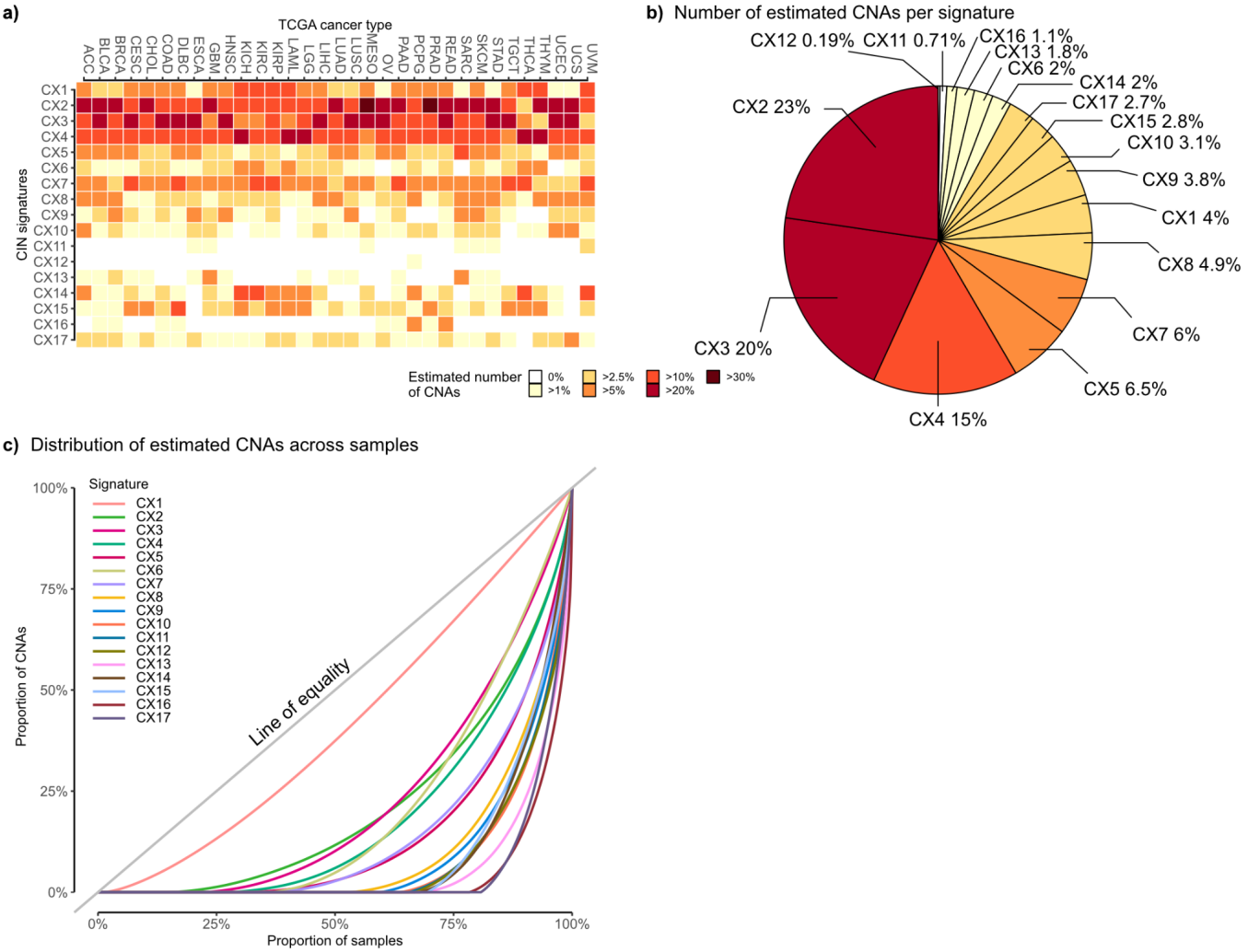
Supplementary Fig. 36: Kaplan-Meier curves and Cox proportional hazards model for ICGC breast cancer cohort

Left column: Univariate survival modelling by a Kaplan-Meier estimate for the ICGC breast cancer cohort. **Right column:** Hazard ratios and their 95% confidence interval obtained from a Cox proportional hazards model correcting for age and tumour stage (multivariate survival modelling). **Top row:** Clinical classification predicting sensitivity to platinum-based chemotherapy done by our CX2/CX3 classifier. **Bottom row:** Clinical classification performed by HRDetect.



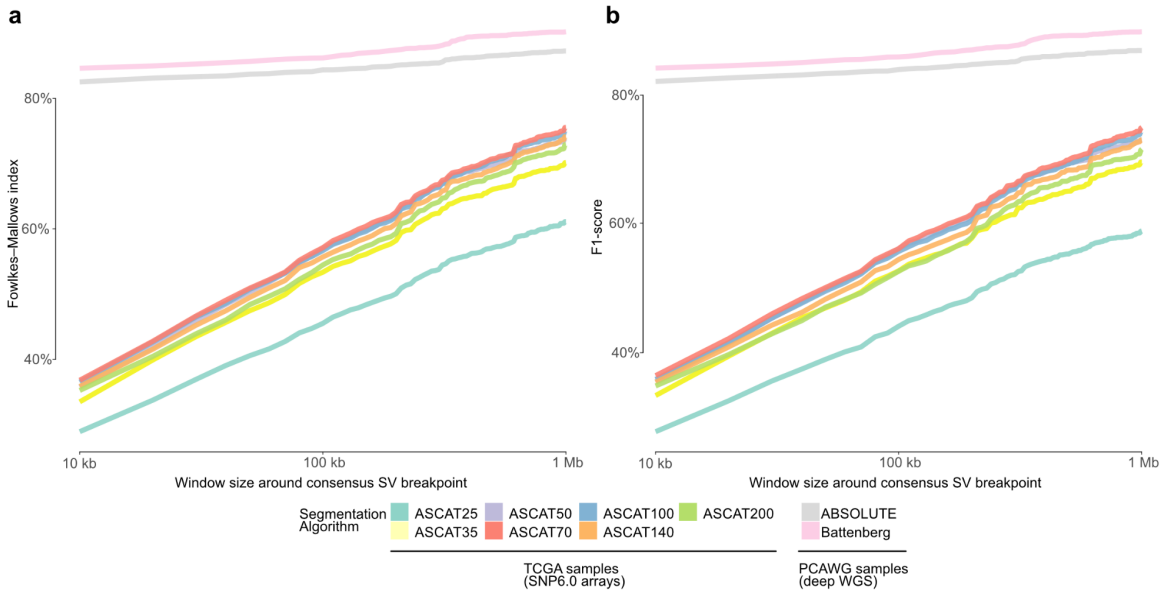
Supplementary Fig. 37: Kaplan-Meier curves and Cox proportional hazards model for three alternative classifiers utilising CX2, CX3 and CX5

Left column: Univariate survival modelling by a Kaplan-Meier estimate for the TCGA ovarian cancer cohort. **Right column:** Hazard ratios and their 95% confidence interval obtained from a Cox proportional hazards model correcting for age and tumour stage (multivariate survival modelling). **Top row:** Clinical classification predicting sensitivity to platinum-based chemotherapy done by a logistic regression model taking all three impaired HR signatures into account. **Middle row:** Clinical classification predicting sensitivity to platinum-based chemotherapy done by support vector machine (SVM). **Bottom row:** Clinical classification predicting sensitivity to platinum-based chemotherapy done by a SVM on clustered samples.



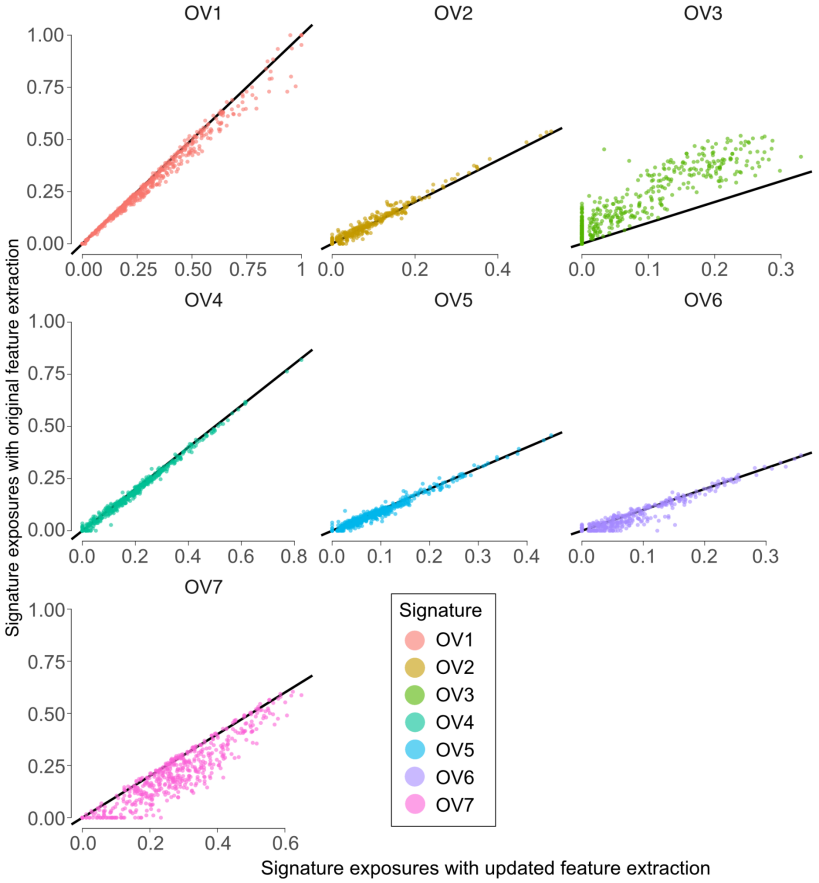
Supplementary Fig. 38: Estimated number of CNAs produced by each signature

a: Heatmap showing estimated number of CNAs produced by the 17 signatures split by the 33 TCGA cancer types. **b:** Pie chart of the estimated number of CNAs produced by the 17 signatures. **c:** Lorenz curve of inequality showing the distribution of estimated number of CNAs produced by each signature across the 6,335 TCGA samples.



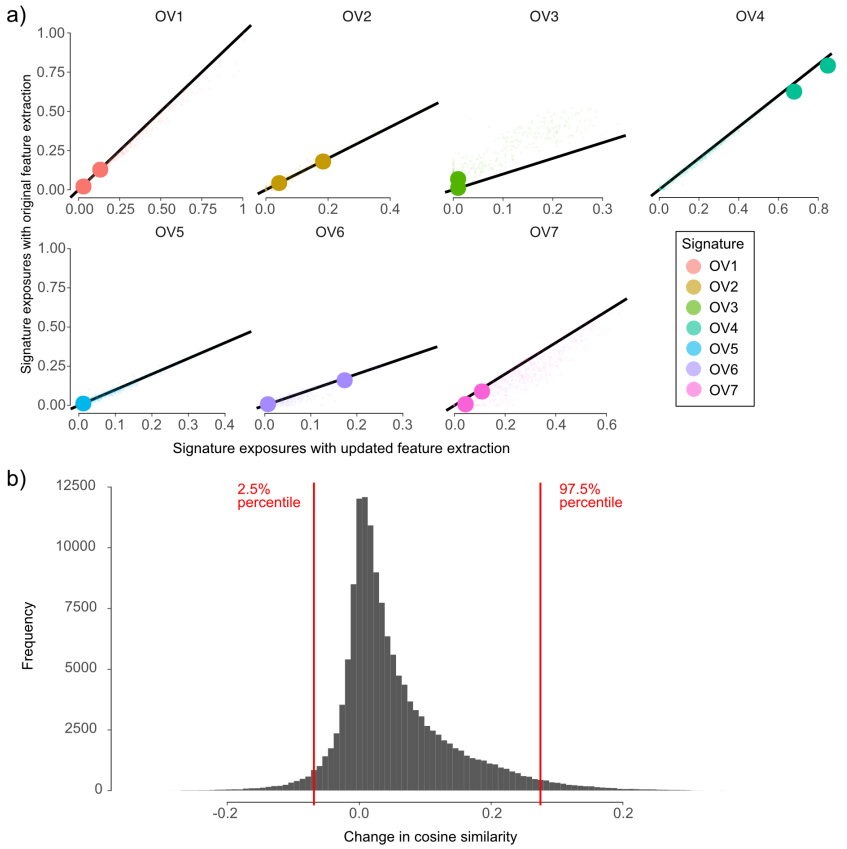
Supplementary Fig. 39: Agreement between segmentations on SNP 6.0 arrays and whole-genome sequencing

Comparison of ASCAT segmentation on SNP 6.0 arrays against the consensus segmentation from PCAWG performed on whole-genome sequenced samples. As breakpoints were unlikely to be on the exact same genomic location, a window was extended on either side of the consensus structural variant-identified breakpoint.



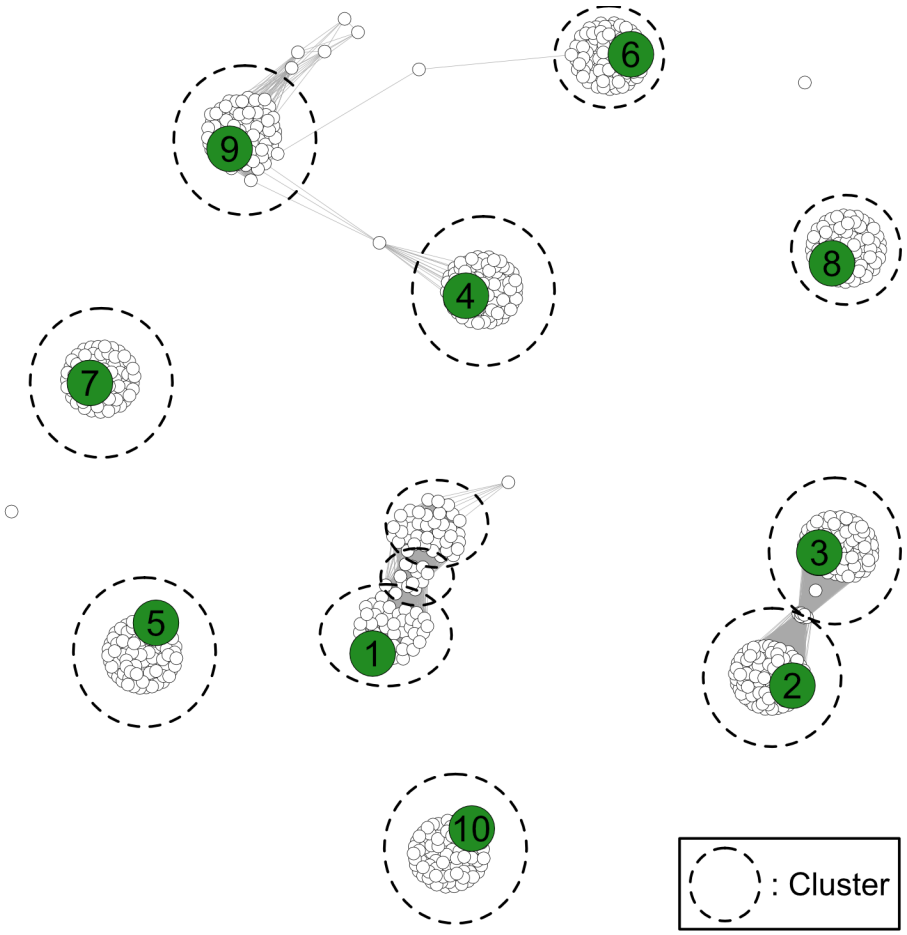
Supplementary Fig. 40: Ovarian signature activities under two feature extraction modalities

Comparing the signature activities (previously called exposures) for the ovarian copy number signatures influenced by two different feature extraction methods: the updated feature extraction used in this study (x-axis) and the original feature extraction method used in Macintyre et al.(y-axis).

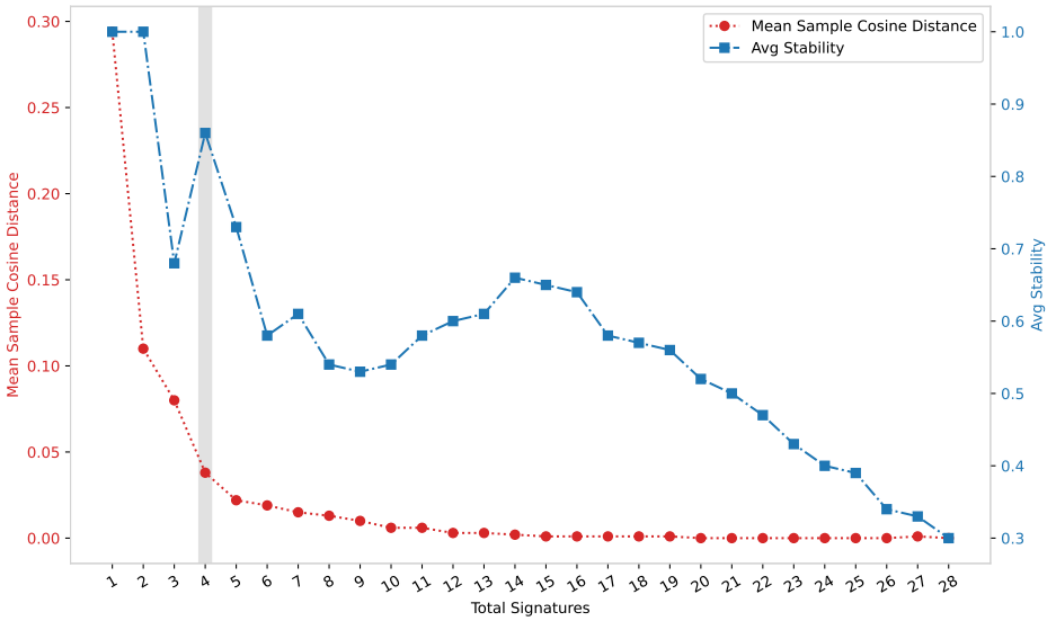


Supplementary Fig. 41: Similarities of ovarian signature activities before and after removing diploid segments

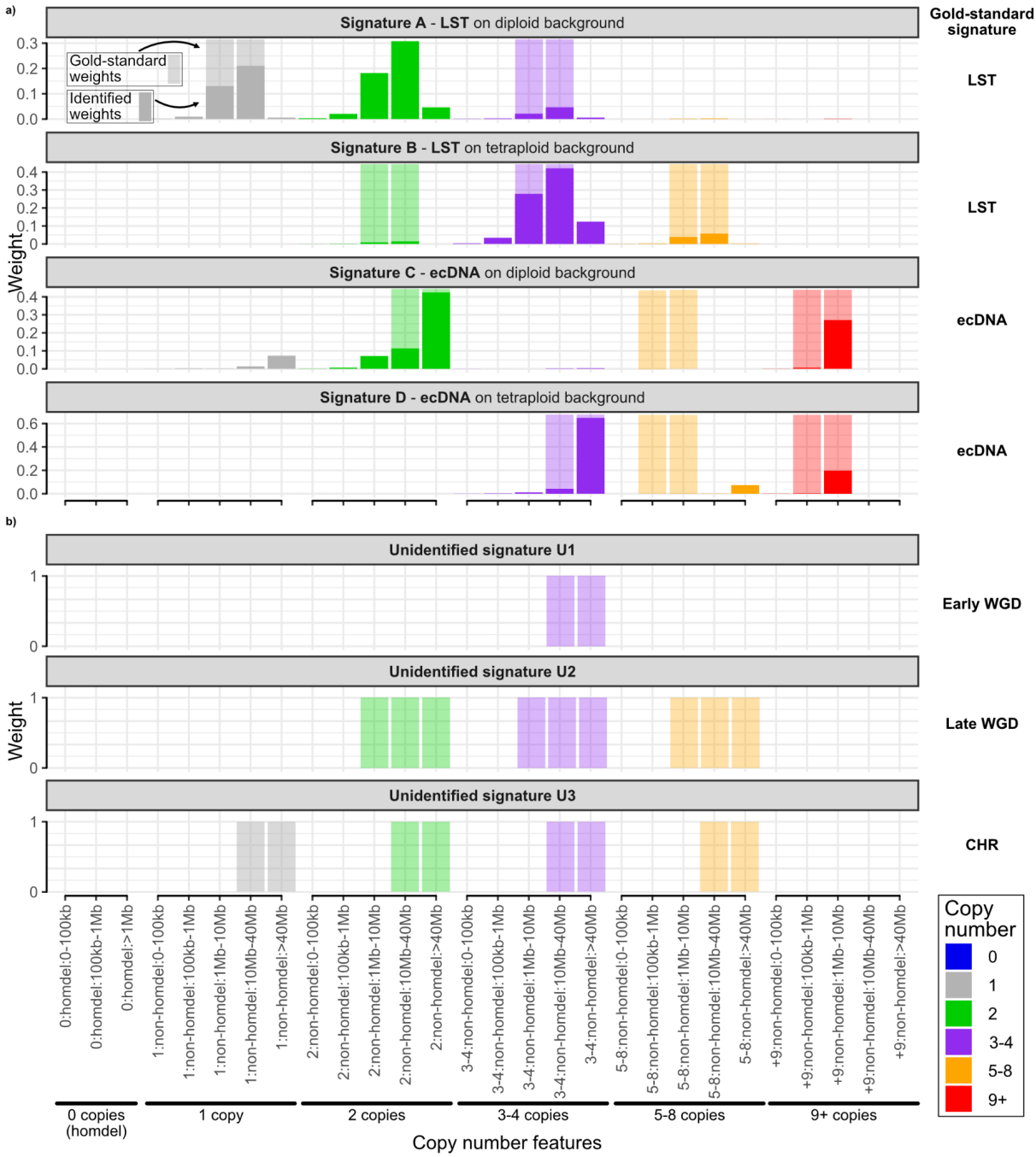
a: Signature activities before (y-axis) and after (x-axis) removing diploid segments with samples TCGA-09-1763 and TCGA-24-2026 highlighted. **b:** Histogram of the changes in cosine similarity before and after removing diploid segments.



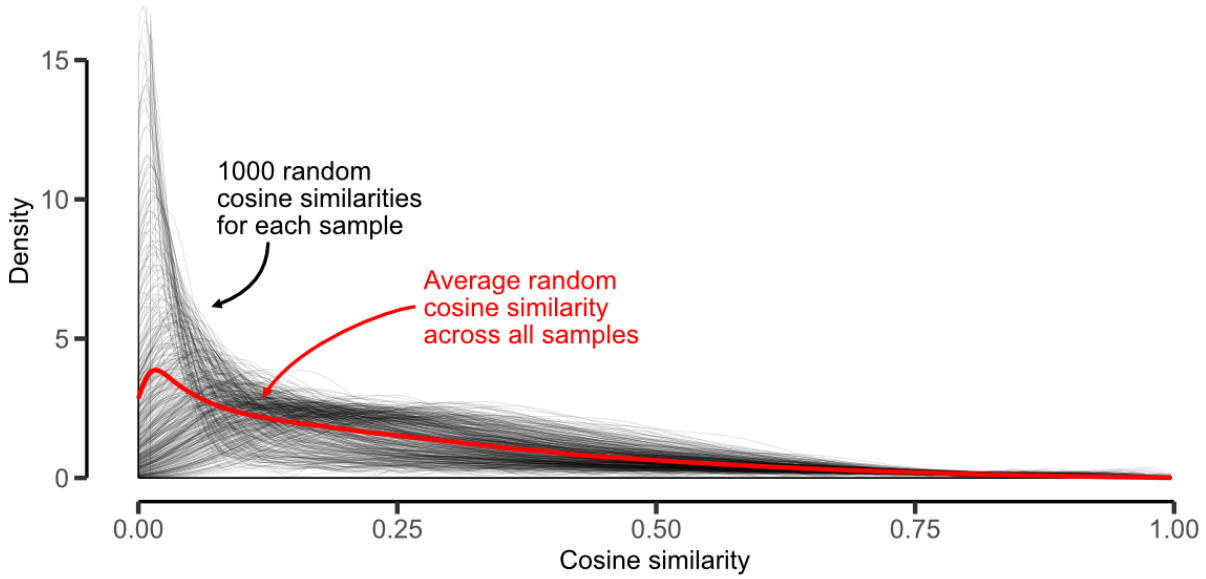
Supplementary Fig. 42: Hairball plot of cosine similarities of 95 sets with K=10 solutions
From 200 BayesNMF runs, most of them settled on K=10 as final K (n=95). This figure shows the cosine similarities between all signatures of the 95 sets. Similarities above 0.85 are shown by lines. A white dot represents a signature. The green dots represent the final ten pan-cancer signatures.



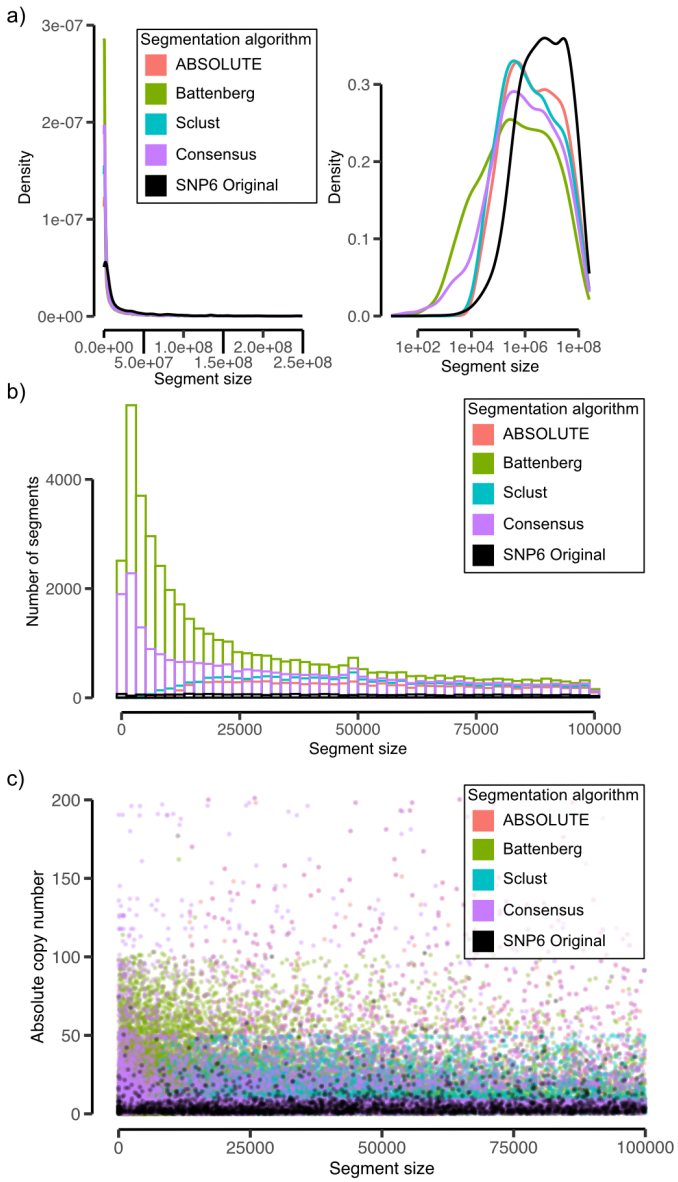
Supplementary Fig. 43: Optimal number of signatures according to SigProfilerExtractor
Plot shows the stability and mean sample cosine distance for different solutions. Grey bar denotes the optimal solution.



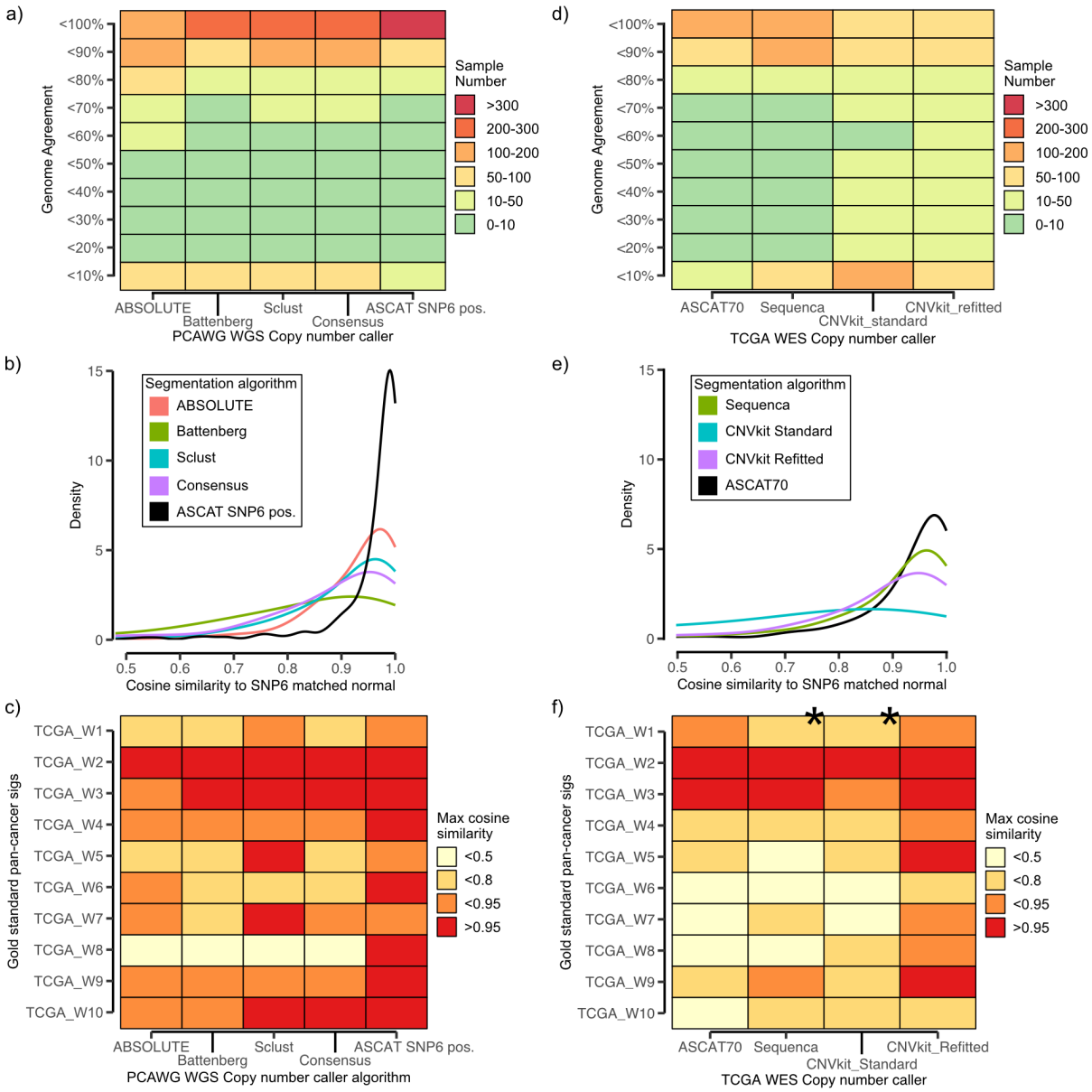
Supplementary Fig. 44: Fig. R7. Barplots showing the proportion of the mutations for each signature identified from simulated genomes using an alternative single-feature approach
Gold-standard signatures of the five mutational processes are shown with shaded bars. Their patterns reflect how the mutational processes were designed and the background ploidy status is taken into account. a) The four signatures identified by the single-feature encoding with SigProfilerExtractor. b) The three signatures not identified by the single-feature encoding method.



Supplementary Fig. 45: Densities of cosine similarities of simulated random signature activities
Black lines depict cosine similarities of 1000 sets of simulated signature activities for each of the 478 samples. Red line is the average across all samples.

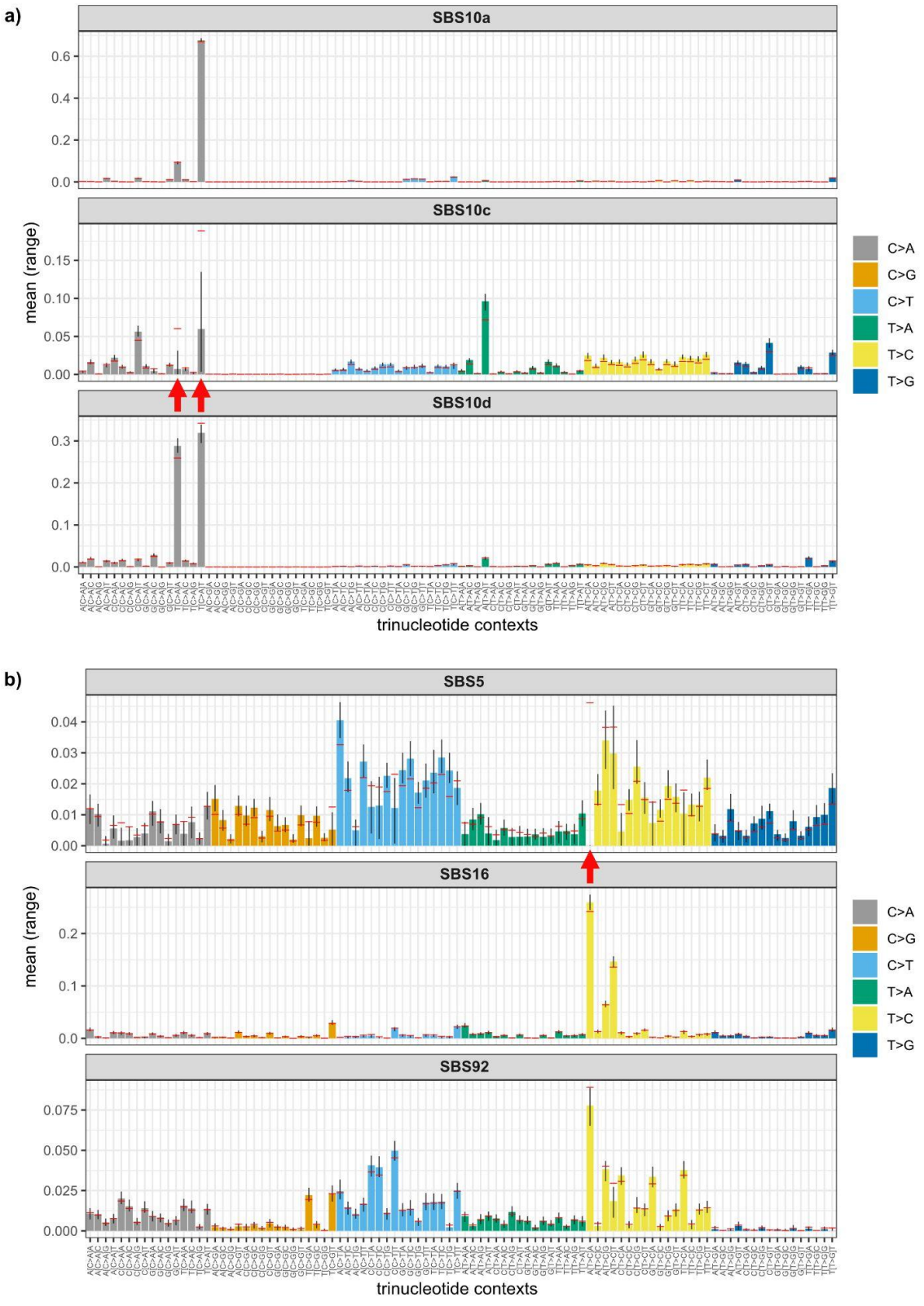


Supplementary Fig. 46: Segment size distributions (<100kb) across technology platforms and algorithms
a: Density plots of the number of copy number segments for 451 samples that were present for all segmentation algorithms. **b:** Histogram of segment sizes for segments smaller than 100kB. **c:** Scatterplot of segments by their segment size (smaller than 100kB) and their absolute copy number for the 451 TCGA/PCAWG samples.

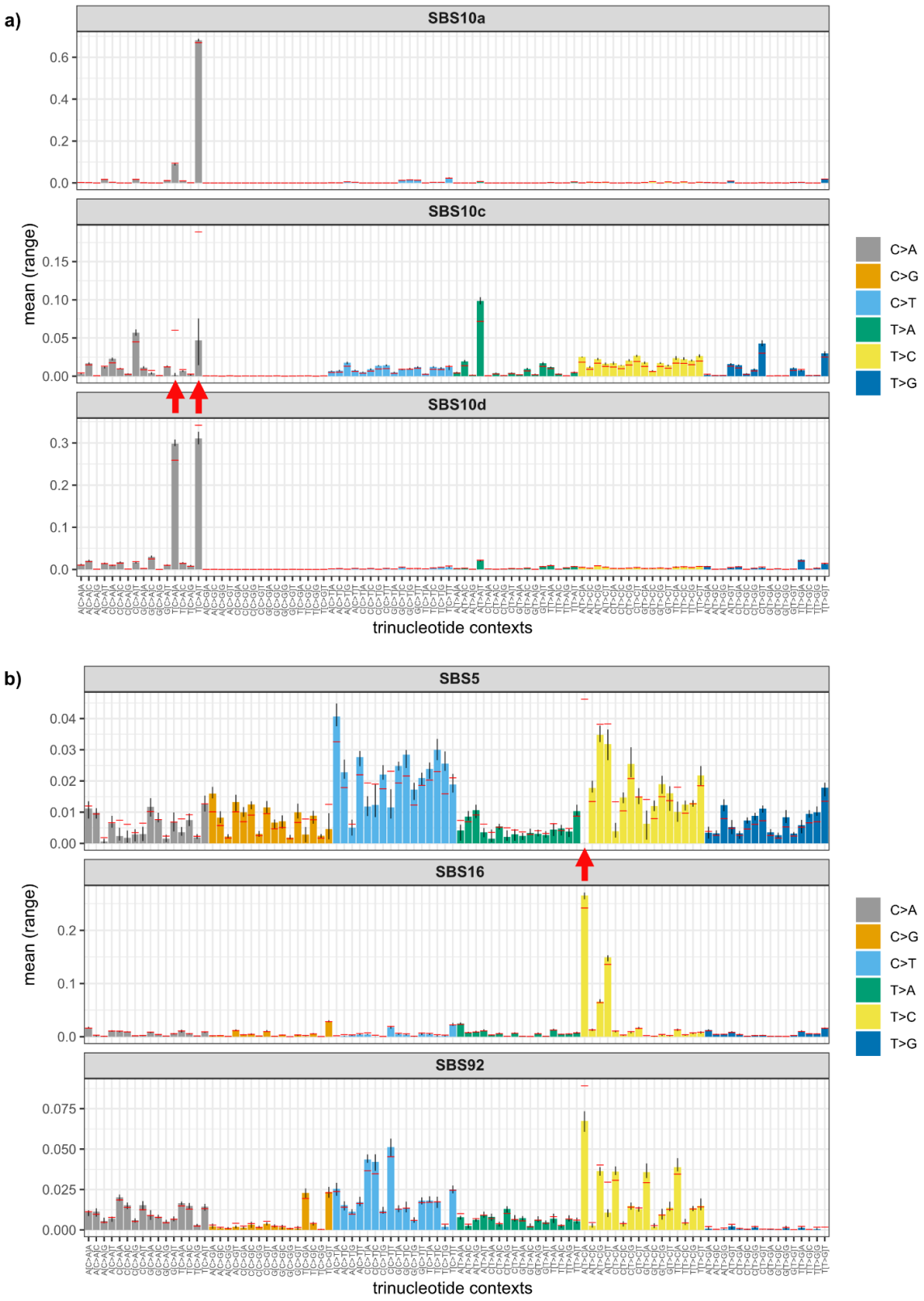


Supplementary Fig. 47: Assessment of signature quantification and identification using whole-genome and whole-exome data

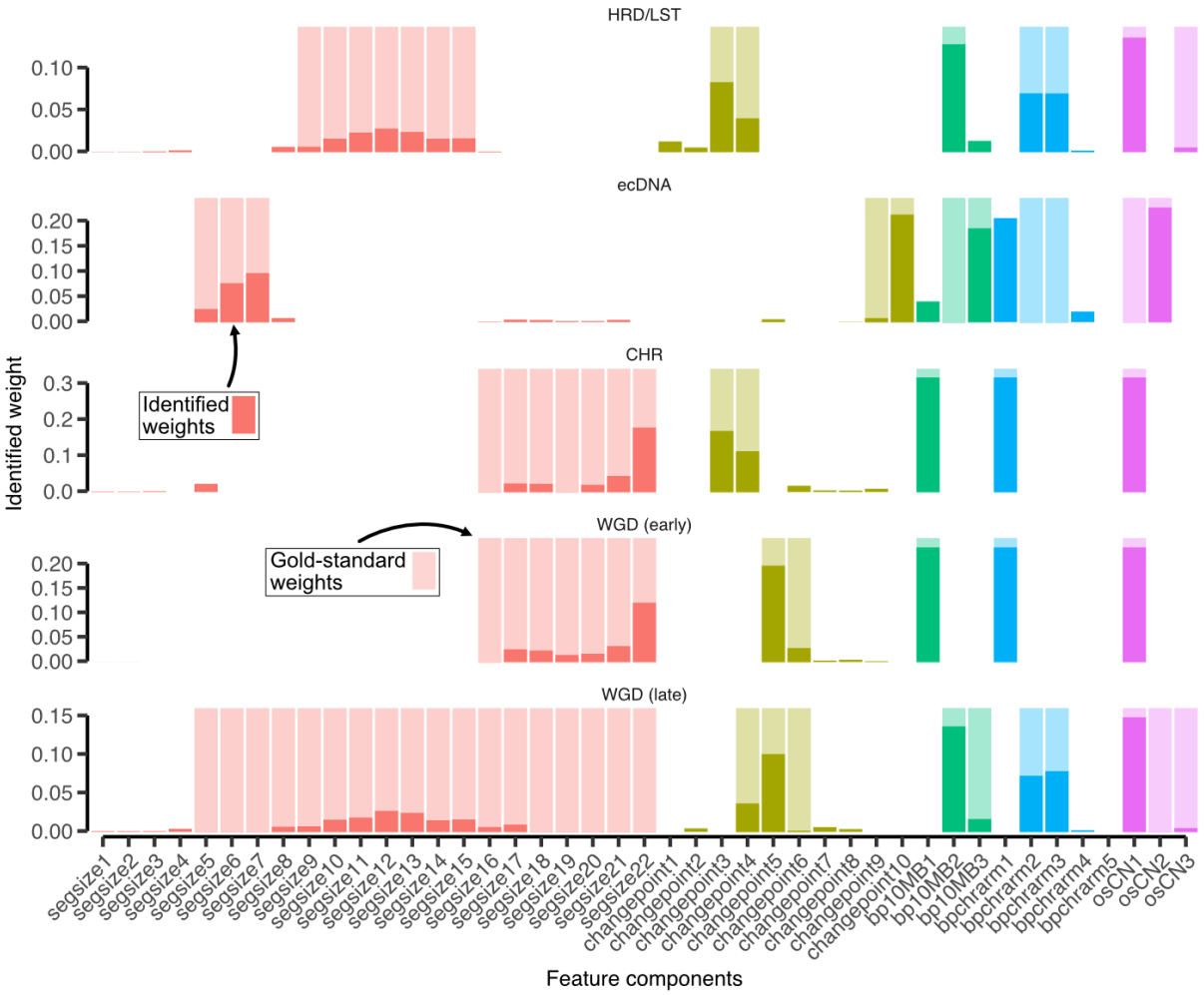
a/d: Heatmap showing goodness of copy number fits on 478 TCGA samples which were WGS (a) or WES (d) sequenced in the PCAWG project and show detectable levels of CIN. **b/e:** Density plots of cosine similarities between SNP6-derived signature activities and WGS-derived (b) or WES-derived (e) activities. **c/f:** Heatmap showing cosine similarities between SNP6-derived signature definitions and WGS-derived (c) or WES-derived (f) signature definitions. (*): For settings marked with a star it was not possible to derive solutions for K=10, instead the optimal number of K was chosen (lower than K=10).



Supplementary Fig. 48: Simulation of SNV signatures shows shrinkage of signature definition weights Coloured bars show the median signature weight obtained by 250 simulations with each including 500 samples. Error bars depict the minimum and maximum values obtained across all 250 simulations. Horizontal red lines show the original component weights and the red arrows depict the components which decreased the strongest compared to the original definitions. The Brunet algorithm was used to decompose signatures. **a:** Combination of SBS10a, SBS10c and SBS10d. **b:** Combination of SBS5, SBS16 and SBS92.

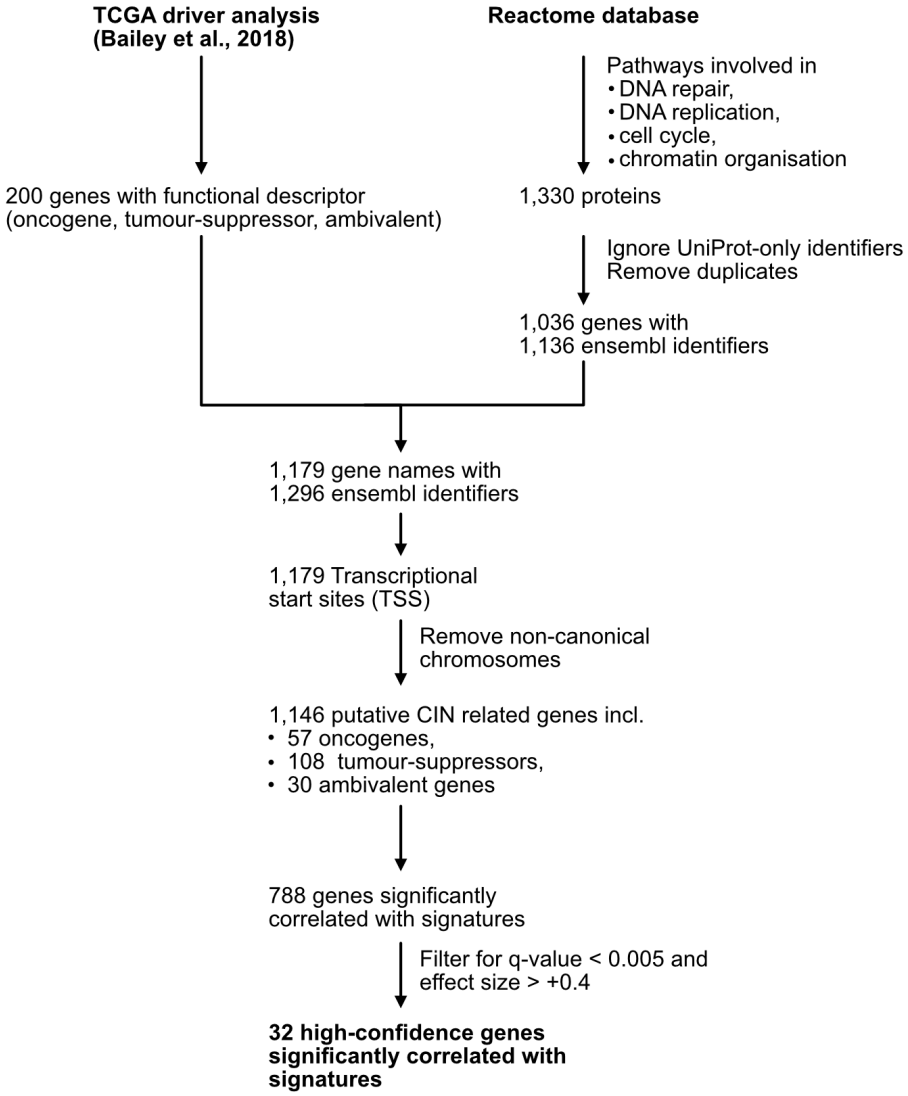


Supplementary Fig. 49: Simulation of SNV signatures shows shrinkage of signature definition weights Coloured bars show the median signature weight obtained by 10 simulations with each including 500 samples. Error bars depict the minimum and maximum values obtained across all 250 simulations. Horizontal red lines show the original component weights and the red arrows depict the components which decreased the strongest compared to the original definitions. The SigProfilerExtractor algorithm was used to decompose signatures. **a:** Combination of SBS10a, SBS10c and SBS10d. **b:** Combination of SBS5, SBS16 and SBS92.

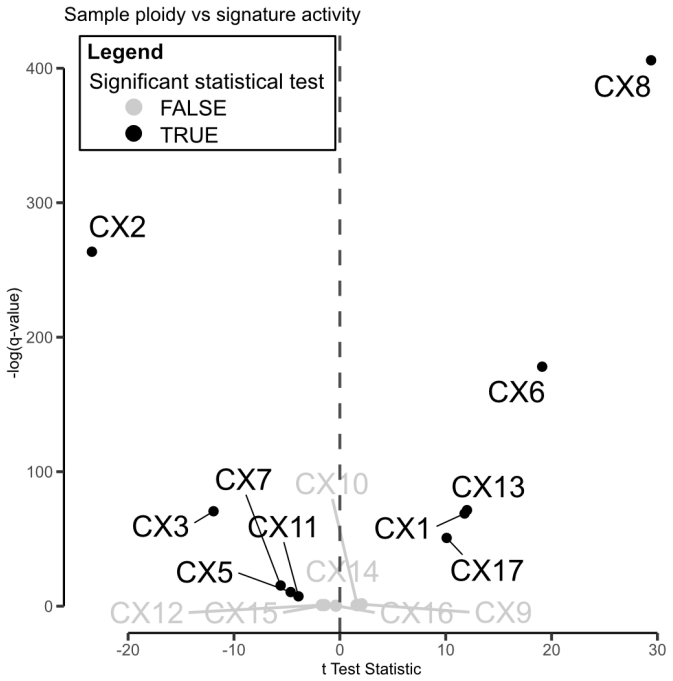


Supplementary Fig. 50: Gold-standard and signatures from the signature interpretation matrix of five mutational processes modelled for the simulation of genomes

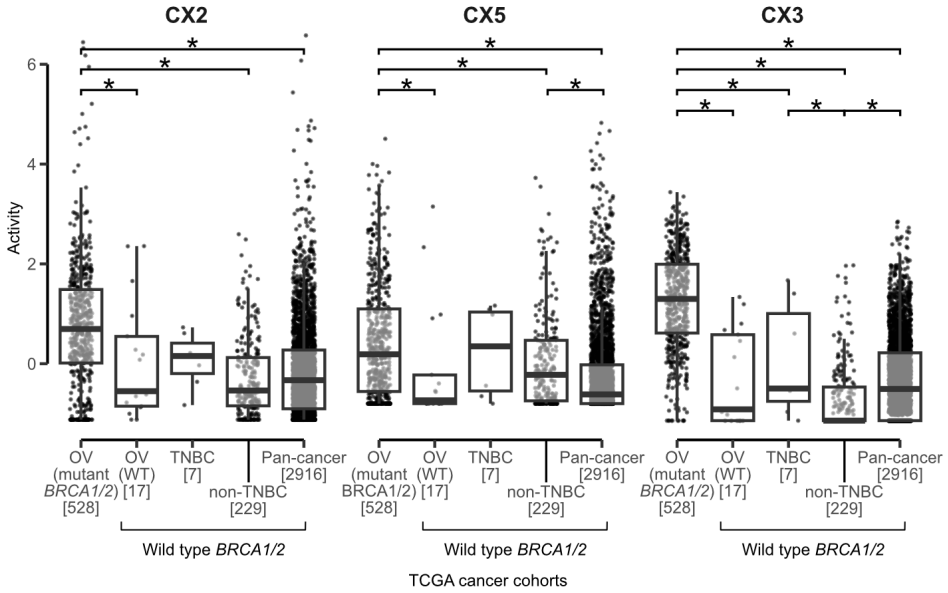
Shaded bars: Components received weight depending on the importance of this component to reflect the mutational process. Fully coloured bars: Identified signatures weights.



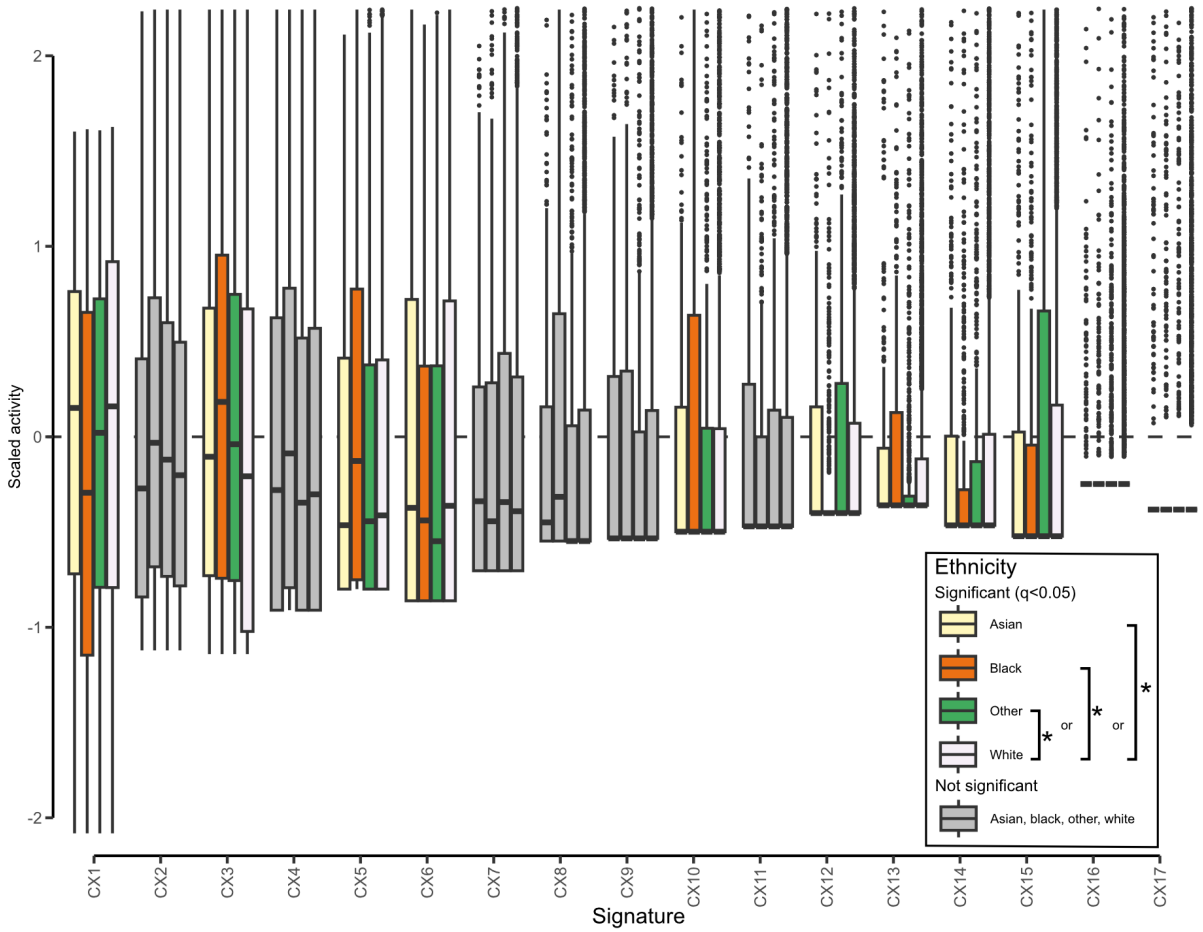
Supplementary Fig. 51: Flowchart depicting how genes were selected and filtered for their associations with the CIN signatures.



Supplementary Fig. 52: Associations of signature activities and sample ploidy
Volcano plot of the t-Test statistic and -log(q-value) from a multivariate regression model testing the association between signatures and sample ploidy by correcting for CX4 activity and whole-genome duplication status. CX4 is therefore omitted.



Supplementary Fig. 53: IHR signatures in different tumour types
Boxplots showing the signature activity of impaired HR signatures (CX2, CX5, CX3, ordered by increased severity) for TCGA samples with wild type BRCA1/2 genes (also without loss of heterozygosity at either loci). Samples from ovarian cancer (OV with and without mutant BRCA1/2), positive triple-negative breast cancers (TNBC) and negative triple-negative breast cancers (non-TNBC) were highlighted. Differences in mean activity was tested with two-sided Welch's t-test and corrected for false discovery rate (q-value < 0.05). Mutant BRCA1/2 samples included germline, somatic and promoter hypermethylation mutants (of BRCA1, BRCA2 or RAD51C) as well as samples with a loss of heterozygosity at either the BRCA1 or BRCA2 loci. Only samples with no known mutation or LOH were treated as true wild type for this analysis. Boxes represent the interquartile range (IQR) with the median as a bolded line. The whiskers extend to the largest/smallest value no further than 1.5 * IQR from the hinge. Outliers beyond the end of the whiskers are marked individually as points.



Supplementary Fig. 54: Associations of signature activities with patient ethnicity

Boxplots comparing signature activity between groups of samples split by their ethnicity. Statistical significance was tested by two-sided Welch's t-test and coloured when there were significant differences between any of the three comparisons with the white patient population (q-value <0.001). Boxes represent the interquartile range (IQR) with the median as a bolded line. The whiskers extend to the largest/smallest value no further than 1.5 * IQR from the hinge. Outliers beyond the end of the whiskers are marked individually as points.