doi:10.1038/nature10983

Supplementary Information for Curtis *et al.*: The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups

Table of Contents

Supplementary Methods

1 Biospecimen collection and ethical consent

We assembled a collection of over 2,000 clinically annotated primary fresh frozen breast cancer specimens and a subset of normals that passed initial selection criteria from tumour banks in the UK and Canada (for which a subset of 2,136 expression arrays and 2,477 Affymetrix SNP 6.0 arrays are reported on here). Both primary breast tumours, with linked pseudo-anonymised clinical data, and normal breast tissue were obtained with appropriate ethical approval from the relevant institutional review board. The METABRIC study protocol, detailing the molecular profiling methodology, was also approved by the ethics committees in Cambridge and Vancouver, the two sites responsible for the molecular analysis of the samples. As described below, tumours were primary invasive breast carcinomas for which clinical information could be categorically linked to DNA and RNA specimens. Initial quality control involved assessment of array quality and the flagging of mismatches between DNA and RNA using a novel eQTL-based approach (outlined below). Following the exclusion of cases on histopathological grounds (benign cases, those with ductal or lobular carcinoma in situ (DCIS, LCIS), or low tumour cellularity), incomplete clinical/pathological data (absence of ER status, grade, or tumour size), or apparently related individuals (based on genotype calls), paired DNA and RNA profiles were available from tumours derived from 997 female patients. DNA from adjacent normal breast tissue (or peripheral blood for some cases) was available from 485 samples, a subset of which match tumours in the discovery set. High quality RNA derived from adjacent normal tissue was available from 144 samples. The demographics, clinical, and pathological characteristics of the patients described in the discovery set are presented in Tables S1 and S2. A second cohort of 995 cases was later assembled for which matched DNA or RNA profiles or clinical information was not available at the time of the initial analyses, and these included low cellularity tumours, DCIS, and three benign cases. This cohort represents a validation set, which was employed for the purposes of testing the reproducibility of the integrative clusters and clinical outcome associations. Genotype analysis for the full set of cases subsequently revealed that eight individuals were represented both in the discovery and validation set (MB: 0667/0025, 0546/0326, 0327/0547, 0549/0329, 0559/0335, 0573/0355, 0408/0407, 0432/0433), and four were represented twice in the validation set (MB: 0110/0196, 0552/0330, 6213/6206, 2820/2720), but were supplied as unique accessions by the tumour bank. These sample pairs represent multiple primary tumours from the same individual and different sections of the same tumour. These comprise only 1.2% of samples in the validation set and do not alter the conclusions from these analyses. The demographics, clinical, and pathological characteristics of the patients described in the validation cohort are presented in Tables S1 and S3.

2 Histopathological review

The frozen tissue sections from which nucleic acids were isolated were subject to expert histopathological review to assess the presence of invasive tumour, pre-malignant or benign changes, tumour cellularity, and lymphocytic infiltration in specific subgroups (Table S45). Tumour cellularity was scored visually in a semiquantitative fashion on three sections (taken at the beginning, middle, and end of cryosectioning) per tumour for UK samples and two sections (taken at the beginning and end of cryosectioning) per tumour for Canadian samples, where cellularity values were binned so that ‘low cellularity’ corresponds to samples with <40% tumour DNA, ‘moderate cellularity’ corresponds to 40% - 70% tumour DNA, and samples with >70% tumour DNA were considered to have ‘high cellularity’. As noted above, samples that were classified as ductal or lobular carcinoma *in situ* (DCIS, LCIS) or as benign were excluded from the discovery set.

3 Experimental assays and orthogonal validation

3.1 *Microarray data description*

Matched DNA and RNA were extracted from each specimen and subject to copy number and genotype analysis on the Affymetrix SNP 6.0 platform and transcriptional profiling on the Illumina HT-12 v3 platform (Illumina_Human_WG-v3). A flow chart depicting the main analytical approaches that utilise these data is presented in Figure S1 (supporting files are available at: http://www.compbio.group.cam.ac.uk/resources.html). The associated genotype and expression data have been deposited at the European Genome-Phenome Archive (EGA, http://www.ebi.ac.uk/ega/), which is hosted by the European Bioinformatics Institute, under accession number EGAS00000000083.

3.2 *Nucleic acid isolation and quality assessment*

For UK samples, DNA and RNA were extracted from ten 30 μm sections each from fresh frozen tumours using the DNeasy Blood and Tissue Kit and the miRNeasy Kit (Qiagen, Crawley, UK) on the QIAcube (Qiagen) according to the manufacturer's instructions. For Canadian samples, DNA and RNA were extracted from 10 eight μm sections each from fresh frozen tumours using the MagAttract DNA Mini M48 Kit and miRNeasy 96 Kit (Qiagen) manually in a 96-well format according to the manufacturer's instructions. Nucleic acids were quantified with a NanoDrop ND-8000 spectrophotometer (NanoDrop Technologies, Wilmington, DE, USA) and quality assessed by agarose gel electrophoresis. RNA quality was also assessed using the Agilent 2100 Bioanalyser Nanochip (Agilent Technologies, Wokingham, UK). Tumour samples for which the RNA had an RNA Integrity Number (RIN) > 7 were hybridized to expression arrays, whereas a less stringent RIN > 5 was required for normal RNA. Random DNA and RNA samples were selected for genotyping purposes to ensure sample uniqueness using the AmpF *l* STRIdentifiler PCR Amplification Kit (Applied Biosystems, Foster City, CA, USA).

3.3 *DNA labelling and microarray hybridisations*

DNA was hybridized to Affymetrix SNP 6.0 arrays per the manufacturer's instructions (Affymetrix, Santa Clara, CA) at AROS Applied Biotechnology (Aarhus, Denmark). Samples that met the quality control criteria established by AROS and suggested by the Affymetrix Genotyping Console v2.1 were subject to further in-house quality assessment (as described below).

3.4 *RNA labelling and microarray hybridisations*

Total RNA was used to generate biotin-labelled cRNA using the Illumina Totalprep RNA amplification kit (Ambion, Warrington, UK) and hybridized onto Illumina Human HT-12 v3 Expression Beadchips per the manufacturer's instructions and was scanned on the Illumina BeadArray Reader.

3.5 *Multiple Ligation-Dependent Probe Amplification (MPLA) assay*

Multiplex ligation-dependent probe amplification (MLPA) was used to validate copy number changes observed in the SNP-CGH data. Two commercially available kits were used, namely, the SALSA MLPA P005 Human Chromosomal Aberration-1 panel and the SALSA MLPA P078 Breast tumour panel. Assays were performed according to the manufacturer's instructions (MRC-Holland, Amsterdam, the Netherlands) on 40 tumours.

3.6 *Long-range PCR validation of germline CNVs and somatic homozygous deletions*

Primers were designed for copy number neutral regions adjacent to the homozygous deletion using Primer 3 software (http://frodo.wi.mit.edu/primer3) (Table S10). PCR reactions were performed using the SequalPrep Long PCR kit and Enhancer B Buffer according to the manufacturer's instructions (Invitrogen, Paisley, UK). Prior to sequencing, LR-PCR products were confirmed by gel electrophoresis. Several CNVs (and somatic homozygous deletions) were validated to nucleotide resolution in DNA from the tumours and/or MN tissue via PCR and sequencing (Table S9) as illustrated for a few exemplary cases in Figure S3.

3.7 *TP53 sequencing and analysis*

As *TP53* mutations also contribute to GI, we surveyed the TP53 mutational spectrum of 820 cases from our cohort by Sanger sequencing (Table S2). Mutations in the 11 TP53 exons were examined as previously described [1] (with the exception that exon 7 was sequenced in only one direction). Mutations were scored using Mutation Surveyor software (v3.23, SoftGenetics, LLC, State College, PA, USA). To increase the mutation detection sensitivity, we applied the following modifications to the default settings: 1.) Data were analysed with unidirectional parameters instead of bidirectional parameters. This allows the detection of mutations in areas of high noise; 2.) The "Check 2D Small Peaks (Mosaic)" box was checked. This option enables the software to display potential somatic mutations which often have low intensity; 3.) The 'dropping factor' was decreased from 0.2 to 0.1. The dropping factor measures the level of decrease in the sample's normal allele. By decreasing the dropping factor from 0.2 to 0.1, the normal allele concentration can drop less before a mutation would be detected. 4.) The Mutation Height was decreased from 500 to 300. All the mutations called by Mutation Surveyor were inspected visually to eliminate false positives and false negatives, but note that this automated method is likely to under call frameshift mutations and thus the mutation frequency reported is likely to be an underestimate. Splice site mutations were not included, but would result in a negligible increase in the number of affected cases (2 cases). Sequences that failed in one direction were scored as unidirectional. Mutations were then manually confirmed using Lasergene's SeqMan Pro (DNASTAR, Inc. Madison, WI, USA). In total, 18,075 Mutation Surveyor reports covering 820/997 cases were parsed and projected onto genomic coordinates using RefSeq accession NC_000017 region 7512445-7531642. All variants were filtered against dbSNP (v129) and amino acid annotations were obtained using MutationAssessor (http://mutationassessor.org). All mutations required support from both directions to be called. A small subset of the mutations were validated by a second PCR reaction and Sanger sequencing. The overall frequency of *TP53* mutation was ~12% (n = 99; 69 missense, 30 truncating) with 84% of mutations present in histologically high-grade (Grade 3) tumours predominantly of the Basal-like subgroup (34% of cases, n = 33) (Figure S5). As such, the genome-wide copy number profiles of *TP53* mutated cases reflect the distinct pattern of the Basal-like tumours (Figures S32, S33), as has been noted by others [2].

4 Gene expression analysis

4.1 *Data pre-processing and normalisation*

A custom R [3] script was written to process each BeadChip once scanning was complete and raw data were available. The script relied on existing functionality within the *beadarray* package [4], including the generation of quality assessment (QA) information and adjustment for spatial artefacts with the BASH tool [5]. Once all 12 arrays on the chip had been processed, the bead-level data were summarized, yielding a series of 48,803 × 12 matrices of $\log_2$ intensities, standard errors, and number of observations. After all chips had been processed, the summarized matrices were combined.

Potential outlier arrays were removed by considering the bead-level QA scores derived using the control probes on each array. We resisted the temptation to use methods that compare array intensities and distributions between arrays to allow for the possibility that these represent biologically interesting differences. Arrays with a P95 scan metric less than 200 were excluded as they were considered to have failed hybridisation. These arrays were then removed from the dataset so as to not influence the following QA. A multivariate outlier testing procedure [6] implemented in the *arrayMvout* Bioconductor package was used to identify arrays with poor quality based on the bead-level QA scores of all arrays. Additionally, this approach was applied to bead-level QA information for each location separately. All arrays that remained after this three-step procedure were retained in the subsequent analysis.

We relied on the comprehensive re-annotation of the Illumina HT-12 v3 platform [7] as we have noticed the sub-optimal performance of certain probes, and therefore wanted to avoid the influence of such probes on the normalisation. To do so, we first generated a list of suitable probes based on the following criteria: must be a perfect genomic match, does not target the sex chromosomes, must not contain a SNP, must be $3'$ matching and without multiple genomic matches, must not contain a polyG tail at the end of probe sequence, GC content must be between 38% and 64%, and the probe must not target genes from the PAM50 list. As we did not note any differences in the target distributions from different tumour banks, a single target distribution was generated using the list of suitable probes. ER-positive and ER-negative samples were quantile normalised separately and averaged to obtain the final target distribution. Each array was then normalised to the target by quantile normalising probes belonging to the target distribution, while the values for the remaining probes were obtained by interpolation using the weighted normalised intensities of the target distribution probes with most similar intensities prior to normalisation. A linear model was then fit using the *limma* [8] Bioconductor package to remove any batch effects associated with the position of an array on the Illumina BeadChip.

Differential expression analysis for contrasts of interest was performed on a subset of the $\log_2$ normalised intensity matrix that excluded any probes mapping to non-transcribed regions. This approach has recently been shown to improve the sensitivity of a differential expression analysis [7]. A probe-wise linear model was fit to the data using *limma* with coefficients estimated for particular clinical variables of interest (for example, ER-positive or ER-negative cases). Differences in expression level arising from characteristics of the different tumour banks were accounted for by incorporating a coefficient for tumour-bank in the linear model. Furthermore, a weight matrix, defined to be the square root of the number of observations on each array, was used to down-weight arrays with fewer observations in the model. After empirical Bayes' moderation of the variance, adjusted *P*-values, log-ratios, and log-odds were used to assess the evidence for differential expression in each contrast of interest.

4.2 *PAM50 classification*

Samples were classified into the five intrinsic subtypes [9, 10] based on PAM50 [11]. As we have previously noted probe annotation to be an important consideration in sample classification using microarray data [12], the PAM50 gene-list was refined so that only genes for which a corresponding probe with perfect annotation on the Illumina HT-12 v3 BeadChip [7] were used for classification. As a result, *BAG1* and *TMEM45B* were not included in the classification. For genes with more than one probe, probes were selected on the basis of their annotation. For example, probes containing a SNP were avoided as were those which target introns, have secondary targets or mismatches, lie in repeat-masked regions, or target the wrong genomic strand. As previously recommended [13], all probes were median centred prior to classification. Due to the imbalance in ER status, we defined 100 random reference distributions consisting of all ER-negative samples and randomly selected ER-positive samples during the median centring step. This resulted in 100 different classifications and the final subtype calls were derived by taking a consensus across all 100 trials. Samples were then assigned to one of the five intrinsic subtypes using the Spearman correlation to the published centroids and the transformed intensities, where samples with correlations < 0.1 for all subtypes were not classified (NC) (0 samples in the discovery set, 6 samples in the validation set).

4.3 *GENIUS classification*

Samples were classified into the four Gene Expression progNostic Index Using Subtypes (GENIUS) subgroups (HER2+, ER-/HER2-, ER+/HER2- high proliferation, ER+/HER2- low proliferation) [14] using the *genefu* [15] package in Bioconductor.

4.4 *Expression-based classification*

ER, PR, and HER2 status were classified based on their empirical expression distributions using a 2-component Gaussian mixture model as implemented in *MCLUST* [16, 17] in an approach similar to that described by Lehmann *et al.* [18].

5 Genotype calling

Affymetrix SNP 6.0 arrays were pre-processed and genotyped using the SNP-RMA algorithm [19], available in the *crlmm* [20] Bioconductor package. The goal of pre-processing is to convert raw intensity values into quantities proportional to the amount of DNA in the target sample associated with each of the alleles, A and B, for each SNP. Briefly, feature intensities were corrected for fragment length and sequence effects, followed by quantile normalisation to a predefined reference distribution. Intensities were then summarized by median polish, resulting in a single value for each allele. A mixture model was then used to adjust for remaining fragment length and intensity-dependent biases on the log-ratio of the summarized intensities. Samples with a signal-to-noise ratio < 5 were flagged in downstream analyses.

6 Inference of sample ethnicity

Genotype calls for each sample were projected onto the HapMap principal component loadings for Affymetrix in order to infer sample ethnicity. The HapMap PCA loadings for Affymetrix SNP probes (snpload.aff.map) were downloaded from: http://www.stats.ox.ac.uk/ davison/software/shellfish/shellfish.php and read into R. Probe annotations and allele frequencies for each of the four HapMap populations (CEPH (CEU); European, Yoruban (YRI); African, Han Chinese (CHB), and Japanese (JPT)) were extracted from the Affymetrix GenomeWideSNP_6.na28.annot.csv file. In total, 168,984 SNPs in the PCA loadings file map to the SNP 6.0 platform and were used in this analysis. For all SNPs in the loading file, the positive strand was taken as a reference. Assuming a coding of 0, 1, and 2 for the SNPs and that they occur in Hardy-Weinberg equilibrium at a ratio of p^2:$2p(1$-$p)$:$(1$-$p)^2$, the relevant allele proportions were extracted and a population score computed for each SNP. Multiplying these scores by the PCA loadings yielded the principal component scores and for each individual, the relevant entries were again multiplied by the PCA loadings. Plotting these data revealed that the two Asian populations were largely overlapping. As such, three main clusters were defined (Africa, Asia, and Europe), and the distance from the cluster to each individual was calculated. Ethnicity calls were made based on the minimal distance between a given sample and population cluster (Figure S7, Table S16).

7 Confirmation of sample identity

We employed a novel eQTL-like approach to ensure that no plating errors had occurred, and that when we claim that the same sample was run on the expression and genotyping platforms we were indeed mapping between the two correctly. This approach we have termed BeadArray Diagnostic for Genotype and Expression Relationships (BADGER) (Lynch *et al.*, manuscript in preparation). Briefly, a panel of 383 strong (-$\log_{10}$ P-value > 15) relationships between measurements of log-expression and reported genotypes was collated from the data. Many of these will be technical associations (e.g. the expression probe covers a SNP), and not genuine eQTLs, but this is irrelevant for this purpose. For each expression array, using this panel and these relationships, a predicted genotype profile was generated. The predicted profile was then compared to each observed profile for those 383 SNPs to ensure that the mappings were correct. BADGER revealed a small number of plating errors that have now been corrected. For all but two of the samples, the match between predicted and observed genotype was exceptional, with the remaining discrepancies probably due to mixture of relatively noisy expression data and possibly the non-diploid nature of the tumour DNA.

8 Segmentation and copy number alteration calling using CBS

Tumour sample copy number $\log_2$ ratios were segmented with two analytical methods, namely circular binary segmentation (CBS) [21] and an adapted hidden Markov model (HMM) [22] to provide contrasts between sensitivity and specificity (see *Comparison of segmentation methods based on MPLA*), and we present the results based on both methods. The maintext figures are based on the CBS-derived copy number profiles with the exception of the outlier expression analysis (Figure 2). Where applicable, the HMM-based results are presented in the Supplementary Information. As outlined below the copy number data were rigorously validated on a subset of cases using several independent techniques (see Methods, Figure S3, Table S9, S10), and allowed for estimation of the Type I and Type II error rates for both segmentation methods (Figure S4, Table S11).

8.1 *Normalisation of intensities*

Affymetrix SNP 6.0 arrays were pre-processed for copy number segmentation using *aroma.affymetrix*. Both tumour and normal samples were independently normalised using the single array method referred to as Copy-number estimation using Robust Multichip Analysis (CRMAv2) [23], as was a publicly available SNP 6.0 dataset consisting of 270 HapMap individuals. We applied the default settings using the following tags: ACC,ra,-XY,BPN,-XY,AVG,A+B,FLN,-XY. For each sample, allelic-crosstalk calibration, probe sequence effects normalisation, probe-level summarisation, and PCR fragment length normalisation were performed in order to obtain $\log_2$ intensity values for total copy number estimation. The following annotation files were used (Feb 14, 2008):

- Chip definition file (GenomeWideSNP_6,Full.cdf)
- Unit fragment-length (GenomeWideSNP_6,Full,na24,HB20080214.ufl)
- Unit genome position (GenomeWideSNP_6,Full,na24,HB20080214.ugp)

Following these steps, probes were sorted according to their genomic position, replicate probes were summarized by their median value, and missing values (produced by negative intensities in the normalisation) were imputed using the loess procedure included in the *snapCGH* [24] Bioconductor package. Various QC metrics were employed to assess the SNP 6.0 copy number data, including the normalised unscaled standard error, relative log expression, and signal-to-noise ratio of the $\log_2$ intensity data. Samples for which any of these metrics were flagged were visually inspected in order to determine whether to retain them in the analysis. Two pooled references were created, one using the median intensities across the HapMap individuals and another for the normals and tumours, using the median intensity values from a set of 473 normals. Different pools based on gender were created for the X chromosome. Next, $\log_2$ ratios were generated for the HapMap samples by subtracting the pooled value from the $\log_2$ intensities. Similarly, $\log_2$ ratios were obtained for the 473 normals using the corresponding pool. For the 997 tumour samples, two data sets were produced: one using the normal pool as the reference for all the tumours and another using the matched normal for each tumour when available, and the normal pool for the remainder. A similar approach was taken for the validation set. Following genotype analysis, we noted that a small number (13) of the normals were not correctly matched to tumours in the validation set, but this would have a nominal impact on the copy number calls due to the strict criteria for calling common CNVs (outlined below). The HapMap and normal datasets were employed to estimate the frequency of germline CNVs in the cohort, while the tumour samples were used for estimating somatic CNAs. After computing the $\log_2$ ratios for each probe, samples were segmented using the circular binary segmentation (CBS) algorithm [21] implemented in the *DNAcopy* [25] Bioconductor package and alterations were called as described below.

8.2 *Calling of copy number alterations and adjustment for tumour cellularity*

DNAcopy with prior smoothing and default parameters was run on the set of 473 normals and 270 HapMap samples. Although Ostrovnaya *et al.* [26] noted that not smoothing facilitated the detection of CNVs, we found that this led to higher error rates (false positives). For calling alterations, thresholds were set based on the median of the $\log_2$ ratio for each array $+2\sigma$ or -2.5σ (where σ is the standard deviation of the $\log_2$ ratio for each array) for calling gains and losses, respectively. These asymmetric thresholds were based on the assumption that the expected $\log_2$ ratio for a single copy gain is $\log_2(3/2)$, which is smaller in absolute value than the expected $\log_2$ ratio for a single copy loss ($\log_2(1/2)$). This was also confirmed by an independent analysis of the MLPA data (see below).

In order to discover common CNVs, at least five samples in a particular population were required to exhibit an alteration in the same probe. We did not put any restrictions on the length or number of probes required to call a CNV, although implicitly there is a restriction on the minimum length of a segment required by *DNAcopy* (2 probes). Consecutive probes were merged to call copy number variable regions and separate lists were obtained for the different HapMap populations (CEPH, YOR, CH-JA) and the 473 normals. Finally, the three HapMap populations were merged with the set of 473 normals to generate a list of CNVs present in these populations (HapMap+Normals CNV list).

For tumour samples, the data were smoothed and *DNAcopy* was run with default parameters. We then applied the MergeLevels algorithm [27] to the segmented data. It can generally be assumed that there is a mixture of normal (benign) and tumour cells present within a given sample. In order to remove the dependence between cellularity and the proportion of alterations, we employed different thresholds for calling alterations and high-levels events according to the cellularity of each sample, resulting in the following somatic copy number states:

$$K_{CNA} = \{HOMD, HETD, NEUT, GAIN, AMP\}\,. \tag{1}$$

For ‘high cellularity’ tumours, the median of the $\log_2$ ratio $+2\sigma$ or $+6\sigma$ was computed for the 50% of the central probes (ordered by their log ratios) to call gains and amplifications, respectively. For losses, the median of the $\log_2$ ratio -2.5σ or -7σ was used to call heterozygous and homozygous losses, respectively. For ‘moderate cellularity’ tumours, the median of the $\log_2$ ratio of each array $+2\sigma$ or $+6\sigma$ was computed for the 45% of the central probes and used to call gains and amplifications, respectively. For losses, the median of the $\log_2$ ratios -2.5σ or -7σ was used to call heterozygous and homozygous losses, respectively.

8.3 *Identification of germline CNVs*

For the tumour samples, any segmented mean that fell within a region included in the HapMap+Normals CNV list was flagged as an inherited CNV, yielding the following germline copy number states:

$$K_{CNV} = \{CNVLOSS, CNVGAIN\} \tag{2}$$

In order to remove all possible CNVs, the frequencies of CNAs in the tumour samples were obtained after removing these CNVs from the data normalised with the pool and the corresponding matched normals, when available. In order to retain all possible CNVs, the frequencies of CNVs in the tumour samples were obtained with these CNVs flagged in the data normalised only with the pool. We further compared the distributions of CNVs derived from tumour-adjacent normal breast tissue to those obtained from constitutional blood DNA in a separate cohort of 1,999 breast cancer patients (unpublished data), and found that 92.8% of CNVs identified in these normals were also called in blood, suggesting that adjacent normal breast tissue is a suitable surrogate for germline CNV profiling. The list of CNV frequencies, gene-centric Ensembl-based CNV mappings, and population distributions are presented in Tables S12-S15, and a plot illustrating that the CNV landscape was the same irrespective of the intrinsic subtype is presented in Figure S6.

8.4 *Gene-centric alterations*

In order to identify the genes and features that were altered by copy number aberrations, we searched for the overlap of segments with gene regions. The complete set of gene annotations and coordinates given by Ensembl 54 (hg18) that correspond to Illumina HT-12 v3 probes was employed, resulting in 30,566 features.

Two patient-by-gene copy number matrices were generated to capture different perspectives of gene alterations. The call matrix, C, was populated with values representing discrete copy number states, whereas the log-ratio matrix, L, contains the segmented mean. For each patient $p \in P$ and each gene $g \in G$, we identified segments s that overlap g and assign $C(p, g)$ with the copy number state of s and $L(p, g)$ with the mean log-ratio of the probes in s. If gene g overlaps or is broken by a set of segments, $S = s_1, \ldots, s_k$, where $k \geq 2$, we assign $L(p, g)$ with the maximal log-ratio (Equation 3). For $C(p, g)$, the copy number state of the segment with maximal *severity* was assigned based on the relationship defined in (Equations 4 and 5), where ties were broken in samples exhibiting both a loss and gain according to the maximal absolute value of the segmented mean.

$$\begin{aligned}
L(p,g) &= \Big(\sum_{s\in S} length(s) * meanLogR(s)\Big) / \sum_{s\in S} length(s) && (3)\\
C(p,g) &= CNstate\Big(\underset{s\in S}{argmax}\{severity(CNstate(s))\Big) && (4)\\
severity &= \{\text{‘}NEUT\text{’} < \text{‘}HETD\text{’} < \text{‘}HOMD\text{’}, \\
&\quad \text{‘}NEUT\text{’} < \text{‘}GAIN\text{’} < \text{‘}AMP\text{’}, \\
&\quad \text{‘}CNVNEUT\text{’} < \text{‘}CNVLOSS\text{’} = \text{‘}CNVGAIN\text{’}\} && (5)
\end{aligned}$$

Note that somatic and germline events were treated separately, but that in some cases both types of events can occur within different segments of the same gene. On occasion, an expression probe may fall in the middle of a copy number breakpoint for a particular sample. In this case there is no way to interpolate the value, so the copy number value for the probe is deemed missing. The discrete call matrix was subsequently used in ANOVA to identify copy number events significantly associated with changes in gene expression, the top associations of which were selected as features for integrative clustering as described below.

As the call matrix only allows for one copy number state per gene, when multiple states are in fact possible, we also report the frequency of samples showing all possible levels of alteration within a gene for the purpose of understanding the entire spectrum of gene copy number alterations in this cohort. Thus a sample can exhibit both a gain and neutral segment spanning the same gene or, alternatively, a region of loss and region of gain within a gene. Here, a sample may contain a region of loss and a region of gain in a particular gene, but for a sample to be considered neutral it must be so for the entire length of the gene. These frequencies were computed for all tumours, as well as by ER status, PAM50 subtype, and the novel integrative cluster-based subgroups defined in this study. For these summaries, we extended the feature set to include ncRNAs and other genomic regions not necessarily targeted by the Illumina HT-12 v3 array by using the complete set of annotations and coordinates given by Ensembl 54 (hg18), with the exclusion of ChrNT and ChrMT genes, resulting in 36,656 features (36,209 when ChrY is excluded). The frequency with which a gene (feature) exhibited a copy number aberration in a particular subtype is reported in Table S4 and the frequency with which a gene was broken by a copy number event is reported in Table S6. Subtype-specific copy number profiles for the PAM50 intrinsic subtypes are plotted in Figure S32. As described below, we also computed a number of summaries of the copy number data that were not restricted to a gene scaffold.

8.5 *Probe-level alterations*

We computed the frequency of alterations for each probe in all tumours and according to ER status, PAM50 subtype and the novel integrative subgroups defined in this study. Frequencies of alteration were computed for germline CNV states:

$$K_{CNV} = \{CNVLOSS, CNVGAIN\} \tag{6}$$

and for each of the 5 discrete somatic states:

$$K_{CNA} = \{HOMD, HETD, NEUT, GAIN, AMP\} \tag{7}$$

in addition to the collapsed summaries of alterations represented by loss (HOMD & HETD), neutral (NEUT), and gain (GAIN & AMP). Several probe-level plots of putative homozygous deletions are illustrated in Figure S17.

8.6 *Measures of genomic instability*

The extent of genomic instability (GI) was computed for each sample using four related measures summarized from the CBS segmented data, namely the area under the segmented mean (area), proportion of probes altered (proportion), the proportion of genome altered (proportion genome), and the sum of changes in altered segmented means divided by the sum of all changes in segment means (e.g. the ratio of the mean change in log-ratio due to alterations relative to variability) (jump). These indices were computed for each sample and each chromosome band and clustering was performed jointly on the proportion of genome altered and jump measure using the Euclidean distance metric and Ward's method in order to identify subgroups partitioned by genomic instability (Figure S26).

8.7 *Identification of recurrent alterations*

We employed the method of Rouveirol *et al.* [28] to compute the minimal common region of alteration for CNAs and CNVs present in the tumours according to ER status and PAM50 subtypes (Tables S5, S12). Due to the difference in their inherent structure, for the CNAs a minimum frequency of 0.05 and 3 probes was required, while for the CNVs no minimum length was imposed and the same minimum frequency was required. Regions of gains and losses were computed independently and were annotated for gene content based on Ensembl 54 (hg18).

9 Segmentation and copy number alteration calling using HMM

9.1 *Normalisation of intensities*

The Affymetrix SNP 6.0 arrays for both tumour and normal samples were independently normalised using the single array method CRMAv2, implemented in *aroma.affymetrix* [23], as described above for CBS. Log-ratios were computed for both tumours and normals by normalising each array independently against a reference. Due to the incomplete pairing of tumours and matched normals, we generated a pooled reference from all the available normals. However, because we are also interested in the identification of CNVs in addition to CNAs in the tumours, using a simple pooled reference could potentially result in the 'subtraction' of frequent (> 50%) copy number events in the normals from the tumours. Therefore, we generated a 'masked' reference from the normal dataset that was free of inherent copy number polymorphisms, while accounting for systematic biases. Starting with the CRMAv2 normalised intensities for the normal samples, we computed log-ratios using a randomly permuted reference. Probe-level discrete copy number calls were then made made using a 6-state hidden Markov model (described below) such that for each sample, probes that were predicted to span regions of gain or loss were identified. Next, we retrieved the normalised intensities and masked out the probes identified in the previous step. Finally, we computed the median (pooling) intensity value for each probe across all samples using only the remaining values after masking, resulting in a vector of median values, which makes up the masked reference. For HMM, the masked reference improved the analysis of germline CNVs in the tumours by serving as a base-line vector of intensities for log-ratio calculations.

9.2 *6-state HMM for segmentation and discrete copy number prediction in normals*

We applied a 6-state HMM to the set of normal samples in order to detect germline CNVs by returning probe-level and segment copy number. The 6-state HMM is a modified version of CNA-HMMer [22] that has been adapted to analysing high-density genotyping arrays such as the Affymetrix SNP 6.0 platform. Another extension is the inclusion of additional copy number states, which offers a more intuitive interpretation of DNA dosage in cancer. The new discrete copy number state space included neutral (NEUT), homozygous (HOMD) and hemizygous (HETD) deletions, gain (GAIN), amplification, (AMP), and high-level amplification (HLAMP).

$$K_{CNA} = \{HOMD, HETD, NEUT, GAIN, AMP, HLAMP\} \tag{8}$$

9.3 *HMM-Dosage for segmentation and discrete copy number prediction in tumours*

We performed copy number analysis on the 997 tumour samples using HMM-Dosage (Ha *et al.*, manuscript in preparation; http://compbio.bccrc.ca). The algorithm was designed to detect and distinguish the complete set of somatic and germline copy number events in cancer genomes interrogated by SNP array data. This model extends the 6-state HMM (above) by using 5 additional CNV states to represent the analogous copy number status

$$K_{CNV} = \{CNVHOMD, CNVHETD, CNVGAIN, CNVAMP, CNVHLAMP\}. \tag{9}$$

The model performs segmentation on log-ratios of intensity data and assigns a discrete copy number states $k_t \in \{K_{CNA}, K_{CNV}\}$ to each latent variable, Z_t, for all probes $t \in \{1 \ldots T\}$. Log-ratios, $Y = \{y_1, \ldots, y_T\}$, are observed values modeled using Student-t densities in the emission component of the HMM, conditional on $Z_t = k_t$. Spatial information is captured using a non-stationary transition matrix, A_t, that encodes position-specific probabilities of CNVs and CNAs at each probe t. In order to distinguish between somatic copy number alterations (CNAs) and germline copy number variants (CNVs), HMM-Dosage probabilistically incorporates CNV information as a prior to the transition matrix. The CNV prior was computed as probe-level CNV frequencies by the weighted averaging of two datasets: the set of normal samples and an external dataset of 450 HapMap normal samples whose CNVs were predicted by Conrad *et al.* [29]. Because the HapMap dataset contains males, we excluded the X chromosome frequencies and instead used the X chromosome frequencies from the set of normals (all of which were female).

The parameters of the Student's t distributions are unobserved and estimated using the expectation maximisation (EM) algorithm for each sample, independently. Initial parameters for the K_{CNA} states of the Student's t mixture were empirically determined using 45 SNP 6.0 breast cancer cell line samples in the COSMIC [30] dataset. First, independently for each cell-line sample, we fit the log-ratios to a 6-state Gaussian mixture model using EM. Subsequently, the converged Gaussian mixture parameters were used as initial parameters for fitting a 6-state Student's t mixtures in the 6-state HMM (again, using EM). Finally, we averaged the converged Student-t parameters for each state across the 45 cell lines. These averaged parameters became the initial K_{CNA} parameters to Student's t emission of HMM-Dosage. Initial parameters for K_{CNV} were set by hand to theoretical values $\log_2([0.5, 1, 3, 5, 7]/2)$. The optimal state sequence, $Z_{1:T}$, which represents the copy number prediction for each probe, was computed using the Viterbi algorithm. HMM-Dosage was tuned to be conservative in calling CNVs so as to avoid misclassifying CNA calls. Therefore, the results were post-processed by comparison against the CNV map (derived from the set of normals and 270 HapMap individuals) such that CNA segments that had at least 25% reciprocal genomic overlap with a segment of the same copy number state in the CNV map were converted to the germline state.

9.4 *Gene-centric alterations*

In order to identify the genes that were altered by copy number changes, we searched for the overlap of segments with gene regions. The gene annotations and coordinates are given by Ensembl 54 (hg18) and the analysis was restricted to those that are protein-coding and overlap with the annotations for the Illumina HT-12 v3 gene expression arrays, resulting in 18,733 genes.

We generated two patient-by-gene copy number matrices to capture two perspectives of gene alterations. The first is the call matrix, C, which is populated with values representing discrete copy number state calls for each patient. The second is the log-ratio matrix, L, which contains segment median log-ratios. For each patient $p \in P$ and each gene $g \in G$, we identify segments s that overlap with g and assign $C(p, g)$ with the copy number state of s and $L(p, g)$ with the median log-ratio of the probes in s. If gene g overlaps or is broken by a set of segments, $S = s_1, \ldots, s_k$, where $k \geq 2$, we assign $L(p, g)$ and $C(p, g)$ with the weighted sum of the median LogRs (Equation 10) and the copy number state of the segment with maximal *severity* based on the relationship denoted in Equations 11 and 12.

$$L(p,g) = \left(\sum_{s \in S} length(s) * medianLogR(s)\right) / \sum_{s \in S} length(s) \tag{10}$$

$$C(p,g) = CNstate\left(\underset{s \in S}{argmax}\{severity(CNstate(s))\right) \tag{11}$$

$$\begin{aligned} severity = \; & \{(\text{‘}NEUT\text{’}, 0), \\ & (\text{‘}HOMD\text{’}, 8), (\text{‘}HETD\text{’}, 6), \\ & (\text{‘}GAIN\text{’}, 5), (\text{‘}AMP\text{’}, 6), (\text{‘}HLAMP\text{’}, 8), \\ & (\text{‘}CNVHOMD\text{’}, 4), (\text{‘}CNVHETD\text{’}, 3), \\ & (CNVGAIN\text{’}, 2), (\text{‘}CNVAMP\text{’}, 3), (\text{‘}CNVHLAMP\text{’}, 4)\} \end{aligned} \tag{12}$$

The frequency with which a gene exhibited a copy number aberration in a particular subtype is reported in Tables S7 and the frequency with which a gene was broken by a copy number event is reported in Table S8. Subtype-specific copy number profiles for the PAM50 and integrative subtypes are plotted in Figures S33 and S28, respectively.

10 Comparison of segmentation methods based on MPLA

We validated the accuracy of both segmentation methods using the MLPA data, which assayed a panel of 90 commonly aberrant cancer loci in 40 tumours. MLPA produces a copy number estimate for each probe (gene) according to the following 6 states: homozygous loss [HOLoss] (0 copies), heterozygous loss [LOH] (1 copy), neutral [Normal] (2 copies), gain [Gain] (3 copies), amplification [Amplification] (more than 3 copies), and ambiguous (borderline value). Note that probes classified as ambiguous by MLPA were not included in the computations. MLPA probes were aligned with the genome using the exonerate program and the resulting coordinates were used to extract the corresponding CBS and HMM copy number states for each of the 90 × 40 MLPA values. The most extreme value was selected when multiple states were present in a given region in accordance with their respective call matrices, C.

The following definitions were employed for these analyses:

- Sensitivity: TP / (TP + FN)

- Specificity: TN / (TN + FP)
- Positive predictive value (PPV): TP / (TP + FP)
- Negative predictive value (NPV): TN / (TN + FN)

where TP = number of true positives, FN = number of false negatives, TN = number of true negatives, and FP = number of false positives. We used different definitions of TP, TN, FP, FN to assess the ability of the two copy number segmentation methods to detect different types of alterations. The notation ([Gain],[G,L]) indicates the number of probes with a gain in MLPA and a gain or loss according to the segmentation method:

1. Global:
    - TP = ([Amplification],[AMP]) + ([Gain], [GAIN]) + ([LOH, HETD]) + ([HOLoss], [HOMD])
    - FN = ([Amplification], [GAIN, NEUT, HETD, HOMD]) + ([Gain], [AMP. NEUT, HETD, HOMD]) + ([LOH], [AMP, GAIN, NEUT, HOMD]) + ([HOLoss], [AMP, GAIN, NEUT, HETD])
    - FP = ([Normal], [AMP, GAIN, HETD, HOMD])
    - TN = ([Normal], [NEUT])
2. High Level:
    - TP = ([Amplification],[AMP]) + ([HOLoss], [HOMD])
    - FN = ([Amplification],[GAIN, NEUT, HETD]) + ([HOLoss],[GAIN, NEUT, HETD])
    - FP = ([Gain, Normal, LOH, HOLoss],[AMP]) + ([Amplification, Gain, Normal, LOH],[HOMD])
    - TN = ([Gain],[GAIN, NEUT, HETD]) + ([NEUT], [GAIN, NEUT, HETD]) + ([LOH], [GAIN, NEUT, HETD])
3. Amplifications:
    - TP = ([Amplification],[AMP])
    - FN = ([Amplification],[GAIN, NEUT, HETD, HOMD])
    - FP = ([Gain, Normal, LOH, HOLoss],[AMP])
    - TN = ([Gain],[GAIN, NEUT, HETD, HOMD]) + ([NEUT], [GAIN, NEUT, HETD, HOMD]) + ([LOH], [GAIN, NEUT, HETD, HOMD] + ([HOLoss], [GAIN, NEUT, HETD, HOMD])
4. Homozygous deletions:
    - TP = ([HOLoss],[HOMD])
    - FN = ([HOLoss],[AMP, GAIN, NEUT, HETD])
    - FP = ([Amplification, Gain, Normal, LOH],[HOMD])
    - TN = ([Amplifcation], [AMP, GAIN, NEUT, HETD]) + ([Gain],[AMP, GAIN, NEUT, HETD) + ([NEUT], [AMP, GAIN, NEUT, HETD]) + ([LOH], [AMP, GAIN, NEUT, HETD])
5. Alterations:
    - TP = ([Amplification, Gain], [AMP, GAIN]) + ([LOH, HOLoss], [HETD, HOMD])

- FN = ([Amplification, Gain], [NEUT]) + ([LOH, HOLoss], [NEUT])
- FP = ([Normal, LOH, HOLoss], [GAIN, AMP]) + ([Amplification, Gain, Normal], [HETD, HOMD]) + ([Amplification, Gain], [HETD, HOMD]) + ([HOLoss, LOH], [GAIN, AMP])
- TN = ([Normal, NEUT])

6. Gains:

   - TP = ([Amplification, Gain], [AMP, GAIN])
   - FN = ([Amplification, Gain], [NEUT, HETD, HOMD])
   - FP = ([Normal, LOH, HOLoss], [GAIN, AMP])
   - TN = ([Normal, LOH, HOLoss], [NEUT, HETD, HOMD])

7. Losses:

   - TP = ([LOH, HOLoss], [HETD, HOMD])
   - FN = ([LOH, HOLoss], [AMP, GAIN, NEUT])
   - FP = ([Amplification, Gain, Normal], [HETD, HOMD])
   - TN = ([Amplification, Gain, Normal], [AMP, GAIN, NEUT])

Based on these indices and definitions, accuracy and precision values, as well as, sensitivity and specificity and the positive and negative predictive value, were estimated for CBS and HMM-derived somatic copy number calls (Figure S4a and Table S11). Additionally, we computed ROC curves for the CBS-derived CNAs using different thresholds to call alterations (Figure S4b).

11 The copy number landscape of breast cancer

11.1 *The germline CNV landscape*

We interrogated the genomic architecture of primary tumours and normal samples to generate the first genome-wide breast cancer CNV map (Figure S6, Tables S12-S15). Since CNVs are generally thought to localise to non-genic regions, it is noteworthy that numerous cancer genes (including *BCAS1, CCND2, EPHA3, ERBB4, ETV6, JAK1, JAK2, MET, PDGFRA, PML, PTEN, RET, TMPRSS2, WNK1*) were each targeted in $>5\%$ of samples (Figure S6a), usually resulting in gain or deletion of only a portion of the gene. Notably, the CNV landscape was found to be conserved across the intrinsic breast cancer subtypes (Figure S6b), whereas distinct somatic copy number profiles characterised the subgroups. A complete list of CNV frequencies, gene-centric Ensembl-based CNV mappings, and population distributions are presented in Tables S12-S15. Taken together with the observed moderate impact on expression variation noted in the eQTL studies, these findings hint that the common CNVs that can be identified on this platform are unlikely to account for a substantial portion of the missing heritability of breast cancer, as has been noted by others in the context of cancer and other complex diseases [31].

11.2 *The somatic CNA landscape*

The minimal common regions of alterations [28] were computed according to ER status and for the PAM50 [11] expression-based intrinsic subtypes (Table S5), and recurrent alterations were summarized using a gene-centric approach (Tables S4, S7). We then compared the underlying copy number profiles for each of the intrinsic subtypes (Figures S32, S33). Despite the observation that distinct copy number profiles characterised the subgroups, unsupervised clustering of genome-wide copy number data in the discovery cohort revealed additional heterogeneity within the subgroups (Figure S31). We also clustered the data on a combined measure of GI (Figure S26) to reveal nine major subgroups, three of which associated with poor outcome. In contrast to previous reports [32], these clusters did not recapitulate the intrinsic subtypes. For example, while the Basal-like tumours (n = 118) were relatively homogeneous, a low-GI subgroup was evident within them, as we noted previously [33]. *ERBB2* copy number independent cases (n = 20) were also apparent within the PAM50 HER2-enriched subtype (n = 87). Examination of the Luminal tumours indicated a high number of cases with *ERBB2* amplification (Luminal A, n = 54; Luminal B, n = 74), suggesting that refinement of the HER2-enriched intrinsic subtype is needed.

Using refinements of CNA boundaries we enumerated the location and frequency of genes that may be disrupted by amplicon events. Certain genes were repeatedly broken, as a result of CNAs that occur within the gene such that a segment of copy number gain or loss is accompanied by a copy number neutral segment, disrupting its contiguous sequence. Recurrently broken genes include *EGFR, ERBB2, FGFR1, IGF1R, PPP1R12B* and *PTPRT*, whereas others (*CCND1, ZNF703*) were never broken and reflect amplicons that typically span many genes. The complete lists are enumerated in Tables S6 and S8.

12 eQTL analysis

Several eQTL analyses were performed using ANOVA, where the ranked expression values for each probe were regressed, in turn, on the set of CNV, CNA and SNP calls. The primary analysis relied on measurements derived from tumour tissue in the cohort of 997 individuals. However, as the individuals were of mixed ancestry, we also evaluated the potential influence of population structure on the results, by performing the analysis on an ethnically homogeneous population. Performing the analysis on a smaller cohort of European ancestry and for whom normal genotype data were available substantially reduced the sample size, and a decrease in the number of significant associations was anticipated. However, the relative contribution of different predictors was stable (Figure S12, Table S18), suggesting that the significance of the eQTLs was not inflated by population stratification. This analysis also lends support to our approach for dealing with the uncertainty of genotype calls derived from tumour tissue, wherein as a result of non-diploid genotype states in the tumours, SNP genotypes coincident with regions of copy number change might be altered. To address potential biases this might introduce, SNPs within 500 bp of a non-diploid CNV or CNA were set to missing. As this choice was heuristic, the suitability of this threshold was assessed by repeating the analysis on a set of 244 individuals of European ancestry and for whom 'unbiased' genotype calls had been obtained from normal breast tissue. As described below, this analysis indicates that only a minimal increase in precision could be expected had normal genotype data been available for all samples (Table S19). Finally, we investigated the robustness of the germline/somatic partitioning by considering a set of 85 samples, for which gene expression data were also available from normal breast tissue. Although the power for this analysis was markedly lower, it was assumed that strong germline associations present in the main analysis would persist, while associations due to somatic copy number changes would vanish, and indeed this was observed (Figure S13, Table S20).

12.1 *Preprocessing of predictors*

Based on the annotation files employed, the SNP 6.0 chip provided measurements on 1,876,300 probes, including 906,600 polymorphic probes and 969,700 copy number, non-polymorphic probes. Since the CBS-derived copy number calls provided the complete enumeration of germline CNVs and somatic CNAs for each sample, these events were analysed separately. Copy number states were collapsed into three levels, namely, loss (HETD, HOMD, or CNVLOSS), neutral (NEUT) or gain (GAIN, AMP, or CNVGAIN), coded as 0, 1, and 2, respectively. As noted above, it was possible for a region to contain both a CNV and CNA, typically due to a narrow CNV being flanked on either side by a CNA. Based on these data, CNV and CNA matrices were populated, both of size 997 × 1,876,300, where each row corresponds to a patient and each column a copy number probe, while the element records the copy number state (0, 1, or 2).

Similarly, a SNP matrix of size 997 × 906,600 was constructed from the genotype calls derived from the SNP (polymorphic) probes on the chip, where each row corresponds to patient, each column a SNP probe, and each element records whether the genotype is homozygous (AA), heterozygous (AB), or homozygous (BB), coded as 0, 1, 2, respectively. Due to the fact that these genotypes derive from tumour tissue, we anticipate biases to arise when a SNP probe corresponds to a non-diploid copy number region, as a deletion/amplification will likely reduce/increase the allele count, respectively. For this reason, any individual's SNP calls within 500 bp of such a region were set to missing, resulting in the loss of approximately 9.7% of genotypes. The second analysis using SNPs derived from non-cancerous tissue sought to investigate the extent of these biases. The logic for selecting a cut-off of 500 bp is noted below.

The three (raw) predictor matrices were then condensed to obtain processed predictor matrices through a series of filtering steps. The first step removed trivial columns such as CNVs, CNAs, or SNPs that exhibited no variation across the 997 individuals. The second step removed duplicate columns. For copy number predictors, these typically corresponded to probes that lie within the same copy number region; for genotypes, these corresponded to SNPs in very high linkage disequilibrium (LD). Following these processing steps, 11,538 CNV, 193,873 CNA and 874,649 SNP predictors remained.

12.2 *Preprocessing of responses*

Given the large number of tests to be performed in the genome-wide setting, we sought to reduce this by eliminating potentially spurious expression probes. Following the strict criteria for exclusion of probes from the normalisation of the expression data, we also excluded probes that either contained a SNP, did not represent a perfect genomic match, were not $3'$ matching and without multiple genomic matches, contained a polyG tail at the end of probe sequence, or whose GC content was not between 38% and 64% (but retained those that targeted the sex chromosomes or were on the PAM50 genelist), resulting in 28,609 probes. The gene expression measurements were stored in a matrix of size 997 × 28,609, with again one row for each individual, and a column for each of the filtered responses. The eQTL analysis subsequently involved regressing each column of the response matrix against the columns of each predictor matrix.

In order to identify the linear combinations of responses that exhibited the most variation across individuals, principal component analysis (PCA) was applied to the expression matrix. As samples derived from multiple sites, it was anticipated that this would be apparent in the top principal components, as was indeed the case. Additionally, ER status was expected to have a sizeable effect on expression. To account for these factors, the expression values were regressed on site and ER status using a categorical model ($2 \times 3 = 6$ degrees of freedom). From this point on, the actual expression values were replaced by the corresponding residuals from this regression. Though it would have been possible to include site and ER status as cofactors in the subsequent association analysis, this approach reflects the belief that these two cofactors are likely to explain a given probe's variation more than any single predictor and additionally it reduces the computation time significantly.

Our original intention was to make use of the quantitative nature of gene expression measures, and regress the raw values directly on the predictors. However, we noted that the response matrix contained a large number of outliers that might be due to large changes in expression, but could also be attributed to measurement error. Such extreme values are highly undesirable when regressing on low-variance predictors. For example, suppose a predictor is observed in *state* 0 for only one individual (e.g. a copy number loss, which is observed only once). If a response were to have a single extreme value, then at random, the chance of this coinciding with the *state* 0 individual is 1 in 997, but when it does, basic regression methods will typically return a far more extreme *P*-value, usually leading to spurious results due to false associations. In these cases, generating *P*-values through permutation testing will often be more informative, as they will recognise the chance of such an occurrence [34]. However, when regressing each of many thousands of responses against many tens of thousands of predictors, meaningful permutation testing is not feasible. For this reason, we experimented with using raw values or introducing various corrections for outliers, and ultimately decided to convert the expression measurements to their corresponding ranked values such that the resulting analysis was essentially equivalent to performing a Kruskal-Wallis test, which is based on the one-way ANOVA formula using ranks [35].

We compared the ANOVA on ranks and the KW test and find that the latter produces, on average, slightly less extreme *P*-values. However, the conclusions from the eQTL analysis were nearly unchanged. The high degree of concordance between the *P*-values from ANOVA and those from a KW test, for CNV, CNA, and SNP associations is shown in Figure S11a. In all cases except a single SNP eQTL, the genes declared not to have a significant association by the ANOVA test were similarly not declared to have an eQTL association using KW. Relatively few (3/78, 180/10,863, 122/1,998) of the genes declared by ANOVA to have a significant CNV, CNA or SNP association were not significant in the KW test. Although there is a drop in concordance for extreme values (-$\log_{10}$ *P*-value >20), as with any asymptotic test, the accuracy of *P*-values deep into the tails is inevitably less reliable. We also compared the Venn diagrams showing the relative number of significant association for each predictor based on ANOVA (Figure 1) and the KW test (Figure S11b). While slightly fewer genes have significant eQTLs based on the KW test, the percentages in each cell stay almost constant. Below we explain how the use of ranks impacted the analysis.

12.3 *eQTL analysis of tumour gene expression profiles*

ANOVA was performed to regress each gene expression, in turn, on the three predictor sets (CNVs, CNAs, SNPs) (see Figure 1, Table S17) as described here for the case of CNVs. For each response, an ANOVA test was performed for each of the 11,538 CNV predictors, in turn. A likelihood ratio test was used to compare the null model of no association against a categorical alternative model. In particular, we assume that we have k random samples, one from each of k populations, where $X_{j1}, \ldots, X_{jn_j}$ is a random sample of size n_j from the jth population, $j = 1, \ldots, k$. We assume the jth population has median μ_j and variance σ^2, and that the k random samples are independent. Here we note that under the null hypothesis that samples come from populations with identical locations, the distribution of the response is independent of the predictor state; while under the alternative hypothesis that at least one group differs in its distribution, its median varies according to which state is observed.
Therefore if μ_0, μ_1, μ_2 represent the median of each population group, we test

$$\text{H0: } \mu_0 = \mu_1 = \mu_2 \qquad\qquad \text{H1: } \mu_i \neq \mu_j \text{ for some } i \neq j$$

The null model has degrees of freedom 2, one for the location and one for the variance. The alternative model, with degrees of freedom d+1 equal to the number of states observed for that predictor (typically d = 3) and one for the variance, allows each state to have an independent effect on the response. A test statistic was obtained as a function of the percentage of variance explained by the full model, which was compared to a χ^2 distribution with degrees of freedom d-1 to obtain a P-value. For each response, the smallest of the 11,538 P-values was retained.

Having performed this analysis for all 28,609 gene expressions, we assessed the evidence for significant CNV associations (likewise for CNAs and SNPs). Due to the large number of tests performed, it was necessary to correct for multiple testing. The Šidák correction [36] adjusts each P-value using a transformation of the form: $p \rightarrow 1\text{-}(1\text{-}p)^{Nr}$, where Nr represents the effective number of tests: N is the number of predictors, while r is a ratio between 0 and 1. The Šidák method is less conservative than the Bonferroni method, while guaranteeing strict control of the family-wise error rate (when comparisons are independent). If the tests for each of the 11,538 CNVs were independent, r = 1 would be suitable. However, as the tests were not independent, such a correction would be overly cautious, resulting in a bias of adjusted P-values close to 1 (Figure S10a). We attempted to find a suitable value of r using median mapping such that the median P-value is transformed to 0.5, similar to the genomic control approach [37]. This supposes that for most of the responses, the null hypothesis is true, in which case the corresponding P-value will be drawn from a uniform distribution on [0,1]. This transformation appears to perform reasonably well for the CNVs, where we anticipate that not many predictors will be associated, but again seems too conservative for CNAs and SNPs: if many of the P-values correspond to true alternative hypotheses, then the median of the correctly adjusted P-values will be significantly less than 0.5, so the r calculated from this method will be too large (Figure S10b). Q-Q plots of the adjusted P-values based on Šidák correction (r = 1) and median mapping were evaluated (Figure S10c). If P-values were correctly adjusted, then for values falling between quantiles 0.65-0.93, the QQ line should be straight in this region. The choice of quantiles, while fairly arbitrary, arose from considering that 8,000 P-values (0.28%), which should be more than sufficient to obtain a straight line, and should allow for a large enough window to judge a flat part of the distribution, an approach inspired by that used to calculate the q-value [38, 39]. We also sought to avoid P-values in the right tail, close to the upper boundary, in case the distribution was skewed for being too close to 1. We selected the value which maximised this correlation and so resulted in the straightest portion of line (Figure S10d). The values of r for CNVs, CNAs and SNPs (0.285, 0.045 and 0.67, respectively) correspond to the effective number of independent tests for each of these predictors (3,289, 8,724 and 586,014) and result in histograms with more desirable properties (Figure S10e).

12.4 *Effect of ranking expression values*

We performed a small study to test the effect of regression using ranked expression values as opposed to raw expression values. For a set of 300 expression probes, we concluded there was sufficient evidence for an association if the P-value was less than 0.001. For each probe, we then calculated an empirical P-value by regressing the raw expression values on the predictors and using 10,000 permutations. The permuted P-value was considered the gold standard, from which the true presence/absence of a CNV association was determined according to whether the empirical P-value was less than 0.001 or not (Figure S11c). When instead using P-values obtained by regressing ranked values on predictors, for 40 of the 300 probes (13%), a different conclusion was reached depending on the approach used. This figure should not be considered representative of an overall misclassification rate, as the 300 gene probes were selected on the basis that they were potentially troublesome (returning P-values close to the significance threshold). For the same set of 300 gene probes, the analysis was repeated using the raw expression values. This time, contradictory conclusions were reached for 173 of the 300 probes (58%), suggesting that outliers were common-place and that regression with ranked values was more reliable.

12.5 *Definition of cis/trans windows*

In the context of these association analyses, the definition of proximity is somewhat arbitrary, and the proportion of associations in *cis* will change as this window is varied. As the definition of the *cis* window is increased from 3 Mb (as reported for the main analysis), to 10 Mb, to the same chromosome, the number of *trans* associations will decrease, while the number of genome-wide associations remains constant (Figure S9a).

An alternative approach would have been to perform the analysis twice: first considering only proximal predictors and second considering only distal predictors (Figure S9b). The first pass will be similar in spirit to the approach used by Stranger *et al.* [34], and would have greater power to detect *cis* associations by considering them on their own, rather than with all other predictors. However, since two eQTL analyses are being conducted for each predictor type, and potentially two predictors can be declared associated for each expression probe, it seemed prudent to tighten the declaration threshold to 0.00005, (i.e. half the previous cut-off). As expected, the values here are larger than when searching genome-wide (e.g. performing a single analysis and then partitioning associations into proximal and distal according to window size, as described above). However, they decrease as the definition of the proximal window is increased, showing that the increase in penalty for multiple testing is only partly offset by the extra signal included when widening the search. As the multiple testing penalty is nearly twice that used in the original analysis, it might be surprising that more associations are found for the distal associations as the proximal window is widened. A clue as to why this might be the case comes from considering the effect of increasing the definition of proximal. Here we observed that ‘distal predictors’ close to the proximal boundary are over-represented among those declared associated within the *trans*-associated group. Thus if the definition of proximal as events that occur on the same chromosome is considered overly cautious, this suggests there some *trans* signals will be obscured in the main analysis, presumably because they are overcome by more dominant *cis* signals.

12.6 *Assessment of genotype calls in non-diploid regions and the impact of population stratification*

Another analysis considered a set of 244 individuals of European-apparent ancestry and for whom SNP genotyping had also been performed on normal breast tissue (Figure S12, Table S18). As mentioned above, an element of doubt in the veracity of the genotype calls was introduced when a SNP coincided with a non-diploid (non-neutral) region of copy number variation. In the main analysis, these calls, and those within 500 bp of the region’s boundaries, were set to ‘missing’. The threshold of 500 bp was selected heuristically, by comparing the accuracy of SNP calls close to regions of copy number change.

For the 244 individuals, the accuracy of the genotyping performed on tumour DNA was measured by assuming the calls obtained from non-tumour tissue were correct. The baseline miscalling rate calculated from all calls at least 50 kb away from the nearest copy number change was estimated to be 2.39%. The accuracy of calls with varying proximity to a non-diploid copy number region (CNV or CNA) was also computed (Table S19). For example, for calls lying within a region of copy number change (threshold 0), the miscalling rate was 8.13% for CNVs and 9.03% for CNAs, and these values diminished as the threshold was increased. These two figures are not independent, as SNPs can be near both a CNV and a CNA. To judge the best threshold cutoff, we considered the miscalling rate between consecutive thresholds. Ideally we would stop when this matched the baseline, however, as the decay of this rate to baseline seemed quite gradual for high thresholds, we selected 500 bp. Although at this distance the rate appears at least 1 point above baseline, we considered this a fair trade-off between accuracy and introducing too many missing values.

The subset of 244 individuals was analysed via ANOVA in an identical fashion to the first analysis of 997 individuals. Although the detection power would be decreased, the expectation was that the ratio of gene probes found to have CNV, CNA, and SNP associations would remain similar, suggesting that the issue of dubious SNP genotyping calls in the presence of copy number variation was dealt with appropriately, and that population stratification did not inflate the significance of the eQTL results. When compared to the main analysis on 997 individuals, the number of gene probes with one or more association fell from 11,198 to 2,111. However, the relative contributions of different predictor types remain reasonably stable. Of the 2,111 gene probes, 90.5% have a CNA association, 9.8% have a SNP association and 0.5% have a CNV association (compared to 97%, 14.7% and 0.7%). However, the fact that the proportion has risen for CNAs while dropping for SNPs might suggest that with only 244 individuals, the chance of detecting two signals for a particular probe is greatly reduced and this is demonstrated by the sharp reduction in the proportion of gene probes with both a CNA and SNP association recorded (11.5% to 0.3%).

Of the 208 gene probes found to have a SNP association, 204 were reported in the first analysis. That four were not reported reflects the fact that at a false positive rate (FPR) of 0.0001, we would expect 2.9 false calls from each analysis. Further examination of these results revealed that 3 of these 4 associations only just exceeded the declaration threshold (*P*-values between 0.00004 and 0.0001). As is evident from the variance histograms (Figure S12b), many of the top associations from the analysis of the 997 individuals (Figure 1) have persisted. These plots also highlight the effect of reducing the sample size. For example, for CNVs and CNAs the proportion of variance a predictor must explain to register as an association rises from approximately 3% to 12%, shown by movement of the left boundary of the CNA histograms. For SNPs, the cut-off rises from about 4% to 16%. These results lend support to our approach for dealing with the uncertainty surrounding genotype calls for SNPs in proximity to regions of copy number change, suggesting that the results from the main analysis are sensible and only a minimal increase in precision could be expected had normal genotypes have been available for all samples.

Additionally, since the ethnic composition of the cohort was mixed, by restricting this analysis to samples of European ancestry (and for which normal genotype calls were available), we indirectly assessed whether population stratification inflated the significance of the eQTLs. Due to the diminished sample size, we anticipated that the number of significant associations would drop substantially (as a given predictor would be required to explain a greater proportion of the variance). However, as noted above, the relative contributions of each predictor remains stable. Since the accuracy of SNP calls, and hence SNP associations would be predominantly affected by non-diploid copy number regions, whereas CNA and CNV associations would be relatively unaffected, these results also suggest that population stratification has not significantly biased the results.

12.7 *eQTL analysis of normal gene expression profiles*

An additional analysis considered a set of 85 individuals of European ancestry for whom genotyping was performed on non-cancerous tissue and gene expression values from matched normal tissue were available (Figure S13, Table S20). The expectation was that strong germline associations would persist in this reduced sample set, while somatic associations would be ablated. As anticipated, the detection power was massively reduced, with an expression association declared for only 6 CNVs and 73 SNPs (compared to 78 and 1,617 from the main analysis). Of the 6 genes with a CNV association, 2 were identified in the main analysis, whereas 63 of the 73 SNP associations were also found by the main analysis, suggesting reasonably high concordance. As illustrated in the variance histograms (Figure S13b), the strongest CNV associated gene, *GSTM1*, is again identified, while *IPO8* also persists in having the strongest SNP association. Reassuringly, expression probes found to have a CNA association have all but disappeared. Of the 5 that remain, 2 were also reported in the first analysis, suggesting only a small element of misclassification of somatic and germline events; the remaining 3 were expected when 28,609 tests are performed with a significance threshold of 0.0001.

12.8 *Trans-acting aberration hotspots*

Several *trans*-acting aberration hotspots that modulated the expression of many mRNAs were identified through the eQTL analyses, including the TCR-deletion hotspots on chromosomes 7 and 14 (Figure 3). As an aside, we note that the chromosome 14 hotspot was also detected when a novel approach to detect sparse regulatory networks [40] was applied to these data. Here, a cutoff of 30 was selected such that CNAs associated with >30 mRNAs were termed a 'hotspot', resulting in 36 regions for both the 3 Mb window and 10 Mb window, respectively (33 of which overlap) (Table S25). We then performed enrichment analysis of the *trans*-associated mRNAs for the 10 Mb window size in order to identify pathways that might be coordinately regulated by a particular CNA hotspot (Table S26). Here, cognate mRNAs with a *P*-value $< 10^{-4}$, were taken to be significant for a given hotspot and were included in the analysis. Enrichment was assessed using the *GOstats* [41] package, where the set of 28,609 genes were used as the universe. Additionally, enrichment maps corresponding to modules enriched amongst *trans* associated mRNAs were generated with cytoscape. Briefly, genes were first projected onto a network using associations encoded in the Reactome FI Cytoscape plugin. The network was then clustered into modules using an edge-betweenness algorithm [42] maximising within module connectivity, while minimising between module connectivity, where pathways enriched (false discovery rate (FDR) < 0.001) for genes clustered in each module are shown. For each pathway, the source database is indicated as follows: B: NCI-PID BioCarta, C: cancer cell map, K: KEGG, N: NCI-PID curated pathways, P: PANTHER, R: Reactome.

13 Analysis of somatic copy number aberrations and gene expression

Motivated by our observation that proximal CNAs are a dominant feature of the tumour expression architecture (Figure 1, Figure 3a), we examined this trend in further detail. In particular, we collapsed the genome to a gene-centric scaffold and performed ANOVA, Kruskal-Wallis, and Spearman correlation tests of association between copy number and expression at the same gene (in *cis*) (Figures S29, S30, Tables S30-S37) and χ^2 tests to determine subtype-specific CNAs within the integrative subgroups (Tables S38, S39). We also examined genome-wide patterns of copy-number and expression correlation (Figure S21). As was observed when the proximal window was set at 3 Mb, many of the top associations localise to genes on chromosomes 1, 8, 11, 16, and 17. Several examples of the relationship between CNAs and expression are illustrated in Figure S14. In particular, subtype-specific CNA-expression correlations are highlighted for *ERBB2* and *RSF1*, which are amplified in HER2/Luminal and Luminal tumours, respectively. The expression of *MAP2K4*, which is deleted in Luminal cases, also shows copy number dependence, whereas *TFF1* exhibits expression differences amongst the subtypes independent of copy number.

13.1 *Subtype-specific analysis of copy number*

A χ^2 (Chi-Squared) test of independence was used to determine subtype association with copy number states derived from CBS or HMM. For each gene g, copy number values were divided into 3 levels: loss (HOMD & HETD), gain (GAIN & AMP/HLAMP) and neutral (NEUT). For the omnibus test, patients were divided into n (number of subtypes) levels; in the case of PAM50 subtypes, there are 5 (Basal, HER2, LumA, LumB, Normal) levels. For the subtype-specific test, patients were divided into 2 levels, subtype-of-interest or not. For these analyses, we employed the list of genes on the Illumina HT-12 v3 array that overlap with protein-coding genes from Ensembl 54 (hg18) (18,733; 18,670 excluding Y chromosome probes). Genes with insufficient representation in at least 2 copy number groupings (with a threshold of 10 cases) to perform the χ^2 test (in the case of 3 groups) or Wilcoxon test (in the case of 2 groups) were excluded.

13.2 *Cis-acting copy number and differential expression associations between groups*

We collapsed the genome to a gene-based scaffold and investigated whether genes are differentially expressed in the tumours when samples are grouped based on predicted copy number states (derived from either CBS or HMM) as captured in the call matrix, C. To this end, we used several hypothesis-testing methods, namely ANOVA/Tukey-HSD and pairwise Student's t tests (parametric) and Kruskal-Wallis/Wilcoxon rank sum tests (non-parametric). For each gene, samples were grouped into 3 levels based on their copy number alteration call: loss (HOMD & HETD), gain (GAIN & AMP/HLAMP) and neutral (NEUT). Then for each gene g and each probe $p \in P_g$, tests of differential expression were performed using Kruskal-Wallis and ANOVA between the groups defining the null hypothesis as drawing samples from the same population or as having the same rank means between the 3 groups, respectively. Often, genes were not altered with sufficient frequency ($< 1\%$) in one level, hence reducing the problem to a Student's t and Wilcoxon rank sum test for 2 groups. If the combined frequencies for both gains and losses were below 1%, then no hypothesis test was performed. Multiple testing correction was applied to all P-values using the Bonferroni method. For these analyses, we report the set of features given by Ensembl that correspond to Illumina HT-12 v3 probes (30,566) or those that map uniquely (29,033).

13.3 *Cis copy number and expression correlation analysis*

We also investigated the effect of copy number alterations on gene expression using Spearman's rank correlation. Because the full enumeration of somatic and germline events is available, genes altered by CNAs and CNVs were analysed separately for both CBS and HMM. Spearman's rank correlation tests were performed on each gene using the median log-ratios in the log matrix L and $\log_2$ expression values across samples.

Because each gene g on the Illumina HT-12 v3 array may contain multiple probes $P_g = \{p_1, \ldots, p_i\}$, we assign gene g with the correlation coefficient, ρ (rho), and P-value from the probe $p\prime \in P_g$ with minimum P-value,

$$p\prime = \underset{p \in P_g}{\operatorname{argmin}}\{pvalue(p)\} \tag{13}$$

There may be multiple $p\prime$ in the case of ties; thus, we select the probe

$$p\prime\prime = \underset{p \in p\prime}{\operatorname{argmax}}\{rho(p)\} \tag{14}$$

with max ρ to break the tie. P-values were corrected for multiple-testing using the Bonferroni adjustment. The list of 18,670 genes on the Illumina HT-12 v3 array that overlap with protein-coding genes from Ensembl 54 (hg18) was employed in this analysis.

13.4 *Genome-wide copy number and expression correlation analysis*

We also investigated genome-wide patterns of correlation between copy number aberrations and gene expression. Given expression and copy number data for 18,733 genes across 997 patients, we computed the

Pearson correlation between pairs of genes using the median log-ratios in log matrix L and $\log_2$ expression values across patients. CNVs were ignored in this analysis and a 18,733 × 18,733 matrix with each element representing the correlation between copy number aberrations and expression of two genes was generated. Correlation P-values were adjusted for multiple hypothesis testing using the Benjamini-Hochberg method [43]. Observations with P-values > 0.05 were excluded and the resulting correlation matrix was used for downstream analysis. The matrix of copy number and expression correlation profiles were plotted for both CBS and HMM-derived CNAs (Figure S21). Enrichment maps corresponding to modules enriched amongst *trans* associated mRNAs were generated with cytoscape. Briefly, genes were first projected onto a network using associations encoded in the Reactome FI Cytoscape plugin. As in the eQTL analyses, the network was then clustered into modules using an edge-betweenness algorithm maximising within module connectivity, while minimising between module connectivity, where pathways enriched (FDR < 0.001) for genes clustered in each module are shown. For each pathway, the source database is indicated as follows: B: NCI-PID BioCarta, C: cancer cell map, K: KEGG, N: NCI-PID curated pathways, P: PANTHER, R: Reactome.

14 Cluster analysis

14.1 *Integrative clustering*

Given that additional heterogeneity was evident within the intrinsic subgroups (Figure S31), and that the analysis of paired copy number and gene expression data suggested subtype-specific CNA-driven expression changes in *cis*, we sought to investigate whether this relationship could be used to further delineate breast cancer subgroups. To this end, we employed an integrative clustering framework (iCluster) described by Shen *et al.* [44] and explored various feature selection paradigms. The integrative clustering approach is based on a joint latent variable model that flexibly handles the associations between multiple data types, and reduces the dataset dimensionality. In particular, the K-Means approach is framed as a Gaussian joint latent variable model, where the maximum likelihood estimate (MLE) of the latent tumour subtypes is obtained through the expectation-maximisation (EM) algorithm. Following EM convergence, the class indicator matrix is obtained by performing K-Means on the resultant estimates. By penalising the log-likelihood with an L_1-norm regularisation parameter, λ, feature selection is inherent to the process and a sparse solution to the clustering is obtained.

Here, we focus on the the use of *cis*-acting genes that exhibit significant associations with CNAs (CBS-derived) across the entire cohort of tumours as determined by gene-centric ANOVA, as this reflects our assumption that copy number aberrations influence phenotypes through concomitant changes in expression. In particular, the 1,000 most significant Bonferroni adjusted P-value probes from the set of 30,566 Illumina probes (Table S30) were employed. Both copy number segmented means and normalised gene expression values for the selected features were used as input to the clustering, after removing copy number probes that exhibited high correlation and poorly annotated probes (caution was exercised in removing probes annotated to contain SNPs as these probes can interrogate key transcripts, and the frequency of the SNP is often sufficiently low in the population so not to impact the analysis). We examined differing numbers of clusters $k \in [2, 18]$, where the sparsity parameter λ was determined by optimising the proportion of deviance (POD), defined by Shen *et al.* [44]. Selecting an appropriate number of features is somewhat arbitrary, so here we empirically assessed the adjusted Rand index (a measure of similarity between clusterings) for different integrative clustering runs with $g = 500$, 1,000, or 2,000 features as input. We observed that $g = 1{,}000$ allowed for a sufficient number of features without saturating the analysis, hence this feature set was employed in the subsequent analyses.

As noted by Handl *et al.* [45] cluster analysis is a complicated and interactive process, and no entirely reliable method exists to identify the number of clusters in a dataset. Hence cluster analysis should always be performed for a sensible range of clusters as this facilitates an understanding of the operations of the algorithm and the identification of trends in the data. Furthermore, most internal cluster validity measures exhibit bias with respect to the underlying structure of the partitioning [45]. We evaluated the optimal number of clusters using Dunn's index [46], which defines the ratio of the smallest distance between observations not in the same cluster to the largest intra-cluster distance, and thus assesses both intra-cluster homogeneity and inter-cluster separation. In particular, Dunn's index was computed for the complete set of features used as input for integrative clustering using the Euclidean distance metric. Evaluation of the indices for $k = 2, \ldots, 15$, indicated a global maximum at $k = 14$ (ignoring the $k = 2$ peak) and a local maximum at $k = 10$ (Figure S23a). The adjusted Rand index (ARI) was computed for all pairwise comparisons of $k = 2, \ldots, 18$. Here we observe a relatively high adjusted Rand index for the $k = 10$ versus $k = 9$, 13, or 14 result, whereas that for $k = 11$ or 12 was lower (Figure S23b). We also inspected the similarities and differences in terms of cluster membership and genome-wide copy number profiles for alternate values of k, focusing in detail on $k = 9$, 13, 14 relative to the $k = 10$ solution. In particular, cross tabulation of the sample membership for the two clusterings indicate that the $k = 9$ partition exhibits a high degree of overlap with the $k = 10$ IntClust 1, 2, 4, 5, 6, 7, 8, and 9 (Figure S23c). However, IntClust 3 is split between several groups in the $k = 9$ clustering as is IntClust 10, which represented predominantly Basal-like tumours. Examination of the genome-wide copy number profiles associated with $k = 9$ suggest that they are less homogeneous than those observed for $k = 10$. A comparison of the $k = 10$ versus $k = 13$ result suggests that the 10 clusters (for the $k = 10$ solution) are generally recovered, and that the 3 new clusters labelled 6, 7, and 9 for $k = 13$ are not clearly distinct. For example, subgroup 6 is a mixture of IntClust 4 and IntClust 10 in the $k = 10$ grouping, and is similar to subgroup 5 for the $k = 13$ result. Similarly, subgroup 7 ($n = 34$) is composed primarily of IntClust 4 members, whereas group 9 ($n = 56$) is composed of IntClust 3 and IntClust 8, but exhibits a similar copy number profile to subgroup 3. Thus, $k = 10$ appears to capture the major features of the data observed for $k = 13$. A comparison of $k = 10$ and $k = 14$ yielded similar findings. Again, while the 10 clusters (for the $k = 10$ solution) are generally recovered, the 4 new clusters labelled 3, 5, 6, and 14 are not clearly distinct. Here, subgroup 3 ($n = 87$) is a mixture of IntClust 4 and IntClust 10 in the $k = 10$ grouping and is quite similar to subgroup 11. Subgroup 5 ($n = 41$) is a mixture of IntClust 6 and IntClust 9, whereas subgroup 6 is a mixture of IntClust 8 and IntClust9 ($n = 60$), and subgroup 14 is composed of multiple IntClust groupings, but primarily IntClust 3 and IntClust 8 and has a similar profile to subgroup 13. Thus, $k = 10$ also appears to capture the major features of the $k = 14$ clustering, and represents a conservative solution.

A similar analysis was repeated for the HMM-derived CNAs (Figure S24). While many of the features of the CBS-derived integrative clustering results are also observed in this analysis, we note the absence of the 11q13/14 *cis*-acting subgroup for this partition. Thus, while alternative clusterings exist, we have presented one possible representation of the data that captures many salient features, providing insight into breast cancer biology. Perhaps most importantly, the clusters identified were found to reproduce in an external dataset as described below.

In order to evaluate the reproducibility of the clustering results on an external dataset, we trained a classifier based on the 10 integrative cluster subgroups identified in the original dataset of 997 patients using the nearest shrunken centroids method described in Tibshirani *et al.* [47] and implemented in the Prediction Analysis of Microarrays (PAMR) software. Given that the integrative clustering approach has an inherent feature selection step, we used the set of 754 features selected by this method (where expression values were scaled) (Tables S42, S43) to build the PAM classifier, as any reduction significantly increased the error rate of the algorithm. Using a threshold value of 0.582 for shrinkage, the cross-validation error with 44 folds (as the smallest class size was $n = 44$) yielded a value of 0.087. Based on this set of centroids, we then classified the validation set of 995 tumours into the 10 groups (Table S44) and compared several attributes, such as the proportion of samples belonging to each group, their genome wide copy number profiles, Kaplan-Meier survival curves, and Cox models stratified by site with the following variables: grade, tumour size (modelled with a spline), number of lymph nodes (modelled with a spline), age at diagnosis (modelled with a spline), and integrative cluster membership (Figures 4, 5). Similar molecular characteristics were observed for each of the 10 clusters in both datasets, despite the fact that the validation set included low cellularity tumours. Moreover, Cox regression analysis shows similar fits for both datasets and the 95% confidence interval of the hazard ratios overlap (Table S40).

The identification of reproducible clusters in an external dataset is related to prediction accuracy, which can be defined as the proportion of data whose predicted classifications are identical to the true classifications. Based on this relationship, Kapp and Tibshirani [48] define a procedure for the validation of clusters in a validation dataset, and a measure of cluster quality termed the in-group proportion (IGP), where IGP refers to the proportion of samples in a group whose nearest neighbours are also in the same group. Validation of groups found in an external dataset is performed by comparison against a null distribution of IGPs as implemented in the *clusterRepro* package in R, to provide a measure of individual cluster significance (as the IGP depends on the proportion of individuals belonging to each cluster and cluster sizes vary, the *P*-value should be used to assess cluster quality). These results demonstrate the reproducibility of the 10 integrative clusters, as all are deemed to be high-quality, although we note that the *P*-value associated with IntClust 3 is borderline significant. Thus the integrative clusters discovered in the training set can be reproduced in a validation set of tumours.

It is worth noting that *de novo* clustering on a separate dataset does not address cluster reproducibility, but rather asks what can also be *discovered* in another cohort. Such an approach cannot be disentangled from the clustering procedure. It is not an approach that is powered for the validation of smaller patient sub-groupings because the inherent variability of clustering will not reveal all features in all subgroups, and will likely reveal additional groups. It is also not used in clinical practice, as one typically would want to assign a new (single) patient to a subgroup as can be done using a classifier. For this reason, internal cross validation and external validation on a separate dataset are the accepted methods for demonstrating the reproducibility of subgroups discovered by unsupervised clustering. Despite the limitations of this approach, *de novo* clustering can be performed on the validation set employed in this study. In doing so, we again observe that 10 represents the optimal number of clusters, confirming the clustering parameters (Figure S27). We also examined their molecular profiles and outcomes (Figure S27, Table S33), and note that the majority of the clusters are identified and have similar molecular profiles, although two of the smaller clusters originally identified in the discovery set (IntClust 2 and 6) are not well represented, presumably because the clustering favours a minimal number of samples per subgroup. A comparison with the predicted validation set assignments indicates reasonable agreement with 71.55% concordance. Analogous to the cluster validation procedure we describe above, one can build a classifier based on the features derived from integrative clustering and the cluster memberships. Using a threshold value of 0 for shrinkage, cross-validation error of 0.085 was obtained. Applying PAMR to the validation set we classified the discovery set and obtain 72.4% concordance with the *de novo* cluster membership in this dataset. All but the two smallest (IntClust 2 and 6) integrative subgroups were represented (as expected), and found to be reproducible based on the IGP. Thus, despite the fact that this approach would not be performed in practice to demonstrate subtype reproducibility, it recapitulates all of the larger clusters identified in the discovery set and highlight the robustness of these clusters since the validation set includes low cellularity tumours, which may influence feature selection and the clustering.

The reproducibility of these integrative subgroups, which are characterised by distinct outcomes, suggests that by integrating multiple genomic features it may be possible to derive more robust patient classifiers that overcome the limitations of expression-only approaches. However, the development of a predictive signature is beyond the scope of this study. There are important conceptual differences between unsupervised class discovery methods, such as that applied here, and supervised clinical outcome prediction models or multigene prognostic signatures. The fundamental difference is that the latter employs supervised class prediction methods that take into account the known clinical outcome of cases during development of the predictor. These tasks are typically the subject of separate studies since they address distinct biological questions and utilise the data in very different ways. Nonetheless, it is possible to compare the predictive value of multivariable Cox models that include the integrative subtypes with the standard clinico-pathological variables using the concordance index (C-index) [49, 50]. Briefly, the multivariable Cox proportional hazards models were as follows (see also *Survival analysis* below): the basic model (m.basic) with clinical variables included the following components: Grade (numerical, linear), Size (numerical, spline), Lymph nodes (numerical, spline), and Age (numerical, spline), stratified by site; the PAM50 intrinsic subtype model is m.basic + Pam50Subtype, and the integrative cluster model is m.basic + IntClust. Cox models based on disease-specific survival were built for all cases as well as separately for ER-positive and ER-negative cases (as the proportional hazards assumption does not hold for ER status). The C-index computed for the discovery set indicates that the IntClust model has greater predictive power than the standard clinico-pathological variables alone and than the PAM50 model (Table S41a). For the validation set, the IntClust model offers improved predictive power over the standard clinico-pathological variables, and performs comparably to the PAM50 model (Table S41b).

14.2 *K-Means clustering*

We designed various feature sets for the K-Means clustering and examined different number of clusters $k \in [5, 15]$. Essentially, we employed gene expression profiles as features in the K-Means algorithm, and investigated different methods for identifying the most informative genes:

- **Entropy.** Select those genes for which the entropy of the copy number profile is high, i.e. above half of the entropy of a uniform distribution.
- **Outlying.** Select those genes which have been identified as outliers for extreme copy number states, based on the *Detection of outlier expression profiles* described below. We filtered the outlying genes based on the threshold of average count per gene $\tau = \frac{\sum_{g \in O_k} \#(g,k)}{|O_k|}$ where O_k is the set of outlier genes for the copy number state k, and $\#(g, k)$ is the frequency of the outliers for gene g in the copy number state k.
- ***cis*-acting.** Select *cis*-acting genes which exhibit high (absolute) correlation between copy number and expression, where the correlation is computed as described in the section on *Correlation analysis*. Here we employed the top 2,000 genes with the highest absolute Spearman correlation value.
- **Kurtosis.** Select the top 2,000 genes with the least kurtosis.

Comparison of these methods will be the subject of another contribution. For this paper, we report the results of the outlying features (Figure S25, Table S32). We note that while this approach captured some of the structure of the population, several key clusters identified based on integrative clustering were not observed.

15 Detection of outlier expression profiles

To identify genes whose expression was driven in *cis* by extreme copy number events, Gaussian distributions were fitted to the 997 $\log_2$ expression values for each gene across the cohort, using the maximum likelihood estimates of the mean and variance. For genes with multiple probes, we summarized the expression level using the median. This resulted in 18,733 Gaussian distributions, each corresponding to the expression profile of a gene. We then sought to capture (gene, sample) events where large amplitude copy number aberrations (such as putative homozygous deletions and high-level amplifications) drive values into the extreme tails of the expression profile. Outlier events in the 5% right tail whose copy number state was either amplification or high-level amplification (AMP & HLAMP) as determined by HMM-Dosage and those events in the 5% left tail whose copy number state was homozygous deletion (HOMD) were reported along with the frequency of copy number aberration at that gene as determined by HMM and CBS (Tables S23, S24). We also illustrate the patterns of CNA-driven outlying expression at select regions throughout the genome in Figures S15 and S16.

16 Pathway analysis

Multiple tools and databases were employed to project genes that were significantly differentially expressed in a particular tumour subgroup relative to normal breast tissue (see the section on Gene expression analysis) onto pathways. Enrichment for Gene Ontology groups was performed using the NIH DAVID Bioinformatics resource (v6.7) [51, 52] and visualised with the Cytoscape software (v2.6.3) [53]. Pathway analysis was performed with Ingenuity Pathway Analysis software (v8.7) (Ingenuity Systems, www.ingenuity.com). The MetaCore (GeneGO) software, KEGG Pathways [54] and Wikipathways [55] databases were also used to cross-check the pathway components.

Genes with a $\log_2$ ratio > 0.58 and an FDR adjusted P-value < 0.01, and that were associated with a canonical pathway in the Ingenuity Knowledge Base of canonical pathways and curated gene sets, were identified. The significance of the association between the data set and a canonical pathway was determined by computing the ratio of the number of features from the data set that map to the pathway divided by the total number of features that map to the canonical pathway. The probability that the association between the genes in the dataset and the canonical pathway could be explained by chance alone was assessed using Fisher's exact test. The results of differential expression analysis and pathway enrichment in the integrative subgroups are presented in Tables S46 and S47, respectively.

The relationships between molecules were represented graphically as a network, where molecules are represented as nodes, and the biological relationship between two nodes is represented as an edge (line). All edges are supported by at least one reference from the literature or canonical information stored in the Ingenuity Pathways Knowledge Base. Human, mouse, and rat orthologues are stored as separate objects in the Ingenuity Pathways Knowledge Base, but are represented as a single node in the network. The intensity of the node colour indicates the degree of up (red) or down (blue) regulation for a particular contrast. Nodes were displayed using various shapes that represent the functional class of the gene product. As described in the text, several subtype-specific gene networks were evident from pathway analysis of the CNA-expression landscape (Figures S36-S39). Note that for pathways that exhibited a major *trans*-acting component as determined by either the eQTL or correlation analyses, the relevant molecules are indicated with a green outline (Figures S36, S38).

17 Survival analysis

We fit a series of Cox regression models for both disease-specific and overall survival using the *survival* [56] and *Design* [49,50] packages in R with several different perspectives in mind. First, we fit a global model for breast cancer specific survival for all 997 samples in the discovery cohort (for which 17 were removed due to missing values as they were lost to follow-up), including the most relevant clinical variables as covariates (grade, size, age at diagnosis, number of lymph nodes positive, ER status, and integrative cluster subtype). Models were also fitted to the validation cohort of 995 cases to evaluate cluster reproducibility in the context of clinical outcome (here 3 cases were lost to follow-up). Other variables were also missing for some cases. Analyses were stratified by tissue bank (site) to account for differences in the basal hazard due to geographical locations. We included spline terms for more complex numerical variables (tumour size, age at diagnosis, and number of lymph nodes positive). The residual analysis showed that ER status also violated the proportional hazards assumption, but the effect of this variable did not seem to be relevant and it was retained. For the analysis of genomic instability (GI), the effects for the GI index (jump) were added both globally and individually for each of the chromosome arms, keeping the significant ones (P-value < 0.05) for the multivariate model. Likelihood ratio tests were used to compare the models of clinical subtype, and combined clinical and molecular variables. Confidence intervals are reported for each of the variables, and differences in the survival distributions were also assessed using Kaplan-Meier curves and the non-parametric logrank test implemented in the *survival* package in R.

References

[1] Zhou, W. *et al.* Full sequencing of TP53 identifies identical mutations within in situ and invasive components in breast cancer suggesting clonal evolution. *Mol Oncol* **3**, 214–219 (2009).

[2] Fridlyand, J. *et al.* Breast tumor copy number aberration phenotypes and genomic instability. *BMC Cancer* **6**, 96 (2006).

[3] R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria (2009).

[4] Dunning, M. J., Smith, M. L., Ritchie, M. E. & Tavaré, S. beadarray: R classes and methods for Illumina bead-based data. *Bioinformatics* **23**, 2183–2184 (2007).

[5] Cairns, J. M., Dunning, M. J., Ritchie, M. E., Russell, R. & Lynch, A. G. BASH: a tool for managing BeadArray spatial artefacts. *Bioinformatics* **24**, 2921–2922 (2008).

[6] Asare, A. L., Gao, Z., Carey, V. J., Wang, R. & Seyfert-Margolis, V. Power enhancement via multivariate outlier testing with gene expression arrays. *Bioinformatics* **25**, 48–53 (2009).

[7] Barbosa-Morais, N. L. *et al.* A re-annotation pipeline for Illumina BeadArrays: improving the interpretation of gene expression data. *Nucleic Acids Res* **38**, e17 (2010).

[8] Smyth, G. K. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Stat Appl Genet Mol Biol* **3**, Article3 (2004).

[9] Perou, C. M. *et al.* Molecular portraits of human breast tumours. *Nature* **406**, 747–752 (2000).

[10] Sorlie, T. *et al.* Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proc Natl Acad Sci U S A* **98**, 10869–10874 (2001).

[11] Parker, J. S. *et al.* Supervised risk predictor of breast cancer based on intrinsic subtypes. *J Clin Oncol* **27**, 1160–1167 (2009).

[12] Dunning, M. J. *et al.* The importance of platform annotation in interpreting microarray data. *Lancet Oncol* **11**, 717 (2010).

[13] Sørlie, T. *et al.* The importance of gene-centring microarray data. *Lancet Oncol* **11**, 719–720 (2010).

[14] Haibe-Kains, B. *et al.* A fuzzy gene expression-based computational approach improves breast cancer prognostication. *Genome Biol* **11**, R18 (2010).

[15] Haibe-Kains, B., Bontempi, G. & Sotiriou, C. *genefu: Relevant Functions for Gene Expression Analysis, Especially in Breast Cancer.* (2010).

[16] Fraley, C. & Raftery, A. E. Model-based clustering, discriminant analysis and density estimation. *Journal of the American Statistical Association* **97**, 611–631 (2002).

[17] Fraley, C. & Raftery, A. E. MCLUST version 3 for R: Normal mixture modeling and model-based clustering. Technical Report 504, University of Washington, Department of Statistics (2006).

[18] Lehmann, B. D. *et al.* Identification of human triple-negative breast cancer subtypes and preclinical models for selection of targeted therapies. *J Clin Invest* **121**, 2750–2767 (2011).

[19] Carvalho, B., Bengtsson, H., Speed, T. P. & Irizarry, R. A. Exploration, normalization, and genotype calls of high-density oligonucleotide SNP array data. *Biostatistics* **8**, 485–499 (2007).

[20] Irizarry, R. A., Carvalho, B. S., Scharpf, R. & Ritchie, M. *crlmm: Genotype Calling (CRLMM) and Copy Number Analysis tool for Affymetrix SNP 5.0 and 6.0 and Illumina arrays.* (2010).

[21] Venkatraman, E. S. & Olshen, A. B. A faster circular binary segmentation algorithm for the analysis of array CGH data. *Bioinformatics* **23**, 657–663 (2007).

[22] Shah, S. P. *et al.* Integrating copy number polymorphisms into array CGH analysis using a robust HMM. *Bioinformatics* **22**, e431–9 (2006).

[23] Bengtsson, H., Wirapati, P. & Speed, T. P. A single-array preprocessing method for estimating full-resolution raw copy numbers from all Affymetrix genotyping arrays including GenomeWideSNP 5 & 6. *Bioinformatics* **25**, 2149–2156 (2009).

[24] Smith, M. L., Marioni, J. C., McKinney, S., Hardcastle, T. & Thorne, N. P. *snapCGH: Segmentation, normalisation and processing of aCGH data.* (2009).

[25] Seshan, V. E. & Olshen, A. *DNAcopy: DNA copy number data analysis* (2010).

[26] Ostrovnaya, I., Nanjangud, G. & Olshen, A. B. A classification model for distinguishing copy number variants from cancer-related alterations. *BMC Bioinformatics* **11**, 297 (2010).

[27] Willenbrock, H. & Fridlyand, J. A comparison study: applying segmentation to array CGH data for downstream analyses. *Bioinformatics* **21**, 4084–4091 (2005).

[28] Rouveirol, C. *et al.* Computation of recurrent minimal genomic alterations from array-CGH data. *Bioinformatics* **22**, 849–856 (2006).

[29] Conrad, D. F. *et al.* Origins and functional impact of copy number variation in the human genome. *Nature* **464**, 704–712 (2010).

[30] Bignell, G. R. *et al.* Signatures of mutation and selection in the cancer genome. *Nature* **463**, 893–898 (2010).

[31] Wellcome Trust Case Control Consortium *et al.* Genome-wide association study of CNVs in 16,000 cases of eight common diseases and 3,000 shared controls. *Nature* **464**, 713–720 (2010).

[32] Habermann, J. K. *et al.* The gene expression signature of genomic instability in breast cancer is an independent predictor of clinical outcome. *Int J Cancer* **124**, 1552–1564 (2009).

[33] Chin, S. F. *et al.* High-resolution aCGH and expression profiling identifies a novel genomic subtype of ER negative breast cancer. *Genome Biol* **8**, R215 (2007).

[34] Stranger, B. E. *et al.* Relative impact of nucleotide and copy number variation on gene expression phenotypes. *Science* **315**, 848–853 (2007).

[35] Kruskal, W. H. A nonparametric test for the several sample problem. *The Annals of Mathematical Statistics* **23**, 525–540 (1952).

[36] Šidák, Z. Rectangular confidence regions for the means of multivariate normal distributions. *American Statistical Association* **62**, 626–633 (1967).

[37] Devlin, B. & Roeder, K. Genomic control for association studies. *Biometrics* **55**, 997–1004 (1999).

[38] Storey, J. D. & Tibshirani, R. Statistical significance for genomewide studies. *Proc Natl Acad Sci U S A* **100**, 9440–9445 (2003).

[39] Storey, J. D. The positive false discovery rate: a Bayesian interpretation and the q-value. *Annals of Statistics* **32**, 2013–2035 (2003).

[40] Yuan, Y., Curtis, C., Caldas, C. & Markowetz, F. A sparse regulatory network of copy-number driven gene expression reveals putative breast cancer oncogenes. *IEEE/ACM Trans Comput Biol Bioinform* (2011).

[41] Falcon, S. & Gentleman, R. Using GOstats to test gene lists for GO term association. *Bioinformatics* **23**, 257–258 (2007).

[42] Wu, G., Feng, X. & Stein, L. A human functional protein interaction network and its application to cancer data analysis. *Genome Biol* **11**, R53 (2010).

[43] Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society* **B**, 289–300 (1995).

[44] Shen, R., Olshen, A. B. & Ladanyi, M. Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis. *Bioinformatics* **25**, 2906–12 (2009).

[45] Handl, J., Knowles, J. & Kell, D. B. Computational cluster validation in post-genomic data analysis. *Bioinformatics* **21**, 3201–12 (2005).

[46] Dunn, J. C. Well separated clusters and optimal fuzzy partitions. *Journal of Cybernetics* **4**, 95–104 (1974).

[47] Tibshirani, R., Hastie, T., Narasimhan, B. & Chu, G. Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proc Natl Acad Sci U S A* **99**, 6567–6572 (2002).

[48] Kapp, A. V. & Tibshirani, R. Are clusters found in one dataset present in another dataset? *Biostatistics* **8**, 9–31 (2007).

[49] Harrell, F. E. *Regression modeling strategies: with applications to linear models, logistic regression, and survival analysis* (Springer, New York, 2001).

[50] Harrell, F. E. *Design: Design Package* (2009).

[51] Huang, D. W., Sherman, B. T. & Lempicki, R. A. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nature Protocols* **4**, 44–57 (2009).

[52] Dennis, G., Jr *et al.* DAVID: Database for annotation, visualization, and integrated discovery. *Genome Biol* **4**, P3 (2003).

[53] Cline, M. S. *et al.* Integration of biological networks and gene expression data using Cytoscape. *Nat Protoc* **2**, 2366–2382 (2007).

[54] Kanehisa, M., Goto, S., Furumichi, M., Tanabe, M. & Hirakawa, M. KEGG for representation and analysis of molecular networks involving diseases and drugs. *Nucleic Acids Res* **38**, D355–360 (2010).

[55] Pico, A. R. *et al.* Wikipathways: pathway editing for the people. *PLoS Biol* **6**, e184 (2008).

[56] Therneau, T. & original R port by Thomas Lumley. *survival: Survival analysis, including penalised likelihood.* (2009).

18 Supplementary Figures

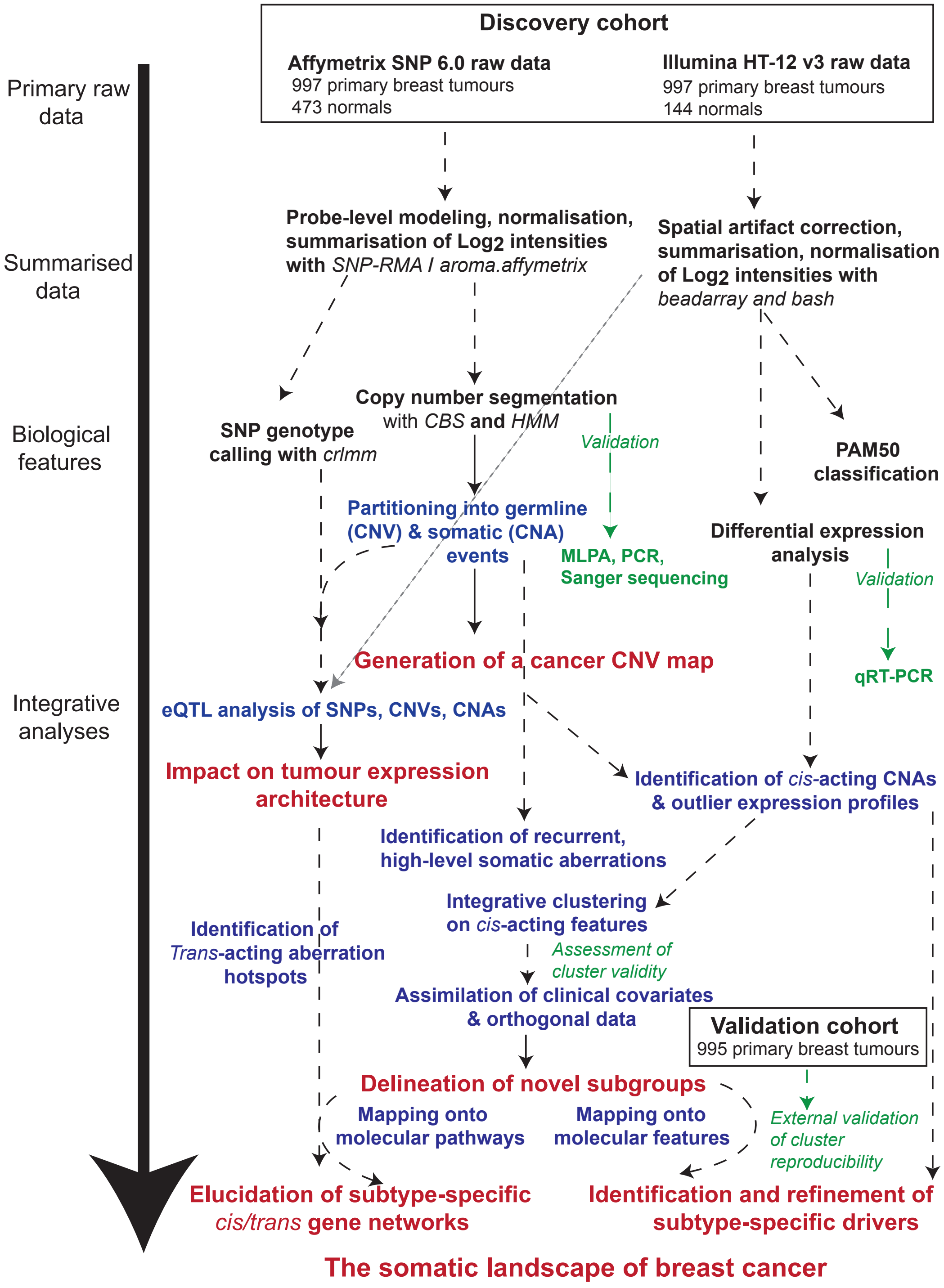
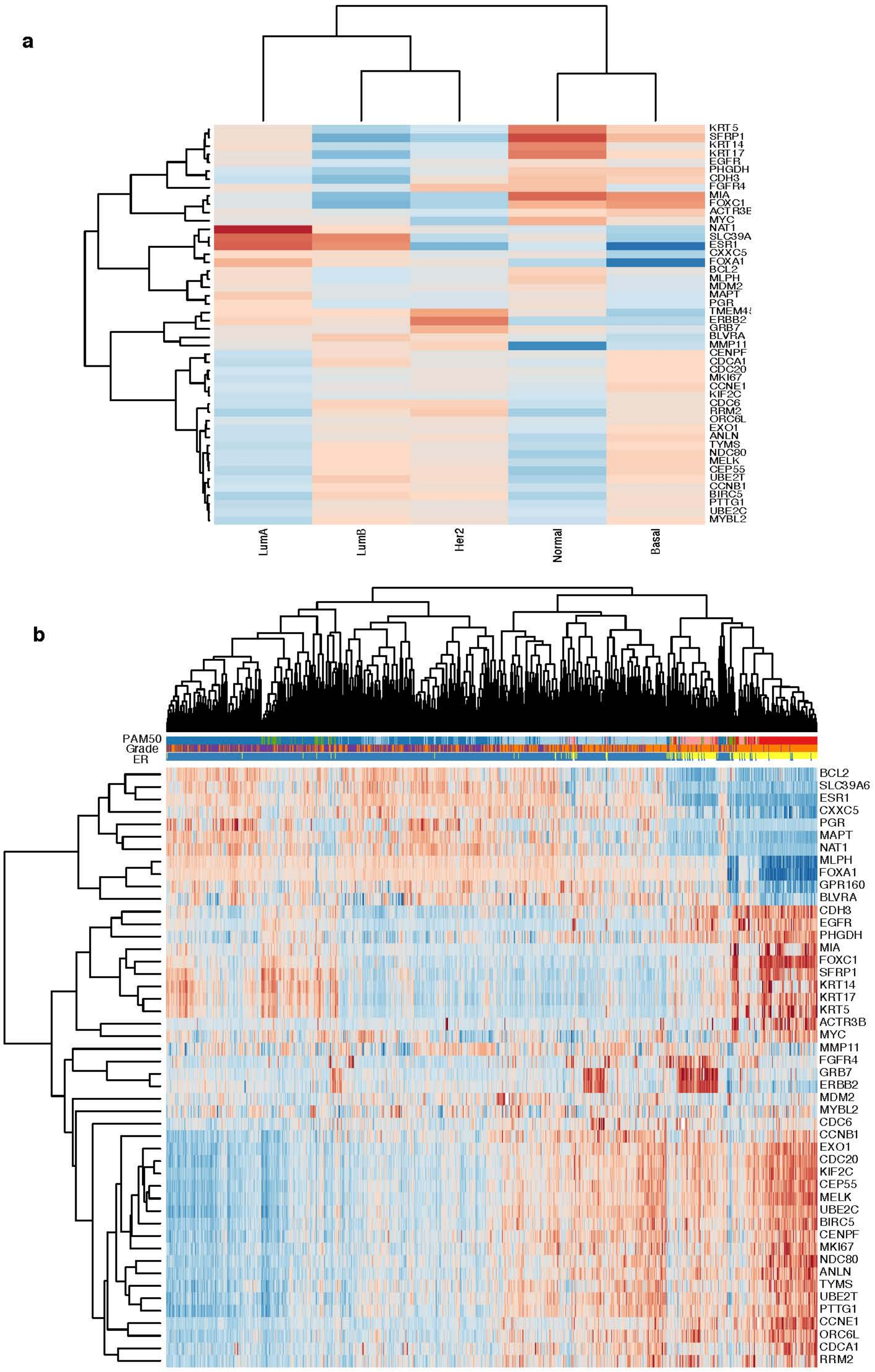

Primary raw data
Summarised data
Biological features
Integrative analyses
Discovery cohort
Affymetrix SNP 6.0 raw data
997 primary breast tumours
473 normals
Illumina HT-12 v3 raw data
997 primary breast tumours
144 normals
Probe-level modeling, normalisation, summarisation of Log2 intensities with SNP-RMA / aroma.affymetrix
Spatial artifact correction, summarisation, normalisation of Log2 intensities with beadarray and bash
SNP genotype calling with crlmm
Copy number segmentation with CBS and HMM
Validation
PAM50 classification
Partitioning into germline (CNV) & somatic (CNA) events
MLPA, PCR, Sanger sequencing
Differential expression analysis
Validation
qRT-PCR
Generation of a cancer CNV map
eQTL analysis of SNPs, CNVs, CNAs
Impact on tumour expression architecture
Identification of cis-acting CNAs & outlier expression profiles
Identification of recurrent, high-level somatic aberrations
Integrative clustering on cis-acting features
Identification of Trans-acting aberration hotspots
Assessment of cluster validity
Assimilation of clinical covariates & orthogonal data
Validation cohort
995 primary breast tumours
Delineation of novel subgroups
Mapping onto molecular pathways
Mapping onto molecular features
External validation of cluster reproducibility
Elucidation of subtype-specific cis/trans gene networks
Identification and refinement of subtype-specific drivers
The somatic landscape of breast cancer

Figure S1: Workflow describing the main analytical approaches used in this study.


a
KRT5
SFRP1
KRT14
KRT17
EGFR
PHGDH
CDH3
FGFR4
MIA
FOXC1
ACTR3E
MYC
NAT1
SLC39A
ESR1
CXXC5
FOXA1
BCL2
MLPH
MDM2
MAPT
PGR
TMEM4!
ERBB2
GRB7
BLVRA
MMP11
CENPF
CDCA1
CDC20
MKI67
CCNE1
KIF2C
CDC6
RRM2
ORC6L
EXO1
ANLN
TYMS
NDC80
MELK
CEP55
UBE2T
CCNB1
BIRC5
PTTG1
UBE2C
MYBL2
LumA
LumB
Her2
Normal
Basal
b
PAM50
Grade
ER
BCL2
SLC39A6
ESR1
CXXC5
PGR
MAPT
NAT1
MLPH
FOXA1
GPR160
BLVRA
CDH3
EGFR
PHGDH
MIA
FOXC1
SFRP1
KRT14
KRT17
KRT5
ACTR3B
MYC
MMP11
FGFR4
GRB7
ERBB2
MDM2
MYBL2
CDC6
CCNB1
EXO1
CDC20
KIF2C
CEP55
MELK
UBE2C
BIRC5
CENPF
MKI67
NDC80
ANLN
TYMS
UBE2T
PTTG1
CCNE1
ORC6L
CDCA1
RRM2

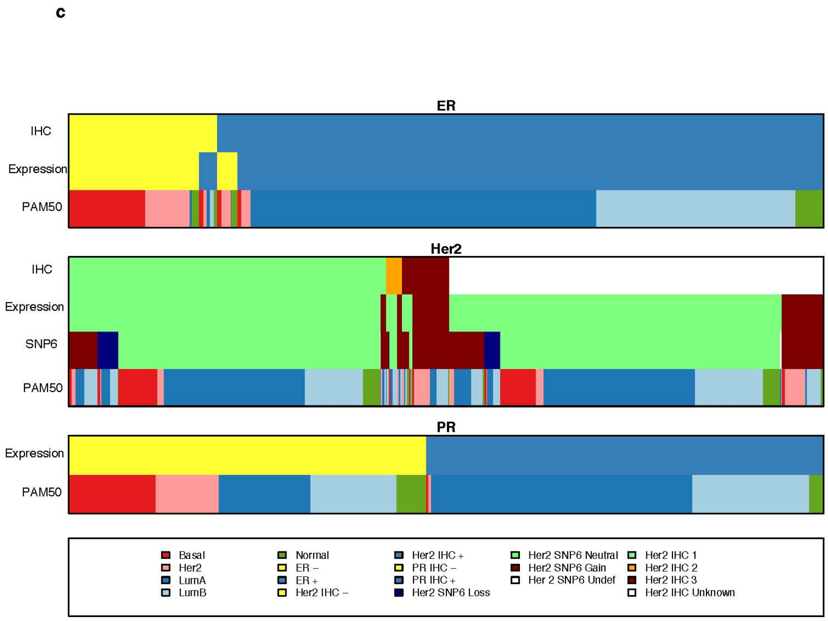
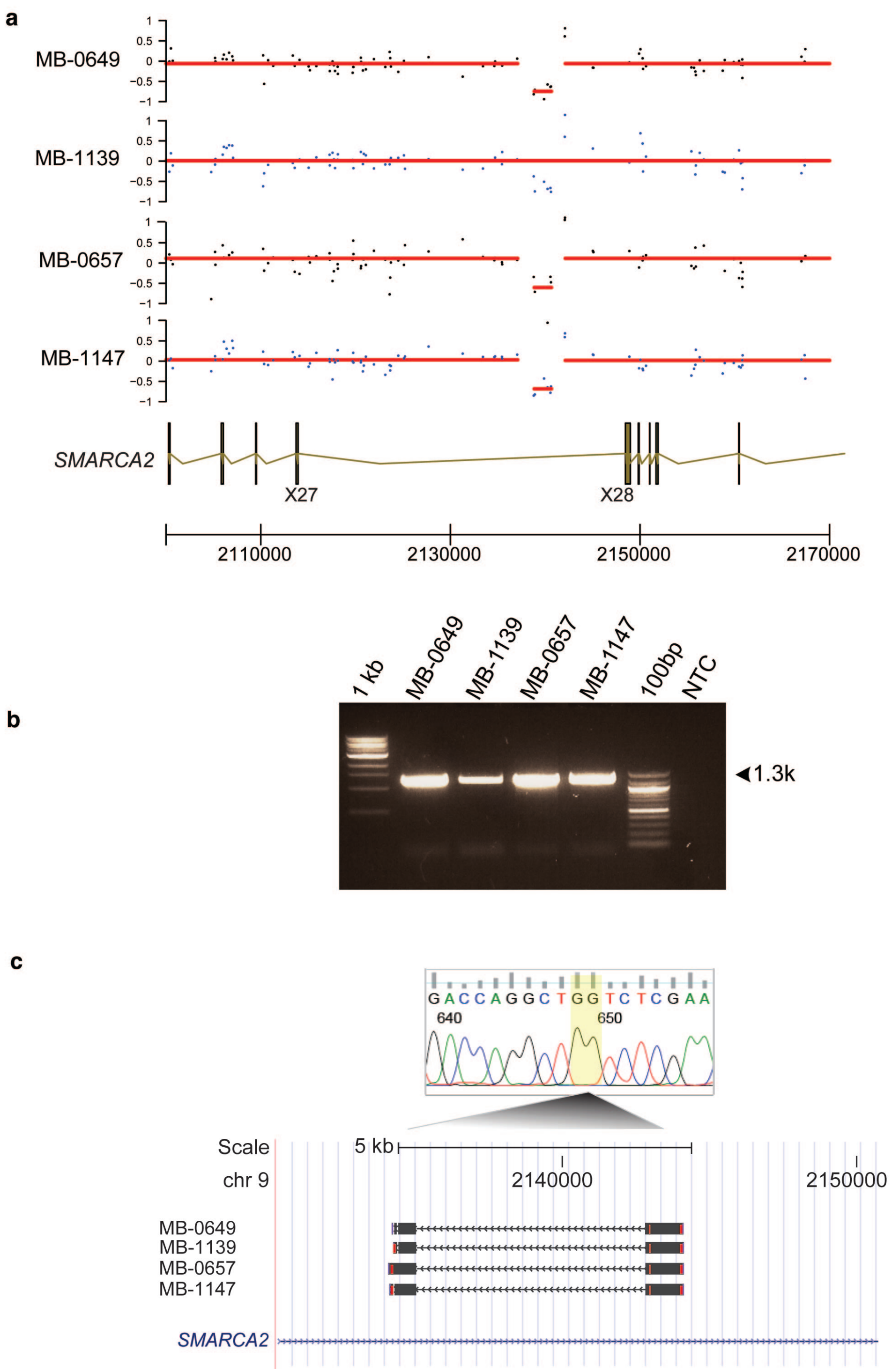


Figure S2: Assignment of cases to the PAM50 intrinsic subtypes. **a.** Heatmap illustration of gene expression values for the published PAM50 centroids. **b.** Unsupervised clustering of samples according to the PAM50 genes, where the dendrogram reflects the PAM50 assignments, grade, and ER status. **c.** Comparison of various assignments for key clinical markers based on immunohistochemistry (IHC), gene expression, PAM50, and copy number calls (SNP6) for ER status (upper panel), HER2 status (middle panel), and PR status (lower panel).


a
MB-0649
MB-1139
MB-0657
MB-1147
1
0.5
0
−0.5
−1
SMARCA2
X27
X28
2110000
2130000
2150000
2170000
b
1 kb
MB-0649
MB-1139
MB-0657
MB-1147
100bp
NTC
◄1.3k
c
G A C C A G G C T G G T C T C G A A
640
650
Scale
5 kb
chr 9
2140000
2150000
MB-0649
MB-1139
MB-0657
MB-1147
SMARCA2

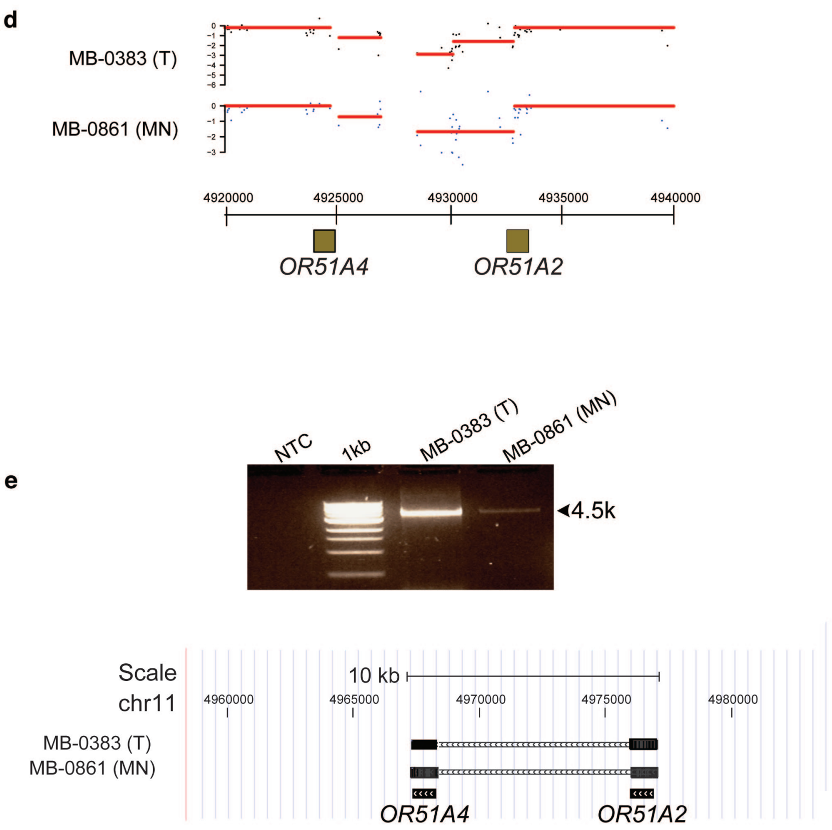
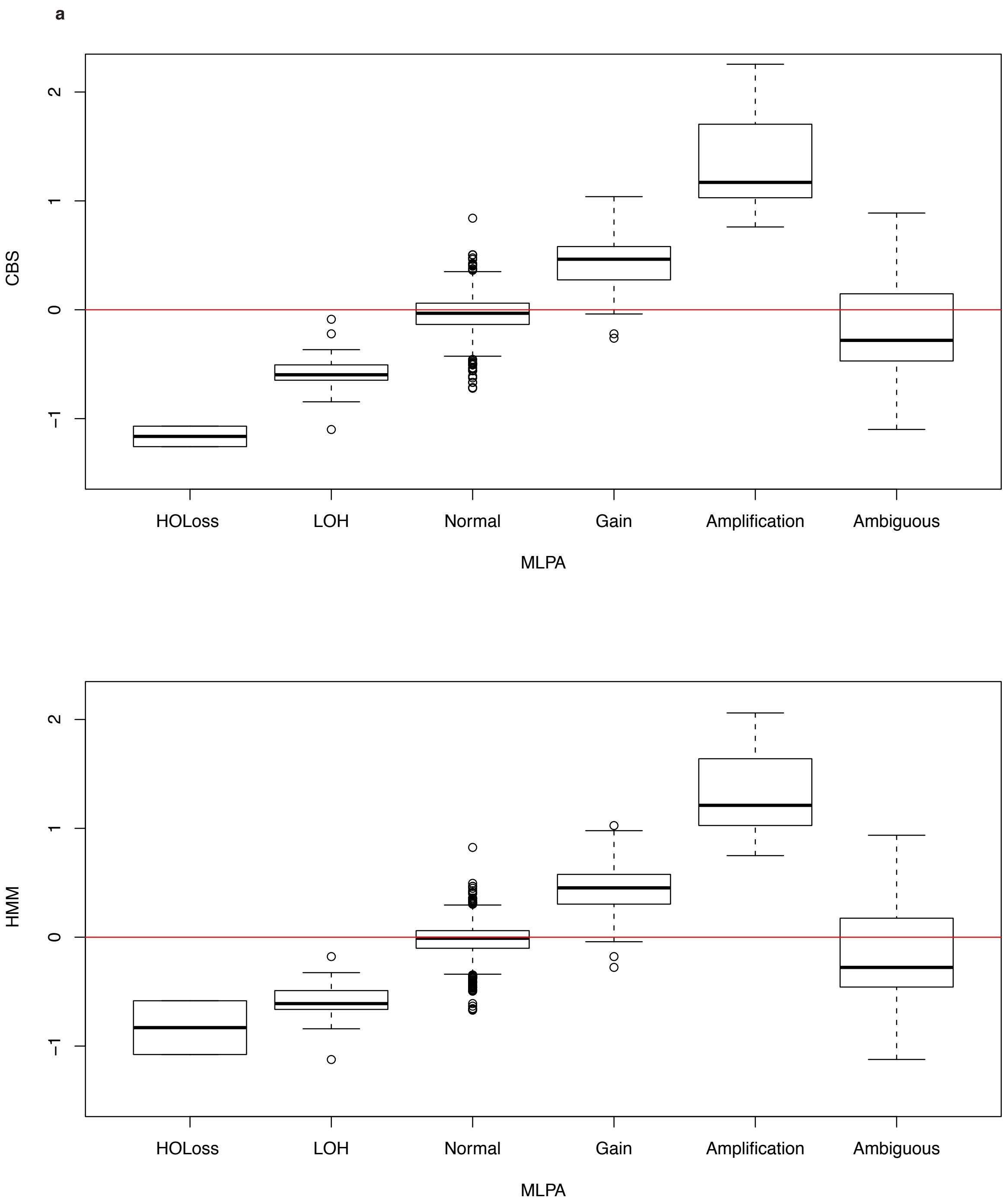


Figure S3: Verification of germline CNVs identified from the analysis of Affymetrix SNP 6.0 copy number data. **a.** Probe-level copy number plots (derived from CBS) illustrating several samples with *SMARCA2* intronic deletion. **b.** Long-range PCR confirmed a 3,914 bp deletion. **c.** The deletion was mapped to nucleotide resolution by sequencing. **d.** Probe-level copy number plots (derived from CBS) illustrating a deletion between two olfactory receptor genes, *OR51A4* and *OR51A2* in a tumour and its matched normal. **e.** The deletion was captured using long-range PCR, and sequencing revealed a 8,585 bp deletion resulting in fusion of both olfactory receptor genes.


a
CBS
2
1
0
−1
HOLoss
LOH
Normal
Gain
Amplification
Ambiguous
MLPA
HMM
2
1
0
−1
HOLoss
LOH
Normal
Gain
Amplification
Ambiguous
MLPA

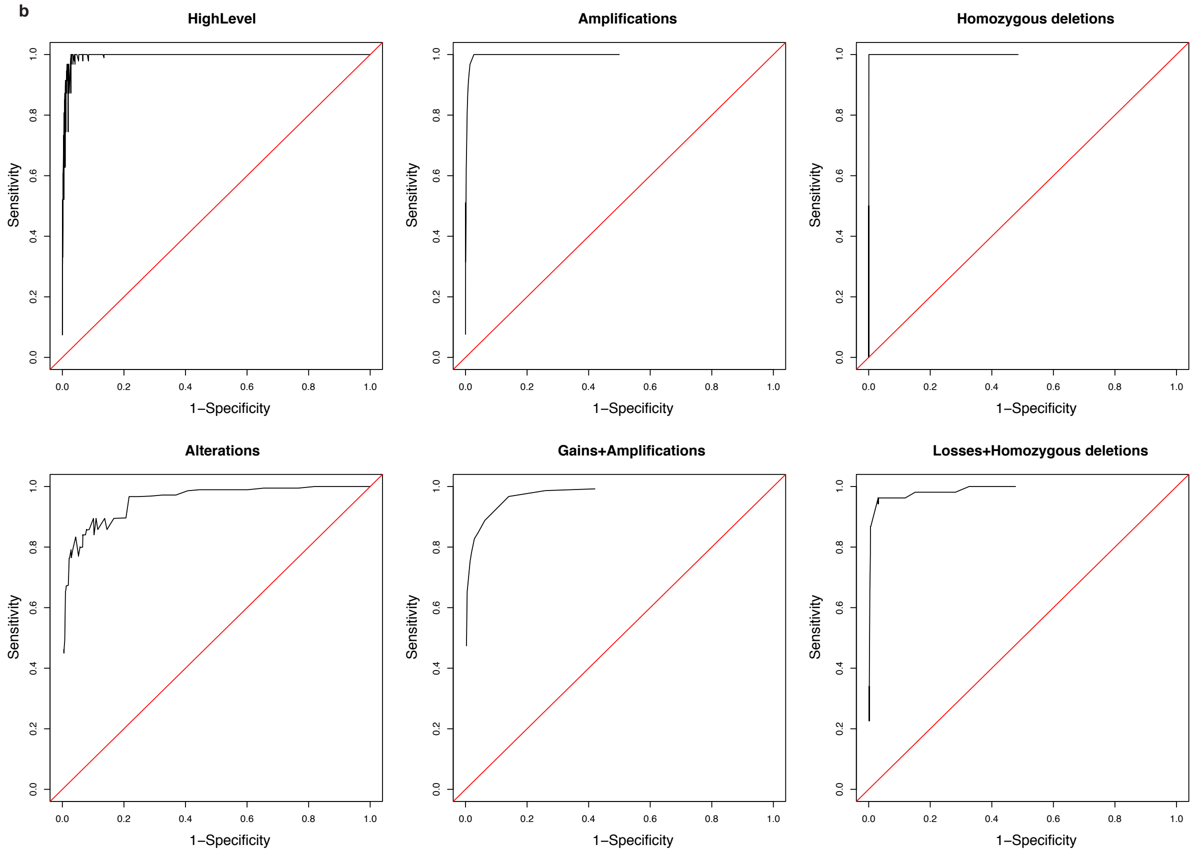


Figure S4: Assessment of somatic copy number segmentation for CBS and HMM. **a.** Comparison of CNA calling accuracy for CBS and HMM based on MLPA data. **b.** ROC curves for calling alterations based on CBS segmentation.

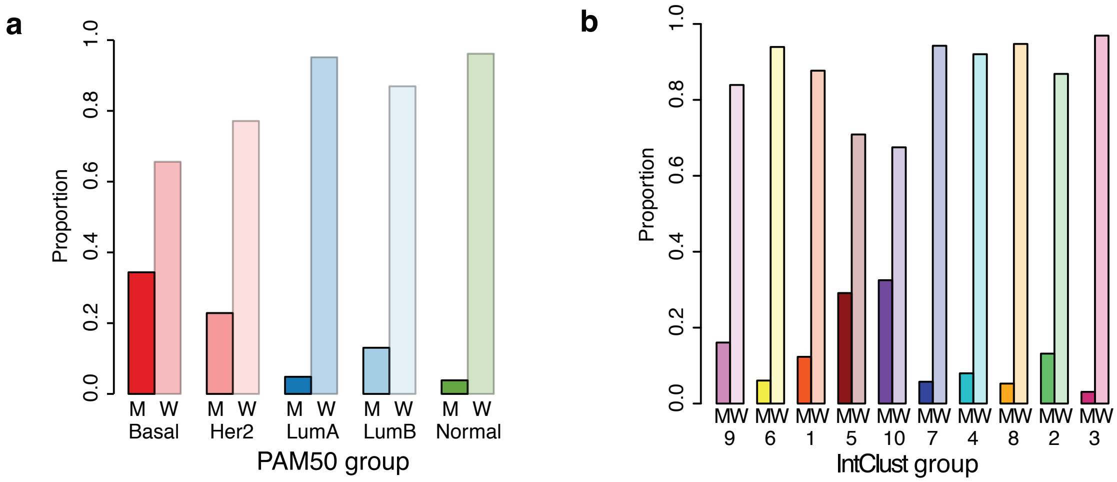
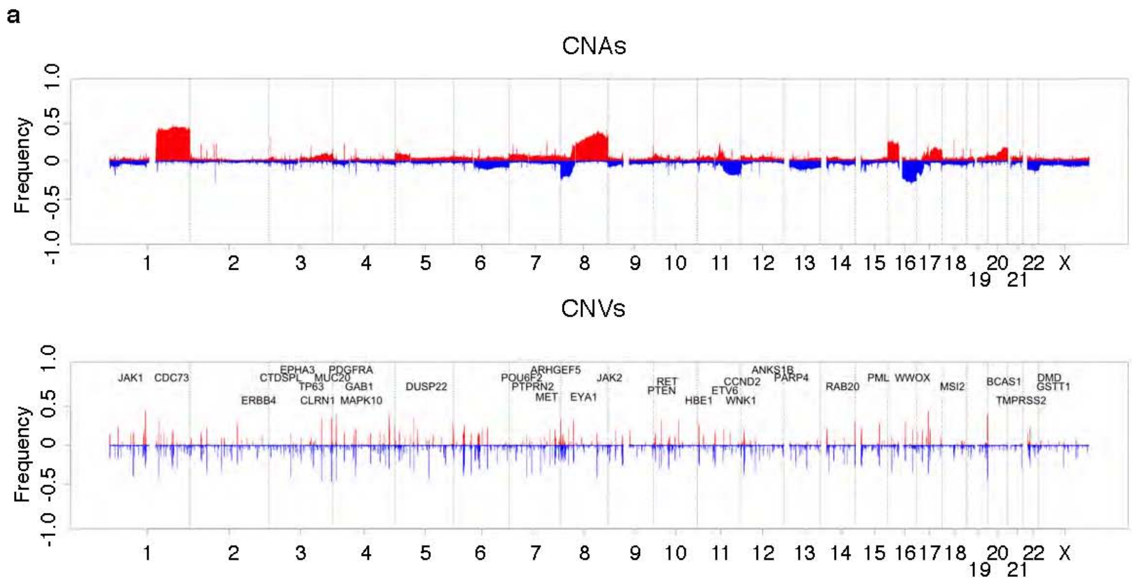


Figure S5: The proportion of *TP53* mutated versus wildtype cases. **a.** Mutational frequencies across the PAM50 intrinsic subtypes. **b.** As in (a), but for the integrative cluster subgroups.



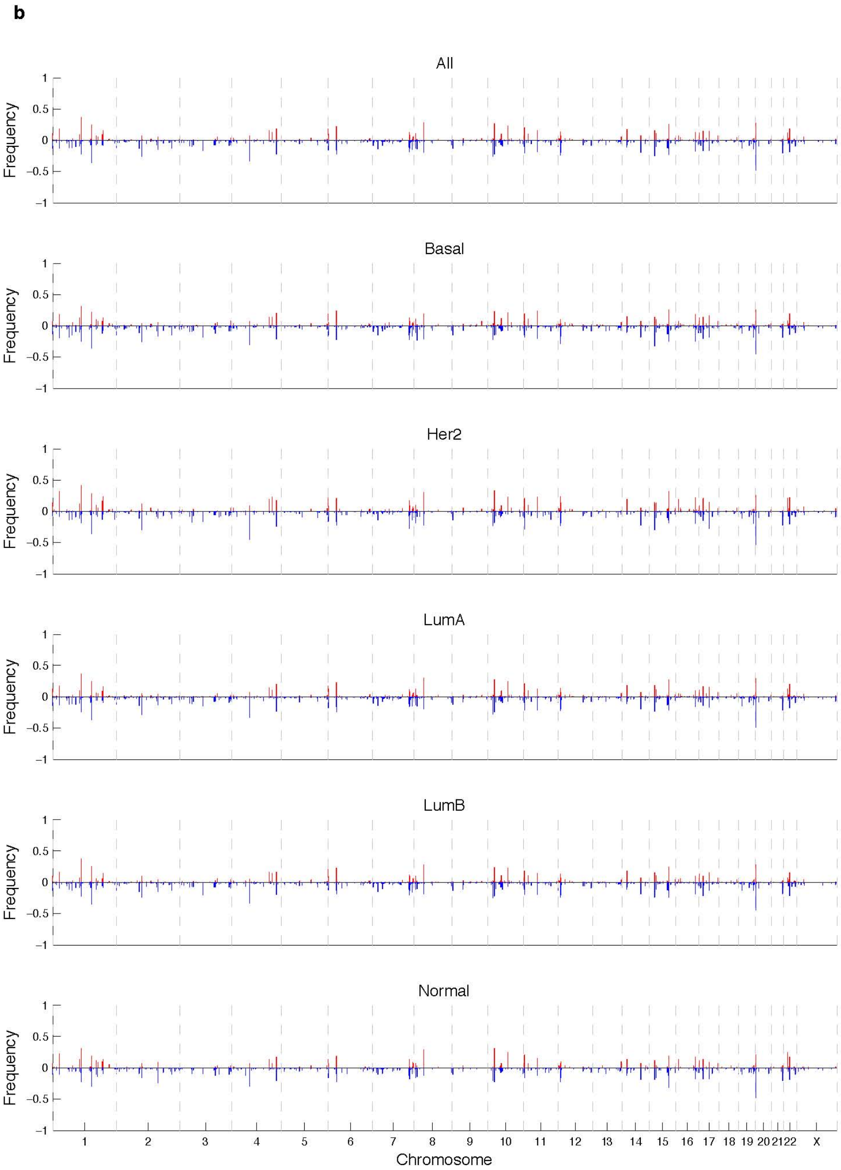


Figure S6: The copy number landscape of breast cancer. **a.** Global map of copy number variation in breast cancer. Genome-wide probe-level frequencies of somatically acquired CNAs (upper panel) and inherited germline CNVs (lower panel) are indicated for gains (red) and losses (blue) (derived from CBS), where known cancer genes targeted by CNVs in > 5% of cases are labelled. **b.** Germline CNV landscape across the PAM50 intrinsic subtypes. Genome-wide probe-level frequencies of inherited germline CNVs, where gains are indicated in red and losses in blue.

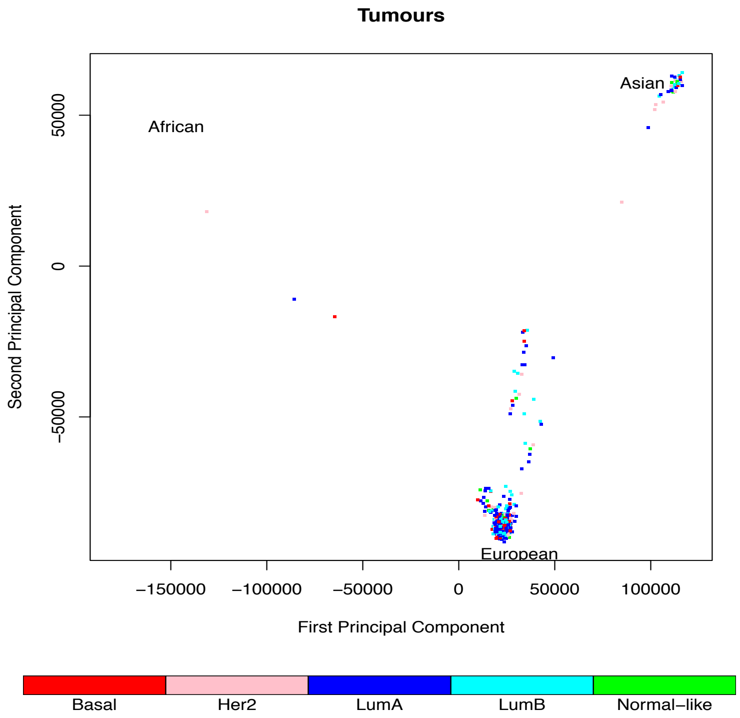


Figure S7: Primary HapMap population clusters for the METABRIC cohort.

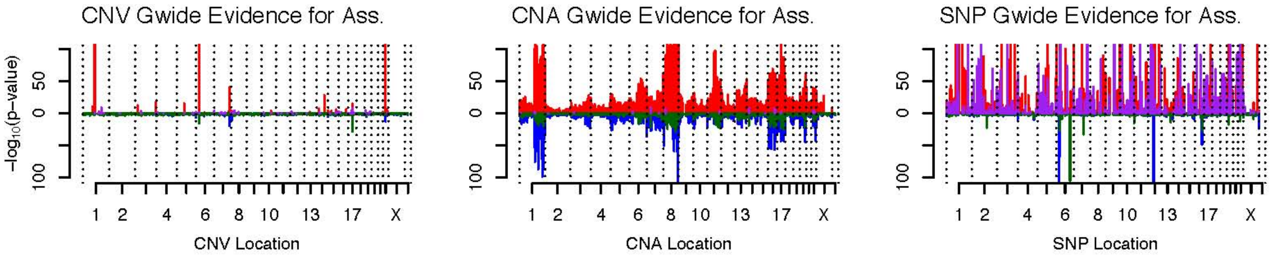
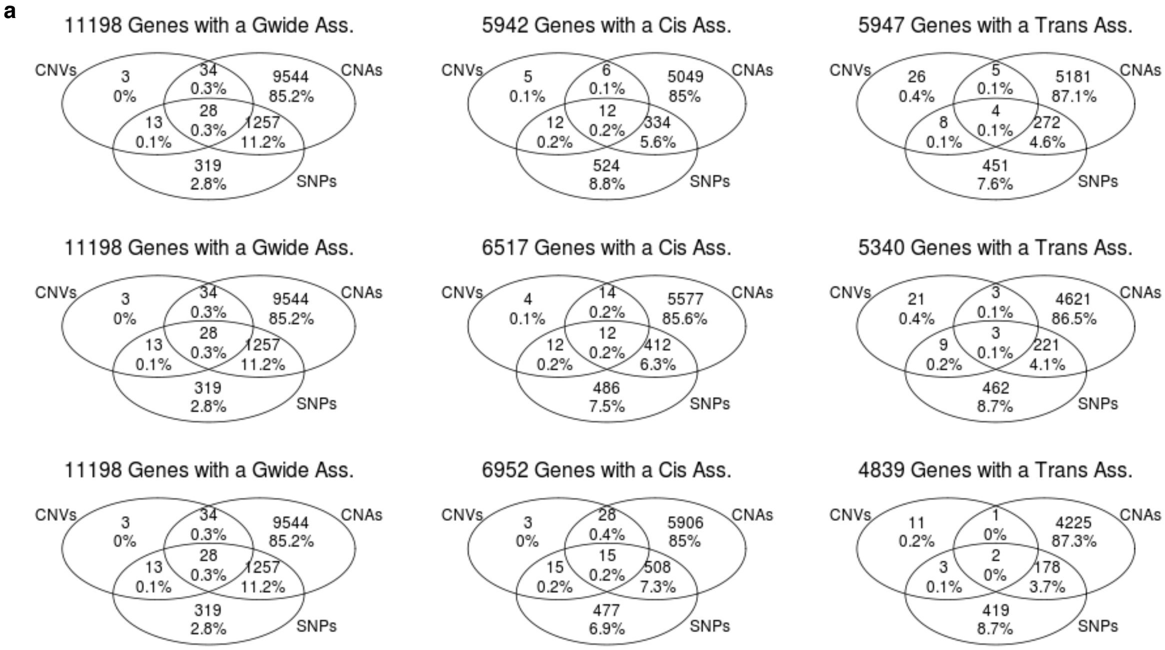


Figure S8: Germline and somatic variants influence tumour expression architecture. Manhattan plots indicate the significance of genome-wide associations for each predictor type. The directionality of the associations is indicated as follows: *cis* - positive (red), negative (pink); *trans*- positive (blue), negative (green).


a
11198 Genes with a Gwide Ass.
CNVs
CNAs
SNPs
3 0%
34 0.3%
9544 85.2%
28 0.3%
13 0.1%
1257 11.2%
319 2.8%
5942 Genes with a Cis Ass.
CNVs
CNAs
SNPs
5 0.1%
6 0.1%
5049 85%
12 0.2%
12 0.2%
334 5.6%
524 8.8%
5947 Genes with a Trans Ass.
CNVs
CNAs
SNPs
26 0.4%
5 0.1%
5181 87.1%
4 0.1%
8 0.1%
272 4.6%
451 7.6%
11198 Genes with a Gwide Ass.
CNVs
CNAs
SNPs
3 0%
34 0.3%
9544 85.2%
28 0.3%
13 0.1%
1257 11.2%
319 2.8%
6517 Genes with a Cis Ass.
CNVs
CNAs
SNPs
4 0.1%
14 0.2%
5577 85.6%
12 0.2%
12 0.2%
412 6.3%
486 7.5%
5340 Genes with a Trans Ass.
CNVs
CNAs
SNPs
21 0.4%
3 0.1%
4621 86.5%
3 0.1%
9 0.2%
221 4.1%
462 8.7%
11198 Genes with a Gwide Ass.
CNVs
CNAs
SNPs
3 0%
34 0.3%
9544 85.2%
28 0.3%
13 0.1%
1257 11.2%
319 2.8%
6952 Genes with a Cis Ass.
CNVs
CNAs
SNPs
3 0%
28 0.4%
5906 85%
15 0.2%
15 0.2%
508 7.3%
477 6.9%
4839 Genes with a Trans Ass.
CNVs
CNAs
SNPs
11 0.2%
1 0%
4225 87.3%
2 0%
3 0.1%
178 3.7%
419 8.7%

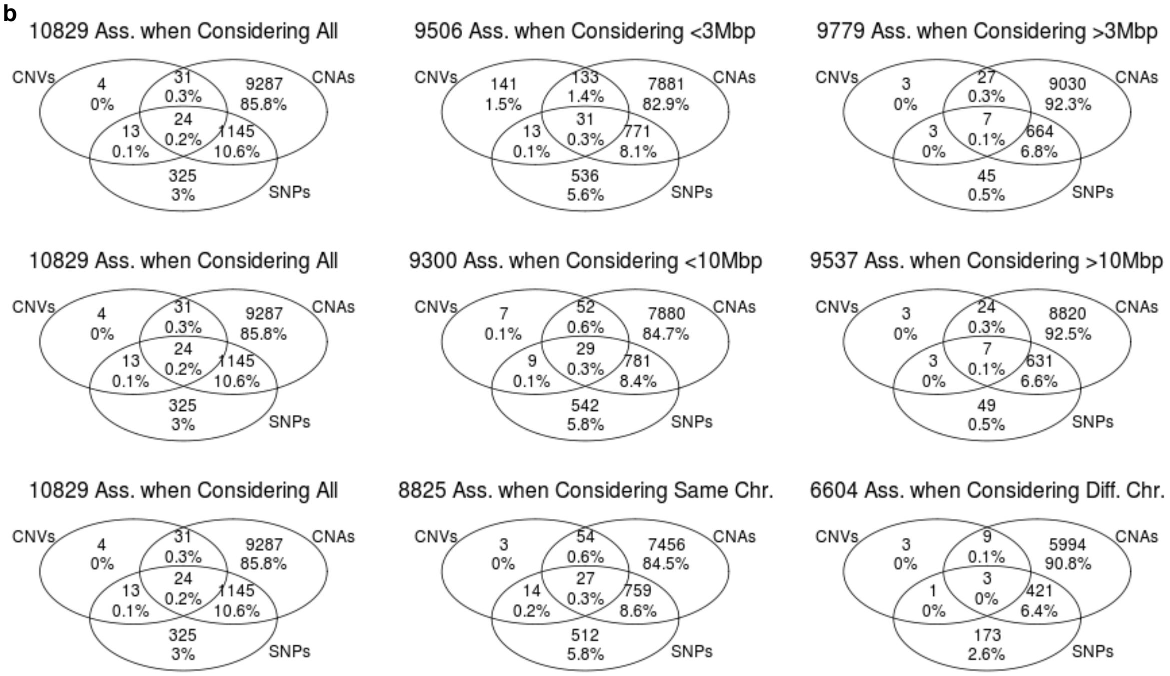


Figure S9: Venn diagrams based on alternative definitions of the proximal *cis*-window. **a.** Venn diagrams illustrate the relative contribution of germline variants (SNPs, CNVs) and somatic aberrations (CNAs) on genome-wide, *cis*, and *trans* tumour expression variation for three different proximal window sizes (3 Mb, 10 Mb, and same chromosome, rows 1-3, respectively) at a 0.0001 significance threshold after Šidák adjustment. This figure illustrates how the number of proximal associations increases, while the number of distal associations decreases, as the definition of proximal is enlarged. Note that for all three thresholds, the number of genome-wide association remains the same (column 1). An alternative approach would have been to perform the analysis twice: first considering only proximal predictors and second considering only distal predictors. Note that for this analysis the significance threshold was halved (reduced to 0.00005) to account for two separate tests being performed. **b.** The results of such an approach are illustrated here. Here, the three rows consider different definitions of proximal (3Mb, 10Mb and same chromosome). The first column reports the types of associations found when examining all predictors, the second column reports the types of associations found when examining only proximal predictors, and the third column reports the types of associations found when examining only distal predictors.

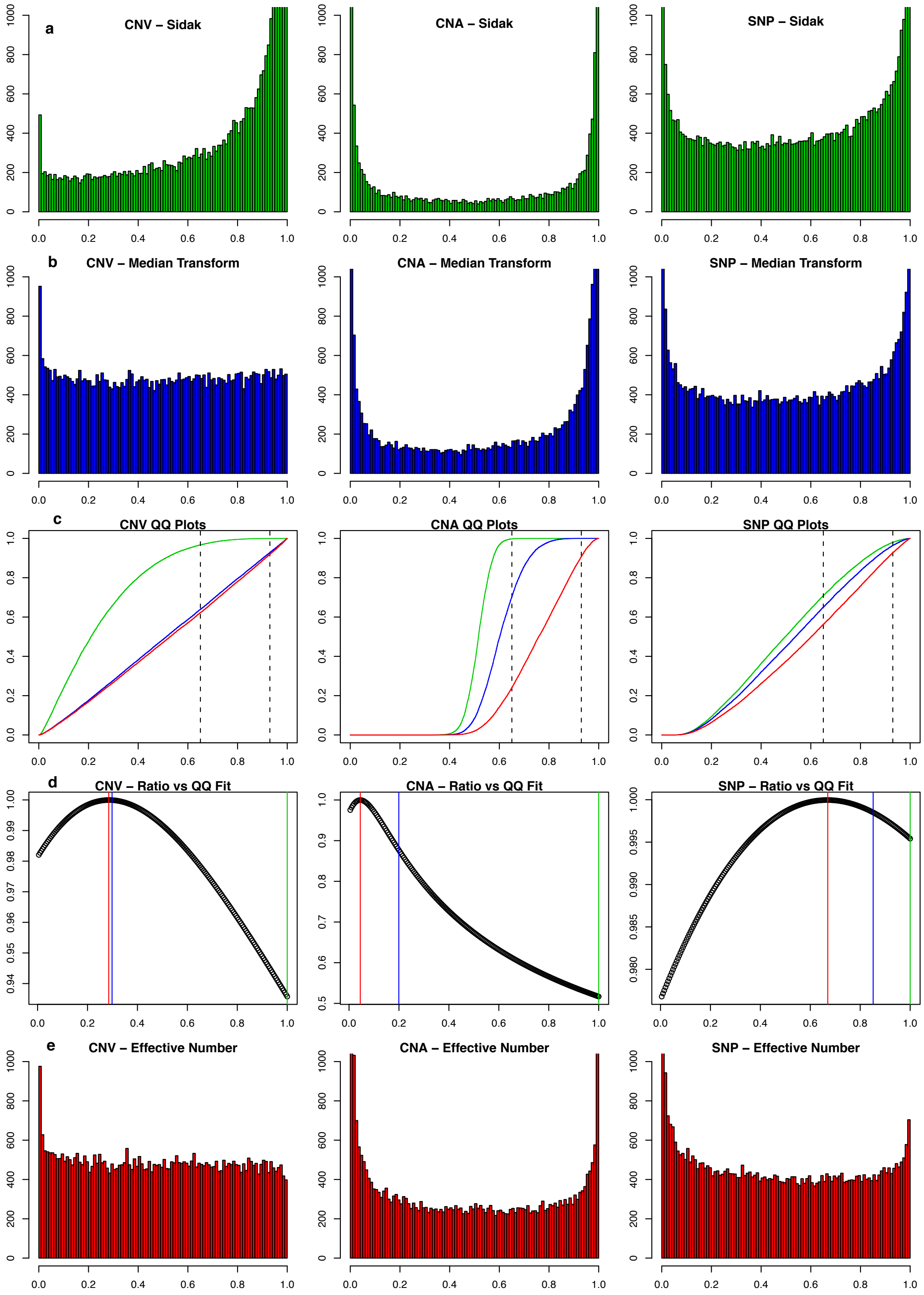
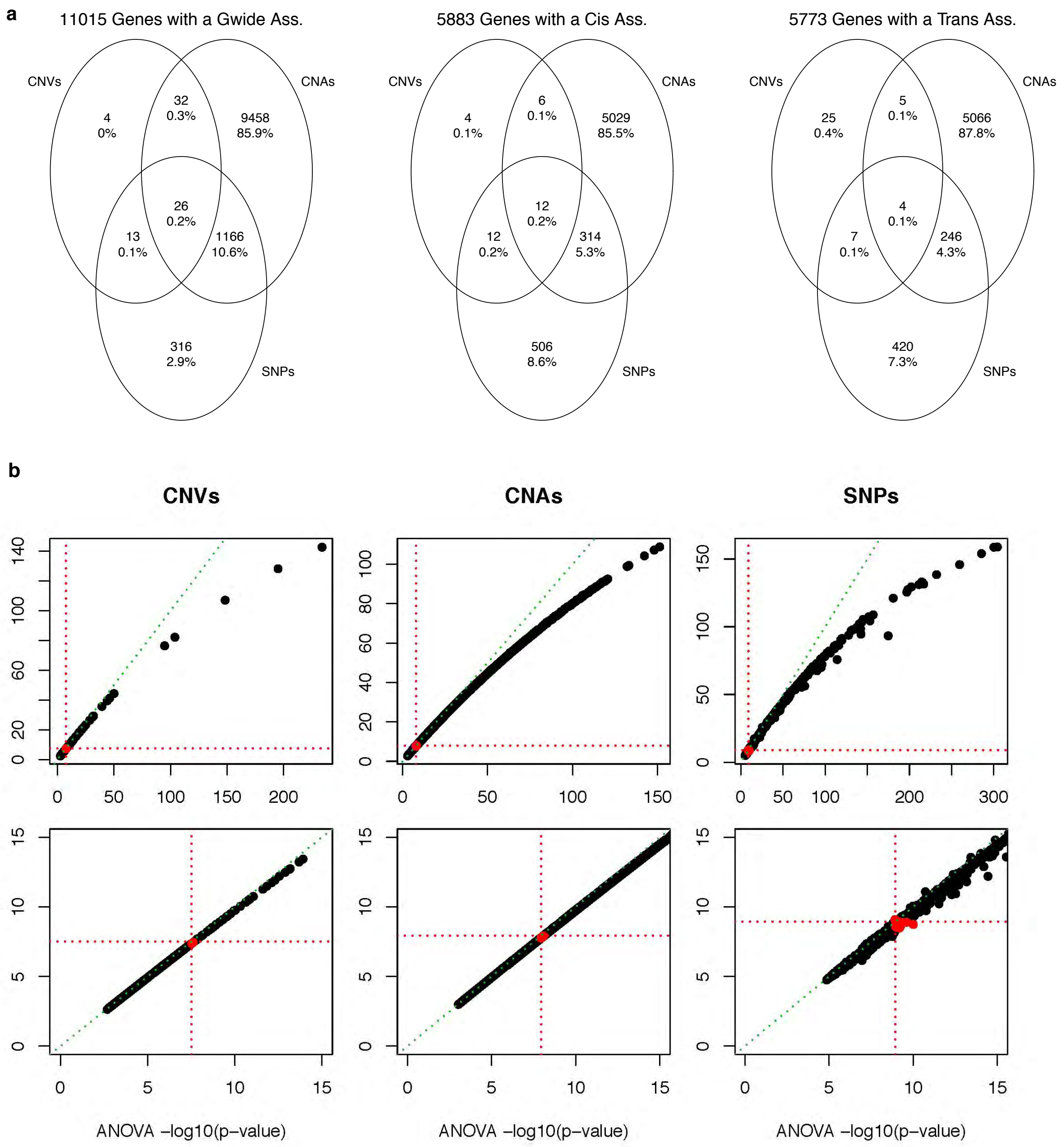

a
CNV – Sidak
CNA – Sidak
SNP – Sidak
b
CNV – Median Transform
CNA – Median Transform
SNP – Median Transform
c
CNV QQ Plots
CNA QQ Plots
SNP QQ Plots
d
CNV – Ratio vs QQ Fit
CNA – Ratio vs QQ Fit
SNP – Ratio vs QQ Fit
e
CNV – Effective Number
CNA – Effective Number
SNP – Effective Number
0 200 400 600 800 1000
0.0 0.2 0.4 0.6 0.8 1.0
0.94 0.95 0.96 0.97 0.98 0.99 1.00
0.5 0.6 0.7 0.8 0.9 1.0
0.980 0.985 0.990 0.995 1.000

Figure S10: Assessment of various multiple testing adjustments for the eQTL analysis. **a.** Histograms illustrating the effect of the Šidák correction, which adjusts each *P*-value according to a transformation of the form: $p \rightarrow 1\text{-}(1\text{-}p)^{Nr}$, where Nr represents the effective number of tests, and r is a value between 0 and 1, shown here (r = 1). This appears to bias the adjusted *P*-values towards 1, and hence is overly cautious. **b.** Histograms illustrating attempts to estimate a suitable value of r via median mapping such that the median *P*-value is transformed to 0.5. While this approach seems suitable for the CNVs, where for many of the responses the null hypothesis is true, it is overly conservative for CNAs and SNPs. **c.** Q-Q plots of adjusted *P*-values, where the overly-cautious nature of setting r = 1 is indicated by the green line, whereas the blue line represents median mapping, which forces the Q-Q line to pass through the point (0.5,0.5). The adjusted *P*-values falling between the two dashed lines correspond to those ranked 18,610 to 26,609 (between quantiles 0.65 to 0.93). The assumption is that these 8,000 *P*-values likely correspond to true null hypotheses, so the line in this region should be straight. **d.** Q-Q plots illustrating a measure of 'straightness' for this region based on the R^2 for different values of r. The value which maximised this correlation and so resulted in the straightest portion of line was selected. Notably, the values of r for CNVs, CNAs and SNPs were 0.285, 0.045 and 0.67, respectively and corresponds roughly to the effective number of independent tests (3,289, 8,724 and 586,014, respectively). **e.** This panel illustrates that the approach based on the effective number of tests resulted in histograms with more desirable properties for all three predictor types.


a
11015 Genes with a Gwide Ass.
CNVs
CNAs
SNPs
4
0%
32
0.3%
9458
85.9%
26
0.2%
13
0.1%
1166
10.6%
316
2.9%
5883 Genes with a Cis Ass.
CNVs
CNAs
SNPs
4
0.1%
6
0.1%
5029
85.5%
12
0.2%
12
0.2%
314
5.3%
506
8.6%
5773 Genes with a Trans Ass.
CNVs
CNAs
SNPs
25
0.4%
5
0.1%
5066
87.8%
4
0.1%
7
0.1%
246
4.3%
420
7.3%
b
CNVs
CNAs
SNPs
KW −log10(p−value)
ANOVA −log10(p−value)
0
20
60
100
140
50
150
200
40
80
300
5
10
15

c

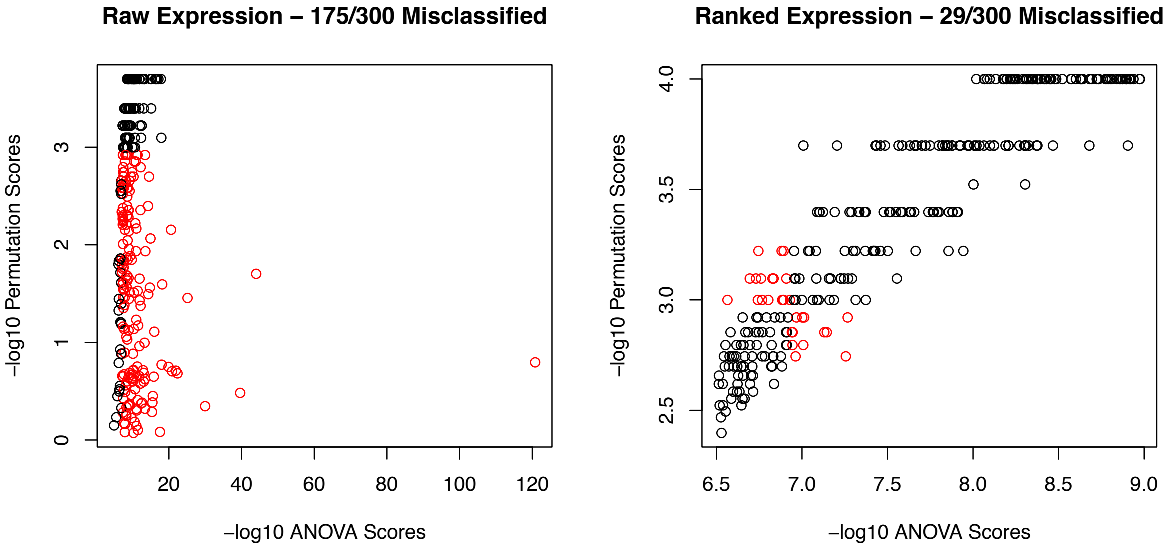


Figure S11: Comparison of the results of ANOVA on ranked values versus the Kruskal-Wallis test and assessment of the use of ranked versus raw expression values for handling outlier responses in the eQTL analysis. **a.** Venn diagrams showing the relative number of significant association for each predictor based on the KW test. Note that although slightly fewer genes exhibit significant eQTLs relative to the ANOVA on ranked expression results, the percentages in each cell remain relatively constant. **b.** Scatterplot of -$\log_{10}$ P-values for the Kruskal-Wallis test versus ANOVA on ranked expression values indicates a high degree of concordance, although there is a slight drop for extreme -$\log_{10}$ P-values (> 20), as expected with any asymptotic test. The red lines indicate the significance threshold (0.0001 after multiple correction). **c.** Plots illustrating the difference between the sets of P-values obtained when ranked versus raw expression values are employed for the ANOVA analyses based on the computation of empirical P-values via permutation testing for a subset of 300 expression probes (see Methods). The permuted P-value was considered the ground truth, from which the presence/absence of an association was determined according to whether the empirical P-value exceeded 0.001 or not. Red points indicate a misclassification.

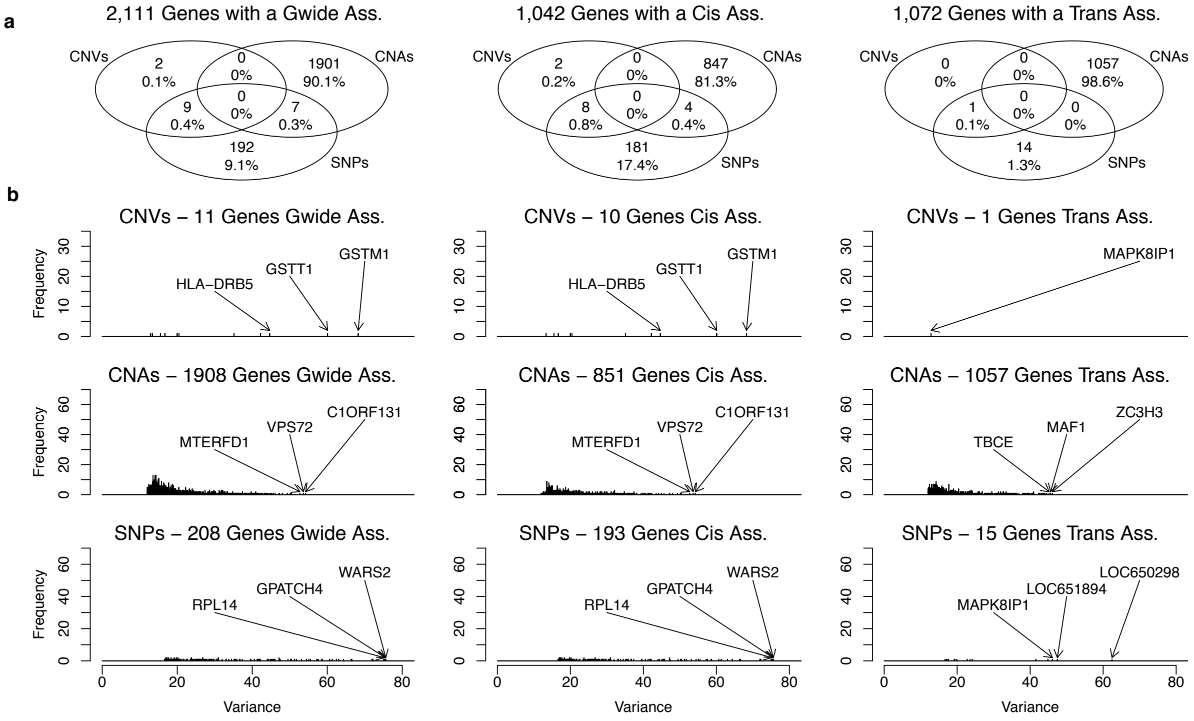


Figure S12: eQTL analysis in cases of European ancestry for which normal genotype calls were available (*n* = 244). **a.** Venn diagrams depict the relative contribution of germline variants (SNPs, CNVs) and somatic aberrations (CNAs) on genome-wide, *cis*, and *trans* tumour expression variation. Expression profiles that were significantly associated with either a CNV, CNA, or SNP at an adjusted *P*-value of 0.0001 are reported. **b.** Histograms illustrate the proportion of variance explained by the most significantly associated predictor for each predictor type, where several of the top associations are indicated. As above, the results are shown for the genome-wide analysis as well as for proximal (*cis*) and distal (*trans*) associations.

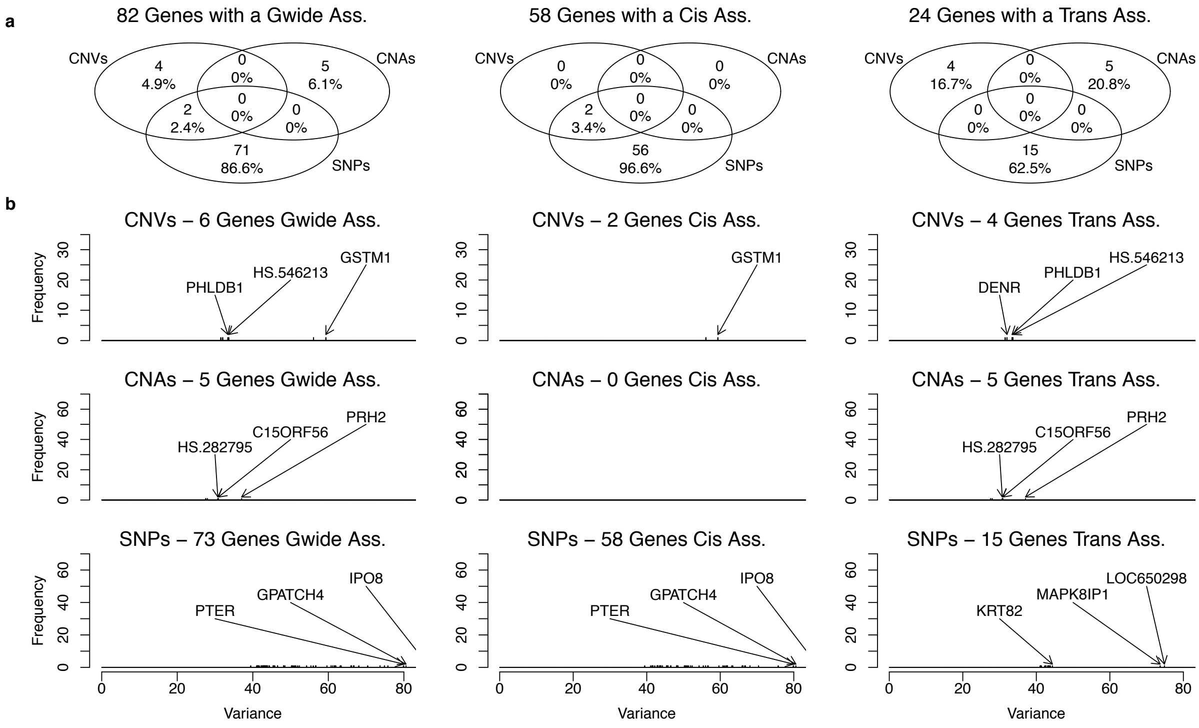


Figure S13: eQTL analysis of normal expression data. This analysis was performed in cases of European ancestry for which normal genotype calls and normal gene expression profiles were available (n = 85). **a.** Venn diagrams depict the relative contribution of germline variants (SNPs, CNVs) and somatic aberrations (CNAs) on genome-wide, *cis*, and *trans* tumour expression variation. Expression profiles that were significantly associated with either a CNV, CNA, or SNP at an adjusted P-value of 0.0001 are reported. **b.** Histograms illustrate the proportion of variance explained by the most significantly associated predictor for each predictor type, where several of the top associations are indicated. As above, the results are shown for the genome-wide analysis as well as for proximal (*cis*) and distal (*trans*) associations.

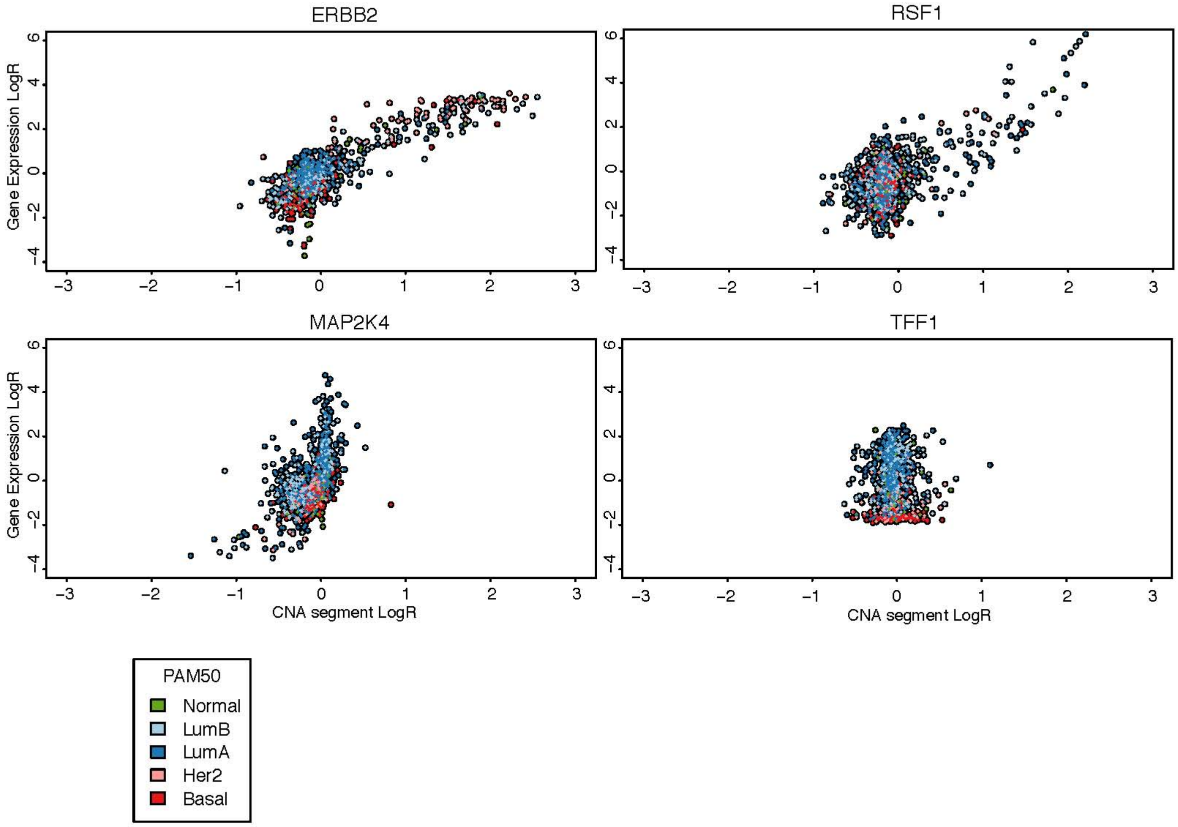
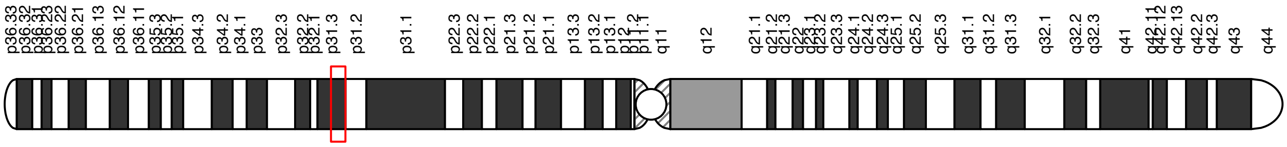
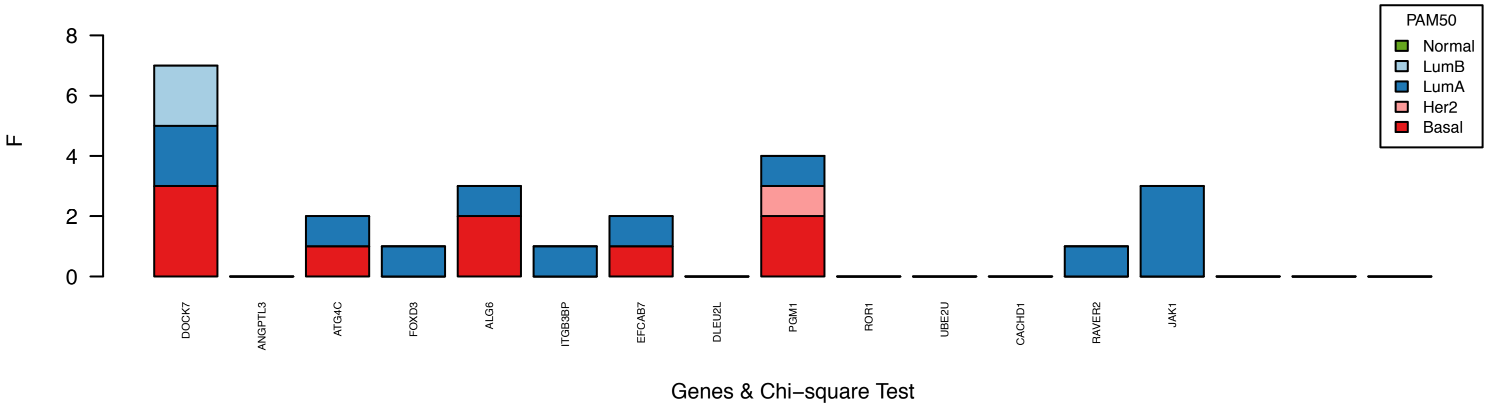


Figure S14: Illustration of *cis*-acting copy number expression associations. Scatterplots of scaled gene expression values versus segmented copy number values (log base 2, CBS-derived) for several exemplary genes, including *ERBB2* and *RSF1*, which are amplified, *MAP2K4*, which is homozygously deleted, and *TFF1*, which shows copy number independent expression changes. Data points are colour-coded according to the PAM50 subtypes.

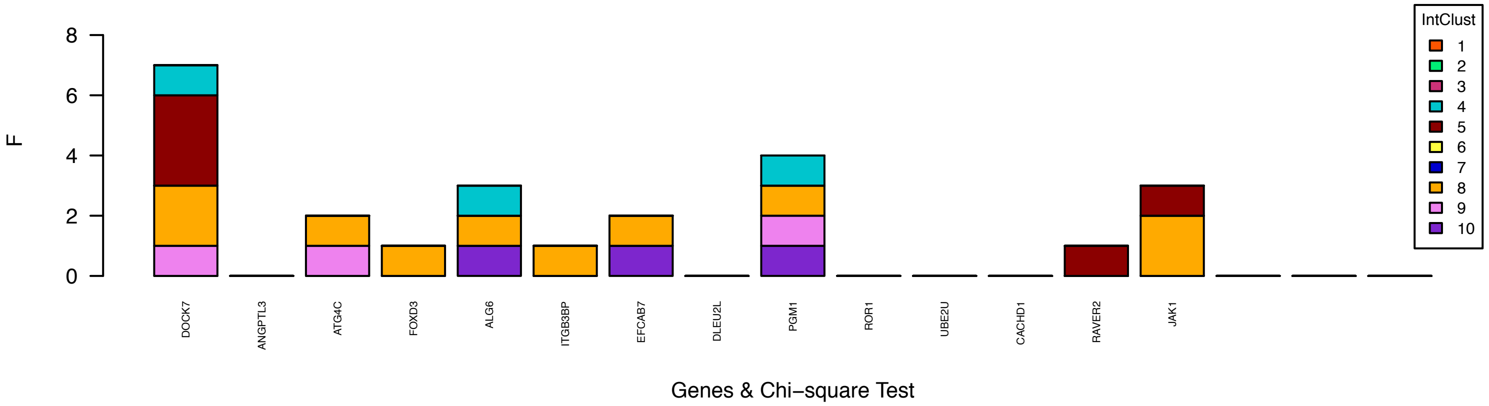
Chromosome 1 amplification outliers



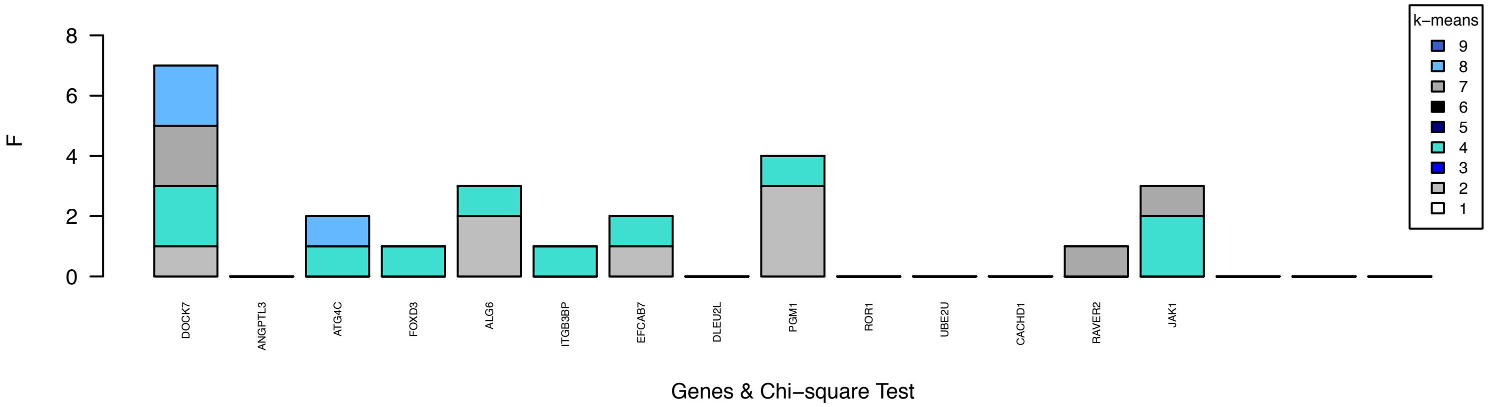
PAM50 region1 chr1:62692985–65465763 Subtypes Segregation – from DOCK7 to JAK1



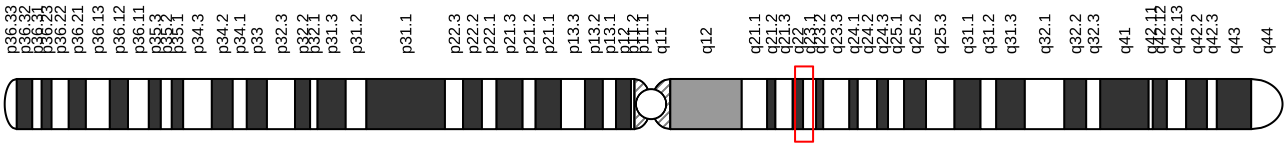
IntClust region1 chr1:62692985–65465763 Subtypes Segregation – from DOCK7 to JAK1



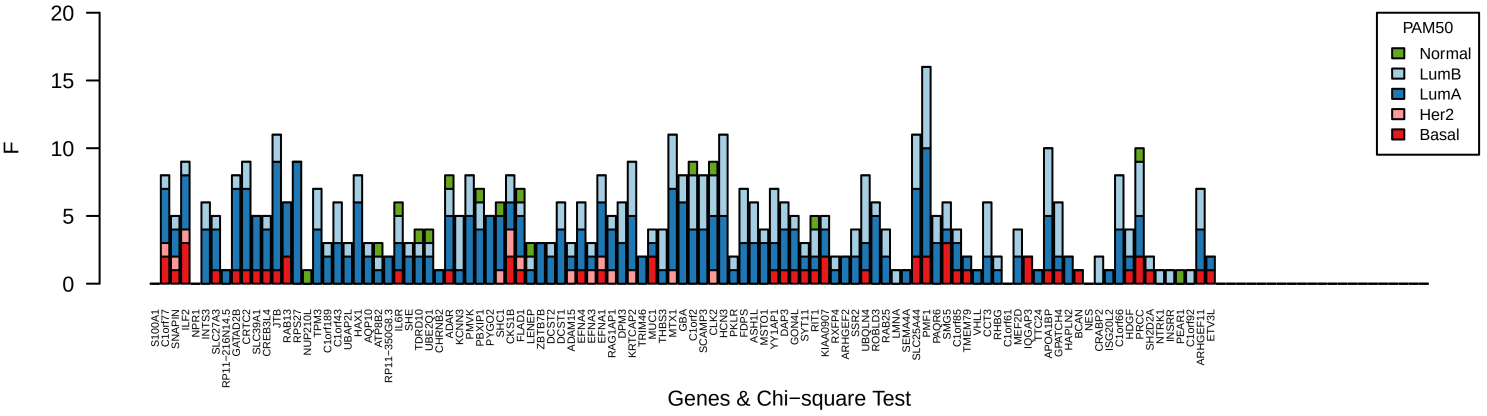
k–means region1 chr1:62692985–65465763 Subtypes Segregation – from DOCK7 to JAK1



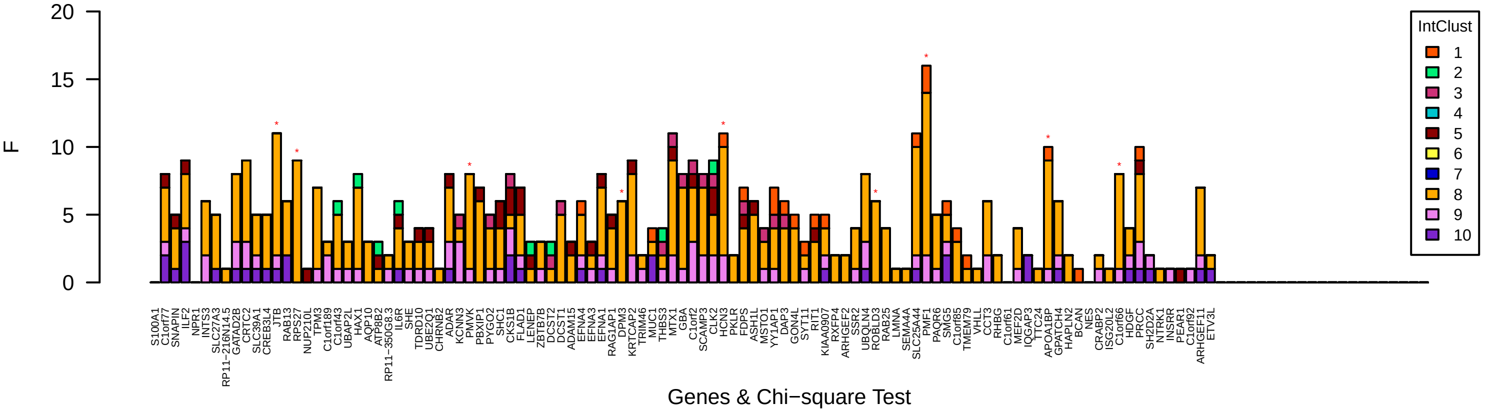
Chromosome 1 amplification outliers



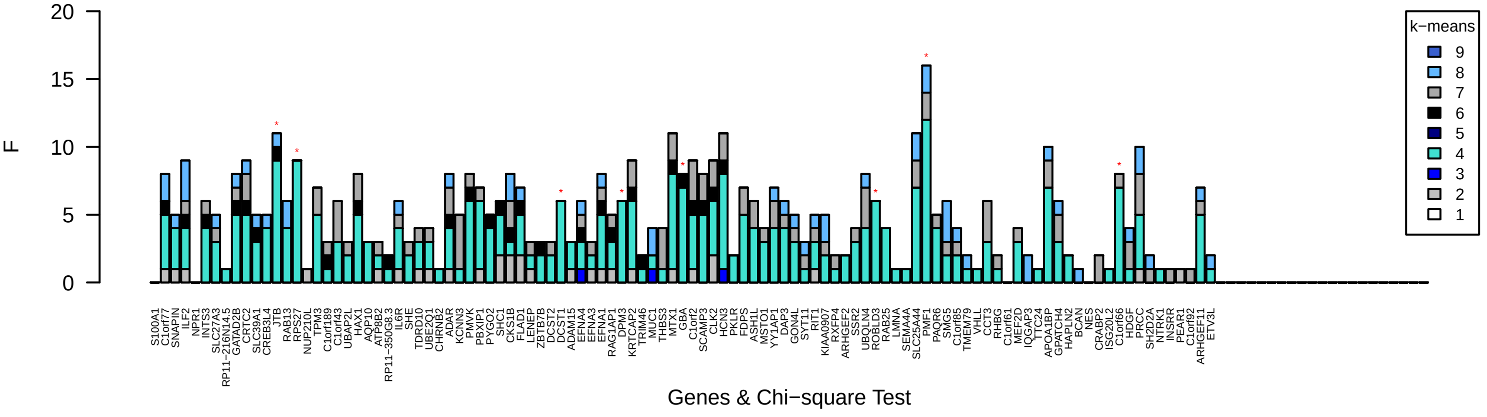
PAM50 region2 chr1:151857900−155336224 Subtypes Segregation − from S100A1 to ETV3L



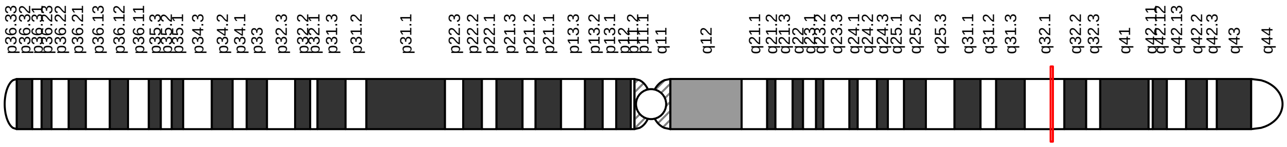
IntClust region2 chr1:151857900−155336224 Subtypes Segregation − from S100A1 to ETV3L



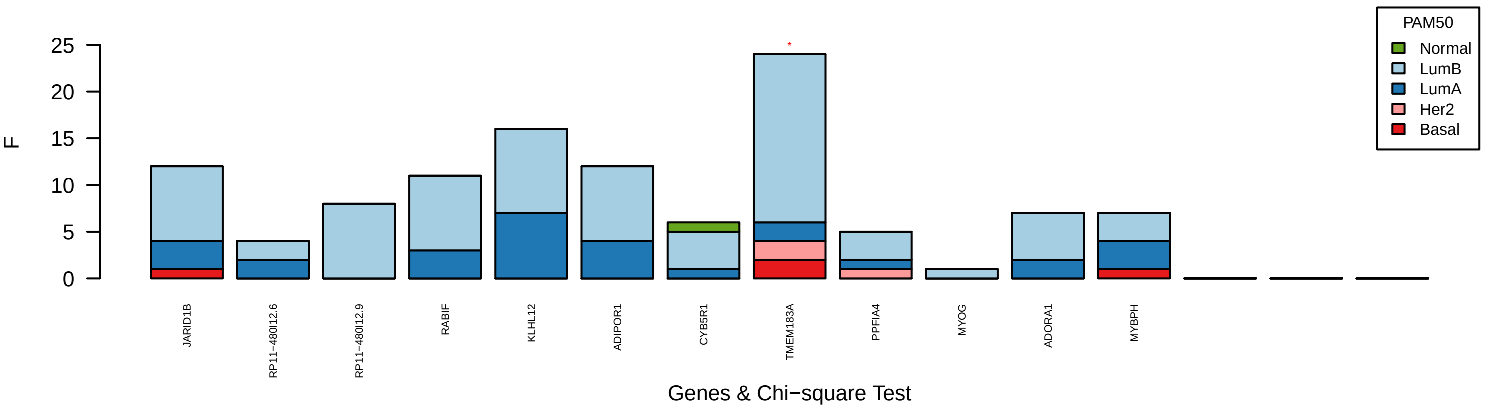
k−means region2 chr1:151857900−155336224 Subtypes Segregation − from S100A1 to ETV3L



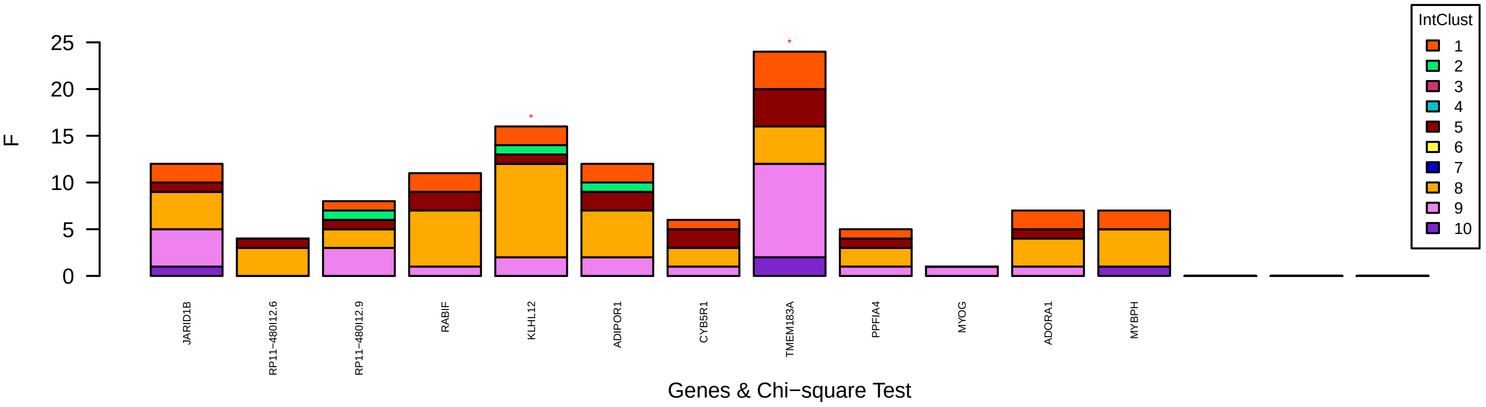
Chromosome 1 amplification outliers


p36.33 p36.32 p36.31 p36.23 p36.22 p36.21 p36.13 p36.12 p36.11 p35.3 p35.2 p35.1 p34.3 p34.2 p34.1 p33 p32.3 p32.2 p32.1 p31.3 p31.2 p31.1 p22.3 p22.2 p22.1 p21.3 p21.2 p21.1 p13.3 p13.2 p13.1 p12 p11.2 p11.1 q11 q12 q21.1 q21.2 q21.3 q22 q23.1 q23.2 q23.3 q24.1 q24.2 q24.3 q25.1 q25.2 q25.3 q31.1 q31.2 q31.3 q32.1 q32.2 q32.3 q41 q42.11 q42.12 q42.13 q42.2 q42.3 q43 q44

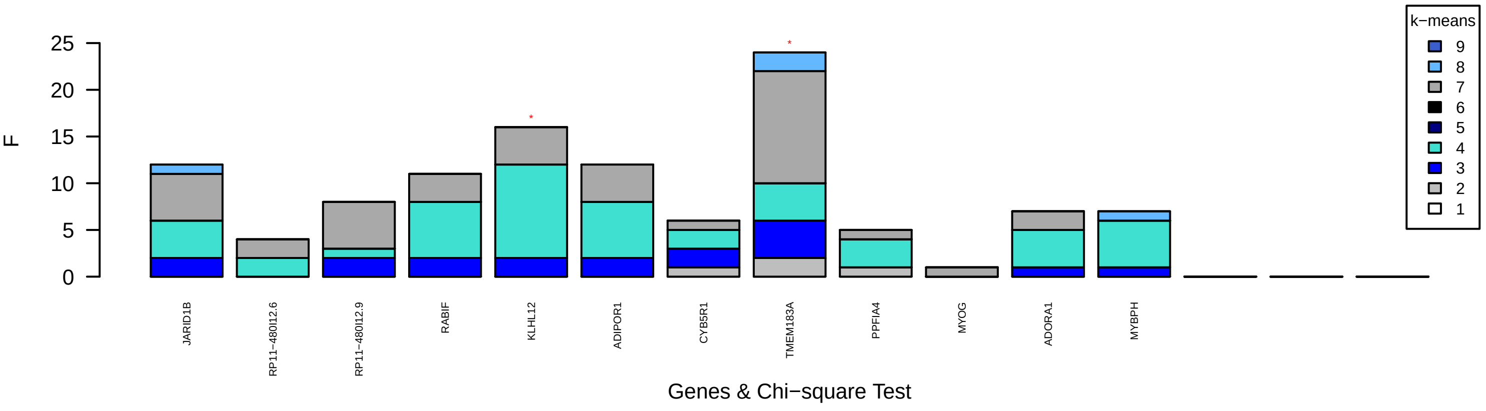
PAM50 region3 chr1:200963155−201411565 Subtypes Segregation − from JARID1B to MYBPH



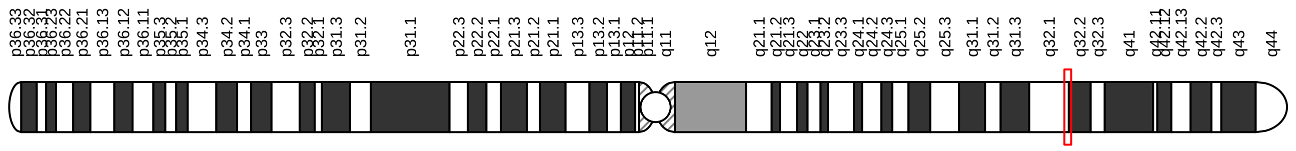
IntClust region3 chr1:200963155−201411565 Subtypes Segregation − from JARID1B to MYBPH



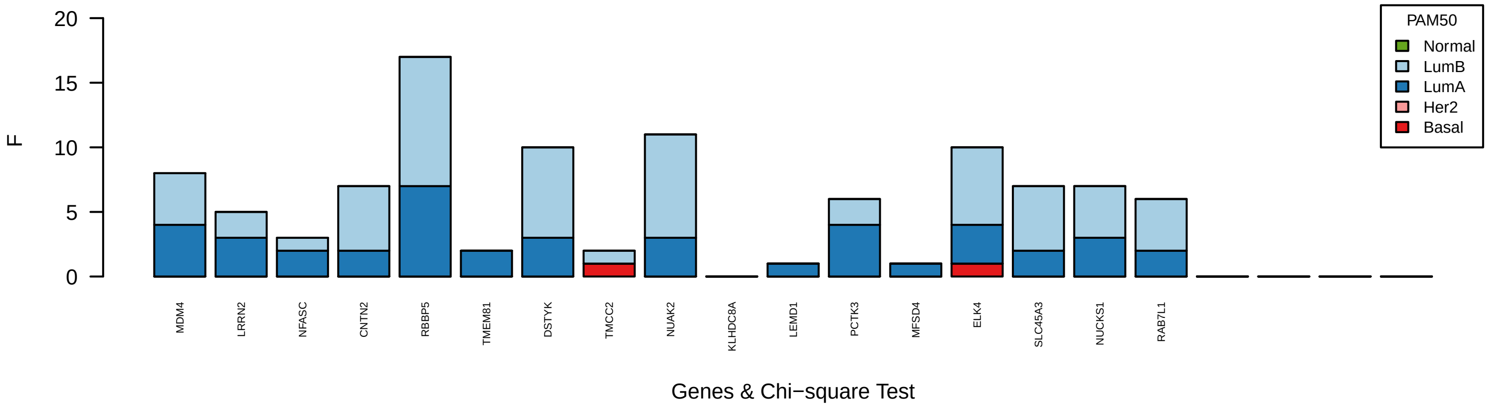
k−means region3 chr1:200963155−201411565 Subtypes Segregation − from JARID1B to MYBPH



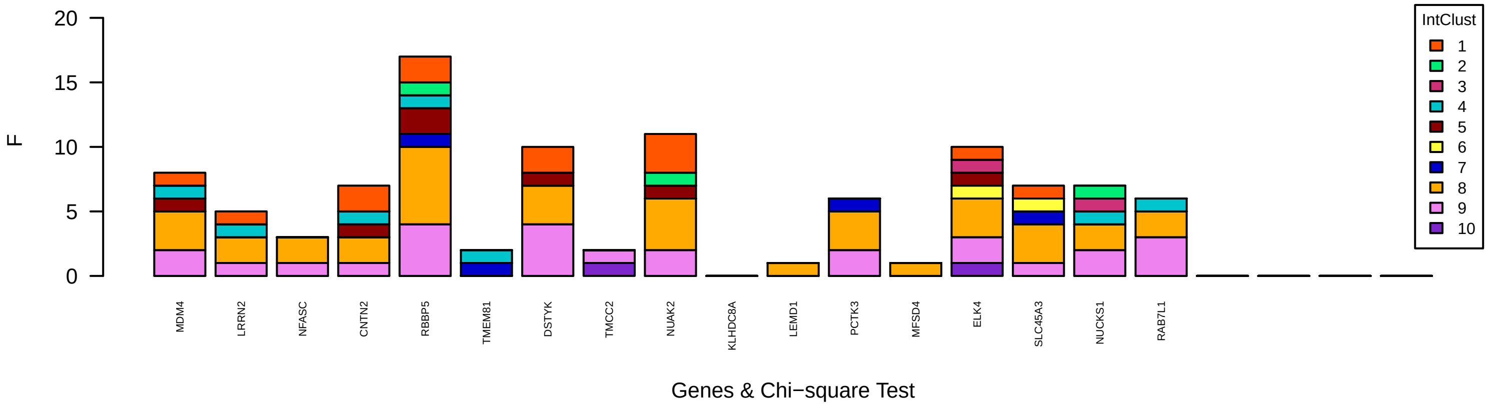
Chromosome 1 amplification outliers



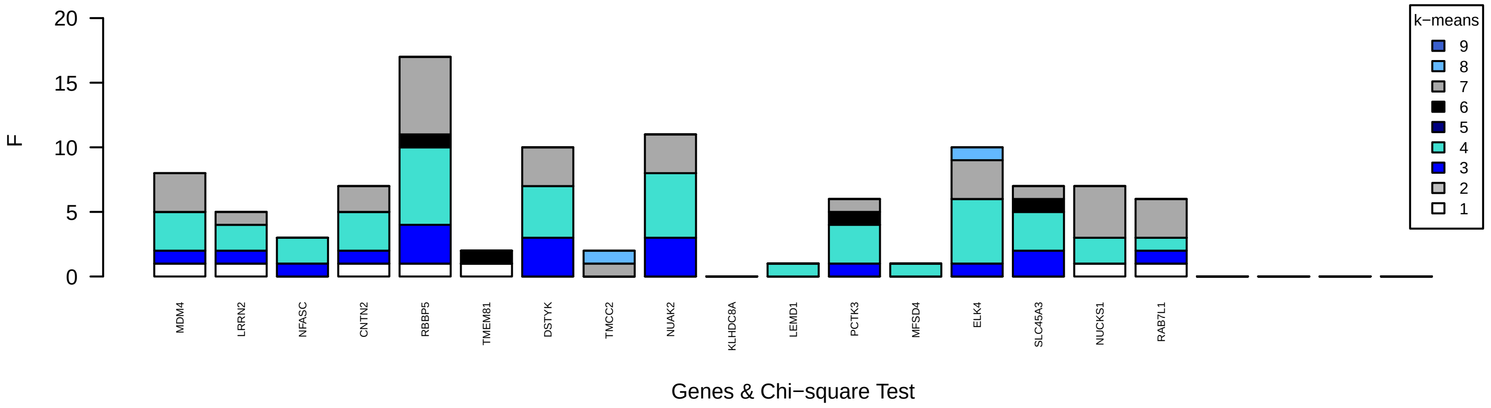
PAM50 region4 chr1:202752134−204048784 Subtypes Segregation − from MDM4 to RAB7L1



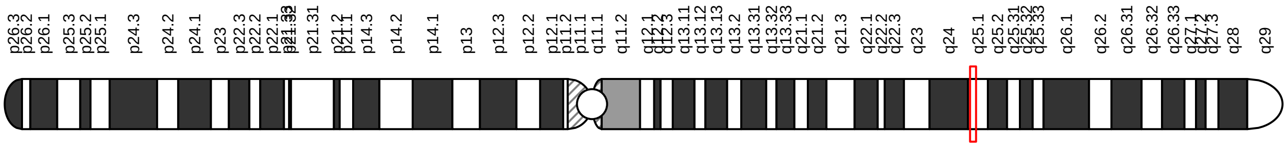
IntClust region4 chr1:202752134−204048784 Subtypes Segregation − from MDM4 to RAB7L1



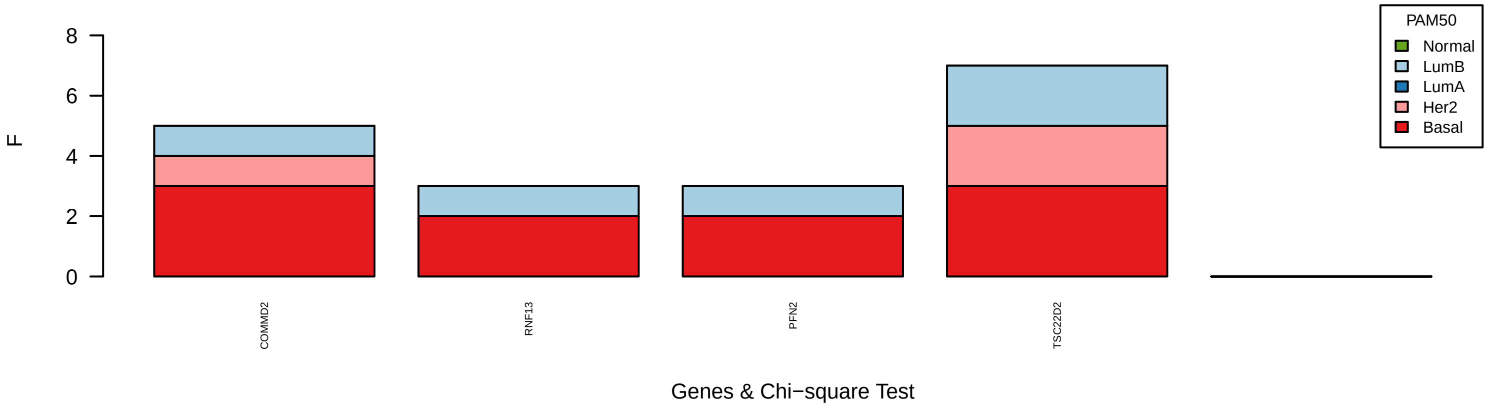
k−means region4 chr1:202752134−204048784 Subtypes Segregation − from MDM4 to RAB7L1



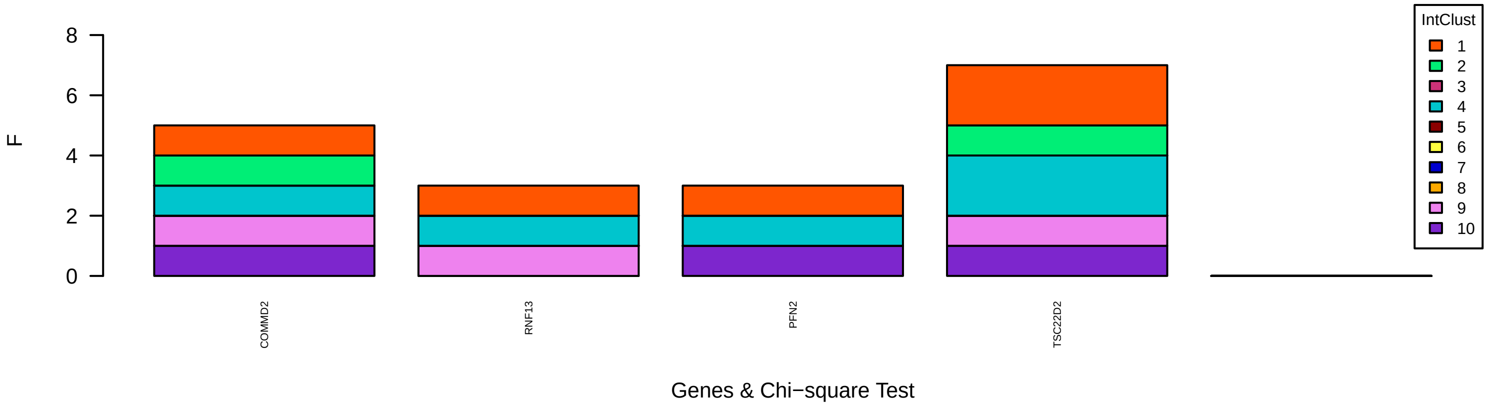
Chromosome 3 amplification outliers



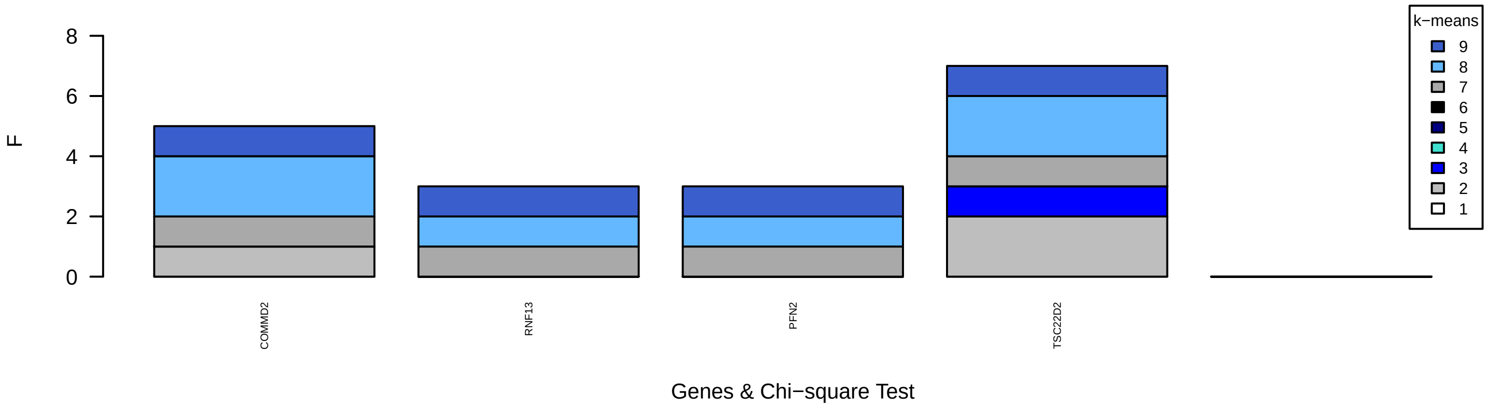
PAM50 region5 chr3:150720722−151660305 Subtypes Segregation − from COMMD2 to TSC22D2



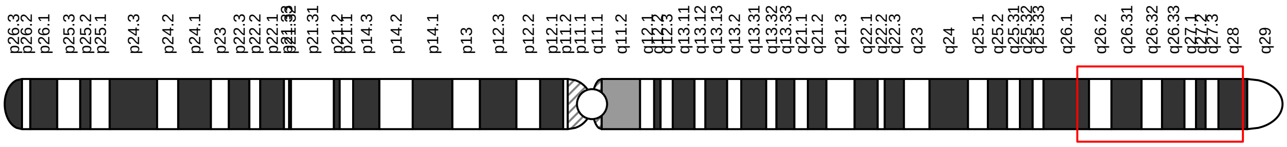
IntClust region5 chr3:150720722−151660305 Subtypes Segregation − from COMMD2 to TSC22D2



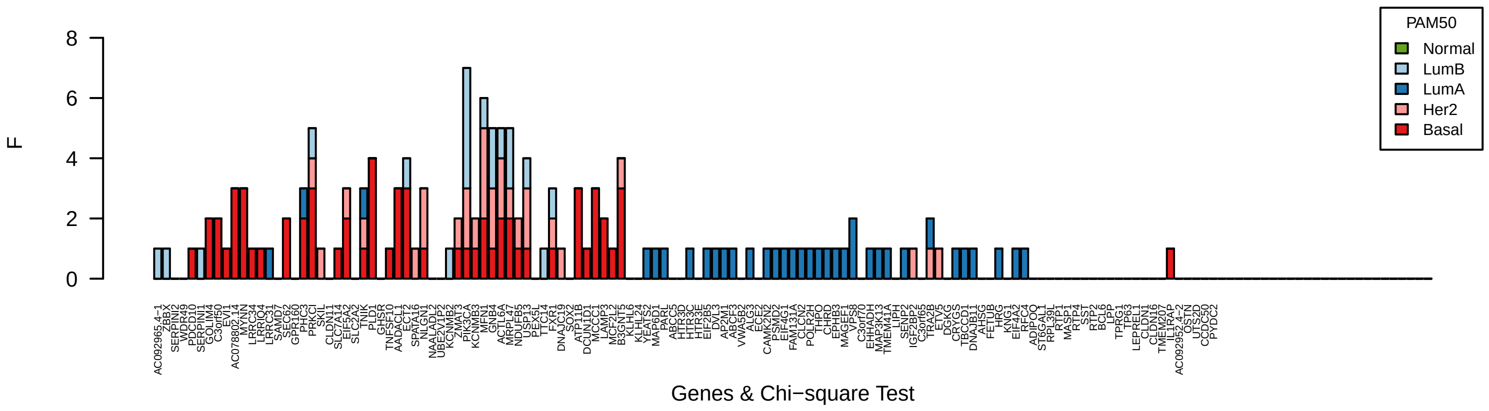
k−means region5 chr3:150720722−151660305 Subtypes Segregation − from COMMD2 to TSC22D2



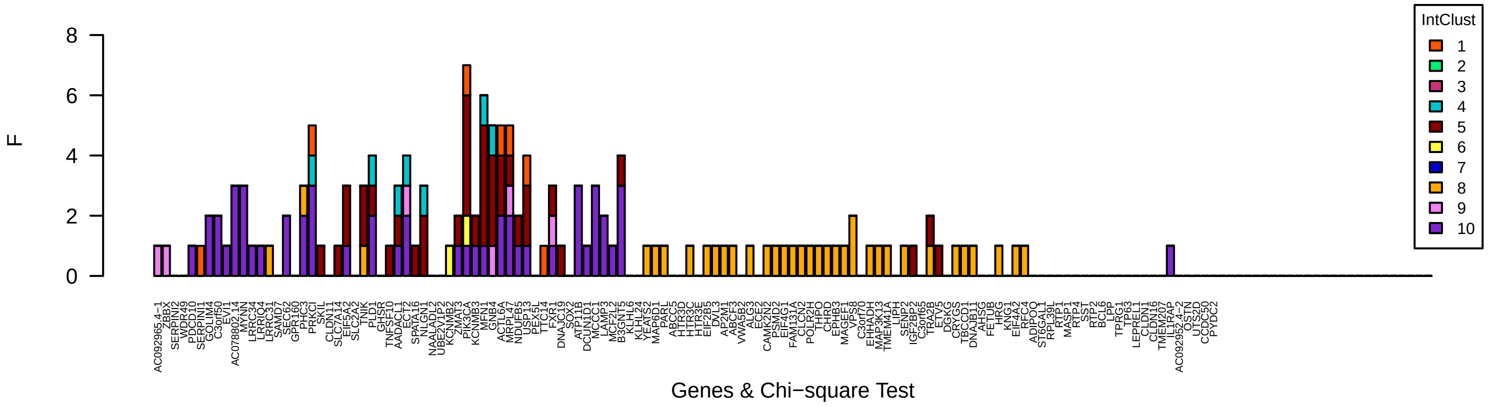
Chromosome 3 amplification outliers



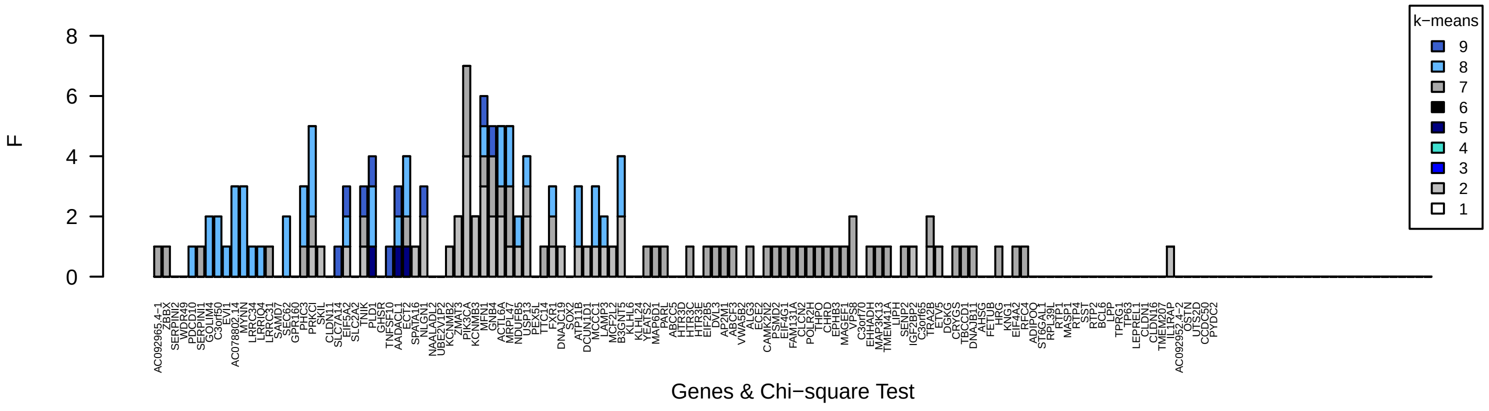
PAM50 region6 chr3:167464048−193320147 Subtypes Segregation − from AC092965.4−1 to PYDC2



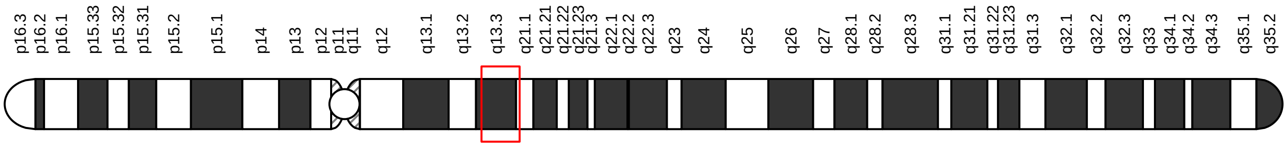
IntClust region6 chr3:167464048−193320147 Subtypes Segregation − from AC092965.4−1 to PYDC2



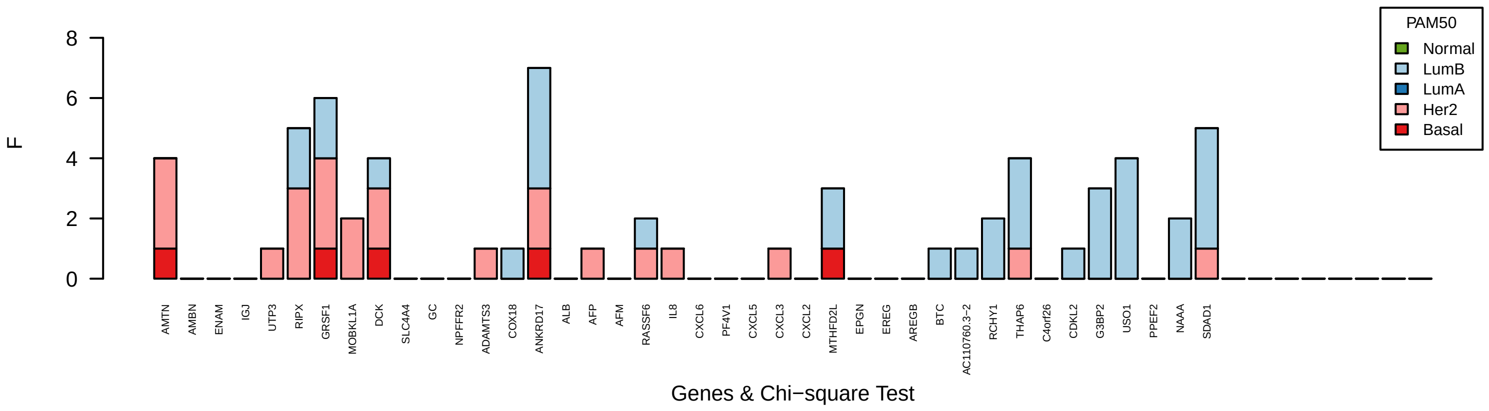
k−means region6 chr3:167464048−193320147 Subtypes Segregation − from AC092965.4−1 to PYDC2



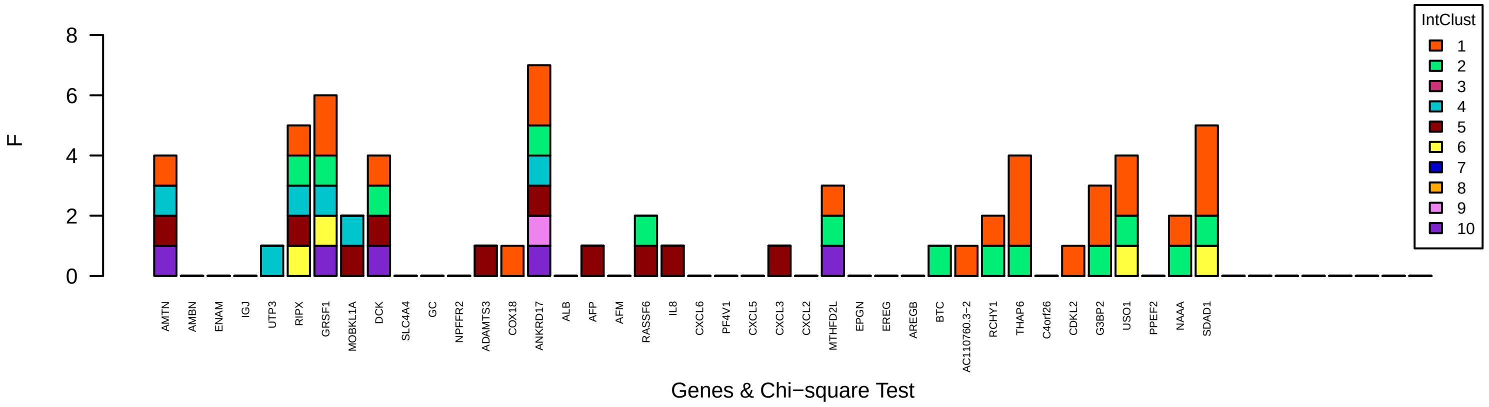
Chromosome 4 amplification outliers



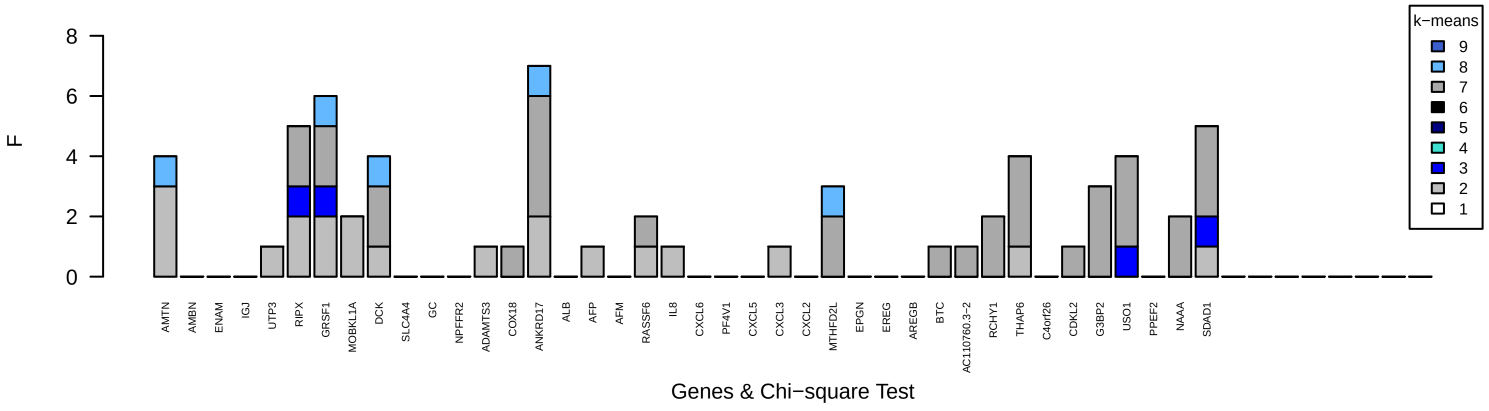
PAM50 region7 chr4:71418887−77131137 Subtypes Segregation − from AMTN to SDAD1



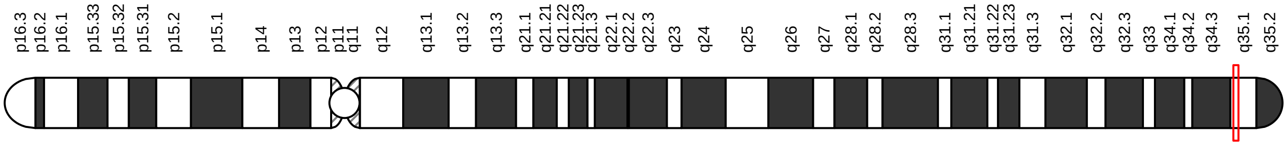
IntClust region7 chr4:71418887−77131137 Subtypes Segregation − from AMTN to SDAD1



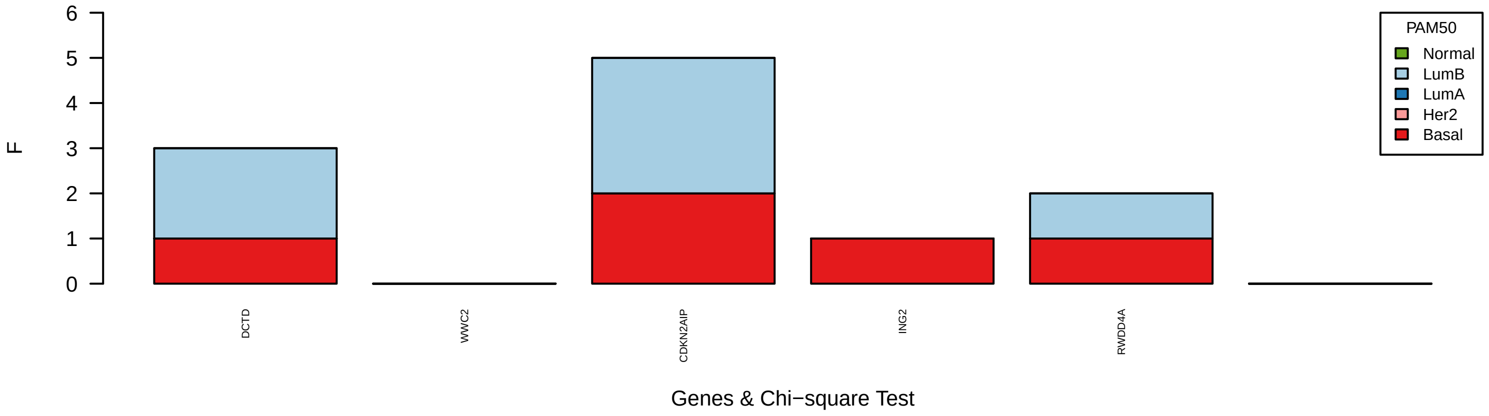
k−means region7 chr4:71418887−77131137 Subtypes Segregation − from AMTN to SDAD1



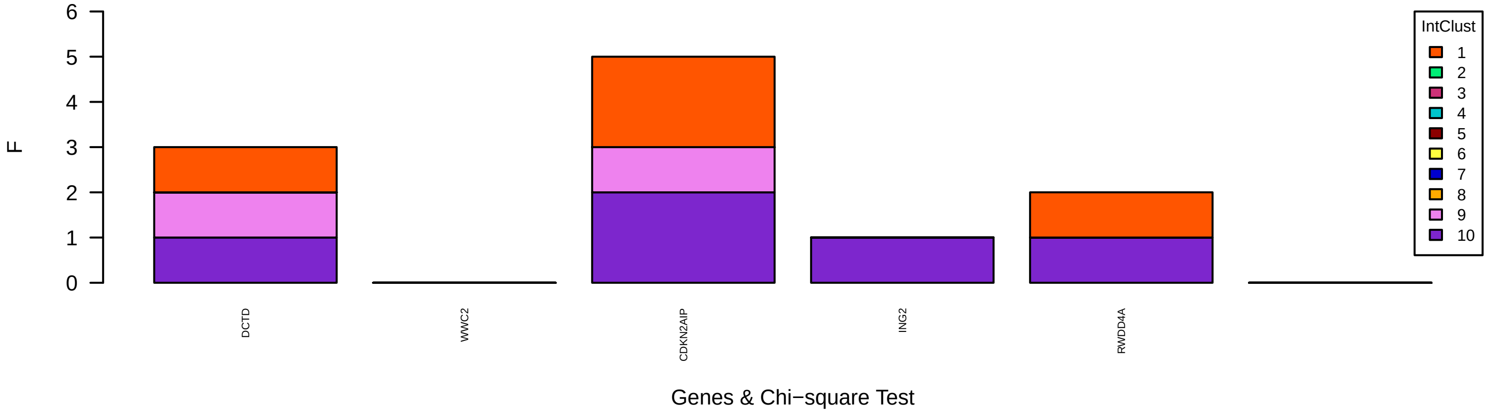
Chromosome 4 deletion outliers



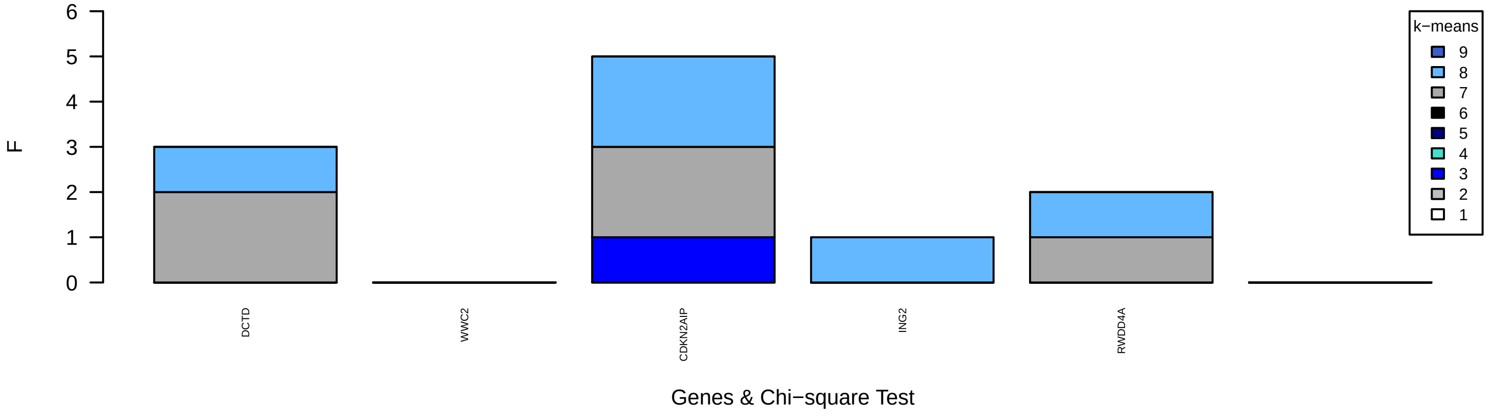
PAM50 region8 chr4:184048238−184817325 Subtypes Segregation − from DCTD to RWDD4A



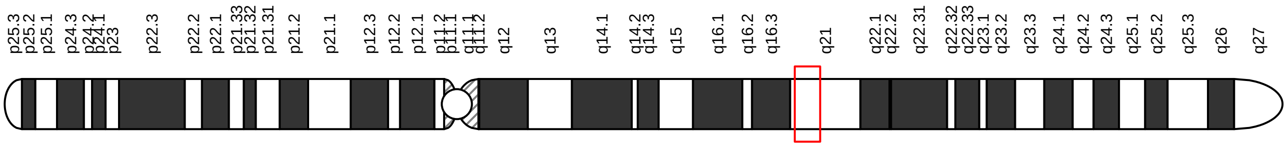
IntClust region8 chr4:184048238−184817325 Subtypes Segregation − from DCTD to RWDD4A



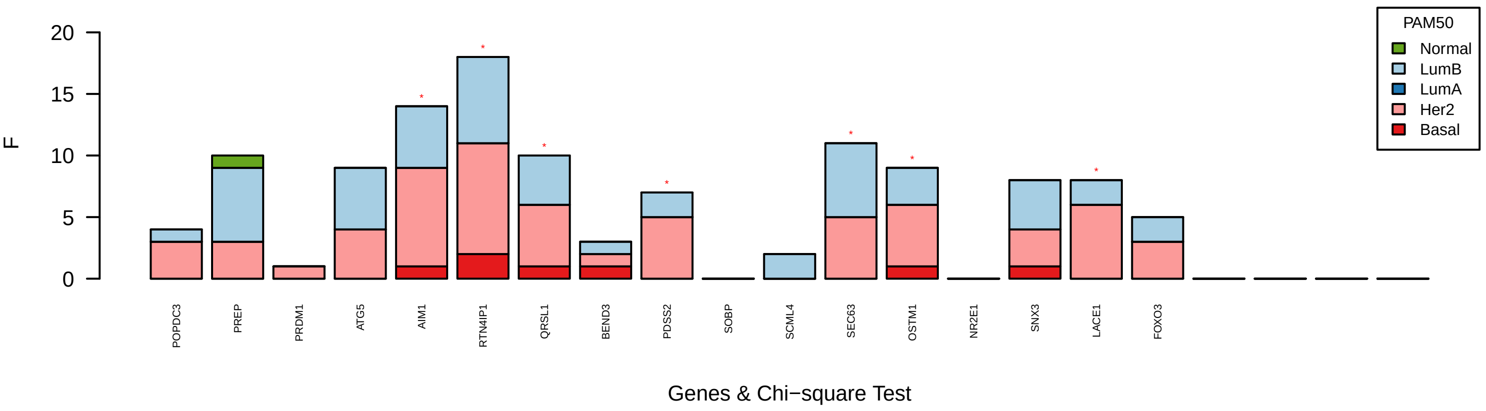
k−means region8 chr4:184048238−184817325 Subtypes Segregation − from DCTD to RWDD4A



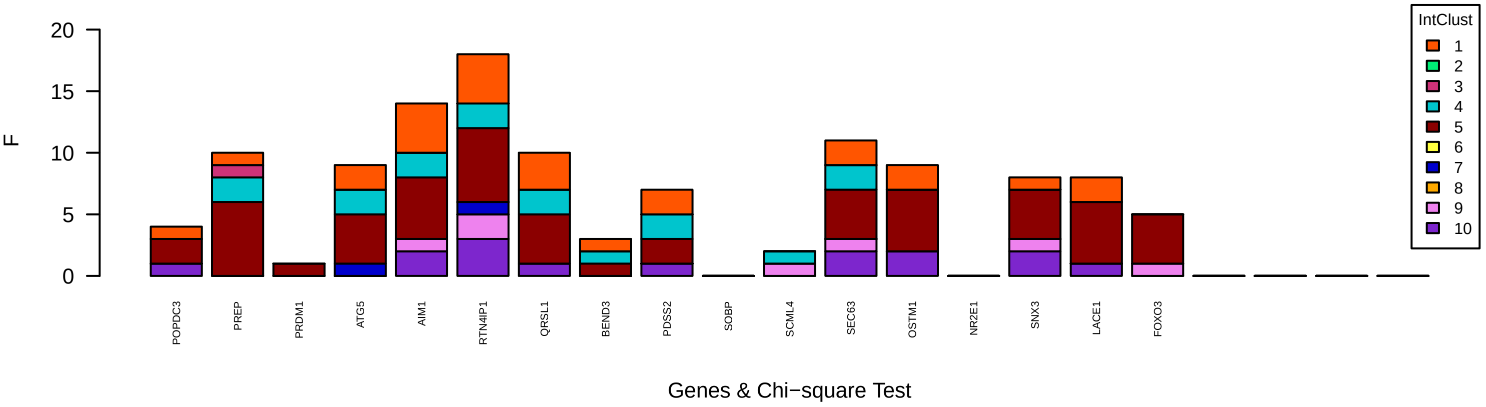
Chromosome 6 amplification outliers



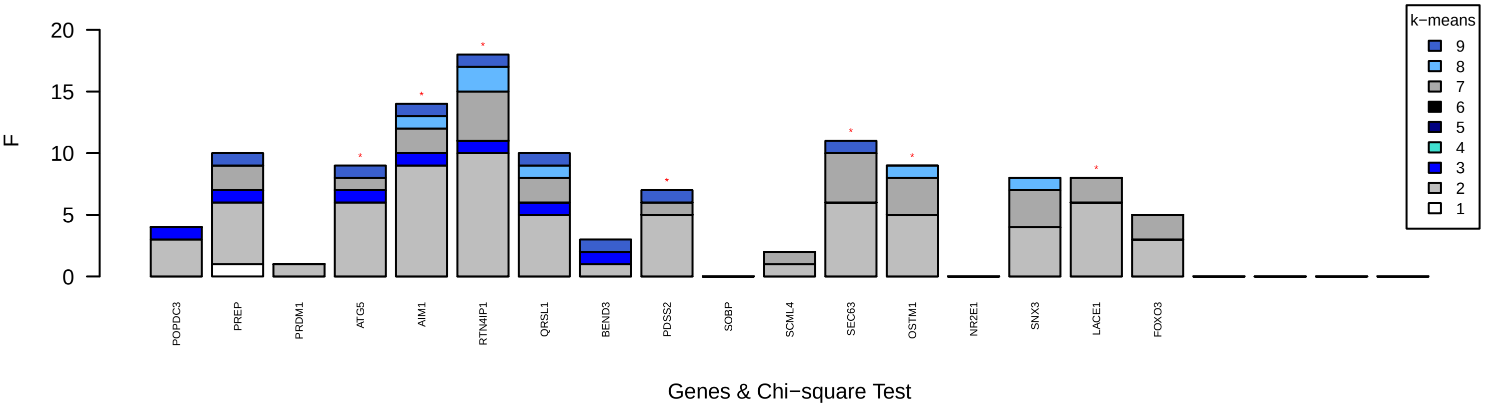
PAM50 region9 chr6:105712470−109112664 Subtypes Segregation − from POPDC3 to FOXO3



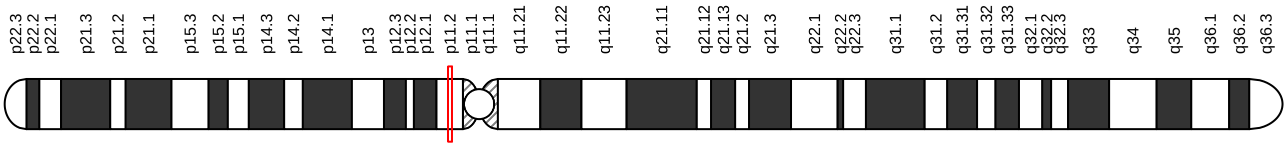
IntClust region9 chr6:105712470−109112664 Subtypes Segregation − from POPDC3 to FOXO3



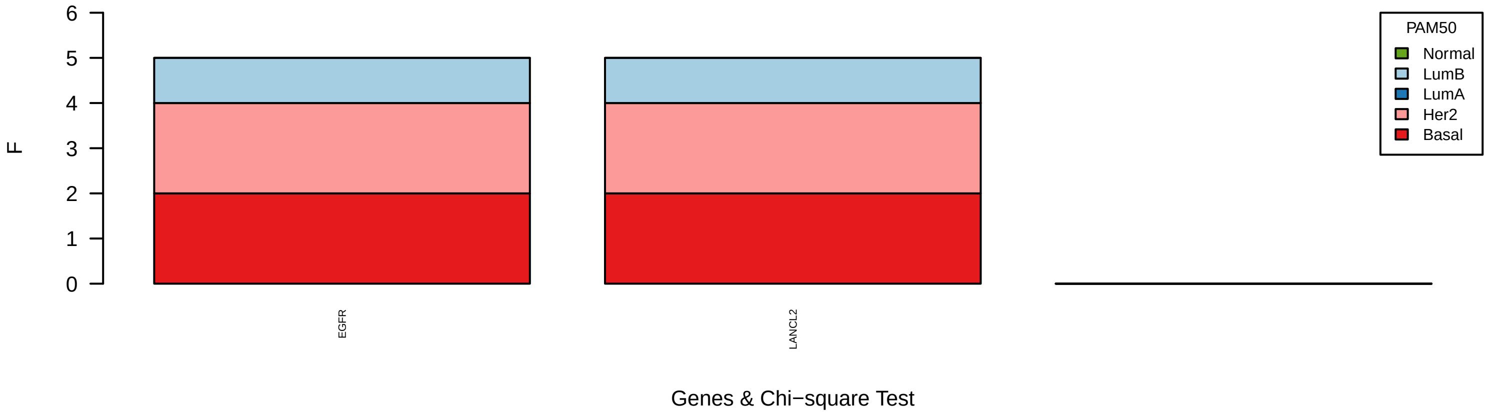
k−means region9 chr6:105712470−109112664 Subtypes Segregation − from POPDC3 to FOXO3



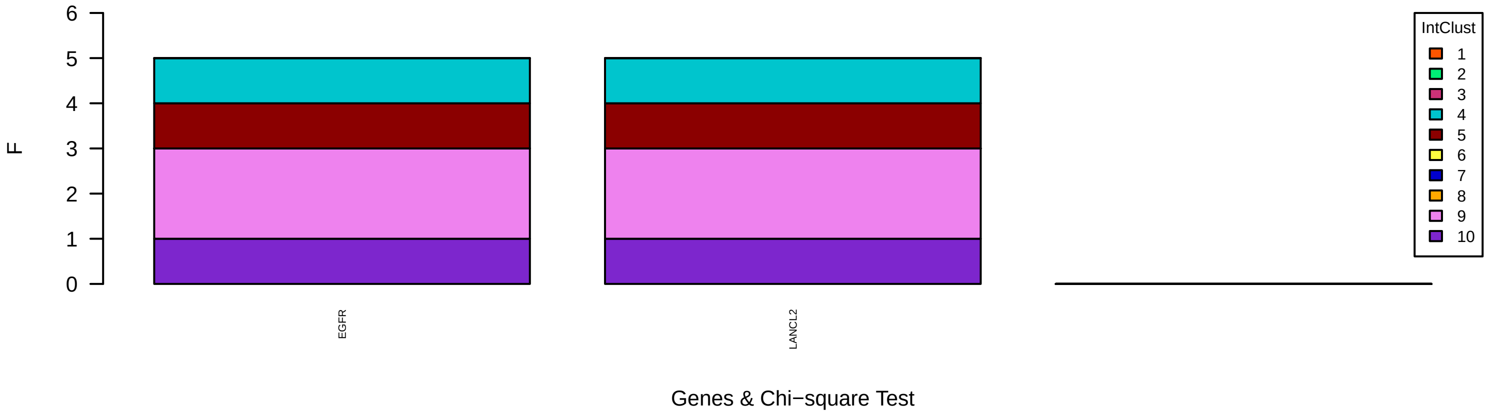
Chromosome 7 amplification outliers



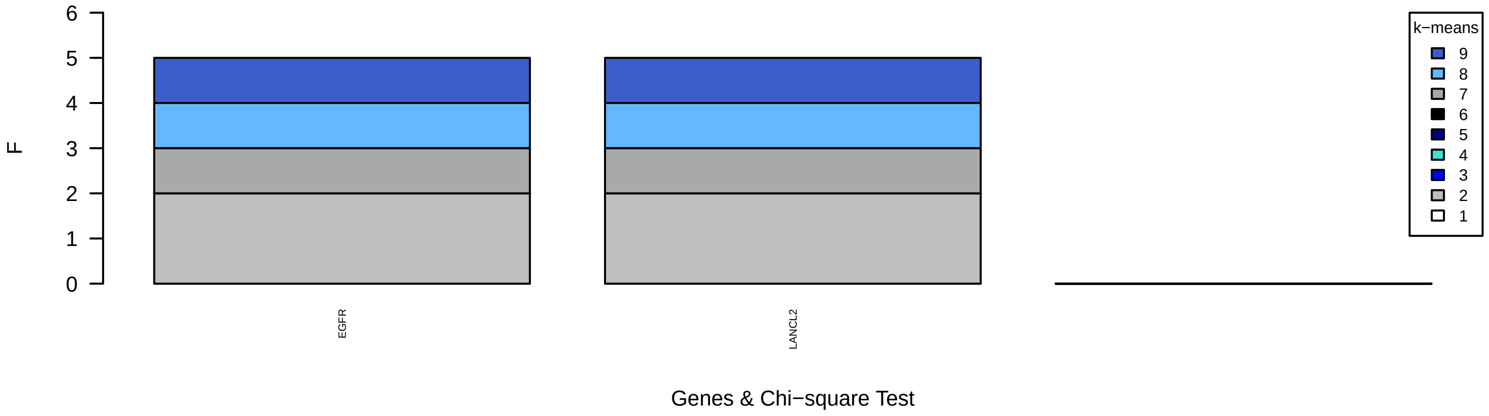
PAM50 region10 chr7:55054219−55545075 Subtypes Segregation − from EGFR to LANCL2



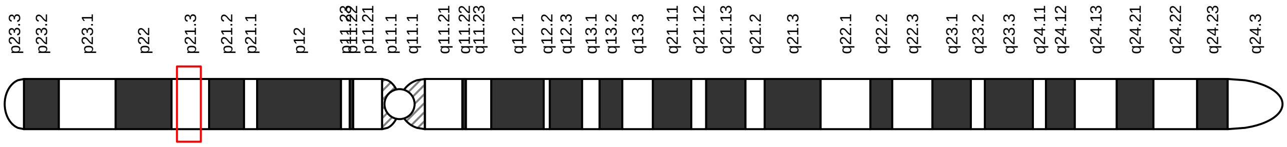
IntClust region10 chr7:55054219−55545075 Subtypes Segregation − from EGFR to LANCL2



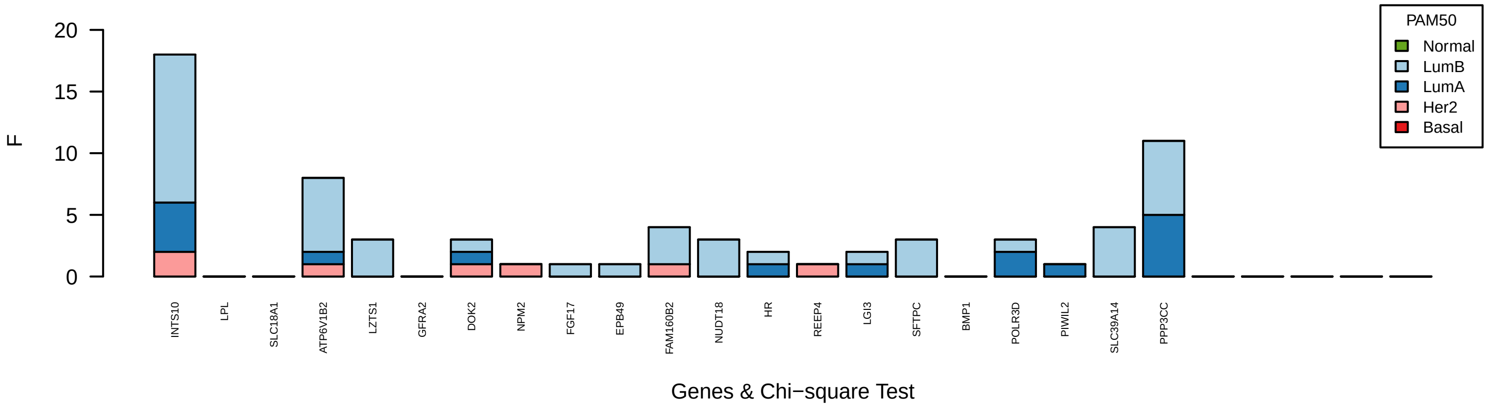
k−means region10 chr7:55054219−55545075 Subtypes Segregation − from EGFR to LANCL2



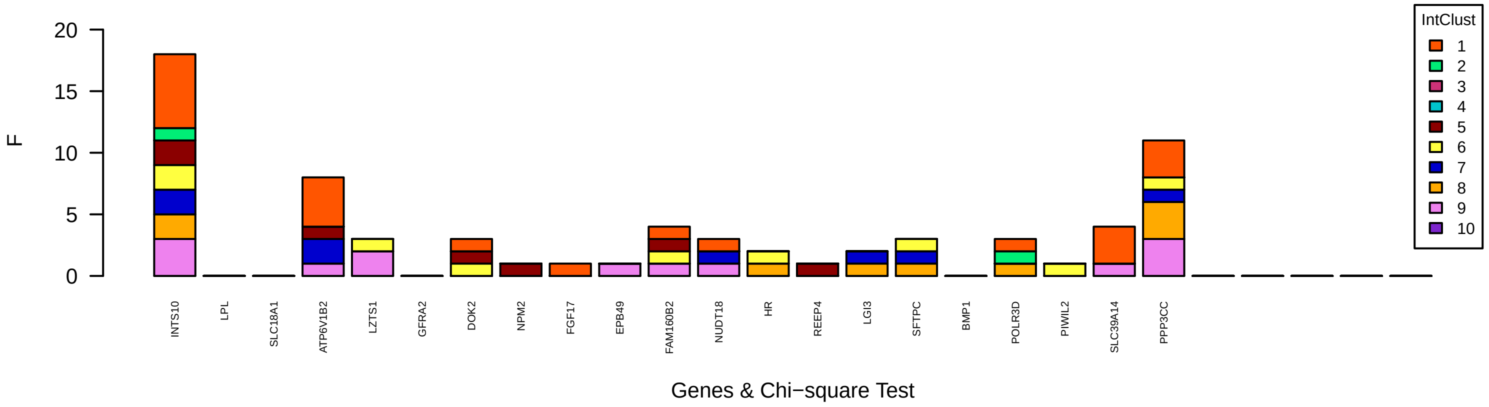
Chromosome 8 deletion outliers



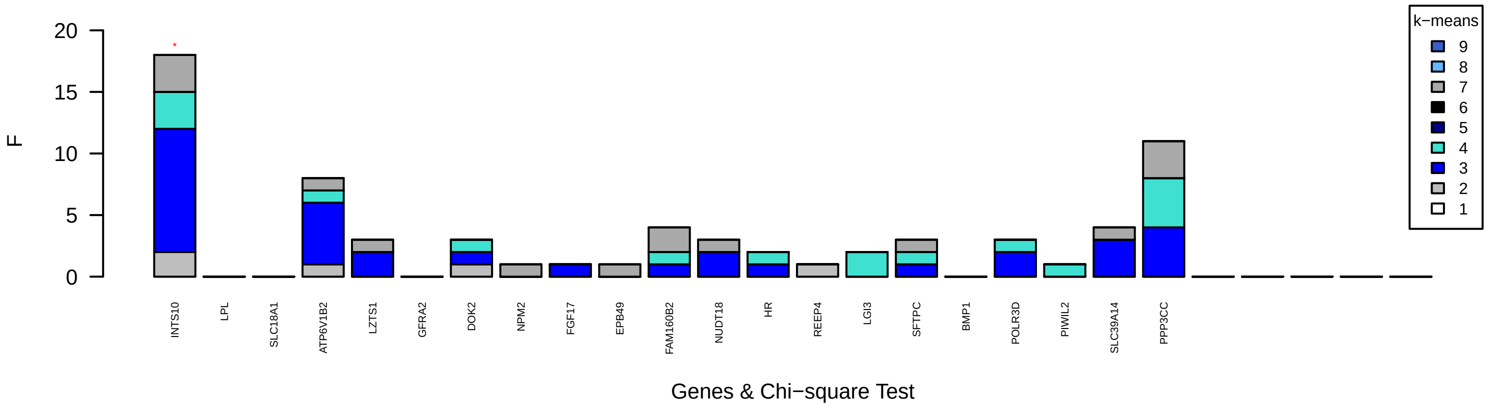
PAM50 region11 chr8:19719198−22454582 Subtypes Segregation − from INTS10 to PPP3CC



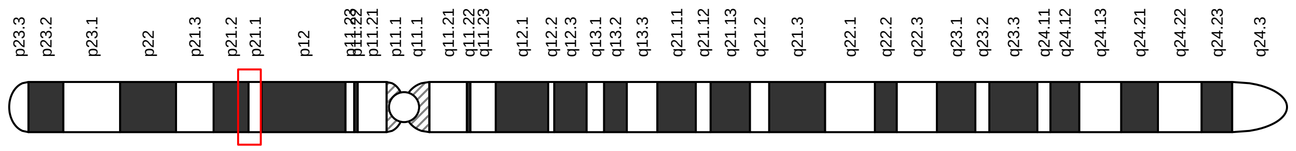
IntClust region11 chr8:19719198−22454582 Subtypes Segregation − from INTS10 to PPP3CC



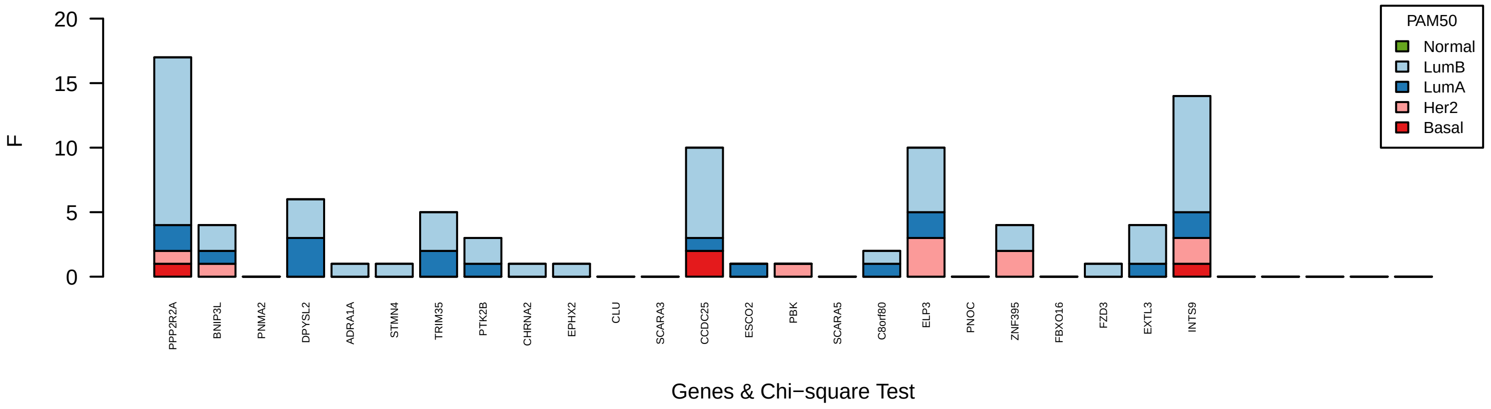
k−means region11 chr8:19719198−22454582 Subtypes Segregation − from INTS10 to PPP3CC



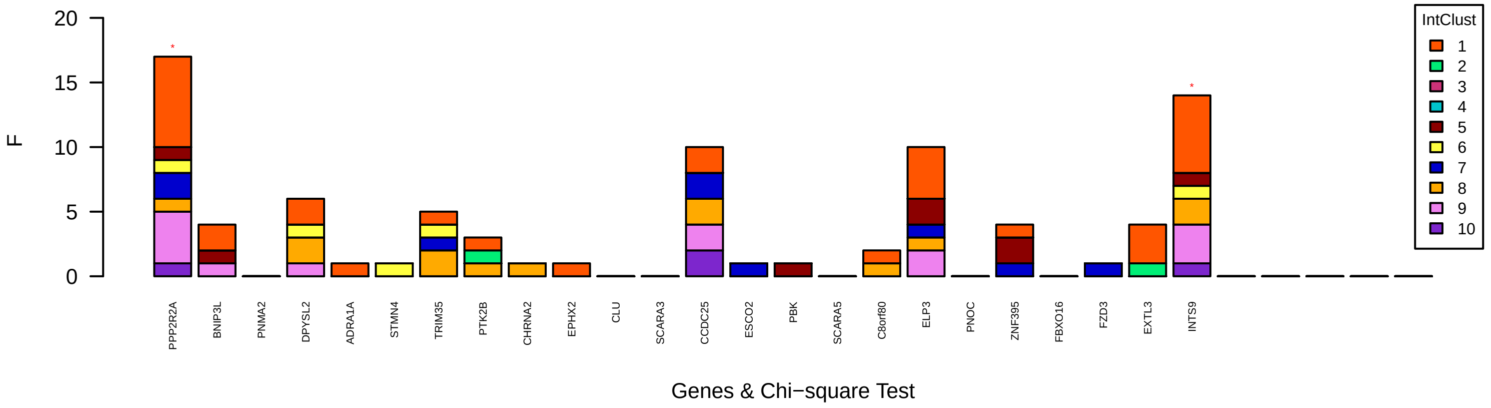
Chromosome 8 deletion outliers



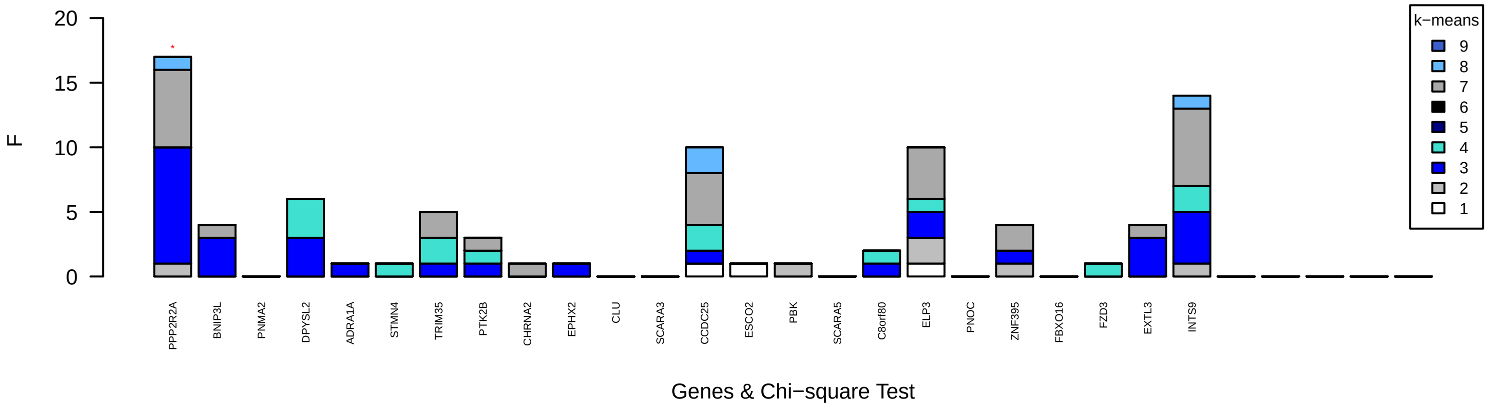
PAM50 region12 chr8:26204951−28803398 Subtypes Segregation − from PPP2R2A to INTS9



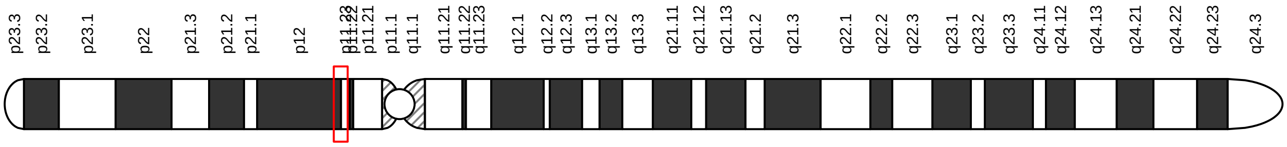
IntClust region12 chr8:26204951−28803398 Subtypes Segregation − from PPP2R2A to INTS9



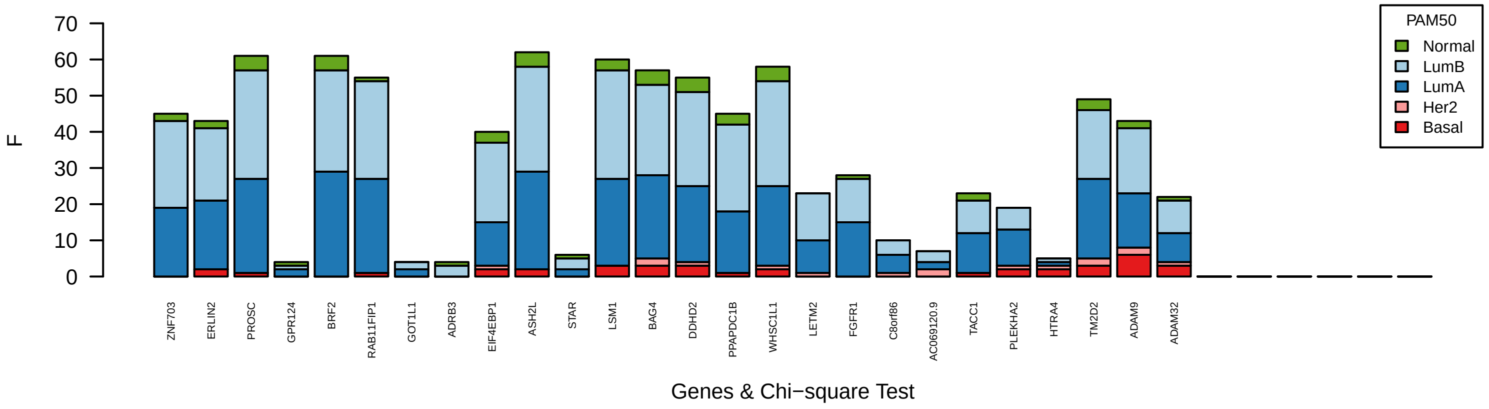
k−means region12 chr8:26204951−28803398 Subtypes Segregation − from PPP2R2A to INTS9



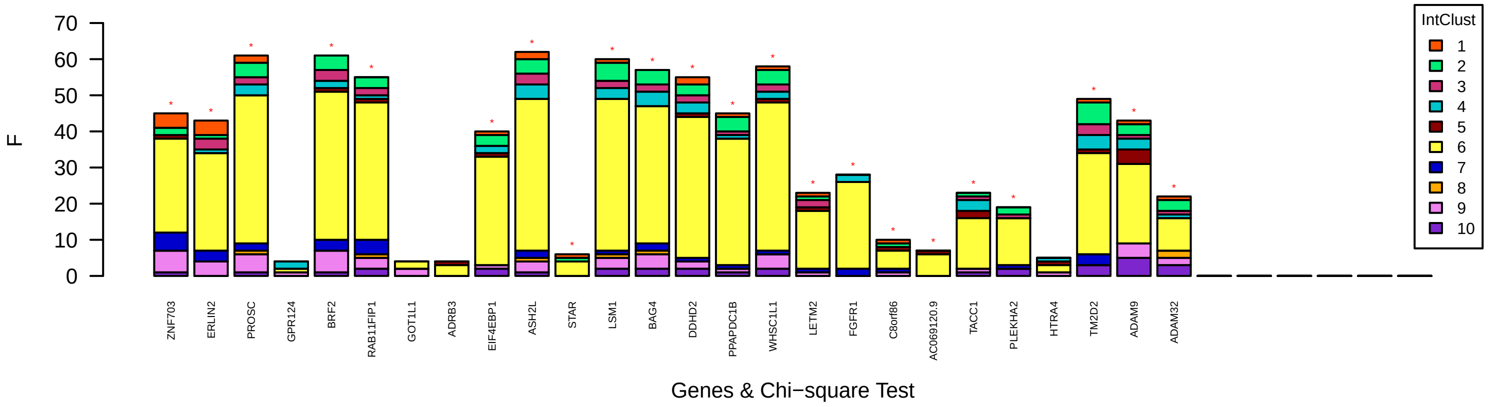
Chromosome 8 amplification outliers



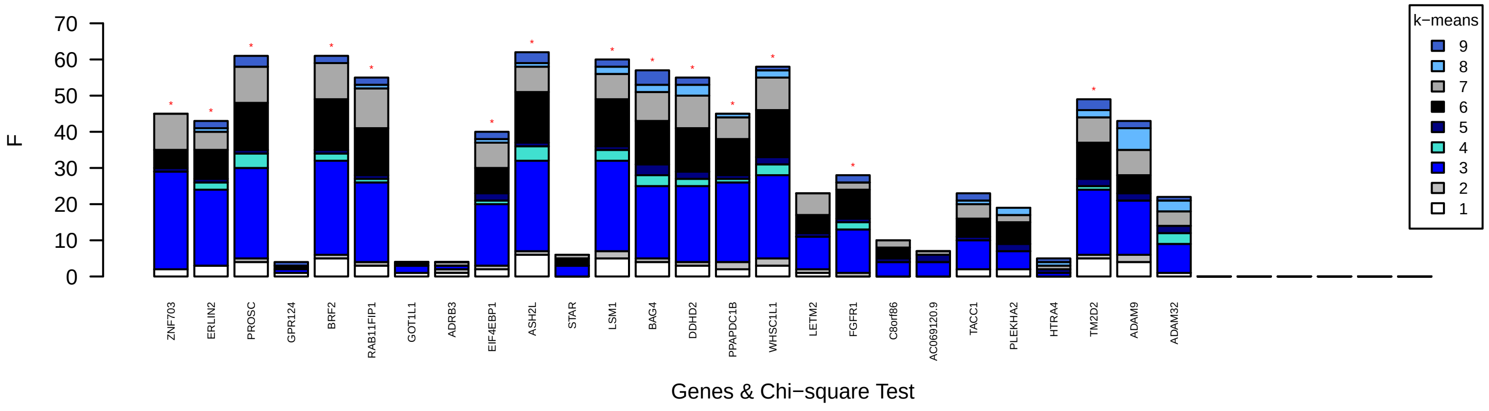
PAM50 region13 chr8:37672459−39261593 Subtypes Segregation − from ZNF703 to ADAM32



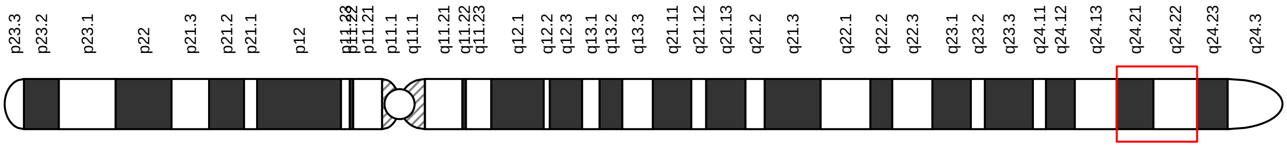
IntClust region13 chr8:37672459−39261593 Subtypes Segregation − from ZNF703 to ADAM32



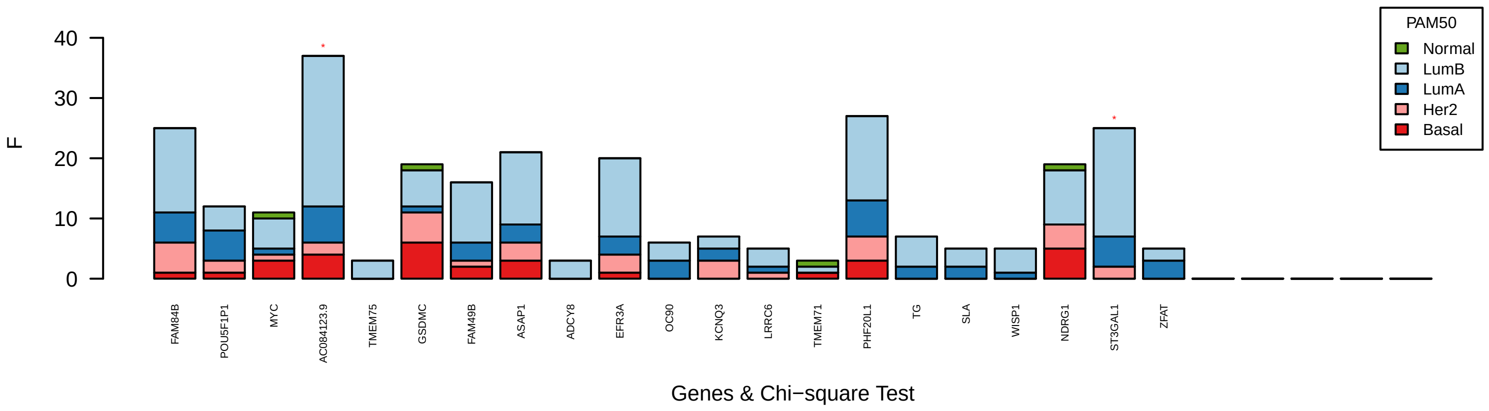
k−means region13 chr8:37672459−39261593 Subtypes Segregation − from ZNF703 to ADAM32



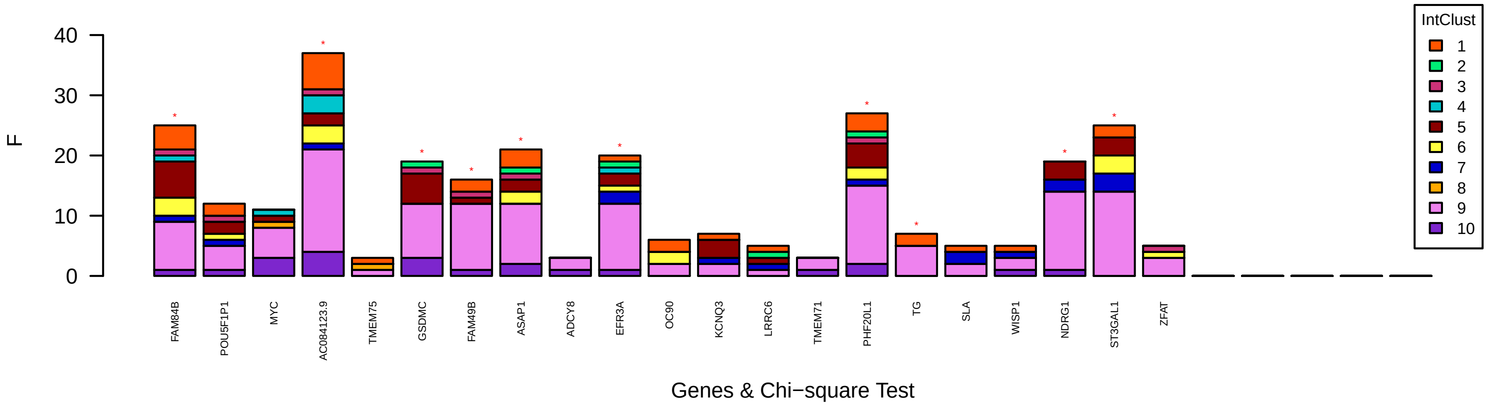
Chromosome 8 amplification outliers



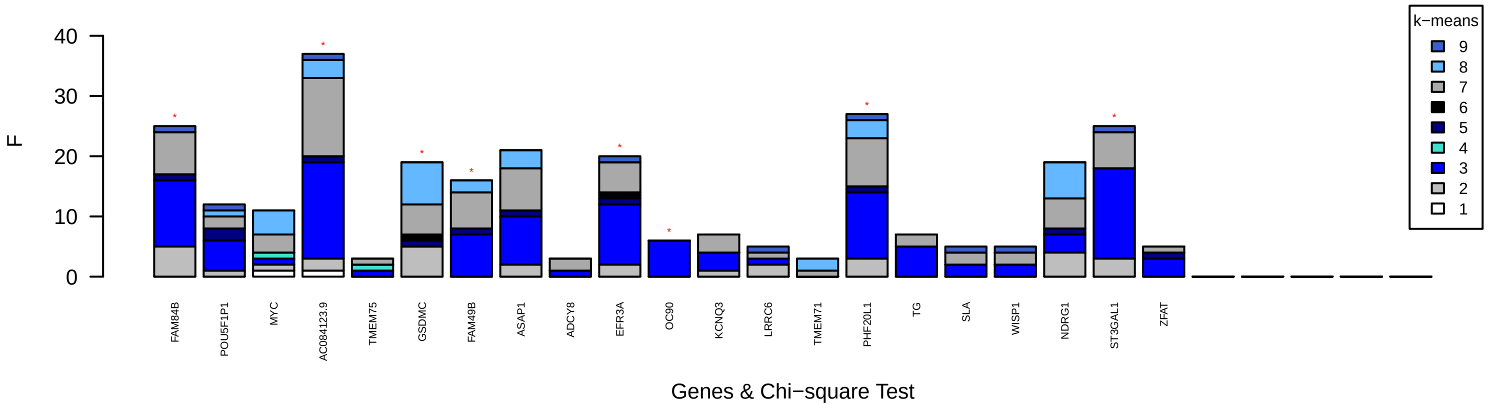
PAM50 region14 chr8:127300001−136500000 Subtypes Segregation − from FAM84B to ZFAT



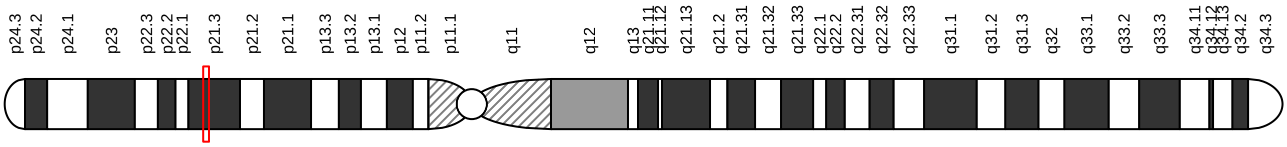
IntClust region14 chr8:127300001−136500000 Subtypes Segregation − from FAM84B to ZFAT



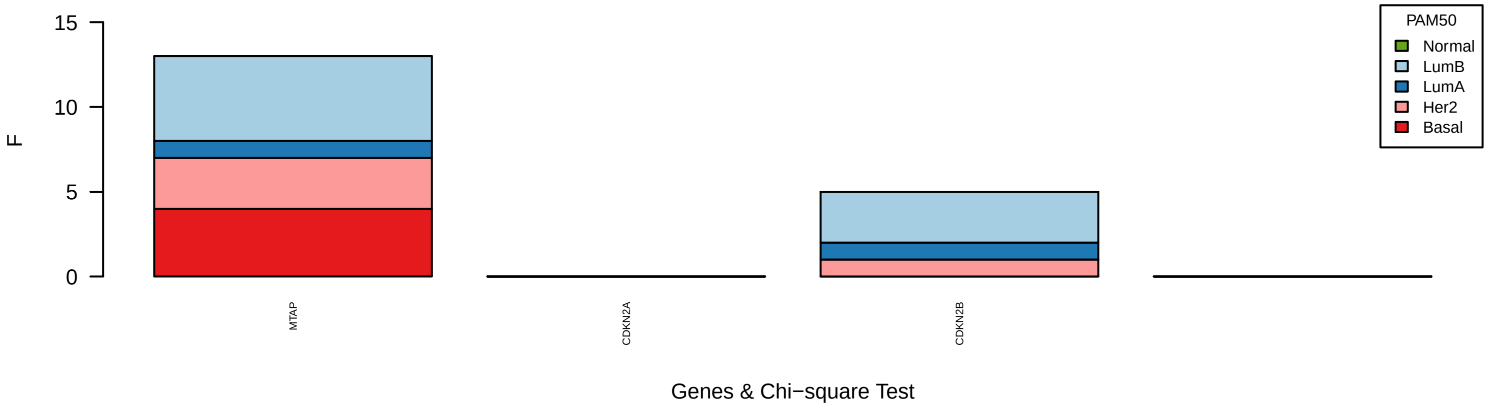
k−means region14 chr8:127300001−136500000 Subtypes Segregation − from FAM84B to ZFAT



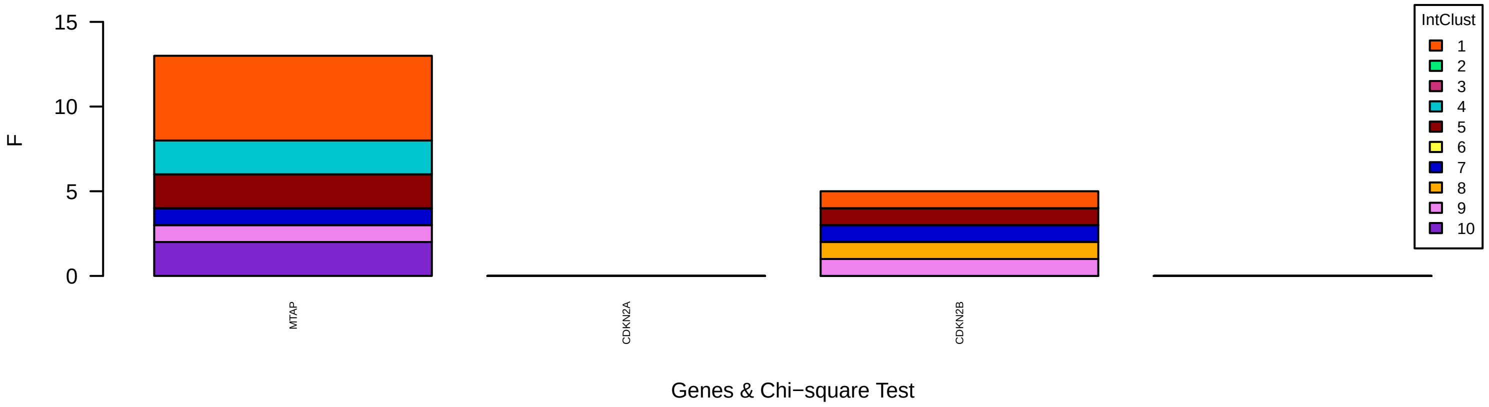
Chromosome 9 deletion outliers



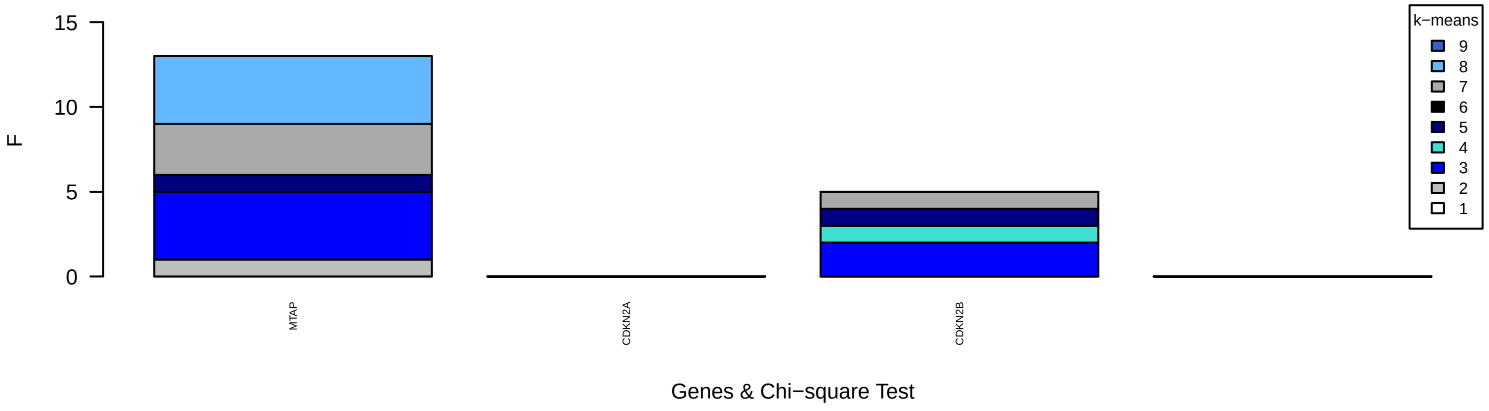
PAM50 region15 chr9:21507627−22140976 Subtypes Segregation − from MTAP to CDKN2B



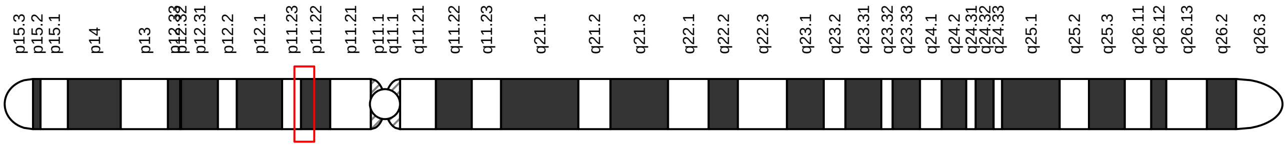
IntClust region15 chr9:21507627−22140976 Subtypes Segregation − from MTAP to CDKN2B



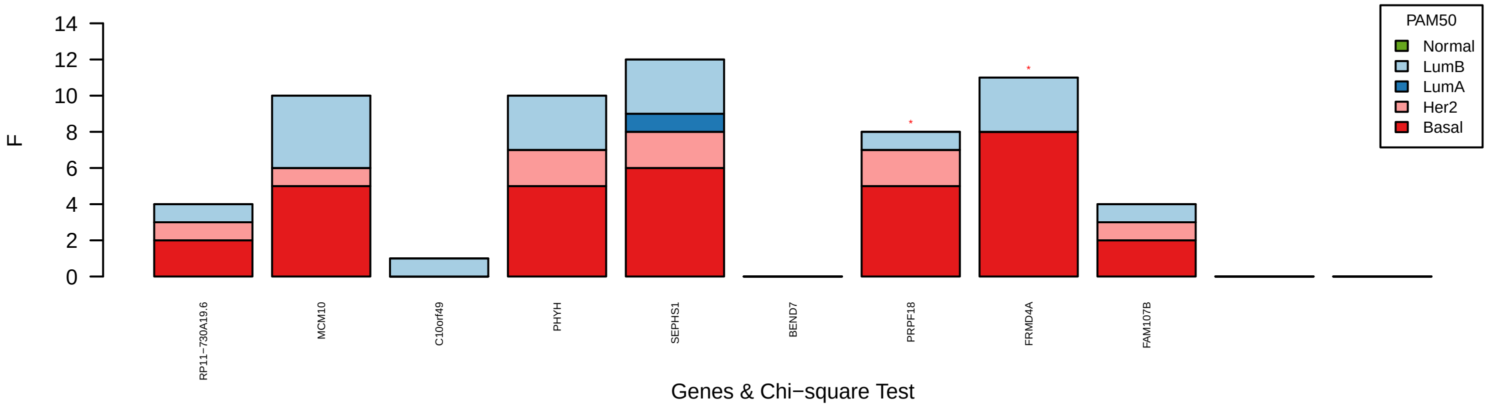
k−means region15 chr9:21507627−22140976 Subtypes Segregation − from MTAP to CDKN2B



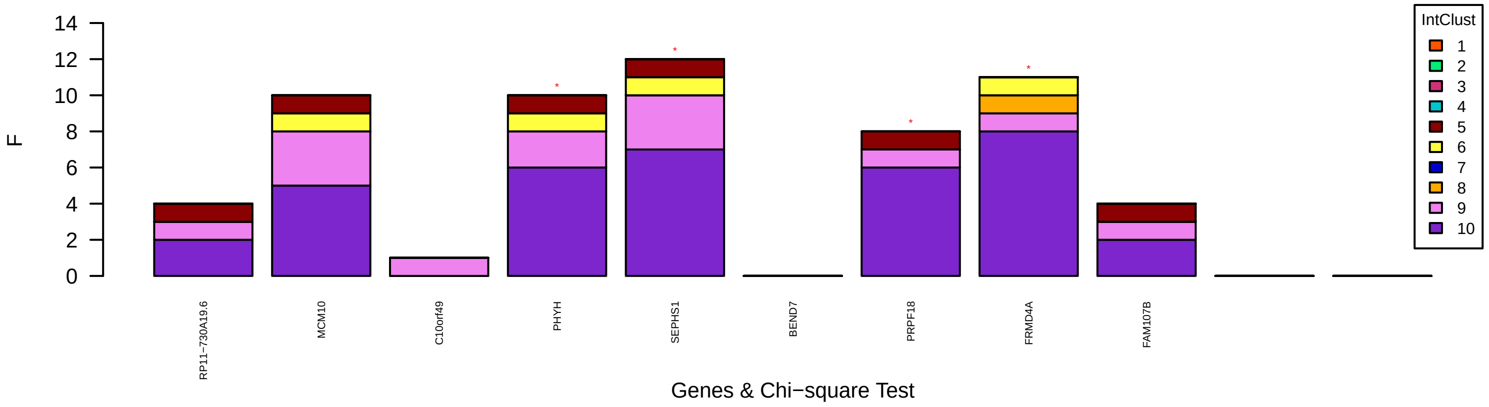
Chromosome 10 amplification outliers



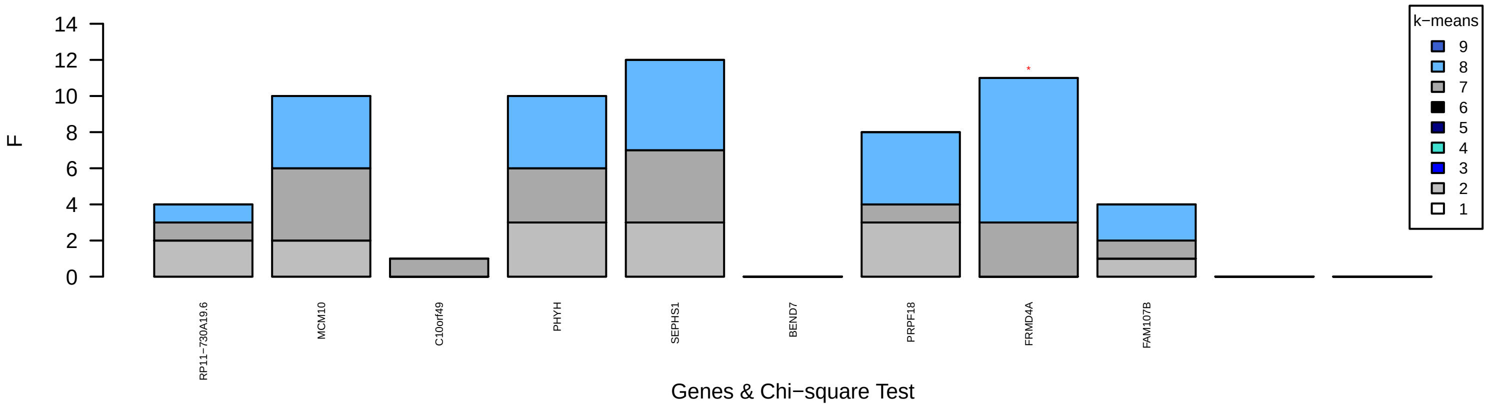
PAM50 region16 chr10:30702999−32803238 Subtypes Segregation − from RP11−730A19.6 to FAM107B



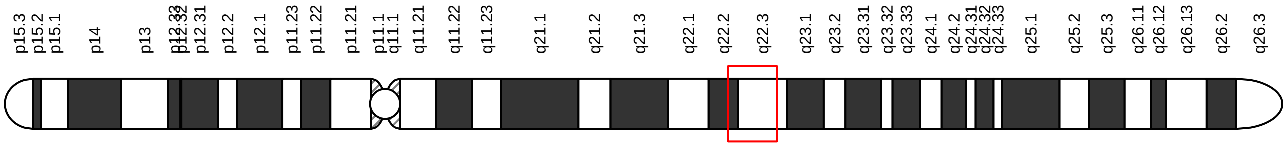
IntClust region16 chr10:30702999−32803238 Subtypes Segregation − from RP11−730A19.6 to FAM107B



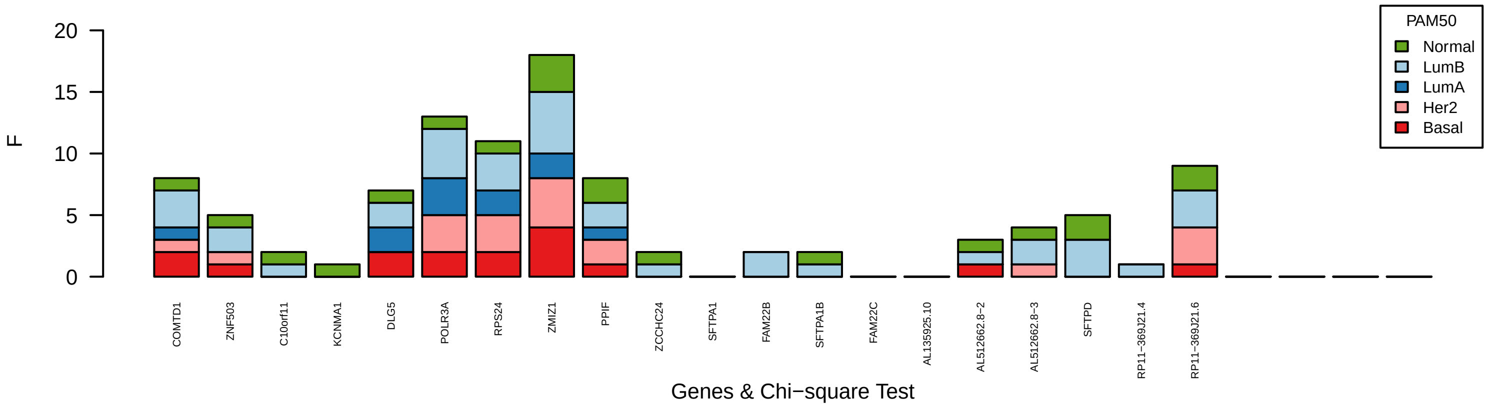
k−means region16 chr10:30702999−32803238 Subtypes Segregation − from RP11−730A19.6 to FAM107B



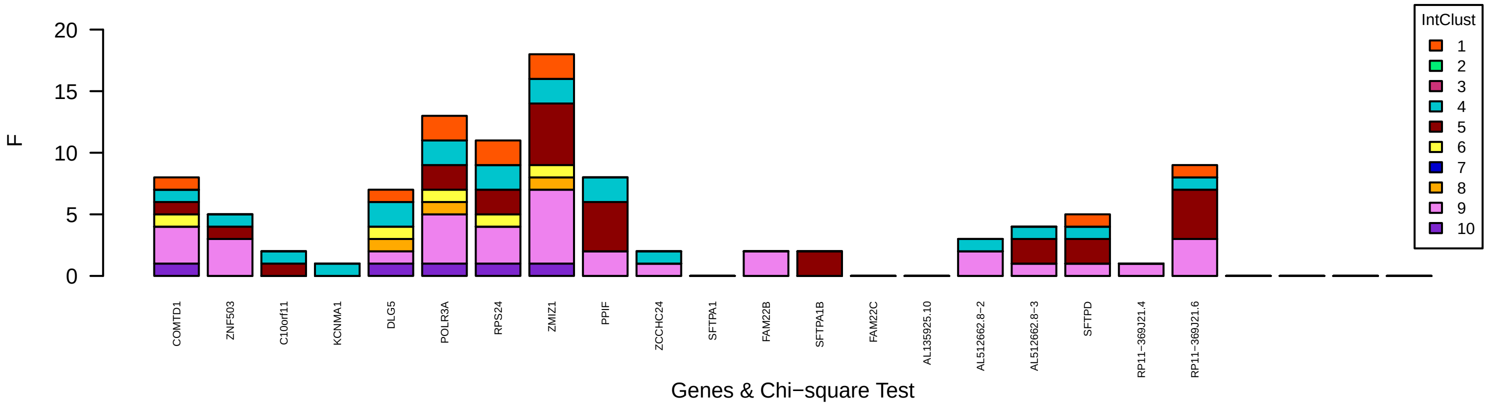
Chromosome 10 amplification outliers



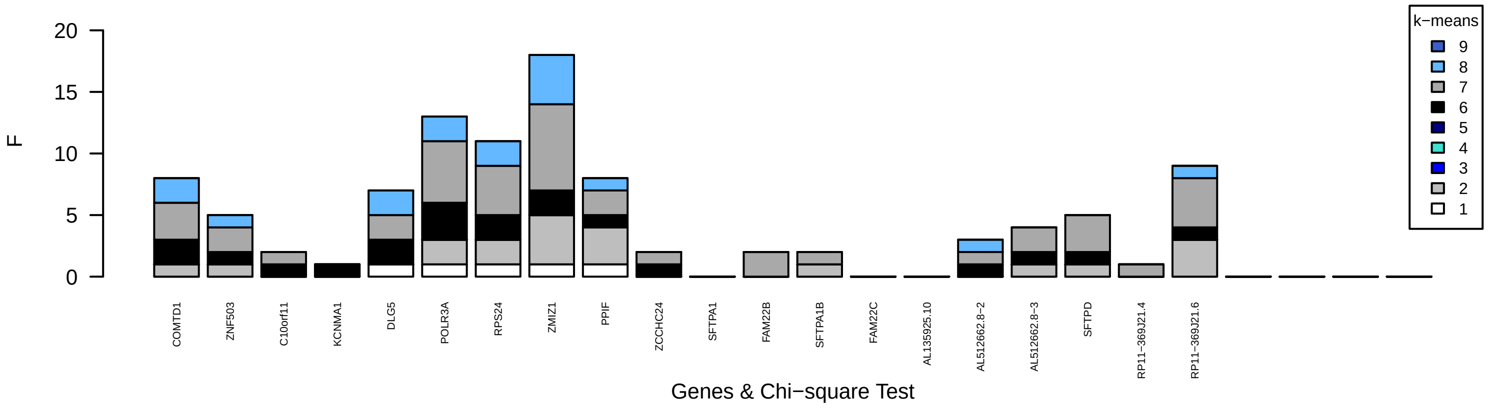
PAM50 region17 chr10:76663735−81842287 Subtypes Segregation − from COMTD1 to RP11−369J21.6

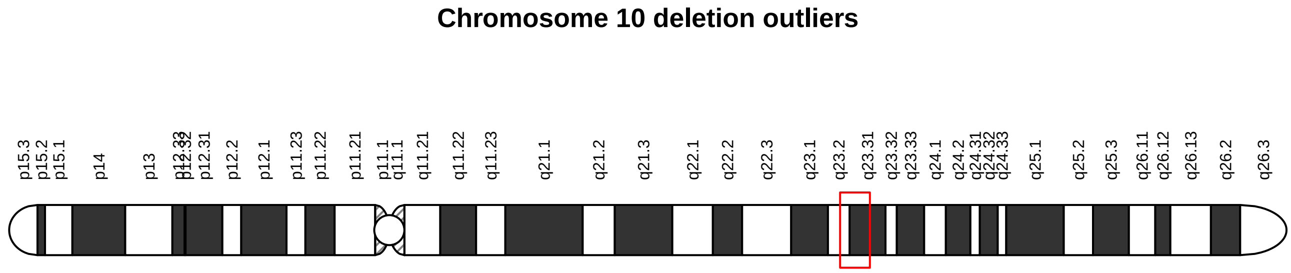


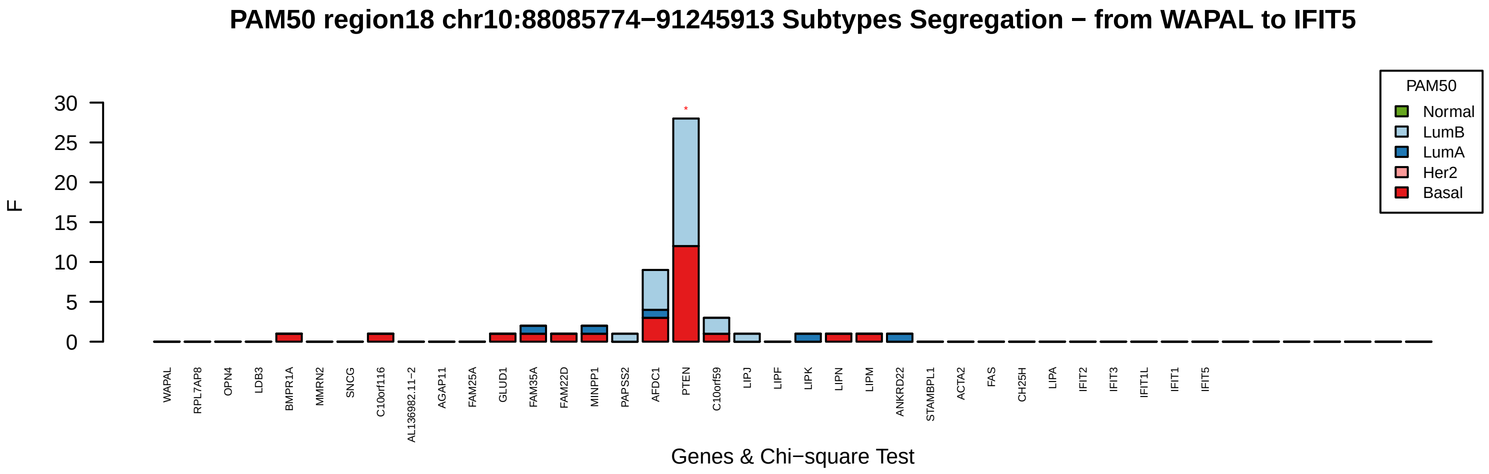
IntClust region17 chr10:76663735−81842287 Subtypes Segregation − from COMTD1 to RP11−369J21.6

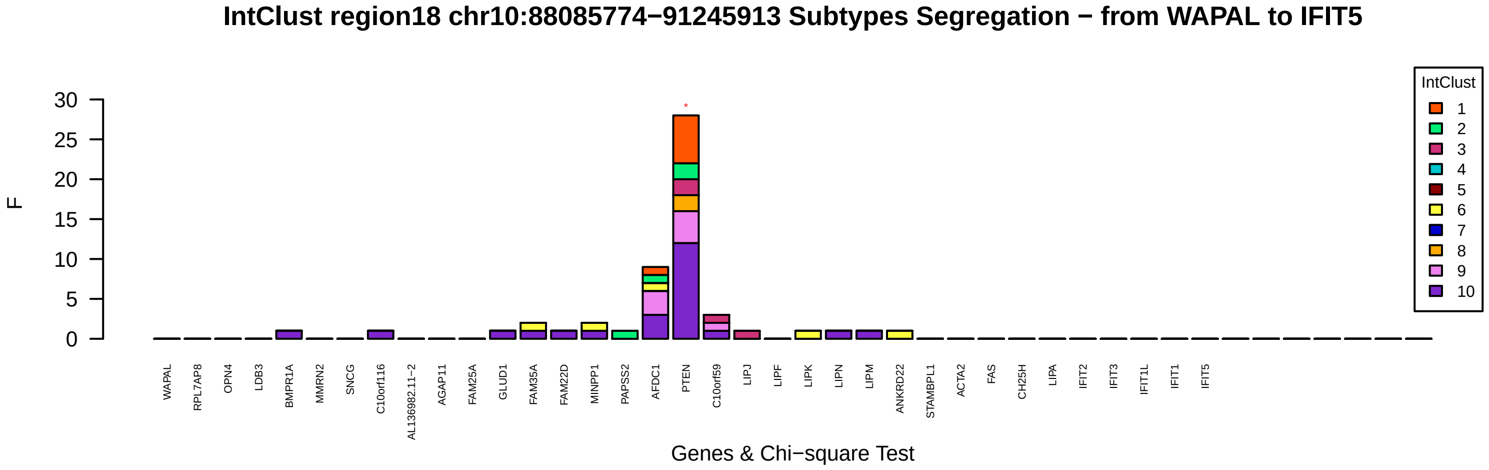


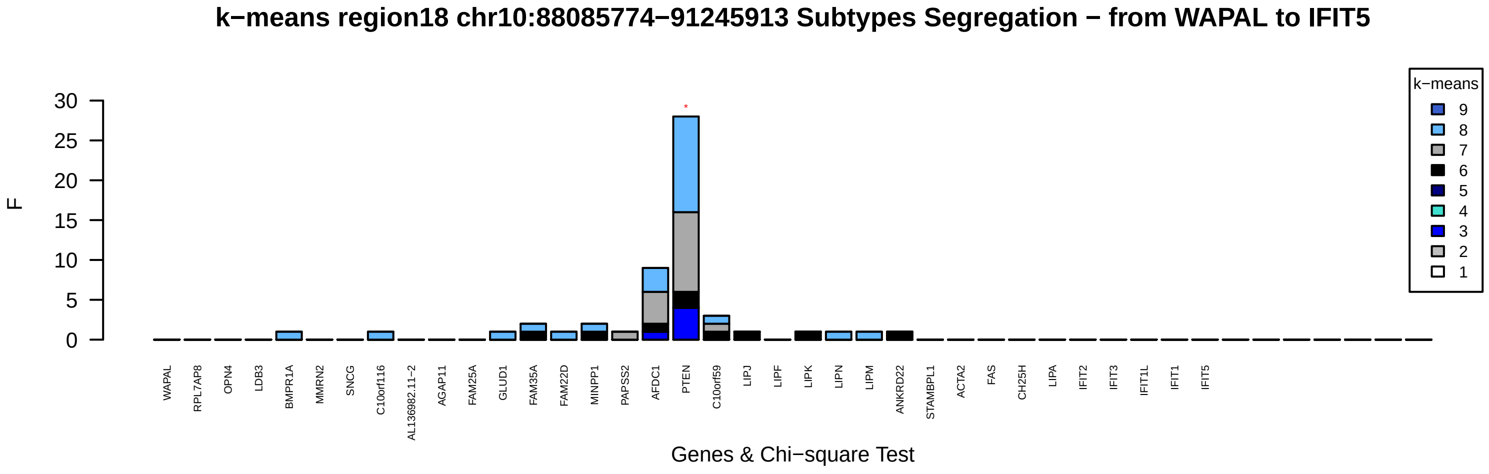
k−means region17 chr10:76663735−81842287 Subtypes Segregation − from COMTD1 to RP11−369J21.6



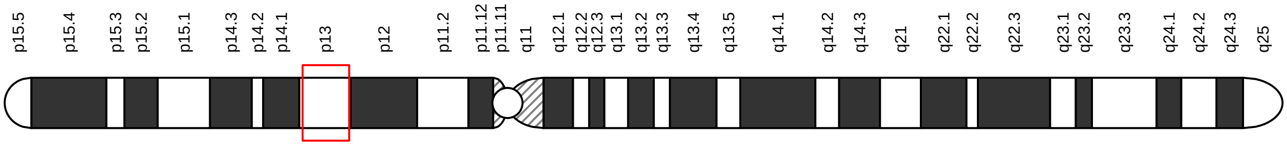

Chromosome 10 deletion outliers
p15.3
p15.2
p15.1
p14
p13
p12.33
p12.32
p12.31
p12.2
p12.1
p11.23
p11.22
p11.21
p11.1
q11.1
q11.21
q11.22
q11.23
q21.1
q21.2
q21.3
q22.1
q22.2
q22.3
q23.1
q23.2
q23.31
q23.32
q23.33
q24.1
q24.2
q24.31
q24.32
q24.33
q25.1
q25.2
q25.3
q26.11
q26.12
q26.13
q26.2
q26.3


PAM50 region18 chr10:88085774−91245913 Subtypes Segregation − from WAPAL to IFIT5
F
30
25
20
15
10
5
0
*
PAM50
Normal
LumB
LumA
Her2
Basal
WAPAL
RPL7AP8
OPN4
LDB3
BMPR1A
MMRN2
SNCG
C10orf116
AL136982.11−2
AGAP11
FAM25A
GLUD1
FAM35A
FAM22D
MINPP1
PAPSS2
AFDC1
PTEN
C10orf59
LIPJ
LIPF
LIPK
LIPN
LIPM
ANKRD22
STAMBPL1
ACTA2
FAS
CH25H
LIPA
IFIT2
IFIT3
IFIT1L
IFIT1
IFIT5
Genes & Chi−square Test

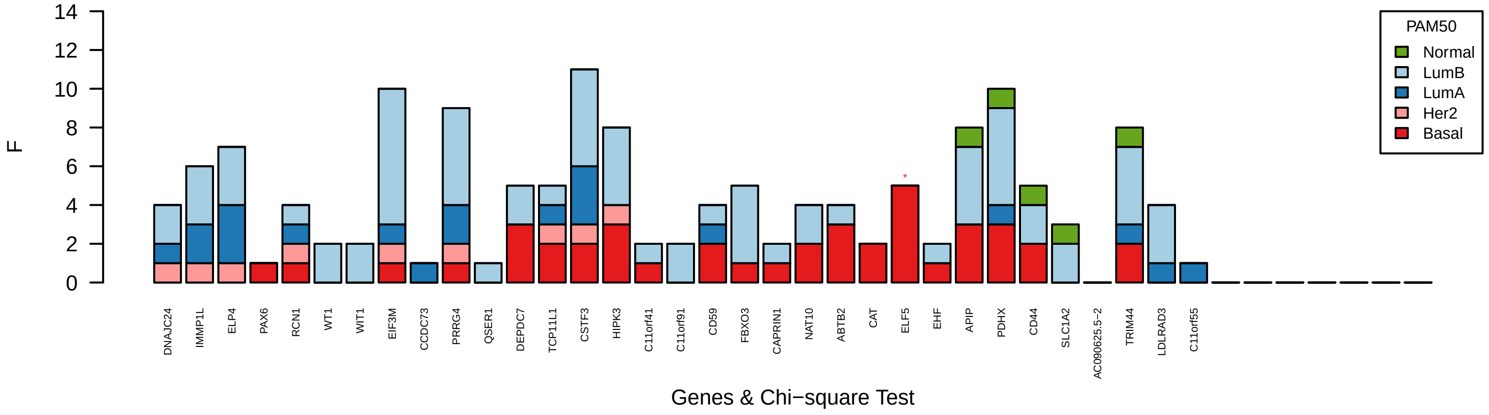

IntClust region18 chr10:88085774−91245913 Subtypes Segregation − from WAPAL to IFIT5
F
30
25
20
15
10
5
0
*
IntClust
1
2
3
4
5
6
7
8
9
10
WAPAL
RPL7AP8
OPN4
LDB3
BMPR1A
MMRN2
SNCG
C10orf116
AL136982.11−2
AGAP11
FAM25A
GLUD1
FAM35A
FAM22D
MINPP1
PAPSS2
AFDC1
PTEN
C10orf59
LIPJ
LIPF
LIPK
LIPN
LIPM
ANKRD22
STAMBPL1
ACTA2
FAS
CH25H
LIPA
IFIT2
IFIT3
IFIT1L
IFIT1
IFIT5
Genes & Chi−square Test


k−means region18 chr10:88085774−91245913 Subtypes Segregation − from WAPAL to IFIT5
F
30
25
20
15
10
5
0
*
k−means
9
8
7
6
5
4
3
2
1
WAPAL
RPL7AP8
OPN4
LDB3
BMPR1A
MMRN2
SNCG
C10orf116
AL136982.11−2
AGAP11
FAM25A
GLUD1
FAM35A
FAM22D
MINPP1
PAPSS2
AFDC1
PTEN
C10orf59
LIPJ
LIPF
LIPK
LIPN
LIPM
ANKRD22
STAMBPL1
ACTA2
FAS
CH25H
LIPA
IFIT2
IFIT3
IFIT1L
IFIT1
IFIT5
Genes & Chi−square Test

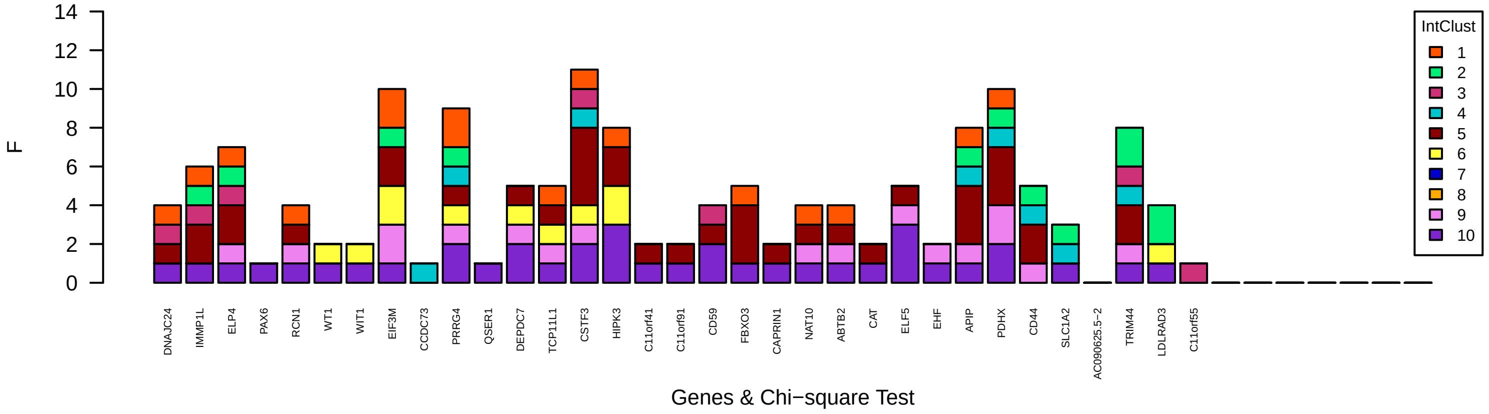
Chromosome 11 amplification outliers



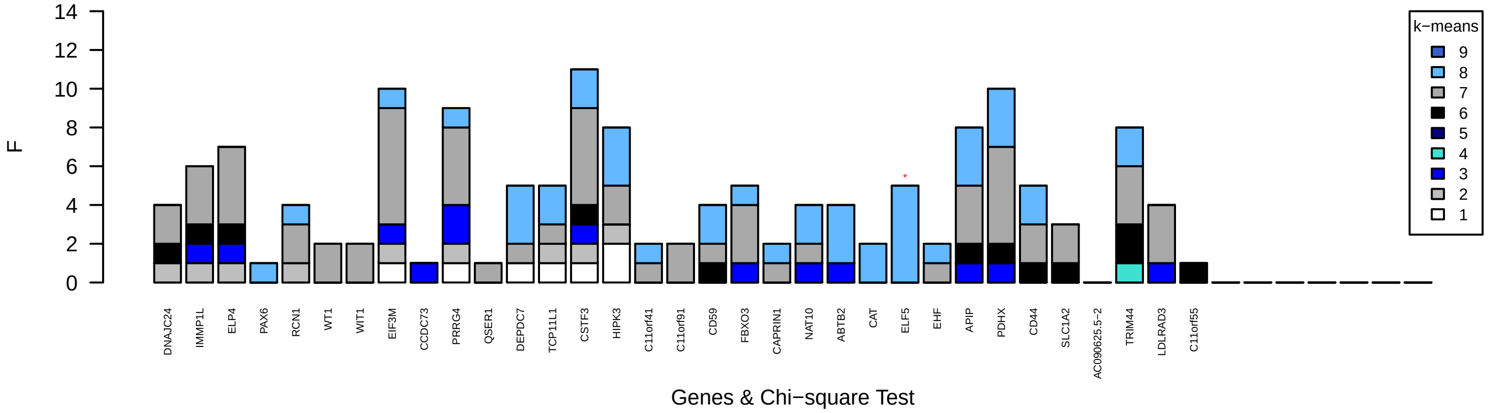
PAM50 region19 chr11:31347953−36251485 Subtypes Segregation − from DNAJC24 to C11orf55



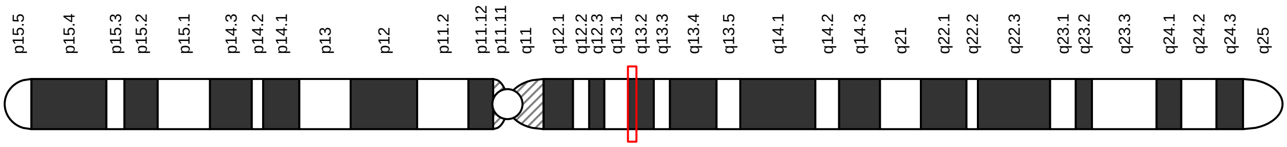
IntClust region19 chr11:31347953−36251485 Subtypes Segregation − from DNAJC24 to C11orf55

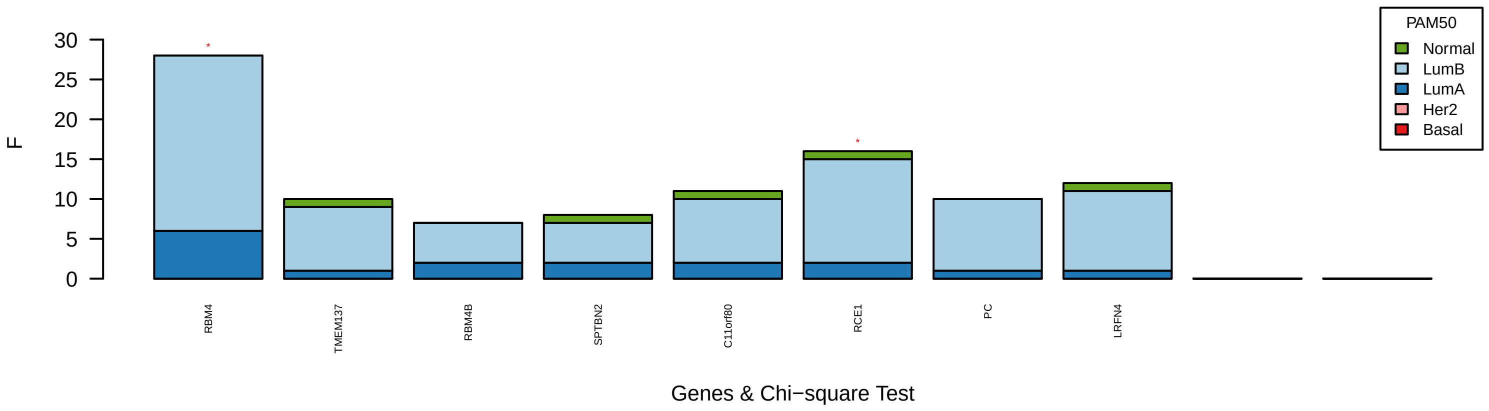


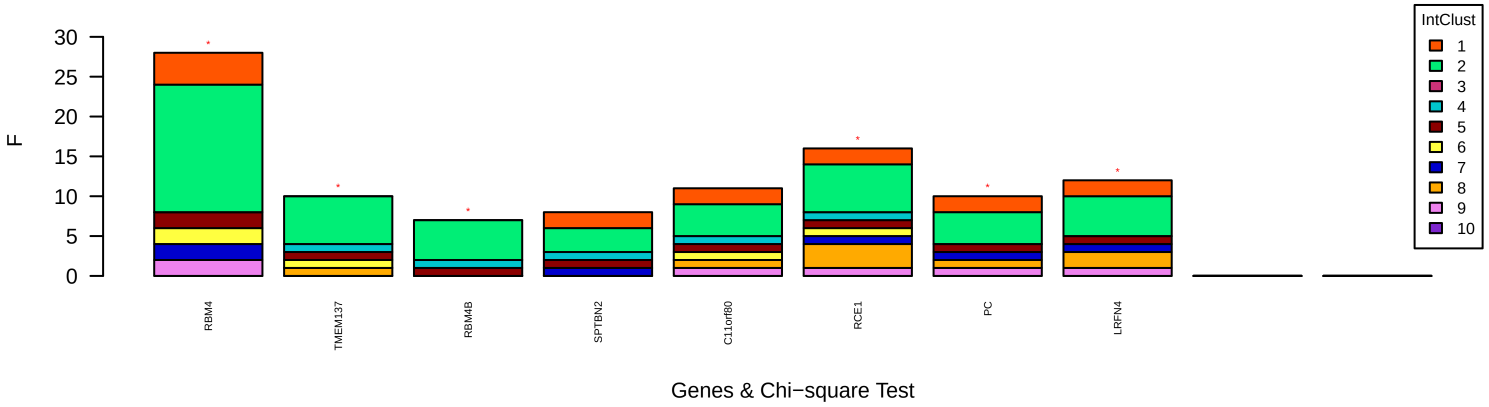
k−means region19 chr11:31347953−36251485 Subtypes Segregation − from DNAJC24 to C11orf55

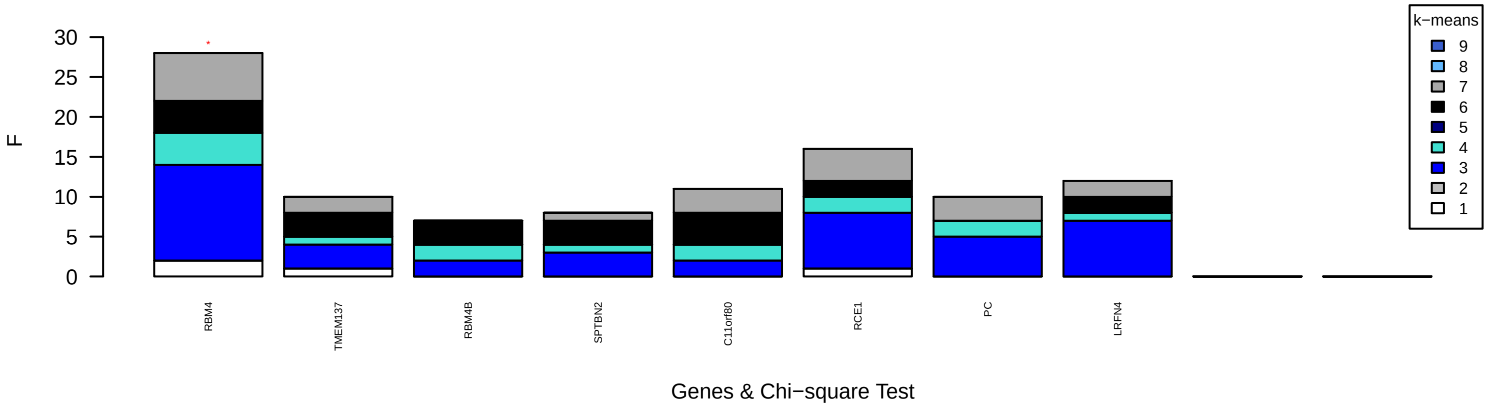


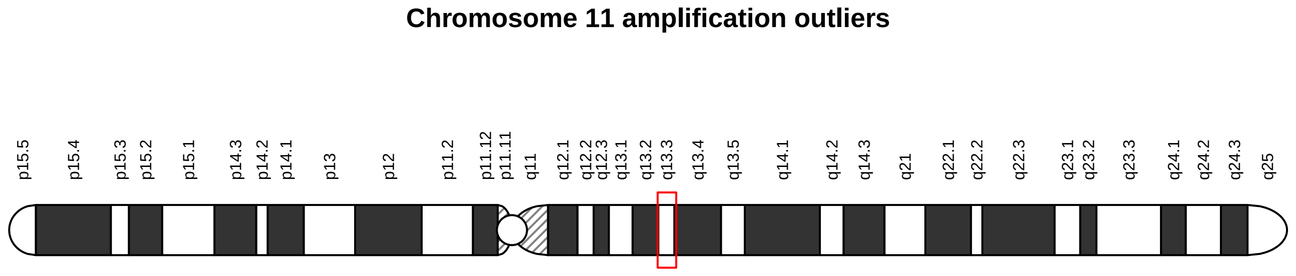
Chromosome 11 amplification outliers

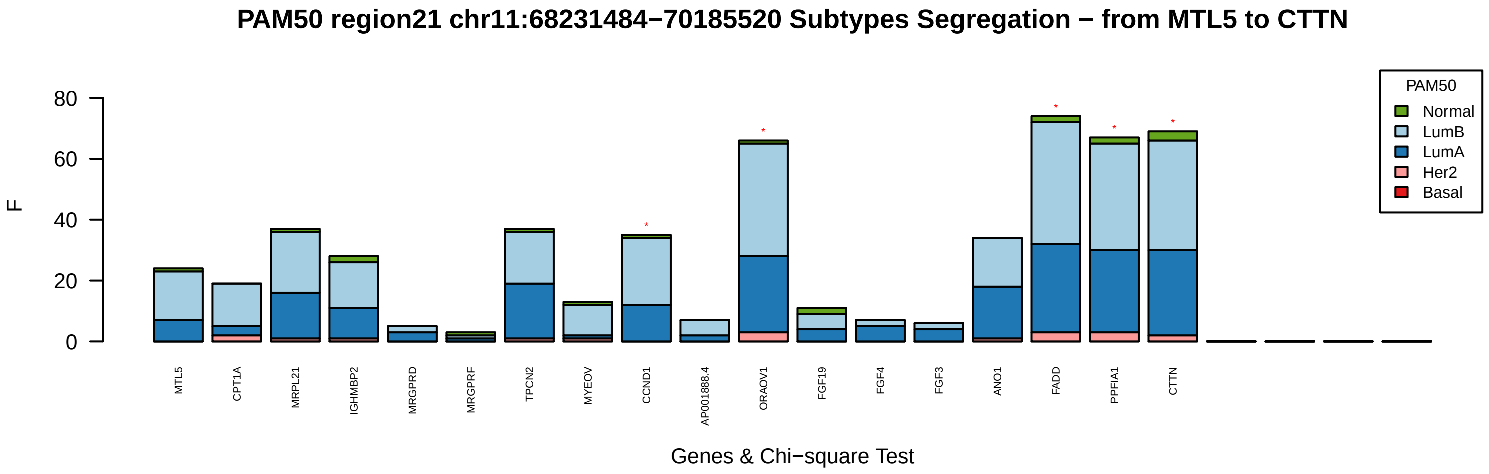


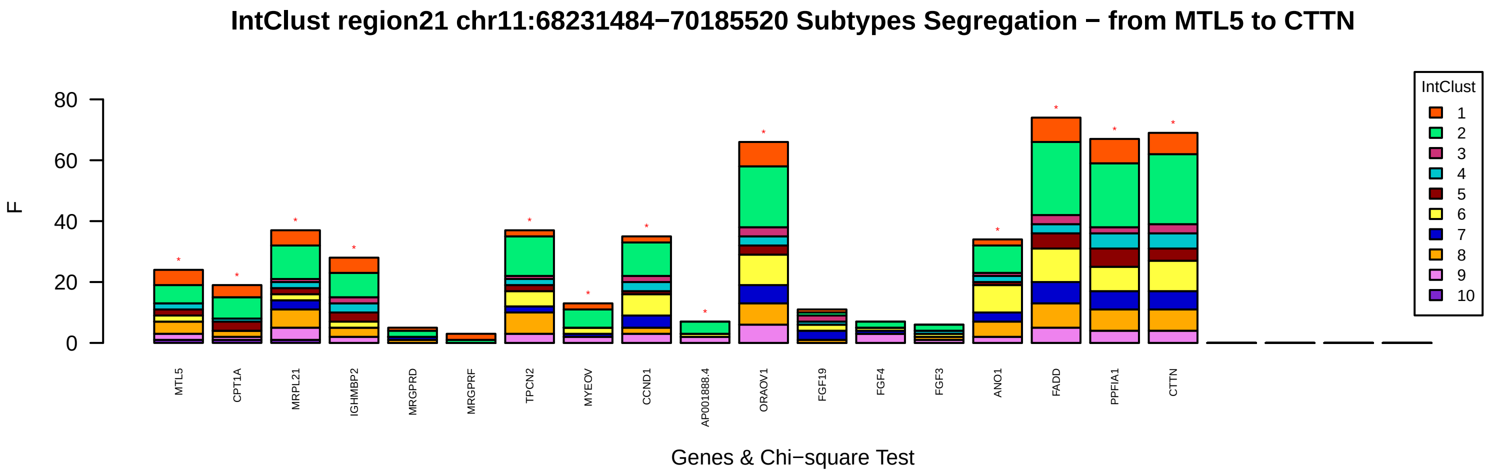

PAM50 region20 chr11:65576392−66471916 Subtypes Segregation − from RBM4 to LRFN4

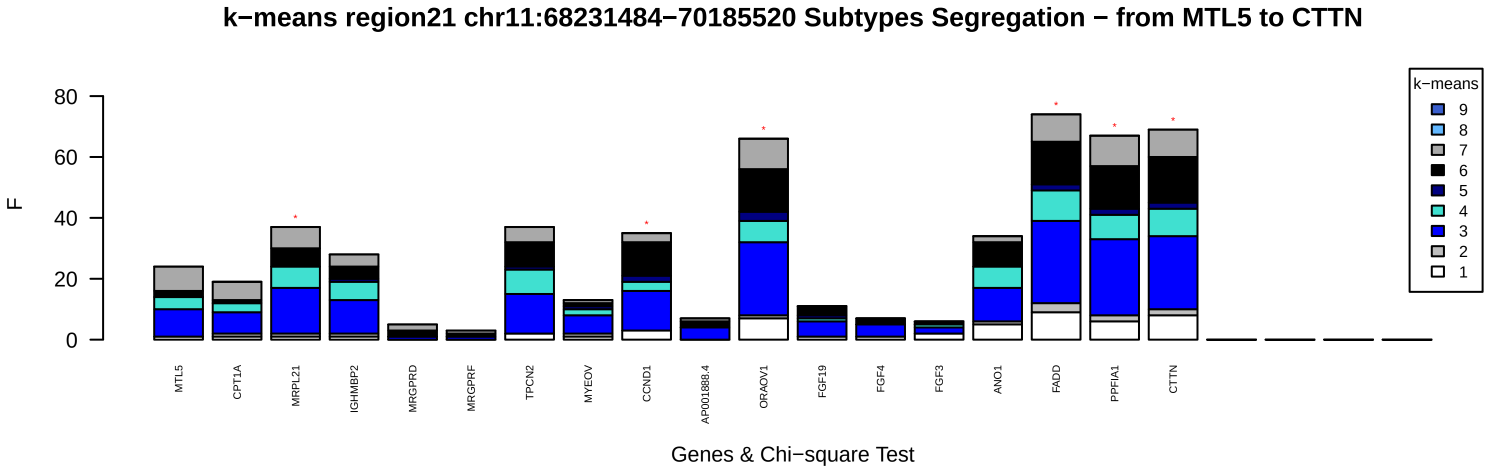

IntClust region20 chr11:65576392−66471916 Subtypes Segregation − from RBM4 to LRFN4


k−means region20 chr11:65576392−66471916 Subtypes Segregation − from RBM4 to LRFN4

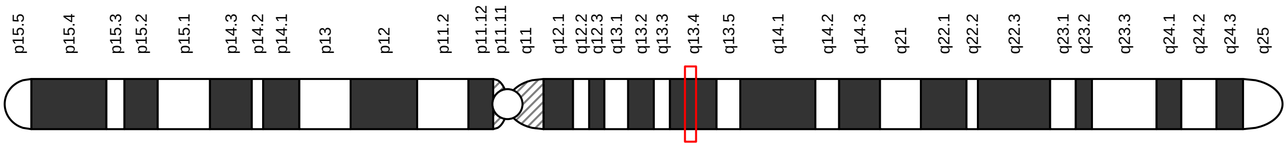

Chromosome 11 amplification outliers
p15.5
p15.4
p15.3
p15.2
p15.1
p14.3
p14.2
p14.1
p13
p12
p11.2
p11.12
p11.11
q11
q12.1
q12.2
q12.3
q13.1
q13.2
q13.3
q13.4
q13.5
q14.1
q14.2
q14.3
q21
q22.1
q22.2
q22.3
q23.1
q23.2
q23.3
q24.1
q24.2
q24.3
q25


PAM50 region21 chr11:68231484−70185520 Subtypes Segregation − from MTL5 to CTTN
F
0
20
40
60
80
PAM50
Normal
LumB
LumA
Her2
Basal
MTL5
CPT1A
MRPL21
IGHMBP2
MRGPRD
MRGPRF
TPCN2
MYEOV
CCND1
AP001888.4
ORAOV1
FGF19
FGF4
FGF3
ANO1
FADD
PPFIA1
CTTN
Genes & Chi−square Test

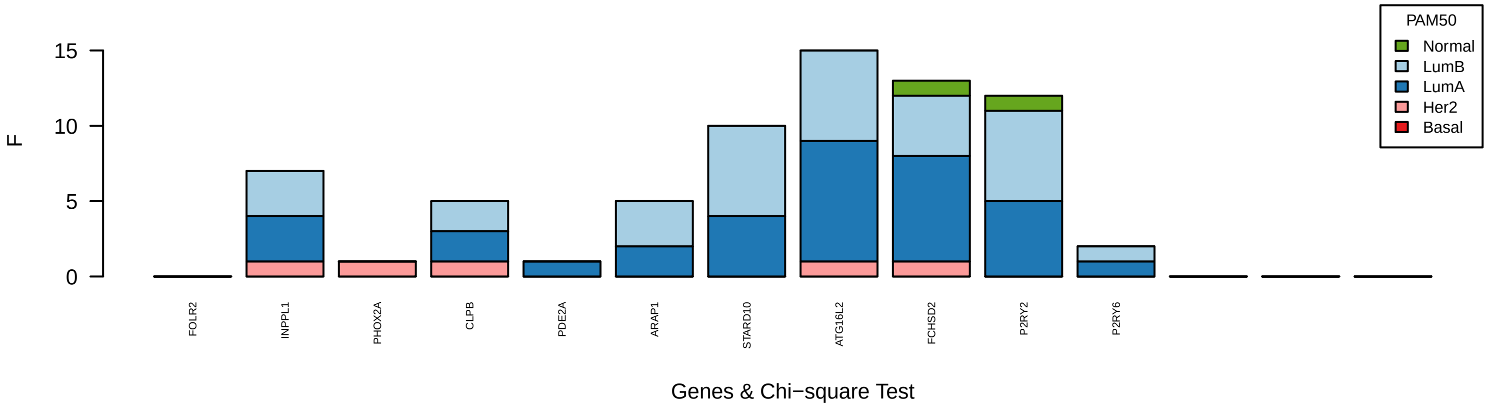

IntClust region21 chr11:68231484−70185520 Subtypes Segregation − from MTL5 to CTTN
F
0
20
40
60
80
IntClust
1
2
3
4
5
6
7
8
9
10
MTL5
CPT1A
MRPL21
IGHMBP2
MRGPRD
MRGPRF
TPCN2
MYEOV
CCND1
AP001888.4
ORAOV1
FGF19
FGF4
FGF3
ANO1
FADD
PPFIA1
CTTN
Genes & Chi−square Test


k−means region21 chr11:68231484−70185520 Subtypes Segregation − from MTL5 to CTTN
F
0
20
40
60
80
k−means
9
8
7
6
5
4
3
2
1
MTL5
CPT1A
MRPL21
IGHMBP2
MRGPRD
MRGPRF
TPCN2
MYEOV
CCND1
AP001888.4
ORAOV1
FGF19
FGF4
FGF3
ANO1
FADD
PPFIA1
CTTN
Genes & Chi−square Test

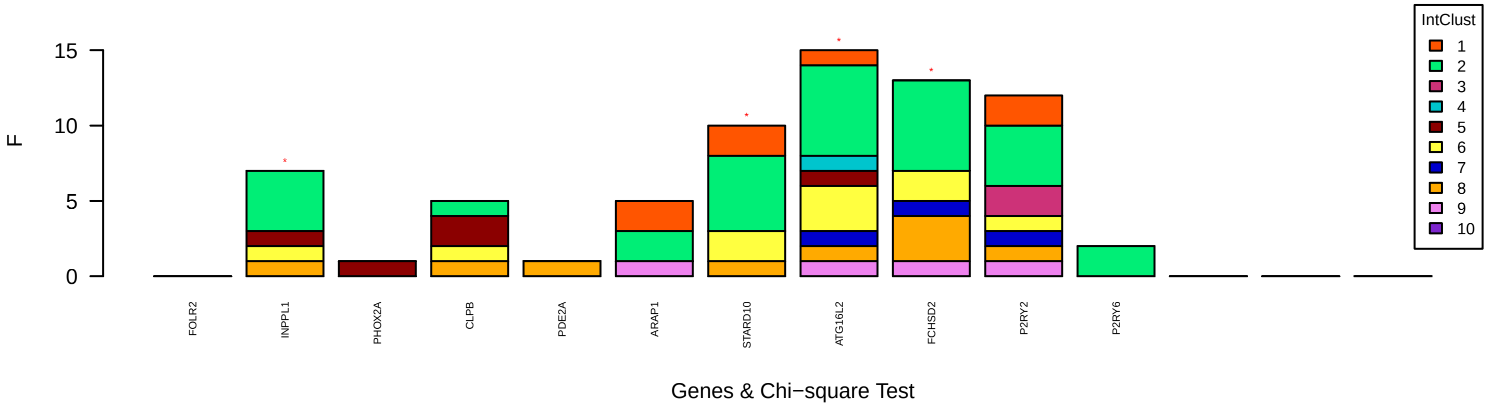
Chromosome 11 amplification outliers



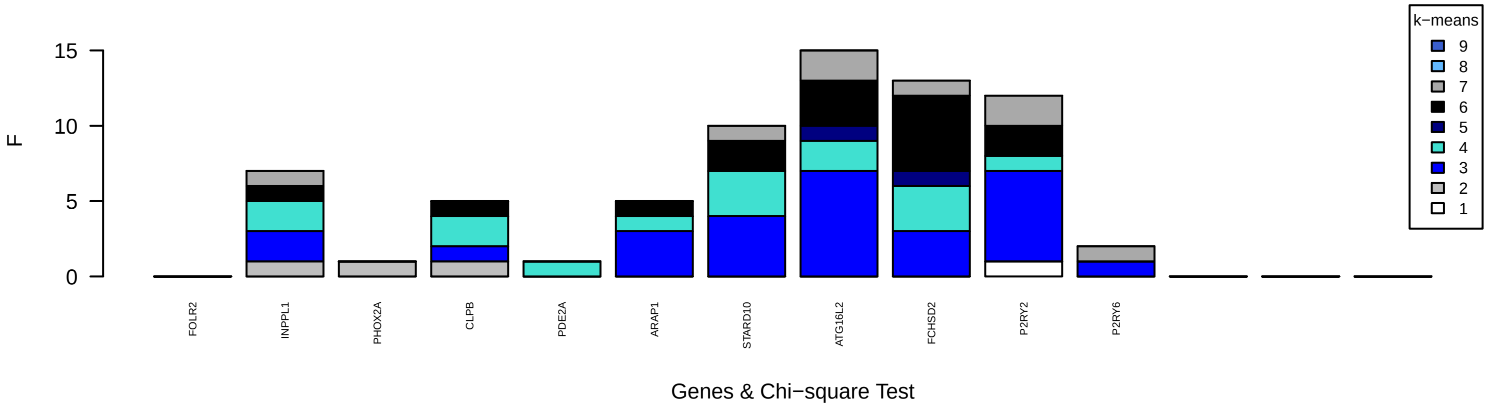
PAM50 region22 chr11:71578255−72747564 Subtypes Segregation − from FOLR2 to P2RY6



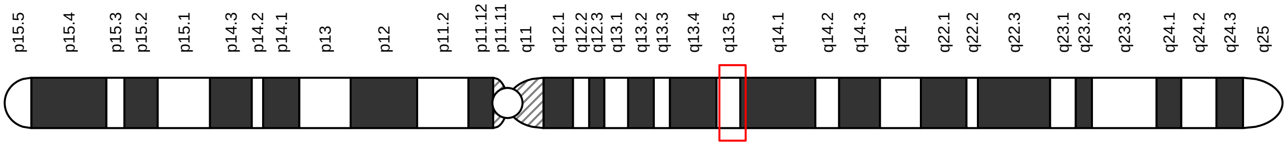
IntClust region22 chr11:71578255−72747564 Subtypes Segregation − from FOLR2 to P2RY6



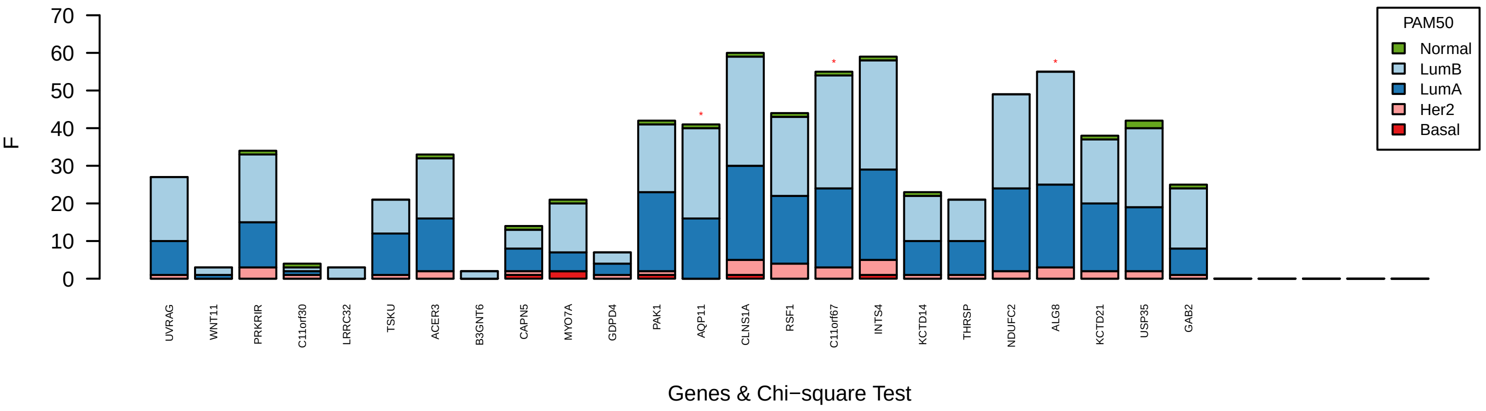
k−means region22 chr11:71578255−72747564 Subtypes Segregation − from FOLR2 to P2RY6



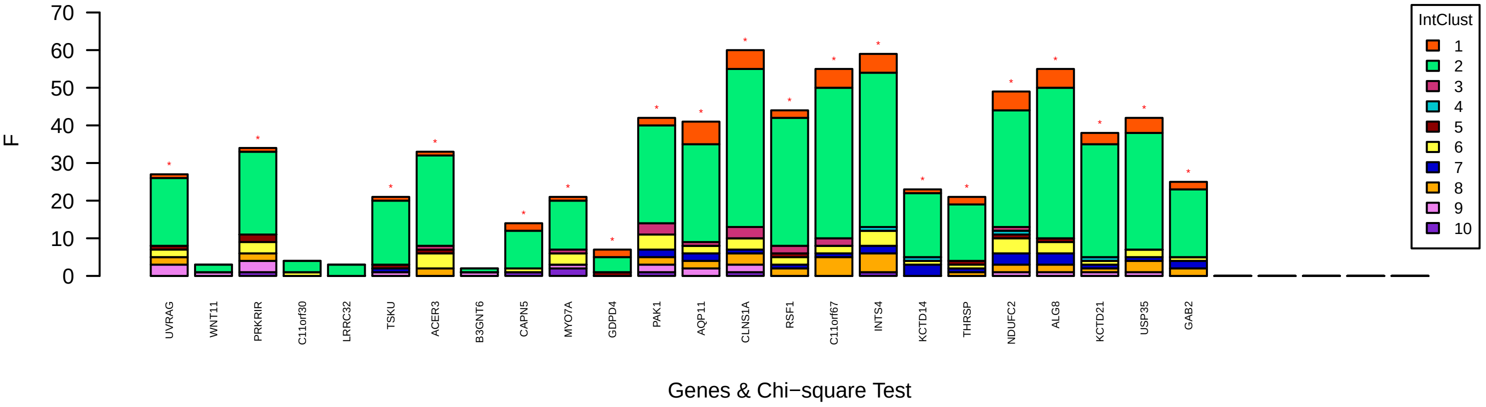
Chromosome 11 amplification outliers



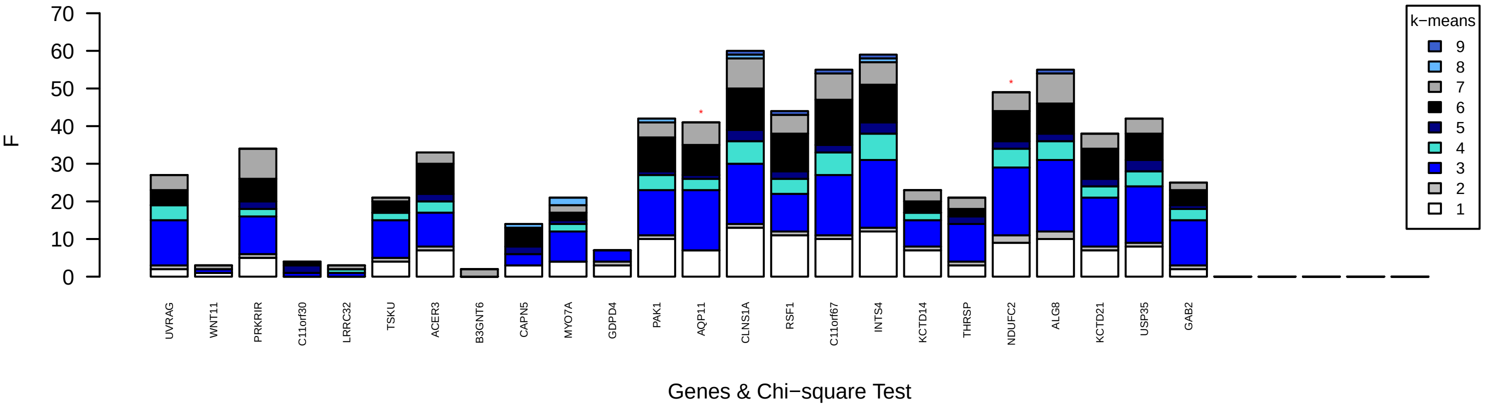
PAM50 region23 chr:75203860−77963367 Subtypes Segregation − from UVRAG to GAB2

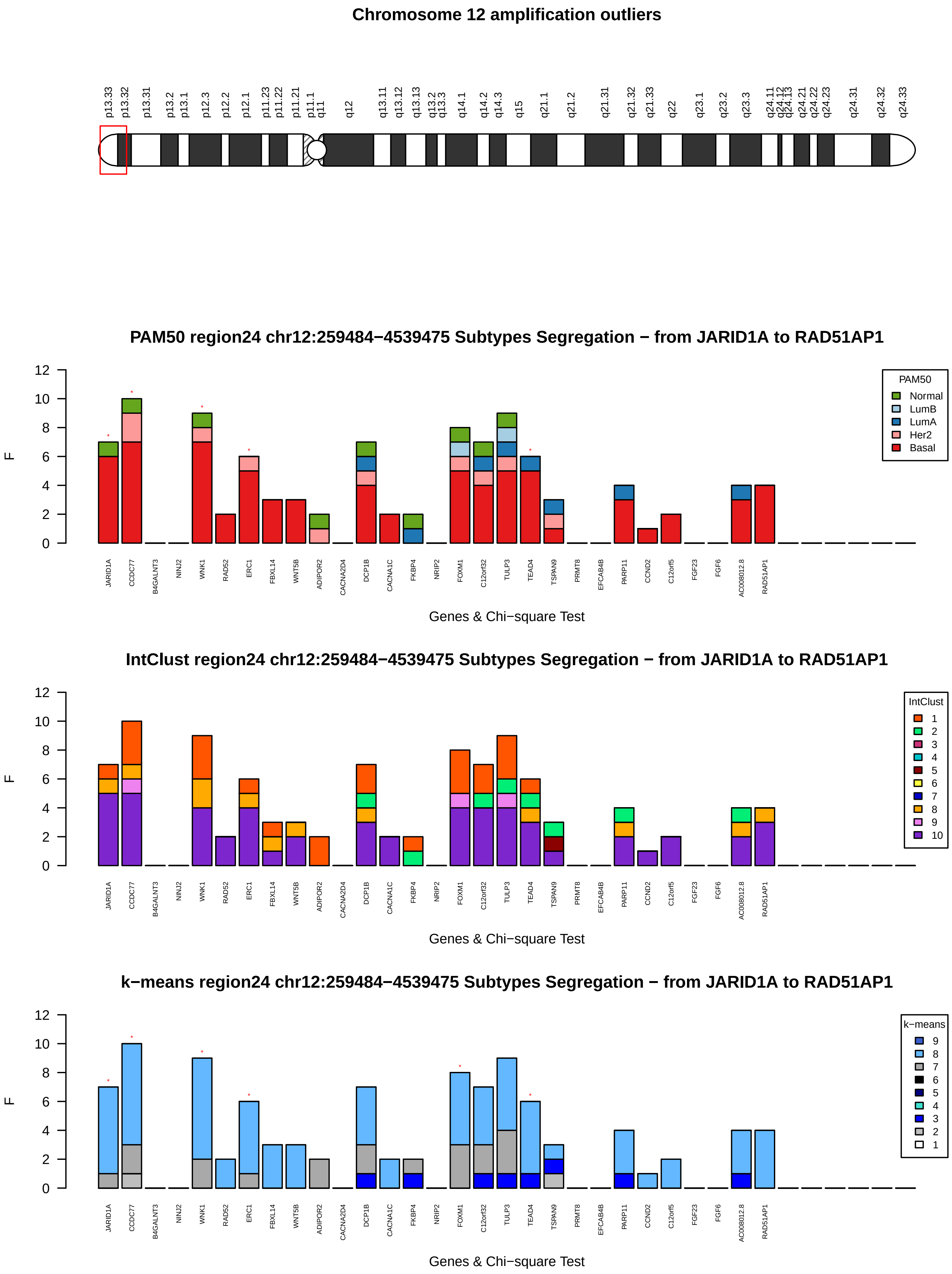


IntClust region23 chr:75203860−77963367 Subtypes Segregation − from UVRAG to GAB2



k−means region23 chr:75203860−77963367 Subtypes Segregation − from UVRAG to GAB2





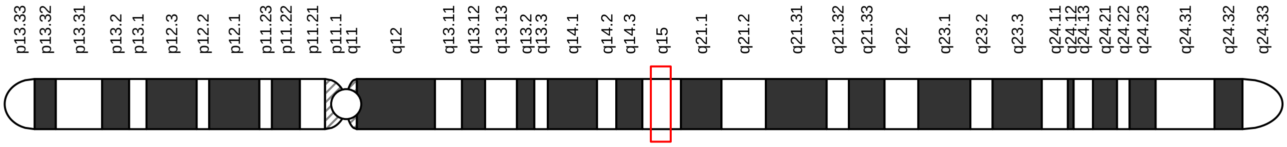
Chromosome 12 amplification outliers

PAM50 region24 chr12:259484−4539475 Subtypes Segregation − from JARID1A to RAD51AP1

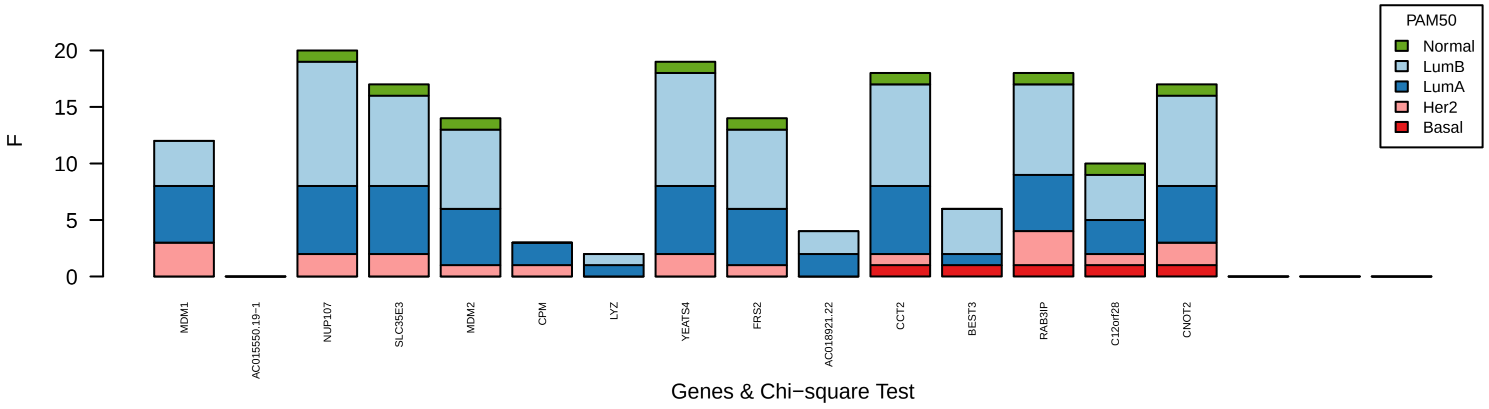
IntClust region24 chr12:259484−4539475 Subtypes Segregation − from JARID1A to RAD51AP1

k−means region24 chr12:259484−4539475 Subtypes Segregation − from JARID1A to RAD51AP1

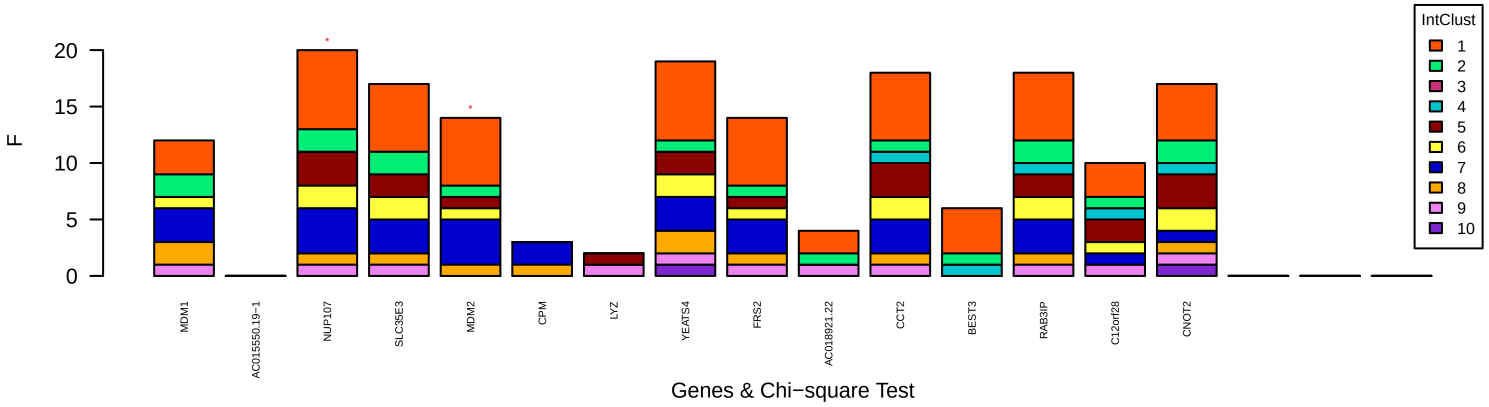
Chromosome 12 amplification outliers



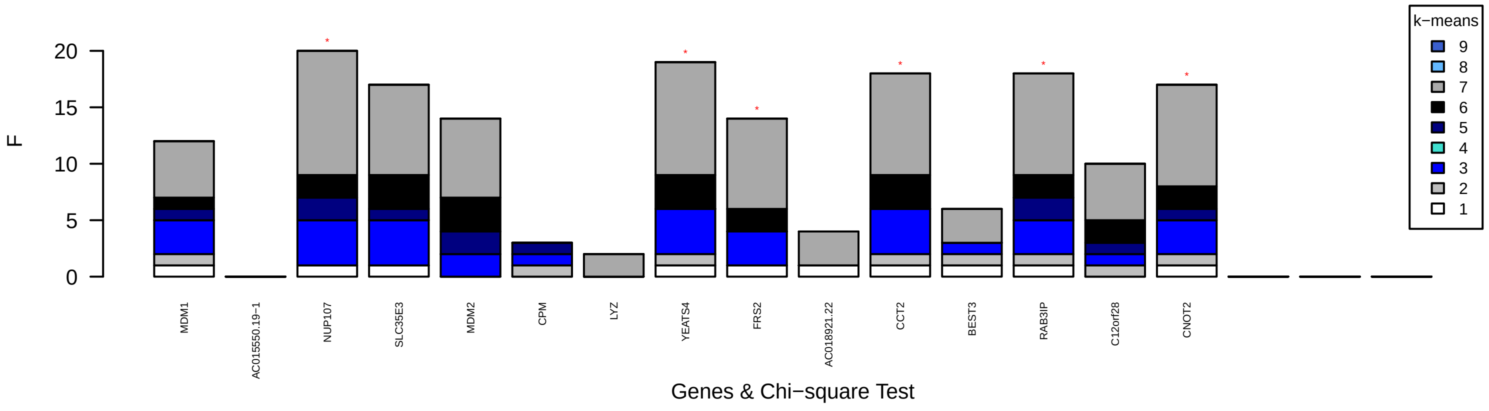
PAM50 region25 chr12:66974613−69035040 Subtypes Segregation − from MDM1 to CNOT2



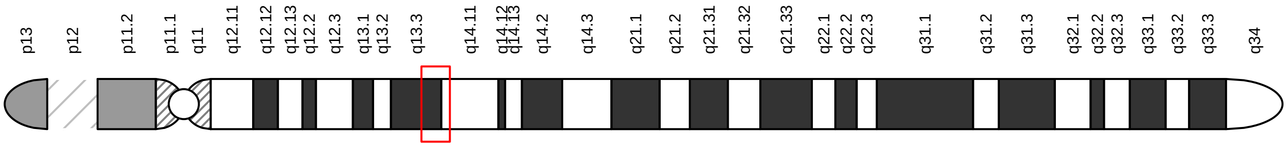
IntClust region25 chr12:66974613−69035040 Subtypes Segregation − from MDM1 to CNOT2



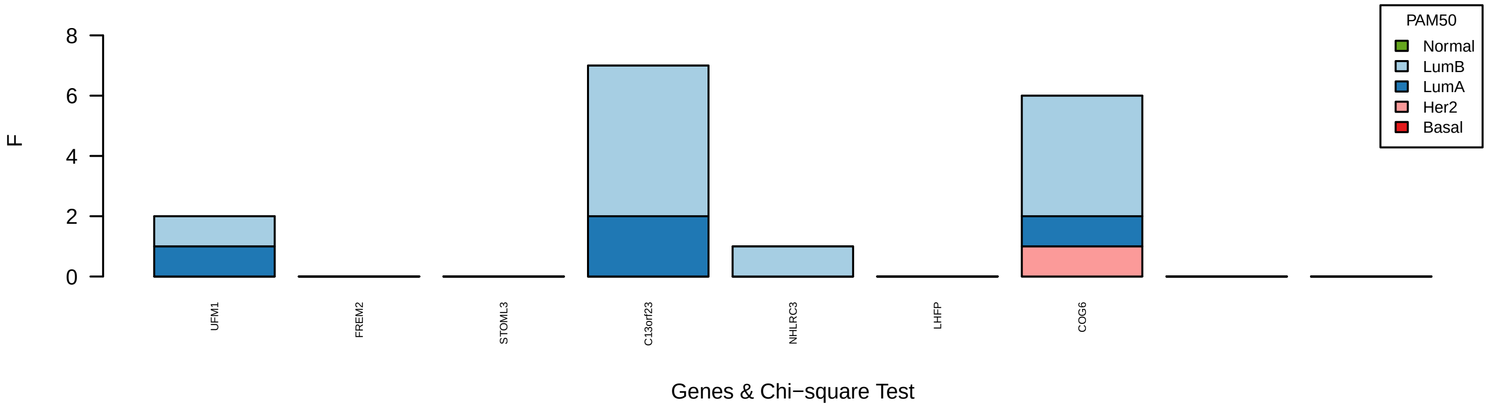
k−means region25 chr12:66974613−69035040 Subtypes Segregation − from MDM1 to CNOT2



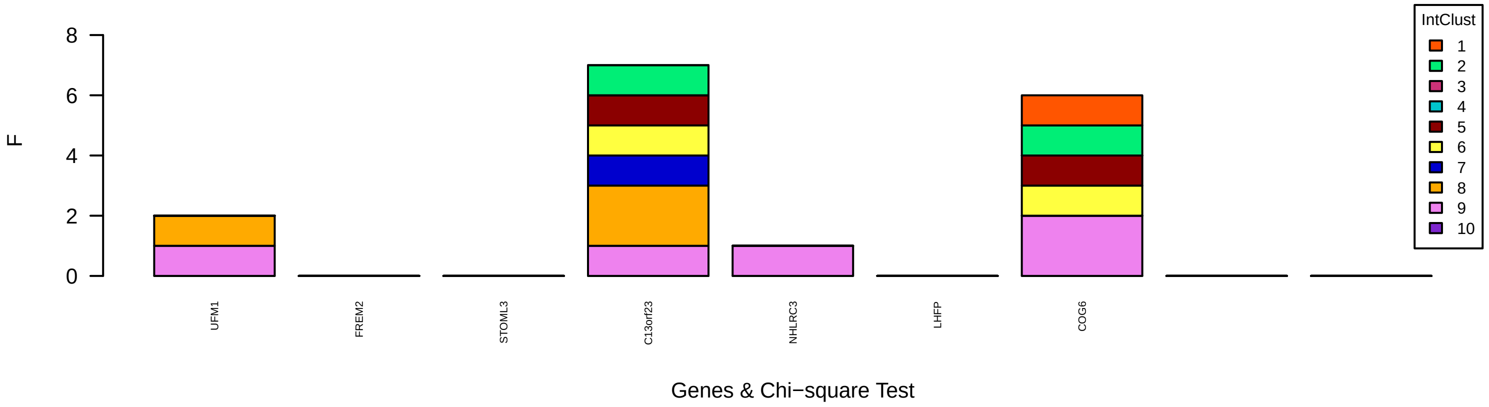
Chromosome 13 deletion outliers



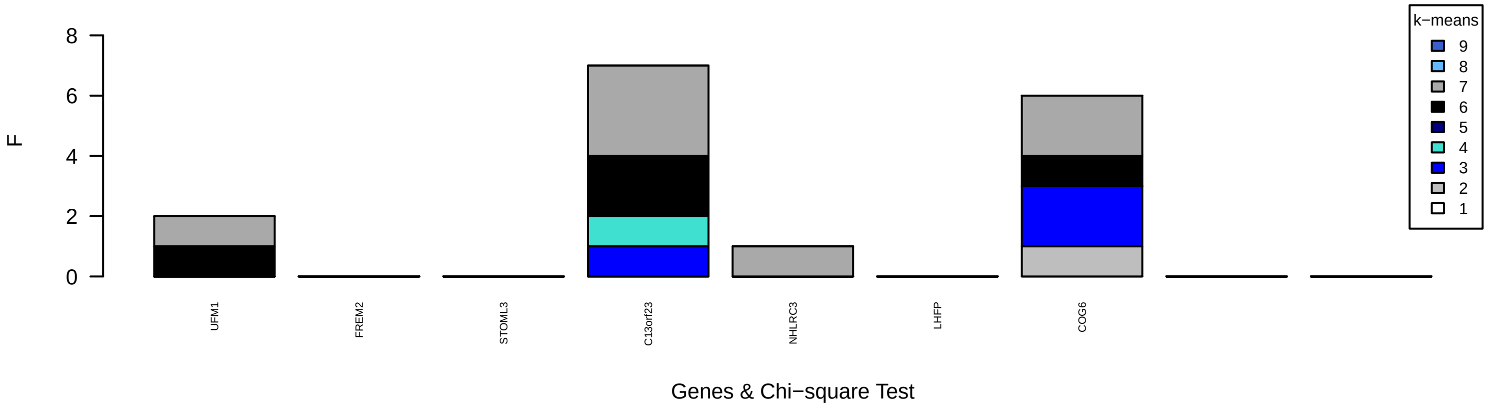
PAM50 region26 chr13:37224832−39767421 Subtypes Segregation − from UFM1 to COG6



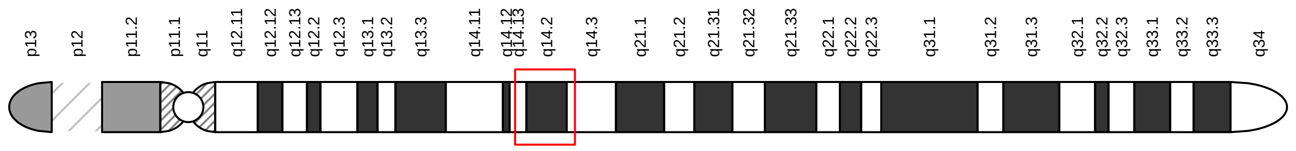
IntClust region26 chr13:37224832−39767421 Subtypes Segregation − from UFM1 to COG6



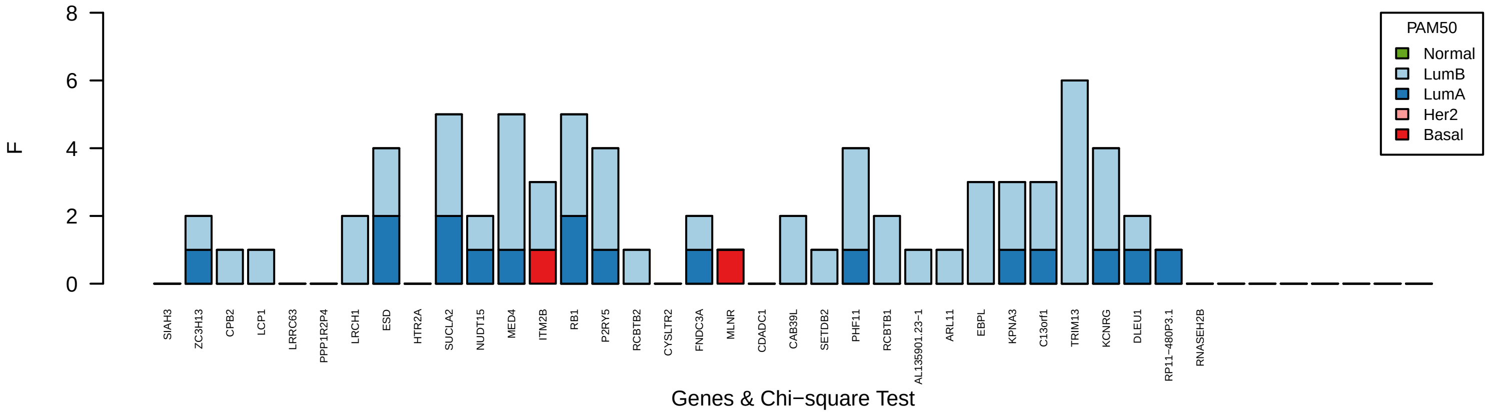
k−means region26 chr13:37224832−39767421 Subtypes Segregation − from UFM1 to COG6



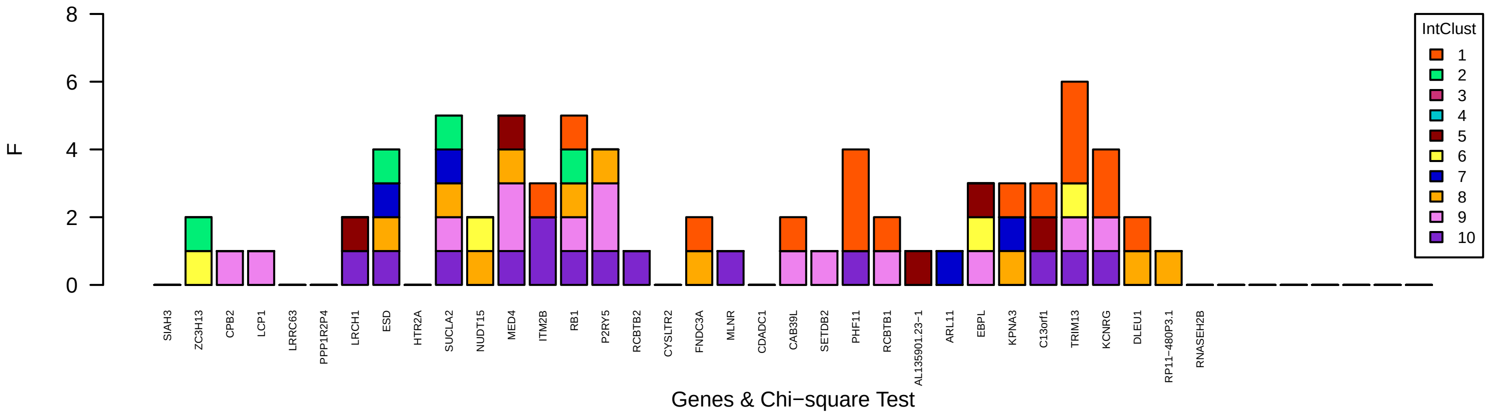
Chromosome 13 deletion outliers



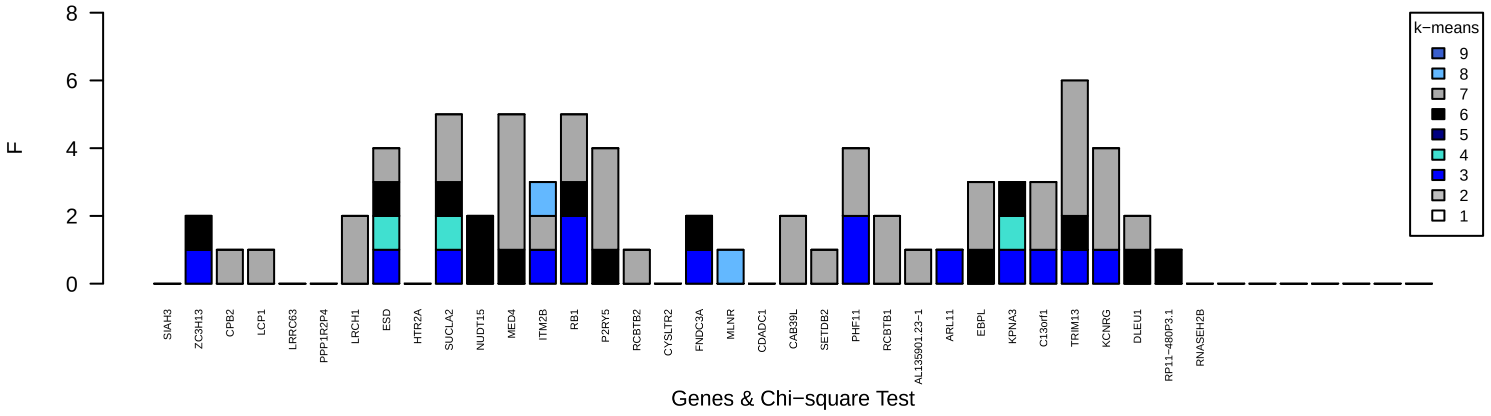
PAM50 region27 chr13:45192796−50537115 Subtypes Segregation − from SIAH3 to RNASEH2B



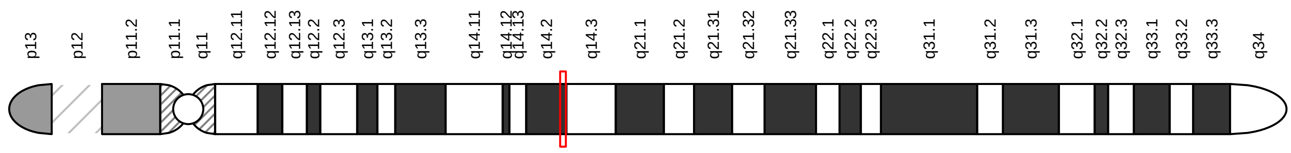
IntClust region27 chr13:45192796−50537115 Subtypes Segregation − from SIAH3 to RNASEH2B



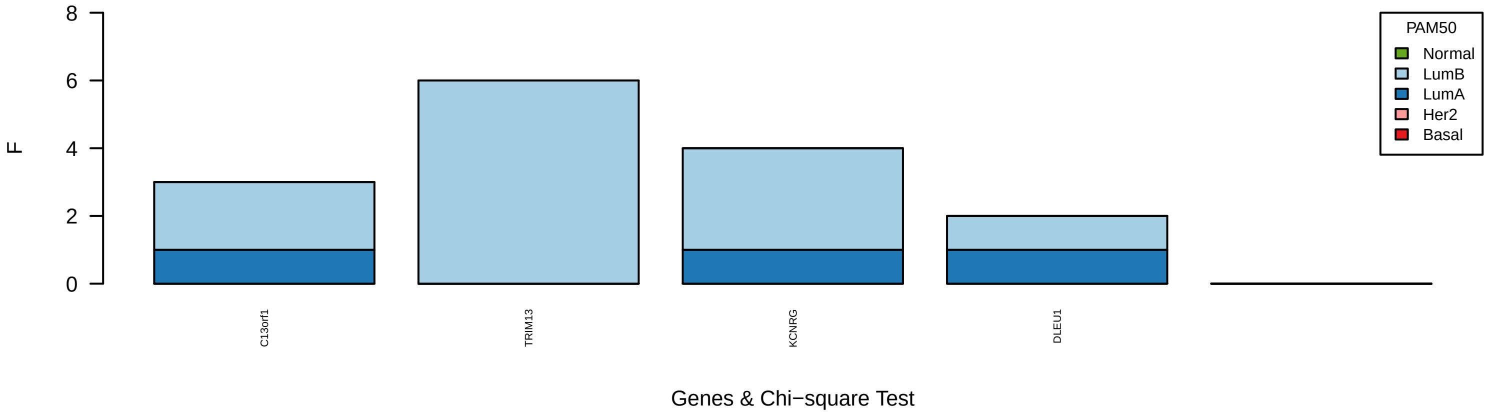
k−means region27 chr13:45192796−50537115 Subtypes Segregation − from SIAH3 to RNASEH2B



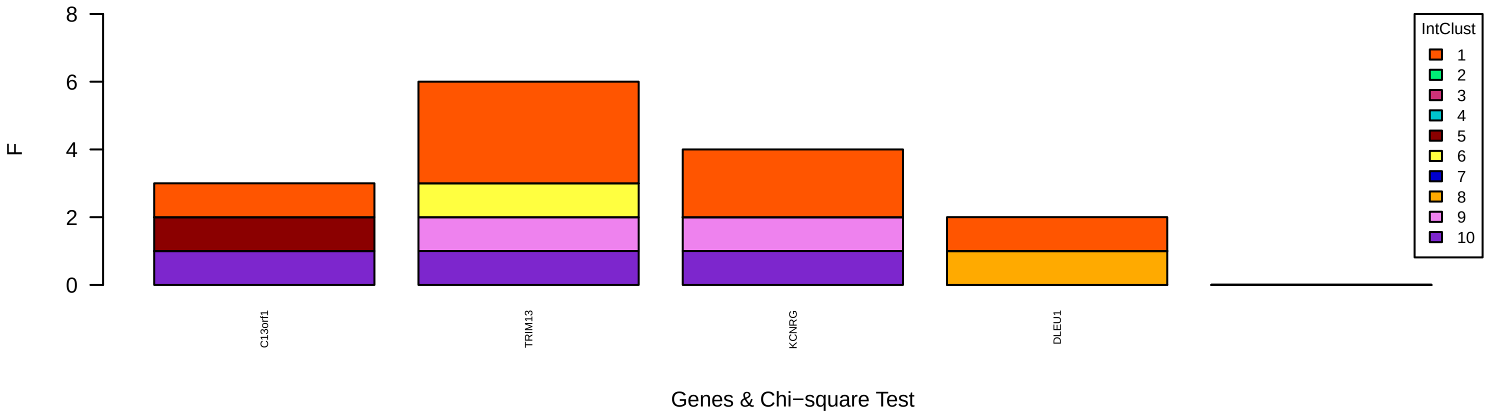
Chromosome 13 deletion outliers



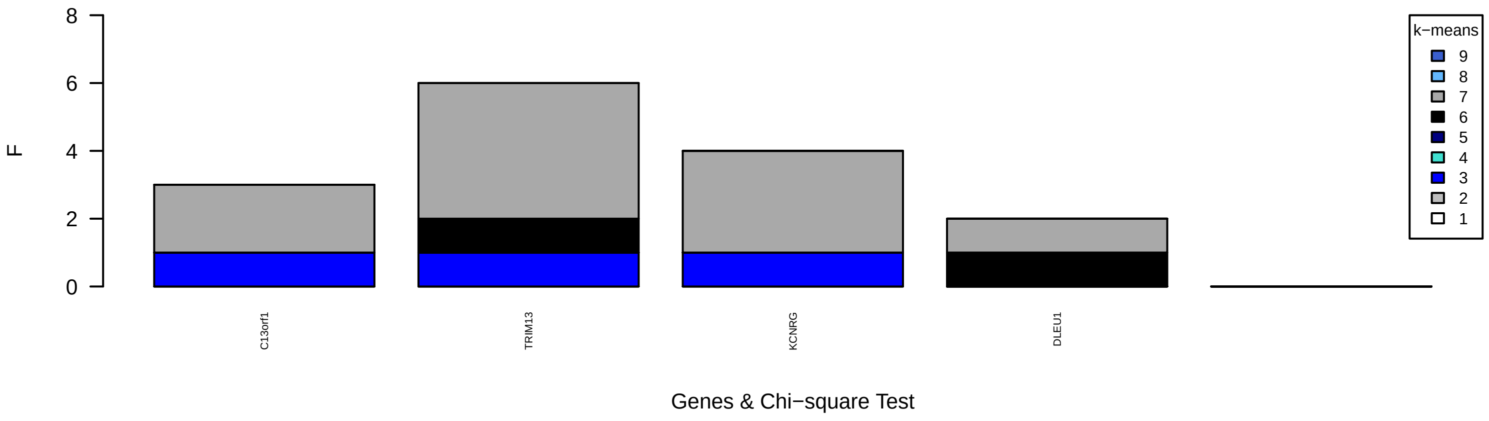
PAM50 region28 chr13:49235120−49745329 Subtypes Segregation − from C13orf1 to DLEU1



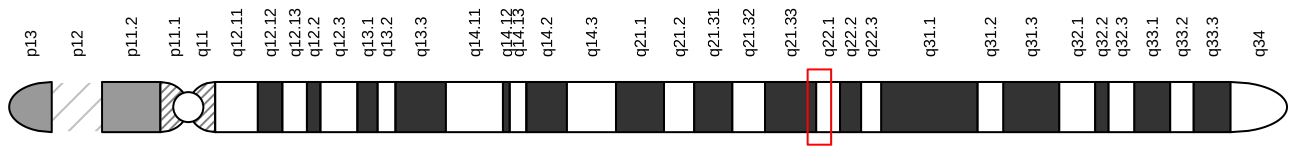
IntClust region28 chr13:49235120−49745329 Subtypes Segregation − from C13orf1 to DLEU1



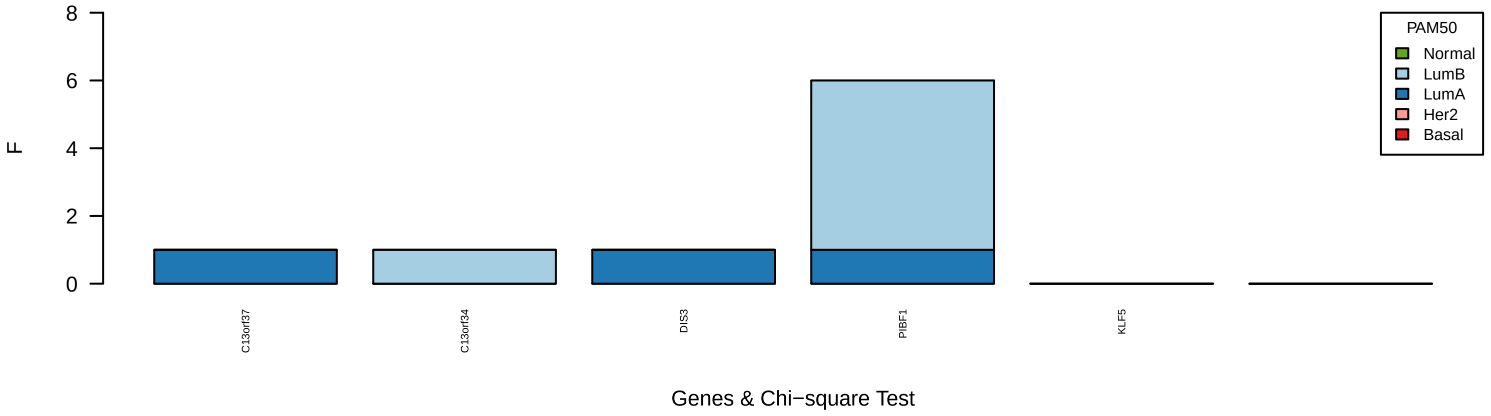
k−means region28 chr13:49235120−49745329 Subtypes Segregation − from C13orf1 to DLEU1



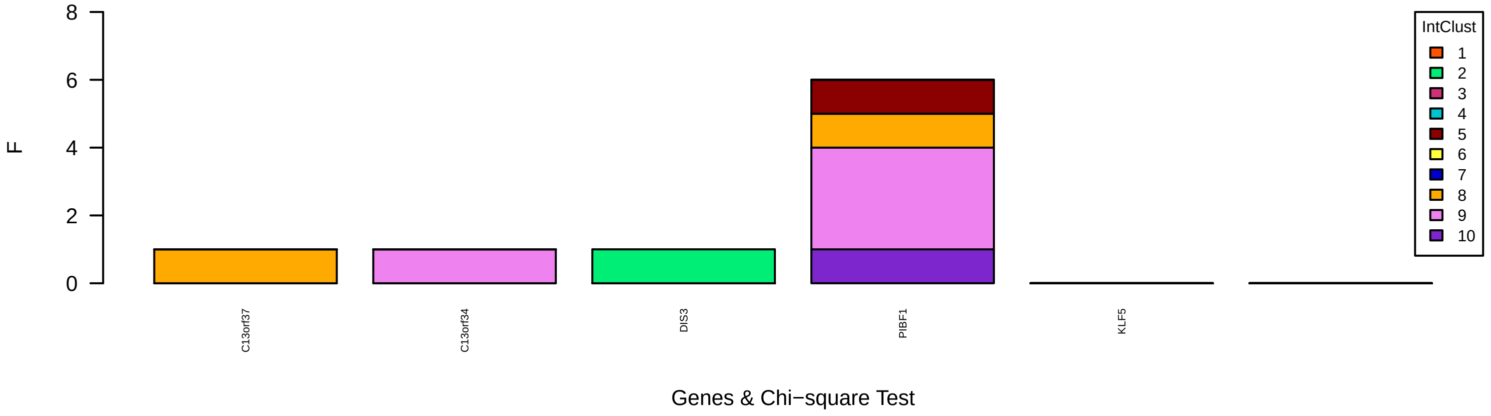
Chromosome 13 deletion outliers



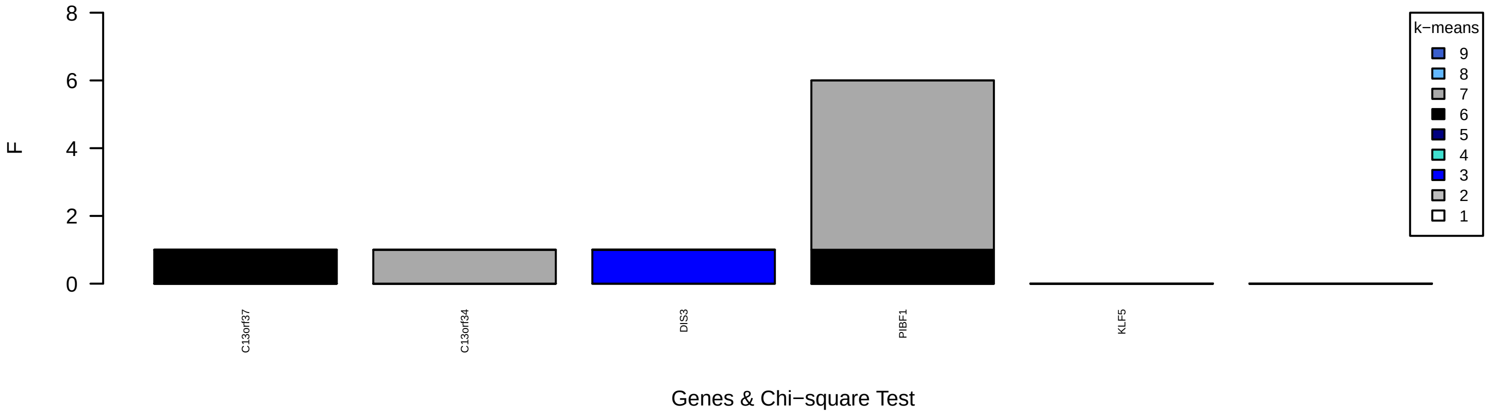
PAM50 region29 chr13:71316783−73426040 Subtypes Segregation − from C13orf37 to KLF5



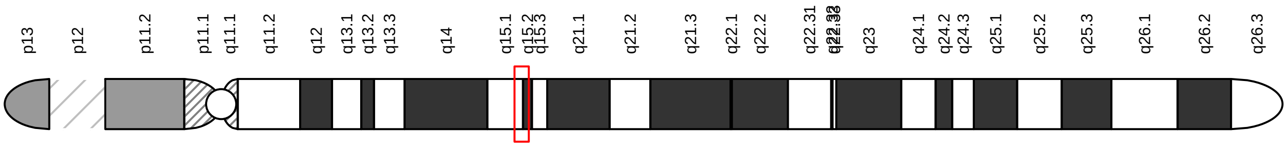
IntClust region29 chr13:71316783−73426040 Subtypes Segregation − from C13orf37 to KLF5



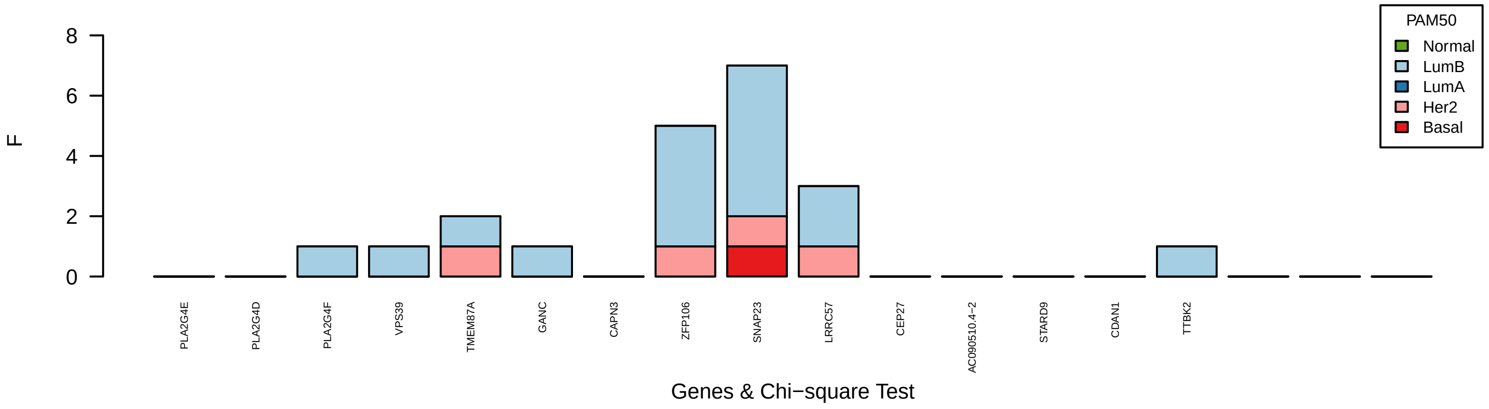
k−means region29 chr13:71316783−73426040 Subtypes Segregation − from C13orf37 to KLF5



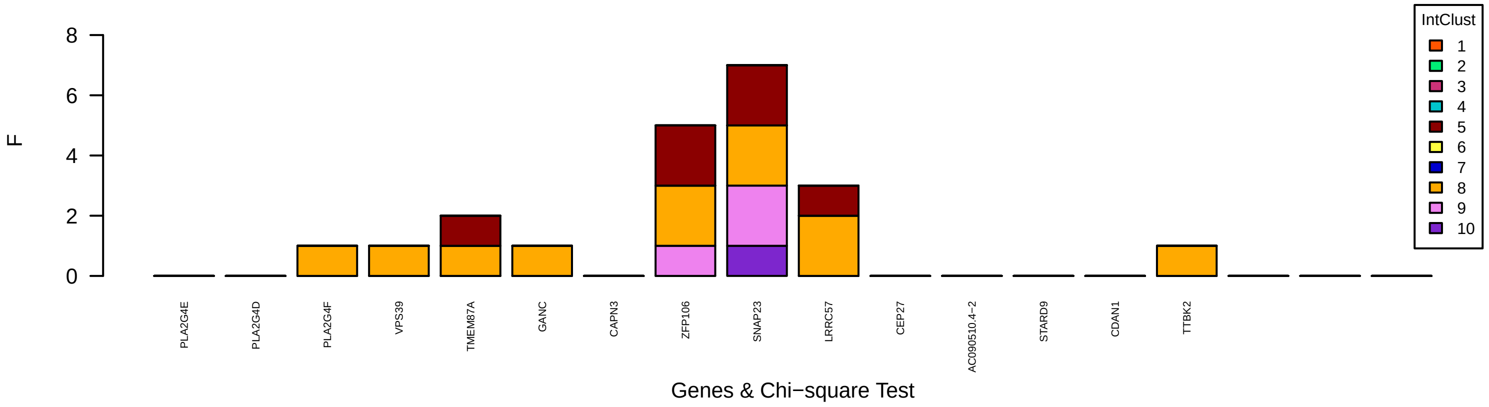
Chromosome 15 deletion outliers



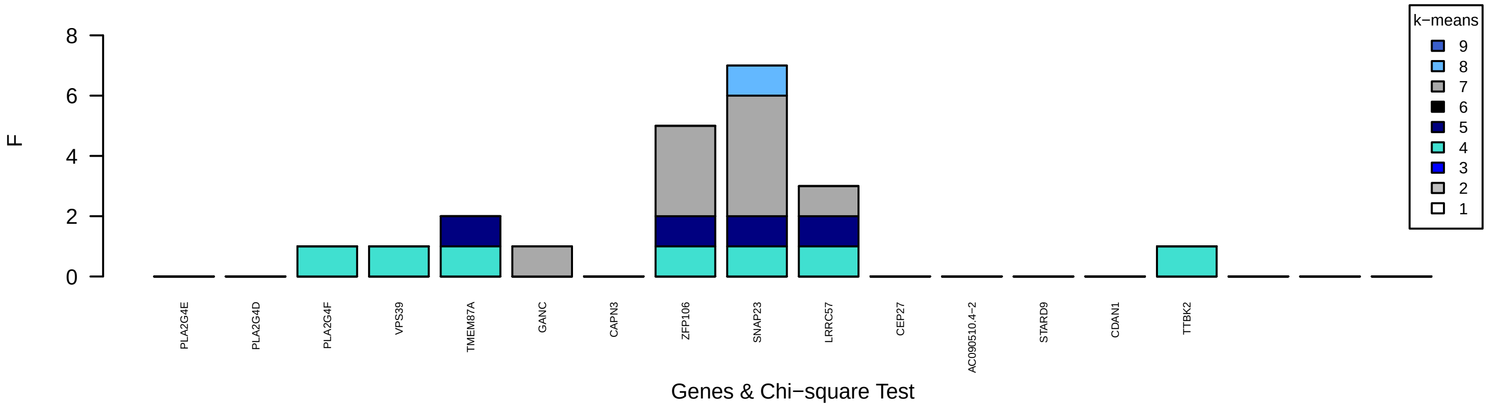
PAM50 region30 chr15:40032508−41155167 Subtypes Segregation − from PLA2G4E to TTBK2

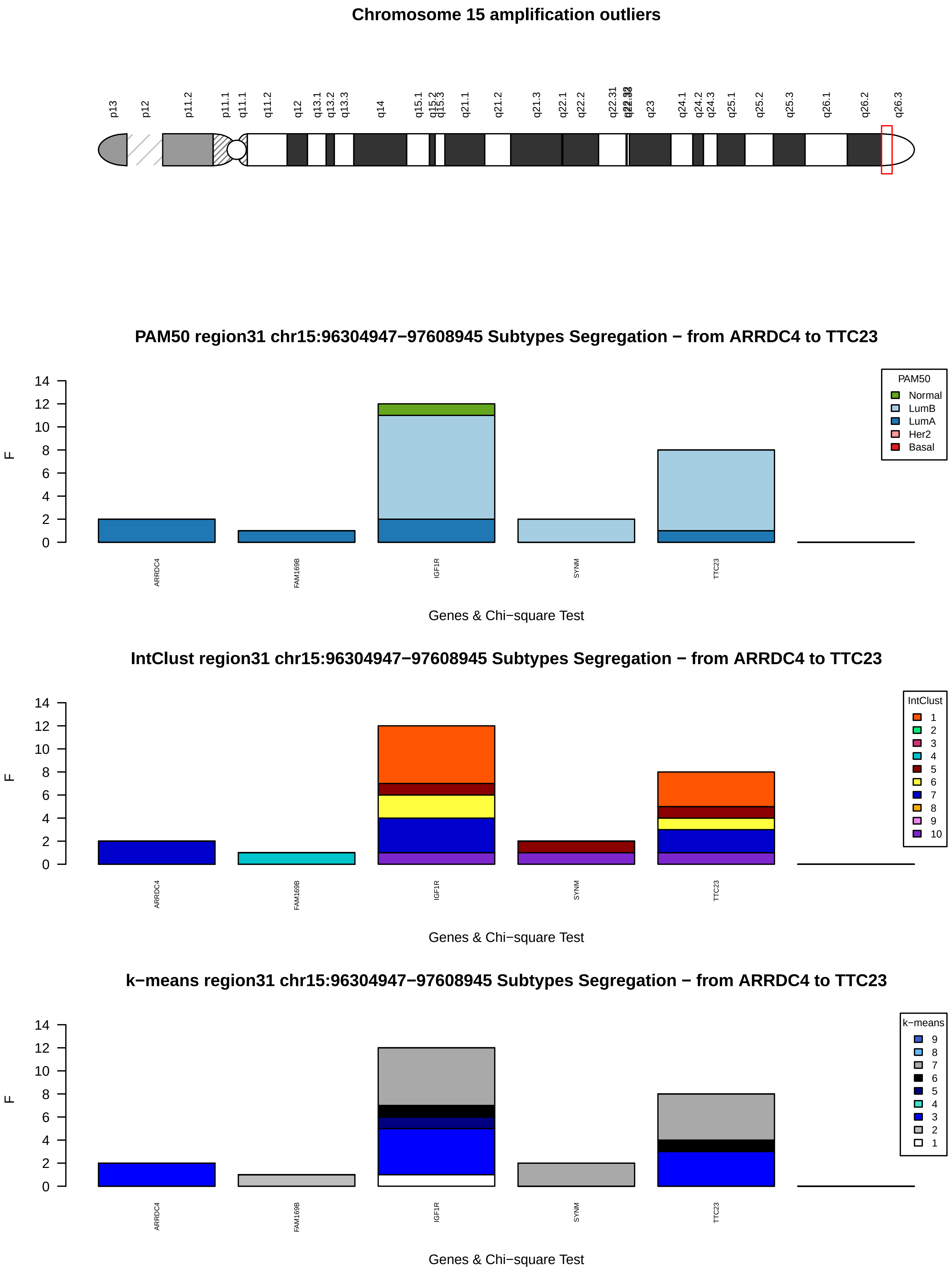


IntClust region30 chr15:40032508−41155167 Subtypes Segregation − from PLA2G4E to TTBK2



k−means region30 chr15:40032508−41155167 Subtypes Segregation − from PLA2G4E to TTBK2





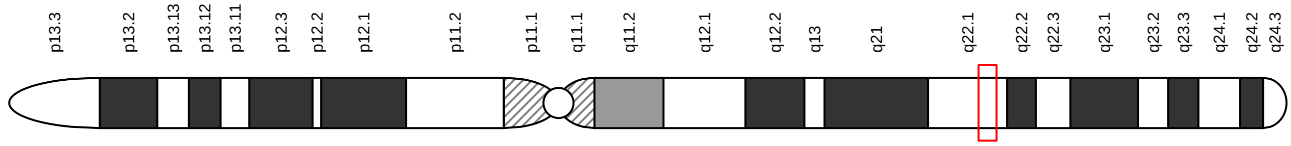
Chromosome 15 amplification outliers

PAM50 region31 chr15:96304947−97608945 Subtypes Segregation − from ARRDC4 to TTC23

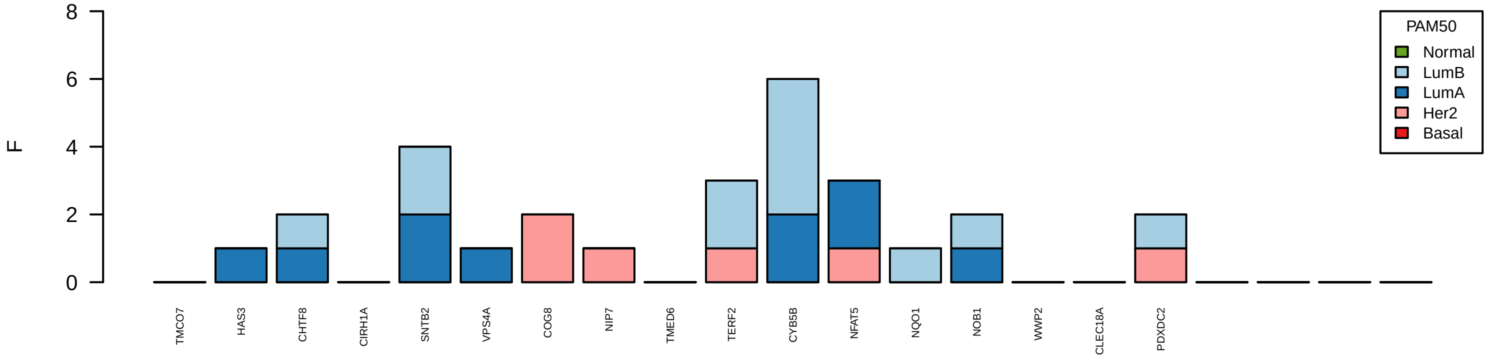
IntClust region31 chr15:96304947−97608945 Subtypes Segregation − from ARRDC4 to TTC23

k−means region31 chr15:96304947−97608945 Subtypes Segregation − from ARRDC4 to TTC23

Chromosome 16 deletion outliers

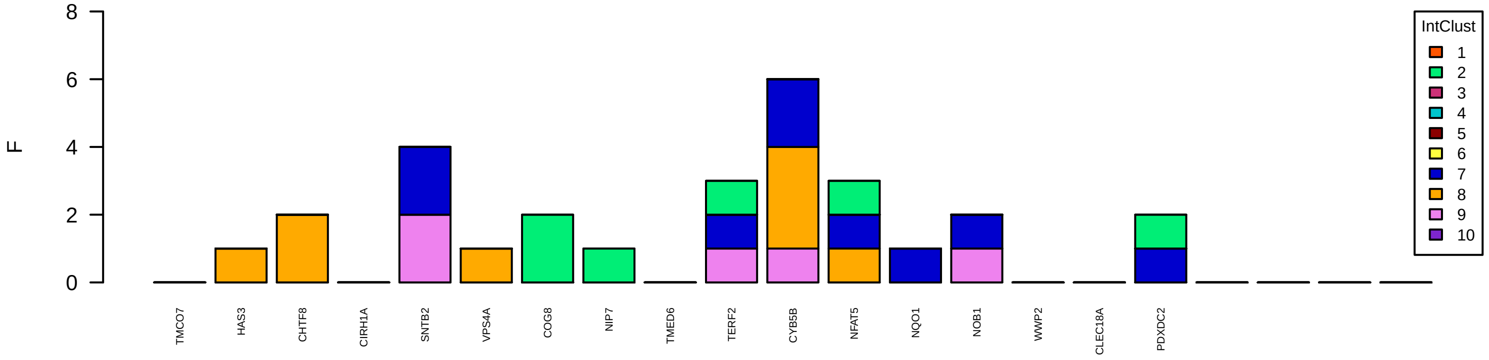


PAM50 region32 chr16:67411784−68661883 Subtypes Segregation − from TMCO7 to PDXDC2



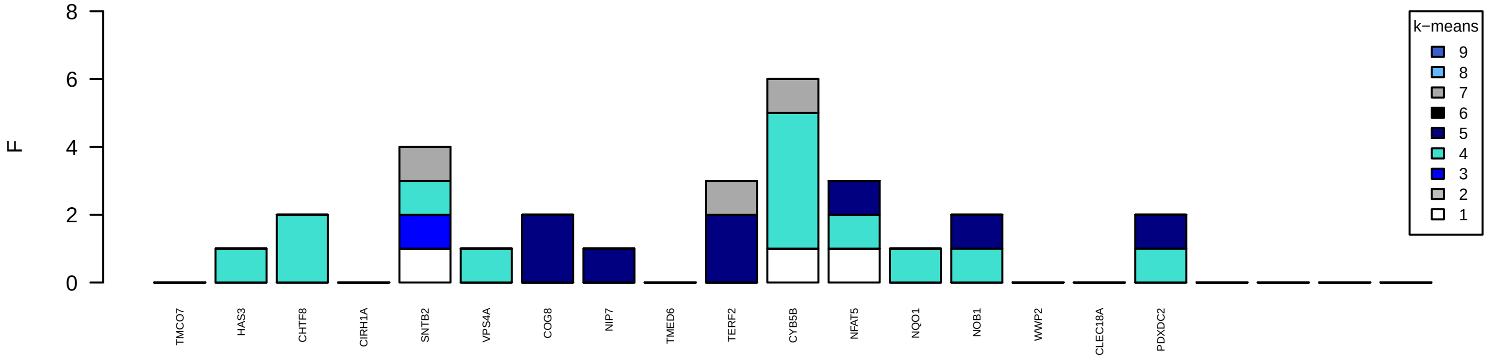
Genes & Chi−square Test

IntClust region32 chr16:67411784−68661883 Subtypes Segregation − from TMCO7 to PDXDC2

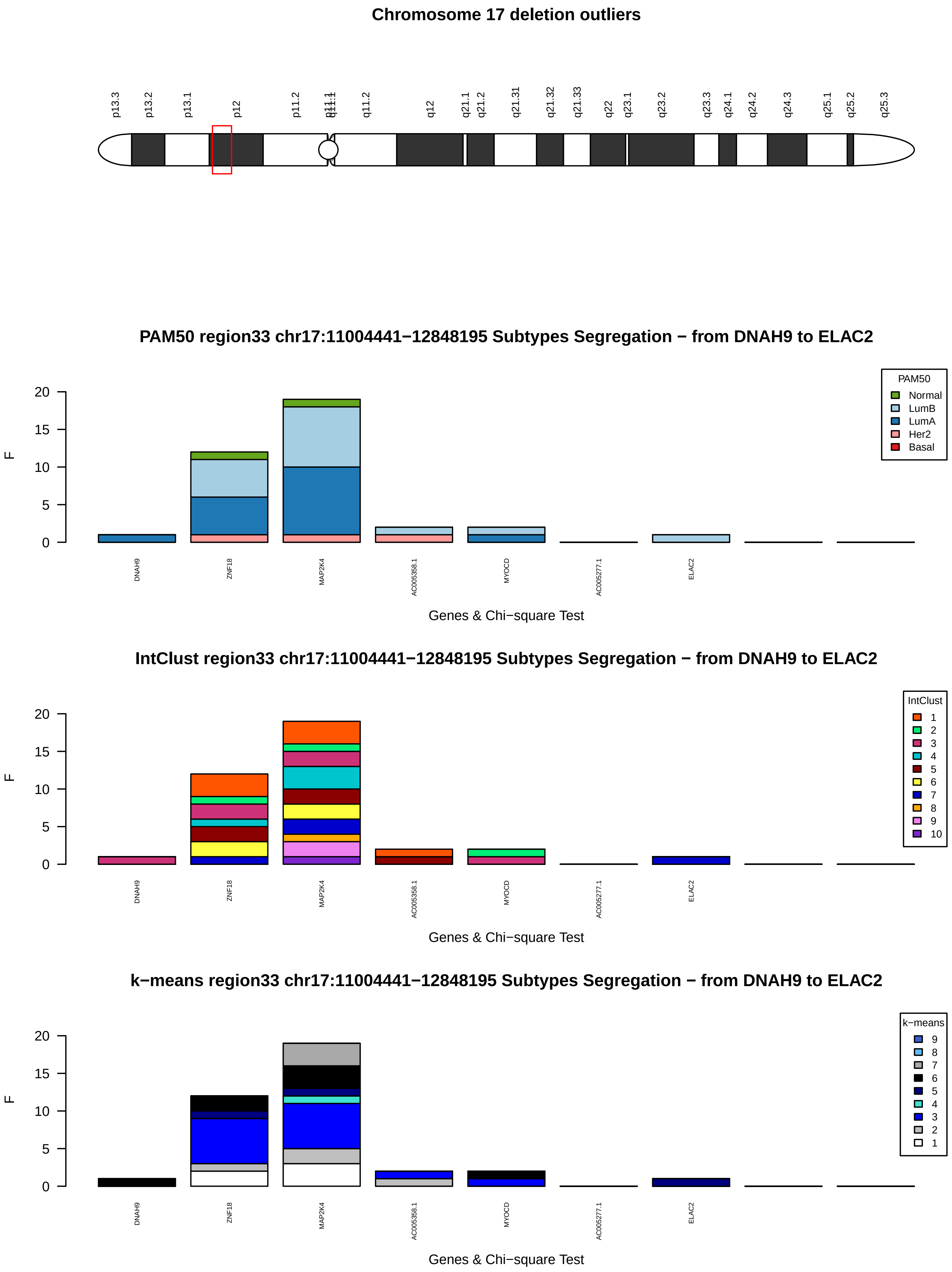


Genes & Chi−square Test

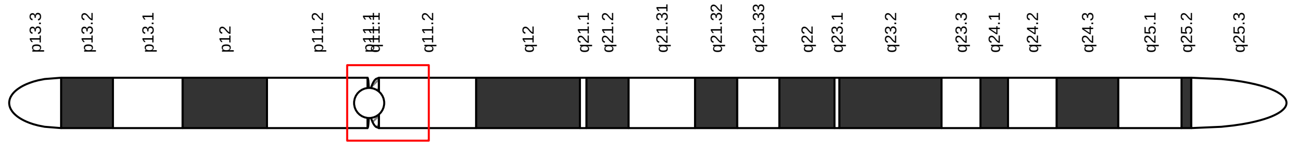
k−means region32 chr16:67411784−68661883 Subtypes Segregation − from TMCO7 to PDXDC2



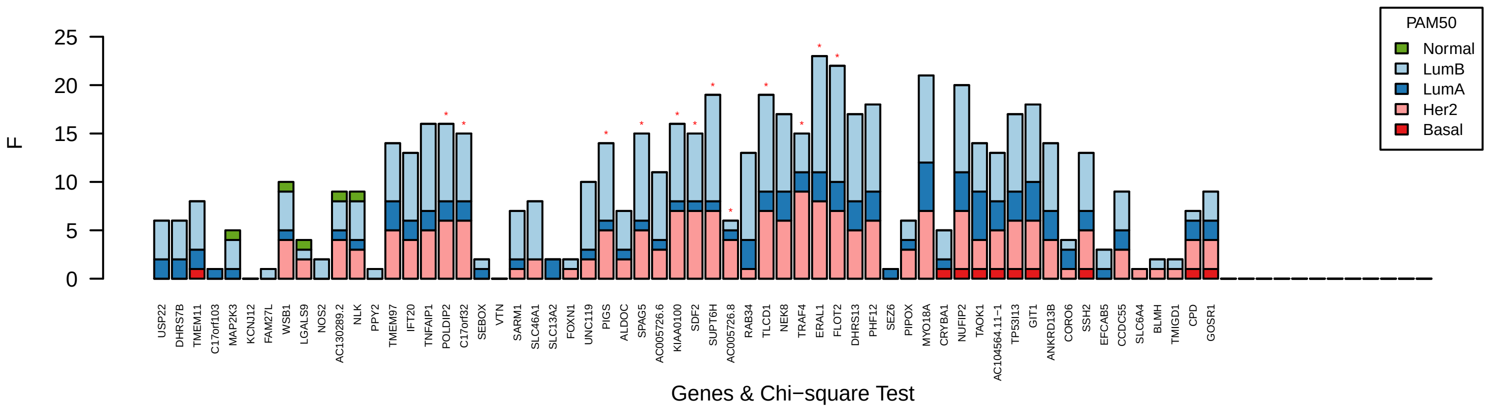
Genes & Chi−square Test


Chromosome 17 deletion outliers
p13.3
p13.2
p13.1
p12
p11.2
p11.1
q11.1
q11.2
q12
q21.1
q21.2
q21.31
q21.32
q21.33
q22
q23.1
q23.2
q23.3
q24.1
q24.2
q24.3
q25.1
q25.2
q25.3
PAM50 region33 chr17:11004441−12848195 Subtypes Segregation − from DNAH9 to ELAC2
PAM50
Normal
LumB
LumA
Her2
Basal
F
0
5
10
15
20
DNAH9
ZNF18
MAP2K4
AC005358.1
MYOCD
AC005277.1
ELAC2
Genes & Chi−square Test
IntClust region33 chr17:11004441−12848195 Subtypes Segregation − from DNAH9 to ELAC2
IntClust
1
2
3
4
5
6
7
8
9
10
F
0
5
10
15
20
DNAH9
ZNF18
MAP2K4
AC005358.1
MYOCD
AC005277.1
ELAC2
Genes & Chi−square Test
k−means region33 chr17:11004441−12848195 Subtypes Segregation − from DNAH9 to ELAC2
k−means
9
8
7
6
5
4
3
2
1
F
0
5
10
15
20
DNAH9
ZNF18
MAP2K4
AC005358.1
MYOCD
AC005277.1
ELAC2
Genes & Chi−square Test

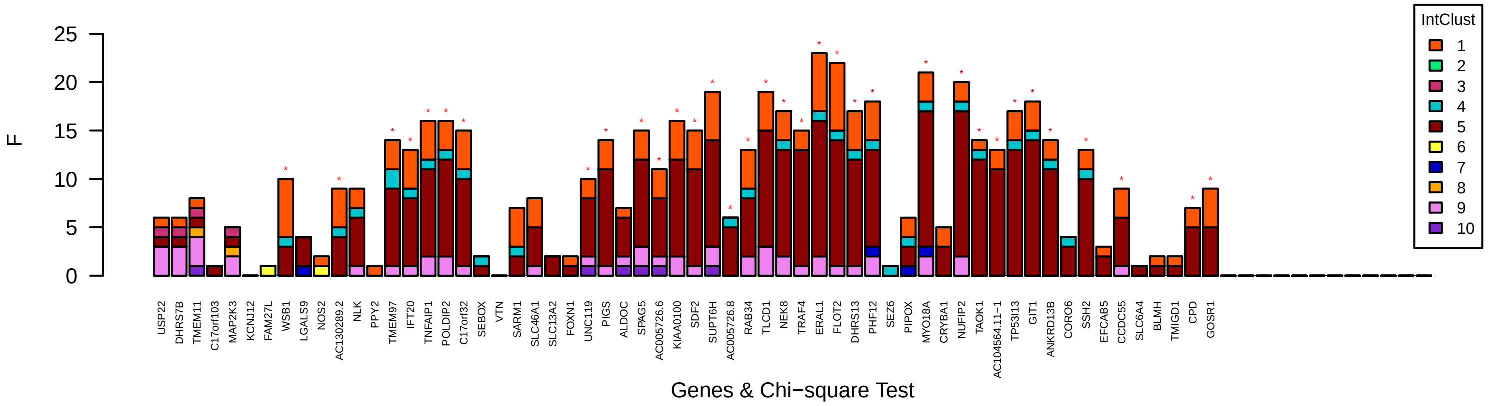
Chromosome 17 amplification outliers



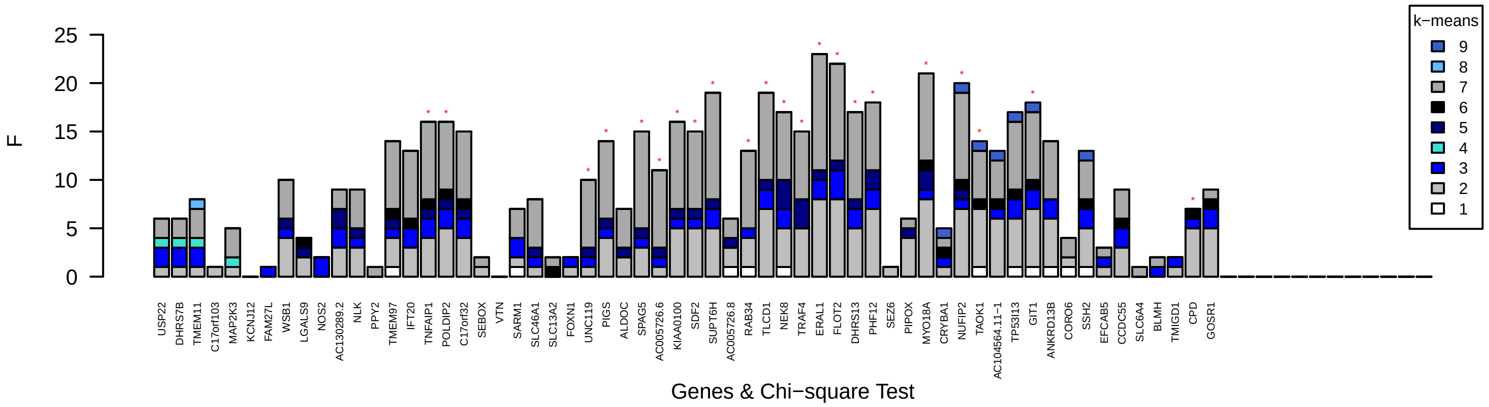
PAM50 region34 chr17:20843498−25877958 Subtypes Segregation − from USP22 to GOSR1



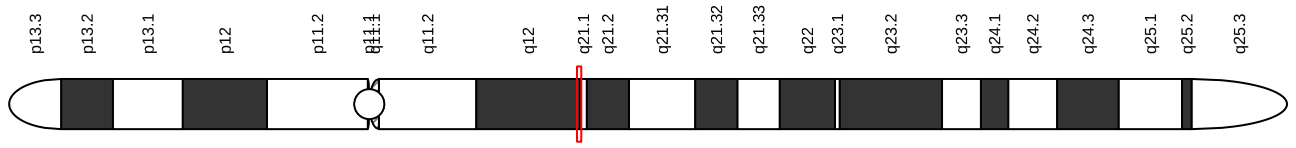
IntClust region34 chr17:20843498−25877958 Subtypes Segregation − from USP22 to GOSR1



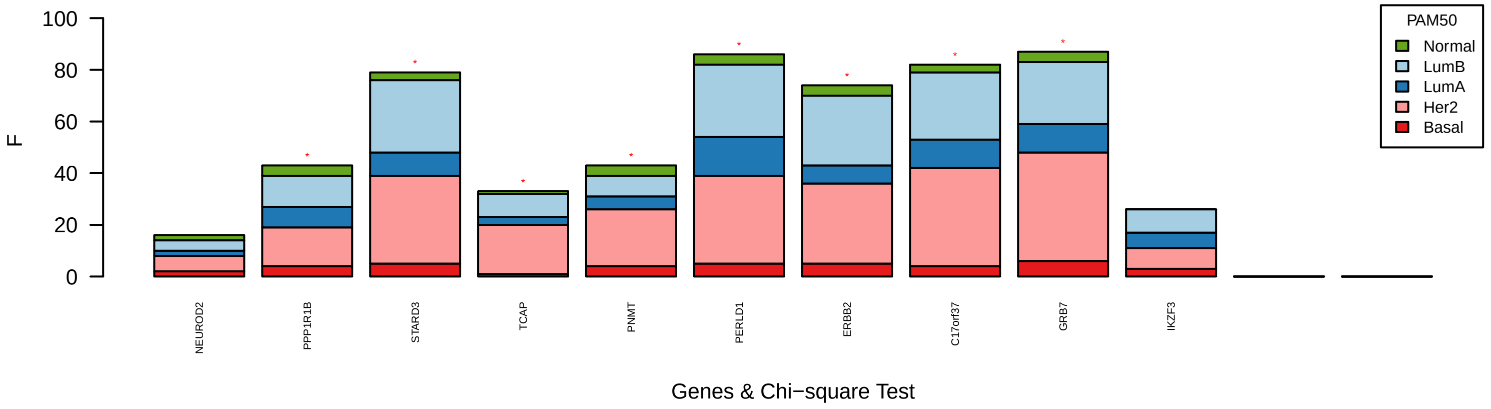
k−means region34 chr17:20843498−25877958 Subtypes Segregation − from USP22 to GOSR1



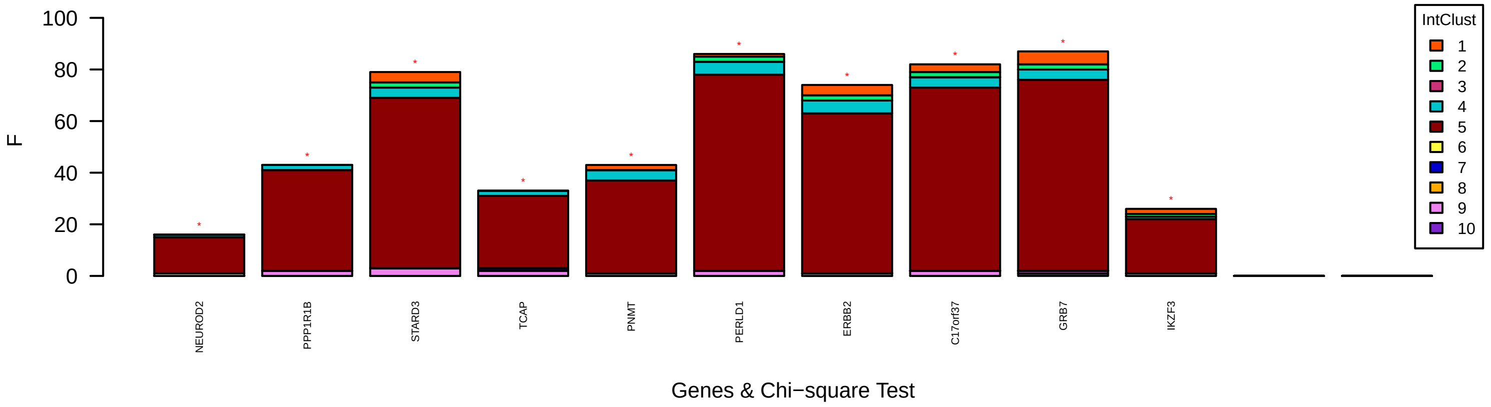
Chromosome 17 amplification outliers



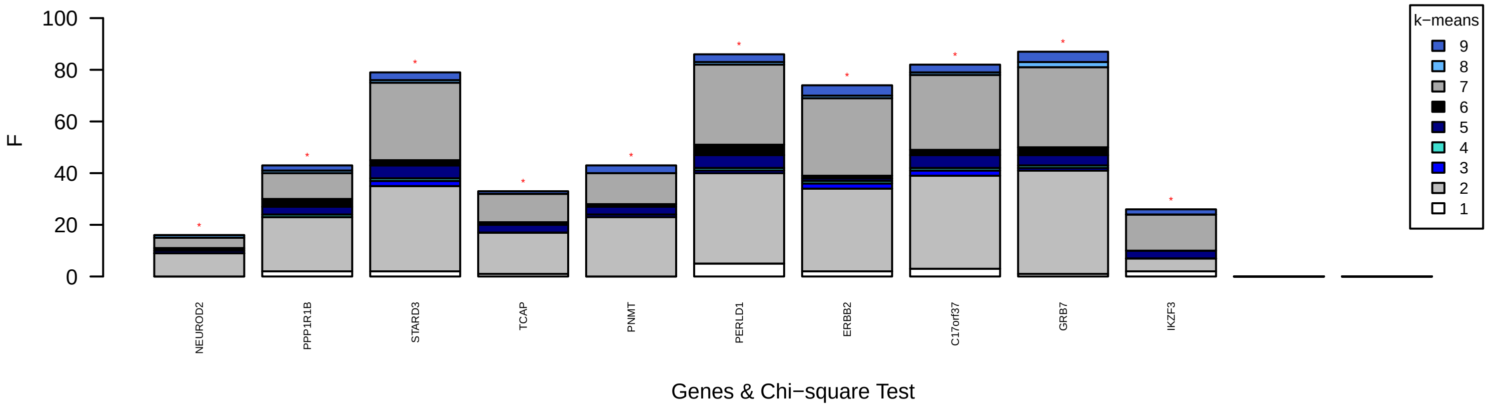
PAM50 region35 chr17:35015230−35273967 Subtypes Segregation − from NEUROD2 to IKZF3



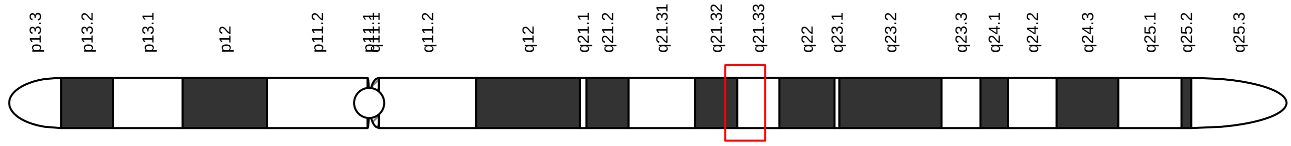
IntClust region35 chr17:35015230−35273967 Subtypes Segregation − from NEUROD2 to IKZF3



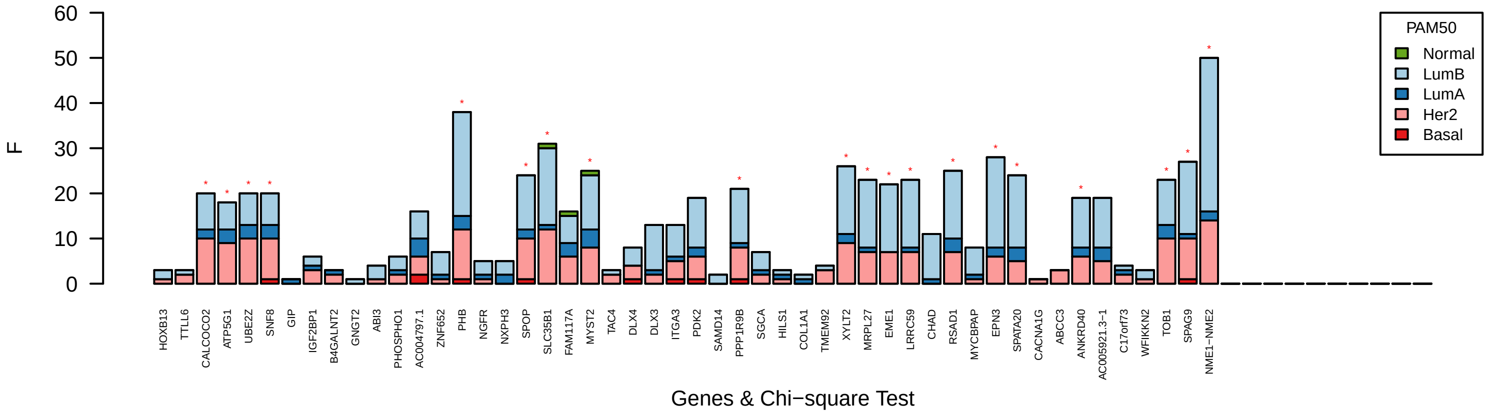
k−means region35 chr17:35015230−35273967 Subtypes Segregation − from NEUROD2 to IKZF3



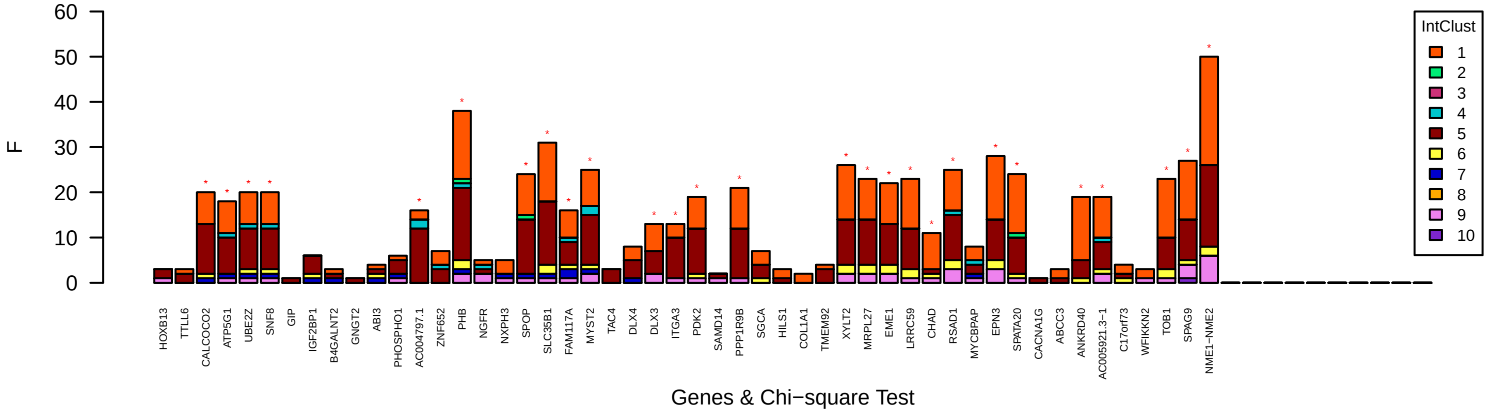
Chromosome 17 amplification outliers



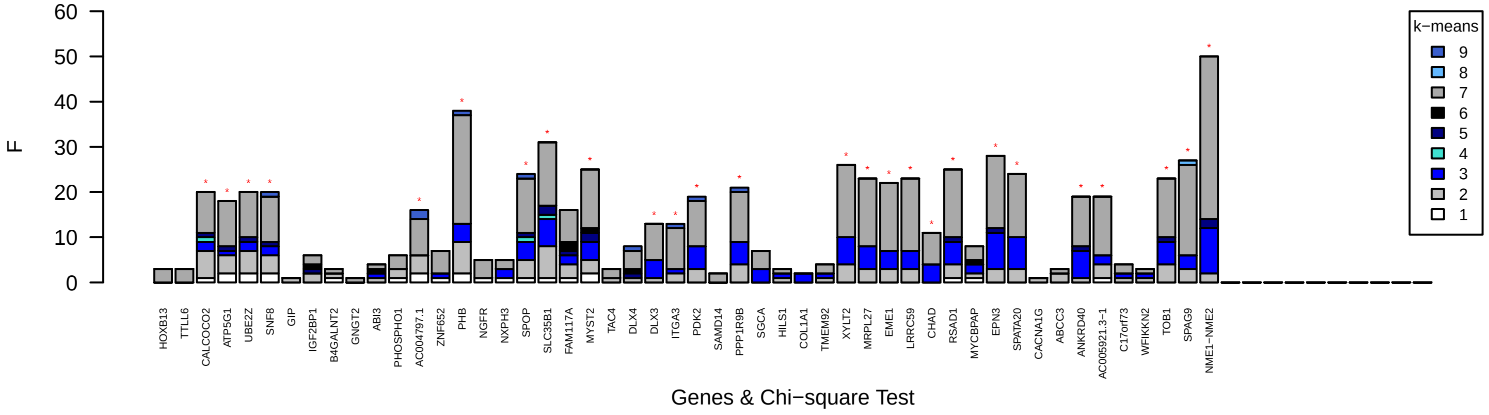
PAM50 region36 chr17:44157125−46621138 Subtypes Segregation − from HOXB13 to NME1−NME2



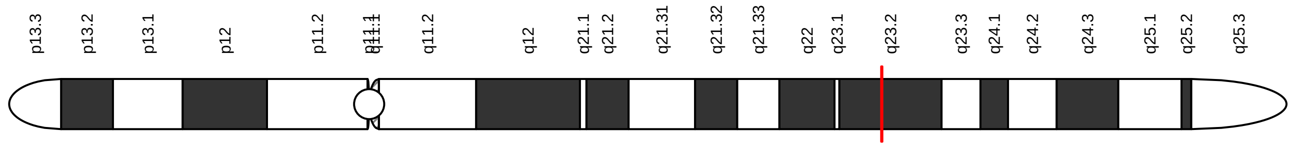
IntClust region36 chr17:44157125−46621138 Subtypes Segregation − from HOXB13 to NME1−NME2



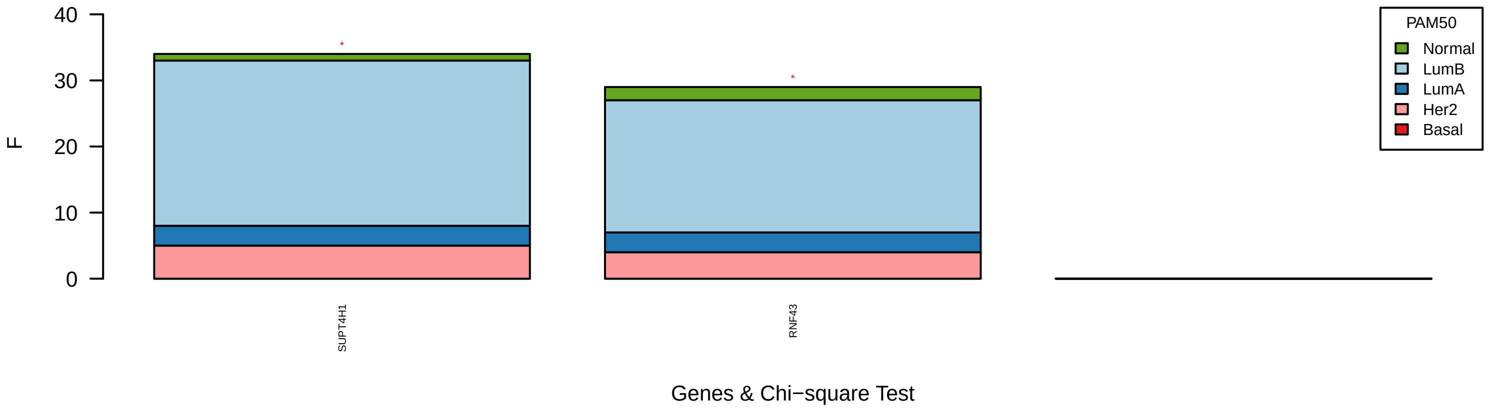
k−means region36 chr17:44157125−46621138 Subtypes Segregation − from HOXB13 to NME1−NME2



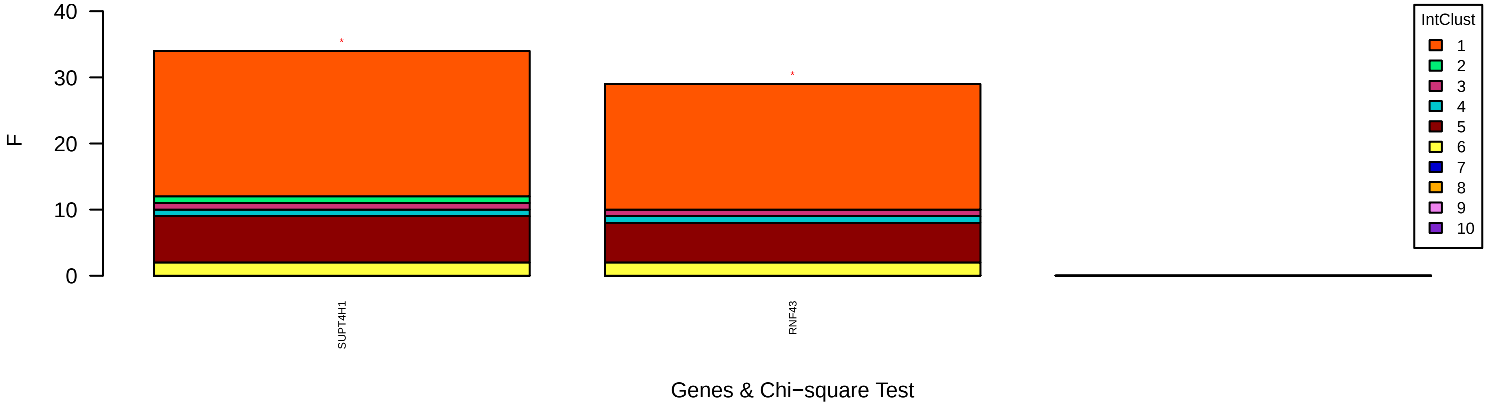
Chromosome 17 amplification outliers



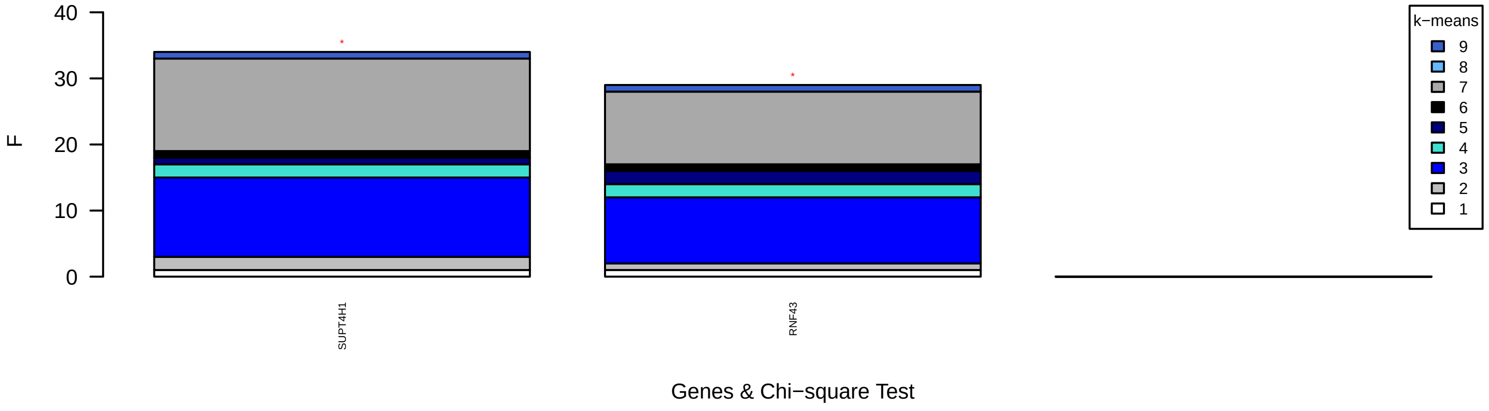
PAM50 region37 chr17:53777538−53849930 Subtypes Segregation − from SUPT4H1 to RNF43



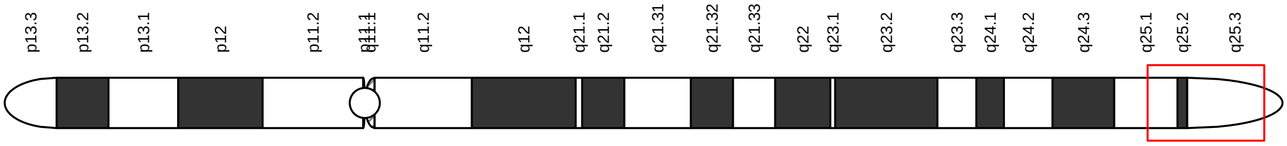
IntClust region37 chr17:53777538−53849930 Subtypes Segregation − from SUPT4H1 to RNF43



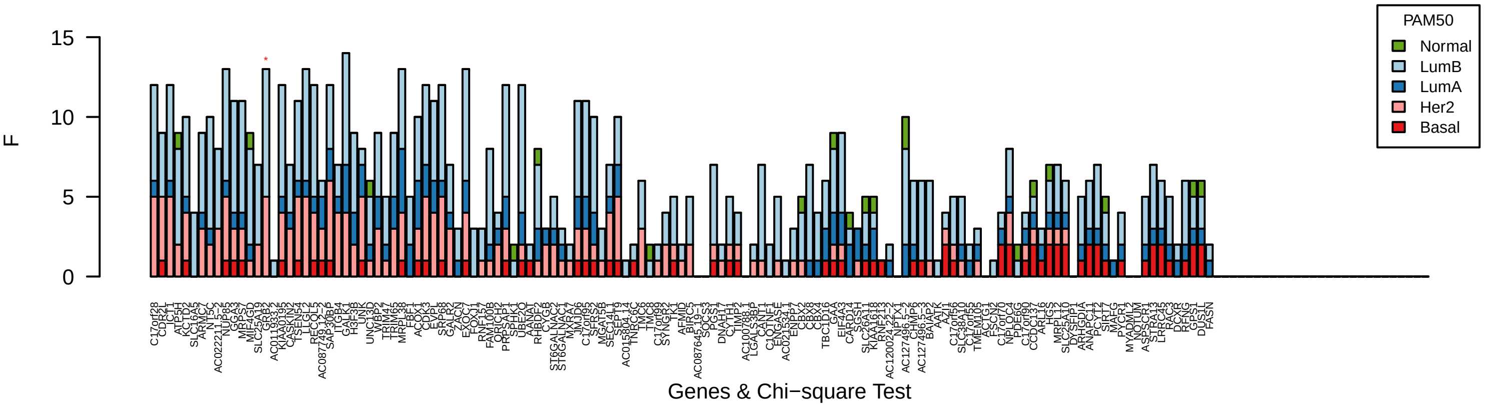
k−means region37 chr17:53777538−53849930 Subtypes Segregation − from SUPT4H1 to RNF43



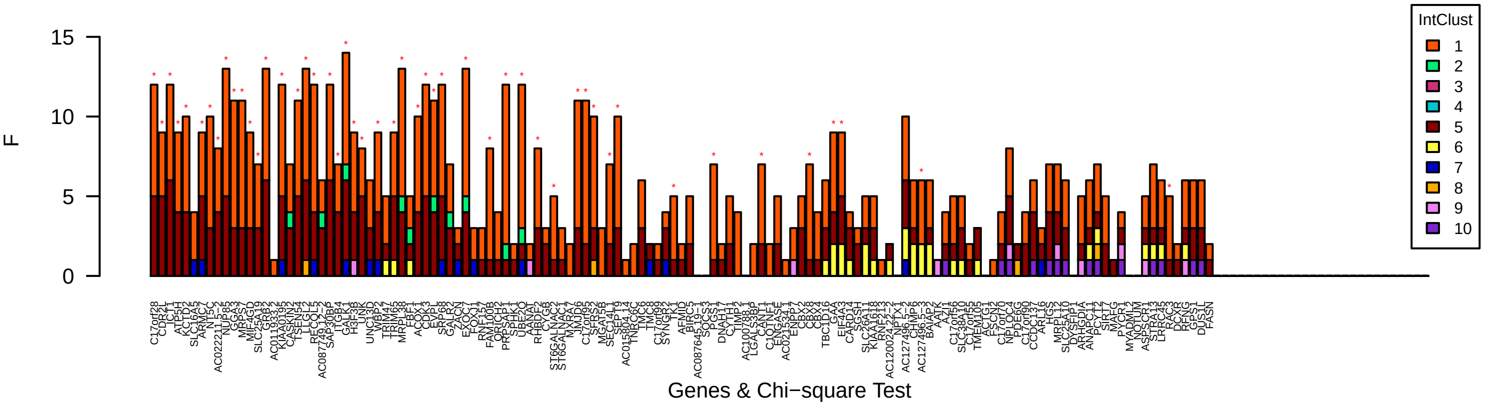
Chromosome 17 amplification outliers



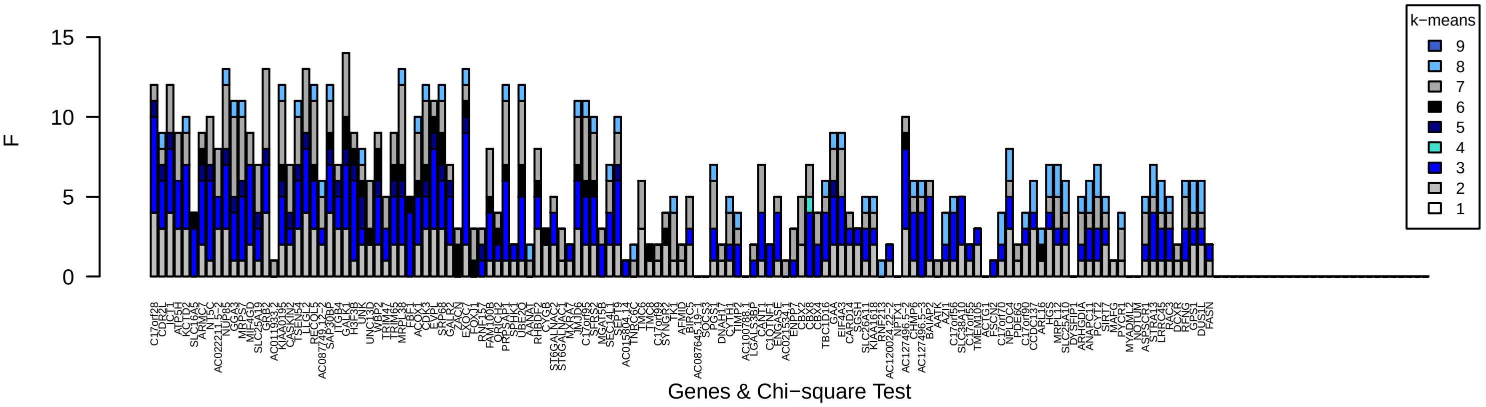
PAM50 region38 chr17:70458434−77649395 Subtypes Segregation − from C17orf28 to FASN



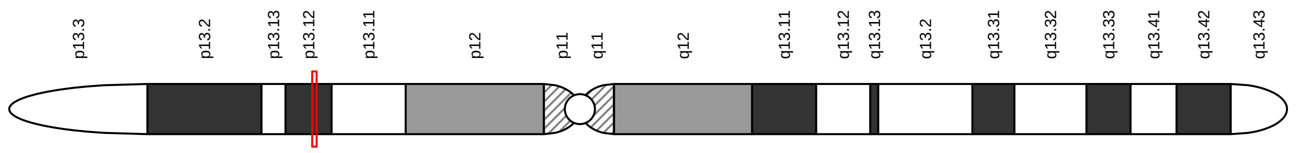
IntClust region38 chr17:70458434−77649395 Subtypes Segregation − from C17orf28 to FASN



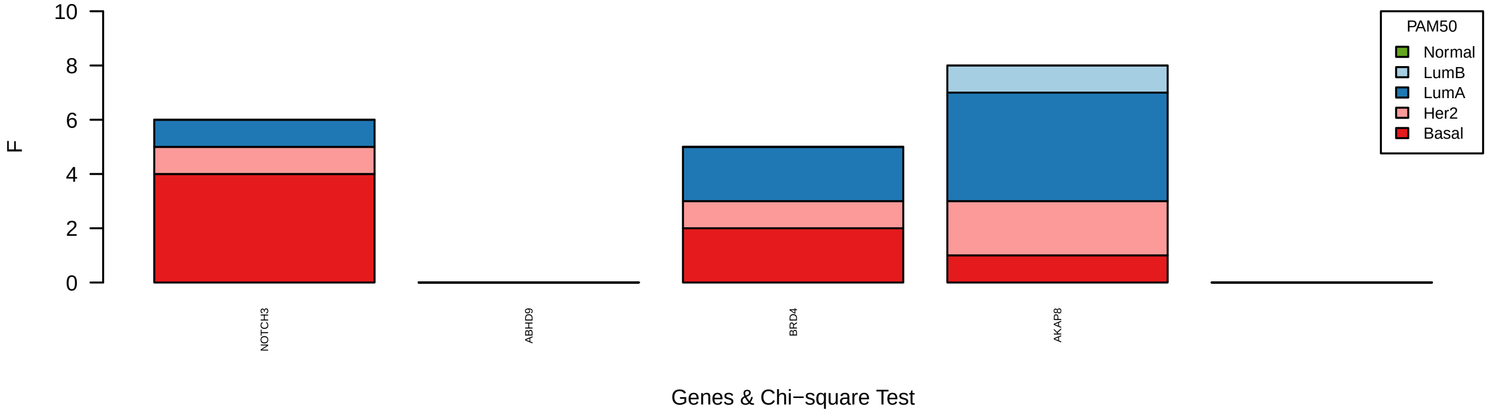
k−means region38 chr17:70458434−77649395 Subtypes Segregation − from C17orf28 to FASN



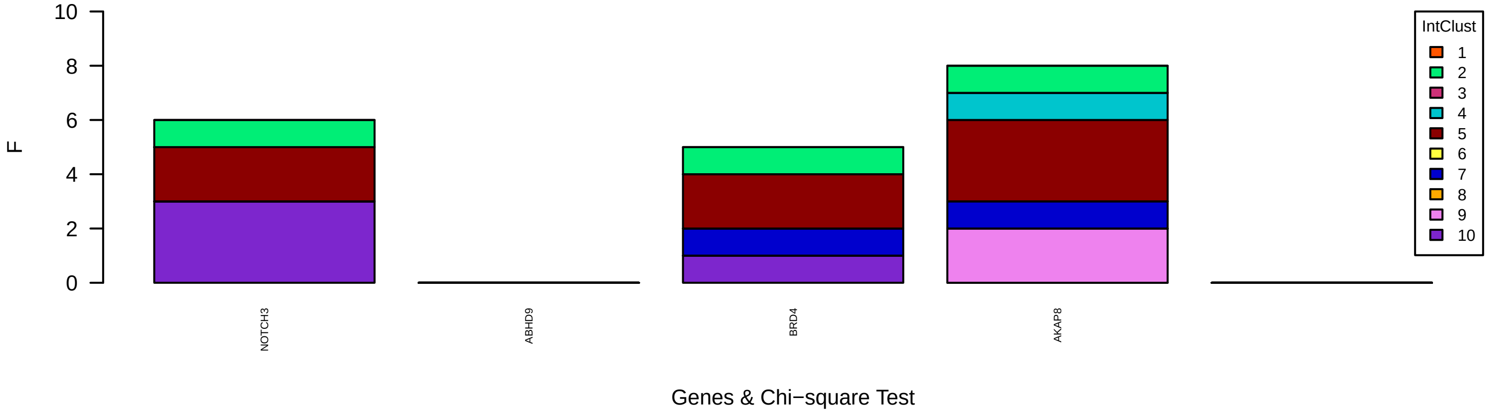
Chromosome 19 amplification outliers



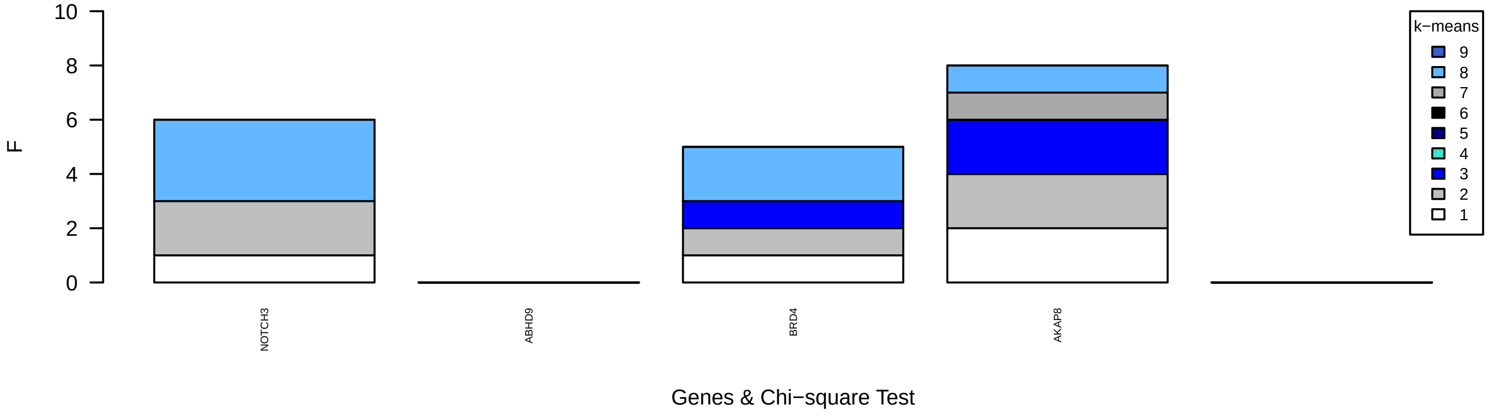
PAM50 region39 chr19:15131444−15351603 Subtypes Segregation − from NOTCH3 to AKAP8



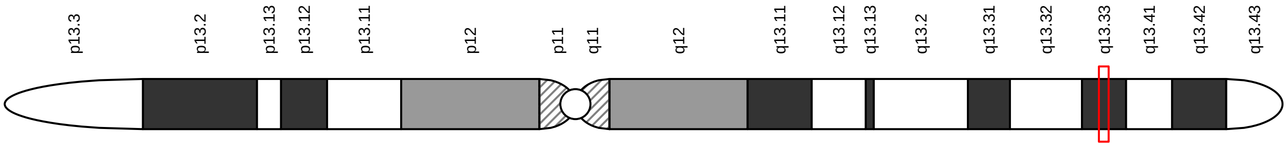
IntClust region39 chr19:15131444−15351603 Subtypes Segregation − from NOTCH3 to AKAP8



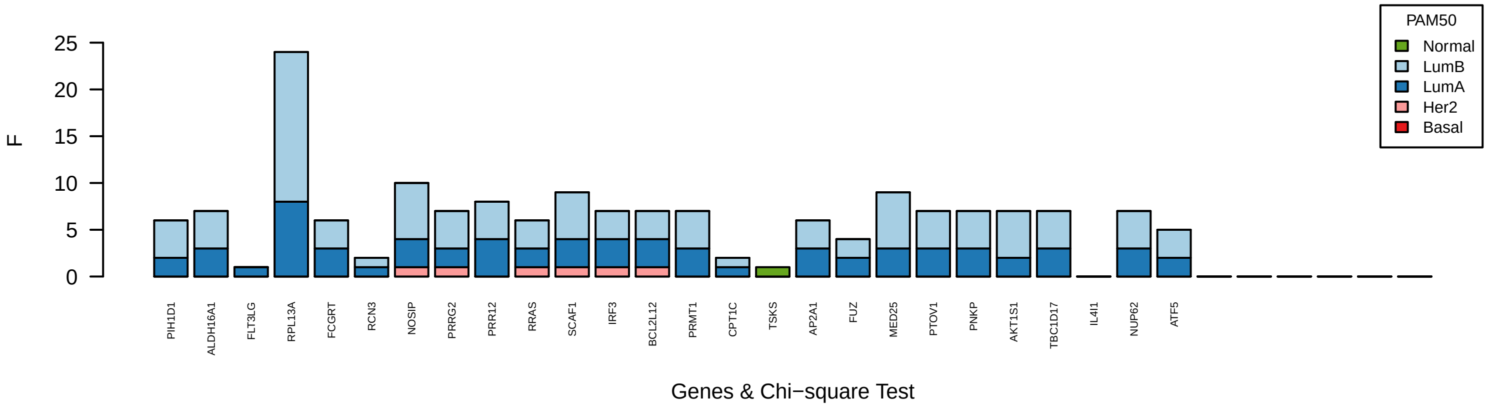
k−means region39 chr19:15131444−15351603 Subtypes Segregation − from NOTCH3 to AKAP8



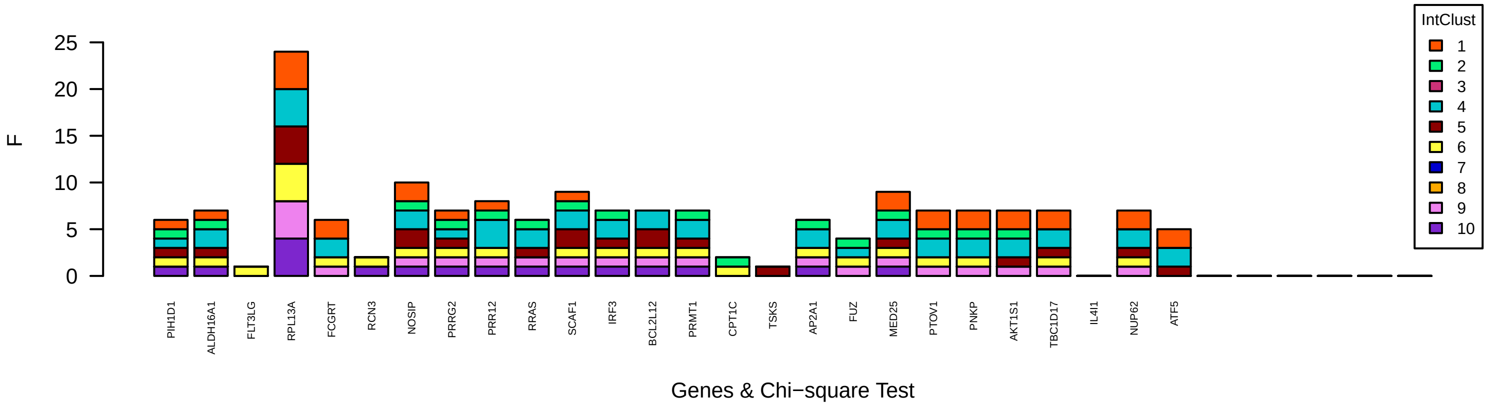
Chromosome 19 amplification outliers



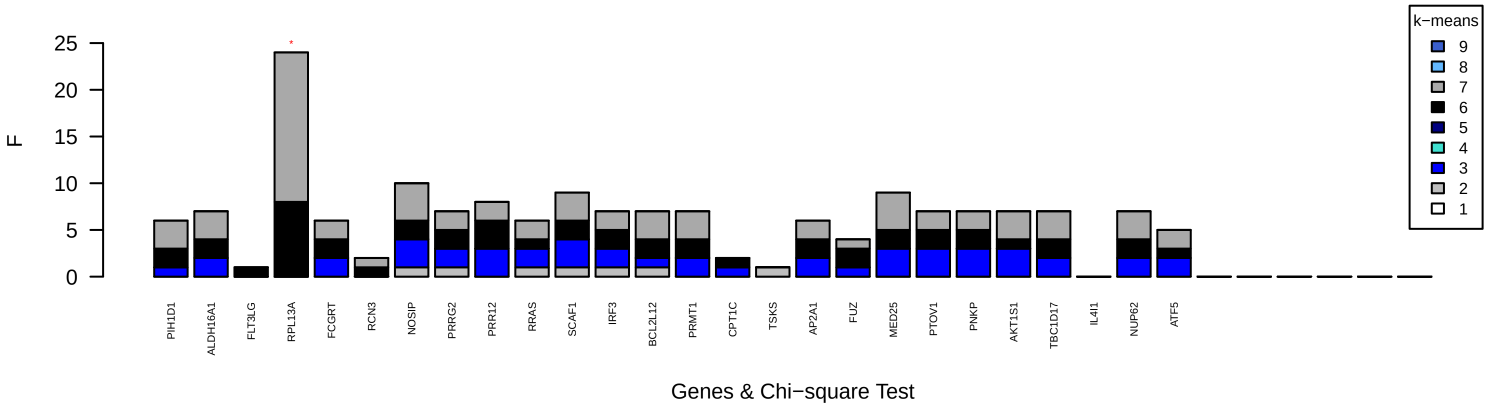
PAM50 region40 chr19:54641368−55129004 Subtypes Segregation − from PIH1D1 to ATF5



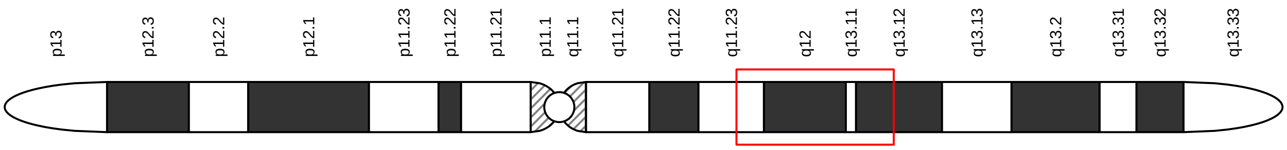
IntClust region40 chr19:54641368−55129004 Subtypes Segregation − from PIH1D1 to ATF5



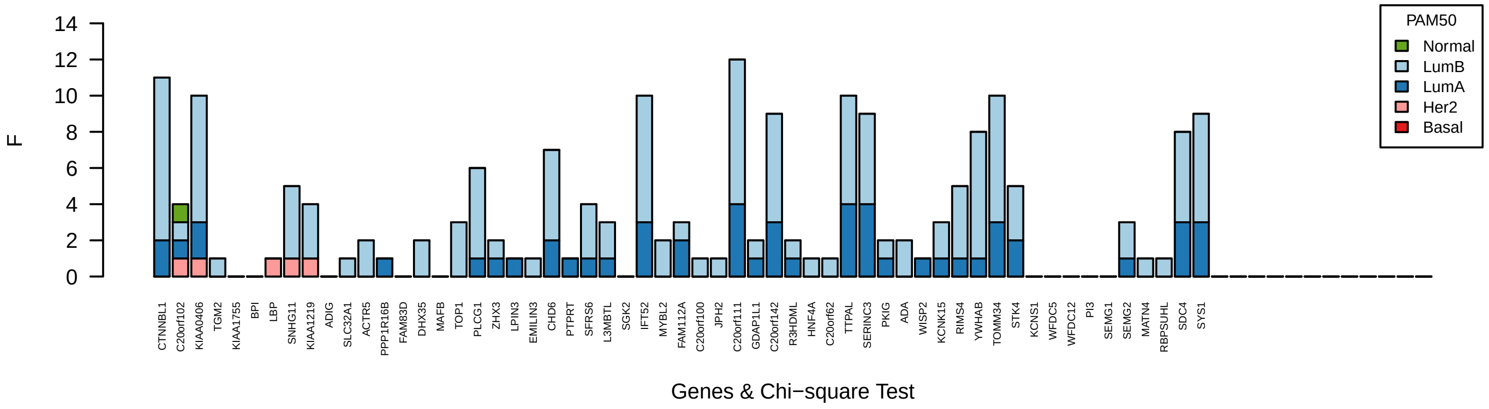
k−means region40 chr19:54641368−55129004 Subtypes Segregation − from PIH1D1 to ATF5



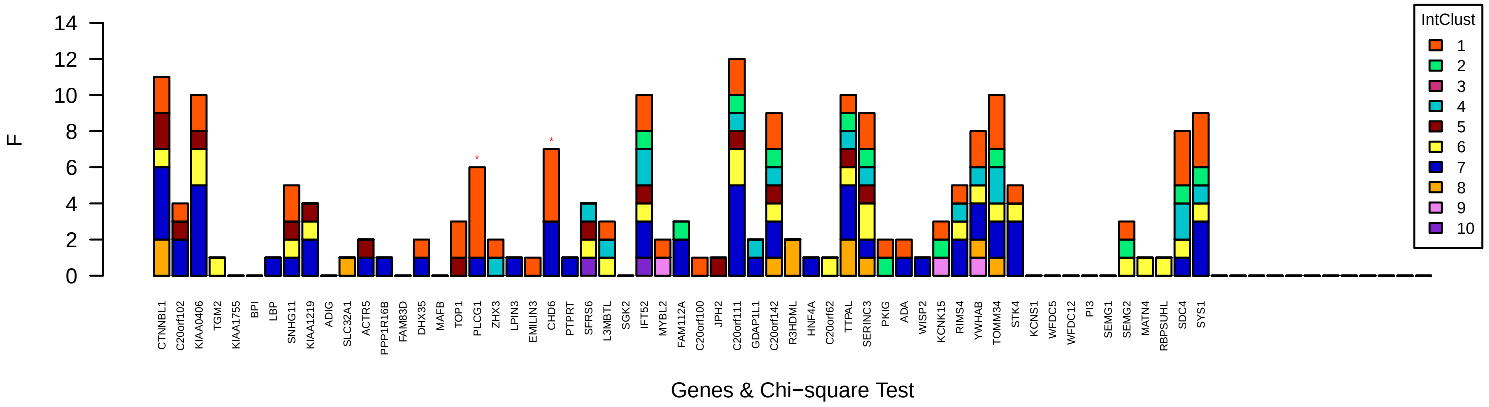
Chromosome 20 amplification outliers



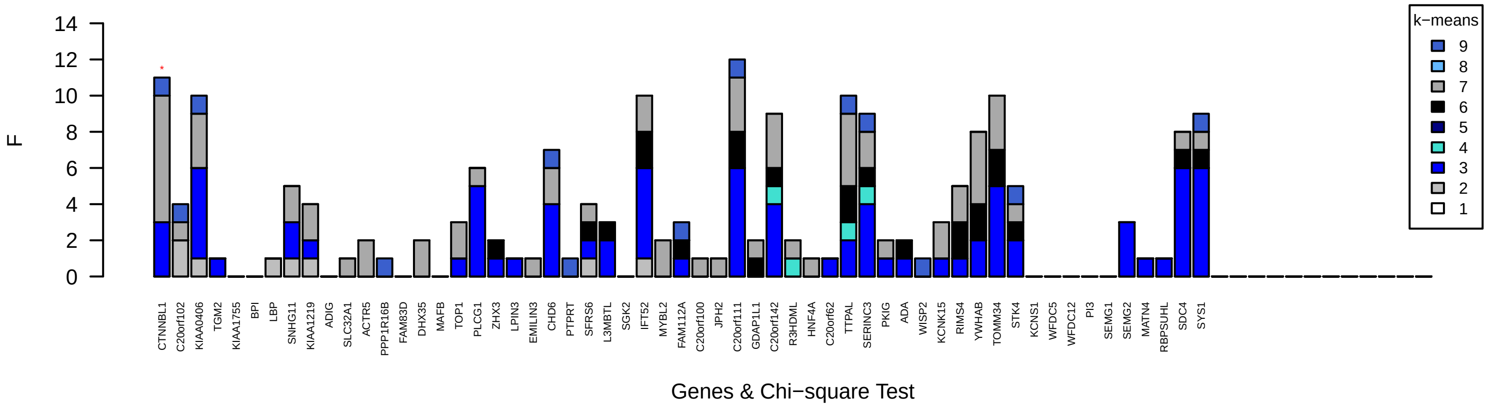
PAM50 region41 chr20:35755848−43440371 Subtypes Segregation − from CTNNBL1 to SYS1



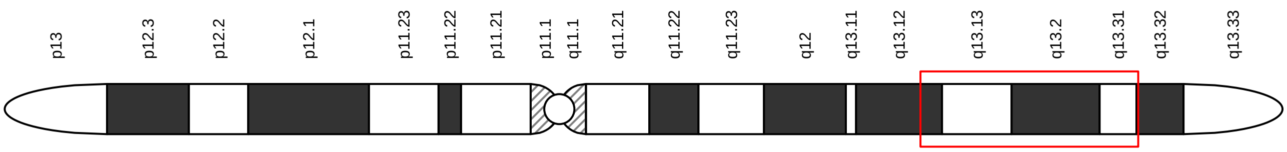
IntClust region41 chr20:35755848−43440371 Subtypes Segregation − from CTNNBL1 to SYS1



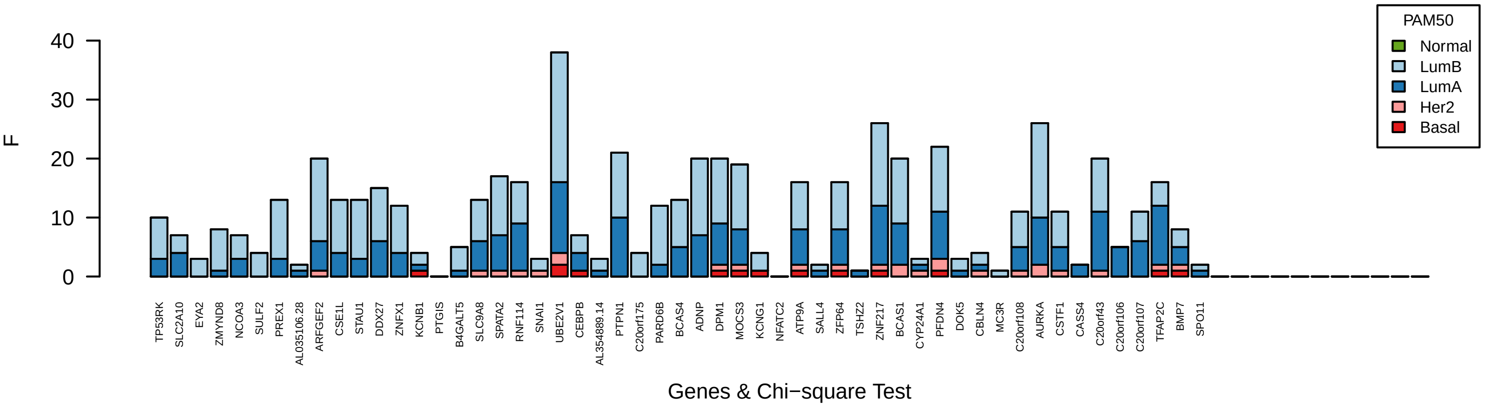
k−means region41 chr20:35755848−43440371 Subtypes Segregation − from CTNNBL1 to SYS1



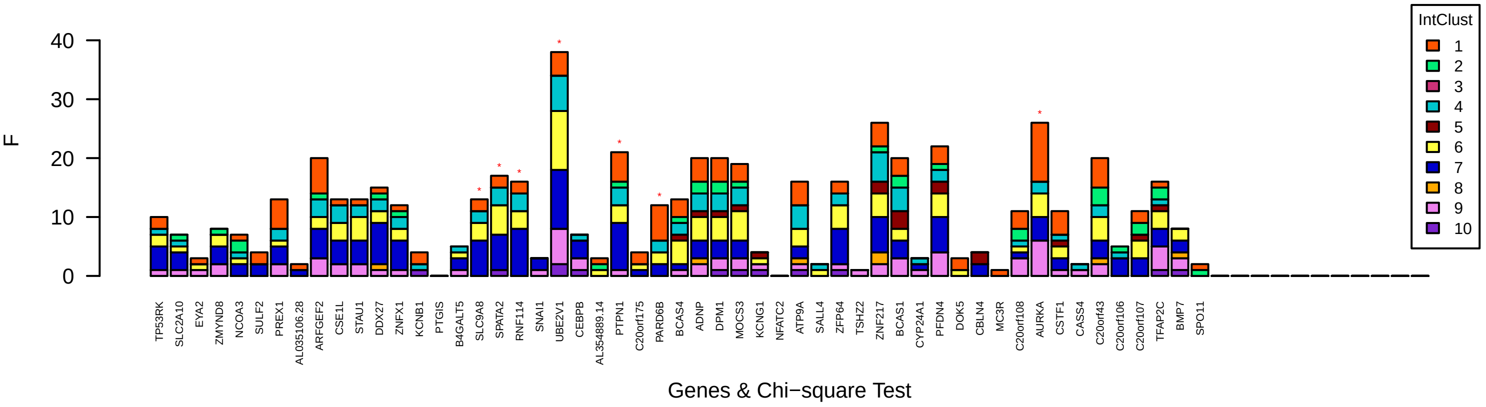
Chromosome 20 amplification outliers



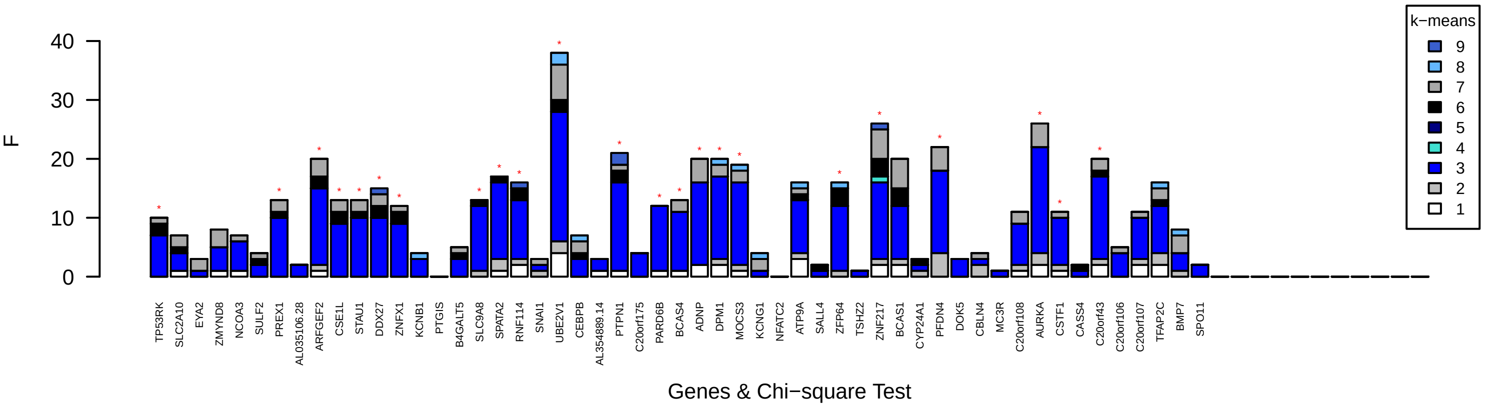
PAM50 region42 chr20:44746411−55386926 Subtypes Segregation − from TP53RK to SPO11



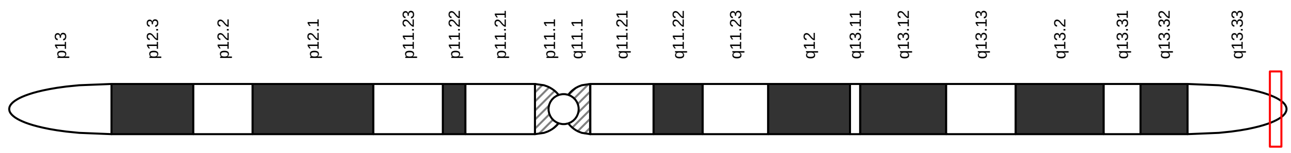
IntClust region42 chr20:44746411−55386926 Subtypes Segregation − from TP53RK to SPO11



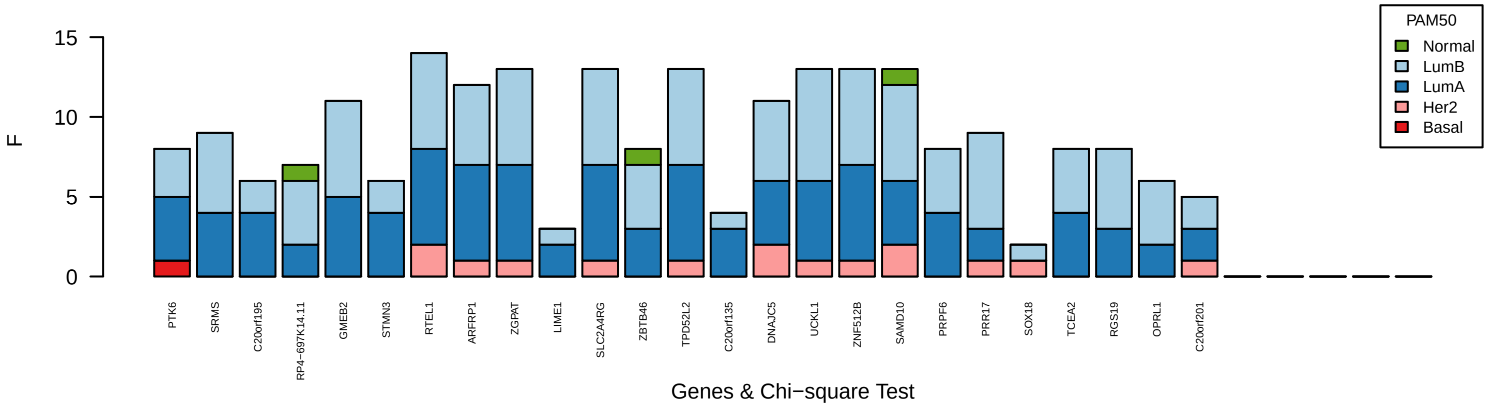
k−means region42 chr20:44746411−55386926 Subtypes Segregation − from TP53RK to SPO11



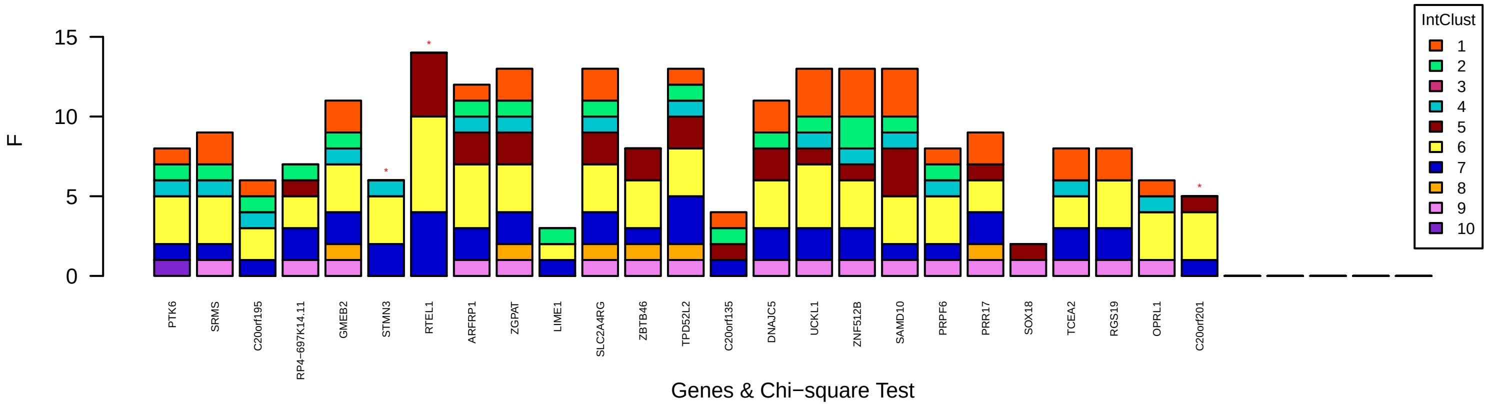
Chromosome 20 amplification outliers



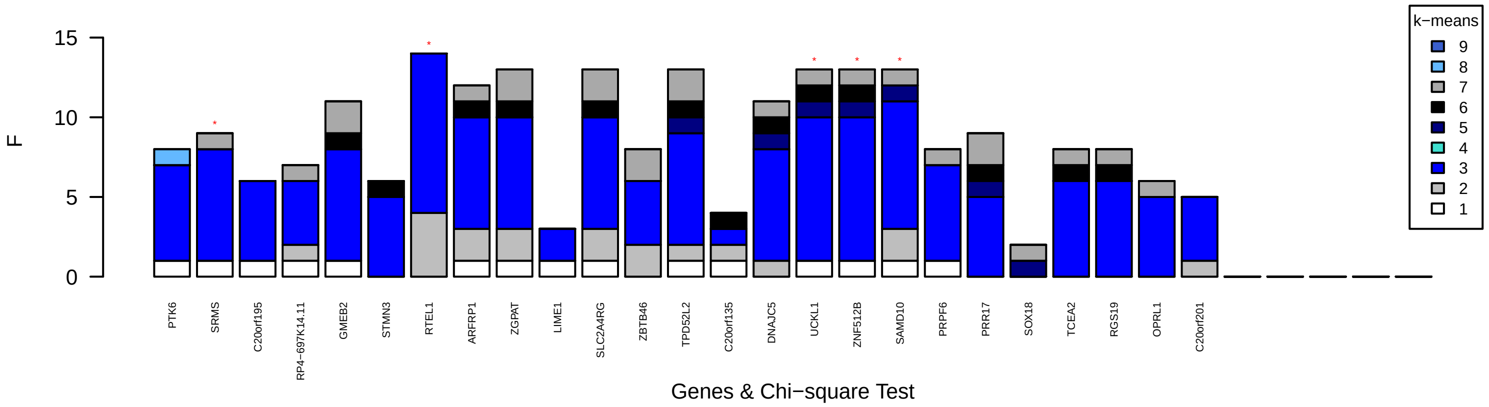
PAM50 region43 chr20:61622577−62186156 Subtypes Segregation − from PTK6 to C20orf201



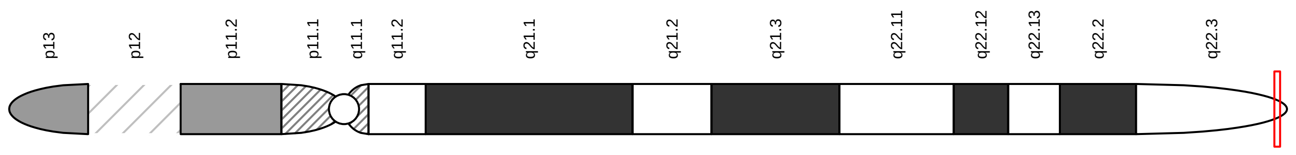
IntClust region43 chr20:61622577−62186156 Subtypes Segregation − from PTK6 to C20orf201



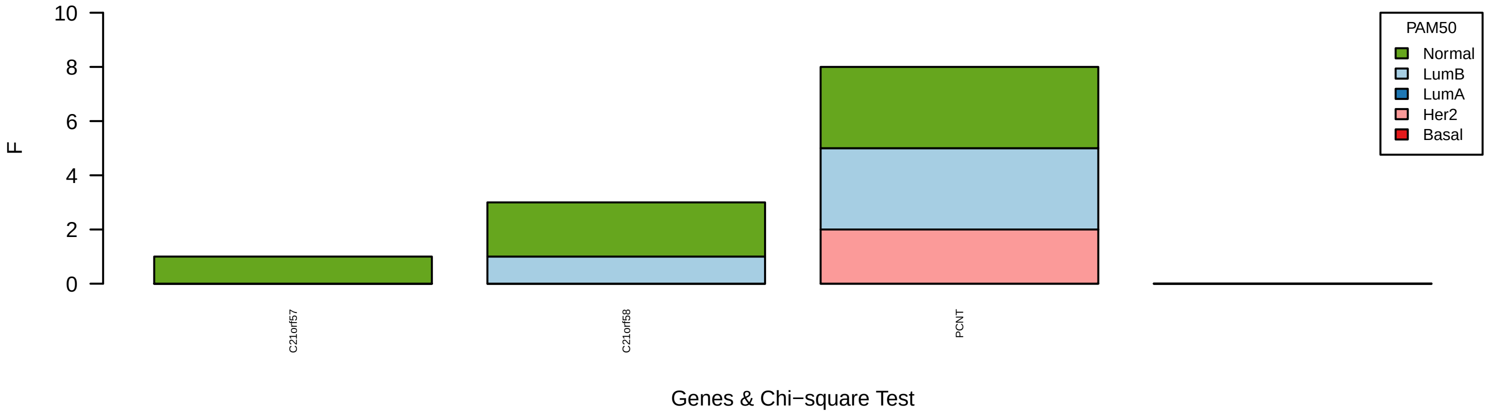
k−means region43 chr20:61622577−62186156 Subtypes Segregation − from PTK6 to C20orf201



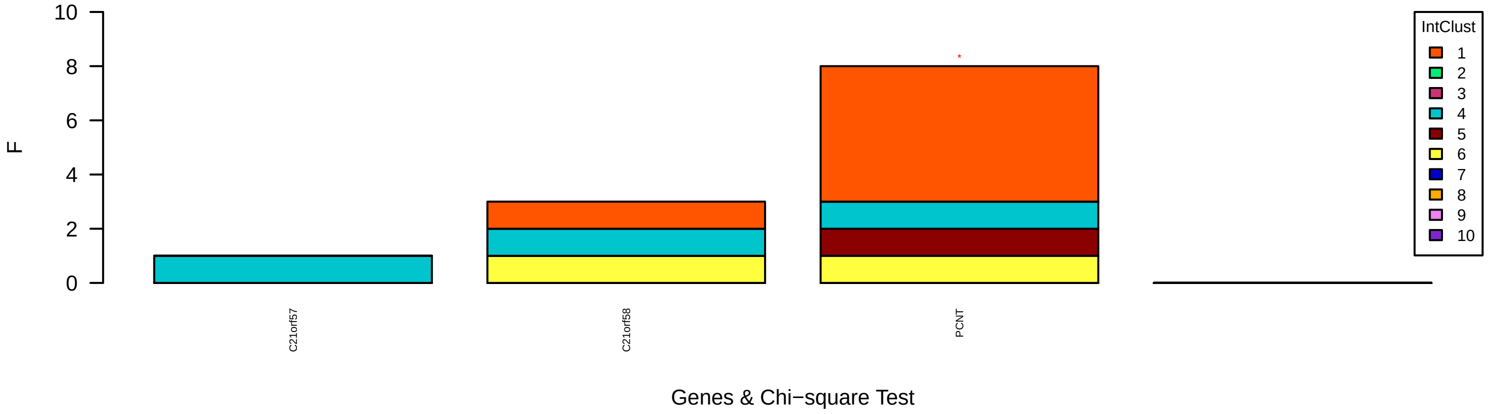
Chromosome 21 amplification outliers



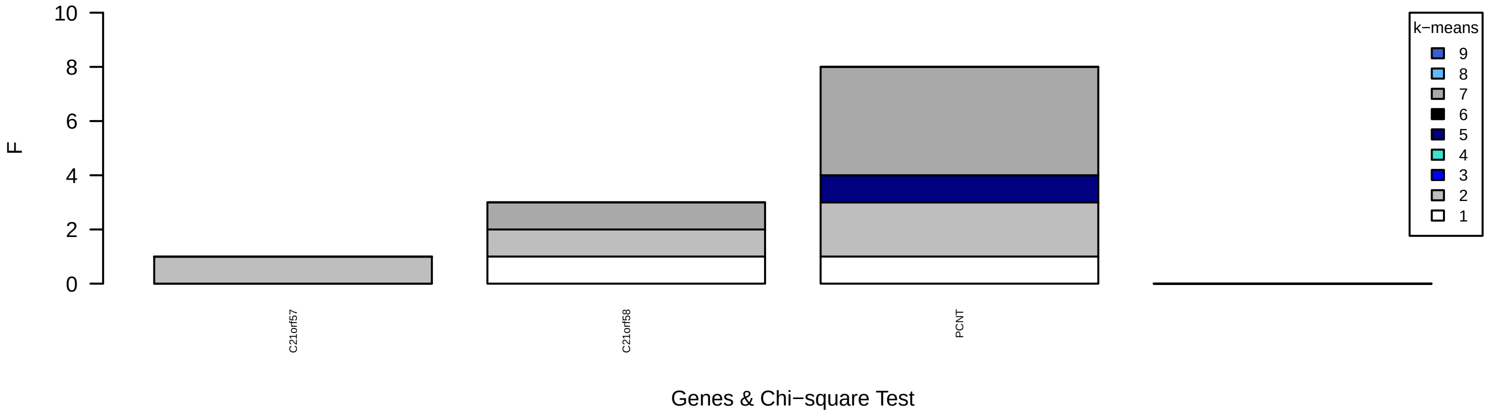
PAM50 region44 chr21:46479476−46690110 Subtypes Segregation − from C21orf57 to PCNT



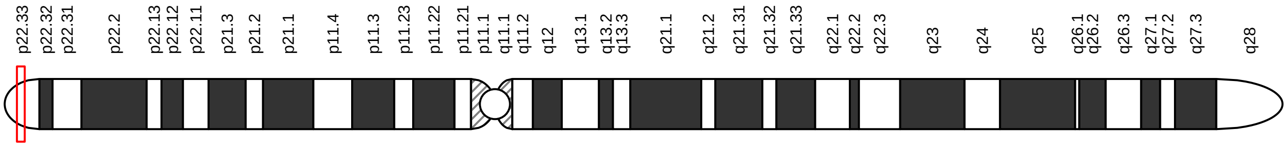
IntClust region44 chr21:46479476−46690110 Subtypes Segregation − from C21orf57 to PCNT



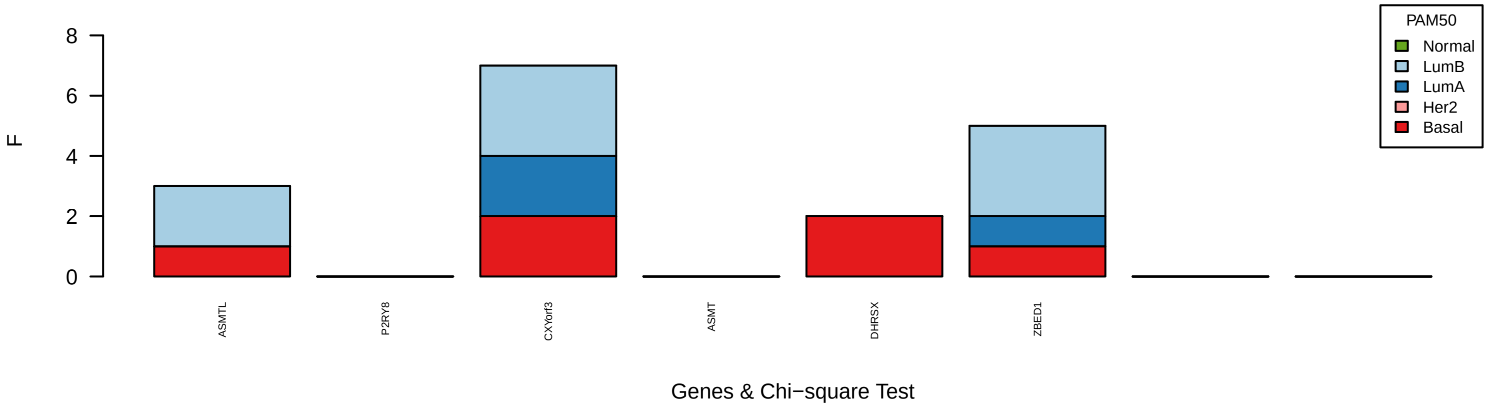
k−means region44 chr21:46479476−46690110 Subtypes Segregation − from C21orf57 to PCNT



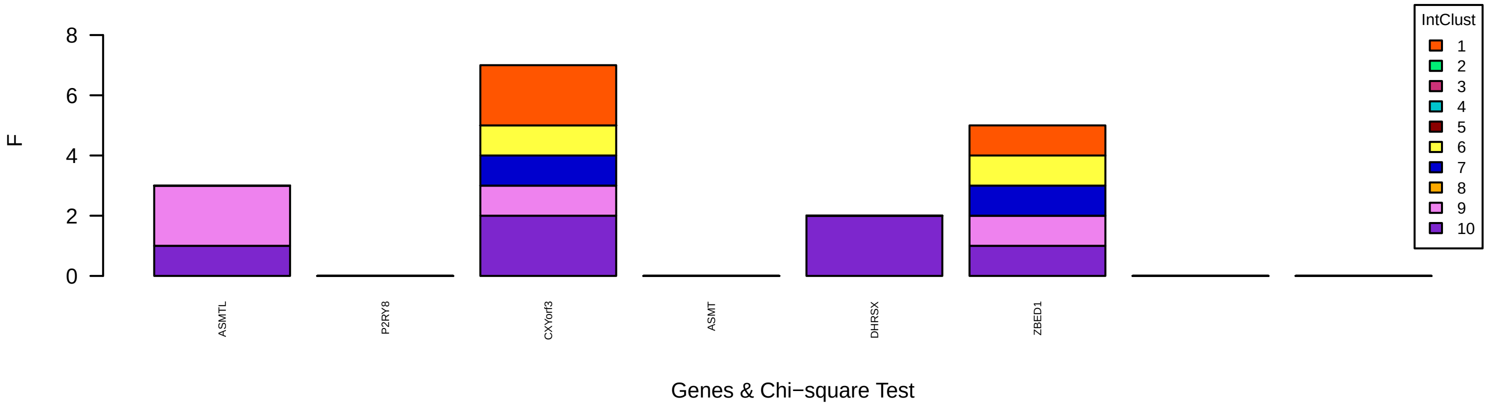
Chromosome X deletion outliers



PAM50 region45 chrX:1482032−2429015 Subtypes Segregation − from ASMTL to ZBED1



IntClust region45 chrX:1482032−2429015 Subtypes Segregation − from ASMTL to ZBED1



k−means region45 chrX:1482032−2429015 Subtypes Segregation − from ASMTL to ZBED1

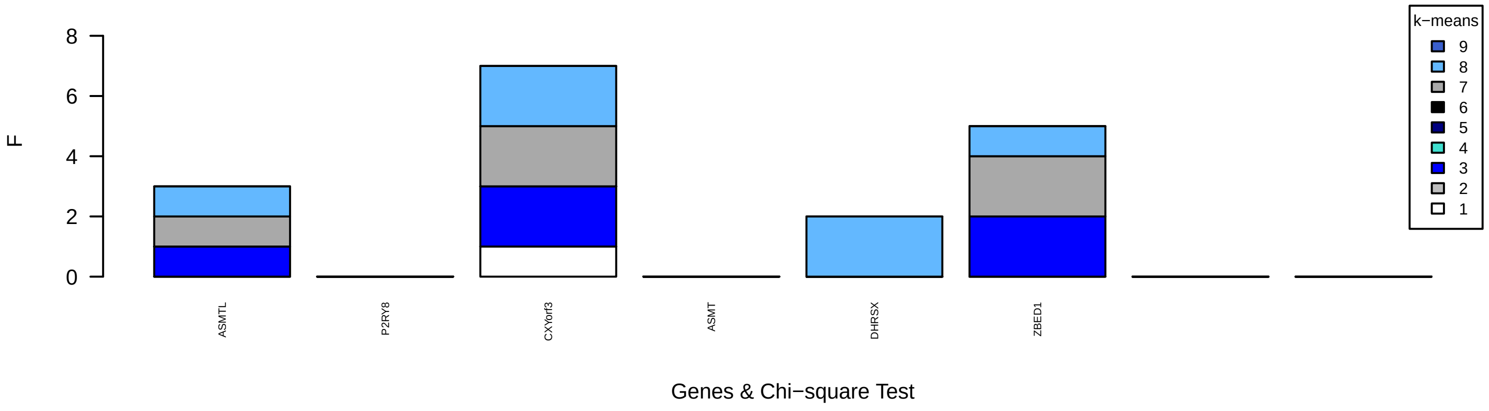
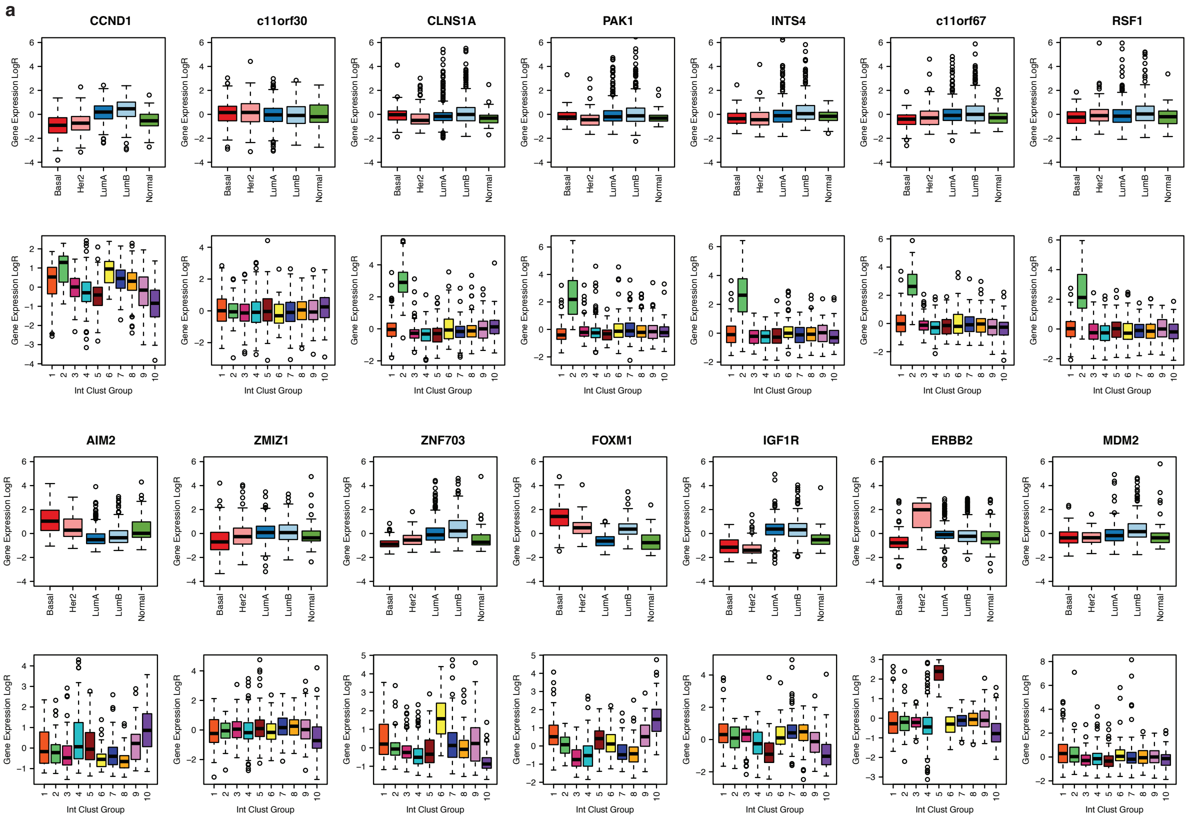


Figure S15: Patterns of CNA driven (HMM-based) outlying expression at selected regions throughout the genome. The top track shows the region of interest (red box) on a chromosome ideogram. The three sets of barplots indicate the subtype distribution for PAM50, IntClust, and K-Means derived subgroups for each gene found within the region of interest. Red asterisks above the bar plots indicate significantly different observed distributions than expected based on the overall population frequency (χ^2, P-value < 0.0001).


a
CCND1
c11orf30
CLNS1A
PAK1
INTS4
c11orf67
RSF1
AIM2
ZMIZ1
ZNF703
FOXM1
IGF1R
ERBB2
MDM2
Gene Expression LogR
Basal
Her2
LumA
LumB
Normal
1 2 3 4 5 6 7 8 9 10
Int Clust Group

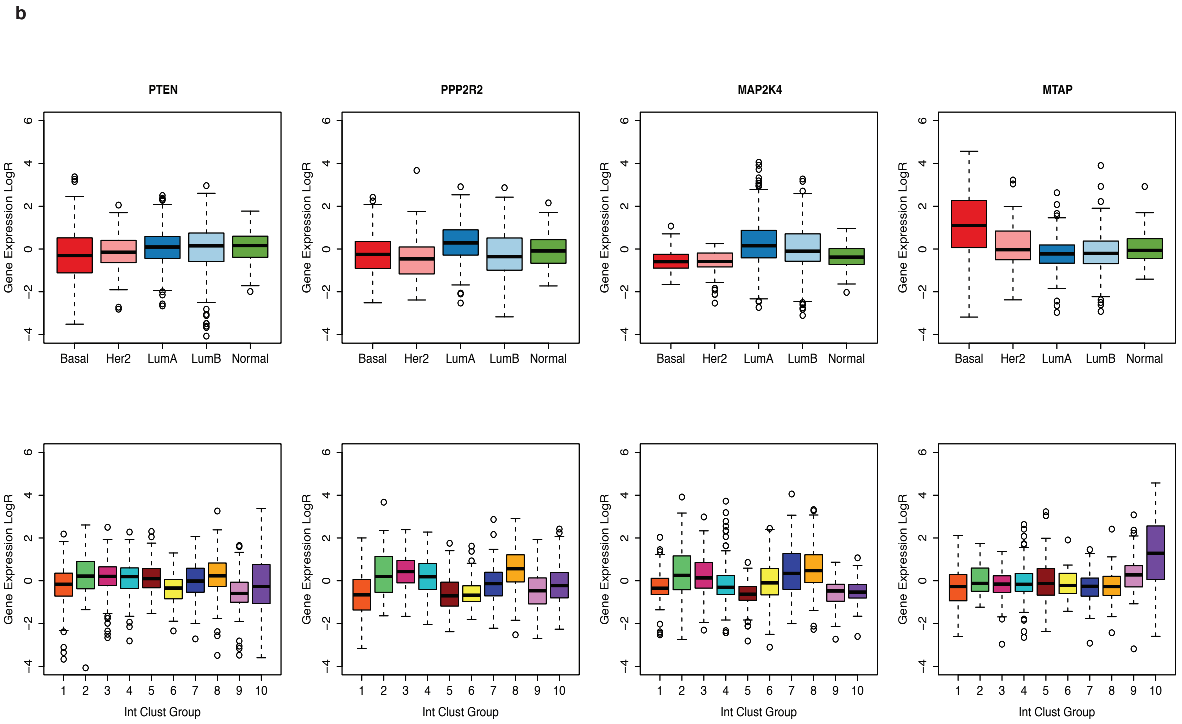
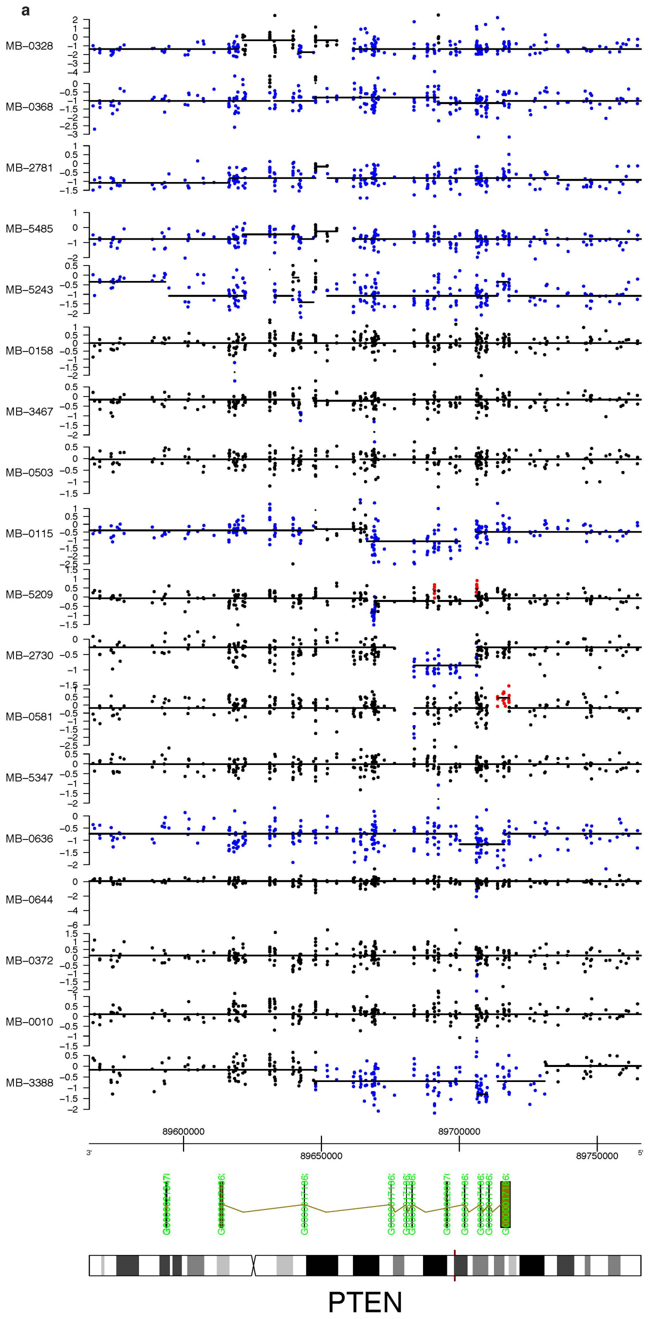
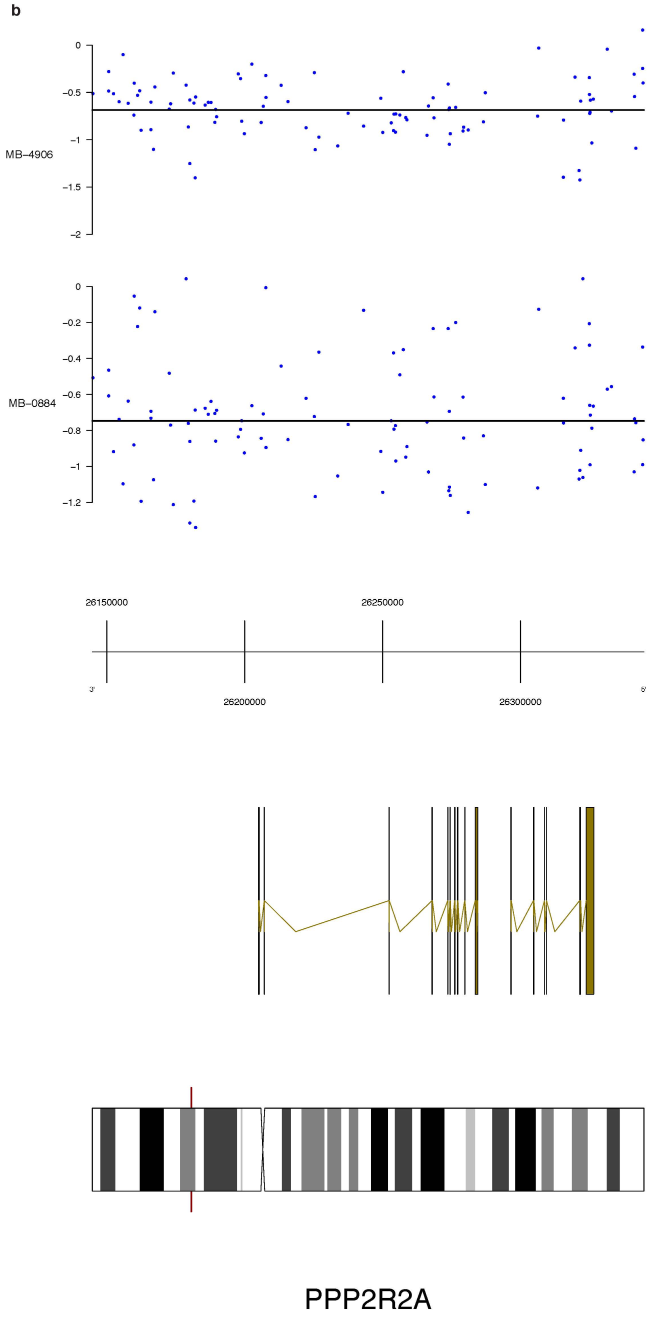
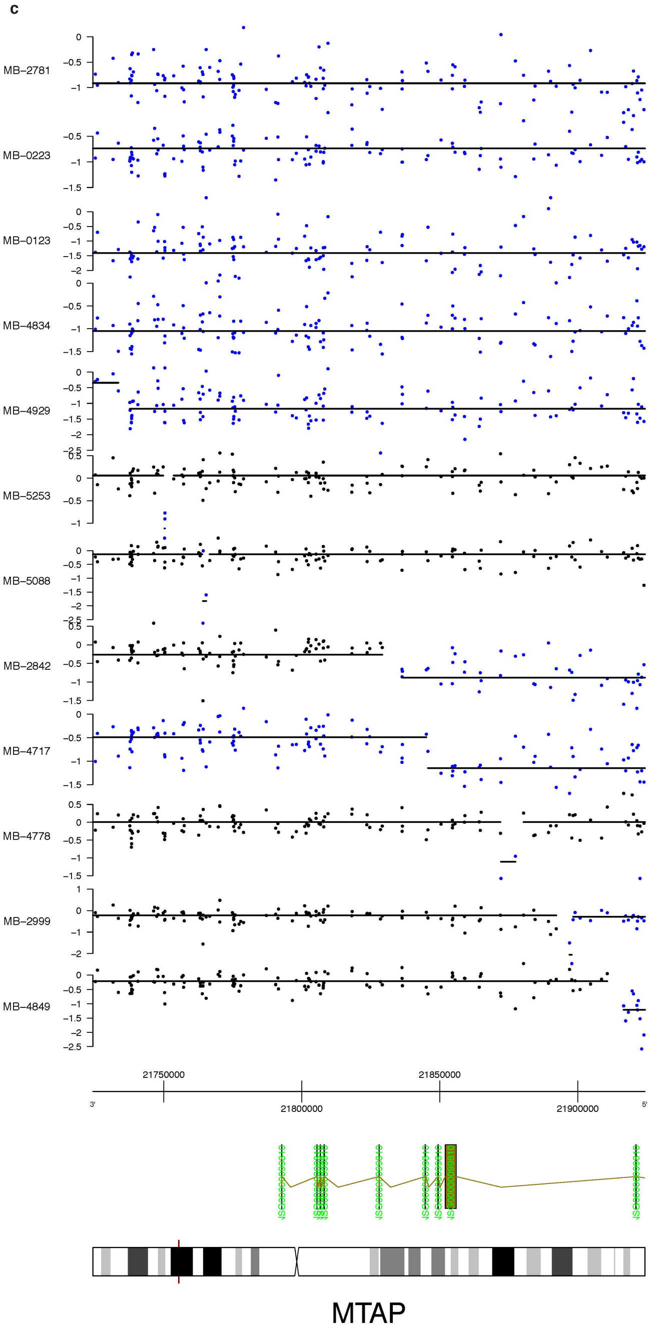


Figure S16: Boxplots of $\log_2$ expression stratified by the PAM50 or integrative subtypes are shown for select genomic regions. As noted in the text, these regions harbour **a.** predicted amplifications or **b.** homozygous deletions.


a
MB–0328
MB–0368
MB–2781
MB–5485
MB–5243
MB–0158
MB–3467
MB–0503
MB–0115
MB–5209
MB–2730
MB–0581
MB–5347
MB–0636
MB–0644
MB–0372
MB–0010
MB–3388
89600000
89650000
89700000
89750000
3'
5'
PTEN


b
0
-0.5
-1
-1.5
-2
MB–4906
0
-0.2
-0.4
-0.6
-0.8
-1
-1.2
MB–0884
26150000
26250000
3'
5'
26200000
26300000
PPP2R2A


c
MB–2781
MB–0223
MB–0123
MB–4834
MB–4929
MB–5253
MB–5088
MB–2842
MB–4717
MB–4778
MB–2999
MB–4849
0
−0.5
−1
−1.5
−2
−2.5
0.5
1
21750000
21850000
3'
21800000
21900000
5'
MTAP

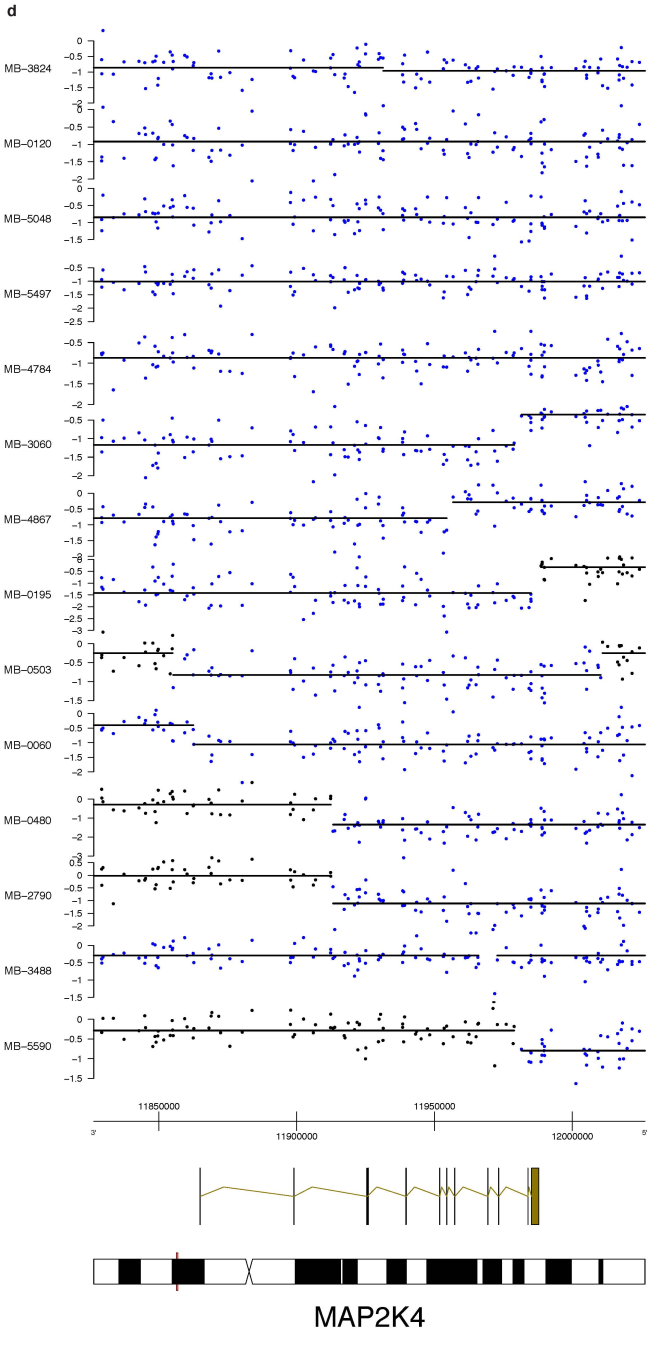


Figure S17: Probe-level plots of putative homozygous deletions referred to in the text. **a.** Putative homozygous deletions as predicted by CBS analysis are shown for *PTEN*, where red indicates regions of gain, black indicates copy number neutral regions, and blue indicates regions of loss. **b.** As in (a) for *PPP2R2A*. **c.** As in (a) for *MTAP*. **d.** As in (a) for *MAP2K4*.

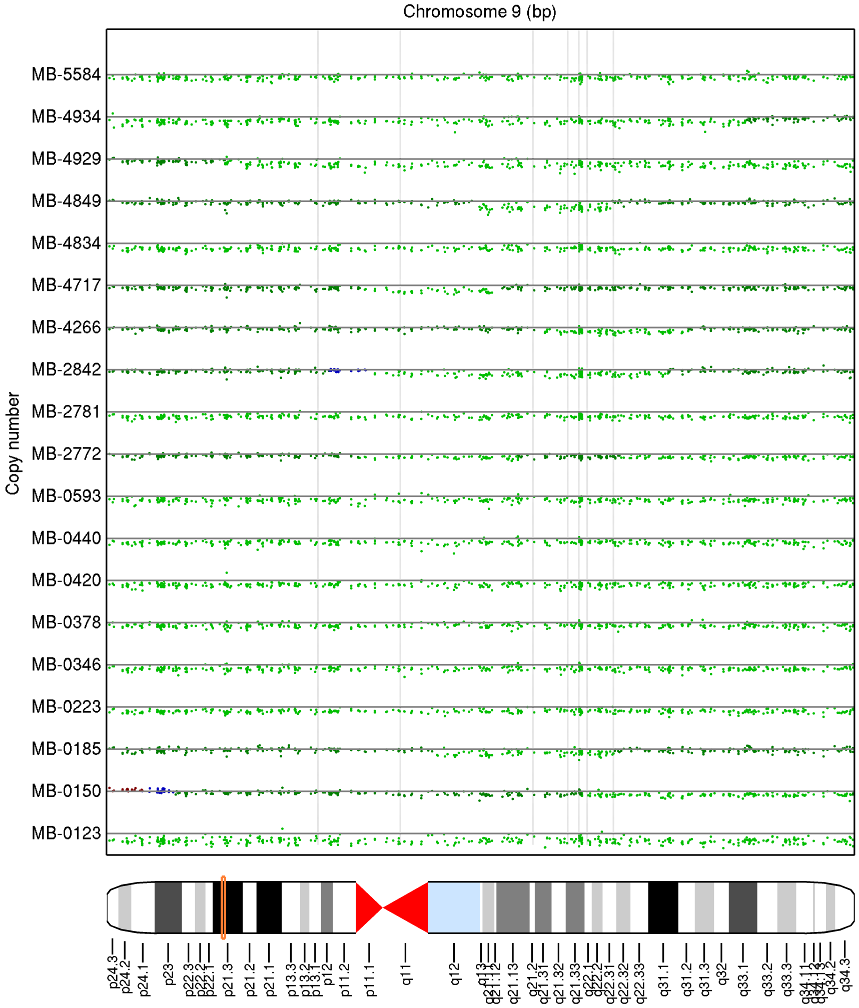


Figure S18: Putative *MTAP* homozygous deletions as predicted by HMM analysis. Here red indicates regions of gain, blue indicates copy number neutral regions, and green indicates regions of loss, where more intense shading corresponds to higher amplitude events.

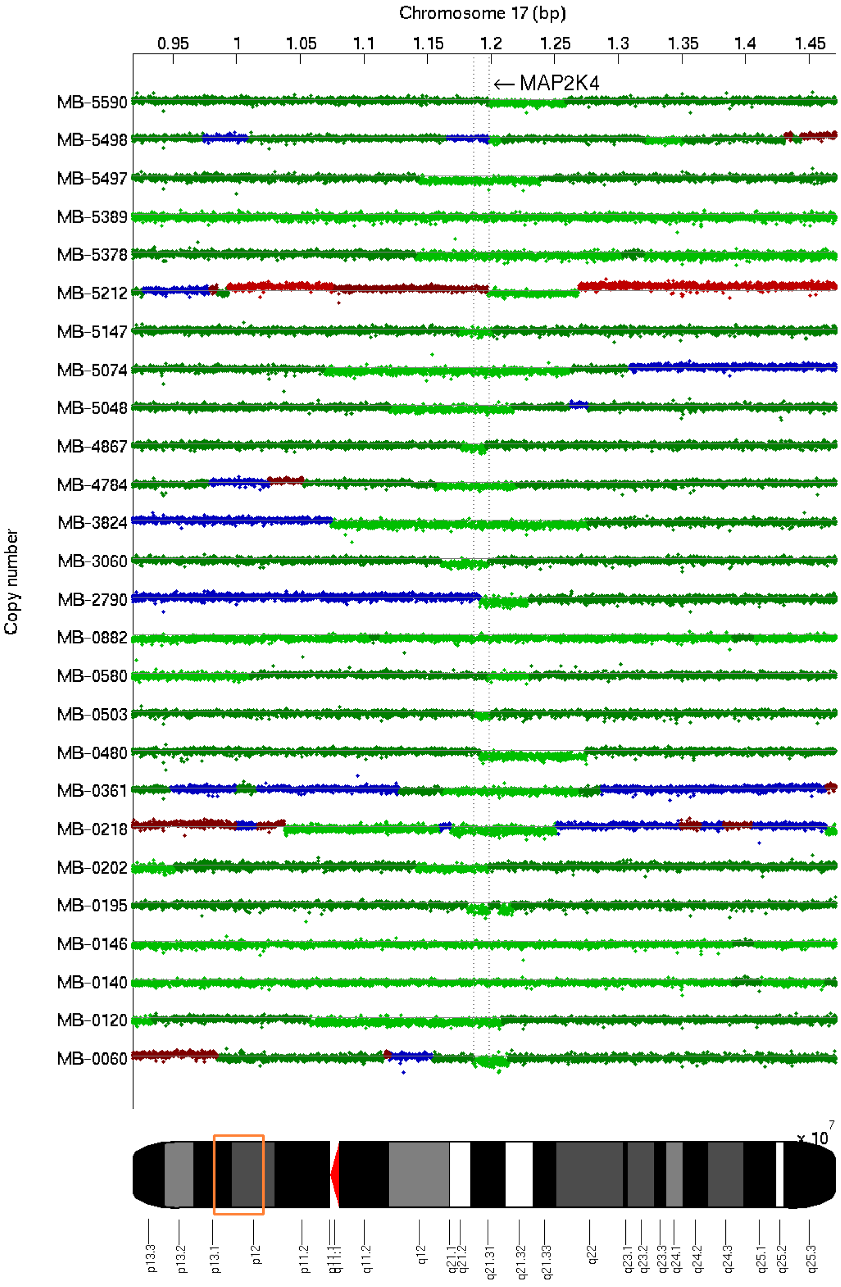
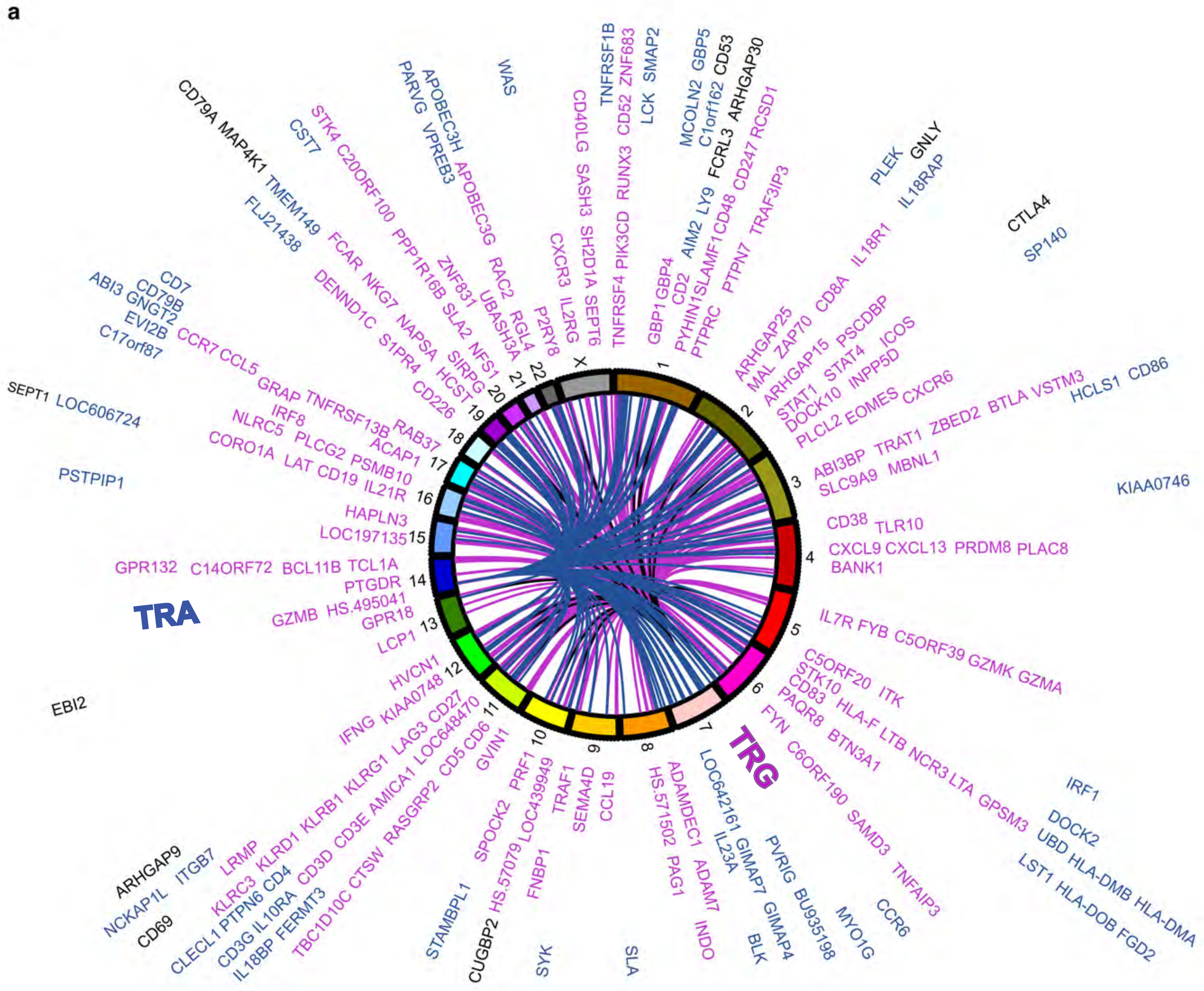
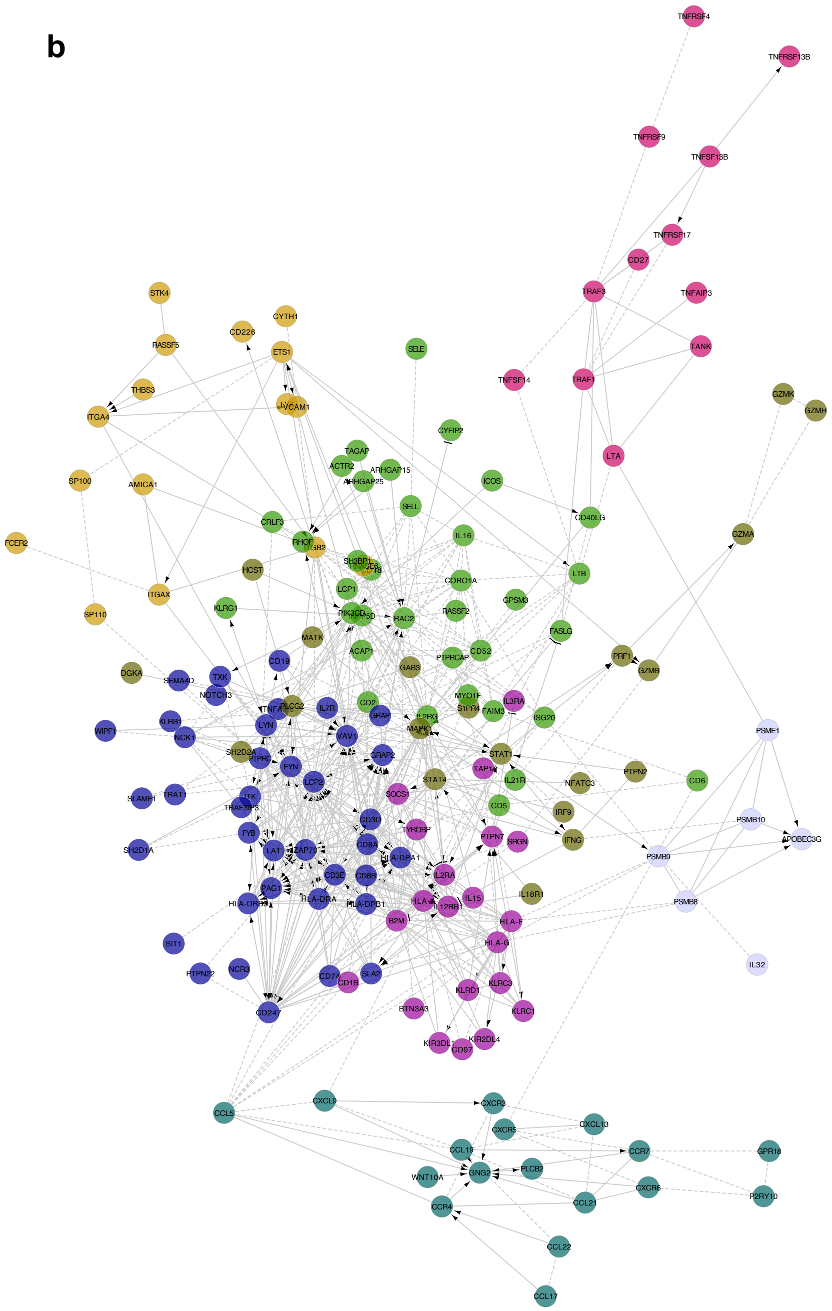


Figure S19: Putative *MAP2K4* homozygous deletions as predicted by HMM analysis. Here red indicates region of gain, blue indicates copy number neutral regions, and green indicates regions of loss, where more intense shading corresponds to higher amplitude events.


a
TRA
TRG
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
X
TNFRSF1B
CD52 ZNF683
LCK SMAP2
MCOLN2 GBP5
C1orf162 CD53
FCRL3 ARHGAP30
CD247 RCSD1
CD48
AIM2 LY9
TRAF3IP3
PTPN7
PYHIN1SLAMF1
CD2
GBP1 GBP4
PTPRC
PLEK GNLY
IL18RAP
CTLA4
SP140
IL18R1
CD8A
ARHGAP25
MAL ZAP70
ARHGAP15 PSCDBP
STAT1 STAT4 ICOS
DOCK10 INPP5D
PLCL2 EOMES CXCR6
HCLS1 CD86
ZBED2 BTLA VSTM3
ABI3BP TRAT1
SLC9A9 MBNL1
KIAA0746
CD38 TLR10
CXCL9 CXCL13 PRDM8 PLAC8
BANK1
IL7R FYB C5ORF39 GZMK GZMA
C5ORF20 ITK
STK10
CD83 HLA-F LTB NCR3 LTA GPSM3
PAQR8 BTN3A1
FYN
C6ORF190 SAMD3 TNFAIP3
IRF1
DOCK2
UBD HLA-DMB HLA-DMA
LST1 HLA-DOB FGD2
CCR6
MYO1G
PVRIG BU935198
BLK
LOC642161 GIMAP7 GIMAP4
IL23A
ADAMDEC1 ADAM7 INDO
HS.571502 PAG1
SLA
CCL19
SEMA4D
TRAF1
PRF1
LOC439949
FNBP1
SYK
SPOCK2
HS.57079
CUGBP2
STAMBPL1
GVIN1
CD5 CD6
LOC648470
AMICA1
RASGRP2
CTSW
TBC1D10C
LAG3 CD27
KIAA0748
IFNG
HVCN1
KLRB1 KLRG1
CD3E
CD3D
KLRD1
FERMT3
IL10RA
IL18BP
LRMP
KLRC3
PTPN6 CD4
CD3G
CLECL1
ITGB7
ARHGAP9
NCKAP1L
CD69
EBI2
LCP1
GPR18
GZMB HS.495041
PTGDR
GPR132 C14ORF72 BCL11B TCL1A
LOC197135
HAPLN3
PSTPIP1
CORO1A LAT CD19 IL21R
NLRC5 PLCG2 PSMB10
IRF8
ACAP1
GRAP TNFRSF13B
RAB37
SEPT1 LOC606724
CCR7 CCL5
CD226
HCST
S1PR4
DENND1C
C17orf87
EVI2B
ABI3 GNGT2
CD79B
CD7
FLJ21438
TMEM149
FCAR NKG7
NAPSA
SIRPG
NFS1
CD79A MAP4K1
CST7
STK4 C20ORF100
PPP1R16B SLA2
ZNF831
UBASH3A
APOBEC3H
APOBEC3G
PARVG VPREB3
RAC2 RGL4
P2RY8
WAS
CD40LG SASH3 SH2D1A SEPT6
CXCR3 IL2RG
RUNX3
TNFRSF4 PIK3CD



Module	GeneSet
0	TCR signaling(R)
0	TCR signaling in naive CD4+ T cells(N)
0	T cell receptor signaling pathway(K)
0	TCR signaling in naive CD8+ T cells(N)
0	T cell activation(P)
0	Primary immunodeficiency(K)
0	Hematopoietic cell lineage(K)
0	Fc-epsilon receptor I signaling in mast cells(N)
0	Antigen processing and presentation(K)
0	t cell receptor signaling pathway(B)
0	role of mef2d in t-cell apoptosis(B)
0	Natural killer cell mediated cytotoxicity(K)
0	lck and fyn tyrosine kinases in initiation of tcr activation(B)
0	Class I PI3K signaling events(N)
0	activation of csk by camp-dependent protein kinase inhibits signaling through the t cell receptor(B)
0	Cell adhesion molecules (CAMs)(K)
0	the co-stimulatory signal during t-cell activation(B)
0	TRAIL signaling pathway(N)
0	IL12-mediated signaling events(N)
0	IL12 signaling mediated by STAT4(N)
0	Downstream signaling in naive CD8+ T cells(N)
0	Immunoregulatory interactions between a Lymphoid and a non-Lymphoid cell(R)
0	Asthma(K)
0	Viral myocarditis(K)
0	Fc epsilon RI signaling pathway(K)
0	Allograft rejection(K)
0	Graft-versus-host disease(K)
0	Type I diabetes mellitus(K)
0	Intestinal immune network for IgA production(K)
2	Antigen processing and presentation(K)
2	Immunoregulatory interactions between a Lymphoid and a non-Lymphoid cell(R)
2	ras-independent pathway in nk cell-mediated cytotoxicity(B)
2	Natural killer cell mediated cytotoxicity(K)
2	Graft-versus-host disease(K)
2	IL12-mediated signaling events(N)
2	Downstream signaling in naive CD8+ T cells(N)
2	Jak-STAT signaling pathway(K)
3	IL12-mediated signaling events(N)
3	Downstream signaling in naive CD8+ T cells(N)
3	Natural killer cell mediated cytotoxicity(K)
3	TCR signaling in naive CD8+ T cells(N)
3	IL12 signaling mediated by STAT4(N)
3	TNF receptor signaling pathway(N)
3	EGF receptor signaling pathway(P)
3	granzyme a mediated apoptosis pathway(B)
3	il12 and stat4 dependent signaling pathway in th1 development(B)
3	TRAIL signaling pathway(N)
3	Inflammation mediated by chemokine and cytokine signaling pathway(P)
3	IFN-gamma pathway(N)
3	IL23-mediated signaling events(N)
3	VEGF signaling pathway(K)
3	IL27-mediated signaling events(N)
4	Chemokine signaling pathway(K)
4	Cytokine-cytokine receptor interaction(K)
4	Receptor-ligand complexes bind G proteins(R)
4	Class A/1 (Rhodopsin-like receptors)(R)
4	Inflammation mediated by chemokine and cytokine signaling pathway(P)
5	Integrin cell surface interactions(R)
5	Immunoregulatory interactions between a Lymphoid and a non-Lymphoid cell(R)
5	Cell surface interactions at the vascular wall(R)
6	Cytokine-cytokine receptor interaction(K)
7	Host Interactions of HIV factors(R)
7	Proteasome(K)
7	Signaling by Wnt(R)
7	M/G1 Transition(R)
7	APC/C-mediated degradation of cell cycle proteins(R)
7	Metabolism of amino acids(R)
7	DNA Replication(R)
7	G1/S Transition(R)
7	S Phase(R)
7	Cell Cycle Checkpoints(R)
7	Apoptosis(R)

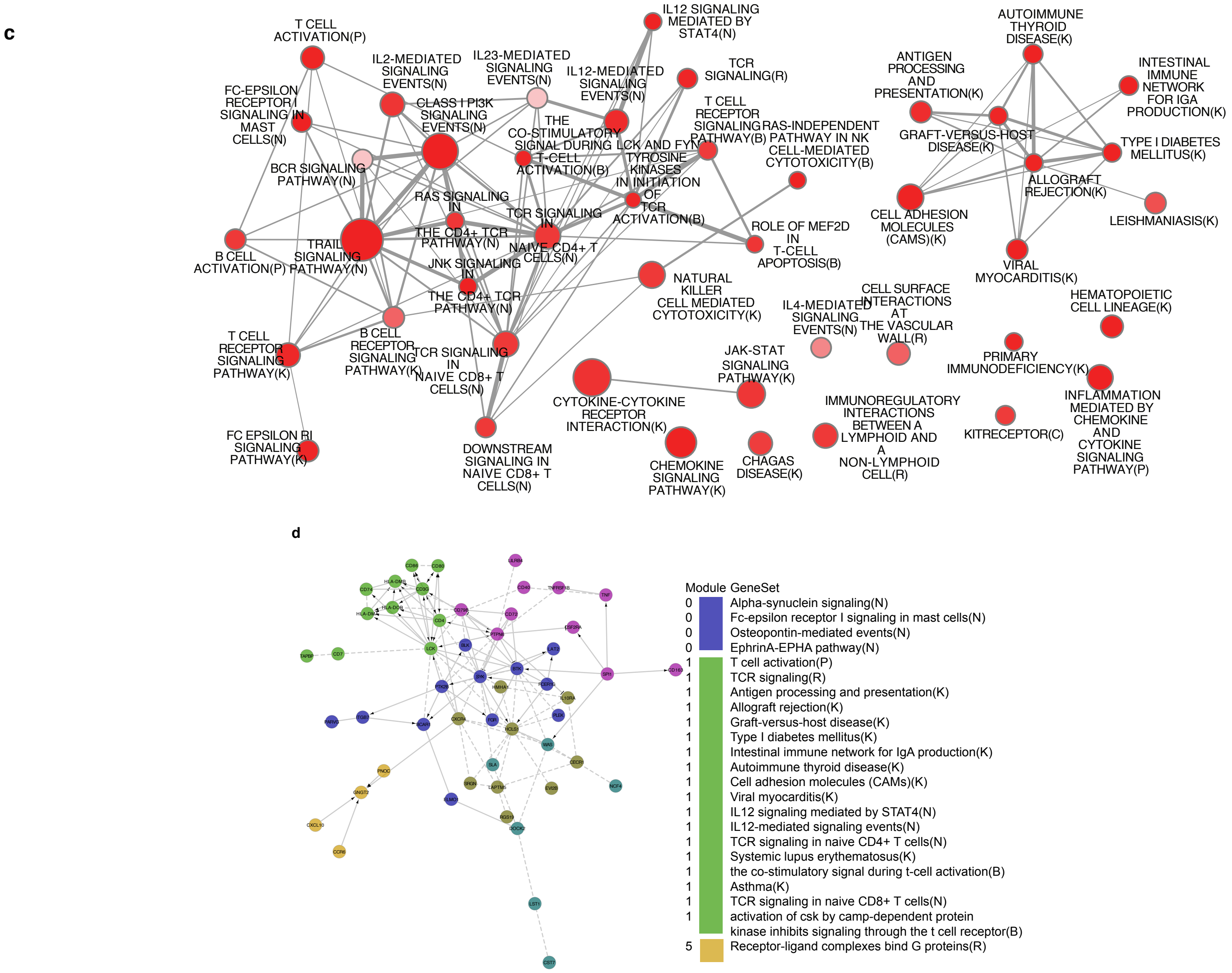
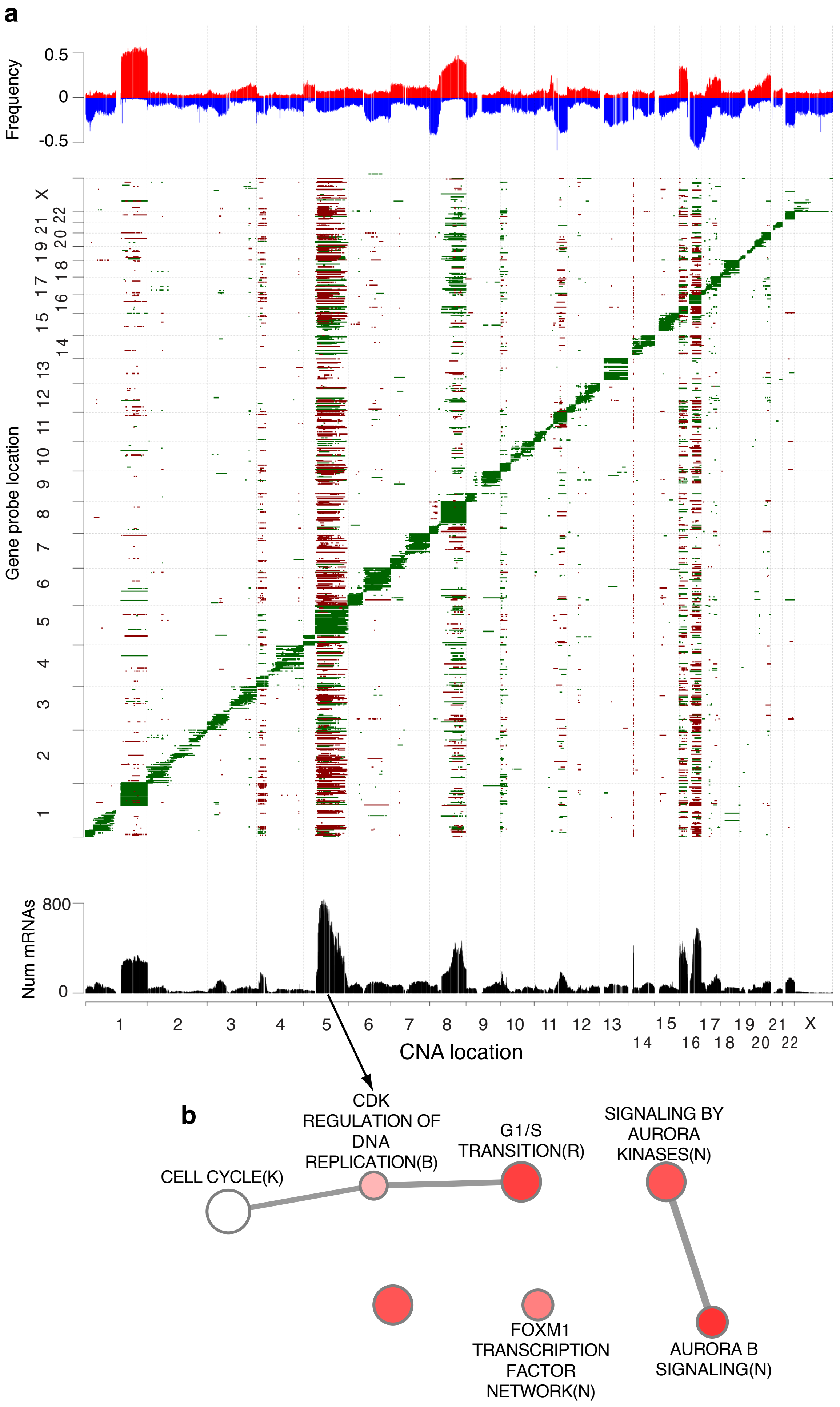
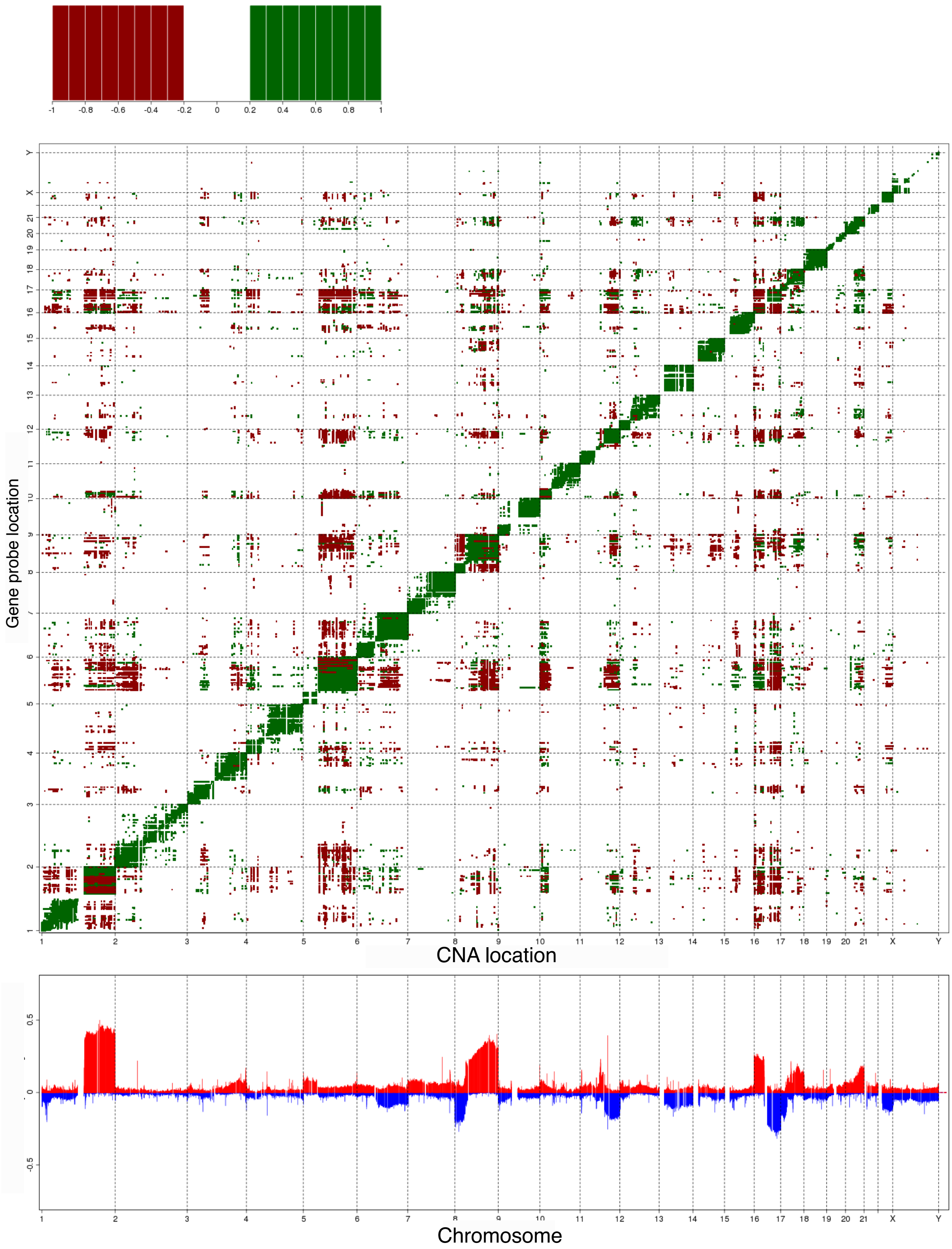

c
T CELL ACTIVATION(P)
IL2-MEDIATED SIGNALING EVENTS(N)
IL23-MEDIATED SIGNALING EVENTS(N)
IL12 SIGNALING MEDIATED BY STAT4(N)
TCR SIGNALING(R)
IL12-MEDIATED SIGNALING EVENTS(N)
FC-EPSILON RECEPTOR I SIGNALING IN MAST CELLS(N)
CLASS I PI3K SIGNALING EVENTS(N)
THE CO-STIMULATORY SIGNAL DURING T-CELL ACTIVATION(B)
LCK AND FYN TYROSINE KINASES IN INITIATION OF TCR ACTIVATION(B)
T CELL RECEPTOR SIGNALING PATHWAY(B)
RAS-INDEPENDENT PATHWAY IN NK CELL-MEDIATED CYTOTOXICITY(B)
BCR SIGNALING PATHWAY(N)
RAS SIGNALING IN THE CD4+ TCR PATHWAY(N)
TCR SIGNALING IN NAIVE CD4+ T CELLS(N)
ROLE OF MEF2D IN T-CELL APOPTOSIS(B)
B CELL ACTIVATION(P)
TRAIL SIGNALING PATHWAY(N)
JNK SIGNALING IN THE CD4+ TCR PATHWAY(N)
NATURAL KILLER CELL MEDIATED CYTOTOXICITY(K)
T CELL RECEPTOR SIGNALING PATHWAY(K)
B CELL RECEPTOR SIGNALING PATHWAY(K)
TCR SIGNALING IN NAIVE CD8+ T CELLS(N)
CYTOKINE-CYTOKINE RECEPTOR INTERACTION(K)
JAK-STAT SIGNALING PATHWAY(K)
IL4-MEDIATED SIGNALING EVENTS(N)
CELL SURFACE INTERACTIONS AT THE VASCULAR WALL(R)
FC EPSILON RI SIGNALING PATHWAY(K)
DOWNSTREAM SIGNALING IN NAIVE CD8+ T CELLS(N)
CHEMOKINE SIGNALING PATHWAY(K)
CHAGAS DISEASE(K)
IMMUNOREGULATORY INTERACTIONS BETWEEN A LYMPHOID AND A NON-LYMPHOID CELL(R)
AUTOIMMUNE THYROID DISEASE(K)
ANTIGEN PROCESSING AND PRESENTATION(K)
INTESTINAL IMMUNE NETWORK FOR IGA PRODUCTION(K)
GRAFT-VERSUS-HOST DISEASE(K)
TYPE I DIABETES MELLITUS(K)
ALLOGRAFT REJECTION(K)
CELL ADHESION MOLECULES (CAMS)(K)
LEISHMANIASIS(K)
VIRAL MYOCARDITIS(K)
HEMATOPOIETIC CELL LINEAGE(K)
PRIMARY IMMUNODEFICIENCY(K)
INFLAMMATION MEDIATED BY CHEMOKINE AND CYTOKINE SIGNALING PATHWAY(P)
KITRECEPTOR(C)
d
LILRB4
CD86
CD80
HLA-DMB
CD74
CD3G
CD40
TNFRSF1B
TNF
HLA-DOB
HLA-DMA
CD79B
CD72
CD4
CSF2RA
PTPN6
TAPBP
CD7
LCK
BLK
LAT2
BTK
SYK
SPI1
CD163
PTK2B
HMHA1
FCER1G
IL10RA
PARVG
ITGB7
BCAR1
CXCR4
FGR
PLEK
HCLS1
WAS
SLA
CECR1
PNOC
GNGT2
SRGN
LAPTM5
EVI2B
NCF4
ELMO1
RGS19
DOCK2
CXCL10
CCR6
LST1
CST7
Module GeneSet
0 Alpha-synuclein signaling(N)
0 Fc-epsilon receptor I signaling in mast cells(N)
0 Osteopontin-mediated events(N)
0 EphrinA-EPHA pathway(N)
1 T cell activation(P)
1 TCR signaling(R)
1 Antigen processing and presentation(K)
1 Allograft rejection(K)
1 Graft-versus-host disease(K)
1 Type I diabetes mellitus(K)
1 Intestinal immune network for IgA production(K)
1 Autoimmune thyroid disease(K)
1 Cell adhesion molecules (CAMs)(K)
1 Viral myocarditis(K)
1 IL12 signaling mediated by STAT4(N)
1 IL12-mediated signaling events(N)
1 TCR signaling in naive CD4+ T cells(N)
1 Systemic lupus erythematosus(K)
1 the co-stimulatory signal during t-cell activation(B)
1 Asthma(K)
1 TCR signaling in naive CD8+ T cells(N)
1 activation of csk by camp-dependent protein kinase inhibits signaling through the t cell receptor(B)
5 Receptor-ligand complexes bind G proteins(R)

Figure S20: The T-cell receptor loci on chromosomes 7 and 14 represent *trans*-acting aberration 'hotspots' that modulate an immune response network. **a.** Circos plot illustrating the connectivity between *trans*-acting deletion hotspots at the TCR loci on chromosomes 7 and 14 with their cognate mRNAs. Genes that are significantly associated (P-value $< 10^{-4}$) with deletion of the *TRG* locus are indicated by purple links, whereas those associated with deletion of *TRA* are indicated by blue links, and those genes dually regulated by both loci are shown in black. **b.** Network representation of mRNAs correlated *TRG* deletion. Coloured subnetworks represent clustering of the network into modules based on an edge-betweenness algorithm (see Methods for additional details). Pathways enriched (FDR < 0.001) for genes clustered in each module are shown on the right. **c.** Enrichment maps of immune response modules in the *trans*-associated TRG network. **d.** As in (b), but for the *TRA* locus.



Chr 5, Basal cancer specific

c



-1
-0.8
-0.6
-0.4
-0.2
0
0.2
0.4
0.6
0.8
1
Gene probe location
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
X
Y
CNA location
0.5
0
-0.5
Chromosome

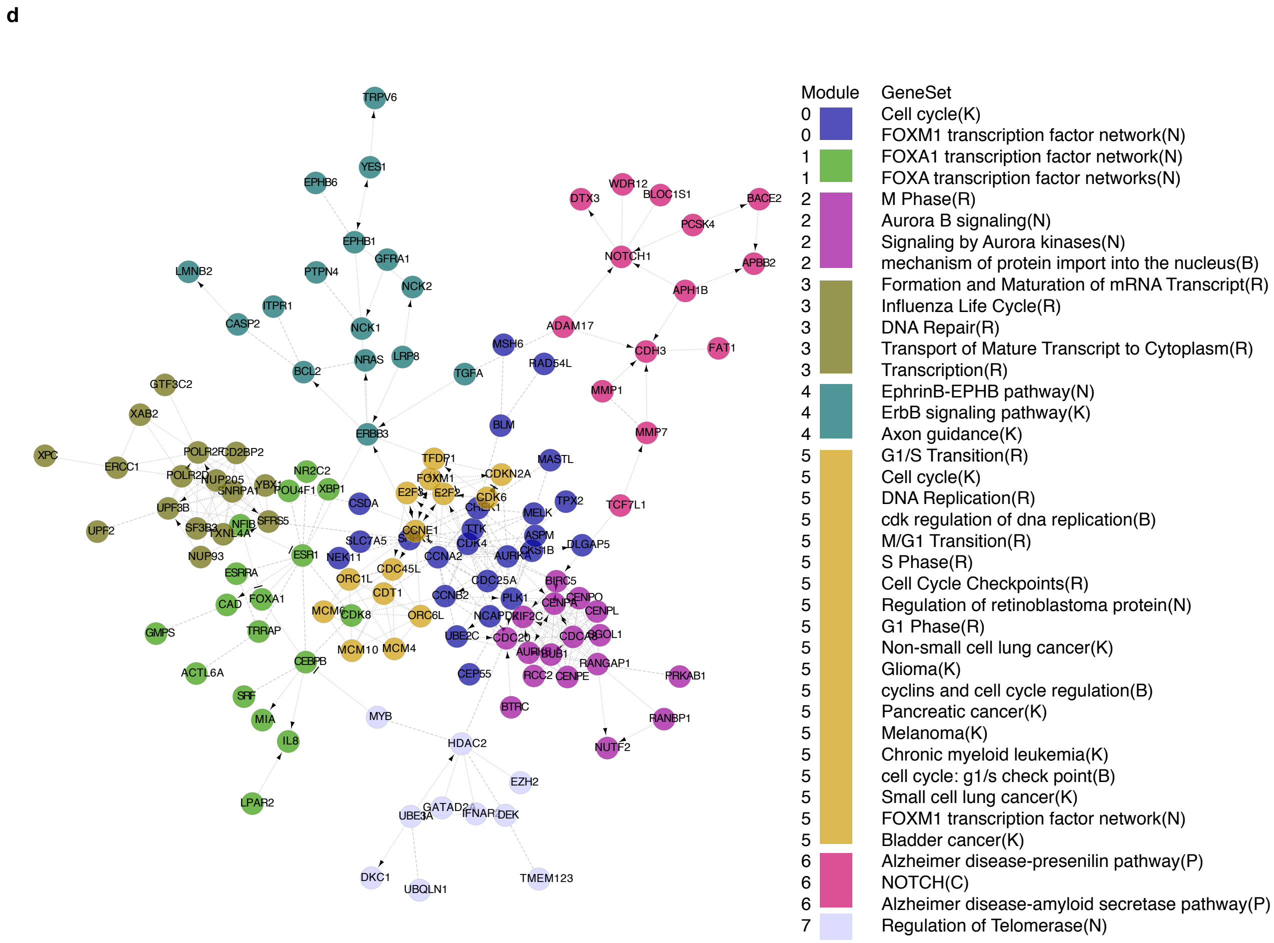


Figure S21: Copy number-expression correlation matrices for CBS and HMM, and the chromosome 5q deletion network module. **a.** Frequency of genomic copy number aberrations (HMM-derived) (upper panel) and corresponding matrix of copy number and gene expression correlations > 0.3 (middle panel). The frequency of mRNAs associated with a particular copy number aberration is also indicated (lower panel). **b.** Enrichment maps of cell-cycle and DNA damage response modules for the chromosome 5 deletion *trans*-associated genes. **c.** The matrix of copy number (CBS-derived) and gene expression profiles with correlation > 0.2 (upper panel) and frequency of copy number events across the entire cohort (lower panel). **d.** Network representation of mRNAs correlated with deletion of chromosome 5q. Coloured subnetworks represent clustering of the network into modules based on an edge-betweenness algorithm (see Methods for additional details). Pathways enriched (FDR < 0.001) for genes clustered in each module are shown on the right.

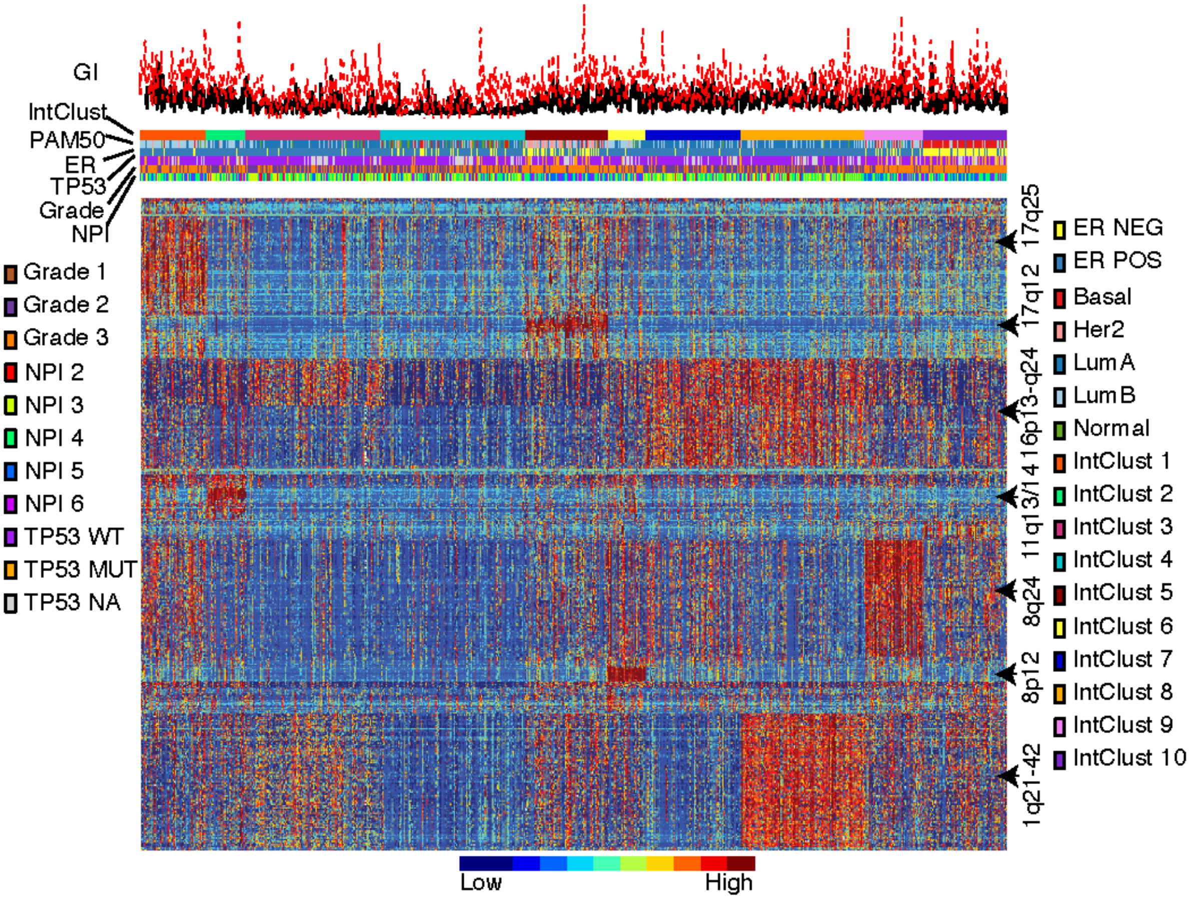
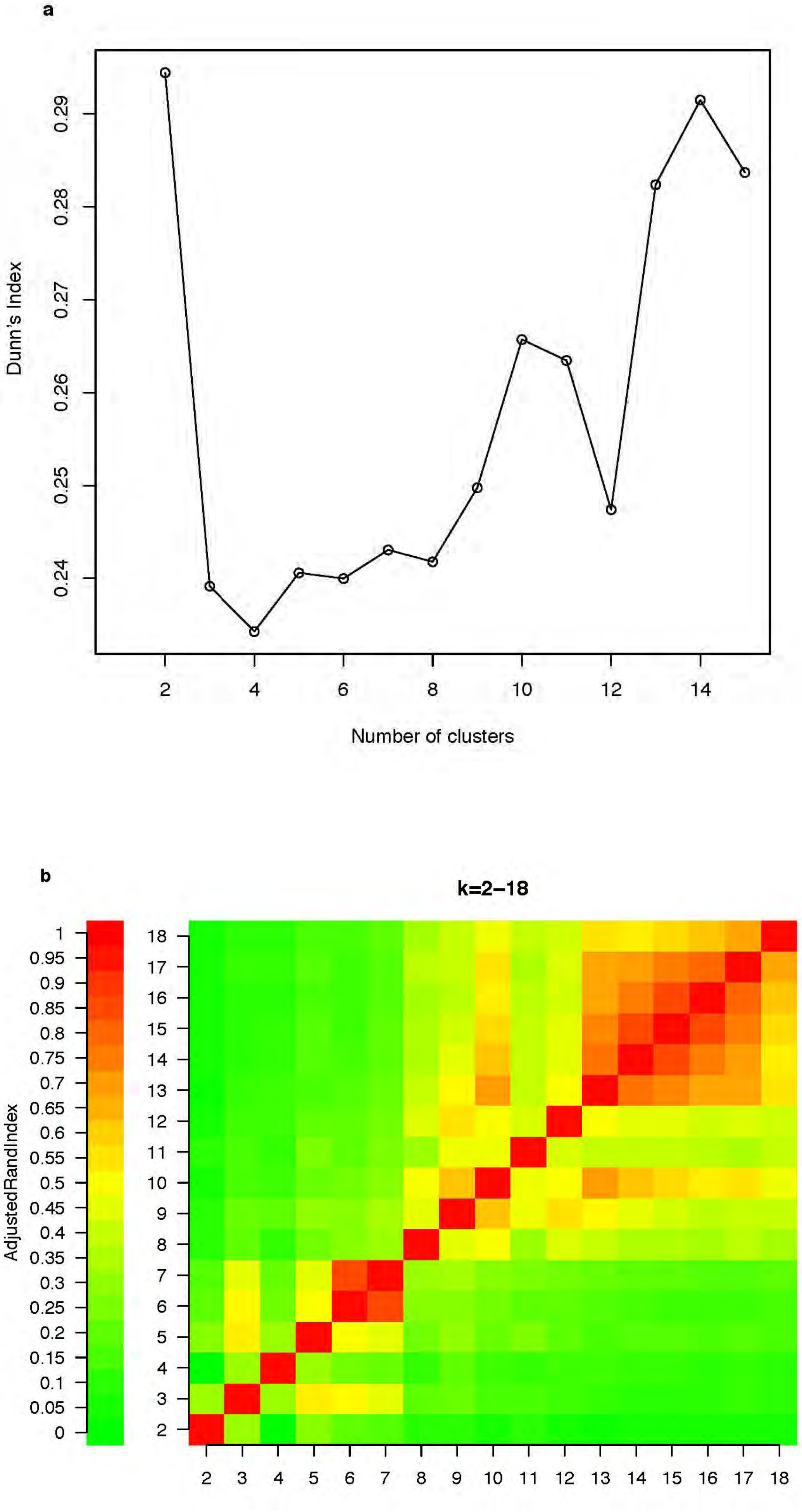


Figure S22: Joint clustering of copy number and expression data for the top 1,000 *cis*-acting copy number-expression associations for CBS-derived CNAs in the discovery cohort (n = 997). Heatmap representation of the joint data matrix for the k =10 clustering, where the product of scaled gene expression and copy number values are shown for the union set of selected features. Genomic instability (GI) based on the proportion of genome altered (black line) and jump measure (red line) is indicated. NPI refers to Nottingham prognostic index.


a
Dunn's Index
0.29
0.28
0.27
0.26
0.25
0.24
2
4
6
8
10
12
14
Number of clusters
b
k=2−18
AdjustedRandIndex
1
0.95
0.9
0.85
0.8
0.75
0.7
0.65
0.6
0.55
0.5
0.45
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
18
17
16
15
14
13
12
11
10
9
8
7
6
5
4
3
2
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18

c

		IntClust k9								
		1	2	3	4	5	6	7	8	9
IntClust k10	1	0	0	0	1	5	0	68	2	0
	2	0	0	0	1	0	0	0	44	0
	3	1	0	80	0	19	48	2	0	6
	4	0	0	0	0	3	162	0	0	2
	5	0	1	0	90	1	2	0	0	0
	6	42	2	0	0	0	0	0	0	0
	7	0	0	1	0	0	2	0	0	106
	8	0	0	127	0	13	0	1	1	1
	9	1	57	0	1	3	0	1	1	1
	10	0	12	4	0	69	9	2	0	0

		IntClust k13												
		1	2	3	4	5	6	7	8	9	10	11	12	13
IntClust k10	1	0	0	0	1	0	3	1	0	7	0	3	2	59
	2	0	0	1	0	0	0	0	0	0	0	44	0	0
	3	0	0	130	0	1	2	6	0	13	4	0	0	0
	4	0	2	2	0	114	30	19	0	0	0	0	0	0
	5	1	0	0	84	0	7	1	0	1	0	0	0	0
	6	0	0	0	0	0	0	0	41	1	0	0	2	0
	7	0	99	1	0	0	1	4	0	0	3	0	0	1
	8	0	0	0	0	0	0	0	0	24	110	0	0	0
	9	4	0	1	0	0	2	2	0	4	0	1	48	5
	10	44	0	3	0	0	41	1	0	6	0	0	1	0

		IntClust k14													
		1	2	3	4	5	6	7	8	9	10	11	12	13	14
IntClust k10	1	1	2	3	0	3	0	0	0	0	0	0	59	2	6
	2	0	43	0	0	0	0	0	1	1	0	0	0	0	0
	3	0	0	1	5	0	5	0	0	125	1	0	0	8	11
	4	0	0	23	0	0	0	0	1	4	1	117	0	21	0
	5	84	0	6	0	0	0	2	0	0	0	0	0	1	1
	6	0	1	0	0	10	1	0	31	0	0	0	0	0	1
	7	0	0	2	0	4	5	0	0	2	88	1	1	6	0
	8	0	0	0	95	0	27	0	0	1	1	0	0	0	19
	9	1	0	6	0	24	21	7	0	0	0	0	3	3	2
	10	0	0	46	0	0	1	46	0	0	0	0	0	0	3

Figure S23: Internal evaluation of cluster validity. **a.** Dunn's index was computed for the complete set of features used as input for integrative clustering using the Euclidean distance metric. Evaluation of the indices for $k = 2, \ldots, 15$ indicated a global maximum at $k = 14$ (ignoring the $k = 2$ peak) and a local maximum at $k = 10$ **b.** The adjusted Rand index (ARI) was computed for all pairwise comparisons of $k = 2, \ldots, 18$ as illustrated in this heatmap, where the value of the ARI is as indicated in the legend. **c.** Cross tabulation matrices for comparison of $k = 9$, $k = 13$, and $k = 14$ with k = 10.

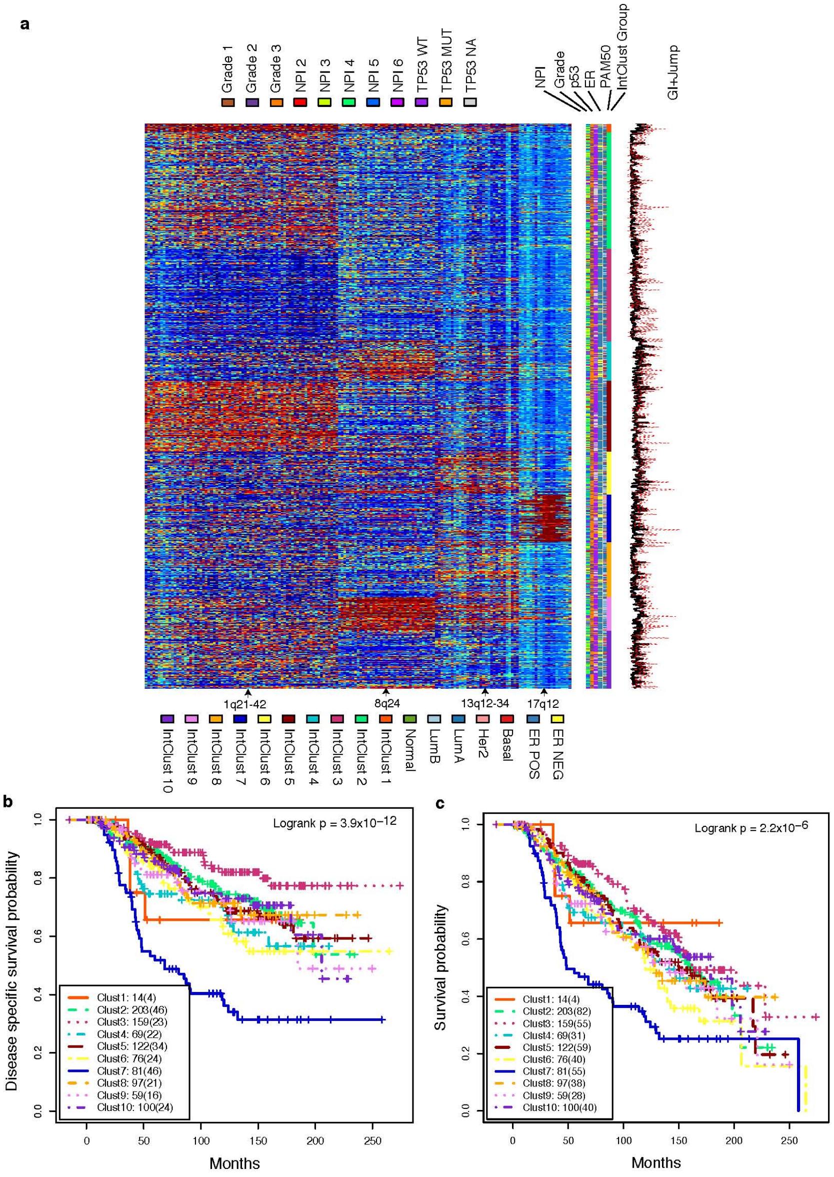


Figure S24: Joint clustering of copy number and expression data for the top 1,000 *cis*-acting copy number-expression associations for HMM-derived CNAs in the discovery cohort (*n* = 997). **a.** Heatmap representation of the joint data matrix of the union set of selected gene expression and copy number features for the *k* =10 clustering. **b.** Kaplan-Meier plots of breast cancer specific survival for the integrative cluster subgroups. **c**. Kaplan-Meier plots of overall survival for the integrative cluster subgroups.

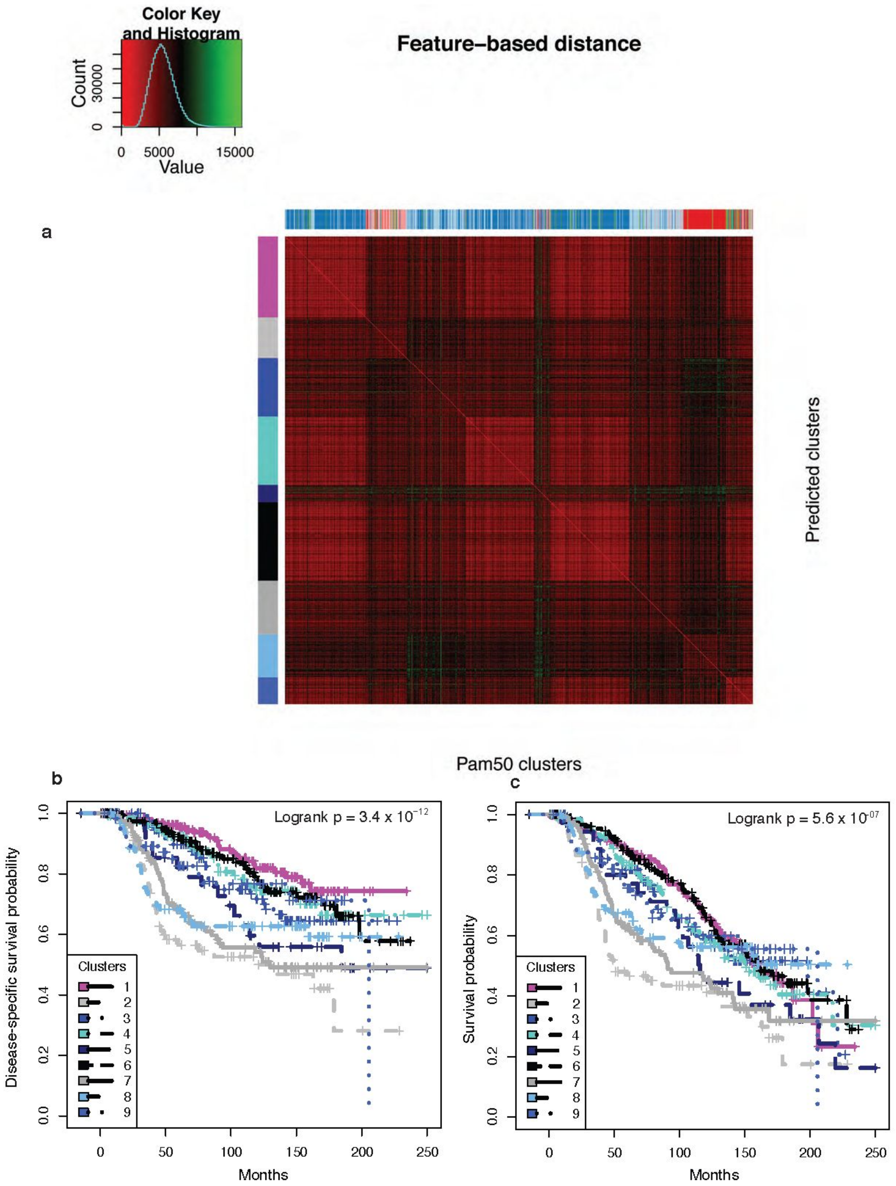
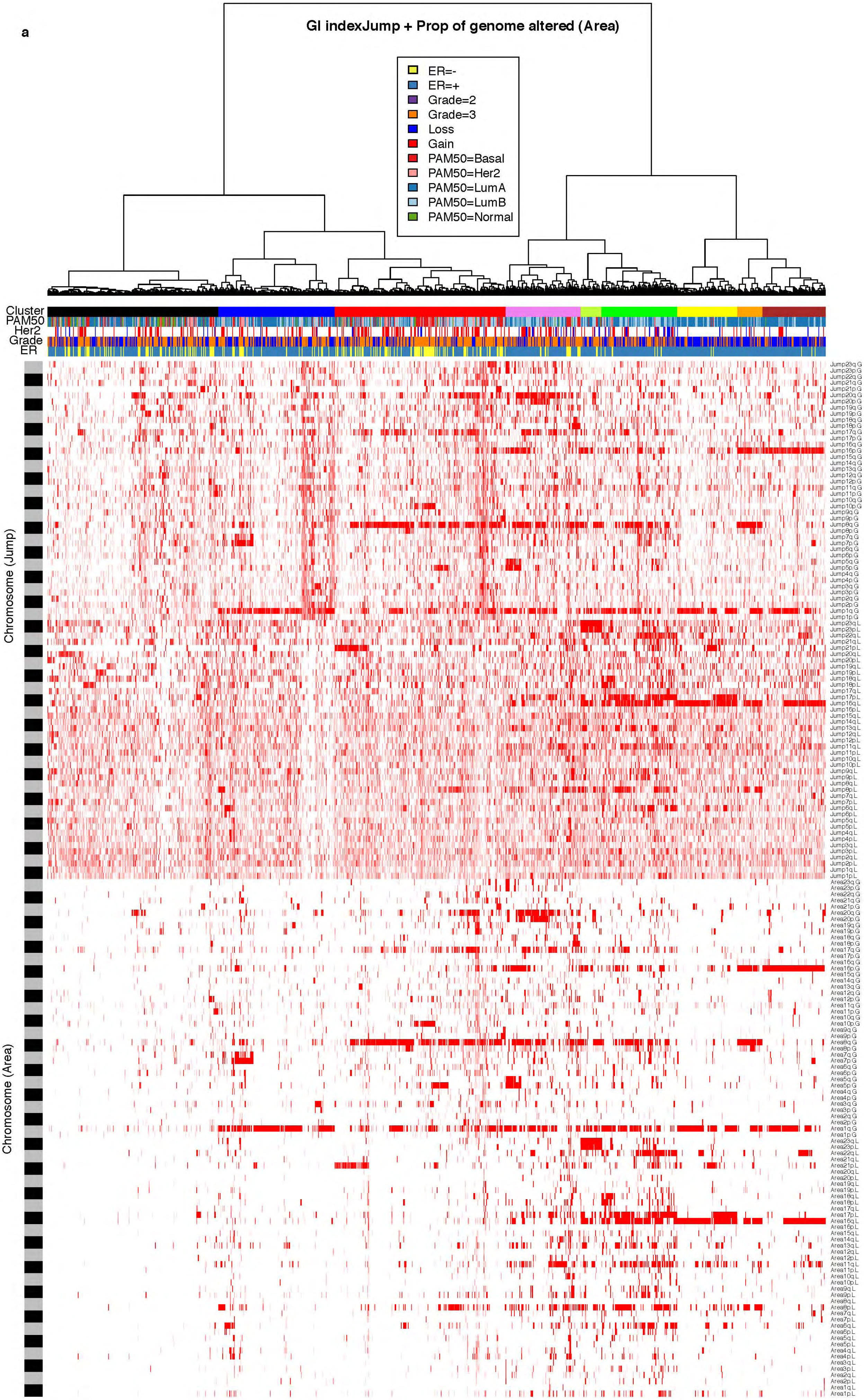


Figure S25: K-Means clustering on outlier gene expression profiles. **a.** Heatmap representation of K-Means clustering (k = 9) on outlier expression profiles. **b.** Kaplan-Meier plots of disease-specific survival for the K-Means clusters. **c**. Kaplan-Meier plots of overall survival for the K-Means clusters.


a
GI indexJump + Prop of genome altered (Area)
ER=-
ER=+
Grade=2
Grade=3
Loss
Gain
PAM50=Basal
PAM50=Her2
PAM50=LumA
PAM50=LumB
PAM50=Normal
Cluster
PAM50
Her2
Grade
ER
Chromosome (Jump)
Chromosome (Area)
Jump23q.G
Jump23p.G
Jump22q.G
Jump21q.G
Jump21p.G
Jump20q.G
Jump20p.G
Jump19q.G
Jump19p.G
Jump18q.G
Jump18p.G
Jump17q.G
Jump17p.G
Jump16q.G
Jump16p.G
Jump15q.G
Jump14q.G
Jump13q.G
Jump12q.G
Jump12p.G
Jump11q.G
Jump11p.G
Jump10q.G
Jump10p.G
Jump9q.G
Jump9p.G
Jump8q.G
Jump8p.G
Jump7q.G
Jump7p.G
Jump6q.G
Jump6p.G
Jump5q.G
Jump5p.G
Jump4q.G
Jump4p.G
Jump3q.G
Jump3p.G
Jump2q.G
Jump2p.G
Jump1q.G
Jump1p.G
Jump23q.L
Jump23p.L
Jump22q.L
Jump21q.L
Jump21p.L
Jump20q.L
Jump20p.L
Jump19q.L
Jump19p.L
Jump18q.L
Jump18p.L
Jump17q.L
Jump17p.L
Jump16q.L
Jump16p.L
Jump15q.L
Jump14q.L
Jump13q.L
Jump12q.L
Jump12p.L
Jump11q.L
Jump11p.L
Jump10q.L
Jump10p.L
Jump9q.L
Jump9p.L
Jump8q.L
Jump8p.L
Jump7q.L
Jump7p.L
Jump6q.L
Jump6p.L
Jump5q.L
Jump5p.L
Jump4q.L
Jump4p.L
Jump3q.L
Jump3p.L
Jump2q.L
Jump2p.L
Jump1q.L
Jump1p.L
Area23q.G
Area23p.G
Area22q.G
Area21q.G
Area21p.G
Area20q.G
Area20p.G
Area19q.G
Area19p.G
Area18q.G
Area18p.G
Area17q.G
Area17p.G
Area16q.G
Area16p.G
Area15q.G
Area14q.G
Area13q.G
Area12q.G
Area12p.G
Area11q.G
Area11p.G
Area10q.G
Area10p.G
Area9q.G
Area9p.G
Area8q.G
Area8p.G
Area7q.G
Area7p.G
Area6q.G
Area6p.G
Area5q.G
Area5p.G
Area4q.G
Area4p.G
Area3q.G
Area3p.G
Area2q.G
Area2p.G
Area1q.G
Area1p.G
Area23q.L
Area23p.L
Area22q.L
Area21q.L
Area21p.L
Area20q.L
Area20p.L
Area19q.L
Area19p.L
Area18q.L
Area18p.L
Area17q.L
Area17p.L
Area16q.L
Area16p.L
Area15q.L
Area14q.L
Area13q.L
Area12q.L
Area12p.L
Area11q.L
Area11p.L
Area10q.L
Area10p.L
Area9q.L
Area9p.L
Area8q.L
Area8p.L
Area7q.L
Area7p.L
Area6q.L
Area6p.L
Area5q.L
Area5p.L
Area4q.L
Area4p.L
Area3q.L
Area3p.L
Area2q.L
Area2p.L
Area1q.L
Area1p.L

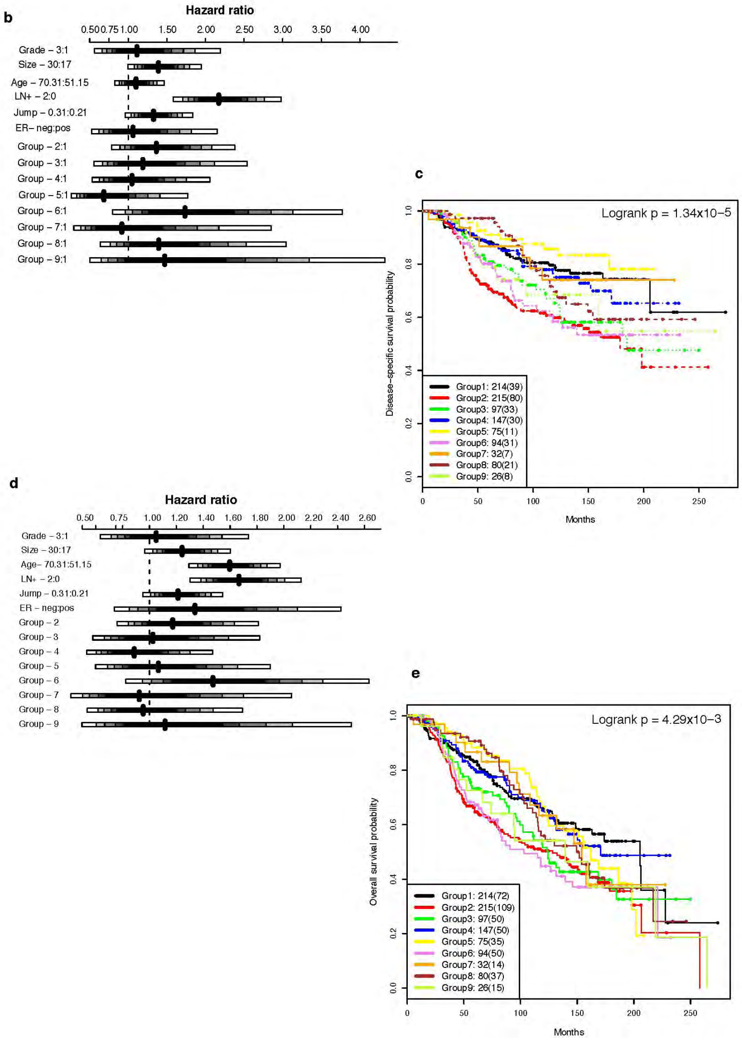
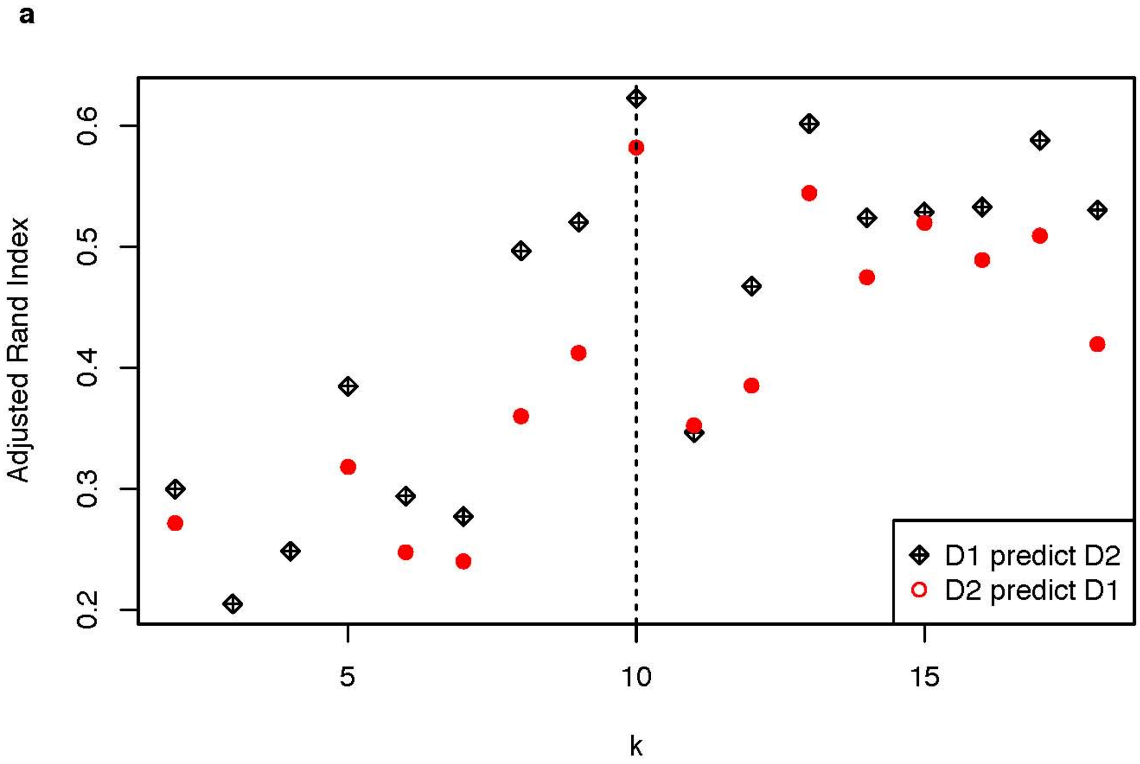
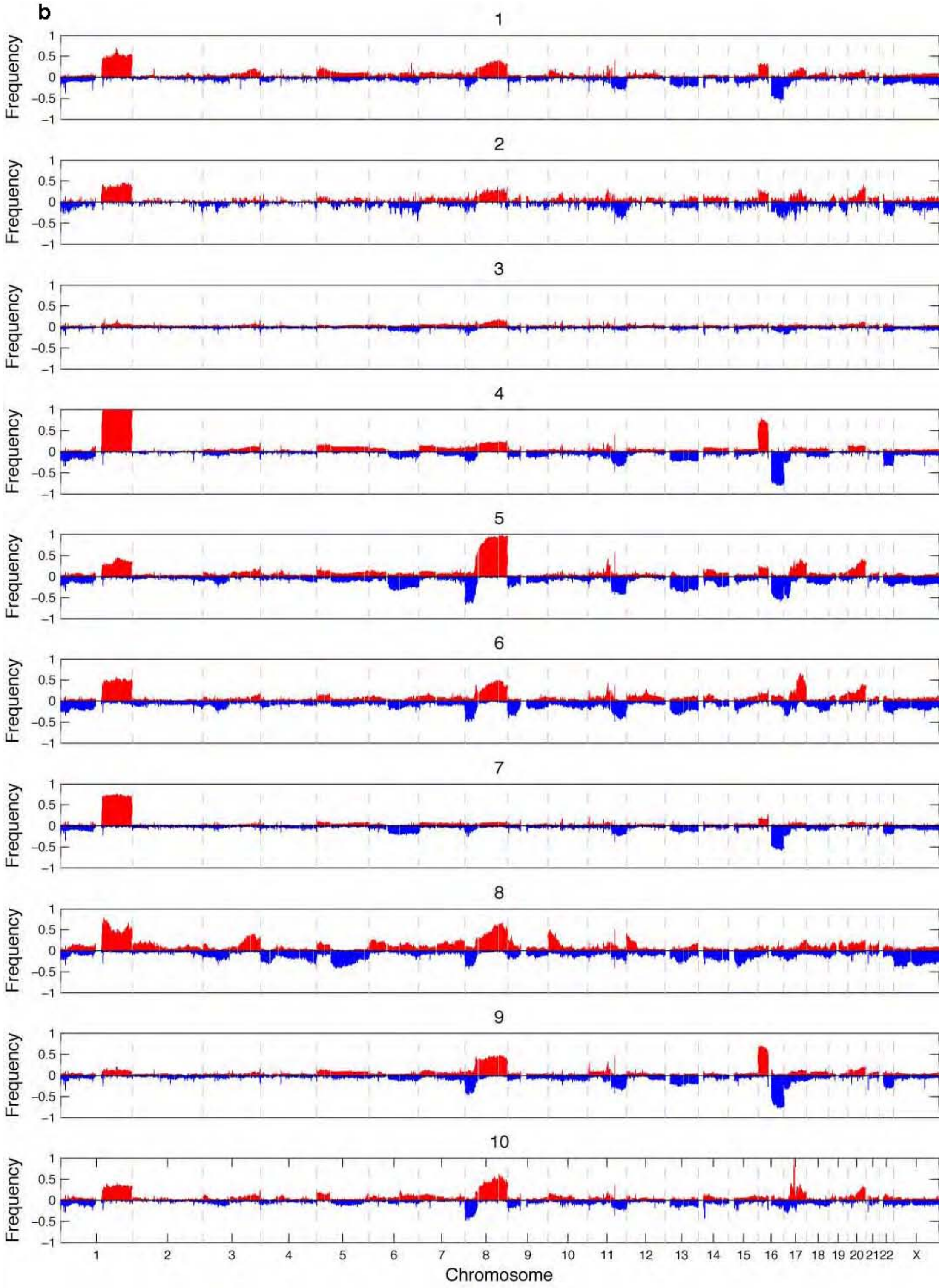
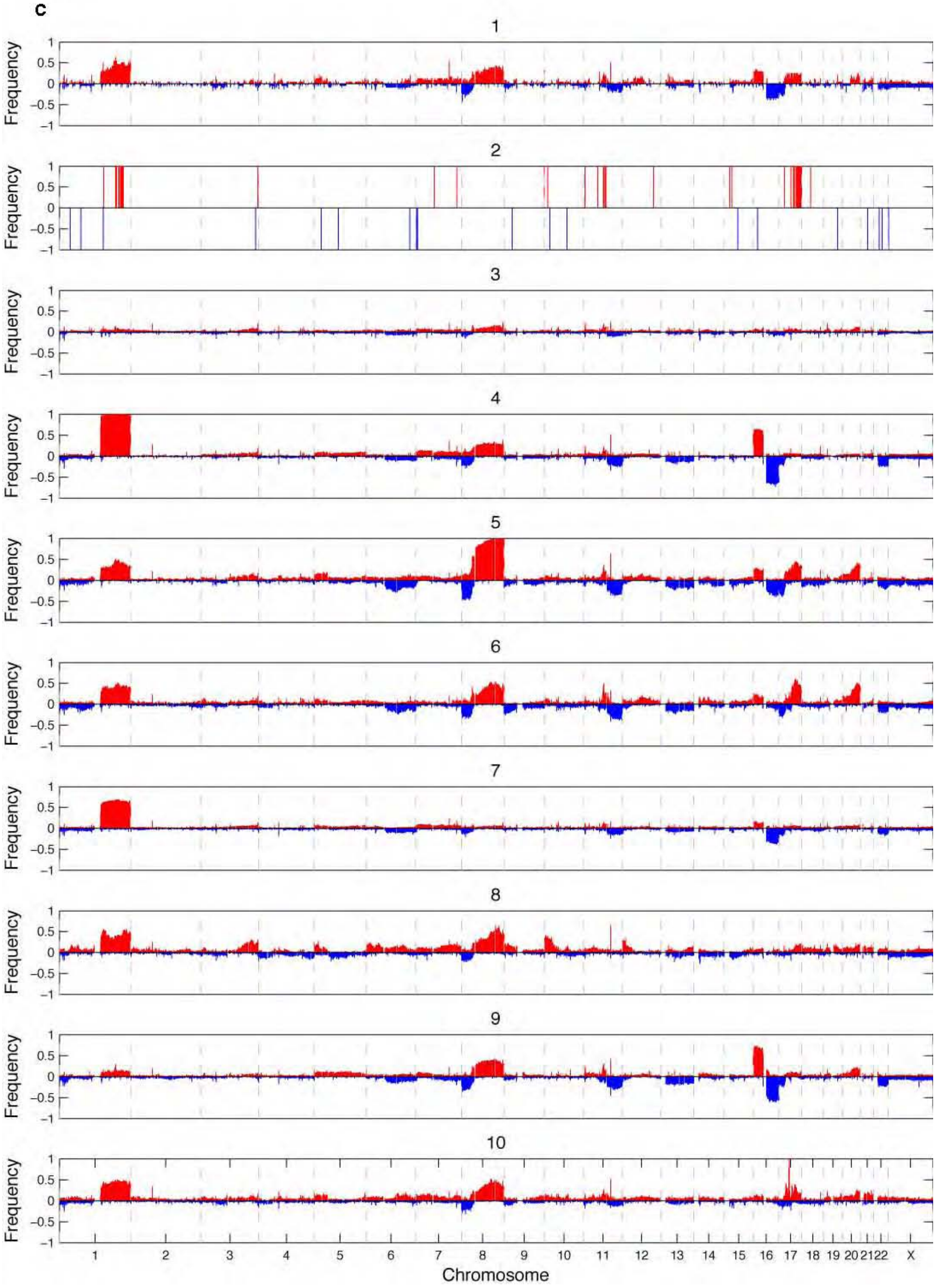
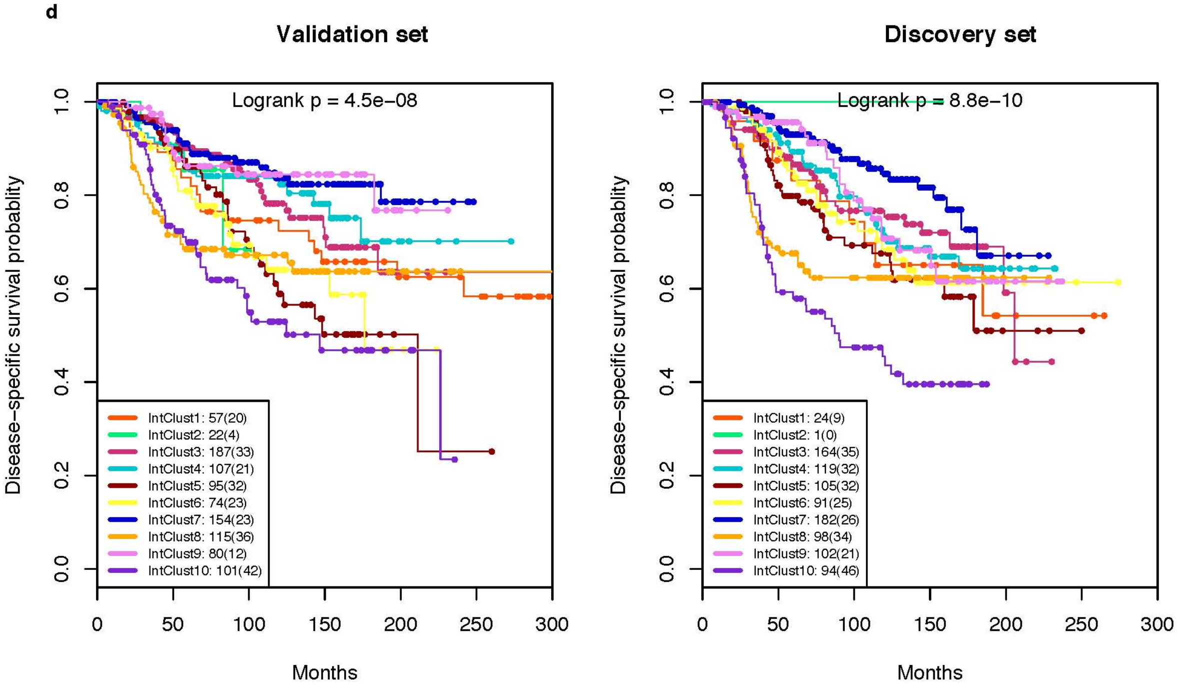


Figure S26: Unsupervised clustering of GI based on two metrics, namely the proportion of genome altered (area), and the ratio between the mean change in log ratio due to alterations versus relative variability (jump) reveals 9 clusters, some of which associate with outcome. **a.** Unsupervised clustering on GI (derived from CBS) using the Euclidean distance metric and Ward's method indicates 9 clusters as illustrated in this heatmap representation of the data. **b.** Graphical representation of the Cox proportional hazard ratios for disease-specific survival for several key variables including, grade, tumour size, age at diagnosis, number of lymph nodes positive, jump, and ER status. The hazard ratio is illustrated for selected values of the covariates, as indicated, where each subgroup was compared against Group 1. Confidence levels correspond to those indicated in the legend. **c.** Kaplan-Meier plots of disease-specific survival for the GI cluster subgroups. **d.** Graphical representation of the Cox proportional hazard ratios for overall survival for several key variables including grade, tumour size, age at diagnosis, number of lymph nodes positive, jump, and ER status. The hazard is illustrated for selected values of the covariates, as indicated, where each subgroup was compared against Group 1. Confidence levels correspond to those indicated in the legend. **e.** Kaplan-Meier plots of overall survival for the GI cluster subgroups.


a
Adjusted Rand Index
0.2
0.3
0.4
0.5
0.6
5
10
15
k
D1 predict D2
D2 predict D1


b
1
2
3
4
5
6
7
8
9
10
Frequency
1
0.5
0
-0.5
−1
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
X
Chromosome


c
1
2
3
4
5
6
7
8
9
10
Frequency
1
0.5
0
−0.5
−1
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 X
Chromosome



e

IntClust	IGP	P.val
1	0.6865672	0
2	0.5818182	0
3	0.7090909	0
4	0.8098592	0
5	0.6454545	0
6	0.5072464	0
7	0.6015038	0
8	0.8235294	0
9	0.7280702	0
10	0.4487179	0

Figure S27: *De novo* clustering of the validation set and prediction of the discovery set. **a.** Plot of the adjusted Rand index over all possible *k* values indicates that *k* = 10 is a global optimum when linear discriminant analysis (LDA) is used to predict subgroup assignment in the discovery (D1) cohort based on the validation cohort (D2) and vice-versa. Here the adjusted Rand index is used to assess predictive accuracy. **b.** Genome-wide gene-centric frequencies (F) of somatic copy number aberrations (CBS-derived) are illustrated for each of the integrative clusters discovered in the validation. Note that the cluster numbering does not correspond to the IntClust groups discovered in the discovery set. **c.** Genome-wide gene-centric frequencies (F) of somatic copy number aberrations (CBS-derived) are illustrated for each of the integrative clusters predicted in the discovery set based on *de novo* clustering of the validation set. Note that the cluster numbering does not correspond to the IntClust groups discovered in the discovery set. **d.** Kaplan-Meier plots of disease-specific survival for the validation cohort and discovery cohort (predicted by the validation cohort). **e.** The in-group proportion (IGP) and associated *P*-value (based on 10,000 permutations) indicate the quality of each cluster reproduced in the discovery set.

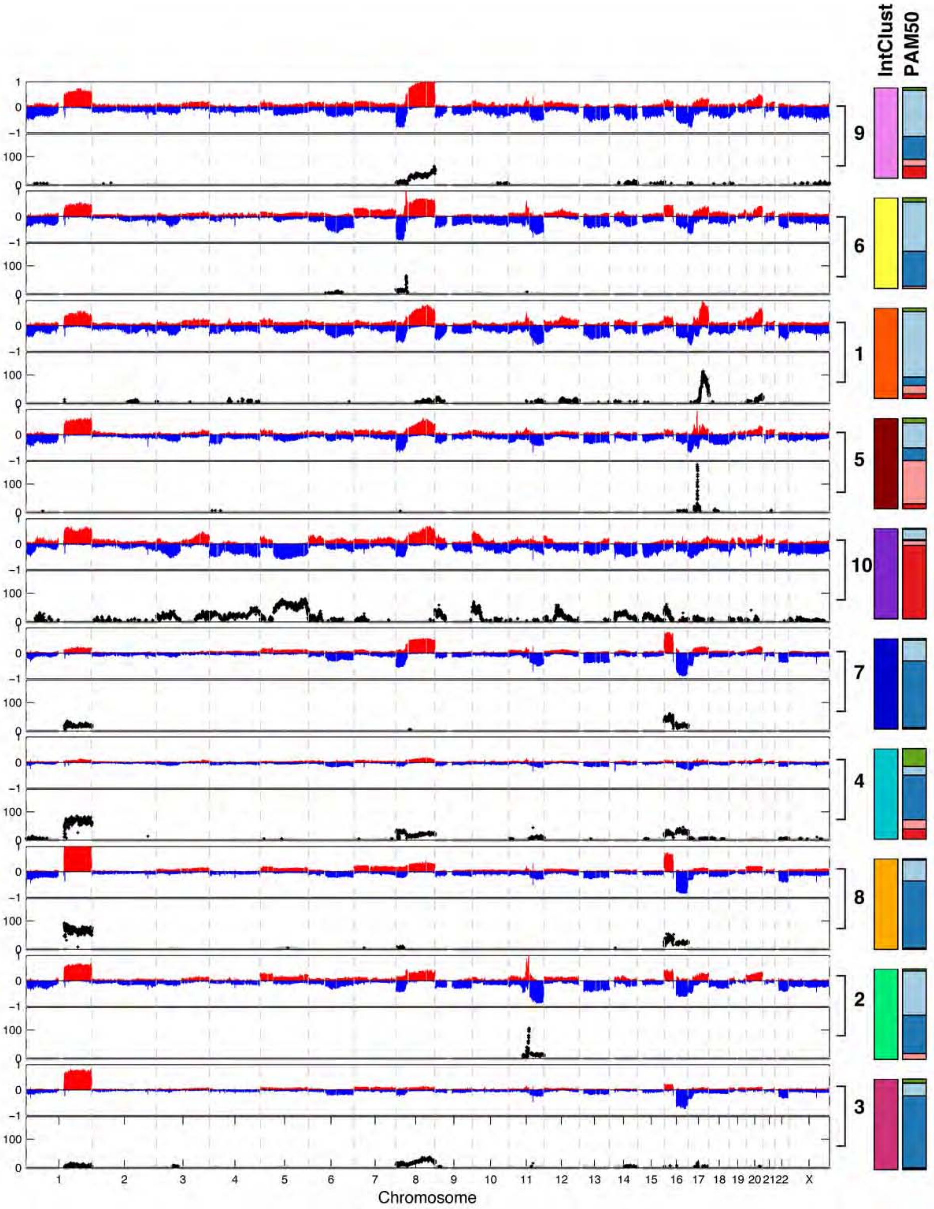
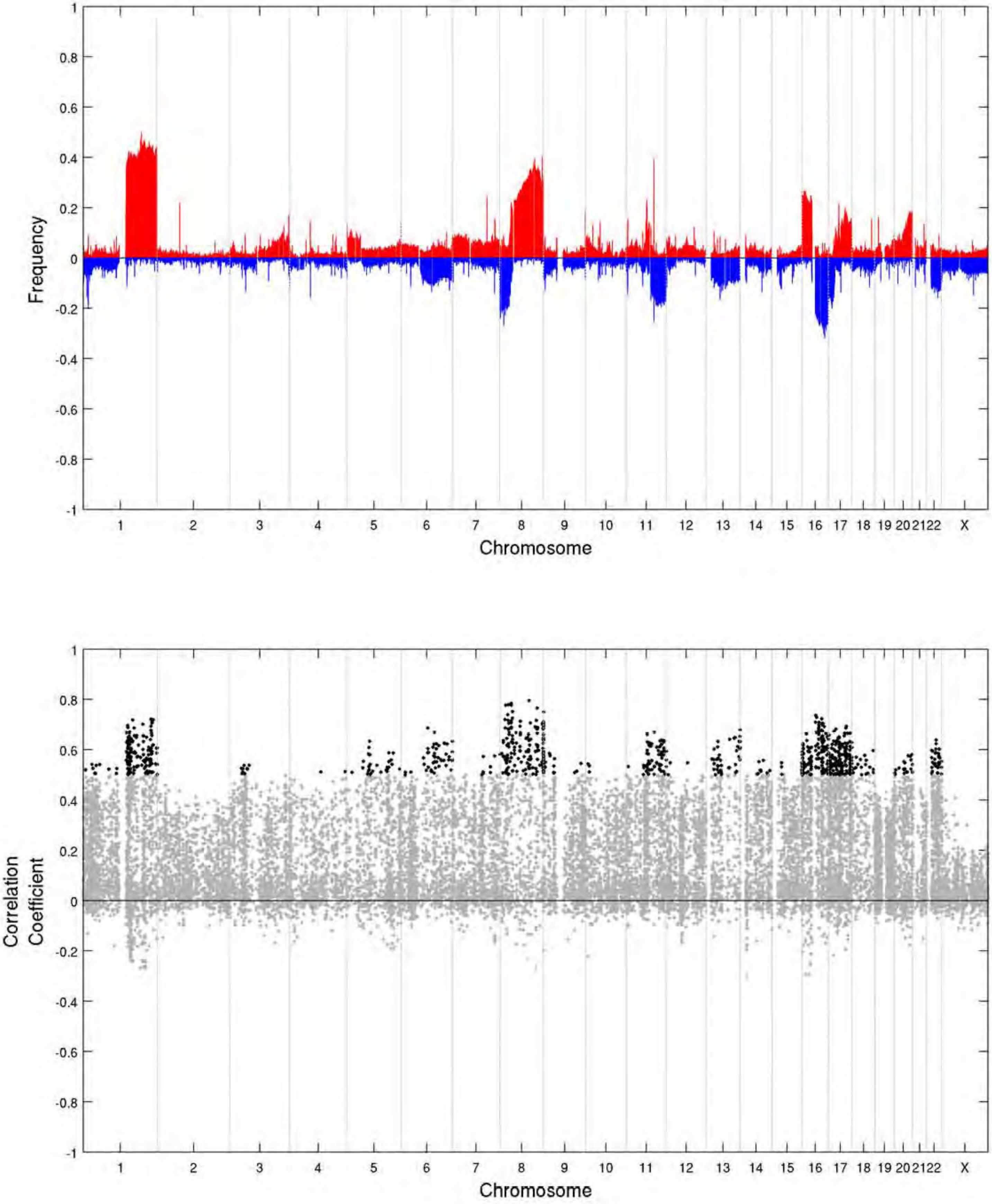


Figure S28: The integrative cluster groups have distinct copy number profiles. Genome-wide gene-centric frequencies (F) of somatic copy number aberrations (HMM-derived) are illustrated for each of the integrative clusters as is the strength of association between an aberration and a particular subtype (-$\log_{10}$ P-value) based on a χ^2 test of independence. Subgroups have been ordered by the similarity (based on hierarchical clustering) of their copy number profiles.

a


1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1
Frequency
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
X
Chromosome
1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1
Correlation Coefficient
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
X
Chromosome

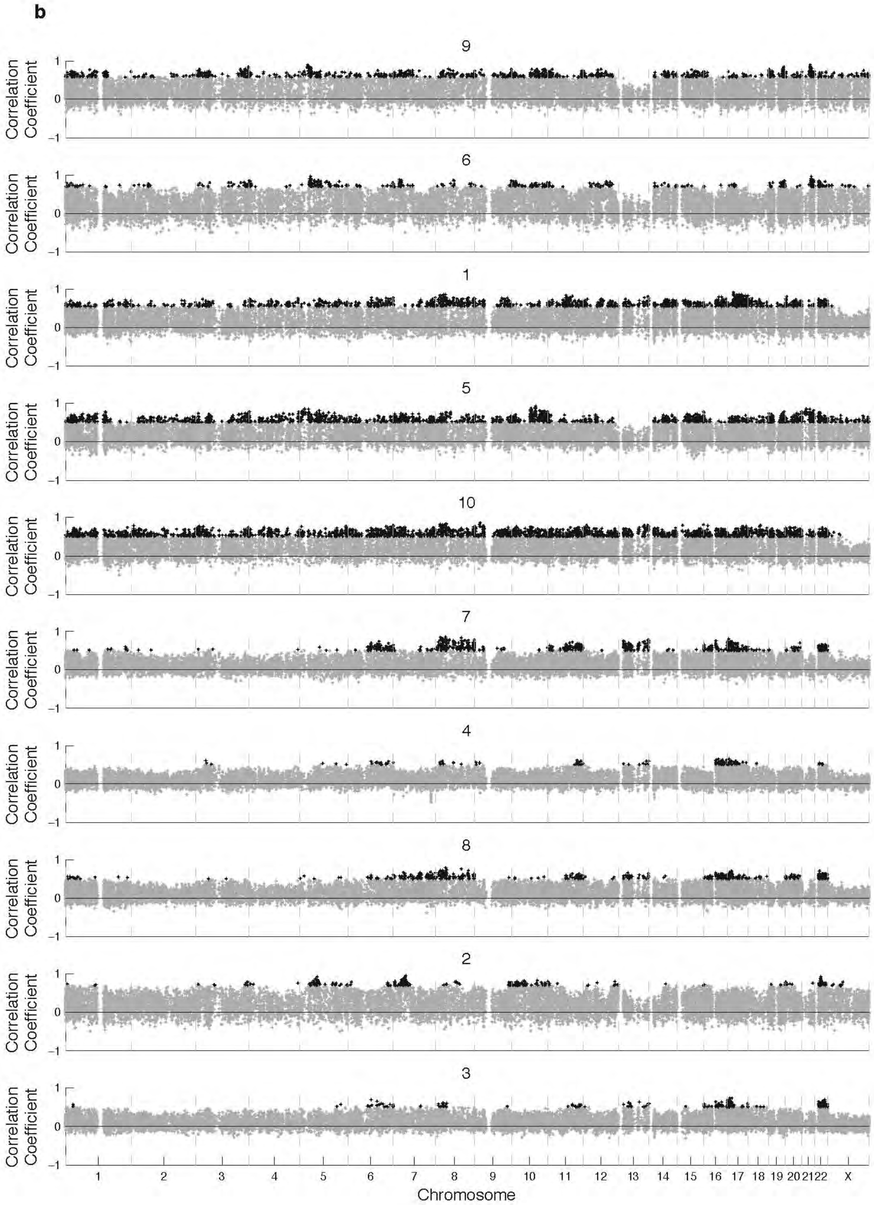
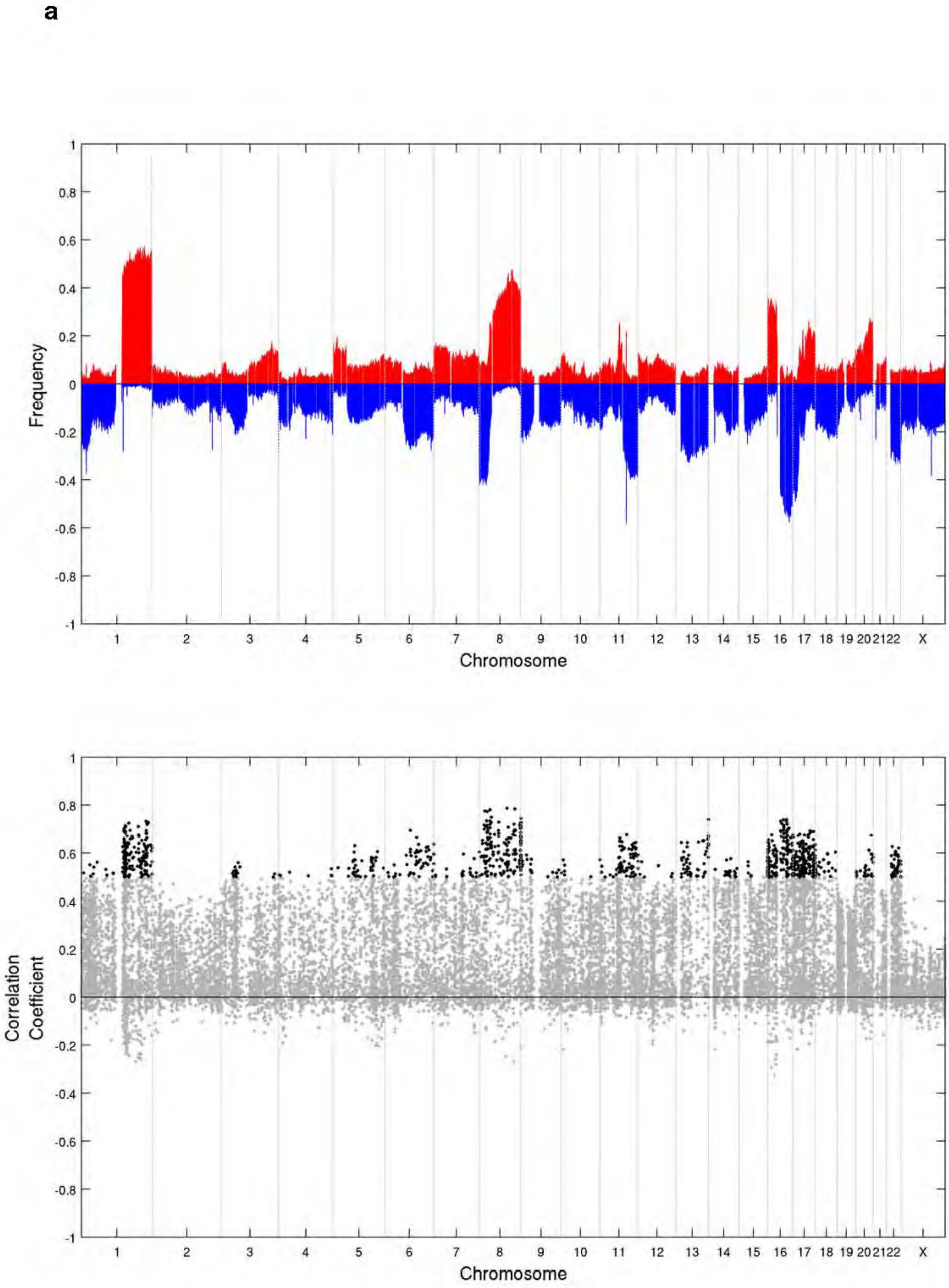


Figure S29: Genome-wide plot of Spearman correlations between CBS-derived CNAs and expression in *cis*. **a.** Genome-wide frequencies and *cis*-acting profiles are illustrated for all tumours, where black points indicate significant correlations ($\rho > 0.5$ and Bonferroni adjusted P-value < 0.05). **b.** As in (a), *cis*-acting profiles are illustrated for each of the integrative subtypes.


a
Frequency
1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 X
Chromosome
Correlation Coefficient
1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 X
Chromosome

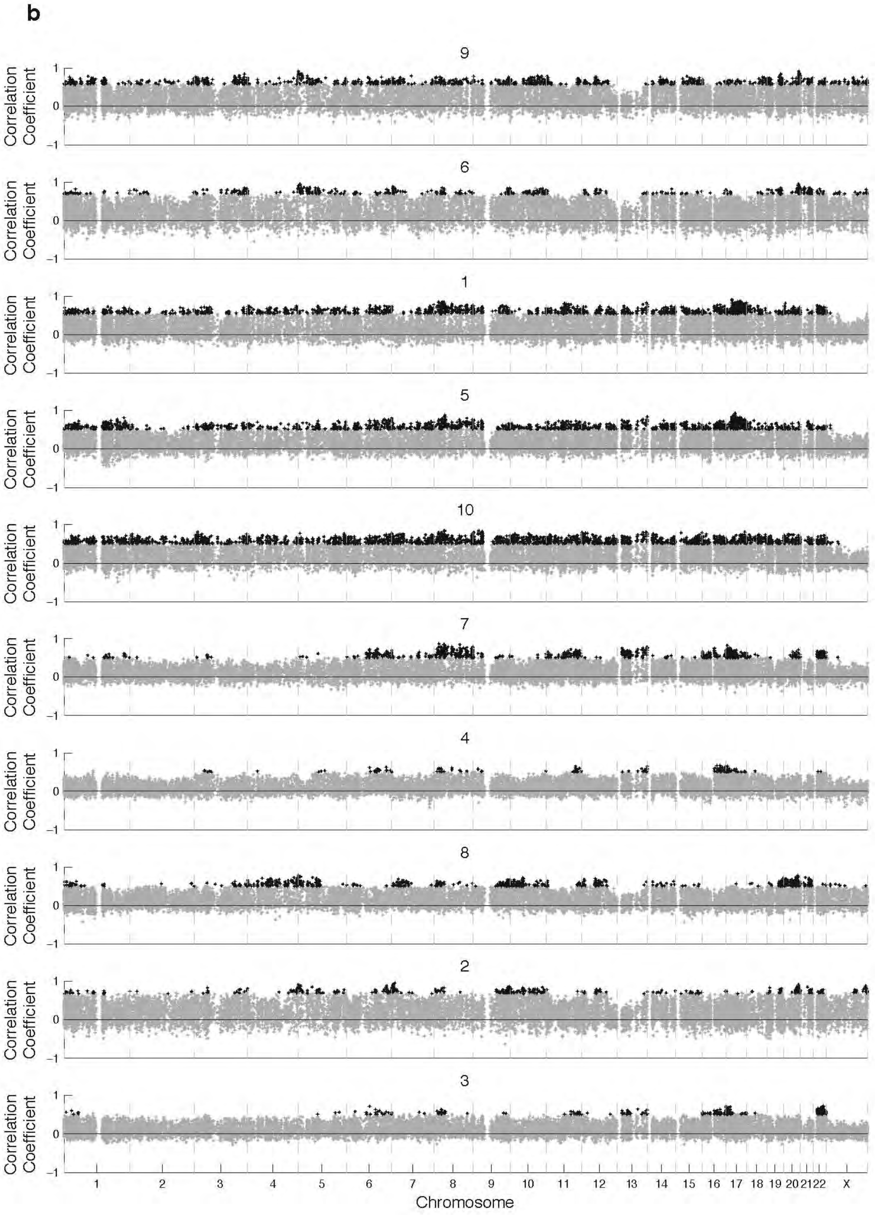


Figure S30: Genome-wide plot of Spearman-correlations between HMM-derived CNAs and expression in *cis*. **a.** Genome-wide frequencies and *cis*-acting profiles are illustrated for all tumours, where black points indicate significant correlations ($\rho > 0.5$ and Bonferroni adjusted P-value < 0.05). **b.** As in (a), *cis*-acting profiles are illustrated for each of the integrative subtypes.

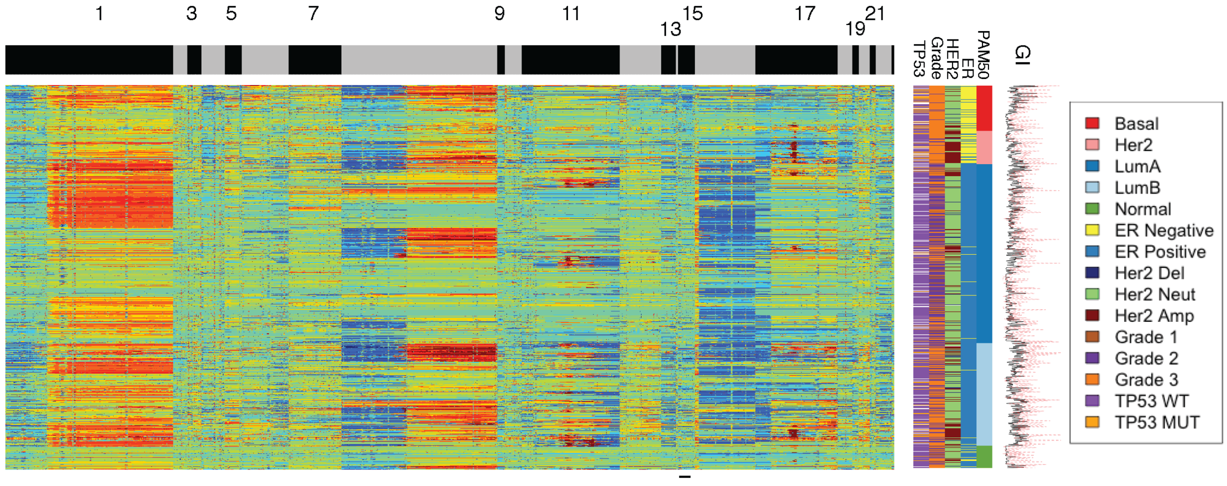


Figure S31: Unsupervised clustering of genome-wide copy number data in 997 tumours from the discovery set indicates heterogeneity within the intrinsic subtypes. Samples were sorted according to their intrinsic subtype classification and clustered based on the copy number profiles (CBS-derived) for 2,995 merged regions using the Euclidean distance metric and the Ward method (red, gain; blue, loss). Genomic instability (GI) based on the proportion of genome altered (black line) and jump measure (red line) is indicated.

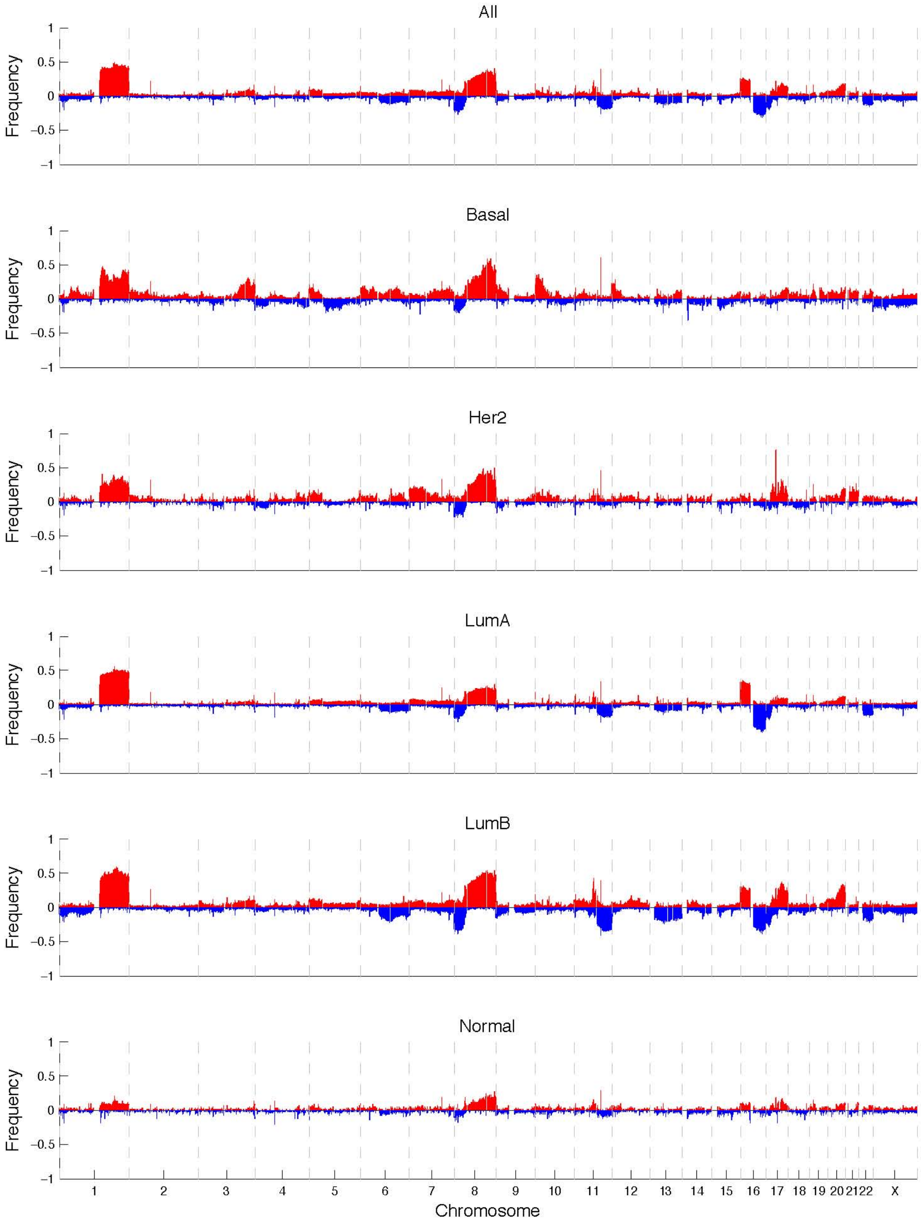


Figure S32: PAM50 subtype-specific copy number profiles for CBS.

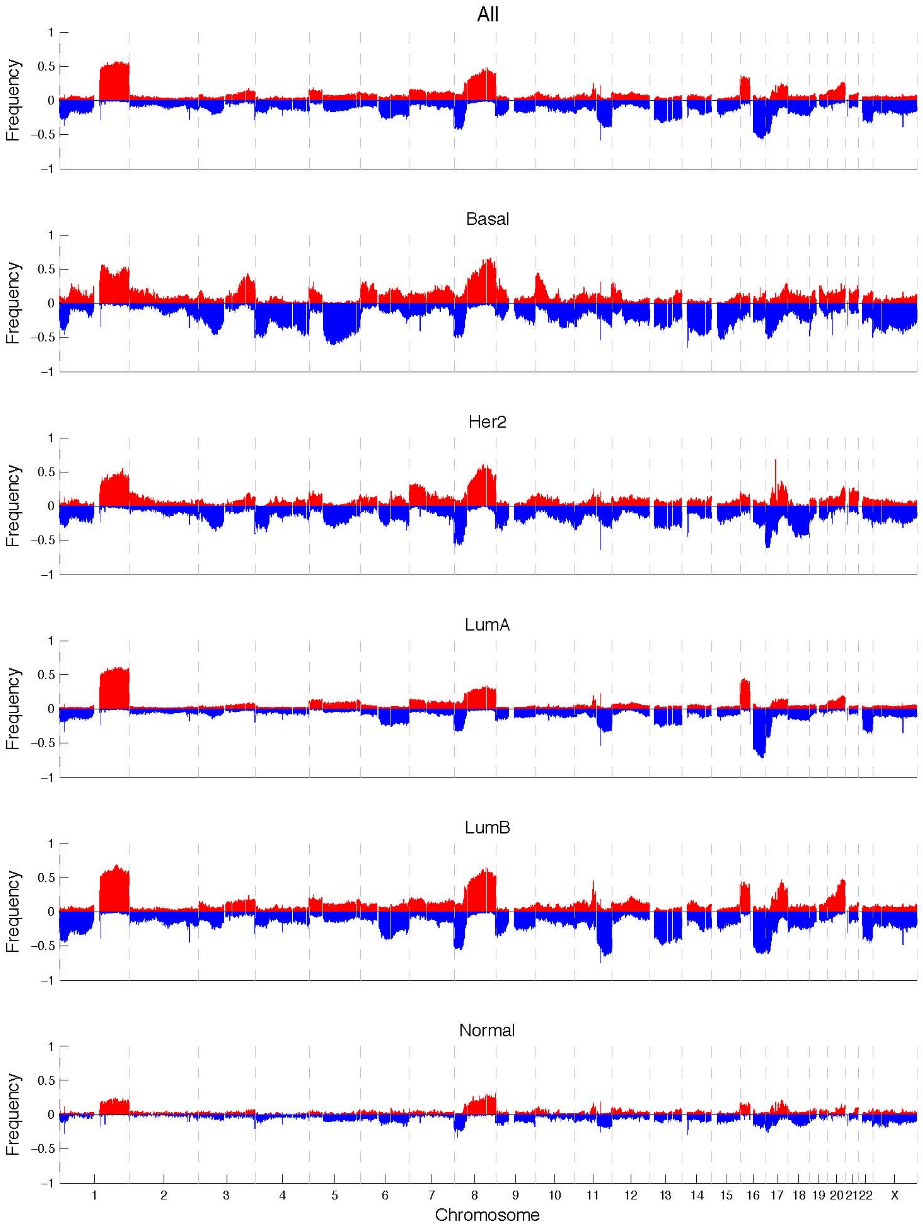


Figure S33: PAM50 subtype-specific copy number profiles for HMM.

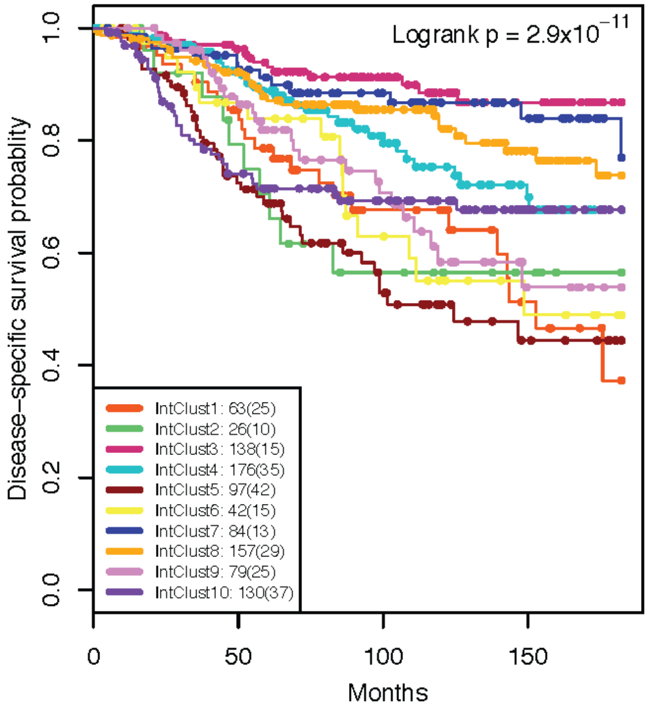


Figure S34: Kaplan-Meier plot of disease-specific survival for the integrative subgroups in the validation cohort.

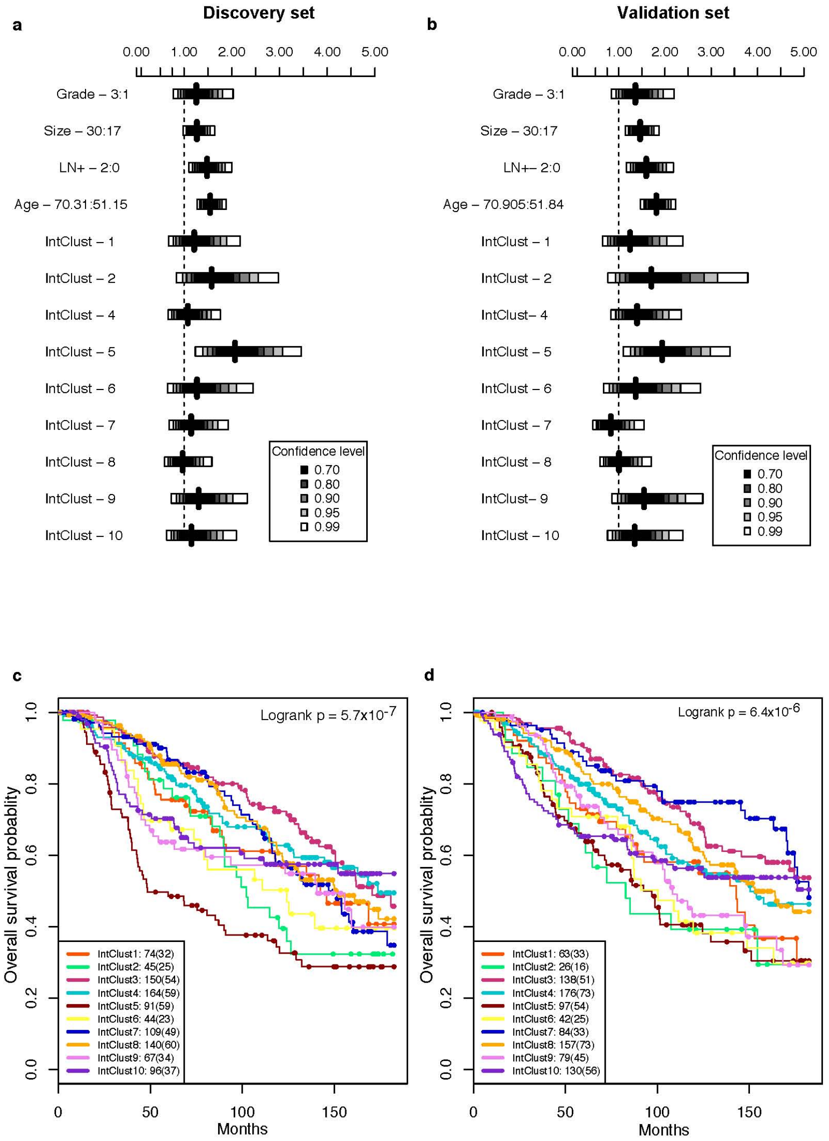


Figure S35: Multivariable Cox regression analysis and Kaplan-Meier plots of overall survival for the integrative subtypes. **a.** Graphical representation of the Cox proportional hazard ratios for overall survival in the discovery cohort for several key variables including, grade, tumour size, age at diagnosis, number of lymph nodes positive, as well as the integrative cluster subgroups. The hazard ratio is illustrated for selected values of the covariates, as indicated where each subgroup was compared against IntClust 3. Confidence levels correspond to those indicated in the legend. **b.** As in (b), but for the validation cohort. **c.** Kaplan-Meier plots of overall survival for the integrative cluster subgroups in the discovery cohort. **d.** Kaplan-Meier plots of overall survival for the integrative cluster subgroups in the validation cohort.

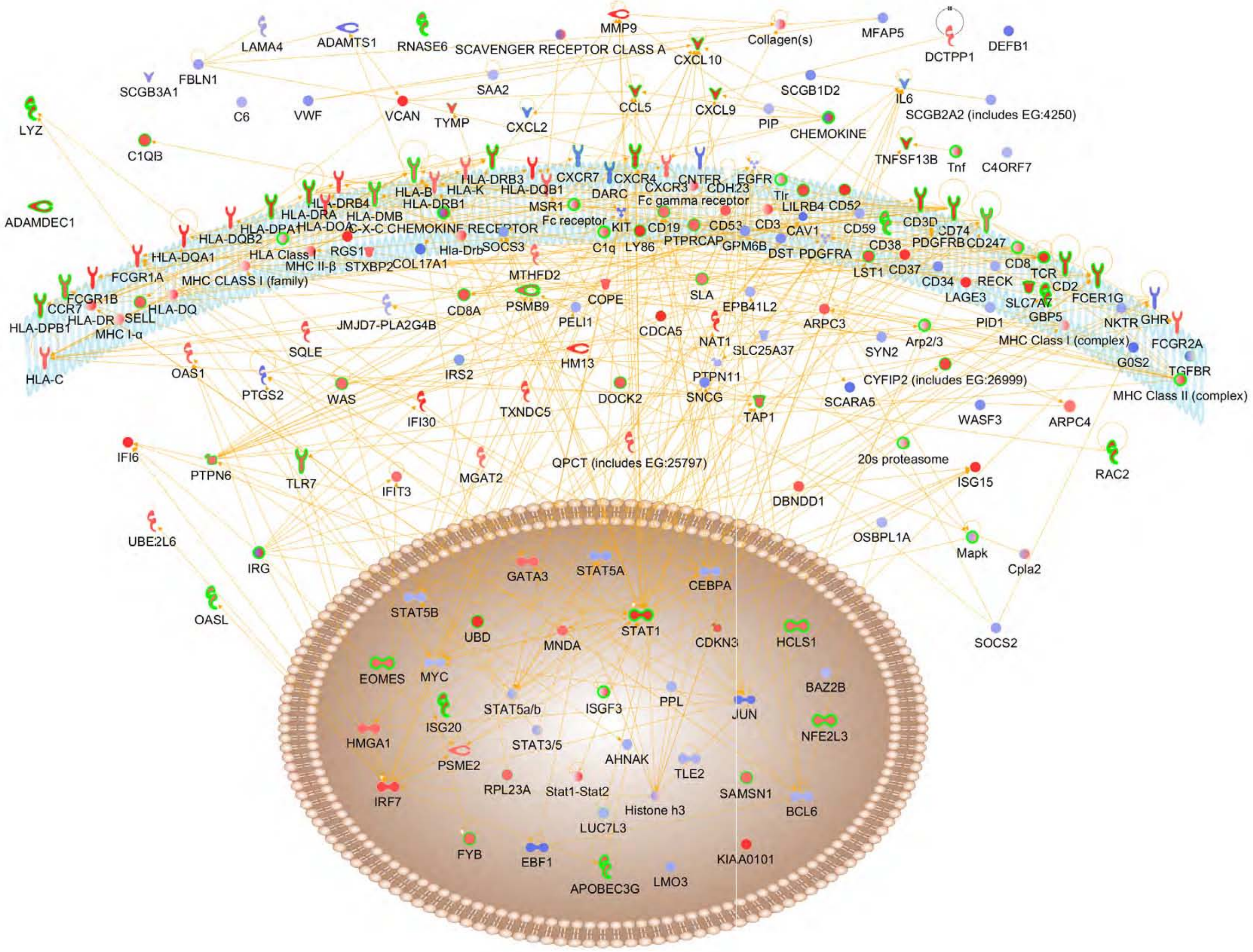


Figure S36: Subtype-specific gene networks are revealed by the CNA-expression landscape. The CNA-devoid subgroup (IntClust 4) exhibits a strong adaptive immune and inflammatory response, which is driven by a *trans*-acting TCR deletion signature. Genes common to the immune response signature and that are modulated in *trans* by TCR deletion events are outlined in green. Up-regulated genes are indicated in red, whereas those that are down-regulated are shown in blue.

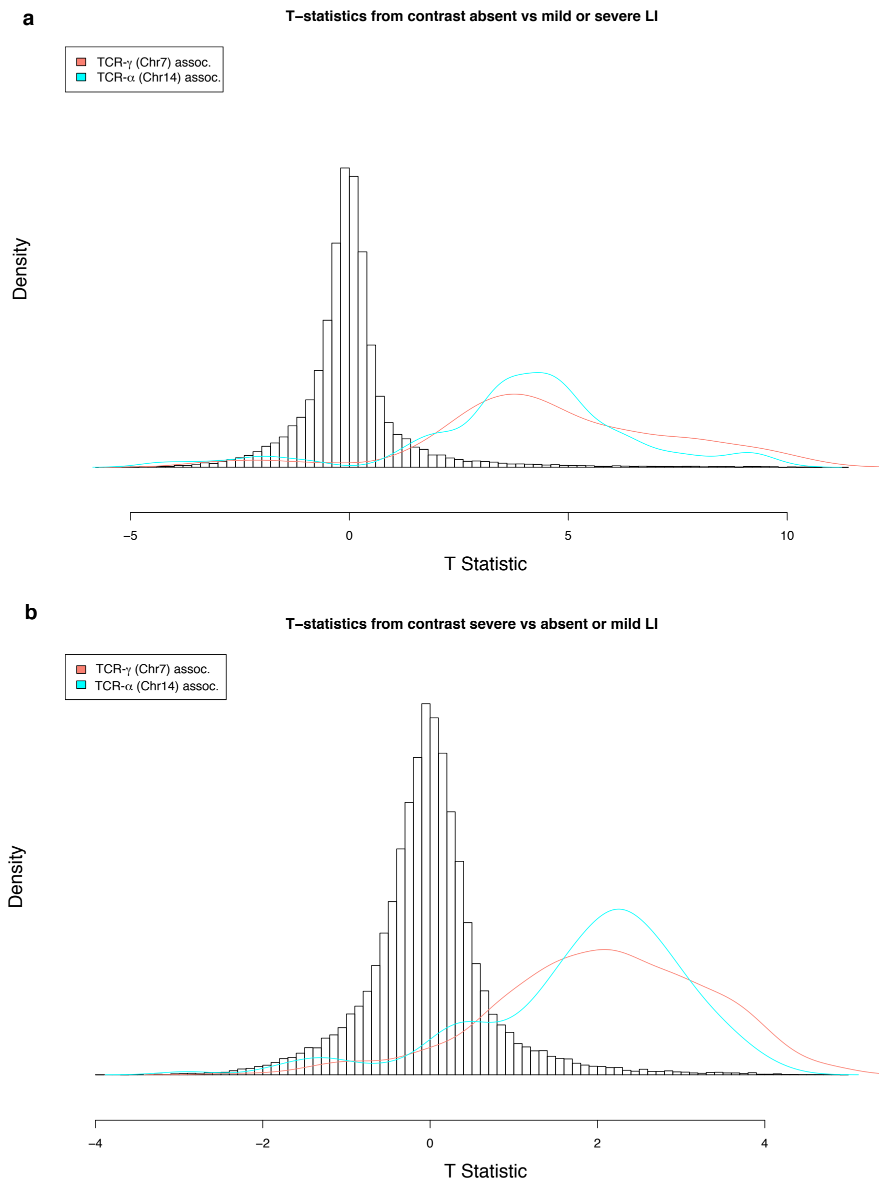


Figure S37: Histograms illustrating the enrichment of *TRG* and *TRA trans*-associated mRNAs amongst genes that were differentially expressed in the CNA-devoid subgroup according to the extent of lymphocytic infiltration. **a.** Histograms illustrating the distribution of T-statistics for mRNAs associated in *trans* with the TCR deletion events in CNA-devoid cases with no LI versus those with mild or severe LI. Genes associated with the *TRG* deletion on chromosome 7 are indicated in orange and the *TRA* deletion on chromosome 14 are indicated in turquoise. **b.** As in (a), but for the comparison of severe LI versus absent or mild LI.

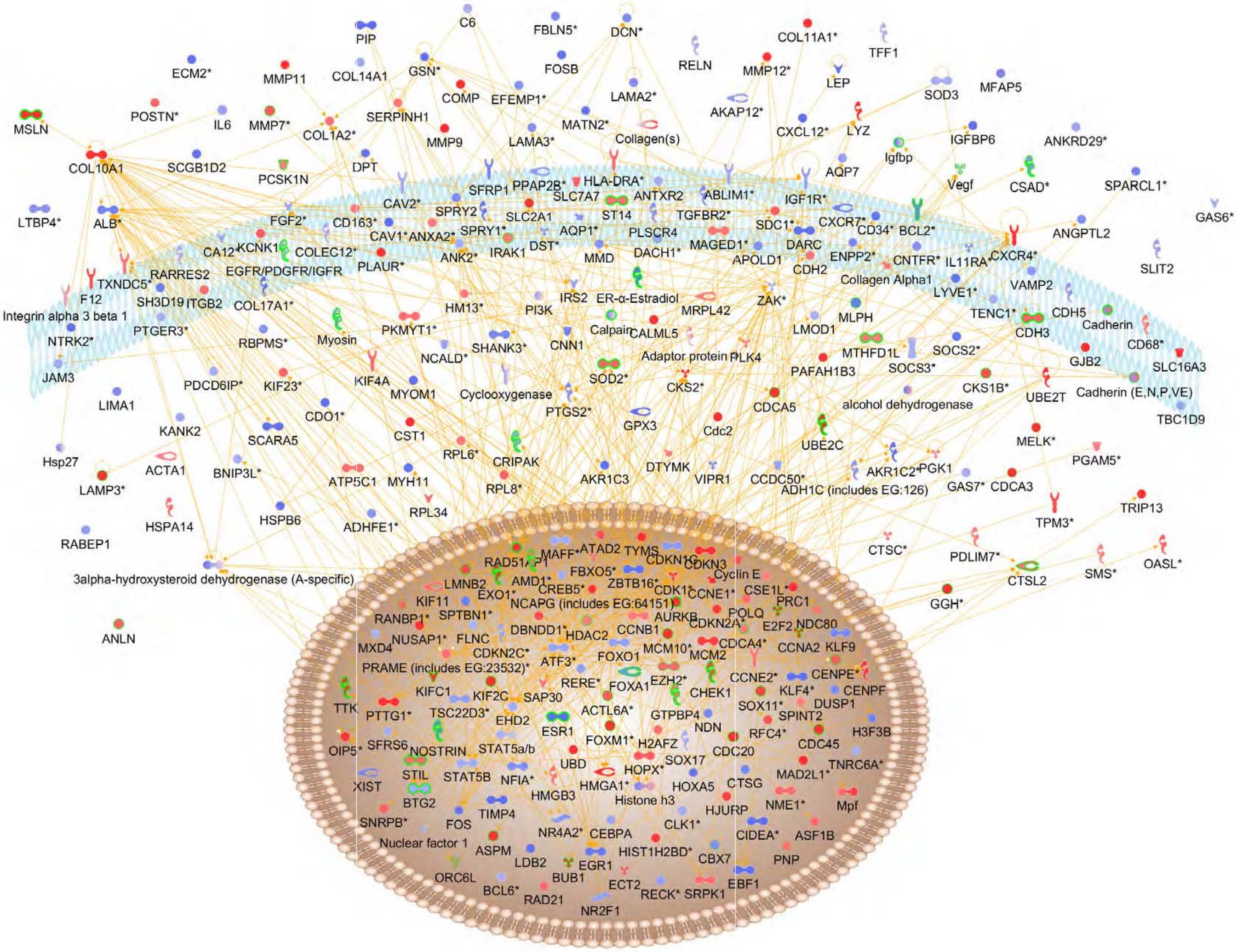


Figure S38: Subtype-specific gene networks are revealed by the CNA-expression landscape. The Basal subgroup (IntClust 10) is characterised by cell cycle and DNA repair processes that largely overlap with a Basal-specific chromosome 5 *trans*-acting deletion. Genes common to the pathway signature and that are modulated in *trans* by chromosome 5q deletions are outlined in green. Up-regulated genes are indicated in red, whereas those that are down-regulated are shown in blue.

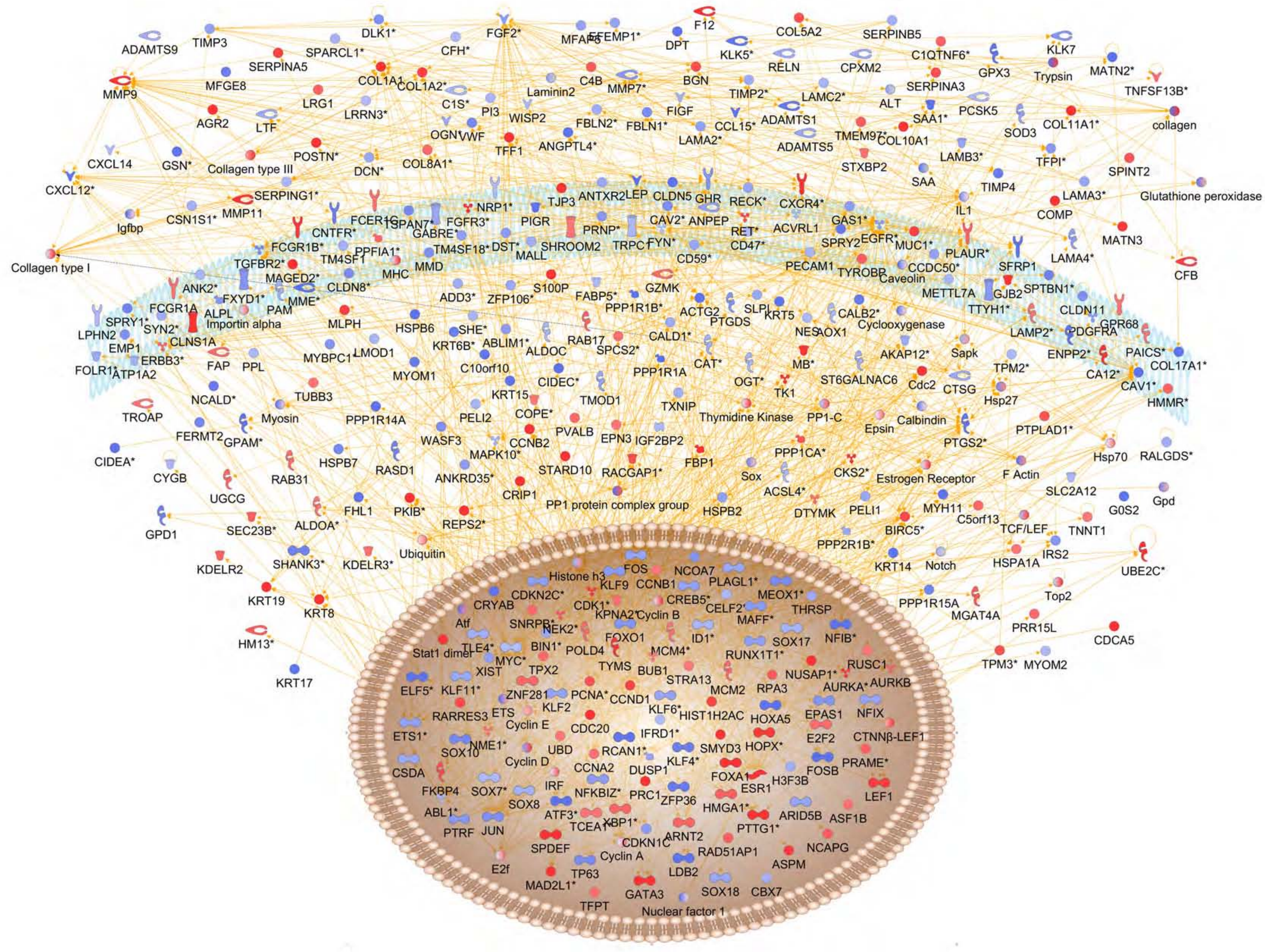


Figure S39: Subtype-specific gene networks are revealed in the CNA-expression landscape. The 11q13/14 *cis*-acting, poor prognosis subgroup (IntClust 2) is enriched for genes implicated in the cell cycle and developmental related processes. Up-regulated genes are indicated in red, whereas those that are down-regulated are shown in blue.

19 Supplementary Table Legends

Table 1: Clinical and demographic characteristics of the discovery and validation tumour cohorts

Table 2: Clinical annotation for the discovery cohort of 997 breast cancer patients

Table 3: Clinical annotation for the validation cohort of 995 breast cancer patients

Table 4: Gene-centric frequencies of somatic aberrations (derived from CBS) across various subtypes

Table 5: Minimal common regions of alteration for CNAs (derived from CBS), stratified by ER status and PAM50 subtype for regions with frequency > 0.05

Table 6: Broken gene frequency summarized across CBS-derived copy number changes, where the pattern of breakage is indicated along with the total number of samples in which a break was observed

Table 7: Gene-centric frequencies of somatic aberrations (derived from HMM) across various subtypes

Table 8: Broken gene frequency summarized across HMM-derived copy number changes, where the total number of samples in which a break was observed is indicated

Table 9: Verification of germline CNVs and somatic homozygous deletions

Table 10: List of primer sequences for verification of homozygous deletions

Table 11: Comparison of CNA calling accuracy between CBS and HMM based on MLPA and corresponding frequencies of alteration

Table 12: Minimal common regions of alteration for germline CNVs (derived from CBS), stratified by ER status and PAM50 subtype for regions with frequency > 0.05

Table 13: Probe-level regions of germline CNVs (derived from CBS) across various cohorts

Table 14: Gene-centric frequencies of germline CNVs (derived from CBS) across various subtypes

Table 15: Distribution and genomic locations of germline CNVs (derived from CBS) in normal breast tissue, tumour samples, and the HapMap population

Table 16: Inferred ethnicity based on projection of tumour genotype data onto HapMap population clusters

Table 17: Genome-wide associations of tumour expression with SNPs, CNVs, and CNAs (derived from CBS)

Table 18: Genome-wide associations of tumour expression with SNPs, CNVs, and CNAs (derived from CBS) in cases of European ancestry for which normal genotype calls were available

Table 19: Accuracy of genotyping calls in non-diploid copy number regions

Table 20: Genome-wide associations of normal expression with SNPs, CNVs, and CNAs (derived from CBS) in cases of European ancestry for which both normal genotype calls and normal gene expression profiles were available

Table 21: Mean and median variance explained by SNP, CNV, and CNA tumour expression associations

Table 22: Genomic regions exhibiting high-level amplification, amplification and putative homozygous deletions (derived from HMM) driving expression to the extreme tails of the population-level distribution

Table 23: High-level amplifications driving expression to the extreme tails of the population-level distribution (derived from HMM) and the frequency of the derivative copy number event for both HMM and CBS analysis

Table 24: Putative homozygous deletions driving expression to the extreme tails of the population-level distribution (derived from HMM) and the frequency of the derivative copy number event for both HMM and CBS analysis

Table 25: Enrichment analysis of *trans*-acting aberration hotspots. Lists of *trans*-acting aberration hotspots (denoted by CNA_ID), their cognate mRNAs (denoted by GENE_NAME), and the -$\log_{10}$ P-value of association between a copy number aberration and gene for varying window sizes

Table 26: Enrichment analysis of *trans*-associated mRNAs for the 10Mb window, labelled according to the corresponding CNA_ID in the aberration hotspot file (Table S25)

Table 27: Pathway enrichment of genes with mRNAs correlated with deletion of the *TRG* locus

Table 28: Pathway enrichment of genes with mRNAs correlated with deletion of the *TRA* locus

Table 29: Pathway enrichment of genes with mRNAs correlated with deletion of chromosome 5q

Table 30: Gene-centric ANOVA to determine copy number associated gene expression changes in *cis* for CBS

Table 31: List of gene expression and copy number features input and selected for the $k = 10$ partitioning based on integrative clustering of CBS-derived copy number changes

Table 32: List of outlying features used as input for K-Means clustering on expression data

Table 33: Cox proportional hazard ratios for **a.** clusters based on the *de novo* clustering of the validation cohort and **b.** clusters predicted in the discovery cohort from the validation cohort

Table 34: Gene-centric Spearman correlation analysis to determine copy number induced gene expression changes in *cis* amongst the integrative subtypes (CBS)

Table 35: Gene-centric Spearman correlation analysis to determine copy number induced gene expression changes in *cis* amongst the integrative subtypes (HMM)

Table 36: Gene-centric analysis to determine copy number associated gene expression changes based on both non-parametric (Kruskal-Wallis rank sum test and Wilcoxon rank sum test) and parametric (ANOVA, Tukey honest significant difference) tests (CBS)

Table 37: Gene-centric analysis to determine copy number associated gene expression changes based on both non-parametric (Kruskal-Wallis rank sum test and Wilcoxon rank sum test) and parametric (ANOVA, Tukey honest significant difference) tests (HMM)

Table 38: Gene-centric χ^2 analysis to determine subtype-specific copy number changes amongst the integrative subtypes (CBS)

Table 39: Gene-centric χ^2 analysis to determine subtype-specific copy number changes amongst the integrative subtypes (HMM)

Table 40: Cox proportional hazard ratios for **a.** the discovery set, **b.** the predicted memberships of the validation set, and **c.** a comparison of the validation set relative to the discovery set

Table 41: Comparison of C-index values for the IntClust model, PAM50 model, and clinico-pathological model

Table 42: Centroids based on the nearest-shrunken centroids method for classifying samples according to the 10 integrative clusters

Table 43: Matrix of features from integrative clustering for the discovery and validation set used for classification

Table 44: Predicted cluster membership for the validation cohort based on PAMR

Table 45: Scoring of lymphocytic infiltration in the CNA-devoid subgroup (IntClust 4)

Table 46: Differentially expressed genes for contrasts comparing each of the integrative subtypes versus normal breast tissue

Table 47: Summary of pathway enrichment amongst the integrative subgroups