doi:10.1038/nature16478

Supplementary Data Table of Contents

Supplementary Table 1 *Driver genes prediction in mouse primary and recurrent tumours*.

(a) Next generation sequencing of transposon insertion sites from primary untreated tumours, local recurrences, and metastatic recurrences followed by determination of commonly inserted genes (gCIS analyses), reveals extremely divergent putative driver events (gCISes) between primary and recurrent disease. Trp53, Arid1b and Tcf4 insertions (highlighted in yellow) were found in both local and metastatic recurrences, but not in therapy naïve tumours. Crebbp, highlighted in grey, was targeted by SB transposition across all compartments. Yellow highlights gCISes in common between local and metastatic recurrences, grey shading denotes gCISes shared across all compartments. (b) Initiator driver genes predicted by our complementary computational method. This per-mouse driver modeling approach predicts initiator genes for each compartment. Gray and Yellow shading denote initiator genes shared by all three or just two compartments, respectively. (*) denotes En2 gene, this is a previously identified screening artifact and not considered an initiator event. (c) Relative quantification of lineage abundance in the matched primary, local recurrent and metastatic recurrent samples by a per-mouse genetic algorithm as detailed in Methods and Supplementary Note. Initiator genes are found at different levels of dominance in the different compartments. Dominant clones in the primary tumours are less prevalent in the recurrences. Patterns suggestive of clonal selection are also present when estimating relative prevalence in the metastatic recurrences. (d) A subset of the initiator genes predicted by our computational methods (displaying >10% relative prevalence or being

predicted in the founder lineage) were analyzed by MSigDB to assess pathway enrichment. Primary tumours show enrichment for the Shh pathway, but this is lost at the time of recurrence. Locally recurrent samples are instead enriched for Trp53 pathway genes and DNA repair genes, neither of which are observed in the untreated primary tumours.

Supplementary Table 2 *Patient information, analyses summary, and associated clinical data*. (a) Clinical information and (b) data summary for the cohort of 46 patients. (c) Total number of SNVs and indels called by Strelka run in paired mode. i.e. rec-germ indicates that mutations were called in the recurrent tumour versus the germline sample. (d) Damaging SNVs and Indels called by Strelka are summarized by compartment. When possible, mutations are called in paired mode in each compartment (pre and post-therapy samples) versus the germline DNA, yielding somatic mutations. To enrich for somatic variants when no germline DNA is available, paired mutation calling is done using the pre versus post-therapy sample (to identify mutations specific to the naive tumour), and in the post-therapy vs naive compartment (to identify mutations specific to the post-therapy tumour). (e) Copy Number events by tumour compartment, called using CNAseq (0: complete loss; 4: gain; 5: high-level amplification) (f) Structural variations called using ABySS. (g) Sequencing statistics for exome libraries. (h) Sequencing statistics for WGS libraries. (i) Neoplastic drug - gene interactions from the Drug Gene Interaction database, targeting those genes identified as mutated or aberrant at the copy number or structural level. Copy number events included the subset of focal events

(<12Gb) that were complete losses (state=0), or high-level gains and amplifications (state=4 or 5). (j) Damaging SNVs were classified based on variant allele frequencies (VAF) into clusters of clonal and subclonal. The median VAF of each cluster, and the size of each cluster is shown for each SNV. (k). Primary-specific SNVs are annotated with clonal cluster information (i.e. the cluster each SNV belongs to, according to a Gaussian mixture model approach), and CNV and LOH calls. Fractional Copy Number is provided from Control-Freec results, and LOH state and allelic summary are indicated along with the LOH confidence (LOH; state_allele_conf). Only SNVs with confident SNV calls are shown (<10 in at least one compartment). (l) Seven patients with sufficient material were used to study the transcriptome by RNA-Seq. For each compartment the expression values are shown (RPKM) as well as the rounded absolute difference between compartments (ABS_difference) and the Log2FC transformation (and absolute Log2FC) for each gene (recurrence vs primary). A gene is considered upregulated when the absolute difference between compartments is>10, and differentially expressed when the Log2FC > 2. Each gene that is differentially expressed is denoted as "Upregulated_Rec" if the expression is higher in the recurrent tumour, and "Upregulated_Pri" if the expression is higher in the primary tumour. (m) Differentially expressed genes in the recurrent samples of at least 2 SHH patients. (n) GSEA analysis (mSigDB) on differentially expression genes in SHH patients. (o) Differentially expressed genes in the recurrent samples of at least 2 Group4 patients. (p) GSEA analysis (mSigDB) on differentially expression genes in Group4 patients. (q) Somatic SNVs, indels, structural variations, homozygous deletions, and high-level amplifications specific to the therapy naive or the recurrent tumours were summarized by patient, and then tested for

recurrence in the cohort. Genes with recurrent events in >2 untreated primary tumours, or in >2 recurrent tumours are short-listed. Genes can appear in both lists if they have different deleterious events in the untreated or recurrent tumour compartments of distinct patients.

Supplementary Table 3 *Positions verified by deep sequencing*.

(a) Fourteen patients with therapy-naïve, post-therapy, and matching germline DNA were analyzed by EXPANDS to infer the clonal composition of the compartments.

(b) The Shannon diversity index was calculated for each compartment using the inferred cellular frequencies of subpopulations defined by EXPANDS. Higher values correspond to greater clonal heterogeneity.

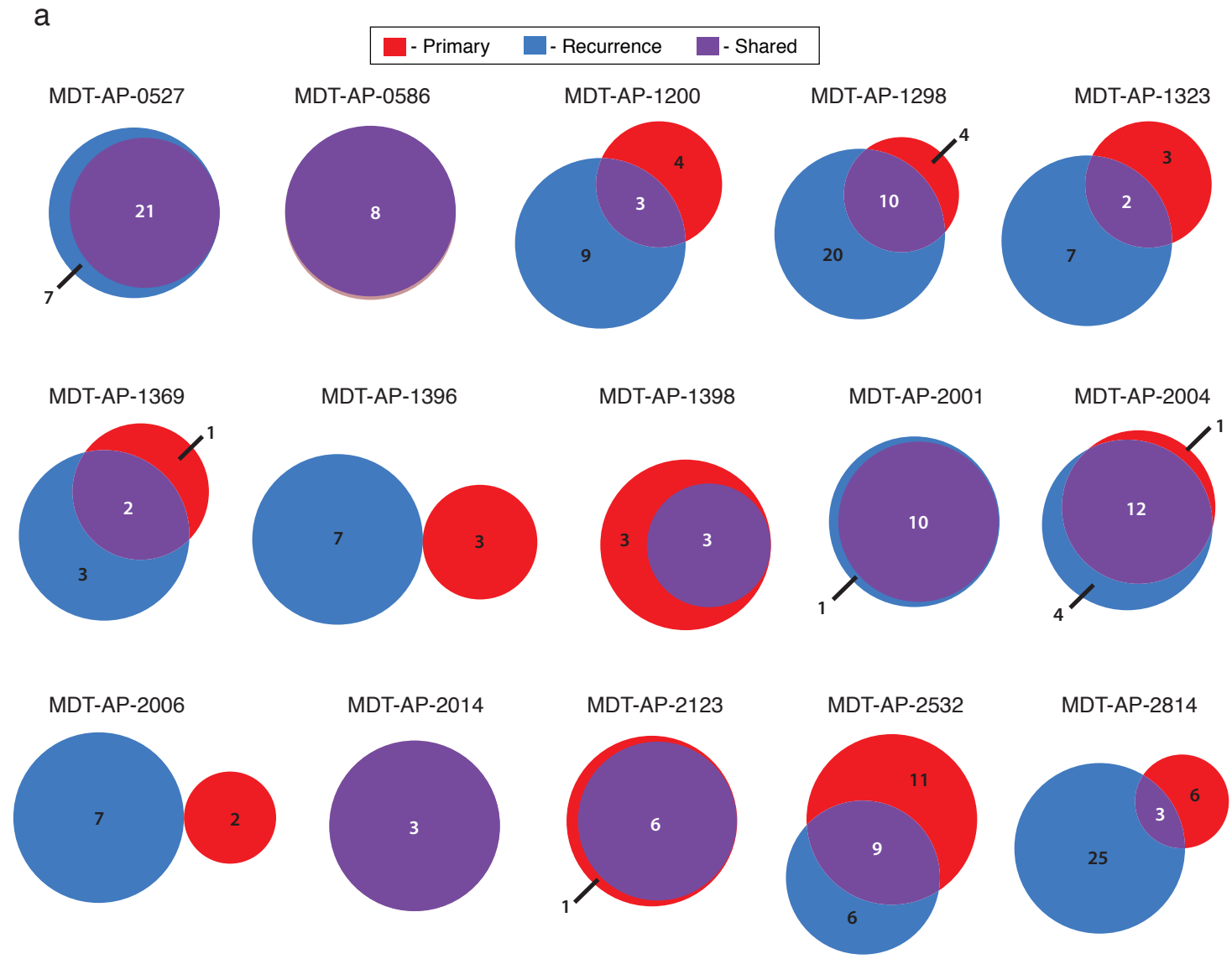
(c) 192 SNVs and indels were selected for deep sequencing in patients with sufficient DNA material remaining after WGS library construction. Primers flanking each position were used to amplify template from the naïve and post-treatment tumours, and matching germline DNA when available. For each position, the allelic frequency in the corresponding WGS library is shown alongside the allelic frequency in the deeply sequenced libraries. A mutation is considered verified when there is sufficient coverage (>10 reads), and when the allelic frequency is >0.1 in the deepseq data. If a mutation an interest was initially identified as compartment-specific (e.g. in the post-therapy sample), it is only verified in that sample (thus, the ICGC verification status flag is set to NOT_TESTED in the naive tumour sample). If a mutation was detected by WGS and

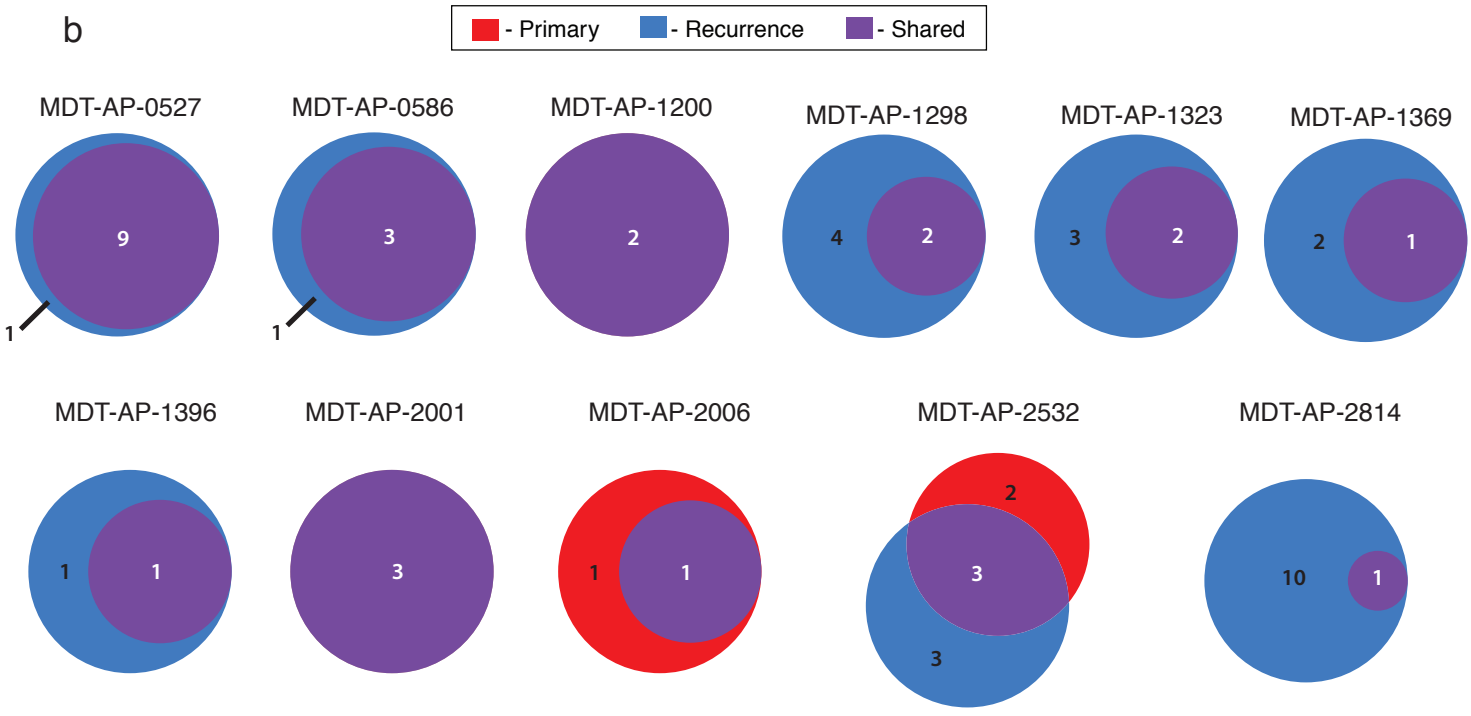
has suffuicient coverage in the deepseq data but an allelic frequency less than 0.1 the flag is set to TESTED_AND_FALSE.

Supplementary Table 4 *Prognostic implications of chr14 loss in Shh MB*.

(a) 14q-associated genes were identified in Shh MBs using ANOVA (n=82 patients). 284 differentially expressed genes between samples with chr14q balanced versus those with chr14q loss are ranked by pvalue. The top 20 ranking genes were used for further analysis.(b) Clinical information of patient samples from an independent MB cohort from Boston. (c) K-means clustering results (k=2) for Shh tumours in the Boston cohort, using the top 20 ranking genes. (d) Follow up (years) and current status of patients in the chr14q-balanced vs chr14q-loss clusters. (e) Relapse status and time to relapse in patients with balanced chr14q vs chr14q loss. (f) Clinical information of patients in the chr14q-balanced vs chr14q-loss clusters. of patient samples from the MAGIC cohort.

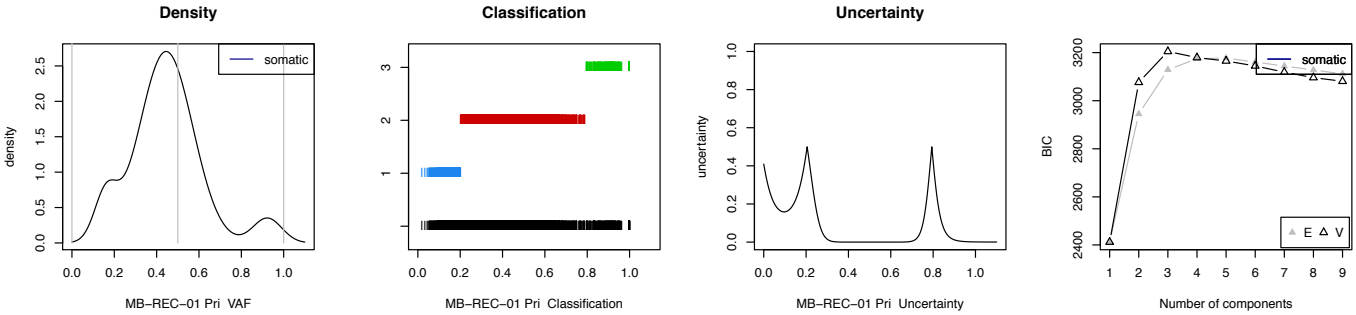
Structural variations by tumor compartment. (a) Copy number variations (CNVs) and (b) loss of heterozygosity, by compartment, using cna-seq.


a
- Primary
- Recurrence
- Shared
MDT-AP-0527
21
7
MDT-AP-0586
8
MDT-AP-1200
4
3
9
MDT-AP-1298
4
10
20
MDT-AP-1323
3
2
7
MDT-AP-1369
1
2
3
MDT-AP-1396
7
3
MDT-AP-1398
3
3
MDT-AP-2001
10
1
MDT-AP-2004
1
12
4
MDT-AP-2006
7
2
MDT-AP-2014
3
MDT-AP-2123
6
1
MDT-AP-2532
11
9
6
MDT-AP-2814
6
3
25

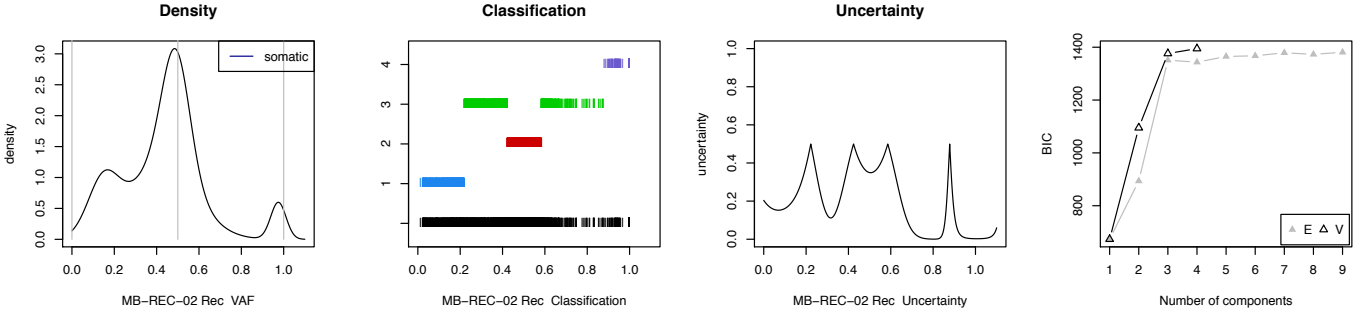

b
- Primary
- Recurrence
- Shared
MDT-AP-0527
9
1
MDT-AP-0586
3
1
MDT-AP-1200
2
MDT-AP-1298
4
2
MDT-AP-1323
3
2
MDT-AP-1369
2
1
MDT-AP-1396
1
1
MDT-AP-2001
3
MDT-AP-2006
1
1
MDT-AP-2532
2
3
3
MDT-AP-2814
10
1

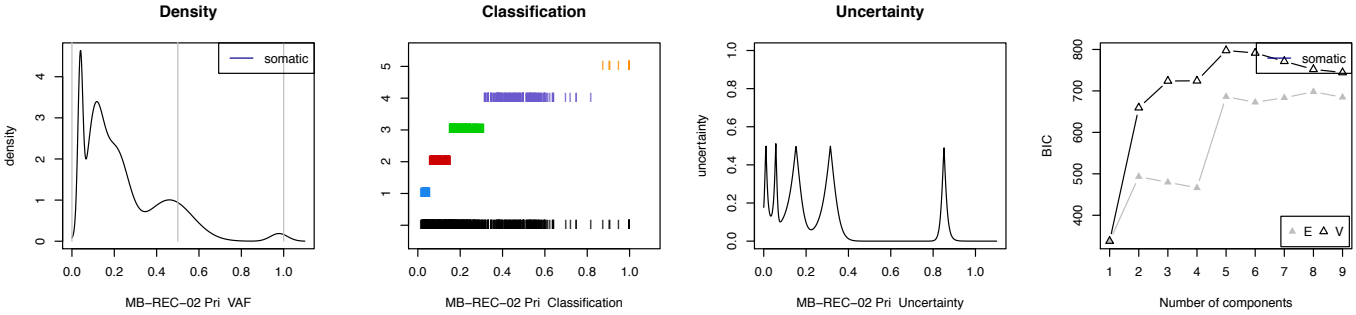
SNV classification using variant allelic frequencies: Classification results for mutations identified as somatic in the primary and recurrent compartments of each patient. Variant allele frequencies (VAF) of damaging SNVs (Strelka) were classified using finite mixture estimation using iterative Expectation Maximization steps (EM) and the Bayesian Information Criterion (BIC). The left panel (Density) of each row is a density plot of the variant allele frequency (VAF) of mutations in that compartment, where the compartment is labelled on the x-axis. Mutations are classified into clusters (Classification panel) that correspond to populations of distinct prevalence. The number of distinct clusters is shown on the y-axis. The uncertainty of classification and the BIC for each set of parameters are plotted in two panels on the right. In patients with matched germline, somatic mutations are used in this analysis. When matched germline was not available, to enrich for somatic mutations, we analyzed separately those mutations found uniquely in one compartment (i.e. Pri-Rec indicates mutations found uniquely in the primary tumor when compared to the recurrent tumor).

MB-REC-01

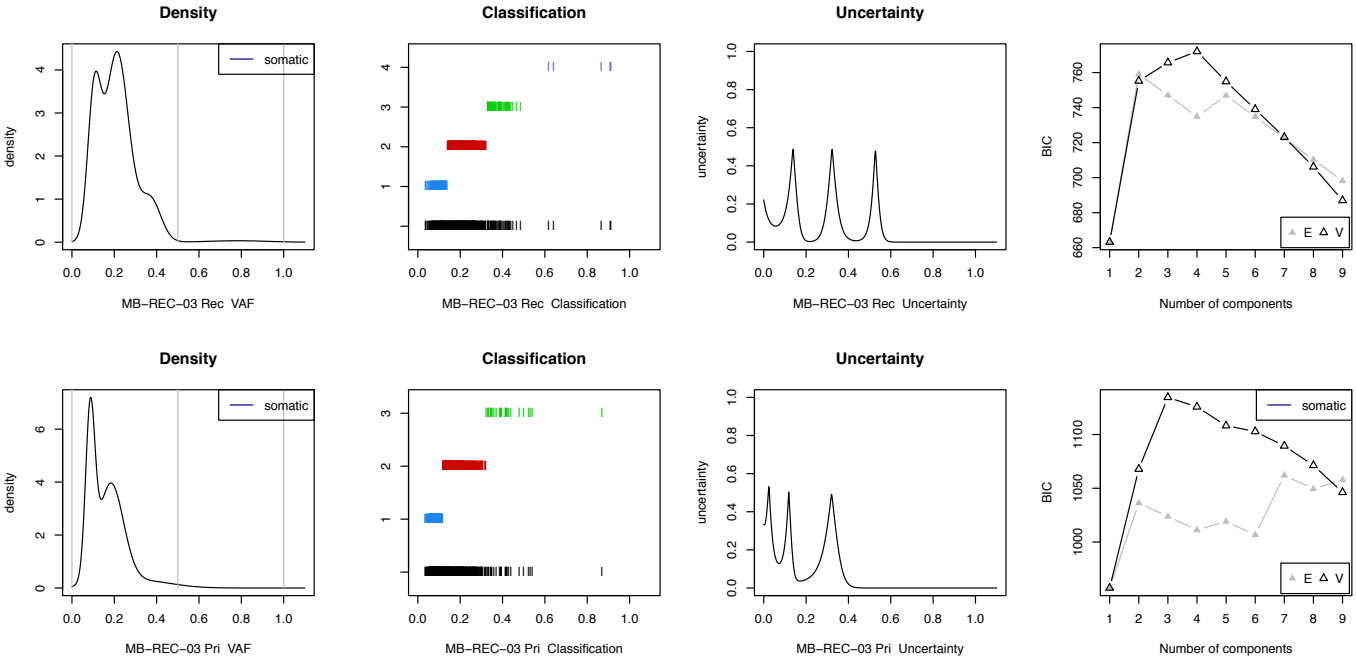


MB-REC-02

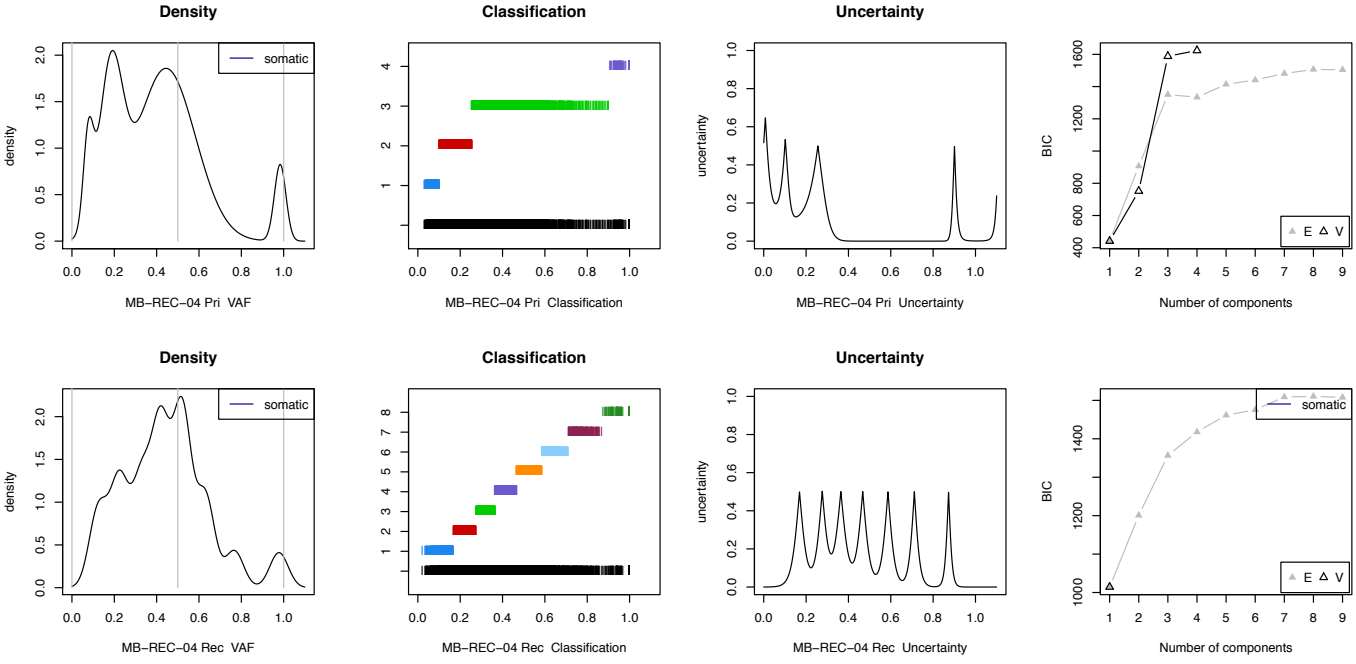




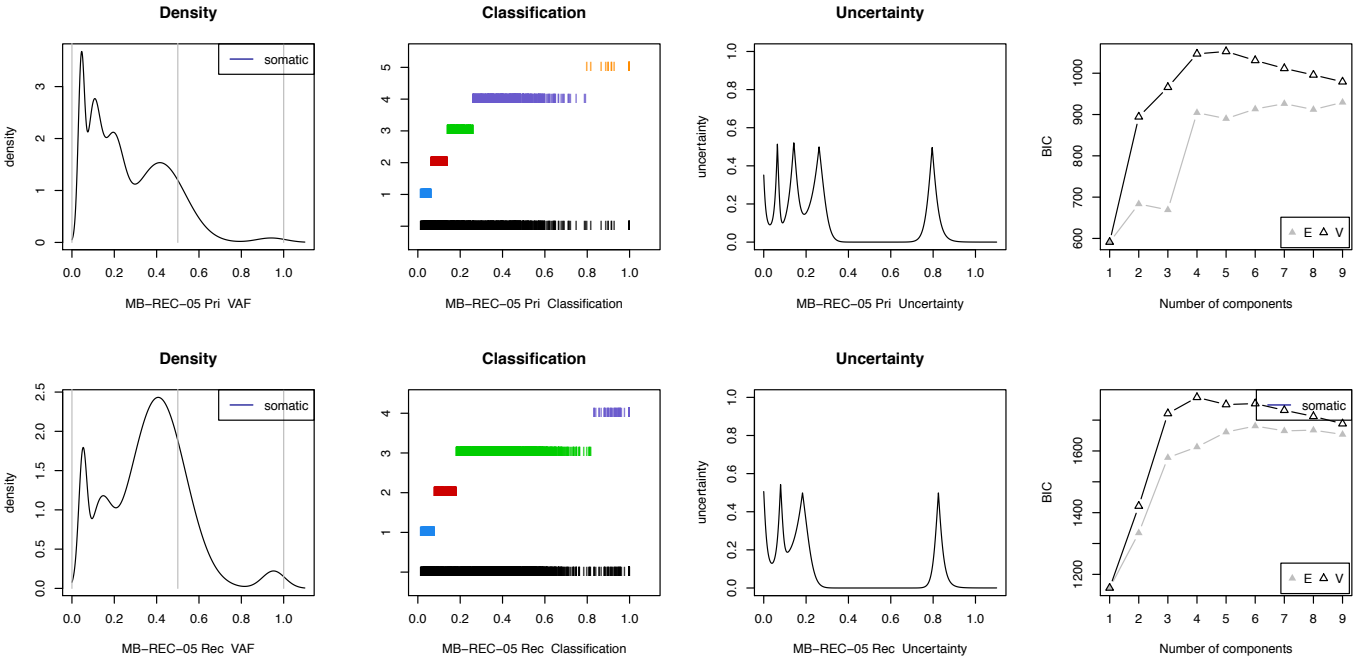
MB-REC-03



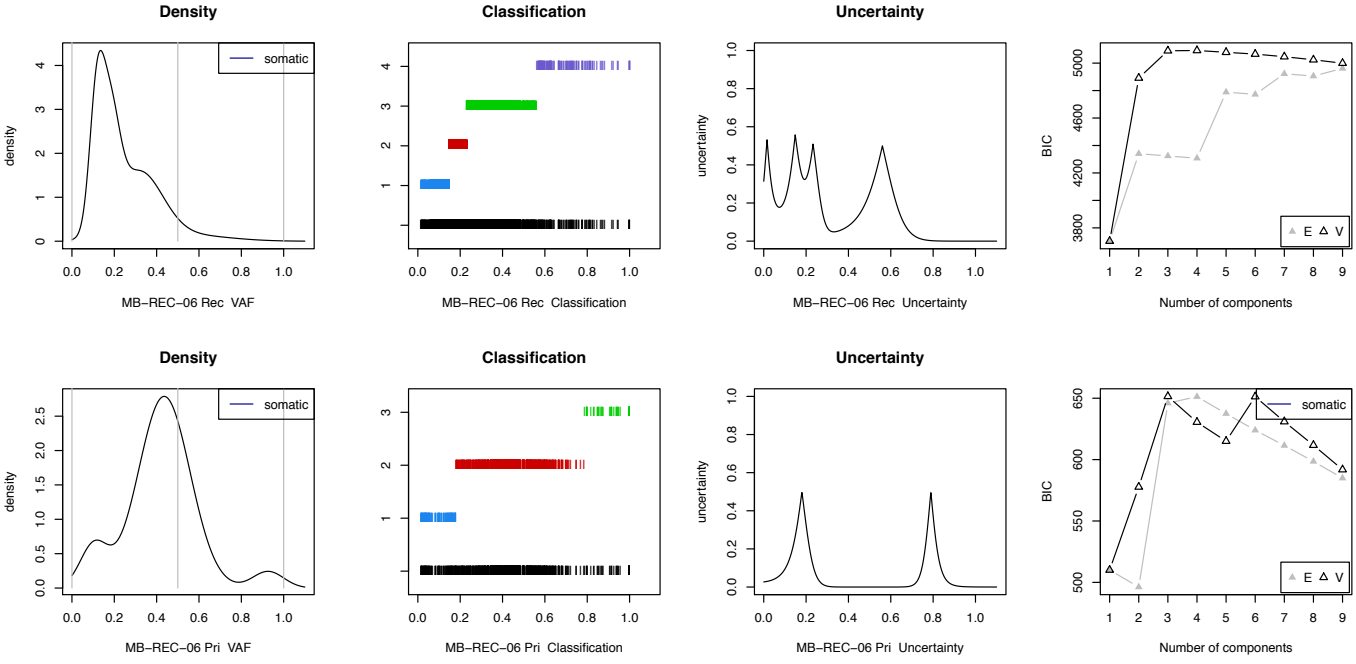
MB-REC-04



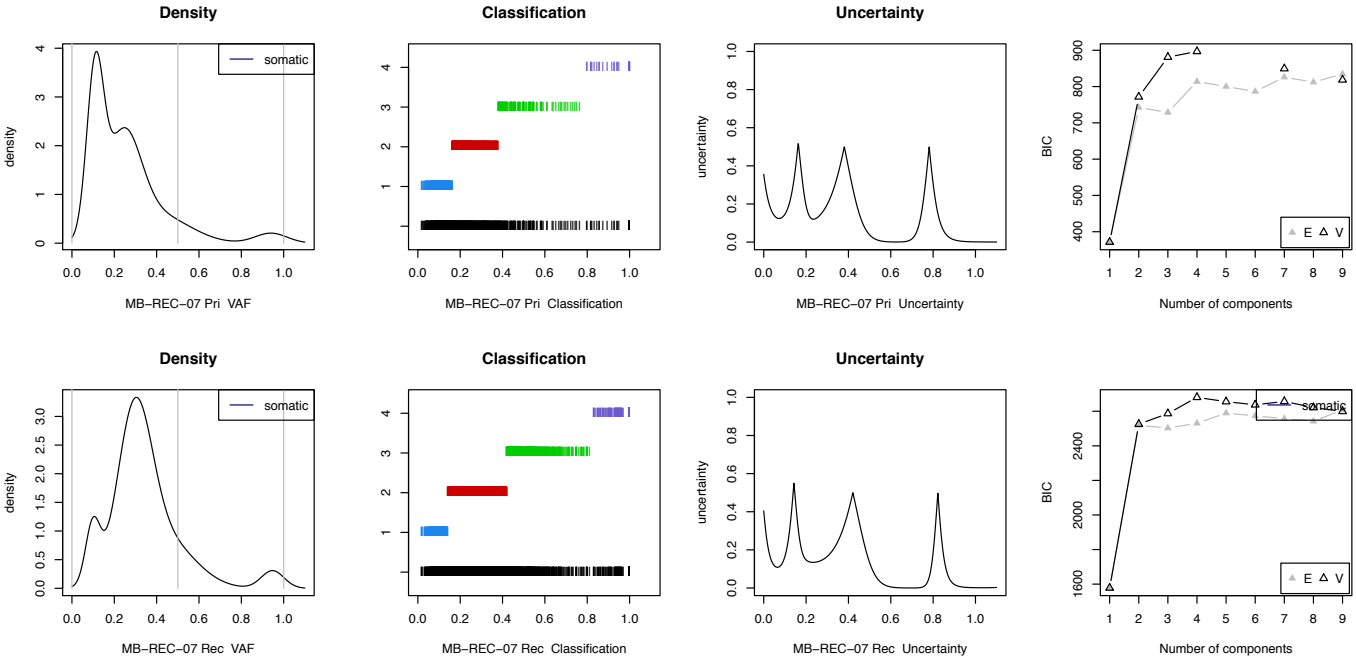
MB-REC-05



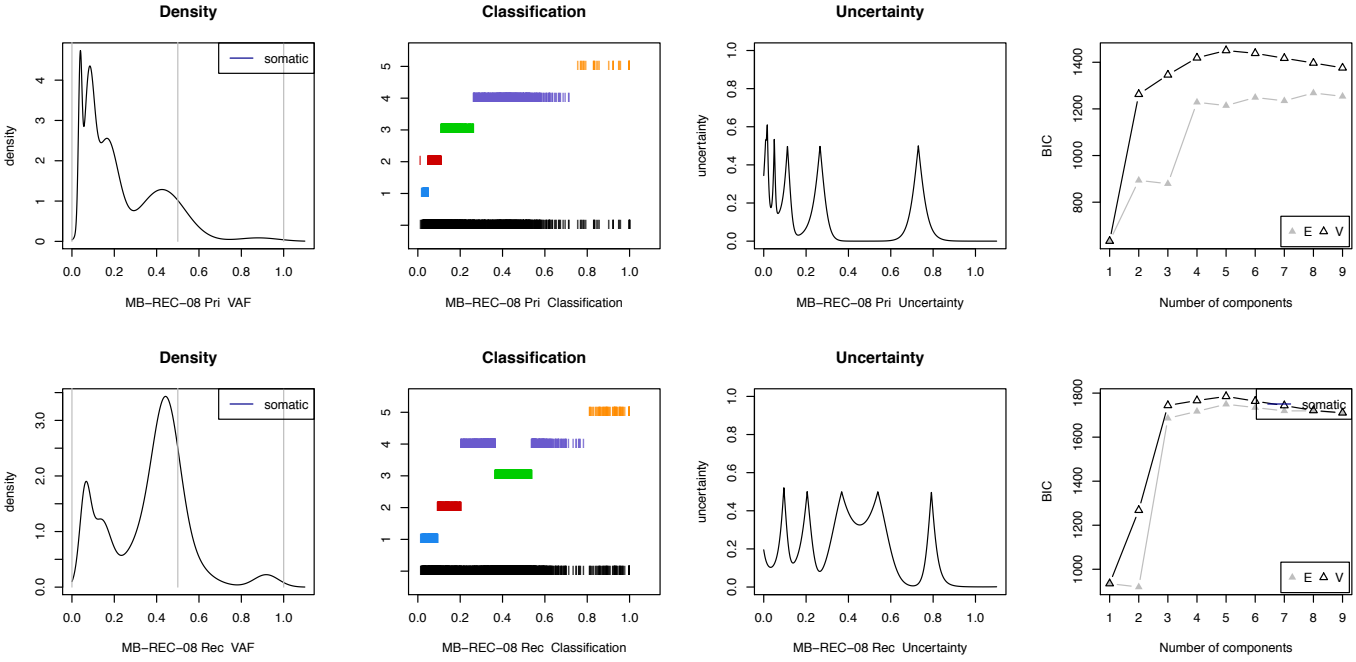
MB-REC-06



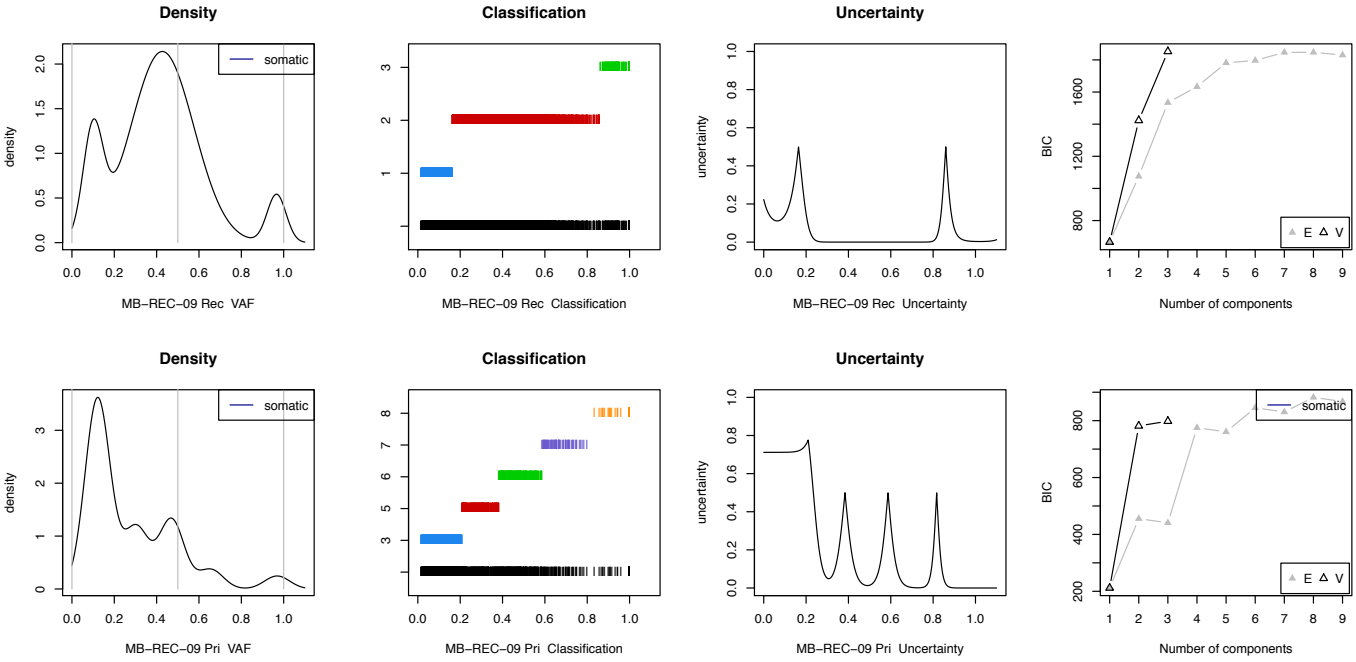
MB-REC-07



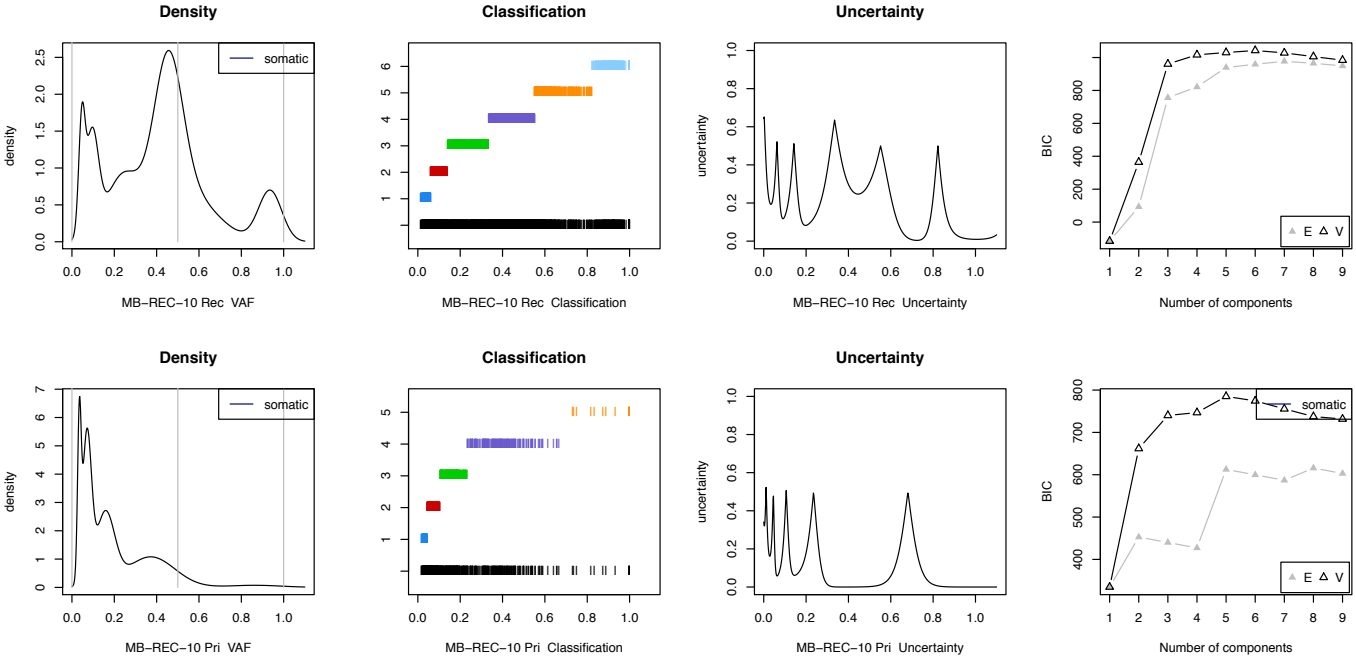
MB-REC-08



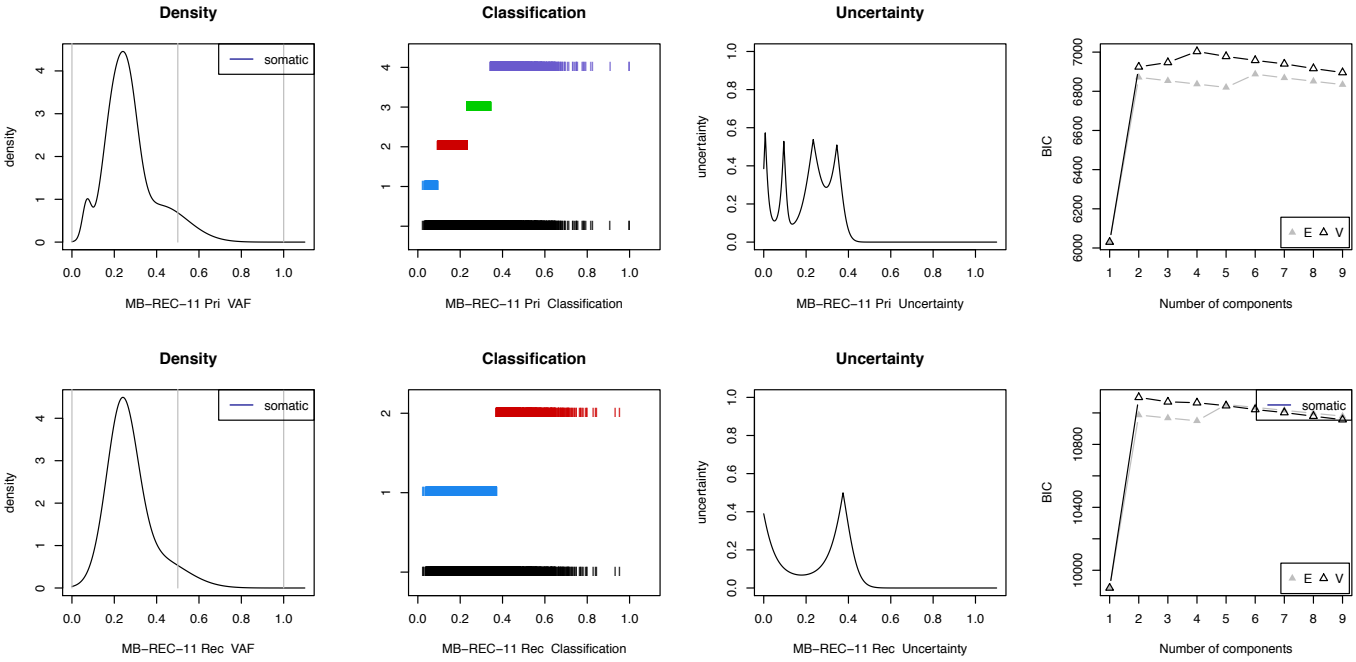
MB-REC-09



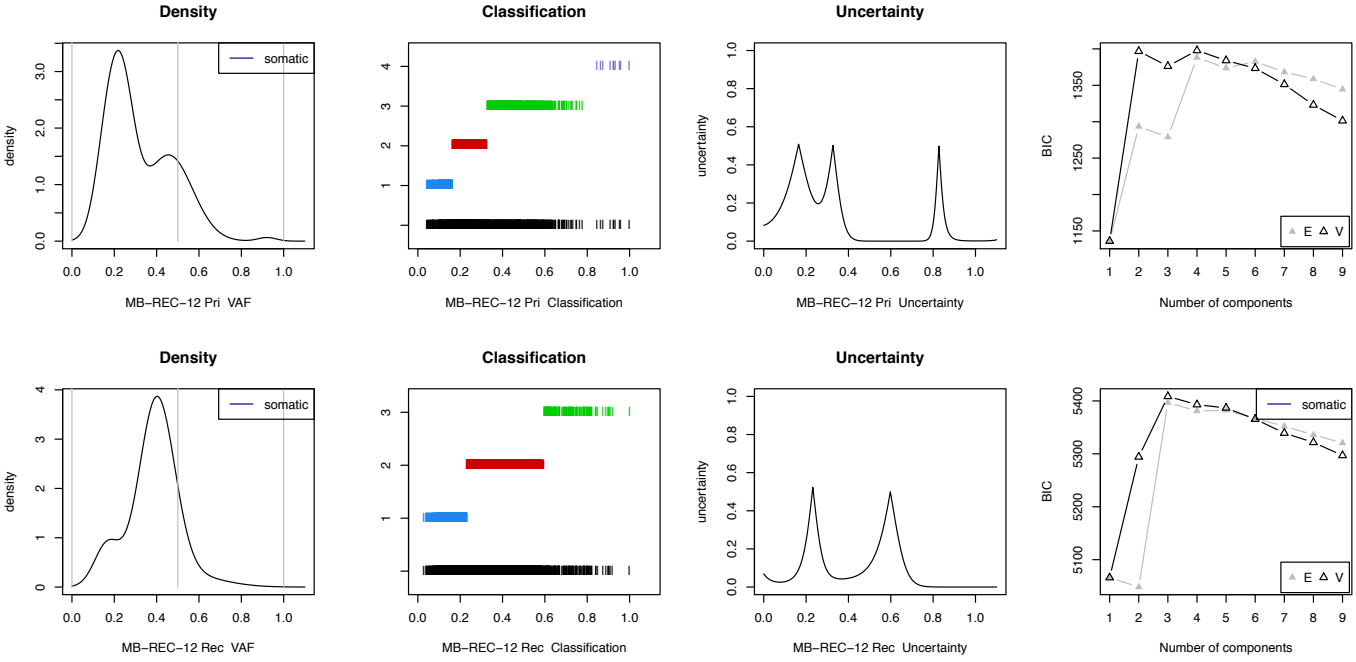
MB-REC-10



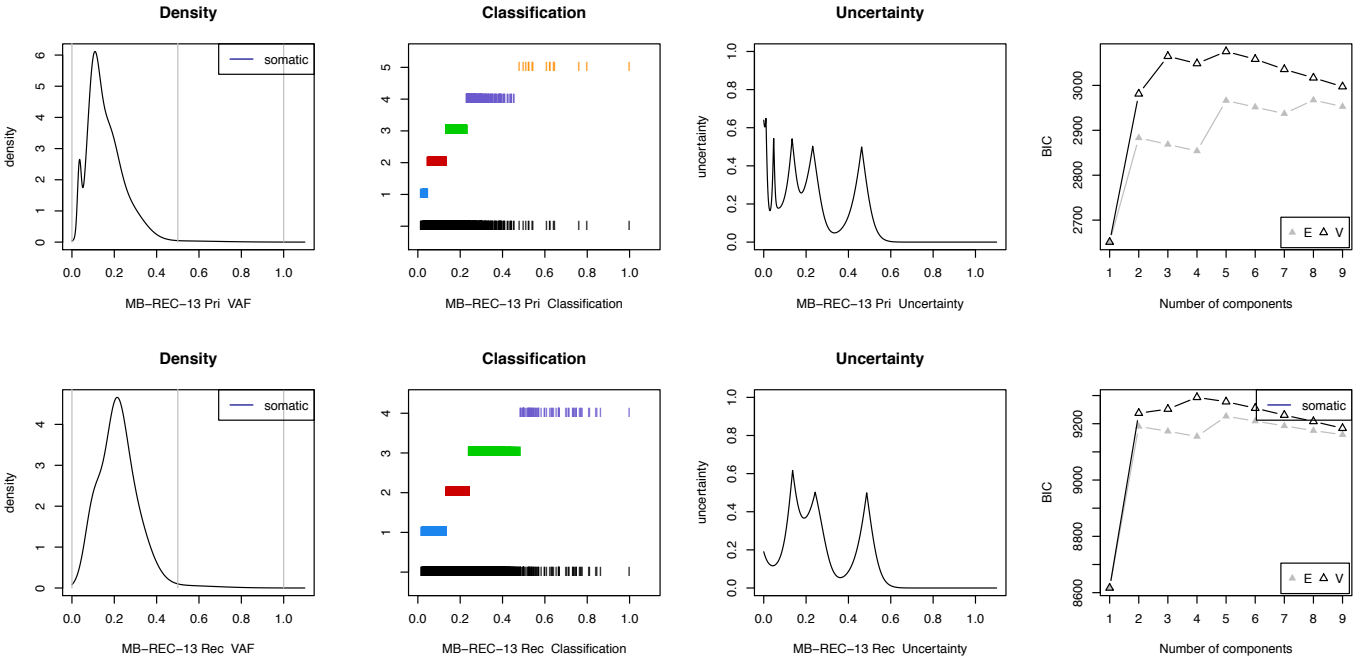
MB-REC-11



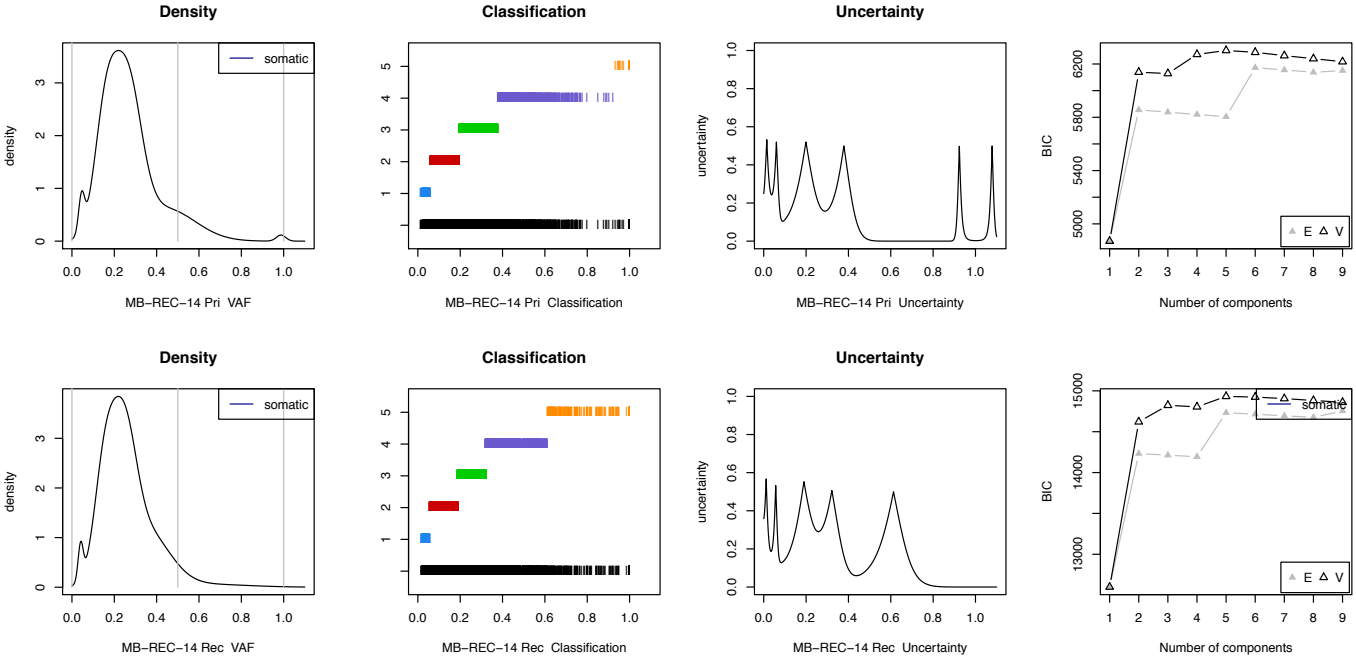
MB-REC-12



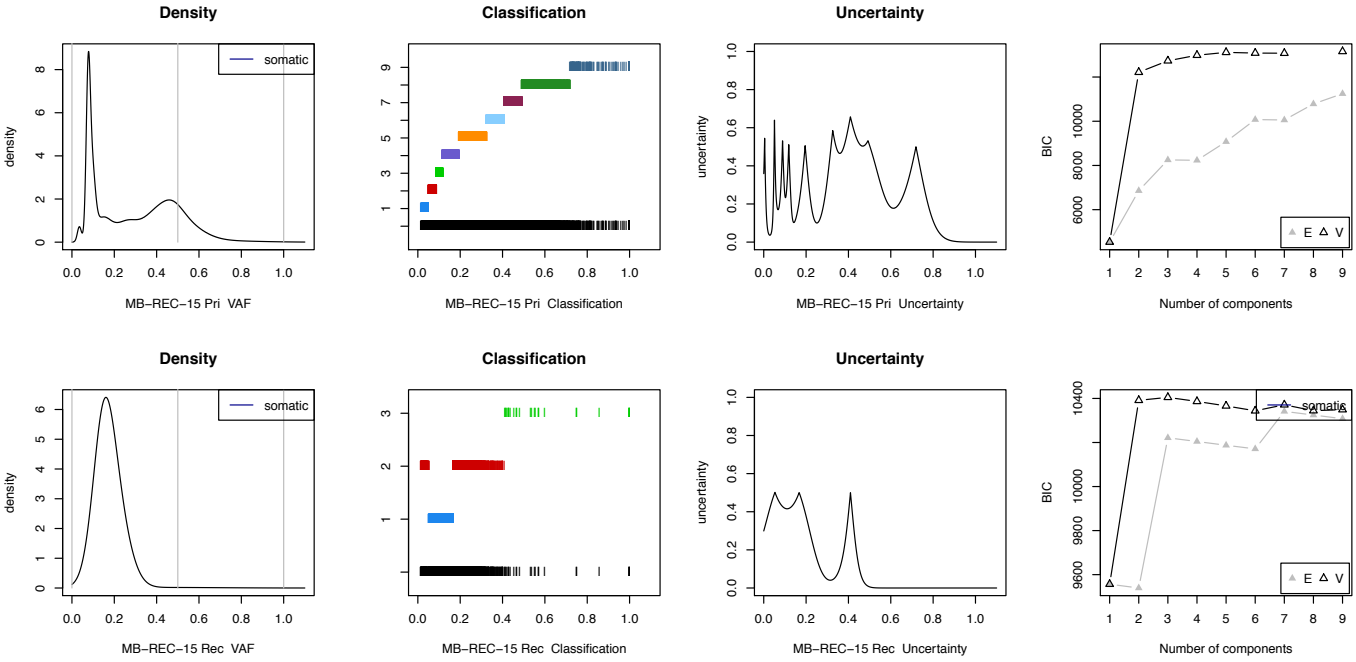
MB-REC-13



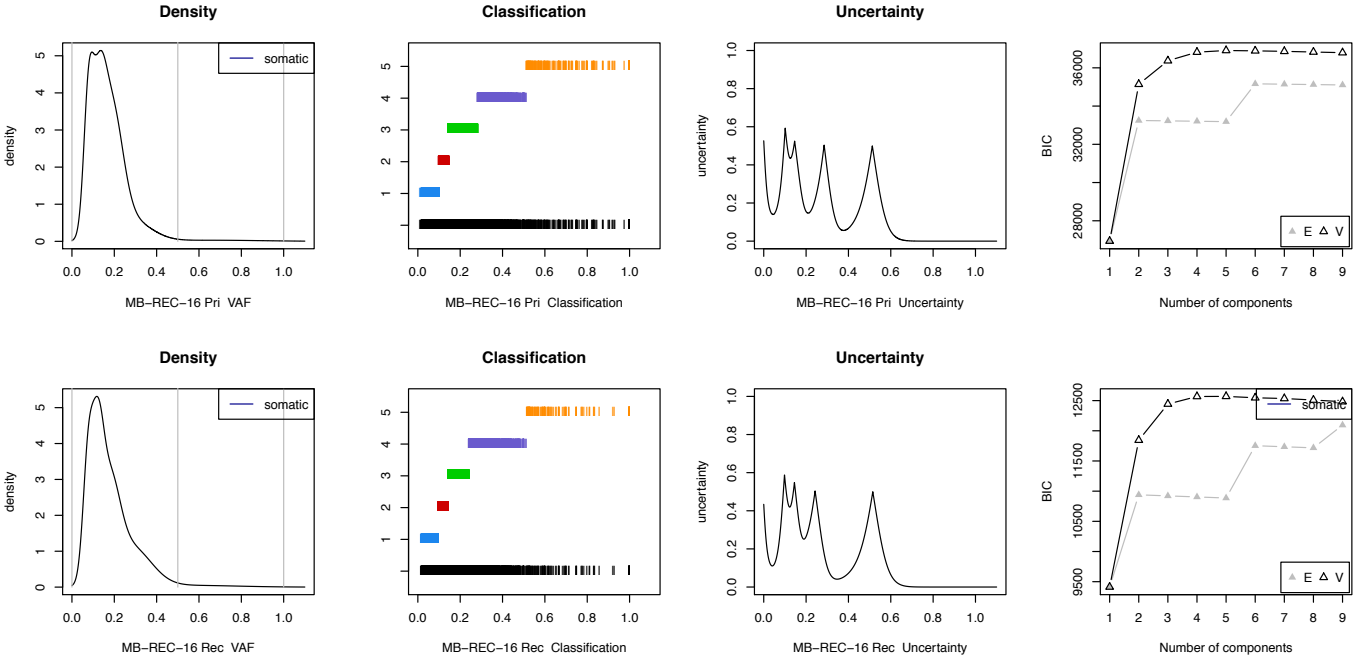
MB-REC-14



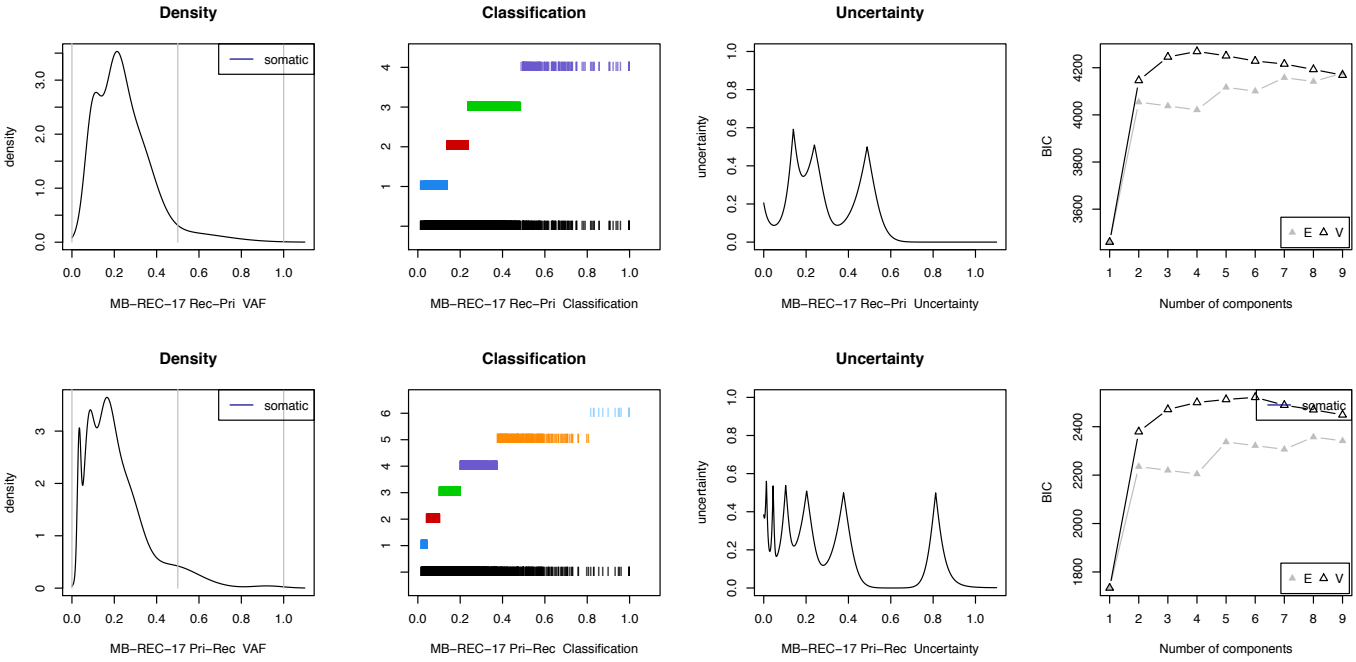
MB-REC-15



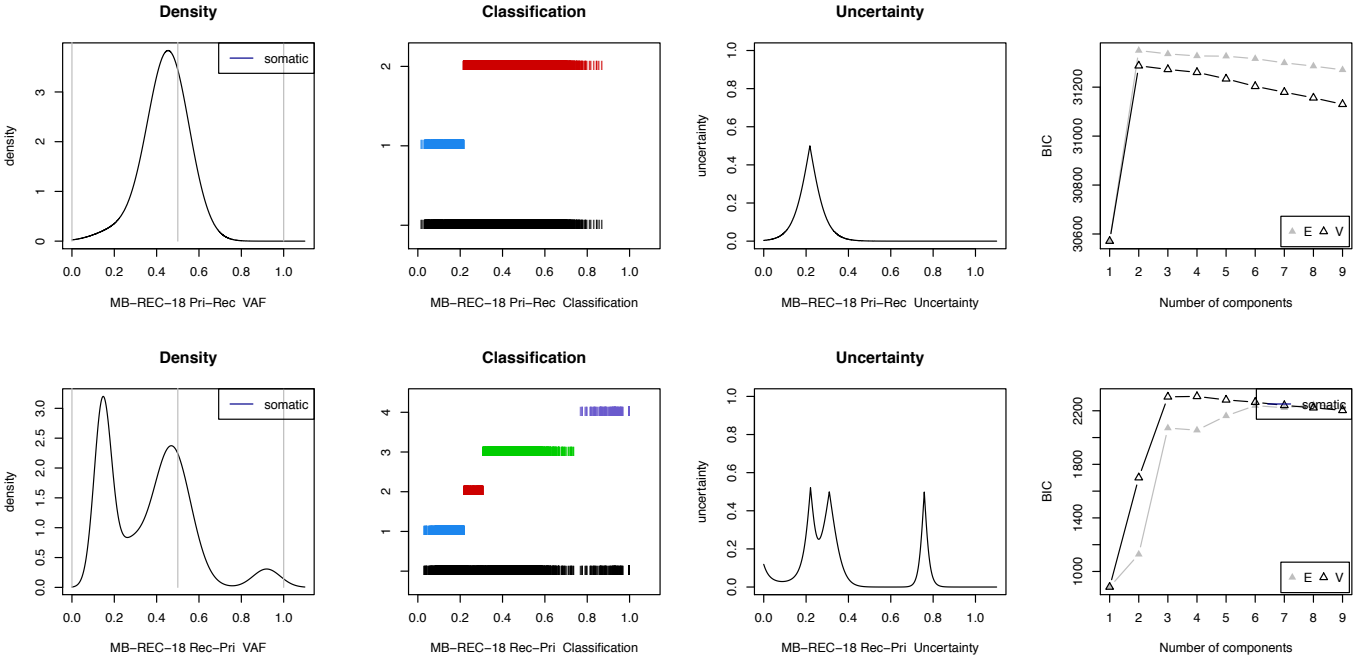
MB-REC-16



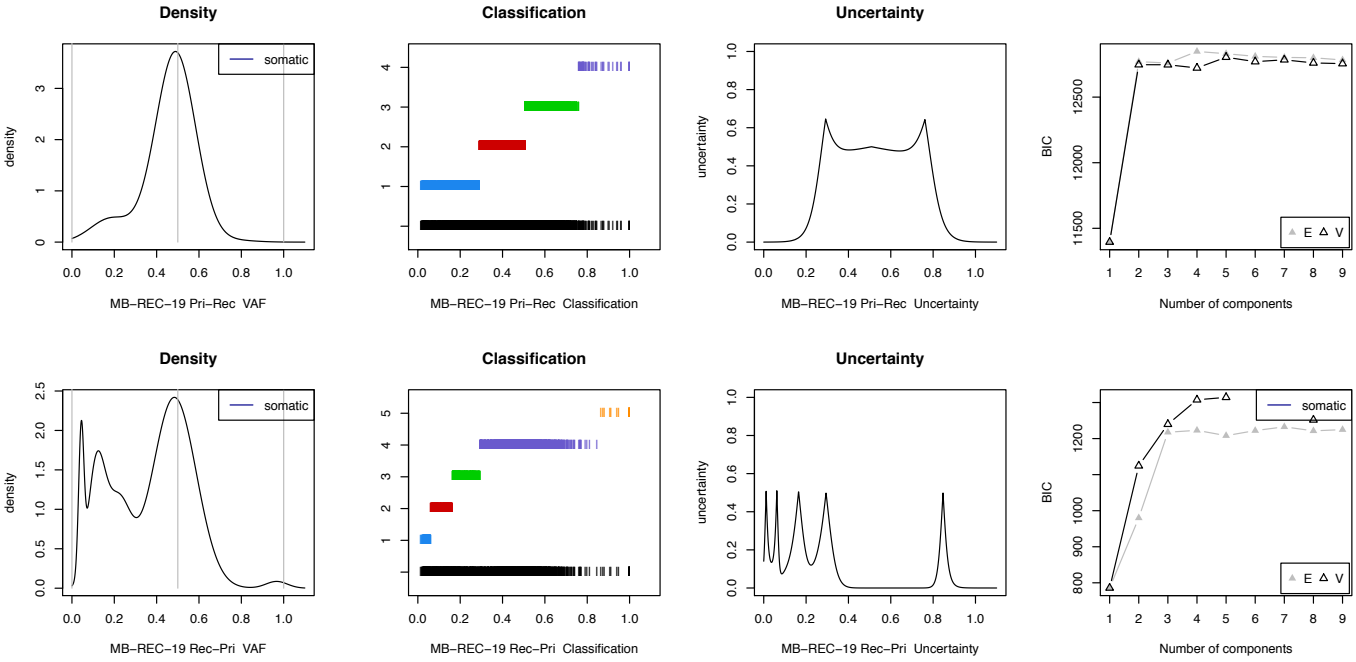
MB-REC-17



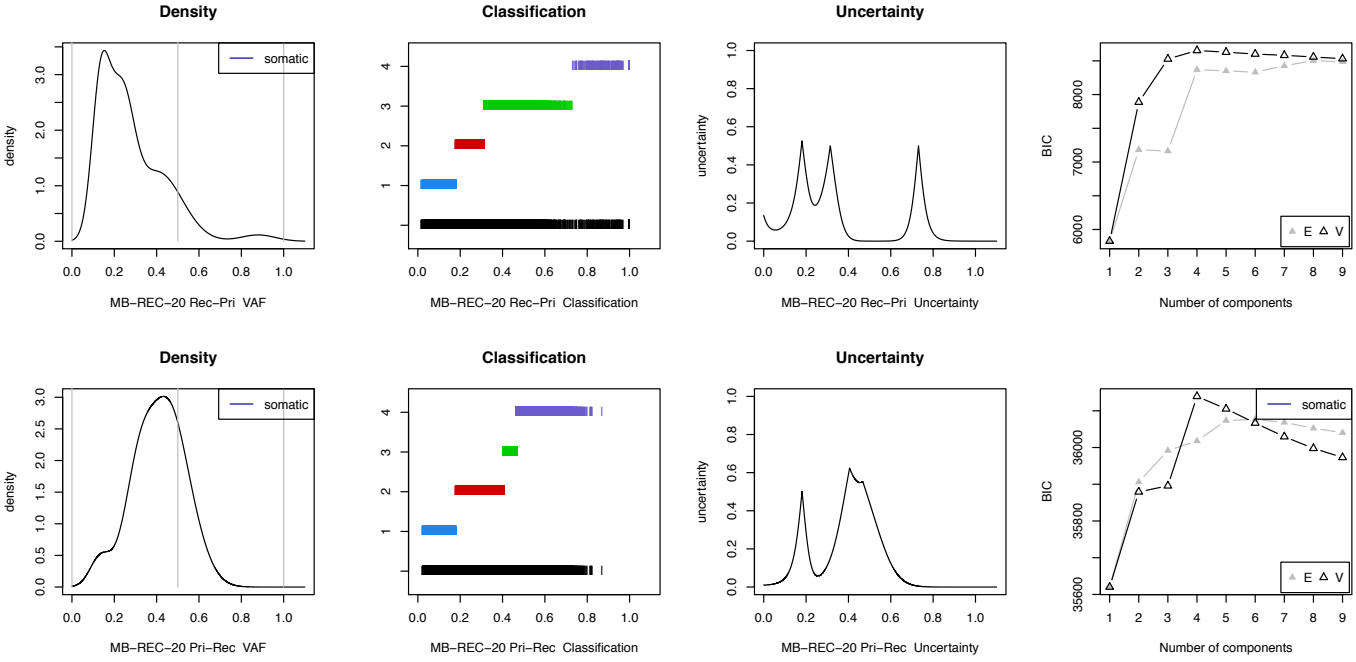
MB-REC-18



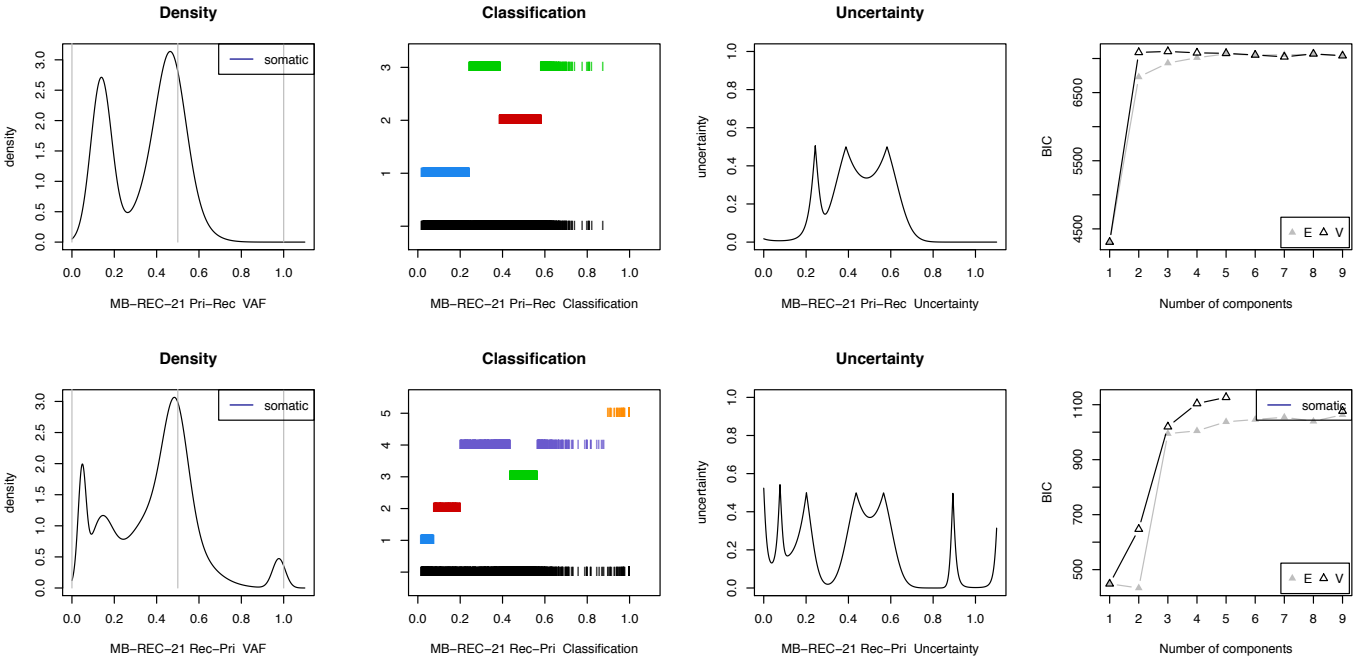
MB-REC-19



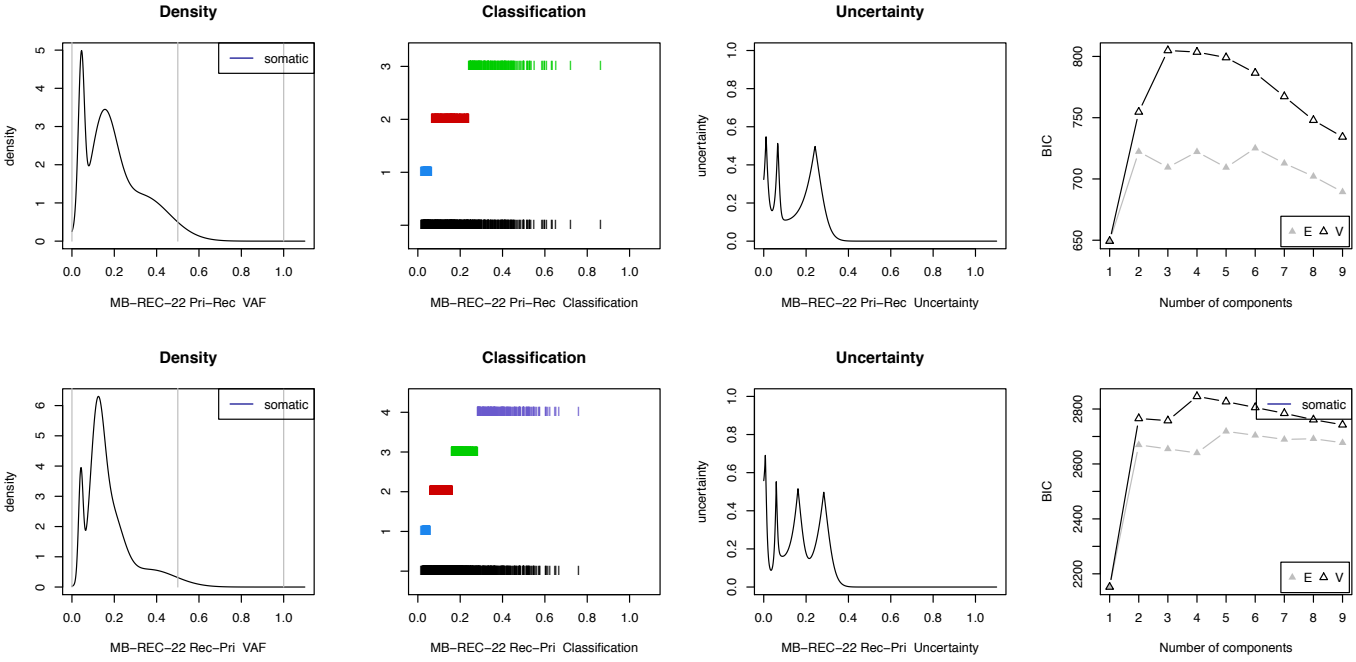
MB-REC-20



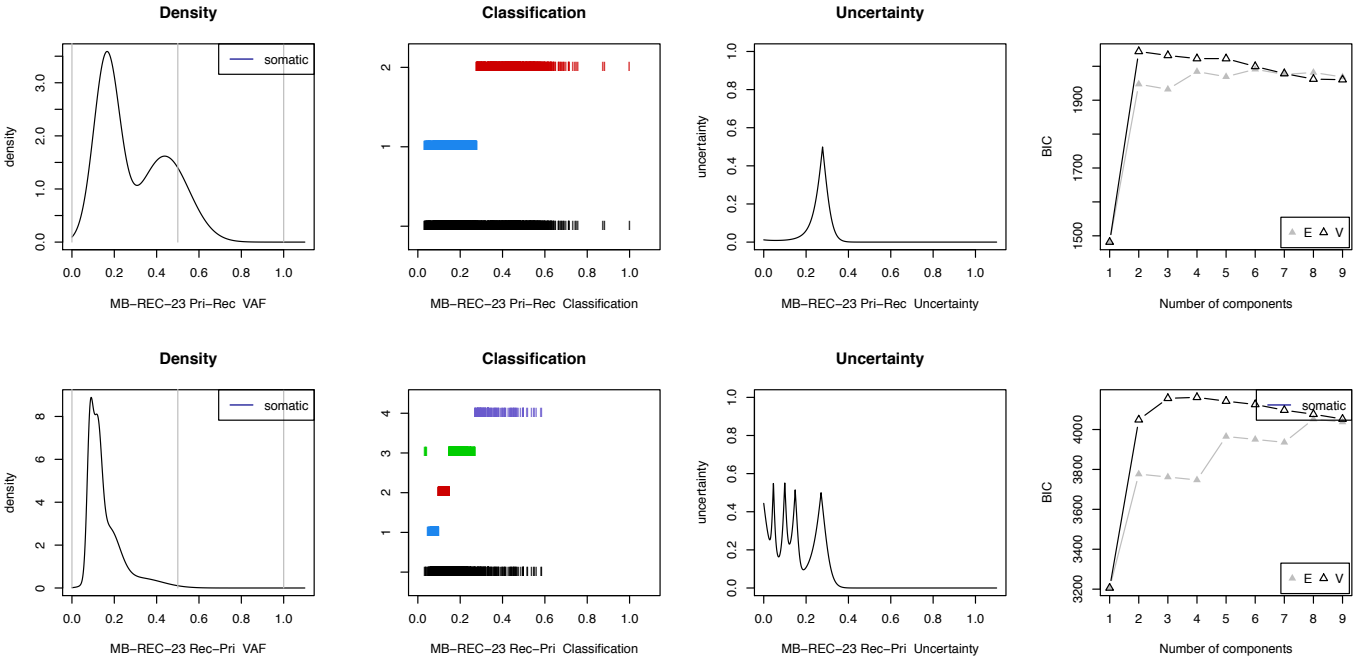
MB-REC-21



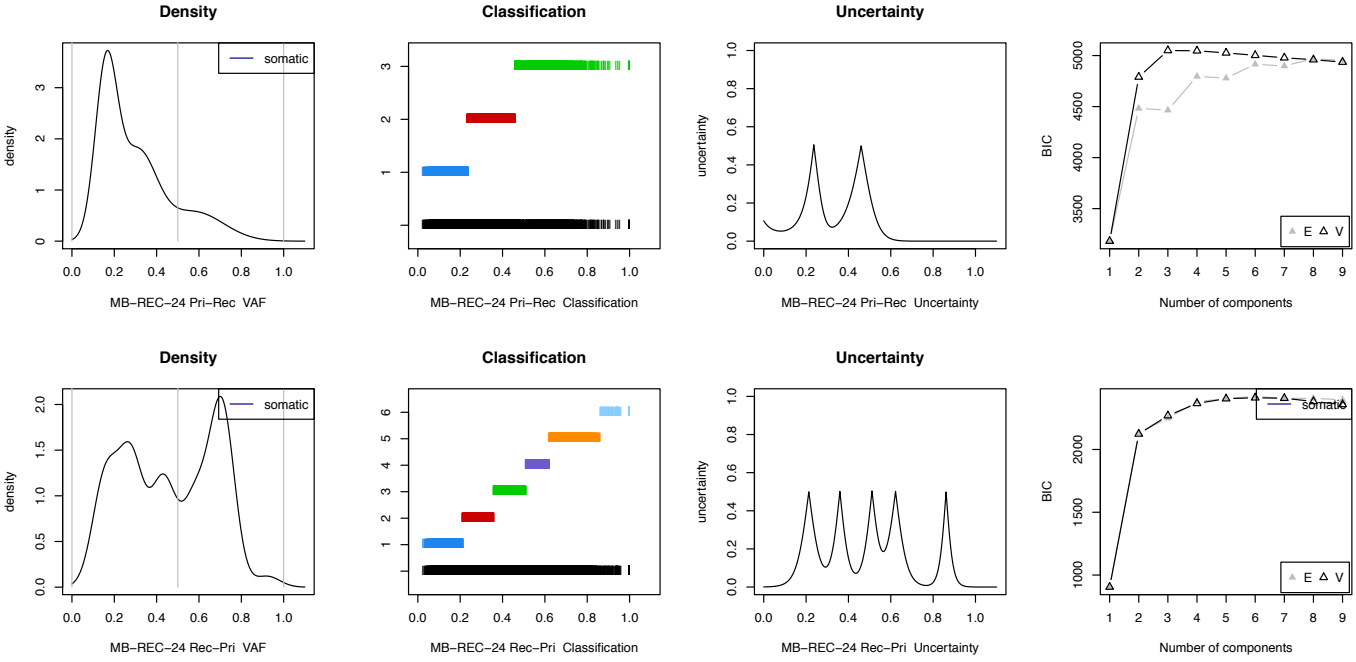
MB-REC-22



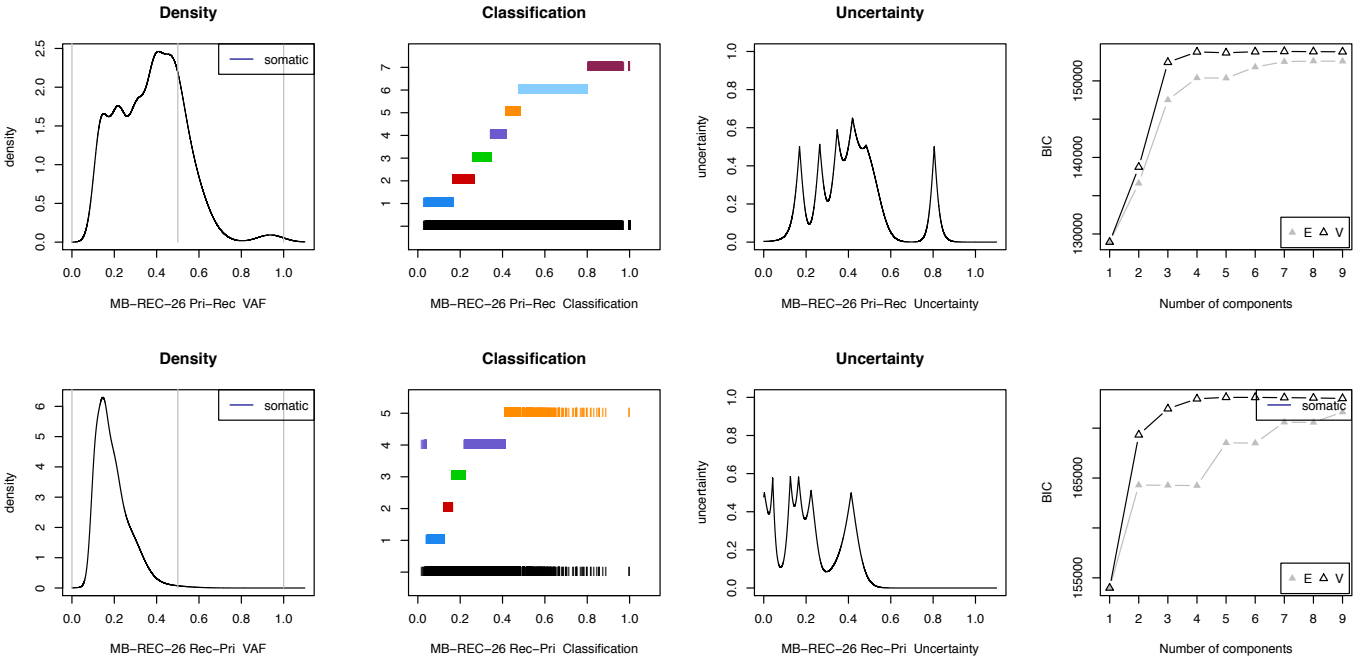
MB-REC-23



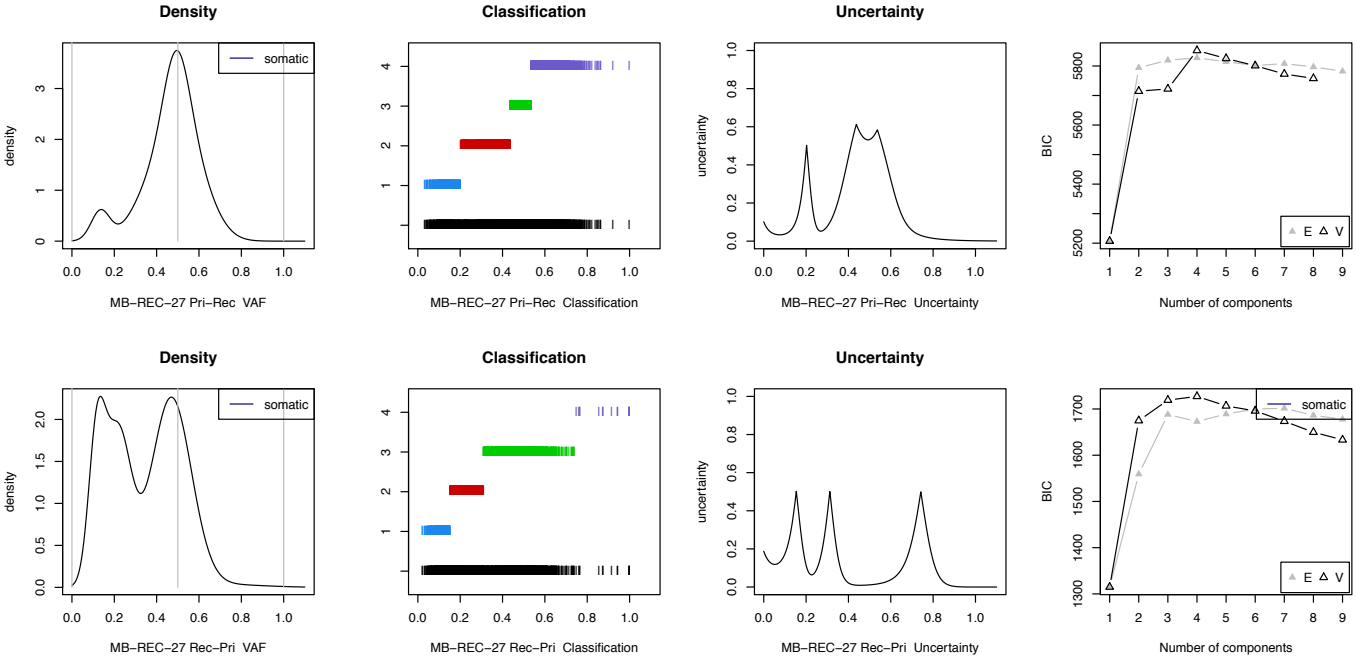
MB-REC-24



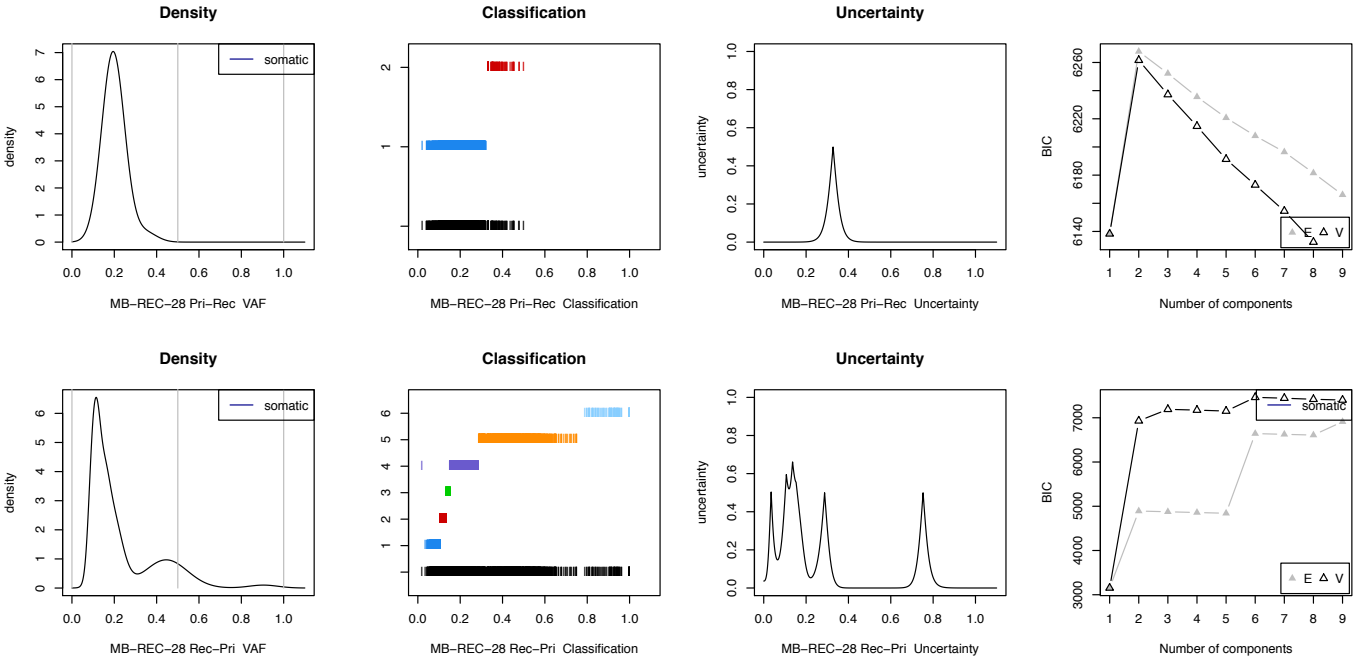
MB-REC-26



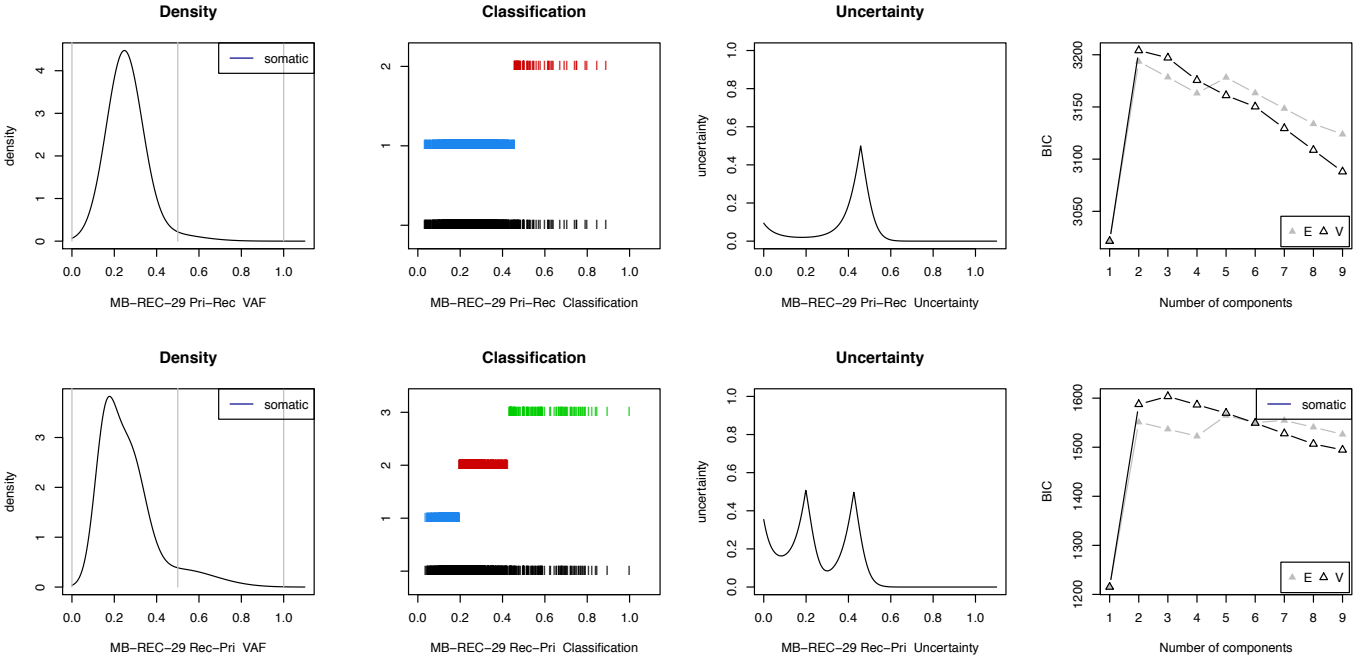
MB-REC-27



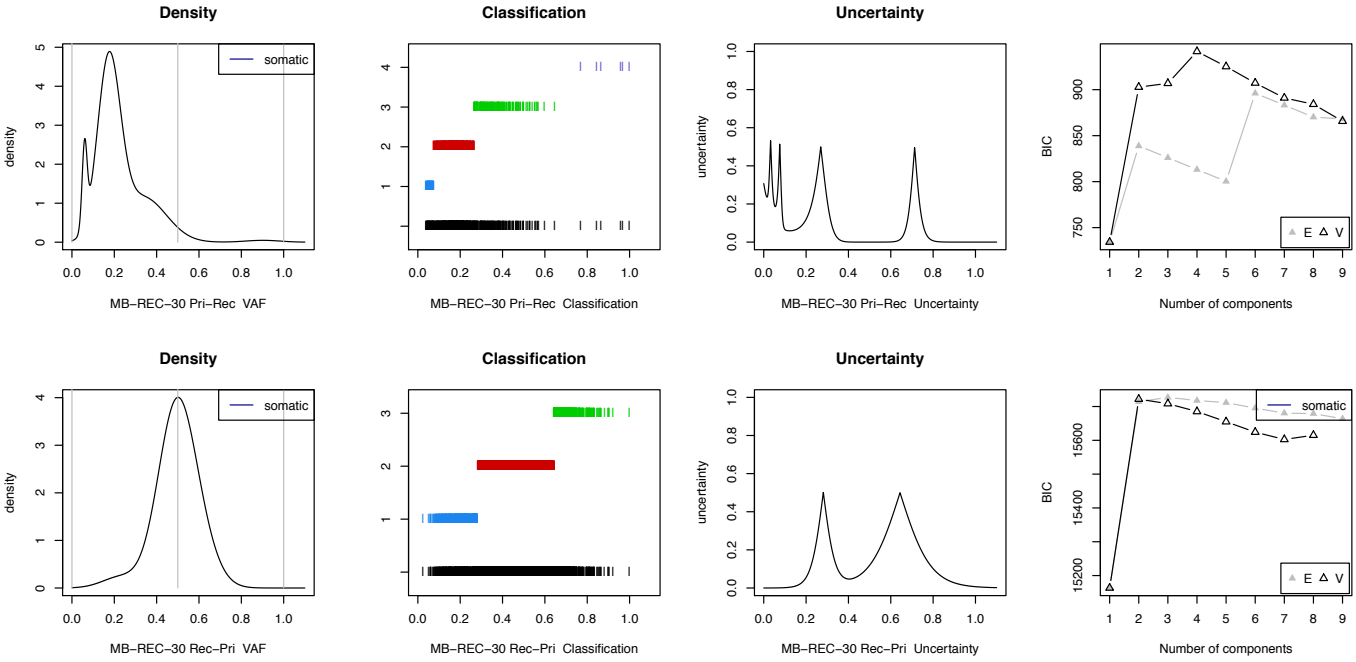
MB-REC-28



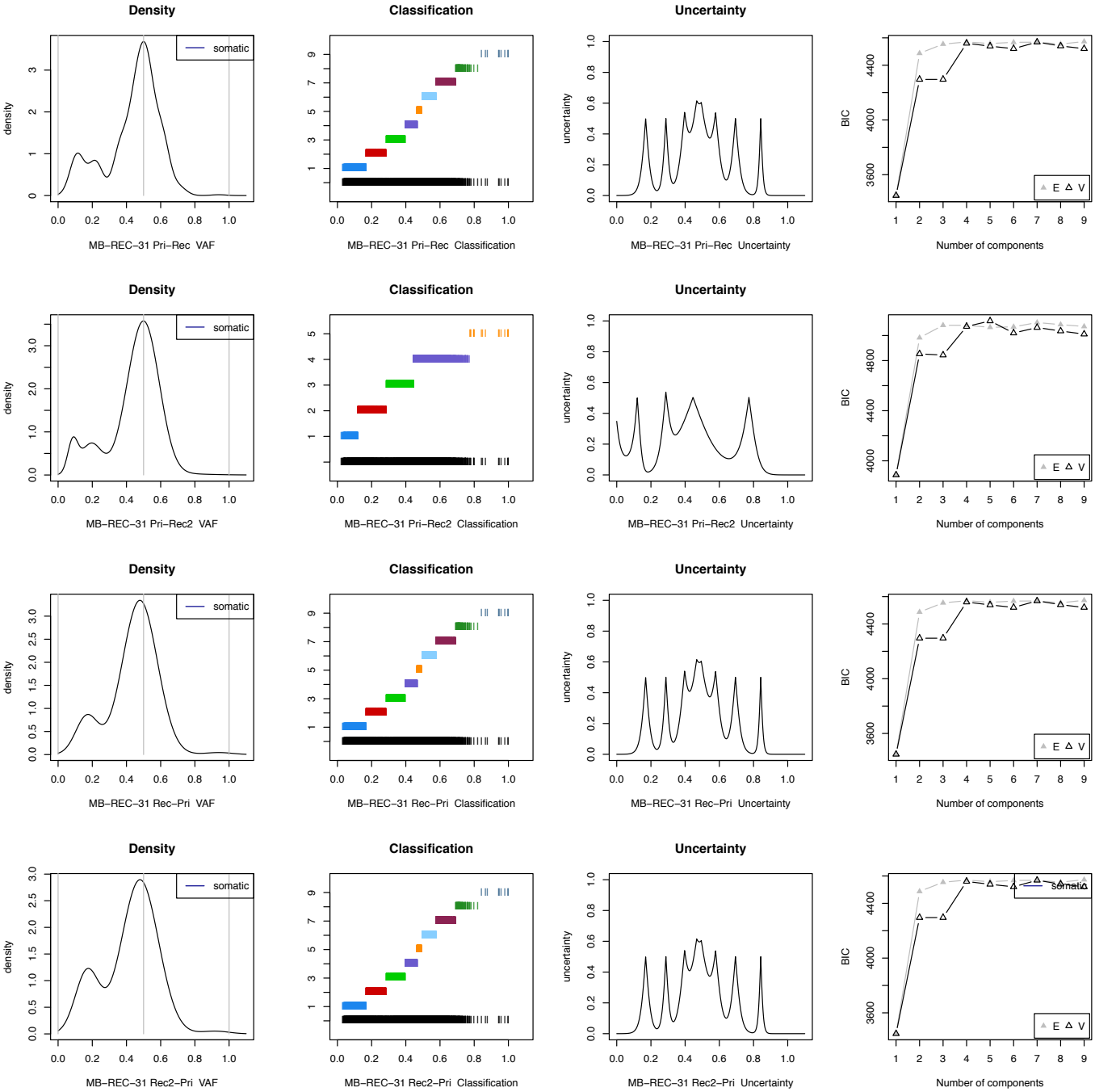
MB-REC-29



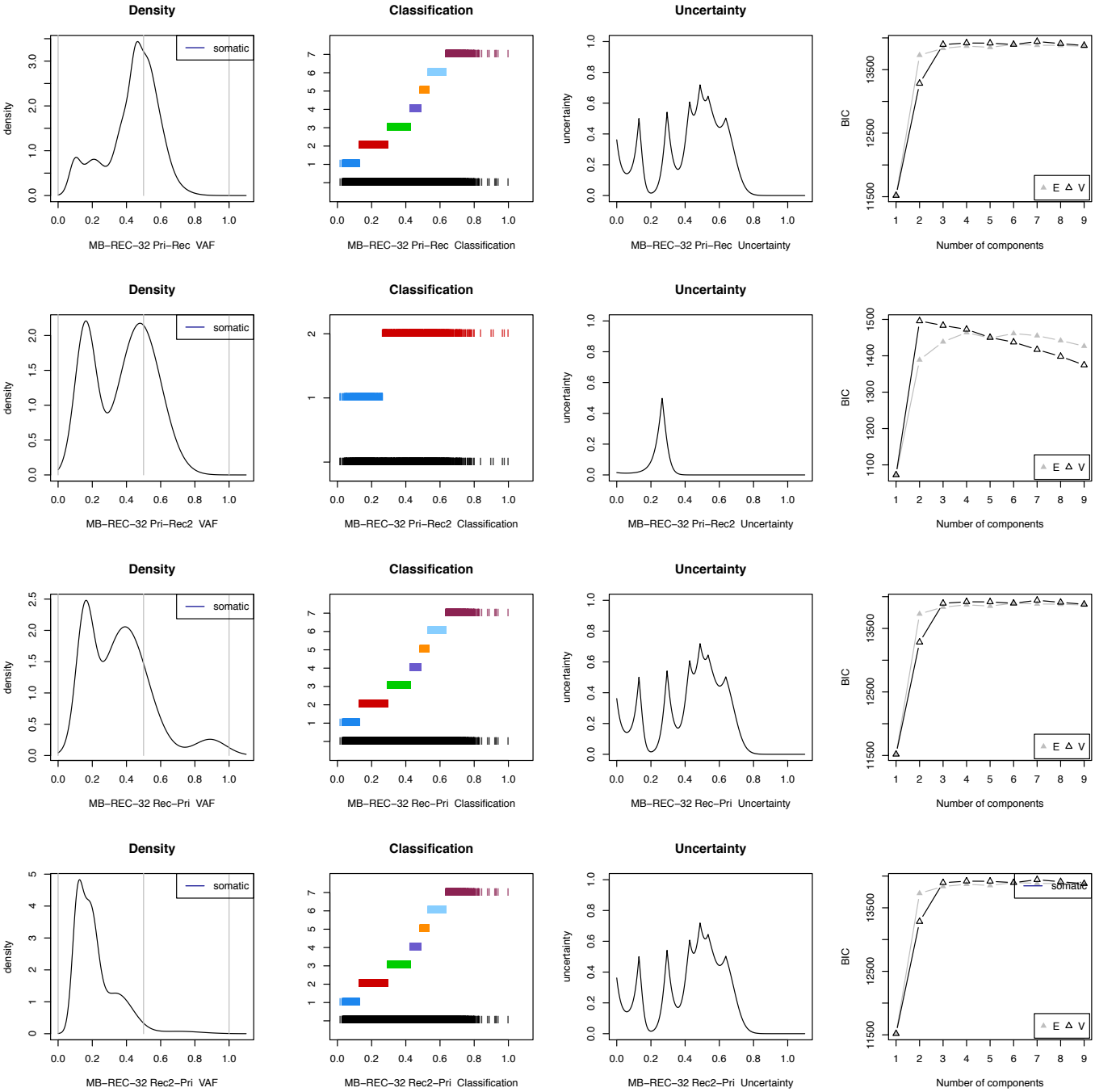
MB-REC-30


Density
somatic
density
MB–REC–30 Pri–Rec VAF
Classification
MB–REC–30 Pri–Rec Classification
Uncertainty
uncertainty
MB–REC–30 Pri–Rec Uncertainty
BIC
E Δ V
Number of components
Density
somatic
density
MB–REC–30 Rec–Pri VAF
Classification
MB–REC–30 Rec–Pri Classification
Uncertainty
uncertainty
MB–REC–30 Rec–Pri Uncertainty
BIC
somatic
E Δ V
Number of components

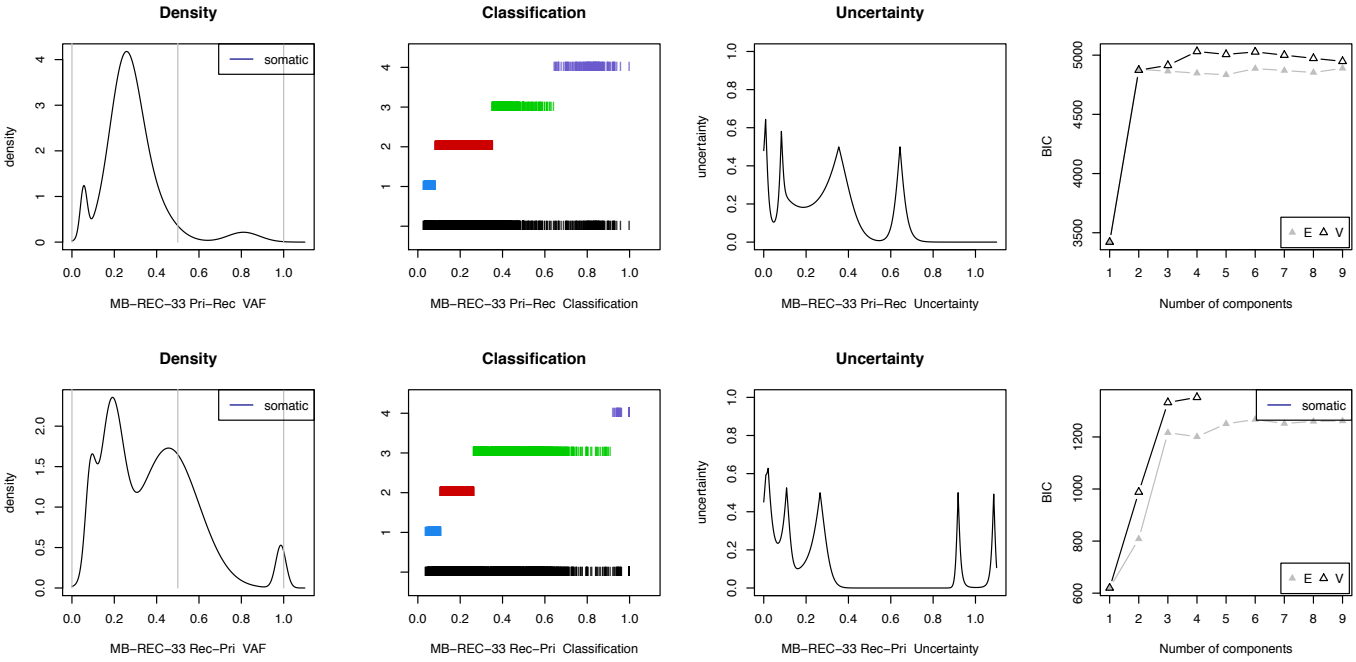
MB-REC-31


Density
Classification
Uncertainty
somatic
density
uncertainty
BIC
E V
MB−REC−31 Pri−Rec VAF
MB−REC−31 Pri−Rec Classification
MB−REC−31 Pri−Rec Uncertainty
Number of components
MB−REC−31 Pri−Rec2 VAF
MB−REC−31 Pri−Rec2 Classification
MB−REC−31 Pri−Rec2 Uncertainty
MB−REC−31 Rec−Pri VAF
MB−REC−31 Rec−Pri Classification
MB−REC−31 Rec−Pri Uncertainty
MB−REC−31 Rec2−Pri VAF
MB−REC−31 Rec2−Pri Classification
MB−REC−31 Rec2−Pri Uncertainty

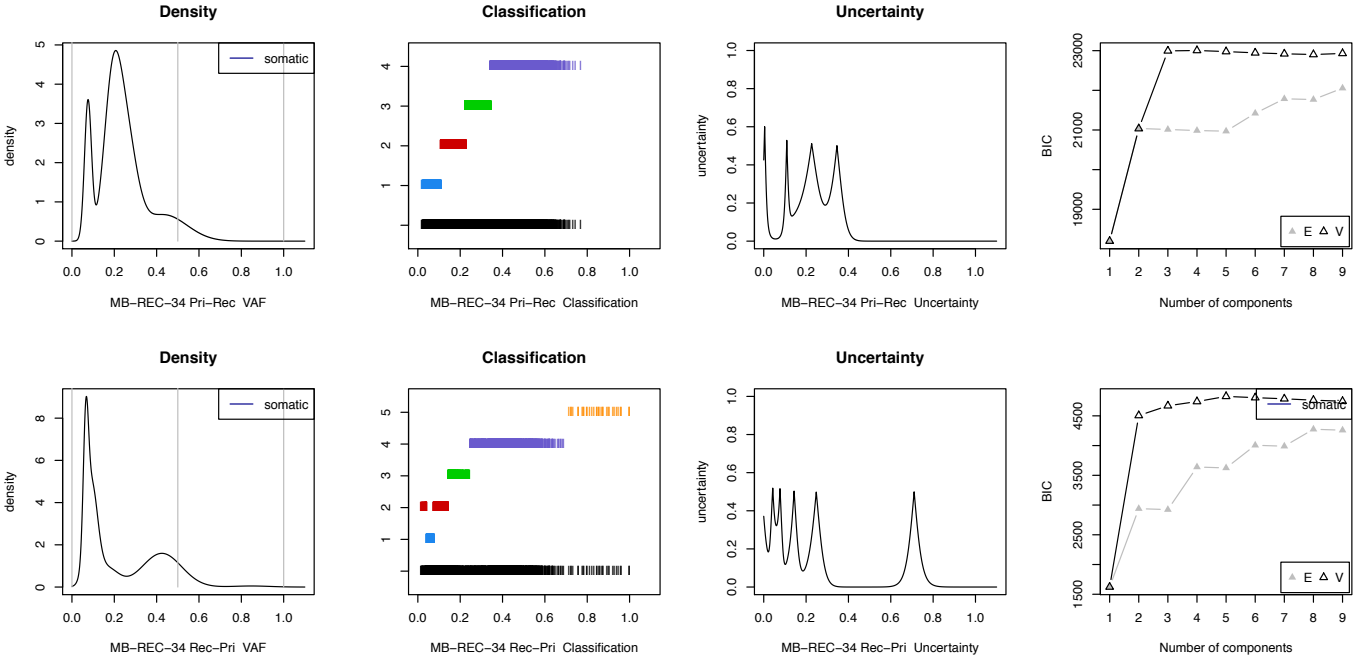
MB-REC-32


Density
Classification
Uncertainty
somatic
density
uncertainty
BIC
E V
MB−REC−32 Pri−Rec VAF
MB−REC−32 Pri−Rec Classification
MB−REC−32 Pri−Rec Uncertainty
Number of components
MB−REC−32 Pri−Rec2 VAF
MB−REC−32 Pri−Rec2 Classification
MB−REC−32 Pri−Rec2 Uncertainty
MB−REC−32 Rec−Pri VAF
MB−REC−32 Rec−Pri Classification
MB−REC−32 Rec−Pri Uncertainty
MB−REC−32 Rec2−Pri VAF
MB−REC−32 Rec2−Pri Classification
MB−REC−32 Rec2−Pri Uncertainty

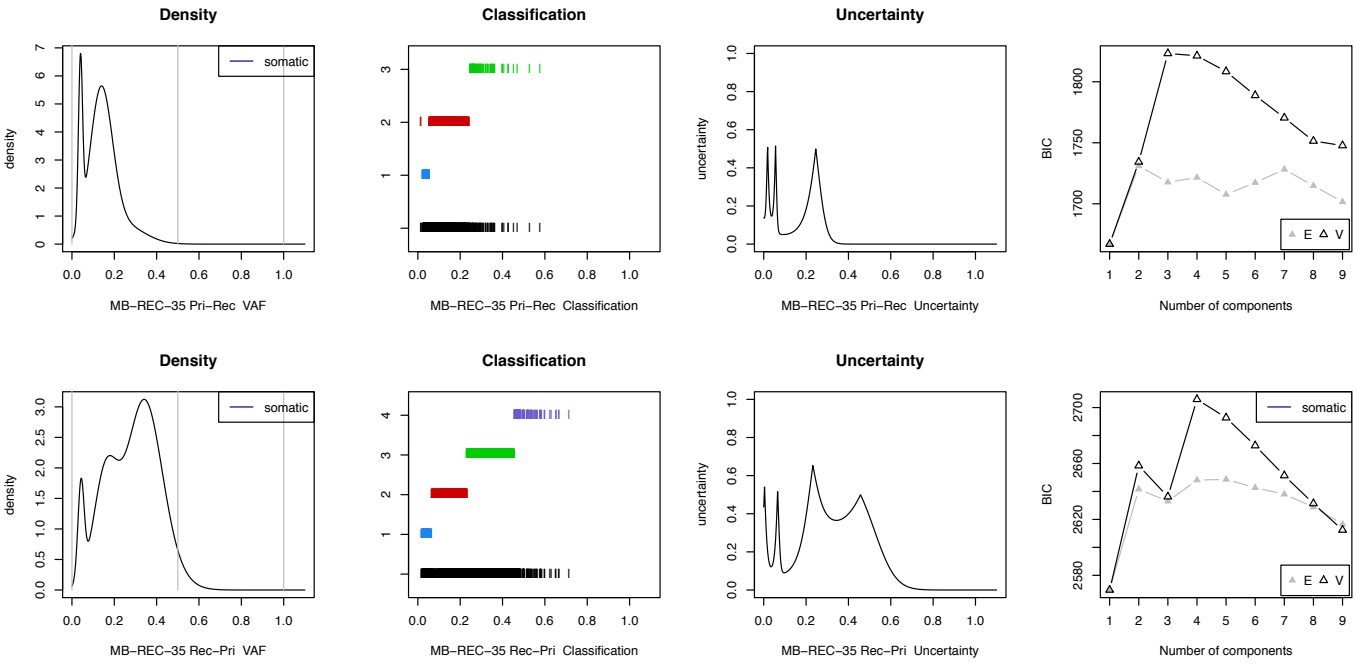
MB-REC-33



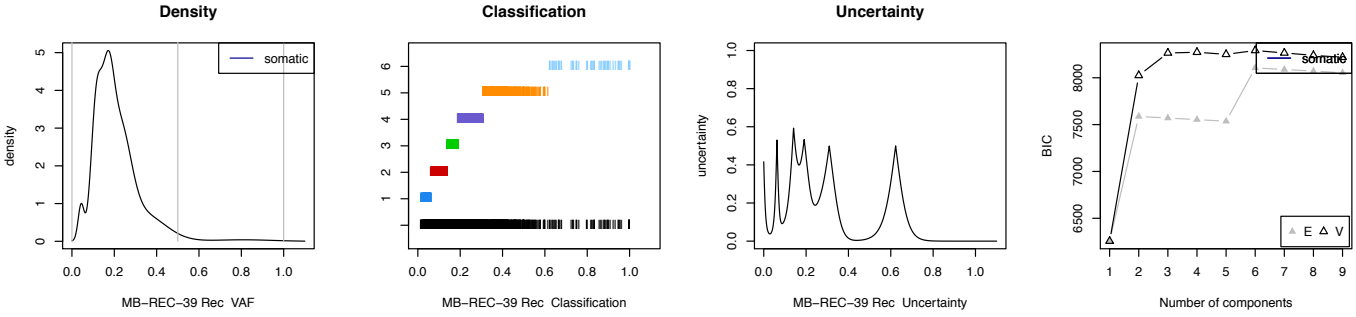
MB-REC-34



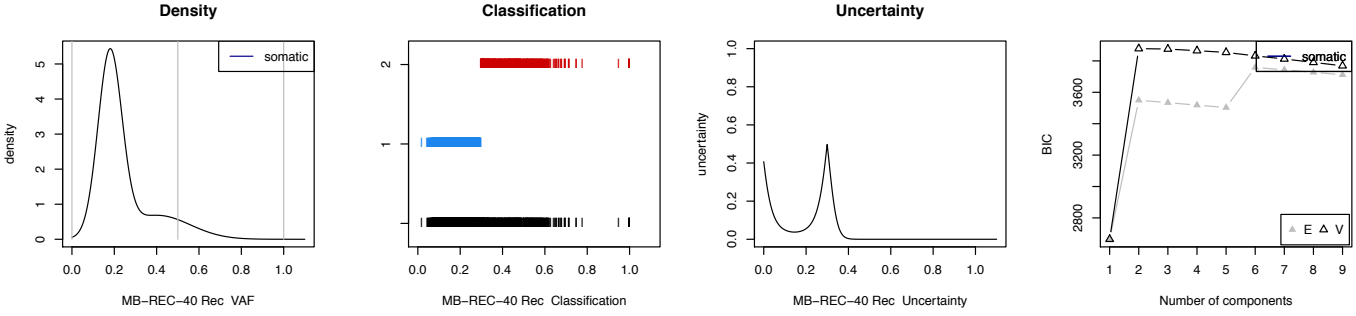
MB-REC-35



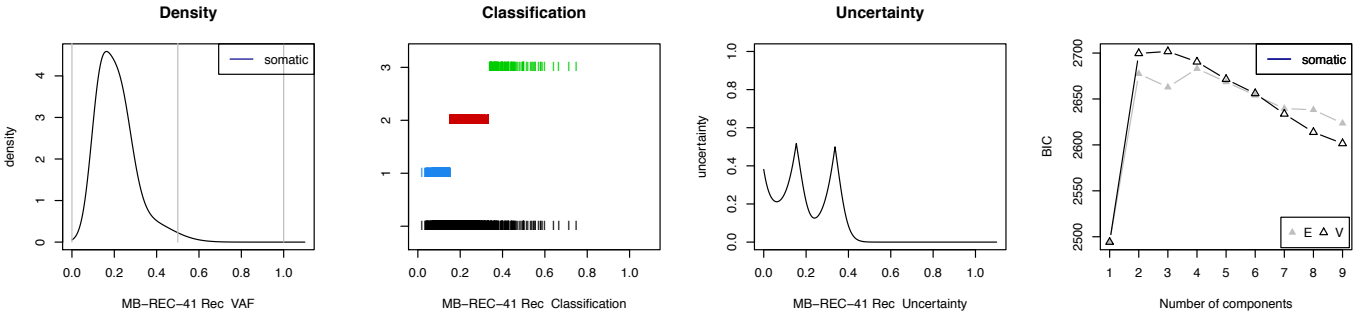
MB-REC-39



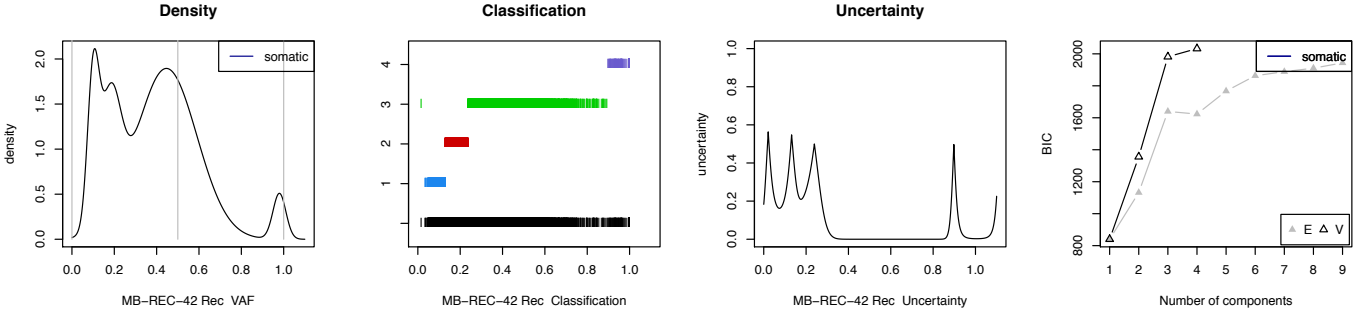
MB-REC-40



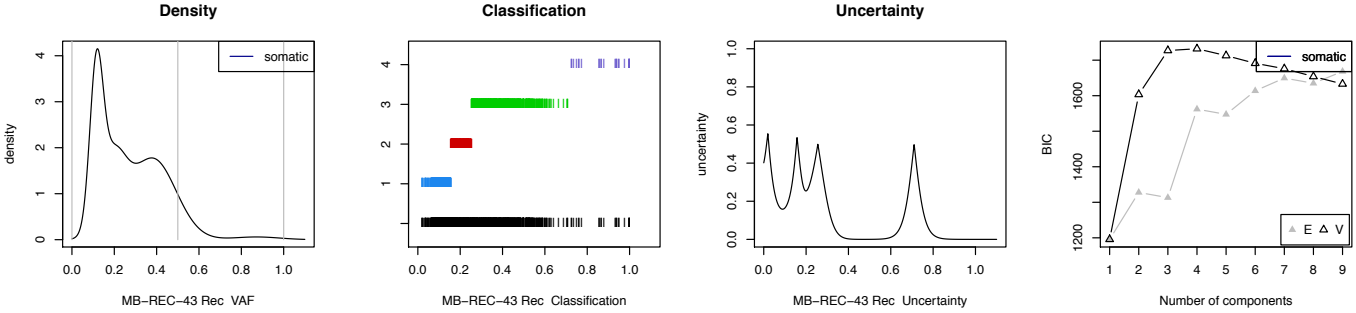
MB-REC-41



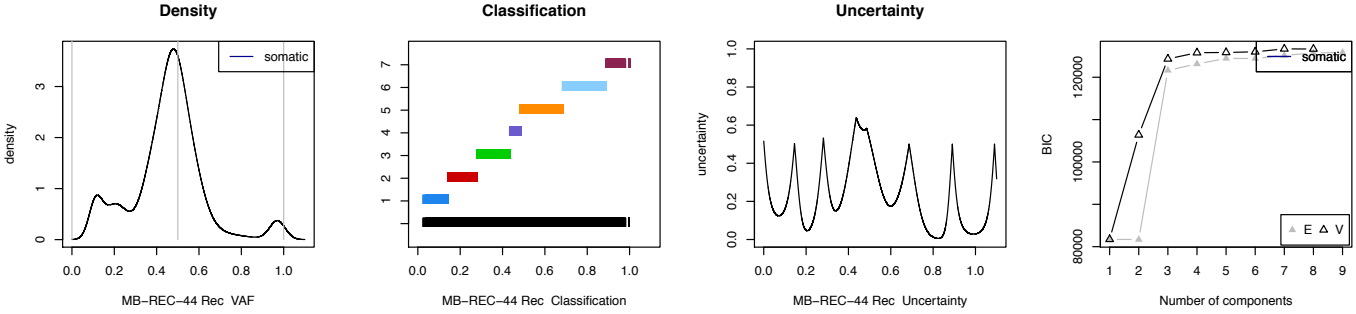
MB-REC-42



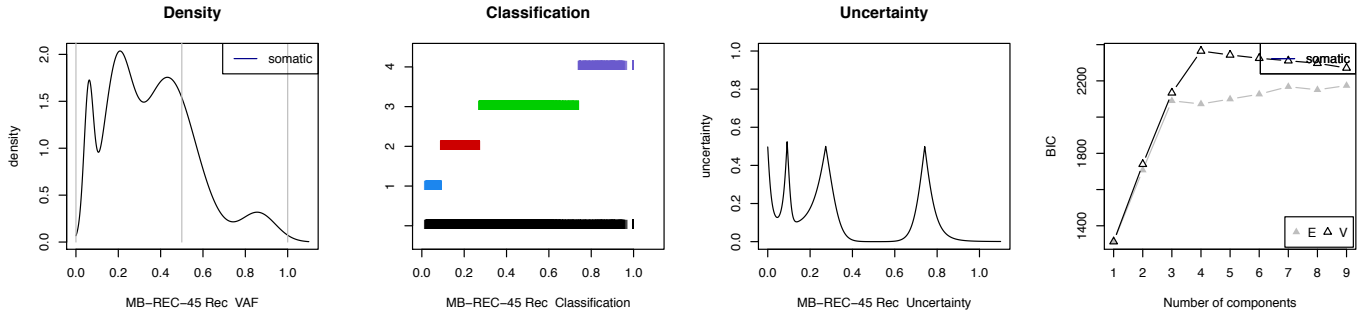
MB-REC-43



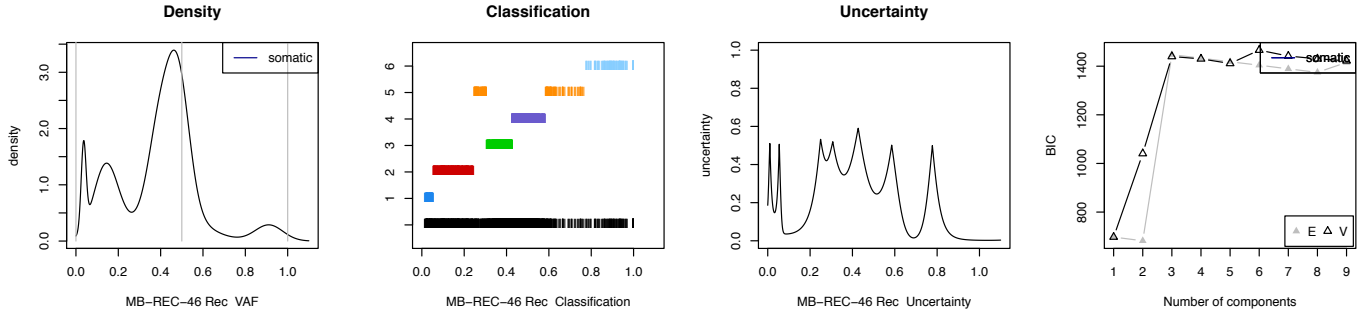
MB-REC-44



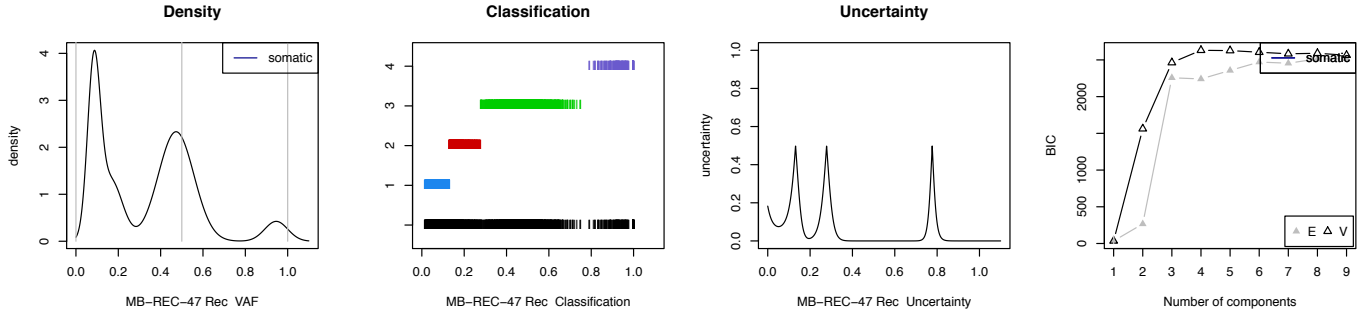
MB-REC-45



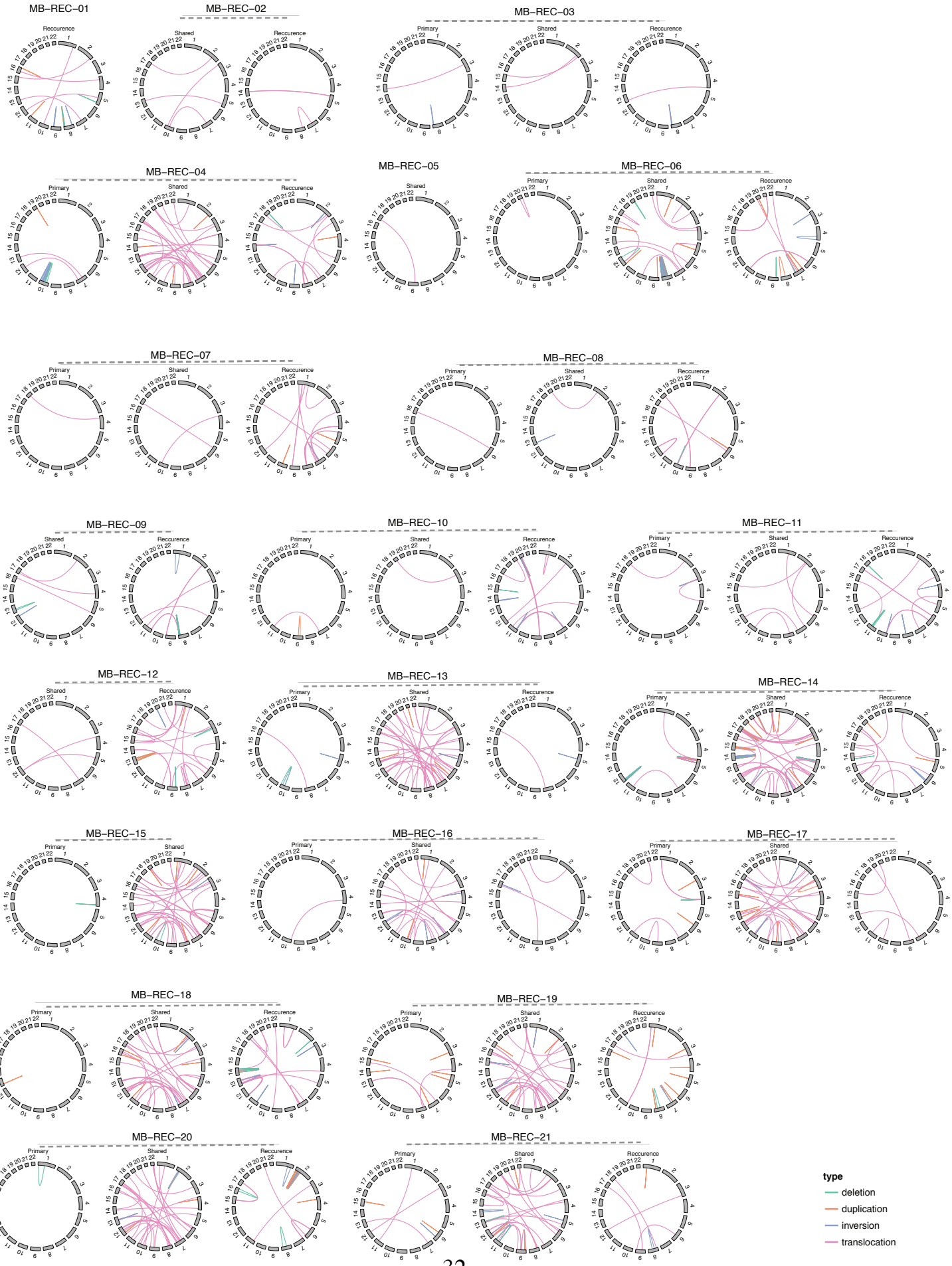
MB-REC-46

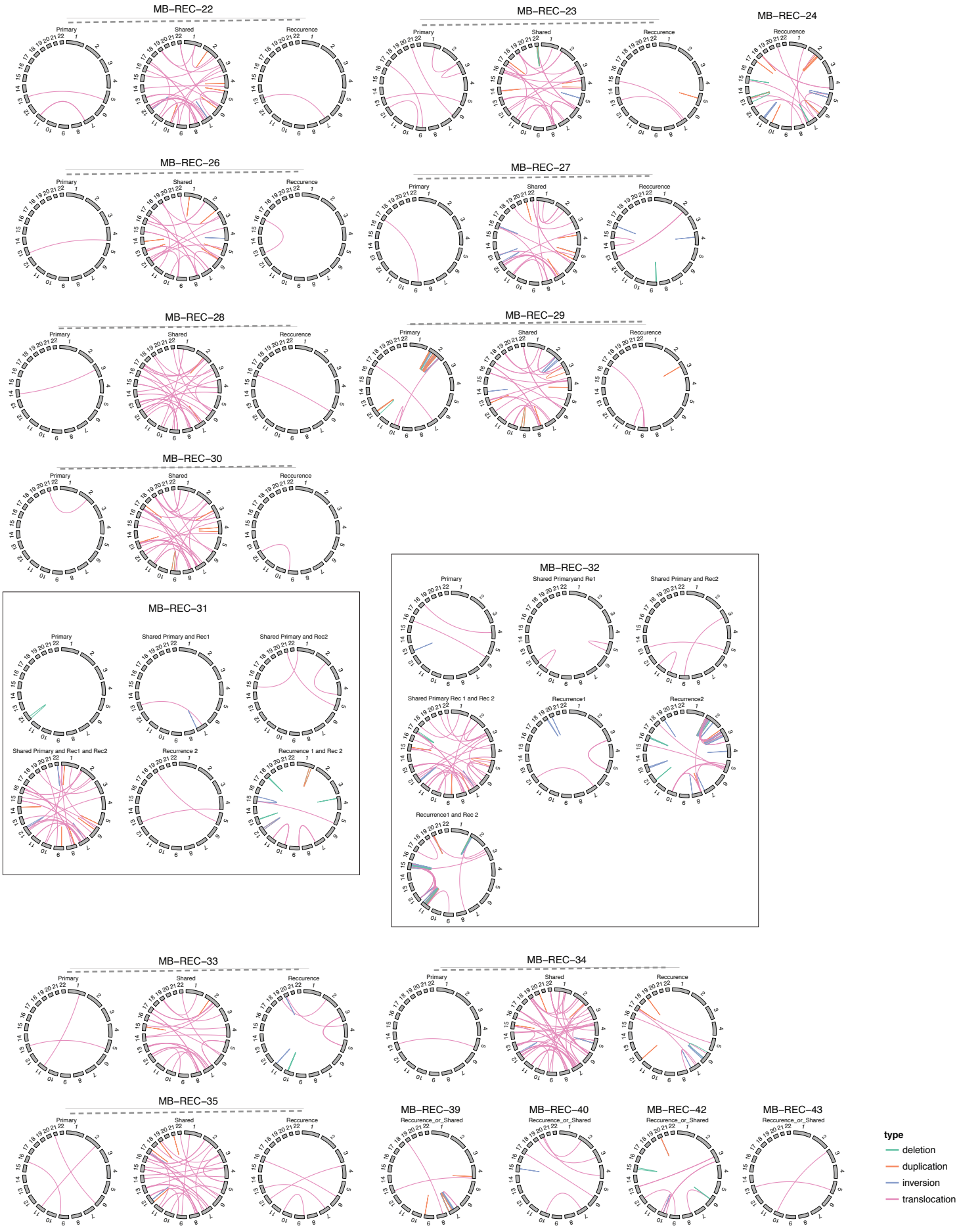


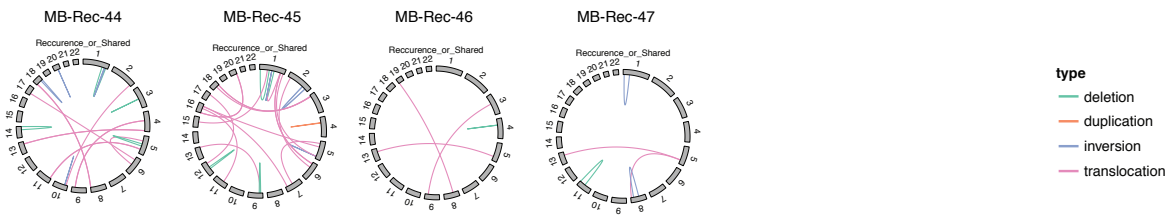
MB-REC-47



Structural aberrations: Circos plots demonstrating compartment-specific somatic structural variants (SV) specific to the untreated tumour, the recurrent tumour, or shared across compartments. Lines denote the breakpoints of translocations, inversions, deletions, and duplications. In nearly every patient in the cohort, post-therapy samples harboured new structural events not identified in the diagnostic sample.

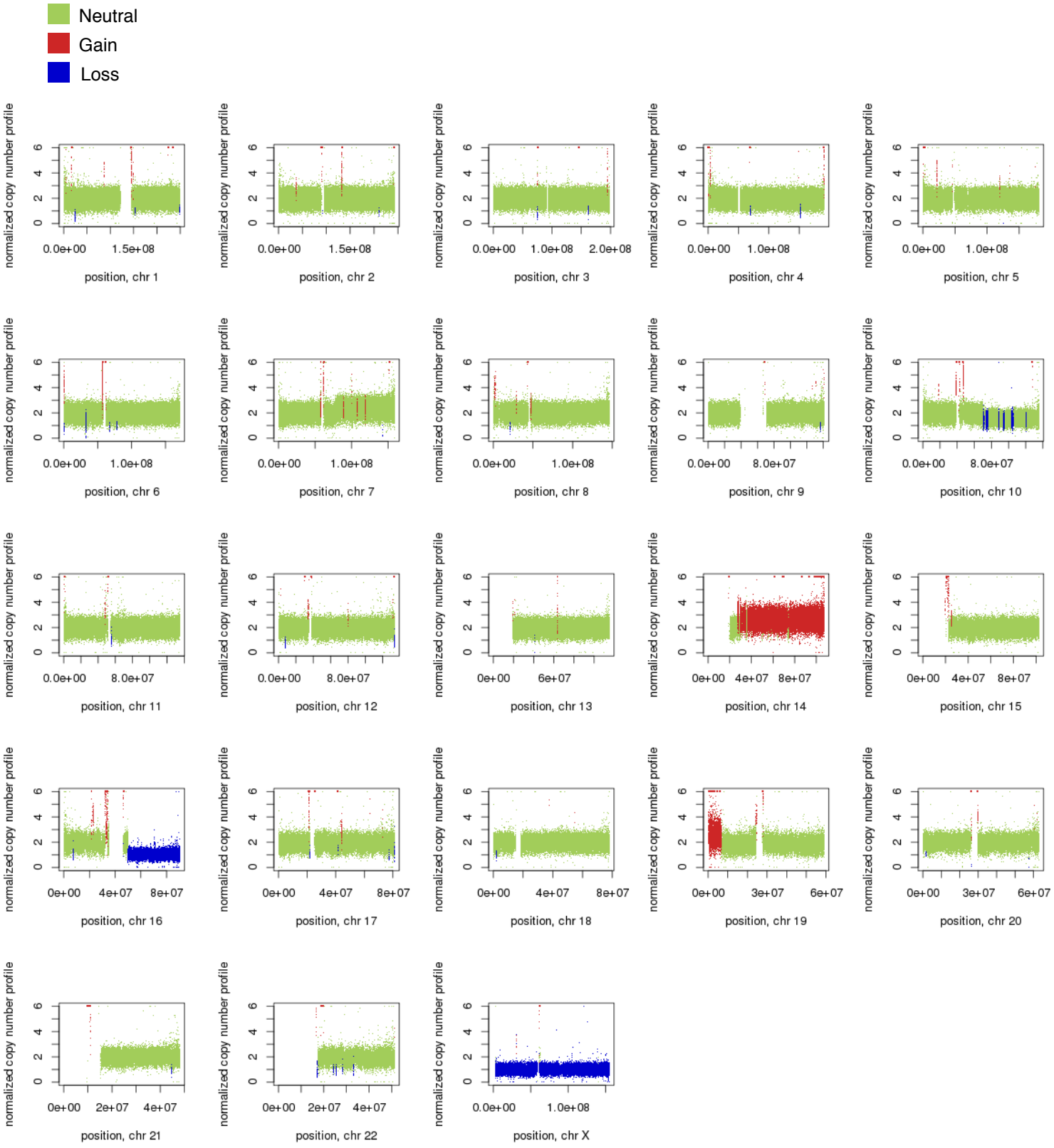

MB–REC–01
Reccurence
MB–REC–02
Shared
Reccurence
MB–REC–03
Primary
Shared
Reccurence
MB–REC–04
Primary
Shared
Reccurence
MB–REC–05
Shared
MB–REC–06
Primary
Shared
Reccurence
MB–REC–07
Primary
Shared
Reccurence
MB–REC–08
Primary
Shared
Reccurence
MB–REC–09
Shared
Reccurence
MB–REC–10
Primary
Shared
Reccurence
MB–REC–11
Primary
Shared
Reccurence
MB–REC–12
Shared
Reccurence
MB–REC–13
Primary
Shared
Reccurence
MB–REC–14
Primary
Shared
Reccurence
MB–REC–15
Primary
Shared
MB–REC–16
Primary
Shared
MB–REC–17
Primary
Shared
MB–REC–18
Primary
Shared
Reccurence
MB–REC–19
Primary
Shared
Reccurence
MB–REC–20
Primary
Shared
Reccurence
MB–REC–21
Primary
Shared
Reccurence
type
deletion
duplication
inversion
translocation


MB–REC–22
Primary
Shared
Reccurence
MB–REC–23
Primary
Shared
Reccurence
MB–REC–24
Reccurence
MB–REC–26
Primary
Shared
Reccurence
MB–REC–27
Primary
Shared
Reccurence
MB–REC–28
Primary
Shared
Reccurence
MB–REC–29
Primary
Shared
Reccurence
MB–REC–30
Primary
Shared
Reccurence
MB–REC–31
Primary
Shared Primary and Rec1
Shared Primary and Rec2
Shared Primary and Rec1 and Rec2
Recurrence 2
Recurrence 1 and Rec 2
MB–REC–32
Primary
Shared Primaryand Re1
Shared Primary and Rec2
Shared Primary Rec 1 and Rec 2
Recurrence1
Recurrence2
Recurrence1 and Rec 2
MB–REC–33
Primary
Shared
Reccurence
MB–REC–34
Primary
Shared
Reccurence
MB–REC–35
Primary
Shared
Reccurence
MB–REC–39
Reccurence_or_Shared
MB–REC–40
Reccurence_or_Shared
MB–REC–42
Reccurence_or_Shared
MB–REC–43
Reccurence_or_Shared
type
deletion
duplication
inversion
translocation

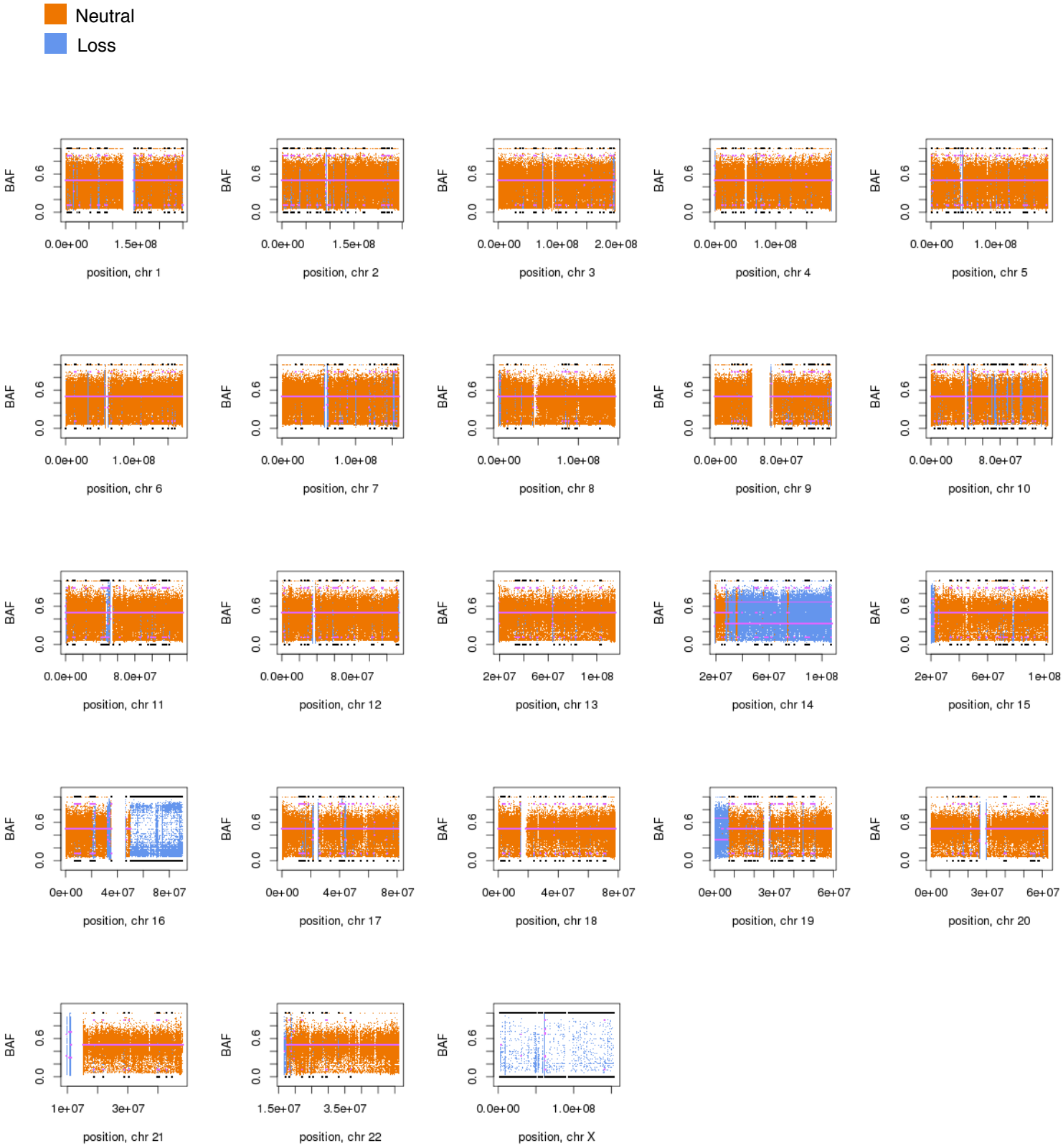

MB-Rec-44
MB-Rec-45
MB-Rec-46
MB-Rec-47
Reccurence_or_Shared
type
deletion
duplication
inversion
translocation

Control-FreeC CNV and LOH plots: Copy number and loss of heterozygosity (LOH) was determined for each tumor sample with a patient-matched germline sample (n=14). Per-chromosome copy number plots indicate areas of somatic gains (red) and losses (blue) relative to the germline, and areas where no copy number differences are evident (green). The copy number is indicated on the y-axis. LOH plots distinguish areas of normal heterozygosity (orange) from that of LOH (blue), with magenta lines indicating the median LOH value for segments. A beta-allele frequency (BAF; y-axis) of ~0.5 corresponds to normal heterozygosity.

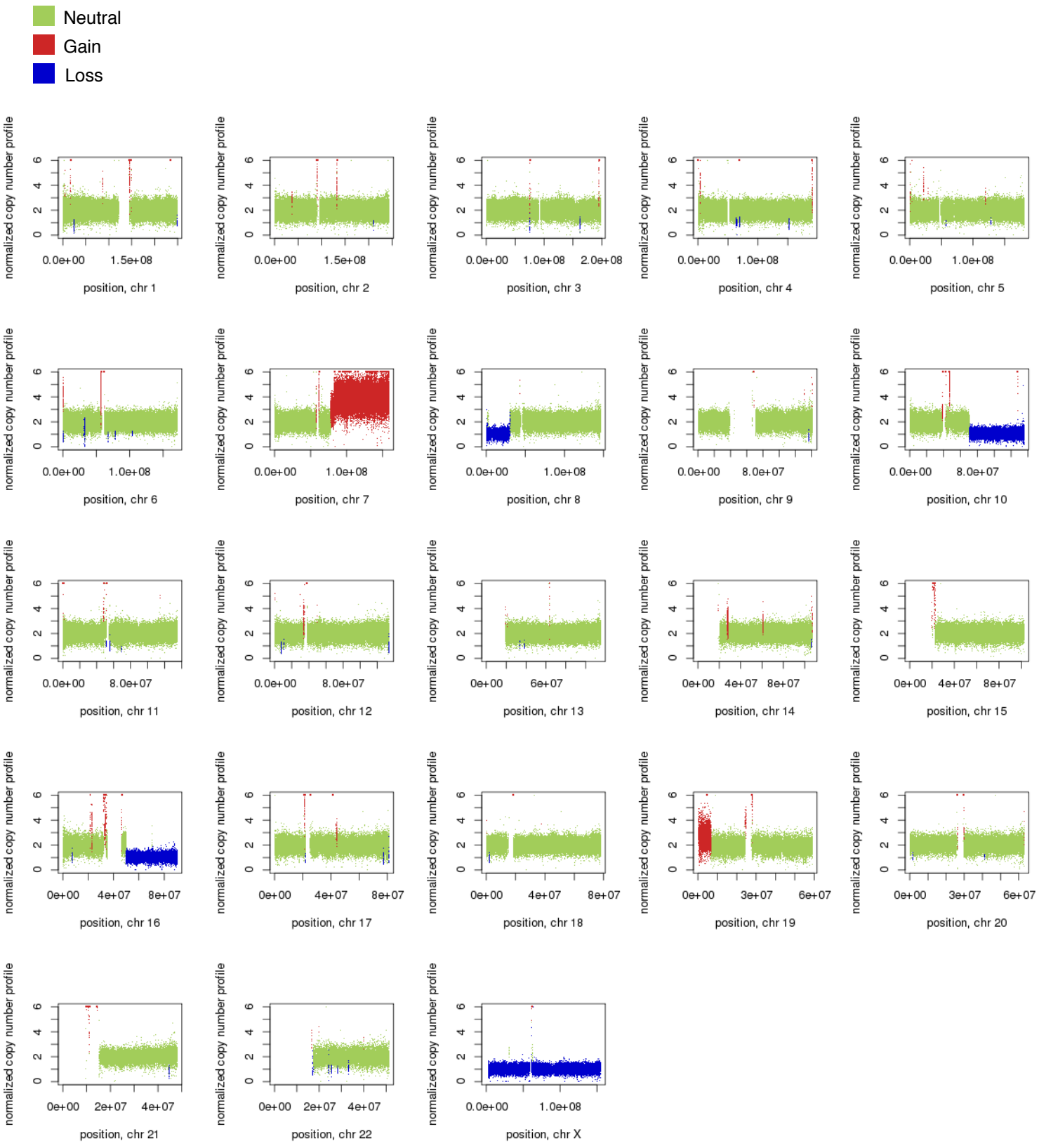
MB-Rec-02 Therapy naive tumor CNV



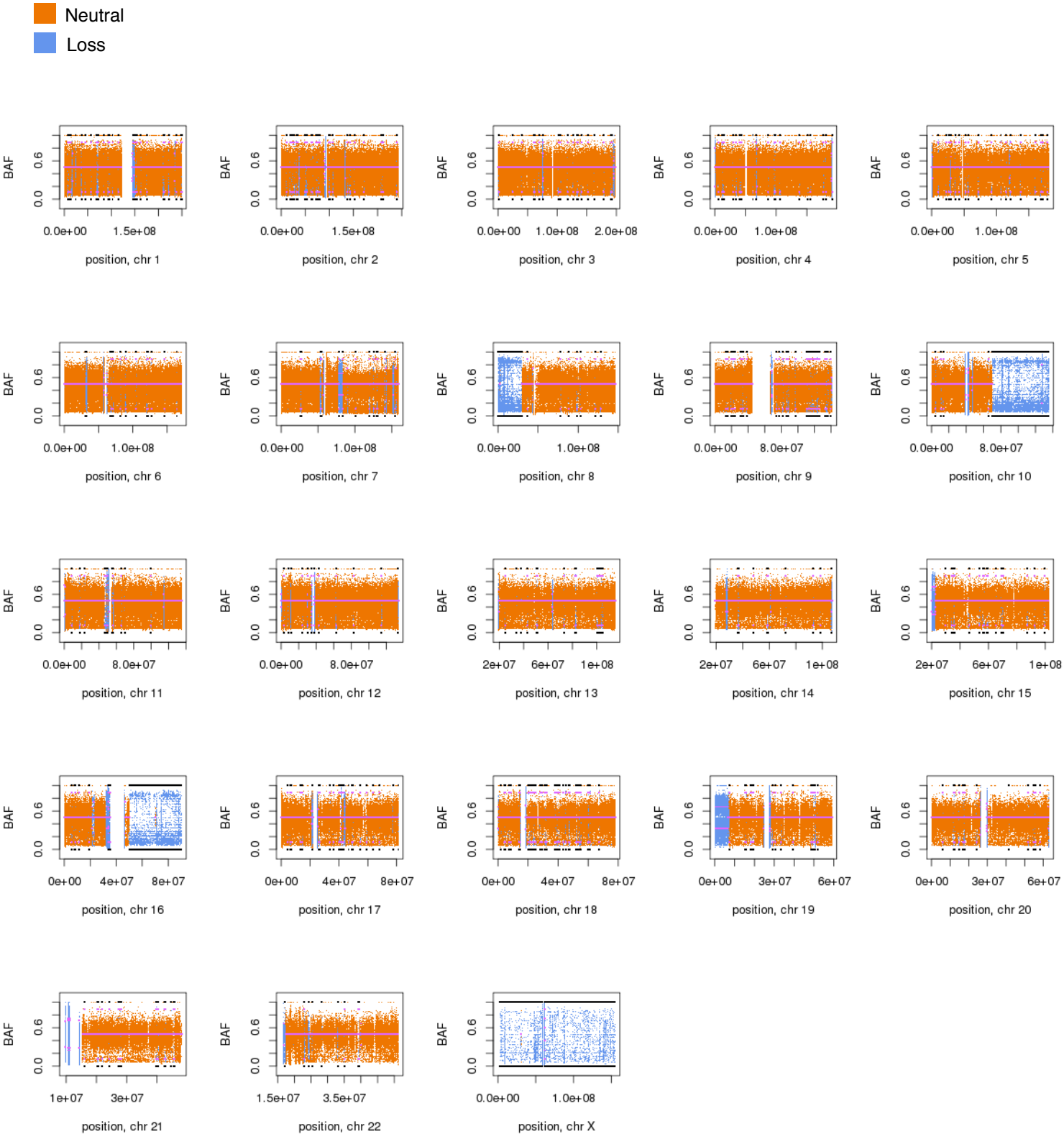
MB-Rec-02 Therapy naive tumor LOH



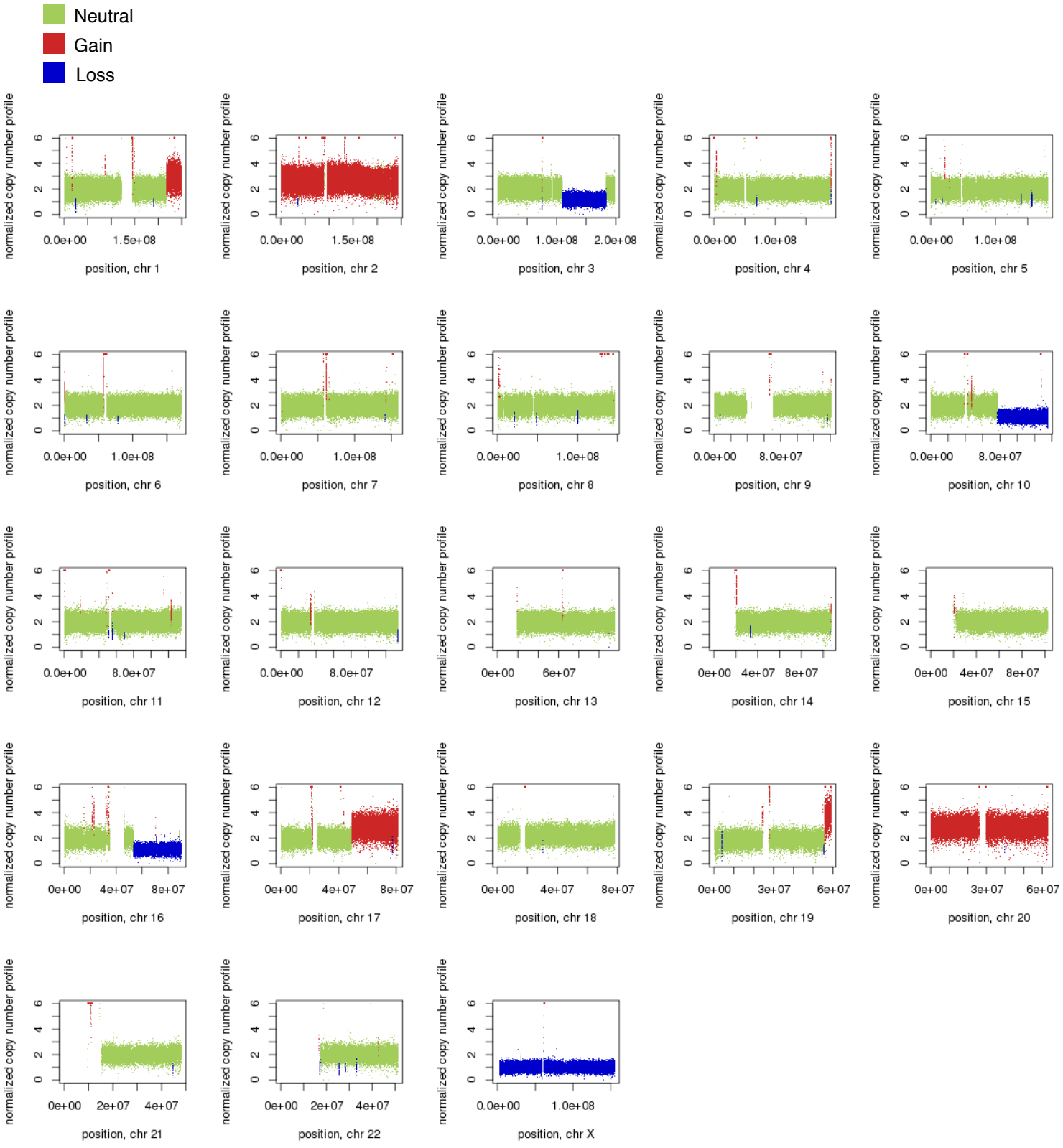
MB-Rec-02 Recurrence CNV



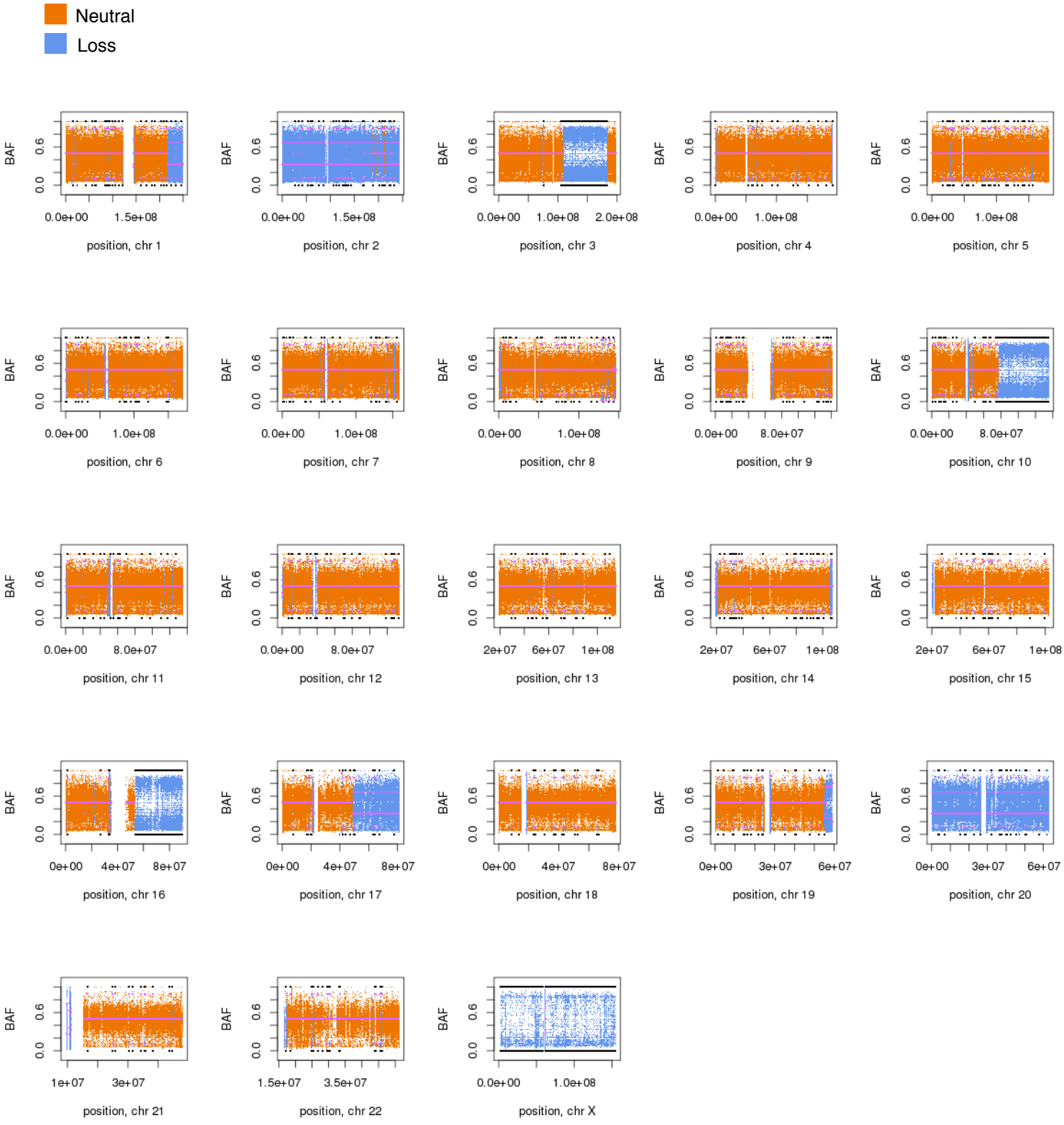
MB-Rec-02 Recurrence LOH



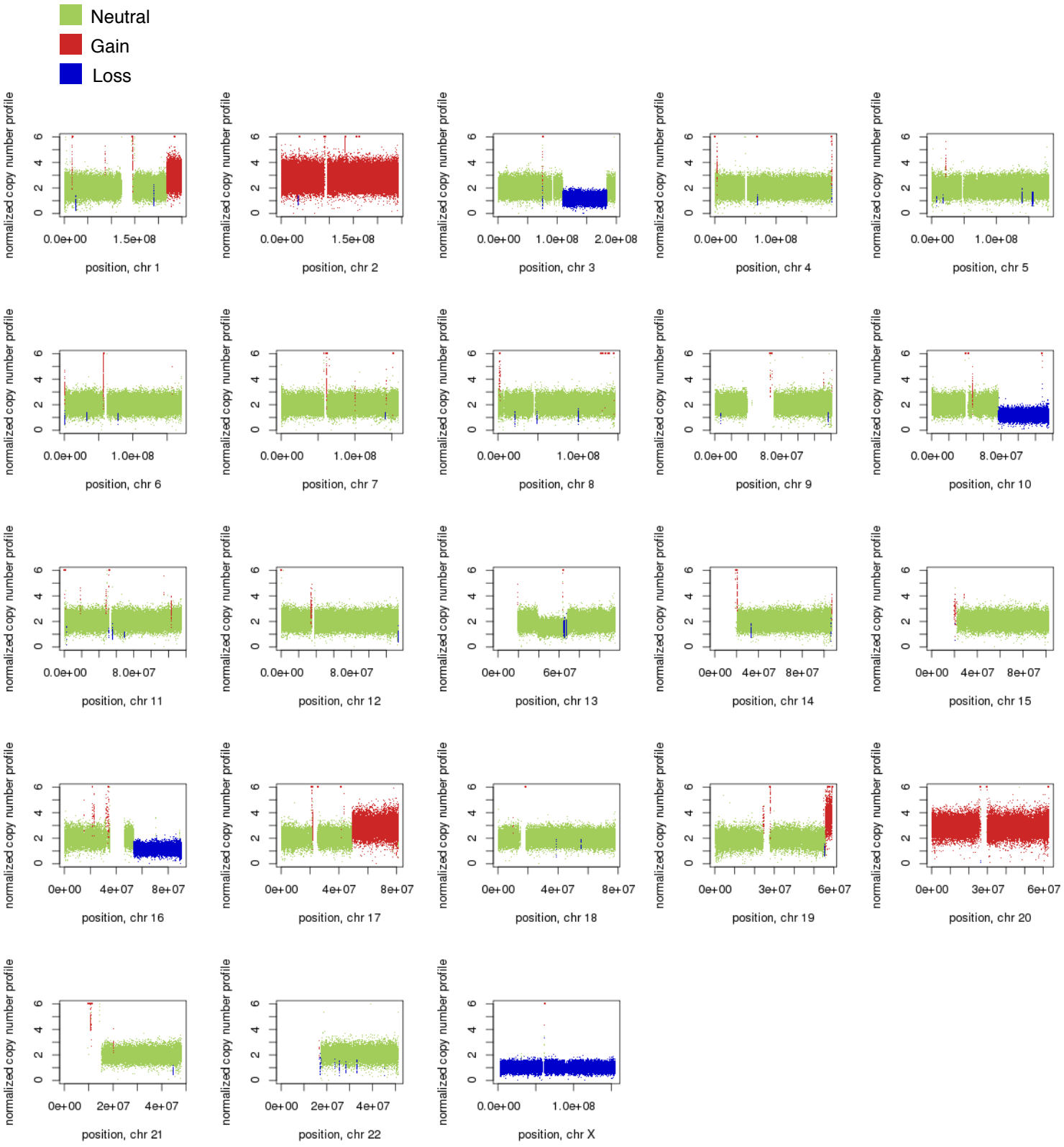
MB-Rec-03 Therapy naive tumor CNV



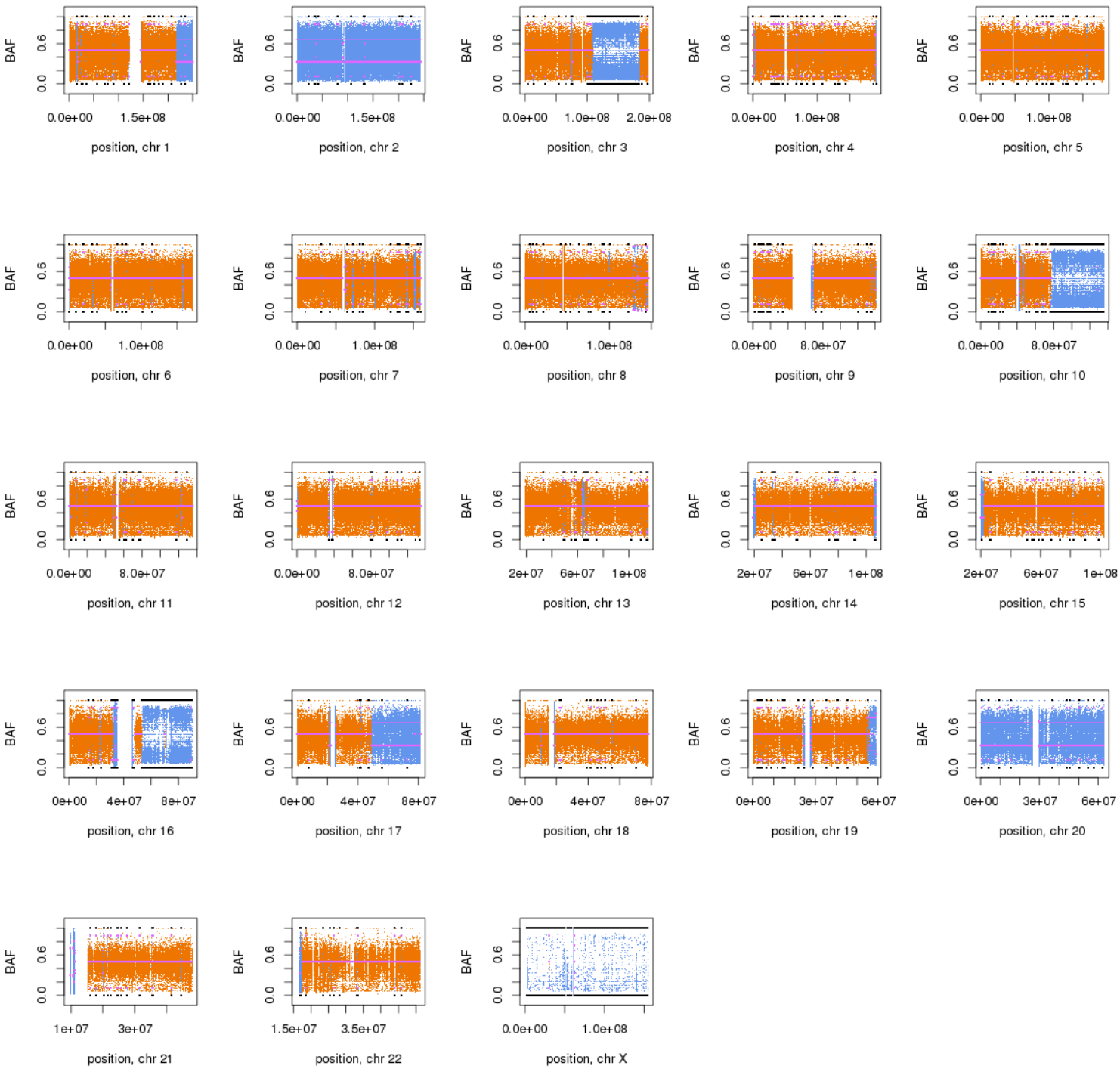
MB-Rec-03 Therapy naive tumor LOH



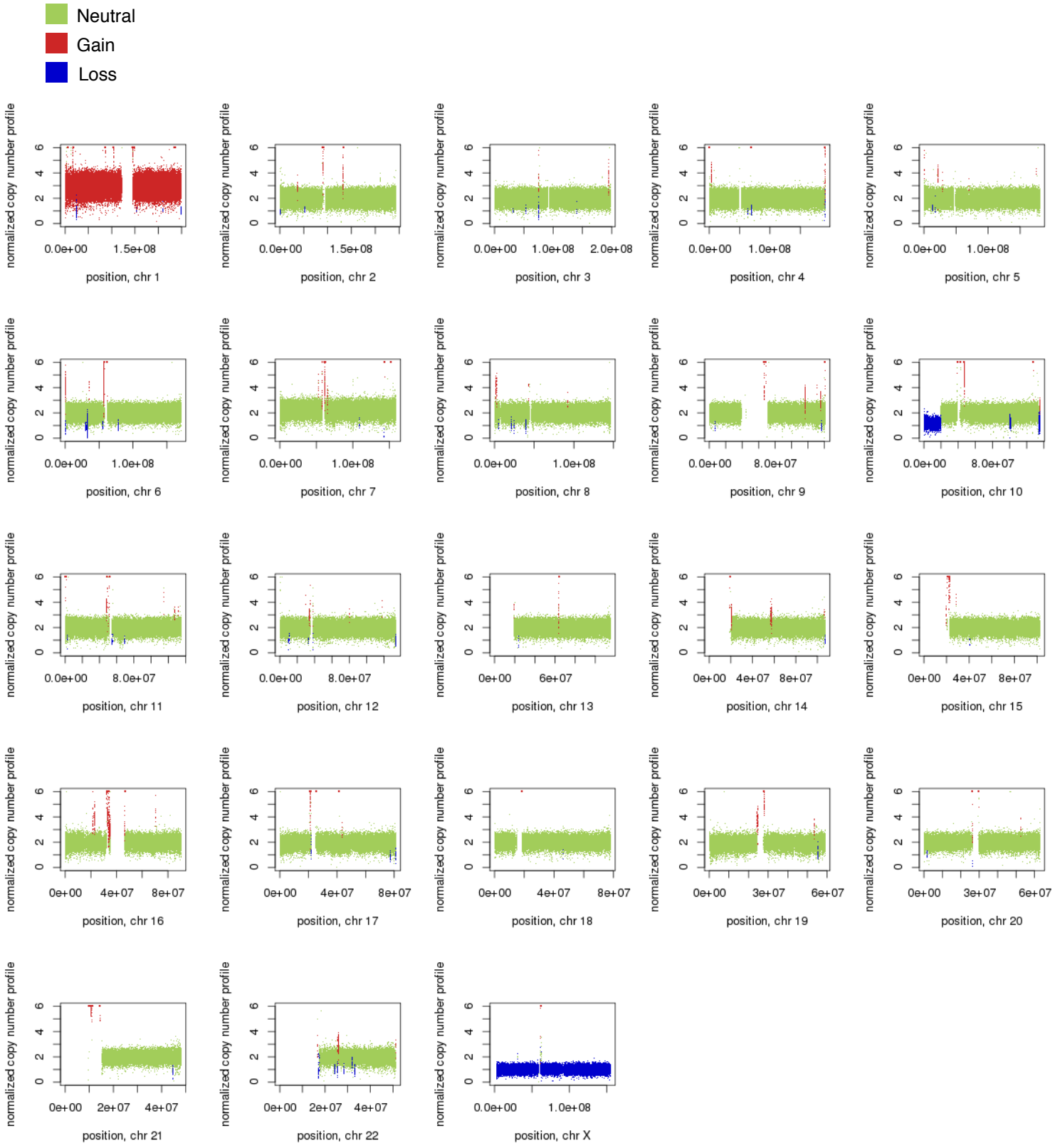
MB-Rec-03 Recurrence CNV



MB-Rec-03 Recurrence LOH

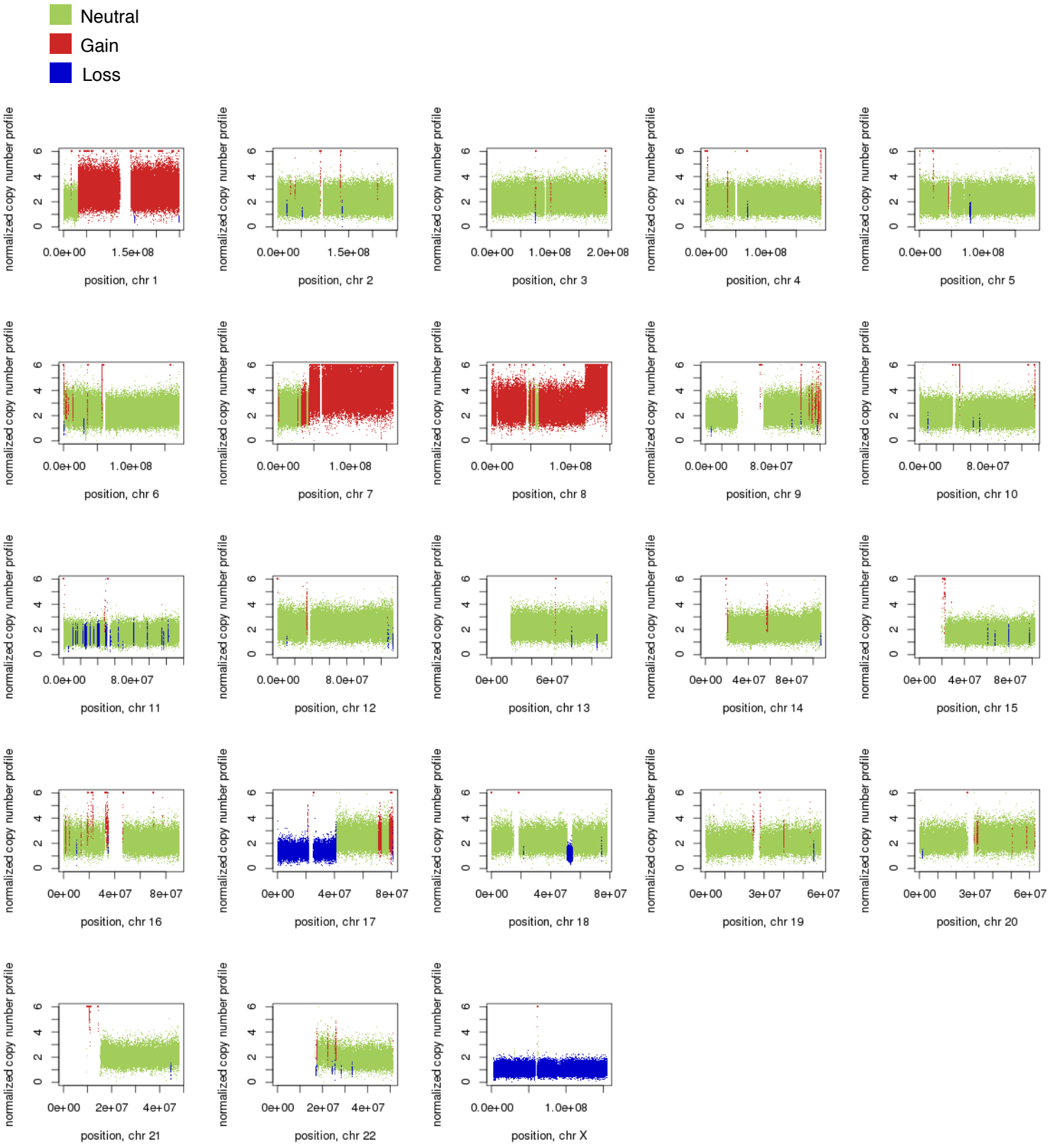


MB-Rec-04 Therapy naive tumor CNV

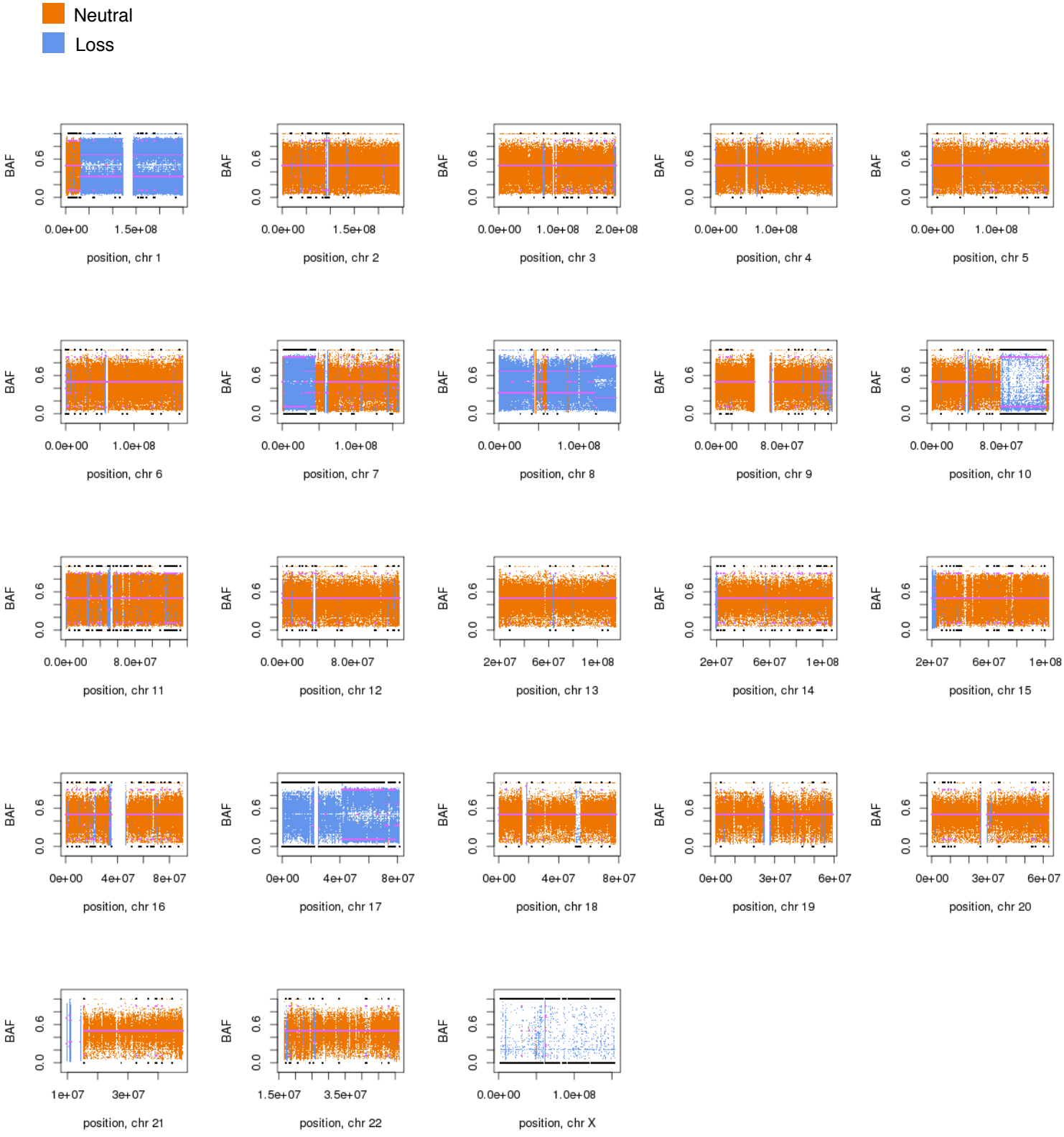


MB-Rec-04 Therapy naive tumor LOH

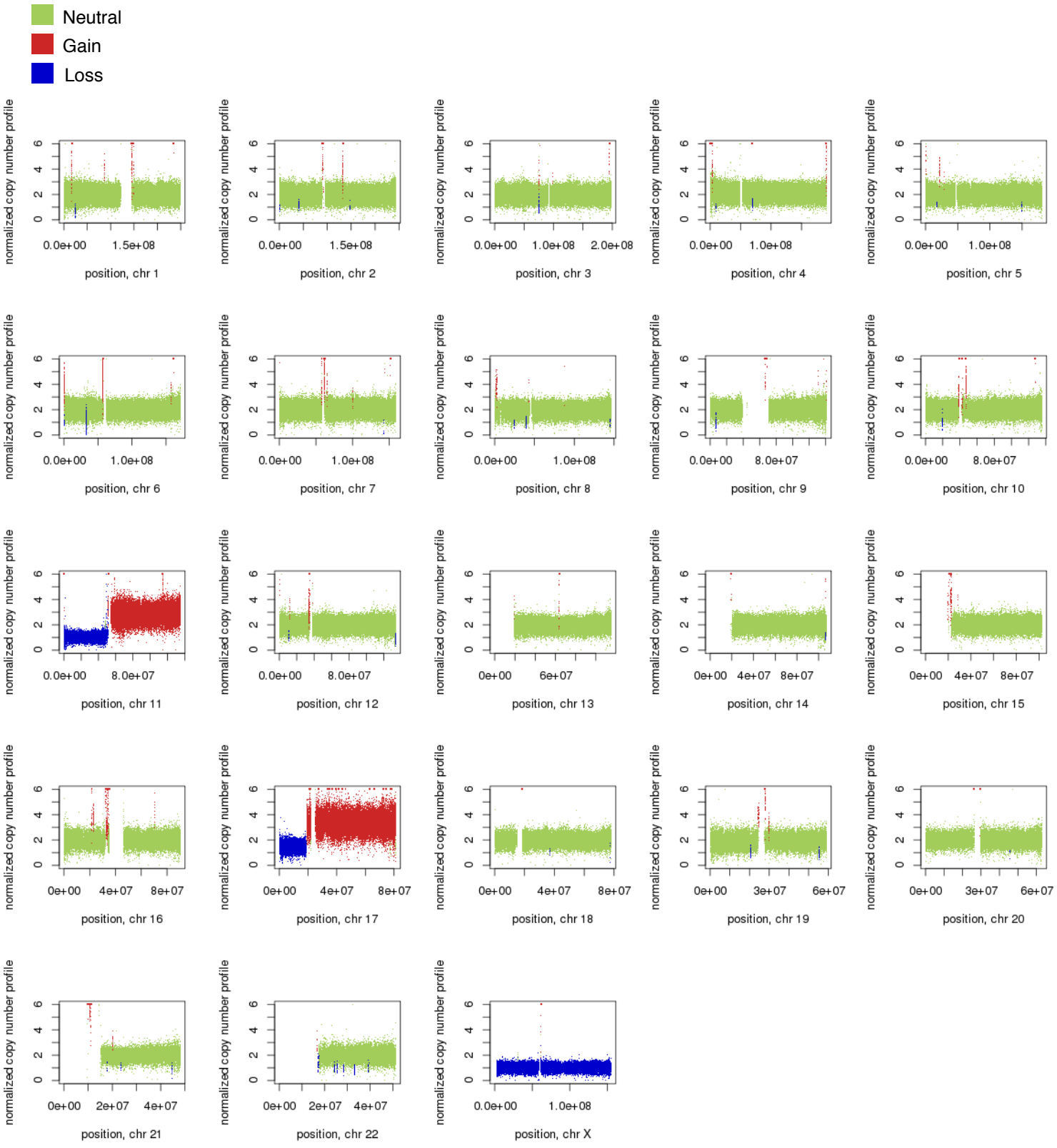
MB-Rec-04 Recurrence CNV



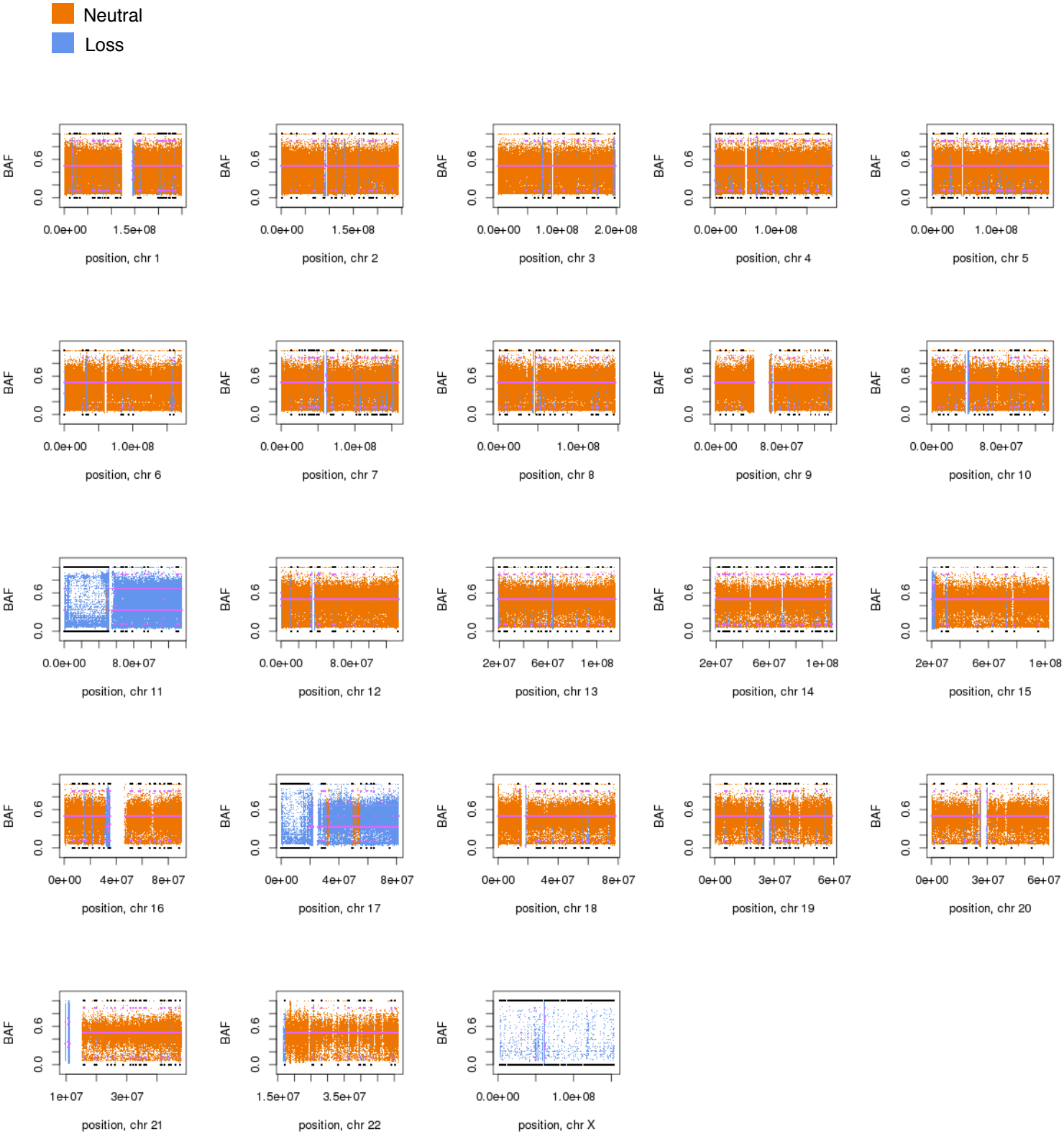
MB-Rec-04 Recurrence LOH



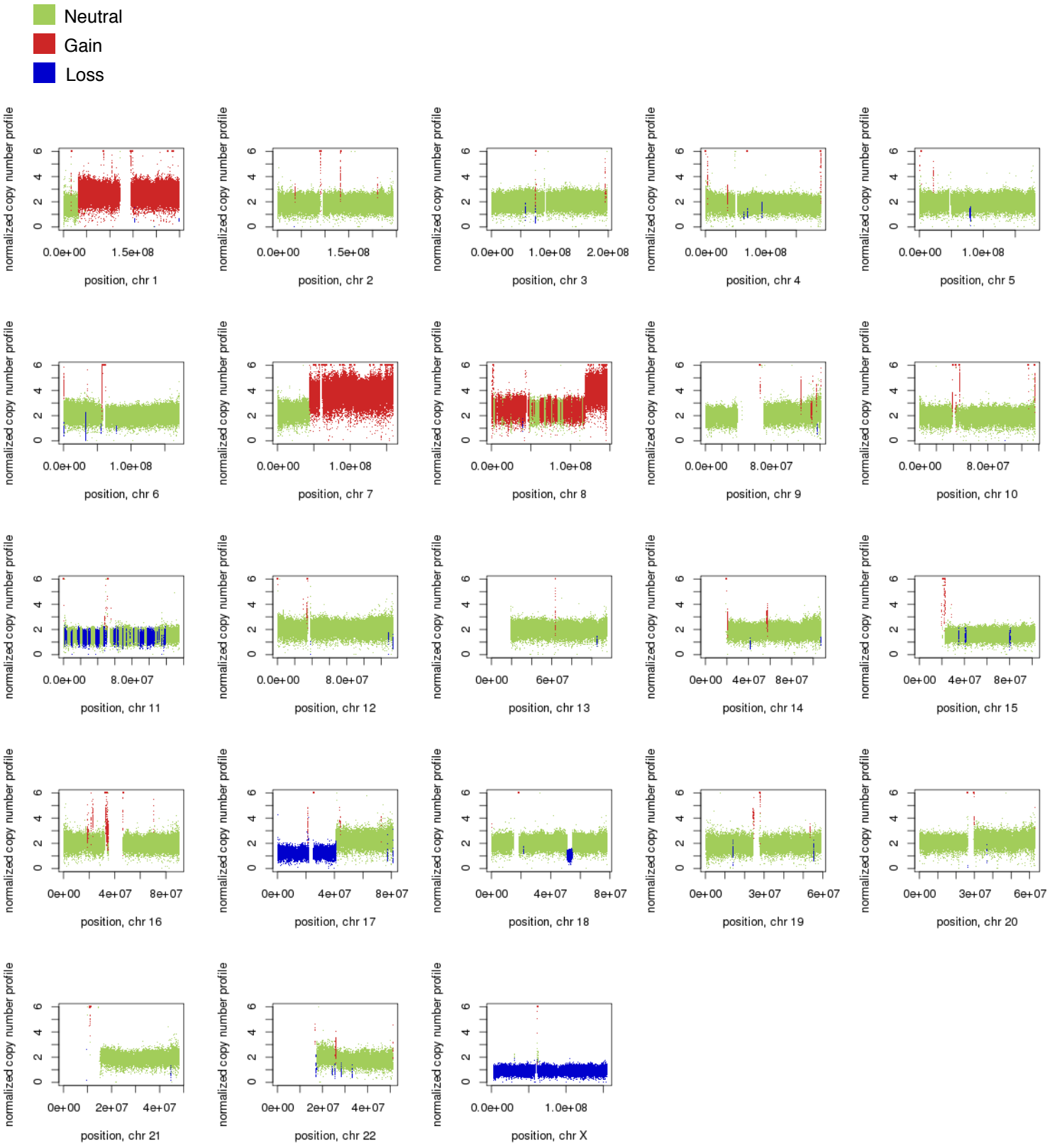
MB-Rec-05 Therapy naive tumor CNV



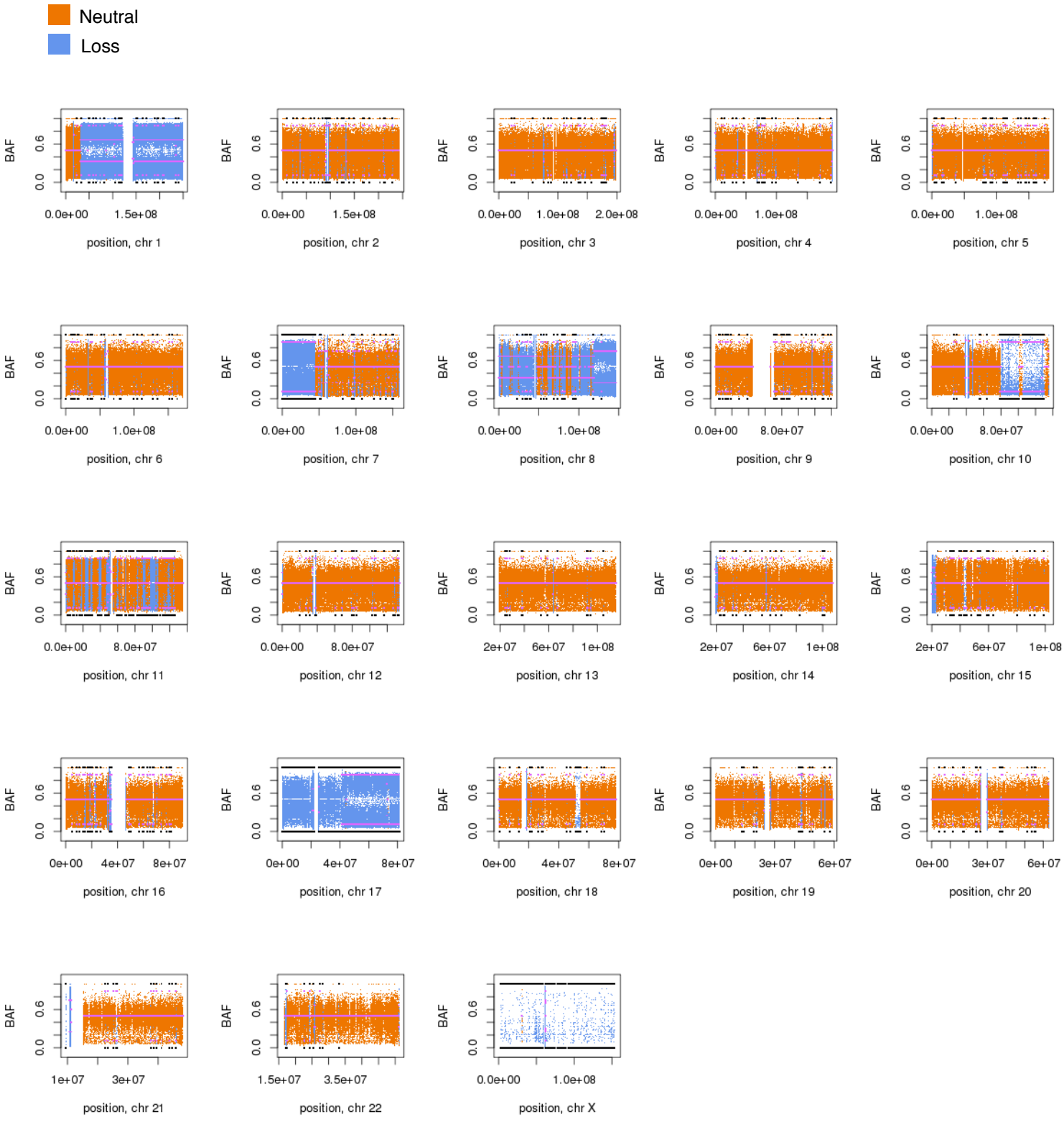
MB-Rec-05 Therapy naive tumor LOH



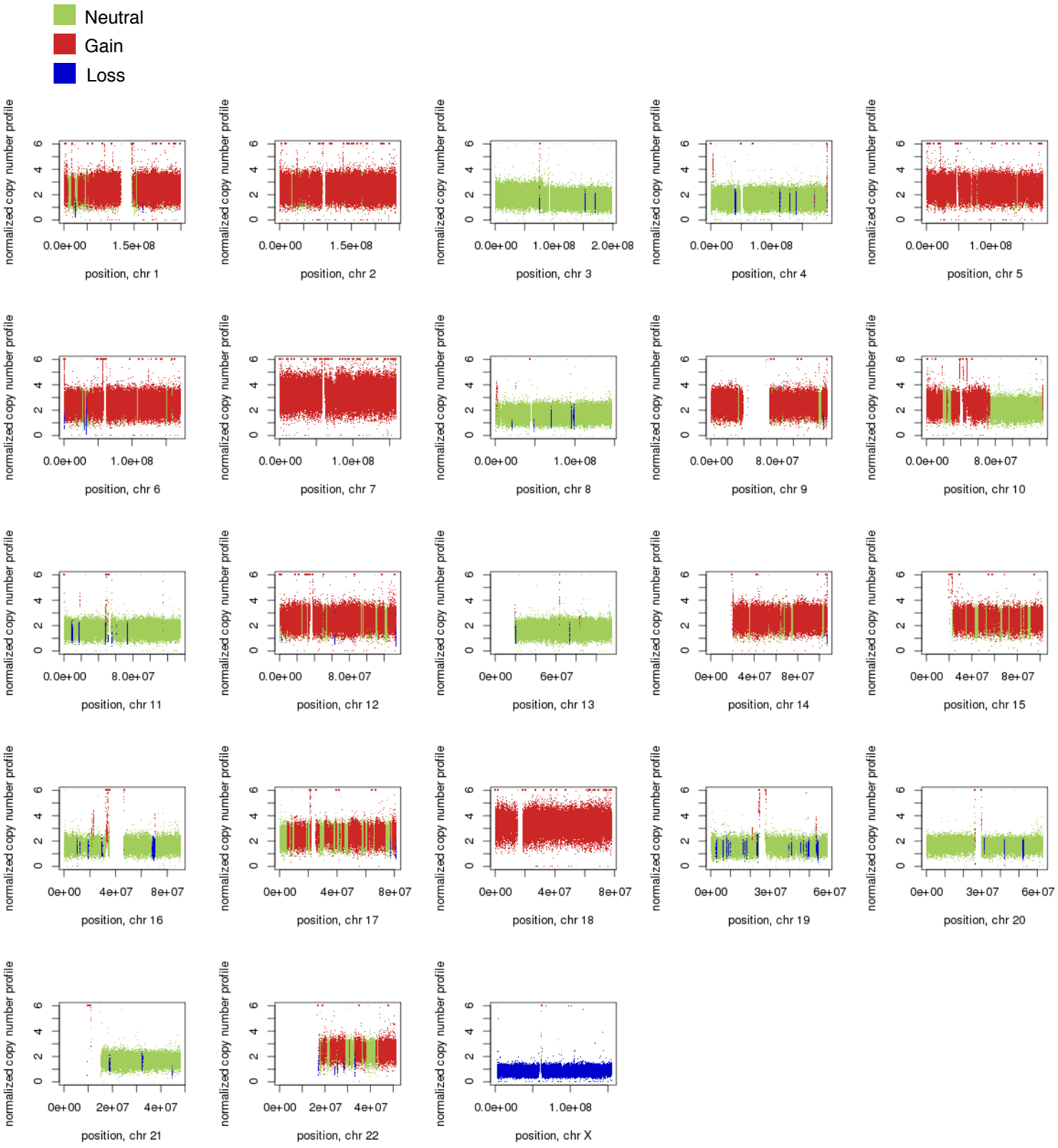
MB-Rec-05 Recurrence CNV



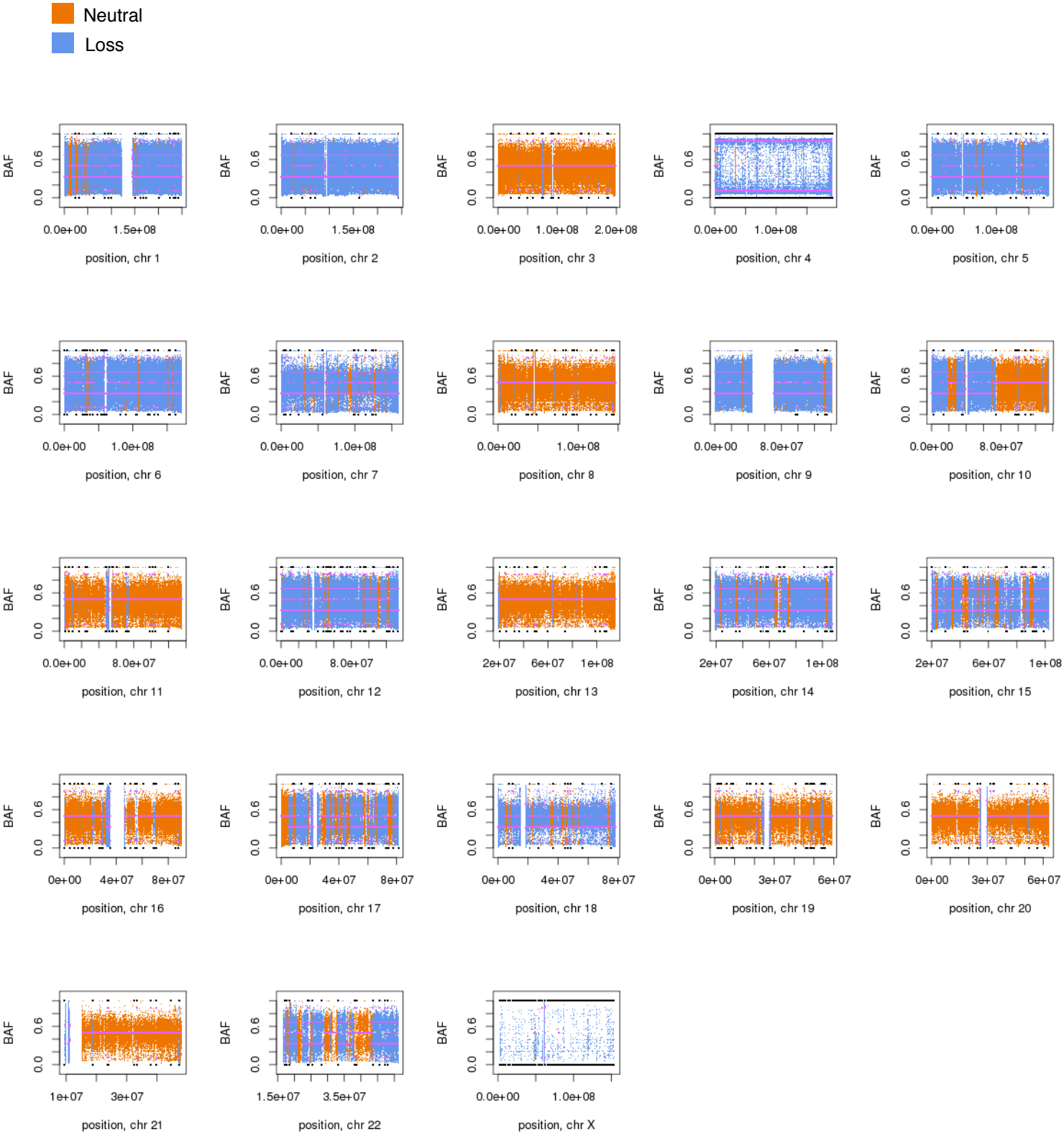
MB-Rec-05 Recurrence LOH



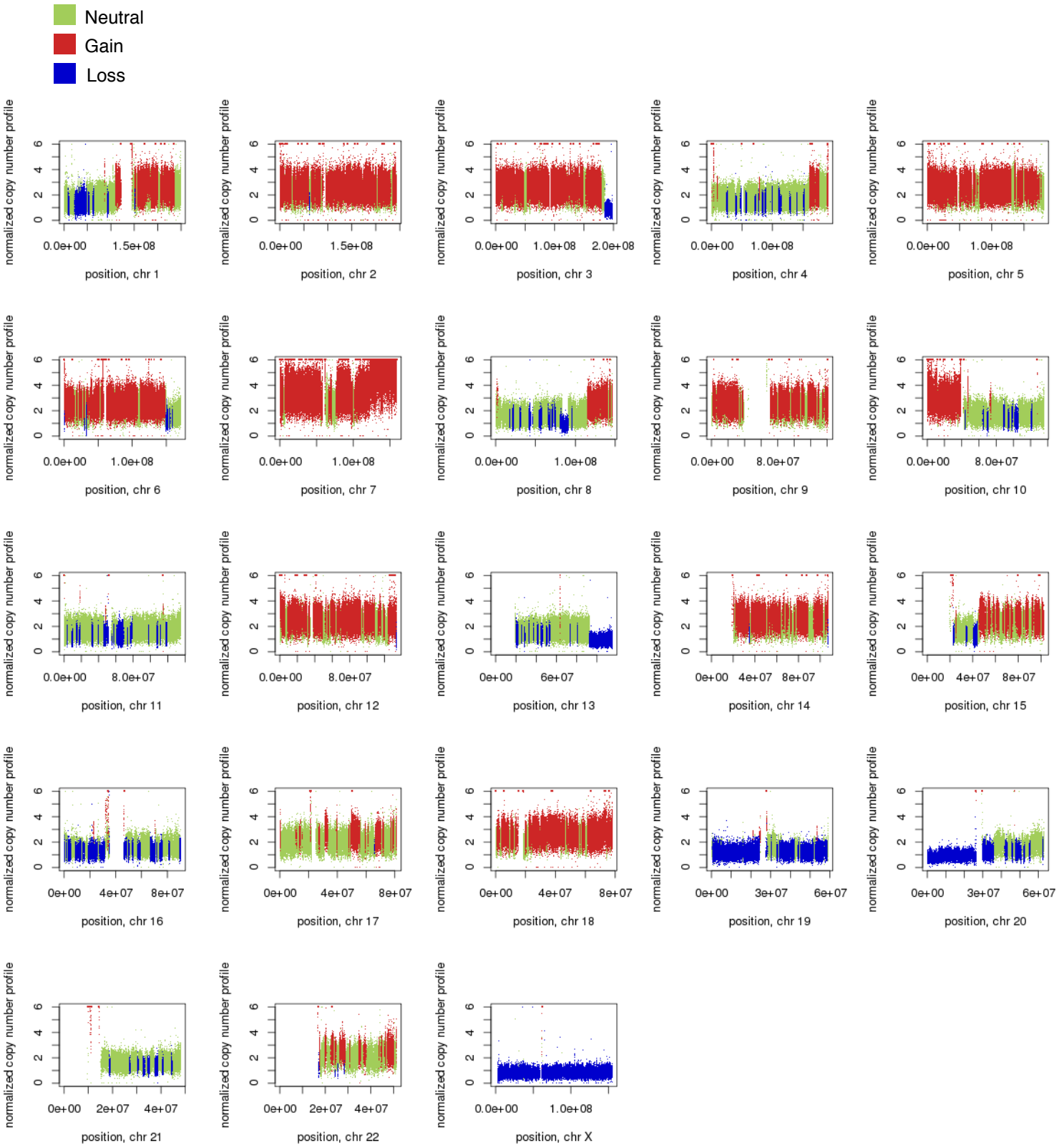
MB-Rec-07 Therapy naive tumor CNV



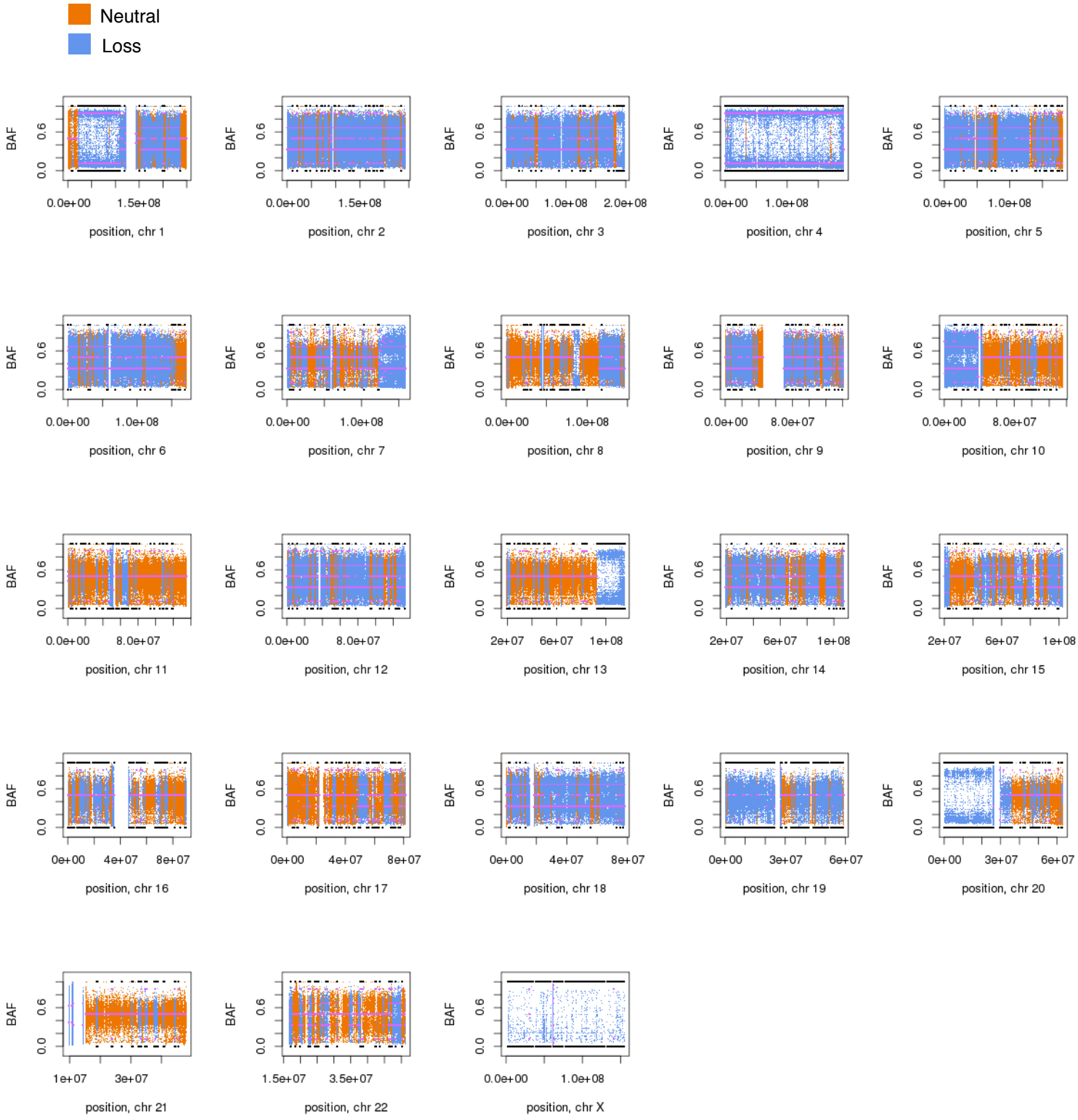
MB-Rec-07 Therapy naive tumor LOH



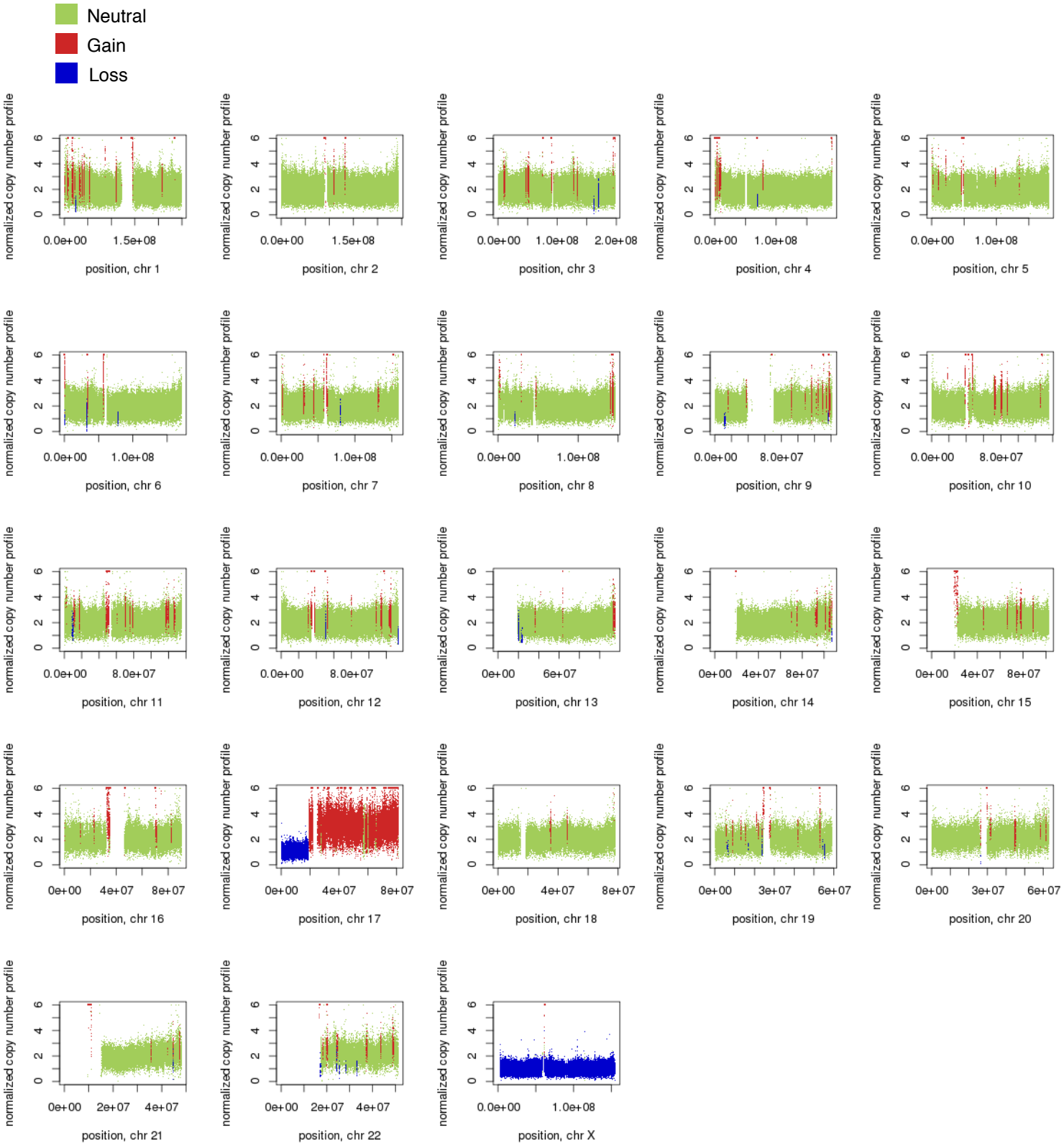
MB-Rec-07 Recurrence CNV



MB-Rec-07 Recurrence LOH

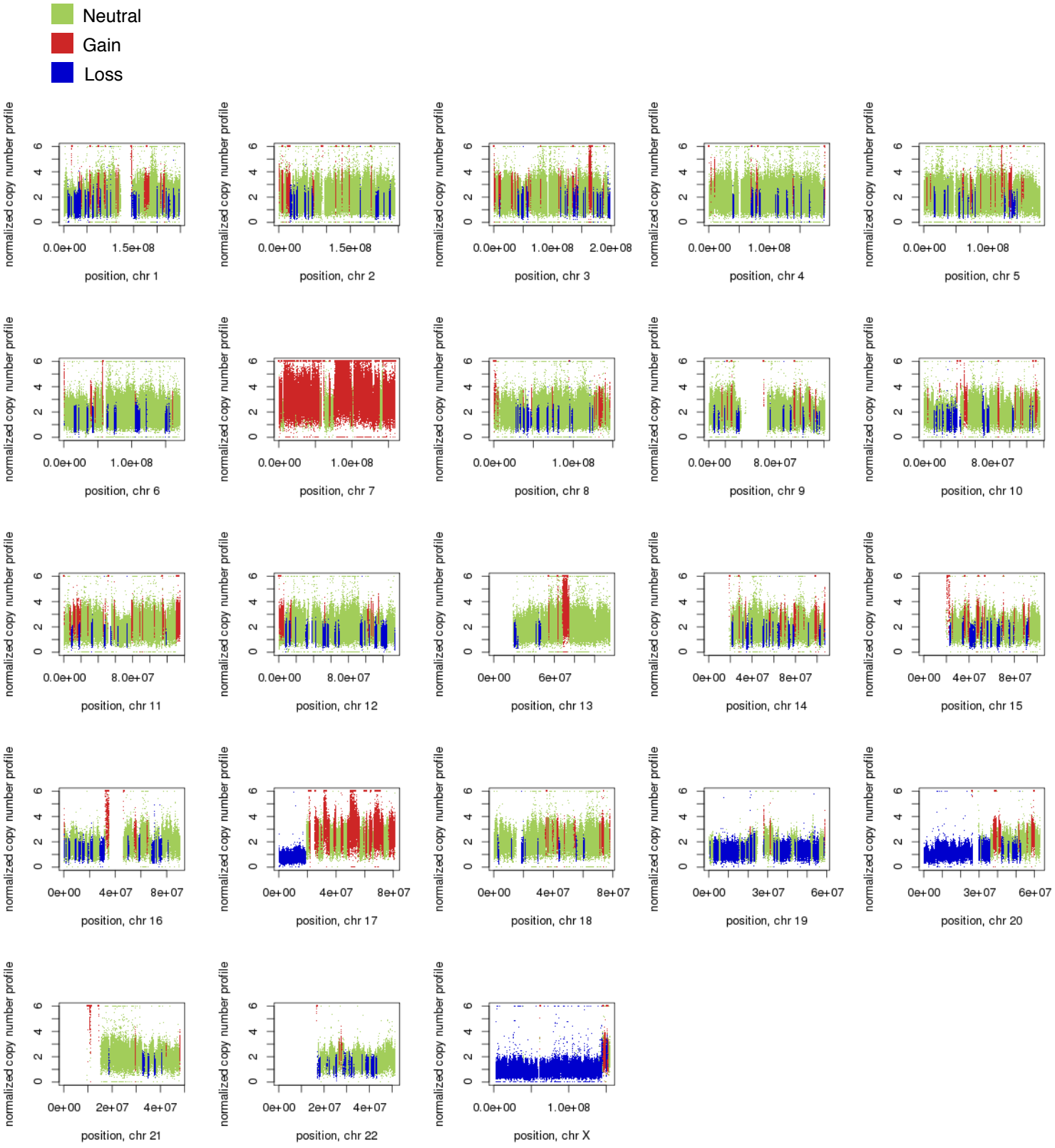


MB-Rec-08 Therapy naive tumor CNV

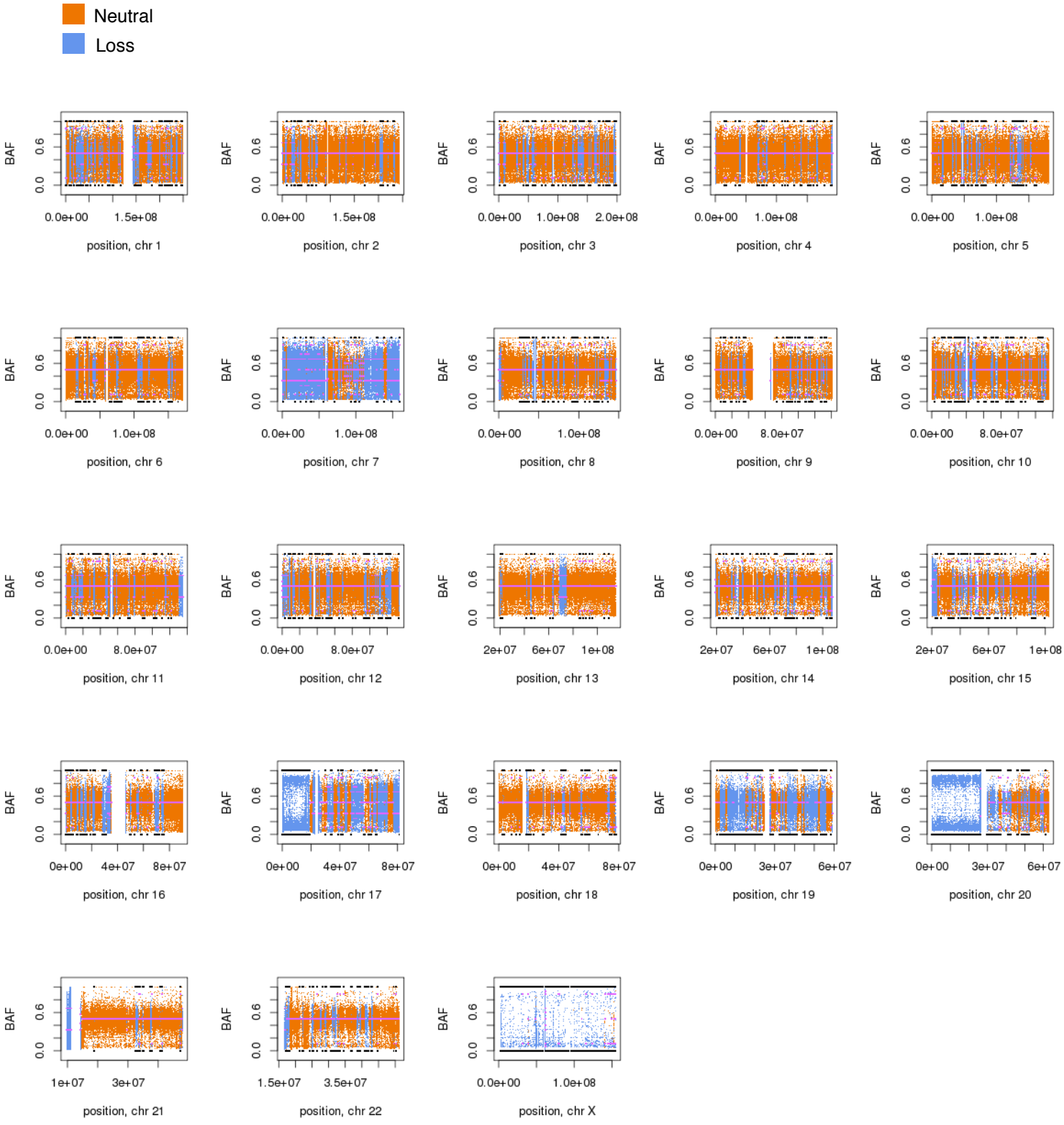


MB-Rec-08 Therapy naive tumor LOH

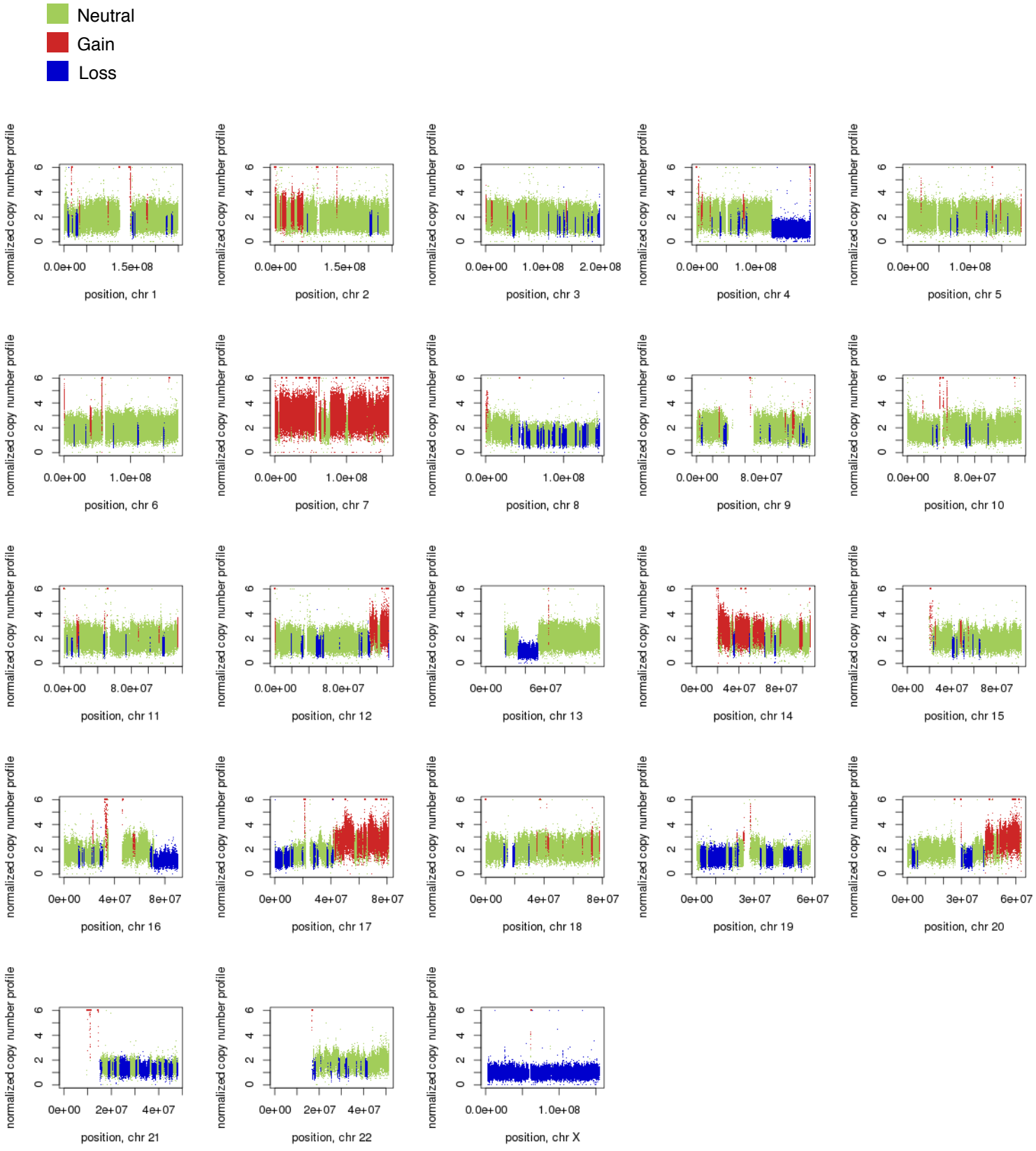
MB-Rec-08 Recurrence CNV



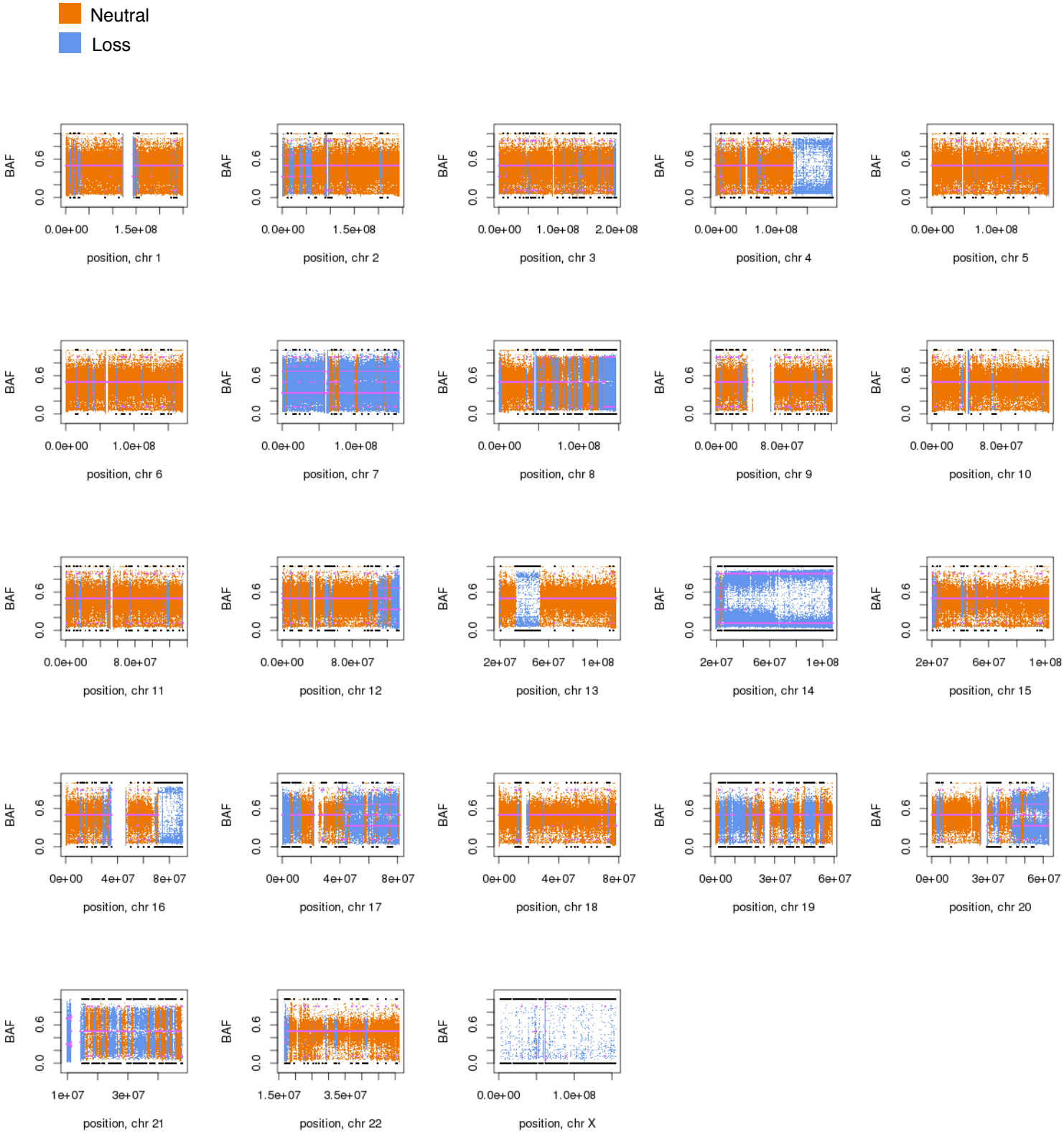
MB-Rec-08 Recurrence LOH



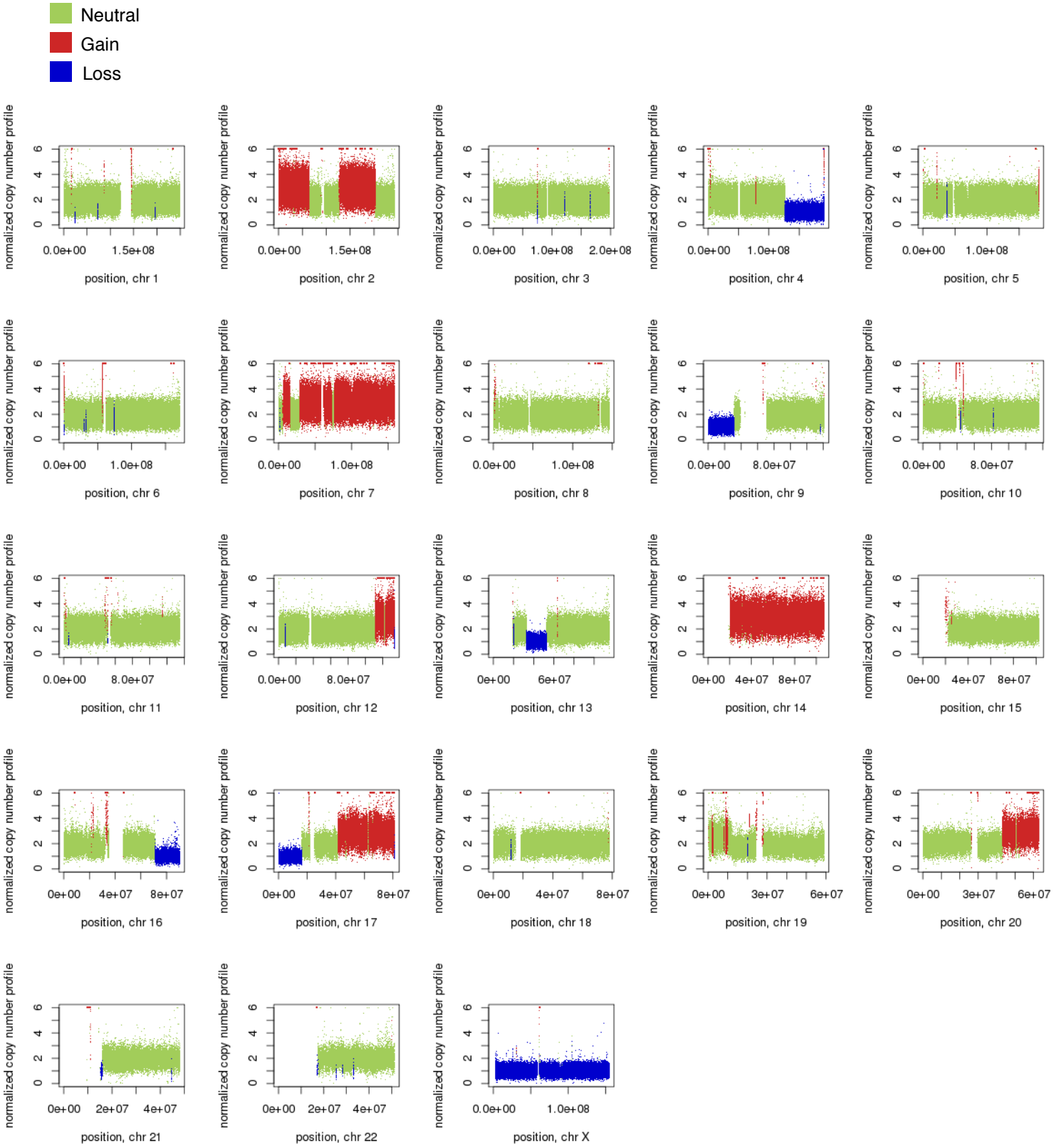
MB-Rec-09 Therapy naive tumor CNV



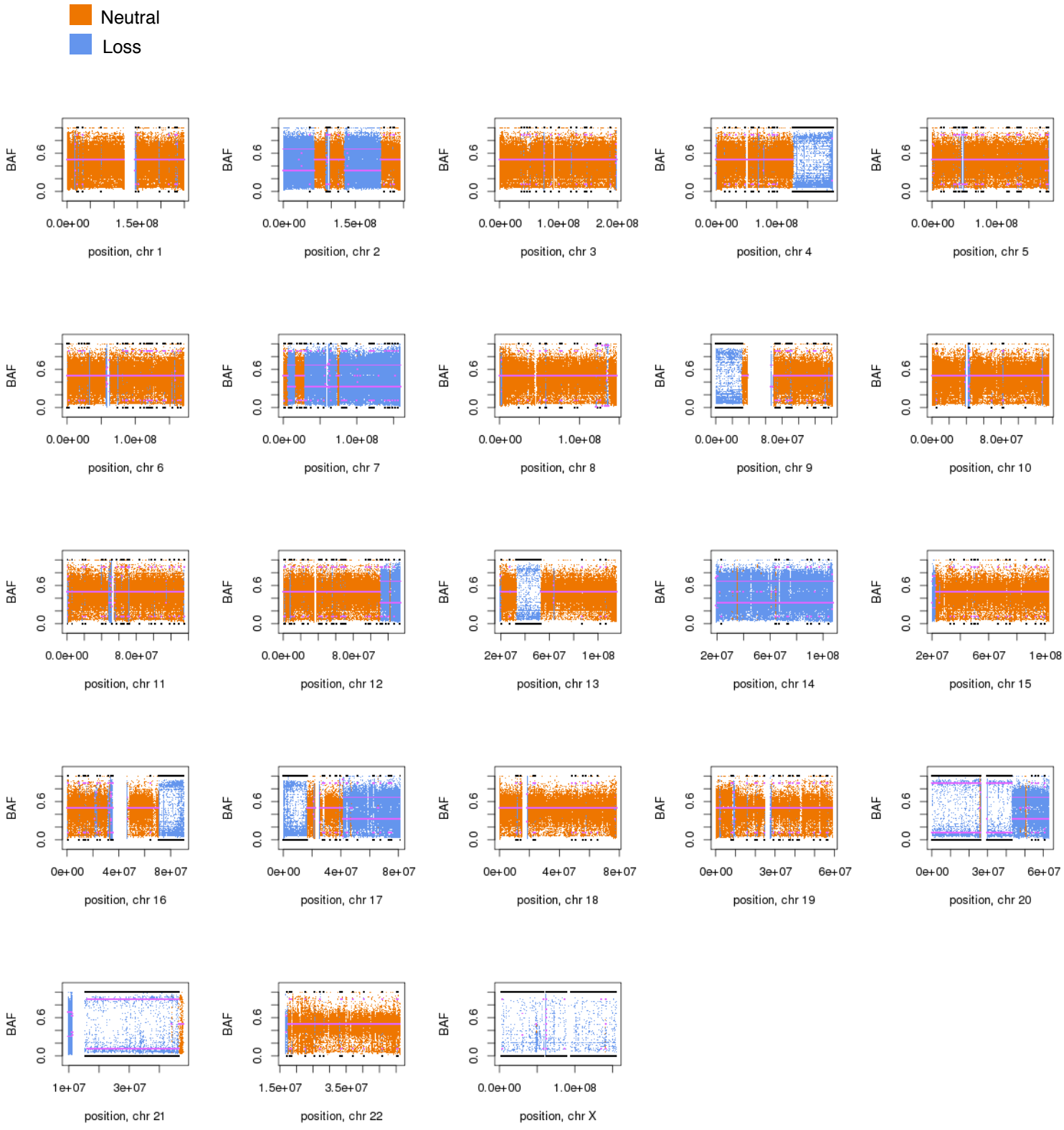
MB-Rec-09 Therapy naive tumor LOH

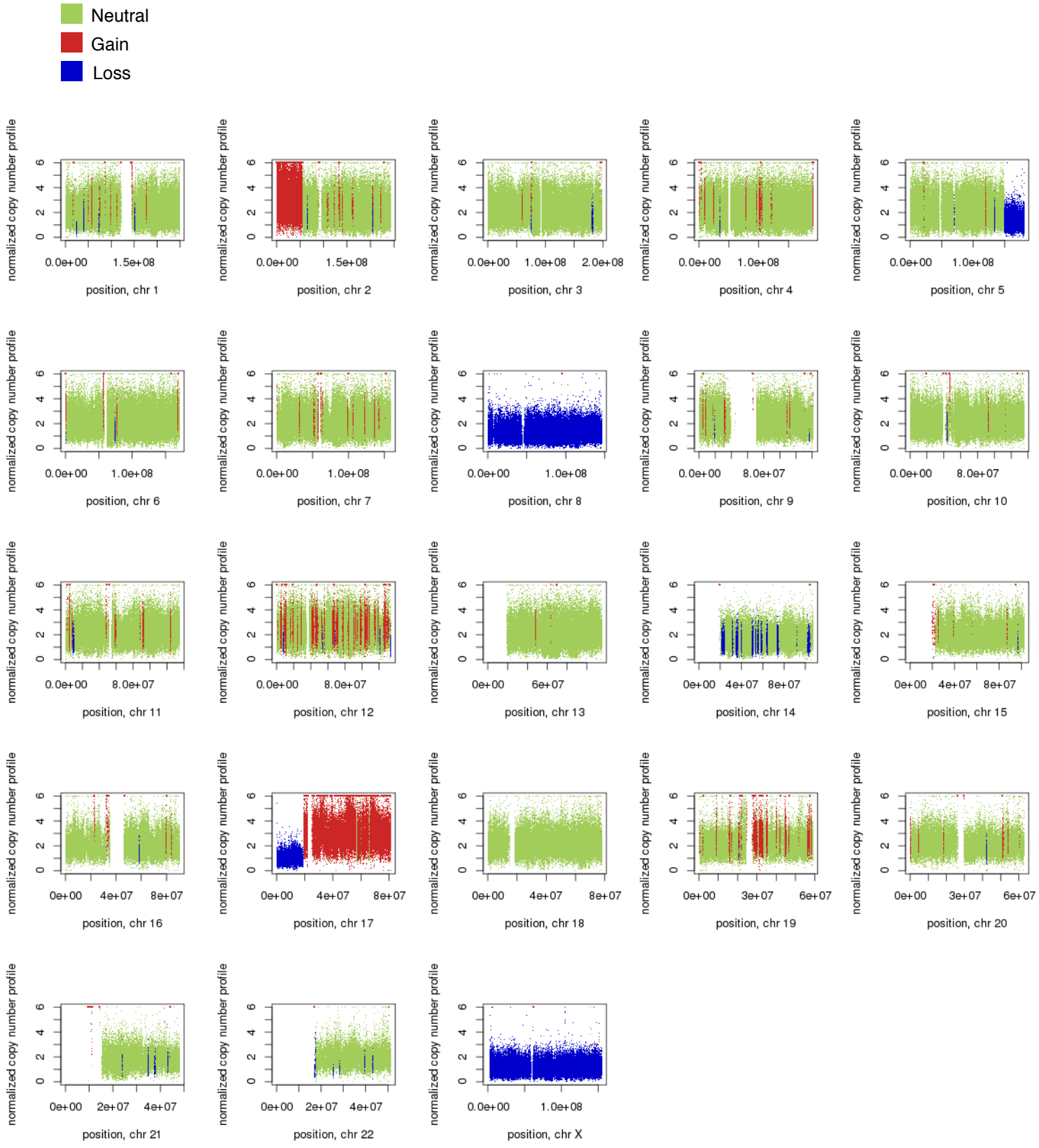


MB-Rec-09 Recurrence CNV

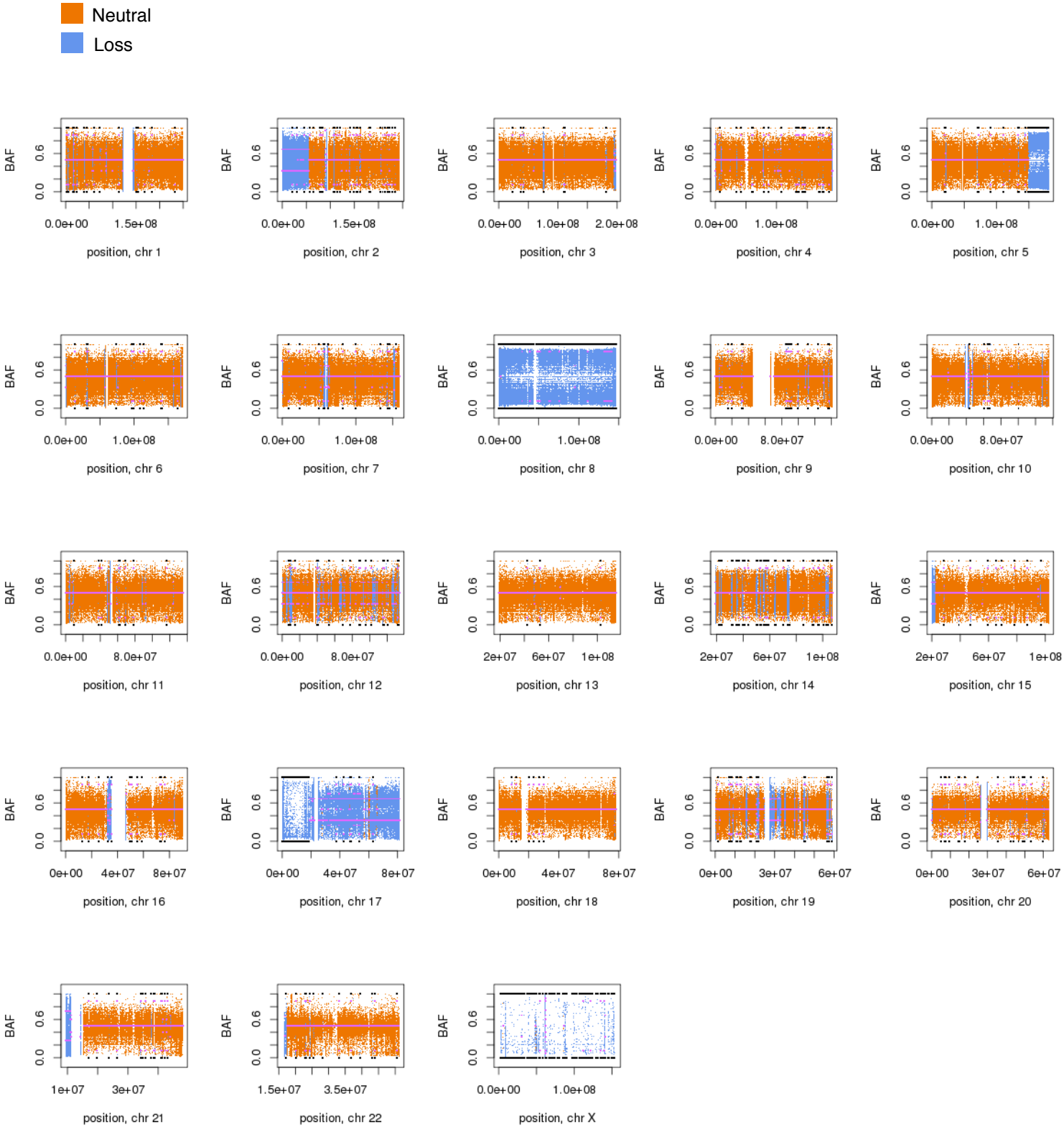


MB-Rec-09 Recurrence LOH

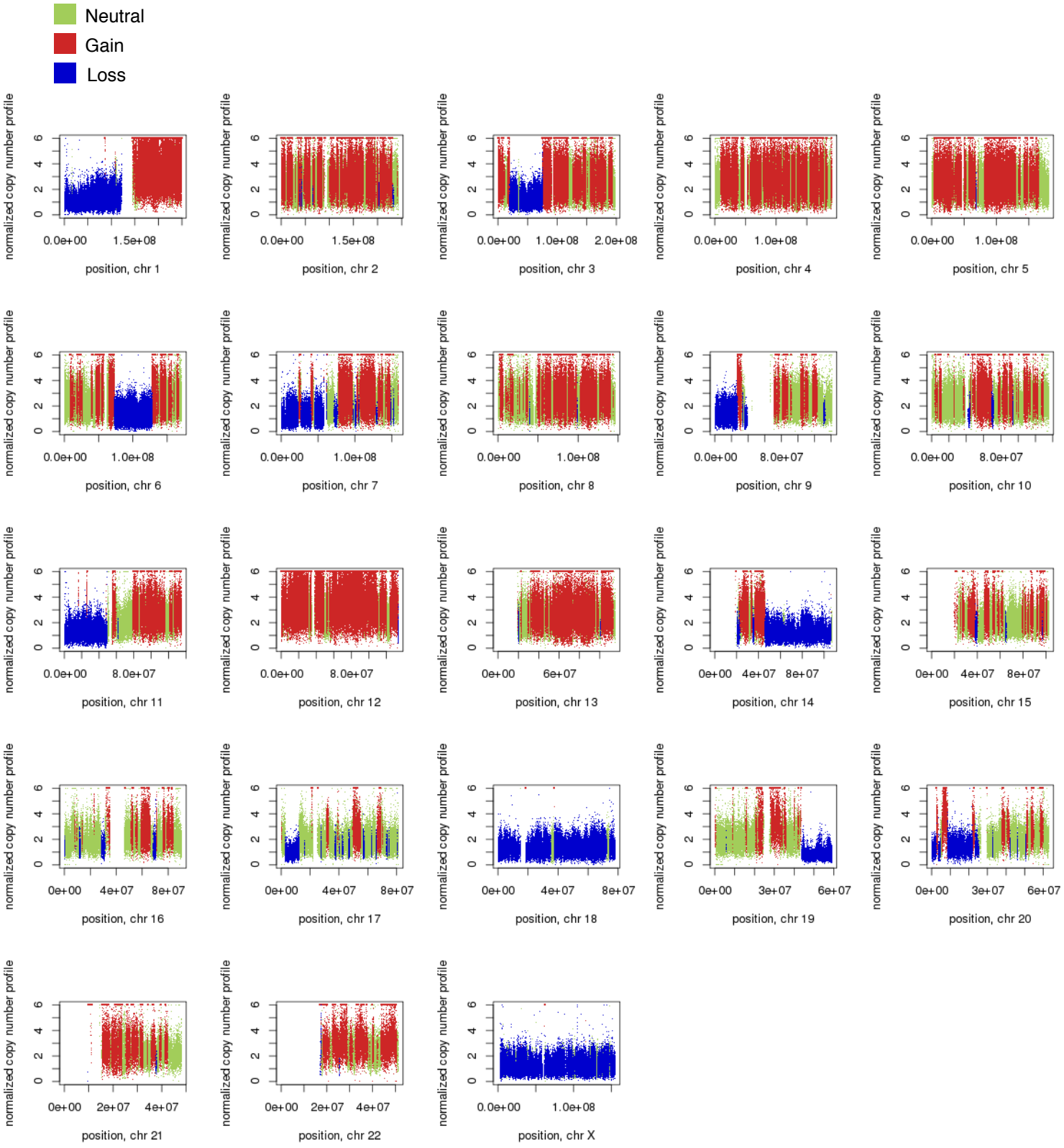



MB-Rec-10 Therapy naive tumor CNV
Neutral
Gain
Loss
normalized copy number profile
position, chr 1
position, chr 2
position, chr 3
position, chr 4
position, chr 5
position, chr 6
position, chr 7
position, chr 8
position, chr 9
position, chr 10
position, chr 11
position, chr 12
position, chr 13
position, chr 14
position, chr 15
position, chr 16
position, chr 17
position, chr 18
position, chr 19
position, chr 20
position, chr 21
position, chr 22
position, chr X

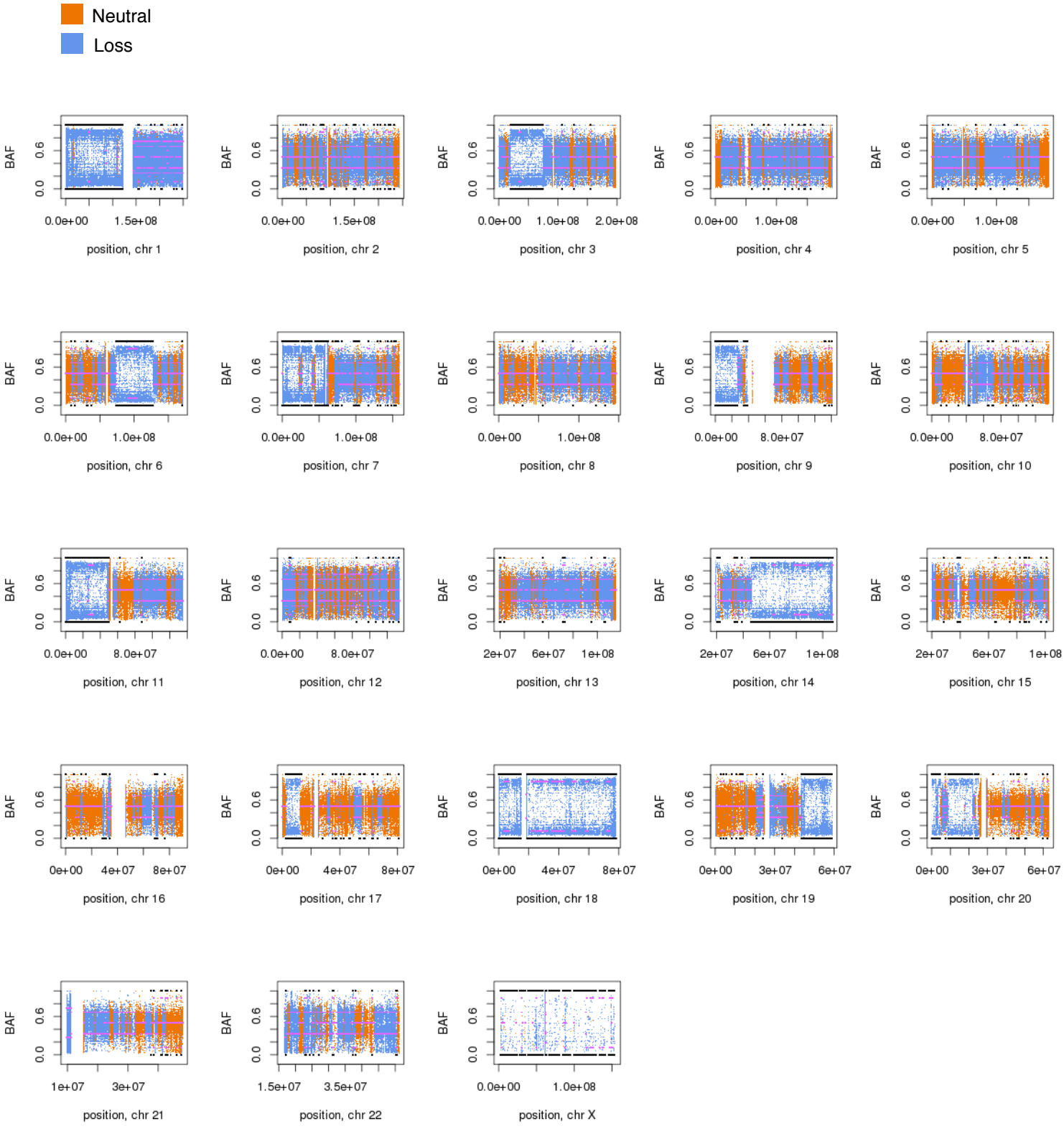
MB-Rec-10 Therapy naive tumor LOH



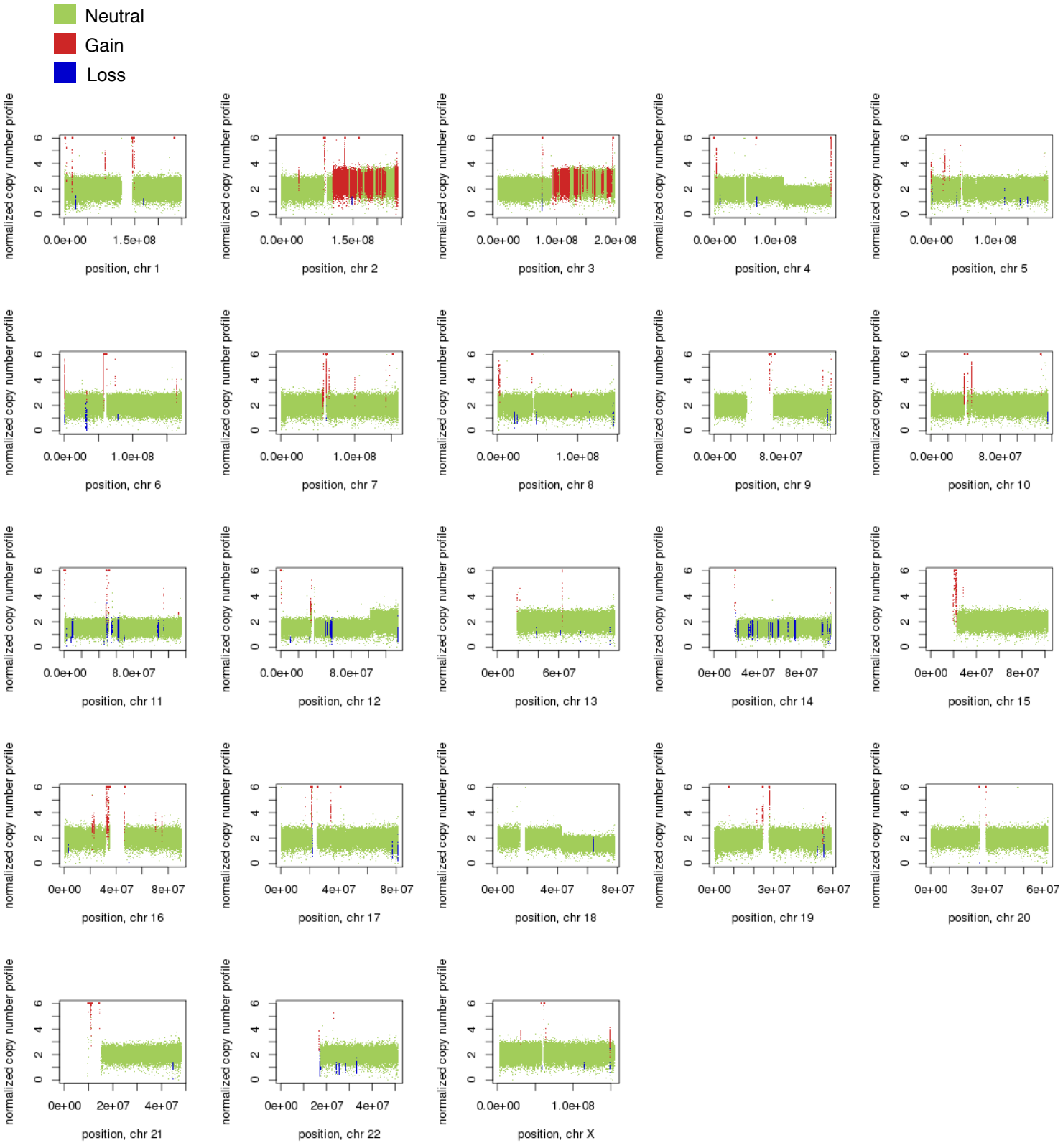
MB-Rec-10 Recurrence CNV



MB-Rec-10 Recurrence LOH

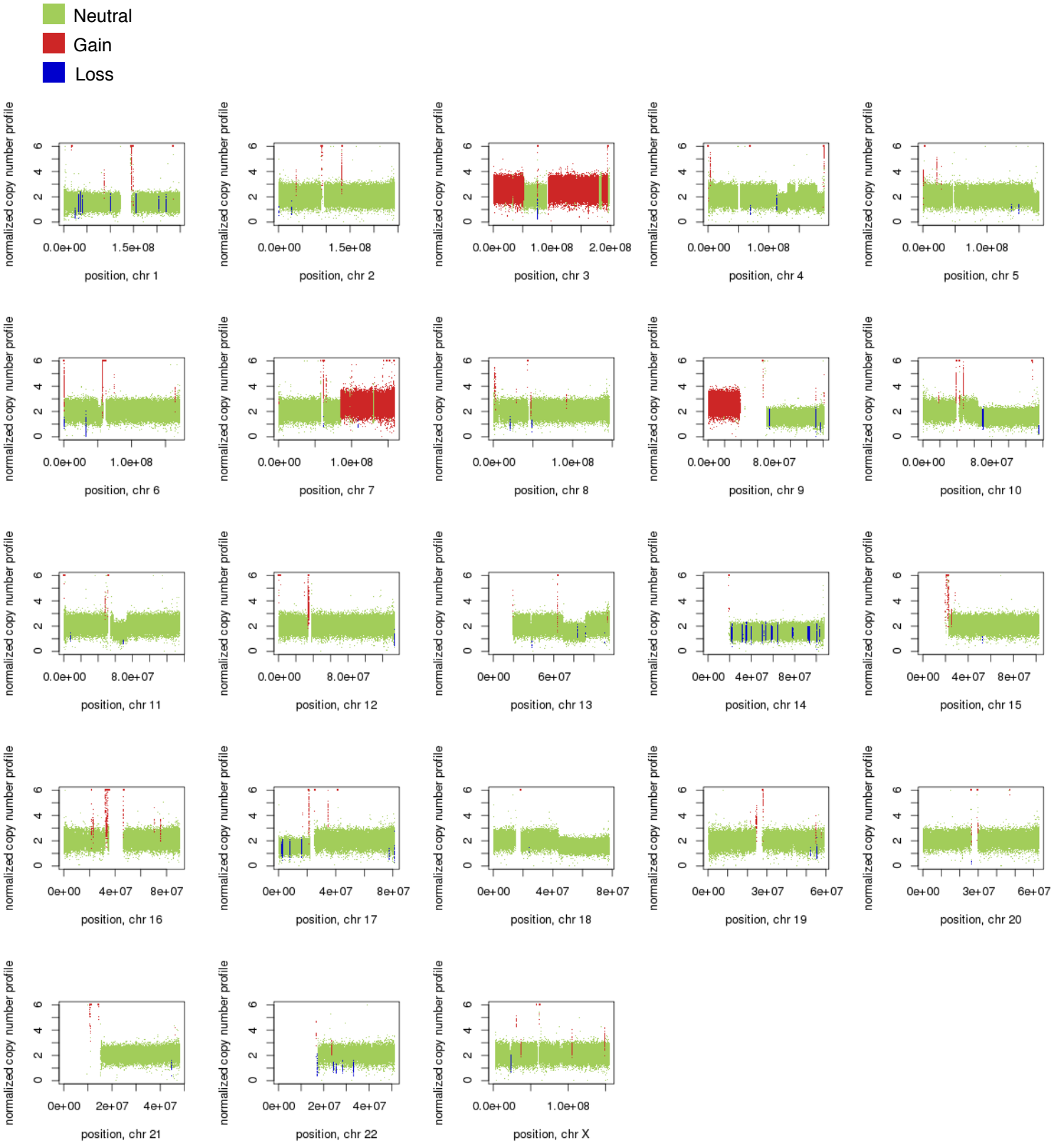


MB-Rec-11 Therapy naive tumor CNV



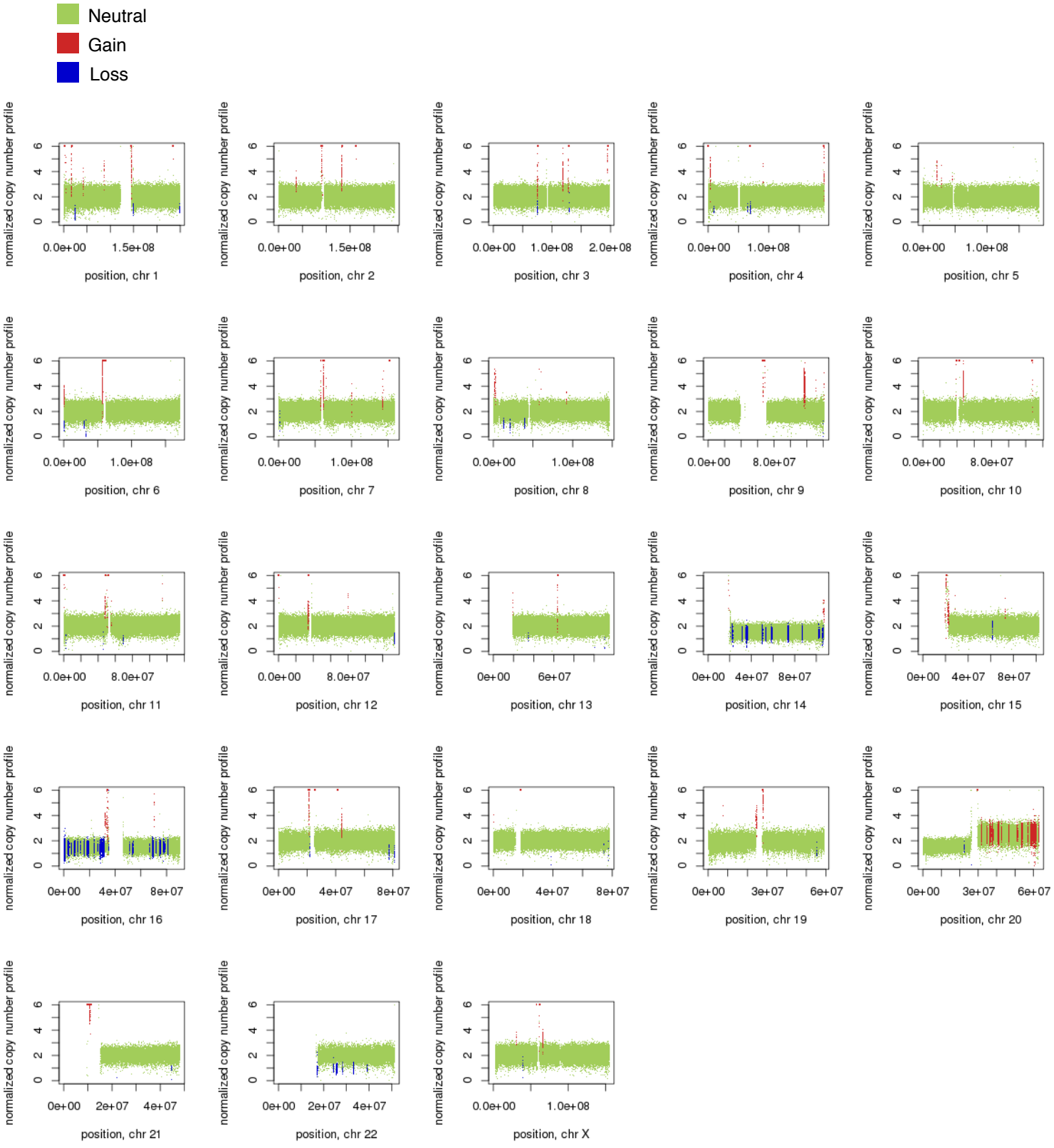
MB-Rec-11 Therapy naive tumor LOH

MB-Rec-11 Recurrence CNV



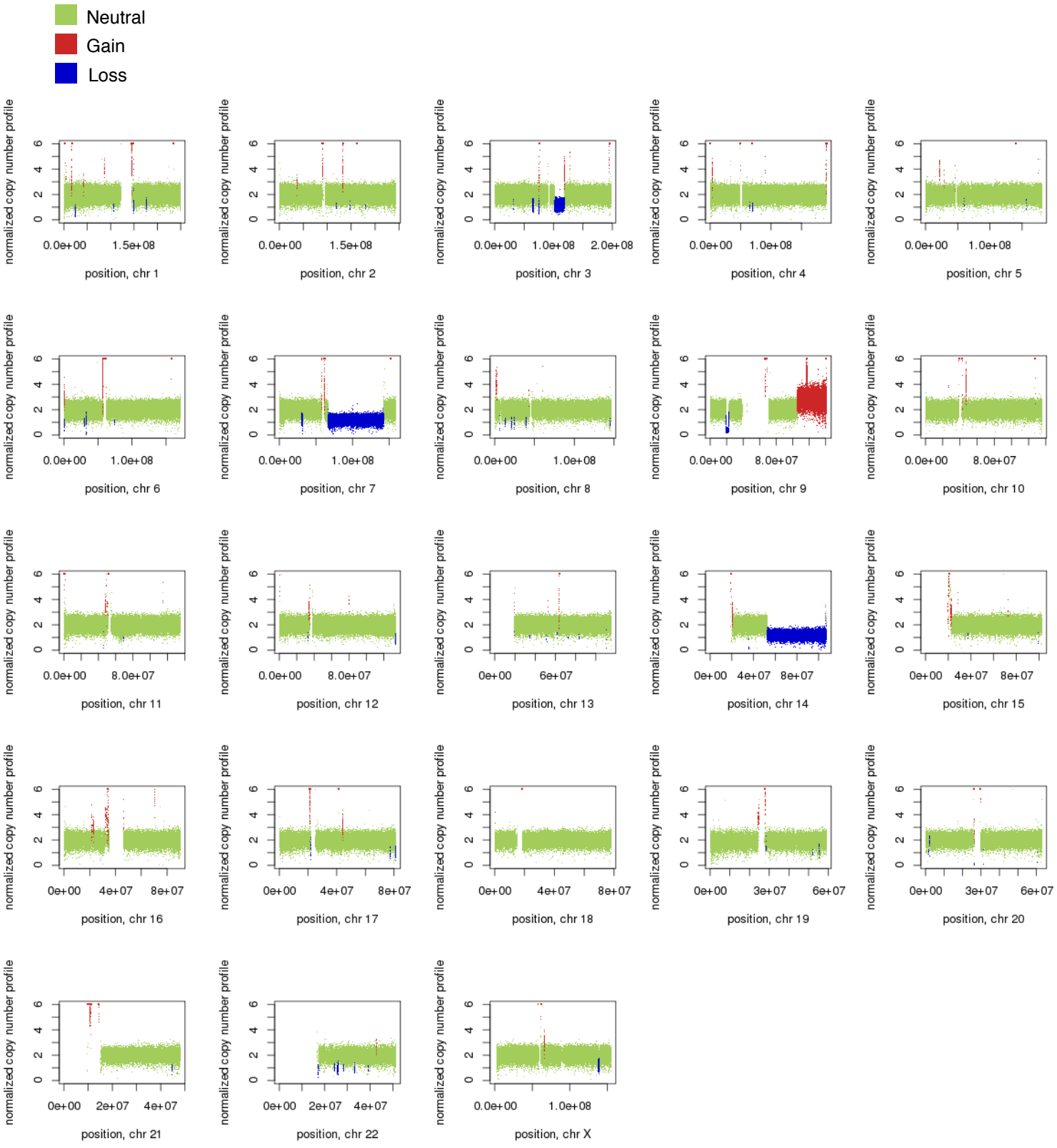
MB-Rec-11 Recurrence LOH

MB-Rec-12 Therapy naive tumor CNV



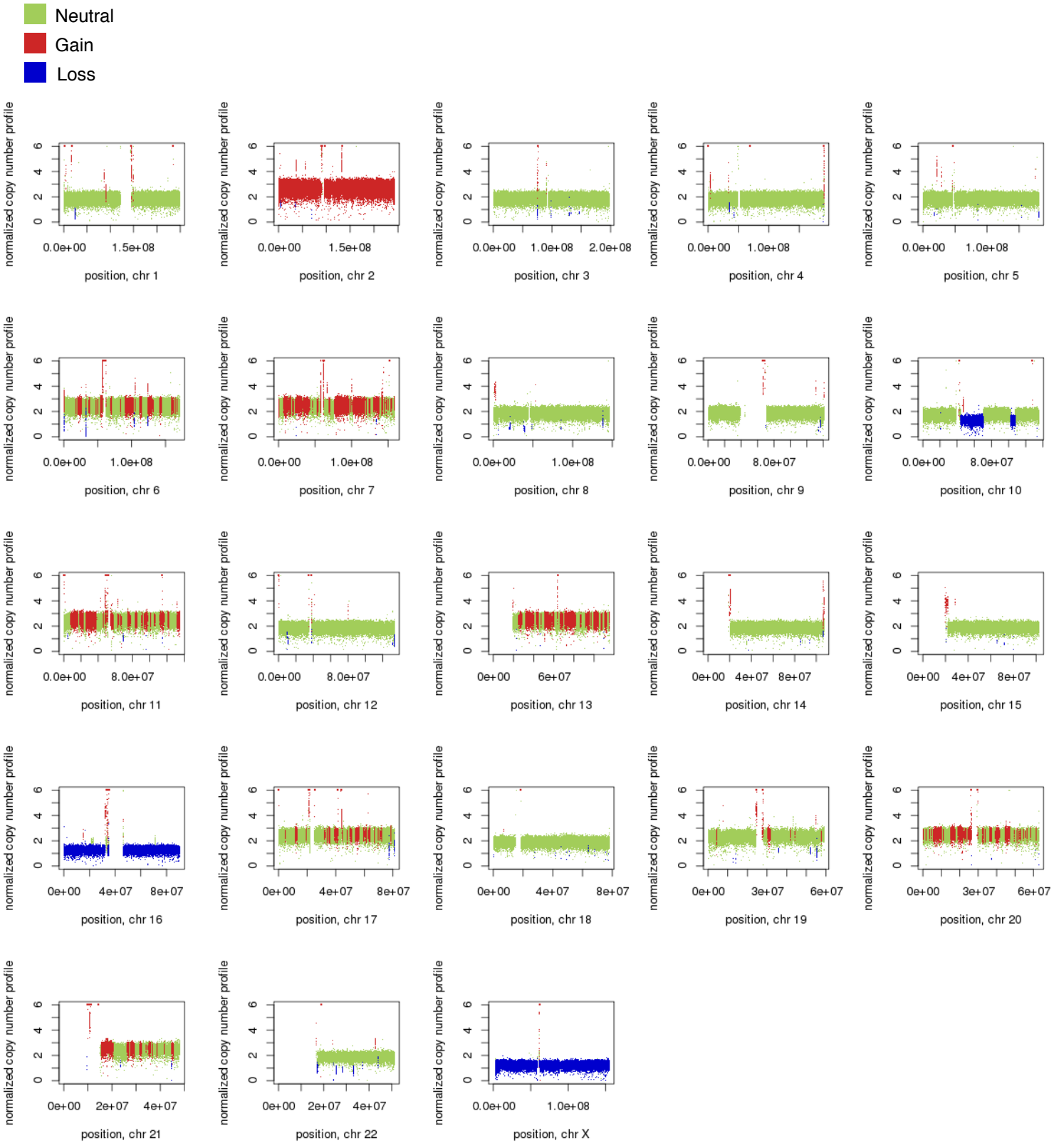
MB-Rec-12 Therapy naive tumor LOH

MB-Rec-12 Recurrence CNV

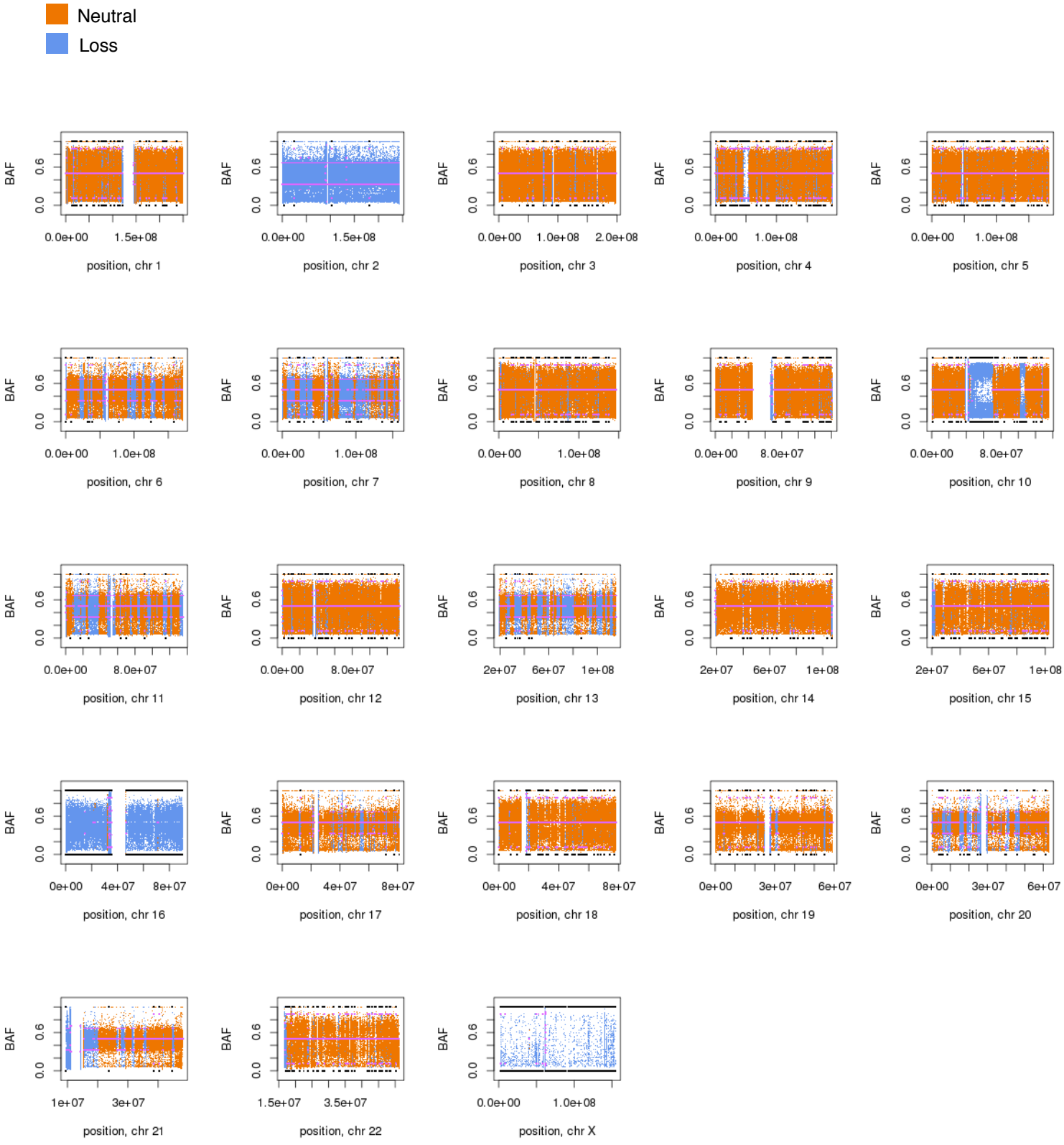


MB-Rec-12 Recurrence LOH

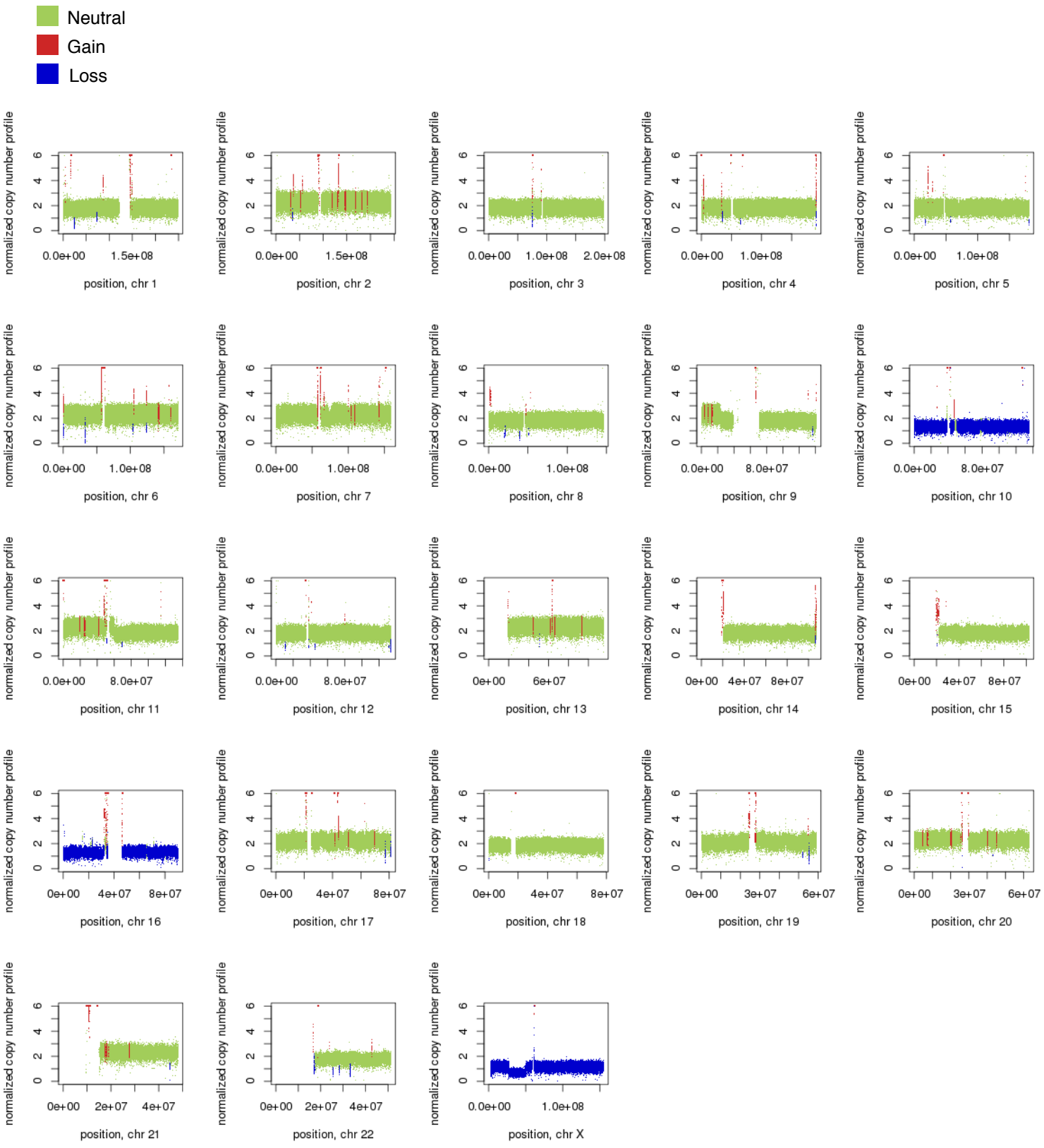
MB-Rec-13 Therapy naive tumor CNV



MB-Rec-13 Therapy naive tumor LOH

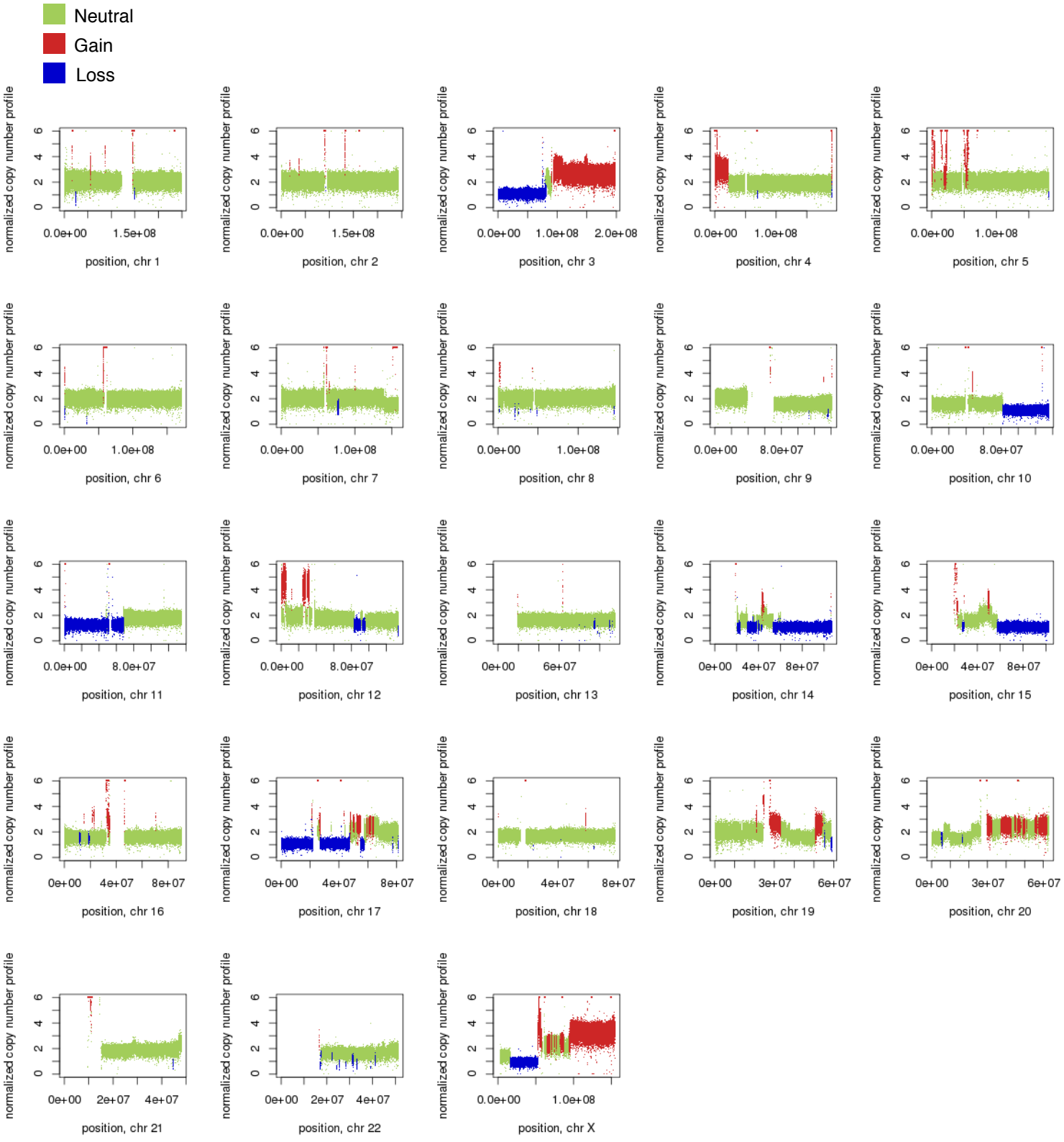


MB-Rec-13 Recurrence CNV

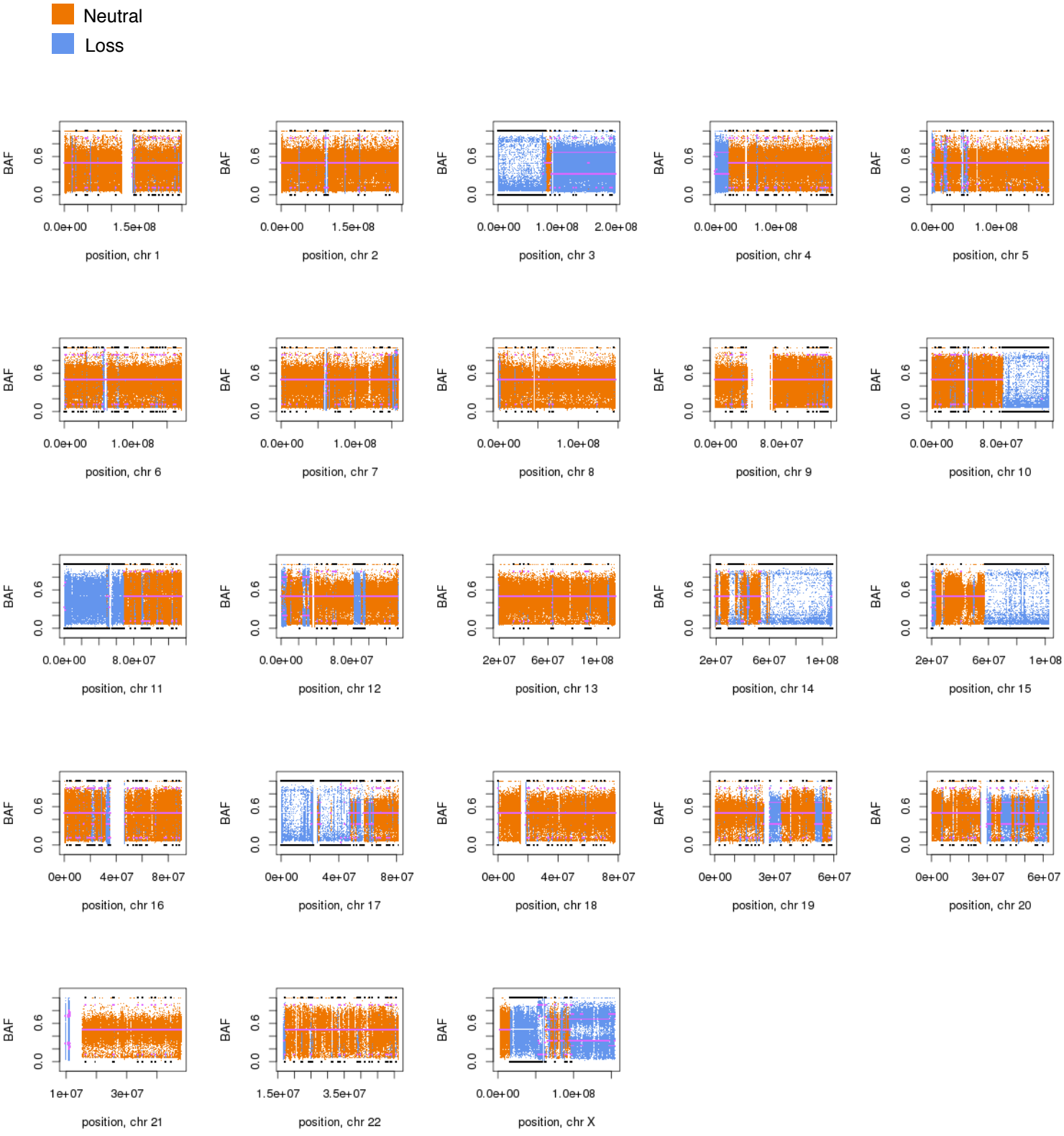


MB-Rec-13 Recurrence LOH

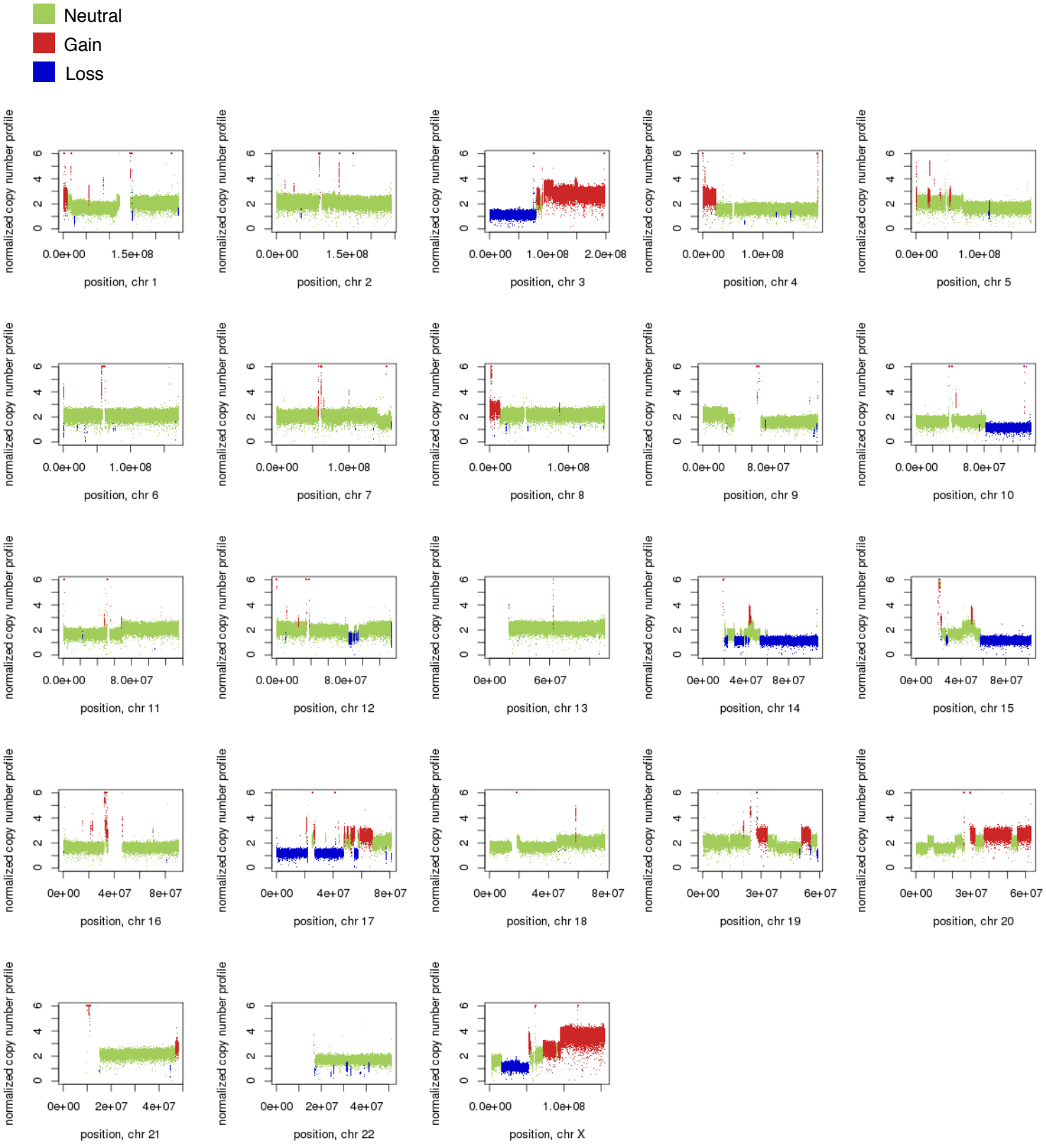
MB-Rec-14 Therapy naive tumor CNV



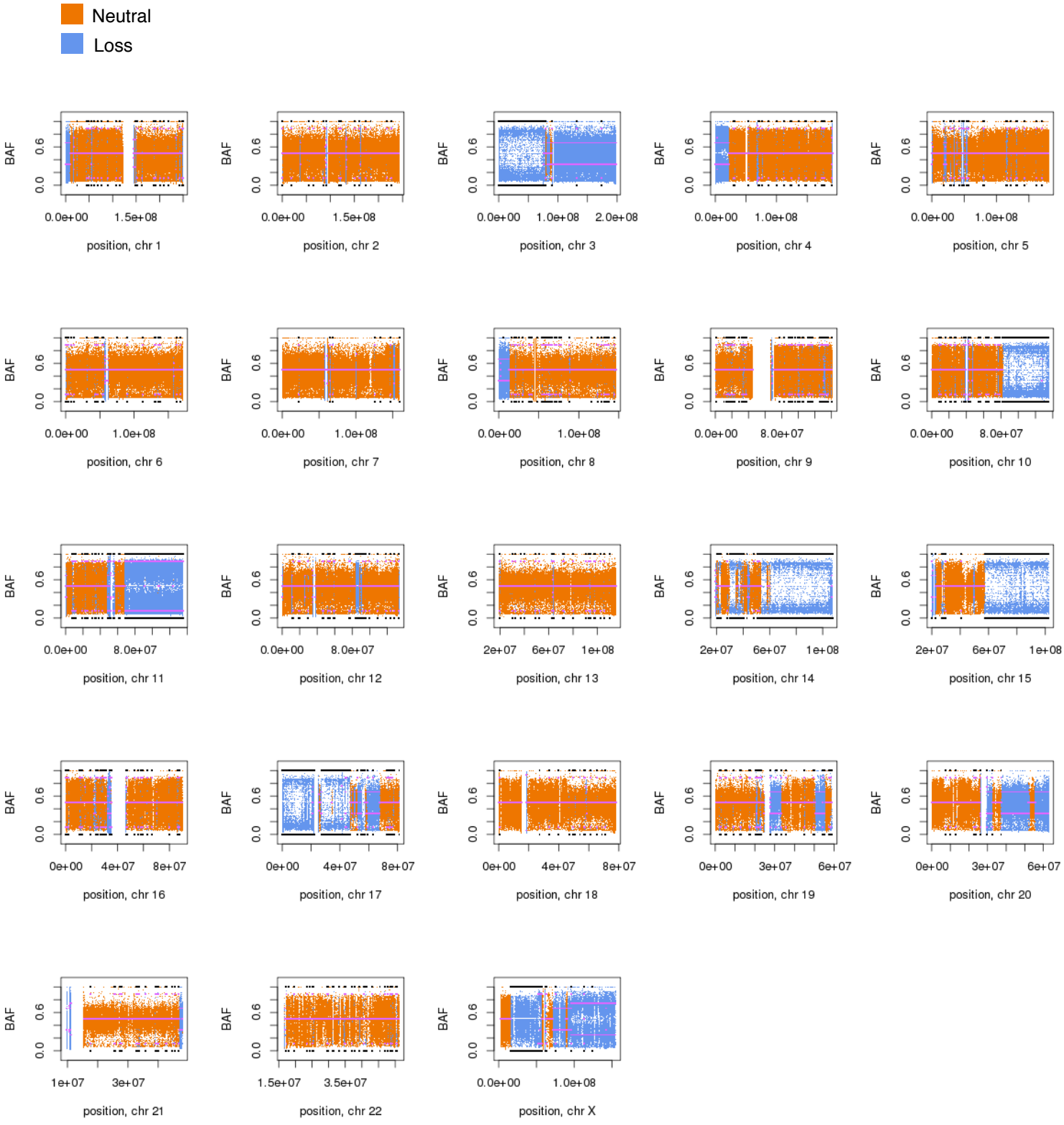
MB-Rec-14 Therapy naive tumor LOH



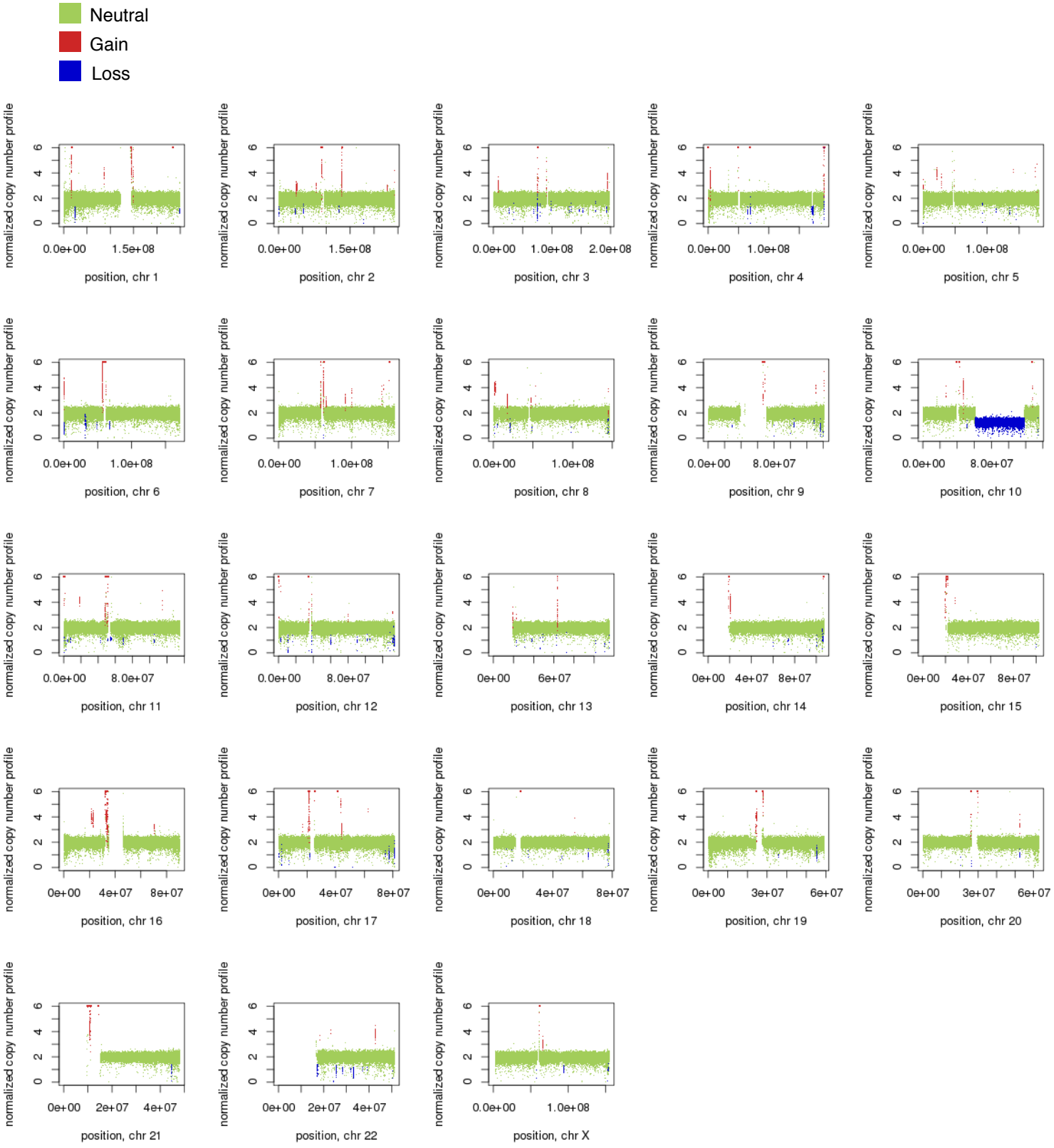
MB-Rec-14 Recurrence CNV



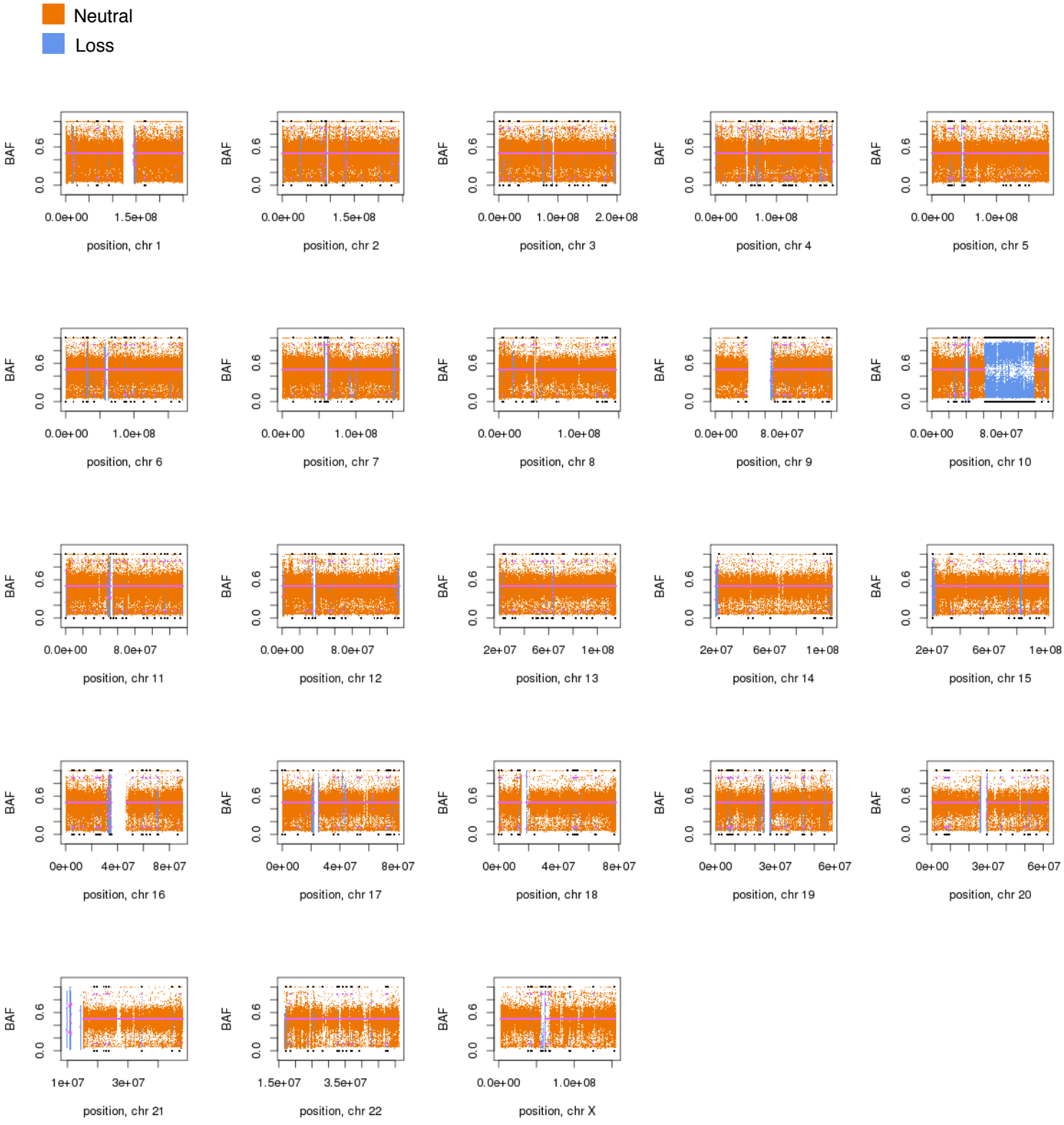
MB-Rec-14 Recurrence LOH



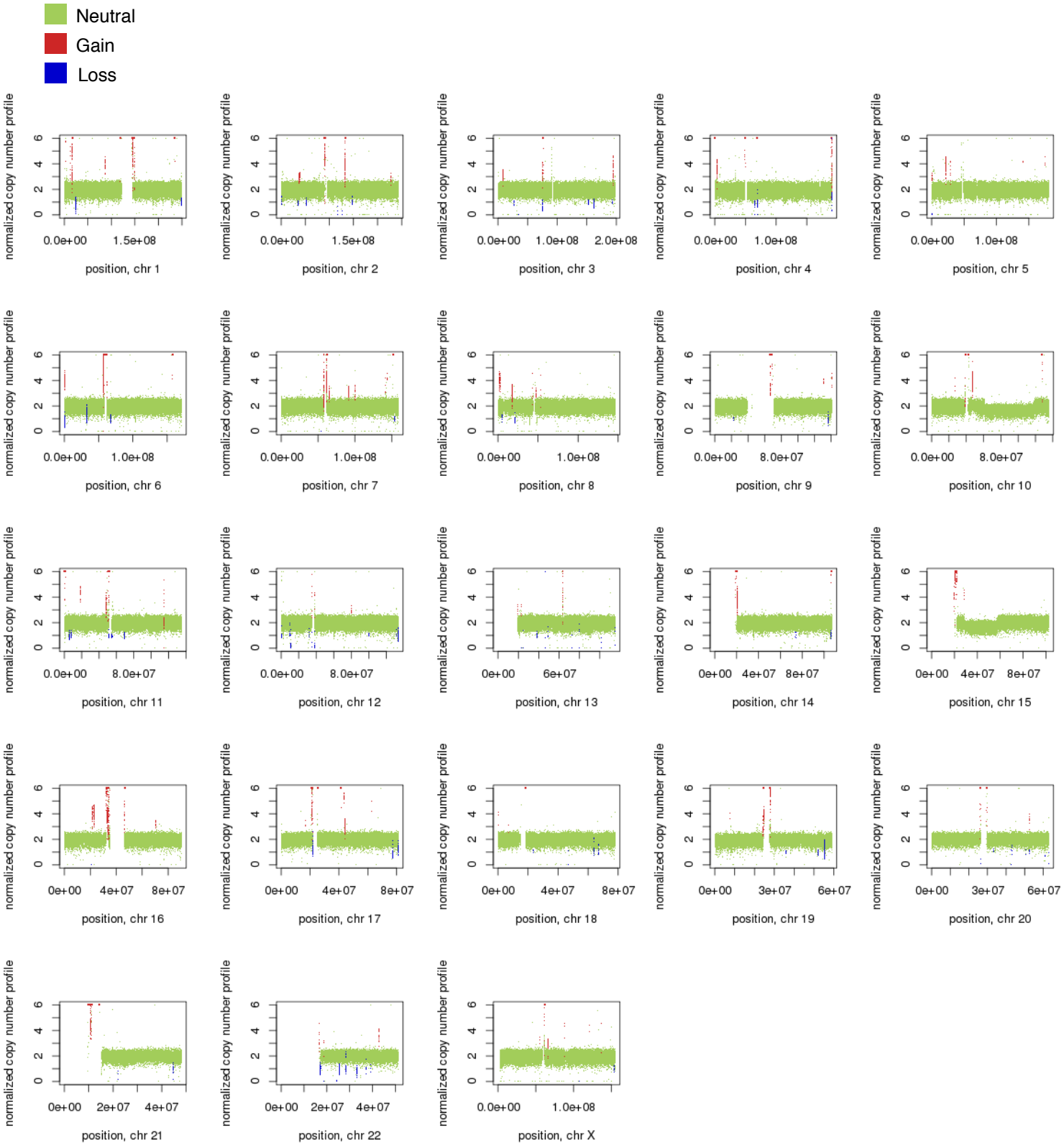
MB-Rec-15 Therapy naive tumor CNV



MB-Rec-15 Therapy naive tumor LOH

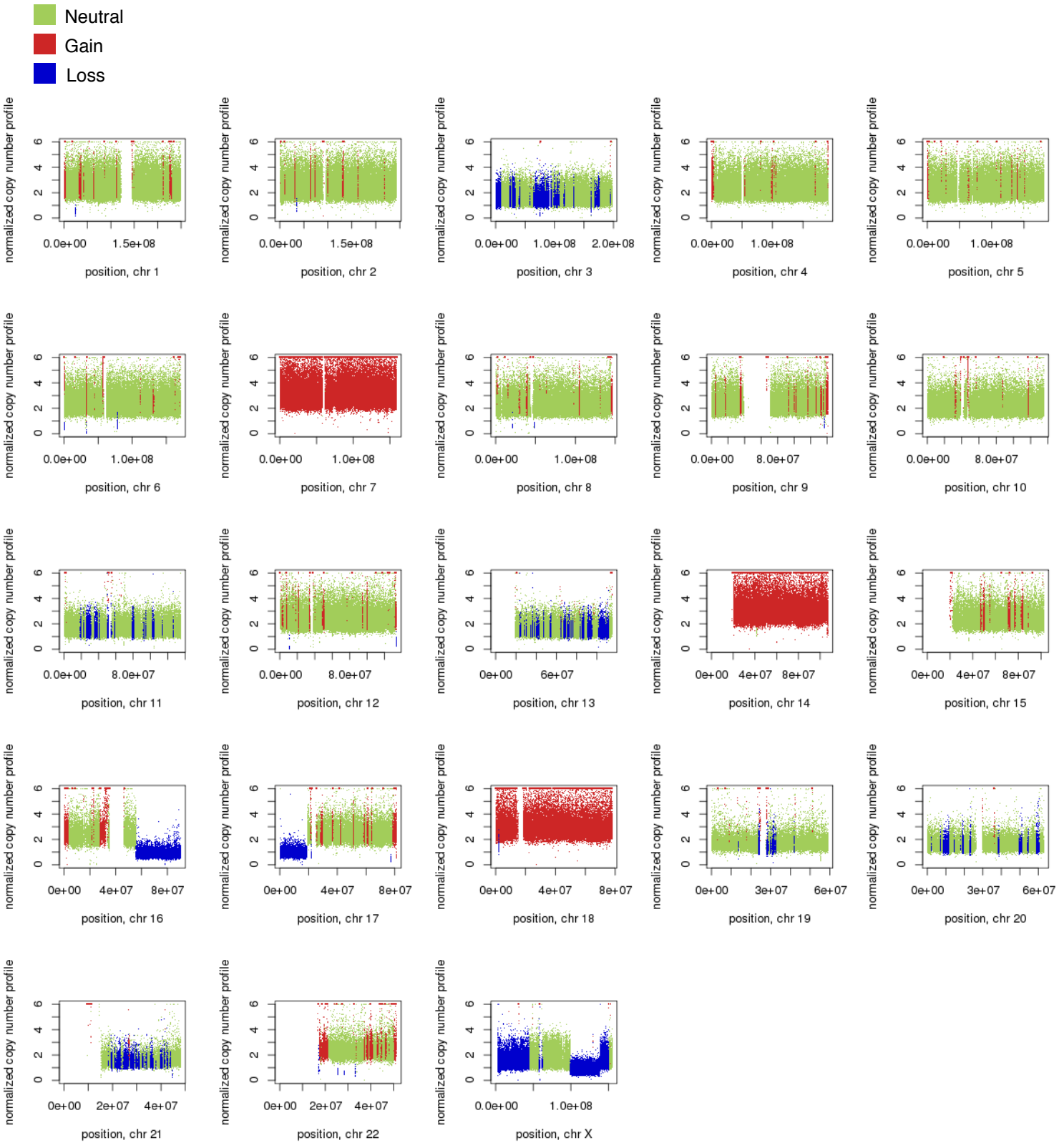


MB-Rec-15 Recurrence CNV

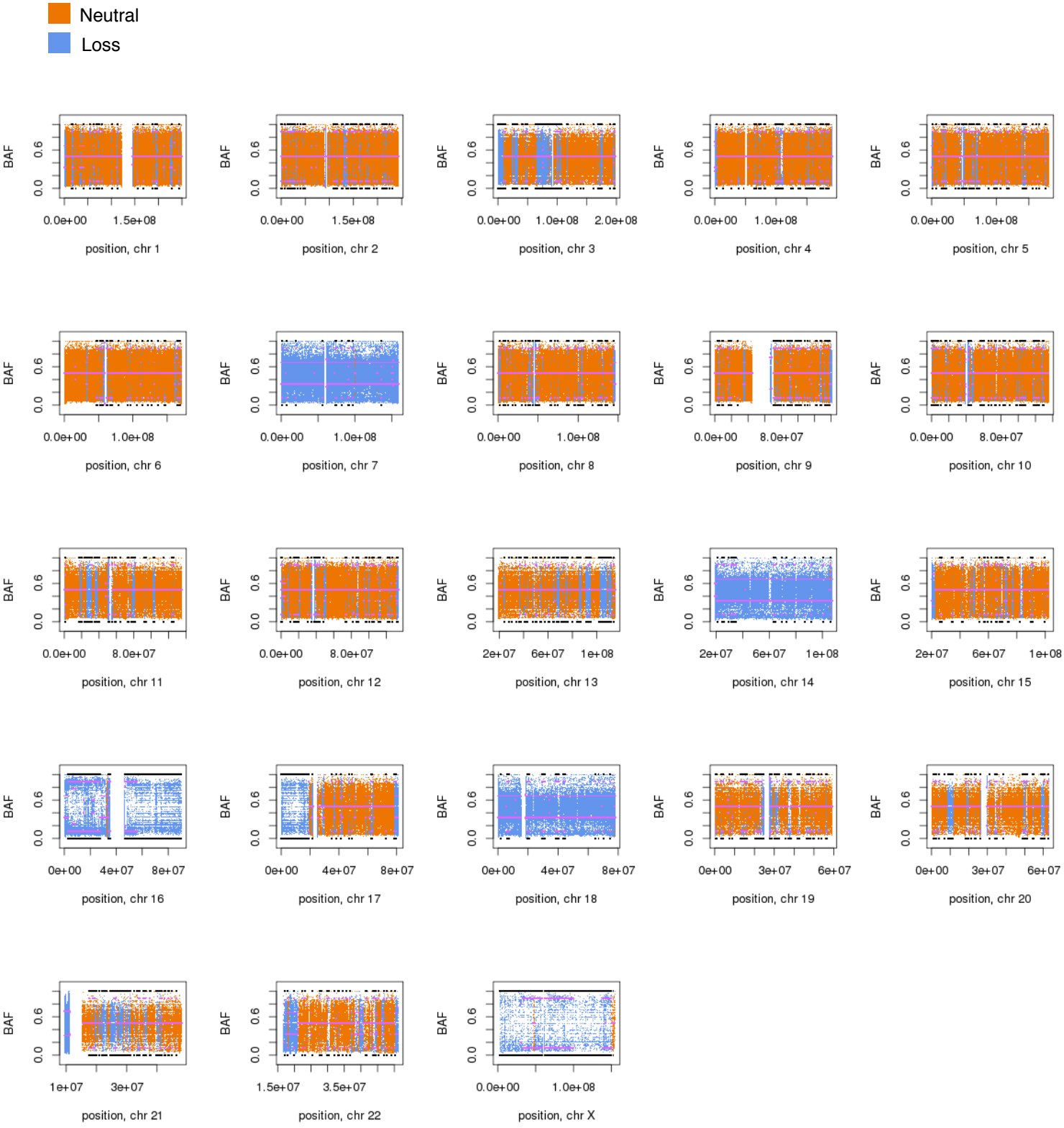


MB-Rec-15 Recurrence LOH

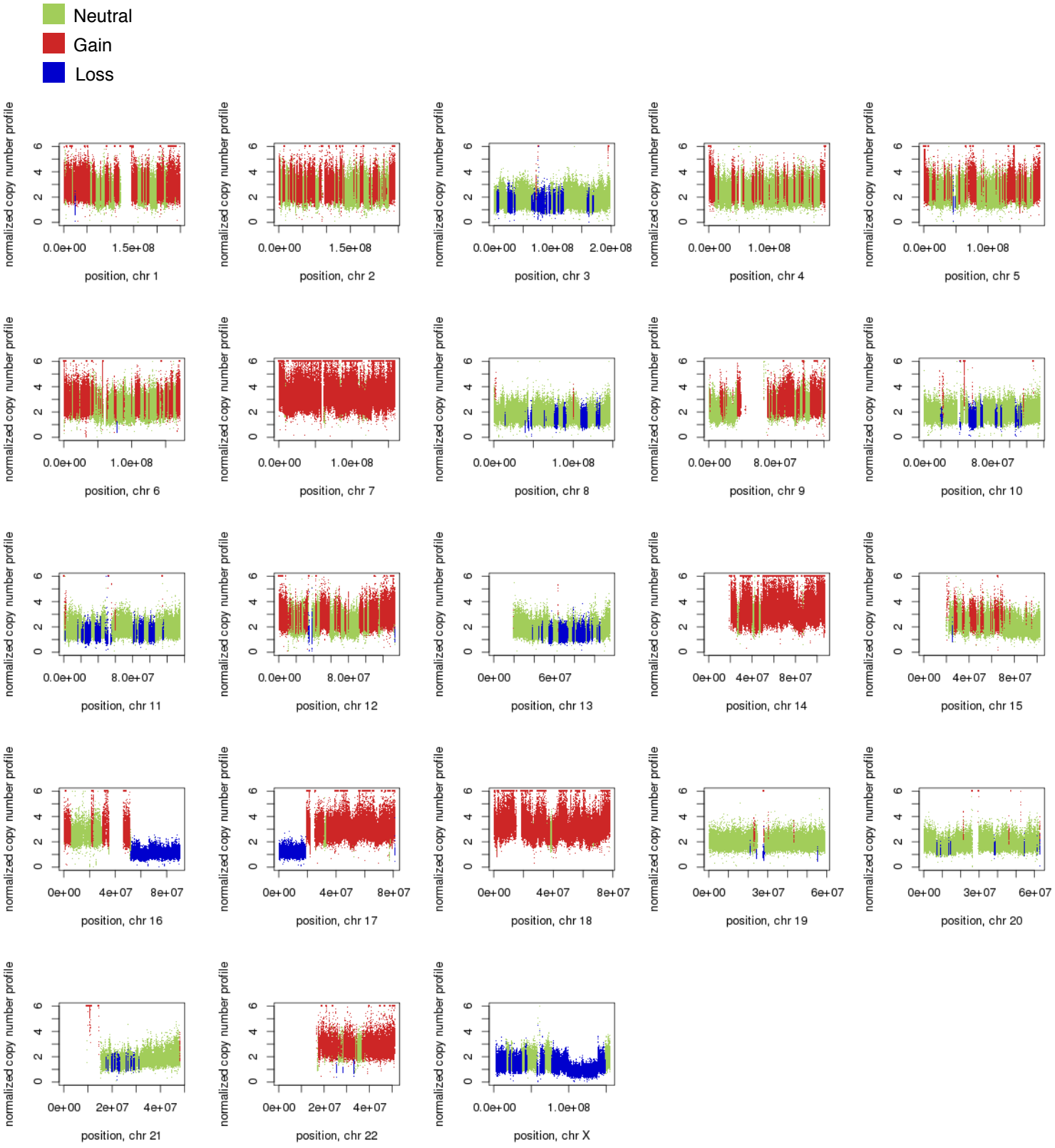
MB-Rec-16 Therapy naive tumor CNV



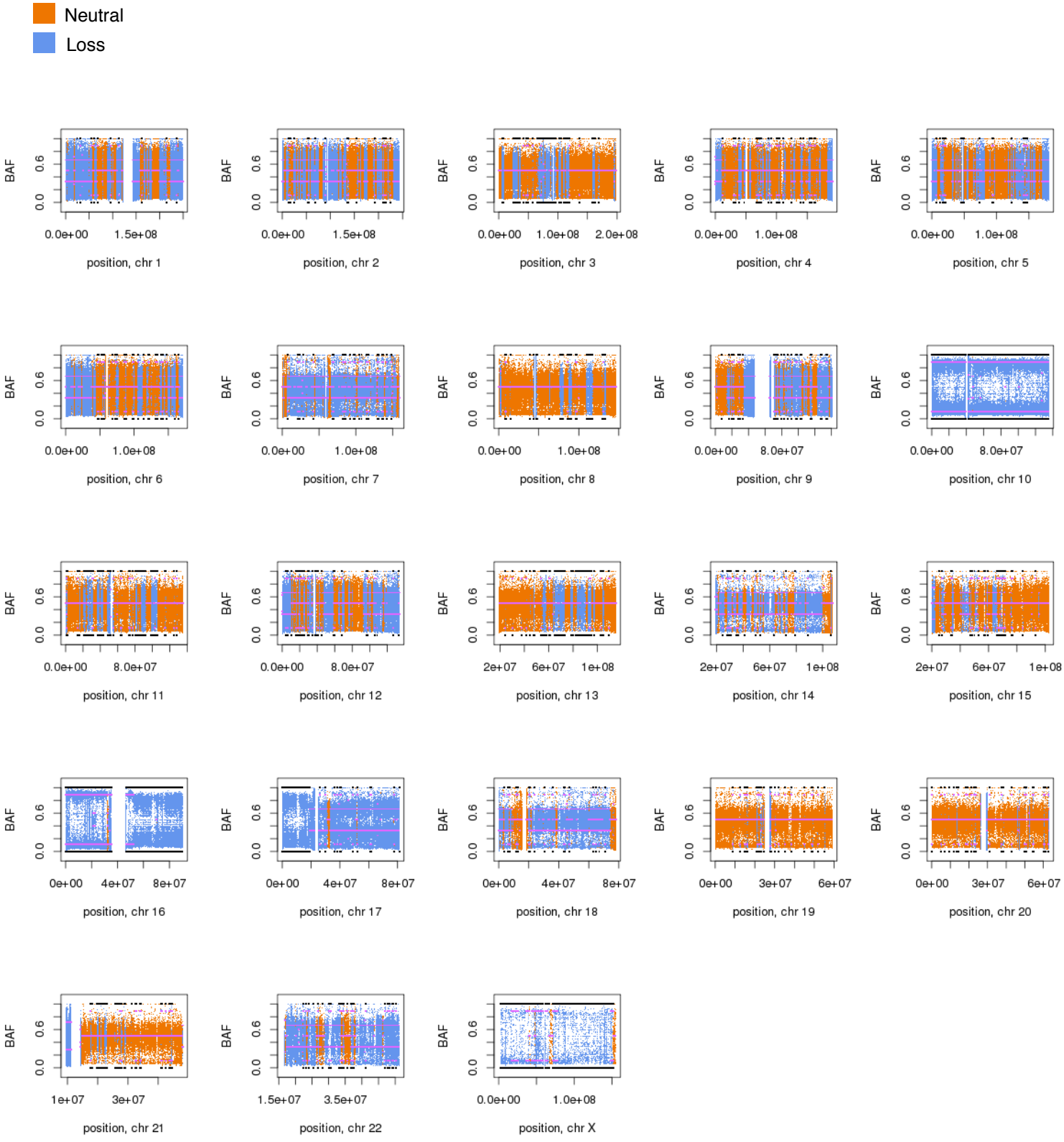
MB-Rec-16 Therapy naive tumor LOH



MB-Rec-16 Recurrence CNV



MB-Rec-16 Recurrence LOH



Driving Initiating Transposon Insertion Site Prediction

1 Rationale

While examining our SB insertion site frequency data, we noted that the read counts of the highest frequency insertions tend to amass around values that are often related to one another by a factor of two. This led us to establish the following set of assumptions:

1. causal insertions initiate or promote tumorigenesis in different cells resulting in several **distinct subpopulations**, each with a different set of insertions
2. **additional insertions are acquired** during the early growth of the tumor while the sleeping beauty system is still active
3. the prevalence of a given insertion is dictated by
   (a) the **growth profile** of the tumor cell subpopulation carrying the insertion
   (b) the **number of cell divisions** that occurred before an additional, passenger insertion event; each division results in a halving of the number cells with a given passenger insertion relative to the founder insertion(s)

The following table shows the top read counts for a single mouse annotated from this perspective:

Mouse 06-29-11 (Primary)

Reads	Gene	Note	Cell Divisions				
44989	Gli2		0				
20376	Pax7			0			
9690	Gpn1	∼20376/2		1			
5735	Hivep3	∼44989/8	2				
5113	B020004J07Rik	∼20376/4		2			
4198	Nav2				0		
3874	Impad1					0	
3818	Col4a3bp					0	
2972	Crebbp						0
2712	Ncoa1	∼44989/16	3				

The Reads column indicates the number of reads that include a transposon insertion site in the specified gene. The last section (5 columns) shows the pattern of driving initiating insertion site counts. Each column in this section is its own series. For instance, the zero in the first column shows that Gli2 is affected by a founding or driving insertion and after two cell divisions, Hivep3 was affected by the transposon and then Ncoa1 was affected in cell division three. Similarly, Pax7 (driver), Gpn1 and B020004J07Rik form another series of insertions over a number of cell divisions. The Note column indicates our explanation of the relationship of the read counts within a cell division series according to our assumptions above. In this specific example, Gli2, Pax7, Nav2, one or both of Impad1 and Col4a3bp, and Crebbp would be considered driving insertions; whereas the remaining genes would be considered passenger insertions.

2 Defining and Fitting the Model

Let us assume that tumor cell division and growth are initiated by a founder driving transposon insertion event. Let us also assume that additional insertions can occur in this cell or its daughter cells, but these are passenger insertions that introduce no selective pressure and have no effect on the growth profile of the tumor. If growth proceeds through an overall doubling of the cell population, passenger insertions should be sampled at a frequency drawn from a geometric sequence with common ratio $r = \frac{1}{2}$ and some scale factor a.

$$ar + ar^2 + ar^3 + ar^4 + ... : r = \frac{1}{2} \tag{1}$$

The observed data is then a mixture of such sequences, each with a different scale factor a, which depends on the growth rate and time of initiation relative to other subpopulations.

This model is mathematically rearranged slightly to make it easier to work with. Let us define g as a subpopulation-specific scaling factor, and d as the number of cell divisions that have occurred since a given insertion event. The number of transposons at site i in some subpopulation j, according to an exponential growth model, will be:

$$\theta_i = 2^{d_i} e^{g_j} : d_i \in \mathbb{N}, 0 \leq g_j < \log 2 \tag{2}$$

Since transposons are randomly sampled and sequenced, the variation in the observed number will be binomially distributed. However, the number of reads captured for any single transposon is small compared to the total number of reads, and so it is acceptable and more convenient to model this variation using the Poisson distribution:

$$y_i \sim \text{Poisson}(\theta_i) \tag{3}$$

The task is then to estimate, given the data, an appropriate set of values G for the subpopulation-specific scaling factors. For a given solution, the likelihood function used is:

$$L(Y|G) = \prod_i \max(\{P(y_i|g,d) : g \in G, d \in \mathbb{N}\}) \tag{4}$$

The function has no apparent closed-form solution, so a simple genetic algorithm was used to search for a set of values G in an attempt to maximize the likelihood.

Although the obvious danger of overfitting arises from using a solution where the cardinality of G is too large (which was assessed using both the Akaike and Bayesian information criterions), we found it was easier to generate a solution that didn't differ significantly from the null model where values for G will be equally spaced within its allowed range $[0, \log 2)$. This was checked by applying a standard likelihood ratio test between the found solution and the "best-case" null model (G with the same cardinality and equally spaced values starting at an optimal offset) and ensuring the p-value was less than 0.05.

Each transposon could then be associated with a value of g and d that would most likely give rise to the observed value. The site(s) with the highest d for each subpopulation was considered the location of a founder insertion.

3 Relevant Biological Themes Are Enriched In Genes With Putative Founder Insertions

Confidence in the suitability of the model was sought by comparing the biological themes associated with genes identified as containing founder insertions versus those containing passenger insertions. We collected all genes identified as founding drivers and compared them to the passenger genes interspersed between founders (this ameliorated potential bias arising from the magnitude of the insertion count). Enrichment analysis was performed using the online DAVID resource [Huang et al., 2009]. This analysis was performed on the primary and locally recurrent libraries from seven mice (02-23-112_4, 02-23-11W, 03-04-11, 04-15-11, 06-28-11, 06-29-11, 09-16-10). The results are summarized in the table below. Themes consistent with the topic of study (e.g. pathways in cancer, transcription regulation, Basal cell carcinoma) were significantly enriched for genes containing putative founder insertions. Genes with putative passenger insertions were enriched for themes that were both more general and less statistically significant. This was taken as a strong indication that this model was appropriate for the study data.

Founder Insertions (N=127)

Annotation	Number of Genes	Source	p-value	FDR
pathways in cancer	12	KEGG	1.4×10^{-6}	0.00010
DNA binding	30	GO	6.6×10^{-7}	0.00014
transcription regulation	26	PIR	4.2×10^{-6}	0.00070
transcription regulator activity	22	GO	1.4×10^{-5}	0.0015
transcription factor activity	16	GO	9.3×10^{-5}	0.0065
transcription	25	GO	1.2×10^{-4}	0.0066
alternative splicing	46	PIR	1.1×10^{-4}	0.0095
Basal cell carcinoma	5	KEGG	2.8×10^{-4}	0.011
DNA-binding	20	PIR	7.1×10^{-4}	0.029
small cell lung cancer	5	KEGG	1.5×10^{-4}	0.037
nucleus	38	PIR	1.3×10^{-3}	0.041

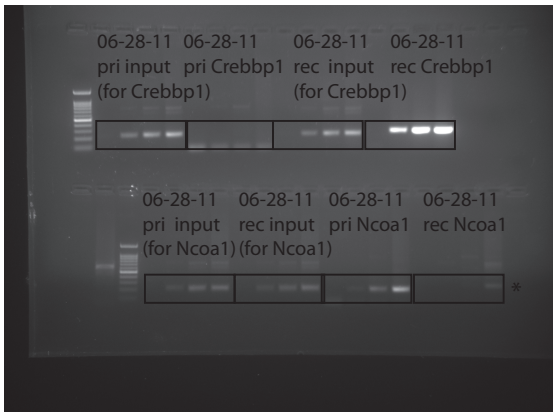
Passenger Insertions (N=119)

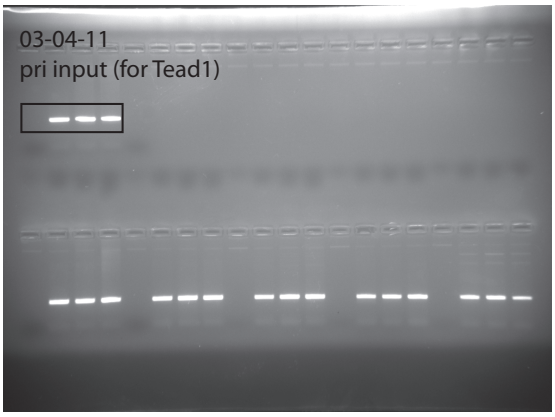
Annotation	Number of Genes	Source	p-value	FDR
positive regulation of macromolecule metabolic process	15	GO	3.4×10^{-6}	0.0032
positive regulation of transcription	12	GO	2.8×10^{-5}	0.0045
positive regulation of biosynthetic process	13	GO	2.4×10^{-5}	0.0045
positive regulation of macromolecule biosynthetic process	13	GO	1.5×10^{-5}	0.0046
positive regulation of gene expression	12	GO	3.6×10^{-5}	0.0049
positive regulation of nucleobase, nucleoside, nucleotide and nucleic acid metabolic process	12	GO	5.4×10^{-5}	0.0051
positive regulation of cellular biosynthetic process	13	GO	2.2×10^{-5}	0.0052
positive regulation of RNA metabolic process	11	GO	5.3×10^{-5}	0.0055
positive regulation of transcription, DNA-dependent	11	GO	5.0×10^{-5}	0.0059
positive regulation of nitrogen compound metabolic process	13	GO	1.4×10^{-5}	0.0064
positive regulation of transcription from RNA polymerase II promoter	10	GO	8.7×10^{-5}	0.0074

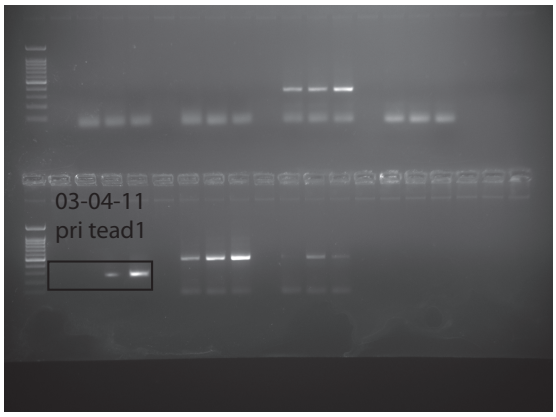
References

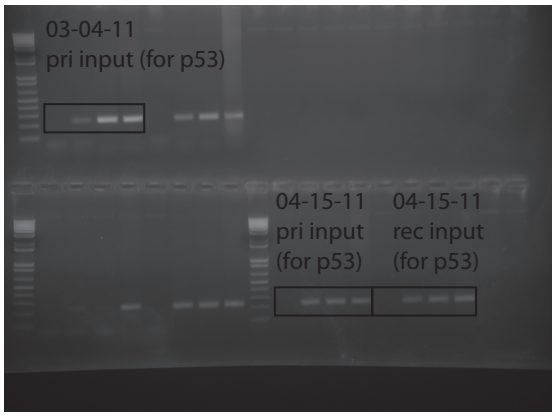
Da Wei a. . W. Huang, Brad T. Sherman, and Richard A. Lempicki. Bioinformatics enrichment tools: paths toward the comprehensive func-

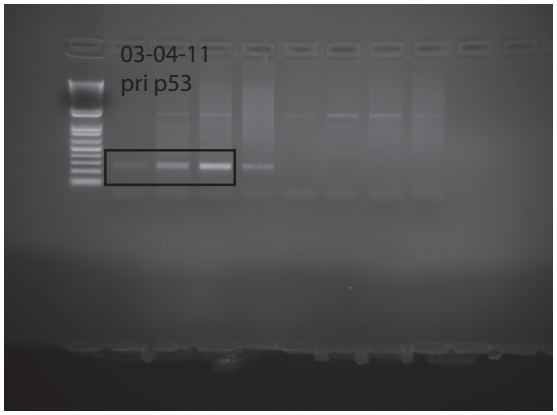
tional analysis of large gene lists. *Nucleic acids research*, 37(1):1–13, January 2009. ISSN 1362-4962. doi: 10.1093/nar/gkn923. URL http://dx.doi.org/10.1093/nar/gkn923.

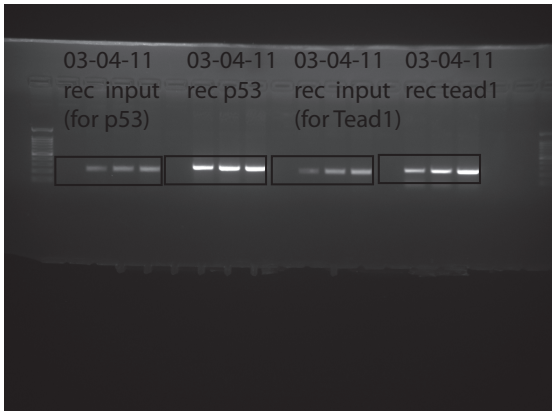

06-28-11
pri input
(for Crebbp1)
06-28-11
pri Crebbp1
06-28-11
rec input
(for Crebbp1)
06-28-11
rec Crebbp1
06-28-11
pri input
(for Ncoa1)
06-28-11
rec input
(for Ncoa1)
06-28-11
pri Ncoa1
06-28-11
rec Ncoa1
*


03-04-11
pri input (for Tead1)


03-04-11
pri tead1


03-04-11
pri input (for p53)
04-15-11
pri input
(for p53)
04-15-11
rec input
(for p53)


03-04-11
pri p53


03-04-11
rec input
(for p53)
03-04-11
rec p53
03-04-11
rec input
(for Tead1)
03-04-11
rec tead1


04-15-11
pri p53
04-15-11
rec p53