Supplementary Notes and Tables for "**Chromosomal alterations among hematopoietic clones in Japan**"

Chikashi Terao, Akari Suzuki, Yukihide Momozawa, Masato Akiyama, Kazuyoshi Ishigaki, Kazuhiko Yamamoto, Koichi Matsuda, Yoshinori Murakami, Steven A McCarroll, Michiaki Kubo, Po-Ru Loh, Yoichiro Kamatani

Contents

Supplementary Notes

1. Detection of autosomal mCA

1.1. Calculation of BAF and LRR from genotype intensity

1.1.1. Computation of cluster median of X and Y signal intensities.

We calculated median of X and Y intensities in each genotype. If a cluster contained fewer than 10 calls, we set its median to missing.

1.1.2. Affine-normalization and correction by GC and CpG contents.

We took a similar approach to Jacobs et al[1], and our method is the same as Loh et al[10] regarding this calculation. A pair of multiple variate linear regression was carried out to correct effects of GC and CpG contents. Median of centers (X, Y) were set as two dependent variables to be expected (X, Y) values and modeled with the use of X and Y values in each subject of the genotype, GC and CpG contents in 9 windows of 50, 0.1k, 0.5k, 1k, 10k, 50k, 100k, 250k and 1Mbp at the center of SNP_m as covariates as follows.

$$\mathrm{X}_{m,i,exp} = \mathrm{e}_{x,i} + \mathrm{X_{m,i}}\alpha_{\mathrm{x,i}} + \mathrm{Y_{m,i}}\alpha_{\mathrm{y,i}} + \sum_{k=1}^{9}\sum_{p=1}^{2}[(f_{m,k}^{GC})^{p}\cdot\alpha_{i,k,p}^{GC} + (f_{m,k}^{CpG})^{p}\cdot\alpha_{i,k,p}^{CpG} \quad]$$

$$\mathrm{Y}_{m,i,exp} = \mathrm{e}_{y,i} + \mathrm{X_{m,i}}\beta_{\mathrm{x,i}} + \mathrm{Y_{m,i}}\beta_{\mathrm{y,i}} + \sum_{k=1}^{9}\sum_{p=1}^{2}[(f_{m,k}^{GC})^{p}\cdot\beta_{i,k,p}^{GC} + (f_{m,k}^{CpG})^{p}\cdot\beta_{i,k,p}^{CpG} \quad]$$

where $X_{m,i,exp}$ and $Y_{m,i,exp}$ are median of X and Y intensity in a cluster of a genotype of i-th individual in m-th variant, respectively, $X_{m,i}$ and $Y_{m,i}$ are X and Y values in i-th individual for m-th variant, respectively, $\alpha_{x,i}$, $\beta_{y,i}$ are effect sizes of calculated X and Y for i-th individual, respectively, $f_{m,k}^{GC}$ or and $f_{m,k}^{CpG}$ are fraction of GC and CpG in k-th window, $\alpha_{i,k,p}^{CpG}$ and $\beta_{i,k,p}^{GC}$, are effect size of GC and CpG in i-th individual for k-th window, respectively and $\mathrm{e}_{x,i}$ and $\mathrm{e}_{y,i}$ are the error terms of X and Y for i-th individual, respectively.

The multi-variate linear regression analyses were conducted per individual (~179k sets of models), assuming fixed effects of X, Y intensities, GC and CpG waves across variants.

The GC content was computed with the use of bedtools on the hg19 reference. The CpG content was calculated with the use of EpiGRAPH CpG annotation[49]. Corrected X, Y values were calculated by expected X and Y values minus residuals of X and Y.

1.1.3. Computation of means of corrected X and Y values.

Means of corrected X and Y calculated above in the three cluster centers for each genotype were computed.

1.1.4. Transformation of corrected X and Y intensities to θ and R values.

We transformed corrected X and Y to θ and R as follows;

$$\theta = \frac{2}{\pi} \cdot \arctan\left(\frac{Y}{X}\right)$$

$$log_2 R = X + Y$$

This follows the method by Staaf et al[43] and differs from the method by Loh et al[10].

1.1.5. Transformation of θ and R to LRR and BAF values.

We computed cluster centers of each genotype to obtain θ and R in three cluster centers. We conducted linear interpolation between the cluster centers to estimate expected $\log_2$R based on θ in each individual. If cluster centers of homozygotes were missing, we used reflection of the cluster center in the opposite homozygote genotype across the vertical line on the center of heterozygote genotype.

1.1.6. Mean shift of LRR.

We noticed that LRR in our dataset had slightly downward bias. To avoid noise in mosaic calls due to this bias, we shifted LRR values in each genotype for each variant to have mean 0 (mean shift).

1.2. Calling mosaic events with the use of BAF and LRR

1.2.1. Filtering constitutional duplications

The 25 states corresponding to phased BAF deviation (from -0.24 to 0.24 with interval of 0.02) were used in HMM. We assume events in the state revealed mean BAF deviation equal to state value (-0.24 – 0.24 with interval of 0.02) with empirical standard deviation computed across genotyping results. We capped z-score of 4. Transition probability was determined 0.003 from

zero to non-zero state, 0.001 from a state to its negative value (phase switch error). Regions to mask were selected by computing maximum likelihood (Viterbi path) and examining contiguous non-zero states. We masked regions of likely constitutional duplications with <2Mb with |ΔBAF|>0.1 and LRR>0.1 and their 2Mb nearby regions.

1.2.2. Parameterized hidden Markov model for event detection

We used a family of 3 states HMMs parametrized by deviation of BAF, θ, namely, {-θ,0, θ} taking phase switch errors into account with transition matrix slightly different from 1st step (from±θ to 0, 0.0003, from 0 to ±θ, 0.004 * 0.0003 and switch error of 0.001). For acrocentric chromosomes (no p-arm genotypes), starting probability of non-zero state was decreased by a factor of 0.2. Like the 1st step, we assume events in the state revealed mean BAF deviation equal to state value with empirical standard deviation computed across genotyping results. We capped z-score of 2.

1.2.3. Calling existence of an event: likelihood ratio test statistics

Based on a sequence of phased BAF deviation (ΔBAF) on each chromosome, we can analyze whether the sequence of observed ΔBAF can be explained by presence of mosaic events with the states defined above. We compute the likelihood ratio statistics with the use of a family of HMM paths parameterized by θ and modeled a total probability observing the sequence ofΔBAF. Likelihood ratio for θ can be modeled by $\mathrm{f}(\Delta\mathrm{BAF}) = \frac{\mathrm{L}(0 \mid \Delta\mathrm{BAF})}{sup_{\theta}\{L(\theta \mid \Delta BAF)\}}$

0 of θ indicates no mosaic events. We discretized θ to run from 0.001 to 0.25 in 50 multiplicative steps in practice.

1.2.4. Calling event boundaries

Boundaries of called events were determined based on the consensus of mosaic calls in five samples taken from the posterior of the HMM using the likelihood-maximizing value for θ.

1.2.5. Calling copy number

Since we noticed that the steepness of slopes in BAF-LRR space to distinguish possible gains and losses from CN-LOH in the current data set were quite different from the previous study (this is not limited to raw or transformed BAF and LRR values in the current study), we did not

use the previously-defined threshold of slopes. We defined 1.3 and -1.05 to distinguish gains and losses from CN-LOH, respectively. We called copy number only if the most likely call was at least 10 times more likely than the next-most likely call (~90% confidence) in which we hypothesized that LRR of mosaic events follow normal distribution of (mu, delta). We evaluated which components affect undetection rate of mCAs (unconfidently classifying events) after applying the downstream filters mentioned below. The results of the analyses are shown in section 1.5 below.

1.2.6. Filtering possible constitutional duplications after calling mosaic events.

We excluded possible constitutional duplications with length >10 Mb with LRR > 0.35 or LRR >0.2 and |ΔBAF|>0.16. We also filtered events with length <10Mb with LRR >0.2 or LRR>0.1 and |ΔBAF|>0.1. These thresholds were defined by the previous study[10].

We further excluded events classified as gain or unknown and satisfying any of the following criteria: (1) less than 5Mbp length and contained in segmental duplications reported in 1000 genome projects with extension of 0.1Mbp in both directions (2) length <5Mb with LRR>0.13, or (3) length<2Mbp and LRR>0.02.

We excluded events with LRR>-0.1 and observed heterozygosity within events less than one third of expected heterozygosity.

1.2.7. Estimating fractions of cells carrying mosaic events.

Cell fraction with mosaic events was calculated by the method previously developed[1] as follows.

$muDiff = 2 \times 0.01 \times BAF$

$AF_{loss} = 2*muDiff / (1+muDiff)$

$AF_{CN\text{-}LOH} = muDiff$

$AF_{gain} = 2*muDiff / (1-muDiff)$

where $AF_{mosaic\ type}$ indicates cell fraction of the corresponding mosaic types.

1.3. Exclusion of samples of possible contamination.

We excluded three samples with mosaic events in more than 7 chromosomes among which they

did not carry loss or gain events but many unknown events strongly suggesting contamination.

1.4. Exclusion of events of possible non-mosaic whole chromosomal trisomy.

We additionally filtered mosaic events on chromosomes with chromosome-wide mean LRR more than 0.175. We noticed that long-range phasing and phasing itself did not always produce a clear sequence of deviation of BAF across chromosomal positions in subjects having possible trisomy. This is compatible with previous findings of trisomy 21 composed of one paternal and two maternal chromosomes without a duplicated chromosome. In addition to excess departure of BAF and LRR in a single chromosome, this type of signal characterizes the existence of a trisomy originating from three different chromosomes.

1.5. Classification of mCAs depends on BAF and length of mCAs.

Copy number was determined if we confidently classified the events into one specific mosaic type (the most likely mosaic type is more than 10 times more likely than the 2nd most likely mosaic type). We analyzed which parameters affect classification rate. We stratified deviation of BAF of the mCAs and showed that classifiability is large driven by BAF deviation (Fig.S1.5.1).

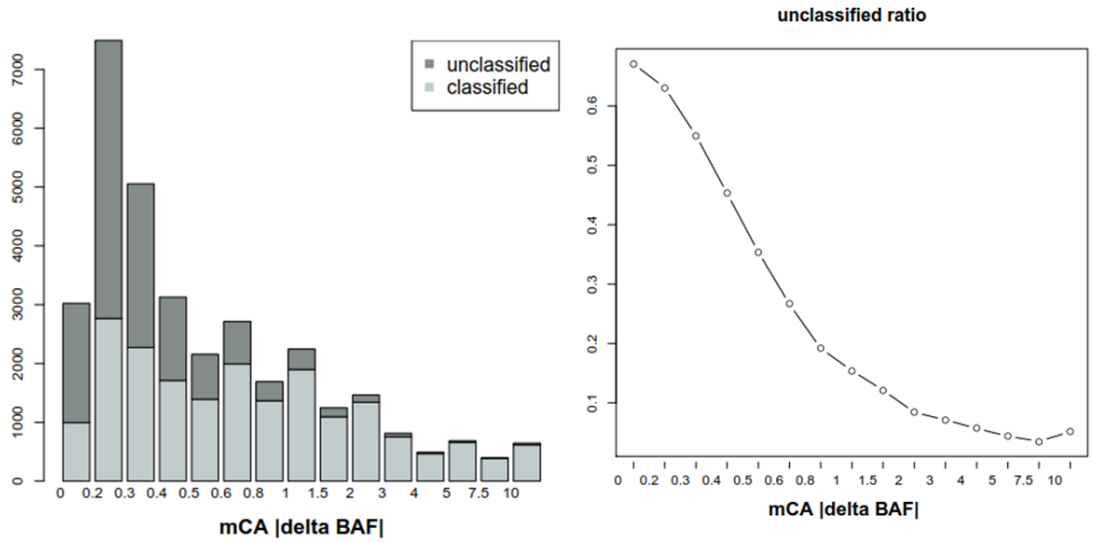

Fig. S1.5.1. Classifiability of events as a function of BAF deviation.

The numbers of classified and unclassified events are shown according to BAF deviation of the events in the left panel. The right panel indicates ratio of unclassified events among the events in the bins of BAF deviation. BAF deviations are absolute values and multiplied by 100. Unclassified events are heavily skewed towards mCAs with small BAF.

Furthermore, in stratified BAF bins, we observed that short mCAs are more difficult to classify (Fig.S1.5.2). This is very reasonable since short mosaic events give us limited information for classification due to limited numbers of genotyping probes spanned by the events.

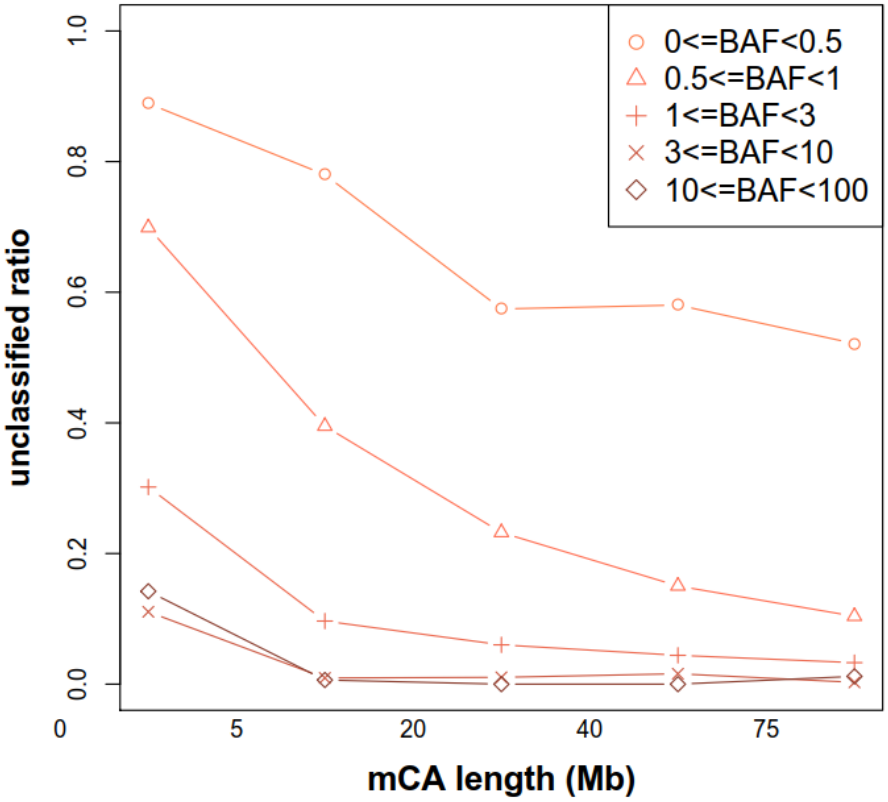

Fig. S1.5.2. Classifiability of events as a function of mCA length.

We first divide events according to their BAF deviation and then subdivide them based on length of the events. Unclassified ratio is computed in each subdivision of the mCAs. BAF deviations are absolute values and multiplied by 100. Shorter mCAs are consistently more difficult to classify. This trend is consistently observed across different BAF bins.

2.Mosaic events in each chromosome and breakpoints of loss events

2.1. mosaic events in each chromosome

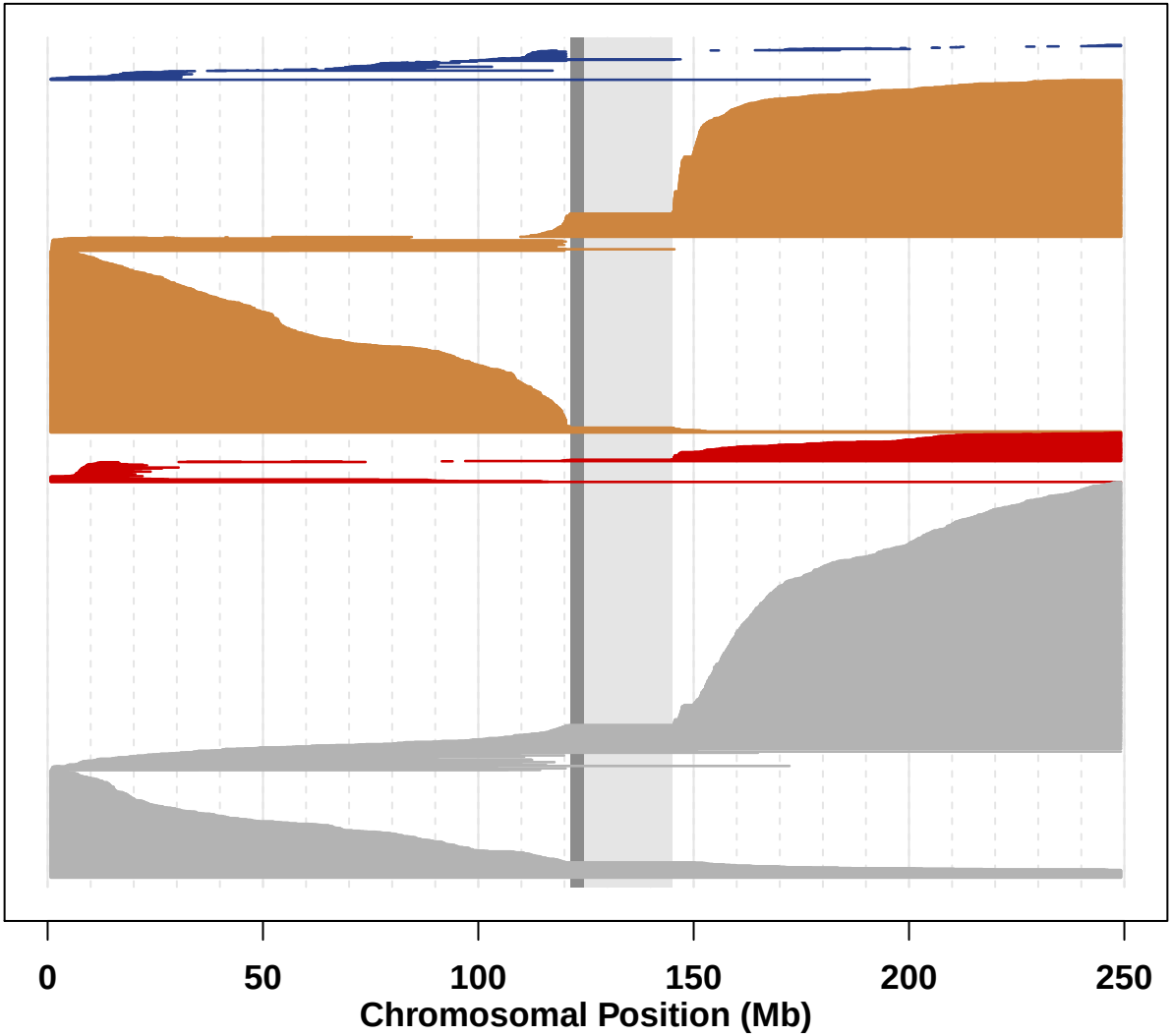

Fig. S2.1.1 A landscape of mosaic events in chromosome 1.

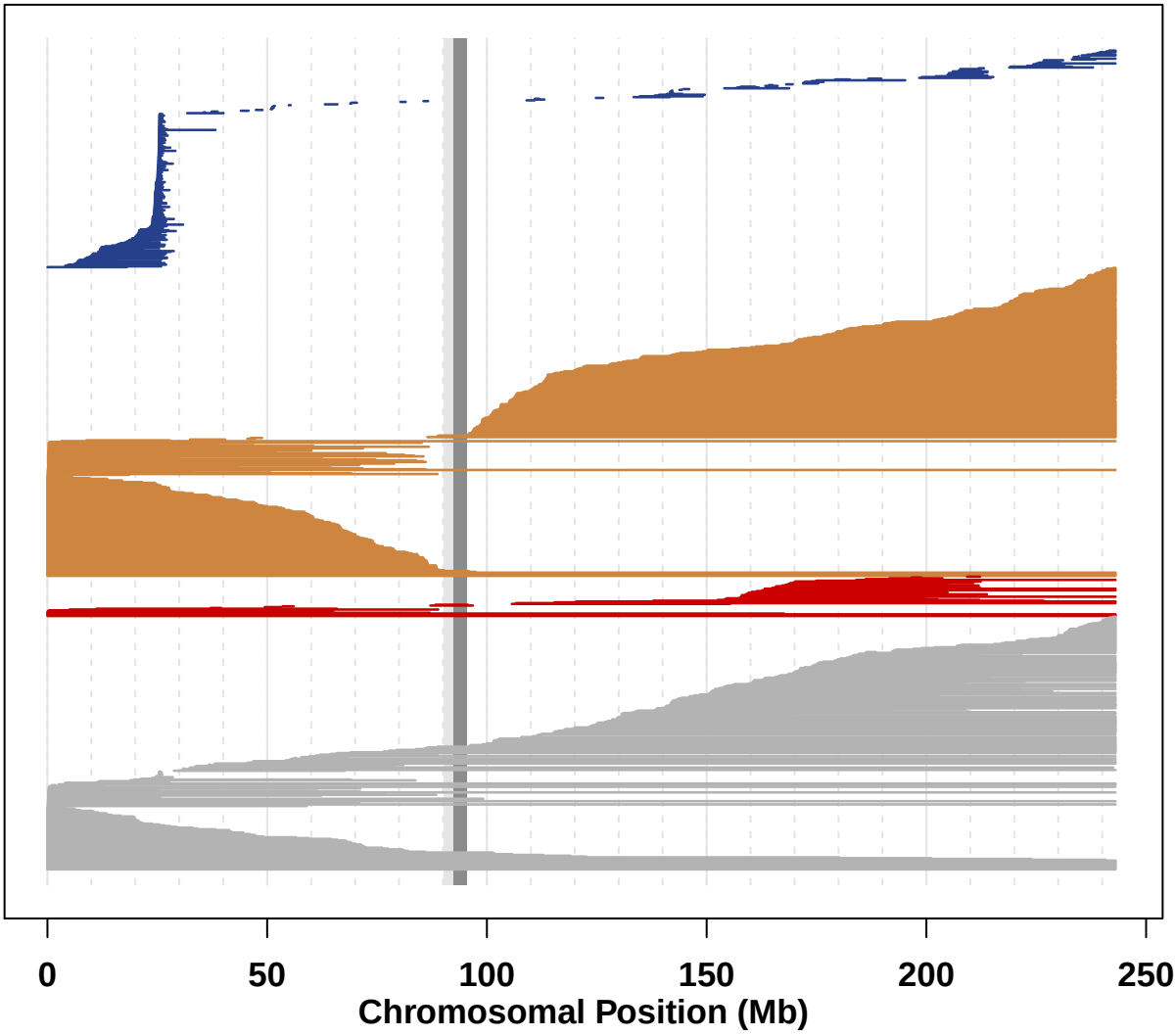

Fig. S2.1.2 A landscape of mosaic events in chromosome 2.

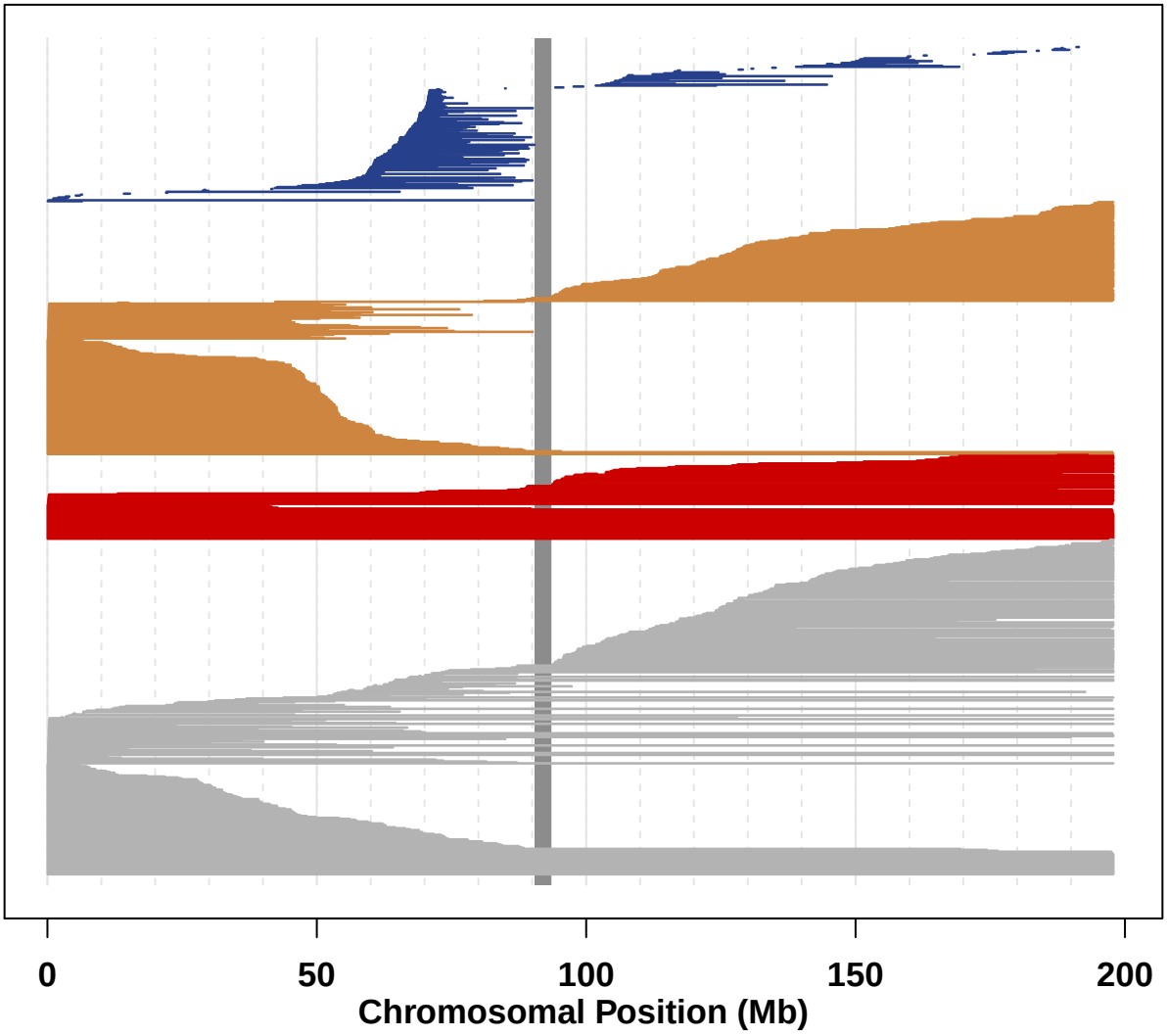

Fig. S2.1.3 A landscape of mosaic events in chromosome 3.

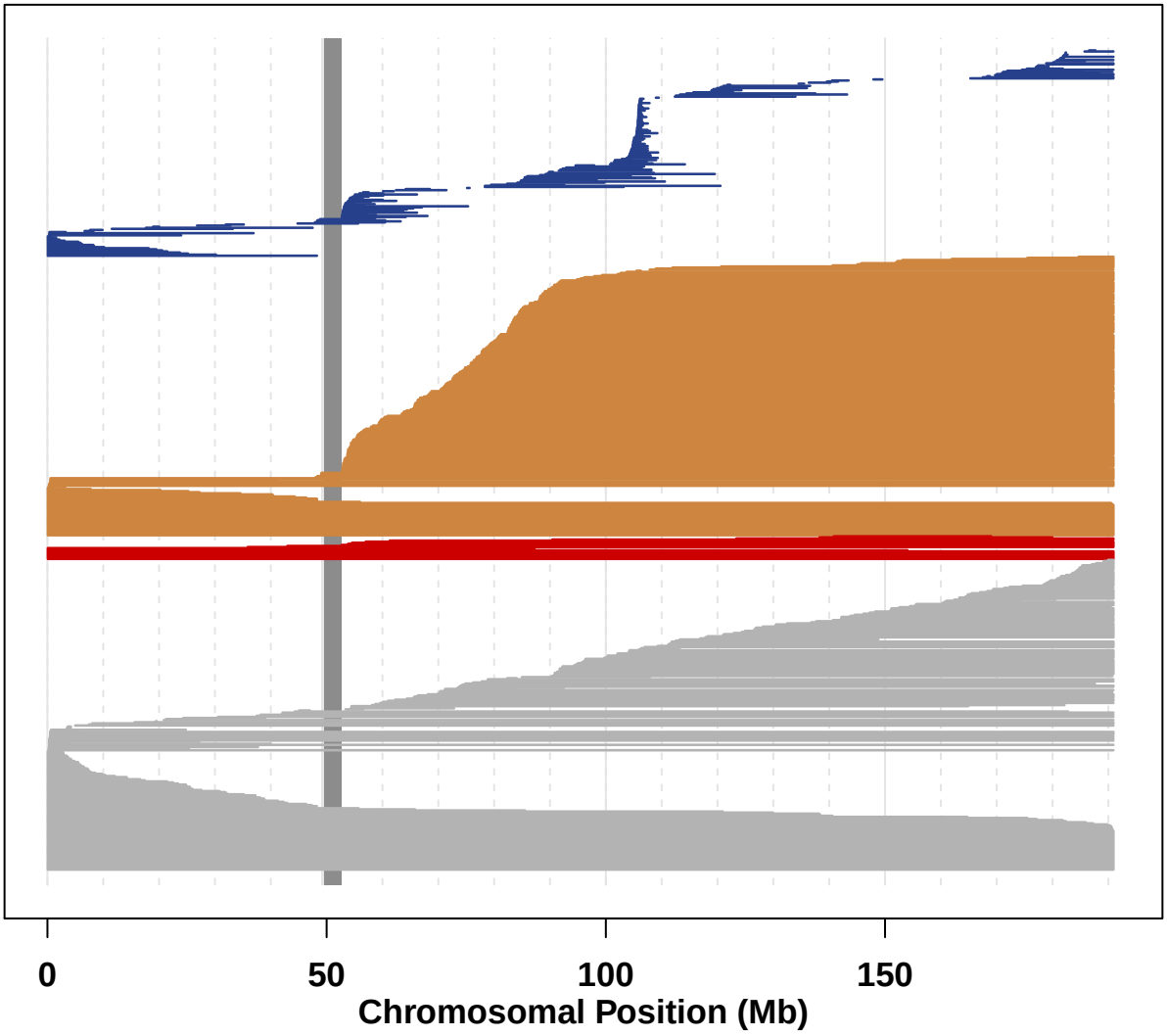

Fig. S2.1.4 A landscape of mosaic events in chromosome 4.

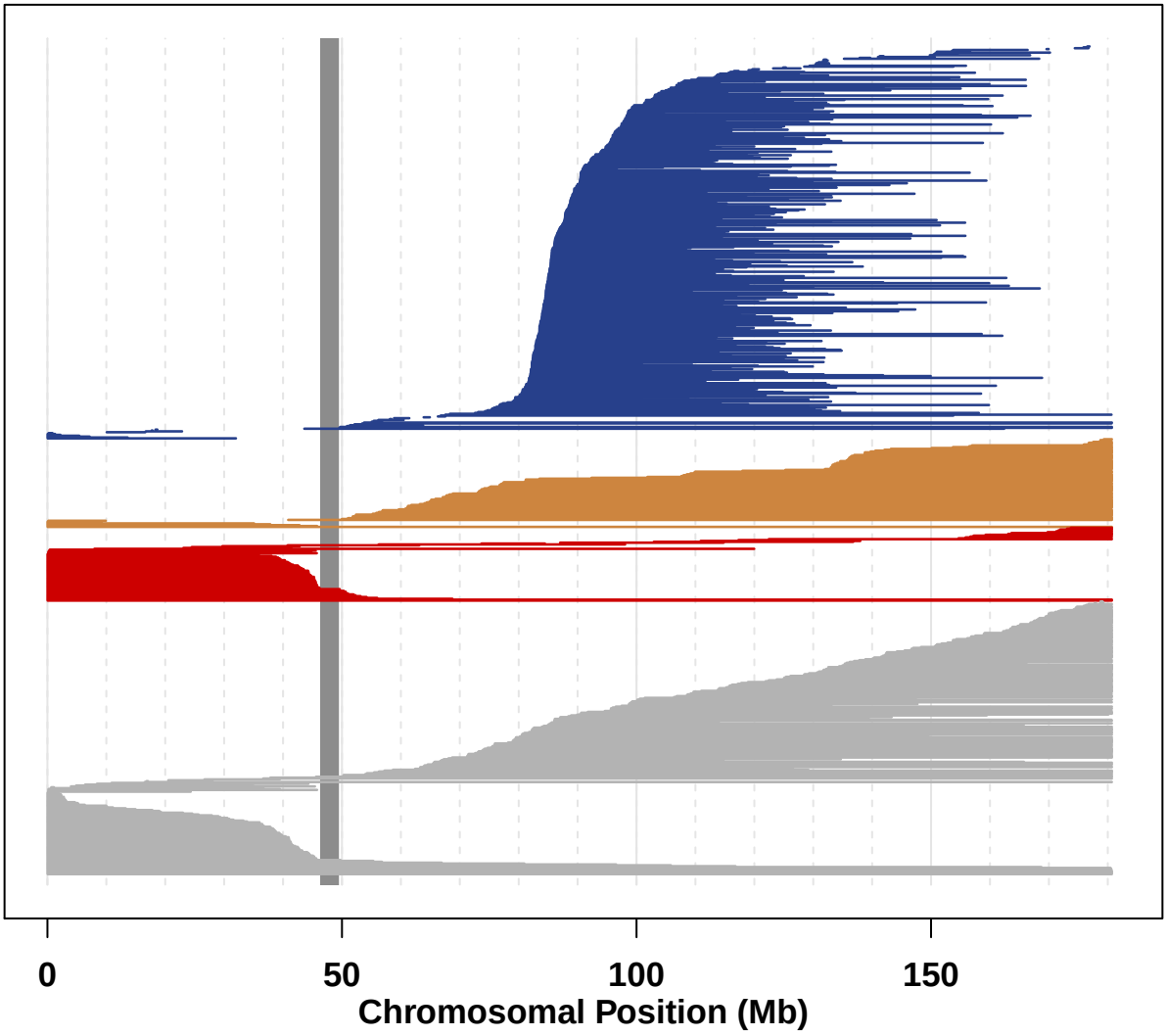

Fig. S2.1.5 A landscape of mosaic events in chromosome 5.

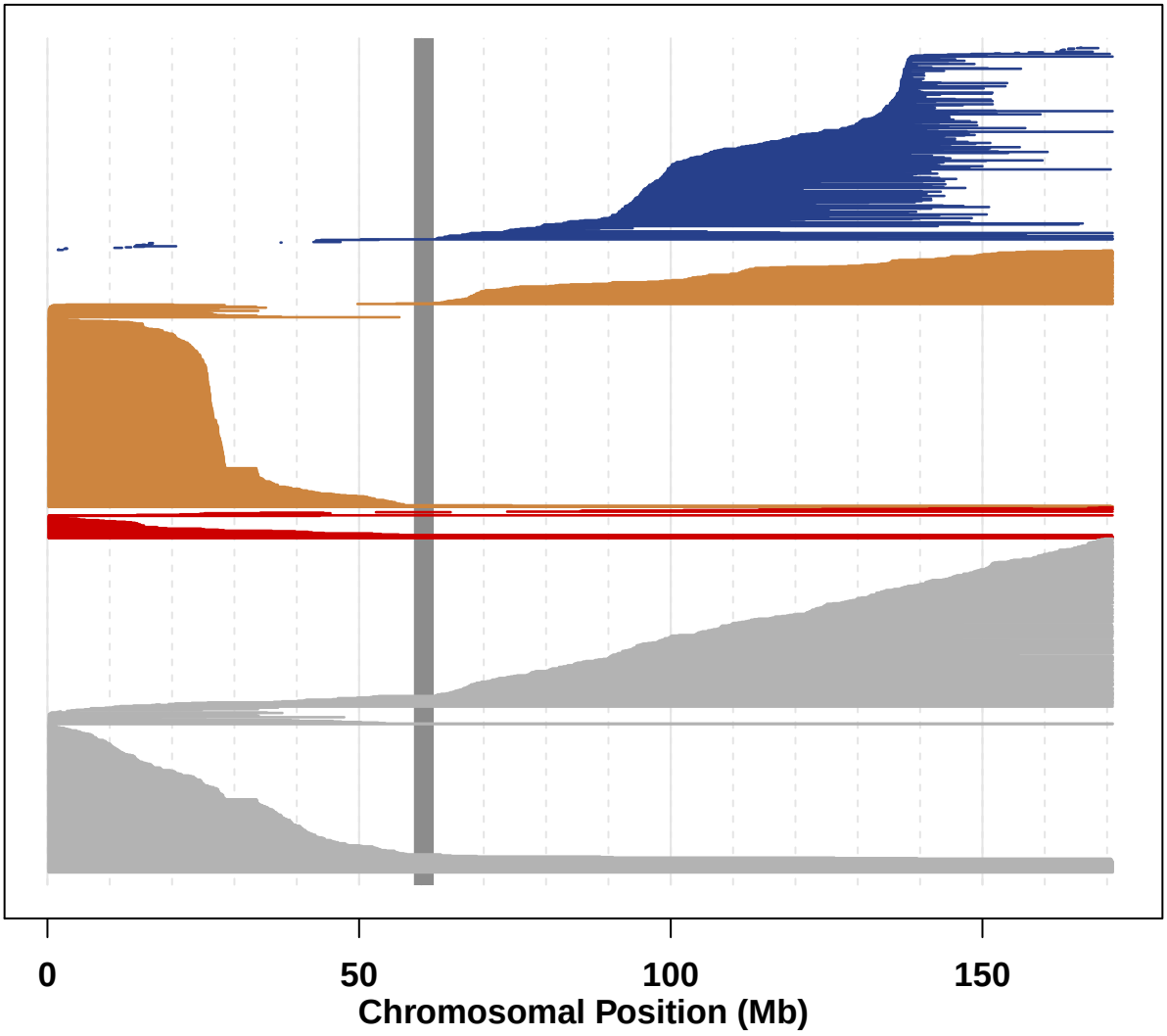

Fig. S2.1.6 A landscape of mosaic events in chromosome 6.

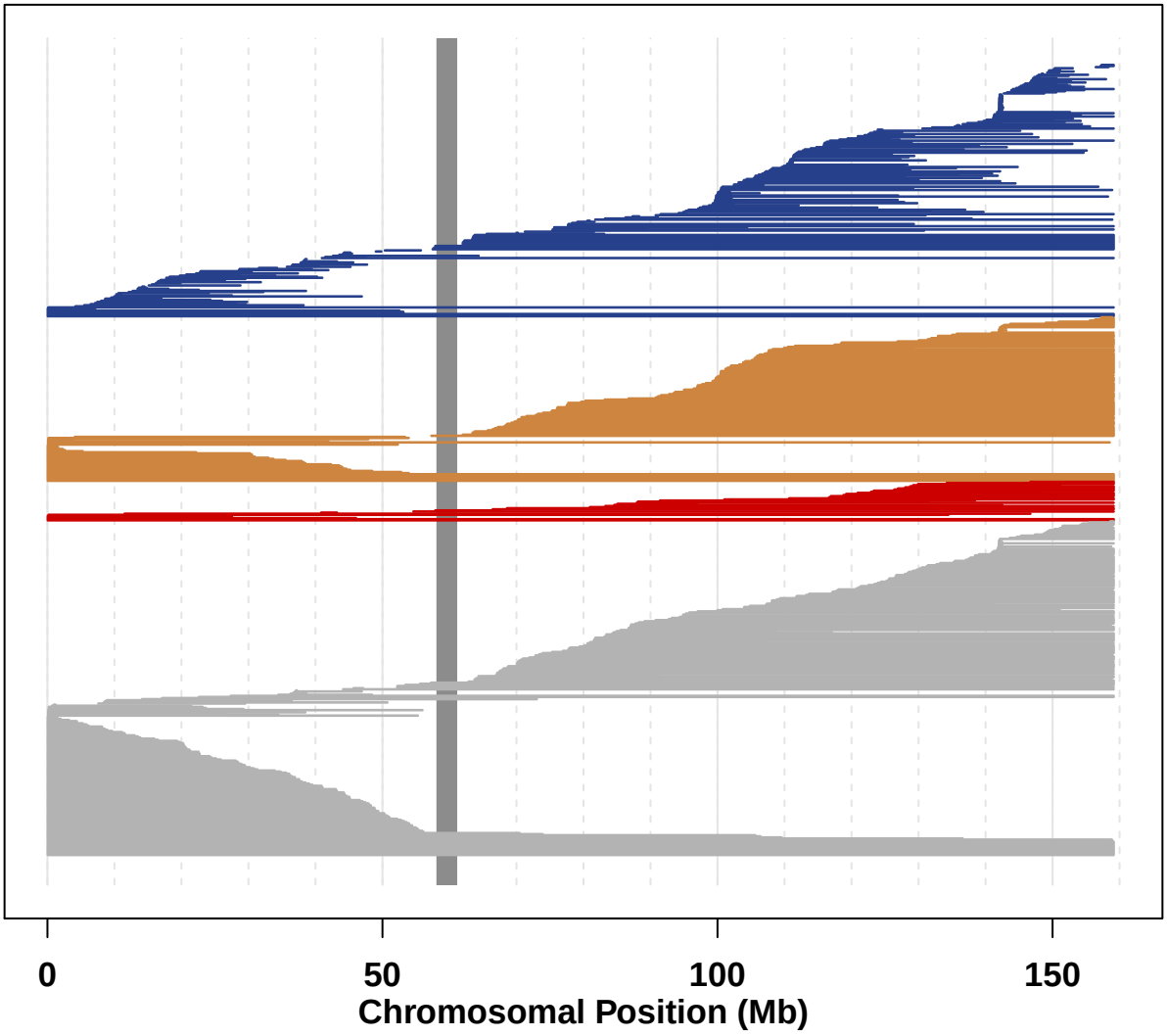

Fig. S2.1.7 A landscape of mosaic events in chromosome 7.

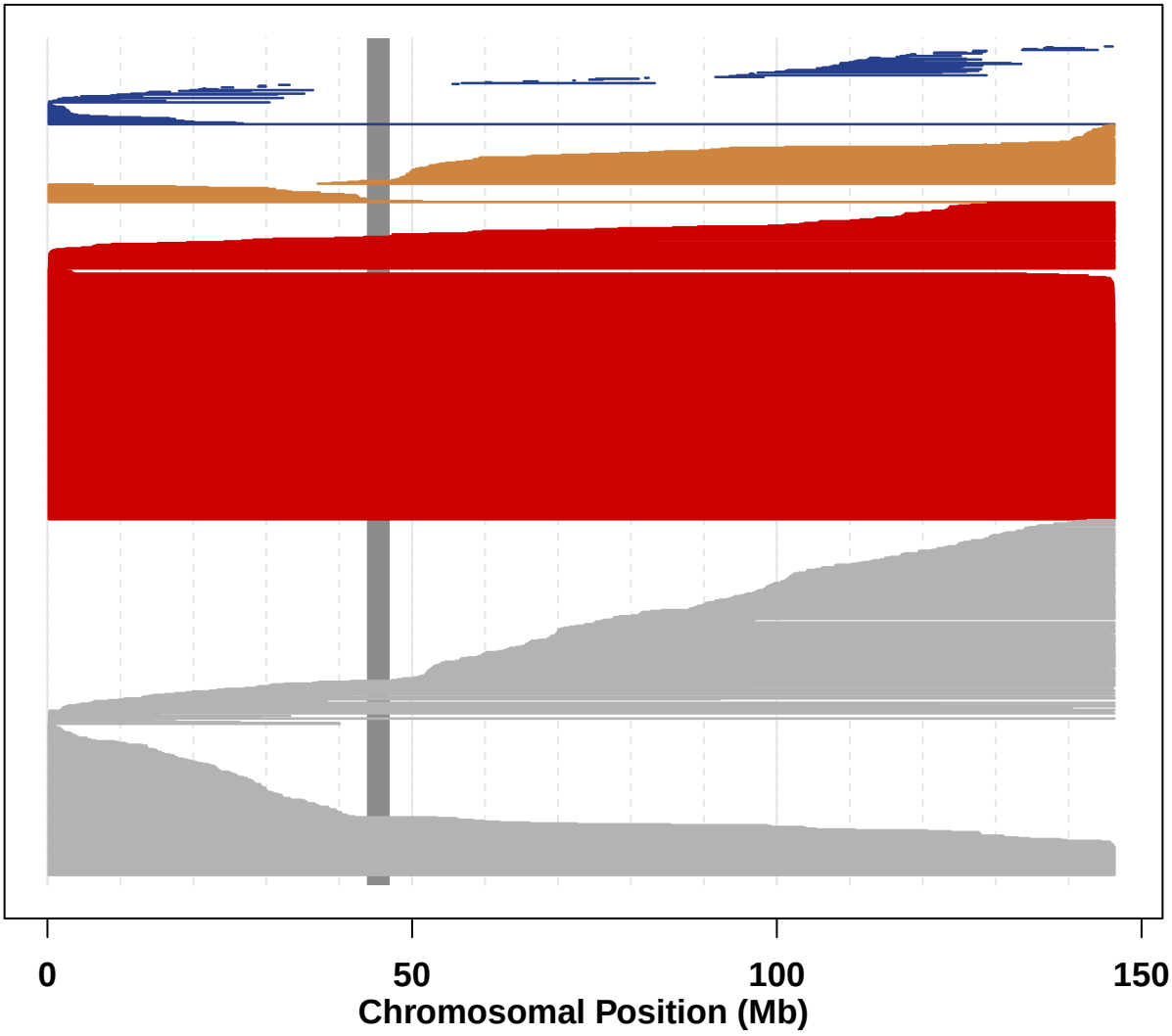

Fig. S2.1.8 A landscape of mosaic events in chromosome 8.

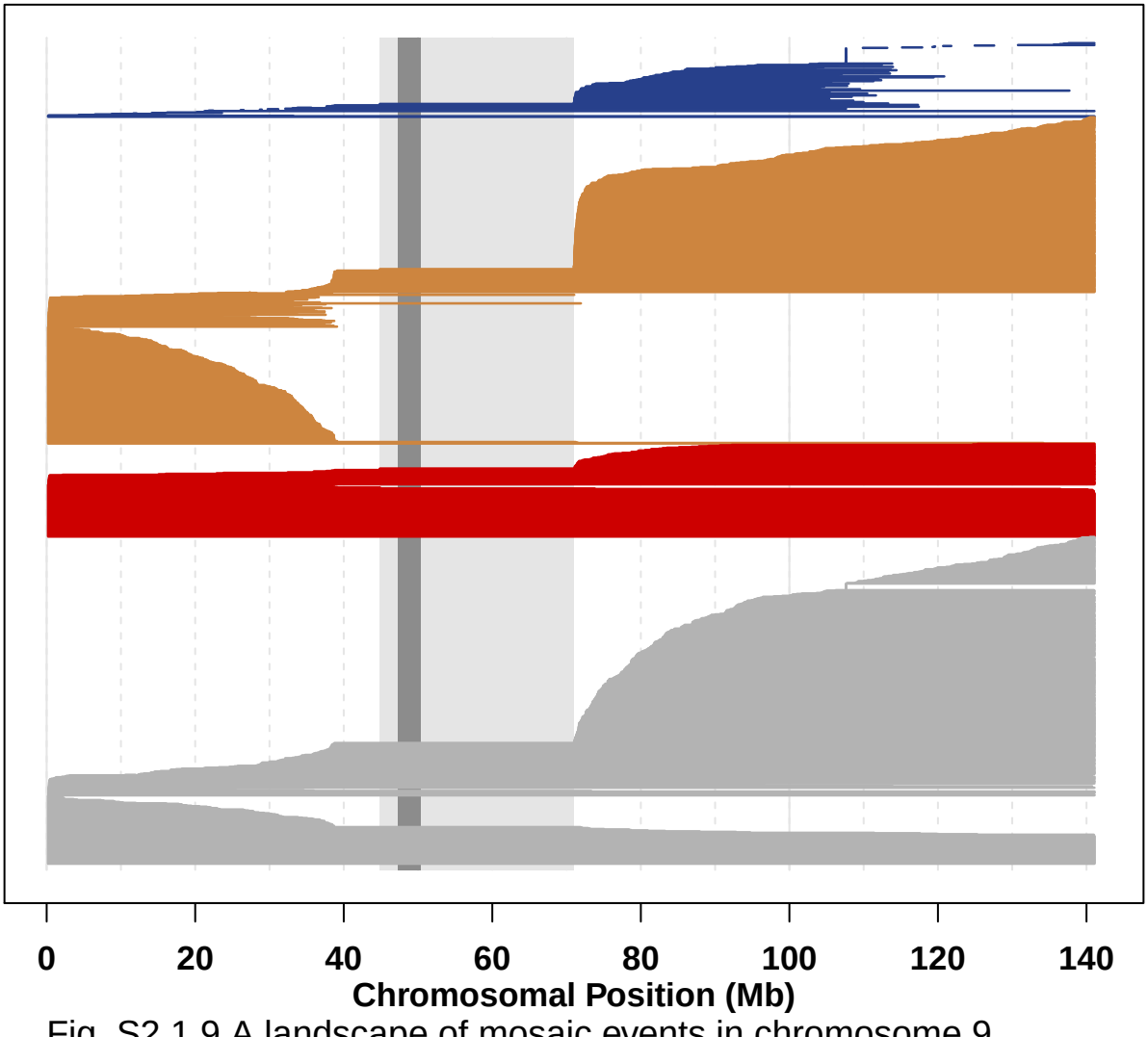

Fig. S2.1.9 A landscape of mosaic events in chromosome 9.

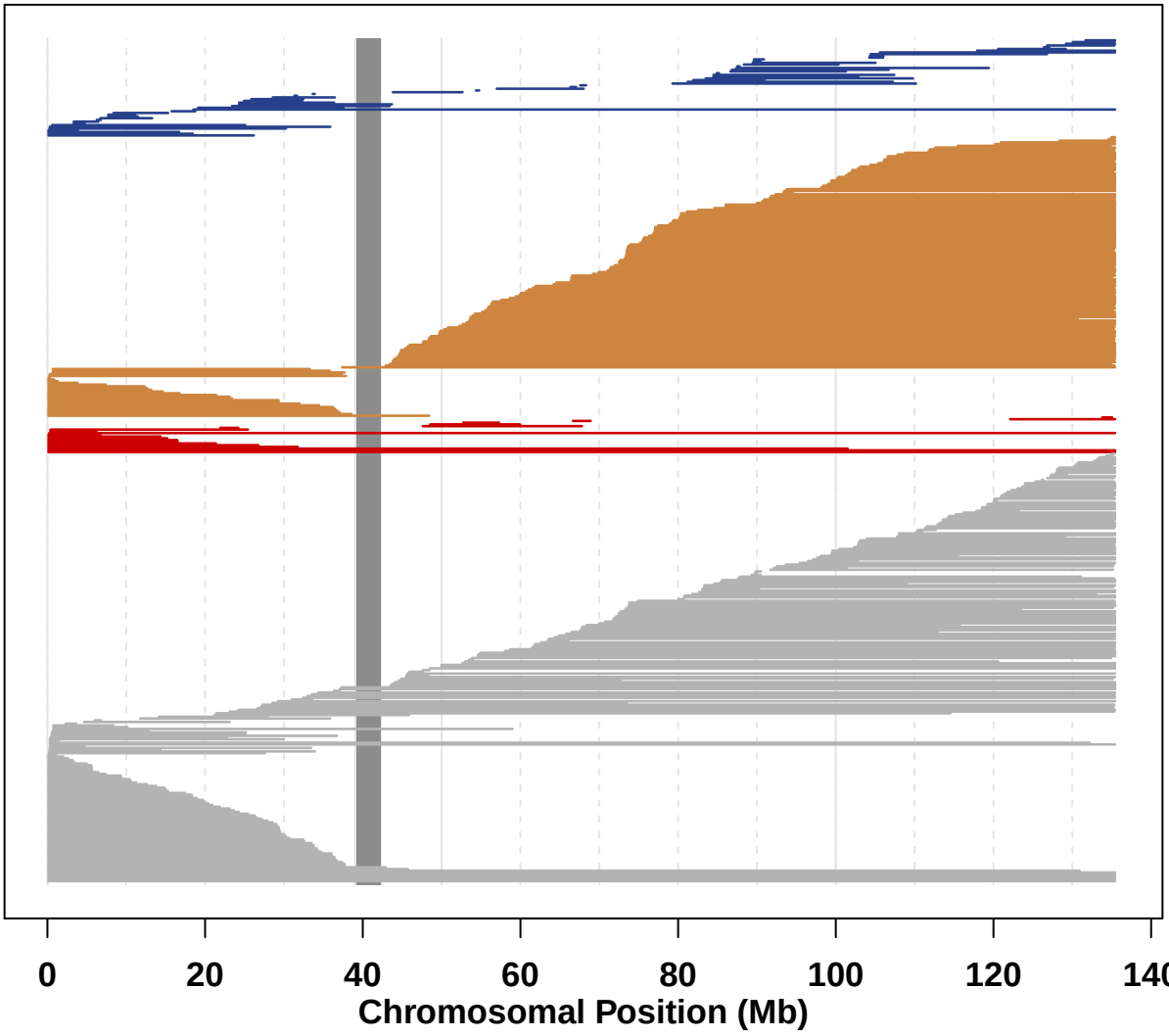

Fig. S2.1.10 A landscape of mosaic events in chromosome 10.

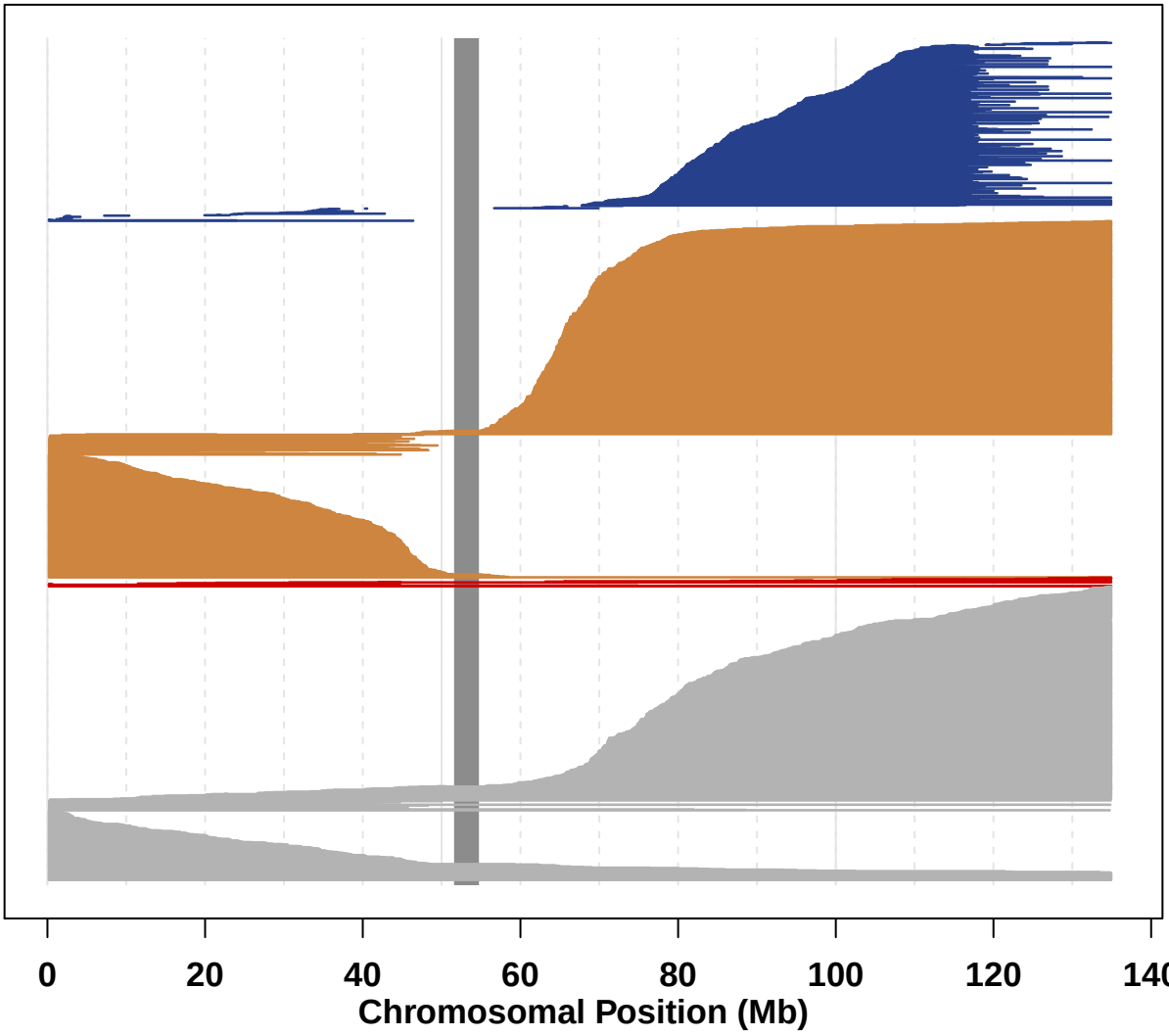

Fig. S2.1.11 A landscape of mosaic events in chromosome 11.

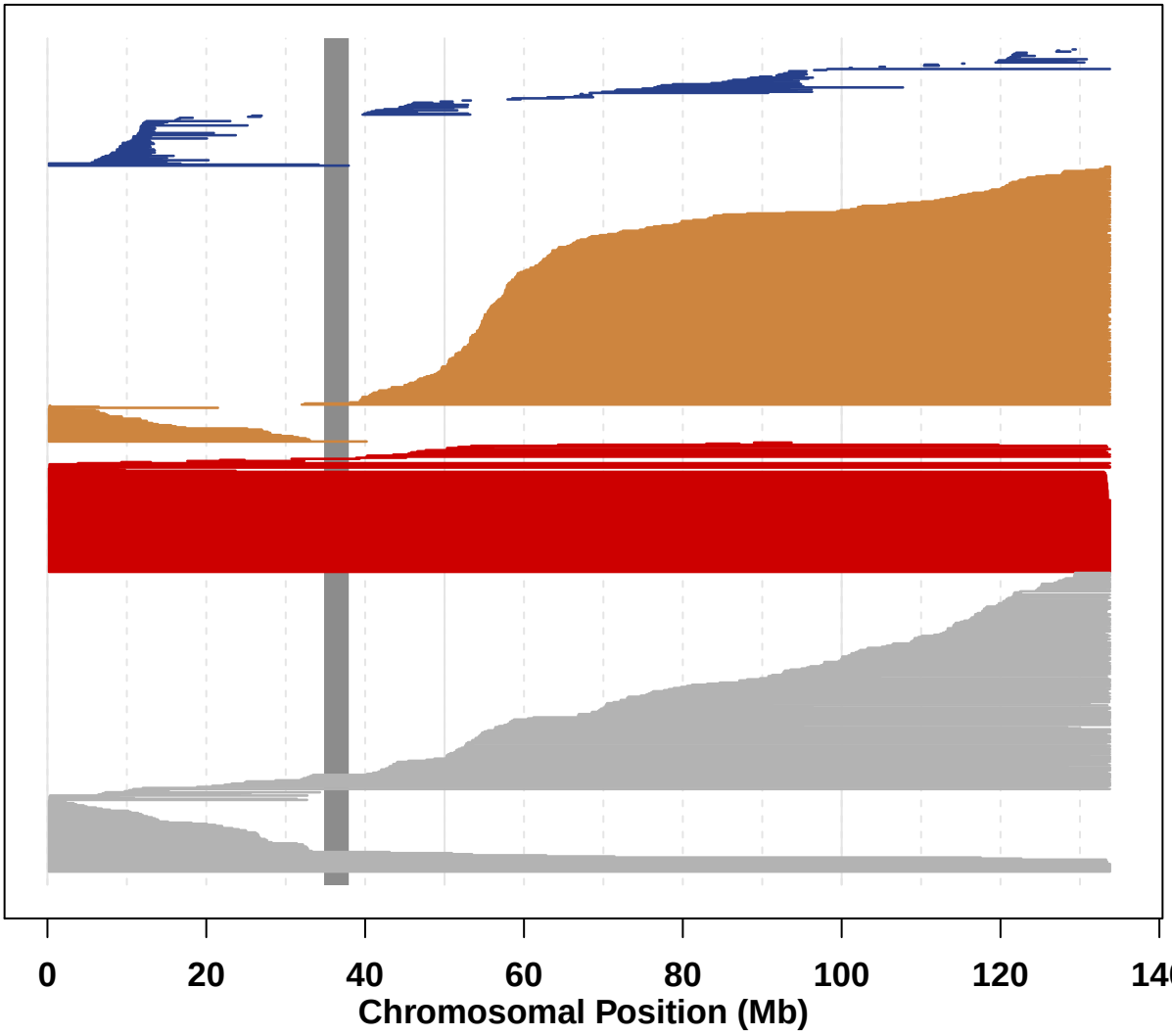

Fig. S2.1.12 A landscape of mosaic events in chromosome 12.

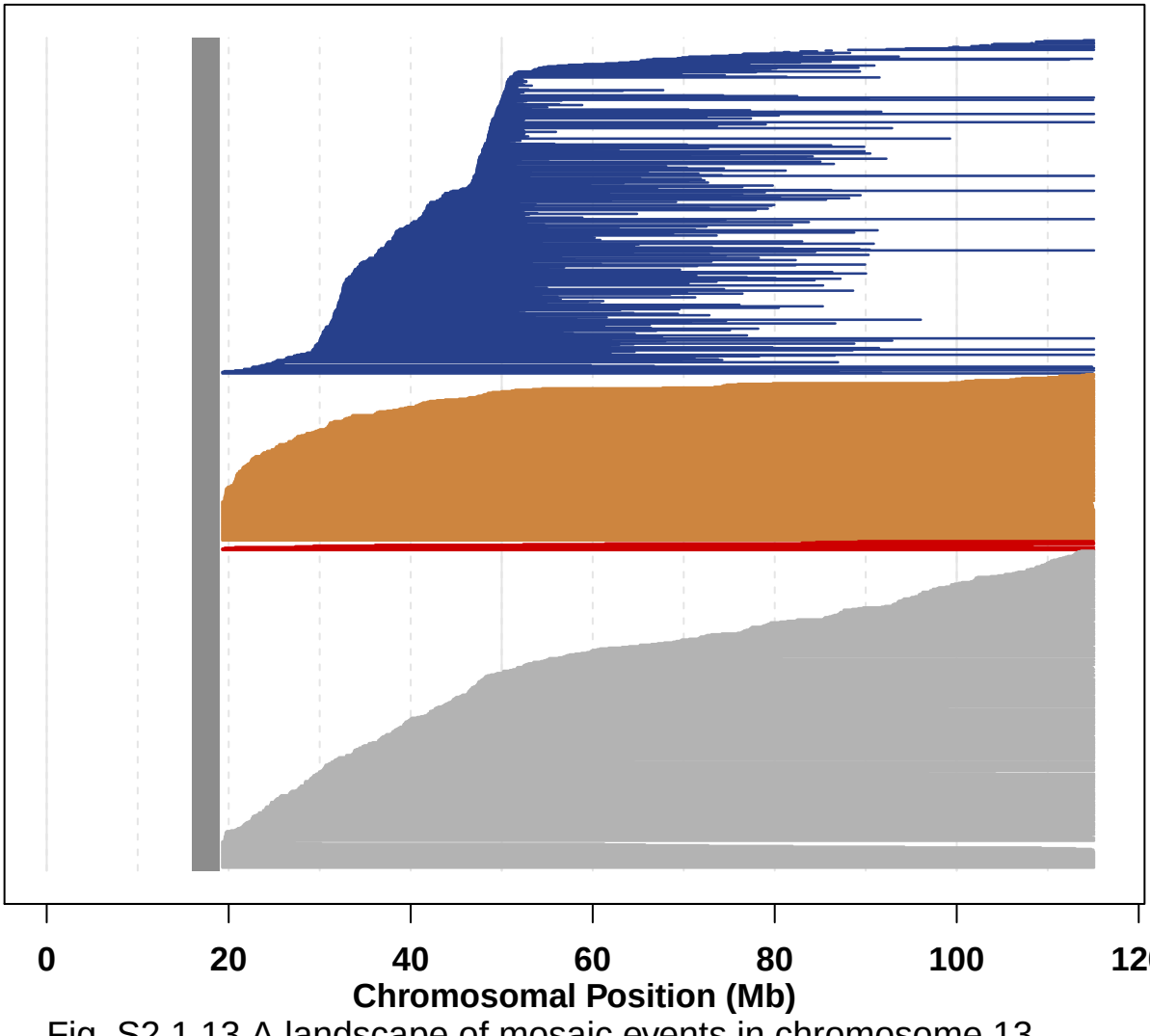

Fig. S2.1.13 A landscape of mosaic events in chromosome 13.

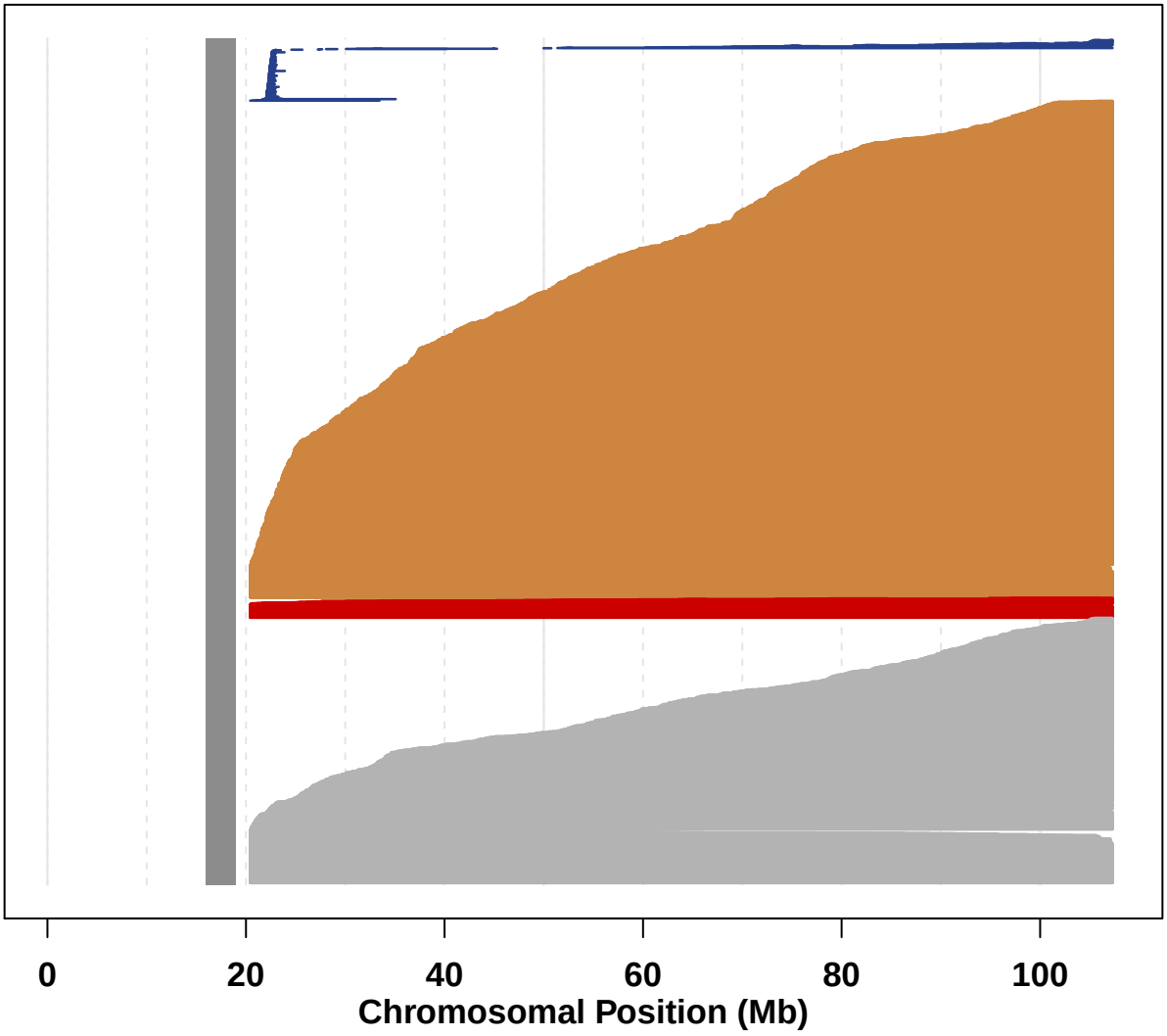

Fig. S2.1.14 A landscape of mosaic events in chromosome 14.

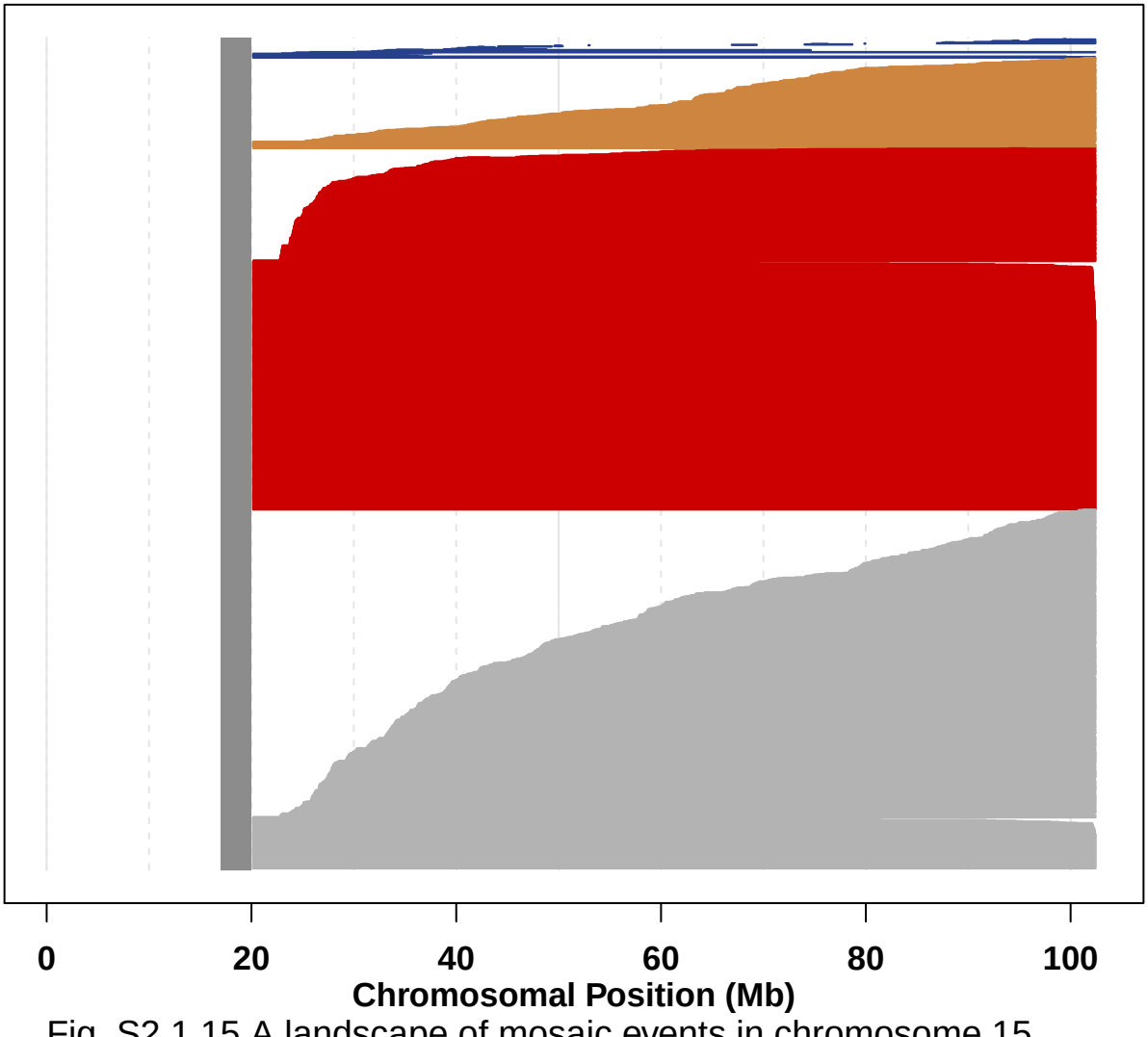

Fig. S2.1.15 A landscape of mosaic events in chromosome 15.

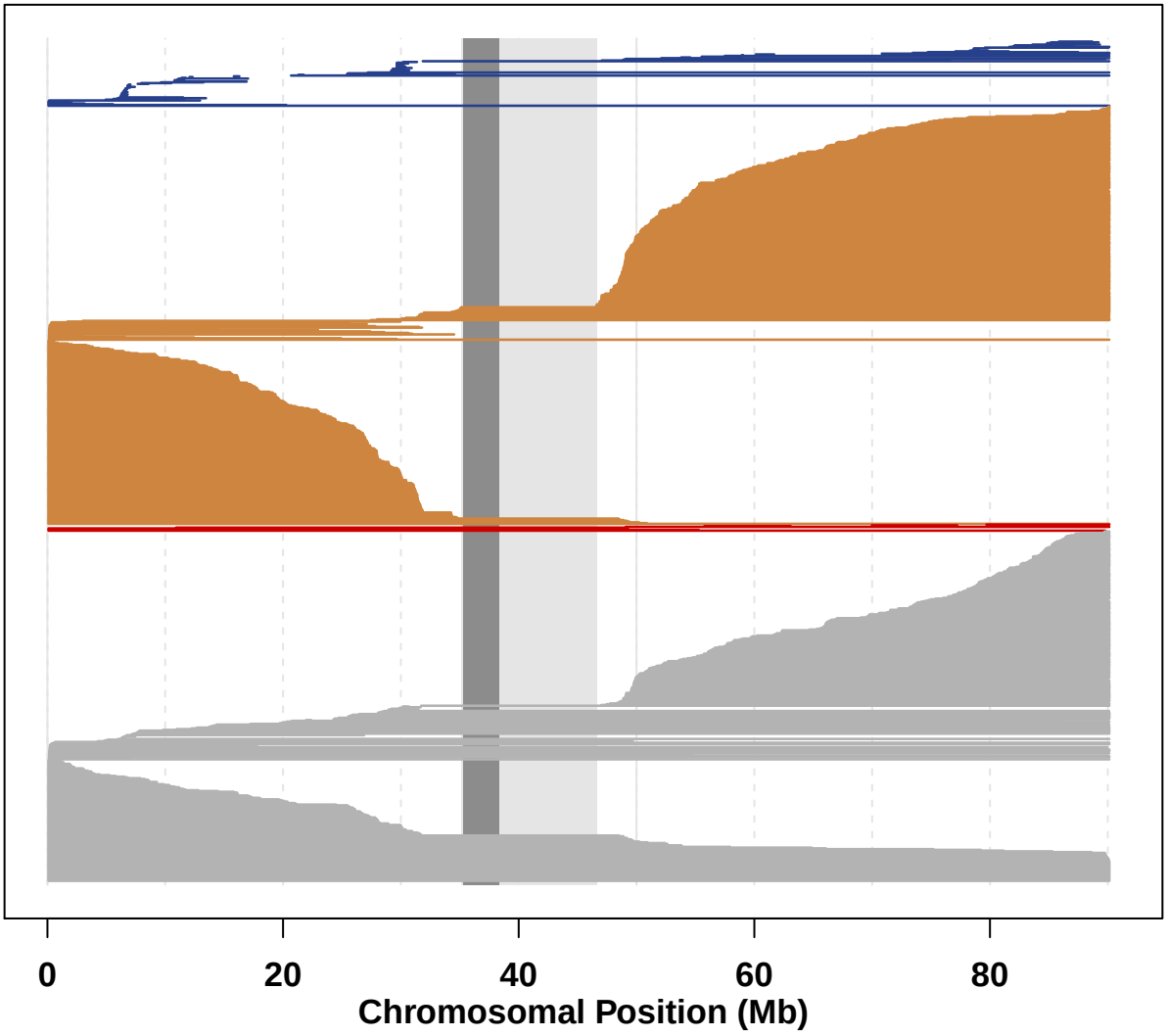

Fig. S2.1.16 A landscape of mosaic events in chromosome 16.

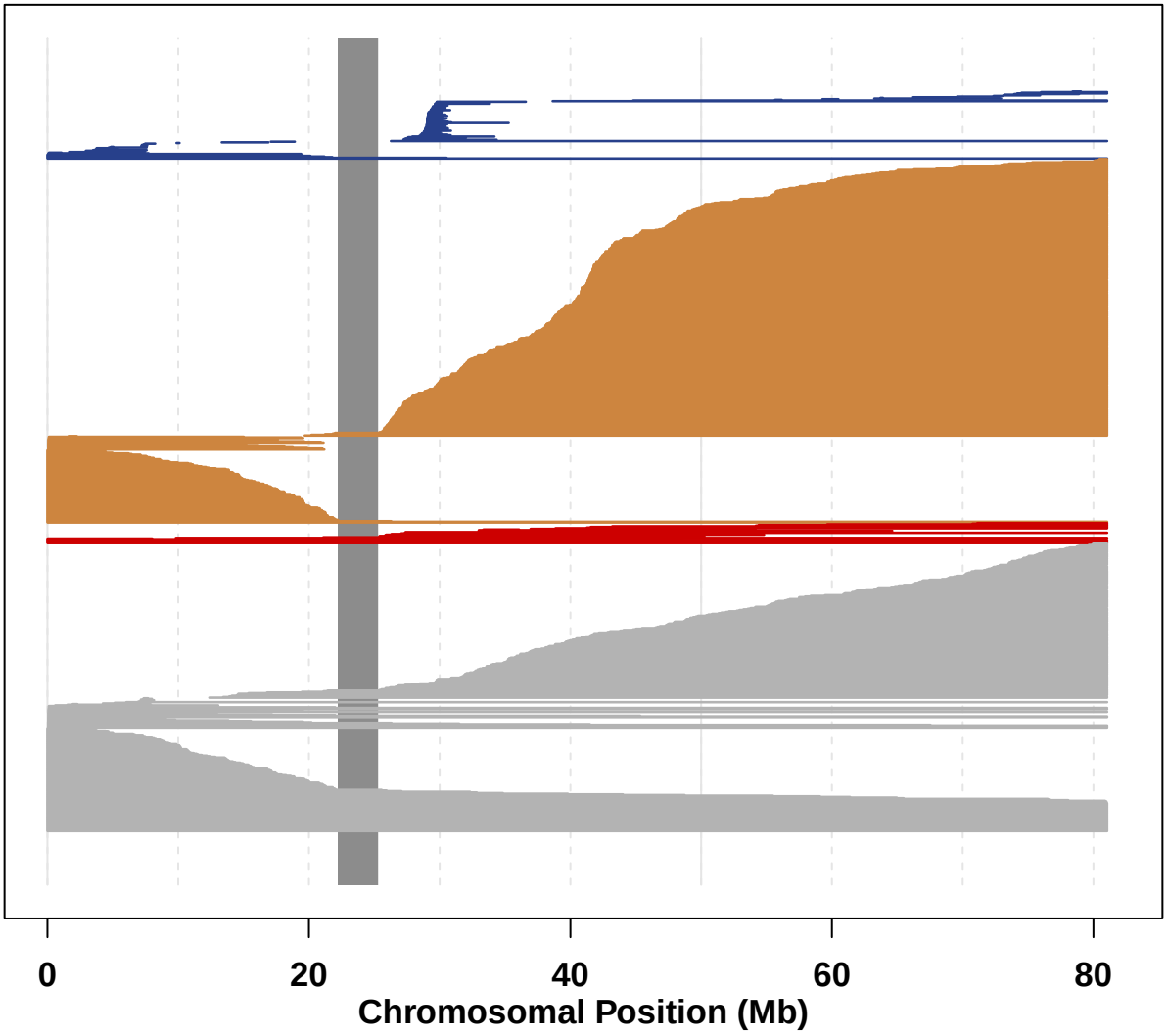

Fig. S2.1.17 A landscape of mosaic events in chromosome 17.

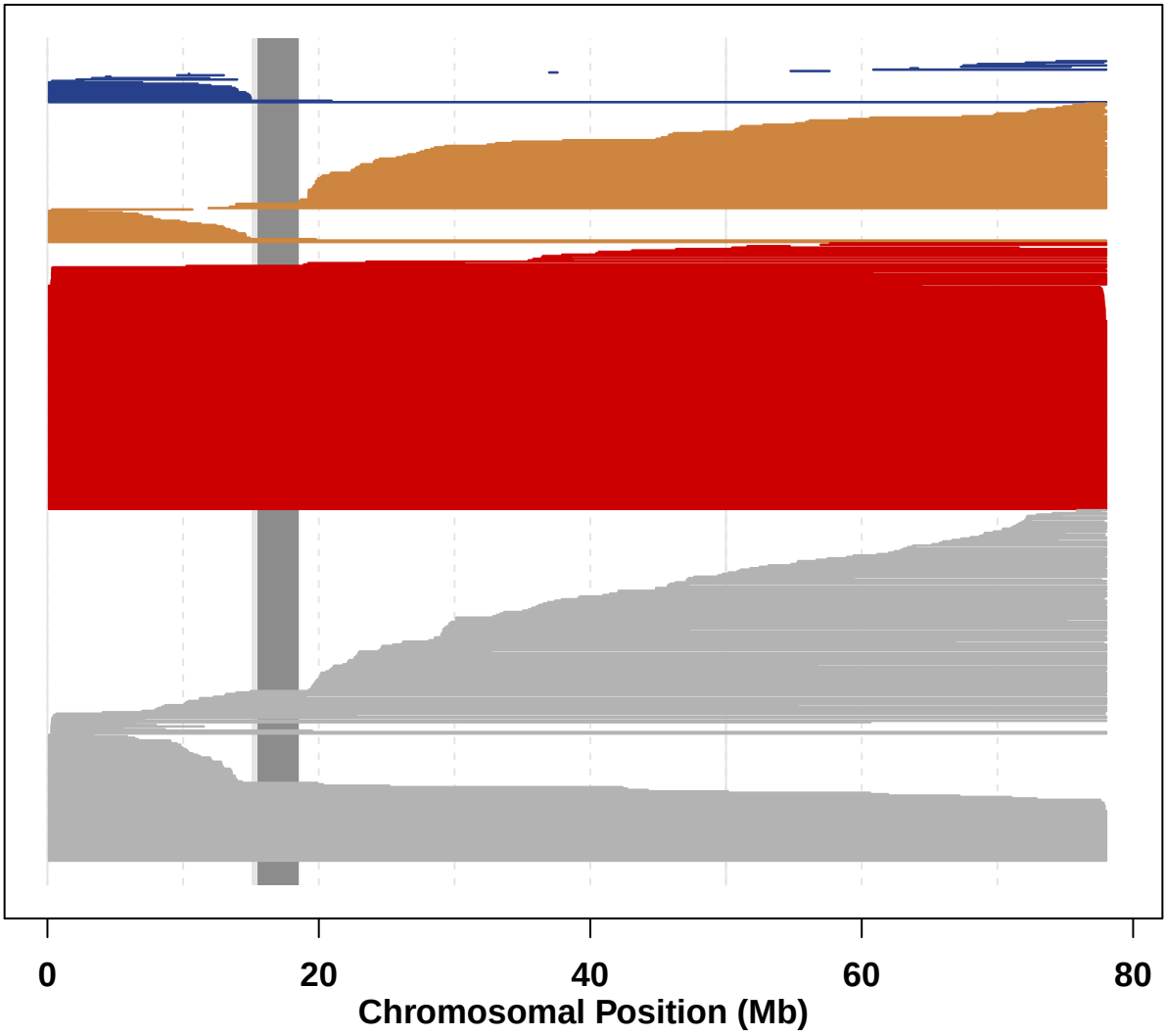

Fig. S2.1.18 A landscape of mosaic events in chromosome 18.

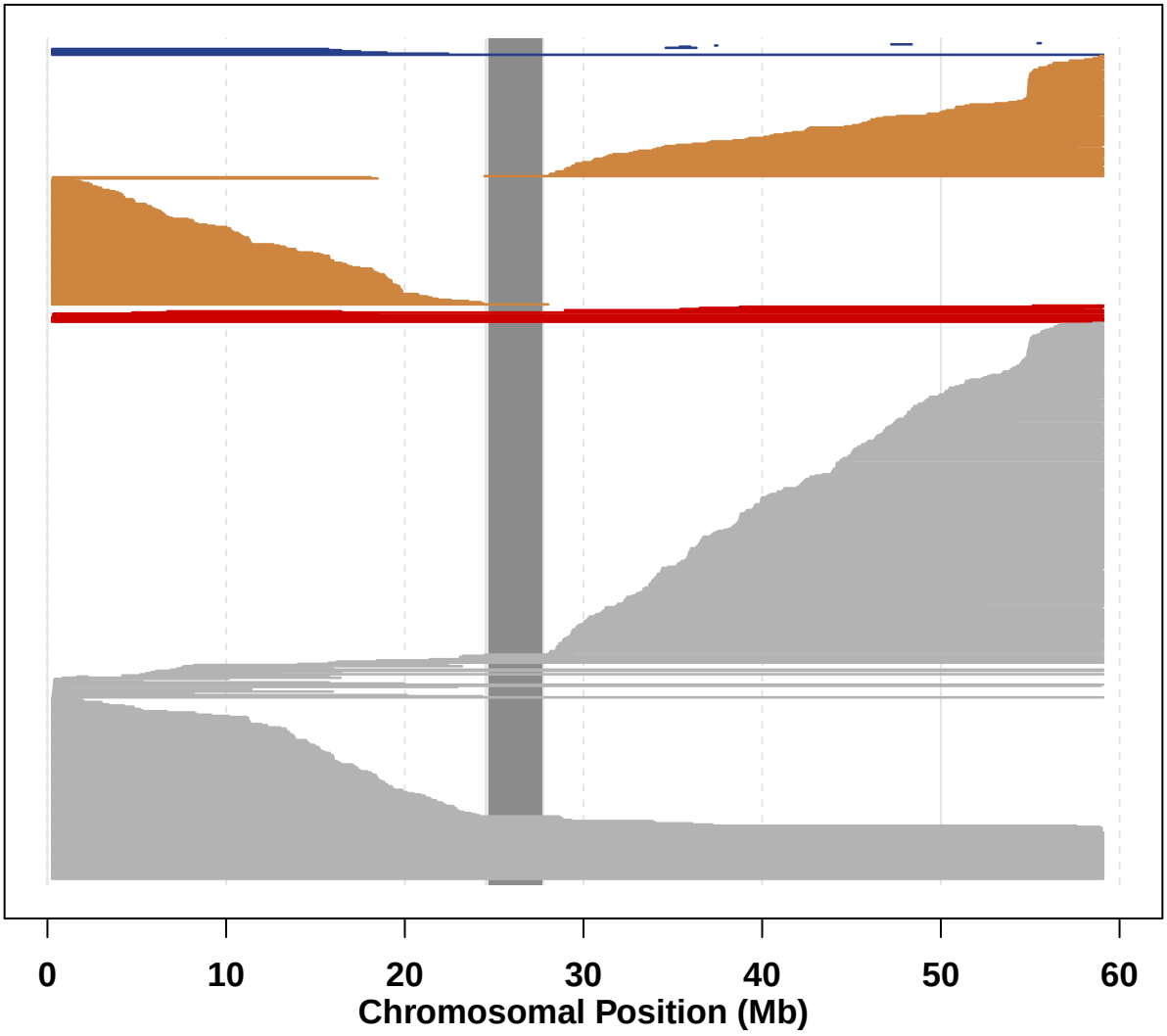

Fig. S2.1.19 A landscape of mosaic events in chromosome 19.

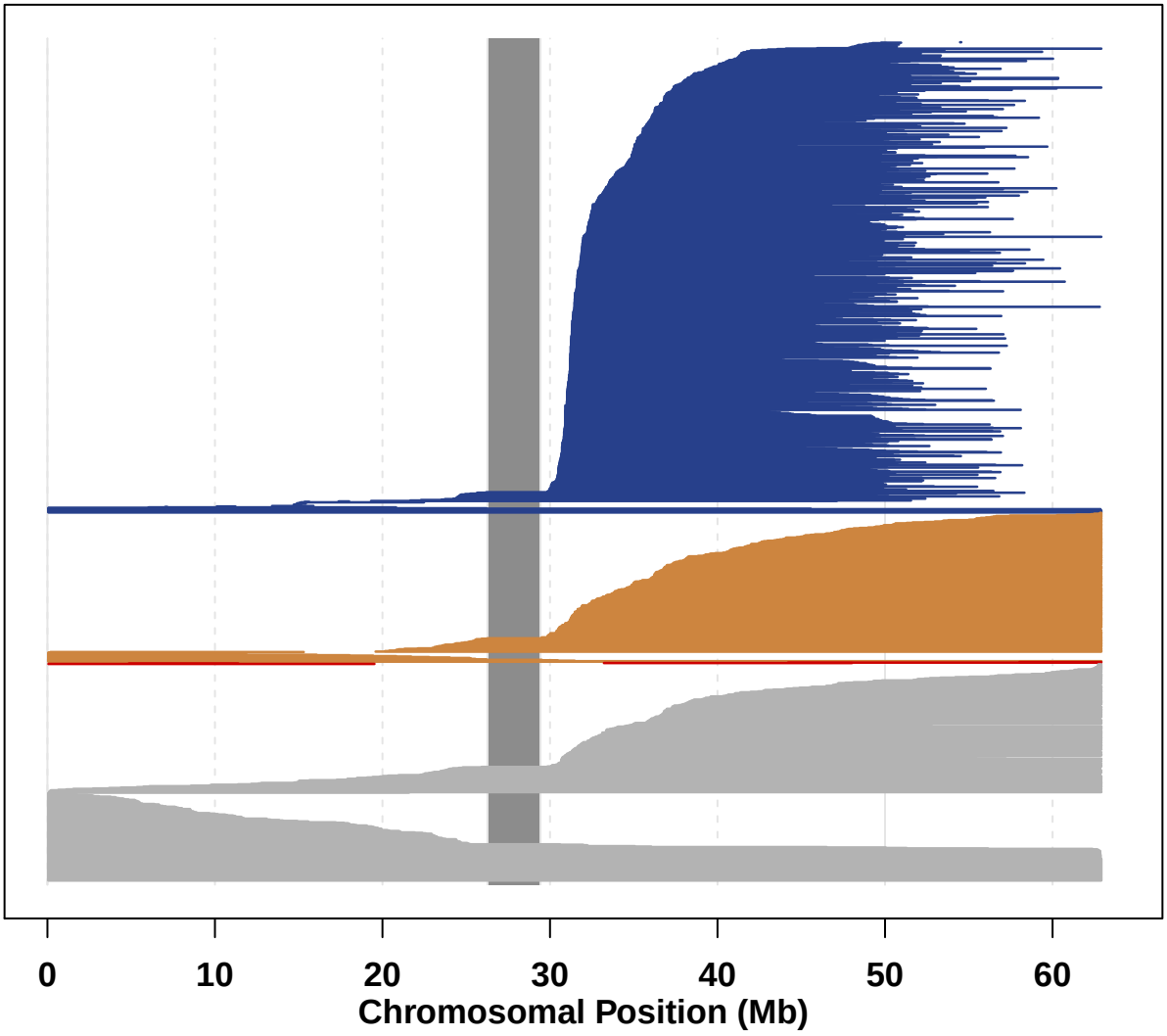

Fig. S2.1.20 A landscape of mosaic events in chromosome 20.

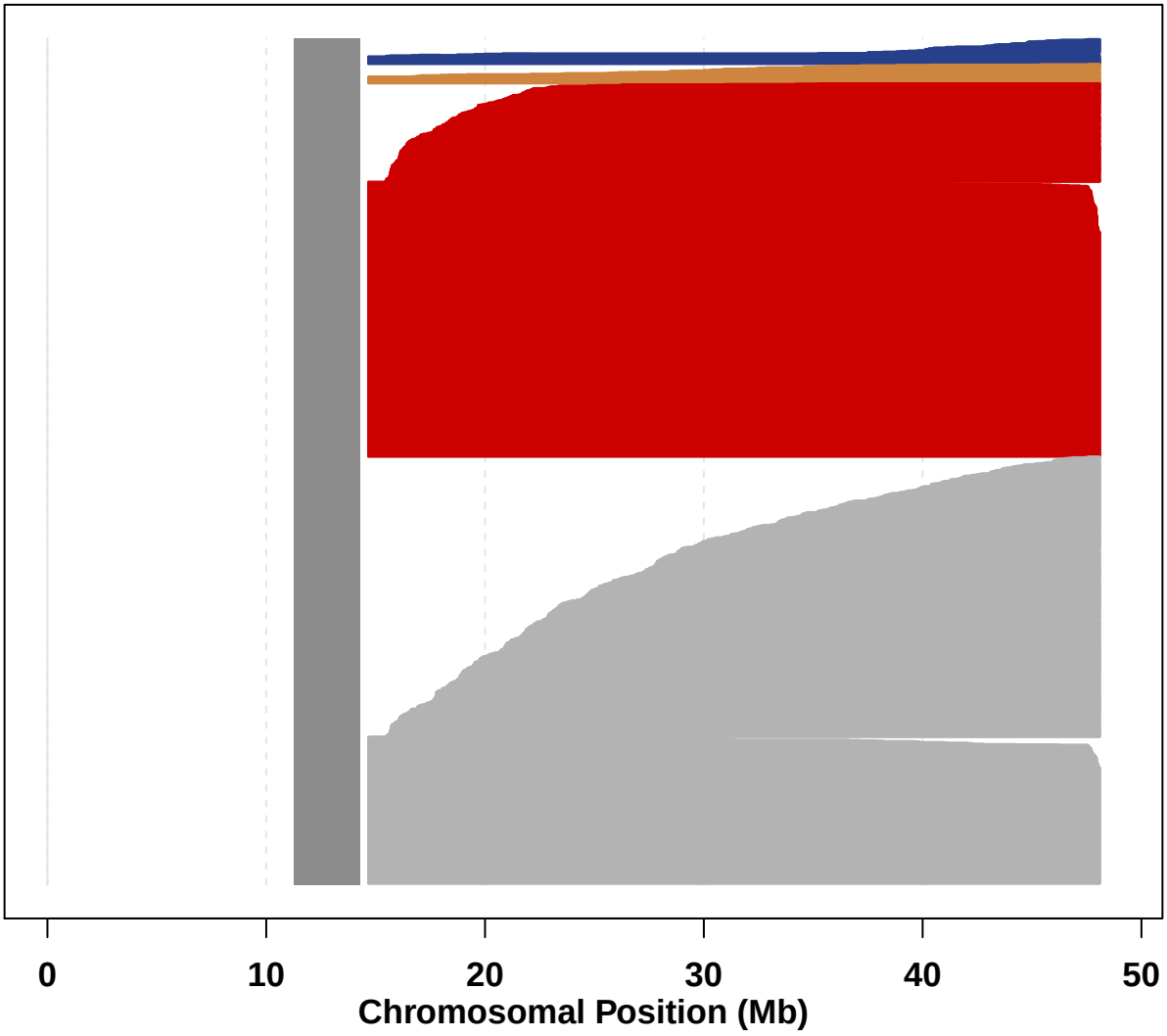

Fig. S2.1.21 A landscape of mosaic events in chromosome 21.

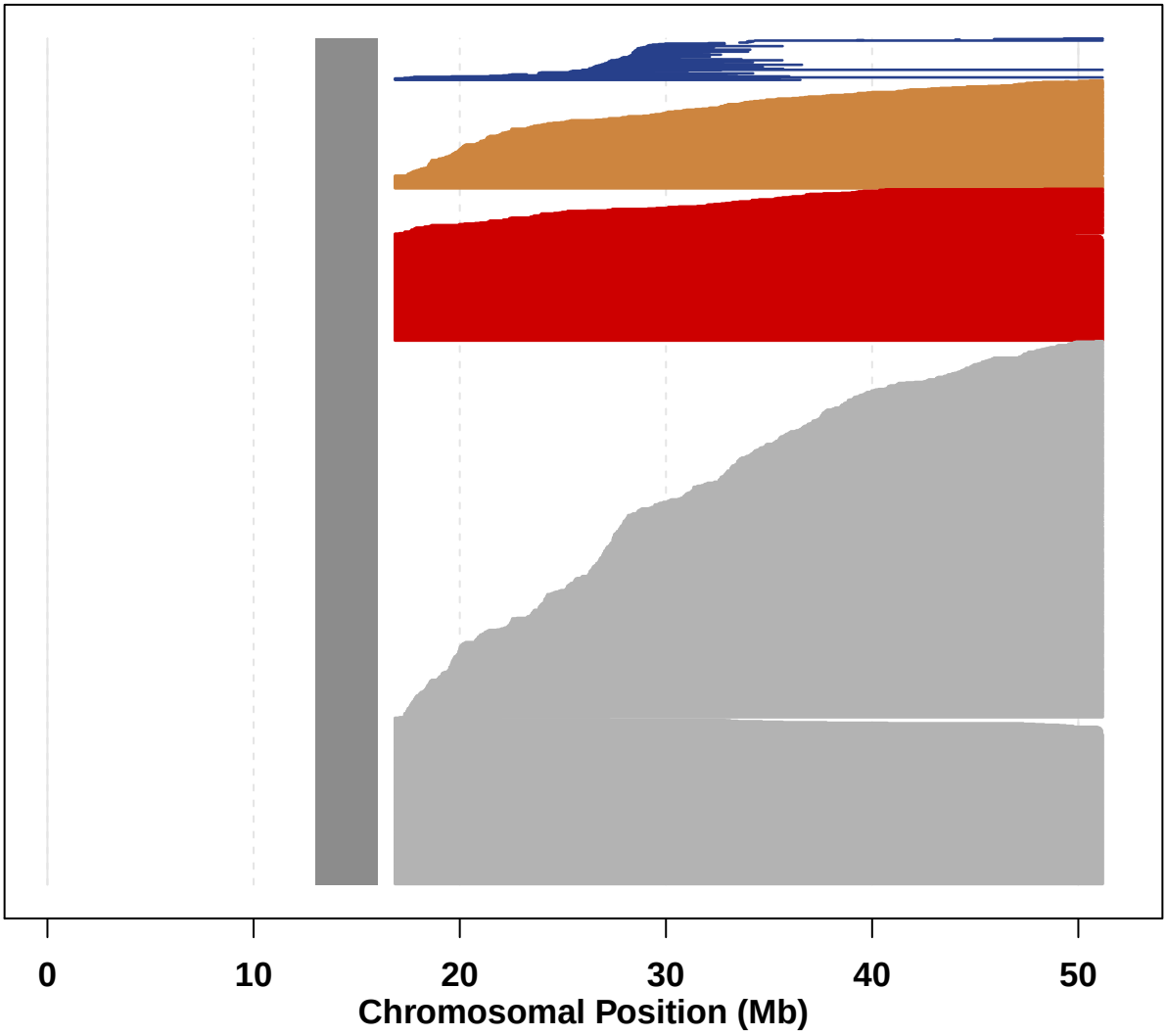

Fig. S2.1.22 A landscape of mosaic events in chromosome 22.

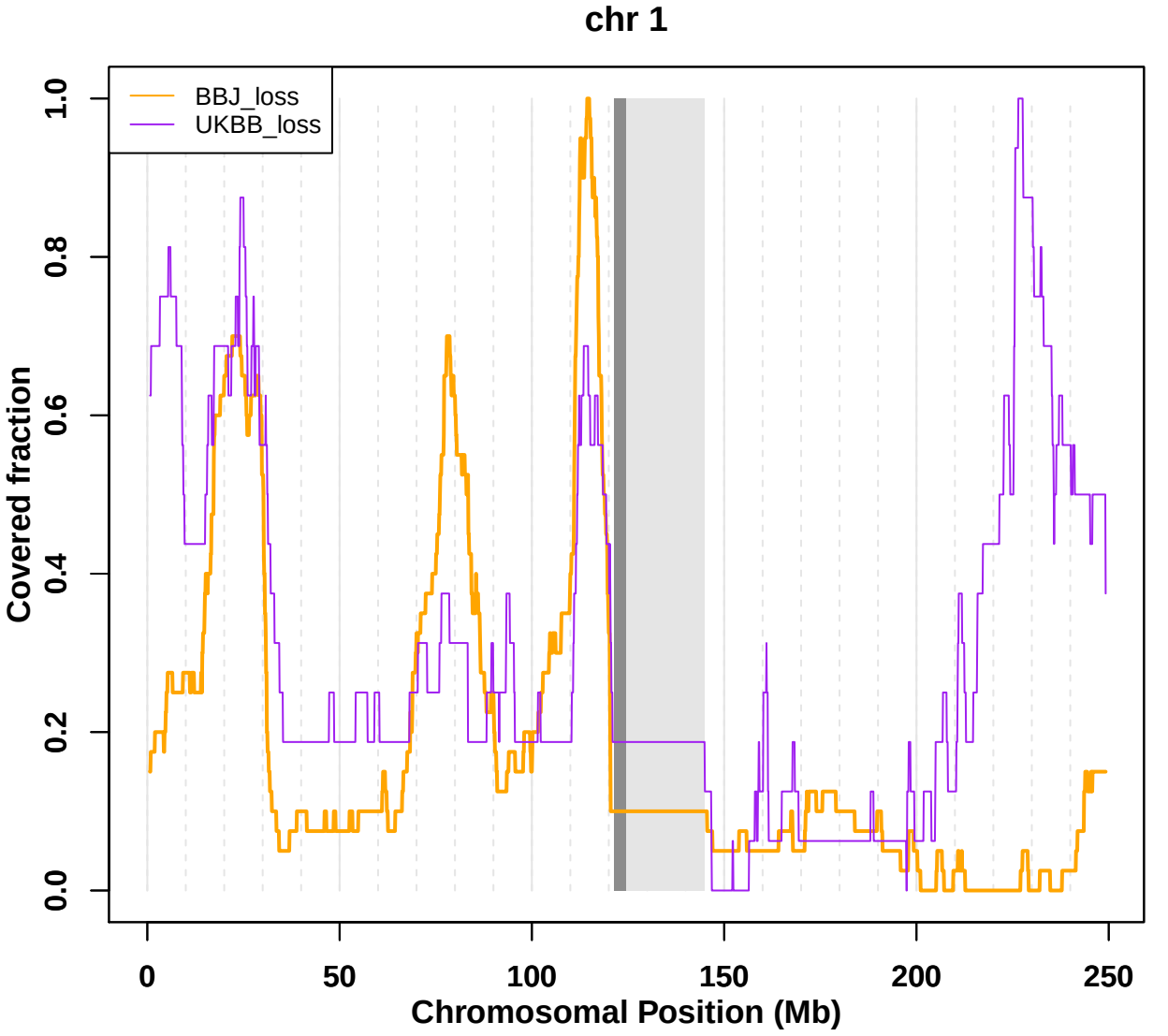

Fig. S2.2.1 Coverage of mosaic loss in chromosome 1.

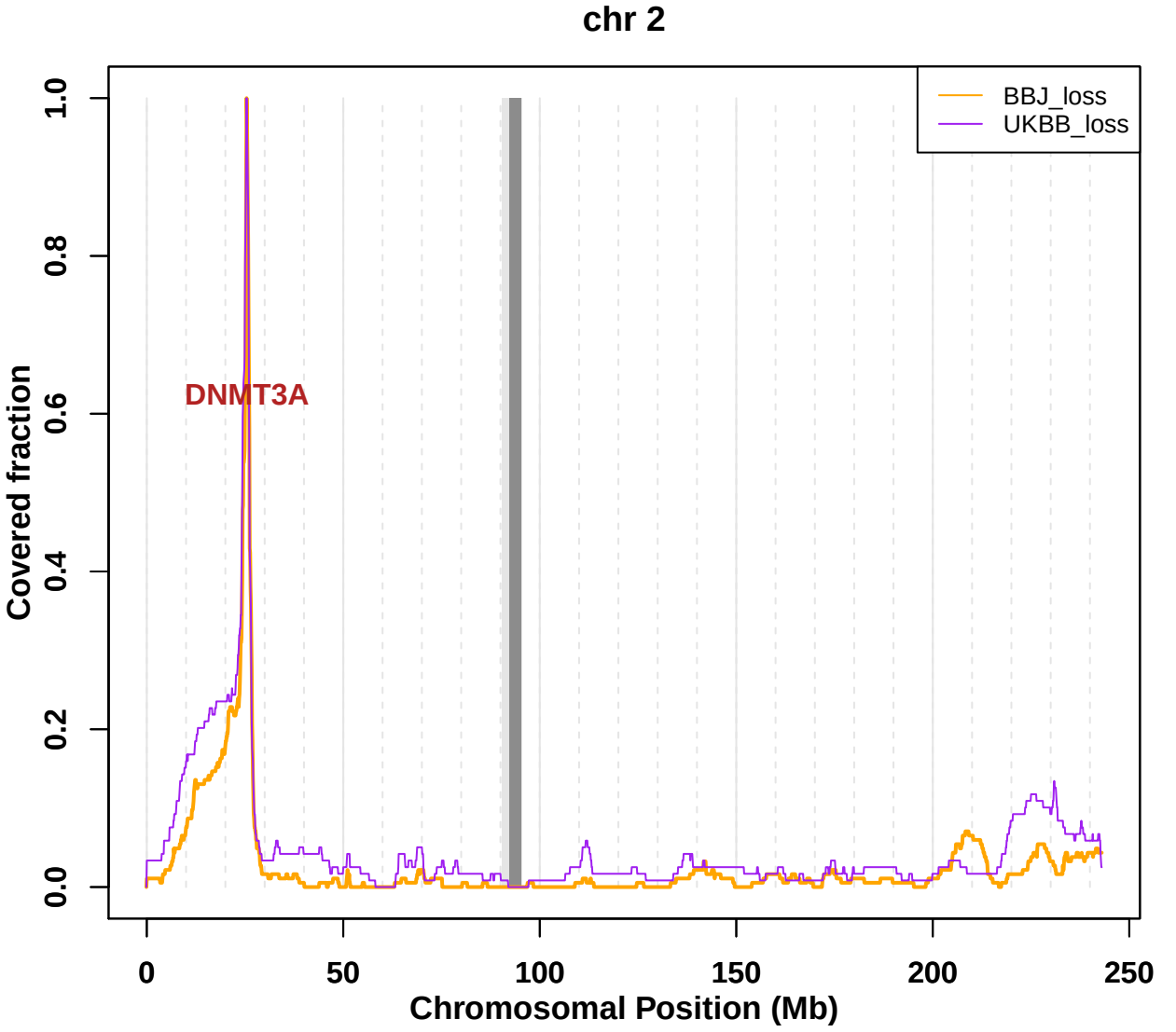

Fig. S2.2.2 Coverage of mosaic loss in chromosome 2.

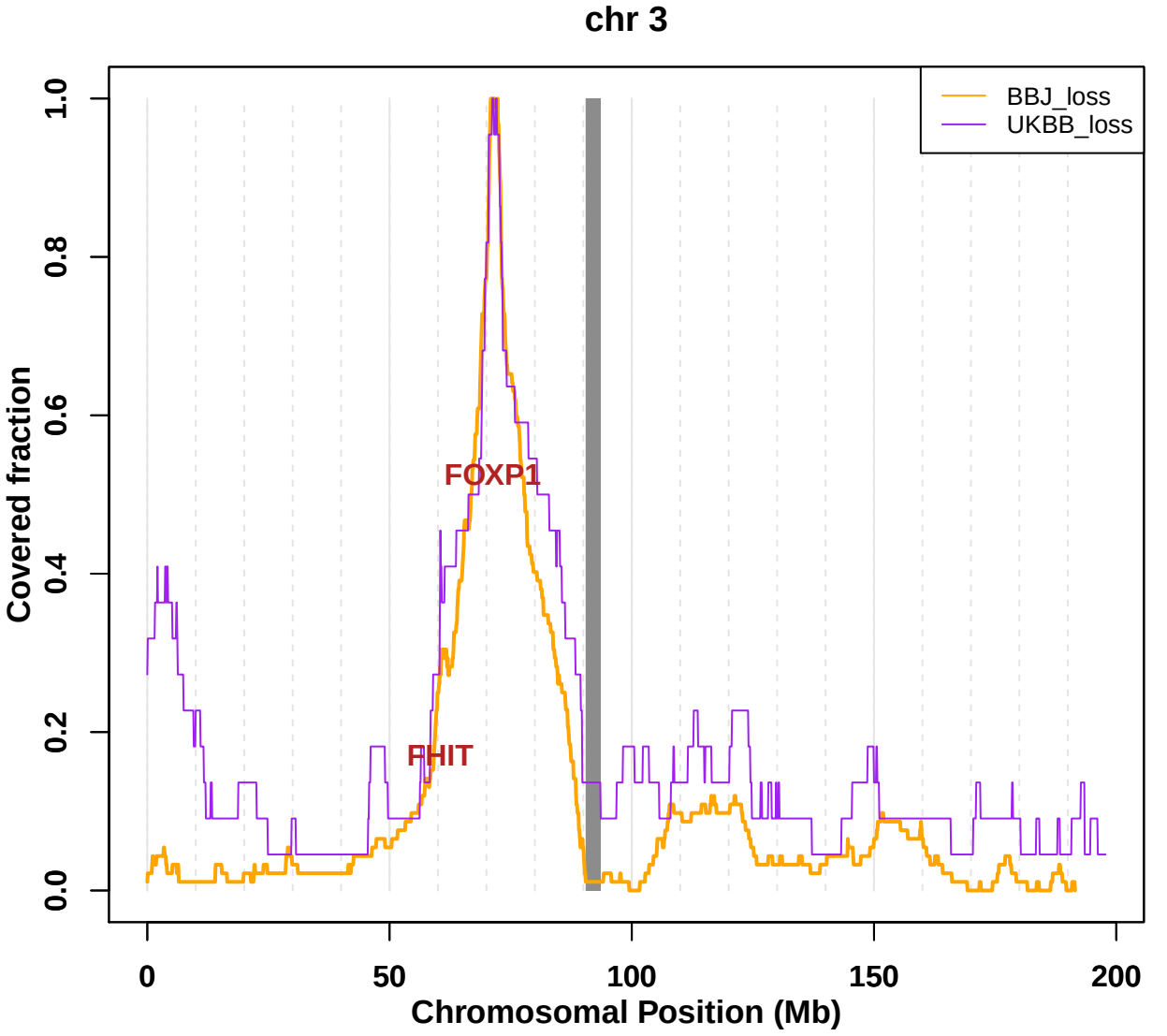

Fig. S2.2.3 Coverage of mosaic loss in chromosome 3.

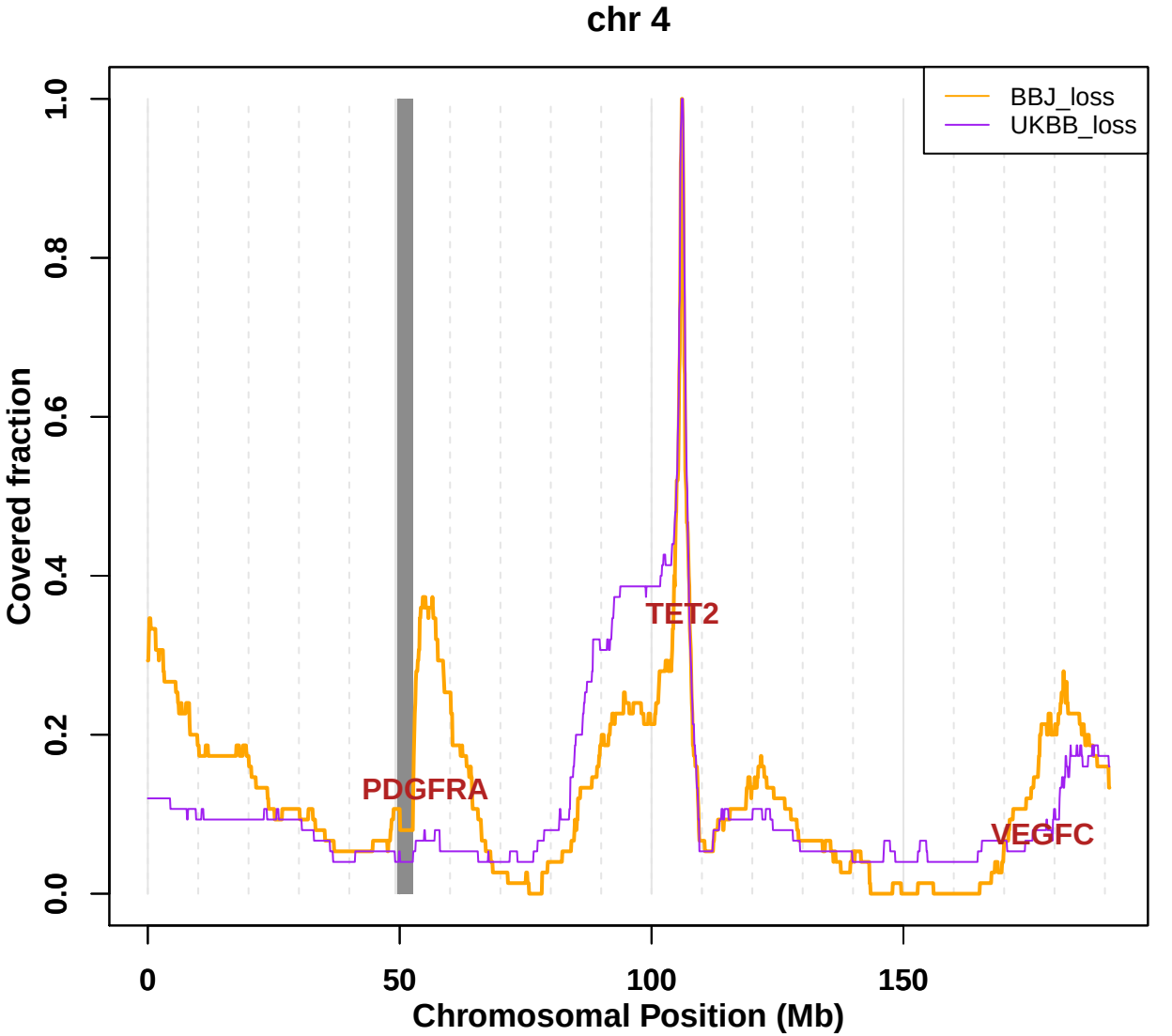

Fig. S2.2.4 Coverage of mosaic loss in chromosome 4.

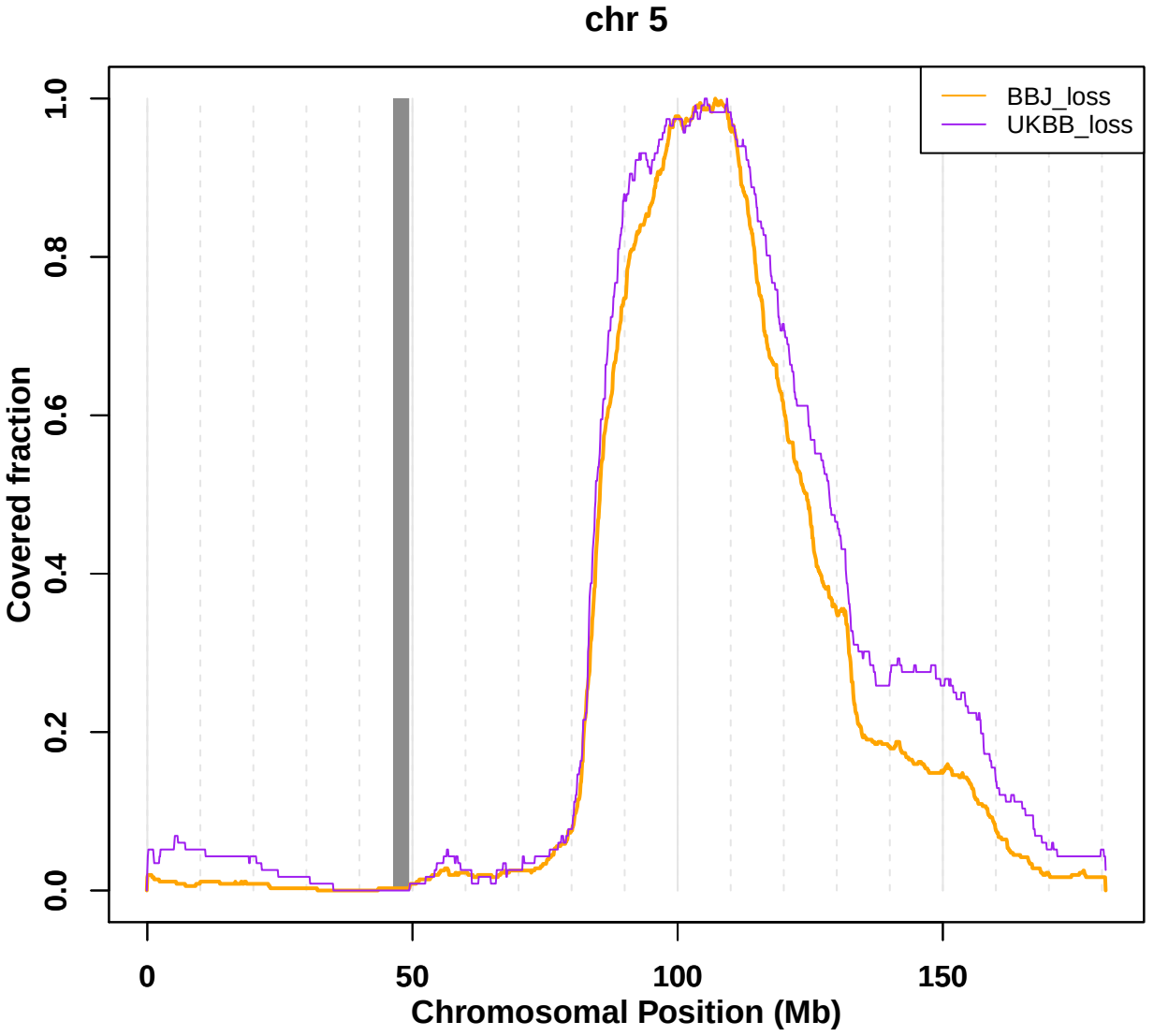

Fig. S2.2.5 Coverage of mosaic loss in chromosome 5.

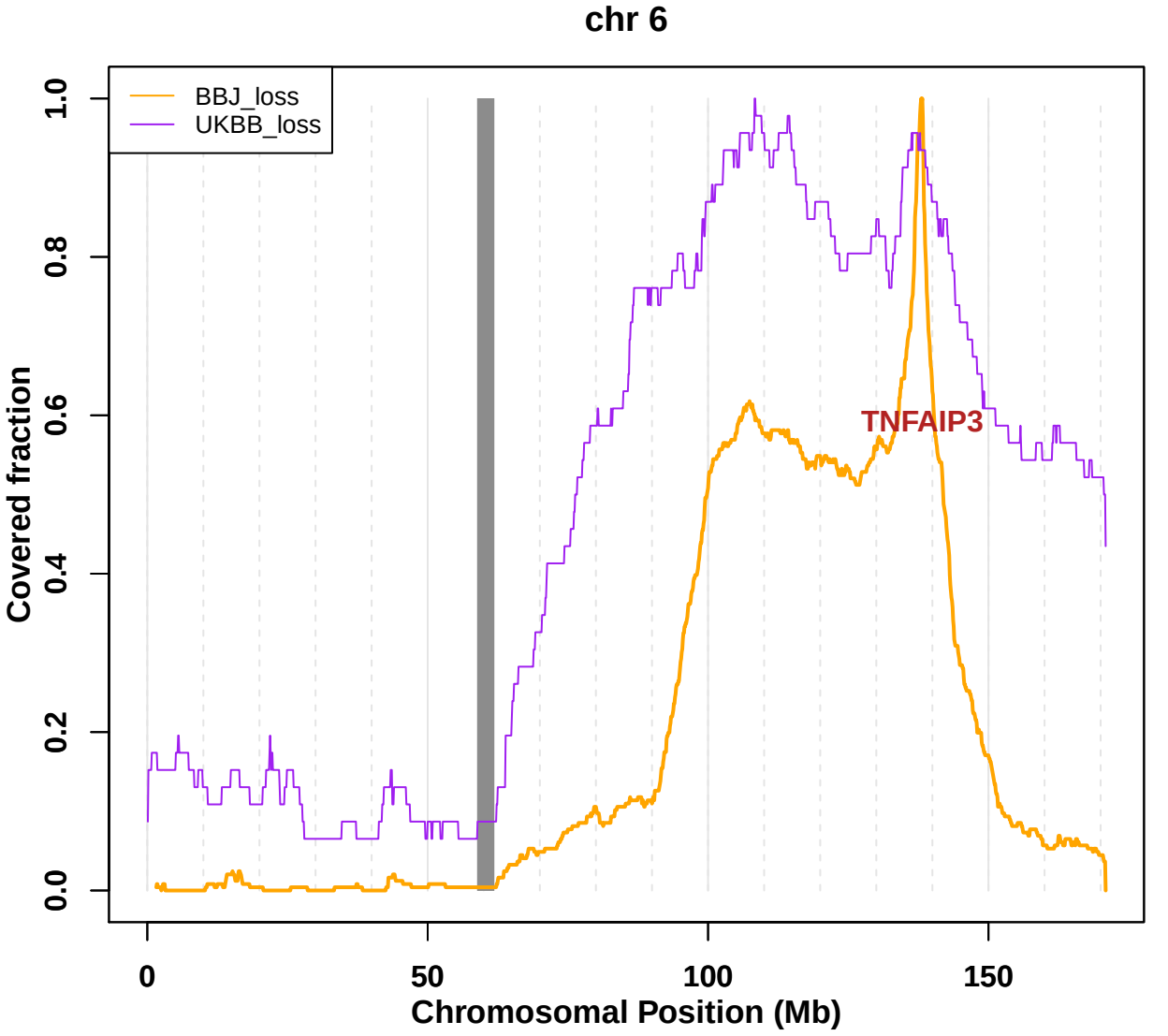

Fig. S2.2.6 Coverage of mosaic loss in chromosome 6.

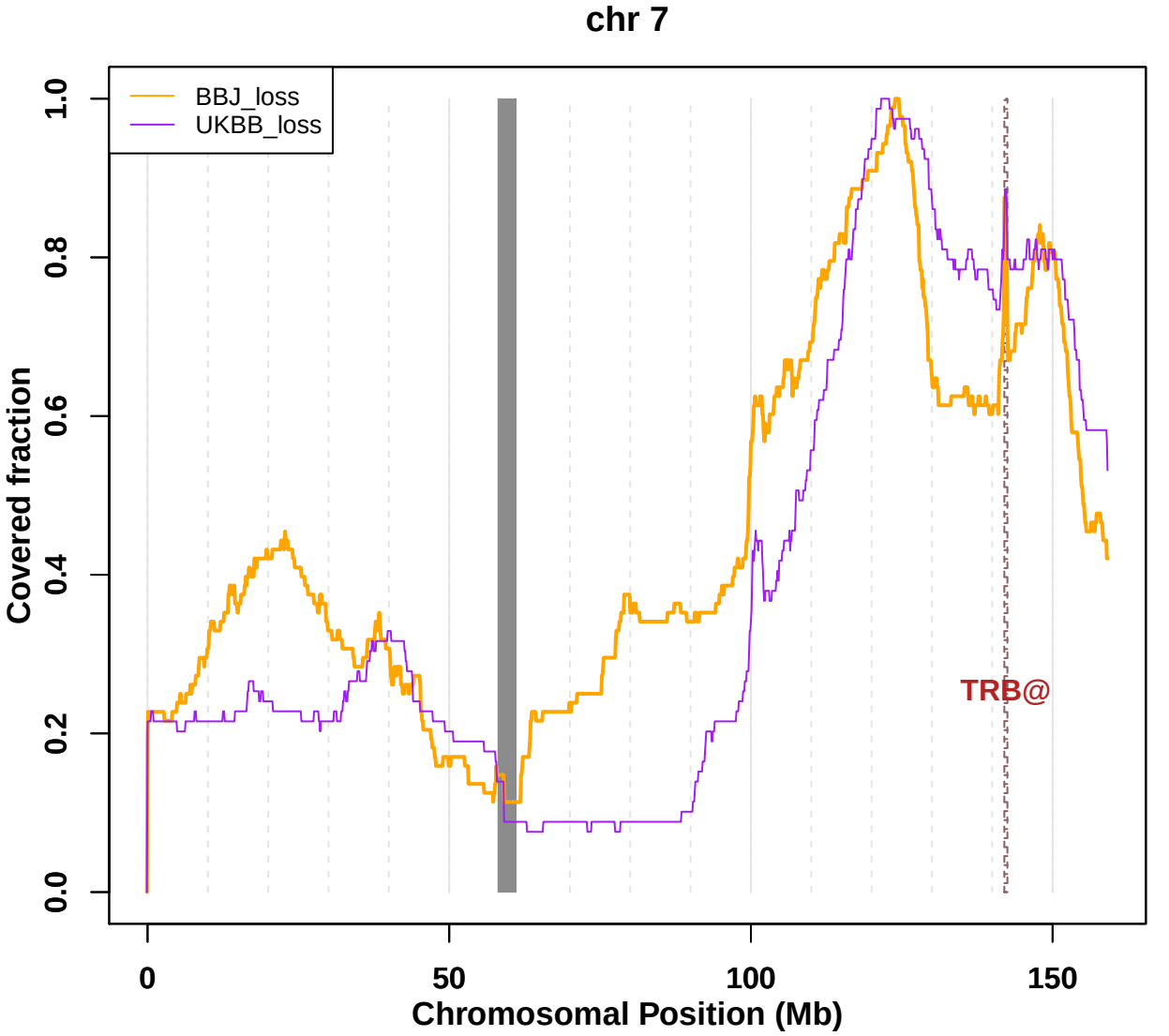

Fig. S2.2.7 Coverage of mosaic loss in chromosome 7.

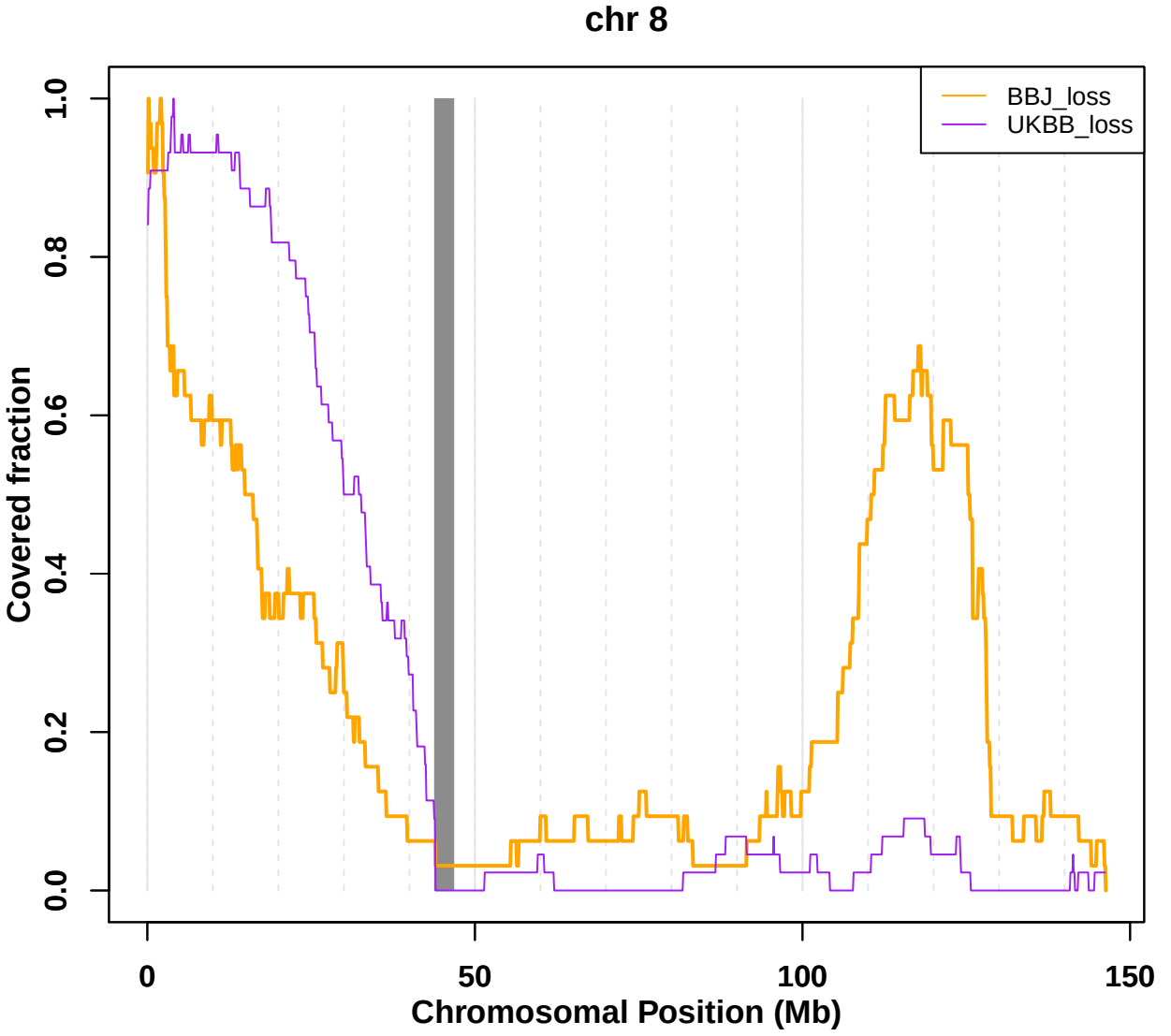

Fig. S2.2.8 Coverage of mosaic loss in chromosome 8.

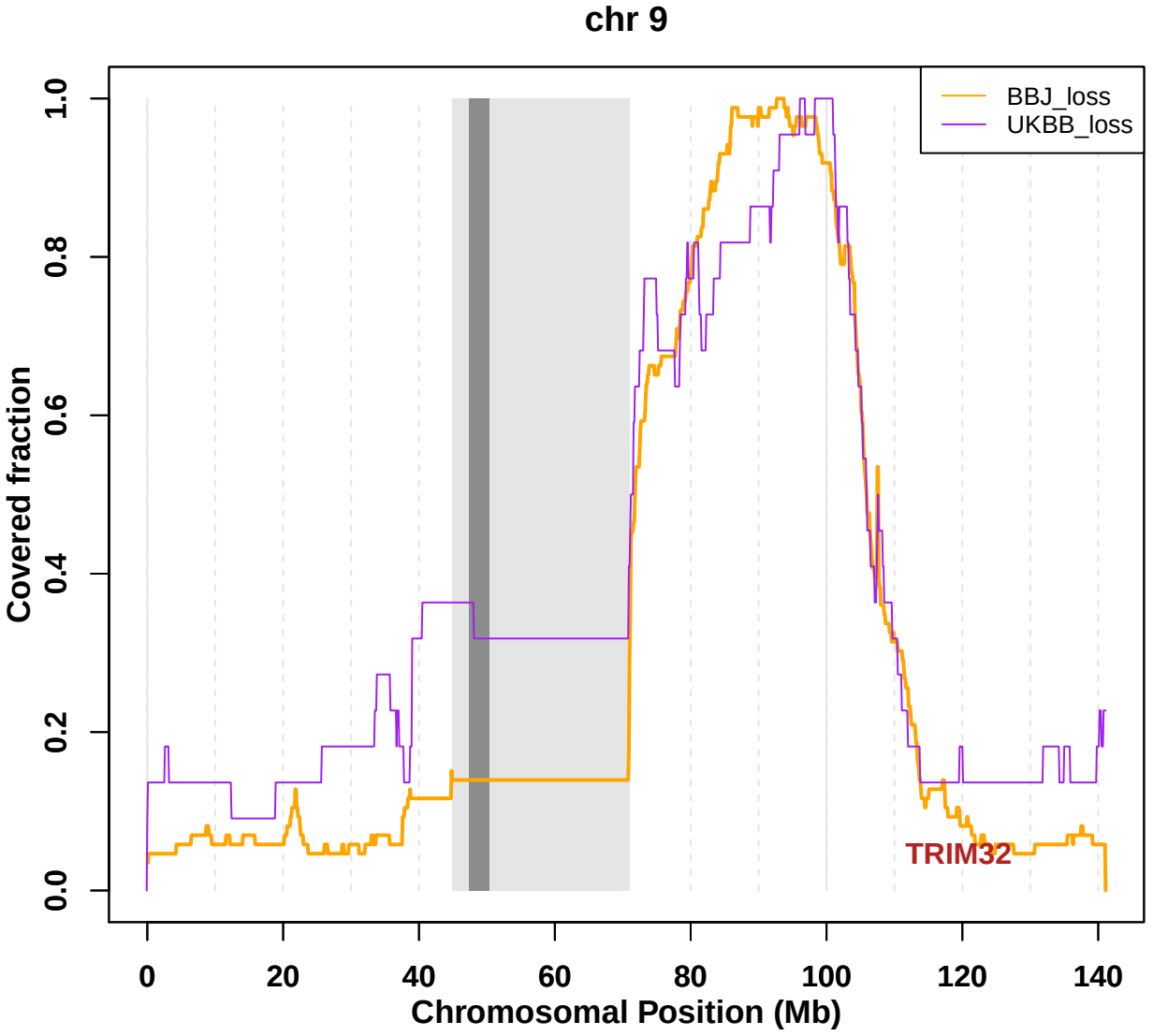

Fig. S2.2.9 Coverage of mosaic loss in chromosome 9.

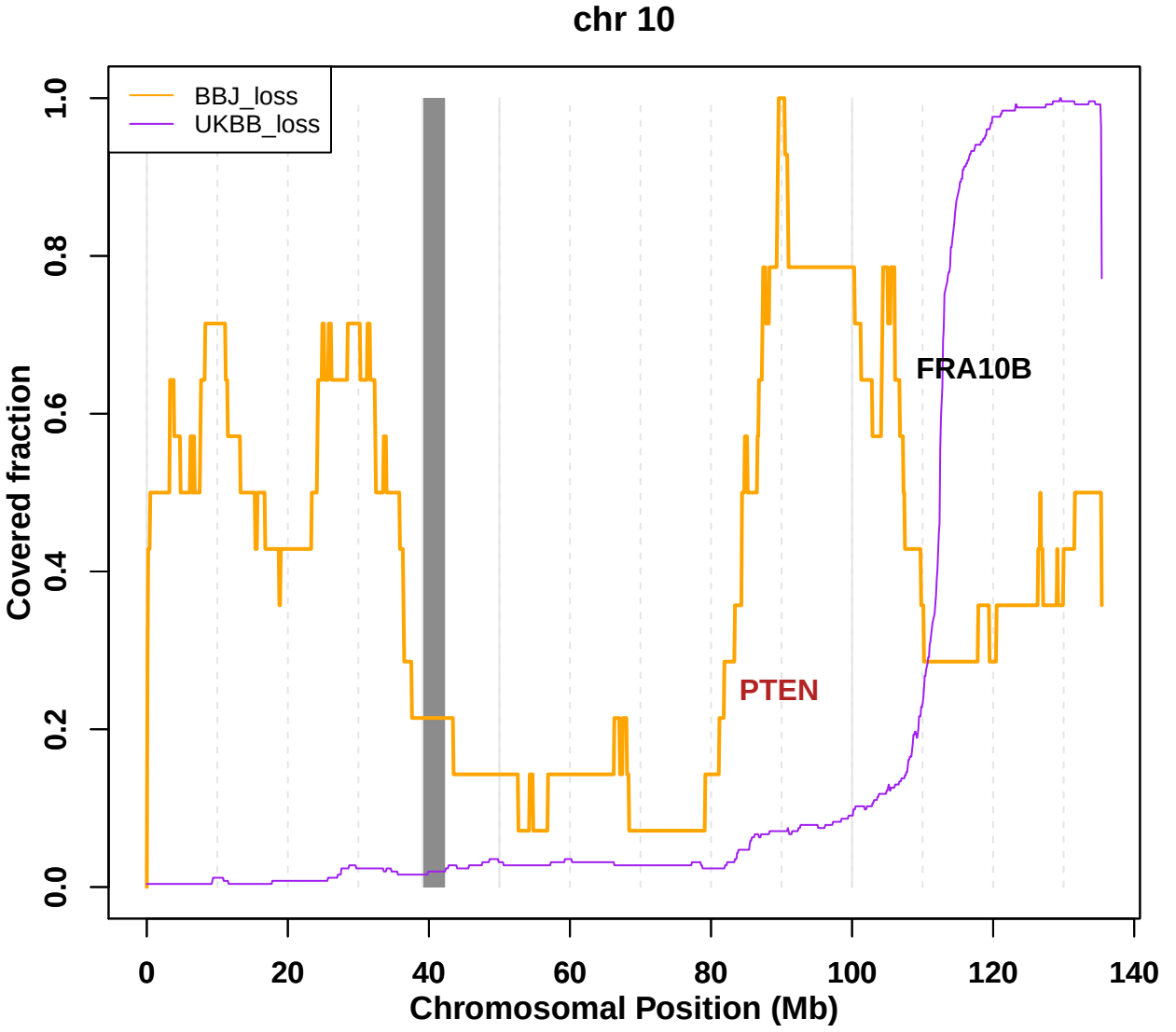

Fig. S2.2.10 Coverage of mosaic loss in chromosome 10.

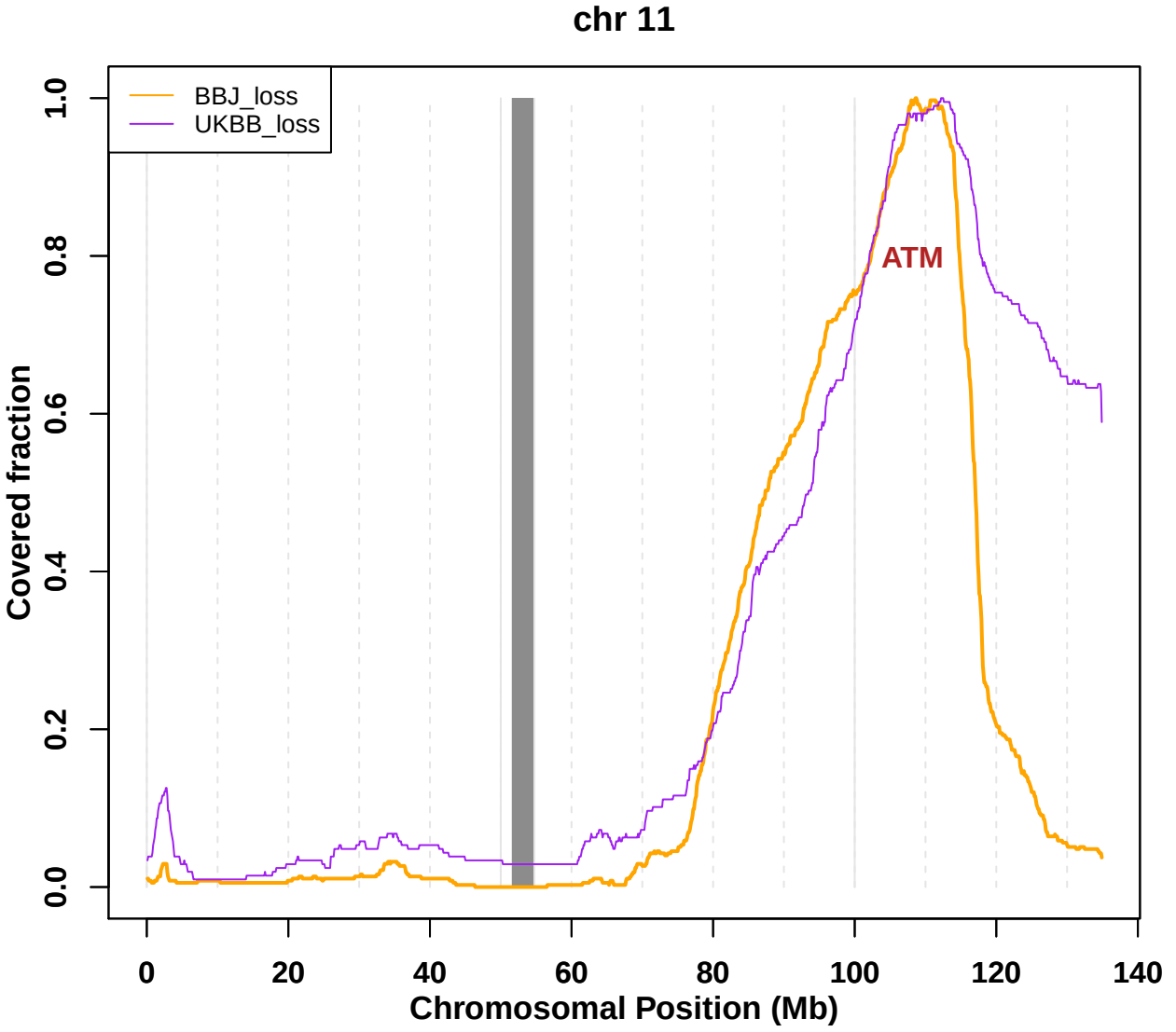

Fig. S2.2.11 Coverage of mosaic loss in chromosome 11.

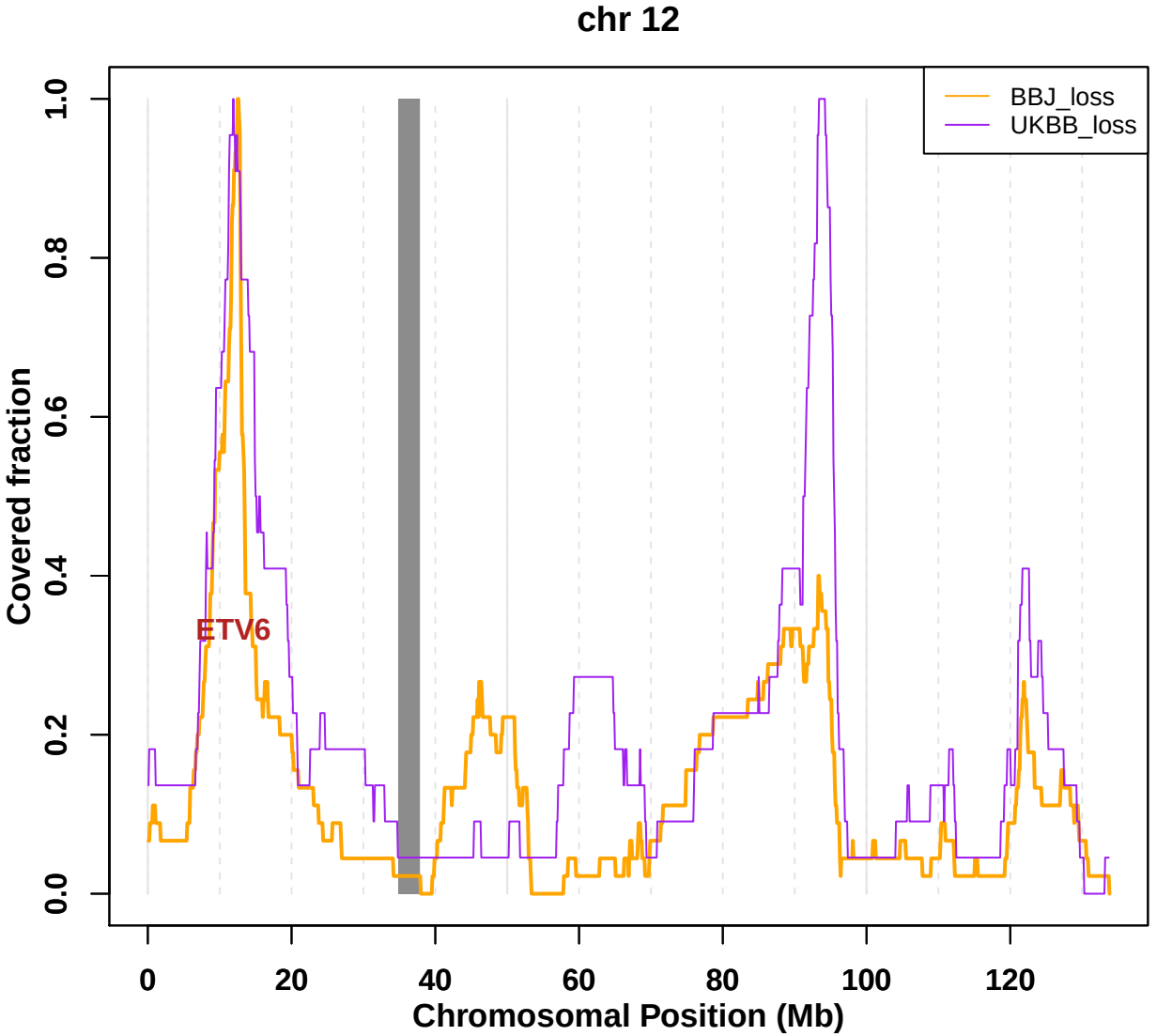

Fig. S2.2.12 Coverage of mosaic loss in chromosome 12.

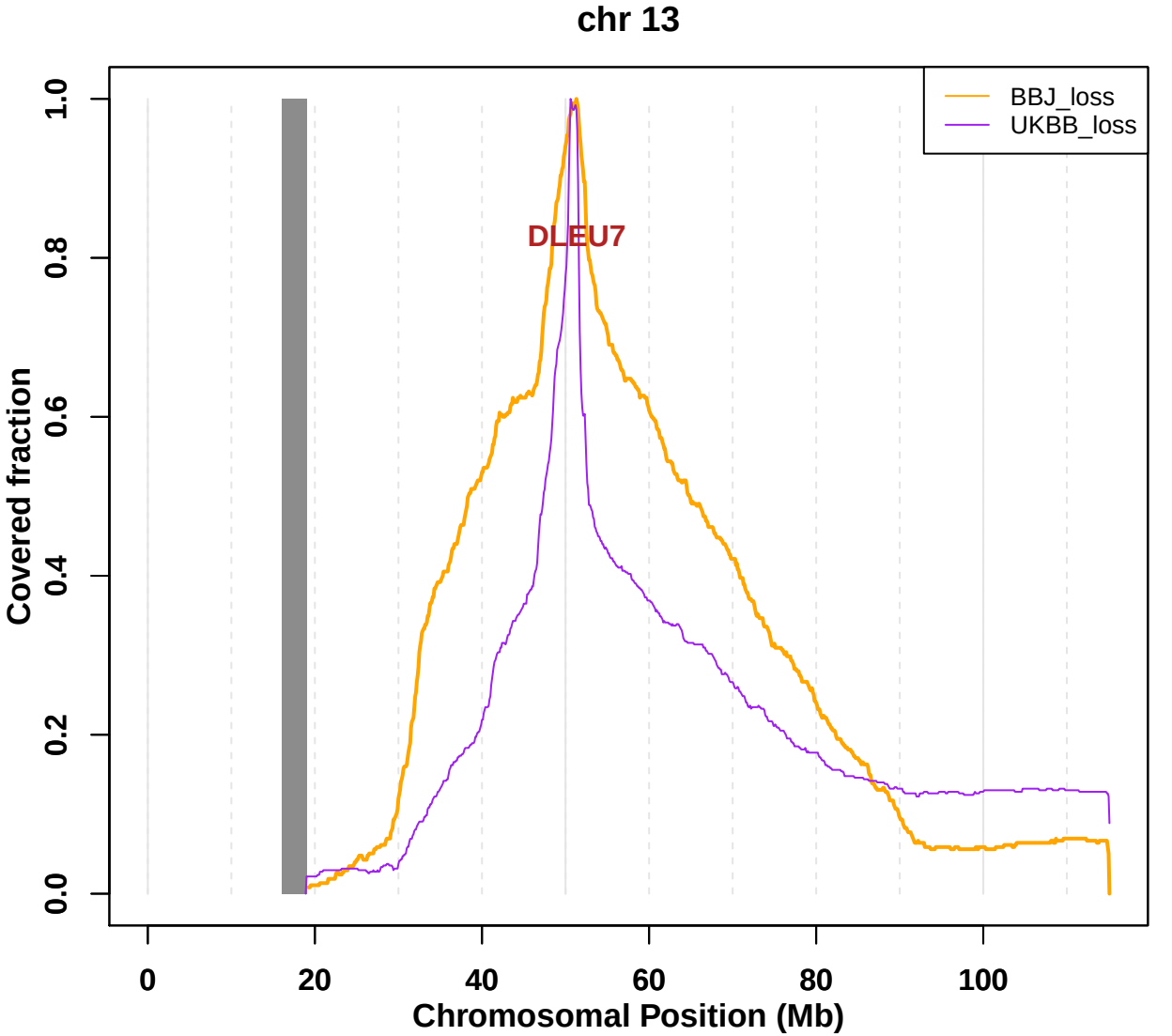

Fig. S2.2.13 Coverage of mosaic loss in chromosome 13.

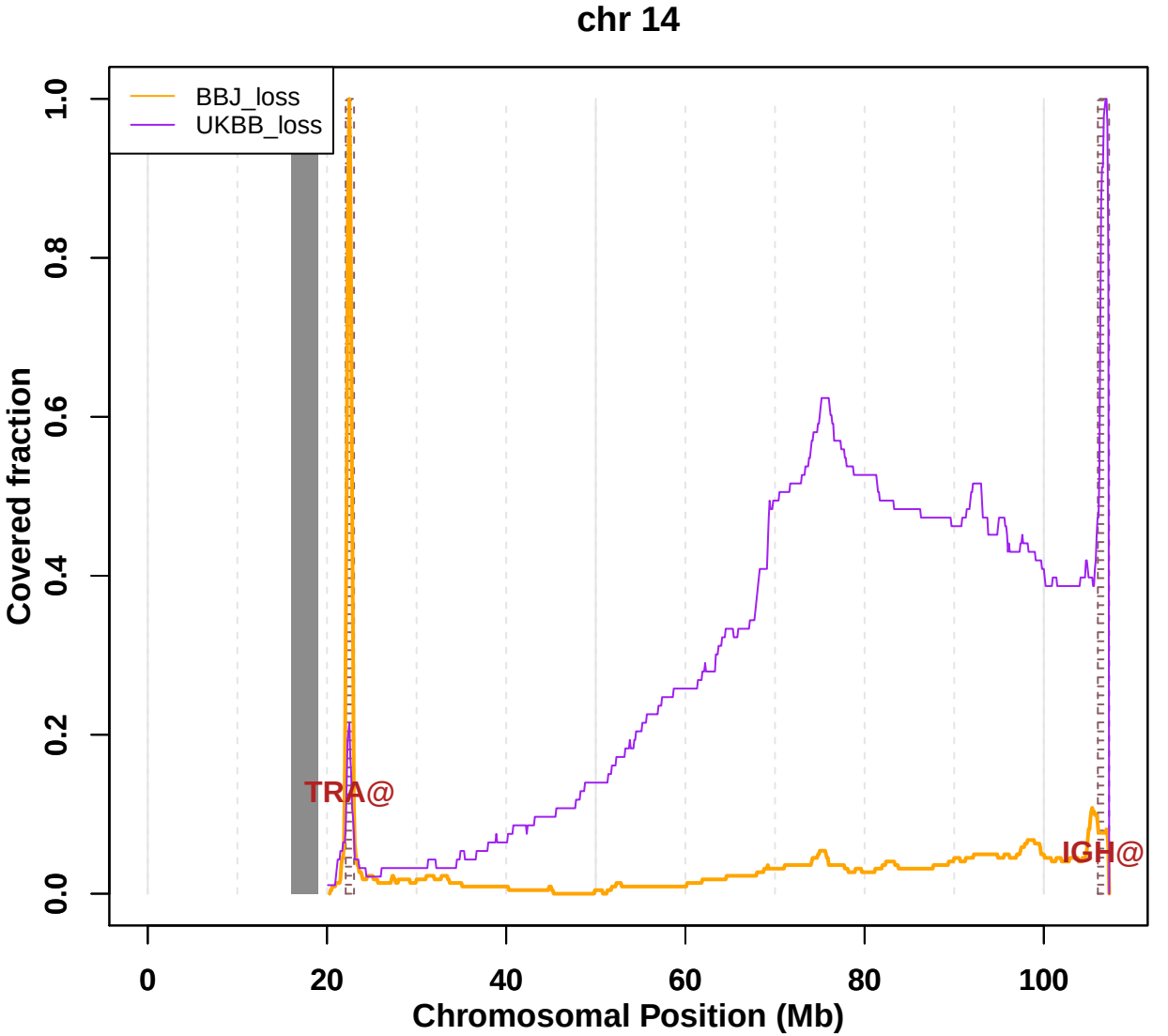

Fig. S2.2.14 Coverage of mosaic loss in chromosome 14.

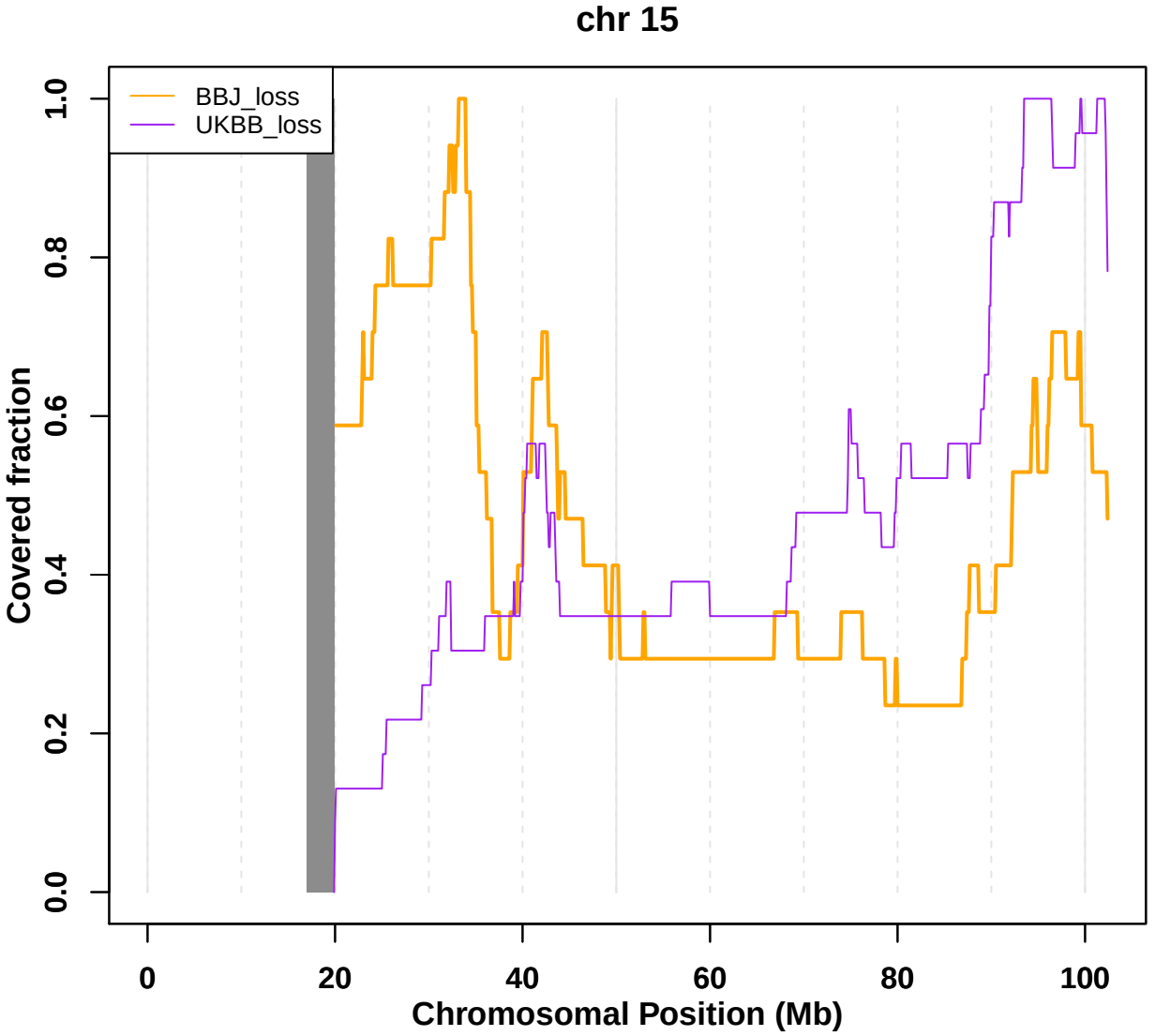

Fig. S2.2.15 Coverage of mosaic loss in chromosome 15.

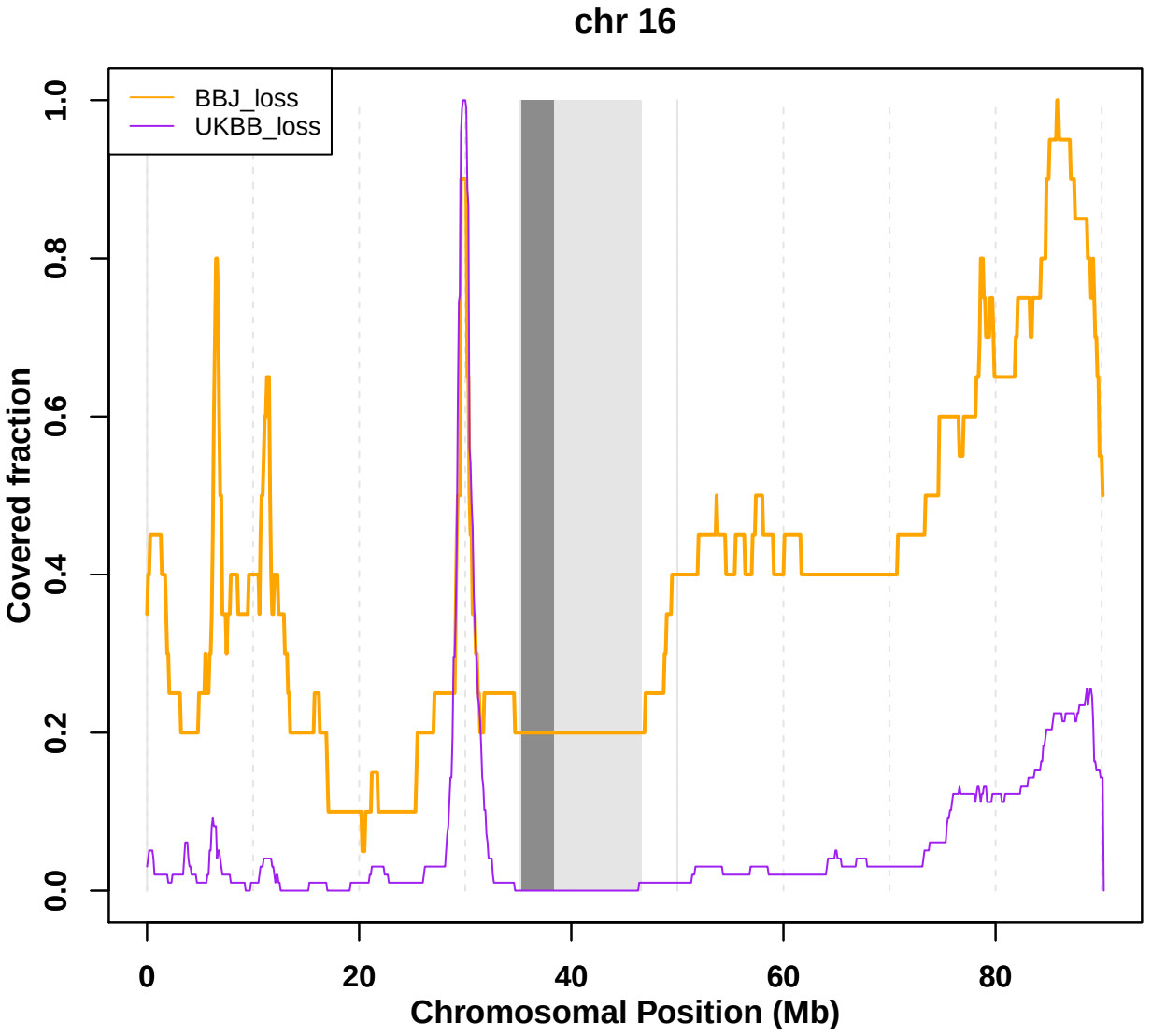

Fig. S2.2.16 Coverage of mosaic loss in chromosome 16.

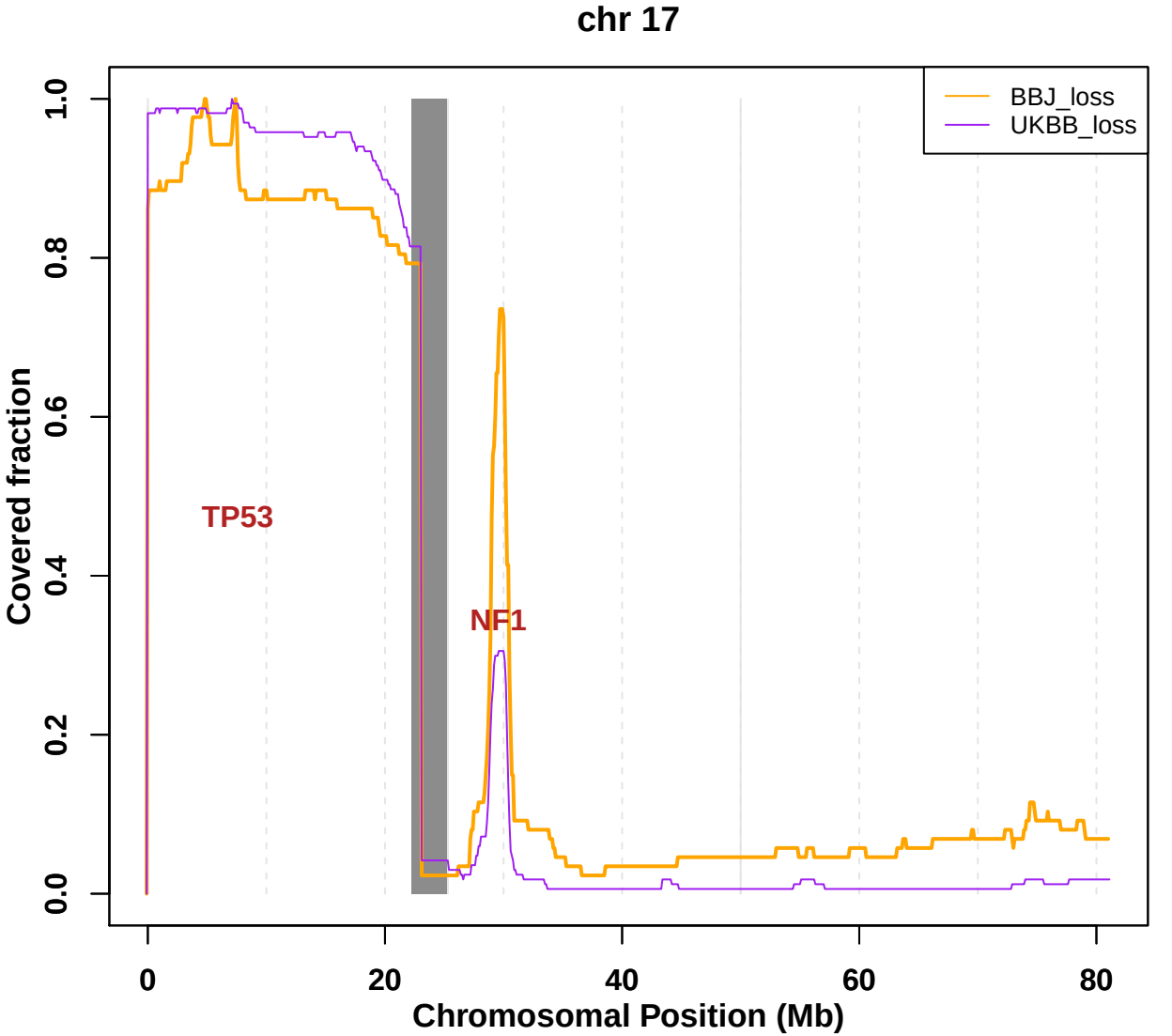

Fig. S2.2.17 Coverage of mosaic loss in chromosome 17.

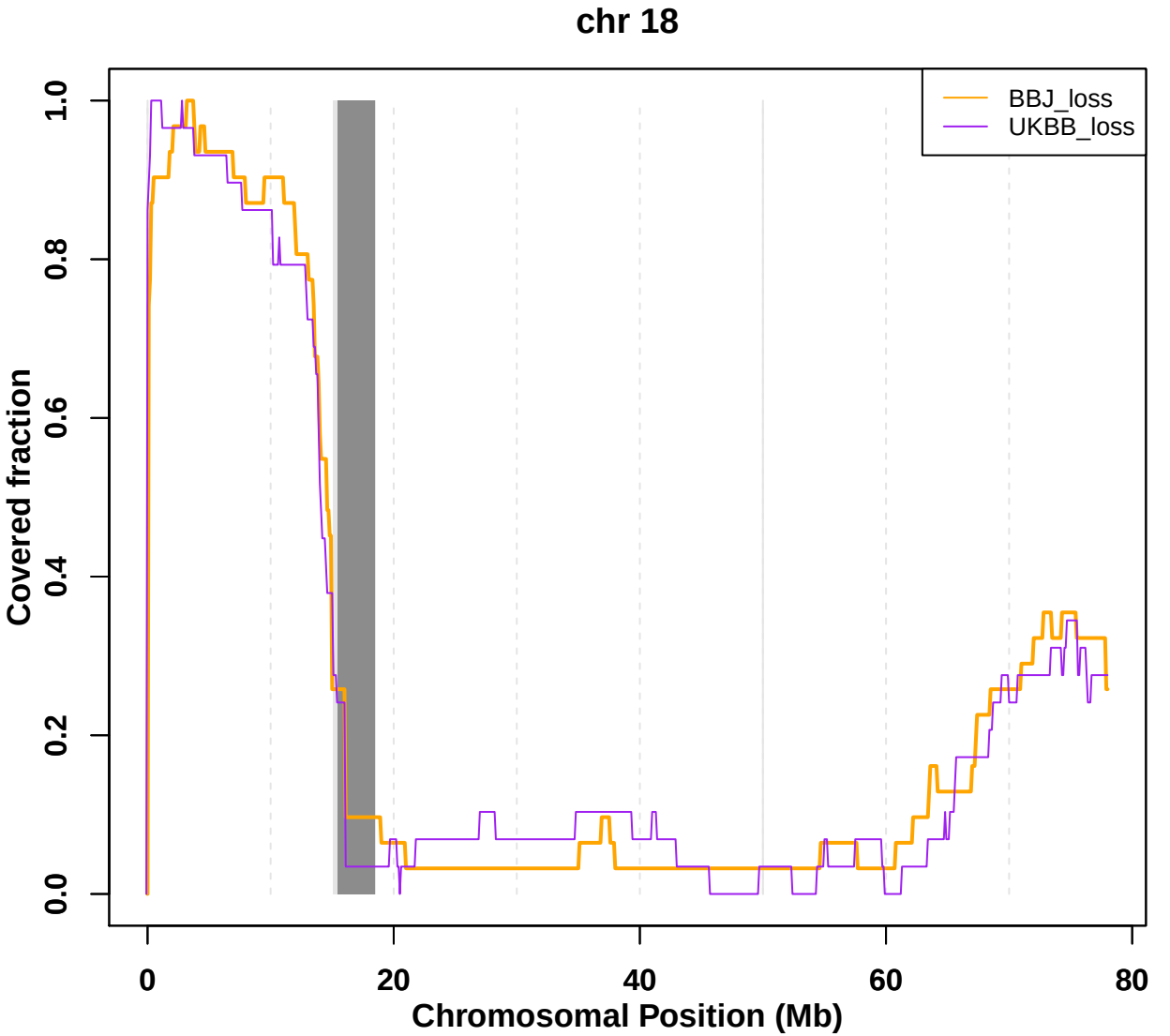

Fig. S2.2.18 Coverage of mosaic loss in chromosome 18.

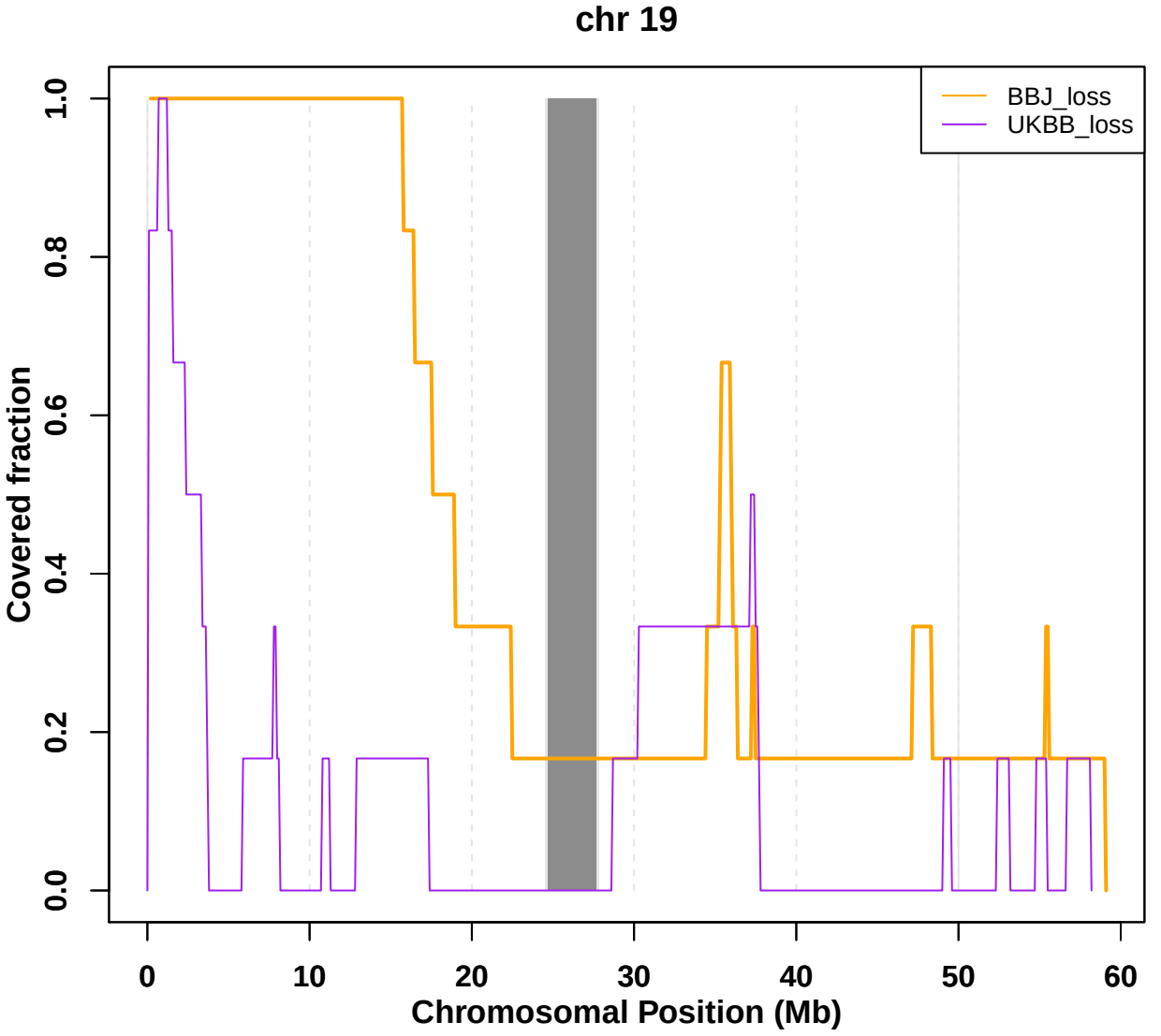

Fig. S2.2.19 Coverage of mosaic loss in chromosome 19.

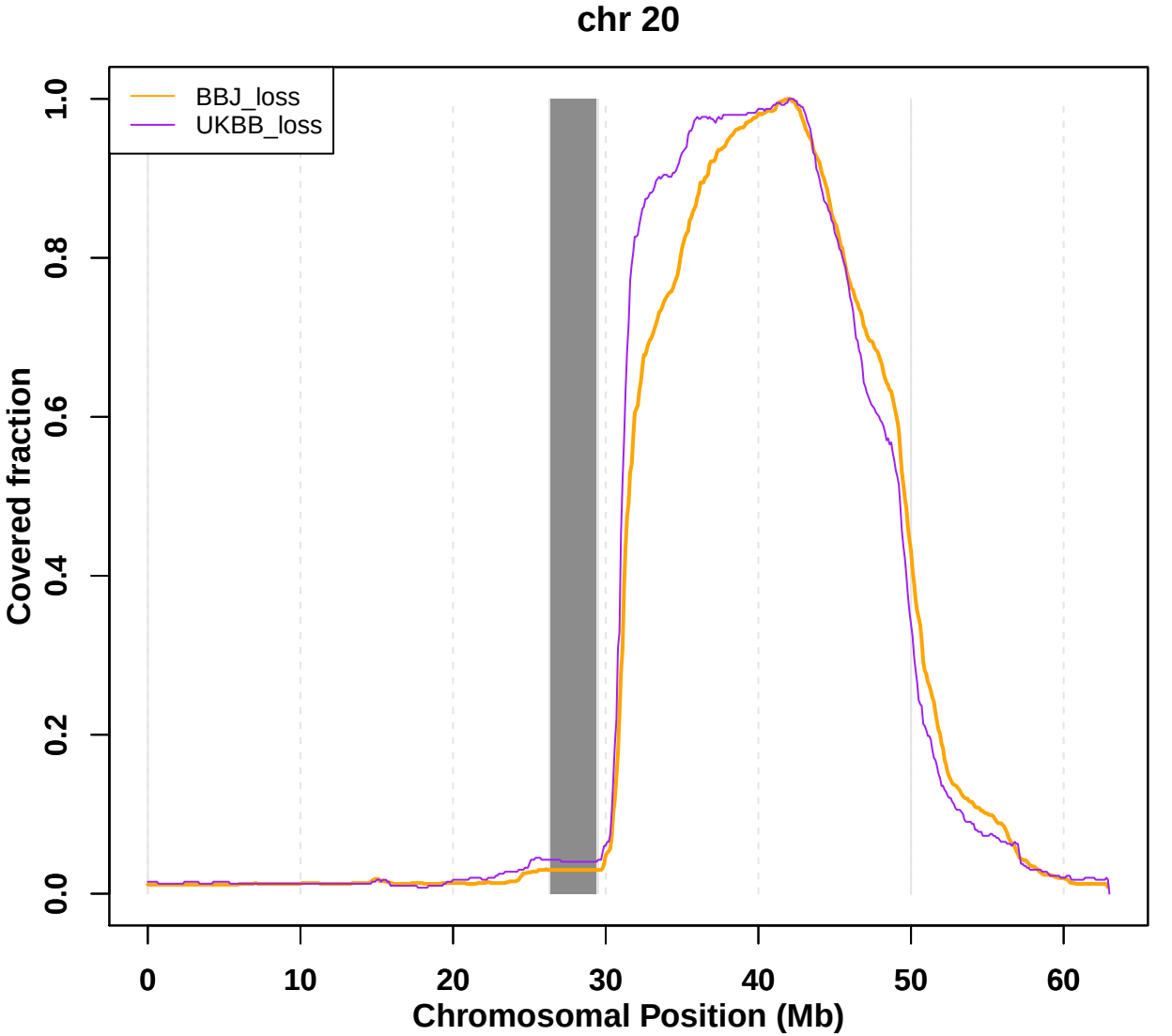

Fig. S2.2.20 Coverage of mosaic loss in chromosome 20.

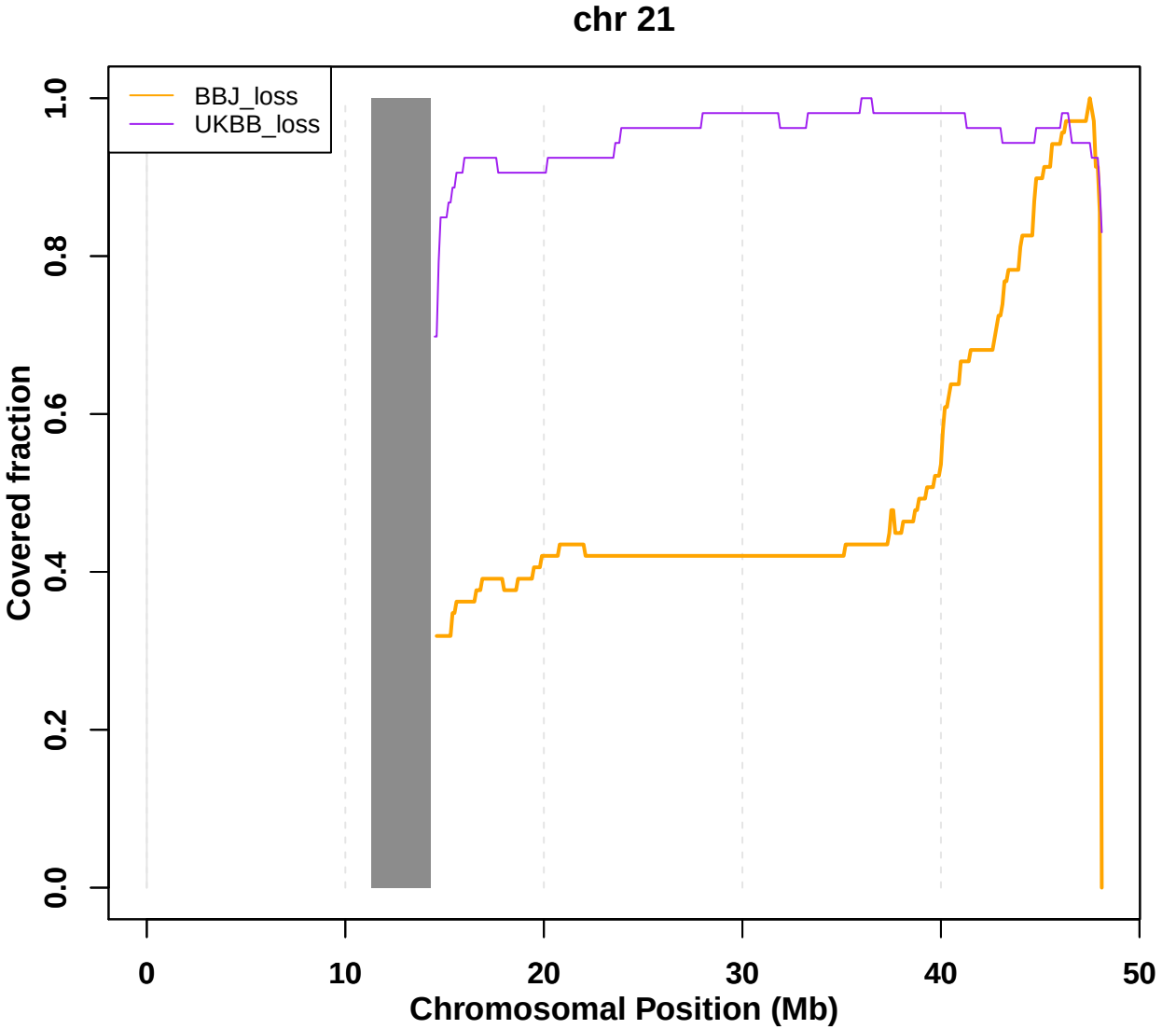

Fig. S2.2.21 Coverage of mosaic loss in chromosome 21.

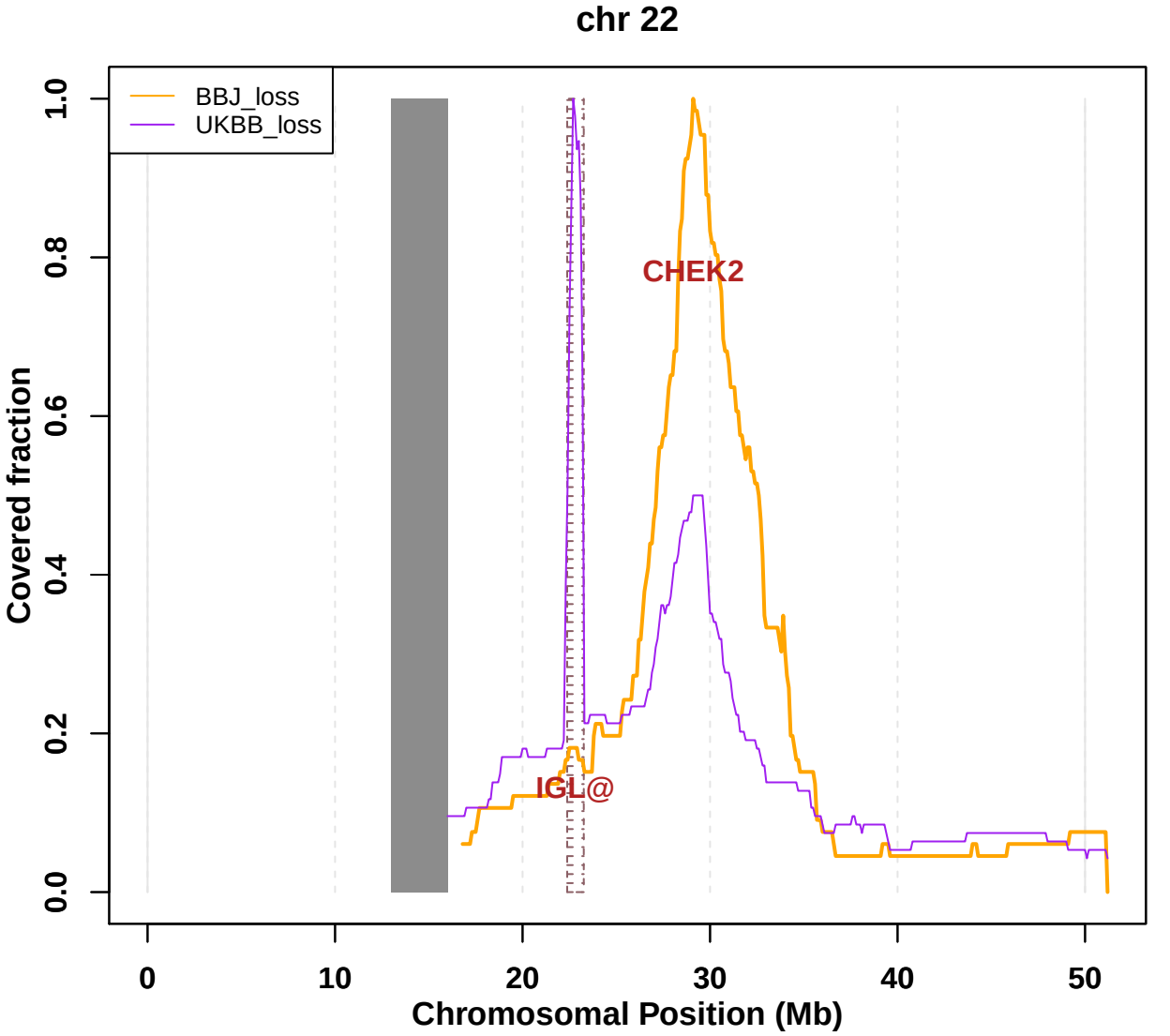

Fig. S2.2.22 Coverage of mosaic loss in chromosome 22.

3. Multiple clones, breakpoints and clone sizes of mosaic events.

3.1. Co-occurrence of mCAs in different chromosomes.

We assessed co-occurrence of classified mosaic events in single individual. We excluded co-occurrence of mosaic events in the same chromosome (for instance 1. multiple LOSS in the same chromosome, 2. LOSS in chromosome 17p and GAIN in chromosome 17q). We evaluated odds ratio (OR) of co-occurrence by Fisher's exact test based on 2x2 tables (rows and columns corresponding to individuals with and without the two mosaic events). We excluded subjects having mosaic events more than 5.

3.1.1. Common and specific pattern of co-occurrence of mCAs

We observed 30 combinations of mosaics which significantly co-occur in the BBJ (FigS3.1.1 and Supplementary Table 24). Five out of the 30 combinations, including a combination of chromosome 3 gain and chromosome 18 gain with the strongest association, were also reported in the UKB. The common combinations include a combination of chromosome 12 gain and chromosome 13q loss, both of which were associated with CLL. These suggest common and population-specific mechanisms underlying co-occurrence of mosaic events.

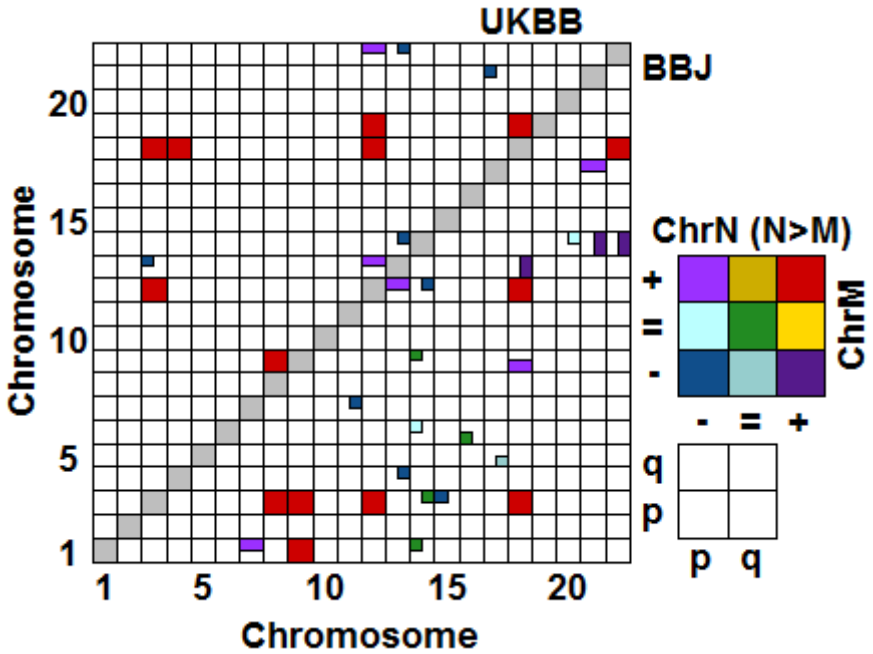

Fig. S3.1.1 Co-occurrence patterns of mosaic events in different chromosomes in BBJ and UKB. +, =, and – indicates gain, CN-LOH and loss, respectively. UKB reports co-occurrence of the same chromosomal combinations (1) 13q- and 3p- or 3+ (2) 14q- and 13q- or 13q=, and (3)

22q- and 13q- or 13q=. If there are multiple co-occurrences in the same chromosomal combinations in a single population, we show one of two co-occurrence as a representative.

3.1.2. Cell fractions in multiple mosaic events in subjects.

We extracted subjects having two mosaic events in different chromosomes. We compared cell fractions of two mosaic events in subjects (Fig. S3.1.2). At least 54.5% subjects were estimated to have different cell fractions, suggesting multiple clones with mosaic events. We further analyzed whether specific combinations of chromosomes showed skewness of lack of difference in cell fractions, but we did not find specific chromosomal combinations. These results suggest that mosaic events do not usually occur at the same time in a single cell in spite of findings of specific mosaic combinations frequently observed in a single subject.

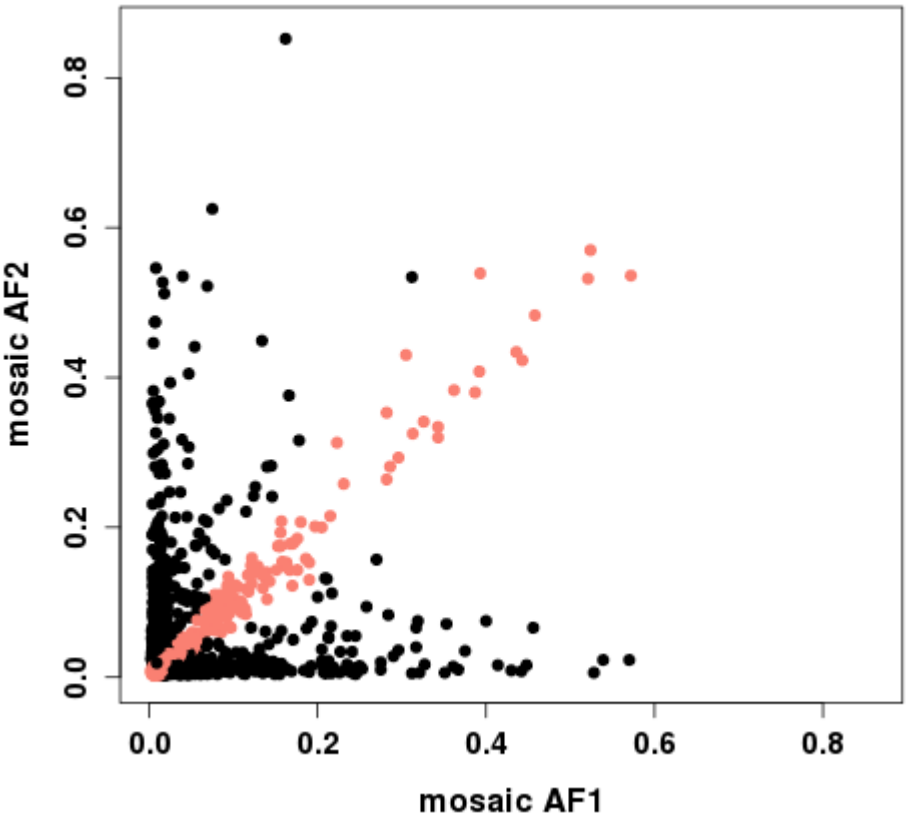

Fig. S3.1.2. Cell fractions of multiple mosaic events occurring in the same subject.

Mosaic allelic fractions (AF) are plotted in subjects with 2 mosaic events in different chromosomes. We computed difference in cell fractions between two mosaic events and regarded cell fractions as different if the difference in fractions satisfied the two conditions; 1)

difference is more than 50% of the smaller AF and 2) difference is more than 0.01. As a result, 54.5% of subjects were estimated to have different cell fractions for mosaic events, suggesting multiple clones. Since it is hard to distinguish different clones with similar fractions in the remaining subjects, 54.5% should be considered as a minimum number.

3.2. Evidence for population differences in clonal selection on CLL-associated mCAs

The differences in frequencies of CLL-associated mCAs between Japanese and UK population may not necessarily indicate different clonal selection of the mCAs between the populations. Another possible explanation is a difference in mutation rates in the CLL-associated loci between the populations. However, we identified four lines of reasoning indicating that differences in clonal selection are a more likely explanation than differences in mutation rate as follows (3.2.1-3.2.4).

3.2.1. No inherited variants associated with CLL-associated mCAs.

We analyzed whether formation of CLL-associated mCAs was associated with genetic variants in the Japanese population. As a result, we did not find any significant associations. Lack of significant associations was also observed in the analogous analyses in the latest study of the UKB[10].

3.2.2. No population-specific fragile sites associated with CLL-associated mCAs.

We analyzed whether Japanese or European-specific fragile sites are present in chromosome 13 (which in theory could explain a difference in rates of 13q loss or 13q CN-LOH analogous to fragile alleles at *FRA10B* associated with 10q deletion in UK Biobank). As a result, we did not find fragile alleles associated with chr13 mCAs in Japanese or UK population, nor did we observe enrichment of breakpoints in any specific location on chromosome 13 for mCAs in Japanese vs. UK individuals (Fig. S3.2.2). (Fragile sites cannot be involved in trisomy 12, as these events arise from missegregation rather than DNA breaks.)

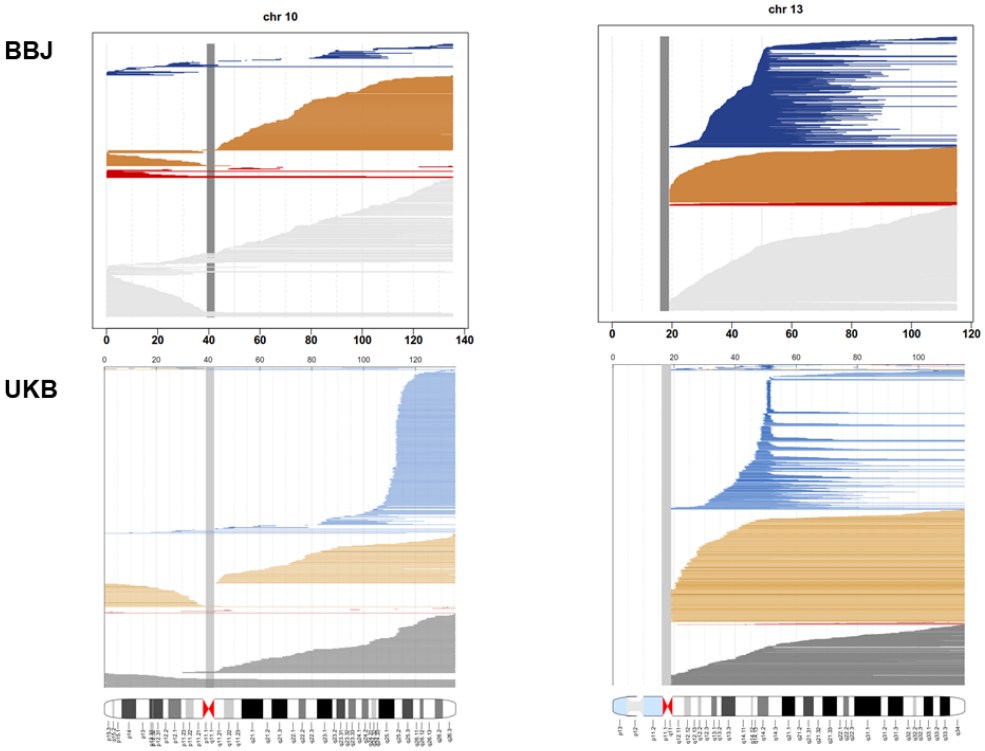

Fig. S3.2.2. No population-specific fragile sites associated with CLL-associated mCAs in chr13. We plotted the distribution of mCAs in chromosomes 10 and 13 in the Japanese and UK populations. The UK population shows a distinct fragile site in chr10 (*FRA10B*) associated with loss events. No fragile sites are observed in chr13 in the Japanese and UK population.

3.2.3. Overlap between CLL-associated mCAs and mCAs with different frequencies between the two populations

Chr12 gain, chr13q loss and chr13q CN-LOH are three of the four mCAs most strongly associated with CLL in the previous UKB study. These mCAs are also three of the four mCAs showing different frequencies between Japanese and UK population. Considering the number of chromosomes, chromosomal arms and type of mCAs arising from very different mutational processes, the consistency of this enrichment of CLL-associated mCAs in European vs. Japanese argues for a difference in selective pressure on CLL-associated clones rather than a difference in mutational propensities.

3.2.4. Smaller clone sizes for trisomy 12, 13q loss, and 13q LOH events in BBJ than in UKB

We compared clone sizes (estimated allelic fractions of mCAs) of CLL-related mCAs between BBJ and UKB. As a result, we observed smaller clone sizes in the BBJ (Fig. S3.2.4). We caution that comparison of the distributions of clone sizes in BBJ vs. UKB is complicated by differing detection sensitivity (due to different genotyping arrays and different length distributions for 13q loss

events), but this observation again broadly suggests stronger selection in the UK population.

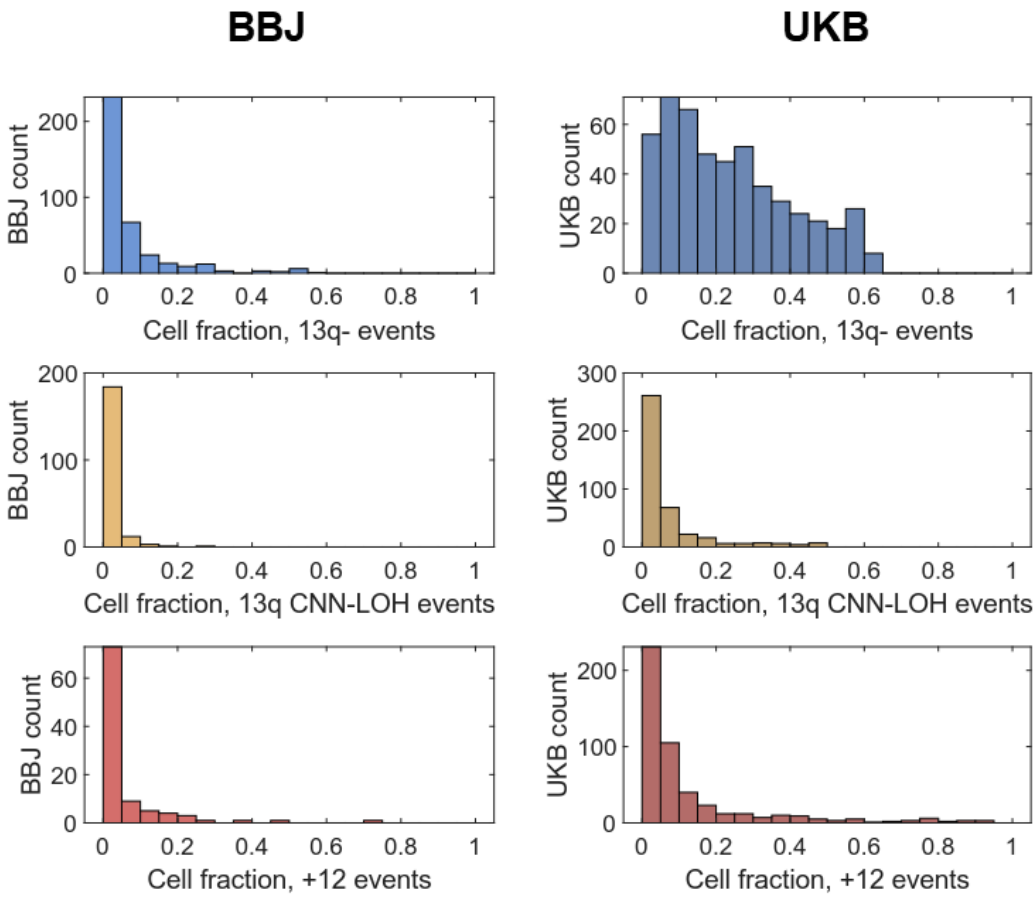

Fig. S3.2.4 Smaller clone sizes of CLL-precursor mCAs in the BBJ than the UKB.

We compared clone sizes (fraction of mCAs) of CLL-precursor mCAs between BBJ and UKB. We observed different distribution of clone sizes between the two cohorts. The BBJ tends to be smaller than the UKB. This trend is especially obvious for 13q- events.

3.3. Analysis of breakpoints in mCAs

3.3.1. Consistent breakpoint and coverage of CN-LOH between BBJ and UKB.

CN-LOH events, which typically extend from an interstitial breakpoint to a telomere, exhibited similar quantitative distributions across the autosomes in BBJ and UKB (Fig. 3c and Extended Data Fig. 3). Correlations of CN-LOH chromosomal coverage between BBJ and UKB ranged from 0.80–1.00 (Supplementary Table 13), suggesting cross-population consistency in the mutational process (mitotic recombination) and selective pressures that lead to CN-LOH events in clonal hematopoiesis.

3.3.2. Multiple breakpoints in the same individuals

We analyzed whether there are multiple CN-LOH clones with different breakpoints affecting the same chromosome arm in a single individual as previously reported in the UKB. To detect these events, we took the same approach as the UKB study. Briefly, in Viterbi decoding, we introduced transitions of BAF deviations between non-zero status (namely, mosaic status) with probability of 10^{-7}. Based on the results of the UKB study, we assessed this phenomenon by searching for a signal of increasing BAF deviation (defined as a step-function increase of more than 0.005) toward a telomere. As a result, we found evidence of multiple clones in 185 subjects (Extended Data Fig. 7). Later, we analyzed whether the presence of multiple clones in the same subjects was driven by inherited genetic variants (see Supplementary Note 6.4.6).

4. Associations between mCA and non-genetic phenotypes.

4.1. Inevitable development of mosaic events in the elderly.

The trend toward inevitability in the elderly has previously been observed for mosaic of point mutations[7,26]; however, previous studies of large chromosomal alterations have detected autosomal mCAs in at most 13% of very elderly individuals[1,2,5,6,10]. Here, the observation of mCA rates reaching 40% was driven by the combination of sensitive statistical methods[10], low noise in measuring allele-specific copy number, and representation of very elderly individuals in the BBJ cohort.

4.2. Common skewing age and sex associations with mCA between BBJ and UKB.

The BBJ data revealed that chr15 gain strongly skewed towards males and the elderly, which was also observed in the UKB[10]. Chr20q loss also showed the same pattern in both populations. In the BBJ data, we did not observe any mosaic types with female dominance which was different from that in the UKB.

4.3. Associations between quantitative hematologic traits and mosaic events.

We found significant associations of multiple loss and gain events with hematologic traits (Supplementary Table 9). Since no CN-LOH reached statistical significance in spite of the CN-LOH events accounting for most mosaic events, this result suggests larger influences of loss or gain, which decrease or increase copy number, on hematopoiesis.

4.4. mCAs in association with diseases at registry, especially with Graves' disease.

While we observed a strong association between hematopoietic malignancy at registry and presence of mCAs, this association indicates a direct observation of the presence of (pre-)malignant cells with chromosomal alterations. Associations with other diseases are expected to have smaller effect sizes (and may nonetheless be of interest).

Since we found a trend of protective association between mCAs and Graves' disease (GD), we analyzed whether specific mCAs were enriched/less enriched for GD. As a result, we did not find any specific mCAs associated with GD ($p>0.0082$; mCAs in chr14 (of any type) showed the smallest p-value, consistent with chr14 events being the most frequently observed type of mCA), indicating that the association pattern between GD and mCAs was not driven by a specific type of mCA.

We did not observe GD-associated variants in LD with mCA-associated variants, indicating that the possible protective association with GD was not genetically supported. While a non-significant trend of this protective association was also observed in the UKB ($p=0.42$, OR:0.93 (95%CI:0.77-1.11)), the association should be regarded as inconclusive.

5. Analysis of focal deletions

We focused on mosaic deletion events in each chromosome. Focal mosaic events are determined based on plots of mosaic coverage in the two populations. We excluded chromosome 19 because deletions on chr19 are rare in both populations. We also evaluated importance of genes by taking numbers of gene involved in loss events into account. We counted the number of genes involved in each loss event and defined a score of each loss event as one divided by the number (when a loss event contained only one gene, the gene received a score of one). We summed up all scores across all loss events in each gene. To pick up genes frequently involved with focal deletions only in Japanese, we picked up genes covered by at least 5% of loss events in a chromosome, more than 10 times of scores than UKB, and scores more than 0.5.

5.1. Genes frequently involved in focal deletions in Japanese but not in UK population.

We found Japanese-specific focal deletions (Supplementary Tables 14-15). A total of 37 regions (defined by 1Mbp margin of each region) across 15 chromosomes were identified. *TNFAIP3* showed the highest scores (see Methods) among Japanese specific genes. Other genes which draw our attention were *FHIT* in chromosome 3 (which encompasses the fragile site *FRA3B*) and *VEGFC* in chromosome 4. Note that in this analysis, we focus on absolute coverage of genes by focal deletions and coverage was not scaled (different from Fig S2.2.1-22).

5.2. Focal deletions of TCR genes

Since focal deletions at the TCR alpha locus (*TRA*) in chromosome 14 were frequently found in the BBJ, we analyzed whether this trend was consistent in *TRB* in chromosome 7. We observed more frequent focal deletions in *TRB* in the BBJ than the UKB (Fig. S2.1.7 and S2.2.7).

Since genetic recombination in TCR occurs in all of lymphocytes, focal deletions in TCR genes may not necessarily indicate clonal expansion (possibly reflecting common recombination of adjacent regions among TCR genes in majority of lymphocytes). To address this point, we analyzed co-occurrence of mosaic focal deletions in *TRA* and *TRB* in the same individual. We found significant co-occurrence of these two deletions (OR 305, $p=3.5\times10^{-11}$), indicating clonal expansion.

We confirmed that these two events were not enriched for subjects with cancer at registry.

6. Genetic association studies

6.1. Mosaic types, subjects and variants for genetic associations

We analyzed mosaic events in each chromosome as distinct phenotypes, treating loss, CN-LOH and gain separately. We divided loss and CN-LOH events in each chromosome into p-arm and q-arm events. Gain was treated as a single category per chromosome. We set a threshold of at least 20 event carriers to consider an event in genetic association studies. This led to a total of 88 copy number-chromosome pairs analyzed.

We excluded subjects showing high degree of kinship (1st-degree or closer as detected by plink[42]) with other subjects, leaving 173,599 subjects for genetic association studies. Among related pairs, we retained subjects having mosaic events. We excluded subjects not carrying the mosaic event being studied but carrying any other mosaic event on the chromosome from each analysis.

We tested associations at 26.6 million variants imputed with R2>0.3 and best-guess minor allele count at least 5. Imputation details are provided in Methods.

6.2. CN-LOH for genetic association

Since the previous UKB study reported genetic associations with CN-LOH events that extended to telomeres and did not span whole chromosomes, we included in the CN-LOH category unclassified events extending to one telomere with |LRR|<0.02 in order to maximize power to identify significant associations with CN-LOH.

These subjects carrying CN-LOH or unclassified events satisfying the conditions above were used as cases for trans-associations. Since the previous UKB study demonstrated that variants associated with CN-LOH events in cis associated specifically with events spanning the variants, we further refined case definitions for cis-associations in a variant-specific manner (section 6.4.1).

6.3. Statistical method for genetic association study

We conducted Fisher's exact tests using plink software (plink --fisher --ci 0.95). We used Fisher's exact test to suppress inflation of statistics especially for rare variants. To confirm significant associations were not driven by confounding factors, we re-analyzed significant associations (detected by Fisher's exact test) in logistic regression with top 10 PCs, disease affection status at registry of the BBJ, age, sex, smoking and genotype batches as covariates. We set a stringent cut-off of significance of $p<5x10^{-9}$ and $5.7x10^{-11}$ for cis-association and trans associations, respectively, based on the following reasons.

The 26.6M variants considered in our association tests are not independent (due to linkage disequilbrium). A previous study[50] estimated at most 10M independent tests under various association testing scenarios that included both common and rare variants. Thus we set a significance threshold for cis-associations as 0.05/10M ($5.0x10^{-9}$). While we tested three types of mCAs for each variant, we did not additionally correct for three mCA types because the previous UKB study found significant associations mainly in one type of mCA (CN-LOH) and we could check significant associations for evidence of allelic imbalance to obtain further confirmation. Regarding trans-associations, we corrected for the number of tests applied (88 different mCA types) and set 0.05/10M/88 = $5.7x10^{-11}$ as the threshold for genome-wide significance.

6.4. Cis-association

6.4.1. CN-LOH for genetic cis-association

The previous UKB study demonstrated that variants associated with CN-LOH in cis showed associations specifically with events spanning the variants. Thus, we were interested in searching for associations between variants and CN-LOH events spanning the variants. Strictly speaking, for each variant, we had to identify the individuals carrying CN-LOH events spanning the variant, defined as cases. In other words, even when we analyze associations between variants and one specific mosaic type of interest (say chr 1p CN-LOH), the number of cases varies from variant to variant (in chromosome 1 p-arm). Given that the boundaries of most mosaic event calls only had megabase-scale resolution, we only recomputed cases every 1Mbp for computational efficiency. The fact that the case definitions changed every 1Mbp did not increase the number of event types being studied in that every variant was still only tested against one CN-LOH event definition (such that in particular, we did not incur additional multiple hypothesis testing burden).

6.4.2. Allelic imbalance study in cis associations for CN-LOH.

We conducted allelic imbalance analyses in mosaic events[10] (analogous to allele-specific expression in gene expression) to assess whether one of the alleles at each variant was preferentially duplicated by mosaic CN-LOH events. We took advantage of phase information of each individual and assessed allelic imbalance in each site by extracting subjects heterozygous for risk variants and carrying mosaic events. P-values of allelic imbalance were

calculated based on a binomial test under the null hypothesis that the allelic shift of mosaic events at heterozygote sites is at random.

6.4.3. Significant cis loci associated with mCA in the BBJ.

We identified five new loci showing cis-associations (with presence of mCA), namely, *NBN*, *MRE11*, *CTU2*, *NEDD8*/*TINF2* and *DLK1*. We also found *TCL1A* showing significant cis association with allele selection in mCA. The associations of *MPL* and *JAK2* were replicated in the BBJ. All of the eight associations were observed in CN-LOH, compatible with the previous study[10].

DLK1 encodes a noncanonical NOTCH ligand and is an imprinted gene[51] associated with both hematopoietic[52] and non-hematopoietic malignancies[53]. *NEDD8* encodes a ubiquitin-like protein also reported in the context of both hematopoietic and non-hematopoietic malignancy[54] and a therapeutic target for AML[55]. *TINF2* encodes a protein of a member of telosome complex which protects telomere. A mutation of TINF2 is known to cause congentical bone marrow failure[56]. *TCL1A* (T-cell leukemia/lymphoma 1A) is a susceptibility gene to mosaic of chromosome Y and hematopoietic and somatic cancer[13] and encodes a protein which expresses in fetal tissues and early stage of lymphocytes, interacts with many partners including ATM and is involved in multiple signaling including NFkB[57].

The association between a *JAK2* variant and presence of clonal expansion of V617F was previously reported. We found consistent associations between chr9p CN-LOH and variants associated with *JAK2* V617F, indicating that chr9p CN-LOH involving *JAK2* probably involve *JAK2* V617F.

A total of four associations (three cis and one trans) were identified in chr14q CN-LOH. This is compatible with chr14q CN-LOH being the most common event among autosomal mCAs and may suggest high susceptibility of mCA in chr 14 and high heritability of chr14q CN-LOH.

6.4.4. An enhanced strong association of NBN

At *NBN*, we observed a very penetrating association between the rare stop gain variant rs756831345 and chr8q CN-LOH (OR=240 (129-472), $p=1.1\times10^{-22}$) when we called mosaic events at FDR of 0.025 and restricted to events confidently called and did not include unclassified events (possible CN-LOH) in the association study. This restriction apparently reduces false-positive signals (at the cost of possibly weaker p-values). This trend of higher OR was observed in the

other two novel rare variant associations (*MRE11*: OR=52 (22-124) and *CTU2*: OR=43 (25-73)), but quite prominent in the association of chr8q suggesting that the *NBN* association is associated with mosaic events with higher cell fraction.

6.4.5. Evaluation of variants reported in the previous UKB study.

We evaluated whether variants which were significantly associated with mosaic events in the previous UKB study were present in our data and associated with mosaic events. We extracted from the current results a total of six variants, namely, three variants in *MPL* (rs144279563, rs182971382 and rs369156948 (nonsense mutation)), rs118137427 in *FRA10B*, rs532198118 in *ATM*, and rs182643535, tagging 70kb deletion of *TM2D3*/*TARSL2* region. We also analyzed whether these variants were included in the Japanese WGS used in the reference panel in the current study. As a result, we did not find these variants in our data and/or the Japanese WGS, indicating these variants being population-private.

6.4.6. Multiple breakpoints driven by rare penetrating variants.

We further assessed whether the rare variants with highly penetrating effects on mCAs in *MPL*, *NBN*, *MRE11* and *CTU2* drive multiple clones. We compared numbers of subjects carrying a risk allele between those who had single clone and multiple clones spanning the variant. We used the same number of clones (mCAs) as the cis genetic association study, namely, CN-LOH spanning the variant and unclassified events satisfying the conditions (likely CN-LOH) spanning the variant. As a result, we found statistically significant evidence that the risk haplotypes in *MRE11* and *MPL* further increase risk of this phenomenon (OR:65 (8-447) and 5.5 (1-21), $p=8.0x10^{-5}$ and 0.027, respectively, Fisher's exact test, Extended Data Table 1). The *NBN* variant showed suggestive evidence of enrichment for multiple clones (only one individual with mCA (CN-LOH) spanning this region had multiple clones and this individual carried the rare variant, p=0.11, OR:Inf (0.22-Inf), Extended Data Table 1). None of the four subjects with multiple clones as 16q CN-LOH spanning *CTU2* region carried the rare variant (p=1, Extended Data Table 1).

In contrast, we did not find significant associations between the rare risk alleles we reported and presence of multiple mCAs in different chromosomes (multiple clones in trans) where we compared numbers of subjects carrying a risk allele between those who had single mCA and multiple mCAs in different chromosomes. We did find a significant association at the common risk variant rs12699483 in *MAD1L1* (p=0.0034, Fisher's exact test, OR:1.07 (95%CI:1.02-1.12)), supporting its trans effects across chromosomes (Supplementary Table 17).

6.5. Trans-association

After evaluating cis-associations, we conducted trans-associations (between chromosome and arm-specific mosaic events and variants outside the chromosome and arm of the mosaic events).

6.6. Candidate analyses of associations between mosaic events and variants associated with MPN, CLL or mLOY.

We combined the 86 variants in the previous study[10] which were associated with MPN, CLL or mLOY in the European population with 4 variants we recently found to be associated with mLOY as the 2nd hit in the known genes[58]. We excluded variants in *TERT*, *DLK1*, *JAK2* and *TCL1A* since these genes were shown to be significantly associated with mosaic events in Table 1. As a result, we analyzed a total of 63 variants. We evaluated whether variants showed associations with loss, CN-LOH, or gain in any chromosomes or any mosaic types in any chromosomes. We set a significance level using Bonferroni's correction. When we found a variant satisfying the significance level, we further analyzed detailed mosaic types and detailed chromosomes (we tested additional 88 phenotypes corresponding to mosaic types with more than 20 carriers).

6.7. Pleiotropic associations of *TERT*

At *TERT*, a SNP previously associated with mosaic *JAK2* V617F mutation[12] associated with 14q CN-LOH events ($p=1.5 \times 10^{-22}$, OR =1.27 (1.21-1.33); Table 1 and Extended Data Fig.5e). Risk alleles at *TERT* have previously been observed to associate with clonal hematopoiesis involving a variety of mosaic mutations[10,12,26]; consistent with this finding, we observed that the *TERT* SNP also exhibited nominal association (after Bonferroni correction) with mosaic 20q- events ($p=1.8 \times 10^{-7}$) and gains on chromosome 15 ($p=0.00011$) (curiously with the opposite allele increasing risk of +15). Candidate variant association tests of previously-reported risk variants for clonal hematopoiesis and hematological malignancies revealed additional, weaker *trans* associations with mCAs that will likely reach genome-wide significance in future studies (Supplementary Tables 25-27).

7. Functional analyses of gene expression in significant variants in *MRE11* and *MPL*

We regarded the *NBN* and *CTU2* variants as likely to be causal since they showed stop-gain and amino acid alteration predicted as deleterious by multiple prediction methods, respectively. We conducted functional analyses for significant variants whose functions were not easily interpreted, namely, the *MRE11* and *MPL* variants. We evaluated alteration of gene expression in risk alleles in contrast to reference alleles by the following methods.

7.1. Vector construction and luciferase reporter assay

The section containing intron in *MRE11* gene with polymorphism and upstream region of *MPL* gene with polymorphism were generated by synthetic oligonucleotides. Annealed oligonucleotides were digested by NheI and HindIII, and linked into the vector pGL4.24 minP vector and pGL4.11 basic vector (Promega, Madison, WI), respectively. As the result, pGL4minP-G and pGL4minP-A which encodes the C/T variant located on 94160189 in *MRE11* gene intron (chr11), and pGL4basic-G and pGL4basic-A which encodes the G/A variant located on 43799207 in upstream of *MPL* gene (chr1) were constructed. These vectors were transformed into Escherichia coli strain DH5α and recovered using the EndoFree Plasmid Maxi Kit (Qiagen, West Sussex, UK). The presences of polymorphisms were verified by sequencing. The recombinant plasmids were used for luciferase assay with pGL4minP and pGL4basic as mock vectors. The Jurkat E6.1 cell line and THP-1 cell line (purchased from ATCC) were maintained in RPMI1640 with 10%FBS at 37°C in a 5% CO_2 incubator. Transfection of Jurkat cells with the reporter vectors carrying each variants was carried out using Neon® Transfection System (Invitrogen, Carlsbad, CA). 0.2 Million cells were resuspended in 100 µl of electroporation buffer R that contained 0.5µg of pGL4.74 (Renilla luciferase-TK control reporter vector, Promega) and 2.5 µg of each test vector with polymorphism or empty vector for control. The procedure was conducted according to the manufacturer's protocol with electroporation options recommended for each cell line (three 10 ms, 1350 mV pulses for Jurkat E6.1 and THP-1). Luciferase activity was measured 24 h after transfection using Dual-Luciferase Reporter Assay System (Promega) according to the manufacturer's protocol. Luminescence was detected by TriStar LB941 (Berthold, Bad Wildbad, Germany). Information of oligo nucleotide sequence used for the current study was provided in Supplementary Table 28.

The luciferase assay revealed that risk variants of 1:43799207 and 11:94160189 at the *MPL* and *MRE11*, respectively, were associated with slight decrease and increase in gene expressions, respectively (Fig. S7.2). 1:43799207 is a lead SNP in the *MPL* region.

7.2. Electrophoretic mobility shift assay (EMSA)

The following protocol has been used to make nuclear extracts from Jurkat E6-1.1x 107 Jurkat E6-1, washed with PBS and used for nuclear extraction as described previously[59].Two double-stranded 51-nucleotide biotin-labeled DNA probes were prepared by annealing (Supplementary Table 28). EMSA experiments were carried out using the Lightshift Chemiluminescent EMSA kit (Thermo Fisher), as recommended by the supplier. In brief, 2 µL Binding buffer was mixed with 6.4 µg nuclear extract, then 50 fmol biotin-labeled probe was added, and hybridization was carried out for 20 min at room temperature. The mixtures were then loaded into a 6% Polyacrylamide gel, separated by electrophoresis at 4°C, and transferred onto a nylon membrane. As competitors, non-labeled oligo nucleotides were incubated with nuclear extracts before adding the labeled probe.

EMSA assay suggested binding of transcription factor with 11:94160189 at *MRE11* (Fig. S7.2).

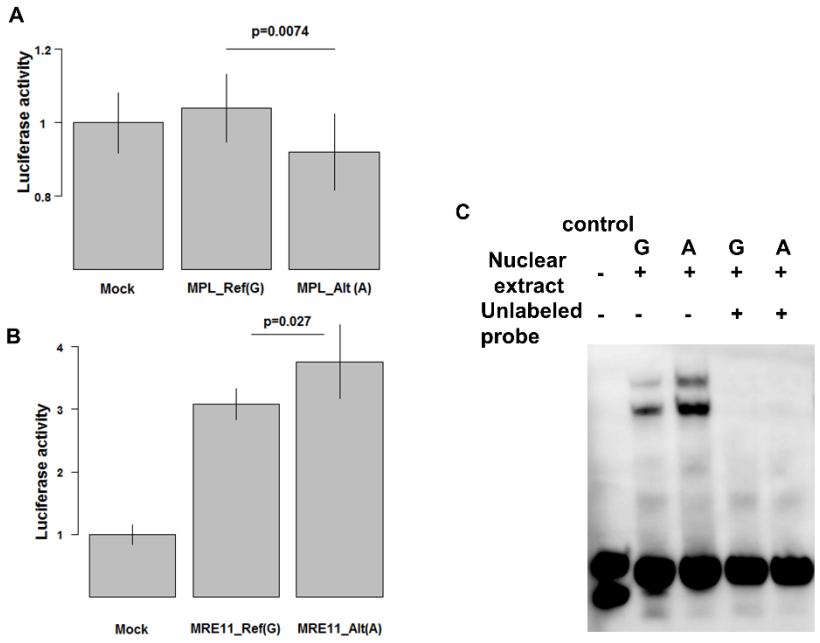

Fig. S7.2. Functional analyses suggesting alteration of gene expression in subjects carrying risk variants of *MPL* and *MRE11*.

A.The *MPL* variant is suggestively associated with decreased expression of *MPL*.

A result of luciferase assay is indicated with the use of synthesized oligonucleotide centering the associated *MPL* variant.

B and C. *MRE11* variant is suggested to be associated with increased expression of *MRE11*.

A result of luciferase assay is indicated with the use of synthesized oligonucleotide centering the associated *MRE11* variant in B. A result of EMSA is indicated in C.

However, previous eQTL studies revealed common variants with much stronger effects on *MRE11*. For instance, the GTEx project reported rs509744 with minor allele frequency of 0.30 associated with alteration of *MRE11* in whole blood ($p=1.0x10^{-30}$ for 369 subjects).

Considering the high penetrance of the associations between the risk variants and CN-LOH, the causal relationship between alteration of the gene expression via these rare variants and mechanism underlying CN-LOH is inconclusive. Alternatively, it may be more likely that rare causal variants introducing functional impairment of proteins are present but not well imputed in the current study.

8. Associations between mCAs and death of subtypes of leukemia in Japanese

Since we observed strong associations between presence of mCAs and death of leukemia, we further analyzed in details associations between mCAs and subgroups of leukemia. We divided leukemia into two groups, myeloid leukemia and lymphoid leukemia. Death of D46.9 MDS, C92.0 AML, C92.1 CML, C92.4 PML and D47.1 CML were classified as myeloid leukemia. Death of C91.5 ATL, C91.0 ALL and C91.1 CLL were classified as lymphoid leukemia. There were leukemia deaths which could not be classified (C95.0 and C95.9). We further divided death of lymphoid leukemia into T cell lymphoid leukemia (C91.5 ATL) and B cell lymphoid leukemia (C91.0 ALL and C91.1 CLL).

References

1 Jacobs, K. B. *et al.* Detectable clonal mosaicism and its relationship to aging and cancer. *Nat. Genet.* **44**, 651-658, doi:10.1038/ng.2270 (2012).

2 Laurie, C. C. *et al.* Detectable clonal mosaicism from birth to old age and its relationship to cancer. *Nat. Genet.* **44**, 642-650, doi:10.1038/ng.2271 (2012).

3 Genovese, G. *et al.* Clonal hematopoiesis and blood-cancer risk inferred from blood DNA sequence. *N. Engl. J. Med.* **371**, 2477-2487, doi:10.1056/NEJMoa1409405 (2014).

4 Jaiswal, S. *et al.* Age-related clonal hematopoiesis associated with adverse outcomes. *N. Engl. J. Med.* **371**, 2488-2498, doi:10.1056/NEJMoa1408617 (2014).

5 Machiela, M. J. *et al.* Characterization of large structural genetic mosaicism in human autosomes. *Am. J. Hum. Genet.* **96**, 487-497, doi:10.1016/j.ajhg.2015.01.011 (2015).

6 Vattathil, S. & Scheet, P. Extensive Hidden Genomic Mosaicism Revealed in Normal Tissue. *Am. J. Hum. Genet.* **98**, 571-578, doi:10.1016/j.ajhg.2016.02.003 (2016).

7 Young, A. L., Challen, G. A., Birmann, B. M. & Druley, T. E. Clonal haematopoiesis harbouring AML-associated mutations is ubiquitous in healthy adults. *Nature communications* **7**, 12484, doi:10.1038/ncomms12484 (2016).

8 Forsberg, L. A., Gisselsson, D. & Dumanski, J. P. Mosaicism in health and disease - clones picking up speed. *Nat Rev Genet* **18**, 128-142, doi:10.1038/nrg.2016.145 (2017).

9 Abelson, S. *et al.* Prediction of acute myeloid leukaemia risk in healthy individuals. *Nature* **559**, 400-404, doi:10.1038/s41586-018-0317-6 (2018).

10 Loh, P. R. *et al.* Insights into clonal haematopoiesis from 8,342 mosaic chromosomal alterations. *Nature* **559**, 350-355, doi:10.1038/s41586-018-0321-x (2018).

11 Zhou, W. *et al.* Mosaic loss of chromosome Y is associated with common variation near TCL1A. *Nat. Genet.* **48**, 563-568, doi:10.1038/ng.3545 (2016).

12 Hinds, D. A. *et al.* Germ line variants predispose to both JAK2 V617F clonal hematopoiesis and myeloproliferative neoplasms. *Blood* **128**, 1121-1128, doi:10.1182/blood-2015-06-652941 (2016).

13 Wright, D. J. *et al.* Genetic variants associated with mosaic Y chromosome loss highlight cell cycle genes and overlap with cancer susceptibility. *Nat. Genet.* **49**, 674-679, doi:10.1038/ng.3821 (2017).

14 Nagai, A. *et al.* Overview of the BioBank Japan Project: Study design and profile. *J. Epidemiol.* **27**, S2-S8, doi:10.1016/j.je.2016.12.005 (2017).

15 Loh, P.-R., Genovese, G. & McCarroll, S. A. Monogenic and polygenic inherited causes of clonal hematopoiesis. *bioRxiv* (2019).

16 Sudlow, C. *et al.* UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Med.* **12**, e1001779, doi:10.1371/journal.pmed.1001779 (2015).

17 Bycroft, C. *et al.* The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203-209, doi:10.1038/s41586-018-0579-z (2018).

18 Iwanaga, M., Watanabe, T. & Yamaguchi, K. Adult T-cell leukemia: a review of epidemiological evidence. *Front Microbiol* **3**, 322, doi:10.3389/fmicb.2012.00322 (2012).

19 Tamura, K. *et al.* Chronic lymphocytic leukemia (CLL) is rare, but the proportion of T-CLL is high in Japan. *Eur. J. Haematol.* **67**, 152-157 (2001).

20 Li, Y., Wang, Y., Wang, Z., Yi, D. & Ma, S. Racial differences in three major NHL subtypes: descriptive epidemiology. *Cancer Epidemiol.* **39**, 8-13, doi:10.1016/j.canep.2014.12.001 (2015).

21 Landau, D. A. *et al.* Mutations driving CLL and their evolution in progression and relapse. *Nature* **526**, 525-530, doi:10.1038/nature15395 (2015).

22 Puente, X. S. *et al.* Non-coding recurrent mutations in chronic lymphocytic leukaemia. *Nature* **526**, 519-524, doi:10.1038/nature14666 (2015).

23 Iwai, M. *et al.* Expression and methylation status of the FHIT gene in acute myeloid leukemia and myelodysplastic syndrome. *Leukemia* **19**, 1367-1375, doi:10.1038/sj.leu.2403805 (2005).

24 Schmitz, R. *et al.* TNFAIP3 (A20) is a tumor suppressor gene in Hodgkin lymphoma and primary mediastinal B cell lymphoma. *J. Exp. Med.* **206**, 981-989, doi:10.1084/jem.20090528 (2009).

25 Liu, Y. *et al.* The genomic landscape of pediatric and young adult T-lineage acute lymphoblastic leukemia. *Nat. Genet.* **49**, 1211-1218, doi:10.1038/ng.3909 (2017).

26 Zink, F. *et al.* Clonal hematopoiesis, with and without candidate driver mutations, is common in the elderly. *Blood* **130**, 742-752, doi:10.1182/blood-2017-02-769869 (2017).

27 Knudson, A. G., Jr. Mutation and cancer: statistical study of retinoblastoma. *Proc. Natl. Acad. Sci. U. S. A.* **68**, 820-823, doi:10.1073/pnas.68.4.820 (1971).

28 Genomes Project, C. *et al.* A global reference for human genetic variation. *Nature* **526**, 68-74, doi:10.1038/nature15393 (2015).

29 Okada, Y. *et al.* Deep whole-genome sequencing reveals recent selection signatures linked to evolution and disease risk of Japanese. *Nature communications* **9**, 1631, doi:10.1038/s41467-018-03274-0 (2018).

30 Kilpivaara, O. *et al.* A germline JAK2 SNP is associated with predisposition to the development of JAK2(V617F)-positive myeloproliferative neoplasms. *Nat. Genet.* **41**, 455-459, doi:10.1038/ng.342 (2009).

31 Jones, A. V. *et al.* JAK2 haplotype is a major risk factor for the development of myeloproliferative neoplasms. *Nat. Genet.* **41**, 446-449, doi:10.1038/ng.334 (2009).

32 Olcaydu, D. *et al.* A common JAK2 haplotype confers susceptibility to myeloproliferative neoplasms. *Nat. Genet.* **41**, 450-454, doi:10.1038/ng.341 (2009).

33 Lee, J. H. & Paull, T. T. ATM activation by DNA double-strand breaks through the Mre11-Rad50-Nbs1 complex. *Science* **308**, 551-554, doi:10.1126/science.1108297 (2005).

34 Dewez, M. *et al.* The conserved Wobble uridine tRNA thiolase Ctu1-Ctu2 is required to maintain genome integrity. *Proc. Natl. Acad. Sci. U. S. A.* **105**, 5459-5464, doi:10.1073/pnas.0709404105 (2008).

35 Adzhubei, I., Jordan, D. M. & Sunyaev, S. R. Predicting functional effect of human missense mutations using PolyPhen-2. *Curr Protoc Hum Genet* **Chapter 7**, Unit7 20, doi:10.1002/0471142905.hg0720s76 (2013).

36 Sim, N. L. *et al.* SIFT web server: predicting effects of amino acid substitutions on proteins. *Nucleic Acids Res.* **40**, W452-457, doi:10.1093/nar/gks539 (2012).

37 DeAntoni, A., Sala, V. & Musacchio, A. Explaining the oligomerization properties of the spindle assembly checkpoint protein Mad2. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **360**, 637-647, discussion 447-638, doi:10.1098/rstb.2004.1618 (2005).

38 Karczewski, K. J. *et al.* Variation across 141,456 human exomes and genomes reveals the spectrum of loss-of-function intolerance across human protein-coding genes. *bioRxiv* (2019).

39 Jaiswal, S. *et al.* Clonal Hematopoiesis and Risk of Atherosclerotic Cardiovascular Disease. *N. Engl. J. Med.* **377**, 111-121, doi:10.1056/NEJMoa1701719 (2017).

40 Hirata, M. *et al.* Overview of BioBank Japan follow-up data in 32 diseases. *J. Epidemiol.* **27**, S22-S28, doi:10.1016/j.je.2016.12.006 (2017).

41 Consortium, G. P. A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061-1073, doi:10.1038/nature09534 (2010).

42 Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 559-575, doi:10.1086/519795 (2007).

43 Staaf, J. *et al.* Normalization of Illumina Infinium whole-genome SNP data improves copy number estimates and allelic intensity ratios. *BMC Bioinformatics* **9**, 409, doi:10.1186/1471-2105-9-409 (2008).

44 Loh, P. R. *et al.* Reference-based phasing using the Haplotype Reference Consortium panel. *Nat. Genet.* **48**, 1443-1448, doi:10.1038/ng.3679 (2016).

45 Kanai, M. *et al.* Genetic analysis of quantitative traits in the Japanese population links cell types to complex human diseases. *Nat. Genet.* **50**, 390-400, doi:10.1038/s41588-018-0047-6 (2018).

46 Loh, P. R., Palamara, P. F. & Price, A. L. Fast and accurate long-range phasing in a UK Biobank cohort. *Nat. Genet.* **48**, 811-816, doi:10.1038/ng.3571 (2016).

47 Das, S. *et al.* Next-generation genotype imputation service and methods. *Nat. Genet.* **48**, 1284-1287, doi:10.1038/ng.3656 (2016).

48 Ma, C., Blackwell, T., Boehnke, M., Scott, L. J. & Go, T. D. i. Recommended joint and meta-analysis strategies for case-control association testing of single low-count variants. *Genet. Epidemiol.* **37**, 539-550, doi:10.1002/gepi.21742 (2013).

49 Bock, C., Walter, J., Paulsen, M. & Lengauer, T. CpG island mapping by epigenome prediction. *PLoS Comput. Biol.* **3**, e110, doi:10.1371/journal.pcbi.0030110 (2007).

50 Hoggart, C. J., Clark, T. G., De Iorio, M., Whittaker, J. C. & Balding, D. J. Genome-wide significance for dense SNP and resequencing data. *Genet. Epidemiol.* **32**, 179-185, doi:10.1002/gepi.20292 (2008).

51 Kagami, M. *et al.* Deletions and epimutations affecting the human 14q32.2 imprinted region in individuals with paternal and maternal upd(14)-like phenotypes. *Nat. Genet.* **40**, 237-242, doi:10.1038/ng.2007.56 (2008).

52 Miyazato, A. *et al.* Identification of myelodysplastic syndrome-specific genes by DNA microarray analysis with purified hematopoietic stem cell fraction. *Blood* **98**, 422-427 (2001).

53 Falix, F. A., Aronson, D. C., Lamers, W. H. & Gaemers, I. C. Possible roles of DLK1 in the Notch pathway during development and disease. *Biochim. Biophys. Acta* **1822**, 988-995, doi:10.1016/j.bbadis.2012.02.003 (2012).

54 Soucy, T. A., Dick, L. R., Smith, P. G., Milhollen, M. A. & Brownell, J. E. The NEDD8 Conjugation Pathway and Its Relevance in Cancer Biology and Therapy. *Genes Cancer* **1**, 708-716, doi:10.1177/1947601910382898 (2010).

55 Swords, R. T. *et al.* Pevonedistat, a first-in-class NEDD8-activating enzyme inhibitor, combined with azacitidine in patients with AML. *Blood* **131**, 1415-1424, doi:10.1182/blood-2017-09-805895 (2018).

56 Jones, M. *et al.* The shelterin complex and hematopoiesis. *J. Clin. Invest.* **126**, 1621-1629, doi:10.1172/JCI84547 (2016).

57 Paduano, F. *et al.* T-Cell Leukemia/Lymphoma 1 (TCL1): An Oncogene Regulating Multiple Signaling Pathways. *Front. Oncol.* **8**, 317, doi:10.3389/fonc.2018.00317 (2018).

58 Terao, C. *et al.* GWAS of mosaic loss of chromosome Y highlights genetic effects on blood cell differentiation. *Nature communications* **10**, 4719, doi:10.1038/s41467-019-12705-5 (2019).

59 Satoh, S. *et al.* AXIN1 mutations in hepatocellular carcinomas, and growth suppression in cancer cells by virus-mediated transfer of AXIN1. *Nat. Genet.* **24**, 245-250, doi:10.1038/73448 (2000).

Supplementary Tables

Supplementary Table 1. The 179,417 subjects used to call mosaic events in the current study.

	Set1	Set2	Set3
Samples	34,256	111,422	33,739
Arrays	OmniExpressExome v1.0	OmniExpressExome v1.2	OmniExpress v1.0 and HumanExome v1.0, 1.1
Age	69.4+/-10.3	61.6+/-14.8	59.9+/-15.4
Female ratio	0.46	0.46	0.45

Supplementary Table 2. Chromosomal distribution of classified mosaic events.

CHR	LOSS	p-arm LOSS	q-arm LOSS	CN-LOH	p-arm CN-LOH	q-arm CN-LOH	GAIN
chr1	156	123	28	1489	813	652	221
chr2	293	219	69	388	163	210	52
chr3	175	125	47	270	158	105	91
chr4	212	34	164	260	16	205	24
chr5	460	12	443	102	7	93	86
chr6	413	22	377	485	380	100	63
chr7	263	70	179	152	33	104	43
chr8	95	50	44	92	20	67	370
chr9	169	19	136	690	317	366	198
chr10	56	22	31	163	28	134	21
chr11	466	35	428	907	362	535	23
chr12	123	51	68	254	33	218	122
chr13	453	0	451	224	0	157	13
chr14	335	0	332	2688	0	2478	108
chr15	48	0	47	225	0	199	886
chr16	92	57	29	555	267	284	9
chr17	177	95	80	567	133	428	95
chr18	50	32	15	99	22	75	195
chr19	12	5	6	217	112	105	16
chr20	1029	16	984	332	20	304	8
chr21	73	0	73	57	0	37	1098
chr22	83	0	83	215	0	186	302
Sum	5233	987	4114	10431	2884	7042	4044

Since p and q indicates whether an event extends to a telomere, the sum of p and q mosaics is not always the same as LOSS or CN-LOH in the corresponding chromosomes.

Supplementary Table 3. Number of mosaic events in each individual.

Mosaic N	Individual
0	151507
1	23754
2	3376
3	554
4	144
5	39
6	23
7	7
8	3
9	6
10	1
11	2
14	1

Supplementary Table 4. Mosaic detection rate across batches

	N	P	OR (95%CI)
Set1	32,959	0.97	1.00 (0.95-1.05)
Set2	107,767	Reference	1
Set3	32,873	2.0×10^{-13}	0.82 (0.77-0.86)

OR: odds ratio, CI: confidence interval

OR was calculated by logistic regression with mosaic detection as a dependent variable and age, sex, smoking, arrays, top 10 PCs and disease status at registry as independent variables. Since we conducted association studies for 173,599 subjects after exclusion of samples based on kinship and evidence of contamination, the number of subjects in the three sets were slightly different from those in Supplementary Table 1.

Supplementary Table 5. Associations between disease status and mosaic events

Disease group	Disease	N	P	OR (95%CI)
Malignant tumors	Lung cancer	3397	0.91	0.99 (0.91-1.09)
	Esophageal cancer	1141	0.93	1.01 (0.86-1.19)
	Gastric cancer	5613	$1.3x10^{-5}$	1.17 (1.09-1.26)
	Colorectal cancer	5909	0.016	1.09 (1.02-1.17)
	Liver cancer	1334	0.24	1.09 (0.94-1.25)
	Pancreas cancer	279	0.80	1.04 (0.75-1.44)
	Gallbladder/Cholangiocarcinoma	224	0.71	0.93 (0.65-1.34)
	Prostate cancer	4519	0.53	0.98 (0.90-1.06)
	Breast cancer	4933	0.68	0.98 (0.89-1.08)
	Cervical cancer	546	0.12	1.25 (0.94-1.65)
	Uterine cancer	917	0.25	0.87 (0.69-1.10)
	Ovarian cancer	663	0.88	1.02 (0.78-1.34)
	Hematopoietic tumor	1075	$2.1x10^{-17}$	1.93 (1.66-2.25)
Cerabral diseases	Cerebral infarction	14862	0.65	1.01 (0.97-1.06)
	Cerebral aneurysm	2473	0.0048	0.83 (0.73-0.95)
	Epilepsy	1908	0.51	0.93 (0.78-1.09)
Respiratory diseases	Bronchial asthma	7304	0.29	0.96 (0.89-1.03)
	Pulmonary tuberculosis	386	0.20	1.17 (0.92-1.50)
	Chronic obstructive pulmonary disease	2525	0.47	1.03 (0.94-1.14)
	Interstitial lung disease/Pulmonary fibrosis	450	0.74	0.96 (0.76-1.23)
Cardiovascular diseases	Myocardial infarction	11881	0.015	0.94 (0.89-0.99)
	Unstable angina	3875	0.41	0.97 (0.89-1.05)
	Stable angina	13402	0.18	1.03 (0.99-1.08)
	Arrhythmia	14447	0.79	0.99 (0.95-1.04)
	Heart failure	6799	0.44	0.97 (0.91-1.04)
	Peripheral arterial diseases	2424	0.0027	1.16 (1.05-1.28)
Liver diseases	Chronic hepatitis B	1197	0.029	1.21 (1.02-1.43)

	Chronic hepatitis C	5144	0.38	1.04 (0.96-1.12)
	Liver cirrhosis	1413	0.86	1.01 (0.88-1.17)
Urologic diseases	Nephrotic syndrome	863	0.98	1.0 (0.80-1.26)
	Urolithiasis	5800	0.0011	0.87 (0.79-0.94)
Metabolic diseases	Osteoporosis	5906	0.060	0.93 (0.87-1.0)
	Diabetes mellitus	35856	0.075	0.97 (0.94-1.0)
	Dyslipidemia	39459	0.88	1.0 (0.97-1.04)
Endocrine diseases	Graves' disease	1973	$2.6x10^{-7}$	0.58 (0.48-0.72)
Connective tissue diseases	Rheumatoid arthritis	3777	$8.9x10^{-5}$	0.81 (0.73-0.90)
Allergic diseases	Hay fever	5062	0.096	0.91 (0.82-1.02)
Dermatologic diseases	Drug eruption	213	0.18	1.28 (0.9-1.82)
	Atopic dermatitis	2507	0.46	1.07 (0.89-1.3)
	Keloid	735	0.31	1.14 (0.89-1.46)
Gynecologic diseases	Uterine fibroid	5473	0.28	1.06 (0.95-1.19)
	Endometriosis	656	0.6	0.91 (0.64-1.29)
Pediatric diseases	Febrile seizure	15	0.83	0 (0-Inf)
Ophthalmologic diseases	Glaucoma	4205	0.46	0.97 (0.89-1.05)
	Cataract	18010	0.29	1.03 (0.98-1.08)
Dental diseases	Periodontitis	2773	0.26	1.07 (0.95-1.20)
Other	Amyotrophic lateral sclerosis	10	0.31	2.03 (0.52-8.02)

OR: odds ratio, CI: confidence interval

Results of logistic regression analysis with presence of mosaic events as a dependent variable and disease status at registry, age, sex, smoking, genotyping array and top 10 PCs as independent variables. Significance level was set at 0.05/1,034 ($p<4.8x10^{-5}$)

Supplementary Table 6. Fraction of mosaic presence according to age and sex

Age range	% of males with autosomal event (s.e.)	% of females with autosomal event (s.e.)
<30	4.6% (0.4%)	4.2% (0.4%)
30-39	5.7% (0.4%)	4.7% (0.3%)
40-49	7.6% (0.3%)	6.5% (0.3%)
50-59	11.4% (0.2%)	8.9% (0.2%)
60-69	16.3% (0.2%)	12.5% (0.2%)
70-79	24.7% (0.3%)	17.6% (0.3%)
80-89	31.8% (0.6%)	24.1% (0.5%)
90+	40.7% (2.3%)	31.5% (1.7%)

This table contains numeric data plotted in Fig. 2b. s.e.:standard error

Supplementary Table 7. Average age and sex of carriers of mosaic event types

	p-arm LOSS		q-arm LOSS		p-arm CN-LOH		q-arm CN-LOH		GAIN	
CHR	AGE	Male ratio	AGE	Male ratio	AGE	Male ratio	AGE	Male ratio	AGE	Male ratio
1	72.7 (0.07)	0.64 (0.04)	72.9 (0.3)	0.61 (0.09)	68.6 (0.01)	0.6 (0.01)	69.3 (0.01)	0.62 (0.01)	67.5 (0.06)	0.62 (0.03)
2	71.2 (0.04)	0.52 (0.03)	69.3 (0.13)	0.61 (0.06)	69.2 (0.05)	0.6 (0.03)	66.6 (0.04)	0.55 (0.03)	70.3 (0.24)	0.42 (0.07)
3	73.6 (0.08)	0.73 (0.04)	67.8 (0.22)	0.55 (0.07)	67.7 (0.04)	0.54 (0.03)	69.7 (0.06)	0.55 (0.03)	70.1 (0.11)	0.68 (0.05)
4	66.7 (0.4)	0.59 (0.08)	71.3 (0.07)	0.64 (0.04)	66.8 (0.16)	0.64 (0.05)	71.9 (0.04)	0.62 (0.03)	69.8 (0.43)	0.83 (0.08)
5	65.2 (0.86)	0.67 (0.14)	72.1 (0.02)	0.53 (0.02)	68.2 (0.13)	0.67 (0.05)	67.4 (0.05)	0.61 (0.03)	68.5 (0.13)	0.72 (0.05)
6	73 (0.48)	0.64 (0.1)	69.8 (0.03)	0.68 (0.02)	67.3 (0.02)	0.6 (0.02)	67.1 (0.04)	0.64 (0.03)	68 (0.17)	0.74 (0.05)
7	65 (0.18)	0.66 (0.06)	70.7 (0.06)	0.61 (0.04)	66.4 (0.08)	0.57 (0.04)	66.5 (0.05)	0.57 (0.03)	65.8 (0.34)	0.53 (0.08)
8	67.7 (0.28)	0.63 (0.07)	70.4 (0.24)	0.57 (0.07)	66.3 (0.08)	0.59 (0.04)	69.8 (0.05)	0.61 (0.03)	74 (0.03)	0.63 (0.02)
9	73.3 (0.51)	0.63 (0.11)	71.7 (0.07)	0.7 (0.04)	68.2 (0.03)	0.64 (0.02)	68.9 (0.01)	0.59 (0.02)	71.5 (0.05)	0.67 (0.03)
10	69.8 (0.65)	0.59 (0.1)	69 (0.4)	0.61 (0.09)	66.1 (0.12)	0.6 (0.05)	68.6 (0.05)	0.61 (0.03)	63.6 (0.71)	0.48 (0.11)
11	67.3 (0.31)	0.57 (0.08)	71.3 (0.02)	0.63 (0.02)	65.8 (0.02)	0.6 (0.02)	69.9 (0.01)	0.67 (0.01)	64 (0.58)	0.65 (0.09)
12	71.8 (0.21)	0.71 (0.06)	69 (0.16)	0.63 (0.06)	68.5 (0.14)	0.61 (0.05)	67.4 (0.04)	0.62 (0.02)	71.4 (0.07)	0.54 (0.05)
13			71 (0.02)	0.65 (0.02)			68.2 (0.06)	0.56 (0.03)	64.9 (1.35)	0.77 (0.12)
14			68.7 (0.04)	0.66 (0.03)			71.8 (0)	0.64 (0.01)	69.7 (0.1)	0.69 (0.04)
15			66.7 (0.29)	0.64 (0.07)			66.9 (0.06)	0.52 (0.03)	74.7 (0.01)	0.77 (0.01)
16	67 (0.23)	0.63 (0.06)	69.1 (0.48)	0.57 (0.09)	68.1 (0.03)	0.58 (0.02)	68.5 (0.02)	0.63 (0.02)	68.4 (1.56)	0.67 (0.16)
17	72.4 (0.11)	0.62 (0.05)	69 (0.14)	0.46 (0.06)	70.5 (0.04)	0.67 (0.03)	67.4 (0.02)	0.56 (0.02)	70.5 (0.11)	0.58 (0.05)
18	71.7 (0.29)	0.66 (0.08)	72 (0.4)	0.53 (0.13)	68.5 (0.15)	0.6 (0.06)	68.1 (0.06)	0.55 (0.03)	69.1 (0.06)	0.59 (0.04)
19	60 (4.4)	1 (0)	64.7 (2.66)	0.33 (0.19)	70.2 (0.05)	0.59 (0.03)	67.9 (0.03)	0.59 (0.03)	63.2 (0.85)	0.63 (0.12)
20	73.9 (0.6)	0.63 (0.12)	72.7 (0.01)	0.73 (0.01)	65.4 (0.09)	0.58 (0.04)	68 (0.02)	0.61 (0.02)	72.8 (1.78)	0.67 (0.16)

21	65.5 (0.16)	0.63 (0.06)	67.4 (0.21)	0.46 (0.07)	68.5 (0.01)	0.66 (0.01)
22	71.7 (0.13)	0.77 (0.05)	69.8 (0.06)	0.65 (0.03)	70.5 (0.03)	0.56 (0.03)

CHR:chromosome, mean (s.e.m) is indicated

Supplementary Table 8. Average age and sex of carriers of focal mosaic deletions.

Focal deletion	Mean AGE (s.e.)	Male ratio (s.e.)
chr2:20-30Mb	71.6 (0.06)	0.52 (0.04)
chr3:55-95Mb	74.7 (0.08)	0.73 (0.04)
chr4:100-110Mb	73.4 (0.17)	0.57 (0.06)
chr6:125-150Mb	69.7 (0.09)	0.69 (0.04)
chr12:7-14Mb (*ETV6*)	73 (0.35)	0.85 (0.07)
chr13:40-60Mb	71.6 (0.07)	0.67 (0.04)
chr14:20-24Mb	68.9 (0.04)	0.65 (0.03)
chr16:24-31Mb	69.7 (0.63)	0.54 (0.14)
chr17:24-31Mb	68.5 (0.22)	0.4 (0.07)
chr22:24-35Mb (*CHEK2*)	72.6 (0.18)	0.81 (0.05)

s.e.:standard error

Supplementary Table 9. Associations between mosaic events and hematopoietic traits.

pheno	Mosaic	Beta	SE	P
RBC	chr15_GAIN	-0.211	0.043	$8.4x10^{-7}$
Plt	chr20q_LOSS	-0.216	0.045	$1.7x10^{-6}$
RBC	chr9_GAIN	-0.413	0.086	$1.7x10^{-6}$
MCV	chr1_GAIN	0.414	0.088	$2.6x10^{-6}$
Lym	chr14q_LOSS	0.491	0.105	$2.9x10^{-6}$
RBC	chr8_GAIN	-0.292	0.063	$4.0x10^{-6}$
Lym	chr21_GAIN	0.261	0.057	$4.0x10^{-6}$
Neutro	chr14q_LOSS	-0.509	0.113	$6.9x10^{-6}$
MCHC	chr9_GAIN	-0.391	0.089	$1.0x10^{-5}$
RBC	chr20q_LOSS	-0.172	0.04	$1.6x10^{-5}$
Lym	chr22_GAIN	0.494	0.117	$2.4x10^{-5}$
Neutro	chr22_GAIN	-0.528	0.127	$3.1x10^{-5}$
MCV	chr9_GAIN	0.392	0.095	$3.7x10^{-5}$

Lym: lymphocyte count, RBC: red blood cell count, MCV: mean corpuscular volume, MCHC: mean corpuscular hemoglobin concentration, Plt: platelet count, Neutro neutrophil count. Significant associations beyond Bonferroni's correction ($p<0.05/13/88$) are indicated. AGE, SEX, smoking, top 10 PCs, array batch and disease status at registry are used as covariates

Supplementary Table 10. Increased lymphocyte counts associated with presence of mosaic of V(D)J deletion in *TRA*.

	TRA del (+) (N=84)	*TRA* del (-) (N=61992)	P
lymphocyte count	2078+/-811	1791+/-703	0.0017

Mean+/-sd (standard deviation) is indicated

Supplementary Table 11. Significant correlation between cell fraction of V(D)J deletion in *TRA* and lymphocyte counts.

Spearman rho (95%CI)	P
0.33 (0.12-0.51)	0.0037

CI: confidence interval

Supplementary Table 12. **Distribution of mCAs by chromosome and copy number in BioBank Japan and UK Biobank.**

Chromsome	Loss freq, BBJ	Loss freq, UKB	CN-LOH freq, BBJ	CN-LOH freq, UKB	Gain freq, BBJ	Gain freq, UKB
1	0.8%	0.7%	7.6%	8.2%	1.1%	0.5%
2	1.5%	1.5%	2.0%	1.8%	0.3%	0.2%
3	0.9%	0.5%	1.4%	1.6%	0.5%	1.2%
4	1.1%	1.0%	1.3%	1.7%	0.1%	0.2%
5	2.3%	1.2%	0.5%	0.9%	0.4%	0.6%
6	2.1%	0.8%	2.5%	2.2%	0.3%	0.2%
7	1.3%	1.2%	0.8%	1.2%	0.2%	0.2%
8	0.5%	0.5%	0.5%	0.9%	1.9%	1.1%
9	0.9%	0.4%	3.5%	4.9%	1.0%	0.8%
10	0.3%	2.0%	0.8%	0.9%	0.1%	0.1%
11	2.4%	2.1%	4.6%	6.1%	0.1%	0.0%
12	0.6%	0.6%	1.3%	1.8%	0.6%	3.7%
13	2.3%	4.2%	1.1%	3.1%	0.1%	0.1%
14	1.7%	1.1%	13.6%	4.9%	0.5%	1.1%
15	0.2%	0.3%	1.1%	2.8%	4.5%	1.6%
16	0.5%	1.3%	2.8%	3.3%	0.0%	0.1%
17	0.9%	1.6%	2.9%	3.0%	0.5%	0.9%
18	0.3%	0.4%	0.5%	0.7%	1.0%	1.4%
19	0.1%	0.1%	1.1%	2.4%	0.1%	0.3%
20	5.2%	3.2%	1.7%	1.4%	0.0%	0.1%
21	0.4%	0.4%	0.3%	1.0%	5.6%	1.1%

22	0.4%	1.0%	1.1%	2.3%	1.5%	1.3%

This table provides numeric data plotted in Fig. 3a,b. Frequencies indicate the contribution of each event type to the total number of mCAs classified as loss, CN-LOH, or gain in each data set. Data for UK Biobank events are from parallel work on the UK Biobank cohort[17].

Supplementary Table 13. Concordance of positional coverage of mosaic events between Japanese and UK population

Chr	All mosaic	Loss	CN-LOH	Gain
1	0.97 (0.97-0.97)	0.41 (0.38-0.45)	0.98 (0.97-0.98)	0.91 (0.9-0.92)
2	0.99 (0.99-0.99)	0.96 (0.96-0.96)	0.97 (0.97-0.97)	0.23 (0.19-0.27)
3	0.8 (0.79-0.82)	0.92 (0.91-0.93)	0.97 (0.97-0.97)	0.98 (0.98-0.98)
4	0.99 (0.99-0.99)	0.72 (0.69-0.74)	0.99 (0.99-0.99)	0.8 (0.78-0.82)
5	0.85 (0.84-0.86)	0.99 (0.99-0.99)	0.96 (0.95-0.96)	0.99 (0.99-1)
6	0.91 (0.91-0.92)	0.86 (0.84-0.87)	0.99 (0.99-0.99)	0.89 (0.88-0.9)
7	0.98 (0.98-0.98)	0.9 (0.9-0.91)	0.97 (0.96-0.97)	0.83 (0.81-0.84)
8	0.88 (0.87-0.89)	0.63 (0.6-0.66)	0.96 (0.95-0.96)	0.95 (0.94-0.95)
9	0.84 (0.82-0.85)	0.97 (0.96-0.97)	0.91 (0.9-0.92)	0.95 (0.94-0.96)
10	0.79 (0.77-0.81)	-0.03 (-0.08-0.03)	0.94 (0.93-0.94)	0.09 (0.04-0.15)
11	0.87 (0.86-0.89)	0.82 (0.8-0.84)	0.8 (0.78-0.82)	-0.21 (-0.26--0.16)
12	0.99 (0.99-0.99)	0.79 (0.77-0.81)	0.98 (0.98-0.98)	0.9 (0.88-0.91)
13	0.87 (0.86-0.89)	0.91 (0.89-0.92)	0.98 (0.98-0.98)	0.86 (0.84-0.87)
14	0.99 (0.99-0.99)	0.13 (0.06-0.19)	0.99 (0.99-0.99)	0.73 (0.7-0.76)
15	0.96 (0.95-0.96)	-0.04 (-0.11-0.03)	0.98 (0.98-0.98)	0.99 (0.99-0.99)
16	0.92 (0.91-0.93)	0.56 (0.51-0.6)	0.93 (0.92-0.94)	0.02 (-0.04-0.09)
17	0.99 (0.99-0.99)	0.99 (0.99-0.99)	1 (1-1)	1 (1-1)
18	0.94 (0.93-0.94)	0.99 (0.98-0.99)	0.98 (0.98-0.98)	0.77 (0.74-0.8)
19	0.91 (0.89-0.92)	0.31 (0.24-0.38)	0.9 (0.89-0.92)	-0.42 (-0.48--0.35)
20	0.99 (0.99-0.99)	0.99 (0.99-0.99)	0.97 (0.96-0.97)	0.7 (0.65-0.73)
21	0.62 (0.55-0.69)	0.16 (0.05-0.26)	0.94 (0.92-0.95)	0.76 (0.71-0.8)
22	0.93 (0.92-0.94)	0.61 (0.54-0.68)	0.97 (0.96-0.98)	0.85 (0.82-0.88)
Mean	0.91±0.09	0.66±0.35	0.96±0.04	0.66±0.42

Pearson's correlation coefficients between BBJ and UKB are indicated.

Mean indicates mean correlation across chromosomes (not concatenating all chromosomes).

Chr: chromosome

Supplementary Table 14. Genes most frequently involved in focal deletions in BBJ.

Chr	Top gene
1	PTPN22, LOC100287722, BCL2L15, AP4B1, DCLRE1B, HIPK1, OLFML3, RPL13AP10, SYT6, MRP63P1, LOC100421116
2	DNMT3A
3	FOXP1
4	TET2
5	FBXL17
6	LOC100130476, TNFAIP3
7	LOC136157, RPS2P31, GPR37, LOC154872, POT1
8	FAM87A, FBXO25, C8orf42, DLGAP2, LOC100130321, LOC100507448, CLN8, MIR3674, MIR596, ARHGEF10, LOC100131395, KBTBD11, MYOM2
9	LOC286370, MIR4290, IL6RP1, OR7E31P, OR7E116P, LOC340515, DIRAS2, OR7E109P, OR7E108P, SYK
10	PTEN, RPL11P3, LOC100128990, VN1R55P, RNLS, LIPJ, RPL7P34
11	DDX10, CYCSP29
12	LOH12CR1, DUSP16
13	DLEU7
14	TRAV24
15	FMN1, LOC100421433, LOC100652815, LOC100652857, RYR3
16	RPL10AP12, IRF8
17	PFN1, ENO3, SPAG7, CAMTA2, INCA1, KIF1C
18	DLGAP1
20	PTPRT
21	COL6A2, FTCD
22	CHEK2, CCDC117, XBP1, ZNRF3

Since chr19 has fewer than 20 loss events, we do not show results in chr 19.

Supplementary Table 15. Genes frequently involved with focal deletions in Japanese and not in UK population.

Chr	Genes
1	LOC100533666,ST13P20,LPHN2,CDK4PS
3	LOC100421672,FHIT,LOC100421670
4	RPL21P46,SCFD2,FIP1L1,LNX1,LOC100129728,RPL21P44,CHIC2,RPL22P13,PDGFRA,LOC100421808,MIR548AG1,LOC100421630,VEGFC,NEIL3,AGA
6	PEX7,SLC35D3,RPL35AP3,NHEG1,IL20RA,IL22RA2,IFNGR1,OLIG3,LOC391040,LOC442263,LOC100507406,LOC100507429,LOC100130476,TNFAIP3,RPSAP42,PERP,KIAA1244,PBOV1,HEBP2,NHSL1,MIR3145
7	NXPH1,RPL9P19,LOC100287551,NDUFA4,DGKB,EEF1A1P26,LOC100533714,VWC2,MAGI2,MAGI2-AS3,RPL10P11,GNAI1,LOC100420647,IMMP2L,LRRN3
8	RPL23AP53,ZNF596,FAM87A,FBXO25,C8orf42,CSMD3,LOC100289099,MIR2053,EXT1
9	LOC100128505
10	CTNNA3,LOC100533794
11	LOC729790
14	LOC100288613,TRA@,TRAV1-1,OR10G2,TRAV1-2,ARL6IP1P1,OR4E2,OR4E1,TRAV2,TRAV3,TRAV4,TRAV5,RPL4P1,TRAV6,TRAV7,TRAV8-1,TRAV9-1,TRAV10,TRAV11,TRAV12-1,TRAV8-2,TRAV8-3,TRAV13-1,TRAV12-2,TRAV8-4,TRAV8-5,TRAV13-2,TRAV14DV4,TRAV9-2,TRAV15,TRAV12-3,TRAV8-6,TRAV16,TRAV17,TRAV18,TRAV19,TRAV20,TRAV21,TRAV22,TRAV23DV6,TRDV1,TRAV24,TRAV25,TRAV26-1,TRAV8-7,TRAV27,TRAV28,TRAV29DV5,TRAV30,TRAV31,TRAV32,TRAV33,TRAV26-2,TRAV34,TRAV35,TRAV36DV7,TRAV37,TRAV38-1,TRAV38-2DV8,TRAV39,TRAV40,TRAV41,TRD@,TRDV2,TRDD1,TRDD2,TRDD3,TRDJ1,TRDJ4,TRDJ2,TRDJ3
15	KIAA1370,LINC00052,NTRK3
16	LOC100131080
18	LOC100422496
22	LARGE,MIR4764,LOC100506195

Supplementary Table 16. No associations between chr5q CN-LOH and variants in *RAD50*.

Chr:Pos	Ref	Alt	Case freq	Cont freq	P	OR (95%CI)
5:131954134	A	G	0.0065	0.00078	0.0062	8.3 (2.7-26)
5:131977046	C	T	0.067	0.041	0.0095	1.7 (1.2-2.4)
5:131874403	T	G	0.011	0.0028	0.010	3.9 (1.6-9.5)
5:131916654	G	A	0.0065	0.0011	0.015	6 (1.9-18.8)
5:131991821	C	A	0.011	0.0035	0.024	3.1 (1.3-7.6)
5:131875296	G	C	0.011	0.0035	0.025	3.1 (1.3-7.5)
5:131940799	T	C	0.0022	$5.5x10^{-5}$	0.027	39.2 (5.2-293.8)
5:131891706	G	A	0.0065	0.0014	0.030	4.6 (1.5-14.3)
5:131962216	A	G	0.0022	$6.7x10^{-5}$	0.032	32.4 (4.4-240.6)
5:131893543	T	C	0.017	0.0079	0.034	2.2 (1.1-4.4)
5:131926388	G	A	0.0065	0.0016	0.040	4 (1.3-12.6)
5:131993276	T	TCTGAA	0.0022	$9.6x10^{-5}$	0.045	22.6 (3.1-165.5)

Variants showing p-values less than 0.05 in the region of *RAD50* are indicated.

Chr: chromosome, Pos: position, Freq: frequency, OR: odds ratio, CI: confidence interval

Supplementary Table 17. Associations of *MAD1L1* variant rs12699483 with gain events.

Chr	Events	N	P	OR (95%CI)
chr1		5	0.34	2.05 (0.58-7.25)
chr2		6	0.38	1.91 (0.61-6.03)
chr3		**59**	**0.025**	**1.51 (1.05-2.17)**
chr4		18	0.027	2.15 (1.1-4.2)
chr5		3	0.7	1.37 (0.28-6.77)
chr6		9	0.34	1.71 (0.67-4.33)
chr7		9	0.63	1.37 (0.54-3.44)
chr8		**320**	**0.15**	**1.13 (0.96-1.31)**
chr9		**129**	**0.05**	**1.29 (1.01-1.64)**
chr10		4	1	0.82 (0.2-3.43)
chr11		5	0.11	3.19 (0.82-12.33)
chr12		**105**	**0.4**	**1.13 (0.86-1.48)**
chr13		9	0.24	0.53 (0.19-1.47)
chr14		**96**	**0.12**	**1.26 (0.95-1.67)**
chr15		**855**	**$4.3x10^{-22}$**	**1.6 (1.46-1.76)**
chr16		2	1	1.37 (0.19-9.7)
chr17		**71**	**0.87**	**1.03 (0.74-1.43)**
chr18		**182**	**0.96**	**0.99 (0.8-1.22)**
chr19		8	0.13	2.28 (0.83-6.27)
chr20		0	1	NA
chr21		**1060**	**0.085**	**1.08 (0.99-1.18)**
chr22		**264**	**0.17**	**1.13 (0.95-1.34)**

Gain events covering >50% of the genotyped span of the chromosome were tested for association. Events occurring in at least 20 individuals are indicated in bold.

OR: odds ratio, CI: confidence interval

Supplementary Table 18. Allele frequencies of BBJ mosaicism risk variants in European population

Mosaic	Gene	Chr	Pos	Variant	Ref	Var	Jpn AF	Eur AF
Cis								
chr1p_CN-LOH	*MPL*	1	45444734	rs560932816	G	A	0.00016	0.00033
		1	44074454	rs190159566	C	T	0.0049	0
		1	47704269	rs556241419	G	A	0.000062	0
		1	44579360	rs184778092	C	T	0.00005	0
chr8q_CN-LOH	*NBN*	8	90949282	rs756831345	C	A	0.00061	0
chr9p_CN-LOH	*JAK2*	9	5026293	rs2183137	A	G	0.24	0.29
chr11q_CN-LOH	*MRE11*	11	94160189	11:94160189	G	A	0.00011	0
chr14q_CN-LOH	*NEDD8/ TINF2*	14	24711798	rs28372734	C	G	0.073	0.0020
chr14q_CN-LOH	*TCL1A*	14	96180242	rs1122138	C	A	0.05	0.15
chr14q_CN-LOH	*DLK1*	14	101175967	rs10873520	G	A	0.30	0.20
chr16q_CN-LOH	*CTU2*	16	88781475	rs200779411	C	T	0.00065	0.00014
trans								
chr14q_CN-LOH	*TERT*	5	1287194	rs2853677	A	G	0.31	0.41
Chr15_gain	*MAD1L1*	7	1975624	rs12699483	C	G	0.42	0.35

Chr:chromosome, Pos: chromosomal base pair position, Ref: reference allele, Var: variant (tested) allele, Jpn AF: Japanese allele frequency calculated by control subjects, Eur MAF: European (non-Finnish) allele frequency obtained from gnomAD(v.2.1.1).

Supplementary Table 19. Associations between mosaic events and mortality.

	HR (95%CI)	P
Overall mortality	1.10 (1.05-1.16)	$2.7x10^{-5}$
All cancer mortality	1.13 (1.03-1.25)	0.014
Blood cancer mortality	2.85 (2.15-3.78)	$4.1x10^{-13}$
Leukemia mortality	4.70 (3.26-6.78)	$1.0x10^{-16}$
Malignant lymphoma mortality	1.39 (0.78-2.47)	0.26
Multiple myeloma mortality	0.87 (0.26-2.88)	0.82
Other cancer mortality	1.04 (0.94-1.16)	0.46
Cardiovascular mortality	1.07 (0.94-1.22)	0.32
Coronary heart disease mortality	1.10 (0.93-1.30)	0.27
Ischemic stroke mortality	1.03 (0.83-1.27)	0.79

CI:confidence interval, HR: hazard ratio

Supplementary Table 20. Associations between mosaic events and leukemia mortality

	pLOSS		qLOSS		pCN-LOH		qCN-LOH		GAIN	
	P	OR (95%CI)	P	OR (95%CI)	P	OR (95%CI)	P	OR (95%CI)	P	OR (95%CI)
chr1	1	0 (0-48.3)	1	0 (0-211.7)	0.026	8.4 (1-32.4)	1	0 (0-14.3)	$2.6x10^{-5}$	27.4 (6.8-79.8)
chr2	1	0 (0-21)	1	0 (0-79)	1	0 (0-55.6)	1	0 (0-51.9)	1.0	0 (0-129.4)
chr3	1	0 (0-36)	1	0 (0-317.7)	1	0 (0-72.5)	0.0015	40.6 (4.4-176.3)	1.0	0 (0-154.8)
chr4	1	0 (0-308.5)	0.0083	15.8 (1.8-62.9)	-	-	1	0 (0-36.7)	1.0	0 (0-571.1)
chr5	-	-	0.29	3 (0.1-17.8)	-	-	1	0 (0-77.4)	1.0	0 (0-68.5)
chr6	0.022	52.6 (1.1-445.5)	0.0052	9.1 (1.8-28.5)	1	0 (0-18.7)	1	0 (0-89.6)	1.0	0 (0-146)
chr7	0.058	17.9 (0.4-117.2)	0.00065	19.5 (3.8-63.1)	1	0 (0-145.6)	1	0 (0-108.7)	0.020	56.5 (1.3-422.4)
chr8	0.038	30.5 (0.7-238.7)	1	0 (0-180.8)	1	0 (0-821.4)	0.054	19.8 (0.5-134.7)	0.21	4.4 (0.1-26.6)
chr9	-	-	1	0 (0-28.7)	1	0 (0-24.6)	1	0 (0-24.5)	0.00024	28.8 (5.4-97.6)
chr10	0.035	32.7 (0.7-259.1)	1	0 (0-162.4)	1	0 (0-318.3)	1	0 (0-177.1)	1.0	0 (0-238.5)
chr11	1	0 (0-338.5)	$4.9x10^{-5}$	14 (4.3-35.4)	0.12	8.4 (0.2-50.8)	1	0 (0-18.6)	1.0	0 (0-426.7)
chr12	0.048	22.7 (0.5-160.1)	0.044	24.2 (0.6-160)	1	0 (0-264.9)	0.00015	33.1 (6.3-111)	0.076	13.7 (0.3-90.5)
chr13	-	-	0.00060	11.4 (3-31.6)	-	-	1	0 (0-35)	-	-
chr14	-	-	0.24	3.8 (0.1-22.5)	-	-	$6.8x10^{-5}$	7.8 (3-17.2)	0.087	11.6 (0.3-73.9)
chr15	-	-	1	0 (0-104.9)	-	-	1	0 (0-31.1)	0.57	1.2 (0-7.1)
chr16	0.062	17.2 (0.4-117.1)	1	0 (0-128.9)	1	0 (0-41.2)	0.019	10 (1.2-39.2)	-	-
chr17	1	0 (0-58.4)	0.0019	35 (3.9-149.3)	$4.5x10^{-7}$	82 (19.5-259)	0.21	4.2 (0.1-25.2)	1.0	0 (0-60.9)
chr18	0.041	27.2 (0.6-198.8)	-	-	1	0 (0-692.8)	1	0 (0-109.5)	0.12	8.5 (0.2-52.3)
chr19	-	-	-	-	1	0 (0-40.2)	1	0 (0-57.3)	-	-
chr20	-	-	0.0021	6 (1.9-15)	1	0 (0-507.1)	0.12	8.1 (0.2-49.3)	-	-

chr21	-	-	0.063	16.3 (0.4-103.6)	-	-	1	0 (0-123)	0.058	3.5 (0.7-10.9)
chr22	-	-	1	0 (0-75.6)	-	-	1	0 (0-22)	1.0	0 (0-13.1)

CI:confidence interval, OR:odds ratio

Supplementary Table 21. Associations between leukemia mortality and cell fraction of mosaic events.

cell fraction	case subjects	coeff	SE	HR (95%CI)	P
1-3%	ALL	0.97	0.28	2.64 (1.51-4.6)	$6.2x10^{-4}$
3-5%	ALL	1.43	0.43	4.19 (1.82-9.64)	$7.5x10^{-4}$
5%-	ALL	2.07	0.23	7.95 (5.06-12.48)	$<2.2x10^{-16}$

Subjects having hematopoietic malignancy are excluded. Results in Cox proportional hazard model with age, age^2, sex, smoking ,disease status and genotyping arrays in covariates.
Coeff: coefficient, SE: standard error, CI:confidence interval, HR: hazard ratio

Supplementary Table 22. Associations between leukemia mortality and presence of multiple mosaic events.

cell fraction	case subjects	coeff	SE	HR (95%CI)	P
1-3%	multiple	-	-	-	-
1-3%	single	0.95	0.29	2.58 (1.45-4.57)	0.0012
3-5%	multiple	2.69	0.59	14.74 (4.59-47.32)	$6.1x10^{-6}$
3-5%	single	0.87	0.59	2.4 (0.76-7.62)	0.14
5%-	multiple	2.71	0.35	15.02 (7.61-29.63)	$5.6x10^{-15}$
5%-	single	1.67	0.29	5.33 (3-9.47)	$1.1x10^{-8}$

Subjects having hematopoietic malignancy are excluded. Results in Cox proportional hazard model with age, age^2, sex, smoking ,disease status and genotyping arrays in covariates.
Coeff: coefficient, SE: standard error, CI:confidence interval, HR: hazard ratio

Supplementary Table 23. Associations between mosaic events and overall mortality.

	pLOSS		qLOSS		pCN-LOH		qCN-LOH		GAIN	
	P	OR (95%CI)	P	OR (95%CI)	P	OR (95%CI)	P	OR (95%CI)	P	OR (95%CI)
chr1	0.09	1.7 (0.9-3.1)	0.19	1.9 (0.6-5.5)	0.12	1.3 (0.9-1.8)	0.27	1.2 (0.9-1.7)	0.41	1.2 (0.7-1.9)
chr2	0.43	1.2 (0.8-1.9)	0.84	0.8 (0.3-2)	0.72	1.1 (0.5-2.4)	1	1 (0.5-1.9)	0.17	1.8 (0.7-4.5)
chr3	0.55	1.2 (0.6-2.2)	0.18	1.9 (0.6-5.4)	0.71	1.2 (0.5-2.4)	0.58	1.3 (0.6-2.7)	0.040	2.3 (1-5)
chr4	0.74	1.3 (0.3-5.4)	0.12	1.4 (0.9-2.3)	-	-	0.78	1.1 (0.6-1.9)	1.0	1.2 (0.1-10)
chr5	-	-	0.03	1.4 (1-1.9)	-	-	0.71	1.2 (0.5-2.5)	0.40	1.3 (0.7-2.6)
chr6	0.25	2 (0.6-7.1)	0.68	0.9 (0.7-1.3)	0.18	0.7 (0.4-1.1)	0.36	1.4 (0.7-2.9)	0.34	1.6 (0.6-4)
chr7	0.57	0.7 (0.3-1.6)	0.81	1.1 (0.6-1.7)	1	0.8 (0.2-2.8)	0.83	1.1 (0.4-2.6)	0.40	1.6 (0.4-4.8)
chr8	1	0.8 (0.2-2.8)	0.79	0.7 (0.2-2.4)	0.66	1.3 (0.1-9.7)	0.54	1.3 (0.6-3.1)	0.018	1.5 (1.1-2.2)
chr9	-	-	0.9	1 (0.6-1.8)	0.061	1.5 (1-2.4)	0.024	1.6 (1-2.4)	0.0094	2 (1.2-3.4)
chr10	0.75	0.7 (0.1-2.9)	1	0.8 (0.2-3.1)	0.72	1.2 (0.2-5.9)	0.35	1.5 (0.6-3.9)	0.70	0.4 (0-2.8)
chr11	0.41	1.5 (0.4-4.2)	0.52	1.1 (0.8-1.5)	0.61	1.1 (0.7-1.9)	0.31	1.2 (0.8-1.8)	0.30	2.1 (0.5-8.1)
chr12	0.29	1.6 (0.6-3.8)	0.55	1.3 (0.5-2.9)	0.77	1.3 (0.3-4.3)	0.77	1.1 (0.6-2)	0.26	1.5 (0.7-3.2)
chr13	-	-	0.28	1.2 (0.9-1.6)	-	-	0.88	1.1 (0.6-1.9)	-	-
chr14	-	-	0.18	1.3 (0.9-1.8)	-	-	0.00082	1.4 (1.1-1.6)	0.29	1.4 (0.7-2.5)
chr15	-	-	0.18	0.4 (0.1-1.5)	-	-	0.18	0.7 (0.3-1.2)	0.20	1.1 (0.9-1.4)
chr16	0.82	0.8 (0.3-2.2)	0.41	0.5 (0.1-2)	0.57	0.8 (0.4-1.5)	0.91	1 (0.6-1.5)	-	-
chr17	0.21	1.5 (0.8-2.8)	0.7	1.2 (0.5-2.6)	0.0028	2.3 (1.3-4.2)	0.56	0.9 (0.6-1.3)	0.31	1.4 (0.7-2.8)
chr18	0.43	0.5 (0.1-1.9)	-	-	0.037	3.6 (0.9-14.4)	0.68	1.2 (0.5-2.7)	1.0	1 (0.6-1.8)
chr19	-	-	-	-	0.62	1.2 (0.6-2.3)	1	1 (0.5-1.9)	-	-
chr20	-	-	0.01	1.3 (1.1-1.6)	0.7	1.4 (0.2-6.4)	0.49	1.2 (0.7-2)	-	-

chr21	-	-	0.32	0.6 (0.2-1.4)	-	-	0.81	1.1 (0.3-3)	0.67	1 (0.8-1.2)
chr22	-	-	0.29	1.5 (0.7-3)	-	-	0.55	0.8 (0.5-1.4)	0.10	0.7 (0.5-1.1)

CI:confidence interval, OR:odds ratio

Supplementary Table 24. Combinations of significantly co-occurring mosaic events in different chromosomes

mosaic1	mosaic2	P	OR (95%CI)	Shared by UK
chr3_GAIN	chr18_GAIN	$3.8x10^{-25}$	198 (98-365)	1
chr1_GAIN	chr7q_LOSS	$1.3x10^{-16}$	66 (32-123)	0
chr20q_LOSS	chr14q_CN-LOH	$1.1x10^{-13}$	4 (3-5)	0
chr14q_LOSS	chr21_GAIN	$2.0x10^{-13}$	11 (6-17)	0
chr1_GAIN	chr9_GAIN	$8.3x10^{-11}$	40 (17-81)	0
chr12_GAIN	chr13q_LOSS	$7.0x10^{-10}$	30 (13-62)	1
chr1p_CN-LOH	chr14q_CN-LOH	$2.1x10^{-9}$	3 (2-4)	0
chr3_GAIN	chr12_GAIN	$2.3x10^{-9}$	109 (34-274)	1
chr6p_CN-LOH	chr16p_CN-LOH	$3.0x10^{-9}$	10 (5-17)	0
chr14q_LOSS	chr22_GAIN	$4.2x10^{-9}$	18 (8-35)	0
chr3_GAIN	chr9_GAIN	$3.1x10^{-8}$	63 (20-157)	0
chr18_GAIN	chr22_GAIN	$3.9x10^{-8}$	24 (9-50)	0
chr17_GAIN	chr21q_LOSS	$5.9x10^{-8}$	123 (32-338)	0
chr12q_LOSS	chr14q_LOSS	$1.5x10^{-7}$	46 (14-115)	0
chr12_GAIN	chr18_GAIN	$1.9x10^{-7}$	44 (14-107)	1
chr3p_LOSS	chr9p_CN-LOH	$4.8x10^{-7}$	23 (8-51)	0
chr3_GAIN	chr8_GAIN	$7.2x10^{-7}$	33 (10-81)	0
chr3q_CN-LOH	chr14q_CN-LOH	$8.0x10^{-7}$	5 (3-9)	0
chr1_GAIN	chr3_GAIN	$2.5x10^{-6}$	47 (12-127)	0
chr6p_CN-LOH	chr14q_LOSS	$2.8x10^{-6}$	8 (4-16)	0
chr9_GAIN	chr18p_LOSS	$3.0x10^{-6}$	124 (24-419)	0
chr3p_LOSS	chr15q_LOSS	$3.4x10^{-6}$	116 (23-377)	0
chr7q_LOSS	chr11q_LOSS	$3.5x10^{-6}$	16 (6-36)	0
chr4q_LOSS	chr13q_LOSS	$3.7x10^{-6}$	16 (6-35)	0
chr5q_LOSS	chr17p_CN-LOH	$5.0x10^{-6}$	11 (4-24)	0
chr9p_CN-LOH	chr14q_CN-LOH	$5.5x10^{-6}$	3 (2-5)	0
chr17p_LOSS	chr21q_LOSS	$6.5x10^{-6}$	92 (18-288)	1
chr7_GAIN	chr9p_LOSS	$6.6x10^{-6}$	627 (67-2791)	0
chr11q_LOSS	chr14q_CN-LOH	$7.3x10^{-6}$	3 (2-5)	0
chr18_GAIN	chr13q_LOSS	$9.5x10^{-6}$	13 (5-30)	0

We assess co-occurrence of mosaic events (more than 10 carriers) in different chromosome. Loss and CN-LOH are evaluated in p,q arm-basis. As a result, 4,299 combinations remain for evaluation. Significance level is set of 0.05/4,299. OR: odds ratio, CI: confidence interval

Supplementary Table 25. Variants associated with *JAK2* V617F demonstrating associations with chr9p CN-LOH.

Gene	SNP	Chrband	Pos	ref	risk	Freqcont	OR (95%CI)	P	shared risk
[JAK2]	rs59384377	9p24.1	5005034	A	T	0.35/0.24	1.65 (1.43-1.90)	$3.2x10^{-11}$	Yes
[TERT]	rs7705526	5p15.33	1285974	C	A	0.43/0.36	1.32 (1.14-1.53)	$1.9x10^{-4}$	Yes
[TERT]	rs2853677	5p15.33	1287194	A	G	0.36/0.31	1.27 (1.10-1.46)	0.0012	Yes
[SH2B3]	rs7310615	12q24.12	111865049	G	C	NA	NA	NA	NA
CXXC4—[]—TET2	rs1548483	4q24	105749895	C	T	NA	NA	NA	NA
[CHEK2]	rs555607708	22q12.1	29091857	I	D	NA	NA	NA	NA
[ATM]	rs1800056	11q22.3	108138003	T	C	NA	NA	NA	NA
[PINT]	rs58270997	7q32.3	130729394	C	T	0.84/0.80	1.27 (1.05-1.52)	0.011	Yes
GFI1B-[]–GTF3C5	rs621940	9q34.13	135870130	C	G	0.058/0.050	1.17 (0.87-1.56)	0.29	Yes

Chrband: chromosome band, Pos: position, ref: reference allele, risk: risk allele, OR: odds ratio, CI: confidence interval, shared risk: shared risk allele (increasing presence of mosaic) between chromosome 9p CN-LOH and *JAK2* V617F in Hinds et al. 2016.

Supplementary Table 26.Candidate analyses of variants associated with MPN, CLL and mLOY in the current study.

Variant	Chr:Pos	trait	gene	loss	CN-LOH	gain	ANY
rs2736609	1:156202640	mLOY	*PMF1,SEMA4A*	0.40	0.51	0.067	0.089
rs11125529	2:54475866	telo	*ACYP2*	0.48	0.020	0.57	0.0029
rs13401811	2:111616104	CLL	*ACOXL,BCL2L11*	0.15	0.61	0.21	0.65
rs17483466	2:111797458	CLL	*ACOXL,BCL2L11*	0.74	0.33	0.87	0.88
rs58055674	2:111831793	CLL	*ACOXL*	0.53	0.53	0.97	0.71
rs1439287	2:111871897	CLL	*ACOXL,BCL2L11*	0.49	0.28	0.29	0.50
rs9308731	2:111908262	CLL	*BCL2L11*	0.31	0.79	0.42	0.25
rs13015798	2:201909515	CLL	*FAM126B,CASP8*	0.52	0.0037	0.42	0.18
rs3769825	2:202111380	CLL	*CASP8,CASP10*	0.22	0.012	0.16	0.44
rs13397985	2:231091223	CLL	*SP140*	0.22	0.021	0.73	0.012
rs9880772	3:27777779	CLL	*EOMES*	0.45	0.082	0.061	0.00079
rs115854006	3:48388170	mLOY	*TREX1,PLXNB1*	0.87	0.35	0.71	0.56
rs13088318	3:101242751	mLOY	*SENP7*	0.28	0.057	0.15	0.79
rs59633341	3:150018880	mLOY	*TSC22D2*	0.14	0.0062	0.40	0.0020
rs2201862	3:168648039	MPN	*EGFEM1P,MECOM*	0.79	0.24	0.88	0.73
rs10936599	3:169492101	CLL,telo	*MYNN*	0.44	0.15	0.25	0.0093
rs9815073	3:188115682	CLL	*LPP*	0.12	0.099	0.14	0.036
rs898518	4:109016824	CLL	*LEF1*	0.13	0.56	0.77	0.084
rs6858698	4:114683844	CLL	*CAMK2D*	0.21	0.84	0.94	0.89
rs56084922	5:111061883	mLOY	*NR*	0.59	0.087	0.36	0.90
rs9391997	6:409119	CLL	*IRF4*	0.83	0.90	0.11	0.40

rs872071	6:411064	CLL	*IRF4*	0.88	0.91	0.10	0.41
rs926070	6:32257566	CLL	*HLA*	0.45	0.60	0.63	0.54
rs674313	6:32578082	CLL	*HLA-DRB5*	0.27	0.47	0.32	0.38
rs9273363	6:32626272	CLL	*HLA*	0.91	0.45	0.84	0.10
rs210142	6:33546837	CLL	*BAK1*	0.66	0.11	0.44	0.43
rs9487023	6:109590004	mLOY	*C6orf183*	0.012	0.42	0.68	0.95
rs13191948	6:109634599	mLOY	*SMPD2,CCDC162P*	0.019	0.34	0.61	0.92
rs2236256	6:154478440	CLL	*IPCEF1*	0.25	0.41	0.41	0.54
rs381500	6:164478388	mLOY	*QKI*	0.34	0.12	0.63	0.41
rs17246404	7:124462661	CLL	*POT1*	0.089	0.60	0.50	0.86
rs58270997	7:130729394	MPN	*PINT*	0.22	0.70	0.83	0.71
rs2511714	8:103578874	CLL	*ODF1,KLF10*	0.37	0.20	0.53	0.58
rs2466035	8:128211229	CLL	*MYC*	0.021	0.019	0.53	0.99
rs1679013	9:22206987	CLL	*AS1,CDKN2B*	0.92	0.53	0.32	0.75
rs1359742	9:22336996	CLL	*DMRTA1,CDKN2B-AS1*	0.36	0.78	0.46	0.92
rs621940	9:135870130	MPN	*GFI1B*	0.96	0.29	0.65	0.39
rs1800682	10:90749963	CLL	*ACTA,FAS*	0.033	0.11	0.04	0.20
rs4406737	10:90759724	CLL	*ACTA2,FAS*	0.025	0.049	0.13	0.17
rs9420907	10:105676465	telo	*OBFC1*	0.83	0.46	0.059	0.93
rs7944004	11:2311152	CLL	*TSPAN32*	0.39	0.19	0.0080	0.74
rs4754301	11:108048541	mLOY	*NPAT,ATM,ACAT1*	0.037	0.0055	0.35	0.00014*
rs35923643	11:123355391	CLL	*GRAMD1B*	0.0086	0.097	0.11	0.012
rs735665	11:123361397	CLL	*SCN3B,GRAMD1B*	0.0097	0.10	0.11	0.017

rs2953196	11:123368333	CLL	*NR*	0.75	0.95	0.94	0.25
rs4251697	12:12874462	mLOY	*CDKN1B*	0.029	0.67	0.88	0.86
rs8024033	15:40403657	CLL	*BMF*	0.81	0.05	0.66	0.90
rs11636802	15:56775597	CLL	*MNS1,RFXDC2*	0.66	0.19	0.56	0.27
rs72742684	15:56780767	CLL	*MNS1,RFX7*	0.66	0.22	0.59	0.28
rs2052702	15:69989505	CLL	*PCAT29*	0.41	0.72	0.14	0.62
rs7176508	15:70018990	CLL	*RPLP1*	0.43	0.71	0.18	0.59
rs12448368	16:81044947	mLOY	*CENPN,ATMIN*	0.69	0.90	0.11	0.58
rs391023	16:85927814	CLL	*IRF8*	0.82	0.13	0.70	0.53
rs391855	16:85928621	CLL	*IRF8*	0.80	0.29	0.96	0.64
rs391525	16:85944439	CLL	*IRF8*	0.058	0.49	0.87	0.10
rs1044873	16:85955671	CLL	*IRF8*	0.62	0.63	0.64	0.35
rs77522818	17:47817373	mLOY	*FAM117A*	0.20	0.11	0.025	0.36
rs11082396	18:42080720	mLOY	*SETBP1*	0.011	0.19	0.45	0.035
rs8088824	18:42151261	mLOY	*LINC01601;SETBP1*	$1.4x10^{-5}$*	$8.3x10^{-5}$*	0.21	$2.1x10^{-5}$*
rs4368253	18:57622287	CLL	*PMAIP1*	0.092	0.66	0.69	0.75
rs4987852	18:60793921	CLL	*BCL2*	0.37	0.11	0.39	0.23
rs8105767	19:22215441	telo	*ZNF208*	0.062	0.85	0.85	0.46
rs755017	20:62421622	telo	*RTEL1*	0.99	0.26	0.15	0.87

*indicates significant associations based on Bonferroni's correction ($p \leq 0.00020$ (0.05/63/4)).

Supplementary Table 27. The significant association between rs8088824 at *LINC01601*/*SETBP1* and mosaic events is largely explained by an association of chr20q loss and chr14q CN-LOH.

	p_loss	q_loss	p_CN-LOH	q_CN-LOH	gain
chr1	0.68	0.48	0.49	0.12	0.12
chr2	0.80	0.83	0.52	0.92	0.25
chr3	0.070	0.42	0.87	0.15	0.079
chr4	0.75	0.0078	-	0.57	1
chr5	-	0.34	-	0.011	1
chr6	0.55	0.64	0.19	0.27	0.64
chr7	0.063	1	0.33	0.21	0.32
chr8	0.70	0.41	0.77	0.32	1
chr9	-	0.18	0.50	0.36	0.43
chr10	1	0.098	0.48	0.088	1
chr11	0.76	0.79	0.47	0.84	0.85
chr12	0.0099	0.15	0.56	0.30	0.31
chr13	-	0.26	-	0.14	-
chr14	-	0.048	-	$1.1x10^{-7}$*	0.86
chr15	-	0.89	-	0.21	0.78
chr16	0.015	0.61	0.16	0.14	-
chr17	0.11	0.15	0.87	0.11	0.18
chr18	0.33	-	0.40	0.85	0.74
chr19	-	-	0.16	0.47	-
chr20	-	0.00015*	0.88	0.80	-
chr21	-	0.28	-	0.37	0.55
chr22	-	0.42	-	0.093	0.26

*significant after Bonferroni's correction

Supplementary Table 28. Oligonucleotide sequences used for luciferase assay.

oligonucleotide	sequence
MRE11_C allele	TTCGTAGAATGAGTTAGGGAGGAGC**C**TCCCTTGATTTTTTTGGAATAATTT
MRE11_C allele_c	AAATTATTCCAAAAAAATCAAGGGA**G**GCTCCTCCCTAACTCATTCTACGAA
MRE11_T allele	TTCGTAGAATGAGTTAGGGAGGAGC**T**TCCCTTGATTTTTTTGGAATAATTT
MRE11_T allele_c	AAATTATTCCAAAAAAATCAAGGGA**A**GCTCCTCCCTAACTCATTCTACGAA
MPL_G allele	GGAAGTATTTACACCACAAAAAGCA**G**CAAATGCTACCAAACAGAACACCCT
MPL_G allele_c	AGGGTGTTCTGTTTGGTAGCATTTG**C**TGCTTTTTGTGGTGTAAATACTTCC
MPL_A allele	GGAAGTATTTACACCACAAAAAGCA**A**CAAATGCTACCAAACAGAACACCCT
MPL_A allele_c	AGGGTGTTCTGTTTGGTAGCATTTG**T**TGCTTTTTGTGGTGTAAATACTTCC

_c indicates complementary sequences. The position of variants are shown in bold.