nature

SUPPLEMENTARY INFORMATION

https://doi.org/10.1038/s41586-018-0321-x

In the format provided by the authors and unedited.

Insights into clonal haematopoiesis from 8,342 mosaic chromosomal alterations

Po-Ru Loh[1,2,14]*, Giulio Genovese[2,3,4,14]*, Robert E. Handsaker[2,3,4], Hilary K. Finucane[2,5], Yakir A. Reshef[6], Pier Francesco Palamara[7], Brenda M. Birmann[8], Michael E. Talkowski[2,3,9,10], Samuel F. Bakhoum[11,12], Steven A. McCarroll[2,3,4,15]* & Alkes L. Price[2,13,15]*

[1]Division of Genetics, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, MA, USA. [2]Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, MA, USA. [3]Stanley Center for Psychiatric Research, Broad Institute of MIT and Harvard, Cambridge, MA, USA. [4]Department of Genetics, Harvard Medical School, Boston, MA, USA. [5]Schmidt Fellows Program, Broad Institute of MIT and Harvard, Cambridge, MA, USA. [6]Department of Computer Science, Harvard University, Cambridge, MA, USA. [7]Department of Statistics, University of Oxford, Oxford, UK. [8]Channing Division of Network Medicine, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, MA, USA. [9]Center for Genomic Medicine, Massachusetts General Hospital, Boston, MA, USA. [10]Department of Neurology, Massachusetts General Hospital and Harvard Medical School, Boston, MA, USA. [11]Department of Radiation Oncology, Memorial Sloan Kettering Cancer Center, New York, NY, USA. [12]Sandra and Edward Meyer Cancer Center, Weill Cornell Medicine, New York, NY, USA. [13]Departments of Epidemiology and Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA, USA. [14]These authors contributed equally: Po-Ru Loh, Giulio Genovese. [15]These authors jointly supervised this work: Steven A McCarroll, Alkes L Price. *e-mail: poruloh@broadinstitute.org; giulio.genovese@gmail.com; mccarroll@genetics.med.harvard.edu; aprice@hsph.harvard.edu

Supplementary Notes and Tables for "Insights into clonal haematopoiesis from 8,342 mosaic chromosomal alterations"

Po-Ru Loh, Giulio Genovese, Robert E Handsaker, Hilary K Finucane, Yakir A Reshef, Pier Francesco Palamara, Brenda M Birmann, Michael E Talkowski, Samuel F Bakhoum, Steven A McCarroll, Alkes L Price

Contents

Supplementary Notes

1 Mosaic chromosomal alteration detection

Our procedure for calling mCA is overviewed in Methods; here we provide additional details omitted in Methods for brevity.

1.1 Computation of LRR and BAF from genotyping intensities

We converted UK Biobank genotyping intensity data (i.e., A and B allele probe set intensities, A_{int} and B_{int}) into $\log_2$ R ratio (LRR) and B allele frequency (BAF) values [51] using an analysis pipeline similar to Jacobs et al. [1] consisting of the following four steps.

1. **For each genotyping batch, for each cluster of called genotypes (AA, AB, BB), compute cluster median in (*X, Y*) = (contrast, size)-space [70]:**

$$X = \log_2 A_{int} - \log_2 B_{int} \quad (1)$$

$$Y = (\log_2 A_{int} + \log_2 B_{int})/2. \quad (2)$$

   We computed batch-level cluster centers to account for possible batch effects (given that the UK Biobank genotyping was done in batches of ≈4,800 samples). If a cluster contained fewer than 10 calls, we set its median intensities to missing. For chromosome X, we considered only genotypes of female samples.

2. **For each individual, affine-normalize and GC-correct (*X, Y*) transformed intensities.**

   This procedure corrects for systematic variation in probe intensities across SNPs for a particular individual (e.g., broadly elevated or reduced intensity levels) as well as for "GC-wave" artifacts [52]. Explicitly, in a manner similar to Jacobs et al. [1], we set up a pair of multivariate linear regressions

$$X_{m,\text{exp}} = \alpha + X_m\beta_X + Y_m\beta_Y + \sum_{k=1}^{9}\sum_{p=1}^{2}\left[\left(f_{m,k}^{\text{GC}}\right)^p \cdot \beta_{k,p}^{\text{GC}} + \left(f_{m,k}^{\text{CpG}}\right)^p \cdot \beta_{k,p}^{\text{CpG}}\right] \quad (3)$$

$$Y_{m,\text{exp}} = \gamma + X_m\delta_X + Y_m\delta_Y + \sum_{k=1}^{9}\sum_{p=1}^{2}\left[\left(f_{m,k}^{\text{GC}}\right)^p \cdot \delta_{k,p}^{\text{GC}} + \left(f_{m,k}^{\text{CpG}}\right)^p \cdot \delta_{k,p}^{\text{CpG}}\right], \quad (4)$$

   where m indexes SNPs, (X_m, Y_m) are intensity values in (constrast, size)-space for the current individual at SNP m, $(X_{m,\text{exp}}, Y_{m,\text{exp}})$ is the cluster center (computed in Step 1) corresponding to the individual's called genotype at SNP m, and $\{f_{m,k}^{\text{GC}}, f_{m,k}^{\text{CpG}}\}_{k=1}^{9}$ are proportions

of GC and CpG content in 9 windows of 50, 100, 500, 1k, 10k, 50k, 100k, 250k, and 1M bp centered around SNP m. We computed GC content using bedtools [71] on the human reference (hg19), and we computed CpG content using the EpiGRAPH CpG annotation [72].

Equations (3) and (4) without the GC and CpG terms amount to an affine transformation of the individual's observed intensity values (X_m, Y_m) to best match the "expected" intensity values $(X_{m,\exp}, Y_{m,\exp})$ based on the individual's called genotypes. The GC and CpG terms constitute a polynomial (quadratic) model for artifactual variation due to effects of local GC and CpG content on measured probe intensities [52].

We performed least-squares regression on equations (3) and (4) (ignoring SNPs at which the individual's genotype was uncalled or the relevant cluster center was set to missing) to obtain corrected (X, Y) values, defined as the regression predictions (i.e., $(X_{m,\exp}, Y_{m,\exp})$ minus the least-squares residuals).

3. **For each genotyping batch, for each cluster of called genotypes (AA, AB, BB), compute means of corrected (*X, Y*) values.**

   In this step we recomputed cluster centers on the affine-normalized and GC-corrected (X, Y) values (taking means rather than medians but otherwise following Step 1).

4. **For each genotype, transform corrected (*X, Y*) values to LRR and BAF.**

   Lastly, we transformed corrected (*X, Y*) values using a polar-like transformation followed by linear interpolation in a manner similar to Peiffer et al. [51]. We set

$$\theta = \frac{2}{\pi} \cdot \arctan\left(2^{X_{AB}-X}\right) \quad (5)$$

$$\log_2 R = Y, \quad (6)$$

   where in the first equation X_{AB} denotes the mean corrected $X = \log_2 A_{int}/B_{int}$ value for genotypes called as hets at the current SNP. (We filtered out SNPs for which X_{AB} was missing.)

   We transformed cluster centers in the same manner to obtain $(\theta_{AA}, \log_2 R_{AA})$, $(\theta_{AB}, \log_2 R_{AB})$, and $(\theta_{BB}, \log_2 R_{BB})$. We then performed linear interpolation between cluster centers [51] in $(\theta, \log_2 R)$-space to estimate BAF and expected $\log_2 R$ for each genotype, from which we obtained LRR as $\log_2 R$ - $\log_2 R_{\text{exp}}$. (If one of the cluster centers $(\theta_{AA}, \log_2 R_{AA})$ and $(\theta_{BB}, \log_2 R_{BB})$ was missing, we set it to the reflection of the opposite cluster center across the vertical line $\theta = \theta_{AB}$.)

QC filters on anomalous BAF and LRR. For each sample, we computed s.d.(BAF) within each autosome, and we removed 320 samples with median s.d.(BAF)>0.11. We further ignored

chromosomes with mean LRR > 0.3 (possible non-mosaic trisomy) or mean LRR $<$ -0.5 (possible non-mosaic monosomy).

Masked genomic regions. Following Laurie et al. [2], we excluded genotype measurements in the HLA region on chromosome 6 (28,477,797–33,448,354, build 37) and the X Translocation Region (XTR) on chromosome X (88,575,629–92,308,067).

1.2 Filtering constitutional segmental duplications

Before testing for mCAs, we first ran a pre-processing step in which we identified and masked likely constitutional (i.e., inherited) segmental duplications. Constitutional duplications can create false positive detections of mCAs because they have the same effect on BAF and LRR as a somatic gain event at 100% cell fraction. (Constitutional deletions also behave like somatic loss events at 100% cell fraction, but because our mCA detection algorithm only uses BAF at heterozygous sites, segmental deletions were not a concern: deletions result in hemizygosity with no heterozygous sites.)

Fortunately, constitutional duplications are relatively easy to filter as they are characteristically short (typically $<$1Mb) and produce extreme shifts in genotyping intensities: heterozygous sites have AAB or ABB genotypes with $|\Delta\text{BAF}|\sim0.17$, and all sites have triploid total copy number with LRR$\sim$0.36 (Fig. 2a and Fig. S1.2-1). To call and mask such regions, we modeled observed phased BAF deviations (pBAF) across a chromosome using a 25-state hidden Markov model (HMM) with states corresponding to pBAF values in [–0.24, +0.24] at intervals of 0.02. We assumed each state emitted a normally distributed observed pBAF with mean equal to the state value and standard deviation equal to the empirical s.d.(BAF) at each site (measured across all individuals within a genotyping batch), capping z-scores at 4 to reduce outlier influence. We allowed transitions between the 0 state and each nonzero state with probability 0.003 (modeling event boundaries) and between each nonzero state and its negative with probability 0.001 (modeling phase switch errors). At the telomeres, we assigned a probability of 0.01 to starting/ending in each nonzero state (to favor calls that end at the telomeres).

We selected regions to mask by computing the Viterbi (maximum likelihood) path through the above HMM and examining contiguous regions of nonzero states. We masked regions of $<$2Mb with $|\Delta\text{BAF}|>0.1$ and LRR$>$0.1, which we deemed to be likely constitutional duplications, and we further masked gaps (of $<$2Mb) between nearby regions of this form (assuming that the 1Mb flanks of the merged region had no apparent mosaicism, i.e., $|\Delta\text{BAF}|<0.05$). In total we masked 267,666 likely constitutional duplications among 151,202 individuals. We believe that this procedure filtered out most constitutional duplications of sufficient size to impact our analyses. At the end of our mCA calling pipeline, we performed further QC (Fig. S1.2-1) to eliminate a small minority of uncaught likely constitutional duplications.

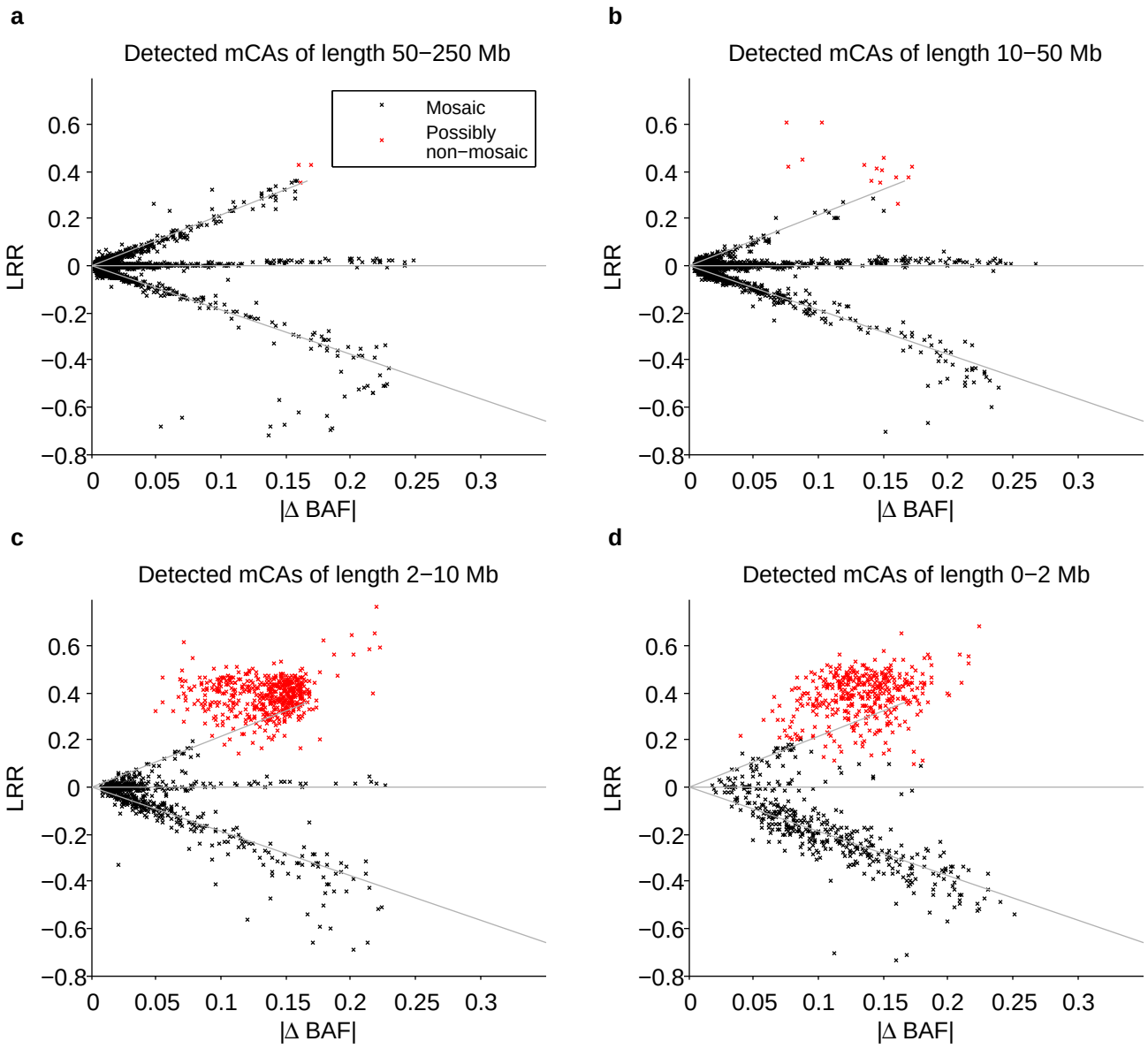
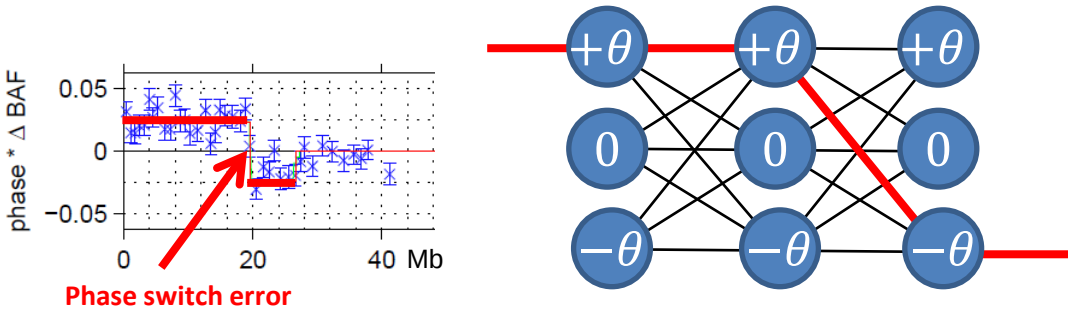


Figure S1.2-1. Exclusion of possible constitutional duplications. We filtered events of length >10Mb with LRR>0.35 or with LRR>0.2 and |ΔBAF|>0.16, and we filtered events of length <10Mb with LRR>0.2 or with LRR>0.1 and |ΔBAF|>0.1. Constitutional duplications have expected |ΔBAF|=1/6, corresponding to LRR of roughly 0.36. The bottom two panels (corresponding to event calls 2–10Mb and <2Mb) each clearly contain a cluster of calls around |ΔBAF|=1/6, LRR=0.36. We chose exclusion thresholds to conservatively discard all calls that might belong to this cluster, applying more stringent filtering to shorter events because (i) most constitutional duplications are short and (ii) shorter events have noisier LRR and |ΔBAF| estimates.



- Hidden Markov model:
  - 1 parameter: θ = |ΔBAF| in mosaic region
  - 3 states: E[phase*ΔBAF] = $+\theta, 0, -\theta$
- Detection procedure:
  - Compute LRT statistic for testing $\theta \neq 0$
  - Calibrate empirically using permutation

Figure S1.3-1. Hidden Markov model for detecting mCAs. Mosaic chromosomal alterations, which alter the balance of maternal vs. paternal chromosome content in a cell population, cause deviations in allelic balance ($|\Delta\text{BAF}|$) at heterozygous sites. In computationally phased genotyping intensity data, these deviations manifest as stretches of signed deviations with the same absolute value (θ) but with sign flips at phase switch errors. A three-state Hidden Markov model with the single parameter θ captures this behavior and enables computation of a likelihood ratio test statistic.

1.3 Parameterized hidden Markov model for event detection

The above approach of performing Viterbi decoding on a many-state hidden Markov model works well for finding constitutional duplications, but to define a formal, well-calibrated statistical test sensitive to mCAs at low cell fractions, we took the following more principled approach. We replaced the single 25-state HMM described above with a *family* of 3-state HMMs parameterized by a single parameter θ representing mean $|\Delta\text{BAF}|$ within a mosaic event (i.e., the states of the HMM are $\{-\theta, 0, +\theta\}$; Fig. S1.3-1). The key advantages of this approach are that (i) it naturally produces a likelihood ratio test statistic for testing $\theta \stackrel{?}{=} 0$ (described in the following section); and (ii) the derived test statistic integrates over uncertainty in phase switches and mCA boundaries (unlike maximum likelihood estimation).

Aside from the reduction in the number of states, the 3-state HMM that we used for event detection differs from the 25-state HMM described above only in values of a few constants. We

reduced the $\pm\theta \rightarrow 0$ "stop" transition probability to 3×10^{-4} in autosomes and 1×10^{-4} in chromosome X, reflecting the fact that most somatic events of interest span tens of megabases. We reduced the $0 \rightarrow \pm\theta$ "start" transition probability to 0.004 (resp. 0.08) times the stop probability in autosomes (resp. chromosome X). (The asymmetry in start vs. stop probabilities reflects the fact that the HMM should not expect to be spend equal amounts of time in the mosaic vs. non-mosaic states; most portions of most chromosomes are expected to be non-mosaic.) We kept the $-\theta \leftrightarrow +\theta$ switch error probability at 0.001, roughly reflecting our estimated rate of large-scale phase switches [23, 24]. We did not assess a probabilistic penalty to starting/ending in nonzero states except in acrocentric chromosomes, for which we reduced the probability of starting in a nonzero state (at the centromere, given that we had no p-arm genotypes) by a factor of 0.2. As above, we assumed each state emitted a normally distributed observed pBAF; here we capped z-scores at 2 to further reduce outlier influence.

We note that a potential criticism of this 3-state HMM is that it does not properly model chromosomes with multiple mCAs of differing $|\Delta\text{BAF}|$. However, the primary purpose of this model is event discovery (particularly for mCAs at low cell fractions); after we called chromosomes containing events, we performed additional post-processing (described below) to pick up complex mCAs. Additionally, we re-estimated $|\Delta\text{BAF}|$ within mCA boundaries after making event calls.

1.4 Calling existence of an event: likelihood ratio test statistic

For a given sequence of phased BAF deviations (denoted x) on a chromosome, the family of HMMs parameterized by θ gives rise to a likelihood ratio test statistic as follows. For a given θ, we can compute the likelihood $L(\theta \mid x)$ as the total probability of observing x under the HMM with nonzero states $\pm\theta$. (This computation can be performed efficiently using dynamic programming.) The likelihood ratio for $\theta \stackrel{?}{=} 0$ is then given by

$$\Lambda(x) = \frac{L(0 \mid x)}{\sup_\theta\{L(\theta \mid x)\}}, \tag{7}$$

where the numerator is the likelihood under the model in which all states collapse to 0 (i.e., no mCA is present) and the denominator is the likelihood under the best choice of θ. (In practice, we discretized θ to run from 0.0025 to 0.25 in 40 multiplicative steps.)

Producing a hypothesis test for $\theta \stackrel{?}{=} 0$ takes one more step. While asymptotic theory can often be invoked to assert that $-2\log\Lambda$ is approximately χ^2 distributed under the null hypothesis, we have two issues here. Most importantly, our hidden Markov model is imperfect, and in particular, different choices of probability constants within the model can substantially change the absolute magnitude of the test statistic. Second, our null hypothesis θ=0 is at the boundary of the parameter space.

For these reasons, we chose to estimate an empirical null distribution for the test statistic

$-2\log\Lambda$ rather than relying on theory. We approximated the null distribution simply by taking observed pBAF sequences and randomizing phase at each heterozygous site (keeping $|\Delta\text{BAF}|$ fixed). We performed 5 independent randomizations per individual, computed $-2\log\Lambda$ for each replicate, and used the resulting distribution of null test statistics to determine the cutoff value that would achieve a false discovery rate of 0.05 in light of the test statistics observed on real data. We performed this calibration independently for each autosome and chromosome X, yielding critical values from 1.41–3.87. In Supplementary Data, we provide q-values from this procedure for each event in our call set.

We note that this calibration procedure assumes that the only source of autocorrelation in pBAF is a true mosaic event, whereas in reality, other sources of autocorrelation exist; in particular, we found that sample contamination produced autocorrelation in regions of long-range LD (resulting in unusual false positive calls that we subsequently filtered). While we believe that our filtering eliminated most samples affected by spurious autocorrelation, the true FDR achieved by this calibration is slightly larger than 5% due to residual artifacts. We explore this issue in detail in Supplementary Note 3.1.

1.5 Calling event boundaries

We have thus far described a method that, for a given sequence of phased BAF deviations on a given chromosome, performs a hypothesis test indicating whether or not a mCA somewhere on the chromosome is needed to explain the observed BAF deviations. However, if so (i.e., if the null hypothesis is rejected), the algorithm that we have described thus far makes no indication of where on the chromosome the mCA is located. The reason is that for the purpose of detection, we wanted to integrate over all possible mCA boundaries (and all possible phase switches). Now, after detecting an event on a chromosome, we need a separate algorithm to call its boundaries.

To estimate mCA boundaries on a chromosome deemed to contain an event, we took 5 samples from the posterior of the HMM using the likelihood-maximizing choice of θ. (We resampled if the state path for any posterior sample contained no nonzero states.) We then called the boundaries of the mCA using the consensus of the 5 samples. In Supplementary Data, we provide the ranges (among the 5 samples) for the left boundary and right boundary of each call. We analyze the coverage of these intervals for *FRA10B*-associated del(10q) events in Supplementary Note 4.2 and observe that the intervals achieve $\approx$73% coverage.

1.6 Calling copy number

The above detection procedure uses only BAF data and ignores LRR measurements by design (to be maximally robust to genotyping artifacts, e.g., "GC waves" that produce local shifts in genotyping intensities [52]); however, after detecting events, we incorporated LRR data to call

copy number. As in previous work [1, 2, 8], we observed that mean LRR in called mCAs either increased or decreased linearly with estimated BAF deviation (for losses and gains) or was near zero (for CNN-LOHs) (Fig. 2a and Fig. S1.6-1). These trend lines allowed us to estimate the expected LRR/|ΔBAF| slopes corresponding to gains and losses (approximately 2.16 and –1.89, respectively). For a particular event with estimated BAF deviation |ΔBAF|, mean LRR $\hat{\mu}$, and standard error of LRR $\hat{\sigma}$, we could then compute the relative probabilities that the event was a loss, CNN-LOH, or gain (assuming that $\hat{\mu}$ had been drawn from a normal distribution with mean $|\Delta\text{BAF}|\times\{-1.89, 0, 2.16\}$ and standard error $\hat{\sigma}$).

We implemented an improvement upon the above approach by leveraging chromosome-specific frequencies of loss, CNN-LOH, and gain. Specifically, we observed that some chromosomes contained many of one type of event and very few of another (Fig. 1), and we reasoned that this information should be helpful for calling events with uncertain copy number (i.e., events with low |ΔBAF| and therefore little separation between the expected mean LRRs corresponding to loss, CNN-LOH, or gain). To guard against circular reasoning, we first split the LRR vs. |ΔBAF| space into three zones bisecting the loss/CNN-LOH/gain trend lines: letting s = LRR/|ΔBAF|, we required that events with $s < -0.94$ be called either as loss or undetermined, events with $-0.94 \leq s < 1.08$ be called either as CNN-LOH or undetermined, and events with $1.08 \leq s$ be called either as gain or undetermined. We further required that in order to call an event within one of these zones, its mean LRR $\hat{\mu}$ needed to be either (i) at least twice as close to its expectation according to the closest trend line vs. the next closest; or (ii) within two standard errors $\hat{\sigma}$ of its expectation. With these rules in place, we assigned preliminary calls to each event, calling copy number for an event if the requirements above were satisfied and if the most likely call was at least 20 times more likely than the next-most likely (based on $\hat{\mu}$ and $\hat{\sigma}$ and the normal model described in the previous paragraph). We then re-called all events by performing the same procedure but incorporating a prior on call probabilities: for a given event, we put a prior on its copy number derived from the preliminary calls made for up to 20 events with similar boundaries (differing by <10Mb and <10% of chromosome length), adding a pseudo-count of 0.5 to prevent copy numbers from being assigned zero probability. Again, we only called copy number if the most likely call was at least 20 times more likely than the next-most likely ($\sim$95% confidence).

We note one special case that we handled separately: isochromosomes, which involve simultaneous loss of one chromosomal arm and gain of the other (most notably i(17q); Fig. S2-17) were initially called as single whole-chromosome events by our detection procedure (which only considered BAF). We therefore included a separate check for whole-chromosome events examining whether LRR was significantly different for the p vs. q arms, and if so, we split the event at the centromere. We also performed manual review more generally to search for events with multiple |ΔBAF| and/or LRR levels within a call, but did not find such events beyond subclonal CNN-LOHs (Section 1.8).

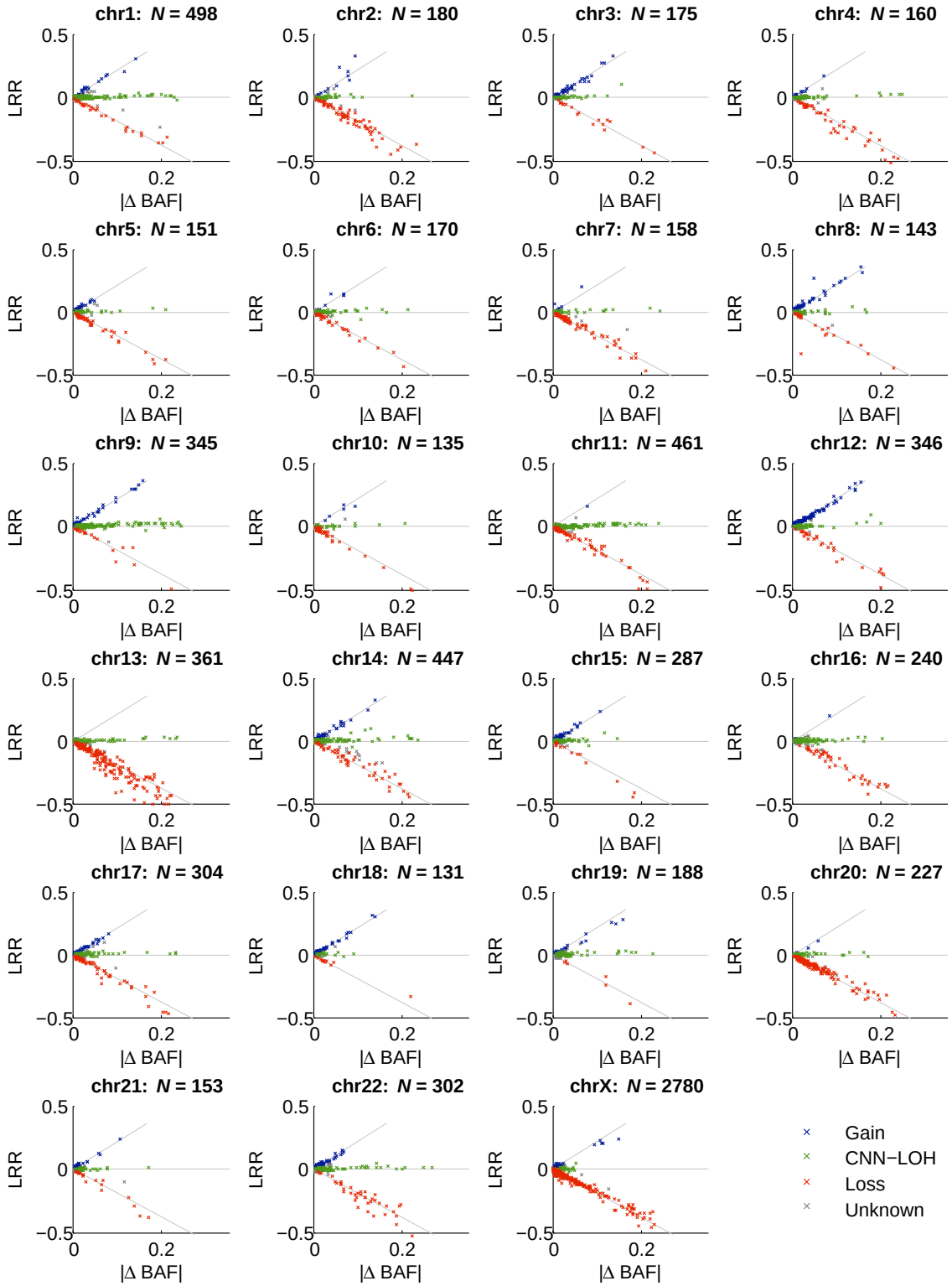


Figure S1.6-1. Total vs. relative allelic intensities of mCAs detected on each chromosome. For each of N mCAs, mean $\log_2$ R ratio (LRR) of each detected mCA is plotted against estimated change in B allele frequency at heterozygous sites ($|\Delta\text{BAF}|$). The data exhibits the characteristic "arrowhead" pattern in which LRR/$|\Delta\text{BAF}|$ approximately equals a positive constant for gain events, zero for CNN-LOH events, and a negative constant for loss events. This pattern is very consistent across chromosomes.

1.7 QC filters on anomalous mCA calls

We found the approach that we have described to be quite robust, with the overall genomic distribution of detected events broadly consistent with previous work [1, 2, 7, 8]. However, in our initial analysis, we did detect several hundred apparent short interstitial CNN-LOH events indicative of technical artifacts (given that CNN-LOHs are generally produced by mitotic recombination and stretch to a telomere). On inspection, we discovered that the overwhelming majority of these artifactual events occurred at five specific regions of the genome: chr3:∼45Mb (11 events), chr6:∼30Mb (709 events), chr8:∼45Mb (12 events), chr10:∼80Mb (40 events), chr17:∼40Mb (40 events). We also noticed that multiple such detections often occurred in the same sample; the union of all carriers contained 717 samples, nearly all of which carried the chr6 artifact at *HLA* (which we did not mask from this initial analysis). The chr3, chr6, and chr8 regions have all been previously noted to harbor long-range LD [73], which suggested sample contamination [8] as the likely culprit: if a sample were contaminated with cells from another individual, then in regions of long-range LD (i.e., long haplotypes), allelic balance could shift in favor of one of the original sample's parental haplotypes (whichever one was a closer match to the foreign DNA). To be safe, we therefore excluded all 717 of these samples from our analysis, and we further excluded 6 individuals with three or more interstitial CNN-LOH calls and 2 individuals with three or more calls with high implied switch error rates, for a total of 725 exclusions.

Independent of the above issue, we also observed a rarer technical artifact in which short interstitial CNN-LOH calls were made in runs of homozygosity (ROH) in which a small fraction of sites had been incorrectly called as hets and subsequently phased on the same haplotype, resulting in very strong phase-aligned BAF deviations. These calls were easy to filter; we used a criterion of low heterozygosity ($<$1/3 the expected heterozygosity in the region) and LRR$>$–0.1 (guaranteeing that the region could not possibly be hemizygous due to a loss event). After applying these filters, we were left with only 32 interstitial CNN-LOH calls among all samples with no obvious artifacts upon manual review.

1.8 Hidden Markov model for detecting multiple subclonal CNN-LOH events

The framework that we have described is aimed at identifying and calling sporadic mCAs arising in a population cohort for which most individuals with detectable clonality have a single simple event (a single clonal loss, CNN-LOH, or gain) at low-to-modest cell fraction. However, for a small subset of individuals (mostly with prevalent or incident cancer diagnoses), we detected multiple events, giving rise to the possibility that some samples might carry overlapping or contiguous events that require more careful treatment. On closer inspection, we observed one common form of additional complexity not properly handled by our approach described thus far: multiple subclonal CNN-LOH events on the same chromosome arm (Extended Data Fig. 8).

To treat this special case, we performed a post-processing step in which we re-analyzed detected events using Viterbi decoding on a 51-state HMM with $|\Delta\text{BAF}|$ levels ranging from 0.01 to 0.25 in multiplicative increments. In this HMM, in addition to start/stop transitions between the 0 state and nonzero states (with probability 10^{-4}) and switch error transitions between each state and its negative (with probability 0.001), we also introduced $|\Delta\text{BAF}|$-shift transitions between different nonzero states (with probability 10^{-7}). At the telomeres, we assigned a probability of 0.01 to starting/ending in each nonzero state. We examined all calls for which the posterior decoding resulted in more than one $|\Delta\text{BAF}|$ state, and we observed that in nearly all of these cases, the event in question had originally been called as a CNN-LOH but exhibited a step function of increasing BAF deviations toward the telomere (consistent with multiple subclonal CNN-LOH events covering varying segments of a chromosome arm). We describe all such events in Extended Data Fig. 8.

We note that all five individuals in Extended Data Fig. 8 with multiple CNN-LOH events on chr13q appear to contain switch errors over 13q14. In reality, these individuals all also contain 13q14 deletions (evident in LRR data) and are mixtures of the following cell populations:

1. Normal cells: 1 paternal chr13, 1 maternal chr13.

2. del(13q14) cells, say on paternal chr13: 0 paternal 13q14, 1 maternal 13q14 (and normal elsewhere on chr13).

3. del(13q14) CNN-LOH cells: 0 paternal 13q14, 2 paternal rest of chr13, 0 maternal chr13.

The result is maternal>paternal allelic imbalance in 13q14 but paternal>maternal imbalance in the rest of chr13, resulting in the observed phased BAF profiles.

The individuals with multiple CNN-LOH events that carry germline risk alleles in *cis* (Table 1) are as follows:

- 1 of 8 individuals with multiple 1p CNN-LOH events carries a germline risk haplotype: individual 165 carries the rs182971382 risk allele. An additional 3 individuals (39, 90, and 25) belong to IBD clusters at the *MPL* locus (Extended Data Fig. 5c).

- 10 of 12 individuals with multiple 9p CNN-LOH events (all but 1678 and 1623) carry the *JAK2* 46/1 risk haplotype.

- 0 of 2 individuals with multiple 11q CNN-LOH events carry the rs532198118 risk allele in the *ATM* locus.

- 5 of 7 individuals with multiple 15q CNN-LOH events (all but 3508 and 3454) carry the $\sim$70kb deletion at 15q26.3.

2 Per-chromosome plots of mosaic event calls

On the following pages, we provide per-chromosome “pile-up” plots of all mosaic chromosomal alterations called on each chromosome.

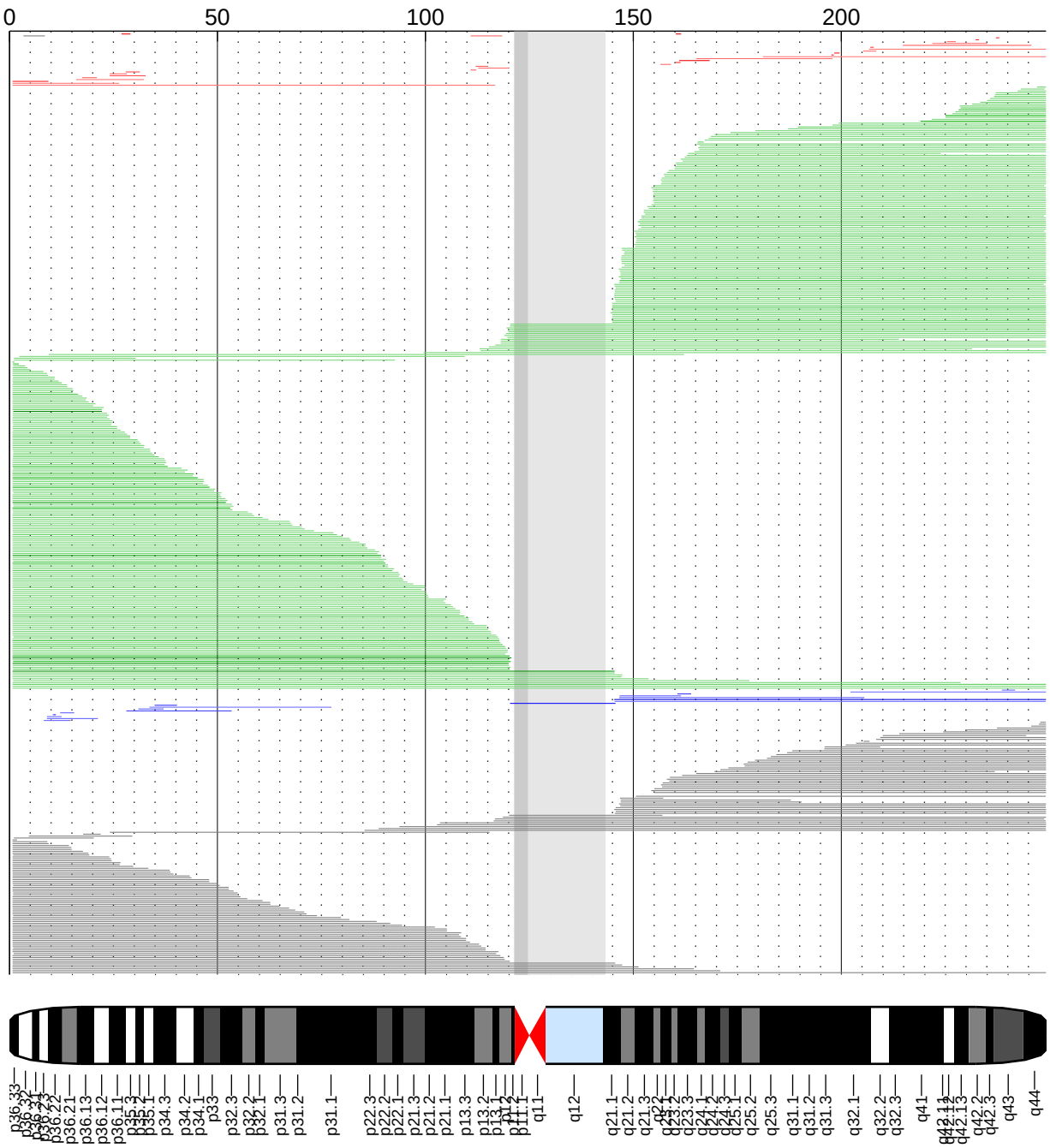


Figure S2-1. Detected mCAs on chromosome 1. Events are color-coded by copy-number: loss (red), CNN-LOH (green), gain (blue), undetermined (grey). Darker coloring indicates higher allelic fraction. Multiple events within a single individual are plotted with the same y-coordinate (at the top of the plot). Note that events with unknown copy number also generally have greater uncertainty in their boundaries due to low allelic fraction.

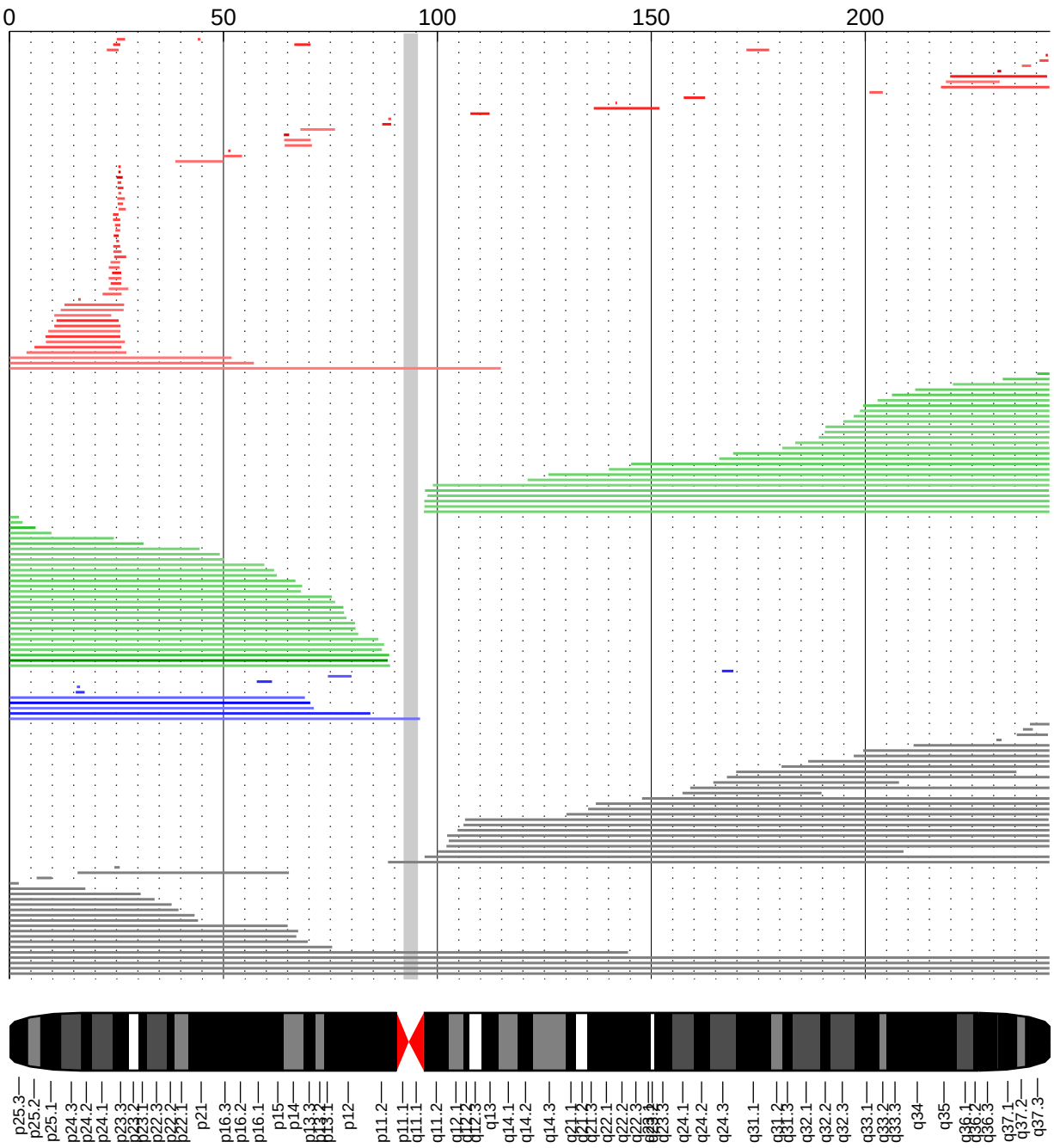


Figure S2-2. Detected mCAs on chromosome 2. Events are color-coded by copy-number: loss (red), CNN-LOH (green), gain (blue), undetermined (grey). Darker coloring indicates higher allelic fraction. Multiple events within a single individual are plotted with the same y-coordinate (at the top of the plot). Note that events with unknown copy number also generally have greater uncertainty in their boundaries due to low allelic fraction.

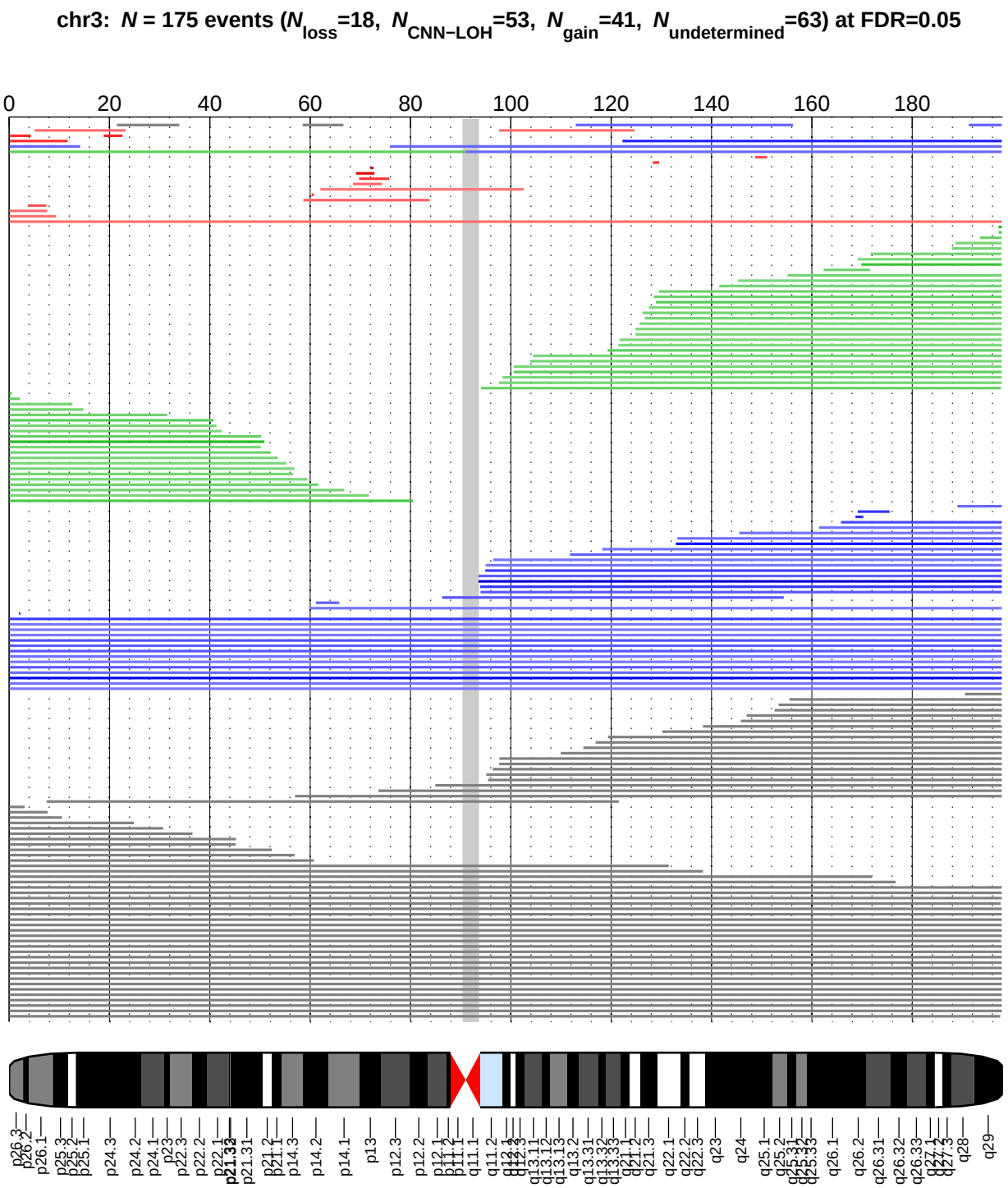


Figure S2-3. Detected mCAs on chromosome 3. Events are color-coded by copy-number: loss (red), CNN-LOH (green), gain (blue), undetermined (grey). Darker coloring indicates higher allelic fraction. Multiple events within a single individual are plotted with the same y-coordinate (at the top of the plot). Note that events with unknown copy number also generally have greater uncertainty in their boundaries due to low allelic fraction.

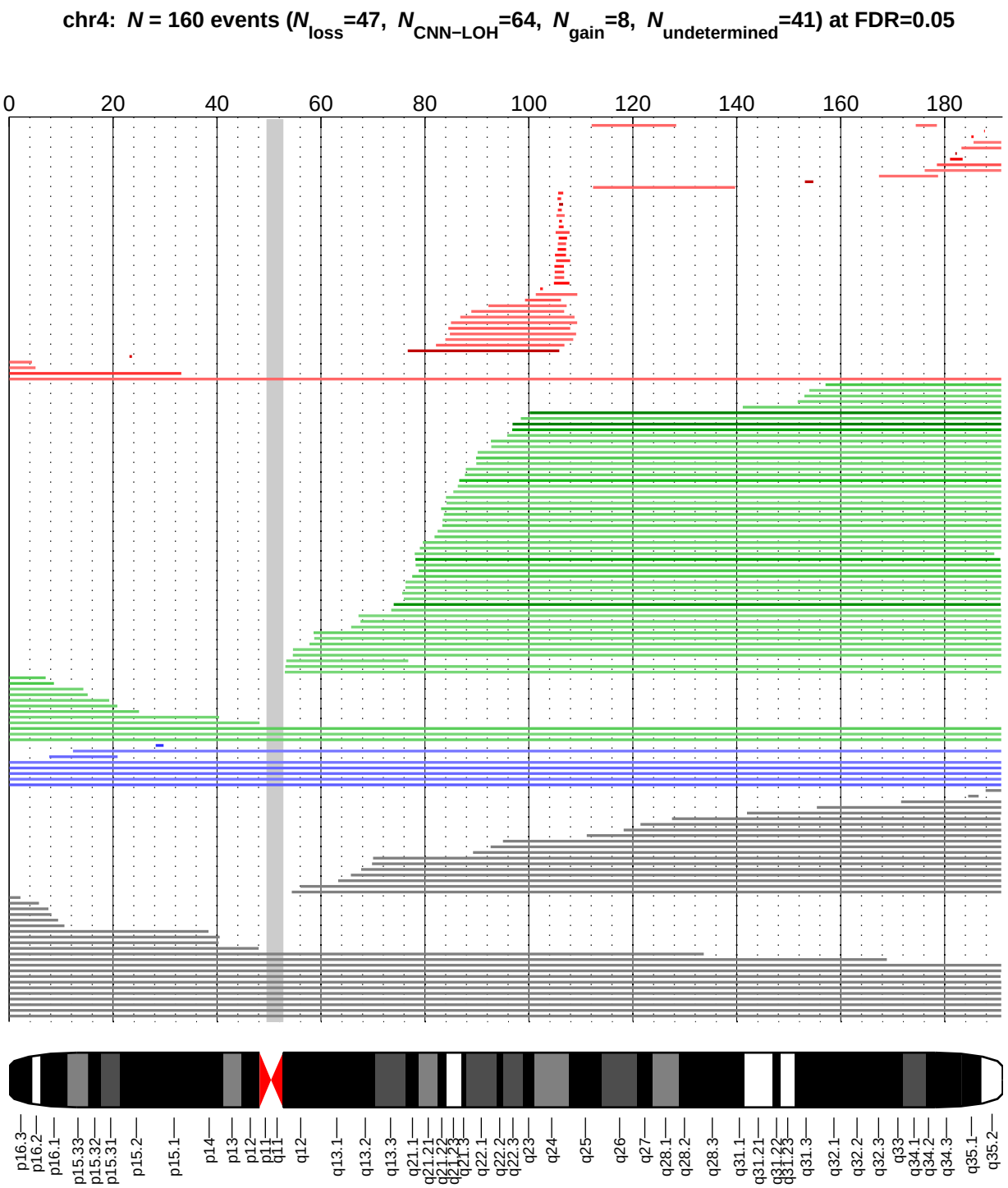


Figure S2-4. Detected mCAs on chromosome 4. Events are color-coded by copy-number: loss (red), CNN-LOH (green), gain (blue), undetermined (grey). Darker coloring indicates higher allelic fraction. Multiple events within a single individual are plotted with the same y-coordinate (at the top of the plot). Note that events with unknown copy number also generally have greater uncertainty in their boundaries due to low allelic fraction.

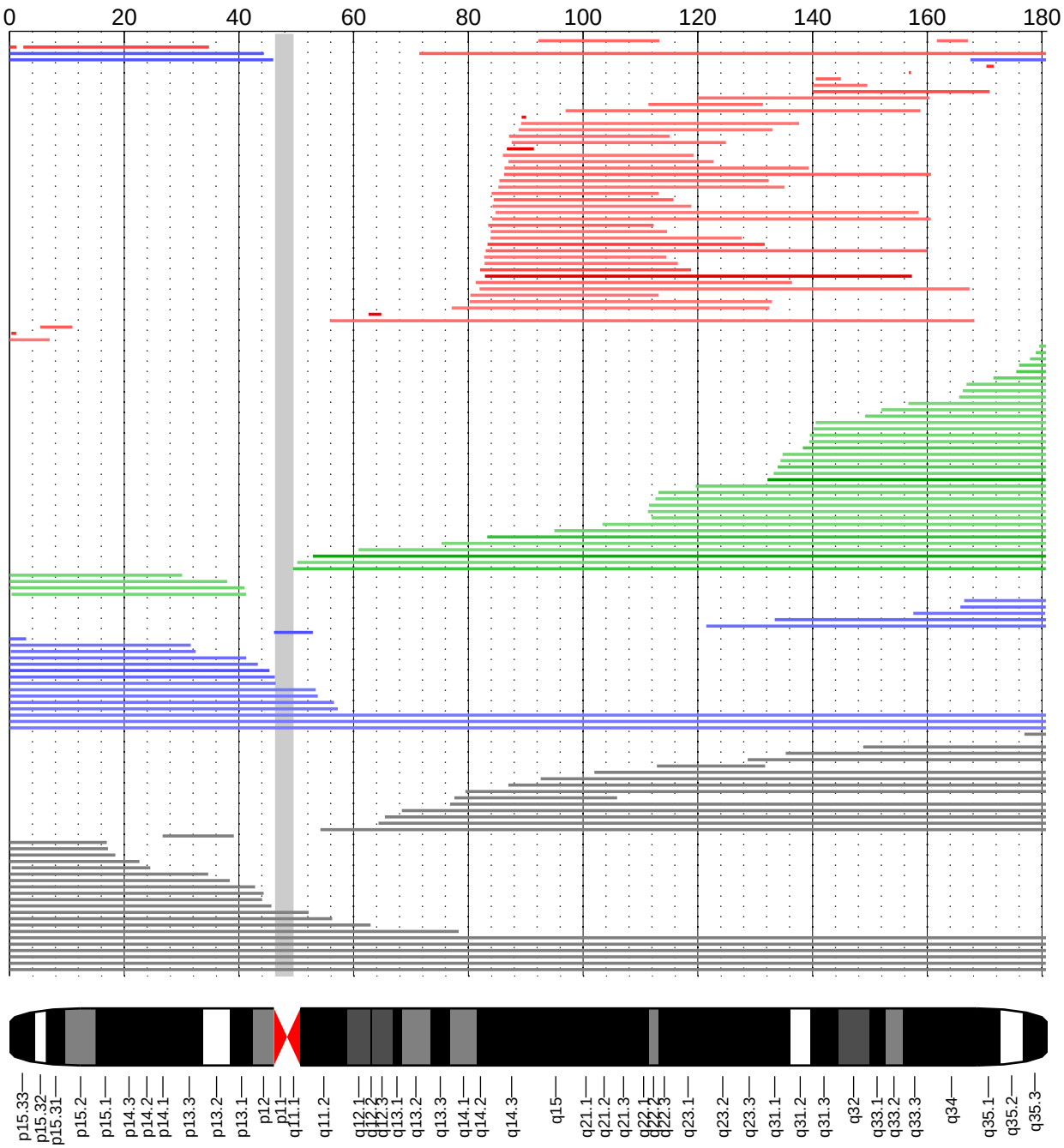


Figure S2-5. Detected mCAs on chromosome 5. Events are color-coded by copy-number: loss (red), CNN-LOH (green), gain (blue), undetermined (grey). Darker coloring indicates higher allelic fraction. Multiple events within a single individual are plotted with the same y-coordinate (at the top of the plot). Note that events with unknown copy number also generally have greater uncertainty in their boundaries due to low allelic fraction.

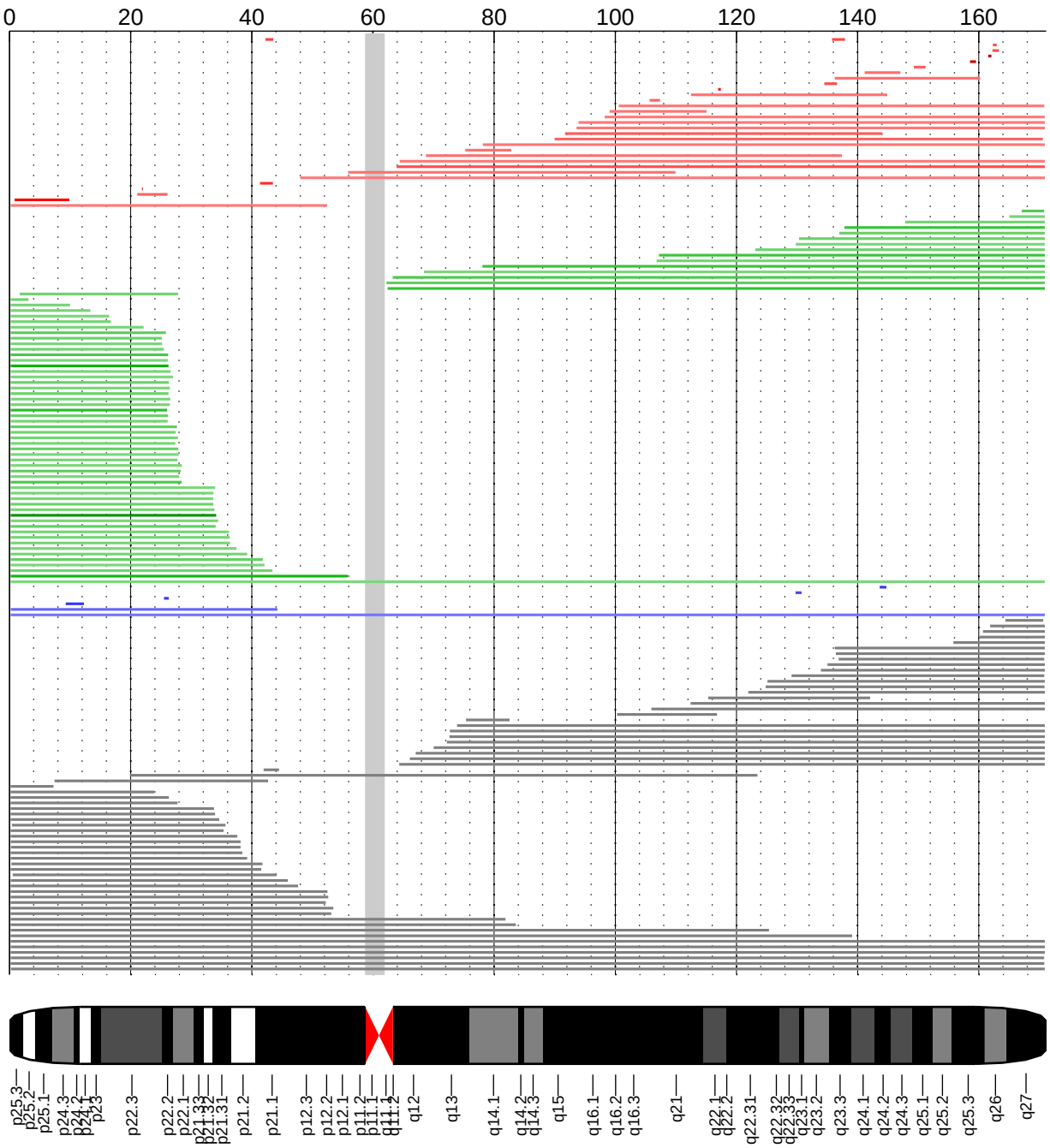


Figure S2-6. Detected mCAs on chromosome 6. Events are color-coded by copy-number: loss (red), CNN-LOH (green), gain (blue), undetermined (grey). Darker coloring indicates higher allelic fraction. Multiple events within a single individual are plotted with the same y-coordinate (at the top of the plot). Note that events with unknown copy number also generally have greater uncertainty in their boundaries due to low allelic fraction.

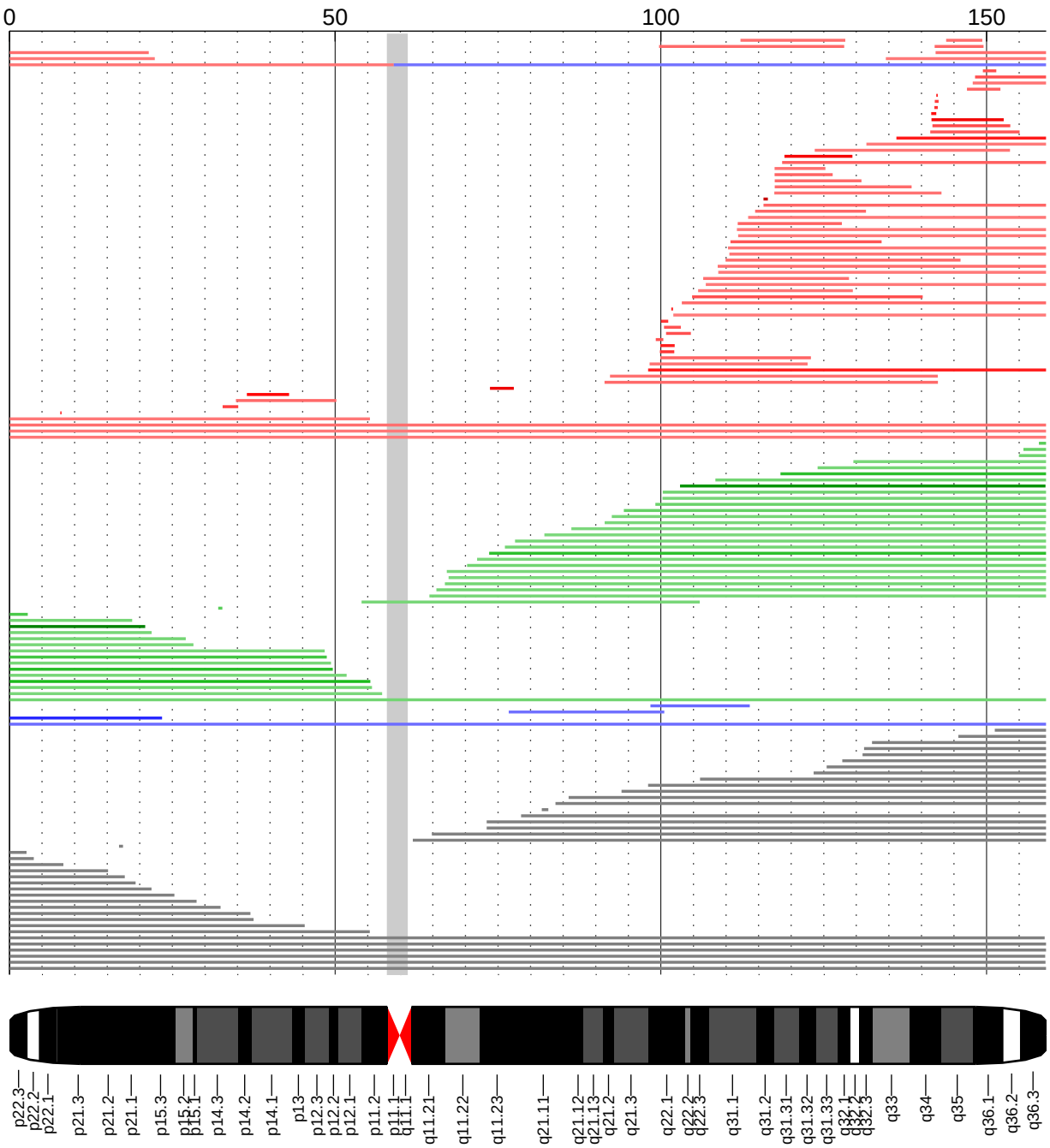


Figure S2-7. Detected mCAs on chromosome 7. Events are color-coded by copy-number: loss (red), CNN-LOH (green), gain (blue), undetermined (grey). Darker coloring indicates higher allelic fraction. Multiple events within a single individual are plotted with the same y-coordinate (at the top of the plot). Note that events with unknown copy number also generally have greater uncertainty in their boundaries due to low allelic fraction.

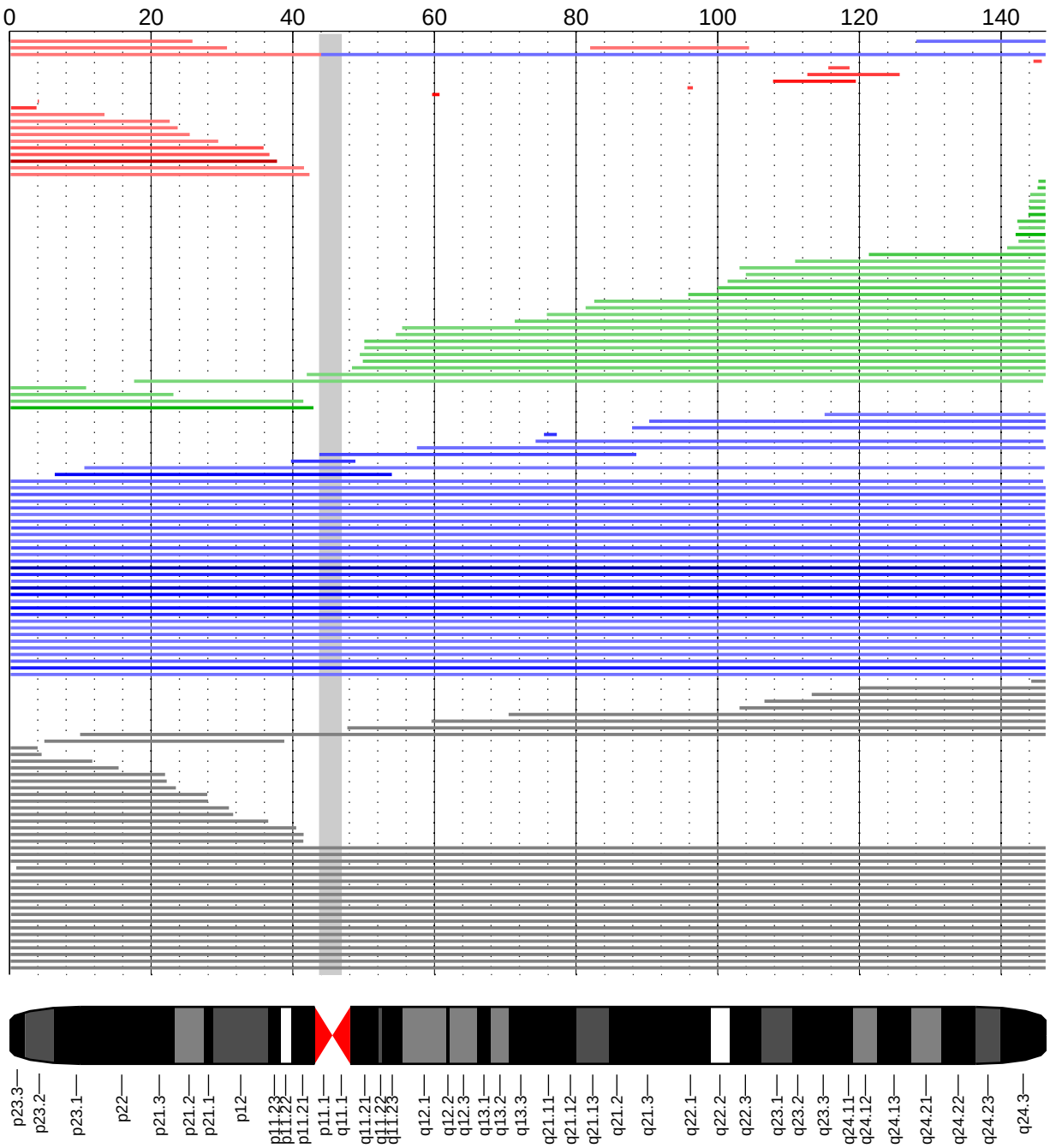


Figure S2-8. Detected mCAs on chromosome 8. Events are color-coded by copy-number: loss (red), CNN-LOH (green), gain (blue), undetermined (grey). Darker coloring indicates higher allelic fraction. Multiple events within a single individual are plotted with the same y-coordinate (at the top of the plot). Note that events with unknown copy number also generally have greater uncertainty in their boundaries due to low allelic fraction.

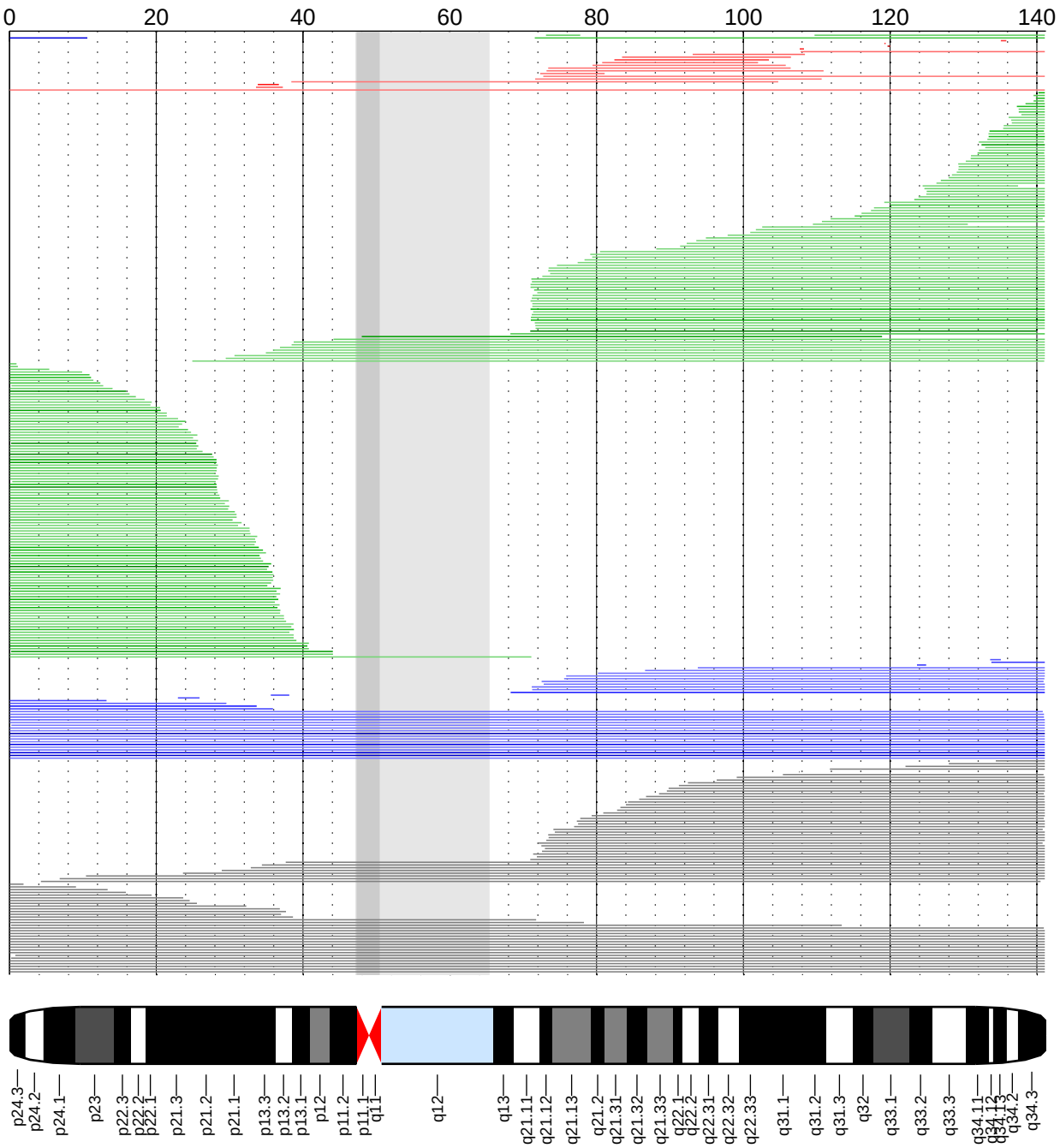


Figure S2-9. Detected mCAs on chromosome 9. Events are color-coded by copy-number: loss (red), CNN-LOH (green), gain (blue), undetermined (grey). Darker coloring indicates higher allelic fraction. Multiple events within a single individual are plotted with the same y-coordinate (at the top of the plot). Note that events with unknown copy number also generally have greater uncertainty in their boundaries due to low allelic fraction.

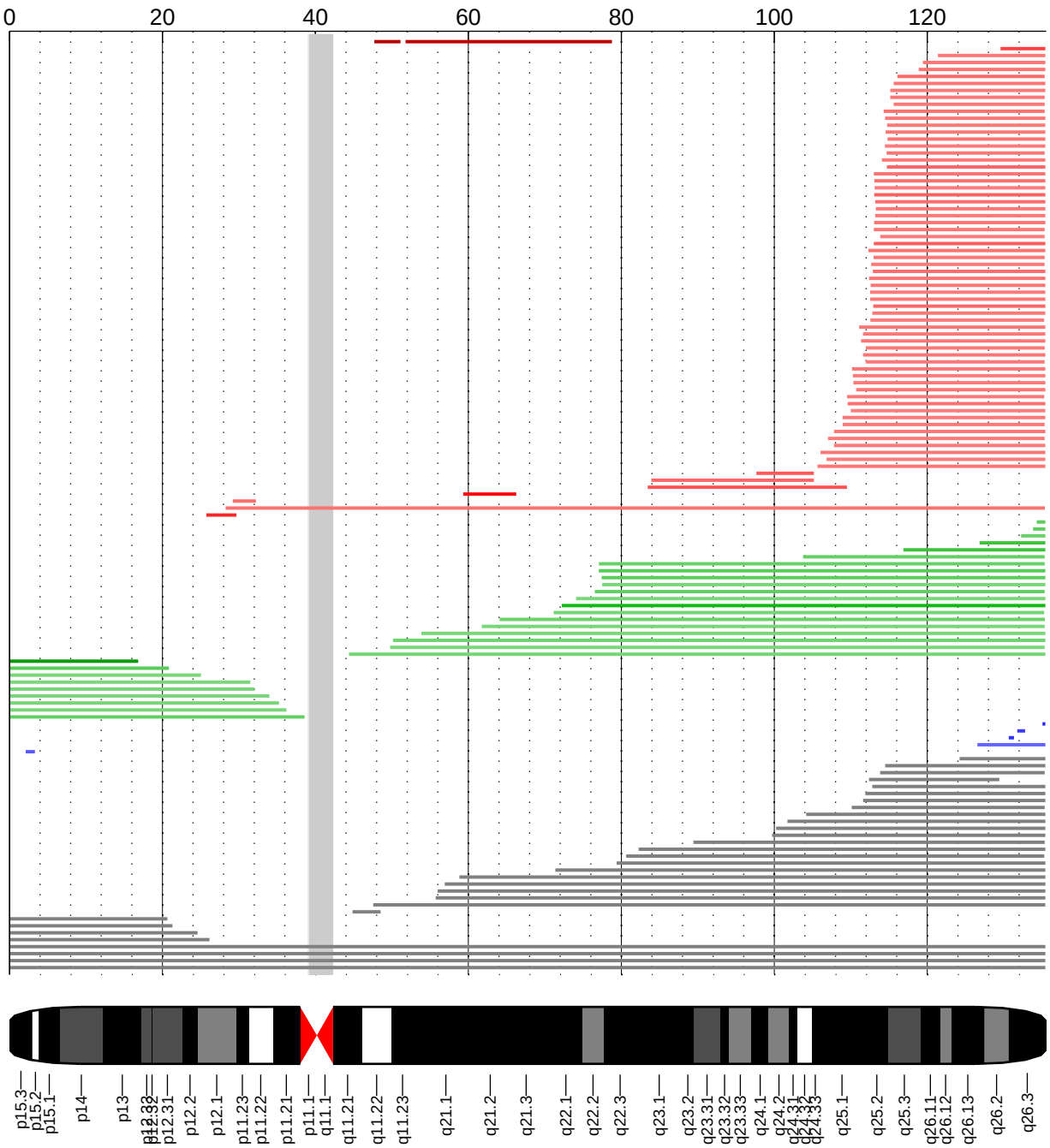


Figure S2-10. Detected mCAs on chromosome 10. Events are color-coded by copy-number: loss (red), CNN-LOH (green), gain (blue), undetermined (grey). Darker coloring indicates higher allelic fraction. Multiple events within a single individual are plotted with the same y-coordinate (at the top of the plot). Note that events with unknown copy number also generally have greater uncertainty in their boundaries due to low allelic fraction.

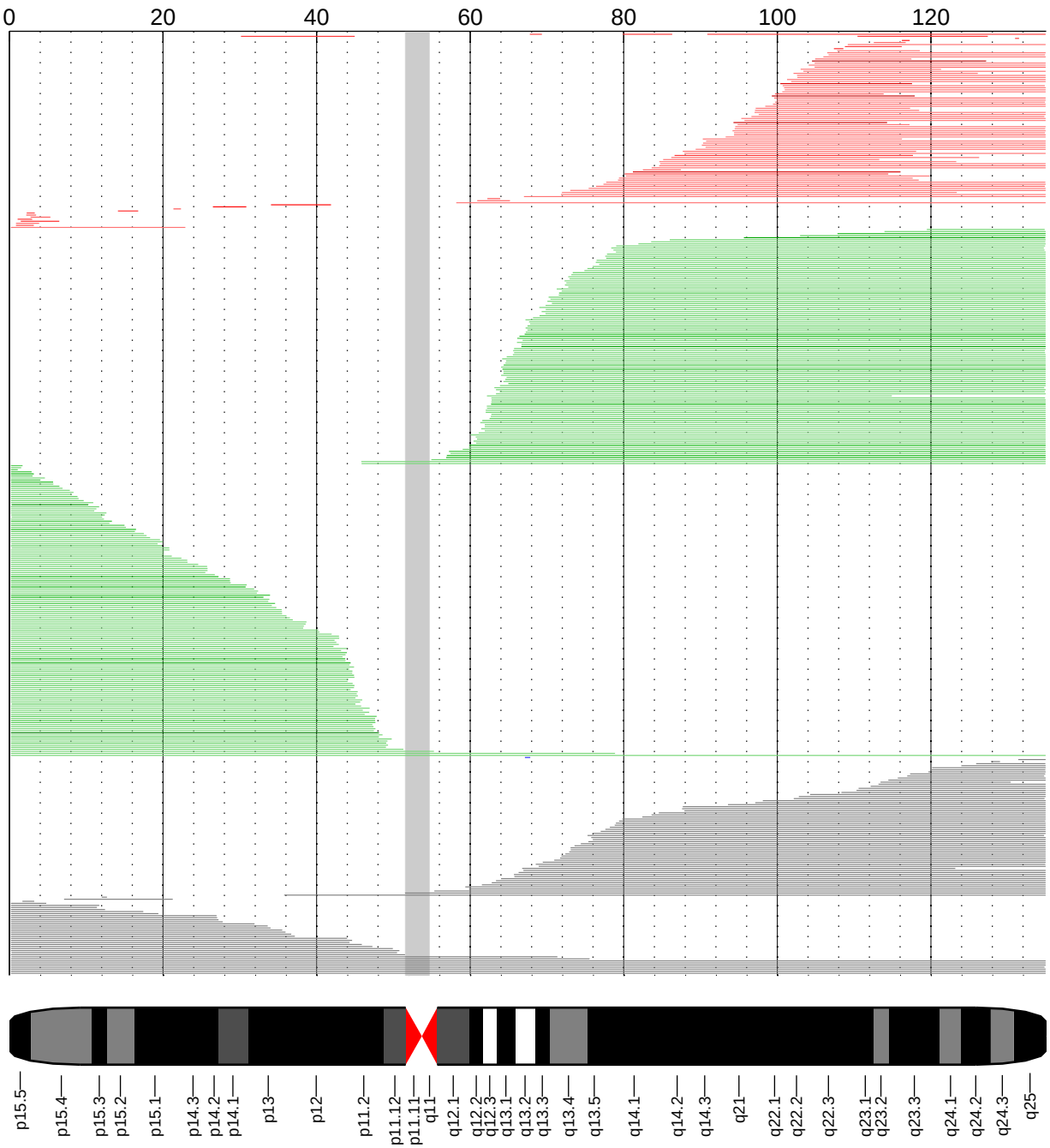


Figure S2-11. Detected mCAs on chromosome 11. Events are color-coded by copy-number: loss (red), CNN-LOH (green), gain (blue), undetermined (grey). Darker coloring indicates higher allelic fraction. Multiple events within a single individual are plotted with the same y-coordinate (at the top of the plot). Note that events with unknown copy number also generally have greater uncertainty in their boundaries due to low allelic fraction.

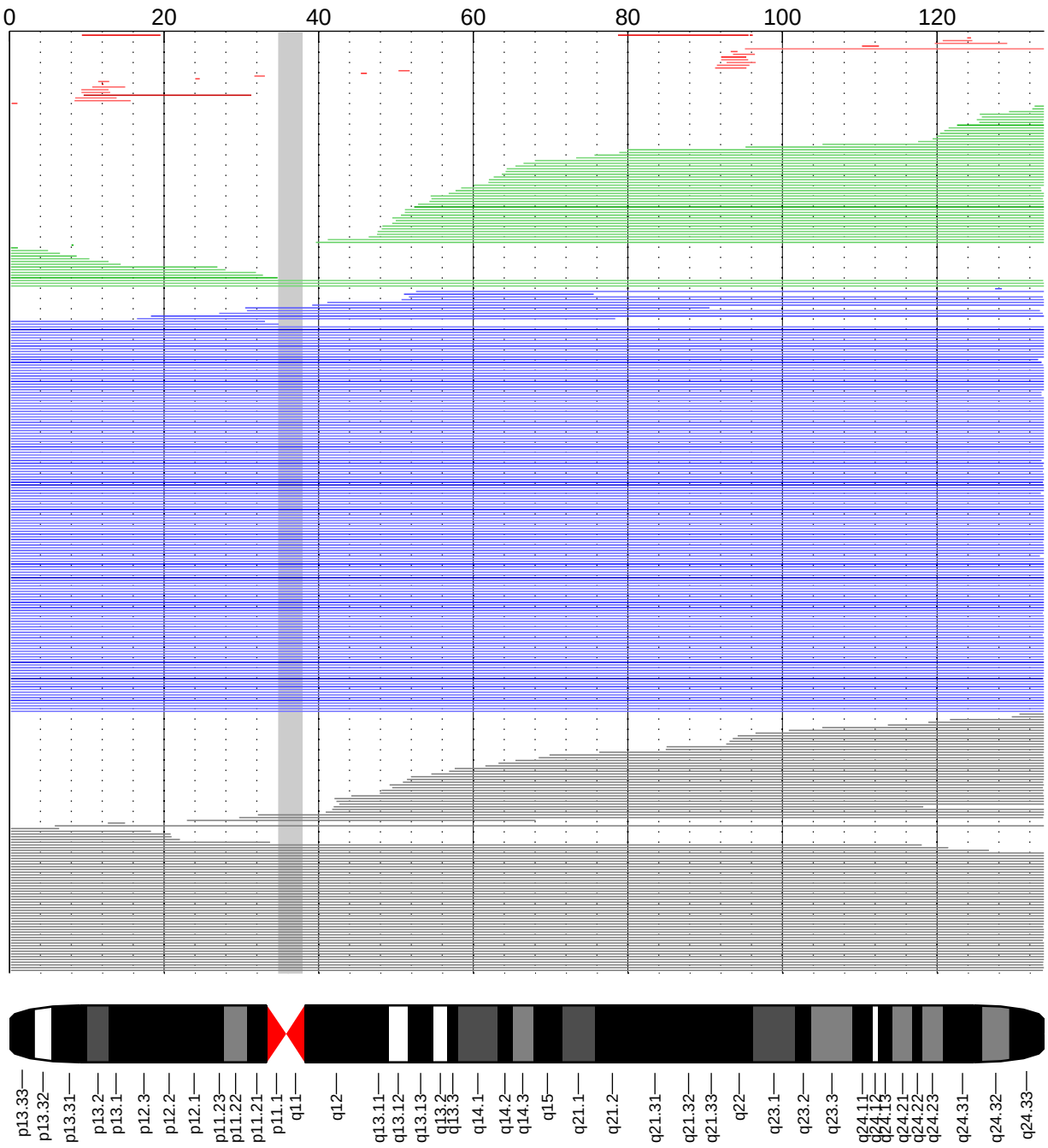


Figure S2-12. Detected mCAs on chromosome 12. Events are color-coded by copy-number: loss (red), CNN-LOH (green), gain (blue), undetermined (grey). Darker coloring indicates higher allelic fraction. Multiple events within a single individual are plotted with the same y-coordinate (at the top of the plot). Note that events with unknown copy number also generally have greater uncertainty in their boundaries due to low allelic fraction.

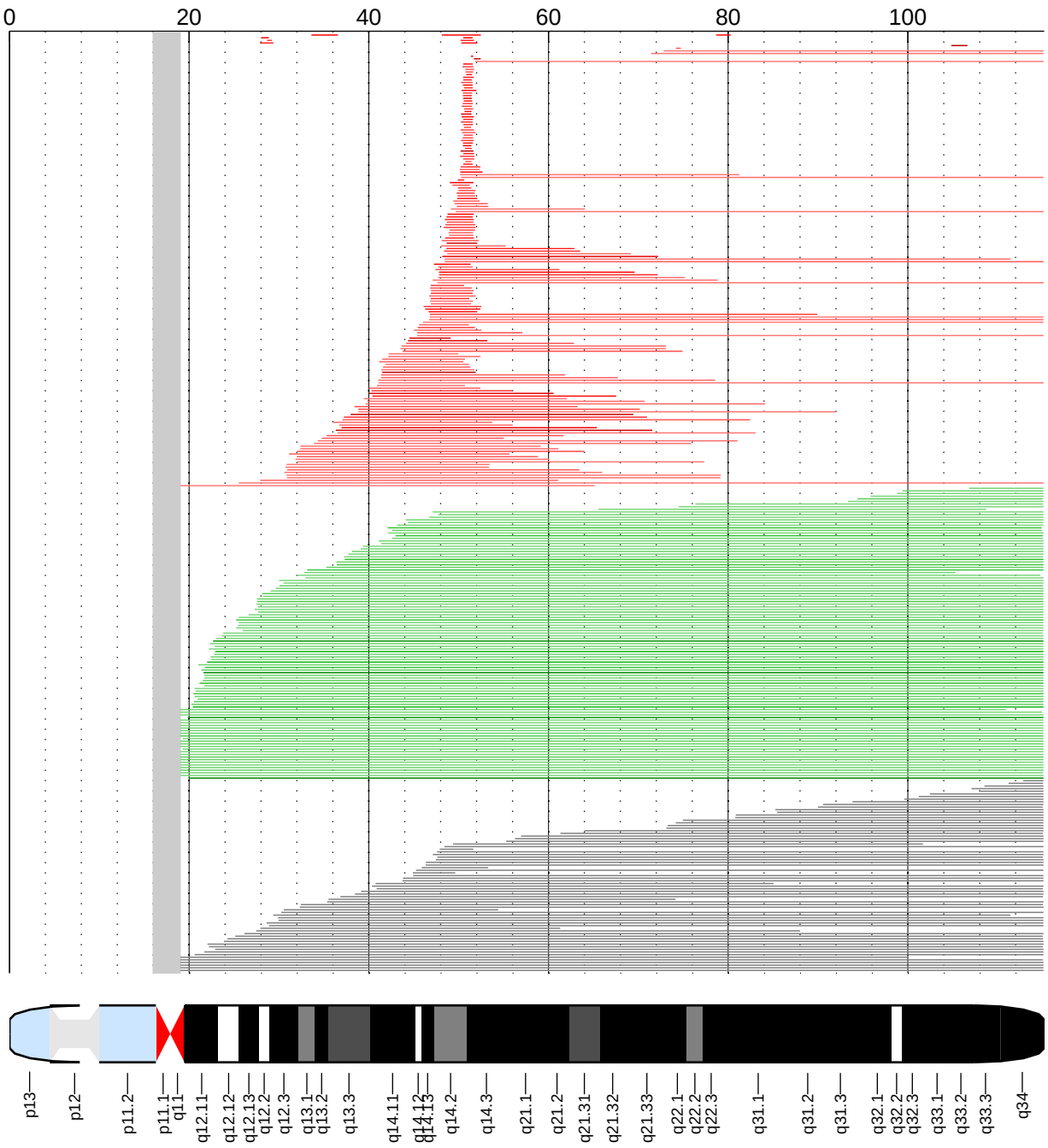


Figure S2-13. Detected mCAs on chromosome 13. Events are color-coded by copy-number: loss (red), CNN-LOH (green), gain (blue), undetermined (grey). Darker coloring indicates higher allelic fraction. Multiple events within a single individual are plotted with the same y-coordinate (at the top of the plot). Note that events with unknown copy number also generally have greater uncertainty in their boundaries due to low allelic fraction.

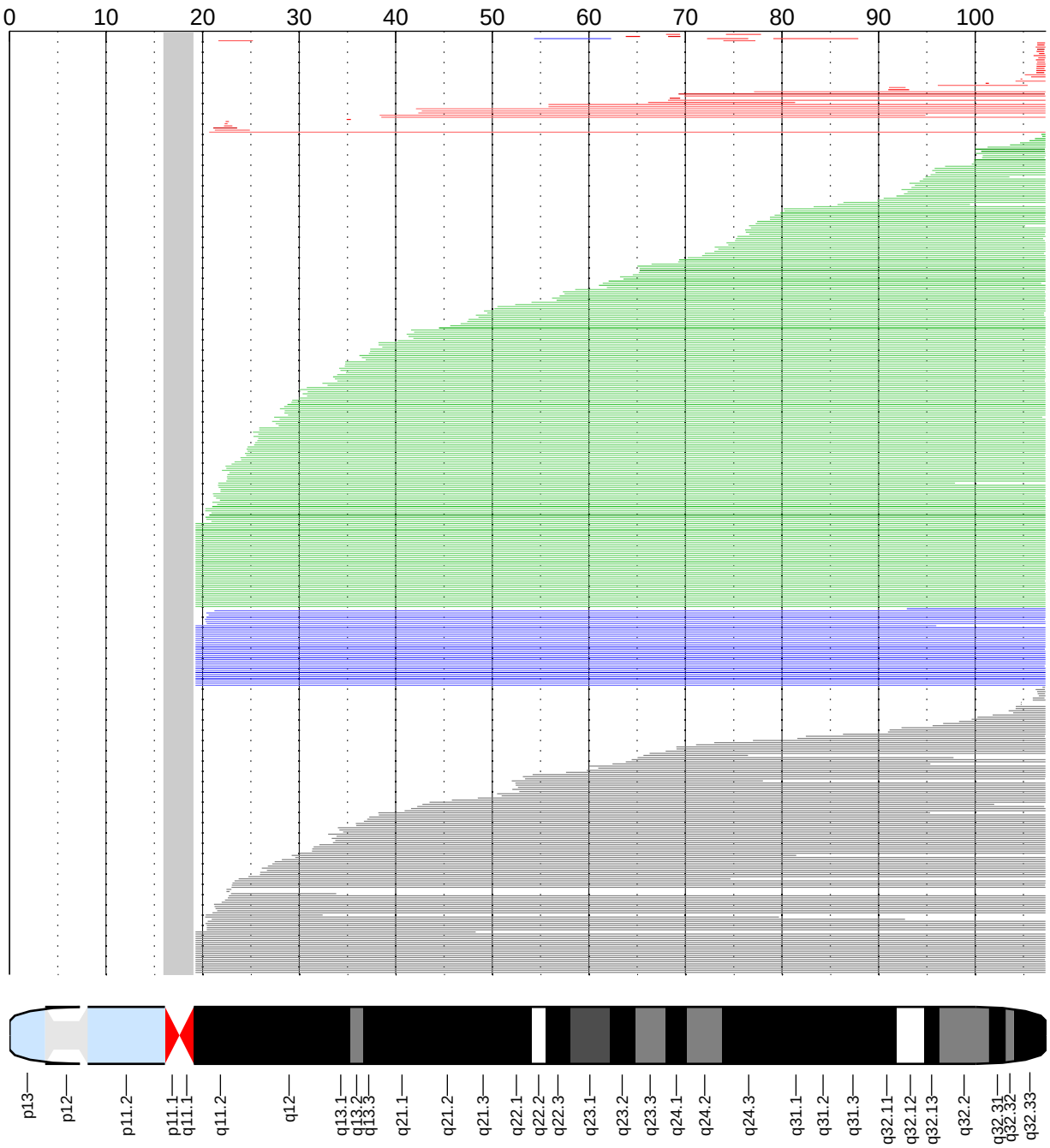


Figure S2-14. Detected mCAs on chromosome 14. Events are color-coded by copy-number: loss (red), CNN-LOH (green), gain (blue), undetermined (grey). Darker coloring indicates higher allelic fraction. Multiple events within a single individual are plotted with the same y-coordinate (at the top of the plot). Note that events with unknown copy number also generally have greater uncertainty in their boundaries due to low allelic fraction.

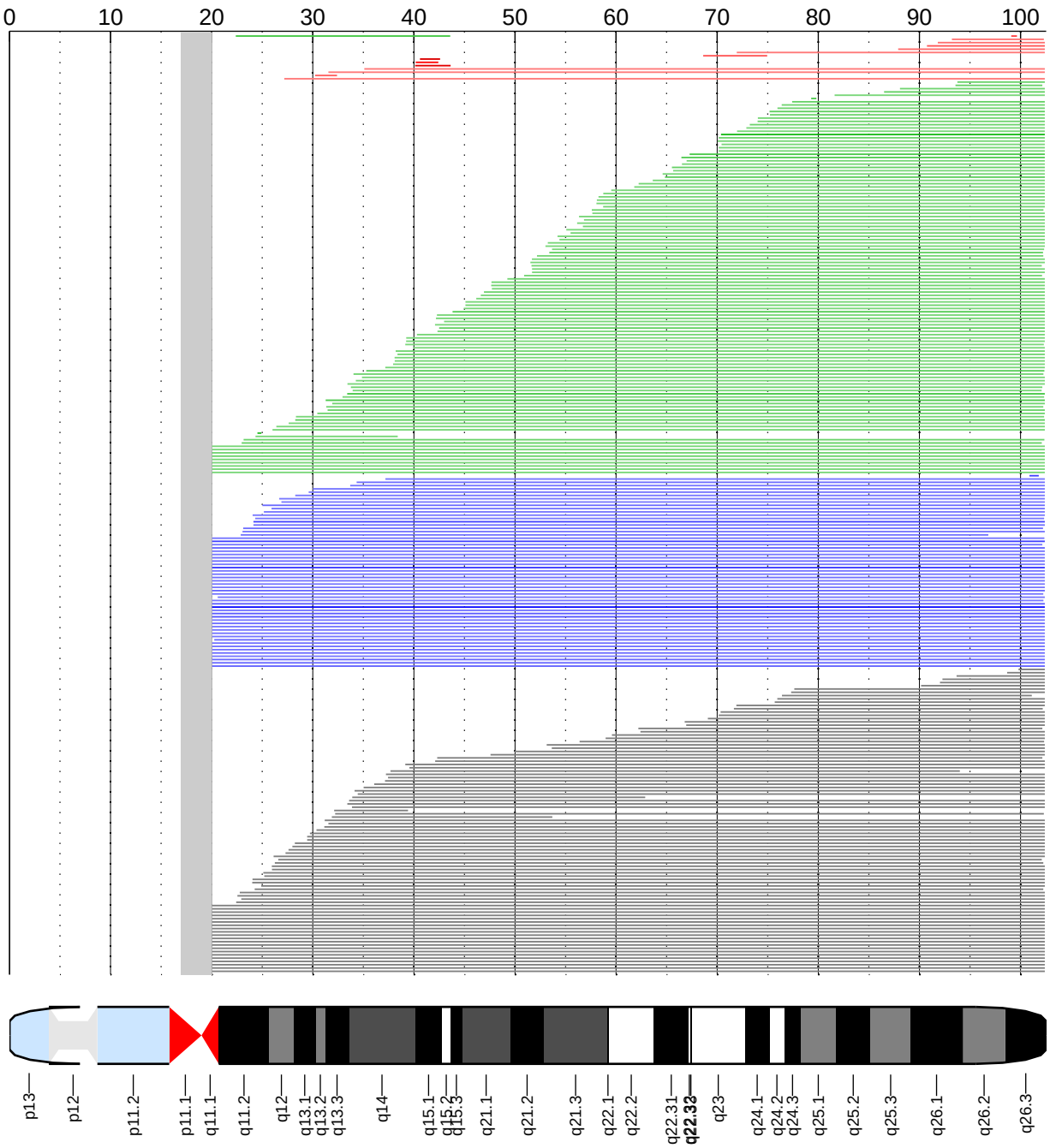


Figure S2-15. Detected mCAs on chromosome 15. Events are color-coded by copy-number: loss (red), CNN-LOH (green), gain (blue), undetermined (grey). Darker coloring indicates higher allelic fraction. Multiple events within a single individual are plotted with the same y-coordinate (at the top of the plot). Note that events with unknown copy number also generally have greater uncertainty in their boundaries due to low allelic fraction.

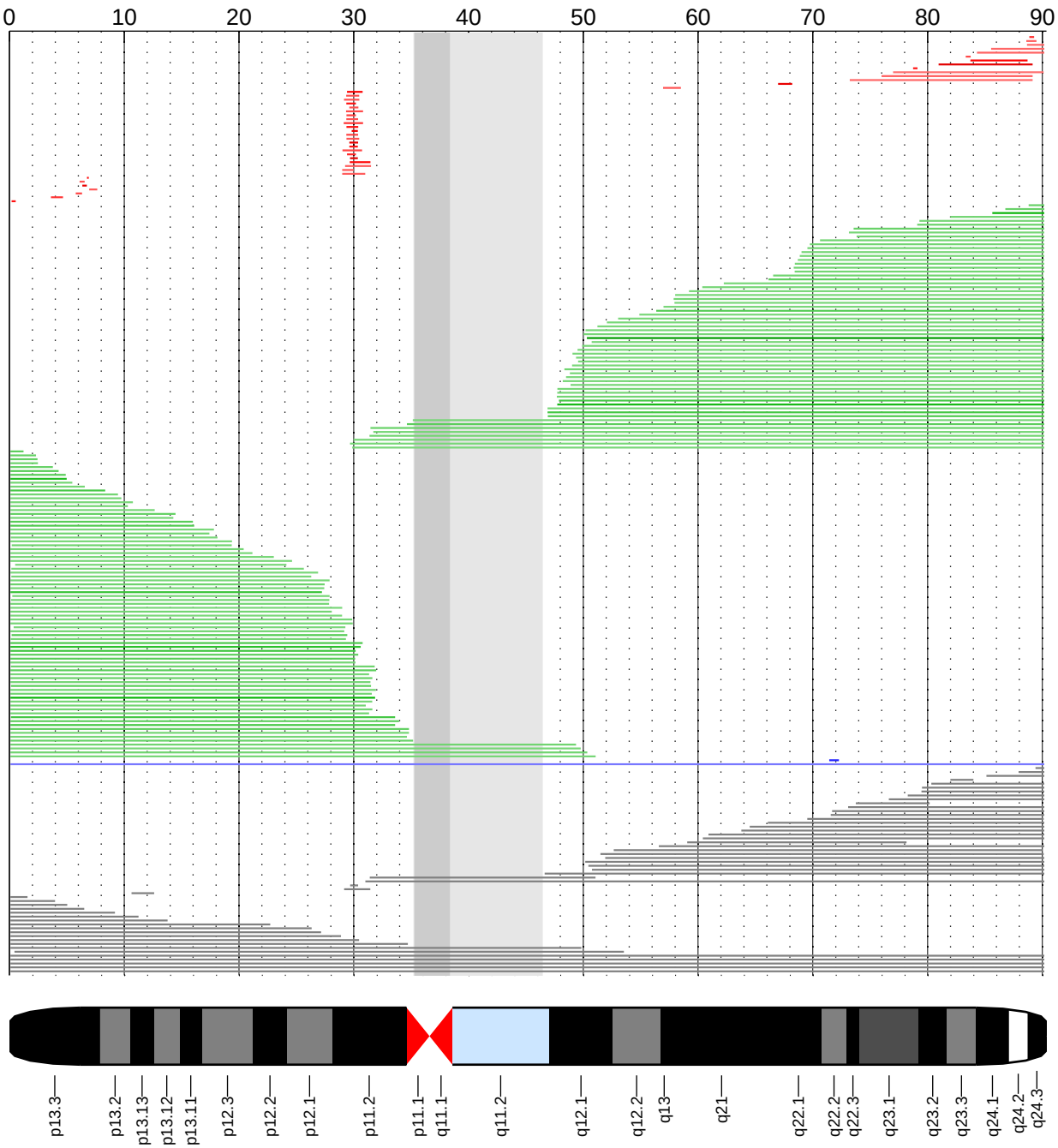


Figure S2-16. Detected mCAs on chromosome 16. Events are color-coded by copy-number: loss (red), CNN-LOH (green), gain (blue), undetermined (grey). Darker coloring indicates higher allelic fraction. Multiple events within a single individual are plotted with the same y-coordinate (at the top of the plot). Note that events with unknown copy number also generally have greater uncertainty in their boundaries due to low allelic fraction.

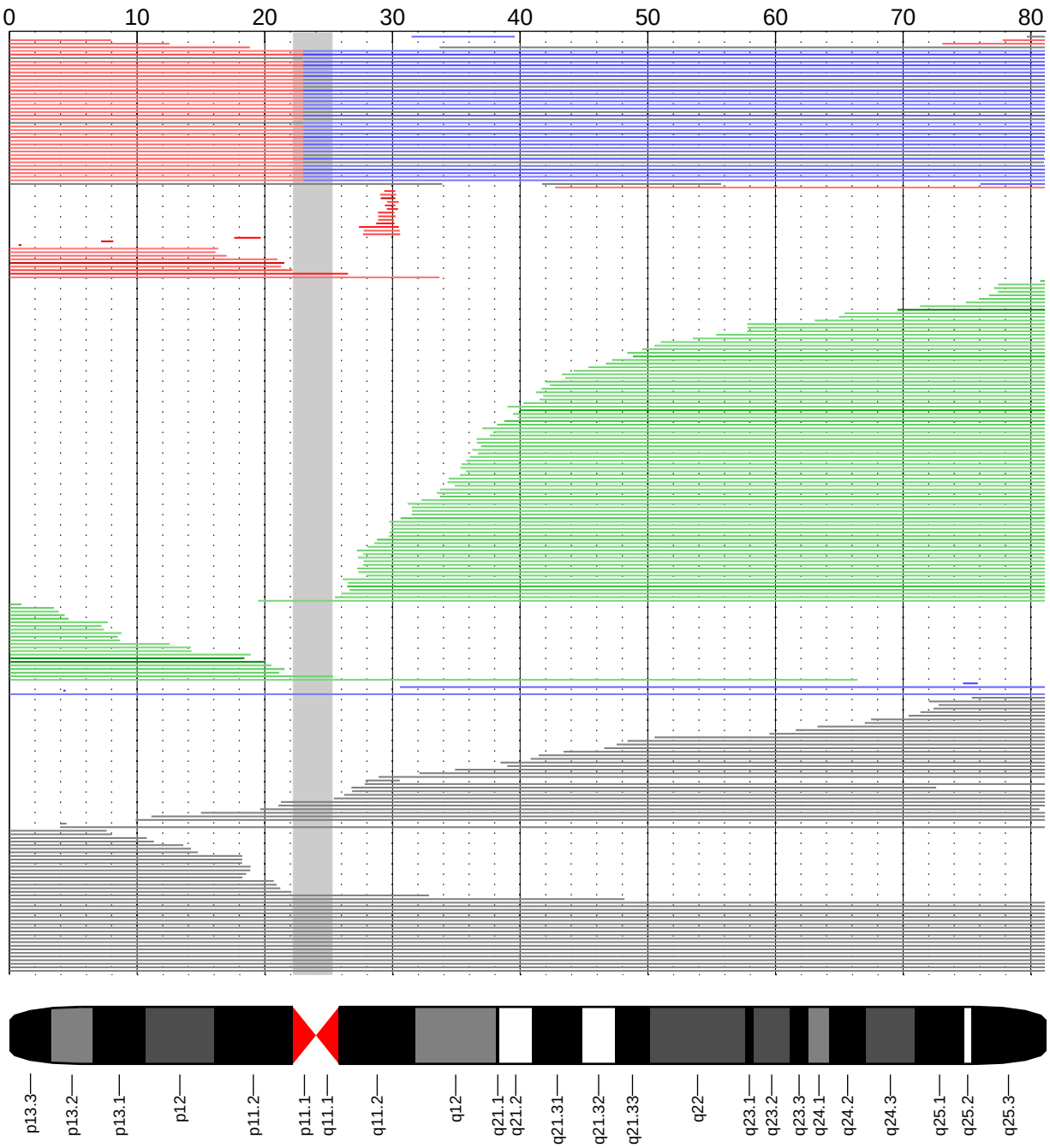


Figure S2-17. Detected mCAs on chromosome 17. Events are color-coded by copy-number: loss (red), CNN-LOH (green), gain (blue), undetermined (grey). Darker coloring indicates higher allelic fraction. Multiple events within a single individual are plotted with the same y-coordinate (at the top of the plot). Note that events with unknown copy number also generally have greater uncertainty in their boundaries due to low allelic fraction.

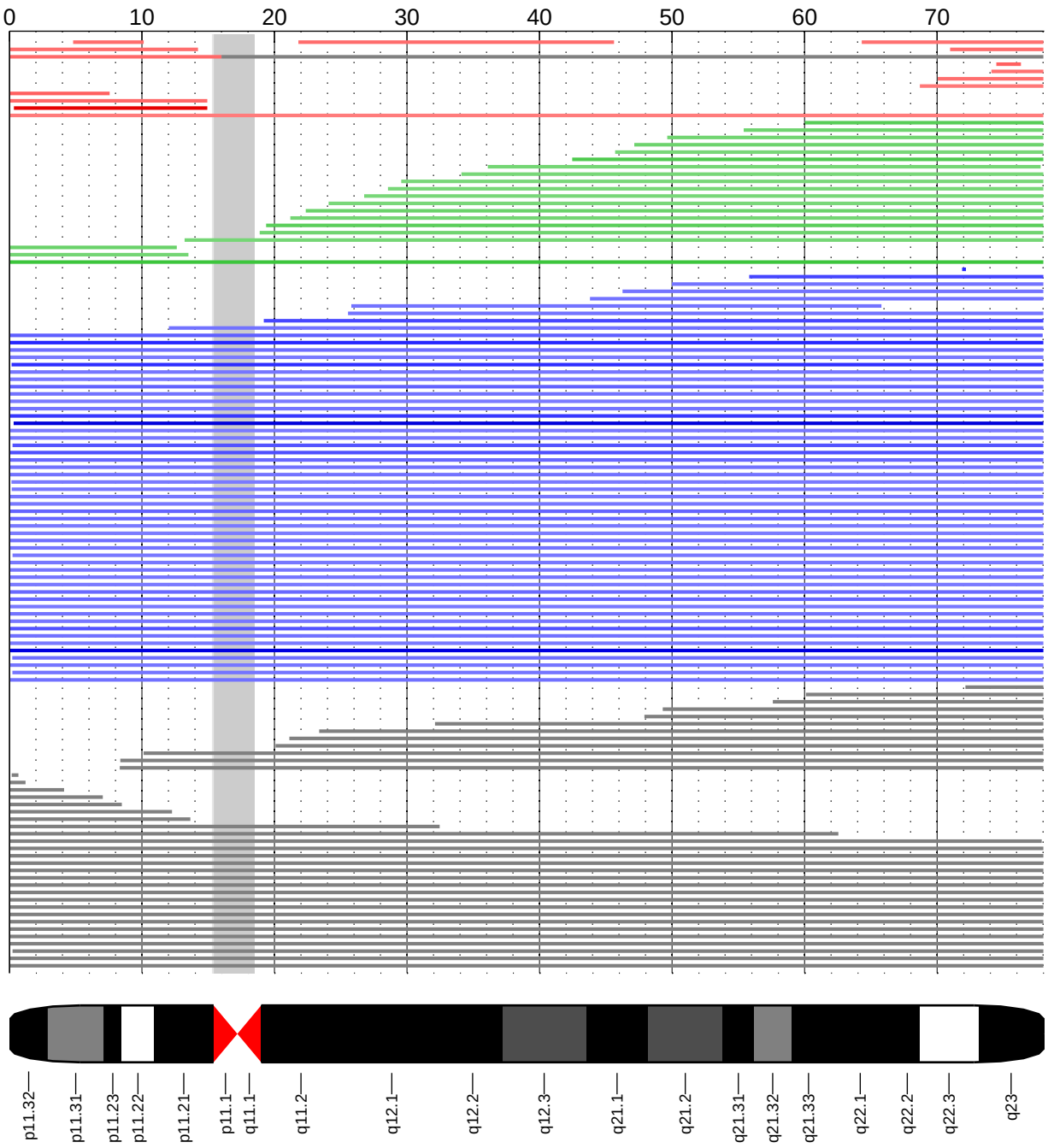


Figure S2-18. Detected mCAs on chromosome 18. Events are color-coded by copy-number: loss (red), CNN-LOH (green), gain (blue), undetermined (grey). Darker coloring indicates higher allelic fraction. Multiple events within a single individual are plotted with the same y-coordinate (at the top of the plot). Note that events with unknown copy number also generally have greater uncertainty in their boundaries due to low allelic fraction.

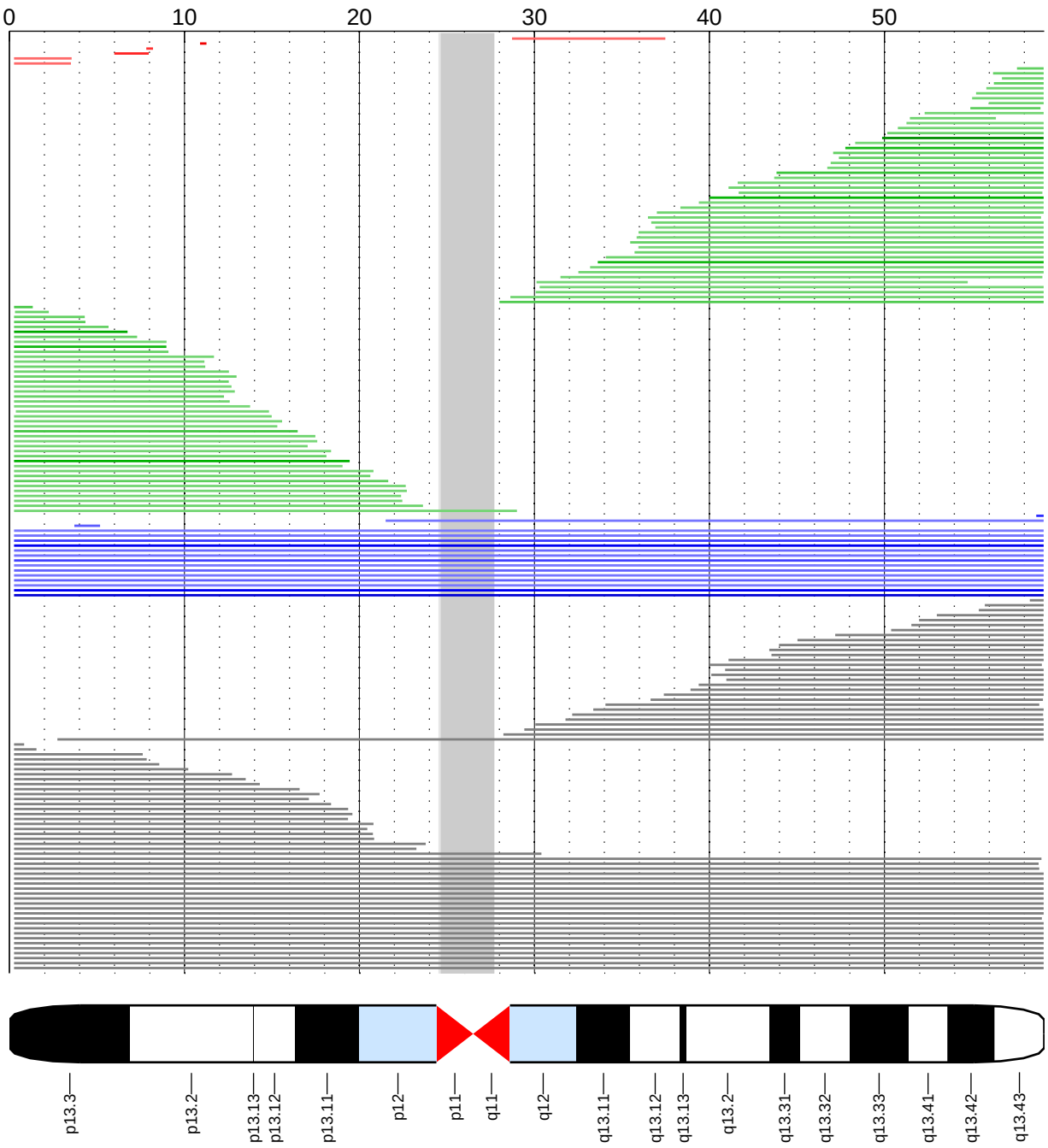


Figure S2-19. Detected mCAs on chromosome 19. Events are color-coded by copy-number: loss (red), CNN-LOH (green), gain (blue), undetermined (grey). Darker coloring indicates higher allelic fraction. Multiple events within a single individual are plotted with the same y-coordinate (at the top of the plot). Note that events with unknown copy number also generally have greater uncertainty in their boundaries due to low allelic fraction.

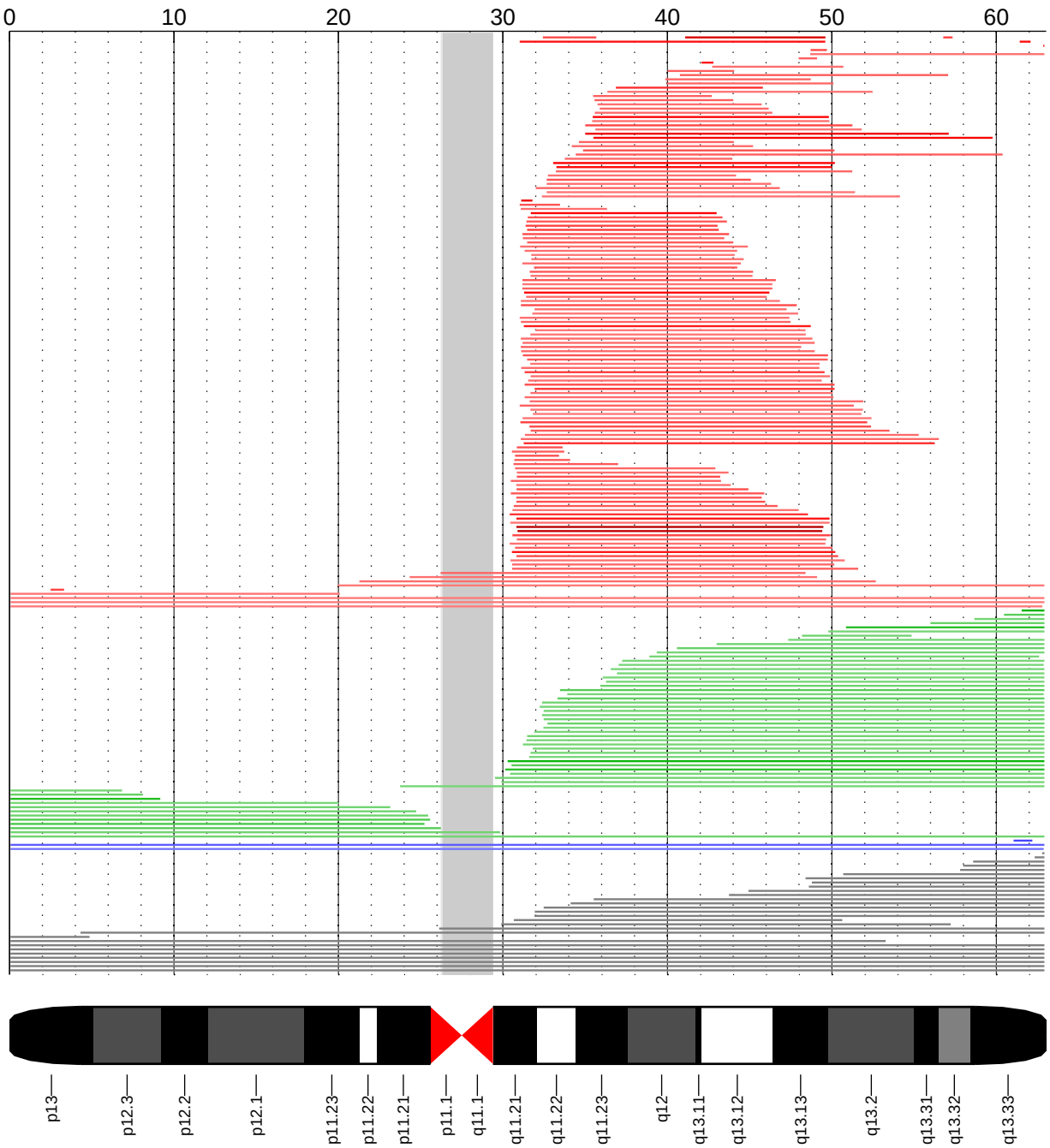


Figure S2-20. Detected mCAs on chromosome 20. Events are color-coded by copy-number: loss (red), CNN-LOH (green), gain (blue), undetermined (grey). Darker coloring indicates higher allelic fraction. Multiple events within a single individual are plotted with the same y-coordinate (at the top of the plot). Note that events with unknown copy number also generally have greater uncertainty in their boundaries due to low allelic fraction.

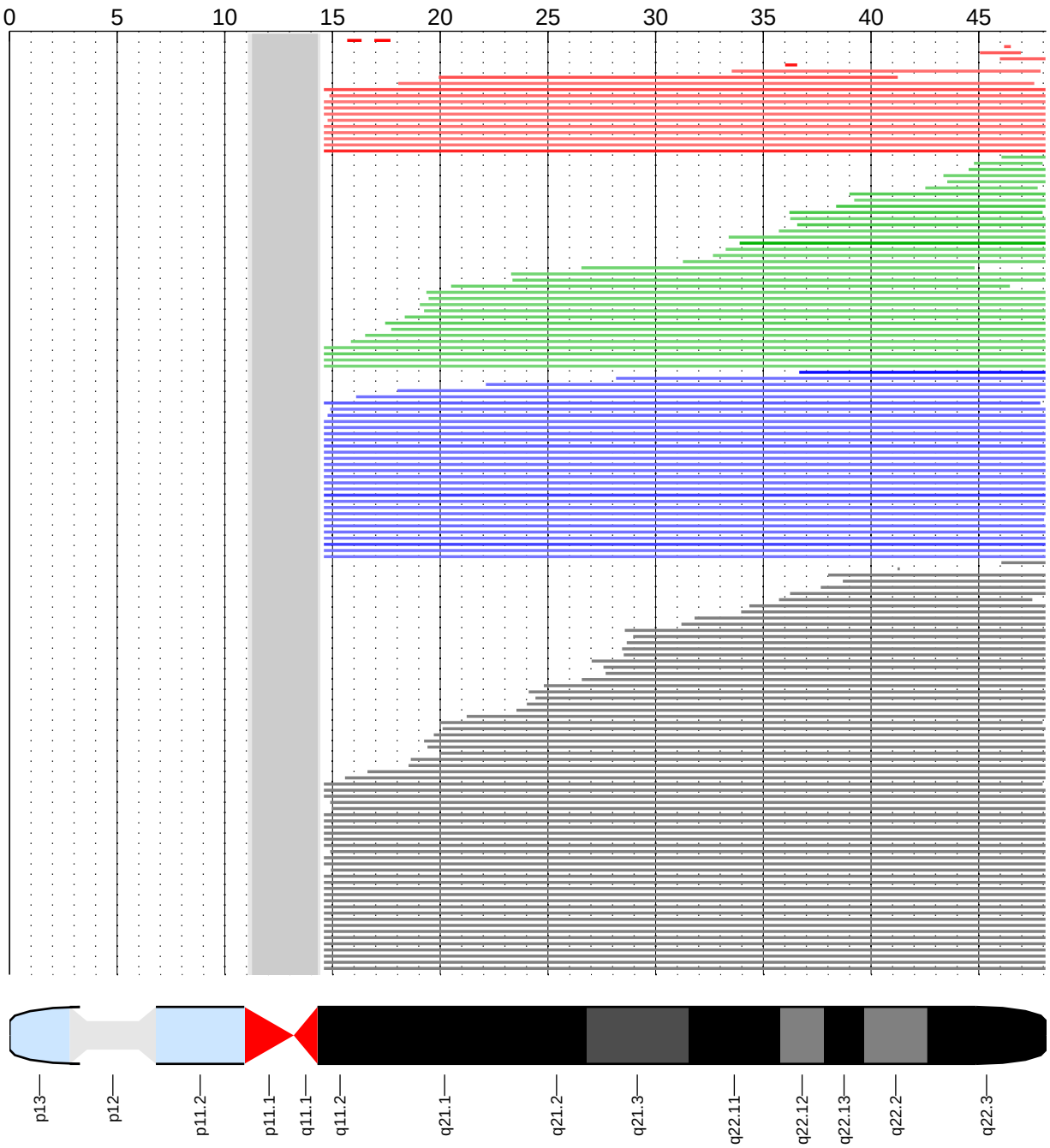


Figure S2-21. Detected mCAs on chromosome 21. Events are color-coded by copy-number: loss (red), CNN-LOH (green), gain (blue), undetermined (grey). Darker coloring indicates higher allelic fraction. Multiple events within a single individual are plotted with the same y-coordinate (at the top of the plot). Note that events with unknown copy number also generally have greater uncertainty in their boundaries due to low allelic fraction.

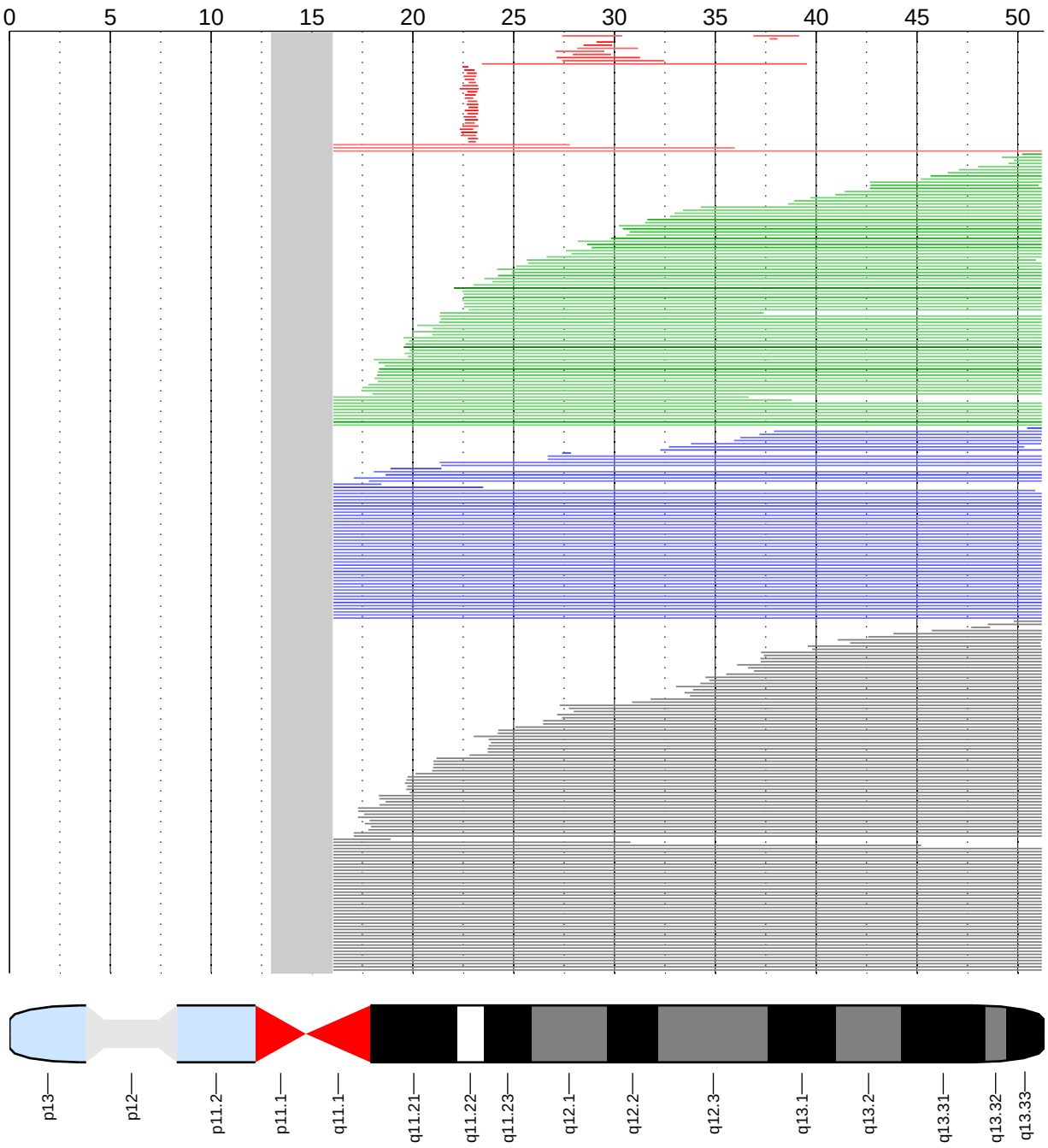


Figure S2-22. Detected mCAs on chromosome 22. Events are color-coded by copy-number: loss (red), CNN-LOH (green), gain (blue), undetermined (grey). Darker coloring indicates higher allelic fraction. Multiple events within a single individual are plotted with the same y-coordinate (at the top of the plot). Note that events with unknown copy number also generally have greater uncertainty in their boundaries due to low allelic fraction.

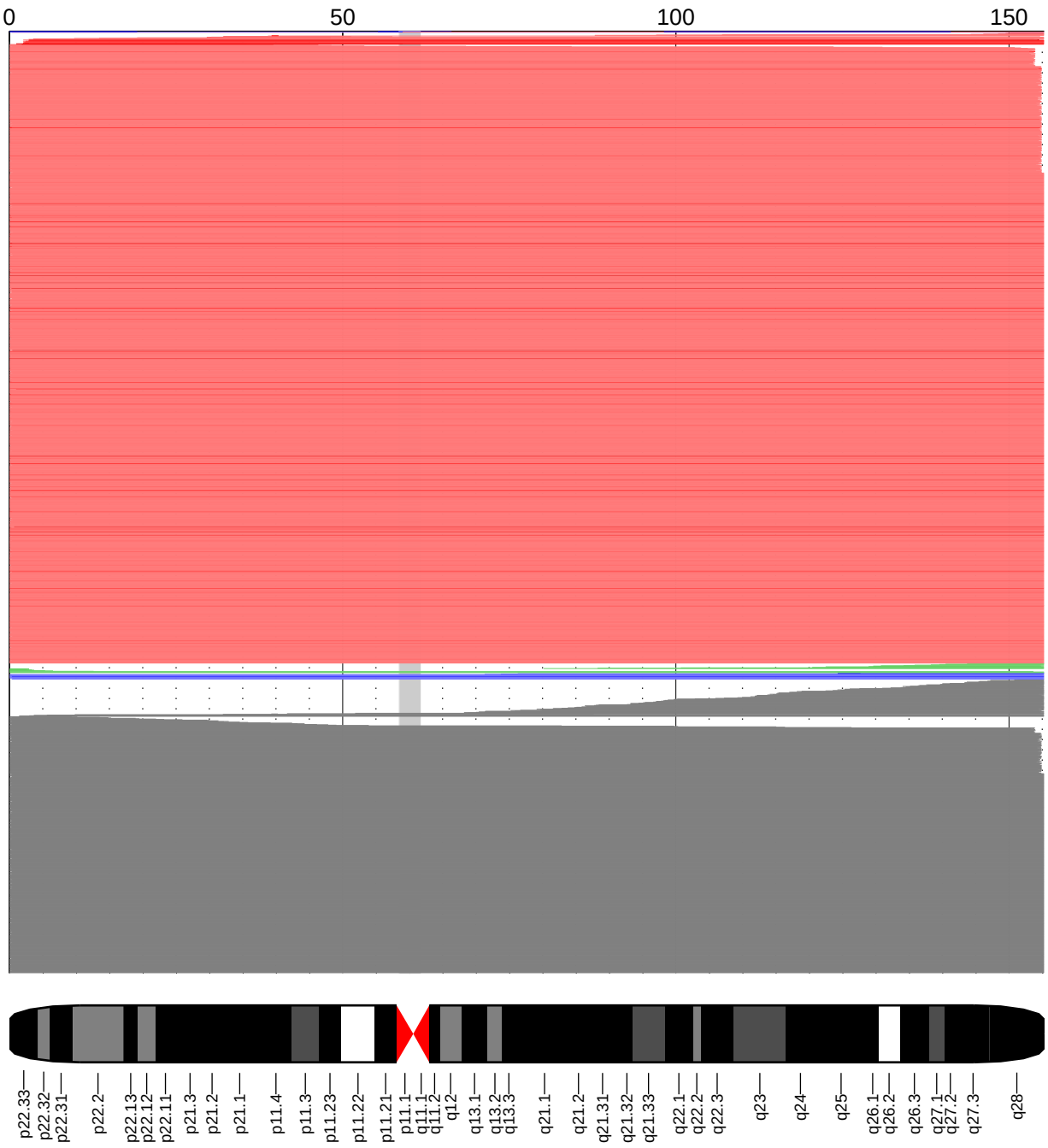


Figure S2-23. Detected mCAs on chromosome X. Events are color-coded by copy-number: loss (red), CNN-LOH (green), gain (blue), undetermined (grey). Darker coloring indicates higher allelic fraction. Multiple events within a single individual are plotted with the same y-coordinate (at the top of the plot). Note that events with unknown copy number also generally have greater uncertainty in their boundaries due to low allelic fraction.

3 Confirmatory analyses for event calls

While performing direct molecular validation of our mosaic event calls would have been ideal, we were unable to do so as we did not have access to the original DNA samples. We therefore conducted a series of confirmatory analyses aimed at (i) validating our false positive control and (ii) replicating our GWAS results and distributional results.

3.1 Estimation of true false discovery rate

Our procedure for calling the existence of a mosaic event (Supplementary Note 1.4) involved identifying significant autocorrelation in phased BAF deviations using a likelihood ratio test statistic. We calibrated these test statistics empirically using a permutation-based procedure (phase randomization) to obtain a nominal 5% false discovery rate (FDR) threshold. However, this permutation-based 5% FDR threshold assumed that the only source of autocorrelation in phased BAF is a true mosaic event. In reality, other sources of autocorrelation exist; in particular, we found that sample contamination produced autocorrelation in regions of long-range LD (resulting in unusual false positive calls that we subsequently filtered). While we believe that our filtering eliminated most samples affected by spurious autocorrelation, our true FDR is likely to be slightly larger than 5% due to residual artifacts.

Fortunately, we can estimate our true FDR by leveraging the fact that true-positive events should be observed more frequently in the genomes of older people, while false-positive calls (which have no relation to age) should be observed in individuals whose age distribution matches that of the study population. This observation allows us to estimate FDR by comparing the age distributions of the highest-confidence calls (6,543 calls passing a permutation-based FDR of 1%) vs. medium-confidence calls (1,797 additional calls passing a permutation-based FDR of 5% when combined with the high-confidence calls, but failing the 1% threshold). The medium-confidence call set is expected to have a false positive rate of ≈20% based on the permutation-based FDRs—meaning that its age distribution is expected to be an 80:20 mixture of (i) the age distribution of high-confidence calls and (ii) the age distribution of the study population. That is, the age distribution of medium-confidence calls should relax toward the age distribution of the overall study due to the inclusion of false positives—which is precisely what we see (Extended Data Fig. 2). (The figure also includes low-confidence calls at FDR 10% for additional context, although we did not analyze these calls.)

Upon fitting the age distribution of medium-confidence calls as a mixture of the age distribution of high-confidence calls and the overall study distribution, the regression fit gives mixture proportions of ≈70:30 rather than 80:20, implying a true FDR of 7.5% (6.2–8.8%, 95% CI) when combined with the high-confidence calls—slightly higher than the permutation-based FDR of 5%, as expected. We note that this estimate is contingent on two assumptions: (i) the high-confidence

call set predominantly contains true positives (which is supported by the observation that changing the high-confidence FDR threshold from 1% to 0.1% results in a near-identical "gold standard" age distribution and inferred true FDR of 7.1% (5.8–8.5%); and (ii) the true positives in the high-confidence and medium-confidence call set have the same age distribution. While we acknowledge that these assumptions are imperfect, this analysis gives good evidence that our FDR is well-controlled. (We also note that while we cannot completely rule out the possibility that our FDR is higher than we estimated, the key results of our paper are robust to higher FDRs than estimated; e.g., we would only expect a higher-than-estimated FDR to weaken GWAS associations and decrease effect sizes.)

3.2 Allelic evidence for validity of 10q event calls

We can also provide one other line of evidence giving us confidence in our FDR control: in chromosome 10q, our event calls display striking specificity for the FRA10B risk haplotype that appears to be required for 10q25 breakage. Of 69 event calls on 10q with estimated breakpoints near FRA10B (104–122Mb) and extending to the q-telomere, 60/60 loss calls carry the rs118137427:G risk allele (RAF=5%), 0/1 CNN-LOH calls carry the risk allele, and 7/8 calls with undetermined copy number carry the risk allele. In contrast, false positive calls (and CNN-LOH calls) would have a 90% chance of being homozygous for the non-risk allele—providing strong evidence that our FDR control is working as expected. (This analysis was uniquely possible for the del(10q) association; the other loci we identified were associated with CNN-LOH events, only a minority of which were related to risk alleles.)

3.3 Replication of distributional results

We confirmed our results concerning the age and sex distributions of particular events by analyzing the largest previously published tables of event calls (Jacobs et al. [1], Laurie et al. [2], and Vattathil & Scheet [8]).

- **del(10q) individuals are younger (vs. other events) and are predominantly female.** We replicated this finding in the Vattathil & Scheet data set (P=0.003; binomial test for enrichment of <50-year-old females). Specifically, we identified three individuals with event calls matching the profile of the del(10q) events of interest (estimated left breakpoint near *FRA10B*, right endpoint at the q-telomere, negative mean LRR), and we compared the age and sex of these individuals to the rest of the Vattathil & Scheet call set. (The Jacobs et al. and Laurie et al. data sets did not contain any such calls, which was not surprising given that most del(10q) events have cell fractions <5%, below the limit of these studies' sensitivity; see Supplementary Note 5.2.)

- **del(16p11.2) individuals are younger (vs. other events) and are predominantly female.** Although we were underpowered to replicate this finding, we found support for it in the Laurie et al. data set, albeit only from one individual. Specifically, we identified one del(16p11.2) event in the Laurie et al. data, which was in a 37-year-old female. For context, the Laurie et al. call set was 72% male, and an age of 37 corresponded to the 13th percentile of the call set. (The published data sets we analyzed contained two other 16p11.2 event calls, but both appeared to be 16p11.2 duplications.)

- **CNN-LOH events do not show a male skew (unlike losses and gains).** We replicated the sex difference between CNN-LOH vs. other events in the Jacobs et al. data (P=0.001) and Laurie et al. data (P=0.1). (The Vattathil & Scheet call set only includes 30 CNN-LOH calls.) In our UK Biobank calls, we also noticed a trend for CNN-LOH events to occur in younger individuals vs. losses and gains, but this age difference did not replicate (P=0.5 in Jacobs et al., P=0.2 in Laurie et al.). One possible reason that the age difference failed to replicate is limited power; however, another possible reason is that our analysis was more sensitive to CNN-LOH events than other events (because CNN-LOH events cause twice as large a BAF deviation), resulting in more detections of small clones in younger individuals. Supplementary Table 14 provides a comparison of age and sex for loss, CNN-LOH, and gain events across studies.

3.4 Replication of GWAS results

We replicated the associations we identified at 10q and 15q (Table 1) in the WGS cohort (2,079 people). (The other associations we found are too weak to replicate in a cohort of that size; e.g., $<$1 *MPL*-associated 1p CNN-LOH event and $<$1 *ATM*-associated 11q CNN-LOH event are expected.) Specifically, restricting our analysis to the unrelated parents in the WGS cohort, we replicated the association of rs118137427 with *FRA10B*-related 10q deletions (P=0.01; both del(10q) parents carry the minor risk allele), and we replicated the association of the 70kb germline deletion at *TM2D3*/*TARSL2* with 15q CNN-LOH (P=0.001; the single 15q CNN-LOH individual carries the 70kb germline CNV).

We also note that our UK Biobank analyses already achieved an independent confirmation of each reported association by virtue of the fact that we ran two orthogonal types of association analysis: (i) a standard GWAS testing for association between inherited variants and presence of nearby somatic events, and (ii) an allelic association test checking whether somatic events preferentially deleted or duplicated one allele. We verified that each of these tests produced well-calibrated P-values (Fig. S3.4-1). We found that all hits that reached $P<1\times10^{-8}$ in either test reached nominal significance in the other test, providing strong evidence of true association. In particular, for each of our key results (1p, 10q, 11q, 15q), the allelic bias was either perfect or near-perfect (Table 1).

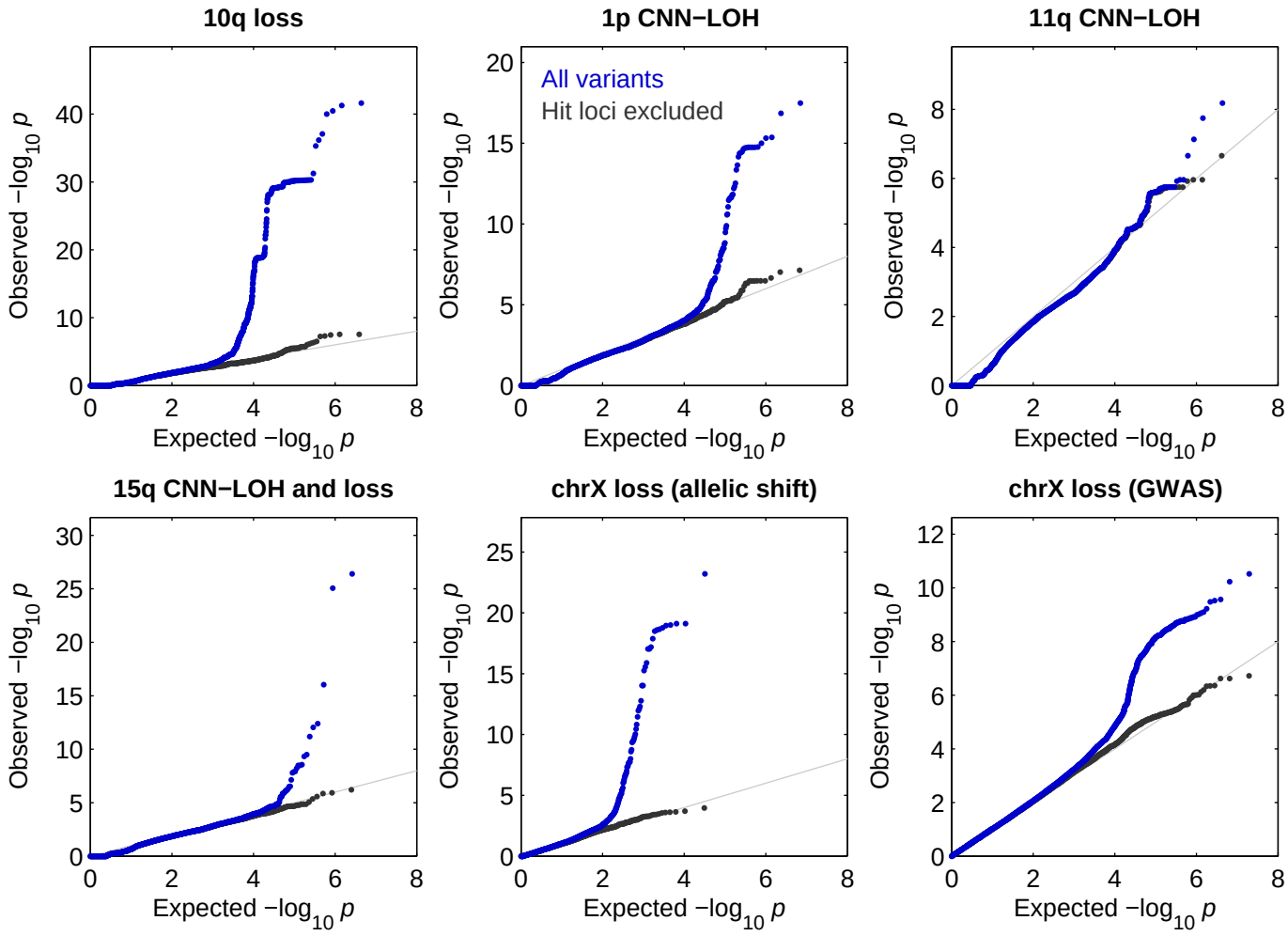


Figure S3.4-1. Quantile-quantile plots of P-values produced by association analyses. These plots verify the calibration of the statistical tests we used to identify the genome-wide significant associations reported in Table 1 (see Methods for details). In each plot, the blue dots correspond to an analysis of all variants tested, while the black dots correspond to an analysis in which regions surrounding significant associations were excluded. Specifically, the plots respectively exclude 10:105–120Mb, 1:40–50Mb, 11:105.5–110.5Mb, 15:100Mb-qter (for the autosomal GWAS on n=120,664 individuals), X:55–66Mb and 114–116Mb (for the X loss allelic shift association analysis on n=3,220 females), and 2:231–232Mb and 6:30–33Mb (for the X loss GWAS on n=66,685 females). In all cases, exclusion of the hit regions (which account for a small fraction of the variants tested) results in a distribution close to the expected null.

4 Statistical properties of event calls

In this note we examine the statistical properties of our detection methodology, focusing on the size distribution of events we detect (with comparisons to previous studies) and the resolution of event boundaries our method estimates.

4.1 Size and clonal fraction distribution of events

We first examine the size distribution of our autosomal mosaic event calls (stratified by copy number) in comparison to Jacobs et al. [1], Laurie et al. [2], Machiela et al. [7], and Vattathil & Scheet [8]). The overall distribution of event sizes we detect is broadly consistent with these previous studies (Fig. S4.1-1). Noticeable differences can be explained by differences in detection methodology (e.g., Jacobs et al. and Machiela et al. restricted to events >2Mb) and sample ascertainment (e.g., most of the calls from Machiela et al. come from cancer cases, in which short gain events are much more common than in healthy elderly individuals). Across all studies, detected mosaic events are generally much larger than inherited structural variants (which have a median length of ≈2.5kb for deletions and ≈36kb for duplications [74]), although this difference is presumably driven in part by detection sensitivity.

We next examine the minimum size of detectable events as a function of clonal cell fraction. Our minimum detectable event size was ≈100kb for events at high clonal fractions (Fig. S4.1-2). In general, the size threshold scales with the inverse square of the clonal fraction, as we show in Supplementary Note 5.1 and is borne out empirically in Fig. S4.1-2. At a clonal fraction of ≈0.1, events >1Mb are detectable, while at a clonal fraction of ≈0.01, events >100Mb are detectable. Conversely, 100Mb events are detectable down to a clonal fraction of ≈0.01, while 1Mb events are detectable down to a clonal fraction of ≈0.1. (We caution, however, that these numbers are specific to the phasing quality and genotyping platform of UK Biobank.) We also note that CNN-LOH events are twice as easy to detect as loss and gain events because CNN-LOH events produce twice the BAF shift of loss and gain events. Overall, the majority of events we detected were present at low clonal fractions (Fig. S4.1-3).

One possible consequence of differential detection sensitivity between methods is that the relative frequencies of different types of events may appear to differ across studies. As a case in point, we consider 20q deletions. In most previous studies of mCAs, del(20q) events have been the most common detected loss event, and sometimes the most common mosaic event altogether [1, 2, 7, 8]. In contrast, in our call set, many events are detected at frequencies similar to del(20q) (although it is still the second-most common loss event after del(13q14); Fig. 1). However, on closer inspection, we realized that our call rate for 20q deletions is actually very similar to previous studies: we call 20q deletions in 130 of 151,202 individuals (0.09%), very similar to the 91 / 82,483 (0.11%) call rate for del(20q) in ref. [25].

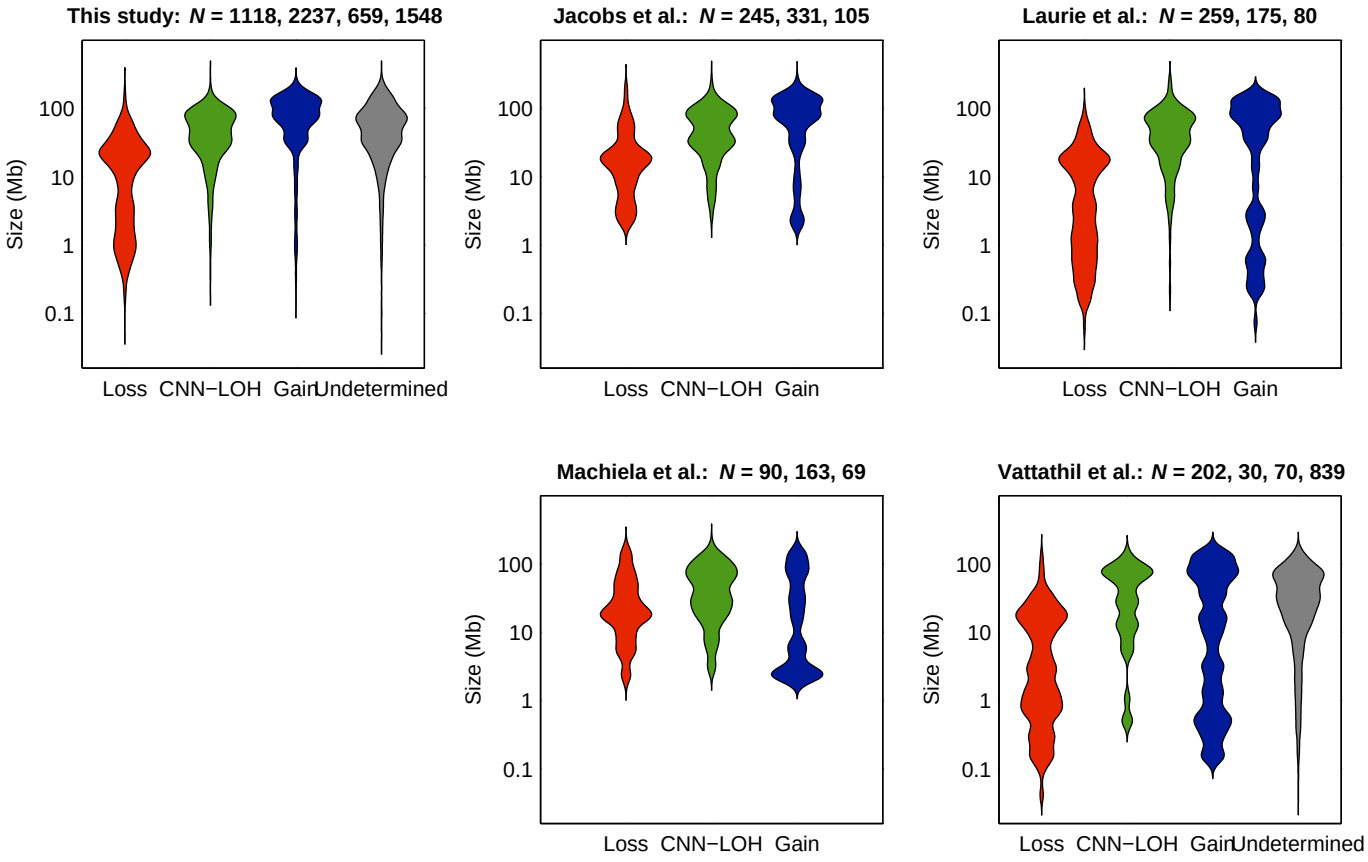


Figure S4.1-1. Size distributions of mCA calls in this study and previous work. We compare the sizes of autosomal mosaic events called in this work and the four largest previous studies of mosaic chromosomal alterations (Jacobs et al. [1], Laurie et al. [2], Machiela et al. [7], and Vattathil & Scheet [8]). Events are stratified by copy number (loss, CNN-LOH, gain); our study and Vattathil & Scheet also call substantial numbers of low-clonal-fraction events for which copy number is undetermined. Violin plots show size distributions over N mCAs with each copy number call. The overall distributions of detected event sizes are broadly consistent. Factors that may contribute to differences between studies include differences in methodology (e.g., Jacobs et al. [1] and Machiela et al. [7] restricted to events >2Mb) and sample ascertainment (e.g., age, sex, cancer status).

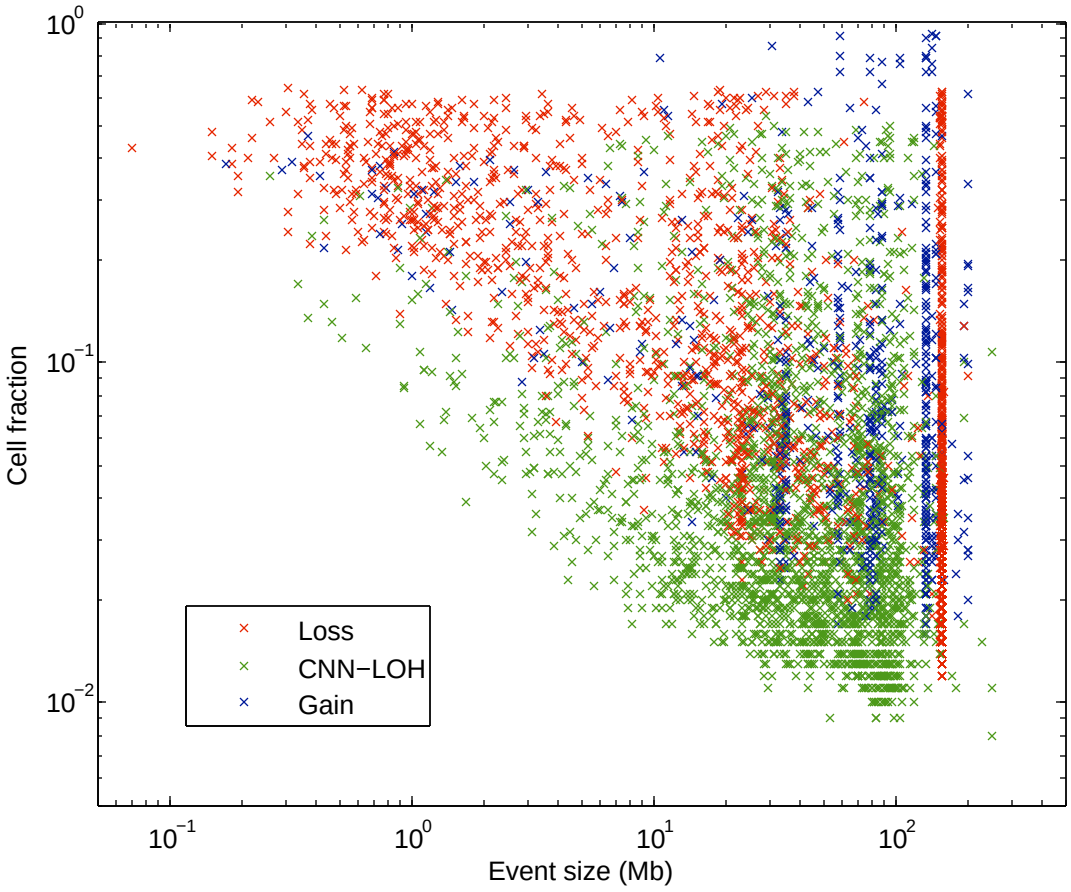


Figure S4.1-2. Scatter plot of clonal cell fraction vs. event size for detected mCAs. Events are color-coded by copy number (red=loss, green=CNN-LOH, blue=gain). (Events with undetermined copy number are not plotted because the relationship between LRR, BAF, and cell fraction is unclear for these events.) Events forming vertical stripes on the far right of the plot correspond to whole-chromosome losses (e.g., loss of X) and trisomies. The scatter plot has a triangular shape because the minimum detectable clonal cell fraction scales as the inverse square root of event size (Supplementary Note 5.1). We also note that CNN-LOH events are twice as easy to detect as loss and gain events because CNN-LOH events produce twice the BAF shift of loss and gain events.

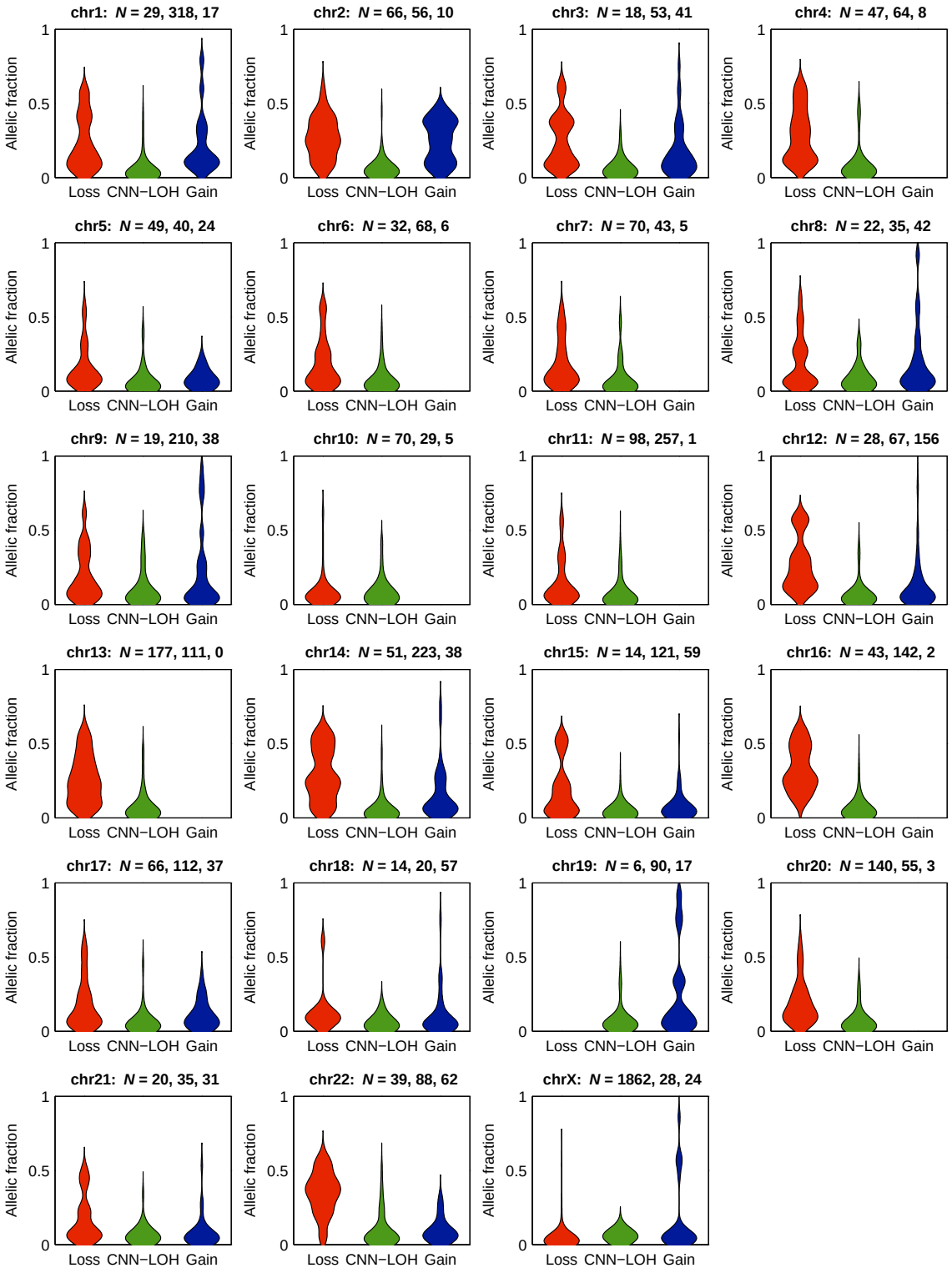


Figure S4.1-3. Extent of clonal proliferation of mCAs detected on each chromosome. For each of N mCAs called as a loss, CNN-LOH, or gain, we estimate its allelic fraction (i.e., fraction of blood cells with the mCA) from LRR and $|\Delta\text{BAF}|$. The violin plots show allelic fraction distributions stratified by chromosome and copy number (whenever at least ten events were called).

We suspect that the reason for the lower *relative* call rate for del(20q) events in our data could be a combination of (i) differences in genotyping coverage or performance (as UK Biobank used Affymetrix whereas previous studies used Illumina); (ii) increased sensitivity of our approach for detecting very long events at low cell fractions, resulting in relatively more detections of long CNN-LOH events and trisomies vs. focal deletions; and (iii) differences in minimum lengths of events analyzed (e.g., Jacobs et al. and Machiela et al. examined >2Mb events, whereas we did not impose an explicit size limit), resulting in our method producing relatively more deletion calls at tightly focal deletion regions (e.g., *DNMT3A*, *TET2*, *DLEU2*) vs. 20q, at which deletions tend to be less focal (several Mb). For example, at *DLEU2* on 13q, we call 166 deletions, of which 48 (29%) are <2Mb. In contrast, only 7 of 130 del(20q) events are <2Mb, such that if we restricted to >2Mb events, del(20q) would be the most common deletion in our call set.

4.2 Breakpoint resolution of events

To estimate the error in our breakpoint calls and the coverage of our confidence intervals, we analyzed the 60 del(10q) calls associated with breakage at the fragile site *FRA10B* (Fig. 3). These calls provide a unique opportunity for measuring breakpoint uncertainty because they are readily confirmed as very likely to be *FRA10B*-associated (all 60 involve carriers of the rs118137427:G risk haplotype at 5% frequency in the population), and for all of these events, the true breakpoint is probably in or very near *FRA10B* (chr10:113Mb). Using this information, we computed an RMSE of 3.0Mb across the 60 del(10q) breakpoint calls, and we observed that 44 of 60 confidence intervals spanned *FRA10B* (≈73% coverage). This coverage increased to 53 of 60 (≈88% coverage) upon expanding interval sizes by 1Mb in each direction.

We also expected that breakpoint uncertainty should exhibit an inverse relationship with cell fraction. Plotting breakpoint calls and confidence intervals against cell fractions confirmed this expectation (Fig. S4.2-1). For the 10 del(10q) calls with highest cell fractions (0.068–0.162), 7 of 10 breakpoints were correct within 0.2Mb, and 5 of 10 within 0.1Mb.

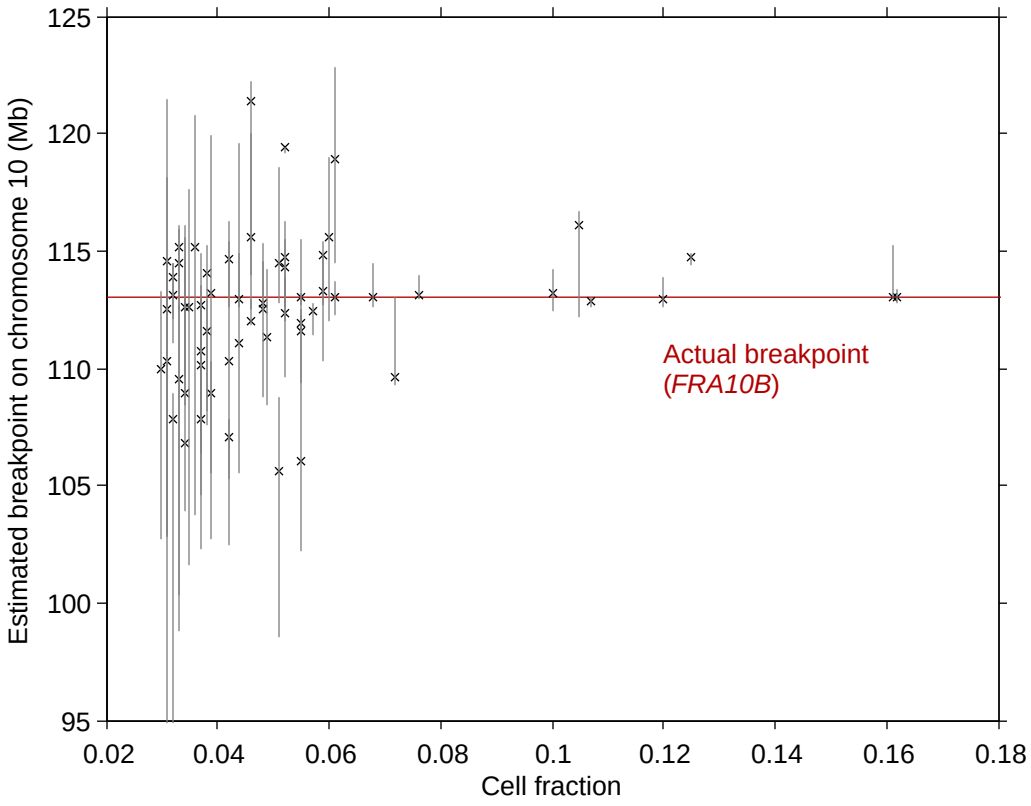


Figure S4.2-1. Estimated breakpoints of *FRA10B*-related del(10q) events. Breakpoints and breakpoint uncertainty estimates (Supplementary Note 1.5) are plotted for each of the 60 del(10q) events we detected that were associated with breakage at the fragile site *FRA10B* (Fig. 3). These calls provide a unique opportunity for measuring breakpoint uncertainty because they are readily confirmed as very likely to be *FRA10B*-associated (all 60 carry the rs118137427:G risk haplotype at 5% frequency in the population), and for all of these events, the true breakpoint is probably in or very near *FRA10B* (chr10:113Mb).

5 Detection sensitivity using long-range phasing vs. previous approaches

In this note we compare the statistical sensitivity of our long-range phase-based mCA detection approach (Supplementary Note 1) to previous approaches. We focus on comparisons with the hapLOH method [8, 54], which was previously shown to be more sensitive for detection of large events at low cell fractions compared to methods that do not incorporate phase information (e.g., circular binary segmentation, CBS [56] and Genomic Alteration Detection Analysis, GADA [57]); however, we also explore the amount of statistical signal available to the latter approaches in our data. (We note that for detection of shorter constitutional or high-cell-fraction CNVs—for which CBS and GADA were originally designed—the relative performance of methods is likely to be very different.)

5.1 Theoretical comparison of statistical tests

While our method applies a principle similar to hapLOH (which demonstrated the value of phase information for event detection [8, 54]), our model and statistical test are quite different from hapLOH. Specifically, whereas hapLOH tabulates and tests "switch consistency" between consecutive heterozygous SNPs, our method applies a hidden Markov model to fully harness long-range phase available across very many SNPs in large data sets such as UK Biobank (Supplementary Note 1).

To understand the effects of these statistical frameworks on detection sensitivity, a mathematical derivation is helpful. Consider a sequence of M consecutive correctly-phased heterozygous SNPs spanning a region in which a mosaic event has created a small BAF shift of δ standard deviations away from 0.5. That is, within the mosaic region, phased BAF (pBAF) has the distribution

$$\text{pBAF} \sim N(0.5 + \delta\sigma, \sigma^2), \tag{8}$$

where σ^2 denotes BAF measurement noise. We can then compute expected z-scores using the switch consistency statistic of hapLOH vs. a long-range phase-based approach that aggregates the pBAF shift across the entire region:

- **Switch consistency (hapLOH).** Equation (8) implies that at each heterozygous SNP,

$$P(\,\text{pBAF} > 0.5\,) \approx 0.5 + \frac{1}{\sqrt{2\pi}} \cdot \delta, \tag{9}$$

  where $\frac{1}{\sqrt{2\pi}}$ comes from the normal probability density (assuming δ is small). Consequently,

the probability of switch consistency between two consecutive SNPs (indexed 1 and 2) is

$$\begin{aligned} & P(\,\text{pBAF}_1 > 0.5\,) \cdot P(\,\text{pBAF}_2 > 0.5\,) + P(\,\text{pBAF}_1 < 0.5\,) \cdot P(\,\text{pBAF}_2 < 0.5\,) \\ = \;& \left(0.5 + \frac{1}{\sqrt{2\pi}} \cdot \delta\right)^2 + \left(0.5 - \frac{1}{\sqrt{2\pi}} \cdot \delta\right)^2 = 0.5 + \frac{\delta^2}{\pi}. \end{aligned} \tag{10}$$

That is, within the mosaic region, switch consistency behaves like a biased coin with a bias of $\frac{\delta^2}{\pi}$. It follows that the expected z-score for detecting elevated switch consistency across M consecutive observations is approximately

$$E[z_{\text{hapLOH}}] \approx \frac{2\delta^2}{\pi}\sqrt{M}. \tag{11}$$

- **Long-range phase.** In contrast, if we instead directly aggregate our signal of a δ-s.d. pBAF shift (equation (8)) across the whole M-SNP region—essentially what our hidden Markov model allows us to do—we obtain a z-score of

$$E[z_{\text{LRP}}] = \delta\sqrt{M}. \tag{12}$$

The key difference between the z-score formulas derived in equations (11) and (12) is the exponent of δ. The difference in exponents implies that hapLOH is sensitive to events with BAF shift $\delta > M^{-1/4}$ (up to a constant factor), while our approach is sensitive to events with BAF shift $\delta > M^{-1/2}$ (which is much smaller than $M^{-1/4}$ for large M, i.e., long events).

We note that for simplicity, we did not consider switch errors in this derivation, which highlights the difference between the methods in the limit of perfect phasing. In practice, the above derivation should be treated as an approximation given that switch errors in inferred phase slightly reduce the sensitivity of both approaches. However, the approximation is quite good in UK Biobank given that our phasing is accurate to tens of megabases [23, 24].

5.2 Empirical power comparison

To compare the sensitivity of different detection approaches in practice, we implemented the switch consistency test used by hapLOH [8, 54] and a mean LRR test using the same basic principle as CBS [56] and GADA [57]. (CBS and GADA are both segmentation methods for identifying regions of copy alteration; within a region, the methods check for consistent allelic intensity deviations.) We then computed test statistics for each approach on the event calls produced by our method, checking which of our events could also have been detected by the other approaches. (We considered directly running each of the other methods but realized that extensive post-processing and parameter-tuning are generally required to QC the output of mosaic event callers; see e.g.

Laurie et al. (2012), Supplementary Note pp. 6-14 [2], and Jacobs et al. (2012), Methods and Supplementary Note pp. 15-16 [1].)

We observed that fewer than half of the events we called (39%) reached nominal $P<0.05$ significance using the hapLOH switch consistency test (Fig. S5.2-1). Much stronger significance would be required to control false discovery rate in a genome-wide detection setting. We observed that requiring $P<0.0001$ significance reduced the detectable proportion of events to 23%. Detection sensitivity improved as a function of clonal cell fraction: among events with $>2\%$ (resp. $>5\%$) cell fraction, the hapLOH test achieved $P<0.0001$ for 40% (resp. 72%) of calls. We observed similar results for the mean LRR statistic (restricted to copy-changing events): among losses and gains with $>2\%$ (resp. $>5\%$) cell fraction, the mean LRR test achieved $P<0.0001$ for 55% (resp. 88%) of calls. (Without further QC, this test would likely produce false positives in practice; LRR is generally prone to local shifts in genotyping intensities [52].)

We note that these quantitative comparisons are undoubtedly specific to the phasing quality and genotyping platform of UK Biobank: in the data we analyzed, phasing quality is exceptionally high while BAF precision appears to be much lower than in previous studies (perhaps because of the Affymetrix genotyping platform used here vs. the Illumina arrays used by previous studies), giving our method an advantage over others. In more typical data sets (with lower-quality phase and more precise genotyping intensities), the performance difference is likely to be smaller.

The above analyses are subject to the caveat that not all of the event calls made by our method are correct: we estimate that our call set has an FDR of 6–9% (Supplementary Note 3.1), but we cannot completely rule out the possibility that our FDR is higher. However, one particular event—terminal deletion of 10q—uniquely provides a gold standard test set and allows comparison of sensitivity across studies (genotyping and phasing differences notwithstanding).

The del(10q) event is unique in its genomic specificity (breakage at *FRA10B* with subsequent deletion of 10q25.2–10qter) and ease of corroboration (via checking for the rs118137427:G risk haplotype (RAF=5%), on which all del(10q) events we observed occurred). Using our methodology, we detected 60 occurrences of this event in 151,202 UK Biobank individuals (with 60/60 carrying the risk haplotype), nearly always at low cell fraction (mean 5.5%, s.d. 2.9%). Only 18 of these events reach nominal significance ($P<0.05$) using the hapLOH switch consistency test. Consistent with these results, ref. [8] detected only 3 such events among 31,100 individuals, and earlier studies that applied CBS or GADA, which had sensitivity limits of $>5\%$ cell fraction [1, 2, 7], did not detect any such events among a total of 127,179 individuals analyzed.

We caution that in general, comparing the statistical performance of detection methods that have been applied to different data sets is complicated by genotyping differences as well as differences in sample ascertainment (e.g., age, sex, and cancer status), but for completeness, Supplementary Table 15 provides a breakdown of detection rates by copy number for our study and previous large studies [1, 2, 7, 8].

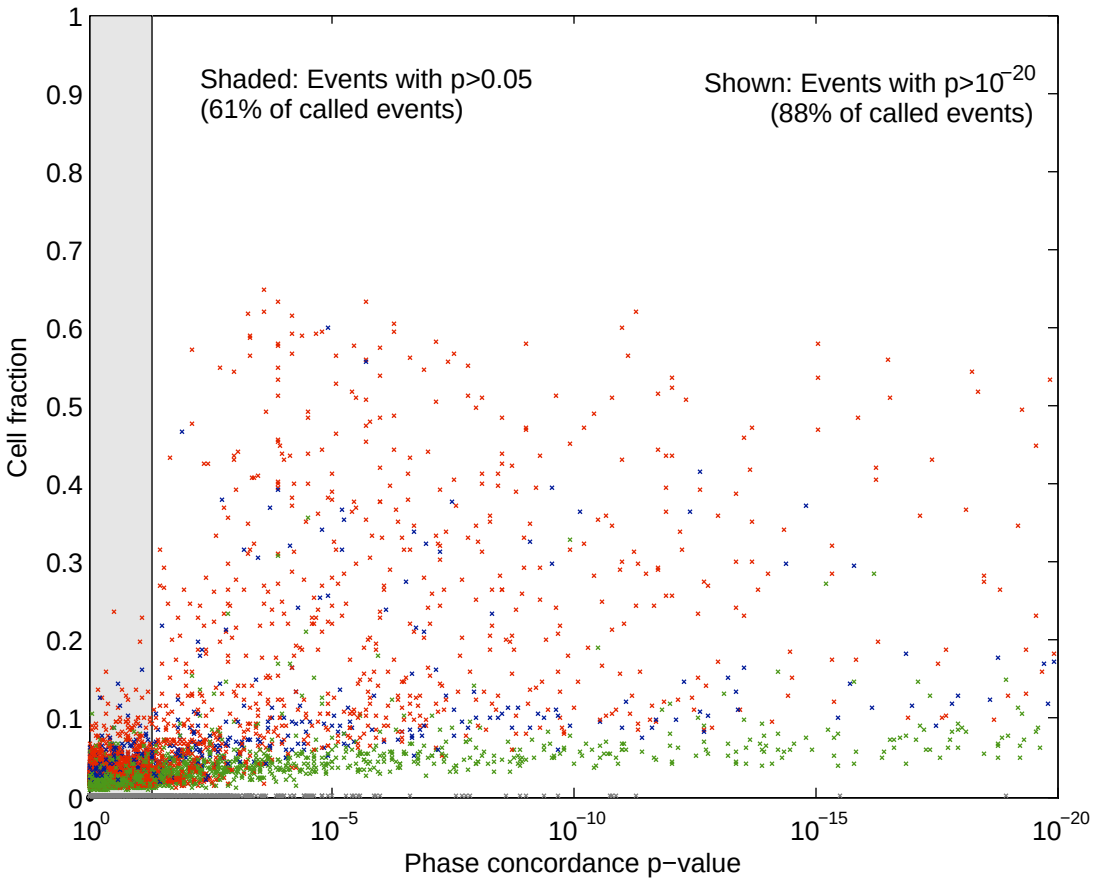


Figure S5.2-1. Sensitivity of phase concordance-based statistical test for detecting mCAs. For each mCA called by our algorithm (red=loss, green=CNN-LOH, blue=gain, grey=undetermined copy number), we computed a binomial *P*-value using the phase concordance test of ref. [54]. This test makes use of relative haplotype phase between successive heterozygous SNPs but does not take advantage of long-range phase information. We plotted the inferred cell fraction of each mCA against its phase concordance *P*-value. (For events with uncertain copy number, we did not infer a cell fraction, so these events are plotted on the x-axis.) We observe that the majority of events detectable by our analysis do not reach nominal significance using the phase concordance test, as expected for subtle allelic imbalances that must be aggregated in-phase over tens of megabases in order to be detectable.

6 Analysis of co-occurring mosaic events

Some kinds of somatic mutations could in principle have synergistic growth-promoting effects, a hypothesis suggested by earlier observations that individuals acquire multiple mCAs much more frequently than expected by chance [1, 2, 7, 8] (Fig. 2b and Supplementary Table 2). We identified three clusters of significantly co-occurring mCAs (Bonferroni $P<0.05$; Fisher's exact test), one of which included events commonly observed together in chronic lymphocytic leukemia (CLL) [32, 33]: trisomy 12, 13q LOH (including deletion and CNN-LOH), and clonal V(D)J deletions on chromosomes 14 and 22 (Fig. 2b and Supplementary Table 3). (The V(D)J deletions may be markers for the cell populations in which the other events are selected.) The co-occurring events generally exhibited cell fractions suggesting co-occurence within the same clonal cell population (Extended Data Fig. 3) and could be explained by synergistic effects of proliferation, by shared genetic, cell-biological, or environmental drivers, or by sequential progression from one event to the other.

7 Analysis of focal deletions

The genomic distribution of mCAs is highly non-random, and commonly deleted regions (CDRs) <1Mb in length are of particular interest as they may indicate haploinsufficient genes for which loss of one copy leads to excessive cell proliferation [2]. Excluding V(D)J recombination regions in 14q11.2, 14q32.33, and 22q11.22, the three most commonly deleted regions targeted *DNMT3A* on 2p, *TET2* on 4q, and *DLEU2*/*DLEU7* on 13q, matching observations in previous studies [2, 8]; we further observed that large majorities of CNN-LOH events on these chromosome arms included these genes, suggesting convergent patterns of selection (Fig. 1 and Fig. S7-1). (We observed a similar pattern with longer deletions and CNN-LOH events spanning *ATM* on 11q; Fig. S2-11.) We also observed CDRs at three genes not previously noted in population studies of mCAs but commonly mutated in cancers: *ETV6* on 12p (mutated in hematological malignancies), *NF1* on 17q (deleted in neurofibromatosis type 1), and *CHEK2* on 22q (involved in the DNA damage response and mutated in many cancers) (Figures S2-12, S2-17, and S2-22). Additionally, we observed two new CDRs for which literature search implicated putative target genes: *RPA2*, which is one of six genes in a 300kb region of 1p36.11–1p35.3 contained in six deletions and is involved in DNA damage response [75], and *RYBP*, which is the only gene in a 620kb region of 3p13 contained in seven deletions and has been reported to be a tumor suppressor gene [76] (Figures S2-1 and S2-3).

To detect CDRs, we needed to identify short genomic regions covered by many loss events; however, we also needed to require that the losses be somewhat specific to a focal region (e.g., a short deletion should carry much more weight than a deletion of an entire arm). To capture this intuition, we gave each loss event a weight equal to 6Mb / [event length], with a maximum weight of 1 (for events shorter than 6Mb). We then examined all regions with a total weight exceeding 4 and checked whether the pileup of losses at these regions was sufficiently focal to be deemed a CDR.

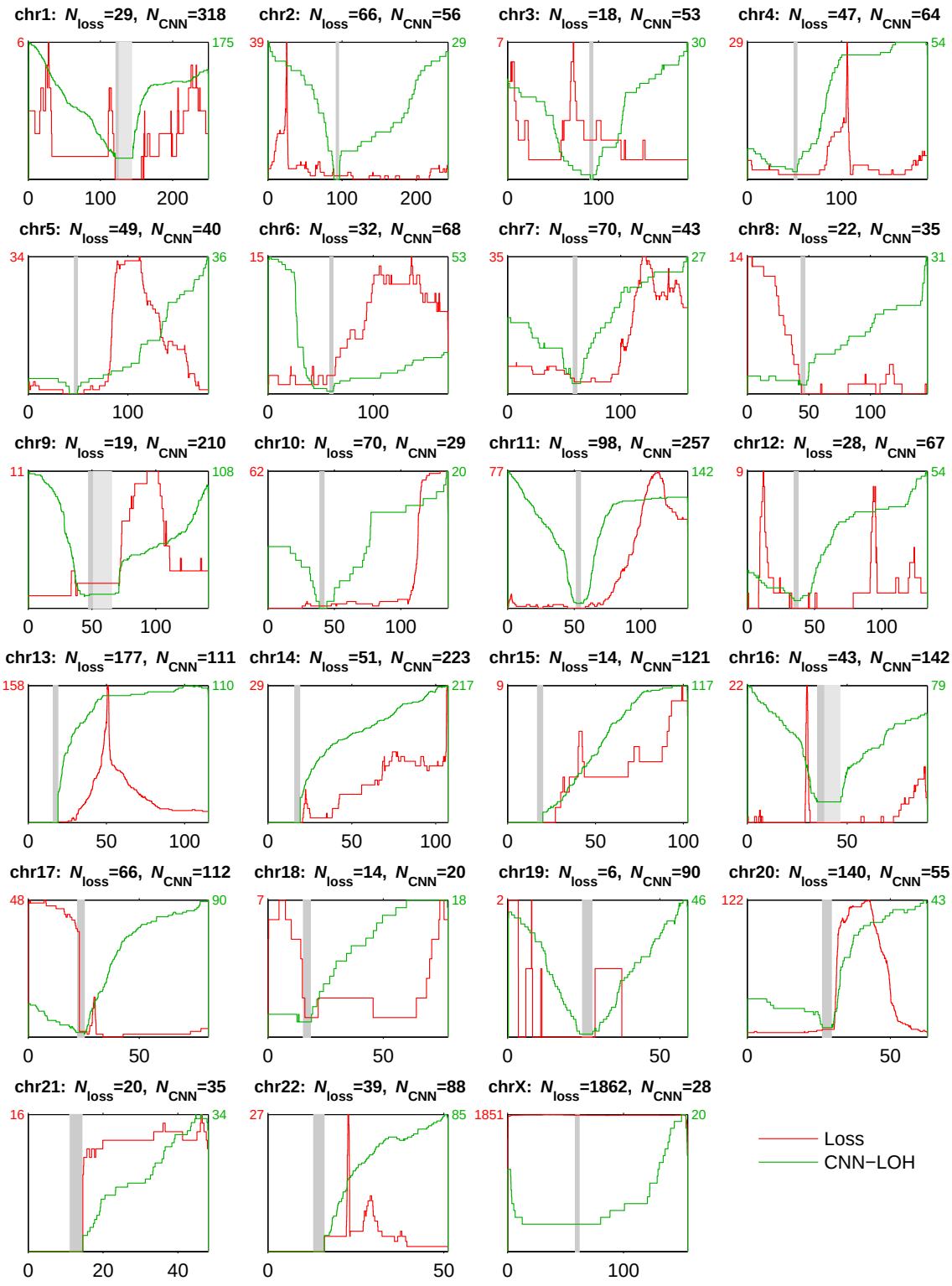


Figure S7-1. Genomic coverage by somatic loss and CNN-LOH events. The red and green curves indicate the total numbers of detected somatic losses (red) and CNN-LOHs (green) covering each position in the genome.

8 Non-age-related mosaic events in ASDs and the general population

Two mCAs (deletion of 16p11.2 and 10q25.2–qter) exhibited no increase in frequency with advancing age, deviating from typical age-related clonal hematopoiesis (Fig. 2e and Supplementary Table 5) and suggesting the possibility of acquisition early in development. Given the well-established relationship of 16p11.2 events to autism [77–79] and the presence of many (16) genes in the deleted 10q region with elevated expression in brain [80], we evaluated their relationships to ASDs in the Simons Simplex Consortium (SSC) [26] WGS data.

8.1 Analysis of del(16p11.2) events

Copy-number variation at 16p11.2 is one of the strongest known genetic effects on ASDs, occurring most often as a *de novo* mutation [77]. Inherited 16p11.2 deletions have been reported to produce a macrocephalic phenotype, while inherited duplications produce a microcephalic phenotype [78, 79].

Surprisingly, we observed 16p11.2 deletions in mosaic form in the general population (22 observations among 151,202 individuals from UK Biobank; Fig. S2-16). Detected events were present at cell fractions of 19–60%. Intriguingly, such mosaic deletions were much more common among females than males (19 females versus 3 males), and as noted above, mosaic del(16p11.2) carriers had an average age similar to the overall study cohort—contrary to the usual skew to the elderly (Fig. 2e and Supplementary Table 5). The lack of an age skew and the high observed cell fractions suggest that these mutations might be developmentally-acquired rather than adult-acquired (although other data will be needed to make a confident determination).

We searched for mosaic 16p11.2 events in the SSC WGS data using our sensitive, haplotype-based mCA detection approach (capable of detecting such events at low cell fractions), but we did not observe any mosaic 16p11.2 mutations among 519 ASD probands or 1,560 family members (parents and unaffected siblings) (Fig. S8.1-1). However, this observation does not preclude the possibility that such mutations might occur at lower than 1/519 frequency among ASD cases (especially given that meiotic, constitutional 16p11.2 mutations are presumably more common and explain only $\approx$1% of cases).

While more data will be needed to evaluate the potential relationship between mosaic 16p11.2 deletions and ASDs, the fact that somatic 16p11.2 deletions give rise to clonality at high cell fractions (in UK Biobank samples) provides a clue to further understanding their effects on development: intriguingly, our observation of mosaic 16p11.2 deletions—but not duplications—aligns with previous work suggesting that 16p11.2 deletions may affect proliferation of a progenitor cell [78, 79]. Specifically, our observation of clonal mosaic 16p11.2 deletions in UK Biobank

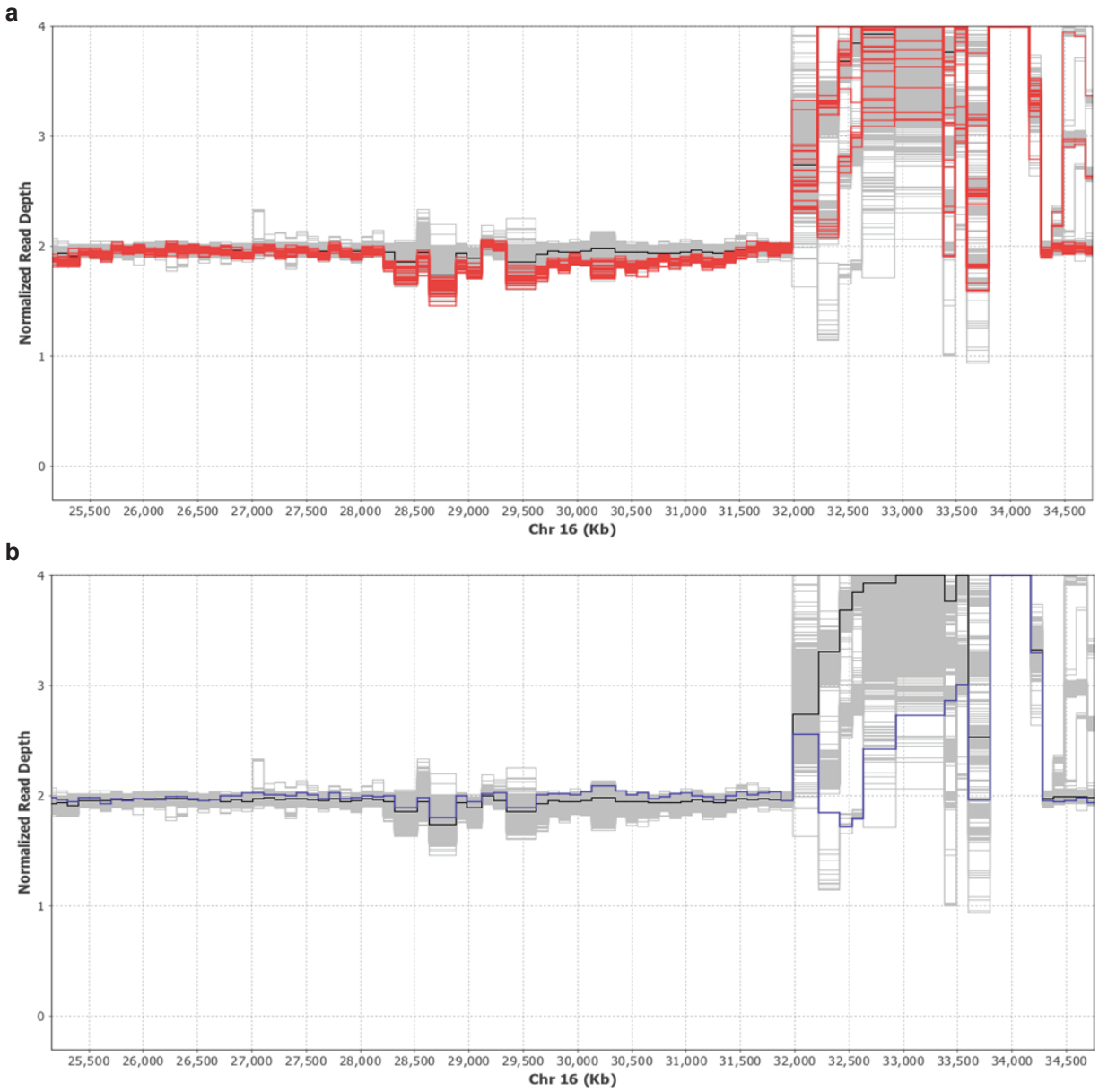


Figure S8.1-1. No evidence for mosaic 16p11.2 deletion in SSC samples. Read depth profile plots in chr6:25-35Mb (one line per SSC individual) show no evidence of individuals carrying the 16p11.2 deletions we observed in UK Biobank (Fig. S2-16). (**a**) Roughly 30 samples (red) exhibit read dropout throughout the region, likely due to technical effects. (**b**) One sample has a candidate mosaic duplication from ∼26.8–31.9Mb.

The fact that we did not detect any examples of mosaic 16p11.2 deletion in the SSC cohort could be due to chance (given that the detection frequency in UK Biobank was 1 in ∼6,000 individuals) or due to ascertainment of the SSC for non-carriers of constitutional 16p11.2 CNVs (which may also have excluded mosaic carriers).

samples suggests that 16p11.2 deletion causes mutant progenitors or stem cells to either increase proliferation or resist differentiation, with the result that clonal progeny of the mutant cell expand in numbers relative to other cells (making the mutation detectable). In contrast, we notably did not observe clonal mosaicism involving the reciprocal mutation (16p11.2 duplication), suggesting that cells with 16p11.2 deletion have a proliferative advantage but cells with the duplication do not. (Assuming these mutations are produced by non-allelic homologous recombination of sister chromatids, both mutations should arise equally frequently.) This proliferative hypothesis is consistent with the macrocephalic phenotype of 16p11.2 deletions and the microcephalic phenotype of 16p11.2 duplications [78, 79], leading us to speculate that 16p11.2 mutation may have analogous biological effects during hematopoiesis and brain development.

8.2 Analysis of del(10q) events and fragile site *FRA10B*

Applying our methodology to detect mosaic del(10q) events in SSC revealed two parent-child duos in which both parent and child had acquired the 10q terminal deletion (in mosaic form). While both children in the duos were unaffected siblings, this observation of Mendelian inheritance for an acquired event nonetheless informs our thinking about ASDs (which are highly heritable), as it shows that acquired mutations can exhibit heritable behavior.

Our association analysis in UK Biobank showed that the heritable acquisition of 10q deletions was linked to a common risk haplotype (allele frequency=5% in the population) tagged by rs118137427 near *FRA10B*, a known genomic fragile site [34, 35] at the estimated common breakpoint of the 10q deletions (Fig. 3). In the SSC cohort, we observed that all four mosaic del(10q) individuals possessed expanded AT-rich repeats at *FRA10B* on the rs118137427:G risk haplotype (Fig. 3c and Extended Data Fig. 5a,b). To further investigate the repeat structure of *FRA10B* alleles, we undertook a detailed analysis of the variable number tandem repeat (VNTR) sequence at *FRA10B* in the SSC WGS data. This analysis (detailed in the following subsections) revealed a diversity of novel VNTR sequence motifs (12 distinct primary repeat units carried by 26 SSC individuals from 14 families); all of these novel VNTR motifs were present on the rs118137427:G haplotype background, despite the low frequency of that haplotype in the population (5%) (Extended Data Fig. 5a,b, and Fig. S8.2-1). We did not observe an association between the VNTR motifs and autism status in the SSC cohort.

8.2.1 Overview of previous work on *FRA10B*

Sutherland et al. [34] discovered the fragile site *FRA10B*, observing that a small fraction of individuals ($\approx$1 in 40 Australians [81]) carry a polymorphism resulting in chromosomal gaps or breakage at 10q25 in lymphocyte culture under bromodeoxyuridine (BrdU) treatment. Hewett et al. [35] characterized the molecular structure of *FRA10B* using Sanger sequencing, obtaining the

following key findings:

- All alleles at the *FRA10B* locus contain an extremely (≈91%) AT-rich region of at least 1kb. This region contains a wide variety of AT-rich repeats of length 16–52 bp.
- Roughly one-third of alleles contain expanded repeats extending the length of the AT-rich region to 1-4kb.
- ≈1% of alleles—those that express *FRA10B* fragility under BrdU treatment—are very long (5–20kb). These expanded *FRA10B* alleles contain repeated variations of a 42bp consensus motif; slight variations exist among the repeat units present within an individual and between individuals. Each expanded allele likely contains >75 repeat copies (based on total allele length and the assumption that expanded alleles are primarily comprised of ≈42bp repeats). Expanded alleles are highly unstable, exhibiting both intergenerational and somatic mutation.

8.2.2 Overview of approach to analyzing WGS data

In this work, we identified a new, much rarer genetically-induced anomaly: breakage at *FRA10B in vivo*, resulting in mosaic loss of 10q25.2–10qter in normal blood DNA. We detected del(10q) mosaicism of this form in 60 of 151,202 genotyped UK Biobank participants and 4 of 2,079 whole-genome-sequenced SSC participants, always on a low-frequency haplotype (rs118137427:G, MAF 5%) at *FRA10B*.

To investigate the genomic structure of *FRA10B* alleles implicated in del(10q) mosaicism vs. normal alleles, we examined Illumina short-read sequencing data available for the WGS cohort (SSC). This task was challenging because of the repetitive, AT-rich sequence composition of the *FRA10B* locus: the reads observed at *FRA10B* likely depend on several factors including (i) the length of the *FRA10B* alleles present in each individual, (ii) technical variation (across sequencing libraries) in efficiency of capturing reads at very low GC fractions (≈10%), and (iii) technical biases in sampling reads from repeat units with different GC compositions (most likely favoring reads covering repeat units with more G's and C's). In particular, despite the 37.8X median coverage of the WGS data, many samples exhibited extreme read dropout at *FRA10B*, with few or no reads aligning to the *FRA10B* locus.

These limitations precluded interrogation of full *FRA10B* sequences: we were unable to infer total *FRA10B* size from read counts (due to unknown extent of read dropout at *FRA10B*), and we were unable to infer relative fractions of constituent repeat unit variants within a *FRA10B* allele (due to likely GC bias in sampling different repeats).

Instead, we undertook the following conservative analysis strategy: for each individual with at least 10 reads mapping to the *FRA10B* locus, we attempted to identify a primary repeat motif based

on assembling the available reads. Intuitively, each individual's primary motif indicates the "most represented" repeat unit within that individual (subject to potential GC bias). (In general, many different repeat units should be present in each *FRA10B* allele based on the analysis of Hewett et al. [35].) We then compared these primary motifs to the reference sequence, ultimately identifying likely carriers of long variable number tandem repeat (VNTR) sequences with mutations away from the reference.

8.2.3 Identification of non-reference VNTR motifs in 26 individuals

To carry out the strategy outlined above, we first identified a 150bp target region (10:113002151–113002300, hg19) at which del(10q) samples exhibited deep read pileups. This region is a poly-AT region in hg19, and the reference sequence contains three tandem repeats of a 40bp motif (Extended Data Fig. 5a) at this locus. We used this region as "bait," counting the number of reads in each individual that aligned to the region (allowing for mismatches in alignment).

We identified 399 individuals with 10 or more reads mapping to the target region. For each individual, we attempted to assemble the reads of interest by performing an all-to-all pairwise gap-free alignment, finding the most-connected read, and pulling in other reads to form an assembly. We then evaluated the assemblies for repeating VNTR motifs. Most samples either did not assemble or contained only short VNTR motifs (15bp or less) with small numbers of tandem repeats. For 102 samples, we identified 40bp VNTR motifs; 99 matched the hg19 reference and the other 3 had distinct 1bp differences. All of these 102 samples had moderate coverage (10 to 29 reads mapping to the target region).

For 26 samples from 14 whole-genome-sequenced families, we confidently identified a primary VNTR motif with length 38bp, 39bp, 42bp, or 43bp and evidence of three or more tandem repeat copies (Supplementary Table 16 and Extended Data Fig. 5a). Eleven samples had read counts greater than 100 at the target locus, with the highest over 1,000, suggesting very long repeat expansions (Supplementary Table 16). Our assemblies revealed a large range of diversity in VNTR sequences across individuals: we identified 12 distinct primary motifs, only one of which was shared among more than one family (VNTR-42-a, carried in families 11336, 11542, and 13777; families 11336 and 13777 contain the del(10q) individuals in the WGS cohort). No motifs exactly matched repeat units from ref. [35], although many were very similar (Extended Data Fig. 5a). The overall sequence diversity underscored the high mutability of the *FRA10B* locus.

All 26 samples with high-confidence non-reference VNTR motifs carried the rs118137427:G low-frequency allele. Based on haplotype transmission within quartets, we identified 7 additional family members who shared haplotypes with the 26 high-confidence non-reference VNTR carriers. Examination of the k-mer composition of reads from these 7 individuals and the 26 high-confidence individuals showed that k-mer profiles clustered in families and by VNTR motifs almost perfectly, lending support to the accuracy of the VNTR assemblies we generated (Fig. S8.2-1). Family 13892

was the lone exception; one individual (09339, a son) has a very different k-mer profile from his family members (09326 and 09330) who carry the same rs118137427:G haplotype. One possible explanation is intergenerational *FRA10B* expansion, as observed by Hewett et al. [35].

8.2.4 Imputation of VNTRs into UK Biobank

We used Minimac3 [61] to impute non-reference VNTR motifs into UK Biobank individuals based on haplotype sharing at the *FRA10B* locus (using the 26 high-confidence individuals as cases and excluding the 7 additional related individuals from the analysis). Although the VNTR motifs were estimated to collectively be present in just 0.7% of the UK Biobank cohort, they were imputed into 24 of 60 mosaic del(10q) individuals (16 with VNTR-42-a, 5 with VNTR-43-b, and 3 with VNTR-38-a; Extended Data Fig. 5b and Supplementary Table 7).

8.2.5 Possible models for del(10q) mosaicism

While the above analyses strongly implicate *FRA10B* expanded alleles as the source of chromosomal breakage in individuals with mosaic loss of 10q25.2–10qter, the mechanism by which del(10q) cells reach a detectable allelic fraction in whole blood DNA remains unclear. We can imagine three possible routes to del(10q) mosaicism:

1. Mutation early in development.
2. Repeated mutation in many different cells.
3. Clonal expansion of a cell (or cells) that have lost 10q25–10qter.

The first two possibilities do not require clonal expansion of del(10q) cells, while the third would imply that loss of 10q25–10qter confers a proliferative advantage to blood cells.

We have limited ability to distinguish between these possible scenarios (which are also not mutually exclusive). Beyond the association we observe with *FRA10B* alleles, our only other observations on del(10q) individuals are the lack of an age bias, a sex bias toward female cases, and low to very low fractions of del(10q) cells (Fig. S4.2-1). Based on the last observation, we speculate that the second scenario—repeated mutation in many different cells, converging to a cell fraction of a few percent—may be most likely, but additional work will be necessary to resolve this question.

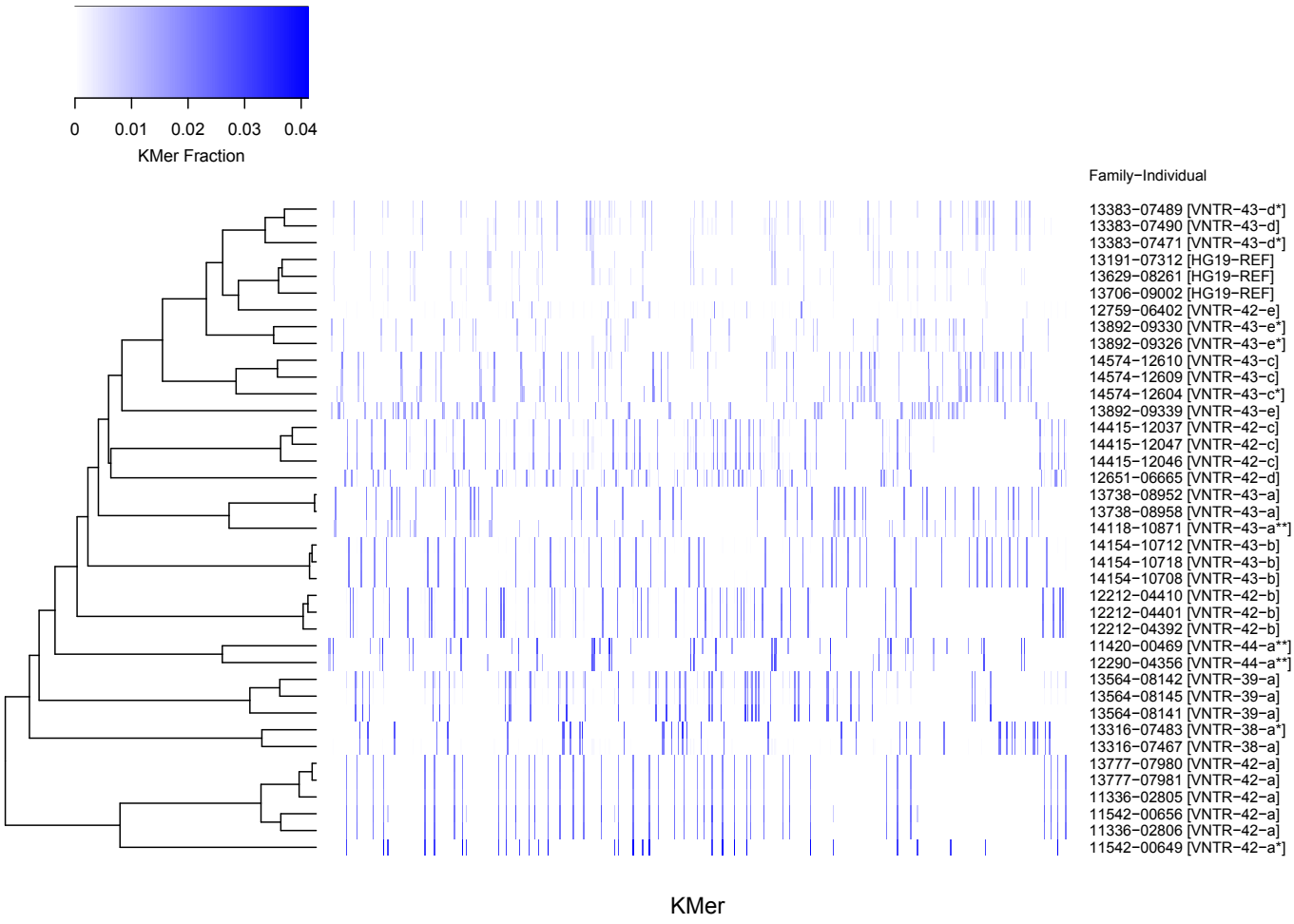


Figure S8.2-1. *FRA10B* read profiles cluster concordantly with primary motifs from VNTR assemblies. To assess the accuracy of the assembly procedure we used to identify VNTR motifs in WGS data, we analyzed k-mer profiles of reads mapping to *FRA10B* in individuals we identified as probable carriers of non-reference VNTR motifs. For each individual of interest (y-axis), we constructed a "barcode" based on 38-mer representation at a set of informative 38-mers (x-axis). This plot contains 3 reference individuals, 26 individuals we identified as high-confidence non-reference VNTR motif carriers, 7 family members of the 26 high-confidence individuals sharing their VNTR haplotypes (indicated with asterisks), and 3 additional medium-confidence VNTR motif carriers (indicated with double asterisks). (The last three individuals only have evidence for two tandem copies of the repeat unit based on 11–12 reads, and they do not carry the rs118137427:G minor allele present in all other non-reference VNTR carriers.)

9 Analysis of biased X chromosome loss

In addition to performing standard GWAS on mosaic status, we also searched our detected mCAs for a different type of association: shift in allelic balance in favor of one allele versus the other in heterozygous individuals (analogous to allele-specific expression). We were well-powered to run this analysis on female chromosome X owing to the high frequency of X loss (Fig. 1), and to further increase association power, we performed X loss association analyses using an expanded set of 3,462 likely X loss calls at an FDR of 0.1. We observed a striking association (P=6.6$\times$10^{-27}, 1.9:1 bias in the lost haplotype) at Xp11.1 near *DXZ1* and a weaker association (P=1.0$\times$10^{-9}, 1.5:1 bias in the lost haplotype) at Xq23 near *DXZ4* (Table 1, Fig. S9-1, and Supplementary Table 9). At both loci, we also observed nominal associations (P=1$\times$10^{-3}) between allele count and X loss (Table 1). The Xp11.1 and Xq23 bias signals appear to be independent (2.7:1 bias when heterozygous risk haplotypes are in phase and 1.2:1 bias when out of phase). We initially suspected that these observations could be explained by biased X chromosome inactivation (XCI) [39], especially given the role of Xp11.1 and Xp23 in XCI [82], but we did not find any evidence of biased XCI in GEUVADIS RNA-seq data [64] (Supplementary Table 10). Interestingly, we observed weak evidence that the lead SNP rs2942875 at Xp11.1 appeared to have similar effects on gain of X (Supplementary Table 9), suggesting a mechanism involving X missegregation, but larger sample sizes will be required to investigate this possibility; we only called 29 likely X gains at FDR 0.1.

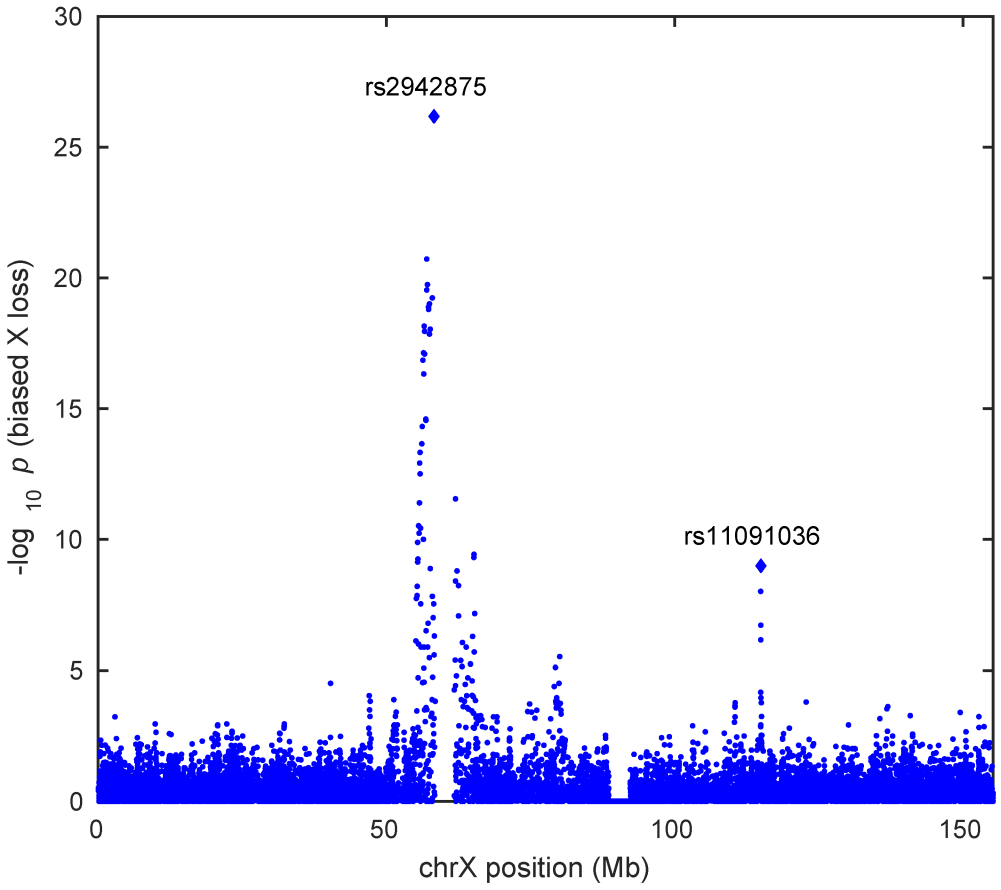


Figure S9-1. Manhattan plot of *cis* associations with biased female chrX loss. For each chrX SNP, a binomial test was run on heterozygous individuals among n=3,220 females with X loss calls at a false discovery threshold of 0.10. The gaps in the plot correspond to the chrX centromere and X-transposed region (XTR); we masked the latter from our analyses, following Laurie et al. [2].

References

1. Jacobs, K. B. *et al.* Detectable clonal mosaicism and its relationship to aging and cancer. *Nature Genetics* **44**, 651–658 (2012).

2. Laurie, C. C. *et al.* Detectable clonal mosaicism from birth to old age and its relationship to cancer. *Nature Genetics* **44**, 642–650 (2012).

3. Genovese, G. *et al.* Clonal hematopoiesis and blood-cancer risk inferred from blood DNA sequence. *New England Journal of Medicine* **371**, 2477–2487 (2014).

4. Jaiswal, S. *et al.* Age-related clonal hematopoiesis associated with adverse outcomes. *New England Journal of Medicine* **371**, 2488–2498 (2014).

5. Xie, M. *et al.* Age-related mutations associated with clonal hematopoietic expansion and malignancies. *Nature Medicine* **20**, 1472–1478 (2014).

6. McKerrell, T. *et al.* Leukemia-associated somatic mutations drive distinct patterns of age-related clonal hemopoiesis. *Cell Reports* **10**, 1239–1245 (2015).

7. Machiela, M. J. *et al.* Characterization of large structural genetic mosaicism in human autosomes. *American Journal of Human Genetics* **96**, 487–497 (2015).

8. Vattathil, S. & Scheet, P. Extensive hidden genomic mosaicism revealed in normal tissue. *American Journal of Human Genetics* **98**, 571–578 (2016).

9. 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** (2016).

10. Forsberg, L. A., Gisselsson, D. & Dumanski, J. P. Mosaicism in health and disease—clones picking up speed. *Nature Reviews Genetics* **18**, 128–142 (2017).

11. Zink, F. *et al.* Clonal hematopoiesis, with and without candidate driver mutations, is common in the elderly. *Blood* **130**, 742–752 (2017).

12. Jaiswal, S. *et al.* Clonal hematopoiesis and risk of atherosclerotic cardiovascular disease. *New England Journal of Medicine* **377**, 111–121 (2017).

13. Acuna-Hidalgo, R. *et al.* Ultra-sensitive sequencing identifies high prevalence of clonal hematopoiesis-associated mutations throughout adult life. *American Journal of Human Genetics* **101**, 50–64 (2017).

14. Laken, S. J. *et al.* Familial colorectal cancer in Ashkenazim due to a hypermutable tract in *APC*. *Nature Genetics* **17**, 79–83 (1997).

15. Jones, A. V. *et al. JAK2* haplotype is a major risk factor for the development of myeloproliferative neoplasms. *Nature Genetics* **41**, 446–449 (2009).

16. Kilpivaara, O. *et al.* A germline *JAK2* SNP is associated with predisposition to the development of *JAK2*V617F-positive myeloproliferative neoplasms. *Nature Genetics* **41**, 455–459 (2009).

17. Olcaydu, D. *et al.* A common *JAK2* haplotype confers susceptibility to myeloproliferative neoplasms. *Nature Genetics* **41**, 450–454 (2009).

18. Koren, A. *et al.* Genetic variation in human DNA replication timing. *Cell* **159**, 1015–1026 (2014).

19. Zhou, W. *et al.* Mosaic loss of chromosome Y is associated with common variation near *TCL1A*. *Nature Genetics* **48**, 563–568 (2016).

20. Hinds, D. A. *et al.* Germ line variants predispose to both *JAK2* V617F clonal hematopoiesis and myeloproliferative neoplasms. *Blood* **128**, 1121–1128 (2016).

21. Wright, D. J. *et al.* Genetic variants associated with mosaic Y chromosome loss highlight cell cycle genes and overlap with cancer susceptibility. *Nature Genetics* **49**, 674–679 (2017).

22. 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 Medicine* **12**, 1–10 (2015).

23. Loh, P.-R., Palamara, P. F. & Price, A. L. Fast and accurate long-range phasing in a UK Biobank cohort. *Nature Genetics* **48**, 811–816 (2016).

24. Loh, P.-R. *et al.* Reference-based phasing using the Haplotype Reference Consortium panel. *Nature Genetics* **48**, 1443–1448 (2016).

25. Machiela, M. J. *et al.* Mosaic chromosome 20q deletions are more frequent in the aging population. *Blood Advances* **1**, 380–385 (2017).

26. Fischbach, G. D. & Lord, C. The Simons Simplex Collection: a resource for identification of autism genetic risk factors. *Neuron* **68**, 192–195 (2010).

27. Werling, D. M. *et al.* An analytical framework for whole-genome sequence association studies and its implications for autism spectrum disorder. *Nature Genetics* **50**, 727–736 (2018).

28. Beroukhim, R. *et al.* The landscape of somatic copy-number alteration across human cancers. *Nature* **463**, 899–905 (2010).

29. Davoli, T. *et al.* Cumulative haploinsufficiency and triplosensitivity drive aneuploidy patterns and shape the cancer genome. *Cell* **155**, 948–962 (2013).

30. Machiela, M. J. *et al.* Female chromosome X mosaicism is age-related and preferentially affects the inactivated X chromosome. *Nature Communications* **7** (2016).

31. Sinclair, E. J., Potter, A. M., Watmore, A. E., Fitchett, M. & Ross, F. Trisomy 15 associated with loss of the Y chromosome in bone marrow: a possible new aging effect. *Cancer Genetics and Cytogenetics* **105**, 20–23 (1998).

32. Landau, D. A. *et al.* Mutations driving CLL and their evolution in progression and relapse. *Nature* **526**, 525–530 (2015).

33. Puente, X. S. *et al.* Non-coding recurrent mutations in chronic lymphocytic leukaemia. *Nature* **526**, 519–524 (2015).

34. Sutherland, G., Baker, E. & Seshadri, R. Heritable fragile sites on human chromosomes. V. A new class of fragile site requiring BrdU for expression. *American Journal of Human Genetics* **32**, 542–548 (1980).

35. Hewett, D. R. *et al. FRA10B* structure reveals common elements in repeat expansion and chromosomal fragile site genesis. *Molecular Cell* **1**, 773–781 (1998).

36. Richards, R. I. & Sutherland, G. R. Dynamic mutations: a new class of mutations causing human disease. *Cell* **70**, 709–712 (1992).

37. Gurney, A. L., Carver-Moore, K., de Sauvage, F. J. & Moore, M. W. Thrombocytopenia in c-mpl-deficient mice. *Science* **265**, 1445–1448 (1994).

38. Tefferi, A. Novel mutations and their functional and clinical relevance in myeloproliferative neoplasms: *JAK2*, *MPL*, *TET2*, *ASXL1*, *CBL*, *IDH* and *IKZF1*. *Leukemia* **24**, 1128–1138 (2010).

39. Tukiainen, T. *et al.* Landscape of X chromosome inactivation across human tissues. *Nature* **550**, 244–248 (2017).

40. Loh, P.-R. *et al.* Contrasting genetic architectures of schizophrenia and other complex diseases using fast variance components analysis. *Nature Genetics* **47**, 1385–1392 (2015).

41. Oddsson, A. *et al.* The germline sequence variant rs2736100_C in *TERT* associates with myeloproliferative neoplasms. *Leukemia* **28**, 1371–1374 (2014).

42. Stacey, S. N. *et al.* A germline variant in the *TP53* polyadenylation signal confers cancer susceptibility. *Nature Genetics* **43**, 1098–1103 (2011).

43. Rawstron, A. C. *et al.* Monoclonal B-cell lymphocytosis and chronic lymphocytic leukemia. *New England Journal of Medicine* **359**, 575–583 (2008).

44. Landgren, O. *et al.* B-cell clones as early markers for chronic lymphocytic leukemia. *New England Journal of Medicine* **360**, 659–667 (2009).

45. Landau, D. A. *et al.* Evolution and impact of subclonal mutations in chronic lymphocytic leukemia. *Cell* **152**, 714–726 (2013).

46. Ojha, J. *et al.* Monoclonal B-cell lymphocytosis is characterized by mutations in CLL putative driver genes and clonal heterogeneity many years before disease progression. *Leukemia* **28**, 2395–2398 (2014).

47. Berndt, S. I. *et al.* Meta-analysis of genome-wide association studies discovers multiple loci for chronic lymphocytic leukemia. *Nature Communications* **7** (2016).

48. O'Keefe, C., McDevitt, M. A. & Maciejewski, J. P. Copy neutral loss of heterozygosity: a novel chromosomal lesion in myeloid malignancies. *Blood* **115**, 2731–2739 (2010).

49. Chase, A. *et al.* Profound parental bias associated with chromosome 14 acquired uniparental disomy indicates targeting of an imprinted locus. *Leukemia* **29**, 2069–2074 (2015).

50. Choate, K. A. *et al.* Mitotic recombination in patients with ichthyosis causes reversion of dominant mutations in *KRT10*. *Science* **330**, 94–97 (2010).

51. Peiffer, D. A. *et al.* High-resolution genomic profiling of chromosomal aberrations using Infinium whole-genome genotyping. *Genome Research* **16**, 1136–1148 (2006).

52. Diskin, S. J. *et al.* Adjustment of genomic waves in signal intensities from whole-genome SNP genotyping platforms. *Nucleic Acids Research* **36**, e126 (2008).

53. Nik-Zainal, S. *et al.* The life history of 21 breast cancers. *Cell* **149**, 994–1007 (2012).

54. Vattathil, S. & Scheet, P. Haplotype-based profiling of subtle allelic imbalance with SNP arrays. *Genome Research* **23**, 152–158 (2013).

55. Genovese, G., Leibon, G., Pollak, M. R. & Rockmore, D. N. Improved IBD detection using incomplete haplotype information. *BMC Genetics* **11**, 58 (2010).

56. Olshen, A. B., Venkatraman, E., Lucito, R. & Wigler, M. Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics* **5**, 557–572 (2004).

57. Pique-Regi, R., Cáceres, A. & González, J. R. R-gada: a fast and flexible pipeline for copy number analysis in association studies. *BMC Bioinformatics* **11**, 380 (2010).

58. Huang, J. *et al.* Improved imputation of low-frequency and rare variants using the UK10K haplotype reference panel. *Nature Communications* **6** (2015).

59. Chang, C. C. *et al.* Second-generation PLINK: rising to the challenge of larger and richer datasets. *GigaScience* **4**, 1–16 (2015).

60. Gusev, A. *et al.* Whole population, genome-wide mapping of hidden relatedness. *Genome Research* **19**, 318–326 (2009).

61. Das, S. *et al.* Next-generation genotype imputation service and methods. *Nature Genetics* **48**, 1284–1287 (2016).

62. Loh, P.-R. *et al.* Efficient Bayesian mixed model analysis increases association power in large cohorts. *Nature Genetics* **47**, 284–290 (2015).

63. Lee, S. H., Wray, N. R., Goddard, M. E. & Visscher, P. M. Estimating missing heritability for disease from genome-wide association studies. *American Journal of Human Genetics* **88**, 294–305 (2011).

64. Lappalainen, T. *et al.* Transcriptome and genome sequencing uncovers functional variation in humans. *Nature* **501**, 506–511 (2013).

65. Castel, S. E., Levy-Moonshine, A., Mohammadi, P., Banks, E. & Lappalainen, T. Tools and best practices for data processing in allelic expression analysis. *Genome Biology* **16**, 195 (2015).

66. Turner, J. J. *et al.* InterLymph hierarchical classification of lymphoid neoplasms for epidemiologic research based on the WHO classification (2008): update and future directions. *Blood* **116**, e90–e98 (2010).

67. Arber, D. A. *et al.* The 2016 revision to the World Health Organization (WHO) classification of myeloid neoplasms and acute leukemia. *Blood* **127**, 2391–2405 (2016).

68. Chatterjee, N., Shi, J. & García-Closas, M. Developing and evaluating polygenic risk prediction models for stratified disease prevention. *Nature Reviews Genetics* **17**, 392–406 (2016).

69. Dumanski, J. P. *et al.* Smoking is associated with mosaic loss of chromosome Y. *Science* **347**, 81–83 (2015).

70. Affymetrix, Inc. Axiom® genotyping solution data analysis guide (2016). URL `http://media.affymetrix.com/support/downloads/manuals/axiom_genotyping_solution_analysis_guide.pdf`.

71. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).

72. Bock, C., Walter, J., Paulsen, M. & Lengauer, T. CpG island mapping by epigenome prediction. *PLOS Computational Biology* **3**, e110 (2007).

73. Price, A. L. *et al.* Long-range LD can confound genome scans in admixed populations. *American Journal of Human Genetics* **83**, 132–135 (2008).

74. Sudmant, P. H. *et al.* An integrated map of structural variation in 2,504 human genomes. *Nature* **526**, 75–81 (2015).

75. Lee, D.-H. *et al.* A PP4 phosphatase complex dephosphorylates RPA2 to facilitate DNA repair via homologous recombination. *Nature Structural & Molecular Biology* **17**, 365–372 (2010).

76. Chen, D. *et al.* RYBP stabilizes p53 by modulating MDM2. *EMBO Reports* **10**, 166–172 (2009).

77. Weiss, L. A. *et al.* Association between microdeletion and microduplication at 16p11.2 and autism. *New England Journal of Medicine* **358**, 667–675 (2008).

78. Shinawi, M. *et al.* Recurrent reciprocal 16p11.2 rearrangements associated with global developmental delay, behavioral problems, dysmorphism, epilepsy, and abnormal head size. *Journal of Medical Genetics* **47**, 332–341 (2010).

79. Golzio, C. *et al. KCTD13* is a major driver of mirrored neuroanatomical phenotypes of the 16p11.2 copy number variant. *Nature* **485**, 363–367 (2012).

80. Uhlén, M. *et al.* Tissue-based map of the human proteome. *Science* **347**, 1260419 (2015).

81. Sutherland, G. R. Heritable fragile sites on human chromosomes. IX. Population cytogenetics and segregation analysis of the BrdU-requiring fragile site at 10q25. *American Journal of Human Genetics* **34**, 753–756 (1982).

82. Rao, S. S. *et al.* A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* **159**, 1665–1680 (2014).

83. Di Bernardo, M. C. *et al.* A genome-wide association study identifies six susceptibility loci for chronic lymphocytic leukemia. *Nature Genetics* **40**, 1204–1210 (2008).

84. Slager, S. L. *et al.* Genome-wide association study identifies a novel susceptibility locus at 6p21.3 among familial CLL. *Blood* **117**, 1911–1916 (2011).

85. Slager, S. L. *et al.* Common variation at 6p21.31 (*BAK1*) influences the risk of chronic lymphocytic leukemia. *Blood* **120**, 843–846 (2012).

86. Berndt, S. I. *et al.* Genome-wide association study identifies multiple risk loci for chronic lymphocytic leukemia. *Nature Genetics* **45**, 868–876 (2013).

87. Speedy, H. E. *et al.* A genome-wide association study identifies multiple susceptibility loci for chronic lymphocytic leukemia. *Nature Genetics* **46**, 56–60 (2014).

88. Tapper, W. *et al.* Genetic variation at *MECOM*, *TERT*, *JAK2* and *HBS1L-MYB* predisposes to myeloproliferative neoplasms. *Nature Communications* **6** (2015).

89. Codd, V. *et al.* Identification of seven loci affecting mean telomere length and their association with disease. *Nature Genetics* **45**, 422–427 (2013).

90. Machiela, M. J. & Chanock, S. J. LDlink: a web-based application for exploring population-specific haplotype structure and linking correlated alleles of possible functional variants. *Bioinformatics* **31**, 3555–3557 (2015).

Supplementary Table 1. Number of mCAs detected per chromosome.

Chromosome	N_{loss}	$N_{\text{CNN-LOH}}$	N_{gain}	$N_{\text{undetermined}}$	N_{total}
chr1	29	318	17	134	498
chr2	66	56	10	48	180
chr3	18	53	41	63	175
chr4	47	64	8	41	160
chr5	49	40	24	38	151
chr6	32	68	6	64	170
chr7	70	43	5	40	158
chr8	22	35	42	44	143
chr9	19	210	38	78	345
chr10	70	29	5	31	135
chr11	98	257	1	105	461
chr12	28	67	156	95	346
chr13	177	111	0	73	361
chr14	51*	223	38	135	447
chr15	14	121	59	93	287
chr16	43	142	2	53	240
chr17	66	112	37	89	304
chr18	14	20	57	40	131
chr19	6	90	17	75	188
chr20	140	55	3	29	227
chr21	20	35	31	67	153
chr22	39*	88	62	113	302
All autosomes	1118	2237	659	1548	5562
Female chrX	1862	28	24	866	2780

*Deletions on chr14 and chr22 include V(D)J recombination events (25 events on chr14 and 25 events on chr22).

Supplementary Table 2. Distribution of the number of detected somatic autosomal mCAs per individual.

mCA count	Frequency
0	146313
1	4448
2	295
3	103
4	27
5	7
6	4
7	0
8	2
9	1
10	0
11	1
12	1

Most individuals with several detected mCAs have prevalent or incident cancers.

Supplementary Table 3. Co-occurrence enrichment among mCAs.

mCA1	mCA2	*P*	OR (95% CI)
3+	12+	3.1×10^{-10}	170 (65–444)
3p–	13q–	1.4×10^{-7}	410 (105–1598)
3+	13q–	7.1×10^{-8}	120 (42–344)
3+	18+	2.7×10^{-18}	829 (345–1991)
4+	18+	1.3×10^{-9}	2361 (515–10832)
8+	9+	1.1×10^{-7}	381 (112–1298)
12+	13q–	1.5×10^{-8}	41 (18–94)
12+	18+	1.1×10^{-33}	473 (253–884)
12+	19+	8.9×10^{-34}	3331 (1061–10457)
12+	22q–	4.5×10^{-8}	135 (47–388)
13q–	13q=	4.1×10^{-67}	208 (137–313)
13q–	14q–	3.7×10^{-19}	288 (135–616)
13q=	14q–	3.2×10^{-6}	120 (36–396)
13q–	22q–	6.3×10^{-8}	124 (43–356)
13q=	22q–	2.1×10^{-6}	139 (42–460)
13q–	X+	8.8×10^{-10}	403 (130–1255)
17p–	21q–	2.7×10^{-12}	1919 (565–6522)
18+	19+	3.7×10^{-21}	2671 (953–7489)

We report pairs of mCA types (grouped by chromosome arm and copy number) with significant co-occurrence ($P<8\times10^{-6}$, Fisher's exact test with Bonferroni correction, and at least three individuals carrying both events). (We subdivided loss and CNN-LOH events by p-arm vs. q-arm, but we did not subdivide gain events by arm because most gain events are whole-chromosome trisomies; e.g., "3+" combines all gains—partial or complete—on chromosome 3.) We excluded individuals with >3 detected mCAs from these calculations to prevent individuals with large numbers of mCAs (typically cancer cases) from dominating the results, leaving n=151,159 individuals. Co-occurrence of 13q– and 13q= events (i.e., 13q14 deletion and 13q CNN-LOH, a frequent combination in chronic lymphocytic leukemia) was computed using a slightly different procedure than the rest of the table because these events affect both homologous copies of chr13, creating a special case not considered by our detection algorithm (which calls only 13q CNN-LOH in this circumstance). Specifically, we called 13q14 deletions based on mean total intensity (LRR) in 13q14 (50.6–51.6Mb); we then computed co-occurrence with 13q CNN-LOH events.

Supplementary Table 4. Fraction of individuals with detected mCAs as a function of age.

Age range	% with autosomal event	% of females with chrX event
<45	1.7% (0.1%)	0.9% (0.1%)
45-50	2.0% (0.1%)	1.1% (0.1%)
50-55	2.3% (0.1%)	1.7% (0.1%)
55-60	3.0% (0.1%)	3.0% (0.1%)
60-65	4.0% (0.1%)	4.7% (0.2%)
>65	4.9% (0.1%)	7.2% (0.2%)

This table provides numerical data plotted in Fig. 2d.

Supplementary Table 5. Age and sex distributions of individuals with detected mCAs on each chromosome.

	Loss events				CNN-LOH events				Gain events	
	p-arm		q-arm		p-arm		q-arm			
chr	Mean age	Frac. male	Mean age	Frac. male	Mean age	Frac. male	Mean age	Frac. male	Mean age	Frac. male
1	61.0 (1.9)	0.54 (0.14)	58.8 (1.8)	0.69 (0.12)	59.5 (0.5)	0.49 (0.04)	59.5 (0.6)	0.50 (0.04)	61.4 (1.5)	0.41 (0.12)
2	62.0 (0.8)	0.40 (0.07)	61.0 (2.3)	0.62 (0.14)	60.6 (1.1)	0.38 (0.09)	58.0 (1.3)	0.26 (0.09)	54.7 (2.7)	0.40 (0.16)
3	57.1 (2.3)	0.50 (0.15)	–	–	59.8 (1.6)	0.45 (0.11)	59.1 (1.6)	0.47 (0.09)	61. 5 (1.0)	0.74 (0.07)
4	–	–	61.8 (1.0)	0.56 (0.08)	53.3 (2.7)	0.56 (0.18)	62.4 (0.9)	0.50 (0.07)	63. 2 (2.3)	0.62 (0.18)
5	–	–	60.3 (1.1)	0.49 (0.08)	–	–	57.9 (1.4)	0.50 (0.08)	61.5 (1.2)	0.57 (0 .11)
6	64.4 (1.3)	0.17 (0.17)	60.8 (1.5)	0.58 (0.10)	56.2 (1.0)	0.43 (0.07)	58.3 (2.3)	0.47 (0.13)	57.7 (3.4)	0.50 (0.22)
7	61.4 (2.3)	0.25 (0.16)	62.0 (0.8)	0.56 (0.07)	61.4 (1.5)	0.50 (0.14)	57.6 (1.9)	0.62 (0.10)	59.1 (4.6)	0.20 (0.20)
8	61.2 (2.0)	0.47 (0.13)	63.5 (1.1)	0.71 (0.18)	–	–	57.2 (1.2)	0.48 (0.09)	61. 2 (1.0)	0.50 (0.08)
9	–	–	59.1 (2.6)	0.47 (0.13)	59.7 (0.7)	0.56 (0.05)	59.3 (0.8)	0.51 (0.05)	61. 2 (1.1)	0.55 (0.08)
10	–	–	56.8 (1.0)	0.20 (0.05)	61.2 (2.8)	0.33 (0.17)	58.8 (1.9)	0.30 (0.11)	60. 6 (4.6)	0.40 (0.24)
11	57.5 (2.5)	0.54 (0.14)	62.0 (0.7)	0.60 (0.05)	58.3 (0.6)	0.54 (0.04)	61.7 (0.6)	0.55 (0.05)	–	–
12	62.0 (1.9)	0.25 (0.13)	60.0 (1.5)	0.47 (0.13)	58.2 (2.7)	0.42 (0.15)	60.5 (1.0)	0.47 (0.07)	62.4 (0.5)	0.54 (0.04)
13	–	–	61.5 (0.4)	0.64 (0.04)	–	–	59.5 (0.8)	0.59 (0.05)	–	–
14	–	–	61.1 (0.8)	0.72 (0.07)	–	–	59.9 (0.5)	0.46 (0.03)	62.9 (0.7)	0.61 (0 .08)
15	–	–	62.5 (2.0)	0.64 (0.13)	–	–	59.5 (0.7)	0.51 (0.05)	65.7 (0.4)	0.83 (0 .05)
16	56.1 (1.4)	0.28 (0.08)	63.2 (1.5)	0.71 (0.13)	59.1 (0.9)	0.54 (0.06)	60.1 (0.9)	0.48 (0.06)	–	–
17	61.1 (1.0)	0.52 (0.07)	59.5 (1.9)	0.56 (0.13)	58.5 (1.6)	0.41 (0.11)	58.1 (0.8)	0.44 (0.05)	60.3 (1.2)	0.46 (0.08)
18	55.5 (2.9)	0.67 (0.21)	61.2 (2.6)	0.50 (0.22)	–	–	61.5 (1.7)	0.35 (0.12)	62. 2 (0.8)	0.70 (0.06)
19	60.8 (2.6)	0.80 (0.20)	–	–	59.2 (1.2)	0.43 (0.08)	60.6 (1.0)	0.53 (0.07)	60. 9 (1.5)	0.76 (0.11)
20	–	–	62.1 (0.6)	0.70 (0.04)	59.1 (2.6)	0.45 (0.16)	57.9 (1.3)	0.38 (0.08)	–	–
21	–	–	59.2 (1.8)	0.37 (0.11)	–	–	57.4 (1.5)	0.56 (0.09)	60.8 (1.1)	0.81 (0 .07)
22	–	–	62.8 (0.7)	0.66 (0.08)	–	–	60.7 (0.8)	0.36 (0.05)	61.2 (0.8)	0.52 (0 .06)
X	60.3 (2.3)	–	59.0 (2.5)	–	61.4 (3.0)	–	60.3 (1.1)	–	56.8 (2.0)	–

This table provides numerical data plotted in Fig. 2e. (Events detected in fewer than 15 individuals and female chrX events were excluded from Fig. 2e for clarity, and events detected in fewer than 5 individuals are excluded here.)

Supplementary Table 6. Enrichment of mCAs in individuals with anomalous (top 1%) blood indices.

mCA	Blood index	*P*-value (one-sided Fisher)	*q*-value	OR (95% CI)
1p–	Lymphocyte #	0.0027	0.047	33.1 (6.7–163.9)
1p–	Lymphocyte %	0.0027	0.047	33.1 (6.7–163.9)
2p=	Monocyte #	0.0027	0.047	11.9 (3.6–39.5)
3p–	Lymphocyte #	0.002	0.038	39.7 (7.7–204.6)
3p–	Lymphocyte %	0.002	0.038	39.7 (7.7–204.6)
3+	Lymphocyte #	3.6e-6	0.00015	26.1 (9.7–70.1)
3+	Lymphocyte %	3.6e-6	0.00015	26.1 (9.7–70.1)
4q=	Monocyte %	2.3e-7	1.2e-5	19.3 (8.6–43.5)
7q–	Lymphocyte #	3.3e-5	0.00097	15.5 (6.0–39.9)
7q–	Lymphocyte %	3.3e-5	0.00097	15.5 (6.0–39.9)
9p=	Red #	1.1e-13	7.6e-12	17.7 (10.2–30.6)
9p=	Hematocrit	3e-11	2e-9	14.9 (8.3–26.8)
9p=	RBC dist. width	2.8e-16	2.5e-14	20.5 (12.1–34.7)
9p=	Platelet #	1.9e-32	4.8e-30	39.3 (25.3–61.0)
9p=	Platelet crit	4.7e-34	1.6e-31	41.3 (26.7–63.8)
9p=	Platelet dist. width	7e-5	0.0019	7.5 (3.5–16.2)
9+	Neutrophil #	1.1e-5	0.0004	19.9 (7.6–52.0)
9+	Neutrophil %	0.00022	0.0054	15.3 (5.3–43.8)
9+	RBC dist. width	1.1e-5	0.0004	19.9 (7.6–52.0)
9+	Platelet #	0.00022	0.0054	15.3 (5.3–43.8)
11q–	Lymphocyte #	4.2e-8	2.3e-6	14.5 (7.2–29.2)
11q–	Lymphocyte %	8.1e-5	0.0021	9.2 (4.0–21.2)
11q–	Platelet dist. width	8.1e-5	0.0021	9.2 (4.0–21.2)
11q=	Lymphocyte #	0.0001	0.0026	7.0 (3.3–15.2)
12+	Lymphocyte #	2.2e-20	3.2e-18	22.2 (13.8–35.7)
12+	Lymphocyte %	3.7e-15	3e-13	17.2 (10.3–28.9)
13q–	Lymphocyte #	3.3e-117	3.3e-114	163.4 (113.3–235.7)
13q–	Lymphocyte %	8e-96	4e-93	116.3 (81.3–166.4)
13q–	Basophil #	4.2e-10	2.6e-8	11.8 (6.6–21.0)
13q–	Basophil %	0.0016	0.03	5.1 (2.2–11.6)
13q–	Monocyte #	3.7e-5	0.001	6.9 (3.4–14.2)
13q=	Lymphocyte #	5.2e-17	5.2e-15	23.0 (13.6–39.1)
13q=	Lymphocyte %	2.5e-14	1.9e-12	19.7 (11.3–34.4)
14q–	Lymphocyte #	6.4e-20	7.1e-18	73.7 (36.9–147.3)
14q–	Lymphocyte %	6.4e-20	7.1e-18	73.7 (36.9–147.3)
14q–	Basophil #	0.00032	0.0075	13.7 (4.8–39.0)
14q=	Monocyte %	0.00085	0.018	4.3 (2.1–8.7)
16p–	Monocyte %	0.0022	0.04	12.9 (3.9–43.2)
16q–	Lymphocyte #	4.6e-6	0.00018	49.7 (14.9–165.1)
16q–	Lymphocyte %	4.6e-6	0.00018	49.7 (14.9–165.1)
16p=	Monocyte %	0.0009	0.019	7.2 (2.9–17.9)
17p–	Lymphocyte #	4.6e-9	2.7e-7	25.7 (11.8–56.0)
17p–	Lymphocyte %	0.00062	0.013	11.3 (4.0–32.0)
17q–	Platelet dist. width	0.00033	0.0076	27.1 (7.5–97.1)
18+	Lymphocyte #	0.00056	0.012	11.7 (4.1–33.0)
19+	Lymphocyte #	6.6e-6	0.00024	44.1 (13.6–143.5)
19+	Lymphocyte %	0.00026	0.0063	29.8 (8.2–108.3)
20q–	Neutrophil %	0.001	0.02	5.6 (2.4–12.7)
20q–	RBC dist. width	2e-5	0.00062	7.6 (3.7–15.6)
20q–	Platelet dist. width	0.001	0.02	5.6 (2.4–12.7)
22q–	Lymphocyte #	1.6e-31	3.2e-29	190.7 (88.5–410.9)
22q–	Lymphocyte %	5.5e-25	9.1e-23	123.3 (59.2–256.8)
22+	Lymphocyte #	5e-8	2.6e-6	18.1 (8.5–38.5)
22+	Lymphocyte %	1.4e-5	0.00044	13.0 (5.5–30.4)
–X	Lymphocyte #	1.5e-6	7.1e-5	2.4 (1.8–3.4)
–X	Lymphocyte %	3.7e-6	0.00015	2.4 (1.7–3.3)

This table provides numerical data plotted in Fig. 2f. Mosaic chromosomal alterations significantly enriched (at an FDR threshold of 0.05) in individuals with anomalous blood indices (top 1% of n=144,637 self-reported white individuals) are reported. Events were grouped by chromosome and copy number, with loss and CNN-LOH events subdivided by p-arm vs. q-arm. (We did not subdivide gain events by arm because most gain events are whole-chromosome trisomies; e.g., "3+" combines all gains—partial or complete—on chromosome 3.)

Supplementary Table 7. Association of *FRA10B* variable number tandem repeat motifs with breakage at 10q25.2.

(a) Variable number tandem repeats imputed into UK Biobank

Variant	MAF	#del(10q)	*P*	Imputation R^2
VNTR-38-a	0.0007	3/60	5×10^{-5}	0.55
VNTR-39-a	0.0000	0/60	0.5	0.16
VNTR-42-a	0.0010	16/60	3×10^{-27}	0.64
VNTR-42-b	0.0001	0/60	0.5	0.26
VNTR-42-c	0.0002	0/60	0.5	0.79
VNTR-42-d	0.0001	0/60	0.5	0.63
VNTR-42-e	0.0000	0/60	0.5	0.15
VNTR-43-a	0.0003	0/60	0.5	0.35
VNTR-43-b	0.0027	5/60	$9\times10^{-}6$	0.64
VNTR-43-c	0.0004	0/60	0.5	0.58
VNTR-43-d	0.0003	0/60	0.5	0.75
VNTR-43-e	0.0000	0/60	0.5	0.14

(b) Lead associated SNPs typed or imputed in UK Biobank

Variant	MAF	#del(10q)	*P*	INFO
rs118137427	0.0527	60/60	6×10^{-42}	1.000 (typed)
rs758889647	0.0015	13/60	4×10^{-21}	0.695

Results are from Fisher's exact test on n=120,664 individuals. All 12 high-confidence non-reference VNTR motifs we identified (Extended Data Fig. 5a,b and Supplementary Note 8) occur on the rs118137427:G haplotype background, which is carried by all chromosomes with detected mosaic breakage at 10q25.2. VNTR-42-a, carried by the four del(10q) individuals in the WGS cohort, is well-tagged by the rare rs758889647:A allele and imputes into 16 of 60 UK Biobank del(10q) individuals. VNTR-43-b imputes into five del(10q) individuals, and VNTR-38-a imputes into an IBD cluster of three del(10q) individuals (Extended Data Fig. 5a,b).

Supplementary Table 8. SNPs at *MPL* and *ATM* associated with *cis* somatic CNN-LOH at $p<10^{-7}$.

SNP	hg19 coordinates	Alleles	RAF	*P*	OR (95% CI)
***MPL* locus: associations with chr1p CNN-LOH**					
rs543652228	1:43640972	A/G	0.0003	2.4×10^{-9}	51 (22–118)
rs777132997	1:43669098	A/G	0.0002	2.0×10^{-10}	79 (34–187)
rs757080968	1:43720418	C/G	0.0002	2.6×10^{-10}	76 (32–178)
rs547321640	1:43752900	T/C	0.0002	1.0×10^{-8}	71 (28–180)
rs538358508	1:43753105	T/G	0.0002	1.0×10^{-8}	71 (28–180)
rs549761468	1:43788667	C/T	0.0002	2.1×10^{-10}	79 (34–187)
rs143549194	1:43815673	G/T	0.0015	2.1×10^{-8}	14 (7–27)
rs369156948	1:43817942	C/T	0.0001	7.3×10^{-8}	103 (35–300)
rs576674585	1:43892277	A/C	0.0001	4.9×10^{-9}	83 (32–214)
rs558677971	1:43895592	G/A	0.0002	2.4×10^{-8}	59 (23–149)
rs566497062	1:43897662	C/T	0.0002	2.4×10^{-8}	59 (23–149)
rs143305686	1:44134295	A/G	0.0018	1.7×10^{-12}	17 (10–30)
rs773168056	1:44156366	A/G	0.0003	4.2×10^{-9}	46 (20–106)
rs182971382	1:44167774	A/G	0.0003	3.0×10^{-11}	63 (29–139)
rs554498272	1:44190215	G/A	0.0001	4.8×10^{-11}	103 (43–248)
rs765697775	1:44546545	C/T	0.0006	9.5×10^{-15}	41 (22–76)
rs540740393	1:45126775	C/A	0.0018	3.1×10^{-10}	15 (8–27)
rs553066968	1:45129752	A/T	0.0019	5.9×10^{-10}	14 (8–26)
rs572698005	1:45129772	C/T	0.0019	5.9×10^{-10}	14 (8–26)
rs565464974	1:45170759	G/A	0.0009	2.4×10^{-13}	30 (16–55)
rs748989559	1:45173569	A/G	0.0005	6.7×10^{-16}	53 (28–98)
rs548041003	1:45175146	C/T	0.0021	6.3×10^{-13}	16 (9–27)
rs144279563	1:45294379	C/T	0.0005	6.2×10^{-16}	53 (28–99)
rs572162077	1:45354774	G/C	0.0010	1.0×10^{-15}	31 (18–55)
***ATM* locus: associations with chr11q CNN-LOH**					
rs535473237	11:108074178	A/G	0.0004	1.8×10^{-8}	61 (25–152)
rs532198118	11:108355523	A/G	0.0007	7.4×10^{-9}	41 (18–94)

Results are from Fisher's exact test on n=120,664 individuals. Alleles: risk lowering/risk increasing allele. RAF: risk allele frequency (in UK Biobank European-ancestry individuals).

Supplementary Table 9. *cis* associations with biased loss of X ($P_{bias}<10^{-6}$) and X gain data.

				Loss of female chrX					Gain of female chrX				
SNP	Location	A1/A2	A2F	$A2F_{case}$	P_{GWAS}	N_{A1+}	N_{A2+}	P_{bias}	$A2F_{case}$	P_{GWAS}	N_{A1+}	N_{A2+}	P_{bias}
rs954958	X:55129982	C/T	0.471	0.452	4.9×10^{-3}	540	716	7.6×10^{-7}	0.407	0.25	4	6	0.75
rs10521478	X:55208161	A/G	0.417	0.397	7.7×10^{-4}	515	713	1.8×10^{-8}	0.370	0.38	5	5	1.00
rs1927307	X:55337294	G/A	0.294	0.278	4.1×10^{-3}	436	621	1.4×10^{-8}	0.241	0.33	1	5	0.22
rs5914315	X:55354496	T/C	0.316	0.299	3.0×10^{-3}	447	639	6.2×10^{-9}	0.296	0.65	2	5	0.45
rs12559108	X:55422562	T/C	0.260	0.243	1.4×10^{-3}	374	572	1.3×10^{-10}	0.204	0.46	1	4	0.38
rs7892090	X:55432212	T/C	0.259	0.242	1.5×10^{-3}	379	569	7.3×10^{-10}	0.241	0.88	1	4	0.38
rs57620007	X:55476740	T/C	0.259	0.242	1.1×10^{-3}	377	568	5.6×10^{-10}	0.222	0.79	1	4	0.38
rs3126241	X:55601683	T/C	0.253	0.234	2.3×10^{-4}	360	562	3.0×10^{-11}	0.222	0.72	1	4	0.38
rs149700928	X:55684550	G/C	0.251	0.232	2.3×10^{-4}	357	555	5.8×10^{-11}	0.222	0.75	1	4	0.38
rs5913856	X:55747717	A/G	0.249	0.230	1.4×10^{-4}	349	558	4.0×10^{-12}	0.222	0.77	1	4	0.38
rs1007153	X:55778139	C/T	0.272	0.251	7.0×10^{-5}	363	592	1.2×10^{-13}	0.259	0.96	1	4	0.38
rs5914476	X:55852696	T/G	0.271	0.250	2.3×10^{-5}	358	590	4.7×10^{-14}	0.259	0.98	1	4	0.38
rs6612385	X:55853321	A/G	0.272	0.251	4.5×10^{-5}	364	589	3.1×10^{-13}	0.259	0.96	1	4	0.38
rs10855058	X:55936822	G/A	0.273	0.254	1.4×10^{-4}	385	592	3.7×10^{-11}	0.222	0.50	1	5	0.22
rs6417935	X:55960724	C/T	0.135	0.126	9.9×10^{-3}	219	352	2.9×10^{-8}	0.018	0.05	0	1	1.00
rs6612472	X:56152985	A/G	0.241	0.222	1.1×10^{-4}	322	547	2.2×10^{-14}	0.167	0.30	2	3	1.00
rs4826461	X:56226649	A/G	0.234	0.218	4.5×10^{-4}	311	539	4.8×10^{-15}	0.148	0.22	2	2	1.00
rs6521388	X:56345127	A/G	0.218	0.206	4.8×10^{-3}	289	533	1.4×10^{-17}	0.111	0.11	1	1	1.00
rs5913935	X:56428273	T/C	0.135	0.124	4.4×10^{-3}	203	356	9.9×10^{-11}	0.037	0.09	1	1	1.00
rs5914638	X:56456144	T/C	0.233	0.218	1.6×10^{-3}	305	557	7.3×10^{-18}	0.185	0.56	3	1	0.62
rs1332731	X:56495976	T/C	0.249	0.233	5.3×10^{-4}	327	579	4.7×10^{-17}	0.204	0.59	3	2	1.00
rs721963	X:56558810	A/C	0.225	0.211	4.7×10^{-3}	294	551	7.0×10^{-19}	0.130	0.17	2	1	1.00
rs766912	X:56630987	A/G	0.224	0.210	1.7×10^{-3}	293	548	1.1×10^{-18}	0.130	0.20	2	1	1.00
rs74503599	X:56640134	C/T	0.240	0.223	3.5×10^{-4}	312	566	8.1×10^{-18}	0.148	0.19	2	2	1.00
rs5914806	X:56847280	A/G	0.180	0.169	7.2×10^{-3}	249	459	2.5×10^{-15}	0.074	0.09	1	1	1.00
rs5914815	X:56870961	T/C	0.179	0.169	8.6×10^{-3}	250	460	2.8×10^{-15}	0.074	0.10	1	1	1.00
rs5960832	X:56894267	C/T	0.210	0.222	7.9×10^{-3}	501	351	3.1×10^{-7}	0.167	0.38	2	4	0.69
rs5914035	X:57008216	T/C	0.225	0.212	3.3×10^{-3}	292	560	2.9×10^{-20}	0.148	0.28	3	2	1.00
rs912956	X:57010138	T/C	0.207	0.195	5.1×10^{-3}	265	532	1.9×10^{-21}	0.093	0.08	1	1	1.00
rs5914052	X:57129959	A/G	0.225	0.213	3.6×10^{-3}	293	563	1.8×10^{-20}	0.148	0.27	3	2	1.00
rs5960927	X:57241324	G/A	0.209	0.222	6.7×10^{-3}	500	347	1.6×10^{-7}	0.185	0.69	2	4	0.69
rs2516023	X:57313357	T/C	0.226	0.212	2.3×10^{-3}	291	553	1.3×10^{-19}	0.148	0.28	3	2	1.00
rs6611612	X:57329089	A/G	0.227	0.213	1.3×10^{-3}	290	551	1.6×10^{-19}	0.148	0.26	3	2	1.00
rs2060113	X:57478582	C/T	0.221	0.209	6.8×10^{-3}	288	550	9.8×10^{-20}	0.130	0.18	3	1	0.62
rs1594503	X:57480930	C/T	0.244	0.231	8.6×10^{-4}	318	581	1.4×10^{-18}	0.167	0.29	3	2	1.00
rs1997715	X:57622607	G/A	0.225	0.213	3.7×10^{-3}	294	550	9.1×10^{-19}	0.148	0.28	3	2	1.00
rs112877950	X:57624653	C/T	0.028	0.027	7.9×10^{-1}	30	98	1.3×10^{-9}	0.018	0.67	0	0	1.00
rs73226048	X:57979353	T/C	0.221	0.209	5.7×10^{-3}	283	545	5.8×10^{-20}	0.111	0.10	2	1	1.00
rs55950555	X:57985647	A/G	0.302	0.313	5.6×10^{-2}	618	434	1.5×10^{-8}	0.333	0.50	1	4	0.38
rs113699645	X:58121440	A/G	0.026	0.025	6.9×10^{-1}	29	86	9.8×10^{-8}	0.018	0.72	0	0	1.00
rs4625204	X:58216902	A/G	0.202	0.215	4.2×10^{-3}	499	338	2.9×10^{-8}	0.222	0.77	1	5	0.22
rs111318471	X:58328362	C/A	0.026	0.026	6.8×10^{-1}	29	82	4.9×10^{-7}	0.018	0.76	0	0	1.00
rs2942875	X:58339545	C/T	0.447	0.429	9.7×10^{-4}	423	796	6.6×10^{-27}	0.315	0.07	6	1	0.12
rs112064215	X:61994151	C/T	0.053	0.050	2.8×10^{-1}	70	159	3.9×10^{-9}	0.056	0.96	1	0	1.00
rs60576970	X:61999396	A/C	0.493	0.513	9.4×10^{-4}	753	505	2.8×10^{-12}	0.500	0.88	1	5	0.22
rs62597976	X:62261609	G/T	0.300	0.322	1.1×10^{-4}	646	446	1.6×10^{-9}	0.259	0.44	1	6	0.12
rs56329621	X:62520485	G/A	0.032	0.029	3.4×10^{-1}	35	103	5.8×10^{-9}	0.037	0.33	1	0	1.00
rs1221064	X:62529141	A/G	0.085	0.078	2.6×10^{-2}	126	227	8.4×10^{-8}	0.074	0.87	1	0	1.00
rs112933767	X:63195237	A/G	0.042	0.041	9.2×10^{-1}	63	132	8.7×10^{-7}	0.056	0.25	1	1	1.00
rs73213355	X:64965828	C/T	0.060	0.061	6.0×10^{-1}	196	108	5.1×10^{-7}	0.074	0.76	1	1	1.00
rs3848896	X:65182724	G/A	0.096	0.096	7.0×10^{-1}	287	156	4.9×10^{-10}	0.111	0.79	3	1	0.62
rs7056244	X:65206855	G/A	0.070	0.074	1.9×10^{-1}	240	121	3.7×10^{-10}	0.111	0.32	3	1	0.62
rs5918586	X:65328292	A/G	0.136	0.136	6.8×10^{-1}	358	227	6.8×10^{-8}	0.130	0.78	4	1	0.38
rs12836051	X:114924811	A/G	0.160	0.148	5.5×10^{-3}	257	405	9.7×10^{-9}	0.125	0.50	2	4	0.69
rs73224841	X:114931929	T/G	0.022	0.022	7.6×10^{-1}	32	86	6.9×10^{-7}	0.018	0.81	1	0	1.00
rs73224844	X:114945104	G/A	0.022	0.022	5.3×10^{-1}	30	86	1.9×10^{-7}	0.018	0.83	1	0	1.00
rs11091036	X:115023111	G/C	0.266	0.249	1.1×10^{-3}	369	555	1.0×10^{-9}	0.304	0.50	6	6	1.00

N=66,685 females were analyzed. A1, A2: major/minor allele. A2F: minor allele frequency. $A2F_{case}$: A2 frequency in individuals with loss (resp. gain) of X. P_{GWAS}: association with increased risk of X event. N_{A1+}: number of heterozygous individuals with X loss (resp. gain) in which the A1/A2 allelic balance shifts toward the A1 allele (and analogously for N_{A2+}). P_{bias}: P-value for biased shift.

Supplementary Table 10. No evidence for rs2942875-biased X inactivation in GEUVADIS RNA-seq data.

HG00122	Read counts			HG00130	Read counts			HG00133	Read counts			HG00158	Read counts		
rs2516023 T/C	2	1		rs2516023 T/C	8	0		rs2516023 T/C	2	2		rs2516023 T/C	3	1	
rs1367830 C/T	3	2		rs1367830 C/T	9	0		rs1367830 C/T	6	8		rs1367830 C/T	2	5	
rs2060113 C/T	1	1		rs2060113 C/T	1	0		rs2060113 C/T	2	1		rs2060113 C/T	1	2	
Total maj/min	6	4	0.60	Total maj/min	18	0	1.00	Total maj/min	10	11	0.48	Total maj/min	6	8	0.43
HG00231	Read counts			HG00232	Read counts			HG00239	Read counts			HG00257	Read counts		
rs2516023 T/C	0	5		rs2516023 T/C	0	1		rs2516023 T/C	3	2		rs2516023 T/C	1	0	
rs1367830 C/T	0	8		rs1367830 C/T	0	6		rs1367830 C/T	4	3		rs1367830 C/T	1	1	
rs2060113 C/T	0	4		rs2060113 C/T	0	4		rs2060113 C/T	1	2		rs2060113 C/T	0	1	
Total maj/min	0	17	0.00	Total maj/min	0	11	0.00	Total maj/min	8	7	0.53	Total maj/min	2	2	0.50
HG00266	Read counts			HG00276	Read counts			HG00315	Read counts			HG00323	Read counts		
rs2516023 T/C	2	0		rs2516023 T/C	0	2		rs2516023 T/C	2	3		rs2516023 T/C	4	4	
rs1367830 C/T	10	0		rs1367830 C/T	1	10		rs1367830 C/T	6	2		rs1367830 C/T	3	3	
rs2060113 C/T	9	0		rs2060113 C/T	0	3		rs2060113 C/T	1	1		rs2060113 C/T	1	0	
Total maj/min	21	0	1.00	Total maj/min	1	15	0.06	Total maj/min	9	6	0.60	Total maj/min	8	7	0.53
HG00327	Read counts			HG00332	Read counts			HG00334	Read counts			HG00337	Read counts		
rs2516023 T/C	0	4		rs2516023 T/C	0	8		rs2516023 T/C	0	4		rs2516023 T/C	2	1	
rs1367830 C/T	0	4		rs1367830 C/T	1	6		rs1367830 C/T	0	8		rs1367830 C/T	2	2	
rs2060113 C/T	0	2		rs2060113 C/T	1	3		rs2060113 C/T	0	3		rs2060113 C/T	0	0	
Total maj/min	0	10	0.00	Total maj/min	2	17	0.11	Total maj/min	0	15	0.00	Total maj/min	4	3	0.57
HG00353	Read counts			HG00362	Read counts			HG00364	Read counts			HG00381	Read counts		
rs2516023 T/C	0	0		rs2516023 T/C	0	2		rs2516023 T/C	8	2		rs2516023 T/C	1	0	
rs1367830 C/T	0	12		rs1367830 C/T	3	5		rs1367830 C/T	7	6		rs1367830 C/T	1	4	
rs2060113 C/T	1	4		rs2060113 C/T	2	1		rs2060113 C/T	3	3		rs2060113 C/T	1	3	
Total maj/min	1	16	0.06	Total maj/min	5	8	0.38	Total maj/min	18	11	0.62	Total maj/min	3	7	0.30
HG01790	Read counts			NA06985	Read counts			NA07037	Read counts			NA07056	Read counts		
rs2516023 T/C	0	0		rs2516023 T/C	2	0		rs2516023 T/C	7	0		rs2516023 T/C	0	3	
rs1367830 C/T	3	2		rs1367830 C/T	4	0		rs1367830 C/T	13	0		rs1367830 C/T	1	1	
rs2060113 C/T	0	2		rs2060113 C/T	6	0		rs2060113 C/T	7	0		rs2060113 C/T	0	1	
Total maj/min	3	4	0.43	Total maj/min	12	0	1.00	Total maj/min	27	0	1.00	Total maj/min	1	5	0.17
NA11830	Read counts			NA11832	Read counts			NA11892	Read counts			NA11931	Read counts		
rs2516023 T/C	1	2		rs2516023 T/C	0	6		rs2516023 T/C	3	0		rs2516023 T/C	0	4	
rs1367830 C/T	3	6		rs1367830 C/T	0	9		rs1367830 C/T	4	0		rs1367830 C/T	0	1	
rs2060113 C/T	1	3		rs2060113 C/T	0	1		rs2060113 C/T	2	0		rs2060113 C/T	0	0	
Total maj/min	5	11	0.31	Total maj/min	0	16	0.00	Total maj/min	9	0	1.00	Total maj/min	0	5	0.00
NA12058	Read counts			NA12156	Read counts			NA12234	Read counts			NA12275	Read counts		
rs2516023 T/C	0	10		rs2516023 T/C	1	4		rs2516023 T/C	1	0		rs2516023 T/C	0	6	
rs1367830 C/T	0	11		rs1367830 C/T	4	5		rs1367830 C/T	5	1		rs1367830 C/T	0	12	
rs2060113 C/T	0	3		rs2060113 C/T	0	1		rs2060113 C/T	1	0		rs2060113 C/T	0	7	
Total maj/min	0	24	0.00	Total maj/min	5	10	0.33	Total maj/min	7	1	0.88	Total maj/min	0	25	0.00
NA12283	Read counts			NA12341	Read counts			NA12383	Read counts			NA12489	Read counts		
rs2516023 T/C	2	0		rs2516023 T/C	7	1		rs2516023 T/C	2	0		rs2516023 T/C	0	0	
rs1367830 C/T	10	0		rs1367830 C/T	9	0		rs1367830 C/T	10	1		rs1367830 C/T	1	5	
rs2060113 C/T	3	0		rs2060113 C/T	6	0		rs2060113 C/T	4	0		rs2060113 C/T	2	1	
Total maj/min	15	0	1.00	Total maj/min	22	1	0.96	Total maj/min	16	1	0.94	Total maj/min	3	6	0.33
NA12718	Read counts			NA12815	Read counts			NA12843	Read counts			NA12890	Read counts		
rs2516023 T/C	0	2		rs2516023 T/C	0	3		rs2516023 T/C	1	6		rs2516023 T/C	3	0	
rs1367830 C/T	0	9		rs1367830 C/T	1	7		rs1367830 C/T	1	5		rs1367830 C/T	10	0	
rs2060113 C/T	0	4		rs2060113 C/T	0	3		rs2060113 C/T	1	4		rs2060113 C/T	5	0	
Total maj/min	0	15	0.00	Total maj/min	1	13	0.07	Total maj/min	3	15	0.17	Total maj/min	18	0	1.00
NA20502	Read counts			NA20503	Read counts			NA20505	Read counts			NA20507	Read counts		
rs2516023 T/C	2	0		rs2516023 T/C	0	0		rs2516023 T/C	4	1		rs2516023 T/C	3	0	
rs1367830 C/T	4	0		rs1367830 C/T	1	0		rs1367830 C/T	7	0		rs1367830 C/T	6	4	
rs2060113 C/T	0	0		rs2060113 C/T	1	0		rs2060113 C/T	3	0		rs2060113 C/T	5	2	
Total maj/min	6	0	1.00	Total maj/min	2	0	1.00	Total maj/min	14	1	0.93	Total maj/min	14	6	0.70
NA20508	Read counts			NA20514	Read counts			NA20529	Read counts			NA20531	Read counts		
rs2516023 T/C	3	0		rs2516023 T/C	2	2		rs2516023 T/C	5	0		rs2516023 T/C	4	1	
rs1367830 C/T	3	1		rs1367830 C/T	3	3		rs1367830 C/T	11	1		rs1367830 C/T	6	7	
rs2060113 C/T	1	0		rs2060113 C/T	2	1		rs2060113 C/T	3	0		rs2060113 C/T	3	4	
Total maj/min	7	1	0.88	Total maj/min	7	6	0.54	Total maj/min	19	1	0.95	Total maj/min	13	12	0.52
NA20541	Read counts			NA20582	Read counts			NA20585	Read counts			NA20589	Read counts		
rs2516023 T/C	5	0		rs2516023 T/C	4	2		rs2516023 T/C	0	2		rs2516023 T/C	0	0	
rs1367830 C/T	4	0		rs1367830 C/T	12	4		rs1367830 C/T	0	5		rs1367830 C/T	6	0	
rs2060113 C/T	0	0		rs2060113 C/T	4	2		rs2060113 C/T	0	1		rs2060113 C/T	2	0	
Total maj/min	9	0	1.00	Total maj/min	20	8	0.71	Total maj/min	0	8	0.00	Total maj/min	8	0	1.00
NA20756	Read counts			NA20761	Read counts			NA20771	Read counts			NA20797	Read counts		
rs2516023 T/C	2	13		rs2516023 T/C	1	6		rs2516023 T/C	4	2		rs2516023 T/C	11	0	
rs1367830 C/T	0	8		rs1367830 C/T	3	8		rs1367830 C/T	3	6		rs1367830 C/T	9	1	
rs2060113 C/T	0	0		rs2060113 C/T	1	2		rs2060113 C/T	2	0		rs2060113 C/T	4	0	
Total maj/min	2	21	0.09	Total maj/min	5	16	0.24	Total maj/min	9	8	0.53	Total maj/min	24	1	0.96
NA20799	Read counts			NA20800	Read counts			NA20807	Read counts			NA20813	Read counts		
rs2516023 T/C	0	4		rs2516023 T/C	0	1		rs2516023 T/C	1	3		rs2516023 T/C	0	4	
rs1367830 C/T	0	8		rs1367830 C/T	0	11		rs1367830 C/T	3	8		rs1367830 C/T	1	7	
rs2060113 C/C	–	–		rs2060113 C/T	0	4		rs2060113 C/T	3	4		rs2060113 C/T	1	4	
Total maj/min	0	12	0.00	Total maj/min	0	16	0.00	Total maj/min	7	15	0.32	Total maj/min	2	15	0.12
NA20819	Read counts														
rs2516023 T/C	4	0													
rs1367830 C/T	5	2													
rs2060113 C/T	3	1													
Total maj/min	12	3	0.80												

RNA-seq reads at three coding SNPs in LD with rs2942875 (the strongest *cis* association for biased loss of X) show consistent allele-specific expression within most individuals, as expected from X-chromosome inactivation that favors one homologous chromosome. However, across individuals, neither haplotype appears to be favored (30 individuals have more major-haplotype reads and 30 have more minor reads).

Supplementary Table 11. *trans* association with classes of mCAs at SNPs previously reported to be associated with related phenotypes.

SNP	Location	Gene(s) reported	MAF	GWAS trait	P_{any}	P_{loss}	$P_{CNN\text{-}LOH}$	P_{gain}	P_{auto}	$P_{auto\ loss}$	$P_{X\ loss}$
rs2736609	1:156202640	*PMF1, SEMA4A*	0.36	mLOY	0.5	0.69	0.47	0.92	0.68	0.62	0.95
rs11125529	2:54475866	*ACYP2*	0.14	telo	0.55	0.35	0.082	1	0.21	0.95	0.25
rs13401811	2:111616104	*ACOXL, BCL2L11*	0.18	CLL	0.57	0.67	0.71	0.74	0.51	0.73	0.84
rs17483466	2:111797458	*ACOXL, BCL2L11*	0.2	CLL	0.12	0.76	0.11	0.92	0.15	0.72	0.5
rs58055674	2:111831793	*ACOXL*	0.18	CLL	0.2	0.45	0.75	0.78	0.56	0.95	0.28
rs1439287	2:111871897	*ACOXL, BCL2L11*	0.49	CLL	0.28	0.28	0.71	0.59	0.92	0.21	0.36
rs9308731	2:111908262	*BCL2L11*	0.45	CLL	0.37	0.55	0.51	0.4	0.96	0.14	0.21
rs13015798	2:201909515	*FAM126B, CASP8*	0.33	CLL	0.0067	0.59	0.11	0.061	0.015	0.87	0.16
rs3769825	2:202111380	*CASP8, CASP10*	0.43	CLL	0.14	0.032	0.78	0.21	0.49	0.24	0.095
rs13397985	2:231091223	*SP140*	0.19	CLL	0.028	0.00026	0.91	0.25	0.13	0.0049	0.015
rs9880772	3:27777779	*EOMES*	0.45	CLL	0.69	0.16	0.59	0.14	0.97	0.6	0.87
rs115854006	3:48388170	*TREX1, PLXNB1*	0.036	mLOY	0.4	0.55	0.81	0.28	0.17	0.075	0.9
rs13088318	3:101242751	*SENP7*	0.34	mLOY	0.75	0.55	0.24	0.15	0.24	0.29	0.68
rs59633341	3:150018880	*TSC22D2*	0.16	mLOY	0.47	0.44	0.26	0.14	0.31	0.96	0.8
rs2201862	3:168648039	*EGFEM1P, MECOM*	0.5	MPN	0.13	0.38	0.75	0.0091	0.35	0.34	0.36
rs10936599	3:169492101	*MYNN*	0.25	CLL,telo	0.095	0.22	0.4	0.6	0.16	0.28	0.62
rs9815073	3:188115682	*LPP*	0.34	CLL	0.26	0.49	0.041	0.066	0.054	0.53	0.54
rs1548483	4:105749895	*TET2*	0.034	MPN	0.67	0.19	0.3	0.34	0.71	0.13	0.48
rs898518	4:109016824	*LEF1*	0.42	CLL	0.95	0.95	0.58	0.58	0.39	0.59	0.76
rs6858698	4:114683844	*CAMK2D*	0.16	CLL	0.63	0.57	0.24	0.54	0.76	0.052	0.69
rs7675998	4:164007820	*NAF1*	0.22	telo	0.48	0.6	0.69	0.62	0.42	0.085	0.67
rs34002450	5:1280940	*TERT*	0.38	CH	0.0031	0.092	0.0012	0.026	$\mathbf{7.8\times10^{-5}}$	0.0019	0.75
rs7705526	5:1285974	*TERT*	0.33	MPN	0.00052	0.036	8.6×10^{-5}	0.16	$\mathbf{4.8\times10^{-5}}$	0.0092	0.2
rs2736100	5:1286516	*TERT*	0.5	MPN,telo	0.0014	0.069	0.00095	0.12	0.00098	0.062	0.24
rs2853677	5:1287194	*TERT*	0.42	MPN	0.0043	0.44	0.00036	0.44	0.0014	0.38	0.92
rs56084922	5:111061883	NR	0.078	mLOY	0.58	0.38	0.73	0.19	0.64	0.36	0.78
rs9391997	6:409119	*IRF4*	0.47	CLL	0.92	0.62	0.38	0.93	0.66	0.73	0.68
rs872071	6:411064	*IRF4*	0.47	CLL	0.99	0.7	0.35	0.97	0.69	0.73	0.75
rs73718779	6:2969278	*SERPINB6*	0.11	CLL	0.59	0.86	0.85	0.57	0.57	0.73	0.02
rs926070	6:32257566	*HLA*	0.34	CLL	1	0.94	0.16	0.12	0.87	0.29	0.52
rs674313	6:32578082	*HLA-DRB5*	0.24	CLL	0.86	0.14	0.19	0.95	0.37	0.58	0.082
rs9273363	6:32626272	*HLA*	0.3	CLL	0.46	1	0.59	0.07	0.053	0.014	0.19
rs210142	6:33546837	*BAK1*	0.3	CLL	0.63	0.44	0.99	0.9	0.92	0.58	0.4
rs13191948	6:109634599	*SMPD2, CCDC162P*	0.46	mLOY	0.45	0.95	0.87	0.67	0.85	0.47	0.18
rs2236256	6:154478440	*IPCEF1*	0.46	CLL	0.72	0.099	0.41	0.39	0.82	0.2	0.53
rs381500	6:164478388	*QKI*	0.45	mLOY	0.49	0.63	0.17	0.43	0.083	0.068	0.56
rs4721217	7:1973579	*MAD1L1*	0.4	mLOY	0.0055	0.69	0.28	0.01	0.009	0.57	0.45
rs17246404	7:124462661	*POT1*	0.28	CLL	0.99	0.3	0.78	0.029	0.53	0.29	0.58
rs58270997	7:130729394	*PINT*	0.25	MPN	0.049	0.039	0.039	0.45	0.29	0.94	0.34
rs35091702	8:30279470	*RBPMS*	0.26	mLOY	0.58	0.21	0.88	0.85	0.52	0.97	0.055
rs2511714	8:103578874	*ODF1, KLF10*	0.4	CLL	0.034	0.13	0.34	0.46	0.6	0.32	0.011
rs2466035	8:128211229	*MYC*	0.33	CLL	0.59	0.55	0.25	0.65	0.89	0.25	0.34
rs59384377	9:5005034	*JAK2*	0.26	MPN	0.057	0.012	0.97	0.74	0.37	0.024	0.18
rs12339666	9:5063296	*JAK2*	0.26	MPN	0.11	0.027	0.98	0.87	0.4	0.032	0.35
rs10974944	9:5070831	*JAK2*	0.25	MPN	0.036	0.013	0.66	0.99	0.17	0.0097	0.46
rs1679013	9:22206987	*AS1, CDKN2B*	0.46	CLL	0.42	0.5	0.56	0.33	0.47	0.2	0.7
rs1359742	9:22336996	*DMRTA1, CDKN2B-AS1*	0.47	CLL	0.9	0.6	0.26	0.64	0.54	0.042	0.3
rs621940	9:135870130	*GFI1B*	0.16	MPN	0.74	0.52	0.073	0.25	0.44	0.18	0.52
rs1800682	10:90749963	*ACTA, FAS*	0.46	CLL	0.023	0.033	0.12	0.29	0.037	0.39	0.92
rs4406737	10:90759724	*ACTA2, FAS*	0.44	CLL	0.45	0.51	0.3	0.15	0.15	0.35	0.59
rs9420907	10:105676465	*OBFC1*	0.13	telo	0.32	0.057	0.99	0.87	0.45	0.059	0.13
rs7944004	11:2311152	*TSPAN32*	0.49	CLL	0.69	0.5	0.66	0.27	0.29	0.021	0.37
rs2521269	11:2321095	*C11orf21*	0.46	CLL	0.095	0.27	0.76	0.18	0.099	0.18	0.3
rs4754301	11:108048541	*NPAT, ATM, ACAT1*	0.45	mLOY	0.95	0.9	0.44	0.19	0.51	0.46	0.74
rs1800056	11:108138003	*ATM*	0.013	MPN	0.099	0.26	0.25	0.54	0.093	0.77	0.77
rs35923643	11:123355391	*GRAMD1B*	0.2	CLL	0.027	0.045	0.11	0.049	0.0091	0.071	0.31
rs735665	11:123361397	*SCN3B, GRAMD1B*	0.19	CLL	0.055	0.049	0.17	0.034	0.016	0.08	0.34
rs2953196	11:123368333	NR	0.25	CLL	0.049	0.1	0.81	0.22	0.06	0.31	0.87
rs7310615	12:111865049	*SH2B3*	0.48	MPN	0.39	0.47	0.85	0.86	0.86	0.33	0.25
rs10687116	13:41678081	*WBP4*	0.2	mLOY	0.76	0.59	0.72	0.6	0.8	0.99	0.73
rs1122138	14:96180242	*TCL1A*	0.16	mLOY	0.33	0.37	0.23	0.54	0.07	0.051	0.48
rs2887399	14:96180695	*TCL1A*	0.2	mLOY	0.31	0.79	0.088	0.61	0.064	0.095	0.49
rs137952017	14:101176090	*DLK1*	0.15	mLOY	0.018	0.15	0.25	0.0031	0.071	0.68	0.36
rs8024033	15:40403657	*BMF*	0.5	CLL	0.083	0.83	0.029	0.45	0.011	0.068	0.4
rs11636802	15:56775597	*MNS1, RFXDC2*	0.11	CLL	0.32	0.79	0.65	0.37	0.36	0.8	0.84
rs72742684	15:56780767	*MNS1, RFX7*	0.11	CLL	0.35	0.89	0.6	0.34	0.35	0.92	0.7
rs2052702	15:69989505	*PCAT29*	0.38	CLL	0.85	0.98	0.75	0.96	0.7	0.46	0.47
rs7176508	15:70018990	*RPLP1*	0.38	CLL	0.93	0.86	0.62	0.89	0.54	0.42	0.37
rs12448368	16:81044947	*CENPN, ATMIN*	0.13	mLOY	0.034	0.26	0.24	0.34	0.075	0.37	0.24
rs391023	16:85927814	*IRF8*	0.36	CLL	0.077	0.37	0.0067	0.31	0.064	0.84	0.012
rs391855	16:85928621	*IRF8*	0.42	CLL	0.0099	0.18	0.0013	0.37	0.015	0.85	0.016
rs391525	16:85944439	*IRF8*	0.34	CLL	0.025	0.045	0.0073	0.92	0.023	0.076	0.24
rs1044873	16:85955671	*IRF8*	0.39	CLL	0.034	0.13	0.0055	0.97	0.024	0.15	0.4
rs78378222	17:7571752	*TP53*	0.013	mLOY	0.037	$\mathbf{3.2\times10^{-5}}$	0.99	0.29	0.42	0.0044	0.0059
rs77522818	17:47817373	*FAM117A*	0.043	mLOY	0.011	0.077	0.08	0.53	0.013	0.091	0.36
rs11082396	18:42080720	*SETBP1*	0.13	mLOY	0.22	0.37	0.5	0.42	0.44	0.99	0.78
rs4368253	18:57622287	*PMAIP1*	0.32	CLL	0.59	0.87	0.89	0.086	0.54	0.55	0.83
rs4987856	18:60793494	*BCL2*	0.097	CLL	0.25	0.49	0.083	0.29	0.19	0.15	0.44
rs4987855	18:60793549	*BCL2*	0.097	CLL	0.34	0.52	0.14	0.37	0.28	0.14	0.44
rs4987852	18:60793921	*BCL2*	0.07	CLL	0.85	0.99	0.7	0.68	0.8	0.91	0.4
rs17758695	18:60920854	*BCL2*	0.03	mLOY	0.61	0.2	0.45	0.036	0.83	0.32	0.23
rs8105767	19:22215441	*ZNF208*	0.29	telo	0.62	0.98	0.18	0.12	0.22	0.72	0.81
rs11083846	19:47207654	*PRKD2, STRN4*	0.23	CLL	0.088	0.36	0.025	0.51	0.14	0.4	0.36
rs60084722	20:30355738	*TPX2, BCL2L1, HM13*	0.21	mLOY	0.018	0.0051	0.049	0.77	0.17	0.51	0.16
rs755017	20:62421622	*RTEL1*	0.13	telo	0.0047	0.0064	0.16	0.61	0.023	0.15	0.14
rs555607708	22:29091856	*CHEK2*	0.0019	MPN	0.0038	0.01	0.00012	0.3	$\mathbf{7.7\times10^{-5}}$	$\mathbf{1.8\times10^{-6}}$	0.76

See next page for extended caption.

Extended caption for Supplementary Table. 11. We examined SNPs previously associated with chronic lymphocytic leukemia (CLL) [47, 83–87], myeloproliferative neoplasms (MPN) [15–17, 20, 88], loss of chromosome Y [19, 21], clonal hematopoiesis (CH) [11], and telomere length [89] for association with classes of mCAs, hypothesizing that similar mechanisms could be perturbed. Of the 88 unique SNPs collectively reported for these traits, 86 were imputed in the N=150K UK Biobank data; we report associations (Fisher's exact test) of these SNPs with:

- Mosaic status for mCAs on any chromosome (P_{any})
- Mosaic status for loss events (P_{loss})
- Mosaic status for CNN-LOH events ($P_{\text{CNN-LOH}}$)
- Mosaic status for gain events (P_{gain})
- Mosaic status for mCAs on any autosome (P_{auto})
- Mosaic status for loss events on any autosome ($P_{\text{auto loss}}$)
- Mosaic status for female loss of chrX ($P_{\text{X loss}}$).

We stratified events by autosome/chrX in the manner above because nearly all female chrX events are losses (Fig. 1).

Four SNPs reach Bonferroni significance ($P<8.3\times10^{-5}$ for 86 SNPs $\times$ 7 phenotypes):

- rs34002450 (chr5:1280940), a common intronic deletion in *TERT* previously associated with clonal hematopoiesis [11]. This SNP is most strongly associated with autosomal events (P=7.8×10^{-5}).

- rs7705526 (chr5:1285974), a common SNP in *TERT* previously associated with somatic *JAK2* V617F mutation [20] and in strong LD with rs2736100, previously associated with telomere length [89]. This SNP is also in LD with rs34002450 (European R^2=0.53 computed using LDlink [90]) and is most strongly associated with autosomal events (P=4.8×10^{-5}). The alleles of these SNPs that were previously associated with longer telomeres are the risk alleles for mosaic status.

- rs78378222 (chr17:7571752), a low-frequency 3' UTR SNP in *TP53* previously associated with mosaic loss of Y [21]. This SNP is most strongly associated with loss events (P=3.2×10^{-5}).

- rs555607708 (chr22:29091856), a rare frameshift SNP in *CHEK2* previously associated with somatic *JAK2* V617F mutation [20]. This SNP is most strongly associated with autosomal loss events (P=1.8×10^{-6}).

Supplementary Table 12. Risk increase for incident cancers conferred by mCAs.

(a) Analyses restricted to n=36 cases and 113,923 controls with normal blood counts at assessment

	CLL		MPN		Any blood cancer	
mCA	*P*	OR (95% CI)	*P*	OR (95% CI)	*P*	OR (95% CI)
1p=	1	0 (0–85.8)	0.025	41.2 (0.99–260)	0.32	2.64 (0.07–15.2)
1q=	1	0 (0–123)	1	0 (0–204)	0.25	3.52 (0.09–20.4)
2p–	1	0 (0–415)	1	0 (0–786)	1	0 (0–48.5)
3+	1.7×10^{-5}	421 (42–2.05e+03)	1	0 (0–1.39e+03)	0.0016	39.2 (4.19–180)
3q=	1	0 (0–817)	1	0 (0–1.76e+03)	1	0 (0–93.8)
4q–	1	0 (0–292)	1	0 (0–608)	0.11	9.39 (0.23–58.6)
4q=	1	0 (0–344)	1	0 (0–656)	0.0063	18.3 (2.07–75)
5q–	1	0 (0–366)	1	0 (0–701)	0.095	10.5 (0.25–66.3)
5q=	1	0 (0–491)	1	0 (0–814)	1	0 (0–59.2)
6p=	1	0 (0–373)	1	0 (0–548)	1	0 (0–34.5)
7q–	1	0 (0–331)	1	0 (0–571)	1	0 (0–34.6)
8+	0.0072	155 (3.52–1.11e+03)	1	0 (0–1.27e+03)	2.4×10^{-5}	68.3 (11.9–272)
8q=	1	0 (0–534)	1	0 (0–848)	1	0 (0–56.1)
9+	1	0 (0–740)	1	0 (0–1.38e+03)	1	0 (0–75.3)
9p=	1	0 (0–201)	1.6×10^{-10}	609 (144–1.91e+03)	8.3×10^{-6}	36.7 (9.16–108)
9q=	1	0 (0–153)	1	0 (0–270)	1	0 (0–16.4)
10q–	1	0 (0–397)	1	0 (0–674)	1	0 (0–45.1)
11q–	1	0 (0–324)	1	0 (0–495)	0.11	8.55 (0.21–52.5)
11p=	1	0 (0–113)	1	0 (0–182)	1	0 (0–11.6)
11q=	1	0 (0–132)	0.018	58 (1.37–376)	0.0025	11.8 (2.35–36.9)
12+	2.2×10^{-10}	191 (55.1–527)	1	0 (0–234)	1.9×10^{-8}	27.6 (10.5–61.8)
12q=	1	0 (0–270)	1	0 (0–538)	1	0 (0–32)
13q–	0.016	66 (1.57–419)	1	0 (0–257)	0.12	8.19 (0.2–50.3)
13q=	0.00024	97.2 (11–404)	1	0 (0–282)	0.18	5.15 (0.13–30.4)
14+	1	0 (0–302)	1	0 (0–520)	1	0 (0–34.6)
14q–	0.006	187 (4.21–1.42e+03)	1	0 (0–1.07e+03)	0.049	22 (0.5–154)
14q=	1	0 (0–79.1)	0.03	34.3 (0.83–215)	0.071	4.7 (0.56–17.6)
15+	1	0 (0–151)	1	0 (0–308)	0.21	4.33 (0.11–26)
15q=	1	0 (0–126)	1	0 (0–220)	0.25	3.48 (0.09–20.3)
16p=	1	0 (0–198)	1	0 (0–336)	0.00083	17.7 (3.48–56.4)
16q=	1	0 (0–250)	1	0 (0–415)	1	0 (0–25.2)
17+	1	0 (0–504)	1	0 (0–754)	0.071	14.6 (0.34–95.1)
17p–	1	0 (0–412)	1	0 (0–588)	0.087	11.7 (0.28–74.4)
17q=	1	0 (0–194)	1	0 (0–295)	1	0 (0–19.6)
18+	0.013	85.1 (1.96–590)	1	0 (0–529)	0.00023	29.1 (5.46–99.9)
19p=	1	0 (0–374)	1	0 (0–690)	1	0 (0–41)
19q=	1	0 (0–278)	1	0 (0–585)	1	0 (0–32.4)
20q–	1	0 (0–107)	1	0 (0–215)	0.0003	13.6 (3.55–37)
20q=	1	0 (0–460)	1	0 (0–660)	1	0 (0–47.4)
21+	1	0 (0–483)	1	0 (0–909)	0.077	13.4 (0.32–87)
21q=	1	0 (0–540)	1	0 (0–935)	1	0 (0–57)
22+	1	0 (0–239)	1	0 (0–452)	1	0 (0–26.7)
22q–	1	0 (0–1.23e+03)	1	0 (0–947)	1	0 (0–126)
22q=	1	0 (0–182)	1	0 (0–308)	1	0 (0–20.3)
–X	1	0 (0–8.14)	1	0 (0–20.3)	0.44	0.5 (0.06–1.85)

This table provides numerical data plotted in Fig. 5a. Events were grouped by chromosome and copy number, with loss and CNN-LOH events subdivided by p-arm vs. q-arm; events observed in $\geq$30 individuals were tested for association with incident CLL, MPN, and any blood cancer (diagnosed $>$1 year after DNA collection in individuals with no previous cancer diagnosis).

(b) Analyses of n=78 cases and 118,481 controls with no restrictions on blood counts at assessment

	CLL		MPN		Any blood cancer	
mCA	*P*	OR (95% CI)	*P*	OR (95% CI)	*P*	OR (95% CI)
1p=	1	0 (0–40)	0.046	22.1 (0.54–133)	0.4	1.96 (0.05–11.3)
1q=	1	0 (0–51.9)	1	0 (0–110)	0.34	2.44 (0.06–14.1)
2p–	0.027	38.1 (0.91–241)	1	0 (0–436)	0.13	7.55 (0.18–46.6)
3+	7.8×10^{-5}	190 (19.6–936)	1	0 (0–749)	8.5×10^{-5}	43.2 (7.76–161)
3q=	1	0 (0–423)	1	0 (0–780)	1	0 (0–74.3)
4q–	1	0 (0–133)	1	0 (0–316)	0.15	6.34 (0.15–38.8)
4q=	1	0 (0–159)	1	0 (0–328)	0.011	13.4 (1.53–54.7)
5q–	1	0 (0–167)	0.011	93.4 (2.21–614)	0.0082	16 (1.81–65.8)
5q=	1	0 (0–230)	1	0 (0–417)	1	0 (0–40.9)
6p=	1	0 (0–165)	1	0 (0–286)	1	0 (0–26.5)
7q–	1	0 (0–137)	1	0 (0–323)	0.15	6.25 (0.15–38.5)
8+	0.018	60.8 (1.41–410)	1	0 (0–606)	6.8×10^{-8}	62.6 (17.5–186)
8q=	1	0 (0–257)	1	0 (0–460)	1	0 (0–44.9)
9+	1	0 (0–324)	1	0 (0–665)	1	0 (0–54.3)
9p=	1	0 (0–89.4)	1.6×10^{-21}	560 (225–1.26e+03)	1.1×10^{-11}	39.5 (16.8–83.1)
9q=	1	0 (0–69.3)	1	0 (0–155)	1	0 (0–12.9)
10q–	1	0 (0–205)	1	0 (0–310)	1	0 (0–34.7)
11q–	0.0006	61.2 (6.93–251)	1	0 (0–271)	0.00099	16.9 (3.29–54.8)
11p=	1	0 (0–52.5)	1	0 (0–96.5)	1	0 (0–8.84)
11q=	1	0 (0–53.6)	0.032	32.6 (0.79–202)	0.0076	7.88 (1.57–24.3)
12+	1.2×10^{-20}	173 (78.1–355)	1	0 (0–131)	2×10^{-15}	33.9 (17–62.7)
12q=	1	0 (0–126)	1	0 (0–296)	1	0 (0–24.2)
13q–	3.4×10^{-19}	185 (80.2–392)	1	0 (0–134)	1.1×10^{-11}	29.5 (13.3–58.9)
13q=	3.3×10^{-7}	81.5 (20.7–233)	1	0 (0–149)	0.00026	14 (3.67–38.4)
14+	1	0 (0–118)	1	0 (0–291)	1	0 (0–22.7)
14q–	0.00017	123 (13.3–540)	1	0 (0–488)	0.00023	29.4 (5.48–102)
14q=	1	0 (0–34.7)	0.0014	38.4 (4.45–151)	0.0035	6.74 (1.8–17.9)
15+	1	0 (0–65.7)	1	0 (0–160)	0.28	3.13 (0.08–18.6)
15q=	1	0 (0–57)	1	0 (0–116)	0.32	2.65 (0.07–15.4)
16p=	1	0 (0–84.4)	1	0 (0–190)	0.0022	12.4 (2.45–39.1)
16q=	1	0 (0–112)	1	0 (0–228)	1	0 (0–19.6)
17+	1	0 (0–181)	1	0 (0–487)	0.11	9.2 (0.22–58.1)
17p–	1	0 (0–140)	1	0 (0–389)	0.01	14.1 (1.61–57.3)
17q=	1	0 (0–83)	1	0 (0–169)	1	0 (0–14.4)
18+	0.031	33.6 (0.8–214)	1	0 (0–306)	0.00075	19 (3.63–63.5)
19p=	1	0 (0–159)	1	0 (0–419)	1	0 (0–30.2)
19q=	1	0 (0–133)	1	0 (0–314)	1	0 (0–24.9)
20q–	1	0 (0–47.3)	1	0 (0–108)	0.0013	9.1 (2.4–24.6)
20q=	1	0 (0–187)	1	0 (0–360)	1	0 (0–34.1)
21+	1	0 (0–225)	1	0 (0–437)	0.1	9.59 (0.23–61.3)
21q=	1	0 (0–236)	1	0 (0–462)	1	0 (0–41.9)
22+	0.042	24.4 (0.59–151)	1	0 (0–218)	0.2	4.5 (0.11–26.9)
22q–	1.2×10^{-8}	207 (49–654)	1	0 (0–494)	8.7×10^{-6}	37.4 (9.1–115)
22q=	1	0 (0–80.7)	1	0 (0–172)	1	0 (0–14.6)
–X	1	0.82 (0.02–4.99)	1	0 (0–13)	0.38	0.54 (0.11–1.63)

This table provides results of analogous analyses removing the restrictions we imposed on blood counts in our primary analyses (lymphocyte count 1–3.5×10^9/L, red cell count $<6.1\times10^{12}$/L for males and $<5.4\times10^{12}$/L for females, platelet count $<450\times10^9$/L, RBC distribution width $<15\%$).

Supplementary Table 13. Risk increase for mortality during ∼7-year follow-up conferred by mCAs.

(a) All-cause mortality risk increase conferred by mCAs

mCA type	Cancer status at assessment	*P*	HR (95% CI)
Loss	No previous Dx	1.3×10^{-7}	2.08 (1.58–2.73)
Loss	Previous Dx	5.4×10^{-10}	2.76 (2.00–3.80)
CNN-LOH	No previous Dx	0.01	1.36 (1.07–1.71)
CNN-LOH	Previous Dx	6.2×10^{-5}	1.81 (1.35–2.42)
Gain	No previous Dx	0.00021	1.92 (1.36–2.70)
Gain	Previous Dx	0.0055	1.97 (1.22–3.19)

(b) Non-cancer mortality risk increase conferred by mCAs

mCA type	Cancer status at assessment	*P*	HR (95% CI)
Loss	No previous Dx	0.0017	1.93 (1.28–2.92)
Loss	Previous Dx	0.00015	3.22 (1.76–5.89)
CNN-LOH	No previous Dx	0.19	1.26 (0.89–1.79)
CNN-LOH	Previous Dx	0.04	1.84 (1.03–3.28)
Gain	No previous Dx	0.096	1.59 (0.92–2.75)
Gain	Previous Dx	0.31	1.67 (0.62–4.50)

The first table provides numerical data plotted in Fig. 5d (from analyses of n=128,854 individuals without previous cancer diagnoses and n=15,782 with prevalent cancer), and the second provides analogous results excluding 2,687 of 4,619 deaths reported to be due to cancer.

Supplementary Table 14. Comparison of age and sex of mosaic individuals across studies.

(a) This study

Copy change	*N* (unique)	Mean age (s.e.m.)	Fraction male (s.e.)
Loss	941	60.3 (0.2)	0.542 (0.016)
CNN-LOH	2208	58.8 (0.2)	0.490 (0.011)
Gain	578	61.5 (0.3)	0.587 (0.021)

(b) Jacobs et al. [1]

Copy change	*N* (unique)	Mean age (s.e.m.)	Fraction male (s.e.)
Loss	186	68.2 (0.6)	0.790 (0.030)
CNN-LOH	278	68.0 (0.6)	0.665 (0.028)
Gain	87	66.9 (1.1)	0.793 (0.044)

(c) Laurie et al. [2]

Copy change	*N* (unique)	Mean age (s.e.m.)	Fraction male (s.e.)
Loss	192	66.0 (1.0)	0.776 (0.030)
CNN-LOH	150	61.4 (1.6)	0.693 (0.038)
Gain	65	56.8 (3.1)	0.692 (0.058)

These tables compare age and sex among unique carriers of loss, CNN-LOH, and gain events in the current study, Jacobs et al. [1], and Laurie et al. [2]. We included all individuals for which both age and sex information were available.

Supplementary Table 15. Comparison of mosaic event detection rates across studies.

Study	*N*	Events detected	Mosaic individuals	Mosaic Rate	Number of each event type			
					Loss	CNN-LOH	Gain	Undetermined
This study (autosomal events)	151,202	5562	4889	3.23%	1118	2237	659	1548
Jacobs et al. (2012)	57,853	681	517	0.89%	245	331	105	0
Laurie et al. (2012)	50,222	514	404	0.80%	259	175	80	0
Machiela et al. (2015) (TGSII)	24,849	341	168	0.68%	90	163	69	19
Vattathil & Scheet (2016)	31,100	1141	901	2.90%	202	30	70	839

Here we compare the number of autosomal events we identified to previous studies of mCAs using SNP genotyping arrays. We note that different studies have multiple differences that impact event detection, including (i) age distributions, (ii) cancer case/control distributions, (iii) genotyping platforms (previous studies used Illumina arrays), and (iv) analysis methods (only our study and Vattathil & Scheet [8] used haplotype phase). The first two differences affect mosaicism rate, while the last two affect detection sensitivity.

Supplementary Table 16. Families with high-confidence non-reference variable number tandem repeat (VNTR) motifs at *FRA10B*.

Family	Individual	Relationship	*FRA10B* reads	Primary motif	del(10q)?
11336	02805	father	562	VNTR-42-a	detected
11336	02806	daughter	76	VNTR-42-a	detected
11542	00649	daughter	6	NA	
11542	00656	father	15	VNTR-42-a	
12212	04392	son	148	VNTR-42-b	
12212	04401	mother	65	VNTR-42-b	
12212	04410	father	74	VNTR-42-b	
12651	06665	mother	162	VNTR-42-d	
12759	06402	father	101	VNTR-42-e	
13316	07467	son	55	VNTR-38-a	
13316	07483	father	8	NA	
13383	07471	son	15	NA	
13383	07489	father	15	NA	
13383	07490	daughter	31	VNTR-43-d	
13564	08141	son	16	VNTR-39-a	
13564	08142	father	26	VNTR-39-a	
13564	08145	son	52	VNTR-39-a	
13738	08952	son	908	VNTR-43-a	
13738	08958	mother	1057	VNTR-43-a	
13777	07980	father	1160	VNTR-42-a	detected
13777	07981	son	881	VNTR-42-a	detected
13892	09326	son	19	NA	
13892	09330	mother	14	NA	
13892	09339	son	16	VNTR-43-e	
14154	10708	son	391	VNTR-43-b	
14154	10712	mother	371	VNTR-43-b	
14154	10718	daughter	346	VNTR-43-b	
14415	12037	son	49	VNTR-42-c	
14415	12046	son	30	VNTR-42-c	
14415	12047	father	18	VNTR-42-c	
14574	12604	mother	10	NA	
14574	12609	son	22	VNTR-43-c	
14574	12610	daughter	34	VNTR-43-c	

We identified 12 distinct high-confidence primary VNTR motifs in 26 individuals from 14 families (Supplementary Note 8). We also list 7 additional family members sharing haplotypes containing non-reference VNTR motifs; the primary motif for these individuals is listed as NA (not assembled), but most of these individuals have read support for the VNTR motif on the shared haplotype (Fig. S8.2-1). The column "*FRA10B* reads" indicates the number of reads mapping to the target region 10:113002151–113002300 in hg19. Mosaic loss of 10q25.2–10qter was detectable in four individuals (Fig. 3).