Monogenic and polygenic inheritance become instruments for clonal selection

https://doi.org/10.1038/s41586-020-2430-6

Received: 1 June 2019

Accepted: 23 April 2020

Published online: 24 June 2020

Check for updates

Po-Ru Loh[1,2]✉, Giulio Genovese[2,3,4]✉ & Steven A. McCarroll[2,3,4]✉

Clonally expanded blood cells that contain somatic mutations (clonal haematopoiesis) are commonly acquired with age and increase the risk of blood cancer[1–9]. The blood clones identified so far contain diverse large-scale mosaic chromosomal alterations (deletions, duplications and copy-neutral loss of heterozygosity (CN-LOH)) on all chromosomes[1,2,5,6,9], but the sources of selective advantage that drive the expansion of most clones remain unknown. Here, to identify genes, mutations and biological processes that give selective advantage to mutant clones, we analysed genotyping data from the blood-derived DNA of 482,789 participants from the UK Biobank[10]. We identified 19,632 autosomal mosaic chromosomal alterations and analysed these for relationships to inherited genetic variation. We found 52 inherited, rare, large-effect coding or splice variants in 7 genes that were associated with greatly increased vulnerability to clonal haematopoiesis with specific acquired CN-LOH mutations. Acquired mutations systematically replaced the inherited risk alleles (at *MPL*) or duplicated them to the homologous chromosome (at *FH*, *NBN*, *MRE11*, *ATM*, *SH2B3* and *TM2D3*). Three of the genes (*MRE11*, *NBN* and *ATM*) encode components of the MRN–ATM pathway, which limits cell division after DNA damage and telomere attrition[11–13]; another two (*MPL* and *SH2B3*) encode proteins that regulate the self-renewal of stem cells[14–16]. In addition, we found that CN-LOH mutations across the genome tended to cause chromosomal segments with alleles that promote the expansion of haematopoietic cells to replace their homologous (allelic) counterparts, increasing polygenic drive for blood-cell proliferation traits. Readily acquired mutations that replace chromosomal segments with their homologous counterparts seem to interact with pervasive inherited variation to create a challenge for lifelong cytopoiesis.

We identified 17,111 cases of clonal haematopoiesis involving 19,632 autosomal mosaic chromosomal alterations (mCAs) (Extended Data Figs. 1, 2, Supplementary Tables 1–3) by analysing single-nucleotide polymorphism (SNP)-array intensity data from 482,789 participants from the UK Biobank aged between 40 and 70 years[10]. To identify these cases, we applied a method previously developed to analyse the UK Biobank interim release (approximately 31% of the current cohort)[9]; our approach finds imbalances in the abundances of homologous chromosomal segments by combining allele-specific intensity data with long-range chromosomal phase information[17,18] (Methods, Supplementary Note 1). We classified 73% of the detected mCAs as loss (3,718 events), gain (2,389 events) or CN-LOH (8,185 events)—that is, the replacement of one chromosomal segment by its homologous (allelic) counterpart (Supplementary Table 1). (Another 5,340 mCAs could not be confidently classified because the power to detect imbalances exceeded the power to distinguish copy-neutral from copy-number-altering mCAs[9] (Extended Data Fig. 2a–d).) Of the 19,632 detected mCAs, 12,683 were present at cell fractions from 0.7% to 5%, and 6,949 were present at cell fractions greater than 5%. Consistent with previous work[1,2,5,6,9], mCAs on different chromosomes exhibited different recurrence rates and size distributions (Extended Data Fig. 1a) and a range of tendencies to be more common in one sex (usually males, although with clear exceptions) and the elderly (Extended Data Fig. 1b, Supplementary Table 4). Clones also tended to be found in individuals with anomalous counts for one or more blood-cell types (Extended Data Fig. 1c, Supplementary Table 5).

Monogenic inherited forms of clonal haematopoiesis

We next sought to identify specific genes and variants that might propel clonal selection. Three loci (*MPL*, *ATM* and *TM2D3–TARS3*) have recently been identified at which inherited rare variants increase the risk of developing clones with acquired CN-LOH mutations that affect the rare inherited risk allele in a predictable way[9]. To detect loci targeted by CN-LOH mutations in this manner, and to identify probable causal inherited variants at these loci, we searched the genome for associations

[1] Division of Genetics, Department of Medicine, Brigham and Women's Hospital, 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. ✉e-mail: poruloh@broadinstitute.org; giulio.genovese@gmail.com; mccarroll@hms.harvard.edu

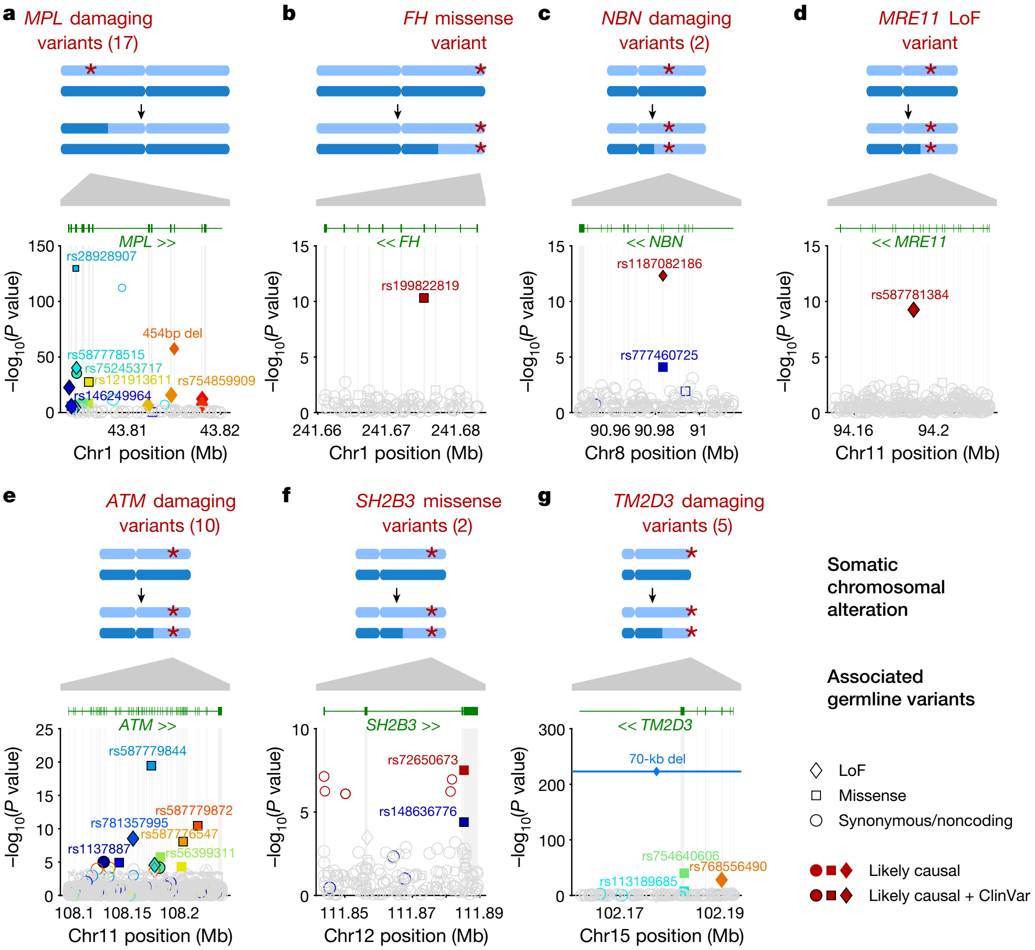


Fig. 1 | Fine-mapped inherited sequence alleles associated with the acquisition/selection of CN-LOH mutations in *cis*. a, *MPL*; **b**, *FH*; **c**, *NBN*; **d**, *MRE11*; **e**, *ATM*; **f**, *SH2B3*; **g**, *TM2D3*. At each locus, the CN-LOH mutations acquired by expanded clones tend to have deleted (**a**) or duplicated (**b**–**g**) the inherited alleles in a predictable manner as shown. Each panel is organized as follows: top, genomic modifications observed in clones; bottom, association *P* values (two-sided Fisher's exact test on $n \geq 378{,}307$ individuals) (Methods) versus chromosomal position. All variants with filled symbols are probable causal coding or splice variants (Extended Data Table 1); black marker edges indicate evidence of pathogenicity in ClinVar[21]. Distinct colours are used to indicate the statistical independence of variants; any variants in linkage disequilibrium with likely causal variants ($R^2 > 0.2$ in cases) are indicated by open symbols with a border colour that matches that of the probable causal variant. Symbol shapes indicate the effects of the indicated variant on encoded protein (such as LoF or missense); symbol sizes scale inversely with minor allele frequency.

between inherited variants and CN-LOH mutations acquired in *cis*. To avoid potential confounding from population stratification, we restricted these analyses to 455,009 individuals who reported European ancestry (Extended Data Fig. 3, Supplementary Note 2).

Inherited rare variants at seven loci (*MPL*, *ATM*, *TM2D3*, *FH*, *NBN*, *MRE11* and *SH2B3*) associated at genome-wide significance with the development of blood clones in which an acquired CN-LOH mutation had affected the inherited risk allele in a consistent way (Fig. 1, Extended Data Table 1, Extended Data Fig. 4, Supplementary Table 6). At six loci (all loci other than *MPL*), the inherited rare alleles were consistently made homozygous by somatic CN-LOH mutations (149 out of 153 cases; binomial $P = 3.9 \times 10^{-39}$). Associations at all seven loci appeared to be driven by rare coding variants with large effect sizes (odds ratios 11–555; 95% confidence intervals 5.8–724): the lead associated variants at six of the seven loci were coding mutations, and the lead variant at the remaining locus, *MRE11* (rs762019591; Fisher's exact $P = 3.0 \times 10^{-11}$), was in strong linkage disequilibrium with a nonsense SNP in *MRE11* (rs587781384) (Extended Data Table 1).

The functions of five of the seven implicated genes converged on two probable mechanisms of clonal advantage. Three implicated genes (*MRE11*, *NBN* and *ATM*) encode proteins that act together to limit cell growth after DNA damage and telomere attrition[13]. Specifically, *MRE11* and *NBN* encode two of the three proteins of the MRN complex, which recognizes double-strand breaks and activates the checkpoint kinase encoded by *ATM*[11,12]. Thus, strong-effect inherited variants (including protein-truncating variants) at *MRE11*, *NBN* and *ATM*—made homozygous by CN-LOH (Fig. 1c–e)—seem to disrupt a key pathway that limits proliferation in cells that have experienced DNA damage or telomere shortening.

Two other implicated genes encode proteins that regulate the self-renewal of haematopoietic stem cells: *MPL*, which encodes the myeloproliferative leukaemia protein that positively regulates the self-renewal of stem cells[14,15] in addition to its roles in thrombocytopoiesis; and *SH2B3*, which encodes a signalling protein (LNK) that negatively regulates haematopoietic signalling through MPL[16]. Clonally selected CN-LOH mutations appeared to have opposite effects on rare inherited (putative function-reducing) variants in *MPL* and *SH2B3*: the acquired 1p CN-LOH mutations eliminated rare inherited variants (including protein-truncating variants) in *MPL* (Fig. 1a), whereas the acquired 12q CN-LOH mutations duplicated *SH2B3* variants to the other homologue (Fig. 1f). The primary *SH2B3* risk allele (rs72650673:A) has previously been found to increase platelet counts in carriers[19], which suggests roles at at least two levels of haematopoietic differentiation.

Inherited mutations in *FH*, which encodes the fumarate hydratase protein that functions in the Krebs cycle, are an established cause of benign and malignant neoplasms in several tissues[20]. The molecular function(s) of *TM2D3* are unknown.

To identify other variants at all seven loci for which CN-LOH mutations led to subsequent clonal proliferation, we comprehensively examined coding and splice variants in these genes by integrating information from SNP arrays, imputation and exome sequencing (Methods, Extended Data Fig. 5, Supplementary Notes 3, 4). Among 616 missense, predicted loss-of-function (LoF) or probable pathogenic[21] variants tested (Methods), 52 variants associated independently with CN-LOH mosaicism in *cis* (at a false discovery rate (FDR) < 0.05 significance per locus; odds ratios 11–758, 95% confidence intervals 4–2,618), including several variants in *MPL* (28 variants), *ATM* (13), *TM2D3* (5), *NBN* (2) and *SH2B3* (2); 38 of the 52 individual variants reached Bonferroni significance (Fisher's exact $P < 8.1 \times 10^{-5}$ for 616 variants tested) (Fig. 1, Extended Data Table 1, Supplementary Tables 7, 8). All 52 variants were rare (population allele frequency < 0.2%). Notably, 23 of the 52 variants had been reported as clinically significant[21] in hereditary blood disorders (eight variants in *MPL* and one in *SH2B3*) or cancer (eleven variants in *ATM* and one each in *MRE11*, *NBN* and *FH*). All 28 *MPL* variants were removed from the genomes of expanded clones by CN-LOH mutations (244 out of 244 cases, binomial $P = 7.1 \times 10^{-74}$), consistent with a model in which the inherited alleles (with reduced MPL function) have a hypo-proliferative effect that is rescued by CN-LOH[9]. The 24 inherited variants at the other six loci were systematically made homozygous by CN-LOH (233 out of 239 cases, binomial $P = 5.6 \times 10^{-61}$), consistent with pro-proliferative effects of reduced ATM, MRE11, NBN, SH2B3, TM2D3 and FH function (Fig. 1, Extended Data Table 1, Supplementary Table 7). Sharing of long haplotypes among individuals with 1p CN-LOH mutations spanning *MPL* and among individuals with 11q CN-LOH mutations spanning *ATM* indicated that, although the identified risk variants (Extended Data Table 1, Supplementary Table 7) are likely to be the primary drivers of heritable clonal haematopoiesis risk at these loci, the full allelic series probably includes many more risk variants (Extended Data Figs. 6, 7).

To detect additional potential risk variants and estimate the fraction of CN-LOH clones attributable to inherited protein-altering variants (including still-rarer variants) at each locus, we examined exome-sequence data available for 49,960 of the UK Biobank participants[22]. Among 271 exome-sequenced individuals of European ancestry with unexplained mosaic CN-LOH events spanning the 7 loci (that is, not carrying any of the 52 variants already identified), 22 individuals carried 21 distinct ultra-rare coding or splice variants that altered the encoded proteins (versus 1.28 individuals expected by chance, $P = 2.8 \times 10^{-20}$; 'ultra-rare' refers to population allele frequency less than 0.0001) (Methods, Supplementary Tables 9–11). Collectively, *MPL* variants identified by these association and burden analyses were present in 39 out of 71 exome-sequenced individuals with 1p CN-LOH events spanning *MPL* (versus 0.5 expected), which suggests that approximately 54% of acquired 1p CN-LOH events are driven by inherited coding or splice variants at *MPL* (Supplementary Table 11). Similarly, inherited variants at *ATM*, *NBN*, *SH2B3* and *TM2D3* appeared to drive approximately 17–33% of CN-LOH events spanning these loci (Supplementary Table 11).

We estimate that about 5% of clones with CN-LOH arose from monogenic inherited vulnerabilities at one of these seven loci.

Common inherited variants at five loci conferred more modest risk of mCA (odds ratio of 1.07–1.24). Common variants at *TCL1A* and *DLK1* on 14q associated with acquired 14q CN-LOH mutations (Supplementary Table 12, Supplementary Note 5), whereas common variants at *TERC*, *SP140* and the previously implicated *TERT* locus[7] broadly increased the risk of clonal haematopoiesis involving any autosomal mCA (Supplementary Table 13, Supplementary Note 6). Notably, *TERC* and *TERT* both encode proteins with key roles in the maintenance and elongation of telomeres (Supplementary Table 14).

Some CN-LOH events provided 'second hits' to acquired point mutations. At the frequently mutated *DNMT3A*, *TET2* and *JAK2* loci[3,4], approximately 24–60% of CN-LOH mutations appeared to provide second hits to somatic point mutations detectable from exome-sequencing reads (Extended Data Fig. 8, Supplementary Table 11); additional CN-LOH events spanning these loci might be explained by point mutations present at lower cell fractions that we could not detect among 10–40 sequencing reads per haplotype (Methods). Among 33 exome-sequenced individuals with 9p CN-LOH events, 20 individuals had at least one read that suggested the *JAK2*^{V617F} mutation; conversely, 18 out of 46 individuals with *JAK2*^{V617F} calls had a detectable mCA on 9p (15 CN-LOH events and three chromosome 9 duplications). Together, the putative second-hit clones at these loci accounted for about 0.3% of all detected CN-LOH clones.

Clonal CN-LOH mutations increase polygenic drive

The majority of the 17,111 haematopoietic clones we observed in UK Biobank still had unknown causes; most clones had CN-LOH mutations, which were numerous on every chromosome arm (Extended Data Fig. 1a). This posed the question of what genetic effect propels detected clones in a manner that is so distributed across the genome. The broad distribution of acquired mCA events reminded us of recent work in human and agricultural genetics that has revealed that many phenotypes are shaped by polygenic effects from alleles of modest effect at hundreds of loci spread across all chromosomes[23–25].

We hypothesized that inherited haplotypes along a chromosome arm can themselves be instruments for clonal selection (Fig. 2a). To evaluate this possibility, we tested whether the haplotypes duplicated and deleted by probable CN-LOH mutations (Methods) tended to differ systematically in polygenic drive for blood-cell abundance phenotypes, as estimated from combinations of many inherited alleles and the relationships of these alleles to blood-cell abundances in the general population. We evaluated this by building polygenic statistical models[26] for blood-cell abundance traits (using data on blood-cell counts from UK Biobank participants) and for clonal loss of the Y chromosome, a frequent marker of haematopoietic clones[27]. On the basis of these models, we estimated haematopoietic polygenic risk scores (HPRS) for the combinations of common alleles along the haplotypes gained and lost by CN-LOH mutations in expanded clones (Methods).

CN-LOH mutations in expanded clones tended to have caused chromosomal segments with higher HPRS values to replace homologous (allelic) counterparts with lower HPRS values (Fig. 2b). Averaging across all autosomal CN-LOH events, the allelic substitutions produced by CN-LOH mutations significantly increased polygenic scores for clonality with the loss of the Y chromosome ($P = 1.2 \times 10^{-13}$; $P = 4.3 \times 10^{-8}$ and $P = 5.2 \times 10^{-7}$ for CN-LOH in men and women separately, respectively) and also tended to increase polygenic scores for the individual blood-cell abundance traits (most significant: neutrophil counts, $P = 7.5 \times 10^{-6}$; eosinophil counts, $P = 1.4 \times 10^{-4}$). This effect was observed throughout the genome: 14 distinct combinations of chromosome arms and cell-abundance traits exhibited significant upward shifts in HPRS values (at an FDR of 0.05), and 209 out of all 312 combinations exhibited a positive mean increase ($P = 2.0 \times 10^{-9}$, sign test) (Fig. 2b, Supplementary Table 15). These effects were specific to polygenic scores for blood-cell abundance traits: CN-LOH mutations did not tend to affect polygenic scores for control traits such as height and BMI (Supplementary Table 16), and results were mixed for blood-cell morphology traits (Extended Data Fig. 9). CN-LOH mutations also seemed to act on risk alleles for myeloproliferative neoplasms (Supplementary Table 17, Supplementary Note 7).

These results raised the possibility that the direction of mosaic CN-LOH mutations—that is, which haplotype has been made homozygous in a clone that rises to detectable frequency—can be predicted from inherited variation. To test this idea, we performed cross-validated

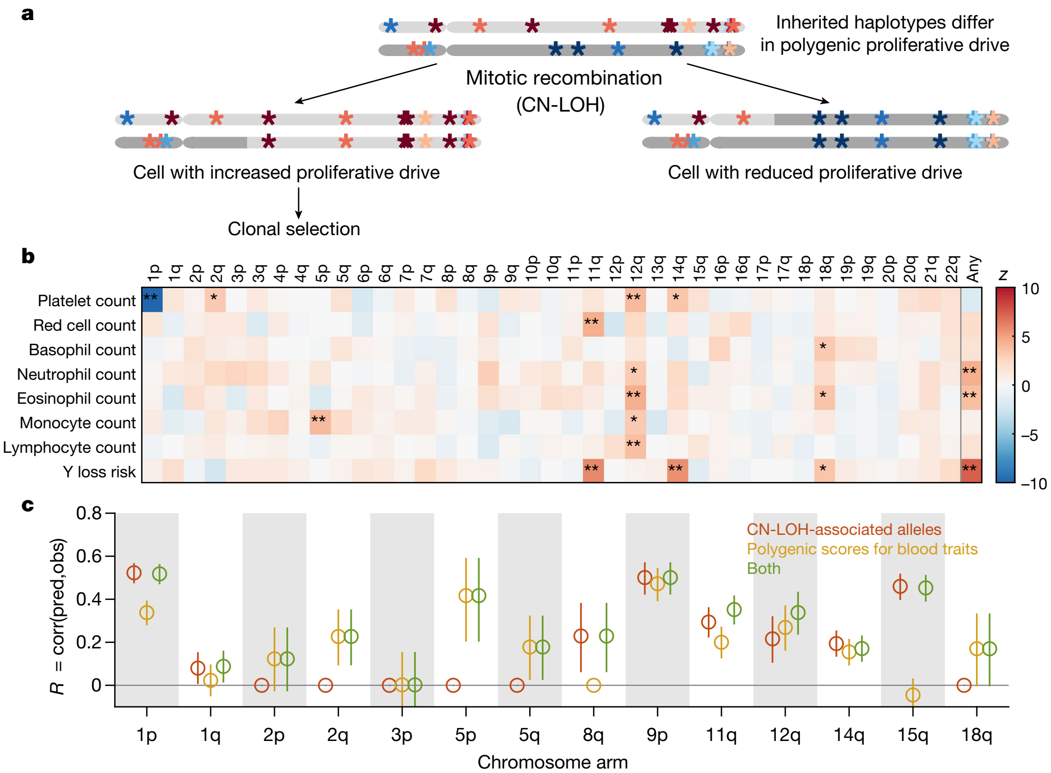


Fig. 2 | Polygenic and monogenic influences on clonal proliferation of cells with CN-LOH mutations. a, Two cellular outcomes of a CN-LOH mutation (mitotic recombination) involving homologous chromosome arms that bear inherited alleles with differing proliferative potentials. In one cell, the CN-LOH mutation has duplicated the chromosomal arm that has alleles that more strongly promote proliferation; proliferative polygenic drive increases, potentially resulting in clonal selection of the mutant cell. By contrast, the cell with the complementary CN-LOH mutation may have reduced tendency to proliferate. **b**, CN-LOH mutations in expanded clones broadly increase polygenic risk scores for increased blood-cell counts and risk of mosaic Y chromosome loss (a marker for clonal haematopoiesis[27]). The heat map displays changes in polygenic scores for each trait, averaged across all ascertained (expanded) CN-LOH mutations observed on each chromosome arm. Colour bar denotes the *z*-score. *FDR < 0.05; ***P* < 0.05 Bonferroni-corrected. **c**, Prediction of the direction of CN-LOH mutations (in expanded clones) from inherited alleles on the affected chromosome arms. Prediction accuracy (the correlation between predicted and observed CN-LOH direction) is plotted for predictions made using: only CN-LOH-associated alleles (red) (Extended Data Table 1, Supplementary Table 7); polygenic score differentials on affected chromosomal segments (orange); or both sources of information (green). Error bars denote 95% confidence intervals. Results are plotted for 14 chromosome arms for which at least one predictor was available. Numeric data and sample sizes are provided in Supplementary Tables 15, 18. Analyses of polygenic scores for control traits such as height and BMI are provided in Supplementary Table 16.

prediction using logistic regression on (i) the CN-LOH-associated alleles we had found (Extended Data Table 1); (ii) polygenic score differentials on chromosomal segments affected by CN-LOH; or (iii) both CN-LOH-associated alleles and polygenic score differentials (Methods). Polygenic scores and specific inherited CN-LOH-associated alleles each helped predict CN-LOH directions; combining both sources of information yielded the most predictive information, reaching significance (FDR < 0.05) for 12 out of 14 chromosome arms tested (Fig. 2c, Supplementary Table 18; we tested the 14 arms for which the prediction algorithm nominated at least one predictor for testing in a non-overlapping dataset, as described in Methods and Supplementary Table 19). The directions of CN-LOH mutations were correctly predicted for 59% ($P = 2.3 \times 10^{-44}$) of 5,582 CN-LOH events on these 14 arms (range 50–70%). Stronger inherited imbalances correlated with greater predictability: after restricting to events that involve larger imbalances in polygenic scores (top quintile), the accuracy of prediction increased to 72% ($P = 1.1 \times 10^{-82}$).

Cancer and cardiovascular risk associated with mCAs

Clonal haematopoiesis increases the risk of adverse health outcomes, including blood cancers, cardiovascular disease and mortality[1–4,8,28]. The size of the full UK Biobank dataset allowed us to further determine the extent to which different mCAs associate with distinct health outcomes (Methods). Thirteen specific mCAs significantly associated (FDR < 0.05) with subsequent haematological cancer diagnoses during 4–9 years of follow-up. The +12, 13q− and 14q− events conferred a more than 100-fold higher risk of chronic lymphocytic leukaemia, and *JAK2*-related 9p CN-LOH events conferred a 260-fold (89–631-fold) higher risk of myeloproliferative neoplasms, replicating previous results[9]; 4q and 7q CN-LOH events conferred a more than 70-fold higher risk of myelodysplastic syndromes (Fig. 3a, Supplementary Table 20). The +12 and 13q LOH events exhibited shared genetic risk with chronic lymphocytic leukaemia (Supplementary Table 21, Supplementary Note 8). The more common CN-LOH events on other chromosome arms also significantly increased the risk of blood cancer (aggregate hazard ratio = 2.84 (2.14–3.78), even after excluding the very-strong-effect *JAK2*-related events on 9p). (We corrected these analyses for age and sex and restricted to individuals with normal blood counts at assessment, no previous cancer diagnoses and no cancer diagnoses within one year of assessment.) We did not find a significant increase in cardiovascular risk among individuals with most categories of clones (with the notable exception of *JAK2*-related 9p CN-LOH events (Fig. 3b, Supplementary Table 22)), which suggests that the relationship between clonal haematopoiesis and cardiovascular disease[4,28] arises from clones that contain specific mutations (such as point mutations in the driver genes *DNMT3A*, *TET2*, *ASXL1* and *JAK2*).

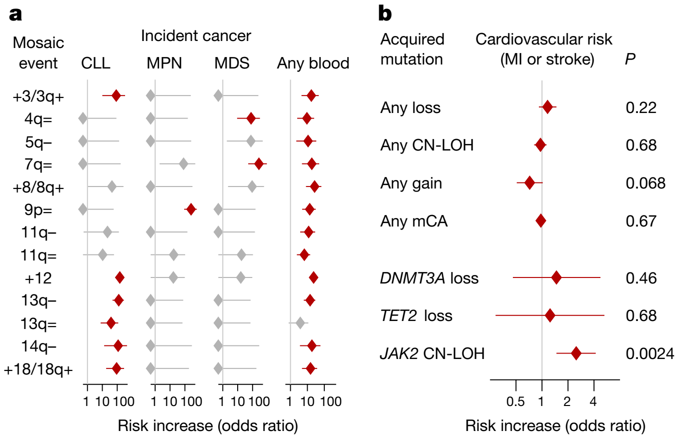


Fig. 3 | Associations of mCAs with incident cancers and cardiovascular disease. a, Clones with specific mCAs confer increased risk of incident blood cancers diagnosed more than 1 year after DNA collection in individuals with normal blood counts at assessment (Cochran–Mantel–Haenszel test adjusting for age and sex; error bars denote 95% confidence intervals). Seven out of nine associations that were previously reported[9] (all but 16p = and 20q−) replicate here. '=' denotes CN-LOH. CLL, chronic lymphocytic leukaemia; MDS, myelodysplastic syndrome; MPN, myeloproliferative neoplasm. **b**, Loss, CN-LOH and gain events (on any autosome) do not broadly increase the risk for incident myocardial infarction (MI) or stroke, but CN-LOH events on 9p (containing *JAK2*) do increase cardiovascular risk[28]. *P* values determined by two-sided Fisher's exact test on cases and controls matched for assessment year, age, sex, smoking, hypertension, BMI and type 2 diabetes status. Error bars denote 95% confidence intervals. See Methods for statistical tests. Numeric data and sample sizes are provided in Supplementary Tables 20, 22.

alleles that cause maternal and paternal haplotypes to differ in their tendency to promote proliferation. Furthermore, this estimate is probably only a lower bound on the strength and prevalence of polygenic selective pressure; as polygenic risk scores become informed by larger samples and lower-frequency alleles, their predictive accuracy tends to increase[24,25,34].

We were initially surprised that even a modest fraction of an individual's polygenic risk—arising from a single chromosome arm—could apparently create substantial clonal advantage. We believe that this results from an important aspect of clonal evolution: mutated cells compete with nearly isogenic cells in a common, shared environment. Estimates of the effects of common alleles and polygenic risk—which are usually made in the context of diverse genetic backgrounds and abundant environmental variation—are likely to underestimate the potential of such alleles to become instruments for clonal selection.

Because human populations contain abundant heterozygosity, and mitotic recombination events occur frequently over an individual's lifetime[9,32], imbalances in the proliferative potential of the homologous chromosome arms inherited from one's two parents provide a context in which clonal selection is almost inevitable. Managing this dynamic may present challenges for cytopoiesis throughout the lifespan in any genetically diverse species.

Discussion

These results demonstrate the clonal advantages conferred by CN-LOH, the common substitution of one chromosome arm for its homologous counterpart, which was present in most of the clones ascertained by mCAs (Extended Data Fig. 1a). Although the first-order gene-dosage effects of deletions and duplications are clear[1,2,5,6,29], clonal expansions of copy-neutral mutations are more common (Extended Data Fig. 1a) and have been more mysterious: the substitution of one chromosome arm for its inherited homologue does not modify gene dosage, so why would a cell that has undergone such a mutation gain a proliferative advantage? Our results, obtained from many genomic loci, point to a core principle: clonally expanded CN-LOH events routinely replace inherited chromosomal segments with homologous segments that more strongly promote proliferation. Examples of potent CN-LOH events have previously been observed in disease studies at a few loci where CN-LOH events provide second hits to acquired mutations[30], disrupt imprinting[31] or revert pathogenic mutations in rare monogenic disorders of the skin and blood[32,33]. It has recently been observed that CN-LOH mutations can also lead to clonal selection in healthy blood by modifying the allelic dosage of inherited rare variants at three loci[9]. The analyses described here suggest that this proliferative mechanism is in fact at work throughout the genome: we identified six more loci (*FH*, *NBN*, *MRE11*, *SH2B3*, *TCL1A* and *DLK1*) at which CN-LOH mutations gain advantage from at least 50 inherited alleles (some with sufficiently large effects to produce multiple clonal expansions in the same individual (Supplementary Table 23, Supplementary Note 9), and we observed a pervasive polygenic effect attributable to combinations of inherited alleles along chromosome arms. The finding here that the direction of 5,582 CN-LOH mutations (across 14 chromosome arms) could be predicted with 59% accuracy—based only on the alleles inherited on each arm—suggests that a substantial fraction of clonal expansions with CN-LOH (at least 18%—that is, 59% − 41%) are influenced by inherited

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-020-2430-6.

1. Jacobs, K. B. et al. Detectable clonal mosaicism and its relationship to aging and cancer. *Nat. Genet*. **44**, 651–658 (2012).
2. Laurie, C. C. et al. Detectable clonal mosaicism from birth to old age and its relationship to cancer. *Nat. Genet*. **44**, 642–650 (2012).
3. Genovese, G. et al. Clonal hematopoiesis and blood-cancer risk inferred from blood DNA sequence. *N. Engl. J. Med*. **371**, 2477–2487 (2014).
4. Jaiswal, S. et al. Age-related clonal hematopoiesis associated with adverse outcomes. *N. Engl. J. Med*. **371**, 2488–2498 (2014).
5. Machiela, M. J. et al. Characterization of large structural genetic mosaicism in human autosomes. *Am. J. Hum. Genet*. **96**, 487–497 (2015).
6. Vattathil, S. & Scheet, P. Extensive hidden genomic mosaicism revealed in normal tissue. *Am. J. Hum. Genet*. **98**, 571–578 (2016).
7. Zink, F. et al. Clonal hematopoiesis, with and without candidate driver mutations, is common in the elderly. *Blood* **130**, 742–752 (2017).
8. Abelson, S. et al. Prediction of acute myeloid leukaemia risk in healthy individuals. *Nature* **559**, 400–404 (2018).
9. Loh, P.-R. et al. Insights into clonal haematopoiesis from 8,342 mosaic chromosomal alterations. *Nature* **559**, 350–355 (2018).
10. Bycroft, C. et al. The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203–209 (2018).
11. Uziel, T. et al. Requirement of the MRN complex for ATM activation by DNA damage. *EMBO J*. **22**, 5612–5621 (2003).
12. Lee, J.-H. & Paull, T. T. ATM activation by DNA double-strand breaks through the Mre11-Rad50-Nbs1 complex. *Science* **308**, 551–554 (2005).
13. Deng, Y., Guo, X., Ferguson, D. O. & Chang, S. Multiple roles for MRE11 at uncapped telomeres. *Nature* **460**, 914–918 (2009).
14. Kimura, S., Roberts, A. W., Metcalf, D. & Alexander, W. S. Hematopoietic stem cell deficiencies in mice lacking c-Mpl, the receptor for thrombopoietin. *Proc. Natl Acad. Sci. USA* **95**, 1195–1200 (1998).
15. Solar, G. P. et al. Role of c-mpl in early hematopoiesis. *Blood* **92**, 4–10 (1998).
16. Seita, J. et al. Lnk negatively regulates self-renewal of hematopoietic stem cells by modifying thrombopoietin-mediated signal transduction. *Proc. Natl Acad. Sci. USA* **104**, 2349–2354 (2007).
17. Loh, P.-R., Palamara, P. F. & Price, A. L. Fast and accurate long-range phasing in a UK Biobank cohort. *Nat. Genet*. **48**, 811–816 (2016).
18. Loh, P.-R. et al. Reference-based phasing using the Haplotype Reference Consortium panel. *Nat. Genet*. **48**, 1443–1448 (2016).
19. Auer, P. L. et al. Rare and low-frequency coding variants in *CXCR2* and other genes are associated with hematological traits. *Nat. Genet*. **46**, 629–634 (2014).
20. Schultz, K. A. P. et al. *PTEN*, *DICER1*, *FH*, and their associated tumor susceptibility syndromes: clinical features, genetics, and surveillance recommendations in childhood. *Clin. Cancer Res*. **23**, e76–e82 (2017).

21. Landrum, M. J. et al. ClinVar: improving access to variant interpretations and supporting evidence. *Nucleic Acids Res*. **46**, D1062–D1067 (2018).
22. Van Hout, C. V. et al. Whole exome sequencing and characterization of coding variation in 49,960 individuals in the UK Biobank. Preprint at https://www.bioRxiv.org/content/10.1101/572347v1 (2019).
23. Meuwissen, T. H., Hayes, B. J. & Goddard, M. E. Prediction of total genetic value using genome-wide dense marker maps. *Genetics* **157**, 1819–1829 (2001).
24. Purcell, S. M. et al. Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature* **460**, 748–752 (2009).
25. Khera, A. V. et al. Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. *Nat. Genet*. **50**, 1219–1224 (2018).
26. Loh, P.-R. et al. Efficient Bayesian mixed-model analysis increases association power in large cohorts. *Nat. Genet*. **47**, 284–290 (2015).
27. Thompson, D. J. et al. Genetic predisposition to mosaic Y chromosome loss in blood. *Nature* **575**, 652–657 (2019).
28. Jaiswal, S. et al. Clonal hematopoiesis and risk of atherosclerotic cardiovascular disease. *N. Engl. J. Med*. **377**, 111–121 (2017).
29. Davoli, T. et al. Cumulative haploinsufficiency and triplosensitivity drive aneuploidy patterns and shape the cancer genome. *Cell* **155**, 948–962 (2013).
30. 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).
31. 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).
32. Choate, K. A. et al. Mitotic recombination in patients with ichthyosis causes reversion of dominant mutations in *KRT10*. *Science* **330**, 94–97 (2010).
33. Tesi, B. et al. Gain-of-function *SAMD9L* mutations cause a syndrome of cytopenia, immunodeficiency, MDS, and neurological symptoms. *Blood* **129**, 2266–2279 (2017).
34. Schizophrenia Working Group of the Psychiatric Genomics Consortium. Biological insights from 108 schizophrenia-associated genetic loci. *Nature* **511**, 421–427 (2014).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Methods

Data reporting

No statistical methods were used to predetermine sample size. The experiments were not randomized and the investigators were not blinded to allocation during experiments and outcome assessment.

UK Biobank cohort and genotyping

The UK Biobank is a very large prospective study of individuals aged 40–70 years at assessment[35]. Participants attended assessment centres between 2006 and 2010, where they contributed blood samples for genotyping and blood analysis and answered questionnaires about medical history and environmental exposures. In the years since assessment, health outcome data for these individuals (for example, diagnoses of cancer and cardiovascular disease) have been accruing via UK national registries and hospital records managed by the NHS.

We analysed genetic data from the full UK Biobank cohort, which consists of 488,377 individuals genotyped on the Affymetrix UK BiLEVE and UK Biobank Axiom arrays. The BiLEVE and Biobank arrays have more than 95% overlap and contain a total of 784,256 unique autosomal variants; 49,950 individuals were genotyped on the BiLEVE array[36] and the remaining individuals on the Biobank array. We restricted our analyses to 487,409 individuals passing previous genotyping quality control and previously imputed to approximately 93 million autosomal variants[10]; we re-phased these individuals using Eagle2[18] to improve phasing accuracy and imputed them to the union of the BiLEVE and Biobank arrays using Minimac3[37] (Supplementary Note 3). We further removed 427 individuals with low genotyping quality (B-allele frequency s.d. > 0.11 at heterozygous sites), 4,111 individuals with evidence of possible sample contamination (Supplementary Note 1) and 82 individuals who had withdrawn consent, leaving 482,789 individuals for analysis. We performed data processing using plink[38].

We also analysed exome-sequencing data available for 49,960 individuals[22]. To extend our rare variant association analyses to include variants identified in exome-sequenced individuals, we phased these individuals using Eagle2 and imputed into the full cohort using Minimac4 (Supplementary Note 3).

Detection of mCAs using genotyping intensities and long-range haplotype phase

As described previously[9], we detected mCAs in genotyping intensity data from blood DNA samples using an approach that leverages the chromosome-scale accuracy of statistical phasing in the UK Biobank cohort[17,18] (Supplementary Note 3). In brief, our approach harnesses long-range phase information to search for local imbalances between maternal and paternal allelic fractions in a cell population, enabling considerable gains in sensitivity for detection of large events at low cell fractions[9]. A full description of the method and a detailed exploration of its statistical properties compared to previous approaches are found in supplementary notes of ref. [9]. As described previously, we applied our approach to genotyping intensities that we transformed to $\log_2$ (R ratio) (LRR) and B-allele frequency (BAF) values[39] (which measure total and relative allelic intensities, respectively) after affine-normalization and GC wave-correction[1,9,40]. We estimated cell fractions of mCAs using the formulas relating BAF to cell fraction presented in table 1 of the supplementary note of ref. [1].

In analysing the full cohort, we made two minor modifications to our original approach. First, we halved the switch error rate parameter of our hidden Markov model (HMM) for BAF deviations, reflecting improved phasing accuracy in the full cohort. Second, we performed a few additional QC steps on the event calls to filter potential technical artefacts that we identified in the full data set; these filters affected <1% of the call set (Supplementary Note 1) and only affected four event calls from our previous analysis[9].

Our detection procedure produced a final call set of 19,632 autosomal mCAs at a nominal FDR of 0.05 (based on our phase randomization approach to estimate statistical significance)[9]. We verified that our FDR was well-controlled using an independent FDR estimation procedure based on the age distribution of event carriers[9]; this approach produced a concordant FDR estimate of 6.6% (4.5–8.6%) (Extended Data Fig. 2e, Supplementary Note 1.3). We also verified that rates of mosaic events on each chromosome were very consistent with our previous call set on the interim UK Biobank data[9]. For our current study, we re-analysed the interim samples for mosaicism using improved haplotype phasing in the full UK Biobank cohort; the increased phasing accuracy led to slightly higher detection sensitivity, such that the overall autosomal mCA detection rate increased by approximately 10%. As before, we observed that lower-confidence events tended to have uncertain copy number (because our power to detect allelic imbalances exceeds our power to distinguish CN-LOH from copy-number alterations) and less-precise event boundaries[9]; we provide information on the uncertainty of each event call in Supplementary Data. Our replication here of results previously reported from the interim UK Biobank release (for example, genomic distribution of mCAs, age and sex distribution of mCAs, relationships to blood cell indices, mCA risk loci and associations with haematological cancers) lends support to the validity of our methods.

Identifying variants associated with CN-LOH mutations in *cis*

We performed two types of association tests to identify inherited variants that influence mosaic CN-LOH mutations in *cis*. First, for each variant, we performed a Fisher test for association with a case–control phenotype specific to that variant: we considered samples to be cases if they carried a likely CN-LOH event containing the variant or within 4 Mb (to allow for uncertainty in event boundaries). We considered an event to be a likely CN-LOH event if it either (i) was called as a CN-LOH event or (ii) had undetermined copy number, extended to a telomere, and had $|\text{LRR}| < 0.02$. We performed this test on all typed and imputed variants and applied a genome-wide significance threshold of 5×10^{-8} for coding variants and 1×10^{-9} for all other variants.

Second, we searched for variants for which CN-LOH mutations in individuals heterozygous for the variant tended to preferentially duplicate one allele and remove the other allele from the genome. For each variant, we examined heterozygous individuals with a likely CN-LOH event overlapping the variant, and then performed a binomial test to check whether the CN-LOH direction tended to favour one allele versus the other. We restricted the binomial test to individuals in which the variant was confidently phased relative to the mosaic event (that is, there was no disagreement in five random resamples from the HMM used to call the event).

Given that the two association tests described above are independent, the second test provided a means of validating associations identified by the first test, as any spurious associations from the first test would have no correlation with CN-LOH direction, whereas variants truly associated with CN-LOH mutations in *cis* typically have strong associations with CN-LOH direction (Extended Data Table 1). We also performed a combined test to identify common variants that did not reach genome-wide significance in the first test alone (which was underpowered for common variants owing to small case counts) but reached significance using both tests together (Fisher's combined $P < 1 \times 10^{-8}$).

We restricted our association analyses to 455,009 individuals who reported European ancestry. Among these individuals, 96,590 pairs had previously been identified to be third-degree or closer relatives[10,41]. For each chromosome, we pruned the samples to an unrelated subset by removing one individual from each related pair, preferentially keeping (i) individuals with a likely CN-LOH on the chromosome and (ii) older controls. This pruning decreased total sample sizes to slightly less than 380,000 individuals (Supplementary Table 6). We verified that filtering on ancestry and relatedness in this way produced well-calibrated association test statistics (Extended Data Fig. 4, Supplementary Note 2).

Fine-mapping loci associated with CN-LOH mutations in *cis*

Given that our association analyses identified rare, large-effect coding variants in seven genes (*FH*, *NBN*, *MRE11*, *SH2B3*, *MPL*, *ATM* and *TM2D3*), we undertook fine-mapping analyses at these loci to uncover additional coding or splice variants in these genes likely to be objects of clonal selection (upon modification of allelic dosage via CN-LOH mutation). We tested variants in these genes in three categories: (i) missense variants with a Combined Annotation Dependent Depletion (CADD; v.1.3)[42] score above 20; (ii) predicted LoF variants (that is, stop-gained, frameshift, splice acceptor, or splice donor sites in any transcript annotated by VEP[43]); and (iii) probable pathogenic variants (according to ClinVar[21], downloaded 25 March 2019). We restricted these analyses to rare variants with MAF between 5×10^{-6} and 0.01. For directly genotyped variants, we required missingness < 0.01; for imputed variants, we required INFO > 0.2 (for variants imputed by UK Biobank using IMPUTE4[10]) or Minimac $R^2 > 0.4$ (for variants we imputed) (Supplementary Note 3). In addition to variants available from genotyping and imputation, we also tested two structural variants: a 454-bp deletion that we discovered in *MPL* by analysing exome-sequencing reads using IGV[44] and mosdepth[45] (Extended Data Fig. 5, Supplementary Note 4) and an approximately 70-kb deletion of *TM2D3* that was previously identified[9]. In total, 616 variants across the seven loci satisfied these criteria.

Of these 616 variants, 38 variants reached Bonferroni significance ($P < 8.1 \times 10^{-5}$) (Extended Data Table 1) and 52 variants reached FDR < 0.05 significance (assessed per gene) (Supplementary Table 7). We determined that all 52 FDR-significant variants were likely to causally drive independent associations with CN-LOH events in *cis*, based on the following lines of evidence. First, CN-LOH events acted on all 52 variants in the expected direction (consistently removing rare variants in *MPL* and duplicating rare variants in the other six genes) (Supplementary Table 7); by contrast, variants associated by chance would have random phase relative to CN-LOH events. Second, none of the 52 variants tagged other nearby variants with stronger associations (Fig. 1). By contrast, nearby variants in linkage disequilibrium (computed in-sample) with the 52 variants had weaker associations that are explained by tagging of the 52 variants (Fig. 1), and we verified that the variants in the *MPL* and *ATM* loci previously reported[9] each tagged one of the 52 variants (Supplementary Table 8). Third, none of the 52 variants tagged each other. The association signals at the 52 variants were driven by almost entirely non-overlapping sets of carriers who also had CN-LOH events in *cis*; the only overlap occurred between 11q CN-LOH individuals carrying the rs587779872 *ATM* missense variant (6 carriers with 11q CN-LOH) and the rs786204751 *ATM* stop-gain variant (2 carriers with 11q CN-LOH, both also carrying rs587779872) (Extended Data Fig. 7). The rs587779872 association remained significant in non-carriers of rs786204751, whereas the rs786204751 stop-gain mutation nullified the effect of the rs587779872 missense mutation (occurring downstream in *ATM*), which led us to conclude that these associations were likely to be independent.

Burden analyses to detect ultra-rare variants targeted by CN-LOH events

To identify CN-LOH events potentially explained by variants too rare to reach significance in single-variant association analyses, we analysed variant calls from exome sequencing of 49,960 participants from the UK Biobank[22] for a burden of ultra-rare coding and splice variants in individuals with CN-LOH events. As in our other association analyses, we restricted to individuals who reported European ancestry. Because these variant calls potentially contained a small fraction of somatic variants that had risen to cell fractions higher than approximately 20%, we included *DNMT3A*, *TET2* and *JAK2* in these analyses in addition to the seven genes at which we found inherited variants influencing clonal haematopoiesis. Beyond being frequently mutated in clonal haematopoiesis[3,4], *DNMT3A*, *TET2* and *JAK2* are also frequently overlapped by CN-LOH events (Extended Data Fig. 1a), which suggests that some CN-LOH events act on previously acquired point mutations in these genes.

As in our fine-mapping analyses, we considered variants annotated as (i) missense with CADD score > 20; (ii) predicted LoF; or (iii) probable pathogenic variants in ClinVar. We restricted to ultra-rare variants (MAF < 1×10^{-4}), with the exception of *JAK2*^{V617F}, which was called in 46 exome-sequenced individuals (MAF = 4.6×10^{-4}). (For *JAK2* and *ATM*, we used exome variant calls generated by UK Biobank using the 'functionally equivalent' (FE) pipeline[46], which we found provided slightly better power at these loci; for all other analyses, we used variant calls from Regeneron's Seal Point Balinese (SPB) pipeline[22].) For each gene, we examined individuals with CN-LOH events that span the gene (not already explained by any of the 52 variants identified in our association analyses) and tabulated the number of such individuals who carried any of the rare variants under consideration (Supplementary Table 10). We then computed a burden *P* value using a one-sided binomial test comparing the observed count to expectation (based on variant frequencies among 46,633 exome-sequenced individuals who reported European ancestry).

For each variant call potentially targeted by a CN-LOH event, we further examined allelic read depths from the exome-sequencing data to assess whether the variant was likely to be of inherited or acquired origin. Although read depths were generally insufficient to make a confident assessment on a per-variant level (and making this determination is complicated by mapping bias towards the reference allele[3]), the allelic depths broadly indicated that all or most variants implicated at our seven inherited risk loci were indeed inherited, and all or most variants implicated at *DNMT3A*, *TET2* and *JAK2* had been acquired somatically (Extended Data Fig. 8).

Genome-wide association studies for *trans* associations with any autosomal mCA

We tested common variants for *trans* associations with the presence of any detectable autosomal mCA. We computed association test statistics using BOLT-LMM[26,47] on 452,469 individuals (of which 16,366 were cases) who reported European ancestry and had imputation data available on autosomes and the X chromosome[10]. We included 20 principal components, age, age-squared, sex, smoking status, genotyping array and assessment centre as covariates in the linear mixed model to guard against confounding and to improve power by removing phenotypic variance explained by covariates.

Polygenic scores for blood cell traits

We analysed 29 blood count traits: counts and percentages of basophils, eosinophils, lymphocytes, monocytes, neutrophils, platelets, red cells, reticulocytes and high light scatter reticulocytes; white cell count, platelet and red cell distribution widths, immature reticulocyte fraction, haemoglobin concentration, mean corpuscular haemoglobin, mean corpuscular haemoglobin concentration, mean corpuscular volume, mean platelet volume, mean reticulocyte volume and mean sphered cell volume. (These traits constituted all available blood count traits except nucleated red blood cell indices, which were mostly zero.) We performed basic quality control and normalization on these traits using the following steps: (i) remove outliers (more than seven times farther from median than the nearest quartile); (ii) stratify into males, pre-menopausal females and post-menopausal females; (iii) within each stratum: (a) inverse-normal transform; (b) regress out age, age^2, height, height^2, BMI, BMI^2, ethnic group, alcohol use and smoking status; (c) inverse-normal transform again.

We computed polygenic score coefficients (that is, 'betas' in a linear predictor) for the traits listed above using the predBetasFile option of

BOLT-LMM[26,47], which estimates polygenic score coefficients using a Bayesian linear mixed model that accounts for linkage disequilibrium among variants. We computed coefficients for 709,999 autosomal and X chromosome variants in the intersection of the Biobank and BiLEVE arrays that passed quality control filters (allele frequency deviation < 0.02 between the arrays, missingness < 0.05, failed quality control in no more than one genotyping batch[10]). For each blood count phenotype, we restricted the sample set to individuals of self-reported European ancestry with non-missing phenotype (437,009–445,438 individuals depending on the phenotype). We ran BOLT-LMM using the same set of covariates we used in our *trans* genome-wide association studies. We computed polygenic risk coefficients for the loss of Y chromosome in blood cells using an analogous analysis restricted to males[27].

Among the 29 blood count parameters we considered, some of the parameters corresponding to abundances of blood cell types might be surrogates for enhanced cellular fitness (in many cases of mitotic progenitors rather than the cell types themselves). However, we also considered other parameters that reflect cell size or morphology (some of which had polygenic scores that tended to be decreased in expanded CN-LOH clones) (Extended Data Fig. 9). These relationships may reflect the production of abnormal cells by biologically altered stem cells, rather than cellular fitness itself (which may be a property of the unobserved haematopoietic stem cells); for example, mean platelet volume (MPV) has been reported to be a marker of myeloproliferative disorders. In our analyses that predict the direction of CN-LOH events, we allowed the logistic model to consider polygenic scores for all 29 parameters, the idea being that it would treat the polygenic scores as proxies for a variety of proliferative or cell-production tendencies and learn from the data how to weight them appropriately.

Polygenic score differentials for CN-LOH events

The polygenic score coefficients we computed for blood cell traits allowed us to estimate the extent to which CN-LOH mutations modified the genetic components of these traits. For each CN-LOH mutation, we computed the difference in polygenic score carried by the haplotype that was duplicated versus the haplotype that was removed. (This quantity is equal to the difference between the polygenic load of the mutant CN-LOH genome versus the original genome.) We determined which haplotype was duplicated and which was deleted using our hidden Markov model of phased BAF deviations[9], averaging across five posterior samples from the HMM. To identify chromosome arms in which CN-LOH events tended to increase polygenic load for specific blood cell traits, we averaged polygenic score differentials across all CN-LOH events on each arm and computed means and z-scores (independently for each blood cell trait) (Fig. 2b, Supplementary Table 15). To maximize power, we included all 'likely-CN-LOH' events in these analyses (that is, events called as CN-LOH as well as events with undetermined copy number that extended to a telomere and had $|LRR| < 0.02$, as in our *cis* association analyses), comprising a total of 11,638 likely-CN-LOH events on 39 chromosome arms containing at least 20 such events.

Prediction of CN-LOH directions using CN-LOH-associated alleles and polygenic scores

To assess the extent to which the direction of a CN-LOH event (that is, which affected haplotype is duplicated and which one is deleted) can be predicted based on the alleles inherited on each haplotype, we fit logistic models on the CN-LOH events on each chromosome arm using tenfold cross-validation. For each fold, we performed logistic regression using stepwise forward selection on three possible sets of predictors: (i) a single variable containing the difference in the number of CN-LOH-associated alleles (Extended Data Table 1, Supplementary Tables 7, 12) carried by the two affected haplotypes; (ii) 31 variables containing polygenic score differentials (for the 29 blood count indices, the Y-chromosome loss trait and myeloproliferative neoplasms) (Supplementary Note 7) between the two affected haplotypes; and (iii) all 32 variables together. We started forward selection using the 'number of CN-LOH-associated alleles' variable in analyses (i) and (iii) and an empty set of variables in analysis (ii). We stopped forward selection when model improvement was no longer significant at a 0.01 level. We restricted our prediction analyses to chromosome arms for which at least one variable was selected (on average across folds).

For each chromosome arm, we merged prediction results across the ten held-out folds and then assessed accuracy in two ways. First, we computed the Pearson correlation (R) between observed and predicted CN-LOH directions (using continuous-valued prediction probabilities from logistic regression). Second, we computed raw prediction accuracy (using binary, hard-called predictions). As in our analyses of polygenic score differentials, we included all likely-CN-LOH events (as defined above) to maximize power in these analyses.

Evaluating the ability of polygenic scores to predict CN-LOH directions in the same samples in which polygenic scores were computed does not result in overfitting. The reason is that we are evaluating a different kind of prediction accuracy: ability to predict which of an individual's two haplotypes is more likely to be made homozygous by a clonal CN-LOH event. This 'directionality' information is independent of the unphased genotype and phenotype information used to build the polygenic scores.

Enrichment of mCA types in specific blood lineages

To identify classes of mCAs linked to different blood cell types[9], we first classified mCAs based on chromosomal location and copy number. For each autosome, we defined five disjoint categories of mCAs that comprised most detected events: loss on p-arm, loss on q-arm, CN-LOH on p-arm, CN-LOH on q-arm and gain. We subdivided loss and CN-LOH events by arm but did not subdivide gain events because most gain events are whole-chromosome trisomies (Extended Data Fig. 1a). (We excluded the chr17 gain category because nearly all of these events arise from i(17q) isochromosomes already counted as 17p− events[9].)

For each mCA type, we computed enrichment among individuals with anomalous (top 1%) values of each of 14 normalized blood indices (counts and percentages of lymphocytes, basophils, monocytes, neutrophils, red cells and platelets, as well as distribution widths of red cells and platelets) using Fisher's exact test (two-sided; all P values reported are from two-sided statistical tests unless explicitly stated otherwise). We restricted these analyses to individuals who reported European ancestry, and reported significant enrichments passing an FDR threshold of 0.05 (Extended Data Fig. 1c, Supplementary Table 5).

UK Biobank cancer phenotypes

We analysed UK cancer registry data provided by UK Biobank for 81,401 individuals in our sample set who had one or more prevalent or incident cancer diagnoses. Cancer registry data included date of diagnosis and ICD-O-3 histology and behaviour codes, which we used to identify individuals with diagnoses of CLL, MPN, MDS or any blood cancer[48,49]. Because our focus was on the prognostic power of mCAs to predict diagnoses of incident cancers more than 1 year after DNA collection, we excluded all individuals with cancers reported before this time (from either cancer registry data or self-reporting of prevalent cancers). We also restricted our attention to the first diagnosis of cancer in each individual, and censored diagnoses after 30 September 2014, as suggested by UK Biobank (resulting in a median follow-up time of 5.7 years, s.d. 0.8 years, range 4–9 years). Finally, we restricted analyses to individuals who reported European ancestry. These exclusions reduced the total counts of incident cases to 199 (CLL), 138 (MPN), 70 (MDS) and 1,383 (any blood cancer). In our primary analyses, we further eliminated individuals with any evidence of potential undiagnosed blood cancer based on anomalous relevant blood indices (lymphocyte count outside the normal range of $1\text{–}3.5 \times 10^9$ cells per litre, red cell count $> 6.1 \times 10^{12}$ cells per litre for males or $>5.4 \times 10^{12}$ cells per litre for females, platelet count >450 × 10^9 cells per litre, red cell distribution width >15%), leaving incident case counts of 107 (CLL), 67 (MPN), 56 (MDS) and 1,055 (any blood cancer).

Estimation of cancer risk conferred by mCAs

To identify classes of mCAs associated with incident cancer diagnoses, we classified mCAs based on chromosomal location and copy number as described above. We then restricted our attention to the 78 classes with at least 30 carriers (to reduce our multiple hypothesis burden, given that we would be underpowered to detect associations with the rarer events). For each mCA class, we considered a sample to be a case if it contained only the mCA or if the mCA had highest cell fraction among all mCAs detected in the sample (that is, we did not count carriers of subclonal events as cases). We computed odds ratios and *P* values for association between mCA classes and incident cancers using Cochran–Mantel–Haenszel (CMH) tests to stratify by sex and by age in six 5-year bins. We used the CMH test to compute odds ratios (for incident cancer any time during follow-up) rather than using a Cox proportional hazards model to compute hazard ratios because both the mCA phenotypes and the incident cancer phenotypes were rare, violating assumptions of normality underlying regression. We reported significant associations passing an FDR threshold of 0.05 (Fig. 3a, Supplementary Table 20).

UK Biobank cardiovascular disease phenotypes

We analysed algorithmically defined cardiovascular events (myocardial infarction and stroke) identified by UK Biobank for 26,873 individuals in our sample set. Events had been identified based on information from baseline questionnaires and/or nurse-led interviews and from linked hospital admission and death registry datasets. We restricted our analyses to individuals with no missing cardiovascular covariates, self-reported European ancestry, and no prevalent cardiovascular disease, leaving 433,339 individuals, of which 8,094 had incident cardiovascular events during 5–10 years of follow-up.

Estimation of cardiovascular risk conferred by mCAs

To increase statistical power and limit the multiple hypothesis testing burden, we grouped all incident cardiovascular events into a single case–control phenotype and tested this phenotype for association with detectable mCAs. We considered mosaicism phenotypes defined by grouping all autosomal mCAs into one phenotype or by grouping mCAs by copy number (loss, CN-LOH or gain), and we also examined specific mCAs related to common mosaic point mutations[3,4,28]: focal deletions at *DNMT3A*, focal deletions at *TET2*, and CN-LOH mutations on 9p (which often duplicate a *JAK2*^{V617F} mutation[50–53]) (Extended Data Fig. 1a). For each category of mCAs, we created a subsample of mCA carriers and noncarriers matched on assessment year, age (in 1-year bins), sex, smoking status (current/ever/never), hypertension status, BMI (<25, 25–30, >30) and type 2 diabetes status, selecting carrier/noncarrier ratios to maximize power. We estimated cardiovascular risk conferred by each category of mCAs by performing Fisher's exact test on the matched sample sets.

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Data availability

Mosaic event calls are available in Supplementary Data in anonymized form. The mCA call set has also been returned to UK Biobank (as Return 2062) to enable individual-level linkage to approved UK Biobank applications. Access to the UK Biobank Resource is available by application (http://www.ukbiobank.ac.uk/).

Code availability

A standalone software implementation (MoChA) of the algorithm used to call mCAs is available at https://github.com/freeseek/mocha. Code used to perform the specific analyses in this study is available from the authors upon request (but unlike MoChA, this code is not immediately portable to other computing environments).

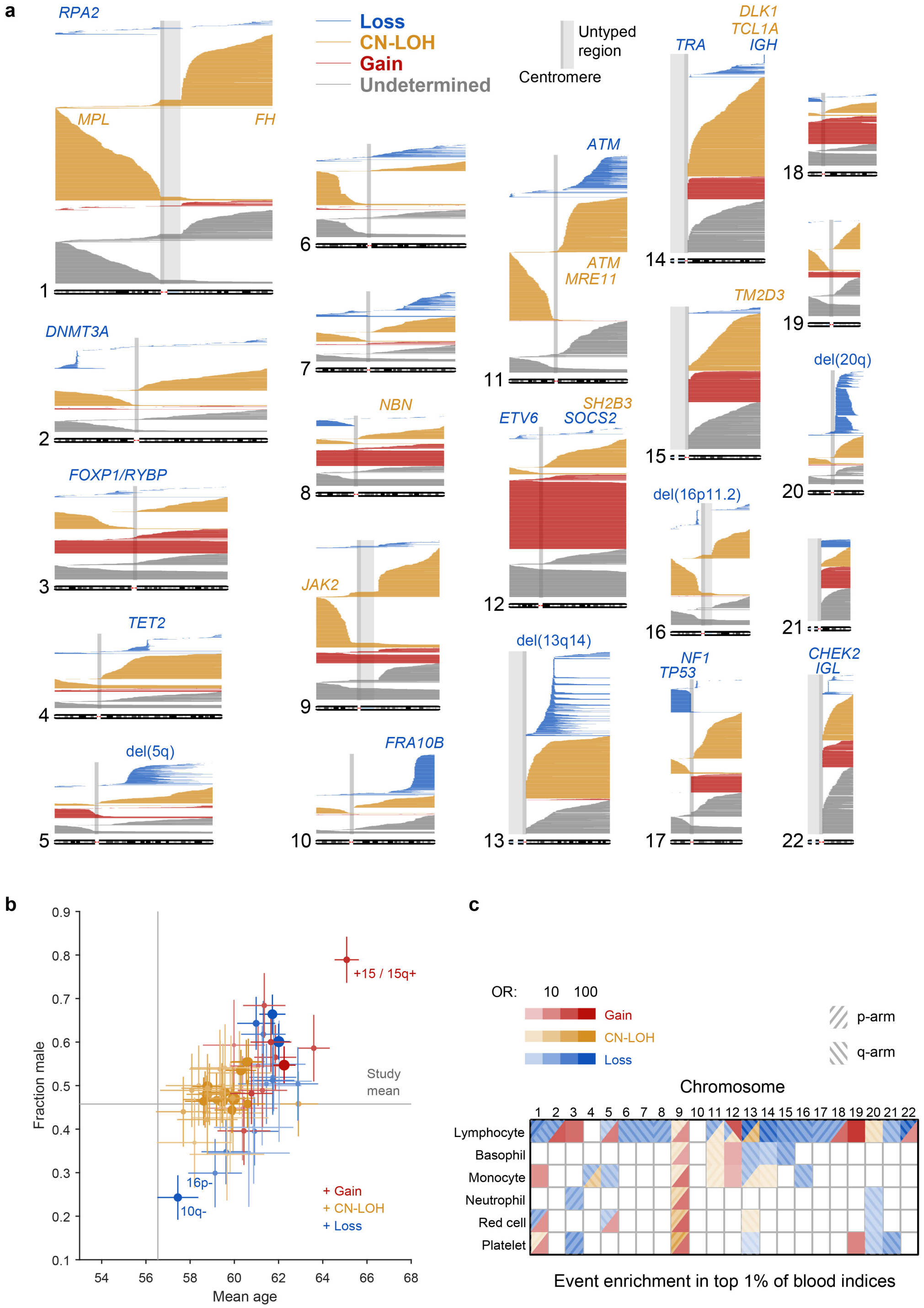
35. Sudlow, C. et al. UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Med*. **12**, e1001779 (2015).
36. Wain, L. V. et al. Novel insights into the genetics of smoking behaviour, lung function, and chronic obstructive pulmonary disease (UK BiLEVE): a genetic association study in UK Biobank. *Lancet Respir. Med*. **3**, 769–781 (2015).
37. Das, S. et al. Next-generation genotype imputation service and methods. *Nat. Genet*. **48**, 1284–1287 (2016).
38. Chang, C. C. et al. Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* **4**, 7 (2015).
39. Peiffer, D. A. et al. High-resolution genomic profiling of chromosomal aberrations using Infinium whole-genome genotyping. *Genome Res*. **16**, 1136–1148 (2006).
40. Diskin, S. J. et al. Adjustment of genomic waves in signal intensities from whole-genome SNP genotyping platforms. *Nucleic Acids Res*. **36**, e126 (2008).
41. Manichaikul, A. et al. Robust relationship inference in genome-wide association studies. *Bioinformatics* **26**, 2867–2873 (2010).
42. Rentzsch, P., Witten, D., Cooper, G. M., Shendure, J. & Kircher, M. CADD: predicting the deleteriousness of variants throughout the human genome. *Nucleic Acids Res*. **47**, D886–D894 (2019).
43. McLaren, W. et al. The Ensembl Variant Effect Predictor. *Genome Biol*. **17**, 122 (2016).
44. Thorvaldsdóttir, H., Robinson, J. T. & Mesirov, J. P. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Brief. Bioinform*. **14**, 178–192 (2013).
45. Pedersen, B. S. & Quinlan, A. R. Mosdepth: quick coverage calculation for genomes and exomes. *Bioinformatics* **34**, 867–868 (2018).
46. Regier, A. A. et al. Functional equivalence of genome sequencing analysis pipelines enables harmonized variant calling across human genetics projects. *Nat. Commun*. **9**, 4038 (2018).
47. Loh, P.-R., Kichaev, G., Gazal, S., Schoech, A. P. & Price, A. L. Mixed-model association for biobank-scale datasets. *Nat. Genet*. **50**, 906–908 (2018).
48. 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).
49. Arber, D. A. et al. The 2016 revision to the World Health Organization classification of myeloid neoplasms and acute leukemia. *Blood* **127**, 2391–2405 (2016).
50. Jones, A. V. et al. *JAK2* haplotype is a major risk factor for the development of myeloproliferative neoplasms. *Nat. Genet*. **41**, 446–449 (2009).
51. Kilpivaara, O. et al. A germline *JAK2* SNP is associated with predisposition to the development of *JAK2*^{V617F}-positive myeloproliferative neoplasms. *Nat. Genet*. **41**, 455–459 (2009).
52. Olcaydu, D. et al. A common *JAK2* haplotype confers susceptibility to myeloproliferative neoplasms. *Nat. Genet*. **41**, 450–454 (2009).
53. Koren, A. et al. Genetic variation in human DNA replication timing. *Cell* **159**, 1015–1026 (2014).
54. Gusev, A. et al. Whole population, genome-wide mapping of hidden relatedness. *Genome Res*. **19**, 318–326 (2009).

Acknowledgements We thank S. Bakhoum, S. Raychaudhuri, M. Sherman, S. Elledge and C. Terao for discussions. This research was conducted using the UK Biobank Resource under application no. 19808. P.-R.L. was supported by US National Institutes of Health (NIH) grant DP2 ES030554, a Burroughs Wellcome Fund Career Award at the Scientific Interfaces, the Next Generation Fund at the Broad Institute of MIT and Harvard, a Glenn Foundation for Medical Research and AFAR Grants for Junior Faculty award, and a Sloan Research Fellowship. G.G. and S.A.M. were supported by US NIH grant R01 HG006855. G.G. was supported by US Department of Defense Breast Cancer Research Breakthrough Award W81XWH-16-1-0316. Computational analyses were performed on the O2 High Performance Compute Cluster, supported by the Research Computing Group, at Harvard Medical School (http://rc.hms.harvard.edu), and on the Genetic Cluster Computer (http://www.geneticcluster.org) hosted by SURFsara and financially supported by the Netherlands Scientific Organization (NWO 480-05-003 PI: Posthuma) along with a supplement from the Dutch Brain Foundation and the VU University Amsterdam. We thank S. Elledge, B. Ebert, and C. Patil for helpful comments on the manuscript.

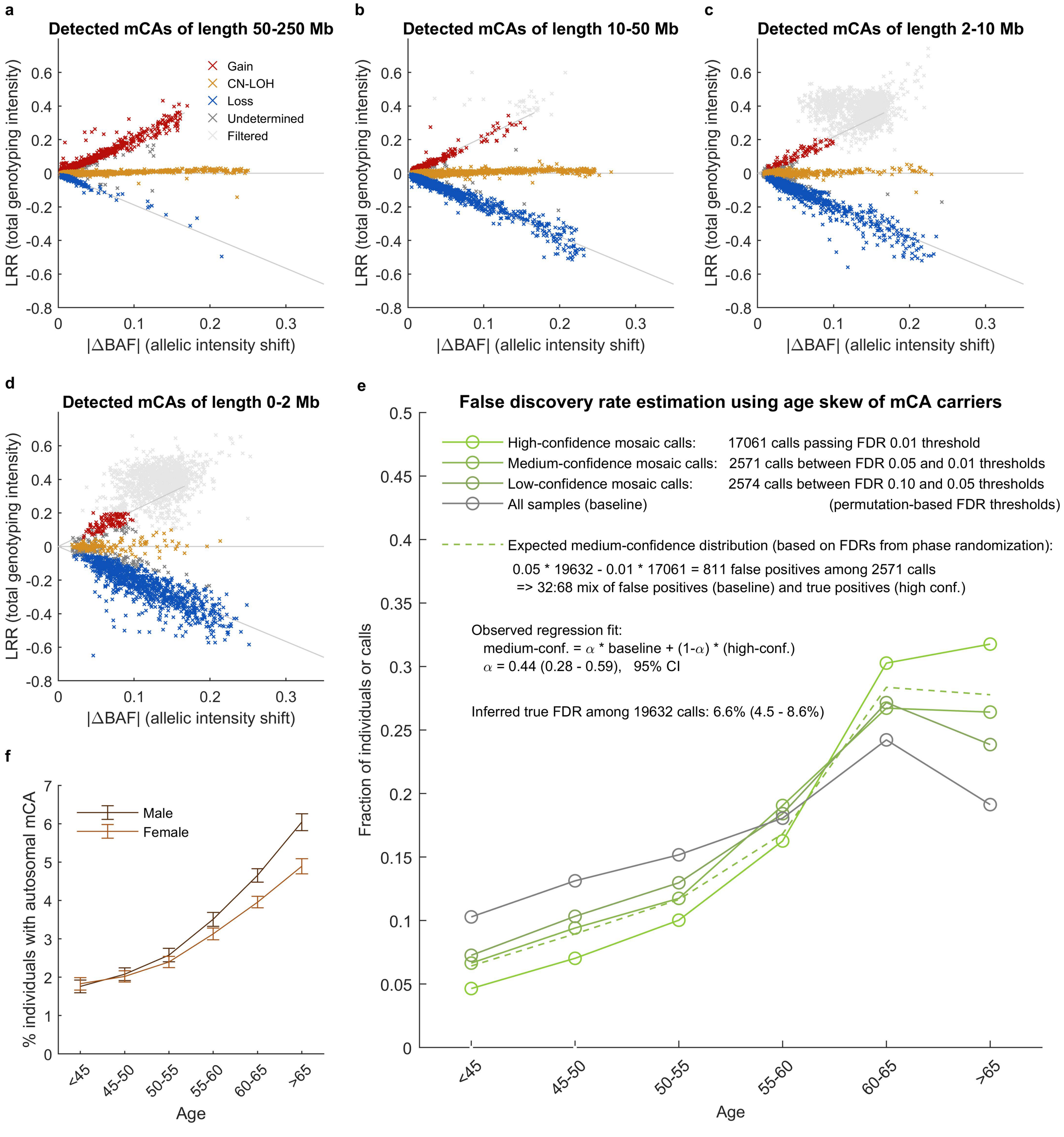
Author contributions P.-R.L., G.G. and S.A.M. designed the study. P.-R.L. performed computational analyses. P.-R.L., G.G. and S.A.M. wrote the paper.

Competing interests Patent application PCT/WO2019/ 079493 has been filed on the mCA detection method used in this work.

Additional information
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-020-2430-6.
Correspondence and requests for materials should be addressed to P.-R.L., G.G. or S.A.M.
Peer review information *Nature* thanks Paul Scheet, George Vassiliou and John Witte for their contribution to the peer review of this work.
Reprints and permissions information is available at http://www.nature.com/reprints.



Extended Data Fig. 1 | Mosaic chromosomal alterations detected among 482,789 UK Biobank participants. a, Each horizontal line corresponds to an mCA; a total of 19,632 autosomal events in 17,111 unique individuals are displayed. Detected events are colour-coded by copy number of the affected chromosome or segment (orange, LOH; blue, loss/deletion; red, gain/duplication). Focal deletions are labelled in blue with the names of putative target genes. Loci containing inherited variants influencing somatic events in *cis* are labelled in the same colour as the corresponding mCA (orange for CN-LOH-associated loci, blue for losses). **b**, Sex and age distributions of individuals with detected mosaic events. Marker size and colour intensity increase with event frequency. Error bars denote 95% confidence intervals. Sample sizes are provided in Supplementary Table 1 and numeric data are provided in Supplementary Table 4. Three events with unusual sex biases (gains on chromosome 15, 16p11.2 deletions and 10q terminal deletions) were previously reported[9], all of which replicated here. We have not identified a mechanism that could explain the sex biases. The overall tendency of male enrichment for most mCAs raises the possibility that environmental exposures could result in genomic insults that lead to mCAs; however, the heterogeneity of the level of male enrichment across different mCAs suggests that the mechanisms producing sex biases may be event-specific. **c**, Enrichment of mosaic chromosomal alterations in individuals with anomalously high blood indices. Different mCAs are significantly enriched (FDR of 0.05; one-sided Fisher's exact test) among $n = 455{,}009$ individuals with anomalous blood counts in different blood lineages (adjusted for age, sex and smoking status). Events were grouped by chromosome and copy number, with loss and CN-LOH events subdivided by p-arm versus q-arm. (We did not subdivide gain events by arm because most gain events are whole-chromosome trisomies.) Numeric data are provided in Supplementary Table 5.



High-confidence mosaic calls: 17061 calls passing FDR 0.01 threshold

Medium-confidence mosaic calls: 2571 calls between FDR 0.05 and 0.01 thresholds

Low-confidence mosaic calls: 2574 calls between FDR 0.10 and 0.05 thresholds

All samples (baseline) (permutation-based FDR thresholds)

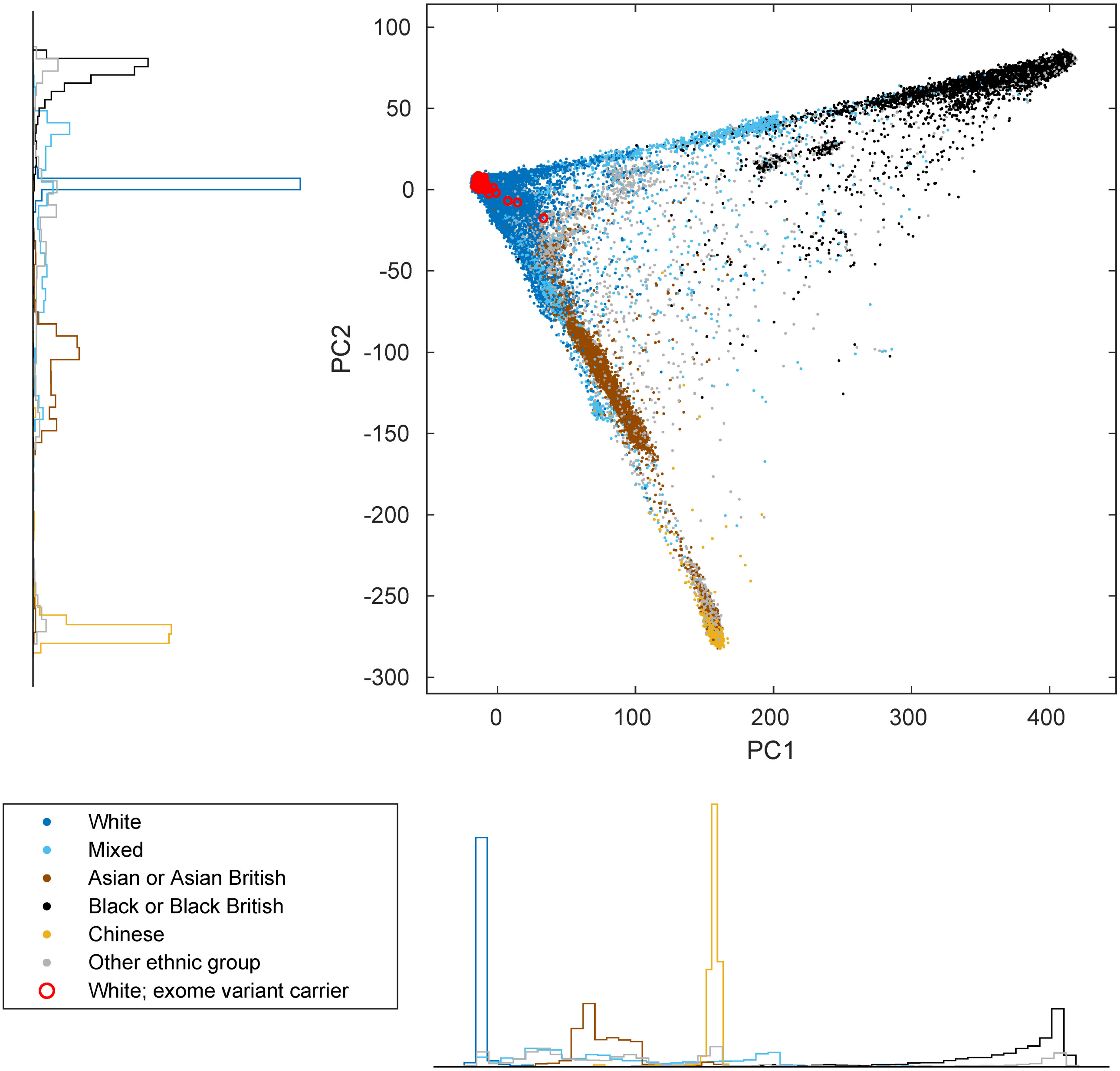
Expected medium-confidence distribution (based on FDRs from phase randomization):
0.05 * 19632 - 0.01 * 17061 = 811 false positives among 2571 calls
=> 32:68 mix of false positives (baseline) and true positives (high conf.)

Observed regression fit:
medium-conf. = α * baseline + (1-α) * (high-conf.)
α = 0.44 (0.28 - 0.59), 95% CI

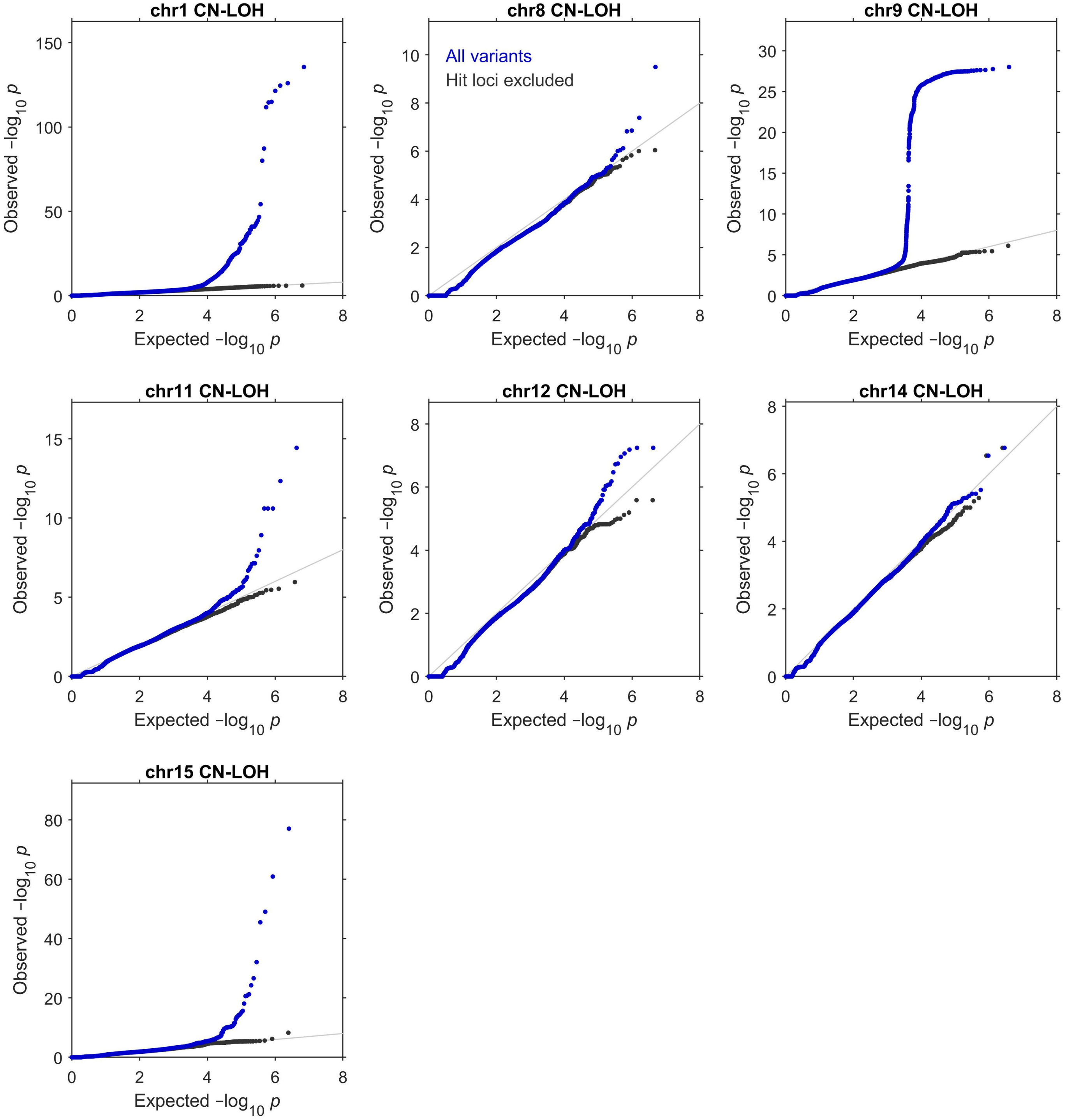
Inferred true FDR among 19632 calls: 6.6% (4.5 - 8.6%)

Extended Data Fig. 2 | See next page for caption.

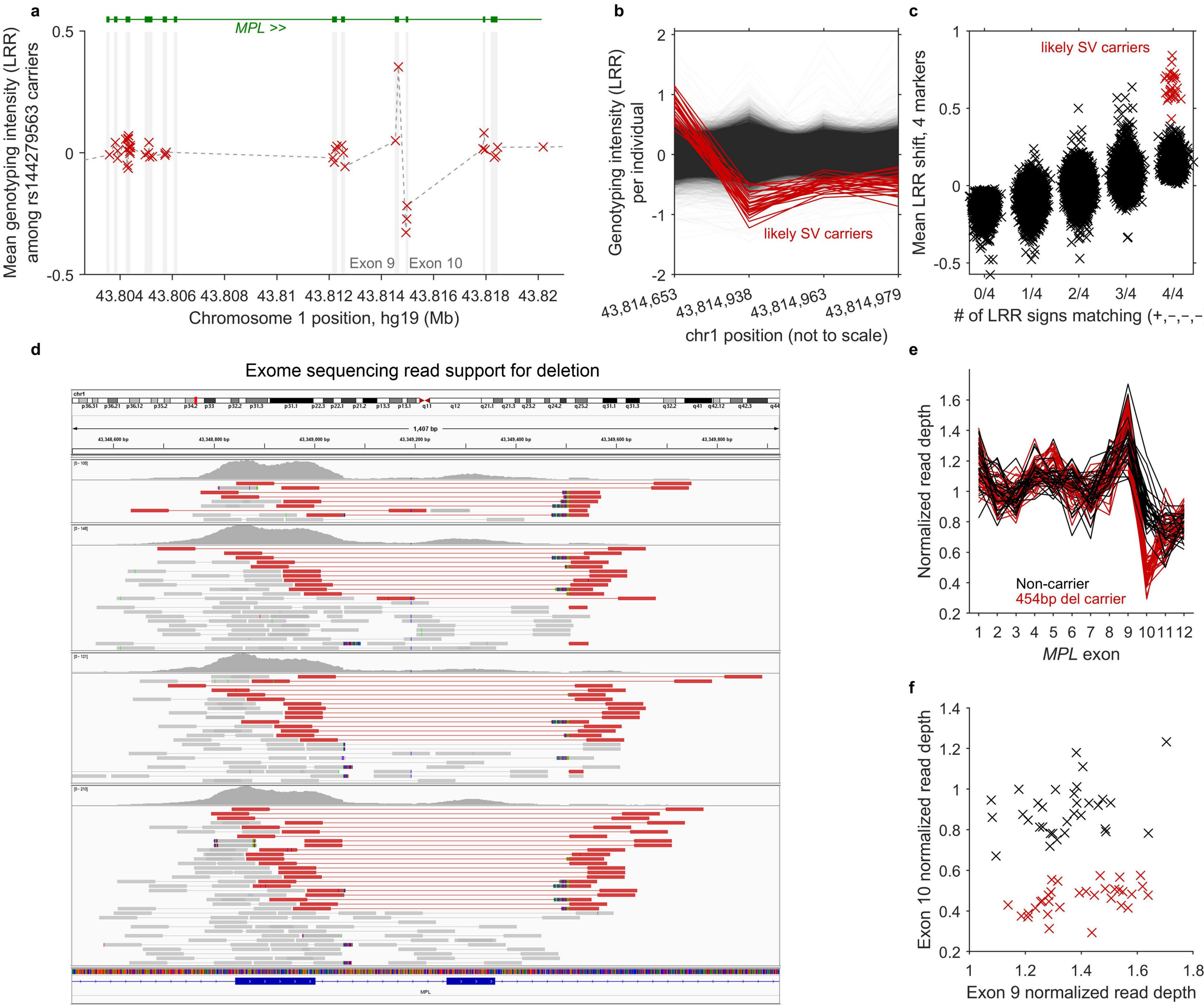
Extended Data Fig. 2 | Copy number determination and quality control of mosaic chromosomal alteration calls. a–d, Total versus relative allelic intensities of mCAs detected on each chromosome. Mean $\log_2$(R ratio) (LRR) of each detected mCA is plotted against estimated change in B allele frequency at heterozygous sites (|ΔBAF|). The data exhibit the characteristic 'arrowhead' pattern in which LRR/|ΔBAF| approximately equals a positive constant for gain events, zero for CN-LOH events, and a negative constant for loss events. Possible constitutional duplications were filtered according to thresholds on LRR and |ΔBAF| defined in Supplementary Note 1. Constitutional duplications have expected |ΔBAF| = 1/6 and have LRR values of approximately 0.36 in this dataset. 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. **e**, Estimation of FDR using age distributions of individuals with mCA calls. We generated age distributions for (i) 'high confidence' events passing a permutation-based FDR threshold of 0.01 (bright green); (ii) 'medium confidence' events below the FDR threshold of 0.01 but passing an FDR threshold of 0.05 (darker green); and (iii) 'low confidence' events below the FDR threshold of 0.05 but passing an FDR threshold of 0.10 (darkest green; excluded from our call set but plotted for context). We compared these distributions to the overall age distribution of UK Biobank participants (grey). On the basis of the numbers of events in each category, approximately 32% of medium-confidence detected events are expected to be false positives. To estimate our true FDR, we regressed the medium-confidence age distribution on the high-confidence and overall age distributions, reasoning that the medium-confidence age distribution should be a mixture of correctly called events (with age distribution similar to that of the high-confidence events) and spurious calls (with age distribution similar to the overall cohort). We observed a regression weight of 0.44 for the component corresponding to spurious calls, in good agreement with expectation, and indicating a true FDR of 6.6% (4.5–8.6%, 95% confidence interval based on regression fit on $n = 6$ age bins). **f**, Fractions of individuals with at least one detected autosomal mCA stratified by age and sex. Error bars denote 95% confidence intervals. Numeric data are provided in Supplementary Table 3.



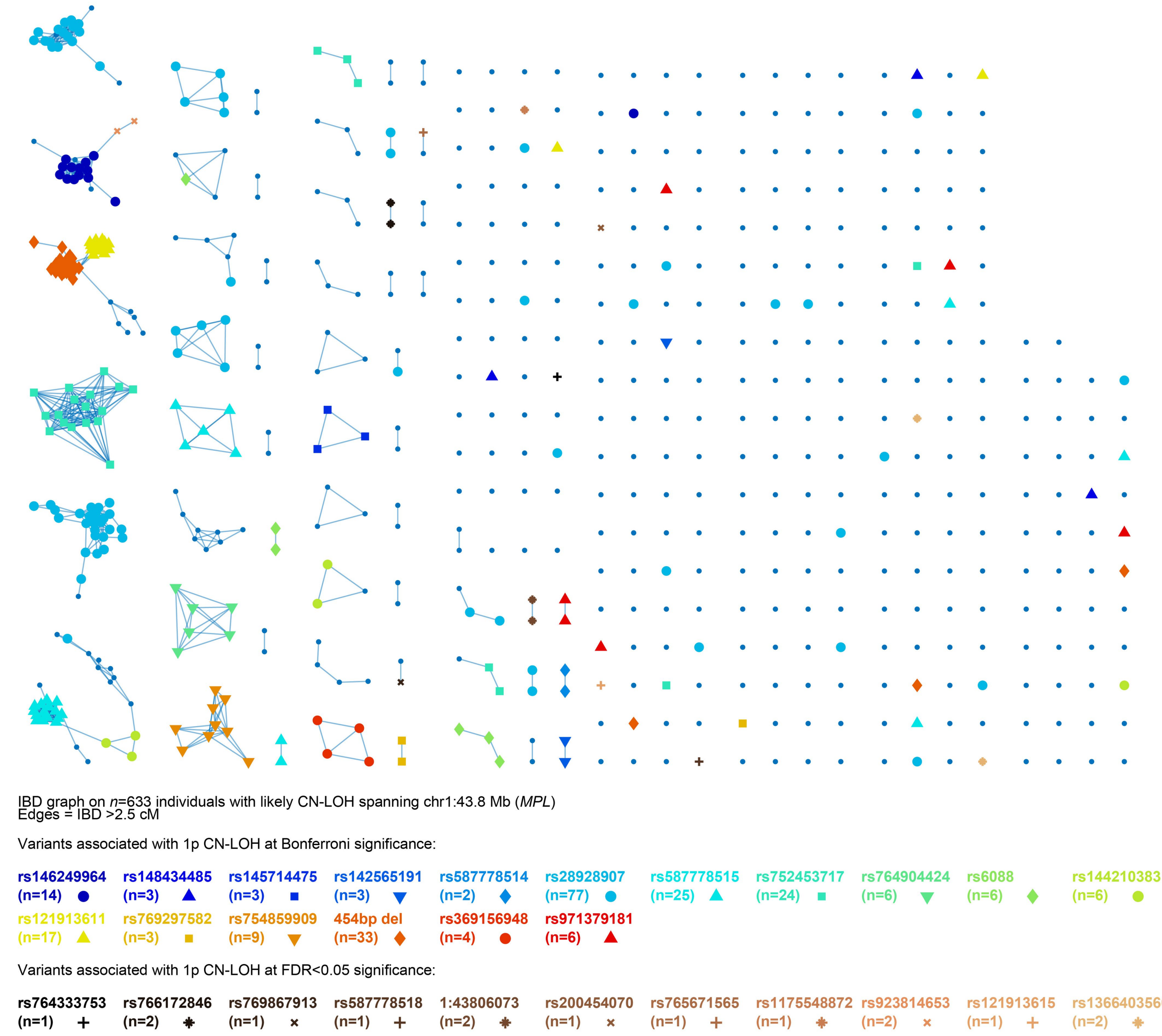
Extended Data Fig. 3 | Principal component plot of UK Biobank participants. Individuals are plotted by their first two genetic principal component coordinates as computed by UK Biobank[10] and coloured according to self-reported ethnic background. Red circles indicate individuals identified in our exome analyses (of self-reported white individuals with mosaic CN-LOH events) as carriers of rare coding or splice variants in frequently-targeted genes. Marginal density histograms stratified by self-reported ethnic background are provided next to the PC1 and PC2 axes.



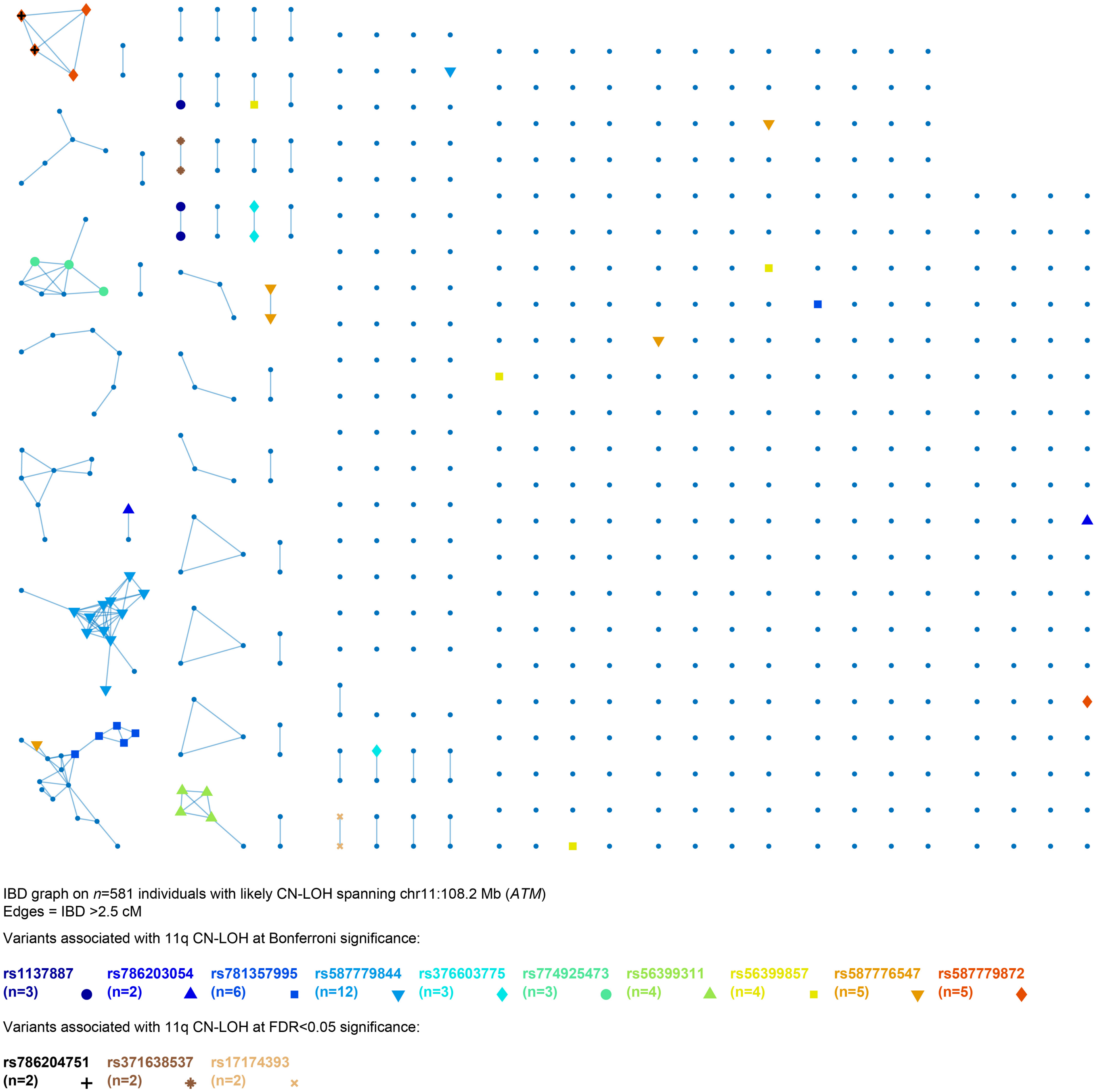
Extended Data Fig. 4 | Quantile–quantile plots of *P* values produced by association analyses. These plots verify the calibration of the statistical tests we used to identify the genome-wide significant associations reported in Extended Data Table 1 (see legend for details of statistical tests and sample sizes). In each plot, the blue dots correspond to an analysis of all variants tested, and the black dots correspond to an analysis in which regions surrounding significant associations were excluded. Specifically, the plots respectively exclude chr1:35–55 Mb (*MPL*), chr1:239–244 Mb (*FH*), chr8:88–93 Mb (*NBN*), chr9:2.5–7.5 Mb (*JAK2*), chr11:92–97 Mb (*MRE11*), chr11:103–113 Mb (*ATM*), chr12:109–114 Mb (*SH2B3*), chr14:92.5–102.5 Mb (*TCL1A* and *DLK1*) and chr15:100Mb–qter (*TM2D3*) (hg19 coordinates). In all cases, exclusion of the hit regions (which account for a small fraction of the variants tested) resulted in a distribution close to the expected null.



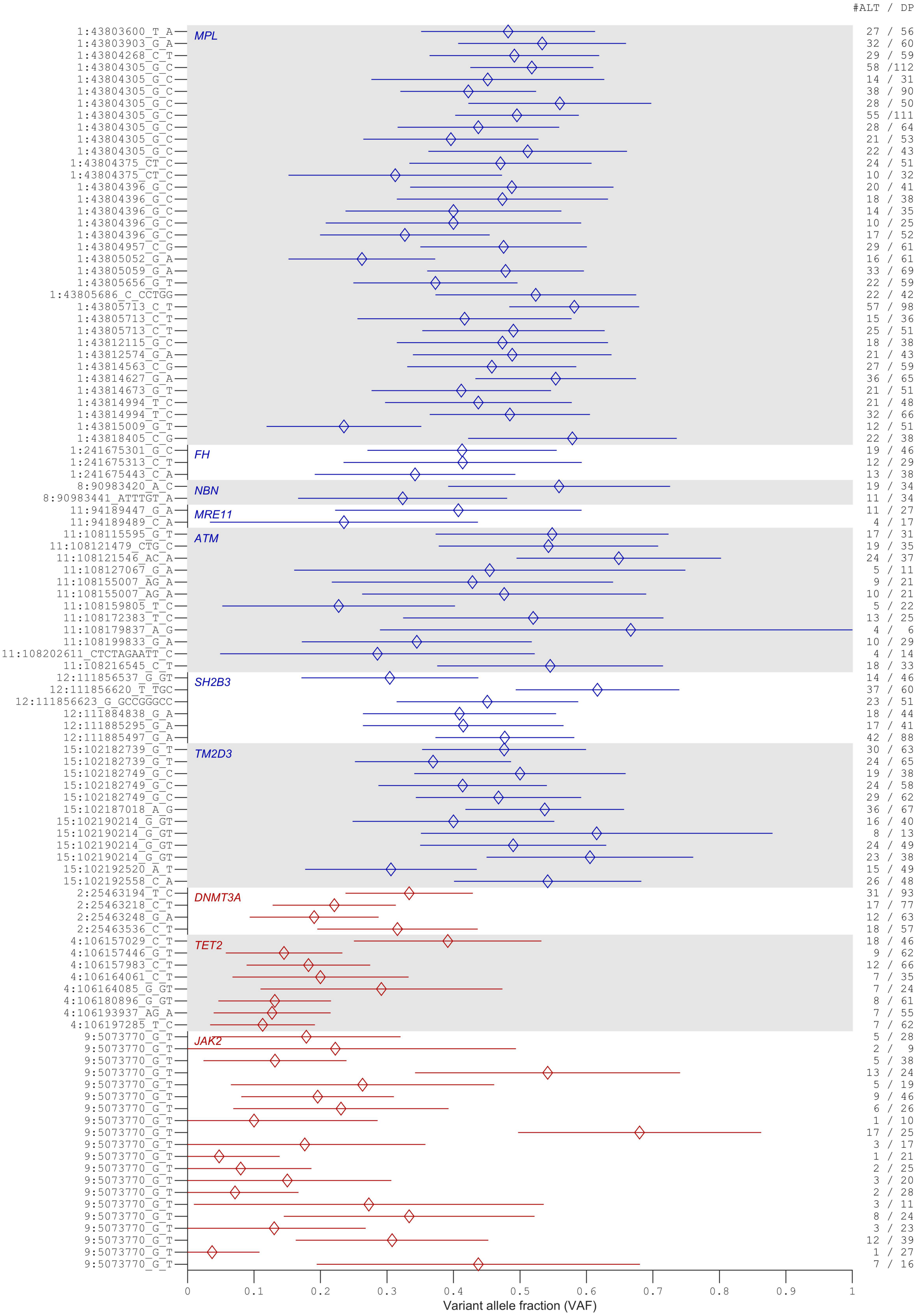
Extended Data Fig. 5 | Identification and validation of an inherited *MPL* structural variant. We suspected that an association between rs144279563 and acquired 1p CN-LOH mutations might tag a causal structural variant in *MPL*. (Although rs144279563 is approximately 1.5 Mb downstream of *MPL*, it is sufficiently rare to be in linkage disequilibrium with variants several megabases away.) We therefore examined genotyping intensities at *MPL* from 49,950 individuals typed on the BiLEVE chip (which contains more probes within *MPL* than the Biobank chip, on which the remaining individuals were typed). **a**, Mean genotyping intensities over 42 carriers of the rs144279563 rare allele exhibit a sharp increase at the end of *MPL* exon 9 (1 genotyping probe) followed by a sharp decrease in exon 10 (3 genotyping probes). **b**, **c**, Closer inspection of genotyping intensities at the 4 probes across all BiLEVE individuals enabled identification of 27 individuals likely to carry an inherited structural variant (20 of which carry the rs144279563 rare allele). We called this variant in the BiLEVE cohort using two criteria: (i) correct sign of LRR at the 4 probes (+, −, −, −); and (ii) mean signed LRR shift >0.4 over the four probes. **d**, Read support for a 454-bp deletion spanning *MPL* exon 10 in exome-sequenced individuals. We used IGV[44] to plot paired-end reads aligning in or near *MPL* exons 9 and 10 in four exome-sequenced individuals imputed to carry the *MPL* structural variant (and also mosaic for 1p CN-LOH events). Read pairs highlighted in red have unusually long insert sizes, consistent with a deletion of genomic sequence between the aligned reads. Multicoloured read segments indicate clipped reads in which one end of a read stops aligning to the reference genome. On the left side of the deletion, clipped reads align through hg19 base pair 43,814,728 (…AGGGA**CTGGG**; last five matching bases in bold for comparison to sequences below), with mismatches consistently occurring starting from 43,814,729 rightward (hg19: CGCCG…). On the right side of the deletion, clipped reads align starting from 43,815,178 (**CTGGG**ACTCG…), with mismatches starting from 43,815,177 leftward (hg19: …CACCT). Examination of individual clipped reads revealed sequence matching …AGGGA**CTGGG**ACTCG…, indicating deletion of 5 bp (**CTGGG**) in addition to the 449 bp between aligning read segments. In this legend we have used hg19 coordinates for consistency with the rest of this Article; the IGV plot uses hg38 coordinates because reads had been aligned to hg38 (amounting to an offset of −465,671 bp relative to hg19 at *MPL*). **e**, **f**, Decreased read depth at exon 10 in all 32 imputed carriers of the *MPL* exon 10 deletion who had been exome-sequenced. We used mosdepth[45] to compute mean read depth across all 12 *MPL* exons in the 32 exome-sequenced imputed deletion carriers along with 32 controls. We normalized read depth in each individual by dividing by mean read depth across exons 1–8 and 11–12. All 32 imputed carriers of the exon 10 deletion had lower exon 10 normalized read depths than all 32 controls. We did not observe any evidence of increased read depth in exon 9 in carriers versus controls.



Extended Data Fig. 6 | Identity-by-descent graph at *MPL* among individuals with likely 1p CN-LOH events spanning *MPL*. We called identity-by-descent (IBD) tracts using GERMLINE with haplotype extension[54]. Coloured nodes indicate carriers of the 28 rare coding or splice variants we observed to be independently (and probably causally) associated with 1p CN-LOH mutations (always replacing the rare allele with the reference allele) (Extended Data Table 1, Supplementary Table 7). (The numbers of carriers listed for each variant here are slightly higher than in the 'allelic shift' columns of Extended Data Table 1 and Supplementary Table 7 because allelic shifts could only be confidently ascertained for a subset of carriers.) The presence of additional IBD clusters not carrying any of the 28 highlighted variants suggests that even more causal variants in *MPL* remain to be discovered.



Extended Data Fig. 7 | Identity-by-descent graph at *ATM* among individuals with likely 11q CN-LOH events spanning *ATM*. We called IBD tracts using GERMLINE with haplotype extension[54]. Coloured nodes indicate carriers of the eight rare coding or splice variants we observed to be independently (and probably causally) associated with 11q CN-LOH mutations (always making the rare allele homozygous) (Extended Data Table 1, Supplementary Table 7). The presence of additional IBD clusters not carrying any of the highlighted variants suggests that even more causal variants in *ATM* remain to be discovered. The two carriers of rs786204751 are also carriers of rs587779872, as discussed in Methods.



Extended Data Fig. 8 | Variant allele fractions of rare coding or splice variants likely to be targets of CN-LOH mutations in exome-sequenced individuals. Variant allele fractions (VAF; the number of reads matching the alternative allele divided by the total number of reads matching either the reference or the alternative allele) are plotted for each variant call identified as the potential target of a CN-LOH event (from either association analyses or burden analyses). Error bars denote 95% confidence intervals approximated using binomial standard errors multiplied by 1.96. Allelic read depths for variants identified at *DNMT3A*, *TET2* and *JAK2* are broadly indicative of somatic origin (VAF < 0.5), whereas read depths for variants at the seven inherited risk loci are broadly consistent with inherited variation (VAF ≈ 0.5). Read depths were generally insufficient to make a confident assessment of somatic versus inherited origin on a per-variant level, as evidenced by wide VAF error bars; in addition, making this determination is further complicated by mapping bias towards the reference allele, which can produce VAF lower than 0.5 even for inherited variants[3].

z-scores for changes in polygenic scores produced by CN-LOH mutations on each chromosome arm

	Cell abundance traits																					Cell morphology traits							
Arm	PLT#	Pct	RBC#	Hgb	Hct	RET#	RET%	HLR#	HLR%	IRF	WBC#	NEU#	NEU%	EOS#	EOS%	MON#	MON%	BAS#	BAS%	LYM#	LYM%	PDW	MPV	RDW	MCV	MRV	MSCV	MCH	MCHC
1p	-11.5	-10.5	1.46	-0.32	-0.52	-0.12	-0.32	-0.57	-0.7	-0.93	-0.14	0.25	0.28	0.61	0.6	0.97	0.34	-0.45	-0.94	-0.57	-0.74	2.38	5.93	-0.38	-2.21	-1.85	-2.01	-1.62	0.89
1q	1.39	3.1	-0.96	-0.8	-1.22	-1.3	-1.1	-2.07	-2	-2.61	1.48	1.27	2.01	-1.96	-2.15	-1.18	-2.32	0.25	-0.58	0.08	-0.96	0.41	1.02	-0.55	0.15	-0.78	-0.41	0.03	-0.22
2p	0.48	0.36	0.63	0.4	-0.03	0.15	-0.22	-0.61	-0.7	-1.61	1.34	1.56	0.81	1.88	0.91	-0.65	-3.06	1.73	0.01	0.84	-0.63	0.39	0.12	0.88	-0.34	-0.4	-0.37	-0.02	0.9
2q	3.24	4.03	1.18	2.13	2.06	1.42	1.23	1.06	1	0.82	2.27	2.44	2.39	1.01	0.65	1.37	-0.17	0.94	-1.56	1	-2.17	-0.91	0.11	-2.02	1.08	1.33	1.27	0.98	-0.44
3p	0.12	0.41	1.26	2.31	2.12	2.37	1.9	2	1.58	2.09	2.05	2.72	0.91	-1.54	-2.37	0.77	-0.17	0.74	-0.5	0.37	-1.46	0.97	0.61	-1.92	0.22	-0.57	-0.69	0.49	0.78
3q	1.35	2.31	-1.7	-0.76	-0.51	-0.07	0.42	1.04	1.56	2.02	2.6	2.41	1.62	1.15	-0.1	1.97	0.24	1.18	0.17	0.42	-2.25	-0.88	1.07	-1.23	0.83	0.89	1.28	0.74	-0.44
4p	0.3	1.08	1.09	2.51	2.6	-0.84	-0.86	-0.54	-0.55	0.32	1.45	1.27	0.37	0.03	-0.42	2.69	1.51	0.01	-0.1	0.64	-1.07	-1.08	0.7	-0.99	0.4	-0.66	0.33	0.86	1.87
4q	0.1	1.44	-0.3	-1.21	-1.34	-2.28	-2.29	-2.97	-2.58	-1.65	0.23	0.21	0.44	0.48	0.09	0.17	0.33	0.01	-0.64	-0.32	-0.36	1.99	2.15	0.05	-1.82	-1.56	-0.81	-2.01	-0.81
5p	0.09	0.55	0.07	0.27	0.44	-0.23	-0.39	-0.23	-0.21	0.03	0.86	1.5	2.11	-1.82	-1.26	3.98	1.04	-0.18	-1.28	-0.62	-2.03	0.56	1.1	-0.2	-0.15	-1.18	-0.64	-0.18	-0.51
5q	1.81	3.34	-0.03	0.44	0.32	0.46	0.79	0.53	0.86	0.19	0.67	0.01	-0.72	0.17	0.02	2.21	1.43	1.54	0.37	0.15	0.14	-1.17	0.17	-2.49	-0.1	-0.56	-1.45	0.74	0.64
6p	-2.08	-1.47	-1.4	0.13	-0.27	1.32	1.12	1.44	1.46	1.63	-0.36	-0.22	-0.99	1.26	1.16	1	1.74	0.42	0.78	0.13	0.26	2.1	1.68	-0.67	1.03	-0.33	0.2	1.04	0.85
6q	-0.43	-0.46	-0.11	0.14	0.41	1.89	1.83	1.42	1.19	0.69	0.9	0.12	-0.02	0.72	1.06	0.53	0.16	0.36	-0.13	0.64	0.14	-0.05	-0.72	0.11	0.51	0.75	0.04	0.27	-0.29
7p	0.87	1.5	0.52	-0.7	-0.33	-1.26	-1.45	-1.34	-1.51	-1.23	-0.28	-0.94	-0.81	0.17	0.52	0.89	1.26	-0.68	-0.08	0.04	0.9	1.22	-0.07	0.68	-1.73	-1.29	-1.22	-1.42	-0.93
7q	0.12	1.22	-0.6	-0.81	-0.62	-0.08	0.03	-0.47	-0.49	-1.78	1.05	0.77	1.25	0.91	0.85	1.08	-0.01	-1.04	-0.84	0.18	-0.96	0.79	1.51	-0.01	-0.28	-0.03	-0.48	0.06	-0.17
8p	1.67	1.09	0.35	-0.27	-0.33	-0.01	0.08	-0.2	-0.11	-0.01	-0.89	-0.88	-0.07	-0.82	-0.87	0.34	0.87	-1.03	-0.37	-0.7	-0.3	-2.65	-1.4	1.25	-1.65	-1.77	-1.24	-0.58	0.52
8q	-0.44	-0.01	-0.06	1	0.69	0.28	0.29	0.5	0.37	-0.06	0.25	0.03	0.52	0.22	0.1	-0.17	-0.43	0.16	0.57	-0.09	-0.42	-0.78	0.21	0.1	-0.11	-0.2	-0.61	1.24	1.6
9p	-1.51	-1.55	1.82	-0.17	-0.53	0.03	-0.62	-0.38	-0.62	0.46	1.01	2.84	3.17	1.95	1.71	-1.46	-3.32	0.99	0.55	-1.82	-3.21	1.25	0.88	0.59	-2.6	-1.8	-1.5	-2.87	-0.32
9q	1.29	1.85	-0.87	-1.43	-1.71	-1.06	-0.86	-0.62	-0.37	-0.22	-0.19	0.15	1.2	0.91	0.98	-1.93	-1.89	-0.36	-0.12	0.16	-0.28	-1.73	0.34	0.52	-0.64	-1.25	-0.81	-1.17	-1.34
10p	0.99	0.67	-0.08	0.2	-0.5	-0.12	-0.05	0.33	0.54	0.37	0.61	1.28	0.9	0.67	0.32	0.2	0.16	-0.08	-0.45	-0.57	-0.77	-2.39	-0.8	-1.17	0.35	0.14	0.26	1.1	-0.68
10q	-0.38	-0.24	-0.44	-0.69	-1.36	0.93	0.65	0.05	-0.39	-0.62	1.54	1.57	0.41	1.96	1.74	0.26	-0.38	-0.23	-0.75	0.4	-0.82	0.93	0.66	-0.27	-0.74	-1.49	-1.68	-0.14	1.45
11p	0.39	1.01	0.65	0.2	0.38	-1.97	-1.87	-1.19	-0.95	0.12	-0.13	0.32	-0.13	1.82	1.46	-0.87	-0.77	1.48	-0.16	0.13	-0.6	-1.33	0.04	-0.29	0	1.49	2.09	0.16	-0.26
11q	1.52	2.94	4.41	3.2	3.07	0.25	-0.92	-0.08	-0.74	-1.12	1.5	1.87	0.91	1.03	0.33	0.3	-1.36	-0.36	-0.72	0.69	-0.72	2.15	0.9	-1.81	-4.2	-3.27	-3.09	-4.39	-0.49
12p	0.15	1.29	-2.22	-2.78	-2.32	0.32	0.6	0.31	0.54	0.62	1.79	0.46	-0.99	-0.2	-0.55	-1.37	-2.37	-0.01	-1.16	1.9	1.68	0.78	2.21	0.03	0.51	0.3	0.5	-0.2	-0.7
12q	3.99	5.15	2.58	3.85	4.05	2.27	1.88	1.39	1.15	-0.1	3.6	3.31	-1.18	4.34	3.98	3.26	-1.03	1.4	-1.63	3.79	0.72	1.21	0.64	0.74	1.46	0.75	-1.04	1.44	-0.11
13q	-0.06	0.67	-1.21	-0.97	-1.01	-0.6	-0.31	-0.97	-0.69	-0.78	-0.58	-0.3	0.59	-0.29	-0.19	-1.91	-2.04	0.01	0.27	-0.1	-0.56	0.56	1.62	-0.83	1.33	0.74	0.75	1.6	0.92
14q	3.4	4.83	1.32	0.03	-0.71	0.78	0.15	-0.43	-0.78	-0.79	2.26	2.47	1.39	2.13	0.94	1.26	-0.59	-1.57	-2.4	0.17	-1.75	-1.14	-0.72	-0.49	-3.01	-3.4	-3.55	-2.33	0.39
15q	1.66	2.07	-0.67	-0.4	-0.38	-0.02	0.18	-0.51	-0.14	0.18	-0.82	-1.52	-2.47	0.75	1.15	-0.05	0.66	0.35	-0.39	1.59	2	-0.12	-0.22	0.28	1.13	0.79	1.06	1.05	0.68
16p	-0.77	0.33	1.39	-0.14	-0.35	0.81	0.59	1.1	0.88	0.39	0.4	0.62	-0.17	0.63	0.59	0.4	0.09	2.75	0.19	0.38	-0.2	-0.26	0.28	1.97	-1.94	-1.64	-1.31	-1.91	-0.2
16q	-0.46	-0.08	2.13	1.73	2.21	1.2	0.51	0.86	0.71	0.32	0.5	0.53	1.03	0.1	-0.39	-0.5	-0.61	0.21	1.33	0.33	-0.24	0.66	0.52	0.52	-1.36	-0.04	-0.29	-0.9	-0.44
17p	0.67	0.53	-0.91	-0.13	-0.52	0.28	0.55	0.36	0.42	0.11	-0.58	-0.79	-0.34	0.55	1.18	-0.82	-0.59	-1.04	-0.27	-0.81	0.23	0.32	-0.44	-0.11	0.14	-0.06	0.04	1.04	2.11
17q	0.35	1.48	-0.53	-0.21	-0.36	0.97	1.14	1.28	1.32	1.49	0.31	0.13	-0.22	-0.74	-0.31	-0.2	-0.74	0.6	0.99	0.83	0.31	0.65	1.87	-0.3	1.17	0.1	0.75	1.31	0.31
18p	-0.24	0.36	-0.79	-1.13	-1.12	1.74	1.67	1.43	1.44	0.73	-0.79	-0.73	-0.69	0.85	1.09	0.56	0.16	-0.6	-0.5	-1.11	0.38	0.89	-0.4	-0.7	-0.28	-0.93	-1.16	-0.51	1.76
18q	0.94	3.13	1.34	-0.21	-0.61	-0.16	-0.68	-0.99	-1.26	-2.09	2.25	2.15	1.38	3.36	2.89	1.35	-0.2	3.23	3.01	-0.08	-1.8	-0.26	1.58	-1.01	-2.24	-1.94	-1.81	-2.08	-0.67
19p	0.03	0.9	0.61	0.5	0.25	-0.36	-0.53	-0.29	-0.24	-0.67	0.3	0.56	0.4	-0.76	-1.58	-0.61	-0.96	1.91	1.06	-0.66	-0.15	0.38	1.25	-1.35	-0.88	-1.46	-1.29	-0.67	0
19q	0.6	1.28	0.46	0.21	0.63	-0.21	-0.75	-0.86	-1.19	-1.58	0.35	0.78	0.91	-0.37	-0.25	-0.24	-0.16	1.77	0.31	-0.2	-0.91	0.11	0.86	0.95	0.07	0.69	0.27	-0.66	-1.13
20p	-0.37	-0.49	-0.35	-0.26	-0.36	-1.13	-1.04	-1.06	-1.07	-0.96	-1.24	-0.82	-0.11	-1.09	-0.56	-0.72	-0.59	0.19	1.06	-0.78	0.26	-0.88	0.41	-0.25	0.41	0.37	0.57	-0.03	1.01
20q	1.79	2.06	0.63	-1.07	-0.56	0.03	-0.16	-0.53	-0.55	-1	0.89	0.57	-0.37	0.43	0.21	1.21	1.1	0.58	0.54	0.04	-0.43	-0.4	-0.13	1.56	-1.84	-1.31	-1.52	-1.84	-1.32
21q	2.15	2.51	0.41	-0.24	-0.23	-0.65	-0.39	-0.97	-0.75	-1.08	1.73	2.01	0.6	1.94	1.62	1	-0.79	-0.61	-0.2	1.22	-0.45	-0.99	-0.27	-0.96	-1.31	-2.01	-1.4	-0.89	0.5
22q	1.76	1.24	1.67	0.77	1.05	-0.15	-0.94	0.06	-1.03	1.26	0.77	0.99	0.56	-0.35	-0.52	0.27	-1.47	1.44	0.96	0.55	-0.25	-1.02	-1.37	-0.07	-0.73	-0.45	-0.2	-0.43	0.29
Any	**-1.69**	**1.72**	**2.21**	**0.94**	**0.37**	**0.45**	**-0.29**	**-0.8**	**-1.1**	**-1.68**	**4.23**	**4.48**	**2.61**	**3.8**	**2.42**	**0.62**	**-3.1**	**1.7**	**-1.2**	**1.81**	**-2.76**	**1.01**	**4.44**	**-1.49**	**-2.95**	**-3.75**	**-3.33**	**-2.06**	**0.77**

Extended Data Fig. 9 | Tendencies of CN-LOH mutations to modify polygenic scores for 29 blood cell parameters. For each blood count parameter and each chromosome arm, the heat map reports the *z*-score for the mean change in polygenic score across all CN-LOH mutations detected on the arm. Among the 29 blood count parameters we considered, some of the parameters corresponding to abundances of blood cell types might be surrogates for enhanced cellular fitness (in many cases of mitotic progenitors rather than the cell types themselves). Other parameters reflect cell size or morphology. Effects of CN-LOH mutations on polygenic scores for these parameters may reflect the production of abnormal cells by biologically altered stem cells, rather than cellular fitness itself (which may be a property of the unobserved haematopoietic stem cells). Columns: platelet count and crit (PLT#, Pct); red blood cell count (RBC#), haemoglobin (Hgb) and haematocrit (Hct) (both strongly correlated with red blood cell count); reticulocyte count and percentage (RET#, RET%); high light scatter reticulocyte count and percent (HLR#, HLR%); immature reticulocyte fraction (IRF); white blood cell count (WBC#); neutrophil count and percentage (NEU#, NEU%); eosinophil count and percentage (EOS#, EOS%); monocyte count and percentage (MON#, MON%); basophil count and percentage (BAS#, BAS%); lymphocyte count and percentage (LYM#, LYM%); platelet distribution width (PDW), mean platelet volume (MPV), RBC distribution width (RDW), mean corpuscular volume (MCV), mean reticulocyte volume (MRV), mean sphered cell volume (MSCV), mean corpuscular haemoglobin (MCH) and mean corpuscular haemoglobin concentration (MCHC).

Extended Data Table 1 | Associations of mosaic CN-LOH mutations with inherited rare coding or splice variants in *cis*.

							GWAS		Allelic shift in hets		
Arm	Gene	Position[a]	Variant	Effect[b]	Alleles[c]	AF[d]	*P*	OR (95% CI)	N_{REF}[e]	N_{ALT}	*P*
Novel loci at which rare variants associate with CN-LOH events in *cis*											
1q	*FH*	241675301	rs199822819	missense	G/C	0.0003	4.9×10^{-11}	28 (14–55)	1	8	0.039
8q	*NBN*	90983441	rs1187082186	frameshift	ATTTGT/A	0.0002	4.8×10^{-13}	210 (92–484)	0	6	0.031
11q	*MRE11*	94189489	rs587781384	stop gained	C/A	4×10^{-5}	5.6×10^{-10}	130 (50–338)	0	5	0.062
12q	*SH2B3*	111885310	rs72650673	missense	G/A	0.002	3.1×10^{-8}	11 (5.8–20)	1	8	0.039
Previously reported loci at which rare variants associate with CN-LOH events in *cis*											
1p	*MPL*	43804305	rs28928907	missense	G/C	0.0006	1.9×10^{-130}	142 (111–184)	70	0	1.7×10^{-21}
11q	*ATM*	108172425	rs587779844	missense	C/T	0.0001	3.5×10^{-20}	96 (52–177)	0	12	0.00049
15q	*TM2D3*	102151467	70kb del[f]	gene deletion	ref/del	0.0003	9.8×10^{-224}	555 (425–724)	2	110	2.4×10^{-30}
Additional independently associated likely causal coding or splice variants											
1p	*MPL*	43803600	rs146249964	splice donor	T/A	0.0001	2.8×10^{-23}	97 (55–171)	12	0	0.00049
		43803817	rs148434485	stop gained	C/T	2×10^{-5}	1.6×10^{-6}	128 (37–446)	2	0	0.5
		43803824	rs145714475	missense	T/C	2×10^{-5}	1.9×10^{-6}	120 (35–414)	3	0	0.25
		43803903	rs142565191	splice donor	G/A	4×10^{-5}	7.5×10^{-6}	72 (22–238)	3	0	0.25
		43804234	rs587778514	frameshift	CCT/C	1×10^{-5}	3.9×10^{-5}	199 (40–987)	2	0	0.5
		43804375	rs587778515	frameshift	CT/C	0.0002	7.0×10^{-41}	105 (68–161)	24	0	1.2×10^{-7}
		43804396	rs752453717	splice modifier	G/C	0.0003	5.8×10^{-36}	74 (48–113)	24	0	1.2×10^{-7}
		43804957	rs764904424	missense	C/G	0.0001	2.1×10^{-8}	35 (15–79)	6	0	0.031
		43805052	rs6088	missense	G/A	9×10^{-5}	8.3×10^{-10}	61 (26–141)	6	0	0.031
		43805656	rs144210383	missense	G/T	0.0001	5.3×10^{-9}	44 (19–101)	6	0	0.031
		43805713	rs121913611	missense	C/T	0.0002	3.3×10^{-28}	102 (61–171)	17	0	1.5×10^{-5}
		43812115	rs769297582	splice acceptor	G/C	2×10^{-5}	5.1×10^{-7}	199 (54–737)	3	0	0.25
		43814627	rs754859909	stop gained	G/A	7×10^{-5}	1.7×10^{-16}	126 (61–258)	9	0	0.0039
		43814729	454bp del[g]	exon 10 deletion	ref/del	0.0002	3.6×10^{-58}	153 (104–225)	31	0	9.3×10^{-10}
		43817942	rs369156948	stop gained	C/T	3×10^{-5}	4.8×10^{-8}	114 (39–333)	4	0	0.12
		43817973	rs971379181	frameshift	CG/C	3×10^{-5}	5.8×10^{-13}	240 (93–618)	6	0	0.031
8q	*NBN*	90983420	rs777460725	missense	A/C	0.0001	8.1×10^{-5}	114 (28–465)	0	2	0.5
11q	*ATM*	108127067	rs1137887	splice modifier	G/A	4×10^{-5}	9.6×10^{-6}	65 (20–214)	0	2	0.5
		108141801	rs786203054	missense	T/G	7×10^{-6}	1.2×10^{-5}	437 (73–2618)	0	2	0.5
		108155007	rs781357995	frameshift	AG/A	0.0001	3.0×10^{-9}	48 (21–111)	0	6	0.031
		108175528	rs376603775	stop gained	C/T	6×10^{-5}	2.8×10^{-5}	44 (14–143)	0	4	0.12
		108179837	rs774925473	splice modifier	A/G	8×10^{-5}	6.8×10^{-5}	33 (10–104)	0	3	0.25
		108181006	rs56399311	missense	A/G	8×10^{-5}	1.7×10^{-6}	44 (16–120)	0	4	0.12
		108201108	rs56399857	missense	T/G	0.0002	4.9×10^{-5}	18 (6.6–48)	0	4	0.12
		108202611	rs587776547[h]	inframe deletion	C...T/C	7×10^{-5}	8.5×10^{-9}	73 (29–183)	0	5	0.062
		108216545	rs587779872	missense	C/T	2×10^{-5}	3.6×10^{-11}	251 (89–706)	0	5	0.062
12q	*SH2B3*	111885295	rs148636776	missense	G/A	0.0004	4.0×10^{-5}	19 (7–50)	0	5	0.062
15q	*TM2D3*	102182739	rs113189685	missense	G/T	3×10^{-5}	2.8×10^{-8}	132 (45–389)	1	3	0.62
		102182749	rs754640606	missense	G/C	5×10^{-5}	1.2×10^{-40}	544 (289–1025)	0	19	3.8×10^{-6}
		102182761	rs976377433	missense	A/G	3×10^{-5}	2.3×10^{-8}	140 (47–413)	0	4	0.12
		102190214	rs768556490	frameshift	G/GT	3×10^{-5}	8.2×10^{-29}	758 (327–1759)	1	11	0.0063

P values from two independent statistical tests are reported: (i) a two-sided Fisher's exact test treating individuals with a mosaic CN-LOH mutation in *cis* as cases ($n \geq 378{,}307$ individuals varying slightly among tests; Supplementary Table 6); and (ii) a binomial test for biased allelic imbalance in heterozygous cases. Loci reaching genome-wide significance in the first test are reported. At these loci, additional independently associated coding or splice variants reaching Bonferroni significance are also reported.

[a] Base pair position in hg19 coordinates.

[b] Variant effects (using evidence reported in ClinVar[21] for splice variants).

[c] Reference or alternative allele.

[d] Alternative allele frequency (in UK Biobank individuals of European ancestry).

[e] Number of mosaic individuals heterozygous for the variant in which the somatic event shifted the allelic balance in favour of the reference allele (by duplication of its chromosomal segment and loss of the homologous segment).

[f] This approximately 70-kb deletion spans 15:102.15–102.22Mb, deleting *TM2D3* and part of *TARS3*[9].

[g] This 454-bp deletion spans 1:43,814,729—43,815,182, deleting *MPL* exon 10 (Extended Data Fig. 5).

[h] This 9-base-pair inframe deletion in *ATM* has alleles CTCTAGAATT/C.

nature research

Corresponding author(s): Po-Ru Loh, Giulio Genovese, Steven A. McCarroll

Last updated by author(s): Mar 14, 2020

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
□	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☒	□	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
□	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
□	☒	A description of all covariates tested
□	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
□	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
□	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	□	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	□	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
□	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	Data was previously collected by UK Biobank.
Data analysis	Custom code used to identify mosaic chromosomal alterations was previously described in Loh*, Genovese* et al. 2018 Nature. To facilitate similar analyses of other data sets, a standalone software implementation of the approach (MoChA) is available at https://github.com/freeseek/mocha. The following open-source software packages were also used: Eagle (v2.3.5), BOLT-LMM (v2.3.2), Minimac3 (v2.0.1), Minimac4 (v1.0.1), plink (v1.9), IGV (v2.7.2), mosdepth (v0.2.5), GERMLINE (v1.5.2).

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

Mosaic event calls are available in Supplementary Data in anonymized form. The mCA call set has also been returned to UK Biobank (as Return 2062) to enable individual-level linkage to approved UK Biobank applications. Access to the UK Biobank Resource is available by application (http://www.ukbiobank.ac.uk/).

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences □ Behavioural & social sciences □ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	We analyzed all available UK Biobank samples.
Data exclusions	We excluded ~1% of samples that failed quality control criteria we previously established in our analysis of the UK Biobank interim data. Specifically, we excluded 427 individuals with low genotyping quality (B-allele frequency s.d.>0.11 at heterozygous sites), 4,111 individuals with evidence of possible sample contamination, and 82 individuals who had withdrawn consent.
Replication	N/A; we did not attempt experimental replication.
Randomization	N/A; we analyzed all samples together in one group.
Blinding	N/A; data was previously collected by UK Biobank, and we analyzed all samples together in one group.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study	
☒	□	Antibodies
☒	□	Eukaryotic cell lines
☒	□	Palaeontology
☒	□	Animals and other organisms
☒	□	Human research participants
☒	□	Clinical data

Methods

n/a	Involved in the study	
☒	□	ChIP-seq
☒	□	Flow cytometry
☒	□	MRI-based neuroimaging