Article

Chromosomal alterations among age-related haematopoietic clones in Japan

https://doi.org/10.1038/s41586-020-2426-2

Received: 1 June 2019

Accepted: 2 April 2020

Published online: 24 June 2020

Check for updates

Chikashi Terao[1,2,3]✉, Akari Suzuki[4], Yukihide Momozawa[5], Masato Akiyama[1,6], Kazuyoshi Ishigaki[1], Kazuhiko Yamamoto[4], Koichi Matsuda[7,8], Yoshinori Murakami[9], Steven A. McCarroll[10,11,12], Michiaki Kubo[5], Po-Ru Loh[10,13,15] & Yoichiro Kamatani[1,14,15]✉

The extent to which the biology of oncogenesis and ageing are shaped by factors that distinguish human populations is unknown. Haematopoietic clones with acquired mutations become common with advancing age and can lead to blood cancers[1–10]. Here we describe shared and population-specific patterns of genomic mutations and clonal selection in haematopoietic cells on the basis of 33,250 autosomal mosaic chromosomal alterations that we detected in 179,417 Japanese participants in the BioBank Japan cohort and compared with analogous data from the UK Biobank. In this long-lived Japanese population, mosaic chromosomal alterations were detected in more than 35.0% (s.e.m., 1.4%) of individuals older than 90 years, which suggests that such clones trend towards inevitability with advancing age. Japanese and European individuals exhibited key differences in the genomic locations of mutations in their respective haematopoietic clones; these differences predicted the relative rates of chronic lymphocytic leukaemia (which is more common among European individuals) and T cell leukaemia (which is more common among Japanese individuals) in these populations. Three different mutational precursors of chronic lymphocytic leukaemia (including trisomy 12, loss of chromosomes 13q and 13q, and copy-neutral loss of heterozygosity) were between two and six times less common among Japanese individuals, which suggests that the Japanese and European populations differ in selective pressures on clones long before the development of clinically apparent chronic lymphocytic leukaemia. Japanese and British populations also exhibited very different rates of clones that arose from B and T cell lineages, which predicted the relative rates of B and T cell cancers in these populations. We identified six previously undescribed loci at which inherited variants predispose to mosaic chromosomal alterations that duplicate or remove the inherited risk alleles, including large-effect rare variants at *NBN*, *MRE11* and *CTU2* (odds ratio, 28–91). We suggest that selective pressures on clones are modulated by factors that are specific to human populations. Further genomic characterization of clonal selection and cancer in populations from around the world is therefore warranted.

Clonal expansions of blood cells with genomic alterations commonly occur in older individuals and confer an increased risk of haematological malignancies and overall mortality[1–10]. Clones can contain diverse mutations—ranging from point mutations to the gains or losses of whole chromosomes—on every chromosome.

Although populations can differ greatly in their rates of various cancers, the genomic landscape of mosaicism in the absence of known cancer remains, to our knowledge, unexplored outside of European-ancestry cohorts[10–13].

Mosaic chromosomal alterations in Japan

We searched for mosaic chromosomal alterations (mCAs) in blood-derived DNA microarray data from 179,417 participants of the BioBank Japan (BBJ) cohort, which recruited patients with 47 diseases[14] (including 13 cancers found in 16.7% of participants) (Methods). We found mCAs by analysing allele-specific hybridization intensities for 515,355 genotyped autosomal single-nucleotide polymorphisms (SNPs) (Supplementary Table 1). We analysed these data using a recently

*A list of affiliations appears at the end of the paper

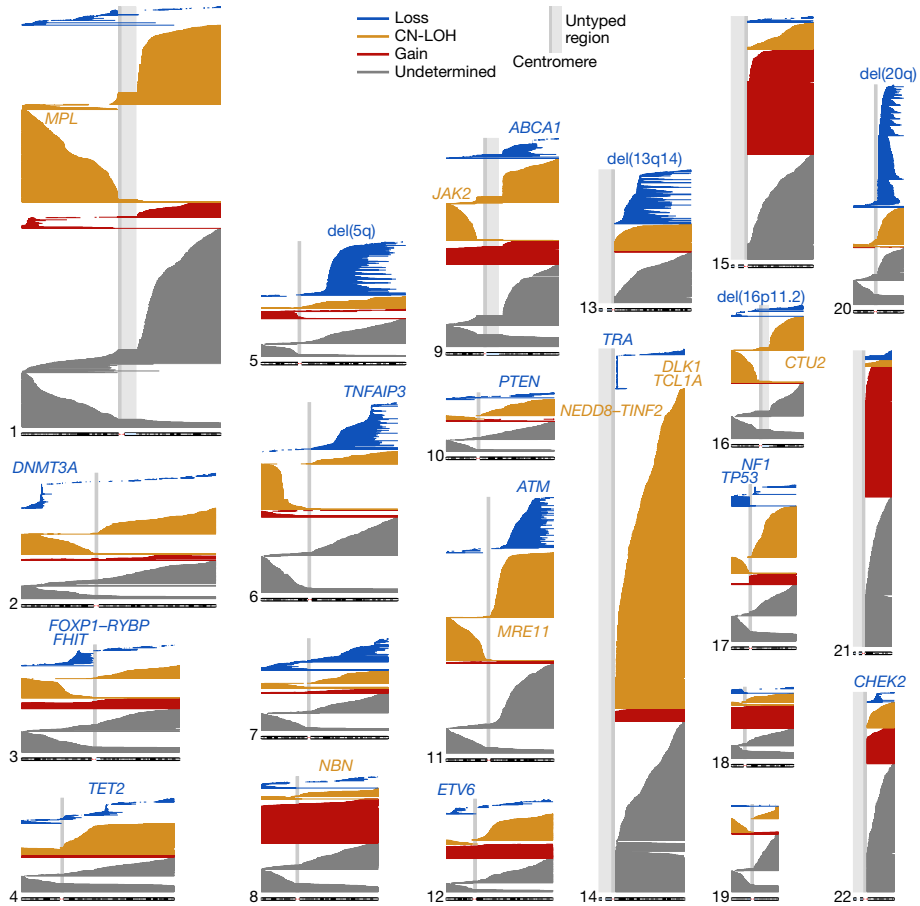


Fig. 1 | Genomic locations of 33,250 autosomal mCAs detected in 27,910 unique BBJ participants. Loss, CN-LOH and gain events are plotted as blue, orange and red horizontal lines, respectively. Events with undetermined copy numbers are plotted in grey. Commonly deleted regions are labelled in blue; loci associated with CN-LOH mutations in *cis* are labelled in orange.

developed approach that detects an imbalance in the abundance of two inherited haplotypes of an individual by using the long-range haplotype phase information that can be inferred from large population samples[10] (Methods and Supplementary Note 1).

This analysis detected 33,250 autosomal mCAs (at a false-discovery rate (FDR) of 0.05) in 27,910 unique individuals (Fig. 1 and Supplementary Note 2). The high rate of events, relative to a contemporaneous analysis of 482,789 participants in the UK Biobank (UKB), reflects the fact that BBJ participants were older (mean age at enrolment, 62.8 years of age in the BBJ compared with 57 years of age in the UKB; s.d., 14.5 years; range, 0–113 years) and a larger fraction of participants was male (54.1% in the BBJ compared with 45.8% in the UKB) and the use of different genotyping arrays[14]. Of these mutations, 5,233 were confidently classified as mosaic deletions, 10,431 as copy-neutral loss of heterozygosity (CN-LOH) events and 4,044 as duplications (Fig. 2a and Supplementary Table 2); the remaining 13,452 events were present at cell fractions that were too low or spanned too few genotyping probes to confidently determine the copy number (Supplementary Note 1). A total of 4,156 individuals had two or more non-overlapping mCAs (Supplementary Table 3); analysis of mosaic cell fractions suggested that these events were usually present in distinct clones (Supplementary Note 3). The mCA-detection rate was broadly consistent across genotyping arrays (Supplementary Table 4) and across cases of the 47 diseases that were systematically surveyed by the BBJ (Supplementary Table 5); mCAs were strongly enriched (odds ratio (95% confidence interval), 1.93 (1.66–2.25); $P = 2.1 \times 10^{-17}$) among individuals with haematological cancers at registry (that is, at the time of DNA sampling) as expected from previous studies[1,2,10] (Supplementary Note 4), but not among individuals with other illnesses.

Inevitability of mCAs in elderly individuals

The long-lived Japanese population revealed that clonal haematopoiesis with mCAs becomes extremely common in very old individuals: detectable mosaicism reached 40.7% (s.e.m., 2.3%) in men and 31.5% (s.e.m., 1.7%) in women over the age of 90 (Fig. 2b and Supplementary Table 6), which suggests that mCAs are inevitable in elderly individuals (Supplementary Note 4). mCAs on different chromosomes and with different copy-number changes exhibited various degrees of enrichment in men and in elderly individuals (Extended Data Fig. 1, Supplementary Tables 7, 8 and Supplementary Note 4) and in individuals with anomalous blood counts (Supplementary Table 9); this suggests that a spectrum of biological processes is involved in the development of different clones.

Population differences in mCA distributions

To compare the genomic distributions of mCAs in the Japanese and British populations, we co-analysed BBJ mCAs together with 19,632 autosomal mCAs detected in a parallel study[15] in the UKB cohort[16,17] (Fig. 2c–e, Supplementary Note 3, 5 and Supplementary Table 10).

The Japanese individuals have a tenfold higher incidence of adult T cell leukaemias[18] and fivefold lower incidence of chronic lymphocytic leukaemia (CLL, a B cell malignancy) compared to European individuals[19,20]. Our analysis indicated that, even among people without cancer, Japanese and British populations have markedly different rates of haematopoietic clones that arise from the B and T cell lineages, as shown by deletions produced during V(D)J recombination in developing T

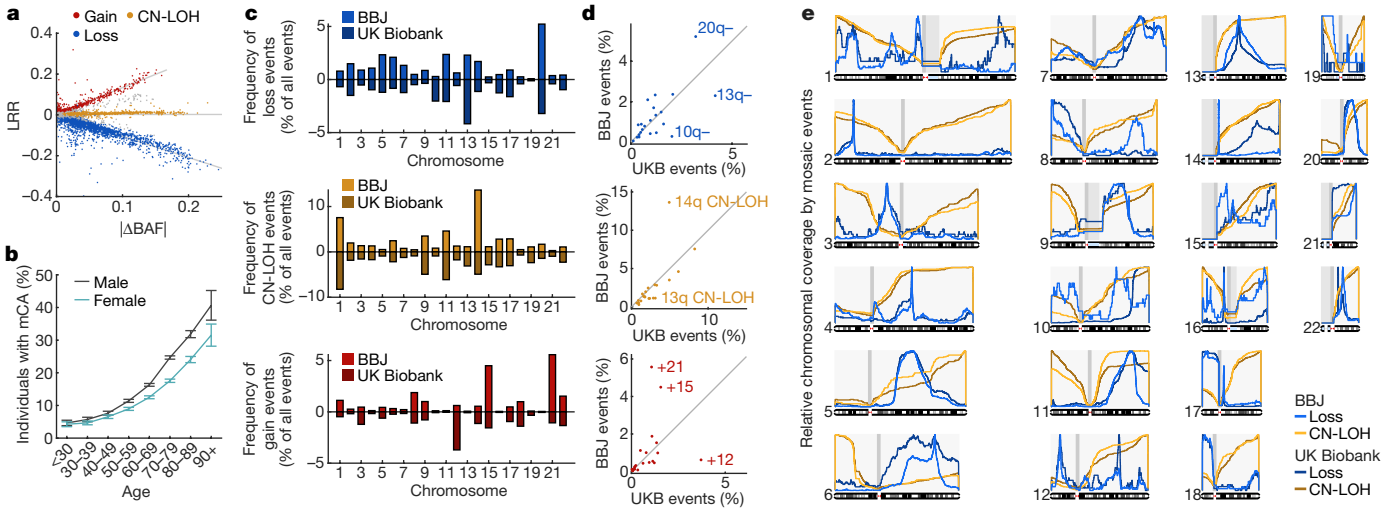


Fig. 2 | Classification of mCAs, frequency as a function of age and comparison of genomic distributions between BBJ and UKB. a, Classification of mCAs as loss, CN-LOH or gain events using log-transformed *R* ratio (LRR, measuring total DNA abundance) and B allele frequency deviation from 0.5 (|ΔBAF|, measuring allelic imbalance) (Methods). Unclassified events are indicated in grey. **b**, Frequency of detectable mosaicism stratified by age and sex. Frequencies (means) and error bars for 95% confidence intervals are indicated for the 179,417 participants analysed. **c**, **d**, Distribution of mCAs by chromosome (**c**) and copy number (**d**) in BBJ and UKB. **e**, Chromosomal coverage of loss and CN-LOH events in BBJ and UKB. Curves indicate the frequencies at which each chromosomal position is contained in loss or CN-LOH events, normalized to 1 on each chromosome. Numeric data are provided in Supplementary Tables 6, 13.

and B lymphocytes that thus identify clonal expansions in the T and B cell lineages. Mosaic deletions at the *TRA* locus on chromosome 14q (indicating clonal expansion in the T cell lineage) (Supplementary Note 5) were common in the BBJ but rare in the UKB dataset (82% versus 11% of loss events on chromosome 14 in the BBJ and UKB datasets, respectively); by contrast, deletions at the *IGH* and *IGL* immunoglobulin loci (indicating clonal expansion in the B cell lineage) were common in the UKB but rare in the BBJ dataset (5% versus 39% of loss events on chromosome 14 and 2% versus 58% of loss events on chromosome 22 in the BBJ and UKB datasets, respectively) (Fig. 2e and Supplementary Note 5). We verified that these differences did not arise from differences in genomic coverage by the genotyping arrays used by the BBJ and UKB projects (Extended Data Fig. 2). Clones that arose from the T cell lineage (as shown by deletions at *TRA*) were also associated with increased lymphocyte counts (Supplementary Tables 11, 12). Therefore, the differences in rates of B and T cell malignancies between Japanese and British populations seem to be preceded by distinct relative rates of subclinical clonal expansions in these lineages.

mCAs affect the various human chromosomes at different frequencies. The frequency of CN-LOH varied across chromosome arms in a way that strongly correlated between BBJ and UKB data ($R = 0.73$, $P = 0.00013$), with the exception of chromosomes 14q (more common in BBJ) and 13q (more common in UKB) (Fig. 2c, d, Extended Data Fig. 3 and Supplementary Table 13). By contrast, the most common loss and gain events in each population (including loss of 20q, 13q and 10q events and gains on chromosomes 21, 15 and 12) tended to be much more common in one population than the other (Fig. 2c, d and Supplementary Table 13).

A clear pattern among the most strongly population-differentiated mutations involved the two- to sixfold lower frequency in the BBJ dataset of chromosome 12 gain, 13q loss and 13q CN-LOH events (Fig. 2d): all three mutations are commonly observed in CLL[21,22] and in individuals who later develop CLL[10]. Considering the 4–5-times lower incidence of CLL in East Asian individuals, the observation that all three of these precursor mutations are also less common in Japanese haematopoietic clones that have expanded to detectable cell fractions suggests that this population difference in CLL risk originates in a reduced selective advantage for clones with (diverse) CLL precursor mutations. Consistent with this hypothesis, we observed that clonal sizes for these events tended to be lower in BBJ than in UKB participants (Supplementary Note 3).

The subchromosomal distributions of mosaic deletion events were broadly similar between BBJ and UKB participants but exhibited a few notable differences (Fig. 2e, Supplementary Table 14 and Supplementary Note 5). Focal deletions frequently targeted *DNMT3A*, *TET2*, *ETV6*, *NF1* and *CHEK2*, as shown in UKB and previous studies[1,2,5,6,10] (Figs. 1, 2e). Notably, the CLL-related deletion region at 13q14 was less focal in BBJ than in UKB data (Fig. 2e), involving longer deletions in a pattern more similar to the chromosome 20q, 5q and 11q deletion regions. We also observed previously undescribed focal deletion regions in the BBJ dataset: at *FHIT* on chromosome 3p, *TNFAIP3* on chromosome 6q, *ABCA1* on chromosome 9q and *PTEN* on chromosome 10q (Fig. 2e and Supplementary Tables 15, 16); *FHIT*, *TNFAIP3* and *PTEN* are known tumour-suppressor genes associated with blood cancers[23–25].

Inherited risk variants for mCAs in *cis*

Recent studies have established an inherited component of clonal haematopoiesis that involves both common variants that slightly increase risk (of clones with any mutation)[11–13,26] and rare variants that strongly predispose to developing clones with specific mCAs in *cis*[10]. The large number of mCAs detected among the Japanese population, together with the presence of distinct low-frequency alleles in Japan, could enable the detection of additional risk loci. To identify inherited variants associated with mCAs, we first performed association tests aimed at detecting CN-LOH events in *cis* that promoted clonal expansion by making risk alleles homozygous or removing them from the genome[10] (the two-hit model[27]). We tested variants imputed into the BBJ dataset using the 1000 Genomes phase 3 reference panel[28] together with 1,037 sequenced Japanese samples[29], setting a significance threshold of $P < 5 \times 10^{-9}$ (Methods). We further performed binomial tests to determine whether each risk allele was consistently duplicated or removed by CN-LOH events (in individuals heterozygous for the risk allele) (Methods).

We identified five new loci at which inherited variants associated with mosaic CN-LOH events in *cis* (we also replicated previously reported

Table 1 | Genome-wide significant associations between inherited variants and mosaic chromosomal alterations

Mosaic event	Locus	Chr.	Position	Variant	REF	ALT	AF	*P*	OR (95% CI)	Allelic imbalance in HET	
										n_{REF}:n_{ALT}	*P*
Novel associations with mCAs in *cis*											
8q CN-LOH	*NBN*	8	90949282	rs756831345	C	A	0.00061	9.8×10^{-23}	91 (52–159)	0:14	0.00012
11q CN-LOH	*MRE11*	11	94160189	11:94160189	G	A	0.00011	2.6×10^{-9}	37 (17–84)	0:7	0.016
14q CN-LOH	*NEDD8–TINF2*	14	24711798	rs28372734	C	G	0.073	1.0×10^{-11}	1.62 (1.42–1.85)	58:176	5.2×10^{-15}
14q CN-LOH	*TCL1A*	14	96180242	rs1122138	C	A	0.05	0.015	0.88 (0.79–0.98)	231:107	1.3×10^{-11}
14q CN-LOH	*DLK1*	14	101175967	rs10873520	G	A	0.30	7.1×10^{-39}	1.38 (1.31–1.44)	1,121:689	2.5×10^{-24}
16q CN-LOH	*CTU2*	16	88781475	rs200779411	C	T	0.00065	7.3×10^{-20}	28 (17–45)	2:11	0.022
Previously reported loci associated with mCAs in *cis*											
1p CN-LOH	*MPL*	1	45444734	rs560932816	G	A	0.00016	5.3×10^{-18}	54 (30–100)	14:0	0.00012
		1	44074454	rs190159566	C	T	0.0049	2.5×10^{-13}	4.8 (3.4–6.7)	29:6	0.00012
		1	47704269	rs556241419	G	A	0.000062	2.2×10^{-11}	81 (34–191)	7:0	0.016
		1	44579360	rs184778092	C	T	0.00005	5.0×10^{-9}	53 (22–128)	6:0	0.031
9p CN-LOH	*JAK2*	9	5026293	rs2183137	A	G	0.24	1.1×10^{-11}	1.68 (1.46–1.95)	54:155	1.7×10^{-12}
Novel associations with mCAs in *trans*											
14q CN-LOH	*TERT*	5	1287194	rs2853677	A	G	0.31	1.5×10^{-22}	1.27 (1.21–1.33)	NA	NA
Chr. 15 gain	*MAD1L1*	7	1975624	rs12699483	C	G	0.42	6.9×10^{-23}	1.61 (1.46–1.77)	NA	NA

Chr., chromosome; position, base-pair position (hg19); REF, reference allele; ALT, alternative allele; AF, allele frequency in controls; *P*, association *P* value; OR, odds ratio (using the alternative allele as the effect allele); 95% CI, 95% confidence interval; n_{REF} and n_{ALT}, number of individuals heterozygous for the variant in which a mosaic CN-LOH event in *cis* made the reference or alternative allele homozygous and removed the other allele. At the *MPL* locus, *P* values for the second, third and fourth SNPs are from stepwise conditional analysis (Methods); the four risk haplotypes are present in disjoint sets of carriers of chromosome 1p CN-LOH. Analysis results using 173,599 individuals are shown. Individuals carrying other types of mCAs on the same chromosome are excluded from controls (Methods). Fisher's exact tests (two-sided) were used for the associations and binomial tests (two-sided) were used for allelic imbalance associations in heterozygote individuals. NA, not applicable.

associations at *JAK2*[30–32] and *MPL*[10]) (Table 1, Extended Data Figs. 4–6 and Supplementary Note 6). Three of the new loci—*NBN*, *MRE11* and *CTU2*—involved rare variants with large effects. At *NBN*, the rare stop-gained variant rs756831345 on chromosome 8q associated strongly (odds ratio, 91 (52–159); $P = 9.8\times10^{-23}$) with chromosome 8q CN-LOH events, which consistently made the *NBN* risk allele homozygous ($P = 0.00012$) (Table 1 and Extended Data Figs. 5, 6). At *MRE11*, a very rare intronic variant (probably tagging a different causal variant) (Supplementary Note 7 and Supplementary Table 17) on chromosome 11q associated strongly (odds ratio, 37 (17–84); $P = 2.6\times10^{-9}$) with chromosome 11q CN-LOH events, which always made the *MRE11* risk allele homozygous ($P = 0.016$) (Table 1 and Extended Data Figs. 5, 6). Consistent with the strong proliferative advantage of these clones, we observed that these rare risk alleles further associated with the detection of multiple CN-LOH clones on the same chromosome arm (with different proximal breakpoints) (Extended Data Fig. 7, Extended Data Table 1 and Supplementary Notes 3, 6). *NBN*, *MRE11* and *RAD50* (which did not exhibit a similar association) (Supplementary Table 18) encode the components of the MRN double-strand break-repair complex, which recruits ATM in response to DNA damage, leading to the phosphorylation of p53 and CHK2 and the initiation of cell-cycle arrest, apoptosis or DNA repair[33]. Together with the observations of focal deletions at *ATM*, *TP53* and *CHEK2* (Fig. 1) and rare *ATM* risk alleles for CN-LOH events in *cis*[10], these results indicate a key role of DNA damage-response dysfunction in clonal selection.

At *CTU2*, the rare missense variant rs200779411 associated strongly ($P = 7.3\times10^{-20}$; odds ratio, 28 (17–45)) with chromosome 16q CN-LOH events, which consistently made the *CTU2* risk allele homozygous ($P = 0.022$) (Table 1 and Extended Data Figs. 5, 6). *CTU2* encodes a component of the cytosolic thiouridylase complex, which is required for maintenance of genome integrity[34]. The missense variant rs200779411 was predicted to be probably damaging by PolyPhen-2[35] and deleterious by SIFT[36], suggesting that impaired CTU2 function may promote clonal expansion by reducing genome stability.

Inherited risk variants for mCAs in *trans*

To additionally detect inherited variants associated with mCAs in *trans*, we performed genome-wide association tests on each mCA type (classifying events by chromosome and copy number), setting a genome-wide significance threshold of $P < 5.7\times10^{-11}$ to account for multiple hypotheses tested (Methods). Two *trans* associations reached significance: common variants in *MAD1L1* associated with gains on chromosome 15 and common variants in *TERT* (previously associated with mosaic *JAK2*^{V617F} mutation[12]) associated with chromosome 14q CN-LOH (Table 1, Extended Data Figs. 4, 5 and Supplementary Note 6). At *MAD1L1*, a cluster of five SNPs in near-perfect linkage disequilibrium (including the missense variant rs1801368) associated ($P = 6.9\times10^{-23}$; odds ratio, 1.61 (1.46–1.77)) (Table 1 and Extended Data Fig. 5d) with chromosome 15 gain events (mostly full trisomies) (Fig. 1). We replicated this association in the UKB cohort with a slightly reduced effect size ($P = 5.1\times10^{-4}$; odds ratio, 1.40 (1.16–1.69) for rs1801368). *MAD1L1* encodes a component of the mitotic-spindle assembly checkpoint that ensures proper chromosome segregation[37]. The *MAD1L1* risk allele was also previously observed to increase risk of mosaic Y chromosome loss[13], which is consistent with a mechanism that involves the mis-segregation of chromosomes during mitosis owing to the impaired function of the spindle assembly checkpoint. Lending further support to this hypothesis, the risk haplotype was estimated to also increase risk for large (arm-level or whole-chromosome) gain events in 9 out of 10 chromosomes with at least 50 such events (binomial $P = 0.02$) (Supplementary Table 19).

Population-specific mCA risk alleles

A comparison of the mCA risk loci detected in the BBJ dataset with previously reported loci from the UKB dataset[10] revealed ways in which genetic background can differentially shape clonal haematopoiesis in different populations (Supplementary Note 6 and Supplementary

Tables 20–23). Four of the risk variants that we found in BBJ participants (at *NBN*, *MRE11*, *NEDD8–TINF2* and *CTU2*) were present at much lower allele frequencies in European individuals[38] (Supplementary Table 20). Conversely, all rare variants that were previously associated with mCAs in UKB participants (at *MPL*, *FRA10B*, *ATM* and *TM2D3–TARSL2*)[10] were absent from Japanese individuals in the whole-genome sequencing imputation panels, with the absence of *FRA10B* fragile alleles explaining the lack of 10q25.2-qter deletions in BBJ participants (Fig. 1). Notably, *MPL* variants were associated with chromosome 1p CN-LOH events in both the BBJ and UKB datasets despite the fact that most risk alleles in each cohort were population-specific, which indicated that a shared path to mosaicism was initiated by different variants in different populations.

mCAs and mortality in Japan

Clonal haematopoiesis has previously been linked to poorer health outcomes, with various types of mosaic events observed to increase the risk of future blood cancers, mortality and cardiovascular disease[1–4,10,39]. To investigate the link between mCAs and mortality, we analysed mortality outcomes (including cause of death), which were available for around 72% of the cohort[40] (Methods).

We observed a nearly fivefold increase in the risk of death due to leukaemia (hazard ratio, 4.70 (3.26–6.78)) (Extended Data Fig. 8, Extended Data Table 2, Supplementary Table 24 and Supplementary Note 8). The increased risk of mortality caused by leukaemia did not appear to extend to other haematological malignancies (malignant lymphoma and multiple myeloma) (Extended Data Fig. 8 and Supplementary Table 24). We also did not observe a significantly increased risk of mortality attributable to cardiovascular disease, suggesting that previous associations of clonal haematopoiesis (which primarily involved point mutations in *DNMT3A*, *TET2*, *JAK2* and *ASXL1*) with cardiovascular outcomes[4,39] may be limited to specific mosaic events (Extended Data Fig. 8 and Supplementary Table 24). To refine the association between mosaic status and leukaemia mortality, we partitioned mosaic events by chromosome and copy-number change (Methods) and identified six mCAs with significant ($P < 0.05/88$, Cochran–Mantel–Haenszel test), large effects on leukaemia mortality risk (Extended Data Fig. 8b and Supplementary Table 25). Mosaic cell fraction and the number of mosaic events carried by an individual each associated with further increases in leukaemia mortality risk (Extended Data Fig. 8c, d and Supplementary Tables 26, 27). Mosaic status increased the risk of overall mortality (hazard ratio, 1.10 (1.05–1.16); $P = 2.7 \times 10^{-5}$) (Supplementary Table 24), an association that was driven by mCAs in chromosomes 9 and 14 (hazard ratio > 1.4) (Supplementary Table 28), underscoring the heterogeneity in the clinical effects of different mCAs.

Discussion

Our study of haematopoietic clones with mCAs in Japan provides a detailed comparison of the genomic landscape of clonal haematopoiesis between populations, revealing broad overall similarities as well as important population differences. A clear pattern among these results showed that population differences in blood cancer rates are preceded by population differences in subclinical clonal expansions, at multiple levels: both in specific cell lineages (including B and T cell lineages) and with specific cancer-associated mutations (for example, gain of chromosome 12, loss of chromosome 13q and chromosome 13q CN-LOH, which are hallmarks of CLL). These results point towards population-specific differences in the clonal advantages that are gained by the same chromosomal mutations in different genetic and environmental contexts.

The interplay between acquired and inherited genetic variation in Japan enabled further insights into the influences of inherited variation on clonal haematopoiesis: population-specific variants at several risk loci pointed to a key role of the maintenance of genomic integrity; with corroborating evidence from loci targeted by focal deletions. These results point to the need for larger and more diverse cohorts in genomic studies of cancer and subclinical clonal expansions as well as inherited variation.

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/s41586-020-2426-2.

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. Young, A. L., Challen, G. A., Birmann, B. M. & Druley, T. E. Clonal haematopoiesis harbouring AML-associated mutations is ubiquitous in healthy adults. *Nat. Commun.* **7**, 12484 (2016).
8. Forsberg, L. A., Gisselsson, D. & Dumanski, J. P. Mosaicism in health and disease — clones picking up speed. *Nat. Rev. Genet.* **18**, 128–142 (2017).
9. Abelson, S. et al. Prediction of acute myeloid leukaemia risk in healthy individuals. *Nature* **559**, 400–404 (2018).
10. Loh, P.-R. et al. Insights into clonal haematopoiesis from 8,342 mosaic chromosomal alterations. *Nature* **559**, 350–355 (2018).
11. Zhou, W. et al. Mosaic loss of chromosome Y is associated with common variation near *TCL1A*. *Nat. Genet.* **48**, 563–568 (2016).
12. Hinds, D. A. et al. Germ line variants predispose to both *JAK2* V617F clonal hematopoiesis and myeloproliferative neoplasms. *Blood* **128**, 1121–1128 (2016).
13. Wright, D. J. et al. Genetic variants associated with mosaic Y chromosome loss highlight cell cycle genes and overlap with cancer susceptibility. *Nat. Genet.* **49**, 674–679 (2017).
14. Nagai, A. et al. Overview of the BioBank Japan project: study design and profile. *J. Epidemiol.* **27**, S2–S8 (2017).
15. Loh, P.-R., Genovese, G. & McCarroll, S. A. Monogenic and polygenic inheritance become instruments for clonal selection. *Nature* https://doi.org/10.1038/s41586-020-2430-6 (2020).
16. Sudlow, C. et al. UK Biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Med.* **12**, e1001779 (2015).
17. Bycroft, C. et al. The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203–209 (2018).
18. Iwanaga, M., Watanabe, T. & Yamaguchi, K. Adult T-cell leukemia: a review of epidemiological evidence. *Front. Microbiol.* **3**, 322 (2012).
19. Tamura, K. et al. Chronic lymphocytic leukemia (CLL) is rare, but the proportion of T-CLL is high in Japan. *Eur. J. Haematol.* **67**, 152–157 (2001).
20. Li, Y., Wang, Y., Wang, Z., Yi, D. & Ma, S. Racial differences in three major NHL subtypes: descriptive epidemiology. *Cancer Epidemiol.* **39**, 8–13 (2015).
21. Landau, D. A. et al. Mutations driving CLL and their evolution in progression and relapse. *Nature* **526**, 525–530 (2015).
22. Puente, X. S. et al. Non-coding recurrent mutations in chronic lymphocytic leukaemia. *Nature* **526**, 519–524 (2015).
23. Iwai, M. et al. Expression and methylation status of the *FHIT* gene in acute myeloid leukemia and myelodysplastic syndrome. *Leukemia* **19**, 1367–1375 (2005).
24. Schmitz, R. et al. *TNFAIP3* (A20) is a tumor suppressor gene in Hodgkin lymphoma and primary mediastinal B cell lymphoma. *J. Exp. Med.* **206**, 981–989 (2009).
25. Liu, Y. et al. The genomic landscape of pediatric and young adult T-lineage acute lymphoblastic leukemia. *Nat. Genet.* **49**, 1211–1218 (2017).
26. Zink, F. et al. Clonal hematopoiesis, with and without candidate driver mutations, is common in the elderly. *Blood* **130**, 742–752 (2017).
27. Knudson, A. G. Jr. Mutation and cancer: statistical study of retinoblastoma. *Proc. Natl Acad. Sci. USA* **68**, 820–823 (1971).
28. The 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
29. Okada, Y. et al. Deep whole-genome sequencing reveals recent selection signatures linked to evolution and disease risk of Japanese. *Nat. Commun.* **9**, 1631 (2018).
30. Kilpivaara, O. et al. A germline *JAK2* SNP is associated with predisposition to the development of *JAK2*^{V617F}-positive myeloproliferative neoplasms. *Nat. Genet.* **41**, 455–459 (2009).
31. Jones, A. V. et al. *JAK2* haplotype is a major risk factor for the development of myeloproliferative neoplasms. *Nat. Genet.* **41**, 446–449 (2009).
32. Olcaydu, D. et al. A common *JAK2* haplotype confers susceptibility to myeloproliferative neoplasms. *Nat. Genet.* **41**, 450–454 (2009).
33. Lee, J. H. & Paull, T. T. ATM activation by DNA double-strand breaks through the Mre11–Rad50–Nbs1 complex. *Science* **308**, 551–554 (2005).

34. Dewez, M. et al. The conserved Wobble uridine tRNA thiolase Ctu1–Ctu2 is required to maintain genome integrity. *Proc. Natl Acad. Sci. USA* **105**, 5459–5464 (2008).
35. Adzhubei, I., Jordan, D. M. & Sunyaev, S. R. Predicting functional effect of human missense mutations using PolyPhen-2. *Curr. Protoc. Hum. Genet.* **76**, 7.20.1–7.20.41 (2013).
36. Sim, N. L. et al. SIFT web server: predicting effects of amino acid substitutions on proteins. *Nucleic Acids Res.* **40**, W452–W457 (2012).
37. DeAntoni, A., Sala, V. & Musacchio, A. Explaining the oligomerization properties of the spindle assembly checkpoint protein Mad2. *Phil. Trans. R. Soc. Lond. B* **360**, 637–648 (2005).
38. Karczewski, K. J. et al. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**, 434–443 (2020).
39. Jaiswal, S. et al. Clonal hematopoiesis and risk of atherosclerotic cardiovascular disease. *N. Engl. J. Med.* **377**, 111–121 (2017).
40. Hirata, M. et al. Overview of BioBank Japan follow-up data in 32 diseases. *J. Epidemiol.* **27**, S22–S28 (2017).

[1]Laboratory for Statistical and Translational Genetics, RIKEN Center for Integrative Medical Sciences, Yokohama, Japan. [2]Clinical Research Center, Shizuoka General Hospital, Shizuoka, Japan. [3]The Department of Applied Genetics, The School of Pharmaceutical Sciences, University of Shizuoka, Shizuoka, Japan. [4]Laboratory for Autoimmune Diseases, RIKEN Center for Integrative Medical Sciences, Yokohama, Japan. [5]Laboratory for Genotyping Development, RIKEN Center for Integrative Medical Sciences, Yokohama, Japan. [6]Department of Ophthalmology, Graduate School of Medical Sciences, Kyushu University, Fukuoka, Japan. [7]Laboratory of Genome Technology, Human Genome Center, Institute of Medical Science, The University of Tokyo, Tokyo, Japan. [8]Laboratory of Clinical Genome Sequencing, Department of Computational Biology and Medical Sciences, Graduate School of Frontier Sciences, The University of Tokyo, Tokyo, Japan. [9]Division of Molecular Pathology, Institute of Medical Science, The University of Tokyo, Tokyo, Japan. [10]Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, MA, USA. [11]Stanley Center for Psychiatric Research, Broad Institute of MIT and Harvard, Cambridge, MA, USA. [12]Department of Genetics, Harvard Medical School, Boston, MA, USA. [13]Division of Genetics, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, MA, USA. [14]Laboratory of Complex Trait Genomics, Department of Computational Biology and Medical Sciences, Graduate School of Frontier Sciences, The University of Tokyo, Tokyo, Japan. [15]These authors jointly supervised this work: Po-Ru Loh, Yoichiro Kamatani. ✉e-mail: chikashi.terao@riken.jp; yoichiro.kamatani@riken.jp

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.

The BBJ cohort
All of the individuals analysed in this study were participants of the BBJ project. The BBJ is a multi-hospital-based registry that collected clinical information, DNA and serum samples from approximately 200,000 patients with one or more of 47 target diseases (including 13 cancers) at a total of 66 hospitals between fiscal years 2003 and 2007[14]. The case proportions correlated well with prevalence in the Japanese population and all of the study participants were diagnosed by medical doctors as described elsewhere[14]. We complied with all relevant ethical regulations. This project was approved by the ethics committees of RIKEN Center for Integrative Medical Sciences and the Institute of Medical Sciences, the University of Tokyo. Written informed consent was obtained from all of the participants.

Genotyping individuals in the BBJ
Participants were genotyped in three batches using different arrays or set of arrays, namely: (1) a combination of Illumina Infinium Omni Express and Human Exome; (2) Infinium Omni Express Exome v.1.0; and (3) Infinium Omni Express Exome v.1.2 (Supplementary Table 1). The SNP content of the three methods was very similar. DNA was obtained from blood samples for all but one individual (for which DNA was obtained from oral mucosa; this sample was negative for mosaic events).

We excluded outliers from East Asian clusters in a plot in which we projected BBJ participants in combination with 1000 Genomes Project[41] samples in the principal component (PC)1 and PC2 space. We also excluded samples genetically identical to another sample, samples with call rates less than 0.98, and samples for which the reported sex information was not supported by genotypes in the X chromosome. We further excluded three samples with evidence of potential contamination (as suggested by low cell-fraction mosaic events called on many chromosomes[6,10]), leaving 179,417 samples for analysis. We used plink v.1.9 software[42] to handle the genotyping data.

Genotyping intensity data used for calling mosaic events
To call mosaic events, we analysed genotyping intensity data for variants in the intersection of the three primary arrays used for BBJ genotyping (namely, Illumina Infinium Omni Express and Infinium Omni Express Exome v.1.0 and v.1.2) to enable the analysis of the same set of variants in all individuals (to avoid the possibility of differing detection sensitivity across batches due to different numbers of genotyping probes analysed). When calling mosaic events, we did not include variants typed on the Human Exome array in some samples (see above) to minimize the potential for batch effects arising from different arrays. We did use variants from the Human Exome array in genetic association analyses (see below) as association tests are robust to genotyping heterogeneity when potential confounders are appropriately controlled by correcting for batch covariates and principal components.

Calculation of BAF and LRR from genotype intensity
We computed BAF and LRR values with the use of the BBJ genotyping intensity data[43]. We modified previously published methods[1,10] to fit the current dataset. We computed LRR and BAF values on a per-array basis in which all of the participants genotyped in the same arrays were clustered together. Details are provided in Supplementary Note 1.

Phasing of genotype data for calling mosaic events
We phased the filtered genotypes mentioned above with the use of Eagle2 software[44], which enabled us to conduct accurate long-range phasing. This phasing information was used for calling mosaic events (Supplementary Note 1).

Filtering possible non-mosaic trisomy or monosomy events
We excluded chromosomes with mean LRR > 0.2 or mean LRR < −0.5 (possible trisomy and monosomy, respectively) (Supplementary Note 1).

Calling mosaic events with the use of BAF and LRR
We used the same method to call mosaic events as previously described[10]. This calling method is composed of the following steps: (1) filtering constitutional duplications; (2) evaluating the phased BAF for variants on each chromosome using a parameterized hidden Markov model; (3) calling the existence of events using a likelihood ratio test; (4) calling the event boundaries; (5) calling the copy number; (6) filtering remaining possible constitutional duplications; (7) estimating the cell fraction of mosaic events. Details of each step are provided in Supplementary Note 1.

Associations between array batches or disease status at registry and detectable mosaicism
We conducted logistic regression analyses to evaluate associations between detectable mosaicism and either array batches or disease status at the time of participant recruitment (47 diseases, a binary trait for each of the diseases). For array batches (Supplementary Table 4), we put mosaic detection status as a dependent binary variable and age, sex, smoking, genotyping arrays and 10 principal components as independent variables. For disease status at registry (Supplementary Table 5), we put disease presence as a dependent variable and presence of mosaic events, age, sex, genotyping arrays and 10 principal components as independent variables.

Associations between haematological traits and mosaic events
We extracted data from the BBJ for 13 haematological traits, namely, red blood cell count, haemoglobin, haematocrit, mean corpuscular volume, mean corpuscular haemoglobin, mean corpuscular cell haemoglobin, white blood cell count, neutrophil count, lymphocyte count, monocyte count, eosinophil count, basophil count and platelet count. Associations between 13 haematological quantitative traits and the presence of 88 types of mosaic events (see below) were analysed in logistic regression models with event presence as outcomes. Before analysis in logistic models, the 13 traits were regressed out by covariates specified in the previous BBJ study[45] for men and women. Residuals were normalized and used as independent variables one by one (a total of 13 models for each mCA). In each logistic model, disease status at registry (for each of the 47 diseases in the BBJ study design), age, sex, smoking, genotyping arrays and 10 principal components were used as covariates. We took this approach to control for effects of covariates associated with both mCAs and haematological traits.

We subdivided mosaic events by copy-number state (loss, CN-LOH or gain) and by p versus q arm for loss and CN-LOH events. To reduce multiple testing burden, we restricted analyses to mosaic events with more than 20 carriers (Supplementary Table 2). As a result, 88 mosaic events were analysed in association with 13 haematological traits. The statistical significance threshold was set to $P < 0.05/88/13$ (4.4×10^{-5}); results are reported in Supplementary Table 9.

Comparison of mosaic frequency between BBJ and UKB
We co-analysed BBJ mosaic calls with mosaic calls in UKB data from 482,857 individuals[15]. We calculated the frequencies of mosaic events subdivided by chromosome arm and copy number among all mosaic events in both datasets. We assessed the correlation of event frequencies in the two datasets using Spearman's correlation coefficients.

Relative coverage of the genome by mosaic events in the BBJ and UKB

We determined mosaic coverage as follows. We divided chromosomes into 0.1-Mb bins and calculated the fraction of loss or CN-LOH events that covered each bin to compute mosaic coverage. We scaled the coverage in each mosaic type in each chromosome (to set maximum coverage as 1). We compared mosaic coverage in the BBJ and UKB datasets using Pearson's correlation coefficients.

Genomic coverage by genotyping arrays in BBJ and UKB

We computed the mean numbers of heterozygous genotyped sites across individuals in each 1-Mb region of the genome for the BBJ and UKB genotyping arrays to confirm that the difference in mosaic frequency between the two populations was not driven by different coverage of the genome by DNA microarrays.

Association between mosaic events indicating T cell expansions and lymphocyte counts

We used a Wilcoxon rank-sum test to compare lymphocyte counts between individuals who carried *TRA* deletions (indicating clonal expansions of T cells) and individuals without *TRA* deletion. We also evaluated Spearman correlations between the cell fraction of *TRA* deletions and lymphocyte counts.

Distribution of breakpoints of CN-LOH in BBJ and UKB

We computed relative frequencies of estimated CN-LOH breakpoint locations in each chromosome in BBJ and UKB. We smoothed breakpoints over ±2 Mb and rescaled to 1.

Genes affected by focal deletion

We evaluated the importance of genes by taking the numbers of genes involved in loss events into account. We counted the number of genes involved in each loss event and defined a score of each loss event as one divided by the number of genes (that is, when a loss event contained only one gene, the gene received a score of 1). We summed scores of all loss events containing each gene. To pick up genes that were frequently involved with focal deletions only in the Japanese population, we identified genes covered by at least 5% of loss events in a chromosome, having a tenfold larger score in BBJ than in UKB, and scoring more than 0.5.

Genetic association studies

We excluded participants who showed a high degree of kinship (first degree or closer as detected by plink[42]) with other individuals, leaving 173,599 participants for genetic association studies. Among related pairs, we retained individuals who had mosaic events. We also integrated the genotyping data used for calling mosaic events with genotyping data from additional variants typed on the Human Exome Array in some samples when also available on the Omni Express Exome Arrays in other samples (Supplementary Table 1) to maximize the number of variants used for imputation. We did not integrate these data at the stage of calling mosaic events to minimize the potential for batch effects. We phased the integrated data using Eagle2 software[46]. The phased genotypes were imputed using a reference panel containing 2,504 1000 Genomes phase 3 samples and 1,037 Japanese high-depth (30×) whole-genome sequencing samples (dataset 1 of a previous study[29]) using Minimac3 software[47]. Variants imputed with $R^2 > 0.3$ were used for the association studies. We filtered variants with minor allele count less than 5. Best-guess data were used to conduct Fisher's exact tests using plink software (plink --fisher --ci 0.95). We used Fisher's exact tests to prevent inflated type I errors when testing associations between rare variants and rare mosaic events[48]. To confirm that significant associations were not driven by confounding factors, we reanalysed significant associations (detected by Fisher's exact test) using logistic regression with and without covariates (10 principal components, disease status at registry, age, sex, smoking and genotype batches) and verified that the associations were robust. We used genotyping data from DNA microarrays if available to rescue rare variants that were not included in the reference panel, that were not well-imputed or that had low allele frequency. As a result, 26.6 million variants were used for association studies.

We analysed mosaic events in each chromosome as distinct phenotypes, treating loss, CN-LOH and gain separately. To maximize the power to identify significant associations with CN-LOH, we included unclassified 'likely CN-LOH' events (that is, events that extended to one telomere with $|LRR| < 0.02$) when testing variants for association with CN-LOH events. We subdivided loss and CN-LOH events in each chromosome into p-arm and q-arm events. We set a threshold of at least 20 event carriers to consider an event in genetic association studies. This led to a total of 88 copy number–chromosome pairs analysed (Supplementary Table 2). We tested each of these phenotypes for association with variants in *cis* (that is, on the same chromosome and contained within a mosaic event) or in *trans* (that is, on any chromosome). For *cis* associations, we also conducted allelic imbalance analyses to assess whether one of the alleles at each variant was preferentially duplicated by mosaic CN-LOH events. Details of each test and corresponding significance thresholds are described in Supplementary Note 6.

At significantly associated loci, we additionally performed stepwise conditional analyses (by iteratively removing carriers of high-risk rare alleles) to test for additional independently associated variants.

Associations between risk variants and presence of multiple CN-LOH clones with different breakpoints

In a small fraction of individuals, we detected evidence of multiple clonal expansions of CN-LOH events that affected the same chromosome arm but with different breakpoints. To detect such events, we applied a modified hidden Markov model as described previously[10] (Supplementary Note 3). In brief, this analysis searched for evidence of CN-LOH with increasing BAF deviation towards the telomere. We evaluated associations between the presence of risk variants found above and the presence of single or multiple breakpoints among individuals with CN-LOH spanning the variants using Fisher's exact test. Details are described in Supplementary Note 6.4.6.

Associations between mosaic events and mortality

The BBJ project has follow-up data to survey mortality and cause of death[40]. A total of 141,612 BBJ participants who have one of 32 out of 47 diseases were prospectively followed up after DNA collection. For participants who died, further detailed surveillance was carried out to identify causes of death (coded with codes from the tenth revision of the International Statistical Classification of Diseases and Related Health Problems (ICD10)) by accessing national vital registration system used for input survey of medical and social welfare at Ministry of Health, Labour and Welfare of the Japanese Government.

We restricted participants to those who were followed for at least 1 year after registry and free from malignancy at blood collection. We found 86,546 participants in the current study who were included in the follow-up data for mortality. Among them, 16,812 deaths were recorded during the follow-up period. The average follow-up period was 7.6 years (median, 8.3 years; s.d., 2.8 years).

Associations between mortality (overall or specific causes) and the presence of mosaic events (regardless of mosaic types) were analysed as an initial evaluation. We analysed overall mortality, haematopoietic malignancy mortality and non-haematopoietic malignancy mortality. We compared individuals with mosaic events (loss, CN-LOH or gain) at cell fraction ≥1% to individuals without mosaic events on any chromosomes. Cox regression analysis was used for the analyses conditioning for age, age^2, sex, disease status, genotyping array and smoking. We used follow-up period as a censoring factor. When we analysed specific causes of death (for example, non-haematopoietic malignancy),

we only used participants whose deaths were not reported during follow-up as controls to use consistent control samples across analyses. We used significance thresholds based on Bonferroni's correction (based on the number of tested mortality phenotypes).

After evaluating associations between mortality phenotypes and the presence of any mosaic event in any chromosome, we searched for associations between specific mosaic event types and mortality. We analysed the same set of 88 mosaic event types (defined by copy-number state and chromosomal location) that we used when testing associations with haematological traits and inherited variants. In these analysis of associations between mortality phenotypes and specific mosaic types, we divided participants based on age, sex and smoking status and computed associations using Cochran–Mantel–Haenszel tests to avoid inflation of statistics arising from small number of individuals who carried mosaic events. We set a significance level based on Bonferroni's correction ($P < 0.00057$, 0.05/88).

We also analysed cardiovascular mortality (defined as ischaemic heart diseases and ischaemic stroke) as previous studies have reported associations between mosaic point mutations and cardiovascular outcomes.

Definition of cancers based on ICD10 codes for causes of death

We categorized causes of death to decrease multiple testing burden. Haematopoietic malignancy was defined by ICD10 codes C81–C96 and D45, D46 and D47. Leukaemic diseases were defined by ICD10 codes C91–C96, D45 and D46. Malignant lymphoma was defined by ICD10 codes C81–C88. Multiple myeloma was defined by C90. Cancers were defined as ICD10 codes starting with 'C' together with haematopoietic malignancies defined not starting with 'C'. We did not regard other ICD10 codes starting with 'D' as cancer as most of those are benign tumours.

Associations between multiple mosaic events and mortality

We extended the mortality analyses to investigate the effect of multiple mosaic events within a single individual. We limited analyses to individuals with at most three mosaic events. We divided participants into three groups: (1) individuals without mosaic events; (2) individuals with a single mosaic event; (3) individuals with multiple mosaic events (two or three mosaic events in different chromosomes). We analysed an association of the presence of multiple mosaic events with leukaemia mortality in comparison with the presence of a single mosaic event. The analyses were conditioned on age, sex, disease status, genotyping array and smoking status.

Associations between cell fraction of mosaic events and mortality

We also extended the mortality analyses to investigate the effect of mosaic cell fraction. For individuals with multiple mosaic events, we took the highest cell fraction. We divided participants into categories according to the cell fraction of mosaic events and analysed associations between cell fractions and outcomes with which the presence of mosaic events were significantly associated.

Reporting summary

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

Data availability

A table for mosaic events detected in the current study is available as Supplementary Data 1. The BBJ genotype is available from the Japanese Genotype-phenotype Archive (JGA; http://trace.ddbj.nig.ac.jp/jga/index_e.html) under accession code JGAD00000000123. Individual-level linkage of mosaic events can be provided by the BBJ project upon request (https://biobankjp.org/english/index.html).

Code availability

All computational codes are available upon request from the corresponding authors (although they are not immediately portable to other computing environments). A standalone software implementation (MoChA) of the algorithm used to call mCAs is available at https://github.com/freeseek/mocha.

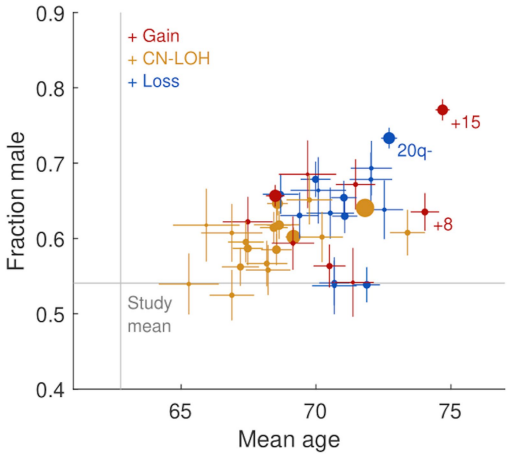
41. The 1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061–1073 (2010).
42. Purcell, S. et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).
43. Staaf, J. et al. Normalization of Illumina Infinium whole-genome SNP data improves copy number estimates and allelic intensity ratios. *BMC Bioinformatics* **9**, 409 (2008).
44. Loh, P.-R. et al. Reference-based phasing using the Haplotype Reference Consortium panel. *Nat. Genet.* **48**, 1443–1448 (2016).
45. Kanai, M. et al. Genetic analysis of quantitative traits in the Japanese population links cell types to complex human diseases. *Nat. Genet.* **50**, 390–400 (2018).
46. Loh, P.-R., Palamara, P. F. & Price, A. L. Fast and accurate long-range phasing in a UK Biobank cohort. *Nat. Genet.* **48**, 811–816 (2016).
47. Das, S. et al. Next-generation genotype imputation service and methods. *Nat. Genet.* **48**, 1284–1287 (2016).
48. Ma, C., Blackwell, T., Boehnke, M. & Scott, L. J. Recommended joint and meta-analysis strategies for case-control association testing of single low-count variants. *Genet. Epidemiol.* **37**, 539–550 (2013).

Acknowledgements We thank the staff of the BBJ for collecting and managing samples and clinical information. This study was funded by the BioBank Japan project, which was supported by the Ministry of Education, Culture, Sports, Sciences and Technology of the Japanese Government and AMED under grant numbers 17km0305002 and 18km0605001. This research was conducted using the UK Biobank Resource under application no. 19808. P.-R.L. was supported by 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.

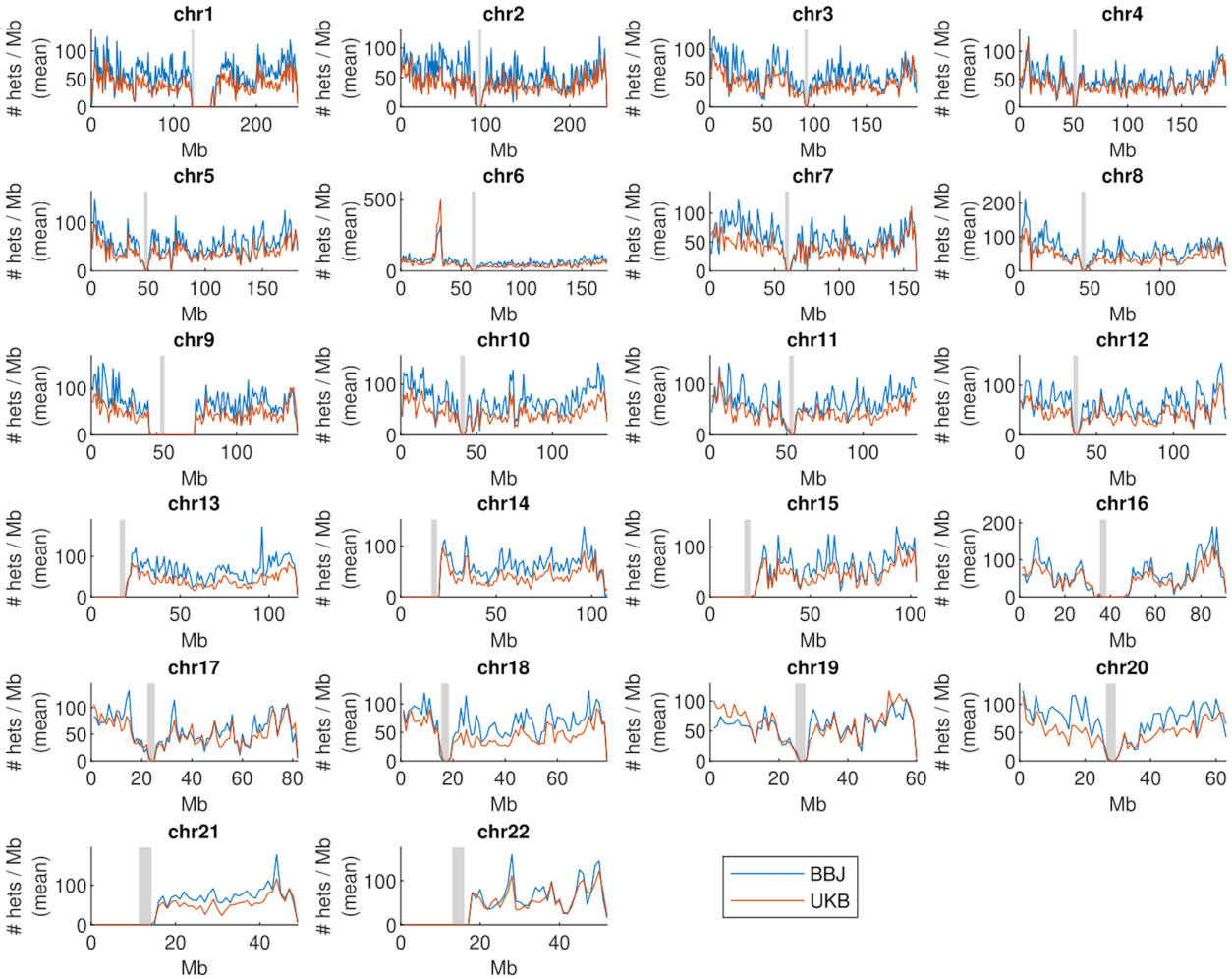
Author contributions C.T., P.-R.L. and Y.K. conceived the study design. P.-R.L. and Y.K. supervised the project. C.T. and P.-R.L. analysed the data. A.S. and K.Y. conducted functional analyses. M.A. and K.I. contributed to the construction of the Japanese reference panel for genotype imputation. Y. Momozawa, K.M., Y. Murakami and M.K. contributed to the generation of the BBJ data. C.T., S.A.M., P.-R.L. and Y.K. wrote the manuscript. All the authors critically reviewed the manuscript and approved the final version.

Competing interests The authors declare no competing interests.

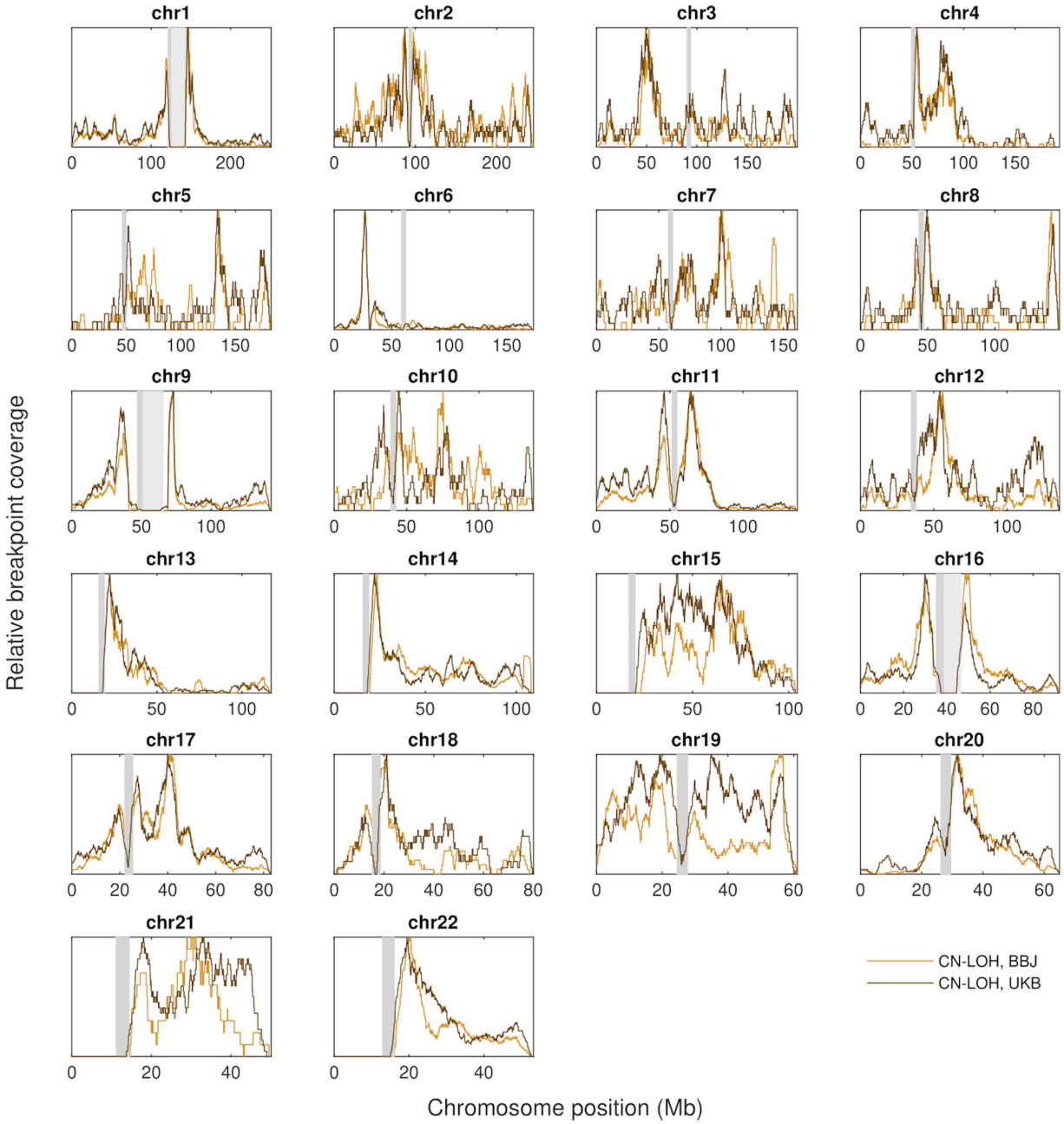
Additional information
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-020-2426-2.
Correspondence and requests for materials should be addressed to C.T. or Y.K.
Peer review information *Nature* thanks Paul Scheet, George Vassiliou and John Witte for their contribution to the peer review of this work.
Reprints and permissions information is available at http://www.nature.com/reprints.



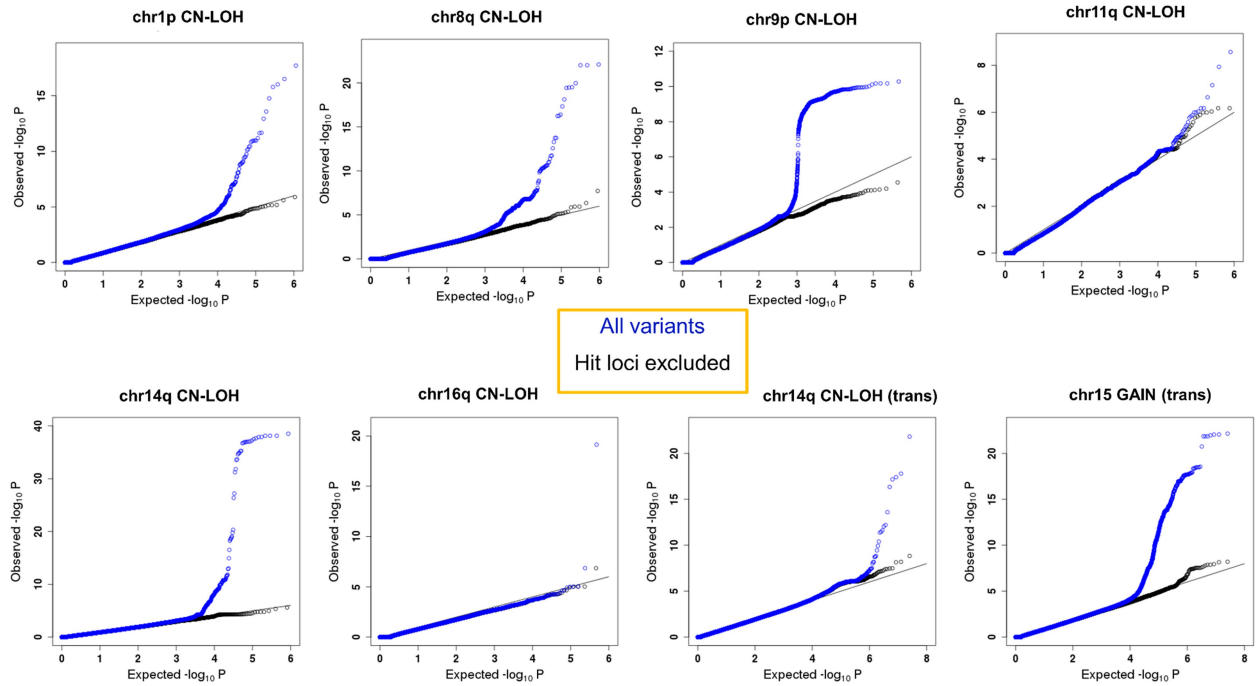
Extended Data Fig. 1 | Age and sex of the carriers of mosaic event types. Mean age and sex of carriers of specific mCA types (defined by chromosome and copy number) with at least 100 carriers in the 179,417 participants. Marker sizes are proportional to mCA frequencies. Data are mean ± s.e.m. Numeric data are provided in Supplementary Table 7.



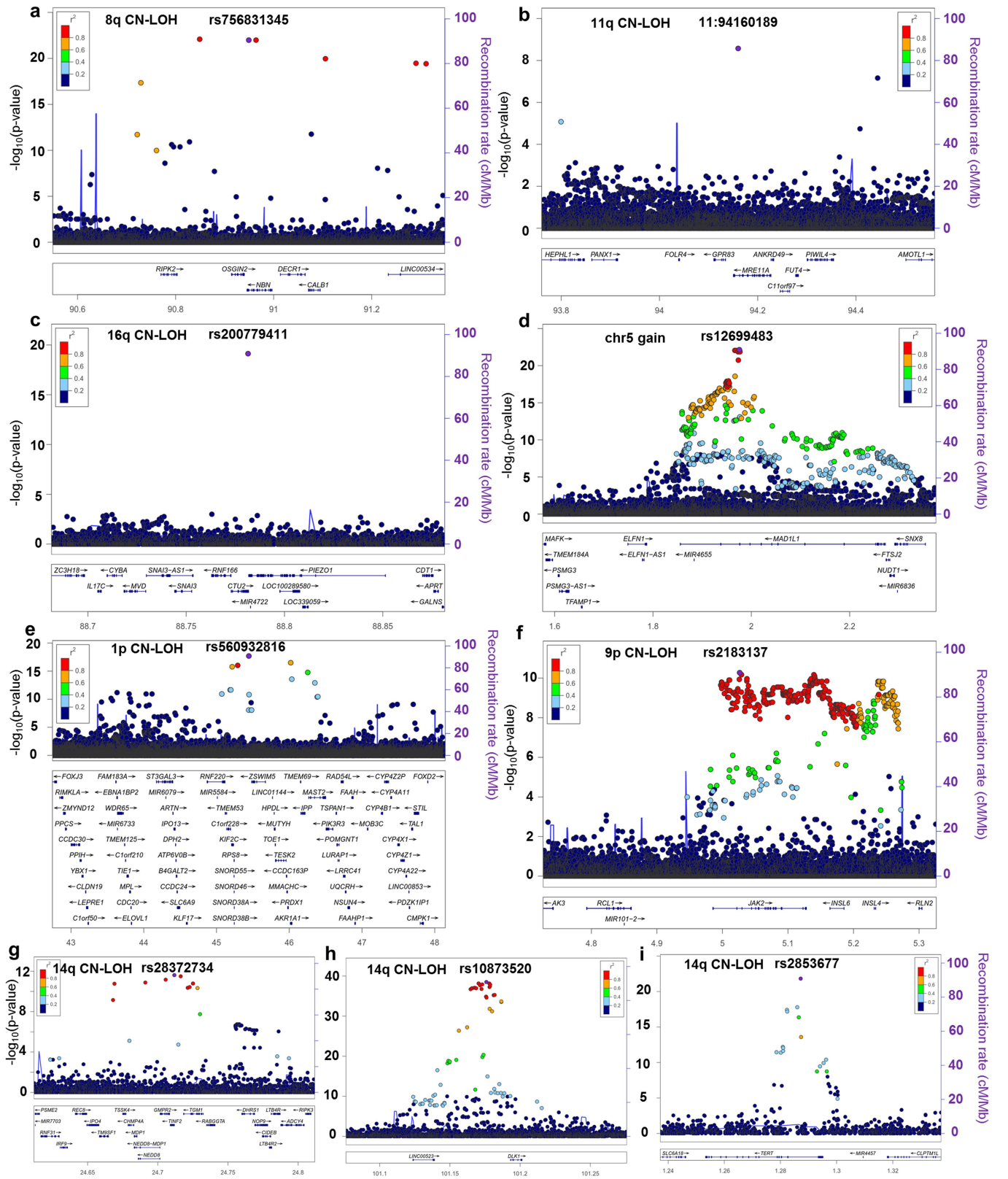
Extended Data Fig. 2 | Comparable chromosomal coverage by heterozygous genotypes in the BBJ and UKB data. Average numbers of heterozygous genotyped sites (averaged across individuals) in each 1-Mb region of the genome for the BBJ and UKB genotyping arrays. hets, heterozygous sites.



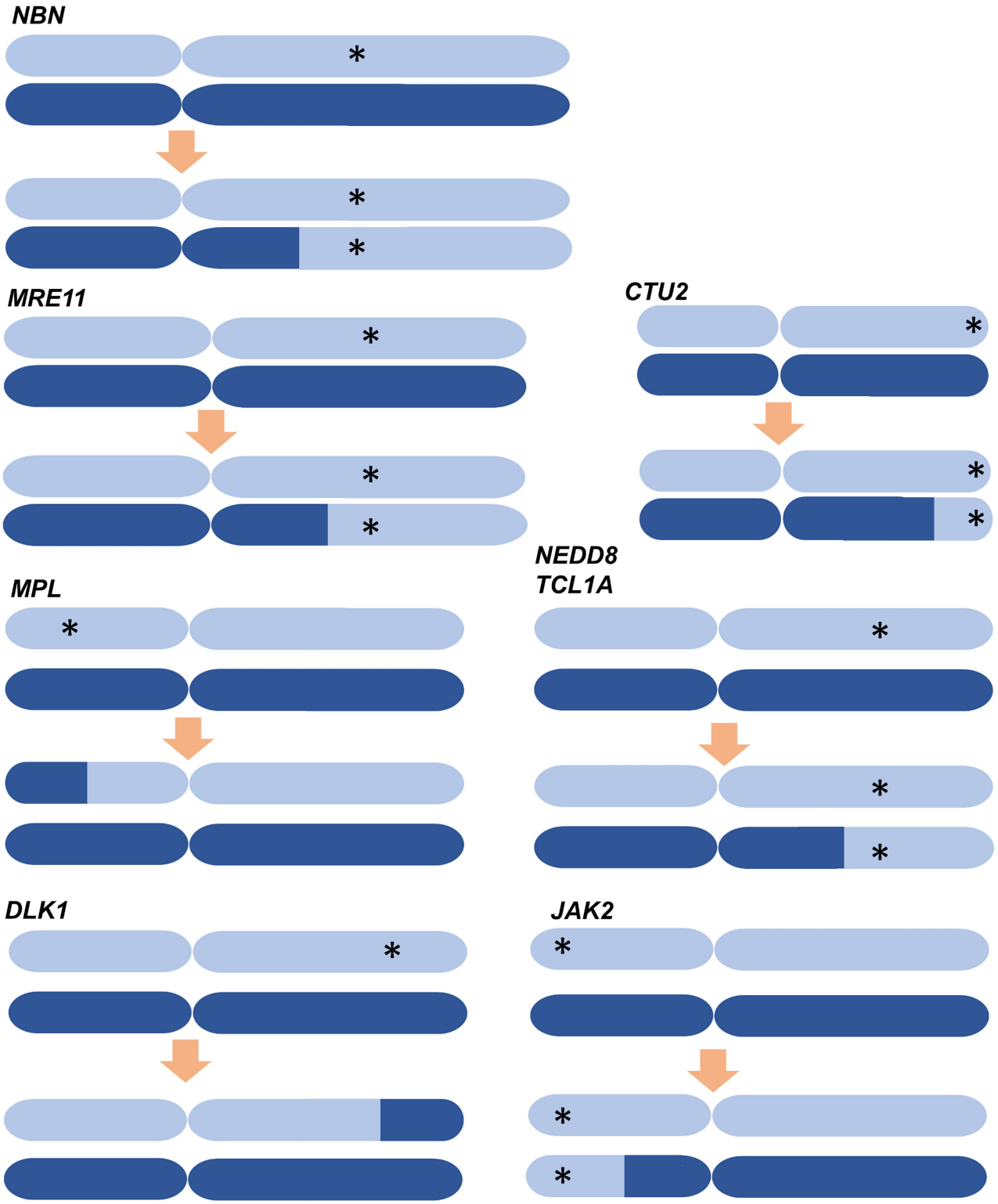
Extended Data Fig. 3 | Similar breakpoint distributions of CN-LOH events in the BBJ and UKB data. Relative frequencies of estimated CN-LOH breakpoint locations in the BBJ and UKB data. Breakpoints were smoothed over ±2 Mb to enable plotting of frequency curves, which were rescaled to 1.



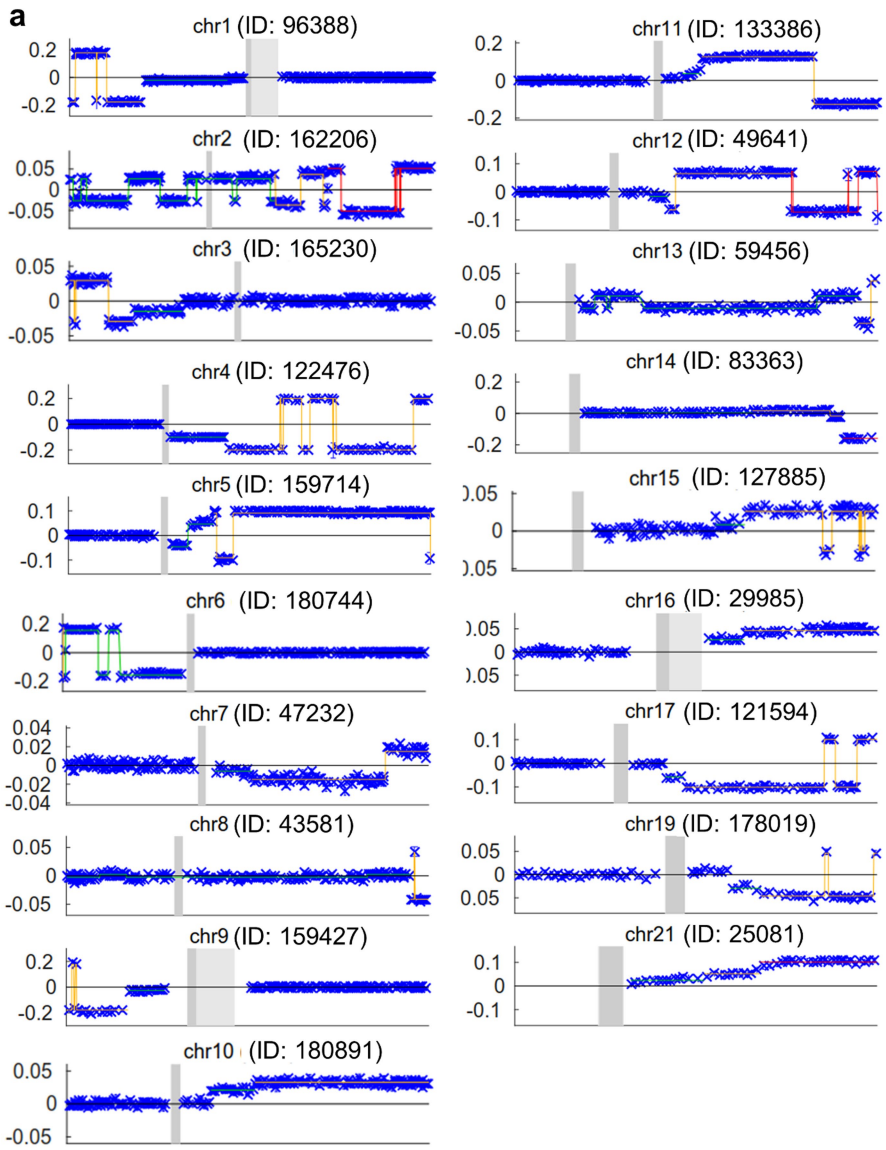
Extended Data Fig. 4 | Quantile–quantile plots of mosaic events with significant associations demonstrate that there is no inflation of association statistics. Quantile–quantile plots of results for mosaic events with significant associations. Analysis results of Fisher's exact test (two-sided, nominal *P* values) using 173,599 participants are shown. We defined the following hits as hit loci: 42–49 Mb at chromosome 1 (1p CN-LOH), 88–94 Mb at chromosome 8 (8q CN-LOH), 92–96 Mb at chromosome 11 (11q CN-LOH), 88–90 Mb at chromosome 16 (16q CN-LOH), 23–26 Mb and 100–103 Mb at chromosome 14 (*cis* association of 14q CN-LOH), 4–6 Mb at chromosome 9 (9p CN-LOH), 0–2 Mb at chromosome 5 (*trans* association of 14q CN-LOH) and 1–3 Mb at chromosome 7 (*trans* association of chromosome 15 gain).



Extended Data Fig. 5 | Local plots for *cis* and *trans* associations. **a**–**i**, Associations of inherited variants with 8q CN-LOH (**a**), 11q CN-LOH (**b**), 16q CN-LOH (**c**), chromosome 15 gain (**d**), 1p CN-LOH (**e**), 9p CN-LOH (**f**) and 14q CN-LOH (**g**–**i**) are shown for regions containing the *NBN*, *MRE11*, *CTU2*, *MAD1L1*, *MPL*, *JAK2*, *NEDD8–TINF2*, *DLK1* and *TERT* loci, respectively. **a**–**c**, **e**, Loci are rare *cis* associations. **f**–**h**, Loci are common *cis* associations. **d**, **i**, Loci are *trans* associations. **a**–**d**, **g**–**i**, Loci are in previously unreported regions. Purple points indicate lead variants. Other variants are colour-coded according to the linkage disequilibrium r^2 with lead variants. The *TCL1A* variant that significantly associated with 14q CN-LOH allelic imbalance is not shown here because it did not significantly associate with 14q CN-LOH risk. Analysis results of Fisher's exact test (two-sided, nominal *P* values) using 173,599 participants are shown.



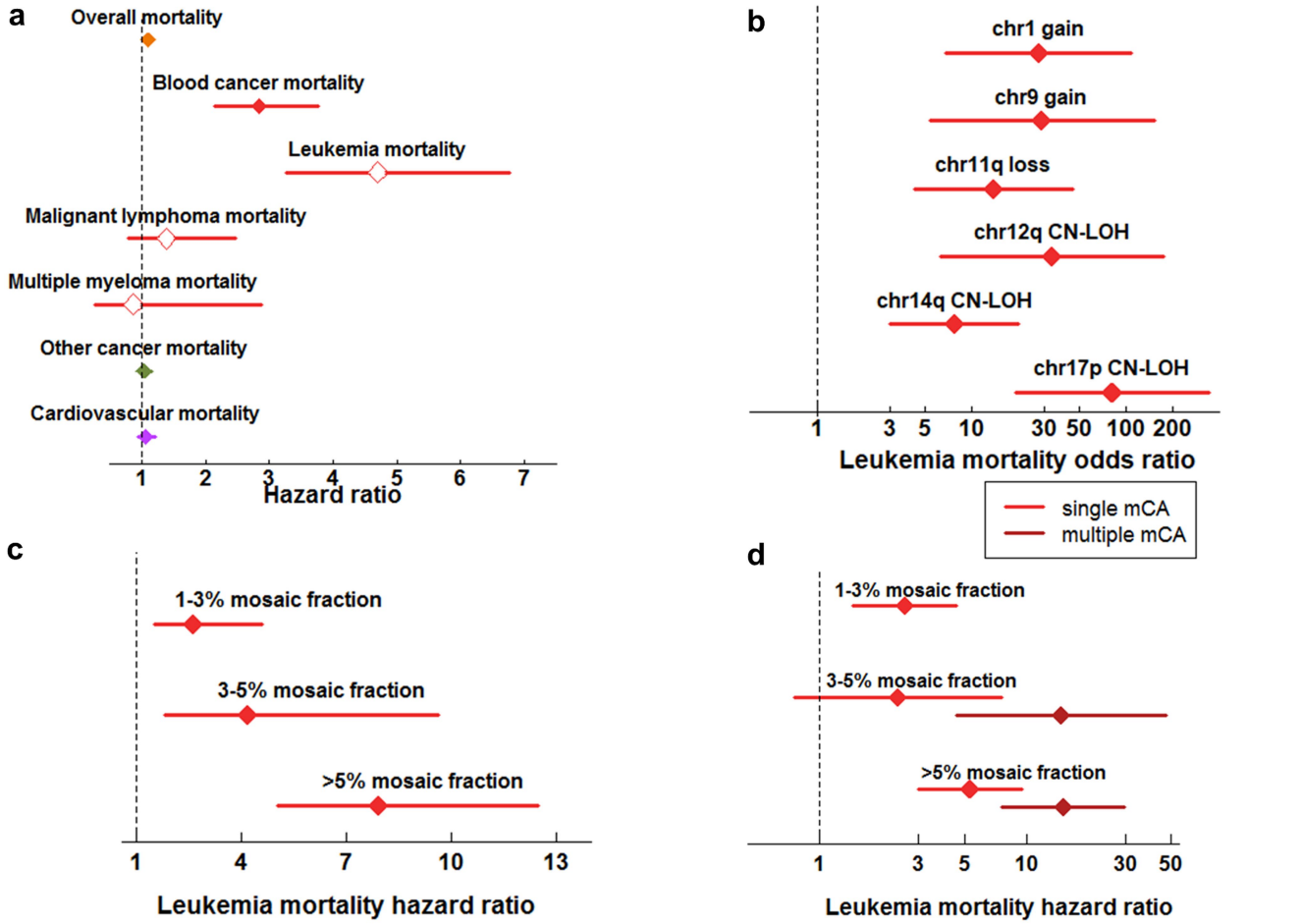
Extended Data Fig. 6 | Action of CN-LOH events on rare and common inherited variants. Schematics show the patterns of selection or elimination of inherited variants by CN-LOH events. Asterisks indicate risk alleles. For the *TCL1A* locus, which did not significantly associate with the presence of 14q CN-LOH, we depict *TCL1A* as a gene for which CN-LOH mutations select an allele.



b

chr	p	q
1	26	2
2	0	3
3	1	3
4	0	6
5	0	5
6	11	0
7	0	1
8	0	2
9	28	1
10	0	1
11	1	13
12	0	2
13	0	1
14	0	57
15	0	5
16	2	4
17	2	5
18	0	0
19	1	2
20	0	0
21	0	2
22	0	0

Extended Data Fig. 7 | Examples of multiple overlapping CN-LOH clones in a single chromosome. We identified 185 individuals who carried multiple CN-LOH clones on a single chromosome. **a**, Multiple clones were observed in at least one individual for all chromosomes except chromosomes 18, 20 and 22. The plots show phased BAF deviations (*y* axis) as a function of chromosome position (*x* axis) for the individual with the largest clone per chromosome (among all individuals with multiple CN-LOH clones on that chromosome). Coloured horizontal lines of different colours indicate distinct BAF deviations corresponding to overlapping CN-LOH events. **b**, The number of participants carrying multiple CN-LOH clones on a single chromosome is shown for each chromosomal arm.



Extended Data Fig. 8 | Mortality risk conferred by mosaic chromosomal alterations. a, Risk of mortality from various causes conferred by presence of an mCA at >1% cell fraction. Leukaemia, malignant lymphoma and multiple myeloma are subdivisions of blood cancer. Cardiovascular mortality includes deaths from coronary artery disease and ischaemic stroke. **b**, Risk of leukaemia mortality conferred by specific mCAs (grouped by chromosomal location and copy-number change) reaching Bonferroni-corrected significance. **c**, Risk of leukaemia mortality conferred by mosaic status stratified by mosaic cell fraction. **d**, Risk of leukaemia mortality conferred by mosaic status stratified by mosaic cell fraction and number of mosaic events detected (one versus two or more). All analyses were restricted to individuals with no previous cancer diagnosis and were corrected for age, sex, smoking status and genotyping array (Methods). Data are hazard ratio or odds ratio and 95% confidence intervals. Numeric data are provided in Supplementary Tables 24–27. Results using 86,546 participants are indicated. Cox proportional hazard models (two-sided) were used for **a**, **b** and **d**. A Cochran–Mantel–Haenszel test was used for **c**.

Extended Data Table 1 | Rare variants associated with CN-LOH further increase risk of multiple overlapping CN-LOH clones

Variant	Gene	mCA	$N_{carrier}$ / $N_{multi\text{-}CN\text{-}LOH}$	p	OR (95%CI)
Rare variants in novel genes					
rs756831345	*NBN*	8q CN-LOH	1 / 1	0.11	Inf(0.22-Inf)
11:94160189	*MRE11*	11q CN-LOH	3 / 12	8.0×10^{-5}	64.8 (8.3-446.9)
rs200779411	*CTU2*	16q CN-LOH	0 / 4	1	-
Previously-reported *MPL* locus					
rs560932816	*MPL*	1p CN-LOH	1 / 20	0.29	3.2 (0.071-23.8)
rs190159566			0 / 20	1	-
rs556241419			1 / 20	0.17	6.4 (0.13-57.1)
rs184778092			1 / 20	0.15	7.7 (0.16-73.9)
Aggregated very rare variants*			3 / 20	0.027	5.5 (1.0-21.0)
Common variants					
rs2183137	*JAK2*	9p CN-LOH	20 / 25	0.086	2.50 (0.88-8.69)
rs28372734	*NEDD8/TINF2*	14q CN-LOH	3 / 17	1	0.76 (0.14-2.75)
rs10873520	*DLK1*	14q CN-LOH	42 / 57	0.074	1.77 (0.96-3.44)

$N_{carrier}/N_{multi\text{-}CN\text{-}LOH}$ indicates the fraction of individuals who carry multiple clones that span the variant who also carried the indicated risk variant(s). OR, odds ratio for carrying multiple clones spanning the variant compared with carrying a single clone. *P*, *P* value obtained using Fisher's exact test (two-sided). *Aggregated very rare variants included rs560932816, rs556241419 and rs184778092, which had variant frequencies that were less than 0.001.

Extended Data Table 2 | Breakdown of associations between mCAs and death attributable to leukaemia

Phenotype	OR (95%CI)	p value
All leukemia	4.70 (3.26-6.78)	$1.0x10^{-16}$
1) myeloid leukemia	5.18 (3.35-8.01)	$1.4x10^{-13}$
2) lymphoid leukemia	3.77 (1.72-8.26)	0.00093
2-1) B cell lymphoid leukemia	4.11 (1.42-11.85)	0.0090
2-2) T cell lymphoid leukemia	3.41 (1.07-10.92)	0.039

Cochran–Mantel–Haenszel tests were used for 86,546 individuals.

nature research

Corresponding author(s): Chikashi Terao, Yoichiro Kamatani

Last updated by author(s): Mar 22, 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	BBJ data and UKB data
Data analysis	Minimac3, Eagle2, R (v3.4, v3.5, v3.6, depending on machines and the time of data analyses), plink1.9 and plink2, matlab v7.9, private C++ codes compiled by gcc-c++v4.8. All computational codes are available upon request to corresponding authors (but not immediately portable to other computing environments). A standalone software implementation (MoChA) of the algorithm used to call mCAs is available at https://github.com/freeseek/mocha.

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

A table for mosaic events detected in the current study is available as Supplementary Data 1. The BBJ genotype is available from the Japanese Genotype-phenotype Archive (JGA; http://trace.ddbj.nig.ac.jp/jga/index_e.html) by application with an accession code JGAD00000000123. Individual-level linkage of mosaic events can be provided by the BBJ project upon request (https://biobankjp.org/english/index.html).

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	179,417 Japanese subjects were used for the current study and they are the maximum number of the BBJ when we started this study. No sample size calculation was performed since we would like to maximize the statistical power to call mosaic in the data set.
Data exclusions	We excluded outliers from East Asian clusters in a plot in which we projected BBJ subjects in combination with 1000 Genomes Project samples in the principal component (PC)1 and PC2 space. We also excluded samples genetically identical to another sample, samples with call rates less than 0.98, and samples whose reported sex information was not supported by genotypes in the X chromosome. We further excluded three samples with evidence of potential contamination (as suggested by low-cell-fraction mosaic events called on many chromosomes).
Replication	not a pure replication, we compared with UKB data and obtained very similar results with population-specific findings
Randomization	Since this was a descriptive study and not an experimental study, no randomisation was performed.
Blinding	Since this was a descriptive study and not an experimental study, we did not undertake blinding.

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

Eukaryotic cell lines

Policy information about cell lines

Cell line source(s)	Jurkat E6.1 cell line and THP-1 were purchased from ATCC (https://www.atcc.org/en.aspx).
Authentication	The cell lines used in this study were obtained from ATCC (https://www.atcc.org/en.aspx) and contamination was consistently monitored. None of the cell lines were authenticated.
Mycoplasma contamination	No contamination when the cell lines were obtained.
Commonly misidentified lines (See ICLAC register)	No commonly misidentified cell lines were used

Human research participants

Policy information about studies involving human research participants

Population characteristics	The BBJ is composed of ~200k participant who had one of the 47 diseases predefined. This cohort has trend of old age (mean age at enrollment 62.8 years and s.d. 14.5 years) and relatively high fraction of male subjects (54.1%). BBJ design is in detail described in Nagai, A. et al. Overview of the BioBank Japan Project: Study design and profile. J. Epidemiol. 27, S2-S8, doi:10.1016/j.je.2016.12.005 (2017).
Recruitment	The BBJ is a multi-hospital-based registry that collected clinical information, DNA, and serum samples from approximately

Recruitment	200,000 patients with one or more of 47 target diseases (including 13 cancers) at a total of 66 hospitals between fiscal years 2003 and 2007. BBJ design is in detail described in Nagai, A. et al. Overview of the BioBank Japan Project: Study design and profile. J. Epidemiol. 27, S2-S8, doi:10.1016/j.je.2016.12.005 (2017). Subjects were recruited based on affection status of the target diseases and not necessarily recruited in a consecutive manner. While we confirmed that most of the disease affection status was not associated with the results, the background of sample collection is worth being kept in mind.
Ethics oversight	the ethics committees of RIKEN Center for Integrative Medical Sciences and the Institute of Medical Sciences, the University of Tokyo

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Clinical data

Policy information about clinical studies

All manuscripts should comply with the ICMJE guidelines for publication of clinical research and a completed CONSORT checklist must be included with all submissions.

Clinical trial registration	This study is not a clinical trial.
Study protocol	BBJ design is in detail described in Nagai, A. et al. Overview of the BioBank Japan Project: Study design and profile. J. Epidemiol. 27, S2-S8, doi:10.1016/j.je.2016.12.005 (2017).
Data collection	The BBJ is a multi-hospital-based registry that collected clinical information, DNA, and serum samples from approximately 200,000 patients with one or more of 47 target diseases (including 13 cancers) at a total of 66 hospitals between fiscal years 2003 and 2007. BBJ design is in detail described in Nagai, A. et al. Overview of the BioBank Japan Project: Study design and profile. J. Epidemiol. 27, S2-S8, doi:10.1016/j.je.2016.12.005 (2017).
Outcomes	We set mortality as a primary outcome in the analyses between mosaic and clinical information since mortality is the best to infer clinical significance of mosaic. The BBJ project has follow-up data to survey mortality and cause of death. For subjects who died, further detailed surveillance was made to identify causes of death coded by ICD10 by accessing national vital registration system used for input survey of medical and social welfare at Ministry of Health, Labor, and Welfare Japanese Government. Further details are available in Hirata, M. et al. Overview of BioBank Japan follow-up data in 32 diseases. J. Epidemiol. 27, S22-S28, doi:10.1016/j.je.2016.12.006 (2017).