Article

Genetic drivers and cellular selection of female mosaic X chromosome loss

https://doi.org/10.1038/s41586-024-07533-7

Received: 20 January 2023

Accepted: 7 May 2024

Published online: 12 June 2024

Check for updates

Aoxing Liu[1,2,3,4,5,31] ✉, Giulio Genovese[4,5,6,31] ✉, Yajie Zhao[7,31], Matti Pirinen[1,8,9], Seyedeh M. Zekavat[4,10,11], Katherine A. Kentistou[7], Zhiyu Yang[1], Kai Yu[12], Caitlyn Vlasschaert[13], Xiaoxi Liu[14], Derek W. Brown[12,15], Georgi Hudjashov[16], Bryan R. Gorman[17,18], Joe Dennis[19], Weiyin Zhou[12,20], Yukihide Momozawa[21], Saiju Pyarajan[17,22], Valdislav Tuzov[16], Fanny-Dhelia Pajuste[16,23], Mervi Aavikko[1], Timo P. Sipilä[1], Awaisa Ghazal[1], Wen-Yi Huang[12], Neal D. Freedman[12], Lei Song[12], Eugene J. Gardner[7], FinnGen*, Estonian Biobank Research Team*, Breast Cancer Association Consortium*, Million Veteran Program*, Vijay G. Sankaran[4,24,25], Aarno Palotie[1,2,4,5], Hanna M. Ollila[1,3,4,26], Taru Tukiainen[1], Stephen J. Chanock[12], Reedik Mägi[16], Pradeep Natarajan[3,4,10], Mark J. Daly[1,2,3,4,5], Alexander Bick[27], Steven A. McCarroll[4,5,6], Chikashi Terao[14,28,29], Po-Ru Loh[4,22,30,32] ✉, Andrea Ganna[1,2,4,5,32] ✉, John R. B. Perry[7,32] ✉ & Mitchell J. Machiela[12,32] ✉

Mosaic loss of the X chromosome (mLOX) is the most common clonal somatic alteration in leukocytes of female individuals[1,2], but little is known about its genetic determinants or phenotypic consequences. Here, to address this, we used data from 883,574 female participants across 8 biobanks; 12% of participants exhibited detectable mLOX in approximately 2% of leukocytes. Female participants with mLOX had an increased risk of myeloid and lymphoid leukaemias. Genetic analyses identified 56 common variants associated with mLOX, implicating genes with roles in chromosomal missegregation, cancer predisposition and autoimmune diseases. Exome-sequence analyses identified rare missense variants in *FBXO10* that confer a twofold increased risk of mLOX. Only a small fraction of associations was shared with mosaic Y chromosome loss, suggesting that distinct biological processes drive formation and clonal expansion of sex chromosome missegregation. Allelic shift analyses identified X chromosome alleles that are preferentially retained in mLOX, demonstrating variation at many loci under cellular selection. A polygenic score including 44 allelic shift loci correctly inferred the retained X chromosomes in 80.7% of mLOX cases in the top decile. Our results support a model in which germline variants predispose female individuals to acquiring mLOX, with the allelic content of the X chromosome possibly shaping the magnitude of clonal expansion.

Female humans carry a maternal and paternal copy of the X chromosome in which one copy is partially rendered transcriptionally inactive early in development[3]. The inactivation process is random in relation to which X chromosome is inactivated, and the resulting inactive state is irreversible and transmitted to daughter cells[4]. X chromosome inactivation has evolved as a mechanism to compensate for gene dosage imbalances between XX female individuals and XY male individuals, although some genes are only partially inactivated[5]. Analytic challenges associated with X inactivation and haploid male X chromosomes have led to fewer studies of the X chromosome, potentially missing critical germline and somatic variations relevant to disease risk.

With age, the expected 1:1 ratio of inactivated maternal to paternal X chromosome copies can become skewed. Skewing of X chromosome inactivation is observed in various tissues, with high frequencies present in leukocytes[6,7]. Detectable skewed X chromosome inactivation is heritable[8] (heritability (h^2) = 0.34) and can indicate depletion of haematopoietic stem cells, selection pressures on leukocytes, or clonal haematopoiesis. Recent investigations of age-related clonal haematopoiesis have described increased rates of mosaic sex chromosome aneuploidies in population-based surveys of healthy adults[1,9–13]. mLOX in female individuals is elevated in frequency compared with mosaic losses in the autosomes[14], preferentially affects the inactivated X chromosome[1] and is associated with increased leukaemia risk[2,15]. This contrasts with the X chromosome in male individuals, which has very low rates of aneuploidy[16]. As the X chromosome encompasses approximately 5% of the genome and contains genes relevant to immunity and cancer susceptibility, loss of a homologous copy and subsequent hemizygous selection could lead to downstream consequences on female health, as observed in Turner syndrome[17]; however, no study has systematically examined longitudinal associations of mLOX with disease risk.

As mLOX is a clonal pro-proliferative genomic alteration, understanding the mechanisms that drive susceptibility to mLOX could provide insights into the effect of ageing on haematopoiesis as well as

A list of affiliations appears at the end of the paper. *Lists of authors and their affiliations appear at the end of the paper.

haematologic cancer risk. The X chromosome, particularly the inactive X chromosome, is more frequently mutated in cancer genomes[18] and is late-replicating relative to autosomes, potentially increasing susceptibility to chromosomal alterations[19]. Although few genome-wide association studies (GWAS) of mLOX have been reported to date[14,20], GWAS of mosaic loss of the Y chromosome (mLOY) in male humans has identified hundreds of susceptibility loci[11–13,21], many of which highlight genes involved in cell cycle regulation and cancer susceptibility. Here we describe insights from epidemiologic and genetic analyses of mLOX for a combined meta-analysis of 883,574 female partricipants (Extended Data Fig. 1). We identify 56 independent common susceptibility variants across 42 loci, rare missense variants of *FBXO10* associated with mLOX, and 44 X chromosome loci that are strongly associated with the X chromosome that is retained in mLOX. The identified signals only partially overlap with known signals for other age-related clonal haematopoiesis. These data indicate that mLOX, along with other forms of clonal haematopoiesis, are important pre-clinical indicators of haematologic cancer risk and identify genes associated with mitotic missegregation, autoimmunity, blood cell traits and cancer predisposition as core aetiologic components for mLOX susceptibility and selection.

Detectable mLOX in eight biobanks

We used genetic data from a total of 883,574 female participants from 8 biobanks worldwide, including European ancestry participants from FinnGen[22], Estonian Biobank[23] (EBB), UK Biobank[24,25] (UKBB), Breast Cancer Association Consortium[26,27] (BCAC), Million Veteran Program[28,29] (MVP), Mass General Brigham Biobank[30,31] (MGB) and Prostate, Lung, Colorectal and Ovarian Cancer Screening Trial[32] (PLCO), as well as participants with East Asian ancestry from Biobank Japan[33] (BBJ) (Extended Data Table 1). The mLOX analysis was restricted to individuals who are genetically female and have two copies of the X choromosome at birth. The age at genotyping ranged from 44 ± 16.3 years for EBB to 65 ± 15.8 years for BBJ (median ± s.d.). We identified mLOX using the mosaic chromosomal alterations (MoChA) WDL pipeline (https://github.com/freeseek/mochawdl), which uses raw signal intensities from single-nucleotide polymorphism (SNP) array data. Out of 883,574 female participants, 105,286 (11.9%) were classified as cases with detectable mLOX (Methods). Overall, the cell fraction of mLOX (that is, the estimated fraction of peripheral leukocytes with X chromosome loss) was low (median = 1.5%) with expanded clones having frequency of 5% or more infrequently observed (0.6% of female participants) (Supplementary Fig. 1 and Supplementary Table 1). A subset of UKBB participants (243,520 out of 261,145) also had whole-exome sequencing (WES) data available, which enabled us to assess the performance of mLOX calling from MoChA. A high correlation ($r = -0.86$) was observed between the cell fraction derived from SNP array data (by MoChA) and X dosage derived from WES data (Supplementary Fig. 2). In addition to the MoChA-generated dichotomous measure used by all biobanks, in UKBB data, we generated a three-way combined quantitative measure by integrating independent information from both SNP array and WES data (Methods).

Lifestyle factors and clinical outcomes

Similar to many other types of somatic mutations[13,14], the frequency of female participants with detectable mLOX in peripheral leukocytes is age-related, with a frequency of 3.0% in female participants below 40 years of age and reaching more than 35.0% after 80 years of age (Supplementary Table 2). Across biobanks, differences were seen in the frequency of mLOX, with the highest age-adjusted frequency presented in EBB and the lowest in MVP (Extended Data Fig. 2a). However, such variation in frequencies was largely reduced when restricted to expanded mLOX with cell fraction above 5% (Extended Data Fig. 2b). To investigate the effect of lifestyle factors on the risk of acquiring detectable mLOX, we assessed associations of smoking and body mass index (BMI) with mLOX in FinnGen and UKBB. Overall, ever-smokers had no increased risk of mLOX ($P = 0.56$ in FinnGen and $P = 0.28$ in UKBB); however, an increased risk was observed among ever-smokers having expanded mLOX with cell fraction of at least 5% (odds ratio (OR) = 1.3 [1.2–1.5], $P = 6.9 \times 10^{-5}$ in FinnGen and OR = 1.3 [1.1–1.5], $P = 4.6 \times 10^{-4}$ in UKBB) (Supplementary Table 3 and Supplementary Figs. 3 and 4). We observed limited evidence for an association between BMI and mLOX (Supplementary Table 4).

To evaluate disease outcomes associated with detectable mLOX, we performed Cox proportional hazards regression for incident disease cases in FinnGen, UKBB, MVP and MGB, adjusting for genotyping age and ever-smoking status as covariates and meta-analysing across biobanks with a fixed-effect model (Methods). Out of the 1,253 diseases that we examined, we identified significant associations ($P < 4.0 \times 10^{-5}$) with leukaemia overall (hazard ratio (HR) = 1.7 [1.5–2.1], $P = 3.5 \times 10^{-10}$) and chronic lymphoid leukaemia (CLL) (HR = 3.3 [2.4–4.4], $P = 8.4 \times 10^{-15}$) and suggestive evidence for acute myeloid leukaemia (AML) (HR = 1.9 [1.3–2.8], $P = 1.8 \times 10^{-3}$) (Supplementary Table 5).

As the median fraction of cells affected by mLOX is approximately 2%, we proposed that expanded clones could result in stronger disease associations. We focused on mLOX with cell fractions of at least 10%, as this threshold has been empirically determined to be aetiologically relevant for detecting diseases associated with mosaic chromosomal alterations[2,15] (mCAs). Restricting to expanded mLOX, we observed evidence for increased associations with leukaemia overall (HR = 6.3 [3.9–10.2], $P = 7.3 \times 10^{-14}$), CLL (HR = 14.7 [6.5–33.3], $P = 9.5 \times 10^{-11}$) and AML (HR = 10.6 [3.1–36.1], $P = 1.5 \times 10^{-4}$) (Supplementary Table 6). To examine the potential effects of other types of clonal haematopoiesis on mLOX associations with leukaemia, we performed sensitivity analyses in UKBB where we had available calls on autosomal mCAs as well as clonal haematopoiesis mutations in driver genes, commonly referred to as clonal haematopoiesis of indeterminate potential[34] (CHIP). We observed attenuations in associations for expanded mLOX when removing individuals with autosomal mCAs (HR = 3.8 [1.6–9.3], $P = 2.7 \times 10^{-3}$), CHIP (HR = 6.2 [3.1–12.4], $P = 3.1 \times 10^{-7}$), and both mCAs and CHIP (HR = 4.5 [1.9–10.8], $P = 8.6 \times 10^{-4}$) (Supplementary Table 7); however, significant associations with expanded mLOX and overall leukaemia risk remained, indicating that mLOX is independently associated with leukaemia risk. Associations for other lymphoid and myeloid leukaemias display similar patterns, albeit losing statistical significance, probably owing to reduced sample size.

We further assessed the relationship between mLOX and a broad range of quantitative phenotypes in UKBB (Methods and Supplementary Table 8) and observed enrichment of associations with blood count traits, such as higher lymphocyte count ($P = 9.3 \times 10^{-126}$) and lower neutrophil count ($P = 3.3 \times 10^{-62}$). As for blood biomarkers or biochemistry, acquiring mLOX was associated with shorter telomere length (for example, $P = 2.8 \times 10^{-14}$ for adjusted telomere to single copy gene (T/S) ratio) and higher levels of total protein ($P = 1.9 \times 10^{-8}$). We noted that, unlike disease associations that usually exerted more significant effects in expanded mLOX (for example, in various subtypes of leukaemia), for quantitative phenotypes, most of the identified associations did not hold for expanded clones, suggesting that mLOX of different cell fraction ranges might not reflect the same medical or biological conditions in female participants.

Common and rare germline contributors

We performed a GWAS to identify common and low-frequency germline variants (minor allele frequency (MAF) > 0.1%) associated with the risk of developing detectable mLOX in peripheral leukocytes. We examined the autosomes (chromosomes 1–22) and X chromosome in each of the 8 contributing biobanks independently, for a total of 883,574 female participants (Methods). To increase power, we used enhanced 3-way

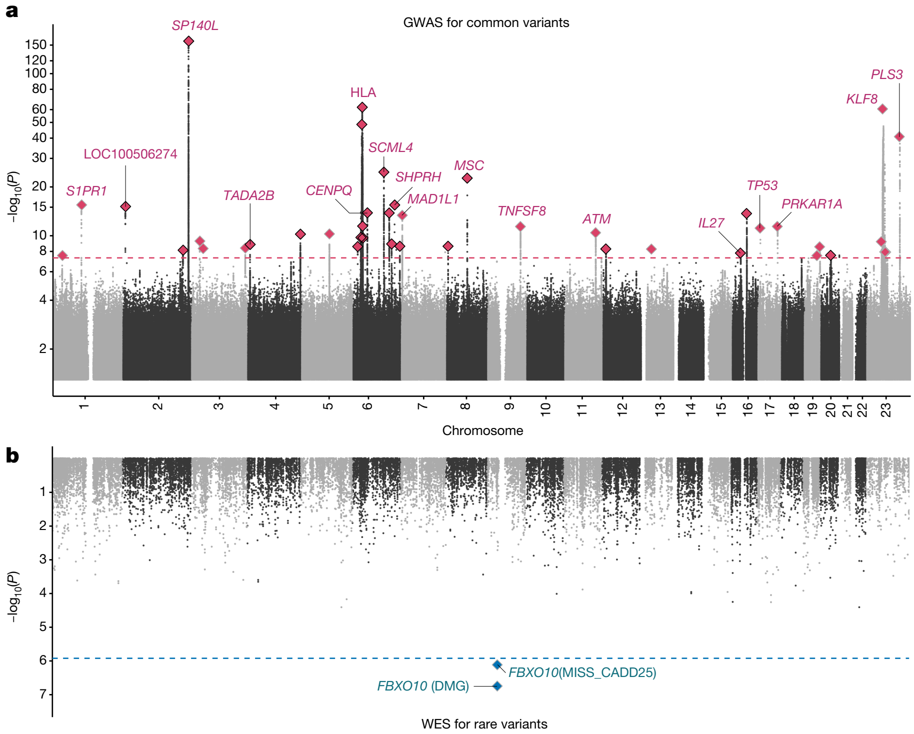


Fig. 1 | Common and rare genetic contributors to mLOX susceptibility. **a**, GWAS $-\log_{10}(P)$ for the association of common variants (MAF > 0.1%) with mLOX in a combined meta-analysis of 883,574 female biobank participants with a weighted *z*-score method. Labels are assigned for candidate genes of the top 10 lead variants from meta-analysis or the top 10 candidate genes from gene prioritization. **b**, Gene burden test $-\log_{10}(P)$ for the associations of rare variants (MAF < 0.1%) with mLOX in 226,125 female UKBB participants with available WES data. In **a**,**b**, the *y* axis shows the log scale of *P* values from a two-sided test. Dashed lines denote the statistical significance after multiple comparison adjustments: 5.0×10^{-8} for GWAS (**a**) and 1.2×10^{-6} for the gene burden test (**b**).

combined calls for UKBB and meta-analysed summary statistics across different mLOX measures with a weighted *z*-score method (Methods). Among the 33,737,925 variants that we examined, we identified 56 independent genome-wide significant variants ($P < 5.0 \times 10^{-8}$) across 42 loci associated with mLOX susceptibility (Fig. 1a, Supplementary Table 9 and Methods). Most independent variants were located on chromosomes 6 (17 variants), 2 (9 variants) and X (7 variants), with these chromosomes explaining more heritability than expected for their chromosome length (Supplementary Fig. 5). The mLOX variant effects were consistent across the 8 biobanks (Cochran's *Q*-test, $P > 0.05/56 = 8.9 \times 10^{-4}$) (Supplementary Table 10), with the exception of rs78378222 (*TP53*; meta-analysis, $P = 7.2 \times 10^{-12}$; heterogeneity test, $P = 6.7 \times 10^{-4}$) and three X chromosome variants (X:51749114:C:CGT, rs141849992 and rs58638231). For rs78378222, the heterogeneity of effects was likely to be due to differences in mLOX cell fraction by contributing studies. When stratifying by cell fraction in FinnGen, the odds ratio for the risk allele of rs78378222 was 1.1 [1.0–1.2] ($P = 0.01$) for cell fractions below 5% but reached 1.7 [1.3–2.3] ($P = 1.4 \times 10^{-4}$) for expanded mLOX with cell fractions above 5% (effect-size difference from a two-sided *t*-test, $P = 2.5 \times 10^{-5}$) (Supplementary Table 11 and Supplementary Fig. 6). Between participants with European ($N = 806,257$) and East Asian ($N = 77,317$) ancestry, we found that mLOX signals were largely shared in the two groups except for four variants rs11686798, rs57760309, rs6521410 and rs141849992, which had mLOX effects in the same directions but heterogeneous effect sizes (Supplementary Fig. 7 and Supplementary Table 12).

We deployed a range of variant-to-gene mapping approaches to rank genes proximal to each of our hits by their strength of evidence for causality (Methods), highlighting the highest-scoring gene at each locus (Supplementary Table 13). The most significantly associated mLOX locus is at 2q37.1, replicating previous UKBB mLOX GWAS signals at that locus[14,20]. We mapped the hit to *SP140L*, a gene that is predicted to be involved in regulation of transcription by RNA polymerase II and active in the nucleus. Nearby genetic variants are associated with lymphocyte percentage[35]. Several identified mLOX loci implicated plausible causal genes relevant to cancer predisposition, including *EOMES* (3p24.1), *JARID2* (6p22.3), *MYB* (6q23.3), *MAD1L1* (7p22.3), *TNFSF8* (9q32–q33.1), *ATM* (11q22.3), *HEATR3* (16q12.1), *TP53* (17p13.1), *PRKAR1A* (17q24.2) and *KLF8* (Xp11.21), many of which (such as *EOMES*[36,37], *JARID2*[38], *MYB*[39], *ATM*[40], *TP53*[41] and *PRKAR1A*[42]) are directly relevant to leukaemia predisposition or progression. Additionally, highlighted genes at several mLOX loci are important for mitotic spindle assembly and kinetochore function including *MAD1L1* (7p22.3), *CENPU* (4q35.1), *CENPQ* (6p12.3) and *CENPW* (6q22.32), all of which are relevant to mitotic missegregation errors leading to loss of an X chromosome at a single cell level. Several mLOX-associated loci also implicate genes related to immunity and autoimmune disorders including *EOMES* (3p24.1), *LPP-AS1* (3q28), *CENPU* (4q35.1), *ERAP2* (5q15), *HLA-A* (6p22.1), *HSPA1A* (6p21.33), *ITPR3* (6p21.31), *CENPW* (6q22.32), *MYB* (6q23.3), *MSC* (8q13.3), *TNFSF8* (9q32–q33.1), *IL27* (16p12.1–p11.2) and *LILRA1* (19q13.42), suggesting a shared aetiologic relationship between mLOX and immune cell function. Similar to these locus-specific results, the genome-wide pathway-based analysis identified enrichment in pathways related to DNA damage response, cell cycle regulation, cancer susceptibility and immunity (Methods and Supplementary Table 14).

We next investigated whether the identified common variants for mLOX susceptibility in female participants were associated with mLOY, the most common leukocyte sex chromosome mosaicism in male participants, and similarly, whether mLOY loci were associated with mLOX. We utilized a Bayesian model to assign 56 independent common variants identified from mLOX GWAS and 147 variants (9 variants were dropped owing to missing in mLOX GWAS) from the published mLOY

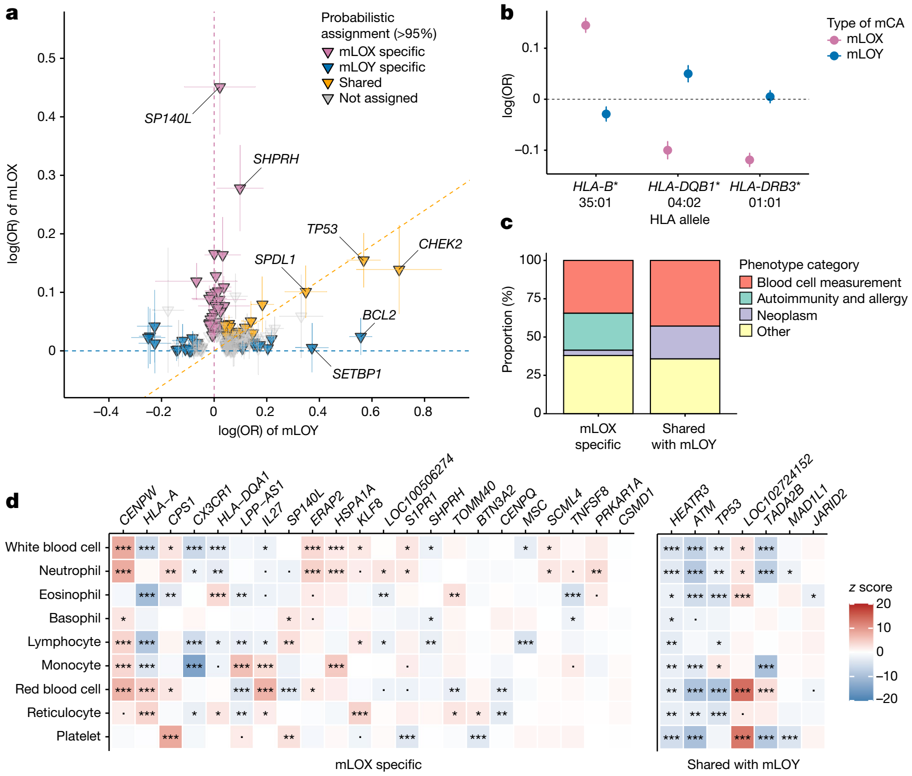


Fig. 2 | Shared and distinct genetic contributors to mLOX susceptibility in female participants and mLOY susceptibility in male participants. a, The effects on mLOX and mLOY of mLOX susceptibility variants ($N = 56$) and mLOY susceptibility variants[13] ($N = 147$). Variants are assigned to 'mLOX specific', 'mLOY specific' or 'shared' by applying a Bayesian model with posterior probability >95%. **b**, Fine-mapping of imputed HLA alleles for mLOX ($N = 168{,}838$) and mLOY in FinnGen ($N = 174{,}404$), for 3 HLA alleles that are significantly associated with mLOX from stepwise conditional analyses. Data are mean ± s.e.m. **c**,**d**, Phenotype associations for lead variants of 29 independent mLOX susceptibility loci that were assigned to either mLOX specific or shared with mLOY. **c**, Phenotype associations (GWAS lead variants ($r^2 > 0.6$)) from Open Targets genetics. To avoid pleiotropic effects, we categorized phenotypes into blood cell measurement, autoimmunity and allergy, neoplasm and others. The association with each phenotype category was first examined at a variant level and then summarized over all variants assigned to the same category in terms of the relationship with mLOY. To avoid the associations driven by HLA signals, we excluded all identified variants from the extended MHC region (GRCh38: chr. 6: 25.7–33.4 Mb). **d**, Heat map for associations with nine blood cell count traits[46], with significance levels from the original GWAS expressed by asterisks (two-sided exact P; $***P \leq 0.001$, $**P \leq 0.01$, $*P \leq 0.05$). Absolute z scores were cropped to the range of [0–20].

GWAS[13] into 3 groups: specific to mLOX, specific to mLOY, and shared between mLOX and mLOY (Fig. 2a and Methods). Out of 56 variants identified from the mLOX GWAS, we assigned 34 variants as specific for mLOX and 7 as shared with mLOY, with greater than 95% probability (Supplementary Table 15). Among 3 centromere protein genes identified for mLOX susceptibility, *CENPQ* (rs9395493; OR = 1.04 [1.03–1.05] for mLOX and 0.99 [0.98–1.01] for mLOY; effect-size difference, $P = 4.1 \times 10^{-9}$) and *CENPW* (rs9372840; OR = 1.04 [1.03–1.06] for mLOX and 1.02 [1.01–1.04] for mLOY; effect-size difference, $P = 0.01$) were specific to mLOX with posterior probability greater than 95%, whereas for *CENPU* (4:184696883:C:CT; OR = 0.96 [0.94–0.97] for mLOX and 0.97 [0.95–0.98] for mLOY; effect-size difference, $P = 0.11$) the probability of being mLOX-specific was 83%. When similarly examining the 147 mLOY susceptibility variants, we further identified 8 variants (prioritized genes such as *SPDL1*, *HLA-A*, *CHEK2* and *MAGEH1*) to be shared with mLOX susceptibility, in addition to the 6 variants that are exactly mLOX GWAS lead variants (prioritized genes *GRPEL1*, *QKI*, *TP53* and *MAD1L1*) or in high linkage disequilibrium (LD) ($r^2 > 0.6$) with mLOX GWAS lead variants (prioritized genes *ATM* and *HEATR3*). Notably, for variants that are shared between mLOX and mLOY, ORs were attenuated for mLOX relative to mLOY, possibly owing to lower cell fractions observed for mLOX compared with mLOY (Supplementary Fig. 1). For example, for rs78378222 (*TP53*), the effect size for mLOX (OR = 1.17 [1.11–1.22]) was lower than for mLOY (OR = 1.77 [1.65–1.88]) (effect-size difference, $P = 6.0 \times 10^{-35}$). Similarly, for rs2280548 (*MAD1L1*), the effect for mLOX (OR = 1.04 [1.03–1.05]) was also lower than for mLOY (OR = 1.13 [1.11–1.14]) (effect-size difference, $P = 1.1 \times 10^{-25}$). This smaller effect size together with the lower frequency of mLOX (for example, 6.2% for 261,145 female UKBB participants aged 40–70 years at genotyping) relative to mLOY (for example, 20.4% for 205,011 male UKBB participants aged 40–70 years at genotyping[13]) indicates that a large meta-analysis was needed to identify susceptibility variants for mLOX. The partially shared genetic architecture from common variants between mLOX and mLOY was also supported by the moderate genetic correlation ($r = 0.30$ [0.21–0.39], $P = 1.7 \times 10^{-10}$) (Methods and Supplementary Table 16). We note that in addition to potential differences in biological mechanisms, the differences between mLOX and mLOY could also be related to differences in cell fractions, as calling algorithms can detect smaller cell fractions of mLOX events relative to mLOY events.

We then explored the overlaps of mLOX susceptibility variants with autosomal mosaicism, a more heterogeneous group comprising multiple types of detectable mosaic events (loss, gain and copy-neutral loss of heterozygosity) on chromosomes 1–22, and whether the reported autosomal mCA *trans* variants in UKBB[43] (3.6% of autosomal mCA cases among 452,469 participants) act in mLOX acquisition in female participants. Of the 55 mLOX variants (one missing) available in the UKBB autosomal mCA GWAS, no variant reached genome-wide significance for autosomal mCAs (Supplementary Table 17). Together with the identified effects on mLOY, our analysis suggested that seven of the mLOX variants were specific for mLOX susceptibility (prioritized

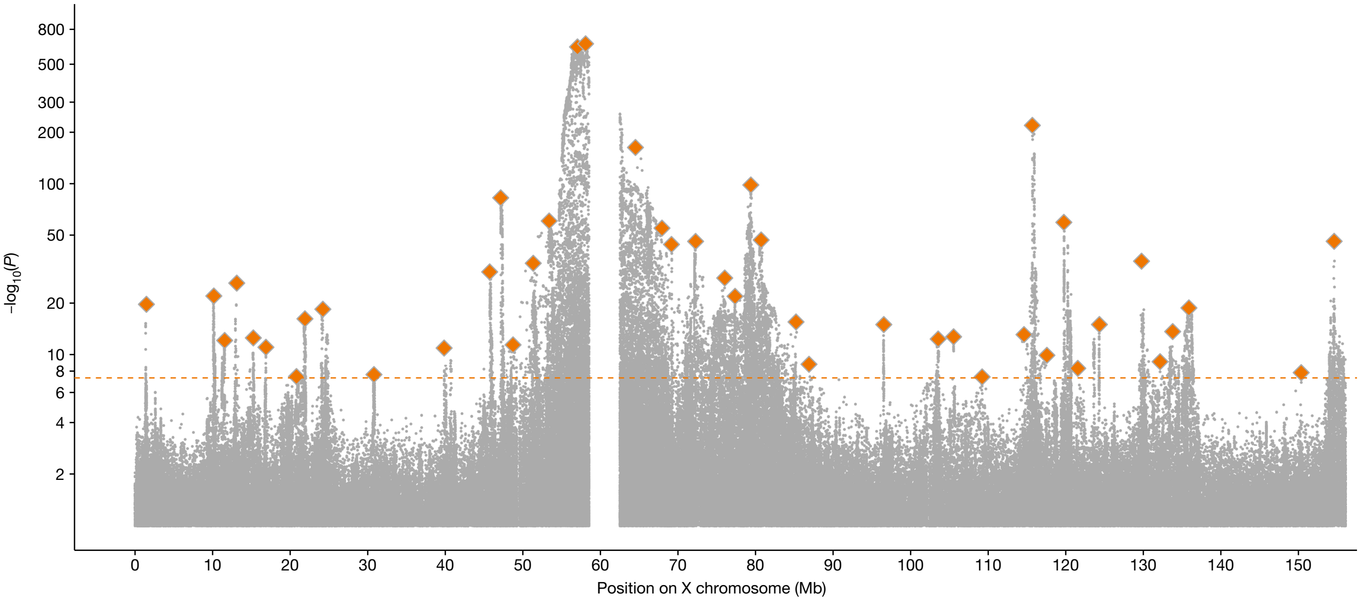


Fig. 3 | Allelic shift of X chromosome alleles among mLOX cases. $-\log_{10}(P)$ of X chromosome variants from allelic shift analysis (two-sided test) by meta-analysis of data from 83,320 mLOX cases from 7 biobanks, with lead variants of 44 independent loci highlighted. The dashed line denotes the statistical significance after multiple comparison adjustments ($P = 5.0 \times 10^{-8}$, the same as the significance threshold for the GWAS).

genes *LOC100506274*, *SP140L*, *HSPA1A*, *CENPW*, *SHPRH*, *TOMM40* and *KLF8*) and three were shared with both mLOY and autosomal mCAs (prioritized genes *MAD1L1*, *ATM* and *TP53*). Additionally, for the three loci reported as associated with any detectable autosomal mCAs in *trans*[43], only the lead variant (rs62191195 (*SP140*)) exerted shared effects with mLOX (OR = 1.05 [1.04–1.06] for mLOX and 1.08 [1.05–1.10] for autosomal mCAs; effect-size difference, $P = 0.08$), whereas the other two variants (rs12638862 (*TERC*) and rs7705526 (*TERT*)) presented limited effects on mLOX.

Given the many associations of HLA genes with mLOX, we fine-mapped HLA alleles at a unique protein sequence level on 10 genes commonly used for HLA marker matching in organ transplantation for a set of 168,838 Finnish female participants (mLOX cases, $N = 27,001$) and 128,729 Finnish male participants (mLOY cases, $N = 45,675$) (Methods and Supplementary Fig. 8). Out of 156 examined HLA alleles, 16 alleles were associated with the odds of developing detectable mLOX ($P < 5.0 \times 10^{-8}$), including alleles from both major histocompatibility complex (MHC) class I (6 out of 74 examined alleles locating on *HLA-A*, *HLA-B* or *HLA-C*) and class II molecules (10 out of 82 examined alleles locating on *HLA-DR*, *HLA-DP* and *HLA-DQ*) (Supplementary Table 18). The most significant HLA allele HLA-B*35:01 increased the risk of mLOX (OR = 1.16 [1.12–1.19], $P = 1.1 \times 10^{-23}$), but had no effect on mLOY (OR = 0.97 [0.94–1.00], $P = 0.03$; effect difference with mLOX, $P = 3.6 \times 10^{-18}$) (Fig. 2b). This association with HLA-B*35:01 was independently replicated in BBJ (OR = 1.10 [1.05–1.15], $P = 1.5 \times 10^{-5}$). The HLA-B*35:01 allele is well established as the major driver for the progression of HIV[44] and is also associated with several autoimmune diseases (for example, subacute thyroiditis[45] (OR = 4.36 [3.25–5.85])). Using stepwise conditional analyses in FinnGen, we identified two independent genome-wide significant HLA associations at HLA-DRB3*01:01 (copy number variation that presents only in a subset of individuals) (OR = 0.89 [0.87–0.91], $P = 2.8 \times 10^{-19}$) and HLA-DQB1*04:02 (OR = 0.90 [0.87–0.94], $P = 6.5 \times 10^{-9}$). For mLOY in male participants, despite a larger effective sample size, no HLA allele reached the genome-wide significant threshold suggesting that HLA has a larger role in mLOX than mLOY. Likewise, we observed no evidence for associations of HLA alleles with autosomal mCAs. Additionally, we conducted conditional GWAS analyses in FinnGen by adjusting for the three lead variants (rs74615740 (*HLA-B*) ($r^2 = 0.45$ with HLA-B*35:01), rs9275511 (*HLA-DQA2*) and rs2734971 (*HLA-G*)) identified from the Finnish population GWAS. The results suggested that the associations with mLOX observed in the extended MHC region (GRCh38: chromosome (chr.) 6:25.7–33.4 Mb) were probably due to HLA signals instead of nearby non-HLA variants (Supplementary Fig. 9).

To understand potential mechanisms relevant to mLOX susceptibility revealed by each identified mLOX variant, we examined associations with additional phenotypes documented in the Open Target genetics platform. Out of 56 independent variants, 30 were in LD ($r^2 > 0.6$) with at least one GWAS lead variant from Open Target (5.0×10^{-8}) (Supplementary Table 19). Notably, more than half of the phenotype associations were with variants associated with blood cell trait measurements, autoimmunity and allergy, and neoplasms (Fig. 2c). Several mLOX-specific variants are GWAS lead variants of multiple autoimmune diseases such as type 1 diabetes (rs9372840 (*CENPW*) and rs181206 (*IL27*)), coeliac disease (rs13080752 (*LPP-AS1*)) and rheumatoid arthritis (rs2887944 (*EOMES*)). On the basis of Open Target genetics, none of the mLOX variants shared with mLOY were reported to be associated with any autoimmune disease. Additionally, the group of variants shared with mLOY have more associations with neoplasms (for example, rs751343 (*ATM*) for breast cancer and rs2280548 (*MAD1L1*) for prostate cancer) and blood cell measurements than the group of variants specific for mLOX. We then examined the associations between each identified mLOX susceptibility locus and the counts of different types of blood cells[46]. Of 42 independent mLOX loci (only considering the lead variant of each locus), 39 were associated with at least one of the 9 blood cell count traits examined ($P < 0.05$), suggesting a shared genetic aetiology between haematopoiesis and development of detectable mLOX (Fig. 2d). Again, the mLOX variants shared with mLOY were among the variants associated with the largest number of blood cell traits (an average of 5.0 traits over 7 variants) compared to mLOX specific variants (an average of 3.3 traits over 22 variants).

To identify rare germline variants (MAF < 0.1%) associated with the susceptibility of detectable mLOX, we performed gene burden tests for our newly proposed 3-way combined calls in 226,125 UKBB female participants with available WES data (Methods). Only one gene, *FBXO10* (encoding F-box protein 10), was associated with mLOX susceptibility ($P < 1.2 \times 10^{-6}$) (Fig. 1b), with the strongest association observed

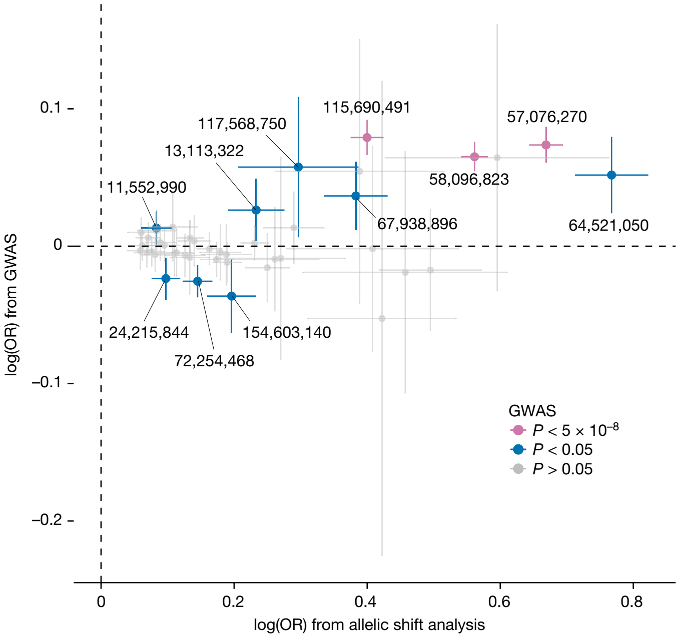


Fig. 4 | X chromosome alleles under *cis* selection and their effects on mLOX susceptibility. Lead variants identified from the allelic shift analysis ($N = 44$) after multiple comparison adjustments ($P < 5.0 \times 10^{-8}$) with a two-sided binomial test of their effects from the allelic shift analysis (*x* axis) plotted against the GWAS effects (*y* axis). Variants are categorized on the basis of exact *P* values from the GWAS.

in carriers of missense variants with CADD scores at least 25 (581 carriers, beta = 0.059, $P = 1.8 \times 10^{-7}$) (Supplementary Table 20). Logistic regression for the dichotomous mLOX status observed a consistent effect of *FBXO10* missense variants associated with a twofold increased risk of acquiring mLOX (OR = 2.1 [1.6–2.7], $P = 1.4 \times 10^{-7}$). We further confirmed this association using a distinct analytical pipeline implementing STAAR[47] ($P = 2.5 \times 10^{-7}$) and SAIGE-GENE+[48] ($P = 9.5 \times 10^{-8}$ for the 3-way combined quantitative measure and $P = 3.0 \times 10^{-7}$ for the dichotomous status). A leave-one-out analysis confirmed this association was not restricted to a single coding variant ($P < 3.0 \times 10^{-7}$). FBXO10 is the substrate-recognition component of the SKP1–CUL1–F-box protein (SCF)-type E3 ubiquitin ligase complex. The SCF complex mediates ubiquitination and degradation of the anti-apoptotic regulator BCL2, and thereby has a role in apoptosis by controlling the stability of BCL2[49].

cis selection of X chromosome alleles

As several germline variants reside on the X chromosome, we sought to investigate whether—for a given X chromosome variant—mLOX cells with one allele retained in a hemizygous state confer a propensity to be retained or a selective advantage over mLOX cells with the alternate X allele retained (Extended Data Fig. 1b). Conditional on mLOX having been detected, for each variant on the X chromosome, we tested whether there is a higher frequency of a given allele retained compared with the alternate allele being retained[14] (Methods). This allelic shift analysis is similar to a transmission disequilibrium test[50] which is robust to the presence of population structure, with only heterozygous genotypes being informative. Of the 1,645,601 X chromosome variants we examined, 25,370 (1.5%) reached the significance threshold ($P < 5.0 \times 10^{-8}$). We identified 44 independent X chromosome variants with shifted allelic fractions on the retained X chromosome (Methods and Supplementary Table 21). The allelic shift signals spanned the length of the X chromosome (Fig. 3), with the strongest signals observed near the centromere (lead variant rs6612886; out of 39,246 heterozygous rs6612886 genotypes examined, 25,035 had the alternative C allele lost while 14,211 had the reference T allele lost, OR = 1.76 [1.73–1.80], $P = 4.0 \times 10^{-659}$). To investigate if the observed associations were driven by variant density, we explored the relationship between the number of markers being statistically significant and the total number of markers we examined within a window size of 1 kb and found no relationship between the two measures (Supplementary Fig. 10). Finally, signals were consistent across seven biobanks further supporting the robustness of the results (Supplementary Fig. 11 and Supplementary Table 22).

Similar to GWAS lead variants, 35 out of 43 lead variants (one variant was dropped owing to no appropriate proxy variant available in blood cell phenotype GWAS[46]) identified from allelic shift analyses were associated with at least one blood cell phenotype (prioritized genes *P2RY8*, *WAS*, *PJA1*, *PLS3*, *ITM2A*, *TMEM255A* and *SOWAHD*) (Supplementary Table 23), especially for several variants near the centromere region (Extended Data Fig. 3).

Among variants exhibiting significant allelic shifts in mLOX cases, 59 were coding variants (Supplementary Table 24) including 16 variants from 11 genes (*P2RY8*, *FANCB, UBA1*, *WAS*, *USP27X*, *VSIG4*, *PJA1*, *CITED1*, *POF1B*, *SAGE1* and *MAP7D3*) that are likely to be lead signals (Supplementary Fig. 12). The genes *VSIG4* (rs41307375, rs41306131 and rs17315645, $r^2 < 0.001$) and *SAGE1* (rs41301507 and rs4829799, $r^2 = 0.30$) each contained more than one independent missense variant. On the basis of the Human Protein Atlas (https://www.proteinatlas.org/), several genes with identified missense variants were also associated with cancer risk/progression (*P2RY8*, *UBA1*, *WAS* and *SAGE1*), mental disorders (for example, *USP27X* for intellectual disability and *PJA1* for schizophrenia[51]), or had relevance to DNA damage and repair (*FANCB*) and apoptosis (*CITED1)*. Additionally, several genes were involved in X-linked recessive disorders (for example, *FANCB* for Fanconi anaemia, *WAS* for Wiskott–Aldrich syndrome, and *POF1B* for X-linked premature ovarian failure) or are known to escape from X chromosome inactivation[5] (for example, *P2RY8*, *UBA1*, *WAS*, *VSIG4* and *POF1B*).

Most X chromosome variants identified from the allelic shift analysis were not shared with the variants from the GWAS of mLOX (Fig. 4), except for rs4029980 (X:57044373:T:C, proxy SNP X:57076270:G:A, $r^2 = 0.87$) and rs6612886 (X:58090464:T:C, proxy SNP X:58096823:A:C, $r^2 = 0.98$) near the centromere and rs12836051 (X:115690491:A:G). Unlike GWAS, which can identify germline variants related to both chromosome missegregation and subsequent clonal selection, X chromosome signals identified from allelic shift analysis suggests that in many female participants, mLOX strongly favours one X chromosome over the other based on the differing allelic content. This preference could arise from the clonal selection on retained alleles or could be owing to allelic influences on X inactivation skewing (Extended Data Fig. 4), which later manifests as an allelic shift if mLOX occurs since mLOX mostly affects the inactive X chromosome[1].

We then investigated how accurately we could predict which X chromosome is likely to be retained when detectable mLOX occurs. An X chromosome differential score was constructed on the basis of the 44 independent variants identified from allelic shift analysis by generating a chromosome-specific score for each X chromosome and calculating the difference between scores of two X chromosomes (Methods). To avoid overfitting, the prediction performance was tested with data from 27,001 FinnGen mLOX cases, with effect sizes of lead variants estimated from the allelic shift analysis of 56,319 mLOX cases from six biobanks excluding FinnGen. The fraction of mLOX cases with the retained X chromosome correctly inferred was 63.7% across all mLOX cases and up to 80.7% for mLOX cases within the top 10th percentile (Fig. 5). When partitioning the contribution at a variant level, starting from the most significant variants (Extended Data Fig. 5), the fraction correctly inferred reached >60% when including the first four lead variants (rs58090464, rs57044373, rs115690491 and rs79395749), and the improvement of prediction accuracy from adding another 40 lead variants increased performance but was smaller in comparison (fraction from 60.3% to 63.7%). We also performed simulation analyses to

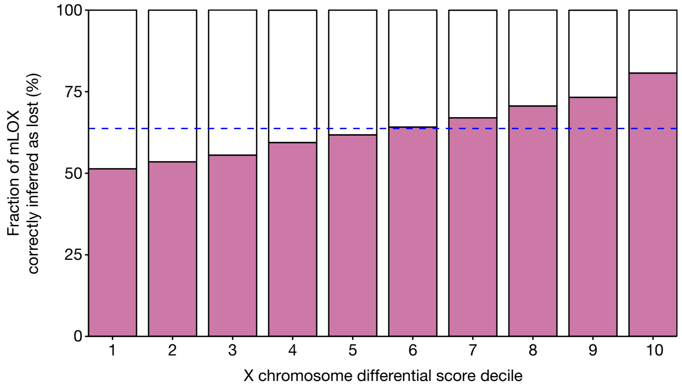


Fig. 5 | Inferring the retained X chromosome in female biobank participants with mLOX. Fraction of mLOX cases with the retained X chromosome correctly inferred using an X chromosome differential score constructed from allelic shift analysis signals. To avoid overfitting, the effects of 44 lead variants were estimated from allelic shift analysis of 56,319 mLOX cases from 6 biobanks excluding FinnGen, and the prediction performance was tested with 27,001 FinnGen mLOX cases.

assess the upper limit of prediction performance that can be reached in FinnGen mLOX cases, given the distribution of allele frequencies of 44 lead variants (Methods). Overall, the fraction of mLOX cases correctly inferred from real data analysis (63.7%) approached that obtained from simulation analysis (65.0%) (Supplementary Figs. 13 and 14). To further understand whether female individuals carrying higher X chromosome differential scores would have an increased lifetime disease risk, we examined its association with 1,630 disease endpoints in 27,001 FinnGen mLOX cases (Methods) and identified significant associations with cardiovascular diseases (for example, for major coronary heart disease event, HR = 1.13 [1.07–1.20] for a 1× s.d. change in the score, $P = 2.1 \times 10^{-5}$) and suggestive evidence for associations with myeloproliferative diseases such as polycythaemia vera (HR = 1.7 [1.2–2.4], $P = 1.3 \times 10^{-3}$) (Supplementary Table 25).

Discussion

This population-based analysis of approximately 900,000 female participants with European and Asian ancestries indicates that detectable mLOX can be observed in a substantial fraction of middle-aged and older female participants, but typically affects less than 5% of circulating leukocytes. For non-genetic risk factors, we replicated prior mLOX associations with age and identified an association with tobacco smoking among high cell fraction mLOX. Our large sample size coupled with an improved mLOX detection approach enabled the identification of 56 common independent germline susceptibility signals across 42 loci and rare coding variations in *FBXO10* associated with mLOX. The mLOX germline susceptibility signals implicate genes involved in kinetochore and spindle function, blood cell measurements, cancer predisposition and immunity as aetiologically relevant to mLOX susceptibility. Little heterogeneity was noted in these loci across contributing studies or ancestry.

We identified shared and, more surprisingly, distinct genetic aetiologies of mLOX with mLOY, which occurs frequently in ageing male individuals—albeit at higher cell fractions. The two traits are moderately correlated genome-wide and 7 out of the 56 mLOX variants demonstrated evidence for shared effects for both mLOX and mLOY. Shared mLOX and mLOY variants were enriched for genes that are important for cancer susceptibility and blood cell traits; however, effects observed for mLOX were noticeably attenuated from effects observed for mLOY. This attenuation could be owing to differences in our ability to detect mLOX at lower cell fractions relative to mLOY or could be a biological impact since mLOX is often present at lower cell fractions relative to mLOY. Variants specific to mLOX demonstrated unique evidence for associations with immunity, including HLA alleles, which could have a role in the selection of X-linked cell surface antigens, in addition to genes relevant to mitotic missegregation (proposed mechanisms in Supplementary Fig. 15).

In addition to GWAS, we also performed allelic shift analyses on X chromosome germline variants to identify signals of *cis* clonal selection. These analyses identified strong independent signals of *cis* selection near the centromere as well as multiple additional signals spanning across the X chromosome. Interestingly, the majority of the allelic shift loci were not detected in the GWAS, demonstrating the ability to identify signals of selection by utilizing this approach. Although the centromeric signals for allelic shift were strongly associated with several blood cell phenotypes, their location near the centromere could tag germline variation with relevance for kinetochore formation and spindle attachment in this region and may predispose specific X chromosomes to missegregation errors, although, there is limited knowledge on how germline variation in DNA sequences could affect centrosomal protein binding and spindle formation[52,53]. Other loci identified by allelic shift analyses provide support for genes involved in escaping X inactivation, cancer susceptibility and blood cell traits being relevant to mLOX. Scores created that aggregate information across allelic shift loci correctly predicted which X chromosome was more likely to be retained in a high percentage of female participants with mLOX in which the difference in X chromosome scores was high. Thus, we have demonstrated the utility of a score that takes into account multiple germline variants to predict which chromosome will be affected if a somatic event occurs. Our approach for identifying variation important for X chromosome loss may be extendable to investigating other somatic events with relevance for cancer risk.

In conclusion, we provide evidence for a strong germline component of somatically occurring mLOX in which genes related to cancer susceptibility, blood cell traits, autoimmunity and chromosomal missegregation events are relevant to mLOX susceptibility. Further, we identify many strong *cis* effects for X chromosome loci that are associated with X chromosome retention and promotion of clonal expansion. Genetic insights from mLOX could also be relevant to better understanding skewed X inactivation, another commonly observed X chromosome abnormality in middle-aged and older female individuals.

Online content

Any methods, additional references, Nature Portfolio 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/s41586-024-07533-7.

1. Machiela, M. J. et al. Female chromosome X mosaicism is age-related and preferentially affects the inactivated X chromosome. *Nat. Commun.* **7**, 11843 (2016).
2. Zekavat, S. M. et al. Hematopoietic mosaic chromosomal alterations increase the risk for diverse types of infection. *Nat. Med.* **27**, 1012–1024 (2021).
3. Brown, C. J. et al. A gene from the region of the human X inactivation centre is expressed exclusively from the inactive X chromosome. *Nature* **349**, 38–44 (1991).
4. Lyon, M. F. Gene action in the X-chromosome of the mouse (*Mus musculus* L.). *Nature* **190**, 372–373 (1961).
5. Tukiainen, T. et al. Landscape of X chromosome inactivation across human tissues. *Nature* **550**, 244–248 (2017).
6. Busque, L. et al. Nonrandom X-inactivation patterns in normal females: lyonization ratios vary with age. *Blood* **88**, 59–65 (1996).
7. Gale, R. E. & Linch, D. C. Interpretation of X-chromosome inactivation patterns. *Blood* **84**, 2376–2378 (1994).
8. Zito, A. et al. Heritability of skewed X-inactivation in female twins is tissue-specific and associated with age. *Nat. Commun.* **10**, 5339 (2019).
9. Forsberg, L. A. et al. Mosaic loss of chromosome Y in peripheral blood is associated with shorter survival and higher risk of cancer. *Nat. Genet.* **46**, 624–628 (2014).
10. Dumanski, J. P. et al. Smoking is associated with mosaic loss of chromosome Y. *Science* **347**, 81–83 (2015).
11. Zhou, W. et al. Mosaic loss of chromosome Y is associated with common variation near TCL1A. *Nat. Genet.* **48**, 563–568 (2016).

12. 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).
13. Thompson, D. J. et al. Genetic predisposition to mosaic Y chromosome loss in blood. *Nature* **575**, 652–657 (2019).
14. Loh, P. R. et al. Insights into clonal haematopoiesis from 8,342 mosaic chromosomal alterations. *Nature* **559**, 350–355 (2018).
15. Lin, S. H. et al. Incident disease associations with mosaic chromosomal alterations on autosomes, X and Y chromosomes: insights from a phenome-wide association study in the UK Biobank. *Cell Biosci.* **11**, 1–11 (2021).
16. Zhou, W. et al. Detectable chromosome X mosaicism in males is rarely tolerated in peripheral leukocytes. *Sci. Rep.* **11**, 1193 (2021).
17. Sybert, V. P. & McCauley, E. Turner's syndrome. *N. Engl. J. Med.* **351**, 1227–1238 (2004).
18. Jäger, N. et al. Hypermutation of the inactive X chromosome is a frequent event in cancer. *Cell* **155**, 567–581 (2013).
19. Koren, A. & McCarroll, S. A. Random replication of the inactive X chromosome. *Genome Res.* **24**, 64–69 (2014).
20. Kessler, M. D. et al. Common and rare variant associations with clonal haematopoiesis phenotypes. *Nature* **612**, 301–309 (2022).
21. Terao, C. et al. GWAS of mosaic loss of chromosome Y highlights genetic effects on blood cell differentiation. *Nat. Commun.* **10**, 4719 (2019).
22. Kurki, M. I. et al. FinnGen provides genetic insights from a well-phenotyped isolated population. *Nature* **613**, 508–518 (2023).
23. Leitsalu, L. et al. Cohort profile: Estonian biobank of the Estonian genome center, University of Tartu. *Int. J. Epidemiol.* **44**, 1137–1147 (2015).
24. 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).
25. Bycroft, C. et al. The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203–209 (2018).
26. Michailidou, K. et al. Large-scale genotyping identifies 41 new loci associated with breast cancer risk. *Nat. Genet.* **45**, 353–361 (2013).
27. Michailidou, K. et al. Association analysis identifies 65 new breast cancer risk loci. *Nature* **551**, 92–94 (2017).
28. Gaziano, J. M. et al. Million Veteran Program: a mega-biobank to study genetic influences on health and disease. *J. Clin. Epidemiol.* **70**, 214–223 (2016).
29. Hunter-Zinck, H. et al. Genotyping array design and data quality control in the Million Veteran Program. *Am. J. Hum. Genet.* **106**, 535–548 (2020).
30. Karlson, E. W., Boutin, N. T., Hoffnagle, A. G. & Allen, N. L. Building the partners healthcare biobank at partners personalized medicine: informed consent, return of research results, recruitment lessons and operational considerations. *J. Pers. Med.* **6**, 2 (2016).
31. Boutin, N. T. et al. The evolution of a large biobank at Mass General Brigham. *J. Pers. Med.* **12**, 1323 (2022).
32. Machiela, M. et al. GWAS Explorer: an open-source tool to explore, visualize, and access GWAS summary statistics in the PLCO Atlas. *Sci. Data* **10**, 25 (2023).
33. Nagai, A. et al. Overview of the BioBank Japan project: study design and profile. *J. Epidemiol.* **27**, S2–S8 (2017).
34. Vlasschaert, C. et al. A practical approach to curate clonal hematopoiesis of indeterminate potential in human genetic datasets. *Blood* **141**, 2214–2223 (2023).
35. Vuckovic, D. et al. The polygenic and monogenic basis of blood traits and diseases. *Cell* **182**, 1214–1231 (2020).
36. Frampton, M. et al. Variation at 3p24. 1 and 6q23. 3 influences the risk of Hodgkin's lymphoma. *Nat. Commun.* **4**, 2549 (2013).
37. Berndt, S. I. et al. Meta-analysis of genome-wide association studies discovers multiple loci for chronic lymphocytic leukemia. *Nat. Commun.* **7**, 10933 (2016).
38. Celik, H. et al. JARID2 functions as a tumor suppressor in myeloid neoplasms by repressing self-renewal in hematopoietic progenitor cells. *Cancer Cell* **34**, 741–756 (2018).
39. Pattabiraman, D. R. & Gonda, T. J. Role and potential for therapeutic targeting of MYB in leukemia. *Leukemia* **27**, 269–277 (2013).
40. Schaffner, C., Stilgenbauer, S., Rappold, G. A., Döhner, H. & Lichter, P. Somatic ATM mutations indicate a pathogenic role of ATM in B-cell chronic lymphocytic leukemia. *Blood* **94**, 748–753 (1999).
41. Zenz, T. et al. *TP53* mutation and survival in chronic lymphocytic leukemia. *J. Clin. Oncol.* **28**, 4473–4479 (2010).
42. Catalano, A. et al. The *PRKAR1A* gene is fused to *RARA* in a new variant acute promyelocytic leukemia. *Blood* **110**, 4073–4076 (2007).
43. Loh, P. R., Genovese, G. & McCarroll, S. A. Monogenic and polygenic inheritance become instruments for clonal selection. *Nature* **584**, 136–141 (2020).
44. Luo, Y. et al. A high-resolution HLA reference panel capturing global population diversity enables multi-ancestry fine-mapping in HIV host response. *Nat. Genet.* **53**, 1504–1516 (2021).
45. Ritari, J., Koskela, S., Hyvärinen, K. & Partanen, J. HLA-disease association and pleiotropy landscape in over 235,000 Finns. *Hum. Immunol.* **83**, 391–398 (2022).
46. Bao, E. L. et al. Inherited myeloproliferative neoplasm risk affects haematopoietic stem cells. *Nature* **586**, 769–775 (2020).
47. Li, X. et al. Dynamic incorporation of multiple in silico functional annotations empowers rare variant association analysis of large whole-genome sequencing studies at scale. *Nat. Genet.* **52**, 969–983 (2020).
48. Zhou, W. et al. SAIGE-GENE+ improves the efficiency and accuracy of set-based rare variant association tests. *Nat. Genet.* **54**, 1466–1469 (2022).
49. Chiorazzi, M. et al. Related F-box proteins control cell death in *Caenorhabditis elegans* and human lymphoma. *Proc. Natl Acad. Sci. USA* **110**, 3943–3948 (2013).
50. Spielman, R. S., McGinnis, R. E. & Ewens, W. J. Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM). *Am. J. Hum. Genet.* **52**, 506 (1993).
51. Trubetskoy, V. et al. Mapping genomic loci implicates genes and synaptic biology in schizophrenia. *Nature* **604**, 502–508 (2022).
52. Yang, C. H., Tomkiel, J., Saitoh, H., Johnson, D. H. & Earnshaw, W. C. Identification of overlapping DNA-binding and centromere-targeting domains in the human kinetochore protein CENP-C. *Mol. Cell. Biol.* **16**, 3576–3586 (1996).
53. Du, Y., Topp, C. N. & Dawe, R. K. DNA binding of centromere protein C (CENPC) is stabilized by single-stranded RNA. *PLoS Genet.* **6**, e1000835 (2010).

FinnGen
Aoxing Liu[1,2,3,4,5,31], Mervi Aavikko[1], Timo P. Sipilä[1], Awaisa Ghazal[1], Andrea Ganna[1,2,4,5,32], Aarno Palotie[1,2,4,5] & Mark J. Daly[1,2,3,4,5]

Estonian Biobank Research Team
Georgi Hudjashov[16], Andres Metspalu[16], Tõnu Esko[16], Mari Nelis[16], Reedik Mägi[16] & Lili Milani[16]

Breast Cancer Association Consortium
Joe Dennis[19]

Million Veteran Program
Saiju Pyarajan[17,22]

[1]Institute for Molecular Medicine Finland (FIMM), HiLIFE, University of Helsinki, Helsinki, Finland. [2]Analytic and Translational Genetics Unit, Massachusetts General Hospital, Boston, MA, USA. [3]Center for Genomic Medicine, Massachusetts General Hospital, Boston, MA, USA. [4]Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, MA, USA. [5]Stanley Center for Psychiatric Research, Broad Institute of MIT and Harvard, Cambridge, MA, USA. [6]Department of Genetics, Harvard Medical School, Boston, MA, USA. [7]MRC Epidemiology Unit, Institute of Metabolic Science, University of Cambridge, Cambridge, UK. [8]Department of Public Health, University of Helsinki, Helsinki, Finland. [9]Department of Mathematics and Statistics, University of Helsinki, Helsinki, Finland. [10]Cardiovascular Research Center, Massachusetts General Hospital, Boston, MA, USA. [11]Department of Ophthalmology, Massachusetts Eye and Ear, Harvard Medical School, Boston, MA, USA. [12]Division of Cancer Epidemiology and Genetics, National Cancer Institute, Rockville, MD, USA. [13]Department of Medicine, Queen's University, Kingston, Ontario, Canada. [14]Laboratory for Statistical and Translational Genetics, RIKEN Center for Integrative Medical Sciences, Yokohama, Japan. [15]Cancer Prevention Fellowship Program, Division of Cancer Prevention, National Cancer Institute, Rockville, MD, USA. [16]Estonian Genome Centre, Institute of Genomics, University of Tartu, Tartu, Estonia. [17]Center for Data and Computational Sciences (C-DACS), VA Cooperative Studies Program, VA Boston Healthcare System, Boston, MA, USA. [18]Booz Allen Hamilton, McLean, VA, USA. [19]Centre for Cancer Genetic Epidemiology, Department of Public Health and Primary Care, University of Cambridge, Cambridge, UK. [20]Cancer Genomics Research Laboratory, Frederick National Laboratory for Cancer Research, Frederick, MD, USA. [21]Laboratory for Genotyping Development, RIKEN Center for Integrative Medical Sciences, Yokohama, Japan. [22]Department of Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, USA. [23]Institute of Molecular and Cell Biology, University of Tartu, Tartu, Estonia. [24]Division of Hematology/Oncology, Boston Children's Hospital, Harvard Medical School, Boston, MA, USA. [25]Department of Pediatric Oncology, Dana-Farber Cancer Institute, Harvard Medical School, Boston, MA, USA. [26]Anesthesia, Critical Care and Pain Medicine, Massachusetts General Hospital, Boston, MA, USA. [27]Division of Genetic Medicine, Department of Medicine, Vanderbilt University Medical Center, Nashville, TN, USA. [28]Clinical Research Center, Shizuoka General Hospital, Shizuoka, Japan. [29]Department of Applied Genetics, School of Pharmaceutical Sciences, University of Shizuoka, Shizuoka, Japan. [30]Center for Data Sciences, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, USA. [31]These authors contributed equally: Aoxing Liu, Giulio Genovese, Yajie Zhao. [32]These authors jointly supervised this work: Po-Ru Loh, Andrea Ganna, John R. B. Perry, Mitchell J. Machiela. ✉e-mail: liuaoxin@broadinstitute.org; giulio@broadinstitute.org; poruloh@broadinstitute.org; andrea.ganna@helsinki.fi; john.perry@mrc-epid.cam.ac.uk; mitchell.machiela@nih.gov

Methods

Definition of mLOX

Detection of mLOX events from SNP array data in eight biobanks. All DNA samples were obtained from peripheral leukocytes of female participants and typed with SNP arrays. mLOX analyses were restricted to individuals who are genetically female and have two copies of the X choromosome at birth. The median (s.d.) age at sample collection for genotyping ranged from 44 (16.3) for EBB to 65 (15.8) for BBJ. The calling of mLOX was performed using the MoChA pipeline (https://github.com/freeseek/mochawdl), with GRCh38 assembly as the reference genome build. The mLOX detection ability is related to X chromosome probe density, missing genotype frequency, clarity of raw probe intensity signals, and phasing accuracy – all of which can be linked to the molecular approach and number of X chromosome probes on the genotyping platform used by each biobank for genotyping. As such, the MoChA pipeline was run separately within each biobank, and biobank results were then meta-analysed for all association analyses to avoid potential cohort effects, except where noted.

The raw genotyping array signal intensities of each variant were first transformed to B allele frequency (BAF) (relative intensity of the B allele) and log *R* ratio (LRR) (total intensity of both alleles). Then, haplotype phasing was performed using SHAPEIT4[54] across all batches of a biobank, except for BBJ and BCAC, for which phasing was done separately within each biobank sub-cohort (for BBJ, 4 sub-cohorts, with cohort sizes ranging from 3,888 to 45,877; for BCAC, 2 sub-cohorts of breast cancer cases and controls by genotyping array platform, with cohort size of 72,145 and 105,177). Utilizing long-range haplotype phasing can improve the sensitivity of detecting large mosaic events with low cell fractions[14], which is characteristic of mLOX. To avoid issues with phasing and the subsequent mLOX calling, we excluded variants with poor genotyping quality such as segmental duplications with low divergence (<2%) and SNPs with high levels of missingness (>3%) or heterozygote excess ($P < 1.0 \times 10^{-6}$). Finally, the calling of mLOX events was performed within each batch based on the imbalance of phased BAF of heterozygous sites over the whole X chromosome. To filter out 47,XXY and 47,XXX samples, we restricted to X chromosome events with estimated ploidy less than 2.5, where the estimated ploidy is estimated by first computing the median LRR across the assayed X chromosome SNPs and then by computing the value $2^{1+(\mathrm{LRR}/\mathrm{LRR\text{-}hap2dip})}$ with LRR-hap2dip (the difference between LRR for haploid and diploid) set at 0.45 by default. We further removed events with length <100 Mb to exclude partial X chromosome loss (for example, 2.0% in FinnGen) as they might be caused by different mechanisms compared to the major type of full mLOX events. For each mLOX event that passed quality control, the fraction of cells (cf) with X loss was calculated as $4 \times \mathrm{bdev}/(1 + 2 \times \mathrm{bdev})$, where bdev is the estimated BAF deviation of heterozygous sites.

The 2022-01-14 version of MoChA was used to detect the dichotomous mLOX status for all biobanks, except for BBJ (versions 2021-08-17 and 2021-09-07) and BCAC (version 2022-12-21). The priors of MoChA have been updated since version 2021-05-14 to improve the detection of low cell fraction mLOX calls, and thus, the biobanks that used the updated MoChA pipeline (all biobanks that contributed to this study) are expected to yield higher age-adjusted mLOX frequencies than those that used the previous version. For BCAC, we included both those diagnosed as breast cancer cases ($N = 99{,}043$) and cancer-free controls ($N = 78{,}279$) in the analyses. A brief description of each contributed biobank (for example, continental ancestry, sample size, age structures and SNP array) is available in Supplementary Table 1.

Estimation of X chromosome dosages from UKBB WES data. For UKBB, the WES data was released in late 2021[55], which enabled identification of X loss from sequencing allelic dosage data in combination with array data. The relative X chromosome dosage at the individual level was estimated following the steps described previously[56]. In brief, we first generated mean coverages from the original WES data for variants on the autosomes and the X chromosome non-pseudoautosomal regions, separately; then, we obtained the relative X chromosome dosage by adjusting for the mean coverage of autosomes. Therefore, for UKBB, three ways were available to define the mLOX phenotype, including the dichotomous mLOX status derived from the phased BAF method (by MoChA) and two quantitative measures employing either median $\log_2$ R ratio (mLRR) from SNP array data or allele dosage from WES data. To assess the performances of the three mLOX measures in UKBB, we compared either mLRR or X dosage between the case and the control groups defined by MoChA (Supplementary Fig. 2a–c). As shown in Supplementary Fig. 2b,c, the participants identified as mLOX cases by MoChA exhibited lower mLRR (ANOVA test, $P = 1.5 \times 10^{-5}$) and X dosage value ($P < 1.0 \times 10^{-250}$) than mLOX controls. Then, for mLOX cases, we examined the relationships between three measures representing the extent of mosaicism (Supplementary Fig. 2d–f), including cell fraction (from MoChA), mLRR and X dosage. Overall, significant correlations were observed across the 3 measures, with the absolute Pearson correlation coefficient ranging from 0.42 between mLRR and X dosage to 0.86 between mLOX cell fraction and X dosage. Again, given that mLRR is a noisier measure than X dosage, for mLOX cell fraction, a stronger correlation was observed with X dosage ($r = -0.86$) than with mLRR (−0.48).

Enhanced 3-way combined mLOX calls in UKBB. In addition to the dichotomous mLOX status defined by the phased BAF method, for UKBB, we proposed a new quantitative measure by combining the 3 methods of mLOX calling for UKBB, that is, the mLOX combined call (3-way) = mLOX status + 2 × cf − 2 × mLRR − 4 × (dosage − 2) (cropped to the range [0,2]). The intuition behind this newly proposed measure was to emphasize mLOX cases with larger cell fractions (similar to the strategy used by a recent mLOY study[57]) while obtaining enhanced mLOX calls from integrating independent information of both SNP array and WES data. As not all participants with SNP array data had WES data available, we imputed the missing 3-way mLOX combined calls with 2-way combined calls, defined as mLOX status + 3 × cf − 3 × mLRR (also cropped to the range [0,2]). As age is strongly associated with mLOX, we evaluated the age–mLOX association for MoChA calls versus the enhanced 3-way combined mLOX calls. Compared to the dichotomous mLOX status derived from MoChA, the *t*-test statistic for association with age was increased by 29.2% when using the 3-way combined calls, suggesting increased power to detect mLOX. Enhanced 3-way combined mLOX calls were used for UKBB in the GWAS meta-analysis and the exome-wide rare variant gene burden test.

Environmental determinants and epidemiological consequences

To investigate the effect of lifestyle factors on the odds of acquiring mLOX, we assessed the associations between smoking and BMI with mLOX in the FinnGen cohort. In FinnGen data freeze 9, 50.3% of female participants had smoking status ($N = 84{,}926$) and 18.4% had measurements for BMI ($N = 31{,}101$) recorded at enrolment. We applied a logistic regression model adjusting for age (at genotyping), age^2, and the first ten principal components as covariates. As sensitivity analyses, we restricted the analyses to expanded mLOX calls having cf >5%. Given that we identified a significant association between ever-smoking and expanded mLOX, we further adjusted for ever-smoking status when assessing the effect of BMI on mLOX. To examine whether the environmental determinants were shared or distinct between mLOX in female participants and mLOY in male participants, we also extended the association analyses to mLOY ($N = 76{,}808$ for smoking, $N = 33{,}668$ for BMI). To validate our findings identified from FinnGen, we performed the same analyses for smoking ($N = 241{,}761$) and BMI ($N = 242{,}024$) in UKBB.

To assess the clinical consequences of acquiring mLOX, we performed a Cox proportional hazards regression for incident cases in FinnGen, UKBB, MVP and MGB independently, with time on study as the time scale. For covariates, we recommended each biobank adjust for age, age[2], smoking and the first ten principal components. Meta-analysis across four biobanks was carried out with a fixed-effect model applied in the meta package[58]. For each disease, we applied Cochran's *Q*-test to assess heterogeneity across biobanks with different healthcare systems. In total, we examined 1,253 phecodes covering 13 disease categories. Accordingly, the multiple-testing corrected *P* value threshold was set to $P < 4.0 \times 10^{-5}$. In the main analysis, we used all detectable mLOX calls without restriction for cell fraction. For a sensitivity analysis, we considered mLOX having cf >10% as expanded calls, following the definition used by Zekavat et al.[2].

To further understand the phenotypic associations for mLOX, we applied a linear regression model adjusting for age, age[2], smoking and the first ten principal components as covariates for a broad range of representative quantitative traits across anthropometry, reproductive health, lung function, blood cell parameters, blood biomarkers, urine biomarkers, cognitive function, and telomere length using the data from UKBB. The same analyses were performed for all detectable mLOX calls without restriction for cf as well as for expanded calls having cf >10%.

Common and rare germline variants associated with detectable mLOX susceptibility

GWAS of dichotomous mLOX status in eight contributed biobanks. To identify common germline variants (MAF > 0.1%) associated with risk of detectable mLOX in peripheral leukocytes, we performed a GWAS on chromosomes 1–22 and the X chromosome in each of eight contributing biobanks independently, for a total of 883,574 female participants. For the dichotomous mLOX status (derived from MoChA), GWAS was conducted for FinnGen and BCAC using the scalable and accurate implementation of generalized mixed model (SAIGE)[59] and for the other six biobanks (including UKBB) using REGENIE[60] applied in the assoc.wdl pipeline (part of the MoChA pipeline; https://github.com/freeseek/mochawdl). Both SAIGE and REGENIE are feasible to account for sample relatedness and extreme case–control imbalances of a dichotomous phenotype. For covariates, each biobank adjusted for age (at genotyping), age[2], and the first 20 genetic principal components. The effective sample size, presented in Extended Data Table 1, was calculated as $(4 \times N_{\text{case}} \times N_{\text{control}})/(N_{\text{case}} + N_{\text{control}})$.

GWAS of 3-way combined quantitative mLOX measure in UKBB. For UKBB, to improve the power of GWAS, we used the new quantitative measure that combined the three ways of mLOX calling. For the proposed quantitative mLOX measure, GWAS was performed with the linear mixed model applied in BOLT-LMM[61].

GWAS meta-analysis. For each contributed biobank, we filtered out variants with MAF < 0.1% or imputation INFO score <0.6. We also inspected allele frequencies of each biobank versus Genome Aggregation Database (gnomAD) 3.0 as well as the relationship between standard errors and effective sample sizes across biobanks, as applied by the COVID-19 Host Genetics Initiative meta-analysis[62]. Given that no biobank deviated from the expected pattern, we conducted meta-analyses across biobanks. In addition to the dichotomous mLOX measure used by all biobanks, UKBB was able to run GWAS with an additional quantitative measure that combined information of three ways of mLOX calling and thus was expected to yield increased power in GWAS. Depending on which mLOX measure was used in the UKBB GWAS, we applied two fixed-effect meta-analysis models accordingly. When using the dichotomous measure, we applied the inverse variance weighting (IVW) method, which weighted the effect size estimated from an individual biobank by its inverse variance. When UKBB used the 3-way combined measure as the GWAS phenotype, we employed the weighted *z*-score method (weighted by the square root of the effective sample size) applied in the METAL (v.2011-03-25) software[63], which can manage the different units of dichotomous and quantitative measures. As the main analysis, we meta-analysed summary statistics across all eight biobanks regardless of ancestry and applied Cochran's *Q*-test to assess the heterogeneity. To further investigate the effect of ancestry, we also conducted a meta-analysis for seven biobanks containing only participants with European ancestry (excluding BBJ participants with East Asian ancestry).

Independent loci identification and gene prioritization. To identify independent signals and prioritize candidate causal genes, we applied the GWAStoGenes pipeline for variants presented in at least half of the contributed biobanks. In brief, primary independent signals associated with mLOX susceptibility at a genome-wide significance level ($P < 5 \times 10^{-8}$) were initially selected in 2-Mb windows[64] (spanning a ±1-Mb region around the most significant variant). Secondary independent signals were identified by using an approximate conditional analysis applied in GCTA[64], with LD structures constructed from UKBB samples. Secondary signals were only considered if they were genome-wide significant, in low LD ($r^2 < 0.05$) with primary signals, and having association statistics unchanged with the conditional analysis. We also excluded variants without any nearby genes (within 500 kb) documented in the NCBI RefSeq dataset[65]. In total, we identified 56 independent common susceptibility variants across 42 loci.

Candidate genes were prioritized using the following criteria and scored by their strength of evidence for causality. First, signals were annotated with their physically closest genes. Second, signals and their closely linked variants ($r^2 > 0.8$) were annotated if they were predicted deleterious coding variants, or if the paired genes exhibited a gene-level association when collapsing all predicted deleterious coding variants within a gene using multi-marker analysis of genomic annotation (MAGMA)[66]. Third, non-coding signals and closely linked variants were then annotated if they could be mapped to known enhancers using the activity-by-contact enhancer maps[67], but restricted to available cells and tissue types where each gene was actively expressed. Fourth, colocalization between GWAS and expression quantitative trait locus (eQTL) data was performed using the summary data-based Mendelian randomization (SMR) and heterogeneity in dependent instruments (HEIDI) test (version 0.68)[68] and the approximate Bayes factor method applied in the coloc package (version 5.1.0)[69]. These two tools were used in conjunction, as using a combination of colocalization methods has been shown to outperform single approaches[70]. To identify tissues exhibiting a significant genome-wide enrichment, we used LD score regression applied to specifically expressed gene (LDSC-SEG)[71] approach, with eQTL datasets from cross-tissue meta-analysed GTEx eQTL v.7[72], eQTLGen[73] and Brain-eMeta[74]. The same set of analyses were also applied to a protein quantitative trait locus (pQTL) dataset[75]. Finally, by integrating GWAS summary statistics with data from gene expression, biological pathway, and predicted protein–protein interaction, candidate genes were identified using the gene-level polygenic priority score (PoPS) method[76].

Independent loci in UKBB with different mLOX measures. Among the 56 mLOX susceptibility variants identified from the GWAS meta-analysis, in UKBB, 47 out of 55 (85%, one missing in UKBB) have a lower *P* value when using the enhanced 3-way combined mLOX calling method compared to the standard MoChA calling method, suggesting the enhanced 3-way combined approach is recommended for mLOX detection when WES data is available. We noted that the meta-analysis signals might favour the 3-way combined measure over the binary MoChA calls given the 3-way combined calls were used for UKBB in the GWAS meta-analysis.

Gene burden test for rare variants causing detectable mLOX. To identify rare germline variants (MAF < 0.1%) associated with the risk of detectable mLOX, we performed gene burden tests on chromosomes

1–22 and the X chromosome in 226,125 UKBB female participants with WES data available. We performed WES data pre-processing and quality control following Gardner et al.[77]. We annotated variants using the ENSEMBL Variant Effect Predictor (VEP) v104[78] and defined protein-truncating variants (PTVs) as high-confidence (HC, as defined by LOFTEE) stop gained, splice donor/acceptor, and frameshift consequences. We then utilized CADDv1.6 to score a variant based on its predicted deleteriousness[79]. Only non-synonymous variants with MAF < 0.1% were included in the analysis. As the main analysis, we used BOLT-LMM[59] to perform the gene burden test. For each gene, we defined individuals with high-confidence PTVs, missense variants with CADD scores ≥ 25 (MISS_CADD25), and damaging variants (HC_PTV + MISS_CADD25) (DMG) as carriers. Then, carriers with non-synonymous variants were defined as heterozygous and non-carriers as homozygous. For covariates, we adjusted for age, age^2, batches and the first ten principal components. We further excluded the genes with less than 50 non-synonymous variant carriers for each setting, resulting in 8,702 genes for HC_PTV, 15,144 for MISS_CADD25 and 16,493 for DMG, for a total of 40,339 genes. Accordingly, the Bonferroni corrected exome-wide significant threshold was set to $0.05/40{,}339 = 1.2 \times 10^{-6}$. To avoid the identified association dominated by a single variant, as sensitivity analysis, we conducted a leave-one-out analysis using a generalized linear model for each significant gene. In addition, we reproduced the associations detected by BOLT-LMM[61] with STAAR (variant-set test for association using annotation information)[47] and SAIGE_GENE+ (scalable generalized mixed-model region-based association test plus)[48] to address the potential case–control imbalance issue.

Pathway and gene set analysis. To identify gene sets enriched in the same biological process, we performed pathway-based analysis using the summary data-based adaptive rank truncated product (sARTP) method[80]. We used summary statistics from meta-analysis of seven biobanks of European ancestry (without BBJ) and linkage disequilibrium (LD) structures constructed from European ancestry samples of the 1000 Genomes project[81]. We considered a total of 6,285 gene sets available in GSEA (https://www.gseamsigdb.org/gsea/msigdb/). Accordingly, the Bonferroni corrected *P* value was set to $0.05/6{,}285 = 8.0 \times 10^{-6}$.

Genetic correlation. To investigate whether there are traits that are genetically correlated with mLOX susceptibility, we estimated genetic correlations between mLOX and 60 phenotypes (including both major diseases and blood cell phenotypes) using LD score regression (LDSC)[82]. For LDSC, we used HapMap3[83] SNPs and LD structures constructed from 1000 Genomes project[81] samples of European ancestry.

Per-chromosome heritability. To examine whether the observed heritability for each chromosome was proportional to chromosome length, we estimated per-chromosome heritability for 3-way combined mLOX measure in UKBB using BOLT-REML[84]. Given the large associations of HLA genes, we further examined how heritability explained by chromosome 6 changed after excluding variants from the extended MHC region (GRCh38: chr. 6: 25.7–33.4 Mb).

Shared and distinct mechanisms between mLOX in female participants and mLOY in male participants

Bayesian models to cluster variants by effects on mLOX and mLOY. We utilized a Bayesian line model framework (https://github.com/mjpirinen/linemodels) to assign each of the 56 independent common variants identified from mLOX GWAS and 147 variants (9 variants dropped due to missing in mLOX GWAS) from the published mLOY GWAS[13] into 3 groups: specific to mLOX, specific to mLOY, and shared between mLOX and mLOY. In general, each variant was fitted into the model separately and assigned to a specific group mainly based on its estimated effect sizes on mLOX and mLOY (variances of the estimates were considered as well to capture the uncertainty, but not for directly deciding the group) rather than *P* values or effective sample sizes of the GWAS discovery populations. The slopes of the line models were set to 0 for the group of variants specific for mLOY and infinite for variants specific for mLOX. For variants shared between mLOX and mLOY, the slope was set to 0.3, based on the effects of four variants (rs568868093, rs381500, rs2280548, rs78378222) that were genome-wide significant in both mLOX GWAS and mLOY GWAS. For all three line models, the prior s.d. determining the magnitude of the effects was set to 0.15 and the correlation parameter determining the allowed deviations from the lines to 0.995. The correlation between mLOX and mLOY GWAS statistics was set to 0 given that there was no overlap between samples used in the two GWAS. We assumed a uniform prior for the three models and obtained the posterior probabilities for each data point separately within a Bayesian framework. Probability assignment threshold was set to 95%.

Fine-mapping of HLA alleles in FinnGen. Given the large associations with mLOX and the high polymorphism of HLA genes, we fine-mapped HLA alleles at a unique protein sequence level in the FinnGen cohort. In FinnGen data freeze 9, a total of 172 HLA alleles of 10 transplantation genes were imputed using a Finnish-specific reference panel, as described in Ritari et al.[85]. We conducted the association analysis between each imputed HLA allele and the dichotomous mLOX status in 168,838 Finnish female participants (27,001 cases) using a multivariate logistic regression model, considering age, age^2 and the first 10 principal components as covariates. Only HLA alleles with more than five mLOX cases carrying the minor alleles were included in the analysis. Ultimately, we considered 156 HLA alleles for mLOX, including 18 alleles for *HLA-A*, 36 for *HLA-B*, 20 for *HLA-C*, 29 for *HLA-DRB1*, 14 for *HLA-DQA1*, 14 for *HLA-DQB1*, 18 for *HLA-DPB1*, 3 for *HLA-DRB3*, and 2 each for *HLA-DRB4* and *HLA-DRB5*. To identify independent HLA alleles, a stepwise conditional analysis was performed with each step adding the most significant HLA allele obtained from the previous step as an additional covariate, until no HLA allele can reach the significant threshold. To examine whether the HLA associations are shared with other types of mCAs, we extended the HLA fine-mapping analyses to mLOY in male participants (total, $N = 128{,}729$; cases, $N = 45{,}675$) for 157 HLA alleles (including HLA-A*02:02 compared to the 156 alleles used by mLOX association analyses) and for autosomal mCAs in both sexes (total, $N = 297{,}567$; cases, $N = 9{,}302$) for 155 HLA alleles (missing HLA-C*15:05 compared to the 156 alleles used by mLOX association analyses).

Allelic shift analysis for *cis* clonal selection of X chromosome alleles

Allelic shift analysis. Conditional on mLOX having been detected, for each variant on the X chromosome we tested whether there is a propensity for X chromosomes with a given allele to be identified as lost more often than X chromosomes with the other allele. Similar to a transmission disequilibrium test[50], this test is robust to the presence of population structure. Rather than measuring the over-transmission of an allele from heterozygous parents to offspring, we measured the propensity of alleles to be on the retained X chromosome homologue. Therefore, we carried out a binomial test for each variant with a sample size equal to the number of female participants with detected mLOX who were heterozygous for that variant, with no need to correct for covariates or relatedness. Given the large number of X chromosome signals observed from the allelic shift analysis, we inspected whether variant density may have contributed to the signals. We hypothesized that if the signals were random, then the number of variants being significant would be related to the number of variants being examined in that region. We therefore checked the number of variants per 1-kb region across the whole X chromosome.

Identification of independent loci. Given the complexity of LD structures for X chromosomes especially for centromere and pseudoautosomal regions, we defined index variants by iteratively spanning the ±500-kb region around the most significant variant until no further variants reached a genome-wide significant level ($P < 5.0 \times 10^{-8}$). Then, we calculated LD between every 2 index variants and kept the variant with a lower P value if a pair of index variants with $r^2 > 0.1$.

Polygenic score to predict the retained X chromosome. To assess how well the identified allelic shift signals can predict which X chromosome is retained when mLOX occurs, we constructed polygenic scores (PGSs) in FinnGen mLOX cases (N = 27,001). In brief, we extracted the effect size for 44 independent loci from the allelic shift analysis of 6 biobanks excluding FinnGen. Given that MoChA was able to detect which alleles were lost at heterozygous sites, for each mLOX case, we computed the PGS for the retained X chromosome ($\mathrm{PGS}_{\mathrm{retained}}$) and the lost X chromosome ($\mathrm{PGS}_{\mathrm{lost}}$) separately and obtained the difference in PGS between the two X chromosomes ($\mathrm{PGS}_{\mathrm{diff}} = \mathrm{PGS}_{\mathrm{lost}} - \mathrm{PGS}_{\mathrm{retained}}$). A negative $\mathrm{PGS}_{\mathrm{diff}}$ indicates that the retained X chromosome of the mLOX case was correctly predicted.

To assess the upper limit of prediction performance for the proposed PGS, we performed simulation analyses in FinnGen mLOX cases. We first simulated genotypes for the 44 loci we identified as independently associated using the allele frequency calculated from the biobank meta-analysis (weighted by the effective sample size of each contributing biobank) and assuming all genotypes were independent (that is, $r^2 = 0$). For a given FinnGen female sample and each one of the 44 loci, we defined OR_i as the odds ratio between the likelihood of losing the paternal X chromosome and the likelihood of losing the maternal X chromosome, as inferred by the meta-analysis and with $\mathrm{OR}_i = 1$ when the ith locus is homozygous. We then defined the X chromosome differential score $\mathrm{PGS}_{\mathrm{diff}}$ with the equation: $\mathrm{PGS}_{\mathrm{diff}} = \Sigma_i \log(\mathrm{OR}_i) = \Sigma_i$ heterozygous $\log(\mathrm{OR}_i)$, by aggregating variant effects at all simulated heterozygous genotypes. Assuming that $\mathrm{PGS}_{\mathrm{diff}}$ is positive (negative), we estimated the probability P of the paternal (maternal) X chromosome being lost using the logistic function for $|\mathrm{PGS}_{\mathrm{diff}}|$, with $P = P/(1{-}P + P) = P/(1 - P)/(1 + P/(1 - P)) = \prod_i \mathrm{OR}_i/(1 + \prod_i \mathrm{OR}_i) = \exp(|\mathrm{PGS}_{\mathrm{diff}}|)/(1 + \exp(|\mathrm{PGS}_{\mathrm{diff}}|))$. Given an estimated $|\mathrm{PGS}_{\mathrm{diff}}|$, we think of P, with $0.5 \leq P < 1$, as the probability of inferring which X chromosome was lost conditional on one X chromosome being lost, that is, our prediction accuracy. As we independently simulated genotypes without modelling LD and variant effects without assuming possible interactions, we expected the simulation to overestimate the prediction accuracy from real data and to effectively estimate a best-case scenario for how predictive our proposed PGS could be.

Lifetime disease risk for female participants with a high X differential score. We then evaluated whether female participants carrying higher X differential scores would have an elevated lifetime disease risk by examining the association between the score and 1,630 disease endpoints in 27,001 FinnGen mLOX cases (FinnGen data freeze 9). In FinnGen, disease endpoints were defined by a clinical expert group by harmonizing International Classification of Diseases (ICD) codes of version 8 (1968–1986), 9 (1987–1995) and 10 (1996–) archived in nationwide healthcare registers[22]. Given that the nature of our proposed X differential score is a PGS, it reflects the germline risk an individual acquires at birth. Therefore, we performed a Cox proportional hazards regression model considering the chronological age as the time scale, with the follow-up time starting from birth rather than the age at genotyping, and censoring at disease onset, death, or the end of follow-up, whichever occurs first. For covariates, similar to the epidemiological association analyses we performed for the dichotomous mLOX status, we considered genotyping age, age^2, smoking and the top ten principal components.

Ethics statement

Patients and control subjects in FinnGen provided informed consent for biobank research, based on the Finnish Biobank Act. Alternatively, separate research cohorts, collected prior the Finnish Biobank Act came into effect (in September 2013) and start of FinnGen (August 2017), were collected based on study-specific consents and later transferred to the Finnish biobanks after approval by Fimea (Finnish Medicines Agency), the National Supervisory Authority for Welfare and Health. Recruitment protocols followed the biobank protocols approved by Fimea. The Coordinating Ethics Committee of the Hospital District of Helsinki and Uusimaa (HUS) statement number for the FinnGen study is Nr HUS/990/2017. The FinnGen study is approved by Finnish Institute for Health and Welfare (permit numbers: THL/2031/6.02.00/2017, THL/1101/5.05.00/2017, THL/341/6.02.00/2018, THL/2222/6.02.00/2018, THL/283/6.02.00/2019, THL/1721/5.05.00/2019 and THL/1524/5.05.00/2020), Digital and Population Data Service Agency (permit numbers: VRK43431/2017-3, VRK/6909/2018-3, VRK/4415/2019-3), the Social Insurance Institution (permit numbers: KELA 58/522/2017, KELA 131 /522/2018, KELA 70/522/2019, KELA 98/522/2019, KELA 134/522/2019, KELA 138/522/2019, KELA 2/522/2020, KELA 16/522/2020), Findata permit numbers THL/2364/14.02/2020, THL/4055/14.06.00/2020,THL/3433/14.06.00/2020, THL/4432/14.06/2020, THL/5189/14.06/2020, THL/5894/14.06.00/2020, THL/6619/14.06.00/2020, THL/209/14.06.00/2021, THL/688/14.06.00/2021, THL/1284/14.06.00/2021, THL/1965/14.06.00/2021, THL/5546/14.02.00/2020, THL/2658/14.06.00/2021, THL/4235/14.06.00/202, Statistics Finland (permit numbers: TK-53-1041-17 and TK/143/07.03.00/2020 (earlier TK-53-90-20) TK/1735/07.03.00/2021, TK/3112/07.03.00/2021) and Finnish Registry for Kidney Diseases permission/extract from the meeting minutes on 4 July 2019. The Biobank Access Decisions for FinnGen samples and data utilized in FinnGen Data Freeze 9 include: THL Biobank BB2017_55, BB2017_111, BB2018_19, BB_2018_34, BB_2018_67, BB2018_71, BB2019_7, BB2019_8, BB2019_26, BB2020_1, Finnish Red Cross Blood Service Biobank 7.12.2017, Helsinki Biobank HUS/359/2017, HUS/248/2020, Auria Biobank AB17-5154 and amendment 1 (17 August 2020), AB20-5926 and amendment 1 (23 April 2020) and its modification (22 September 2021), Biobank Borealis of Northern Finland_2017_1013, Biobank of Eastern Finland 1186/2018 and amendment 22 § /2020, Finnish Clinical Biobank Tampere MH0004 and amendments (21.02.2020 and 06.10.2020), Central Finland Biobank 1-2017, and Terveystalo Biobank STB 2018001 and amendment 25 Aug 2020. The UKBB analyses were conducted using applications 7089, 9905 and 21552. The activities of the EBB are regulated by the Human Genes Research Act, which was adopted in 2000 specifically for the operations of the EBB. Individual level data analysis in the EBB was carried out under ethical approval 1.1-12/624 from the Estonian Committee on Bioethics and Human Research (Estonian Ministry of Social Affairs), using data according to release application N05 from the EBB. For BCAC, the detailed ethics statement is available in the Supplementary Information.

Reporting summary

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

Data availability

Overall and population-level GWAS summary statistics generated from the mLOX meta-analysis are available on the GWAS catalogue (accession numbers GCST90328147, GCST90328148, GCST90328149 and GCST90328150). Requests for access to individual-level data differ for each contributing biobank. For FinnGen, researchers can apply for health data from the Finnish Data Authority Findata (https://findata.fi/en/permits/) and individual-level genotype data available through the Fingenious portal (https://site.fingenious.fi/en/). These resources are hosted by the Finnish Biobank Cooperative FINBB (https://finbb.fi/en/). Access can only be provided for research projects within the scope of the Finnish Biobank Act, which includes health promotion, understanding disease mechanisms or developing medical products or treatment practices. For EBB, individual-level health, lifestyle, demographic and genetic data are anonymized and available for research projects. Data sharing is conducted in accordance with the regulations of the Estonian Genome Center of the University of Tartu (HGRA). A data application form can be found at https://www.biobank.ee. The research project has to obtain approval from the Ethics Review Committee on Human Research of the University of Tartu as well as approval from the EGCUT scientific committee. For UKBB, all individual-level data used in the analysis is available by application to the UKBB Access Management System (https://www.ukbiobank.ac.uk). Approved researchers can submit applications for review and assessments are made to determine if the research proposal qualifies as health-related research in line with public interest. For BCAC, data for some of the samples are available on dbGAP (https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs001265.v1.p1). Requests for BCAC data can be made to the Data Access Coordination Committee (DACC) of BCAC (http://bcac.ccge.medschl.cam.ac.uk/bcacdata/). BCAC DACC approval is required to access individual-level phenotype and genotype data from the ABCFS, ABCS, ABCTB, BBCC, BBCS, BCEES, BCFR-NY, BCFR-PA, BCFR-UTAH, BCINIS, BIGGS, BREOGAN, BSUCH, CBCS, CCGP, CECILE, CGPS, CNIO-BCS, CPSII, CTS, EPIC, DIETCOMP LYF, ESTHER, GC-HBOC, GENICA, HABCS, HCSC, HEBCS, HMBCS, HUBCS, KARBAC, KARMA, KBCP, KCONFAB/AOCS, LMBC, MABCS, MARIE, MBCSG, MCBCS, MCCS, MEC, MISS, MMHS, MTLGEBCS, NBCS, NC-BCFR, NBHS, NCBCS, NHS, NHS2, OBCS, OFBCR, ORIGO, PBCS, PKARMA, PLCO, POSH, RBCS, SASBAC, SBCS, SEARCH, SISTER, SKKD KFZS, SMC, SZBCS, UCIBCS, UKBGS, UKOPS and USRT studies. For MVP, summary statistics are available on dbGaP under the MVP accession number phs001672. Additional data supporting the findings of this study are available upon reasonable request from MVP. These data are not publicly available due to restrictions of the US Government and Department of Veterans Affairs concerning privacy and participant consent. For MGB, a portion of individual-level genomic data are available in dbGAP as part of the eMERGE consortium (phs001584.v2.p2) and as part of the Center Common Disease Genomics (phs002018.v1.p1). Additional MGB data are not currently publicly available due to data restrictions. For PLCO, individual-level genotype data is available in dbGaP (phs001286.v2.p2). Permitted data use includes discovery and hypothesis generation in the investigation of genetic contributions to cancer risk and risk of other diseases as well as development of novel analytical approaches for GWAS. Individual-level phenotype data can be requested through the NCI Cancer Data Access System (CDAS) (https://cdas.cancer.gov/plco/). For BBJ, information on the cohort is available at the RIKEN website (http://jenger.riken.jp/en/). While individual-level genetic data are not accessible, all other individual-level data are available upon request.

Code availability

The MoChA pipelines used for mLOX calling (mocha.wdl), GWAS (assoc.wdl), allelic shift analysis (impute.wdl and shift.wdl) and X chromosome differential score estimation (score.wdl) are available at https://doi.org/10.5281/zenodo.10892520[86] (please see the detailed and most updated version at https://github.com/freeseek/mochawdl). The GWAS meta-analysis was performed by using the pipeline developed by the COVID-19 Host Genetics Initiative, available at https://github.com/covid19-hg/META_ANALYSIS. The codes used for the Bayesian line model are available at https://github.com/dsgelab/Mosaic-loss-of-chromosome-X/tree/main/BayesLineModel.

54. Delaneau, O., Zagury, J. F., Robinson, M. R., Marchini, J. L. & Dermitzakis, E. T. Accurate, scalable and integrative haplotype estimation. *Nat. Commun.* **10**, 5436 (2019).
55. Backman, J. D. et al. Exome sequencing and analysis of 454,787 UK Biobank participants. *Nature* **599**, 628–634 (2021).
56. Zhao, Y. et al. Detection and characterization of male sex chromosome abnormalities in the UK Biobank study. *Genet. Med.* **24**, 1909–1919 (2022).
57. Zhao, Y. et al. *GIGYF1* loss of function is associated with clonal mosaicism and adverse metabolic health. *Nat. Commun.* **12**, 4178 (2021).
58. Balduzzi, S., Rücker, G. & Schwarzer, G. How to perform a meta-analysis with R: a practical tutorial. *Evid. Based Ment. Health* **22**, 153–160 (2019).
59. Zhou, W. et al. Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies. *Nat. Genet.* **50**, 1335–1341 (2018).
60. Mbatchou, J. et al. Computationally efficient whole-genome regression for quantitative and binary traits. *Nat. Genet.* **53**, 1097–1103 (2021).
61. Loh, P. R. et al. Efficient Bayesian mixed-model analysis increases association power in large cohorts. *Nat. Genet.* **47**, 284–290 (2015A).
62. COVID-19 Host Genetics Initiative. Mapping the human genetic architecture of COVID-19. *Nature* **600**, 472–477 (2021).
63. Willer, C. J., Li, Y. & Abecasis, G. R. METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics* **26**, 2190–2191 (2010).
64. Yang, J. et al. Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. *Nat. Genet.* **44**, 369–375 (2012).
65. O'Leary, N. A. et al. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.* **44**, D733–D745 (2016).
66. de Leeuw, C. A., Mooij, J. M., Heskes, T. & Posthuma, D. MAGMA: generalized gene-set analysis of GWAS data. *PLoS Comput. Biol.* **11**, e1004219 (2015).
67. Nasser, J. et al. Genome-wide enhancer maps link risk variants to disease genes. *Nature* **593**, 238–243 (2021).
68. Zhu, Z. et al. Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. *Nat. Genet.* **48**, 481–487 (2016).
69. Giambartolomei, C. et al. Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. *PLoS Genet.* **10**, e1004383 (2014).
70. Barbeira, A. N. et al. Exploiting the GTEx resources to decipher the mechanisms at GWAS loci. *Genome Biol.* **22**, 49 (2021).
71. Finucane, H. K. et al. Heritability enrichment of specifically expressed genes identifies disease-relevant tissues and cell types. *Nat. Genet.* **50**, 621–629 (2018).
72. GTEx Consortium. et al. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science* **348**, 648–660 (2015).
73. Võsa, U. et al. Large-scale *cis*-and *trans*-eQTL analyses identify thousands of genetic loci and polygenic scores that regulate blood gene expression. *Nat. Genet.* **53**, 1300–1310 (2021).
74. Qi, T. et al. Identifying gene targets for brain-related traits using transcriptomic and methylomic data from blood. *Nat. Commun.* **9**, 2282. (2018).
75. Pietzner, M. et al. Mapping the proteo-genomic convergence of human diseases. *Science* **374**, eabj1541 (2021).
76. Weeks, E. M. et al. Leveraging polygenic enrichments of gene features to predict genes underlying complex traits and diseases. *Nat. Genet.* **55**, 1267–1276 (2023).
77. Gardner, E. J. et al. Damaging missense variants in *IGF1R* implicate a role for IGF-1 resistance in the aetiology of type 2 diabetes. *Cell Genomics* **2**, 100208 (2022).
78. McLaren, W. et al. The ensembl variant effect predictor. *Genome Biol.* **17**, 122 (2016).
79. Kircher, M. et al. A general framework for estimating the relative pathogenicity of human genetic variants. *Nat. Genet.* **46**, 310–315 (2014).
80. Zhang, H. et al. A powerful procedure for pathway-based meta-analysis using summary statistics identifies 43 pathways associated with type II diabetes in European populations. *PLoS Genet.* **12**, e1006122 (2016).
81. 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature* **526**, 68 (2015).
82. Bulik-Sullivan, B. et al. An atlas of genetic correlations across human diseases and traits. *Nat. Genet.* **47**, 1236–1241 (2015).
83. International HapMap 3 Consortium. Integrating common and rare genetic variation in diverse human populations. *Nature* **467**, 52–58 (2010).
84. Loh, P. R. et al. Contrasting genetic architectures of schizophrenia and other complex diseases using fast variance-components analysis. *Nat. Genet.* **47**, 1385–1392 (2015).
85. Ritari, J. et al. Increasing accuracy of HLA imputation by a population-specific reference panel in a FinnGen biobank cohort. *NAR Genomics Bioinformatics* **2**, lqaa030 (2020).
86. Genovese, G. MoChA WDL pipelines 2022-12-21. *Zenodo* https://doi.org/10.5281/zenodo.10892520 (2022).

Acknowledgements The authors thank J. Karjalainen and M. Cordioli for assistance in GWAS meta-analysis; S. J. Andrews and J. Leinonen for sharing formatted GWAS summary statistics used in genetic correlation analyses; S. Jukarainen and A. Gerussi for insightful discussion on phenome-wide association studies (pheWAS) analyses from a clinical standpoint; S. Jones and M. Kanai for valuable feedback on HLA and fine-mapping; J. Koskela and M. Myllymäki for discussion on clonal haematopoiesis; Y. Fu and A. Preussner for discussion on genetic analyses of sex chromosomes; G. Kops for discussion on the mechanism causing chromosome missegregation; A. Kouno and the members of the BBJ Project for supporting the BBJ analyses; and B. Wheeler for his assistance in running the pathway analyses. We acknowledge the participants and investigators of each contributing biobank. The FinnGen project is funded by two grants from Business Finland (HUS 4685/31/2016 and UH 4386/31/2016) and the following industry partners: AbbVie, AstraZeneca UK, Biogen MA, Bristol Myers Squibb (and Celgene Corporation & Celgene International II), Genentech, Merck Sharp & Dohme, Pfizer, GlaxoSmithKline Intellectual Property Development, Sanofi US Services, Maze Therapeutics, Janssen Biotech, Novartis and Boehringer Ingelheim International. The following biobanks are acknowledged for delivering biobank samples to FinnGen: Auria Biobank (www.auria.fi/biopankki), THL Biobank (www.thl.fi/biobank), Helsinki Biobank (www.helsinginbiopankki.fi), Biobank Borealis of Northern Finland (https://www.ppshp.fi/Tutkimus-ja-opetus/Biopankki/Pages/Biobank-Borealis-briefly-in-English.aspx), Finnish Clinical Biobank Tampere (www.tays.fi/en-US/Research_and_development/Finnish_Clinical_Biobank_Tampere), Biobank of Eastern Finland (www.ita-suomenbiopankki.fi/en), Central Finland Biobank (www.ksshp.fi/fi-FI/Potilaalle/Biopankki), Finnish Red Cross Blood Service Biobank (www.veripalvelu.fi/verenluovutus/biopankkitoiminta), Terveystalo Biobank (www.terveystalo.com/fi/Yritystietoa/Terveystalo-Biopankki/Biopankki/) and Arctic Biobank (https://www.oulu.fi/en/university/faculties-and-units/faculty-medicine/northern-finland-birth-cohorts-and-arctic-biobank). All Finnish Biobanks are members of the BBMRI.fi infrastructure (www.bbmri.fi). Finnish Biobank Cooperative (FINBB) (https://finbb.fi/) is the coordinator of BBMRI-ERIC operations in Finland. The Finnish biobank data can be accessed through Fingenious services (https://site.fingenious.fi/en/) managed by FINBB. The work related to EBB was supported by the Estonian Research Council grants PRG1911 and TK (TK214) and the European Union through the European Regional Development Fund Project no. 2014-2020.4.01.15-0012 GENTRANSMED. The EBB data analysis was carried out in part in the High-Performance Computing Center of University of Tartu. For BCAC and MVP, a detailed acknowledgement is available in the Supplementary Information. This work was supported by the Intramural Research Program of the National Cancer Institute, National Institutes of Health, and the Medical Research Council (unit programmes: MC_UU_12015/2, MC_UU_00006/2). G.G. was supported by NIH grants R01 MH104964 and R01 MH123451; A. Ganna was supported by the Academy of Finland (grant no. 323116) and by the European Research Council under the European Union's Horizon 2020 Research and Innovation Programme (grant no. 945733); 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, and a Sloan Research Fellowship; J.R.B.P. receives research fundings from GSK; C.T. was supported by Japan Agency for Medical Research and Development (AMED) grants JP21ek0109555, JP21tm0424220, JP21ck0106642, JP22wm0425008, JP23ek0410114 and JP23tm0424225, and Japan Society for the Promotion of Science (JSPS) KAKENHI grant JP20H00462; P.N. reports research grants from Allelica, Amgen, Apple, Boston Scientific, Genentech/Roche and Novartis; and S.P. acknowledges research funding from the mvp000 grant. The manuscript does not necessarily represent the views of the Department of Veterans Affairs or the US Government.

Author contributions This project was initiated and led by A.L., G.G., P.-R.L., A. Ganna, J.R.B.P. and M.J.M. A.L. and M.J.M. wrote the first draft of the manuscript with input from all lead authors. A.L. coordinated the analyses of each contributing biobank, conducted across-biobank meta-analysis (including GWAS, allelic shift analysis and pheWAS) and FinnGen-specific analyses, organized post-GWAS analyses, designed and generated all figures and tables (except where noted), and wrote Results, Methods and part of the introduction and Discussion sections. G.G. developed the MoChA pipelines for mLOX calling, GWAS, allelic shift analysis, and X chromosome differential score estimation, guided the analyses of each contributing biobank, performed mLOX calling, GWAS and allelic shift analysis for UKBB and MGB, and wrote the manuscript. Y.Z. performed WES analyses and three-way combined call GWAS in UKBB, generated Supplementary Figs. 2 and 5, prepared Supplementary Tables 8 and 19, and drafted the relevant Results and Methods paragraphs. M.P. developed the Bayesian line model to cluster mLOX and mLOY loci and wrote the relevant Methods paragraph. S.M.Z. performed pheWAS for UKBB and MGB. K.A.K. performed the GWASToGenes pipeline, prepared Supplementary Table 13, and drafted the relevant Methods paragraphs. Z.Y. estimated heritability and genetic correlations and prepared Supplementary Table 16. K.Y. and L.S. performed the pathway analysis and prepared Supplementary Table 14. C.V. performed the sensitivity analyses for associations with leukaemia in UKBB and prepared Supplementary Table 7. X.L. performed mLOX calling, GWAS, allelic shift analysis and HLA fine-mapping replication analysis in BBJ. D.W.B. performed GWAS for PLCO and formatted inputs for blood cell trait heat maps (Fig. 2d and Extended Data Fig. 3b). G.H. performed mLOX calling, GWAS and allelic shift analysis for EBB. B.R.G. and S.P. performed mLOX calling, GWAS, allelic shift analysis and pheWAS for MVP. J.D. performed mLOX calling and GWAS for BCAC. W.Z. performed mLOX calling, GWAS and allelic shift analysis for PLCO. Y.M. participated in BBJ analyses. V.T. and F.-D.P. participated in EBB analyses. M.A., T.P.S. and A. Ghazal participated in FinnGen analyses. W.-Y.H. and N.D.F. participated in PLCO analyses. E.J.G. participated in UKBB WES analyses. V.G.S. assisted in interpretating findings related to clonal haematopoiesis. A.P. coordinated the FinnGen project. H.M.O. advised in the HLA fine-mapping analysis and assisted in interpretating findings related to HLA. T.T. assisted in interpretating findings related to skewed X chromosome inactivation and escape from X chromosome inactivation. S.J.C. coordinated the PLCO project. R.M. supervised EBB analyses. P.N. supervised pheWAS for UKBB, MGB and MVP. M.J.D. initialized and conceptualized the mCA project in FinnGen and assisted in interpreting findings, especially those related to mLOY in male participants. A.B. supervised pheWAS in UKBB, MGB and MVP, and the sensitivity analyses for associations with leukaemia in UKBB. S.A.M. supervised the development of MoChA pipelines. C.T. supervised BBJ analyses and advised the HLA fine-mapping analysis. P.-R.L., A. Ganna, J.R.B.P. and M.J.M. co-supervised the project, interpreted the findings and wrote the manuscript. All authors reviewed the manuscript.

Competing interests G.G., P.-R.L. and S.A.M. declare competing interests: patent application PCT/WO2019/079493 has been filed on the mCA detection method used in this work. J.R.B.P. and E.J.G. are employees and shareholders of Insmed. Y.Z. is a UK University Worker of GSK. A.B. reports scientific advisory board membership for TenSixteen Bio. P.N. reports personal fees from Allelica, Apple, AstraZeneca, Blackstone Life Sciences, Creative Education Concepts, CRISPR Therapeutics, Eli Lilly & Co., Foresite Labs, Genentech/Roche, GV, HeartFlow, Magnet Biomedicine, Merck and Novartis, scientific advisory board membership of Esperion Therapeutics, Preciseli, and TenSixteen Bio, scientific co-founder of TenSixteen Bio, equity in MyOme, Preciseli and TenSixteen Bio, and spousal employment at Vertex Pharmaceuticals, all unrelated to the present work.

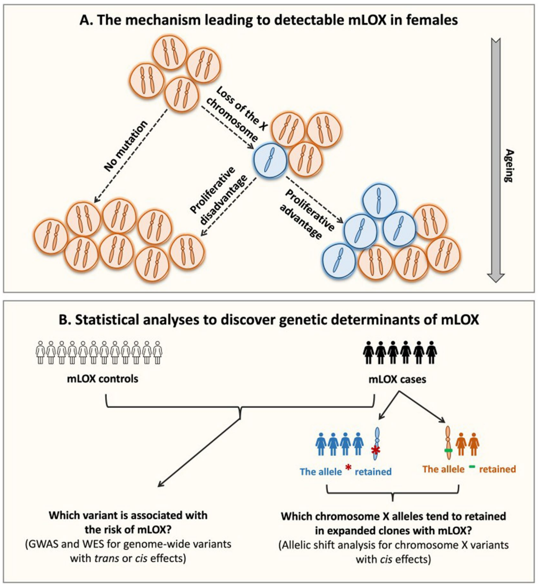
Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41586-024-07533-7.

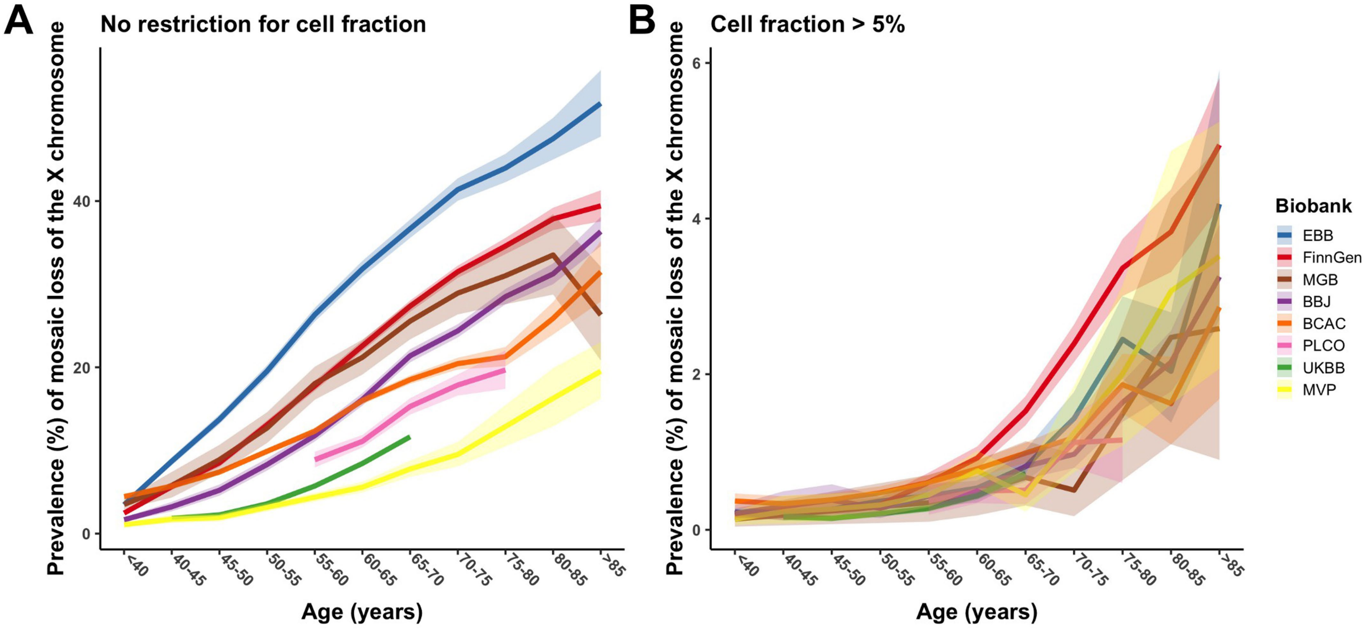
Correspondence and requests for materials should be addressed to Aoxing Liu, Giulio Genovese, Po-Ru Loh, Andrea Ganna, John R. B. Perry or Mitchell J. Machiela.

Peer review information *Nature* thanks Eric Jorgenson and Siddhartha Kar for their contribution to the peer review of this work. Peer reviewer reports are available.

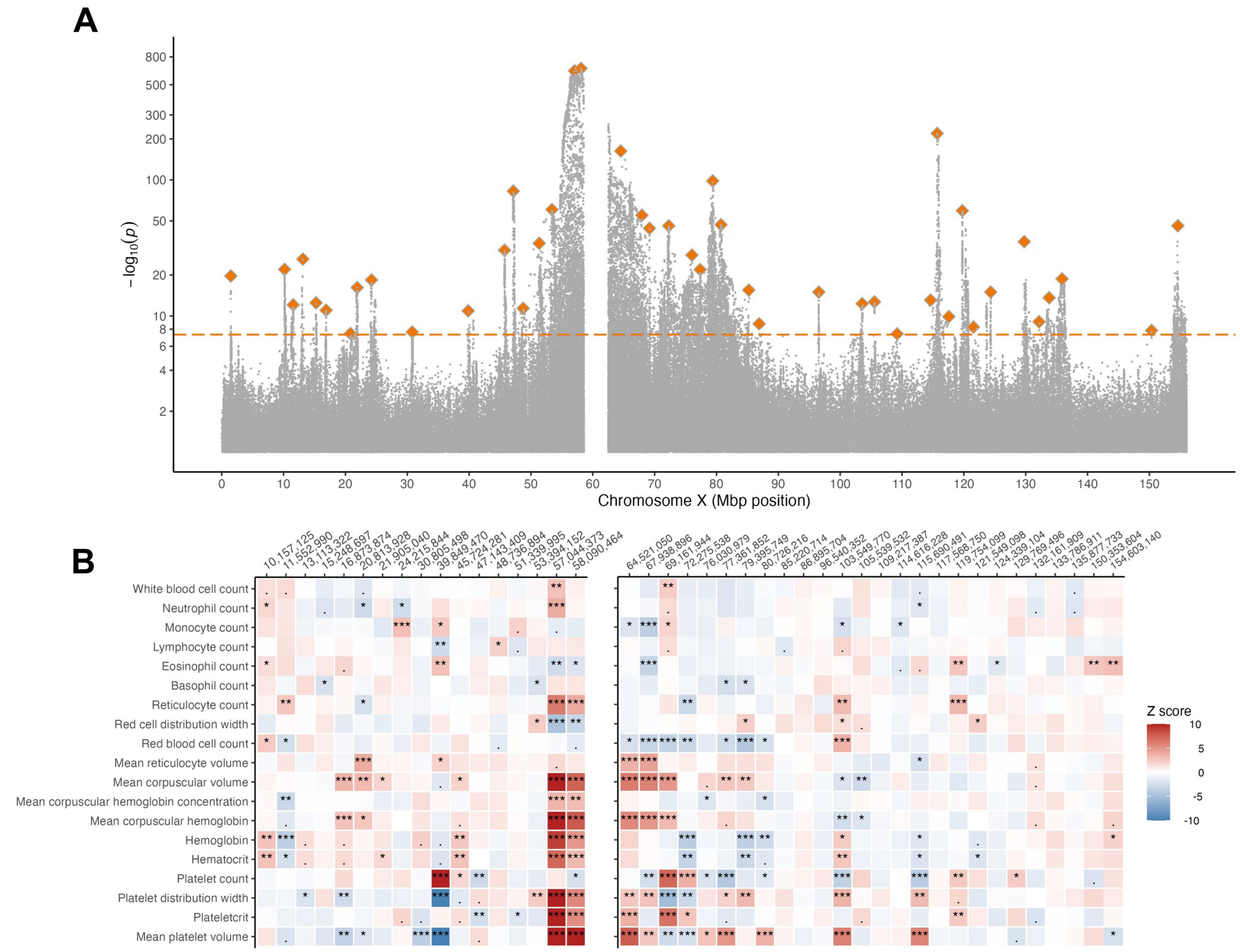
Reprints and permissions information is available at http://www.nature.com/reprints.


A. The mechanism leading to detectable mLOX in females
Loss of the X chromosome
No mutation
Proliferative disadvantage
Proliferative advantage
Ageing
B. Statistical analyses to discover genetic determinants of mLOX
mLOX controls
mLOX cases
The allele * retained
The allele - retained
Which variant is associated with the risk of mLOX?
(GWAS and WES for genome-wide variants with *trans* or *cis* effects)
Which chromosome X alleles tend to retained in expanded clones with mLOX?
(Allelic shift analysis for chromosome X variants with *cis* effects)

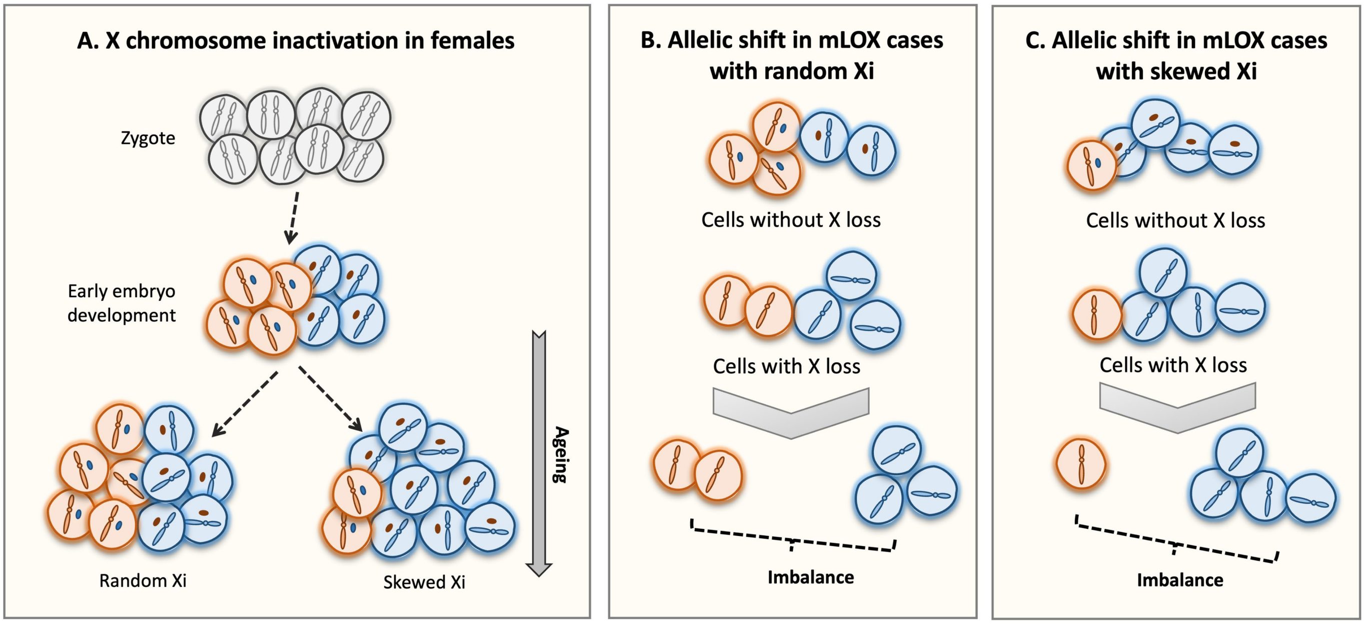
Extended Data Fig. 1 | Theoretical framework of the mLOX study. Panel (A) depicts the etiologic process leading to detectable mosaic loss of the X chromosome (mLOX) in females. Detectable age-related mLOX develops only if the mutant haematopoietic stem cell (HSC) survives loss of the X chromosome and the mutation confers a proliferative advantage over normal cells. Panel (B) shows the statistical approaches used to discover the genetic determinants of mLOX. Variants associated with susceptibility to mLOX, acting as either trans or cis factors, are examined using a genome-wide association study (GWAS), for common variants with minor allele frequency (MAF) > 0.1%, and a gene-burden test performed for whole-exome sequencing (WES) data for rare variants with MAF < 0.1%. Among samples with detectable mLOX, allelic shift analysis is used to detect chromosome X alleles exhibiting *cis* selection, that is, more likely to be clonally selected for when detectable mLOX retains these alleles.



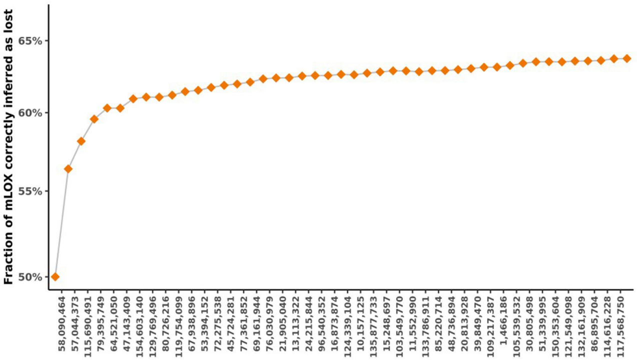
Extended Data Fig. 2 | Prevalence of mLOX by age at genotyping in each contributed biobank. Panel (A) is for all detectable mLOX in peripheral leukocytes, while Panel (B) is restricted to expanded mLOX with cell fraction >5%. Data are presented as mean values +/− SEM.



Extended Data Fig. 3 | Allelic shift of chromosome X alleles among mLOX cases. Panel (A) shows -log10(P) of chromosome X variants from allelic shift analysis by meta-analyzing data of 83,320 mLOX cases from seven biobanks, with lead variants of 44 independent loci highlighted. The y axis is the log scale of P values from a two-sided test and the dashed line denotes the statistical significance after multiple comparison adjustments (5.0×10^{-8}, which is the same as the GWAS significance level). Panel (B) is a heat map for associations of 43 allelic shift analysis lead variants with 19 blood cell phenotypes[46], with significance levels from the original GWAS expressed by asterisks (*** for two-sided exact $P \leq 0.001$, ** for $P \leq 0.01$, * for $P \leq 0.05$). One variant was dropped due to no appropriate proxy variant available in blood cell phenotype GWAS. The absolute Z scores were cropped to the range of [0–10].



Extended Data Fig. 4 | Allelic shift in the context of X chromosome inactivation. Panel (A) depicts the main mechanism of X chromosome inactivation (Xi) in females. To compensate for gene dosage imbalances between XX females and XY males, one of the two X chromosomes in females is randomly inactivated early in embryonic development and this inactivation status is passed down to daughter cells. As some females age, the expected 1:1 ratio of inactivated maternal to paternal X chromosome copies can become skewed, if cells harboring one of the active X chromosomes is more frequent than the other. Panel (B) and (C) depict the pattern of allelic shift in mLOX cases in terms of the status of Xi, with Panel (B) for random Xi and panel (C) for skewed Xi. As mLOX preferentially affects the inactivated X chromosome[2], the imbalance between chromosome X alleles in mLOX cases can be seen as the combined *cis* effects of both skewed Xi and mLOX. In other words, the imbalance of chromosome X alleles in mLOX cases could also be shaped by alleles that have *cis* effects solely on the process of skewed Xi.



Extended Data Fig. 5 | Contribution of each X chromosome allelic shift loci to the prediction of the retained X chromosome in females with mLOX. We proposed a novel polygenic score including the 44 loci identified from allelic shift analysis to infer the retained X chromosome in detectable mLOX. To avoid overfitting, the effects of the 44 loci were estimated from allelic shift analysis of 56,319 mLOX cases from six biobanks excluding FinnGen while the prediction performance was tested in 27,001 FinnGen mLOX cases. The plot shows the contribution of each of the 44 loci to the prediction, starting with the most significant variants.

Extended Data Table 1 | Descriptive characteristics of the eight biobanks contributing to the mLOX analysis

Biobank	Median age (SD)	mLOX Cases	Controls	Effective sample size	Continental ancestry groups
FinnGen	54 (18.2)	27,001	141,837	90,732	European, Finnish
Breast Cancer Association Consortium (BCAC)	57 (11.9)	21,966	155,356	76,980	European
Estonian Biobank (EBB)	44 (16.3)	20,232	110,547	68,408	European, Estonians
UK Biobank (UKBB)	57 (8.0)	16,214	244,931	60,829	European, British
Biobank Japan (BBJ)	65 (15.8)	13,597	63,720	44,823	East Asian, Japanese
Million Veteran Program (MVP)	54 (13.9)	1,496	33,192	5,726	European
Mass General Brigham Biobank (MGB)	54 (17.3)	2,108	11,527	7,128	European
Prostate, Lung, Colorectal and Ovarian Cancer Screening Trial (PLCO)	64 (5.4)	2,672	17,178	9,249	European

nature portfolio

Corresponding author(s): Aoxing Liu, Giulio Genovese, Po-Ru Loh, Andrea Ganna, John Perry, Mitchell Machiela

Last updated by author(s): Jan 5, 2024

Reporting Summary

Nature Portfolio 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 Portfolio policies, see our Editorial Policies 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	MoChA pipeline (versions 2022-10-14, 2021-05-14, 2021-08-17 and 2021-09-07)(https://github.com/freeseek/mochawdl)
Data analysis	SAIGE (0.45.1), regenie (3.2.5), BOLT-LMM (2.4), METAL (2011-03-25)(https://github.com/covid19-hg/META_ANALYSIS), GCTA (1.94.1), MAGMA (1.10), HEIDI (0.68), coloc (5.1.0), LDSC (1.0.1), STAAR (0.9.6.2), sARTP (0.9.45), linemodels (0.2.0)(https://github.com/dsgelab/Mosaic-loss-of-chromosome-X/tree/main/BayesLineModel)

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 and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio 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 description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

Overall and population level GWAS summary statistics generated from the mLOX meta-analysis are available on the GWAS Catalog (accession numbers: GCST90328147, GCST90328148, GCST90328149, GCST90328150). Requests for access to individual-level data differ for each contributing biobank. For FinnGen, researchers can apply for health data from the Finnish Data Authority Findata (https://findata.fi/en/permits/) and individual-level genotype data available through the Fingenious portal (https://site.fingenious.fi/en/). These resources are hosted by the Finnish Biobank Cooperative FINBB (https://finbb.fi/en/). Access can only be provided for research projects within the scope of the Finnish Biobank Act, which includes health promotion, understanding disease mechanisms or developing medical products or treatment practices. For EBB, individual level health, lifestyle, demographic and genetic data are anonymized and available for research projects. Data sharing is conducted in accordance with the regulations of the Estonian Genome Center of the University of Tartu (HGRA). A data application form can be found at www.biobank.ee. The research project has to obtain approval from the Ethics Review Committee on Human Research of the University of Tartu as well as approval from the EGCUT scientific committee. For UKBB, all individual-level data used in the analysis is available by application to the UKBB Access Management System (https://www.ukbiobank.ac.uk). Approved researchers can submit applications for review and assessments are made to determine if the research proposal qualifies as health-related research in line with public interest. For BCAC, data for some of the samples is available on dbGAP (https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs001265.v1.p1). Requests for BCAC data can be made to the Data Access Coordination Committee (DACC) of BCAC (http://bcac.ccge.medschl.cam.ac.uk/bcacdata/). BCAC DACC approval is required to access individual level phenotype and genotype data from the ABCFS, ABCS, ABCTB, BBCC, BBCS, BCEES, BCFR-NY, BCFR-PA, BCFR-UTAH, BCINIS, BIGGS, BREOGAN, BSUCH, CBCS, CCGP, CECILE, CGPS, CNIO-BCS, CPSII, CTS, EPIC, DIETCOMPLYF, ESTHER, GC-HBOC, GENICA, HABCS, HCSC, HEBCS, HMBCS, HUBCS, KARBAC, KARMA, KBCP, KCONFAB/AOCS, LMBC, MABCS, MARIE, MBCSG, MCBCS, MCCS, MEC, MISS, MMHS, MTLGEBCS, NBCS, NC-BCFR, NBHS, NCBCS, NHS, NHS2, OBCS, OFBCR, ORIGO, PBCS, PKARMA, PLCO, POSH, RBCS, SASBAC, SBCS, SEARCH, SISTER, SKKDKFZS, SMC, SZBCS, UCIBCS, UKBGS, UKOPS and USRT studies. For MVP, summary statistics are available on dbGaP under the Million Veteran Program accession number phs001672. Additional data supporting the findings of this study are available upon reasonable request from MVP. These data are not publicly available due to restrictions of the US Government and Department of Veterans Affairs concerning privacy and participant consent. For MGB, a portion of individual level genomic data are available in dbGAP as part of the eMERGE consortium (phs001584.v2.p2) and as part of the Center Common Disease Genomics (phs002018.v1.p1). Additional MGB data are not currently publicly available due to data restrictions. For PLCO, individual level genotype data is available in dbGaP (phs001286.v2.p2). Permitted data use includes discovery and hypothesis generation in the investigation of genetic contributions to cancer risk and risk of other diseases as well as development of novel analytical approaches for GWAS. Individual level phenotype data can be requested through the NCI Cancer Data Access System (CDAS) (https://cdas.cancer.gov/plco/). For BBJ, information on the cohort is available at the RIKEN website (http://jenger.riken.jp/en/). While individual-level genetic data are not accessible, all other individual-level data are available upon request.

Human research participants

Policy information about studies involving human research participants and Sex and Gender in Research.

Reporting on sex and gender	All analyses of mLOX were restricted to participants of female sex as determined by genetic markers. All analyses comparing mLOX to mLOY used mLOY summary statistics from participants of male sex as determined by genetic markers.
Population characteristics	Our primary study population was apparently healthy female participants from international biobanks that had existing genotype array data used to call mosaic chromosomal alterations. Median age ranged from 44 (sd=16.3) for the Estonian Biobank to 65 (sd=15.8) for Biobank Japan.
Recruitment	Recruitment for each contributing biobank was intended to capture study participants that were representative of the underlying study population of interest. Specific details on recruitment are available in the study-specific references we provided for each contributing biobank.
Ethics oversight	The local Institutional Review Boards of each participating biobank provided ethical oversight of the respective biobanks.

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

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	The total combined sample size of our investigation was 883,574 female participants. These samples were recruited from international biobank studies with existing genotype array data for calling mLOX, for which data use permissions permitted the investigation of mLOX, and for which collaboration for calling mLOX from raw intensity data was possible. This sample size was sufficiently powered to detect the number of mLOX cases needed for investigating common and low frequency variation associated with mLOX.
Data exclusions	We excluded samples with male sex, samples with poor genotyping completion rate and samples with high deviations in raw intensity characteristics (e.g., B allele frequency deviation).
Replication	No formal replication was present in the mLOX GWAS, allelic shift analysis, and epidemiological association analyses, although identified associations were tested for heterogeneity across contributing studies when possible.
Randomization	This study did not have a randomization component. Covariates for known and suspected confounders were controlled for using statistical adjustment.

Blinding	No blinding was present in this study. Blinding was unnecessary as knowledge of mLOX status was unknown to the study participants and could not impact the germline genetic variables for which we associations were tested.

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 and archaeology
☒	□	Animals and other organisms
□	☒	Clinical data
☒	□	Dual use research of concern

Methods

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

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	n/a
Study protocol	All contributing biobanks utilized a common MoChA wdl pipeline for calling mLOX.
Data collection	Genotyping and collection of clinical data was carried out independently by each contributing biobank. More details are available in the study specific references.
Outcomes	The primary outcome of interest was mLOX as called by the MoChA pipeline.