Population analyses of mosaic X chromosome loss identify genetic drivers and widespread signatures of cellular selection

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

[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 Publich 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]Division of Cancer Epidemiology and Genetics, National Cancer Institute, Rockville, MD, USA. [12]Department of Medicine, Queen's University, Kingston, ON, Canada. [13]Laboratory for Statistical and Translational Genetics, RIKEN Center for Integrative Medical Sciences, Yokohama, Japan. [14]Cancer Prevention Fellowship Program, Division of Cancer Prevention, National Cancer Institute, Rockville, MD, USA. [15]Estonian Genome Centre, Institute of Genomics, University of Tartu, Tartu, Estonia. [16]Center for Data and Computational Sciences (C-DACS), VA Cooperative Studies Program, VA Boston Healthcare System, Boston, MA, USA. [17]Booz Allen Hamilton, McLean, VA, USA. [18]Centre for Cancer Genetic Epidemiology, Department of Public Health and Primary Care, University of Cambridge, Cambridge, UK. [19]Laboratory for Genotyping Development, RIKEN Center for Integrative Medical Sciences, Yokohama, Japan. [20]Department of Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, USA. [21]Division of Hematology/Oncology, Boston Children's Hospital, Harvard Medical School, Boston, MA, USA. [22]Department of Pediatric Oncology, Dana-Farber Cancer Institute, Harvard Medical School, Boston, MA, USA. [23]Anesthesia, Critical Care, and Pain Medicine, Massachusetts General Hospital, Boston, MA, USA. [24]Division of Genetic Medicine, Department of Medicine, Vanderbilt University Medical Center, Nashville, TN, USA. [25]Clinical Research Center, Shizuoka General Hospital, Shizuoka, Japan. [26]Department of Applied Genetics, School of Pharmaceutical Sciences, University of Shizuoka, Shizuoka, Japan. [27]Center for Data Sciences, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, USA. [28]These authors contributed equally: Aoxing Liu, Giulio Genovese, Yajie Zhao. [29]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

Mosaic loss of the X chromosome (mLOX) is the most commonly occurring clonal somatic alteration detected in the leukocytes of women, yet little is known about its genetic determinants or phenotypic consequences. To address this, we estimated mLOX in > 880,000 women across eight biobanks, identifying 12% of women with detectable X loss in approximately 2% of their leukocytes. Out of 1,253 diseases examined, women with mLOX had an elevated risk of myeloid and lymphoid leukemias and pneumonia. Genetic analyses identified 56 common variants influencing mLOX, implicating genes with established roles in chromosomal missegregation, cancer predisposition, and autoimmune diseases. Complementary exome-sequence analyses identified rare missense variants in *FBXO10* which confer a two-fold increased risk of mLOX. A small fraction of these associations were shared with mosaic Y chromosome loss in men, suggesting different biological processes drive the formation and clonal expansion of sex chromosome missegregation events. Allelic shift analyses identified alleles on the X chromosome which are preferentially retained, demonstrating that variation at many loci across the X chromosome is under cellular selection. A novel polygenic score including 44 independent X chromosome allelic shift loci correctly inferred the retained X chromosomes in 80.7% of mLOX cases in the top decile. Collectively our results support a model where germline variants predispose women to acquiring mLOX, with the allelic content of the X chromosome possibly shaping the magnitude of subsequent clonal expansion.

Introduction

Females carry a maternal and paternal copy of the X chromosome in which one copy is partially rendered transcriptionally inactive early in development by expression of Xist[1] and epigenetic modifications. The inactivation process is random as to which X chromosome is chosen with the resulting inactive state being irreversible and clonally transmitted to daughter cells[2]. X chromosome inactivation has evolved as a mechanism to compensate for gene dosage imbalances between XX females and XY males, although some genes are only partially inactivated[3], including several tumor suppressor genes (e.g., *ATRX*, *KDM5C*)[4]. 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 variation relevant to disease risk.

With age, the expected 1:1 ratio of inactivated maternal to paternal X chromosome copies can become skewed. X chromosome inactivation skewing is observed in various tissues with high frequencies observed in leukocytes[5,6]. Detectable skewed X chromosome inactivation in leukocytes is heritable (h^2=0.34)[7] and can indicate depletion of haematopoietic stem cells, selection pressures on leukocytes, or clonal hematopoiesis (CH). Recent investigations of age-related CH have described elevated rates of mosaic sex chromosome aneuploidies in population-based surveys of apparently healthy adults[8-13]. Mosaic loss of the female X chromosome (mLOX) is elevated in frequency compared to the

autosomes[14], preferentially impacts the inactivated X chromosome[10] and is associated with elevated leukemia risk[15,16]. This contrasts with the male X chromosome which has very low rates of aneuploidy[17]. As the X chromosome encompasses approximately 5% of the genome and contains genes relevant to immunity, cancer susceptibility, and cardiovascular diseases, loss of a homologous copy and subsequent hemizygous selection could lead to downstream consequences on female health, as observed in Turner syndrome (45,XO)[18]; however, no study has systematically examined longitudinal associations of mLOX with disease risk.

As mLOX is a clonal pro-proliferative genomic alteration, understanding the molecular mechanisms driving susceptibility to mLOX could provide new insights into the impact of aging on hematopoiesis as well as hematologic cancer risk. The X chromosome, particularly the inactive X, is more frequently mutated in cancer genomes[19] and is late replicating relative to autosomes, potentially increasing susceptibility to chromosomal alterations[20]. While few genome-wide association studies (GWAS) of mLOX have been reported to date[14,21], GWAS of mosaic loss of the Y chromosome (mLOY) in men has identified hundreds of susceptibility loci[11-13,22], many of which highlight genes involved in cell cycle regulation and cancer susceptibility. Here we describe insights from epidemiologic and genetic analyses of X chromosome loss from a combined meta-analysis of 883,574 women. We identify 56 independent common susceptibility variants across 42 loci, rare missense variants of *FBXO10* associated with mLOX, and 44 X chromosome loci that strongly influence which X chromosome is retained. The identified signals only partially overlap with known signals for other age-related types of CH. These data indicate mLOX, along with other age-related types of CH, are important pre-clinical indicators of hematologic cancer risk and identify mitotic missegregation, autoimmunity, blood cell trait, and cancer predisposition genes as core etiologic components for mLOX susceptibility and selection.

Results

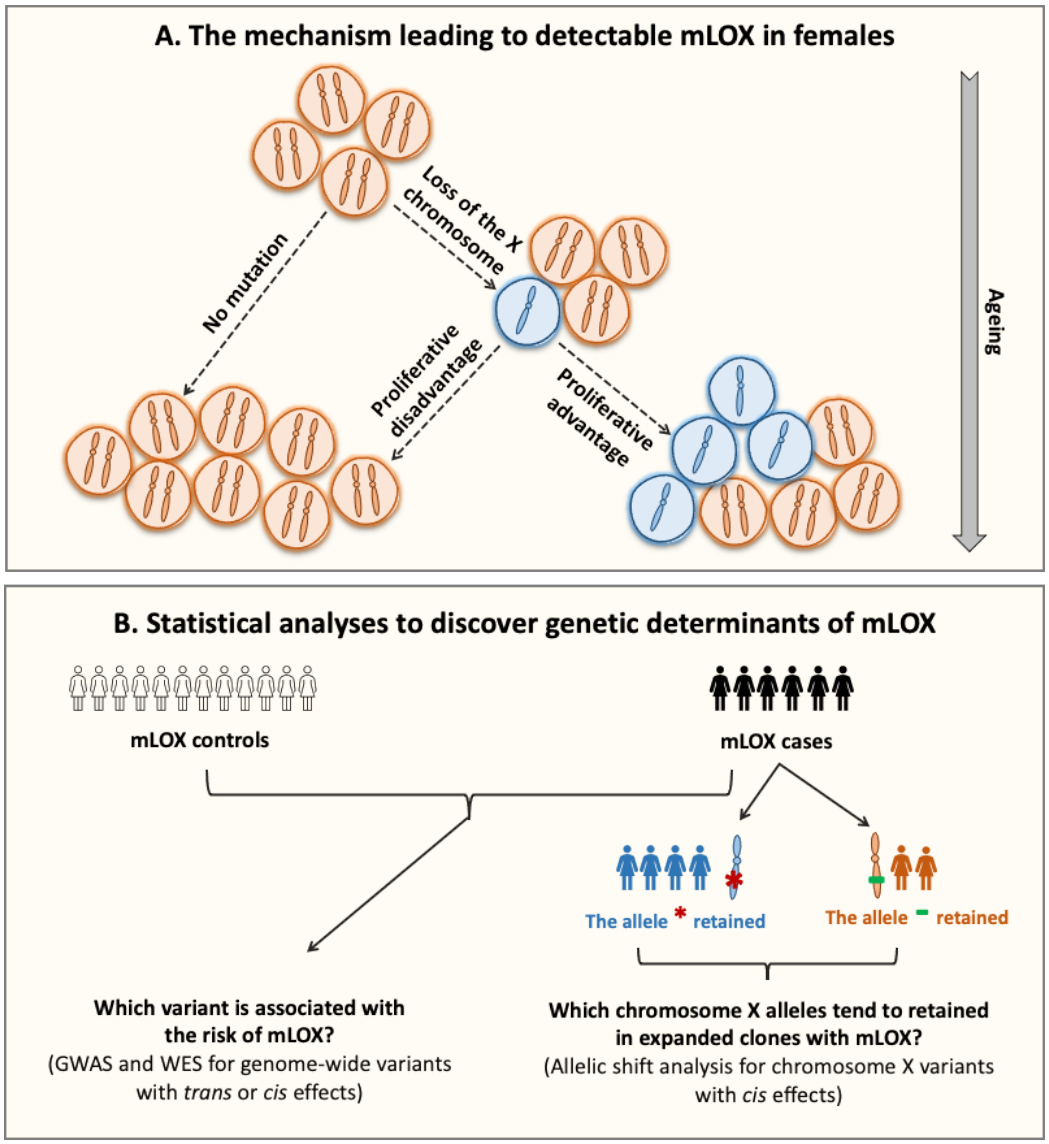

Figure 1. Theoretical framework of the mLOX study.

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.

Mosaic loss of the X chromosome in eight contributed biobanks

We leveraged genetic data in a total of 883,574 women from eight biobanks worldwide, including European ancestry participants from FinnGen[23], Estonian Biobank (EBB)[24], UK Biobank (UKBB)[25,26], Breast Cancer Association Consortium (BCAC)[27,28], Million Veteran Program (MVP)[29,30], Mass General Brigham Biobank (MGB)[31,32], and Prostate, Lung, Colorectal and Ovarian Cancer Screening Trial (PLCO)[33], as well as East Asian ancestry participants from Biobank Japan (BBJ)[34] (**Supplementary Table S1**). The median (SD) age at sample collection for genotyping ranged from 44 (16.3) for EBB to 65 (15.8) for BBJ. 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 women, 105,286 (11.9%) were classified as cases with detectable mLOX (**Methods**; **Table 1**). Overall, the cell fraction of mLOX (i.e., the estimated fraction of peripheral leukocytes with X loss) was low (median=1.5%) with expanded clones having frequency ≥5% infrequently observed (0.6% of women) (**Supplementary Figure S1**). A subset of UKBB participants (243,520 out of 261,145) also had whole-exome sequencing (WES) data available which allowed us to assess the performance of mLOX calling from MoChA. Among UKBB mLOX cases classified by MoChA, a high correlation (r=-0.86) was observed between cell fraction derived from SNP array data (by MoChA) and X dosage derived from WES data (**Supplementary Figure S2**). In addition to the MoChA generated dichotomous measure used by all biobanks, in UKBB we generated a 3-way combined quantitative measure by integrating independent information from both SNP array and WES data (**Methods**). As increasing age is a well-established causal factor for acquiring all types of CH including mLOX, we further assessed the performance of different mLOX measures in UKBB by their associations with age. We observed an increase in t-test statistics by 29.2% with the 3-way calls but noted that the SNP array-only calls with MoChA were still a powerful approach for defining mLOX.

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

Environmental determinants and epidemiological consequences

Like many other types of somatic mutations[13,14], the frequency of women with detectable mLOX in peripheral leukocyte is age-related, with a frequency of 3.0% in women aged <40 and reaching >35.0% after 80, averaged over all contributing biobanks (**Supplementary Table S2**). Across biobanks, differences were seen in the frequency of mLOX, with the highest age-adjusted frequency presented in EBB and the lowest in MVP (**Supplementary Figure S3A**). However, such variation in frequencies was largely reduced when restricted to expanded mLOX with cell fraction >5% (**Supplementary Figure S3B**), suggesting that mLOX detection differences were more prominent for low cell fraction clones. 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 ≥5% (OR=1.3 [1.2-1.5], $P=6.9\times10^{-5}$ in FinnGen and OR=1.3 [1.1-1.5], $P=4.6\times10^{-4}$ in UKBB) (**Supplementary Table S3** and **Figure S4-S5**). The relationship between smoking and skewed X-inactivation has not been established, as smoking was suggested as a modulator for skewed X inactivation in the whole-blood tissue for women older than age 55[7] but not associated in the TwinsUK cohort[35]. We observed limited evidence for an association between BMI and mLOX in FinnGen and UKBB (**Supplementary Table S4**).

To evaluate disease outcomes associated with detectable mLOX, we performed Cox proportional hazards regression for incident disease cases in FinnGen, UKBB, MVP, and MGB independently considering genotyping age and ever-smoking status as covariates and meta-analyzed across biobanks with a fixed-effect model (**Methods**). Out of the 1,253 diseases we examined, we identified significant associations ($P<4.0\times10^{-5}$) with leukemia overall (HR=1.7 [1.5-2.1], $P=3.5\times10^{-10}$) and chronic lymphoid leukemia (CLL) (HR=3.3 [2.4-4.4], $P=8.4\times10^{-15}$) and suggestive evidence for acute myeloid leukemia (AML) (HR=1.9 [1.3-2.8], $P=1.8\times10^{-3}$) (**Supplementary Table S5**). Unlike the germline loss of the X chromosome in women with Turner syndrome (45,XO), which can cause various medical and developmental problems[18], we noted limited clinical consequences for women with detectable mLOX in blood.

As the median mLOX cell fraction impacted is approximately 2%, we proposed that investigating expanded clones could result in stronger disease associations. Here, we focused on mLOX with cell fraction ≥10% as this threshold has been empirically determined to be etiologically relevant for detecting diseases associated with mCAs[15,16]. Restricting to expanded mLOX, we observed evidence for elevated associations with leukemia overall (HR=6.3 [3.9-10.2], $P=7.3\times10^{-14}$), CLL (HR=14.7 [6.5-33.3], $P=9.5\times10^{-11}$), and AML (HR=10.6 [3.1-36.1], $P=1.5\times10^{-4}$) (**Supplementary Table S6**).

We also observed suggestive evidence for associations with vitamin B complex deficiency (HR=3.7 [1.8-7.9], $P=6.0\times10^{-4}$) and pneumonia (HR=1.5 [1.2-1.8], $P=4.7\times10^{-4}$), especially pneumonia caused by bacterial infections (HR=1.8 [1.3-2.3], $P=3.9\times10^{-5}$). Similarly, in UKBB[16], an increased risk of incident pneumonia was observed for both women with expanded mLOX (HR=1.8 [1.0-3.2], P=0.035) and men with expanded mLOY (HR=1.2 [1.1-1.4], $P=1.1\times10^{-4}$).

To examine the potential impacts of other types of CH on mLOX associations with leukemia, we performed sensitivity analyses in UKBB where we had available calls on autosomal mosaic chromosomal alterations (mCAs) as well as CH mutations in driver genes, commonly referred to as clonal hematopoiesis of indeterminate potential (CHIP)[36]. We observed attenuations in associations for expanded mLOX when removing individuals with autosomal mCAs (HR=3.8 [1.6-9.3], $P=2.7\times10^{-3}$), CHIP (HR=6.2 [3.1-12.4], $P=3.1\times10^{-7}$), and both mCAs and CHIP (HR=4.5 [1.9-10.8], $P=8.6\times10^{-4}$) (**Supplementary Table S7**); however, significant associations with expanded mLOX and overall leukemia risk remained indicating mLOX is independently associated with leukemia risk. Associations for other lymphoid and myeloid leukemias display similar patterns, albeit losing statistical significance likely due to reduced sample size.

We further assessed the relationship between mLOX and a broad range of quantitative phenotypes in UKBB (**Methods**; **Supplementary Table S8**) and observed enrichment of associations with blood count traits, such as higher levels of lymphocyte count ($P=9.3\times10^{-126}$) and monocyte count ($P=4.9\times10^{-4}$) and lower levels of neutrophil count ($P=3.3\times10^{-62}$) and red blood cell count ($P=4.4\times10^{-4}$). As for blood biomarkers or biochemistry, acquiring mLOX was associated with shorter telomere length (e.g., $P=2.8\times10^{-14}$ for adjusted T/S ratio) and higher levels of total protein ($P=1.9\times10^{-8}$), triglycerides ($P=1.1\times10^{-5}$), aspartate aminotransferase ($P=1.1\times10^{-7}$), and gamma-glutamyl transferase ($P=3.0\times10^{-4}$). We noted that, unlike disease associations that usually exerted more significant effects in expanded mLOX (e.g., various subtypes of leukemia), 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 women.

Common and rare variants associated with mLOX susceptibility

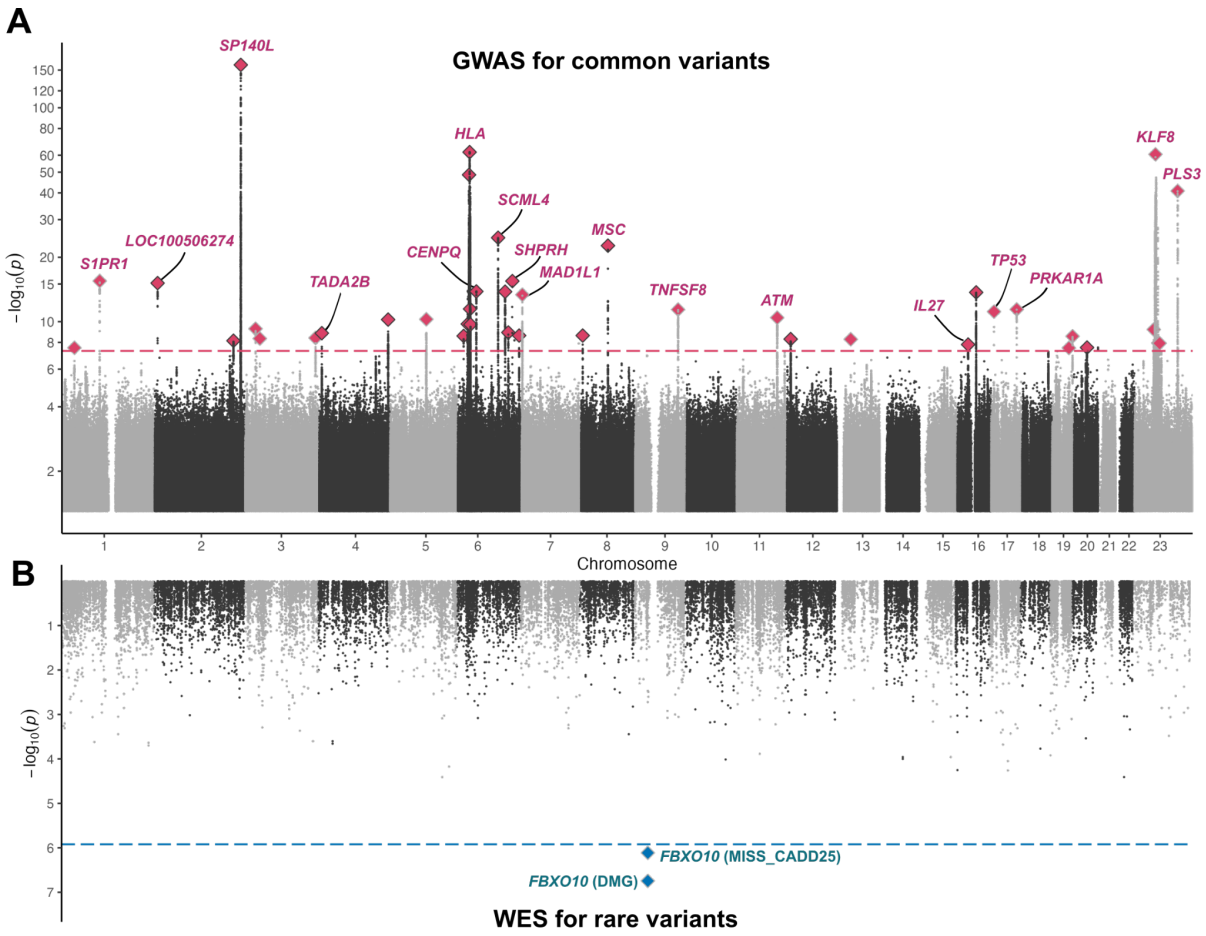

Figure 2. Common and rare genetic contributors to mLOX susceptibility.

Panel (**A**) shows genome-wide association study -$\log_{10}$(P) for the association of common variants (MAF>0.1%) with mLOX. Labels are only assigned for candidate genes of the top 10 lead variants from meta-analysis or the top 10 candidate genes from gene prioritization and the y-axis is log scale. Panel (**B**) presents gene burden test -$\log_{10}$(P) for the rare variants (MAF<0.1%) associations with mLOX. The dashed lines denote the statistical significance, which is 5.0×10^{-8} for GWAS (**A**) and 1.2×10^{-6} for the gene-burden test (**B**).

We performed a genome-wide association study (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 eight contributing biobanks independently, for a total of 883,574 women (**Methods**). To increase GWAS power, we used enhanced 3-way combined calls for UKBB and meta-analyzed summary statistics across different mLOX measures with a weighted z-score method (**Methods**). Of the 33,737,925 variants examined, we identified 56 independent genome-wide significant variants ($P<5.0\times10^{-8}$) across 42 loci associated with mLOX susceptibility (**Methods**; **Figure 2A**; **Supplementary Table S9**). Most independent variants were located on chromosomes 6

(17 variants), 2 (9 variants), X (7 variants), 3 (3 variants), and 17 (3 variants), with chromosomes 6, 2, and X explaining more heritability than expected by their chromosome length (**Supplementary Figure S6**). Despite differences in age-adjusted mLOX frequencies, mLOX variant effects were consistent across the eight biobanks and across European and East Asian ancestry (P from Cochran's Q-test > $0.05/56 = 8.9\times10^{-4}$) (**Supplementary Table S10**), with the exception of rs78378222 (*TP53*, P from meta-analysis = 7.2×10^{-12}, P from heterogeneity test = 6.7×10^{-4}) and three X chromosome variants (X:51749114:C:CGT, rs141849992, and rs58638231). For rs78378222, the heterogeneity of variant effects across biobanks was likely due to differences in mLOX cell fraction by contributing studies. When stratifying by cell fraction in FinnGen, the OR 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] ($P=1.4\times10^{-4}$) for expanded mLOX with cell fraction above 5% (P for effect size difference from a two-sided t-test = 2.5×10^{-5}) (**Supplementary Table S11** and **Figure S7**).

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 S12**). The most significantly associated mLOX locus is at 2q37.1, replicating previous UKBB mLOX GWAS signals at that locus[14,21]. We mapped the hit to *SP140L*, a gene 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[37]. 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 (e.g., *EOMES*[38,39], *JARID2*[40], *MYB*[41], *ATM*[42], *TP53*[43], and *PRKAR1A*[44]) are directly relevant to leukemia 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 highly 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 etiologic 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**; **Supplementary Table S13**).

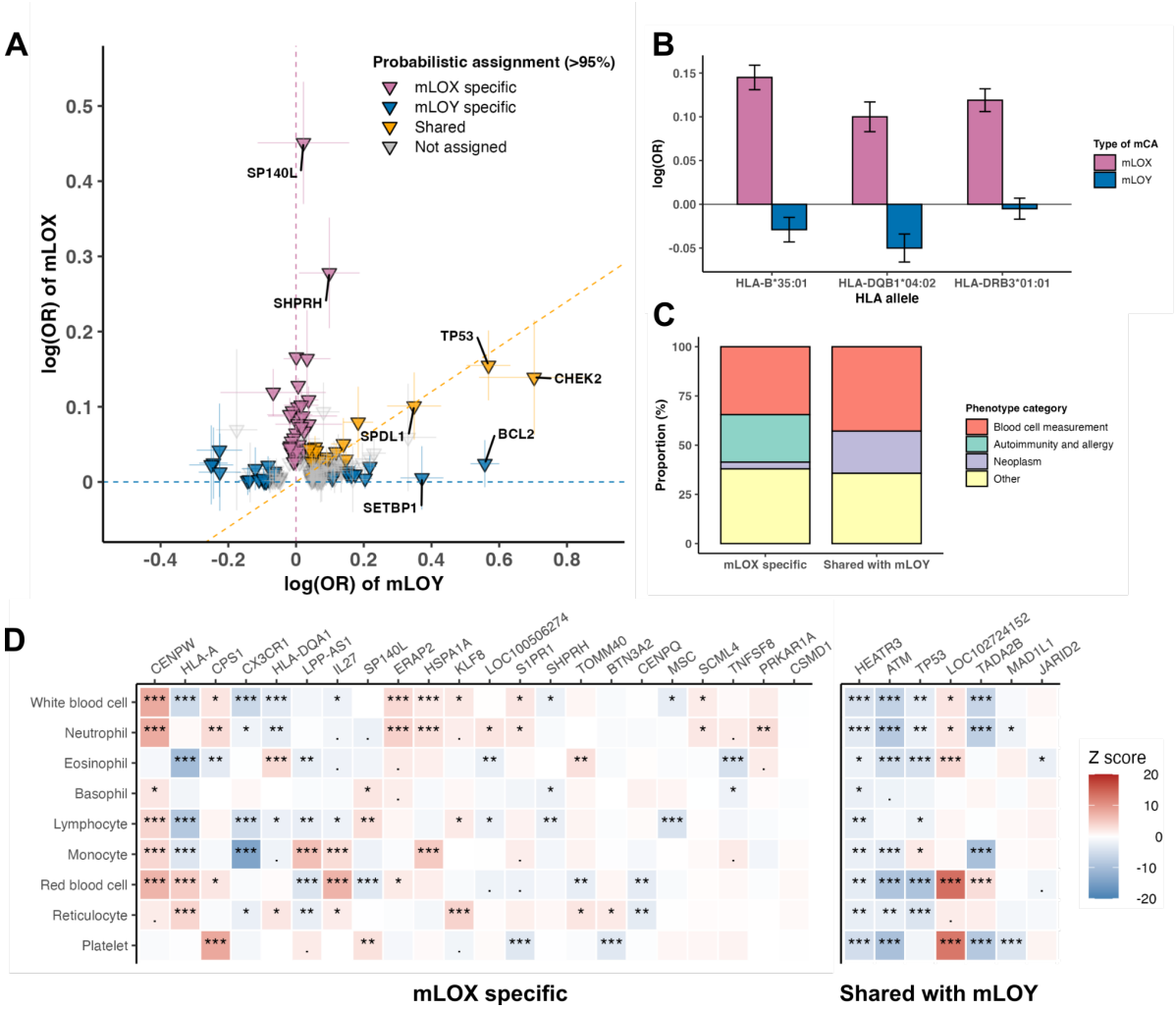

Figure 3. Shared and distinct genetic contributors to mLOX susceptibility in women and mLOY susceptibility in men.

Examination of the shared and distinct genetic contributors to mLOX in women and mLOY in men. Panel (**A**) is a scatterplot of mLOX susceptibility variants (N=56) and mLOY susceptibility variants[13] (N=147) and their effects on mLOX and mLOY. Variants are assigned to mLOX specific, mLOY specific, and shared by applying a Bayesian model with posterior probability >95%. (**B**) Fine-mapping of imputed HLA alleles for mLOX and mLOY in FinnGen, for three HLA alleles that are significantly associated with mLOX from step-wise conditional analyses. Panel (**C**) and (**D**) depict 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 the impact of 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: chr6:25.7-33.4 Mb). (**D**) Associations with nine blood cell count traits[48]. The absolute Z scores were cropped to the range of [0-20].

We next investigated if the identified common variants for mLOX susceptibility in women were associated with mLOY, the most common leukocyte sex chromosome mosaicism in men (**Supplementary Figure S8**) and likewise if mLOY loci were associated with mLOX. We employed a Bayesian model to assign 56 independent common variants identified from mLOX GWAS and 147 variants (nine variants dropped due to missing in mLOX GWAS) from the published mLOY GWAS[13] into three groups: specific to mLOX, specific to mLOY, and shared between mLOX and mLOY (**Methods; Figure 3A**). Out of 56 variants identified from the mLOX GWAS, we assigned 34 variants as specific for mLOX and seven as shared with mLOY, with greater than 95% probability (**Supplementary Table S14**). Among three centromere protein genes identified for mLOX susceptibility, *CENPQ* (for rs9395493, OR=1.04 [1.03-1.05] for mLOX and 0.99 [0.98-1.01] for mLOY, P for effect size difference=4.1×10^{-9}) and *CENPW* (for rs9372840, OR=1.04 [1.03-1.06] for mLOX and 1.02 [1.01-1.04] for mLOY, P for effect size difference=0.01) were specific to mLOX with posterior probability > 95%, while for *CENPU* (for 4:184696883:C:CT, OR=0.96 [0.94-0.97] for mLOX and 0.97 [0.95-0.98] for mLOY, P for effect size difference=0.11) the probability to be mLOX specific was 83%. When likewise examining the 147 mLOY susceptibility variants, we further identified eight variants (prioritized genes such as *SPDL1*, *HLA-A, CHEK2*, and *MAGEH1*) to be shared with mLOX susceptibility, in addition to the six variants that are exactly mLOX GWAS lead variants (prioritized genes *GRPEL1*, *QKI*, *TP53*, and *MAD1L1*) or in high 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 due to lower cell fractions observed for mLOX as compared to mLOY (**Supplementary Figure S1**). 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]) (P for effect size difference=6.0×10^{-35}). Likewise 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]) (P for effect size difference=1.1×10^{-25}). This smaller effect size together with the lower frequency of mLOX (e.g., 6.2% for 261,145 women in UKBB aged 40-70 at genotyping) relative to mLOY (e.g., 20.4% for 205,011 men in UKBB aged 40-70 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=1.7×10^{-10}) (**Methods**; **Supplementary Table S15**). We noted 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 lower cell fraction mLOX events relative to mLOY events (**Supplementary Figure S1**).

We then wanted to understand the overlaps of mLOX susceptibility variants with autosomal mosaicism, a more heterogeneous group composing any 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 (3.6% of autosomal mCA cases among 452,469 participants)[45] act in acquiring of mLOX in women. Of the 55 mLOX variants (one dropped) available in the UKBB autosomal mCA GWAS, no variant reached genome-wide significance for autosomal mCAs (**Supplementary Table S16**). Together with the identified effects on mLOY, we suggested seven of the mLOX variants as specific for mLOX susceptibility (prioritized genes *LOC100506274*, *SP140L*, *HSPA1A*, *CENPW*, *SHPRH*, *TOMM40*, and *KLF8*) and three as 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*[45], 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, P for effect size difference=0.08) while 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 (N of mLOX cases=27,001) and 128,729 Finnish male participants (N of mLOY cases=45,675) (**Methods**; **Supplementary Figure S8**). Out of 156 examined HLA alleles, 16 alleles were associated with the odds of developing detectable mLOX ($P<5.0\times10^{-8}$), including alleles from both MHC class I (six out of 74 examined alleles locating on HLA-A, -B, and -C) and class II molecules (10 out of 82 examined alleles locating on HLA-DR, -DP, and -DQ) (**Supplementary Table S17**). The most significant HLA allele HLA-B*35:01 increased the risk of mLOX (OR=1.16 [1.12-1.19], $P=1.1\times10^{-23}$), but had no effect on mLOY (OR=0.97 [0.94-1.00], P for mLOY=0.03, P for effect difference with mLOX = 3.6×10^{-18}) (**Figure 3B**). This association with HLA-B*35:01 was independently replicated in BBJ (OR= 1.10 [1.05-1.15], $P=1.5\times10^{-5}$). The HLA-B*35:01 allele is well established as the major driver for the progression of human immunodeficiency virus (HIV)[46] and also associated with several autoimmune diseases (e.g., subacute thyroiditis (OR=4.36 [3.25-5.85])[47]). With 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], $P=2.8\times10^{-19}$) and HLA-DQB1*04:02 (OR=0.90 [0.87-0.94], $P=6.5\times10^{-9}$). For mLOY in males, 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*) (r2=0.45 with HLA-B*35:01), rs9275511 (*HLA-DQA2*), rs2734971 (*HLA-G*)) identified from the Finnish population

GWAS. The results suggested that the associations with mLOX observed in the extended MHC region (GRCh38: chr6:25.7-33.4 Mb) were likely due to HLA signals instead of nearby non-HLA variants (**Supplementary Figure S9**).

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 S18**). Notably, more than half of the phenotype associations were with variants associated with blood cell trait measurements, autoimmunity and allergy, and neoplasms (**Figure 3C**). Several mLOX specific variants are GWAS lead variants of multiple autoimmune diseases such as type 1 diabetes (rs9372840 (*CENPW*) and rs181206 (*IL27*)), celiac disease (rs13080752 (*LPP-AS1*)), and rheumatoid arthritis (rs2887944 (*EOMES*)). Based on 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 (e.g., 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[48]. Of 42 independent mLOX loci (only considering the lead variant of each locus), 39 were associated with at least one of the nine blood cell count traits examined (P<0.05), suggesting a shared genetic etiology between hematopoiesis and development of detectable mLOX (**Figure 3D**). Again, the mLOX variants shared with mLOY were among the variants associated with the most number of blood cell traits (5.0 traits average over seven variants) compared to mLOX specific variants (3.3 traits average over 22 variants).

To identify rare autosomal and X chromosome germline variants (MAF < 0.1%) associated with the susceptibility of detectable mLOX, we performed gene-burden tests for our newly proposed mLOX metric which utilized information from both SNP array and WES data (mLOX 3-way combined calls) in 226,125 UKBB female participants with available WES data (**Methods**). Three non-synonymous variant functional categories were used in our analysis: high-confidence protein truncating variants (HC_PTVs), missense variants with CADD scores ≥ 25 (MISS_CADD25), and damaging variants (HC_PTV+MISS_CADD25). Only one gene, *FBXO10* (F-Box Protein 10), was associated with mLOX susceptibility ($P<1.2\times10^{-6}$) (**Figure 2B**), with the strongest association observed in carriers of missense variants with CADD scores ≥25 (N of carriers=581, beta=0.059, $P=1.8\times10^{-7}$) (**Supplementary Table S19**). Logistic regression for the dichotomous mLOX status observed a consistent effect of *FBXO10* missense variants associated with a 2-fold increased risk of acquiring mLOX (OR=2.1 [1.6-2.7], $P=1.4\times10^{-7}$), and we further confirmed this association using a distinct analytical pipeline implementing STAAR (variant-set test for association using annotation information)[49] ($P=2.5\times10^{-7}$) and SAIGE-GENE+ (scalable generalized mixed-model region-based

association test plus)[50] ($P=9.5\times10^{-8}$ for the 3-way combined quantitative measure and $P=3.0\times10^{-7}$ for the dichotomous status). A leave-one-out analysis confirmed this association was not restricted to a single coding variant ($P<3.0\times10^{-7}$). *FBXO10* is the substrate-recognition component of the SCF (SKP1-CUL1-F-box protein)-type E3 ubiquitin ligase complex. The SCF (*FBXO10*) complex mediates ubiquitination and degradation of the anti-apoptotic protein, *BCL2* (BCL2 apoptosis regulator), thereby playing a role in apoptosis by controlling the stability of *BCL2*[51].

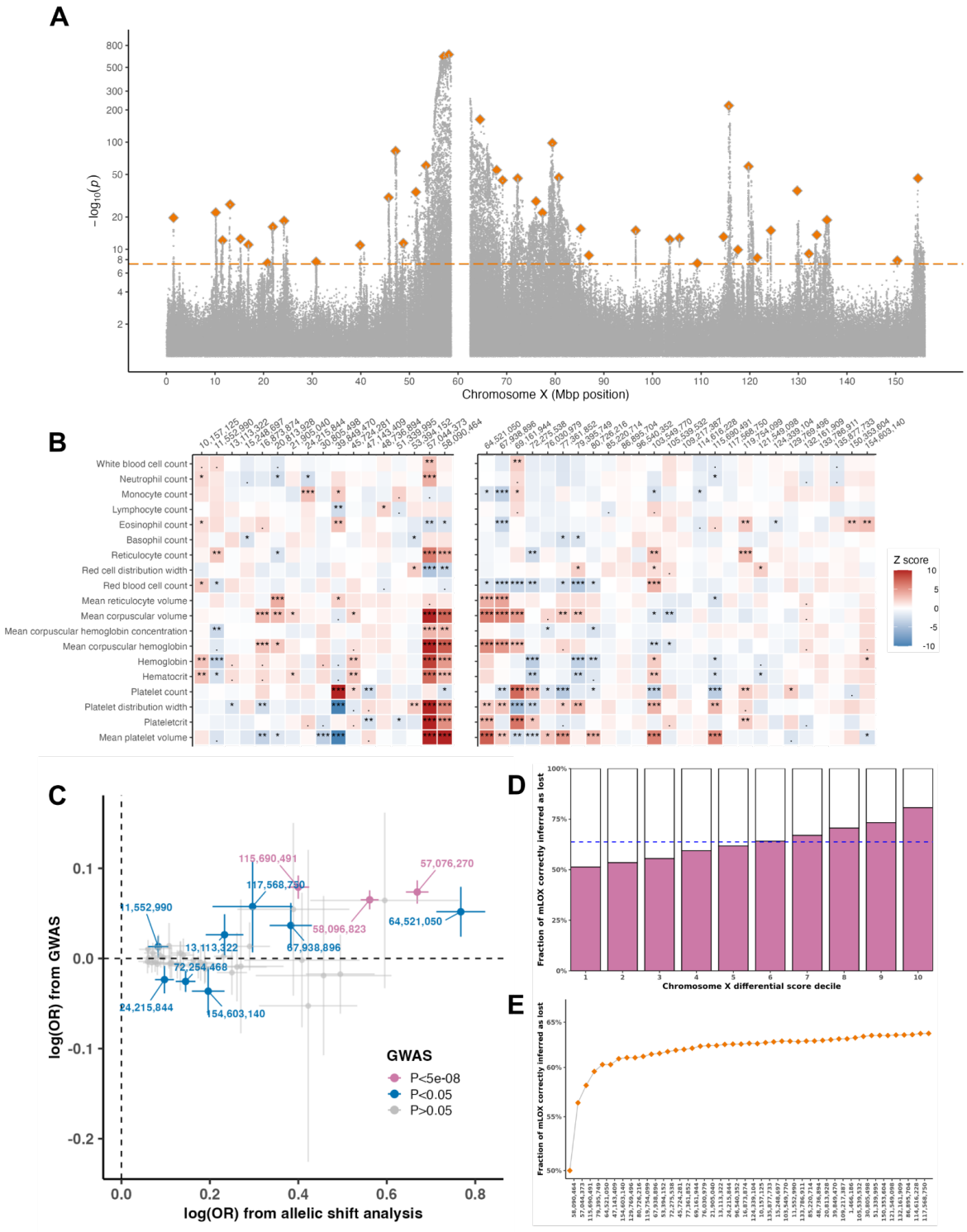

Figure 4. Allelic shift of chromosome X alleles among mLOX cases.

Panel (**A**) shows -$\log_{10}$(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 dashed line denotes the statistical significance (5.0×10^{-8}, which is the same as the GWAS significance level) and the y axis is log scale. Panel (**B**) depicts associations of 43 allelic shift analysis lead variants with 19 blood cell phenotypes[48]. 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 [0-20]. Panel (**C**) is a scatterplot of lead variants identified from allelic shift analysis (N=44) and their effects from allelic shift analysis (x axis) and GWAS (y axis). Variants are categorized based on P values from GWAS. Panel (**D**) and (**E**) show the 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 six biobanks excluding FinnGen while the prediction performance was tested in 27,001 FinnGen mLOX cases. Panel (**D**) stratifies prediction performance by differential decile of each X chromosome prediction score. Panel (**E**) shows the contribution of each lead variant to the prediction, starting with the most significant variants.

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

As several germline variants reside on the X chromosome, we sought to investigate for a given X chromosome variant whether mLOX cells with one allele retained in a hemizygous state confers a propensity to be retained or a selective advantage over mLOX cells with the alternate X allele retained (**Figure 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 in comparison to the alternate allele being retained[14] (**Methods**). This allelic shift analysis is similar to a transmission disequilibrium test[52] 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 ($P<5.0\times10^{-8}$). We identified 44 independent chromosome X variants with shifted allelic fractions on the retained X chromosome (**Methods**; **Supplementary Table S20**). The allelic shift signals spanned the length of the X chromosome (**Figure 4A**), 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], $P=4.0\times10^{-659}$). 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 1k bp and found no relationship between the two measures (**Supplementary Figure S10**). Finally, signals were consistent

across seven biobanks further supporting the robustness of the results (**Supplementary Figure S11**; **Supplementary Table S21**).

Similar to GWAS lead variants, 35 out of 43 lead variants (one variant dropped due to no appropriate proxy variant available in blood cell phenotype GWAS[48]) identified from allelic shift analyses were associated with at least one of blood cell phenotypes (prioritized genes *P2RY8*, *WAS*, *PJA1*, *PLS3*, *ITM2A*, *TMEM255A*, and *SOWAHD*) (**Supplementary Table S22**), especially for several variants near the centromere region (**Figure 4B**).

Among variants exhibiting significant allelic shifts in mLOX cases, 59 were missense variants (**Supplementary Table S23**) including 16 variants from 11 genes (*P2RY8*, *FANCB, UBA1*, *WAS*, *USP27X*, *VSIG4*, *PJA1*, *CITED1*, *POF1B*, *SAGE1*, and *MAP7D3*) likely to be lead signals (**Supplementary Figure S12**). The genes *VSIG4* (rs41307375/rs41306131 and rs17315645, r2<0.001) and *SAGE1* (rs41301507 and rs4829799, r2=0.30) each contained more than one independent missense variant. Based on 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 (e.g., *USP27X* for intellectual disability and *PJA1* for schizophrenia[53]), or had relevance to DNA damage/repair (*FANCB*) and apoptosis (*CITED1)*. Additionally, several genes were involved in X-linked recessive disorders (e.g., *FANCB* for Fanconi anemia, *WAS* for Wiskott–Aldrich syndrome, and *POF1B* for X-linked premature ovarian failure) or known to escape from X-inactivation (e.g., *P2RY8*, *UBA1*, *WAS*, *VSIG4*, and *POF1B*)[3].

Most chromosome X variants identified from the allelic shift analysis were not shared with the variants from the GWAS of mLOX (**Figure 4C**), 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, a large amount of chromosome X signals identified from allelic shift analysis suggests that in many women mLOX strongly favors one X chromosome over the other based on the differing allelic content of the two X chromosomes. This preference could arise from the clonal selection on retained alleles or could be due to allelic influences on X inactivation skewing (**Supplementary Figure S13**), which later manifests as an allelic shift if mLOX occurs since mLOX mostly affects the inactive X chromosome[10].

We then investigated how accurately we can predict which X chromosome is likely to be retained when detectable mLOX occurs. An X chromosome differential score was constructed based on 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 in 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 10^{th} percentile (**Figure 4D**). When partitioning the contribution at a variant level, starting from the most significant variants (**Figure 4E**), the fraction correctly inferred reached >60% when including the first four lead variants (rs58090464, rs57044373, rs115690491, rs79395749), while 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 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 Figure S14-S15**). To further understand whether women carrying higher X chromosome differential scores would have an elevated 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 (e.g., for major coronary heart disease event, HR=1.1 [1.1-1.2] for a one SD change in the score, $P=2.1\times10^{-5}$) and suggestive evidence for associations with myeloproliferative diseases such as polycythaemia vera (HR=1.7 [1.2-2.4], $P=1.3\times10^{-3}$) (**Supplementary Table S24**).

Discussion

This population-based analysis of approximately 900K European and Asian ancestry women indicates detectable mLOX can be observed in a substantial fraction of middle-aged and elderly women, but typically impacts less than 5% of circulating leukocytes. In an analysis of 1,253 diseases extracted from electronic health records or registry data, we identified prospective associations of mLOX with leukemia risk, specifically myeloid leukemia, and provided additional evidence for susceptibility to infectious disease such as pneumonia. Our results indicated that the value of mLOX as a diagnostic marker could be limited to blood cancers. 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 etiologically relevant to mLOX susceptibility. Little heterogeneity was noted in these loci across contributing studies or ancestry.

We identified shared and, more surprisingly, distinct genetic etiologies of mLOX with mLOY, which

occurs frequently in aging men – albeit at higher cell fractions. The two traits are moderately correlated genome-wide and seven of the 56 mLOX variants demonstrated evidence for shared effects for both mLOX and mLOY. Shared mLOX and mLOY variants were enriched for genes 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 due 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 play a role in the selection of X-linked cell surface antigens, in addition to genes relevant to mitotic missegregation (**Supplementary Figure S16**). The biological implications of shared germline susceptibility of mLOX and immunological traits could indicate the observed increased risk of pneumonia among females with mLOX is driven by pleiotropic effects; however, the mLOX-pneumonia association was restricted to a subset of mLOX females with high clonality (>10% cell fraction), suggesting mLOX could be associated with elevated infectious disease risk among high-cell fraction mLOX carriers independent of the effects of germline variation in immune-related genes.

In addition to conducting a GWAS, we also performed allelic shift analyses on X chromosome germline variants to identify signals of *cis* clonal selection. Allelic shift tests are similar to transmission disequilibrium tests commonly used in family trios and are robust to population stratification. 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. While the allelic shift centromeric signals 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, limited is known as to how germline variation in DNA sequences could impact centrosomal protein binding and spindle formation[54,55]. Other loci identified by allelic shift analyses provide support for genes involved in escaping X inactivation, cancer susceptibility, and blood cell traits as relevant to mLOX. Scores created that aggregate information across allelic shift loci correctly classified which X chromosome was more likely retained in a high percentage of mLOX women in which the difference in X chromosome scores was high. To our knowledge, this is the first demonstration of the utility of a score consisting of multiple germline variants to predict which chromosome will be impacted if a somatic event occurs. Our approach for identifying variation important for chromosome X loss may be extendable to investigating other somatic events with relevance for cancer risk.

In conclusion, we provide evidence for a strong germline component to 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 chromosome X loci that impact which X chromosome is retained and promote 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 elderly women.

Reference

1. Brown, C.J., Ballabio, A., Rupert, J.L., Lafreniere, R.G., Grompe, M., Tonlorenzi, R. and Willard, H.F., 1991. A gene from the region of the human X inactivation centre is expressed exclusively from the inactive X chromosome. Nature, 349(6304), pp.38-44.
2. Lyon, M.F., 1961. Gene action in the X-chromosome of the mouse (Mus musculus L.). Nature, 190(4773), pp.372-373.
3. Tukiainen T, Villani AC, Yen A, Rivas MA, Marshall JL, Satija R, Aguirre M, Gauthier L, Fleharty M, Kirby A, Cummings BB. Landscape of X chromosome inactivation across human tissues. Nature. 2017 Oct;550(7675):244-8.
4. Dunford, A., Weinstock, D.M., Savova, V., Schumacher, S.E., Cleary, J.P., Yoda, A., Sullivan, T.J., Hess, J.M., Gimelbrant, A.A., Beroukhim, R. and Lawrence, M.S., 2017. Tumor-suppressor genes that escape from X-inactivation contribute to cancer sex bias. Nature genetics, 49(1), pp.10-16.
5. Busque, L., Mio, R., Mattioli, J., Brais, E., Blais, N., Lalonde, Y., Maragh, M. and Gilliland, D.G., 1996. Nonrandom X-inactivation patterns in normal females: lyonization ratios vary with age. Blood, 88, 59–65.
6. Gale, R.E. and Linch, D.C., 1994. Interpretation of X-chromosome inactivation patterns. Blood, 84, 2376–2378.
7. Zito, A., Davies, M.N., Tsai, P.C., Roberts, S., Andres-Ejarque, R., Nardone, S., Bell, J.T., Wong, C.C. and Small, K.S., 2019. Heritability of skewed X-inactivation in female twins is tissue-specific and associated with age. Nature communications, 10(1), pp.1-11.
8. Forsberg, L.A., Rasi, C., Malmqvist, N., Davies, H., Pasupulati, S., Pakalapati, G., Sandgren, J., de Ståhl, T.D., Zaghlool, A., Giedraitis, V. and Lannfelt, L., 2014. Mosaic loss of chromosome Y in peripheral blood is associated with shorter survival and higher risk of cancer. Nature genetics, 46(6), pp.624-628.
9. Dumanski, J.P., Rasi, C., Lönn, M., Davies, H., Ingelsson, M., Giedraitis, V., Lannfelt, L., Magnusson, P.K., Lindgren, C.M., Morris, A.P. and Cesarini, D., 2015. Smoking is associated with mosaic loss of chromosome Y. Science, 347(6217), pp.81-83.
10. Machiela, M.J., Zhou, W., Karlins, E., Sampson, J.N., Freedman, N.D., Yang, Q., Hicks, B., Dagnall, C., Hautman, C., Jacobs, K.B. and Abnet, C.C., 2016. Female chromosome X

mosaicism is age-related and preferentially affects the inactivated X chromosome. Nature communications, 7(1), pp.1-9.

11. Zhou, W., Machiela, M.J., Freedman, N.D., Rothman, N., Malats, N., Dagnall, C., Caporaso, N., Teras, L.T., Gaudet, M.M., Gapstur, S.M. and Stevens, V.L., 2016. Mosaic loss of chromosome Y is associated with common variation near TCL1A. Nature genetics, 48(5), pp.563-568.
12. Wright, D.J., Day, F.R., Kerrison, N.D., Zink, F., Cardona, A., Sulem, P., Thompson, D.J., Sigurjonsdottir, S., Gudbjartsson, D.F., Helgason, A. and Chapman, J.R., 2017. Genetic variants associated with mosaic Y chromosome loss highlight cell cycle genes and overlap with cancer susceptibility. Nature genetics, 49(5), pp.674-679.
13. Thompson, D.J., Genovese, G., Halvardson, J., Ulirsch, J.C., Wright, D.J., Terao, C., Davidsson, O.B., Day, F.R., Sulem, P., Jiang, Y. and Danielsson, M., 2019. Genetic predisposition to mosaic Y chromosome loss in blood. Nature, 575(7784), pp.652-657.
14. Loh, P.R., Genovese, G., Handsaker, R.E., Finucane, H.K., Reshef, Y.A., Palamara, P.F., Birmann, B.M., Talkowski, M.E., Bakhoum, S.F., McCarroll, S.A. and Price, A.L., 2018. Insights into clonal haematopoiesis from 8,342 mosaic chromosomal alterations. Nature, 559(7714), pp.350-355.
15. Lin, S.H., Brown, D.W., Rose, B., Day, F., Lee, O.W., Khan, S.M., Hislop, J., Chanock, S.J., Perry, J.R. and Machiela, M.J., 2021. 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 & bioscience, 11(1), pp.1-11.
16. Zekavat, S.M., Lin, S.H., Bick, A.G., Liu, A., Paruchuri, K., Wang, C., Uddin, M.M., Ye, Y., Yu, Z., Liu, X. and Kamatani, Y., 2021. Hematopoietic mosaic chromosomal alterations increase the risk for diverse types of infection. Nature medicine, 27(6), pp.1012-1024.
17. Zhou, W., Lin, S.H., Khan, S.M., Yeager, M., Chanock, S.J. and Machiela, M.J., 2021. Detectable chromosome X mosaicism in males is rarely tolerated in peripheral leukocytes. Scientific reports, 11(1), pp.1-5.
18. Sybert, V.P. and McCauley, E., 2004. Turner's syndrome. *New England Journal of Medicine*, *351*(12), pp.1227-1238.
19. Jäger, N., Schlesner, M., Jones, D.T., Raffel, S., Mallm, J.P., Junge, K.M., Weichenhan, D., Bauer, T., Ishaque, N., Kool, M. and Northcott, P.A., 2013. Hypermutation of the inactive X chromosome is a frequent event in cancer. *Cell*, *155*(3), pp.567-581.
20. Koren, A. and McCarroll, S.A., 2014. Random replication of the inactive X chromosome. *Genome Research*, *24*(1), pp.64-69.
21. Kessler, M.D., Damask, A., O'Keeffe, S., Banerjee, N., Li, D., Watanabe, K., Marketta, A., Van Meter, M., Semrau, S., Horowitz, J. and Tang, J., 2022. Common and rare variant associations with clonal haematopoiesis phenotypes. *Nature*, pp.1-9.

22. Terao, C., Momozawa, Y., Ishigaki, K., Kawakami, E., Akiyama, M., Loh, P.R., Genovese, G., Sugishita, H., Ohta, T., Hirata, M. and Perry, J.R., 2019. GWAS of mosaic loss of chromosome Y highlights genetic effects on blood cell differentiation. Nature communications, 10(1), pp.1-10.
23. Kurki, M.I., Karjalainen, J., Palta, P., Sipilä, T.P., Kristiansson, K., Donner, K.M., Reeve, M.P., Laivuori, H., Aavikko, M., Kaunisto, M.A. and Loukola, A., 2023. FinnGen provides genetic insights from a well-phenotyped isolated population. Nature, 613(7944), pp.508-518.
24. Leitsalu, L., Haller, T., Esko, T., Tammesoo, M.L., Alavere, H., Snieder, H., Perola, M., Ng, P.C., Mägi, R., Milani, L. and Fischer, K., 2015. Cohort profile: Estonian biobank of the Estonian genome center, university of Tartu. *International journal of epidemiology*, *44*(4), pp.1137-1147.
25. Sudlow, C., Gallacher, J., Allen, N., Beral, V., Burton, P., Danesh, J., Downey, P., Elliott, P., Green, J., Landray, M. and Liu, B., 2015. UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS medicine*, *12*(3), p.e1001779.
26. Bycroft, C., Freeman, C., Petkova, D., Band, G., Elliott, L.T., Sharp, K., Motyer, A., Vukcevic, D., Delaneau, O., O'Connell, J. and Cortes, A., 2018. The UK Biobank resource with deep phenotyping and genomic data. *Nature*, *562*(7726), pp.203-209.
27. Michailidou, K., Hall, P., Gonzalez-Neira, A., Ghoussaini, M., Dennis, J., Milne, R.L., Schmidt, M.K., Chang-Claude, J., Bojesen, S.E., Bolla, M.K. and Wang, Q., 2013. Large-scale genotyping identifies 41 new loci associated with breast cancer risk. *Nature genetics*, *45*(4), pp.353-361.
28. Michailidou, K., Lindström, S., Dennis, J., Beesley, J., Hui, S., Kar, S., Lemaçon, A., Soucy, P., Glubb, D., Rostamianfar, A. and Bolla, M.K., 2017. Association analysis identifies 65 new breast cancer risk loci. *Nature*, *551*(7678), pp.92-94.
29. Gaziano, J.M., Concato, J., Brophy, M., Fiore, L., Pyarajan, S., Breeling, J., Whitbourne, S., Deen, J., Shannon, C., Humphries, D. and Guarino, P., 2016. Million Veteran Program: A mega-biobank to study genetic influences on health and disease. *Journal of clinical epidemiology*, *70*, pp.214-223.
30. Hunter-Zinck, H., Shi, Y., Li, M., Gorman, B.R., Ji, S.G., Sun, N., Webster, T., Liem, A., Hsieh, P., Devineni, P. and Karnam, P., 2020. Genotyping array design and data quality control in the million veteran program. *The American Journal of Human Genetics*, *106*(4), pp.535-548.
31. Karlson, E.W., Boutin, N.T., Hoffnagle, A.G. and Allen, N.L., 2016. Building the partners healthcare biobank at partners personalized medicine: informed consent, return of research results, recruitment lessons and operational considerations. *Journal of personalized medicine*, *6*(1), p.2.

32. Boutin, N.T., Schecter, S.B., Perez, E.F., Tchamitchian, N.S., Cerretani, X.R., Gainer, V.S., Lebo, M.S., Mahanta, L.M., Karlson, E.W. and Smoller, J.W., 2022. The Evolution of a Large Biobank at Mass General Brigham. *Journal of Personalized Medicine*, *12*(8), p.1323.
33. Machiela, M.J. et al., 2023. GWAS Explorer: an open-source tool to explore, visualize, and access GWAS summary statistics in the PLCO Atlas. Scientific Data.
34. Nagai, A., Hirata, M., Kamatani, Y., Muto, K., Matsuda, K., Kiyohara, Y., Ninomiya, T., Tamakoshi, A., Yamagata, Z., Mushiroda, T. and Murakami, Y., 2017. Overview of the BioBank Japan Project: study design and profile. *Journal of epidemiology*, *27*, pp.S2-S8.
35. Roberts, A.L., Morea, A., Amar, A., Zito, A., Moustafa, J.S.E.S., Tomlinson, M., Bowyer, R., Zhang, X., Christiansen, C., Costeira, R. and Steves, C.J., 2022. Age acquired skewed X Chromosome Inactivation is associated with adverse health outcomes in humans. medRxiv.
36. Vlasschaert, C., Mack, T., Heimlich, J.B., Niroula, A., Uddin, M.M., Weinstock, J.S., Sharber, B., Silver, A.J., Xu, Y., Savona, M.R. and Gibson, C.J., 2022. A practical approach to curate clonal hematopoiesis of indeterminate potential in human genetic datasets. *medRxiv*.
37. Vuckovic, D., Bao, E.L., Akbari, P., Lareau, C.A., Mousas, A., Jiang, T., Chen, M.H., Raffield, L.M., Tardaguila, M., Huffman, J.E. and Ritchie, S.C., 2020. The polygenic and monogenic basis of blood traits and diseases. Cell, 182(5), pp.1214-1231.
38. Frampton, M., da Silva Filho, M.I., Broderick, P., Thomsen, H., Försti, A., Vijayakrishnan, J., Cooke, R., Enciso-Mora, V., Hoffmann, P., Nöthen, M.M. and Lloyd, A., 2013. Variation at 3p24. 1 and 6q23. 3 influences the risk of Hodgkin's lymphoma. Nature communications, 4(1), p.2549.
39. Berndt, S.I., Camp, N.J., Skibola, C.F., Vijai, J., Wang, Z., Gu, J., Nieters, A., Kelly, R.S., Smedby, K.E., Monnereau, A. and Cozen, W., 2016. Meta-analysis of genome-wide association studies discovers multiple loci for chronic lymphocytic leukemia. Nature communications, 7(1), pp.1-9.
40. Celik, H., Koh, W.K., Kramer, A.C., Ostrander, E.L., Mallaney, C., Fisher, D.A., Xiang, J., Wilson, W.C., Martens, A., Kothari, A. and Fishberger, G., 2018. JARID2 functions as a tumor suppressor in myeloid neoplasms by repressing self-renewal in hematopoietic progenitor cells. *Cancer cell*, *34*(5), pp.741-756.
41. Pattabiraman, D.R. and Gonda, T.J., 2013. Role and potential for therapeutic targeting of MYB in leukemia. *Leukemia*, *27*(2), pp.269-277.
42. Schaffner, C., Stilgenbauer, S., Rappold, G.A., Döhner, H. and Lichter, P., 1999. Somatic ATM mutations indicate a pathogenic role of ATM in B-cell chronic lymphocytic leukemia. Blood, The Journal of the American Society of Hematology, 94(2), pp.748-753.
43. Zenz, T., Eichhorst, B., Busch, R., Denzel, T., Häbe, S., Winkler, D., Bühler, A., Edelmann, J., Bergmann, M., Hopfinger, G. and Hensel, M., 2010. TP53 mutation and survival in chronic lymphocytic leukemia. *Journal of Clinical Oncology*, *28*(29), pp.4473-4479.

44. Catalano, A., Dawson, M.A., Somana, K., Opat, S., Schwarer, A., Campbell, L.J. and Iland, H., 2007. The PRKAR1A gene is fused to RARA in a new variant acute promyelocytic leukemia. *Blood, The Journal of the American Society of Hematology*, *110*(12), pp.4073-4076.
45. Loh, P.R., Genovese, G. and McCarroll, S.A., 2020. Monogenic and polygenic inheritance become instruments for clonal selection. Nature, 584(7819), pp.136-141.
46. Luo, Y., Kanai, M., Choi, W., Li, X., Sakaue, S., Yamamoto, K., Ogawa, K., Gutierrez-Arcelus, M., Gregersen, P.K., Stuart, P.E. and Elder, J.T., 2021. A high-resolution HLA reference panel capturing global population diversity enables multi-ancestry fine-mapping in HIV host response. Nature Genetics, 53(10), pp.1504-1516.
47. Ritari, J., Koskela, S., Hyvärinen, K. and Partanen, J., 2022. HLA-disease association and pleiotropy landscape in over 235,000 Finns. Human Immunology, 83(5), pp.391-398.
48. Bao, E.L., Nandakumar, S.K., Liao, X., Bick, A.G., Karjalainen, J., Tabaka, M., Gan, O.I., Havulinna, A.S., Kiiskinen, T.T., Lareau, C.A. and de Lapuente Portilla, A.L., 2020. Inherited myeloproliferative neoplasm risk affects haematopoietic stem cells. Nature, 586(7831), pp.769-775.
49. Li, X., Li, Z., Zhou, H., Gaynor, S.M., Liu, Y., Chen, H., Sun, R., Dey, R., Arnett, D.K., Aslibekyan, S. and Ballantyne, C.M., 2020. Dynamic incorporation of multiple in silico functional annotations empowers rare variant association analysis of large whole-genome sequencing studies at scale. Nature genetics, 52(9), pp.969-983.
50. Zhou, W., Bi, W., Zhao, Z., Dey, K.K., Jagadeesh, K.A., Karczewski, K.J., Daly, M.J., Neale, B.M. and Lee, S., 2022. SAIGE-GENE+ improves the efficiency and accuracy of set-based rare variant association tests. Nature genetics, 54(10), pp.1466-1469.
51. Chiorazzi, M., Rui, L., Yang, Y., Ceribelli, M., Tishbi, N., Maurer, C.W., Ranuncolo, S.M., Zhao, H., Xu, W., Chan, W.C.C. and Jaffe, E.S., 2013. Related F-box proteins control cell death in Caenorhabditis elegans and human lymphoma. Proceedings of the National Academy of Sciences, 110(10), pp.3943-3948.
52. Spielman, R.S., McGinnis, R.E. and Ewens, W.J., 1993. Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM). American journal of human genetics, 52(3), p.506.
53. Trubetskoy, V., Pardiñas, A.F., Qi, T., Panagiotaropoulou, G., Awasthi, S., Bigdeli, T.B., Bryois, J., Chen, C.Y., Dennison, C.A., Hall, L.S. and Lam, M., 2022. Mapping genomic loci implicates genes and synaptic biology in schizophrenia. *Nature*, *604*(7906), pp.502-508.
54. Yang, C.H., Tomkiel, J., Saitoh, H., Johnson, D.H. and Earnshaw, W.C., 1996. Identification of overlapping DNA-binding and centromere-targeting domains in the human kinetochore protein CENP-C. Molecular and cellular biology, 16(7), pp.3576-3586.
55. Du, Y., Topp, C.N. and Dawe, R.K., 2010. DNA binding of centromere protein C (CENPC) is

stabilized by single-stranded RNA. PLoS genetics, 6(2), p.e1000835.

Online Methods

Definition of mosaic loss of the X chromosome (mLOX)

Detection of mLOX events from SNP array data in eight biobanks

All DNA samples were obtained from peripheral leukocytes and typed with single nucleotide polymorphism (SNP) arrays. The median (SD) age at sample collection for genotyping ranged from 44 (16.3) for EBB to 65 (15.8) for BBJ. The calling of mosaic loss of the X chromosome (mLOX) was performed using the Mosaic Chromosomal Alterations (MoChA) pipeline (https://github.com/freeseek/mochawdl), with GRCh38 assembly as the reference genome build. The mLOX detection ability is related to chromosome X 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 chromosome X 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-analyzed 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[56] across all batches of a biobank, except for BBJ and BCAC for which phasing was done separately within each biobank sub-cohort (for BBJ, four sub-cohorts, with cohort sizes ranging from 3,888 to 45,877; for BCAC, two 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 single-nucleotide polymorphisms (SNPs) with high levels of missingness (>3%) or heterozygote excess ($P<1.0\times10^{-6}$). 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 chromosome X events with estimated ploidy less than 2.5, where the estimated ploidy is estimated by first computing the median LRR across the assayed chromosome X SNPs and then by computing the value $2^{1+(LRR/LRR\text{-}hap2dip)}$ 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 (e.g., 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*bdev/(1+2*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 (version: 2021-08-17 and 2021-09-07) and BCAC (version: 2022-12-21). The priors of MoChA have been updated since 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 (e.g., continental ancestry, sample size, age structures, and SNP array) is available in **Supplementary Table S1**.

Estimation of X chromosome dosages from UKBB whole-exome sequence data

For UKBB, the whole-exome sequence (WES) data was released in late 2021[57], which permitted 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[58]. 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 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 (**Figure S2A-C**). As shown in **Figure S2B** and **S2C**, the participants identified as mLOX cases by MoChA exhibited lower mLRR (P from the Analysis of Variance (ANOVA) test=1.5×10^{-5}) and X dosage value ($P<1.0\times10^{-250}$) than mLOX controls. Then, for mLOX cases, we examined the relationships between three measures representing the extent of mosaicism (**Figure S2D-F**), including cell fraction (from MoChA), mLRR, and X dosage. Overall, significant correlations were observed across the three 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 three 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 mosaic loss of the Y chromosome (mLOY) study[59]) 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 (cropped to the range [0,2] as well). 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. As such, enhanced 3-way combined mLOX calls were used for UKBB in the genome-wide association study (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 body mass index (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 enrollment. We applied a logistic regression model adjusting for age (at genotyping), age^2, and the first 10 PCs 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 women and mLOY in men, 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 the time-on study as the time scale. For covariates, we recommended each biobank adjust for age, age^2, smoking, and the first 10 PCs. Meta-analysis across four biobanks was carried out with a fixed-effect model applied in the meta package[60]. 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 $P<4.0\times10^{-5}$. 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[16].

To further understand the phenotypic associations for mLOX, we applied a linear regression model adjusting for age, age^2, smoking, and the first 10 PCs 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 (minor allele frequency (MAF)>0.1%) associated with risk of detectable mLOX in peripheral leukocytes, we performed a GWAS on chromosomes 1-22 and X in each of eight contributing biobanks independently, for a total of 883,574 women. 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)[61] and for the other six biobanks (including UKBB) using regenie[62] 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 (PCs). The effective sample size, presented in Table 1, was calculated as $(4*N_{case}*N_{control})/(N_{case}+N_{control})$.

GWAS of 3-way combined quantitative mLOX measure in UKBB

For UKBB, to improve the power of GWAS, we used the new quantitative measure which 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[63].

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 HGI meta-analysis[64]. 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 software[65] which can manage the different units of dichotomous and quantitative measures. As the main analysis, we meta-analyzed summary statistics across all eight biobanks regardless of ancestry and applied Cochran's Q-test to assess the heterogeneity. To further investigate the impact of ancestry, we also conducted a meta-analysis for 7 biobanks containing only participants of European ancestry (without BBJ of 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 ($P<5\times10^{-8}$) were initially selected in 2Mb windows[66] (spanning ± 1Mb region around the most significant variant). Secondary independent signals were identified by using an approximate conditional analysis applied in GCTA[66], with LD structures constructed from UKBB samples. Secondary signals were only considered if they were genome-wide significant, in low LD ($r2<0.05$) 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[67]. 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)[68]. Third, non-coding signals and closely linked variants were then annotated if they could be mapped to known enhancers using the activity-by-contact (ABC) enhancer maps[69], 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)[70] and the Approximate Bayes Factor (ABF) method applied in the "coloc" package (version 5.1.0)[71]. These two tools were used in conjunction, as using a combination of colocalization methods has been shown to outperform single approaches[72]. To identify tissues exhibiting a significant genome-wide enrichment, we used LD score regression applied to specifically expressed gene (LDSC-SEG)[73] approach, with eQTL datasets from cross-tissue meta-analyzed GTEx

eQTL v.7[74], eQTLGen[75], and Brain-eMeta[76]. The same set of analyses were also applied to a protein quantitative trait locus (pQTL) dataset[77]. 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[78].

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 favor 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 X in 226,125 UKBB female participants with WES data available. We performed WES data pre-processing and quality control following Gardner et al.[79]. We annotated variants using the ENSEMBL Variant Effect Predictor (VEP) v104[80] 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[81]. Only non-synonymous variants with MAF < 0.1% were included in the analysis. As the main analysis, we used BOLT-LMM[61] to perform the gene-burden test. For each gene, we defined individuals with HC 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, sex, and the first 10 PCs. 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 $0.05/40{,}339=1.24\times10^{-6}$. 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[63] with STAAR (variant-set test for association using annotation information)[49] and SAIGE_GENE+ (scalable generalized mixed-model region-based association test plus)[50] 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[82]. We used summary statistics from meta-analysis of seven biobanks of European ancestry (without BBJ) and LD structures constructed from European ancestry samples of the 1000 Genomes project[83]. 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 $0.05/6,285=8.0\times10^{-6}$.

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)[84]. For LDSC, we used HapMap3[85] SNPs and LD structures constructed from 1000 Genomes project[83] 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[86]. 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: chr6:25.7-33.4 Mb).

Shared and distinct mechanisms between mLOX in women and mLOY in men

Bayesian models to cluster variants by effects on mLOX and mLOY

We employed 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 (nine variants dropped due to missing in mLOX GWAS) from the published mLOY GWAS[13] into three 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 SD 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.[87]. We conducted the association analysis between each imputed HLA allele and the dichotomous mLOX status in 168,838 Finnish female participants (N of cases = 27,001) using a multivariate logistic regression model, considering age, age^2, and the first 10 PCs as covariates. Only HLA alleles with more than 5 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 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 men (total N = 128,729, N of cases = 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, N of cases = 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 chromosome X 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[52], 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 chromosome X homologue. Therefore, we carried out a binomial test for each variant with a sample size equal to the number of women 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 1kb region across the whole X chromosome.

Identification of independent loci

Given the complexity of LD structures for X chromosomes especially for centromere and pseudoautosomal (PAR) 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 ($P<5.0\times10^{-8}$). Then, we calculated LD between every two index variants and kept the variant with a lower P value if a pair of index variants with r2<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 six 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 ($PGS_{retained}$) and the lost X chromosome (PGS_{lost}) separately and obtained the difference in PGS between the two X chromosomes ($PGS_{diff}=PGS_{lost}-PGS_{retained}$). A negative PGS_{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 (i.e., $r^2=0$). 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 $OR_i=1$ when the *i*th locus is homozygous. We then defined the X chromosome differential score PGS_{diff} with the equation: $PGS_{diff} = \Sigma_i\ log(OR_i) = \Sigma_i\ heterozygous\ log(OR_i)$, by aggregating variant effects at all simulated heterozygous genotypes. Assuming that PGS_{diff} is positive (negative), we estimated the probability P of the paternal (maternal) X chromosome being lost using the logistic function for $|PGS_{diff}|$, with $P = P / (1-P + P) = P/(1-P) / (1 + P/(1-P)) = \prod_i OR_i / (1 + \prod_i OR_i) = \exp(|PGS_{diff}|) / (1+\exp(|PGS_{diff}|))$. Given an estimated $|PGS_{diff}|$, 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 modeling linkage disequilibrium 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 women with high X differential score

We then evaluated whether women 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[23]. 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 10 PCs.

Data availability

Summary statistics generated from meta-analysis will be uploaded to the GWAS Catalog after publication. The access to individual-level data can be requested directly from each contributing biobank.

Code availability

The Mosaic Chromosomal Alterations (MoChA) pipelines used for mosaic loss of the X chromosome 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://github.com/freeseek/mochawdl. The GWAS meta-analysis was performed by using the pipeline developed by COVID-19 HGI, 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.

56. Delaneau, O., Zagury, J.F., Robinson, M.R., Marchini, J.L. and Dermitzakis, E.T., 2019. Accurate, scalable and integrative haplotype estimation. Nature communications, 10(1), pp.1-10.
57. Backman, J.D., Li, A.H., Marcketta, A., Sun, D., Mbatchou, J., Kessler, M.D., Benner, C.,

Liu, D., Locke, A.E., Balasubramanian, S. and Yadav, A., 2021. Exome sequencing and analysis of 454,787 UK Biobank participants. Nature, 599(7886), pp.628-634.

58. Zhao, Y., Gardner, E.J., Tuke, M.A., Zhang, H., Pietzner, M., Koprulu, M., Jia, R.Y., Ruth, K.S., Wood, A.R., Beaumont, R.N. and Tyrrell, J., 2022. Detection and characterization of male sex chromosome abnormalities in the UK Biobank study. Genetics in Medicine.
59. Zhao, Y., Stankovic, S., Koprulu, M., Wheeler, E., Day, F.R., Lango Allen, H., Kerrison, N.D., Pietzner, M., Loh, P.R., Wareham, N.J. and Langenberg, C., 2021. GIGYF1 loss of function is associated with clonal mosaicism and adverse metabolic health. Nature Communications, 12(1), pp.1-6.
60. Balduzzi, S., Rücker, G. and Schwarzer, G., 2019. How to perform a meta-analysis with R: a practical tutorial. Evidence-based mental health, 22(4), pp.153-160.
61. Zhou, W., Nielsen, J.B., Fritsche, L.G., Dey, R., Gabrielsen, M.E., Wolford, B.N., LeFaive, J., VandeHaar, P., Gagliano, S.A., Gifford, A. and Bastarache, L.A., 2018. Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies. Nature genetics, 50(9), pp.1335-1341.
62. Mbatchou, J., Barnard, L., Backman, J., Marcketta, A., Kosmicki, J.A., Ziyatdinov, A., Benner, C., O'Dushlaine, C., Barber, M., Boutkov, B. and Habegger, L., 2021. Computationally efficient whole-genome regression for quantitative and binary traits. Nature genetics, 53(7), pp.1097-1103.
63. Loh, P.R., Tucker, G., Bulik-Sullivan, B.K., Vilhjalmsson, B.J., Finucane, H.K., Salem, R.M., Chasman, D.I., Ridker, P.M., Neale, B.M., Berger, B. and Patterson, N., 2015A. Efficient Bayesian mixed-model analysis increases association power in large cohorts. Nature genetics, 47(3), pp.284-290.
64. COVID-19 Host Genetics Initiative. Mapping the human genetic architecture of COVID-19. Nature 600, 472–477 (2021).
65. Willer, C.J., Li, Y. and Abecasis, G.R., 2010. METAL: fast and efficient meta-analysis of genomewide association scans. Bioinformatics, 26(17), pp.2190-2191.
66. Yang, J., Ferreira, T., Morris, A.P., Medland, S.E., Madden, P.A., Heath, A.C., Martin, N.G., Montgomery, G.W., Weedon, M.N., Loos, R.J. and Frayling, T.M., 2012. Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. Nature genetics, 44(4), pp.369-375.
67. O'Leary, N.A., Wright, M.W., Brister, J.R., Ciufo, S., Haddad, D., McVeigh, R., Rajput, B., Robbertse, B., Smith-White, B., Ako-Adjei, D. and Astashyn, A., 2016. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic acids research, 44(D1), pp.D733-D745.
68. de Leeuw, C.A., Mooij, J.M., Heskes, T. and Posthuma, D., 2015. MAGMA: generalized gene-set analysis of GWAS data. PLoS computational biology, 11(4), p.e1004219.

69. Nasser, J., Bergman, D.T., Fulco, C.P., Guckelberger, P., Doughty, B.R., Patwardhan, T.A., Jones, T.R., Nguyen, T.H., Ulirsch, J.C., Lekschas, F. and Mualim, K., 2021. Genome-wide enhancer maps link risk variants to disease genes. Nature, 593(7858), pp.238-243.
70. Zhu, Z., Zhang, F., Hu, H., Bakshi, A., Robinson, M.R., Powell, J.E., Montgomery, G.W., Goddard, M.E., Wray, N.R., Visscher, P.M. and Yang, J., 2016. Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. Nature genetics, 48(5), pp.481-487.
71. Giambartolomei, C., Vukcevic, D., Schadt, E.E., Franke, L., Hingorani, A.D., Wallace, C. and Plagnol, V., 2014. Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. PLoS genetics, 10(5), p.e1004383.
72. Barbeira, A.N., Bonazzola, R., Gamazon, E.R., Liang, Y., Park, Y., Kim-Hellmuth, S., Wang, G., Jiang, Z., Zhou, D., Hormozdiari, F. and Liu, B., 2021. Exploiting the GTEx resources to decipher the mechanisms at GWAS loci. Genome biology, 22, pp.1-24.
73. Finucane, H.K., Reshef, Y.A., Anttila, V., Slowikowski, K., Gusev, A., Byrnes, A., Gazal, S., Loh, P.R., Lareau, C., Shoresh, N. and Genovese, G., 2018. Heritability enrichment of specifically expressed genes identifies disease-relevant tissues and cell types. Nature genetics, 50(4), pp.621-629.
74. GTEx Consortium, Ardlie, K.G., Deluca, D.S., Segrè, A.V., Sullivan, T.J., Young, T.R., Gelfand, E.T., Trowbridge, C.A., Maller, J.B., Tukiainen, T. and Lek, M., 2015. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. Science, 348(6235), pp.648-660.
75. Võsa, U., Claringbould, A., Westra, H.J., Bonder, M.J., Deelen, P., Zeng, B., Kirsten, H., Saha, A., Kreuzhuber, R., Brugge, H. and Oelen, R., 2021. Large-scale cis-and trans-eQTL analyses identify thousands of genetic loci and polygenic scores that regulate blood gene expression. Nature genetics, 53(9), pp.1300-1310.
76. Qi, T., Wu, Y., Zeng, J., Zhang, F., Xue, A., Jiang, L., Zhu, Z., Kemper, K., Yengo, L., Zheng, Z. and Marioni, R.E., 2018. Identifying gene targets for brain-related traits using transcriptomic and methylomic data from blood. Nature communications, 9(1), pp.1-12.
77. Pietzner, M., Wheeler, E., Carrasco-Zanini, J., Cortes, A., Koprulu, M., Wörheide, M.A., Oerton, E., Cook, J., Stewart, I.D., Kerrison, N.D. and Luan, J.A., 2021. Mapping the proteo-genomic convergence of human diseases. Science, 374(6569), p.eabj1541.
78. Weeks, E.M., Ulirsch, J.C., Cheng, N.Y., Trippe, B.L., Fine, R.S., Miao, J., Patwardhan, T.A., Kanai, M., Nasser, J., Fulco, C.P. and Tashman, K.C., 2020. Leveraging polygenic enrichments of gene features to predict genes underlying complex traits and diseases. medRxiv.
79. Gardner, E.J., Kentistou, K.A., Stankovic, S., Lockhart, S., Wheeler, E., Day, F.R., Kerrison, N.D., Wareham, N.J., Langenberg, C., O'Rahilly, S., Ong, K.K. and Perry J.R.B., 2022. Damaging missense variants in IGF1R implicate a role for IGF-1 resistance in the aetiology

of type 2 diabetes. Cell Genomics.
80. McLaren, W., Gil, L., Hunt, S.E., Riat, H.S., Ritchie, G.R., Thormann, A., Flicek, P. and Cunningham, F., 2016. The ensembl variant effect predictor. Genome biology, 17(1), pp.1-14.
81. Kircher, M., Witten, D.M., Jain, P., O'roak, B.J., Cooper, G.M. and Shendure, J., 2014. A general framework for estimating the relative pathogenicity of human genetic variants. Nature genetics, 46(3), pp.310-315.
82. Zhang, H., Wheeler, W., Hyland, P.L., Yang, Y., Shi, J., Chatterjee, N. and Yu, K., 2016. A powerful procedure for pathway-based meta-analysis using summary statistics identifies 43 pathways associated with type II diabetes in European populations. PLoS genetics, 12(6), p.e1006122.
83. 1000 Genomes Project Consortium, 2015. A global reference for human genetic variation. Nature, 526(7571), p.68.
84. Bulik-Sullivan, B., Finucane, H.K., Anttila, V., Gusev, A., Day, F.R., Loh, P.R., Duncan, L., Perry, J.R., Patterson, N., Robinson, E.B. and Daly, M.J., 2015. An atlas of genetic correlations across human diseases and traits. Nature genetics, 47(11), pp.1236-1241.
85. International HapMap 3 Consortium, 2010. Integrating common and rare genetic variation in diverse human populations. Nature, 467(7311), p.52.
86. Loh, P.R., Bhatia, G., Gusev, A., Finucane, H.K., Bulik-Sullivan, B.K., Pollack, S.J., de Candia, T.R., Lee, S.H., Wray, N.R., Kendler, K.S. and O'Donovan, M.C., 2015B. Contrasting genetic architectures of schizophrenia and other complex diseases using fast variance-components analysis. Nature genetics, 47(12), pp.1385-1392.
87. Ritari, J., Hyvärinen, K., Clancy, J., FinnGen, Partanen, J. and Koskela, S., 2020. Increasing accuracy of HLA imputation by a population-specific reference panel in a FinnGen biobank cohort. NAR genomics and bioinformatics, 2(2), p.lqaa030.

Acknowledgments We thank Juha Karjalainen (Institute for Molecular Medicine Finland (FIMM), Finland) and Mattia Cordioli (FIMM, Finland) for assistance in GWAS meta-analysis, Shea J. Andrews (Icahn School of Medicine at Mount Sinai, USA) and Jaakko Leinonen (FIMM, Finland) for kindly sharing formatted GWAS summary statistics used in genetic correlation analyses, Sakari Jukarainen (FIMM, Finland) and Alessio Gerussi (University of Milano-Bicocca, Italy) for insightful discussion on pheWAS analyses from a clinical standpoint, Samuel Jones (FIMM, Finland) and Masahiro Kanai (Broad Institute of MIT and Harvard, USA) for valuable feedback on HLA and fine-mapping, Jukka Koskela (FIMM, Finland) and Mikko Myllymäki (FIMM, Finland) for discussion on clonal hematopoiesis, Yu Fu (FIMM, Finland) and Annina Preussner (FIMM, Finland) for discussion on genetic analyses of sex chromosomes, and Geert Kops for discussion on mechanism causing chromosome missegregation. We thank Ms. Azusa Kouno in RIKEN Center for Integrative Medical

Sciences and the members of the BioBank Japan Project, headquartered in the University of Tokyo Institute of Medical Science, for supporting the BBJ analyses. We thank Bill Wheeler for his assistance in running the pathway analyses. We want to 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 Inc., AstraZeneca UK Ltd, Biogen MA Inc., Bristol Myers Squibb (and Celgene Corporation & Celgene International II Sàrl), Genentech Inc., Merck Sharp & Dohme LCC, Pfizer Inc., GlaxoSmithKline Intellectual Property Development Ltd., Sanofi US Services Inc., Maze Therapeutics Inc., Janssen Biotech Inc, Novartis AG, and Boehringer Ingelheim International GmbH. 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 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 the Fingenious® services (https://site.fingenious.fi/en/) managed by FINBB. For BCAC and MVP, the detailed acknowledgement is available in Supplementary materials.

Author contributions This project is initialized and led by A.L., G.G., P.-R.L, A.G., J.R.B.P., and M.M.. A.L. and M.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 3-way combined call GWAS in UKBB, generated Supplementary Figure S2 and S6, prepared Supplementary Table S8 and S19, and drafted the relevant result and method paragraphs. M.P

developed the Bayesian line model to cluster mLOX and mLOY loci and wrote the relevant method paragraph. M.M.Z. performed pheWAS for UKBB, MGB, and MVP and GWAS for MGB. K.K. performed the GWAS to gene pipeline, prepared Supplementary Table S12, and drafted the relevant method paragraphs. Z.Y. estimated heritability and genetic correlations and prepared Supplementary Table S15. K.Y. and L.S. performed the pathway analysis and prepared Supplementary Table S13. C.V. performed the sensitivity analyses for associations with leukemia in UKBB and prepared Supplementary Table S7. 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 generated inputs for blood cell trait heat-map (Figure 3D and 4B). G.H. performed mLOX calling, GWAS, and allelic shift analysis for EBB. B.G. and S.P. performed mLOX calling, GWAS, and allelic shift analysis 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.G. participated in FinnGen analyses. W-Y.H. and N.F. participated in PLCO analyses. E.J.G. participated in UKBB WES analyses. V.G.S. assisted in interpretating findings related to clonal hematopoiesis. A.P. coordinated the FinnGen project. H.M.O advised 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 escaping 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/conceptualized the mosaic chromosomal alteration project in FinnGen and assisted in interpretating findings especially those related to mLOY in men. A.B. supervised pheWAS in UKBB, MGB, and MVP and the sensitivity analyses for associations with leukemia 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.G., J.R.B.P, and M.M. co-supervised the project, interpreted the findings, and wrote the manuscript. For FinnGen, BCAC, and MVP, detailed author lists are available in supplementary materials. All authors reviewed the manuscript.

Funding This work was supported by the Intramural Research Program of the National Cancer Institute, National Institutes of Health, and the Medical Research Council (unit programs: MC_UU_12015/2, MC_UU_00006/2). G.G. was supported by NIH grants R01 MH104964 and R01 MH123451. A.G. 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. 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.

Competing interests G.G., P.-R.L., and S.A.M. declare competing interests: patent application PCT/WO2019/079493 has been filed on the mosaic chromosomal alterations detection method used in this work. J.R.B.P and E.J.G are employee of and hold shares in Adrestia Therapeutics. A.B. reports scientific advisory board membership for TenSixteen Bio. P.N. reports grant support from Apple, Amgen, Boston Scientific, AstraZeneca, and Novartis, personal fees from Apple, AstraZeneca, Blackstone Life Sciences, Foresite Labs, Genentech/Roche, Allelica, Novartius, scientific advisory board membership for geneXwell, Esperion Therapeutics, and TenSixteen Bio, is a scientific co-founder of TenSixteen Bio, and spousal employment at Vertex, all unrelated to the present study.

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 4th 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 (August 17 2020), AB20-5926 and amendment #1 (April 23 2020) and it´s modification (Sep 22 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 & 06.10.2020), Central Finland Biobank 1-2017, and Terveystalo Biobank STB 2018001 and amendment 25th Aug 2020. The UKBB analyses were conducted using applications 7089, 9905, and 21552.