

1 Population analyses of mosaic X chromosome loss identify genetic drivers and 2 widespread signatures of cellular selection

3 Aoxing Liu^{1,2,28}, Giulio Genovese^{2,3,4,28}, Yajie Zhao^{5,28}, Matti Pirinen^{1,6,7}, Maryam M. Zekavat^{2,8},
4 Katherine Kentistou⁵, Zhiyu Yang¹, Kai Yu⁹, Caitlyn Vlasschaert¹⁰, Xiaoxi Liu¹¹, Derek W. Brown^{9,12},
5 Georgi Hudjashov¹³, Bryan Gorman^{14,15}, Joe Dennis¹⁶, Weiyin Zhou⁹, Yukihide Momozawa¹⁷, Saiju
6 Pyarajan^{14,18}, Vlad Tuzov¹³, Fanny-Dhelia Pajuste¹³, Mervi Aavikko¹, Timo P. Sipilä¹, Awaisa
7 Ghazal¹, Wen-Yi Huang⁹, Neal Freedman⁹, Lei Song⁹, Eugene J. Gardner⁵, FinnGen, BCAC, MVP,
8 Vijay G. Sankaran^{2,19,20}, Aarno Palotie^{1,2,3,21}, Hanna M. Ollila^{1,2,22,23}, Taru Tukiainen¹, Stephen J.
9 Chanock⁹, Reedik Mägi¹³, Pradeep Natarajan^{2,8,23}, Mark J. Daly^{1,2,3,21}, Alexander Bick²⁴, Steven A.
10 McCarroll^{2,3,4}, Chikashi Terao^{11,25,26}, Po-Ru Loh^{2,18,27,29}, Andrea Ganna^{1,2,3,21,29}, John R.B. Perry^{5,29},
11 Mitchell J. Machiela^{9,29}

12 ¹Institute for Molecular Medicine Finland (FIMM), HiLIFE, University of Helsinki, Helsinki, Finland.
13 ²Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, MA,
14 USA. ³Stanley Center, Broad Institute of Harvard and MIT, Cambridge, MA, USA. ⁴Department of
15 Genetics, Harvard Medical School, Boston, MA, USA. ⁵MRC Epidemiology Unit, Institute of
16 Metabolic Science, University of Cambridge, Cambridge, UK. ⁶Department of Public Health,
17 University of Helsinki, Helsinki, Finland. ⁷Department of Mathematics and Statistics, University of
18 Helsinki, Helsinki, Finland. ⁸Cardiovascular Research Center, Massachusetts General Hospital,
19 Boston, MA, USA. ⁹Division of Cancer Epidemiology and Genetics, National Cancer Institute,
20 Rockville, MD, USA. ¹⁰Department of Medicine, Queen's University, Kingston, ON, Canada.
21 ¹¹Laboratory for Statistical and Translational Genetics, RIKEN Center for Integrative Medical
22 Sciences, Yokohama, Japan. ¹²Cancer Prevention Fellowship Program, Division of Cancer
23 Prevention, National Cancer Institute, Rockville, MD, USA. ¹³Estonian Genome Centre, Institute of
24 Genomics, University of Tartu, Tartu, Estonia. ¹⁴Center for Data and Computational Sciences (C-
25 DACS), VA Cooperative Studies Program, VA Boston Healthcare System, Boston, MA, USA. ¹⁵Booz
26 Allen Hamilton, McLean, VA, USA. ¹⁶Centre for Cancer Genetic Epidemiology, Department of
27 Public Health and Primary Care, University of Cambridge, Cambridge, UK. ¹⁷Laboratory for
28 Genotyping Development, RIKEN Center for Integrative Medical Sciences, Yokohama, Japan.
29 ¹⁸Department of Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, MA,
30 USA. ¹⁹Division of Hematology/Oncology, Boston Children's Hospital, Harvard Medical School,
31 Boston, MA, USA. ²⁰Department of Pediatric Oncology, Dana-Farber Cancer Institute, Harvard
32 Medical School, Boston, MA, USA. ²¹Analytic and Translational Genetics Unit, Massachusetts
33 General Hospital, Boston, MA, USA. ²²Anesthesia, Critical Care, and Pain Medicine, Massachusetts
34 General Hospital, Boston, MA, USA. ²³Center of Genomic Medicine, Massachusetts General
35 Hospital, Boston, MA, USA. ²⁴Division of Genetic Medicine, Department of Medicine, Vanderbilt
36 University Medical Center, Nashville, TN, USA. ²⁵Clinical Research Center, Shizuoka General
37 Hospital, Shizuoka, Japan. ²⁶Department of Applied Genetics, School of Pharmaceutical Sciences,
38 University of Shizuoka, Shizuoka, Japan. ²⁷Center for Data Sciences, Brigham and Women's
39 Hospital, Harvard Medical School, Boston, MA, USA. ²⁸These authors contributed equally: Aoxing
40 Liu, Giulio Genovese, Yajie Zhao. ²⁹These authors jointly supervised this work: Po-Ru Loh, Andrea
41 Ganna, John R.B. Perry, Mitchell J. Machiela.

42 e-mail: aoxing.liu@helsinki.fi, giulio.genovese@gmail.com, poruloh@broadinstitute.org;
43 aganna@broadinstitute.org, john.perry@mrc-epid.cam.ac.uk, mitchell.machiela@nih.gov

44

45 Mosaic loss of the X chromosome (mLOX) is the most commonly occurring clonal somatic alteration
46 detected in the leukocytes of women, yet little is known about its genetic determinants or phenotypic
47 consequences. To address this, we estimated mLOX in >900,000 women across eight biobanks,
48 identifying 10% of women with detectable X loss in approximately 2% of their leukocytes. Out of
49 1,253 diseases examined, women with mLOX had an elevated risk of myeloid and lymphoid
50 leukemias and pneumonia. Genetic analyses identified 49 common variants influencing mLOX,
51 implicating genes with established roles in chromosomal missegregation, cancer predisposition, and
52 autoimmune diseases. Complementary exome-sequence analyses identified rare missense variants in
53 *FBXO10* which confer a two-fold increased risk of mLOX. A small fraction of these associations
54 were shared with mosaic Y chromosome loss in men, suggesting different biological processes drive
55 the formation and clonal expansion of sex chromosome missegregation events. Allelic shift analyses
56 identified alleles on the X chromosome which are preferentially retained, demonstrating that variation
57 at many loci across the X chromosome is under cellular selection. A novel polygenic score including
58 44 independent X chromosome allelic shift loci correctly inferred the retained X chromosomes in
59 80.7% of mLOX cases in the top decile. Collectively our results support a model where germline
60 variants predispose women to acquiring mLOX, with the allelic content of the X chromosome
61 possibly shaping the magnitude of subsequent clonal expansion.

62

63 **Introduction**

64 Females carry a maternal and paternal copy of the X chromosome in which one copy is partially
65 rendered transcriptionally inactive early in development by expression of *Xist*¹ and epigenetic
66 modifications. The inactivation process is random as to which X chromosome is chosen with the
67 resulting inactive state being irreversible and clonally transmitted to daughter cells². X chromosome
68 inactivation has evolved as a mechanism to compensate for gene dosage imbalances between XX
69 females and XY males, although some genes are only partially inactivated³, including several tumor
70 suppressor genes (e.g., *ATRX*, *KDM5C*)⁴. Analytic challenges associated with X inactivation and
71 haploid male X chromosomes have led to fewer studies of the X chromosome, potentially missing
72 critical germline and somatic variation relevant to disease risk.

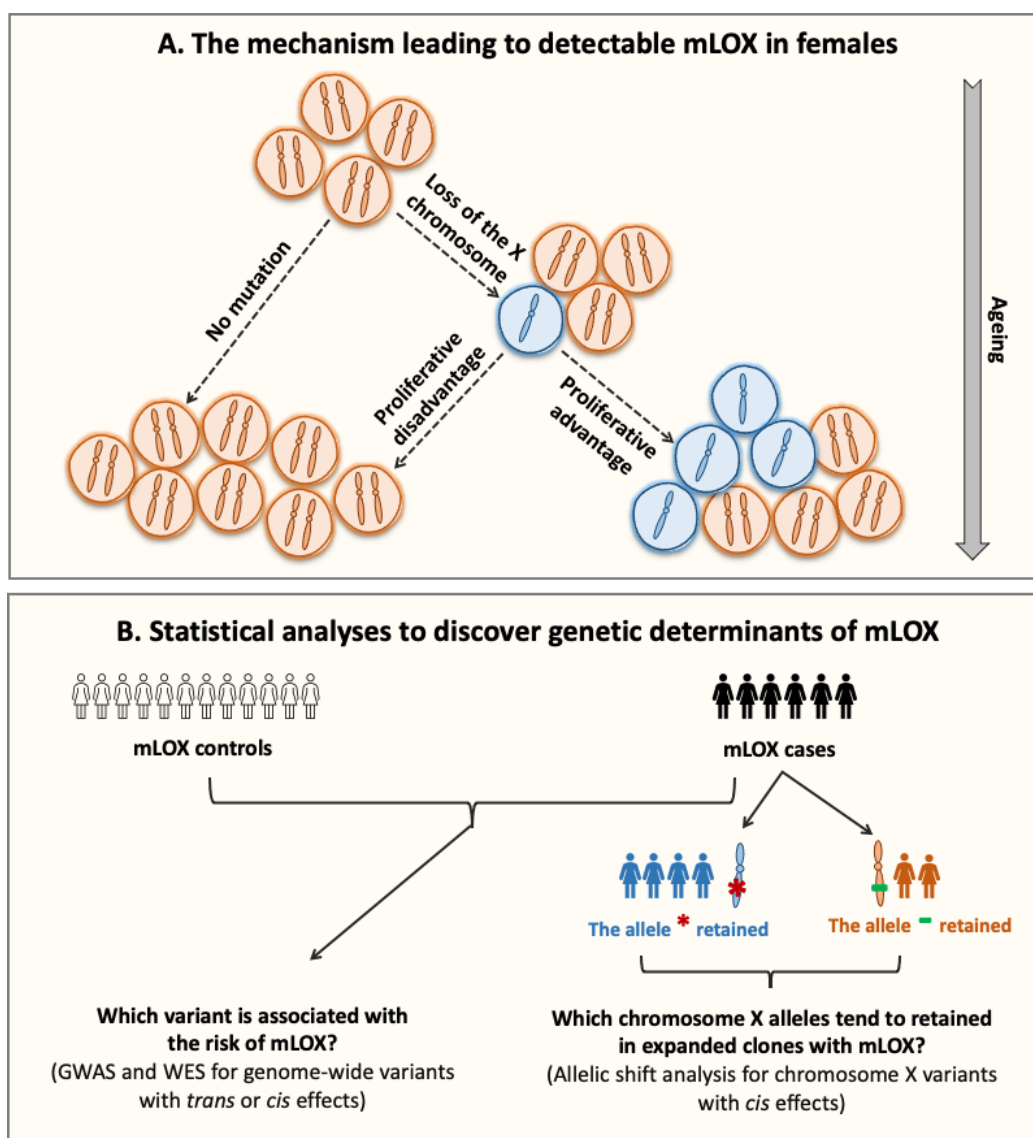
73 With age, the expected 1:1 ratio of inactivated maternal to paternal X chromosome copies can become
74 skewed. X chromosome inactivation skewing is observed in various tissues with high frequencies
75 observed in leukocytes^{5,6}. Detectable skewed X chromosome inactivation in leukocytes is heritable
76 ($h^2=0.34$)⁷ and can indicate depletion of haematopoietic stem cells, selection pressures on leukocytes,
77 or clonal hematopoiesis (CH). Recent investigations of age-related CH have described elevated rates
78 of mosaic sex chromosome aneuploidies in population-based surveys of apparently healthy adults⁸⁻¹³.
79 Mosaic loss of the female X chromosome (mLOX) is elevated in frequency compared to the

80 autosomes¹⁴, preferentially impacts the inactivated X chromosome¹⁰ and is associated with elevated
81 leukemia risk^{15,16}. This contrasts with the male X chromosome which has very low rates of
82 aneuploidy¹⁷. As the X chromosome encompasses approximately 5% of the genome and contains
83 genes relevant to immunity, cancer susceptibility, and cardiovascular diseases, loss of a homologous
84 copy and subsequent hemizygous selection could lead to downstream consequences on female health,
85 as observed in Turner syndrome (45,XO)¹⁸; however, no study has systematically examined
86 longitudinal associations of mLOX with disease risk.

87 As mLOX is a clonal pro-proliferative genomic alteration, understanding the molecular mechanisms
88 driving susceptibility to mLOX could provide new insights into the impact of aging on hematopoiesis
89 as well as hematologic cancer risk. The X chromosome, particularly the inactive X, is more frequently
90 mutated in cancer genomes¹⁹ and is late replicating relative to autosomes, potentially increasing
91 susceptibility to chromosomal alterations²⁰. While few genome-wide association studies (GWAS) of
92 mLOX have been reported to date^{14,21}, GWAS of mosaic loss of the Y chromosome (mLOY) in men
93 has identified hundreds of susceptibility loci^{11-13,22}, many of which highlight genes involved in cell
94 cycle regulation and cancer susceptibility. Here we describe insights from epidemiologic and genetic
95 analyses of X chromosome loss from a combined meta-analysis of 904,524 women. We identify 49
96 independent common susceptibility variants across 35 loci, rare missense variants of *FBXO10*
97 associated with mLOX, and 44 X chromosome loci that strongly influence which X chromosome is
98 retained. The identified signals only partially overlap with known signals for other age-related types
99 of CH. These data indicate mLOX, along with other age-related types of CH, are important pre-
100 clinical indicators of hematologic cancer risk and identify mitotic missegregation, autoimmunity,
101 blood cell trait, and cancer predisposition genes as core etiologic components for mLOX
102 susceptibility and selection.

103

104 **Results**



105

106 **Figure 1. Theoretical framework of the mLOX study.**

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

117

118 Mosaic loss of the X chromosome in eight contributed biobanks

119 We leveraged genetic data in a total of 904,524 women from eight biobanks worldwide, including
 120 European ancestry participants from FinnGen²³, Estonian Biobank (EBB)²⁴, UK Biobank (UKBB)^{25,26},
 121 Breast Cancer Association Consortium (BCAC)^{27,28}, Million Veteran Program (MVP)^{29,30}, Mass
 122 General Brigham Biobank (MGB)^{31,32}, and Prostate, Lung, Colorectal and Ovarian Cancer Screening
 123 Trial (PLCO)³³, as well as East Asian ancestry participants from Biobank Japan (BBJ)³⁴
 124 (**Supplementary Table S1**). The median (SD) age at sample collection for genotyping ranged from
 125 44 (16.3) for EBB to 67.2 (12.9) for FinnGen. We identified mLOX using the Mosaic Chromosomal
 126 Alterations (MoChA) WDL pipeline (<https://github.com/freeseek/mochawdl>), which uses raw signal
 127 intensities from single-nucleotide polymorphism (SNP) array data. Out of 904,524 women, 86,093
 128 (9.5%) were classified as cases with detectable mLOX (**Methods; Table 1**). Overall, the cell fraction
 129 of mLOX (i.e., the estimated fraction of peripheral leukocytes with X loss) was low (median=2.0%)
 130 with expanded clones having frequency $\geq 5\%$ infrequently observed (0.5% of women)
 131 (**Supplementary Figure S1**). A subset of UKBB participants (243,520 out of 261,145) also had
 132 whole-exome sequencing (WES) data available which allowed us to assess the performance of mLOX
 133 calling from MoChA. Among UKBB mLOX cases classified by MoChA, a high correlation ($r=-0.86$)
 134 was observed between cell fraction derived from SNP array data (by MoChA) and X dosage derived
 135 from WES data (**Supplementary Figure S2**). In addition to the MoChA generated dichotomous
 136 measure used by all biobanks, in UKBB we generated a 3-way combined quantitative measure by
 137 integrating independent information from both SNP array and WES data (**Methods**). The t-test
 138 statistic for association with age was increased by 29.2% with the 3-way calls, indicating improved
 139 performance relative to SNP array-only calls.

140

141 **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	67.2 (12.9)	27,001	141,837	90,732	European, Finnish
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
Breast Cancer Association Consortium (BCAC)	57 (11.3)	2,773	195,499	10,937	European
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)	64.0 (5.4)	2,672	17,178	9,249	European

142

143 Environmental determinants and epidemiological consequences

144 Like many other types of somatic mutations^{13,14}, the frequency of women with detectable mLOX in
145 peripheral leukocyte is age-related, with a frequency of 2.5% in women aged <40 and
146 reaching >32.2% after 80, averaged over all contributing biobanks (**Supplementary Figure S3** and
147 **Table S2**). To investigate the effect of lifestyle factors on the risk of acquiring detectable mLOX, we
148 assessed associations of smoking and body mass index (BMI) with mLOX in the FinnGen cohort,
149 which had an available smoking status for 50.3% of women and BMI for 18.4% of women
150 (**Methods**). Overall, ever-smokers had no increased risk of mLOX ($P=0.56$); however, an increased
151 risk was observed among ever-smokers for acquiring expanded mLOX with cell fraction $\geq 5\%$
152 ($OR=1.3$ [$1.2-1.5$], $P=6.9 \times 10^{-5}$) (**Supplementary Table S3** and **Figure S4-5**). The relationship
153 between smoking and skewed X-inactivation has not been established, as smoking was suggested as a
154 modulator for skewed XCI in the whole-blood tissue for women older than age 55⁷ but not associated
155 in the TwinsUK cohort³⁵. No associations were observed between BMI and mLOX (**Supplementary**
156 **Table S4**).

157 To evaluate disease outcomes associated with detectable mLOX, we performed Cox proportional
158 hazards regression for incident disease cases in FinnGen, UKBB, MVP, and MGB independently
159 considering genotyping age and ever-smoking status as covariates and meta-analyzed across biobanks
160 with a fixed-effect model (**Methods**). Out of the 1,253 diseases we examined, we identified mLOX
161 associations ($P < 4.0 \times 10^{-5}$) with leukemia overall ($HR=1.7$ [$1.5-2.1$], $P=3.5 \times 10^{-10}$) and chronic
162 lymphoid leukemia ($HR=3.3$ [$2.4-4.4$], $P=8.4 \times 10^{-15}$) and suggestive evidence for acute myeloid
163 leukemia (AML) ($HR=1.9$ [$1.3-2.8$], $P=1.8 \times 10^{-3}$) (**Supplementary Table S5**). Unlike the germline
164 loss of the X chromosome in women with Turner syndrome (45,XO), which can cause various
165 medical and developmental problems¹⁸, we noted limited clinical consequences for women with
166 detectable mLOX in blood.

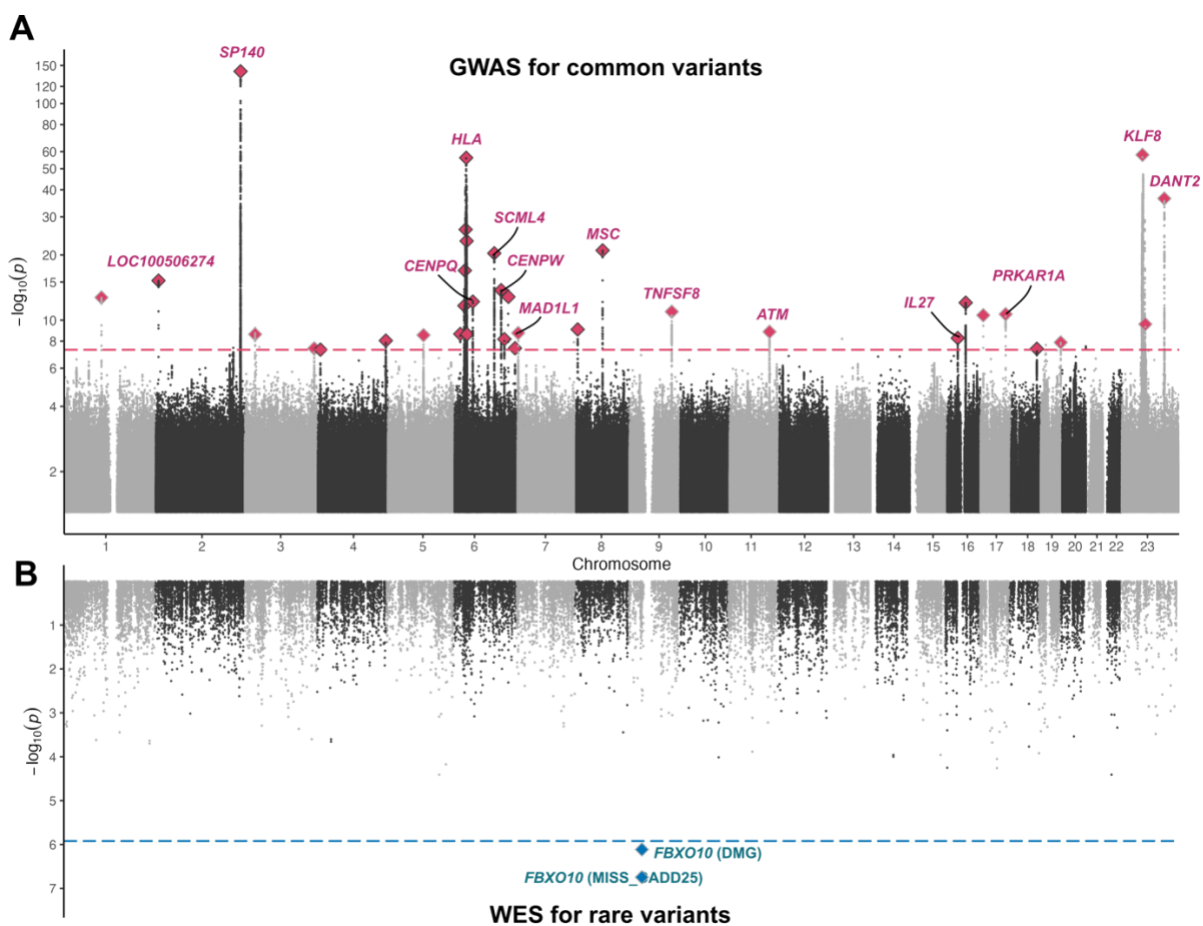
167 As the average mLOX cell fraction impacted is approximately 2%, we proposed that investigating
168 expanded mLOX with higher cell fraction ($\geq 10\%$ as previously defined¹⁶) could result in stronger
169 disease associations. Restricting to expanded mLOX, we observed evidence for elevated associations
170 with leukemia overall ($HR=6.3$ [$3.9-10.2$], $P=7.3 \times 10^{-14}$) and AML ($HR=10.6$ [$3.1-36.1$], $P=1.5 \times 10^{-4}$)
171 (**Supplementary Table S6**). We also observed suggestive evidence for associations with vitamin B
172 complex deficiency ($HR=3.7$ [$1.8-7.9$], $P=6.0 \times 10^{-4}$) and pneumonia ($HR=1.5$ [$1.2-1.8$], $P=4.7 \times 10^{-4}$),
173 especially pneumonia caused by bacterial infections ($HR=1.8$ [$1.3-2.3$], $P=3.9 \times 10^{-5}$). Similarly, in
174 UKBB¹⁶, an increased risk of incident pneumonia was observed for both women with expanded
175 mLOX ($HR=1.8$ [$1.0-3.2$], $P=0.035$) and men with expanded mLOY ($HR=1.2$ [$1.1-1.4$], $P=1.1 \times 10^{-4}$).

176 To examine the potential impacts of other types of CH on mLOX associations with leukemia, we
177 performed sensitivity analyses in UKBB where we had available calls on autosomal mosaic

178 chromosomal alterations (mCAs) as well as CH mutations in driver genes, commonly referred to as
179 clonal hematopoiesis of indeterminate potential (CHIP)³⁶. We observed attenuations in associations
180 for expanded mLOX when removing individuals with autosomal mCAs (HR=3.8 [1.6-9.3],
181 $P=2.7 \times 10^{-3}$), CHIP (HR=6.2 [3.1-12.4], $P=3.1 \times 10^{-7}$) and both mCAs and CHIP (HR=4.5 [1.9-10.8],
182 $P=8.6 \times 10^{-4}$) (**Supplementary Table S7**); however, significant associations with expanded mLOX
183 and overall leukemia risk remained indicating mLOX is independently associated with leukemia risk.
184 Associations for other lymphoid and myeloid leukemias display similar patterns, albeit losing
185 statistical significance likely due to reduced sample size (**Supplementary Table S7**).

186

187 Common and rare variants associated with mLOX susceptibility



188

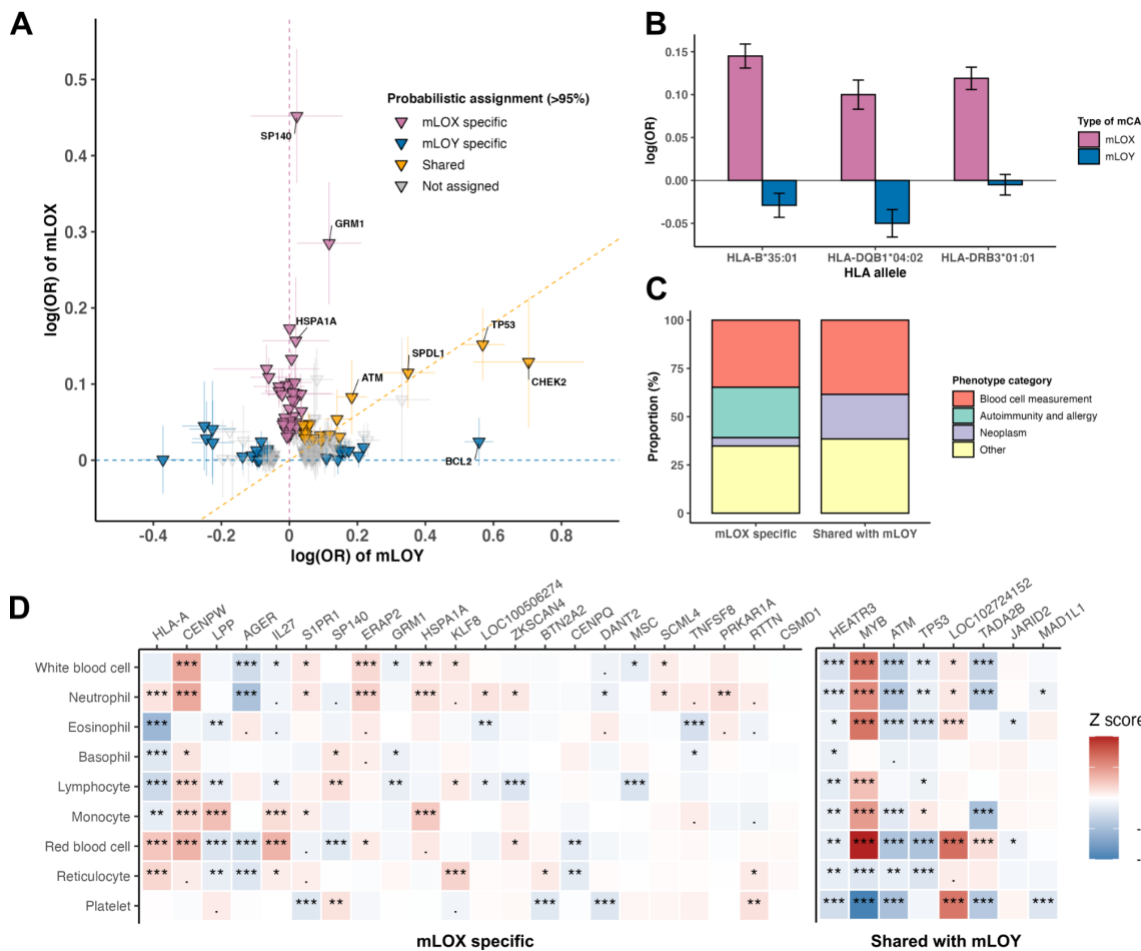
189 Figure 2. Common and rare genetic contributors to mLOX susceptibility.

190 Panel (A) shows genome-wide association study $-\log_{10}(P)$ for the association of common variants
191 (MAF>0.1%) with mLOX. Labels are only assigned for candidate genes of the top 10 lead variants
192 from meta-analysis or the top 10 candidate genes from gene prioritization and the y-axis is log scale.
193 Panel (B) presents gene burden test $-\log_{10}(P)$ for the rare variants (MAF<0.1%) associations with
194 mLOX. The dashed lines denote the statistical significance, which is 5.0×10^{-8} for GWAS (A) and
195 1.2×10^{-6} for the gene-burden test (B).

196 We performed a genome-wide association study (GWAS) to identify common and low-frequency
197 germline variants (minor allele frequency (MAF)>0.1%) associated with the risk of developing
198 detectable mLOX in peripheral leukocytes. We examined the autosomes (chromosomes 1-22) and X
199 chromosome in each of the eight contributing biobanks independently, for a total of 904,524 women
200 (**Methods**). To increase GWAS power, we used enhanced 3-way combined calls for UKBB and meta-
201 analyzed summary statistics across different mLOX measures with a weighted z-score method
202 (**Methods**). Of the 33,737,466 variants examined, we identified 49 independent genome-wide
203 significant variants ($P < 5.0 \times 10^{-8}$) across 35 loci associated with mLOX susceptibility (**Methods**;
204 **Figure 2A; Supplementary Table S8**). Most independent variants were located on chromosome 6
205 (21 variants), 2 (7 variants), 17 (3 variants), and X (3 variants). Despite differences in age-adjusted
206 mLOX frequencies, mLOX variant effects were consistent across the eight biobanks and across
207 European and East Asian ancestry (P from Cochran's Q-test $< 0.05/49 = 0.001$) (**Supplementary**
208 **Table S9**), with the exception of rs9267499 (*HSPA1A*, P from meta-analysis = 1.4×10^{-21} , P from
209 heterogeneity test = 5.8×10^{-4}) and rs78378222 (*TP53*, P from meta-analysis = 3.3×10^{-10} , P from
210 heterogeneity test = 7.1×10^{-4}). For rs9267499, the association was predominantly driven by FinnGen
211 (P from GWAS = 1.3×10^{-20}), for which the risk allele C had higher frequency (16%) in this Finnish
212 European population compared with other biobanks with either non-Finnish European (8-10%) or
213 East Asian (3%) ancestry. For rs78378222, the heterogeneity of variant effects across biobanks was
214 likely due to differences in mLOX cell fraction by contributing studies. When stratifying by cell
215 fraction in FinnGen, the OR for the risk allele of rs78378222 was 1.12 [1.03-1.21] (P=0.01) for cell
216 fractions below 5% but reached 1.73 [1.30-2.29] (P= 1.4×10^{-4}) for expanded mLOX with cell fraction
217 above 5% (P for effect size difference from a two-sided t-test = 2.5×10^{-5}) (**Supplementary Table S10**
218 and **Figure S7**).

219 We deployed a range of variant to gene mapping approaches to rank genes proximal to each of our
220 hits by their strength of evidence for causality (**Methods**), highlighting the highest-scoring gene at
221 each locus (**Supplementary Table S11**). The most significantly associated mLOX locus is at 2q37.1
222 which we mapped to *SPI40*, an interferon-inducible gene expressed at high levels in leukocytes with
223 nearby genetic variants associated with chronic lymphocytic leukemia³⁷ and autoimmune diseases^{38,39}.
224 Several identified mLOX loci implicated plausible causal genes relevant to cancer predisposition
225 including *JARID2* (6p22.3), *MYB* (6q23.3), *TNFSF8* (9q32-q33.1), *ATM* (11q22.3), *TP53* (17p13.1),
226 *PRKARIA* (17q24.2), and *KLF8* (Xp11.21), many of which (e.g., *JARID2*⁴⁰, *MYB*⁴¹, *ATM*⁴², *TP53*⁴³,
227 and *PRKARIA*⁴⁴) are directly relevant to leukemia predisposition or progression. Additionally,
228 highlighted genes at several mLOX loci are important for mitotic spindle assembly and kinetochore
229 function including *MAD1L1* (7p22.3), *CENPU* (4q35.1), *CENPQ* (6p12.3), and *CENPW* (6q22.32),
230 all of which are highly relevant to mitotic missegregation errors leading to loss of an X chromosome
231 at a single cell level. Several mLOX associated loci also implicate genes related to immunity and

232 autoimmune disorders including *EOMES* (3p24.1), *ERAP2* (5q15), *HLA-A* (6p22.1), *HLA-B*
 233 (6p21.33), *AGER* (6p21.32), *HLA-DPA1* (6p21.32), *IL27* (16p12.1-p11.2), and *LILRA1* (19q13.42),
 234 suggesting a shared etiologic relationship between mLOX and immune cell function. Similar to these
 235 locus-specific results, the genome-wide pathway-based analysis identified enrichment in pathways
 236 related to DNA damage response, cell-cycle regulation, cancer susceptibility, and immunity
 237 (Methods; Supplementary Table S12).
 238



239
 240 **Figure 3. Shared and distinct genetic contributors to mLOX susceptibility in women and mLOY**
 241 **susceptibility in men.**

242 Examination of the shared and distinct genetic contributors to mLOX in women and mLOY in men.
 243 Panel (A) is a scatterplot of mLOX susceptibility variants (N=49) and mLOY susceptibility variants¹³
 244 (N=147) and their effects on mLOX and mLOY. Variants are assigned to mLOX specific, mLOY
 245 specific, and shared by applying a Bayesian model with posterior probability >95%. (B) Fine-
 246 mapping of imputed HLA alleles for mLOX and mLOY in FinnGen, for three HLA alleles that are
 247 significantly associated with mLOX from step-wise conditional analyses. Panel (C) and (D) depict
 248 phenotype associations for lead variants of 30 independent mLOX susceptibility loci that were

249 assigned to either mLOX specific or shared with mLOY. (C) Phenotype associations (GWAS lead
250 variants ($r^2 > 0.6$)) from Open Targets genetics. To avoid the impact of pleiotropic effects, we
251 categorized phenotypes into blood cell measurement, autoimmunity and allergy, neoplasm, and
252 others. The association with each phenotype category was first examined at a variant level and then
253 summarized over all variants assigned to the same category in terms of the relationship with mLOY.
254 To avoid the associations driven by HLA signals, we excluded all identified variants from the
255 extended MHC region (GRCh38: chr6:25.7-33.4 Mb). (D) Associations with nine blood cell count
256 traits⁴⁷. The absolute Z scores were cropped to the range of [0-20].

257
258 We next investigated if the identified common variants for mLOX susceptibility in women were
259 associated with mLOY, the most common leukocyte sex chromosome mosaicism in men
260 (**Supplementary Figure S8**) and likewise if mLOY loci were associated with mLOX. We employed a
261 Bayesian model to assign 49 independent common variants identified from mLOX GWAS and 147
262 variants (nine variants dropped due to missing in mLOX GWAS) from the published mLOY GWAS¹³
263 into three groups: specific to mLOX, specific to mLOY, and shared between mLOX and mLOY
264 (**Methods; Figure 3A**). Out of 49 variants identified from the mLOX GWAS, we assigned 36
265 variants as specific for mLOX and eight as shared with mLOY, with greater than 95% probability
266 (**Supplementary Table S13**). Among three centromere protein genes identified for mLOX
267 susceptibility, *CENPQ* (for rs2448705, OR=0.96 [0.95-0.97] for mLOX and 1.00 [0.99-1.02] for
268 mLOY, P for effect size difference= 6.16×10^{-8}) and *CENPW* (for rs9398805, OR=1.04 [1.03-1.06] for
269 mLOX and 1.02 [1.01-1.04] for mLOY, P for effect size difference=0.01) were specific to mLOX
270 with posterior probability > 95%, while for *CENPU* (for rs2705883, OR=1.04 [1.03-1.06] for mLOX
271 and 1.03 [1.01-1.04] for mLOY, P for effect size difference=0.09) the probability to be mLOX
272 specific was 91%. When likewise examining the 147 mLOY susceptibility variants, we further
273 identified nine variants (prioritized genes such as *SPDL1*, *HLA-A*, *CHEK2*, and *MAGEH1*) to be
274 shared with mLOX susceptibility, in addition to the six variants that are exactly mLOX GWAS lead
275 variants (prioritized genes *GRPELI*, *QKI*, *TP53*, and *MAD1L1*) or in high LD ($r^2 > 0.6$) with mLOX
276 GWAS lead variants (prioritized genes *ATM* and *HEATR3*). Notably, for variants that are shared
277 between mLOX and mLOY, ORs were attenuated for mLOX relative to mLOY, possibly due to lower
278 cell fractions observed for mLOX as compared to mLOY (**Supplementary Figure S1**). For example,
279 for rs78378222 (*TP53*), the effect size for mLOX (OR=1.16 [1.11-1.22]) was lower than for mLOY
280 (OR=1.77 [1.65-1.88]) (P for effect size difference= 3.25×10^{-10}). Likewise for rs2280548 (*MAD1L1*),
281 the effect for mLOX (OR=1.03 [1.02-1.05]) was also lower than for mLOY (OR=1.13 [1.11-1.14]) (P
282 for effect size difference= 9.13×10^{-9}). This smaller effect size together with the lower frequency of
283 mLOX (e.g., 6.2% for 261,145 women in UKBB aged 40-70 at genotyping) relative to mLOY (e.g.,
284 20.4% for 205,011 men in UKBB aged 40-70 at genotyping¹³) indicates that a large meta-analysis was

285 needed to identify susceptibility variants for mLOX. The partially shared genetic architecture from
286 common variants between mLOX and mLOY was also supported by the moderate genetic correlation
287 ($r=0.30$ [0.20-0.40], $P=2.9\times 10^{-9}$) (**Methods; Supplementary Table S14**). We note that, in addition to
288 potential differences in biological mechanisms, the differences between mLOX and mLOY could also
289 be related to differences in cell fractions as calling algorithms can detect lower cell fraction mLOX
290 events relative to mLOY events (**Supplementary Figure S1**).

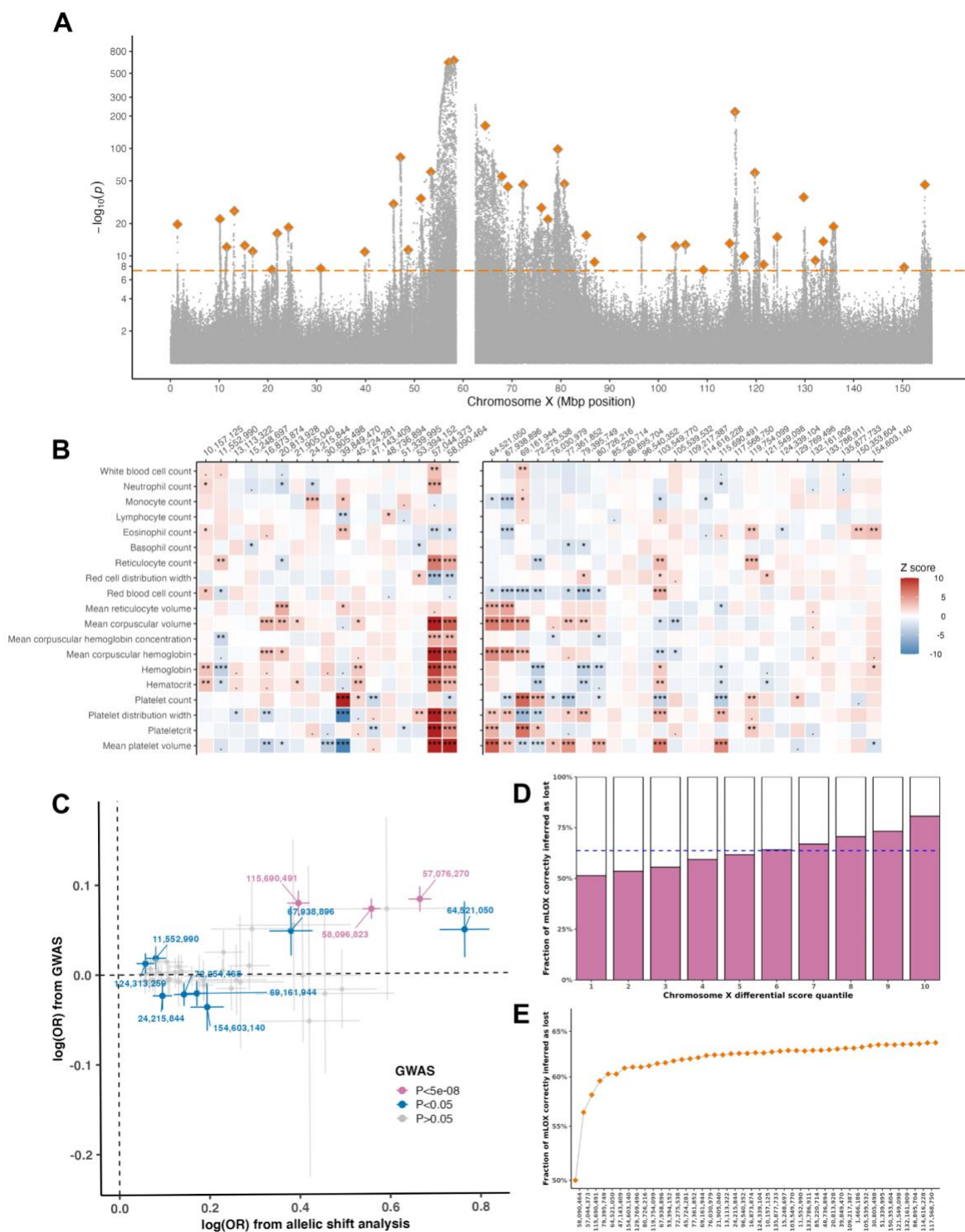
291 Given the many associations of HLA genes with mLOX, we fine-mapped HLA alleles at a unique
292 protein sequence level on 10 genes commonly used for HLA marker matching in organ
293 transplantation for a set of 168,838 Finnish female participants (N of mLOX cases=27,001) and
294 128,729 Finnish male participants (N of mLOY cases=45,675) (**Methods**). Out of 156 examined HLA
295 alleles, 16 alleles were associated with the odds of developing detectable mLOX ($P<5.0\times 10^{-8}$),
296 including alleles from both MHC class I (6 out of 74 examined alleles locating on HLA-A, -B, and -
297 C) and class II molecules (10 out of 82 examined alleles locating on HLA-DR, -DP, and -DQ)
298 (**Supplementary Table S15**). The most significant HLA allele HLA-B*35:01 increased the risk of
299 mLOX (OR=1.16 [1.12-1.19], $P=1.1\times 10^{-23}$), but had no effect on mLOY (OR=0.97 [0.94-1.00], P for
300 mLOY=0.03, P for effect difference with mLOX = 3.6×10^{-18}) (**Figure 3B**). This association with
301 HLA-B*35:01 was independently replicated in BBJ (OR= 1.10 [1.05-1.15], $P=1.5\times 10^{-5}$). The HLA-
302 B*35:01 allele is well established as the major driver for the progression of human immunodeficiency
303 virus (HIV)⁴⁵ and also associated with several autoimmune diseases (e.g., subacute thyroiditis
304 (OR=4.36 [3.25-5.85])⁴⁶). With stepwise conditional analyses in FinnGen, we identified two
305 independent genome-wide significant HLA associations at HLA-DRB3*01:01 (copy number variation
306 that presents only in a subset of individuals) (OR=0.89 [0.87-0.91], $P=2.8\times 10^{-19}$) and HLA-
307 DQB1*04:02 (OR=0.90 [0.87-0.94], $P=6.5\times 10^{-9}$). For mLOY in males, despite a larger effective
308 sample size, no HLA allele reached the genome-wide significant threshold suggesting that HLA has a
309 larger role in mLOX than mLOY. Additionally, we conducted conditional GWAS analyses in
310 FinnGen by adjusting for the three lead variants (rs74615740 (*HLA-B*) ($r^2=0.45$ with HLA-B*35:01),
311 rs9275511 (*HLA-DQA2*), rs2734971 (*HLA-G*)) identified from the Finnish population GWAS. The
312 results suggest that the associations with mLOX observed in the extended MHC region (GRCh38:
313 chr6:25.7-33.4 Mb) were likely due to HLA signals instead of nearby non-HLA variants
314 (**Supplementary Figure S9**).

315 To understand potential mechanisms relevant to mLOX susceptibility revealed by each identified
316 mLOX variant, we examined associations with additional phenotypes documented in the Open Target
317 Genetics platform (**Methods**). Out of 49 independent variants, 26 were in LD ($r^2>0.6$) with at least
318 one GWAS lead variant from Open Target (5.0×10^{-8}) (**Supplementary Table S16**). Notably, more
319 than half of the phenotype associations were with variants associated with blood cell trait
320 measurements, autoimmunity and allergy, and neoplasms (**Figure 3C**). Several mLOX specific

321 variants are GWAS lead variants of multiple autoimmune diseases such as type 1 diabetes (rs9398805
322 (*CENPW*) and rs4788084 (*IL27*)), celiac disease (rs13080752 (*LPP*)), and rheumatoid arthritis
323 (rs2371109 (*EOMES*)). Based on Open Target Genetics, none of the mLOX variants shared with
324 mLOY were reported to be associated with any autoimmune disease. Additionally, the group of
325 variants shared with mLOY have more associations with neoplasms (e.g., rs751343 (*ATM*) for breast
326 cancer and rs2280548 (*MAD1L1*) for prostate cancer) and blood cell measurements than the group of
327 variants specific for mLOX. We then examined the associations between each identified mLOX
328 susceptibility locus and the counts of different types of blood cells⁴⁷. Of 35 independent mLOX loci
329 (only considering the lead variant of each locus), 33 were associated with at least one of the nine
330 blood cell count traits examined ($P < 0.05$), suggesting a shared genetic etiology between
331 hematopoiesis and development of detectable mLOX (**Figure 3D**). Again, the mLOX variants shared
332 with mLOY were among the variants associated with the most number of blood cell traits (5.5 traits
333 average over eight variants) compared with mLOX specific variants (3 traits average over 22
334 variants).

335 To identify rare autosomal and X chromosome germline variants ($MAF < 0.1\%$) associated with the
336 risk of detectable mLOX, we performed gene-burden tests for our newly proposed mLOX metric
337 which utilized information from both SNP array and WES data (mLOX 3-way combined calls) in
338 226,125 UKBB female participants with available WES data (**Methods**). Three non-synonymous
339 variant functional categories were used in our analysis: high-confidence protein truncating variants
340 (HC PTVs), missense variants with CADD scores ≥ 25 (MISS_CADD25), and damaging variants
341 (HC_PTV+MISS_CADD25). Only one gene, *FBXO10* (F-Box Protein 10), was associated with
342 mLOX susceptibility ($P < 1.2 \times 10^{-6}$) (**Figure 2B**), with the strongest association observed in carriers of
343 missense variants with CADD scores ≥ 25 (N of carriers=581, $\beta = 0.059$, $P = 1.8 \times 10^{-7}$)
344 (**Supplementary Table S17**). Logistic regression for mLOX status observed a consistent effect of
345 *FBXO10* missense variants associated with a 2-fold increased risk of mLOX (OR=2.06 [1.59-2.68],
346 $P = 1.4 \times 10^{-7}$), and we further confirmed this association using a distinct analytical pipeline
347 implementing STAAR ($P = 2.5 \times 10^{-7}$)⁴⁸. A leave-one-out analysis confirmed this association was not
348 restricted to a single coding variant ($P < 8.5 \times 10^{-6}$). *FBXO10* is the substrate-recognition component of
349 the SCF (SKP1-CUL1-F-box protein)-type E3 ubiquitin ligase complex. The SCF (*FBXO10*) complex
350 mediates ubiquitination and degradation of the anti-apoptotic protein, *BCL2* (BCL2 apoptosis
351 regulator), thereby playing a role in apoptosis by controlling the stability of *BCL2*⁴⁹.

352



353

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

355 Panel (A) shows $-\log_{10}(P)$ of chromosome X variants from allelic shift analysis by meta-analyzing
 356 data of 83,320 mLOX cases from seven biobanks, with lead variants of 44 independent loci
 357 highlighted. The dashed line denotes the statistical significance (5.0×10^{-8} , which is the same as the
 358 GWAS significance level) and the y axis is log scale. Panel (B) depicts associations of 43 allelic shift
 359 analysis lead variants with 19 blood cell phenotypes⁴⁷. One variant was dropped due to no appropriate

360 proxy variant available in blood cell phenotype GWAS. The absolute Z scores were cropped to the
361 range of [0-20]. Panel (C) is a scatterplot of lead variants identified from allelic shift analysis (N=44)
362 and their effects from allelic shift analysis (x axis) and GWAS (y axis). Variants are categorized based
363 on P values from GWAS. Panel (D) and (E) show the fraction of mLOX cases with the retained X
364 chromosome correctly inferred using an X chromosome differential score constructed from allelic
365 shift analysis signals. To avoid overfitting, the effects of 44 lead variants were estimated from allelic
366 shift analysis of 56,319 mLOX cases from six biobanks excluding FinnGen while the prediction
367 performance was tested in 27,001 FinnGen mLOX cases. Panel (D) stratifies prediction performance
368 by differential quantile of each X chromosome prediction score. Panel (E) shows the contribution of
369 each lead variant to the prediction, starting with the most significant variants.

370

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

372 As several germline variants reside on the X chromosome, we sought to investigate for a given X
373 chromosome variant whether mLOX cells with one allele retained in a hemizygous state confers a
374 propensity to be retained or a selective advantage over mLOX cells with the alternate X allele retained
375 (**Figure 1B**). Conditional on mLOX having been detected, for each variant on the X chromosome, we
376 tested whether there is a higher frequency of a given allele retained in comparison to the alternate
377 allele being retained¹⁴ (**Methods**). This allelic shift analysis is similar to a transmission disequilibrium
378 test⁵⁰ which is robust to the presence of population structure, with only heterozygous genotypes being
379 informative. Of the 1,645,601 X chromosome variants we examined, 25,370 (1.5%) reached the
380 significance threshold ($P < 5.0 \times 10^{-8}$). We identified 44 independent chromosome X variants with
381 shifted allelic fractions on the retained X chromosome (**Methods; Supplementary Table S18**). The
382 allelic shift signals spanned the length of the X chromosome (**Figure 4A**), with the strongest signals
383 observed near the centromere (lead variant rs6612886; out of 39,246 heterozygous rs6612886
384 genotypes examined, 25,035 had the alternative C allele lost while 14,211 had the reference T allele
385 lost, OR=1.76 [1.73-1.80], $P=4.0 \times 10^{-659}$). To investigate if the observed associations were being
386 driven by inflation of the test statistic, we examined the relationship between the number of markers
387 being statistically significant and the marker density within a window size of 1k bp and found no
388 relationship between the two measures (**Supplementary Figure S10**). Similar to GWAS lead
389 variants, 35 out of 43 lead variants (one variant dropped due to no appropriate proxy variant available
390 in blood cell phenotype GWAS⁴⁷) identified from allelic shift analyses were associated with at least
391 one of blood cell phenotypes (prioritized genes *P2RY8*, *WAS*, *PJA1*, *PLS3*, *ITM2A*, *TMEM255A*, and
392 *SOWAHD*), especially for several variants near the centromere region (**Figure 4B**). Finally, signals
393 were consistent across seven biobanks further supporting the robustness of the results.

394 Among variants exhibiting significant allelic shifts in mLOX cases, 59 were missense variants

395 **(Supplementary Table S19)** including 16 variants from 11 genes (*P2RY8*, *FANCB*, *UBAI*, *WAS*,
396 *USP27X*, *VSIG4*, *PJAI*, *CITED1*, *POF1B*, *SAGE1*, and *MAP7D3*) likely to be lead signals
397 **(Supplementary Figure S11)**. The genes *VSIG4* (rs41307375/rs41306131 and rs17315645,
398 $r^2 < 0.001$) and *SAGE1* (rs41301507 and rs4829799, $r^2 = 0.30$) each contained more than one
399 independent missense variant. Based on the Human Protein Atlas (<https://www.proteinatlas.org/>),
400 several genes with identified missense variants were also associated with cancer risk/progression
401 (*P2RY8*, *UBAI*, *WAS*, and *SAGE1*), mental disorders (e.g., *USP27X* for intellectual disability and
402 *PJAI* for schizophrenia⁵¹), or had relevance to DNA damage/repair (*FANCB*) and apoptosis
403 (*CITED1*). Additionally, several genes were involved in X-linked recessive disorders (e.g., *FANCB*
404 for Fanconi anemia, *WAS* for Wiskott–Aldrich syndrome, and *POF1B* for X-linked premature ovarian
405 failure) or known to escape from X-inactivation (e.g., *P2RY8*, *UBAI*, *WAS*, *VSIG4*, and *POF1B*)³.

406 Most chromosome X variants identified from the allelic shift analysis were not shared with the
407 variants from the GWAS of mLOX **(Figure 4C)**, except for rs4029980 (X:57044373:T:C, proxy SNP
408 X:57076270:G:A, $r^2 = 0.87$) and rs6612886 (X:58090464:T:C, proxy SNP X:58096823:A:C, $r^2 = 0.98$)
409 near the centromere and rs12836051 (X:115690491:A:G). Unlike GWAS, which can identify
410 germline variants related to both chromosome missegregation and subsequent clonal selection, a large
411 amount of chromosome X signals identified from allelic shift analysis suggests that in many women
412 mLOX strongly favors one X chromosome over the other based on the differing allelic content of the
413 two X chromosomes. This preference could arise from the clonal selection on retained alleles or could
414 be due to allelic influences on X inactivation skewing **(Supplementary Figure S12)**., which later
415 manifests as an allelic shift if mLOX occurs since mLOX affects the inactive X chromosome¹⁰.

416 We then investigated how accurately we can predict which X chromosome is likely to be retained
417 when detectable mLOX occurs. The X chromosome differential score was constructed based on the
418 44 independent variants identified from allelic shift analysis by generating a chromosome-specific
419 score for each X chromosome and calculating the difference between scores of two X chromosomes
420 **(Methods)**. To avoid overfitting, the prediction performance was tested in 27,001 FinnGen mLOX
421 cases, with effect sizes of lead variants estimated from the allelic shift analysis of 56,319 mLOX cases
422 from six biobanks excluding FinnGen. The fraction of mLOX cases with the retained X chromosome
423 correctly inferred was 63.7% across all mLOX cases and up to 80.7% for mLOX cases within the top
424 10th percentile **(Figure 4D)**. When partitioning the contribution at a variant level, starting from the
425 most significant variants **(Figure 4E)**, the fraction correctly inferred reached >60% when including
426 the first four lead variants (rs58090464, rs57044373, rs115690491, rs79395749), while the
427 improvement of prediction accuracy from adding another 40 lead variants increased performance but
428 was smaller in comparison (fraction from 60.3% to 63.7%). We also performed simulation analyses to
429 assess the upper limit of prediction performance that can be reached in FinnGen mLOX cases, given
430 the distribution of allele frequencies of 44 lead variants **(Methods)**. Overall, the fraction of mLOX

431 cases correctly inferred from real data analysis (63.7%) approached that obtained from simulation
432 analysis (65.0%) (**Supplementary Figure S13-S14**).

433

434 **Discussion**

435 This population-based analysis of over 900K European and Asian ancestry women indicates
436 detectable mLOX can be observed in a substantial fraction of middle-aged and elderly women, but
437 typically impacts less than 5% of circulating leukocytes. In an analysis of 1,253 diseases extracted
438 from electronic health records or registry data, we identified prospective associations of mLOX with
439 leukemia risk, specifically myeloid leukemia, and provide additional evidence for susceptibility to
440 infectious disease. Our results indicate that the value of mLOX as a diagnostic marker could be
441 limited to blood cancers. For non-genetic risk factors, we replicated prior mLOX associations with
442 age and identified an association with tobacco smoking among high cell fraction mLOX. Our large
443 sample size coupled with an improved mLOX detection approach enabled the identification of 49
444 common independent germline susceptibility signals across 35 loci and rare coding variations in
445 *FBXO10* associated with mLOX. Little heterogeneity was noted in these loci across contributing
446 studies or ancestry. The mLOX germline susceptibility signals implicate genes involved in
447 kinetochore and spindle function, blood cell measurements, cancer predisposition, and immunity as
448 etiologically relevant to mLOX susceptibility.

449 We identified shared and, more surprisingly, distinct genetic etiologies of mLOX with mLOY, which
450 occurs frequently in aging men – albeit at higher cell fractions. The two traits are moderately
451 correlated genome-wide and eight of the 49 mLOX variants demonstrated evidence for shared effects
452 for both mLOX and mLOY. Shared mLOX and mLOY variants were enriched for genes important for
453 cancer susceptibility and blood cell traits; however, effects observed for mLOX were noticeably
454 attenuated from effects observed for mLOY. This attenuation could be due to differences in our
455 ability to detect mLOX at lower cell fractions relative to mLOY, or could be a biological impact since
456 mLOX is often present at lower cell fractions relative to mLOY. Variants specific to mLOX
457 demonstrated unique evidence for associations with immunity, including HLA alleles which could
458 play a role in the selection of X-linked cell surface antigens, as well as genes relevant to mitotic
459 missegregation (**Supplementary Figure S15**).

460 In addition to conducting a GWAS, we also performed allelic shift analyses on X chromosome
461 germline variants to identify signals of *cis* clonal selection. Allelic shift tests are similar to
462 transmission disequilibrium tests commonly used in family trios and are robust to population
463 stratification. These analyses identified strong independent signals of *cis* selection near the centromere
464 as well as multiple additional signals spanning across the X chromosome. Interestingly, the majority
465 of the allelic shift loci were not detected in the GWAS, demonstrating the ability to identify signals of

466 selection by utilizing this approach. While the allelic shift centromeric signals were strongly
467 associated with several blood cell phenotypes, their location near the centromere could tag germline
468 variation with relevance for kinetochore formation and spindle attachment in this region and may
469 predispose specific X chromosomes to missegregation errors; although, limited is known as to how
470 germline variation in DNA sequences could impact centrosomal protein binding and spindle
471 formation^{52,53}. Other loci identified by allelic shift analyses provide support for genes involved in
472 escaping X inactivation, cancer susceptibility, and blood cell traits as relevant to mLOX. Scores
473 created that aggregate information across allelic shift loci correctly classified which X chromosome
474 was more likely retained in a high percentage of mLOX women in which the difference in X
475 chromosome scores was high. To our knowledge, this is the first demonstration of the utility of a
476 score consisting of multiple germline variants to predict which chromosome will be lost if a somatic
477 event occurs. Our approach for identifying variation important for chromosome X loss may be
478 extendable to investigating other chromosomal loss events with relevance for cancer risk.

479 In conclusion, we provide evidence for a strong germline component to somatically occurring mLOX
480 in which genes related to cancer susceptibility, blood cell traits, autoimmunity, and chromosomal
481 missegregation events are relevant to mLOX susceptibility. Further, we identify many strong *cis*
482 effects for chromosome X loci that impact which X chromosome is retained and promote clonal
483 expansion. Genetic insights from mLOX could also be relevant to better understanding skewed X
484 inactivation, another commonly observed X chromosome abnormality in middle-aged and elderly
485 women.

486

487 **Reference**

- 488 1. Brown, C.J., Ballabio, A., Rupert, J.L., Lafreniere, R.G., Grompe, M., Tonlorenzi, R. and
489 Willard, H.F., 1991. A gene from the region of the human X inactivation centre is expressed
490 exclusively from the inactive X chromosome. *Nature*, 349(6304), pp.38-44.
- 491 2. Lyon, M.F., 1961. Gene action in the X-chromosome of the mouse (*Mus musculus* L.).
492 *Nature*, 190(4773), pp.372-373.
- 493 3. Tukiainen T, Villani AC, Yen A, Rivas MA, Marshall JL, Satija R, Aguirre M, Gauthier L,
494 Fleharty M, Kirby A, Cummings BB. Landscape of X chromosome inactivation across human
495 tissues. *Nature*. 2017 Oct;550(7675):244-8.
- 496 4. Dunford, A., Weinstock, D.M., Savova, V., Schumacher, S.E., Cleary, J.P., Yoda, A.,
497 Sullivan, T.J., Hess, J.M., Gimelbrant, A.A., Beroukhir, R. and Lawrence, M.S., 2017.
498 Tumor-suppressor genes that escape from X-inactivation contribute to cancer sex bias. *Nature*
499 *genetics*, 49(1), pp.10-16.
- 500 5. Busque, L., Mio, R., Mattioli, J., Brais, E., Blais, N., Lalonde, Y., Maragh, M. and Gilliland,

- 501 D.G., 1996. Nonrandom X-inactivation patterns in normal females: lyonization ratios vary
502 with age. *Blood*, 88, 59–65.
- 503 6. Gale, R.E. and Linch, D.C., 1994. Interpretation of X-chromosome inactivation patterns.
504 *Blood*, 84, 2376–2378.
- 505 7. Zito, A., Davies, M.N., Tsai, P.C., Roberts, S., Andres-Ejarque, R., Nardone, S., Bell, J.T.,
506 Wong, C.C. and Small, K.S., 2019. Heritability of skewed X-inactivation in female twins is
507 tissue-specific and associated with age. *Nature communications*, 10(1), pp.1-11.
- 508 8. Forsberg, L.A., Rasi, C., Malmqvist, N., Davies, H., Pasupulati, S., Pakalapati, G., Sandgren,
509 J., de Ståhl, T.D., Zaghlool, A., Giedraitis, V. and Lannfelt, L., 2014. Mosaic loss of
510 chromosome Y in peripheral blood is associated with shorter survival and higher risk of
511 cancer. *Nature genetics*, 46(6), pp.624-628.
- 512 9. Dumanski, J.P., Rasi, C., Lönn, M., Davies, H., Ingelsson, M., Giedraitis, V., Lannfelt, L.,
513 Magnusson, P.K., Lindgren, C.M., Morris, A.P. and Cesarini, D., 2015. Smoking is associated
514 with mosaic loss of chromosome Y. *Science*, 347(6217), pp.81-83.
- 515 10. Machiela, M.J., Zhou, W., Karlins, E., Sampson, J.N., Freedman, N.D., Yang, Q., Hicks, B.,
516 Dagnall, C., Hautman, C., Jacobs, K.B. and Abnet, C.C., 2016. Female chromosome X
517 mosaicism is age-related and preferentially affects the inactivated X chromosome. *Nature*
518 *communications*, 7(1), pp.1-9.
- 519 11. Zhou, W., Machiela, M.J., Freedman, N.D., Rothman, N., Malats, N., Dagnall, C., Caporaso,
520 N., Teras, L.T., Gaudet, M.M., Gapstur, S.M. and Stevens, V.L., 2016. Mosaic loss of
521 chromosome Y is associated with common variation near *TCL1A*. *Nature genetics*, 48(5),
522 pp.563-568.
- 523 12. Wright, D.J., Day, F.R., Kerrison, N.D., Zink, F., Cardona, A., Sulem, P., Thompson, D.J.,
524 Sigurjonsdottir, S., Gudbjartsson, D.F., Helgason, A. and Chapman, J.R., 2017. Genetic
525 variants associated with mosaic Y chromosome loss highlight cell cycle genes and overlap
526 with cancer susceptibility. *Nature genetics*, 49(5), pp.674-679.
- 527 13. Thompson, D.J., Genovese, G., Halvardson, J., Ulirsch, J.C., Wright, D.J., Terao, C.,
528 Davidsson, O.B., Day, F.R., Sulem, P., Jiang, Y. and Danielsson, M., 2019. Genetic
529 predisposition to mosaic Y chromosome loss in blood. *Nature*, 575(7784), pp.652-657.
- 530 14. Loh, P.R., Genovese, G., Handsaker, R.E., Finucane, H.K., Reshef, Y.A., Palamara, P.F.,
531 Birmann, B.M., Talkowski, M.E., Bakhoun, S.F., McCarroll, S.A. and Price, A.L., 2018.
532 Insights into clonal haematopoiesis from 8,342 mosaic chromosomal alterations. *Nature*,
533 559(7714), pp.350-355.
- 534 15. Lin, S.H., Brown, D.W., Rose, B., Day, F., Lee, O.W., Khan, S.M., Hislop, J., Chanock, S.J.,
535 Perry, J.R. and Machiela, M.J., 2021. Incident disease associations with mosaic chromosomal
536 alterations on autosomes, X and Y chromosomes: insights from a phenome-wide association
537 study in the UK Biobank. *Cell & bioscience*, 11(1), pp.1-11.

- 538 16. Zekavat, S.M., Lin, S.H., Bick, A.G., Liu, A., Paruchuri, K., Wang, C., Uddin, M.M., Ye, Y.,
539 Yu, Z., Liu, X. and Kamatani, Y., 2021. Hematopoietic mosaic chromosomal alterations
540 increase the risk for diverse types of infection. *Nature medicine*, 27(6), pp.1012-1024.
- 541 17. Zhou, W., Lin, S.H., Khan, S.M., Yeager, M., Chanock, S.J. and Machiela, M.J., 2021.
542 Detectable chromosome X mosaicism in males is rarely tolerated in peripheral leukocytes.
543 *Scientific reports*, 11(1), pp.1-5.
- 544 18. Sybert, V.P. and McCauley, E., 2004. Turner's syndrome. *New England Journal of Medicine*,
545 351(12), pp.1227-1238.
- 546 19. Jäger, N., Schlesner, M., Jones, D.T., Raffel, S., Mallm, J.P., Junge, K.M., Weichenhan, D.,
547 Bauer, T., Ishaque, N., Kool, M. and Northcott, P.A., 2013. Hypermutation of the inactive X
548 chromosome is a frequent event in cancer. *Cell*, 155(3), pp.567-581.
- 549 20. Koren, A. and McCarroll, S.A., 2014. Random replication of the inactive X chromosome.
550 *Genome Research*, 24(1), pp.64-69.
- 551 21. Kessler, M.D., Damask, A., O’Keeffe, S., Banerjee, N., Li, D., Watanabe, K., Marketta, A.,
552 Van Meter, M., Semrau, S., Horowitz, J. and Tang, J., 2022. Common and rare variant
553 associations with clonal haematopoiesis phenotypes. *Nature*, pp.1-9.
- 554 22. Terao, C., Momozawa, Y., Ishigaki, K., Kawakami, E., Akiyama, M., Loh, P.R., Genovese,
555 G., Sugishita, H., Ohta, T., Hirata, M. and Perry, J.R., 2019. GWAS of mosaic loss of
556 chromosome Y highlights genetic effects on blood cell differentiation. *Nature*
557 *communications*, 10(1), pp.1-10.
- 558 23. Kurki, M.I., Karjalainen, J., Palta, P., Sipilä, T.P., Kristiansson, K., Donner, K., Reeve, M.P.,
559 Laivuori, H., Aavikko, M., Kaunisto, M.A. and Loukola, A., 2022. FinnGen: Unique genetic
560 insights from combining isolated population and national health register data. *medRxiv*.
- 561 24. Leitsalu, L., Haller, T., Esko, T., Tammesoo, M.L., Alavere, H., Snieder, H., Perola, M., Ng,
562 P.C., Mägi, R., Milani, L. and Fischer, K., 2015. Cohort profile: Estonian biobank of the
563 Estonian genome center, university of Tartu. *International journal of epidemiology*, 44(4),
564 pp.1137-1147.
- 565 25. Sudlow, C., Gallacher, J., Allen, N., Beral, V., Burton, P., Danesh, J., Downey, P., Elliott, P.,
566 Green, J., Landray, M. and Liu, B., 2015. UK biobank: an open access resource for
567 identifying the causes of a wide range of complex diseases of middle and old age. *PLoS*
568 *medicine*, 12(3), p.e1001779.
- 569 26. Bycroft, C., Freeman, C., Petkova, D., Band, G., Elliott, L.T., Sharp, K., Motyer, A.,
570 Vukcevic, D., Delaneau, O., O’Connell, J. and Cortes, A., 2018. The UK Biobank resource
571 with deep phenotyping and genomic data. *Nature*, 562(7726), pp.203-209.
- 572 27. Michailidou, K., Hall, P., Gonzalez-Neira, A., Ghoussaini, M., Dennis, J., Milne, R.L.,
573 Schmidt, M.K., Chang-Claude, J., Bojesen, S.E., Bolla, M.K. and Wang, Q., 2013. Large-
574 scale genotyping identifies 41 new loci associated with breast cancer risk. *Nature genetics*,

- 575 45(4), pp.353-361.
- 576 28. Michailidou, K., Lindström, S., Dennis, J., Beesley, J., Hui, S., Kar, S., Lemaçon, A., Soucy,
577 P., Glubb, D., Rostamianfar, A. and Bolla, M.K., 2017. Association analysis identifies 65 new
578 breast cancer risk loci. *Nature*, 551(7678), pp.92-94.
- 579 29. Gaziano, J.M., Concato, J., Brophy, M., Fiore, L., Pyarajan, S., Breeling, J., Whitbourne, S.,
580 Deen, J., Shannon, C., Humphries, D. and Guarino, P., 2016. Million Veteran Program: A
581 mega-biobank to study genetic influences on health and disease. *Journal of clinical*
582 *epidemiology*, 70, pp.214-223.
- 583 30. Hunter-Zinck, H., Shi, Y., Li, M., Gorman, B.R., Ji, S.G., Sun, N., Webster, T., Liem, A.,
584 Hsieh, P., Devineni, P. and Karnam, P., 2020. Genotyping array design and data quality
585 control in the million veteran program. *The American Journal of Human Genetics*, 106(4),
586 pp.535-548.
- 587 31. Karlson, E.W., Boutin, N.T., Hoffnagle, A.G. and Allen, N.L., 2016. Building the partners
588 healthcare biobank at partners personalized medicine: informed consent, return of research
589 results, recruitment lessons and operational considerations. *Journal of personalized medicine*,
590 6(1), p.2.
- 591 32. Boutin, N.T., Schechter, S.B., Perez, E.F., Tchamitchian, N.S., Cerretani, X.R., Gainer, V.S.,
592 Lebo, M.S., Mahanta, L.M., Karlson, E.W. and Smoller, J.W., 2022. The Evolution of a
593 Large Biobank at Mass General Brigham. *Journal of Personalized Medicine*, 12(8), p.1323.
- 594 33. Machiela, M.J. et al., 2023. GWAS Explorer: an open-source tool to explore, visualize, and
595 access GWAS summary statistics in the PLCO Atlas. *Scientific Data*.
- 596 34. Nagai, A., Hirata, M., Kamatani, Y., Muto, K., Matsuda, K., Kiyohara, Y., Ninomiya, T.,
597 Tamakoshi, A., Yamagata, Z., Mushirola, T. and Murakami, Y., 2017. Overview of the
598 BioBank Japan Project: study design and profile. *Journal of epidemiology*, 27, pp.S2-S8.
- 599 35. Roberts, A.L., Morea, A., Amar, A., Zito, A., Moustafa, J.S.E.S., Tomlinson, M., Bowyer, R.,
600 Zhang, X., Christiansen, C., Costeira, R. and Steves, C.J., 2022. Age acquired skewed X
601 Chromosome Inactivation is associated with adverse health outcomes in humans. medRxiv.
- 602 36. Vlasschaert, C., Mack, T., Heimlich, J.B., Niroula, A., Uddin, M.M., Weinstock, J.S.,
603 Sharber, B., Silver, A.J., Xu, Y., Savona, M.R. and Gibson, C.J., 2022. A practical approach
604 to curate clonal hematopoiesis of indeterminate potential in human genetic datasets. medRxiv.
- 605 37. Berndt, S.I., Skibola, C.F., Joseph, V., Camp, N.J., Nieters, A., Wang, Z., Cozen, W.,
606 Monnereau, A., Wang, S.S., Kelly, R.S. and Lan, Q., 2013. Genome-wide association study
607 identifies multiple risk loci for chronic lymphocytic leukemia. *Nature genetics*, 45(8), pp.868-
608 876.
- 609 38. International Multiple Sclerosis Genetics Consortium*†, ANZgene, IIBDGC and WTCCC2,
610 2019. Multiple sclerosis genomic map implicates peripheral immune cells and microglia in
611 susceptibility. *Science*, 365(6460), p.eaav7188.

- 612 39. Liu, J.Z., Van Sommeren, S., Huang, H., Ng, S.C., Alberts, R., Takahashi, A., Ripke, S., Lee,
613 J.C., Jostins, L., Shah, T. and Abedian, S., 2015. Association analyses identify 38
614 susceptibility loci for inflammatory bowel disease and highlight shared genetic risk across
615 populations. *Nature genetics*, 47(9), pp.979-986.
- 616 40. Celik, H., Koh, W.K., Kramer, A.C., Ostrander, E.L., Mallaney, C., Fisher, D.A., Xiang, J.,
617 Wilson, W.C., Martens, A., Kothari, A. and Fishberger, G., 2018. JARID2 functions as a
618 tumor suppressor in myeloid neoplasms by repressing self-renewal in hematopoietic
619 progenitor cells. *Cancer cell*, 34(5), pp.741-756.
- 620 41. Pattabiraman, D.R. and Gonda, T.J., 2013. Role and potential for therapeutic targeting of
621 MYB in leukemia. *Leukemia*, 27(2), pp.269-277.
- 622 42. Schaffner, C., Stilgenbauer, S., Rappold, G.A., Döhner, H. and Lichter, P., 1999. Somatic
623 ATM mutations indicate a pathogenic role of ATM in B-cell chronic lymphocytic leukemia.
624 *Blood, The Journal of the American Society of Hematology*, 94(2), pp.748-753.
- 625 43. Zenz, T., Eichhorst, B., Busch, R., Denzel, T., Häbe, S., Winkler, D., Bühler, A., Edelmann,
626 J., Bergmann, M., Hopfinger, G. and Hensel, M., 2010. TP53 mutation and survival in
627 chronic lymphocytic leukemia. *Journal of Clinical Oncology*, 28(29), pp.4473-4479.
- 628 44. Catalano, A., Dawson, M.A., Somana, K., Opat, S., Schwarzer, A., Campbell, L.J. and Iland,
629 H., 2007. The PRKAR1A gene is fused to RARA in a new variant acute promyelocytic
630 leukemia. *Blood, The Journal of the American Society of Hematology*, 110(12), pp.4073-
631 4076.
- 632 45. Luo, Y., Kanai, M., Choi, W., Li, X., Sakaue, S., Yamamoto, K., Ogawa, K., Gutierrez-
633 Arcelus, M., Gregersen, P.K., Stuart, P.E. and Elder, J.T., 2021. A high-resolution HLA
634 reference panel capturing global population diversity enables multi-ancestry fine-mapping in
635 HIV host response. *Nature Genetics*, 53(10), pp.1504-1516.
- 636 46. Ritari, J., Koskela, S., Hyvärinen, K. and Partanen, J., 2022. HLA-disease association and
637 pleiotropy landscape in over 235,000 Finns. *Human Immunology*, 83(5), pp.391-398.
- 638 47. Bao, E.L., Nandakumar, S.K., Liao, X., Bick, A.G., Karjalainen, J., Tabaka, M., Gan, O.I.,
639 Havulinna, A.S., Kiiskinen, T.T., Lareau, C.A. and de Lapuente Portilla, A.L., 2020. Inherited
640 myeloproliferative neoplasm risk affects haematopoietic stem cells. *Nature*, 586(7831),
641 pp.769-775.
- 642 48. Li, X., Li, Z., Zhou, H., Gaynor, S.M., Liu, Y., Chen, H., Sun, R., Dey, R., Arnett, D.K.,
643 Aslibekyan, S. and Ballantyne, C.M., 2020. Dynamic incorporation of multiple in silico
644 functional annotations empowers rare variant association analysis of large whole-genome
645 sequencing studies at scale. *Nature genetics*, 52(9), pp.969-983.
- 646 49. Chiorazzi, M., Rui, L., Yang, Y., Ceribelli, M., Tishbi, N., Maurer, C.W., Ranuncolo, S.M.,
647 Zhao, H., Xu, W., Chan, W.C.C. and Jaffe, E.S., 2013. Related F-box proteins control cell
648 death in *Caenorhabditis elegans* and human lymphoma. *Proceedings of the National Academy*

- 649 of Sciences, 110(10), pp.3943-3948.
- 650 50. Spielman, R.S., McGinnis, R.E. and Ewens, W.J., 1993. Transmission test for linkage
651 disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM).
652 American journal of human genetics, 52(3), p.506.
- 653 51. Trubetskoy, V., Pardiñas, A.F., Qi, T., Panagiotaropoulou, G., Awasthi, S., Bigdeli, T.B.,
654 Bryois, J., Chen, C.Y., Dennison, C.A., Hall, L.S. and Lam, M., 2022. Mapping genomic loci
655 implicates genes and synaptic biology in schizophrenia. *Nature*, 604(7906), pp.502-508.
- 656 52. Yang, C.H., Tomkiel, J., Saitoh, H., Johnson, D.H. and Earnshaw, W.C., 1996. Identification
657 of overlapping DNA-binding and centromere-targeting domains in the human kinetochore
658 protein CENP-C. *Molecular and cellular biology*, 16(7), pp.3576-3586.
- 659 53. Du, Y., Topp, C.N. and Dawe, R.K., 2010. DNA binding of centromere protein C (CENPC) is
660 stabilized by single-stranded RNA. *PLoS genetics*, 6(2), p.e1000835.
- 661

662 Online Methods

663 Definition of mosaic loss of the X chromosome (mLOX)

664 Detection of mLOX events from SNP array data in eight biobanks

665 All DNA samples were obtained from peripheral leukocytes and typed with single nucleotide
666 polymorphism (SNP) arrays. The median (SD) age at sample collection for genotyping ranged from
667 44 (16.3) for EBB to 67.2 (12.9) for FinnGen. The calling of mosaic loss of the X chromosome
668 (mLOX) was performed within each biobank using the Mosaic Chromosomal Alterations (MoChA)
669 pipeline (<https://github.com/freeseek/mochawdl>), with GRCh38 assembly as the reference genome
670 build. The raw genotyping array signal intensities of each variant were first transformed to B allele
671 frequency (BAF) (relative intensity of the B allele) and Log R Ratio (LRR) (total intensity of both
672 alleles). Then, haplotype phasing was performed using SHAPEIT4⁵⁴ across all batches of a biobank,
673 except for BBJ and BCAC for which phasing was done separately for each of the four sub-cohorts
674 (cohort sizes ranged from 3,888 to 45,877 for BBJ and from 42,360 to 62,889 for BCAC). Utilizing
675 long-range haplotype phasing can improve the sensitivity of detecting large mosaic events with low
676 cell fractions¹⁴, which is characteristic of mLOX. To avoid issues with phasing and the subsequent
677 mLOX calling, we excluded variants with poor genotyping quality such as segmental duplications
678 with low divergence (<2%) and single-nucleotide polymorphisms (SNPs) with high levels of
679 missingness (>3%) or heterozygote excess ($P < 1 \times 10^{-6}$). Finally, the calling of mLOX events was
680 performed within each batch based on the imbalance of phased BAF of heterozygous sites over the
681 whole X chromosome. To filter out 47,XXY and 47,XXX samples, we restricted to chromosome X
682 events with estimated ploidy less than 2.5, where the estimated ploidy is estimated by first computing
683 the median LRR across the assayed chromosome X SNPs and then by computing the value $2^{1+(LRR/LRR-$
684 $\text{hap2dip})}$ with LRR-hap2dip (the difference between LRR for haploid and diploid) set at 0.45 by default.
685 We further removed X loss events with length < 100 Mb to exclude other mosaic events (e.g., copy
686 number neutral loss of heterozygosity) on the X chromosome. For each mLOX event that passed
687 quality control, the fraction of cells (cf) with X loss was calculated as $4 * \text{bdev} / (1 + 2 * \text{bdev})$, where
688 bdev is the estimated BAF deviation of heterozygous sites. In addition to the dichotomous mLOX
689 status defined by the phased BAF method, for UKBB, the mean LRR (mLRR) of variants on X
690 chromosome non-pseudoautosomal (non-PAR) regions has also been used as a quantitative measure
691 of mLOX.

692 The 2022-01-14 version of MoChA was used to detect the dichotomous mLOX status for all
693 biobanks, except for BCAC (version: 2021-05-14) and BBJ (version: 2021-08-17 and 2021-09-07).
694 The priors of MoChA have been updated since 2021-05-14 to improve the detection of low cell
695 fraction mLOX calls, and thus, the biobanks that used the updated MoChA pipeline (all biobanks
696 except for BCAC) were expected to yield higher age-adjusted mLOX prevalences than biobanks that

697 used the previous version (only BCAC). A brief description of each contributed biobank (e.g.,
698 continental ancestry, sample size, age structures, and SNP array) is available in **Supplementary**
699 **Table S1**.

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

701 For UKBB, the whole-exome sequence (WES) data was released in late 2021⁵⁵, which permitted
702 identification of X loss from sequencing allelic dosage data in combination with array data. The
703 relative X chromosome dosage at the individual level was estimated following the steps described
704 previously⁵⁶. In brief, we first generated mean coverages from the original WES data for variants on
705 the autosomes and the X chromosome non-PAR regions, separately; then, we obtained the relative X
706 chromosome dosage by adjusting for the mean coverage of autosomes.

707 Comparison of different mLOX measures in UKBB

708 As mentioned above, for UKBB, three ways were used to define the mLOX phenotype, including the
709 dichotomous mLOX status derived from the phased BAF method (by MoChA) and two quantitative
710 measures employing either mLRR from SNP array data or allele dosage from WES data. To assess the
711 performances of the three mLOX measures in UKBB, we compared either mLRR or X dosage
712 between the case and the control groups defined by MoChA (**Figure S2A-C**). As shown in **Figure**
713 **S2B** and **S2C**, the participants identified as mLOX cases by MoChA exhibited lower mLRR (P from
714 the Analysis of Variance (ANOVA) test= 1.5×10^{-5}) and X dosage value ($P < 1.0 \times 10^{-250}$) than mLOX
715 controls. Then, for mLOX cases, we examined the relationships between three measures representing
716 the extent of mosaicism (**Figure S2D-F**), including cell fraction (from MoChA), mLRR, and X
717 dosage. Overall, significant correlations were observed across the three measures, with the absolute
718 Pearson correlation coefficient ranging from 0.42 between mLRR and X dosage to 0.86 between
719 mLOX cell fraction and X dosage. Again, given that mLRR is a noisier measure than X dosage, for
720 mLOX cell fraction, a stronger correlation was observed with X dosage ($r = -0.86$) than with mLRR (-
721 0.48).

722

723 **Environmental determinants and epidemiological consequences**

724 To investigate the effect of lifestyle factors on the odds of acquiring mLOX, we assessed the
725 associations between smoking and body mass index (BMI) with mLOX in the FinnGen cohort. In
726 FinnGen data freeze 9, 50.3% of female participants had smoking status (N=84,926) and 18.4% had
727 measurements for BMI (N=31,101) recorded at enrollment. We applied a logistic regression model
728 adjusting for age (at genotyping), age², and the first 10 PCs as covariates. As sensitivity analyses, we
729 restricted the analyses to expanded mLOX calls having cf > 5%. Given that we identified a significant
730 association between ever-smoking and expanded mLOX, we further adjusted for ever-smoking status

731 when assessing the effect of BMI on mLOX. To examine whether the environmental determinants
732 were shared or distinct between mLOX in women and mLOY in men, we also extended the
733 association analyses to mLOY (N=76,808 for smoking, N=33,668 for BMI).
734 To assess the clinical consequences of acquiring expanded mLOX, we performed a Cox proportional
735 hazards regression for incident cases in FinnGen, UKBB, MVP, and MGB independently, with the
736 time-on study as the time scale. For covariates, we recommended each biobank adjust for age, age²,
737 smoking, and the first 10 PCs. Meta-analysis across four biobanks was carried out with a fixed-effect
738 model applied in the meta package⁵⁷. For each disease, we applied Cochran's Q-test to assess
739 heterogeneity across biobanks with different healthcare systems. In total, we examined 1,253
740 phecodes covering 13 disease categories. Accordingly, the multiple-testing corrected P value
741 threshold was set to $P < 4.0 \times 10^{-5}$. In the main analysis, we used all detectable mLOX calls without
742 restriction for cell fraction. For a sensitivity analysis, we considered mLOX having cf >10% as
743 expanded calls, following the definition used by Zekavat et al¹⁶.

744

745 **Common and rare germline variants associated with detectable mLOX susceptibility**

746 GWAS of dichotomous mLOX status in eight contributed biobanks

747 To identify common germline variants (minor allele frequency (MAF) > 0.1%) associated with risk of
748 detectable mLOX in peripheral leukocytes, we performed a genome-wide association study (GWAS)
749 on chromosomes 1-22 and X in each of eight contributing biobanks independently, for a total of
750 904,524 women. For the dichotomous mLOX status (derived from MoChA), GWAS was conducted
751 for FinnGen and BCAC using the Scalable and Accurate Implementation of Generalized mixed model
752 (SAIGE)⁵⁸ and for the other six biobanks (including UKBB) using regenie⁵⁹ applied in the assoc.wdl
753 pipeline (part of the MoChA pipeline; <https://github.com/freeseek/mochawdl>). Both SAIGE and
754 regenie are feasible to account for sample relatedness and extreme case-control imbalances of a
755 dichotomous phenotype. For covariates, each biobank adjusted for age (at genotyping), age², and the
756 first 20 genetic principal components (PCs). The effective sample size, presented in Table 1, was
757 calculated as $(4 * N_{\text{case}} * N_{\text{control}}) / (N_{\text{case}} + N_{\text{control}})$.

758 GWAS of 3-way combined quantitative mLOX measure in UKBB

759 For UKBB, to improve the power of GWAS, we proposed a new quantitative measure by combining
760 the three methods of mLOX calling, that is, the mLOX combined call (3-way) = mLOX-status + 2*cf
761 - 2*mLRR - 4*(dosage-2) (cropped to the range [0,2]). The intuition behind this newly proposed
762 measure was to emphasize mLOX cases with larger cell fractions (similar to the strategy used by a
763 recent mosaic loss of the Y chromosome (mLOY) study⁶⁰) while obtaining enhanced mLOX calls
764 from integrating independent information of both SNP array and WES data. Compared to the

765 dichotomous mLOX status derived from MoChA, the t-test statistic for association with age was
766 increased by 29.2% when using the 3-way combined calls. As not all participants with SNP array data
767 had WES data available, we imputed the missing 3-way mLOX combined calls with 2-way combined
768 calls, defined as $mLOX\text{-status} + 3*cf - 3*mLRR$ (cropped to the range [0,2] as well). For the proposed
769 quantitative mLOX measure, GWAS was performed with the linear mixed model applied in BOLT-
770 LMM⁶¹.

771 Meta-analysis

772 For each contributed biobank, we filtered out variants with $MAF < 0.1\%$ or imputation INFO score $<$
773 0.6 . We also inspected allele frequencies of each biobank versus Genome Aggregation Database
774 (gnomAD) 3.0 as well as the relationship between standard errors and effective sample sizes across
775 biobanks, as applied by the covid-19 HGI meta-analysis⁶². Given that no biobank deviated from the
776 expected pattern, we conducted meta-analyses across biobanks. In addition to the dichotomous mLOX
777 measure used by all biobanks, UKBB was able to run GWAS with an additional quantitative measure
778 that combined information of three ways of mLOX calling and thus was expected to yield increased
779 power in GWAS. Depending on which mLOX measure was used in the UKBB GWAS, we applied
780 two fixed-effect meta-analysis models accordingly. When using the dichotomous measure, we applied
781 the inverse variance weighting (IVW) method which weighted the effect size estimated from an
782 individual biobank by its inverse variance. When UKBB used the 3-way combined measure as the
783 GWAS phenotype, we employed the weighted z-score method (weighted by the square root of the
784 effective sample size) applied in the METAL software⁶³ which can manage the different units of
785 dichotomous and quantitative measures. As the main analysis, we meta-analyzed summary statistics
786 across all eight biobanks regardless of ancestry and applied Cochran's Q-test to assess the
787 heterogeneity. To further investigate the impact of ancestry, we also conducted a meta-analysis for 7
788 biobanks containing only participants of European ancestry (without BBJ of East Asian ancestry).

789 Independent loci identification and gene prioritization

790 To identify independent signals and prioritize candidate causal genes, we applied the GWAS_{to}Genes
791 pipeline for variants presented in at least half of the contributed biobanks. In brief, primary
792 independent signals associated with mLOX susceptibility at a genome-wide significant level
793 ($P < 5 \times 10^{-8}$) were initially selected in 1Mb windows. Secondary independent signals were identified by
794 using an approximate conditional analysis applied in GCTA⁶⁴, with LD structures constructed from
795 UKBB samples. Secondary signals were only considered if they were genome-wide significant, in low
796 LD ($r^2 < 0.05$) with primary signals, and having association statistics unchanged with the conditional
797 analysis. We also excluded variants without any nearby genes (within 500 kb) documented in the
798 NCBI RefSeq dataset⁶⁵.

799 Candidate genes were prioritized using the following criteria and scored by their strength of evidence

800 for causality. First, signals were annotated with their physically closest genes. Second, signals and
801 their closely linked variants ($R^2 > 0.8$) were annotated if they were predicted deleterious coding
802 variants, or if the paired genes exhibited a gene-level association when collapsing all predicted
803 deleterious coding variants within a gene using Multi-marker Analysis of GenoMic Annotation
804 (MAGMA)⁶⁶. Third, non-coding signals and closely-linked variants were then annotated if they could
805 be mapped to known enhancers via the activity-by-contact (ABC) model⁶⁷. Fourth, colocalization
806 between GWAS and expression quantitative trait locus (eQTL) data was performed using the
807 summary data-based Mendelian randomization (SMR) and heterogeneity in dependent instruments
808 (HEIDI) test (version 0.68)⁶⁸ and the Approximate Bayes Factor (ABF) method applied in the “coloc”
809 package (version 5.1.0)⁶⁹. To identify tissues exhibiting a significant genome-wide enrichment, we
810 used LD score regression applied to specifically expressed gene (LDSC-SEG)⁷⁰ approach, with eQTL
811 datasets from cross-tissue meta-analyzed GTEx eQTL v.7⁷¹, eQTLGen⁷², and Brain-eMeta⁷³. The
812 same set of analyses were also applied to a protein quantitative trait locus (pQTL) dataset⁷⁴. Finally,
813 by integrating GWAS summary statistics with data from gene expression, biological pathway, and
814 predicted protein-protein interaction, candidate genes were identified using the gene-level Polygenic
815 Priority Score (PoPS) method⁷⁵.

816 Gene-burden test for rare variants causing detectable mLOX

817 To identify rare germline variants (minor allele frequency (MAF) < 0.1%) associated with the risk of
818 detectable mLOX, we performed gene-burden tests on chromosomes 1-22 and X in 226,125 UKBB
819 female participants with WES data available. We performed WES data pre-processing and quality
820 control following Gardner et al.⁷⁶. We annotated variants using the ENSEMBL Variant Effect
821 Predictor (VEP) v104⁷⁷ and defined protein-truncating variants (PTVs) as high-confidence (HC, as
822 defined by LOFTEE) stop gained, splice donor/acceptor, and frameshift consequences. We then
823 utilized CADDv1.6 to score a variant based on its predicted deleteriousness⁷⁸. Only non-synonymous
824 variants with MAF < 0.1% were included in the analysis. As the main analysis, we used BOLT-
825 LMM⁶¹ to perform the gene-burden test. For each gene, we defined individuals with HC PTVs,
826 missense variants with CADD scores ≥ 25 (MISS_CADD25), and damaging variants (HC_PTV +
827 MISS_CADD25) (DMG) as carriers. Then, carriers with non-synonymous variants were defined as
828 heterozygous and non-carriers as homozygous. For covariates, we adjusted for age, age², batches, sex,
829 and the first ten PCs. We further excluded the genes with less than 50 non-synonymous variant
830 carriers for each setting, resulting in 8,702 genes for HC_PTV, 15,144 for MISS_CADD25, and
831 16,493 for DMG, for a total of 40,339 genes. Accordingly, the Bonferroni corrected exome-wide
832 significant threshold was set to $0.05/40,339=1.24 \times 10^{-6}$. To avoid the identified association dominated
833 by a single variant, as sensitivity analysis, we conducted a leave-one-out analysis using a generalized
834 linear model for each significant gene. In addition, we reproduced the associations detected by BOLT-
835 LMM with STAAR⁴⁸.

836 Pathway and gene set analysis

837 To identify gene sets enriched in the same biological process, we performed pathway-based analysis
838 using the summary data-based adaptive rank truncated product (sARTP) method⁷⁹. We used summary
839 statistics from meta-analysis of seven biobanks of European ancestry (without BBJ) and LD structures
840 constructed from European ancestry samples of the 1000 Genomes project (1000 Genomes Project
841 Consortium, Nature, 2015). We considered a total of 6,285 gene sets available in GSEA
842 (<https://www.gseamsigdb.org/gsea/msigdb/>). Accordingly, the Bonferroni corrected P value was set to
843 $0.05/6,285=8.0\times 10^{-6}$.

844 Genetic correlation

845 To investigate whether there are traits that are genetically correlated with mLOX susceptibility, we
846 estimated genetic correlations between mLOX and 60 phenotypes (including both major diseases and
847 blood cell phenotypes) using LD score regression (LDSC)⁸⁰. For LDSC, we used HapMap3⁸¹ SNPs
848 and LD structures constructed from 1000 Genomes project⁸² samples of European ancestry.

849 Per-chromosome heritability

850 To examine whether the observed heritability for each chromosome was proportional to chromosome
851 length, we estimated per-chromosome heritability for 3-way combined mLOX measure in UKBB
852 using BOLT-REML⁸³. Given the large associations of HLA genes, we further examined how
853 heritability explained by chromosome 6 changed after excluding variants from the extended MHC
854 region (GRCh38: chr6:25.7-33.4 Mb).

855

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

857 Bayesian models to cluster variants by effects on mLOX and mLOY

858 We employed a Bayesian line model framework (<https://github.com/mjpirinen/linemodels>) to assign
859 each of the 49 independent common variants identified from mLOX GWAS and 147 variants (nine
860 variants dropped due to missing in mLOX GWAS) from the published mLOY GWAS¹³ into three
861 groups: specific to mLOX, specific to mLOY, and shared between mLOX and mLOY. The slopes of
862 the line models were set to 0 for the group of variants specific for mLOY and infinite for variants
863 specific for mLOX. For variants shared between mLOX and mLOY, the slope was set to 0.3, based on
864 the effects of four variants (rs568868093, rs381500, rs2280548, rs78378222) that were genome-wide
865 significant in both mLOX GWAS and mLOY GWAS. For all three line models, the prior SD
866 determining the magnitude of the effects was set to 0.15 and the correlation parameter determining the
867 allowed deviations from the lines to 0.995. The correlation between mLOX and mLOY GWAS
868 statistics was set to 0 given that there was no overlap between samples used in the two GWAS. We
869 assumed a uniform prior for the three models and obtained the posterior probabilities for each data

870 point separately within a Bayesian framework. Probability assignment threshold was set to 95%.

871 Fine-mapping of HLA alleles for mLOX and mLOY in FinnGen

872 Given the large associations with mLOX and the high polymorphism of HLA genes, we fine-mapped
873 HLA alleles at a unique protein sequence level in the FinnGen cohort. In FinnGen data freeze 9, a
874 total of 172 HLA alleles of 10 transplantation genes were imputed using a Finnish-specific reference
875 panel, as described in Ritari et al.⁸⁴. We conducted the association analysis between each imputed
876 HLA allele and the dichotomous mLOX status in 168,838 Finnish female participants (N of cases =
877 27,001) using a multivariate logistic regression model, considering age, age², and the first 10 PCs as
878 covariates. Only HLA alleles with more than 5 mLOX cases carrying the minor alleles were included
879 in the analysis. Ultimately, we considered 156 HLA alleles for mLOX, including 18 alleles for HLA-
880 A, 36 for HLA-B, 20 for HLA-C, 29 for HLA-DRB1, 14 for HLA-DQA1, 14 for HLA-DQB1, 18 for
881 HLA-DPB1, 3 for HLA-DRB3, and 2 each for HLA-DRB4, and DRB5. To identify independent HLA
882 alleles, a stepwise conditional analysis was performed with each step adding the most significant HLA
883 allele obtained from the previous step as an additional covariate, until no HLA allele can reach the
884 significant threshold. To examine whether the HLA associations are shared between mLOX and
885 mLOY, we extended the HLA fine-mapping analyses to mLOY in men (total N = 128,729, N of cases
886 = 45,675) for 157 HLA alleles (including HLA-A*02:02 compared to the 156 alleles used by mLOX
887 association analyses).

888

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

890 Allelic shift analysis

891 Conditional on mLOX having been detected, for each variant on the X chromosome we tested
892 whether there is a propensity for X chromosomes with a given allele to be identified as lost more
893 often than X chromosomes with the other allele. Similar to a transmission disequilibrium test⁵⁰, this
894 test is robust to the presence of population structure. Rather than measuring the over-transmission of
895 an allele from heterozygous parents to offspring, we measured the propensity of alleles to be on the
896 retained chromosome X homologue. Therefore, we carried out a binomial test for each variant with a
897 sample size equal to the number of women with detected mLOX who were heterozygous for that
898 variant, with no need to correct for covariates or relatedness.

899 Given the large number of X chromosome signals observed from the allelic shift analysis, we
900 inspected whether inflation may have contributed to the signals. We hypothesized that if the signals
901 were random, then the number of variants being significant can be related to the number of variants in
902 that region. Therefore, we checked the number of variants per 1kb region across the whole X
903 chromosome.

904 Identification of independent loci

905 Given the complexity of LD structures for X chromosomes especially for centromere and
906 pseudoautosomal (PAR) regions, we defined index variants by iteratively spanning the ± 500 kb
907 region around the most significant variant until no further variants reached a genome-wide significant
908 level ($P < 5 \times 10^{-8}$). Then, we calculated LD between every two index variants and kept the variant with
909 a lower P value if a pair of index variants with $r^2 < 0.1$.

910 Polygenic score to predict the retained X chromosome

911 To assess how well the allelic shift analysis polygenic score (PGS) can predict which X chromosome
912 is retained when mLOX occurs, we constructed PGSs in FinnGen mLOX cases (N=27,001). In brief,
913 we extracted the effect size for 44 independent loci from the allelic shift analysis of 6 biobanks
914 excluding FinnGen. Given that MoChA was able to detect which alleles were lost at heterozygous
915 sites, for each mLOX case, we computed the PGS for the retained X chromosome (PGS_{retained}) and the
916 lost X chromosome (PGS_{lost}) separately and obtained the difference in PGS between the two X
917 chromosomes (PGS_{diff}=PGS_{lost}-PGS_{retained}). A negative PGS_{diff} indicates that the retained X
918 chromosome of the mLOX case was correctly predicted. To assess the upper limit of prediction
919 performance, we performed simulation analysis in FinnGen mLOX cases based on the distribution of
920 allele frequencies of 44 lead variants in the Finnish population.

921

922 **Data availability**

923 Summary statistics generated from meta-analysis will be uploaded to GWAS Catalog after
924 publication. Individual level data can be requested directly from each contributing biobank.

925

926 **Code availability**

927 The Mosaic Chromosomal Alterations (MoChA) pipelines used for mosaic loss of the X chromosome
928 calling (mocha.wdl), GWAS (assoc.wdl), allelic shift analysis (impute.wdl and shift.wdl), and X
929 chromosome differential score estimation (score.wdl) are available at

930 <https://github.com/freeseeek/mochawdl>. The GWAS meta-analysis was performed by using the
931 pipeline developed by COVID-19 HGI, available at [https://github.com/covid19-](https://github.com/covid19-hgi/META_ANALYSIS)
932 [hg/META_ANALYSIS](https://github.com/covid19-hgi/META_ANALYSIS).

933

934 54. Delaneau, O., Zagury, J.F., Robinson, M.R., Marchini, J.L. and Dermitzakis, E.T., 2019.
935 Accurate, scalable and integrative haplotype estimation. Nature communications, 10(1), pp.1-
936 10.

- 937 55. Backman, J.D., Li, A.H., Marcketta, A., Sun, D., Mbatchou, J., Kessler, M.D., Benner, C.,
938 Liu, D., Locke, A.E., Balasubramanian, S. and Yadav, A., 2021. Exome sequencing and
939 analysis of 454,787 UK Biobank participants. *Nature*, 599(7886), pp.628-634.
- 940 56. Zhao, Y., Gardner, E.J., Tuke, M.A., Zhang, H., Pietzner, M., Koprulu, M., Jia, R.Y., Ruth,
941 K.S., Wood, A.R., Beaumont, R.N. and Tyrrell, J., 2022. Detection and characterization of
942 male sex chromosome abnormalities in the UK Biobank study. *Genetics in Medicine*.
- 943 57. Balduzzi, S., Rücker, G. and Schwarzer, G., 2019. How to perform a meta-analysis with R: a
944 practical tutorial. *Evidence-based mental health*, 22(4), pp.153-160.
- 945 58. Zhou, W., Nielsen, J.B., Fritsche, L.G., Dey, R., Gabrielsen, M.E., Wolford, B.N., LeFaive,
946 J., VandeHaar, P., Gagliano, S.A., Gifford, A. and Bastarache, L.A., 2018. Efficiently
947 controlling for case-control imbalance and sample relatedness in large-scale genetic
948 association studies. *Nature genetics*, 50(9), pp.1335-1341.
- 949 59. Mbatchou, J., Barnard, L., Backman, J., Marcketta, A., Kosmicki, J.A., Ziyatdinov, A.,
950 Benner, C., O'Dushlaine, C., Barber, M., Boutkov, B. and Habegger, L., 2021.
951 Computationally efficient whole-genome regression for quantitative and binary traits. *Nature*
952 *genetics*, 53(7), pp.1097-1103.
- 953 60. Zhao, Y., Stankovic, S., Koprulu, M., Wheeler, E., Day, F.R., Lango Allen, H., Kerrison,
954 N.D., Pietzner, M., Loh, P.R., Wareham, N.J. and Langenberg, C., 2021. GIGYF1 loss of
955 function is associated with clonal mosaicism and adverse metabolic health. *Nature*
956 *Communications*, 12(1), pp.1-6.
- 957 61. Loh, P.R., Tucker, G., Bulik-Sullivan, B.K., Vilhjalmsson, B.J., Finucane, H.K., Salem, R.M.,
958 Chasman, D.I., Ridker, P.M., Neale, B.M., Berger, B. and Patterson, N., 2015A. Efficient
959 Bayesian mixed-model analysis increases association power in large cohorts. *Nature genetics*,
960 47(3), pp.284-290.
- 961 62. COVID-19 Host Genetics Initiative. Mapping the human genetic architecture of COVID-19.
962 *Nature* 600, 472–477 (2021).
- 963 63. Willer, C.J., Li, Y. and Abecasis, G.R., 2010. METAL: fast and efficient meta-analysis of
964 genomewide association scans. *Bioinformatics*, 26(17), pp.2190-2191.
- 965 64. Yang, J., Ferreira, T., Morris, A.P., Medland, S.E., Madden, P.A., Heath, A.C., Martin, N.G.,
966 Montgomery, G.W., Weedon, M.N., Loos, R.J. and Frayling, T.M., 2012. Conditional and
967 joint multiple-SNP analysis of GWAS summary statistics identifies additional variants
968 influencing complex traits. *Nature genetics*, 44(4), pp.369-375.
- 969 65. O'Leary, N.A., Wright, M.W., Brister, J.R., Ciufo, S., Haddad, D., McVeigh, R., Rajput, B.,
970 Robbertse, B., Smith-White, B., Ako-Adjei, D. and Astashyn, A., 2016. Reference sequence
971 (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation.
972 *Nucleic acids research*, 44(D1), pp.D733-D745.
- 973 66. de Leeuw, C.A., Mooij, J.M., Heskes, T. and Posthuma, D., 2015. MAGMA: generalized

- 974 gene-set analysis of GWAS data. *PLoS computational biology*, 11(4), p.e1004219.
- 975 67. Nasser, J., Bergman, D.T., Fulco, C.P., Guckelberger, P., Doughty, B.R., Patwardhan, T.A.,
976 Jones, T.R., Nguyen, T.H., Ulirsch, J.C., Lekschas, F. and Mualim, K., 2021. Genome-wide
977 enhancer maps link risk variants to disease genes. *Nature*, 593(7858), pp.238-243.
- 978 68. Zhu, Z., Zhang, F., Hu, H., Bakshi, A., Robinson, M.R., Powell, J.E., Montgomery, G.W.,
979 Goddard, M.E., Wray, N.R., Visscher, P.M. and Yang, J., 2016. Integration of summary data
980 from GWAS and eQTL studies predicts complex trait gene targets. *Nature genetics*, 48(5),
981 pp.481-487.
- 982 69. Giambartolomei, C., Vukcevic, D., Schadt, E.E., Franke, L., Hingorani, A.D., Wallace, C. and
983 Plagnol, V., 2014. Bayesian test for colocalisation between pairs of genetic association
984 studies using summary statistics. *PLoS genetics*, 10(5), p.e1004383.
- 985 70. Finucane, H.K., Reshef, Y.A., Anttila, V., Slowikowski, K., Gusev, A., Byrnes, A., Gazal, S.,
986 Loh, P.R., Lareau, C., Shores, N. and Genovese, G., 2018. Heritability enrichment of
987 specifically expressed genes identifies disease-relevant tissues and cell types. *Nature genetics*,
988 50(4), pp.621-629.
- 989 71. GTEx Consortium, Ardlie, K.G., Deluca, D.S., Segrè, A.V., Sullivan, T.J., Young, T.R.,
990 Gelfand, E.T., Trowbridge, C.A., Maller, J.B., Tukiainen, T. and Lek, M., 2015. The
991 Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans.
992 *Science*, 348(6235), pp.648-660.
- 993 72. Võsa, U., Claringbould, A., Westra, H.J., Bonder, M.J., Deelen, P., Zeng, B., Kirsten, H.,
994 Saha, A., Kreuzhuber, R., Brugge, H. and Oelen, R., 2021. Large-scale cis-and trans-eQTL
995 analyses identify thousands of genetic loci and polygenic scores that regulate blood gene
996 expression. *Nature genetics*, 53(9), pp.1300-1310.
- 997 73. Qi, T., Wu, Y., Zeng, J., Zhang, F., Xue, A., Jiang, L., Zhu, Z., Kemper, K., Yengo, L.,
998 Zheng, Z. and Marioni, R.E., 2018. Identifying gene targets for brain-related traits using
999 transcriptomic and methylomic data from blood. *Nature communications*, 9(1), pp.1-12.
- 1000 74. Pietzner, M., Wheeler, E., Carrasco-Zanini, J., Cortes, A., Koprulu, M., Wörheide, M.A.,
1001 Oerton, E., Cook, J., Stewart, I.D., Kerrison, N.D. and Luan, J.A., 2021. Mapping the proteo-
1002 genomic convergence of human diseases. *Science*, 374(6569), p.eabj1541.
- 1003 75. Weeks, E.M., Ulirsch, J.C., Cheng, N.Y., Trippe, B.L., Fine, R.S., Miao, J., Patwardhan, T.A.,
1004 Kanai, M., Nasser, J., Fulco, C.P. and Tashman, K.C., 2020. Leveraging polygenic enrichments
1005 of gene features to predict genes underlying complex traits and diseases. *medRxiv*.
- 1006 76. Gardner, E.J., Kentistou, K.A., Stankovic, S., Lockhart, S., Wheeler, E., Day, F.R., Kerrison,
1007 N.D., Wareham, N.J., Langenberg, C., O’Rahilly, S., Ong, K.K. and Perry J.R.B., 2022.
1008 Damaging missense variants in IGF1R implicate a role for IGF-1 resistance in the aetiology
1009 of type 2 diabetes. *Cell Genomics*.
- 1010 77. McLaren, W., Gil, L., Hunt, S.E., Riat, H.S., Ritchie, G.R., Thormann, A., Flicek, P. and

- 1011 Cunningham, F., 2016. The ensembl variant effect predictor. *Genome biology*, 17(1), pp.1-14.
- 1012 78. Kircher, M., Witten, D.M., Jain, P., O'roak, B.J., Cooper, G.M. and Shendure, J., 2014. A
1013 general framework for estimating the relative pathogenicity of human genetic variants. *Nature*
1014 *genetics*, 46(3), pp.310-315.
- 1015 79. Zhang, H., Wheeler, W., Hyland, P.L., Yang, Y., Shi, J., Chatterjee, N. and Yu, K., 2016. A
1016 powerful procedure for pathway-based meta-analysis using summary statistics identifies 43
1017 pathways associated with type II diabetes in European populations. *PLoS genetics*, 12(6),
1018 p.e1006122.
- 1019 80. Bulik-Sullivan, B., Finucane, H.K., Anttila, V., Gusev, A., Day, F.R., Loh, P.R., Duncan, L.,
1020 Perry, J.R., Patterson, N., Robinson, E.B. and Daly, M.J., 2015. An atlas of genetic
1021 correlations across human diseases and traits. *Nature genetics*, 47(11), pp.1236-1241.
- 1022 81. International HapMap 3 Consortium, 2010. Integrating common and rare genetic variation in
1023 diverse human populations. *Nature*, 467(7311), p.52.
- 1024 82. 1000 Genomes Project Consortium, 2015. A global reference for human genetic variation.
1025 *Nature*, 526(7571), p.68.
- 1026 83. Loh, P.R., Bhatia, G., Gusev, A., Finucane, H.K., Bulik-Sullivan, B.K., Pollack, S.J., de
1027 Candia, T.R., Lee, S.H., Wray, N.R., Kendler, K.S. and O'Donovan, M.C., 2015B.
1028 Contrasting genetic architectures of schizophrenia and other complex diseases using fast
1029 variance-components analysis. *Nature genetics*, 47(12), pp.1385-1392.
- 1030 84. Ritari, J., Hyvärinen, K., Clancy, J., FinnGen, Partanen, J. and Koskela, S., 2020. Increasing
1031 accuracy of HLA imputation by a population-specific reference panel in a FinnGen biobank
1032 cohort. *NAR genomics and bioinformatics*, 2(2), p.lqaa030.

1033

1034 **Acknowledgments** We thank Juha Karjalainen (Institute for Molecular Medicine Finland (FIMM),
1035 Finland) and Mattia Cordioli (FIMM, Finland) for assistance in GWAS meta-analysis, Shea J.
1036 Andrews (Icahn School of Medicine at Mount Sinai, USA) and Jaakko Leinonen (FIMM, Finland) for
1037 kindly sharing formatted GWAS summary statistics used in genetic correlation analyses, Sakari
1038 Jukarainen (FIMM, Finland) and Alessio Gerussi (University of Milano-Bicocca, Italy) for insightful
1039 discussion on pheWAS analyses from a clinical standpoint, Samuel Jones (FIMM, Finland) and
1040 Masahiro Kanai (Broad Institute of MIT and Harvard, USA) for valuable feedback on HLA and fine-
1041 mapping, Jukka Koskela (FIMM, Finland) and Mikko Myllymäki (FIMM, Finland) for discussion on
1042 clonal hematopoiesis, Yu Fu (FIMM, Finland) and Annina Preussner (FIMM, Finland) for discussion
1043 on genetic analyses of sex chromosomes, and Geert Kops for discussion on mechanism causing
1044 chromosome missegregation. We thank Ms. Azusa Kouno in RIKEN Center for Integrative Medical
1045 Sciences and the members of the BioBank Japan Project, headquartered in the University of Tokyo
1046 Institute of Medical Science, for supporting this project. We want to acknowledge the participants and

1047 investigators of each contributing biobank. The FinnGen project is funded by two grants from
1048 Business Finland (HUS 4685/31/2016 and UH 4386/31/2016) and the following industry partners:
1049 AbbVie Inc., AstraZeneca UK Ltd, Biogen MA Inc., Bristol Myers Squibb (and Celgene Corporation
1050 & Celgene International II Sàrl), Genentech Inc., Merck Sharp & Dohme LCC, Pfizer Inc.,
1051 GlaxoSmithKline Intellectual Property Development Ltd., Sanofi US Services Inc., Maze
1052 Therapeutics Inc., Janssen Biotech Inc, Novartis AG, and Boehringer Ingelheim International GmbH.
1053 Following biobanks are acknowledged for delivering biobank samples to FinnGen: Auria Biobank
1054 (www.auria.fi/biopankki), THL Biobank (www.thl.fi/biobank), Helsinki Biobank
1055 (www.helsinginbiopankki.fi), Biobank Borealis of Northern Finland ([https://www.ppshep.fi/Tutkimus-
1056 ja-opetus/Biopankki/Pages/Biobank-Borealis-briefly-in-English.aspx](https://www.ppshep.fi/Tutkimus-ja-opetus/Biopankki/Pages/Biobank-Borealis-briefly-in-English.aspx)), Finnish Clinical Biobank
1057 Tampere (www.tays.fi/en-US/Research_and_development/Finnish_Clinical_Biobank_Tampere),
1058 Biobank of Eastern Finland (www.ita-suomenbiopankki.fi/en), Central Finland Biobank
1059 (www.ksshp.fi/fi-FI/Potilaille/Biopankki), Finnish Red Cross Blood Service Biobank
1060 (www.veripalvelu.fi/verenluovutus/biopankkitoiminta), Terveystalo Biobank
1061 (www.terveystalo.com/fi/Yritystietoa/Terveystalo-Biopankki/Biopankki/) and Arctic Biobank
1062 ([https://www.oulu.fi/en/university/faculties-and-units/faculty-medicine/northern-finland-birth-
1063 cohorts-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
1064 (www.bbmri.fi). Finnish Biobank Cooperative -FINBB (<https://finbb.fi/>) is the coordinator of
1065 BBMRI-ERIC operations in Finland. The Finnish biobank data can be accessed through the
1066 Fingenious® services (<https://site.fingenious.fi/en/>) managed by FINBB. For BCAC and MVP, the
1067 detailed acknowledgement is available in Supplementary materials.

1068

1069 **Author contributions** This project is initialized and led by A.L., G.G., P.-R.L., A.G., J.R.B.P., and
1070 M.M.. A.L. and M.M. wrote the first draft of the manuscript. A.L. coordinated the analyses of each
1071 contributing biobank, performed FinnGen specific analyses, conducted meta-analysis (including
1072 GWAS, allelic shift analysis, and pheWAS) and post-GWAS analyses, generated the figures and
1073 tables, and wrote the manuscript. G.G. developed the MoChA pipelines for mLOX calling, GWAS,
1074 allelic shift analysis, and X chromosome differential score estimation, guided the analyses of each
1075 contributing biobank, performed mLOX calling, GWAS, and allelic shift analysis for UKBB and
1076 MGB, and wrote the manuscript. Y.Z. performed WES analyses and 3-way combined call GWAS in
1077 UKBB, generated Supplementary Figure S2 and S6, prepared Supplementary Table S17, and wrote
1078 the relevant result and method paragraphs. M.P developed the Bayesian line model to cluster mLOX
1079 and mLOY loci and wrote the relevant method paragraph. M.M.Z. performed pheWAS for UKBB,
1080 MGB, and MVP and GWAS for MGB. K.K. performed the GWAS to gene pipeline, prepared
1081 Supplementary Table S11, and wrote the relevant method paragraphs. Z.Y. estimated heritability and
1082 genetic correlations and prepared Supplementary Table S14. K.Y. and L.S. performed the pathway

1083 analysis and prepared Supplementary Table S12. C.V. performed the sensitivity analyses for
1084 associations with leukemia in UKBB and prepared Supplementary Table S7. X.L. performed mLOX
1085 calling, GWAS, allelic shift analysis, and HLA fine-mapping replication analysis in BBJ. D.W.B.
1086 performed GWAS for PLCO and generated inputs for blood cell trait heat-map (Figure 3D and 4B).
1087 G.H. performed mLOX calling, GWAS, and allelic shift analysis for EBB. B.G. and S.P. performed
1088 mLOX calling, GWAS, and allelic shift analysis for MVP. J.D performed mLOX calling and GWAS
1089 for BCAC. W.Z. performed mLOX calling, GWAS, and allelic shift analysis for PLCO. Y.M.
1090 participated in BBJ analyses. V.T. and F.-D.P participated in EBB analyses. M.A., T.P.S, and A.G.
1091 participated in FinnGen analyses. W.-Y.H. and N.F. participated in PLCO analyses. E.J.G. participated
1092 in UKBB WES analyses. V.G.S. assisted in interpreting findings related to clonal hematopoiesis.
1093 A.P. coordinated the FinnGen project. H.M.O advised the HLA fine-mapping analysis and assisted in
1094 interpreting findings related to HLA. T.T. assisted in interpreting findings related to skewed X-
1095 chromosome inactivation and escaping from X-chromosome inactivation. S.J.C. coordinated the
1096 PLCO project. R.M. supervised EBB analyses. P.N. supervised pheWAS for UKBB, MGB, and
1097 MVP. M.J.D. initialized/conceptualized the mosaic chromosomal alteration project in FinnGen and
1098 assisted in interpreting findings especially those related to mLOY in men. A.B. supervised pheWAS
1099 in UKBB, MGB, and MVP and the sensitivity analyses for associations with leukemia in UKBB.
1100 S.A.M. supervised the development of MoChA pipelines. C.T. supervised BBJ analyses and advised
1101 the HLA fine-mapping analysis. P.-R.L., A.G., J.R.B.P, and M.M. co-supervised the project,
1102 interpreted the findings, and wrote the manuscript. For FinnGen, BCAC, and MVP, detailed author
1103 lists are available in supplementary materials. All authors reviewed the manuscript.

1104

1105 **Funding** This work was supported by the Intramural Research Program of the National Cancer
1106 Institute, National Institutes of Health, and the Medical Research Council (unit programs:
1107 MC_UU_12015/2, MC_UU_00006/2). G.G. was supported by NIH grants R01 MH104964 and R01
1108 MH123451. A.G. was supported by the Academy of Finland (grant no. 323116) and by the European
1109 Research Council under the European Union's Horizon 2020 Research and Innovation Programme
1110 (grant no. 945733). P.-R.L. was supported by NIH grant DP2 ES030554, a Burroughs Wellcome Fund
1111 Career Award at the Scientific Interfaces, the Next Generation Fund at the Broad Institute of MIT and
1112 Harvard, and a Sloan Research Fellowship. C.T. was supported by Japan Agency for Medical
1113 Research and Development (AMED) grants JP21kk0305013, JP21tm0424220, and JP21ck0106642,
1114 and Japan Society for the Promotion of Science (JSPS) KAKENHI grant JP20H00462.

1115

1116 **Competing interests** G.G., P.-R.L., and S.A.M. declare competing interests: patent application
1117 PCT/WO2019/079493 has been filed on the mosaic chromosomal alterations detection method used

1118 in this work. J.R.B.P and E.J.G are employee of and hold shares in Adrestia Therapeutics. A.B.
1119 reports scientific advisory board membership for TenSixteen Bio. P.N. reports grant support from
1120 Apple, Amgen, Boston Scientific, AstraZeneca, and Novartis, personal fees from Apple, AstraZeneca,
1121 Blackstone Life Sciences, Foresite Labs, Genentech/Roche, Allelica, Novartius, scientific advisory
1122 board membership for geneXwell, Esperion Therapeutics, and TenSixteen Bio, is a scientific co-
1123 founder of TenSixteen Bio, and spousal employment at Vertex, all unrelated to the present study.

1124

1125 **Ethics statement** Patients and control subjects in FinnGen provided informed consent for biobank
1126 research, based on the Finnish Biobank Act. Alternatively, separate research cohorts, collected prior
1127 the Finnish Biobank Act came into effect (in September 2013) and start of FinnGen (August 2017),
1128 were collected based on study-specific consents and later transferred to the Finnish biobanks after
1129 approval by Fimea (Finnish Medicines Agency), the National Supervisory Authority for Welfare and
1130 Health. Recruitment protocols followed the biobank protocols approved by Fimea. The Coordinating
1131 Ethics Committee of the Hospital District of Helsinki and Uusimaa (HUS) statement number for the
1132 FinnGen study is Nr HUS/990/2017. The FinnGen study is approved by Finnish Institute for Health
1133 and Welfare (permit numbers: THL/2031/6.02.00/2017, THL/1101/5.05.00/2017,
1134 THL/341/6.02.00/2018, THL/2222/6.02.00/2018, THL/283/6.02.00/2019, THL/1721/5.05.00/2019
1135 and THL/1524/5.05.00/2020), Digital and population data service agency (permit numbers:
1136 VRK43431/2017-3, VRK/6909/2018-3, VRK/4415/2019-3), the Social Insurance Institution (permit
1137 numbers: KELA 58/522/2017, KELA 131/522/2018, KELA 70/522/2019, KELA 98/522/2019, KELA
1138 134/522/2019, KELA 138/522/2019, KELA 2/522/2020, KELA 16/522/2020), Findata permit
1139 numbers THL/2364/14.02/2020, THL/4055/14.06.00/2020,,THL/3433/14.06.00/2020,
1140 THL/4432/14.06/2020, THL/5189/14.06/2020, THL/5894/14.06.00/2020, THL/6619/14.06.00/2020,
1141 THL/209/14.06.00/2021, THL/688/14.06.00/2021, THL/1284/14.06.00/2021,
1142 THL/1965/14.06.00/2021, THL/5546/14.02.00/2020, THL/2658/14.06.00/2021,
1143 THL/4235/14.06.00/202, Statistics Finland (permit numbers: TK-53-1041-17 and
1144 TK/143/07.03.00/2020 (earlier TK-53-90-20) TK/1735/07.03.00/2021, TK/3112/07.03.00/2021) and
1145 Finnish Registry for Kidney Diseases permission/extract from the meeting minutes on 4th July 2019.
1146 The Biobank Access Decisions for FinnGen samples and data utilized in FinnGen Data Freeze 9
1147 include: THL Biobank BB2017_55, BB2017_111, BB2018_19, BB_2018_34, BB_2018_67,
1148 BB2018_71, BB2019_7, BB2019_8, BB2019_26, BB2020_1, Finnish Red Cross Blood Service
1149 Biobank 7.12.2017, Helsinki Biobank HUS/359/2017, HUS/248/2020, Auria Biobank AB17-5154
1150 and amendment #1 (August 17 2020), AB20-5926 and amendment #1 (April 23 2020) and it's
1151 modification (Sep 22 2021), Biobank Borealis of Northern Finland_2017_1013, Biobank of Eastern
1152 Finland 1186/2018 and amendment 22 § /2020, Finnish Clinical Biobank Tampere MH0004 and
1153 amendments (21.02.2020 & 06.10.2020), Central Finland Biobank 1-2017, and Terveystalo Biobank

1154 STB 2018001 and amendment 25th Aug 2020. The UKBB analyses were conducted using applications
1155 7089, 9905, and 21552.