

Multi-ancestry genome-wide association meta-analysis of mosaic loss of chromosome Y in the Million Veteran Program identifies 167 novel loci

Michael Francis^{1,2†}, Bryan R. Gorman^{1,2†}, Tim B. Bigdeli^{3,4†}, Giulio Genovese^{5,6,7†}, Georgios Voloudakis^{8,9}, Jaroslav Bendl⁹, Biao Zeng⁹, Sanan Venkatesh^{8,9}, Chris Chatzinakos⁴, Erin McAuley^{1,2}, Sun-Gou Ji^{1,33}, Kyriacos Markianos¹, Patrick A. Schreiner^{1,2}, Elizabeth Partan¹⁰, Yunling Shi¹¹, Poornima Devineni¹¹, VA Million Veteran Program*, Jennifer Moser¹², Sumitra Muralidhar¹², Rachel Ramoni¹², Alexander G. Bick¹³, Pradeep Natarajan^{5,14,15}, Themistocles L. Assimes^{16,17}, Philip S. Tsao^{16,17,18}, Derek Klarin^{19,20}, Catherine Tcheandjieu^{21,22,23,24}, Neal S. Peachey^{25,26,27}, Sudha K. Iyengar^{25,28,29,30,31}, Panos Roussos^{8,9}, Saiju Pyarajan^{1,32†}

¹Center for Data and Computational Sciences (C-DACS), VA Cooperative Studies Program, VA Boston Healthcare System, Boston, MA, 02130, USA, ²Booz Allen Hamilton, McLean, VA, 22102, USA, ³Research Service, VA New York Harbor Healthcare System, Brooklyn, NY, 11209, USA, ⁴Department of Psychiatry and Behavioral Sciences, SUNY Downstate Health Sciences University, Brooklyn, NY, 11203, USA, ⁵Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, MA, 02142, USA, ⁶Stanley Center, Broad Institute of MIT and Harvard, Cambridge, MA, 02142, USA, ⁷Department of Genetics, Harvard Medical School, Boston, MA, 02115, USA, ⁸Center for Precision Medicine and Translational Therapeutics; VISN 2 Mental Illness Research, Education, and Clinical Center (MIRECC), James J Peters Veteran Affairs Medical Center, New York/New Jersey VA Health Care Network, Bronx, NY, 10468, USA, ⁹Center for Disease Neurogenomics; Department of Psychiatry; Friedman Brain Institute; Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, NY, 10029, USA, ¹⁰Center for Data and Computational Sciences (C-DACS), VA Cooperative Studies Program, VA Boston Healthcare System, Boston, MA, USA, ¹¹Center for Data and Computational Sciences (C-DACS), VA Boston Healthcare System, Boston, MA, USA, ¹²Office of Research and Development, Department of Veterans Affairs, Washington, DC, 20420, USA, ¹³Division of Genetic Medicine, Vanderbilt University Medical Center, Nashville, TN, 37232, USA, ¹⁴Department of Medicine, Cardiology Division, Massachusetts General Hospital, Boston, MA, 02114, USA, ¹⁵Cardiovascular Research Center, Massachusetts General Hospital, Boston, MA, 02114, USA, ¹⁶VA Palo Alto Epidemiology Research and Information Center for Genomics, VA Palo Alto Health Care System, Palo Alto, CA, 94304, USA, ¹⁷Department of Medicine, Stanford University School of Medicine, Stanford, CA, 94305, USA, ¹⁸Stanford Cardiovascular Institute, Stanford University School of Medicine, Stanford, CA, 94305, USA, ¹⁹VA Palo Alto Health Care System, Palo Alto, CA, USA, ²⁰Department of Surgery, Stanford University School of Medicine, Stanford, CA, USA, ²¹VA Palo Alto Health Care System, ²²Department of Medicine, Division of Cardiovascular Medicine, Stanford University School of Medicine, Stanford, CA, USA, ²³Gladstone Institute of Data Science and Biotechnology, Gladstone Institutes, San Francisco, CA, USA, ²⁴Department of Epidemiology and Biostatistics, University of California San Francisco, San Francisco, CA, USA, ²⁵Research Service, VA Northeast Ohio Healthcare System, Cleveland, OH, 44106, USA, ²⁶Cole Eye Institute, Cleveland Clinic, Cleveland, OH, 44195, USA, ²⁷Department of Ophthalmology, Cleveland Clinic Lerner College of Medicine of Case Western Reserve University, Cleveland, OH, 44195, USA, ²⁸Department of Population and Quantitative Health Sciences, Case Western Reserve University, Cleveland, OH, 44106,

46 USA, ²⁹Department of Ophthalmology and Visual Sciences, University Hospitals Eye Institute,
47 Cleveland, OH, 44106, USA, ³⁰Department of Genetics & Genome Sciences, Case Western
48 Reserve University, Cleveland, OH, 44106, USA, ³¹Cleveland Institute for Computational
49 Biology, Case Western Reserve University, Cleveland, OH, 44106, USA, ³²Department of
50 Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, 02115, USA,
51 ³³Present address: BridgeBio Pharma, Palo Alto, CA

52 † Denotes equal contribution

53 ‡ Saiju Pyarajan: saiju.pyarajan@va.gov

54 *Lists of authors and their affiliations appear at the end of the paper.

55

56

57 Abstract

58 Mosaic loss of chromosome Y (mLOY) is a common somatic mutation in leukocytes of
59 older males. mLOY was detected in 126,108 participants of the Million Veteran
60 Program: 106,054 European (EUR), 13,927 admixed African (AFR), and 6,127
61 Hispanic. In multi-ancestry genome-wide association analysis, we identified 323
62 genome-wide significant loci, 167 of which were novel—more than doubling the number
63 of known mLOY loci. Tract-based ancestry deconvolution resolved local inflation at AFR
64 lead SNPs. Transcriptome-wide associations yielded 2,297 significant genes, including
65 seven additional novel genes; integrative eQTL analyses highlighted 51 genes that
66 causally influence mLOY via differential expression. Thirty-two significant traits found in
67 a phenome-wide polygenic score scan were used in Mendelian randomization (MR).
68 MR implicated six traits as causal influences on mLOY: triglycerides, high-density
69 lipoprotein, smoking, body mass index, testosterone, and sex hormone-binding globulin;
70 and found influence of mLOY on plateletcrit, prostate cancer, lymphocyte percentage,
71 and neutrophil percentage. These results mark a major step forward in our
72 understanding of the genetic architecture of mLOY and its associated risks.

73

74 **Introduction**

75 Mosaic loss of Y chromosome (mLOY) is the most common type of mosaic
76 chromosomal alteration (mCA), observable in upwards of 40% of males above age
77 70^{1,2}, and 70% of males over 85³. mLOY is the most readily detected mCA in
78 leukocytes, which are the primary source of blood-derived DNA. Adaptive immunity
79 causes high turnover rates in the hematopoietic stem cell compartment, enabling clonal
80 expansion of mosaic cell subpopulations with selective advantages⁴. Chromosomal
81 aberrations in these rapidly expanding cell subpopulations can produce mLOY, though it
82 is unclear if mLOY itself confers selective advantages.

83 The gene-poor and repetitive-element-rich composition of the Y chromosome
84 initially led researchers to believe its role was restricted to spermatogenesis and sex
85 determination⁵. Congruently, mLOY was considered a benign condition, and a
86 consequence of the broader genomic instability that occurs with aging⁶. But in the past
87 decade, epidemiological studies have highlighted associations between mLOY and a
88 broad range of health outcomes; these include all-cause mortality, hematological
89 malignancies and other types of cancer, Alzheimer's disease, type 2 diabetes (T2D),
90 obesity, and cardiovascular disease (CVD)⁷. However, it has yet to be determined
91 whether mLOY is a driving causal factor in these conditions, a passenger (i.e. a
92 consequence), or a symptom of a shared, underlying cause, such as genetic
93 susceptibility to DNA replication errors, or exogenous exposure to mutagens
94 (particularly via smoking cigarettes)^{7,8}. There have also been inconsistent associations
95 with the comorbidities observed across studies, although this may be related to
96 differences in sample collection and mLOY classification methods (e.g. low versus high
97 cell fraction detection)².

98 The mechanisms of mLOY in producing disease phenotypes are gradually being
99 elucidated. Many genetic risk loci for clonal hematopoiesis of indeterminate potential
100 (CHIP)^{9,10} have also been identified as mLOY risk loci¹¹, and these two types of
101 mosaicism can co-occur (even in men without observable hematological disease¹²).
102 However, the cellular and epidemiological outcomes of CHIP and mLOY appear to be
103 distinct. For example, a study which induced *TET2*-associated CHIP in mice suggested
104 a mechanistic link to CVD based on expression of inflammatory chemokines in
105 macrophages¹³, while a mouse model of mLOY demonstrated a mechanism of
106 producing CVD through fibrotic deposition in the extracellular matrix¹⁴. Additionally,
107 deletion of chromosome Y by CRISPR–Cas9 produced more aggressive bladder cancer
108 tumors by means of T-cell exhaustion¹⁵. It has also been suggested that there are bi-
109 directional relationships between mLOY and transcriptional dysregulation that lead to
110 differences in disease phenotypes that are dependent on mLOY cell lineage⁸.

111 Estimates of single nucleotide polymorphism-based heritability (SNP- h^2) for
112 mLOY detected via the pseudo-autosomal region 1 (PAR1) are as high as 31.7%,
113 highlighting mLOY as substantially heritable compared to most human traits^{1,16}.
114 Germline genetics govern many processes which can lead to somatic mCA acquisition
115 and clonal expansion, particularly via cell-cycle regulation, DNA damage response,
116 apoptosis, and susceptibility to cancer. Advances in integrating long-range phasing with
117 genotype have enabled sensitive and accurate identification of mLOY in large-scale
118 cohorts¹. A genome-wide association study (GWAS) of mLOY status, performed in
119 European ancestry (EUR) UK Biobank (UKB) participants¹, replicated all 19 previously
120 reported genetic risk loci¹⁷, including the oncogene *TCL1A*¹⁸, and identified 137 novel

121 loci. A GWAS of mLOY in Biobank Japan (BBJ) using mLRR-Y intensity as a
122 quantitative trait measure identified 46 loci, 35 of which were novel¹⁹.

123 In this study we analyzed 544,112 male participants in the Million Veteran
124 Program (MVP), a biobank in the Department of Veterans Affairs (VA) healthcare
125 system which combines ancestrally diverse genetic data with extensive electronic health
126 records²⁰. In addition to a large EUR cohort, we present the first GWAS of mLOY status
127 in African (AFR) and Hispanic/Latino (HIS) ancestries. MVP provides a uniquely
128 valuable resource to perform multi-ancestry mLOY analyses, as it avoids technical
129 issues related to genotyping and mosaicism calling that may be introduced by
130 combining data from separate biobanks. Our results highlight the benefits of inclusive
131 population studies in advancing our understanding of mLOY.

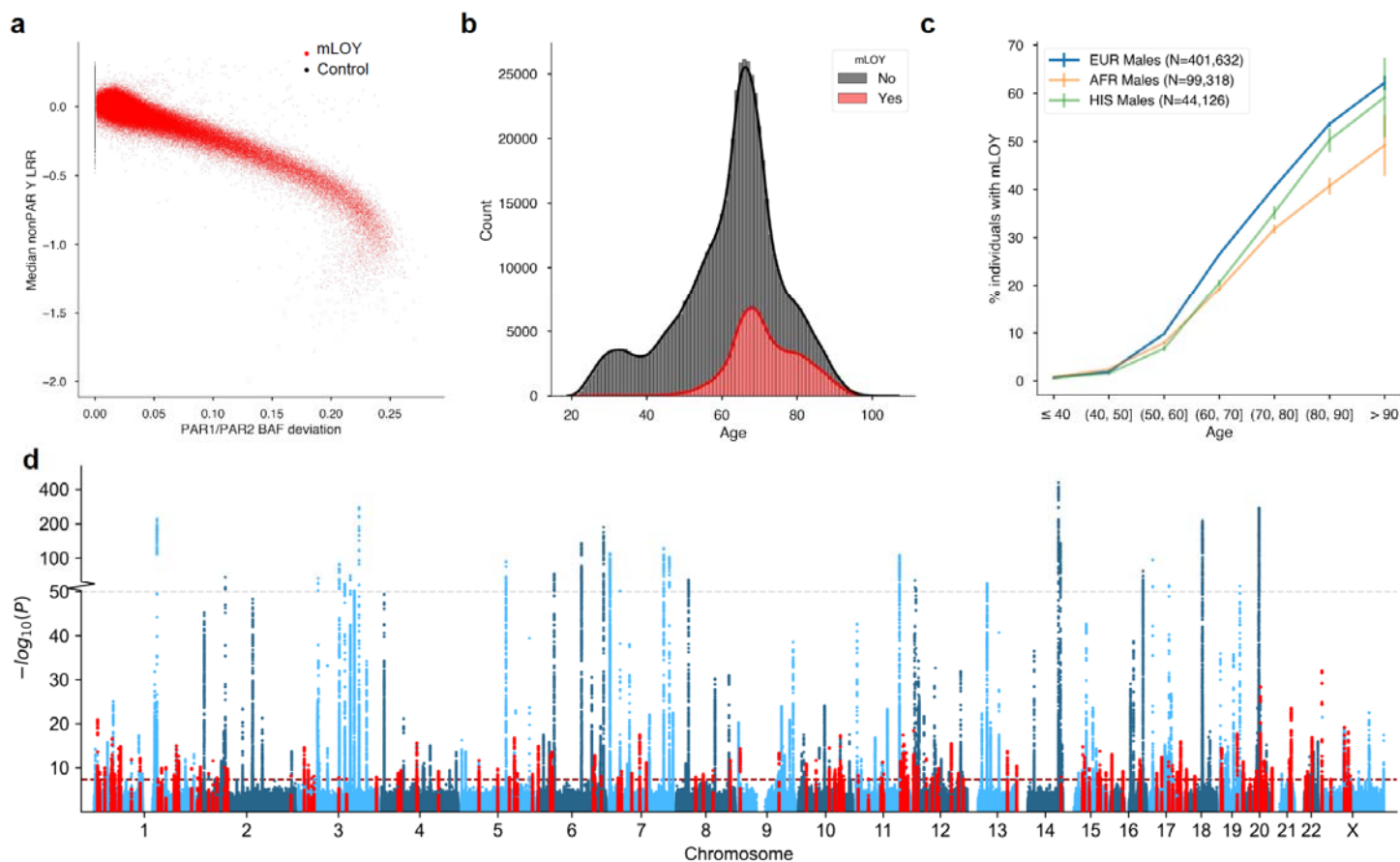
132

133

134 **Results**

135 *mLOY phenotyping and participant characteristics*

136 In this study we identified mLOY in MVP participants and performed GWAS with
137 subsequent functional analyses to place our findings in biological context
138 (Supplementary Fig. 1). We used a case-control design to classify mLOY as any
139 detectable mosaicism, using allelic ratio genotyping intensities in PAR1 and PAR2
140 shared by the X and Y chromosomes (Fig. 1a), similar to a previous GWAS in UKB¹. In
141 MVP, 106,054 of 400,970 (26.4%) EUR men (median age 66) showed evidence of
142 mLOY. A lower prevalence of mLOY was observed in AFR (13,927 of 99,103; 14.1%;
143 median age 60) and HIS (6,127 of 44,039; 13.9%; median age 60) (Supplementary
144 Data 1). The prevalence of mLOY increased with age, from 10% among participants
145 aged 50-60 to upwards of 50% in octogenarians (Fig. 1b-c). mLOY cell fraction
146 percentage increased with age across all ancestries (Supplementary Fig. 2a). Lifetime
147 smoking was associated with a higher odds ratio (OR) of mLOY (adjusted for age and
148 age-squared) at 1.33 [95% confidence interval (CI)=1.30,1.35] in EUR; this finding was
149 consistent across AFR and HIS (Supplementary Data 1), and with previous reports from
150 UKB¹. Current and former smoking status were also associated with higher mLOY cell
151 fraction percentages (Supplementary Fig. 2b).



152

153 **Fig. 1. mLOY in the Million Veteran Program.** **a**, Median genotyping probe intensity log R ratio (LRR) vs. phased B
 154 Allele Frequency (BAF) in the pseudo-autosomal regions (PAR) 1 and 2. **b**, Percentage of individuals with mLOY per ten-
 155 year age bin for MVP European (EUR), African (AFR), and Hispanic (HIS) cohorts. Error bars represent 95% confidence
 156 intervals. **c**, Density of age distribution in all MVP mLOY cases and controls. **d**, Manhattan plots show the $-\log_{10}(P)$ for
 157 associations of genetic variants with mLOY in the multi-ancestry meta-analysis. Novel mLOY index variants and variants
 158 within ± 50 Kb are highlighted in red. The red line indicates the genome-wide significance threshold ($P < 5 \times 10^{-8}$). The grey
 159 dotted line represents a transition from linear to log-scale on the y-axis.

160 *Genome-wide significant associations*

161 We performed a GWAS for mLOY case-control status in each ancestry group
162 (Supplementary Data 2). We identified 336 conditionally independent genome-wide
163 significant (GWS; $P < 5 \times 10^{-8}$) signals in 203 distinct loci for EUR, 50 signals in 46 loci for
164 AFR, and 17 signals in 15 loci for HIS (Supplementary Fig. 3a-c, Supplementary Data 3-
165 5). Of the EUR signals, 220 were within 1Mb of a previously reported mLOY index
166 variant^{17,19}, including 148 of the 156 previously identified in UKB¹, and 116 were novel.
167 In a variant-level replication of 327 EUR signals that were available in UKB¹, all but one
168 (rs925301) had the same effect direction (Supplementary Data 3, Supplementary Fig.
169 4a). Of the 116 novel MVP signals, 17 had $P < 1 \times 10^{-5}$ and 97 had $P < 0.05$ in UKB¹.

170 The most significant association signals in MVP EUR and HIS were in *TCL1A*; in
171 EUR, we identified the same *TCL1A* lead variant as in previous GWAS^{1,18} (rs2887399;
172 OR=0.708 [0.697, 0.719]; $P = 3.18 \times 10^{-419}$; Supplementary Fig. 5a). The strongest effect
173 of an allele in EUR was at *TP53*, an oncogene associated with mCAs and CHIP
174 (rs78378222; OR=1.664 [1.584, 1.749]; $P = 1.44 \times 10^{-90}$). In AFR, the most significant
175 association was in *RPN1* at rs113336380, which has a minor allele frequency (MAF) of
176 ~6% in AFR and <0.01% in EUR (Supplementary Fig. 5b). Effect directions were largely
177 concordant between EUR and AFR associations ($r = 0.656$, $P = 1.13 \times 10^{-46}$), with the
178 exception of a group of novel EUR variants that were not significant in AFR, and
179 rs6018599 (*GGT5/CABIN1*), which was GWS only in AFR (Supplementary Fig. 4b).
180 Four additional GWS novel signals were specific to AFR: *MPL*, *NKX2-3*, *ETV6*, and
181 *BLCAP* (Supplementary Fig. 6; Supplementary Data 4). All 17 HIS signals were GWS in
182 EUR (Supplementary Data 5); 15 of these were GWS in AFR, and all HIS signals were

183 GWS in UKB¹. Effect direction at significant HIS signals between HIS and EUR were
184 consistent (Supplementary Fig. 4c). We improved our variant selection by fine-mapping
185 and estimating credible sets of candidate causal variants in EUR and AFR. In EUR, we
186 found 11,242 variants in 334 high-quality credible sets, with a median of 8.5 variants per
187 credible set (Supplementary Data 6). In AFR, we found 533 variants in 45 high-quality
188 credible sets, with a median of 5 variants per credible set (Supplementary Data 7).

189 We extended our association analyses for all ancestries to a genotype panel
190 enriched in protein-altering rare variants ($MAF < 0.001$)²¹, and identified four novel GWS
191 variants in EUR (Supplementary Data 8), including a frameshift mutation at
192 *DCXR*:c.583del (p.His195fs), and somatic missense mutations in *DNMT3A* (R882H),
193 *JAK2* (V617F), and *IDH2* (R140Q), which are known to be associated with hematologic
194 malignancies such as CHIP and acute myeloid leukemia (AML)^{22,23}. The estimated
195 effects of these rare variants were negative and, consistent with expectation, stronger
196 than those of common variants (Supplementary Fig. 7).

197 Fixed effects (FE) multi-ancestry meta-analysis of the three MVP ancestry
198 groups identified 298 GWS loci, including 157 novel loci, 42 of which did not reach GWS
199 in any individual ancestry (Fig. 1d; Supplementary Fig. 3d; Supplementary Data 9). After
200 meta-analysis there remained 8 EUR loci and 2 AFR loci that were GWS only within
201 their respective ancestries. Of the 42 added meta-analysis novel lead variants, 32 had
202 $P < 0.05$ in UKB¹ and two had $P < 1 \times 10^{-5}$. For all 298 meta-analysis lead variants, 211
203 were available in the BBJ mLOY GWAS¹⁹; though BBJ used a less sensitive
204 quantitative mLOY measure (logarithm of R ratio) as opposed to our case-control
205 designation, the effect of 188 of the 211 variants were aligned, including all 88 BBJ

206 variants with $P < 0.01$ (Supplementary Fig. 4d). BBJ¹⁹ shared 28 GWS variants with our
207 meta-analysis (Supplementary Data 9). Within MVP meta-analysis index variants,
208 rs2887399 (*TCL1A*) remained the most significant association ($P = 5.25 \times 10^{-459}$). We also
209 performed random effects (RE) meta-analyses using the Han-Eskin method (RE2)²⁴,
210 and observed similar P-values to FE (Supplementary Data 9).

211 *Tract-based association analysis for AFR*

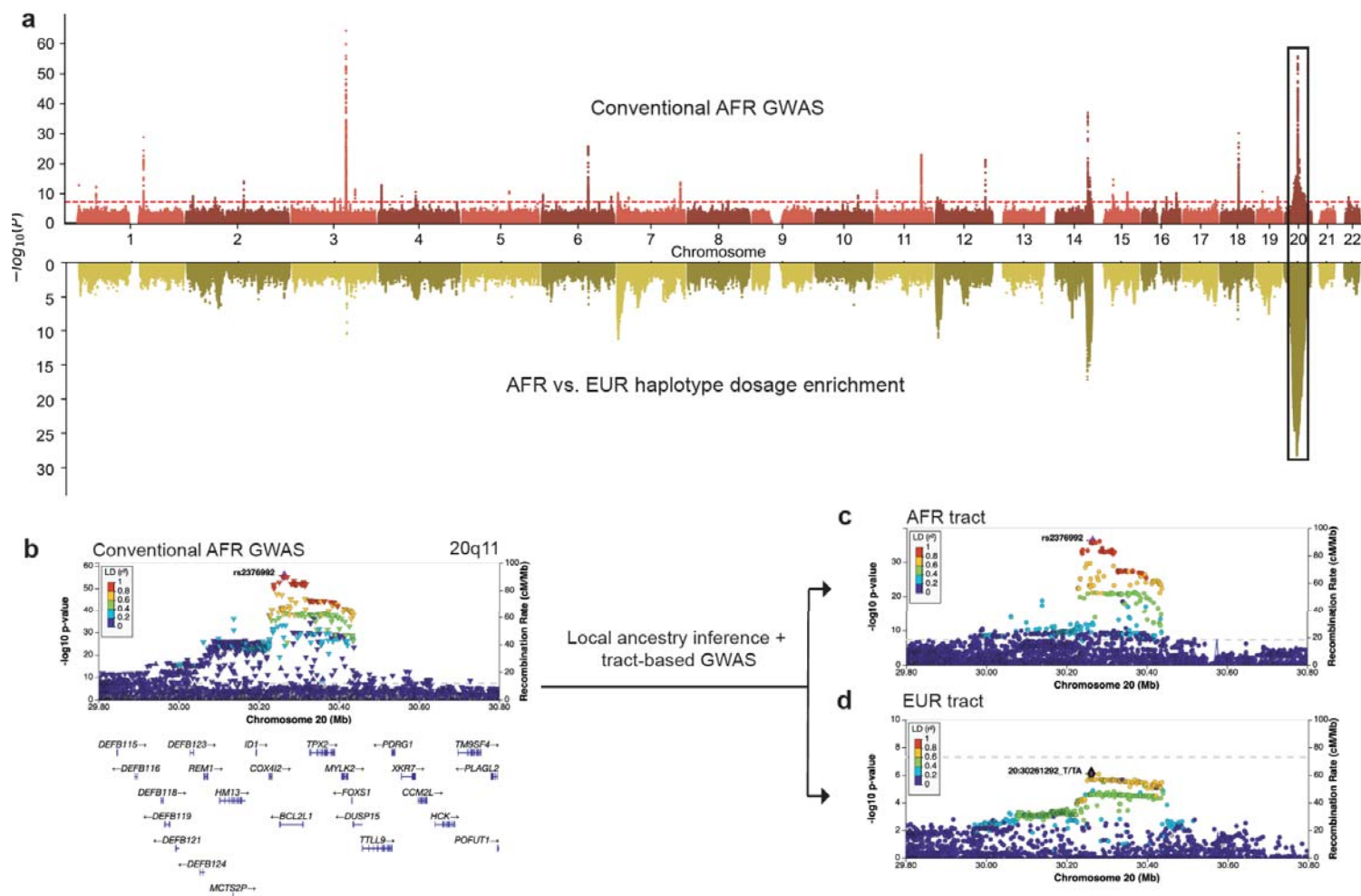
212 We used local ancestry inference and performed a GWAS using the Tractor
213 method²⁵ as a secondary analysis to resolve ancestry-specific signals in AFR
214 (Supplementary Fig. 8). Signals resulting from recent admixture with large allele
215 frequency differences between ancestries can be disentangled with this method; we
216 highlight two AFR loci to illustrate. First, inflation due to admixture was observed at
217 20q11.21 near *BCL2L1* (Fig. 2a). Because global adjustment for ancestry by including
218 20 principal components (PCs) in this GWAS model did not sufficiently resolve this
219 locus (Fig. 2b), we inferred that the inflation was due to the differences in risk allele
220 frequency at the lead SNP rs2376992 across ancestral haplotypes (51% AFR vs 22%
221 EUR, Supplementary Fig. 9a) at this large effect size locus (OR=0.798 [0.776, 0.820];
222 $P = 1.7 \times 10^{-56}$).

223 This was further supported by 20q11.21 having a highly significant difference in
224 AFR and EUR haplotype dosages ($P = 5.6 \times 10^{-29}$; Fig 2a). The secondary cluster of
225 significant variants downstream of *BCL2L1* was resolved in the AFR tract relative to the
226 EUR tract (Fig. 2cd; Supplementary Fig. 9b). Strikingly, a meta-analysis of the AFR and
227 EUR tracts for chromosome 20 yielded a genomic control (λ) of 1.08, as compared to
228 $\lambda = 1.40$ in the conventional AFR GWAS. We then fine-mapped the smaller credible set

229 of SNPs identified in the AFR tract. The risk allele of the SNP with the highest posterior
230 probability, rs2376992, is found in a known promoter region for *BCL2L1*
231 (ENSR00001234227).

232 At 18q12.3 (*SETBP1*), a known EUR mLOY locus, we observed a complex LD
233 structure with multiple causal SNPs which inhibited fine mapping of this locus in AFR
234 (Supplementary Fig. 9c). Tract-based association analysis resolved the overlapping LD
235 structures, and revealed the primary AFR signal at rs4414576 (Supplementary Fig. 9d);
236 this allele had 32% frequency in AFR and only 3% in EUR. The EUR tract at this locus
237 (Supplementary Fig. 9e) had a similar structure to the EUR GWAS (Supplementary Fig.
238 9f).

239 Additionally, we performed a tract-based analysis in HIS mLOY cases
240 (Supplementary Fig. 10). In this model we accounted for the two main ancestral
241 contributors, EUR and Native American (NAT). Similar to AFR, we found the most
242 significant differential enrichment of haplotypes at 20q11.21 (Supplementary Fig. 11).
243 This locus in HIS is inflated by the low frequency of the lead variant rs2376992 in NAT
244 haplotypes (~0.3%, compared to 49% in AFR and 78% in EUR).



245
 246 **Fig. 2. Global- vs. local-ancestry-adjusted GWAS of mosaic loss of Y in admixed African (AFR) ancestry Million**
 247 **Veteran Program participants.** **a**, Miami plot showing conventional AFR GWAS globally adjusted by principal
 248 components (top) and AFR vs. European (EUR) haplotype dosage enrichment (bottom). **b**, At 20q11 (*BCL2L1*) we
 249 observed inflation due to local ancestry. This inflation was resolved by separating the AFR and EUR tracts as shown in **c**
 250 and **d**.

251 *Gene set, tissue and cell-type enrichment of mLOY genes*

252 Using the multi-ancestry meta-analysis, genes mapped within 10Kb of GWS
253 multi-ancestry LD blocks were significantly enriched for two GTEx v8 general tissue
254 types, blood ($P=3.11\times 10^{-10}$) and spleen ($P=3.85\times 10^{-6}$), consistent with expectations of
255 mLOY primarily occurring in leukocytes (Supplementary Fig. 8a). These mapped genes
256 were then used to compare gene set enrichment in all loci (Supplementary Data 10) and
257 novel loci (Supplementary Data 11). The most highly enriched Gene Ontology Biological
258 Process (GO BP) for novel loci was Cell Cycle (47 novel loci), indicating genes involved
259 in the replication and segregation of genetic material and cell division; this was in
260 addition to 64 Cell Cycle known loci. Hallmark gene sets, representing well-defined
261 biological processes, were used as a framework for categorization of novel mLOY loci.
262 Novel genes associated with the G2/M checkpoint were the most significantly enriched
263 with 18 novel loci and FDR-adjusted P -value ($\text{adj}P=3.16\times 10^{-6}$), followed by the
264 PI3K/AKT/mTOR (10 novel loci; $\text{adj}P=1.38\times 10^{-4}$), and heme metabolism (11 novel loci;
265 $\text{adj}P=9.07\times 10^{-3}$) gene sets (Supplementary Fig. 12bc).

266 We then tested for enrichment of EUR index variants located in cell-specific open
267 chromatin regions, by intersecting our genetic associations with data from two catalogs
268 of the human epigenome that profile major human body lineages and blood cell
269 lines^{26,27}. At the tissue level, we found significant enrichment only in myeloid/erythroid
270 cells (Supplementary Fig. 13a; $\text{adj}P=1.2\times 10^{-4}$). Of the blood cell lines, the highest
271 enrichment was measured for multipotent progenitors (MPP; $\text{adj}P=6.4\times 10^{-4}$) and their
272 subsequent differentiation stages, i.e. common myeloid progenitors (CMP; $\text{adj}P=1.2\times 10^{-3}$)
273 and lymphoid-primed multipotent progenitors (Supplementary Fig. 13b; LMP; BH-
274 corrected $P=1.1\times 10^{-3}$), thus supporting the established role of mLOY genetic effects on

275 blood cell differentiation¹⁹. Interestingly, among the six differentiated cell types
276 encompassing myeloid, erythroid, and lymphoid cells (Supplementary Fig. 13b), only
277 erythroid cells exhibited significant enrichment ($\text{adj}P=0.017$). This enrichment pattern of
278 mLOY-related effects on differentiating blood cells contrasts starkly with other diseases
279 characterized by perturbations in immune responses, chronic inflammation, or
280 autoimmune mechanisms, such as Crohn's disease, rheumatoid arthritis, systemic
281 lupus erythematosus, multiple sclerosis or Alzheimer's disease (Supplementary Fig.
282 13c).

283 *eQTLs from single-cell RNA-seq*

284 To identify how SNPs associated with mLOY in EUR associate with gene
285 expression in immune cells, we used a recently published expression quantitative trait
286 loci (eQTL) dataset derived from single-cell RNA-seq data across different immune cell
287 populations²⁸. Across the fourteen immune cell subsets, we found that 197 of 327 EUR
288 mLOY SNPs spanning 251 eQTLs reached at least nominal significance ($P<0.05$;
289 Supplementary Data 12). Of these, 34 eQTLs (22 unique genes; 8 unique SNPs)
290 reached the significance threshold corrected for multiple testing ($P<1.1\times 10^{-5}$); 20 of
291 these eQTLs were associated with two SNPs in the major histocompatibility complex
292 (MHC) region (6p22.1 to 6p21.3).

293

294 *Multi-tissue TWAS and SMR*

295 We linked our EUR GWAS signals with functional gene units in a multi-tissue
296 transcriptome-wide association study (TWAS). Our TWAS leveraged 43 tissue models

297 including STARNET Blood²⁹ and a high-powered dorsolateral prefrontal cortex (DLPFC)
298 dataset³⁰ (Supplementary Data 13); this yielded 2,297 unique significant gene features
299 at $P_{Bonferroni} < 0.05$. In the STARNET blood model, 117 features were significant at
300 $P_{Bonferroni} < 0.05$, including one novel gene that did not appear in GWAS, *MED19*, a
301 component of the Mediator complex involved in the regulated transcription of RNA
302 polymerase II-dependent genes (Supplementary Data 14, Supplementary Fig. 14a). In
303 the DLPFC model, 191 features were significant at $P_{Bonferroni} < 0.05$; the novel genes
304 identified in DLPFC were *IL21R*, (cytokine receptor for interleukin 21) and *COX7A2L*
305 (Supplementary Data 15). All tissues were then meta-analyzed using ACAT³¹, yielding a
306 total of 683 genes with $P_{Bonferroni} < 0.05$ (Supplementary Data 16, Supplementary Fig.
307 14b). ACAT revealed an additional five novel genes: *PSTPIP2*, *CCNK*, *RAD54L2*,
308 *PARP10*, and *G3BP1*, plus the non-coding gene *LINC01933* and pseudogene
309 *AC091982.1*. We found that mLOY-associated gene expression was highly correlated
310 across the imputed transcriptomes of all tissues (Supplementary Fig. 15).

311 We further performed summary-data-based Mendelian randomization (SMR)
312 experiments to provide support for inference of causality. Across 33 tissue types, we
313 identified 1,870 significant genes with $FDR < 0.05$ and $P_{HEIDI} \geq 0.05$, and of these, 234
314 were identified in Blood SMR (Supplementary Data 17). SMR in Blood provided causal
315 support for 23 significant genes (20%) from STARNET Blood, and 51 genes (7%)
316 across the combined TWAS findings from STARNET Blood, DLPFC, and ACAT meta-
317 analysis (Supplementary Fig. 16).

318 *Genetic correlations of mLOY*

319 Genetic correlations (r_g) between mLOY in EUR and 750 traits were tested
320 (Supplementary Data 18). At the multiple testing corrected significance threshold
321 $P < 6.67 \times 10^{-5}$, 36 traits were significantly correlated. Many metabolite measures had a
322 significant negative r_g with mLOY at this threshold, including particle sizes of cholesterol,
323 phospholipids, triglycerides, and total lipids in VLDL, which had r_g ranging from -0.34 (se
324 = 0.07) to -0.28(0.06). Maternal smoking around birth was positively correlated at $r_g =$
325 0.12 (0.03). Significant and negative r_g were observed for hip circumference and the
326 related anthropometric measures obesity class 2, and BMI.

327 *Association of PGS Catalog-based polygenic scores with mLOY*

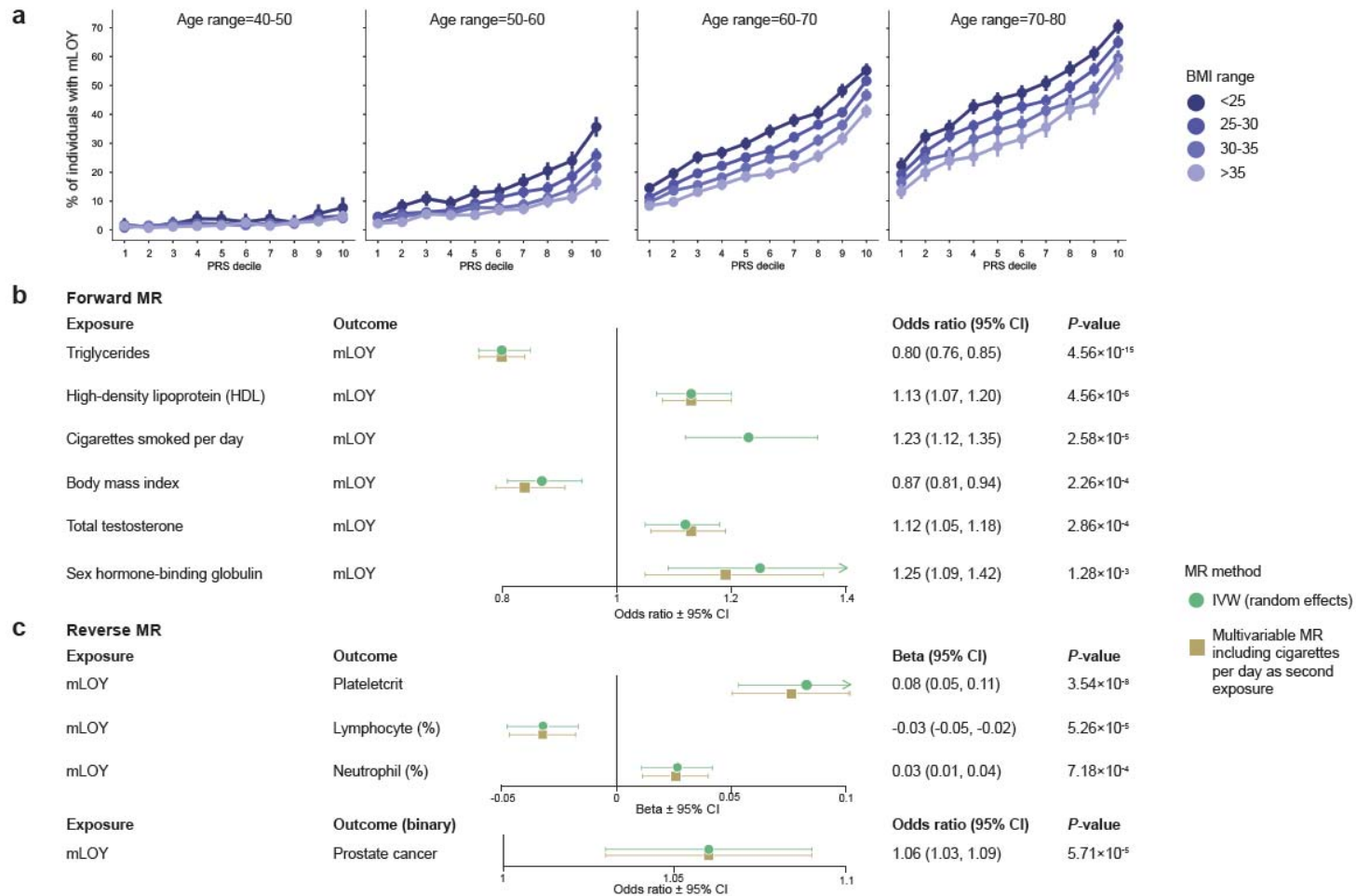
328 We calculated polygenic scores (PGS) for every trait in the PGS Catalog³², for all
329 EUR subjects, and performed a phenome-wide scan for the association of normalized
330 PGS scores with mLOY case-control status (PGS-WAS) to discover shared genetic
331 etiology (Supplementary Data 19). A total of 2,644 scores corresponding to 562
332 uniquely mapped traits in the Experimental Factor Ontology (EFO) were tested; we
333 found 82 mapped traits significant after multiple testing correction ($\alpha = 0.05/562$). Many of
334 the most significant traits were blood measures, including increased platelet crit (1.11
335 [1.10,1.12]), leukocyte count (1.07 [1.06,1.08]), monocyte count (1.07 [1.06,1.08]), and
336 neutrophil count (1.06 [1.05,1.07]), and decreased mean corpuscular hemoglobin
337 concentration (0.94 [0.93,0.95]). Several metabolic traits were also associated, such as
338 triglycerides (0.925 [0.914,0.935]), HDL (1.06 [1.04,1.07]), BMI (0.945 [0.94,0.96]),
339 SHBG (1.05 [1.04,1.06]), T2D (0.93 [0.91,0.94]), as well as smoking status (1.05
340 [1.04,1.06]).

341 *Multi-trait conditioning of mLOY on cigarettes per day*

342 In MVP participants we observed a strong observational association of mLOY
343 with smoking, as well as in the PGS-WAS. To parse out the residual effects of smoking
344 on mLOY susceptibility, we conducted multi-trait-based conditional and joint association
345 analysis³³ (mtCOJO) in EUR, conditioning on cigarettes per day³⁴. Across the genome,
346 only the 15q25 region was primarily impacted by the conditional analysis
347 (Supplementary Fig. 17a). This signal, which was GWS for mLOY in the primary GWAS,
348 was attenuated towards the null after conditioning (Supplementary Fig. 17bc). This
349 locus contains a well-known cluster of smoking-related genes, including *CHRNA5* which
350 encodes a nicotinic acetylcholine receptor subunit that has been frequently associated
351 with smoking in GWAS³⁵.

352 *Polygenic risk and BMI influence penetrance of mLOY in an age-dependent manner*

353 Because we observed a significant negative genetic correlation and PGS-WAS
354 association between mLOY and BMI, an association that has also been reported in
355 previous studies^{36,37}, we sought to examine the prevalence of mLOY in MVP-EUR as a
356 function of age, BMI, and PRS decile derived from UKB¹. In all age bins, we observed
357 that decreasing BMI and increasing PRS were associated with higher mLOY prevalence
358 (Fig. 3a). Overall, our results suggest that this association between mLOY and BMI is
359 not merely confounded by the strong age-dependence of mLOY. We also observed a
360 stronger genetic correlation between mLOY and BMI in MVP ($r_g=-0.110$ (0.026)) than
361 previously reported in UKB ($r_g=-0.052$ (0.032); Supplementary Data 18); this may be a
362 result of greater co-morbidities in the MVP cohort versus UKB.



363

364 **Fig. 3. Mosaic loss of Y by PRS decile and Mendelian randomization (MR).** a, Percentage of individuals with mLOY
 365 by PRS decile, stratified by age decile (plot grid) and BMI range (color). b, Forest plot showing exposure traits with
 366 significant results from random effects inverse-variance weighted (RE IVW) MR and corresponding multivariable MR
 367 (MVMR) with cigarettes per day as an additional exposure, on mLOY outcome in Million Veteran Program Europeans. c,
 368 Significant results from RE IVW MR and MVMR considering mLOY exposure on trait outcomes. CI, confidence interval.

369 *Univariable and multivariable Mendelian randomization*

370 We used the significant PGS Catalog trait associations with mLOY to inform the
371 selection of 32 traits for MR in the EUR cohort, using male-only summary statistics
372 where available, to infer the direction of causality between these exposures and mLOY.
373 Random effects inverse-variance weighted (IVW) forward MR supported six significant
374 traits ($\alpha=0.05/32$) with non-significant MR-Egger intercept ($P \geq 0.05$) as causal influences
375 on mLOY: triglycerides, high-density lipoprotein (HDL), cigarettes per day³⁸, body mass
376 index (BMI), total testosterone, and sex hormone-binding globulin (SHBG) (Fig. 3b;
377 Supplementary Fig. 18; Supplementary Data 20). Reverse MR indicated a significant
378 causal influence of mLOY on plateletcrit, lymphocyte percentage, prostate cancer, and
379 neutrophil percentage (Fig. 3c; Supplementary Fig. 19; Supplementary Data 21). Our
380 finding that mLOY status increases the risk of prostate cancer (OR=1.061 [1.031,
381 1.092]) is in agreement with a recent study using the PRACTICAL Consortium³⁹.

382 Because mLOY is strongly associated with smoking, and because the pleiotropic
383 effects of tobacco smoking instruments have large effects on human health, we
384 conducted multivariable MR using cigarettes per day³⁸ as a second exposure in the
385 models of significant forward and reverse MR associations (Fig. 3bc; Supplementary
386 Data 22). The direct effect of each exposure on their respective outcomes in
387 multivariable MR were highly similar to the univariable model, and remained significant,
388 with the exception of SHBG ($P=8.92 \times 10^{-3}$). Additionally, SHBG had $P > 0.05$ in all of MR-
389 Egger, weighted-median and weighted-mode sensitivity analyses, and so this
390 association is considered less robust than those of the other traits. Overall, multivariable

391 MR demonstrates that the inferred causality identified in univariable MR was
392 independent of cigarette smoking.

393 Discussion

394 In this multi-ancestry meta-analysis we have more than doubled the number of
395 genetic loci associated with mLOY, adding 167 novel loci. The large number of new
396 mLOY cases (N=126,108) that powered our discovery was enabled by several factors.
397 First, we utilized the MVP biobank²⁰, which consists of mostly aging males, many of
398 whom were current or previous cigarette smokers. Next, the MoChA^{4,40} software, which
399 can detect chromosome-length events at a lower cell fraction threshold than in previous
400 GWAS^{1,18}, enabled the inclusion of cases from early stages of mosaicism proliferation.
401 Lastly, the sample size achieved by combining cohorts into a large multi-ancestry meta-
402 analysis increased the number of GWS loci compared to the largest individual cohort
403 (EUR) by about 50%. This mLOY GWAS is the first to include AFR and HIS
404 populations; we identified 5 AFR-specific signals, and additionally found 17 loci in HIS
405 which were replicated in EUR. An additional benefit of MVP is that the extensive
406 electronic health records within this biobank enabled a thorough post-GWAS exploration
407 of the relationships of mLOY with other phenotypes.

408 We found that cell cycle was the most highly enriched Biological Process (GO)
409 for genes positionally mapped to novel loci (47 novel loci), strengthening the
410 mechanistic suppositions of previous studies that reduced cell cycle efficacy is a
411 primary driver of mLOY^{1,17,18}. The most significant gene set in novel loci was G2/M
412 checkpoint, responsible for blocking damaged and incompletely replicated DNA from
413 progressing through the cell cycle. The G2/M checkpoint is regulated in part by p53⁴¹, a
414 tumor suppressor involved with mLOY, CHIP, and other mCAs³. The next most highly
415 enriched novel gene set, the PI3K/AKT/mTOR pathway (13 novel loci), is a regulator of

416 cell growth and survival, particularly in the context of cancer progression⁴². The Heme
417 Metabolism gene set (11 novel loci) also includes genes involved in erythroblast
418 differentiation; the expansion of leukocytes in mLOY has previously been associated
419 with reduced erythrocytes⁴³. Additionally, Xenobiotic Metabolism (11 novel loci) of
420 foreign substances, including cigarettes, environmental pollutants, and chemotherapy
421 drugs, has also been strongly associated with mCAs in previous studies^{11,44,45}.

422 Our forward MR associations were all directionally consistent with a recent
423 observational study in UKB⁴⁶; additionally, their MR analysis of SHBG is in agreement
424 with our finding that SHBG exerts a positive causal influence on mLOY in forward MR⁴⁶.
425 Interestingly, higher triglycerides had a protective causal influence on mLOY (OR=0.84
426 [0.80, 0.89]; $P=5.38 \times 10^{-11}$) while higher HDL conferred risk (OR=1.10 [1.05, 1.15];
427 $P=3.71 \times 10^{-5}$). Though this is opposite of the pattern commonly observed in CVD,
428 previous MR studies⁴⁷ have identified robust associations between HDL and increased
429 risk of breast cancer,^{48,49} and between triglycerides and decreased risk of breast
430 cancer.⁵⁰ Triglycerides also had significant negative genetic correlation and PGS-WAS
431 score with mLOY, and HDL had significant positive PGS-WAS score with mLOY. Two
432 novel mLOY-risk-increasing SNPs are also lead SNPs for triglycerides, at *BUD13*⁵¹ and
433 *SNX17*⁵². Overall, more studies are necessary to uncover the mechanisms underlying
434 this phenomenon. We also found that red blood cell distribution width (RDW) exerts a
435 positive causal influence on mLOY that was nearly significant at the multiple testing
436 threshold (OR= 1.11 [1.04, 1.18]; $P=0.002$). RDW is an index which reflects impaired
437 erythropoiesis and abnormal red blood cell survival. Multiple mLOY-associated cyclin
438 genes are related to RDW mechanism: *CDK6* (novel) promotes G1/S transition, *CCND3*

439 (known) encodes cyclin D3, alters cell cycle progression and reduce control of cell
440 size⁵³, and novel TWAS gene *CCNK* is a cyclin activator.

441 In addition to replicating the result that mLOY can increase risk of prostate
442 cancer³⁹; reverse MR indicated a significant causal influence of mLOY on increased
443 plateletcrit, increased neutrophil percentage, and decreased lymphocyte percentage.
444 The directions of these relationships are concordant with the previous observational
445 report⁴³. Neutrophil-to-lymphocyte ratio (NLR) in peripheral blood is an emerging
446 prognostic factor in many diseases, especially cancers^{43,54}, and elevated NLR indicates
447 neutrophilic inflammatory response, impaired cell-mediated immunity, and is suggestive
448 of overall poor prognosis^{43,54,55}.

449 Our study was not without limitations. First, we performed a cross-sectional study
450 at a single time point. Future studies may benefit from a prospective study design as
451 mLOY is associated with aging⁵⁶. Additionally, it has been shown in MDS that young
452 and old cases have distinct genetic landscapes⁵⁷, which should also be examined in
453 future studies. We also did not consider the effects of environmental exposures aside
454 from smoking and BMI. Veterans may be disproportionately exposed to pollutants and
455 other toxins over their lifetimes compared to the general public⁵⁸. This could have
456 caused our mLOY prevalence estimates to be inflated (although they were in agreement
457 with previous studies), and could exacerbate potential gene-environment interactions on
458 mLOY risk. Next, we did not distinguish between high- and low-cell fraction of mLOY
459 when defining cases. Our classification method could detect very low cell fractions as
460 opposed to most existing studies which used high cell fraction detection methods such
461 as intensity thresholding (i.e. on mLRR-Y). Next, our analysis evaluated DNA from

462 peripheral blood mononuclear cells (PMBCs) only, and did not consider mLOY from
463 other tissues. Finally, as always, significant GWAS results are associations, and not
464 proof of causal disease mechanisms.

465 We found broad concordance across multiple ancestral populations in our meta-
466 analysis, as well as with the previous BBJ cohort¹⁹, strengthening the generalizability of
467 our findings. Future multi-ancestry meta-analyses may enable increased power and
468 associated loci discovery. The new risk loci identified in this study will lead to improved
469 genetic risk prediction, diagnosis, and understanding of the cellular mechanisms
470 surrounding mLOY.

471

472 **Methods**

473 *Ethics/study approval*

474 All participants provided informed consent, and the studies conducted at
475 participating centers received approval from the Institutional Review Boards.

476 *Genotyping, imputation, and ancestry assignment*

477 Genomic data processing was performed for >650,000 MVP participants
478 (releases 1-4). Genotyping was performed using the Thermo Fisher MVP 1.0 Affymetrix
479 Axiom Biobank array²¹. Samples with >2.5% missing genotype calls, excess
480 heterozygosity, those that were potential duplicates, and those with discordance
481 between genetic sex and self-identified gender, were excluded. SNPs with missingness
482 >5% or minor allele frequency (MAF) that deviated by >10% from the 1000 Genomes
483 Project Phase 3 (1KGP3) data⁵⁹ were excluded. Pairwise genetic relatedness was
484 estimated using KING⁶⁰; one individual was removed at random from each pair of first-
485 degree relatives, preferentially retaining cases from case-control pairs. Ancestry was
486 algorithmically assigned using HARE (Harmonized ancestry and race/ethnicity)⁶¹, which
487 incorporates self-reported race and ethnicity data to train a genetic ancestry classifier.
488 Using HARE, we grouped 544,112 male participants in MVP according to European
489 (EUR), African (AFR), or Hispanic (HIS) ancestry.

490 Genotypes were statistically phased over the entire cohort using SHAPEIT4
491 version 4.1.3⁶² with PBWT depth 8. Phased genotypes were imputed to the African
492 Genome Resources (AGR) reference panel⁶³ using Minimac 4. The AGR panel consists
493 of all 5,008 1KGP3 haplotypes and an additional 2,862 haplotypes from unrelated pan-
494 African samples. As AGR contains biallelic SNPs only, a second imputation was

495 performed using 1KGP3, with indels and other complex variants merged into the
496 primary imputation. Imputation for chrX for EUR, AFR, and HIS was performed using
497 TOPMed (hg38); significant loci on chrX were lifted over⁶⁴ to hg19 for reporting.

498 *Detection of mLOY using long-range haplotype phase*

499 We used SHAPEIT4⁶² to infer haplotypes from array genotypes for the whole
500 MVP cohort and we utilized MoChA^{4,40}, an extension to the BCFtools software suite⁶⁵,
501 to infer the presence of mLOY by detecting shifts in allelic ratios between the phased
502 PAR1 and PAR2 haplotypes, similar to what was done previously in UKB¹. This
503 methodology allows to infer the presence of mLOY for cell fractions as low as ~1%. Cell
504 fraction was estimated from B allele frequency deviation (bdev) using the formula
505 $4*bdev / (1+2*bdev)$.

506 *GWAS*

507 Presence or absence of any detectable mLOY cell fraction was used as a case-
508 control trait in all analyses, performed using male participants only. Single variant
509 genome-wide association testing was carried out with REGENIE v1.0.6.7⁶⁶ using age,
510 age-squared, and twenty PCs. REGENIE step 1 was performed using leave-one out
511 cross validation (--loocv). Approximate Firth likelihood ratio test (LRT) was applied as
512 fallback for associations with $P < 0.05$, with SE computed based on LRT where applied.
513 We kept common variants with minor allele frequency (MAF) $\geq 0.1\%$ and minimum
514 imputation quality (INFO) of 0.3. Two significant chrX loci in PAR1 (rs2857319) and
515 PAR2 (rs306890), both near the boundary with the nonPAR, had large frequency
516 differences between X and Y chromosomes in the Genome Aggregation Database

517 (gnomAD v3.1.2). These loci were removed after determining the mLOY allele was
518 exclusively in high cell fraction participants, indicating likely genotyping error
519 (Supplementary Fig. 20). For chrY, genotype calls were tested for associations using
520 PLINK2 (alpha v20211217), with Firth correction applied to all variants.

521 Within each ancestry group, we performed conditional association analyses
522 using Genome-wide Complex Trait Analysis multi-SNP-based conditional and joint
523 association analysis (GCTA-COJO)⁶⁷ to identify secondary association signals at
524 associated loci, using LD reference panels consisting of 100,000 randomly selected
525 participants for EUR and AFR, and all 52,183 participants in HIS. COJO SNPs with $r^2 \geq$
526 0.05 were iteratively retained based on lowest P -value.

527 For replication, MVP-EUR COJO association signals were compared to summary
528 statistics from the previous mLOY GWAS in UKB¹. An updated version of chrX UKB
529 summary statistics were utilized for this comparison ($N_{\text{case}}=40,466$; $N_{\text{control}}=146,066$;
530 personal.broadinstitute.org/giulio/mLOY); MACH R2 values were considered for variant
531 quality in lieu of INFO scores.

532 *Fine-mapping*

533 We performed Bayesian fine-mapping of each genome-wide significant locus in
534 the EUR and AFR using SuSiE⁶⁸. Pairwise SNP correlations were calculated directly
535 from imputed dosages on 320,831 European-ancestry samples in MVP using
536 LDSTORE 2.0⁶⁹. The maximum number of allowed causal SNPs at each locus was set
537 to 10 (the default used in the FinnGen fine-mapping pipeline:
538 <https://github.com/FINNGEN/finemapping-pipeline>). Fine-mapping regions which
539 overlapped the major histocompatibility complex (MHC; chr6:25,000,000-34,000,000)

540 were excluded. High quality credible sets were defined as those with minimum $r^2 < 0.5$
541 between variants (88/422 discarded in EUR, 24/69 in AFR).

542 *Rare variant analysis*

543 We conducted association analyses for rare variants in each MVP ancestry
544 group using REGENIE and the same covariates as in standard GWAS. We considered
545 only variants genotyped on the MVP 1.0 array²¹, which is enriched in protein-altering
546 rare variants, and applied the Rare Heterozygous Adjustment algorithm⁷⁰ to improve the
547 positive predictive value of rare genotype calls. We further restricted the included
548 markers to directly genotyped ultra-rare variants (MAF < 0.1% in controls) classified as
549 “high-impact”⁷¹. Rare variants were categorized as somatic or germline based on allele
550 balance for heterozygotes obtained from the Genome Aggregation Database
551 (gnomAD)⁷².

552 *GWAS multi-ancestry meta-analysis*

553 MVP EUR, AFR, and HIS cohorts were filtered by INFO > 0.5 to retain only high
554 quality variants, and meta-analyzed for fixed effects using METAL (v20200505),
555 weighting effect sizes by the inverse of their corresponding standard errors. Only
556 variants present in two or more ancestries were retained. Loci were defined for all
557 cohorts (including counting loci in UKB¹ for comparison purposes), using the two-stage
558 “clumping” procedure implemented in the Functional Mapping and Annotation (FUMA)
559 platform⁷³. In this process, genome-wide significant variants are collapsed into LD
560 blocks ($r^2 > 0.6$) and subsequently re-clumped to yield approximately independent
561 ($r^2 < 0.1$) signals; adjacent signals separated by < 250kb are ligated to form independent

562 loci. Novel variants were defined as COJO signals in independent ancestry cohorts, or
563 meta-analysis index variants, located >1Mb from a previously reported GWAS
564 association with mLOY. For the multi-ancestry meta-analysis, we further performed a
565 sensitivity analysis using the Han-Eskin random effects model (RE2) in METASOFT
566 v2.0.1²⁴. FE and RE2 *P*-values at top loci were highly similar. We compared our meta-
567 analysis lead variants to UKB¹ as described above and to BBJ¹⁹ where available. BBJ
568 reported their results as mLRR-Y intensity thresholding as a proxy for mean Y
569 chromosome dosage in circulating blood cells of subjects.

570 *Local ancestry deconvolution and tract-based GWAS*

571 We inferred local ancestry within AFR participants assuming two-way (AFR/EUR)
572 admixture, and within HIS assuming three-way (AFR/EUR/NAT) admixture. The 1000
573 Genomes YRI (N=108) and CEU (N=99) populations, were used as the AFR and EUR
574 reference, respectively, and 43 Native American samples from Mao et al.^{74,75} were used
575 as the NAT reference. We used RFMIX⁷⁶ version 2 to generate local ancestry calls for
576 phased genotypes. We then extracted ancestry-specific dosages from the imputed data
577 into PLINK 2.0-compatible files⁷⁷ using custom scripts based on the Tractor workflow²⁵.
578 For the AFR analysis, EUR-specific dosages were put into a PGEN file, and African-
579 specific dosages and EUR haplotype counts were interlaced in a zstandard-compressed
580 table. For the HIS analysis, EUR-specific dosages were put into a PGEN file, with
581 African and NAT-specific dosages and EUR and AFR haplotype counts interlaced into a
582 zstandard-compressed table. We used these files to conduct a local ancestry-aware
583 GWAS using the PLINK 2.0 local covariates feature, obtaining ancestry-specific
584 marginal effect size estimates.

585 *Gene set, tissue and cell type enrichment analysis*

586 FUMA GENE2FUNC⁷³ was performed using multi-ancestry meta-analysis
587 summary statistics in genes that were positionally mapped to significant variants (within
588 10 Kbp) excluding the MHC gene region; this analysis was also stratified by GWAS
589 locus novelty. Benjamini-Hochberg (FDR) was used as the gene set enrichment multiple
590 test correction method. Hallmark gene sets⁷⁸ were used to categorize genes.

591 To further evaluate whether the genomic loci implicated in mLOY were enriched
592 in any particular cell type, we intersected common mLOY risk variants with broad and
593 blood-specific epigenomic catalogs of cell-specific open chromatin^{26,27} using an LD
594 score partitioned heritability approach (LDSC)⁷⁹ (Fig. S_ctype_a-b). For the broad
595 epigenome catalog encompassing various human tissues²⁶, we re-used the open
596 chromatin regions associated with each tissue from the lists provided by the creators of
597 the atlas
598 (https://www.meuleman.org/DHS_Index_and_Vocabulary_hg38_WM20190703.txt.gz).

599 To identify cell-specific chromatin regions within the epigenome map of human blood
600 lineages²⁷, we conducted differential analysis on sequencing data sourced from the
601 Gene Expression Omnibus (GEO) under accession GSE74912. To ensure a consistent
602 evaluation of the generated LD-sc statistics, which rely on the overall genomic coverage
603 of the tested chromatin regions, we selected an identical number of the most specific
604 open chromatin regions from each blood lineage for subsequent heritability analysis by
605 LDSC. For contextual comparison of heritability signal with the other diseases, we
606 acquired summary statistics of Crohn's disease⁸⁰, rheumatoid arthritis⁸¹, systemic lupus
607 erythematosus⁸², multiple sclerosis⁸³, or Alzheimer's disease⁸⁴. Similarly to the FUMA

608 analysis, the MHC region was excluded but otherwise the default parameters of LDSC
609 were used for the analysis.

610 *eQTL analysis using published datasets*

611 We interrogated a previously published dataset²⁸ for SNPs associated with
612 mLOY in our European ancestry cohort. Chromosome, position, and alleles were used
613 as unique identifiers with which to cross-reference SNPs across different immune cell
614 subsets.

615 *Transcriptomic imputation model construction and transcriptome-wide association study*

616 Transcriptomic imputation models were constructed as previously described^{85,86}
617 for tissues of the GTEx⁸⁷ v8, STARNET²⁹ and PsychENCODE^{30,88} cohorts. For GTEx
618 and STARNET cohorts, we considered adipose tissue: subcutaneous (GTEx &
619 STARNET) and visceral (GTEx & STARNET); arterial tissue: aorta (GTEx &
620 STARNET), coronary (GTEx), mammary (STARNET), and tibial (GTEx); blood (GTEx &
621 STARNET); cell lines (GTEx): EBV-transformed lymphocytes and transformed
622 fibroblasts; endocrine (GTEx): adrenal gland, pituitary, and thyroid; colon (GTEx):
623 sigmoid and trasverse; esophagus (GTEx): gastroesophageal junction, mucosa and
624 muscularis; pancreas (GTEx); salivary gland minor (GTEx); stomach (GTEx); terminal
625 ileum (GTEx); heart (GTEx): atrial appendage and left ventricle; liver (GTEx &
626 STARNET), skeletal muscle (GTEx & STARNET); nerve tibial (GTEx); reproductive
627 (GTEx): mammary tissue, ovary, prostate, testis, uterus, vagina; lung (GTEx); skin
628 (GTEx): not sun exposed suprapubic and sun exposed lower leg; and spleen (GTEx).
629 From PsychENCODE^{30,88} we considered brain: dorsolateral prefrontal cortex (DLPFC)
630 genes. The genetic datasets of the GTEx⁸⁷, STARNET²⁹ and PsychENCODE⁸⁸ cohorts

631 were uniformly processed for quality control (QC) steps before genotype imputation as
632 previously described^{85,86}. We restricted our analysis to samples with European ancestry
633 as previously described⁸⁵. Genotypes were imputed using the University of Michigan
634 server⁸⁹ with the Haplotype Reference Consortium (HRC) reference panel⁹⁰. Gene
635 expression information was derived from RNA-seq gene level counts, which were
636 adjusted for known and hidden confounders, followed by quantile normalization. For
637 GTEx, we used publicly available, quality-controlled, gene expression datasets from the
638 GTEx consortium (<http://www.gtexportal.org/>). RNA-seq data for STARNET were
639 obtained in the form of residualized gene counts from a previously published study²⁹.
640 For the dorsolateral prefrontal cortex from PsychENCODE we used post-quality-control
641 RNA-seq data that were fully processed, filtered, normalized, and extensively corrected
642 for all known biological and technical covariates except the diagnosis status³⁰ as
643 previously described⁸⁶. Feature types queried include genes, long non-coding RNA
644 (lincRNA), microRNA, processed transcripts, pseudogenes, RNA, small nucleolar RNA
645 (snoRNA), plus constant (C), joining (J), and variable (V) gene segments.

646 For population classification we used individuals of known ancestry from 1000
647 Genomes. We excluded variants in regions of high linkage disequilibrium, variants with
648 $MAF < 0.05$, variant with high missingness (> 0.01), and variants with Hardy-Weinberg
649 equilibrium $P < 1 \times 10^{-10}$; the remaining variants were pruned (--indep-pairwise 1000 10
650 0.02 with PLINK⁹¹) and PCA was performed with PLINK⁷⁷ version 2.0. We used the first
651 (PC1), second (PC2) and third (PC3) ancestral PCs to define an ellipsoid based on
652 1000Gp3v5 EUR samples⁵⁹ and samples within 3 SD from the ellipsoid center were
653 classified as EUR; based on this definition of EUR samples, we excluded one non-

654 European ancestry individual. In the remaining samples ($n = 405$), we performed
655 additional sample-level quality control by retaining non-related samples (--king-cutoff
656 0.0884 with PLINK⁷⁷ version 2.0) with sample-level missingness < 0.015 for variants
657 with variant-level missingness < 0.02 , and heterozygosity rate of $< 3SD$ away from the
658 mean; of note, no samples were excluded by these steps. For the next step of our
659 pipeline, we performed outlier testing in the gene expression data. After performing
660 counts per million filtering (> 0.5 counts per million in at least 30% of samples) and
661 voom normalization, PCA was performed, and we excluded individuals located more
662 than 4 SD away from the mean of the ellipsoid defined by PC1 to PC3. This did not
663 remove any individuals but assured us that our data did not contain any outliers. In this
664 final set of individuals, we performed variant-level quality control of the genotypes by
665 removing variants with less than 0.01 minor allele frequency (for all variants possible,
666 we utilized minor allele frequencies reported by Allele Frequency Aggregator European
667 population⁹², to reduce minor allele frequency bias from the comparatively small
668 imputation model training population), 5 minor allele counts and 0.02 missingness rate;
669 only variants present in the reference panel of the Haplotype Reference Consortium
670 were retained to ensure good representation of variants in the target GWAS⁹⁰. We used
671 this final set of quality-controlled genotypes in conjunction with our normalized
672 expression data to discover the optimal number of PEER factors to find expression
673 quantitative trait loci. Our analysis led to the decision to utilize 15 PEER factors, which
674 had resulted in the discovery of 4,299 significant eQTLs⁹³. This was the closest value to
675 90% of the maximum value of eQTLs discovered by any chosen number of PEER
676 factors (4,844 significant eQTLs from 50 PEER factors). This allowed us to retain the

677 maximum signal for gene expression prediction without overcorrecting our data. After
678 residualization for 15 PEER factors, expression data were quantile normalized.
679 Genotypes were then converted to dosages, and missing values were replaced with
680 twice the variant's minor allele frequency before dosages were rounded to the nearest
681 whole number. For training, we used PrediXcan⁹⁴ for the construction of the retinal
682 transcriptomic imputation model due to a lack of SNP epigenetic annotation information;
683 for all other models, we used EpiXcan⁸⁵.

684 *Multi-tissue transcriptome-wide association study (TWAS)*

685 We performed a gene-trait association analysis as previously described⁸⁵. We
686 applied the S-PrediXcan method⁹⁵ to integrate the summary statistics and the
687 transcriptomic imputation models constructed above to obtain gene-level association
688 results. *P*-values were adjusted for multiple testing using the Benjamini & Hochberg
689 (FDR) method and Bonferroni correction. *P*-values across tissues were meta-analyzed
690 using ACAT³¹ ≤ 0.05 and predictive $r^2 > 0.01$ to control for both significance and variance
691 explained.

692 *Summary-data-based Mendelian randomization*

693 To test for joint associations between GWAS summary statistics SNPs and
694 eQTL, the SMR method⁹⁶, a Mendelian randomization approach, was used. Top SNPs
695 used in SMR for each probe were selected as the most significant SNP in the eQTL
696 data which was also present in the GWAS data. The SMR software (v1.03) was run
697 using the default settings using GTEx Consortium⁸⁷ v8 whole blood tissue. European
698 samples of the 1KGP were used as a reference panel. Bonferroni multiple-testing

699 correction was applied on SMR P -values (PSMR). Moreover, a post-filtering step was
700 applied by conducting heterogeneity in dependent instruments (HEIDI) test. The HEIDI
701 test distinguishes the causality and pleiotropy models from the linkage model by
702 considering the pattern of associations using all the SNPs that are significantly
703 associated with gene expression in the cis-eQTL region. The null hypothesis is that a
704 single variant is associated with both trait and gene expression, while the alternative
705 hypothesis ($P_{\text{HEIDI}} < 0.05$) is that trait and gene expression are associated with two
706 distinct variants. The same tissues as in the TWAS section from the V8 release of the
707 GTEx Consortium⁸⁷ were queried in SMR.

708 *Heritability and genetic correlation analyses*

709 Genetic correlation analyses were performed using linkage disequilibrium score
710 regression (LDSC)⁹⁷ using the provided European-ancestry LD scores derived from
711 1KGP, as implemented in LDHub (v1.9.0)⁹⁸. Bonferroni multiple testing correction was
712 applied. SNPs from the MHC region (chr6:26M~34M) were removed.

713 *Association of PGS Catalog-based polygenic scores with mLOY status*

714 Phenome-wide polygenic score files for 2,652 traits were obtained from
715 European Molecular Biology Laboratory's European Bioinformatics Institute (EMBL-EBI)
716 PGS Catalog (September 2022 version)³². All EUR-ancestry subjects in MVP were
717 scored across all available PGSs, excluding those derived from other MVP studies (20),
718 using the +score plugin (<https://github.com/freeseeek/score>) of bcftools⁶⁵. PGSs were
719 loaded into the dosage format field of VCFs readable by SAIGE v1.1.6.2⁹⁹ for
720 association testing. Logistic regression was used to examine associations of PGSs on

721 MVP EUR mLOY cases and controls, adjusting for the same covariates as in GWAS
722 (sex, age, mean-centered age-squared, and 20 ancestry-specific PCs).

723 *Conditional meta-analysis (mtCOJO)*

724 In order to assess the residual effects of genetic predisposition to cigarette
725 smoking from mLOY susceptibility, we conducted a multi-trait meta-analysis²⁵
726 conditioned on cigarettes per day³⁴. The conditional meta-analyses were performed
727 using the EUR mLOY summary statistics using GCTA-mtCOJO⁴⁶. The EUR LD panel
728 described above for use with COJO was also used in this analysis.

729 *Polygenic risk scoring*

730 UKB summary statistics¹ were used to construct a polygenic risk score for EUR
731 MVP participants with PRS-CS¹⁰⁰ (v20210604) with a global shrinkage prior of 1×10^{-4} .
732 European samples of the 1KGP were used as a reference panel. Variants were filtered
733 to include only those with $R^2 > 0.8$ and $MAF > 1\%$.

734 *Univariable and multivariable Mendelian randomization*

735 Forward and reverse two-sample Mendelian randomization was performed using
736 summary statistics from previous European GWAS. Summary statistics were accessed
737 through the OpenGWAS database API¹⁰¹ via the GWAS codes listed in **Supplementary**
738 **Data 20-22**, except body mass index in males only from the GIANT (Genetic
739 Investigation of ANthropometric Traits) consortium¹⁰², and cigarettes per day from the
740 GSCAN (GWAS & Sequencing Consortium of Alcohol and Nicotine use) consortium³⁸,
741 which were downloaded separately. The genome-wide significance threshold of
742 $P < 5 \times 10^{-8}$ was used for the selection of genetic instrumental variables. LD clumping of

743 $r^2 < 0.001$ within a 10 Mb window was used to identify independent instruments.
744 Selection, clumping, and harmonization of instruments was performed using
745 TwoSampleMR (v0.5.7)¹⁰³. Primary analyses used the random-effect inverse-variance
746 weighted (IVW) method. Sensitivity analyses were performed with the
747 MendelianRandomization (v0.6.0) R package¹⁰⁴ using fixed effect IVW, and by
748 achieving a nominal significance threshold ($P < 0.05$) using one of either MR-Egger,
749 weighted median or weighted mode methods. Additionally we required $P > 0.05$ for MR-
750 Egger intercept. MR-PRESSO was conducted using MRPRESSO (v1.0) to test for
751 horizontal pleiotropy¹⁰⁵.

752 To control for the possibility that genetic instruments related to mLOY displayed
753 possible horizontal pleiotropic effects via smoking behavior, we conducted multivariable
754 Mendelian randomization (MVMR) using the MVMR R package v0.4¹⁰⁶ and included a
755 second exposure of cigarettes per day³⁸. We report inverse-variance weighted
756 multivariable MR results along with the test for heterogeneity from a modified form of
757 Cochran's Q statistic with respect to differences in MVMR estimates across the set of
758 instruments. Covariance between the effect of genetic variants derived from the two
759 exposures was fixed to zero due to the use of non-overlapping samples. The F -statistic
760 for instrument strength achieved $F > 10$ for all tests.

761

762 ***Consortium authors and affiliations**

763 **VA Million Veteran Program**

764 J. Michael Gaziano^{33,34}, Philip S. Tsao^{16,17,18}, Saiju Pyarajan^{1,32}

765 ³³Massachusetts Veterans Epidemiology Research and Information Center (MAVERIC), VA
766 Boston Healthcare System, Boston, MA, USA; ³⁴Division of Aging, Department of Medicine,
767 Brigham and Women's Hospital, Harvard Medical School, Boston, MA, USA
768

769 **Data availability**

770 The full summary level association data from the meta-analysis and individual
771 population association analyses in MVP will be available via the dbGaP study accession
772 number phs001672. Full transcriptome-wide association study results are available
773 upon request.

774

775

776 **Competing interests**

777 A.G.B. is on the scientific advisory board of TenSixteen Bio unrelated to the present
778 work. P.N. reports research grants from Allelica, Amgen, Apple, Boston Scientific,
779 Genentech / Roche, and Novartis, personal fees from Allelica, Apple, AstraZeneca,
780 Blackstone Life Sciences, Creative Education Concepts, CRISPR Therapeutics, Eli Lilly
781 & Co, Foresite Labs, Genentech / Roche, GV, HeartFlow, Magnet Biomedicine, Merck,
782 and Novartis, scientific advisory board membership of Esperion Therapeutics, Preciseli,
783 and TenSixteen Bio, scientific co-founder of TenSixteen Bio, equity in MyOme,
784 Preciseli, and TenSixteen Bio, and spousal employment at Vertex Pharmaceuticals, all
785 unrelated to the present work. D.K. is a scientific advisor and reports consulting fees
786 from Bitterroot Bio, Inc unrelated to the present work. The other authors declare no
787 competing interests.

788 References

- 789 1. Thompson, D. J. *et al.* Genetic predisposition to mosaic Y chromosome loss in blood.
790 *Nature* **575**, 652–657 (2019).
- 791 2. Forsberg, L. A. *et al.* Mosaic loss of chromosome Y in leukocytes matters. *Nature genetics*
792 vol. 51 4–7 (2019).
- 793 3. Zekavat, S. M. *et al.* Hematopoietic mosaic chromosomal alterations increase the risk for
794 diverse types of infection. *Nat. Med.* **27**, 1012–1024 (2021).
- 795 4. Loh, P.-R., Genovese, G. & McCarroll, S. A. Monogenic and polygenic inheritance become
796 instruments for clonal selection. *Nature* **584**, 136–141 (2020).
- 797 5. Quintana-Murci, L. & Fellous, M. The Human Y Chromosome: The Biological Role of a
798 ‘Functional Wasteland’. *J Biomed Biotechnol.* **1**, 18–24 (2001).
- 799 6. Pierre, R. V. & Hoagland, H. C. Age-associated aneuploidy: loss of Y chromosome from
800 human bone marrow cells with aging. *Cancer* **30**, 889–894 (1972).
- 801 7. Hubbard, A. K., Brown, D. W. & Machiela, M. J. Clonal hematopoiesis due to mosaic
802 chromosomal alterations: Impact on disease risk and mortality. *Leuk. Res.* **126**, 107022
803 (2023).
- 804 8. Dumanski, J. P. *et al.* Immune cells lacking Y chromosome show dysregulation of
805 autosomal gene expression. *Cell. Mol. Life Sci.* **78**, 4019–4033 (2021).
- 806 9. Kar, S. P. *et al.* Genome-wide analyses of 200,453 individuals yield new insights into the
807 causes and consequences of clonal hematopoiesis. *Nat. Genet.* **54**, 1155–1166 (2022).
- 808 10. Kessler, M. D. *et al.* Common and rare variant associations with clonal haematopoiesis
809 phenotypes. *Nature* **612**, 301–309 (2022).
- 810 11. Jakubek, Y. A., Reiner, A. P. & Honigberg, M. C. Risk factors for clonal hematopoiesis of
811 indeterminate potential and mosaic chromosomal alterations. *Transl. Res.* **255**, 171–180
812 (2023).

- 813 12. Ljungström, V. *et al.* Loss of Y and clonal hematopoiesis in blood—two sides of the same
814 coin? *Leukemia* **36**, 889–891 (2021).
- 815 13. Jaiswal, S. *et al.* Clonal Hematopoiesis and Risk of Atherosclerotic Cardiovascular Disease.
816 *N. Engl. J. Med.* **377**, 111–121 (2017).
- 817 14. Sano, S. *et al.* Hematopoietic loss of Y chromosome leads to cardiac fibrosis and heart
818 failure mortality. *Science* **377**, 292–297 (2022).
- 819 15. Abdel-Hafiz, H. A. *et al.* Y chromosome loss in cancer drives growth by evasion of adaptive
820 immunity. *Nature* **619**, 624–631 (2023).
- 821 16. Pan-UKB team. <https://pan.ukbb.broadinstitute.org>. (2020).
- 822 17. Wright, D. J. *et al.* Genetic variants associated with mosaic Y chromosome loss highlight
823 cell cycle genes and overlap with cancer susceptibility. *Nat. Genet.* **49**, 674–679 (2017).
- 824 18. Zhou, W. *et al.* Mosaic loss of chromosome Y is associated with common variation near
825 TCL1A. *Nat. Genet.* **48**, 563–568 (2016).
- 826 19. Terao, C. *et al.* GWAS of mosaic loss of chromosome Y highlights genetic effects on blood
827 cell differentiation. *Nat. Commun.* **10**, 4719 (2019).
- 828 20. Gaziano, J. M. *et al.* Million Veteran Program: A mega-biobank to study genetic influences
829 on health and disease. *J. Clin. Epidemiol.* **70**, 214–223 (2016).
- 830 21. Hunter-Zinck, H. *et al.* Genotyping Array Design and Data Quality Control in the Million
831 Veteran Program. *Am. J. Hum. Genet.* **106**, 535–548 (2020).
- 832 22. Scheller, M. *et al.* Hotspot DNMT3A mutations in clonal hematopoiesis and acute myeloid
833 leukemia sensitize cells to azacytidine via viral mimicry response. *Nat Cancer* **2**, 527–544
834 (2021).
- 835 23. Cerchione, C. *et al.* IDH1/IDH2 Inhibition in Acute Myeloid Leukemia. *Front. Oncol.* **11**,
836 639387 (2021).
- 837 24. Han, B. & Eskin, E. Random-effects model aimed at discovering associations in meta-
838 analysis of genome-wide association studies. *Am. J. Hum. Genet.* **88**, 586–598 (2011).

- 839 25. Atkinson, E. G. *et al.* Tractor uses local ancestry to enable the inclusion of admixed
840 individuals in GWAS and to boost power. *Nat. Genet.* **53**, 195–204 (2021).
- 841 26. Meuleman, W. *et al.* Index and biological spectrum of human DNase I hypersensitive sites.
842 *Nature* **584**, 244–251 (2020).
- 843 27. Corces, M. R. *et al.* Lineage-specific and single-cell chromatin accessibility charts human
844 hematopoiesis and leukemia evolution. *Nat. Genet.* **48**, 1193–1203 (2016).
- 845 28. Yazar, S. *et al.* Single-cell eQTL mapping identifies cell type–specific genetic control of
846 autoimmune disease. *Science* **376**, eabf3041 (2022).
- 847 29. Franzén, O. *et al.* Cardiometabolic risk loci share downstream cis- and trans-gene
848 regulation across tissues and diseases. *Science* **353**, 827–830 (2016).
- 849 30. Gandal, M. J. *et al.* Transcriptome-wide isoform-level dysregulation in ASD, schizophrenia,
850 and bipolar disorder. *Science* **362**, (2018).
- 851 31. Liu, Y. *et al.* ACAT: A Fast and Powerful p Value Combination Method for Rare-Variant
852 Analysis in Sequencing Studies. *Am. J. Hum. Genet.* **104**, 410–421 (2019).
- 853 32. Lambert, S. A. *et al.* The Polygenic Score Catalog as an open database for reproducibility
854 and systematic evaluation. *Nat. Genet.* **53**, 420–425 (2021).
- 855 33. Zhu, Z. *et al.* Causal associations between risk factors and common diseases inferred from
856 GWAS summary data. *Nat. Commun.* **9**, 1–12 (2018).
- 857 34. Liu, M. *et al.* Association studies of up to 1.2 million individuals yield new insights into the
858 genetic etiology of tobacco and alcohol use. *Nat. Genet.* **51**, 237–244 (2019).
- 859 35. Ickick, R. *et al.* Genetic susceptibility to nicotine addiction: Advances and shortcomings in
860 our understanding of the CHRNA5/A3/B4 gene cluster contribution. *Neuropharmacology*
861 **177**, 108234 (2020).
- 862 36. Lofffield, E. *et al.* Mosaic Y Loss Is Moderately Associated with Solid Tumor Risk. *Cancer*
863 *Res.* **79**, 461–466 (2019).
- 864 37. Lofffield, E. *et al.* Predictors of mosaic chromosome Y loss and associations with mortality

- 865 in the UK Biobank. *Sci. Rep.* **8**, 12316 (2018).
- 866 38. Saunders, G. R. B. *et al.* Genetic diversity fuels gene discovery for tobacco and alcohol
867 use. *Nature* **612**, 720–724 (2022).
- 868 39. Kobayashi, T., Hachiya, T., Ikehata, Y. & Horie, S. Genetic association of mosaic loss of
869 chromosome Y with prostate cancer in men of European and East Asian ancestries: a
870 Mendelian randomization study. *Front Aging* **4**, 1176451 (2023).
- 871 40. Loh, P.-R. *et al.* Insights into clonal haematopoiesis from 8,342 mosaic chromosomal
872 alterations. *Nature* **559**, 350–355 (2018).
- 873 41. Taylor, W. R. & Stark, G. R. Regulation of the G2/M transition by p53. *Oncogene* **20**, 1803–
874 1815 (2001).
- 875 42. Porta, C., Paglino, C. & Mosca, A. Targeting PI3K/Akt/mTOR Signaling in Cancer. *Front.*
876 *Oncol.* **4**, 64 (2014).
- 877 43. Lin, S.-H. *et al.* Mosaic chromosome Y loss is associated with alterations in blood cell
878 counts in UK Biobank men. *Sci. Rep.* **10**, 3655 (2020).
- 879 44. Hsu, J. I. *et al.* PPM1D Mutations Drive Clonal Hematopoiesis in Response to Cytotoxic
880 Chemotherapy. *Cell Stem Cell* **23**, 700–713.e6 (2018).
- 881 45. Wong, J. Y. Y. *et al.* Outdoor air pollution and mosaic loss of chromosome Y in older men
882 from the Cardiovascular Health Study. *Environ. Int.* **116**, 239–247 (2018).
- 883 46. Dawoud, A. A. Z., Tapper, W. J. & Cross, N. C. P. Age-related loss of chromosome Y is
884 associated with levels of sex hormone binding globulin and clonal hematopoiesis defined by
885 TET2, TP53, and CBL mutations. *Sci Adv* **9**, eade9746 (2023).
- 886 47. Markozannes, G. *et al.* Systematic review of Mendelian randomization studies on risk of
887 cancer. *BMC Med.* **20**, 1–22 (2022).
- 888 48. Nowak, C. & Ärnlöv, J. A Mendelian randomization study of the effects of blood lipids on
889 breast cancer risk. *Nat. Commun.* **9**, 3957 (2018).
- 890 49. Johnson, K. E. *et al.* The relationship between circulating lipids and breast cancer risk: A

- 891 Mendelian randomization study. *PLoS Med.* **17**, e1003302 (2020).
- 892 50. Orho-Melander, M. *et al.* Blood lipid genetic scores, the HMGCR gene and cancer risk: a
893 Mendelian randomization study. *Int. J. Epidemiol.* **47**, 495–505 (2018).
- 894 51. Hu, Y. *et al.* Minority-centric meta-analyses of blood lipid levels identify novel loci in the
895 Population Architecture using Genomics and Epidemiology (PAGE) study. *PLoS Genet.* **16**,
896 e1008684 (2020).
- 897 52. Richardson, T. G. *et al.* Evaluating the relationship between circulating lipoprotein lipids and
898 apolipoproteins with risk of coronary heart disease: A multivariable Mendelian
899 randomisation analysis. *PLoS Med.* **17**, e1003062 (2020).
- 900 53. Owoicho, O. *et al.* Red blood cell distribution width as a prognostic biomarker for viral
901 infections: prospects and challenges. *Biomark. Med.* **16**, 41–50 (2022).
- 902 54. Faria, S. S. *et al.* The neutrophil-to-lymphocyte ratio: a narrative review.
903 *Ecancermedicalscience* **10**, 702 (2016).
- 904 55. Ethier, J.-L., Desautels, D., Templeton, A., Shah, P. S. & Amir, E. Prognostic role of
905 neutrophil-to-lymphocyte ratio in breast cancer: a systematic review and meta-analysis.
906 *Breast Cancer Res.* **19**, 2 (2017).
- 907 56. Danielsson, M. *et al.* Longitudinal changes in the frequency of mosaic chromosome Y loss
908 in peripheral blood cells of aging men varies profoundly between individuals. *Eur. J. Hum.*
909 *Genet.* **28**, 349–357 (2020).
- 910 57. Lee, W.-H. *et al.* Distinct genetic landscapes and their clinical implications in younger and
911 older patients with myelodysplastic syndromes. *Hematol. Oncol.* (2022)
912 doi:10.1002/hon.3109.
- 913 58. Teichman, R. Exposures of concern to veterans returning from Afghanistan and Iraq. *J.*
914 *Occup. Environ. Med.* **54**, 677–681 (2012).
- 915 59. 1000 Genomes Project Consortium *et al.* A global reference for human genetic variation.
916 *Nature* **526**, 68–74 (2015).

- 917 60. Manichaikul, A. *et al.* Robust relationship inference in genome-wide association studies.
918 *Bioinformatics* **26**, 2867–2873 (2010).
- 919 61. Fang, H. *et al.* Harmonizing Genetic Ancestry and Self-identified Race/Ethnicity in Genome-
920 wide Association Studies. *Am. J. Hum. Genet.* **105**, 763–772 (2019).
- 921 62. Delaneau, O., Zagury, J.-F., Robinson, M. R., Marchini, J. L. & Dermitzakis, E. T. Accurate,
922 scalable and integrative haplotype estimation. *Nat. Commun.* **10**, 5436 (2019).
- 923 63. Gurdasani, D. *et al.* The African Genome Variation Project shapes medical genetics in
924 Africa. *Nature* **517**, 327–332 (2015).
- 925 64. Hinrichs, A. S. *et al.* The UCSC Genome Browser Database: update 2006. *Nucleic Acids*
926 *Res.* **34**, D590–8 (2006).
- 927 65. Danecek, P. *et al.* Twelve years of SAMtools and BCFtools. *Gigascience* **10**, (2021).
- 928 66. Mbatchou, J., Barnard, L., Backman, J. & Marcketta, A. Computationally efficient whole
929 genome regression for quantitative and binary traits. *bioRxiv* (2020).
- 930 67. Yang, J. *et al.* Conditional and joint multiple-SNP analysis of GWAS summary statistics
931 identifies additional variants influencing complex traits. *Nat. Genet.* **44**, 369–75, S1–3
932 (2012).
- 933 68. Wang, G., Sarkar, A., Carbonetto, P. & Stephens, M. A simple new approach to variable
934 selection in regression, with application to genetic fine mapping. *J. R. Stat. Soc. Series B*
935 *Stat. Methodol.* **82**, 1273–1300 (2020).
- 936 69. Benner, C. *et al.* FINEMAP: efficient variable selection using summary data from genome-
937 wide association studies. *Bioinformatics* **32**, 1493–1501 (2016).
- 938 70. Mizrahi Man, O. *et al.* Novel genotyping algorithms for rare variants significantly improve
939 the accuracy of Applied Biosystems™ Axiom™ array genotyping calls. *bioRxiv* (2021)
940 doi:10.1101/2021.09.13.459984.
- 941 71. Cirulli, E. T. *et al.* Genome-wide rare variant analysis for thousands of phenotypes in over
942 70,000 exomes from two cohorts. *Nat. Commun.* **11**, 542 (2020).

- 943 72. Karczewski, K. J. *et al.* Author Correction: The mutational constraint spectrum quantified
944 from variation in 141,456 humans. *Nature* **590**, E53 (2021).
- 945 73. Watanabe, K., Taskesen, E., van Bochoven, A. & Posthuma, D. Functional mapping and
946 annotation of genetic associations with FUMA. *Nat. Commun.* **8**, 1826 (2017).
- 947 74. Mao, X. *et al.* A genomewide admixture mapping panel for Hispanic/Latino populations.
948 *Am. J. Hum. Genet.* **80**, 1171–1178 (2007).
- 949 75. Martin, A. R. *et al.* Human Demographic History Impacts Genetic Risk Prediction across
950 Diverse Populations. *Am. J. Hum. Genet.* **100**, 635–649 (2017).
- 951 76. Maples, B. K., Gravel, S., Kenny, E. E. & Bustamante, C. D. RFMix: a discriminative
952 modeling approach for rapid and robust local-ancestry inference. *Am. J. Hum. Genet.* **93**,
953 278–288 (2013).
- 954 77. Chang, C. C. *et al.* Second-generation PLINK: rising to the challenge of larger and richer
955 datasets. *Gigascience* **4**, 7 (2015).
- 956 78. Liberzon, A. *et al.* The molecular signatures database (MSigDB) hallmark gene set
957 collection. *Cell Syst.* 2015; 1 (6): 417--25.
- 958 79. Finucane, H. K. *et al.* Partitioning heritability by functional annotation using genome-wide
959 association summary statistics. *Nat. Genet.* **47**, 1228–1235 (2015).
- 960 80. Liu, J. Z. *et al.* Association analyses identify 38 susceptibility loci for inflammatory bowel
961 disease and highlight shared genetic risk across populations. *Nat. Genet.* **47**, 979–986
962 (2015).
- 963 81. Okada, Y. *et al.* Genetics of rheumatoid arthritis contributes to biology and drug discovery.
964 *Nature* **506**, 376–381 (2014).
- 965 82. Bentham, J. *et al.* Genetic association analyses implicate aberrant regulation of innate and
966 adaptive immunity genes in the pathogenesis of systemic lupus erythematosus. *Nat. Genet.*
967 **47**, 1457–1464 (2015).
- 968 83. International Multiple Sclerosis Genetics Consortium. Multiple sclerosis genomic map

- 969 implicates peripheral immune cells and microglia in susceptibility. *Science* **365**, (2019).
- 970 84. Kunkle, B. W. *et al.* Genetic meta-analysis of diagnosed Alzheimer's disease identifies new
971 risk loci and implicates A β , tau, immunity and lipid processing. *Nat. Genet.* **51**, 414–430
972 (2019).
- 973 85. Zhang, W. *et al.* Integrative transcriptome imputation reveals tissue-specific and shared
974 biological mechanisms mediating susceptibility to complex traits. *Nat. Commun.* **10**, 3834
975 (2019).
- 976 86. Fullard, J. F. *et al.* Single-nucleus transcriptome analysis of human brain immune response
977 in patients with severe COVID-19. *Genome Med.* **13**, 118 (2021).
- 978 87. GTEx Consortium. The GTEx Consortium atlas of genetic regulatory effects across human
979 tissues. *Science* **369**, 1318–1330 (2020).
- 980 88. Wang, D. *et al.* Comprehensive functional genomic resource and integrative model for the
981 human brain. *Science* **362**, (2018).
- 982 89. Das, S. *et al.* Next-generation genotype imputation service and methods. *Nat. Genet.* **48**,
983 1284–1287 (2016).
- 984 90. McCarthy, S. *et al.* A reference panel of 64,976 haplotypes for genotype imputation. *Nat.*
985 *Genet.* **48**, 1279–1283 (2016).
- 986 91. Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based
987 linkage analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).
- 988 92. Phan, L. *et al.* ALFA: allele frequency aggregator. *National Center for Biotechnology*
989 *Information, US National Library of Medicine* (2020).
- 990 93. Ongen, H., Buil, A., Brown, A. A., Dermitzakis, E. T. & Delaneau, O. Fast and efficient QTL
991 mapper for thousands of molecular phenotypes. *Bioinformatics* **32**, 1479–1485 (2016).
- 992 94. Gamazon, E. R. *et al.* A gene-based association method for mapping traits using reference
993 transcriptome data. *Nat. Genet.* **47**, 1091–1098 (2015).
- 994 95. Barbeira, A. N. *et al.* Exploring the phenotypic consequences of tissue specific gene

- 995 expression variation inferred from GWAS summary statistics. *Nat. Commun.* **9**, 1825
996 (2018).
- 997 96. Zhu, Z. *et al.* Integration of summary data from GWAS and eQTL studies predicts complex
998 trait gene targets. *Nat. Genet.* **48**, 481–487 (2016).
- 999 97. Bulik-Sullivan, B. K. *et al.* LD Score regression distinguishes confounding from polygenicity
1000 in genome-wide association studies. *Nat. Genet.* **47**, 291–295 (2015).
- 1001 98. Zheng, J. *et al.* LD Hub: a centralized database and web interface to perform LD score
1002 regression that maximizes the potential of summary level GWAS data for SNP heritability
1003 and genetic correlation analysis. *Bioinformatics* **33**, 272–279 (2017).
- 1004 99. Zhou, W. *et al.* Efficiently controlling for case-control imbalance and sample relatedness in
1005 large-scale genetic association studies. *Nat. Genet.* **50**, 1335–1341 (2018).
- 1006 100. Ge, T., Chen, C.-Y., Ni, Y., Feng, Y.-C. A. & Smoller, J. W. Polygenic prediction via
1007 Bayesian regression and continuous shrinkage priors. *Nat. Commun.* **10**, 1–10 (2019).
- 1008 101. Elsworth, B. *et al.* The MRC IEU OpenGWAS data infrastructure. *bioRxiv*
1009 2020.08.10.244293 (2020) doi:10.1101/2020.08.10.244293.
- 1010 102. Pulit, S. L. *et al.* Meta-analysis of genome-wide association studies for body fat distribution
1011 in 694,649 individuals of European ancestry. Preprint at <https://doi.org/10.1101/304030>.
- 1012 103. Hemani, G. *et al.* The MR-Base platform supports systematic causal inference across the
1013 human phenome. *Elife* **7**, e34408 (2018).
- 1014 104. Broadbent, J. R. *et al.* MendelianRandomization v0.5.0: updates to an R package for
1015 performing Mendelian randomization analyses using summarized data. *Wellcome Open*
1016 *Research* **5**, (2020).
- 1017 105. Verbanck, M., Chen, C.-Y., Neale, B. & Do, R. Detection of widespread horizontal
1018 pleiotropy in causal relationships inferred from Mendelian randomization between complex
1019 traits and diseases. *Nat. Genet.* **50**, 693–698 (2018).
- 1020 106. Sanderson, E., Spiller, W. & Bowden, J. Testing and correcting for weak and pleiotropic

- 1021 instruments in two-sample multivariable Mendelian randomization. *Stat. Med.* **40**, 5434–
- 1022 5452 (2021).
- 1023

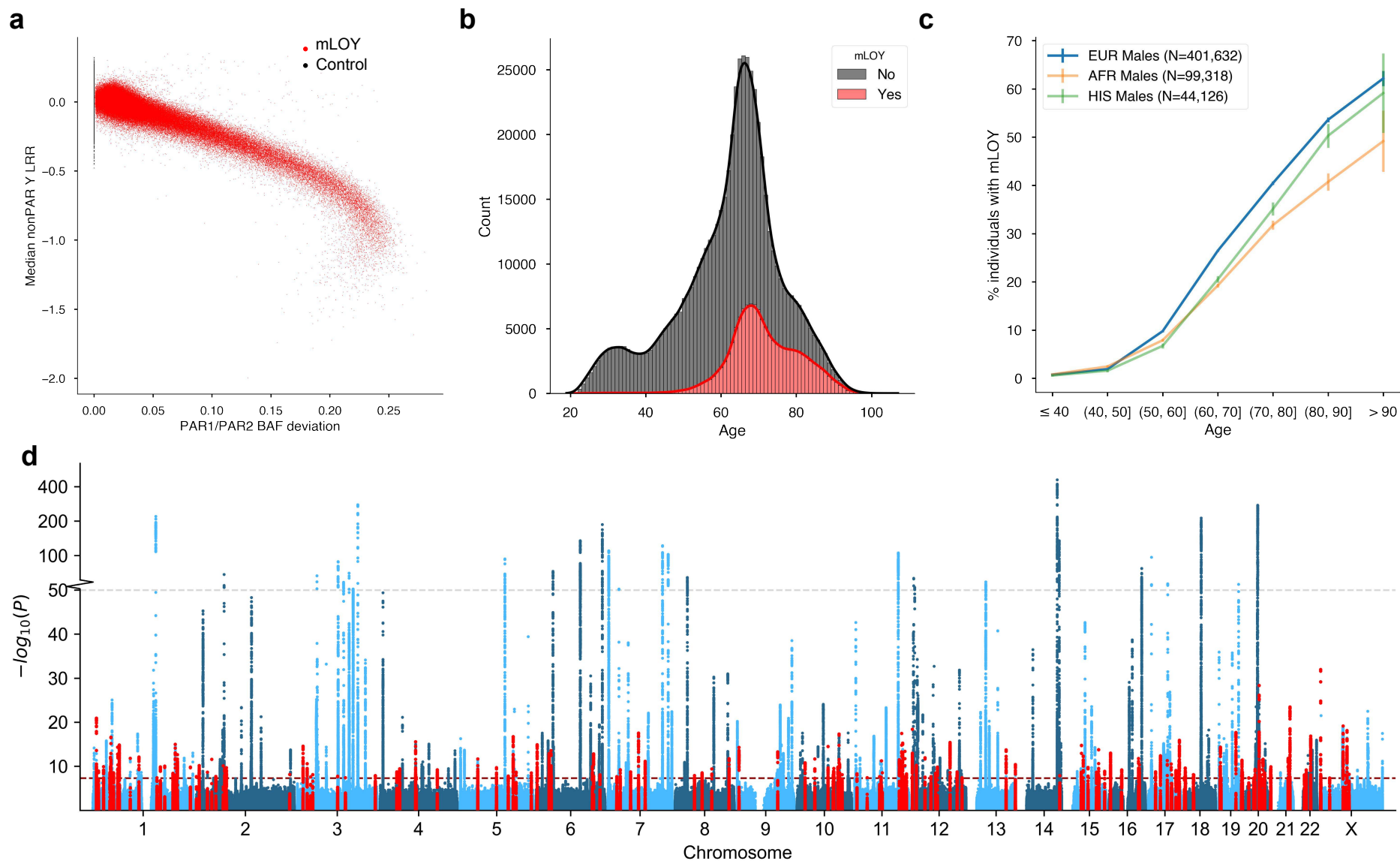


Fig. 1. Mosaic loss of Y chromosome (mLOY) in the Million Veteran Program (MVP). **a**, Median genotyping probe intensity log R ratio (LRR) vs. phased B Allele Frequency (BAF) in the pseudo-autosomal regions (PAR) 1 and 2. **b**, Density of age distribution in all MVP mLOY cases and controls. **c**, Percentage of individuals with mLOY per ten-year age bin for MVP European (EUR), African (AFR), and Hispanic (HIS) cohorts. Error bars represent 95% confidence intervals. **d**, Manhattan plot shows the $-\log_{10}(P)$ for associations of genetic variants with mLOY in the multi-ancestry meta-analysis. Novel mLOY index variants and variants within ± 50 Kb are highlighted in red. The red line indicates the genome-wide significance threshold ($P < 5 \times 10^{-8}$). The grey dotted line represents a transition from linear to log-scale on the y-axis.

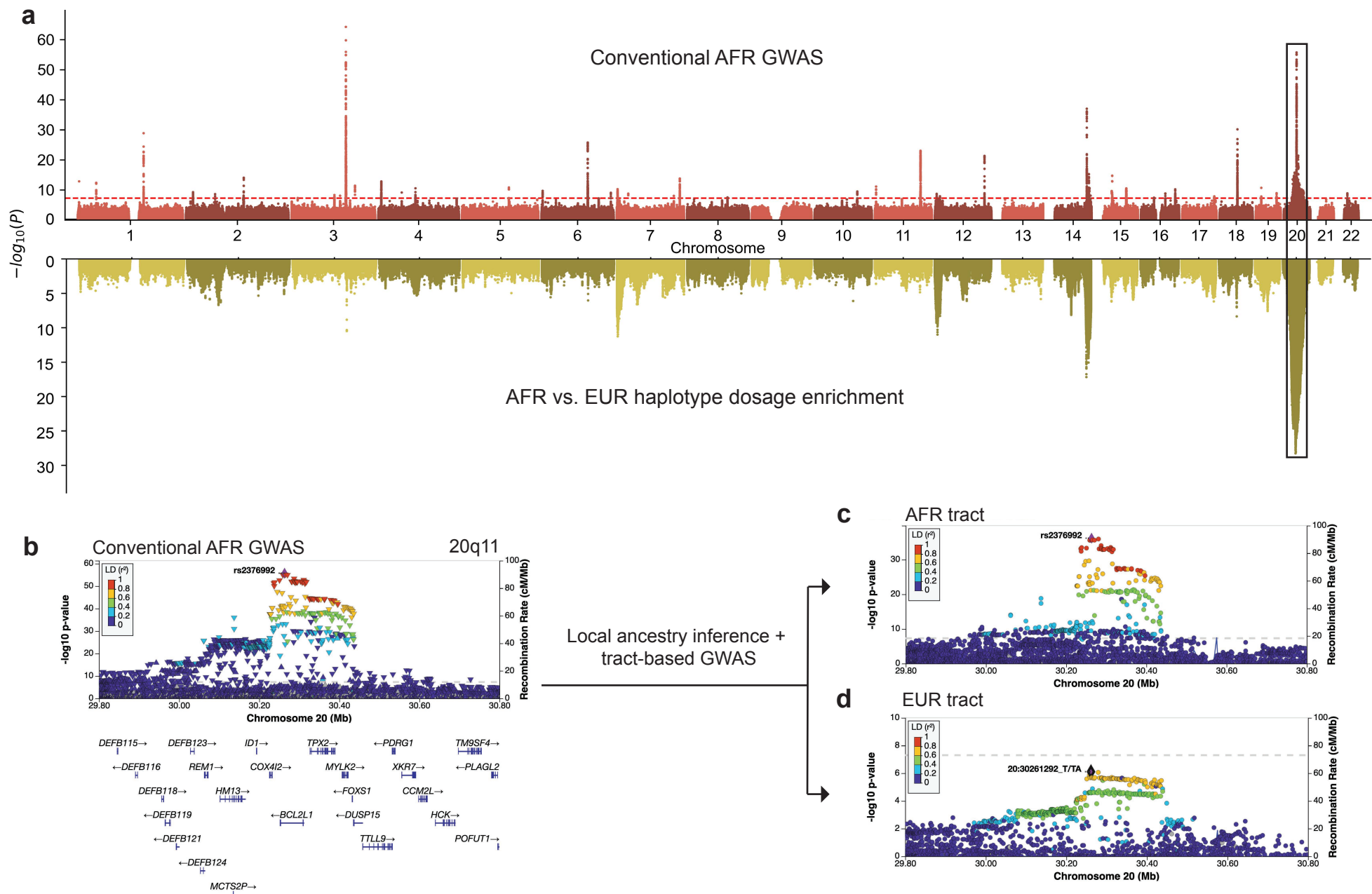


Fig. 2. Global- and local-ancestry-adjusted GWAS of mosaic loss of Y in admixed African (AFR) ancestry Million Veteran Program participants. a, Manhattan plot showing conventional AFR GWAS (top) and AFR vs. European (EUR) haplotype dosage enrichment (bottom). **b**, At 20q11 (*BCL2L1*) we observed inflation due to local ancestry. This inflation was resolved by separating the AFR and EUR tracts as shown in **c** and **d**.

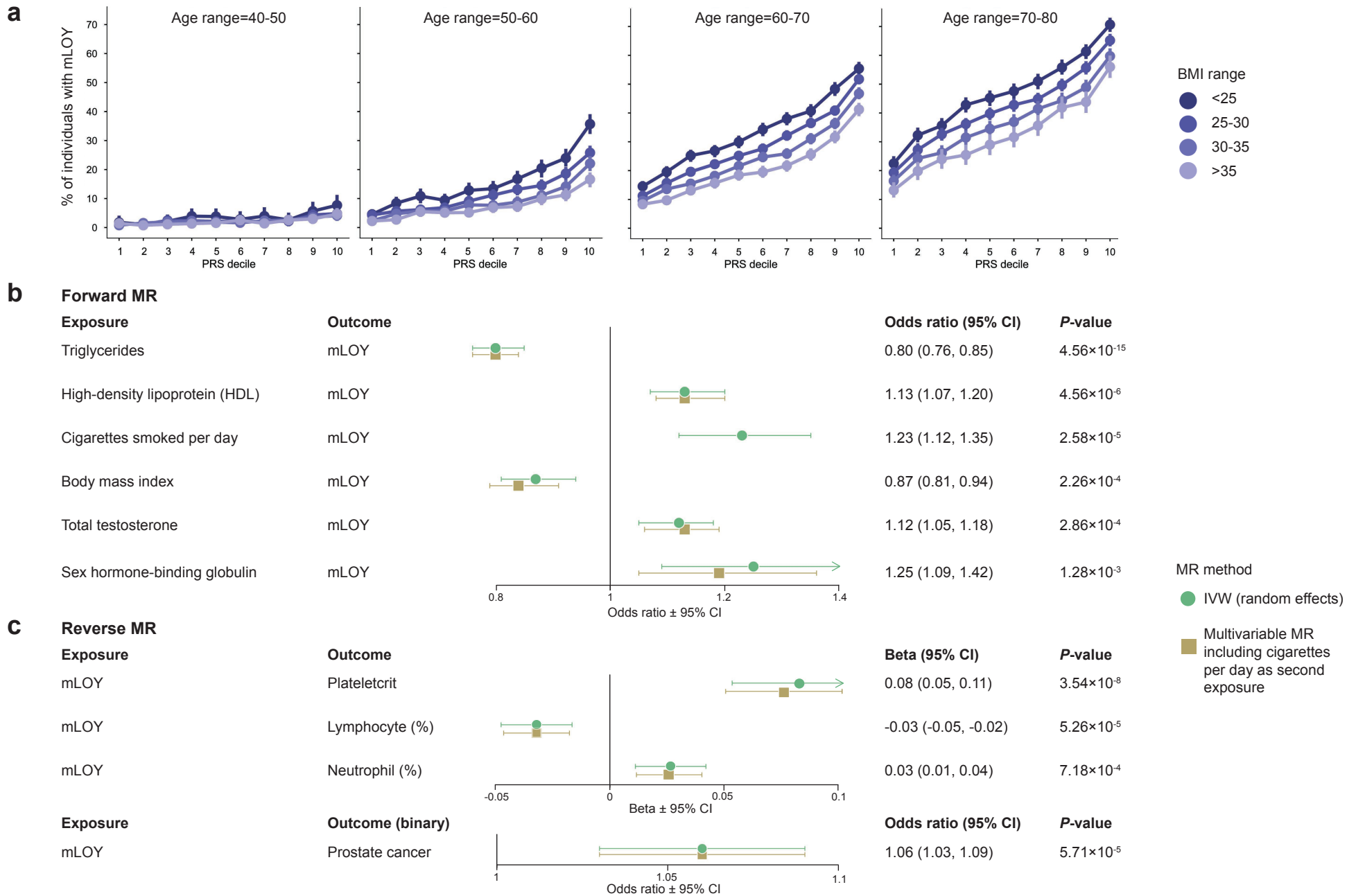


Fig. 3. Mosaic loss of Y by PRS decile and Mendelian randomization (MR). **a**, Percentage of individuals with mLOY by PRS decile, stratified by age decile (plot grid) and BMI range (color). **b**, Forest plot showing exposure traits with significant results from random effects inverse-variance weighted (RE IVW) MR and corresponding multivariable MR (MVMR) with cigarettes per day as an additional exposure, on mLOY outcome in Million Veteran Program Europeans. **c**, Significant results from RE IVW MR and MVMR considering mLOY exposure on trait outcomes. CI, confidence interval.