medRxiv preprint doi: https://doi.org/10.1101/2024.04.24.24306301; this version posted April 25, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. This article is a US Government work. It is not subject to copyright under 17 USC 105 and is also made available for use under a CC0 license.

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 Bendl[9], Biao Zeng[9], Sanan Venkatesh[8,9], Chris Chatzinakos[4], Erin McAuley[1,2], Sun-Gou Ji[1,33], Kyriacos Markianos[1], Patrick A. Schreiner[1,2], Elizabeth Partan[10], Yunling Shi[11], Poornima Devineni[11], VA Million Veteran Program*, Jennifer Moser[12], Sumitra Muralidhar[12], Rachel Ramoni[12], Alexander G. Bick[13], 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‡]

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

NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.

medRxiv preprint doi: https://doi.org/10.1101/2024.04.24.24306301; this version posted April 25, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. This article is a US Government work. It is not subject to copyright under 17 USC 105 and is also made available for use under a CC0 license.

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

[†] Denotes equal contribution

[‡] Saiju Pyarajan: saiju.pyarajan@va.gov

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

Abstract

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

medRxiv preprint doi: https://doi.org/10.1101/2024.04.24.24306301; this version posted April 25, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. This article is a US Government work. It is not subject to copyright under 17 USC 105 and is also made available for use under a CC0 license.

Introduction

Mosaic loss of Y chromosome (mLOY) is the most common type of mosaic chromosomal alteration (mCA), observable in upwards of 40% of males above age 70[1,2], and 70% of males over 85[3]. mLOY is the most readily detected mCA in leukocytes, which are the primary source of blood-derived DNA. Adaptive immunity causes high turnover rates in the hematopoietic stem cell compartment, enabling clonal expansion of mosaic cell subpopulations with selective advantages[4]. Chromosomal aberrations in these rapidly expanding cell subpopulations can produce mLOY, though it is unclear if mLOY itself confers selective advantages.

The gene-poor and repetitive-element-rich composition of the Y chromosome initially led researchers to believe its role was restricted to spermatogenesis and sex determination[5]. Congruently, mLOY was considered a benign condition, and a consequence of the broader genomic instability that occurs with aging[6]. But in the past decade, epidemiological studies have highlighted associations between mLOY and a broad range of health outcomes; these include all-cause mortality, hematological malignancies and other types of cancer, Alzheimer's disease, type 2 diabetes (T2D), obesity, and cardiovascular disease (CVD)[7]. However, it has yet to be determined whether mLOY is a driving causal factor in these conditions, a passenger (i.e. a consequence), or a symptom of a shared, underlying cause, such as genetic susceptibility to DNA replication errors, or exogenous exposure to mutagens (particularly via smoking cigarettes)[7,8]. There have also been inconsistent associations with the comorbidities observed across studies, although this may be related to differences in sample collection and mLOY classification methods (e.g. low versus high cell fraction detection)[2].

medRxiv preprint doi: https://doi.org/10.1101/2024.04.24.24306301; this version posted April 25, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. This article is a US Government work. It is not subject to copyright under 17 USC 105 and is also made available for use under a CC0 license.

The mechanisms of mLOY in producing disease phenotypes are gradually being elucidated. Many genetic risk loci for clonal hematopoiesis of indeterminate potential (CHIP)[9,10] have also been identified as mLOY risk loci[11], and these two types of mosaicism can co-occur (even in men without observable hematological disease[12]). However, the cellular and epidemiological outcomes of CHIP and mLOY appear to be distinct. For example, a study which induced *TET2*-associated CHIP in mice suggested a mechanistic link to CVD based on expression of inflammatory chemokines in macrophages[13], while a mouse model of mLOY demonstrated a mechanism of producing CVD through fibrotic deposition in the extracellular matrix[14]. Additionally, deletion of chromosome Y by CRISPR–Cas9 produced more aggressive bladder cancer tumors by means of T-cell exhaustion[15]. It has also been suggested that there are bi-directional relationships between mLOY and transcriptional dysregulation that lead to differences in disease phenotypes that are dependent on mLOY cell lineage[8].

Estimates of single nucleotide polymorphism-based heritability (SNP-h^2) for mLOY detected via the pseudo-autosomal region 1 (PAR1) are as high as 31.7%, highlighting mLOY as substantially heritable compared to most human traits[1,16]. Germline genetics govern many processes which can lead to somatic mCA acquisition and clonal expansion, particularly via cell-cycle regulation, DNA damage response, apoptosis, and susceptibility to cancer. Advances in integrating long-range phasing with genotype have enabled sensitive and accurate identification of mLOY in large-scale cohorts[1]. A genome-wide association study (GWAS) of mLOY status, performed in European ancestry (EUR) UK Biobank (UKB) participants[1], replicated all 19 previously reported genetic risk loci[17], including the oncogene *TCL1A*[18], and identified 137 novel

medRxiv preprint doi: https://doi.org/10.1101/2024.04.24.24306301; this version posted April 25, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. This article is a US Government work. It is not subject to copyright under 17 USC 105 and is also made available for use under a CC0 license.

loci. A GWAS of mLOY in Biobank Japan (BBJ) using mLRR-Y intensity as a quantitative trait measure identified 46 loci, 35 of which were novel[19].

In this study we analyzed 544,112 male participants in the Million Veteran Program (MVP), a biobank in the Department of Veterans Affairs (VA) healthcare system which combines ancestrally diverse genetic data with extensive electronic health records[20]. In addition to a large EUR cohort, we present the first GWAS of mLOY status in African (AFR) and Hispanic/Latino (HIS) ancestries. MVP provides a uniquely valuable resource to perform multi-ancestry mLOY analyses, as it avoids technical issues related to genotyping and mosaicism calling that may be introduced by combining data from separate biobanks. Our results highlight the benefits of inclusive population studies in advancing our understanding of mLOY.

medRxiv preprint doi: https://doi.org/10.1101/2024.04.24.24306301; this version posted April 25, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. This article is a US Government work. It is not subject to copyright under 17 USC 105 and is also made available for use under a CC0 license.

Results

mLOY phenotyping and participant characteristics

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

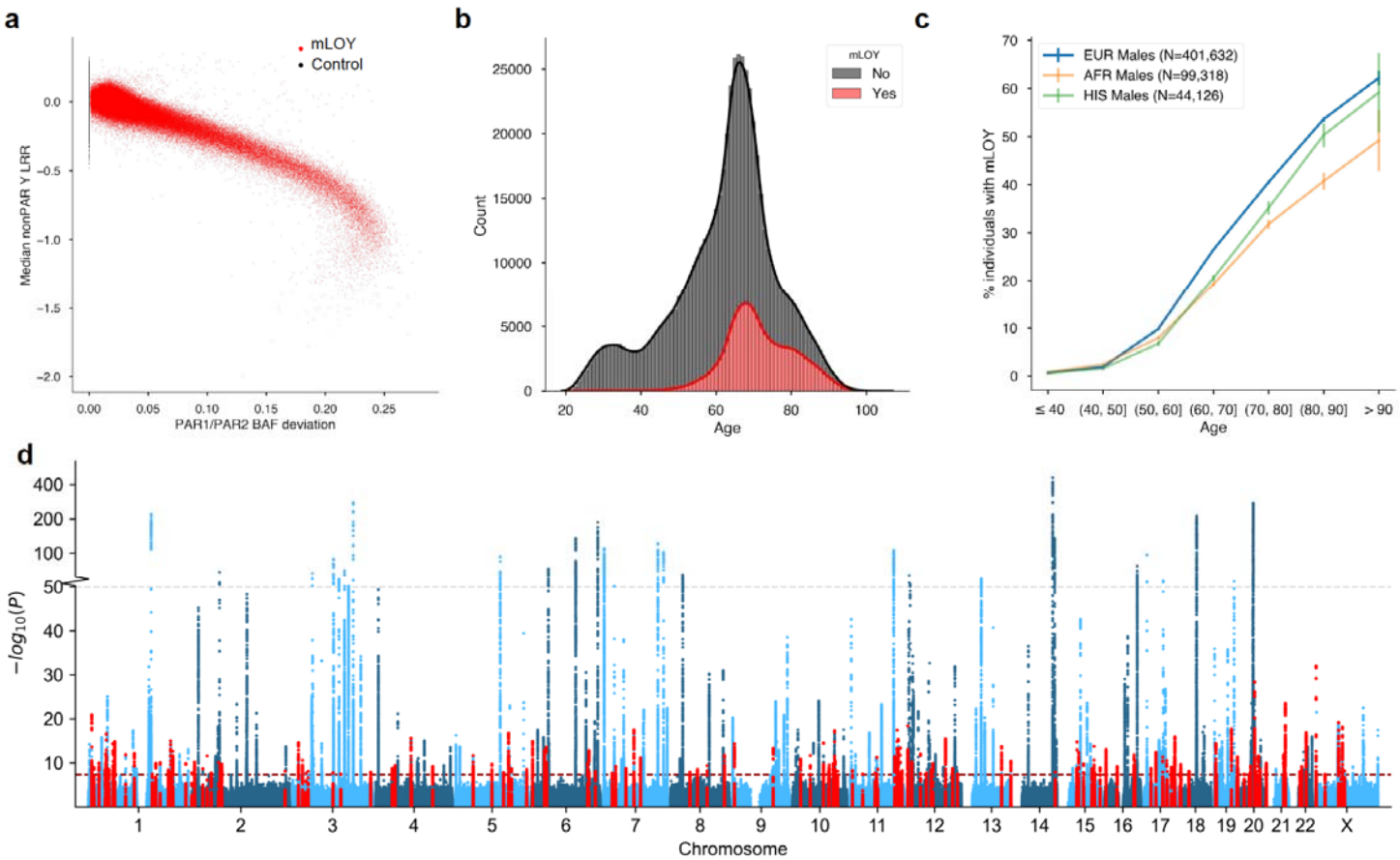

Fig. 1. mLOY in the Million Veteran Program. a, Median genotyping probe intensity log R ratio (LRR) vs. phased B Allele Frequency (BAF) in the pseudo-autosomal regions (PAR) 1 and 2. **b**, Percentage of individuals with mLOY per ten-year age bin for MVP European (EUR), African (AFR), and Hispanic (HIS) cohorts. Error bars represent 95% confidence intervals. **c**, Density of age distribution in all MVP mLOY cases and controls. **d**, Manhattan plots show the -log10(*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 ($P<5\times10^{-8}$). The grey dotted line represents a transition from linear to log-scale on the y-axis.

medRxiv preprint doi: https://doi.org/10.1101/2024.04.24.24306301; this version posted April 25, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. This article is a US Government work. It is not subject to copyright under 17 USC 105 and is also made available for use under a CC0 license.

medRxiv preprint doi: https://doi.org/10.1101/2024.04.24.24306301; this version posted April 25, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. This article is a US Government work. It is not subject to copyright under 17 USC 105 and is also made available for use under a CC0 license.

Genome-wide significant associations

We performed a GWAS for mLOY case-control status in each ancestry group (Supplementary Data 2). We identified 336 conditionally independent genome-wide significant (GWS; $P<5\times10^{-8}$) signals in 203 distinct loci for EUR, 50 signals in 46 loci for AFR, and 17 signals in 15 loci for HIS (Supplementary Fig. 3a-c, Supplementary Data 3-5). Of the EUR signals, 220 were within 1Mb of a previously reported mLOY index variant[17,19], including 148 of the 156 previously identified in UKB[1], and 116 were novel. In a variant-level replication of 327 EUR signals that were available in UKB[1], all but one (rs925301) had the same effect direction (Supplementary Data 3, Supplementary Fig. 4a). Of the 116 novel MVP signals, 17 had $P<1\times10^{-5}$ and 97 had $P<0.05$ in UKB[1].

The most significant association signals in MVP EUR and HIS were in *TCL1A*; in EUR, we identified the same *TCL1A* lead variant as in previous GWAS[1,18] (rs2887399; OR=0.708 [0.697, 0.719]; $P=3.18\times10^{-419}$; Supplementary Fig. 5a). The strongest effect of an allele in EUR was at *TP53*, an oncogene associated with mCAs and CHIP (rs78378222; OR=1.664 [1.584, 1.749]; $P=1.44\times10^{-90}$). In AFR, the most significant association was in *RPN1* at rs113336380, which has a minor allele frequency (MAF) of ~6% in AFR and <0.01% in EUR (Supplementary Fig. 5b). Effect directions were largely concordant between EUR and AFR associations ($r=0.656$, $P=1.13\times10^{-46}$), with the exception of a group of novel EUR variants that were not significant in AFR, and rs6018599 (*GGT5*/*CABIN1*), which was GWS only in AFR (Supplementary Fig. 4b). Four additional GWS novel signals were specific to AFR: *MPL*, *NKX2-3*, *ETV6*, and *BLCAP* (Supplementary Fig. 6; Supplementary Data 4). All 17 HIS signals were GWS in EUR (Supplementary Data 5); 15 of these were GWS in AFR, and all HIS signals were

medRxiv preprint doi: https://doi.org/10.1101/2024.04.24.24306301; this version posted April 25, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. This article is a US Government work. It is not subject to copyright under 17 USC 105 and is also made available for use under a CC0 license.

GWS in UKB[1]. Effect direction at significant HIS signals between HIS and EUR were consistent (Supplementary Fig. 4c). We improved our variant selection by fine-mapping and estimating credible sets of candidate causal variants in EUR and AFR. In EUR, we found 11,242 variants in 334 high-quality credible sets, with a median of 8.5 variants per credible set (Supplementary Data 6). In AFR, we found 533 variants in 45 high-quality credible sets, with a median of 5 variants per credible set (Supplementary Data 7).

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

Fixed effects (FE) multi-ancestry meta-analysis of the three MVP ancestry groups identified 298 GWS loci, including 157 novel loci, 42 of which did not reach GWS in any individual ancestry (Fig. 1d; Supplementary Fig. 3d; Supplementary Data 9). After meta-analysis there remained 8 EUR loci and 2 AFR loci that were GWS only within their respective ancestries. Of the 42 added meta-analysis novel lead variants, 32 had $P<0.05$ in UKB[1] and two had $P<1\times10^{-5}$. For all 298 meta-analysis lead variants, 211 were available in the BBJ mLOY GWAS[19]; though BBJ used a less sensitive quantitative mLOY measure (logarithm of R ratio) as opposed to our case-control designation, the effect of 188 of the 211 variants were aligned, including all 88 BBJ

medRxiv preprint doi: https://doi.org/10.1101/2024.04.24.24306301; this version posted April 25, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. This article is a US Government work. It is not subject to copyright under 17 USC 105 and is also made available for use under a CC0 license.

variants with $P<0.01$ (Supplementary Fig. 4d). BBJ[19] shared 28 GWS variants with our meta-analysis (Supplementary Data 9). Within MVP meta-analysis index variants, rs2887399 (*TCL1A*) remained the most significant association ($P=5.25\times10^{-459}$). We also performed random effects (RE) meta-analyses using the Han-Eskin method (RE2)[24], and observed similar P-values to FE (Supplementary Data 9).

Tract-based association analysis for AFR

We used local ancestry inference and performed a GWAS using the Tractor method[25] as a secondary analysis to resolve ancestry-specific signals in AFR (Supplementary Fig. 8). Signals resulting from recent admixture with large allele frequency differences between ancestries can be disentangled with this method; we highlight two AFR loci to illustrate. First, inflation due to admixture was observed at 20q11.21 near *BCL2L1* (Fig. 2a). Because global adjustment for ancestry by including 20 principal components (PCs) in this GWAS model did not sufficiently resolve this locus (Fig. 2b), we inferred that the inflation was due to the differences in risk allele frequency at the lead SNP rs2376992 across ancestral haplotypes (51% AFR vs 22% EUR, Supplementary Fig. 9a) at this large effect size locus (OR=0.798 [0.776, 0.820]; $P=1.7\times10^{-56}$).

This was further supported by 20q11.21 having a highly significant difference in AFR and EUR haplotype dosages ($P=5.6\times10^{-29}$; Fig 2a). The secondary cluster of significant variants downstream of *BCL2L1* was resolved in the AFR tract relative to the EUR tract (Fig. 2cd; Supplementary Fig. 9b). Strikingly, a meta-analysis of the AFR and EUR tracts for chromosome 20 yielded a genomic control (λ) of 1.08, as compared to $\lambda=1.40$ in the conventional AFR GWAS. We then fine-mapped the smaller credible set

medRxiv preprint doi: https://doi.org/10.1101/2024.04.24.24306301; this version posted April 25, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. This article is a US Government work. It is not subject to copyright under 17 USC 105 and is also made available for use under a CC0 license.

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

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

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

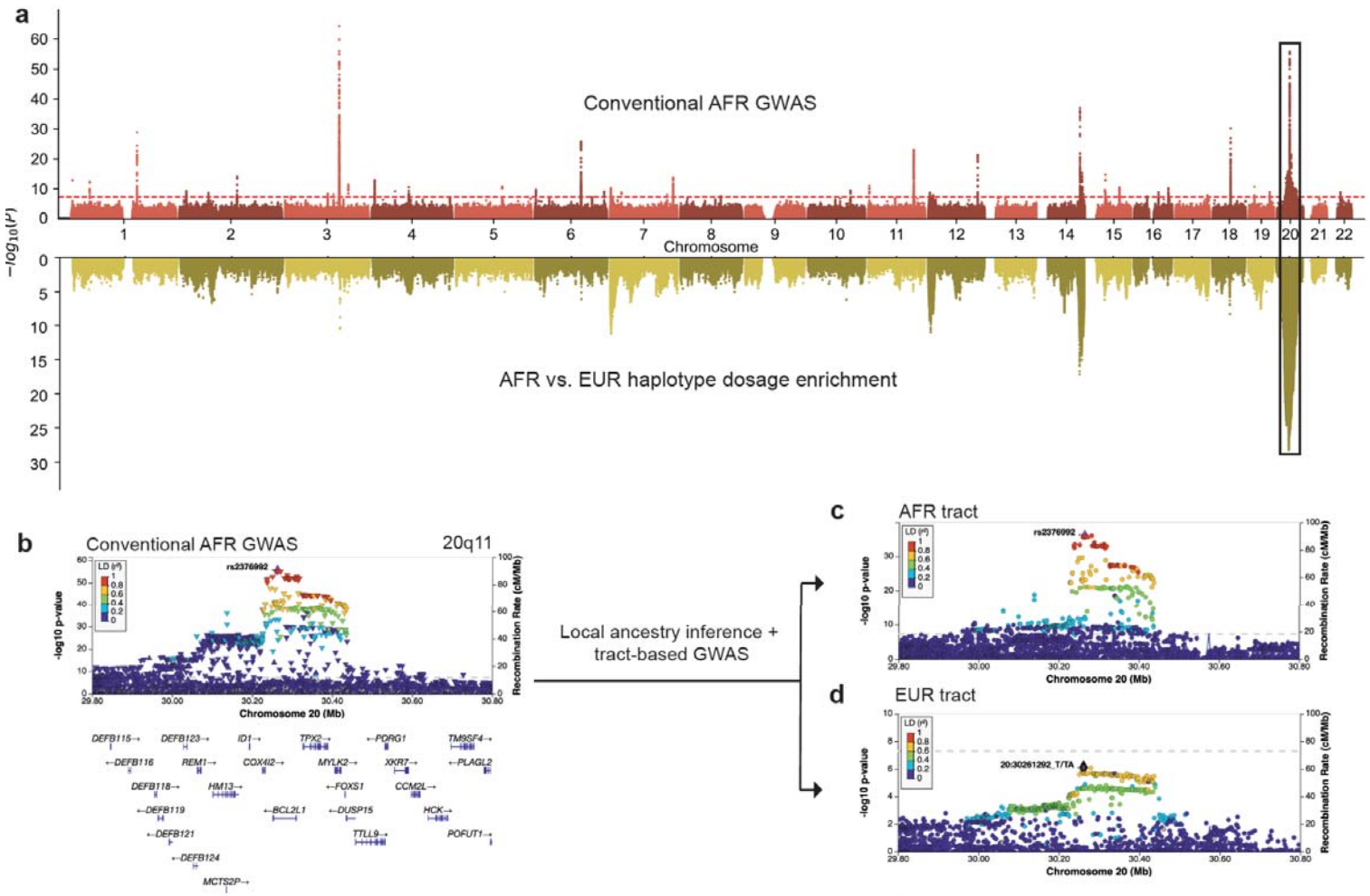

Fig. 2. Global- vs. local-ancestry-adjusted GWAS of mosaic loss of Y in admixed African (AFR) ancestry Million Veteran Program participants. a, Miami plot showing conventional AFR GWAS globally adjusted by principal components (top) and AFR vs. European (EUR) haplotype dosage enrichment (bottom). **b**, 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**.

medRxiv preprint doi: https://doi.org/10.1101/2024.04.24.24306301; this version posted April 25, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. This article is a US Government work. It is not subject to copyright under 17 USC 105 and is also made available for use under a CC0 license.

medRxiv preprint doi: https://doi.org/10.1101/2024.04.24.24306301; this version posted April 25, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. This article is a US Government work. It is not subject to copyright under 17 USC 105 and is also made available for use under a CC0 license.

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

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

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

medRxiv preprint doi: https://doi.org/10.1101/2024.04.24.24306301; this version posted April 25, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. This article is a US Government work. It is not subject to copyright under 17 USC 105 and is also made available for use under a CC0 license.

blood cell differentiation[19]. Interestingly, among the six differentiated cell types encompassing myeloid, erythroid, and lymphoid cells (Supplementary Fig. 13b), only erythroid cells exhibited significant enrichment (adj*P*=0.017). This enrichment pattern of mLOY-related effects on differentiating blood cells contrasts starkly with other diseases characterized by perturbations in immune responses, chronic inflammation, or autoimmune mechanisms, such as Crohn's disease, rheumatoid arthritis, systemic lupus erythematosus, multiple sclerosis or Alzheimer's disease (Supplementary Fig. 13c).

eQTLs from single-cell RNA-seq

To identify how SNPs associated with mLOY in EUR associate with gene expression in immune cells, we used a recently published expression quantitative trait loci (eQTL) dataset derived from single-cell RNA-seq data across different immune cell populations[28]. Across the fourteen immune cell subsets, we found that 197 of 327 EUR mLOY SNPs spanning 251 eQTLs reached at least nominal significance ($P<0.05$; Supplementary Data 12). Of these, 34 eQTLs (22 unique genes; 8 unique SNPs) reached the significance threshold corrected for multiple testing ($P<1.1\times10^{-5}$); 20 of these eQTLs were associated with two SNPs in the major histocompatibility complex (MHC) region (6p22.1 to 6p21.3).

Multi-tissue TWAS and SMR

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

medRxiv preprint doi: https://doi.org/10.1101/2024.04.24.24306301; this version posted April 25, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. This article is a US Government work. It is not subject to copyright under 17 USC 105 and is also made available for use under a CC0 license.

including STARNET Blood[29] and a high-powered dorsolateral prefrontal cortex (DLPFC) dataset[30] (Supplementary Data 13); this yielded 2,297 unique significant gene features at $P_{Bonferroni}$<0.05. In the STARNET blood model, 117 features were significant at $P_{Bonferroni}$<0.05, including one novel gene that did not appear in GWAS, *MED19*, a component of the Mediator complex involved in the regulated transcription of RNA polymerase II-dependent genes (Supplementary Data 14, Supplementary Fig. 14a). In the DLPFC model, 191 features were significant at $P_{Bonferroni}$<0.05; the novel genes identified in DLPFC were *IL21R*, (cytokine receptor for interleukin 21) and *COX7A2L* (Supplementary Data 15). All tissues were then meta-analyzed using ACAT[31], yielding a total of 683 genes with $P_{Bonferroni}$<0.05 (Supplementary Data 16, Supplementary Fig. 14b). ACAT revealed an additional five novel genes: *PSTPIP2*, *CCNK*, *RAD54L2*, *PARP10*, and *G3BP1*, plus the non-coding gene *LINC01933* and pseudogene *AC091982.1.* We found that mLOY-associated gene expression was highly correlated across the imputed transcriptomes of all tissues (Supplementary Fig. 15).

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

medRxiv preprint doi: https://doi.org/10.1101/2024.04.24.24306301; this version posted April 25, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. This article is a US Government work. It is not subject to copyright under 17 USC 105 and is also made available for use under a CC0 license.

Genetic correlations of mLOY

Genetic correlations (r_g) between mLOY in EUR and 750 traits were tested (Supplementary Data 18). At the multiple testing corrected significance threshold $P<6.67\times10^{-5}$, 36 traits were significantly correlated. Many metabolite measures had a significant negative r_g with mLOY at this threshold, including particle sizes of cholesterol, phospholipids, triglycerides, and total lipids in VLDL, which had r_g ranging from -0.34 (se = 0.07) to -0.28(0.06). Maternal smoking around birth was positively correlated at r_g = 0.12 (0.03). Significant and negative r_g were observed for hip circumference and the related anthropometric measures obesity class 2, and BMI.

Association of PGS Catalog-based polygenic scores with mLOY

We calculated polygenic scores (PGS) for every trait in the PGS Catalog[32], for all EUR subjects, and performed a phenome-wide scan for the association of normalized PGS scores with mLOY case-control status (PGS-WAS) to discover shared genetic etiology (Supplementary Data 19). A total of 2,644 scores corresponding to 562 uniquely mapped traits in the Experimental Factor Ontology (EFO) were tested; we found 82 mapped traits significant after multiple testing correction (α=0.05/562). Many of the most significant traits were blood measures, including increased platelet crit (1.11 [1.10,1.12]), leukocyte count (1.07 [1.06,1.08]), monocyte count (1.07 [1.06,1.08]), and neutrophil count (1.06 [1.05,1.07]), and decreased mean corpuscular hemoglobin concentration (0.94 [0.93,0.95]). Several metabolic traits were also associated, such as triglycerides (0.925 [0.914,0.935]), HDL (1.06 [1.04,1.07]), BMI (0.945 [0.94,0.96]), SHBG (1.05 [1.04,1.06]), T2D (0.93 [0.91,0.94]), as well as smoking status (1.05 [1.04,1.06]).

medRxiv preprint doi: https://doi.org/10.1101/2024.04.24.24306301; this version posted April 25, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. This article is a US Government work. It is not subject to copyright under 17 USC 105 and is also made available for use under a CC0 license.

Multi-trait conditioning of mLOY on cigarettes per day

In MVP participants we observed a strong observational association of mLOY with smoking, as well as in the PGS-WAS. To parse out the residual effects of smoking on mLOY susceptibility, we conducted multi-trait-based conditional and joint association analysis[33] (mtCOJO) in EUR, conditioning on cigarettes per day[34]. Across the genome, only the 15q25 region was primarily impacted by the conditional analysis (Supplementary Fig. 17a). This signal, which was GWS for mLOY in the primary GWAS, was attenuated towards the null after conditioning (Supplementary Fig. 17bc). This locus contains a well-known cluster of smoking-related genes, including *CHRNA5* which encodes a nicotinic acetylcholine receptor subunit that has been frequently associated with smoking in GWAS[35].

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

Because we observed a significant negative genetic correlation and PGS-WAS association between mLOY and BMI, an association that has also been reported in previous studies[36,37], we sought to examine the prevalence of mLOY in MVP-EUR as a function of age, BMI, and PRS decile derived from UKB[1]. In all age bins, we observed that decreasing BMI and increasing PRS were associated with higher mLOY prevalence (Fig. 3a). Overall, our results suggest that this association between mLOY and BMI is not merely confounded by the strong age-dependence of mLOY. We also observed a stronger genetic correlation between mLOY and BMI in MVP (r_g=-0.110 (0.026)) than previously reported in UKB (r_g=-0.052 (0.032); Supplementary Data 18); this may be a result of greater co-morbidities in the MVP cohort versus UKB.

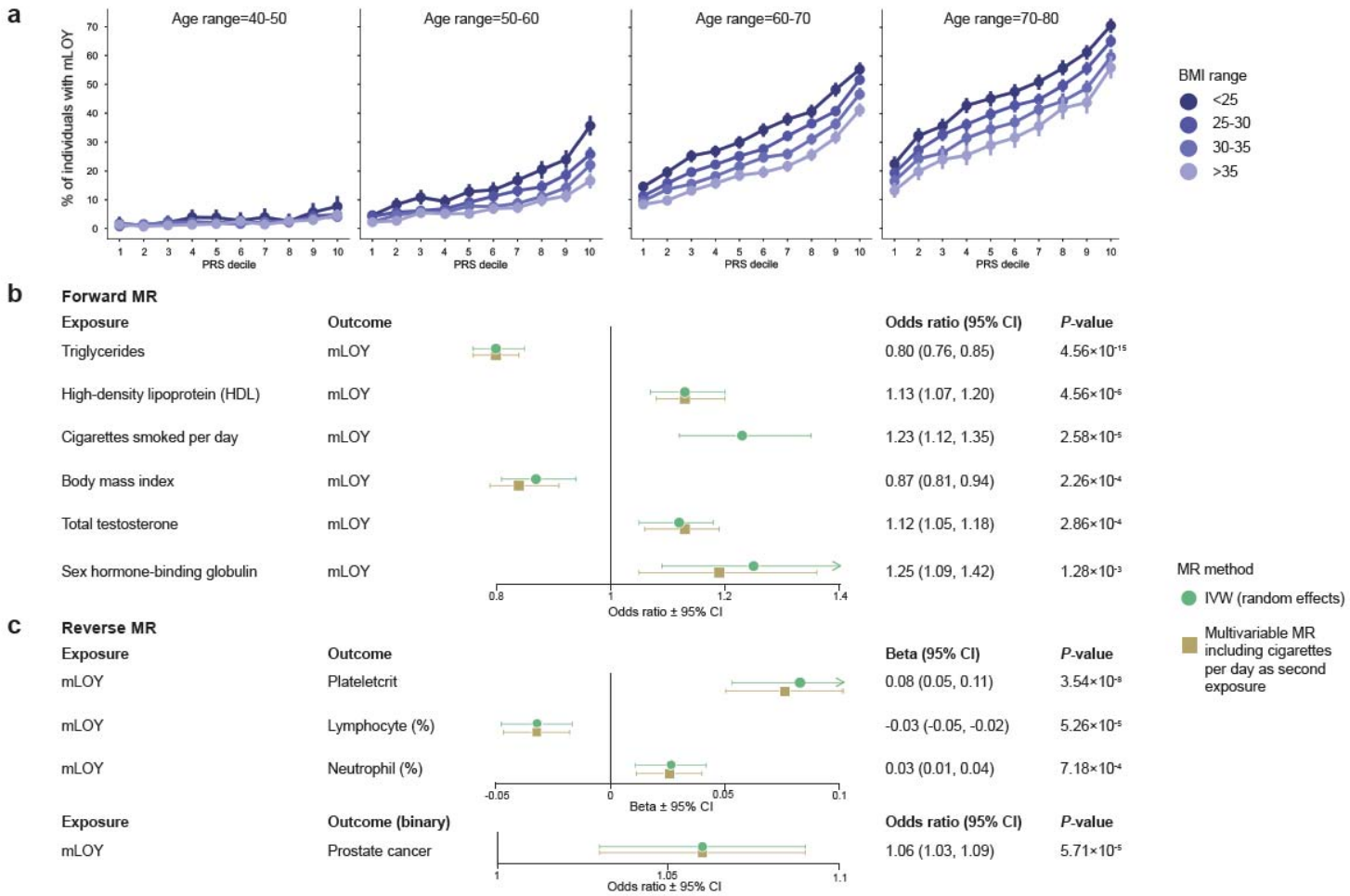

Fig. 3. Mosaic loss of Y by PRS decile and Mendelian randomization (MR). a, Percentage of individuals with mLOY by PRS decile, stratified by age decile (plot grid) and BMI range (color). **b**, Forest plot showing exposure traits with significant results from random effects inverse-variance weighted (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.

medRxiv preprint doi: https://doi.org/10.1101/2024.04.24.24306301; this version posted April 25, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. This article is a US Government work. It is not subject to copyright under 17 USC 105 and is also made available for use under a CC0 license.

medRxiv preprint doi: https://doi.org/10.1101/2024.04.24.24306301; this version posted April 25, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. This article is a US Government work. It is not subject to copyright under 17 USC 105 and is also made available for use under a CC0 license.

Univariable and multivariable Mendelian randomization

We used the significant PGS Catalog trait associations with mLOY to inform the selection of 32 traits for MR in the EUR cohort, using male-only summary statistics where available, to infer the direction of causality between these exposures and mLOY. Random effects inverse-variance weighted (IVW) forward MR supported six significant traits (α=0.05/32) with non-significant MR-Egger intercept ($P \geq 0.05$) as causal influences on mLOY: triglycerides, high-density lipoprotein (HDL), cigarettes per day[38], body mass index (BMI), total testosterone, and sex hormone-binding globulin (SHBG) (Fig. 3b; Supplementary Fig. 18; Supplementary Data 20). Reverse MR indicated a significant causal influence of mLOY on plateletcrit, lymphocyte percentage, prostate cancer, and neutrophil percentage (Fig. 3c; Supplementary Fig. 19; Supplementary Data 21). Our finding that mLOY status increases the risk of prostate cancer (OR=1.061 [1.031, 1.092]) is in agreement with a recent study using the PRACTICAL Consortium[39].

Because mLOY is strongly associated with smoking, and because the pleiotropic effects of tobacco smoking instruments have large effects on human health, we conducted multivariable MR using cigarettes per day[38] as a second exposure in the models of significant forward and reverse MR associations (Fig. 3bc; Supplementary Data 22). The direct effect of each exposure on their respective outcomes in multivariable MR were highly similar to the univariable model, and remained significant, with the exception of SHBG ($P=8.92\times10^{-3}$). Additionally, SHBG had $P>0.05$ in all of MR-Egger, weighted-median and weighted-mode sensitivity analyses, and so this association is considered less robust than those of the other traits. Overall, multivariable

medRxiv preprint doi: https://doi.org/10.1101/2024.04.24.24306301; this version posted April 25, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. This article is a US Government work. It is not subject to copyright under 17 USC 105 and is also made available for use under a CC0 license.

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

medRxiv preprint doi: https://doi.org/10.1101/2024.04.24.24306301; this version posted April 25, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. This article is a US Government work. It is not subject to copyright under 17 USC 105 and is also made available for use under a CC0 license.

Discussion

In this multi-ancestry meta-analysis we have more than doubled the number of genetic loci associated with mLOY, adding 167 novel loci. The large number of new mLOY cases (N=126,108) that powered our discovery was enabled by several factors. First, we utilized the MVP biobank[20], which consists of mostly aging males, many of whom were current or previous cigarette smokers. Next, the MoChA[4,40] software, which can detect chromosome-length events at a lower cell fraction threshold than in previous GWAS[1,18], enabled the inclusion of cases from early stages of mosaicism proliferation. Lastly, the sample size achieved by combining cohorts into a large multi-ancestry meta-analysis increased the number of GWS loci compared to the largest individual cohort (EUR) by about 50%. This mLOY GWAS is the first to include AFR and HIS populations; we identified 5 AFR-specific signals, and additionally found 17 loci in HIS which were replicated in EUR. An additional benefit of MVP is that the extensive electronic health records within this biobank enabled a thorough post-GWAS exploration of the relationships of mLOY with other phenotypes.

We found that cell cycle was the most highly enriched Biological Process (GO) for genes positionally mapped to novel loci (47 novel loci), strengthening the mechanistic suppositions of previous studies that reduced cell cycle efficacy is a primary driver of mLOY[1,17,18]. The most significant gene set in novel loci was G2/M checkpoint, responsible for blocking damaged and incompletely replicated DNA from progressing through the cell cycle. The G2/M checkpoint is regulated in part by p53[41], a tumor suppressor involved with mLOY, CHIP, and other mCAs[3]. The next most highly enriched novel gene set, the PI3K/AKT/mTOR pathway (13 novel loci), is a regulator of

medRxiv preprint doi: https://doi.org/10.1101/2024.04.24.24306301; this version posted April 25, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. This article is a US Government work. It is not subject to copyright under 17 USC 105 and is also made available for use under a CC0 license.

cell growth and survival, particularly in the context of cancer progression[42]. The Heme Metabolism gene set (11 novel loci) also includes genes involved in erythroblast differentiation; the expansion of leukocytes in mLOY has previously been associated with reduced erythrocytes[43]. Additionally, Xenobiotic Metabolism (11 novel loci) of foreign substances, including cigarettes, environmental pollutants, and chemotherapy drugs, has also been strongly associated with mCAs in previous studies[11,44,45].

Our forward MR associations were all directionally consistent with a recent observational study in UKB[46]; additionally, their MR analysis of SHBG is in agreement with our finding that SHBG exerts a positive causal influence on mLOY in forward MR[46]. Interestingly, higher triglycerides had a protective causal influence on mLOY (OR=0.84 [0.80, 0.89]; P=5.38×10^{-11}) while higher HDL conferred risk (OR=1.10 [1.05, 1.15]; P=3.71×10^{-5}). Though this is opposite of the pattern commonly observed in CVD, previous MR studies[47] have identified robust associations between HDL and increased risk of breast cancer,[48,49] and between triglycerides and decreased risk of breast cancer.[50]. Triglycerides also had significant negative genetic correlation and PGS-WAS score with mLOY, and HDL had significant positive PGS-WAS score with mLOY. Two novel mLOY-risk-increasing SNPs are also lead SNPs for triglycerides, at *BUD13*[51] and *SNX17*[52]. Overall, more studies are necessary to uncover the mechanisms underlying this phenomenon. We also found that red blood cell distribution width (RDW) exerts a positive causal influence on mLOY that was nearly significant at the multiple testing threshold (OR= 1.11 [1.04, 1.18]; P=0.002). RDW is an index which reflects impaired erythropoiesis and abnormal red blood cell survival. Multiple mLOY-associated cyclin genes are related to RDW mechanism: *CDK6* (novel) promotes G1/S transition, *CCND3*

medRxiv preprint doi: https://doi.org/10.1101/2024.04.24.24306301; this version posted April 25, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. This article is a US Government work. It is not subject to copyright under 17 USC 105 and is also made available for use under a CC0 license.

(known) encodes cyclin D3, alters cell cycle progression and reduce control of cell size[53], and novel TWAS gene *CCNK* is a cyclin activator.

In addition to replicating the result that mLOY can increase risk of prostate cancer[39]; reverse MR indicated a significant causal influence of mLOY on increased plateletcrit, increased neutrophil percentage, and decreased lymphocyte percentage. The directions of these relationships are concordant with the previous observational report[43]. Neutrophil-to-lymphocyte ratio (NLR) in peripheral blood is an emerging prognostic factor in many diseases, especially cancers[43,54], and elevated NLR indicates neutrophilic inflammatory response, impaired cell-mediated immunity, and is suggestive of overall poor prognosis[43,54,55].

Our study was not without limitations. First, we performed a cross-sectional study at a single time point. Future studies may benefit from a prospective study design as mLOY is associated with aging[56]. Additionally, it has been shown in MDS that young and old cases have distinct genetic landscapes[57], which should also be examined in future studies. We also did not consider the effects of environmental exposures aside from smoking and BMI. Veterans may be disproportionately exposed to pollutants and other toxins over their lifetimes compared to the general public[58]. This could have caused our mLOY prevalence estimates to be inflated (although they were in agreement with previous studies), and could exacerbate potential gene-environment interactions on mLOY risk. Next, we did not distinguish between high- and low-cell fraction of mLOY when defining cases. Our classification method could detect very low cell fractions as opposed to most existing studies which used high cell fraction detection methods such as intensity thresholding (i.e. on mLRR-Y). Next, our analysis evaluated DNA from

medRxiv preprint doi: https://doi.org/10.1101/2024.04.24.24306301; this version posted April 25, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. This article is a US Government work. It is not subject to copyright under 17 USC 105 and is also made available for use under a CC0 license.

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

We found broad concordance across multiple ancestral populations in our meta-analysis, as well as with the previous BBJ cohort[19], strengthening the generalizability of our findings. Future multi-ancestry meta-analyses may enable increased power and associated loci discovery. The new risk loci identified in this study will lead to improved genetic risk prediction, diagnosis, and understanding of the cellular mechanisms surrounding mLOY.

medRxiv preprint doi: https://doi.org/10.1101/2024.04.24.24306301; this version posted April 25, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. This article is a US Government work. It is not subject to copyright under 17 USC 105 and is also made available for use under a CC0 license.

Methods

Ethics/study approval

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

Genotyping, imputation, and ancestry assignment

Genomic data processing was performed for >650,000 MVP participants (releases 1-4). Genotyping was performed using the Thermo Fisher MVP 1.0 Affymetrix Axiom Biobank array[21]. Samples with >2.5% missing genotype calls, excess heterozygosity, those that were potential duplicates, and those with discordance between genetic sex and self-identified gender, were excluded. SNPs with missingness >5% or minor allele frequency (MAF) that deviated by >10% from the 1000 Genomes Project Phase 3 (1KGP3) data[59] were excluded. Pairwise genetic relatedness was estimated using KING[60]; one individual was removed at random from each pair of first-degree relatives, preferentially retaining cases from case-control pairs. Ancestry was algorithmically assigned using HARE (Harmonized ancestry and race/ethnicity)[61], which incorporates self-reported race and ethnicity data to train a genetic ancestry classifier. Using HARE, we grouped 544,112 male participants in MVP according to European (EUR), African (AFR), or Hispanic (HIS) ancestry.

Genotypes were statistically phased over the entire cohort using SHAPEIT4 version 4.1.3[62] with PBWT depth 8. Phased genotypes were imputed to the African Genome Resources (AGR) reference panel[63] using Minimac 4. The AGR panel consists of all 5,008 1KGP3 haplotypes and an additional 2,862 haplotypes from unrelated pan-African samples. As AGR contains biallelic SNPs only, a second imputation was

medRxiv preprint doi: https://doi.org/10.1101/2024.04.24.24306301; this version posted April 25, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. This article is a US Government work. It is not subject to copyright under 17 USC 105 and is also made available for use under a CC0 license.

performed using 1KGP3, with indels and other complex variants merged into the primary imputation. Imputation for chrX for EUR, AFR, and HIS was performed using TOPMed (hg38); significant loci on chrX were lifted over[64] to hg19 for reporting.

Detection of mLOY using long-range haplotype phase

We used SHAPEIT4[62] to infer haplotypes from array genotypes for the whole MVP cohort and we utilized MoChA[4,40], an extension to the BCFtools software suite[65], to infer the presence of mLOY by detecting shifts in allelic ratios between the phased PAR1 and PAR2 haplotypes, similar to what was done previously in UKB[1]. This methodology allows to infer the presence of mLOY for cell fractions as low as ~1%. Cell fraction was estimated from B allele frequency deviation (bdev) using the formula 4*bdev / (1+2*bdev).

GWAS

Presence or absence of any detectable mLOY cell fraction was used as a case-control trait in all analyses, performed using male participants only. Single variant genome-wide association testing was carried out with REGENIE v1.0.6.7[66] using age, age-squared, and twenty PCs. REGENIE step 1 was performed using leave-one out cross validation (--loocv). Approximate Firth likelihood ratio test (LRT) was applied as fallback for associations with $P<0.05$, with SE computed based on LRT where applied. We kept common variants with minor allele frequency (MAF) ≥0.1% and minimum imputation quality (INFO) of 0.3. Two significant chrX loci in PAR1 (rs2857319) and PAR2 (rs306890), both near the boundary with the nonPAR, had large frequency differences between X and Y chromosomes in the Genome Aggregation Database

medRxiv preprint doi: https://doi.org/10.1101/2024.04.24.24306301; this version posted April 25, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. This article is a US Government work. It is not subject to copyright under 17 USC 105 and is also made available for use under a CC0 license.

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

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

For replication, MVP-EUR COJO association signals were compared to summary statistics from the previous mLOY GWAS in UKB[1]. An updated version of chrX UKB summary statistics were utilized for this comparison (N_{case}=40,466; $N_{control}$=146,066; personal.broadinstitute.org/giulio/mLOY); MACH R2 values were considered for variant quality in lieu of INFO scores.

Fine-mapping

We performed Bayesian fine-mapping of each genome-wide significant locus in the EUR and AFR using SuSiE[68] . Pairwise SNP correlations were calculated directly from imputed dosages on 320,831 European-ancestry samples in MVP using LDSTORE 2.0[69]. The maximum number of allowed causal SNPs at each locus was set to 10 (the default used in the FinnGen fine-mapping pipeline: https://github.com/FINNGEN/finemapping-pipeline). Fine-mapping regions which overlapped the major histocompatibility complex (MHC; chr6:25,000,000-34,000,000)

medRxiv preprint doi: https://doi.org/10.1101/2024.04.24.24306301; this version posted April 25, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. This article is a US Government work. It is not subject to copyright under 17 USC 105 and is also made available for use under a CC0 license.

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

Rare variant analysis

We conducted association analyses for rare variants in each MVP ancestry group using REGENIE and the same covariates as in standard GWAS. We considered only variants genotyped on the MVP 1.0 array[21], which is enriched in protein-altering rare variants, and applied the Rare Heterozygous Adjustment algorithm[70] to improve the positive predictive value of rare genotype calls. We further restricted the included markers to directly genotyped ultra-rare variants (MAF<0.1% in controls) classified as "high-impact"[71]. Rare variants were categorized as somatic or germline based on allele balance for heterozygotes obtained from the Genome Aggregation Database (gnomAD)[72].

GWAS multi-ancestry meta-analysis

MVP EUR, AFR, and HIS cohorts were filtered by INFO>0.5 to retain only high quality variants, and meta-analyzed for fixed effects using METAL (v20200505), weighting effect sizes by the inverse of their corresponding standard errors. Only variants present in two or more ancestries were retained. Loci were defined for all cohorts (including counting loci in UKB[1] for comparison purposes), using the two-stage "clumping" procedure implemented in the Functional Mapping and Annotation (FUMA) platform[73]. In this process, genome-wide significant variants are collapsed into LD blocks (r^2>0.6) and subsequently re-clumped to yield approximately independent (r^2<0.1) signals; adjacent signals separated by <250kb are ligated to form independent

medRxiv preprint doi: https://doi.org/10.1101/2024.04.24.24306301; this version posted April 25, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. This article is a US Government work. It is not subject to copyright under 17 USC 105 and is also made available for use under a CC0 license.

loci. Novel variants were defined as COJO signals in independent ancestry cohorts, or meta-analysis index variants, located >1Mb from a previously reported GWAS association with mLOY. For the multi-ancestry meta-analysis, we further performed a sensitivity analysis using the Han-Eskin random effects model (RE2) in METASOFT v2.0.1[24]. FE and RE2 *P*-values at top loci were highly similar. We compared our meta-analysis lead variants to UKB[1] as described above and to BBJ[19] where available. BBJ reported their results as mLRR-Y intensity thresholding as a proxy for mean Y chromosome dosage in circulating blood cells of subjects.

Local ancestry deconvolution and tract-based GWAS

We inferred local ancestry within AFR participants assuming two-way (AFR/EUR) admixture, and within HIS assuming three-way (AFR/EUR/NAT) admixture. The 1000 Genomes YRI (N=108) and CEU (N=99) populations, were used as the AFR and EUR reference, respectively, and 43 Native American samples from Mao et al.[74,75] were used as the NAT reference. We used RFMIX[76] version 2 to generate local ancestry calls for phased genotypes. We then extracted ancestry-specific dosages from the imputed data into PLINK 2.0-compatible files[77] using custom scripts based on the Tractor workflow[25]. For the AFR analysis, EUR-specific dosages were put into a PGEN file, and African-specific dosages and EUR haplotype counts were interlaced in a zstandard-compressed table. For the HIS analysis, EUR-specific dosages were put into a PGEN file, with African and NAT-specific dosages and EUR and AFR haplotype counts interlaced into a zstandard-compressed table. We used these files to conduct a local ancestry-aware GWAS using the PLINK 2.0 local covariates feature, obtaining ancestry-specific marginal effect size estimates.

medRxiv preprint doi: https://doi.org/10.1101/2024.04.24.24306301; this version posted April 25, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. This article is a US Government work. It is not subject to copyright under 17 USC 105 and is also made available for use under a CC0 license.

Gene set, tissue and cell type enrichment analysis

FUMA GENE2FUNC[73] was performed using multi-ancestry meta-analysis summary statistics in genes that were positionally mapped to significant variants (within 10 Kbp) excluding the MHC gene region; this analysis was also stratified by GWAS locus novelty. Benjamini-Hochberg (FDR) was used as the gene set enrichment multiple test correction method. Hallmark gene sets[78] were used to categorize genes.

To further evaluate whether the genomic loci implicated in mLOY were enriched in any particular cell type, we intersected common mLOY risk variants with broad and blood-specific epigenomic catalogs of cell-specific open chromatin[26,27] using an LD score partitioned heritability approach (LDSC)[79] (Fig. S_ctype_a-b). For the broad epigenome catalog encompassing various human tissues[26], we re-used the open chromatin regions associated with each tissue from the lists provided by the creators of the atlas (https://www.meuleman.org/DHS_Index_and_Vocabulary_hg38_WM20190703.txt.gz). To identify cell-specific chromatin regions within the epigenome map of human blood lineages[27], we conducted differential analysis on sequencing data sourced from the Gene Expression Omnibus (GEO) under accession GSE74912. To ensure a consistent evaluation of the generated LD-sc statistics, which rely on the overall genomic coverage of the tested chromatin regions, we selected an identical number of the most specific open chromatin regions from each blood lineage for subsequent heritability analysis by LDSC. For contextual comparison of heritability signal with the other diseases, we acquired summary statistics of Crohn's disease[80], rheumatoid arthritis[81], systemic lupus erythematosus[82], multiple sclerosis[83], or Alzheimer's disease[84]. Similarly to the FUMA

medRxiv preprint doi: https://doi.org/10.1101/2024.04.24.24306301; this version posted April 25, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. This article is a US Government work. It is not subject to copyright under 17 USC 105 and is also made available for use under a CC0 license.

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

eQTL analysis using published datasets

We interrogated a previously published dataset[28] for SNPs associated with mLOY in our European ancestry cohort. Chromosome, position, and alleles were used as unique identifiers with which to cross-reference SNPs across different immune cell subsets.

Transcriptomic imputation model construction and transcriptome-wide association study

Transcriptomic imputation models were constructed as previously described[85,86] for tissues of the GTEx[87] v8, STARNET[29] and PsychENCODE[30,88] cohorts. For GTEx and STARNET cohorts, we considered adipose tissue: subcutaneous (GTEx & STARNET) and visceral (GTEx & STARNET); arterial tissue: aorta (GTEx & STARNET), coronary (GTEx), mammary (STARNET), and tibial (GTEx); blood (GTEx & STARNET); cell lines (GTEx): EBV-transformed lymphocytes and transformed fibroblasts; endocrine (GTEx): adrenal gland, pituitary, and thyroid; colon (GTEx): sigmoid and trasverse; esophagus (GTEx): gastroesophageal junction, mucosa and mascularis; pancreas (GTEx); salivary gland minor (GTEx); stomach (GTEx); terminal ileum (GTEx); heart (GTEx): atrial appendage and left ventricle; liver (GTEx & STARNET), skeletal muscle (GTEx & STARNET); nerve tibial (GTEx); reproductive (GTEx): mammary tissue, ovary, prostate, testis, uterus, vagina; lung (GTEx); skin (GTEx): not sun exposed suprapubic and sun exposed lower leg; and spleen (GTEx). From PsychENCODE[30,88] we considered brain: dorsolateral prefrontal cortex (DLPFC) genes. The genetic datasets of the GTEx[87], STARNET[29] and PsychENCODE[88] cohorts

medRxiv preprint doi: https://doi.org/10.1101/2024.04.24.24306301; this version posted April 25, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. This article is a US Government work. It is not subject to copyright under 17 USC 105 and is also made available for use under a CC0 license.

were uniformly processed for quality control (QC) steps before genotype imputation as previously described[85,86]. We restricted our analysis to samples with European ancestry as previously described[85]. Genotypes were imputed using the University of Michigan server[89] with the Haplotype Reference Consortium (HRC) reference panel[90]. Gene expression information was derived from RNA-seq gene level counts, which were adjusted for known and hidden confounders, followed by quantile normalization. For GTEx, we used publicly available, quality-controlled, gene expression datasets from the GTEx consortium (http://www.gtexportal.org/). RNA-seq data for STARNET were obtained in the form of residualized gene counts from a previously published study[29]. For the dorsolateral prefrontal cortex from PsychENCODE we used post-quality-control RNA-seq data that were fully processed, filtered, normalized, and extensively corrected for all known biological and technical covariates except the diagnosis status[30] as previously described[86]. Feature types queried include genes, long non-coding RNA (lincRNA), microRNA, processed transcripts, pseudogenes, RNA, small nucleolar RNA (snoRNA), plus constant (C), joining (J), and variable (V) gene segments.

For population classification we used individuals of known ancestry from 1000 Genomes. We excluded variants in regions of high linkage disequilibrium, variants with MAF<0.05, variant with high missingness (>0.01), and variants with Hardy-Weinberg equilibrium $P<1\times10^{-10}$; the remaining variants were pruned (--indep-pairwise 1000 10 0.02 with PLINK[91]) and PCA was performed with PLINK[77] version 2.0. We used the first (PC1), second (PC2) and third (PC3) ancestral PCs to define an ellipsoid based on 1000Gp3v5 EUR samples[59] and samples within 3 SD from the ellipsoid center were classified as EUR; based on this definition of EUR samples, we excluded one non-

medRxiv preprint doi: https://doi.org/10.1101/2024.04.24.24306301; this version posted April 25, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. This article is a US Government work. It is not subject to copyright under 17 USC 105 and is also made available for use under a CC0 license.

European ancestry individual. In the remaining samples (n = 405), we performed additional sample-level quality control by retaining non-related samples (--king-cutoff 0.0884 with PLINK[77] version 2.0) with sample-level missingness < 0.015 for variants with variant-level missingness < 0.02, and heterozygosity rate of < 3SD away from the mean; of note, no samples were excluded by these steps. For the next step of our pipeline, we performed outlier testing in the gene expression data. After performing counts per million filtering (> 0.5 counts per million in at least 30% of samples) and voom normalization, PCA was performed, and we excluded individuals located more than 4 SD away from the mean of the ellipsoid defined by PC1 to PC3. This did not remove any individuals but assured us that our data did not contain any outliers. In this final set of individuals, we performed variant-level quality control of the genotypes by removing variants with less than 0.01 minor allele frequency (for all variants possible, we utilized minor allele frequencies reported by Allele Frequency Aggregator European population[92], to reduce minor allele frequency bias from the comparatively small imputation model training population), 5 minor allele counts and 0.02 missingness rate; only variants present in the reference panel of the Haplotype Reference Consortium were retained to ensure good representation of variants in the target GWAS[90]. We used this final set of quality-controlled genotypes in conjunction with our normalized expression data to discover the optimal number of PEER factors to find expression quantitative trait loci. Our analysis led to the decision to utilize 15 PEER factors, which had resulted in the discovery of 4,299 significant eQTLs[93]. This was the closest value to 90% of the maximum value of eQTLs discovered by any chosen number of PEER factors (4,844 significant eQTLs from 50 PEER factors). This allowed us to retain the

medRxiv preprint doi: https://doi.org/10.1101/2024.04.24.24306301; this version posted April 25, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. This article is a US Government work. It is not subject to copyright under 17 USC 105 and is also made available for use under a CC0 license.

maximum signal for gene expression prediction without overcorrecting our data. After residualization for 15 PEER factors, expression data were quantile normalized. Genotypes were then converted to dosages, and missing values were replaced with twice the variant's minor allele frequency before dosages were rounded to the nearest whole number. For training, we used PrediXcan[94] for the construction of the retinal transcriptomic imputation model due to a lack of SNP epigenetic annotation information; for all other models, we used EpiXcan[85].

Multi-tissue transcriptome-wide association study (TWAS)

We performed a gene-trait association analysis as previously described[85]. We applied the S-PrediXcan method[95] to integrate the summary statistics and the transcriptomic imputation models constructed above to obtain gene-level association results. *P*-values were adjusted for multiple testing using the Benjamini & Hochberg (FDR) method and Bonferroni correction. *P*-values across tissues were meta-analyzed using ACAT[31] ≤0.05 and predictive r^2>0.01 to control for both significance and variance explained.

Summary-data-based Mendelian randomization

To test for joint associations between GWAS summary statistics SNPs and eQTL, the SMR method[96], a Mendelian randomization approach, was used. Top SNPs used in SMR for each probe were selected as the most significant SNP in the eQTL data which was also present in the GWAS data. The SMR software (v1.03) was run using the default settings using GTEx Consortium[87] v8 whole blood tissue. European samples of the 1KGP were used as a reference panel. Bonferroni multiple-testing

medRxiv preprint doi: https://doi.org/10.1101/2024.04.24.24306301; this version posted April 25, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. This article is a US Government work. It is not subject to copyright under 17 USC 105 and is also made available for use under a CC0 license.

correction was applied on SMR *P*-values (PSMR). Moreover, a post-filtering step was applied by conducting heterogeneity in dependent instruments (HEIDI) test. The HEIDI test distinguishes the causality and pleiotropy models from the linkage model by considering the pattern of associations using all the SNPs that are significantly associated with gene expression in the cis-eQTL region. The null hypothesis is that a single variant is associated with both trait and gene expression, while the alternative hypothesis (P_{HEIDI}<0.05) is that trait and gene expression are associated with two distinct variants. The same tissues as in the TWAS section from the V8 release of the GTEx Consortium[87] were queried in SMR.

Heritability and genetic correlation analyses

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

Association of PGS Catalog-based polygenic scores with mLOY status

Phenome-wide polygenic score files for 2,652 traits were obtained from European Molecular Biology Laboratory's European Bioinformatics Institute (EMBL-EBI) PGS Catalog (September 2022 version)[32]. All EUR-ancestry subjects in MVP were scored across all available PGSs, excluding those derived from other MVP studies (20), using the +score plugin (https://github.com/freeseek/score) of bcftools[65]. PGSs were loaded into the dosage format field of VCFs readable by SAIGE v1.1.6.2[99] for association testing. Logistic regression was used to examine associations of PGSs on

medRxiv preprint doi: https://doi.org/10.1101/2024.04.24.24306301; this version posted April 25, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. This article is a US Government work. It is not subject to copyright under 17 USC 105 and is also made available for use under a CC0 license.

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

Conditional meta-analysis (mtCOJO)

In order to assess the residual effects of genetic predisposition to cigarette smoking from mLOY susceptibility, we conducted a multi-trait meta-analysis[25] conditioned on cigarettes per day[34]. The conditional meta-analyses were performed using the EUR mLOY summary statistics using GCTA-mtCOJO[46]. The EUR LD panel described above for use with COJO was also used in this analysis.

Polygenic risk scoring

UKB summary statistics[1] were used to construct a polygenic risk score for EUR MVP participants with PRS-CS[100] (v20210604) with a global shrinkage prior of 1×10^{-4}. European samples of the 1KGP were used as a reference panel. Variants were filtered to include only those with $R^2>0.8$ and MAF>1%.

Univariable and multivariable Mendelian randomization

Forward and reverse two-sample Mendelian randomization was performed using summary statistics from previous European GWAS. Summary statistics were accessed through the OpenGWAS database API[101] via the GWAS codes listed in **Supplementary Data 20-22**, except body mass index in males only from the GIANT (Genetic Investigation of ANthropometric Traits) consortium[102], and cigarettes per day from the GSCAN (GWAS & Sequencing Consortium of Alcohol and Nicotine use) consortium[38], which were downloaded separately. The genome-wide significance threshold of $P<5\times10^{-8}$ was used for the selection of genetic instrumental variables. LD clumping of

medRxiv preprint doi: https://doi.org/10.1101/2024.04.24.24306301; this version posted April 25, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. This article is a US Government work. It is not subject to copyright under 17 USC 105 and is also made available for use under a CC0 license.

r^2<0.001 within a 10 Mb window was used to identify independent instruments. Selection, clumping, and harmonization of instruments was performed using TwoSampleMR (v0.5.7)[103]. Primary analyses used the random-effect inverse-variance weighted (IVW) method. Sensitivity analyses were performed with the MendelianRandomization (v0.6.0) R package[104] using fixed effect IVW, and by achieving a nominal significance threshold (*P*<0.05) using one of either MR-Egger, weighted median or weighted mode methods. Additionally we required *P*>0.05 for MR-Egger intercept. MR-PRESSO was conducted using MRPRESSO (v1.0) to test for horizontal pleiotropy[105].

To control for the possibility that genetic instruments related to mLOY displayed possible horizontal pleiotropic effects via smoking behavior, we conducted multivariable Mendelian randomization (MVMR) using the MVMR R package v0.4[106] and included a second exposure of cigarettes per day[38]. We report inverse-variance weighted multivariable MR results along with the test for heterogeneity from a modified form of Cochran's Q statistic with respect to differences in MVMR estimates across the set of instruments. Covariance between the effect of genetic variants derived from the two exposures was fixed to zero due to the use of non-overlapping samples. The *F*-statistic for instrument strength achieved *F*>10 for all tests.

medRxiv preprint doi: https://doi.org/10.1101/2024.04.24.24306301; this version posted April 25, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. This article is a US Government work. It is not subject to copyright under 17 USC 105 and is also made available for use under a CC0 license.

***Consortium authors and affiliations**

VA Million Veteran Program

J. Michael Gaziano[33,34], Philip S. Tsao[16,17,18], Saiju Pyarajan[1,32]

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

Data availability

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

Competing interests

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

medRxiv preprint doi: https://doi.org/10.1101/2024.04.24.24306301; this version posted April 25, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. This article is a US Government work. It is not subject to copyright under 17 USC 105 and is also made available for use under a CC0 license.

References

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

medRxiv preprint doi: https://doi.org/10.1101/2024.04.24.24306301; this version posted April 25, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. This article is a US Government work. It is not subject to copyright under 17 USC 105 and is also made available for use under a CC0 license.

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

medRxiv preprint doi: https://doi.org/10.1101/2024.04.24.24306301; this version posted April 25, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. This article is a US Government work. It is not subject to copyright under 17 USC 105 and is also made available for use under a CC0 license.

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

medRxiv preprint doi: https://doi.org/10.1101/2024.04.24.24306301; this version posted April 25, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. This article is a US Government work. It is not subject to copyright under 17 USC 105 and is also made available for use under a CC0 license.

in the UK Biobank. *Sci. Rep.* **8**, 12316 (2018).

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

medRxiv preprint doi: https://doi.org/10.1101/2024.04.24.24306301; this version posted April 25, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. This article is a US Government work. It is not subject to copyright under 17 USC 105 and is also made available for use under a CC0 license.

Mendelian randomization study. *PLoS Med.* **17**, e1003302 (2020).

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

medRxiv preprint doi: https://doi.org/10.1101/2024.04.24.24306301; this version posted April 25, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. This article is a US Government work. It is not subject to copyright under 17 USC 105 and is also made available for use under a CC0 license.

60. Manichaikul, A. *et al.* Robust relationship inference in genome-wide association studies. *Bioinformatics* **26**, 2867–2873 (2010).

61. Fang, H. *et al.* Harmonizing Genetic Ancestry and Self-identified Race/Ethnicity in Genome-wide Association Studies. *Am. J. Hum. Genet.* **105**, 763–772 (2019).

62. Delaneau, O., Zagury, J.-F., Robinson, M. R., Marchini, J. L. & Dermitzakis, E. T. Accurate, scalable and integrative haplotype estimation. *Nat. Commun.* **10**, 5436 (2019).

63. Gurdasani, D. *et al.* The African Genome Variation Project shapes medical genetics in Africa. *Nature* **517**, 327–332 (2015).

64. Hinrichs, A. S. *et al.* The UCSC Genome Browser Database: update 2006. *Nucleic Acids Res.* **34**, D590–8 (2006).

65. Danecek, P. *et al.* Twelve years of SAMtools and BCFtools. *Gigascience* **10**, (2021).

66. Mbatchou, J., Barnard, L., Backman, J. & Marcketta, A. Computationally efficient whole genome regression for quantitative and binary traits. *bioRxiv* (2020).

67. Yang, J. *et al.* Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. *Nat. Genet.* **44**, 369–75, S1–3 (2012).

68. Wang, G., Sarkar, A., Carbonetto, P. & Stephens, M. A simple new approach to variable selection in regression, with application to genetic fine mapping. *J. R. Stat. Soc. Series B Stat. Methodol.* **82**, 1273–1300 (2020).

69. Benner, C. *et al.* FINEMAP: efficient variable selection using summary data from genome-wide association studies. *Bioinformatics* **32**, 1493–1501 (2016).

70. Mizrahi Man, O. *et al.* Novel genotyping algorithms for rare variants significantly improve the accuracy of Applied Biosystems ™ Axiom ™ array genotyping calls. *bioRxiv* (2021) doi:10.1101/2021.09.13.459984.

71. Cirulli, E. T. *et al.* Genome-wide rare variant analysis for thousands of phenotypes in over 70,000 exomes from two cohorts. *Nat. Commun.* **11**, 542 (2020).

medRxiv preprint doi: https://doi.org/10.1101/2024.04.24.24306301; this version posted April 25, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. This article is a US Government work. It is not subject to copyright under 17 USC 105 and is also made available for use under a CC0 license.

72. Karczewski, K. J. *et al.* Author Correction: The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **590**, E53 (2021).

73. Watanabe, K., Taskesen, E., van Bochoven, A. & Posthuma, D. Functional mapping and annotation of genetic associations with FUMA. *Nat. Commun.* **8**, 1826 (2017).

74. Mao, X. *et al.* A genomewide admixture mapping panel for Hispanic/Latino populations. *Am. J. Hum. Genet.* **80**, 1171–1178 (2007).

75. Martin, A. R. *et al.* Human Demographic History Impacts Genetic Risk Prediction across Diverse Populations. *Am. J. Hum. Genet.* **100**, 635–649 (2017).

76. Maples, B. K., Gravel, S., Kenny, E. E. & Bustamante, C. D. RFMix: a discriminative modeling approach for rapid and robust local-ancestry inference. *Am. J. Hum. Genet.* **93**, 278–288 (2013).

77. Chang, C. C. *et al.* Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* **4**, 7 (2015).

78. Liberzon, A. *et al.* The molecular signatures database (MSigDB) hallmark gene set collection. Cell Syst. 2015; 1 (6): 417--25.

79. Finucane, H. K. *et al.* Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nat. Genet.* **47**, 1228–1235 (2015).

80. Liu, J. Z. *et al.* Association analyses identify 38 susceptibility loci for inflammatory bowel disease and highlight shared genetic risk across populations. *Nat. Genet.* **47**, 979–986 (2015).

81. Okada, Y. *et al.* Genetics of rheumatoid arthritis contributes to biology and drug discovery. *Nature* **506**, 376–381 (2014).

82. Bentham, J. *et al.* Genetic association analyses implicate aberrant regulation of innate and adaptive immunity genes in the pathogenesis of systemic lupus erythematosus. *Nat. Genet.* **47**, 1457–1464 (2015).

83. International Multiple Sclerosis Genetics Consortium. Multiple sclerosis genomic map

medRxiv preprint doi: https://doi.org/10.1101/2024.04.24.24306301; this version posted April 25, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. This article is a US Government work. It is not subject to copyright under 17 USC 105 and is also made available for use under a CC0 license.

implicates peripheral immune cells and microglia in susceptibility. *Science* **365**, (2019).

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

medRxiv preprint doi: https://doi.org/10.1101/2024.04.24.24306301; this version posted April 25, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. This article is a US Government work. It is not subject to copyright under 17 USC 105 and is also made available for use under a CC0 license.

expression variation inferred from GWAS summary statistics. *Nat. Commun.* **9**, 1825 (2018).

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

medRxiv preprint doi: https://doi.org/10.1101/2024.04.24.24306301; this version posted April 25, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. This article is a US Government work. It is not subject to copyright under 17 USC 105 and is also made available for use under a CC0 license.

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

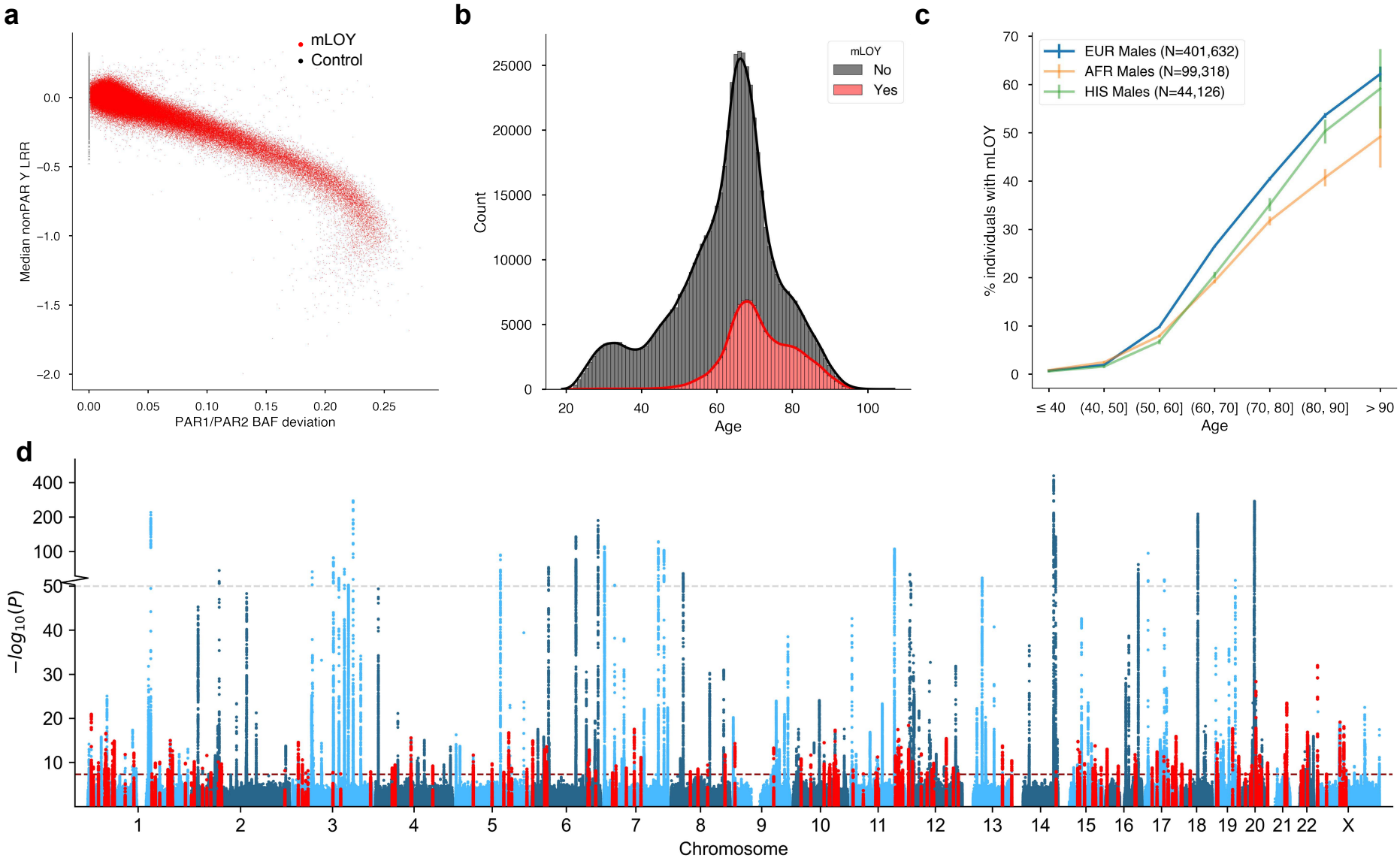

Fig. 1. Mosaic loss of Y chromosome (mLOY) in the Million Veteran Program (MVP). a, Median genotyping probe intensity log R ratio (LRR) vs. phased B Allele Frequency (BAF) in the pseudo-autosomal regions (PAR) 1 and 2. **b**, Density of age distribution in all MVP mLOY cases and controls. **c**, Percentage 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 -log10(*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 ($P<5\times10^{-8}$). The grey dotted line represents a transition from linear to log-scale on the y-axis.

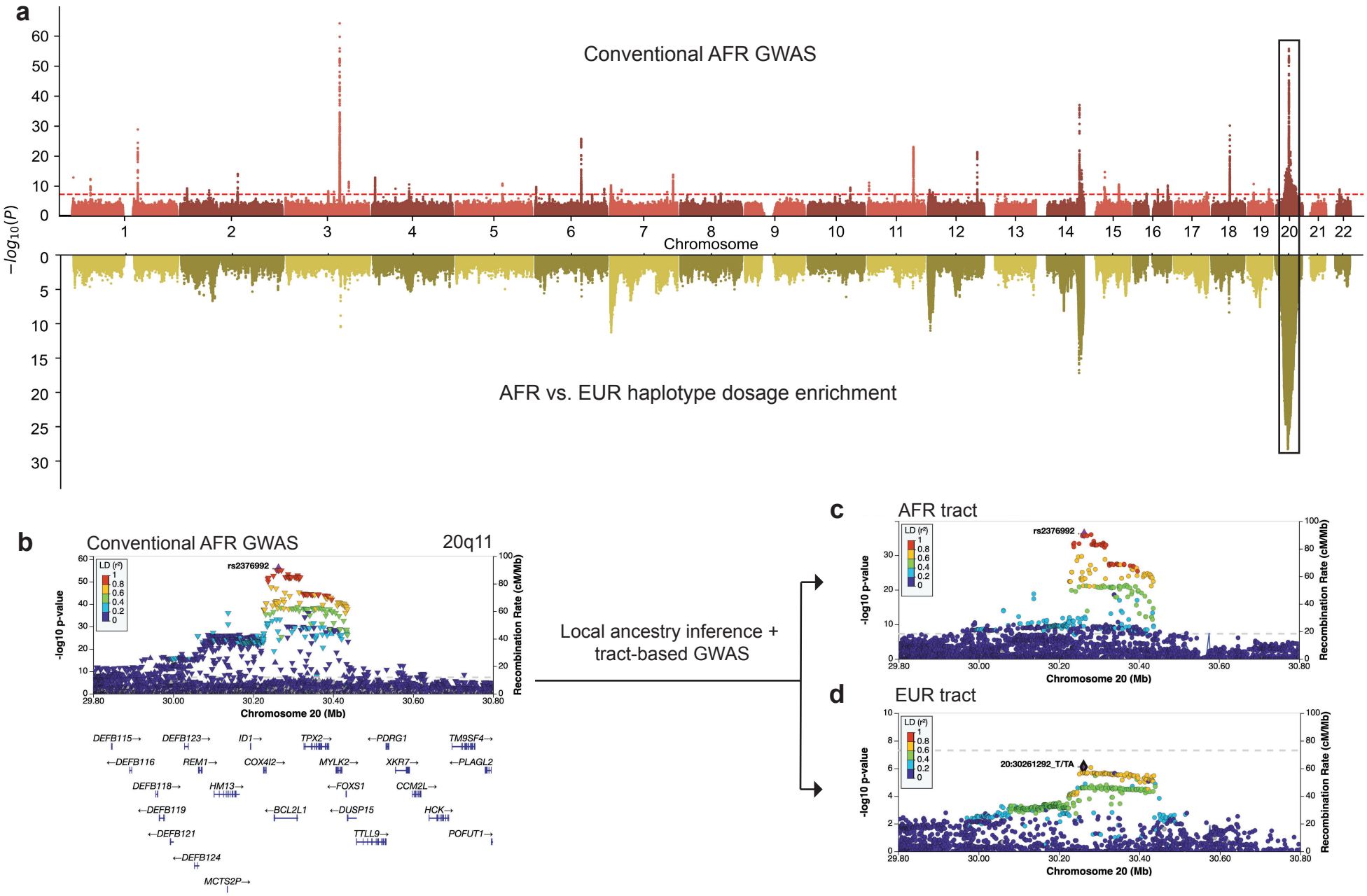

Fig. 2. Global- and local-ancestry-adjusted GWAS of mosaic loss of Y in admixed African (AFR) ancestry Million Veteran Program participants. a, Miami plot showing conventional AFR GWAS (top) and AFR vs. European (EUR) haplotype dosage enrichment (bottom). **b**, At 20q11 (*BCL2L1*) we observed inflation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

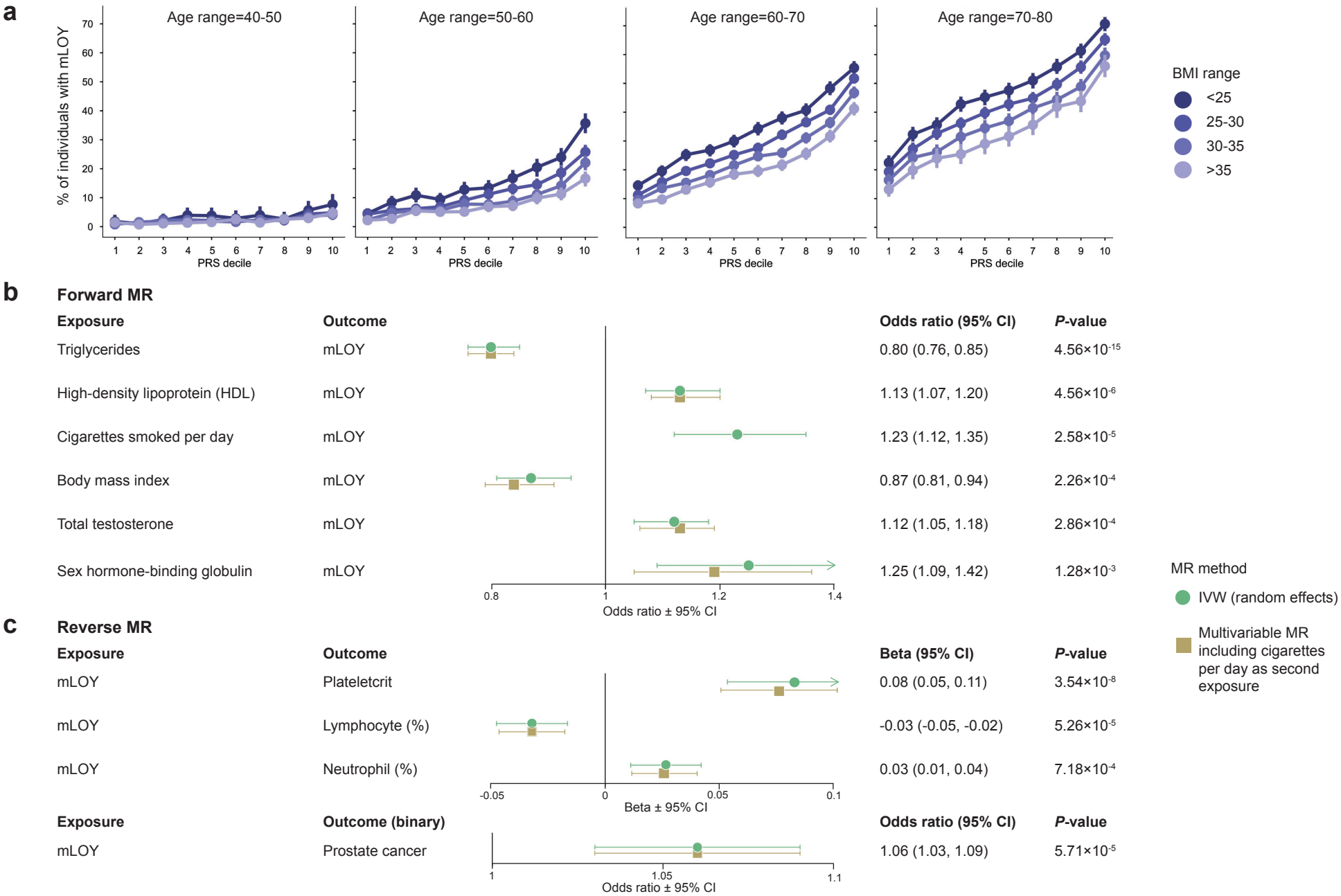

Fig. 3. Mosaic loss of Y by PRS decile and Mendelian randomization (MR). a, Percentage of individuals with mLOY by PRS decile, stratified by age decile (plot grid) and BMI range (color). **b**, Forest plot showing exposure traits with significant results from random effects inverse-variance weighted (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.