Multi-ancestry genome-wide association meta-analysis of mosaic loss of chromosome Y in the Million Veteran Program identifies 240 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], Gabriel Hoffman[8,9], 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,

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 the most commonly detected mosaic chromosomal alteration, and is associated with a range of health outcomes in males. We detected mLOY in 106,054 European (EUR), 13,927 admixed African (AFR), and 6,127 Hispanic (HIS) participants of the Million Veteran Program (MVP; N_{total}=544,112). mLOY was positively associated with cigarette smoking, and negatively with obesity and type 2 diabetes. Multi-ancestry genome-wide association meta-analysis of MVP and UK Biobank mLOY summary statistics, jointly comprising 167,899 cases and 581,224 controls, identified 380 independent risk loci–240 of which were novel. GWAS in MVP high cell fraction (CF>10%) mLOY cases exhibited stronger effect sizes across risk loci. Local-ancestry-aware GWAS resolved ancestry-specific signals at *BCL2L1* and *SETBP1*. Integrative eQTL analyses highlighted 51 genes that causally influence mLOY via differential expression. Finally, traits with genetic correlations and polygenic scores significantly associated with mLOY were selected for Mendelian randomization (MR), implicating thirteen traits as causal influences on mLOY, including triglycerides, high-density lipoprotein, and body mass index.

Introduction

Mosaic loss of chromosome Y (mLOY, also abbreviated LOY) is the most commonly detected mosaic chromosomal alteration (mCA)[1–4]. mLOY is readily detected in leukocytes[5], which are the primary source of blood-derived DNA, and is observable in upwards of 40% of males above age 70[1,6], and 70% of males over 85[7]. High turnover in the hematopoietic stem cell compartment drives clonal expansion in mosaic cell subpopulations possessing selective advantages, though it is unknown if mLOY itself contributes to this selection[5,8,9].

Historically, the gene-poor and repetitive-element-rich composition of the Y chromosome led researchers to believe its role was restricted to spermatogenesis and sex determination[10]. Congruently, mLOY was initially considered to be a benign condition, and a consequence of the broader genomic instability that occurs with aging[11]. In the past decade, epidemiological studies have reported associations between mLOY and a wide range of health outcomes[5], including positive associations with all-cause mortality[12–14], hematological malignancies and other cancers[12,13,15], Alzheimer's disease[16], and cardiovascular disease (CVD)[5,12,17]; negative associations with body mass index (BMI)[17] and gout[17]; and mixed associations with type 2 diabetes (T2D)[12,17]. These observed associations with comorbidities have also been inconsistent across studies[5], perhaps owing to differences in sample collection and mLOY classification methods, such as low versus high cell fraction (CF) detection[6]. Importantly, it has yet to be determined whether mLOY is a driving causal factor of these conditions, a "passenger" (i.e., a consequence), or is symptomatic of a shared,

underlying cause, such as genetic susceptibility to DNA replication errors, or exogenous exposure to mutagens (particularly via smoking cigarettes)[5,18].

The mechanisms by which mLOY contributes to pathogenesis are potentially wide-ranging and gradually being uncovered. For example, in a murine model of mLOY, fibrotic deposition in the extracellular matrix was shown to induce CVD[14]. In another study, CRISPR–Cas9 deletion of chromosome Y produced more aggressive bladder cancer by means of T-cell exhaustion[19]. It has also been suggested that cell lineage influences the relationship between mLOY and transcriptional dysregulation, which in turn produces distinct disease phenotypes[18].

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]. Germline genetics govern many processes which can lead to somatic mCA acquisition and clonal expansion, particularly via genes involved in cell-cycle regulation, DNA damage response, apoptosis, and cancer susceptibility[1,8]. Advances in leveraging genotype phase to detect subtle imbalances in allelic ratio 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[20], including the oncogene *TCL1A*[15], and identified 137 novel loci. A GWAS of mLOY in BioBank Japan (BBJ), which used the mean logarithm of R ratio probe intensity in chromosome Y (mLRR-Y) as a quantitative proxy for mean Y chromosome dosage, identified 50 signals

in 46 loci, 35 of which were reported as novel[21]. Subsequent mLOY GWAS did not reveal additional novel loci[22,23].

In this study we analyzed 544,112 male participants (126,108 cases and 418,004 controls) from 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 (EHRs)[24]. In addition to a large EUR cohort, we present the first GWAS of mLOY status in African (AFR) and Hispanic/Latino (HIS) ancestries. We then meta-analyzed the MVP GWAS results with a previous GWAS from UKB[1], producing the largest mLOY study to-date (N=749,123). 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. Additionally, our results highlight the benefits of inclusive population studies in advancing our understanding of mLOY.

Results

mLOY phenotyping and prevalence across ancestries

In all analyses performed in this study (Fig. 1) we classified mLOY cases in MVP as having any detectable CF of mosaicism in chromosome Y. MoChA[8,25] was used to detect shifts in allelic ratios of phased PAR1 and PAR2 haplotypes shared by the X and Y chromosomes (Fig. 2a), similar to a previous mLOY GWAS in UKB[1]. In MVP EUR, 106,054 of 400,970 men (26.4%; median age 66) had detectable mLOY. Lower prevalences were observed in MVP AFR (13,927 of 99,103; 14.1%; median age 60) and MVP HIS (6,127 of 44,039; 13.9%; median age 60) (Supplementary Data 1). As expected, mLOY prevalence increased with age, from 10% among participants aged 50-60 to upwards of 50% in octogenarians (Fig. 2b-c). CF also increased with age across all ancestries (Extended Data Fig. 1a); high CF mLOY (>10%) was observed in 56%, 37%, and 44% of cases in EUR, AFR, and HIS, respectively (Extended Data Fig. 1b).

To evaluate the impact of ancestry on mLOY prevalence within the three MVP ancestry groups, which had varying age distributions, we generated sensitivity models with sequential covariate adjustment, and applied balancing weights to estimate effects using weighted samples (Extended Data Fig. 2a-c; Supplementary Data 2-3). Upon adjusting for age, age-squared, BMI, smoking status, interactions of these covariates with ancestry, and age*BMI interaction, we observed a reduced marginal odds ratio (OR) of mLOY in AFR (OR=0.72, 95% confidence interval = [0.71, 0.74]) and HIS (OR=0.81 [0.79, 0.84]) compared to EUR (Extended Data Fig. 2d; Supplementary Data 3). The consistency of these ORs in both weighted and unweighted models increase our

confidence in the observed associations. Additionally these findings replicate the results of a recent study in TOPMed[23]. Using the same modeling strategy, we observed an increased association with mLOY in former smokers (OR=1.26 [1.24, 1.28]) and current smokers (OR=1.64 [1.59, 1.70]) (Extended Data Fig. 2e; Supplementary Data 3). Similarly, former and current smokers had higher mLOY CF in all ancestries (Extended Data Fig. 1c). These findings that smoking is associated with increased risk of mLOY are consistent with previous reports[12,15,21,23,26].

Phenome-wide associations between mLOY and comorbidity prevalence

We performed phenome-wide association studies (PheWAS) to identify associations between prevalent disease and mLOY (Supplementary Data 4-6). We observed 344 significant associations in EUR; top associations included increased tobacco use disorder (OR=1.65 [1.62, 1.68]) and chronic airway obstruction (OR=1.35 [1.32, 1.37]), and decreased obesity (OR=0.71 [0.70, 0.72]), T2D (OR=0.75 [0.74, 0.76], and gout (OR=0.68 (0.67, 0.70) (Extended Data Fig. 3a). Expectedly, the associations observed in the full EUR cohort related to respiratory traits and substance use were greatly attenuated in a stratified analysis of non-smokers; patterns of decreased obesity, diabetes, gout, and renal and heart diseases remained significant (Extended Data Fig. 3b). The strongest hematopoietic trait association was decreased thrombocytopenia (OR=0.70 [0.67, 0.73])[27]. Highly similar PheWAS results were observed in AFR and HIS (Supplementary Data 5-6; Supplementary Figs. 1-3), as well as in stratified analyses of current smokers and age-matched high CF cases (Supplementary Data 4-7).

Genome-wide significant associations in MVP EUR, AFR, and HIS

We performed a GWAS for mLOY case-control status in each MVP ancestry group (Supplementary Data 8), and, using COJO (conditional and joint analysis), identified 336 conditionally independent association signals reaching genome-wide significance (GWS; $P<5\times10^{-8}$) in EUR, 50 in AFR, and 17 in HIS (Extended Data Fig. 4; Supplementary Fig. 4a-c; Supplementary Data 9-11). The most significant signals in EUR and HIS were in *TCL1A*; in EUR, we identified the same *TCL1A* lead risk allele as in previous GWAS[1,15] (rs2887399-T; OR=0.71 [0.70, 0.72]; $P=3.18\times10^{-419}$; Supplementary Fig. 5a). The strongest effect observed in the MVP EUR GWAS, in variants with minor allele frequency (MAF) ≥0.1, was in *TP53* (rs78378222; OR=1.66 [1.58, 1.75]; $P=1.44\times10^{-90}$). Of the MVP EUR signals, 222 were within 1 Mb of a previously reported mLOY index variant[1,20,21], including 148 of the 156 previously reported in UKB[1], and effect sizes were highly concordant between these two European cohorts (r=0.91; Supplementary Fig. 6a). In AFR, the most significant association was in *RPN1* (rs113336380), which has a MAF of ~6% in AFR and <0.01% in EUR (Supplementary Fig. 5b).

We improved our variant selection by fine-mapping and estimating credible sets of candidate causal variants in each ancestry. In EUR, we identified 11,242 variants in 334 high-quality credible sets, with a median of 8.5 variants per set (Supplementary Data 12); in AFR, 533 variants in 45 high-quality credible sets, with a median of 5 variants per set (Supplementary Data 13); and in HIS, 713 variants in 16 high-quality credible sets, with a median of 12.5 variants per set (Supplementary Data 14).

Effect directions were concordant between EUR and AFR associations (r=0.66, P=1.13×10^{-46}), with the exception of a small group of novel variants that were exclusively found in EUR (Supplementary Fig. 6b). All GWS HIS signals were also GWS in EUR, with concordant effects (r=0.85; Supplementary Fig. 6c). In these three MVP single ancestry GWAS, we report 114 novel signals in EUR, and five additional GWS novel signals specific to AFR: rs2840287 (*MPL*), rs10786559 (*NKX2-3*), rs2724631 (*ETV6*), rs6018599 (*BLCAP*) and rs113887465 (*GGT5*/*CABIN1*) (Supplementary Fig. 7; Supplementary Data 9-10).

We extended our association analyses for all ancestries to a genotype panel enriched in protein-altering rare variants (MAF<0.1%)[28], and identified four novel rare variants in EUR (Supplementary Data 15). These include 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 clonal hematopoiesis of indeterminate potential (CHIP) and acute myeloid leukemia (AML)[29,30]. The estimated effects of these rare variants were negative and, consistent with expectation, stronger than those of common variants (Supplementary Fig. 8).

High cell fraction GWAS

To explore whether higher CF cases of mLOY are driven by a unique genetic risk profile, we performed GWAS using only age-matched cases with CF>10% (Extended Data Fig. 4; Supplementary Fig. 4d-f; Supplementary Data 16-18). Interestingly, COJO signals in high CF GWAS had stronger effect sizes in a variant-level comparison with the full cohorts (Extended Data Fig. 4); this was consistent with expectations of a

refined phenotypic signal in high CF cases. Overall, we observed strong correlations when comparing COJO signal effects across high CF and full cohorts (r=0.82 to 0.99; Supplementary Data 19). The high CF GWAS revealed two novel signals in AFR (Extended Data Fig. 4b; Supplementary Data 17). Additionally, there were three novel loci in HIS, however, we consider these to be lower confidence due because the lead variants were not replicated (P>0.05) in our other GWAS (Extended Data Fig. 4c; Supplementary Data 18).

Meta-analyses of MVP and UKB

We performed a fixed-effects meta-analysis of MVP EUR, AFR, and HIS ancestries, for a total of 126,108 cases and 418,004 controls. This yielded 239 risk loci, 40 of which did not reach GWS in any individual MVP ancestry (Supplementary Fig. 9ab; Supplementary Data 20). We then performed a combined meta-analysis of MVP with UKB[1,31], for a total of 749,123 samples (167,899 cases and 581,224 controls). In MVP+UKB, we identified 380 risk loci at GWS, 230 of which were novel in this study (Fig 2d; Supplementary Fig. 9c; Supplementary Data 21). Loci were defined using the two-stage clumping procedure implemented by the Functional Mapping and Annotation (FUMA) platform[32]. Locus novelty was defined as the absence of previously reported mLOY associations within the larger of either these locus boundaries or a ±1 Mb window. (The ±1 Mb window was also used to assess novelty of COJO index variants and rare variants). All novel loci in this study are represented in the MVP+UKB meta-analysis, with the exception of ten additional signals only found in a single GWAS (including the three rare variants), for a total of 240 novel mLOY loci (Supplementary Data 22).

Across MVP multi-ancestry and MVP+UKB meta-analysis index variants, rs2887399 (*TCL1A*) remained the most significant association ($P_{MVP+UKB}$=7.93×10^{-573}; Supplementary Fig. 5a). The strongest effects were observed at *CHEK2* (rs186430430; OR=1.78 [1.61, 1.97]) and *TP53* (rs78378222; OR=1.70 [1.64, 1.77]), whose gene products coordinate (together with *ATM*, which contains a rare variant associated with mLOY) in a crucial DNA damage response pathway. For both of these meta-analyses we also generated random-effects (RE) estimates using the Han-Eskin method (RE2)[33], and observed similar *P*-values to FE (Supplementary Data 20-21).

Replication in UK Biobank and BioBank Japan

We performed external replication of our GWAS results using the two previous largest mLOY GWAS, in UKB[1] European and BBJ[21] Japanese ancestry samples. We observed highly similar effects and *P*-values between MVP and UKB[1]. In a variant-level replication of the 327 MVP EUR signals that were available in UKB[1], all but one (rs925301) had the same effect direction, and beta values were highly correlated (*r*=0.91; Supplementary Fig. 6a; Supplementary Data 9, 23). Of the 114 novel MVP EUR signals, 83 had *P*<0.01 and 16 had *P*<1×10^{-5} in UKB[1]. Highly similar replicability was observed between the predominantly European MVP multi-ancestry meta-analysis and UKB[1] (Supplementary Data 20, 23).

We additionally queried BBJ[21] for replicability using the 172 of 239 index variants from the MVP meta-analysis available in that dataset. Although BBJ used a less sensitive quantitative mLOY measure (mLRR-Y) as opposed to our case-control designation, the effects of 151 of 172 index variants were aligned, including all 71 variants with *P*<0.01 in BBJ (Supplementary Fig. 6d; Supplementary Data 20, 23).

Finally, we compared our MVP+UKB meta-analysis with BBJ[21], and found similar effect size and *P*-value concordance as with the MVP multi-ancestry meta-analysis (Supplementary Data 22-23). BBJ[21] had 135 of 286 available MVP+UKB meta-analysis index variants (47.2%) with $P<0.05$ and shared 29 GWS signals (10%; Supplementary Data 22-23).

Local-ancestry-aware associations in admixed groups

Next, we performed local ancestry inference[34] in the two admixed MVP groups in this study, AFR and HIS. On average, AFR had 81.3% African and 18.7% European admixture; HIS had 58.1% European, 33.2% Native American (NAT), and 8.8% African admixture (Extended Data Fig. 5; Supplementary Data 25). We performed local-ancestry-aware GWAS using the Tractor model[35] to resolve ancestry-specific mLOY association signals (Fig. 3a; Supplementary Figs. 10-11). Signals with large allele frequency differences between ancestries resulting from recent admixture can be separated with this method; we highlight two AFR loci at 20q11 and 18q12 to illustrate.

Inflation due to admixture was observed at 20q11.21 near *BCL2L1* (Fig. 3ab). Global adjustment for ancestry by including 20 principal components (PCs) in our conventional GWAS model did not sufficiently resolve this locus. We therefore inferred that the inflation was due to the differences in risk allele frequency at the lead SNP (rs2376992) between ancestral haplotypes (51% AFR vs 22% EUR; Supplementary Fig. 10a) at this large effect size locus (OR=0.80 [0.78, 0.82]; $P=1.7\times10^{-56}$). 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). In the local-ancestry-aware model, the secondary cluster of significant variants downstream of *BCL2L1* was resolved in the AFR tract relative to

the EUR tract (Fig. 3cd; Supplementary Fig. 10b). Strikingly, a meta-analysis of the AFR and EUR tracts for chromosome 20 yielded an inflation factor (λ) of 1.08, compared to 1.40 from the conventional AFR GWAS for this chromosome (Supplementary Data 26). By fine-mapping the smaller credible set of SNPs identified in the AFR tract, we found that the lead risk allele rs2376992 also had the highest posterior probability, and 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. 11c). Local-ancestry-aware association analysis resolved the overlapping LD structures, and revealed the primary AFR signal at rs4414576 (Supplementary Fig. 11d); this allele has 32% frequency in AFR and only 3% in EUR. The EUR tract at this locus (Supplementary Fig. 11e) has a similar structure to the EUR GWAS (Supplementary Fig. 11f).

Additionally, we performed local-ancestry-aware GWAS in HIS (Extended Data Fig. 6). In this model we assessed the haplotype enrichment of 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, and the same lead variant (rs2376992). This locus in HIS is inflated by the low frequency of rs2376992 in NAT haplotypes (0.3%, compared to 51% in AFR and 22% in EUR).

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

Using *P*-values from the MVP+UKB meta-analysis, we performed MAGMA gene-property tissue expression analysis; this indicated strong specificity for blood (P=2.57×10^{-16}) and spleen (P=1.04×10^{-9}) tissues, consistent with our measurement of

mLOY in leukocytes (Supplementary Fig. 12a). The majority of independent GWS signals and variants in LD were annotated as intronic (proportion=0.52; Supplementary Fig 12b). Genes mapped within 10Kb of GWS LD blocks were sent to gene set enrichment analysis (GSEA; Supplementary Data 27). The most significantly enriched Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways were p53 signaling, cell cycle, and chronic myeloid leukemia, which reflect key biological processes proximal to mLOY. Hallmark gene sets, representing well-defined biological processes, were used to categorize known and novel mLOY loci (Supplementary Fig. 12c). Genes associated with the G2/M checkpoint were the most significantly enriched with 14 known and 20 novel loci (adj$P$=3.16×10^{-6}), followed by genes encoding cell cycle related targets of E2F transcription factors, and genes mediating programmed cell death (apoptosis).

We then tested for enrichment of MVP 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[36,37]. 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 blood cell differentiation[21]. 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 pattern of mLOY

genetic variant enrichment in differentiating blood cells differs markedly from enrichment patterns observed for other diseases characterized by perturbations in immune responses, 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 compared our GWAS signals with a recently published expression quantitative trait loci (eQTL) dataset derived from single-cell RNA-seq data across immune cell populations in Northern Europeans[38]. Across the fourteen immune cell subsets, we found that 197 of 327 MVP EUR mLOY signals spanning 251 eQTLs reached at least nominal significance ($P<0.05$; Supplementary Data 28). 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 (6:29842409:T:C, 6:31150435:A:G) in the major histocompatibility complex (MHC) region (6p22.1 to 6p21.3).

Multi-tissue TWAS and SMR

We linked our MVP EUR GWAS signals with functional gene units in a multi-tissue transcriptome-wide association study (TWAS). Our TWAS leveraged 43 tissue models including STARNET blood[39] and a high-powered dorsolateral prefrontal cortex (DLPFC) dataset[40] (Supplementary Data 29); this yielded 2,297 unique significant gene features at $P_{Bonferroni}<0.05$. In the STARNET blood model, 117 features were significant at $P_{Bonferroni}<0.05$ (Supplementary Data 30, Supplementary Fig. 14a). In the DLPFC

model, 191 features were significant at $P_{Bonferroni}$<0.05; revealing one novel gene, *IL21R* (cytokine receptor for interleukin 21) (Supplementary Data 31). All tissues were then meta-analyzed using ACAT[41], yielding a total of 683 genes with $P_{Bonferroni}$<0.05 (Supplementary Data 32, Supplementary Fig. 14b). ACAT revealed an additional four novel genes: *PSTPIP2*, *CCNK*, *PARP10*, and *G3BP1.* We found that mLOY-associated gene expression was highly correlated across the imputed transcriptomes of all tissues (Supplementary Fig. 15), similar to what has been previously observed for other traits[42,43].

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 for MVP EUR (Supplementary Data 33). 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); The most significant of these 51 genes from TWAS with causal support include *ATM*, whose gene product repairs DNA double-strand breaks, *NFE2*, a transcription factor that plays a crucial role in hematopoietic differentiation, and *SLC22A5*, whose gene product transports carnitine (required to move fatty acids into mitochondria) into cells (Supplementary Data 33). Of the novel TWAS genes listed above, *PARP10*, whose gene product regulates transcription and DNA repair, had causal support from SMR in blood.

Genetic correlations of mLOY

Genetic correlations (r_g) between mLOY in MVP EUR and 750 traits were tested (Supplementary Data 34). At the multiple-testing-corrected threshold ($\alpha=6.67\times10^{-5}$), 36 traits were significantly correlated. Several metabolite measures had a significant negative r_g with mLOY, including particle sizes of cholesterol, phospholipids, triglycerides, and total lipids in VLDL, which had r_g ranging from -0.34 (standard error = 0.07) to -0.28 (0.06). Significant and negative r_g were observed for hip circumference, and the related anthropometric measures obesity class 2 and BMI. We also generated r_g using UKB[1] summary statistics, and found highly similar results as in MVP, except among a small group of obesity-related traits, which had a stronger negative r_g in MVP (Supplementary Data 34).

Association of PGS Catalog-based polygenic scores with mLOY

We calculated polygenic scores (PGS) for every trait in the PGS Catalog[44], 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 35). A total of 2,644 scores corresponding to 562 uniquely mapped traits in the Experimental Factor Ontology were tested; we found 82 mapped traits significant after multiple testing correction ($\alpha=0.05/562$). Many of the most significant traits were blood measures, including increased OR (associated with a standard deviation increase in the PGS) of plateletcrit (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 (0.93 [0.91,0.94]), HDL (1.06 [1.04,1.07]), BMI (0.95 [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]).

Multi-trait conditioning (mtCOJO) 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[45] (mtCOJO) in EUR, conditioning on cigarettes per day[46]. Across the genome, only the 15q25 region was primarily impacted by the conditional analysis (Extended Data Fig. 7a). This signal, which was GWS for mLOY in the primary GWAS, was attenuated towards the null after conditioning (Extended Data Fig. 7bc). 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 and lung cancer in GWAS[47,48]. The *CHRNA5* locus did not reach GWS in MVP AFR nor HIS cohorts. However, the top SNP in the MVP multi-ancestry meta-analysis (rs11852372) reached nominal significance in AFR (*P*=0.03; OR=1.05 [1.01, 1.09]), with a similar effect as observed in EUR (OR=1.04 [1.02, 1.05]).

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

We observed significant negative associations between mLOY and obesity in PheWAS and PGS-WAS, as well as a negative genetic correlation; this phenotypic association has also been previously reported[12,49]. Therefore, we sought to examine the prevalence of mLOY in MVP EUR as a function of age, BMI, and mLOY 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. 4a). Our results suggest that the 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 EUR (r_g=-0.11 (0.03)) than in UKB (r_g=-0.05 (0.03); Supplementary Data 34); this may partly be a result of 'healthy volunteer bias' in UKB[50].

Univariable and multivariable Mendelian randomization

We used the significant PGS-WAS associations to inform the selection of 32 traits for two-sample MR in the MVP EUR cohort, using male-only summary statistics where available, to infer potential causality between these exposures and mLOY. Bidirectional causality was assessed, using mLOY as the outcome ("forward MR") and exposure ("reverse MR"). Random effects inverse-variance weighted (IVW) forward MR supported thirteen traits as potential causal influences on mLOY (P<0.05/32; Fig. 4b; Supplementary Fig. 17; Supplementary Data 36). Our analysis indicates that triglycerides, mean corpuscular hemoglobin concentration (MCHC), urate, BMI, reticulocyte percentage, and creatinine exert negative causal influence on mLOY, while cigarettes per day, high-density lipoprotein (HDL), plateletcrit, mean sphered cell volume (MSCV), vitamin D, white blood cell count, and apolipoprotein A (ApoA) showed positive causal influence. Reverse MR supported potential causal influence of mLOY on six traits; indicating negative causal influence upon lymphocyte percentage, MSCV, and MCHC, and positive causal influence upon plateletcrit, vitamin D, and monocyte percentage (Fig. 4c; Supplementary Fig. 18; Supplementary Data 37). All MR results remained robust to multiple sensitivity analyses; all tests except ApoA had P<0.05 in

weighted median MR[51], and all tests passed MR-Egger[52] and MR-PRESSO[53] tests for evidence of horizontal pleiotropy. The bidirectional MR associations with plateletcrit, vitamin D, and MSCV suggest a shared heritability may underlie those associations.

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[54] as a second exposure in the models of significant forward and reverse MR associations (Fig. 4bc; Supplementary Data 38). 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 Vitamin D (forward and reverse) and MCHC (reverse) dropping marginally below the multiple testing threshold. Overall, multivariable MR demonstrates that the inferred causality identified in univariable MR was independent of cigarette smoking. Finally, we sent the significant forward MR traits to mtCOJO and observed modest conditioning of mLOY by triglycerides, particularly at *BUD13* (Supplementary Fig. 19).

Discussion

In this study we have more than doubled the number of loci associated with mLOY, adding 240 novel risk loci which predispose individuals to this highly polygenic acquired trait. Our discovery was powered by a large cohort of older males within MVP[24] (N_{total}=544,112; N_{cases}=126,108), which is further enriched for mLOY risk factors such as cigarette smoking (Supplementary Data 1). This GWAS of mLOY also includes the largest admixed AFR and HIS cohorts to date, for whom we have detailed overlapping genetic risk factors as well as ancestry-specific loci. An additional benefit of MVP is that the extensive EHRs within this biobank enabled a thorough exploration of the associations of mLOY with other phenotypes.

Even with the addition of new mLOY risk loci, our top GSEA results (Supplementary Fig. 12c, Supplementary Data 27) strongly align with the mechanistic hypotheses of earlier GWAS, namely that disruptions in cell cycle regulation and apoptosis are key germline risk factors for mLOY[1,20,211,15,20]. Our most significantly enriched gene set was the G2/M checkpoint (14 known and 20 novel loci), responsible for blocking damaged and incompletely replicated DNA from progressing through the cell cycle. The next most highly enriched gene set, E2F transcription factors, are regulators of the cell cycle[55], DNA replication, and apoptosis. Apoptosis was the third most highly enriched gene set, and our findings strengthen previous inferences regarding the crucial role of programmed cell death in limiting expansion of mosaic cell populations[20].

Forward MR analyses revealed several significant blood lipid traits, with triglycerides exhibiting the most significant inferred causal effect on mLOY risk. Triglycerides also had a negative genetic correlation and PGS association with mLOY that were among the strongest we observed, and this trait is also closely linked with the top PheWAS negative associations, obesity and T2D. These relationships can potentially be explained by the ability of triglycerides to induce apoptosis in monocytes[56–58], a process that can weaken innate immunity but also limit the expansion of clonal cell populations. A similar mechanism may explain the negative causal influence of urate upon mLOY, and the corresponding negative PheWAS and PGS associations with gout, as urate induces apoptosis in monocytes both directly and via reactive oxygen species production[59]. In contrast, the positive causal relationship and PGS association of HDL may be explained by the anti-apoptotic, cytoprotective effects it exerts upon monocytes[60].

Our significant biomarker traits from forward MR (triglycerides, urate, creatinine, HDL, vitamin D, and ApoA) all replicated the findings from a recent UK Biobank study[61]. That study also highlighted sex hormone-binding globulin (SHBG) as having the strongest inferred causal relationship with mLOY, using a fixed effects IVW model. This relationship was not replicable using a RE model, which is more robust to model assumptions and heterogeneity; in fact, SHBG had among the highest heterogeneity *P*-value of all the traits we tested ($P_Q=7.6\times10^{-130}$).

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 highly associated with age[62]. We also did not consider the effects of

environmental exposures aside from smoking and BMI. Veterans may be disproportionately exposed to mutagens over their lifetimes compared to the general public[63], and these environmental effects may add to or interact with inherited mLOY risk. Next, the majority of mLOY studies, including ours, are based on DNA from peripheral blood mononuclear cells only, limiting our view of mLOY in other disease-relevant tissues. Finally, potential causal relationships inferred by MR should be interpreted cautiously given the high polygenicity of mLOY and potential for residual pleiotropy among genetic instruments.

In summary, we report the largest genomic and phenotypic association scans of mLOY, to date. We identified common genetic architecture across multiple MVP ancestral populations, as well as with the previous UKB[1] and BBJ[21] cohorts, strengthening the generalizability of our findings. We hope the novel risk loci identified in this study can enable future researchers to develop improved genetic risk prediction, diagnosis, and understanding of the cellular mechanisms surrounding mLOY.

Methods

Ethics/study approval

The VA Central Institutional Review Board (IRB) approved the study protocol. All participants provided informed consent, and all studies conducted at participating centers received approval from IRBs, in accordance with the Declaration of Helsinki. Only previously generated data were analyzed in this study.

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[28]. Sample exclusion criteria included: >2.5% missing genotype calls, excess heterozygosity, potential duplicates, and discordance between genetic sex and self-identified gender. SNPs with missingness >5% or MAF that deviated by >10% from the 1000 Genomes Project Phase 3 (1KGP3) data[64] were excluded. Ancestry was defined using a machine learning algorithm that harmonizes self-reported ethnicity and genetic ancestry (HARE)[65]. Using HARE, we assigned 544,112 male participants in MVP to either European (EUR), African (AFR), or Hispanic (HIS) ancestry.

Genotypes were statistically phased over the entire cohort using SHAPEIT4[66] (v4.1.3) with Positional Burrow Wheeler Transform (PBWT) depth set to 8. AFR phased genotypes were imputed to the African Genome Resources (AGR) reference panel[67] using Minimac 4[68] (v4.1.0). 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 performed using 1KGP3 to include indels and other complex variants merged into the primary imputation. Imputation for chrX was

performed using the TOPMed panel (hg38); these data became available to us while we were performing this study, and yielded a pronounced improvement in imputation quality for chrX variants. Significant mLOY loci on chrX were lifted over[69] to hg19 and reported with both build positions.

Detection of mLOY using long-range haplotype phase

We used SHAPEIT4[66] to infer haplotypes from array genotypes for the whole MVP cohort and we utilized MoChA[8,25] (v20200825), an extension to the BCFtools software suite[70,71], to infer the presence of mLOY by detecting shifts in allelic ratios between phased PAR1 and PAR2 haplotypes, similar to what was done previously in UKB[1]. Briefly, we selected XY males who had chrX mCA calls with length >1 Mb and lod_baf_phase > 2 as mLOY cases. Samples with chrX calls with 1 ≤ lod_baf_phase ≤ 2 were empirically determined to have borderline evidence of mLOY and were not included as cases or controls. This methodology allowed us to infer the presence of mLOY for CFs as low as ~1%. CF was estimated from B allele frequency deviation (bdev) using the formula 4*bdev / (1+2*bdev).

Estimating effect of ancestry and smoking status

We estimated the effects of ancestry and smoking status on mLOY. Effects of AFR and HIS ancestries were estimated with reference to EUR. Smoking status (current, former, never) was ascertained from smoking-related health factors and ICD-9/10 codes for tobacco dependence, using an algorithm previously validated in the VA[72]. Covariate sensitivity models included age, age-squared, BMI, and ancestry or smoking status where applicable and logistic regression was performed in R (v4.2.2).

Because of the imbalance of covariate distribution, particularly age, between MVP ancestries, we generated balancing weights for participants as an additional effects estimation sensitivity analysis. We used WeightIt[73] (v1.3.0) with the average treatment effect in the population (ATE) estimand and two weighting methods, generalized linear model (glm) propensity score weighting and entropy balancing. Marginal odds ratios were generated with g-computation using marginaleffects[74] (v0.23.0) for R.

Phenome-wide association studies

ICD-9/10 codes were grouped hierarchically into 1,605 phecode-based disease traits. For each phecode, individuals with two or more relevant codes were set as cases, and those without codes were considered controls. Associations of phecode-based outcomes on mLOY status were tested within HARE groups using the R PheWAS package[75] (v0.1). Age, age-squared, and the first 20 ancestry-specific PCs were included as covariates. A Bonferroni-corrected threshold was used to assess significance (α=0.05/1605).

GWAS

Presence or absence of any detectable mLOY CF as determined by MoChA[8,25] was used as the case-control designation in all GWAS. Single variant genome-wide association testing was carried out with REGENIE[76] (v1.0.6.7) using age, age-squared, and the first 20 ancestry-specific PCs. REGENIE applies a whole-genome regression model to control for relatedness and population structure, and includes a Firth correction to control for bias in rare SNPs. 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 (--firth-se). We kept common variants with MAF≥0.1% and imputation quality (INFO) >0.3. A higher INFO cutoff of >0.5 was used for chrX, due to greater challenges with X chromosome genotyping and imputation, and availability of fewer reference haplotypes due to male haploidy. GWAS Manhattan plots were generated using gwaslab[77] (v3.4.7).

Two significant chrX loci in PAR1 (rs2857319, *XG*) and PAR2 (rs306890, *SPRY3*), both near the boundary with the nonPAR, had large frequency differences between X and Y chromosomes in the Genome Aggregation Database (gnomAD v3.1.2). These loci were removed after determining the mLOY-associated allele was exclusively in high CF participants, indicating likely genotyping error (Supplementary Fig. 20).

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)[78] 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 MVP participants in HIS. Default values for all parameters were used. COJO output SNPs with r^2≥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 UKB chrX summary statistics were utilized for this comparison (N_{case}=40,466; $N_{control}$=146,066). These were run natively on GRCh38 using the pipelines ukb2txt.wdl, mocha.wdl, impute.wdl, and assoc.wdl from https://github.com/freeseek/mochawdl. MACH R2

values were considered for variant quality in lieu of INFO scores. The updated chrX sumstats contained an additional signal which did not previously reach GWS[1], X:1565178:C:T (hg38).

GWAS sensitivity analysis in high CF mLOY cases

We performed an additional GWAS sensitivity analysis in only mLOY cases with high CF (>10%). GWAS methods, including QC, model parameters, and post-processing with COJO, were conducted identically to the full cohort models described above. Because selecting for high CF cases exacerbated the age difference between cases and controls (>1 standard deviation in AFR and HIS), we matched cases and controls in each ancestral group on age and age-squared before performing this GWAS. Each case was pre-processed to match two controls in EUR, or three controls in AFR and HIS, with nearest-neighbor (greedy) matching without replacement, using MatchIt[79] (v4.5.5) with matching distance between participants calculated by general linearized model. After matching, absolute standardized mean difference (ASMD) values were <0.05 in AFR and approximately equal to 0.2 in EUR and HIS, indicating improved covariate balance.

Fine-mapping

We adapted the FinnGen fine-mapping pipeline[80] (https://github.com/FINNGEN/finemapping-pipeline) to perform Bayesian fine-mapping at each genome-wide significant locus in EUR, AFR, and HIS using Sum of Single Effects (SuSiE) modeling[81]. Pairwise SNP correlations were calculated directly from imputed dosages on 320,831 European-ancestry samples in MVP using LDSTORE 2.0[82]. The maximum number of allowed causal SNPs at each locus was kept at the

default of 10. Fine-mapping regions which overlapped the MHC region (chr6:25,000,000-34,000,000) 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, 3/21 in HIS).

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 directly genotyped on the MVP 1.0 array[28], which is enriched in protein-altering rare variants, and applied the Rare Heterozygous Adjustment algorithm[83] 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"[84]. Rare variants were categorized as somatic or germline based on allele balance for heterozygotes obtained from the Genome Aggregation Database (gnomAD)[85]. For a heterozygous germline variant, the distribution of allelic balance in heterozygotes has a peak at 50%, while for somatic variants the allelic balance distribution has a peak closer to 0%, with a wide range of allelic balances from 0-100% representing varying cell fractions.

GWAS multi-ancestry meta-analysis

For meta-analysis, MVP EUR, AFR, and HIS cohorts were filtered by INFO>0.5 to retain only high quality variants. Similarly, for the combined meta-analysis of MVP+UKB, the three MVP ancestry groups were meta-analyzed together with UKB[1] summary statistics comprising 41,791 mLOY cases and 163,220 controls. Each set of summary statistics was converted into GWAS-VCFs using the +munge plug-in of

bcftools[71] (v1.21). Fixed-effect meta-analyses were performed using the +metal bcftools plug-in with an IVW scheme. Only variants present in two or more ancestries were retained in meta-analysis results.

Loci were defined using the two-stage "clumping" procedure implemented in the Functional Mapping and Annotation (FUMA) platform[32] (v1.5.2). In this process, genome-wide significant variants were 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 were ligated into a single locus.

Overlapping loci between cohorts were defined using a ±500 Kbp window from COJO index variants, or in the case of meta-analysis loci, the larger of a ±500 Kbp window and the locus boundaries defined by FUMA as described above. Novelty was defined as the absence of previously reported mLOY associations in the larger of either these locus boundaries or a ±1 Mb window. (The ±1 Mb window was also used to assess novelty of COJO index variants and rare variants).

In the MVP multi-ancestry meta-analysis and meta-analysis of MVP+UKB, we further performed a sensitivity analysis using the Han-Eskin random effects model (RE2) in METASOFT[33] (v2.0.1). FE and RE2 *P*-values at top loci were highly similar. We replicated our meta-analysis lead variants in UKB[1] as described above and in BBJ[21] 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 inference (LAI) and local-ancestry-aware GWAS

To improve interpretation of GWAS results in admixed groups[35], we generated local ancestry calls in MVP within 125,846 AFR participants assuming two-way

(AFR/EUR) admixture, and within 53,820 HIS assuming three-way (AFR/EUR/NAT) admixture; these were later subsetted by those passing GWAS QC described above. RFMIX2[34] (v2.03) was used to query MVP release 4 phased genotypes using the gnomAD (v3.1.2) Human Genome Diversity Project (HGDP) + 1KGP subset release as reference. Expectation-maximization (EM) optimization was applied using two and three iterations of the AFR and HIS models, respectively. Global ancestry proportion estimates were derived by taking the average per-chromosome Q estimates (weighted by chromosome length) for each component ancestry. HIS admixture was plotted using ggtern[86] (v2.2.0).

We then extracted ancestry-specific dosages from the imputed data into files compatible with the PLINK 2.0[87] (v2.00a3) local covariates feature (--local-covar; --local-pvar; --local-psam). For AFR, EUR-specific allele dosages were loaded into a .pgen (PLINK 2.0 binary genotype) file, and African-specific dosages and EUR haplotype counts were interlaced in a zstandard-compressed table. For HIS, 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. local-ancestry-aware GWAS was then conducted to obtain ancestry-specific marginal effect estimates.

Gene set, tissue and cell type enrichment analysis

FUMA GENE2FUNC[32] was run using MVP+UKB meta-analysis summary statistics in genes that were positionally mapped to significant variants (within 10 Kbp) excluding the MHC gene region. Benjamini-Hochberg (FDR) was used as the gene set enrichment multiple test correction method. Hallmark gene sets[88] 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[36,37] using the LD score partitioned heritability approach of LDSC[89] (v1.01). For the broad epigenome catalog encompassing various human tissues[36], 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[37], 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 LDSC 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[90], rheumatoid arthritis[91], systemic lupus erythematosus[92], multiple sclerosis[93], or Alzheimer's disease[94]. Default parameters of LDSC were used and the MHC region was excluded.

eQTL analysis using published datasets

We interrogated a previously published dataset[38] 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[42,95] for tissues of the GTEx[96] (v8), STARNET[39] and PsychENCODE[40,97] 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[40,97] we considered brain: dorsolateral prefrontal cortex (DLPFC) genes. The genetic datasets of the GTEx[96], STARNET[39] and PsychENCODE[97] cohorts were uniformly processed for quality control (QC) steps before genotype imputation as previously described[42,95]. We restricted our analysis to samples with European ancestry as previously described[42]. Genotypes were imputed using the University of Michigan server[68] with the Haplotype Reference Consortium (HRC) reference panel[98]. 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[39]. 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[40] as previously described[95]. 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 1KGP3. 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 2.0[99]) and PCA was performed with PLINK 2.0[87]. We used the first three ancestral PCs to define an ellipsoid based on 1KGP3 (v5) EUR samples[64] and samples <3 s.d. from the ellipsoid center were classified as EUR; based on this definition of EUR samples, we excluded one non-EUR 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 2.0[87]) with sample-level missingness < 0.015, variants with variant-level missingness < 0.02, and heterozygosity rate of <3 s.d. from the mean; 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 >4 s.d. away

from the mean of the ellipsoid defined by PC1 to PC3. This did not remove any individuals and 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 MAF<0.01 (for all variants possible, we utilized MAFs reported by Allele Frequency Aggregator European population[100], to reduce MAF bias from the comparatively small imputation model training population), 5 minor allele counts and 0.02 missingness rate; only variants present in the HRC reference panel were retained to ensure good representation of variants in the target GWAS[98]. 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 eQTLs. Our analysis led to the decision to utilize 15 PEER factors, which had resulted in the discovery of 4,299 significant eQTLs[101]. 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 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 MAF before dosages were rounded to the nearest whole number. For training, we used PrediXcan[102] 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[42].

Multi-tissue transcriptome-wide association study (TWAS)

We performed a gene-trait association analysis as previously described[42]. We applied the S-PrediXcan method[103] 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[41] ≤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[104], an MR 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. SMR[104] (v1.03) was run using the default settings using the same GTEx[96] (v8) tissues as in the TWAS section. European 1KGP3 samples were used as a reference panel. Bonferroni multiple-testing correction was applied on SMR *P*-values (PSMR). 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.

Heritability and genetic correlation analyses

Genetic correlation analyses were performed using LDSC[105] (v1.01) using the provided European-ancestry LD scores derived from 1KGP3, as implemented in LDHub

(v1.9.0)[106]. Bonferroni multiple testing correction was applied. SNPs from the MHC region 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[44] (September 2022 version). All EUR-ancestry subjects in MVP were scored across all available PGSs, excluding those derived from other MVP studies (20), using the +score plugin of bcftools[71]. PGSs were loaded into the dosage format field of VCFs readable by SAIGE[107] (v1.1.6.2) for association testing. Logistic regression was used to examine associations of PGSs on MVP EUR mLOY cases and controls, adjusting for the same covariates as in GWAS (age, mean-centered age-squared, and 20 ancestry-specific PCs).

Multi-trait-based conditional and joint analysis (mtCOJO)

In order to assess the residual effects of genetic predisposition to cigarette smoking from mLOY susceptibility, we conducted a multi-trait-based conditional and joint analysis (mtCOJO)[45], conditioned on cigarettes per day[46]. The conditional meta-analyses were performed using the MVP EUR mLOY summary statistics. The EUR LD panel described above for use with COJO was also used in this analysis.

Polygenic risk scoring

UKB[1] summary statistics were used to construct a polygenic risk score for EUR MVP participants with PRS-CS[108] (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 MR was performed using summary statistics from MVP EUR and publicly available European GWAS. Summary statistics were retrieved through the OpenGWAS database API[109] via the accession codes listed in **Supplementary Data 36-37**, except BMI in males only which was derived from the GIANT (Genetic Investigation of ANthropometric Traits) consortium study[110], and cigarettes per day from the GSCAN (GWAS & Sequencing Consortium of Alcohol and Nicotine use) consortium[54], which were downloaded separately. The GWS threshold of $P<5\times10^{-8}$ was used for the selection of genetic instrumental variables. LD clumping of $r^2<0.001$ within a 10 Mb window was used to identify independent instruments. HEIDI filtering of pleiotropic SNP outliers ($P<0.01$) was implemented using the GSMR2[111] (v1.1.1) R package.

Selection, clumping, and harmonization of instruments was performed using TwoSampleMR[112] (v0.6.8). Primary analyses used the random-effect IVW method. Sensitivity analyses were performed with the MendelianRandomization[113] (v0.9.0) R package using fixed effect IVW, using a nominal significance threshold ($P<0.05$) in the weighted median method and $P>0.05$ for MR-Egger[114] intercept. MR-PRESSO[53] was conducted using the MRPRESSO (v1.0) R package to test for horizontal pleiotropy. MR-PRESSO was not performed if either the MR-PRESSO global test was insignificant or the number of IVs retained after outlier removal was insufficient. All *P*-values from MR tests were derived from two-sided tests.

To control for the possibility that genetic instruments related to mLOY exhibited possible horizontal pleiotropic effects via smoking behavior, we conducted multivariable MR using the MVMR[115] (v0.4) R package and included a second exposure of cigarettes per day[54]. We report IVW multivariable MR results along with the test for heterogeneity from a modified form of Cochran's Q statistic with respect to differences in multivariable MR 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.

Significant forward MR traits (triglycerides, HDL, BMI, total testosterone, and SHBG) were sent to conditional analysis with GCTA-mtCOJO[45]. mtCOJO was performed using the MVP EUR mLOY summary statistics and EUR LD panel as described above.

***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.

References

1. Thompson, D. J. *et al.* Genetic predisposition to mosaic Y chromosome loss in blood. *Nature* **575**, 652–657 (2019).
2. Sano, S., Thel, M. C. & Walsh, K. Mosaic Loss of Y Chromosome in White Blood Cells: Its Impact on Men's Health. *Physiology (Bethesda)* **38**, 0 (2023).
3. Jobling, M. A. & Tyler-Smith, C. Human Y-chromosome variation in the genome-sequencing era. *Nat Rev Genet* **18**, 485–497 (2017).
4. Guo, X. *et al.* Mosaic loss of human Y chromosome: what, how and why. *Hum Genet* **139**, 421–446 (2020).
5. 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).
6. Forsberg, L. A. *et al.* Mosaic loss of chromosome Y in leukocytes matters. *Nat. Genet.* **51**, 4–7 (2019).
7. Zekavat, S. M. *et al.* Hematopoietic mosaic chromosomal alterations increase the risk for diverse types of infection. *Nat. Med.* **27**, 1012–1024 (2021).
8. Loh, P.-R., Genovese, G. & McCarroll, S. A. Monogenic and polygenic inheritance become instruments for clonal selection. *Nature* **584**, 136–141 (2020).
9. Silver, A. J., Bick, A. G. & Savona, M. R. Germline risk of clonal haematopoiesis. *Nat. Rev. Genet.* **22**, 603–617 (2021).
10. Quintana-Murci, L. & Fellous, M. The Human Y Chromosome: The Biological Role of a 'Functional Wasteland'. *J Biomed Biotechnol.* **1**, 18–24 (2001).
11. 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).
12. Loftfield, E. *et al.* Predictors of mosaic chromosome Y loss and associations with mortality

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

13. Forsberg, L. A. *et al.* Mosaic loss of chromosome Y in peripheral blood is associated with shorter survival and higher risk of cancer. *Nat. Genet.* **46**, 624–628 (2014).
14. Sano, S. *et al.* Hematopoietic loss of Y chromosome leads to cardiac fibrosis and heart failure mortality. *Science* **377**, 292–297 (2022).
15. Zhou, W. *et al.* Mosaic loss of chromosome Y is associated with common variation near TCL1A. *Nat. Genet.* **48**, 563–568 (2016).
16. Dumanski, J. P. *et al.* Mosaic loss of chromosome Y in blood is associated with Alzheimer disease. *Am. J. Hum. Genet.* **98**, 1208–1219 (2016).
17. Lin, S.-H. *et al.* Incident disease associations with mosaic chromosomal alterations on autosomes, X and Y chromosomes: insights from a phenome-wide association study in the UK Biobank. *Cell Biosci* **11**, 143 (2021).
18. Dumanski, J. P. *et al.* Immune cells lacking Y chromosome show dysregulation of autosomal gene expression. *Cell. Mol. Life Sci.* **78**, 4019–4033 (2021).
19. Abdel-Hafiz, H. A. *et al.* Y chromosome loss in cancer drives growth by evasion of adaptive immunity. *Nature* **619**, 624–631 (2023).
20. 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).
21. Terao, C. *et al.* GWAS of mosaic loss of chromosome Y highlights genetic effects on blood cell differentiation. *Nat. Commun.* **10**, 4719 (2019).
22. Kessler, M. D. *et al.* Common and rare variant associations with clonal haematopoiesis phenotypes. *Nature* **612**, 301–309 (2022).
23. Jakubek, Y. A. *et al.* Genomic and phenotypic correlates of mosaic loss of chromosome Y in blood. *The American Journal of Human Genetics* **112**, 276–290 (2025).
24. 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).

25. Loh, P.-R. *et al.* Insights into clonal haematopoiesis from 8,342 mosaic chromosomal alterations. *Nature* **559**, 350–355 (2018).
26. Dumanski, J. P. *et al.* Mutagenesis. Smoking is associated with mosaic loss of chromosome Y. *Science* **347**, 81–83 (2015).
27. Ouseph, M. M. *et al.* Genomic alterations in patients with somatic loss of the Y chromosome as the sole cytogenetic finding in bone marrow cells. *Haematologica* **106**, 555–564 (2021).
28. 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).
29. 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).
30. Cerchione, C. *et al.* IDH1/IDH2 Inhibition in Acute Myeloid Leukemia. *Front. Oncol.* **11**, 639387 (2021).
31. Bycroft, C. *et al.* The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203–209 (2018).
32. Watanabe, K., Taskesen, E., van Bochoven, A. & Posthuma, D. Functional mapping and annotation of genetic associations with FUMA. *Nat. Commun.* **8**, 1826 (2017).
33. 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).
34. 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).
35. 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).
36. Meuleman, W. *et al.* Index and biological spectrum of human DNase I hypersensitive sites.

Nature **584**, 244–251 (2020).

37. Corces, M. R. *et al.* Lineage-specific and single-cell chromatin accessibility charts human hematopoiesis and leukemia evolution. *Nat. Genet.* **48**, 1193–1203 (2016).
38. Yazar, S. *et al.* Single-cell eQTL mapping identifies cell type–specific genetic control of autoimmune disease. *Science* **376**, eabf3041 (2022).
39. Franzén, O. *et al.* Cardiometabolic risk loci share downstream cis- and trans-gene regulation across tissues and diseases. *Science* **353**, 827–830 (2016).
40. Gandal, M. J. *et al.* Transcriptome-wide isoform-level dysregulation in ASD, schizophrenia, and bipolar disorder. *Science* **362**, (2018).
41. 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).
42. Zhang, W. *et al.* Integrative transcriptome imputation reveals tissue-specific and shared biological mechanisms mediating susceptibility to complex traits. *Nat. Commun.* **10**, 3834 (2019).
43. Voloudakis, G. *et al.* A translational genomics approach identifies IL10RB as the top candidate gene target for COVID-19 susceptibility. *NPJ Genom Med* **7**, 52 (2022).
44. Lambert, S. A. *et al.* The Polygenic Score Catalog as an open database for reproducibility and systematic evaluation. *Nat. Genet.* **53**, 420–425 (2021).
45. Zhu, Z. *et al.* Causal associations between risk factors and common diseases inferred from GWAS summary data. *Nat. Commun.* **9**, 1–12 (2018).
46. 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).
47. 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).
48. Gorman, B. R. *et al.* Multi-ancestry GWAS meta-analyses of lung cancer reveal

susceptibility loci and elucidate smoking-independent genetic risk. *Nat Commun* **15**, 8629 (2024).

49. Loftfield, E. *et al.* Mosaic Y Loss Is Moderately Associated with Solid Tumor Risk. *Cancer Res.* **79**, 461–466 (2019).
50. Fry, A. *et al.* Comparison of Sociodemographic and Health-Related Characteristics of UK Biobank Participants With Those of the General Population. *Am J Epidemiol* **186**, 1026–1034 (2017).
51. Bowden, J., Davey Smith, G., Haycock, P. C. & Burgess, S. Consistent Estimation in Mendelian Randomization with Some Invalid Instruments Using a Weighted Median Estimator. *Genet Epidemiol* **40**, 304–314 (2016).
52. Burgess, S. & Thompson, S. G. Interpreting findings from Mendelian randomization using the MR-Egger method. *Eur J Epidemiol* **32**, 377–389 (2017).
53. 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).
54. Saunders, G. R. B. *et al.* Genetic diversity fuels gene discovery for tobacco and alcohol use. *Nature* **612**, 720–724 (2022).
55. Thurlings, I. & de Bruin, A. E2F Transcription Factors Control the Roller Coaster Ride of Cell Cycle Gene Expression. *Methods Mol Biol* **1342**, 71–88 (2016).
56. Jung, B. C., Kim, H.-K., Kim, S. H. & Kim, Y. S. Triglyceride induces DNA damage leading to monocyte death by activating caspase-2 and caspase-8. *BMB Rep.* **56**, 166–171 (2023).
57. López, S. *et al.* Dietary oleic and palmitic acids modulate the ratio of triacylglycerols to cholesterol in postprandial triacylglycerol-rich lipoproteins in men and cell viability and cycling in human monocytes. *J Nutr* **137**, 1999–2005 (2007).
58. Aflaki, E. *et al.* Triacylglycerol accumulation activates the mitochondrial apoptosis pathway in macrophages. *J Biol Chem* **286**, 7418–7428 (2011).

59. Kimura, Y., Tsukui, D. & Kono, H. Uric Acid in Inflammation and the Pathogenesis of Atherosclerosis. *Int J Mol Sci* **22**, (2021).
60. Kontush, A. HDL-mediated mechanisms of protection in cardiovascular disease. *Cardiovasc Res* **103**, 341–349 (2014).
61. 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).
62. 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).
63. Teichman, R. Exposures of concern to veterans returning from Afghanistan and Iraq. *J. Occup. Environ. Med.* **54**, 677–681 (2012).
64. 1000 Genomes Project Consortium *et al.* A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
65. 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).
66. 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).
67. Gurdasani, D. *et al.* The African Genome Variation Project shapes medical genetics in Africa. *Nature* **517**, 327–332 (2015).
68. Das, S. *et al.* Next-generation genotype imputation service and methods. *Nat. Genet.* **48**, 1284–1287 (2016).
69. Hinrichs, A. S. *et al.* The UCSC Genome Browser Database: update 2006. *Nucleic Acids Res.* **34**, D590–8 (2006).
70. Danecek, P. *et al.* Twelve years of SAMtools and BCFtools. *Gigascience* **10**, (2021).
71. Genovese, G. *et al.* BCFtools/liftover: an accurate and comprehensive tool to convert

genetic variants across genome assemblies. *Bioinformatics* **40**, (2024).

72. Song, R. J. *et al.* Abstract 18809: Development of an Electronic Health Record-Based Algorithm for Smoking Status Using the Million Veteran Program (MVP) Cohort Survey Response. *Circulation* (2016).
73. Greifer, N. WeightIt: Weighting for Covariate Balance in Observational Studies. Preprint at https://ngreifer.github.io/WeightIt/ (2024).
74. Arel-Bundock, V., Greifer, N. & Heiss, A. How to Interpret Statistical Models Using marginaleffects for *R* and *Python*. *J. Stat. Softw.* **111**, (2024).
75. Carroll, R. J., Bastarache, L. & Denny, J. C. R PheWAS: data analysis and plotting tools for phenome-wide association studies in the R environment. *Bioinformatics* **30**, 2375–2376 (2014).
76. Mbatchou, J., Barnard, L., Backman, J. & Marcketta, A. Computationally efficient whole genome regression for quantitative and binary traits. *bioRxiv* (2020).
77. He, Y., Koido, M., Shimmori, Y. & Kamatani, Y. GWASLab: a Python package for processing and visualizing GWAS summary statistics.
78. 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).
79. Ho, D. E., Imai, K., King, G. & Stuart, E. A. MatchIt: Nonparametric Preprocessing for Parametric Causal Inference. *J. Stat. Softw.* **42**, (2011).
80. Kanai, M. *et al.* Insights from complex trait fine-mapping across diverse populations. *bioRxiv* (2021) doi:10.1101/2021.09.03.21262975.
81. 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).
82. Benner, C. *et al.* FINEMAP: efficient variable selection using summary data from genome-

wide association studies. *Bioinformatics* **32**, 1493–1501 (2016).

83. 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.

84. 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).

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

86. Hamilton, N. E. & Ferry, M. Ggtern: Ternary diagrams using ggplot2. *J. Stat. Softw.* **87**, (2018).

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

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

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

90. 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).

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

92. 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).

93. International Multiple Sclerosis Genetics Consortium. Multiple sclerosis genomic map implicates peripheral immune cells and microglia in susceptibility. *Science* **365**, (2019).

94. 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).
95. 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).
96. GTEx Consortium. The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science* **369**, 1318–1330 (2020).
97. Wang, D. *et al.* Comprehensive functional genomic resource and integrative model for the human brain. *Science* **362**, (2018).
98. McCarthy, S. *et al.* A reference panel of 64,976 haplotypes for genotype imputation. *Nat. Genet.* **48**, 1279–1283 (2016).
99. 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).
100. Phan, L. *et al.* ALFA: allele frequency aggregator. *National Center for Biotechnology Information, US National Library of Medicine* (2020).
101. 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).
102. Gamazon, E. R. *et al.* A gene-based association method for mapping traits using reference transcriptome data. *Nat. Genet.* **47**, 1091–1098 (2015).
103. Barbeira, A. N. *et al.* Exploring the phenotypic consequences of tissue specific gene expression variation inferred from GWAS summary statistics. *Nat. Commun.* **9**, 1825 (2018).
104. Zhu, Z. *et al.* Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. *Nat. Genet.* **48**, 481–487 (2016).
105. Bulik-Sullivan, B. K. *et al.* LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat. Genet.* **47**, 291–295 (2015).

106. 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).
107. 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).
108. 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).
109. Elsworth, B. *et al.* The MRC IEU OpenGWAS data infrastructure. *bioRxiv* 2020.08.10.244293 (2020) doi:10.1101/2020.08.10.244293.
110. 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.
111. Xue, A. *et al.* Unravelling the complex causal effects of substance use behaviours on common diseases. *Commun Med (Lond)* **4**, 43 (2024).
112. Hemani, G. *et al.* The MR-Base platform supports systematic causal inference across the human phenome. *Elife* **7**, e34408 (2018).
113. 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).
114. Bowden, J., Davey Smith, G. & Burgess, S. Mendelian randomization with invalid instruments: effect estimation and bias detection through Egger regression. *Int J Epidemiol* **44**, 512–525 (2015).
115. Sanderson, E., Spiller, W. & Bowden, J. Testing and correcting for weak and pleiotropic instruments in two-sample multivariable Mendelian randomization. *Stat. Med.* **40**, 5434–5452 (2021).

Main Figure Captions

Fig. 1. Overview of genome-wide association study for mosaic loss of chromosome Y (mLOY) in Million Veteran Program (MVP) participants. Genome-wide association studies (GWAS) were conducted in MVP European, African, and Hispanic ancestries. These were then meta-analyzed with UK Biobank[1]. Replication was conducted using UK Biobank[1] and BioBank Japan[21]. Several follow-up functional analyses were performed, indicated by bulleted lists. BMI, body mass index; eQTL, expression quantitative trait loci; GSEA, gene set enrichment analysis; mLRR-Y, mean logarithm of R ratio in chromosome Y; mtCOJO, multi-trait-based conditional & joint association analysis; MVMR, multivariable Mendelian randomization; PGS, polygenic score; PheWAS, phenome-wide association study.

Fig. 2. Mosaic loss of chromosome Y (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 plots show the $-\log_{10}(P)$ for associations of genetic variants with mLOY in the multi-ancestry meta-analysis. Novel mLOY index variants and surrounding 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. The 25 most significant novel (bold) and known loci are annotated with the nearest gene.

Fig. 3. Local-ancestry-aware association analysis of admixed African (AFR) ancestry Million Veteran Program participants. a, Stacked Miami plot showing the $-\log_{10}(P)$ values for components of the local-ancestry-aware mLOY GWAS for MVP AFR (from top to bottom): conventional AFR GWAS, AFR tract, AFR haplotype dosage enrichment, European (EUR) haplotype dosage enrichment, and EUR tract. The 20q11 locus is highlighted and shown in the regional Manhattan plots below. **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. 4. 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. MCHC, mean corpuscular hemoglobin concentration; MSCV, mean sphered cell volume; CI, confidence interval.

Extended Data Figure Captions

Extended Data Fig. 1. Cell fraction (CF) of mLOY across groups. a, CF of mLOY in Million Veteran Program participants by ten-year age bins, stratified by ancestry. Boxplots show first decile, first quartile, median, third quartile, and last decile; dots indicate mean. **b**, Density ridgeline plots show distribution of CF for each ancestry. CF 10% is marked to indicate the cutoff for high CF sensitivity analyses. **c**, CF of mLOY, stratified ancestry and by current, former, and never smoking status. Full cohort and CF>10% are shown. Error bars indicate 95% confidence intervals.

Extended Data Fig. 2. Estimating effects of ancestry and smoking status on mLOY. a, Love plot of covariate balance before and after weighting using general linearized model propensity scoring. **b**, Love plot of covariate balance before and after weighting using entropy balancing. **c**, Density plots of age distribution across ancestries before and after each weighting method. **d**, Forest plot of sequential covariate adjustment sensitivity analyses for effect of ancestry on mLOY. **e**, Forest plot of sequential covariate adjustment sensitivity analyses for effect of smoking status on mLOY. BMI, body mass index; c.i., confidence interval. Parenthesis around model 2 covariates indicate the interactions of these terms and ancestry were included in the model.

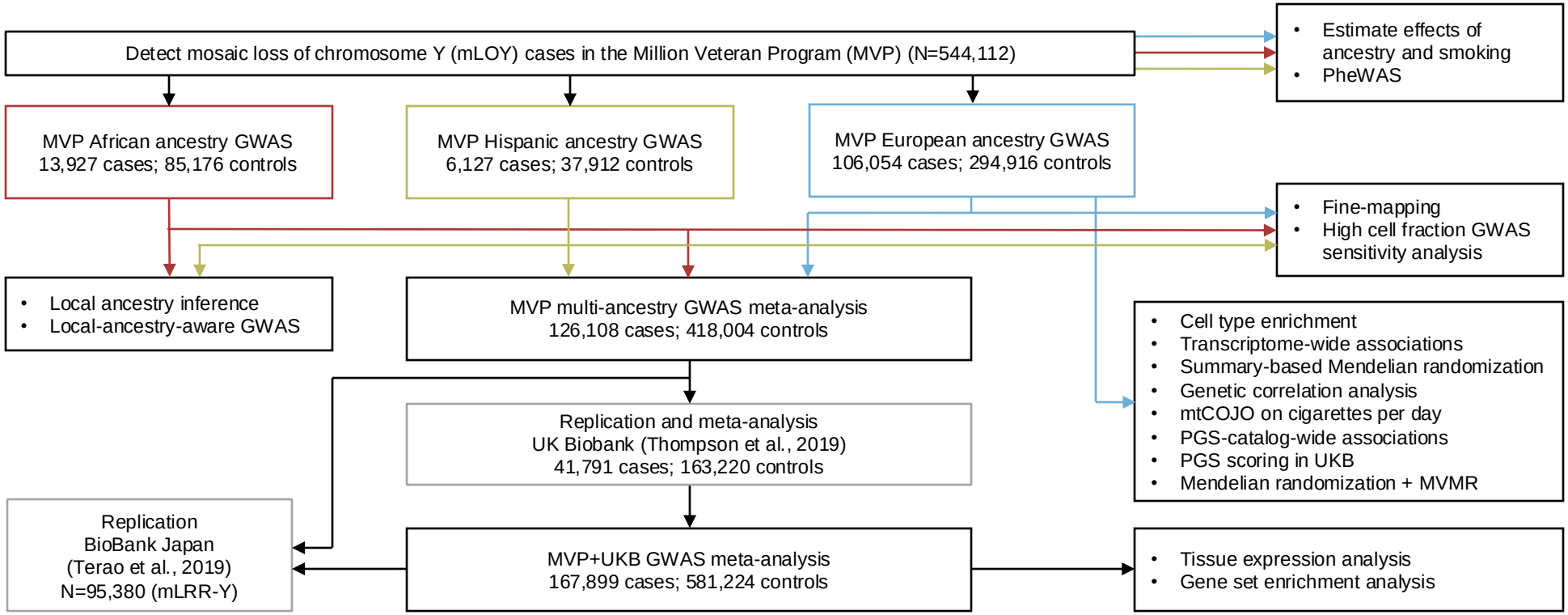
Extended Data Fig. 3. Phenome-wide association studies (PheWAS) of mLOY status. PheWAS for the largest Million Veteran Program ancestry, European (EUR), in: **a**, Full cohort; **b**, Never smokers only. Association *P*-values were obtained from a two-sided test of the z-statistic, calculated using logistic regression. The red line in each plot indicates Bonferroni-corrected significance (α=0.05/1605).

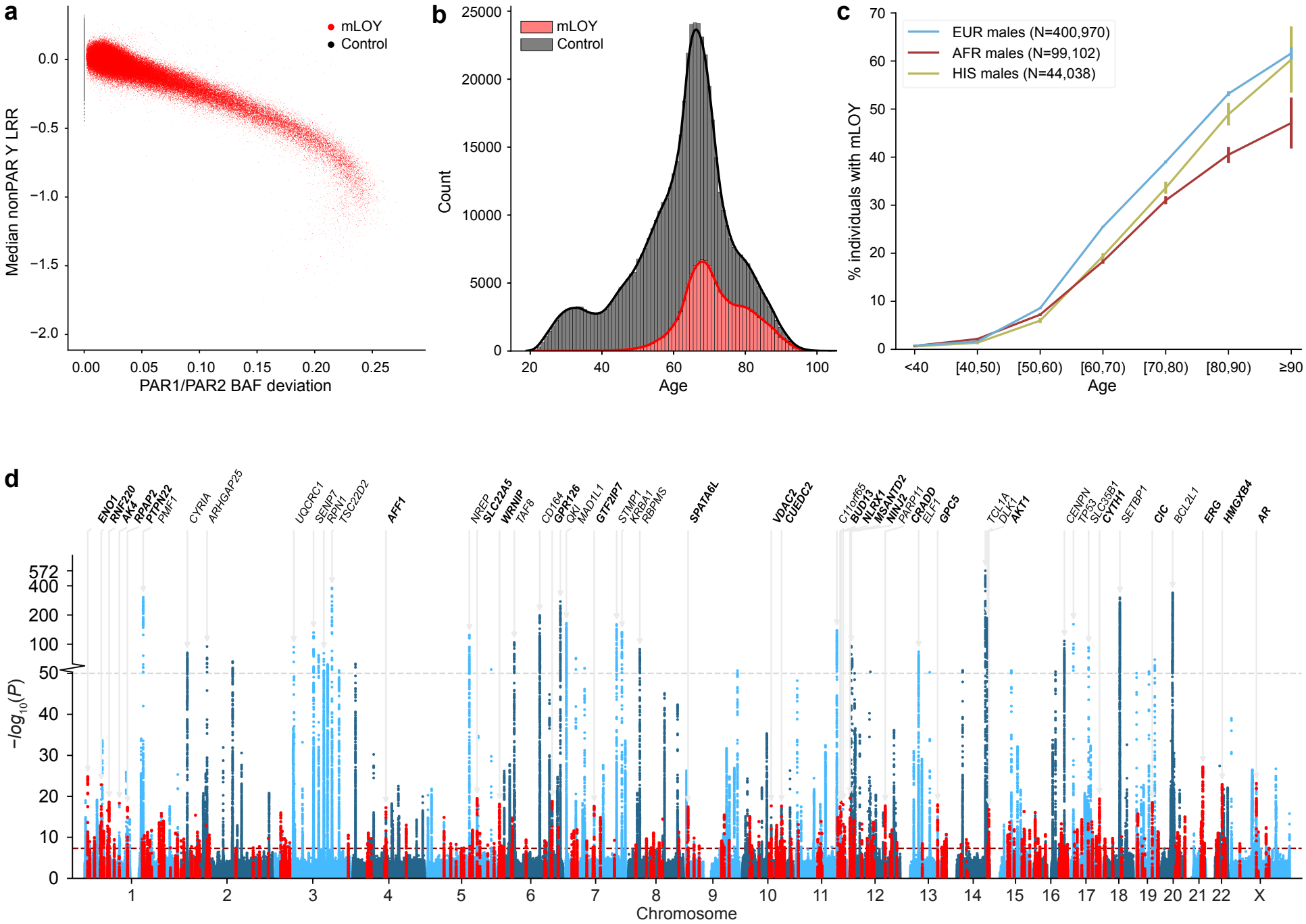
Extended Data Fig. 4. Miami plots of mosaic loss of chromosome Y (mLOY) risk loci in Million Veteran Program participants by ancestry. Miami plots (left) and effect concordance scatterplots (right) for GWAS of mLOY in full cohort and high cell fraction (CF>10%) sensitivity analyses. Miami plots show the -$\log_{10}(P)$ for associations of genetic variants with mLOY in **a**, European ancestry, **b**, African ancestry, and **c**, Hispanic ancestry. Top plots show full cohorts, bottom plots show high cell fraction (CF>10%) sensitivity analyses. Numbers central to plots indicate chromosomes. Novel mLOY index variants (± 50 Kb) are highlighted in top, those additional loci found in CF>10% only are highlighted in bottom plots. The dotted line indicates the genome-wide significance threshold ($P<5\times10^{-8}$). Light grey dotted lines represent a transition from linear to log-scale on the y-axis. c.i., confidence interval; *r*, Pearson correlation coefficient; *m*, slope.

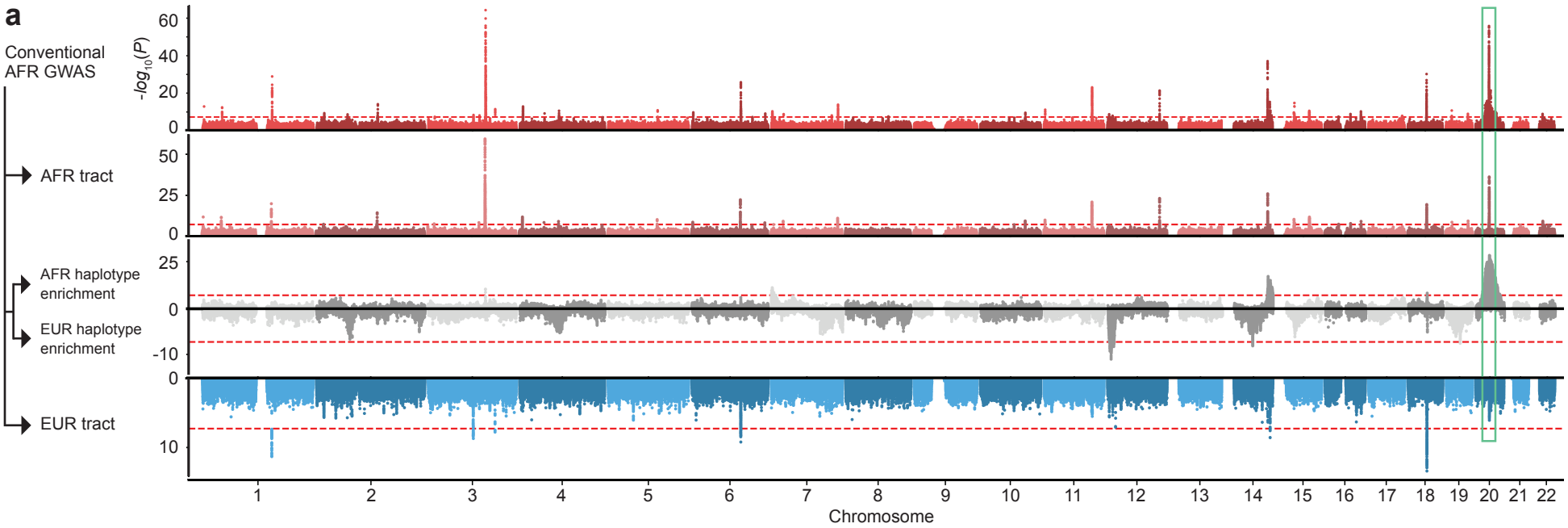
Extended Data Fig. 5. Admixture proportions in Million Veteran Program (MVP) African (AFR) and Hispanic (HIS) cohorts. a, Two-way admixture proportion across MVP AFR participants. **b**, MVP AFR admixture binned by percentile of AFR ancestry (%). **c**, Three way admixture proportion across MVP HIS participants. **d**, MVP HIS admixture binned by percentile of European (EUR), AFR, and Native American (NAT) ancestry (%). **e**, Ternary plot of HIS admixture.

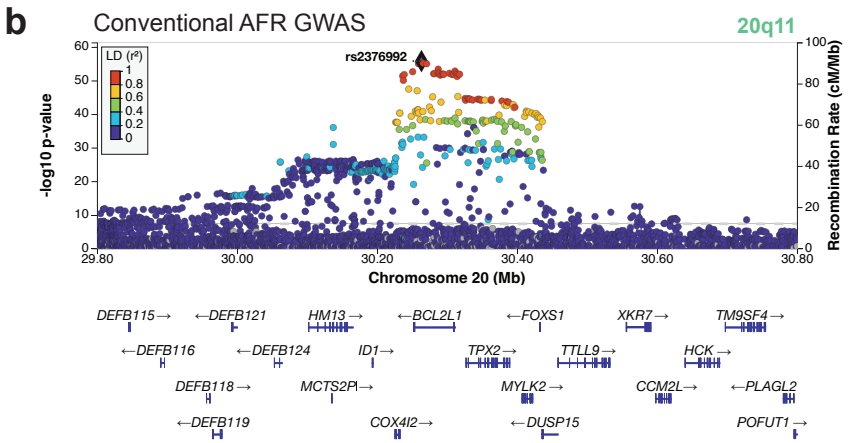
Extended Data Fig. 6. Local-ancestry-aware association analysis of Million Veteran Program (MVP) Hispanic (HIS) participants. Stacked Manhattan plots show the -$\log_{10}(P)$ values for components of the local-ancestry-aware GWAS for MVP HIS (from top to bottom): conventional HIS GWAS, European (EUR) tract, haplotype dosage enrichment of EUR with reference to African (AFR) and Native American (NAT) haplotype dosage, NAT tract, haplotype dosage enrichment of NAT with reference to EUR and AFR haplotype dosage, and AFR tract.

Extended Data Fig. 7. Multi-trait meta-analysis (mtCOJO) conditional analysis on smoking. a, Scatterplot of -$\log_{10}(P)$ from conditioning on cigarettes-per-day with mtCOJO (LPc) vs. -$\log_{10}(P)$ from standard mLOY GWAS (LP) in MVP European ancestry (EUR) demonstrates the region containing the *CHRNA5* gene (chr15:78,711,000-79,258,000), a known smoking locus, was primarily affected by conditioning. Red dashed lines indicate genome-wide significance at $P<5\times10^{-8}$. **b**, Regional Manhattan plot for *CHRNA5* in MVP EUR. **c**, Regional Manhattan plot for *CHRNA5* in MVP EUR after conditioning mLOY on cigarettes-per-day.

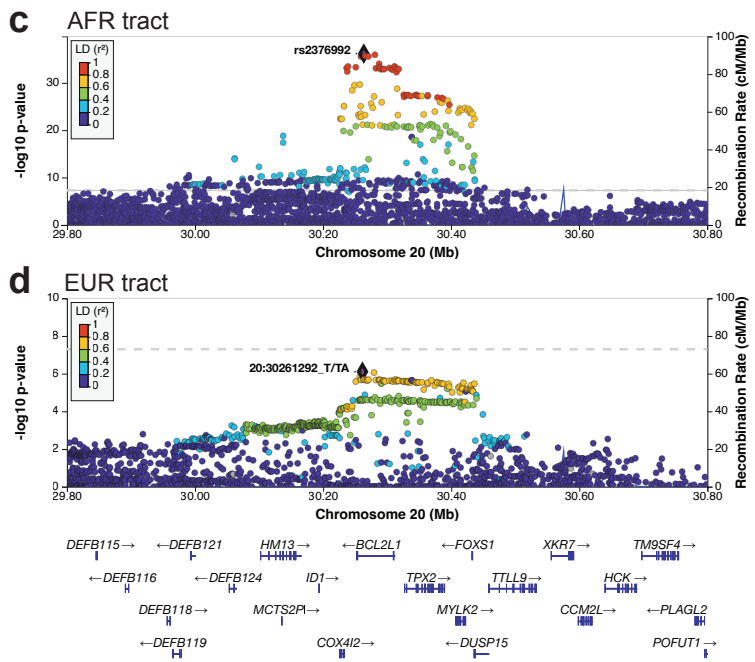

Detect mosaic loss of chromosome Y (mLOY) cases in the Million Veteran Program (MVP) (N=544,112)
Estimate effects of ancestry and smoking
PheWAS
MVP African ancestry GWAS
13,927 cases; 85,176 controls
MVP Hispanic ancestry GWAS
6,127 cases; 37,912 controls
MVP European ancestry GWAS
106,054 cases; 294,916 controls
Fine-mapping
High cell fraction GWAS sensitivity analysis
Local ancestry inference
Local-ancestry-aware GWAS
MVP multi-ancestry GWAS meta-analysis
126,108 cases; 418,004 controls
Cell type enrichment
Transcriptome-wide associations
Summary-based Mendelian randomization
Genetic correlation analysis
mtCOJO on cigarettes per day
PGS-catalog-wide associations
PGS scoring in UKB
Mendelian randomization + MVMR
Replication and meta-analysis
UK Biobank (Thompson et al., 2019)
41,791 cases; 163,220 controls
Replication
BioBank Japan
(Terao et al., 2019)
N=95,380 (mLRR-Y)
MVP+UKB GWAS meta-analysis
167,899 cases; 581,224 controls
Tissue expression analysis
Gene set enrichment analysis

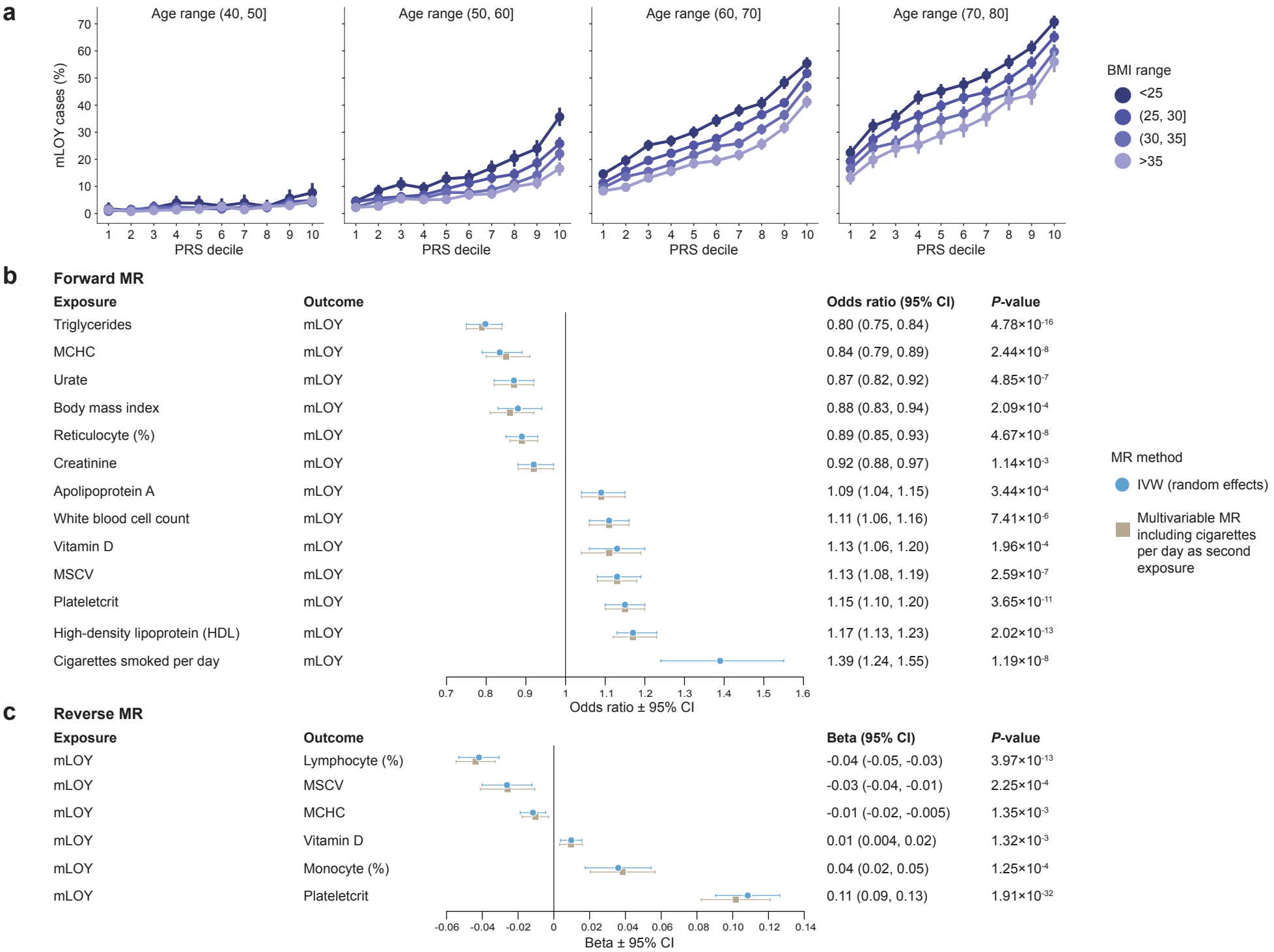

a
mLOY
Control
Median nonPAR Y LRR
PAR1/PAR2 BAF deviation
b
mLOY
Control
Count
Age
c
EUR males (N=400,970)
AFR males (N=99,102)
HIS males (N=44,038)
% individuals with mLOY
Age
<40
[40,50)
[50,60)
[60,70)
[70,80)
[80,90)
≥90
d
ENO1
RNF220
AK4
RPAP2
PTPN22
PMF1
CYRIA
ARHGAP25
UQCRC1
SENP7
RPN1
TSC22D2
AFF1
NREP
SLC22A5
WRNIP
TAF8
CD164
GPR126
QKI
MAD1L1
GTF2IP7
STMP1
KRBA1
RBPMS
SPATA6L
VDAC2
CUEDC2
C11orf65
BUD13
NLRX1
MSANTD2
NINJ2
PARP11
CRADD
ELF1
GPC5
TCL1A
DLK1
AKT1
CENPN
TP53
SLC35B1
CYTH1
SETBP1
CIC
BCL2L1
ERG
HMGXB4
AR
$-log_{10}(P)$
Chromosome





Local-ancestry-aware regional associations




a
Age range (40, 50]
Age range (50, 60]
Age range (60, 70]
Age range (70, 80]
mLOY cases (%)
PRS decile
BMI range
<25
(25, 30]
(30, 35]
>35
b
Forward MR
Exposure
Outcome
Odds ratio (95% CI)
P-value
Triglycerides mLOY 0.80 (0.75, 0.84) 4.78×10-16
MCHC mLOY 0.84 (0.79, 0.89) 2.44×10-8
Urate mLOY 0.87 (0.82, 0.92) 4.85×10-7
Body mass index mLOY 0.88 (0.83, 0.94) 2.09×10-4
Reticulocyte (%) mLOY 0.89 (0.85, 0.93) 4.67×10-8
Creatinine mLOY 0.92 (0.88, 0.97) 1.14×10-3
Apolipoprotein A mLOY 1.09 (1.04, 1.15) 3.44×10-4
White blood cell count mLOY 1.11 (1.06, 1.16) 7.41×10-6
Vitamin D mLOY 1.13 (1.06, 1.20) 1.96×10-4
MSCV mLOY 1.13 (1.08, 1.19) 2.59×10-7
Plateletcrit mLOY 1.15 (1.10, 1.20) 3.65×10-11
High-density lipoprotein (HDL) mLOY 1.17 (1.13, 1.23) 2.02×10-13
Cigarettes smoked per day mLOY 1.39 (1.24, 1.55) 1.19×10-8
Odds ratio ± 95% CI
MR method
IVW (random effects)
Multivariable MR including cigarettes per day as second exposure
c
Reverse MR
Exposure
Outcome
Beta (95% CI)
P-value
mLOY Lymphocyte (%) -0.04 (-0.05, -0.03) 3.97×10-13
mLOY MSCV -0.03 (-0.04, -0.01) 2.25×10-4
mLOY MCHC -0.01 (-0.02, -0.005) 1.35×10-3
mLOY Vitamin D 0.01 (0.004, 0.02) 1.32×10-3
mLOY Monocyte (%) 0.04 (0.02, 0.05) 1.25×10-4
mLOY Plateletcrit 0.11 (0.09, 0.13) 1.91×10-32
Beta ± 95% CI

Main Figure Captions

Fig. 1. Overview of genome-wide association studies (GWAS) for mosaic loss of chromosome Y (mLOY) in Million Veteran Program (MVP) participants. GWAS were conducted in MVP European, African, and Hispanic ancestries. These were then meta-analyzed with UK Biobank[1]. Replication was conducted using UK Biobank[1] and BioBank Japan[21]. Several follow-up functional analyses were performed, indicated by bulleted lists. BMI, body mass index; eQTL, expression quantitative trait loci; GSEA, gene set enrichment analysis; mLRR-Y, mean logarithm of R ratio in chromosome Y; mtCOJO, multi-trait-based conditional & joint association analysis; MVMR, multivariable Mendelian randomization; PGS, polygenic score; PheWAS, phenome-wide association study.

Fig. 2. Mosaic loss of chromosome Y (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 plots show the $-\log_{10}(P)$ for associations of genetic variants with mLOY in the multi-ancestry meta-analysis. Novel mLOY index variants and surrounding 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. The 25 most significant novel (bold) and known loci are annotated with the nearest gene.

Fig. 3. Local-ancestry-aware association analysis of admixed African (AFR) ancestry Million Veteran Program participants. a, Stacked Miami plot showing the $-\log_{10}(P)$ values for components of the local-ancestry-aware mLOY GWAS for MVP AFR (from top to bottom): conventional AFR GWAS, AFR tract, AFR haplotype dosage enrichment, European (EUR) haplotype dosage enrichment, and EUR tract. The 20q11 locus is highlighted and shown in the regional Manhattan plots below. **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. 4. 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

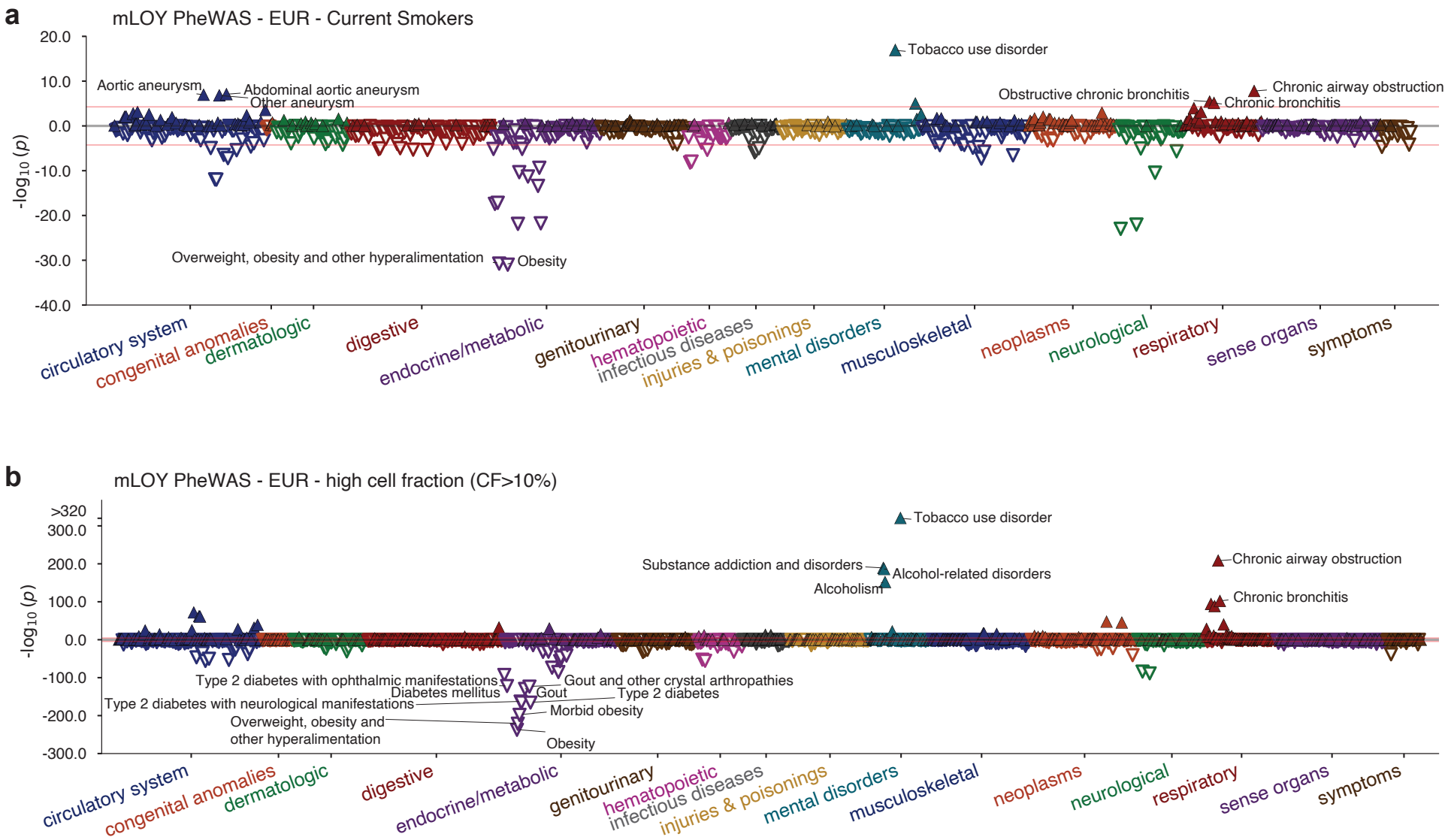
weighted (RE IVW) MR and corresponding multivariable MR (MVMR) with cigarettes per day as an additional exposure, on mLOY outcome in Million Veteran Program Europeans. **c**, Significant results from RE IVW MR and MVMR considering mLOY exposure on trait outcomes. MCHC, mean corpuscular hemoglobin concentration; MSCV, mean sphered cell volume; CI, confidence interval.

Supplementary information

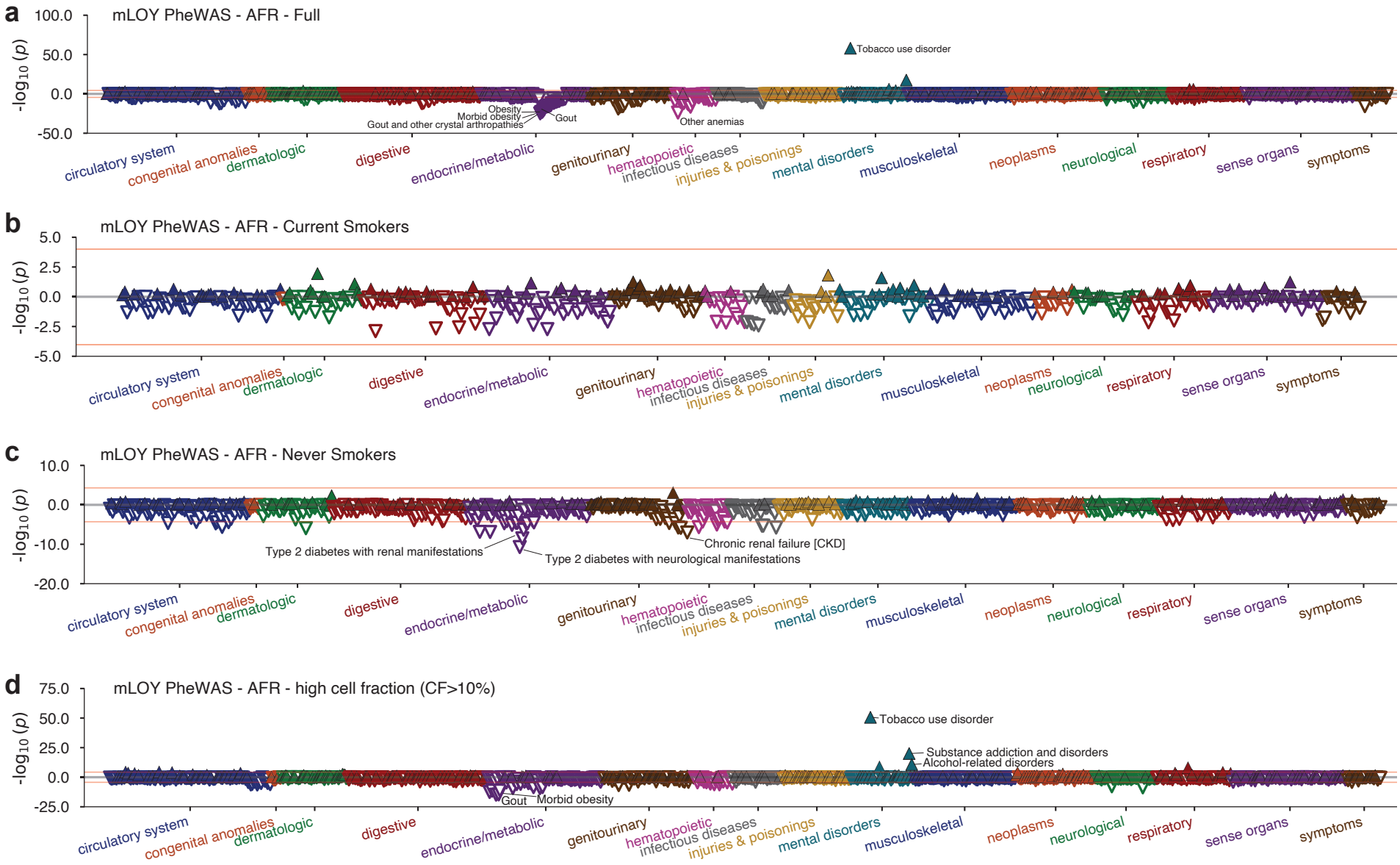
Multi-ancestry genome-wide association meta-analysis of mosaic loss of chromosome Y in the Million Veteran Program identifies 240 novel loci

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], Gabriel Hoffman[8,9], 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‡]

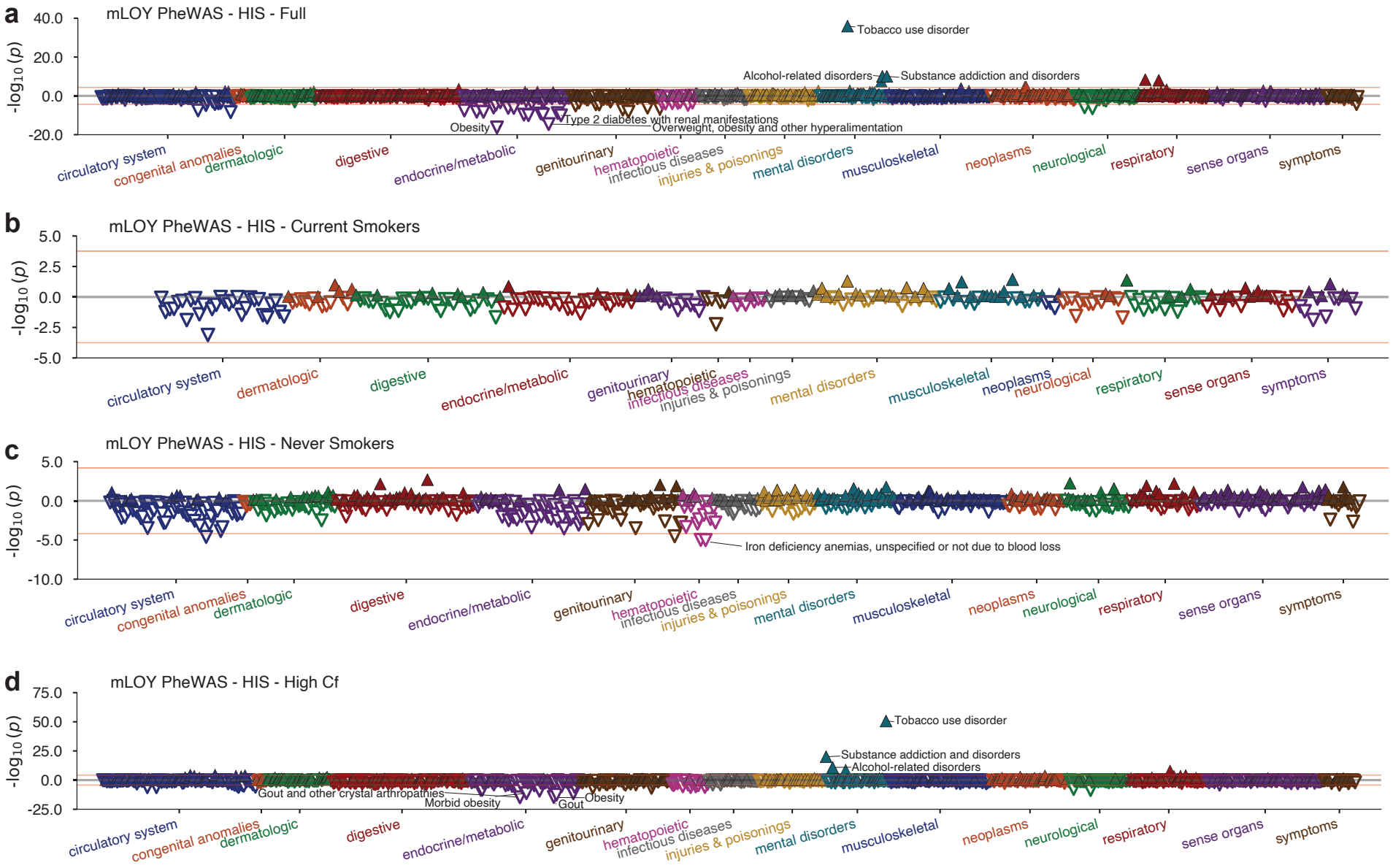
[‡]e-mail: Saiju Pyarajan: saiju.pyarajan@va.gov



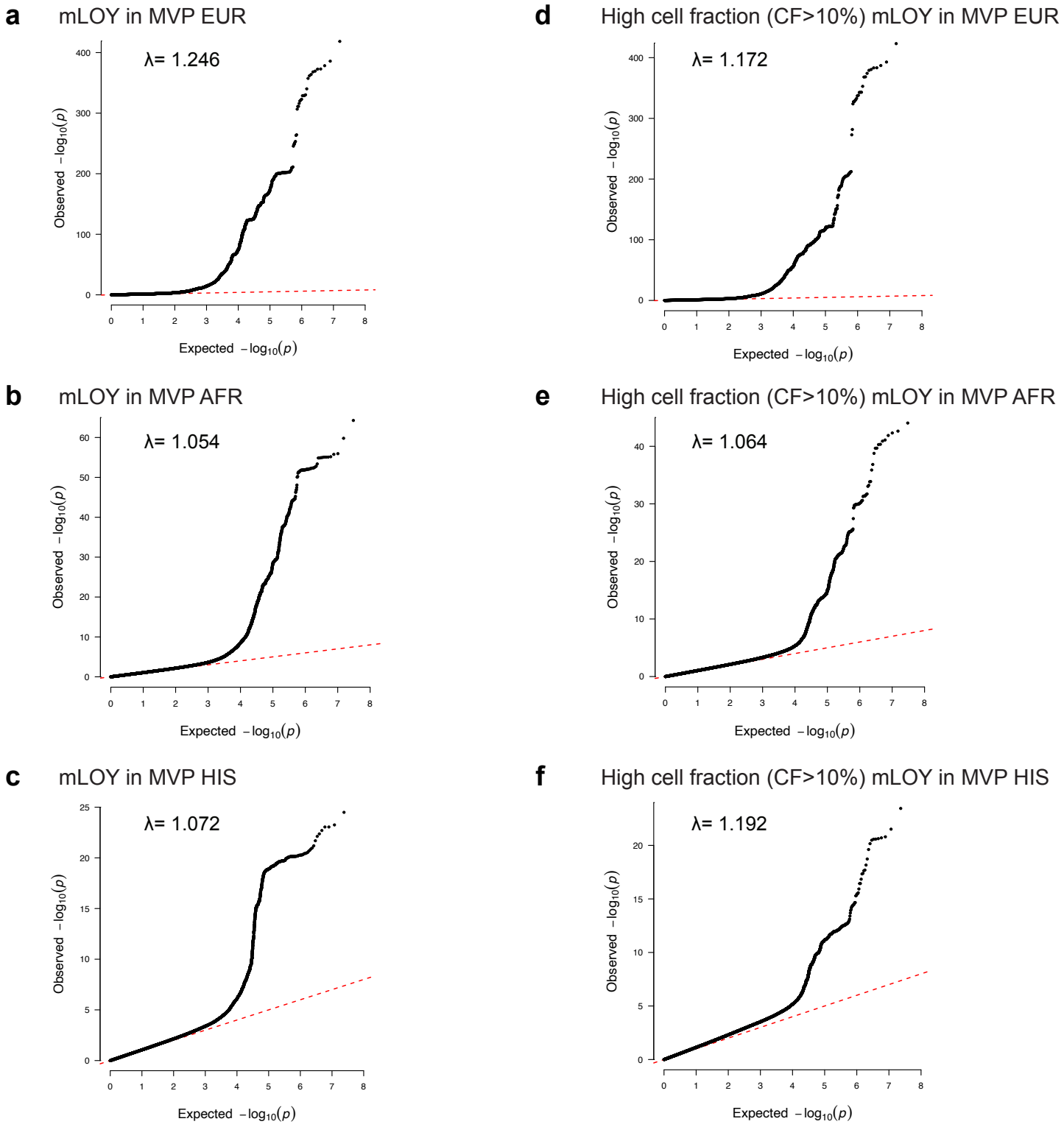
Supplementary Fig. 1. Phenome-wide association scan Manhattan plots for Million Veteran Program European ancestry. Red lines indicate significance threshold with Bonferroni correction for number of traits.



Supplementary Fig. 2. Phenome-wide association scan Manhattan plots for mosaic loss of chromosome Y in Million Veteran Program African ancestry. a, Full cohort ; **b**, current smokers only; **c**, never smokers only; **d**, high cell fraction cases (CF>10%). Association p-values were obtained from a two-sided test of the z-statistic, calculated using logistic regression. The red line in each plot indicates Bonferroni-corrected significance.



Supplementary Fig. 3. Phenome-wide association scan Manhattan plots for mosaic loss of chromosome Y in Million Veteran Program Hispanic ancestry. a, Full cohort ; **b**, current smokers only; **c**, never smokers only; **d**, high cell fraction cases (CF>10%). Association p-values were obtained from a two-sided test of the z-statistic, calculated using logistic regression. The red line in each plot indicates Bonferroni-corrected significance.

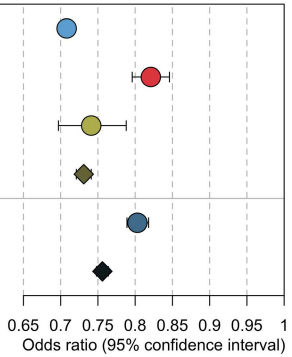


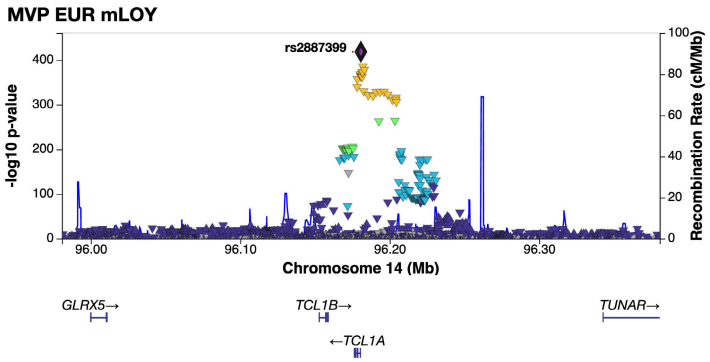
Supplementary Fig. 4. Quantile-quantile (QQ) plots of mLOY GWAS performed in this study. QQ plots show observed versus expected distributions of negative log transformed *P*-values for associations with mLOY in: **a**, Million Veteran Program (MVP) European ancestry (EUR); **b**, MVP admixed African ancestry (AFR); **c**, MVP Hispanic ancestry (HIS); **d**, High cell fraction (CF>10%) MVP EUR; **e**, High CF MVP AFR; **f**, High CF MVP HIS.

a

rs2887399-T (TCL1A)

Cohort	EA freq.	P-value	Odds ratio (95% c.i.)
MVP European	0.19	3.18E-419	0.708 (0.697, 0.719)
MVP African	0.39	9.40E-38	0.821 (0.796, 0.846)
MVP Hispanic	0.19	6.42E-22	0.741 (0.697, 0.788)
MVP multi-ancestry meta-analysis	0.23	5.25E-459	0.731 (0.721, 0.741)
UK Biobank (Thompson et al., 2019)	0.20	4.30E-136	0.803 (0.789, 0.818)
MVP+UKB meta-analysis	0.22	7.93E-573	0.756 (0.748, 0.764)

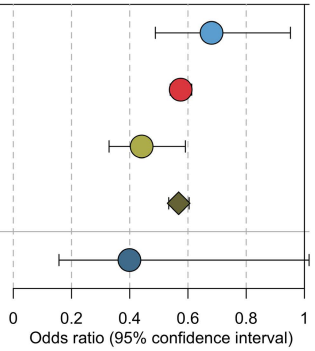


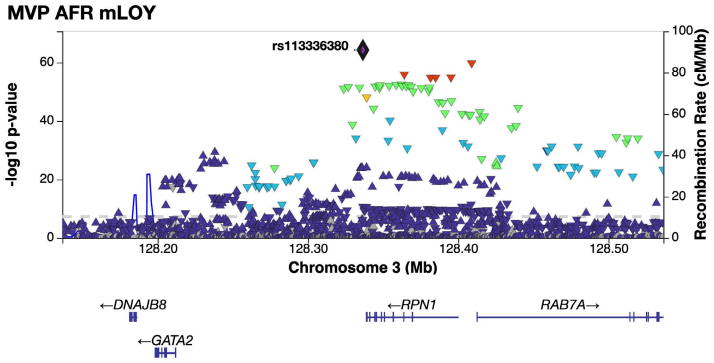


b

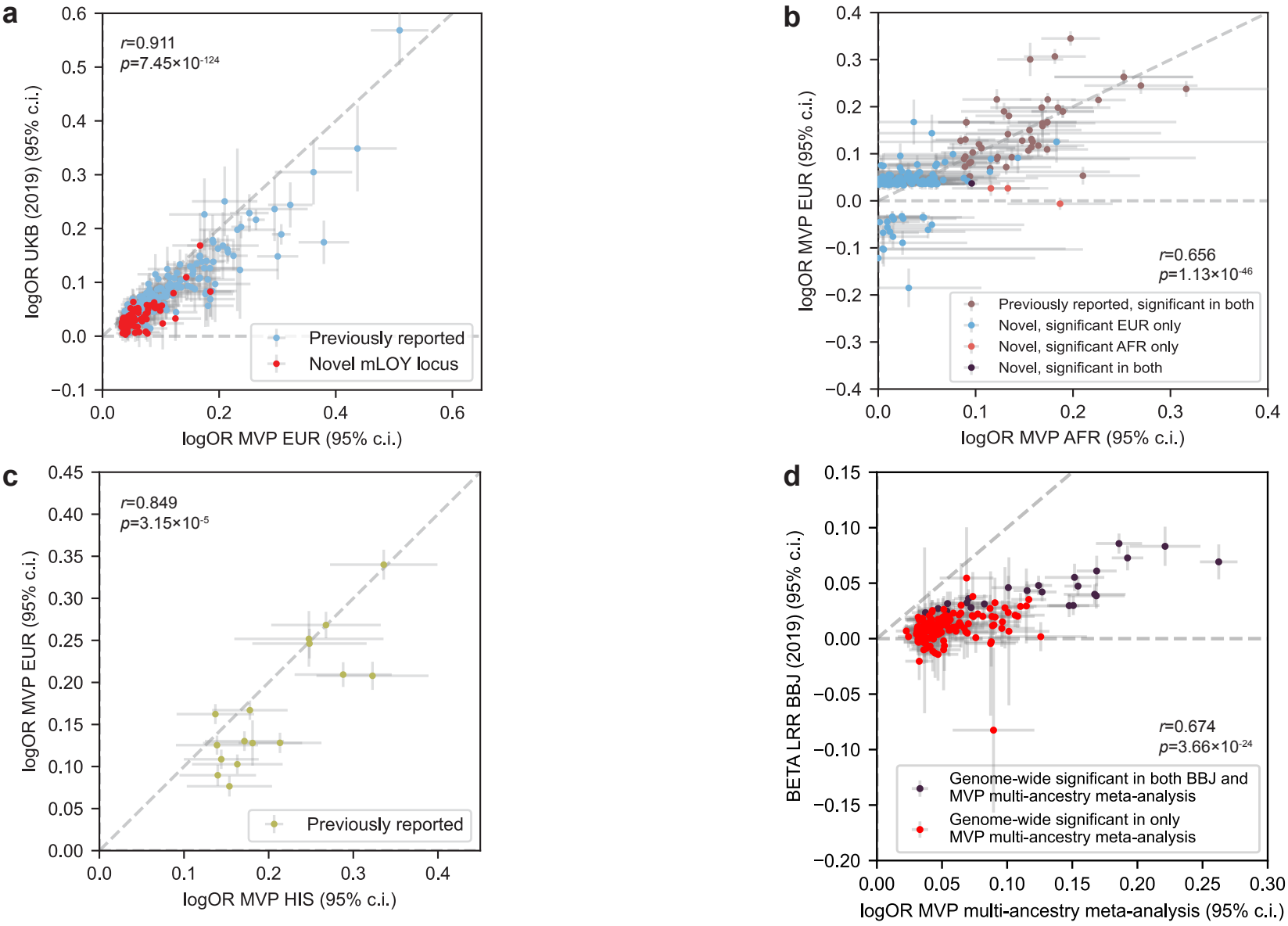
rs113336380-A (RPN1)

Cohort	EA freq.	P-value	Odds ratio
MVP European	0.00067	0.025	0.681 (0.488, 0.952)
MVP African	0.0624	5.36E-65	0.575 (0.539, 0.613)
MVP Hispanic	0.0078	4.69E-08	0.441 (0.329, 0.591)
MVP multi-ancestry meta-analysis	0.06	7.59E-71	0.568 (0.534, 0.604)
UK Biobank (Thompson et al., 2019)	7e-05	0.043	0.399 (0.157, 1.016)





Supplementary Fig. 5. Forest plots and regional Manhattan plots for most significant mosaic loss of Y associations. a, rs2887399, the top SNP at *TCL1A*; **b**, rs113336380, the top African ancestry SNP at *RPN1*. Left: forest plots showing frequency of effect allele, *P*-value, and Odds ratio for each Million Veteran Program (MVP) cohort and multi-ancestry meta-analysis, plus replication cohort UK Biobank[1] (UKB) and MVP+UKB meta-analysis. Right: Regional Manhattan plots for the MVP ancestry in which the SNP was most significant. rs113336380 was not meta-analyzed with UKB due to low frequency (MAF<0.001).

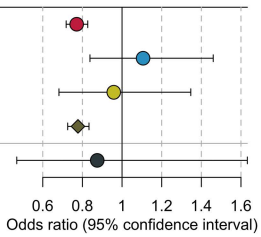


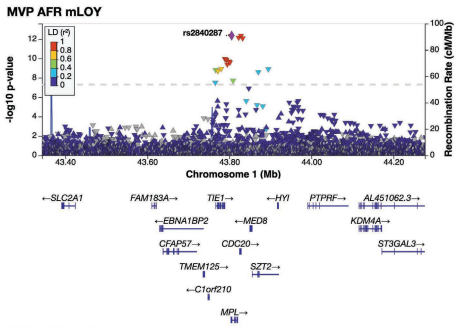
Supplementary Fig. 6. Effect comparisons between cohorts. Scatter plots comparing effect sizes (logarithm of odds ratios and 95% confidence interval) of genome-wide significant conditionally independent hits. **a**, UK Biobank[1] versus Million Veteran Program (MVP) European (EUR), indicating previously reported loci (blue) and novel loci (pink). **b**, MVP EUR versus MVP African ancestry (AFR), indicating previously reported loci, novel loci, and novel loci exclusive to EUR or AFR. **c**, MVP EUR versus MVP Hispanic ancestry (HIS), where all HIS signals were also found in EUR. **d**, MVP multi-ancestry meta-analysis versus BioBank Japan (BBJ)[2], indicating variants that were genome-wide significant in both studies or only MVP. BBJ effect is measured using log R ratio (LRR) intensity thresholding.

a

rs2840287 (MPL)

Cohort	EA freq.	P-value	Odds ratio
MVP African	0.0440	4.41E-13	0.77
MVP European	0.0004	0.476	1.11
MVP Hispanic	0.0046	0.806	0.96
MVP multi-ancestry meta	0.0426	9.12E-13	0.78
UK Biobank (White European)	0.0001	0.650	0.87

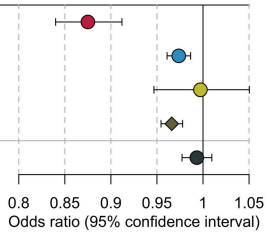


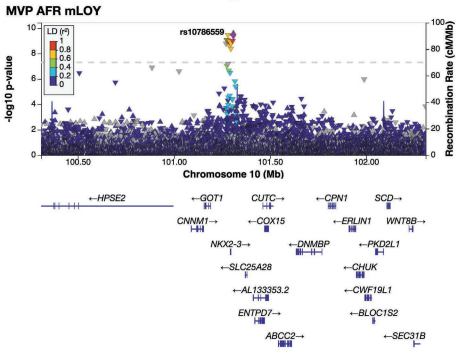


b

rs10786559 (NKX2-3)

Cohort	EA freq.	P-value	Odds ratio
MVP African	0.14	3.32E-10	0.88
MVP European	0.25	6.86E-05	0.97
MVP Hispanic	0.25	0.915	1.00
MVP multi-ancestry meta	0.24	2.75E-08	0.97
UK Biobank (White European)	0.26	0.600	0.99

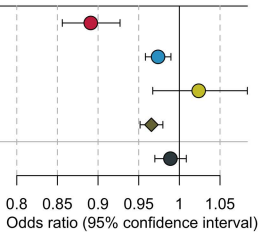


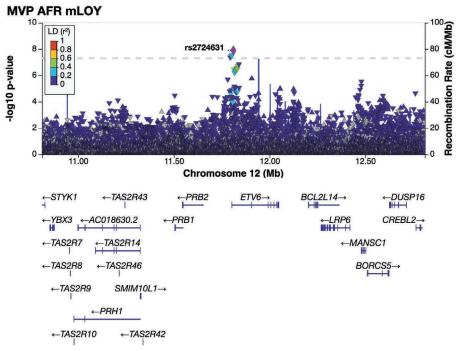


c

rs2724631 (ETV6)

Cohort	EA freq.	P-value	Odds ratio
MVP African	0.20	1.26E-08	0.89
MVP European	0.82	0.001	0.97
MVP Hispanic	0.76	0.417	1.02
MVP multi-ancestry meta	0.73	2.00E-06	0.97
UK Biobank (White European)	0.84	0.280	0.99

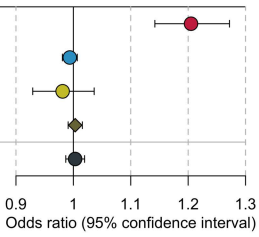


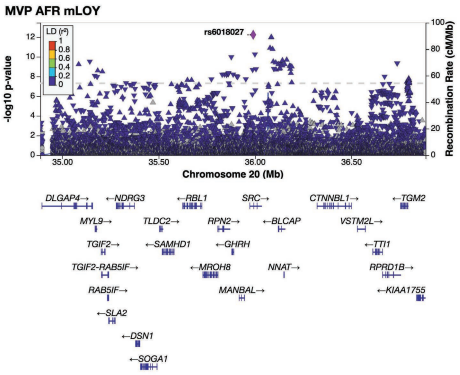


d

rs6018599 (BLCAP)

Cohort	EA freq.	P-value	Odds ratio
MVP African	0.07	1.27E-11	1.21
MVP European	0.26	0.348	0.99
MVP Hispanic	0.21	0.497	0.98
MVP multi-ancestry meta	0.25	0.617	1.00
UK Biobank (White European)	0.26	0.770	1.00

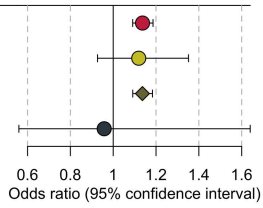


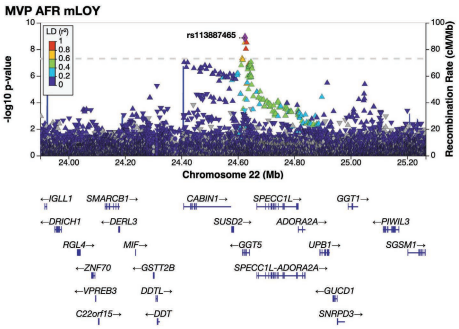


e

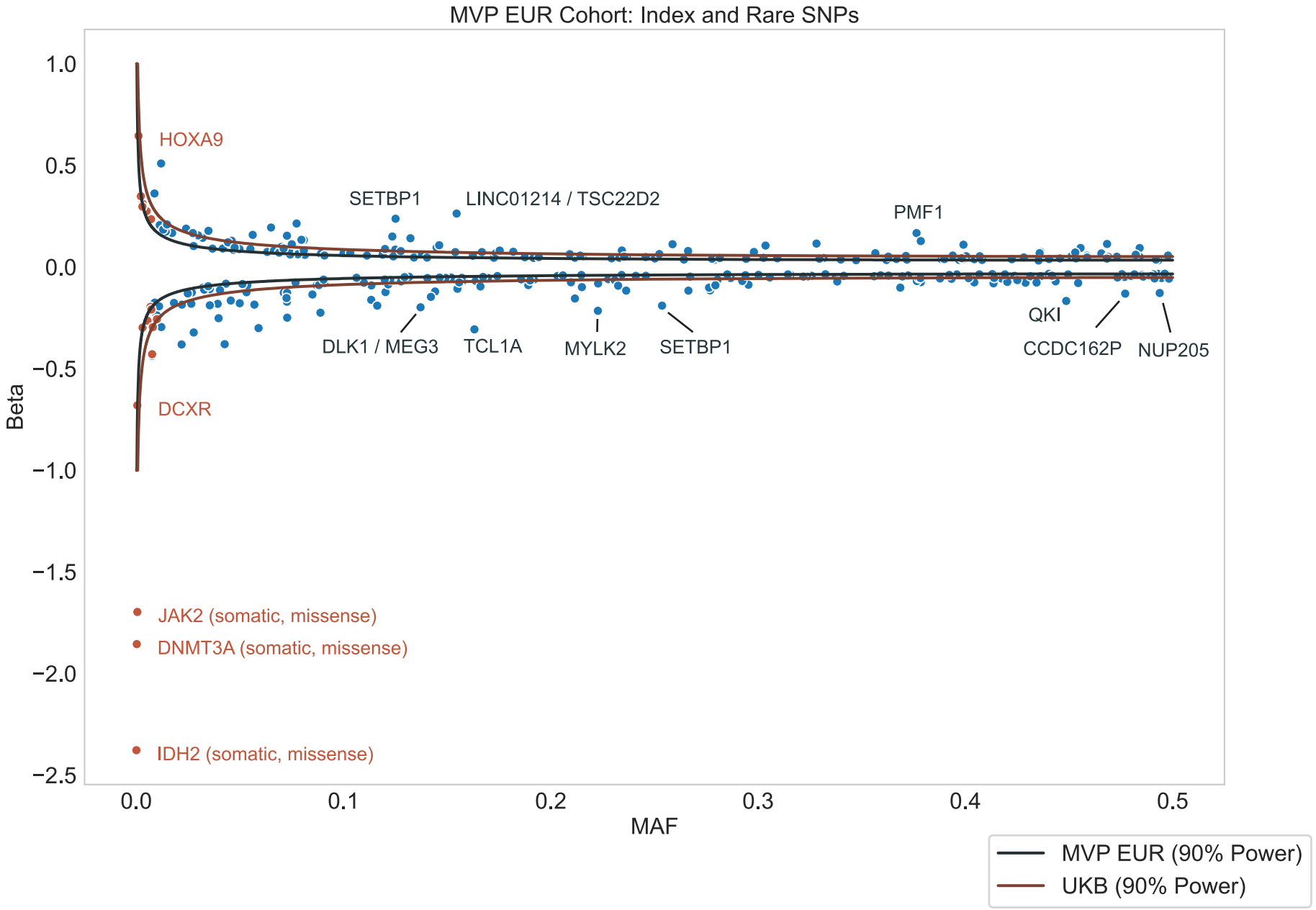
rs113887465 (GGT5/CABIN1)

Cohort	EA freq.	P-value	Odds ratio
MVP African	0.1390	1.37E-09	1.14
MVP Hispanic	0.0178	0.242	1.12
MVP multi-ancestry meta	0.1331	6.92E-10	1.14
UK Biobank (White European)	0.0002	0.930	0.96

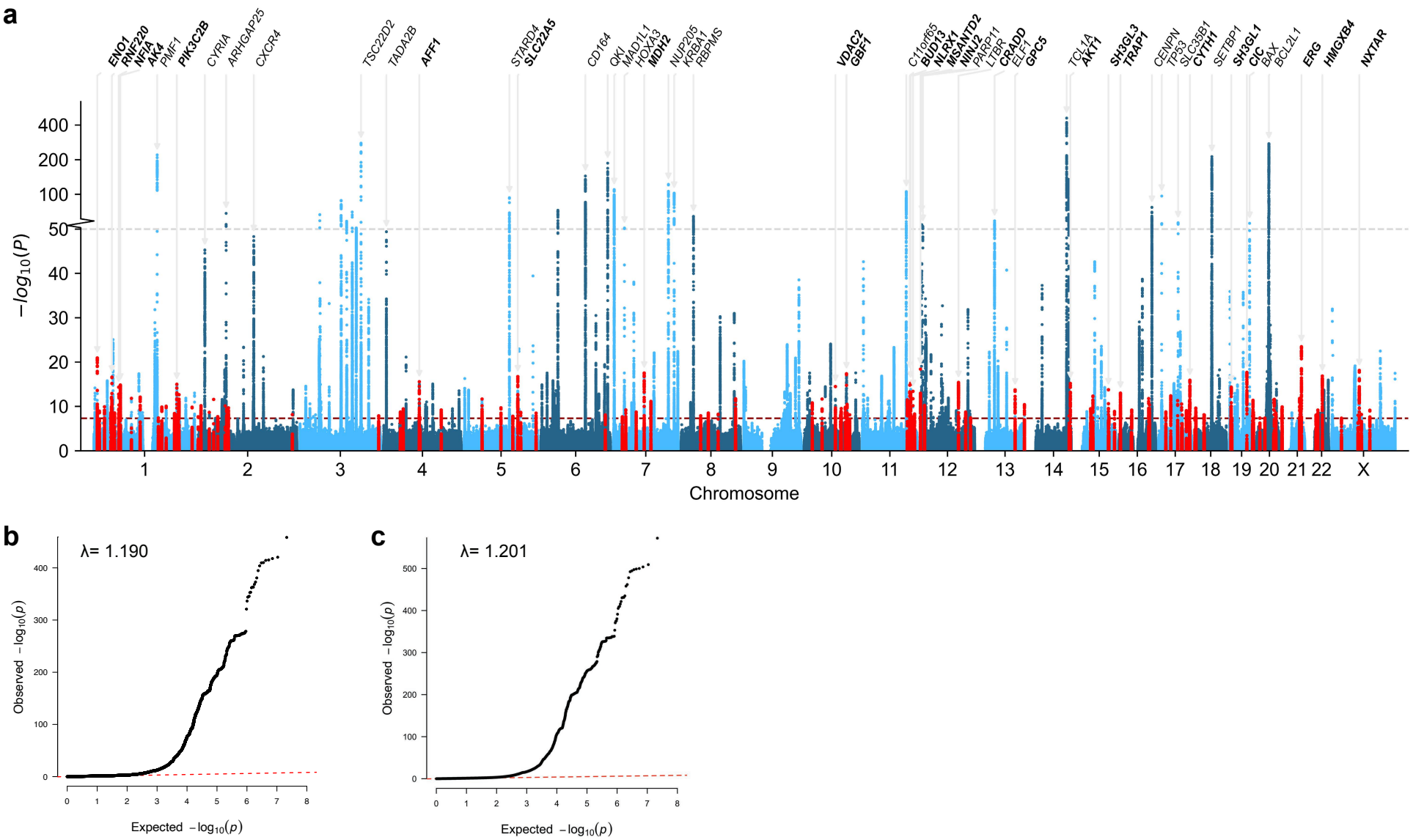




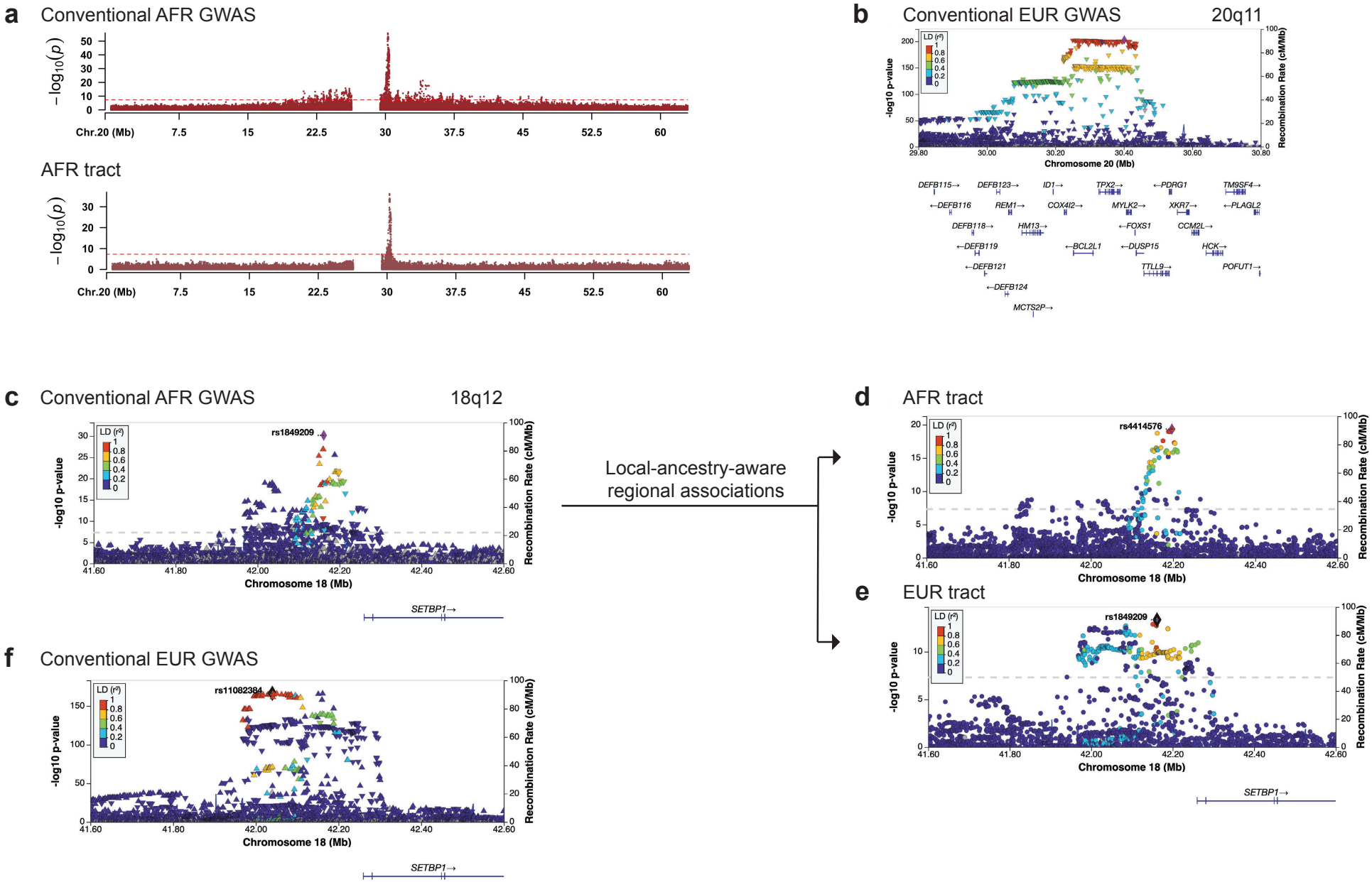
Supplementary Figure 7. Forest plots and regional Manhattan plots for African-ancestry-specific novel mLOY loci. Five novel mLOY loci which were only genome-wide significant in Million Veteran program African Ancestry (within 1Mb): **a,** *MPL*, **b,** *NKX2-3*, **c,** *ETV6*, **d,** *BLCAP*, and **e,** *GGT5*/*CABIN1* (not available in MVP European).



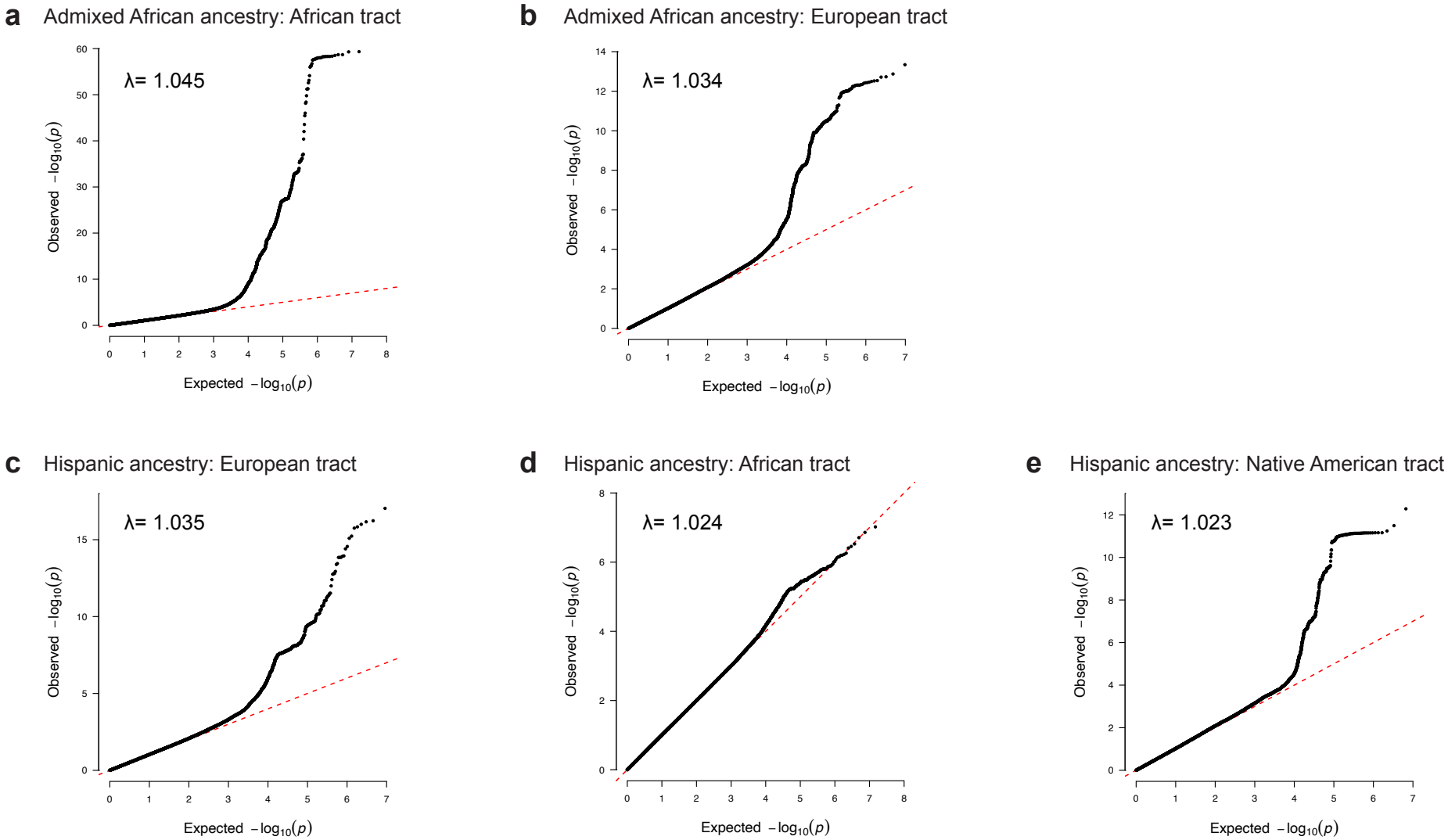
Supplementary Fig. 8. Relationship between minor allele frequency and effect size. For 327 genome-wide significant conditionally independent SNPs, plotting effect size (beta in log odds) versus minor allele frequency (MAF). Simulated power curves are shown for MVP and UKB[1] MAFs representing the expected magnitude of beta detectable at $P<5\times10^{-8}$. Significant rare variants (MAF<0.001 in MVP) are shown in orange. Gene names provided for SNPs with $P<1\times10^{-100}$.



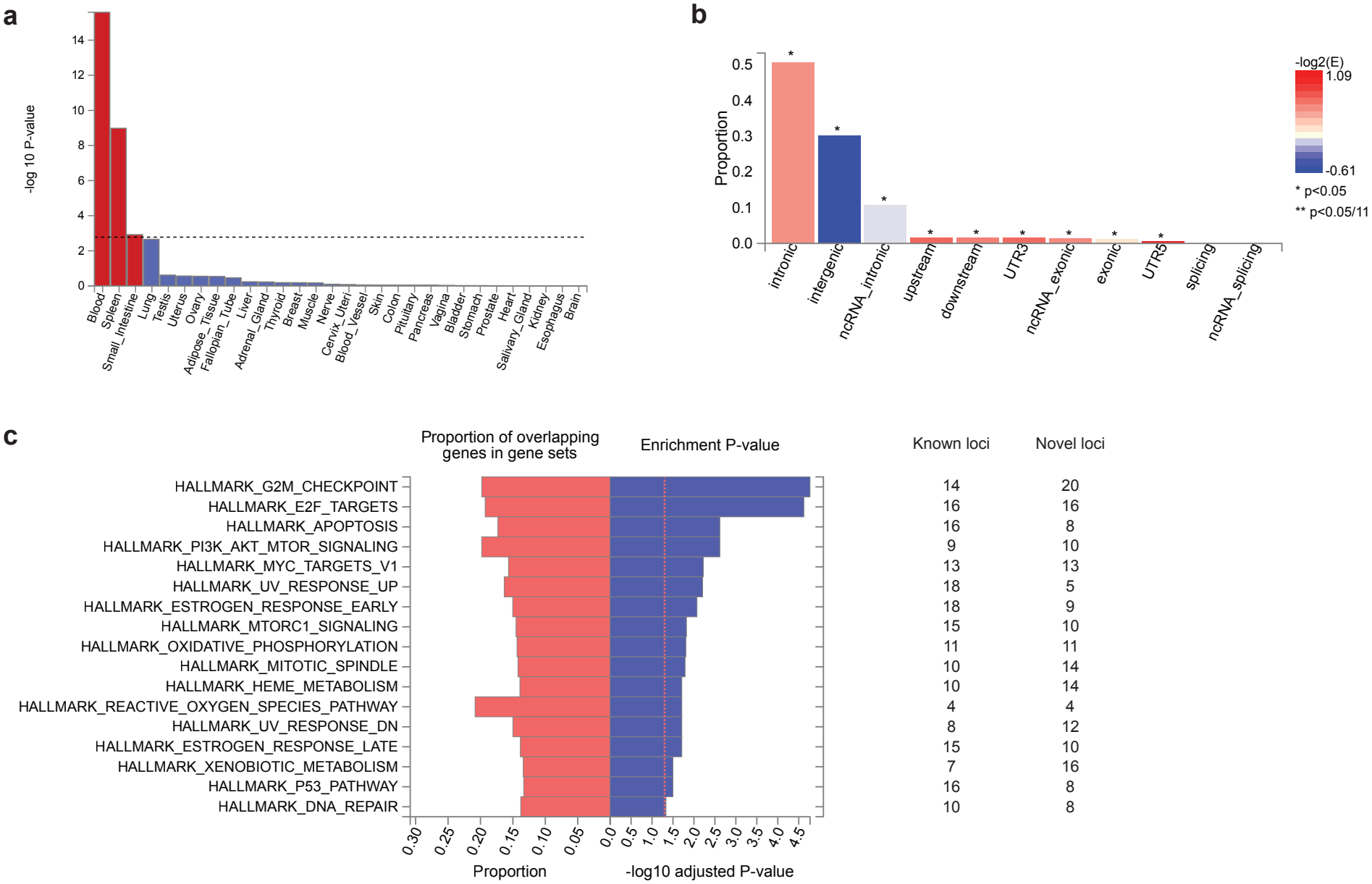
Supplementary Fig. 9. Manhattan and quantile-quantile (QQ) plots of Million Veteran Program (MVP) multi-ancestry meta-analysis. a, Manhattan plot shows the -log10(*P*) for associations of genetic variants with mLOY in the MVP meta-analysis of European, African, and Hispanic ancestries. Novel mLOY index 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.The 25 most significant known and novel loci are annotated with the nearest gene. **b**, QQ plot shows observed versus expected distributions of *P*-values in MVP multi-ancestry meta-analysis. **c**, QQ plot shows observed versus expected distributions of *P*-values in MVP and UK Biobank meta-analysis (corresponding to the Manhattan plot shown in Main Fig. 2d).



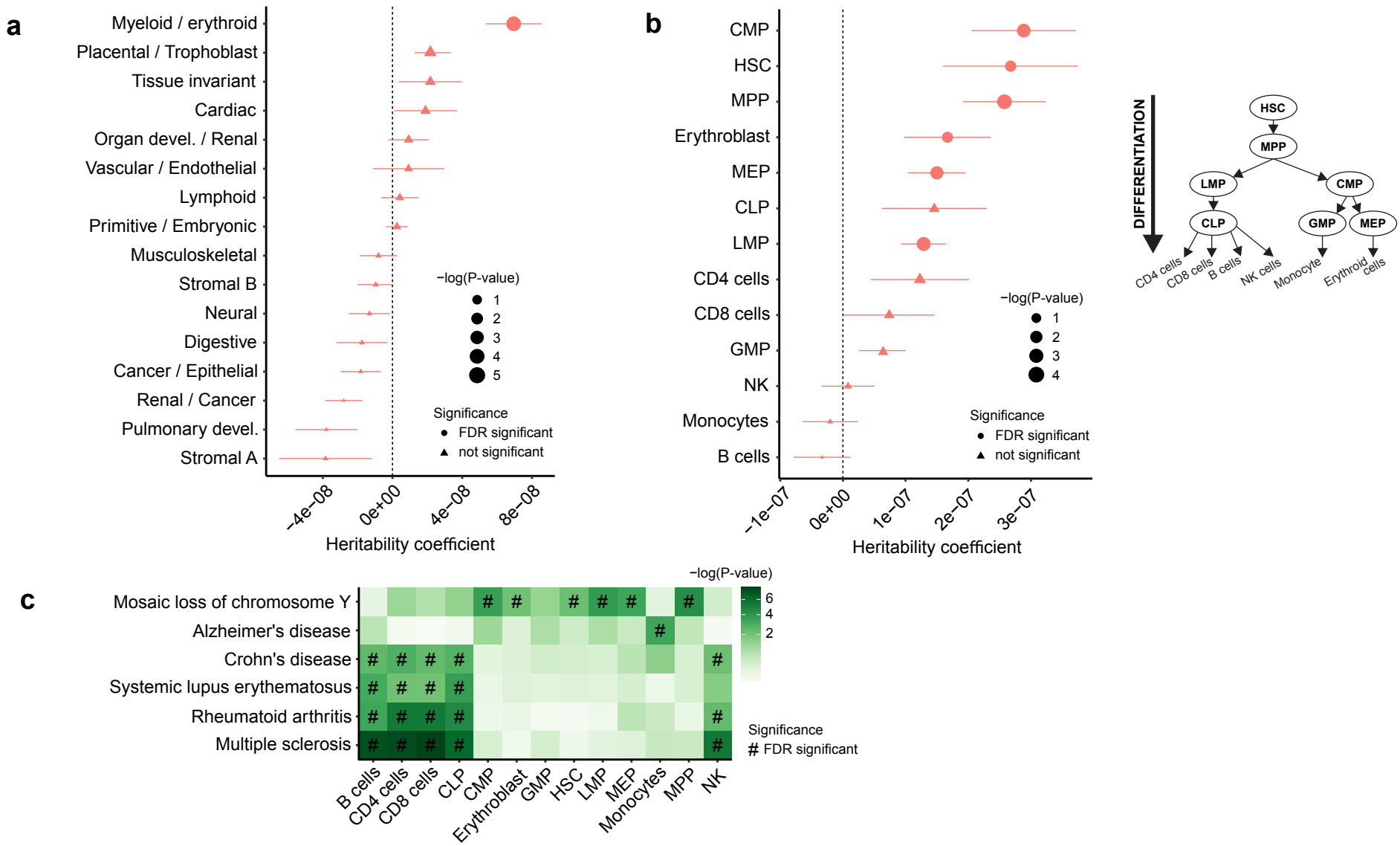
Supplementary Fig. 10. Supporting evidence in Million Veteran Program admixed African samples for local ancestry associations with mLOY. a, Chromosome 20 Manhattan plots for conventional admixed African (AFR) and AFR tract association *P*-values. **b,** Conventional European (EUR) regional Manhattan plot for the *BCL2L1* locus. **c,** At 18q12 (*SETBP1*), we observed complex linkage disequilibrium structure; this was resolved by deconvolving the **d,** AFR and **e,** EUR tracts. **f,** Regional Manhattan plot for *SETBP1* in conventional EUR association analysis.



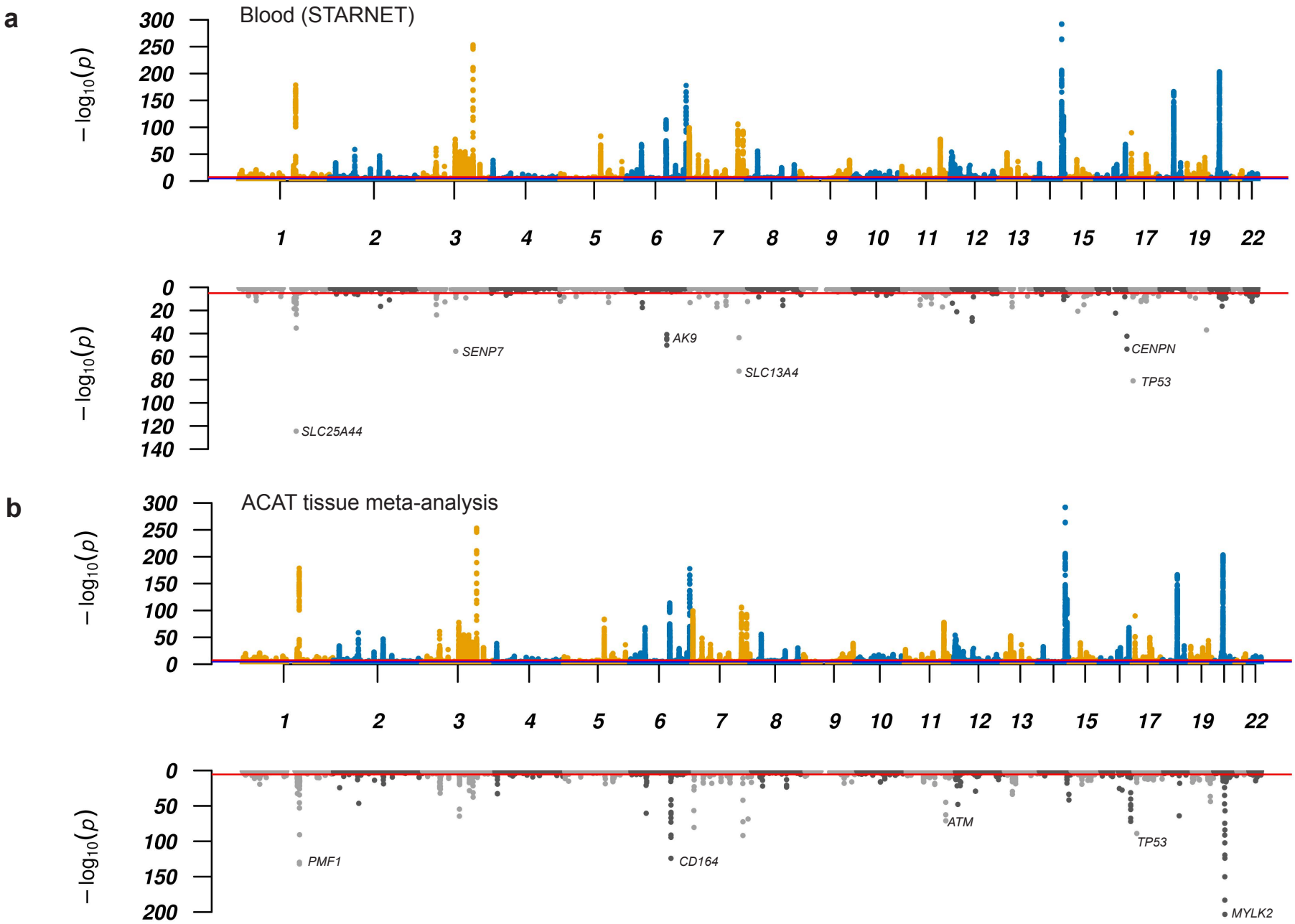
Supplementary Fig. 11. Local ancestry tract quantile-quantile (QQ) plots. QQ plots show observed versus expected distributions of *P*-values in Million Veteran Program (MVP) admixed African tracts for **a**, African, and **b**, European ancestral contributions; and in MVP Hispanic tracts for **c**, European, **d**, African, and **e**, Native American ancestral contributions.



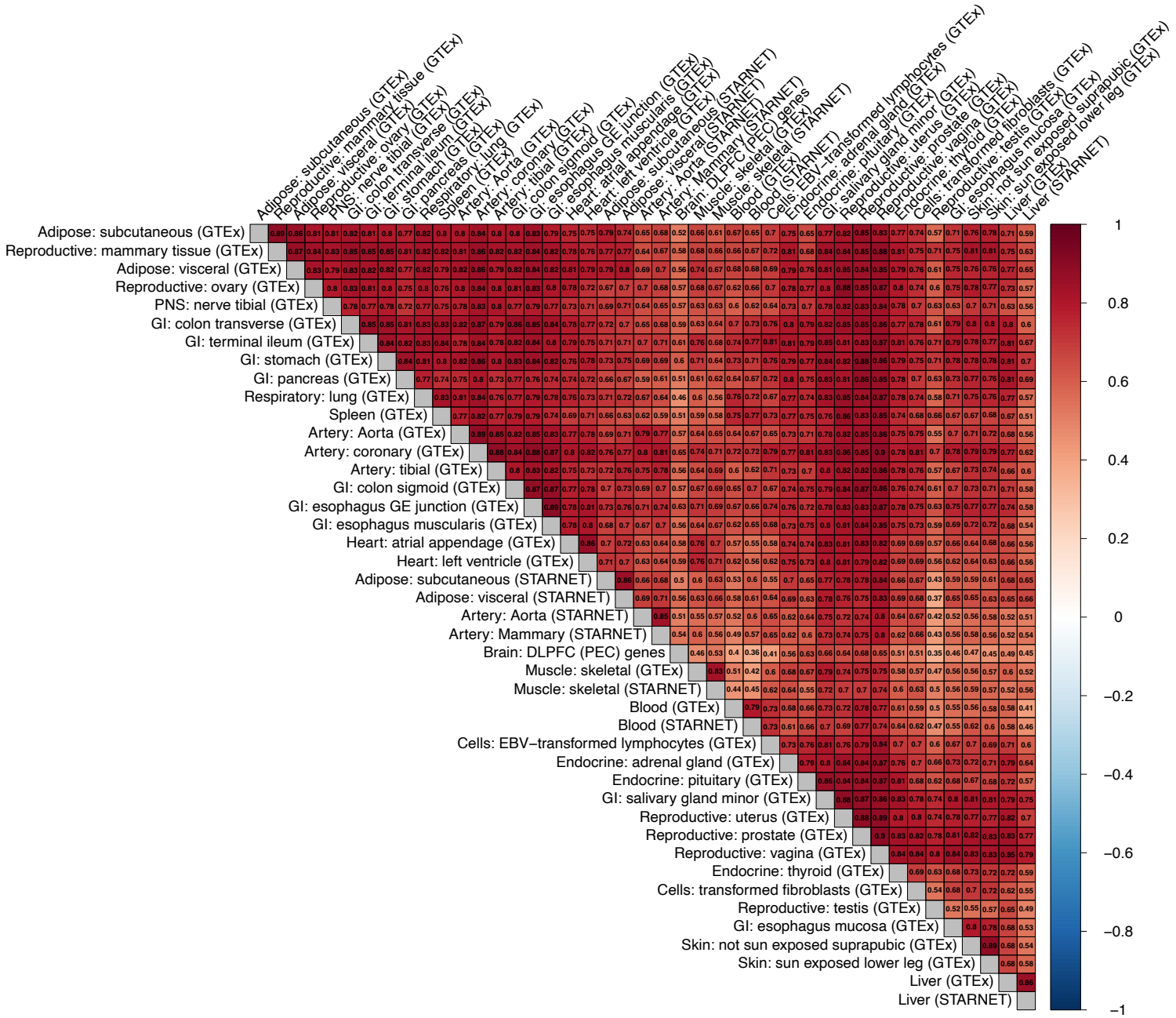
Supplementary Fig. 12. Gene set enrichment of mLOY associations. For genes positionally mapped to significant mLOY loci in the Million Veteran Program + UK Biobank meta-analysis: **a,** -$\log_{10}(P)$ values for enrichment of GTEx v8 tissue types. **b**, Functional consequences of SNPs on genes, and **c,** FDR adjusted -$\log_{10}(P)$ for enrichment of Hallmark gene sets, showing counts of known and novel loci in each gene set.



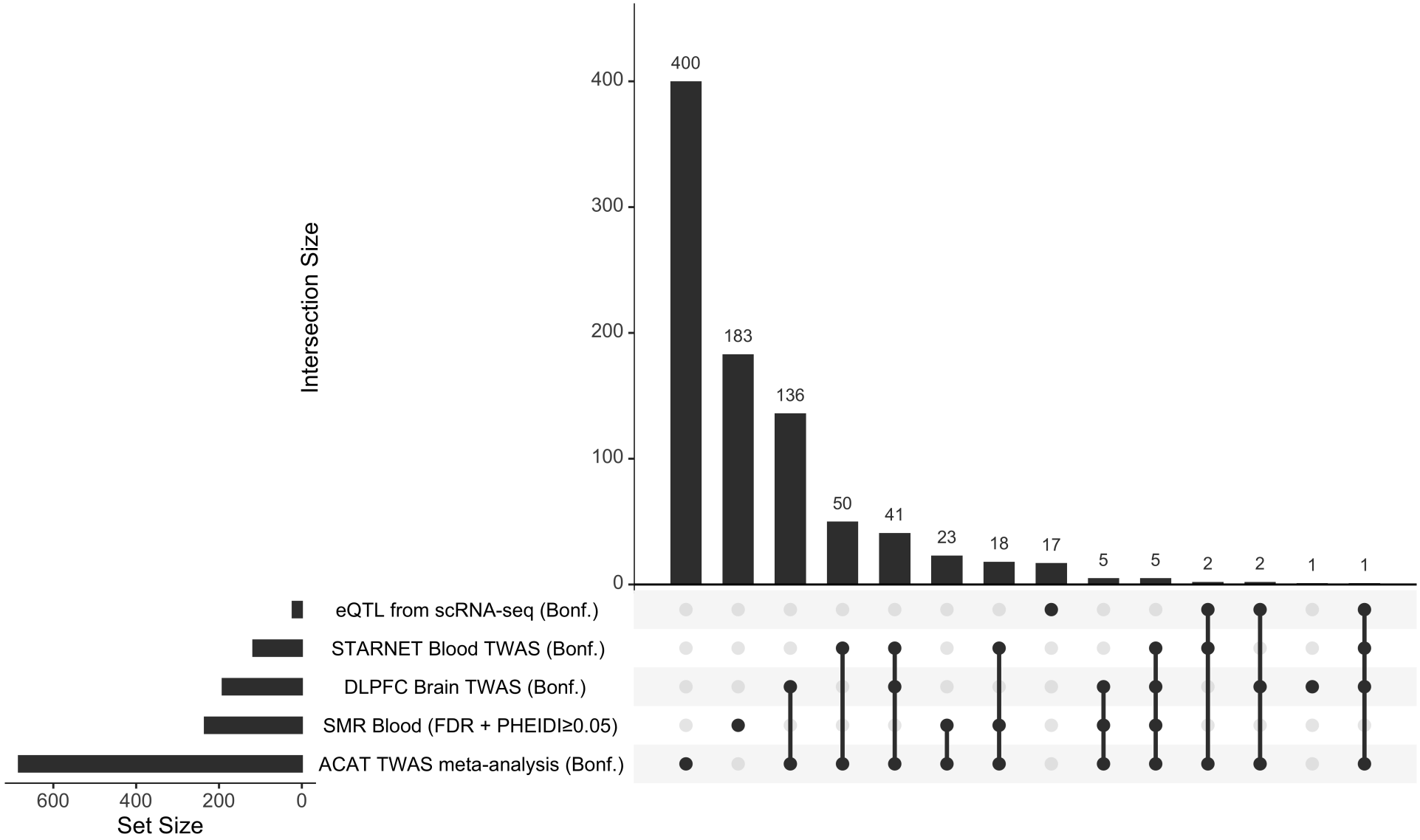
Supplementary Fig. 13. Genomic enrichment of mLOY risk variants in open chromatin regions across human cell types. (a-b) Enrichment of mLOY risk variants in cell-specific open chromatin regions from **a,** main human body lineages and **b,** blood lineages. Symbols (circle/triangle) represent the LD score coefficient and horizontal bars reflect standard error. The positive LD score coefficient signifies enrichment in heritability (horizontal bars reflect standard error). Symbol size reflects the value of LD score regression. (Right) Diagram illustrating the hierarchical structure of the human hematopoietic system, highlighting the blood lineages from the atlas. **c,** Comparison of enrichment of blood lineages across a panel of diseases. CLP, common lymphoid progenitor cells; CMP, common myeloid progenitor cells; GMP, granulocyte-macrophage progenitor cells; HSC, hematopoietic stem cells; LMP, lymphoid-primed progenitor cells; MEP, megakaryote-erythroid progenitor cells; MMP, multipotent progenitor cells; NK, natural killer cells.



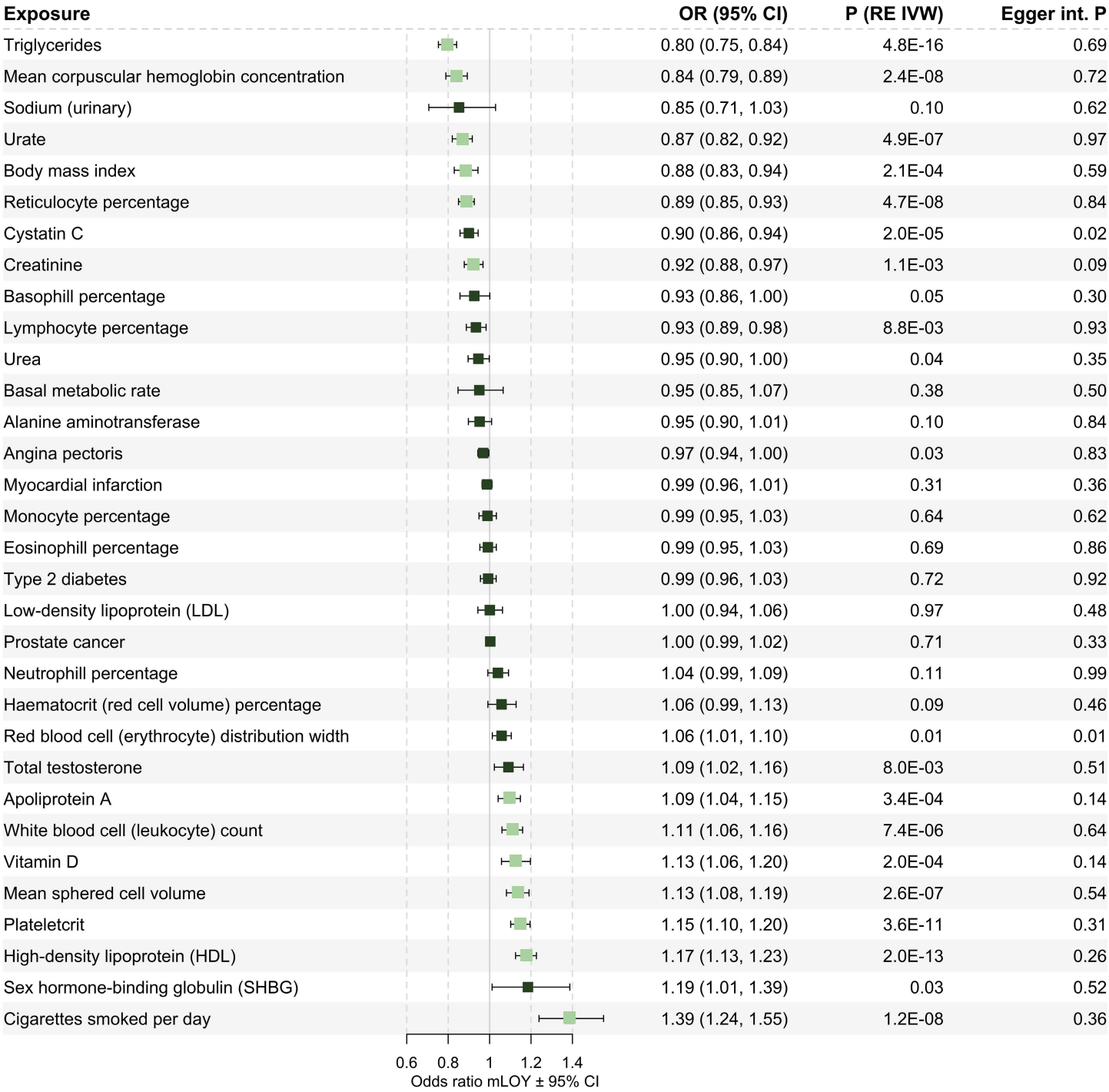
Supplementary Fig. 14. Global Miami plots for transcriptome-wide association study. Miami plots with Million Veteran Program European participants -log10(*P*) for associations of genetic variants from GWAS (top) and -log10(*P*) for associations of gene features from TWAS (bottom). Plots show TWAS results for **a**, Blood (STARNET) and **b**, ACAT tissue meta-analysis.



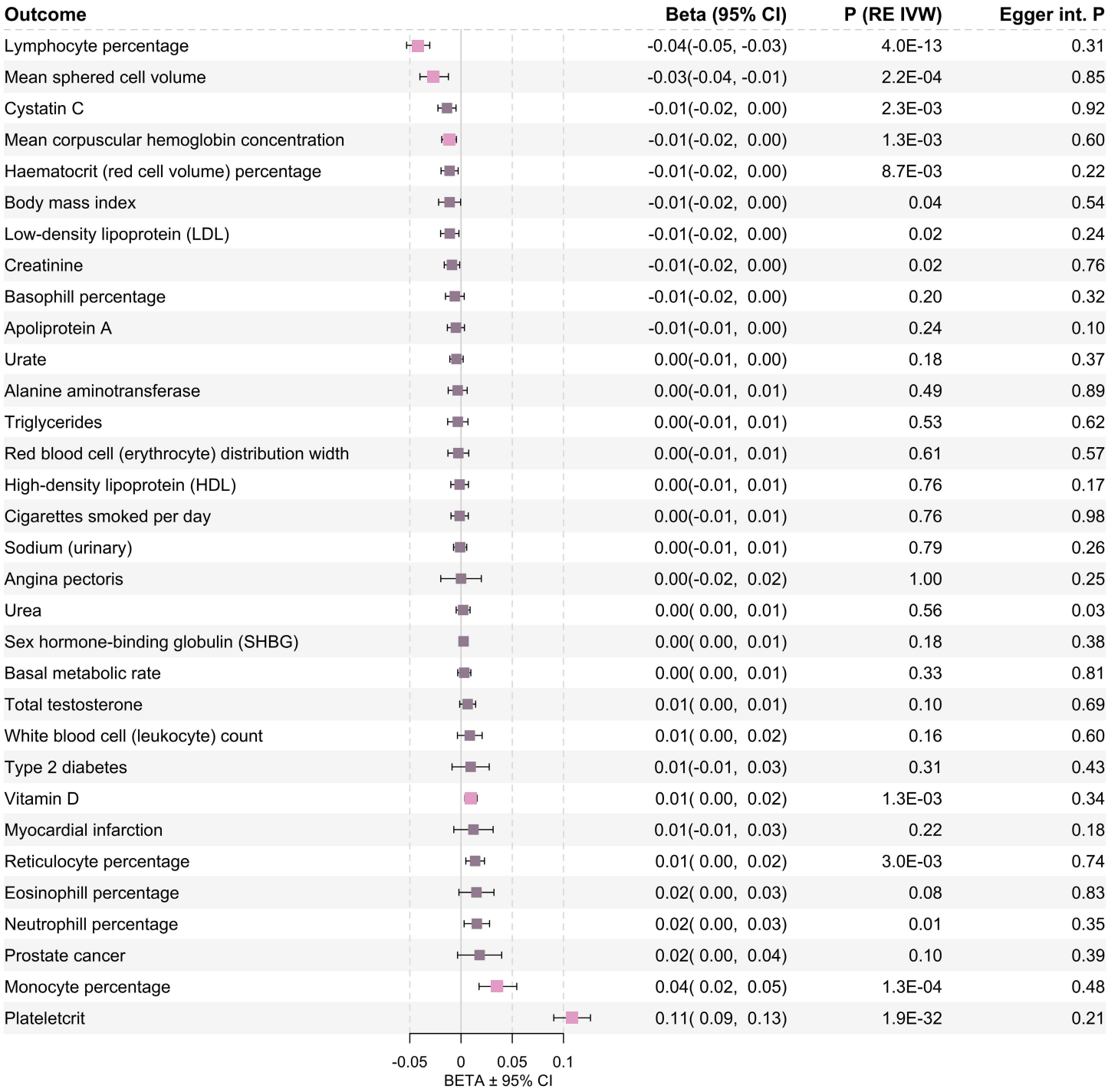
Supplementary Figure 15. Correlation of gene expression across tissues from transcriptome-wide association study (TWAS). Values represent Pearson correlation coefficient for imputed expression changes.



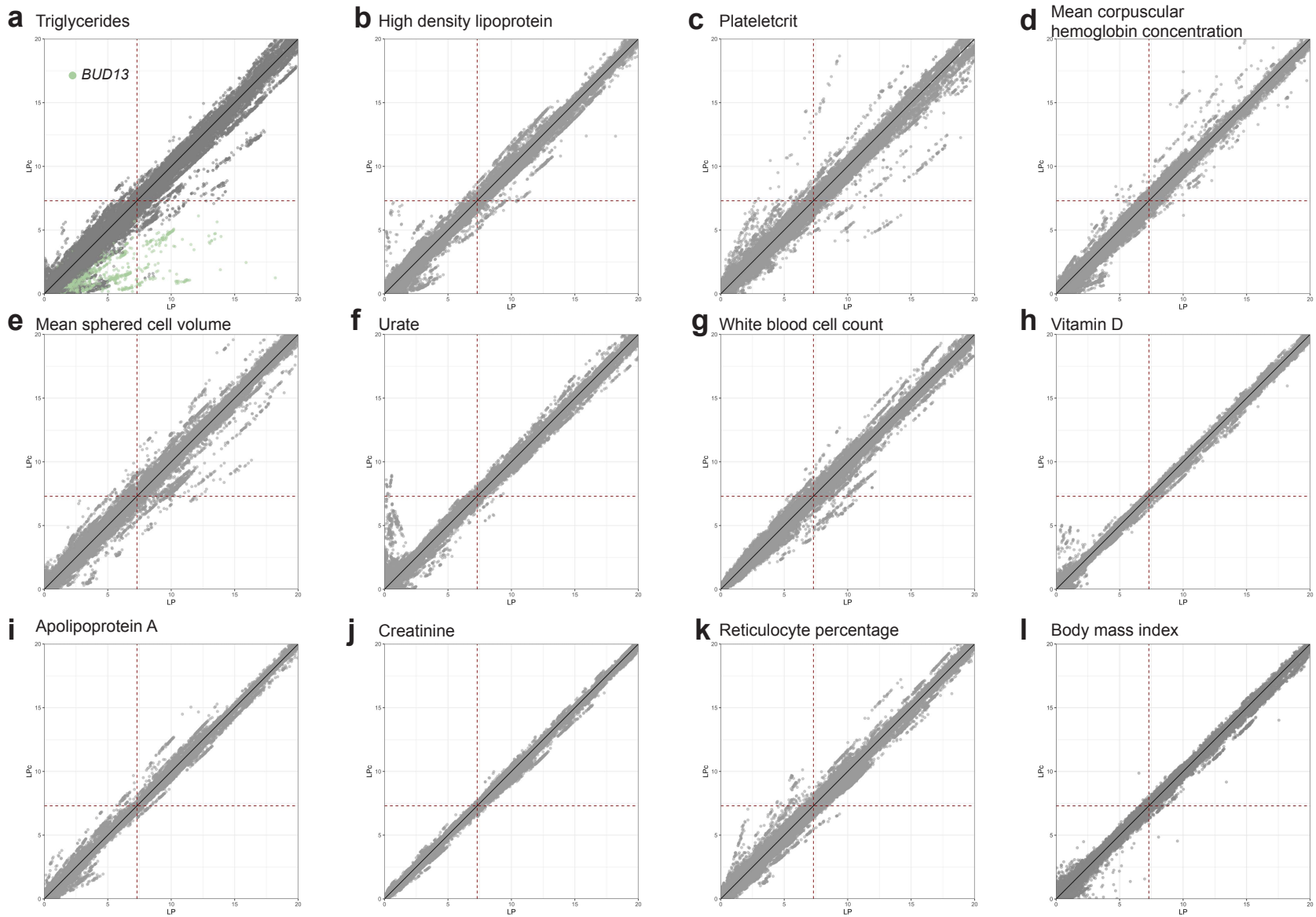
Supplementary Fig. 16. UpSet plot showing overlap between gene transcript association methods. Genes are derived from eQTLs with single cell RNA-seq (scRNA-seq) data across different immune cell populations, STARNET Blood transcriptome-wide association study (TWAS), dorsolateral prefrontal cortex TWAS, Summary Mendelian Randomization (SMR), and ACAT TWAS meta-analysis. Bonf., Bonferroni-corrected significance; FDR, false discovery rate.



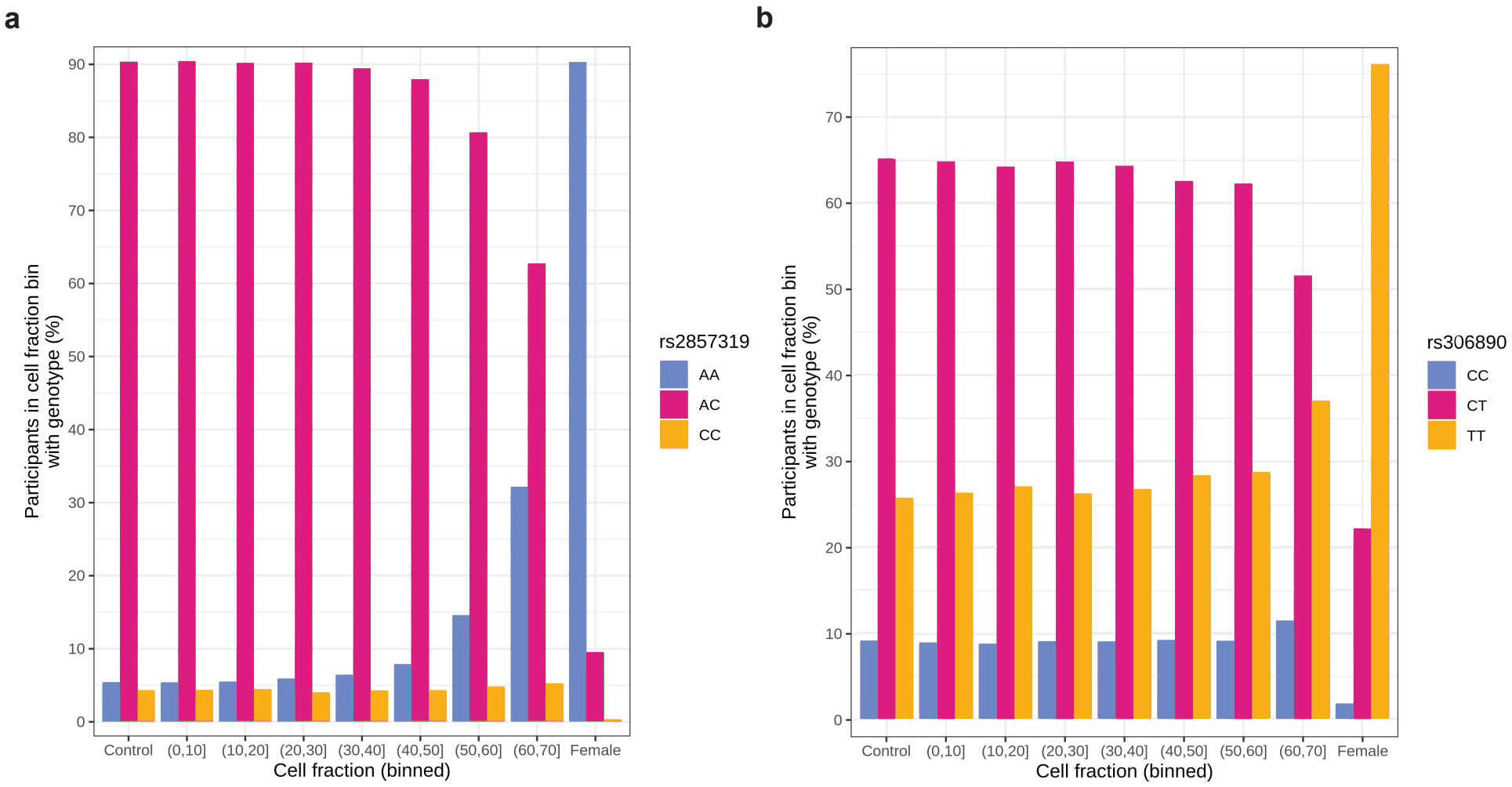
Supplementary Fig. 17. Forward Mendelian Randomization (MR) for exposure traits on mLOY. Forest plots showing random effects inverse-variance weighted (RE IVW) MR for 32 exposure traits on mLOY outcome in Million Veteran Program Europeans. Lighter box color denotes the exposure reached Bonferroni-corrected significance in RE IVW (P<0.05/32) and P MR-Egger intercept (int.)≥0.05. OR, odds ratio; CI, confidence interval.



Supplementary Fig. 18. Reverse Mendelian Randomization (MR) for mLOY exposure on all trait outcomes. Forest plots showing random effects inverse-variance weighted (RE IVW) MR for mLOY exposure on 32 trait outcomes in Million Veteran Program Europeans. Lighter box color denotes the exposure reached Bonferroni-corrected significance in RE IVW (P<0.05/32) and P MR-Egger intercept (int.)≥0.05. OR, odds ratio; CI, confidence interval.



Supplementary Fig. 19. Multi-trait meta-analysis (mtCOJO) conditional analysis on additional significant forward Mendelian Randomization (MR) traits. Scatterplot of -log(*P*) from conditioning on significant MR traits using mtCOJO (LPc) vs. -log(*P*) from standard mLOY GWAS (LP) in MVP European ancestry. Red dashed lines indicate genome-wide significance at $P<5\times10^{-8}$. *BUD13* gene region is highlighted in trigylcerides (chr11:116,519,358-117,091,609).



Supplementary Fig. 20. Rejected chromosome X associations. Two genome-wide significant variants on chromosome X, **(a)** rs2857319 and **(b)** rs306890, near the borders of PAR1 and PAR2, respectively, were found to be likely false positives. Plotting the percentage of participants with each genotype across cell fraction bins reveals a heterozygous dropout in higher cell fractions, suggesting genotyping error.

Supplementary References

1. Thompson, D. J. *et al.* Genetic predisposition to mosaic Y chromosome loss in blood. *Nature* **575**, 652–657 (2019).
2. Terao, C. *et al.* GWAS of mosaic loss of chromosome Y highlights genetic effects on blood cell differentiation. *Nat. Commun.* **10**, 4719 (2019).

Million Veteran Program: Consortium Acknowledgement for Manuscripts

MVP Executive Committee

- Co-Chair: J. Michael Gaziano, M.D., M.P.H.
- Co-Chair: Rachel Ramoni, D.M.D., Sc.D.
- Jim Breeling, M.D. (ex-officio)
- Kyong-Mi Chang, M.D.
- Grant Huang, Ph.D.
- Sumitra Muralidhar, Ph.D.
- Christopher J. O'Donnell, M.D., M.P.H.
- Philip S. Tsao, Ph.D.

MVP Program Office

- Sumitra Muralidhar, Ph.D.
- Jennifer Moser, Ph.D.

MVP Recruitment/Enrollment

- Recruitment/Enrollment Director/Deputy Director, Boston
- Stacey B. Whitbourne, Ph.D.; Jessica V. Brewer, M.P.H.
- MVP Coordinating Centers
    - o Clinical Epidemiology Research Center (CERC), West Haven – John Concato, M.D., M.P.H.
    - o Cooperative Studies Program Clinical Research Pharmacy Coordinating Center, Albuquerque - Stuart Warren, J.D., Pharm D.; Dean P. Argyres, M.S.
    - o Genomics Coordinating Center, Palo Alto – Philip S. Tsao, Ph.D.
    - o Massachusetts Veterans Epidemiology Research Information Center (MAVERIC), Boston - J. Michael Gaziano, M.D., M.P.H.
    - o MVP Information Center, Canandaigua – Brady Stephens, M.S.
- Core Biorepository, Boston – Mary T. Brophy M.D., M.P.H.; Donald E. Humphries, Ph.D.
- MVP Informatics, Boston – Nhan Do, M.D.; Shahpoor Shayan
- Data Operations/Analytics, Boston – Xuan-Mai T. Nguyen, Ph.D.

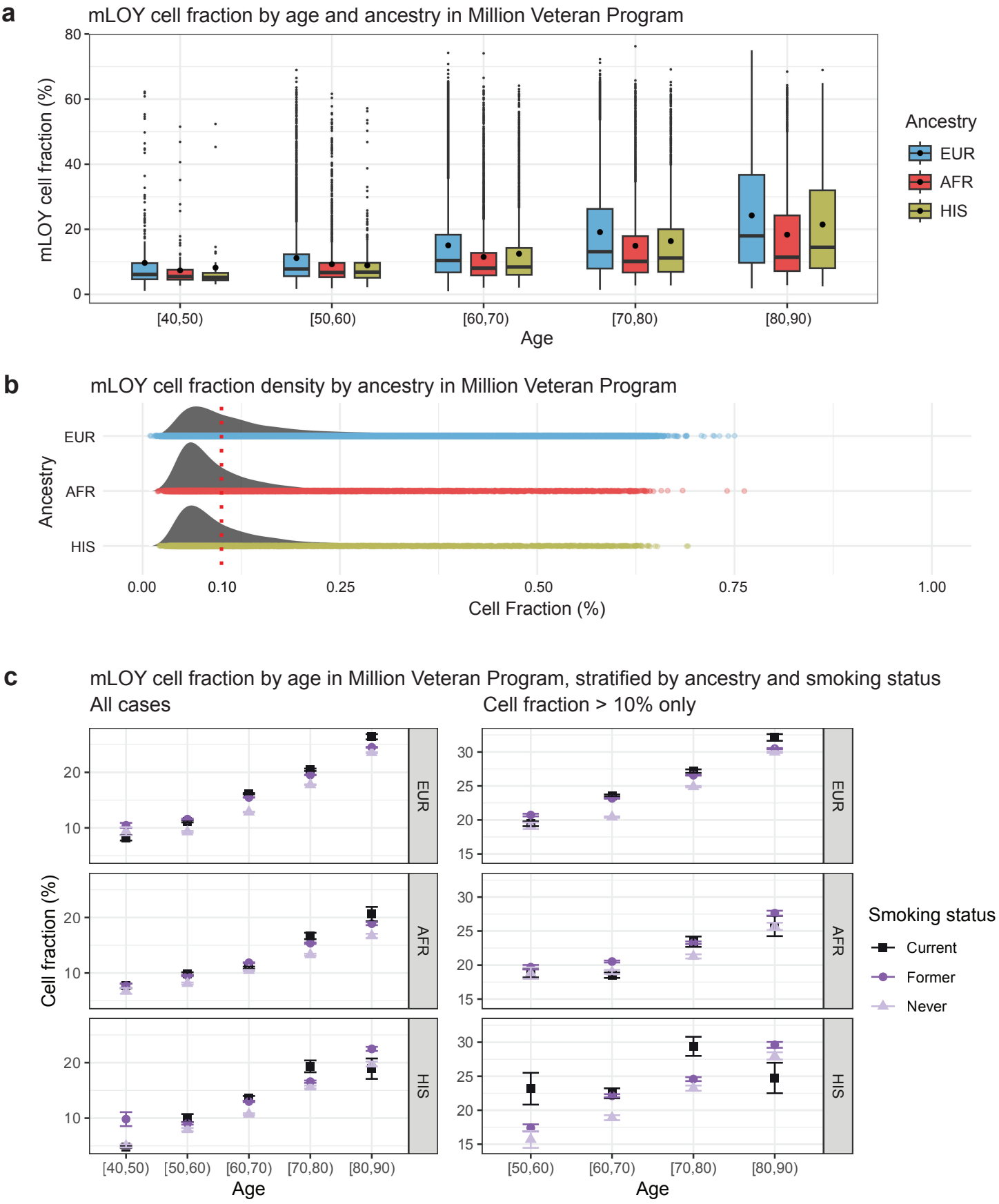
MVP Science

- Genomics - Christopher J. O'Donnell, M.D., M.P.H.; Saiju Pyarajan Ph.D.; Philip S. Tsao, Ph.D.
- Phenomics - Kelly Cho, M.P.H, Ph.D.
- Data and Computational Sciences – Saiju Pyarajan, Ph.D.
- Statistical Genetics – Elizabeth Hauser, Ph.D.; Yan Sun, Ph.D.; Hongyu Zhao, Ph.D.

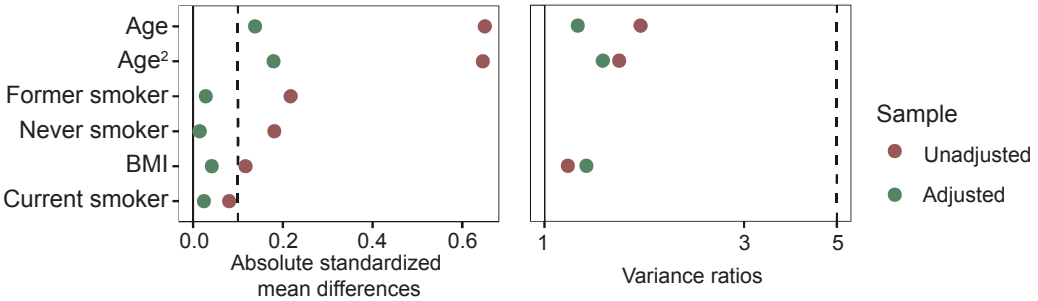
MVP Local Site Investigators

- Atlanta VA Medical Center (Peter Wilson) - Bay Pines VA Healthcare System (Rachel McArdle)
- Birmingham VA Medical Center (Louis Dellitalia)
- Cincinnati VA Medical Center (John Harley)
- Clement J. Zablocki VA Medical Center (Jeffrey Whittle)
- Durham VA Medical Center (Jean Beckham)
- Edith Nourse Rogers Memorial Veterans Hospital (John Wells)
- Edward Hines, Jr. VA Medical Center (Salvador Gutierrez)
- Fayetteville VA Medical Center (Gretchen Gibson)

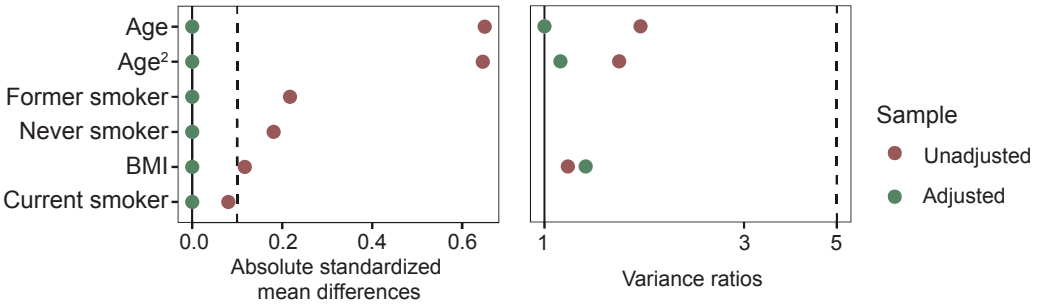
- VA Health Care Upstate New York (Laurence Kaminsky)
- New Mexico VA Health Care System (Gerardo Villareal)
- VA Boston Healthcare System (Scott Kinlay)
- VA Western New York Healthcare System (Junzhe Xu)
- Ralph H. Johnson VA Medical Center (Mark Hamner)
- Wm. Jennings Bryan Dorn VA Medical Center (Kathlyn Sue Haddock)
- VA North Texas Health Care System (Sujata Bhushan)
- Hampton VA Medical Center (Pran Iruvanti)
- Hunter Holmes McGuire VA Medical Center (Michael Godschalk)
- Iowa City VA Health Care System (Zuhair Ballas)
- Jack C. Montgomery VA Medical Center (Malcolm Buford)
- James A. Haley Veterans' Hospital (Stephen Mastorides)
- Louisville VA Medical Center (Jon Klein)
- Manchester VA Medical Center (Nora Ratcliffe)
- Miami VA Health Care System (Hermes Florez)
- Michael E. DeBakey VA Medical Center (Alan Swann)
- Minneapolis VA Health Care System (Maureen Murdoch)
- N. FL/S. GA Veterans Health System (Peruvemba Sriram)
- Northport VA Medical Center (Shing Shing Yeh)
- Overton Brooks VA Medical Center (Ronald Washburn)
- Philadelphia VA Medical Center (Darshana Jhala)
- Phoenix VA Health Care System (Samuel Aguayo)
- Portland VA Medical Center (David Cohen)
- Providence VA Medical Center (Satish Sharma)
- Richard Roudebush VA Medical Center (John Callaghan)
- Salem VA Medical Center (Kris Ann Oursler)
- San Francisco VA Health Care System (Mary Whooley)
- South Texas Veterans Health Care System (Sunil Ahuja)
- Southeast Louisiana Veterans Health Care System (Amparo Gutierrez)
- Southern Arizona VA Health Care System (Ronald Schifman)
- Sioux Falls VA Health Care System (Jennifer Greco)
- St. Louis VA Health Care System (Michael Rauchman)
- Syracuse VA Medical Center (Richard Servatius)
- VA Eastern Kansas Health Care System (Mary Oehlert)
- VA Greater Los Angeles Health Care System (Agnes Wallbom)
- VA Loma Linda Healthcare System (Ronald Fernando)
- VA Long Beach Healthcare System (Timothy Morgan)
- VA Maine Healthcare System (Todd Stapley)
- VA New York Harbor Healthcare System (Scott Sherman)
- VA Pacific Islands Health Care System (Gwenevere Anderson)
- VA Palo Alto Health Care System (Philip Tsao)
- VA Pittsburgh Health Care System (Elif Sonel)
- VA Puget Sound Health Care System (Edward Boyko)
- VA Salt Lake City Health Care System (Laurence Meyer)
- VA San Diego Healthcare System (Samir Gupta)
- VA Southern Nevada Healthcare System (Joseph Fayad)
- VA Tennessee Valley Healthcare System (Adriana Hung)
- Washington DC VA Medical Center (Jack Lichy)
- W.G. (Bill) Hefner VA Medical Center (Robin Hurley)
- White River Junction VA Medical Center (Brooks Robey)
- William S. Middleton Memorial Veterans Hospital (Robert Striker)


a
mLOY cell fraction by age and ancestry in Million Veteran Program
mLOY cell fraction (%)
80
60
40
20
0
[40,50)
[50,60)
[60,70)
[70,80)
[80,90)
Age
Ancestry
EUR
AFR
HIS
b
mLOY cell fraction density by ancestry in Million Veteran Program
Ancestry
EUR
AFR
HIS
0.00
0.10
0.25
0.50
0.75
1.00
Cell Fraction (%)
c
mLOY cell fraction by age in Million Veteran Program, stratified by ancestry and smoking status
All cases
Cell fraction > 10% only
Cell fraction (%)
EUR
AFR
HIS
[40,50)
[50,60)
[60,70)
[70,80)
[80,90)
Age
Smoking status
Current
Former
Never

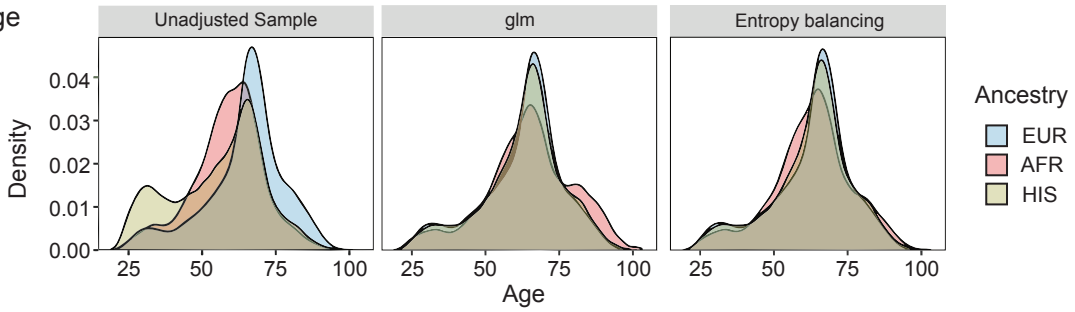
a Generalized linear model propensity score weighting



b Entropy balancing

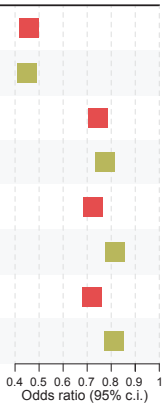


c Distributional balance for Age



d Sensitivity models for effect of ancestry on mLOY

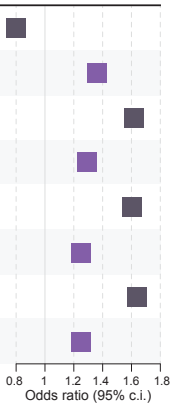
Model	Term	Odds ratio
0	odds(AFR)/odds(EUR)	0.456 (0.447, 0.465)
0	odds(HIS)/odds(EUR)	0.449 (0.436, 0.462)
1	odds(AFR)/odds(EUR)	0.745 (0.731, 0.759)
1	odds(HIS)/odds(EUR)	0.776 (0.755, 0.796)
2	odds(AFR)/odds(EUR)	0.724 (0.709, 0.738)
2	odds(HIS)/odds(EUR)	0.815 (0.793, 0.838)
3	odds(AFR)/odds(EUR)	0.721 (0.707, 0.736)
3	odds(HIS)/odds(EUR)	0.812 (0.790, 0.835)



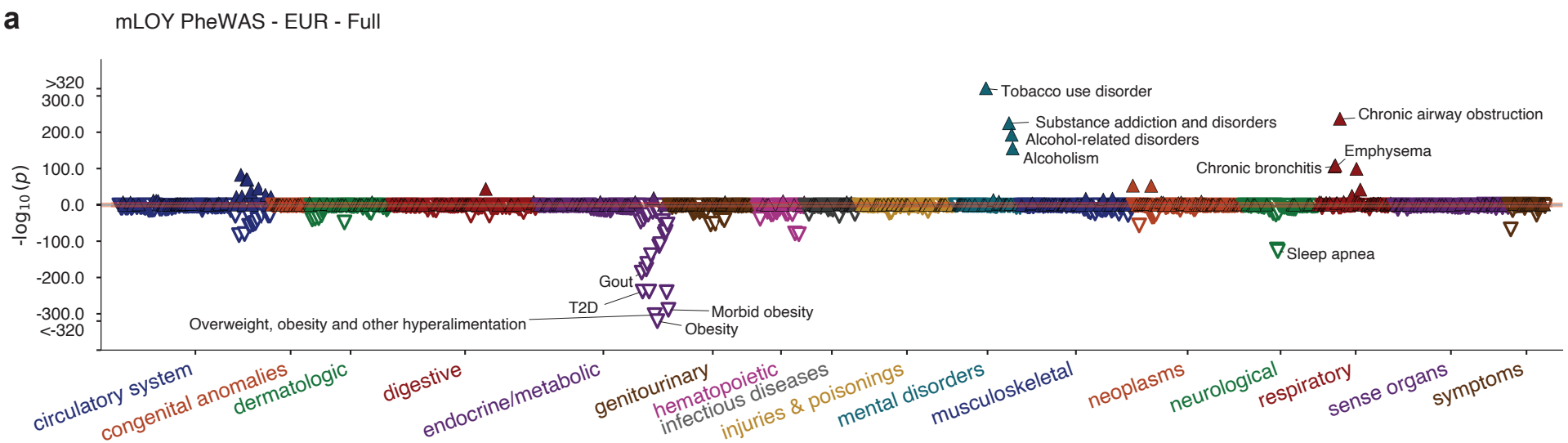
Model 0: mLOY ~ Ancestry
Model 1: mLOY ~ Ancestry + Age
Model 2: mLOY ~ Ancestry * (Age + Age^2 + BMI + Smoking status) + Age * BMI
Model 3: Model 2 weighted with entropy balancing

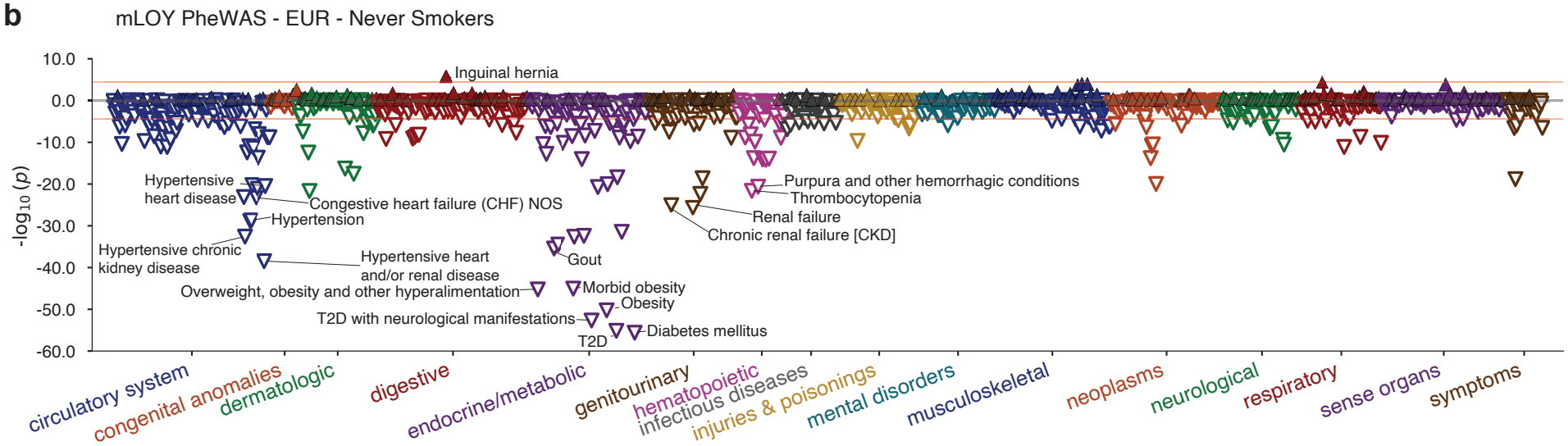
e Sensitivity models for effect of smoking status on mLOY

Model	Term	Odds ratio
0	odds(Current)/odds(Never)	0.813 (0.791, 0.837)
0	odds(Former)/odds(Never)	1.363 (1.343, 1.384)
1	odds(Current)/odds(Never)	1.624 (1.582, 1.667)
1	odds(Former)/odds(Never)	1.290 (1.272, 1.308)
2	odds(Current)/odds(Never)	1.610 (1.561, 1.660)
2	odds(Former)/odds(Never)	1.252 (1.223, 1.269)
3	odds(Current)/odds(Never)	1.640 (1.587, 1.696)
3	odds(Former)/odds(Never)	1.257 (1.239, 1.275)

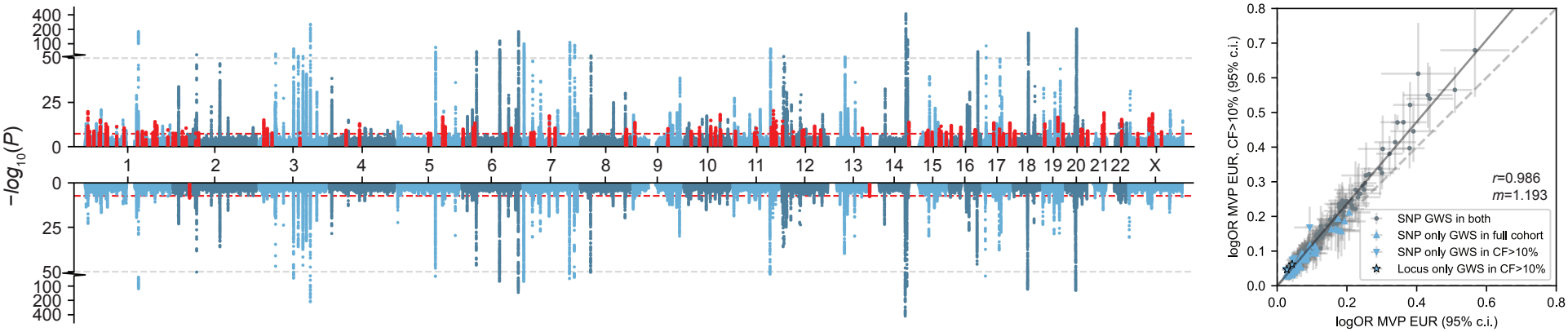


Model 0: mLOY ~ Smoking status
Model 1: mLOY ~ Smoking status + Age
Model 2: mLOY ~ Smoking status * (Age + Age^2 + BMI + Ancestry) + Age * BMI
Model 3: Model 2 weighted with entropy balancing

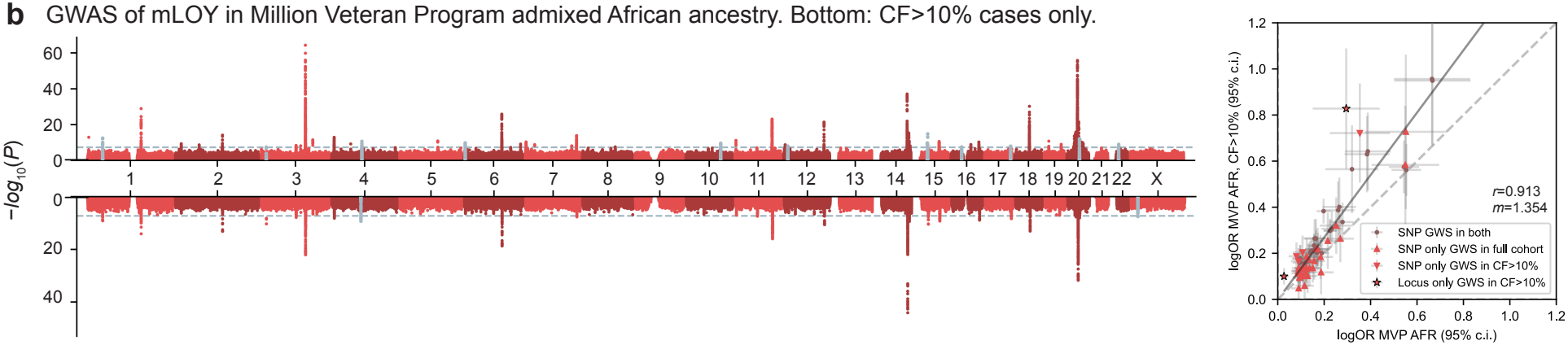

a
mLOY PheWAS - EUR - Full
-log10 (p)
>320
300.0
200.0
100.0
0.0
-100.0
-200.0
-300.0
<-320
Tobacco use disorder
Substance addiction and disorders
Alcohol-related disorders
Alcoholism
Chronic airway obstruction
Emphysema
Chronic bronchitis
Sleep apnea
Gout
T2D
Morbid obesity
Overweight, obesity and other hyperalimentation
Obesity
circulatory system
congenital anomalies
dermatologic
digestive
endocrine/metabolic
genitourinary
hematopoietic
infectious diseases
injuries & poisonings
mental disorders
musculoskeletal
neoplasms
neurological
respiratory
sense organs
symptoms


b
mLOY PheWAS - EUR - Never Smokers
-log10 (p)
10.0
0.0
-10.0
-20.0
-30.0
-40.0
-50.0
-60.0
Inguinal hernia
Hypertensive heart disease
Congestive heart failure (CHF) NOS
Hypertension
Hypertensive chronic kidney disease
Hypertensive heart and/or renal disease
Overweight, obesity and other hyperalimentation
Gout
Morbid obesity
Obesity
T2D with neurological manifestations
T2D
Diabetes mellitus
Purpura and other hemorrhagic conditions
Thrombocytopenia
Renal failure
Chronic renal failure [CKD]
circulatory system
congenital anomalies
dermatologic
digestive
endocrine/metabolic
genitourinary
hematopoietic
infectious diseases
injuries & poisonings
mental disorders
musculoskeletal
neoplasms
neurological
respiratory
sense organs
symptoms

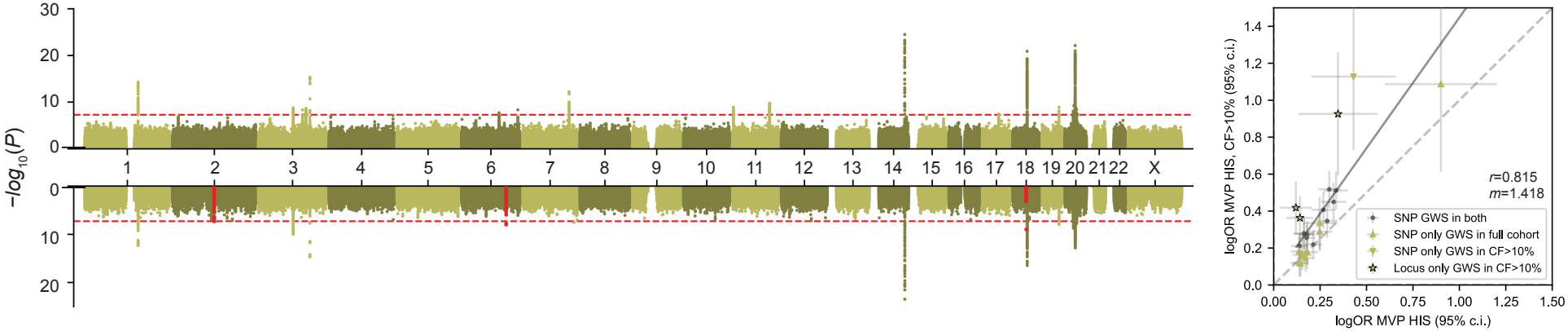
a GWAS of mLOY in Million Veteran Program European ancestry. Bottom: high cell fraction (CF>10%) cases only.

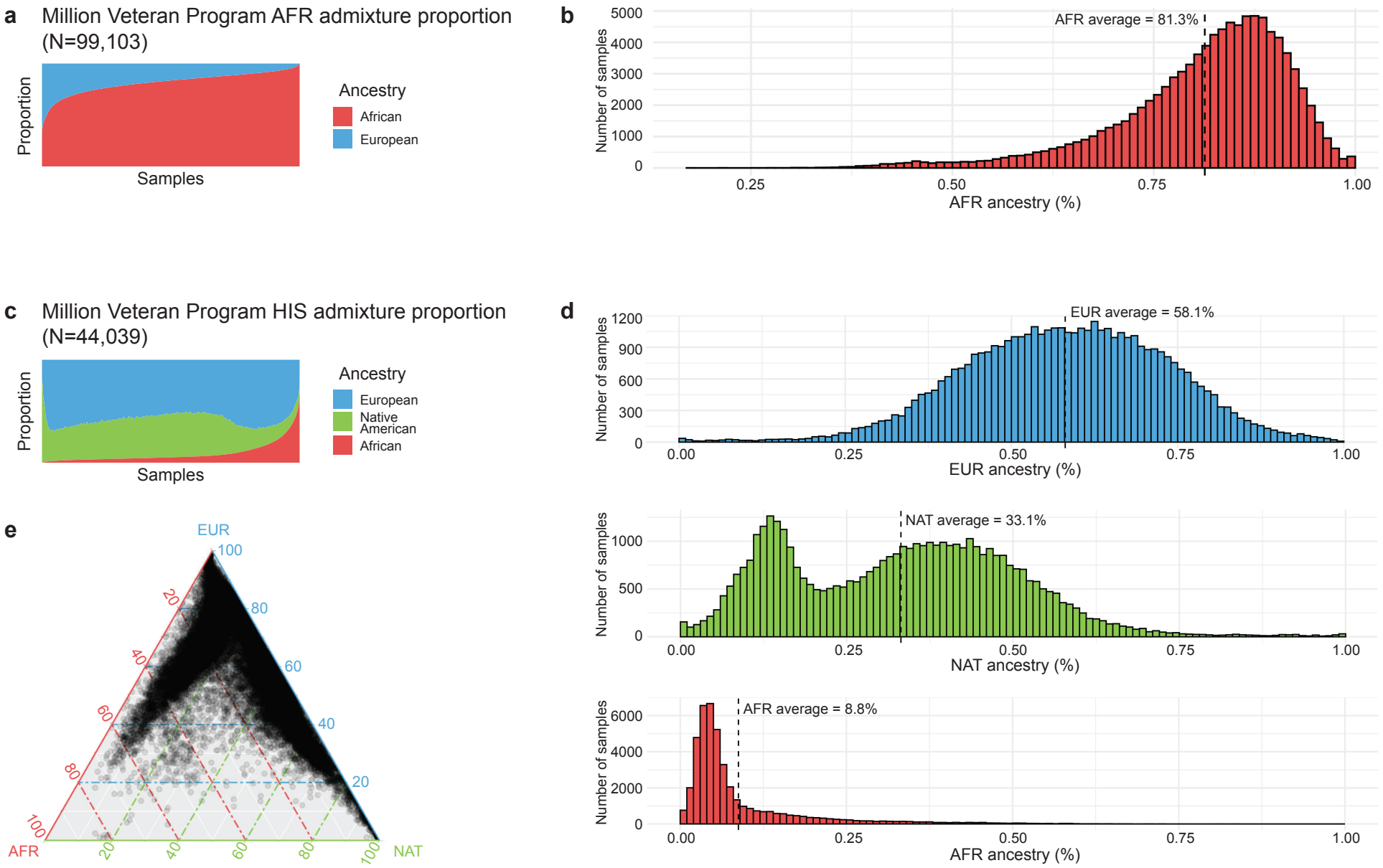


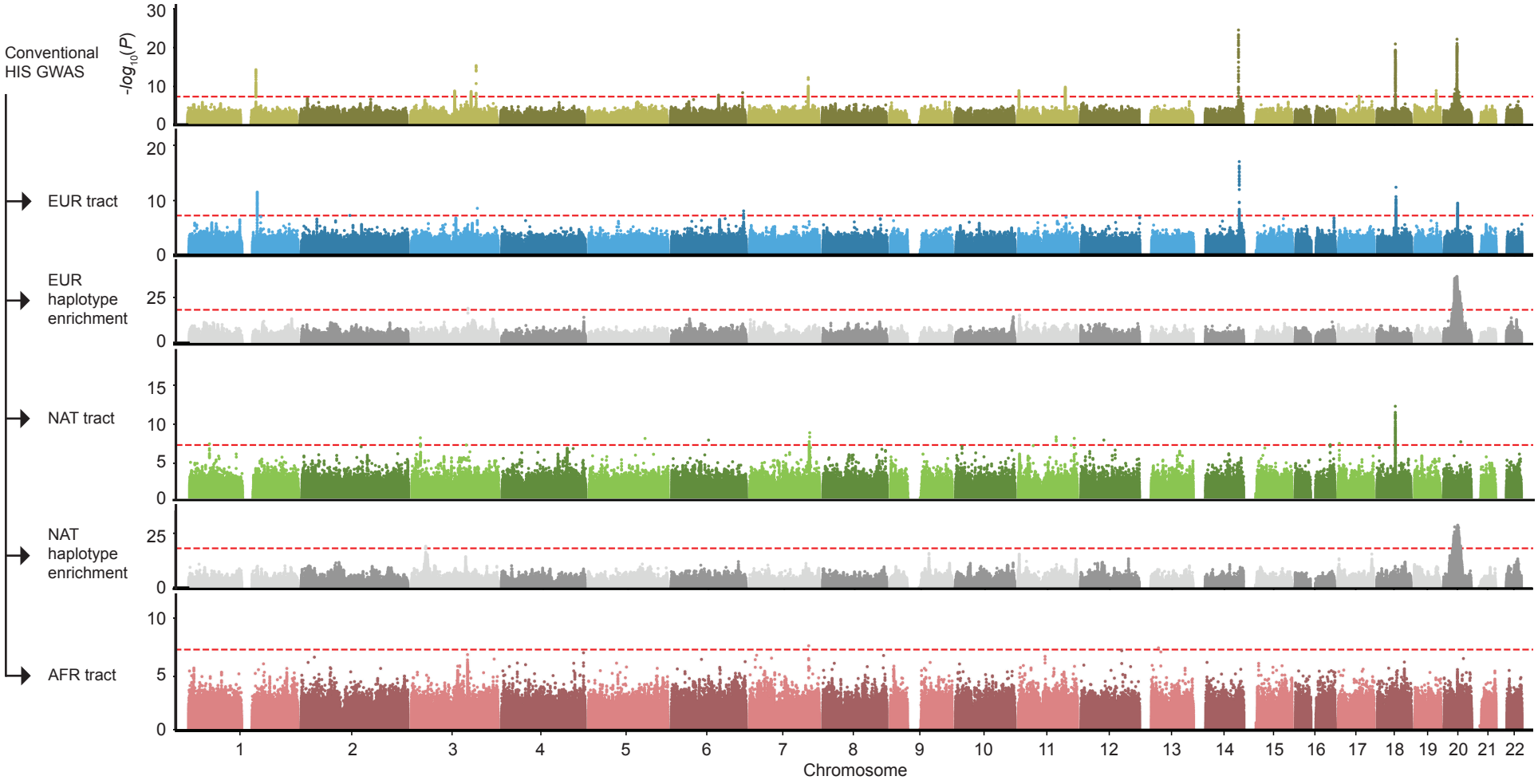
b GWAS of mLOY in Million Veteran Program admixed African ancestry. Bottom: CF>10% cases only.

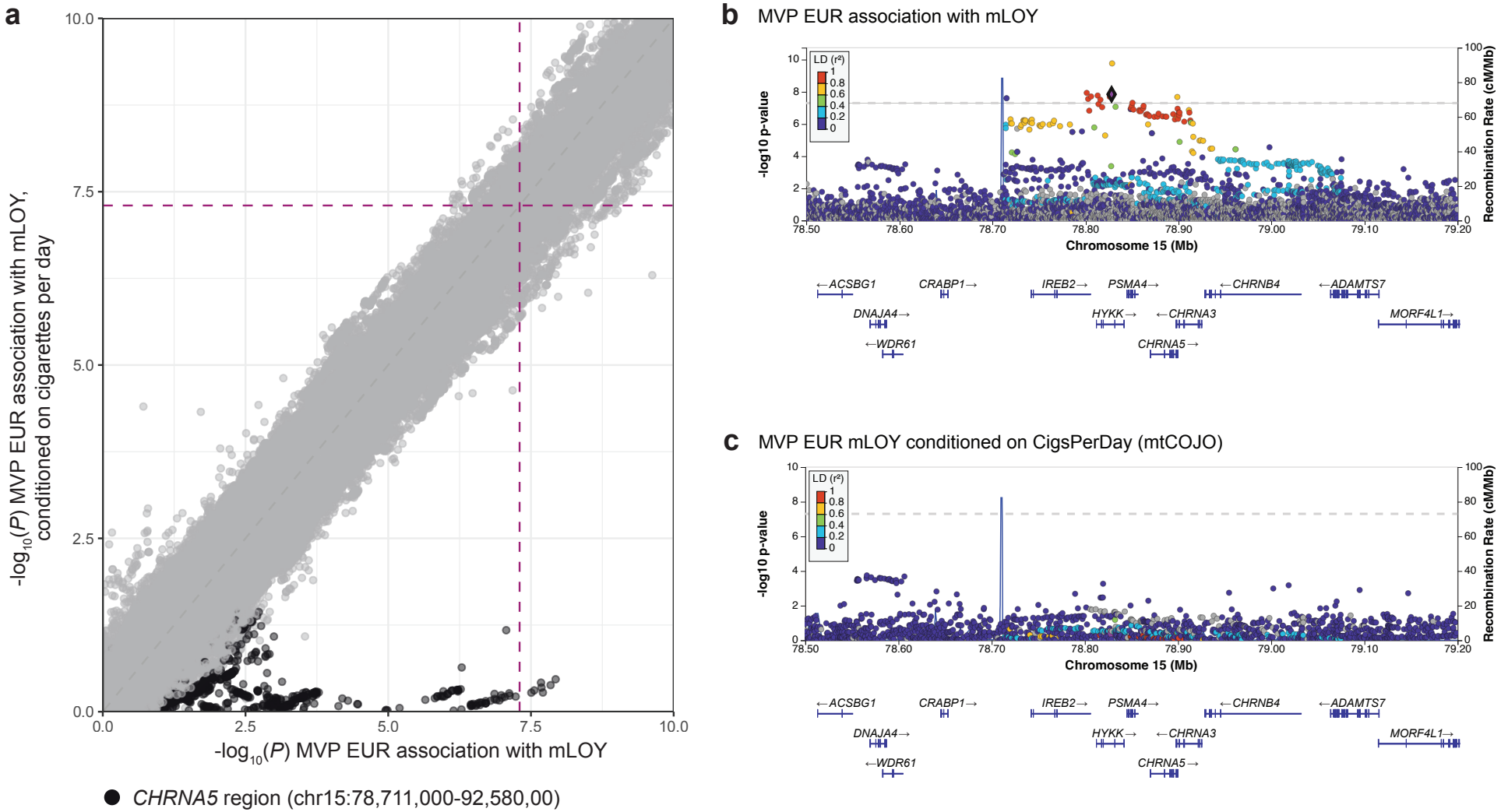


c GWAS of mLOY in Million Veteran Program Hispanic ancestry. Bottom: CF>10% cases only.




a
Million Veteran Program AFR admixture proportion (N=99,103)
Proportion
Samples
Ancestry
African
European
b
AFR average = 81.3%
Number of samples
AFR ancestry (%)
c
Million Veteran Program HIS admixture proportion (N=44,039)
Proportion
Samples
Ancestry
European
Native American
African
d
EUR average = 58.1%
Number of samples
EUR ancestry (%)
NAT average = 33.1%
Number of samples
NAT ancestry (%)
AFR average = 8.8%
Number of samples
AFR ancestry (%)
e
EUR
AFR
NAT


Conventional HIS GWAS
EUR tract
EUR haplotype enrichment
NAT tract
NAT haplotype enrichment
AFR tract
-log10(P)
Chromosome
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22


a
-log10(P) MVP EUR association with mLOY, conditioned on cigarettes per day
-log10(P) MVP EUR association with mLOY
CHRNA5 region (chr15:78,711,000-92,580,00)
b
MVP EUR association with mLOY
-log10 p-value
Recombination Rate (cM/Mb)
Chromosome 15 (Mb)
LD (r²)
ACSBG1
CRABP1
IREB2
PSMA4
CHRNB4
ADAMTS7
DNAJA4
HYKK
CHRNA3
MORF4L1
WDR61
CHRNA5
c
MVP EUR mLOY conditioned on CigsPerDay (mtCOJO)

Extended Data Figure Captions

Extended Data Fig. 1. Cell fraction (CF) of mLOY across groups. a, CF of mLOY in Million Veteran Program participants by ten-year age bins, stratified by ancestry. Boxplots show first decile, first quartile, median, third quartile, and last decile; dots indicate mean. **b**, Density ridgeline plots show distribution of CF for each ancestry. CF 10% is marked to indicate the cutoff for high CF sensitivity analyses. **c**, CF of mLOY, stratified ancestry and by current, former, and never smoking status. Full cohort and CF>10% are shown. Error bars indicate 95% confidence intervals.

Extended Data Fig. 2. Estimating effects of ancestry and smoking status on mLOY. a, Love plot of covariate balance before and after weighting using general linearized model propensity scoring. **b**, Love plot of covariate balance before and after weighting using entropy balancing. **c**, Density plots of age distribution across ancestries before and after each weighting method. **d**, Forest plot of sequential covariate adjustment sensitivity analyses for effect of ancestry on mLOY. **e**, Forest plot of sequential covariate adjustment sensitivity analyses for effect of smoking status on mLOY. BMI, body mass index; c.i., confidence interval. Parenthesis around model 2 covariates indicate the interactions of these terms and ancestry were included in the model.

Extended Data Fig. 3. Phenome-wide association studies (PheWAS) of mLOY status. PheWAS for the largest Million Veteran Program ancestry, European (EUR), in: **a**, Full cohort; **b**, Never smokers only. Association *P*-values were obtained from a two-sided test of the z-statistic, calculated using logistic regression. The red line in each plot indicates Bonferroni-corrected significance (α=0.05/1605).

Extended Data Fig. 4. Miami plots of mosaic loss of chromosome Y (mLOY) risk loci in Million Veteran Program participants by ancestry. Miami plots (left) and effect concordance scatterplots (right) for GWAS of mLOY in full cohort and high cell fraction (CF>10%) sensitivity analyses. Miami plots show the $-\log_{10}(P)$ for associations of genetic variants with mLOY in **a**, European ancestry, **b**, African ancestry, and **c**, Hispanic ancestry. Top plots show full cohorts, bottom plots show high cell fraction (CF>10%) sensitivity analyses. Numbers central to plots indicate chromosomes. Novel mLOY index variants (± 50 Kb) are highlighted in top, those additional loci found in CF>10% only are highlighted in bottom plots. The dotted line indicates the genome-wide significance threshold ($P<5\times10^{-8}$). Light grey dotted lines represent a transition from

linear to log-scale on the y-axis. c.i., confidence interval; *r*, Pearson correlation coefficient; *m*, slope.

Extended Data Fig. 5. Admixture proportions in Million Veteran Program (MVP) African (AFR) and Hispanic (HIS) cohorts. a, Two-way admixture proportion across MVP AFR participants. **b**, MVP AFR admixture binned by percentile of AFR ancestry (%). **c**, Three way admixture proportion across MVP HIS participants. **d**, MVP HIS admixture binned by percentile of European (EUR), AFR, and Native American (NAT) ancestry (%). **e**, Ternary plot of HIS admixture.

Extended Data Fig. 6. Local-ancestry-aware association analysis of Million Veteran Program (MVP) Hispanic (HIS) participants. Stacked Manhattan plots show the $-\log_{10}(P)$ values for components of the local-ancestry-aware GWAS for MVP HIS (from top to bottom): conventional HIS GWAS, European (EUR) tract, haplotype dosage enrichment of EUR with reference to African (AFR) and Native American (NAT) haplotype dosage, NAT tract, haplotype dosage enrichment of NAT with reference to EUR and AFR haplotype dosage, and AFR tract.

Extended Data Fig. 7. Multi-trait meta-analysis (mtCOJO) conditional analysis on smoking. a, Scatterplot of $-\log_{10}(P)$ from conditioning on cigarettes-per-day with mtCOJO (LPc) vs. $-\log_{10}(P)$ from standard mLOY GWAS (LP) in MVP European ancestry (EUR) demonstrates the region containing the *CHRNA5* gene (chr15:78,711,000-79,258,000), a known smoking locus, was primarily affected by conditioning. Red dashed lines indicate genome-wide significance at $P<5\times10^{-8}$. **b**, Regional Manhattan plot for *CHRNA5* in MVP EUR. **c**, Regional Manhattan plot for *CHRNA5* in MVP EUR after conditioning mLOY on cigarettes-per-day.