Supplementary Methods

PrediXcan method:

PrediXcan (Gamazon et al., 2015) is a gene-based association test that prioritizes genes which are likely to be causal for the phenotype. It implements an elastic net-based method for selecting variants associated with gene expression in a given reference panel, and then uses those variants to predict gene expression in a cohort with only genotype data. We downloaded the PrediXcan software (see URLs) along with its prepackaged weights for gene expression data from PredictDB (see URLs). Weights for gene expression using RNA sequencing data were obtained from the Genotype-Tissue Expression project (version 7) (Zhang & Lin, 2013) (whole blood, genes= 6208; and EBV transformed lymphocytes, genes=3000), Depression Genes and Networks (Battle et al., 2014) (whole blood, genes=11538, n=922), and Multi-Ethnic Study of Atherosclerosis (Europeans only, monocytes, genes=4647) (Mogil et al., 2018). Imputed genotypes for all cohorts were filtered for imputation quality based on $R^2 > 0.3$; variants not meeting this threshold were excluded from the analysis. We use DGN as our primary reference panel for all TWAS analyses as it is the largest single whole blood RNA-seq dataset.

cpgen:

We used the R package cpgen to perform conditional analysis of TWAS-significant genes, while accounting for a KING kinship matrix. However, cpgen is designed in such a way that it performs eigenvalue decomposition on the cohort sample for every function call. Since we had 239 TWAS-significant associations, this would have required eigenvalue decomposition on a

sample of N ~ 55,000 for each of those 239 associations, a computationally burdensome calculation. Thus, we slightly modified the cpgen script. Specifically, we computed the eigenvalue decomposition on the GERA sample outside of the cpgen script (for each phenotype), and then subsequently loaded the appropriate eigenvectors and eigenvalues into the program, modifying the script so that it could take these eigenvectors and eigenvalues as input.

Included cohorts:

These TWAS analyses were limited to self-reported white or European ancestry participants, for easy comparability with the DGN European ancestry eQTL panel, including input of LD information into the R Shiny application (see R Shiny Methods), and with the largest single-ancestry blood cell trait GWAS.

Genetic Epidemiology Research on Adult Health and Aging (GERA). The GERA cohort includes over 100,000 adults who are members of the Kaiser Permanente Medical Care Plan, Northern California Region (KPNC) and consented to research on the genetic and environmental factors that affect health and disease, linking together clinical data from electronic health records, survey data on demographic and behavioral factors, and environmental data with genetic data(Banda et al., 2015; Kvale et al., 2015). By self-report, the GERA cohort is 81% White and 19% minority. Each GERA participant provided a saliva sample for extraction of DNA, which was conducted at KPNC using Oragene kits (DNA Genotek Inc., Ottawa, ON, Canada). DNA samples were genotyped at the Genomics Core Facility of UCSF. Genotyping was completed as previously described (Kvale et al., 2015) using 4 different custom Affymetrix Axiom arrays with

ethnic-specific content to increase genomic coverage. In addition to the QC protocols performed during genotyping, a total of six subjects, all female, were dropped due to sex non-agreement according to the Plink v1.07 --geno option and variants with more than 10% missingness were removed. Genotype data were phased without external reference using Eagle v2.3 and then imputed to 1000 Genomes Phase 3 v5 using Minimac3. Principal components analysis was used to characterize genetic structure in this European sample (Banda et al., 2015). Hematological measures were extracted from medical records. In individuals with multiple measurements, the first visit with complete white blood cell differential (if any) was used for each participant. Otherwise, the first visit was used. In total, 54,542 non-Hispanic White individuals with hematological measures were included in the analysis.

GERA GWAS results were included in the R Shiny app as well. In the app, GERA phenotypes (log10 transformed for WBC subtypes) were based on inverse normalized residuals and adjusted for sex, age, age-squared, and the first 10 genetic principal components; analysis was done with Bolt LMM as implemented in rvtests (Zhan, Hu, Li, Abecasis, & Liu, 2016), as used in the meta-analyses reported in (Vuckovic et al., 2020). We excluded those without a valid date of blood cell count measurement, with age < 18 years, or with discordant genotypic and phenotypic sex, as well as those with no blood cell trait data. The cohort also has longitudinal data; we preferentially selected the first visit with complete data for all measures. If no visit had complete data, we used the first available visit. We also excluded extreme blood cell measures: WBC>200x109 cells/L, HGB>20 g/dL, HCT>60%, and PLT>1000x109 cells/L. For WBC subtypes, we analyzed log10-transformed absolute counts obtained by multiplying relative counts with total WBC count. Custom Axiom arrays used for GERA genotyping have

been previously described (Hoffmann, Kvale, et al., 2011; Hoffmann, Zhan, et al., 2011), as has genotype calling with apt-probeset-genotype and generation of PCs using EIGENSOFT4.2 (Banda et al., 2015).

Women's Health Initiative (WHI). WHI originally enrolled 161,808 women aged 50-79 between 1993 and 1998 at 40 centers across the US, including both a clinical trial (including three trials for hormone therapy, dietary modification, and calcium/vitamin D) and an observational study arm (The Women's Health Initiative Study Group). WHI recruited a socio-demographically diverse population representative of US women in this age range. Two WHI extension studies conducted additional follow-up on consenting women from 2005-2010 and 2010-2015. Genotyping was available on some WHI participants through the WHI SNP Health Association Resource (SHARe) resource, which used the Affymetrix 6.0 array and on other participants through the MEGA array (Wojcik et al., 2019). Imputation and association analysis was performed separately in individuals with Affymetrix only, MEGA only, and both Affymetrix and MEGA data. For variants with both Affymetrix and MEGA genotypes available, MEGA genotypes were used. In total, 18,100 self-reported white women with hematological phenotypes were included. All WHI subcohorts were imputed to 1000 Genomes Phase 3. Six sub-cohorts from the WHI study were included in the meta-analysis and phenotypes were not collected uniformly across the cohorts. Sample size information for each phenotype is contained in Supplementary Table 8.

Atherosclerosis Risk in Communities Study (ARIC). The ARIC study was initiated in 1987 and recruited participants age 45-64 years from 4 field centers (Forsyth County, NC; Jackson, MS; northwestern suburbs of Minneapolis, MN; Washington County, MD) to study cardiovascular

disease and its risk factors ("The Atherosclerosis Risk in Communities (ARIC) Study: design and objectives. The ARIC investigators," 1989), including the participants of self-reported European ancestry included here. Standardized physical examinations and interviewer-administered questionnaires were conducted at baseline (1987-89), three triennial follow-up examinations, a fifth examination in 2011-13, a sixth exam in 2016-2017 and a seventh exam in 2018-2019. Genotyping was performed through the CARe consortium Affymetrix 6.0 array (Musunuru et al., 2010). ARIC European American genotype data were imputed to Haplotype Reference Consortium (HRC) (McCarthy et al., 2016). In total, 9,345 European ancestry participants with hematological phenotypes were included in the analysis. All phenotypes were adjusted for study site, age, age squared, sex, and top ten PCs and were inverse normalized.

Mount Sinai Bio*Me* Biobank. The Mount Sinai Bio*Me* Biobank, founded in September 2007, is an ongoing, broadly consented EHR-linked bio- and data repository that enrolls participants non-selectively from the Mount Sinai Medical Center patient population. The Bio*Me* Biobank draws from a population of over 70,000 inpatient and 800,000 outpatient visits annually from over 30 broadly selected clinical sites of the Mount Sinai Medical Center (MSMC). As of September 2020, Bio*Me* has enrolled more than 50,000 patients that represent a broad racial, ethnic and socioeconomic diversity with a distinct and population-specific disease burden, characteristic of the communities served by Mount Sinai Hospital. Bio*Me* participants are predominantly of African (AA, 24%), Hispanic/Latino (HL, 35%), European (EA, 32%), and other ancestry (OA, 10%). All blood cell phenotype data, as well as demographic variables, were extracted from the patients' EHRs. Genotyping was performed using the Illumina GSA array (~640.000 variants) and genotype data were imputed using the "1000G Phase 3 v5" reference

panel. In total, 8,455 European ancestry participants with hematological phenotypes were included in the analysis. All phenotypes were adjusted for study site, age, age squared, sex, and top ten PCs and were inverse normalized. The Bio*Me* Biobank Program operates under a Mount Sinai Institutional Review Board-approved research protocol. All study participants provided written informed consent.

R Shiny:

Additional details relevant to the production of the LocusXcanR application in R Shiny follow. Correlation of predicted expression among genes at the locus was calculated using R's *cor* function, and LD among variants was computed using plink --r2 (https://zzz.bwh.harvard.edu/plink/ld.shtml). We used the *visNetwork* function for network visualizations and the *ggplot2* function to produce all other figures. Tables were produced using the DT package (https://www.rdocumentation.org/packages/DT/versions/0.16). The IdeogramTrack (https://rdrr.io/bioc/Gviz/man/IdeogramTrack-class.html) uses Genome Reference Consortium Human Build 37 (GRCh37) and UCSC cytogenetic bands from http://hgdownload.cse.ucsc.edu/goldenpath/hg19/database/.

References

The Atherosclerosis Risk in Communities (ARIC) Study: design and objectives. The ARIC investigators. (1989). *Am J Epidemiol, 129*(4), 687-702.

Banda, Y., Kvale, M. N., Hoffmann, T. J., Hesselson, S. E., Ranatunga, D., Tang, H., . . . Risch, N. (2015). Characterizing Race/Ethnicity and Genetic Ancestry for 100,000 Subjects in the Genetic

Epidemiology Research on Adult Health and Aging (GERA) Cohort. *Genetics, 200*(4), 1285-1295. doi:10.1534/genetics.115.178616

Battle, A., Mostafavi, S., Zhu, X., Potash, J. B., Weissman, M. M., McCormick, C., . . . Koller, D. (2014). Characterizing the genetic basis of transcriptome diversity through RNA-sequencing of 922 individuals. *Genome Res, 24*(1), 14-24.

Gamazon, E. R., Wheeler, H. E., Shah, K. P., Mozaffari, S. V., Aquino-Michaels, K., Carroll, R. J., . . . Im, H. K. (2015). A gene-based association method for mapping traits using reference transcriptome data. *Nat Genet*. doi:10.1038/ng.3367

Hoffmann, T. J., Kvale, M. N., Hesselson, S. E., Zhan, Y., Aquino, C., Cao, Y., . . . Risch, N. (2011). Next generation genome-wide association tool: design and coverage of a high-throughput European-optimized SNP array. *Genomics, 98*(2), 79-89. doi:10.1016/j.ygeno.2011.04.005

Hoffmann, T. J., Zhan, Y., Kvale, M. N., Hesselson, S. E., Gollub, J., Iribarren, C., . . . Risch, N. (2011). Design and coverage of high throughput genotyping arrays optimized for individuals of East Asian, African American, and Latino race/ethnicity using imputation and a novel hybrid SNP selection algorithm. *Genomics, 98*(6), 422-430. doi:10.1016/j.ygeno.2011.08.007

Kvale, M. N., Hesselson, S., Hoffmann, T. J., Cao, Y., Chan, D., Connell, S., . . . Risch, N. (2015). Genotyping Informatics and Quality Control for 100,000 Subjects in the Genetic Epidemiology Research on Adult Health and Aging (GERA) Cohort. *Genetics, 200*(4), 1051-1060. doi:10.1534/genetics.115.178905

McCarthy, S., Das, S., Kretzschmar, W., Delaneau, O., Wood, A. R., Teumer, A., . . . Durbin, R. (2016). A reference panel of 64,976 haplotypes for genotype imputation. *Nat Genet, 48*(10), 1279-1283. doi:10.1038/ng.3643

Mogil, L. S., Andaleon, A., Badalamenti, A., Dickinson, S. P., Guo, X., Rotter, J. I., . . . Wheeler, H. E. (2018). Genetic architecture of gene expression traits across diverse populations. *PLoS Genet, 14*(8), e1007586. doi:10.1371/journal.pgen.1007586

Musunuru, K., Lettre, G., Young, T., Farlow, D. N., Pirruccello, J. P., Ejebe, K. G., . . . Gabriel, S. B. (2010). Candidate gene association resource (CARe): design, methods, and proof of concept. *Circ Cardiovasc Genet, 3*(3), 267-275. doi:10.1161/circgenetics.109.882696

The Women's Health Initiative Study Group. (1998). Design of the Women's Health Initiative clinical trial and observational study. *Control Clin Trials, 19*(1), 61-109.

Vuckovic, D., Bao, E. L., Akbari, P., Lareau, C. A., Mousas, A., Jiang, T., . . . Soranzo, N. (2020). The Polygenic and Monogenic Basis of Blood Traits and Diseases. *Cell, 182*(5), 1214-1231.e1211. doi:10.1016/j.cell.2020.08.008

Wojcik, G. L., Graff, M., Nishimura, K. K., Tao, R., Haessler, J., Gignoux, C. R., . . . Carlson, C. S. (2019). Genetic analyses of diverse populations improves discovery for complex traits. *Nature, 570*(7762), 514-518. doi:10.1038/s41586-019-1310-4

Zhan, X., Hu, Y., Li, B., Abecasis, G. R., & Liu, D. J. (2016). RVTESTS: an efficient and comprehensive tool for rare variant association analysis using sequence data. *Bioinformatics, 32*(9), 1423-1426. doi:10.1093/bioinformatics/btw079

Zhang, L., & Lin, X. (2013). Some considerations of classification for high dimension low-sample size data. *Stat Methods Med Res, 22*(5), 537-550. doi:10.1177/0962280211428387