 doi:10.20944/preprints202106.0129.v1

Article

Transcriptome-wide association study of blood cell traits in African ancestry and Hispanic/Latino Populations

Jia Wen[1], Munan Xie[1], Bryce Rowland[2], Jonathan D. Rosen[2], Quan Sun[2], Amanda L. Tapia[2], Huijun Qian[3], Madeline H. Kowalski[2], Yue Shan[2], Kristin L. Young[4], Marielisa Graff[4], Maria Argos[5], Christy L. Avery[4], Stephanie A. Bien[6], Steve Buyske[7], Jie Yin[8], Hélène Choquet[8], Myriam Fornage[9], Chani J. Hodonsky[10], Eric Jorgenson[11], Charles Kooperberg[6], Ruth J.F. Loos[12], Yongmei Liu[13], Jee-Young Moon[14], Kari E. North[4], Stephen S. Rich[10], Jerome I. Rotter[15], Jennifer A. Smith[16], Wei Zhao[16], Lulu Shang[17], Tao Wang[14], Xiang Zhou[17], Alexander P. Reiner[18], Laura M. Raffield[1], Yun Li[1,2]*

[1] Department of Genetics, University of North Carolina at Chapel Hill, Chapel Hill, NC, 27502, USA; jia_wen@med.unc.edu (J.W.); meximus@126.com (M.N.X.); laura_raffield@unc.edu (L.M.R.)
[2] Department of Biostatistics, University of North Carolina at Chapel Hill, Chapel Hill, NC, 27599, USA; bryce.rowland@unc.edu (B.R.); jdrosen@live.unc.edu (J.D.R.); quansun@live.unc.edu (Q.S.); altapia@live.unc.edu (A.L.T.); madeline.kowalski@nyulangone.org (M.H.K.); yshan@live.unc.edu (Y.S.)
[3] Department of Statistics and Operations Research, University of North Carolina at Chapel Hill, Chapel Hill, NC, 27599, USA; hjqian@alumni.unc.edu (H.J.Q.)
[4] Department of Epidemiology, University of North Carolina at Chapel Hill, Chapel Hill, NC, 27516, USA; kristin.young@unc.edu (K.L.Y.); migraff@email.unc.edu (M.G.); christy_avery@unc.edu (C.L.A.); kari_north@unc.edu (K.E.N.)
[5] Division of Epidemiology and Biostatistics, University of Illinois at Chicago, Chicago, IL, 60612, USA; argos@uic.edu (M.A.)
[6] Division of Public Health Sciences, Fred Hutchinson Cancer Research Center, Seattle, WA, 98109, USA; sbien@fredhutch.org (S.A.B.); clk@fredhutch.org (C.K.)
[7] Department of Statistics, Rutgers University, Piscataway, NJ, 08854, USA; buyske@stat.rutgers.edu (S.B.)
[8] Division of Research, Kaiser Permanente Northern California, Oakland, CA, 94612, USA; Jie.Yin@kp.org (J.Y.); Helene.Choquet@kp.org (H.C.)
[9] Institute of Molecular Medicine, McGovern Medical School, The University of Texas Health Science Center, Houston, TX, 77030, USA; Myriam.Fornage@uth.tmc.edu (M.F.)
[10] Center for Public Health Genomics, University of Virginia, Charlottesville VA, 22908, USA; ch2um@virginia.edu (C.J.H.); ssr4n@virginia.edu (S.S.R.)
[11] Regeneron Genetics Center, Tarrytown, NY, 10591, USA; eric.jorgenson@regeneron.com (E.J.)
[12] The Charles Bronfman Institute for Personalized Medicine, Icahn School of Medicine at Mount Sinai, New York, NY, 10029, USA; ruth.loos@mssm.edu (R.J.L)
[13] Molecular Physiology Institute, Duke University, Durham, NC, 27701, USA; yongmei.liu@duke.edu (Y.M.L.)
[14] Department of Epidemiology & Population Health, Albert Einstein College of Medicine, Bronx, NY, 10461, USA; jee-young.moon@einsteinmed.org (J.-Y.M.); tao.wang@einsteinmed.org (T.W.)
[15] The Institute for Translational Genomics and Population Sciences, Department of Pediatrics, The Lundquist Institute for Biomedical Innovation at Harbor-UCLA Medical Center, Torrance, CA, 90502, USA; jrotter@lundquist.org (J.I.R.)
[16] Department of Epidemiology, School of Public Health, University of Michigan, Ann Arbor, MI, 48109, USA; smjenn@umich.edu (J.A.S.); zhaowei@umich.edu (W.Z.)
[17] Department of Biostatistics, School of Public Health, University of Michigan, Ann Arbor, MI, 48109, USA; shanglu@umich.edu (L.L.S.); xzhousph@umich.edu (X.Z.)
[18] Department of Epidemiology, University of Washington, Seattle, WA, 98195, USA; apreiner@uw.edu (A.P.R)
* Correspondence: yunli@med.unc.edu (Y.L.)

Abstract:

Background: Thousands of genetic variants have been associated with hematological traits, though target genes remain unknown at most loci. Also, limited analyses have been conducted in African ancestry and Hispanic/Latino populations; hematological trait associated variants more common in these populations have likely been missed.

 doi:10.20944/preprints202106.0129.v1

Methods: To derive gene expression prediction models, we used ancestry-stratified datasets from the Multi-Ethnic Study of Atherosclerosis (MESA, including N=229 African American and N=381 Hispanic/Latino participants, monocytes) and the Depression Genes and Networks study (DGN, N = 922 European ancestry participants, whole blood). We then performed a transcriptome-wide association study (TWAS) for platelet count, hemoglobin, hematocrit, and white blood cell count in African (N = 27,955) and Hispanic/Latino (N = 28,324) ancestry participants.

Results: Our results revealed 24 suggestive signals ($p < 1\times 10^{-4}$) that were conditionally distinct from known GWAS identified variants and successfully replicated these signals in European ancestry subjects from UK Biobank. We found modestly improved correlation of predicted and measured gene expression in an independent African American cohort (the Genetic Epidemiology Network of Arteriopathy (GENOA) study (N=802), lymphoblastoid cell lines) using the larger DGN reference panel; however, some genes were well predicted using MESA but not DGN.

Conclusions: These analyses demonstrate the importance of performing TWAS and other genetic analyses across diverse populations and of balancing sample size and ancestry background matching when selecting a TWAS reference panel.

Keywords: TWAS; non-European; blood cell traits

1. Introduction

Hematological measures (red blood cell, white blood cell, and platelet traits) play critical roles in oxygen transport, immunity, infection, thrombosis, and hemostasis and are associated with many chronic diseases, including autoimmunity, asthma, cardiovascular disease, and viral infections. Hematological traits vary between self-reported race/ethnicity groups [1-3], in part due to variants which vary in frequency by genetic ancestry group [4,5]. For example, selective pressure from malaria has led to an increased prevalence of sickle cell anemia, G6PD deficiency, and alpha thalassemia in exposed populations [6,7], with impacts on hematological indices [8] in individuals with genetic ancestry from malaria endemic regions. Unfortunately, most existing genome-wide association studies (GWAS) have focused on European ancestry (EA) populations [9-11]. It is essential to also explore hematological trait genetics in underrepresented admixed African (AA) and Hispanic/Latino (HL) populations.

Along with the lack of diversity in included populations, GWAS of hematological traits, similar to other complex phenotypes, also identify regions of associated variants whose biological function and target genes are often not clear. One approach to linking variants to genes (and thus to biological function) that has demonstrated success in various complex traits is transcriptome-wide association studies (TWAS) [12,13]. TWAS methods leverage reference expression quantitative trait loci (eQTL) datasets to select genetic variants which in aggregate associate with to predict tissue-specific gene expression. Weights based on variant associations with a transcript are applied to cohorts with genotype (but not expression) data available [13,14]. Transcripts whose levels can be confidently imputed from genetic variants are then assessed for phenotype associations [15]. However, like GWAS, TWAS analyses have often included only EA populations [16,17], though some recent efforts have included more diverse populations [18-22]. Furthermore, most reference eQTL datasets used for TWAS include predominantly EA American individuals [23,24], limiting the predictive power of this novel method in other populations with different allele frequencies.

We here examine the utility of TWAS methods to understand previously identified GWAS loci for blood cell traits, as well as to identify new candidate loci and genes (Figure 1). We focus on underrepresented AA and HL populations and evaluate whether use of

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 June 2021 doi:10.20944/preprints202106.0129.v1

relevant ancestry-specific eQTL reference panels from the Multi-Ethnic Study of Atherosclerosis (MESA) improves power for blood cell trait genetic discovery (above European-acentric TWAS reference panels).

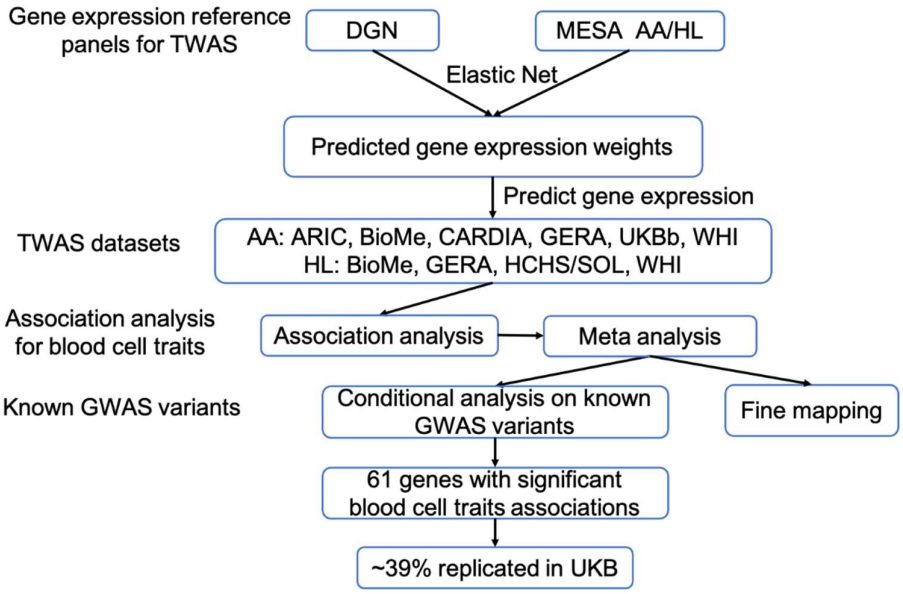

Figure 1. Study design for blood cell trait focused TWAS analyses in African ancestry (AA) and Hispanic/Latino (HL) populations. TWAS: Transcriptome-wide association study; DGN: Depression Genes and Networks; MESA: Multi-ethnic Study of Atherosclerosis. ARIC: Atherosclerosis Risk in Communities; BioMe: BioMe™ Biobank; CARDIA: Coronary Artery Risk Development in Young Adults; GERA: Genetic Epidemiology Research on Adult Health and Aging; UKB: UK Biobank; WHI: Women's Health Initiative; HCHS/SOL: Hispanic Community Health Study/Study of Latinos.

2. Materials and Methods

2.1 *Training of TWAS Prediction Models*

Genotype data Using genotype data from 6 African ancestry cohorts (ARIC, BioMe, CARDIA, GERA, UK Biobank, and WHI) and 4 Hispanic/Latino cohorts (BioMe, GERA, WHI, HCHS/SOL) (Figure 1, Supplementary Table 1, Supplementary Content), we performed genotype imputation to the TOPMed freeze 5b reference panel (September 2017 release, methods described at https://www.nhlbiwgs.org/topmed-whole-genome-sequencing-project-freeze-5b-phases-1-and-2) on the Michigan Imputation Server [25]. After genotype imputation, we selected variants for TWAS model training if they were well-imputed in all GWAS cohorts and in reference expression quantitative trait loci (eQTL) datasets (MAF>0.05, R^2> 0.3). The only exception was for the MESA eQTL reference panel, where direct sequencing data from TOPMed freeze 8 was available for those with monocyte expression microarray data. In brief, the sequencing for MESA was performed with mean genome coverage ≥30x and completed at seven sequencing centers; sequencing data files were transferred from sequencing centers to the TOPMed Informatics Research Center (IRC), where reads were aligned to human genome build GRCh38 using a common pipeline across all centers, followed by joint genotype calling. Additional sample-level quality control (such as detection of sex mismatches, pedigree discrepancies, sample swaps, etc.) was undertaken by the Sequencing Centers, the IRC, and the TOPMed Data Coordinating Center (DCC). Detailed methods for freeze 8 are available at https://www.nhlbiwgs.org/topmed-whole-genome-sequencing-methods-

 doi:10.20944/preprints202106.0129.v1

freeze-8. For MESA, we included all variants with MAF >0.05 for TWAS model training. After quality control exclusions, there were 4,694,075 variants in DGN, 4,948,702 variants in MESA AA ancestry, and 4,265,631 in MESA HL ancestry for TWAS modelling.

Gene expression data DGN RNA-seq (whole blood) data was obtained from [24]. RNA-seq was already normalized using the Hidden Covariates with Prior (HCP) method [26], correcting for technical and biological factors [19]. A total of 15,231 genes were used for training models in the DGN reference panel. The MESA monocyte gene expression data were obtained from the Illumina HumanHT-12 V4.0 expression beadchip (GSE56045) [27,28] [29]. Illumina IDs were converted to Ensembl IDs. For Illumina IDs with multiple Ensembl IDs, each Illumina ID was labeled and treated as an individual gene, leading to multiple TWAS weights being used for 11,056 probe sets (including 8,623 genes). More genes were available for TWAS prediction in DGN versus MESA, likely due to a use of RNA-seq instead of expression microarray, larger sample size, and inclusion of additional cell types in whole blood versus monocytes alone (for example, lymphocyte-specific transcripts). To adjust for confounders and then normalize the MESA expression data, we fit a multivariate linear model for gene expression values in MESA controlling for the following covariates: sex, age, age squared, 10 genotype principal components calculated by PC-AiR method from TOPMed freeze 8 WGS data, and 10 PEER factors from microarray gene expression profiling calculated using the "peer" package in R (v 1.0) [30]. Adjusted gene expression values were computed by inverse-quantile normalizing the residuals of the above multivariate linear model.

Model training We used the DGN EA, MESA AA, and MESA HL eQTL datasets to train gene expression prediction models using elastic net. For each gene, variants located within +/- 1Mb of genes start and end were used in model training. Elastic-net methods have been previously shown to be effective even when there is collinearity due to variants in high linkage disequilibrium (LD), so no LD pruning of variants was conducted prior to elastic net model fitting, except for removal of perfectly correlated variants. Hyperparameters were trained via 10-fold cross validation, similar to the PrediXcan approach [13]. Genes with model $R^2 \geq 0.05$ for elastic net prediction models were carried forward for TWAS analyses. After model training, we separately applied DGN, MESA AA and MESA HL training models to predict gene expression values in each GWAS cohort.

These model training methods are essentially the same as the well-established PrediXcan method (elastic net), for which precomputed weights are already available for DGN and MESA at http://predictdb.org/, but we retrained models using TOPMed freeze 5b imputed data and TOPMed freeze 8 WGS data (in MESA) due to substantial gains in imputation quality in AA and Hispanic/Latino populations versus commonly used reference panels like 1000 Genomes [31].

2.2 Assessment of Trained Expression Prediction Models in GENOA

To assess the performance of the gene expression prediction models trained from the eQTL reference datasets in an independent AA cohort, we applied the gene expression prediction models to N=802 participants from the Genetic Epidemiology Network of Arteriopathy (GENOA) study [32]. GENOA gene expression was measured in lymphoblastoid cell lines (LCLs) using the Affymetrix Human Transcriptome Array 2.0. We used correlation between real and predicted gene expression transcript values (true R^2) to compare the performance of the MESA AA and DGN TWAS reference datasets.

Based on initial findings that, overall, DGN gene expression prediction was better than the prediction results from the smaller, ancestry-matched MESA cohort, we randomly subsampled DGN to the same size as MESA to evaluate the impact of sample

 doi:10.20944/preprints202106.0129.v1

size on the training models. In order to account for variability in prediction accuracy, we randomly performed five subsamplings of the DGN cohort to the MESA cohort sample size. We then retrained TWAS models with elastic net within this smaller subsampled DGN cohort, using the same TWAS training methods described above, and then compared the average of true R^2 from the multiple subsampled-DGN training models with true R^2 from MESA AA (to evaluate if ancestry matching would improve TWAS performance in the same sample size).

2.3 Phenotype

We analyzed four blood cell traits from each major hematological domain (hemoglobin – HGB, hematocrit – HCT, white blood cell count – WBC, platelet counts – PLT). These traits had the largest sample size across included cohorts. Prior to association analyses, we excluded extreme outlier values, notably WBC values >200x10^9/L, HCT >60%, and HGB >20g/dL. For longitudinal cohort studies, all values are from the same exam cycle, with the exam which maximized available sample size chosen (usually baseline). WBC was log transformed due to the skewed distribution. All four traits were adjusted for age, age squared, sex, and top 10 principal components/study specific covariates. The residuals were then inverse normalized.

2.4 TWAS Association and Conditional Analysis

We assessed association between predicted gene expression and the inverse normalized HCT, HGB, WBC, and PLT residuals using the cGWAS.emmax function of R package cpgen v0.1, adjusting for an EPACTS kinship matrix. Additionally, WBC analyses included adjustment for the Duffy (rs2814778) variant, which strongly impacts neutrophil counts and overall WBC, with the null allele much more common in African ancestry populations [4]. Following association testing within each cohort, we performed meta-analysis within ancestry groups using METAL [33] We further investigated genes with a nominal meta-analysis p-value < 1×10^{-4} through a conditional association test on all known variants within +/- 1Mb flanking regions. Strict Bonferroni significance thresholds (0.05/# tested genes) are listed in Supplementary Table 7. We performed conditional analysis by conditioning on these known signals within +/- 1Mb of each significant gene in two steps (Supplementary Table 5). We initially conducted conditional analysis using known variants from the GWAS catalog as of June 2018 (Step 1). If the TWAS signal remained significant after Step 1 conditional analysis, or if the TWAS signal had no GWAS catalog GWAS variants within +/- 1Mb region, we further conditioned on significant ($p < 5 \times 10^{-8}$) variants from recent blood cell traits GWAS [11,34] (Step 2). We then keep the larger (i.e., less significant) conditional p-value as the final results. Note that we used a lenient threshold of 0.05 to declare conditional significance. In addition, we define a locus from our TWAS results as the +/- 1Mb region of most significant sentinel gene's start and end positions. For loci with more than one marginally significant gene identified, we further perform fine-mapping using FINEMAP [35].

2.5 Replication

Our replication procedure was split into two tiers of testing to assess the effect of ancestry-matched gene expression prediction models in TWAS replication. First, we attempted replication using gene expression prediction models trained in DGN, but further restricted to variants that are common and well-imputed (MAF > 0.05, R^2 > 0.8) in both DGN and UKB Europeans (DGN-for-UKB). We retrained gene expression prediction model using DGN-for-UKB with the same elastic net procedure described above.

For genes with models trained in DGN using variants available in all cohorts in the AA or HL TWAS meta-analysis dataset, DGN-for-UKB trained models served as the only model for predicted gene expression considered in initial replication analyses. For MESA AA or MESA HL trained models, if either the gene-trait association was not significant at

 doi:10.20944/preprints202106.0129.v1

the Bonferroni adjusted threshold ($p < 8 \times 10^{-4}$) or the prediction quality of the gene expression prediction model was poor (model $R^2 < 0.01$), we used the MESA trained model to impute gene expression values into UKB Europeans, in a second tier of replication analyses.

Sixty-one gene-trait pairs that demonstrated evidence of conditionally significant association in AA or HL TWAS meta-analyses were included in the UK Biobank replication analysis. We used a Bonferroni-corrected threshold for the number of tests ($0.05/61 = 8 \times 10^{-4}$) to identify replicated signals. Phenotypes were adjusted for covariates and inverse normalized, as described above for AA and HL cohorts (additional details in Supplemental Content, Included Cohorts). REGENIE [36] was used to test the association between predicted gene expression and adjusted phenotypes.

3. Results

3.1. *Train Gene Expression Prediction Models from Reference eQTL Datasets*

In this study, we leveraged data from two studies as reference eQTL datasets: EA individuals from the Depression Genes and Networks (DGN) cohort (N=922 with whole blood RNA-sequencing data [24]), and AA (N=229) and HL (N=381) individuals from the MESA study [27,28]. We attempted to train gene expression prediction models using elastic net for 11,687 genes, 5,958 genes with 7,120 probe sets, and 6,082 genes with 7,316 probe sets in DGN EA, MESA AA, and MESA HL eQTL reference panels respectively, and obtained models with model $R^2 > 0.05$ for 9,861 genes, 5,883 genes with 7,026 probe sets, and 5,522 genes with 6,559 probe sets, respectively. From the training results, based on model $R^2 > 0.05$, 3,331 genes were well-predicted with all reference panels, but 4,859 were well-predicted only with DGN (and 664 and 410 only in MESA AA and HL, respectively, Figure 2). If we restrict to the 3,931 genes present in both the DGN and both MESA reference panels, 3,927 genes are well-predicted (model $R^2 > 0.05$) using any of the three reference panels; 7 genes are well predicted only with DGN; 50 only with MESA AA; and 5 only with MESA HL. Comparing the distribution of model R^2 across all genes and common genes, the MESA AA reference panel performed slightly better than DGN, while DGN performed slightly better than MESA HL (Figure 3, Supplementary Table 2). The large number of genes (4,859) only well-predicted with DGN is likely due to the larger sample size and higher number of expressed genes for this reference panel, as well as the greater number of included cell types in whole blood. Genes not highly expressed in monocytes are unlikely to be well predicted with the MESA reference panel. Differences in model R^2 across reference panels can be quite striking; for example, *LTBP3* is well predicted in DGN (model $R^2 = 0.47$), but not well predicted with MESA HL (model $R^2 = 0.03$), and less than 2 non-zero weights contribute to the *LTBP3* prediction model with MESA AA. The Human Protein Atlas, a database that provides gene expression in different blood cell types, shows that *LTBP3* has higher gene expression in lymphocytes such as the memory B-cells and naive B-cells but very low gene expression in myeloid-lineage white blood cell types such as monocytes, eosinophils, and neutrophils [37], which may explain this poor prediction using MESA monocyte expression data as a reference panel. However, for a subset of genes, the MESA ancestry-specific reference panels had better performance, such as *TNFAIP2* with model $R^2 = 0.10$ in MESA HL but only $R^2 = 0.02$ in DGN. For *TNFAIP2*, variants in the prediction models for MESA HLs were more common in HL than EA populations (14/16, 87.5%) using 1000 Genome phase 3 data as a minor allele frequency (MAF) reference.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 June 2021 doi:10.20944/preprints202106.0129.v1

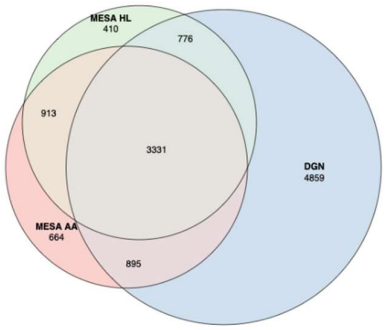

Figure 2. Venn diagram showing the overlap of well-predicted genes by Depression Genes and Networks (DGN) European ancestry and Multi-ethnic Study of Atherosclerosis (MESA) African ancestry (AA) and Hispanic/Latino (HL) reference panels.

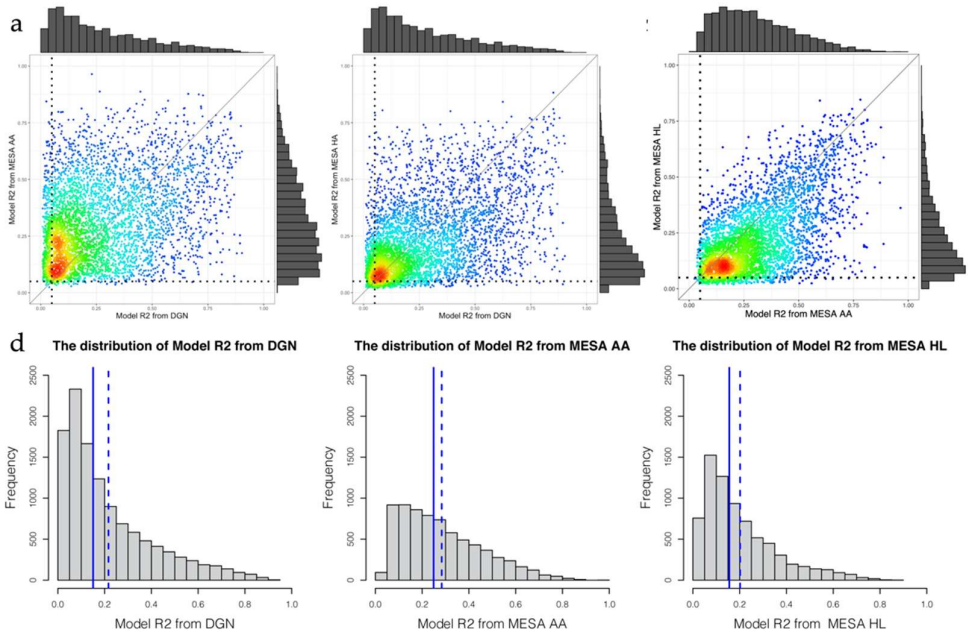

Figure 3. The smooth scatter plots show the model R^2 distribution of common genes available in both the Depression Genes and Networks (DGN) European and Multi-ethnic Study of Atherosclerosis (MESA) reference eQTL datasets. The dashed line denotes the threshold value (model $R^2 = 0.05$) for well-predicted genes. (**a**) Comparison of model R^2 of common genes found in both Depression Genes and Networks (DGN) and Multi-ethnic Study of Atherosclerosis (MESA) African ancestry (AA) reference panels; (**b**) Comparison of model R^2 of common genes between the Depression Genes and Networks (DGN) and Multi-ethnic Study of Atherosclerosis (MESA) Hispanic/Latino (HL) reference panels; (**c**) Comparison of model R^2 of common genes between the Multi-ethnic Study of Atherosclerosis (MESA) African ancestry (AA) and Multi-ethnic Study of Atherosclerosis (MESA) Hispanic/Latino (HL) reference panels; (**d**) Histograms showing the model R^2 distribution of all genes in each reference panel (without model *R2* filtering). The blue solid line denotes the median of model R^2; the blue dashed line denotes the mean of model R^2. All genes, including those which do not meet a model $R^2 = 0.05$ cut-off, are displayed.

 doi:10.20944/preprints202106.0129.v1

3.2. Assessment of Trained Expression Prediction Models in Independent non-European Datasets.

To assess the performance of gene expression prediction models trained from the eQTL reference datasets among minority participants, we applied TWAS prediction models to AA participants (N=802) from the Genetic Epidemiology Network of Arteriopathy (GENOA) study [32]. We compared the performance of prediction models trained from MESA AA and DGN reference eQTL datasets in terms of the correlation between predicted and true (true R^2) (Figure 4a, b, Supplementary Table 3). When comparing genes found in both DGN and MESA reference panels, we found that DGN performs better than MESA AA in prediction of GENOA expression, based on true R^2 (Figure 4a). Restricting to well-predicted genes in both DGN and MESA AA (model $R^2 > 0.05$ in both reference panels), DGN outperforms MESA AA as well (Supplementary Table 3, Figure 4b). Among them, 279 genes had true $R^2 > 0.05$ in GENOA with both reference panels, 584 only with DGN, and 247 only with MESA AA (Figure 4c). We suspected that DGN might be performing slightly better than the ancestry-matched MESA AA reference due to the larger sample size. We therefore sub-sampled DGN to be the same size as the MESA AA panel and predicted into the external GENOA dataset, to directly compare the impact of sample size on the training model. The true R^2, which decreased with the reduced sample size for the sub-sampled DGN panel, was still slightly better, on average, than models trained from MESA AA (Supplementary Table 3) when restricting to common predicted genes between DGN and MESA AA. The decreased true R^2 of sub-sampled DGN models compared to the original full DGN-sample models does, however, directly demonstrate the effect of sample size on the prediction performance (Supplementary Table 3). In addition, we still found GENOA genes which are well-predicted by MESA but very poorly predicted by DGN, such as *POLR2J*, which is highly expressed in most blood cells (including monocytes and lymphocytes) [37]. For *POLR2J*, true R^2 in GENOA is 0.42 using the MESA TWAS reference panel versus $R^2 = 0.007$ using DGN as a TWAS reference panel. Comparing the variants included in prediction models from both reference panels, 39 of 57 variants (~68%) used in MESA AA derived prediction model are more common in AA than EA populations, which may lead to better prediction, while only 4 of 27 (~15%) included variants are more common in AA versus EA populations for the DGN prediction model. Allele frequency differences across populations likely leads to the lower true R^2 using DGN as a reference panel.

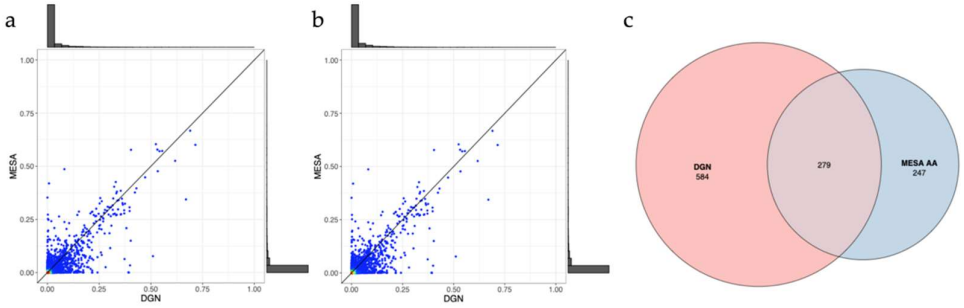

Figure 4. Smooth scatter plots comparing true gene expression from Genetic Epidemiology Network of Arteriopathy (GENOA) lymphoblastoid (LCL) data to predicted gene expression using Depression Genes and Networks (DGN) and Multi-ethnic Study of Atherosclerosis (MESA) African ancestry (AA) reference eQTL datasets. (**a**) True R^2 distribution using the Depression Genes and Networks (DGN) and Multi-ethnic Study of Atherosclerosis (MESA) eQTL reference panel for all genes (# genes = 4,043); (**b**) True R^2 distribution for genes with model $R^2 > 0.05$ in each reference eQTL dataset (# genes = 3,426); (**c**) Venn diagram of well-predicted genes (true $R^2 > 0.05$) by DGN and MESA AA reference panels in GENOA.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 June 2021 doi:10.20944/preprints202106.0129.v1

3.3. Association between Predicted Gene Expression and Blood independent Traits.

We applied TWAS weights from both DGN and MESA to independent cohorts of AA ancestry (N= 27,955) and HL (N= 28,324) participants and performed TWAS for four blood cell traits: HGB, HCT, WBC, and PLT. We selected these four traits to both encompass all three major hematological domains (red and white blood cells, and platelets) and retain maximal sample size in each constituent cohort (due to extensive missing data for most other traits). Results were then meta-analyzed for each trait within each ancestry group (AA or HL).

For the 27,955 AA participants from 6 AA cohorts (Supplementary Content and Supplementary Table 1), we conducted TWAS with the four hematological traits using pre-trained models on 9,861 and 5,883 genes with 7,026 probe sets from the DGN and MESA AA eQTL reference dataset respectively. For the 28,324 HL participants from 4 HL cohorts (cohort details in Supplementary Content and Supplementary Table 1), we performed TWAS with the four blood cell traits for 9,861 and 5,522 genes with 6,559 probe sets trained in DGN and MESA HL eQTL reference datasets, respectively.

Meta-analyses revealed 42 and 27 marginally significant gene-trait pairs in AA and HL, respectively, when DGN expression prediction models were used. When prediction models from MESA AA and HL samples were employed, 22 and 18 marginally significant gene-trait pairs were identified for AA and HL, respectively. In total, we identified 80 unique genes and 90 unique gene-trait pairs across AA and HL. Of 90 unique marginally significant gene-trait pairs, we found 13 gene-trait pairs by both DGN and MESA (AA and HL) prediction models, 51 gene-trait pairs by DGN models only, and 26 gene-trait pairs using MESA AA and/or HL prediction models only. We note that there were 23 unique genes associated with blood cell traits only when using non-EA reference panels, despite the much smaller sample sizes compared to DGN EA reference (detailed in Supplementary Table 4, Supplementary Figure 1).

3.4. TWAS Analysis Conditional on Neighboring GWAS Variants.

For the 90 marginally significant gene-trait pairs, 67 genes (involved in 74 unique gene-trait associations, ~82.2%) have known GWAS-identified variant associations for the same hematological trait within +/- 1Mb. We performed conditional analysis to investigate whether the significant TWAS genes remain significant when conditioned on known GWAS variants, which would suggest that additional genetic variants regulating expression of the TWAS identified gene remain to be discovered (see Methods). The conditional analysis shows that 45 (including 42 genes) out of the total 74 (~60.8%) unique associations are still nominally significant when conditioned on known GWAS variants ($p < 0.05$, Supplementary Table 4). An additional 16 out of 90 unique gene-trait associations have no known GWAS variants within 1 Mb. In the replication analysis using the much larger UKB EA sample, we focused on the 42 conditionally significant genes (45 gene-trait associations) and the 15 genes with no known GWAS variants within +/- 1 Mb (16 gene-trait associations), as these were the genes with evidence of association with hematological traits not captured by existing GWAS, for 61 total gene-trait associations (Supplementary Table 4).

3.5. Replication in UK Biobank.

After establishing a list of conditionally significant gene-trait associations, we performed replication analysis for the conditionally significant gene-trait associations and gene-trait associations without nearby known GWAS variants in 405,782 EA participants from the UKB. We note that African ancestry UKB participants were already included in our discovery TWAS meta-analyses and thus were not available for replication purposes. Using independently trained DGN-for-UKB gene expression prediction models, 21 out of

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 June 2021 doi:10.20944/preprints202106.0129.v1

49 gene-trait association pairs were successfully replicated at significance level of $\alpha < 8 \times 10^{-4}$ ($\alpha = 0.05/61$). For the 16 genes (involved in 16 associations) which did not have a valid DGN-for-UKB gene expression prediction model, we performed replication analysis in UKB-EA using the same models for gene expression prediction used in MESA AA and MESA HL. Of these genes, 3 out of 16 replicated in the secondary analysis at $\alpha < 8 \times 10^{-4}$ ($\alpha = 0.05/61$). We note that almost all of these replicated genes are conditionally distinct signals at known loci; of the 16 gene-trait pairs with no nearby GWAS identified variants, only *APEH*'s association with hematocrit in AA populations compellingly replicates in UK Biobank ($p = 1.7 \times 10^{-6}$). Two findings of interest, both of which replicated in UKB, are described below.

3.6. Example replicated genes still nominally significant after conditioning on known GWAS variants.

***TNFAIP2* (TNF Alpha Induced Protein 2)** In the analyses conducted in HL populations using MESA as reference, *TNFAIP2* remains significant for platelet count when conditioned on known GWAS variants (marginal $p = 6.8 \times 10^{-5}$; conditional $p = 8.1 \times 10^{-3}$). As noted previously, *TNFAIP2* is not well-predicted using the DGN reference panel; the prediction model of *TNFAIP2* using MESA HL as a reference mostly includes variants more common in HL than EA populations, which may lead to better prediction using this ancestry-matched reference panel. We assessed the variants which contribute to *TNFAIP2* prediction in functional annotation data from megakaryocytes, the cells which produce platelets. Five variants (Variants ID of chr:GRCh38-positions:Ref_allele:Alt_allele: 14:103099825:G:A; 14:103112649:T:C; 14:103118337:A:G; 14:103378293:G:A, and 14:103535180:G:A) used in the MESA HL TWAS prediction model for *TNFAIP2* are annotated as being in open chromatin by megakaryocyte ATAC-seq [38]. Megakaryocyte gene expression from the BLUEPRINT consortium [39] shows that *TNFAIP2* has higher expression in this cell type than other nearby GWAS annotated genes (*EXOC3L4* and *TECPR2*) and has comparable expression with *RCOR1* (Figure 5). Moreover, this gene replicates in the UKB EA cohort using variant weights derived from the MESA HL reference panel (marginal $p = 2.9 \times 10^{-30}$).

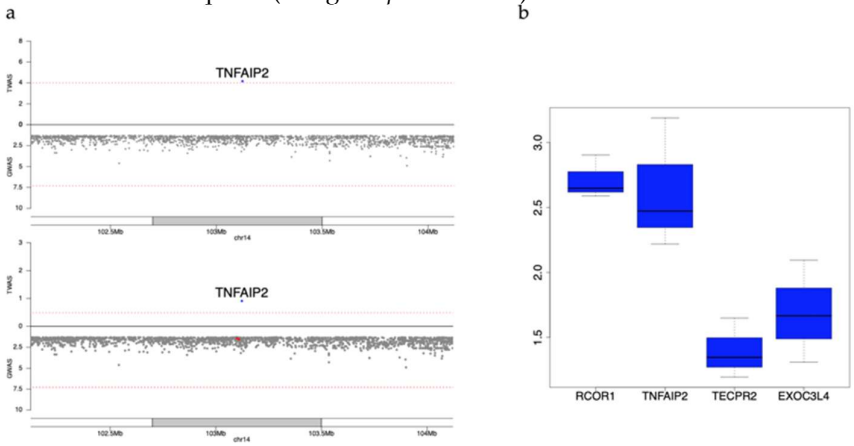

Figure 5. *TNFAIP2* locus. (**a**) Mirror plots showing the conditional analysis for *TNFAIP2*, predicted using the Multi-ethnic Study of Atherosclerosis (MESA) Hispanic/Latino (HL) reference panel, for platelet association meta-analysis in Hispanic/Latino cohorts. The red dots in the bottom panel denote the nearby GWAS signals conditioned on. *TNFAIP2* is still significant when conditioned on nearby GWAS signals (as listed in Supplementary Table 5); (**b**) Gene expression for *TNFAIP2* and the other three GWAS annotated genes in this locus from platelet-producing megakaryocytes (MK) from BLUEPRINT [39].

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 June 2021 doi:10.20944/preprints202106.0129.v1

***ENG* (Endoglin).** *ENG* is a known hemoglobin (HGB)/hematocrit (HCT) associated gene reported in EA populations in previous GWAS [11,34,40]. We identified this gene in our TWAS meta-analyses using weights from the MESA AA panel (marginal $p = 1.27\times10^{-5}$ for HGB, $p = 1.24\times10^{-6}$, which passes the Bonferroni threshold value for HCT), but not the DGN panel (marginal $p = 5\times10^{-3}$ for HGB and $p = 3\times10^{-4}$ for HCT) (Figure 6). When conditioned on known GWAS variants for HGB/HCT, *ENG* is still nominally significant with conditional $p = 3.44\times10^{-3}$ for HGB and $p = 2.6\times10^{-3}$ for HCT. We further investigated the MAF distribution of variants used in the prediction models of the two training panels and found 18/22 (~81.8%) variants used in the MESA-derived prediction model are more common in AA versus EA individuals in 1000G, with similar enrichment not observed for the DGN EA reference panel weights (27/53 variants (~50.9%) more common in AA populations).

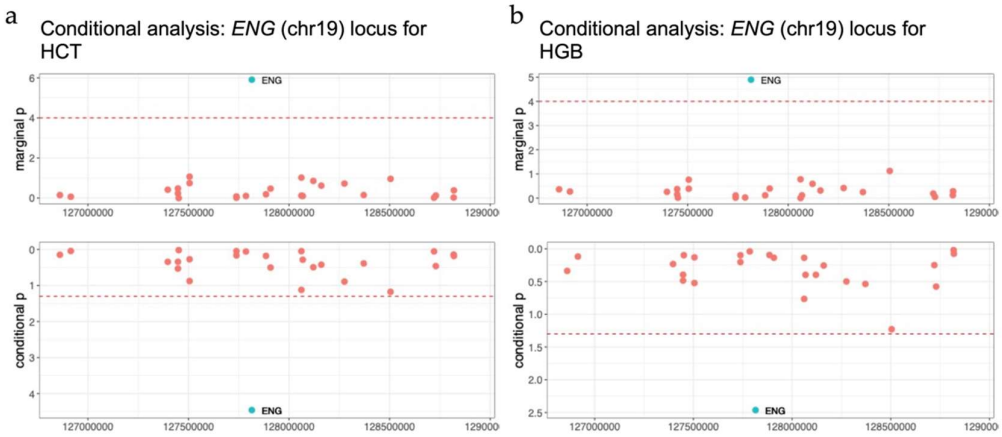

Figure 6. *ENG* locus. (**a**) The marginal and conditional results for *ENG* for hematocrit (HCT) and (**b**) hemoglobin (HGB) predicted using the Multi-ethnic Study of Atherosclerosis (MESA) reference panel in African ancestry (AA) meta-analysis. The green dot denotes the gene *ENG*, the red dots denote other genes within this locus region.

3.6. FINEMAP analysis for significant gene-trait associations.

In total, we identified 66 loci for 4 blood cell traits across AA and HL using both reference panels (Supplementary Table 4), of which 58 loci only include one gene. There are in total 27 gene-trait pairs in the remaining 8 loci which include at least two signals. For prioritizing the candidate causal gene at these loci, we performed fine-mapping using FINEMAP [35], however the findings were difficult to interpret, in particular for the following two examples.

***THBS3* locus.** The *THBS3* locus includes four genes: *THBS3*, *EFNA2*, *TRIM46*, and *GBAP1*. *THBS3* was the most marginally significant gene, passing the Bonferroni threshold value (marginal $p = 9.8\times10^{-8}$) for HCT from TWAS analysis using the DGN reference panel in HL. The correlation of predicted gene expression was very low for these 4 genes (correlation R^2 ranged from 7×10^{-4} to 0.03). However, the fine-mapping results included 6 genes: 4 significant genes (*EFNA2*, *TRIM46*, *THBS3*, and *GBAP1*) plus 2 additional non-significant genes (*MTX1* and *KRTCAP2*), in the 95% credible set (Figure 7a, c). *MTX1* is reported as the mapped gene for the previously reported GWAS lead variant for HCT at this locus (i.e. rs760077) [9,11,34]. However, *MTX1* was not significant in our TWAS results, despite adequate model R^2 for gene expression prediction (marginal $p = 0.001$; model $R^2 = 0.10$). Within this locus the predicted gene expression between the lead gene *THBS3* and GWAS-reported gene *MTX1* was not correlated ($R^2 = 7.2\times10^{-5}$), but the gene prediction correlations are stronger between *MTX1* and *GBAP1* ($R^2 = 0.31$), *MTX1* and *EFNA3* ($R^2 = 0.40$), and *MTX1* and *TRIM46* ($R^2 = 0.78$). We further note that FINEMAP

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 June 2021 doi:10.20944/preprints202106.0129.v1

assigned *MTX1* the highest marginal posterior inclusion probability (PIP =1) [9,40], followed by *GBAP1*. Given the high correlations between *GBAP1*, *EFNA3*, *TRIM46*, and *MTX1*, none of which are correlated with lead TWAS gene (*THBS3*), it is difficult to accurately narrow down a reasonably sized candidate causal gene set at this locus. The low PIP for the most significant *THBS3* gene is also difficult to interpret. In summary, fine-mapping was not informative in pinpointing the candidate causal gene for this locus.

***MFN2* locus.** The *MFN2* locus includes two significant genes: *PLOD1* and *MFN2*. The lead gene *MFN2* is the most significant gene for platelet count in our AA-focused meta-analysis ($p = 1.36\times10^{-5}$, DGN reference panel). *MFN2* is a hemostasis-related gene involved in megakaryocyte development and platelet production [41]. Both of the genes were reported in prior GWAS results based on proximity (the TSS of *MFN2* is ~5kb from the 3'UTR of *PLOD1* on the same strand) [9,34]. Fine-mapping results reported three genes (*PLOD1*, *KIAA2013*, and *MFN2*) in the 95% credible set at this locus. The predicted expression of the lead gene, *MFN2*, is not highly correlated with *KIAA2013* ($R^2 = 0.005$) or *PLOD1* ($R^2 = 0.003$); however, fine-mapping shows that all these three genes are in the 95% credible causal gene set (Figure 7b, d). Additionally, around 50% - 55% of variants are shared in the three gene prediction models, even though the predicted expression correlations are low, which may contribute to the unclear 95% credible set results.

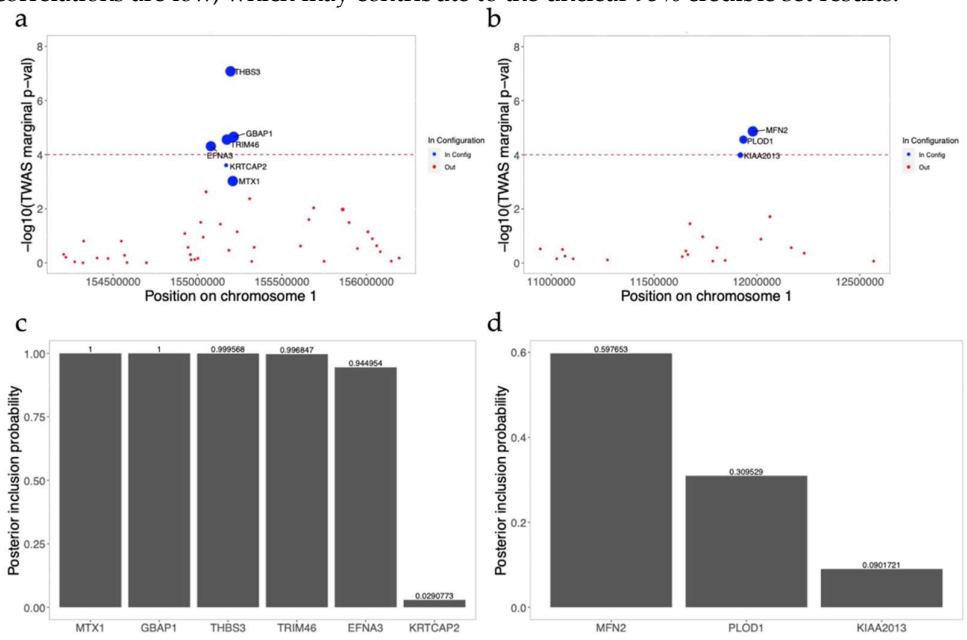

Figure 7. Fine-mapping of the *THBS3* and *MFN2* locus. Blue dots denote genes in the causal gene set configuration; red dots denote the genes outside of the causal gene set configuration. Dot size is proportional to the marginal posterior inclusion probability of each gene in the 95% credible set within the locus. The red dashed line denotes the TWAS significance threshold value. (**a**), (**c**) *THBS3* locus for hematocrit using Depression Genes and Networks (DGN) reference panel in African ancestry (AA) and the posterior inclusion probability of each gene in the 95% credible set within this locus; (**b**), (**d**) *MFN2* locus for platelet count using Depression Genes and Networks (DGN) reference panel in African ancestry (AA) and the posterior inclusion probability of each gene in the 95% credible set within this locus.

We further checked additional 6 loci from our TWAS results and found that the 95% credible causal gene sets were very similar: fine-mapping included all marginal significant

 doi:10.20944/preprints202106.0129.v1

genes within each locus and did not help to distinguish the candidate causal genes (Supplementary Table 6).

4. Discussion

We conducted large-scale TWAS meta-analyses for 4 blood cell traits in AA and HL populations, using a larger whole-blood eQTL reference panel (DGN) in European ancestry populations and smaller but ancestry-matched reference panels from MESA monocytes. To our knowledge, this is the largest genetic discovery effort yet conducted in these populations for blood cell indices, exceeding available sample sizes from recent trans-ethnic blood cell traits GWAS ([34], N=15,171 African ancestry, N=9,368 Hispanic/Latino), recent analyses of red blood cells in the Population Architecture using Genomics and Epidemiology [PAGE] studies ([42], n=16,258 African American, N=20,784 Hispanic/Latino), and our own recent TOPMed imputed GWAS analyses ([43], N=21,513 African ancestry and 21,689 Hispanic/Latino participants). These previous efforts all included some but not all of the cohorts included here. In our comparisons of expression reference panels, using one of the largest expression datasets in African Americans (GENOA LCL data) as a testing dataset for comparing imputed to true gene expression (using the DGN, sub-sampled DGN and MESA TWAS training reference panels), we observed that the larger sample size for DGN (as well as the potentially the use of whole blood instead of monocyte specific data, or RNA-seq versus expression microarray) led to slight improvements in model performance over MESA AAs. The generalizability of these comparisons in GENOA is limited by tissue heterogeneity between the DGN and MESA reference panels, however. However, we see additional significant gene-trait associations using the ancestry-matched monocyte eQTL reference panels not captured using DGN, and these TWAS models generally include a high percentage of variants more common in non-European populations. This is true even for red blood cell and platelet traits; in theory, a monocyte-specific eQTL dataset should be a poorer model for these traits than whole blood, as unlike in whole blood no transcripts from platelets or red blood cells would be included (though even for whole blood, the vast majority of transcripts are for white blood cells (monocytes and lymphocytes)). Finding new, replicated loci in much more modest sample sizes than for current trans-ethnic GWAS efforts (which include >750,000 participants [34]) also highlights the importance of performing TWAS and other genetic analyses in diverse populations.

Previous work has also suggested the value of ancestry-matched TWAS reference panels. Recent work from Geoffroy, et al., [22] found more significant complex trait-associated genes in summary statistics from the diverse PAGE consortium (~50,000 HL, African American, Asian, Native Hawaiian, and Native American individuals) using models trained in African American and HL MESA participants than in European- or all-ancestry TWAS models. Work in the diverse Carolina Breast Cancer Study (CBCS) cohort also suggests that TWAS models perform poorly across race/ethnicity groups, with multiple novel discoveries for breast cancer survival found in AA women using TWAS gene expression models trained in a subset of the cohort with measured expression data [19]. Poor performance comparing real and predicted gene expression in whole blood was also observed for GTEx v7 (>85% European) and DGN TWAS weights in a pediatric African American cohort—the Study of African Americans, Asthma, Genes, and Environment (SAGE)—with substantial declines in prediction accuracy in African ancestry populations from the GEUVADIS LCL expression datasets also observed versus European ancestry datasets [44]. Sample sizes still remain limited for transcriptome data in non-European populations, however. For example, the large blood-based eQTLGen meta-analysis included almost exclusively European-ancestry individuals[45]. Additional data generation in non-European populations is important to improve TWAS performance and gene-trait association discovery.

 doi:10.20944/preprints202106.0129.v1

Our work also identified biologically relevant genes for hematological traits at loci not containing previous GWAS signals, as well as genes which may explain known GWAS signals in gene-dense regions or highlight additional sub-genome-wide significant conditionally distinct signals at known loci, such as *TNFAIP2*. Previous studies reported that *TNFAIP2*, identified by TWAS with the MESA HL reference panel for platelet count, is related to acute promyelocytic leukemia [46,47] and may play a role as mediator of inflammation [48] and angiogenesis [49]. It is also known to be highly expressed in mononuclear progenitor cells of the bone marrow [50] and in mature peripheral blood monocytes [51], and plays a crucial role in apoptosis [48,52,53]. One study reported that *TNFAIP2* was involved with the protective mechanism of Amentoflavone in the hematopoietic system of mice against γ-irradiation [50]. While this broad locus for platelet count was known from GWAS, conditional analysis adjusting for these known GWAS identified variants suggests additional variant-trait associations remain to be discovered, and this gene also has not been commonly identified as the likely target gene in prior GWAS studies [11,34] (where annotated genes were often based on proximity alone). This gene-trait association was not identified using the DGN reference panel, likely due to the allele frequency differences for variants in the gene expression prediction model in European-ancestry and HL populations, further demonstrating the importance of ancestry matched reference panels.

Even though fine-mapping could help to prioritize the candidate causal genes at TWAS loci in some cases, our results show that fine-mapping may not always provide assistance in identifying candidate causal genes. There are several challenges for TWAS fine-mapping. One challenge is that shared eQTLs in the prediction models for multiple genes could bias the fine-mapping results, such as those observed at the *MFN2* locus. Having a large number of genes whose predicted expression is highly correlated to the GWAS reported gene could also bias fine-mapping results, as we observed at the *THBS3* locus. Moreover, FINEMAP, which performs fine-mapping using the correlation matrix based on the gene prediction models, may not reflect the true correlation of gene expression within a locus. Future work on pinpointing the potential causal genes from TWAS may include using measured gene expression, instead of predicted gene expression, to calibrate the correlation matrix in fine-mapping analysis [12].

Our analysis has a number of limitations. First, our TWAS significance threshold value was lenient. Using a strict Bonferroni (0.05/# tested genes, Supplementary Table 7) significance threshold, only 24 out of 90 unique gene-trait associations noted in Supplementary Table 4 (~26.7%) met that criterion, most of which are well known genes from previous GWAS studies. To increase our odds of identifying novel and biologically interesting signals in understudied AA and HL populations, we chose 1×10^{-4} as a nominal significance threshold, but only considered as true findings genes which additionally replicated in UK Biobank, increasing confidence in our results. Larger sample sizes are necessary to identify novel genes using a strict multiple-testing threshold. Secondly, we do not have access to a dataset for comparing real and predicted gene expression in a HL population (as such a publicly available dataset does not, to our knowledge, currently exist); also, the tissue used for GENOA gene expression quantification, LCLs, differs from that in MESA (monocytes) and DGN (whole blood). Comparisons in GENOA are thus complicated by tissue heterogeneity, though we note that prior studies have performed comparisons (for example across eQTLs) across these relatively similar blood cell related tissues (for example LCLs, whole blood, and peripheral blood mononuclear cells are compared in the eQTLGen meta-analysis [45]). In addition, we note that cohorts were classified as HL and AA based primarily on participant self-report; these self-identified race/ethnicity groupings, which are socially defined, encompass a variety of genetic ancestry backgrounds. Therefore, exact ancestry

matching between, for example, MESA HL participants and HL participants in cohorts included in the hematological traits meta-analysis cannot be assumed.

In summary, TWAS analysis can help to enhance our understanding of the genetics of complex traits, including hematological indices. TWAS not only helps to identify potential causal genes at known GWAS loci, but can also increase power for new candidate gene and locus discovery. Some gene-trait associations in our analysis were only revealed by ancestry-matched TWAS reference panels due to the difference in MAF of included variants for expression prediction across populations. More work is needed to both generate large expression datasets in non-European-ancestry populations and apply these datasets to TWAS discovery for complex traits across ancestrally diverse cohorts.

Supplementary Materials: Supplementary Content and figures; Supplementary Tables S1 – S7.

Author Contributions: Conceptualization, A.P.R., L.M.R. and Y.L.; methodology, J.W., M. X., B.R., J.D.R., Q.S., A.L.T., H.Q., M.H.K., Y.S., K.L.Y., M.G.; formal analysis, J.W., M. X. and B.R.; resources, M.A., C.L.A., S.A.B., S.B., J.Y., H.C., M.F., C.J.H., E.J., C.K., R.J.F.L., Y.L., J-Y.M.,K.E.N., S.S.R., J.I.R., J.A.S., W.Z., L.S., T.W. , and X.Z..; writing—original draft preparation, J.W., M.X., B.R. and L.M.R; writing—review and editing, all authors; visualization, J.W.; supervision, L.M.R. and Y.L.; funding acquisition, A.P.R. and Y.L. All authors have read and agreed to the published version of the manuscript.

Funding: This project was funded by R01HL129132. The project described was also supported by the National Center for Advancing Translational Sciences, National Institutes of Health, through Grant KL2TR002490 (L.M.R.). The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH. CJH was additionally funded by T32 HL007284. LMR was additionally funded by R01HG010297 and T32 HL129982. YL was additionally funded by R01 GM105785 and U01 DA052713. This material is additionally based upon work supported by the National Science Foundation Graduate Research Fellowship under Grant No. DGE-1650116 to B.R.

Institutional Review Board Statement: Data collection for all contributing studies was conducted according to the guidelines of the Declaration of Helsinki, and approved by the Institutional Review Board (or Ethics Committee) of the relevant institutions. IRB numbers are: MESA 16-2213; ARIC 16-2213; BioMe 16-2213; CARDIA 16-2213; GERA 14-2427; UK Biobank 16-2213; WHI 16-2213; HCHS/SOL 16-2213.

Informed Consent Statement: Informed consent was provided by all participants in the participating cohort and biobank studies.

Data Availability Statement: Data used in these analyses can be obtained either through the coordinating centers of the participating cohort and biobank studies or through dbGaP. Relevant dbGaP accession numbers are listed below.

BioMe phs000227 (MEGA) and phs000925

ARIC phs000557

CARDIA phs000613

WHI phs000227 (MEGA) and phs000386

HCHS/SOL phs000810

GERA phs000674

MESA phs001416, phs000209

GENOA phs000379

Acknowledgments: This research has been conducted using the UK Biobank Resource under Application Number 25953.

Support for title page creation and format was provided by AuthorArranger, a tool developed at the National Cancer Institute.

Whole genome sequencing (WGS) for the Trans-Omics in Precision Medicine (TOPMed) program was supported by the National Heart, Lung and Blood Institute (NHLBI). WGS for "NHLBI TOPMed: Multi-Ethnic Study of Atherosclerosis (MESA)" (phs001416.v1.p1) was performed at the Broad Institute of MIT and Harvard (3U54HG003067-13S1). Centralized read mapping and genotype calling, along with variant quality metrics and filtering were provided by the TOPMed Informatics Research Center (3R01HL-117626-02S1, contract HHSN268201800002I). Phenotype harmonization, data management, sample-identity QC, and general study coordination, were provided by the TOPMed Data Coordinating Center (R01HL-120393; U01HL-120393; contract HHSN268201800001I).

MESA and the MESA SHARe project are conducted and supported by the National Heart, Lung, and Blood Institute (NHLBI) in collaboration with MESA investigators. Support for MESA is provided by contracts 75N92020D00001, HHSN268201500003I, N01-HC-95159, 75N92020D00005, N01-HC-95160, 75N92020D00002, N01-HC-95161, 75N92020D00003, N01-HC-95162, 75N92020D00006, N01-HC-95163, 75N92020D00004, N01-HC-95164, 75N92020D00007, N01-HC-95165, N01-HC-95166, N01-HC-95167, N01-HC-95168, N01-HC-95169, UL1-TR-000040, UL1-TR-001079, UL1-TR-001420, UL1-TR-001881, and DK063491. The MESA Epigenomics & Transcriptomics Studies were funded by NIH grants 1R01HL101250, 1RF1AG054474, R01HL126477, R01DK101921, and R01HL135009.

Genotyping of the GERA cohort was funded by a grant from the National Institute on Aging, National Institute of Mental Health, and National Institute of Health Common Fund (RC2 AG036607).

The Population Architecture Using Genomics and Epidemiology (PAGE) program is funded by the National Human Genome Research Institute (NHGRI) with co-funding from the National Institute on Minority Health and Health Disparities (NIMHD). The contents of this paper are solely the responsibility of the authors and do not necessarily represent the official views of the NIH. The PAGE consortium thanks the staff and participants of all PAGE studies for their important contributions. We thank Rasheeda Williams and Margaret Ginoza for providing assistance with program coordination. The complete list of PAGE members can be found at http://www.pagestudy.org.

Assistance with data management, data integration, data dissemination, genotype imputation, ancestry deconvolution, population genetics, analysis pipelines, and general study coordination was provided by the PAGE Coordinating Center (NIH U01HG007419). Genotyping services were provided by the Center for Inherited Disease Research (CIDR). CIDR is fully funded through a federal contract from the National Institutes of Health to The Johns Hopkins University, contract number HHSN268201200008I. Genotype data quality control and quality assurance services were provided by the Genetic Analysis Center in the Biostatistics Department of the University of Washington, through support provided by the CIDR contract.

The data and materials included in this report result from collaboration between the following studies and organizations:

HCHS/SOL: Primary funding support to Dr. North and colleagues is provided by U01HG007416. Additional support was provided via R01DK101855 and 15GRNT25880008. The Hispanic Community Health Study/Study of Latinos was carried out as a collaborative study supported by contracts from the National Heart, Lung, and Blood Institute (NHLBI) to the University of North Carolina (N01-HC65233), University of Miami (N01-HC65234), Albert Einstein College of Medicine (N01-HC65235), Northwestern University (N01-HC65236), and San Diego State University (N01-HC65237). The following Institutes/Centers/Offices contribute to the HCHS/SOL through a transfer of funds to the NHLBI: National Institute on Minority Health and Health Disparities, National Institute on Deafness and Other Communication Disorders, National Institute of Dental and Craniofacial Research, National Institute of Diabetes and Digestive and Kidney Diseases, National Institute of Neurological Disorders and Stroke, NIH Institution-Office of Dietary Supplements. The Genetic Analysis Center at the University of Washington was supported by NHLBI and NIDCR contracts (HHSN268201300005C AM03 and MOD03).

WHI: Funding support for the "Exonic variants and their relation to complex traits in minorities of the WHI" study is provided through the NHGRI PAGE program (NIH U01HG007376). The WHI program is funded by the National Heart, Lung, and Blood Institute, National Institutes of Health, U.S. Department of Health and Human Services through contracts HHSN268201100046C,

HHSN268201100001C, HHSN268201100002C, HHSN268201100003C, HHSN268201100004C, and HHSN271201100004C. The authors thank the WHI investigators and staff for their dedication, and the study participants for making the program possible. A listing of WHI investigators can be found at:
https://www.whi.org/researchers/Documents%20%20Write%20a%20Paper/WHI%20Investigator%20Long%20List.pdf.

ARIC: The Atherosclerosis Risk in Communities study has been funded in whole or in part with Federal funds from the National Heart, Lung, and Blood Institute, National Institutes of Health, Department of Health and Human Services, under Contract nos. (HHSN268201700001I, HHSN268201700002I, HHSN268201700003I, HHSN268201700005I, HHSN268201700004I). The authors thank the staff and participants of the ARIC study for their important contributions.

CARDIA:The Coronary Artery Risk Development in Young Adults Study (CARDIA) is conducted and supported by the National Heart, Lung, and Blood Institute (NHLBI) in collaboration with the University of Alabama at Birmingham (HHSN268201800005I & HHSN268201800007I), Northwestern University (HHSN268201800003I), University of Minnesota (HHSN268201800006I), and Kaiser Foundation Research Institute (HHSN268201800004I).

Support for the Genetic Epidemiology Network of Arteriopathy (GENOA) data collection and analysis was provided by the National Heart, Lung and Blood Institute (U01HL054457, RC1HL100185, R01HL119443, R01HL133221, and R01HL141292) and the National Institute of Neurological Disorders and Stroke (R01NS041558) of the National Institutes of Health.

Finally, we would like to acknowledge use of the Trans-Omics in Precision Medicine (TOPMed) program imputation panel (freeze 5 version) supported by the National Heart, Lung and Blood Institute (NHLBI); see www.nhlbiwgs.org. TOPMed study investigators contributed data to the reference panel, which was accessed through the Michigan Imputation Server; see https://imputationserver.sph.umich.edu. The newest version of the imputation panel (which was not used in this study) can be found at https://imputation.biodatacatalyst.nhlbi.nih.gov. The panel was constructed and implemented by the TOPMed Informatics Research Center at the University of Michigan (3R01HL-117626-02S1; contract HHSN268201800002I). The TOPMed Data Coordinating Center (R01HL-120393; U01HL-120393; contract HHSN268201800001I) provided additional data management, sample identity checks, and overall program coordination and support. We gratefully acknowledge the studies and participants who provided biological samples and data for TOPMed.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Segal, J.B.; Moliterno, A.R. Platelet counts differ by sex, ethnicity, and age in the United States. *Annals of epidemiology* **2006**, *16*, 123-130, doi:10.1016/j.annepidem.2005.06.052.
2. Zakai, N.A.; McClure, L.A.; Prineas, R.; Howard, G.; McClellan, W.; Holmes, C.E.; Newsome, B.B.; Warnock, D.G.; Audhya, P.; Cushman, M. Correlates of anemia in American blacks and whites: the REGARDS Renal Ancillary Study. *American journal of epidemiology* **2009**, *169*, 355-364, doi:10.1093/aje/kwn355.
3. Lim, E.M.; Cembrowski, G.; Cembrowski, M.; Clarke, G. Race-specific WBC and neutrophil count reference intervals. *Int J Lab Hematol* **2010**, *32*, 590-597, doi:10.1111/j.1751-553X.2010.01223.x.
4. Reich, D.; Nalls, M.A.; Kao, W.H.; Akylbekova, E.L.; Tandon, A.; Patterson, N.; Mullikin, J.; Hsueh, W.C.; Cheng, C.Y.; Coresh, J.; et al. Reduced neutrophil count in people of African descent is due to a regulatory variant in the Duffy antigen receptor for chemokines gene. *PLoS genetics* **2009**, *5*, e1000360, doi:10.1371/journal.pgen.1000360.
5. Beutler, E.; West, C. Hematologic differences between African-Americans and whites: the roles of iron deficiency and alpha-thalassemia on hemoglobin levels and mean corpuscular volume. *Blood* **2005**, *106*, 740-745, doi:10.1182/blood-2005-02-0713.

6. Insights into malaria susceptibility using genome-wide data on 17,000 individuals from Africa, Asia and Oceania. *Nature communications* **2019**, *10*, 5732, doi:10.1038/s41467-019-13480-z.
7. Clarke, G.M.; Rockett, K.; Kivinen, K.; Hubbart, C.; Jeffreys, A.E.; Rowlands, K.; Jallow, M.; Conway, D.J.; Bojang, K.A.; Pinder, M.; et al. Characterisation of the opposing effects of G6PD deficiency on cerebral malaria and severe malarial anaemia. *eLife* **2017**, *6*, doi:10.7554/eLife.15085.
8. Hu, Y.; Stilp, A.M.; McHugh, C.P.; Rao, S.; Jai, D.; Zheng, X.; Lane, J.; Méric de Bellefon, S.; Raffield, L.M.; Chen, M.H.; et al. Whole-genome sequencing association analysis of quantitative red blood cell phenotypes: The NHLBI TOPMed program. *Am J Hum Genet* **2021**, doi:10.1016/j.ajhg.2021.04.003.
9. Astle, W.J.; Elding, H.; Jiang, T.; Allen, D.; Ruklisa, D.; Mann, A.L.; Mead, D.; Bouman, H.; Riveros-Mckay, F.; Kostadima, M.A.; et al. The Allelic Landscape of Human Blood Cell Trait Variation and Links to Common Complex Disease. *Cell* **2016**, *167*, 1415-1429 e1419, doi:10.1016/j.cell.2016.10.042.
10. Ganesh, S.K.; Zakai, N.A.; van Rooij, F.J.; Soranzo, N.; Smith, A.V.; Nalls, M.A.; Chen, M.H.; Kottgen, A.; Glazer, N.L.; Dehghan, A.; et al. Multiple loci influence erythrocyte phenotypes in the CHARGE Consortium. *Nat Genet* **2009**, *41*, 1191-1198, doi:10.1038/ng.466.
11. Vuckovic, D.; Bao, E.L.; Akbari, P.; Lareau, C.A.; Mousas, A.; Jiang, T.; Chen, M.H.; Raffield, L.M.; Tardaguila, M.; Huffman, J.E.; et al. The Polygenic and Monogenic Basis of Blood Traits and Diseases. *Cell* **2020**, *182*, 1214-1231.e1211, doi:10.1016/j.cell.2020.08.008.
12. Wainberg, M.; Sinnott-Armstrong, N.; Mancuso, N.; Barbeira, A.N.; Knowles, D.A.; Golan, D.; Ermel, R.; Ruusalepp, A.; Quertermous, T.; Hao, K.; et al. Opportunities and challenges for transcriptome-wide association studies. *Nature genetics* **2019**, *51*, 592-599, doi:10.1038/s41588-019-0385-z.
13. Gamazon, E.R.; Wheeler, H.E.; Shah, K.P.; Mozaffari, S.V.; Aquino-Michaels, K.; Carroll, R.J.; Eyler, A.E.; Denny, J.C.; Nicolae, D.L.; Cox, N.J.; et al. A gene-based association method for mapping traits using reference transcriptome data. *Nat Genet* **2015**, *47*, 1091-1098, doi:10.1038/ng.3367.
14. Gusev, A.; Ko, A.; Shi, H.; Bhatia, G.; Chung, W.; Penninx, B.W.; Jansen, R.; de Geus, E.J.; Boomsma, D.I.; Wright, F.A.; et al. Integrative approaches for large-scale transcriptome-wide association studies. *Nat Genet* **2016**, *48*, 245-252, doi:10.1038/ng.3506.
15. Barbeira, A.N.; Dickinson, S.P.; Bonazzola, R.; Zheng, J.; Wheeler, H.E.; Torres, J.M.; Torstenson, E.S.; Shah, K.P.; Garcia, T.; Edwards, T.L.; et al. Exploring the phenotypic consequences of tissue specific gene expression variation inferred from GWAS summary statistics. *Nature communications* **2018**, *9*, 1825, doi:10.1038/s41467-018-03621-1.
16. Wu, L.; Shi, W.; Long, J.; Guo, X.; Michailidou, K.; Beesley, J.; Bolla, M.K.; Shu, X.O.; Lu, Y.; Cai, Q.; et al. A transcriptome-wide association study of 229,000 women identifies new candidate susceptibility genes for breast cancer. *Nat Genet* **2018**, *50*, 968-978, doi:10.1038/s41588-018-0132-x.
17. Mancuso, N.; Gayther, S.; Gusev, A.; Zheng, W.; Penney, K.L.; Kote-Jarai, Z.; Eeles, R.; Freedman, M.; Haiman, C.; Pasaniuc, B. Large-scale transcriptome-wide association study identifies new prostate cancer risk regions. *Nature communications* **2018**, *9*, 4079, doi:10.1038/s41467-018-06302-1.
18. Andaleon, A.; Mogil, L.S.; Wheeler, H.E. Genetically regulated gene expression underlies lipid traits in Hispanic cohorts. *PloS one* **2019**, *14*, e0220827, doi:10.1371/journal.pone.0220827.
19. Bhattacharya, A.; Garcia-Closas, M.; Olshan, A.F.; Perou, C.M.; Troester, M.A.; Love, M.I. A framework for transcriptome-wide association studies in breast cancer in diverse study populations. *Genome biology* **2020**, *21*, 42, doi:10.1186/s13059-020-1942-6.

20. Patel, A.; García-Closas, M.; Olshan, A.F.; Perou, C.M.; Troester, M.A.; Love, M.I.; Bhattacharya, A. Gene-level germline contributions to clinical risk of recurrence scores in Black and White breast cancer patients. *medRxiv* **2021**, 2021.2003.2019.21253983, doi:10.1101/2021.03.19.21253983.
21. Fiorica, P.N.; Schubert, R.; Morris, J.D.; Abdul Sami, M.; Wheeler, H.E. Multi-ethnic transcriptome-wide association study of prostate cancer. *PloS one* **2020**, *15*, e0236209, doi:10.1371/journal.pone.0236209.
22. Geoffroy, E.; Gregga, I.; Wheeler, H.E. Population-Matched Transcriptome Prediction Increases TWAS Discovery and Replication Rate. *iScience* **2020**, *23*, 101850, doi:10.1016/j.isci.2020.101850.
23. GTEx Consortium; Battle, A.; Brown, C.D.; Engelhardt, B.E.; Montgomery, S.B. Genetic effects on gene expression across human tissues. *Nature* **2017**, *550*, 204-213, doi:10.1038/nature24277.
24. Battle, A.; Mostafavi, S.; Zhu, X.; Potash, J.B.; Weissman, M.M.; McCormick, C.; Haudenschild, C.D.; Beckman, K.B.; Shi, J.; Mei, R.; et al. Characterizing the genetic basis of transcriptome diversity through RNA-sequencing of 922 individuals. *Genome research* **2014**, *24*, 14-24.
25. Das, S.; Forer, L.; Schönherr, S.; Sidore, C.; Locke, A.E.; Kwong, A.; Vrieze, S.I.; Chew, E.Y.; Levy, S.; McGue, M.; et al. Next-generation genotype imputation service and methods. *Nat Genet* **2016**, *48*, 1284-1287, doi:10.1038/ng.3656.
26. Mostafavi, S.; Battle, A.; Zhu, X.; Urban, A.E.; Levinson, D.; Montgomery, S.B.; Koller, D. Normalizing RNA-sequencing data by modeling hidden covariates with prior knowledge. *PloS one* **2013**, *8*, e68141, doi:10.1371/journal.pone.0068141.
27. Mogil, L.S.; Andaleon, A.; Badalamenti, A.; Dickinson, S.P.; Guo, X.; Rotter, J.I.; Johnson, W.C.; Im, H.K.; Liu, Y.; Wheeler, H.E. Genetic architecture of gene expression traits across diverse populations. *PLoS genetics* **2018**, *14*, e1007586, doi:10.1371/journal.pgen.1007586.
28. Liu, Y.; Reynolds, L.M.; Ding, J.; Hou, L.; Lohman, K.; Young, T.; Cui, W.; Huang, Z.; Grenier, C.; Wan, M.; et al. Blood monocyte transcriptome and epigenome analyses reveal loci associated with human atherosclerosis. *Nature communications* **2017**, *8*, 393, doi:10.1038/s41467-017-00517-4.
29. Liu, Y.; Ding, J.; Reynolds, L.M.; Lohman, K.; Register, T.C.; De La Fuente, A.; Howard, T.D.; Hawkins, G.A.; Cui, W.; Morris, J.; et al. Methylomics of gene expression in human monocytes. *Hum Mol Genet* **2013**, *22*, 5065-5074, doi:10.1093/hmg/ddt356.
30. Stegle, O.; Parts, L.; Piipari, M.; Winn, J.; Durbin, R. Using probabilistic estimation of expression residuals (PEER) to obtain increased power and interpretability of gene expression analyses. *Nature protocols* **2012**, *7*, 500-507, doi:10.1038/nprot.2011.457.
31. Auton, A.; Brooks, L.D.; Durbin, R.M.; Garrison, E.P.; Kang, H.M.; Korbel, J.O.; Marchini, J.L.; McCarthy, S.; McVean, G.A.; Abecasis, G.R. A global reference for human genetic variation. *Nature* **2015**, *526*, 68-74, doi:10.1038/nature15393.
32. Shang, L.; Smith, J.A.; Zhao, W.; Kho, M.; Turner, S.T.; Mosley, T.H.; Kardia, S.L.R.; Zhou, X. Genetic Architecture of Gene Expression in European and African Americans: An eQTL Mapping Study in GENOA. *Am J Hum Genet* **2020**, *106*, 496-512, doi:10.1016/j.ajhg.2020.03.002.
33. Willer, C.J.; Li, Y.; Abecasis, G.R. METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics* **2010**, *26*, 2190-2191, doi:10.1093/bioinformatics/btq340.
34. Chen, M.H.; Raffield, L.M.; Mousas, A.; Sakaue, S.; Huffman, J.E.; Moscati, A.; Trivedi, B.; Jiang, T.; Akbari, P.; Vuckovic, D.; et al. Trans-ethnic and Ancestry-Specific Blood-Cell Genetics in 746,667 Individuals from 5 Global Populations. *Cell* **2020**, *182*, 1198-1213.e1114, doi:10.1016/j.cell.2020.06.045.

35. Benner, C.; Spencer, C.C.; Havulinna, A.S.; Salomaa, V.; Ripatti, S.; Pirinen, M. FINEMAP: efficient variable selection using summary data from genome-wide association studies. *Bioinformatics* **2016**, *32*, 1493-1501, doi:10.1093/bioinformatics/btw018.
36. Mbatchou, J.; Barnard, L.; Backman, J.; Marcketta, A.; Kosmicki, J.A.; Ziyatdinov, A.; Benner, C.; O'Dushlaine, C.; Barber, M.; Boutkov, B.; et al. Computationally efficient whole genome regression for quantitative and binary traits. *bioRxiv* **2020**, 2020.2006.2019.162354, doi:10.1101/2020.06.19.162354.
37. Uhlén, M.; Fagerberg, L.; Hallström, B.M.; Lindskog, C.; Oksvold, P.; Mardinoglu, A.; Sivertsson, Å.; Kampf, C.; Sjöstedt, E.; Asplund, A.; et al. Proteomics. Tissue-based map of the human proteome. *Science (New York, N.Y.)* **2015**, *347*, 1260419, doi:10.1126/science.1260419.
38. Ulirsch, J.C.; Lareau, C.A.; Bao, E.L.; Ludwig, L.S.; Guo, M.H.; Benner, C.; Satpathy, A.T.; Kartha, V.K.; Salem, R.M.; Hirschhorn, J.N.; et al. Interrogation of human hematopoiesis at single-cell and single-variant resolution. *Nat Genet* **2019**, *51*, 683-693, doi:10.1038/s41588-019-0362-6.
39. Javierre, B.M.; Burren, O.S.; Wilder, S.P.; Kreuzhuber, R.; Hill, S.M.; Sewitz, S.; Cairns, J.; Wingett, S.W.; Várnai, C.; Thiecke, M.J.; et al. Lineage-Specific Genome Architecture Links Enhancers and Non-coding Disease Variants to Target Gene Promoters. *Cell* **2016**, *167*, 1369-1384.e1319, doi:10.1016/j.cell.2016.09.037.
40. Kichaev, G.; Bhatia, G.; Loh, P.R.; Gazal, S.; Burch, K.; Freund, M.K.; Schoech, A.; Pasaniuc, B.; Price, A.L. Leveraging Polygenic Functional Enrichment to Improve GWAS Power. *Am J Hum Genet* **2019**, *104*, 65-75, doi:10.1016/j.ajhg.2018.11.008.
41. Gieger, C.; Radhakrishnan, A.; Cvejic, A.; Tang, W.; Porcu, E.; Pistis, G.; Serbanovic-Canic, J.; Elling, U.; Goodall, A.H.; Labrune, Y.; et al. New gene functions in megakaryopoiesis and platelet formation. *Nature* **2011**, *480*, 201-208, doi:10.1038/nature10659.
42. Hodonsky, C.J.; Baldassari, A.R.; Bien, S.A.; Raffield, L.M.; Highland, H.M.; Sitlani, C.M.; Wojcik, G.L.; Tao, R.; Graff, M.; Tang, W.; et al. Ancestry-specific associations identified in genome-wide combined-phenotype study of red blood cell traits emphasize benefits of diversity in genomics. *BMC genomics* **2020**, *21*, 228, doi:10.1186/s12864-020-6626-9.
43. Kowalski, M.H.; Qian, H.; Hou, Z.; Rosen, J.D.; Tapia, A.L.; Shan, Y.; Jain, D.; Argos, M.; Arnett, D.K.; Avery, C.; et al. Use of >100,000 NHLBI Trans-Omics for Precision Medicine (TOPMed) Consortium whole genome sequences improves imputation quality and detection of rare variant associations in admixed African and Hispanic/Latino populations. *PLoS genetics* **2019**, *15*, e1008500, doi:10.1371/journal.pgen.1008500.
44. Keys, K.L.; Mak, A.C.Y.; White, M.J.; Eckalbar, W.L.; Dahl, A.W.; Mefford, J.; Mikhaylova, A.V.; Contreras, M.G.; Elhawary, J.R.; Eng, C.; et al. On the cross-population generalizability of gene expression prediction models. *PLoS genetics* **2020**, *16*, e1008927, doi:10.1371/journal.pgen.1008927.
45. Võsa, U.; Claringbould, A.; Westra, H.-J.; Bonder, M.J.; Deelen, P.; Zeng, B.; Kirsten, H.; Saha, A.; Kreuzhuber, R.; Kasela, S.; et al. Unraveling the polygenic architecture of complex traits using blood eQTL metaanalysis. *bioRxiv* **2018**, 447367, doi:10.1101/447367.
46. Ramsingh, G.; Koboldt, D.C.; Trissal, M.; Chiappinelli, K.B.; Wylie, T.; Koul, S.; Chang, L.W.; Nagarajan, R.; Fehniger, T.A.; Goodfellow, P.; et al. Complete characterization of the microRNAome in a patient with acute myeloid leukemia. *Blood* **2010**, *116*, 5316-5326, doi:10.1182/blood-2010-05-285395.
47. Rusiniak, M.E.; Yu, M.; Ross, D.T.; Tolhurst, E.C.; Slack, J.L. Identification of B94 (TNFAIP2) as a potential retinoic acid target gene in acute promyelocytic leukemia. *Cancer research* **2000**, *60*, 1824-1829.

48. Zhao, D.; Deng, S.C.; Ma, Y.; Hao, Y.H.; Jia, Z.H. miR-221 alleviates the inflammatory response and cell apoptosis of neuronal cell through targeting TNFAIP2 in spinal cord ischemia-reperfusion. *Neuroreport* **2018**, *29*, 655-660, doi:10.1097/wnr.0000000000001013.
49. Sarma, V.; Wolf, F.W.; Marks, R.M.; Shows, T.B.; Dixit, V.M. Cloning of a novel tumor necrosis factor-alpha-inducible primary response gene that is differentially expressed in development and capillary tube-like formation in vitro. *Journal of immunology (Baltimore, Md. : 1950)* **1992**, *148*, 3302-3312.
50. Qu, X.; Li, Q.; Zhang, X.; Wang, Z.; Wang, S.; Zhou, Z. Amentoflavone protects the hematopoietic system of mice against γ-irradiation. *Archives of pharmacal research* **2019**, *42*, 1021-1029, doi:10.1007/s12272-019-01187-0.
51. Wolf, F.W.; Sarma, V.; Seldin, M.; Drake, S.; Suchard, S.J.; Shao, H.; O'Shea, K.S.; Dixit, V.M. B94, a primary response gene inducible by tumor necrosis factor-alpha, is expressed in developing hematopoietic tissues and the sperm acrosome. *The Journal of biological chemistry* **1994**, *269*, 3633-3640.
52. Cheng, Z.; Wang, H.Z.; Li, X.; Wu, Z.; Han, Y.; Li, Y.; Chen, G.; Xie, X.; Huang, Y.; Du, Z.; et al. MicroRNA-184 inhibits cell proliferation and invasion, and specifically targets TNFAIP2 in Glioma. *Journal of experimental & clinical cancer research : CR* **2015**, *34*, 27, doi:10.1186/s13046-015-0142-9.
53. Liu, Z.; Wei, S.; Ma, H.; Zhao, M.; Myers, J.N.; Weber, R.S.; Sturgis, E.M.; Wei, Q. A functional variant at the miR-184 binding site in TNFAIP2 and risk of squamous cell carcinoma of the head and neck. *Carcinogenesis* **2011**, *32*, 1668-1674, doi:10.1093/carcin/bgr209.