Generalization and fine mapping of red blood cell trait genetic associations to multi-ethnic populations: The PAGE Study

Authorship: Chani Jo Hodonsky, MPH[1], Claudia Schurmann, PhD[2,3], Ursula M Schick, PhD[2,3,4], Jonathan Kocarnik, PhD[4], Ran Tao, PhD[5], Frank JA van Rooij, PhD[6], Christina Wassel, PhD[7], Steve Buyske, PhD[8], Myriam Fornage, PhD[9], Lucia A Hindorff, PhD[10], James S Floyd, MD, MS[11,12], Santhi K Ganesh, MD[13,14], Dan-Yu Lin, PhD[15], Kari E North, PhD[1], Alex P Reiner, MD, MSc[4,12], Ruth JF Loos, PhD[2,3], Charles Kooperberg, PhD[4], Christy L Avery, PhD[1,16]

1 Department of Epidemiology, University of North Carolina Gillings School of Public Health, Chapel Hill, NC
2 The Charles Bronfman Institute for Personalized Medicine, Icahn School of Medicine at Mount Sinai, New York, NY
3 The Genetics of Obesity and Related Metabolic Traits Program, The Icahn School of Medicine at Mount Sinai, New York, NY
4 Division of Public Health Sciences, Fred Hutchinson Cancer Research Center, Seattle, WA
5 Department of Biostatistics, Vanderbilt University Medical Center, Nashville, TN
6 Department of Epidemiology, Erasmus University Medical Center, Rotterdam 3000, the Netherlands
7 Department of Pathology and Laboratory Medicine, College of Medicine, University of Vermont, Burlington, VT
8 Department of Statistics and Biostatistics, Hill Center, Rutgers, The State University of New Jersey, 110 Frelinghuysen Rd. Piscataway, NY
9 Institute of Molecular Medicine and Human Genetics Center, University of Texas Health Science Center at Houston, Houston, TX
10 Division of Genomic Medicine, National Human Genome Research Institute, National institutes of Health, Bethesda, MD
11 Departments of Medicine, University of Washington, Seattle, WA
12 Department of Epidemiology, University of Washington, Seattle, WA
13 Department of Human Genetics, University of Michigan Medical School, Ann Arbor, MI
14 Department of Internal Medicine, University of Michigan Medical School, Ann Arbor, MI
15 Department of Biostatistics, University of North Carolina, Chapel Hill, NC
16 Carolina Population Center, University of North Carolina, Chapel Hill, NC

Corresponding author: Chani J Hodonsky, chani_hodonsky@unc.edu
Voice: (919) 966-4312
Fax: (919) 966-9800
Address: 123 W Franklin St, St 4208-A, Chapel Hill, NC 27514

Abstract word count: 244
Manuscript word count: 4,729
Tables: 2
Figures: 1

Running title: Fine mapping of RBC traits in a multi-ethnic U.S. population

Key words: Genomics, fine-mapping, generalization, RBC traits, trans-ethnic meta-analysis

ABSTRACT

Red blood cell (RBC) traits provide insight into a wide range of physiological states and exhibit moderate to high heritability, making them excellent candidates for genetic studies to inform underlying biologic mechanisms. Previous RBC trait genome-wide association studies were performed primarily in European- or Asian-ancestry populations, missing opportunities to inform understanding of RBC genetic architecture in diverse populations and reduce intervals surrounding putative functional SNPs through fine-mapping. Here, we report the first fine-mapping of six correlated (Pearson's *r* range: |0.04 – 0.92|) RBC traits in up to 19,036 African Americans and 19,562 Hispanic/Latinos participants of the Population Architecture using Genomics and Epidemiology (PAGE) consortium. Trans-ethnic meta-analysis of race/ethnic- and study-specific estimates for approximately 11,000 SNPs flanking 13 previously identified association signals as well as 150,000 additional array-wide SNPs was performed using inverse-variance meta-analysis after adjusting for study and clinical covariates. Approximately half of previously reported index SNP-RBC trait associations generalized to the trans-ethnic study population ($p<1.7x10^{-4}$); previously unreported independent association signals within the *ABO* region reinforce the potential for multiple functional variants affecting the same locus. Trans-ethnic fine-mapping did not reveal additional signals at the *HFE* locus independent of the known functional variants. Finally, we identified a potential novel association in the Hispanic/Latino study population at the *HECTD4/RPL6* locus for RBC count ($p=1.9x10^{-7}$). The identification of a previously unknown association, generalization of a large proportion of known association signals, and refinement of known association signals all exemplify the benefits of genetic studies in diverse populations.

INTRODUCTION

Red blood cell (RBC) trait measurements are used to characterize the physiology of RBCs in both clinical and research settings, are captured by a complete blood count panel, and include the primary traits hematocrit (HCT), hemoglobin (HGB), and RBC count. Accompanying HCT, HGB, and RBC count are three derived traits—mean corpuscular hemoglobin (MCH), MCH concentration (MCHC), and mean corpuscular volume (MCV)—which can be used in combination with primary traits to evaluate RBC development and maintenance. Together, primary and derived RBC trait deficiencies (e.g., abnormally low HGB or excessive RBC count) cause circulatory diseases such as thalassemia, polycythemia, and genetic or nonhereditary anemias[1-5]. Population-specific *HBB* causal alleles for recessive diseases such as sickle-cell anemia and β-thalassemia have also been associated with protection against malaria and myocardial infarction, respectively, in the heterozygous state.[6-8] Additionally, RBC traits have been associated with stroke, cardiovascular disease (CVD) in populations with chronic kidney disease, and all-cause mortality.[9-12] RBC traits are therefore of substantial public health and clinical importance, yet their underlying pathophysiological mechanisms remain incompletely characterized.

As RBC traits exhibit moderate to high heritability (40-90%), population-based genetic analysis of these phenotypes can help identify causal alleles for and inform the underlying biology of RBC-related disorders.[3; 13; 14] To date, over 80 independent association signals with one or more RBC traits have been reported, primarily in studies of European- or Asian-ancestry populations[15-24]. One genome-wide association study (GWAS) performed in over 16,000 African

Americans identified 12 genome-wide-significant loci previously reported in European-ancestry or Japanese populations, indicating a shared role for common variants at RBC trait association signals[16]. However, fine-mapping of RBC trait associations identified in GWAS have had limited success narrowing broad GWAS signals to prioritize functional candidates due to large linkage disequilibrium (LD) blocks or characterize variants that are rare or monomorphic in Europeans or Asians, as has been demonstrated for platelet count[25; 26]. Narrowing and fine-mapping of previously identified association signals may be improved by performing analyses in ancestrally diverse populations with multi-continental admixture, including African Americans and Hispanic/Latinos[16; 17; 27].

Here, we evaluated 32 index SNP-RBC trait associations in 11 fine-mapped Metabochip regions, previously identified in populations of European-, Japanese-, and South Asian descent (*SPTA1, BCL11A, HFE, ABO, HK1, SH2B3/ATXN2, LIPC, PPCDC, NUTF2, NEUROD2,* and *TMPRSS6*) for evidence of generalization and locus refinement in African American and Hispanic/Latino participants of the Population Architecture using Genomics and Epidemiology (PAGE) consortium[28]. Additionally, we evaluated all SNPs genotyped on the Metabochip for associations not previously described in any of the six RBC traits. These efforts will help address gaps in understanding the genetic underpinnings of RBC traits.

MATERIALS AND METHODS

Study Populations

The PAGE consortium is a National Human Genome Research Institute funded effort to examine the epidemiologic architecture of genetic variants associated with human diseases and traits across diverse populations[29]. The following PAGE I studies contributed to this manuscript (Supplemental Materials and Methods): the Atherosclerosis Risk in Communities Study (ARIC),[30] the Coronary Artery Risk Disease in Young Adults study (CARDIA),[31] the Cardiovascular Health Study (CHS),[32] the Hispanic Community Health Study/Study of Latinos (HCHS/SOL),[33] and the Women's Health Initiative (WHI)[34]. The Icahn Mt. Sinai School of Medicine (MSSM) contributed both African American and Hispanic/Latino study populations separately from PAGE I[29]. The Institutional Review Board at all participating institutions approved the study protocol and all participants gave written consent.

Genotype platforms

The Metabochip was a custom Illumina iSELECT array that contained approximately 195,000 SNPs and was designed to support large scale follow up of putative associations for cardiovascular and metabolic traits[28]. Further information on genotyping and quality control is provided in the supplemental material. We defined an index SNP as a SNP reported in the GWAS catalog prior to October 1, 2016 as having a genome-wide significant association ($5x10^{-8}$) with at least one of the six RBC traits we evaluated. Index SNPs that were not directly genotyped on the Metabochip were represented by proxies, defined as SNPs in high ($r^2 \geq 0.80$) with the GWAS index SNP in the ancestral population in which the association was first reported. For one index SNP, rs671 (*ALDH2*), no proxy was available because this variant is

specific to populations of East Asian ancestry. A total of 74% of participants were directly genotyped on the Illumina custom Metabochip array; genotypes for the remaining participants were imputed from the Affymetrix 6.0 panel[35]. After QC and study-population-specific effective heterozygosity criteria were applied, 163,929 SNPs were available for analysis in African Americans and 159,467 SNPs were available for analysis in Hispanics/Latinos.

Statistical Analysis

We performed four types of analysis: (1) generalization, whereby we examined 32 index SNP associations across six RBC traits; (2) fine-mapping of association signals that generalized in (1); (3) testing for independent association signals for any RBC trait within one of the 11 densely genotyped regions; and (4) discovery of previously unreported associations with any RBC trait in all remaining Metabochip regions. Only SNPs meeting an effective heterozygosity of 35 were used within each race/ethnic study population; 4,814 SNPs were excluded in Hispanics/Latinos but included for African Americans, whereas 9,431 were excluded for African Americans but included for Hispanics/Latinos. We examined a maximum of 8,082 SNPs in African Americans and 7,991 SNPs in Hispanics/Latinos (9,201 SNPs total) within one of 11 regions densely genotyped on the Illumina Metabochip for all non-discovery analyses.

To interpret fine-mapping results, LD was calculated in 500kb sliding windows using PLINK (http://pngu.mgh.harvard.edu/purcell/plink) and African American (ARIC data), Hispanic/Latino (HCHS/SOL data), and trans-ethnic panels (randomly sampled ARIC and HCHS/SOL participant data in proportion to the racial/ethnic-specific sample population sizes)[36]. In addition, Metabochip LD and frequency information (but not individual-level information) was provided

by the Malmö Diet and Cancer Study on 2,143 control participants from a Swedish population to facilitate LD and MAF comparisons between PAGE African American and Hispanic/Latino populations and populations of European ancestry[37]. We used NCBI build 36 positions for regional association plots. Recombination rates were estimated from the combined HapMap phase II data.

A weighted version of generalized estimating equations (GEE; HCHS/SOL) accommodating the HCHS/SOL sampling design, relatedness, and household structure was implemented in SUGEN[38]. Race/ethnic-stratified linear regression was performed for all other studies (Atherosclerosis Risk in Communities [ARIC], Coronary Artery Risk Development in Young Adults [CARDIA], Cardiovascular Health Study [CHS], Icahn Mt. Sinai School of Medicine BioMe Biobank [MSSM], and WHI) using PLINK[36]. We evaluate the association between each quantitative RBC trait (see Supplement for RBC trait measurement methods and calculations for derived equations, **Table S2**) and a maximum of 9,201 SNPs (racial/ethnic- and study-specific effective heterozygosity >30, present in more than one study in either African Americans or Hispanics/Latinos) from 11 previously identified RBC trait loci. An additive genetic model was assumed including age, sex, study center/region, and ten ancestry principal components. Racial/ethnic-stratified and trans-ethnic study-specific association results were combined via inverse variance meta-analysis as implemented in METAL[39]. Genomic inflation factors were not calculated as the design of the Metabochip purposefully emphasizes potential functional candidates, leading to expected early departure from a uniform p-value distribution.

Generalization

We defined an "association signal" as a set of SNPs genotyped in a Metabochip fine-mapped region and exhibiting linkage disequilibrium ($r^2 \geq 0.2$ in the Malmö Diet and Cancer Study) with a previously reported genome-wide significant SNP for one or more RBC traits. For two or more previously reported genome-wide-significant SNPs to be considered within the same association signal in our study, those variants had to be in LD.

We next defined an "index SNP" as the most significant previously reported SNP within an association signal for each RBC trait. In instances for which multiple SNPs were published as the most significant SNP within a particular association signal for the same trait, we selected the SNP with the lowest reported p-value as the index SNP. The index SNP within an association signal may vary by trait due to differences in sample size, measurement error, and allelic heterogeneity among other possible reasons related to genetic architecture of the traits. Therefore, we evaluated the most significant SNP reported for each association signal-trait combination rather than selecting one index SNP to evaluate in all traits for which that association signal was previously reported as genome-wide-significant, even though some of the index SNPs likely tag the same genetic association across multiple RBC traits. For example, the *SH2B3/ATXN2* association signal has been reported for multiple RBC traits with the most significant SNP differing by trait, meaning the index SNP for RBC count is rs3184504 whereas the index SNP for hematocrit is rs11065987. These two SNPs are in LD and likely represent the same association signal. Furthermore, while several RBC trait associations examined in this paper were first reported in Japanese populations, those associations have since been generalized to European populations. European LD blocks are typically larger than for African or admixed-ancestry haplotypes, therefore we used European LD to conservatively define loci when

analyzing potential independent associations in fine-mapped regions containing previously reported RBC trait GWAS associations.

We then evaluated whether association signals identified in populations of European, Japanese, or South Asian ancestry generalized to African American and Hispanic/Latino populations. Approximately 44% of all previously reported genome-wide-significant RBC trait SNPs, but only 13% of reported association signals (defined above based on European LD, identified as of September 2016), were located in fine-mapped regions on the Metabochip[15-24]. The generalization significance criterion was then defined as $\alpha = 1.7 \times 10^{-4}$, a Bonferroni-corrected threshold calculated using 294 tag SNPs in African Americans ($r^2 \geq 0.80$; determined using African American LD from the ARIC Study) that captured all SNPs correlated with the index SNPs representing 32 index SNP-trait associations as identified in the Malmö Diet Study population.

Fine-Mapping Generalized Associations

We evaluated association-signal narrowing across ancestral backgrounds by comparing the number of SNPs in high LD with the trans-ethnic lead SNP, as well as the width of the region covered by the high-LD SNPs (**Table 2, Figures 1, S2**). LD for African Americans was calculated using ARIC study participants; LD for Hispanics/Latinos was calculated using HCHS/SOL study participants.

Independent and Discovery SNP Identification

To identify independent SNPs influencing RBC traits, we identified all SNPs at the 11 RBC trait loci that were uncorrelated with the index SNPs ($r^2 < 0.20$ in the Malmö Diet and Cancer Study). Sequential conditional analyses were then performed by adjusting for significant racial/ethnic-specific lead SNPs. If a statistically significant association was identified, defined as 0.05 divided by the number of SNPs in African Americans with MAF ≥ 0.01 that were uncorrelated with the index SNPs (n=8,907; $\alpha = 5.61 \times 10^{-6}$), the SNP was identified as independent and added to the adjustment set. Sequential conditional analysis was repeated until no significant SNPs were identified. We evaluated all remaining SNPs for discovery in African Americans or Hispanics/Latinos using a Metabochip-wide significant threshold of 0.05/155,022 (the number of SNPs available for evaluation after exclusion of SNPs evaluated for generalization), or $\alpha = 3.23 \times 10^{-7}$, and only considered SNPs with an effective heterozygosity >30 in more than one cohort study population per race/ethnicity.

Bioinformatic Characterization of RBC Trait loci

For each of the significant RBC trait SNPs (i.e., any lead SNP that generalized in one or both race/ethnic populations or the trans-ethnic meta-analysis; or any novel SNP identified in either race/ethnic population), all SNPs in LD ($r^2 \geq 0.8$) were identified in the appropriate 1000 Genomes reference superpopulations (AFR [Africans] for African Americans and AMR [Admixed Americans] for Hispanic/Latinos) for functional annotation. Using HaploRegV2[40], all variants in each LD block were characterized with putative functional roles including: conservation; promoter and/or enhancer epigenetic markers, derived from the Roadmap Epigenomics Project[41] and ENCODE;[42] DNAse hypersensitive sites; and transcription factor binding motifs calculated as a library of position weight matrices.[43-45] Evidence of functional

activity considered promoter and enhancer regions based on histone modification patterns in k562 erythroleukemia cells in the Roadmap Epigenome Project; DNase hypersensitive sites in ENCODE tissues including erythroblasts and erythroid leukemia cell lines; and transcription factor binding evidence in k562 erythroid leukemia cells. Evidence of cis-eQTL status was performed using Blueprint for relevant blood tissues[46]. All functional elements that varied by cell type were restricted to RBC-relevant tissues (**Tables S12a, S12b**).

In order to evaluate the relevance of trans-ethnic PAGE lead SNPs across tissue types, we compared the eQTL status of index SNPs in both blood-relevant and other tissues (**Table S13**). We looked up significant eQTLs for each index SNP ($p<1E-06$) in whole blood in **GTEx**, which provides data on a wide array of tissues; and two blood-specific eQTL databases: **the blood eQTL browser**; and the **NESDA NTR Conditional eQTL catalog**[47-49]. Only GTEx tissues which showed an association with at least one SNP are reported. We further reported clinical relevance of the trans-ethnic lead SNPs as described in the literature[50; 51].

RESULTS

We analyzed six correlated RBC traits (Pearson's correlation coefficient range: -0.29 to 0.92 in Hispanic Community Health Study/Study of Latinos (HCHS/SOL) participants; **Tables S1, S2**) in a maximum of 19,036 African American and 19,562 Hispanic/Latino participants from six studies participating in the PAGE consortium (**Table S3**). Females were over-represented among both African Americans (83%) and Hispanic/Latinos (70%). The HCHS/SOL (n=11,675) and Women's Health Initiative (WHI, n=17,363, of which 12,022 are African American) studies contributed the largest proportion of Hispanic/Latino (60%) and African American (63%) participants, respectively.

Generalization and fine-mapping of 11 densely genotyped Metabochip regions

Generalization

First, we examined 11 regions densely genotyped on the Illumina Metabochip and harboring one or more variants previously associated at genome-wide significance ($p<5x10^{-8}$) with at least one RBC trait (**Table S4**). All but two of the 11 regions contained one association signal, with the *HFE* and *ABO* regions each containing two association signals (see Methods). Of these 13 association signals, eight were previously associated with two or more RBC traits and two were previously associated with four traits, for a total of 32 index SNP-trait associations (**Tables 1, S5, S6**).

Seventeen of the 32 index SNP-trait associations (53%) generalized at $p<1.7x10^{-4}$ to the trans-ethnic study population (**Tables 1, S7, S8**), of which six trans-ethnic lead SNPs were identical to the previously reported index SNP. Of the remaining 11 generalized associations, nine trans-ethnic lead SNP p-values exceeded the index SNP p-values by at least an order of magnitude (**Tables 1, S10**). Effect sizes for both generalized and non-generalized association signals for index SNPs and trans-ethnic lead SNPs were consistent with previously reported estimates (**Table S6**)[17].

The first *HFE* association signal (index SNP: rs198846) generalized with the same trans-ethnic lead SNP (rs1799945, the functional H63D hemochromatosis variant) to all three previously reported traits—HGB, MCH, and MCV. Furthermore, both *ABO* association signals (**Figure 1**) and the *SH2B3/ATXN2* association signal generalized to all traits except RBC count. Notably, RBC count was the only trait for which none of the index SNP generalized in the trans-ethnic population; it also was the trait with the smallest sample size (46% of the maximum number of participants). Association signals for *SPTA1, BCL11A, LIPC, NUTF2, PPCDC,* and *NEUROD2* did not generalize. Six non-generalized index SNP-trait associations could not be evaluated for directional consistency because a proxy SNP was used in generalization analyses or the effect size was not reported in the initial publication (**Table S9**). For the remaining eight non-generalized index SNP-trait associations with sufficient information to evaluate directional consistency, six were directionally consistent in the trans-ethnic population. Additionally, when compared to SNP-trait associations from a previously published RBC trait GWAS, seven of 11 PAGE lead SNPs exceeded the generalization significance threshold in 24,167 participants of the CHARGE consortium (**Table S14**)[17].

In race/ethnicity-specific meta-analyses (**Tables 2, S6-S8**), 9% (n=3) of index SNP-trait associations generalized to African Americans and generalization was limited to HGB. Conversely, 38% (n=12) of index SNP-trait associations generalized to Hispanics/Latinos, including the *SH2B3/ATXN2* association with RBC count, representing the only instance where evidence of generalization for the RBC trait was detected. Of note, HCT, HGB, and MCHC were reported for similar numbers of Hispanics/Latino and African American participants in our study population, whereas MCH, MCV, and RBC count were reported for more than twice as many Hispanics/Latinos as African Americans, potentially contributing to the disparity in generalization by race/ethnicity.

Novel independent signals in 11 fine-mapped regions

We next evaluated the 11 fine-mapped regions to identify significant variants independent of published association signals by examining all SNPs that were uncorrelated with any of the index SNPs (see Methods). We identified no independent associations in previously reported regions for any of the six RBC traits (significance threshold: $p<1.3x10^{-5}$).

Fine-Mapping

To fine-map association signals that generalized, we then evaluated the LD structure in the trans-ethnic study population and by race/ethnicity (**Figures 1 & S2, Table 2**). The median reduction in interval width was 75%, likely due to the large reduction in association signals for which the index variant was functional but fell within a large LD block in Europeans. The first *HFE* association signal showed consistent evidence of narrowing across three traits (113kb

decrement), with the same trans-ethnic lead SNP for all traits (the causal H63D variant rs1799945, CAF = 0.97 in African Americans, CAF = 0.88 in Hispanics/Latinos). Both *ABO* association signals (the latter of which has the determining variant for blood type B, rs8176746, as the published index SNP) fine-mapped to a limited number of SNPs in narrow LD blocks in the trans-ethnic study population, with the trans-ethnic lead SNP varying by trait. Rs855791 is a known functional coding variant in *TMPRSS6*, therefore we do not consider this signal to be narrowed.

Discovery

Next, all SNPs outside the 11 previously identified RBC trait-associated regions were evaluated for evidence of discovery associations ($p < 3.03x10^{-7}$, see Methods). No SNP association exceeded Metabochip-wide significance in the total trans-ethnic study population or among African Americans for any RBC trait. However, one previously unreported association met Metabochip-wide significance in Hispanics/Latinos: rs76350043 at *HECTD4* / *RPL6* (chr 12q24.13, $p=2.5x10^{-7}$) for RBC count (**Table S11**). This SNP was also nominally significant for HCT and HGB ($p < 0.05$).

In Silico Bioinformatics Analysis

All SNPs highly correlated ($r^2 \geq 0.9$ in relevant AFR or AMR 1000 Genomes Phase I ancestral populations) with trans-ethnic lead SNPs from generalization analysis were examined using publicly available functional prediction data for erythrocytes or erythroblastoid cell lines, as well as pathogenicity prediction (**Tables S12a, S12b**)[52]. With the exception of the well-established *TMPRSS6* missense variant rs855791 (generalized to HCT, HGB, MCH, MCHC, and MCV), all

trans-ethnic lead variants and their LD proxies were noncoding variants. Lead SNPs and their LD proxies most commonly exhibited potential regulatory effects including disruption of RBC-relevant transcription factor consensus sequences and sites exhibiting DNase I activity. Several SNPs at generalized loci exhibited promise for molecular characterization, including rs198851, the HCT lead SNP in Hispanics/Latinos at the first *HFE* association signal. In k562 erythroid leukemia cells, rs198851 exhibits both DNase and enhancer activity, is an eQTL for *TRIM38,* and is located within an RNA Polymerase II ChIP-seq peak. With the exception of one association signal (*TMPRSS6* in both African Americans and Hispanics/Latinos), all independent signals contained at least one SNP with evidence for a regulatory function or cis-eQTL activity in relevant blood cell types (**Tables S12a, S12b**).

We also evaluated tissue specificity of significant eQTLs (p<1E-06) for each published index SNP or trans-ethnic lead SNP for all generalized association signals, as well as the putative clinical relevance of each SNP when information was available (**Table S13**)[47-51]. EQTL results show varied evidence of tissue expression effects. Lead or index variants for *SH2B3*/*ATXN2* and *TMPRSS6* fine-mapped regions demonstrated no significant association with any gene expression in any tissue type. In contrast, SNPs within the second *ABO* association signal showed evidence of broad *ABO* expression across 30 tissue types. SNPs in the first *ABO* association signal show evidence of expression for multiple genes, but across fewer tissues than the second signal. Lead and index SNPs within the second *HFE* association signal were only associated with expression of genes other than *HFE* across a broad array of GTEx tissues; no lead or index SNPs for the second *HFE* association signal exhibited eQTL activity for the *HFE* gene in any tissue type. The first *HFE* association signal showed some evidence of tissue

specificity in gene expression profiles—both the index SNP and trans-ethnic lead SNP exhibited cis-eQTLs for either *HFE* or several other genes, but overlap by tissue type was uncommon in this association signal.

DISCUSSION

In this study we performed generalization, fine-mapping, and discovery analysis of six RBC traits in a population of over 38,000 African American and Hispanic/Latino PAGE participants. We demonstrated that genetic regions influencing RBC traits identified in European- and Asian-ancestry populations are also applicable to African American and Hispanic/Latino populations. The merits of incorporating multi-ethnic study populations in genomic studies were also displayed via locus refinement and identification of a previously unreported RBC trait association that warrants validation in future studies.

In the eleven fine-mapped regions we evaluated, over half of known index SNP-trait associations generalized to the trans-ethnic study population across all six RBC traits, indicating that the effects of known RBC loci are likely shared across ancestral populations. Additionally, ten of 17 generalized associations (59%) met or exceeded the more stringent genome-wide significance threshold of $5x10^{-8}$ in the trans-ethnic study population. Although some association signals showed variation in lead SNP, the trans-ethnic lead SNPs almost always matched across traits when we restricted to participants with all traits measured (results not shown). The higher proportion of generalized associations in Hispanics/Latinos compared to African Americans suggests that results in Hispanic/Latino populations may contribute disproportionately to the larger trans-ethnic findings. This was not surprising given the Metabochip design that was enriched for European ancestral content as well as Hispanic/Latino genetic architecture, which shares more features with European-ancestry or Asian-ancestry individuals than does African American architecture[53]. Additionally, the SNPs designated as proxies for index SNPs

discovered in European- or Japanese-ancestry individuals were almost always in much lower LD in African Americans than Hispanics/Latinos, suggesting that previously reported index SNPs are not highly effective for characterizing the genetic architecture of RBC traits in African Americans.

We also detected several instances where trans-ethnic lead SNPs showed considerably stronger evidence of association with RBC traits in our study population than previously reported GWAS index SNPs identified in primarily European or East Asian populations. By examining visualizations of generalized association signals, we further identified several cases in which the lead SNP in LD with the European index SNP was not the most significant SNP in the region, indicating differences. These findings are consistent with recent work in large trans-ethnic populations, which demonstrated considerable effect heterogeneity by genetic ancestry in GWAS index SNPs reported in studies of predominantly European populations; considerably less evidence of heterogeneity was detected when examining index SNPs identified in multi-ethnic populations[54]. Of particular relevance are recent demonstrations of inappropriate designation of variants which are rare in European populations as pathogenic when they are in fact common in other ancestral groups[55]. GWAS inclusive of diverse populations can improve the accuracy of identifying functional variants, but fine-mapping is particularly well-suited to this type of exercise. As interest in genetic risk scores for RBC traits increases, e.g., to predict adverse outcomes in pregnancy, cardiovascular and neurologic diseases, and mortality, studies examining generalization of reported loci to global populations will become even more important, particularly in the era of precision medicine[22; 56-59].

Over the past decade, GWAS have identified hundreds of loci associated with RBC traits, but these findings incompletely account for the population-level variability attributable to additive genetic effects. A possible explanation for this missing heritability is that all genes expressed within RBC-relevant tissues play a role in RBC trait biology, but their identification may require infinite statistical power[60]. A recent review described the suite of genes affecting complex traits as including both "core" genes (i.e., those with tissue-specific effects crucial to one or few complex traits) and "peripheral" genes (i.e., those with broad expression profiles playing a role in many traits)[61]. Distinguishing core from peripheral genes may inform canonical pathways for RBC traits, provide mechanistic insight into biology, and inform targets for pharmaceutical intervention[60; 61]. Importantly, this designation occurs on a spectrum, with some genes not clearly predisposed to one class over the other. For example, hexokinase 1 (*HK1*) is highly and ubiquitously expressed, and several GWAS have identified associations within 200kb of *HK1* for psychiatric phenotypes, autoimmune disorders, and blood metabolite levels[17; 62]. However, *HK1* was also the only generalized association signal in our study with evidence of a blood-specific eQTL, which localized to a narrow segment of intron four representing GWAS study populations of multiple ancestries. The approximately 10kb segment contains multiple regulatory elements (e.g., DNase hypersensitivity regions and histone methylation marks), but GWAS findings to-date for this region remain restricted to RBC traits, and RBC trait index SNPs within 500kb of this region remain restricted to this narrow genomic fragment. These results reinforce the concept that tissue-specific regulators may play an important role for individual complex traits in broadly expressed genes. In light of this information and other complex-trait GWAS findings, tissue-specific expression data and genomic information will be particularly relevant when considering candidate variants for functional studies.

Large-scale genetic evaluation of correlated traits is challenging, particularly when evaluating multiple populations and traits with variable sample sizes. Importantly, novel statistical methods scalable to GWAS that leverage correlation among phenotypes for novel locus discovery have been reported[56-59; 63]. Such approaches seem particularly well suited for RBC traits, given evidence of a shared genetic architecture and the number of GWAS associations which have been reported in multiple RBC traits[17; 22]. Regarding fine-mapping, extensions of correlated-phenotype methods were recently described and have similarly shown promise for reducing sets of SNPs for functional evaluation over single trait methods. However, no studies to date have leveraged such innovations for discovery or fine-mapping of RBC traits.

Finally, we identified a potential novel association at the *HECTD4* / *RPL6* locus for RBC count. The *HECTD4* / *RPL6* locus has been previously associated with MCV (which exhibits modest correlation with RBC count and HGB), blood pressure, coronary heart disease, and multiple metabolic traits[64-68]. Additionally, coding mutations within several members of the ribosomal protein gene family have been causally associated with Diamond-Blackfan anemia, making an association with RBC count plausible[69; 70]. This association signal fell within a sparsely genotyped region on the Metabochip, and hence could not be further evaluated via fine-mapping. Evidence of association with multiple non-RBC traits should motivate larger efforts to understand whether the mechanisms underlying these associations are shared across traits or whether, for instance, tissue-specific effects relevant to each trait are represented by the same signal. Multi-ethnic fine-mapping to narrow the association signal for molecular characterization

likely represents an ideal first-step, as functional variants in this region have not been described for other associated traits.

This study faced several limitations which deserve consideration. First, phenotype availability differed by study and smaller sample sizes for MCH, MCV, and RBC count likely reduced power , specifically among the African American study population. Second, five of eleven fine-mapped regions we evaluated (*SPTA1*, *BCL11A*, *HK1*, *LIPC*, and *TMPRSS6*) also were mapped narrowly (<100kb) with few SNPs, potentially providing insufficient coverage of African American or Hispanic/Latino genetic content to perform comprehensive fine-mapping and generalization analyses. Sparse coverage in the *SPTA1* and *BCL11A* regions could also contribute to lack of generalization at these loci, at which the respective genes have established, functional roles in RBC development and maintenance[71-73]. With regard to bioinformatic characterization for functional candidate SNP evaluation, eQTL analysis—while insufficient as the sole determinant of tissue specificity—is an important component for ascertaining functional status of candidate variants. Finally, the Metabochip design emphasized regions identified for cardiometabolic traits, so overlap with RBC-trait associations was coincidental; we therefore could not examine generalization or fine-mapping in several well established RBC trait associations, including *HBS1L/MYB*, *LUC7L/ITGF3*, *HBA1/2*, and *HBB*[17; 18; 21; 74].

CONCLUSION

Population-based GWAS emphasize discovery, and are often the first step toward elucidating the genetic architecture underlying complex quantitative traits like RBC traits. Fine-mapping previously reported associations—particularly associations identified in genetically homogeneous populations, including European- and East Asian-ancestry populations—provides additional information about known association signals and can lead to narrowing of broad association signals to reduce the burden for bioinformatics and molecular functional analysis. Additional characterization of genetic associations contributing to population-level variability of RBC traits through large-scale sequencing and methods exploiting the correlation of RBC traits may further illuminate biological pathways for these complex quantitative traits.

ACKNOWLEDGMENTS

(a) The Population Architecture Using Genomics and Epidemiology (PAGE) program is funded by the National Human Genome Research Institute (NHGRI), supported by U01HG004803 (CALiCo), U01HG004790 (WHI), and U01HG004801 (Coordinating Center), and their respective NHGRI ARRA supplements. The contents of this paper are solely the responsibility of the authors and do not necessarily represent the official views of the NIH. The complete list of PAGE members can be found at http://www.pagestudy.org.

(b) The data and materials included in this report result from a collaboration between the following studies:

Funding support for the "Epidemiology of putative genetic variants: The Women's Health Initiative" study is provided through the NHGRI PAGE program (U01HG004790 and its NHGRI ARRA supplement). The WHI program is funded by the National Heart, Lung, and Blood Institute; NIH; and U.S. Department of Health and Human Services through contracts N01WH22110, 24152, 32100-2, 32105-6, 32108-9, 32111-13, 32115, 32118-32119, 32122, 42107-26, 42129-32, and 44221. The authors thank the WHI investigators and staff for their dedication, and the study participants for making the program possible. A full listing of WHI investigators can be found at: http://www.whiscience.org/publications/ WHI_investigators_shortlist.pdf.

Funding support for the Genetic Epidemiology of Causal Variants Across the Life Course (CALiCo) program was provided through the NHGRI PAGE program (U01HG004803 and its NHGRI ARRA supplement). The following studies contributed to this manuscript and are funded by the following agencies: The Atherosclerosis Risk in Communities (ARIC) Study is carried out as a collaborative study supported by National Heart, Lung, and Blood Institute contracts N01-HC-55015, N01-HC-55016, N01-HC-55018, N01-HC-55019, N01-HC-55020, N01-HC-55021, N01-HC-55022. The Cardiovascular Health Study (CHS) is supported by contracts HHSN268201200036C, HHSN268200800007C, N01 HC55222, N01HC85079, N01HC85080, N01HC85081, N01HC85082, N01HC85083, N01HC85086, and grants HL080295 and HL087652 from the National Heart, Lung, and Blood Institute (NHLBI), with additional contribution from the National Institute of Neurological Disorders and Stroke (NINDS). Additional support was provided by AG023629 from the National Institute on Aging (NIA). A full list of principal CHS investigators and institutions can be found at http://www.chs-nhlbi.org/PI.htm. CHS GWAS DNA handling and genotyping at Cedars-Sinai Medical Center was supported in part by the National Center for Research Resources, grant UL1RR033176, and is now at the National Center for Advancing Translational Sciences, CTSI grant UL1TR000124; in addition the National Institute of Diabetes and Digestive and Kidney Diseases grant DK063491 to the Southern California Diabetes Endocrinology Research Center.

Assistance with phenotype harmonization, SNP selection and annotation, data cleaning, data management, integration and dissemination, and general study coordination was provided by the PAGE Coordinating Center (U01HG004801-01 and its NHGRI ARRA supplement). The National Institutes of Mental Health also contributes to the support for the Coordinating Center.

The PAGE consortium thanks the staff and participants of all PAGE studies for their important contributions.

Works Cited

1. Taliaferro, W.H., and Huck, J.G. (1923). The Inheritance of Sickle-Cell Anaemia in Man. Genetics 8, 594-598.
2. Williamson, G.R., and Crawford, R. (1945). Fatal Mediterranean (Cooley's) anemia. New Orleans Med Surg J 98, 280-284.
3. Whitfield, J.B., and Martin, N.G. (1985). Genetic and environmental influences on the size and number of cells in the blood. Genet Epidemiol 2, 133-144.
4. Lamson, P.D. (1916). The Processes Taking Place in the Body by Which the Number of Erythrocytes Per Unit Volume of Blood is Increased in Acute Experimental Polycythaemia. Proc Natl Acad Sci U S A 2, 365-369.
5. Neel, J.V., and Valentine, W.N. (1947). Further Studies on the Genetics of Thalassemia. Genetics 32, 38-63.
6. Chami, N., and Lettre, G. (2014). Lessons and Implications from Genome-Wide Association Studies (GWAS) Findings of Blood Cell Phenotypes. Genes 5, 51-64.
7. Hashemi, M., Shirzadi, E., Talaei, Z., Moghadas, L., Shaygannia, I., Yavari, M., Amiri, N., Taheri, H., Montazeri, H., and Shamsolkottabi, H. (2007). Effect of heterozygous beta-thalassaemia trait on coronary atherosclerosis via coronary artery disease risk factors: a preliminary study. Cardiovascular journal of Africa 18, 165-168.
8. Wang, C.H., and Schilling, R.F. (1995). Myocardial infarction and thalassemia trait: an example of heterozygote advantage. American journal of hematology 49, 73-75.
9. Franczuk, P., Kaczorowski, M., Kucharska, K., Franczuk, J., Josiak, K., Zimoch, W., Kosowski, M., Reczuch, K., Majda, J., Banasiak, W., et al. (2015). Could an analysis of mean corpuscular volume help to improve a risk stratification in non-anemic patients with acute myocardial infarction? Cardiol J.
10. Panwar, B., Judd, S.E., Warnock, D.G., McClellan, W.M., Booth, J.N., 3rd, Muntner, P., and Gutierrez, O.M. (2016). Hemoglobin Concentration and Risk of Incident Stroke in Community-Living Adults. Stroke 47, 2017-2024.
11. Barlas, R.S., Honney, K., Loke, Y.K., McCall, S.J., Bettencourt-Silva, J.H., Clark, A.B., Bowles, K.M., Metcalf, A.K., Mamas, M.A., Potter, J.F., et al. (2016). Impact of Hemoglobin Levels and Anemia on Mortality in Acute Stroke: Analysis of UK Regional Registry Data, Systematic Review, and Meta-Analysis. J Am Heart Assoc 5.
12. Solak, Y., Yilmaz, M.I., Saglam, M., Demirbas, S., Verim, S., Unal, H.U., Gaipov, A., Oguz, Y., Kayrak, M., Caglar, K., et al. (2013). Mean corpuscular volume is associated with endothelial dysfunction and predicts composite cardiovascular events in patients with chronic kidney disease. Nephrology (Carlton) 18, 728-735.
13. Evans, D.M., Frazer, I.H., and Martin, N.G. (1999). Genetic and environmental causes of variation in basal levels of blood cells. Twin research : the official journal of the International Society for Twin Studies 2, 250-257.
14. Wright, F.A., Sullivan, P.F., Brooks, A.I., Zou, F., Sun, W., Xia, K., Madar, V., Jansen, R., Chung, W., Zhou, Y.H., et al. (2014). Heritability and genomics of gene expression in peripheral blood. Nature genetics 46, 430-437.

15. Chambers, J.C., Zhang, W., Li, Y., Sehmi, J., Wass, M.N., Zabaneh, D., Hoggart, C., Bayele, H., McCarthy, M.I., Peltonen, L., et al. (2009). Genome-wide association study identifies variants in TMPRSS6 associated with hemoglobin levels. Nature genetics 41, 1170-1172.
16. Chen, Z., Tang, H., Qayyum, R., Schick, U.M., Nalls, M.A., Handsaker, R., Li, J., Lu, Y., Yanek, L.R., Keating, B., et al. (2013). Genome-wide association analysis of red blood cell traits in African Americans: the COGENT Network. Human molecular genetics 22, 2529-2538.
17. 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. (2009). Multiple loci influence erythrocyte phenotypes in the CHARGE Consortium. Nature genetics 41, 1191-1198.
18. Kamatani, Y., Matsuda, K., Okada, Y., Kubo, M., Hosono, N., Daigo, Y., Nakamura, Y., and Kamatani, N. (2010). Genome-wide association study of hematological and biochemical traits in a Japanese population. Nature genetics 42, 210-215.
19. Kullo, I.J., Ding, K., Jouni, H., Smith, C.Y., and Chute, C.G. (2010). A genome-wide association study of red blood cell traits using the electronic medical record. PloS one 5.
20. Li, J., Glessner, J.T., Zhang, H., Hou, C., Wei, Z., Bradfield, J.P., Mentch, F.D., Guo, Y., Kim, C., Xia, Q., et al. (2013). GWAS of blood cell traits identifies novel associated loci and epistatic interactions in Caucasian and African-American children. Human molecular genetics 22, 1457-1464.
21. Soranzo, N., Spector, T.D., Mangino, M., Kuhnel, B., Rendon, A., Teumer, A., Willenborg, C., Wright, B., Chen, L., Li, M., et al. (2009). A genome-wide meta-analysis identifies 22 loci associated with eight hematological parameters in the HaemGen consortium. Nature genetics 41, 1182-1190.
22. van der Harst, P., Zhang, W., Mateo Leach, I., Rendon, A., Verweij, N., Sehmi, J., Paul, D.S., Elling, U., Allayee, H., Li, X., et al. (2012). Seventy-five genetic loci influencing the human red blood cell. Nature 492, 369-375.
23. Ferreira, M.A., Hottenga, J.J., Warrington, N.M., Medland, S.E., Willemsen, G., Lawrence, R.W., Gordon, S., de Geus, E.J., Henders, A.K., Smit, J.H., et al. (2009). Sequence variants in three loci influence monocyte counts and erythrocyte volume. Am J Hum Genet 85, 745-749.
24. Yang, Q., Kathiresan, S., Lin, J.P., Tofler, G.H., and O'Donnell, C.J. (2007). Genome-wide association and linkage analyses of hemostatic factors and hematological phenotypes in the Framingham Heart Study. BMC Med Genet 8 Suppl 1, S12.
25. Gravel, S., Henn, B.M., Gutenkunst, R.N., Indap, A.R., Marth, G.T., Clark, A.G., Yu, F., Gibbs, R.A., Genomes, P., and Bustamante, C.D. (2011). Demographic history and rare allele sharing among human populations. Proc Natl Acad Sci U S A 108, 11983-11988.
26. Schick, U.M., Jain, D., Hodonsky, C.J., Morrison, J.V., Davis, J.P., Brown, L., Sofer, T., Conomos, M.P., Schurmann, C., McHugh, C.P., et al. (2016). Genome-wide Association Study of Platelet Count Identifies Ancestry-Specific Loci in Hispanic/Latino Americans. Am J Hum Genet 98, 229-242.

27. McCarthy, M.I., and Hirschhorn, J.N. (2008). Genome-wide association studies: potential next steps on a genetic journey. Human molecular genetics 17, R156-165.
28. Voight, B.F., Kang, H.M., Ding, J., Palmer, C.D., Sidore, C., Chines, P.S., Burtt, N.P., Fuchsberger, C., Li, Y., Erdmann, J., et al. (2012). The metabochip, a custom genotyping array for genetic studies of metabolic, cardiovascular, and anthropometric traits. PLoS Genet 8, e1002793.
29. Matise, T.C., Ambite, J.L., Buyske, S., Carlson, C.S., Cole, S.A., Crawford, D.C., Haiman, C.A., Heiss, G., Kooperberg, C., Marchand, L.L., et al. (2011). The Next PAGE in understanding complex traits: design for the analysis of Population Architecture Using Genetics and Epidemiology (PAGE) Study. American journal of epidemiology 174, 849-859.
30. ARIC Investigators. (1989). The Atherosclerosis Risk in Communities (ARIC) Study: design and objectives. The ARIC investigators. American journal of epidemiology 129, 687-702.
31. Friedman, G.D., Cutter, G.R., Donahue, R.P., Hughes, G.H., Hulley, S.B., Jacobs, D.R., Jr., Liu, K., and Savage, P.J. (1988). CARDIA: study design, recruitment, and some characteristics of the examined subjects. Journal of clinical epidemiology 41, 1105-1116.
32. Fried, L.P., Borhani, N.O., Enright, P., Furberg, C.D., Gardin, J.M., Kronmal, R.A., Kuller, L.H., Manolio, T.A., Mittelmark, M.B., Newman, A., et al. (1991). The Cardiovascular Health Study: design and rationale. Ann Epidemiol 1, 263-276.
33. Daviglus, M.L., Talavera, G.A., Aviles-Santa, M.L., Allison, M., Cai, J., Criqui, M.H., Gellman, M., Giachello, A.L., Gouskova, N., Kaplan, R.C., et al. (2012). Prevalence of major cardiovascular risk factors and cardiovascular diseases among Hispanic/Latino individuals of diverse backgrounds in the United States. JAMA 308, 1775-1784.
34. WHI Study Group. (1998). Design of the Women's Health Initiative clinical trial and observational study. The Women's Health Initiative Study Group. Controlled clinical trials 19, 61-109.
35. Li, L., Li, Y., Browning, S.R., Browning, B.L., Slater, A.J., Kong, X., Aponte, J.L., Mooser, V.E., Chissoe, S.L., Whittaker, J.C., et al. (2011). Performance of genotype imputation for rare variants identified in exons and flanking regions of genes. PloS one 6, e24945.
36. Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A., Bender, D., Maller, J., Sklar, P., de Bakker, P.I., Daly, M.J., et al. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet 81, 559-575.
37. (1993). Minisymposium: The Malmo Diet and Cancer Study. Design, biological bank and biomarker programme. 23 October 1991, Malmo, Sweden. J Intern Med 233, 39-79.
38. Lin, D.Y., Tao, R., Kalsbeek, W.D., Zeng, D., Gonzalez, F., 2nd, Fernandez-Rhodes, L., Graff, M., Koch, G.G., North, K.E., and Heiss, G. (2014). Genetic association analysis under complex survey sampling: the Hispanic Community Health Study/Study of Latinos. Am J Hum Genet 95, 675-688.

39. Li, Y., Willer, C.J., Ding, J., Scheet, P., and Abecasis, G.R. (2010). MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. Genetic epidemiology 34, 816-834.
40. Ward, L.D., and Kellis, M. (2012). HaploReg: a resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants. Nucleic Acids Res 40, D930-934.
41. Roadmap Epigenomics, C., Kundaje, A., Meuleman, W., Ernst, J., Bilenky, M., Yen, A., Heravi-Moussavi, A., Kheradpour, P., Zhang, Z., Wang, J., et al. (2015). Integrative analysis of 111 reference human epigenomes. Nature 518, 317-330.
42. Consortium, E.P. (2012). An integrated encyclopedia of DNA elements in the human genome. Nature 489, 57-74.
43. Badis, G., Berger, M.F., Philippakis, A.A., Talukder, S., Gehrke, A.R., Jaeger, S.A., Chan, E.T., Metzler, G., Vedenko, A., Chen, X., et al. (2009). Diversity and complexity in DNA recognition by transcription factors. Science 324, 1720-1723.
44. Berger, M.F., Badis, G., Gehrke, A.R., Talukder, S., Philippakis, A.A., Pena-Castillo, L., Alleyne, T.M., Mnaimneh, S., Botvinnik, O.B., Chan, E.T., et al. (2008). Variation in homeodomain DNA binding revealed by high-resolution analysis of sequence preferences. Cell 133, 1266-1276.
45. Berger, M.F., Philippakis, A.A., Qureshi, A.M., He, F.S., Estep, P.W., 3rd, and Bulyk, M.L. (2006). Compact, universal DNA microarrays to comprehensively determine transcription-factor binding site specificities. Nature biotechnology 24, 1429-1435.
46. Adams, D., Altucci, L., Antonarakis, S.E., Ballesteros, J., Beck, S., Bird, A., Bock, C., Boehm, B., Campo, E., Caricasole, A., et al. (2012). BLUEPRINT to decode the epigenetic signature written in blood. Nature biotechnology 30, 224-226.
47. Consortium, G.T. (2015). Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. Science 348, 648-660.
48. Jansen, R., Hottenga, J.J., Nivard, M.G., Abdellaoui, A., Laport, B., de Geus, E.J., Wright, F.A., Penninx, B., and Boomsma, D.I. (2017). Conditional eQTL analysis reveals allelic heterogeneity of gene expression. Human molecular genetics 26, 1444-1451.
49. Westra, H.J., Peters, M.J., Esko, T., Yaghootkar, H., Schurmann, C., Kettunen, J., Christiansen, M.W., Fairfax, B.P., Schramm, K., Powell, J.E., et al. (2013). Systematic identification of trans eQTLs as putative drivers of known disease associations. Nature genetics 45, 1238-1243.
50. Landrum, M.J., Lee, J.M., Riley, G.R., Jang, W., Rubinstein, W.S., Church, D.M., and Maglott, D.R. (2014). ClinVar: public archive of relationships among sequence variation and human phenotype. Nucleic Acids Res 42, D980-985.
51. Zhou, X., Maricque, B., Xie, M., Li, D., Sundaram, V., Martin, E.A., Koebbe, B.C., Nielsen, C., Hirst, M., Farnham, P., et al. (2011). The Human Epigenome Browser at Washington University. Nat Methods 8, 989-990.
52. Ward, L.D., and Kellis, M. (2016). HaploReg v4: systematic mining of putative causal variants, cell types, regulators and target genes for human complex traits and disease. Nucleic Acids Research 44, D877-D881.

53. Conomos, M.P., Reiner, A.P., Weir, B.S., and Thornton, T.A. (2016). Model-free Estimation of Recent Genetic Relatedness. Am J Hum Genet 98, 127-148.
54. Wojcik, G.L., Graff, M., Nishimura, K.K., Tao, R., Haessler, J., Gignoux, C.R., Highland, H.M., Patel,, Y.M., S., E.P., Avery, C.L., Belbin, G.M., Bien, S.A., Cheng, I., Hodonsky, C.J., Huckins, L.M., Jeffs, J.,, Justice, A.E., Kocarnik, J.M., Lin, B.M., Lu, Y.K., Nelson, S.C., Park, S.L., Preuss, M., Richard, M.A., Schurmann,, C., S., V.W., Vahi, K., Vishnu, A., Verbanck, M., Walker, R., Young, K.L., Zubair, N., Ambite, J.L.,, Boerwinkle, E., Bottinger, E.P., Bustamante, C.D., Caberto, C., Conomos, M.P., Deelman, E., Do,, R., D., K., Fernandez-Rhodes, L., Fornage, M., Heiss, G., Hindorff, L.A., Jackson, R.D., James, R.,, Laurie, C.A., Laurie, C.C., Li, Y., Lin, D.Y., Nadkarni, G., Pankow, J., Pooler, L.C., Reiner, A.P., Romm,, J., S., C., Sheng, X., Stahl, E., Stram, D.O., Thornton, T.A., Wassel, C.L., Wilkens, L.R., Yoneyama, S.,, Buyske, S., Haiman, C., Kooperberg, C., Le Marchand, L., Loos, R.J.F., Matise, T.C., North, K.E., Peters, U.,, and Kenny, E.E., Carlson, C.S. (2017). Genetic Diversity Turns a New PAGE in Our Understanding of Complex Traits. (In review).
55. Manrai, A.K., Funke, B.H., Rehm, H.L., Olesen, M.S., Maron, B.A., Szolovits, P., Margulies, D.M., Loscalzo, J., and Kohane, I.S. (2016). Genetic Misdiagnoses and the Potential for Health Disparities. N Engl J Med 375, 655-665.
56. Kim, J., Bai, Y., and Pan, W. (2015). An Adaptive Association Test for Multiple Phenotypes with GWAS Summary Statistics. Genet Epidemiol 39, 651-663.
57. Pan, W., Kim, J., Zhang, Y., Shen, X., and Wei, P. (2014). A powerful and adaptive association test for rare variants. Genetics 197, 1081-1095.
58. Pasaniuc, B., and Price, A.L. (2016). Dissecting the genetics of complex traits using summary association statistics. Nat Rev Genet.
59. Kichaev, G., Roytman, M., Johnson, R., Eskin, E., Lindstrom, S., Kraft, P., and Pasaniuc, B. (2016). Improved methods for multi-trait fine mapping of pleiotropic risk loci. Bioinformatics.
60. Chakravarti, A., and Turner, T.N. (2016). Revealing rate-limiting steps in complex disease biology: The crucial importance of studying rare, extreme-phenotype families. Bioessays 38, 578-586.
61. Boyle, E.A., Li, Y.I., and Pritchard, J.K. (2017). An Expanded View of Complex Traits: From Polygenic to Omnigenic. Cell 169, 1177-1186.
62. Rawofi, L., Edwards, M., Krithika, S., Le, P., Cha, D., Yang, Z., Ma, Y., Wang, J., Su, B., Jin, L., et al. (2017). Genome-wide association study of pigmentary traits (skin and iris color) in individuals of East Asian ancestry. PeerJ 5, e3951.
63. Wei, P., Cao, Y., Zhang, Y., Xu, Z., Kwak, I.Y., Boerwinkle, E., and Pan, W. (2016). On Robust Association Testing for Quantitative Traits and Rare Variants. G3 (Bethesda) 6, 3941-3950.
64. 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. (2016). The Allelic Landscape of Human Blood Cell Trait Variation and Links to Common Complex Disease. Cell 167, 1415-1429 e1419.
65. Kato, N., Loh, M., Takeuchi, F., Verweij, N., Wang, X., Zhang, W., Kelly, T.N., Saleheen, D., Lehne, B., Mateo Leach, I., et al. (2015). Trans-ancestry genome-

wide association study identifies 12 genetic loci influencing blood pressure and implicates a role for DNA methylation. Nature genetics 47, 1282-1293.
66. van Rooij, F.J., Qayyum, R., Smith, A.V., Zhou, Y., Trompet, S., Tanaka, T., Keller, M.F., Chang, L.C., Schmidt, H., Yang, M.L., et al. (2017). Genome-wide Trans-ethnic Meta-analysis Identifies Seven Genetic Loci Influencing Erythrocyte Traits and a Role for RBPMS in Erythropoiesis. Am J Hum Genet 100, 51-63.
67. Kato, N., Takeuchi, F., Tabara, Y., Kelly, T.N., Go, M.J., Sim, X., Tay, W.T., Chen, C.H., Zhang, Y., Yamamoto, K., et al. (2011). Meta-analysis of genome-wide association studies identifies common variants associated with blood pressure variation in east Asians. Nature genetics 43, 531-538.
68. Ligthart, S., Vaez, A., Hsu, Y.H., Inflammation Working Group of the, C.C., Pmi Wg, X.C.P., LifeLines Cohort, S., Stolk, R., Uitterlinden, A.G., Hofman, A., Alizadeh, B.Z., et al. (2016). Bivariate genome-wide association study identifies novel pleiotropic loci for lipids and inflammation. BMC Genomics 17, 443.
69. Cmejla, R., Cmejlova, J., Handrkova, H., Petrak, J., Petrtylova, K., Mihal, V., Stary, J., Cerna, Z., Jabali, Y., and Pospisilova, D. (2009). Identification of mutations in the ribosomal protein L5 (RPL5) and ribosomal protein L11 (RPL11) genes in Czech patients with Diamond-Blackfan anemia. Hum Mutat 30, 321-327.
70. Konno, Y., Toki, T., Tandai, S., Xu, G., Wang, R., Terui, K., Ohga, S., Hara, T., Hama, A., Kojima, S., et al. (2010). Mutations in the ribosomal protein genes in Japanese patients with Diamond-Blackfan anemia. Haematologica 95, 1293-1299.
71. Bauer, D.E., and Orkin, S.H. (2015). Hemoglobin switching's surprise: the versatile transcription factor BCL11A is a master repressor of fetal hemoglobin. Curr Opin Genet Dev 33, 62-70.
72. An, X., and Mohandas, N. (2008). Disorders of red cell membrane. Br J Haematol 141, 367-375.
73. Mankelow, T.J., Satchwell, T.J., and Burton, N.M. (2012). Refined views of multi-protein complexes in the erythrocyte membrane. Blood Cells Mol Dis 49, 1-10.
74. Hodonsky, C.J., Jain, D., Schick, U.M., Morrison, J.V., Brown, L., McHugh, C.P., Schurmann, C., Chen, D.D., Liu, Y.M., Auer, P.L., et al. (2017). Genome-wide association study of red blood cell traits in Hispanics/Latinos: The Hispanic Community Health Study/Study of Latinos. PLoS Genet 13, e1006760.