Human Molecular Genetics, 2013, Vol. 22, No. 12 **2529–2538**
doi:10.1093/hmg/ddt087
Advance Access published on February 26, 2013

Genome-wide association analysis of red blood cell traits in African Americans: the COGENT Network

Zhao Chen[1], Hua Tang[3], Rehan Qayyum[4], Ursula M. Schick[5], Michael A. Nalls[6], Robert Handsaker[7], Jin Li[8], Yingchang Lu[10], Lisa R. Yanek[15], Brendan Keating[16], Yan Meng[18], Frank J.A. van Rooij[19], Yukinori Okada[20,21,22], Michiaki Kubo[23], Laura Rasmussen-Torvik[24], Margaux F. Keller[25], Leslie Lange[26], Michele Evans[27], Erwin P. Bottinger[11], Michael D. Linderman[12], Douglas M. Ruderfer[13], Hakon Hakonarson[8,9,17], George Papanicolaou[28], Alan B. Zonderman[29], Omri Gottesman[11], BioBank Japan Project, CHARGE Consortium, Cynthia Thomson[2], Elad Ziv[30], Andrew B. Singleton[25], Ruth J.F. Loos[14], Patrick M.A. Sleiman[8,9,17], Santhi Ganesh[31], Steven McCarroll[32,33], Diane M. Becker[4], James G. Wilson[34], Guillaume Lettre[35] and Alexander P. Reiner[5,36,*]

[1]Division of Epidemiology and Biostatistics, Mel and Enid Zuckerman College of Public Health and [2]Division of Nutrition, Mel and Enid Zuckerman College of Public Health, University of Arizona, Tucson, AZ 85724, USA, [3]Department of Statistics and Department of Genetics, Stanford University, Stanford, CA 94305, USA, [4]GeneSTAR Research Program, Division of General Internal Medicine, Johns Hopkins School of Medicine, Baltimore, MD 21287, USA, [5]Division of Public Health Sciences, Fred Hutchinson Cancer Research Center, Seattle, WA 98195, USA, [6]Molecular Genetics Section, Laboratory of Neurogenetics, National Institute on Aging, Bethesda, MD 20892, USA, [7]Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, MA 02141, USA, [8]Center for Applied Genomics, Abramson Research Center and [9]Division of Human Genetics, The Children's Hospital of Philadelphia, Philadelphia, PA 19104, USA, [10]The Charles Bronfman Institute for Personalized Medicine, The Genetics of Obesity and Related Metabolic Traits Program, [11]The Charles Bronfman Institute for Personalized Medicine, [12]Department of Genetics and Genomic Sciences, Institute of Genomics and Multiscale Biology, [13]Division of Psychiatric Genomics, Department of Psychiatry and [14]The Charles Bronfman Institute for Personalized Medicine, Institute of Child Health and Development, The Genetics of Obesity and Related Metabolic Traits Program, Mount Sinai School of Medicine, New York, NY 10029, USA, [15]Department of Medicine, The Johns Hopkins University School of Medicine, Baltimore, MD, USA, [16]Department of Medicine and [17]Department of Pediatrics, University of Pennsylvania School of Medicine, Philadelphia, PA 19104, USA, [18]Program in Medical and Population Genetics, Broad Institute, Cambridge, MA, USA, [19]Department of Epidemiology, Erasmus MC, University Medical Center Rotterdam, Rotterdam, The Netherlands, [20]Division of Rheumatology, Immunology, and Allergy and [21]Division of Genetics, Brigham and Women's Hospital, Harvard Medical School, Boston, MA 02115, USA, [22]Medical and Population Genetics Program, Broad Institute, Cambridge, MA 02142, USA, [23]Laboratory for Genotyping Development, CGM, RIKEN, Yokohama, Japan, [24]Department of Preventive Medicine, Northwestern University Feinberg School of Medicine, Chicago, IL, USA, [25]Laboratory of Neurogenetics and [26]Department of Genetics, University of North Carolina, Chapel Hill, NC 27599, USA, [27]Health Disparities Research Section, Clinical Research Branch, National Institute on Aging, National Institutes of Health, Baltimore, MD 21225, USA, [28]Division of Cardiovascular Sciences, National Heart, Lung, and Blood Institute (NHLBI), Bethesda, MD, USA, [29]Laboratory of Personality and Cognition, National Institute on Aging, National Institutes of Health, Baltimore, MD 21224, USA, [30]Department of Medicine, University of California, San Francisco, CA 94143, USA, [31]Division of Cardiology, University of Michigan Health System, Ann Arbor, MI 48109, USA, [32]Department of Genetics, Harvard Medical School, Cambridge, MA, USA,

*To whom correspondence should be addressed at: 1100 N Fairview Ave N M3-A410, Seattle, WA, USA. Tel: +1 206 667 2710; Fax: +1 206 667 4142; Email: apreiner@u.washington.edu

[33]Broad Institute of MIT and Harvard, Cambridge, MA, USA, [34]Department of Department of Physiology and Biophysics, University of Mississippi Medical Center, Jackson, MS 39216, USA, [35]Montreal Heart Institute, Montréal, Québec, Canada H1T 1C8, [36]Department of Epidemiology, University of Washington, Seattle, WA 98195, USA

Received December 4, 2012; Revised February 9, 2013; Accepted February 18, 2013

Laboratory red blood cell (RBC) measurements are clinically important, heritable and differ among ethnic groups. To identify genetic variants that contribute to RBC phenotypes in African Americans (AAs), we conducted a genome-wide association study in up to ~16 500 AAs. The alpha-globin locus on chromosome 16pter [lead SNP rs13335629 in *ITFG3* gene; $P < 1E-13$ for hemoglobin (Hgb), RBC count, mean corpuscular volume (MCV), MCH and MCHC] and the G6PD locus on Xq28 [lead SNP rs1050828; $P < 1E-13$ for Hgb, hematocrit (Hct), MCV, RBC count and red cell distribution width (RDW)] were each associated with multiple RBC traits. At the alpha-globin region, both the common African 3.7 kb deletion and common single nucleotide polymorphisms (SNPs) appear to contribute independently to RBC phenotypes among AAs. In the 2p21 region, we identified a novel variant of *PRKCE* distinctly associated with Hct in AAs. In a genome-wide admixture mapping scan, local European ancestry at the 6p22 region containing *HFE* and *LRRC16A* was associated with higher Hgb. *LRRC16A* has been previously associated with the platelet count and mean platelet volume in AAs, but not with Hgb. Finally, we extended to AAs the findings of association of erythrocyte traits with several loci previously reported in Europeans and/or Asians, including *CD164* and *HBS1L-MYB*. In summary, this large-scale genome-wide analysis in AAs has extended the importance of several RBC-associated genetic loci to AAs and identified allelic heterogeneity and pleiotropy at several previously known genetic loci associated with blood cell traits in AAs.

INTRODUCTION

Laboratory red blood cell (RBC) measurements are important for the diagnosis and classification of various hematologic disorders. Some disorders of RBCs, such as sickle cell anemia and alpha thalassemia, are single-gene diseases with higher frequency among populations of African descent (1,2). Even among healthy individuals, African Americans (AAs) have lower hemoglobin (Hgb), hematocrit (Hct) and mean corpuscular volume (MCV) compared with other racial/ethnic groups across all ages (3–5).

Heritability studies suggest that RBC traits are under significant genetic influence. Genome-wide association studies (GWASs) of RBC indices have been reported among European and Japanese populations (6–8), but to our knowledge have not yet been reported for AA. In a gene-centric association study from the CARe consortium, the common African glucose-6-phosphate dehydrogenase (*G6PD*) A-variant on chromosome X and another variant of the α-globin (*HBA2-HBA1*) locus were associated with multiple RBC traits in AAs (9).

The genetic loci reported to date explain only a small fraction of heritability in RBC traits, highlighting the need for larger studies that include ethnic minorities and complementary analytic approaches (10). Thus, we performed a GWA meta-analysis of RBC traits among AA participants from cohorts of the Continental Origins and Genetic Epidemiology Network (COGENT). As AAs are an admixed population, the resulting genomic architecture can be leveraged to identify regions where either African or European ancestral alleles are associated with traits such as Hgb which differ significantly between European and African populations. Therefore, we performed admixture mapping for the association between available RBC traits (Hgb, Hct, MCHC) and local ancestry.

RESULTS

Descriptive analysis

Since not all RBC traits were available in every COGENT cohort, the numbers of individuals available for meta-analysis varied by each RBC trait (Supplementary Material, Table S1). Only Hgb ($n = 16\,485$) and Hct ($n = 16\,496$) were available in all cohorts. MCHC ($n = 12\,152$), MCV ($n = 6438$), RBC count (4818), MCH ($n = 4066$) and RDW ($n = 3811$) were available in subsets of participating cohorts. There were varying degrees of pairwise correlation between RBC traits (Supplementary Material, Table S2). Pearson's correlation coefficients were highest (>0.95) between Hgb and Hct, and between MCV and MCH and were lowest between the RDW and RBC count (0.03).

GWAS of RBC traits in COGENT AAs

The GWA results for each RBC trait are summarized by Manhattan (Fig. 1) and quantile–quantile (Supplementary Material, Fig. S1) plots. The meta-analysis inflation factors were all near unity (0.998–1.005), suggesting that confounders and other technical artifacts were well-controlled. In total, seven independent genomic loci met the experiment-wide significance threshold ($P < 1 \times 10^{-8}$) for one or more RBC traits (Table 1 and Supplementary Material, Table S3). Three loci (1p31.1, 13q31.2, 16p13.3centromeric) have not been previously associated with RBC traits, whereas four loci (2p21,

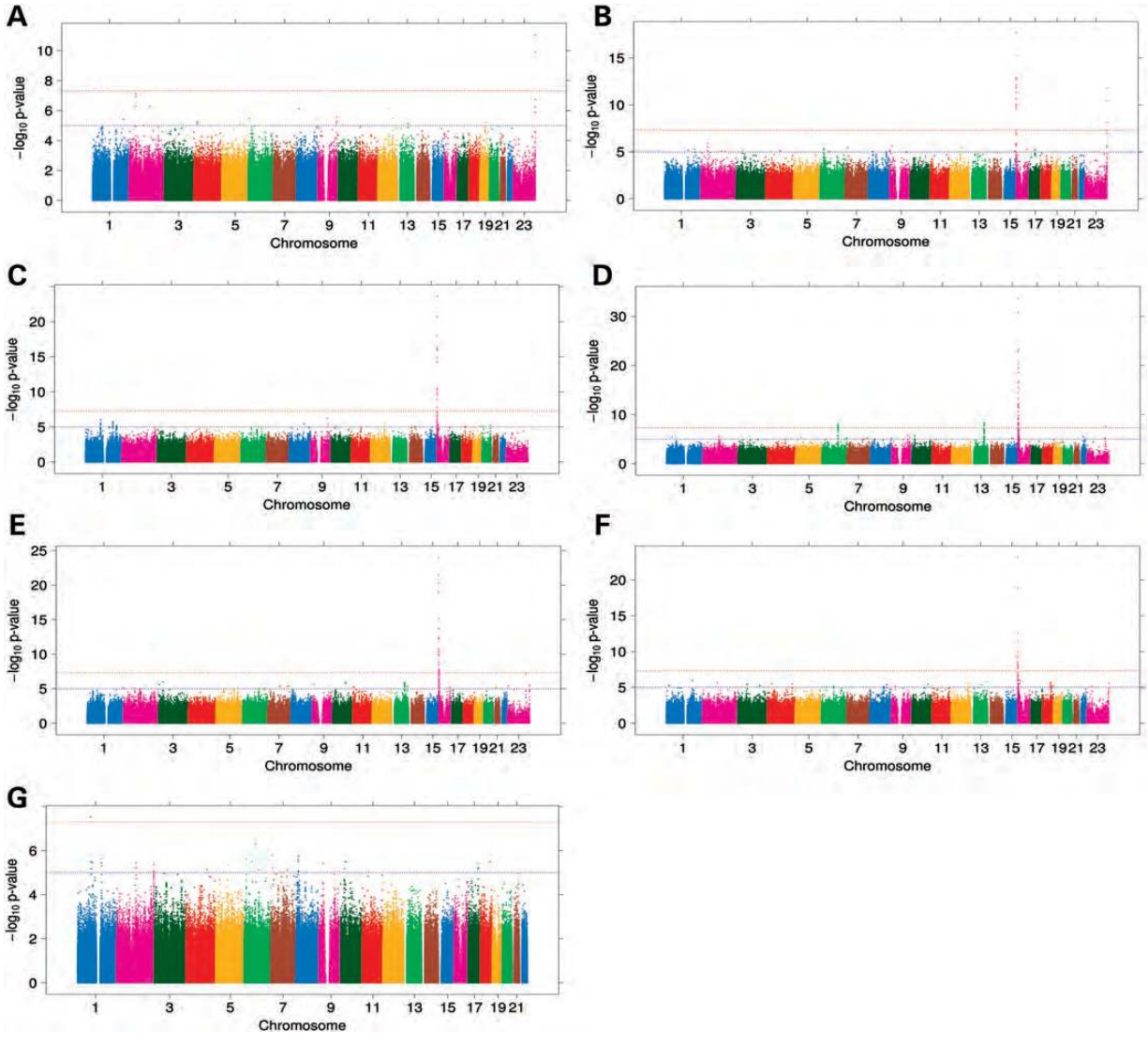


Figure 1. Manhattan plots of GWAS analysis for RBC traits (**A**) Hct; (**B**) Hgb; (**C**) MCHC; (**D**) MCH; (**E**) MCV; (**F**) RBC count and (**G**) RDW. The dashed horizontal red line indicates $P = 1 \times 10^{-8}$. The dashed horizontal blue line indicates $P = 5 \times 10^{-3}$.

6q21, 16p31.3telomeric, and Xq28) have been associated with at least one such trait in populations of European, Japanese or African descent.

Previously reported RBC loci

The two top Xq28 single nucleotide polymorphisms (SNPs) (rs762516, rs1050828) for Hgb, Hct, MCV and RBC count are located in the *G6PD* gene. rs1050828 encodes the G6PD amino acid substitution Val68Met that results in the *G6PD* A^- allele known to cause *G6PD* deficiency (MIM #305900). The G6PD A-variant has been previously associated with lower Hct, Hgb and RBC count, and with higher MCV in AAs (9). Here, we additionally report that the G6PD A^- allele is associated with the lower RDW. Given the extent of the association signal at Xq28, we repeated the Hgb and Hct association analyses in women from Women Health Initiative (WHI), the largest AA cohort ($n = 8304$) for Hgb and conditioning the lead SNPs, rs1050828. After adjusting for rs1050828, the strength of association with Hgb for the remaining SNPs on Xq28 was greatly attenuated and no longer significant (data not shown).

The index SNP on 16p13, which encompasses the α-globin (*HBA2-HBA1*) locus, was rs13335629 within an intron of *ITFG3*. rs13335629 met the genome-wide significance threshold for association with lower Hgb, MCH, MCHC and MCV and also with a higher RBC count. The rs13335629 variant was also nominally associated with lower Hct ($\beta = -0.215 \pm 0.056$; $P = 1.33\text{E}-04$) and higher RDW ($\beta = 0.0053 \pm 0.0021$; $P = 0.01$). Lo *et al*. previously reported a common rs1211375 variant within the 16p13 region associated with lower Hgb, MCH and MCV in AAs, and that these associations were not present in Caucasians (9). Our index SNP rs13335629 is in moderate linkage disequilibrium (LD) with rs1211375 ($r^2 = 0.33$ in HapMap YRI).

Three intronic variants of the protein kinase C (PKC)-epsilon gene *PRKCE* on 2p21 were associated with lower

Table 1. Results of genome-wide significant SNPs for RBC traits in COGENT AA

Trait	Chromo-some	Number of SNPs with $P < 5 \times 10^{-8}$	Top SNP in the region	Position Hg18	Candidate genes	Minor/ major allele	MAF	Effect size (SE)	*P*-value
HCT	Xq28	6	rs762516	153 417 857	*G6PD, TKTL1, MECP2, MPP1*	T/C	0.148	−0.452 (0.055)	2.17E − 16
HCT	2p21	3	rs13008603	46 209 352	*PRKCE*	A/C	0.163	−0.277 (0.047)	4.09E − 09
HGB	16p13.3	11	rs13335629	250 381	*ITFG3, LUCL7, NPRL3, POLR3K, RPL2B, MPG*	A/G	0.120	−0.190 (0.019)	2.63E − 23
HGB	Xq28	9	rs762516	153 417 857	*G6PD, TKTL1, MECP2, MPP1*	T/C	0.146	−0.1614 (0.0186)	3.73E − 18
MCHC	16p13.3	38	rs13335629	250 381	*ITFG3, LUCL7, NPRL3, POLR3K, RPL2B, MPG, NME4, DECR*	A/G	0.117	−0.3298 (0.0227)	8.66E − 48
MCH	16p13.3	28	rs13339636	238 589	*ITFG3, LUCL7, NPRL3, POLR3K, RPL2B, MPG, NME4, DECR*	G/A	0.132	−0.6847 (0.0559)	1.87E − 34
MCH	13q31.2	4	rs9559892[a]	88 166 665	–	A/C	0.253	−0.2589 (0.0444)	5.46E − 09
MCH	16p13.3	1	rs7192051[a]	4 482 118	*HMOX2*	G/T	0.360	−0.2396 (0.0411)	5.70E − 09
MCH	6q21	10	rs9386791	109 715 190	*CD164*	C/T	0.416	−0.2294 (0.0401)	1.20E − 08
MCV	16p13.3	21	rs13335629	250 381	*ITFG3, LUCL7, NPRL3, POLR3K, RPL2B, MPG, NME4, DECR, RHOT2, LMF1, WDR90*	A/G	0.134	−0.648 (0.0669)	3.61E − 22
MCV	Xq28	6	rs762516	153 417 857	*G6PD, FAM3A, F8, MPP1*	T/C	0.137	1.5768 (0.2083)	3.76E − 14
MCV	16p13.3	1	rs7192051[a]	4 482 118	*HMOX2*	G/T	0.363	−0.259 (0.0475)	4.83E − 08
RBC	16p13.3	12	rs13335629	250 381	*ITFG3, LUCL7*	A/G	0.120	0.1699 (0.0169)	7.48E − 24
RBC	Xq28	9	rs1050828	153 417 411	*G6PD, F8, MPP1, MECP2, CTAG2*	T/C	0.108	−0.1424 (0.0159)	4.00E − 19
RDW	Xq28	1	rs1050828	153 417 411	*G6PD*	T/C	0.116	−0.0326 (0.0048)	1.70E − 11
RDW	1p31.1	1	rs10493739[a]	83 698 745	–	T/C	0.334	0.0128 (0.0023)	3.02E − 08

[a]Novel loci.

Hct. The index SNP rs13008603 was also nominally associated with a lower RBC count ($\beta = -0.044 \pm 0.013$; $P = 4.69\text{E}{-}04$), but not with other RBC traits ($P > 0.05$ for Hgb, MCV, MCH, MCHC, RDW). The three Hct-associated *PRKCE* variants are in strong LD (pairwise $r^2 > 0.7$). Another intronic variant of *PRKCE* (rs10495928) was previously associated with Hgb and Hct in Europeans (8) and with the RBC count in Japanese (6), but showed no evidence of association in AAs ($P = 0.50$ and 0.71 for Hct and Hgb, respectively). In European and African HapMap populations, there is no evidence of LD between rs10495928 and any of the three Hct-associated variants observed in COGENT AA. These results strongly suggest ethnicity-specific allelic heterogeneity for RBC traits at the *PRKCE* locus.

At 6q21, a haplotype comprised of 10 SNPs (lead SNP = rs9386791) was associated with a lower MCH, and nominally with a lower MCV ($P = 1.09\text{E}{-}05$), Hgb ($P = 0.007$), Hct ($P = 0.03$), MCHC ($P = 0.02$), RBC count ($P = 0.01$). These variants are located ∼50 kb upstream of *CD164*, which encodes a mucin-like molecule expressed by human CD34(+) hematopoietic progenitor cells that regulate erythropoiesis. Other variants of the *CD164* 5′ flanking region have been associated with RBC, MCH and MCV in Japanese (rs11966072) (6) and with MCV in Europeans (rs9374080) (8). In HapMap CEU, rs9374080 is in LD with our AA index SNP rs9386791 ($r^2 = 0.87$).

Newly discovered RBC loci

Of the three novel loci associated with RBC traits, rs10493739 at 1p31.1 (associated with RDW) and rs9559892 at 13q31.2 (associated with MCH) are both located in regions devoid of known genes. *TTLL7* is the closest gene to rs10493739 (400 kb away) and encodes a tubulin polyglutamylase, which modifies beta-tubulin (11). There are no known genes within 500 kb on either side of rs9559892. The lead SNP at the third locus, rs7192051 is located within the second intron of the heme oxygenase-2 gene (*HMOX2*) and was associated with lower MCH and MCV. Heme oxygenase 2, the protein product of HMOX2, degrades heme and is important in erythropoiesis (12). Although *HMOX2* is located ∼4 Mb centromic to the alpha-globin locus, it is not in LD with the previously identified 16p13 association signals (maximum $r^2 = 0.004$ with rs13335629).

We attempted to validate two of our three novel RBC loci discovered in COGENT in two independent population-based samples: ∼7700 AA youths ages 8–21 years from CHOP and 2010 AA adults from the Mount Sinai eMERGE study. There was no evidence of replication of rs9559892 with MCH, nor of rs7192051 with MCV or MCH (Supplementary Material, Table S4) in the validation sample. It was not possible to pursue replication of rs10493739 in CHOP and eMERGE because this SNP was not genotyped and it could not be imputed in the available replication samples. In over 20 000

Europeans from the CHARGE consortium and 14 000 Japanese from RIKEN, there was no evidence of association of rs9559892 with MCH. Similarly, there was no evidence of association of rs7192051 with MCV or MCH in CHARGE Europeans (Supplementary Material, Table S4).

Admixture mapping analysis of Hgb, Hct and MCHC traits in WHI AA

As a complementary approach to identifying variants associated with RBC traits in AAs that occur at disparate frequencies in ancestral African versus European populations, we performed admixture mapping for Hgb, Hct and MCHC in WHI, the largest cohort comprising COGENT. For MCHC, there was one genome-wide significant association signal at the p-term of 16 containing the alpha-globin locus (Supplementary Material, Fig. S2). Local African ancestry in this region was associated with lower MCHC. The admixture association peak is at rs7203694 ($P = 2.78e-06$) located within *RAB40C*, and the genome-wide significant region spans 0–0.78 mb (build 36). There were no genome-wide significant admixture associations for Hct (data not shown). For Hgb, a 2 mb region on chromosome 6p22.2–6p22.1 (25.2–27.1 mb, build 36) reached genome-wide significance, with increased European ancestry associated with higher Hgb levels (Fig. 2A). The Hgb admixture signal appears to be comprised of two peaks (Fig. 2B). Underlying the centromeric peak is *HFE*, the hemochromatosis protein-coding gene, which regulates iron absorption by modulating the interaction of the transferrin receptor with transferrin. Two known *HFE* mutations C282Y (rs1800562) and H63D (rs1799945) cause hereditary hemochromatosis, an autosomal recessive iron storage disorder (13). Among individuals of European descent, the frequencies of C282Y and H63D are 3.23 and 16.6 %, respectively. Both the mutations are essentially absent in the HapMap YRI populations, and therefore the frequency in AAs is low and is the result of European admixture. C282Y was directly typed in WHI SHARe and other COGENT cohorts (total $N = 15\,584$); the genotype association test yielded an Hgb association $P = 0.0003$ in WHI alone and 4.3×10^{-6} in COGENT overall (minor allele frequency = 0.015; $\beta = 0.239 \pm 0.052$). H63D was not directly typed; however, it is tagged by rs129128 ($r^2 = 1.0$ in CEU), which was associated with Hgb levels in WHI ($P = 0.008$) but not when all COGENT cohorts were analyzed together ($P = 0.07$).

After adjusting for genotypes at C282Y and the H63D proxy rs129128 in WHI, the admixture *P* value for chromosome 6p22 was attenuated, but remained significant (from 3.28×10^{-7} to 1.12×10^{-4}), suggesting the existence of additional variants in this region that contribute to inter-population differences in Hgb levels. Located within the telomeric peak of the admixture signal (Fig. 2B) are a number of additional variants that have Fst > 0.3, including several near *LRRC16A*, which has been associated with both serum transferrin levels in whites and the platelet count in AAs from COGENT (14). The most strongly associated *LRRC16A* variant rs9356970 is located ~25 kb upstream of the 5′ flanking region (MAF = 0.09; $\beta = 0.118 \pm 0.028$; $P = 2.7 \times 10^{-5}$). According to the HapMap, the minor allele is present in 30% of European chromosomes, but only 2.5% of YRI

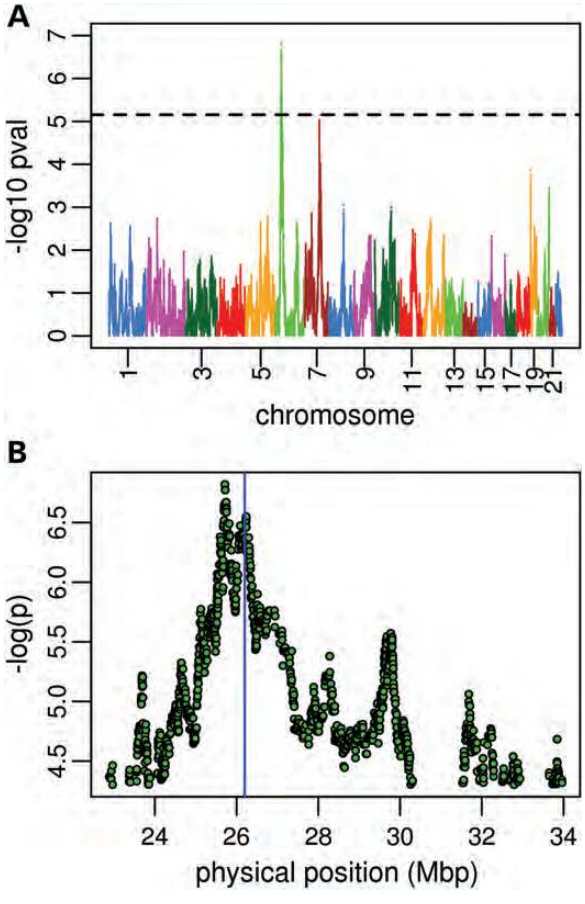


Figure 2. Admixture scan of Hgb concentration. (**A**) shows a genome-wide plot of $-\log(P\text{-values})$ for local-ancestry association with Hgb. The dashed horizontal line indicates the experiment-wide admixture scan significance threshold of $P < 7 \times 10^{-6}$. (**B**) indicates a zoom-in of the genome-wide significant region on chromosome 6, where there appears to be a broad, bimodal admixture peak. The region corresponding to the *HFE* gene is shown in blue.

chromosomes. In a regression model simultaneously adjusting for *LRRC16A* rs9356970 in addition to *HFE* C282Y and rs129128, the association signal for local African ancestry at 6p22 was further attenuated, but remained nominally associated with lower Hgb ($\beta = -0.058 \pm 0.029$; $P = 0.045$). Together, these results suggest that several European-derived alleles in the 6p22 region, including those of *HFE* and *LRRC16A*, may contribute to higher Hgb levels observed in populations of European descent compared with AAs.

CNV analysis and assessment of allelic heterogeneity at 16p13 alpha-globin region

Given the genetic complexity of the alpha-globin locus on chromosome 16p13, including the presence of a common 3.7 kb alpha-thalassemia deletion in AAs (15), and the extent and magnitude of the observed GWAS signal for RBC traits at 16p13, we assessed structural variation at the 16p13 alpha-globin locus using data from 1000 genomes. First, we confirmed the presence of a common deletion ($-\alpha^{3.7}$) among African Americans (AAs) and West Africans that removes one alpha-globin gene copy (*HBA2*)

(Supplementary Material, Fig. S3). Using pooled sequence data from 16 samples (9 YRI, 5 LWK, 1 ASW, 1 CLM) that appear to be homozygous for $-\alpha^{3.7}$ deletion, we further localized the breakpoints, which appear to be bounded by ~300 bp of nearly identical sequence located within the 5′ flanking regions of *HBA1* and *HBA2* (Supplementary Material, Fig. S4). Second, we identified a rare deletion spanning *HBM* through *HBQ1*, in three Han Chinese individuals, and several other possible (but uncertain) rare copy number variations (CNVs) including one duplication (Supplementary Material, Fig. S5) and a very rare deletion that deletes a known regulatory element MCS-R1 (16) (Supplementary Material, Fig. S6).

Among all typed and imputed SNPs in the WHI dataset, the strongest correlation with alpha37 in 63 YRI samples from 1000 Genomes was observed in the region of *ITFG3*. This includes rs13335629 (r-squared = 0.6), which is also the top Hgb- and MCHC-associated SNP in COGENT AA. We repeated the association analyses in WHI ($n = 8304$) for Hgb and MCHC conditioning on rs13335629. The top SNP for Hgb in the conditional analysis was *POLR3K* rs798693 ($P = 8.7\text{E}-06$). For MCHC, rs2541612 in *NPRL3* remained genome-wide significantly associated with lower MCHC ($P = 1.14\text{E}-09$). When both *ITFG3* rs13335629 and *NPRL3* rs2541612 were included as covariates in the conditional analysis, the SNP most strongly associated with lower MCHC was *LUC7L* rs1211375 ($P = 1.50\text{E}-06$). Taken together, the 1000 Genomes CNV analyses and the results of conditional regression analyses for Hgb and MCHC suggest that while some of the red cell GWAS association signal may be due to the common African alpha37 deletion, there appears to be independent signals coming from other structural variants and/or SNP(s) in the region.

Cross-ethnic transferability of previously reported RBC to COGENT AA

We assessed whether 72 SNPs previously associated with RBC traits in European or Japanese populations are associated with RBC traits in COGENT AA (Supplementary Material, Table S5). Using the conservative Bonferroni multiple comparison corrected significance threshold ($P < 0.0001$), we validated four associations. In addition to the association of *HFE* rs1800562 with Hgb, these include *ITFG3* rs1122794 (previously associated with MCH in Europeans) with higher MCHC ($P = 1.5 \times 10^{-8}$), MCH ($P = 7.2 \times 10^{-6}$) and MCV ($P = 7.1 \times 10^{-5}$) in AAs; *ITFG3* rs7189020 (previously associated with MCV in Europeans) with higher MCH ($P = 1.0 \times 10^{-5}$) and MCV ($P = 1.5 \times 10^{-5}$) in AAs; and *HBS1L-MYB* rs7775698 (previously associated with HCT, MCH, MCHC, MCV and RBC count in Japanese) with a lower RBC count ($P = 3.3 \times 10^{-5}$) in AAs.

RBC-associated genetic variants and anemia in AA women

To identify genetic variants associated with anemia, defined dichotomously as Hgb < 12 g/dl, we performed a GWA scan in 8304 AA women 50–79 years old from WHI. Two loci, Xq28 and 16p13, met the threshold of genome-wide significance. The G6PD rs1050828 A-variant was associated with a 1.49-fold (95% CI: 1.33–1.67) increased risk of anemia ($P = 3.3 \times 10^{-12}$). The index SNP at 16p13 (rs1088638) is located ~20 kb 3′ to *POLR3K*, and was associated with a 1.42-fold (95% CI: 1.26–1.60) increased risk of anemia ($P = 1.2 \times 10^{-8}$). We also constructed a composite RBC genetic risk score (GRS) by summing genotyped or imputed allele dosage at the 15 SNPs associated with at least one RBC trait in AA through the GWA scan, admixture mapping scan, conditional analyses or cross-ethnic transferability analyses described above. The GRS ranged from 5 to 19, with a median of 12. When modeled as a quantitative trait, the GRS was strongly associated with anemia ($P = 3.5 \times 10^{-18}$), explaining 1.4% of the anemia phenotypic variance, or 2.2% of the variance in Hgb concentration. When WHI participants were grouped into four GRS categories, those in the highest GRS category had a 1.95-fold increased risk of anemia (95% CI: 1.56–2.42) compared with those in the lowest GRS category ($P = 3.6 \times 10^{-9}$).

DISCUSSION

In this first reported GWAS meta-analysis of RBC traits in AAs, we report genome-wide associations for four loci (*G6PD* on Xq28, alpha-globin locus on 16pter, *PRKCE* on 2p21 and *CD164* on 6q21). We also validated the association in AAs of variants in genes such as *HFE* and *HBS1L-MYB*, which have previously been associated with RBC traits in other ethnicities. At the alpha-globin locus, there appears to be allelic heterogeneity (particularly for MCHC), with both copy number variants and SNPs having apparent independent effects. At *PRKCE*, the variants associated with lower Hct in our AA sample appear to be distinct from another set of *PRKCE* variants that have been associated with Hgb, Hct and RBC count in Europeans and Japanese.

Hemizygous males and in some instances female carriers of the X-linked G6PD A- allele are predisposed to acute episodes of drug- or infection-induced hemolytic anemia. Under basal conditions, however, the G6PD A-allele is not generally thought to be associated with RBC abnormalities, and hemizygous G6PD A-individuals have been reported to have normal baseline red cell survival in the absence of oxidant stress (17). Nonetheless, the association of low RDW with G6PD deficiency may be due to low grade hemolysis resulting in an increase in the MCV with rightward shift of the overall distribution of RBC volume without change in the shape of the distribution (18). The G6PD A-variant is in LD with other nearby genetic variants that plausibly could influence Hgb or RBC morphology. *TKTL1* encodes a transketolase enzyme that links the pentose phosphate pathway with anaerobic glycolysis, which constitutes the two major metabolic pathways for glucose utilization in human erythrocytes. *MPP1* encodes the red cell membrane protein p55, a scaffolding protein that anchors the actin cytoskeleton to the plasma membrane by forming a ternary complex with protein 4.1R and glycophorin C (19).

Aside from genes involved in Hgb synthesis or metabolism, other genetic loci such as *CD164* and *PRKCE* may be associated with RBC traits through effects on erythropoiesis. CD164 (endolyn) is an adhesive receptor present on early hematopoietic progenitors and maturing erythroid cells that

regulates the adhesion of CD34+ cells to bone marrow stroma and affects migration and proliferation of hematopoietic stem cells and progenitor cells (20,21). The upstream region harboring the RBC trait-associated variants contains an erythroleukemia cell line (K562)-specific cluster of histone modifications and ENCODE transcription factor ChIP-seq binding sites including those for GATA-2 and c-Jun. *PRKCE* encodes an isoform of PKC, PKC epsilon, which is expressed in hematopoietic progenitor cells in a lineage- and stage-specific manner and appears to influence erythroid and megakaryocytic progenitor proliferation and differentiation by modulating the response of hematopoietic precursors to a tumor necrosis factor-related apoptosis-inducing ligand (22–24).

Though the finding was not validated in independent AA samples, one of our novel genome-wide significant associations in our discovery cohorts was the association of MCH and MCV with *HMOX2*, which encodes heme oxygenase-2, a constitutively expressed enzyme with a major role in heme catabolism. Heme induces expression of globin genes in erythrocyte progenitor cells and thus plays an important role in erythropoiesis (12,25,26). The lead SNP in this region, rs7192051, is within 5 kb of predicted *HMOX2* regulatory elements such as transcription factor binding sites, DNase sites and histone modification sites (27). Therefore, further study of this variant in larger, independent samples of AA may be warranted.

Our findings have potential clinical implications. Although previous studies have explored the role of common genetic variation in the regulation of these RBC phenotypes in populations of European and Asian descent (6,7,10), no systematic genetic association studies of these traits have been reported in African-ancestry populations. This is particularly important, as there are marked differences in these RBC indices among ethnic groups, and anemia is more prevalent in populations of African descent (28). While it appears that some of the phenotypic variations for RBC and other hematologic traits are controlled by genetic variation shared across ethnic groups (29), other RBC loci are relatively unique to Africans. Rare variants, which are not well captured by GWASs, and undetected common variants of more modest effect may account for additional genetic variance. Discovery and validation of these and additional genetic variants associated with RBC traits in other ethnic populations are likely to uncover new mechanisms and pathways that affect hematopoiesis and RBC turnover, offering insights that may inform further research into red cell biology. Indeed, recent reports have shown that genetic loci uncovered through an unbiased genome-wide study in human populations, together with follow-up functional studies incorporating gene expression, bioinformatic analyses and insights from mouse models and gene knockdown experiments, can greatly contribute to our understanding of the biological mechanisms underlying RBC production (30,31).

MATERIALS AND METHODS

Primary subjects and data collection

We performed GWA analysis of RBC traits in over 16 000 AAs from seven population-based cohorts that comprise the Continental Origins and Genetic Epidemiology Network (COGENT). The characteristics of each cohort were described in previous publications (14,29). Fasting blood samples were drawn and analyzed for RBC traits at designated clinical laboratories using an automated electronic cell counter. These counters directly measure Hgb concentration (in grams per deciliter), RBC count (in millions per microliter) and MCV, the average size of the RBC in femtoliters. Electronic cell counters calculate MCH, MCHC, Hct and RDW. Hct is the percentage of blood by volume that is occupied by RBC and is calculated by multiplying the RBC count in millions/microliter by the MCV in femtoliters. MCH is the average amount of Hgb inside an RBC expressed in picograms and is calculated by dividing the Hgb concentration by the RBC count in millions per microliter, then multiplying by 10. The MCHC is the average concentration of Hgb in RBCs and is calculated by dividing Hgb in grams per deciliter by Hct. The RDW is a measure of the variance in RBC size and is calculated by dividing the standard deviation of RBC volume by the MCV and multiplying by 100.

All participants self-reported their race/ethnicity. Additional clinical information was collected by self-report and clinical examination. Participants provided written informed consent as approved by local Human Subjects Committees. Study participants who were pregnant or had a diagnosis of cancer or AIDS at the time of blood count were excluded.

Replication subjects and data collection

For validation of novel, genome-wide significant associations identified in the COGENT discovery sample, we performed association analyses in two independent population-based samples: ~7700 AA youths ages 8–21 years from Children's Hospital of Philadelphia (CHOP) and 2010 AA adults from the Mount Sinai electronic Medical Records and Genomics (eMERGE) study. We also attempted to replicate novel loci in two other ethnic populations: 14 088 Japanese from RIKEN and up to 30 000 European Americans from CHARGE. Details of each validation cohort are provided under Supplementary Material.

Genotyping and quality-control

Genomic DNA was extracted from peripheral blood leukocytes and genotyping was performed on the Affymetrix 6.0 array or Illumina Omni or 1 M platforms within each cohort using methods described previously (14,29). DNA samples with a genome-wide genotyping success rate of <90% or sex discordance were excluded, as were genetic ancestry outliers (identified by cluster analysis using principal components analysis or multi-dimensional scaling). SNPs with a genotyping success rate of <95% or MAF <1%, monomorphic SNPs and SNPs that map to several genomic locations were removed from the analyses. Participants and SNPs passing basic quality control thresholds were imputed to >2.2 million autosomal SNPs based on HapMap2 haplotype data using a 1:1 mixture of Europeans (CEU) and Africans (YRI) as the reference panel. Details of the genotype imputation procedure have been described previously (14,29). Prior to discovery meta-analyses, SNPs were excluded if imputation quality metrics

(equivalent to the squared correlation between proximal imputed and genotyped SNPs) were <0.30.

Data analyses

For all cohorts, GWA analysis was performed on the raw, untransformed RBC trait using linear regression adjusted for covariates, implemented in either PLINK v1.07 or MACH2QTL v1.08. In GeneSTAR, the family structure was accounted for in the association tests using linear mixed-effects models implemented in R (32). For the 22 autosomes, analysis was performed using genotyped and imputed SNPs. For the X chromosome, only genotyped SNPs were analyzed due to the technical limitations of imputing X-linked SNPs. All analyses were performed under an additive genetic model using allelic dosage (genotyped or imputed) at each SNP, adjusted for age, age-squared, sex and clinic site (if applicable), 4–10 principal components.

For each phenotype, meta-analysis was conducted using inverse-variance weighted fixed-effects models to combine β coefficients and standard errors from study-level regression results for each SNP, to derive pooled estimates. Study-level results were corrected for genomic inflation factors (λ) by multiplying the standard errors (SEs) of the regression coefficients by the square-root of the study-specific λ. Meta-analyses were implemented in the METAL software. Between-study heterogeneity of results was assessed by using Cochran's Q statistic and the I^2 inconsistency metric. A threshold of $\alpha = 1 \times 10^{-8}$ was used to declare genome-wide statistical significance. This statistical threshold accounts for the greater nucleotide diversity and lower LD in African descent populations combined with testing of multiple, correlated RBC traits (31,33). We carried out replication testing of 'suggestive' SNPs selected on the basis of a more liberal significance threshold in our primary AA discovery GWAS ($P < 5 \times 10^{-8}$).

To assess the potential existence of multiple, independent variants influencing a trait at the same locus (allelic heterogeneity), regression analyses were repeated in the largest sample (WHI, $n = 8095$), conditional on the most strongly associated (index) SNP in that region.

We also assessed the transferability to AAs of SNPs previously associated with RBC traits in populations of European or Japanese ancestry by assessing association with RBC traits in the COGENT discovery meta-analyses. For validation, we considered consistency of direction of effect, and assessed statistical significance using a simple Bonferroni adjustment for the total number of SNPs assessed, using a two-sided hypothesis test.

Local ancestry estimation and admixture mapping in WHI

For each AA individual in the WHI sample, locus-specific ancestry was estimated using an extension of the model described by Tang *et al*. (34). We used phased haplotype data from HapMap3 CEU and YRI individuals as reference panels. An admixture mapping analysis was performed in WHI to test for association between Hgb levels and ancestry at each genomic location (local ancestry), while adjusting for the first 10 principal components, regions of recruitment, clinical trial, age and age-squared. The critical value for genome-wide significance level of admixture mapping is substantially lower than the genotype test due to the extensive correlation in local ancestry between adjacent markers that result from the recent admixture in AAs. We therefore adopted an empirically determined genome-wide significance threshold of $P < 7.1 \times 10^{-6}$, which corresponds to a Bonferroni correction of ~7000 independent tests (35).

Copy number variation (CNV) analysis using 1000 genomes data

We used the 1000 Genomes sequencing data to investigate CNVs at the chromosome 16 p31 alpha-globin locus, studying 946 African-ancestry samples at roughly 4× sequencing coverage. As a result of noise in depth-based genotyping at this locus (due to low-pass sequencing, high %GC and potential overlapping variants), some of our analyses were confined to the 76 YRI samples, which have higher sequence coverage in 1000 Genomes data and more complete genotyping (call rate 84 % at 95 % CI).

SUPPLEMENTARY MATERIAL

Supplementary Material is available at *HMG* online.

ACKNOWLEDGMENTS

The authors wish to acknowledge the support of the National Heart, Lung and Blood Institute and the contributions of the involved research institutions, study investigators, field staff and study participants of Atherosclerosis Risk in Communities (ARIC), Coronary Artery Risk in Young Adults (CARDIA), Jackson Heart Study (JHS) and Broad Institute in creating the Candidate-gene Association Resource for biomedical research (CARe; http://public.nhlbi.nih.gov/GeneticsGeno?mics/home/care.aspx).

The authors also wish to thank the investigators, staff and participants of GeneSTAR, Health ABC, Healthy Aging in Neighborhoods of Diversity across the Life Span Study (HANDLS) and Women Health Initiative (WHI) for their important contributions. A listing of WHI investigators can be found at http://www.whiscience.org/publications/WHI_investigators_shortlist.pdf.

We thank all the children who donated blood samples for genetic research purpose. The CHOP study was funded by the Institute Development Funds to the Center for Applied Genomics at the Children's Hospital of Philadelphia and an Adele S. and Daniel S. Kubert Estate gift to the Center for Applied Genomics.

The Mount Sinai IPM Biobank Program is supported by The Andrea and Charles Bronfman Philanthropies.

The authors acknowledge the essential role of the Cohorts for Heart and Aging Research in Genome Epidemiology (CHARGE) Consortium in development and support of this manuscript. CHARGE members include the Rotterdam Study (RS), Framingham Heart Study (FHS), Cardiovascular Health Study (CHS), the NHLBI's Atherosclerosis Risk in Communities (ARIC) Study and the NIA's Iceland Age,

Gene/Environment Susceptibility (AGES) Study. The collaboration of studies such as the Health Aging and Body Composition Study (Health ABC), the Baltimore Longitudinal Study of Aging (BLSA), the Invecchiare in Chianti Study (InChianti), and the Heart and Vascular Health Study (HVH) also played a vital role.

The following parent studies contributed study data, ancillary study data and DNA samples through the Broad Institute (N01-HC-65226) to create this genotype/phenotype data base for wide dissemination to the biomedical research community:

Atherosclerosis Risk in Communities (ARIC): University of North Carolina at Chapel Hill (N01-HC-55015), Baylor Medical College (N01-HC-55016), University of Mississippi Medical Center (N01-HC-55021), University of Minnesota (N01-HC-55019), Johns Hopkins University (N01-HC-55020), University of Texas, Houston (N01-HC-55017), University of North Carolina (N01-HC-55018). Other NIH support contributing to the GWAS in ARIC are: R01HL 087641, R01HL59367, R01HL86694, U01HG004402 and HHSN268200625226C.

Coronary Artery Risk in Young Adults (CARDIA): University of Alabama at Birmingham (N01-HC-48047), University of Minnesota (N01-HC-48048), Northwestern University (N01-HC-48049), Kaiser Foundation Research Institute (N01-HC-48050), University of Alabama at Birmingham (N01-HC-95095), Tufts-New England Medical Center (N01-HC-45204), Wake Forest University (N01-HC-45205), Harbor-UCLA Research and Education Institute (N01-HC-05187), University of California, Irvine (N01-HC-45134, N01-HC-95100).

Jackson Heart Study (JHS): Jackson State University (N01-HC-95170), University of Mississippi (N01-HC-95171), Tougaloo College (N01-HC-95172).

Healthy Aging in Neighborhoods of Diversity across the Life Span Study (HANDLS): this research was supported by the Intramural Research Program of the NIH, National Institute on Aging and the National Center on Minority Health and Health Disparities (intramural project Z01-AG000513 and human subjects protocol 2009-149). Data analyses for the HANDLS study utilized the high-performance computational capabilities of the Biowulf Linux cluster at the National Institutes of Health, Bethesda, MD, USA (http://biowulf.nih.gov).

Health ABC: this research was supported by NIA contracts N01AG62101, N01AG62103 and N01AG62106. The GWAS was funded by NIA grant 1R01AG032098-01A1 to Wake Forest University Health Sciences and 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 HHSN268200782096C. This research was supported in part by the Intramural Research Program of the NIH, National Institute on Aging.

GeneSTAR: this research was supported by the National Heart, Lung and Blood Institute (NHLBI) through the PROGENI (U01 HL72518) and STAMPEED (R01 HL087698-01) consortia. Additional support was provided by grants from the NIH/National Institute of Nursing Research (R01 NR08153), and the NIH/National Center for Research Resources (M01-RR000052) to the Johns Hopkins General Clinical Research Center.

WHI: the WHI program is funded by the National Heart, Lung and Blood Institute, National Institutes of Health, US 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.

AGES: the Age, Gene/Environment Susceptibility Reykjavik Study is funded by NIH contract N01-AG-12100, the NIA Intramural Research Program, Hjartavernd (the Icelandic Heart Association) and the Althingi (the Icelandic Parliament).

Framingham: the National Heart, Lung and Blood Institute's Framingham Heart Study is a joint project of the National Institutes of Health and Boston University School of Medicine and was supported bythe National Heart, Lung, and Blood Institute's Framingham Heart Study (contract No. N01-HC-25195) and its contract with Affymetrix, Inc. for genotyping services (contract No. N02-HL-6-4278). Analyses reflect the efforts and resource development from the Framingham Heart Study investigators participating in the SNP Health Association Resource (SHARe) project. A portion of this research was conducted using the Linux Cluster for Genetic Analysis (LinGA-II) funded by the Robert Dawson Evans Endowment of the Department of Medicine at Boston University School of Medicine and Boston Medical Center.

InChianti: the InChianti Study was supported as a "targeted project" (ICS 110.1RS97.71) by the Italian Ministry of Health, by the US National Institute on Aging (Contracts N01-AG-916413, N01-AG-821336, 263 MD 9164 13 and 263 MD 821336) and in part by the Intramural Research Program, National Institute on Aging, National Institutes of Health, USA.

Rotterdam: Rotterdam Study GWAS database of the Rotterdam Study was funded through the Netherlands Organization of Scientific Research NWO (no. 175.010.2005.011, 911.03.012) and the Research Institute for Diseases in the Elderly (RIDE). This study was supported by the Netherlands Genomics Initiative (NGI)/NWO project number 050 060 810 (Netherlands Consortium for Healthy Ageing). We thank Dr Michael Moorhouse, Pascal Arp, Mila Jhamai, Marijn Verkerk and Sander Bervoets for their help in creating the genetic database. We thank the laboratory technicians Jeannette M Vergeer—Drop, Bernadette H M van Ast—Copier, Andy A L J van Oosterhout, Sue Ellen Mauricia, Andrea J M Vermeij—Verdoold, Els Halbmeijer—van der Plas, Debby M S Lont and Hasna Kariouh for their help in phenotype assessment. The Rotterdam Study is supported by the Erasmus Medical Center and Erasmus University, Rotterdam; the Netherlands organization for scientific research (NWO), the Netherlands Organization for the Health Research and Development (ZonMw), the Research Institute for Diseases in the Elderly (RIDE), the Netherlands Heart Foundation, the Ministry of Education, Culture and Science, the Ministry of Health, Welfare and Sports, the European Commission (DG XII) and the Municipality of Rotterdam.

RIKEN: we would like to thank all the staff of the Laboratory for Statistical Analysis at RIKEN for their technical assistance. The BioBank Japan Project was supported by Ministry of Education, Culture, Sports, Science and Technology, Japan.

Conflict of Interest statement. None declared.

FUNDING

Additional support for this work was provided by NIH (R01 HL71862-06 and ARRA N000949304 to A.P.R.). Some of the results of this paper were obtained by using the program package S.A.G.E., which is supported by a US Public Health Service Resource Grant (RR03655) from the National Center for Research Resources. Additional support came from the National Cancer Institute (grant R25CA094880 to U.M.S.).

REFERENCES

1. Camaschella, C. (2005) Understanding iron homeostasis through genetic analysis of hemochromatosis and related disorders. *Blood*, **106**, 3710–3717.
2. Melis, M.A., Cau, M., Congiu, R., Sole, G., Barella, S., Cao, A., Westerman, M., Cazzola, M. and Galanello, R. (2008) A mutation in the TMPRSS6 gene, encoding a transmembrane serine protease that suppresses hepcidin production, in familial iron deficiency anemia refractory to oral iron. *Haematologica*, **93**, 1473–1479.
3. Patel, K.V., Longo, D.L., Ershler, W.B., Yu, B., Semba, R.D., Ferrucci, L. and Guralnik, J.M. (2009) Haemoglobin concentration and the risk of death in older adults: differences by race/ethnicity in the NHANES III follow-up. *Br. J. Haematol.*, **145**, 514–523.
4. Schechter, G.P. (2006) Hemoglobin levels in African-Americans. *Blood*, **107**, 2208; author reply 2208–2209.
5. Beutler, E. and Duparc, S. (2007) Glucose-6-phosphate dehydrogenase deficiency and antimalarial drug development. *Am. J. Trop. Med. Hyg.*, **77**, 779–789.
6. 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. *Nat. Genet.*, **42**, 210–215.
7. Soranzo, N., Spector, T.D., Mangino, M., Kühnel, 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. *Nat Genet.*, **41**, 1182–1190.
8. 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. *Nat. Genet.*, **41**, 1191–1198.
9. Lo, K.S., Wilson, J.G., Lange, L.A., Folsom, A.R., Galarneau, G., Ganesh, S.K., Grant, S.F., Keating, B.J., McCarroll, S.A., Mohler, E.R. III *et al.* (2011) Genetic association analysis highlights new loci that modulate hematological trait variation in Caucasians and African Americans. *Hum. Genet.*, **129**, 307–317.
10. 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**, e13011.
11. Mukai, M., Ikegami, K., Sugiura, Y., Takeshita, K., Nakagawa, A. and Setou, M. (2009) Recombinant mammalian tubulin polyglutamylase TTLL7 performs both initiation and elongation of polyglutamylation on beta-tubulin through a random sequential pathway. *Biochemistry*, **48**, 1084–1093.
12. Alves, L.R., Costa, E.S., Sorgine, M.H., Nascimento-Silva, M.C., Teodosio, C., Barcena, P., Castro-Faria-Neto, H.C., Bozza, P.T., Orfao, A., Oliveira, P.L. *et al.* (2011) Heme-oxygenases during erythropoiesis in K562 and human bone marrow cells. *PLoS One*, **6**, e21358.
13. Feder, J.N., Gnirke, A., Thomas, W., Tsuchihashi, Z., Ruddy, D.A., Basava, A., Dormishian, F., Domingo, R. Jr., Ellis, M.C., Fullan, A. *et al.* (1996) A novel MHC class I-like gene is mutated in patients with hereditary haemochromatosis. *Nat. Genet.*, **13**, 399–408.
14. Qayyum, R., Snively, B.M., Ziv, E., Nalls, M.A., Liu, Y., Tang, W., Yanek, L.R., Lange, L., Evans, M.K., Ganesh, S. *et al.* (2012) A meta-analysis and genome-wide association study of platelet count and mean platelet volume in african americans. *PLoS Genet.*, **8**, e1002491.
15. Beutler, E. and West, C. (2005) Hematologic differences between African-Americans and whites: the roles of iron deficiency and alpha-thalassemia on hemoglobin levels and mean corpuscular volume. *Blood*, **106**, 740–745.
16. Viprakasit, V., Harteveld, C.L., Ayyub, H., Stanley, J.S., Giordano, P.C., Wood, W.G. and Higgs, D.R. (2006) A novel deletion causing alpha thalassemia clarifies the importance of the major human alpha globin regulatory element. *Blood*, **107**, 3811–3812.
17. Beutler, E. (1994) G6PD deficiency. *Blood*, **84**, 3613–3636.
18. Nakhaee, A., Dabiri, S. and Noora, M. (2009) Survey of the prevalence of glucose-6-phosphate dehydrogenase (G6PD) deficiency in admitted men for premarriage tests in Zahedan-Iran Reference Laboratory. *Zahedan J. Res. Med. Sci.*, **11**, 0–0.
19. Chishti, A.H. (1998) Function of p55 and its nonerythroid homologues. *Curr. Opin. Hematol.*, **5**, 116–121.
20. Watt, S.M., Buhring, H.J., Rappold, I., Chan, J.Y., Lee-Prudhoe, J., Jones, T., Zannettino, A.C., Simmons, P.J., Doyonnas, R., Sheer, D. *et al.* (1998) CD164, a novel sialomucin on CD34(+) and erythroid subsets, is located on human chromosome 6q21. *Blood*, **92**, 849–866.
21. Forde, S., Tye, B.J., Newey, S.E., Roubelakis, M., Smythe, J., McGuckin, C.P., Pettengell, R. and Watt, S.M. (2007) Endolyn (CD164) modulates the CXCL12-mediated migration of umbilical cord blood CD133+ cells. *Blood*, **109**, 1825–1833.
22. Klingmuller, U., Wu, H., Hsiao, J.G., Toker, A., Duckworth, B.C., Cantley, L.C. and Lodish, H.F. (1997) Identification of a novel pathway important for proliferation and differentiation of primary erythroid progenitors. *Proc. Natl Acad. Sci. USA*, **94**, 3016–3021.
23. Gobbi, G., Mirandola, P., Sponzilli, I., Micheloni, C., Malinverno, C., Cocco, L. and Vitale, M. (2007) Timing and expression level of protein kinase C epsilon regulate the megakaryocytic differentiation of human CD34 cells. *Stem Cells*, **25**, 2322–2329.
24. Mirandola, P., Gobbi, G., Ponti, C., Sponzilli, I., Cocco, L. and Vitale, M. (2006) PKC epsilon controls protection against TRAIL in erythroid progenitors. *Blood*, **107**, 508–513.
25. Kollia, P., Noguchi, C.T., Fibach, E., Loukopoulos, D. and Schechter, A.N. (1997) Modulation of globin gene expression in cultured erythroid precursors derived from normal individuals: transcriptional and posttranscriptional regulation by hemin. *Proc. Assoc. Am. Phys.*, **109**, 420–428.
26. Melefors, O., Goossen, B., Johansson, H.E., Stripecke, R., Gray, N.K. and Hentze, M.W. (1993) Translational control of 5-aminolevulinate synthase mRNA by iron-responsive elements in erythroid cells. *J. Biol. Chem.*, **268**, 5974–5978.
27. Rosenbloom, K.R., Dreszer, T.R., Long, J.C., Malladi, V.S., Sloan, C.A., Raney, B.J., Cline, M.S., Karolchik, D., Barber, G.P., Clawson, H. *et al.* (2012) In *Nucleic Acids Res.*, **Vol. 40**, D912–D917.
28. Zakai, N.A., McClure, L.A., Prineas, R., Howard, G., McClellan, W., Holmes, C.E., Newsome, B.B., Warnock, D.G., Audhya, P. and Cushman, M. (2009) Correlates of anemia in American blacks and whites: the REGARDS Renal Ancillary Study. *Am. J. Epidemiol.*, **169**, 355–364.
29. Reiner, A.P., Lettre, G., Nalls, M.A., Ganesh, S.K., Mathias, R., Austin, M.A., Dean, E., Arepalli, S., Britton, A., Chen, Z. *et al.* (2011) Genome-wide association study of white blood cell count in 16,388 African Americans: the continental origins and genetic epidemiology network (COGENT). *PLoS Genet.*, **7**, e1002108.
30. Sankaran, V.G., Ludwig, L.S., Sicinska, E., Xu, J., Bauer, D.E., Eng, J.C., Patterson, H.C., Metcalf, R.A., Natkunam, Y., Orkin, S.H. *et al.* (2012) Cyclin D3 coordinates the cell cycle during differentiation to regulate erythrocyte size and number. *Genes Dev.*, **26**, 2075–2087.
31. 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.
32. Chen, M.H. and Yang, Q. (2010) GWAF: an R package for genome-wide association analyses with family data. *Bioinformatics*, **26**, 580–581.
33. Pe'er, I., Yelensky, R., Altshuler, D. and Daly, M.J. (2008) Estimation of the multiple testing burden for genomewide association studies of nearly all common variants. *Genet. Epidemiol.*, **32**, 381–385.
34. Tang, H., Coram, M., Wang, P., Zhu, X. and Risch, N. (2006) Reconstructing genetic ancestry blocks in admixed individuals. *Am. J. Hum. Genet.*, **79**, 1–12.
35. Tang, H., Siegmund, D.O., Johnson, N.A., Romieu, I. and London, S.J. (2010) Joint testing of genotype and ancestry association in admixed families. *Genet. Epidemiol.*, **34**, 783–791.

SUPPLEMENTAL METHODS

Replication subjects and data collection

Children's Hospital of Philadelphia Study

All participating children were recruited under the research protocol approved by the Institutional Review Board at the Children's Hospital of Philadelphia, and written informed consent was obtained from their parents. We only included the 7,943 genetically inferred African American children (Age in years: 7.394 ± 5.754) in the analysis and further excluded any subject with missing data or hematological traits beyond three standard deviation of the mean from the analysis of the trait studied. Samples were genotyped on the Illumina HumanHap550 or the Human610-Quad platform and only those with call rate greater than 98% were included in the analysis. Only those SNPs met the following quality control criteria were included in the analysis: genotype missing rate < 5%, minor allele frequency > 0.01, as well as HWE-pvalue > 0.0001. Cryptic relatedness was detected by identity-by-descent (IBD) analysis using software PLINK (1) and one sample from each pair was excluded if IBD score is > 0.50. We performed SNP imputation using software IMPUTE2 (2, 3) with HapMap2 CEU and YRI combined data as the reference panel. We further conducted association analysis using missing data likelihood score test implemented in software SNPTEST v2 (3), with age, sex, hematological diseases status and the first three principal components (PC) from EIGENSTRAT (4) analysis as covariates.

eMERGE Mount Sinai Study

The Mount Sinai Biobank Program at the Institute for Personalized Medicine (IPM) is a consented, Electronic Medical Record (EMR)-linked medical care setting biorepository of the Mount Sinai Medical Center (MSMC), with currently more than 22,000 participants. The Mount Sinai Biobank Program (IRB # 07-0529 0001 02 ME) operates under an IRB-approved research protocol with IRB-approved informed consent forms. All study participants provided written informed consent.

The Mount Sinai Biobank populations include 28% African American, 38% Hispanic Latino (predominantly of Caribbean origin) and 23% Caucasian/White. Biobank operations are fully integrated in clinical care processes and recruitment currently occurs at a broad spectrum of over 30 clinical care sites. For the present analyses, we contributed data African American (self-reported) adults for whom we had RBC and genotype data available. Laboratory red blood cell measurements were derived from participants' EMRs.

A total of 888 samples were genotyped using the Affymetrix GeneChip Human Mapping 500K Array Set and 3,478 samples were genotyped using the Illumina OmniExpress. Quality control, imputation and association analyses were performed for the two sub-sets (by genotyping platform) separately. We excluded samples that did not meet the quality control criteria (sample call rate <95%; heterozygosity Z-value >|6|; samples with evidence of (cryptic) relatedness (IBD>0.185); samples that deviate from African ancestry clustering based on the HapMap II genomic using CEU, YRI, JPN, and CHN data. We removed SNPs with MAF <1%, those which distribution was not consistent with the Hardy-Weinberg Equilibrium expectation (P < 0.001), and SNPs with low call rate and those that show evidence of batch/plate effects. As such, 2012 individuals and 711,270 genotyped SNPs were available for imputation, which was performed using IMPUTE2 (2, 3) using the 1000Genome data (March 2012 version 3). Subsequent association between imputed and genotyped SNPs with RBC adjusting age sex and relevant PCs was performed using SNPTEST (3), assuming an additive model of inheritance, using the score-based method.

CHARGE European GWAS Consortium

The European-American GWAS replication sample comprised 30,000 subjects from 7 CHARGE cohorts. Details of the CHARGE consortium including subject details and study designs, are described elsewhere (5, 6). For the current analysis, red blood cell phenotypes were derived from data provided by automated blood cell counters commonly employed in clinical and epidemiological studies to interrogate common hematological elements found in peripheral blood. Each study excluded all participants with any RBC measure outside of +/- 2 standard deviations from the mean value for that trait.

RIKEN Japanese

The Japanese replication sample consists of 14,767 participants with red blood cell phenotypes, originally obtained as part of the BioBank Japan GWAS project (7). The mean age was 62.3 ± 10.5 years and

34.5% were female. For the current analysis, RBC phenotypes were derived from medical records. Genotyping was performed using Illumina HumanHap610-Quad Genotyping BeadChip or Illumina HumanHap550v3 Genotyping BeadChip. Subjects with call rates < 0.98, closely related subjects based on the identity-by-descent (IBD), and subjects who were determined to be of non-Japanese origin by either self-report or by PCA were excluded from analysis. SNPs with MAF < 0.01 or with an exact P-value of the Hardy-Weinberg equilibrium test < 1.0×10^{-7} were excluded. Genotype imputation was performed using MACH 1.0 and genotype data from Phase II HapMap JPT and CHB individuals (release 24) as reference panel. Quality control filters of MAF ≥ 0.01 and *Rsq* values ≥ 0.7 were applied for the imputed SNPs.

References

1 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.

2 Howie, B.N., Donnelly, P. and Marchini, J. (2009) A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet.*, **5**, e1000529.

3 Marchini, J., Howie, B., Myers, S., McVean, G. and Donnelly, P. (2007) A new multipoint method for genome-wide association studies by imputation of genotypes. *Nat. Genet.*, **39**, 906-913.

4 Price, A.L., Patterson, N.J., Plenge, R.M., Weinblatt, M.E., Shadick, N.A. and Reich, D. (2006) Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.*, **38**, 904-909.

5 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. *Nat. Genet.*, **41**, 1191-1198.

6 Psaty, B.M., O'Donnell, C.J., Gudnason, V., Lunetta, K.L., Folsom, A.R., Rotter, J.I., Uitterlinden, A.G., Harris, T.B., Witteman, J.C. and Boerwinkle, E. (2009) Cohorts for Heart and Aging Research in Genomic Epidemiology (CHARGE) Consortium: Design of prospective meta-analyses of genome-wide association studies from 5 cohorts. *Circulation: Cardiovascular Genetics*, **2**, 73-80.

7 Kamatani, Y., Matsuda, K., Okada, Y., Kubo, M., Hosono, N., Daigo, Y., Nakamura, Y. and Kamatani, N. Genome-wide association study of hematological and biochemical traits in a Japanese population. *Nat. Genet.*, **42**, 210-215.

SUPPLEMENTAL TABLES

Table S1. Characteristics of Continental Origins and Genetic Epidemiology Networks (COGENT) African-American GWAS Participants (n=16,485)

Study	Atherosclerosis Risk in Communities (ARIC)	Coronary Artery Risk Development in Young Adults (CARDIA)	Johns Hopkins Genetic Study of Atherosclerosis Risk (GeneSTAR)	Healthy Aging in Neighborhoods of Diversity across the Life Span (HANDLS)	Health, Aging, and Body Composition (Health ABC)	Jackson Heart Study (JHS)	Women's Health Initiative (WHI)
Sample size	2782	943	931	873	773	2145	8095
Study design	Population-based, unrelated	Population-based, unrelated	Population-based, family	Population-based, unrelated	Population-based, unrelated	Population-based, unrelated	Population-based, unrelated
Age, years (SD)	53.4 (5.8)	24.4 (3.8)	44.5 (12.6)	48.2 (9.0)	73.4 (2.8)	50.0 (12.1)	61.6 (7.0)
% Female	63.2	58.7	62.2	56.0	58.8	61.2	100
Hemoglobin (SD), g/dL	13.2 (1.47) 2671	13.8 (1.83) 943	13.0 (1.38) g/dL 931	13.34 (1.47) 863	13.14 (1.22) 772	13.1 (1.49) 2110	12.9 (1.04) 8300
Hematocrit (SD), %	40.2 (4.33) 2675	41.1 (4.51) 943	39.8 (3.88) % 931	40.40 (4.17) 864	39.44 (3.67) 773	39.5 (4.22) 2111	39.0 (3.08) 8294
MCV (SD), fL	86.6 (6.24) 2558	NA	87.7 (5.98) fL 931	88.81 (6.67) 861	88.62 (5.73) 771	86.8 (6.41) 1993	NA
MCH (SD), pg	29.4 (2.45) 189	NA	28.7 (2.32) pg 931	29.33 (2.56) 858	29.59 (2.23) 772	28.8 (2.54) 1993	NA
MCHC (SD), g/dL	33.5 (0.719) 189	NA	32.7 (0.820) g/dL 931	32.98 (0.91) 859	33.36 (0.70) 773	33.1 (0.930) 1993	33.1 (1.18) 8277
RBC count (SD) x10^6/uL	NA	4.75 (0.563) 934	4.54 (0.470) x10^6/uL 931	4.57 (0.49) 868	4.47 (0.47) 773	4.57 (0.514) 1993	NA
RDW (SD), %	NA	NA	13.6 (1.12) 931	14.22 (1.35) 852	13.89 (1.32) 773	13.7 (1.39)1993	NA

SD=standard deviation; NA = not available

Table S2. Pearson's correlation coefficients for red blood cell traits in Jackson Heart Study African-American GWAS Participants (n=1,992)

	hemoglobin	hematocrit	RBC count	mcv	mch	mchc	rdw
hemoglobin	1						
hematocrit	0.9665	1					
RBC count	0.6787	0.7656	1				
mcv	0.3441	0.2673	-0.4086	1			
mch	0.3937	0.2595	-0.3901	0.9589	1		
mchc	0.3567	0.1329	-0.1457	0.4032	0.6443	1	
rdw	-0.3843	-0.313	0.0308	-0.4965	-0.5345	-0.4097	1

Pearson’s correlation coefficients are very similar in GeneSTAR

	hemoglobin	hematocrit	RBC count	mcv	mch	mchc	rdw
hemoglobin	1						
hematocrit	0.9728	1					
RBC count	0.6783	0.7541	1				
mcv	0.3460	0.2631	-0.4140	1			
mch	0.4144	0.2772	-0.3770	0.9588	1		
mchc	0.4222	0.1952	-0.0911	0.4147	0.6504	1	
rdw	-0.3304	-0.2699	0.0066	-0.3864	-0.4341	-0.3687	1

Table S3: Genome-wide significant associations for all red blood cell traits

Marker	Chr.	Position	Genes (location)*	Function (database)	a1	a2	Direction	P-value	Beta	SE	Trait	Freq1	N
rs1050828	X	153417411	G6PD	coding-nonsynonymous (dbSNP)	T	C	---???-	4.84E-15	-0.4835	0.0617	HCT	0.1166	12005
rs762516	X	153417857	G6PD	intron(dbSNP)	T	C	--?-??-	2.17E-16	-0.4525	0.0551	HCT	0.1483	13782

rs5987239	X	153186763	TKTL1	intron(dbSNP)	C	G	+++?+	6.60E-10	0.3238	0.0524	HCT	0.8339	13799
rs7059306	X	152968184	MECP2	intron(dbSNP)	A	G	-----	1.80E-09	-0.3011	0.05	HCT	0.1668	14709
rs2734643	X	152944381	MECP2	utr-3(dbSNP)	T	C	-----	4.80E-09	-0.2969	0.0507	HCT	0.1643	14664
rs5987026	X	153665315	MPP1	intron(dbSNP)	T	C	+-+?+	3.50E-08	0.2408	0.0437	HCT	0.7218	13785
rs13008603	2	46209352	PRKCE	intron(dbSNP)	A	C	--+----	4.02E-09	-0.2769	0.0471	HCT	0.163	16496
rs17034641	2	46226148	PRKCE	intron(dbSNP)	A	G	--+----	3.89E-08	-0.2476	0.0451	HCT	0.1899	16496
rs12990867	2	46225915	PRKCE	intron(dbSNP)	A	G	--+----	4.07E-08	-0.247	0.045	HCT	0.1925	16496
rs13335629	16	250381	ITFG3	intron(dbSNP)	A	G	---?---	2.63E-23	-0.1898	0.0191	HGB	0.1201	15516
rs13339109	16	202300	LUC7L	intron(dbSNP)	A	C	-------	2.28E-18	-0.2213	0.0253	HGB	0.0601	16446
rs1211375	16	180281	LUC7L	intron(dbSNP)	A	C	-------	4.54E-16	-0.1093	0.0135	HGB	0.2796	16447
rs2562181	16	81659	C16orf35	intron(dbSNP)	T	C	-------	1.36E-13	-0.0986	0.0133	HGB	0.3243	16418
rs2541612	16	114196	C16orf35	intron(dbSNP)	A	G	++++???	7.38E-13	0.105	0.0146	HGB	0.7142	13765
rs9929571	16	234387	ITFG3	intron(dbSNP)	A	G	+++++-+	7.30E-13	0.1294	0.018	HGB	0.8724	16428
rs1088638	16	31010	POLR3K (5990-downstream)	intergenic(GVS)	A	G	-----??	2.99E-12	-0.148	0.0212	HGB	0.1034	14554
rs798673	16	35481	POLR3K (1519 downstream)	near-gene-3 (dbSNP)	T	C	------?	1.55E-11	-0.2485	0.0369	HGB	0.0296	15529
rs369322	16	376809	none	intergenic(GVS)	A	G	------?	7.03E-11	-0.1347	0.0207	HGB	0.1824	15567
rs3176398	16	70406	MPG	intron(dbSNP)	C	G	------?	1.34E-10	-0.1561	0.0243	HGB	0.0704	15529
rs9940149	16	240642	ITFG3	intron(dbSNP)	A	G	-------	2.96E-09	-0.0744	0.0125	HGB	0.5984	16442
rs762516	X	153417857	G6PD	intron(dbSNP)	T	C	---?-	3.73E-18	-0.1614	0.0186	HGB	0.1462	13775
rs1050828	X	153417411	G6PD	coding-nonsynonymous (dbSNP)	T	C	?----	6.71E-16	-0.169	0.0209	HGB	0.1151	11991
rs5987239	X	153186763	TKTL1	intron(dbSNP)	C	G	+++?+	2.56E-12	0.1237	0.0177	HGB	0.8337	13800
rs7059306	X	152968184	MECP2	intron(dbSNP)	A	G	-----	6.61E-11	-0.1109	0.017	HGB	0.1668	14699
rs2734643	X	152944381	MECP2	utr-3(dbSNP)	T	C	-----	1.91E-10	-0.1095	0.0172	HGB	0.1641	14654
rs5987026	X	153665315	MPP1	intron(dbSNP)	T	C	+-+?+	5.24E-09	0.086	0.0147	HGB	0.7219	13786
rs1734792	X	152994254	MECP2	intron(dbSNP)	A	C	-----	1.13E-08	-0.0874	0.0153	HGB	0.2277	14687
rs41362344	X	154028266	none	intergenic(GVS)	T	C	---?-	2.19E-08	-0.0908	0.0162	HGB	0.2082	13801

rs5987027	X	153667301	MPP1	intron (UCSC)	T	C	-----	2.35E-08	-0.0823	0.0147	HGB	0.2488	14679
rs13335629	16	250381	ITFG3	intron(dbSNP)	A	G	-?--?-	8.66E-48	-0.3298	0.0227	MCHC	0.1168	11255
rs13339109	16	202300	LUC7L	intron(dbSNP)	A	C	-?----	8.32E-32	-0.3424	0.0292	MCHC	0.058	12134
rs1211375	16	180281	LUC7L	intron(dbSNP)	A	C	------	2.81E-30	-0.1718	0.015	MCHC	0.2776	12326
rs3176398	16	70406	MPG	intron(dbSNP)	C	G	----?-	2.26E-26	-0.2996	0.0282	MCHC	0.0705	11453
rs2541612	16	114196	C16orf35	intron(dbSNP)	A	G	+????+	2.56E-23	0.1731	0.0174	MCHC	0.7154	10268
rs7197554	16	124242	C16orf35	intron(dbSNP)	A	C	+++-++	1.11E-22	0.2486	0.0254	MCHC	0.8774	12327
rs2562181	16	81659	C16orf35	intron(dbSNP)	T	C	------	4.17E-22	-0.1484	0.0153	MCHC	0.3245	11725
rs9940149	16	240642	ITFG3	intron(dbSNP)	A	G	------	7.97E-20	-0.1282	0.0141	MCHC	0.5913	12315
rs2541613	16	102998	C16orf35	intron(dbSNP)	A	G	++++++	1.77E-17	0.1217	0.0143	MCHC	0.5125	12332
rs9929571	16	234387	ITFG3	intron(dbSNP)	A	G	++++++	7.10E-16	0.1635	0.0203	MCHC	0.877	12301
rs11248914	16	233563	ITFG3	intron(dbSNP)	T	C	------	7.34E-16	-0.113	0.014	MCHC	0.6229	12327
rs1061438	16	74457	MPG	intron (UCSC)	A	G	---+--	2.46E-15	-0.1204	0.0152	MCHC	0.4881	12331
rs2239739	16	251854	ITFG3	intron(dbSNP)	A	G	++++++	5.80E-15	0.113	0.0145	MCHC	0.4727	12291
rs2562182	16	73946	MPG	near-gene-3 (dbSNP)	A	G	++++++	8.03E-14	0.12	0.0161	MCHC	0.4331	12319
rs431324	16	388173	NME4	intron(dbSNP)	A	G	+++-++	1.06E-12	0.1441	0.0202	MCHC	0.7733	12333
rs177510	16	103626	C16orf35	intron(dbSNP)	T	C	------	1.30E-12	-0.1245	0.0175	MCHC	0.1817	12330
rs369322	16	376809	none	intergenic(GVS)	A	G	----?-	2.43E-12	-0.1725	0.0246	MCHC	0.1809	11461
rs743725	16	76888	C16orf35	intron(dbSNP)	T	C	---+--	9.03E-12	-0.1032	0.0151	MCHC	0.398	12344
rs7200589	16	289332	AXIN1	intron(dbSNP)	A	G	------	3.77E-11	-0.144	0.0218	MCHC	0.1459	12332
rs1203957	16	181211	LUC7L	intron(dbSNP)	T	G	++++++	5.45E-11	0.1021	0.0156	MCHC	0.254	12325
rs3743883	16	344751	AXIN1 (2286 upstream)	intergenic(GVS)	T	C	++++++	9.77E-11	0.1369	0.0212	MCHC	0.1431	12344
rs9927150	16	764099	MPFL	intron(dbSNP)	T	G	+?++++	1.00E-10	0.1275	0.0197	MCHC	0.2898	12142
rs4141288	16	225314	ITFG3	intron(dbSNP)	T	C	++++++	2.70E-10	0.1137	0.018	MCHC	0.2194	12332
rs11642609	16	192480	LUC7L	intron(dbSNP)	T	C	---+--	2.83E-10	-0.1061	0.0168	MCHC	0.7562	12330
rs1122794	16	249156	ITFG3	intron(dbSNP)	A	C	+++-++	6.82E-10	0.1078	0.0175	MCHC	0.1977	12332

rs367146	16	376258	none	intergenic(GVS)	A	G	---+--	9.75E-10	-0.1668	0.0273	MCHC	0.1097	12333
rs1203980	16	204643	LUC7L	intron(dbSNP)	T	C	++++++	1.13E-09	0.0867	0.0142	MCHC	0.5632	12344
rs2562189	16	127429	C16orf35	intron(dbSNP)	T	C	------	1.21E-09	-0.1073	0.0176	MCHC	0.6972	12330
rs3848368	16	354609	MRPL28 (2788 downstream)	intergenic(GVS)	T	G	+??+++	1.40E-09	0.1109	0.0183	MCHC	0.1971	11345
rs2157115	16	53970	RHBDF1	intron(dbSNP)	T	C	------	2.52E-09	-0.106	0.0178	MCHC	0.5196	12329
rs2857998	16	125123	C16orf35	intron(dbSNP)	A	G	+++-++	4.83E-09	0.1019	0.0174	MCHC	0.2655	12333
rs710080	16	69277	MPG	intron(dbSNP)	A	G	++++++	5.76E-09	0.101	0.0173	MCHC	0.3805	12332
rs2301522	16	299954	AXIN1	intron(dbSNP)	A	G	+++-++	9.98E-09	0.1248	0.0218	MCHC	0.1254	12325
rs17136255	16	340476	AXIN1	intron(dbSNP)	T	C	------	1.30E-08	-0.1035	0.0182	MCHC	0.1855	12332
rs2562147	16	54535	RHBDF1	intron(dbSNP)	A	G	-??---	6.63E-09	-0.087	0.015	MCHC	0.3523	11284
rs7195617	16	316782	AXIN1	intron(dbSNP)	A	G	------	2.01E-08	-0.1131	0.0202	MCHC	0.8678	12297
rs1203974	16	217459	LUC7L	intron(dbSNP)	A	G	+++-++	2.26E-08	0.0779	0.0139	MCHC	0.5959	12264
rs11248850	16	103598	C16orf35	intron(dbSNP)	A	G	++++++	3.75E-08	0.088	0.016	MCHC	0.2354	12330
rs13339636	16	238589	ITFG3	intron(dbSNP)	A	G	+++++	1.87E-34	0.6847	0.0559	MCH	0.8679	4033
rs13335629	16	250381	ITFG3	intron(dbSNP)	A	G	----?	1.77E-31	-0.7159	0.0613	MCH	0.1301	3178
rs13339109	16	202300	LUC7L	intron(dbSNP)	A	C	-----	8.33E-24	-0.7784	0.0774	MCH	0.0654	4051
rs1211375	16	180281	LUC7L	intron(dbSNP)	A	C	-----	1.32E-23	-0.4296	0.0429	MCH	0.2693	4052
rs13336641	16	107408	C16orf35	intron(dbSNP)	T	C	---+-	3.61E-21	-0.6267	0.0664	MCH	0.0956	4055
rs9929571	16	234387	ITFG3	intron(dbSNP)	A	G	+++++	3.28E-20	0.5143	0.0558	MCH	0.8758	4042
rs9940149	16	240642	ITFG3	intron(dbSNP)	A	G	-----	2.66E-19	-0.3641	0.0405	MCH	0.6104	4049
rs7203560	16	124390	C16orf35	intron(dbSNP)	T	G	+++-+	1.89E-17	0.6336	0.0745	MCH	0.9057	4051
rs9926112	16	120719	C16orf35	intron(dbSNP)	A	G	+++-+	3.31E-17	0.6388	0.0757	MCH	0.9079	4054
rs3176398	16	70406	MPG	intron(dbSNP)	C	G	----?	2.87E-16	-0.6562	0.0802	MCH	0.0786	3180
rs2541613	16	102998	C16orf35	intron(dbSNP)	A	G	+++++	2.58E-15	0.3144	0.0397	MCH	0.4973	4054
rs2562181	16	81659	C16orf35	intron(dbSNP)	T	C	-----	5.78E-14	-0.3228	0.043	MCH	0.3413	4001
rs2239739	16	251854	ITFG3	intron(dbSNP)	A	G	+++++	6.19E-13	0.3161	0.0439	MCH	0.4596	4053

rs4141288	16	225314	ITFG3	intron(dbSNP)	T	C	+++++	1.16E-12	0.339	0.047	MCH	0.2187	4054
rs11642609	16	192480	LUC7L	intron(dbSNP)	T	C	---+-	1.59E-12	-0.3265	0.0462	MCH	0.7589	4052
rs2857998	16	125123	C16orf35	intron(dbSNP)	A	G	+++++	3.28E-12	0.3214	0.0461	MCH	0.2502	4055
rs11248914	16	233563	ITFG3	intron(dbSNP)	T	C	-----	2.31E-11	-0.272	0.0407	MCH	0.6255	4052
rs2562189	16	127429	C16orf35	intron(dbSNP)	T	C	---+-	1.59E-10	-0.2884	0.0451	MCH	0.7132	4052
rs407983	16	390760	NME4 (5 downstream)	near-gene-5 (dbSNP)	C	G	+++++	4.50E-10	0.3261	0.0523	MCH	0.2382	4066
rs431324	16	388173	NME4	intron(dbSNP)	A	G	+++-+	6.12E-10	0.2899	0.0469	MCH	0.7341	4055
rs12927713	16	92220	C16orf35	intron(dbSNP)	A	G	+++++	8.38E-10	0.2916	0.0475	MCH	0.2056	4055
rs2238368	16	110328	C16orf35	intron(dbSNP)	T	C	+++-+	4.95E-09	0.2763	0.0472	MCH	0.2226	4055
rs9559892	13	88166665	none	intergenic(GVS)	A	C	-----	5.46E-09	-0.2589	0.0444	MCH	0.2532	4066
rs7192051	16	4482118	HMOX2	near-gene-5 (dbSNP)	T	G	+++++	5.70E-09	0.2396	0.0411	MCH	0.6401	4026
rs7189948	16	350055	AXIN1 (7590 upstream)	intergenic(GVS)	A	G	++++?	6.43E-09	0.4522	0.0779	MCH	0.911	3183
rs11248850	16	103598	C16orf35	intron(dbSNP)	A	G	+++-+	7.24E-09	0.2693	0.0465	MCH	0.2209	4052
rs9386791	6	109715190	C6orf184 (7009 upstream)	intergenic(GVS)	T	C	+++++	1.02E-08	0.2294	0.0401	MCH	0.5835	4066
rs10220154	13	88176841	none	intergenic(GVS)	A	G	-----	1.18E-08	-0.2536	0.0445	MCH	0.244	4044
rs2562182	16	73946	MPG	near-gene-3 (dbSNP)	A	G	+++++	1.26E-08	0.2351	0.0413	MCH	0.3845	4041
rs7197554	16	124242	C16orf35	intron(dbSNP)	A	C	+++-+	1.46E-08	0.3275	0.0578	MCH	0.8481	4049
rs214246	16	289294	AXIN1	intron(dbSNP)	A	G	+++++	1.49E-08	0.3243	0.0573	MCH	0.1624	4030
rs369322	16	376809	none	intergenic(GVS)	A	G	----?	1.61E-08	-0.3487	0.0617	MCH	0.1851	3183
rs6924815	6	109719037	C6orf184 (3162 upstream)	intergenic(GVS)	A	G	-----	1.78E-08	-0.2225	0.0395	MCH	0.4284	4066
rs9386790	6	109715153	C6orf184 (7046 upstream)	intergenic(GVS)	A	C	-----	2.03E-08	-0.2215	0.0395	MCH	0.4283	4063
rs9560016	13	88207654	none	intergenic(GVS)	T	C	+++++	2.09E-08	0.2392	0.0427	MCH	0.7194	4066
rs9301486	13	88183824	none	intergenic(GVS)	A	G	-----	2.09E-08	-0.2526	0.0451	MCH	0.2395	4066
rs1111865	6	109717556	C6orf184 (4643 upstream)	intergenic(GVS)	T	C	+++++	2.35E-08	0.2205	0.0395	MCH	0.5716	4063
rs9400269	6	109712758	C6orf184 (9441 upstream)	intergenic(GVS)	C	G	-----	3.65E-08	-0.2201	0.04	MCH	0.4266	4061
rs9480921	6	109712361	C6orf184 (9838 upstream)	intergenic(GVS)	A	G	+++++	3.70E-08	0.2198	0.0399	MCH	0.5774	4055

rs4601178	6	109717939	C6orf184 (4260 upstream)	intergenic(GVS)	A	G	+++++	3.83E-08	0.2178	0.0396	MCH	0.5718	4063
rs9400271	6	109714249	C6orf184 (7950 upstream)	intergenic(GVS)	A	G	+++++	3.89E-08	0.2179	0.0396	MCH	0.5718	4066
rs9487034	6	109712592	C6orf184 (9607 upstream)	intergenic(GVS)	T	C	-----	4.35E-08	-0.2189	0.04	MCH	0.4264	4063
rs1111866	6	109717539	C6orf184 (4660 upstream)	intergenic(GVS)	T	G	+++++	4.54E-08	0.2166	0.0396	MCH	0.5717	4063
rs13335629	16	250381	ITFG3	intron(dbSNP)	A	G	----?	3.61E-22	-0.648	0.0669	MCV	0.1345	5508
rs13339636	16	238589	ITFG3	intron(dbSNP)	A	G	+++++	4.40E-21	0.6038	0.0641	MCV	0.8628	6405
rs13336641	16	107408	C16orf35	intron(dbSNP)	T	C	---+-	9.37E-20	-0.6847	0.0753	MCV	0.0975	6427
rs1211375	16	180281	LUC7L	intron(dbSNP)	A	C	-----	1.06E-19	-0.4481	0.0493	MCV	0.2695	6389
rs13339109	16	202300	LUC7L	intron(dbSNP)	A	C	-----	6.79E-16	-0.7126	0.0883	MCV	0.0693	6389
rs9929571	16	234387	ITFG3	intron(dbSNP)	A	G	+++++	2.08E-14	0.498	0.0651	MCV	0.8739	6374
rs7203560	16	124390	C16orf35	intron(dbSNP)	T	G	+++-+	4.14E-13	0.609	0.084	MCV	0.9236	6423
rs9926112	16	120719	C16orf35	intron(dbSNP)	A	G	+++-+	4.19E-13	0.6267	0.0864	MCV	0.926	6426
rs3176398	16	70406	MPG	intron(dbSNP)	C	G	----?	4.50E-13	-0.6446	0.089	MCV	0.0798	5517
rs9940149	16	240642	ITFG3	intron(dbSNP)	A	G	-----	7.14E-13	-0.3385	0.0472	MCV	0.6182	6387
rs11642609	16	192480	LUC7L	intron(dbSNP)	T	C	---+-	1.90E-11	-0.3628	0.054	MCV	0.7712	6424
rs2541613	16	102998	C16orf35	intron(dbSNP)	A	G	+++++	3.86E-11	0.3117	0.0472	MCV	0.4855	6426
rs4984681	16	659934	RHOT2	intron(dbSNP)	T	C	++?-+	5.06E-11	1.0629	0.1618	MCV	0.1735	5585
rs4141288	16	225314	ITFG3	intron(dbSNP)	T	C	+++++	1.17E-10	0.3566	0.0553	MCV	0.2202	6426
rs3848368	16	354609	MRPL28 (2788 downstream)	intergenic(GVS)	T	G	++?-+	1.65E-10	1.0458	0.1636	MCV	0.1752	5556
rs7196136	16	875568	LMF1	intron(dbSNP)	T	C	--??-	3.09E-09	-0.9661	0.163	MCV	0.7863	5434
rs7185192	16	85032	C16orf35	intron(UCSC)	A	G	---+-	8.36E-09	-0.7743	0.1344	MCV	0.0271	6427
rs2239739	16	251854	ITFG3	intron(dbSNP)	A	G	+++++	1.44E-08	0.289	0.051	MCV	0.4586	6425
rs2240735	16	3967606	ADCY9	coding-synonymous (dbSNP)	T	C	++?++	1.97E-08	0.9388	0.1672	MCV	0.2639	5630
rs2857998	16	125123	C16orf35	intron(dbSNP)	A	G	+++++	2.54E-08	0.2934	0.0527	MCV	0.2464	6427
rs4984911	16	654156	WDR90	intron(dbSNP)	T	C	++???	3.17E-08	0.977	0.1766	MCV	0.1946	4410
rs2562181	16	81659	C16orf35	intron(dbSNP)	T	C	-----	3.95E-08	-0.2761	0.0503	MCV	0.3413	6373

rs7192051	16	4482118	HMOX2	near-gene-5(dbSNP)	T	G	+++-+	4.83E-08	0.259	0.0475	MCV	0.6367	6343
rs762516	X	153417857	G6PD	intron(dbSNP)	T	C	++?	3.76E-14	1.5768	0.2083	MCV	0.1372	4536
rs1050828	X	153417411	G6PD	coding-nonsynonymous(dbSNP)	T	C	?++	2.52E-11	1.6826	0.2522	MCV	0.1094	2844
rs5987270	X	153384775	FAM3A (2925 downstream)	intergenic(GVS)	T	C	+++	8.88E-10	0.8127	0.1326	MCV	0.3339	5388
rs12014480	X	153820482	F8	intron(dbSNP)	A	G	+++	2.55E-08	0.818	0.1469	MCV	0.2437	5390
rs5987027	X	153667301	MPP1	intron(UCSC)	T	C	+++	2.62E-08	0.8068	0.145	MCV	0.2527	5382
rs1800297	X	153742032	F8	coding-nonsynonymous(dbSNP)	T	C	---	3.84E-08	-0.8175	0.1487	MCV	0.7631	5387
rs13335629	16	250381	ITFG3	intron(dbSNP)	A	G	++++?	7.48E-24	0.1699	0.0169	RBC	0.1202	3928
rs13339636	16	238589	ITFG3	intron(dbSNP)	A	G	-----	1.29E-19	-0.14	0.0154	RBC	0.8625	4785
rs1211375	16	180281	LUC7L	intron(dbSNP)	A	C	+++++	2.59E-13	0.0815	0.0111	RBC	0.2706	4804
rs13339109	16	202300	LUC7L	intron(dbSNP)	A	C	+++++	2.92E-12	0.1465	0.021	RBC	0.0579	4803
rs4141288	16	225314	ITFG3	intron(dbSNP)	T	C	-----	1.29E-10	-0.0819	0.0127	RBC	0.2122	4806
rs2541613	16	102998	C16orf35	intron(dbSNP)	A	G	-----	7.18E-10	-0.0616	0.01	RBC	0.5158	4806
rs11642609	16	192480	LUC7L	intron(dbSNP)	T	C	+++-+	2.91E-09	0.0716	0.0121	RBC	0.7567	4804
rs13336641	16	107408	C16orf35	intron(dbSNP)	T	C	+++-+	5.47E-09	0.1073	0.0184	RBC	0.0929	4807
rs7203560	16	124390	C16orf35	intron(dbSNP)	T	G	---+-	7.41E-09	-0.1212	0.021	RBC	0.9011	4803
rs9929571	16	234387	ITFG3	intron(dbSNP)	A	G	-----	1.30E-08	-0.0827	0.0146	RBC	0.8747	4790
rs2239739	16	251854	ITFG3	intron(dbSNP)	A	G	-----	1.64E-08	-0.0617	0.0109	RBC	0.455	4759
rs9926112	16	120719	C16orf35	intron(dbSNP)	A	G	---+-	2.32E-08	-0.1167	0.0209	RBC	0.9026	4806
rs1050828	X	153417411	G6PD	coding-nonsynonymous(dbSNP)	T	C	---	4.00E-19	-0.1424	0.0159	RBC	0.1082	3742
rs762516	X	153417857	G6PD	intron(dbSNP)	T	C	--?	1.70E-14	-0.1355	0.0177	RBC	0.1363	2917
rs5987027	X	153667301	MPP1	intron(UCSC)	T	C	---	7.00E-12	-0.082	0.012	RBC	0.2488	3760
rs5987026	X	153665315	MPP1	intron(dbSNP)	T	C	++?	2.90E-11	0.0925	0.0139	RBC	0.7191	2922
rs12014480	X	153820482	F8	intron(dbSNP)	A	G	---	7.60E-11	-0.0789	0.0121	RBC	0.2346	3768

rs1800297	X	153742032	F8	coding-nonsynonymous(dbSNP)	T	C	+++	1.90E-10	0.0781	0.0123	RBC	0.7691	3765
rs1734792	X	152994254	MECP2	intron(dbSNP)	A	C	---	9.20E-10	-0.077	0.0126	RBC	0.2358	3769
rs4898348	X	153539588	CTAG2 (4552 upstream)	intergenic(GVS)	T	C	---	5.30E-09	-0.0619	0.0106	RBC	0.4547	3755
rs2734643	X	152944381	MECP2	utr-3(dbSNP)	T	C	---	3.30E-08	-0.0775	0.014	RBC	0.1672	3760
rs10493739	1	83698745	none	intergenic(GVS)	T	C	+??+	3.02E-08	0.0128	0.0023	RDW	0.3337	2781
rs1050828	X	153417411	G6PD	coding-nonsynonymous(dbSNP)	T	C	--	1.70E-11	-0.0326	0.0048	RDW	0.1158	2771

* For intergenic SNPs within 10 kb of gene.

Table S4. Results from the replication of rs9559892 with MCH and rs7192051 with MCV or MCH

	CHOP	Mt. Sinai Affymetrix	Mt. Sinai Illumina Omni	Meta-Analyzed	CHARGE European Americans	RIKEN Japanese
MCH~ rs9559892						
N	7680	378	1634	9692	21020	14088
MAF	0.23	0.23	0.26	0.24	0.04	0.24
Beta (SE)	0.059 (0.042)	-0.27(0.24)	0.057(0.11)	0.050 (0.039)*	-0.0004 (0.0014)	0 (0.001)
P-value	0.17	0.26	0.60	0.20*	0.76	0.54
MCH~ rs7192051						
N	7680	378	1634	9692	21020	NA
MAF	0.36	0.36	0.38	0.36	0.03	NA
Beta (SE)	0.059 (0.038)	-0.28(0.21)	0.12(0.10)	0.056 (0.035)*	0.0011 (0.0016)	NA
P-value	0.12	0.17	0.24	0.11*	0.50	NA
MCV~ rs7192051						
N	7684	378	1634	9696	29646	NA
MAF	0.36	0.36	0.38	0.36	0.03	NA
Beta (SE)	0.033 (0.094)	-0.43(0.53)	0.28(0.26)	0.048 (0.087)*	0.0005 (0.0011)	NA

P-value	0.73	0.42	0.28	0.58*	0.65	NA

*Under fixed effects model

Table S5: Cross-Ethnic Transferability of SNPs Previously Associated with RBC

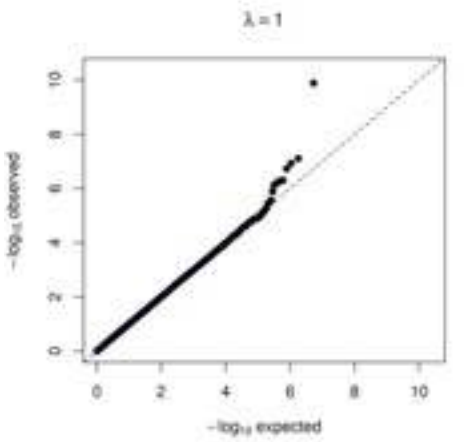
Trait	Marker	Chr:Position	Gene	Function	A1/A2	Freq (A1)	Effect (SE)	Direction	P-Value	Previous Study (Ethnicity*)	Prev_Associated_Trait
HCT	rs1800562	6:26201120	HFE	coding-nonsynonymous (dbSNP)	A/G	0.0151	0.5308 (0.1531)	+-++?++	5.25E-04	Benyamin(C), Ganesh(C)	HB, HCT, MCH, MCV
HB	rs1122794	16:249156	ITFG3	intron(dbSNP)	A/C	0.1965	0.0574 (0.0156)	++++-++	2.32E-04	Ganesh (C)	MCH
HB	rs1800562	6:26201120	HFE	coding-nonsynonymous (dbSNP)	A/G	0.0151	0.2388 (0.0519)	++++?++	4.28E-06	Benyamin(C), Ganesh (C)	HB, HCT, MCH, MCV
MCHC	rs1122794	16:249156	ITFG3	intron(dbSNP)	A/C	0.1984	0.1598 (0.0282)	++++	1.49E-08	Ganesh (C)	MCH
MCHC	rs11966072	6:109741521	C6orf184 (5405 downstream)	intergenic(GVS)	A/G	0.7604	-0.228 (0.0594)	???-	1.24E-04	Kamatani (J)	MCH, MCV, RBC

MCHC	rs837763	16:87381230	None	intergenic(GVS)	T/C	0.5473	-0.0864 (0.0249)	-+--	5.04E-04	Kamatani (J)	MCHC
MCV	rs1122794	16:249156	ITFG3	intron(dbSNP)	A/C	0.1920	0.2396 (0.0603)	+++-+	7.05E-05	Ganesh (C)	MCH
MCV	rs632057	6:139875705	None	intergenic(GVS)	T/G	0.5331	0.1485 (0.0447)	+++++	8.97E-04	Kamatani (J)	MCH, MCV
MCV	rs7189020	16:244804	ITFG3	intron(dbSNP)	A/T	0.1421	0.2997 (0.0693)	+++--	1.54E-05	Ganesh (C)	MCV
MCH	rs1122794	16:249156	ITFG3	intron(dbSNP)	A/C	0.1955	0.2281 (0.0508)	+++-+	7.21E-06	Ganesh (C)	MCH
MCH	rs7189020	16:244804	ITFG3	intron(dbSNP)	A/T	0.1419	0.2606 (0.0591)	++++-	1.04E-05	Ganesh (C)	MCV
MCH	rs9374080	6:109723113	C6orf184	intron(dbSNP)	T/C	0.6427	-0.139 (0.0413)	-----	7.62E-04	Ganesh (C)	MCV
RBC	rs7189020	16:244804	ITFG3	intron(dbSNP)	A/T	0.1365	-0.0573 (0.0154)	----+	1.95E-04	Ganesh (C)	MCV
RBC	rs7775698	6:135460328	None	intergenic(GVS)	T/C	0.1950	-0.0534 (0.0129)	---+-	3.34E-05	Kamatani (J)	HCT, MCH, MCHC, MCV,RBC

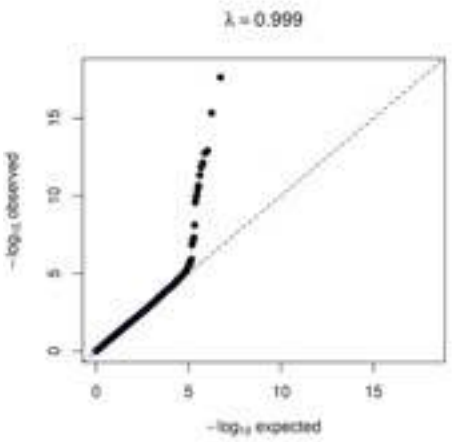
SUPPLEMENTAL FIGURES

Figure S1. QQ plots of GWAS for red cell traits: a) Hematocrit, b) Hemoglobin, c) Mean corpuscular hemoglobin, d) Mean corpuscular hemoglobin concentration, e) Mean corpuscular volume, f) Red blood cell, g) Red blood cell distribution width

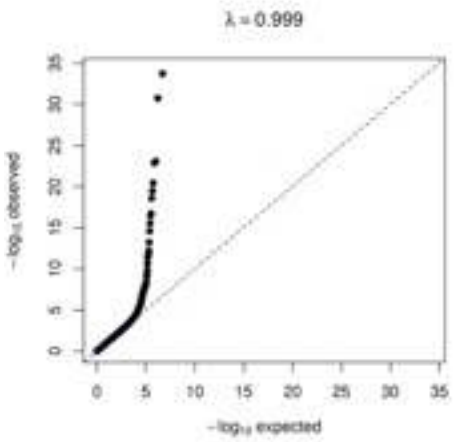
a)



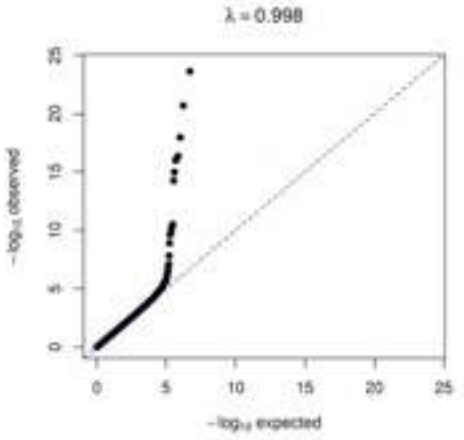
b)



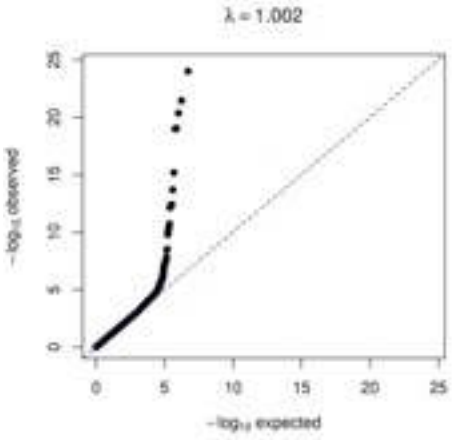
c)



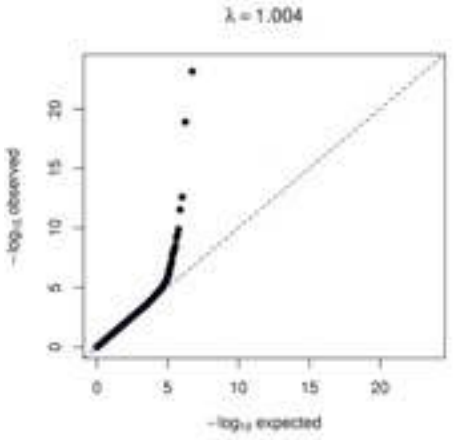
d)



e)



f)



g)

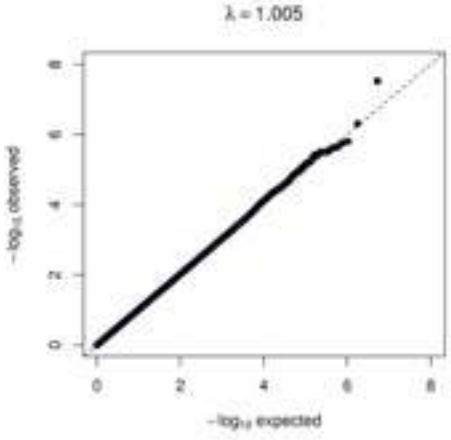

λ = 1.005
−log10 observed
−log10 expected
0
2
4
6
8

Figure S2: MCHC admixture scan in WHI

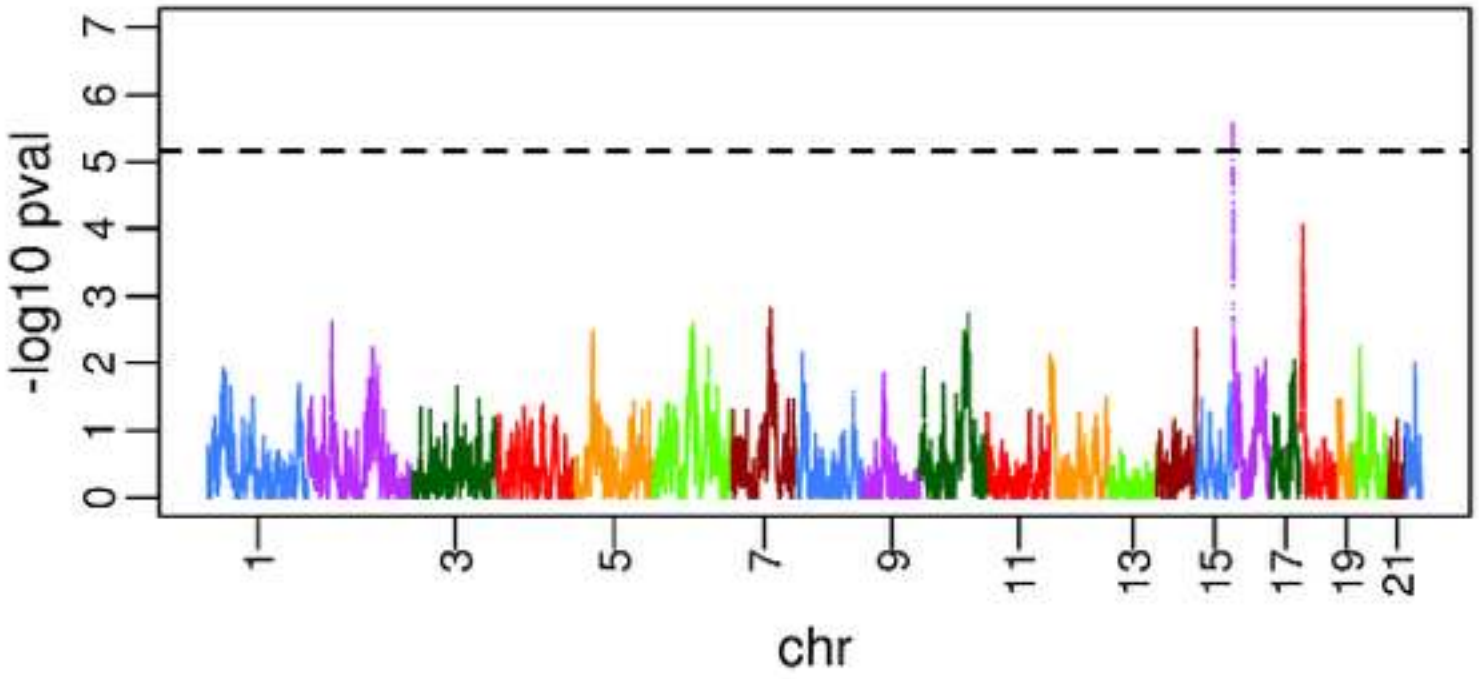


Figure S3: Identification of alpha-globin 3.7 kb deletion CNV in 1000 Genomes data

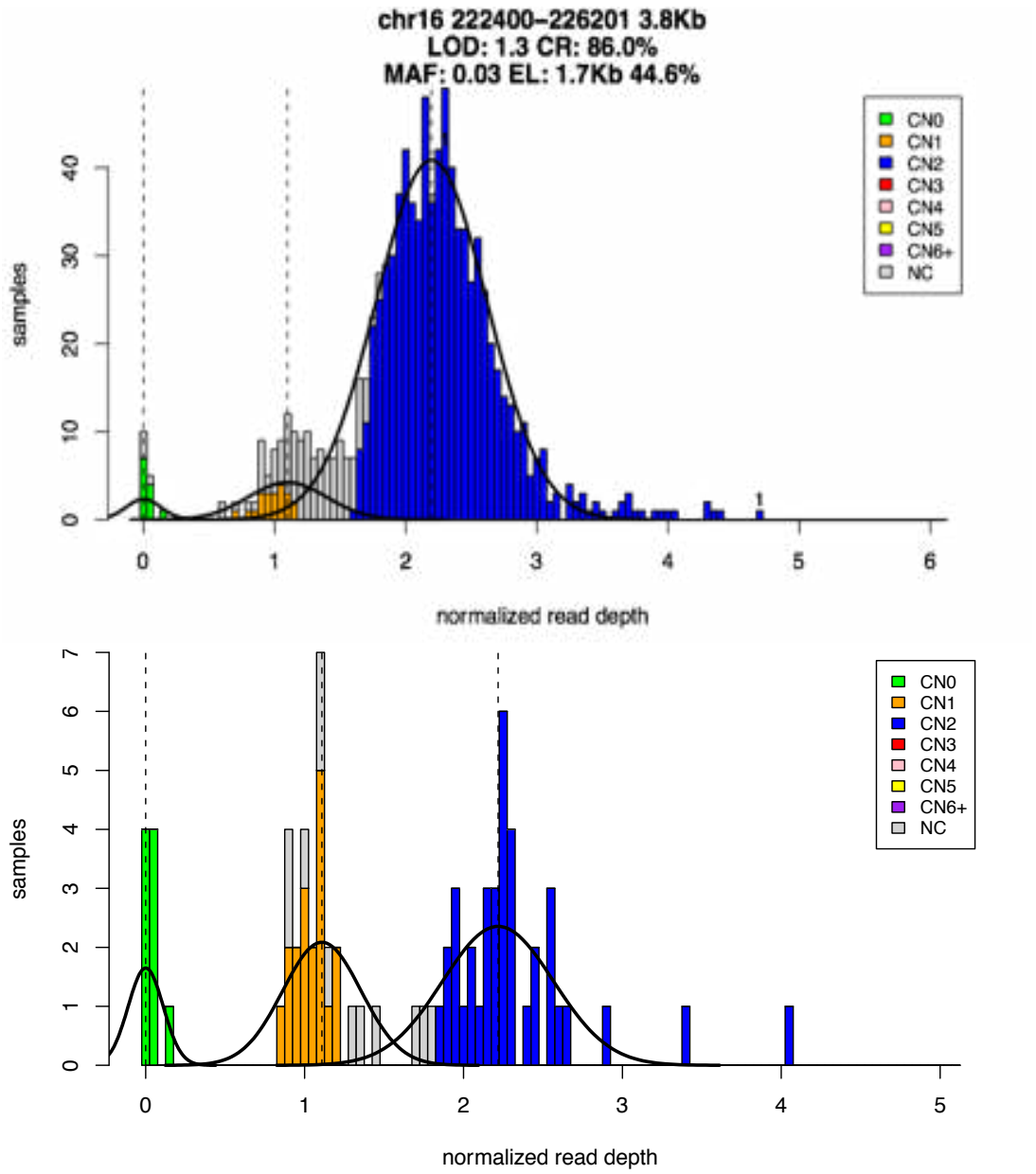
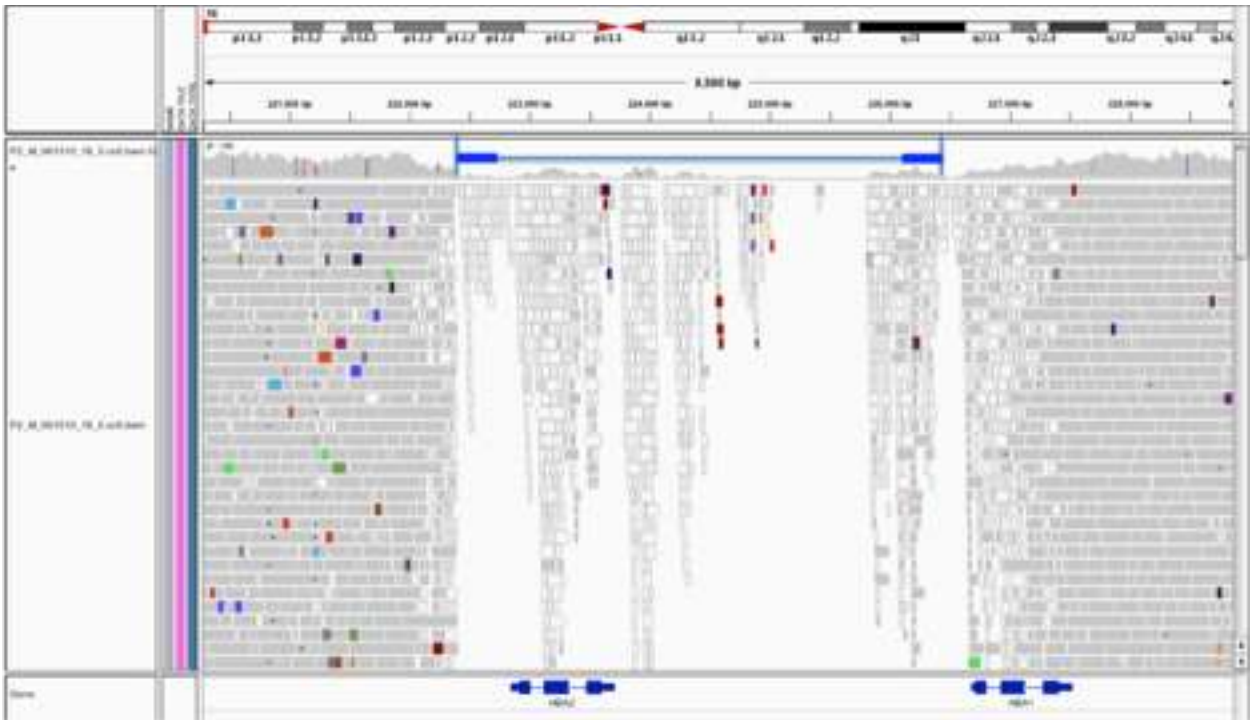
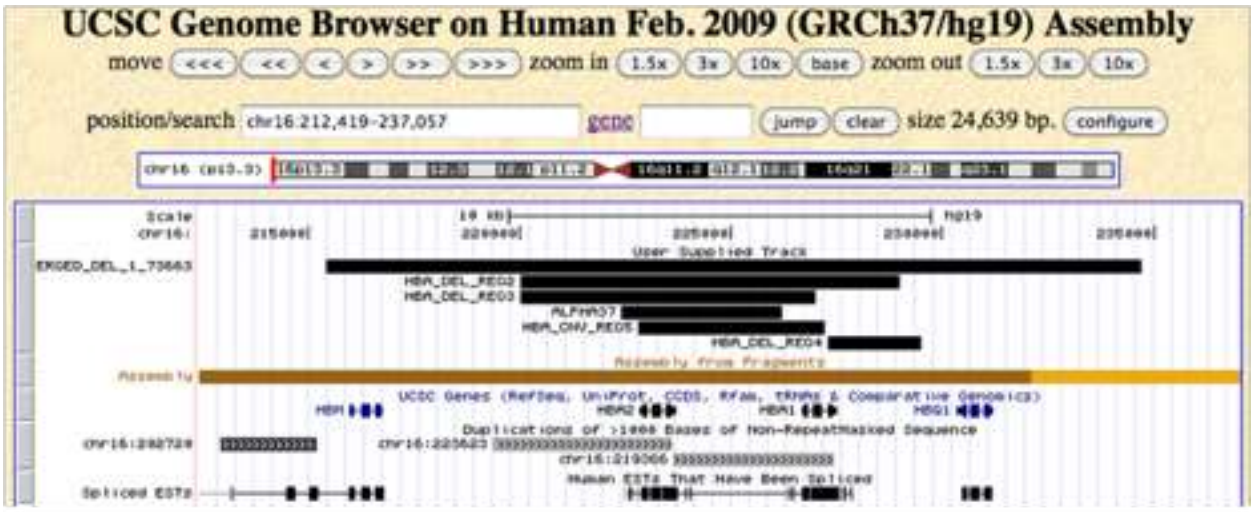


Figure S4: Identification of breakpoints for alpha3.7 deletion



Shown here are 16 pooled samples that appear to be homozygous deleted (9 YRI, 5 LWK, 1 ASW, 1 CLM). The deletion appears to be bounded by ~300bp (thick blue lines) of nearly identical sequence. Light colored reads shown below have low mapping quality and are due to random placement, mis-mapping or sequencing error.

Figure S5: Other probable and possible rare CNV identified in alpha-globin region



More Certain	
Alpha37	Known common deletion of HBA2
MERGED_DEL_1_73663	Rare deletion from 1KG pilot (3 CHS), covers HBAM through HBQ1
MERGED_DEL_1_73647	Very rare deletion from 1KG pilot (chr16:154000-160000, not shown above) covers potential regulatory element MCS-R1
More Speculative	
HBA_DEL_REG2	Possible deletion in 8 LWK samples
HBA_DEL_REG3	Possible deletion in 4 samples (3 ASW, 1 YRI)
HBA_DEL_REG4	Speculative deletion in 3 JPT samples
HBA_CNV_REG5	Speculative duplication of HBA in 6 samples (various populations).

Figure S6: Known alpha-globin gene regulatory regions and possible identification of very rare deletion spanning MCS-R1

Region	Est. Location
MCS-R1	155086-155154
MCS-R2	163508-163751 or 163669-163751
MCS-R3	Unknown (around 165000 – 200000)
MCS-R4	Unknown (around 165000 – 200000)