Genetic Loci Implicated in Erythroid Differentiation and Cell Cycle Regulation Are Associated With Red Blood Cell Traits

Keyue Ding, PhD; Khader Shameer, PhD; Hayan Jouni, MD; Daniel R. Masys, MD; Gail P. Jarvik, MD, PhD; Abel N. Kho, MD; Marylyn D. Ritchie, PhD; Catherine A. McCarty, PhD; Christopher G. Chute, MD, DrPH; Teri A. Manolio, MD, PhD; and Iftikhar J. Kullo, MD

Abstract

Objective: To identify common genetic variants influencing red blood cell (RBC) traits.

Patients and Methods: We performed a genomewide association study from June 2008 through July 2011 of hemoglobin, hematocrit, RBC count, mean corpuscular volume, mean corpuscular hemoglobin, and mean corpuscular hemoglobin concentration in 12,486 patients of European ancestry from the electronic MEDical Records and Genomics (eMERGE) network. We developed an electronic medical record–based algorithm that included individuals who had RBC measurements obtained for clinical care and excluded values measured in the setting of hematopoietic disorders, comorbid conditions, or medications known to affect RBC production or a recent history of blood loss.

Results: We identified 4 new genetic loci and replicated 11 loci previously reported to be associated with one or more RBC traits in individuals of European ancestry. Notably, genes present in 3 of the 4 newly identified loci (*THRB*, *PTPLAD1*, *CDT1*) and in 6 of the 11 replicated loci (*KLF1*, *ALDH8A1*, *CCND3*, *SPTA1*, *FBXO7*, *TFR2/EPO*) are implicated in erythroid differentiation and regulation of cell cycle in hematopoietic stem cells.

Conclusion: Genes in the erythroid differentiation and cell cycle regulation pathways influence interindividual variation in RBC indices. Our results provide insights into the molecular basis underlying variation in RBC traits.

© 2012 Mayo Foundation for Medical Education and Research ■ Mayo Clin Proc. 2012;87(5):461–474

The red blood cell (RBC) acts as a container for hemoglobin (HGB) and thereby assumes the critical function of oxygen transport and delivery from the lungs to tissues throughout the body to support aerobic metabolism.¹ The RBC membrane and metabolic machinery facilitate oxygen transport and delivery function of HGB. Disorders involving RBCs, such as anemia, are common and associated with adverse health outcomes, including cardiovascular events.² The RBC traits have a substantial genetic component, with heritabilities of 0.56, 0.52, and 0.52 reported for RBC count, mean corpuscular volume (MCV), and mean corpuscular hemoglobin (MCH), respectively.³ Recently, genomewide association studies (GWAS) in cohorts of European ancestry,^{4–7} as well as in a Japanese cohort,⁸ have reported multiple quantitative trait loci associated with one or more RBC traits.

The electronic MEDical Records and GENomics (eMERGE) network (www.gwas.net) was established to develop and implement approaches for leveraging biorepositories with electronic medical record (EMR) systems for large-scale

genomic research, including but not limited to GWAS, sequencing, and structural variation.⁹ The 5 participating sites include Group Health Cooperative–University of Washington (GHC), Marshfield Clinic (MC), Mayo Clinic, Northwestern University (NU), and Vanderbilt University Medical Center (VUMC). Each site is conducting GWAS of multiple phenotypes, with Mayo Clinic investigating genetic loci associated with peripheral arterial disease and RBC traits including HGB, hematocrit (HCT), RBC count, MCV, MCH, and mean corpuscular hemoglobin concentration (MCHC).

Using the Mayo eMERGE cohort (n=3336), we demonstrated feasibility of EMR-based genetic studies to replicate genetic loci previously associated with RBC traits.¹⁰ In this article, we report results of an analysis of 12,486 patients in the eMERGE network, performed to identify additional genetic loci influencing interindividual variation in RBC traits. For functional annotation of significant loci, we used a database of uniquely mapped single-nucleotide polymorphisms (SNPs), as well as gene expression and epigenetic data.

From the Division of Cardiovascular Diseases (K.D., K.S., H.J., I.J.K.) and Division of Biomedical Statistics and Informatics (C.G.C.), Mayo Clinic, Rochester, MN; Division of Biomedical and Health Informatics, Department of Medical Education and Biomedical Informatics (D.R.M.), and Department of Medicine (Medical Genetics) and Department of Genome Sciences (G.P.J.), University of Washington, Seattle; Department of Medicine, Northwestern University, Chicago, IL (A.N.K.); Department of Biochemistry and Molecular Biology, Pennsylvania State University, University Park (M.D.R.); Essentia Institute of Rural Health, Duluth, MN (C.A.M.); and the Office of Population Genomics, National Human Genome Research Institute, Bethesda, MD (T.A.M.).

METHODS

Study Design of GWAS for RBC Traits in the eMERGE Consortium

We performed a GWAS for 6 RBC traits in 12,486 patients of European ancestry identified from 5 sites in the eMERGE network (Table 1). Four of these traits (HGB, HCT, RBC count, and MCV) are measured directly by standard methods established in clinical laboratories, whereas MCH and MCHC are derived (ie, $MCH = HGB \times 10/RBC$ count and $MCHC = HGB/HCT$). We used an algorithm to extract RBC traits from the EMR, excluding RBC values affected by comorbid conditions, medications, and blood loss.^{10,11} The algorithm was exported to the remaining eMERGE sites to extract RBC trait values from the EMR. We randomly split the 5 cohorts into a discovery cohort (Mayo + GHC + VUMC, n=7873) and a replication cohort (MC + NU, n=4613). We performed tests of association within each of the 5 sites and the discovery and replication cohorts. Finally, we performed a joint analysis of the combined cohort that included all 5 sites.

Genotyping and Quality Control

Genotyping was performed at the Center for Genotyping and Analysis at the Broad Institute (Mayo, VUMC, and NU) and the Center for Inherited Disease Research at the Johns Hopkins University (MC and GHC) using the Illumina Human660W-Quad v1_A genotyping platform (Illumina, San Diego, CA), consisting of 561,490 SNPs and 95,876 intensity-only probes. Data were cleaned using the quality control pipeline developed by the eMERGE

Genomics Working Group.¹² The process includes evaluation of sample and marker call rate, sex mismatch and anomalies, duplicate and HapMap concordance, batch effects, Hardy-Weinberg equilibrium, sample relatedness, and population stratification. A total of 476,395 SNPs were available for analysis after applying the following quality control criteria: SNP call rate greater than 98%, sample call rate greater than 98%, minor allele frequency greater than 0.05, Hardy-Weinberg equilibrium $P > .001$, and 99.99% concordance rate in duplicates. The first 2 dimensions were derived from multidimensional decomposition analysis (using the `cmdscale` function in R) of the 1-IBS (identity-by-state) matrix based on all eMERGE samples (n=17,358). Samples greater than 6 standard error from the mean of self-reported white ethnicity on dimensions 1 and 2 were excluded. After the quality control steps, 12,486 patients (ie, patients with genetically defined European ancestry) with phenotype and genotype data were available for association analyses.

Statistical Analyses

When multiple measurements of RBC traits were available for an individual patient, we chose the median value and the corresponding age for the genetic analyses. We used an efficient mixed-model association-expedited (EMMAX) algorithm¹³ to correct for sample relatedness and cryptic population substructure. The IBS matrix was calculated for each pair of individuals using the genome-wide genotype data. The generalized least-squares F test was used to estimate the regression coefficient (β) and per-

TABLE 1. Sample Characteristics^{a,b}

Characteristic ^c	Mayo	GHC	MC	VUMC	NU	Total
No. of patients	3021	2266	3648	2586	965	12,486
Women, No. (%)	1148 (38.0)	1324 (58.4)	2153 (59.0)	1527 (59.0)	499 (51.7)	6651 (53.3)
Age, (y) ^d	62±9	77±8	64±12	51±17	52±13	63±15
HGB, (g/dL)	14.06±1.29	13.61±1.23	14.03±1.17	13.72±1.48	13.58±1.56	13.86±1.32
HCT, (%)	40.95±3.64	40.10±3.52	40.92±3.26	41.31±3.81	39.70±4.40	40.76±3.65
RBC count, ($\times 10^{12}/L$)	4.53±0.43	4.40±0.41	4.56±0.41	4.61±0.47	4.45±0.51	4.52±0.44
MCV, (fL)	90.53±4.17	91.56±4.49	89.97±4.11	89.68±4.92	89.55±4.87	90.31±4.48
MCH, (pg)	31.09±1.59	31.05±1.70	30.84±1.55	29.81±1.93	30.63±1.89	30.72±1.76
MCHC, %	34.33±0.50	33.91±0.62	34.25±0.52	33.25±1.07	34.20±0.64	34.01±0.79

^aGHC = Group Health Cooperative; HCT = hematocrit; HGB = hemoglobin; MC = Marshfield Clinic; MCH = mean corpuscular hemoglobin; MCHC = mean corpuscular hemoglobin concentration; MCV = mean corpuscular volume; NU = Northwestern University; RBC = red blood cell; VUMC = Vanderbilt University Medical Center.

^bValues are presented as mean $\pm$ SD unless indicated otherwise.

^cTo convert HGB to g/L, multiply by 10; to convert HCT to a proportion of 1.0, multiply by 0.01.

^dThe age is the age for median HGB level. For the remaining traits, the age was within ± 1 year of the age at median HGB level. In the association analyses for each trait, we used the corresponding median age for an individual.

form association analyses, which were implemented in the EMMAX package, with adjustment for age, sex, and site. The variance components for each RBC trait, as well as the pseudoheritability¹³ (ie, the fraction of phenotypic variance explained by the empirically estimated relatedness matrix), were estimated from the kinship matrix and residual errors (from the IBS matrix) (Supplemental Table 1, available online at <http://www.mayoclinicproceedings.org>). The statistical power of our study was 80% to detect a quantitative trait locus that explained 0.32% variance in an RBC trait, given a sample size of 12,486, and the significance level of 5×10^{-8} . Regional plots of genomic loci associated with RBC traits were plotted using LocusZoom.¹⁴

We imputed all ungenotyped SNPs in 9 chromosomes harboring loci associated with RBC traits, based on HapMap II CEU database (release 21). Imputation-based association for ungenotyped SNPs was performed using the same IBS matrix for the genotyped SNPs by EMMAX. We tested the association ($r^2 > 0.3$) between the imputed SNPs ($P < 5 \times 10^{-8}$) and the most significant (or nonsynonymous) SNP in each locus.

Patterns of Linkage Disequilibrium in the Candidate Regions and Selection of Candidate Genes

The patterns of linkage disequilibrium (LD) in the chromosome regions were illustrated using LocusZoom.¹⁴ The SNPs near the most significant or the nonsynonymous SNP (ie, the reference SNP) were color coded to reflect their LD with this SNP. Pairwise r^2 values were taken from the HapMap CEU database. We assumed that 2 genes are independently affecting the trait if $r^2 < 0.3$ between the reference SNP and a SNP in the other gene, recognizing that there is no standard threshold of r^2 to declare a good correlation. We also used the conditional haplotype test implemented in PLINK¹⁵ to test whether SNPs have independent effects. In a region of LD with multiple genes, we selected the most likely candidate genes based on their function (ie, whether the genes are involved in erythrocyte differentiation, regulation of cell cycle, or iron metabolism).

Functional Annotation of Significant Loci

We created a pipeline to perform functional annotation of SNPs associated with the RBC traits (eFigure 1).¹⁶ First, we performed association analyses of all variants within the region of LD based on imputation of ungenotyped SNPs from the HapMap II CEU database.¹⁷ The imputation was implemented using MACH1.¹⁸ Functional characterization of significant SNPs was performed using the uniquely mapped SNP database (<http://research.nhgri.nih.gov/tools/unisnp/>). We then explored associations between

statistically significant SNPs (genotyped or imputed) and the expression of genes (ie, expression quantitative trait locus [eQTL] analysis) in the HapMap CEU samples by using the SCAN database.¹⁹ Finally, we performed an epigenetic analysis of human embryonic (methylation analysis) and hematopoietic stem cells and erythroid progenitors (histone signatures analysis) in the regions of the significant SNPs (eAppendix 1).

RESULTS

The characteristics of 12,486 patients of European ancestry in the eMERGE cohort are listed in Table 1. We identified 15 chromosomal regions associated with at least one RBC trait in the combined cohort ($P < 5 \times 10^{-8}$) (Figure 1 and Table 2). A list of 142 genotyped SNPs (72 unique) associated with RBC traits, including P values within each individual site and the discovery, replication, and combined cohorts, is in Supplemental Table 2 and Supplemental Table 3, available online at <http://www.mayoclinicproceedings.org>, respectively. We present our GWAS results by chromosome regions because a given chromosome region may contain multiple independent loci (defined based on the pattern of LD in the HapMap CEU population, $r^2 < 0.3$) that contribute to interindividual variation in RBC traits.

New Loci Associated With RBC Traits in Individuals of European Ancestry

In the discovery cohort, 32 new loci not previously reported in individuals of European ancestry were associated with at least 1 RBC trait at a significance level of $P \leq 1 \times 10^{-5}$, 4 each with HGB, and MCHC, 12 each with MCV and MCH, and 3 each with HCT and RBC count. Of these, 3 loci were replicated to be associated with RBC traits after correction for multiple testing in the replication cohort and 4 loci achieved $P < 5 \times 10^{-8}$ in the combined cohort analyses (Table 3). All the lead SNPs at these loci associated with RBC traits had similar effect sizes and direction within each eMERGE study site. The 4 new loci on chromosomes 3p24.2, 15q22.3, 16q24.2, and 16q24.3 associated with RBC traits are summarized in Table 3, and the regional plots of these loci are shown in Figure 2.

Chromosome 3p24.2. An intronic SNP (rs9310736) in the thyroid hormone receptor β -gene (*THRB*) (Mendelian Inheritance of Man [MIM] 190160) was associated with MCV ($\beta = .35$, $P = 6 \times 10^{-9}$). The SNP was associated with MCV and MCH in a Japanese cohort.⁸ *THRB* is a nuclear hormone receptor for triiodothyronine (T_3) and mediates biological activities of the thyroid hormone.²⁰

FIGURE 1. Manhattan plots for genomewide association analysis of red blood cell (RBC) traits. The vertical axis indicates ($-\log_{10}$ transformed) observed P values, and the horizontal line indicates the genomewide significance level of $P < 5E-08$. HCT = hematocrit; HGB = hemoglobin; MCH = mean corpuscular hemoglobin; MCHC = mean corpuscular hemoglobin concentration; MCV = mean corpuscular volume.

TABLE 2. Gene Loci Associated With RBC Traits in the Present and Previous GWAS^a

Locus	Genes ^b	Present study						Previous GWAS						Population	
		HGB	HCT	RBC	MCV	MCH	MCHC	HGB	HCT	RBC	MCV	MCH	MCHC	E	J
1q23.1	<i>SPTA1</i>	—	—	—	—	—	2E-08	—	—	—	—	—	1E-10 ^c	+	—
3p24.2	<i>THRB</i>	—	—	—	6E-09	—	—	—	—	—	3E-08 ^d	4E-10 ^d	—	—	+
4q12	—	—	—	1E-08	3E-12	3E-11	—	—	—	2E-17 ^d	1E-15 ^c	3E-25 ^d	—	+	+
6p22.2~1	<i>TRIM38</i>	—	—	—	1E-09	6E-12	5E-08	^e	^e	^e	^e	^e	^e	+	—
	<i>HFE</i>	2E-09	—	—	2E-21	2E-27	3E-10	6E-19 ^c	2E-09 ^c	—	1E-46 ^c	—	—	+	—
	<i>HIST1H2AC</i>	—	—	—	2E-14	1E-19	4E-10	—	—	—	—	—	—	—	—
6p21.1	<i>CCND3</i>	—	—	—	1E-18	6E-12	—	—	—	1E-10 ^d	1E-31 ^c	8E-20 ^c	—	+	+
6q23.3	<i>HBS1L/MYB</i>	—	1E-08	3E-31	5E-34	5E-38	3E-10	—	3E-15 ^c	1E-47 ^c	7E-86 ^c	7E-69 ^c	7E-14 ^c	+	+
	<i>ALDH8A1</i>	—	—	8E-09	5E-09	5E-09	—	^e	^e	^e	^e	^e	^e	+	—
6q24.1	<i>LOC645434</i>	—	—	—	1E-08	—	—	—	—	—	5E-25 ^c	1E-17 ^c	—	+	—
7q22.1	<i>TFR2</i>	—	—	—	2E-09	2E-08	—	—	4E-10 ^c	5E-10 ^f	3E-11 ^c	—	—	+	—
15q22.3	<i>PTPLAD1, C15orf44</i>	—	—	—	—	9E-09	—	—	—	—	—	3E-09 ^d	—	—	+
16q24.2	—	—	—	—	—	—	2E-08	—	—	—	—	—	—	—	—
16q24.3	<i>CDT1</i>	—	—	—	—	—	2E-08	—	—	—	—	—	4E-13 ^d	—	+
19p13.3	<i>KLF1</i>	—	—	—	—	2E-09	—	—	—	—	—	1E-11 ^c	—	+	—
22q12.3	<i>FBXO7</i>	—	—	—	3E-10	2E-08	—	—	—	—	4E-10 ^f	—	—	+	—
22q12.3	<i>MPST</i>	—	—	—	1E-10	8E-11	—	—	—	—	—	—	—	—	—
	<i>KCTD17/TMPRSS6</i>	5E-11	—	—	2E-31	9E-41	5E-11	3E-25 ^c	—	—	2E-41 ^c	8E-34 ^c	8E-14 ^d	+	+
22q13.3	<i>TYMP</i>	—	—	—	2E-08	—	—	—	—	—	1E-15 ^c	4E-08 ^d	—	+	+

^a The 4 novel genomic loci associated with RBC traits in patients of European ancestry identified in the present study are shown in bold. E = European; GWAS = genomewide association study; HCT = hematocrit; HGB = hemoglobin; J = Japanese; MCH = mean corpuscular hemoglobin; MCHC = mean corpuscular hemoglobin concentration; MCV = mean corpuscular volume; RBC = red blood cell. "—" = the locus was not identified in the present or previous study (in European or Japanese population); "+" = the locus was identified in previous study (in European or Japanese population).

^b On a given chromosome locus, we tabulated genes separately if they were located in different linkage disequilibrium blocks (see the "Methods" section).

^c P value is from Ganesh et al.⁴

^d P value is from Kamatani et al.⁸

^e P values were not available.

^f P value is from Soranzo et al.⁵

TABLE 3. Newly Discovered Loci (for Individuals of European Ancestry) Associated With at Least One of the RBC Traits^a

Trait	Lead SNP	Chromosome	Gene	MAF	β^b	P value							
						Mayo (n = 3021)	GHC (n = 2266)	VUMC (n = 2586)	Discovery ^c (n = 7873)	MC (n = 3648)	NUJ (n = 965)	Replication ^d (n = 4613)	Combined (n = 12,486)
MCV	rs9310736	3p24.2	THRB	0.34 (A)	0.35 (G)	3E-03	6E-02	7E-03	1E-05	8E-05	8E-02	1E-05	6E-09
MCH	rs8035639	15q22.3	PTPLAD1, C15orf44	0.42 (T)	0.13 (C)	4E-02	1E-02	9E-05	2E-06	2E-02	5E-02	3E-03	8E-09
MCHC	rs9937239	16q24.2	-	0.25 (A)	0.06 (G)	6E-05	5E-02	3E-04	9E-08	8E-01	6E-02	2E-01	2E-08
MCHC	rs837763	16q24.3	CDT1	0.43 (G)	-0.06 (A)	3E-04	1E-01	1E-06	5E-10	8E-05	4E-01	2E-04	2E-08

^aGHC = Group Health Cooperative; MAF = minor allele frequency (in the combined cohort); MC = Marshfield Clinic; MCHC = mean corpuscular hemoglobin concentration; MCV = mean corpuscular volume; RBC = red blood cell; VUMC = Vanderbilt University Medical Center; NUJ = Northwestern University; SNP = single-nucleotide polymorphism. The physical location for each SNP is shown in Table S2.

^bThe β is the effect of the major allele in the combined cohort (based on the efficient mixed-model association-expedited [EMMAX] analysis).

^cDiscovery cohort = Mayo, GHC, and VUMC.

^dReplication cohort = MC and NUJ.

^aGHC = Group Health Cooperative; MAF = minor allele frequency (in the combined cohort); MC = Marshfield Clinic; MCHC = mean corpuscular hemoglobin concentration; MCV = mean corpuscular volume; RBC = red blood cell; VUMC = Vanderbilt University Medical Center; NU = Northwestern University; SNP = single-nucleotide polymorphism. The physical location for each SNP is shown in Table S2.

^bThe β is the effect of the major allele in the combined cohort (based on the efficient mixed-model association-expedited [EMMAX] analysis).

^cDiscovery cohort = Mayo, GHC, and VUMC.

^dReplication cohort = MC and NU.

Chromosome 15q22.3. At this locus, 4 SNPs in an LD block of 328 kb were associated with MCH. The region contains several candidate genes: dipeptidyl-peptidase 8 (*DPP8*), protein tyrosine phosphatase-like A domain containing 1 (*PTPLAD1*), solute carrier family 24 (sodium/potassium/calcium exchanger), member 1 (*SLC24A1*), chromosome 15 open reading frame (*C15orf44*), and DENN/MADD domain containing 4A (*DENND4A*). One of the SNPs (rs352476, $\beta=0.13$, $P=9\times 10^{-9}$) is located in the 5' regulatory region of *PTPLAD1*, a component of Rac1-signalling pathway necessary for erythropoiesis in the bone marrow.²¹ Another candidate gene, *C15orf44*, is differentially expressed in hematopoietic stem cells and erythrocyte progenitor cells.²² An intronic SNP (rs6494537) in *DENND4A* was associated with MCH in the Japanese cohort.⁸

Chromosome 16q24.2. An intergenic SNP (rs9937239) in this region was associated with MCHC ($\beta=0.06$, $P=2\times 10^{-8}$). No known gene is present within ± 60 kb of this SNP, although transcripts have been detected by RNA sequencing. Two major CCCTC-binding factor binding sites flank rs9937239 and could act in *trans* to confer enhancer-blocking insulator activity.²³

Chromosome 16q24.3. A SNP (rs837763) in the 5' flanking region of the chromatin licensing and DNA replication factor 1 gene (*CDT1*) was associated with MCHC ($\beta=-0.06$, $P=2\times 10^{-8}$), and the same SNP was associated with MCHC in a Japanese cohort.⁸ *CDT1* cooperates with *CDK6* (cell division cycle 6 homolog) to promote loading of the minichromosome maintenance complex onto chromatin to form the pre-replication complex needed to initiate DNA replication.²⁴

Replicated Loci

Eleven loci previously associated with RBC traits in individuals of European ancestry were replicated in the joint analysis (Table 4). Regional plots for these loci are shown in eFigure 2. We provide a detailed description of GWAS results and LD analyses at these loci. On the basis of LD analysis and conditional haplotype test, we identified additional significant signals, independent of previously reported genes, at 3 chromosome loci: chr6p22.2 (1 gene [*HIST1H2AC*] independent of *HFE*) (eFigure 2c), chr6q23.3 (1 gene [*ALDH8A1*] independent of *HBS1L/MYB*) (eFigure 2e), and chr22q12.3 (1 gene [*MPST*] independent of *TMPRSS6*) (eFigure 2j).

Chromosome 1q23.1. A nonsynonymous SNP (rs857725, Lys1693Gln) and a SNP in the 3' flank-

FIGURE 2. Regional plots of 4 novel loci associated with red blood cell (RBC) traits on chromosomes 3p24.2 (mean corpuscular volume [MCV]), 15q22.3 (mean corpuscular hemoglobin [MCH]), 16q24.2 (mean corpuscular hemoglobin concentration [MCHC]), and 16q24.3 (MCHC). Single-nucleotide polymorphisms (SNPs) are plotted by position on the chromosome (x-axis) vs association with RBC traits ($-\log_{10} P$ value) on the y-axis. The rs number for the most significant SNP in the joint analysis is shown on the plot. Estimated recombination rates (from HapMap) are plotted in cyan to reflect the local linkage disequilibrium (LD) structure. The SNPs near the most significant SNP are color coded to reflect their LD with this SNP (taken from pairwise r^2 values from the HapMap CEU data). Genes, the position of exons and the direction of transcription from the University of California, Santa Cruz genome browser are also plotted. $\blacktriangledown$ = nonsynonymous; $\bullet$ = no annotation; $\boxtimes$ = conserved in mammals; $*$ = conserved TFBS (transcription factor binding site). cM/Mb = centimorgan/megabase; Mb = megabase.

ing region of the spectrin gene (*SPTA1*) was associated with MCHC (rs857725, $\beta=0.07$, $P=2 \times 10^{-8}$). An intronic SNP (rs857721) in *SPTA1* was associated with MCHC in a prior GWAS.⁴ Spectrin is an actin cross-linking and molecular scaffold protein that links the plasma membrane to the actin cytoskeleton²⁶ and determines cell shape, arrangement of transmembrane proteins, and organization of organelles. Mutations in *SPTA1* result in a variety of hereditary RBC disorders, including hemolytic anemia, elliptocytosis, spherocytosis, and pyropoikilocytosis.²⁷ Exon 1 and intron 1 of *SPTA1* have been proposed as candidate regions for mutations in patients with spectrin-linked hemolytic anemia.²⁸

Chromosome 4q12. An intergenic SNP (rs218237) in this region was associated with RBC count ($\beta=0.04$, $P=1 \times 10^{-8}$), MCV ($\beta=-0.58$, $P=3 \times 10^{-12}$), and MCH ($\beta=-0.22$, $P=3 \times 10^{-11}$). The SNP was associ-

ated with RBC count and MCH in a Japanese cohort.⁸ Another SNP (rs172629, 13.5 kb downstream of rs218237) at this locus was associated with MCV in a prior GWAS.⁴ The rs218237 is 230 kb downstream of the platelet-derived growth factor receptor α polypeptide gene (*PDGFRA*) and 130 kb upstream of the human homolog of the proto-oncogene c-kit gene (*KIT*). The former is an attractive regional candidate gene for hematologic traits.

Chromosome 6p22.2-6p22.1. We found multiple SNPs within this locus to be associated with several RBC traits. This genomic region spans approximately 2.4 MB and contains 66 reference genes (eFigure 2c). The nonsynonymous SNP rs1800562 in *HFE* is known to be associated with HGB,^{4,5} HCT,⁴ and MCV.^{4,5} In the present study, the SNP was associated with HGB ($\beta=-0.19$,

TABLE 4. Replication of Loci Previously Reported to Be Associated With RBC Traits^a

Trait	Lead SNP	Chromosome	Gene	eMERGE			β for previous studies and reference ^c
				MAF	β^b	P	
MCHC	rs857725	1q23.1	<i>SPTA1</i>	0.26 (G)	0.07	2E-08	-0.002 (rs857721, A) ⁴
MCV	rs218237	4q12	—	0.13 (T)	0.58	3E-12	-0.005 (rs172629, G) ⁴
HGB	rs1800562	6p22.2	<i>HFE</i>	0.06 (A)	-0.19	2E-09	0.16 ⁴
MCV	rs1800562	6p22.2	<i>HFE</i>	0.06 (A)	-1.17	2E-21	0.012 ⁴ ; 1.41 ⁵
MCH	rs1800562	6p22.2	<i>HFE</i>	0.06 (A)	-0.52	2E-27	-0.32 (rs1408272, T) ²⁵
MCV	rs3218097	6p21.1	<i>CCND3</i>	0.25 (T)	0.58	1E-18	-0.005 (rs9349205, A), ⁴ 0.575 (rs11970772, A) ⁵
MCH	rs3218097	6p21.1	<i>CCND3</i>	0.25 (T)	0.18	6E-12	-0.005 (rs9349205, A) ⁴
HCT	rs7775698	6q22.3	<i>HBS1L, MYB</i>	0.26 (T)	0.27	1E-08	0.22 (rs9483788, C) ⁴
RBC count	rs7775698	6q23.3	<i>HBS1L, MYB</i>	0.26 (T)	0.07	3E-31	0.014 (rs9483788, G) ⁴
MCV	rs7775698	6q23.3	<i>HBS1L, MYB</i>	0.26 (T)	-0.81	5E-34	-0.008 (rs4895441, G), ⁴ 0.818 (rs9402686, A) ⁵
MCH	rs7775698	6q23.3	<i>HBS1L, MYB</i>	0.26 (T)	-0.33	5E-38	-0.01 (rs7776054, G) ⁴
MCHC	rs7775698	6q23.3	<i>HBS1L, MYB</i>	0.26 (T)	-0.07	4E-09	-0.002 (rs9373124, C) ⁴
MCV	rs668459	6q24.1	<i>LOC645434</i>	0.42 (C)	-0.33	1E-08	-0.004 (rs643381, A) ⁴
MCV	rs4434553	7q22.1	<i>TFR2</i>	0.49 (G)	-0.34	2E-09	0.003 (rs7786877, G) ⁴
MCH	rs2293683	19p13.3	<i>KLF1</i>	0.36 (A)	-0.14	2E-09	-0.004 (rs11085824, G) ⁴
MCV	rs5994574	22q12.3	<i>FBXO7</i>	0.34 (C)	0.38	3E-10	0.372 (rs9609565, G) ⁵
HGB	rs855791	22q12.3	<i>TMPRSS6</i>	0.44 (T)	0.10	5E-11	-0.09 ⁴
MCV	rs855791	22q12.3	<i>TMPRSS6</i>	0.44 (T)	0.66	2E-31	-0.005 (rs2413450, T) ⁴
MCH	rs855791	22q12.3	<i>TMPRSS6</i>	0.44 (T)	0.30	9E-41	-0.006 (rs2413450, T) ⁴
MCV	rs140522	22q13.3	<i>TYMP</i>	0.33 (A)	0.34	2E-08	-0.004 (rs131794, A) ⁴

^aeMERGE = electronic Medical Records and Genomics; HCT = hematocrit; HGB = hemoglobin; MAF = minor allele frequency (in the combined cohort); MCH = mean corpuscular hemoglobin; MCHC = mean corpuscular hemoglobin concentration; MCV = mean corpuscular volume; RBC = red blood cell; SNP = single-nucleotide polymorphism.

^bThe β is the effect size of the major allele in efficient mixed-model association-expedited (EMMAX) analysis.

^cThe β for previous studies is the effect size of the minor allele, or the allele listed in the parenthesis.

$P=2\times 10^{-9}$), MCV ($\beta=-1.17$, $P=2\times 10^{-21}$), MCH ($\beta=-0.52$, $P=2\times 10^{-27}$), and MCHC ($\beta=-0.14$, $P=3\times 10^{-10}$).

In addition to the hemochromatosis gene (*HFE*), there are 7 butyrophilin family genes, 4 solute carrier family genes (eg, *SLC17A1*), 40 histone family genes (eg, *HIST1H2BJ*), and 14 other genes (eg, *LRRC16A*, *SCGN*, and *TRIM38*). These genes are in different LD blocks and may therefore independently contribute to variation in RBC traits. For example, there was low LD between SNP rs1800562 in *HFE* (reference SNP) and rs169219 near the tripartite motif containing 38 gene (*TRIM38*) ($r^2=0.04$) and rs129128 in the *HIST1H2AC* gene ($r^2=0.01$).

The SNP (rs169219) in the 5' flanking region of *TRIM38* was associated with MCV ($\beta=0.38$, $P=1\times 10^{-9}$) and MCH ($\beta=0.17$, $P=1\times 10^{-11}$). Three SNPs (rs12216125, rs9379818, and rs9295684) in the 3' flanking region of *TRIM38* were also associated with MCH, and rs12216125 was additionally associated with MCHC. Linkage disequilibrium analysis revealed an LD block spanning approx-

imately 66 kb (26,083-26,150 kb) in this region that contains 6 histone family genes. The *TRIM38* motif includes 3 zinc-binding domains (ie, a RING, a B-box type 1 and a B-box type 2, and a coiled-coil region) and is involved in signal transduction of I- κ B kinase/NF- κ B cascade. However, rs169219 was not independently associated with MCV and MCH by conditional haplotype test ($P=0.23$ and $P=0.21$, respectively).²⁹

Several SNPs (rs707896, rs129128, and rs2856646) within the locus 26,203 to 26,243 kb were associated with MCV (rs129128, $\beta=-0.65$, $P=2\times 10^{-14}$), MCH (rs129128, $\beta=-0.30$, $P=1\times 10^{-19}$), and MCHC (rs129128, $\beta=-0.09$, $P=4\times 10^{-10}$). Although this region is adjacent to *HFE* (rs1800562), LD analysis suggests that these SNPs are located in an LD block (approximately 40 kb) separate from *HFE*. This region contains 4 histone family genes: *HIST1H4C*, *HIST1H1T*, *HIST1H2BC*, and *HIST1H2AC*. The SNP rs129128 was independently associated with MCV ($F=43.1$; $P=5\times 10^{-11}$) and MCH ($F=58.6$; $P=2\times 10^{-14}$).

Chromosome 6p21.1. This region with a strong level of LD contains 4 candidate genes: ubiquitin specific peptidase 49 (*USP49*), mediator complex subunit 20 (*MED20*), bystin-like (*BYSL*), and cyclin D3 (*CCND3*). Multiple intronic SNPs in this region were associated with MCV and MCH. The lead SNP (rs3218097) is in the intron of *CCND3* ($\beta = -0.58$, $P = 1 \times 10^{-18}$ for MCV and $\beta = -0.18$, $P = 6 \times 10^{-12}$ for MCH). The same SNP was associated with RBC count in a Japanese cohort,⁸ and 2 other SNPs (rs9349205 and rs11970772) in an intron of *CCND3* were associated with MCV^{4,5} and MCH.⁴ *CCND3* is involved in hematopoietic stem cell expansion.³⁰ It forms a complex with and functions as a regulatory subunit of cyclin-dependent kinase 4 (CDK4) or cyclin-dependent kinase 6 (CDK6), whose activity is required for cell cycle G1/S transition.

Chromosome 6q23.3. Multiple SNPs in the 5' flanking region of *HBS1L/MYB* were associated with HCT, RBC, MCV, MCH, and MCHC. This locus has been identified in prior GWAS for RBC traits.^{4,5} These SNPs were located in a region (approximately 120 kb) between *HBS1L* and *MYB* (top SNP rs7775698). We also noted several SNPs in the 3' flanking region of the aldehyde dehydrogenase 8 family, member A1 gene (*ALDH8A1*) to be associated with MCV (rs9483769, $\beta = 0.59$, $P = 5 \times 10^{-9}$) and MCH (rs9483769, $\beta = 0.23$, $P = 4 \times 10^{-9}$) (eFigure 2e). We noted low LD ($r^2 = 0.02$) between rs7775698 and rs9483769, and the latter was independently associated with MCV ($F = 14.4$; $P = 1 \times 10^{-4}$) and MCH ($F = 14.1$; $P = 2 \times 10^{-4}$). *ALDH8A1* plays a role in the 9-cis-retinoic acid biosynthesis pathway by converting 9-cis-retinaldehyde into the retinoid X receptor ligand 9-cis-retinoic acid. All-trans retinoic acid and its isomer 9-cis retinoic acid may play a key role in hematopoiesis that occurs in the fetal liver.³¹

Chromosome 6q24.1. Two intergenic SNPs were associated with MCV (lead SNP rs668459, $\beta = -0.33$, $P = 1 \times 10^{-8}$). A SNP (rs643381) in this locus was associated with MCV and another SNP (rs628751) with MCH.⁴ There is only one noncoding gene (*LOC645434*) within ± 60 kb of this locus.

Chromosome 7q22.1. An intronic SNP (rs4434553) in the transferrin receptor 2 gene (*TFR2*) was associated with MCV ($\beta = -0.34$, $P = 2 \times 10^{-9}$) and MCH ($\beta = -0.12$, $P = 2 \times 10^{-8}$). Another intronic SNP (rs7385804) in *TFR2* was associated with HCT and RBC in prior GWAS.^{4,5} Pichler et al³² identified a common variant (rs7385804) in *TFR2* to be associ-

ated with serum iron levels. *TFR2* is a member of the transferrin receptor-like family that mediates cellular uptake of transferrin-bound iron, and mutations in this gene have been associated with hereditary hemochromatosis type III.³³ *TFR2*, a homolog of *TFR1* with expression restricted to hepatocytes and erythroid cells, binds transferrin at lower affinity than *TFR1* and is involved in iron homeostasis.

In addition, imputation-based association analyses identified a SNP (rs551238, $r^2 = 0.31$ with rs4434553 in *TFR2*) in the 3' flanking region of the erythropoietin gene (*EPO*) to be associated with MCV ($P = 4.6 \times 10^{-8}$) and MCH ($P = 9.7 \times 10^{-9}$). *EPO* is a hematopoietic growth factor that regulates RBC mass and stimulates erythropoiesis by binding with the erythropoietin receptor.

Chromosome 19p13.3. Several SNPs at this locus, which spans approximately 45 kb with a high level of LD, were associated with MCH (rs2293683, $\beta = -0.14$, $P = 2 \times 10^{-9}$). Ganesh et al⁴ reported that a SNP (rs11085824) at this locus was associated with MCH ($P = 1 \times 10^{-11}$), which is in LD with rs2293683 ($r^2 = 0.93$). Candidate genes in this region include Kruppel-like factor 1 (erythroid) (*KLF1*), glutaryl-CoA dehydrogenase (*GCDH*), synaptonemal complex central element protein 2 (*SYCE2*), and calreticulin (*CALR*). Of note, *KLF1* (MIM 600599) is a zinc finger transcription factor expressed in erythroid cells and their precursors.³⁴

Chromosome 22q12.3. Multiple SNPs in the F-box protein 7 gene (*FBXO7*) (MIM 605648) were associated with MCV, including a nonsynonymous SNP (rs11107, $\beta = 0.35$, $P = 3 \times 10^{-9}$). A SNP (rs5994574) in the 5' flanking region of *FBXO7* was also associated with MCH. A SNP (rs9609565) in the 5' flanking region of *FBXO7* was previously reported to be associated with MCV.⁵ The F-box proteins constitute one of the 4 subunits of the ubiquitin protein ligase complex called SCFs (SKP1-cullin-F-box), which functions in phosphorylation-dependent ubiquitination.³⁶

Chromosome 22q12.3. The transmembrane protease, serine 6 gene (*TMPRSS6*), has been associated with HGB, MCV, MCH, and MCHC in prior GWAS.^{4,5} We found that multiple SNPs within *TMPRSS6* were associated with HGB, MCV, MCH, and MCHC. We also noted that several SNPs (lead SNP rs130624, 23 kb upstream of *TMPRSS6*) in the potassium channel tetramerisation domain containing 17 gene (*KCTD17*) were associated with HGB and MCV. However, these SNPs were in LD with the top SNP rs855791 in *TMPRSS6*.

A SNP (rs8141597) in the mercaptopyruvate sulfurtransferase gene (*MPST*) was associated with MCV ($\beta = -0.39$, $P = 1 \times 10^{-10}$), and MCH ($\beta = -0.15$, $P = 8 \times 10^{-11}$) (eFigure 2j). *MPST* catalyzes the transfer of a sulfur ion from 3-mercaptopyruvate to cyanide or other thiol compounds and may be involved in cysteine degradation and cyanide detoxification. Mutations in *MPST* result in a rare heritable disorder (mercaptolactate-cysteine disulfiduria) characterized by elevated levels of the mixed disulfide of 3-mercaptolactate and cysteine in the urine,³⁵ RBCs devoid of *MPST* activity, and anemia. We noted low LD ($r^2 = 0.13$) between rs855791 and rs8141597, and the latter was independently associated with MCV ($F = 5.96$; $P = 3 \times 10^{-3}$) and MCH ($F = 4.24$; $P = .01$).

Chromosome 22q13.3. A SNP (rs140522) in the 5' flanking region of the thymidine phosphorylase gene (*TYMP*) was associated with MCV ($\beta = 0.34$, $P = 2 \times 10^{-8}$). Ganesh et al⁴ also found SNP rs131794 in the 5' flanking region of *TYMP* to be associated with MCV. An additional intronic SNP (rs470119) was associated with MCH in a Japanese cohort.⁸ *TYMP* encodes an angiogenic factor that promotes angiogenesis in vivo and stimulates the growth of endothelial cells in vitro.³⁷

Functional Annotation of SNPs Associated With RBC Traits

We created a pipeline to perform functional annotation of genotyped or imputed SNPs associated with the RBC traits (eFigure 1)¹⁶ and identified 32 SNPs associated with RBC traits as potentially functional (Supplemental Table 4, available online at <http://www.mayoclinicproceedings.org>). Seven of the variants were nonsynonymous. The eQTL (Supplemental Table 5 and Supplemental Table 6, available online at <http://www.mayoclinicproceedings.org>), methylation (Supplemental Table 7 and Supplemental Table 8, available online at <http://www.mayoclinicproceedings.org>), and histone modification (Supplemental Table 9 and Supplemental Table 10, available online at <http://www.mayoclinicproceedings.org>) analyses suggested that the remaining SNPs may exert their effect by regulating gene expression. A detailed description of the results of functional characterization of SNPs associated with RBC traits is presented in eAppendix 2.

DISCUSSION

In the present study, we report results of a GWAS for 6 RBC traits (HGB, HCT, RBC count, MCV, MCH, and MCHC) in 12,486 patients of European ancestry from 5 sites that comprise the eMERGE network.^{9,38} We identified 15 chromosomal loci associated with at least one RBC trait, including 11 loci

associated with RBC traits in prior cohort studies.^{4,5,7} Notably, genes present in 3 of the 4 loci newly identified in individuals of European ancestry (*THRB*, *PTPLAD1*, *CDT1*) and in 6 of the 11 replicated loci (*KLF1*, *ALDH8A1*, *CCND3*, *SPTA1*, *FBXO7*, *TFR2/EPO*) are implicated in erythroid differentiation and regulation of cell cycle in hematopoietic stem cells. Our report also highlights the potential use of the EMR in conducting GWAS of quantitative traits of medical importance.

The sample size in this study was similar to that of previous GWAS of RBC traits in individuals of European ancestry,^{4,5} yet we were able to identify 4 new loci and replicate 11 loci that had been identified in at least 2 of the previous reports. Of the 4 new loci that were novel for individuals of European ancestry, 3 have been reported to be associated with RBC traits in Japanese individuals.⁸ An advantage of EMR-based GWAS of medically relevant quantitative traits is that the traits are measured by well-validated and established methods. In addition, multiple measures over time are often available, and these may permit study of the genetic basis of temporal trends in the traits. On the other hand, trait values may be affected by acute illness or chronic comorbid conditions present at the time of measurement, and appropriate EMR-based phenotyping algorithms for excluding such trait values need to be developed (as was done for this study) before genetic analyses.

In the present study, genes (*PTPLAD1*, *THRB*, and *CDT1*) residing in 3 newly identified loci associated with RBC traits are implicated in erythroid differentiation and cell cycle regulation (Figure 3). Coordinated regulation of cell cycle progression and differentiation is critical for normal hematopoiesis. Cells committed to differentiation undergo a programmed loss of proliferative capacity, restricted to only a few divisions because of their irreversible growth arrest in G₁ phase, and terminal erythroid differentiation is accompanied by arrest of the cell cycle in the G₁/G₀ phase.⁴¹ We briefly describe the role of these genes in cell proliferation and differentiation below.

PTPLAD1 potentiates Rac1-induced NF- κ B and c-Jun N-terminal kinase activation and also forms complexes with constitutively activated Rac1.⁴² Rac1 (and Rac2) GTPases are necessary for early erythropoietic expansion in the bone marrow, and erythropoiesis in Rac1^{-/-} and Rac2^{-/-} mice is characterized by abnormal burst-forming units—erythroid colony morphologic findings and decreased numbers of megakaryocyte-erythrocyte progenitors, colony-forming units—erythroid, and erythroblasts in the bone marrow.²¹

THRB is also known as erythroblastic leukemia viral (v-erbA) oncogene homolog 2, avian. The v-erbA oncogene of avian erythroblastosis vi-

FIGURE 3. Genes implicated in erythroid differentiation are associated with red blood cell (RBC) traits. Three novel genes (*THRB*, *PTPLAD1*, *CDT1*) shaded in blue and 6 replicated loci (*KLF1*, *ALDH8A1*, *SPTA1*, *FBXO7*, *CCND3*, *TFR2/EPO*) shaded in brown are shown. A, Different stages of erythroid differentiation. Hematopoietic stem cells committed to erythroid lineage, followed by erythroid terminal differentiation (from burst-forming units–erythroid [BFU-E] to RBC). Adapted and modified from *Curr Opin Hematol*,³⁹ with permission. B, Genes involved in regulation of cell cycle (M, G₁, S, and G₂ phase). The inserted table lists the 3 novel genes associated with RBC traits in the present study. Adapted from *J Mol Cell Biol*,⁴⁰ with permission. Baso EB = basophilic erythroblasts; HSC = hematopoietic stem cell; CFU_E = colony-forming units—erythroid; MCH = mean corpuscular hemoglobin; MCHC = mean corpuscular hemoglobin concentration; MCV = mean corpuscular volume; ORC = origin of replicative complex; Ortho EB = orthochromatic erythroblast; Poly EB = polychromatic erythroblasts; Pro EB = proerythroblasts; RET = reticulocytes; (?) = indicates expression stage cannot be specified.

rus encodes an aberrant version of a gene for a thyroid hormone receptor (c-erbA) and promotes neoplasia by blocking erythroid differentiation. Ligand-activated c-erbA/TR accelerates erythroid differentiation, whereas unliganded c-erbA/TR effectively blocks erythroid differentiation.⁴³ Ex-

pression of *THRB* may vary according to cell cycle, thereby mediating hormone sensitivity and contributing to cell cycle progression during normal development.⁴⁴

Cdk6 has been implicated in the terminal hematopoietic differentiation processes in mice, and loss

of Cdk6 affects the production of terminally differentiated myeloid and erythroid cells.⁴⁵ *CDT1*, which is required for the initiation of DNA replication, together with origin recognition complex and CDC6, constitutes the machinery that loads the minichromosome maintenance complex, a candidate replicative helicase, onto chromatin during the G₁ phase.⁴⁶ In mice, *CDT1* is phosphorylated by CDKs during the cell cycle, which induces the association of *CDT1* with ubiquitination complex SCF-SKP2 and targets *CDT1* for degradation.

Of the genes residing in the replicated loci, 6 were also involved in erythrocyte differentiation and regulation of cell cycle. Erythropoiesis is largely mediated by a relatively small number of lineage-restricted transcription factors, including *KLF1*.⁴⁷ *KLF1* controls the development and differentiation of erythroid lineage by mediating the switch from expression of fetal γ -globin to adult β -globin and regulates transcription of genes that encode cytoskeletal proteins, heme synthesis enzymes, and blood group antigens.^{48,49} *KLF1* also regulates the lineage progression of megakaryocyte-erythroid progenitor cells. Failure of terminal erythroid differentiation in *KLF1*-deficient mice is associated with cell cycle perturbation and reduced expression of *E2F2*.⁴⁹ In humans, haploinsufficiency for *KLF1* causes hereditary persistence of fetal hemoglobin.⁵⁰ In erythrocyte progenitor cells, *KLF1* was upregulated 28-fold when compared with hematopoietic stem cells.²²

ALDH8A1 converts 9-cis-retinaldehyde into the 9-cis-retinoic acid, which influences hematopoiesis in the fetal liver.³¹ *CCND3* interacts with CDK4 and CDK6 to play a lineage-independent role in hematopoiesis,⁵¹ *SPTA1* and *TFR2* are proteins related to erythrocyte structure and function, *EPO* is a hematopoietic growth factor, and *EPO* is the required receptor. In the mouse, *FBXO7* associates specifically with CDK6 to promote CDK6–cyclin D complex formation and cellular transformation.⁵²

The effect of trait-associated alleles may be exerted by multiple mechanisms influencing gene expression.¹⁶ Of the SNPs associated with RBC traits in the present study, 15 SNPs in 7 loci were associated with 27 distant eQTLs (Supplemental Table 5, available online at <http://www.mayoclinicproceedings.org>). The associated SNPs may also influence epigenetic regulation, including methylation and histone modification.

Three of the 4 newly discovered loci associated with RBC traits in our analyses were reported in a Japanese cohort.⁸ Nonetheless, these loci are novel for individuals of European ancestry. The associations of these loci with RBC traits are weaker than the replicated loci, as expected. Our

annotation of likely candidate genes and their function is based on LD patterns and known literature and bioinformatics analyses. Additional work is needed to confirm the role of specific genes at a locus in RBC biology.

CONCLUSION

We identified 4 new genetic loci associated with RBC traits at $P < 5 \times 10^{-8}$ in individuals of European ancestry and replicated 11 previously reported loci. Genes present within 3 of the 4 newly identified loci (*THRB*, *PTPLAD1*, and *CDT1*) and in 6 of the 11 replicated loci (*KLF1*, *ALDH8A1*, *CCND3*, *SPTA1*, *FBXO7*, and *TFR2/EPO*) are implicated in erythroid differentiation/cell cycle of hematopoietic stem cells. The results provide insights into common genetic variants influencing RBC traits and advance our understanding of the mechanisms regulating erythroid differentiation. These findings have implications for understanding the origin of RBC disorders, including anemia and erythroid dysplasia/neoplasia, and developing therapies for such disorders.

SUPPLEMENTAL ONLINE MATERIAL

Supplemental material can be found online at <http://www.mayoclinicproceedings.org>.

Abbreviations and Acronyms: eMERGE = electronic MEDical Records and GENomics; EMMAX = mixed-model association-expedited; EMR = electronic medical record; eQTL = expression quantitative trait locus; GHC = Group Health Cooperative–University of Washington; GWAS = genomewide association study; HCT = hematocrit; HGB = hemoglobin; IBS = identity-by-state; LD = linkage disequilibrium; MC = Marshfield Clinic; MCH = mean corpuscular hemoglobin; MCHC = mean corpuscular hemoglobin concentration; MCV = mean corpuscular volume; MIM = Mendelian Inheritance of Man; NU = Northwestern University; RBC = red blood cell; SNP = single-nucleotide polymorphism; VUMC = Vanderbilt University Medical Center

Grant Support: The eMERGE network was initiated and funded by the National Human Genome Research Institute, with additional funding from National Institute of General Medical Sciences through the following grants: U01-HG-04599 (Mayo Clinic), U01-HG-004610 (GHC), U01-HG-004608 (MC), U01-HG004609 (NU), and U01-HG-04603 (VUMC, also serving as the administrative coordinating center).

Correspondence: Address to Ifthikhar J. Kullo, MD, Mayo Clinic, 200 First St SW, Rochester, MN 55905 (Kullo.iftikhar@mayo.edu).

REFERENCES

1. Ascenzi P, Bellelli A, Coletta M, et al. Multiple strategies for O2 transport: from simplicity to complexity. *IUBMB Life*. 2007; 59(8-9):600-616.
2. Sarnak MJ, Tighiouart H, Manjunath G, et al. Anemia as a risk factor for cardiovascular disease in The Atherosclerosis

- Risk in Communities (ARIC) study. *J Am Coll Cardiol*. 2002;40(1):27-33.
3. Lin JP, O'Donnell CJ, Jin L, Fox C, Yang Q, Cupples LA. Evidence for linkage of red blood cell size and count: genome-wide scans in the Framingham Heart Study. *Am J Hematol*. 2007;82(7):605-610.
4. Ganesh SK, Zakai NA, van Rooij FJ, et al. Multiple loci influence erythrocyte phenotypes in the CHARGE Consortium. *Nat Genet*. 2009;41(11):1191-1198.
5. Soranzo N, Spector TD, Mangino M, et al. A genome-wide meta-analysis identifies 22 loci associated with eight hematological parameters in the HaemGen consortium. *Nat Genet*. 2009;41(11):1182-1190.
6. Chambers JC, Zhang W, Li Y, et al. Genome-wide association study identifies variants in TMPRSS6 associated with hemoglobin levels. *Nat Genet*. 2009;41(11):1170-1172.
7. Benyamin B, Ferreira MA, Willemsen G, et al. Common variants in TMPRSS6 are associated with iron status and erythrocyte volume. *Nat Genet*. 2009;41(11):1173-1175.
8. Kamatani Y, Matsuda K, Okada Y, et al. Genome-wide association study of hematological and biochemical traits in a Japanese population. *Nat Genet*. 2010;42(3):210-215.
9. McCarty CA, Chisholm RL, Chute CG, et al. The eMERGE Network: a consortium of biorepositories linked to electronic medical records data for conducting genomic studies. *BMC Med Genomics*. 2011;4(1):13.
10. Kullo IJ, Ding K, Jouni H, Smith CY, Chute CG. A genome-wide association study of red blood cell traits using the electronic medical record. *PLoS One*. 2010;5(9):e13011.
11. Kullo IJ, Ding K, Shameer K, et al. Complement receptor 1 gene variants are associated with erythrocyte sedimentation rate. *Am J Hum Genet*. 2011;89:181-189.
12. Turner S, Armstrong LL, Bradford Y, et al. Quality control procedures for genome-wide association studies. *Curr Protoc Hum Genet*. 2011;Chapter 1:Unit 1.19.
13. Kang HM, Sul JH, Service SK, et al. Variance component model to account for sample structure in genome-wide association studies. *Nat Genet*. 2010;42(4):348-354.
14. Pruim RJ, Welch RP, Sanna S, et al. LocusZoom: regional visualization of genome-wide association scan results. *Bioinformatics*. 2010;26(18):2336-2337.
15. Purcell S, Neale B, Todd-Brown K, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet*. 2007;81(3):559-575.
16. Freedman ML, Monteiro AN, Gayther SA, et al. Principles for the post-GWAS functional characterization of cancer risk loci. *Nat Genet*. 2011;43(6):513-518.
17. International Hapmap Consortium. A haplotype map of the human genome. *Nature*. 2005;437(7063):1299-1320.
18. Li Y, Willer C, Sanna S, Abecasis G. Genotype imputation. *Annu Rev Genomics Hum Genet*. 2009;10:387-406.
19. Gamazon ER, Zhang W, Konkashbaev A, et al. SCAN: SNP and copy number annotation. *Bioinformatics*. 2010;26(2):259-262.
20. Sakurai A, Nakai A, DeGroot LJ. Structural analysis of human thyroid hormone receptor beta gene. *Mol Cell Endocrinol*. 1990;71(2):83-91.
21. Kalfa TA, Pushkaran S, Zhang X, et al. Rac1 and Rac2 GTPases are necessary for early erythropoietic expansion in the bone marrow but not in the spleen. *Haematologica*. 2010;95(1):27-35.
22. Cui K, Zang C, Roh TY, et al. Chromatin signatures in multipotent human hematopoietic stem cells indicate the fate of bivalent genes during differentiation. *Cell Stem Cell*. 2009;4(1):80-93.
23. Bao L, Zhou M, Cui Y. CTCFBSDB: a CTCF-binding site database for characterization of vertebrate genomic insulators. *Nucleic Acids Res*. 2008;36(Database issue):D83-D87.
24. Nishitani H, Lygerou Z, Nishimoto T, Nurse P. The Cdt1 protein is required to license DNA for replication in fission yeast. *Nature*. 2000;404(6778):625-628.
25. Ferreira MA, Hottenga JJ, Warrington NM, et al. Sequence variants in three loci influence monocyte counts and erythrocyte volume. *Am J Hum Genet*. 2009;85(5):745-749.
26. Ziennicka-Kotula D, Xu J, Gu H, et al. Identification of a candidate human spectrin Src homology 3 domain-binding protein suggests a general mechanism of association of tyrosine kinases with the spectrin-based membrane skeleton. *J Biol Chem*. 1998;273(22):13681-13692.
27. Marchesi SL, Letsinger JT, Speicher DW, et al. Mutant forms of spectrin alpha-subunits in hereditary elliptocytosis. *J Clin Invest*. 1987;80(1):191-198.
28. Gallagher PG, Nilson DG, Steiner LA, Maksimova YD, Lin JY, Bodine DM. An insulator with barrier-element activity promotes alpha-spectrin gene expression in erythroid cells. *Blood*. 2009;113(7):1547-1554.
29. Matsuda A, Suzuki Y, Honda G, et al. Large-scale identification and characterization of human genes that activate NF-kappaB and MAPK signaling pathways. *Oncogene*. 2003;22(21):3307-3318.
30. Kozar K, Ciernych MA, Rebel VI, et al. Mouse development and cell proliferation in the absence of D-cyclins. *Cell*. 2004;118(4):477-491.
31. Tocci A, Parolini I, Gabbianelli M, et al. Dual action of retinoic acid on human embryonic/fetal hematopoiesis: blockade of primitive progenitor proliferation and shift from multipotent/erythroid/monocytic to granulocytic differentiation program. *Blood*. 1996;88(8):2878-2888.
32. Pichler I, Minelli C, Sanna S, et al. Identification of a common variant in the TFR2 gene implicated in the physiological regulation of serum iron levels. *Hum Mol Genet*. 2011;20(6):1232-1240.
33. Roetto A, Totaro A, Piperno A, et al. New mutations inactivating transferrin receptor 2 in hemochromatosis type 3. *Blood*. 2001;97(9):2555-2560.
34. Miller IJ, Bieker JJ. A novel, erythroid cell-specific murine transcription factor that binds to the CACCC element and is related to the Kruppel family of nuclear proteins. *Mol Cell Biol*. 1993;13(5):2776-2786.
35. Billaut-Laden I, Rat E, Allorge D, et al. Evidence for a functional genetic polymorphism of the human mercaptopyruvate sulfurtransferase (MPST), a cyanide detoxification enzyme. *Toxicol Lett*. 2006;165(2):101-111.
36. Chang YF, Cheng CM, Chang LK, Jong YJ, Yuo CY. The F-box protein Fbxo7 interacts with human inhibitor of apoptosis protein cIAP1 and promotes cIAP1 ubiquitination. *Biochem Biophys Res Commun*. 2006;342(4):1022-1026.
37. Leung DW, Cachianes G, Kuang WJ, Goeddel DV, Ferrara N. Vascular endothelial growth factor is a secreted angiogenic mitogen. *Science*. 1989;246(4935):1306-1309.
38. Kho AN, Pacheco JA, Peissig PL, et al. Electronic medical records for genetic research: results of the eMERGE consortium. *Sci Transl Med*. 2011;3(79):79re71.
39. Koury MJ, Sawyer ST, Brandt SJ. New insights into erythropoiesis. *Curr Opin Hematol*. 2002;9(2):93-100.

40. Truong LN, Wu X. Prevention of DNA re-replication in eukaryotic cells. *J Mol Cell Biol*. 2011;3(1):13-22.
41. Tsiftoglou AS, Vizirianakis IS, Strouboulis J. Erythropoiesis: model systems, molecular regulators, and developmental programs. *IUBMB Life*. 2009;61(8):800-830.
42. Courilleau D, Chastre E, Sabbah M, Redeuilh G, Atfi A, Mester J. B-ind1, a novel mediator of Rac1 signaling cloned from sodium butyrate-treated fibroblasts. *J Biol Chem*. 2000;275(23):17344-17348.
43. Bartunek P, Zenke M. Retinoid X receptor and c-cerbA/thyroid hormone receptor regulate erythroid cell growth and differentiation. *Mol Endocrinol*. 1998;12(9):1269-1279.
44. Maruvada P, Dmitrieva NI, East-Palmer J, Yen PM. Cell cycle-dependent expression of thyroid hormone receptor-beta is a mechanism for variable hormone sensitivity. *Mol Biol Cell*. 2004;15(4):1895-1903.
45. Malumbres M, Sotillo R, Santamaria D, et al. Mammalian cells cycle without the D-type cyclin-dependent kinases Cdk4 and Cdk6. *Cell*. 2004;118(4):493-504.
46. Sugimoto N, Tatsumi Y, Tsurumi T, et al. Cdt1 phosphorylation by cyclin A-dependent kinases negatively regulates its function without affecting geminin binding. *J Biol Chem*. 2004;279(19):19691-19697.
47. Cantor AB, Orkin SH. Transcriptional regulation of erythropoiesis: an affair involving multiple partners. *Oncogene*. 2002;21(21):3368-3376.
48. Hodge D, Coghill E, Keys J, et al. A global role for EKLF in definitive and primitive erythropoiesis. *Blood*. 2006;107(8):3359-3370.
49. Pilon AM, Arcasoy MO, Dressman HK, et al. Failure of terminal erythroid differentiation in EKLF-deficient mice is associated with cell cycle perturbation and reduced expression of E2F2. *Mol Cell Biol*. 2008;28(24):7394-7401.
50. Borg J, Papadopoulos P, Georgitsi M, et al. Haploinsufficiency for the erythroid transcription factor *KLF1* causes hereditary persistence of fetal hemoglobin. *Nat Genet*. 2010;42(9):801-805.
51. Uchimaru K, Taniguchi T, Yoshikawa M, et al. Detection of cyclin D1 (bcl-1, PRAD1) overexpression by a simple competitive reverse transcription-polymerase chain reaction assay in t(11;14)(q13;q32)-bearing B-cell malignancies and/or mantle cell lymphoma. *Blood*. 1997;89(3):965-974.
52. Laman H, Funes JM, Ye H, et al. Transforming activity of Fbxo7 is mediated specifically through regulation of cyclin D/cdk6. *EMBO J*. 2005;24(17):3104-3116.