ARIC Manuscript Proposal #2410

PC Reviewed: 8/12/14 **Status: A** **Priority: 2**
SC Reviewed: _____ **Status: _____** **Priority: ____**

Population Architecture using Genomics and Epidemiology (PAGE)
Ver. 06/12/14

PAGE Manuscript Proposal Template
Submit proposals by email to the PAGE Coordinating Center to Rasheeda Williams.
All sections must be completed; incomplete applications will be returned.
Do not exceed 3 pages in length (not including references).

PAGE Ms. Number: 77 **Approval Date:**

Title of Proposed Ms.: Generalization and fine mapping of loci previously associated with RBC traits to multi-ethnic populations: The PAGE Study

I. INVESTIGATOR INFORMATION:

Name of Lead Author: Chani Hodonsky **Junior Investigator? Yes**
Email Address: chani_hodonsky@unc.edu
Telephone Number: 919/966-4312

Names, affiliations and email address of PAGE Investigators proposed as co-authors:

N, N	Affiliation in PAGE	Email
Kari North	CALiCo – ARIC	kari_north@unc.edu
Danyu Lin	CALiCo –SOL	lin@bios.unc.edu
Ran Tao	CALiCo-SOL	taor@live.unc.edu
Christy Avery	CALiCo – ARIC	christy_avery@unc.edu
Lucia Hindorff	NHGRI	hindorffl@mail.nih.gov
Steve Buyske	CC	buyske@stat.rutgers.edu
Alex Reiner	WHI	apreiner@u.washington.edu
Jonathan Kocarnik	WHI	kocarnik@u.washington.edu
Bruce Psaty	CHS	psaty@uw.edu
Myriam Fornage	CARDIA	myriam.Fornage@uth.tmc.edu
Ruth Loos	Mt Sinai Biobank	ruth.loss@mssm.edu
Christina Wassel	CALiCo-SOL	cwassel@pitt.edu
Charles Kooperberg	WHI	clk@fhcrc.org
John Eckfeldt	SOL	eckfe001@umn.edu
Bharat Thyagarajan	SOL	thya0003@umn.edu
Other WHI, CARDIA, and Mt. Sinai authors will be added later.		

Names, affiliations, email address of non-PAGE investigators proposed as co-authors:
N/A

II. SCIENTIFIC RATIONALE (Please be specific and concise) 2-3 paragraphs

Red Blood Cell (RBC) trait deficiencies cause multiple circulatory diseases such as thalassemia, polycythemia, and genetic or nonhereditary anemia. RBC traits are also associated with risk for cardiovascular disease, the leading cause of death in the United States [1-3]. Variation of RBC traits—due to genetic and/or environmental factors—can directly influence disease outcomes by diminishing the ability of RBCs to effectively transport oxygen to the rest of the body. Genetic modifiers of disease severity in several blood-related disorders have also been found, indicating a complex regulatory network for RBC traits [4, 5]. As these traits—specifically hemoglobin (HGB), hematocrit (HCT), and RBC count—exhibit high heritability (40-90% in twins), establishing narrow loci associated with these traits will help elucidate causal alleles and treatment options for RBC-related disorders [6-8]. Related derived traits used in disease diagnosis—namely mean corpuscular volume (MCV), mean corpuscular hemoglobin (MCH), and mean corpuscular hemoglobin concentration (MCHC)—have been uniquely associated with multiple loci. These associations suggest physiological relevance for these traits, which are measured via transformation (Table 2) [9-11]. Causal alleles for several population-specific recessive diseases have been shown to confer protective effects in the heterozygous state, with sickle-cell anemia and (somewhat controversially) β-thalassemia being prime vascular-related examples [4, 12, 13].

Genetic association studies for RBC traits in non-European populations are extremely limited in number. As has been emphasized in recent studies, SNPs identified in studies restricted to European ancestral groups entail only the broadest stroke regarding chromosomal location of causal variants and generalizability to other ethnic groups [9, 14]. Increasing the population base of RBC genetic studies to include populations of African and Hispanic ancestry provides opportunities for narrowing and fine-mapping already established loci, as well as the possibility of identifying population-specific loci not previously described [15]. We therefore propose to evaluate eleven RBC loci previously identified in populations of European- and South Asian descent and fine-mapped on the Metabochip (*SPTA1, BCL11A, HFE, ABO, HK1, SH2B3/ATXN2, LIPC, PPCDC, NUTF2, NEUROD2, and TMPRSS6*) for evidence of generalization and locus refinement in the multi-ethnic PAGE populations.

III. OBJECTIVES AND PLAN (Please be specific and concise)

a. **Study Questions/Hypotheses**

RBC count, HGB, HCT, MCV, MCH, and MCHC are all involved in proper oxygen transportation and absorption by bodily tissues. Pleiotropic effects across multiple cardiovascular phenotypes have been associated with abnormal expression of these traits [2-5, 7]. We intend to elucidate causal variants for RBC trait-related disorders using fine-mapping of SNPs across multiple ethnic groups. First, we will perform fine-mapping in African American, Latino, and Asian populations of eleven loci previously associated with one or more of these traits in European- and South Asian ancestral populations. We will further perform fine-mapping analysis

of all SNPs included in the Metabochip to identify new associations of RBC traits in regions with known cardiovascular and metabolic trait associations [16].

b. **Study populations, study design for each**

All PAGE study participants of Latino, African American, and Asian ancestry with both Metabochip data and measures of RBC traits analyzed via CBC or whole-blood fractionation (WHI only) performed at the study sites.

c. **Variant/SNPs (Specify)**

Approximately 11,900 SNPs fine-mapped on the Metabochip are located near previously identified RBC trait loci (Table 1), restricting to SNPs with population-specific minor allele frequency estimates ≥0.001 (0.1%) [5, 9-11, 17, 18]. We will also be evaluating all Metabochip SNPs for previously unannotated RBC trait associations in all somatic chromosomes.

TABLE 1. Characterization of 11 genomic regions to be fine-mapped for RBC traits.

Previously Identified Loci	Genetic Region	Base Pair Range (Hg37)	N SNPs Fine-mapped[a,b]	Previously Identified Phenotype
SPTA1	1q23.1	158575876 - 158656445	66	HGB
BCL11A / FANCL	2p16.1	60557031 - 60607007	160	HGB, MCV
HFE	6p22.2	25234884 - 26153335	2849	HCT, HGB, MCV
ABO	9q34.2	136039533 - 136482476	1410	MCHC
HK1	10q22.1	71049561 - 71141144	183	HCT, HGB, MCHC
SH2B3 / ATXN2	12q24.12	111290599 - 113206306	3494	HCT, HGB
LIPC	15q21	58666341 - 58710627	184	HGB
PPCDC	15q24	74864568 - 75450557	1353	MCV
NUTF2	16q22	67552105 - 68335392	1053	RBC
NEUROD2	17q12	37387440 - 38082831	1150	RBC
TMPRSS6	22q13.31	37449250 - 37499692	25	HCT, HGB, MCH, MCV

[a] Restricted to SNPs with minor allele frequency > 0.01.
[b] We expect numbers to differ by race/ethnicity.

d. **Phenotype(s) (Specify)**

We will analyze six RBC-relevant phenotypes, all recorded as part of a CBC by CALiCo or WHI investigators. Although HCT, HGB, and RBC count are correlated (in twins: HGB and HCT $r^2 = 0.95$; HGB/RBC and HCT/RBC $r^2 \sim 0.8$), they are not mutually inclusive regarding disease phenotypes, which can affect one trait without a detectable abnormality in the other(s) [7, 8]. We will also analyze MCV, MCH, and MCHC, calculated as outlined in Table 2. These derived traits can be approximated using RBC count, HGB, and HCT; however, unique genetic associations involving these traits have been characterized, indicating the importance of including them in our analysis [9, 10, 14, 18]. All studies involved used comparable procedures for preparing participants, drawing blood, and sample processing/QC. For WHI participants, only HGB and HCT were measured during baseline-visit sampling, so only HGB, HCT, and MCH will be analyzed in this population.

e. **Covariates (Specify)**

To maintain consistency with previous GWA study efforts, we will consider age, sex, study center/region (where appropriate), and principal components measuring global ancestry as covariates [9, 11, 14].

The following participants will be excluded:

i. RBC trait measurements / Metabochip data unavailable
ii. Participant does not self-identify as being of Latino, African American, or Asian ancestry
iii. Pregnancy at time of exam or prior diagnosis of blood cancer, sickle-cell or other congenital hemolytic anemia, HIV, or ESRD to the extent these phenotypes can be identified in the participating cohorts.

f. **Main statistical analysis methods**

Race/ethnic-specific linear regression models will be used to test associations between the RBC traits and approximately 12,000 SNPs from 11 regions fine-mapped for those traits under an additive genetic model and including the above-mentioned covariates. We will transform trait values as previously described using inverse normal transformation to reduce the influence of large outliers [19]. Standard errors of trait distributions will be evaluated for normality before proceeding with linear regression. Examples of figures and tables that will be constructed are presented below in the appendix.

Table 2. Descriptions of Red Blood Cell Traits*

Abbv	Trait	Units	Transformation
RBC Count	Erythrocyte Count	10^6 cells / mm^3	Square root
HGB	Hemoglobin Blood Concentration	g/dL	N/A
HCT	Hematocrit (RBC whole blood fraction)	%	N/A
MCH	Mean Corpuscular Hemoglobin: **MCH = HGB*10 / RBC**	picogram	Natural log
MCHC	MCH Concentration: **MCHC = HGB*100 / HCT**	g/dL	Natural log
MCV	Mean Corpuscular Volume (Erythrocytes): **MCV = HCT*10 / RBC**	femtoliter	Natural log

***Trait calculations and descriptions adopted from Table S1 [9].

For analysis of SOL data, CALiCo/PAGE has been working closely with the SOL GWAS analysis center at University of Washington, and has implemented an analysis strategy that accounts for the sampling design (i.e. sampling weights) and relatedness among SOL participants, with additional covariate adjustment for principal components of global ancestry to control for population stratification. (Local ancestry adjustment will take place if necessary.) This analysis strategy involves modified generalized estimating equation (GEEs) for binary outcomes, and modified linear mixed effects models for continuous traits. This strategy has been shown to adequately control for genomic inflation across several traits, which had been an issue with some analysis strategies in the past for SOL.

Multiple testing thresholds

For each HGB-, HCT-, or RBC-associated locus, we anticipate that SNPs associated with those traits in African Americans, Latinos, and Asians will be correlated with the index SNP reported in Europeans. Therefore, we will first identify and test SNPs that are correlated ($r^2 > 0.20$) with the index signals in Europeans using LD statistics estimated in the Malmö Diet and Cancer Study. For loci with numerous reported index SNPs, we will consider SNPs with $r^2 < 0.20$ as representing independent signals. In order to determine the appropriate multiple testing threshold for declaration of whether the independent signals significantly associated with red blood cell traits are generalizable to PAGE populations, we will then estimate the number of tag SNPs for the most well represented race/ethnic group (African American) needed to capture all common alleles ($r^2 > 0.80$, MAF > 0.05) using LD patterns specific to that race/ethnicity. Using the ethnicity with the highest number of non-LD SNPs for our corrected significance cutoff will allow us to estimate associations conservatively while having a common α threshold for all ethnicities in the analysis. As an example, the multiple testing threshold for declaring generalization that was used in our previous QT fine-mapping effort was $\alpha_a = 0.05/415$, using 415 as the total number of tags identified for African American LD patterns [16].

For all SNPs that are not correlated with the index signal in Europeans, i.e. population-specific SNPs influencing RBC traits, we will use a conservative Bonferroni correction to designate ethnicity-specific alpha levels based on the number of SNPs in all remaining Metabochip loci. Conditional analyses will then be performed to identify independent signals. Specifically, analyses will be repeated for each locus including the SNP with the smallest *p* value as a covariate. This approach will be performed adjusting for successively less significant SNPs until no SNPs with *p* values lower than the Bonferroni-adjusted alpha level are identified. Finally, we will evaluate whether regulatory regions are over-represented in our associated fine-mapped SNPs using ENCODE data.

g. **Ancestry information used? No __ Yes __X__**

How is it used in the analyses?

Global ancestry, as measured by principal components, will be included as a covariate in the analysis.

h. **Anticipated date of draft manuscript to P&P:** 6 months after all data are available

i. **What manuscript proposals listed on** *www.pagestudy.org/index.php/manuscripts* **are most related to the work proposed here? Approved PAGE mss. numbers:**
This is the first PAGE "blood traits" paper, to the best of our knowledge.

If any: Have the lead authors of these proposals been contacted for comments and/or collaboration? Yes __ No __

IV. SOURCE OF DATA TO BE USED (Provide rationale for any data whose relevance to this manuscript is not obvious): **Check all that apply:**

Aggregate/summary data to be generated by investigators of the study(ies) mentioned:

[ ] EAGLE; **[X]** CALiCO; **[ ]** MEC; **[X]** WHI; **[X]** CC; **[ ]**
Other:_____
If CALiCo, specify **[X]** ARIC; **[X]** CARDIA; **[X]** CHS; **[ ]** SHS-Fam; **[ ]** SHS-Cohort; **[X]** SOL

I, _CJH_, affirm that this proposal has been reviewed and approved by all listed investigators.

V. REFERENCES

1. Yang, Q., et al., *Genome-wide association and linkage analyses of hemostatic factors and hematological phenotypes in the Framingham Heart Study.* BMC Med Genet, 2007. **8 Suppl 1**: p. S12.
2. Agre, P., et al., *Partial deficiency of erythrocyte spectrin in hereditary spherocytosis.* Nature, 1985. **314**(6009): p. 380-3.
3. Havlik, R.J., et al., *Evidence for additional blood pressure correlates in adults 20-56 years old.* Circulation, 1980. **61**(4): p. 710-5.
4. Chami, N. and G. Lettre, *Lessons and Implications from Genome-Wide Association Studies (GWAS) Findings of Blood Cell Phenotypes.* Genes (Basel), 2014. **5**(1): p. 51-64.
5. Nuinoon, M., et al., *A genome-wide association identified the common genetic variants influence disease severity in beta0-thalassemia/hemoglobin E.* Hum Genet, 2010. **127**(3): p. 303-14.
6. Wright, F.A., et al., *Heritability and genomics of gene expression in peripheral blood.* Nat Genet, 2014. **46**(5): p. 430-7.
7. Whitfield, J.B. and N.G. Martin, *Genetic and environmental influences on the size and number of cells in the blood.* Genet Epidemiol, 1985. **2**(2): p. 133-44.
8. Evans, D.M., I.H. Frazer, and N.G. Martin, *Genetic and environmental causes of variation in basal levels of blood cells.* Twin Res, 1999. **2**(4): p. 250-7.
9. Ganesh, S.K., et al., *Multiple loci influence erythrocyte phenotypes in the CHARGE Consortium.* Nat Genet, 2009. **41**(11): p. 1191-8.
10. Kullo, I.J., et al., *A genome-wide association study of red blood cell traits using the electronic medical record.* PLoS One, 2010. **5**(9).
11. van der Harst, P., et al., *Seventy-five genetic loci influencing the human red blood cell.* Nature, 2012. **492**(7429): p. 369-75.
12. Hashemi, M., et al., *Effect of heterozygous beta-thalassaemia trait on coronary atherosclerosis via coronary artery disease risk factors: a preliminary study.* Cardiovasc J Afr, 2007. **18**(3): p. 165-8.
13. Wang, C.H. and R.F. Schilling, *Myocardial infarction and thalassemia trait: an example of heterozygote advantage.* Am J Hematol, 1995. **49**(1): p. 73-5.
14. Chen, Z., et al., *Genome-wide association analysis of red blood cell traits in African Americans: the COGENT Network.* Hum Mol Genet, 2013. **22**(12): p. 2529-38.
15. McCarthy, M.I. and J.N. Hirschhorn, *Genome-wide association studies: potential next steps on a genetic journey.* Hum Mol Genet, 2008. **17**(R2): p. R156-65.
16. Avery, C.L., et al., *Fine-mapping and initial characterization of QT interval loci in African Americans.* PLoS Genet, 2012. **8**(8): p. e1002870.
17. Chambers, J.C., et al., *Genome-wide association study identifies variants in TMPRSS6 associated with hemoglobin levels.* Nat Genet, 2009. **41**(11): p. 1170-2.
18. Kamatani, Y., et al., *Genome-wide association study of hematological and biochemical traits in a Japanese population.* Nat Genet, 2010. **42**(3): p. 210-5.
19. Lo, K.S., et al., *Genetic association analysis highlights new loci that modulate hematological trait variation in Caucasians and African Americans.* Hum Genet, 2011. **129**(3): p. 307-17.

PAGE Metabochip fine-mapping example figures and tables. The below figures and tables were excerpted from our PAGE fine-mapping study of QT interval loci that was conducted in African American populations [16]. Although tables and figures are presented below for African American populations, similar tables and figures will be prepared for each race/ethnic group under investigation. Tables will include results from all six RBC phenotypes under investigation.

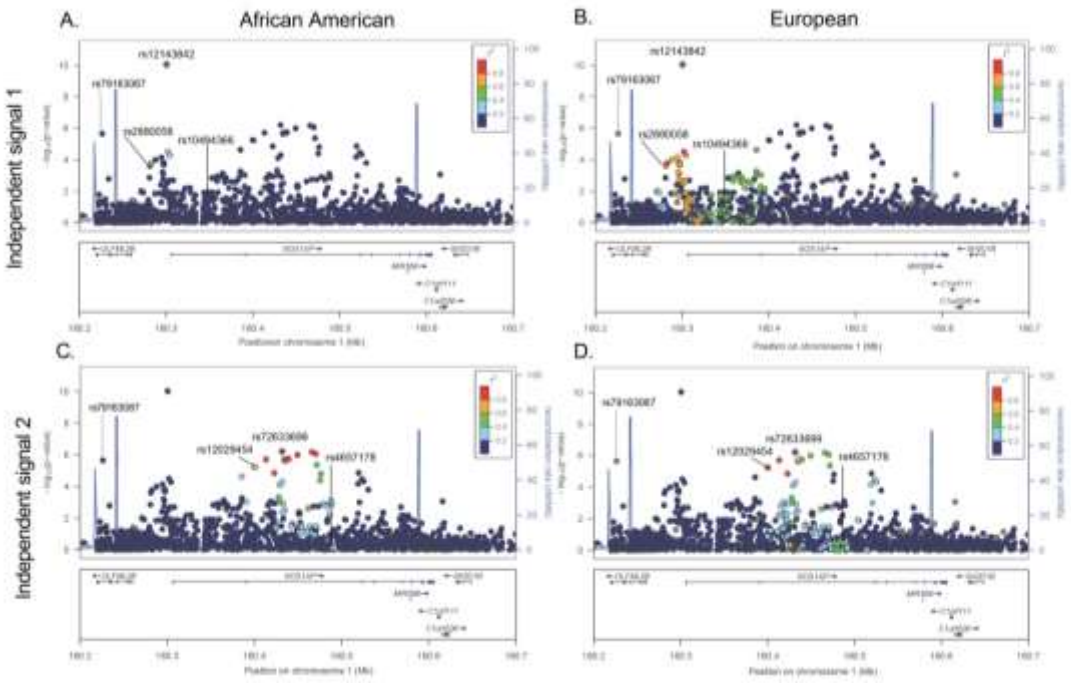

FIGURE 1.–Log *P* plot for common SNPs at the *NOS1AP* independent signal 1 and 2 loci. *P*-values are estimated in African Americans are plotted using linkage disequilibrium estimates from African Americans (panels A and C) and Europeans (panels B and D). SNPs are represented by *circles*, lines indicate index SNPS previously identified in GWA studies of European and Indian Asian populations, and the *large blue diamond* is the best marker in African Americans. Circle color represents correlation with the best marker in African Americans: *blue* indicates weak correlation and *red* indicates strong correlation. Recombination rate is plotted in the background and annotated genes are shown at the bottom of the plot.

TABLE 1. Associations with common variants at known QT loci in n=8,644 African American participants.

Index SNPs from GWA studies in European and Indian Asian populations								Best marker in African Americans[a]					r^2 with index SNP	
					CAF		*P*-value		BP (build			*P*-value		
Locus	Position	Ind. signal	Index SNP	Alleles	EU[b]	AF[c]	(AF)	Marker	36)	Alleles	CAF		EU[b]	AF[c]

[a]Restricted to SNPs with minor allele frequency > 0.01.[b]Calculated in the Malmö Diet and Cancer Study or 1,000 Genomes CEU data when Malmö data unavailable. [c]Calculated in the Atherosclerosis Risk in Communities Study. [d]SNP not present on Metabochip, SNP proxy substituted. [e]SNP not present on Metabochip, but in very high LD with rs2968863 (r^2> 0.95). [f]SNP failed quality control and no proxy was available. AF, African American. BP, base pair. CAF, coded allele frequency. Est, estimate. European. GWA, genome wide association. Ind, independent. NA, not available. SE, standard error. SNP, single nucleotide polymorphism.

TABLE 2. Comparison of linkage disequilibrium patterns between populations of African and European descent for six previously identified QT loci significantly associated with QT in n=8,644 African American participants from four studies.

		African Americans		Europeans		
Locus	Ind. signal	N. SNPs in LD with best marker[a,b]	Region size (kb)	N. SNPs in LD with index SNPs[a,c]	Region size (kb)	Region size difference (kb)[d]

[a]$r^2 \geq 0.50$.[b]Calculatedusing African American LD patterns. [c]Calculated using European LD patterns. [d]Calculated as (African American region size – European region size (kb)). LD, linkage disequilibrium.

SUPPLEMENTAL TABLE 1. Demographic characteristics of n=8,644 African American participants from four studies.

Characteristic	ARIC	WHI PAGE		WHI SHARe
		Wave 1	Wave 2	
N				
Age, years, mean (SD)				
Sex, female, N (%)				
QT duration, ms, mean (SD)				
Heart rate, bpm, mean (SD)				

ARIC, Atherosclerosis Risk in Communities Study. Bpm, beats per minute; Ms, milliseconds; PAGE, Population Architecture using Genomics and Epidemiology. SHARe, SNP Health Association Resource. WHI, Women's Health Initiative.