ARIC Manuscript Proposal #3039

1.a. Full Title: Multi-ethnic GWAS meta-analysis of blood cell traits
   b. Abbreviated Title (Length 26 characters): GWAS of blood traits

2. Writing Group:
   Writing group members:

   Nathan Pankratz (University of Minnesota), Kari North (University of North Carolina), Misa Graff (University of North Carolina), Laura Raffield (UNC), Yun Li (UNC), Leslie Lange (University of Colorado), Linda Polfus (University of Kentucky), Santhi Ganesh (University of Michigan), Alana Morrison (UT)

   I, the first author, confirm that all the coauthors have given their approval for this manuscript proposal. __NP__ [please confirm with your initials electronically or in writing]

   First author: Nathan Pankratz
   Address: University of Minnesota

   Phone
   E-mail: pankr018@umn.edu

   ARIC author to be contacted if there are questions about the manuscript and the first author does not respond or cannot be located (this must be an ARIC investigator).

   Name: Kari North
   Address: 137 E. Franklin St., Suite 306
             Chapel Hill, NC 27514

   Phone: 919/966-2148
   Fax: 919/966-9800
   E-mail: kari_north@unc.edu

3. Timeline: Analyses will begin once study-specific permissions are obtained.

4. Rationale: Circulating blood cell counts represent potentially important intermediate phenotypes for a variety of cardiovascular, pulmonary, hematologic, and immunologic diseases. Recent genome-wide association studies (GWAS) have begun to contribute to our understanding of the genetics of blood cell traits in European and African ancestry populations. The Exome Sequencing Project (ESP), Cohorts for Heart and Aging Research in Genomic Epidemiology
(CHARGE), HaemGen, and Continental Origins and Genetic Epidemiology Network (COGENT) Consortia [1-4] have identified >100 loci associated with blood cell traits including hemoglobin concentration, hematocrit, mean corpuscular volume (MCV), mean corpuscular hemoglobin (MCH), mean corpuscular hemoglobin concentration (MCHC) red blood cell count (RBC), white blood count (WBC), platelet count, and mean platelet volume. Traits such as leukocyte count and hemoglobin are known to be associated with mortality and cardiovascular disease [5, 6].

While a number of GWAS have already been conducted in European populations for blood cell related traits, more comprehensive analyses using novel imputation panels with improved coverage of rarer variants, such as the Haplotype Reference Consortium (HRC) panel [7], and with inclusion of additional cohorts, notably those from non-European populations and those derived from large biobanks, can increase power to detect novel loci, particularly for lower frequency variants. Through a planned consortium effort, combined sample sizes of at least 100-150,000 individuals with blood cell phenotypes should be achieved, with possible addition of nearly 500,000 additional individuals from the UK Biobank.

The overarching goal of this proposal is to determine and investigate novel genetic associations with blood cell traits through a new large consortium effort (BCX) related to blood cell traits using genome-wide imputation data to the HRC panel for European ancestry participants and the 1000 Genomes phase 3 version 5 reference panel for African ancestry participants. We hypothesize this effort will reveal additional genes which play an important role in blood cell traits and function. ARIC is an excellent window into genetic factors relevant to blood cell traits.

5. Main Hypothesis/Study Questions:
We plan to investigate the association of common and low frequency variants with blood cell counts and indices. We hypothesize this effort will reveal additional genes which play an important role in blood cell traits and function.

6. Design and analysis (study design, inclusion/exclusion, outcome and other variables of interest with specific reference to the time of their collection, summary of data analysis, and any anticipated methodologic limitations or challenges if present).

Our analyses will focus on more than a dozen quantitative traits available in ARIC samples: red blood count, red blood cell distribution width, reticulocyte count, hemoglobin, hematocrit, mean corpuscular hemoglobin, mean corpuscular hemoglobin concentration, mean corpuscular volume, WBC and differential counts (neutrophil, eosinophil, basophil, lymphocyte, and monocyte), and platelet count. Study exclusions are current pregnancy and kidney disease (eGFR<60). We will perform a number of preliminary QC steps. We will then conduct standard linear regression analyses for dosage values for each well imputed variant using standard software such as MACH2QTL [8], PLINK [9], or EPACTS. QQ-plots will be examined to identify potential issues that would require additional refining of sample and variant filters.

Pertinent variables: red blood cell count, red blood cell distribution width, hemoglobin, hematocrit, mean corpuscular hemoglobin, mean corpuscular hemoglobin concentration, mean corpuscular volume, WBC and differential counts (neutrophil, eosinophil, basophil, lymphocyte, and monocyte), platelet count, and mean platelet volume.
Covariate data (baseline): Stratify by self-reported race, adjust for sex, age at measurement, age squared, center, and ancestry (as described below)

GWAS data imputed to the HRC imputation panel – The GWAS data for European ancestry participants has already been imputed for all loci in the Haplotype Reference Consortium (HRC) panel using best practices. GWAS data for African ancestry participants has also been previously imputed to 1000 Genomes phase 3 version 5. Dosage values will be used as predictors in the analytic model.

Ancestry: We will estimate the ancestral admixture of ARIC participants using genome-wide genotyping data from the Affymetrix 6.0 platform (after using standard quality control metrics: callrate>95%, in Hardy-Weinberg equilibrium [p>10^{-5}] and not in high linkage disequilibrium with another variant [r^2<0.80]). Based on our prior experience with this data, these quality control measures will leave a final sample of 631,243 autosomal SNPs from which to estimate ancestry using ADMIXTURE.

7.a. Will the data be used for non-CVD analysis in this manuscript? __X__ Yes  ____ No

b. If Yes, is the author aware that the file ICTDER03 must be used to exclude persons with a value RES_OTH = “CVD Research” for non-DNA analysis, and for DNA analysis RES_DNA = “CVD Research” would be used? __X__ Yes  ____ No
(This file ICTDER has been distributed to ARIC PIs, and contains the responses to consent updates related to stored sample use for research.)

8.a. Will the DNA data be used in this manuscript? _X___ Yes  ____ No

8.b. If yes, is the author aware that either DNA data distributed by the Coordinating Center must be used, or the file ICTDER03 must be used to exclude those with value RES_DNA = “No use/storage DNA”? ____ Yes  ____ No

9. The lead author of this manuscript proposal has reviewed the list of existing ARIC Study manuscript proposals and has found no overlap between this proposal and previously approved manuscript proposals either published or still in active status. ARIC Investigators have access to the publications lists under the Study Members Area of the web site at: http://www.cscc.unc.edu/ARIC/search.php

__X__ Yes  _______ No

10. What are the most related manuscript proposals in ARIC (authors are encouraged to contact lead authors of these proposals for comments on the new proposal or collaboration)?

MS 2410 - Generalization and fine mapping of loci previously associated with RBC traits to multi-ethnic populations: The PAGE Study
MS 2879 - A Search for Common Multiethnic- and Rare Ancestry-Specific Genetic Markers of Red Blood Cell Traits Among African Americans and Hispanic/Latinos in PAGE
MS 2482 - Exonic Variants Associated with Blood Cell Traits
The first two proposals are distinct from ours in that they utilize different traits and a different set of genotypic markers. The third was focused on rare variants and is already published.

11.a. Is this manuscript proposal associated with any ARIC ancillary studies or use any ancillary study data? _____ Yes ___X___ No

11.b. If yes, is the proposal
   ___ A. primarily the result of an ancillary study (list number* ________)
   ___ B. primarily based on ARIC data with ancillary data playing a minor role (usually control variables; list number(s)* __________ __________ __________)

*ancillary studies are listed by number at http://www.csc.c.unc.edu/aric/forms/

12a. Manuscript preparation is expected to be completed in one to three years. If a manuscript is not submitted for ARIC review at the end of the 3-years from the date of the approval, the manuscript proposal will expire.

12b. The NIH instituted a Public Access Policy in April, 2008 which ensures that the public has access to the published results of NIH funded research. It is your responsibility to upload manuscripts to PubMed Central whenever the journal does not and be in compliance with this policy. Four files about the public access policy from http://publicaccess.nih.gov/ are posted in http://www.cscc.unc.edu/aric/index.php, under Publications, Policies & Forms. http://publicaccess.nih.gov/submit_process_journals.htm shows you which journals automatically upload articles to PubMed central.

13. Per Data Use Agreement Addendum, approved manuscripts using CMS data shall be submitted by the Coordinating Center to CMS for informational purposes prior to publication. Approved manuscripts should be sent to Pingping Wu at CC, at pingping_wu@unc.edu. I will be using CMS data in my manuscript _____ Yes ___X___ No.
Literature Cited