ARIC Manuscript Proposal #1810

PC Reviewed: 6/14/11	**Status: A**	**Priority: 2**
SC Reviewed: _________	**Status: _____**	**Priority: ____**

1.a. Full Title: **Comprehensive Evaluation of Imputation Quality and Coverage in African Americans**

b. Abbreviated Title (Length 26 characters): **Comprehensive Evaluation of Imputation Quality and Coverage in African Americans**

2. Writing Group: ARIC GWAS Imputation Working Group
Writing group members: Pritam Chanda, Naoya Yuhki, Man Li, Alex Hartz, Dan E. Arking, Linda Kao. Other ARIC study members are welcome to participate.

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

F**irst author**: **Pritam Chanda**
Address: 733 N. Broadway, BRB 311,Baltimore MD 21205
Phone: 410-502-5936
Fax:
E-mail: pchanda2@jhmi.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: **Dan E. Arking**
Address: 733 N. Broadway, BRB 453
Baltimore, MD 21205
Phone: 410-502-4867
Fax: 410-614-8600
E-mail: arking@jhmi.edu

3. Timeline: Analyses to be completed spring of 2011, and manuscript submitted by July 2011.

4. Rationale:
Although considerable progress has been made in imputing genotype data with high imputation accuracy and reliability in samples of European descent, progress has been limited in African American samples primarily due to the complex evolutionary history, high levels of mixed ancestry and admixture, short linkage disequilibrium (LD) blocks and the lack of population specific genotyping microarray systems and study populations. Few studies have been published to impute African American populations using the latest publicly available reference panels from HapMap and 1000 Genomes.

The current literature on imputation of the African Americans is lacking in that the imputation performance of the three popular imputation algorithms – MACH, IMPUTE v2 and BEAGLE using the 1000 Genomes panels and in combinations with HapMap panels have not been studied. Although the 1000 Genomes panels have much larger number of SNPs than the HapMap panels, no studies to our knowledge have been done to compare the imputation accuracies between the 1000 Genomes and HapMap panels using the three imputation methods. Secondly, the metrics for comparison of the imputation accuracies have been limited to using the proportion of concordance between the original and masked genotype calls across all the study samples and averaged over all the masked SNPs (total concordance). We hypothesize that this naive metric is not suitable to adequately differentiate the imputation accuracies between different panels and algorithms. Finally, none of the previous studies have examined dependencies of the imputation accuracy metrics on minor allele frequencies and evaluated the impact of each of the three genotypes - minor allele homozygotes, heterozygotes and major allele homozygotes, separately, on imputation performance.

In our study, we will systematically investigate the imputation performance of the widely used imputation methods - MACH, IMPUTE v2 and BEAGLE on ARIC African Americans using a variety of combinations of reference haplotypes from HapMap and 1000 genome reference panels. We will explore several imputation quality and imputation coverage metrics and also stratify them by minor allele frequencies to provide the research community with a comprehensive guideline of the trade-off between imputation coverage and imputation accuracy for African Americans on different reference panels and imputation algorithms.

5. Main Hypothesis/Study Questions:

Investigation of imputation performance of different combinations of reference panels from HapMap and 1000 Genomes using the popular imputation algorithms – MACH, IMPUTE v2 and BEAGLE to evaluate the dependency of imputation of the ARIC African Americans on minor allele frequencies and several imputation quality metrics.

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).

Study Design : Using each of the methods MACH, IMPUTE v2 and BEAGLE, we propose to investigate the following measures of imputation performance – Concordance Accuracy (CA), Kappa coefficient, imputation yield, and power of detecting imputed SNPs significantly associated with simulated phenotypes at a p-value threshold of 0.001. A variety of combinations of reference panels from HapMap Phase III and 1000 Genomes would be studied (Table 1). For each of the three chromosomes (18, 20 and 22), randomly chosen 10% of all the SNPs in the study sample will be masked by setting their genotypes to 0 (untyped or missing). The total number of SNPs and count of masked SNPs binned by allele frequencies for each chromosome is shown in Table 2. The commonly used statistic to measure quality of imputation is R^2, which represents the ratio of the empirically observed variance of allele dosage to the expected binomial variance at HWE.

For each imputation method and reference panel, we propose to measure the imputation performance in terms of the following statistics using the imputed SNPs that have R^2 greater than or equal to a given cutoff (R^2_{cutoff}):
(1) Concordance Accuracy: Degree of concordance between the observed genotypes from the study sample and the imputed genotypes at the masked SNPs. It is computed as the proportion of matches between genotype calls from imputed data and the study sample at the masked SNPs such that both genotypes are not missing.
(2) Kappa: The Kappa coefficient is used as a measure of overall concordance across the three genotypes using the masked SNPs.
(3) Yield: The yield of an imputation process is defined as the fraction of masked SNPs that have $R^2 \geq R^2_{cutoff}$.
(4) Power: A quantitative phenotype will be simulated for each of the masked SNPs with the fraction of variance explained set at 0.005. The power of each imputation is computed with the mean genotype of each SNP from the imputed data and the corresponding simulated phenotype as the fraction of masked SNPs detected as significant at a p-value threshold of 0.001. This allows us to quantitatively evaluate how the imputation performance affects the ability to detect SNPs significantly associated with a given trait.

Intuitively the SNPs with low allele frequencies are more difficult to impute accurately compared to those with both alleles available abundantly. Also at low MAF, the minor allele homozygotes and the heterozygotes would be rarer and more prone to prediction errors. We, therefore, propose to compute and report the above three performance measures for each of the three genotypes at the masked SNPs after binning by four allele frequency ranges (Table 2): ≤0.05, 0.05-0.1, 0.1-0.3 and > 0.3. Furthermore, for each round of imputation, 10 quality score cutoffs (R^2_{cutoff}) are chosen: 0.0-0.9 in steps of 0.1.

The above four quality measures computed at each of the cutoffs for each allele frequency bin will enable us to visualize the distribution of the quality scores and comprehensively evaluate the imputation methods and the reference panels for the study sample.

Exclusion:

1) individuals without GWAS data
2) individuals who did not consent to genetic research
3) self-reported race that is not “Black”

Table 1: Reference Panels to be used for imputation with each method.

Reference Panels	SNPs (Chr18/20/22)	haps
HapMap Phase III	40824/36258/20085	
ASW		126
YRI		230
CEU+YRI		464
ASW+YRI		356
ASW+CEU+YRI		592
1000 Genomes (Aug 2010)		
AFR	505592/ 400297/ 226376	348
EUR+AFR (consensus SNPs)	261388/213336/129064	904
1000 Genome (June 2010) + HapMap Phase III		
YRI (1000 Genomes) + All HapMap III	291232/229767/130229	2032

Table 2: Count of SNPs in the study sample and count of masked SNPs in brackets. Four allele freqeuncy bins are considered: MAF ≤ 0.05 (denoted ≤ 0.05), 0.05 < MAF ≤ 0.1 (denoted 0.05-0.1), 0.1 < MAF ≤ 0.3 (denoted 0.1-0.3) and MAF > 0.3 (denoted > 0.3).

Chr	Count of all SNPS (masked SNPs)				
	Total	≤0.05	0.05-0.1	0.1-0.3	>0.3
18	23613(2361)	2290(242)	3482(337)	10687(1060)	7154(722)
20	20471(2046)	2048(206)	3126(315)	9003(891)	6294(635)
22	10179(1017)	1237(120)	1549(160)	4447(460)	2946(277)

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 ICTDER03 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"? ___X_ 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)?

None**.**

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

11.b. If yes, is the proposal

___ A. primarily the result of an ancillary study (list number* _2006.03, 2008.09_________)

___ 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.cscc.unc.edu/aric/forms/

12. 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.