ARIC Manuscript Proposal #2790

PC Reviewed: 7/12/16	**Status: A**	**Priority: 2**
SC Reviewed: _________	**Status: _____**	**Priority: ____**

1.a. Full Title:
Practical Approaches for Whole Genome Sequence Analysis of Complex Traits

b. Abbreviated Title (Length 26 characters):
WGS practical approaches

2. Writing Group:
Alanna C. Morrison, Zhuoyi Huang, Bing Yu, Ginger Metcalf, Xiaoming Liu, Fuli Yu, Donna Muzny, Elena Feofanova, Navin Rustagi, Christie Ballantyne, Josef Coresh, Richard Gibbs, Eric Boerwinkle

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

First author: **Alanna C Morrison**
Address: Human Genetics Center
1200 Pressler Street, Suite E-447

Phone: 713-500-9913 Fax: 713-500-0900
E-mail: alanna.c.morrison@uth.tmc.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: Eric Boerwinkle
Address: Human Genetics Center
1200 Pressler Street, Suite W114
Houston, TX 77030

Phone: 713-500-9058 Fax: 713-500-9020
E-mail: eric.boerwinkle@uth.tmc.edu

3. Timeline:
Data are available, and analyses are to start as soon as possible. Manuscript is to be prepared as soon as analyses are completed.

4. Rationale:

Common complex traits, such as blood glucose and cholesterol levels, underlie some of the most common diseases burdening human health. Genetic analysis of these complex traits has followed the development of the fields of genetics and genomics, beginning with familial aggregation and linkage transitioning through candidate genes and genome-wide association studies (GWAS) to the emerging promise of whole genome sequencing (WGS). Declining costs have catalyzed accelerated adoption of WGS in large-scale genetics studies. However, few studies have utilized WGS to assess the contribution of low frequency and rare genetic variation to complex traits.

Morrison et al.(*1*) conducted WGS analysis of high-density lipoprotein cholesterol and was the first to describe initial steps for an unbiased and coordinated approach to evaluating WGS data in a population-based sample of European-Americans (EA). The UK10K Consortium also explored association testing of common, low frequency, and rare variants for quantitative traits using WGS data among European individuals.(*2*) These initial studies, along with the results of numerous GWAS studies (*3*), support the evaluation of non-coding regions in relation to complex quantitative traits, and also suggest that tests of association involving WGS would benefit from variant selection strategies that incorporate annotation of functional genomic elements. However, tests of association involving WGS are challenged by the large number of very rare variants, especially singletons (*4*), and tests that aggregate the cumulative effects of rare variants have been proposed and implemented.(*5*) These aggregate tests require an *a priori* defined region of the genome within which the combined effect of rare variants are assessed, and by far the most common units are the protein-encoding genes. WGS data offers the opportunity to aggregate variants over the full spectrum of annotated motifs, from specifically defined regulatory domains to an agnostic sliding window. In this study, we offer a practical approach to WGS analysis of complex traits using aggregate tests across a variety of annotated functional motifs. In addition, we consider weighted analyses using nucleotide specific information, and provide guidance on p-values for defining thresholds of statistical significance. Because previous applications have focused on samples from populations of European-descent, we provide an example application in a sample of African Americans (AA) measured for multiple heart- and blood-related traits.

5. Main Hypothesis/Study Questions:

Low frequency and rare variation residing in annotated genomic motifs are associated with heart- and blood-related phenotypes.

The field will benefit from a practical guide to whole genome sequence analysis of complex traits.

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).

Inclusion: Those with low-pass WGS, ten phenotypes of interest (including circulating neutrophil count, platelet count, and levels of hemoglobin, lipoprotein(a) (Lp(a)), magnesium (Mg) and phosphorus (P), small dense low-density lipoprotein cholesterol (sdLDL-C), C-reactive protein (CRP), highly sensitive cardiac troponin T (hs-cTnT), and N-terminal pro–B-type natriuretic peptide (NT-proBNP)), and informed consent.

Exclusions: Obvious phenotypic outliers, prevalent coronary heart disease and heart failure cases for some analyses. WGS data has been assessed for quality control and quality assurance measures at the Baylor College of Medicine Human Genome Sequencing Center and as a part of the CHARGE Analysis and Bioinformatics Working Group.

Analysis methods: Each of the 10 cardiovascular risk factor traits will be analyzed separately. Because our primary focus is on low frequency and rare variant sequence analysis, analyses within annotated functional motifs (including promoter, enhancer, 5' and 3' UTRs and first intron) will contain only low frequency and rare variants (MAF ≤ 5%). Within each annotated functional motif, a burden test (T5 (*6*)) and the Sequence Kernel Association Test (SKAT (*7*)) will be used adjusting for age, sex and the first three principal components (PCs), with additional adjustment of body mass index (BMI) for CRP and current smoking status (yes or no) for neutrophil counts. The T5 test collapses variants with MAF ≤ 5% into a single genetic score, while SKAT allows for effects of single nucleotide variants (SNVs) in both directions. In addition, we will analyze low frequency and rare variants using a sliding window approach (4 kb window length, 2 kb skip length). For completeness, we will also conduct an additional survey of the genome investigating all variants with MAF > 5%, evaluated individually, using an additive genetic model with the same adjustments. All analyses will be carried out using the R seqMeta package.(*8*)

Statistical significance: Significance levels will be established based on specific hypotheses and take into account the number of tests. For example, a p-value of 7.5×10^{-8} will be utilized taking into account ~700,000 contiguous sliding windows across an entire genome and utilizing Bonferroni correction. Outline of practical significance thresholds will be a part of the results.

7.a. Will the data be used for non-CVD analysis in this manuscript? ____ Yes __X__ 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? _____ 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"? __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)?

MS# #1530B Morrison et al., Evaluation of whole genome sequence in relation to quantitative traits.

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

__X__ A. primarily the result of an ancillary study (list number* 2009.12)
___ 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/

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.
Agree.

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.
Agree.

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.

References:

1. A. C. Morrison *et al.*, Whole-genome sequence-based analysis of high-density lipoprotein cholesterol. *Nat Genet* **45**, 899-901 (2013).

2. U. K. Consortium *et al.*, The UK10K project identifies rare variants in health and disease. *Nature* **526**, 82-90 (2015).
3. M. L. Freedman *et al.*, Principles for the post-GWAS functional characterization of cancer risk loci. *Nat Genet* **43**, 513-518 (2011).
4. F. Yu *et al.*, Population genomic analysis of 962 whole genome sequences of humans reveals natural selection in non-coding regions. *PLoS One* **10**, e0121644 (2015).
5. S. Lee, G. R. Abecasis, M. Boehnke, X. Lin, Rare-variant association analysis: study designs and statistical tests. *Am J Hum Genet* **95**, 5-23 (2014).
6. B. Li, S. M. Leal, Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. *Am J Hum Genet* **83**, 311-321 (2008).
7. M. C. Wu *et al.*, Rare-variant association testing for sequencing data with the sequence kernel association test. *Am J Hum Genet* **89**, 82-93 (2011).
8. http://cran.r-project.org/web/packages/seqMeta/index.html.