ARIC Manuscript Proposal #2107

PC Reviewed: 4/9/13	**Status: A**	**Priority: 2**
SC Reviewed: _________	**Status: _____**	**Priority: ____**

1.a. Full Title: Analysis of Sequencing Studies Under Multivariate Trait-Dependent Sampling

b. Abbreviated Title (Length 26 characters): Multivariate Trait-Dependent Sampling

2. Writing Group: Ran Tao, Donglin Zeng, Nora Franceschini, Kari E. North, Eric Boerwinkle, Danyu Lin.

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

First author:

Ran Tao,
Ph.D. candidate
Department of Biostatistics
University of North Carolina, Chapel Hill
3103D McGavran-Greenberg Hall
Chapel Hill, NC 27599-7420
(919) 381-8227 (Phone)

Corresponding/senior author (if different from first author correspondence will be sent to both the first author & the corresponding author):

Kari North
UNC Gillings School of Global Public Health
137 E. Franklin St., Ste. 306
Chapel Hill, NC 27514
(919) 966-2148 (Phone)
(919) 966-9800 (Fax)
Email: kari_north@unc.edu

3. Timeline: Analyses will begin upon approval of this manuscript proposal.

4. Rationale:

To develop novel statistical methodology to analyze sequencing data under multivariate quantitative trait dependent sampling design. The ARIC data in CHARGE Targeted

Sequencing Study will be used as a real data example to compare the new method to standard linear regression method ignoring this sampling design.

This study proposes a very general statistical framework to properly analyze data generated under multivariate trait dependent sampling design, with potential application to any type of studies incorporating such design. We plan to demonstrate the advantage of our method in the context of sequencing association studies and, in particular, the CHARGE Targeted Sequencing Study, which motivated the development of this statistical method due to the sample design.

5. **Main Hypothesis/Study Questions**:

Multivariate trait dependent sampling design is a design strategy to preferentially sequence the subjects with the extreme values of a number of quantitative traits, which is cost effective to conduct sequencing association studies. In this case, standard linear regression analysis (univariate or multivariate) can result in bias of parameter estimators and loss of power. We develop a novel statistical model that can properly reflects the sampling mechanism and achieve high efficiency.

To be more specific, we construct a semiparametric likelihood that properly reflects the sampling mechanism. In our formulation, quantitative traits are related to genetic variables and covariates through a multivariate linear regression model while the distributions of genetic variables and covariates are completely arbitrary. We develop a novel EM algorithm to maximize the likelihood and establish the consistency, asymptotical normality and asymptotic efficiency of the resulting estimators. Simulation studies demonstrate the superiority of the proposed method over standard linear regression method (both univariate and multivairate). Upon approval of this manuscript proposal, an application to ARIC data in CHARGE Targeted Sequencing Study will be included as a real data example.

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).

In general, this is a project focusing primarily on statistical methodology rather than epidemiological findings. The ARIC data in CHARGE Targeted Sequencing Study will be used only as an example for application of our new method. Claiming any new epidemiological findings from analyzing this data is beyond the scope of this project.

Subjects: 9103 ARIC cohort participants of European ancestry, who gave informed consent for genetic data usage, had sufficient DNA for sequencing, and had Principle Components (PC) calculated in previous CHARGE GWAS studies will be included. Among them, 1927 individuals had CHARGE targeted sequencing data. For these sequenced subjects, information on the sequencing genotype profile, all phenotypes involved in sampling, and a number of covariates will be used. For the remaining non-sequenced subjects, only information on the phenotypes will be used.

Phenotypes:

1) ECG PR interval
2) ECG QRS interval
3) BMI
4) Blood pressure
5) C-reactive protein level
6) Carotid intima-media thickness
7) Fast insulin
8) Hematocrit
9) Pulmonary function
10) Retinal venule diameter

Retinal venule diameter is obtained from visit 3 data. All other phenotypes are obtained from the baseline visit (visit 1).

Exclusions: Subjects not of European ancestry, who did not give informed consent for genetic data usage, who did not have sufficient DNA for sequencing, and who did not have the GWAS PCs will be excluded.

Exposure: Genetic variants are called from a targeted sequencing of 77 pre-selected genomic regions, which encompassed approximated 2 megabases of the genome. These regions have been previously shown to be associated with one or more of the phenotypes.

Model: We construct a semiparametric likelihood that properly reflects the sampling mechanism. In our formulation, quantitative traits are related to genetic variables and covariates through a multivariate linear regression model while the distributions of genetic variables and covariates are completely arbitrary. We will perform analysis using all available SNPs that pass QC and have a minor allele frequency larger than 5%.

Covariates: age at visit 1, gender, ARIC center indicators, first 5 PCs.

Statistical significance: Epidemiological findings will not be the focus of this project. Results will be presented only to serve as a comparison of the performances of our proposed method and the standard linear regression method.

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 ICTDER02 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 ICTDER02 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 ICTDER02 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)?

We did not identify a statistical methodology project on multivariate trait dependent sampling in sequencing studies.

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.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.