ARIC Manuscript Proposal #2348

PC Reviewed: 4/8/14	**Status: A**	**Priority: 2**
SC Reviewed: _________	**Status: _____**	**Priority: ____**

1.a. Full Title: Comparison of functional prediction methods for nonsynonymous SNPs in exome sequencing studies of human diseases

b. Abbreviated Title (Length 26 characters): SNP prediction comparison

2. Writing Group:

Writing group members: Chengliang Dong, Peng Wei, Xueqiu Jian, Richard A. Gibbs, Eric Boerwinkle, Kai Wang and Xiaoming Liu

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

F**irst author**: Chengliang Dong
Address: Zilkha Neurogenetic Institute, 1501 San Pablo Street, ZNI 221, Los Angeles, CA 90089

Phone: Fax:
E-mail: chenglid@usc.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: Xiaoming Liu
Address: 1200 Herman Pressler ST, E529, Houston, TX 77030

Phone: (713) 500-9820 Fax: 7135000900
E-mail: xiaoming.liu@uth.tmc.edu

3. Timeline: Completion of the manuscript is anticipated in April 2014.

4. Rationale:

Functional prediction methods for non-synonymous SNPs are commonly used in sequencing-based Mendelian disease studies. However, as different methods are based on different gene or SNP features, each has its own advantages and disadvantages. We propose a large-scale comparison of those methods using known SNPs that causing Mendelian diseases and those observed in cohort populations and unlikely causing

Mendelian diseases. Based on the comparison results, we propose to build an ensemble score based on multiple prediction scores to improve the prediction accuracy.

Exome sequencing data of the ARIC cohort random samples will be used to identify non-synonymous SNPs that have not been reported previously (e.g. 1000 genomes project or ESP project). Part of those SNPs will be used as a testing set for comparing different functional prediction methods.

5. Main Hypothesis/Study Questions:

The main study questions are (1) to compare functional prediction methods/scores on the accuracy for predicting the potential that SNP will cause damaging effect for the gene function. (2) to demonstrate the value of combining information from multiple orthologous approaches to achieve better prediction accuracy.

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).

We collected non-synonymous SNPs that causing Mendelian disease and those annotated as polymorphism from the Uniprot database. This dataset will be training dataset for our ensemble scores.

We collected non-synonymous SNPs that are not overlapping with our training data set from the VariBench database. This dataset will be the major testing data set.

Additional testing data set of disease causing non-synonymous SNPs was collected from recent literatures (2011-2013).

Additional testing data set of non-disease causing non-synonymous SNPs was collected from exome sequence data of ARIC cohort random samples from the freeze 3 CHARGE-S data. We used the SNPs after QC. Mapping quality and minor allele frequency were used to further filter SNPs. No individual-level phenotype data or genotype data were used in the study.

We used support vector machine and logistic regression to construct our ensemble scores. Prediction performance was compared using ROC curves and AUC.

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"?
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)?
No related manuscript proposal found.

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* _________)**
x **B. primarily based on ARIC data with ancillary data playing a minor role (usually control variables; list number(s)*** 2009.12**)**

*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.