ARIC Manuscript Proposal # 1185

PC Reviewed: _08__/_15__/06	**Status: __A___**	**Priority: __2__**
SC Reviewed: _08__/_17__/06	**Status: __A__**	**Priority: __2__**

1.a. Full Title:

Use of a Random Forests classifier for variable selection in large-scale genomic association studies.

b. Abbreviated Title (Length 26 characters):

RF applied to Chr 19 dataset.

2. Writing Group:
Writing group members:

Andrei S. Rodin, Anatoliy Litvinenko, Kathy Klos, Alanna C. Morrison, Trevor Woodage, Josef Coresh, Eric Boerwinkle, and other ARIC investigators as desired.

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

First author: Andrei S Rodin
Address:

Human Genetics Center
PO Box 20186
University of Texas
Health Science Center
Houston, TX 77225

Phone: 713-500-9845 Fax: 713-500-0900
E-mail: arodin@uth.tmc.edu

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

Eric Boerwinkle
Human Genetics Center
PO Box 20186
University of Texas
Health Science Center
Houston, TX 77225

Phone: 713-500-9816 Fax: 713-500-0900
E-mail: eboerwinkle@uth.tmc.edu

3. Timeline:

All analyses have been completed. A manuscript will be completed by July 2006.

4. Rationale:

Traditional statistical genetics analysis methods might be ill-equipped to deal with highly multidimensional datasets common to modern genetic epidemiology research. Specifically, a vast amount of potentially interacting weak-effect variables makes it imperative that some *variable selection* procedure is carried out before the actual analysis, for reasons of both computational efficiency and *overfitting*. This is a common practice in *data mining* research. A typical data mining analysis strategy consists of three steps: variable selection (or *feature set reduction*), model building (in context of genetic epidemiology, reverse-engineering of the relationships between predictive variables, genetic and otherwise, and phenotypes of interest) and model validation (by means of traditional statistical hypothesis testing, and/or replication). In this manuscript we aim to explore the first step, namely feature set reduction, as applied to a highly dimensional Chr 19 dataset generated by the ARIC project.

By using a *wrapper* variable selection scheme, we show that the vast majority of potentially predictive variables (SNPs) can be safely removed from the dataset without compromising the predictive accuracy. A highly scalable classifier, *Random Forests*, is used to track the *generalization* (unbiased) predictive accuracy, by means of internal bootstrap resampling and/or external-loop cross-validation. In addition, a statistical *Set Association* approach (Hoh and Ott, 2001) is also applied to the dataset in order to confirm the concordance between the "pruned" SNP sets resulting from different feature set reduction procedures. Subsequently, we apply a specialized *genetic algorithm* to further optimize the resulting pruned SNP sets. Finally, we compare SNP rankings obtained by different (both computer science and statistical) methods and devise practical guidelines for researchers trying to generate a compact and highly predictive subset of SNPs from the highly multidimensional genome-wide datasets.

While the second (model building) step of our analysis strategy is beyond the scope of this manuscript, we will detail the general rationale behind the strategy, and discuss the phenomenon of overfitting. We will also introduce the concept of *real and synthetic positive controls* (used to ascertain the analysis methods' performance) and illustrate its utility on example of ApoE isoform-coding SNPs and an intermediate (LDL) phenotype.

5. Main Hypothesis/Study Questions:

There is no biological "main hypothesis" *per se*. Rather, this study will investigate the applicability of our variable selection algorithm to large-scale genomic data. We will show that only a small portion (sometimes less than 2%) of the typed SNPs in genome-wide association studies are potentially predictive with respect to a complex disease phenotype and that, with care, the remaining SNPs can be safely excluded from the analysis --- based not on the arbitrary cut-off threshold, but rather on the amount of signal contained in the dataset itself. The main technical goal of this manuscript is to test the applicability of a Random Forests classifier to a typical genome-wide dataset.

Additional study questions are as follows: (1) compare bootstrapping and external-loop cross-validation in context of variable selection procedure, (2) compare pruned SNP sets (and rankings) obtained by different, computer science- and statistically- motivated methodologies, (3) briefly introduce a three-step analysis strategy and a genetic algorithm, and (4) devise practical guidelines for reconciling heterogeneous analysis methods in the first step (feature, or SNP, set reduction) of our analysis strategy.

6. Data (variables, time window, source, inclusions/exclusions):

The Chr 19 dataset is based upon a case-cohort design. CHD Cases (N=1,476) were defined by documented MI, unstable angina, CHD death and/or cardiovascular procedures within a 10-year span (1988-1998). Prevalent cases were excluded. A stratified random sample of non-cases from the ARIC cohort was selected as the control reference group (designated as the cohort random sample, or CRS (N=949)). The CHD case group is composed of 1,142 non-Hispanic whites and 334 African-Americans. The CRS reference group consists of 618 non-Hispanic whites and 331 African-Americans. SNPs selected for the chromosome 19 study were identified according to a gene-based method. Specifically, approximately 1,000 genes were identified on chromosome 19, followed by identification of variation (SNPs) on chromosome 19, and these two processes were combined to identify all SNPs within genes (~90%) plus those SNPs that are in tracts for which genes have not yet been identified. After assay development and SNP validation, SNP laboratory quality control measures were satisfactory for an average of six SNPs per gene. SNP genotyping was completed using the TaqMan and SNPlex genotyping platforms (Applied Biosystems, Foster City, CA). At the time of the actual analyses, a dataset containing 4,813 verified and validated SNPs was available. Additional variables included BMI, age, gender and blood lipids (such as HDL)

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)?

N/A

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. primiarly 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.