ARIC Manuscript Proposal #3459

PC Reviewed: 9/10/19 **Status: _____** **Priority: 2**
SC Reviewed: _________ **Status: _____** **Priority: ____**

1.a. Full Title: Omics, cardiac electrophysiology, and arrhythmias: the ARIC study

b. Abbreviated Title (Length 26 characters): **Omics and ECGs**

2. Writing Group:
Writing group members: ThuyVy Duong, Dan Arking, Joel Bader, Alexis Battle, Jen Brody, Nona Sotoodehnia, Alvaro Alonso, Joe Coresh, Alanna Morrison, Eric Boerwinkle, Faye Norby, Lin Yee Chen, others are welcome to join.

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

First author: **ThuyVy Duong**
Address: 733 N. Broadway, MRB 420
Baltimore, MD 21205

Phone: (443) 287-4434 Fax: 410-614-8600
E-mail: vduong4@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, MRB 459
Baltimore, MD 21205

Phone: 410-502-4867 Fax: 410-614-8600
E-mail: arking@jhmi.edu

3. Timeline: ARIC analyses to be completed in Spring 2020, meta-analyses in spring 2021, and manuscript to be submitted in Fall 2021

4. Rationale:

Application of multi-omics to the study of sudden cardiac death (SCD), and related electrocardiographic endophenotypes (e.g. QT interval, PR interval, QRS interval, RR interval, P wave indices) and arrhythmias (e.g. atrial fibrillation [AF]), has been limited. Multi-omics approaches could potentially identify novel biomarkers of SCD and provide new clues regarding the etiopathogenesis of this lethal arrhythmia. We propose to take advantage of the ARIC proteomics, methylomics, and RNAseq data obtained across multiple visits and explore their association with electrocardiographic measures as well as incident AF and SCD. These results will be integrated with additional cohorts obtained under the TOPMed auspices (JHS, WHI, SOL, CHS, MESA, FHS, Amish).

5. Main Hypothesis/Study Questions:

The primary aim of this analysis is to study the association of multi-omics data (methylation, transcript, proteome) with incident AF, incident SCD, and related electrocardiographic 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).

We fully recognize that the omics data are available on different sets of ARIC individuals and at different visits, and will fully take that into account in our analyses. Below is a general outline of the approach we will take.

Subjects: For each ‘omics data analysis, we will use all ARIC participants with the specific ‘omics data available (ranging from ~200 for transcriptome to all individuals for proteomics).

Data: TOPMed Multi-omics data including whole-genome sequencing variant calls, methylation, RNA-seq, proteomics (SomaLogic). For the proteomics data, we will focus on Visit 3 data in order to have sufficient incident SCD and AF cases for analysis. For ECG analysis, we will explore using mixed models to allow us to incorporate both visit 3 and visit 5 data.

Main Outcome Variables:

The primary event outcome will be incident SCD and atrial fibrillation during follow-up.

The second set of outcome variables will be EKG measurements (including PR, QRS, QT, and RR intervals, and P wave indices).

Covariates and exclusion variables: Age, race/ethnicity, sex, field center, height, weight, SBP/DBP, antihypertensive medication, diabetes status, CVD status, pacemaker, Wolff-Parkinson-White (WPW) syndrome, pregnancy, history of myocardial infarction, heart failure, 2nd or 3rd degree AV block, use of class I and/or class III antiarrhythmic drugs, and digoxin use. We will also include covariates that affect the various ‘omics measures, including kidney function and smoking status.

Brief Methods/Analysis Plan:

Analysis for Aim 1: We will examine incident SCD/AF in relation to each omics measure (methylation, transcription, or protein levels). We will use Cox proportional hazards regression models to model the relationship between SCD/AF and each individual omics level, starting with the earliest exam for which omics data is available, and updating to later exam values for time accrued after the initial exam (i.e. we will use the omics value closest in time preceding the event). Analyses will be adjusted for age, sex, field center, and genetic principal components. Primary analyses will combine samples across all racial/ethnic groups, in order to maximize sample size. Secondary analyses will be performed separately by race/ethnic group. Participants with prevalent AF at the baseline visit will be excluded. Analyses of methylation and transcription levels will be additionally adjusted for cell count compositions, as estimated by the Houseman method, or by use of directly measured cell counts (if available). We will use surrogate variable analysis to account for unmeasured artifacts, as needed. To control for type I error, analyses will be corrected for the number of omics levels tested.

Analysis for Aim 2: We will examine each omics measure (methylation, transcription, or protein levels) in relation to each EKG measurement (including PR, QRS, QT, RR intervals, and P wave indices). Analyses will be run using generalized estimating equations (GEE) to account for individuals with EKG and omics levels measured at multiple time points. Participants will be excluded from analyses based on the following criteria: AF on EKG, Wolff-Parkinson-White (WPW) syndrome, pacemaker, pregnancy, myocardial infarction, heart failure, 2nd or 3rd degree AV block, use of class I and/or class III antiarrhythmic drugs, and digoxin use. Analyses will be adjusted for age, sex, field center, RR interval (except in the RR analysis), BMI, and genetic principal components. Analyses of methylation and transcription levels will be additionally adjusted for cell count compositions, as estimated by the Houseman method. As needed, we will use surrogate variable analysis to account for unmeasured artifacts.

All analyses will be corrected for the number of omics levels times the number of outcomes tested (e.g., methylation analyses will be corrected for the number of methylation probes times five outcomes [QRS, QT, RR, PR and P wave indices]).

Detailed analysis plan including all participating cohorts (does not explicitly include proteomics data, but that will be handled in a similar manner):

Aim 1: To test directly the association of novel omics measures separately with electrophysiologic endophenotypes and arrhythmia risk. In cross-sectional design, we will examine the association of WGS data (n>28,000), methylation data (n>14,000), and transcriptomic data (n>7,000) with electrophysiologic endophenotypes and with arrhythmia risk (n~4300 AF and ~400 SCD).

We will perform agnostic genome-wide scans for each omic data type - sequence variation, methylation levels and transcription levels - separately for each electrophysiologic phenotype. We will perform both single-variant tests for common variants (minor allele frequency [MAF] > 1%) and aggregate tests for rare variants. Analyses of DNA methylation levels will test for association with each CpG separately. Transcription levels are assayed either through RNAseq or Affymetrix GeneChip Human Exon 1.0 ST Arrays. We will analyze transcription levels separately by platform and meta-analyze results among the approximately 18,000 overlapping genes. Analyses of rare variants are sensitive to model assumption of normality. For continuous traits (PR, QRS, QT, RR, P wave indices, PACs and PVCs), we will apply an inverse normal transformation on trait residuals from the adjusted models. For incident analyses, we will use Cox-proportional hazards models and their adaptions to SKAT tests. In addition to accounting for population and familial relatedness, all analyses will adjust for age, sex, height, study, assay center and heart rate where appropriate. Both methylation and transcription levels are sensitive to technical artifacts that may not be completely captured by recorded variables. We will use surrogate variable analysis to account for unmeasured artifacts. Methylation analyses will additionally be adjusted for cell type counts where available or through estimated cell type counts using the Houseman method.

We will have good power to discover associations with a range of effect sizes. For methylation analyses using a Bonferroni correction for the number of tests (0.05/450,000), we have 80% power to detect a 1.1 millisecond (ms) change in QT interval for a one standard deviation change in methylation. For transcription analyses with 18,000 tests, we have 80% power to detect a 1.5 ms change in QT. For rare variants, gene-based tests aggregating on functional and genomic annotation will be employed to improve power to detect associations. Using a Bonferroni correction for 40,000 tests, for an aggregate test with a 2% cumulative minor allele frequency, we have 80% power to detect a 3.5 ms change in QT per copy of the alternate allele. With 12 million single-variant tests, we have 80% power to detect a 1.3-1.9 ms change in QT, given minor allele frequencies ranging from 10-30%. Because power for the clinical outcomes of AF and SCD is more limited, the primary analysis will be to examine the endophenotype findings for association with AF and SCD (e.g. PR CpGs for association with AF). However, we will also examine these clinical phenotypes genome-wide.

Aim 2: To identify novel variants, genes, and pathways associated with cardiac electro-physiologic phenotypes, using a multi-pronged systems biology approach. In **Aim 2A,** first we will identify pathways associated with cardiac electrophysiology by integrating multi-omic data from samples with ECG measures using machine learning approaches. We will first use our pathway-based omics data integration approach to identify pathways associated with electrophysiologic phenotypes. Within these enriched pathways, in **Aim 2B**, we then use an extension of our NetGSA framework together with our recently proposed kernel-penalized regression to integrate genomic and omic data in order to identify genes as well as genetic and omic variation within those genes, associated with electrophysiologic phenotypes. In **Aim 2C**, we will evaluate the clinical relevance of our findings from **Aim 2B** by examining their association with the clinical phenotypes of AF and SCD.

In the first stage, biological pathways associated with ECG and arrhythmias are identified using a rank-based pathway-centric integration of genomics, epigenomics and transcriptomics measurements. Briefly, we independently rank the biological pathways (obtained, e.g., from MSigDB) according to their enrichment score based on each omics measurement. The independent rankings are obtained using methods most suitable for each omics measurement, and incorporate all available samples for each omics data type. For instance, we will use GSEA/GSA to obtain pathway rankings based on all TOPMed samples with WGS data. For transcriptomics, we will instead use our recently proposed penalized kernel regression method, which allows us to incorporate the knowledge of gene regulatory interactions. Our kernel-penalized regression can be considered a generalization of ridge regression that allows one to incorporate the knowledge of gene regulatory interactions. Specifically, it amounts to solving a penalized least-square criterion of the form $|| y - Xß ||_H + \mu ||ß||_Q$, where the ‘data’ kernel, H, captures similarities among subjects (for instance through genetic distance), and the ‘parameter’ kernel, Q, captures information from gene regulatory interactions; here µ is a tuning parameter, chosen via cross validation, which controls the degree of penalization. Our kernel-penalized regression method generalizes the SKAT framework to, e.g. incorporate gene regulatory network information. Next, the separate rankings for each pathway are integrated into a single *multi-omics enrichment score* using a resampling (bootstrap) approach that accounts for the intrinsic variability of each omics data source.

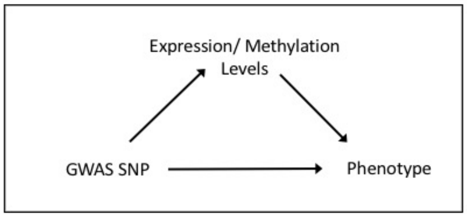
In addition to facilitating the use of all available samples for each omics type, the pathway-centric integration of multi-omics data in this approach reduces the number of hypotheses and improves the power for detecting enriched pathways. We recently used the above rank-based omics integration to identify pathways associated with prostate cancer progression based on joint transcriptomic and metabolomics data. In addition to identifying highly enriched pathways based on each omics platform, this approach also identified novel pathways that demonstrate modest but orchestrated enrichment in each of their omics measurements.

Upon identifying enriched pathways based on the multi-omics integration of the first stage, the second stage of our analysis framework (**Aim 2B**) includes network-based integration of genomics, epigenomics and transcriptomics measures. More specifically, we will use a generalization of our NetGSA framework to model not only gene regulatory interactions in each of the enriched pathways, but also the genetic and epigenetic effects corresponding to each gene. Through detailed modeling of biological networks, this approach accounts for complex interaction mechanisms among all pathway components, as well as genetic and epigenetic effects. While computational and sample size constraints limit the applicability of such an approach at the whole-genome level, our two-stage framework significantly reduces the computational and statistical complexity of the problem by first identifying enriched pathways.

In **Aim 2C,** we will assess the clinical relevance of variants and genes identified in **Aim 2B**. Specifically, we will examine the findings for PR variants and genes for association with incident AF, and findings for QRS and QT interval variants and genes with SCD in a hypothesis driven approach. For specific variants or CpGs, these analyses will be performed using a modified Cox proportional hazards models for case-cohort data, adjusting for age, sex, PCs, and study site. For genes identified through burden/collapsing methods, the set of variants contributing to the signal will be treated as a single variable, and assessed through the modified Cox model. Given the rich covariate data available in each cohort, secondary analyses adjusting for traditional cardiovascular risk factors will be performed to assess whether any of the genetic effects are mediated through these pathways, and whether interactions, such as with age and sex, are present.

Aim 3: To identify causal genes within known GWAS loci. We will examine the association of electrophysiologic phenotypes separately with gene transcription and methylation levels within previously identified GWAS PR, QRS, and QT loci. We will perform mediation analyses to determine if methylation and transcription levels at gene-regulatory sites mediate the effect of GWAS SNPs on ECG phenotypes within the loci. We will also examine GWAS SNPs for association with expression in cardiac tissues in the GTEx data.

Figure 1. Each omic will be tested as a mediator of the GWAS SNP on its associated phenotype.

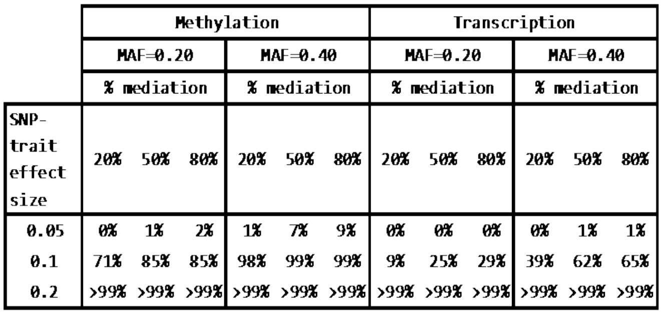


GWAS have identified numerous electrophysiological trait loci. However, rarely is the causal gene or mechanism identified. By analyzing samples that have the genetic variants, phenotypes of interest, and gene-based molecular intermediates such as transcription levels or methylation levels that likely affect transcript levels, we aim to identify which genes are responsible for the GWAS signals. The primary statistical method will be mediation analysis, which can explore how much of the GWAS signal is accounted for by variation in methylation or transcription levels associated with nearby genes (**Figure 1)**. Methylation and transcription levels are measured on a large number of samples from blood. Since the likely cell-types of interest for cardiac phenotypes are in the heart, a parallel analysis will look up each GWAS index SNP and identify any expression quantitative trait loci (eQTL) in atrial appendage or left-ventricle expression data from GTEx where we can examine genotype and transcript levels, though we will not have ECG measures on these samples. Positive mediation signals from blood transcript and methylation analyses described below will be interpreted in the context of eQTLs identified in cardiac tissues.

We will examine local cis- effects for each index SNP by selecting all transcripts within 1MB to examine for mediation of the SNP effect. For those genes, we will also examine for mediation all CpG sites that are within 2000bp of the transcription start site, the first exon of each gene, gene body, 3’ UTR and 5’ UTR. Some CpGs in these regions could possibly act through regulation of adjacent genes.

Mediation analyses using methylation and transcription levels compares statistical models for association between SNP and ECG phenotype that are 1) adjusted and 2) unadjusted for the methylation or transcription level. Specifically, linear mixed models that account for relatedness and population structure via the GRM and PCs will be used to evaluate both adjusted and unadjusted associations between the SNP and the electrophysiologic phenotype. Letting β_U be the SNP coefficient for the unadjusted model and letting β_A be the SNP coefficient for the adjusted model, the ratio $(\beta_U - \beta_A)/\beta_U$ measures the proportion of SNP-phenotype association that is mediated by methylation at the CpG site or transcription of the gene. Because the two components of this proportion come from different models fit to the same data, the bootstrap method is a convenient way to form confidence intervals for this proportion and test whether it is different from zero. CpGs with significant mediation effects will be tested for association with gene transcript levels among the subset of samples with both methylation and expression levels. GWAS SNPs that do not act directly through methylation or transcription levels (e.g. coding variation) may nonetheless be modified by the regulation of gene expression; we will use interaction analyses to examine this hypothesis.

As an example, we provide power calculations for the QT interval GWAS loci (**Table 1)**. There are on average 22 genes within the 1MB window around the index SNPs, with 14 CpGs per gene. The alpha level will correct for the number of tests per phenotype, p=6.5e-5 (0.05/770) for transcription mediation analyses and 4.6e-6 for methylation mediation analyses (0.05/10780). We have good to excellent power to detect mediation for the GWAS SNPs with larger (>= 0.1 standard deviation change) in the phenotype.



	Methylation						Transcription					
	MAF=0.20			MAF=0.40			MAF=0.20			MAF=0.40		
	% mediation			% mediation			% mediation			% mediation		
SNP-trait effect size	20%	50%	80%	20%	50%	80%	20%	50%	80%	20%	50%	80%
0.05	0%	1%	2%	1%	7%	9%	0%	0%	0%	0%	1%	1%
0.1	71%	85%	85%	98%	99%	99%	9%	25%	29%	39%	62%	65%
0.2	>99%	>99%	>99%	>99%	>99%	>99%	>99%	>99%	>99%	>99%	>99%	>99%

Table 2. Mediation Power. Effect sizes are standard deviations of the trait.

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/mantrack/maintain/search/dtSearch.html

___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)?
MP#3189 Metabolomics, cardiac electrophysiology, and sudden cardiac death: the ARIC study
MP#3398 Proteomics and the Risk of Incident Atrial Fibrillation in the Elderly: The Atherosclerosis Risk in Communities (ARIC) study
Lead authors of related manuscripts are included in the author list.

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* _2017.27, 2013.22, 2013.12________)
___ 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 https://www2.cscc.unc.edu/aric/approved-ancillary-studies

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.

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.