1.a. **Full Title**: Title: Exome Meta-Analysis of Drinking and Smoking (EMADS)

b. **Abbreviated Title (Length 26 characters)**: Meta-Analysis of Drinking and Smoking.

2. **Writing Group**: Kari North, Nora Franceschini, Linda Kao, Jim Pankow, Eric Boerwinkle, Megan Grove, and other interested ARIC Authors.

Many other EMADS Authors will also be included in this paper. These Authors will all come from the other partner studies.

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

**First author**: Kari E. North, Ph.D.
Associate Professor
Department of Epidemiology and Carolina Center for Genome Sciences
University of North Carolina Chapel Hill
Bank of America Center
137 E. Franklin St., Suite 306
CB #8050
Chapel Hill, NC 27599-8050
(919) 966-2148 (Voice)
(919) 966-9800 (Fax)
kari_north@unc.edu

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

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

4. **Rationale**: Evidence from twin, family and adoption studies has consistently shown genetic factors to be linked to sizeable individual differences in smoking. First reported by Fisher, concordance rates for smoking have been shown to be higher for MZ than DZ twins. 1,2 Numerous studies have further confirmed these population differences. 3-15 The
heritability estimates for smoking initiation, the Fagerström Test for Nicotine Dependence (FTND) score, and smoking cessation range from 0.50-0.80. Of the variables that comprise FTND score, the average number of cigarettes smoked per day (CPD), and time to first cigarette upon waking exhibit the highest heritability estimates. Although genome-wide linkage studies have identified some genomic regions that predispose or protect against nicotine dependence, the data are not consistent. Similarly, candidate gene association studies of smoking have not been successful in identifying replicable associations and are dependent upon our limited knowledge of addiction biology.

Thorgeirsson and colleagues identified a signal nucleotide polymorphism (SNP), rs1051730, and other cholinergic pathway SNPs within the nicotinic acetylcholine receptor gene cluster (CHRNA5/CHRNA3/CHRNB4) cluster on chromosome 15q24-25.1, associated with both smoking quantity (cigarettes/day; SQ). These results have been replicated in European-ancestry smokers in GWAS datasets of ~1,000 cases and >30,000 controls from Iceland, Spain, the Netherlands, other European nations and the United States by the European Network of Genomic and Genetic Epidemiology (ENGAGE) Consortium. Moreover, a large (>70000 European-ancestry samples) genome-wide meta-analysis by the Tobacco and Genetics (TAG) Consortium confirmed an association between two SNPs in the CHRNA5/CHRNA3/CHRNB4 gene cluster with SQ for rs1051730 and rs1696998 (ps~1x10\(^{-33}\)) as well as genome-wide significant associations with other genes implicated in nicotine dependence and smoking cessation.

Using data from the Women’s Health Initiative (n = 8208) and twelve other study groups forming the Study of Tobacco Use in Minority Populations (STOMP) Genetics Consortium (N = 32829), we conducted the first genome-wide meta-analysis of smoking behaviors in African Americans and have published our results in *Translational Psychiatry* (WHI MS984), with a related cross-population study of chromosome 15q24-25.1 in press in *Genetic Epidemiology* (MS1453). We identified one non-coding SNP (rs2036527) on chromosome 15q25.1 associated with smoking quantity (cigarettes per day - SQ) that exceeded genome-wide significance (β = 0.040, standard error (s.e.) = 0.007, \(P = 1.84 \times 10^{-8}\)) in female and male smokers in a genome-wide meta-analysis of 13 studies, including more than 8000 African-American WHI SHARE females. After back transformation of the beta estimate, mean CPD values for each rs2036527 genotype were 14.6 for AA, 13.5 for AG and 12.8 for GG, suggesting that there is an increase of less than 1 cigarette smoked per day for each copy of the A allele. This SNP accounted for approximately 0.20% of the phenotypic variance of CPD in our sample. This variant is present in the 5’ distal enhancer region of the CHRNA5 gene and defines the primary index signal reported in studies of European ancestry. No other SNP reached genome-wide significance for smoking initiation (ever vs. never smoking), age of smoking initiation, or smoking cessation (former vs. current smoking). Informative associations that approached genome-wide significance included three modestly correlated variants, at 15q25.1 within PSMA4, CHRNA5 and CHRNA3 for smoking quantity, which are associated with a second signal previously reported in studies in European ancestry populations, and a signal represented by three SNPs in the SPOCK2 gene on chr10q22.1.
The only genome-wide significant SNP (rs2036527) in the chromosome 15q24-25.1 region associated with SQ in the STOMP investigation, has also been associated with lung cancer in the only published fine mapping study of lung cancer in African Americans to date. This observation raises the prospect that rs2036527 could be a susceptibility locus for both SQ and lung cancer given the established dose-response relationship between SQ and lung cancer, we seek to examine the hypothesis that the association between this SNP and lung cancer is moderated by SQ by conducting a candidate gene based analysis in a discovery sample of WHI participants (African Americans in WHI SHARe) and seeking replication in other STOMP Consortium samples; and then to compare these results to those observed for other shared susceptibility loci in European-ancestry samples with collaborators in the TAG, ENGAGE and other genetic consortia.

Like smoking, alcohol use and abuse are moderately to highly heritable and have extreme societal, occupational, and personal costs. To date, there has been a single GWAS meta-analysis of alcohol consumption (N>47,000), which found a SNP in the AUTS2 gene associated with the phenotype\(^1\). The SNP was verified in a mouse model.

There are many candidate gene investigations of alcohol-related phenotypes, of which we highlight two here. First, it is well-established that variation in ALDH2\(^1\) is associated with alcohol use and dependence. This variation is functional, in that individuals with deficient copies of the gene show a marked “flushing” response after consumption of alcohol due to impairment in their ability to process acetaldehyde. The buildup of this unprocessed metabolite causes dysphoria, and appears to be a protective factor for alcoholism. This is a common example of the large phenotypic effect possible by studying functional variation.

There are many more genes involved in the metabolism of alcohol, as well as the psychotropic effects, but there are to our knowledge no variants within these genes that have shown a replicable association with alcohol use or dependence. One possible explanation is that genetic association studies to date have been underpowered and have focused on common, likely non-functional genetic variation. The present proposal is an attempt to overcome both of these challenges, by organizing a consortium of investigators to directly investigate the role of nonsynonymous exomic variation in alcohol use and smoking behavior.

The exome is the protein coding portion of the genome. A nonsynonymous variant is one that changes the coding sequence, resulting in either a change in the protein, or production of a dysfunctional protein. Because these damaging variants can alter the product of an entire gene, they are relatively rare and are expected to have larger phenotypic effects.

The exome meta-analysis of drinking and smoking (EMADS) has informal commitments from over 20 studies totaling over 100,000 samples. The consortium began in late December 2012, and we continue to grow as more studies genotype their samples on the exome chip. Many past GWAS meta-analytic efforts have met with considerable
success in elucidating the genetic etiology of complex traits and disease (e.g., for height, BMI, type-2 diabetes, age-related macular degeneration, smoking, drinking, schizophrenia, etc.). Similar to past GWAS meta-analyses, individual studies will generate summary-level statistics for each SNP in their sample. These statistics will then be aggregated through meta-analysis, allowing for the massive sample sizes required to detect what are expected to be small individual genetic effects. We describe the particular objectives of this study, as well as the analysis and phenotype details, below.

5. Main Hypothesis/Study Questions: A. Examine relationship between nonsynonymous and other exomic loci with alcohol and tobacco use; B. Meta-analyze results from ARIC and other cohorts in the EMADS Genetics Consortium

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).

Subjects and sample sizes: The study populations for this project include all participants from ARIC with available smoking and alcohol phenotype and exome chip data who consented to future DNA use. The variables of interest and socio-demographic characteristics of genotyped participants in ARIC and of other cohorts in the STOMP Genetics Consortium are in Table 1. Other studies that plan to contribute to the EMADS consortium are listed in Table 2.

Table 1. Example of Meta-Analysis Studies Descriptive Characteristics

<table>
<thead>
<tr>
<th>Study</th>
<th>N (% female)</th>
<th>Age, mean (s.d.)</th>
<th>Ever smokers (%)</th>
<th>CPD, mean (s.d.)</th>
<th>AOF, mean (s.d.)</th>
<th>Former smokers (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>AABC</td>
<td>5861 (100)</td>
<td>56.6 (12.6)</td>
<td>47.2</td>
<td>11.9 (8.4)</td>
<td>23.3 (9.0)</td>
<td>58.8</td>
</tr>
<tr>
<td>AACP</td>
<td>5866 (3)</td>
<td>63.7 (9.6)</td>
<td>68.7</td>
<td>14.8 (9.9)</td>
<td>23.2 (9.0)</td>
<td>64.9</td>
</tr>
<tr>
<td>CHS</td>
<td>801 (63.2)</td>
<td>72.9 (5.5)</td>
<td>51.2</td>
<td>13.9 (11.5)</td>
<td>19.0 (5.2)</td>
<td>66.8</td>
</tr>
<tr>
<td>CRe</td>
<td>2916 (61.2)</td>
<td>54.1 (3.7)</td>
<td>52.2</td>
<td>14.4 (9.8)</td>
<td>19.5 (6.4)</td>
<td>28.1</td>
</tr>
<tr>
<td>ARIC</td>
<td>CARDIA</td>
<td>953 (61.4)</td>
<td>24.4 (3.8)</td>
<td>59.2</td>
<td>11.8 (8.7)</td>
<td>17.3 (5.1)</td>
</tr>
<tr>
<td>CHS</td>
<td>632 (69.0)</td>
<td>36.5 (19.6)</td>
<td>45.1</td>
<td>13.1 (10.3)</td>
<td>19.0 (5.5)</td>
<td>13.3</td>
</tr>
<tr>
<td>JHS</td>
<td>2145 (60.7)</td>
<td>55.2 (12.8)</td>
<td>33.2</td>
<td>14.9 (10.8)</td>
<td>19.3 (5.7)</td>
<td>17.0</td>
</tr>
<tr>
<td>MESA</td>
<td>1646 (64.7)</td>
<td>62.2 (10.1)</td>
<td>53.5</td>
<td>14.6 (18.3)</td>
<td>18.3 (5.4)</td>
<td>36.0</td>
</tr>
<tr>
<td>GeneSTAR</td>
<td>1175 (61.7)</td>
<td>47.4 (12.3)</td>
<td>72</td>
<td>11.5 (13.3)</td>
<td>16.3 (5.4)</td>
<td>44.0</td>
</tr>
<tr>
<td>HANDLS</td>
<td>918 (54.5)</td>
<td>48.6 (8.9)</td>
<td>65.4</td>
<td>15.7 (32.8)</td>
<td>17.4 (8.2)</td>
<td>29.0</td>
</tr>
<tr>
<td>Health ABC</td>
<td>1137 (57.2)</td>
<td>73.4 (2.9)</td>
<td>56.4</td>
<td>15.7 (12.6)</td>
<td>19.5 (7.0)</td>
<td>68.5</td>
</tr>
<tr>
<td>HyperGEN</td>
<td>1241 (67.3)</td>
<td>45.2 (13.3)</td>
<td>48.7</td>
<td>12.1 (9.8)</td>
<td>19.5 (5.5)</td>
<td>58.0</td>
</tr>
<tr>
<td>WHI (SHARe)</td>
<td>8208 (100)</td>
<td>61.6 (7.0)</td>
<td>50.6</td>
<td>11.5 (9.5)</td>
<td>20.5 (5.9)</td>
<td>39.1</td>
</tr>
</tbody>
</table>

Table 2. Other studies that plan to contribute to EMADS consortium

<table>
<thead>
<tr>
<th>Study</th>
<th>~N (European)</th>
<th>~N (other)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Study</td>
<td>Ancestry</td>
<td>ancestry</td>
</tr>
<tr>
<td>-----------------------</td>
<td>----------</td>
<td>----------</td>
</tr>
<tr>
<td>GECCO &amp; CORECT</td>
<td>20,000</td>
<td>TBD</td>
</tr>
<tr>
<td>Multiethnic Cohort</td>
<td>15,000</td>
<td>TBD</td>
</tr>
<tr>
<td>Malmo Diet and Cancer</td>
<td>12,000</td>
<td>TBD</td>
</tr>
<tr>
<td>Health and Retirement Study</td>
<td>12,000</td>
<td>3000</td>
</tr>
<tr>
<td>METSIM</td>
<td>9,000</td>
<td>0</td>
</tr>
<tr>
<td>FUSION</td>
<td>1,500</td>
<td>0</td>
</tr>
<tr>
<td>ARIC</td>
<td>9,000</td>
<td>TBD</td>
</tr>
<tr>
<td>IMAGEN</td>
<td>2,000</td>
<td>TBD</td>
</tr>
<tr>
<td>Lollypop, NFBC, Police</td>
<td>6,000</td>
<td>TBD</td>
</tr>
<tr>
<td>MCTFR</td>
<td>7,000</td>
<td>800</td>
</tr>
<tr>
<td>SardNIA</td>
<td>6,000</td>
<td>0</td>
</tr>
<tr>
<td>HCBS, FINRISK &amp; Health 2000</td>
<td>6,500</td>
<td>TBD</td>
</tr>
<tr>
<td>COPD</td>
<td>7,000 (all smokers)</td>
<td>TBD</td>
</tr>
</tbody>
</table>

**Genotypes:** All samples have some version of the Exome Chip or exome/whole genome sequences. Individual studies will provide information about the manufacturer and version of the exome chip, or sequencing platform, they are using.

**Inclusion Criteria:** For our first analysis, samples must be between ages 18 and 70 (inclusive) and of European ancestry. We will extend analysis to other ancestral groups later.

**Quality Control:** We leave calling algorithms, marker filters, and sample filters to the discretion of local sites, although we will evaluate the possibility of batch effects (where batch might be a study) during the meta-analysis step.

For reference, four of the five currently participating studies have used Illumina chips and Illumina’s genotype caller in Genome Studio (Gencall). Some studies also implemented some manual curation involving reclustering the intensity data of ~1500 markers.

**Primary Phenotypes**

*Average cigarettes smoked per day, either as a current smoker or former smoker:* Individuals who either never smoked, or on whom we have no data (e.g., someone was a former smoker but former smoking was never assessed) will be excluded from analysis. Only cigarettes will be included in the estimate. If preferable, repeated measures designs (longitudinal data) can use all assessments by scaling and correcting for covariates within waves of assessment, then averaging across assessments.
There was some cross-study variability on this measure. Some studies specified avg smoking during a specific window, such as the last 12 months; most made no such specification. One study allowed respondents to report packs.

**Smoking Initiation:** Every study had some useable measure of whether a respondent has ever regularly smoked. Almost all asked directly. Some have necessary information (e.g., 100 cigs lifetime?, or ever smoked every day for 2 weeks straight?).

Note that we’re among the first groups conducting such meta-analyses, and our analysis pipeline is current restricted to continuous traits. Until methods are developed for binary traits, it is proposed that we analyze smoking initiation as a continuous trait.

**Average drinks per week, either as a current drinker or former drinker:** Individuals who either never drank, or on whom we have no data (e.g., someone was a former drinker but former drinking was not assessed) will be excluded from analysis. All types of liquor will be combined in the total estimate. If preferable, repeated measures designs (longitudinal data) can use all assessments by scaling and correcting for covariates within waves of assessment, then averaging across assessments.

There was some cross-study variability on this measure. Some studies specified avg drinking during a specific window, such as the 12 months or last one month; most made no such specification. Two studies forced the respondent to select ranges.

Secondary Phenotypes

**Pack Years:** Number of packs of cigarettes per day by the number of years the person has smoked, corrected for age.

**Age of Initiation of Smoking:** The age an individual first became a regular smoker.

**Covariates:** Appropriate covariates can often be study-specific. We will depend on local investigators to determine the most appropriate covariates. We list here some covariates that will likely be necessary.

**Main Effects**

1. Age of assessment for current smokers/drinkers
   a. At assessment for current smokers/drinkers
2. Age of smoking/drinking for former smokers/drinkers
   a. Could be age at quitting
3. Age of assessment for Pack Years, Smoking Initiation, and Age of Smoking Initiation
4. Sex
5. Date of birth (or year, or range)
6. Cohort
7. Height, weight, BMI, for drinking (a single beer has different effects on a 200 lb man versus a 100 lb woman)
8. Genetic principle components (or empirical kinships)
9. Adolescence versus adulthood (e.g., < 21 years of age versus >=21)
10. Date of assessment (e.g., the calendar year of the assessment)?

**Interactions**
11. Sex X Adolescence interaction
12. Sex X Age interaction
13. Sex X Weight/Height/BMI interaction
14. Age X Adolescence interaction

**Analysis:** The basic analysis is two-stage. In the first stage, local investigators produce, for each phenotype, a set of single-variant summary statistics using a tool developed at the University of Michigan. In the second stage, these summary statistics are pooled for meta-analysis. All single-variant and gene-based ('burden') tests can be conducted from the summary statistics.

These two stages are now described in more detail.

**Stage 1: Local Sites Produce Summary Statistics Using Rare-Metal-Worker:** The meta-analysis step (stage 2) requires a very specific set of summary statistics, which includes single-variant test statistics and p-values, as well as the test statistic covariance matrix within a sliding window (default: 1Mb). Shuang Feng, Dajiang Liu, and Goncalo Abecasis at the University of Michigan have developed software specifically for this purpose, called Rare-Metal-Worker. Software and usage instructions to generate necessary single variant statistics is available at:

http://genomewiki.expasy.org/wiki/Rare-Metal-Worker

**NOTE:** It is essential that each trait is corrected for covariates, and the residuals are inverse normalized, and then association testing is conducted. The Rare-Metal-Worker software has this functionality when --makeResiduals, and --inverseNormalize are jointly specified.

Alternatively, you could correct for covariates prior to using Rare-Metal-Worker, and then specify --makeResiduals (to ensure the data is centered) and --inverseNormalize. All output files from Rare-Metal-Worker can then be emailed to Scott Vrieze (svrieze at umich dot edu) for centralized meta-analysis.

**Stage 2: Single-Variant and Gene-Based Meta-Analysis:**

**Single-Variant Tests:** We will do meta-analysis of score statistics for individual variants weighting by sample size.

**Gene-Based Tests:** Gene-based tests can be conducted centrally by Scott Vrieze using output from Rare-Metal-Worker.
We will implement two burden tests. First, a Variable Threshold Combined Multivariate and Collapsing count method, where the number of rare alleles is counted in each gene, then the gene is tested for association. Second, we will use SKAT for all rare variants (MAF < .05) within a gene. SKAT allows for variants with opposite directions of effect within the same gene, whereas the variable threshold combined multivariate and collapsing method does not.

Other Analysis Considerations

**Genotype Annotation:** Gene-based burden tests can be augmented with genotype annotation. We currently plan to use only nonsynonymous variants from ANNO-generated annotations relative to GENCODE v7 transcripts. All annotation can be done centrally at the meta-analysis stage to ensure consistency across sites.

**Multivariate Test:** We will pursue development of a multivariate test for drinking and smoking jointly. This could be as simple as, on a per-marker or per-gene basis, averaging effect sizes or p-values for meta-analytic CPD and DPW p-value results.

**Further Downstream Analysis:** Multiple analyses will be implemented to interpret our study findings, including possibly:

1. Pathway-based analysis (e.g., with MAGENTA: http://www.broadinstitute.org/mpg/magenta/)
2. Gene set analysis (e.g., grouping together all nicotinic receptor genes)
3. Human knockout analysis for all individuals with rare variants resulting in effective gene knockouts.
4. Conditional analysis on known variants and any newly discovered rare variants.
5. Sex specific analysis

We more than welcome individual sites to propose additional analysis, as well as to take the lead on additional projects related to the primary aims of this meta-analysis.

7.a. Will the data be used for non-CVD analysis in this manuscript?  
___ Yes    ___ 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.cscce.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)?

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.cscce.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.

13. References:


