TOPMed Smoking Phenotype Analysis Plan, Version 0.2

(Adapted from GSCAN Smoking Sequencing Analysis Plan)

June 5, 2016

# Overview

This analysis plan is intended specifically for analysis of TOPMed sequence genotypes. Analyses herein will be conducted on all PASS genotypes.

# Software

Needed for generating association summary statistics

rvTests: <https://github.com/zhanxw/rvtests>

# Genotypes

All studies have had sequence variant calls made centrally by the TOPMed ICC.

All association analyses will be conducted on all available PASS genotypes (likely a standard output file from the ICC).

# Association analyses will be stratified by ancestry

We want to conduct analyses of all available ancestral groups, including but not limited to African, East Asian, European, Native American, South Asian, Latino/Hispanic, African-American, etc. If you have over a few hundred samples from one or more of these ancestries, please do include them.

We assume ancestry information will be provided directly by the TOPMed ICC and will therefore be simple to incorporate.

# Association analyses **may** be stratified by other variables

If your study is heavily ascertained (e.g., one cohort from a medical center, the other from the community) or has some kind of complicating ascertainment bias (e.g., case-control for heart disease), we ask that you stratify analyses by ascertainment status or use ascertainment status as a covariate. We expect study analysts in collaboration with study PIs to have the best idea of which (stratified analysis versus covariate correction versus other alternative) would be the best approach. If you’re unsure or have questions, please contact Scott (scott.vrieze@colorado.edu).

# Step 1: Define Phenotypes & Covariates

Some studies will have extensive data relevant to these phenotypes. For example, a study may have collected measures of binge drinking during adulthood but also during a participant’s heaviest period of drinking (likely early 20’s). If your study has extensive phenotyping like this, then please contact Scott to develop a phenotype definition plan for your study. He will then bring any issues before the phenotype workgroup.

## Phenotype Definitions

**N.B.:** For binary phenotypes cases are always coded 2 and controls coded 1.

**LIST OF PHENOTYPES**

1. **Cigarettes per day**
   1. Average number of cigarettes smoked per day, either as a current smoker or former smoker. Individuals who either never smoked, or on whom there is no available data (e.g., someone was a former smoker but former smoking was never assessed) will be set to missing.
   2. For studies that collect a quantitative measure of CPD, where the respondent is free to provide any integer (e.g., 13 CPD) we will bin responses as follows:

1 = 1-5

2 = 6-15

3 = 16-25

4 = 26-35

5 = 36+

For studies that already have pre-defined bins, which are different from ours, we will prefer their existing bins.

* 1. Cigarettes per day is almost always measured with a single question:
     1. *How many cigarettes do you smoke per day?*
     2. *How many cigarettes did you smoke per day?*

1. **Smoking initiation (Smoker versus nonsmoker)**
   1. This is a binary phenotype. Code **“2”** for everyone in the study who reports ever being a regular smoker in their life (current or former). Code a **“1”** for everyone who denies ever being a regular smoker in their life. This phenotype is not available in studies that only address current smoking and ignore former smoking.
   2. This can be measured in a variety of ways:
      1. *Have you smoked over 100 cigarettes over the course of your life?*
      2. *Have you ever smoked every day for at least a month?*
      3. *Have you ever smoked regularly?*
      4. *Do you smoke?*
2. **Smoking Cessation (Current versus former smoker)**
   1. Binary phenotype with current smokers coded as **“2”** and former smokers coded as **“1”**.
   2. Usually measured through a combination of questions, including:
      1. *Do you currently smoke?* and *Have you ever smoked regularly?*
      2. *Do you smoke?* and *Have you smoked over 100 cigarettes in your entire life?*
3. **Age at which an individual started smoking regularly**
   1. The age at which an individual first became a regular smoker
   2. This can be measured in a variety of ways:
      1. *At what age did you begin smoking regularly?*
      2. *How long have you smoked?* combined with *What is your current age?*
4. **Drinks per week in individuals who are active drinkers**
   1. The average number of drinks a subject reports drinking each week, aggregated across all types of alcohol. If a study recorded binned response ranges (e.g., instead of quantitative responses your study coded something like 1-4 drinks per week, 5-10 drinks per week) we will use the midpoint of the range. So if an individual reports 1-5 DPW, we assume they drink 2.5 DPW on average.
   2. This can be measured in a variety of ways:
      1. *In the past week, how many alcoholic beverages did you have?*
      2. *Thinking about the past year, on the average how many drinks did you have each week?*
   3. For studies that collect drinks per week separately for different types of alcoholic beverage (e.g., beer, wine, spirits), please contact Scott for details on how to collapse across beverage types.
   4. Please log-transform this variable to pull in outliers (use natural log)
5. **Drinker versus Non-drinker**
   1. If a respondent reports drinking during the recall timeframe used in your study (e.g., your study may have recorded drinker status in the last week, or the last month, or even the last year), then they are coded **“2”**. If they report that they did not drink, they are coded **“1”**.
   2. This can be measured in a variety of ways:
      1. *In the past week (or month, or year) how many drinks did you have on average each week?* (Those reporting zero drinks are considered non-drinkers. Those reporting 1 or more are considered drinkers.)
      2. *Do you currently drink alcohol?*
      3. *Thinking about the last week, on how many days did you drink alcohol?*
6. **Binge Drinking**
   1. This is a complex phenotype that may be measured in a wide variety of ways in different studies. The point is to have a phenotype that measures pathological drinking. Pragmatically, we propose a binary variable where binge drinkers are coded as **“2”** and non-binge-drinkers are coded as **“1”**,
   2. This can be measured in a variety of ways, for example:
      1. Consuming 5+/4+ standard drinks in one sitting (males/females)
      2. *In the last 4 weeks, did you drink so much that you felt very intoxicated (drunk)?*

**COVARIATES**

1. All Phenotypes
   1. Sex
   2. Age
   3. Genetic principal components
   4. Other study specific covariates (e.g., cohort, case/control status). Please contact Scott if you are unsure what other covariates may be appropriate.
2. Additional covariates for specific phenotypes
   1. Drinks per week & Binge Drinking
      1. Height and weight
   2. Cigarettes per day
      1. Current versus former smoker status

**TRANSFORMATIONS**

Drinks per week will be left-anchored at 1 and log-transformed (natural log).

**ABBREVIATIONS**

Phenotype abbreviations used throughout the example code below are:

* CPD = Cigarettes per day
* SI = Smoking initiation
* SC = Smoking cessation
* AI = Age of smoking initiation
* DPW = Drinks per week
* DND = Drinker versus nondrinker
* BDE = Binge drinking in everyone
* BDL = Binge drinking in lifetime drinkers only (if applicable – see phenotype definitions)

## Create one ped file for each ancestry group (study\_gscan\_ANCESTRY\_phen.ped)

If your study is composed only of individuals of European ancestry, then you would create only one ped file and call it “study\_gscan\_EUR\_phen.ped”. If your study is composed of two ancestry groups, say African-Americans and Europeans, then you would create two ped files and call them “study\_gscan\_EUR\_phen.ped” and “study\_gscan\_AFR\_phen.ped”, the first containing only individuals of European ancestry; the second containing only those of African-American ancestry. Repeat this process for other ancestral groups.

Here is an example tab-delimited file with three participants using “x” to denote missing data:

fid iid patid matid sex cpd si sc ai dpw dnd bde bdl

f1 i1 x x 1 3 2 1 15 2.30 2 2 2

f2 i2 x x 2 x 1 x x 0 2 1 1

f3 i3 x x 2 1 2 2 17 x 1 1 x

Key:

**fid** = family ID, **iid** = individual ID, **patid** = father ID, **matid** = mother ID

**cpd** = **cigarettes per day** (binned according to phenotype definitions)

**si** = **smoking initiation** (2=does/has smoked, 1=denies ever smoking)

**sc** = **smoking cessation** (1=has quit; 2=has not quit)

**ai** = **age of initiation of smoking** (normal log of age of initiation)

**dpw** = **drinks per week** (normal log of reported number of drinks per week)

**dnd** = **drinker versus non-drinker** (2=drinker, 1=non-drinker)

**bde** = **binge drinking in everyone** (2=has reported binge drinking, 1=denied binge drinking)

**bdl** **= binge drinking in lifetime drinkers** (2=has reported binge drinking, 1=denied binge drinking)

This example is useful because it shows what values you would expect to have in your pedigree file if you followed the phenotype definition and scale transformations correctly from the Phenotype Definition document.

In this example individual i1 is

* Male (**sex = 1**),
* a FORMER smoker (**si=2**; **sc=1**) who smokes 16-25 cigarettes per day (**cpd = 3**),
* started smoking at 15 [**ai = 15**], and
* has 10 drinks per week [**dpw = ln(10) = 2.30**; **dnd=2**]
* and reports binge drinking (**bde=2, bdl=2**)

Individual i2 is

* female (**sex = 2**),
* a lifelong nonsmoker (**cpd = x**; **sc = x**; **ai = x**; **si = 2**),
* drinks 1 drink per week [**dpw = ln(1) = 0; dnd = 2**],
* and denies binge drinking (**bde=1, bdl=1**)

Individual i3 is

* female (**sex = 2**),
* a CURRENT smoker (**si=2**; **sc=2**) who smokes 1-5 cigarettes per day (**cpd = 1**),
* started smoking at age 17 [**ai = 17**],
* and denies drinking alcoholic beverages (**dpw = x**, **dnd = 1**; **bde =1, bdl=x**)

## Create one covariate file for each ancestry group (study\_gscan\_ANCESTRY\_cov.ped)

For each phenotype file you create you will also create a covariates file, one for each ancestry group in your study.

Here is an example with fake data for individuals i1 and i2:

\*Again, missing values are denoted as “x”. **age2** = age squared, **PC[1-3]** = genetic principal components (if applicable)

fid iid patid matid sex age age2 PC1 PC2 PC3 ... (additional covariates)

f1 i1 x x 1 25 625 1.2 0.8 0.9 ... (additional covariates)

f2 i2 x x 2 40 1600 0.4 0.5 1.0 ... (additional covariates)

f3 i3 x x 2 59 3481 -0.3 1.2 1.4 ... (additional covariates)

# Step 2: Generating Summary Statistics

## **PLEASE NOTE!**

* If your sample is composed of primarily unrelated individuals, proceed to **Step 2a**
* If your sample is composed of a significant number of related individuals (e.g., it is a family study), proceed to **Step 2b**

# Step 2a: **UNRELATED** Individuals

## Run rvTests for each ancestry and trait separately

Example commands for individuals of European ancestry. Note that there are two separate commands for continuous (e.g., CPD) and binary traits (e.g., SI). These commands loop over chromosomes in an attempt to parallelize the analyses. You will likely want to explore other ways of parallelizing.

*####################################################*

*### 1000 Genomes imputation Association Analyses ###*

*####################################################*

*### CONTINUOUS TRAITS (cpd, ai, dpw)*

ancestry=EUR *#replace this as needed with the appropriate ancestry*

for cont\_trait in cpd ai dpw; do *#Loop over continuous phenotypes*

for i in {1..22} X Y; do *#Loop over chromosomes*

rvtest --inVcf yourvcffile.chr${i}.vcf.gz \ *#vcf ( 1000G-imputed)*

--pheno study\_gscan\_${ancestry}\_phen.ped \ *#Input phenotype ped file*

--pheno-name ${cont\_trait} \ *#Name of phenotype (cpd in this case)*

--covar study\_gscan\_${ancestry}\_cov.ped \ *#Name of covariate file*

--meta score \ *#Generate score stats for meta-analysis*

--covar-name sex,age,age2,PC1,PC2,PC3,PC4,PC5,PC6,PC7,PC8,PC9,PC10 \

--xLabel X \ *#Label used for X-chromosome (“X”)*

--useResidualAsPhenotype *#Residualize before testing variants*

--inverseNormal \ *#Inverse normalize the resid distr.*

--qtl \ *#Specify pheno is continuously distr.*

--dosage DS \ *#Specify vcf dosage field (here EC)*

--out STUDY\_${ancestry}\_${cont\_trait}\_chr${i} &

done

done

*###*

*###BINARY traits (si sc dnd bd)*

for binary\_trait in si sc dnd bd; do *#Loop over phenotypes*

for i in {1..22} X Y; do

rvtest --inVcf yourvcffile.chr${i}.vcf.gz \

--pheno study\_gscan\_${ancestry}\_phen.ped \

--pheno-name ${binary\_trait} \

--covar study\_gscan\_${ancestry}\_cov.ped \

--meta score \

--covar-name sex,age,age2,PC1,PC2,PC3,PC4,PC5,PC6,PC7,PC8,PC9,PC10 \

--xLabel X \

--dosage DS \

--out STUDY\_${ancestry}\_${binary\_trait}\_chr${i} &

done

done

The following example command concatenates per-chromosome results into one output file for each ancestry x trait combination. Our hope is that having fewer numbers of files to submit will make file management and transfer easier.

*# Concatenate results into a single file*

for ancestry in EUR AFR EAS LAT; do *#Our ancestry abbreviations; change as needed*

for trait in cpd ai dpw si sc dnd bd; do *#Loop over phenotypes; change as needed*

(zgrep –E '^1\s|#|CHROM' STUDY\_${ancestry}\_${trait}\_chr1.MetaScore.assoc.gz; \

zgrep –E '^2\s' STUDY\_${ancestry}\_${trait}\_chr2.MetaScore.assoc.gz; \

zgrep –E '^3\s' STUDY\_${ancestry}\_${trait}\_chr3.MetaScore.assoc.gz; \

zgrep –E '^4\s' STUDY\_${ancestry}\_${trait}\_chr4.MetaScore.assoc.gz; \

zgrep –E '^5\s' STUDY\_${ancestry}\_${trait}\_chr5.MetaScore.assoc.gz; \

zgrep –E '^6\s' STUDY\_${ancestry}\_${trait}\_chr6.MetaScore.assoc.gz; \

zgrep –E '^7\s' STUDY\_${ancestry}\_${trait}\_chr7.MetaScore.assoc.gz; \

zgrep –E '^8\s' STUDY\_${ancestry}\_${trait}\_chr8.MetaScore.assoc.gz; \

zgrep –E '^9\s' STUDY\_${ancestry}\_${trait}\_chr9.MetaScore.assoc.gz; \

zgrep –E '^10\s' STUDY\_${ancestry}\_${trait}\_chr10.MetaScore.assoc.gz; \

zgrep –E '^11\s' STUDY\_${ancestry}\_${trait}\_chr11.MetaScore.assoc.gz; \

zgrep –E '^12\s' STUDY\_${ancestry}\_${trait}\_chr12.MetaScore.assoc.gz; \

zgrep –E '^13\s' STUDY\_${ancestry}\_${trait}\_chr13.MetaScore.assoc.gz; \

zgrep –E '^14\s' STUDY\_${ancestry}\_${trait}\_chr14.MetaScore.assoc.gz; \

zgrep –E '^15\s' STUDY\_${ancestry}\_${trait}\_chr15.MetaScore.assoc.gz; \

zgrep –E '^16\s' STUDY\_${ancestry}\_${trait}\_chr16.MetaScore.assoc.gz; \

zgrep –E '^17\s' STUDY\_${ancestry}\_${trait}\_chr17.MetaScore.assoc.gz; \

zgrep –E '^18\s' STUDY\_${ancestry}\_${trait}\_chr18.MetaScore.assoc.gz; \

zgrep –E '^19\s' STUDY\_${ancestry}\_${trait}\_chr19.MetaScore.assoc.gz; \

zgrep –E '^20\s' STUDY\_${ancestry}\_${trait}\_chr20.MetaScore.assoc.gz; \

zgrep –E '^21\s' STUDY\_${ancestry}\_${trait}\_chr21.MetaScore.assoc.gz; \

zgrep –E '^22\s' STUDY\_${ancestry}\_${trait}\_chr22.MetaScore.assoc.gz; \

zgrep –E '^X\s' STUDY\_${ancestry}\_${trait}\_chrX.MetaScore.assoc.gz) \

| bgzip -c > STUDY\_${ancestry}\_${trait} &

done

done

# Step 3b: Sample of **RELATED** Individuals (e.g., families)

## CREATE PHENOTYPE/COVARIATE FILES (study\_gscan\_ANCESTRY\_phen.ped & study\_gscan\_ANCESTRY\_cov.ped)

Follow the instruction under Step 3a to define phenotypes and create these files.

## Generate kinship matrices

To account for familial relatedness and population stratification rvtests uses an empirical kinship matrix. We need only generate this kinship matrix once, on the full 1000 Genomes imputed VCF files. That matrix can then be used in all association analyses.

rvTests generates an empirical kinship matrix from the VCF file. Within the rvtest folder there is a script called “vcf2kinship”. However, we want to run it on all common markers genome-wide, so you probably want to reduce your available markers to common SNPs only, and then concatenate into a single genome-wide file

*### Generate kinship matrix (--threads controls the number of parallel threads, adjust as needed)*

vcf2kinship --inVcf yourvcffile.MAF10.SNPs.vcf.gz \

--bn \ *#balding-nichols method*

--out kinship\_matrix \ *#output file name prefix*

--xLabel X \ *#Label we used for the X chromosome*

--xHemi \ *#create kinship for hemizygous region*

--minMAF .05 \ *#min MAF of variants that contribute to the kinship matrix*

--threads 12

## Run rvTests for each ancestry and trait separately

Example commands for individuals of East Asian ancestry. These commands loop over chromosomes in an attempt to parallelize the analyses. You will likely want to explore other ways of parallelizing.

*########################*

*### 1000G imputation ###*

*### CONTINUOUS AND BINARY TRAITS USE SAME COMMAND*

ancestry=EAS *#replace this as needed with the appropriate ancestry*

for trait in cpd ai dpw si sc dnd bd; do *#Loop over phenotypes*

for chr in {1..22} X; do *#Loop over chromosomes*

rvtest --inVcf yourvcffile.chr${i}.vcf.gz \ *#Input vcf*

--pheno study\_gscan\_${ancestry}\_phen.ped \ *#Input phenotype ped file*

--pheno-name ${trait} \ *#Name of phenotype*

--covar study\_gscan\_${ancestry}\_cov.ped \ *#Name of covariate file*

--meta score \ *#Generate score stats for meta-analysis*

--covar-name sex,age,age2,PC1,PC2,PC3,PC4,PC5,PC6,PC7,PC8,PC9,PC10 \

--kinship kinship\_matrix.kinship \ *#Kinship filename*

--xLabel X \ *#Label for X chromosome in vcf file*

--xHemiKinship kinship\_matrix.xHemiKinship \ *#Kinship in hemizygous region*

--useResidualAsPhenotype \ *#Residualize before testing variants*

--inverseNormal \ *#Inverse normalize the resid distr.*

--qtl \ *#Force analysis as if pheno is continuous*

--dosage DS \ *#Specify vcf dosage field (here DS)*

--out STUDY\_${ancestry}\_${trait}\_chr${i} &

done

wait

done

## CONCATENATING RESULTS INTO ONE OUTPUT FILE PER PHENOTYPE

The following example command concatenates per-chromosome results for HRC-imputed files into one output file for each ancestry x trait combination. Our hope is that having fewer numbers of files will make file management and transfer easier.

*# Concatenate results into a single file*

for ancestry in EUR AFR EAS LAT; do *#Our ancestry abbreviations; change as needed*

for trait in cpd ai dpw si sc dnd bd; do *#Loop over phenotypes; change as needed*

(zgrep –E '^1\s|#|CHROM' STUDY\_${ancestry}\_HRC\_${trait}\_chr1.MetaScore.assoc.gz; \

zgrep –E '^2\s' STUDY\_${ancestry}\_${trait}\_chr2.MetaScore.assoc.gz; \

zgrep –E '^3\s' STUDY\_${ancestry}\_${trait}\_chr3.MetaScore.assoc.gz; \

zgrep –E '^4\s' STUDY\_${ancestry}\_${trait}\_chr4.MetaScore.assoc.gz; \

zgrep –E '^5\s' STUDY\_${ancestry}\_${trait}\_chr5.MetaScore.assoc.gz; \

zgrep –E '^6\s' STUDY\_${ancestry}\_${trait}\_chr6.MetaScore.assoc.gz; \

zgrep –E '^7\s' STUDY\_${ancestry}\_${trait}\_chr7.MetaScore.assoc.gz; \

zgrep –E '^8\s' STUDY\_${ancestry}\_${trait}\_chr8.MetaScore.assoc.gz; \

zgrep –E '^9\s' STUDY\_${ancestry}\_${trait}\_chr9.MetaScore.assoc.gz; \

zgrep –E '^10\s' STUDY\_${ancestry}\_${trait}\_chr10.MetaScore.assoc.gz; \

zgrep –E '^11\s' STUDY\_${ancestry}\_${trait}\_chr11.MetaScore.assoc.gz; \

zgrep –E '^12\s' STUDY\_${ancestry}\_${trait}\_chr12.MetaScore.assoc.gz; \

zgrep –E '^13\s' STUDY\_${ancestry}\_${trait}\_chr13.MetaScore.assoc.gz; \

zgrep –E '^14\s' STUDY\_${ancestry}\_${trait}\_chr14.MetaScore.assoc.gz; \

zgrep –E '^15\s' STUDY\_${ancestry}\_${trait}\_chr15.MetaScore.assoc.gz; \

zgrep –E '^16\s' STUDY\_${ancestry}\_${trait}\_chr16.MetaScore.assoc.gz; \

zgrep –E '^17\s' STUDY\_${ancestry}\_${trait}\_chr17.MetaScore.assoc.gz; \

zgrep –E '^18\s' STUDY\_${ancestry}\_${trait}\_chr18.MetaScore.assoc.gz; \

zgrep –E '^19\s' STUDY\_${ancestry}\_${trait}\_chr19.MetaScore.assoc.gz; \

zgrep –E '^20\s' STUDY\_${ancestry}\_${trait}\_chr20.MetaScore.assoc.gz; \

zgrep –E '^21\s' STUDY\_${ancestry}\_${trait}\_chr21.MetaScore.assoc.gz; \

zgrep –E '^22\s' STUDY\_${ancestry}\_${trait}\_chr22.MetaScore.assoc.gz; \

zgrep –E '^X\s' STUDY\_${ancestry}\_${trait}\_chrX.MetaScore.assoc.gz) \

| bgzip -c > STUDY\_${ancestry}\_${trait}.MetaScore.assoc.gz &

done

done

# Step 4. Upload Results

Congratulations! You made it! Please upload to sftp server at the University of Michigan for central meta-analysis -- please email Scott ([scott.vrieze@colorado.edu](mailto:scott.vrieze@colorado.edu)) for the hostname, username, and password.

## Rename all of the following files and place them in a single directory

**rvTests.** Please submit each of the results files. The total number will be a function of the number of ancestries times the number of phenotypes. Use the following naming convention.

STUDY\_ANCESTRY\_TRAIT\_DDMMYY\_INITIALS.MetaScore.assoc.gz

**README**

STUDY\_DDMMYY\_INITIALS.readme

Please submit the README file with the following information:

1. Name, email
2. Study name
3. Array version(s)
4. Basic information about genotype calling procedures (e.g., GenomeStudio, etc.)
5. The actual survey questions and recall periods (last year, last month, period of heaviest drinking, etc.) used to build the phenotypes and any irregularities encountered in phenotype definition
6. Other concerns or uncertainties that arose during the analysis

Key:

STUDY = your study name (please also add any strata - e.g., WHI\_AfricanAmerican)

ANCESTRY = AFR, EUR, EAS, SAS, LAT, AME (for African, European, East Asian, South Asian, Latino, and (Native) American respectively)

TRAIT = CPD, AI, SI, SC, DPW, DND, BDE, BDL (if applicable)

DDMMYY = Day, Month, and Year of submission (January 1, 2016 would be “010116”)

INITIALS = initials of the analyst

## Tarball the directory and transfer to sftp server (ask Scott for log-in details)

Please place all relevant output files into a folder, make a zipped tarball, and upload it. Let Scott know when you’ve done so.

*### Make tarball to hold all the results*

tar -zcvf STUDY\_DDMMYY\_INITIALS.tar.gz yourresultsdirectory