GSCAN EXOME CHIP ANALYSIS PLAN, Version 2
April 15, 2014

Software

All the following software will very likely be needed. rvTests and RareMetalWorker are redundant, so only one is really needed.
Plinkseq: https://atgu.mgh.harvard.edu/plinkseq/
rvTests: https://github.com/zhanxw/rvtests
raremetalworker: http://genome.sph.umich.edu/wiki/Rare-Metal-Worker
VCF check python script: http://genome.sph.umich.edu/wiki/CheckVCF.py
BGZIP and tabix: http://samtools.sourceforge.net/tabix.shtml

Required Files

These files are available on the sftp server in /data/reference-files. Email Scott at svrieze@umich.edu for log in details.

```
Autosomal_x_snps_illumina_v1.0_exome.txt
FORCEALLELES_V1.0
```

Genotypes

Array

All studies for the exome chip meta-analysis have some version of the Exome Chip or whole exome sequencing. Individual studies will provide information about the manufacturer and version of the exome chip, or sequencing platform, they are using.

Genotype QC

We leave calling algorithms, marker filters, and sample filters to the discretion of local sites. We only ask that studies report basic QC information to us when they submit summary statistics.

Strand

Strand is an extremely important issue. Without the correct strand meta-analysis doesn't work.

Illumina forward strand IS NOT build 37 forward strand. An easy way to get build 37 forward strand is to export genotypes from GenomeStudio using TOP allele annotations (typical output from GenomeStudio), which we then ask you to update to the forward strand of build 37 using scripts provided by Will Rayner at Sanger. A description of Illumina's TOP/BOT scheme is here. Will's usage instructions, including scripts, are available here.

CHARGE studies only: please flip alleles to forward strand using the file on the sftp server in /data/reference-files/SNP_LIST_TO_FLIP_CHARGE_POStoFWD.txt.

```
plink --bfile QCD_CHARGE_PLINK_FILE --flip SNP_LIST_TO_FLIP_CHARGE_POStoFWD.txt \
      --make-bed ---out QCD_PLINK_FILE
```

Step 1: Generate VCF file on forward strand of Build 37

Note that this step is the same as for other consortia including GIANT and MAGIC. If you have already generated aligned VCF files for GIANT, there is no need to generate them again, and you can go to Step 3.

Converting PLINK Genotypes to VCF File Format

We keep autosomal and Chromosome X and force the reference allele to be consistent across studies (Note that the file is designed for those genotyped on Illumina Exome chip v1.0. If you have another version of the chip, please create a file with a list of variants in chr1-23 and use that).

We assume your genotypes are stored in PLINK format with filenames QCD_PLINK_FILE.[bim/bam/fam]. To force reference alleles run the following command:

```
plink --bfile QCD_PLINK_FILE --extract autosomal_x_snps_illumina_exome.txt \
      --reference-allele FORCEALLELES_V1.0 --make-bed --out QCD_PLINK_FILE_FINAL
```

Next, use plinkseq to generate a VCF file.

1. Create new project
   ```
   pseq gscan_project new-project
   ```

2. Load existing PLINK files
   ```
   pseq gscan_project load-plink --file QCD_PLINK_FILE_FINAL --id iid \
        --check-reference
   pseq gscan_project write-vcf > study_gscan_chr.vcf
   ```

3. Remove “chr” prefix from the chromosome markers in the vcf file and change 23 to X (the latter step is necessary for the vcf-check file to run successfully)
   ```
   sed 's/^chr//' study_gscan_chr.vcf | sed 's^23/X/' > study_gscan.vcf
   ```

4. bgzip and tabix the vcf file
   ```
   bgzip study_gscan.vcf
   tabix -p vcf study_gscan.vcf.gz
   ```

Checking Strand and Allele Orientation in the VCF File

```
python checkVCF.py -r hs37d5.fa -o study_gscan_vcf_check1 study_gscan.vcf.gz
```

This script will output files where the reference allele in the vcf file doesn’t match the required reference allele. This information will be stored in (using the output name -o {output_prefix} in the above command): study_gscan_vcf_check1.check.ref

If the alleles match, congratualations! You now have a clean VCF file for rvTests or RareMetalWorker.

Step 2: Define phenotypes

Inclusion Criteria

Age -- For all analyses we are restricting to individuals between the ages of 18 and 70, inclusive.
Ancestry -- If a study contains individuals from diverse ancestries we propose to stratify the sample by ancestry and conduct association analysis for each ancestry separately. Then, at the meta-analysis stage, we can conduct within-ancestry and trans-ancestry analysis.

(1) 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.

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 into the following bins: 1-10, 11-20, 21-30, 31+.'" If some study collected binned responses from the outset, and those bins happen to differ from ours (e.g., 1-5, 6-15, etc.), then we will simply use whatever bins the study has collected. Please contact Scott if your study does something completely different.

In analysis, we consider the bins to correspond to the following numerical values.

- 1 = 1-10 cigarettes per day
- 2 = 11-20 cigarettes per day
- 3 = 21-30 cigarettes per day
- 4 = 31+ cigarettes per day

Please note, however, that when we report descriptive statistics about our phenotypes we will want to report the original participant responses. Even though we'll bin the data for analysis, we'll still report quantitative CPD (when possible) when we describe each study's phenotype in eventual publications.

(2) Smoking Initiation

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.

Every study had some usable measure of whether a respondent has ever regularly smoked. Almost all asked directly. Some have necessary information to code this variable (e.g., 100 cigs lifetime? 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 currently restricted to continuous traits. Until methods are developed for binary traits, it is proposed that we analyze smoking initiation as a continuous trait.

(3) Pack Years

Number of cigarettes per day, divided by 20, then multiplied by the number of years the person has smoked. For this measure please use the quantitative CPD, and not the binned responses discussed above under the CPD heading. If your study collected binned responses from the outset, please use the midpoint of the range in calculating Pack Years. For example, individuals stating they smoked 11-20 CPD would be assumed to have smoked 15.5 on average

(4) Age of Initiation of Smoking

The age an individual first became a regular smoker. Please check for obvious outliers and remove them (e.g., someone who claims to be a regular smoker at age 4).

(5) Average drinks per week, either as a current drinker or former drinker

The average number of drinks a subject reports drinking each week. Most studies asked this question directly. Other studies have converted to grams per day, or grams per week. The latter are fine to analyze directly for our purposes.

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. Please combine all types of liquor 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.

If your study forced the respondent to report ranges (e.g., 1-5, 6-10, 11-15, 16-20, etc.) please simply use the midpoint of the range. For example, if one range is 1-5 DPW, we assume they drink 2.5 DPW on average. Then use these midpoints in all subsequent analysis.

Scale Transformation

Please natural log transform the three quantitative phenotypes (Pack Years, Age of Initiation, Drinks Per Week). You may need to left-anchor the phenotypes first, adding/subtracting a constant to all values such that no value is less than 1, which prevents the log-transform from returning nonsense values like negative infinity. (Try taking the log of zero!). We will use these transformed phenotypes in all analyses. No transformations are necessary for CPD or the binary smoking initiation phenotype.

Step 3: Define 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.

Recommended Covariates

- Age
  - At assessment in current smokers/drinkers
  - Age of smoking/drinking for former smokers/drinkers could be age at quitting
  - At assessment for Pack Years, Smoking Initiation, and Age of Initiation, regardless of current/former smoking status
  - Age squared, if appropriate
- Sex
- Genetic principal components (DO NOT use if employing an empirical kinship mixed model to account for population stratification)

- Current versus former smoker as a covariate for cigarettes per day and pack years. This would be a binary covariate and can be coded as 0/1 or 1/2.
- Current versus former drinker for drinking phenotypes. This would be a binary covariate and can be coded as 0/1 or 1/2.

Suggestions for additional covariates to consider

- Date of birth (or year, or range), if appropriate (e.g., in an accelerated cohort design)
- Cohort
- Adolescence versus adulthood (e.g., < 21 years of age versus >=21). Only consider using this covariate if you have a large number of adolescents in your study.
- Date of assessment (e.g., the calendar year of the assessment)?
- For drinking phenotype, consider height, weight, and/or BMI (the idea is that a similar amount of alcohol has different effects on a 200 lb person versus a 100 lb person)

Step 4: Generating summary statistics (CHOOSE between a or b)

Step 4a: rvTests

Use these steps if you wish to use rvTests.

Create Phenotype File with Headers (study_gscan_phen.ped)

Here is an example tab-delimited file with two rows of fake data and "x" to denote missing data:

fid	iid	patid	matid	sex	cpd	py	ai	si	dpw
f1	i1	x	x	1	2	2.30	2.71	1	2.302
f2	i2	x	x	2	x	x	x	0	0.693

*Missing values should be coded as strings, as in the example. fid = family ID, iid = individual id, patid = father id, matid = mother's id, cpd = cigarettes per day, py = pack years, ai = age of initiation of smoking, si = smoking initiation, dpw = drinks per week

In this example individual i1 is male, smokes 11-20 cigarettes per day (cpd bin 2), started smoking at 15, and drinks 10 drinks per week. Individual i2 is female, a lifelong nonsmoker (hence missing for cpd, py, and ai but 0 for si), and drinks 2 drinks per week. Please note that I've used the bins for cpd. For py, ai, and si, I've left-anchored followed by log-transform with natural log.

Create Covariate File with Headers (study_gscan_cov.txt)

Another example with fake data for individuals i1 and i2:

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	
f2	i2	x	x	2	40	1600	0.4	0.5	1.0	

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

For Family Studies Only: Generate Kinship Matrix

rvTests generates an empirical kinship matrix from the VCF file. Within the rvtests folder there is a script called "vcf2kinship".

```
vcf2kinship --inVcf study_gscan.vcf.gz --nb --out output_kin --xLabel X --xHemi
```

Alternatively, if preferable and pedigrees are known

```
vcf2kinship --pedigree study_magic_phen.ped --out output_kin --xHemi
```

Run rvTests for each trait separately

Here are some examples for unrelated samples (i.e., those that don't require kinship matrices).

Cigarettes per day

```
rvtest --inVcf study_gscan.vcf.gz --pheno study_gscan_phen.txt --pheno-name cpd \
      --covar study_gscan_cov.txt --meta score,cov,dominant,recessive \
      --covar-name sex,age,age2,PC1,PC2,PC3,PC4,PC5,PC6,PC7,PC8,PC9,PC10
      --xLabel X --useResidualsAsPhenotype --inverseNormal \
      --out study_gscan_invnorm_cpd
```

Pack years

```
rvtest --inVcf study_gscan.vcf.gz --pheno study_gscan_phen.txt --pheno-name py \
      --covar study_gscan_cov.txt --meta score,cov,dominant,recessive \
      --covar-name sex,age,age2,PC1,PC2,PC3,PC4,PC5,PC6,PC7,PC8,PC9,PC10
      --xLabel X --useResidualsAsPhenotype --inverseNormal \
      --out study_gscan_invnorm_py
```

Age of Initiation of Smoking:

```
rvtest --inVcf study_gscan.vcf.gz --pheno study_gscan_phen.txt --pheno-name ai \
      --covar study_gscan_cov.txt --meta score,cov,dominant,recessive \
      --covar-name sex,age,age2,PC1,PC2,PC3,PC4,PC5,PC6,PC7,PC8,PC9,PC10
      --xLabel X --useResidualsAsPhenotype --inverseNormal \
      --out study_gscan_invnorm_ai
```

Smoking Initiation (BINARY TRAIT):

```
rvtest --inVcf study_gscan.vcf.gz --pheno study_gscan_phen.txt --pheno-name si \
      --covar study_gscan_cov.txt --meta score,cov,dominant,recessive \
      --covar-name sex,age,age2,PC1,PC2,PC3,PC4,PC5,PC6,PC7,PC8,PC9,PC10
      --xLabel X --useResidualsAsPhenotype --inverseNormal \
      --out study_gscan_invnorm_si \
      --qtl
```

Drinks Per Week:

```
rvtest --inVcf study_gscan.vcf.gz --pheno study_gscan_phen.txt --pheno-name dpw \
      --covar study_gscan_cov.txt --meta score,cov,dominant,recessive \
      --covar-name sex,age,age2,PC1,PC2,PC3,PC4,PC5,PC6,PC7,PC8,PC9,PC10
      --xLabel X --useResidualsAsPhenotype --inverseNormal \
      --out study_gscan_invnorm_dpw
```

To include a kinship matrix to correct for family structure in a family study, simply specify the additional command --kinship output_kin.kinship.

Step 4b: RareMetalWorker

Raremetalworker does the same thing as rvTests but the file format is slightly different. Use these steps if you wish to use raremetalworker.

Create Phenotype File without Headers (study_gscan.ped)

Here is an example tab-delimited file with two rows of fake data and "x" to denote missing data. This is a combination of the phenol and covar files described above for rvTests.

```
f1      i1      x       x       1       2       2.30    2.71    1       2.302   25      625     1.2     0.8     0.9
f2      i2      x       x       2       x       x       x       0       0.693   40      1600    0.4     0.5     1.0
```

*Missing values should be coded as strings, as in the example. In this example individual i1 is male, smokes 11-20 cigarettes per day (cpd bin 2), started smoking at 15, and drinks 10 drinks per week. Individual i2 is female, a lifelong nonsmoker (hence missing for cpd, py, and ai but 0 for si), and drinks 2 drinks per week. Please note that I've used the bins for cpd. For py, ai, and si, I've left-anchored followed by log-transform with natural log.

Create dat file listing phenotypes and covariates (study_gscan.dat)

The .dat file simply tells which columns are which in the ped file. Phenotypes are denoted "T" (for trait), and covariates are denoted "C". The first columns of our ped file (IDs and sex) are not denoted in the dat file. In our example file, the dat file would look like this.

```
T cpd
T py
T ai
T si
T dpw
C age
C age2
C PC1
C PC2
C PC3
(and so on for additional
covariates).
```

*cpd = cigarettes per day, py = pack years, ai = age of initiation of smoking, si = smoking initiation, dpw = drinks per week, age2 = age squared, PC[1-3] = genetic principal components (if applicable)

Caveat for Covariates in raremetalworker

One caveat for raremetalworker is that you cannot choose in the analysis step which covariates you wish to correct for. Raremetalworker simply corrects for all covariates listed in the ped/dat files. So if the covariate list is different for a particular phenotype, a separate set of ped/dat files for that phenotype must be generated. For

example, if you correct for weight in drinks per week, then you need to create a separate ped/dat file set for drinks per week that includes weight as a covariate.

Run raremetalworker for each Trait Separately

Here are example commands for unrelated individuals assuming the covariates are the same for all phenotypes.

Cigarettes per day

```
raremetalworker --ped study_gscan.ped --dat study_gscan.dat \
     --vcf study_gscan.vcf.gz --recessive --dominant --traitName cpd \
     --makeResiduals --inverseNormal --zip --prefix study_gscan_cpd_invnorm
```

Pack years

```
raremetalworker --ped study_gscan.ped --dat study_gscan.dat \
     --vcf study_gscan.vcf.gz --recessive --dominant --traitName py \
     --makeResiduals --inverseNormal --zip --prefix study_gscan_py_invnorm
```

Age of initiation of smoking

```
raremetalworker --ped study_gscan.ped --dat study_gscan.dat \
     --vcf study_gscan.vcf.gz --recessive --dominant --traitName ai \
     --makeResiduals --inverseNormal --zip --prefix study_gscan_ai_invnorm
```

Smoking initiation

```
raremetalworker --ped study_gscan.ped --dat study_gscan.dat \
     --vcf study_gscan.vcf.gz --recessive --dominant --traitName si \
     --makeResiduals --inverseNormal --zip --prefix study_gscan_si_invnorm
```

Drinks per week

```
raremetalworker --ped study_gscan.ped --dat study_gscan.dat \
     --vcf study_gscan.vcf.gz --recessive --dominant --traitName dpw \
     --makeResiduals --inverseNormal --zip --prefix study_gscan_dpw_invnorm
```

To use a genetic kinship matrix, one simply adds the --kinGeno --kinSave and --vcX commands. These will generate a kinship matrix and conduct the association analyses. The --kinSave command allows you to save and reuse the kinship matrix, which is handy because generating the matrix can take extended periods of time for large studies (like days). Here is an example for cpd:

```
raremetalworker --ped study_gscan.ped --dat study_gscan.dat \
     --vcf study_gscan.vcf.gz --recessive --dominant --traitName cpd \
     --makeResiduals --inverseNormal --zip --prefix study_gscan_cpd_invnorm
```

Step 5. Upload Results

Please upload to sftp server at the University of Michigan for central analysis -- please email Scott for the hostname, username, and password. One study also used Aspera to transmit results, which worked well.

Please upload the following files:
vcfCheck
STUDY_DDMMYY_INITIALS.check.ref

Raremetalworker
STUDY_TRAIT_DDMMYY_INITIALS_MODEL.singlevar.score.txt.gz
STUDY_TRAIT_DDMMYY_INITIALS_MODEL.singlevar.cov.txt.gz
STUDY_TRAIT_DDMMYY_INITIALS_MODEL.singlevar.RMW.log

rvTests
STUDY_TRAIT_DDMMYY_INITIALS_MODEL.MetaCov.assoc.gz
STUDY_TRAIT_DDMMYY_INITIALS_MODEL.MetaCov.assoc.gz.tbi
STUDY_TRAIT_DDMMYY_INITIALS_MODEL.MetaScore.assoc.gz

README
STUDY_DDMMYY_INITIALS.readme

where
STUDY = your study name (please also add any strata - e.g., COGA_AfricanAmerican)
TRAIT = cpd, py, ai, si, dpw
MODEL = ADDITIVE, RECESSIVE, DOMINANT

README File format

Please submit the README file with the following information:
Name, email
Study name
Exome chip version
Any basic information about genotype calling procedures (e.g., GenomeStudio, zCall, etc.)
Covariates