Analysis plan for primary cohort GWAS for blood lipid levels for the Global Lipids Genetics Consortium

Cristen Willer (✉ cristen@umich.edu)

Method Article

Keywords: GLGC, GWAS, Lipids, meta-analysis, cholesterol

Posted Date: November 29th, 2021

DOI: <https://doi.org/10.21203/rs.3.pex-1687/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Abstract

This protocol describes the guidelines for generating GWAS summary statistics that were provided to all participating cohorts in the Global Lipids Genetics Consortium HRC + 1KGP3 imputation meta-analysis collected jointly with the GIANT consortium. It is applicable for generating the input phenotype and genotype files for use with the rvtests software. Key stages include imputation, generating phenotype files, and running the analysis to generate summary statistics. The exact time needed to carry out this protocol varies depending on the existing genotype files and size of the cohort. The latest version of rvtests is available at: <http://zhanxw.github.io/rvtests/>

Introduction

Reagents

Equipment

Procedure

Global Lipids Genetics/GIANT Consortium HRC and 1KG phase3 Imputation and Analysis plan version 2

Updated: April 27, 2017 (v2a) to include non-HDL-cholesterol and some minor tweaks but analysis results generated using the March 17, 2017 plan (v2) are also acceptable. All changes since March 17, 2017 are highlighted in yellow.

Plan originally drafted by Xueling Sim, Heather Highland, and Scott Vrieze, modified by GIANT 1KG/HRC working group and Global Lipids Genetics Consortium HRC imputation working group, including Adam Locke and Sailaja Vedantam. If you have questions, please reply to the group CCed on the email accompanying this plan.

Overview

There are three major components to this analysis plan:

- 1) Genome-wide genotypes must be on the correct build (37/hg19) and correct strand (forward).
- 2) For ALL studies, imputation of genotypes is performed using the 1KG phase 3 (if you have not already done so) and, for studies with samples of European-ancestry, also to the large haplotype panel

Loading [MathJax]/jax/output/CommonHTML/fonts/TeX/fontdata.js

from the Haplotype Reference Consortium. If you have already imputed to a different large haplotype panel (e.g. UK10K) please contact us.

3) Association analysis is conducted using specific software tools that will provide all necessary summary statistics for flexible central meta-analysis.

This coordinated plan between Global Lipids and GIANT is meant to reduce the burden on primary analysts. We welcome you to join one or both consortia. We thank you for your participation in past projects, and particularly welcome studies that are new to either consortia.

Software

Below is a list of the software you will need to complete this analysis plan. For some tools using the specific version listed may be important.

Phasing algorithms such as SHAPEIT, Eagle, HapiUR, and Minimac are only necessary if you're conducting imputation in-house, rather than using an imputation server. We provide generic instructions and instructions specific for the University of Michigan imputation server. Some may have used or be using other servers (e.g. the Sanger server); please contact them for instructions specific to that imputation server.

rvTests (20170210): <https://github.com/zhanxw/rvtests/releases>

BGZIP and tabix (0.2.6): <http://samtools.sourceforge.net/tabix.shtml>

Bcftools (1.3.1): <http://www.htslib.org/download>

Vcftools (0.1.8): <https://github.com/vcftools/vcftools>

PLINK1.9 (Plink2): <https://www.cog-genomics.org/plink2>

1. Genotypes

Array

All studies must have some version of a genome-wide array, for example with >200,000 genome-wide common variants. Targeted arrays with limited/incomplete coverage of the genome (e.g. MetaboChip, ImmunoChip, Exome Chip, etc.) should not be used unless merged with a genome-wide array. If you are unsure, please contact Joel and/or Cristen for specific advice.

Individual studies should provide information in the Excel spreadsheet about the manufacturer and version of the array(s) they are using.

QC should be done separately for individual studies, for major continental ancestries within a study, and also separately for samples of the same study that were genotyped on different arrays.

Genotype QC

Typical pre-imputation QC criteria:

These are some steps that we recommend for sample QC. Additional QC steps may be needed and should be determined by the local analysts for each study. Studies should provide a brief description of QC criteria in the Excel sheet.

- Sample call rate (cut off >95% threshold recommended)
- Exclude samples with heterozygosity > median + 3*IQR
- Remove gender mismatches
- Remove duplicates
- Remove PCA outliers using a PCA projection of the study samples onto 1KG reference samples.
- Hardy-Weinberg $p > 10^{-6}$, SNP call rate $\geq 98\%$
- Remove monomorphic markers

Additional QC will be conducted centrally at the meta-analysis stage.

Prepare files for imputation

The “HRC/1KG Imputation Preparation and Checking Tool” developed by Will Rayner will check input data for accuracy relative to expected HRC or 1000G inputs prior to imputation.

This process will identify errors in your input data, including incorrect REF/ALT designations, incorrect strand designations, extreme deviations from expected allele frequencies, and palindromic (A/T and G/C) SNPs with allele frequency near 0.5 that are often the source of imputation errors, and generates commands to make files that have fixed or removed these problematic variants.

The tool can be downloaded here: <http://www.well.ox.ac.uk/~wrayner/tools/HRC-1000G-check-bim.v4.2.5.zip>

The tool requires site lists from the reference panels for comparison, which are available for download at the following locations:

HRC: <ftp://ngs.sanger.ac.uk/production/hrc/HRC.r1-1/HRC.r1-1.GRCh37.wgs.mac5.sites.tab.gz> (then unzipped with gunzip)

1000G Phase 3: http://www.well.ox.ac.uk/~wrayner/tools/1000GP_Phase3_combined.legend.gz (then unzipped with gunzip)

The tool also requires frequency files from plink, which can be created as follows:

```
plink -bfile <binary plink prefix> -freq -out input_file_prefix
```

With these input files the tool can be run as follows:

```
HRC: perl HRC-1000G-check-bim.pl -b <bim file> -f <frequency file> -r HRC.r1-1.GRCh37.wgs.mac5.sites.tab -h
```

```
1000G: perl HRC-1000G-check-bim.pl -b <bim file> -f <frequency file> -r 1000GP_Phase3_combined.legend -g -p <population>
```

(ALL, EUR, AFR, AMR, SAS, EAS are the relevant options for the `-p <population>` parameter, with ALL as default)

The perl script automatically produces a shell-script called *Run-plink.sh*. The shell script contains a set of plink commands that should be run to update or remove SNPs based on the checks and to create one updated binary plink file per chromosome. For this, the paths to the original study binary file should be adjusted in the shell script and the script should be started by typing `./Run-plink.sh` at the command prompt.

The cleaned/updated binary files (one for each chromosome) generated by this tool should be ready for upload to the imputation server below after changing them to VCF format using plink2, bgzip and tabix in the UNIX environment:

```
for i in seq123;
do
    plink -bfile Study-updated-chr${i} --keep-allele-order --recode vcf-iid --out Study-updated-chr${i}
        bgzip Study-updated-chr${i}.vcf
        tabix -p vcf Study-updated-chr${i}.vcf.gz
done
```

2. Imputation

There are two ways to conduct imputation: a) an Imputation Server or b) in-house imputation. The following instructions provide detailed instructions for imputation with the Michigan Imputation Server, but others such as the Sanger server are also acceptable. We recommend you use an imputation server if possible.

Note: If you prefer, the imputation server has a 'Quality Control only' option you can use prior to

Loading [MathJax]/jax/output/CommonHTML/fonts/TeX/fontdata.js

uring imputation of autosomes or X

chromosome.

If you run 'QC and imputation', and samples are removed (e.g., due to gender discordance), please make sure that number and order of samples in the X chromosome and autosomes will match as described in section 2.3.

-

Ancestry of samples and imputation panels:

We ask all studies that have not imputed to 1KG phase 3 to use the imputation server to do so. In addition, for studies with individuals of European-ancestry, we ask that they **also** use the imputation server to impute using the HRC panel.

-

Phasing

The imputation server will automatically determine if you provide phased or unphased VCF files. We encourage re-phasing using Eagle (which is implemented by the imputation server). If you would like to provide phased haplotypes please convert them to VCF format.

An example command to convert haplotype format of HAPS/SAMPLE files from SHAPEIT to VCF would be:

```
shapeit -convert -input-haps study.phased -output-vcf study-phased.vcf
```

2.1. Impute to 1KG phase 3 panel

The 1KG phase 3 samples has both European and non-European samples and includes SNPs, indels, and CNVs. To capture these additional markers, and to provide a common reference panel across all samples, we ask that ALL studies, regardless of ancestry, impute to 1KG phase 3.

Please upload VCF files (phased or non-phased). Then select the following options:

- **Reference Panel: 1000G Phase3 v5**
- **Phasing: Eagle v2.3**

Loading [MathJax]/jax/output/CommonHTML/fonts/TeX/fontdata.js or quality control purposes)

- **Mode:** Quality Control & Imputation

Note: Chromosome X will need to be imputed separate from other chromosomes after reformatting plink code 23 to X in the VCF file. Currently, SHAPEIT is the only available method for phasing chrX.

See below for detailed instructions on using the imputation server.

2.2. Impute to HRC reference panel

The Haplotype Reference Consortium (HRC) has assembled over 60,000 haplotypes by combining together sequencing data from multiple cohorts (<http://www.haplotype-reference-consortium.org>). Impute all individuals of European ancestry using the HRC panel.

Please upload VCF files. Then select the following options:

- **Reference Panel:** HRC r1.1 2016
- **Phasing:** Eagle v2.3
- **Population:** EUR (this parameter is for quality control purposes)
- **Mode:** Quality Control & Imputation

NOTE: Chromosome X will need to be imputed separately from other chromosomes after reformatting plink code 23 to X in the VCF file. Currently, SHAPEIT is the only available method for phasing chrX.

Using the imputation server

The Michigan Imputation Server is available at <https://imputationserver.sph.umich.edu>. The server uses standard encryption (SSL or SFTP) to help ensure that genotypes are encrypted during upload and download. It does a few things automatically:

1. Ensures we are all imputing to the exact same reference panel.
2. Automatically performs quality checks (alleles, MAFs, SNP names, etc.)
3. It phases and imputes, at no cost to the user in either computational time or analyst time.

Make an account and follow the instructions on the website. You will upload VCF genotype files to the server (only VCF file format is supported; instructions on converting to VCF are contained on the imputation server website under “Help” and above in the chapter “Prepare files for imputation”). **One VCF file must be submitted for each chromosome.**

Downloading your imputed genotypes, info files, and QC report

When imputation has finished you will receive an email alert. The imputation server will automatically encrypt all your imputed genotypes (for protection during download). The password to decrypt the files will be in the email notification, so don't delete that email!

When you download your imputed genotypes please be sure to download all available files (the qcreport, statistics, zip files, and all the log files). We will ask that you submit most of these files to us along with all other files generated as part of this analysis plan. **Please do a quick check of the qcreport.html and statistics.txt files before proceeding with the analysis plan. If you are unsure of your imputation quality, please contact us. You will also upload these files to the server (see 8c).**

X chromosome

NOTES on chromosome X:

1) If the server detects sex discrepancy between genotypes and the provided PED file, discordant samples will be removed automatically prior to phasing and imputation. You will need to resolve these discrepancies genome-wide before proceeding with analysis.

2) You will receive two output VCF files for chromosome X, one for males and one for females. You will need to merge these into a single chrX VCF.

Both of these issues are addressed below in section 2.3.

-

Imputation In-House

If you have chosen to use an imputation server then you can ignore this step.

Loading [MathJax]/jax/output/CommonHTML/fonts/TeX/fontdata.js

Some studies will not be able to use the imputation server, for example if original participant consents forbid it. These studies will have to conduct imputation on their own. Please contact us if you need assistance.

2.3. Post-imputation sample harmonization and variant pruning

There are a few post-imputation processing steps that are necessary prior to analysis.

A. Harmonize samples and samples order between autosomes and chromosome X

As mentioned above, chrX imputation may have different numbers and order of samples from the autosomes due to automated filtering on the imputation server. You need to reconcile these possible discrepancies before creating a single whole genome VCF file, which will be used to create a genomic relationship matrix (kinship matrix) for the linear mixed model. If the sets of samples are not identical between the X and autosome VCFs (most likely because possibly sex-discordant samples were dropped during imputation of the X) you will need to eliminate any samples present only in autosome or X VCFs, and then reorder the samples in the X chromosome VCF file. The following bcftools commands 1) make a list of samples IDs in the chr22 VCF, 2) merges the male and female chrX VCFs and reorders the resulting chrX VCF according to the order of samples in chr22, then 3) generates an ordered list of IDs from chrX. Compare the two ID lists to confirm they are identical before proceeding. Note that some cohorts have a small number of males who are heterozygous at many X chromosome markers outside the pseudoautosomal region. Because these create problems later in analysis, we recommend removing these individuals as well, in addition to sex-discordant individuals.

- 1) `bcftools query -l chr22.dose.vcf.gz > ID_order_autosomes.txt`
- 2) `bcftools merge chrX.no.auto_male.dose.vcf.gz chrX.no.auto_female.vcf.gz -m both -O u | bcftools view -S ID_order_autosomes.txt -force-samples -O z > ChrX.combined.reorder.vcf.gz`
- 3) `bcftools query -l chrX.combined.reorder.vcf.gz > ID_order_chrX.txt`

If there are samples in the autosomal files that do not appear in the chrX imputed VCFs, then bcftools will report a warning message:

It then proceeds to reorder the file, omitting these observations. If you encounter this message, you will need to use the sample list file ID_order_chrX.txt to select the overlapping samples when creating the polymorphic datasets in section B below.

If you do **NOT** receive the warning message, please check that the sample list files ID_order_chrX.txt and ID_order_autosomes.txt are exactly the same (in content and order). If these two files are identical, you may proceed without the sample selection option in section B. This will speed up the operation of bcftools in section B.

B. Create a list of monomorphic sites and subset polymorphic variants

These reference panels impute a great many variants into your samples, the vast majority of which are rare, and many will not be polymorphic in your sample. To minimize the size of output files in the association analyses below, we ask that you remove all monomorphic sites (monomorphic defined as at least one variant hard call genotype) from input VCF files prior to analysis. In conjunction, we ask that you upload a list of the variants that you drop in this process. The following commands with bcftools will generate both the list of monomorphic SNPs and the pruned VCF files.

#Generate lists of monomorphic sites

```
#!/bin/bash
```

```
for i in seq122;
```

```
do
```

```
bcftools view -q1.0:major -S ID_order_chrX.txt chr${i}.dose.vcf.gz -O u | bcftools query -f  
'%CHROM\t%POS\t%REF,%ALT\n' >> STUDY_PANEL_ANALYST_imputed.monomorphic.txt
```

```
done
```

#Generate VCFs with only polymorphic variants

```
#!/bin/bash
```

```
bcftools view -c 1:minor -S ID_order_chrX.txt chr${i}.dose.vcf.gz -O z > chr${i}.imputed.poly.vcf.gz

tabix -p vcf chr${i}.imputed.poly.vcf.gz

done
```

3. Phenotype definition, covariate adjustment, and trait transformation

The rank-based inverse normal transformation of the residuals can be performed as below; alternatively, `rvtests` can automatically perform inverse normalization for covariate-adjusted trait residuals (instructions for this are in section 4.D. under Genotype-Phenotype Associations).

```
in R: #if you have missing data

      y <- qnorm((rank(x,na.last="keep")-0.5)/sum(!is.na(x)))
```

(Where N is number of observations)

3A. Lipid Traits

Phenotypes:

We will analyze fasting lipid values using two different trait distributions: raw values and after inverse normal transformation. If you only have non-fasting levels then you can contribute to all four lipids analyses, but your study may be excluded from some analyses.

(i.a) Total cholesterol (mg/dl; inverse normal)

(i.b) Total cholesterol (mg/dl; raw trait)

(ii.a) High-density lipoprotein (HDL) cholesterol (mg/dl; inverse normal)

(ii.b) High-density lipoprotein (HDL) cholesterol (mg/dl; raw trait)

(iii.a) Low-density lipoprotein (LDL) cholesterol (inverse normal)

(iii.b) Low-density lipoprotein (LDL) cholesterol (inverse normal)

(iv.a) Triglycerides (mg/dl; natural log -> inverse normal)

(iv.b) Triglycerides (mg/dl; natural log transformed)

(v.a) Non-HDL cholesterol (nonhdlc) (TC-HDL; mg/dl; natural log -> inverse normal)

(v.b) Non-HDL cholesterol (nonhdlc) (TC-HDL; mg/dl; natural log transformed)

If you have both fasting and non-fasting measurements on the same individuals: please use the fasting measurement.

If you have both fasting and non-fasting measurements in different individuals: For LDL cholesterol and triglycerides, analyses should be carried out in fasting and non-fasting blood measurements separately. For total cholesterol and HDL cholesterol, fasting and non-fasting samples can be analyzed together.

For longitudinal studies, please use the baseline exam or earliest exam with fasting measures. Values should be in mg/dl.

Exclusions:

- Exclude known Mendelian cases or obvious outliers likely to be data errors.

Perform phenotypic modeling separated by:

- Race/ethnicity/ancestry
- Sex
- Disease status if appropriate: analyze cases and controls separately if the phenotype is correlated with disease status.
- Adult/child status

(i) Total Cholesterol

Phenotype adjustment (prior to regression).

- Data collected before 1994 – no lipid medication adjustment
- Data collected after 1994 – for subjects on lipid medication replace TC by TC/0.8

(i.a) Inverse normal transformed

Model: Generate residuals and inverse normal transform in men and women separately and separately by case status if appropriate (e.g. where disease status is correlated with cholesterol you will have MenCase, MenControl, WomenCase, WomenControl, SexCombinedCase, and SexCombinedControl). For the sex combined analyses, combine the “men-specific_TC_INV” and women-specific_TC_INV” phenotype values to analyze together (all_TC_INV).

Total cholesterol (raw trait value in mg/dl for men) = $\text{age} + \text{age}^2 + \text{PCs} + \text{other study specific covariates as needed}$ -> residuals -> inverse normal transformation (men-specific_TC_INV).

NOTE: GLGC requests that you include ~4 PCs as study-specific covariates.

(i.b) Raw trait

Model: Only sex-combined. Generate residuals with sexes combined, but separately by case status if appropriate (e.g. where disease status is correlated with cholesterol you will have SexCombinedCase and SexCombinedControl).

Total cholesterol (raw trait value in mg/dl) = $\text{age} + \text{age}^2 + \text{sex} + \text{PCs}$ (+ other study specific covariates as needed) -> residuals (TC_RAW)

NOTE: GLGC requests that you include ~4 PCs as study-specific covariates.

(ii) High-density lipoprotein (HDL) Cholesterol

(ii.a) Inverse normal transformed

Model: Generate residuals and inverse normal transform in men and women separately and separately by case status if appropriate (e.g. where disease status is correlated with cholesterol you will have MenCase, MenControl, WomenCase, WomenControl, SexCombinedCase, and SexCombinedControl). For the sex-combined analyses, combine the “men-specific_HDL_INV” and women-specific_HDL_INV” phenotype values to analyze together (all_HDL_INV).

HDL cholesterol (raw trait value in mg/dl) = $\text{age} + \text{age}^2 + \text{PCs}$ (+ other study specific covariates as needed) -> residuals -> inverse normal transformation (HDL_INV)

NOTE: GLGC requests that you include ~4 PCs as study-specific covariates.

(ii.b) Raw trait

Model: Only sex-combined. Generate residuals with sexes combined, but separately by case status if appropriate (e.g. where disease status is correlated with cholesterol you will have SexCombinedCase and SexCombinedControl).

HDL cholesterol (raw trait value in mg/dl) = age + age² + sex + PCs (+ other study specific covariates as needed) -> residuals (HDL_RAW)

NOTE: GLGC requests that you include ~4 PCs as study-specific covariates.

(iii) Low-density lipoprotein (LDL) Cholesterol

Phenotype adjustment (prior to regression).

- If not measured, LDL cholesterol can be calculated using the Friedewald equation for those with TG < 400 mg/dl
- LDL = TC – HDL – (TG/5); if TC modified as described above for medication use after 1994, the modified TC is used in this formula.
- If LDL was measured after 1994 (measured and not estimated by Friedewald), then adjust LDL values for individuals taking lipid-lowering medication by using LDL/0.7 to approximate pre-medication LDL levels.

(iii.a) Inverse normal transformed

Model: Generate residuals and inverse normal transform in men and women separately and separately by case status if appropriate (e.g. where disease status is correlated with cholesterol you will have MenCase, MenControl, WomenCase, WomenControl, SexCombinedCase, and SexCombinedControl). For the sex combined analyses, combine the “men-specific_LDL_INV” and women-specific_LDL_INV” phenotype values to analyze together (all_LDL_INV).

LDL cholesterol (raw trait value in mg/dl) = age + age² + PCs (+ other study specific covariates as needed) -> residuals -> inverse normal transformation (LDL_INV)

NOTE: GLGC requests that you include ~4 PCs as study-specific covariates.

(iii.b) Raw trait

Model: Only sex-combined. Generate residuals with sexes combined, but separately by case status if appropriate (e.g. where disease status is correlated with cholesterol you will have SexCombinedCase and SexCombinedControl).

LDL cholesterol (raw trait value in mg/dl) = age + age² + sex + PCs (+ other study specific covariates as needed) -> residuals (LDL_RAW)

NOTE: GLGC requests that you include ~4 PCs as study-specific covariates.

(iv) Triglycerides

(iv.a) Natural log + inverse normal transformed

Model: Generate residuals and inverse normal transform in men and women separately and separately by case status if appropriate (e.g. where disease status is correlated with cholesterol you will have MenCase, MenControl, WomenCase, WomenControl, SexCombinedCase, and SexCombinedControl). For the sex combined analyses, combine the “men-specific_logTG_INV” and women-specific_logTG_INV” phenotype values to analyze together (all_logTG_INV).

$\ln(\text{raw Triglycerides trait value in mg/dl}) = \text{age} + \text{age}^2 + \text{PCs} (+ \text{other study specific covariates as needed})$
-> residuals -> inverse normal transformation (logTG_INV)

NOTE: GLGC requests that you include ~4 PCs as study-specific covariates.

(iv.b) Natural log transformed

Model: Only sex-combined. Generate residuals with sexes combined, but separately by case status if appropriate (e.g. where disease status is correlated with cholesterol you will have SexCombinedCase and SexCombinedControl).

$\ln(\text{raw Triglycerides trait value in mg/dl}) = \text{age} + \text{age}^2 + \text{sex} + \text{PCs} (+ \text{other study specific covariates as needed}) \rightarrow \text{residuals (logTG_RAW)}$

NOTE: GLGC requests that you include ~4 PCs as study-specific covariates.

(v) Non-HDL cholesterol

Use medication-adjusted TC generated in i) and subtract HDL-cholesterol

- Data collected before 1994 – no lipid medication adjustment
- Data collected after 1994 – for subjects on lipid medication replace TC by TC/0.8
- Non-HDL = TC (adjusted) – HDL

(v.a) Inverse normal transformed

Model: Generate residuals and inverse normal transform in men and women separately and separately by case status if appropriate (e.g. where disease status is correlated with cholesterol you will have MenCase, MenControl, WomenCase, WomenControl, SexCombinedCase, and SexCombinedControl). For the sex combined analyses, combine the “men-specific_nonHDL_INV” and women-specific_nonHDL_INV” phenotype values to analyze together (all_nonHDL_INV).

Non-HDL cholesterol (raw trait value in mg/dl for men) = $\text{age} + \text{age}^2 + \text{PCs} (+ \text{other study specific covariates as needed}) \rightarrow \text{residuals} \rightarrow \text{inverse normal transformation (men-specific_nonHDL_INV)}$.

(v.b) Raw trait

Model: Only sex-combined. Generate residuals with sexes combined, but separately by case status if appropriate (e.g. where disease status is correlated with cholesterol you will have SexCombinedCase and SexCombinedControl).

Non-HDL cholesterol (raw trait value in mg/dl) = age + age² + sex + PCs (+ other study specific covariates as needed) -> residuals (nonHDL_RAW)

4. Genotype-phenotype association

A. Group samples for analysis

- Please analyze major ancestry groups separately.
- We anticipate that some studies will have data from multiple genotyping arrays on samples from the same cohort. We expect there will likely be three typical situations:
 - o No sample overlap: analyze studies separately (batches with reasonably similar arrays can be analyzed together, using the array type as a covariate).
 - o All samples overlap: either a) select 1 array for imputation or ideally b) merge genotypes prior to imputation and perform a single analysis.
 - o Partial but significant overlap of samples - contact us, if needed, to customize a plan. The goal should be to upload sets of results from non-overlapping samples that can be combined in meta-analysis.
- If a set of samples being analyzed together is particularly large (for example, N>30000), the association plan may need to be modified. Please contact us to discuss further if this is the case.

B. Perform association analyses

Always test additive models using linear regression, using a method that accounts for genotype

Loading [MathJax]/jax/output/CommonHTML/fonts/TeX/fontdata.js cryptic relatedness (see below). Please indicate

in your submission README file what method you have used for association analysis.

Each individual study will perform data quality control (QC) and analysis and provide summary results for meta-analysis. Results files will be deposited to a central repository (details are provided below) where QC/data cleaning and meta-analysis will be performed.

We request association analyses be carried out using rvtests (version 20170210).

This software can be used for samples that include families or when an empirical kinship matrix is required for analysis. The input files required (phenotype, covariates) for rvtests are compatible with PLINK formats. It also calculates Hardy-Weinberg p-value and call rate for quality control purposes. We ask each study to generate kinship matrices (one for the autosomes and one for chromosome X) and fit linear mixed models to deliver single variant analysis results for the additive model. Documentation for rvtests can be found here: <https://github.com/zhanxw/rvtests>

If you are unable to use a linear mixed model for some reason, please correct for ancestry/relatedness. At a minimum, use ~10 principal components as covariates to correct for population stratification and include principal components as indicated in the analysis models below. If you do not use rvtests, your results may not be usable for conditional analyses and aggregated analyses of rare variants. If in doubt, we are happy to advise. Please indicate in the Excel file requested below the method used to account for ancestry/relatedness.

1. Summary level statistics for meta-analysis of variant associations

The following summary level statistics will be generated and shared for meta-analysis of variant associations, ideally using rvtests. rvtests requires an indexed VCF file (for genotypes), a PED file (for phenotypes) as input. Instructions below are for rvtests. If you are using a different method, please contact us.

i. **Basic VCF metrics**, including reference and alternative alleles, chromosome positions and strand. These statistics are not directly used in the computation of the variant tests but are needed for interpretation and meta-analysis.

ii. **Allele frequency for each variant**

Loading [MathJax]/jax/output/CommonHTML/fonts/TeX/fontdata.js

- iii. **Single variant association test statistics, including direction of effect.** We use score statistics calculated at each variant site. Another possible option would be to share estimates of genetic effects – but, when variant frequencies are low, score statistics are more numerically stable and preferred.
- iv. **Covariance matrix for each genetic region.** We compute the genotype covariance for all variants in a sliding window, with a width of 500,000 base pairs. This matrix reflects linkage disequilibrium in the region and will be used for meta-analysis of aggregate region- or gene-based tests of rare/low-frequency variation.

2. Indexing the VCF files

rvtest works with tabix, and takes indexed bgzipped VCF files as input.

If your VCF file is not compressed with bgzip please use the following command

```
(grep ^"#" $your_old_vcf; grep -v ^"#" $your_old_vcf | sed 's:^chr::ig' | sort -k1, 1n -k2, 2n) | bgzip -c > $your_vcf_file
```

Please index your bgzip compressed VCF files using the following command:

```
tabix -f -p vcf $your_vcf_file
```

Please make sure there are no entries with -1's or 0's in the chromosome number column.

3. Specify VCF files

You should use `-inVcf $your_vcf_file` to specify which VCF to use.

4. Specify phenotypes

The rvtest tool requires a simple pedigree file that starts with the standard 5 columns (family id, individual id, father id, mother id and sex) followed by trait or trait residuals.

You can use `-mpheno $phenotypeColumnName` or `-pheno-name` to specify a given phenotype.

An example phenotype file, (example.pheno), has the following format:

```
fid iid fatid matid sex y1 y2 y3 y4
```

```
P1 P1 0 0 0 1.7642934435605 -0.733862638327895 -0.980843608339726 1
```

```
P2 P2 0 0 0 0.457111744989746 0.623297281416372 -2.24266162284447 0
```

```
P3 P3 0 0 0 0.566689682543218 1.44136462889459 -1.6490100777089 0
```

```
P4 P4 0 0 0 0.350528353203767 -1.79533911725537 -1.11916876241804 0
```

```
P5 P5 0 0 1 2.72675074738545 -1.05487747371158 -0.33586430010589 1
```

Phenotype file is specified by the option `-pheno example.pheno`. The default phenotype column header is “y1”. If you want to use alternative columns as phenotype for association analysis (e.g the column with header y2), you may specify the header names using either

- `-mpheno 2` or
- `-pheno-name y2`

NOTE: to use “-pheno-name” the header line must start with “fid iid” as PLINK requires.

In phenotype file, missing values can be denoted by NA or any non-numeric value. Individuals with missing phenotypes will be automatically dropped from subsequent association analysis. For each missing phenotype value, a warning will be generated and recorded in the log file.

Optional: If using `rvtests` for phenotype transformation:

If you would like to calculate residuals using the `rvtest` software please refer to the covariates that need to be included for each trait as described in the phenotype transformation step and use `-covar` and `-covar-name` options to designate covariates, and the `-inverseNormal` and `-useResidualAsPhenotype` while performing the analysis.

The covariate file, (e.g. example.covar) has a similar format as the phenotype file, e.g.:

```
fid iid age bmi covCenter1 covCenter2 covCenter3
```

```
Loading [MathJax]/jax/output/CommonHTML/fonts/TeX/fontdata.js
```

P1 P1 23 24.2 1 0 0

P2 P2 32 29.0 0 1 0

P3 P3 44 32.4 0 0 1

P4 P4 25 28.2 1 0 0

NOTE: Missing data in the covariate file can be labeled by any non-numeric value (e.g. NA). Missing values will automatically be imputed to the mean value in analysis.

5. Generate kinship matrices

We ask all studies to use an **empirical kinship matrix (aka genomic relationship matrix)** for use in analysis with rvtests. If you do not already have a kinship matrix for your study it can be generated with rvtests.

Kinship matrices are needed **for all studies** to account for familial relatedness, cryptic relatedness, and population stratification. rvtests generates an empirical kinship matrix from the VCF file. Within the rvtests folder there is a script called “vcf2kinship”. However, we want to run it on all common markers ($MAF \geq 5\%$) genome-wide so we first must create a concatenated VCF file from all chromosomes. If this step is still too slow, you may want to further prune the VCF file with $MAF \geq 5\%$ markers to include only genotyped variants – please contact us if you need advice on this step.

Concatenate per-chromosome imputed polymorphic genotypes VCF files into single VCF file

(zcat yourvcffile.HRCimputed.chr1.imputed.poly.vcf.gz;

*zgrep -v '^#' yourvcffile.HRCimputed.chr2.imputed.poly.vcf.gz; *

*zgrep -v '^#' yourvcffile.HRCimputed.chr3.imputed.poly.vcf.gz; *

*zgrep -v '^#' yourvcffile.HRCimputed.chr4.imputed.poly.vcf.gz; *

*zgrep -v '^#' yourvcffile.HRCimputed.chr5.imputed.poly.vcf.gz; *

*zgrep -v '^#' yourvcffile.HRCimputed.chr6.imputed.poly.vcf.gz; *

Loading [MathJax]/jax/output/CommonHTML/fonts/TeX/fontdata.js

```

zgrep -v '^#' yourvcffile.HRCimputed.chr7.imputed.poly.vcf.gz; \
zgrep -v '^#' yourvcffile.HRCimputed.chr8.imputed.poly.vcf.gz; \
zgrep -v '^#' yourvcffile.HRCimputed.chr9.imputed.poly.vcf.gz; \
zgrep -v '^#' yourvcffile.HRCimputed.chr10.imputed.poly.vcf.gz; \
zgrep -v '^#' yourvcffile.HRCimputed.chr11.imputed.poly.vcf.gz; \
zgrep -v '^#' yourvcffile.HRCimputed.chr12.imputed.poly.vcf.gz; \
zgrep -v '^#' yourvcffile.HRCimputed.chr13.imputed.poly.vcf.gz; \
zgrep -v '^#' yourvcffile.HRCimputed.chr14.imputed.poly.vcf.gz; \
zgrep -v '^#' yourvcffile.HRCimputed.chr15.imputed.poly.vcf.gz; \
zgrep -v '^#' yourvcffile.HRCimputed.chr16.imputed.poly.vcf.gz; \
zgrep -v '^#' yourvcffile.HRCimputed.chr17.imputed.poly.vcf.gz; \
zgrep -v '^#' yourvcffile.HRCimputed.chr18.imputed.poly.vcf.gz; \
zgrep -v '^#' yourvcffile.HRCimputed.chr19.imputed.poly.vcf.gz; \
zgrep -v '^#' yourvcffile.HRCimputed.chr20.imputed.poly.vcf.gz; \
zgrep -v '^#' yourvcffile.HRCimputed.chr21.imputed.poly.vcf.gz; \
zgrep -v '^#' yourvcffile.HRCimputed.chr22.imputed.poly.vcf.gz; \

zgrep -v '^#' yourvcffile.HRCimputed.chrX.reordered.imputed.poly.vcf.gz) | bgzip -c >
yourvcffile.HRCimputed.chrALL.imputed.poly.vcf.gz

```

This will effectively duplicate the imputed genotype data. If disk space is a concern, this alternative requires less storage space by keeping only variants above 5%. Use this for creating the kinship matrix, but the full polymorphic VCFs for association.

```

(bcftools view -q 0.05:minor yourvcffile.HRCimputed.chr1.imputed.poly.vcf.gz -O u; \
bcftools view -q 0.05:minor yourvcffile.HRCimputed.chr2.imputed.poly.vcf.gz -O u \

```



```
| zgrep -v '^#'; \
```

[ditto through chr3-22]

```
bcftools view -q 0.05:minor yourvcffile.HRCimputed.chrX.imputed.poly.vcf.gz -O u \
```

```
| zgrep -v '^#') \
```

```
| bgzip -c > yourvcffile.HRCimputed.chrALL.imputed.mafge5pct.vcf.gz
```

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

```
#!/bin/bash
```

```
vcf2kinship -inVcf yourvcffile.HRCimputed.chrALL.imputed.poly.vcf.gz \
```

```
-ped yourpedfile.ped \
```

```
-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
```

```
-thread 12
```

NOTE: *vcf2kinship* will produce a warning message if there are heterozygous genotype calls present for males in the hemizygous region, and then sets these genotypes to missing to produce the kinship matrix. These samples should have been removed – see section 2.3 above.

6. Exemplar command:

The following are example commands for using *rvtests* to generate score statistics and covariance matrices. It can be used if you have previously generated residuals that have been inverse normally transformed or if you are using *rvtests* to generate residuals and perform the inverse normal transformation.

Loading [MathJax]/jax/output/CommonHTML/fonts/TeX/fontdata.js

Brief example: *rvtests -inVcf input.vcf -pheno phenotype.ped -pheno-name HT_INV -out output -kinship kinship_matrix -meta score,cov[windowSize=500000] -dosage DS*

IMPORTANT NOTE: The [windowSize=500000] option is shell-specific. It works in bash, but not in csh. It may not work in other Unix shells. Therefore, we recommend running all rvtest jobs in a bash shell.

Analyses will loop through chromosomes

Many of the steps in this analysis plan will be done chromosome by chromosome. One very easy way to run by chromosome is with a *for* loop in bash or c-shell. An example is provided below; modify as appropriate for the phenotypes, sexes, and ancestries available in your study, the panel(s) being used for imputation, and for job submission with your compute resources.

NOTE: Some of these lines are not used if you have already generated residuals and performed inverse normal transformation. These are indicated in red. If using rvtest to generate residuals, input covariates are not consistent for all analyses.

For chrX, male genotypes in rvtest can be coded as 0, 1, 0/0, or 1/1, and dosages should be between 0 and 1. rvtest will convert these to a 0-2 scale automatically in analysis. rvtest already has preset values for the pseudo-autosomal regions (PAR), and properly adjusts analysis accordingly. rvtest will perform analysis on both PAR and non-PAR regions, but only calculates HWE and alleles counts in females for non-PAR regions. rvtest association analysis should be conducted on polymorphic variants only.

#!/bin/bash

#replace the following variables and values as needed to match traits, ancestries, sexes, and reference panels available for your study

###trait=TC_INV, logTG_INV, HDL_INV, LDL_INV, TC_RAW, logTG_RAW, HDL_RAW, LDL_RAW, Height, BMI, WHRadjBMI, WHRunadjusted

###ancestry=EUR, AFR, AMR, SAS, EAS

###sex=MEN, WOMEN, ALL

###panel=1KGP3, HRC

for i in 122 X; do #Loop over chromosomes

rvtest -inVcf yourvcffile.\${panel}.chr\${i}.imputed.poly.vcf.gz \ #Input vcf

-pheno study_\${ancestry}_pheno.ped \ #Input phenotype ped file

-pheno-name \${trait} \ #Name the phenotype column in the phenotype ped file

*-meta score, cov[windowSize=500000] \ #Generate single variant score test #and
genotype covariance matrix with a window size of 500kb around #each variant. Use for
only largest N trait (cases and controls)*

-meta score \ #Use for all other traits (does not generate covariance matrix)

*-kinship kinship_matrix.kinship *

-covar study_\${ancestry}_cov.ped \ #Name of covariate file, if using one

*-covar-name sex, age, age2, \ #Names of covariates only if using covariates (that is,
if #residuals not already calculated and inverse normal transformed)*

*-useResidualsAsPhenotype \ #Residualize before testing variants only if not
already #calculated*

*-inverseNormal \ #Inverse normalize the residual distribution only if not
already #calculated*

-xLabel <value> \ #Flag can be added if chrX is coded other than X in VCF (e.g. 23)

*-xHemiKinship kinship_matrix.xHemiKinship \ #use instead of -kinship to
indicate #chrX analysis*

-dosage DS \ #Specify vcf dosage field

-out STUDY_\${ancestry}_\${panel}_imputation_\${trait}_chr\${i}

done

Combine your results files

The following example command, corresponding to the outputs from the example command provided above, will concatenate all per-chromosome results into one output file for each ancestry x trait combination. For other reference panels, phenotypes, etc. please modify as appropriate.

Please combine the set of score test results and the covariance matrix files across chromosomes to generate a single cumulative file across all autosomes for each data type (i.e., 1 metaScore file including all autosomes and 1 metaCov file for all autosomes). Keep and upload chrX separate from the autosomes. You will only upload one metaCov file in total (not one per trait, see File Naming and Uploading in section 8, below.)

Concatenate association results into a single file

#!/bin/bash

trait=TC_INV, logTG_INV, HDL_INV, LDL_INV, TC_RAW, logTG_RAW, HDL_RAW, LDL_RAW, Height, BMI, WHRadjBMI, WHRunadjusted

###ancestry=EUR,AFR,AMR,SAS,EAS #replace this as needed with the appropriate ancestry

###sex=MEN,WOMEN,ALL

###panel=1KGP3,HRC

(

*zgrep -E -h '^1\s|#[CHROM]' STUDY_\${ancestry}_\${panel}_\${trait}_\${sex}_chr1.MetaScore.assoc.gz; *

*zgrep -E -h '^2\s|' STUDY_\${ancestry}_\${panel}_\${trait}_\${sex}_chr2.MetaScore.assoc.gz; *

*zgrep -E -h '^3\s|' STUDY_\${ancestry}_\${panel}_\${trait}_\${sex}_chr3.MetaScore.assoc.gz; *

*zgrep -E -h '^4\s|' STUDY_\${ancestry}_\${panel}_\${trait}_\${sex}_chr4.MetaScore.assoc.gz; *

*zgrep -E -h '^5\s|' STUDY_\${ancestry}_\${panel}_\${trait}_\${sex}_chr5.MetaScore.assoc.gz; *

Loading [MathJax]/jax/output/CommonHTML/fonts/TeX/fontdata.js

```

zgrep -E -h '^6\s' STUDY_${ancestry}_${panel}_${trait}_${sex}_chr6.MetaScore.assoc.gz; \
zgrep -E -h '^7\s' STUDY_${ancestry}_${panel}_${trait}_${sex}_chr7.MetaScore.assoc.gz; \
zgrep -E -h '^8\s' STUDY_${ancestry}_${panel}_${trait}_${sex}_chr8.MetaScore.assoc.gz; \
zgrep -E -h '^9\s' STUDY_${ancestry}_${panel}_${trait}_${sex}_chr9.MetaScore.assoc.gz; \
zgrep -E -h '^10\s' STUDY_${ancestry}_${panel}_${trait}_${sex}_chr10.MetaScore.assoc.gz; \
zgrep -E -h '^11\s' STUDY_${ancestry}_${panel}_${trait}_${sex}_chr11.MetaScore.assoc.gz; \
zgrep -E -h '^12\s' STUDY_${ancestry}_${panel}_${trait}_${sex}_chr12.MetaScore.assoc.gz; \
zgrep -E -h '^13\s' STUDY_${ancestry}_${panel}_${trait}_${sex}_chr13.MetaScore.assoc.gz; \
zgrep -E -h '^14\s' STUDY_${ancestry}_${panel}_${trait}_${sex}_chr14.MetaScore.assoc.gz; \
zgrep -E -h '^15\s' STUDY_${ancestry}_${panel}_${trait}_${sex}_chr15.MetaScore.assoc.gz; \
zgrep -E -h '^16\s' STUDY_${ancestry}_${panel}_${trait}_${sex}_chr16.MetaScore.assoc.gz; \
zgrep -E -h '^17\s' STUDY_${ancestry}_${panel}_${trait}_${sex}_chr17.MetaScore.assoc.gz; \
zgrep -E -h '^18\s' STUDY_${ancestry}_${panel}_${trait}_${sex}_chr18.MetaScore.assoc.gz; \
zgrep -E -h '^19\s' STUDY_${ancestry}_${panel}_${trait}_${sex}_chr19.MetaScore.assoc.gz; \
zgrep -E -h '^20\s' STUDY_${ancestry}_${panel}_${trait}_${sex}_chr20.MetaScore.assoc.gz; \
zgrep -E -h '^21\s' STUDY_${ancestry}_${panel}_${trait}_${sex}_chr21.MetaScore.assoc.gz; \
zgrep -E -h '^22\s' STUDY_${ancestry}_${panel}_${trait}_${sex}_chr22.MetaScore.assoc.gz; \

zgrep -E -h '^X\s' STUDY_${ancestry}_${panel}_${trait}_${sex}_chrX.MetaScore.assoc.gz) | bgzip -c >
STUDY_${ancestry}_${panel}_${trait}_${sex}_STATUS_DATE_ANALYST_ANALYSIS.MetaScore.assoc.gz

```

Then, for the trait with the largest sample size, merge the *.MetaCov.assoc.gz files.

There will be one covariance matrix per ancestry, and separate matrices if analyses were separated into cases and controls.

7 Output Files

Loading [MathJax]/jax/output/CommonHTML/fonts/TeX/fontdata.js

rvtest will generate three files for sharing. The first has the summary statistics (from single variant score tests); the second has linkage disequilibrium (LD) matrices summarizing covariance between single marker score statistics, and the third provides a log file with all parameter settings. These three files are required for central meta-analysis and QC.

Using “-out *output*” options, the results are stored in *output.MetaScore.assoc* and *output.MetaCov.assoc.gz*. Each output file contains a header with summaries of trait and covariates.

The file *output.MetaScore.assoc.gz* file contains the following columns:

CHROM, POS, REF, ALT: chromosome name, position, reference allele and alternative allele

N_INFORMATIVE: number of samples in the analysis

AF: allele frequency

INFORMATIVE_ALT_AC: number of alternative alleles

CALL_RATE: fraction of non-missing genotypes

HWE_PVALUE: Hardy-Weinberg Equilibrium P-value.

N_REF, N_HET, N_ALT: number of reference/reference, reference/alternative, alternative/alternative genotypes respectively.

U_STAT, SQRT_V_STAT: U statistic and square root of V statistic as in the score test.

ALT_EFFSIZE: estimated effect size using alternative allele.

PVALUE: P-value when testing genetic association using score test.

The *output.MetaCov.assoc.gz* contains following columns:

CHROM: Chromosome name

START_POS, END_POS: Chromosomal positions of the first and the last variant in the sliding window.

NUM_MARKER: number of variants in the sliding window.

MARKER_POS: variant chromosomal positions in the sliding window.

COV: covariance for each pair of test statistics in the sliding window.

Loading [MathJax]/jax/output/CommonHTML/fonts/TeX/fontdata.js

Troubleshooting

Time Taken

Anticipated Results

References