Analysis plan

August 2020 - version 1.1

Table of contents

Initial analytic considerations

We would like to promote as much data and results sharing as possible, and so have developed an initial proposal to facilitate these activities. Again, nothing is definitive, but this should serve as a starting point. We are focused on primary association analyses, rather than potential secondary analysis projects.

We recognize that a diverse range of study designs, recruitment and data generation strategies will be pursued to learn more about COVID-19 related outcomes. This diversity of approach is a real strength of this effort, as it will enable a more thorough characterization of all aspects of COVID-19 infection.

Genetic data types

This analysis plan includes instructions for analyzing common variant analysis from array data. For processing of WES and WGS data, this document is a work in progress, but has more content tailored for sequencing data.

Checklist

1. I have a Google account. Otherwise check instructions here.
2. I have completed this form and we'll give you access to the buckets.
3. I have checked the phenotype file.
4. I have performed quality control, for example following what is described in "Recommended pipeline and QC parameters".
5. I have imputed my genotypes as described in "GWAS imputation".
6. I have run association analyses for phenotypes with N.cases > 50 according to the instructions in "Association analysis".
7. I have run analyses separately for each major ancestry group.
8. I have run sex and age-stratified analyses and check that N.cases > 50.
9. I have included sex chromosome in my analyses according to the instructions in "Sex Chromosomes"
10. I have run the checks described in "Results Quality Checks" to assess my results quality
11. I have formatted the summary statistics according to the instructions in "Results format". Importantly I named the files according to the instructions.
12. I have registered my study at the website of COVID-19 HGI. You can check here if the study is registered or otherwise register here.
13. I filled a few information regarding my study in the spreadsheet as described in "Study characteristics collection format".
14. I uploaded the data according to the instructions in "Results Upload Instructions"

Phenotype definitions

(for any question regarding this chapter, please contact andrea.ganna@helsinki.fi or @andrea ganna (Slack)

The phenotypes for analysis are described in the V2 phenotype definitions.
Analysis should also be stratified by age (≤ 60 and > 60 years old - at the time partecipants became cases) and sex (males vs females).

Recommended pipeline and QC parameters

For processing of genotyping data, we would highly recommend using the Ricopili pipeline. Recommendations for QC parameter specifics can be found here and reported below for simplicity. However, every genotyping dataset is slightly different from the other: some tweaking may be required from dataset to dataset.

> SNP QC: call rate ≥ 0.95 (this criterion is useful when merging case and control datasets from different studies)
> Sample QC: call rate in cases or controls ≥ 0.98
> Sample QC: FHET within +/- 0.20 in cases or controls
> Sample QC: Sex violations (excluded) - genetic sex does not match pedigree sex
> SNP QC: call rate ≥ 0.98
> SNP QC: missing difference ≤ 0.02
> SNP QC: Hardy-Weinberg equilibrium (HWE) in controls p value ≥ 1e-06 (i.e., ≥ log 10(p) of -6)
> SNP QC: Hardy-Weinberg equilibrium (HWE) in cases p value ≥ 1e-10 (i.e., ≥ log 10(p) of -10)
> (HWE step should be done only in females when applying to chromosome X. See “Sex chromosomes”)

GWAS imputation

Please use imputed genotypes for analyses, both hg19 and hg38 genome versions are fine. For genotype imputation, please either use your own reference panel, existing imputation panels or use the TopMed imputation server or the Michigan imputation server when possible. Michigan University has certified GDPR compliance of the Michigan server while the TopMed server (who has a larger imputation panel) is not yet GDPR-certified. However, all input data are deleted after imputation and some European studies have therefore been using this server for imputation.

As part of the initiative you have the opportunity to get upgraded in the queue. To that you will need to email: imputationserver@umich.edu, specify the study is part of the COVID19-HGI initiative and they will manually put study at top of queue.

Association analysis

(for any question regarding with chapter, please contact Kumar Veerapen (veerapen@broadinstitute.org - @kumar (he/him/his) (Slack)) or Juha Karjalainen (jkarjala@broadinstitute.org - @Juha Karjalainen (Slack)) or Wei Zhou (wzhou@broadinstitute.org - @Wei Zhou (Slack))

For all genetic studies, the following standard association model should be adopted if possible:

Phenotype ~ variant + age + age^2 + sex + age*sex + PCs + study_specific_covariates

GWAS will be run by each cohort and summary statistics are shared for joint meta-analysis. We recommend using SAIGE for analysis (see "Appendix" for code example), which takes into account relatedness and case-control unbalance.

- Phenotypes are described above in the "Phenotype definitions" chapter.
- Analysis should be run only in N.cases > 50.
- Analyses should also be run separately for males and females removing the sex and age*sex covariates. Report results only for the sex group with N.cases > 50.
- Analyses should be run separately for individuals with age (at the time they became cases) ≤ 60 and > 60. Report results only for the age group with N.cases > 50.
- Analyses should be run separately for each major ancestry group. Report results only for the ancestry group with N.cases > 50.
- We suggest adjusting for 20 PCs.
- *Study_specific_covariate* indicates covariates used to correct technical artifacts (e.g. batch number) and not risk factors or other comorbidities. Avoid to adjust for "heritable" covariates.
- MAF or INFO filtering are not necessary as both will be part of the uploaded summary statistics.

Note: To conduct analysis for a subset of individuals only, you can just use the same SAIGE jobs for all individuals and then pass a phenotypes file with the subset of individuals to test.

Sex chromosomes

(for any question regarding with chapter, please contact Wei Zhou (wzhou@broadinstitute.org - @Wei Zhou (Slack))

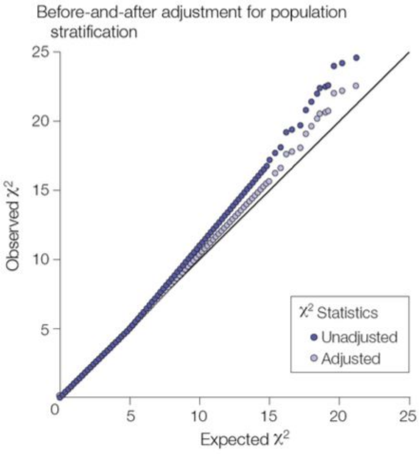
The X chromosome should be included in analyses. If possible, please code females as 0/1/2 and males as 0/2 for X chromosome variants. Pseudoautosomal (PAR) regions should be treated as diploid and analyzed as autosomes. We have modified the SAIGE software to to help process non-PAR regions of chromosome X for association tests so that you don't need to recode the genotypes (see Appendix)

The QC for chromosome X should be done similarly as the autosomal chromosomes. The only QC that should be differentiated is for Hardy-Weinberg Equilibrium (HWE) test: this test should only be done on females. See "Recommended pipeline and QC parameters" for more information about QC.

Results Quality Checks

(for any question regarding with chapter, please contact Kumar Veerapen (veerapen@broadinstitute.org - @kumar (he/him/his) (Slack))

Considering most of the results are churned out via SAIGE, we would suggest adapting 2 scripts available here under the postSAIGE_QC directory. The scripts in this github repository subdirectory will allow you to clean up the header in your SAIGE results output (`postscript.py`) and finally, be able to plot QQ and Manhattan plots using `qqplot.R` (adapted from FinnGen). The package used to plot both plots is `R:qqman`.
Based on this, we would suggest a few steps to determine the quality of your results :



Quantile-Quantile (QQ) plot

The distribution of the analysed $-\log_{10}$(pvalues) is linearly related to the expected values to some point. When there is a deviation from this, it would indicate that there are loci in your dataset which when higher than the expected normal distribution will hold significant (inflated QQ plot) or lower than the expected normal distribution (deflated QQ plot). Using the figure that was graciously adapted from the Analytic and Translational Genetics Unit Workshop 2020, we will describe interpretation of a QQ plot.

When your QQ plot is highly inflated (unadjusted in Figure): the model may need further adjustment i.e. population stratification. Therefore, adjustment with the first 10 principal components (PCs) should adjust stratification out of your model. Another reason for inflation could be from the polygenic architecture in your model. As such, interpretation of the LD score regression intercept would shed light on this situation. Considering the LD score is directly estimating the polygenic effect, the LD score intercept should be approximately 1.0 (or not significant from 1.0). This would remain valid even in the presence of polygenicity. Any deviations from 1.0 would suggest for uncontrolled population stratification. In the Figure, the unadjusted model is highly inflated with a genomic inflation (λ) of 3.2; upon adjustment, the inflation value reduced to 1.2. You can calculate the genomic inflation with the following formula (in R):

median(qchisq(p_value, df=1, lower.tail=FALSE)) / 0.456

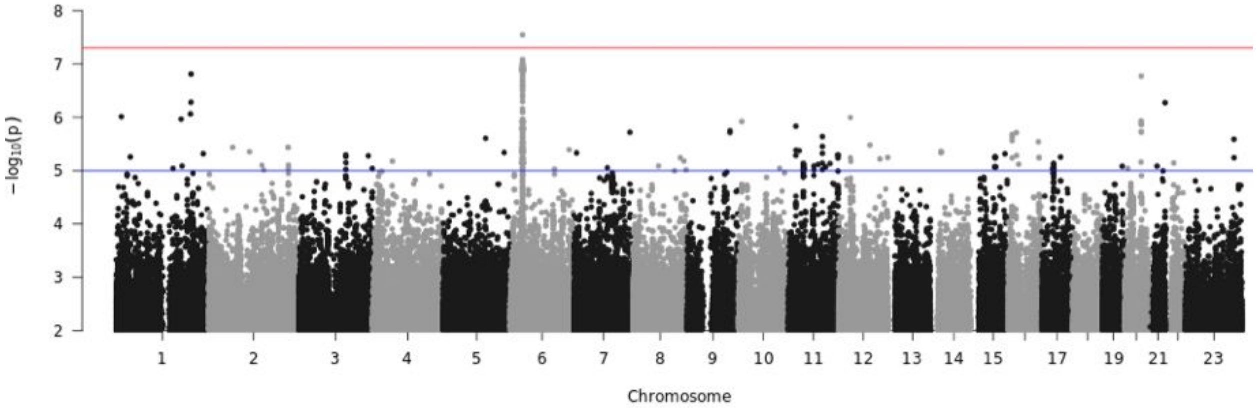
You would generally want a lambda of > 1.0 and less than ~ 1.5. This threshold would imply that you may have significantly associated loci to your trait of interest.

When your QQ plot is deflated (example at end of this section): the model may have some rare variants or that the variance of the tests are insufficient for the model to work well. For the latter, permutations may be needed to improve the quality of your tested model.

Manhattan plot

The easiest way to visualize the results from your analysis would be to plot a Manhattan plot. An example of this would be the plot on the right where the x-axis are chromosomal positions and the y-axis is the -$\log_{10}$(pvalue) from your association analysis. The highest associations will have the smallest p-values and therefore -$\log_{10}$(p value) the highest height in the plot.

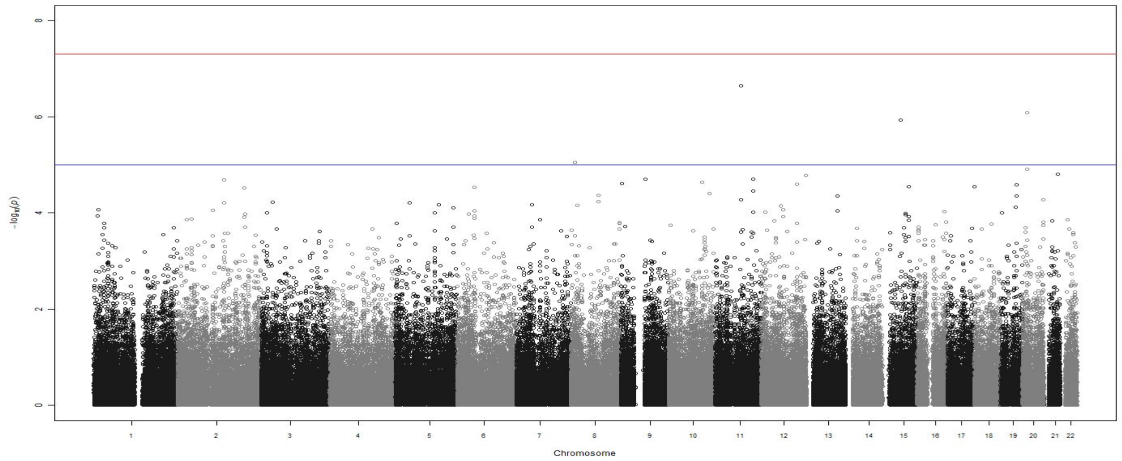
"Good" Manhattan Plot. A Manhattan plot that is considered "good" would have clear LD peaks with few sporadic points i.e. like the Manhattan skyline.



Other than observing significant loci from your association tests, Manhattan plots can also be used to determine if your statistical model needs further adjustments or a more strict quality control is needed. This could be determined by observing the "clarity" of the peaks.

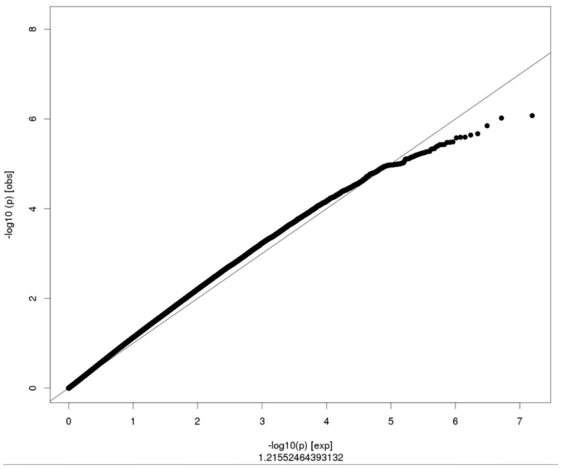
"Bad" Manhattan Plot. If the points are sporadic and lonely loci like the plot below, this might indicate that a more strict quality control is needed. However, with small sample sizes, these sporadic points might simply be due to low allele frequencies variants. Inspect the frequency of those variants. **However**, If they are all < 1%, then it is fine. The following plot was adapted from https://images.app.goo.gl/Js9jg1EpYpJ7NAL77

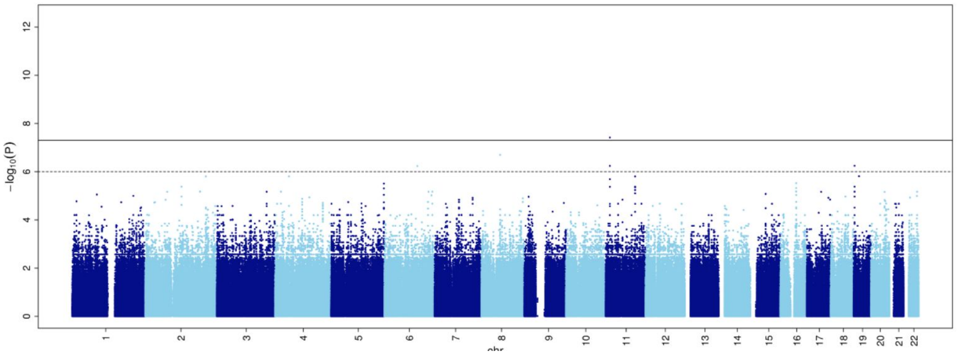
Note: Please share all results without filtering for allele frequencies and INFO score (as specified in the "association analysis" chapter)



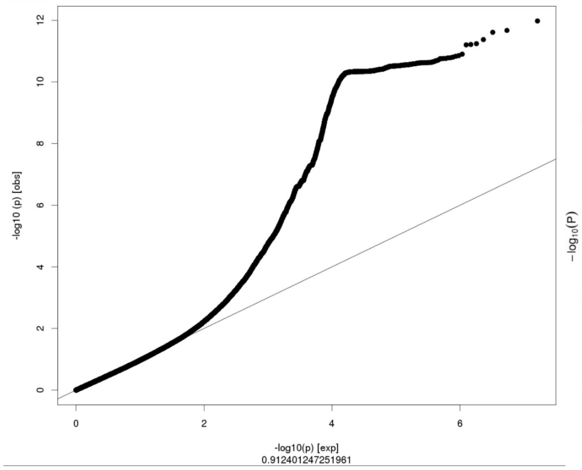
To show the relationship between QQ and Manhattan plots, we elaborate with further examples adapted from the Institute of Behavioral Science, University of Boulder, Colorado.

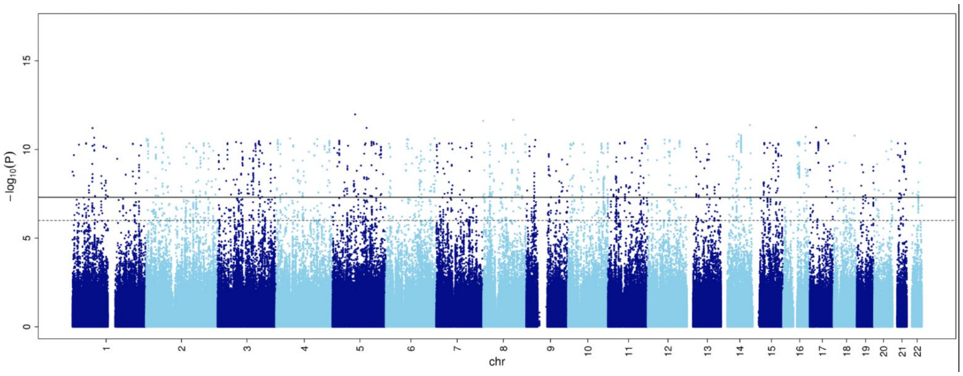
Deflated QQ plot that shows sporadic and poor associations throughout the genome:





Highly inflated QQ plot where everything is seemingly associated (but not) with the trait:





Results format

(for any question regarding with chapter, please contact jkarjala@broadinstitute.org - @Juha Karjalainen (Slack)

The following summary level statistics for GWAS based on an additive genetic model will be generated by each biobank and shared as text files for meta-analysis of variant associations.

1. **Basic variant metrics**, including chromosome positions (GRCh37 (hg19) is preferred, but hg38 is fine as well), reference and alternative alleles (on the forward strand), and allele frequency (for binary traits, allele frequencies in cases and in controls). *Indels: Please list specific nucleotides for indels instead of using I and D and use the chromosome position of the leftmost nucleotide. If coded as R/I/D, please include clarification of what the alleles map to in terms of the leftmost nucleotide.*

2. **Single variant association test statistics,** including the effect size and the standard error of effect size for each variant

   The following fields as returned by the SAIGE GWAS software are recommended for sharing the summary statistics

   CHR POS Allele1 Allele2 AC_Allele2 AF_Allele2 imputationInfo BETA SE p.value Tstat varT AF.Cases AF.Controls N.Cases N.Controls homN_Allele2_cases hetN_Allele2_cases homN_Allele2_ctrls hetN_Allele2_ctrls

   CHR: chromosome
   POS: genome position
   Allele1: allele 1
   Allele2: allele 2 (effect allele)
   AC_Allele2: allele count of allele 2
   AF_Allele2: allele frequency of allele 2
   imputationInfo: imputation quality score
   AF.Cases: allele frequency of allele 2 in cases (only for binary trait)
   AF.Controls: allele frequency of allele 2 in controls (only for binary traits)
   homN_Allele2_cases: homozygote counts in cases
   hetN_Allele2_cases: heterozygote counts in cases
   homN_Allele2_ctrls: homozygote counts in controls
   hetN_Allele2_ctrls: heterozygote counts in controls
   N.Cases: number of cases
   N.Controls: number of controls
   BETA: effect size of allele 2
   SE: standard error of BETA
   p.value: p value
   Tstat: score statistics of allele 2 (if available)
   varT: variance of score statistics (Tstat) (if available)

The minimum set includes

CHR: chromosome
POS: genome position
Allele1: allele 1
Allele2: allele 2 (effect allele)
AF_Allele2: allele frequency of allele 2 in sample
imputationInfo: imputation quality score
BETA: effect size of allele 2
SE: standard error of BETA
p.value: p value

Please use these field names. Chromosome X can be represented by “X” or “23”. Alternative contigs should not be included (only use autosomes and chromosome X). Chromosomes can have “chr” prefix or not (“chr1” or “1”). The file should be sorted by chromosome and base pair position.

To simplify analysts’ life, we will also accept standard output directly from common GWAS softwares such as plink or BOLT-LMM.

Please bgzip or gzip and rename your summary statistics file:

[dataset].[last name].[analysis_name].[freeze_number].[age].[sex].[ancestry].[n_cases].[n_controls].[gwas software].[YYYYMMDD].txt.gz
(*e.g.,* UKBB.Doe.ANA2.1.ALL.ALL.EUR.154.1341.SAIGE.20200414.txt.gz)

[sex] is ALL, MALE or FEMALE.
[age] is ALL, LE_60 or GT_60.

If n_cases and n_controls don’t apply, use

[dataset].[last name].[analysis_name].[freeze_number].[age].[sex].[ancestry].[gwas software].[YYYYMMDD].txt.gz

Ancestry abbreviations are (following the 1000 Genomes definition):
African: AFR
Admixed American: AMR
European: EUR
East Asian: EAS
South Asian: SAS
(Other: please indicate appropriate label and let us know!)

Sex abbreviations are:
Males and Females: ALL
Males: M

Females: F

Age abbreviations are:

All ages: ALL
Lower equal than 60: LE_60
Greater than 60: GT_60

Study characteristics collection format

Few simple sample descriptive statistics should be submitted using this google spreadsheet form.

Simply add your study in a new row in the spreadsheet. **Analysis-specific information will be extracted from the filename**, so please be consistent in the way you name your files (see "Results format")

Results upload instructions

(for any question regarding with chapter, please contact sbryant@broadinstitute.org - @samcbryant (Slack) or jkarjala@broadinstitute.org - @Juha Karjalainen (Slack)

Access to Google Cloud

To upload your data or access data in the COVID-19-hg buckets, you will first need a Google Account. You can also use your existing email address and link it to a Google Account. Here are the instructions on the Google site.

Once you have a Google Account, complete this form and we'll give you access to the buckets.

Bucket for uploading data

There will be an upload bucket created for each group. Use your upload bucket to upload your data. QC checks will be done with uploaded data and the data are then transferred to the analysis bucket.

To upload data, go to your upload bucket (need to be logged in with the given Google account) at https://console.cloud.google.com/storage/browser/covid19-hg-upload-STUDY?forceOnBucketsSortingFiltering=false&project=covid-19-hg replacing STUDY with your study name, and use the Upload files or Upload folder buttons. Upload times vary depending on the network at your institution. If you prefer to use a command line utility, use gsutil. Instructions for gsutil use can be found here. Example command:

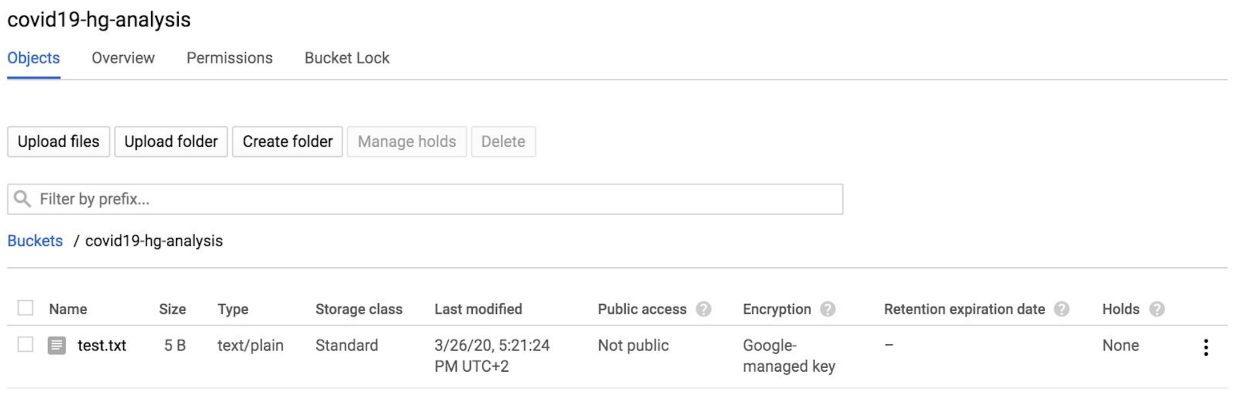
```
gsutil cp UKBB.Doe.ANA2.1.ALL.ALL.EUR.154.1341.SAIGE.20200414.txt.gz
gs://covid19-hg-upload-YOUR_STUDY/
```

For each file that you upload you need to add the corresponding information to the spreadsheet as described in "Study characteristics collection format". Please notice that first you need to create a tab for your study.

Bucket for downloading data

covid19-hg-analysis - use this bucket to read or download data provided by participants, as well as analysis results.

To download data from the covid19-hg-analysis bucket, check the checkboxes next to the files or folders, click the 3 vertical periods (ellipses) at the far right and select Download



covid19-hg-analysis

□	Name	Size	Type	Storage class	Last modified	Public access	Encryption	Retention expiration date	Holds
□	test.txt	5 B	text/plain	Standard	3/26/20, 5:21:24 PM UTC+2	Not public	Google-managed key	–	None

Appendix

(for any question regarding with chapter, please contact wzhou@broadinstitute.org - @Wei Zhou (Slack)

Analysis instructions for conducting GWAS using SAIGE

Here are instructions for using the SAIGE R library which we recommend for the single-variant association tests and gene-based association tests. This approach provides robust statistics for scenarios such as rare variants and highly skewed case:control ratios, while accounting for sample relatedness.

- SAIGE contains two steps to perform single-variant association tests. In step 1, a null logistic mixed model (for binary phenotypes) or a null linear mixed model (for quantitative phenotypes) is fitted. In step 2, for each genetic variant, a score test is conducted based on the results from step 1.

  The detailed tutorial can be found for

  1. Installing the SAIGE R library: https://github.com/weizhouUMICH/SAIGE/wiki/Genetic-association-tests-using-SAIGE#installing-saige

  2. Running SAIGE for single-variant association tests and examples: https://github.com/weizhouUMICH/SAIGE/wiki/Genetic-association-tests-using-SAIGE#single-variant-association-tests

  3. Running SAIGE-GENE for gene-based association tests and examples: https://github.com/weizhouUMICH/SAIGE/wiki/Genetic-association-tests-using-SAIGE#region--or-gene-based-association-tests-saige-gene

  The default setting of parameters are recommended. Below are the example commands (scripts and example data can be found in this folder https://github.com/weizhouUMICH/SAIGE/tree/master/extdata)

  ```
  cd /extdata/
  less -S cmd.sh
  ```

  Input files are in the folder extdata/input, output files are in the folder extdata/output and to obtain help information of the scripts that call functions in the SAIGE library

  ```
  Rscript createSparseGRM.R --help
  Rscript step1_fitNULLGLMM.R --help
  ```

```
Rscript step2_SPAtests.R --help
```

To obtain help information of functions in the SAIGE library

```
#open R
R
library(SAIGE)
?createSparseGRM
?fitNULLGLMM
?SPAGMMATtest
```

- Single-variant association tests

  Step 1: fitting a null mixed model

  - For binary traits, a null logistic mixed model will be fitted (--traitType=binary).
  - For quantitative traits, a null linear mixed model will be fitted (--traitType=quantitative) and needs to be inverse normalized (--invNormalize=TRUE)

    ```
    #check the help info for step 1
    Rscript step1_fitNULLGLMM.R --help
    ```

    Details can be found for input files and output files

```
      Rscript step1_fitNULLGLMM.R     \
             --plinkFile=./input/nfam_100_nindep_0_step1_includeMoreRareVariants_poly \
             --phenoFile=./input/pheno_1000samples.txt_withdosages_withBothTraitTypes.txt \
             --phenoCol=y_binary \
             --covarColList=x1,x2 \
             --sampleIDColinphenoFile=IID \
             --traitType=binary        \
             --outputPrefix=./output/example_binary \
             --nThreads=4
```

  Step 2: single-variant association tests

  Details can be found for input files and output files

  ```
  Rscript step2_SPAtests.R \
  ```

```
        --bgenFile=./input/genotype_100markers.bgen \
        --bgenFileIndex=./input/genotype_100markers.bgen.bgi \
        --minMAF=0.0001 \
        --minMAC=1 \
        --sampleFile=./input/samplefileforbgen_10000samples.txt \
        --GMMATmodelFile=./output/example.rda \
        --varianceRatioFile=./output/example.varianceRatio.txt \
        --SAIGEOutputFile=./output/example.SAIGE.bgen.txt \
        --numLinesOutput=2 \
        --IsOutputNinCaseCtrl=TRUE \
        --IsOutputHetHomCountsinCaseCtrl=TRUE \
        --IsOutputAFinCaseCtrl=TRUE
```

--IsOutputAFinCaseCtrl=TRUE is specified to output the allele frequencies in cases and controls
--IsOutputNinCaseCtrl=TRUE is specified to output the sample sizes of cases and controls
--IsOutputHetHomCountsinCaseCtrl=TRUE is specified to output the homozygote and heterozygote counts in cases and controls

SAIGE has been updated to version 0.39.2 to help process non-PAR regions of chromosome X for association tests. The link contains the instructions for installation
https://github.com/weizhouUMICH/SAIGE/tree/0.39.2

Three arguments have been added in Step 2 to test chromosome X, which multiplies genotypes/dosages of non-PAR regions for males by 2 to convert 0/1 to 0/2

--sampleFile_male=SAMPLEFILE_MALE
Path to the file containing one column for IDs of MALE samples in the bgen or vcf file with NO header. Order does not matter

--X_PARregion=X_PARREGION
Ranges of (pseudoautosomal) PAR region on chromosome X, which are seperated by comma and in the format start:end. ***By default: '60001-2699520,154931044-155260560' in the UCSC build GRCh37/hg19****. For males, there are two X alleles in the PAR region, so PAR regions are treated the same as autosomes. In the NON-PAR regions (outside the specified PAR regions on chromosome X), for males, there is only one X allele. If is_rewrite_XnonPAR_forMales=TRUE, genotypes/dosages of all variants in the NON-PAR regions on chromosome X will be multiplied by 2. If the genome version* ***is in the UCSC build GRCh38/hg38, you need to use --X_PARregion=10001-2781479,155701383-156030895*** ***https://en.wikipedia.org/wiki/Pseudoautosomal_region***

--is_rewrite_XnonPAR_forMales=IS_REWRITE_XNONPAR_FORMALES
Whether to rewrite genotypes or dosages in the NON-PAR regions on chromosome X for males (multiply by 2). By default, FALSE. ***Note, only use is_rewrite_XnonPAR_forMales=TRUE when the specified VCF or Bgen file only has variants on chromosome X. When is_rewrite_XnonPAR_forMales=TRUE, the program DOES NOT check the chromosome value by assuming all variants are on chromosome X***

The following script can be used to run single-variant association tests for chromosome X **if non-PAR regions on chromosome X have not been coded as 0/2 for males in VCF or BGEN**

```
   Rscript step2_SPAtests.R        \

   --vcfFile=./input/genotype_10markers.missingness_chrX.vcf.gz \
   --vcfFileIndex=./input/genotype_10markers.missingness_chrX.vcf.gz.t
bi        \
   --vcfField=GT \
   --chrom=chrX \
   --minMAF=0.0001 \
   --minMAC=1 \
   --GMMATmodelFile=./output/example_binary.rda \
   --varianceRatioFile=./output/example_binary.varianceRatio.txt \
   --SAIGEOutputFile=./output/example_binary.SAIGE.vcf.genotype.missin
   gness_chrX.txt \
   --numLinesOutput=2 \
   --IsOutputAFinCaseCtrl=TRUE     \
   --is_rewrite_XnonPAR_forMales=TRUE      \
   --X_PARregion=1-9,12-15 \      #Note the numbers used here are
   just for example data. If your data are in the UCSC build
   GRCh37/hg19, you don’t need to specify this argument as the default
   setting is 60001-2699520,154931044-155260560. If your data are in
   the UCSC build GRCh38/hg38, you need to specify --X_PARregion=
   10001-2781479,154931044-155260560
   --sampleFile_male=./input/sampleid_males.txt
```