Analysis plan

16th april 2020 - version 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

We will include GWAS, whole exome sequencing (WES) and whole genome sequencing (WGS). These different data types require different data processing, quality control and analytic approaches. Where possible, we would like to align on these steps, but will not require complete consistency across the different contributions. At the end of this document is a series of appendices for more detailed information for processing each data type.

Phenotype definitions

The phenotypes for analysis are described in this document. Studies might not be able to perform analysis for all the phenotypes. We will ask to prioritize phenotypes listed in the “minimal analysis”.

GWAS imputation

Imputation panel

Please use imputed genotypes for analyses. 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.

HLA imputation

The HLA plays a critical role in human immune response. As such, we encourage HLA imputation. In the future, it will be possible to use the multi-ethnic HLA reference panel of ~21,000 individuals constructed based on deep-coverage whole genome sequencing data. This will be available through the Michigan imputation server and the TopMed imputation server. Until then, please use your own reference panel for HLA imputation.

Association analysis

Primary association model

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. When possible, please also run GWAS separately for males and females removing the sex and age*sex covariates. Phenotypes are described above in the "**Phenotype definitions**" chapter.

The X chromosome should be included in analyses. When possible, please code females as 0/1/2 and males as 0/2 for X chromosome variants. Pseudoautosomal (PAR) regions can be treated as diploid.

In addition to GWAS, gene burden tests are recommended to be run by each cohort when possible.

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.

HLA association study

The HLA plays a critical role in human immune response and has been shown to contribute to susceptibility and course for a variety of infections. With that backdrop, we propose two HLA association models using three different types of variants in the HLA region: SNPs, amino acids (AAs; *e.g.*, HLA-A AA position 9) and HLA classical alleles separately for each digit field (*e.g.*, HLA-A*01 for 1-field and HLA-A*01:01 for 2-field).

Single-variant association

Phenotype ~ variant (SNPs/AAs/Classical alleles) + age + age^2 + sex + age*sex + PCs + study_specific_covariates

Similar to the primary association model, we here ask to test association between phenotype and any type of variants (presence or absence) in the HLA.

Joint association per AA position

> Phenotype ~ **{variants}** (AA changes in the same position) + age + age^2 + sex + age*sex + PCs + study_specific_covariates

In addition, please conduct a joint regression analysis for multi-allelic AA changes of the same position (*e.g.,* HLA-A AA position 9 F/S/T/Y) if possible. Here, we test the effects of AA substitutions simultaneously for the position. To avoid collinearity, please exclude the most frequent AA change from the model (based on the multi-ethnic HLA reference panel). Given individual genotype data is not always available to conduct an omnibus test across all the studies, we plan to conduct joint regression meta-analysis (Becker & Wu, 2007; Vaitsiakhovich, et al., 2015) to test an omnibus effect of each AA position (null hypothesis: all betas of AA changes = 0). To this end, please include the variance-covariance matrix of coefficients as well. Analysis script for the HLA-imputed data through the Michigan imputation server will be planned to be distributed.

Results format

Dataset descriptives

Sample descriptive statistics including demographic information, distributions of age, sex breakdown, ancestry composition (e.g. 1KG PCA projections) and phenotypes should be submitted using this form each time data are uploaded. We will develop further guidance on quality control reports.

GWAS results format

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.

Please bgzip or gzip and rename your summary statistics file:

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

[sex] is ALL, MALE or FEMALE. If n_cases and n_controls don’t apply, use

[dataset].[last name].[analysis_name].[freeze_number].[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

Gene-based analysis results format

The summary level statistics below will be generated and shared by each biobank performing gene-based tests from WES and WGS analyses using burden tests (using an additive genetic model) or variance tests. These burden tests are focused on rare variation, rather than a common variant version of a gene-based test.

Considerations: Gene and variant annotations should ideally be performed using GENCODE v19 if on GRCh37/hg19; GENCODE v29 if on GRCh38. For loss-of-function annotations, LOFTEE can be used to automatically filter variants for high confidence LoF variants. These are the defaults implemented in Hail 0.2. However, if not using this pipeline, stop-gained (nonsense), essential splice (donor and acceptor), and frameshift variants should be grouped together when performing pLoF burden tests.

1. **Gene and annotation information:** genes should be identified by HGNC symbols (and additionally Ensembl gene IDs if convenient). Annotations should be one of \`pLoF\` (see grouping above), \`missense\`, \`missense-probably_damaging\` (annotated by PolyPhen-2), or \`synonymous\`. Other annotations should be noted as the original VEP annotations (e.g. \`splice_region_variant\`).

2. **Gene burden association test statistics,** including the effect size and the standard error of effect size for each gene.

   The following elements are recommended for the sharing summary statistics (returned in SAIGE 0.38 or later).

   Gene Pvalue markerIDs markerAFs Pvalue_Burden Pvalue_SKAT BETA_Burden SE_Burden

   Gene: The identifier for the gene-based test. Suggested format: EnsemblID_GeneSymbol_annotation (annotation as above; e.g. ENSG00000197780_TAF13_pLoF)
   Pvalue: p value of SKAT-O test
   markerIDs: semi-colon delimited list of marker IDs used in test in the format: chr1:109065000_A/T
   markerAFs: semi-colon delimited list of allele frequencies corresponding to each entry in markerIDs
   Pvalue_Burden: p value of Burden test
   Pvalue_SKAT: p value of SKAT test
   BETA_Burden: Burden test effect size
   SE_Burden: standard error of BETA_Burden

Chromosomes can have chr prefix or not (“chr1” or “1”), but ideally “1” for GRCh37, “chr1” for GRCh38.

Please **bgzip** and name your summary statistics file as recommended in GWAS results format above.

Study characteristics collection format

To be collected across all the analyses and filled each time data are uploaded using this form

- Study/cohort name
- Contact email
- Uploaded file name
- Study freeze number
- Analysis name
- Sex (all/male/female)
- Analysis type (GWAS/HLA/burden)
- N cases and controls (if applicable) and N individuals
- Association testing software and its version (SAIGE preferred)
- % female
- Age (mean, SD)
- Period covered by the analysis (min date, max date). E.g. period of EHR data extraction
- Ancestry definition (AFR, AMR, EUR, EAS, SAS)
- Chromosome build (37 or 38)
- X chromosome male coding (0/2 (recommended) or 0/1)
- Imputation panel used
- Other analysis specific information

Analysis-specific (analysis names are listed in the phenotype document):

v1.0 ANALYSES - please see phenotype document which is more up to date

Analysis 1:
- Definition of COVID-19 cases (report diagnostic codes used or If PCR/antibody based diagnosis)
- Definition of respiratory support (report diagnostic code or other approaches to collect this information and which respiratory support techniques were included)
- Number of deceased individuals (% total) among those on respiratory support
- Time (in days) from diagnosis to respiratory support (mean, SD)
- Time (in days) from hospitalization to respiratory support (mean, SD)

Analysis 2:
- Definition of COVID-19 cases (report diagnostic codes used or If PCR/antibody based diagnosis)

- Time (in days) from diagnosis to hospitalization (mean, SD)

Analysis 3:

- Definition of COVID-19 cases (report diagnostic codes used or If PCR/antibody based diagnosis)
- Time (in days) from diagnosis to death (mean, SD)

Analysis 4:

- Definition of COVID-19 cases (report diagnostic codes used or If PCR/antibody based diagnosis)
- Definition of ordinal severity scale (specify criteria to define: 1) mild 2) severe 3) critical)
- Severity distribution (% of mild, severe and critical)

Analysis 5:

- Definition of COVID-19 cases (report diagnostic codes used or If PCR/antibody based diagnosis)

Analysis 6:

- Definition of COVID-19 cases (report diagnostic codes used or If PCR/antibody based diagnosis)

Analysis 7:

- List of flu-like symptoms included
- Distribution of flu symptoms (% of individuals reporting each flu symptom)

Results upload instructions

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:

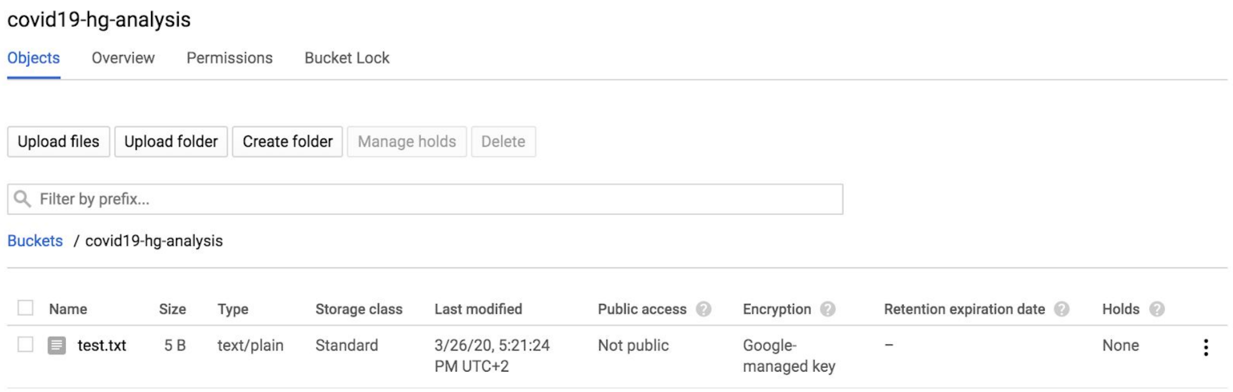
```
gsutil cp UKBB.Doe.ANA2.1.ALL.EUR.154.1341.SAIGE.20200414.txt.gz gs://covid19-hg-upload-YOUR_STUDY/
```

Each time you upload data, please fill in this form, information to be included is listed in the "**Study characteristics collection format**" chapter.

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

Buckets / 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

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_includeMoreRareVa
             riants_poly \
             --phenoFile=./input/pheno_1000samples.txt_withdosages_withBot
             hTraitTypes.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

- Gene-based association tests

Step 0: creating a sparse GRM

- Note: This step is only needed for region- and gene-based tests (SAIGE-GENE) The sparse GRM only needs to be created once for each data set and can be used for all different phenotypes as long as all tested samples are included in it.

Details can be found for input files and output files

```
#check the help info for step 0
Rscript createSparseGRM.R --help

createSparseGRM.R \
      --plinkFile=./input/nfam_100_nindep_0_step1_includeMoreRareVa
riants_poly \
      --nThreads=4  \
      --outputPrefix=./output/sparseGRM    \
      --numRandomMarkerforSparseKin=2000   \
      --relatednessCutoff=0.125
```

Step 1: fitting the null logistic/linear mixed model

- Step 1 model results from the single-variant assoc test can be re-used, except that for gene-based tests, variance ratios for multiple MAC categories and a sparse GRM need to be used (IsSparseKin=TRUE).

Details can be found for input files and output files

- If Step 1 has been done previously for single-variant association tests, for gene-based tests, `--skipModelFitting=TRUE can be used to skip the model fitting for gene-based tests but` variance ratios for multiple MAC categorie still need to be estimated

```
Rscript step1_fitNULLGLMM.R \
      --plinkFile=./input/nfam_100_nindep_0_step1_includeMore
      RareVariants_poly \

      --phenoFile=./input/pheno_1000samples.txt_withdosages_w
      ithBothTraitTypes.txt \
      --phenoCol=y_binary \
      --covarColList=x1,x2 \
      --sampleIDColinphenoFile=IID \
      --traitType=binary       \
      --invNormalize=TRUE     \
      --outputPrefix=./output/example_binary \ ##specify the
prefix of the model file from the previous Step 1
      --outputPrefix_varRatio=./output/example_binary_cate
      \ ##specify the prefix for the variant ratios to be
      estimated
      --sparseGRMFile=./output/example_binary_cate.varianceRa
      tio.txt.sparseGRM.mtx    \
      --sparseGRMSampleIDFile=./output/example_binary.varianc
      eRatio.txt.sparseGRM.mtx.sample  \
      --nThreads=4 \
      --LOCO=FALSE    \
      --skipModelFitting=TRUE \    ##skip the model fitting
      --IsSparseKin=TRUE      \
      --isCateVarianceRatio=TRUE
```

- If Step 1 has not been done previously for single-variant association tests

```
Rscript step1_fitNULLGLMM.R \
      --plinkFile=./input/nfam_100_nindep_0_step1_includeMore
      RareVariants_poly \

      --phenoFile=./input/pheno_1000samples.txt_withdosages_w
      ithBothTraitTypes.txt \
      --phenoCol=y_binary \
      --covarColList=x1,x2 \
      --sampleIDColinphenoFile=IID \
      --traitType=binary       \
      --invNormalize=TRUE     \
      --outputPrefix=./output/example_binary \ ##specify the
      prefix of the model file and file for variant ratios
```

```
        --sparseGRMFile=./output/example_binary_cate.varianceRatio.txt.sparseGRM.mtx    \
        --sparseGRMSampleIDFile=./output/example_binary.varianceRatio.txt.sparseGRM.mtx.sample  \
        --nThreads=4 \
        --LOCO=FALSE    \
        --skipModelFitting=FALSE \    ##NOT skip the model fitting
        --IsSparseKin=TRUE      \
        --isCateVarianceRatio=TRUE
```

Step 2:performing the region- or gene-based association tests

Details can be found for input files and output files

```
Rscript step2_SPAtests.R \
        --vcfFile=./input/seedNumLow_126001_seedNumHigh_127000_nfam_1000_nindep_0.sav \
        --vcfFileIndex=./input/seedNumLow_126001_seedNumHigh_127000_nfam_1000_nindep_0.sav.s1r \
        --vcfField=DS \
        --chrom=chr1 \
        --minMAF=0 \
        --minMAC=0.5 \
        --maxMAFforGroupTest=0.01       \
        --sampleFile=./input/samplelist.txt \
        --GMMATmodelFile=./output/example_binary.rda \
        --varianceRatioFile=./output/example_binary_cate_v2.varianceRatio.txt \
        --SAIGEOutputFile=./output/example_binary.SAIGE.gene.txt \
        --numLinesOutput=1 \
        --groupFile=./input/groupFile_geneBasedtest.txt    \
        --sparseSigmaFile=./output/example_binary_cate_v2.varianceRatio.txt_relatednessCutoff_0.125.sparseSigma.mtx \
        --IsOutputAFinCaseCtrl=TRUE     \
        --IsSingleVarinGroupTest=TRUE    \
        --IsOutputPvalueNAinGroupTestforBinary=TRUE     \
        --IsAccountforCasecontrolImbalanceinGroupTest=TRUE
```

If you need assistance, do not hesitate to contact us.
Wei Zhou <wzhou@broadinstitute.org>