**Analysis plan**

**16th april 2020 - version 1**

**Table of contents**

[**Initial analytic considerations**](#_ph1gwckwigtp) **2**

[**Genetic data types**](#_boylaquxasqd) **2**

[**Phenotype definitions**](#_t8uk7kn78u7y) **2**

[**GWAS imputation**](#_tb1oy6dpogsm) **2**

[**Imputation panel**](#_6y42m4bxp1s0) **2**

[**HLA imputation**](#_7z8rud4owdqm) **3**

[**Association analysis**](#_hsap0amda0z9) **4**

[**Primary association model**](#_87tbvmly0gib) **4**

[**HLA association study**](#_gjp6lmy7qzrt) **4**

[**Results format**](#_n3lsavj9jwgw) **6**

[**GWAS results format**](#_qrdyb34f2ur3) **6**

[**Gene-based analysis results format**](#_clitx31gja7m) **7**

[**Study characteristics collection format**](#_yjg1rri7avhh) **9**

[**Results upload instructions**](#_9jyc3fycronl) **11**

## 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](https://docs.google.com/document/d/1eMdzhO5xk-MACxjz-kOUJLP6Jort5KuwoOa_u-aZPHs/edit?usp=sharing). 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](https://imputation.biodatacatalyst.nhlbi.nih.gov/) or the [Michigan imputation server](https://imputationserver.sph.umich.edu/index.html) 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](mailto: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](https://imputationserver.sph.umich.edu/index.html) and the [TopMed imputation server](https://imputation.biodatacatalyst.nhlbi.nih.gov/). 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 + age2 + sex + age\*sex + PCs + covariates

GWAS will be run by each cohort and summary statistics are shared for joint meta-analysis. When possible, please include sex-stratified GWAS for males and females removing the sex and age\*sex covariates.

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.

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

### 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 (*e.g.*, HLA-A\*01:01).

*Single-variant association*

Phenotype ~ variant (SNPs/AAs/Classical alleles) + age + age2 + sex + age\*sex + PCs + 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 + age2 + sex + age\*sex + PCs + 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 effect of AA changes simultaneously for the position. To avoid collinearity, we exclude the most frequent AA change from the model (based on the multi-ethnic HLA reference panel). When contributing, please include the variance-covariance matrix of coefficients as well (for joint meta-analysis). Analysis script for the HLA-imputed data through the Michigan imputation server will be planned to be distributed.

## Results format

### 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.*

1. **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 AF.Cases AF.Controls N.Cases N.Controls BETA SE p.value Tstat varT**

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)

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

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**](http://www.htslib.org/doc/bgzip.html) or gzip and rename your summary statistics file:

**[dataset].[last name].[analysis\_name].[freeze\_number].[ancestry].[n\_cases].[n\_controls].[gwas software].[YYYYMMDD].txt.gz**

(*e.g.,* UKBB.Doe.ANA2.1.EUR.154.1341.SAIGE.20200414.txt.gz)

If n\_cases and n\_controls don’t apply, use

**[dataset].[last name].[analysis\_name].[freeze\_number].[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!)

# 

### 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.

**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](http://github.com/konradjk/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 0.36.5 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**](http://www.htslib.org/doc/bgzip.html) 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:

- Analysis name

- Analysis type (GWAS/HLA/burden)

- Sex (both/male/female)

- % female

- Age (mean, SD)

- Period covered by the analysis (min date, max date). E.g. period of EHR data extraction

- Ancestry definition (method used, % individuals in each ancestry group)

- N cases and controls or N individuals (if continuous trait)

- Chromosome build (37 or 38)

- X chromosome male coding (0/2 (recommended) or 0/1)

- Imputation panel used

- Uploaded file name

Analysis-specific:

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)

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:

- 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](https://support.google.com/accounts/answer/176347?co=GENIE.Platform%3DDesktop&hl=en) on the Google site.

Once you have a Google Account, complete this form (link to be added) 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) in the [Storage browser of the Google Cloud Console](https://console.cloud.google.com/storage/browser), 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](https://cloud.google.com/storage/docs/gsutil_install). Instructions for gsutil use can be found [here](https://cloud.google.com/storage/docs/gsutil).

Each time you upload data, please fill in this form (link to be added).

Bucket for downloading data

[covid19-hg-analysis](https://console.cloud.google.com/storage/browser/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

