|  |  |
| --- | --- |
| STUDY FULL TITLE | Using whole-genome or whole-exome sequencing to identify host genetic determinants of Covid-19 susceptibility and severity |
| PROGRAM VERSION | 5.0 |
| PROGRAM VERSION DATE | 2021-01-22 |
| Cohorts (alphabetical order) | **Biobanque Québec Covid-19**  Investigator: J Brent Richards, Professor of Medicine, McGill University  Co-Investigators: Guillaume Butler-Laporte, Tomoko Nakanishi, Sirui Zhou, Vincenzo Forgetta  **CHIRP Study**  Investigator: Sulggi Lee  **Columbia COVID-19 Biobank**  Investigators: Krzysztof Kiryluk, David Goldstein  Co-investigator: Gundula Povysil  **Deutsche COVID-19 OMICS Initiative (DeCOI)**  Investigators: Olaf Riess, Eva Schulte, Julien Gagneur, Kerstin Ludwig  **GEN-COVID**  Investigator: Alessandra Renieri  Co-investigator: Simone Furini  **GEN-COVID** from the IDIS (Instituto de Investigación Sanitaria; Spain)  Investigators: Antonio Salas and Federico Martinón  **Helix and Healthy Nevada Project Exome+ Covid-19 Phenotypes**  Investigator: Liz Cirulli  **INTERVAL**  Investigators: Nicole Soranzo, Adam Butterworth, Kousik Kundu, Klaudia Walter  **Japan NCGM-COVID19**  Investigators: Masashi Mizokami, Katsushi Tokunaga  Co-Investigators: Norio Ohmagari, Masaya Sugiyama, Yosuke Kawai, Nicholas F. Parrish, Yoichiro Kamatani  **Penn Medicine Biobank**  Investigators: Anurag Verma, Marylyn Ritchie, Daniel Rader  **POLCOVID-Genomika**  Investigator: Miroslaw Kwasniewski  Co-investigator: Pawel Olszewski  **Qatar Genome**  Investigator: Sail Ismail  Co-investigator: Hamdi Mbarek  **Regeneron Pharmaceuticals Inc.**  Investigator: Manuel Allen Revez Ferreira  Co-investigator: Jack Kosmicki  **Saudi Genome Program**  Investigator: Malak Abedalthagafi  Co-investigator: Manal Alaamery  **Sinai\_Covid**  Investigator: Joseph Buxbaum  **Swedish Covid-19 Cohort**  Investigator: Hugo Zeberg  **UC COVID-19 Host Genetics Consortium**  Investigators: Dan Geschwind, Manish Butte, Bogdan Pasaniuc, Jonathan Sebat  **Weill Cornell Medicine Clinical Omics Cohort**  Investigators: Christopher Mason, Rob Schwartz |

# Table of Contents

[Table of Contents](#_1fob9te) 3

[Abbreviations](#_2et92p0) 4

[Introduction](#_1t3h5sf) 4

[Preface](#_4d34og8) 4

[Scope of the Analyses](#_2s8eyo1) 4

[Principles of Collaboration](#_17dp8vu) 4

[Data Sharing and Participant Confidentiality](#_3rdcrjn) 5

[Study Objectives and Endpoints](#_26in1rg) 5

[Study Objectives](#_lnxbz9) 5

[Modifications from version 4, following data freeze 1](#_35nkun2) 5

[Analysis overview](#_1ksv4uv) 5

[Endpoints, Outcomes, and Measurements](#_2jxsxqh) 6

[Analysis scope](#_z337ya) 6

[Study Methods](#_3j2qqm3) 7

[Participants inclusion criteria](#_1y810tw) 7

[Whole genome or Exome sequencing](#_4i7ojhp) 7

[Genetic variant calling and quality control](#_3whwml4) 7

[Genetic variant annotation](#_2bn6wsx) 7

[Genetic association studies](#_qsh70q) 7

[Individual variants association tests](#_3as4poj) 9

[Meta-analysis](#_1pxezwc) 10

[References](#_147n2zr) 11

[Appendix 1](#_23ckvvd) 12

[Appendix 2](#_31wgor1ihmbq) 13

[Appendix 3](#_3os9uwgfebf4) 15

# 

# Abbreviations

|  |  |
| --- | --- |
| BiPAP | Bilevel positive airway pressure |
| CPAP | Continuous positive airway pressure |
| Covid-19 HGI | Covid-19 Host Genetic Initiative (<https://www.covid19hg.org/>) |
| ECMO | Extracorporeal membrane oxygenation |
| GWAS | Genome-wide association study |
| pLoF | Predicted loss of function genetic variants |
| MAF | Minor allele frequency |
| MAC | Minor allele count |
| NAAT | Nucleic acid amplification test |
| SNP | Single nucleotide polymorphism |
| WES | Whole-exome sequencing |
| WGS | Whole-genome sequencing |

# 

# Introduction

## Preface

The Covid-19 pandemic is likely to be the most disastrous public health emergency of our lifetime, with already hundreds of thousands of deaths worldwide. One of the hallmarks of Covid-19 is the wide range of its clinical presentation and ability to affect seemingly healthy patients. Host genetics is therefore a likely determinant of Covid-19 susceptibility and severity. This is supported by recent heritability studies1, and by the Covid-19 HGI (<https://www.covid19hg.org/>) GWAS results. Therefore, by studying genetic determinants of Covid-19, we may be able to better understand the biology of Covid-19 severity and susceptibility and subsequently prioritize drug development and clinical trials. This study protocol describes the motivation, analysis plan, and data sharing plan for the Covid-19 HGI Working group WES/WGS working group. The working group focuses on the analysis of rare genetic variants, and their impact on Covid-19 susceptibility and outcomes.

## Scope of the Analyses

In this study, we will use WGS or WES data from participants of collaborating cohorts to find genetic determinants of Covid-19 susceptibility, severity, and other associated phenotypes.

## Principles of Collaboration

The WES/WGS working group adheres to the same principles as the Covid-19 HGI (<https://www.covid19hg.org/about/>), including, but not limited to, the principles of transparent and open science, scientific collaboration, and the respect of each group’s data. Hence, while the end goal is to provide freely accessible summary results, each cohort still retains control over their own data and will be able to perform their analysis locally before sharing summary statistics.

## Data Sharing and Participant Confidentiality

Individual patient level data sharing will not be required to participate in this working group. Sharing of summary statistics necessary for pooling of the analysis results will be done using a two-way encrypted HIPAA compliant server (www.sync.com) with a password known only to each cohort and the meta-analysis analyst (GBL, from BQC-19). Any summary data will be deleted from the server once the analysis is complete, and only the pooled summary statistics will be kept and shared.

If data sharing is required between participating cohorts, data sharing agreements will be arranged between these cohorts independently of the working group. Similarly, if researchers are interested in individual patient level data, arrangements will be made with each cohort separately.

# Study Objectives and Endpoints

## Study Objectives

We will perform gene burden tests to find genes associated with Covid-19 susceptibility and adverse outcomes.

## Modifications from version 4, following data freeze 1

Following our first data freeze (December 17, 2020, see protocol here: <https://docs.google.com/document/d/1Rr1mhvY6JViHpkGg_QtoP0VfB0-YNqb2vxscNY7PbVw/edit?usp=sharing>), we have made the following changes to our protocol:

1. This round of analysis will be performed exome wide.
2. The definition of predicted loss of functions has changed to allow for better data harmonization across cohorts.
3. Mask 2 (M2) will be removed, as it was believed to be most prone to false positive associations. M2 included pLOF and missense variants. For clarity, Masks 3 and 4 will still be named M3 and M4 below.
4. We have added a very-rare analysis restricted to genetic variants with minor allele frequencies of less than 0.1%.
5. Association tests of clinical laboratory data was removed

Please see the corresponding sections below for more details.

## Analysis overview

Analysis of rare variants present specific challenges which require tailored solutions. These challenges include:

1. Rare number of genetic events, making single variant association tests difficult to perform.
2. Difficulty in meta-analyzing gene burden tests.
3. Preservation of study participant anonymity.
4. Small sample size.

To alleviate these challenges, we first propose to use a gene burden test that collapses deleterious variants for each gene and hence allows for gene-level analysis, thereby increasing event rate. We will also use Firth penalized likelihood regression2 to provide unbiased effect estimators. Second, our burden test will mimic an additive effect model used in GWAS to allow for meaningful pooling of results and meta-analysis. Third, each cohort will be allowed to perform their own analysis locally before the meta-analysis, and only cohort deidentified meta-analysis results will be shared. This should make patient de-identification from our pooled summary statistics impossible. Therefore, these solutions will allow us to increase the number of Covid-19 cases and considerably increase our sample size.

Another major challenge of WGS and WES data analysis across multiple cohorts is the uniformization of analytic methods to ensure comparable results for the final meta-analysis. A github page with code was made therefore available to guide local cohorts with the analysis, based on the BQC-19 local pipelines, but will still require local adjustments. These are presented as suggestions and may be modified or deviated from based on a local need assessment. It can be accessed here: <https://github.com/DrGBL/WES.WGS>. We also refer to suggested sequencing quality control pipeline here (sections 1 and 2), these should also be tailored based on a local need assessment:

<https://docs.google.com/document/d/1X_qjplH8T4BJXSeMQ_sBfQUTiu_kAisicOqGb6B8hcM/edit>).

## Endpoints, Outcomes, and Measurements

The primary endpoints will follow outcome phenotype definition from the Covid-HGI3. That is, we will study the following three outcomes:

1. Covid-19 susceptibility (C2 in Covid-19 HGI analysis):
   * **Cases**: determined by NAAT (e.g. PCR), serology, medical record review, or self-reported by the patient.
   * **Controls**: every other participant in each cohort that is not a case.
2. Covid-19 hospitalization (B2 in Covid-19 HGI analysis):
   * **Cases**: hospitalized with laboratory confirmed Covid-19. We suggest that a pre-specified timeframe be used to define Covid-19 hospitalization (e.g. from 14 days prior to, and 30 days following the positive test). This can be defined by each site, depending on study design.
   * **Controls**: every other participant in each cohort that is not a case.
3. Severe Covid-19 (A2 in Covid-19 HGI):
   * **Cases**: laboratory confirmed Covid-19 with one or more of the following outcomes:
     1. Death
     2. ECMO requirement
     3. Mechanical ventilation (i.e. intubation) requirement
     4. Non-invasive ventilation requirement (i.e. **new** requirement for BiPAP or CPAP)
     5. High-flow oxygen therapy requirement (e.g. Optiflow)

Again, we suggest that a pre-specified timeframe be used by each study site, depending on study design.

* + **Controls**: every other participant in each cohort that is not a case.

## Analysis scope

The gene burden tests will be performed across all genes of the human genome, except for those located on chromosome Y.

# Study Methods

## Participants inclusion criteria

Individual participant criteria will depend on the enrolling cohort. For this study, there will be no cohort exclusion criteria based on study design, as long as phenotype outcomes can be measured reliably. Given variations in study design, included cohorts may not be able to provide data about each of the outcomes listed in **Section 4.2**. They may still be included in the study if they are able to perform at least one of the analyses. The only requirement will be informed consent (preferably written) from all study participants (or their surrogate decision makers) to be enrolled in their respective cohort, and to have WGS or WES performed.

## Whole genome or Exome sequencing

There will be no restriction on sequencing technology used as long as it has been properly validated, as per local recommendations. Cohort using either WGS or WES may be included in this study.

## Genetic variant calling and quality control

The GRCh38 reference genome will be used for variant calling and for all downstream analyses. Variants will be joint called locally over all study participants genetic sequencing data. There will be no restriction about genome alignment, quality control, and variant calling pipelines, as long as they follow similar steps and recommendations as GATK Best Practices. If needed, each study site is encouraged to follow the steps outline in our WES/WGS analysis pipeline document (sections 1 and 2 of the following: <https://docs.google.com/document/d/1X_qjplH8T4BJXSeMQ_sBfQUTiu_kAisicOqGb6B8hcM/edit>).

## Genetic variant annotation

From the variant-calling results above, we will perform variant annotation using VEP and the following tools, all available using the VEP tool4 (plugins: dbNSFP v4.0 for the first 5 algorithms below):

* SIFT5
* PolyPhen26 with HDIV database
* PolyPhen2 with HVAR database
* LRT7
* MutationTaster8

## Genetic association studies

All analyses will be performed separately for each continental ancestry (AFR, AMR, EAS, EUR, SAS, as well as a general middle eastern ancestry), as per each contributing cohort’s population, then meta-analyzed (see below for meta-analysis plan).

The primary analysis will be aligned with previous work done on the UK Biobank whole-exome sequencing from Regeneron’s genetics group9. That is, we will define a burden test by assigning each gene per individual a score of 0, 1, or 2 in the following way:

* An individual with no deleterious variant at the gene will receive a score of 0 for that gene.
* An individual with any number of deleterious variants at the gene, if they are all heterozygous, will receive a score of 1 for that gene.
* An individual with any amount of homozygous deleterious variants at the gene will receive a score of 2 for that gene.

This score is like an additive effect model for single nucleotide polymorphisms in traditional genome-wide association studies. This is akin to an *in-silico* testing of gene knockdown experiments, where one deleterious variant is assumed to be enough to lead to a loss-of-function, and compounding such variants is hypothesized not to alter function more. To avoid confusion with other types of burden tests, we will refer to this test as the knockdown burden test below.

Using variants that passed quality control described in **Section 5.3**, we will perform two analyses. First using variants with MAF<1% and second using variants with MAF<0.1%. Each set of variants will be further restricted to predicted-deleterious variants according to VEP annotations or five *in-silico* algorithms to define 3 “masks” (see below) to use for our knockdown tests. These five algorithms are the following:

* SIFT
* PolyPhen2 with HDIV database
* PolyPhen2 with HVAR database
* LRT
* MutationTaster

These 5 algorithms are available through the dbSNP4.0a VEP plugin.

All association tests will include the following covariates:

* Age
* Sex
* Age \* Sex
* Age squared
* Age squared \* sex
* 10 PCs obtained from common variants (e.g. from whole-genome genotyping, as in GWASs), using a MAF lower bound of at least 1%.
* 20 PCs obtained from rare variants10, using LD pruned variants with minor allele frequency less than 1% (allele frequency based on each individual cohort’s sample). Lower threshold on MAC to be left at the discretion of each participating cohort.

The following three masks will be used (in all analyses), on each gene’s *canonical transcript*:

* **M1**: using pLOF variants defined by VEP as high impact variants.
* **M3**: using pLOF variants defined in M1 **OR** moderate impacts indels as defined by VEP **OR** moderate impacts SNPs as defined by VEP which are classified as deleterious by all *in-silico* algorithms providing an annotation.
* **M4**: using pLOF variants defined in M1 **OR** moderate impacts indels as defined by VEP **OR** moderate impacts SNPs as defined by VEP which are classified as deleterious by at least 1 *in-silico* algorithm.

The association studies will therefore be performed locally in the following ways:

* Knockdown burden tests at MAF<1%
  + All cohorts will restrict analysis to variants that have a MAF<1% in their sample.
  + To ensure that all cohorts use a common definition of what constitutes a MAF below 1%, we will provide an additional list of excluded variants, made from merging common variants from all cohorts. This will ensure that common variants in one cohort are not used as rare variants by another group. We will proceed with two analyses:
    1. Variants present in any of the gnomAD11 or ESP12 reference ancestries with MAF>1% will be excluded in all studies in the first analysis
    2. An additional variant list made up of each cohort’s common variant (MAF>1%) will serve as a second exclusion list for the second analysis (in addition to gnomAD and ESP).

The different masks above will be used with the remaining variants.

The exclusion list will be centrally compiled by the BQC-19 investigators and built from the list of all variants from each participating cohort. Each cohort will have previously normalized and left aligned their variants to make sure we are comparing the same variants. Code for normalization and alignment using bcftools is provided on the github: <https://github.com/DrGBL/WES.WGS>.

* + P-value threshold for knockdown burden tests will be obtained using Bonferroni adjustments by the number of genes included, with a significant p-value of 0.05, divided by the number of genes tested (on an exome-wide scale this corresponds to around 20,000 genes and a threshold of 0.05/20,000 = 0.000025).

1. Knockdown burden tests at MAF<0.1%
   * The same process as above will be followed using a MAF<0.1% for the choice of deleterious variants to be included with the following changes:
     1. All cohorts will restrict analysis to variants with MAF<0.1% **AND** to singletons (to allow for smaller cohorts to contribute to the analysis).
     2. The gnomAD and ESP list will also be used to remove all variants with MAF>0.1% in any population.
     3. In an additional analysis, a pooled list of common variants will also be used for additional removal of locally common variants that are not annotated as MAF<0.1% by gnomAD or ESP.

To account for imprecise MAF estimates in smaller cohorts and imbalance in cases and controls in some cohorts, while still excluding most common variants, each cohort will report variants with both MAF≥0.1% and MAC≥6. These will be pooled across cohorts for the exclusion list. The MAC≥6 was made in order for a cohort with a sample size of 300 to report variants with MAF≥1%.

See **Appendix 2** for algorithms on how the different analysis and how to select variants based on these multiple MAF thresholds. Also refer to **Appendix 3** for an algorithm on how to select variants to provide to the consortium to build the pooled common variant exclusion list in the MAF<0.1% analysis.

All analyses will be performed using the Regenie software14. Regenie uses a two-step approach to performing genome-wide association studies:

1. Step 1: uses ridge regression to produce a genome-wide prediction for the outcomes. Common variants (i.e. like a GWAS) should be used for this step. Step 1 only needs to be done 1, and multiple phenotypes can be combined at once.
2. Step 2: using results from Step 1, we will analyze each gene and mask combinations as though they were a single nucleotide polymorphism, using the knock-down burden test described above. This step uses Firth penalized likelihood to adjust for rare events.

## Individual variants association tests

If we find significant associations using gene-based tests, we will proceed with single association studies using all variants from those genes. Power curves based on the hospitalized phenotypes, with a significance threshold of 5x10-8, and using the November 17 number of cases reported in the consortium were performed using the Genetics Association Study Power Calculator13. They are shown in **Appendix 1**. These suggest that we should be well powered for relative risks > 2. P-value threshold for these tests will be obtained using Bonferroni adjustments by the number of variants tested (note that this is a likely conservative estimate, given linkage disequilibrium).

## Meta-analysis

Summary statistics from the knockdown burden will be used in an inverse-variance weighted random effect model, to account for the expected ancestry heterogeneity. This will be done separately for every mask in the knockout tests. The same procedure will be used if single variant association tests are done.

For this, we ask that the summary statistics file obtained from Regenie’s Step 2 (with the --htp command) be named with the following convention:

* CohortName.YYYY.MM.DD.Phenotype.Ancestry.WhichExclusionList.txt

Additionally, we ask that one additional file named “CohortName.YYYY.MM.DD.info” containing a table with the following columns be also submitted (as a tab separated, csv separated, or excel sheet). The use of the --htp option allows for a standardized output that includes all necessary information for the meta-analysis. The only information not provided by the --htp output, and hence required in the file name, are the name of the cohort/biobank, the date the analysis was performed, the phenotype (A, B, C), the ancestry, and the variant exclusion list used.

For the exclusion list, write either “gnomad” (for the analysis using gnomad/esp) or “commonList” (for the analysis using gnomad/esp and the common variant merged list). For example, for a cohort named “HelloWorldBioBank”, with an analysis performed on December 17 2020, with south asian ancestry, phenotype A, the file names would be:

* HelloWorldBioBank.2020.12.17.A.SAS.gnomad.txt
* HelloWorldBioBank.2020.12.17.A.SAS.commonList.txt

If other arrangements need to be made, based on local needs, please discuss with Guillaume Butler-Laporte.

The meta-analysis will be performed in two steps:

1. For each separate ancestry, summary statistics will be meta-analyzed using inverse-variance weighted fixed effect meta-analysis.
2. Using results from step 1, we will perform random effect meta-analysis using Dermisonian-Laird estimators.

The sample sizes and carrier counts for cases and controls will be summed for each gene.

# References

1. Williams FMK, Freydin M, Mangino M, et al. Self-reported symptoms of covid-19 including symptoms most predictive of SARS-CoV-2 infection, are heritable. *medRxiv*. Published online January 1, 2020:2020.04.22.20072124. doi:10.1101/2020.04.22.20072124

2. Wang X. Firth logistic regression for rare variant association tests. *Front Genet*. 2014;5:187. doi:10.3389/fgene.2014.00187

3. Covid-19 Host Genetic Initiative Phenotype definitions for analyses. Published 2020. Accessed September 9, 2020. https://docs.google.com/document/d/1okamrqYmJfa35ClLvCt\_vEe4PkvrTwggHq7T3jbeyCI/edit

4. McLaren W, Gil L, Hunt SE, et al. The Ensembl Variant Effect Predictor. *Genome Biol*. 2016;17(1):122. doi:10.1186/s13059-016-0974-4

5. Vaser R, Adusumalli S, Leng SN, Sikic M, Ng PC. SIFT missense predictions for genomes. *Nat Protoc*. 2016;11(1):1-9. doi:10.1038/nprot.2015.123

6. Adzhubei I, Jordan DM, Sunyaev SR. Predicting functional effect of human missense mutations using PolyPhen-2. *Curr Protoc Hum Genet*. 2013;Chapter 7:Unit7.20. doi:10.1002/0471142905.hg0720s76

7. Chun S, Fay JC. Identification of deleterious mutations within three human genomes. *Genome Res*. 2009;19(9):1553-1561. doi:10.1101/gr.092619.109

8. Schwarz JM, Rödelsperger C, Schuelke M, Seelow D. MutationTaster evaluates disease-causing potential of sequence alterations. *Nat Methods*. 2010;7(8):575-576. doi:10.1038/nmeth0810-575

9. Kosmicki JA, Horowitz JE, Banerjee N, et al. Genetic association analysis of SARS-CoV-2 infection in 455,838 UK Biobank participants. *medRxiv*. Published online January 1, 2020:2020.10.28.20221804. doi:10.1101/2020.10.28.20221804

10. Mathieson I, McVean G. Differential confounding of rare and common variants in spatially structured populations. *Nat Genet*. 2012;44(3):243-246. doi:10.1038/ng.1074

11. Karczewski KJ, Francioli LC, Tiao G, et al. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature*. 2020;581(7809):434-443. doi:10.1038/s41586-020-2308-7

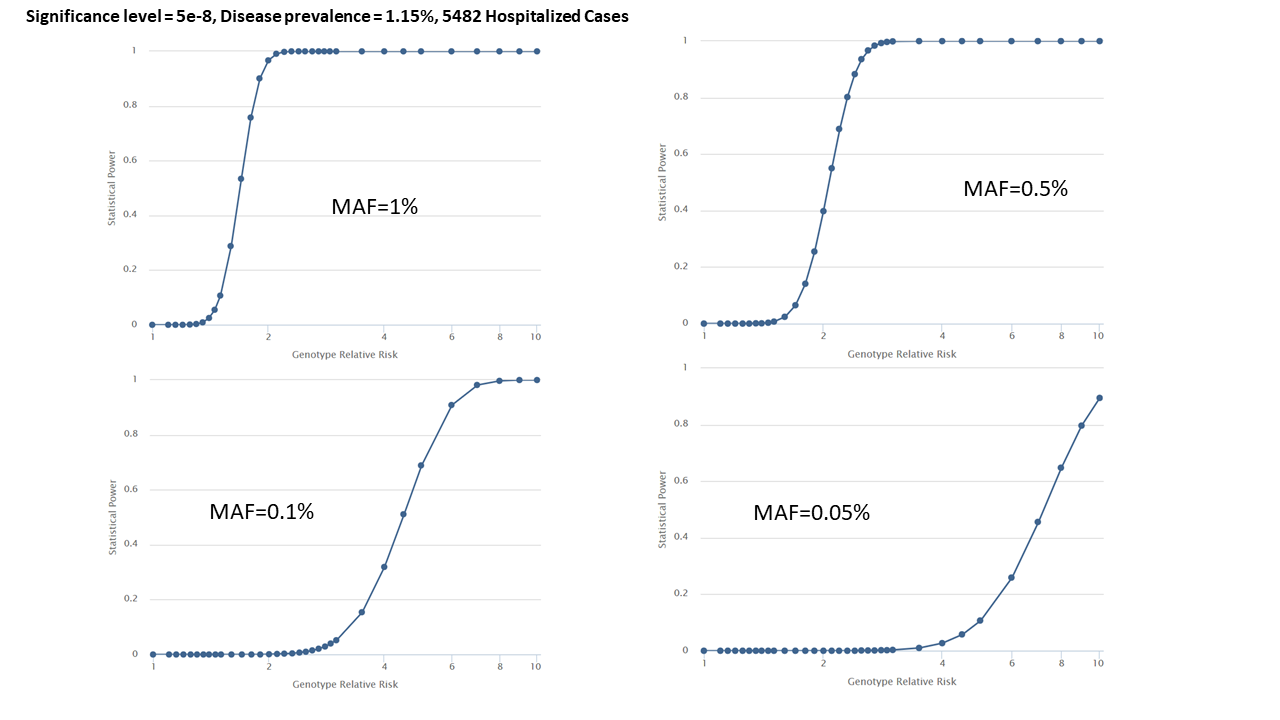
12. Exome Variant Server, NHLBI GO Exome Sequencing Project (ESP), Seattle W. Exome Variant Server.

13. Johnson JL, Abecasis GR. GAS Power Calculator: web-based power calculator for genetic association studies. *bioRxiv*. Published online January 1, 2017:164343. doi:10.1101/164343

14. Mbatchou J, Barnard L, Backman J, et al. Computationally efficient whole genome regression for quantitative and binary traits. *bioRxiv*. Published online January 1, 2020:2020.06.19.162354. doi:10.1101/2020.06.19.162354

# Appendix 1

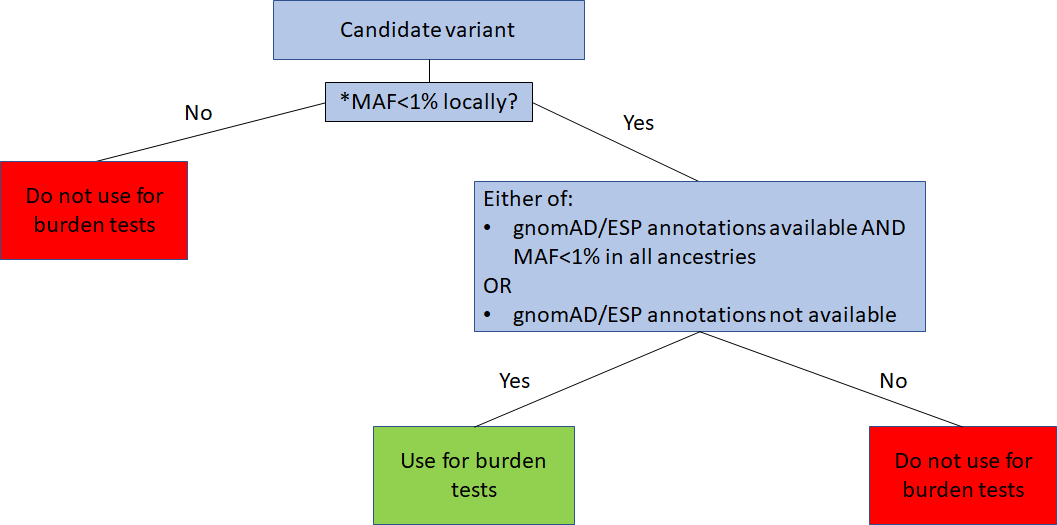
Power calculations using the Genetics Association Study Power Calculation Tool: <http://csg.sph.umich.edu/abecasis/gas_power_calculator/>



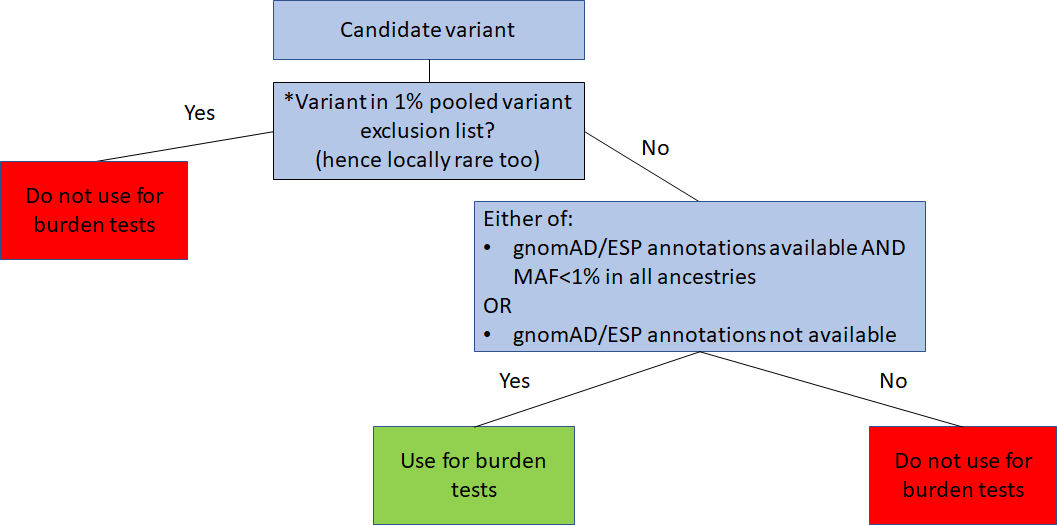
# Appendix 2

Algorithms for variant inclusions based on their minor allele frequencies.

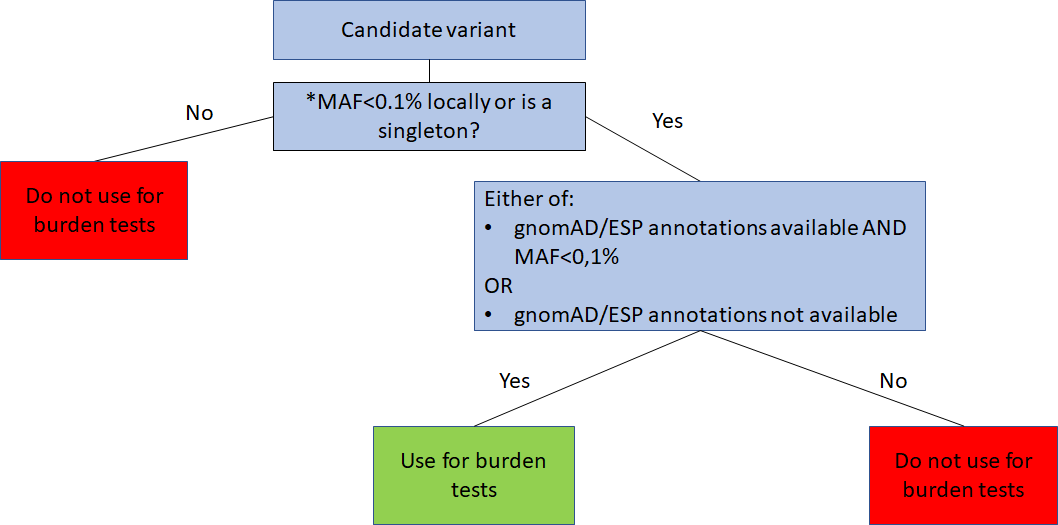
1. MAF<1% analysis without the pooled common variant exclusion list.



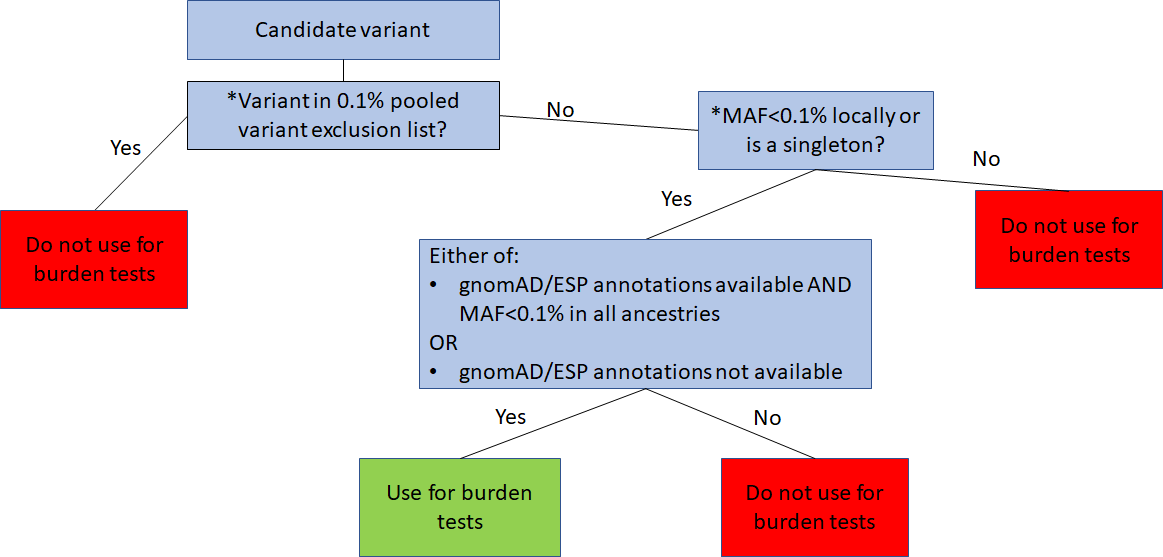
1. MAF<1% analysis with the pooled common variant exclusion list.



1. MAF<0.1% analysis without the pooled common variant exclusion list.



1. MAF<1% analysis with the pooled common variant exclusion list.



# Appendix 3

Algorithm on choosing locally common variants for the MAF<0.1% analysis pooled common variant exclusion list.

