Manual for Conducting a Large-scale GWAS and Whole Genome Sequencing (WGS) of COVID-19 infection (COVNET)

Version 2.5 (August 20, 2020)

NIH Lead Investigator: Stephen Chanock, M.D. (chanocks@mail.nih.gov)
Director, Division of Cancer Epidemiology and Genetics, NCI

NIH Investigators:
Sharon Savage, M.D. (savagesh@mail.nih.gov)
Lisa Mirabello, Ph.D. (mirabellol@mail.nih.gov)
Meredith Yeager, Ph.D. (yeagerm@mail.nih.gov)
Mitchell Machiela, Sc.D. (mitchell.machiela@nih.gov)
Joshua Sampson, Ph.D. (joshua.sampson@nih.gov)
Amy Hutchinson, M.S. (hutchiam@mail.nih.gov)
Belynda Hicks, M.S. (hicksbel@mail.nih.gov)
Division of Cancer Epidemiology and Genetics, NCI

Mary Carrington, Ph.D. (carringm@mail.nih.gov)
Center for Cancer Research, NCI

Leslie Biesecker, M.D. (lesb@mail.nih.gov)
Teri Manolio, M.D., Ph.D. (manoliot@mail.nih.gov)
National Human Genome Research Institute

Steven Holland, M.D. (sholland@niaid.nih.gov)
National Institute of Allergy and Infectious Diseases

Project Manager

Vibha Vij, M.S., MPH (vibha.vij@nih.gov)
Division of Cancer Epidemiology and Genetics, NCI

Website: https://dceg.cancer.gov/research/how-we-study/genomic-studies/covnet

Table of Contents

Sections

Appendices

1.0 Background and Rationale for a Large Crowdsourced GWAS of COVID19 Infection

The infectious disease syndrome known as Coronavirus disease 2019 (COVID-19) is caused by the etiologic agent, severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), which was first described in late 2019 in Wuhan, China. Since then, it has spread widely into the 2019 - 20 coronavirus global pandemic. As of June 30, 2020, there have been over 10 million confirmed COVID-19 cases worldwide and over 500,000 deaths; of these, over 2.6 million cases have been reported in the United States (https://coronavirus.jhu.edu/map.html). Disease expression is highly variable including rapidly fatal infection as well as the more frequently observed mild or asymptomatic infection [1]. Commonly reported symptoms include fever, cough, shortness of breath and loss of taste and smell. Other symptoms such as muscle pain, sputum production, diarrhea, conjunctivitis, and sore throat are seen less frequently. In contrast, a minority of infected individuals progress to severe pneumonia and multi-organ failure, and mortality is mainly driven by a subset of patients that develop severe respiratory failure. The early reports of the rate of deaths per number of diagnosed cases have ranged from 0.2% to 15% according to age group and co-existence of select comorbidities. It is likely that the case fatality ratio will decline as more mildly affected cases are identified through expanded testing. Irrespective of the actual rate of severe disease and death, it is clear that the burden and fear associated with the large number of severe cases necessitates rapid investigation to establish scientific bases for diagnostic, therapeutic and preventive strategies.

The rapid proliferation of infection is directly related to the high transmissibility of the virus, for which virtually no one has natural immunity. SARS-CoV-2 is spread from person to person by respiratory droplets. It can also be spread much less efficiently via hand to mouth (mucous membrane) transmission. The typical time from exposure to symptom onset is typically two to 12 days, with an average of five days. The standard diagnostic method is by reverse transcription polymerase chain reaction (rRT-PCR) from a nasopharyngeal swab. The infection can also be diagnosed clinically from a combination of symptoms, risk factors and a chest CT scan compatible with multifocal ground glass infiltrates or pneumonia.

There is clearly a dramatic range of disease severity associated with this infection. Children appear to rarely experience severe symptoms [2] but recently, it has been noted that SARS-CoV-2 infected children can rarely develop symptoms strongly resembling Kawasaki's Disease . A review of 72,314 cases by the Chinese Center for Disease Control and Prevention showed that <1% of the cases were in children younger than 10 years of age but it appears that many children are asymptomatic and thus not tested [1]. Young adults are uncommonly symptomatic. Overall, the median age of Chinese patients was 47 years and 41.9% were female [3]. 5.0% were admitted to the ICU, 2.3% underwent invasive mechanical ventilation, and 1.4% died. Early on the expression of the disease was described primarily in China and now with many reports from the US and Europe, the importance of age (over 65), male gender and underlying conditions (diabetes, pulmonary and cardiac disease and less so for cancer) contribute to an enhanced risk for severity and higher mortality.

Table 1. Disease severity by age in China.

Age	N (%)	Disease Severity		ICU, ventilator, or death	
		Non-severe N (%)	Severe N (%)	Yes	No
0-14yr	9/1011 (0.9)	8/848 (0.9)	1/163 (0.6)	**0**	9/946 (1.0)
15-49 yr	557/1011 (55.1)	490/848 (57.8)	67/163 (41.1)	**12/65 (18.5)**	545/946 (57.6)
50-64 yr	292/1011 (28.9)	241/848 (28.4))	51/163 (31.3)	**21/65 (32.3)**	271/946 (28.6)
>65 yr	153/1011 (15.1)	109/848 (12.9)	44/163 (27.0)	**32/65 (49.2)**	121/946 (12.8)

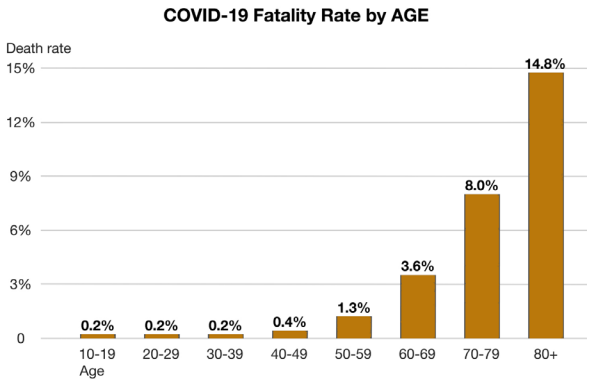
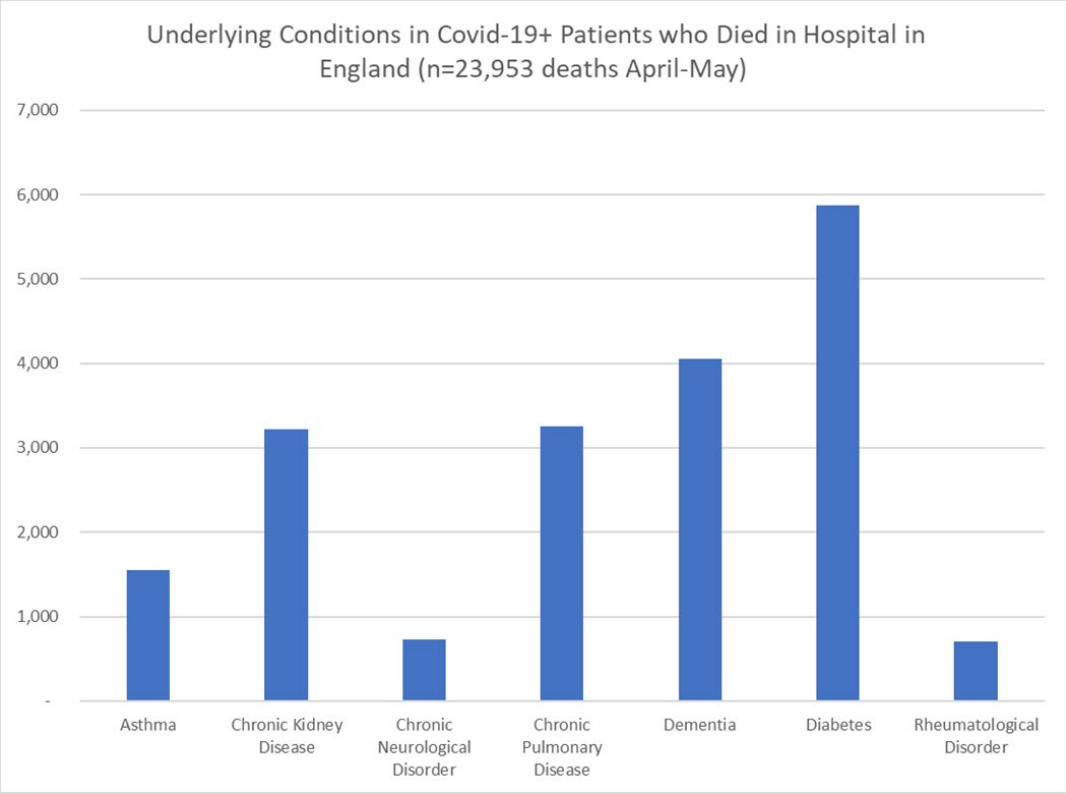


Figure 1. COVID-19 fatality rate by age in China, data from: http://weekly.chinacdc.cn/en/article/id/e53946e2-c6c4-41e9-9a9b-fea8db1a8f51

From Table 1 above, it is clear that there is a major effect of age with nearly half of those hospitalized over 65 years of age requiring ICU care, ventilatory support, or dying. Another presentation of data that includes non-hospitalized patients demonstrates the age skewing pattern that is now worldwide and first described by the Chinese CDC (Figure 1)[4]. The most common complicating issue in the Chinese patients was acute respiratory distress syndrome, with septic shock the next most common. The pathogenesis of severe disease is poorly understood, and despite the strong age skewing, the mechanism of this increased susceptibility in older ages is unknown.

Recent evaluation of data from the UK National Statistics of data drawn from the NHS website in late May, 2020 reveals that select underlying conditions are associated with increased risk for death.



Source: https://www.england.nhs.uk/statistics/statistical-work-areas/covid-19-daily-deaths/

A study of COVID-19 viral genome sequences determined that viral genetic variation did not significantly affect patient outcomes, and suggested instead that disease severity correlated with host factors [5]. Recent studies of adult patients consistently suggest mortality associated with older age, male

gender, and underlying cardiovascular disease traits [6-9]. Other risk factors suggested to be related to COVID-19 severity include obesity, diabetes, and smoking [6-9]. Black, Asian and minority ethnic individuals have also been reported to have an increased risk of infection and worse clinical outcomes [10]. It is clear that overall a minority of individuals develop severe disease and we hypothesize that this susceptibility is, at least in part, due to genetic factors.

A recent case-control genome-wide association study (GWAS) of 1,610 severe COVID-19 patients from Spain and Italy identified a locus at chromosome 3p21.31 associated with COVID-19 respiratory failure at genome-wide significance. The chromosome 3 signal is near a notable candidate gene, *SLC6A20*, that encodes an interaction partner with *ACE2*, the SARS-CoV-2 cell surface receptor. The authors also suggested a locus at 9q34, which harbors the ABO blood group, and the data indicates type A+ could also be a risk factor but more work is needed to confirm this. These data suggest an underlying genetic component to susceptibility to severe disease, which already could lead to new mechanistic insights into the pathophysiology of COVID-19 infection and the reasons for major interindividual differences.

The goal of our project is to conduct large genome-wide studies to identify common and rare germline genetic variants associated with host susceptibility to severe or fatal COVID-19 disease using a case-case design. We are taking an innovative and unconventional approach to recruit using an academic 'crowdsource approach' that will accrue from multiple centers and studies throughout the U.S. and Canada.

Our target GWAS is to genotype as many as 40,000 COVID-19 new positive cases, defined on the basis of positive viral testing or strong epidemiologic data linked to serologic confirmation of antibodies to SARS-CoV-2. The primary aim is to investigate common and uncommon variants associated with differences in disease outcome. While the pandemic is rapidly spreading, it is daunting to identify suitable controls for identification of actual alleles associated with acquisition or prevention of SARS-CoV2 infection due to concerns about misclassification or lack of exposure to the pandemic.

The urgency of this pandemic necessitates finding new ways of organizing multi-institutional collaborations and sharing data in the shortest possible time. We are recruiting participants through direct contact with experts across the country and seek biospecimens for analysis (including DNA, blood or oral swab/sputum) linked to collected data of established covariates, risk factors and clinical outcomes. The recruitment of studies is conducted through several pathways:

1. NIH COVIDcode protocol that will include consented under an NIH IRB approved protocol from NIH patients and employees as well as from local participating hospitals. Currently, the protocol requires a positive viral COVID19 test result plus clinical and covariate data. Please see Appendix 5 for full COVIDcode protocol, recruitment materials and consent forms.
2. Participation of collaborating hospitals and research institutions throughout the U.S and Canada who will locally recruit participants through an IRB-approved protocol and share coded/de-identified samples and clinical data (with no personal identifiers).

All cases will be genotyped on the Illumina GSA v2 array; a subset of approximately 4,000-5,000 (at 30-40X using Illumina short-read technology) will be selected for whole genome sequencing based on communal assessment of possible informative cases. We have adopted an aggressive approach to data sharing consistent with the NIH Data Sharing Policy. After quality control metrics for the GWAS and /or genome sequencing has been completed, the primary genetic data will be returned to the collaborating investigators (e.g., genotype calls or genome sequence data) as part of the signed MTA in a timely manner. The genotyped GWAS data will be imputed on the Michigan server to increase the number of SNPs to the range of 66 Million down to a MAF of 0.015. The coded data will be made available to bonafide researchers and shared with the international COVID-19 Host Genetics

Initiative (https://www.covid19hg.org/), to facilitate rapid analyses. This multi-institutional collaborative approach will enable a sufficiently powered study for a thorough exploration of both common and rare germline genetic variants that may play a role in host susceptibility to severe disease. Detailed analyses will be conducted internally as described below and shared through the COVID19 hg.org network.

2.0 Biospecimens: genomic DNA, blood/blood products, buccal cells/saliva

All processing of received samples and genetic work at NIH will be fully supported by NIH funds as will costs of shipping to the Cancer Genomics Research Laboratory (CGR), DCEG/NCI. Shipping will be arranged on a per study basis and fully covered by the NIH.

2.1 Pre-Extracted DNA Requirements

Pre-extracted DNA specimens should be quantified at the processing lab. Qualification for downstream genotyping will be based on sample concentration and total available dsDNA mass. Given variation in DNA quantification methods, it is important to provide as much mass as possible so that samples qualify for all downstream processing without the need for time-consuming and costly replenishment efforts. Use of a quantification method specifically designed to measure double-stranded DNA (dsDNA) is preferred (e.g., PicoGreen or Qubit). If optical density (OD) readings are used to quantify DNA, it is important to send higher amounts of DNA to account for differences between OD and PicoGreen, which can be greater than 50%.

- Total Mass = 0.5 – 1.0ug [if measured by dsDNA quantification method, such as PicoGreen]
- Total Mass = 1.0 – 1.5ug [if measured by OD, such as NanoDrop]
- Minimum volume = 30ul

NOTE: Due to automation requirements, samples received with less than 30ul will be normalized (diluted) with 1X TE upon receipt.

DNA Sample Format

There are 3 sample format options for sending DNA to the processing lab. Acceptable formats are listed in order of preference by the processing lab, but please select the one that is best suited to your laboratory.

1. SAMPLE KITS: Sample Kits are boxed sets of pre-barcoded tubes, which can be shipped to the site directly by the processing lab at no cost to the user. The user simply adds DNA to the provided tubes, applies the provided caps, and freezes the specimens for shipment. A manifest template will be provided for the user to link relevant specimen data to the tube barcode. ***The use of Sample Kits is strongly encouraged***.

Two types of pre-barcoded Sample Kits are available (type a. is preferred):

a. *96-well SBS format tube racks:* Tube racks provided are FluidX 96-Format, 0.5ml External Thread, Next-Gen Jacket, Tri-Coded Tube (manufactured by Brooks – part # 68-0703-12).

b. *Individual cryovials:* 9x9 sample boxes will be provided containing 1.5mL skirted Sarstedt cryovials, with screw cap (manufactured by Sarstedt – part # 72.703 (tube) + # 65.716 (cap).

NOTE: Sample Kits are available for DNA specimens only. If you are sending biospecimens that need extraction (blood, saliva, etc.), please see Section II.

If you are interested in receiving a Sample Kit for your DNA samples, please contact **Amy Hutchinson** (240-760-6496; amy.hutchinson@nih.gov).

2. PLATES: Standard 96-well plates (PCR plates or similar) are also acceptable. Plates can be full or half-skirted. Plates should be clearly labeled and linked appropriately in the provided manifest template. Plates should be well sealed (either manual or heat-sealed) with a removeable seal. Be

diligent to ensure the well tops are completely sealed to avoid sample loss. Plates should be frozen prior to packaging for shipment.

NOTE: If providing standard PCR-type 96-well plates, please leave well H12 empty for internal controls. Leave the well completely empty, do not add water.

3. TUBES: Samples may be shipped in screw-cap tubes. Once DNA is added, screw caps should be well tightened and boxed in 9x9 or 10x10 boxes with grids and frozen for shipment. If the shipment will contain mixed tube types, please batch the samples by tube type (tube type will be a required value on the provided manifest template if the site is using their own vials rather than a Sample Kit). Specimen vials should be labeled with a linear (1D) barcode corresponding the ID provided on the manifest.

If you are unable to use a Sample Kit or 96-well plates, the processing lab recommends the following vial types:

- NUNC 1.8mL/1.0mL
- Nalgene 2.0mL
- Sarstedt 0.5mL/1.5mL/2.0mL

Manufacturer and part number of tubes utilized must be provided. If that information cannot be determined, additional empty vials must be provided for automation optimization.

NOTE: Flip cap Eppendorf-style microcentrifuge tubes will not be accepted.

2.2 Biological Samples Needing DNA Extraction

Materials needing extraction can also be sent with prior consent. Depending on the number of specimens and the source material type, an appropriate validated protocol will be selected. Current automation systems in place for DNA extraction at the processing lab include those from Qiagen (QIAsymphony) and ThermoFisher (KingFisher).

Blood samples

A variety of blood-derived specimens are acceptable source material for genomic DNA extraction. Material collected from standard phlebotomy vacutainers with anti-coagulant (EDTA/K2, ACD, heparin) are recommended. Acceptable blood specimen types include (in order of preference):

- Whole blood
- Buffy coat
- Buffy coat + RBC (buffy coat layer plus red blood cell layer from standard blood separation)

NOTE: Clotted blood material, serum, or plasma are not possible at this time.

A minimum of 150uL of blood material is required for extraction. Depending on the extraction method selected and the volume provided, an aliquot of the source material may be taken for extraction. Any residual source material can be returned to the sending site at the end of the project upon request (see Section VI. for details).

Buccal Cells/Saliva

Buccal cells and saliva, while the least preferred, are another acceptable source for genomic DNA extraction. Acceptable specimen types include (in order of preference):

- Oragene™ kits (recommended kit: DNA Genotek Oragene•DISCOVER catalog #ORG-500 or #ORG-600)
- Heat Inactivated** Mouthwash buccal collection (recommended mouthwash is Crest/Scope Classic (standard green); please indicate if specimens are pelleted and un-pelleted)
- Heat Inactivated** Oral Rinse buccal collection (non-mouthwash)
- Heat Inactivated** Saliva (no additive or diluent)

**NOTE: Mouthwash, Oral Rinse, and Saliva specimens must be heat inactivated at 56°C for 1 hour prior to shipment.

A minimum of 1000uL (1mL) of saliva/buccal source material is required for extraction. Depending on the extraction method selected and the volume provided, an aliquot of the source material may be taken for extraction. Any residual source material can be returned to the sending site at the end of the project upon request (see Section VI. for details). Details of extraction instruments and protocols utilized for processing specimens will be provided to the sample provider for record keeping.

If you have other material types that you would like considered for extraction, please contact **Amy Hutchinson** (240-760-6496; amy.hutchinson@nih.gov).

Acceptable Sample Formats – Material for Extraction

Source materials for extraction should be provided in screw-cap vials. The processing lab recommends the following vial types:

- NUNC 1.8mL/1.0mL
- Nalgene 2.0mL
- Sarstedt 0.5mL/1.5mL/2.0mL

Manufacturer and part number of tubes utilized must be provided. If that information cannot be determined, additional empty vials must be provided for automation optimization.

If Oragene kits will be sent, the entire collection vial should be shipped (no aliquoting required). If the shipment will contain mixed tube types, please batch the samples by tube type (tube type will be a required value on the provided manifest template if the site is using their own vials rather than a Sample Kit). Specimen vials should be labeled with a linear (1D) barcode corresponding the ID provided on the manifest.

2.3 Sample Preparation for Shipment

Packaging Samples

Care should be taken when packing specimens for transit to avoid sample loss or contamination.

1. SAMPLE KITS: Provided caps should be seated correctly and screwed tightly. Racks should be closed and locked using the side latches. Racks should be placed in provided zip-top bags and frozen prior to packaging. Bags can be placed directly into the container containing dry ice.

2. PLATES: Once 96-well plates have been sealed with a removeable seal (either heat or manual), plates should be frozen prior to packaging. *To avoid piercing the applied seals, plates should not be stacked directly on top of each other for shipment.*

3. TUBES: Boxes (either 9x9 or 10x10) containing vials should be frozen prior to packaging. Boxes may be place on dry ice for shipment.

Shipping containers should be sturdy cardboard with Styrofoam inserts. Many lab suppliers sell shipping cartons that are appropriate for shipping samples. Containers should be sufficiently large to hold all specimens and an appropriate amount of dry ice for the estimated shipping time.

Sample Manifests

Electronic manifest templates (MS Excel) will be provided to each site and will be tailored to the type of specimen being provided. Manifests must be populated and sent via email prior to shipping samples to the processing lab. Please note any fields indicated as mandatory, these must be populated in order to receive shipment clearance.

Prior to shipment, please provide the completed manifest via email to NCICGFDESLReceiving@mail.nih.gov to obtain shipping clearance.

Shipping Samples

Domestic U.S. shipments should be handled via overnight courier (e.g. FedEx) Monday – Wednesday.

International shipments should be handled by World Courier, QuickStat, or other cold-chain logistic expert. These companies will replenish dry ice in transit, should delays occur. It is advised that you DO NOT use FedEx for international shipments as they are not able to replenish dry ice in transit. Please contact the processing lab if you would like assistance setting up an international shipment.

Steps for shipping:

- Email the receiving group list (NCICGFDESLReceiving@mail.nih.gov) at the NCI processing lab and attach the populated sample manifest (either Sample Kit manifest or other manifest that was provided) prior to shipping specimens to request clearance for shipment. Include planned shipment date and estimated delivery date in the email request at least 48 hours prior to shipment.

- The processing lab will reply with clearance for shipment and will provide the shipping address.

- Via email (NCICGFDESLReceiving@mail.nih.gov) provide the carrier, tracking number, and estimated delivery date.

- The processing lab will confirm acceptance of the shipment via email upon receipt.

Sample Return

The processing lab can return excess DNA to sending site by request. A return mass/volume threshold for DNA return will be established prior to beginning the project. Only samples meeting the minimum will be returned. Any residual source material (blood, saliva) remaining following extraction will be returned to the sending site, regardless of volume remaining.

3.0 Patient Phenotypes

We will use an established shared set of phenotype definitions that overlap with the essential elements outlined by the COVID-19 Host Genetics Initiative (https://www.covid19hg.org/) for ease of combining data among groups. SARS-CoV-2 infection should be defined as:

1. Laboratory confirmation of SARS-CoV-2 infection (RNA and/or PCR-based),

   OR

2. ICD/administrative/EHR coding-based definition of SARS-CoV-2 infection where large-scale & rapid clinical testing unavailable (see Appendix 1 for case codes),

   OR

3. Serologic evidence of SARS-CoV-2 infection together with sufficient clinical documentation of symptoms consistent with mild, moderate or severe SARS-CoV-2 infection prior to the serology (and after March 1, 2020).

There will be a preferred set of phenotypes for the primary association analyses and basic covariates for adjustment. There will also be more extensive phenotypes collected when available, but we anticipate these will not be available from all participating groups. Key covariates are sought and are briefly listed in the attached patient questionnaire (see Appendix 2 for questionnaire).

Basic phenotypes and covariates of interest include age at diagnosis, sex, height, weight, ancestry, symptoms (i.e., cough, fever, difficulty breathing, fatigue, muscle pain), days reported a cough, days reported a fever, respiratory rate at admission, pneumonia status, septic shock, organ failure, highest level of respiratory support, if ventilator was used (and, for how many days), and information on basic self-reported comorbidities and risk factors (i.e., smoking, use of medications). We particularly seek sets of cases with known conditions that confer risk—cardiac, pulmonary, diabetic, oncologic, immunologic, or other conditions—as well as cases with no known risk factors. Please let us know if only a small subset of the requisite data is available to determine suitability of phenotype for a subset of the planned analyses.

4.0 Genotyping and Imputation

We will use Illumina's Global Screening Array (GSA) version 2.0 plus consortia defined multi-disease content (GSAv2-MD; Illumina Inc., San Diego, USA) that contains 712,191 variants before quality control (QC) for genotyping. This array has cross-population and population-specific content designed specifically for diverse and admixed American populations, is enriched for low-frequency variants (1-5%), exonic variants and clinical content, and has high imputation accuracy for all populations. Illumina's GenomeStudio v. 2.0 software, LIMS and open source library will be used to generate genotype cluster files and convert the raw intensity datafiles to genotype calls.

Once GSA genotyping is complete, all samples will be run through an established automated QC pipeline and a standardized report will be generated to assess various QC metrics related to the array, including call rate, concordance rate and contamination. Samples that are determined to be contaminated (contamination rate >20% with VerifyIDintensity, https://github.com/gjun/verifyIDintensity), ancestry-specific outliers for heterozygosity (FHET within +/- 0.20), low call rate (>2% missing), and sex-discordant samples will be excluded. Variants that deviate from Hardy-Weinberg equilibrium ($P < 10e^{-6}$), have a minor allele frequency (MAF) <0.01, genotyping rate < 95%, and SNPs with differential

missingness in cases vs. controls (p-value cutoff 0.001) will be excluded. We will evaluate duplicates and relatedness (IBD estimation) using a linkage disequilibrium (LD)-pruned set of autosomal SNPs with a MAF >5%, and one individual will be removed from each pair of duplicates. We will determine the number of relatives and degree of relatedness, but relatives will remain in the association analyses (see below).

To maximize genetic coverage, we will perform single nucleotide polymorphism (SNP) imputation on genome build GRCh38 using the Michigan Imputation Server and 194,512 haplotypes generated by the Trans-Omics for Precision Medicine (TOPMed) program (Freeze 5). After QC and exclusions, a common set of well-genotyped SNPs, filtered of SNPs with alleles AT or CG (the latter often leading to strand issues during imputation), will be uploaded to the Michigan Server for whole-genome imputation. Poorly imputed SNPs with info scores <0.3 will be removed from subsequent analyses and SNPs with posterior probability < 0.9 will be converted to missing. Genotyping rate and MAF filters will be then applied.

The genotyped SNPs within chromosome 6 will be used for HLA imputation. We will apply the well characterized HIBAG software [11] on the European ancestry patients for which a model has been developed and validated through high resolution typing in other similar cohorts, and in other ancestries as methods improve.

Whole genome sequencing (WGS) will be performed in a subset of approximately 4,000-5,000 informative cases. Indexed Illumina NGS libraries will be prepared from the germline DNA using the KAPA HyperPlus (KAPA Biosystems, Wilmington, MA) library preparation kit, and sequenced on a Illumina NovaSeq 6000 (Illumina, San Diego, CA, USA) at an average depth of 30-40× using paired-end readlengths of 2x150bp. Sequencing reads will be trimmed using the Trimmomatic program (v0.36), which marks all low-quality stretches (average quality score < Q15 in a 4-bp sliding window) and reports the longest high-quality stretch of each read. Only read pairs with both ends no shorter than 36bp will be included. Reads will then be aligned to the reference human genome assembly GRCh38 using Novoalign software (v4.02.01). Duplicate reads due to either optical or PCR artifacts will be removed using the MarkDuplicates module of the Picard software (v2.18.11). Analyses will use only properly aligned read pairs (i.e., the two ends of each pair must be mapped to the reference genome in complementary directions and must reflect a reasonable fragment length [300+/-100 bp]). A series of quality control steps will be implemented, including evaluating sequencing reads (FastQC), duplications, contamination, individual subject coverage, cohort coverage, sex concordance, and pre-variant-calling. Single nucleotide variants (SNVs) and small insertions and deletions (INDELs) will be detected using the HaplotypeCaller module from Genome Analysis Toolkit (GATK v4.0), Google's deep learning based variant caller DeepVariant (v0.5.2), and Illumina's Strelka2 (v2.9). An in-house developed Ensemble workflow (based on a majority voting strategy) will be applied to filter and integrate variant calling results from these three callers. A post-variant-calling check will also be conducted to assess the variant calling quality. Variant annotation will be performed using an in-house custom pipeline. SNVs/INDELs of interest will be subjected to manual inspection of the raw bam files for further QC checks to rule out artifacts.

5.0 Analysis Plan

Phenotype definitions

Primary association analyses will be focused on binary phenotypes for mild vs. severe disease that have been harmonized among groups, and consistent with the COVID-19 Host Genetics Initiative (https://www.covid19hg.org), to promote data and results sharing. See Appendix 3 for primary association analysis phenotype definitions v1.

Primary association analysis

For European ancestry individuals, generalized linear mixed models (LMM) will be used to determine the associations of individual imputed SNPs and mild vs. severe outcomes, using SAIGE. Association models will be adjusted for principal components (PCs), sex, age, and appropriate covariates (e.g., underlying conditions from questionnaire) that are significantly associated with the disease outcome.

Principal component analyses (PCA) using PC-AiR (https://rdrr.io/bioc/GENESIS/man/pcair.html) on LD-pruned genotype calls will be performed to evaluate and control for population substructure. PC-AiR directly performs PCA and provides robust population structure inference in the presence of related individuals [12].

Global ancestry estimates (e.g., from PCA) are not sufficient to adjust for local ancestry in genetic association analyses. Therefore, for admixed individuals, those with different proportions of European, African, and Native-American ancestry (such as African-Americans and Latinx), we will use tools (e.g., RFMix [13]) that account for local ancestry across different regions of the genome as traditional GWAS methods are not appropriate for these populations.

Stratified analyses will also be run for specific groups of interest, including by sex and age group (without these covariates in the respective models). Quantile-quantile plots will be used to assess any potential inflation in overall test statistics. We will evaluate heterogeneity across other studies and use random or fixed effects meta-analyses to combine results with available data from other genotyped resources for harmonized outcomes using METAL. SNPs that achieve genome-wide significance (p-values less than 5×10^{-8}) will be confirmed by TaqMan and further investigated with fine-mapping strategies to identify highly correlated variants based on LD patterns and current bioinformatic programs that infer biological function.

Secondary analyses

In addition to the primary analyses, in a subset of individuals with detailed phenotypes we will conduct additional association analyses, including for mortality, an ordinal disease phenotype for severity, and susceptibility (See Appendix 3 for extended association analysis phenotype definitions v1). Rare-variant association tests will also be used to compare cases with mild vs. severe outcomes. Additional analyses include detection of large structural copy number alterations and association with COVID-19 severity, gene-based and pathway-based analyses, and potential polygenic risk scores. ***We expect to provide successive versions of our analysis plans as the data set grows defined by the timing or completion and sample sizes of interest.***

5.1 Statistical power

We estimated the power to detect a significant genetic association for different combinations of sample sizes (N), relative risks (RR), and mild:severe case ratios (R) in Tables 2 and 3 for SNPs with a MAF of 0.10 and 0.20, respectively, and in Table 4 for 30,000 cases and a wider range of MAFs, assuming an α of 5×10^{-8}. With a sample size of 2,500 and 2:1 mild:severe case ratio, we have approximately 80% power to detect a significant SNP with a relative risk of 1.7 if the MAF is 0.10, and a relative risk of 1.5 if the MAF is 0.2. With a sample size of 30,000 and 2:1 mild:severe case ratio, we have greater than 80% power to detect a significant SNP with a relative risk of 1.3 if the MAF is 0.03.

Table 2. Power to detect a significant genetic association for different combinations of sample sizes (N), relative risks (RR), and mild:severe case ratios (R), assuming a MAF of 0.10.

	N = 1000			N = 2500			N=5000		
	R=8:1	5:1	2:1	8:1	5:1	2:1	8:1	5:1	2:1
RR=1.3	0	0	0	0	0	0.02	0	0.01	0.31
1.4	0	0	0	0	0	0.12	0.01	0.05	0.76
1.5	0	0	0.01	0	0.01	0.37	0.03	0.19	0.96
1.6	0	0	0.05	0.01	0.03	0.63	0.1	0.39	1
1.7	0	0	0.1	0.01	0.07	0.83	0.18	0.61	1
1.8	0	0	0.14	0.03	0.15	0.93	0.29	0.78	1
1.9	0	0.01	0.22	0.04	0.26	0.97	0.41	0.88	1
2	0	0.01	0.33	0.06	0.31	0.99	0.55	0.95	1
2.1	0	0.02	0.4	0.08	0.41	1	0.67	0.97	1
2.2	0	0.03	0.47	0.14	0.52	1	0.75	0.99	1
2.3	0	0.03	0.59	0.17	0.57	1	0.82	1	1
2.4	0.01	0.05	0.66	0.22	0.68	1	0.87	1	1
2.5	0.01	0.05	0.73	0.23	0.72	1	0.91	1	1

Note: we assume a MAF=0.10, a multiplicative effect, a Cochran-Armitage test of trend, and an $\alpha = 5 \times 10^{-8}$.

Table 3. Power to detect a significant genetic association for different combinations of sample sizes (N), relative risks (RR), and mild:severe case ratios (R), assuming a MAF of 0.20.

	N = 1000			N = 2500			N=5000		
	R=8:1	5:1	2:1	8:1	5:1	2:1	8:1	5:1	2:1
RR=1.3	0	0	0.01	0	0.01	0.22	0.02	0.1	0.88
1.4	0	0	0.04	0.01	0.04	0.68	0.09	0.44	1
1.5	0	0	0.14	0.02	0.14	0.94	0.31	0.78	1
1.6	0	0.01	0.33	0.07	0.34	0.99	0.57	0.95	1
1.7	0	0.03	0.55	0.14	0.56	1	0.79	0.99	1
1.8	0.01	0.05	0.75	0.28	0.75	1	0.92	1	1
1.9	0.01	0.11	0.84	0.39	0.85	1	0.97	1	1
2	0.03	0.16	0.94	0.53	0.94	1	0.99	1	1
2.1	0.05	0.2	0.97	0.65	0.97	1	1	1	1
2.2	0.07	0.25	0.99	0.74	0.99	1	1	1	1
2.3	0.07	0.38	1	0.82	1	1	1	1	1
2.4	0.1	0.5	1	0.9	1	1	1	1	1
2.5	0.15	0.52	1	0.93	1	1	1	1	1

Note: we assume a MAF=0.20, a multiplicative effect, a Cochran-Armitage test of trend, and an $\alpha = 5 \times 10^{-8}$.

Table 4. Power to detect a significant genetic association for 30,000 cases and different combinations of relative risks (RR) and mild:severe case ratios (R), assuming the following MAFs:

	MAF = 0.01			MAF = 0.03			MAF = 0.05			MAF = 0.1			MAF = 0.2		
	R=8:1	5:1	2:1	8:1	5:1	2:1	8:1	5:1	2:1	8:1	5:1	2:1	8:1	5:1	2:1
RR=1.05	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.01
1.1	0	0	0	0	0	0	0	0	0.02	0	0.01	0.21	0.01	0.06	0.74
1.15	0	0	0	0	0	0.05	0	0.01	0.3	0.02	0.12	0.9	0.15	0.57	1
1.2	0	0	0	0	0.01	0.3	0.01	0.07	0.8	0.13	0.5	1	0.61	0.96	1
1.25	0	0	0.02	0.01	0.04	0.68	0.05	0.25	0.98	0.41	0.87	1	0.93	1	1
1.3	0	0	0.07	0.02	0.13	0.92	0.15	0.52	1	0.71	0.99	1	1	1	1
1.4	0	0.01	0.29	0.11	0.49	1	0.5	0.92	1	0.98	1	1	1	1	1
1.5	0.01	0.04	0.63	0.32	0.79	1	0.82	1	1	1	1	1	1	1	1
1.6	0.01	0.1	0.85	0.56	0.95	1	0.96	1	1	1	1	1	1	1	1
1.7	0.03	0.17	0.96	0.76	0.99	1	0.99	1	1	1	1	1	1	1	1
1.8	0.06	0.3	0.99	0.88	1	1	1	1	1	1	1	1	1	1	1
1.9	0.09	0.38	1	0.95	1	1	1	1	1	1	1	1	1	1	1
2	0.14	0.52	1	0.98	1	1	1	1	1	1	1	1	1	1	1
2.1	0.17	0.6	1	0.99	1	1	1	1	1	1	1	1	1	1	1
2.2	0.25	0.72	1	1	1	1	1	1	1	1	1	1	1	1	1
2.3	0.29	0.76	1	1	1	1	1	1	1	1	1	1	1	1	1
2.4	0.34	0.82	1	1	1	1	1	1	1	1	1	1	1	1	1
2.5	0.39	0.87	1	1	1	1	1	1	1	1	1	1	1	1	1

Note: we assume a multiplicative effect, a Cochran-Armitage test of trend, and an $\alpha = 5 \times 10^{-8}$. MAF, minor allele frequency.

6.0 How to contact leadership of the study components, including access to the standard NIH Material Transfer Agreement (MTA)

If you are interested or think you know someone or a group that could be interested in contributing to this effort, please contact DCEG/NCI:

Stephen Chanock: chanocks@mail.nih.gov
Sharon Savage: savagesh@mail.nih.gov

And our Project Manager

Vibha Vij: vibha.vij@nih.gov

We will arrange for a Material Transfer Agreement using our standard form, revised for this specific project. Please see Appendix 4 for Material Transfer Agreement template, for receipt of de-identified DNA, blood/blood products, buccal cells/salivacollected through a local IRB-approved protocol for enrollment, collection, genetic analysis and sharing of data as per NIH policies.
We require receipt of only de-identified samples that we would register as 'non-human subject research'.

7.0 Data sharing plan

Data will be shared in accordance with the NIH Genomic Data Sharing Policy. We will analyze de-identified samples collected under the auspices of an IRB with clear evidence of permission to share data with the research community through standard portals according the tenets of the NIH Data Sharing Policy. Exceptions can be made but only in very remarkable and unusual circumstances that require discussion and written verification of an agreed upon plan, sanctioned by the local IRB before de-identification.

After QC, the genetic data will be returned to the submitting principal investigators as well as deposited in dbGap. If the local IRB and the principal investigator permit, we will directly deposit in the resource of the COVID19HG.org project.

We will retain samples on request for later distribution (at the study level) at the direction of the submitting principal investigators. Alternatively, once, the genetic analyses are completed, we can return the materials to the submitting investigator's program with the written request of the principal investigator(s).

7.1 Authorship and publication policy

All contributors to COVNET will be offered an opportunity to be a co-author on any COVNET results paper that includes their data. Major authors, namely those conducting the analyses and drafting of the manuscript will be named on the masthead, which will be discussed and agreed upon by co-authors. All papers will include the COVNET Consortium as a named author and all contributors will be included as part of this Consortium name. Preferential submission to journals will be to those that enable the COVNET Consortium to be indexed and expanded on PubMed, listing or recognizing all co-authors (this is the expected outcome).

Authorship will be determined by contributions. Both co-first authors and co-last authors will be liberally used will bear a footnote indicating that they are co-first authors. Between the co-first, co-last and critical genomics and biostatistical analysts, co-authors from participating studies and NIH staff will be listed alphabetically. Each individual cohort will designate the appropriate investigators from their study for inclusion in the manuscript author list. Corresponding authorship(s) will be determined by the first/last co-authors.

Preference will be given to submission to a pre-print archive at or before submission to a peer-review journal. Data will be available according to the NIH Data Sharing Poliy at or before the submission of the manuscript to a peer-review journal.

References

1. Wu Z, McGoogan JM: **Characteristics of and Important Lessons From the Coronavirus Disease 2019 (COVID-19) Outbreak in China: Summary of a Report of 72 314 Cases From the Chinese Center for Disease Control and Prevention**. *JAMA* 2020, **323**(13):1239-1242.
2. Lu X, Zhang L, Du H, Zhang J, Li YY, Qu J, Zhang W, Wang Y, Bao S, Li Y *et al*: **SARS-CoV-2 Infection in Children**. *N Engl J Med* 2020, **382**(17):1663-1665.
3. Guan W-j, Ni Z-y, Hu Y, Liang W-h, Ou C-q, He J-x, Liu L, Shan H, Lei C-l, Hui DSC *et al*: **Clinical Characteristics of Coronavirus Disease 2019 in China**. *N Engl J Med* 2020, **382**(18):1708-1720.
4. The Novel Coronavirus Pneumonia Emergency Response Epidemiology Team: **he Epidemiological Characteristics of an Outbreak of 2019 Novel Coronavirus Diseases (COVID-19) — China, 2020**. In: *China CDC Weekly.* vol. 2; 2020: 113-122.
5. Zhang X, Tan Y, Ling Y, Lu G, Liu F, Yi Z, Jia X, Wu M, Shi B, Xu S *et al*: **Viral and host factors related to the clinical outcome of COVID-19**. *Nature* 2020.
6. Mehra MR, Desai SS, Kuy S, Henry TD, Patel AN: **Cardiovascular Disease, Drug Therapy, and Mortality in Covid-19**. *N Engl J Med* 2020.
7. Li X, Xu S, Yu M, Wang K, Tao Y, Zhou Y, Shi J, Zhou M, Wu B, Yang Z *et al*: **Risk factors for severity and mortality in adult COVID-19 inpatients in Wuhan**. *J Allergy Clin Immunol* 2020:S0091-6749(0020)30495-30494.
8. Zhou F, Yu T, Du R, Fan G, Liu Y, Liu Z, Xiang J, Wang Y, Song B, Gu X *et al*: **Clinical course and risk factors for mortality of adult inpatients with COVID-19 in Wuhan, China: a retrospective cohort study**. *The Lancet* 2020, **395**(10229):1054-1062.
9. Chen R, Liang W, Jiang M, Guan W, Zhan C, Wang T, Tang C, Sang L, Liu J, Ni Z *et al*: **Risk Factors of Fatal Outcome in Hospitalized Subjects With Coronavirus Disease 2019 From a Nationwide Analysis in China**. *Chest* 2020, **S0012-3692**(20):30710-30718.
10. Pan D, Sze S, Minhas JS, Bangash MN, Pareek N, Divall P, Williams CML, Oggioni MR, Squire IB, Nellums LB *et al*: **The impact of ethnicity on clinical outcomes in COVID-19: A systematic review**. *EClinicalMedicine*.
11. Zheng X, Shen J, Cox C, Wakefield JC, Ehm MG, Nelson MR, Weir BS: **HIBAG—HLA genotype imputation with attribute bagging**. *The Pharmacogenomics Journal* 2014, **14**(2):192-200.
12. Conomos MP, Miller MB, Thornton TA: **Robust inference of population structure for ancestry prediction and correction of stratification in the presence of relatedness**. *Genetic epidemiology* 2015, **39**(4):276-293.
13. Maples BK, Gravel S, Kenny EE, Bustamante CD: **RFMix: a discriminative modeling approach for rapid and robust local-ancestry inference**. *American journal of human genetics* 2013, **93**(2):278-288.
14. Lu X, Zhang L, Du H, Zhang J, Li YY, Qu J, Zhang W, Wang Y, Bao S, Li Y *et al*: **SARS-CoV-2 Infection in Children**. *N Engl J Med* 2020.
15. Guan WJ, Ni ZY, Hu Y, Liang WH, Ou CQ, He JX, Liu L, Shan H, Lei CL, Hui DSC *et al*: **Clinical Characteristics of Coronavirus Disease 2019 in China**. *N Engl J Med* 2020.

Appendix 1: COVID-19 Case Codes

Strict		
Type	**Code**	**Description**
ICD10	U07.1	Diagnosis of COVID-19 confirmed by laboratory testing
SNOMED-CT international	3947183016	Disease caused by severe acute respiratory syndrome coronavirus 2 (disorder)
SNOMED-CT international	3947184010	Disease caused by severe acute respiratory syndrome coronavirus 2
SNOMED-CT international	3947185011	COVID-19
SNOMED-CT international	3947189017	Severe acute respiratory syndrome coronavirus 2 (organism)
SNOMED-CT international	3947191013	Severe acute respiratory syndrome coronavirus 2
SNOMED-CT international	3947190014	SARS-CoV-2
SNOMED-CT international	3947180018	Antigen of severe acute respiratory syndrome coronavirus 2 (substance)
SNOMED-CT international	3947181019	Antigen of severe acute respiratory syndrome coronavirus 2
SNOMED-CT international	3947182014	Antigen of SARS-CoV-2
CTV3	Y20d1	Confirmed 2019-nCoV (Wuhan) infection
EMIS	EMISNQCO303	2019-nCoV (Wuhan) infection
SNOMED-CT UK	1240581000000100	2019-nCoV (novel coronavirus) detected
SNOMED-CT UK	1240751000000100	Disease caused by 2019-nCoV (novel coronavirus)
SNOMED-CT UK	1240561000000100	Encephalopathy caused by 2019-nCoV (novel coronavirus)
SNOMED-CT UK	1240571000000100	Gastroenteritis caused by 2019-nCoV (novel coronavirus)
SNOMED-CT UK	1240531000000100	Myocarditis caused by 2019-nCoV (novel coronavirus)
SNOMED-CT UK	1240521000000100	Otitis media caused by 2019-nCoV (novel coronavirus)
SNOMED-CT UK	1240551000000100	Pneumonia caused by 2019-nCoV (novel coronavirus)
SNOMED-CT UK	1240541000000100	Upper respiratory tract infection caused by 2019-nCoV (novel coronavirus)
SNOMED-CT UK	1240511000000100	Detection of 2019-nCoV (novel coronavirus) using polymerase chain reaction technique
SNOMED-CT UK	1240381000000100	Severe acute respiratory syndrome coronavirus 2 (organism)

Broad		
Type	**Code**	**Description**
ICD10	U07.2	Diagnosis of COVID-19 suspected or probable
SNOMED-CT international	3947197012	Suspected disease caused by severe acute respiratory coronavirus 2 (situation)
SNOMED-CT international	3947196015	Suspected disease caused by severe acute respiratory coronavirus 2
SNOMED-CT international	3947195016	Suspected disease caused by SARS-CoV-2
SNOMED-CT international	3950926016	Suspected COVID-19
CTV3	Y20cf	Suspected 2019-nCoV (Wuhan) infection
EMIS	EMISNQSU106	Suspected 2019-nCoV (Wuhan) infection
SNOMED-CT UK	1240761000000100	Suspected disease caused by 2019-nCoV (novel coronavirus)
ICD10	U07.1	Diagnosis of COVID-19 confirmed by laboratory testing
SNOMED-CT international	3947183016	Disease caused by severe acute respiratory syndrome coronavirus 2 (disorder)
SNOMED-CT international	3947184010	Disease caused by severe acute respiratory syndrome coronavirus 2
SNOMED-CT international	3947185011	COVID-19
SNOMED-CT international	3947189017	Severe acute respiratory syndrome coronavirus 2 (organism)
SNOMED-CT international	3947191013	Severe acute respiratory syndrome coronavirus 2
SNOMED-CT international	3947190014	SARS-CoV-2
SNOMED-CT international	3947180018	Antigen of severe acute respiratory syndrome coronavirus 2 (substance)
SNOMED-CT international	3947181019	Antigen of severe acute respiratory syndrome coronavirus 2
SNOMED-CT international	3947182014	Antigen of SARS-CoV-2
CTV3	Y20d1	Confirmed 2019-nCoV (Wuhan) infection
EMIS	EMISNQCO303	2019-nCoV (Wuhan) infection
SNOMED-CT UK	1240581000000100	2019-nCoV (novel coronavirus) detected
SNOMED-CT UK	1240751000000100	Disease caused by 2019-nCoV (novel coronavirus)
SNOMED-CT UK	1240561000000100	Encephalopathy caused by 2019-nCoV (novel coronavirus)
SNOMED-CT UK	1240571000000100	Gastroenteritis caused by 2019-nCoV (novel coronavirus)
SNOMED-CT UK	1240531000000100	Myocarditis caused by 2019-nCoV (novel coronavirus)
SNOMED-CT UK	1240521000000100	Otitis media caused by 2019-nCoV (novel coronavirus)

SNOMED-CT UK	1240551000000100	Pneumonia caused by 2019-nCoV (novel coronavirus)
SNOMED-CT UK	1240541000000100	Upper respiratory tract infection caused by 2019-nCoV (novel coronavirus)
SNOMED-CT UK	1240511000000100	Detection of 2019-nCoV (novel coronavirus) using polymerase chain reaction technique
SNOMED-CT UK	1240381000000100	Severe acute respiratory syndrome coronavirus 2 (organism)

Appendix 2: Questionnaire for confirmed COVID-19 cases

DEMOGRAPHICS

1. What is the date this survey is being conducted?
   - ☐ _____(mm/dd/yyyy)

2. What is the patient's birth year?
   - ☐ _____(yyyy)
   - ☐ Don't know

3. What is the patient's biological sex (sex assigned at birth)?
   - ☐ Male
   - ☐ Female

4. What is the patient's current height (fill in one)?
   - ☐ _____(ft) _____(in)
   - ☐ _____(in)
   - ☐ _____(cm)
   - ☐ Don't know
   - ☐ Prefer not to answer

5. What is the patient's current weight (fill in one)?
   - ☐ _____(lb)
   - ☐ _____ (kg)
   - ☐ Don't know
   - ☐ Prefer not to answer

6. How would you describe the patient's race?
   - ☐ America Indian/Alaska Native
   - ☐ Asian
   - ☐ Native Hawaiian or other Pacific Islander
   - ☐ Black or African American
   - ☐ White
   - ☐ More than one race
   - ☐ Don't know
   - ☐ Prefer not to answer

7. Is the patient Hispanic or Latino?
   - ☐ Yes
   - ☐ No
   - ☐ Don't know
   - ☐ Prefer not to answer

SYMPTOMOLOGY

8. When did the patient first develop symptoms of COVID-19?
   - ☐ _____ (mm/dd/yyyy)
   - ☐ Don't know

9. What date was the patient officially diagnosed with COVID-19?
   - ☐ _____(mm/dd/yyyy)
   - ☐ Don't know

DISEASE COURSE

10. What date is the beginning of this patient's current period of hospitalization (including at another facility if transferred)?
    - ☐ _____(mm/dd/yyyy)
    - ☐ Don't know
    - ☐ This patient was not admitted to a hospital to treat COVID-19

11. What COVID-19 associated symptoms has this patient reported (check all that apply)?
    - ☐ Cough
    - ☐ Fever
    - ☐ Difficulty breathing
    - ☐ Fatigue/lethargy
    - ☐ Muscle pain
    - ☐ Lost sense of smell (anosmia) or taste
    - ☐ None of the above
    - ☐ Don't know

12. How many days did this patient report a cough?
    - ☐ _____ (days)
    - ☐ Don't know
    - ☐ Prefer not to answer
    - ☐ Not applicable

13. How many days did this patient report or have a fever?
    - ☐ _____(days)
    - ☐ Don't know
    - ☐ Prefer not to answer
    - ☐ Not applicable

14. How would you describe the patient's respiratory rate at admission to your facility (or at admission to a prior facility if transferred)?
    - ☐ Not elevated
    - ☐ Slightly elevated
    - ☐ Moderately elevated
    - ☐ Severely elevated
    - ☐ Don't know

15. Describe the patient's current pneumonia status:

- ☐ Not present
- ☐ Suspected, no radiological signs
- ☐ Unilateral
- ☐ Bilateral (dual)
- ☐ Don't know

16. Did the patient present with or does the patient currently have septic shock?
    - ☐ No
    - ☐ Yes
    - ☐ Don't know

17. Did the patient present with or does the patient currently have organ failure?
    - ☐ No
    - ☐ Yes
    - ☐ Don't know

18. Please check the box next to any of the following respiratory supports this patient has required while admitted to your facility (or prior facility if transferred):
    - ☐ Oxygen (mask, nasal cannula)
    - ☐ Non-invasive ventilation (CPAP, BIPAP)
    - ☐ Intubation
    - ☐ None of the above
    - ☐ Don't know
19. How many days has your patient required ventilator support while admitted to your facility (include days at a prior facility if transferred)?
    - ☐ _____(days)
    - ☐ Don't know

EXPOSURES AND RISK FACTORS

20. Please check the box next to any of the following underlying health conditions/co-morbidities this patient has:
    - ☐ Diabetes mellitus
    - ☐ Coronary heart disease
    - ☐ Cerebrovascular Disease (stroke)
    - ☐ Atrial fibrillation
    - ☐ Hypertension
    - ☐ COPD
    - ☐ Asthma
    - ☐ Malignant tumor (cancer)
    - ☐ Chronic renal disease or renal insufficiency
    - ☐ Immunocompromised condition
    - ☐ Autoimmune disease
    - ☐ Thyroid disease
    - ☐ Chronic liver disease

- None of the above
- Don't know

21. Has the patient ever regularly vaped an e-cigarette or similar device?
    - Currently uses every day
    - Currently uses some days of the month
    - Former user
    - Never
    - Don't know

22. Describe the patient's smoking history immediately prior to their admission to your facility:
    - Never smoke
    - Smoked less than 100 cigarettes in lifetime
    - Currently smokes daily
    - Currently smokes some days a moth
    - Stopped smoking in the last year
    - Stopped smoking 1-4 years ago
    - Stopped smoking 5-9 years ago
    - Stopped smoking 10 or more years ago

23. If the patient is a current or former smoker, how many cigarettes did/does the patient smoke daily?
    - The patient does not smoke regularly
    - 1-10
    - 11-20
    - 21-30
    - 31 or more
    - Don't know/prefer not to answer

---

24 – 32. What types of medications has the patient taken? For each medication, please indicate when this was last taken or select "never" if never taken. Please note the selection "don't know" implies the medication has been taken at some point in the past.

---

24. Non-steroidal anti-inflammatory drugs (ex. ibuprofen: Advil, Motrin, etc.)
    - Today or yesterday
    - 2-7 days ago
    - 1-4 weeks ago
    - 1-2 months ago
    - Over a year ago
    - Never
    - Don't know
    - Prefer not to answer

25. Paracetamol/acetaminophen (ex. acetaminophen: Tylenol)
- Today or yesterday
- 2-7 days ago
- 1-4 weeks ago
- 1-2 months ago
- Over a year ago
- Never
- Don't know
- Prefer not to answer

26. Asthma medication (ex. albuterol – Ventolin)
- Today or yesterday
- 2-7 days ago
- 1-4 weeks ago
- 1-2 months ago
- Over a year ago
- Never
- Don't know
- Prefer not to answer

27. Pollen allergy medications (ex. diphenhydramine – Sudafed, ioratadine - Claritin)
- Today or yesterday
- 2-7 days ago
- 1-4 weeks ago
- 1-2 months ago
- Over a year ago
- Never
- Don't know
- Prefer not to answer

28. Antibiotics penicillin, azithromycin (Z-pack)
- Today or yesterday
- 2-7 days ago
- 1-4 weeks ago
- 1-2 months ago
- Over a year ago
- Never
- Don't know
- Prefer not to answer

29. Myocardial infarction or stroke medication (ex. digoxin)
- Today or yesterday
- 2-7 days ago

- 1-4 weeks ago
- 1-2 months ago
- Over a year ago
- Never
- Don't know
- Prefer not to answer

30. Blood thinners (ex. warfarin – Coumadin, rivaroxaban – Xarelto)
    - Today or yesterday
    - 2-7 days ago
    - 1-4 weeks ago
    - 1-2 months ago
    - Over a year ago
    - Never
    - Don't know
    - Prefer not to answer

31. ACE-inhibitors for blood pressure (ex. enalapril, lisinopril)
    - Today or yesterday
    - 2-7 days ago
    - 1-4 weeks ago
    - 1-2 months ago
    - Over a year ago
    - Never
    - Don't know
    - Prefer not to answer

32. ARBs (Angiotensin II Receptor Blockers) (ex. candesartan- Altacand, valsartan-Diovan)
    - Today or yesterday
    - 2-7 days ago
    - 1-4 weeks ago
    - 1-2 months ago
    - Over a year ago
    - Never
    - Don't know
    - Prefer not to answer

33. Did the patient report exposures to others with COVID-19?
    - Never
    - Once
    - Multiple times
    - Don't know

34. Did the patient report wearing a face mask to protect against COVID-19 infection?

- No
- Yes
- Don’t know

Appendix 3: Phenotype definitions, version 1

Phenotypes are consistent with those outlined by the COVID-19 Host Genetics Initiative (https://www.covid19hg.org/).

Patient inclusion criteria: laboratory confirmed SARS-CoV-2 infected individuals with a clinical window of 15-90 days from diagnosis.

Primary analyses

Analysis 1

- Cases: death OR respiratory support (intubation, CPAP, CPN or BiPAP) due to SARS-CoV-2 infection.
- Controls: cases that are alive without respiratory support at any point since diagnosis.

Analysis 2

- Cases: hospitalized (in-patient) individuals due to SARS-CoV-2 related symptoms.
- Controls: non-hospitalized cases.

Extended analyses

Analysis 3 -mortality

- Cases: individuals that died due to SARS-CoV-2 infection.
- Controls: cases that are alive.

Analysis 4 -severity ordinal

- Continuous outcome: 3 level scale (mild, severe, critical) of COVID19 disease severity based on https://jamanetwork.com/journals/jama/fullarticle/2762130.
    - Mild: nonpneumonia and mild pneumonia.
    - Severe: dyspnea, respiratory frequency ≥30/min, blood oxygen saturation ≤93%, partial pressure of arterial oxygen to fraction of inspired oxygen ratio <300, and/or lung infiltrates >50% within 24 to 48 hours.
    - Critical: respiratory failure, septic shock, and/or multiple organ dysfunction or failure.

Analysis 5-susceptibility

- Cases: laboratory confirmed SARS-CoV-2 infected individuals.
- Controls: individuals without documented diagnosis of COVID-19 [Laboratory tested for SARS-CoV-2 infection AND all tests (if multiple tests) negative OR self-reported tested negative for SARS-CoV-2 infection (e.g. by questionnaire)].

Analysis 6-susceptibility-severe

- Cases: laboratory confirmed SARS-CoV-2 infected individuals and hospitalized.
- Controls: individuals without documented diagnosis of COVID-19 [Laboratory tested for SARS-CoV-2 infection AND all tests (if multiple tests) negative OR self-reported tested negative for SARS-CoV-2 infection (e.g. by questionnaire)].

Appendix 4: Material transfer agreement template

MATERIAL TRANSFER AGREEMENT
FOR COVID-19 GENOMICS PROTOCOL

The National Cancer Institute, National Human Genome Research Institute, and National Institute of Allergy and Infectious Disease, institutes of the National Institutes of Health, are conducting a prospective study to explore the genetic contributions to severity of disease among people who are infected with the novel coronavirus SARS-CoV-2. Coded samples with accompanying de-identified data will be accepted from collaborating institutions for genomic analyses, and samples will be stored and may be further distributed for future COVID-19 research.

This Material Transfer Agreement ("Agreement") is between the National Cancer Institute ("RECIPIENT") and the organization listed below ("PROVIDER") and will become effective on the date of the last authorized signature below. Under this Agreement, PROVIDER will transfer to RECIPIENT the following human specimens: _____ with accompanying de-identified human data, if any (collectively "MATERIAL").

RECIPIENT and PROVIDER agree as follows:

1. The above MATERIAL is made available for research on COVID-19, the novel coronavirus, 2019-nCoV, or related topics.

2. THIS MATERIAL IS NOT FOR USE IN HUMAN SUBJECTS.

3. The MATERIAL may be further distributed by RECIPIENT for the purpose described above in Article 1 under terms no more restrictive than this Agreement.

4. The RECIPIENT agrees to acknowledge the source of the MATERIAL in any public disclosures reporting use of the MATERIAL, including any publicly available deposit of genetic sequence data from the MATERIAL.

5. Any MATERIAL delivered pursuant to this Agreement is understood to be experimental in nature and may have hazardous properties. THE PROVIDER MAKES NO REPRESENTATIONS AND EXTENDS NO WARRANTIES OF ANY KIND, EITHER EXPRESSED OR IMPLIED. THERE ARE NO EXPRESS OR IMPLIED WARRANTIES OF MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE, OR THAT THE USE OF THE MATERIAL WILL NOT INFRINGE ANY PATENT, COPYRIGHT, TRADEMARK, OR OTHER PROPRIETARY RIGHTS.

6. The RECIPIENT agrees to use the MATERIAL in compliance with all applicable statutes and regulations.

7. The PROVIDER represents that the MATERIAL has been collected from human subjects in accordance with applicable laws and regulations. PROVIDER will provide coded specimens and de-identified data which will not include patient identifying information. RECIPIENT agrees to comply with all applicable statutes, regulations and ethical requirements to protect the identity and privacy of human subjects from whom the MATERIAL was collected.

SIGNATURES BEGIN ON THE NEXT PAGE

The PROVIDER and RECIPIENT will sign this Agreement and the PROVIDER will then send the MATERIAL.

PROVIDER INFORMATION and AUTHORIZED SIGNATURE

PROVIDER Scientist:
PROVIDER Organization:

_____ _____
Signature of Authorized Official Date
Name:
Title:
Address for Notices:

RECIPIENT INFORMATION and AUTHORIZED SIGNATURE

RECIPIENT Scientist: Dr. Stephen Chanock
RECIPIENT Organization: National Cancer Institute ("NCI")

_____ _____
Signature of Authorized Official Date
Name: Lisa D. Finkelstein, Ph.D.
Title: Supervisory Technology Transfer Manager
Address for Notices:
Technology Transfer Center, NCI
9609 Medical Center Drive
Rm 1E530
Rockville, MD 20850-9702

Appendix 5: COVIDcode Protocol and Attachments

Human Subjects Research Protocol

Project title: Genetics of COVID-19 susceptibility and manifestations

Draft/version number: June 22, 2020

Protocol Number: 20-HG-0090

Principal Investigator: Leslie G. Biesecker, M.D.

Chief, Medical Genomics and Metabolic Genetics Branch
National Human Genome Research Institute
Building 50, Rm 5140
50 South Drive
Bethesda, MD 20892
Phone: 301-402-2041
Fax: 301-480-0353
E-mail: lesb@mail.nih.gov

Title/Section/Branch/Institute: Clinical Genomics Section/Medical Genetics and Metabolic Genomics Branch/National Human Genome Research Institute

Contents

List of Abbreviations

AE	Adverse Event/Adverse Experience
CFR	Code of Federal Regulations
CLIA	Clinical Laboratory Improvement Amendment of 1988
COI	Conflict of Interest
COVID-19	Coronavirus disease 2019
CRADA	Cooperative Research and Development Agreement
DAC	Data Access Committee
dbGaP	Database of Genotypes and Phenotypes
DUC	Data Use Certificate
DHHS	Department of Health and Human Services
DSMB	Data Safety and Monitoring Board
ES	Exome Sequencing
FWA	Federal Wide Assurance
GCP	Good Clinical Practice
GS	Genome Sequencing
GINA	Genetic Information Nondiscrimination Act
GWAS	Genome-Wide Association Study
HIPAA	Health Insurance Portability and Accountability Act
ICF	Informed Consent Form
iPSC	Induced Pluripotent Stem Cells
IRB	Institutional Review Board
MTA	Material Transfer Agreement
N	Number (typically refers to number of subjects/sample size)
NHGRI	National Human Genome Research Institute, NIH
NGS	Next Generation Sequencing
NIH	National Institutes of Health
OHRP	Office for Human Research Protections
OHSRP	Office of Human Subjects Research Program
PHI	Protected Health Information
PII	Personally Identifiable Information
PI	Principal Investigator
PK	Pharmacokinetics
QA	Quality Assurance
QC	Quality Control
SAE	Serious Adverse Event/Serious Adverse Experience
SARS-CoV-2	Severe acute respiratory syndrome coronavirus type 2
SNP	Single nucleotide polymorphism
SOP	Standard Operating Procedure
UP	Unanticipated Problem
UPnonAE	Unanticipated Problem that is not an Adverse Event

STATEMENT OF COMPLIANCE

The protocol will be carried out in accordance with International Conference on Harmonisation Good Clinical Practice (ICH GCP) and the following:

- United States (US) Code of Federal Regulations (CFR) applicable to clinical studies (45 CFR Part 46, 21 CFR Part 50, 21 CFR Part 56, 21 CFR Part 312, and/or 21 CFR Part 812)

National Institutes of Health (NIH)-funded investigators and trial site staff who are responsible for the conduct, management, or oversight of NIH-funded trials have completed Human Subjects Protection and ICH GCP Training.

The protocol, informed consent form(s), recruitment materials, and all participant materials will be submitted to the Institutional Review Board (IRB) for review and approval. Approval of both the protocol and the consent form must be obtained before any participant is enrolled. Any amendment to the protocol will require review and approval by the IRB before the changes are implemented to the study. In addition, all changes to the consent form will be IRB-approved; a determination will be made regarding whether a new consent needs to be obtained from participants who provided consent, using a previously approved consent form.

1.0 Precis

The current SARS-CoV-2 pandemic presents a serious challenge to public health. Individuals infected with SARS-CoV-2 experience extremes in symptomatology ranging from a complete lack of symptoms to rapidly worsening end-stage pulmonary disease. The explanatory mechanism underlying susceptibility to severe disease remains unknown. We hypothesize that underlying genetic factors are at least partially explanatory. We aim to employ a "phenotypic extremes" approach to rapidly ascertain severely and mildly affected COVID-19 patients for genomic interrogation to identify germline and somatic variants that may play a role in host susceptibility to disease to correlate those phenotypic extremes with genetic variants. We will employ both a rare and common variant approach, using both genome sequencing and SNP chip analysis and B and T cell repertoire interrogation.

2.0 Objectives and specific aims

Primary objectives

To identify potential germline susceptibility variants that determine host responses to disease that would allow improved patient stratification and identification of potential therapeutic targets to moderate the severe symptoms of the disease.

1. Identify common and rare germline variants associated with host susceptibility to severe or fatal COVID-19 disease using a case-case design.
2. Deposit and share data as fast as possible to allow community analyses by bona-fide researchers who seek permission to analyze the data according to NIH data sharing precepts.

Secondary objectives

Perform exploratory analyses of epigenetic signatures, serologic immune markers and antibody profiles, and other possible techniques to discover other mechanisms of disease. These will use samples from our prioritized sample collection list (see below).

1. We will collect whole blood specimens for sera and DNA that will support these activities, which will be developed dynamically during the protocol.
2. We will collect cells where possible to explore B and T cell repertoire.
3. We will collect serum or plasma where possible to explore humoral response and soluble mediators such as cytokines.
4. We will collect RNA sample tubes where possible for transcriptomic analysis.

3.0 Brief rationale and background

The infectious disease syndrome known as Coronavirus disease 2019 (COVID-19) is caused by the etiologic agent named acute respiratory syndrome coronavirus 2 (SARS-CoV-2). This condition was originally described in late 2019 in Wuhan, China, and has since become the 2019–20 coronavirus pandemic. The common symptoms include fever, cough, and shortness of breath. Other symptoms such as muscle pain, sputum production, diarrhea, conjunctivitis, and sore throat are seen less frequently. The majority of cases result in mild symptoms and many have little or no symptoms. In contrast, a minority progress to severe pneumonia and multi-organ failure. The rate of deaths per number of diagnosed cases is 4.1%; however, it ranges from 0.2% to 15% according to age group and other health problems. It is likely that case fatality ratio will decline as more mildly affected cases are identified through expanded testing.

Irrespective of the actual rate of severe disease and death, it is clear that there will be enormous numbers of severe cases worldwide and management strategies are desperately needed.

The emergent need for treatment and management strategies is because of the high transmissibility of the virus and the apparently low rate of natural immunity. SARS-CoV-2 is spread from person to person by respiratory droplets produced during coughing. It may also be spread via hand to mouth (mucous membrane) transmission. The typical time from exposure to symptom onset is typically two to 12 days, with an average of five days. The standard diagnostic method is by reverse transcription polymerase chain reaction (rRT-PCR) from a nasopharyngeal swab. The infection can also be diagnosed clinically from a combination of symptoms, risk factors and a chest CT scan compatible with multifocal ground glass infiltrates or pneumonia.

There is clearly a dramatic range of severity associated with this infection. Children appear to rarely experience any severe symptoms[14]. A recent review of 72,314 cases by the Chinese Center for Disease Control and Prevention showed that <1% of the cases were in children younger than 10 years of age, when clearly the pediatric population of that region is many times higher than that. Young adults are uncommonly symptomatic. Overall, the median age of Chinese patients was 47 years and 41.9% were female.[15] 5.0% were admitted to the ICU, 2.3% underwent invasive mechanical ventilation, and 1.4% died.

		Disease severity		ICU, ventilator, or death	
AGE	N (%)	Nonsevere N (%)	Severe	Yes	No
0-14yr	9/1011 (0.9)	8/848 (0.9)	1/163 (0.6)	**0**	9/946 (1.0)
15-49 yr	557/1011 (55.1)	490/848 (57.8)	67/163 (41.1)	**12/65 (18.5)**	545/946 (57.6)
50-64 yr	292/1011 (28.9)	241/848 (28.4))	51/163 (31.3)	**21/65 (32.3)**	271/946 (28.6)
≥65 yr	153/1011 (15.1)	109/848 (12.9)	44/163 (27.0)	**32/65 (49.2)**	121/946 (12.8)

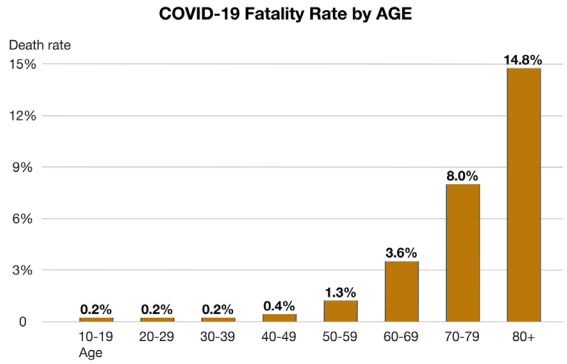


Figure 1 From: http://weekly.chinacdc.cn/en/article/id/e53946e2-c6c4-41e9-9a9b-fea8db1a8f51

From the table above, it is clear that there is a major effect of age with nearly half of those who were over 65 and hospitalized required ICU care, ventilatory support, or dying. No children had this outcome. Of course, these data are limited to observations of hospitalized patients. Another presentation of data that includes non-hospitalized patients to demonstrate the skewing is from the Chinese CDC (Figure 1), which again demonstrates substantial differences by age. The most common complicating issue in the Chinese patients was acute respiratory distress syndrome, with septic shock the next most common. Although the age skewing is clear, at all ages only a minority of individuals develop the severe aspects of this disease and we hypothesize that this susceptibility is, at least in part, due to genetic factors. The purpose of this protocol is to identify genetic variations that contribute to this susceptibility to severity and poor outcome.

We are submitting this proposal as a 'bare bones' initial proposal considering that we are in a public health emergency. The imperative is to begin sample collection and generate data as quickly as we possibly can.

To this end, we outline here a basic study design and eligibility parameters, but we know that some specific processes are highly likely to evolve rapidly in the next weeks and months. Participants will consent to genomic and molecular studies, data sharing, and future research use of data and samples for the purpose of understanding the pathogenesis of COVID-19 related disease. We have described the nature of the data that are acquired and generated and how these data are analyzed to allow for flexibility and refinement as iterative data analysis begins and additional knowledge is generated. This project will adopt an aggressive approach to data sharing with rapid uploads to widely accessible portals such as AnVIL. We will also share coded data and SNP genotypes with the members of the COVID-19 Host Genetics Initiative consortium, again to facilitate the most rapid detection of associations.

4.0 Description of study design

This is a prospective genomic immunologic study to explore the genetic contributions to severity of disease among people who are infected with SARS-CoV-2. We will use a 'phenotypic tails' or 'extreme phenotypes' design to associate variants with severely affected disease presentation. Our hypothesis is that people with severe disease have germline variations in genes that modulate disease severity. We will analyze data from both identifiable participants who directly interact with our research group as well as coded data and samples sent to us by partnering institutions in accordance with their local IRBs' policies. The study will include both SNP-chip testing, currently anticipated for all participants, and a subset analysis of genome sequencing analysis, which will be applied to those with low risk factors and severe presentation versus those with a mild (or asymptomatic) presentation.

5.0 Description of procedures

In this highly fluid environment, we propose this as our initial plan, which will be subject to substantial change in the coming weeks. As outlined below in Section 6, we define "participants" as identifiable individuals with a range of severity of symptoms of COVID-19 and the research activities requiring or resulting from interactions with identifiable participants described herein apply only to this group. We will also accept coded samples and data from partnering institutions with their own local IRB approval – most of these are yet to be identified and, importantly, will not be counted as human subjects as we will not have access to identifying information for these individuals.

We have three "Sampling Groups", listed here:

1) NIHCC inpatients
2) Participants recruited via OMS or otherwise referred to the study
3) Coded samples from regional hospitals

Prospective participants in Sampling Group 1 will be identified by first contacting the primary PI of that protocol (which we expect will be an NIAID investigator). Sampling Group 2 comprises individuals identified via case finding in OMS and individuals who have been made aware of the study by external collaborators or NIH investigators (i.e. participants who are already enrolled on other NIH protocols who have been made aware of this study by virtue of their participation in another NIH protocol). Coded specimens from outside hospitals, which comprises Sampling Group 3, will be ascertained, recruited, and enrolled by each of those centers and these will be reviewed and approved by the local IRB. See Section 6.2 and the study flowchart below for additional details.

Individuals' consent, clinical data, and sample collection will be handled distinctly in each group. For Sampling Group 1, blood drawing will be done by NIHCC staff. Research staff will verbally confirm that they have already consented and then sample them by venous phlebotomy at the participant's point of care. Participants in Sampling Group 2 may have labs collected at the NIHCC or mailed in via sample collection

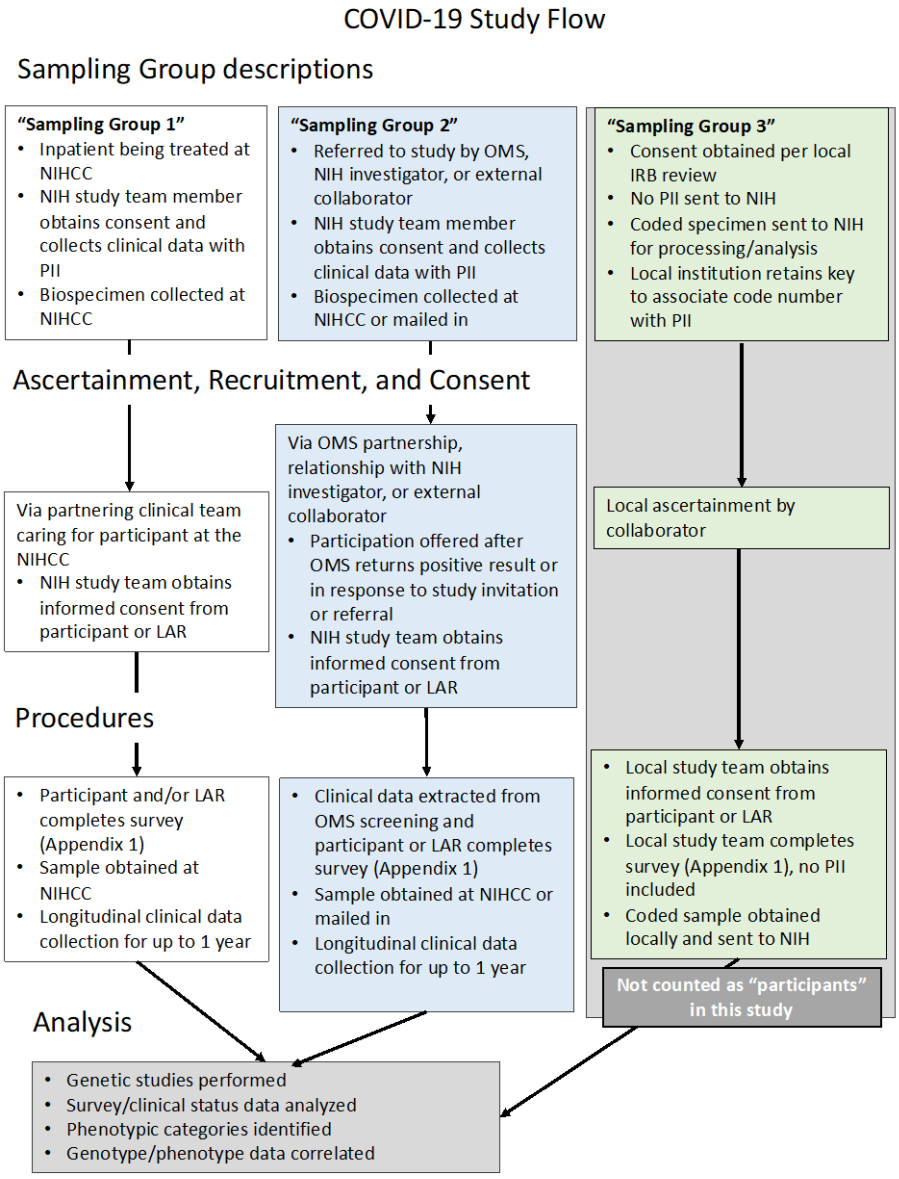
kits we provide. For Sampling Group 3, those processes will be determined by local investigators (and their IRB).

For participants, our sample collections will be based on a prioritized list, recognizing that some participants may be quite ill, and that blood drawing may be limited to levels below that of the NIHCC blood drawing guidelines. Collaborators sending us coded data/specimens from their local institution will be provided with similar guidance re: sample prioritization. In no case would we exceed the NIHCC limits (or even come close to those) but we recognize that with a disease process where PaO_2/FiO_2 may be low due to reduced diffusion capacity caused by the pneumonia, clinicians can (and should) consider limiting phlebotomy to levels well below that of the standard guidelines. As well, clinicians may consider no sampling to be permitted in a case, per good medical practice. For those who can tolerate sample collection we have established this prioritized list:

Priority sample list: most important to least important

1) 6 cc EDTA (for DNA)
2) 6 cc EDTA (for whole cell assays)
3) 8 cc serum separator tube (for serum assays)
4) 8 cc PAXgene tube, 2.5 cc draw volume (for transcriptomics)

Figure 1. Study Flow



5.1 Medical information

5.1.1 Research only medical information

We will work with participants/LARs and/or their providers to obtain the clinical and demographic data listed in Appendix 1.

For coded specimens received as part of specific collaborations, we will not gather personally identifiable information on these patients/samples. Collaborators obtaining samples from Sampling Group 3 will complete a version of Appendix 1 and will be instructed to only append their local codes to this information.

We anticipate that data will be kept for the entire duration of the study. Coded data (including samples/data without PII that we collect from outside institutions) may be

shared with other researchers. At the time the study is closed, a proposal to the IRB will be made to keep the data or destroy it. At no point will NIH investigators seek to obtain PII on individuals whose samples sent without PII are accessioned to the study from outside institutions.

At this time, we anticipate that we may collect follow up clinical data (Appendix 1) on participants for up to one year after enrollment. Further sampling for follow-up assays may be proposed in a future amendment.

5.1.2 Clinically indicated medical information

N/A

5.2 Biological Specimens

5.2.1 Research only collection of biological specimens

We propose to collect peripheral blood for DNA isolation at a minimum. Please see above for our priority sampling list. The total maximum blood drawn may be 22.5 ml. These amounts are well below NIH blood drawing limits of 5 cc/kg.

We plan to accept coded blood and DNA samples from collaborating institutions. DNA will be isolated by standard procedures by the Biesecker or Chanock/NCI Cancer Genomics Research Laboratory. DNA samples will be stored for the duration of the study by the Chanock laboratory biobank and may be retained for future research on COVID-19 related topics.

As mentioned above, we will store samples for the duration of the project, the exact duration of which is difficult to estimate given the emergent nature of the COVID-19 pandemic. We propose an initial sample retention period of five years and will approach the IRB as needed to request up- or downward revision of this estimate.

5.3 Approved Drugs Being Used for Research

N/A

5.4 Unapproved Drugs/Devices

N/A

5.5 Describe questionnaires or other psychological instruments and estimate how long they will take to complete, and whether they address sensitive topics (Submit copies.)

Please see Appendix 1. We will attempt to gather this information from participants and their providers with as few contacts as possible and may also extract this information from records sent to us whenever possible to reduce stress on participants and providers. We may ask participants recruited through OMS to share with us the COVID-19 Monitoring Log they are provided by OMS nurses to keep track of their progression of symptoms.

5.6 Specific results that will be given to participants or their health care providers

At this time, considering the exigencies of a public health emergency, we anticipate returning no individual results to these participants. In this public health emergency, it is not appropriate for us to divert resources towards secondary findings identification and return, as it would be a burden to consent to return of results. As well, we recognize patients and providers are stressed under the current circumstances, and it would be unreasonable to add to their charge. We intend to maintain the option to potentially re-contact and reconsent for return of secondary findings after the pandemic. We would only do that after working with the IRB to determine if it would be practical and appropriate.

6.0 Description of study population

6.1 Estimated number of participants, enrollment ceiling, and anticipated enrollment by year.

We anticipate 2,500 cases for the identified participants. De-identified samples/data received from collaborators will not be counted as participants enrolled in this protocol.

6.2 Description and justification of clinical inclusion/exclusion criteria.

All populations (children, individuals with impaired decision-making ability, etc.) are susceptible to infection with SARS-CoV-2 and may subsequently develop mild or severe symptoms of COVID-19. Any of these individuals may be admitted to the NIHCC for treatment of COVID-19. As well, we plan to actively recruit individuals tested through NIH OMS; these individuals are NIH workers and OMS reports that many actively query OMS as to possible ways they can participate in intramural research. As such, no vulnerable populations are excluded from participation in this protocol.

The Sampling Groups for this protocol are defined here:

Sampling Group 1: Individuals admitted to the NIHCC under a protocol to study COVID-19 (these protocols are primarily housed in NIAID).

Sampling Group 2:

a) NIH workers who test positive for SARS-CoV-2 infection via NIH Occupational Medicine Services
b) Existing participants in NIH studies referred to or made aware of this study by an NIH Investigator/study team
c) Individuals referred to this study by an outside collaborator

Sampling Group 3: De-identified samples/clinical data derived from individuals enrolled in a research study allowing for the sharing and storage of biospecimens and genomic data at an institution whose research team wishes to contribute these materials to our study

Please see Section 7 below for a detailed description of our Analysis Groups and how they relate to our Sampling Groups. The inclusion/exclusion criteria listed below only apply to Sampling Groups 1 and 2 (as each local collaborator may set additional criteria per their local studies).

Inclusion Criterion:

- Positive test for SARS-CoV-2 virus infection (serology or PCR-based)
- Weight $\geq$ 10 kg
- Age $\geq$ 3 years old
- Meets criteria for Sampling Groups 1 and 2 as outlined above

Exclusion Criterion:

Individuals (or LARs/family members when applicable) whom we cannot consent for participation in a language offered by Cyracom

As the study progresses, we may amend our inclusion/exclusion criteria to include COVID-19 disease attributes and will submit these accordingly.

6.3 Location of study

Participants will be recruited from the NIH Clinical Center (Sampling Group 1), through OMS, other NIH protocols, or referred to the study by a collaborator (Sampling Group 2). We will receive exempted samples and clinical data from collaborating hospitals and research institutions (Sampling Group 3).

6.4 Description of recruitment strategies

Participants will be recruited through several pipelines. For Sampling Group 1 we will recruit individuals admitted to the NIH Clinical Center for COVID-19 disease by approaching primary study staff for the admitting protocol. COVID-19 patients with mild to severe symptoms admitted to the NIH Clinical Center will be offered the opportunity to enroll in this protocol. For Sampling Group 2 we will directly recruit individuals tested at OMS, including NIH employees, as well as participants enrolled in other NIH studies. Prospective participants may also be referred to this study by their local care providers and other outside collaborators. We have developed a recruitment flyer (Appendix 2) in English and Spanish to assist in recruitment/offering participation to prospective participants. Prospective participants (Sampling Group 1 and Sampling Group 2) may receive this flyer (Appendix 2) from OMS, their primary study team, or by their local care providers. Referring investigators or providers may develop recruitment materials specific to their patient population (Appendix 2).

For Sampling Group 3 this project will be piloted with collaborating hospitals and research institutions in the surrounding area and may expand to include institutions across the United States and internationally. Potential institutions who have expressed interest in assisting in the recruitment of participants to their locally approved protocol and submitting coded/de-identified samples include Inova Hospitals, Georgetown University (MedStar) Hospital, Children's National Medical Center, and Northwell Hospital System in New York.

6.5 For existing sample/data sets, note whether samples were originally collected for research or clinical practice. If obtained for research, include a description of the original purpose of study and prior plans for sample storage. Was consent obtained that would be applicable to this study?

N/A

6.6 Description of criteria for withdrawal from study.

Participants may withdraw from the study at any time and are encouraged to exercise this right if they are uncomfortable with the study process or our approach to this research. We make it clear to each participant that research is not for everyone and that we will not take offense to such a withdrawal.

If a participant chooses to withdraw, we will discuss his/her preferences regarding ongoing participation in the various components of the study including interactions with study personnel, utilization of participants' samples and data by us, and utilization of participants' samples and data by other researchers (see Section 10.8). We will make reasonable efforts to exercise the preferences indicated by each participant. However, there may be some limitations surrounding our level of oversight of the use of samples and data by other researchers. For example, if a participant's sample has been shared with many other researchers, it will be impossible to ensure that each one of those researchers destroys that sample and stops depositing data on databases. If data have been shared with shared repositories, they may not be taken down. For these complex reasons, we plan on

discussing each withdrawal on a case-by-case basis. If a participant declines to discuss his/her preferences with us, we will discontinue interacting with him/her, we will remove personally identifying information (PII) from the research samples and data, and we may or may not continue to use his/her samples and data. The consent forms explain these limitations of withdrawal.

6.7 Description of study discontinuation and closure.

Participants are free to withdraw from participation in the study at any time upon request (See Section 6.6). This is not an interventional study, and therefore we do not anticipate that the PI will discontinue or withdraw a participant from the study.

This study may be temporarily suspended or prematurely terminated if there is sufficient reasonable cause. If the study is prematurely terminated or suspended, the PI will inform the IRB and will provide the reason(s) for the termination or suspension.

Circumstances that may warrant termination or suspension include, but are not limited to:

- Determination of unexpected, significant, or unacceptable risk to participants
- Demonstration of efficacy that would warrant stopping
- Insufficient compliance to protocol requirements
- Data that are not sufficiently complete and/or evaluable
- Determination that the primary endpoint has been met
- Determination of futility

The study may resume once concerns about safety, protocol compliance, and data quality are addressed, and satisfy the IRB.

7.0 Description of study statistical considerations and/or analytic plan

7.1 Common variant analyses

We are describing this as Analysis Group 1. This will be based on the generation of common and uncommon genotypes using the Illumina GSA version 3 biobank chip (e.g., 700,000 SNP markers). Genotypes will be called from the iDAT file and a standard pipeline for assessing quality control metrics will be performed to remove poor performing samples, sex-discordant samples, highly related subjects and incomplete SNP assays. Genotypes that have met the quality control metrics will be uploaded to the Michigan server for imputation of up to 66 million SNPs. Imputation will include HLA using the multi-ethnic reference panel of 21,000 individuals based on deep-sequencing coverage. After further quality control metrics are applied to the output of the actual and imputed genotype set, the full genotype set will be combined with available phenotypes and covariates to generate an analytic data set. Population substructure will be assessed by standard programs to determine the admixture coefficients for subsets of the population tested based on eigenvectors which are critical to control against spurious genetic associations due to population ancestry.

The initial analytical plan will be to conduct a series of models that take into account underlying conditions and population ancestry. Statistical models will be generated, and analyses conducted to determine if single SNPs or pathways achieve genome wide significance. Since we will be using imputed variants down to estimates of 0.01% allele frequency, effect sizes will be estimates with large confidence intervals- and the need for replication. The basic principles outlined by the COVID-19 Host Genetics Initiative

common analysis pipeline will be followed. The primary approach will be a fixed effects analysis of the primary data set accounting for appropriate covariates (that could be confounders) and focus on case-case analysis- namely mild versus severe cases. Subsequent fixed effects meta-analysis with available data from other genotyped resources for COVID19 outcomes will be pursued.

SNPs that achieve genome wide significance based on standard single SNP assays will be confirmed if imputed by a second laboratory test and further investigated with fine-mapping strategies to identify highly correlate variants based on linkage disequilibrium patterns and current bioinformatic programs that infer biological function (such as dbRegulome). If sets of SNPs achieve or approach genome wide significance, then polygenic risk scores can be estimated by several methods.

7.2 Rare variant analyses

We are describing this as Analysis Group 2. For genome sequence data we will attempt to identify rare variants that may predispose to COVID-19 susceptibility. Sequence variants will be sifted bioinformatically, based on existing algorithms and those that we will develop in collaboration with the COVID-19 Host Genetics Initiative and other collaborators. Our approach to the identification of causative mutations in these patients will evolve process informed by our own initial efforts and success from collaborators and other groups. The approach is to apply filters to the data to reduce plausible candidate genes to a manageable number and then combine by gene bins, pathway bins, genomic space bins, or other concepts and then compare the burden in severe vs. mild cases. Some of the attributes upon which the variants could be filtered include:

1) variants not in dbSNP
2) variants in the same gene in multiple patients with the same disorder
3) variants *not* found in either normal DNA samples or in mildly affected cases
4) variants predicted to be damaging
5) variants in regions linked to the phenotype by future GWAS studies
6) variants in genes that are in a KEGG or GO pathway where mutations in another gene in that pathway cause a related phenotype (e.g., the RAS pathway for Noonan, CFC, etc.)

Regardless of which of these molecular and analytic methods are used, in which combinations, in which order, we believe that the risks and the benefits to the subjects are not significantly altered and that it is appropriate to consent the subjects broadly to genome sequencing and SNP testing.

The issue of follow-up studies downstream of the expected gene identifications described above is challenging to address. As is the case in positional cloning studies, the nature of the process is to identify genes not previously suspected to cause the phenotype at hand. Therefore, it is very difficult to predict what studies may be necessary or useful in the follow-up of those gene identification efforts. Occasionally, gene identification studies indicate involvement of a gene that is already well characterized for another reason (because it is involved in another phenotype, is itself interesting from a basic science perspective, was known to cause a phenotype in an animal model, etc.). In these situations, little downstream research may be necessary and the work that is needed can be accomplished by collaboration. In contrast, a gene could be identified that is an anonymous transcript with no known function, no conserved motifs, no mutant animal, etc. In this case much will need to be done.

Overall, the analysis methods for this study will greatly evolve while it is underway. As members of the COVID-19 Host Genetics Initiative, we will extensively collaborate and share data on rare variant analyses with the consortium, per NIH policies on Genomic Data Sharing.

8.0 Description of potential benefits of study

8.1 Direct benefits to participants

Participants are not expected to benefit directly from this study.

8.1.1 Specific results that will be given to participants or their health care providers

At this time, we do not anticipate that we will return any results to the participants.

8.2 Collateral benefit to participants

We do not anticipate any collateral benefits. We do plan to make a summary of overall results of the study available to participants, likely through the study website.

8.3 Benefits to society

Understanding the genetic contributions to the severity of disease caused by SARS-CoV-2 has the potential to improve diagnostic tools and enhance treatment development.

9.0 Description of likelihood and seriousness of harms and how safety will be maximized:

9.1 Therapeutic interventions

N/A

9.2 Diagnostic interventions

The likelihood of complications from blood draws is low and if these occur, they are unlikely to be other than minor (bruising, infection, etc.).

9.3 Radiation

N/A

9.4 Sedation

N/A

9.5 Risks to family relationships and other psychosocial or economic harms

We do not anticipate at this time that individual results will be returned to participants and thus the harms mentioned in the prompt are not applicable. All individuals who participate in this study will be diagnosed with (or suspected to be infected with) SARS-CoV-2. We recognize that interacting with severely ill individuals and/or their family members or LARs to invite them to participate in a research study requires attention to the individual emotional circumstances of each family. Staff performing these functions will be trained in this aspect of this protocol and will also have access to support (i.e., peer supervision, case conferences, etc.). We have also taken steps to reduce the burden associated with participation by relying on short and empathetic verbal consent (see Section 14), providing participants the resources they need to reach out to us should they desire more information and/or when they feel this is temporally appropriate.

10.0 Description of how privacy and confidentiality of medical information/biological specimens will be maximized

10.1 Description of any clinical/demographic information that will be included. (age, ethnicity, sex, diagnosis, stage, treatment)

We will collect some demographic and clinical information for identifiable participants (Sampling Groups 1 and 2). Staff will be alerted to the fact that many identifiable participants in this study may be NIH employees; although all normal measures to protect participant privacy and confidentiality will certainly employed, the fact that some participants in this study may be our colleagues will require protocol personnel to be especially mindful of these measures. See Section 14.4 for additional details. Participating institutions (Sampling Group 3) will collect this information at their site, per their protocol, using our forms, and return this as coded (exempt) data.

Data generated from de-identified specimens (Sampling Group 3) will be retained in our databases, but we will not count these samples as participants enrolled in this protocol.

10.2 How might this information make specific individuals or families identifiable?

It is unlikely that this information will make specific individuals or families identifiable, except to investigators via their EMR in the NIH Clinical Center (Sampling Groups 1 & 2). Appropriate measures will be taken to de-identify samples. Access to the key linking coded data to PII is restricted.

10.3 If research data will be coded, how will access to the "key" for the code be limited?

All samples from Sampling Groups 1 and 2 will be coded upon arrival in the research laboratory. Samples and data from Sampling Group 3 will already be coded before being sent to NIH. Outside institutions sending de-identified samples/data (i.e., Sampling Group 3) will be offered access to a tracking application currently employed in several NCI protocols which transfers de-identified data to Westat/NCI. The use of this application is voluntary. Local hospitals may collect the data using a local application that is installed on their local PC. Then when needed local hospital staff will be able to upload the de-identified data as a .csv file to Westat/NCI via a secure file transfer process using the local application. The application connects to the Westat/NCI FTP using SFTP (Secure FTP or SSH FTP) which is built on\uses SSH (Secure Shell protocol).

For Sampling Groups 1 and 2, subjects' and prospective subjects' names and other information will be kept in records associated with the NIHCC and in a laboratory sample database, which is password-protected (i.e., LabMatrix, CTDB, NCI, and NIAID databases).

10.4 Will pedigrees be published?

N/A

10.5 Will personally identifiable information be released to third parties?

As per standard NIH policy, this is only performed upon provision of a signed release from a participant or a court order.

10.6 Under what circumstances will data/samples be shared with other researchers or deposited in various repositories, biobanks, and/or databases voluntarily or as mandated by NIH policies (e.g., genomic data sharing policy, available at http://gds.nih.gov)?

This study will be conducted in accordance with publication and data sharing policies and regulations. National Institutes of Health (NIH) Public Access Policy, which ensures that the public has access to the published results of NIH-funded research. It requires scientists to submit final peer-reviewed journal manuscripts that arise from NIH funds to the digital archive PubMed Central upon acceptance for publication. De-identified data from this study will be made available to collaborators and as members of the COVID-19 Host Genetics Initiative and we will extensively collaborate and share data on rare variant analyses with the consortium, per NIH policies on Genomic Data Sharing.

This study will comply with the NIH Genomic Data Sharing Policy, which applies to all NIH-funded research that generates large-scale human or non-human genomic data, as well as the use of these data for subsequent research. Large-scale data include genome-wide association studies (GWAS), single nucleotide polymorphisms (SNP) arrays, and genome sequence, transcriptomic, epigenomic, and gene expression data.

Samples and/or data may be shared with other researchers in coded form if their studies relate to the broadly defined purpose of increasing the understanding of the genetic contributions to disease expression of COVID-19, per standard NIH Genomic Data Sharing principles.

10.7 Describe any additional features to protect confidentiality.

N/A

10.8 What circumstances would prompt the PI to report to the IRB loss or destruction of samples. For additional information, see SOP 5, "NIH Research Activities with Human Data/Specimens" or contact OHSRP at 301-402-3444.

We will report any samples that are lost or destroyed and represent a Privacy Act breach to the NHGRI clinical director in accordance with SOP 5.

11.0 Assessment of risk/benefit ratio

The risks of participating in this study are minimal given the limited protocol-related activities for participants and rigorous measures in place to protect participant confidentiality.

12.0 Unanticipated Problems: Collection, monitoring, analysis and reporting of adverse events and protocol deviations

12.1 Describe plan to monitor and report adverse events and protocol deviations.

Adverse events, non-compliance both serious or continuing, protocol deviations both major and minor, as well as unanticipated problems are defined and described by the NIH Office of Human Subjects Research Protection Policy #801 and will be reported in accordance with this policy.

12.2 Describe whether a Data Safety and Monitoring Board (DSMB) and/or any other additional monitoring measures will be used.

We propose that no DSMB is needed as the study is not therapeutic. The PI will be responsible for monitoring data quality and safety.

13.0 Description of alternatives to participation

Participants could enroll at another site that may be accepting volunteers. See https://covid-19genehostinitiative.net/about for information on other sites.

14.0 Description of consent process

14.1 Who will obtain consent?

Qualified, trained, and authorized Associate Investigators may obtain informed consent for Sampling Groups 1 and 2.

14.2 Setting where consent will be obtained

Consent and assent for Sampling Groups 1 and 2 will typically be obtained over the phone. If a participant is too ill to consent, we will obtain consent from their next of kin/LAR. We may have the ability to obtain samples from deceased individuals in which case we will ask for permission from their next of kin. See Section 14.4 for details.

14.3 What information will be provided to participants?

We will endeavor to make the Study Information Sheet (Appendix 2) available to all participants from Sampling Groups 1 and 2 prior to engaging with our team; this material will be provided in paper or electronic format (i.e., PDF) to prospective participants if they have not previously received it. Close coordination with the primary protocol team (Sampling Group 1) and OMS (Sampling Group 2) will be important and we will establish and maintain appropriate channels to accomplish this. We will establish and maintain a study website (Appendix 3) which outlines procedures and provides study staff contact information for Sampling Groups 1 and 2 and the study URL and a QR code linking directly to the website will appear on the study information sheet. Translations into other languages will be prepared and submitted as needed.

In accordance with OHSRP SOP 14F, Appendix C from this SOP will be provided, either electronically, physically, or verbally to Sampling Group 2 prior to or as part of the consent process for this study (this Appendix is appended to the consent process for Sampling Group 2, see below). All participants in Sampling Groups 1 and 2 will be provided written information about NIH policy regarding the Privacy Act and Certificate of Confidentiality (Appendices 4 and 5) prior to consent.

14.4 Protections for participants who may be vulnerable to coercion or undue influences

<u>Adults who lack capacity:</u>

People with severe symptoms of COVID-19 may be or become unable to consent, and the research questions cannot be answered by only enrolling adults who can consent. The risks of this research for adults who lack capacity to provide consent are minimal.

In accordance with OHSRP SOP 14E, if the subject does not have capacity to provide consent, it is possible to have the participant's legally authorized representative (LAR) provide consent on his or her behalf. For this study, a court appointed guardian, DPA, or LAR identified from the next-of-kin hierarchy can enroll adults who are not able to provide consent. Because this is a minimal risk study, we propose that, for individuals who are unable to consent because of the severity of their illness, the next of kin who are making health care decisions on behalf of the participant be considered as a LAR. In cases when a decisionally impaired subject has not already assigned a DPA and doesn't have a guardian appointed on his/her behalf, the ACAT (Ability to Consent Assessment Team) will be asked to evaluate his/her ability to assign a surrogate.

Once the LAR has been identified, the PI or other authorized and qualified AI will assess the appropriateness of the participant's LAR to provide consent. An appropriate LAR is one who at least: 1) understands that the protocol involves research; 2) understands the

risks, potential benefits (if any), and alternatives to the study; and 3) has sufficient reason to believe participation in the study is consistent with the subject's preferences and values. We do not intend to enroll individuals who have pre-existing status of limited cognition.

Children

For children with cognition above that of a 7-year-old, the team will use a verbal assent process (when possible and appropriate) in addition to obtaining informed consent via a LAR.

For telephone assent, the informed consent and assent documents will be sent to the parents/guardian and child. An explanation of the study will be provided over the telephone after the parents/guardian and child have had the opportunity to read recruitment materials (Appendices 2 and 3). Age-appropriate language will be used to discuss the study with the child. The parents/guardian will verbally consent to the study. The informed consent and assent process will be documented in the research database.

Obtaining reconsent from children at age of majority (waiver request when minor subjects will not be followed as adults):

When a pediatric subject reaches age 18, continued participation (including ongoing interactions with the subject or continued analysis of identifiable data) will require consenting of the now adult with the standard protocol consent document to ensure legally effective informed consent has been obtained.

If reconsent is not feasible, we request waiver of informed consent to continue to use data and/or specimens for those individuals who become lost to follow up or who have been taken off study prior to reaching the age of majority.

Requirements for Waiver of Consent consistent with 45 CFR 46.116 (d):

(1) The research involves no more than minimal risk to the subjects.
    a. Analysis of samples and data from this study involves no additional risks to subjects.
(2) The research could not practicably be carried out without the waiver or alteration.
    a. Considering the length of time between the minor's last contact with the research team and their age of majority, it will likely be very difficult to locate them again. A significant reduction in the number of samples analyzed is likely to impact the quality of the research.
(3) As the research involves using identifiable private information or identifiable biospecimens, the research could not practicably be carried out without using such information or biospecimens in an identifiable format.

    a. Though the purpose of future studies cannot yet be known, they often involve the correlation of clinical outcomes and clinical interventions with laboratory studies. Such information would be unavailable if access to medical record numbers was unavailable.
(4) The waiver or alteration will not adversely affect the rights and welfare of the subjects.

a. Retention of these samples or data does not affect the welfare of subjects.

(5) Whenever appropriate, the subjects will be provided with additional pertinent information after participation.

a. We only request a waiver of consent for those subjects who have been lost to follow-up or who have been taken off study prior to reaching the age of majority.

Statement regarding divorced parents with joint custody:

In cases where parents share joint legal custody for medical decision-making of a child (e.g., by a custody agreement or court order), it is NIH policy (OHSRP SOP 14D) that both parents must give their permission regardless of the risk level of the research. There are limited exceptions when one parent has since died, become incompetent, or is not reasonably available (e.g., in prison). Consent from the second parent can be obtained by telephone per NIH SOP 12.

NIH staff:

NIH staff members are not prohibited from enrollment if they meet the study's eligibility criteria. The study team will make every effort to protect the confidentiality of the NIH staff member's health information, to minimize any pressure on or discomfort of the NIH staff member, and provide a copy of the NIH Information Sheet on Staff Research Participation (SOP 14F, Appendix C), before consent is obtained. In addition, associate investigators working directly with identifiable data/specimens and/or potentially sensitive information from NIH staff member participants will be trained by the PI or lead AI to ensure that privacy and confidentiality is protected as much as possible. However, we will acknowledge to participants that there are limits to these protections. We do not intend to enroll any NIH staff from the labs of the PI and the associated investigators to avoid ethics concerns of enrolling individuals on the research team.

Note that Section 14.4 applies to Sampling Groups 1 and 2. Procedures for this for Sampling Group 3 are to be established locally.

14.5 Are there special circumstances regarding obtaining consent?

We request a waiver of documentation of written consent for this protocol for Sampling Groups 1 and 2 for the following reasons:

1. This is a public health emergency involving a highly transmissible virus with which all participants in this study will be infected. Minimizing the exchange of physical objects (i.e., paper consent forms) to the greatest degree possible is necessary to minimize risk to participants and staff.
2. This research is minimal risk and many of the individuals we will engage with will be severely ill and some may be hospitalized in highly isolated circumstances. The essential components of participation in this study can be adequately, quickly, and compassionately conveyed by phone. All participants will be provided with the study website which we will keep up to date. This will allow participants to control the degree of their interaction and do so when it is convenient and appropriate for them.
3. We do not want to restrict participation in this protocol to English-speakers only. The rapidly evolving nature of the COVID-19 pandemic necessitates quick action. Introducing a delay with certified translations of consents will limit the collection of crucial data.

Please see Appendices 4-6 for consent scripts/processes which will be employed with Sampling Groups 1 & 2 and the assent process for children. The potential participant will have the opportunity to review the study materials (Appendices 2 and 3), consider participation with others (family, friends, physician) and recontact us with a decision. The informed consent and assent processes will be documented in the research database.

When needed, certified and qualified interpreters will be used via Cyracom and the consent/assent scripts/processes interpreted as well. We will document when a translator has been used in the research database.

14.6 If this study involves collaborating sites, indicate if there is a single IRB review or if each site's IRB will review their site's participation in the study. Describe plans for ensuring appropriate IRB review and approval of consent forms at each site.

Coded samples may be obtained via collaborating sites (Sampling Group 3), however, because we will not have a way to link coded data with participant identifiers, this is not a multi-site study. Any site sending us coded data will do so under the auspices of their local IRB.

15.0 Description of any costs or financial compensation

15.1 Describe the rationale for and amount of any proposed compensation, consistent with SOP 13.

Participants will be offered compensation for completing protocol activities. We will offer $30 for the blood draw, $20 for the visit to the NIH Clinical Center or local blood draw site, and $10 for completing the survey. We will offer an additional $10 if participants are asked to fill out, and subsequently complete, a follow up survey. Participants must complete each of these protocol activities in order to receive the compensation for the activity. Compensation will be offered to the parents of a minor participant. For participants who are unable to consent, compensation will be offered to the LAR for completion of the survey and offered to the participant for the blood draw. NIH employees will be provided a copy of the NIH Information Sheet on Staff Research Participation (SOP 14F Appendix C).

15.2 Describe whether compensation will be modified if participant withdraws early.

A participant will not be compensated for any protocol activities they do not complete. Compensation will not be modified if a participant withdraws early, but they must complete each protocol activity to receive the designated compensation.

15.3 Describe any costs to participation.

There are no costs associated with participation.

16.0 References

1. Lu X, Zhang L, Du H, et al. SARS-CoV-2 Infection in Children. *N Engl J Med* 2020
2. Guan WJ, Ni ZY, Hu Y, et al. Clinical Characteristics of Coronavirus Disease 2019 in China. *N Engl J Med* 2020

Appendix 2 - Recruitment Materials

The text of the flyer is provided below and may be modified slightly to reflect minor updates (e.g., contact information, website URL) as needed. The specific version of the recruitment flyer for distribution by OMS contains the [[bracketed]] language below per OMS's request. The format may also be modified by the NHGRI communications team, without changing any of the content.

Why are we doing this study?

Coronavirus 2019 (COVID-19) is a serious public health problem. The goal of this study is to compare biomarkers like genetic differences in people with COVID-19 and understand how these markers may influence symptoms. We hope that we can use this information to help develop therapies to reduce the severity of symptoms of COVID-19.

Who can join?

Anyone who has tested positive for SARS-CoV-2 infection at any point may be eligible to join. We will use the clinical information and samples of people who test positive for our main analyses.

What is involved?

People in this study will:

· Provide a blood sample

· Fill out a brief health questionnaire either by phone or online

· Be contacted to share health and/or symptom updates with us for 30-60 days after joining the study

· Be offered compensation for completing the blood sample and brief health questionnaire (up to $70)

Privacy and confidentiality are very important to us. [[We know this is especially important for NIH Employees and their families.]] All information shared with us will be coded. We will not return individual genetic results, and after you fill out the survey and donate your blood sample, we plan very limited continued contact with you only to get updates on your health. We may recontact you for this information for a year after you have joined the study. [[This study is not connected to NIH Occupational Medical Services (OMS), and your participation in this study will have no impact on the care that OMS offers.]]

We intend for our study to help more patients benefit from this research. We hope that you will consider partnering with us on this important project. If you are interested, please contact us at COVIDcode@nih.gov or (240) 274-6777 (QR code).

Spanish Version (certified translation 22MAY2020)

¿Por qué realizamos este estudio?

La enfermedad por coronavirus de 2019 (COVID-19) es un problema grave de salud pública. El objetivo de este estudio es comparar biomarcadores, como diferencias genéticas en personas con COVID-19, y comprender cómo estos marcadores pueden influir en los síntomas. Esperamos poder usar esta información para ayudar a diseñar terapias a fin de reducir la intensidad de los síntomas de la COVID-19.

¿Quién puede inscribirse?

Toda persona que haya tenido un resultado positivo de la prueba de infección por SARS-CoV-2 puede ser apta para inscribirse. Usaremos la información clínica y las muestras de las personas con resultados positivos para nuestros análisis principales.

¿Qué implica?

Las personas en este estudio harán lo siguiente:

· darán una muestra de sangre;

· completarán un cuestionario médico corto, ya sea por teléfono o en línea;

· las contactaremos para que actualicen su estado de salud o sus síntomas de 30 a 60 días después de haberse inscrito en el estudio.

· recibirán una compensación por completar la muestra de sangre y un breve cuestionario de salud (hasta $70)

La privacidad y la confidencialidad son muy importantes para nosotros. [[Entendemos que este es especialmente importante para los empleados de NIH y sus familias.]] Codificaremos toda la información que nos dé. No le daremos los resultados individuales de las pruebas genéticas, y después de que complete el cuestionario y le tomemos la muestra de sangre, prevemos contactarlo en muy limitadas ocasiones solo para que nos actualice sobre su salud. Es posible que lo volvamos a contactar para obtener esta información 1 año después de que se haya inscrito en el estudio. [[Este studio no está conectado a los NIH Occupational Medical Services (OMS), y su participación en este estudio no tendrá ningún impacto en la atención que ofrece la OMS.]]

Tenemos la intención de ayudar a que más pacientes se beneficien de esta investigación. Esperamos que considere hacer equipo con nosotros para este importante proyecto. Si está interesado, comuníquese con nosotros a COVIDcode@nih.gov o al (240) 274-6777 (código QR).

ClinSeq ® Recruitment Letter

Dear Participant,

I hope that this letter finds you well amidst the current COVID-19 pandemic. Many people in the ClinSeq study have asked if there are research studies they can join to contribute to our understanding of this disease. I am sure that many more of you have had these questions, so I am writing to tell you what we and others at the NIH are doing.

If you have tested positive for COVID-19, we invite you to join a study we are doing called COVIDcode. This project looks at the role genes play in the severity of a person's symptoms of COVID-19. Participation involves one visit to the NIH for a blood draw and a survey. If you are willing to join, please contact the study team at COVIDcode@mail.nih.gov or (240)274-6777. You can learn more by going to https://www.genome.gov/covidcode.

If you had not had a positive test for COVID-19 (including if you had symptoms but were never tested), you may be eligible for other studies. These include projects to develop vaccines and treatments, to learn how many people had COVID-19 without symptoms, and to look at immune response in people with the disease. To learn more, please visit: https://www.niaid.nih.gov/diseases-conditions/covid-19-clinical-research

Whether you join these studies is your choice. If you choose not to join or withdraw from those studies, it will not affect your participation in ClinSeq®. We do not plan to do any COVID-19 research with the samples you gave for the ClinSeq® project. We never share your contact information with researchers unless you say we can.

The NIH also has a website with general information and resources about COVID-19 (www.nih.gov/health-information/coronavirus). It includes information about the symptoms, how to protect yourself (such as making cloth masks and cleaning your home), how to get testing, ways to cope (such as how to talk to manage stress or talk to children) and more.

As always, we are extremely grateful to you for your participation.

Wishing you the best of health,

Priscilla Chan

Research Assistant, ClinSeq

301-443-6160 clinseq@mail.nih.gov

ppendix 4: COVID-19 Verbal Consent Script for NIHCC participants (Sampling Group 1)

Consent will be administered verbally by study staff. Appropriate language describing relationship used throughout. "*You/your*" may be replaced with "*your [spouse/child/loved one]*" as needed. Participant/LAR questions will be answered as needed throughout; consenting study staff will return to the script after answering questions. All participants will be provided with written information about the NIH policies on the Privacy Act and Certificate of Confidentiality. If a participant indicates that they have not received this document, they will be sent the document.

_____

Thank you for your interest in the Genetics of COVID-19 susceptibility and manifestations research study. Please ask me any questions you want as they come up.

Have you reviewed the information sent to you about the Privacy Act and Certificate of Confidentiality? → If participant answers "no," the document will be sent to the participant to review prior to consent. If the participant is unable to receive the written document, it may be read to them.

Do you have any questions about this information?

The goal of this study is to learn more about how genes may play a role in COVID-19 disease. We will do this by studying the DNA of people who have tested positive for the new coronavirus.

It is your choice to join this study – you do not have to join the study. Receiving normal clinical care is an alternative to being in this study. If you decide to join the study, there are two main things we hope to get from you. First, we will work with you or your providers to complete a questionnaire asking about your health history and COVID-19 symptoms. Second, we will ask you to donate a blood sample, which is about 2 tablespoons of blood. We will use this blood sample to study your DNA. In this study, we will "read through" all of your DNA. This is called "sequencing."

As part of this research study, we will put your genetic information in a large database for broad sharing with the research community. Your genetic information will be labeled with a code and not with your name or other information that could be used to easily identify you, and only qualified researchers will be able to access this information. If you do not wish to have your data shared, you should not enroll in this study.

Your privacy and confidentiality are important to us. Your DNA and your health information will be coded. Only the people working directly on this study can connect your sample code to personal information, like your name and address. We plan to store your samples and health information and we may use them to do other research studies in the future. We may share only your coded information with other qualified researchers trying to learn about COVID-19. Other researchers we share your information with will not be able to link your specimens or data back to you.

There are minor risks to being in this study. Giving a blood sample can hurt and there is a small risk of infection. In spite of all of the safety measures that we will use, we cannot guarantee that your identity will never become known. There is a very small chance that you may be identified; for example, through unauthorized access to databases we will use. Every precaution will be taken to minimize this risk. You may not want to answer some of the questions we ask about your health. You may not benefit directly from this study and we won't return any individual genetic results to you.

You will be offered compensation for the completion of each of the protocol activities. You may receive $10 to complete the survey, $30 for the blood draw, and $20 for your visit to get the blood draw. You may be asked to complete a follow up survey. If you complete the follow up survey, you will be offered an additional $10.

You can withdraw from the study at any time. If you choose to withdraw, we need to talk with you about what to do with your samples and information. It may not be possible to destroy all data that have been generated on you. Most of your participation in this study will take place in the month or two after this phone call but we may recontact you for up to a year.

I will share information about how to access the study website with you. It will list all of our contact information; please call us if you have questions or concerns. If we have results to share from this study, we will post them on this website.

Do you have any [more] questions?

Do you want to move forward with joining the study?

"Yes" – participant/LAR agrees to join study
"No" – participant/LAR declines to join the study and is thanked by study staff

_____ Study Staff Signature and Date

Certificate of Confidentiality

To help us protect your privacy, the NIH Intramural Program has received a Certificate of Confidentiality (Certificate). With this certificate, researchers may not release or use data or information about you except in certain circumstances.

NIH researchers must not share information that may identify you in any federal, state, or local civil, criminal, administrative, legislative, or other proceedings, for example, if requested by a court.

The Certificate does not protect your information when it:

1. is disclosed to people connected with the research, for example, information may be used for auditing or program evaluation internally by the NIH; or
2. is required to be disclosed by Federal, State, or local laws, for example, when information must be disclosed to meet the legal requirements of the federal Food and Drug Administration (FDA);
3. is for other research;
4. is disclosed with your consent.

The Certificate does not prevent you from voluntarily releasing information about yourself or your involvement in this research.

The Certificate will not be used to prevent disclosure to state or local authorities of harm to self or others including, for example, child abuse and neglect, and by signing below you consent to those disclosures. Other permissions for release may be made by signing NIH forms, such as the Notice and Acknowledgement of Information Practices consent.

Privacy Act

The Federal Privacy Act generally protects the confidentiality of your NIH medical records we collect under the authority of the Public Health Service Act. In some cases, the Privacy Act protections differ from the Certificate of Confidentiality. For example, sometimes the Privacy Act allows release of information from your medical record without your permission, for example, if it is requested by Congress. Information may also be released for certain research purposes with due consideration and protection, to those engaged by the agency for research purposes, to certain federal and state agencies, for HIV partner notification, for infectious disease or abuse or neglect reporting, to tumor registries, for quality assessment and medical audits, or when the NIH is involved in a lawsuit. However, NIH will only release information from your medical record if it is permitted by both the Certificate of Confidentiality and the Privacy Act.

Your medical records are maintained at the NIH in accordance with the Privacy Act of 1974. Much of the medical information obtained about you will be stored in a computer system and used for research by NIH scientists, some of whom may have no personal contact with you. Much of the information will eventually be used in publications, but your identity will not be

revealed. In addition, certain diseases or conditions, including infectious diseases, may be reported to appropriate representatives of the State or Federal Government as required by law.

Appendix 5: COVID-19 Verbal Consent Script for Sampling Group 2

Consent will be administered verbally by study staff. Appropriate language describing relationship used throughout. "*You/your*" may be replaced with "*your [spouse/child/loved one]*" as needed. Participant/LAR questions will be answered as needed throughout; consenting study staff will return to the script after answering questions. All participants will be provided with written information about the NIH policies on the Privacy Act and Certificate of Confidentiality. If a participant indicates that they have not received this document, they will be sent the document. OMS participants will be provided with a physical or electronic copy of Appendix C from SOP14F prior to verbal consent whenever possible. If a participant indicates that they have not received this Appendix, then the text of the Appendix will be read to them.

_____

(For OMS participants) Thank you for your interest in the Genetics of COVID-19 susceptibility and manifestations research study. Have you reviewed Appendix C from the NIH SOP describing protections for NIH employees participating in research?
→ If participant answers "no," text of Appendix C from SOP14F will be read at this point. See below.

Have you reviewed the information sent to you about the Privacy Act and Certificate of Confidentiality? → If participant answers "no," the document will be sent to the participant to review prior to consent.
Do you have any questions about this information?

Please ask me any questions as they come up. The goal of this study is to learn more about how genes may play a role in COVID-19 disease. We will do this by studying the DNA of people who have tested positive for the new coronavirus.

It is your choice to join this study – you do not have to join the study. Receiving normal clinical care is an alternative to being in this study. If you decide to join the study, there are two main things we hope to get from you. First, we will work with you or your providers to complete a questionnaire asking about your health history and COVID-19 symptoms. Second, we will ask you to donate a blood sample, which is about 2 tablespoons of blood. We will use this blood sample to study your DNA. In this study, we will "read through" all of your DNA. This is called "sequencing."

As part of this research study, we will put your genetic information in a large database for broad sharing with the research community. Your genetic information *will be labeled with a code and not with your name or other information that could be used to easily identify you,* and only qualified researchers will be able to access this information. If you do not wish to have your data shared, you should not enroll in this study.

Your privacy and confidentiality are important to us. Your DNA and your health information will be coded. Only the people working directly on this study can connect your sample code to personal information, like your name and address. We plan to store your samples and health

information and we may use them to do other research studies in the future. We may share only your coded information with other qualified researchers trying to learn about COVID-19. Other researchers we share your information with will not be able to link your specimens or data back to you.

There are minor risks to being in this study. Giving a blood sample can hurt and there is a small risk of infection. In spite of all of the safety measures that we will use, we cannot guarantee that your identity will never become known. There is a very small chance that you may be identified; for example, through unauthorized access to databases we will use. Every precaution will be taken to minimize this risk. You may not want to answer some of the questions we ask about your health. You may not benefit directly from this study and we won't return any individual genetic results to you.

You will be offered compensation for the completion of each of the protocol activities. You may receive $10 to complete the survey, $30 for the blood draw, and $20 for your visit to get the blood draw. You may be asked to complete a follow up survey. If you complete the follow up survey, you will be offered an additional $10.

You can withdraw from the study at any time. If you choose to withdraw, we need to talk with you about what to do with your samples and information. Most of your participation in this study will take place in the month or two after this phone call but we may recontact you for up to a year.

I will share information about how to access the study website with you. It will list all of our contact information; please call us if you have questions or concerns. If we have results to share from this study, we will post them on this website.

Do you have any [more] questions?

Do you want to move forward with joining the study?

"Yes" – participant/LAR agrees to join study
"No" – participant/LAR declines to join the study and is thanked by study staff

_____ Study Staff Signature and Date

Appendix C from SOP 14F

APPENDIX C: NIH INFORMATION SHEET ON STAFF RESEARCH PARTICIPATION (APRIL 2016)

As an NIH employee, contractor, Special Volunteer, Guest Researcher, or trainee, you may participate in intramural research studies unless it is prohibited by your Institute or Center (IC), or if you are excluded by the criteria of the protocol in which you want to enroll. The inclusion of NIH staff in a particular protocol must also be approved by the IRB. You may be motivated by altruism, a commitment to research in your own or related fields, or want access to clinical trials of potential direct therapeutic benefit. When deciding, you should make an informed decision about participation. This information sheet offers some points to consider for NIH staff who are considering research participation at NIH.

First, similar to any individual who is considering research participation, you should seek adequate information about the study purpose, what is required of you in terms of procedures, interventions and time, and the potential risks and benefits of participation. For more information, see the NIH Clinical Center's public website "Are Clinical Studies for You?" at http://www.cc.nih.gov/participate/studies.shtml.

When you are thinking about participation in a research study that is being conducted by your supervisor, or others with whom you work closely in your laboratory, branch, or unit, you should consider some additional factors:

A. **Possible bias**: Are you confident that you can be unbiased about reporting answers, side effects, or other information that could influence the study outcome or risk to you?

B. **Confidentiality**: Has the principal investigator (PI) spoken about what information will be collected from you as part of the study? Has the PI discussed what information will be available to those within, and outside, the study team? If applicable, are you comfortable sharing your medical history (including, for example, mental health history or STDs) and your social history (e.g. substance use) with study investigators who may be your coworkers, or with the possibility of them discovering something about your health during the study (e.g. pregnancy status or a new diagnosis)? Although every effort will be made to protect your information and keep it private and confidential, your information may, depending on the nature of the protocol, become available in medical records or to authorized users outside of the study team. Discuss any concerns with the PI.

C. **Pressure**: Do you perceive any pressure or expectations from your supervisor or colleagues regarding participation? Could that pressure influence your decision or make it difficult for you to choose whether or not to participate? Remember that it is your choice whether or not to participate and that your decision to participate or not should not have an effect, either beneficial or adverse, on your position at NIH.

D. **Time and Compensation**: Can you take time off from work to complete the study requirements or participate solely during non-duty hours? Can you receive compensation for your participation in this study? Will your supervisor give you permission to participate during work hours? See the NIH Policy Manual 2300-630-3 *Leave Policy for NIH Employees Participating in NIH Medical Research Studies*.

E. **Consent Process:** Is the person obtaining your consent for the study your supervisor, a subordinate, or co-worker? If so, is there an independent person monitoring the consent process? If the study PI is a supervisor and intends to obtain consent from you, an independent person (e.g., through Bioethics or the NIMH Human Subjects Protections Unit [HSPU], or others as approved by the IRB) must monitor the consent process. If the person obtaining consent from you is a co-worker then an independent person (e.g., through Bioethics or the NIMH HSPU, or others as approved by the IRB) may be required to monitor the consent process, as determined by the IRB for the specific study.

If you are thinking of enrolling as a subject at the NIH Clinical Center and you have any questions or concerns, please contact the Office of Human Subjects Research Protections (OHSRP) at 301-402-3444 and/ or the Patient Representative if you are thinking of enrolling as a subject at the NIH Clinical Center on 301-496-2626. If you are at a NIH site outside the Clinical Center then please contact local site leadership.

Appendix 6: COVID-19 Verbal Assent Script for Minor Participants (Sampling Groups 1 &2)

Assent will be administered verbally by study staff. Participant answered as needed throughout; consenting study staff will return to the script after answering questions. [COVID-19] is a placeholder for whatever term the minor is most comfortable with in describing the disease associated with SARS-CoV-2 infection, e.g., many children commonly refer to the disease and the infection as "the coronavirus."

_____

We are asking you to join a research study to help us learn more about [COVID-19]. We want to study the DNA of people who have been tested for [COVID-19]. Studying your DNA may tell us why you got sick or did not get sick from [COVID-19].

If you decide to be in the study, we will get a blood sample and ask a grown-up like your mom or dad some questions about your health for up to a year. We will not tell you anything about your DNA that we study.

You do not have to be in this study if you do not want to. Giving blood can hurt. You may not want to answer the questions we ask. Some people do not want to have their DNA studied. These are reasons you may not want to be in the study.

Do you have any questions?

Do you want to move forward with joining the study?

"Yes" – participant agrees to join the study
"No" – participant declines to join the study and is thanked by study staff

_____ Study Staff Signature and Date