Update on COVNET (Genetic Determinants of COVID-19 Outcomes & Susceptibility)

COVNET Team

Division of Cancer Epidemiology and Genetics

NIH NATIONAL CANCER INSTITUTE

May 27, 2021

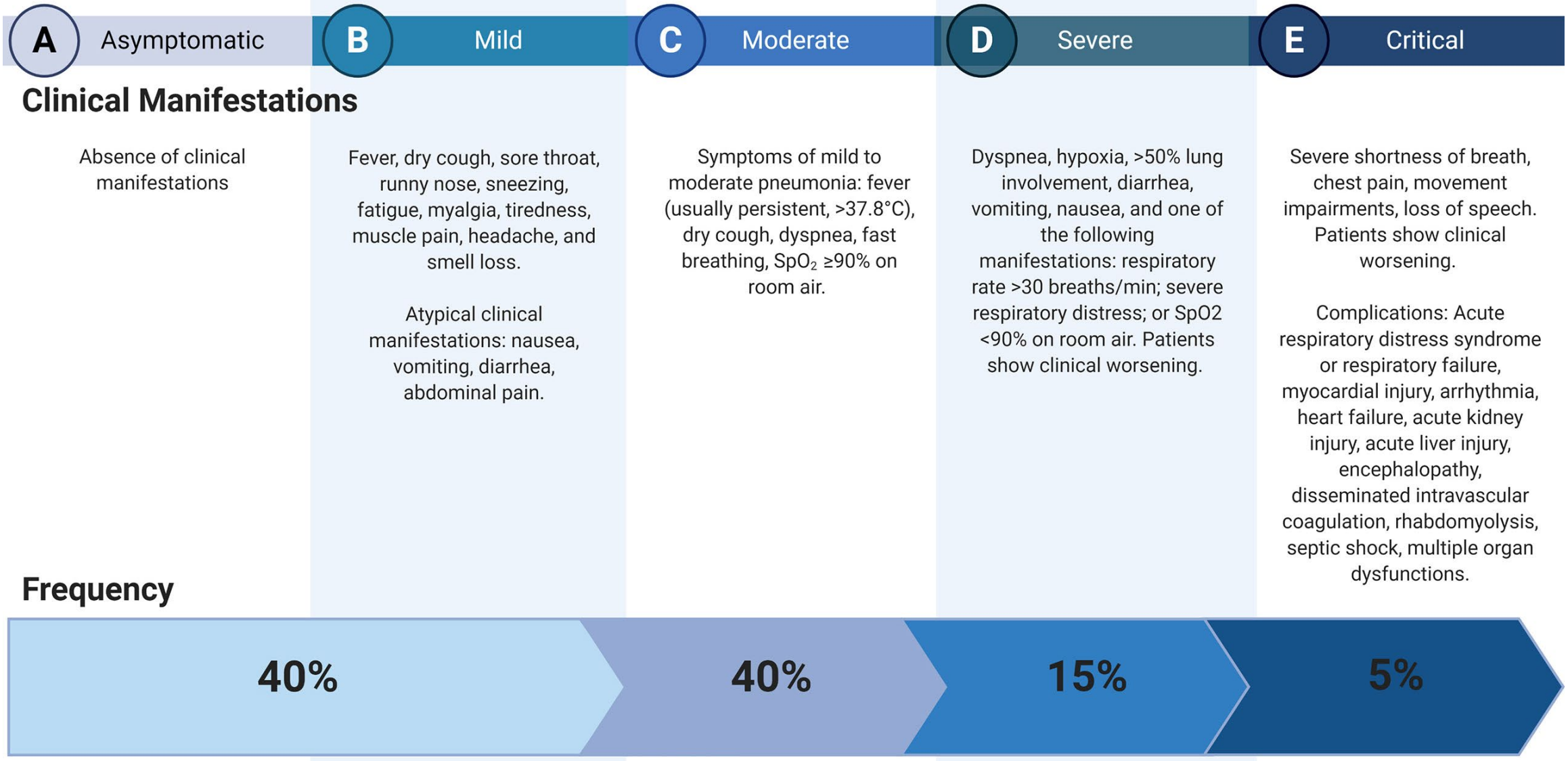
Three NCI/NIH COVID-19 genetic efforts in progress

1. **COVNET**: a large-scale study of the germline genetics of COVID-19 susceptibility and manifestations
   - 40,000 GWAS & 5,000 Whole Genome Sequencing (WGS) of COVID-19 cases

2. **COVIDcode Study**: NIH Clinical Center IRB approved study of 2,500 cases
   - Collaboration between investigators in NHGRI, NCI, and NIAID
   - Genetic analyses and extensive Immunologic assessment

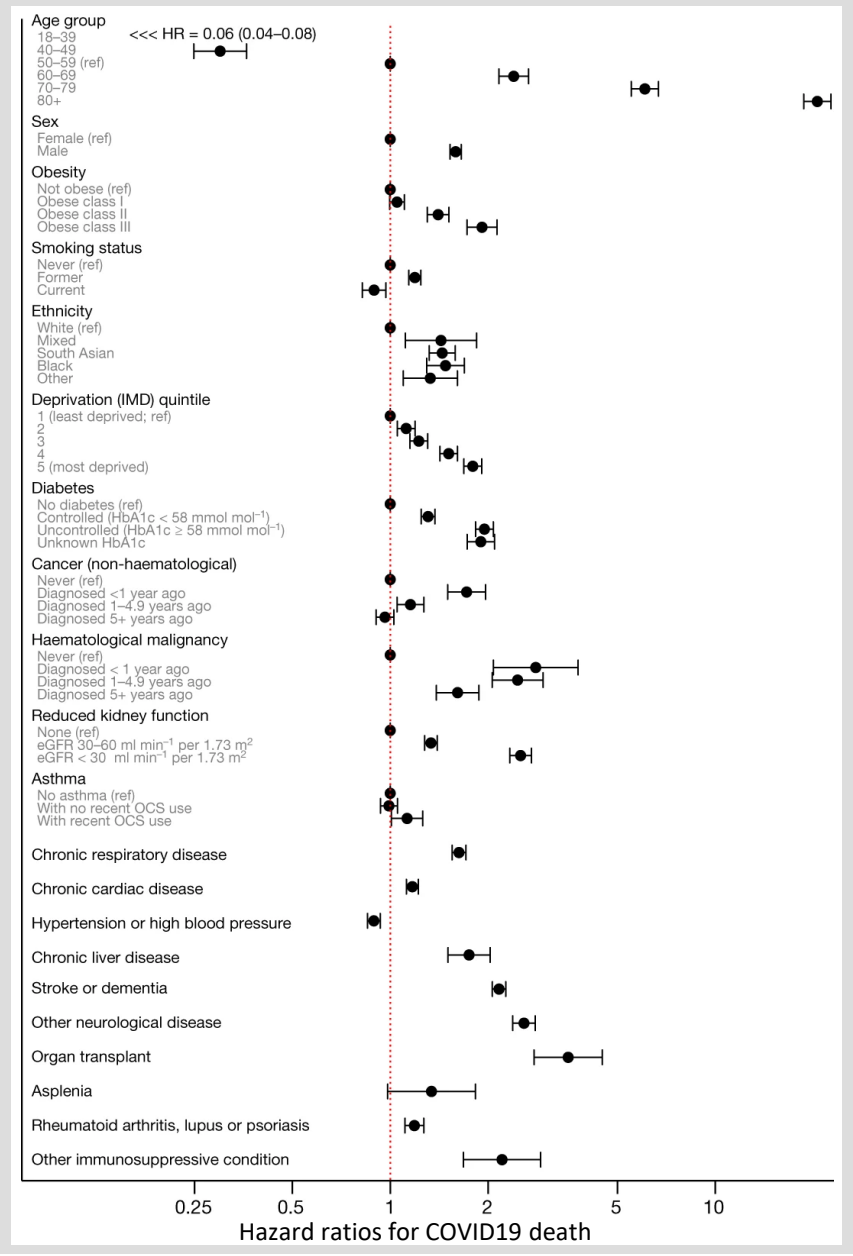
3. **NCI COVID-19 in Cancer Patients Study (NCCAPS)**: Prospective Study of COVID-19 in Cancer Patients
   - Target accrual of 2,000 patients with biospecimens
   - Follow-up and survivorship evaluations by early 2022

Clinical course of COVID-19 is highly variable

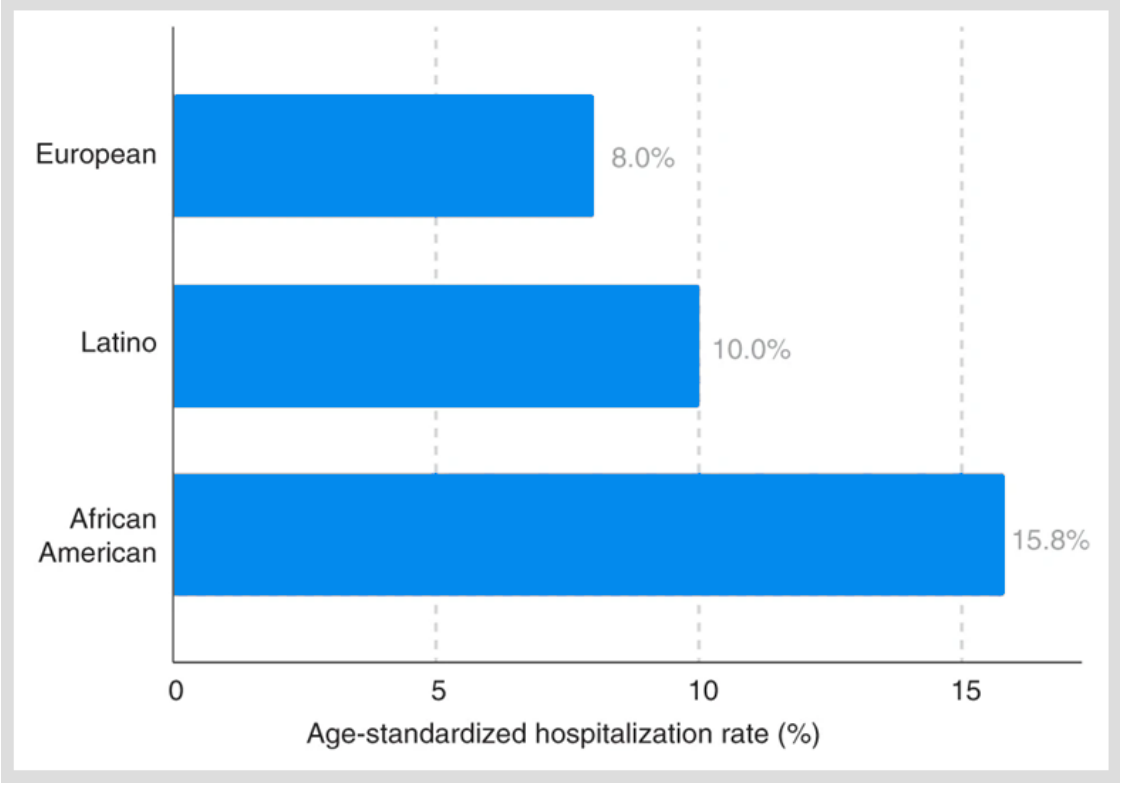


Fricke-Galindo and Falfán-Valencia, 2021.

Factors associated with the clinical course of COVID-19

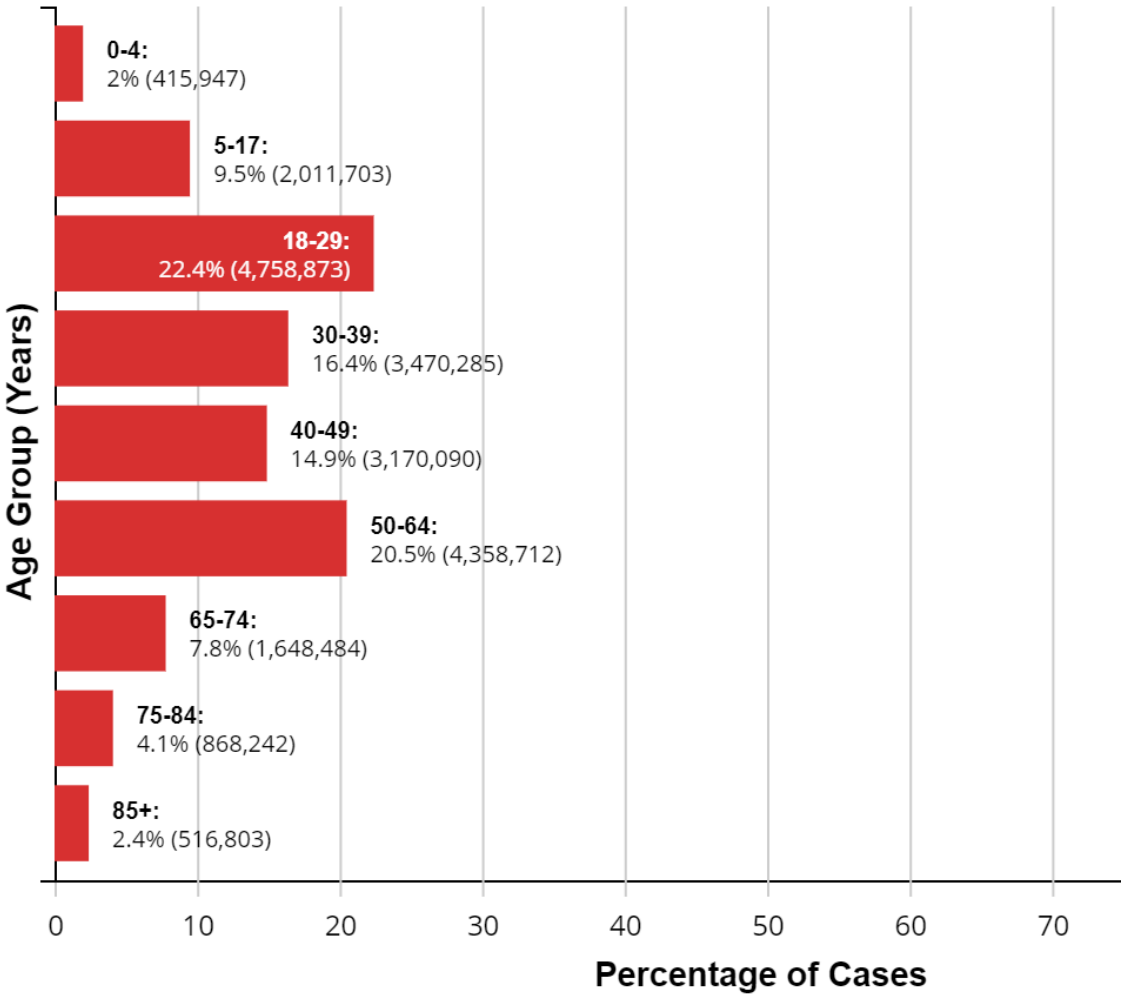


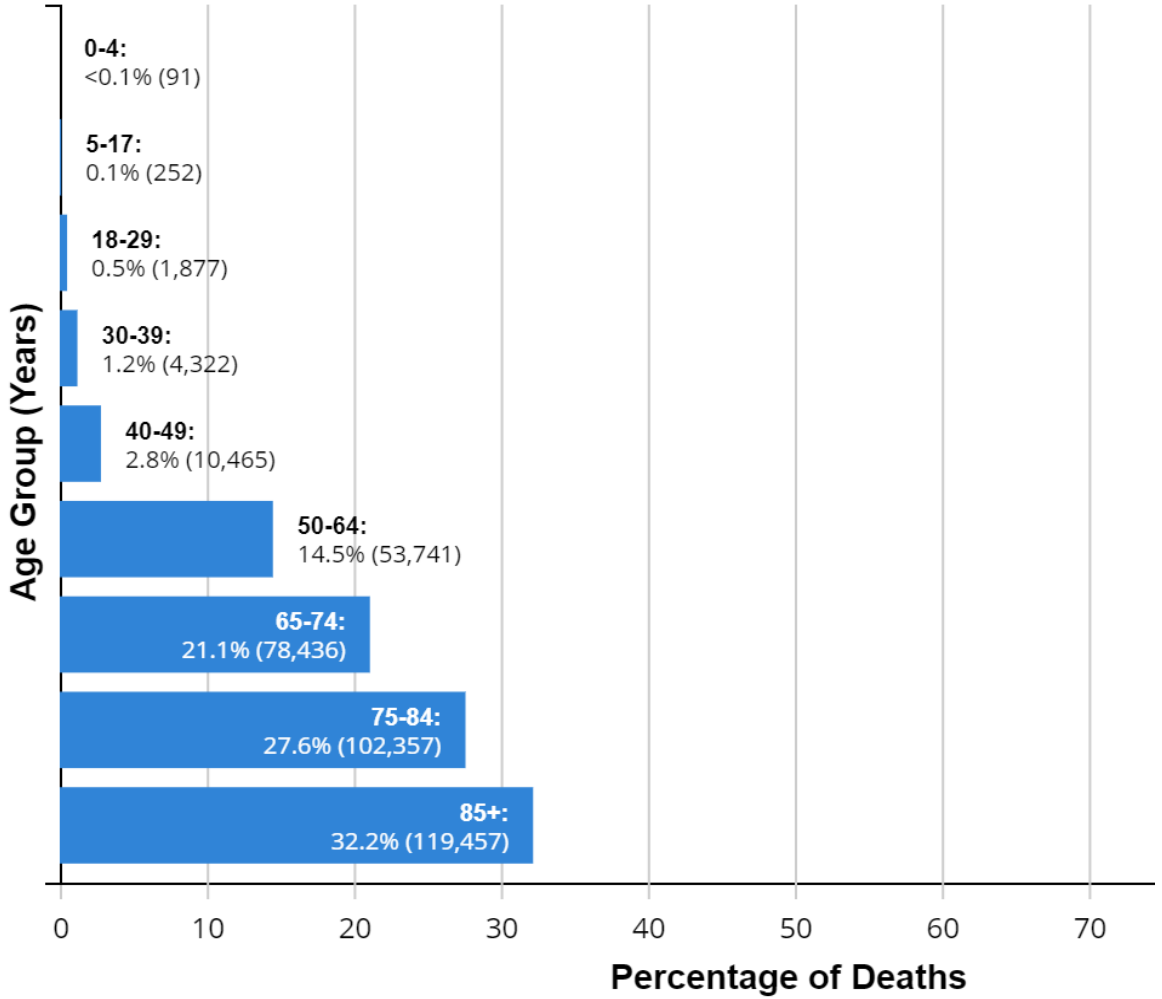
- Hospitalization and death consistently associated with gender (male), age, obesity, lower socioeconomic status, and select pre-existing health conditions*
- Non-European ancestry



Williamson et al., Nature 2021; Shelton et al., Nature Genetics 2021.

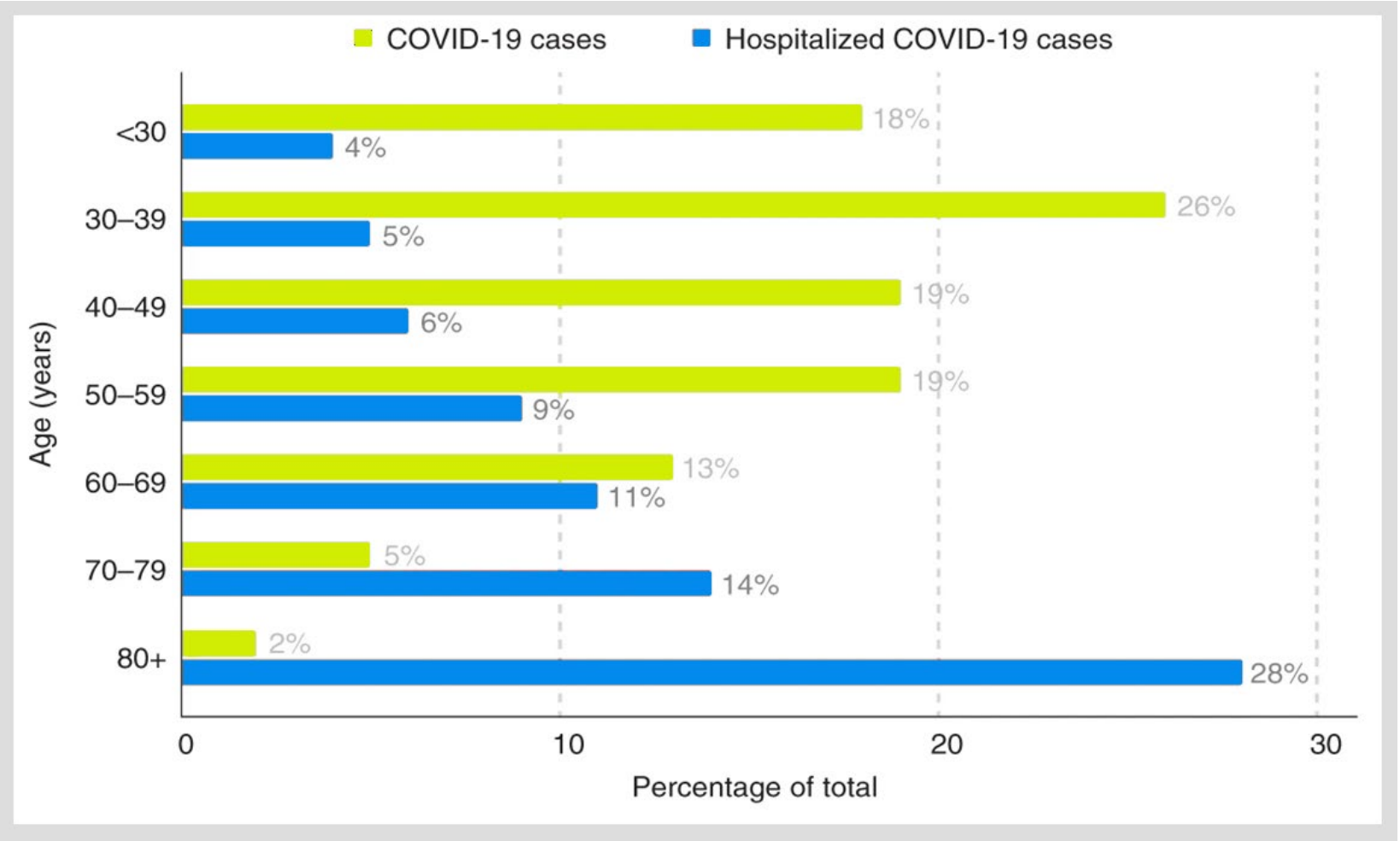
COVID-19 cases in the U.S. by age





CDC COVID Data Tracker: https://covid.cdc.gov/covid-data-tracker/#demographics

COVID-19 cases in the U.S. by age



Shelton et al., Nature Genetics 2021.

Principal Investigator: Les Biesecker, M.D., NHGRI

Associate Investigators:

NCI, DCEG

- Stephen Chanock, Sharon Savage, Lisa Mirabello, Renee Bremer, Mandy Black, Lisa McReynolds, Margarita Aryavand

NIAID

- Stephen Holland, Helen Su, Luigi Notorangelo, and others

Primary objectives

- To identify germline susceptibility variants that determine host responses to COVID-19 disease in case-case design.
- Identify common and rare germline variants associated with host susceptibility to severe or fatal COVID-19 disease.

Secondary objective

- Exploratory analyses of epigenetic signatures, serologic immune markers and antibody profiles using whole blood collected.
  - Collect *(when possible & distributed carefully):*
    - Whole blood specimens for sera, DNA, RNA
    - Explore B and T cell repertoire
    - Serum or plasma to explore humoral response and soluble mediators
    - RNA sample tubes for transcriptomic analysis

Principal Investigator: Les Biesecker, M.D., NHGRI

Associate Investigators:

NCI, DCEG

- Stephen Chanock, Sharon Savage, Lisa Mirabello, Renee Bremer, Mandy Black, Lisa McReynolds, Margarita Aryavand

NIAID

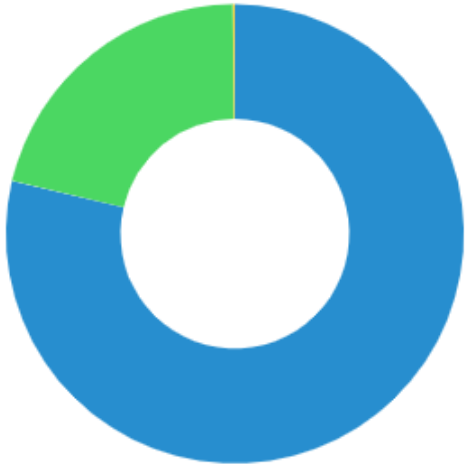
- Stephen Holland, Helen Su, Luigi Notorangelo, and others

Data collection

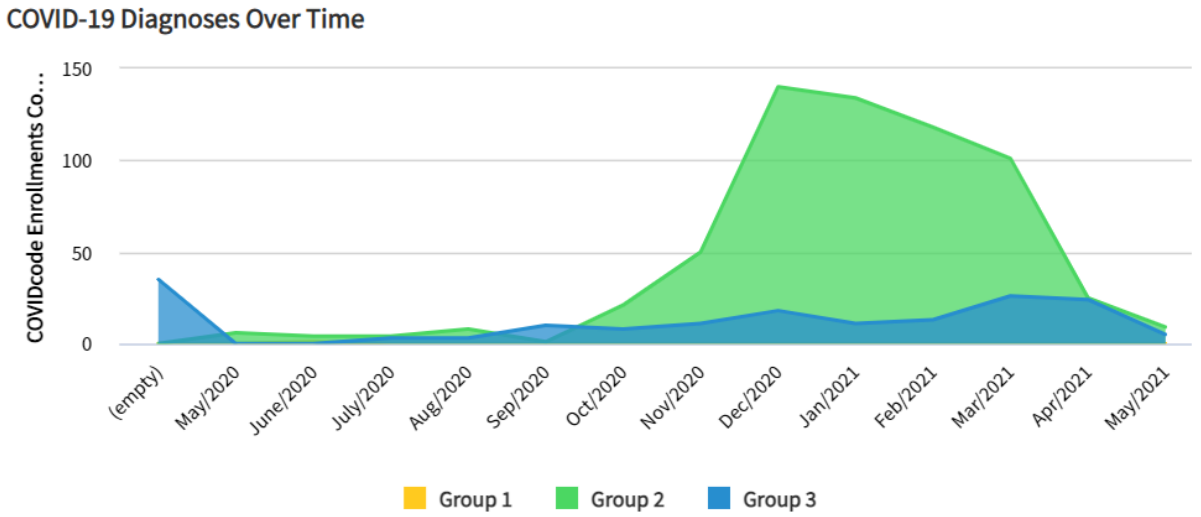
- Target COVID-19 patient recruitment: 2500
  - NIH intramural program: CC and OMS participants
  - Extramural collaborators
  - Participants can self enroll
- Uses CGB's CHARMS (communications hub and research management system)
- Collection of risk factors, rapid oral/on-line questionnaire (30 questions)
  - Extended Questionnaire including more clinical details (150 questions)

COVIDcode progress
Target: 2500

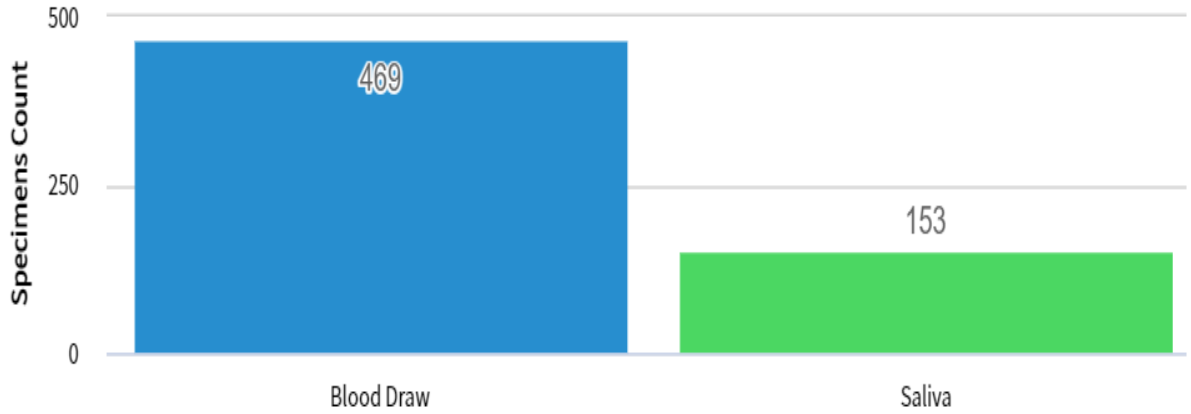
Enrollment by Group



Group 2 = 621 (78.71%) Group 3 = 167 (21.17%) Group 1 = 1 (0.13%)



Specimen Counts



NIH NATIONAL CANCER INSTITUTE

NCI COVID-19 in Cancer Patients Study (NCCAPS)

1. Cohort of cancer patients infected with COVID-19 comprising **all age groups** for collection of a comprehensive dataset on the cancers, treatments, medications, symptoms, course, and recovery, and co-morbidities with longitudinal follow-up until return to pre-morbid status;
2. Follow subset of pts for >1 yr to assess impact of COVID-19 on survivorship and cancer outcomes;
3. Collect blood samples at study entry and then every 2-3 months for 1 yr to estimate antibody and cellular immune response, genetic susceptibility, coagulation abnormalities, and for biomarker development;
4. Collect imaging and QOL data longitudinally;
5. Public database/biospecimens.

NCI COVID-19 in Cancer Patients Study (NCCAPS)

Critical Study Milestones

- **Study opened in late May 2020**
  - 6 weeks from idea to patient entry
- Enroll the first 500 patients within 3 months of trial activation
- Target accrual of 2,000 patients with biospecimens @ Nationwide
- Follow-up and survivorship evaluations by early 2022
- Begin biomarker studies on blood samples soon after initial 500 patients accrued—supported by new Congressional appropriation
- GWAS and NGS on all 2,000 to be done through CGR/DCEG

Breakdown of NCCAPS, May 2021

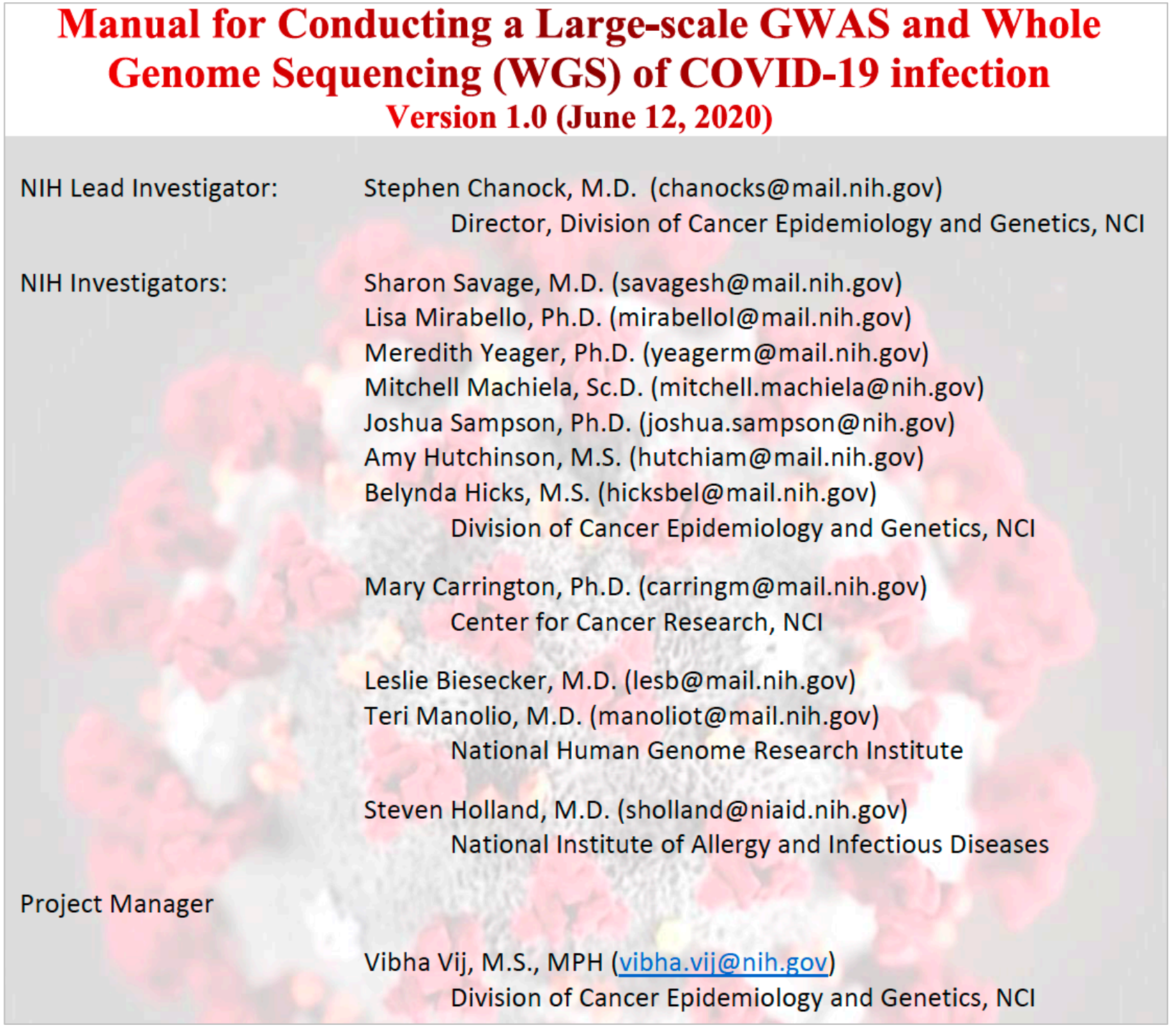
	Not Hispanic or Latino	Hispanic or Latino	Not Reported / Unknown	Total
American Indian or Alaskan Native	8			8
Asian	16			16
Black or African American	123	2	2	127
Native Hawaiian or Pacific Islander	5			5
White	820	72	10	902
Not Reported/Unknown	18	45	6	69
Total	990	119	18	**1,127**

NCI COVID-19 in Cancer Patients Study (NCCAPS)

- NCCAPS pediatric amendment to specifically accrue COVID-19 pediatric patients without the requirement of longitudinal follow-up.
  - Maximize accrual to determine the consequences of COVID-19 on kids with cancer

Germline genetics of COVID-19 susceptibility and manifestations

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



Status of COVNET, May 2021

- Over 150 programs/Geneticists in US approached
  - >100 Teleconferences
- MTAs (marker of commitment in US)
  - 20 Signed
  - 5 in process
- Sampled received (as of 5/19/2021)
  - ~9,000 (15 studies)
  - ~4500 expected in coming weeks
  - ~4000 longer term
- Genotyped
  - 3,764 (phenotypes received for 2,654)

COVNET-Current Collaborators

International Collaborators

Innsbruck Medical University, Austria

Seoul National University Hospital, Korea

Academy of Athens, Greece

Koc University, Istanbul, Turkey

NRCRM, Kiev, Ukraine

Latin American countries (next slide)

COVNET Latin America

Large-scale genome-wide association study and whole genome sequencing of COVID-19 severity

May 2021

LATIN AMERICA - COLLABORATORS

Country	Investigator	Planned sample size	Current samples	MTA	Updates
Chile/Argentina/ Guatemala/Colombia	Luis A. Quiñones, Matias Olguin	1,000	700	Executed	**700 done + 300 from Guatemala/Argentina/South Chile**
Chile	Catterina Ferreccio, Vanessa Van de Wyngard	1,500	600	Executed	**448 done + 900 to DNA extraction**
Colombia/Venezuela	Bladimiro R Orozco	158	158	In progress	**158 done**
Chile	Alvaro Cerda, Monica Aguilar	300	100	Executed	**100 done + 200 to be collected**
Peru	Meddly Santolalla	1,100	700	In progress	**DNA extraction kits Children? Blood?**
Brazil	Eduardo Tarazona, Maria Cássia, Leandro Colli	400	400	In progress	**207 done +193 DNA extraction**
Total expected		**4,458**			

NIH NATIONAL CANCER INSTITUTE

Distinctive Opportunities in COVNET

- First look at susceptibility to SARS-CoV2
  - CDC Studies x 2
  - Austrian Isghl Study
- Special Populations
  - Liquidator Cohort in Ukraine (post Chernobyl)
  - Long term consequences of radiation exposure
- Population Genetics of Latin America

COVNET: Key principles

- Emphasis on harmonization of:
  - Phenotype definitions
  - Genotype analyses
  - Meta-analyses
- Aligned with COVID-19 Host Genetics Initiative: https://www.covid19hg.org
  - Importance of phenotype data (lagging behind)
- Emphasis on disparities
- Data sharing is central to identify and validate risk alleles
  - Hard to share until we have adequate phenotype data

Sample requirements and processing

Amy Hutchinson

Director of Operations, CGR

Sample Requirements Requested by COVNET (pre WGS)

DNA

Mass: 1.0 – 1.5ug (no WGS)

Mass: 2.5 – 3.5 ug (with WGS)*

Volume: ≥ 30ul

Sample Kit use is encouraged.

Blood

Whole Blood, Buffy Coat, or PBMCs

Volume: ≥ 150ul

Buccal

Oragene, Mouthwash, or Saliva

Volume: ≥ 1000ul

*** Currently exploring low input options for WGS**

Sample Processing

Samples are processed in a highly-automated, high-throughput laboratory environment following established SOPs

Laboratory

Tracking via a highly-integrated, customized Laboratory Information Management System (LIMS)

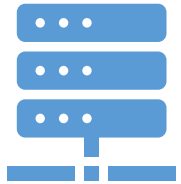
Extraction

KingFisher Flex

DNA QC

Volume verification

PicoGreen quantification



Genotyping

Standard input = 200ng

Minimum input = 50ng

1-3% QC replicates

Approved/consented studies only with phenodata

Whole Genome Sequencing

Standard input = 1200ng

Exploring option of low input WGS @ 300ng

Clifton Dalgard
The American Genome Center

Samples Genotyped

Institute	*Unique subjects*	*Subjects with phenotype classification*	*Samples genotyped (includes replicates)*	*Expected replicates (samples with lower call rate excluded)*	*Sex-discordant*	*Contaminated*	*Low call rate*	*Unexpected Replicates*
Athens	306	305	315	9	1	1	1	8
Austria	342	NA	346	4	7	3	4	0
Carrington/Harvard	495	433	500	5	8	1	0	26
COVIDcode	45	40	47	2	NA	0	0	0
KP - Colorado	549	443	554	5	0	3	1	0
MSKCC	388	NA	391	3	2	0	0	4
NCI/CCR	59	55	58	2	1	0	0	2
NIAID*	392	378	396	4	1	0	0	20
Northwestern	86	50	88	2	0	0	0	0
South Korea	105	82	109	4	0	0	0	0
UAB	576	561	592	16	6	3	2	2
UCSD	3	3	4	1	2	0	0	0
UPenn	358	304	364	10	2	0	0	90*
Total	**3,704**	**2,654**	**3,764**	**67**	**30**	**11**	**8**	**152**

* Includes provided WGS data

Post-Genotyping Data QC for GWAS on GSA v2

- Sample-level QC
  - Array processing: remove samples that fail to generate valid idat/gtc files
  - Completion rate: cutoff 0.8 for samples and 0.8 for loci, followed by 0.95 for samples and 0.95 for loci
  - Sample contamination: exclude samples with >10% contamination as predicted by VerifyIDintensity
  - Expected replicate removal: include replicate with higher call rate

- Subject-level QC
  - Sex verification: reported vs. observed sex based on chrX method-of-moments F coefficient. Cutoff at 0.5 (expected to be 0.0 for females and 1.0 for males)
  - Unexpected replicates: phenotypes are assessed before filtering unexpected replicates

Data Sharing Plan

- Following primary data QC, a delivery package will be provided to each site via NIH secured transfer on the Box platform.
- Data access is restricted to you (and those you designate) plus designated NCI staff only.
- Delivery package includes:
  - **QC Reports (.xls and .doc)** – These reports are generated from the genotyping batch of which your samples were a part. You will therefore notice additional samples included in both reports, this is expected and samples that are not your own can be ignored (no proprietary information is provided here).
  - **Called Genotypes** – Called genotypes for your subjects from CGR's pipeline in standard PLINK format.
  - **Sample ID Linkage File** – This is the master sample ID file and provides a listing of the various IDs contained in the reports and genotyping files for your samples only. This provides the link between IDs we generated internally as part of our laboratory pipelines and the IDs you provided with your samples. Any samples excluded from genotyping are listed at the top of the spreadsheet with a comment regarding the reason for exclusion.
  - **Illumina Infinium Genotyping_V3** – This is CGR's standard material and methods (M&M) document for Illumina genotyping and describes genotyping and data QC workflows currently in place. Specific details related to data QC of this project are included in the QC Report.

Analyses and preliminary data: GWAS

Lisa Mirabello

Senior Investigator, DCEG

Meredith Yeager

Scientific Director, CGR

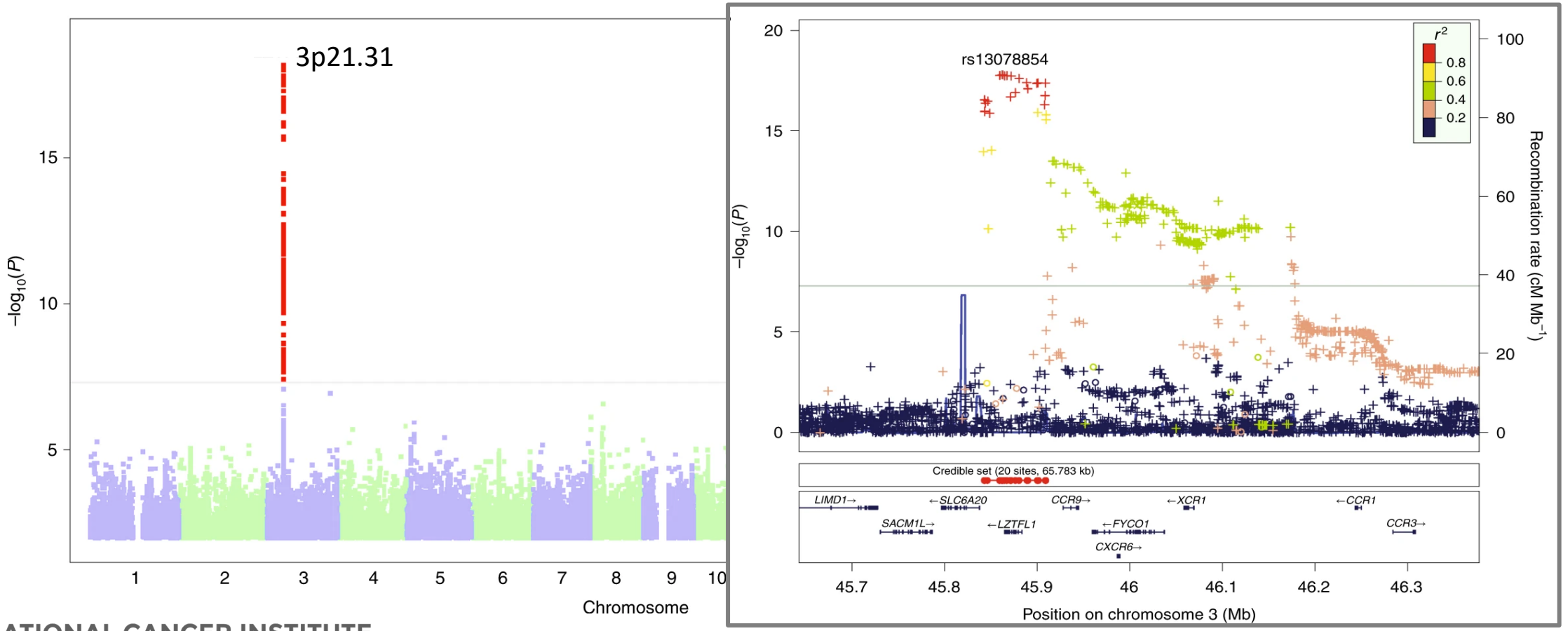
Published germline genetics of severe COVID-19 disease

- Several studies suggest an underlying genetic component to severe disease
- One genetic locus—on chromosome 3p21.31—has been repeatedly associated with hospitalization [1-3]
- *ABO* locus has been reported as a risk factor for both COVID-19 susceptibility [3] and severity [1] in some studies but not in others [2]
- Recent GWAS identified 3 new loci on: chromosome 12q24.13 (*OAS1-3*); chromosome 19p13.3 (*DPP9*); and, chromosome 21q22.1 (*IFNAR2*) [2]
- Compared patients with severe COVID-19 disease to the general population
- Rare variants in IFN genes associated with severe disease [4]

[1] The Severe Covid-19 GWAS Group. N. Engl. J. Med. 2020; [2] Pairo-Castineira et al., Nature 2020; [3] Shelton et al., Nature Genetics 2021; [4] Zhang et al., Science 2020.

Published germline genetics of severe COVID-19 disease

- 23andMe research cohort: 15,434 COVID-19 patients
- Chromosome 3p21.31 contains multiple genes that could be functionally implicated in COVID-19 pathology: *SLC6A20*, *LZFTL1*, *CCR9*, *CXCR6*, *XCR1*, *FYCO1*



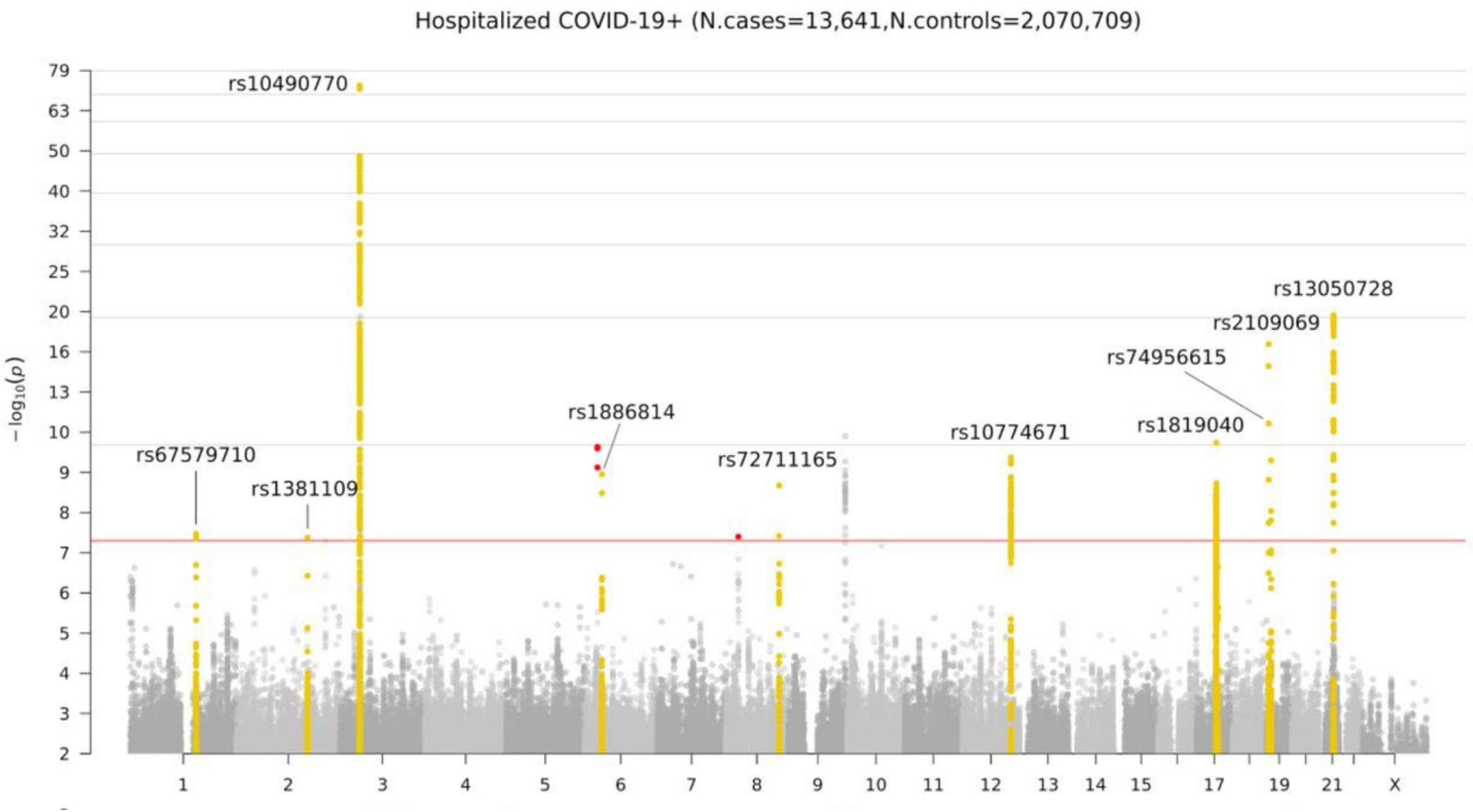
Shelton et al., Nature Genetics 2021.

Published germline genetics of severe COVID-19 disease

- COVID-19 Host Genetics Initiative: 3 case-control GWAS including 49,562 COVID-19 patients from 46 studies
    1. Critically ill cases requiring respiratory support in hospital or who were deceased, N=6,179
    2. Cases with moderate or severe COVID-19 defined as those hospitalized, N=13,641
    3. All cases with reported SARS-CoV-2 infection with or without any symptoms, N=49,562
- Each compared to ~2M population controls: genetically ancestry-matched without SARS-CoV-2

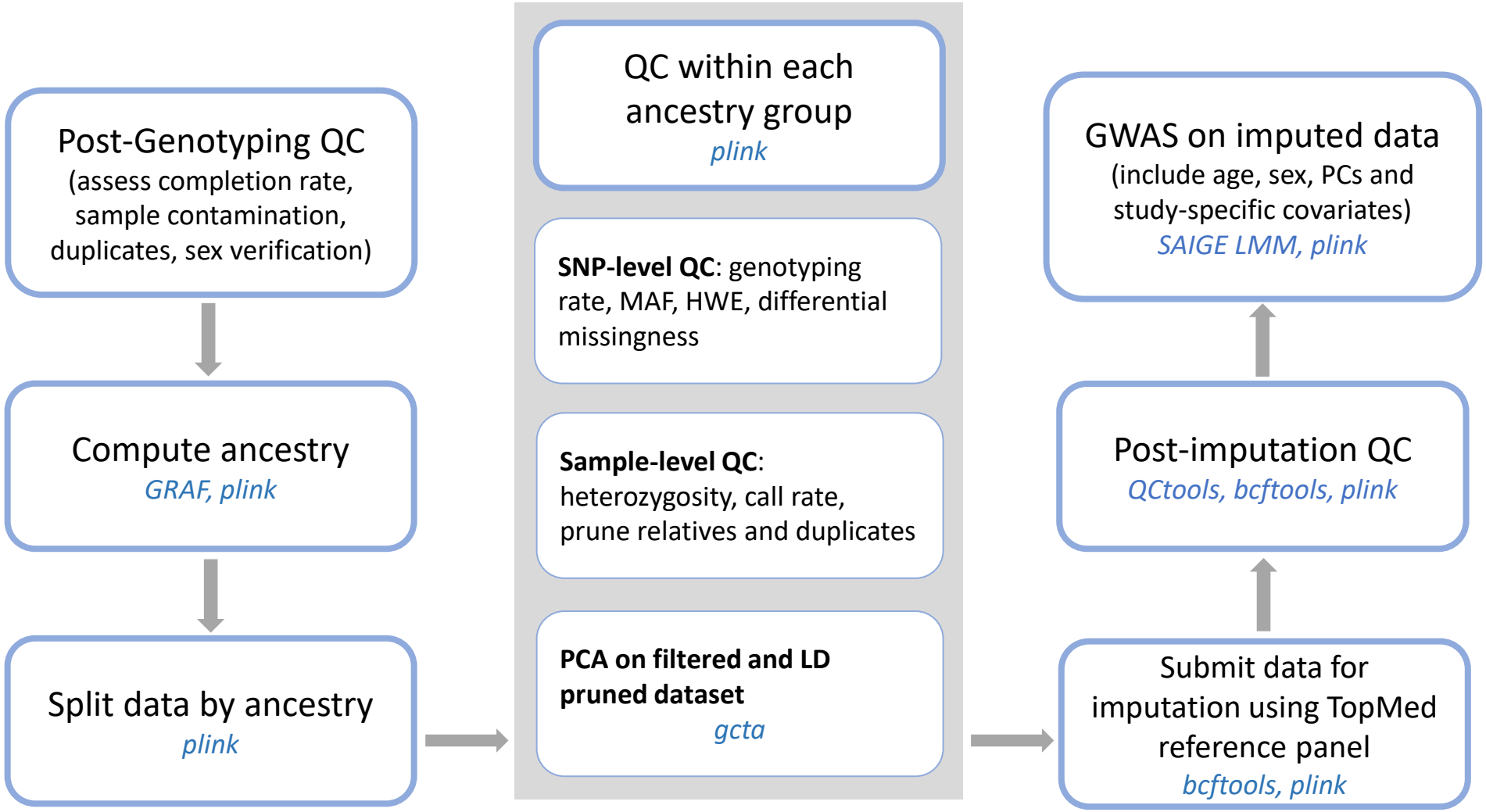
COVID-19 Host Genetics Initiative, medRxiv preprint

Published germline genetics of severe COVID-19 disease



COVID-19 Host Genetics Initiative, medRxiv preprint

GWAS Analysis plan



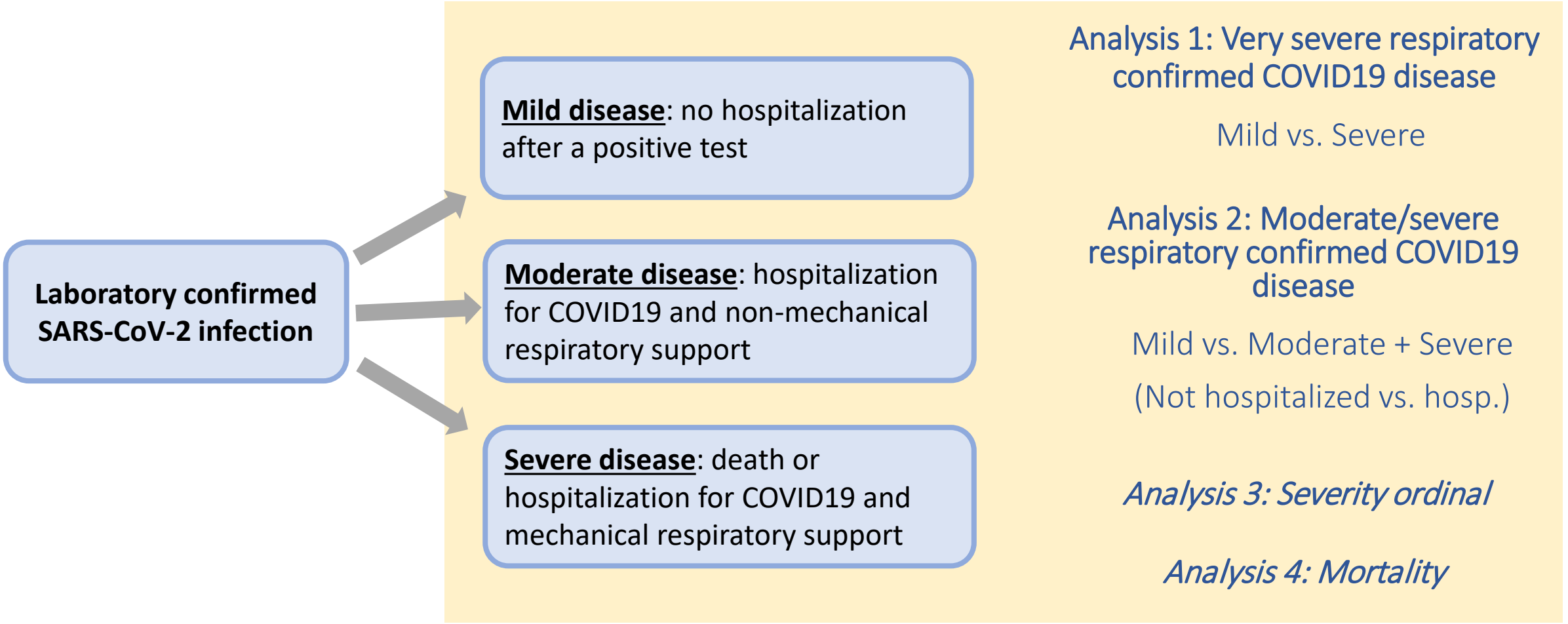
Adapted from Pairo-Castineira et al., *Nature* 2020.

Samples received: Phenotype data required for analyses

Samples	N	% of total
Total received	8,898	
Total genotyped	3,704	
Total with phenotype	2,654	
Unable to classify disease	*213*	*8%*
Missing covariate data (eg, smoking information)	*1,351*	*51%*
Missing co-morbidity data	*2,032*	*77%*

- Phenotype data needed with requested covariates using our provided template in order to harmonize disease severity classification across studies

Analysis Plan: Primary association analyses for disease severity

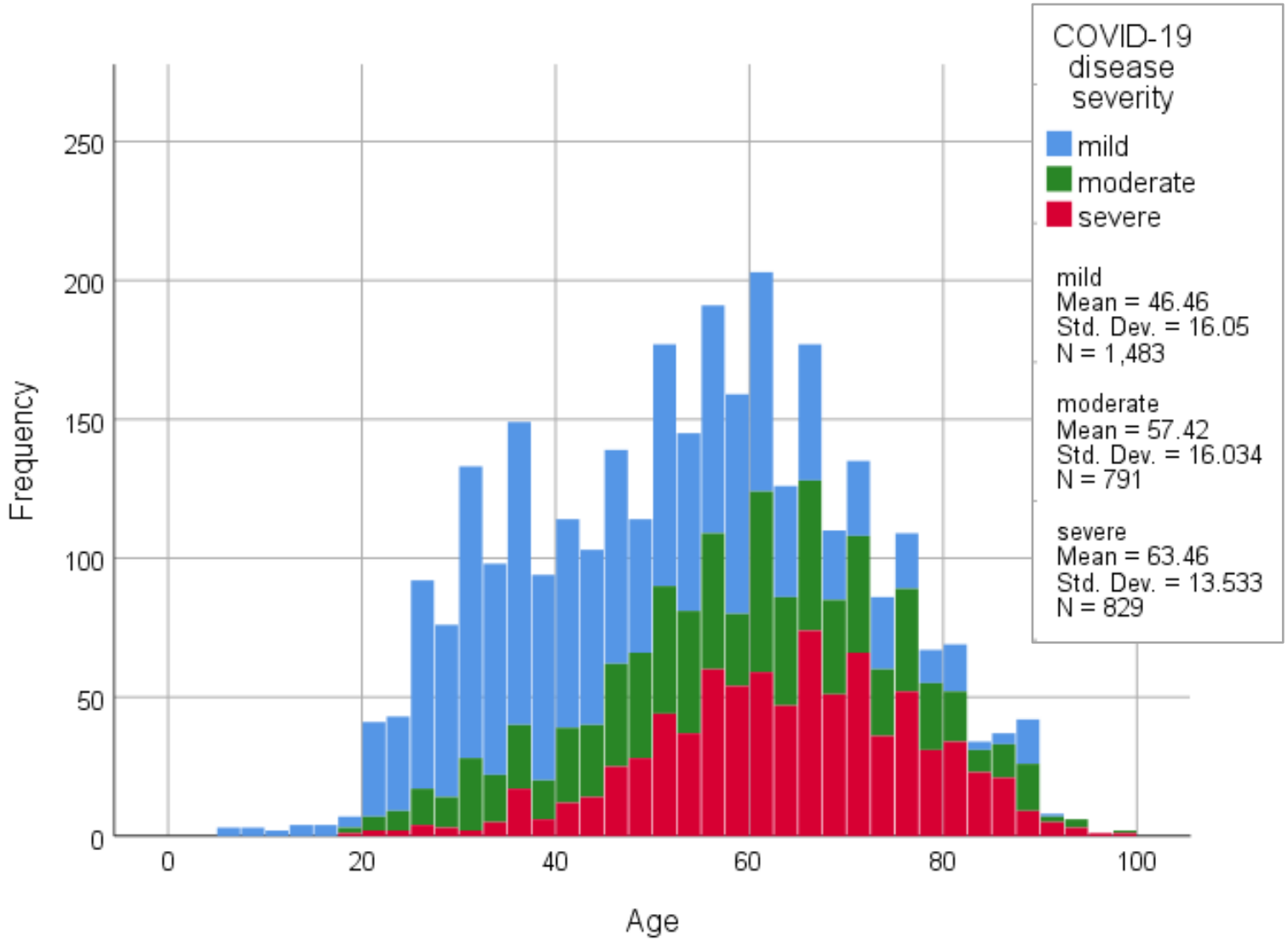


For EUR ancestry individuals: generalized linear mixed models to evaluate imputed SNP associations for the specified outcomes; models adjusted for PCs, sex, age, and appropriate covariates.

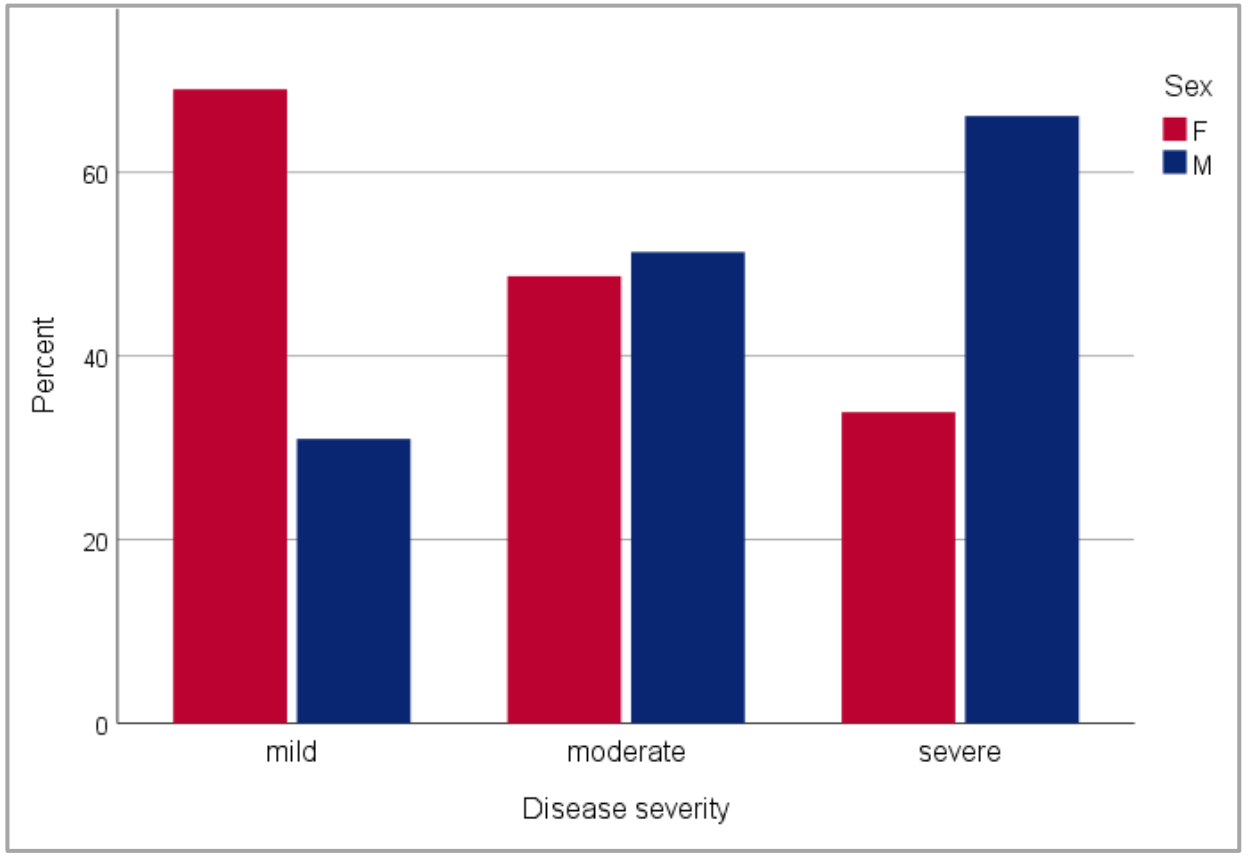
COVNET COVID-19 cases with phenotype data

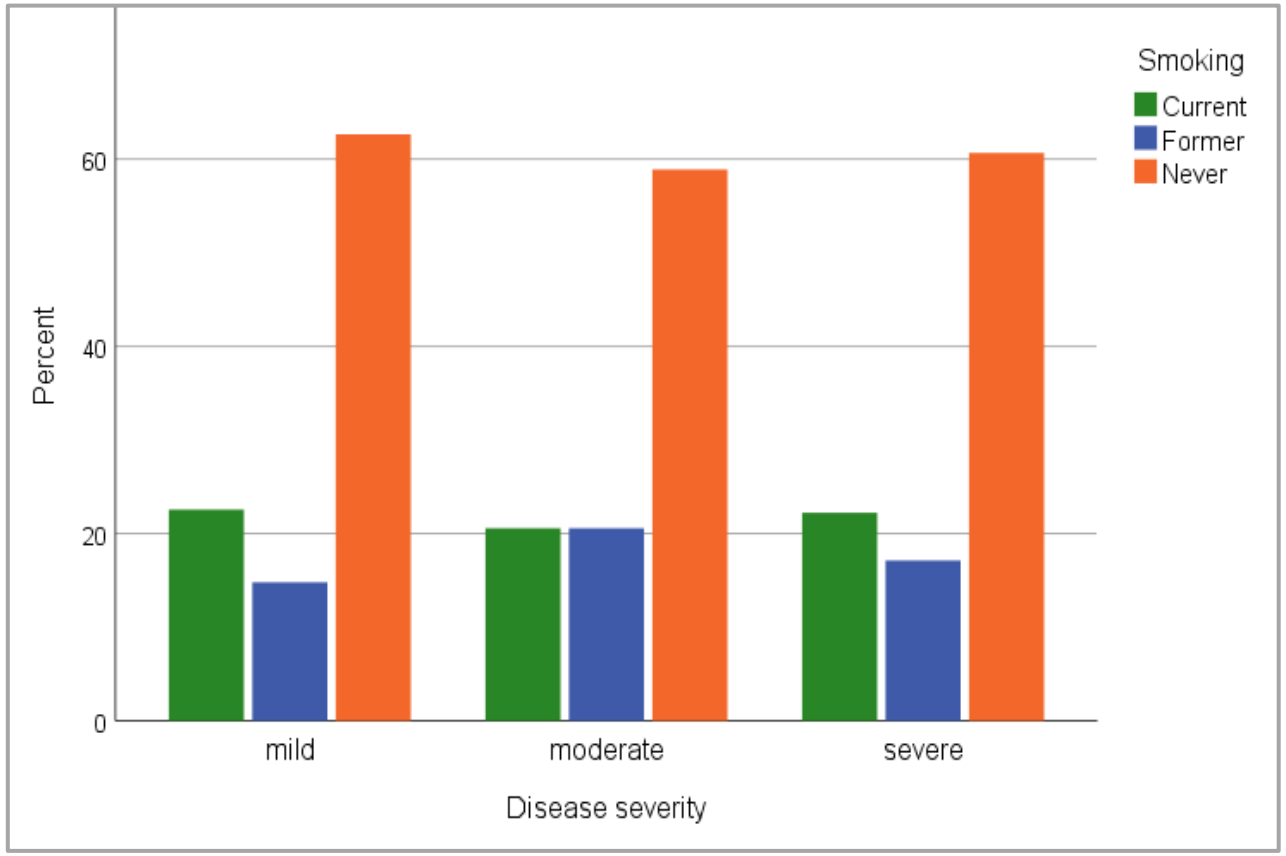
COVID patients		N	Column %
Disease			
	mild	1486	47.8%
	moderate	792	25.5%
	severe	830	26.7%
	*missing**	*213*	
Sex			
	male	1382	44.5%
	female	1726	55.5%
Vital status			
	alive	1132	82.7%
	deceased	236	17.3%
	missing	*1740*	
Smoking			
	current	400	22.1%
	former	298	16.4%
	never	1116	61.5%
	missing	*1351*	

* Severity could not be assigned from phenotype data



COVNET COVID-19 cases with phenotype data

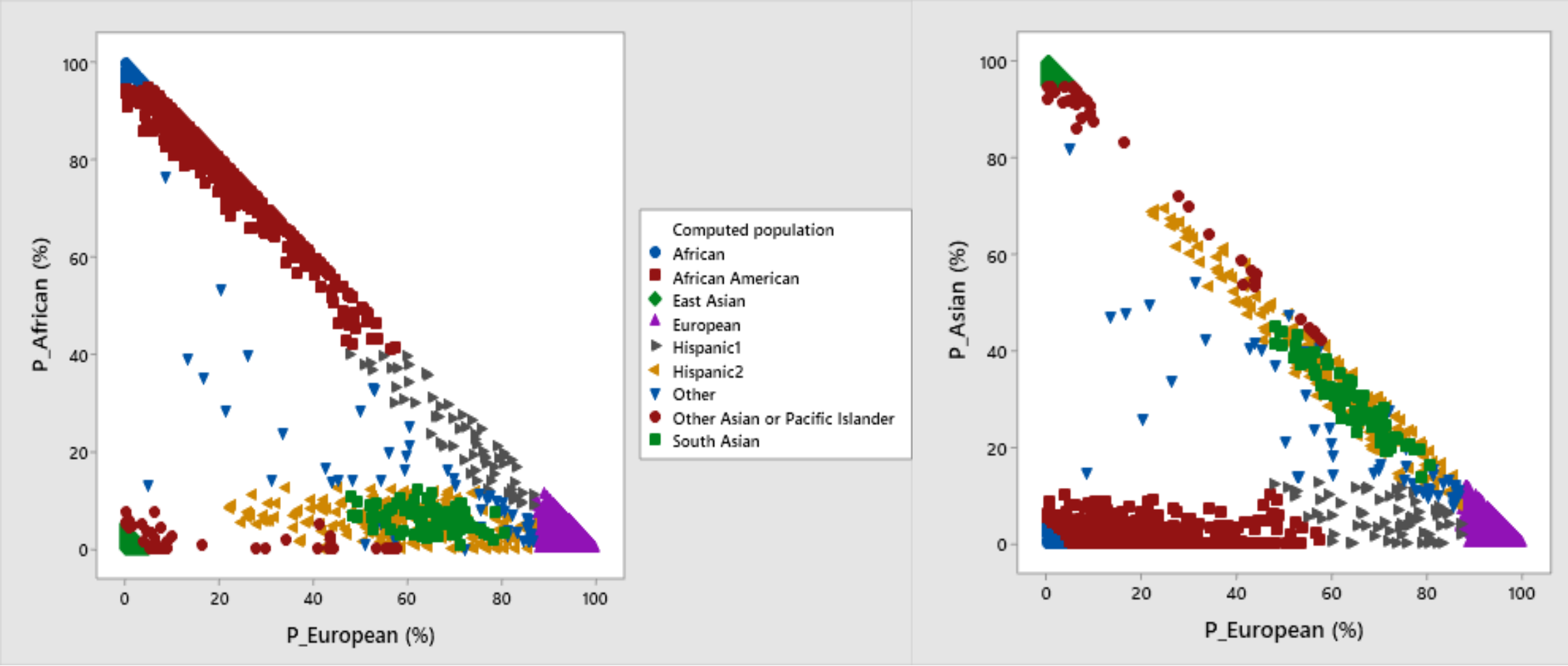


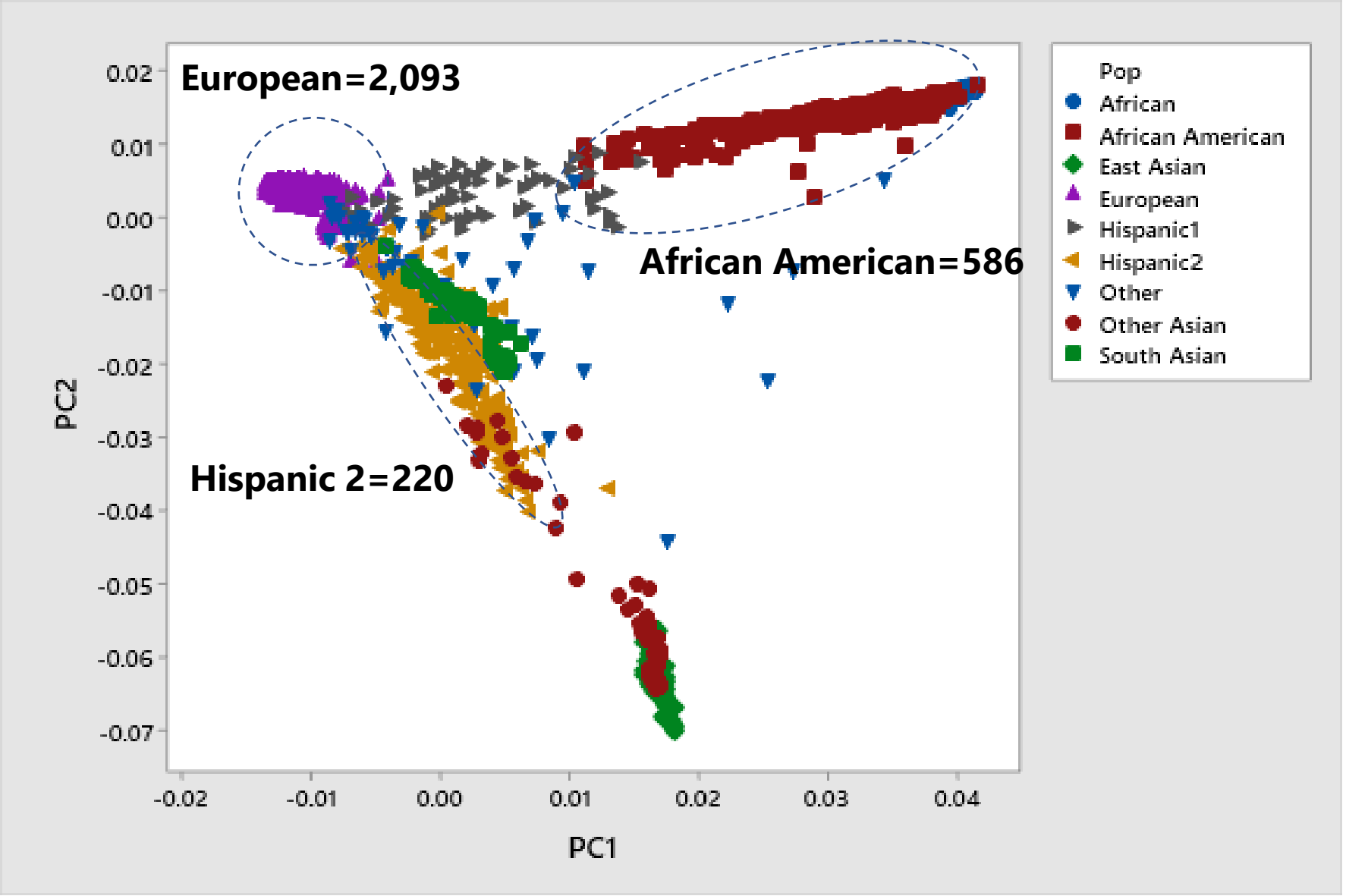


Ancestry computation with GRAF

- Assumes that each subject is an admixture of three ancestries:

  European Asian African
- Estimates proportions of the 3 ancestries and assigns each subject to a population





- PCA performed on filtered (MAF>0.1, HWE p-value>10^{-6}, missingness<1%) and LD-pruned SNPs (1000 markers, a step size of 80 markers and an r2 threshold of 0.1)

Analysis of EUR samples with COVID disease severity classification available (N=1,611 unique subjects)

Institute	Samples genotyped (includes replicates)	EUR ancestry	*EUR: Sample-level exclusions*		Final EUR sample counts
			Missing phenotype data/classification	*QC: kinship, duplicates, sex-discordant*	
Athens	315	257	*7*	*5*	250
Austria	346	129	*129*		
Carrington Lab	500	331	*50*	*16*	281
COVIDcode	47	18	*5*	*2*	13
KP - Colorado	554	268	*0*	*2*	268
MSKCC	391	234	*234*		
NCI/CCR	58	40	*1*	*3*	39
NIAID	396	351	*14*	*15*	337
Northwestern	88	58	*31*	*0*	27
South Korea	109	1	*0*	*0*	1
UAB	592	317	*11*	*16*	306
UCSD	4	0	*0*		
UPenn	364	89	*0*	*7*	89
Total	**3,764**	**2,093**	***482***	***66***	**1,611**

Analysis of EUR samples with COVID disease severity classification available (N=1,611 unique subjects)

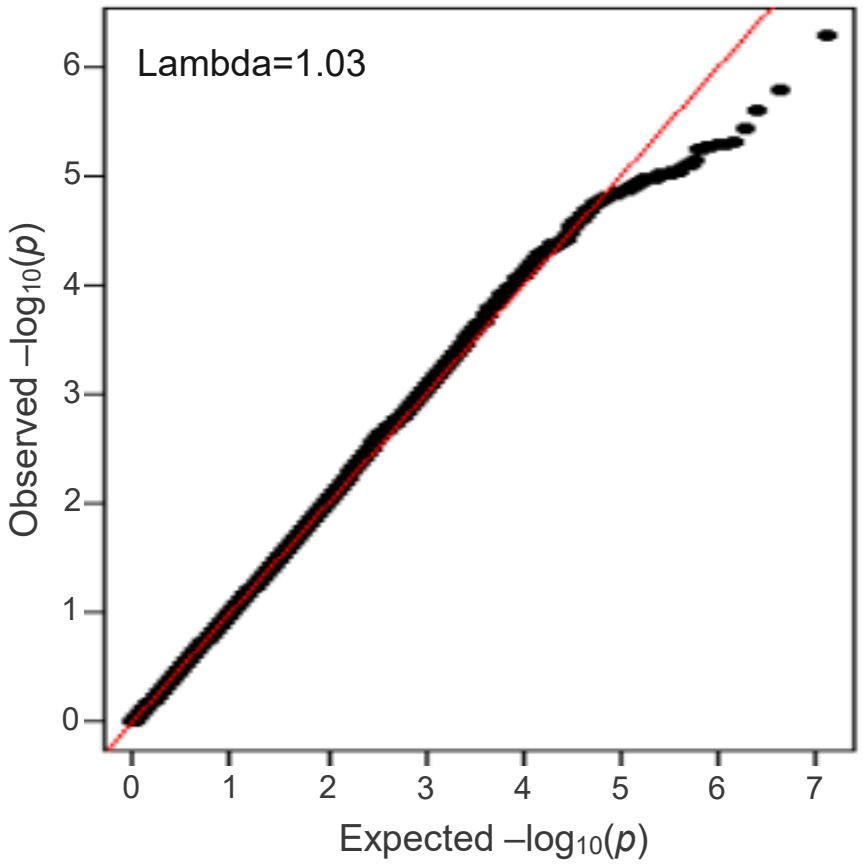
SNP-level QC

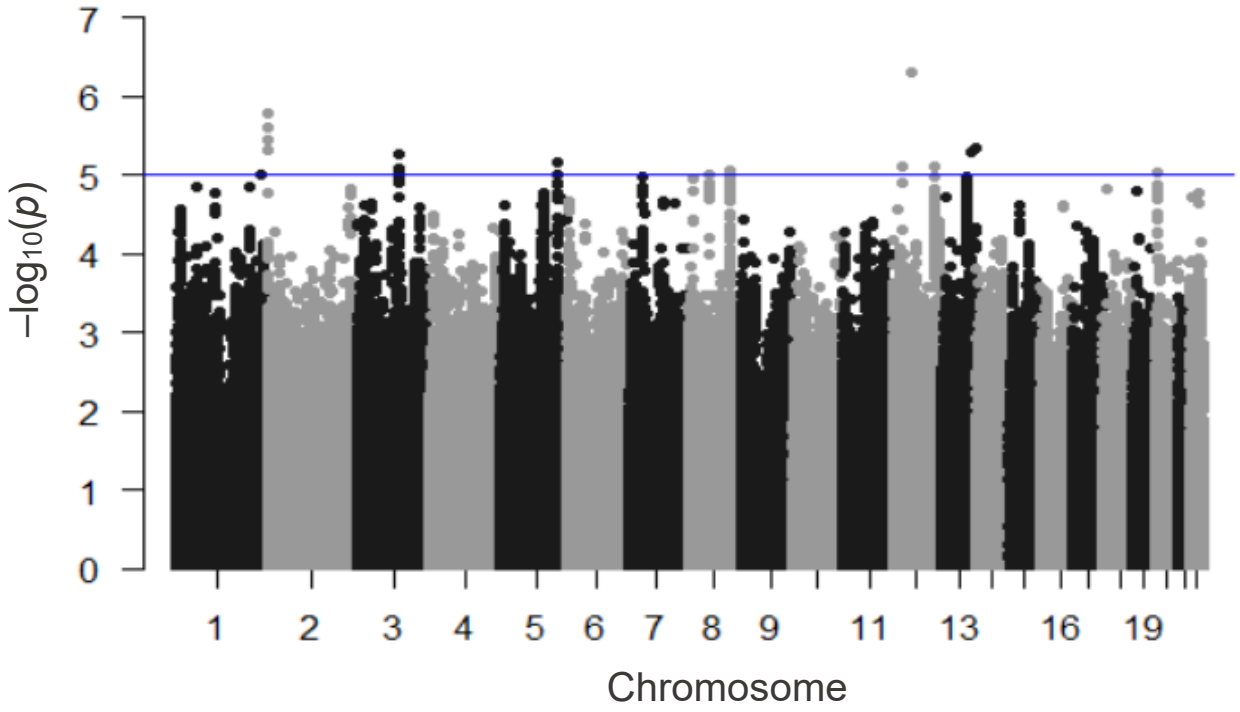
- Select variants with call rate > 99%, minor allele frequency (MAF) > 1%, Hardy-Weinberg equilibrium (HWE) p-value > 10^{-6}
- Exclude SNPs with significant differential missingness (P<0.001) across disease states

- Compute PCs with LD-pruned SNPs using gcta1.9
- Submit QC'd dataset for imputation using TopMed reference panel
- Download and QC imputed data (imputation quality score > 0.9, ~6.8M SNPs remain) and convert to BGEN format

Case-Case GWAS: mild vs. moderate + severe European ancestry cases

- 614 mild cases, 997 moderate/severe cases
- **Top SNP**: chr12, 12q13.12; p-value = 5.09e-07
  - Covariates: age, sex, PC1-10





WGS selection and data sharing

Stephen Chanock

Principles for Whole Genome Sequencing (30X)

- Target: 15-20% of each COVNET study
- All of NCCAPS (extra funds)
- Larger fraction of CovidCode (extra funds)
- Selection of samples
  - NCI COVNET team will work with each study
    - Must meet requirements for adequate DNA and phenotypes
  - Highest priorities
    - Extreme phenotypes (mortality as well as survival with many comorbidities)
    - Population genetics
- Agreement (@ reagent cost)
  - American Genome Center at Uniformed Services University
  - HudsonAlpha Institute for Biotechnology with UAB samples only

Whole Genome Sequencing

- Collaboration with The American Genome Center (TAGC) at Uniformed Services University (USU)
- Currently Pilot 384 subject in progress
- Selection for WGS requires proper study approvals/consent **AND** completed phenotype data delivery

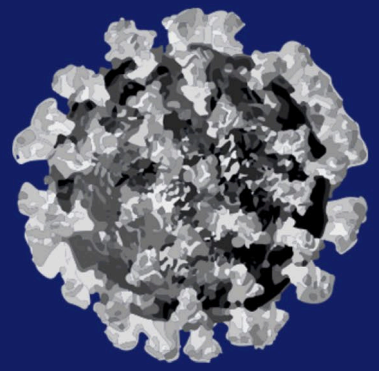
Institute	Subjects
Carrington Lab	1
NCI/CCR	45
COVIDcode	32
Northwestern	51
South Korea	3
UAB	10
UPenn	242
Grand Total	**384**

Whole Genome Sequencing: batch 1

WGS selection criteria	Subjects
Phenotype extremes	40
aged <50years and severe disease	8
aged >65years and mild disease	32
Diverse population (AfrAm, His, Asn)	214
EUR, disease groups for comparison	125
Duplicates	*5*
Total	**384**

Home

https://www.covid19hg.org



The COVID-19 Host Genetics Initiative

Aims

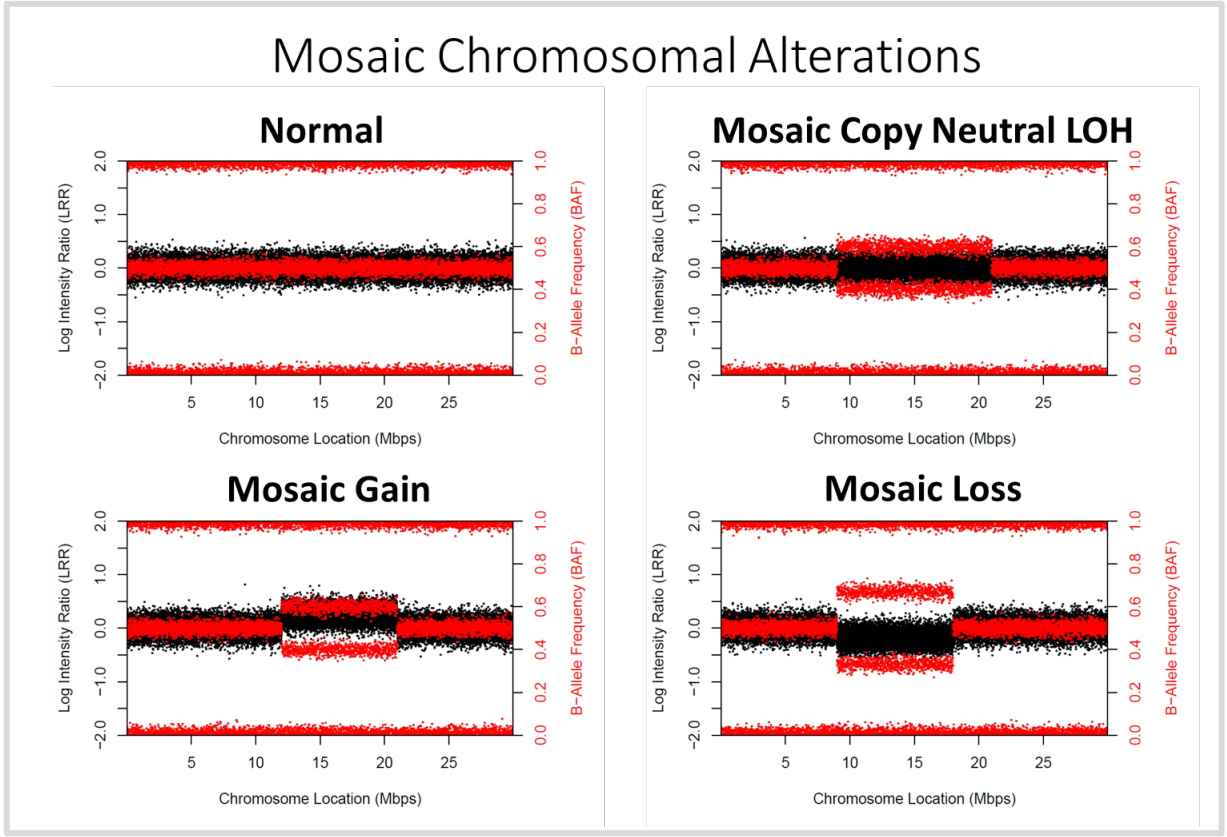
The COVID-19 host genetics initiative is a bottom-up collaborative effort that has three main goals:

1. Provide an environment to foster the sharing of resources to facilitate COVID-19 host genetics research (e.g. protocols, questionnaires).
2. Organize analytical activities across studies to identify genetic determinants of COVID-19 susceptibility and severity.
3. Provide a platform to share the results from such activities, as well as the individual-level data where possible, to benefit the broader scientific community.

Secondary analyses with COVNET genotype data

- Use the genotyped SNPs within chromosome 6 for HLA imputation to determine if specific HLA alleles are associated with COVID-19 disease severity

- Evaluate large structural copy number alterations (i.e., clonal mosaicism)
  - Acquired somatic mutations (mosaic gain, loss, or copy neutral LOH) in a clonal subset of cells that differs from the inherited germline genome



Next Major Steps

- Increase Accrual- target 40,000 for GWAS
- Selecting NGS/WGS for American Genome Center (USUHS)
  - 4-5,000 (COVNET)
    - Selection- extreme phenotypes, URM and population genetics
  - 2,000 (NCCAPS)
- Data sharing with dbGap/AnVil
- COVID Human Genetics Initiative

Important Take Home Points

- Critical need for providing *minimal phenotype data*- focus on acute COVID19 outcomes
  - Major impediment to analyses
- Potential for studying 'Long-COVID19' when follow-up data is available
  - Challenge of defining phenotype
  - NIH Strategic Initiative
- Large capacity for GWAS and NGS
  - Targets
    - 40,000 GWAS
    - 5,000 NGS/WGS

QUESTIONS?

NATIONAL CANCER INSTITUTE

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

www.cancer.gov

www.cancer.gov/espanol