

# Deconvoluting complex correlates of COVID19 severity with local ancestry inference and viral phylodynamics: Results of a multiomic pandemic tracking strategy

Victoria N. Parikh, MD<sup>1\*</sup>; Alexander G. Ioannidis, PhD<sup>2,3\*</sup>; David Jimenez-Morales, PhD<sup>1</sup>; John E. Gorzynski, DVM<sup>1,4</sup>; Hannah N. De Jong, BS<sup>1,4</sup>; Xiran Liu, MS<sup>3</sup>; Jonasel Roque, BS<sup>1</sup>; Victoria P. Cepeda-Espinoza, PhD<sup>2</sup>; Kazutoyo Osoegawa, PhD<sup>5</sup>; Chris Hughes, BS<sup>1,4</sup>; Shirley C. Sutton, MS<sup>1,4</sup>; Nathan Younton, BS<sup>1,4</sup>; Ruchi Joshi, MS<sup>1</sup>; David Amar, PhD<sup>1</sup>; Yosuke Tanigawa, PhD<sup>2</sup>; Douglas Russo, MS<sup>6</sup>; Justin Wong, MS<sup>6</sup>; Jessie T. Lauzon, MS<sup>7</sup>; Jacob Edelson, MS<sup>2</sup>; Daniel Mas Montserrat, PhD<sup>2</sup>; Yongchan Kwon, PhD<sup>2</sup>; Simone Rubinacci, PhD<sup>8</sup>; Olivier Delaneau, PhD<sup>8</sup>; Lorenzo Cappello, PhD<sup>6</sup>; Jaehee Kim, PhD<sup>9</sup>; Massa J. Shoura, PhD<sup>4,10</sup>; Archana N. Raja, PhD<sup>1</sup>; Nathaniel Watson<sup>10</sup>; Nathan Hammond<sup>10</sup>; Elizabeth Spiteri PhD<sup>10</sup>; Kalyan C. Mallempati<sup>5</sup>; Gonzalo Montero-Martín, PhD<sup>5</sup>; Jeffrey Christle, PhD<sup>1</sup>; Jennifer Kim, BS<sup>1</sup>; Anna Kirillova, BS<sup>11</sup>; Kinya Seo, PhD<sup>1</sup>; Yong Huang, MD<sup>1</sup>; Chunli Zhao, PhD<sup>1</sup>; Sonia Moreno-Grau PhD<sup>2</sup>; Steven G. Hershman, MS, MSE, PhD<sup>1</sup>; Karen P. Dalton<sup>1</sup>; Jimmy Zhen<sup>1</sup>; Jack Kamm, PhD<sup>12</sup>; Karan D. Bhatt, MS<sup>12</sup>; Alina Isakova, PhD<sup>13</sup>; Maurizio Morri, PhD<sup>12</sup>; Thanmayi Ranganath, MS<sup>1</sup>; Catherine A. Blish MD PhD<sup>1</sup>; Angela J. Rogers, MD<sup>1</sup>; Kari Nadeau, MD PhD<sup>1,14</sup>; Samuel Yang, MD<sup>15</sup>; Andra Blomkalns, MD<sup>15</sup>; Ruth O'Hara, MD PhD<sup>16</sup>; Norma F. Neff, PhD<sup>12</sup>; Christopher DeBoever, PhD<sup>2</sup>; Sándor Szalma<sup>17</sup>; Matthew T. Wheeler MD PhD<sup>1</sup>; Kyle Farh, MD PhD<sup>18</sup>; Gary P. Schroth, PhD<sup>18</sup>; Phil Febbo, MD<sup>18</sup>; Francis deSouza<sup>18</sup>; Marcelo Fernandez-Vina, PhD<sup>5,10</sup>; Amy Kistler, PhD<sup>12</sup>; Julia Palacios, PhD<sup>2,6</sup>; Benjamin A. Pinsky, MD PhD<sup>1,10</sup>; Carlos D. Bustamante, PhD<sup>2\*</sup>; Manuel A Rivas, PhD<sup>2\*</sup>; Euan A. Ashley, MD PhD<sup>1,4\*</sup>

\*Contributed equally

1. Department of Medicine, Stanford University School of Medicine, Stanford, CA, USA
2. Department of Biomedical Data Science, Stanford University, Stanford, CA, USA
3. Institute for Computational and Mathematical Engineering, Stanford University
4. Department of Genetics, Stanford University School of Medicine, Stanford, CA, USA
5. Histocompatibility & Immunogenetics Laboratory, Stanford Blood Center, Stanford Health Care
6. Department of Statistics, Stanford University, Stanford, CA, USA
7. Department of Aeronautics and Astronautics, Stanford University, Stanford, CA
8. Department of Computational Biology and Swiss Institute of Bioinformatics, University of Lausanne, Lausanne, Switzerland
9. Department of Biology, Stanford University, Stanford, CA
10. Department of Pathology, Stanford University School of Medicine, , Stanford, CA
11. Medical Scientist Training Program, University of Pittsburgh and Carnegie Mellon University, Pittsburgh, PA
12. Chan Zuckerberg Biohub, San Francisco, CA
13. Department of Bioengineering, Stanford University, Stanford, CA
14. Sean N. Parker Center for Allergy and Asthma Research, Stanford University School of Medicine, Stanford, CA
15. Department of Emergency Medicine, Stanford University School of Medicine, Stanford, CA
16. Department of Psychiatry and Behavioral Sciences, Stanford University School of Medicine, Stanford, CA
17. Takeda Development Center, Americas, Inc.
18. Illumina, Inc.

## Correspondence to:

Euan A. Ashley, MD PhD  
euan@stanford.edu

## ABSTRACT

The SARS-CoV-2 pandemic has differentially impacted populations of varied race, ethnicity and socioeconomic status. Admixture mapping and local ancestry inference represent powerful tools to examine genetic risk within multi-ancestry genomes independent of these confounding social constructs. Here, we leverage a pandemic tracking strategy in which we sequence viral and host genomes and transcriptomes from 1,327 nasopharyngeal swab residuals and integrate them with digital phenotypes from electronic health records. We demonstrate over-representation of individuals possessing Oceanian and Indigenous American ancestry in SARS-CoV-2 positive populations. Genome-wide-association disaggregated by admixture mapping reveals regions of chromosomes 5 and 14 associated with COVID19 severity within African and Oceanic local ancestries, respectively, independent of overall ancestry fraction. Phylodynamic tracking of consensus viral genomes reveals no association with disease severity or inferred ancestry. We further present summary data from a multi-omic investigation of human-leukocyte-antigen (HLA) typing, nasopharyngeal microbiome and human transcriptomics that reveal metagenomic and HLA associations with severe COVID19 infection. This work demonstrates the power of multi-omic pandemic tracking and genomic analyses to reveal distinct epidemiologic, genetic and biological associations for those at the highest risk.

## MAIN

Two central questions from the COVID-19 pandemic remain unresolved: who is at risk of severe disease, and why? Genome-wide-association-studies (GWAS) from the COVID-19 Host Genetics Initiative and others have identified up to 13 genetic loci associated with COVID19 infection, hospitalization and critical illness.<sup>1,2,3</sup> Among these loci is a chromosome 3 haplotype that entered the human population from Neanderthals, suggesting that genetic ancestry plays an important role in susceptibility and severity in SARS-CoV-2 infection.<sup>4</sup> At the same time, epidemiologic studies have shown that comorbidities, sex and socioeconomic status are strongly associated with infection prevalence and disease severity.<sup>5-8</sup> Several groups have reported higher incidence of COVID19 infection and higher disease severity among Hispanic/Latino and African American racial and ethnic groups.<sup>5,7</sup> Because social constructs of race and ethnicity covary with ancestry-related genetic markers, such associations may confound the study of COVID19 host genetic susceptibility. Genetic ancestry inference along the genome represents a powerful tool to examine genetic risk within the context of multi-ancestry genomes. These analyses are independent of the socioeconomic factors associated with race and ethnicity, due to the random shuffling of genomic ancestry segments, even between siblings, via recombination. Further, viral phylodynamics, host immunity (e.g., HLA typing), and metagenomics are important potential contributors to disease severity that can vary with ancestry and social factors and can be clarified with this omics-based approach.

To build the complex database necessary for these investigations, residual viral transport media (VTM) from SARS-CoV-2 clinical diagnostic tests were randomly prospectively collected from March 2020 to July 2020 and linked to structured clinical information from the electronic health record. 938 SARS-CoV-2 positive and 389 SARS-CoV-2 negative NP swab residuals were sequenced from a total of 892 (571 positive and 321 negative) patients (**Figure 1A&B**). Host whole genome sequencing was aligned and called using methods for low pass data<sup>9-11</sup> (mean of mean coverages: 2.56X±2.50X(SD), **Figures 1C**), followed by phasing and imputation with GLIMPSE (**Figure S1**).<sup>12</sup> Shotgun RNAseq of initial samples yielded high coverage of much of the viral genome for samples with clinical test CT values <35 regardless of RNA yield, which improved with primer-based capture (**Figure 1D** and **S1D**). For digital phenotype abstraction, we generated a COVID19 clinical severity score based on the ordinal scale proposed by the World Health Organization (**Table S1**). Clinical data were obtained through the STAnford Research Repository (STARR), specifically the STRIDE data management system, which is populated from patients' clinical and biospecimen data.<sup>13</sup> Severity

score was calculated on the date of sample collection and daily for one month before and indefinitely after (**Figures 1E&F and S2**).

We used genetic ancestry inference to identify subpopulations highly impacted by the COVID19 pandemic. Self-reported race and ethnicity re-demonstrates prior reports of over-representation of minority race and ethnic populations within the US amongst COVID19 positive patients (**Figure S3**). Based on genetic ancestry inference, a higher proportion of individuals of Oceanian (Pacific Islander) and Indigenous American genetic ancestry exists in COVID19 cases as compared to negative controls (**Fig 2A**, 44% vs. 26% for individuals having Indigenous American genetic ancestry, associated with Hispanic populations ( $\chi^2 = 99$ ,  $p < 1e-10$ ), and 4.6% vs. 2.3% for individuals having Oceanian ancestry, associated with Pacific Islander origins ( $\chi^2 = 13.9$ ,  $p = 2e-4$ )). Evaluation of temporal trends of genetic ancestry in case positivity reveals early predominance of European ancestry followed by a significant increase in Indigenous American ancestry after May 2020. These findings are recapitulated by self-reported ethnicity, as the majority of COVID19 cases between May and July 2020 self-identified as Hispanic/Latino in the medical record, an ethnicity often associated with mixed Indigenous American and European genetic ancestry (**Fig 2B**).

We also investigated the role of viral phylodynamics in COVID19 severity and its potential interaction with individual ancestry and disease severity. Consensus viral genomes (10X coverage at >99% of the genome) were recovered for 255 samples from unique, unrelated individuals. The estimated time to the most recent common ancestor of observed samples is December 11, 2019 with a 95% Bayesian CI of (2019-10-27, 2020-01-12). However, the phylogenetic reconstruction (**Figure 2C**) reveals an early introduction in the area between late December 2019 and early January of 2020 with several independent introductions later in February 2020. While about 37% of the infected individuals in the sample have Indigenous American ancestry, there is no evidence of exclusive transmission amongst individuals of this ancestry. Other genetic ancestries also do not sort with clades, though a single clade from early in the pandemic had fewer Hispanic individuals, consistent with the first wave prior to May 2020, in which European genetic ancestry individuals were enriched (**Figure 2C**). We also tested the hypothesis of association between viral lineages and disease severity. No association was found between specific clades and severity score at the time of NP swab (**Figure 2C**).

After adjusting for age, sex and BMI (known correlates of disease severity<sup>14–16</sup>) together with overall genetic ancestry proportion, we assessed association of genetic ancestry at a given genomic position with the COVID19 severity score as an ordinal outcome (admixture mapping). Because our captured case population was enriched for non-European ancestry groups, we were able to perform admixture mapping for six ancestries (African, Native American, Oceanian, South Asian, East Asian, and European/West Asian). This analysis revealed loci in chromosomal regions of African and Oceanic ancestry that met genome-wide significance, determined as previously described by Shriner et al.<sup>17</sup> (**Figure 3**). In segments of African ancestry, a region of chromosome 5 (5:34675449-35163719, GRCh38.p13) was associated with increased disease severity ( $p = 0.03$  after multiple test correction), while in segments of Oceanian ancestry, a region of chromosome 14 (14:44073392-44149995) was also associated with high severity COVID19 ( $p = 1.2e-07$  after correction). Two SNPs in these regions (rs13167486 and rs4608241 respectively) were reported in the same genome-wide interaction analysis of pathological hallmarks of Alzheimer's Disease<sup>18</sup> and map to two long noncoding (lnc)RNAs, SPRY4-AS1 and LINC02307. These lncRNAs are predicted to bind a set of miRNAs<sup>19</sup> whose empirically confirmed target genes are enriched in brain, as well as epithelium (e.g., lung).<sup>20, 21</sup> LncRNAs have an important role in antiviral immunity both via their effect in immune cells as well as direct interaction with the viral genome.<sup>22,23</sup>

We next established a resource of summary statistics for COVID19 severity versus host genotype by genetic ancestry, host HLA type, and metagenomic alignments (<https://covid-omics.org/results>), and have contributed host genetic summary data and viral consensus sequences to the COVID-19 Host Genetics Initiative<sup>3</sup> and GISAID,<sup>24</sup> respectively (**Figure 4A**). Using this resource, we explored the potential contribution of the nasopharyngeal microbiome and HLA-type as biological determinants of COVID19 severity. A UMAP plot of microbiome species abundance shows clustering largely independent of severity (**Figure 4B**); however, a regression of species abundance against COVID19 severity (controlling again for age and BMI) revealed enrichment of *Paracoccus yeei* transcripts in high severity cases, a bacterium that causes opportunistic infections in critically ill,<sup>25</sup> organ transplant,<sup>26</sup> and dialysis patients,<sup>27,28</sup> indicating an association with immune compromise and severe illness (**Figure 4C**,  $p = 7e-04$  after Bonferroni correction). The HLA-B\*07:02 allele (common prototype allele for the serotype B7) was associated with elevated risk of high severity score (OR 2.7 [1.4, 5.1],  $p = 2.9e-03$ ), whereas the HLA-C\*15:02 allele (common prototype allele for the serotype Cw15) was associated with risk reduction (OR 0.12 [0.02, 0.82],  $p = 1.41e-02$ ) (**Figure 4D**). HLA-B\*07:02 presents epitopes from the SARS-CoV-2 *N* gene and *Orf1ab*,<sup>29</sup> and the HLA-C\*15:02 allele contains two distinctive amino acid substitutions at residues 113 and 116 located within the peptide binding groove. HLA-C\*15:02 was associated with milder disease in the first SARS epidemic,<sup>30</sup> and is predicted to bind a SARS-CoV-2 Spike protein epitope.<sup>31</sup>

These results represent a substantial effort to assemble host and viral genomic, transcriptomic and digital clinical data from a diverse cross-section of the socioeconomic, racial and ethnic groups hit hardest by the COVID19 pandemic. We show that ancestry inference can be used to track changes in the affected population in real-time, demonstrating that Hispanic/Latino groups (associated with Indigenous American genetic ancestry) were disproportionately affected during a second pandemic wave. This is consistent with the model that this second wave was driven not by introduction from travelers (likely the source of the first wave), but by economic pressure on essential service workers to leave their homes and family units, enabling viral spread.<sup>32</sup> Phylodynamic overlay on this at-risk population further supports this conclusion, demonstrating that viral clades did not differentially affect ancestral groups, nor did they confer differential disease severity during six months of prospective enrollment. Thus, the impact of introduction of viral variants on community spread was likely less than that of exposure related to essential services work.

In addition to the use of ancestry inference to track the impact of the pandemic on ancestral sub-populations, genomic regions associated with COVID19 severity in the context of local African and Oceanian ancestries highlight potentially novel pathobiology. Both regions are near lncRNAs, non-coding transcripts that can interact with other small noncoding RNAs to alter viral immunity and host gene expression, one of which is expressed in lung epithelium, consistent with potential biological significance to SARS-CoV-2 infection.<sup>19,21</sup> Admixture mapping and local ancestry disaggregation were necessary to reveal these markers, which would likely otherwise be masked by social and economic determinants of severity that disproportionately affect these ancestral populations.

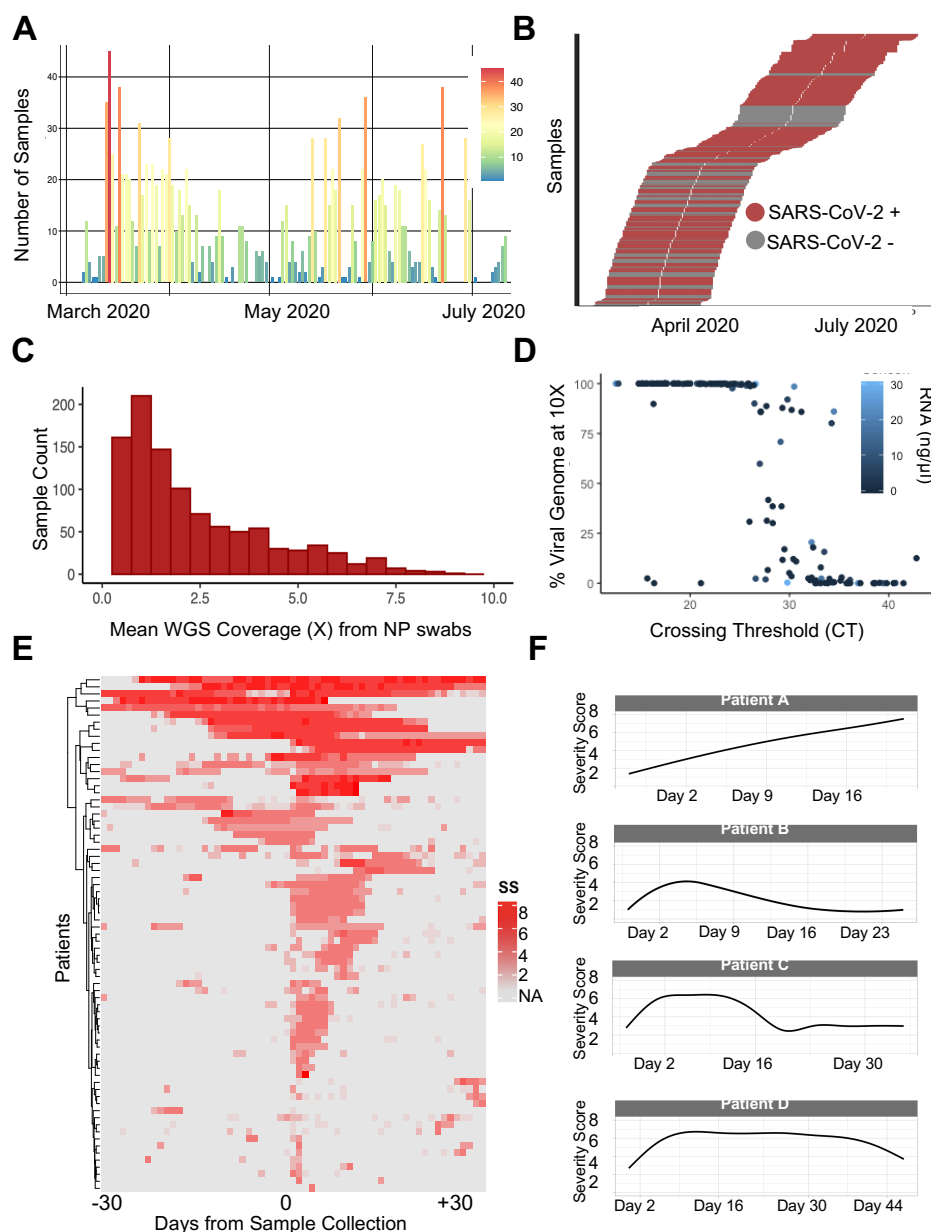
As the populations at highest risk of severe outcomes in a pandemic (due in large part to health and socioeconomic inequities) are not proportionally represented in existing datasets, the development of our real-time data collection strategy from clinical swab residuals was critical to assessing the relevance of ancestry-specific genetic variation.<sup>33</sup> We present a resource combining summary host and viral genomic, metagenomic and transcriptomic data with highly evolved digital phenotype abstraction from extant EHR data to help deconvolve genetic environmental and social factors while tracking spread across the community. This system can be applied in real-time to model individual and population trajectories in future emerging global infectious diseases (<https://covid-omics.org/results>), and can help illuminate underlying biology and medical insights relevant to diverse

populations that can be otherwise missed or ignored due to confounding with the social and economic factors that are also associated with race and ethnicity in diverse populations.



## FIGURES

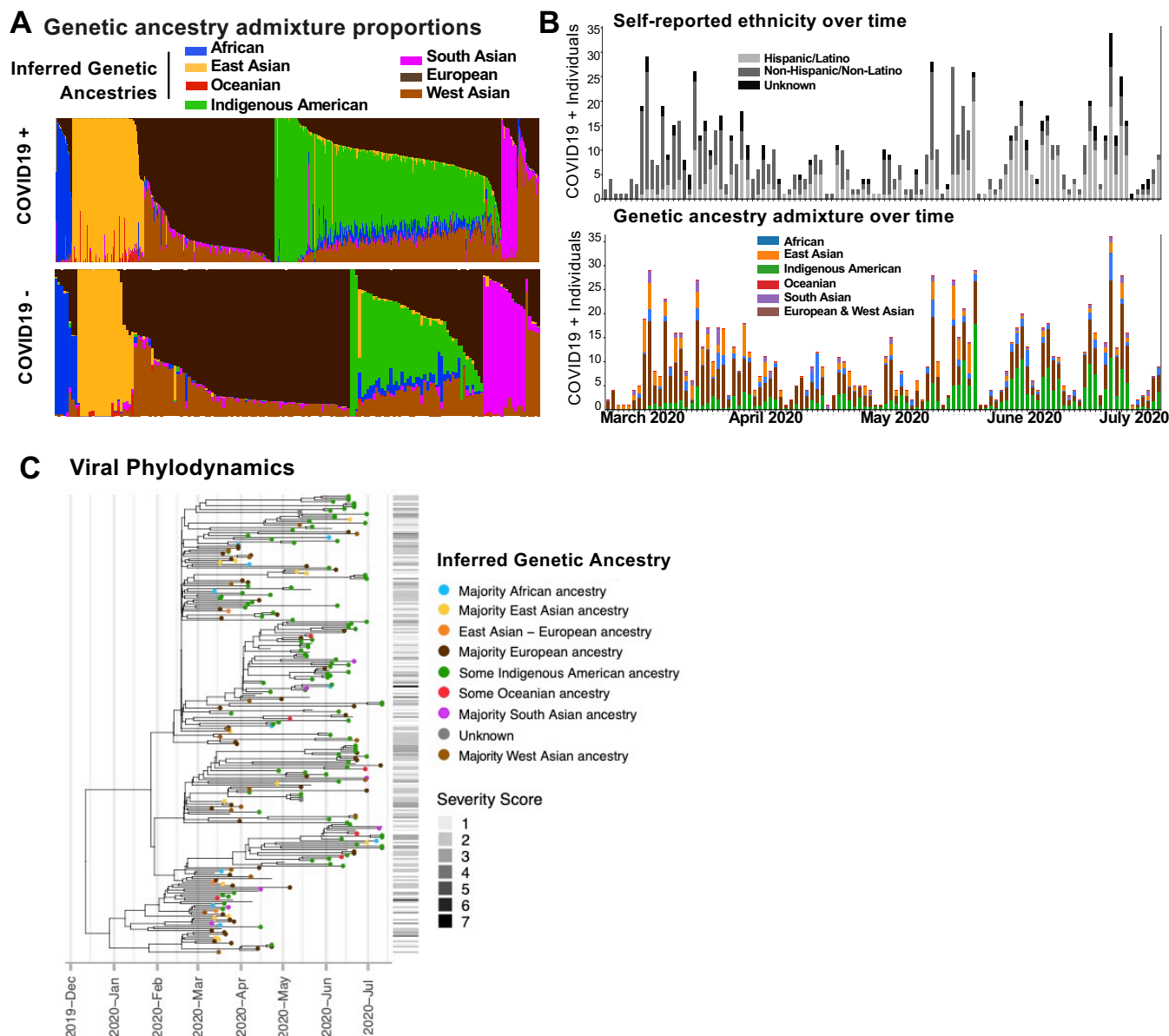
Figure 1



**Figure 1. SARS-CoV-2 Pandemic tracking from residual NP swabs and abstracted EHR data.**

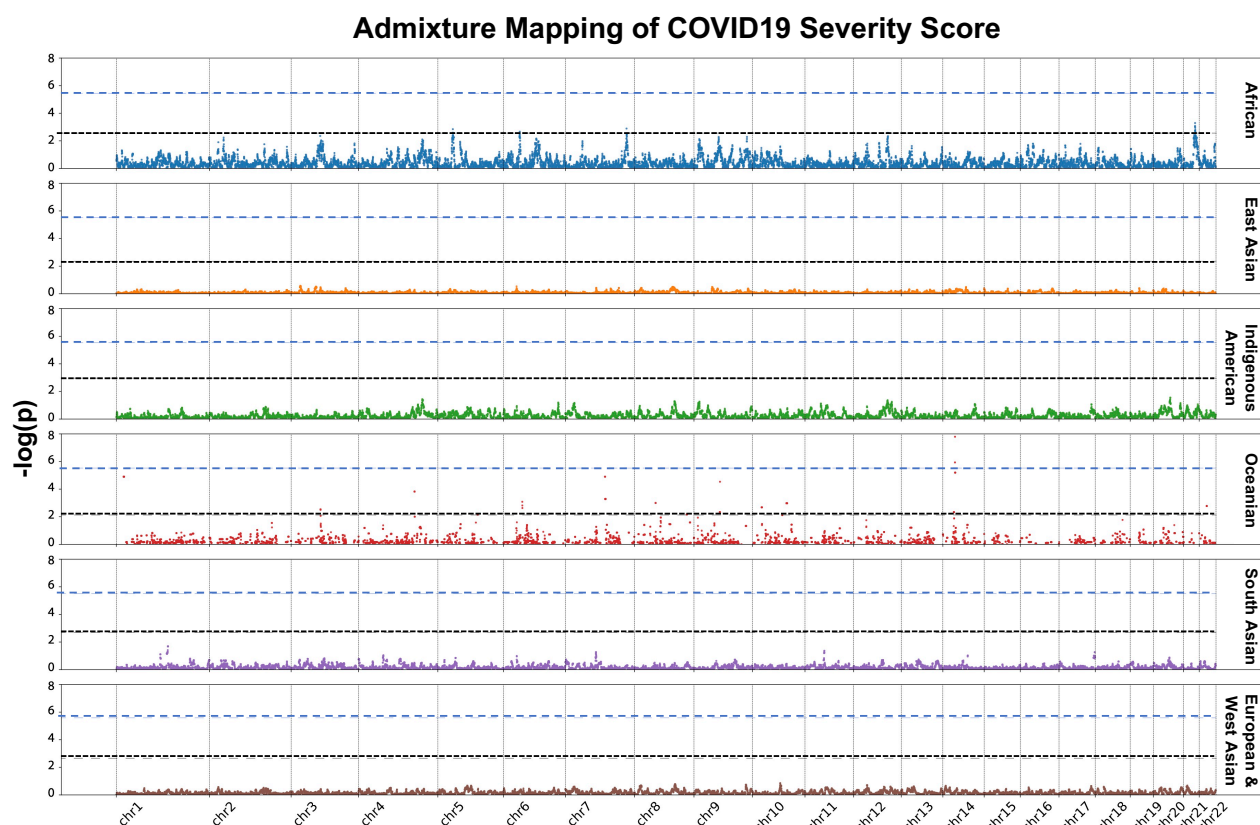
(A) We collected 1327 NP swab residuals from a single Northern California Hospital System between March and July 2020. Colors represent the number of samples collected over time (x-axis, see legend). (B) These included 938 SARS-CoV-2 positive and 389 SARS-CoV-2 negative swabs that were collected in parallel during this period. Time over which clinical data was abstracted for each individual is shown (30 days before and after sample collection date.) (C) Whole genome sequencing from DNA isolated from 150  $\mu$ l of NP swab VTM yielded sequence on >95% of samples with mean of means coverage 2.6X. (D) RNA sequencing using shotgun sequencing recovered consensus SARS-CoV-2 sequence on the majority of NP swabs with a clinical PCR CT value <30. ARTIC primer enrichment increased this yield (Figure S1D). (E and F) Severity scores abstracted directly from the electronic health record daily for thirty days before and after the positive NP swab test on all included patients with severity score  $\geq 4$  (hospitalized, needs oxygen) demonstrates significant variability in patient course.

Figure 2



**Figure 2. Genetic ancestry inference allows identification of high risk populations and tracing of transmission over time through its interaction with viral phylogeny and disease severity. (A)** Genetic ancestry admixture of individuals with positive versus negative COVID19 tests in the present study. Individuals with Indigenous American ancestry are overrepresented in cases, whereas controls show more European and South Asian genetic ancestry. **(B)** Self-reported (top) and genetic ancestry (bottom) of enrolled COVID19+ individuals over time reveals disproportionate representation of Hispanic/Latino ethnicity and Indigenous American ancestries during summer pandemic wave, whereas the first wave is seen to have predominantly affected non-Hispanic individuals and individuals of European genetic ancestry. **(C)** Phylogenetic reconstruction of SARS-CoV-2 sequences. Tip colors correspond to the inferred genetic ancestry of the infected hosts, whose consensus SARS-CoV-2 sequences were isolated and used for inferring the viral phylogeny. Horizontal lines at the bottom of the phylogeny indicate host severity scores corresponding to the tips of the phylogeny. Severity score codes are displayed in Table S1.

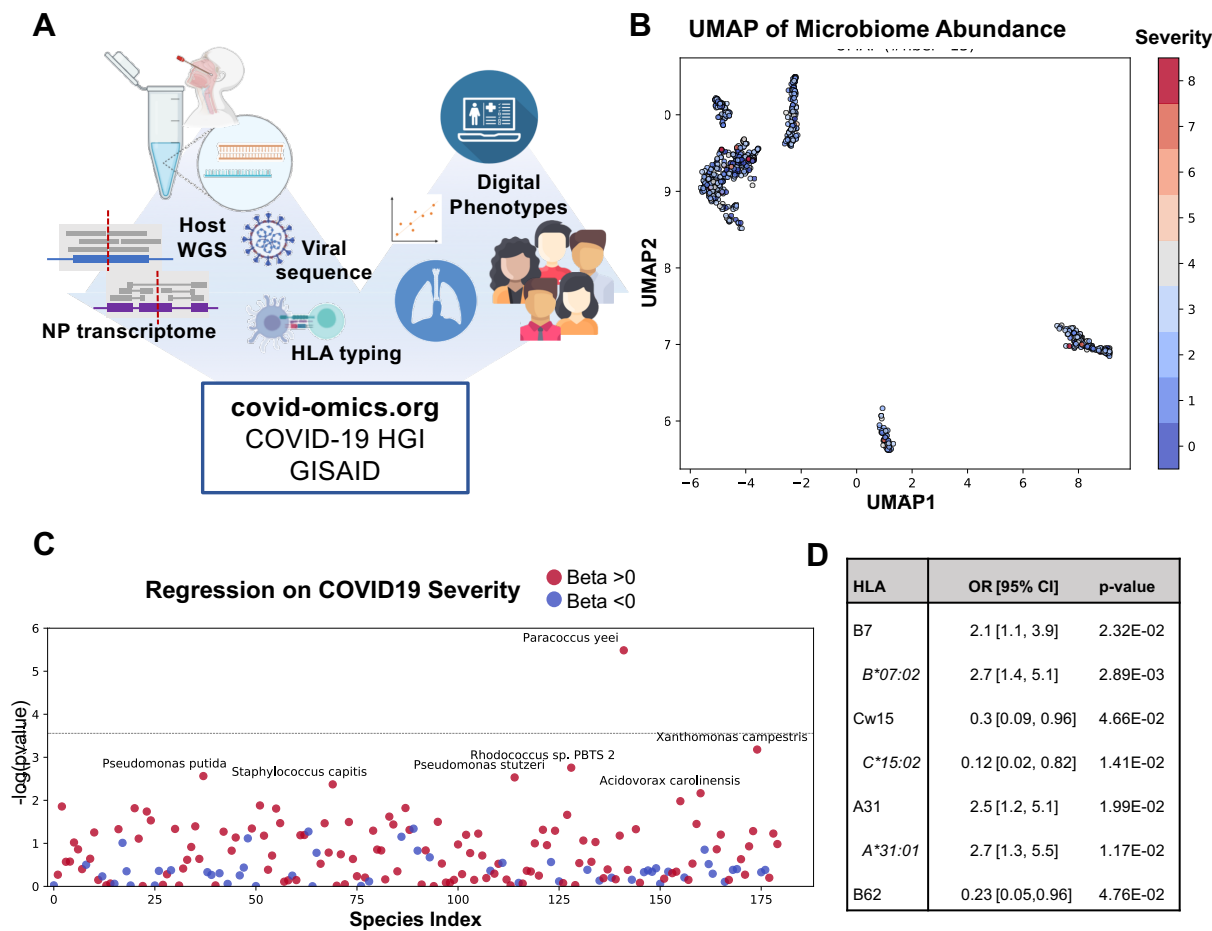
Figure 3



**Figure 3. Admixture mapping shows association of ancestry along the genomes of COVID+ patients with severity score.** Ancestry-specific risk loci found in Chromosome 5 and 14 in African and Oceanian ancestries, respectively, were found to be associated with the same functional mechanism. Genetic ancestry associations were first corrected for overall ancestry proportions of patients, for BMI, and for age. Lower black dotted-lines indicate less stringent ancestry-specific multiple test correction for a p-value of 0.05, using the method of Shriner et al., while higher, more stringent blue dashed lines indicate naive Bonferroni corrections for the same.



Figure 4



**Figure 4. (A)** Figure representing data resource. **(B)** Uniform manifold approximation and projection (UMAP) of Nasal Microbiome abundance colored by severity score. **(C)** Regression of species-specific transcript abundance against continuous disease severity, corrected for BMI and age, identified *P. yeii* abundance in the nasopharyngeal microbiome as associated with high severity COVID19 infections (Bonferroni adjusted  $p=7e-04$ ). **(D)** HLA allele and serotype (*italics*) association with COVID19 severity after ancestry adjustment by Cochran Mantel Haenszel test.

## METHODS

### *Sample Collection and diagnostics*

Residual VTM from SARS-CoV-2 positive nasopharyngeal swabs collected during clinical assessment of asymptomatic and symptomatic patients at Stanford Healthcare were used in accordance with the Stanford School of Medicine Institutional Review Board. RT-qPCR targeting the *envelope* gene or ORF1ab were used to detect infection. 152 samples with detectable SARS-CoV-2 RNA and 8 with undetectable virus by RT-qPCR were included. Positive samples were defined as those crossing threshold (CT) of 40 cycles or less on the RT-qPCR or positive Transcription-Mediated Amplification (TMA) diagnostic tests used at Stanford Health Care clinical laboratory.<sup>34</sup>

### *EHR Data Abstraction and Severity Score Development*

A critical task is to determine for every sampled patient the disease severity from the Electronic Health Records (EHR). To accomplish this task, we used as the “COVID-19 Clinical Severity Scale” an adapted version of the “Ordinal Scale for Clinical Improvement” proposed by the World Health Organization in the COVID-19 Therapeutic Trial Synopsis (Draft February 18, 2020, Table S1). This scale categorized the COVID-19 severity according to the level of care and oxygen support. Scores 1 to 2 include patients not requiring supplemental oxygen support or hospitalization. However, the WHO definition of these scores were modified due to their vague scope. Thus, score “1”, originally described as “no limitation of activities”, was modified to “asymptomatic patient”, and score “2”, from “limitation of activities” to “symptomatic patient” (symptoms were extracted and curated from EHR billed diagnoses). Scores 3 to 4 include patients hospitalized, with score 4 assigned only to those requiring non-invasive supplemental oxygen (oxygen mask). Scores 5 to 7 are defined as “severe disease” based on level of oxygen support. Thus, “score 5” is for patients requiring high flow oxygen and “score 6” mechanical ventilation. “Score 7” includes critically ill patients requiring, in addition to ventilation, the administration of specific medications (pressors), dialysis, or extracorporeal membrane ventilation (ECMO). A custom algorithm was written that abstracted digital phenotypes from each chart (see Table S1, “EHR annotations”). First, SARS-CoV-2 positive status was confirmed based on clinical test reports abstracted from the EHR. SARS-CoV-2 negative patients were assigned a score of zero. For SARS-CoV-2 positive patients, starting with the highest score (8, death) and working down, if criteria were met, the individual was assigned that score. If no clinical notes were available for data abstraction, then a severity score was not assigned and these individuals were not included in severity score based analyses. We calculated the score for any given date and assigned the maximum value according to the EHR annotations defined for every score (Supplementary Table 1). For example, patients with annotations for both ventilation and the administration of pressors received a score “7” for that day. Clinical data were obtained through the STANford Research Repository (STARR), a Stanford Medicine's approved resource for working with clinical data for research purposes extracted from the Epic database management system used by the Stanford hospitals. Specifically, we queried the STRIDE data management system, which is populated from patients' observational clinical, research, and biospecimen data<sup>13</sup>.

### *Nucleic Acid Extraction*

DNA; host genomic DNA was extracted from 200ul of VTM inoculated with nasopharyngeal swabs. Using a modified Qiagen DNEASY blood and tissue kit protocol and quantified using fluorometric readings (Protocols.io doi: [dx.doi.org/10.17504/protocols.io.bi8xkxhn](https://doi.org/10.17504/protocols.io.bi8xkxhn)). Total RNA was extracted from 200ul of VTM using a modified Ambion MirVana mRNA kit protocol (Protocols.io doi: [dx.doi.org/10.17504/protocols.io.bi8ykhxw](https://doi.org/10.17504/protocols.io.bi8ykhxw)) or Zymo Research Quick-Viral RNA extraction kits (R1041) and quantified using fluorometric readings.

### *Host gDNA library preparation and sequencing*

Using 1-10ng of host gDNA, the Illumina Nextera Flex library preparation was performed according to manufacturer's protocol (Protocols.io doi: [dx.doi.org/10.17504/protocols.io.bi8zkhx6](https://doi.org/10.17504/protocols.io.bi8zkhx6)). To allow for multiplexing, gDNA was barcoded using IDT-ILMN Nextera DNA UD Indices, a set of 10bp index adapters from Illumina. Indexed samples were diluted to 4nM, pooled, and analyzed on an Agilent TapeStation to ensure the mean DNA fragment size was ~300bp. Pooling and library quality was further assessed by sequencing the pool using a V3 MiSeq flow cell. 160 samples were pooled and sequenced for 76 cycles, paired end reads. For the purpose of QC, ~ 50 million reads were obtained and Q30 was determined to be >92%. If needed the pool was normalized (balanced) to ensure equal representation of each sample. The library was then sequenced on an Illumina NovaSeq 6000 using an S4 300 cycle flow cell.

#### *Viral RNA library preparation and sequencing*

After extraction, RNA acquired from 100 ul nasal swab media were incubated with recombinant RNase-free DNase (Qiagen, Inc.) per manufacturer's instructions for 15 minutes, followed by SPRI bead (GE Healthcare) purification to remove residual DNA remaining in each sample. A fixed volume (5uL) of the resulting RNA from each sample, together with a fixed mass (25pg) of the External RNA Controls Consortium RNA spike-in mix (ERCC RNA spike-in mix, Thermo Fisher), served as input for SARS-CoV-2 metatranscriptomic next generation sequencing (mNGS) library preparation ([dx.doi.org/10.17504/protocols.io.beshjeb6](https://doi.org/10.17504/protocols.io.beshjeb6); a modification of Deng et al. (2020)<sup>35</sup>).

For samples collected after May 2020, SARS-CoV-2 ARTIC V3 amplicon libraries were made from extracted total nucleic acid for whole genome sequencing using previously reported protocols<sup>[1]</sup>. Briefly, 3 ul of total nucleic acid was used as input for a randomly primed cDNA synthesis reaction. This cDNA served as input for 30 cycles of amplification with ARTIC V3 primers (github repo: [artic-network/artic-ncov2019](https://github.com/artic-network/artic-ncov2019)), and was then diluted 1:100 before tagmentation. Adaptor tagmentation was performed using homebrew Tn5, and 8 cycles of index PCR was performed using unique dual barcode Nextera indices (Detailed protocol: <https://protocols.io/view/artic-neb-tagmentation-protocol-high-throughput-wh-bt66nrhe>). Final libraries were pooled at equal volumes and cleaned at 0.7x (SPRI: Sample) using SPRIselect beads. Library was sequenced on Illumina Novaseq SP platform in a paired-end 2 x 150 cycle run. An incubation step with 1:10 dilution of FastSelect (Qiagen) reagent was included between the RNA fragmentation and first strand synthesis steps of the library prep to deplete highly abundant host rRNA sequences present in each sample. Equimolar pools (n=160-384 samples) of the resulting individual dual-barcoded library preps were subjected to paired-end 2 x 150bp sequence analysis on an Illumina NovaSeq 6000 (S2 or equivalent flow cell) to yield approximately 50 million reads per sample.

#### *Viral and Metagenomic Genome Alignment*

For SARS-CoV-2 genomes, FASTQ sequences were aligned to the SARS-CoV-2 reference genome NC\_045512.2 using minimap2.<sup>36</sup> Non-SARS-CoV-2 reads were filtered out with Kraken2,<sup>37</sup> using an index of human and viral genomes in RefSeq (index downloaded from <https://genexa.ch/sars2-bioinformatics-resources/>). Spiked primers for viral enrichment were trimmed from the ends of short reads using ivar.<sup>38</sup> Finally, a pileup of the aligned reads was generated with samtools<sup>39</sup>, and consensus genomes were called with ivar. The full pipeline used is publicly available on Github (<https://github.com/czbiohub/sc2-illumina-pipeline>). All viral consensus sequences were uploaded to the GISAID database (<https://www.gisaid.org/>).

Host and metagenomic RNA alignment was performed using STAR run against a combined index of the human reference genome grch38, SARS-CoV2 (SARSCoV2\_NC\_045512.2), and ERCC spike-ins. STAR parameters were chosen to avoid bias towards GTAG eukaryotic splice signatures for both the viral RNA and host RNA analyses. Metagenomic classification of reads unmapped to both SARS-CoV2 and human was performed using KrakenUniq (<https://genomebiology.biomedcentral.com/articles/10.1186/s13059-018-1568-0>). KrakenUniq

parameters ( $\geq 100$  kmers and duplication  $\leq$  less kmers ) were chosen to avoid false positives. From the filtered KrakenUniq output, an abundance table was created by finding the kmer percentages (kmers divided by the total kmer count) for relevant taxa detected for each individual. This table included only well-represented taxa, which was defined as those appearing in at least 10% of patients. A uniform manifold approximation (UMAP) plot was then created from this table using fifteen nearest neighbors.

### *Host Genome Sequence Alignment*

Low-coverage FASTQ sequences underwent quality control assessment via FastQC v0.11.8 before alt-aware alignment to GRCh38.p12 using BWA-MEM v0.7.17-r1188. Duplicate sequences were marked with MarkDuplicates of the Picard Tools suite v2.21.2. After duplicate marking, base quality score recalibration was performed with Picard Tools' BaseRecalibrator and high-confidence variant call sets from dbSNP and the 1000 Genomes Project. Quality control metrics, including coverage, were generated with Qualimap BAMQC v2.2.1, Samtools v1.10, and Mosdepth v0.2.9. Finally, quality control reports for each sample were aggregated using MultiQC v1.9. Reproducible code and steps are available at Protocols.io doi:

(<https://www.protocols.io/private/8CFBD1AD8FE611EA815E0A58A9FEAC2A>) All high confidence calls were contributed to the COVID19 Host Genetics Initiative.<sup>3</sup>

### *Variant Calling, Imputation, PCA, Kinship*

BAM files were used for an initial calling with bcftools v1.9 mpileup.<sup>40</sup> To account for the low-coverage sequencing we used the GLIMPSE algorithm v1.0 for imputation and phasing.<sup>12</sup> Briefly, this algorithm uses a reference set of haplotypes (1000 genomes project samples in our case) to compute genotype likelihoods using a Gibbs sampling procedure. The imputed data were filtered for low imputation scores (INFO $>0.8$ ), and were then merged with a reference set that contained samples from: (1) the 1000 genomes project, (2) the Human Genome Diversity Project (HGDP),<sup>41</sup> and (3) the Simons Genome Diversity Project (SGDP).<sup>42</sup> While merging these data, we set minor allele count thresholds of 5 for our data and 20 for the reference set (e.g., MAC $>4$  using bcftools), and a stringent call rate threshold (`--geno 0.01` in PLINK2).<sup>43,44</sup> The resulting VCF was loaded into PLINK2 v2.00a3LM using the following flags: `dosage=DS, --import-dosage-certainty 0.8`. These merged data had 4,111,339 autosomal variants that survived the filters above. PLINK2 was then used for LD pruning (`--indep-pairwise 500 10 0.1`) and PCA (`--maf 0.01 --pca`). We also extracted the kinship matrix of our samples using the King algorithm (`--make-king` in PLINK2).<sup>45</sup>

### *Genetic Ancestry Inference and assignment*

Genetic ancestry was determined by running supervised local ancestry inference (RFMix v2.03)<sup>46</sup> on the phased patient genomes using a training reference panel of single ancestry samples selected from the 1000 genomes project, HGDP, and SGDP via unsupervised genetic clustering (ADMIXTURE)<sup>47</sup> at  $K = 7$ . Only individuals with greater than 0.95 assignment to one of the seven unsupervised clusters in this analysis were used as references for RFMix. The unsupervised cluster labels--African, East Asian, South Asian, Oceanian, European, West Asian, and Indigenous American--were chosen to reflect the biogeographic origin of the reference samples assigned to each cluster. Local genetic ancestry assignments along the genome were then summed to create overall genetic ancestry proportions for each sample which were used for barplots, covariates for regression analyses, and for making individual genetic ancestry assignments. Individual genetic ancestry labels (e.g. for determination of enrichment in cases vs. controls (Figure 2A), association with viral clades (Figure 2C) and controlling HLA associations with severe disease by genetic ancestry) were assigned based on these overall proportions via the following decision sequence: some Oceanian ancestry (Pacific Islanders)  $> 5\%$ , some Indigenous American ancestry  $> 10\%$ , West Asian  $> 50\%$ , South

Asian > 50%, East Asian > 50%, European > 50%, African > 50%. For individuals meeting none of these criteria, and ancestry label was assigned consisting of the two predominant ancestries (e.g. East Asian and European in Figure 2C).

### *Association Analyses*

Admixture mapping association analyses were used to regress the residual of severity of COVID symptoms for each patient--after correcting for associations with overall genetic ancestry proportion, BMI, sex, and age--against the local ancestry of each particular window of the genome for that patient.<sup>48</sup> With the genome subdivided into 19,474 windows for local genomic ancestry assignment, and assuming complete independence between each, a naive Bonferroni corrected p-value of  $2.57 \times 10^{-6}$  is obtained for genome-wide significance at  $p=.05$ ; however, the genomic ancestry of neighboring, linked genomic windows is not independent and depends upon the characteristic length of each ancestry segment distribution, itself a function of the time since admixture in each population. A less stringent multiple-test correction factor that incorporates this distribution was empirically determined by considering the spectral density evaluated at frequency zero of an autoregressive model of local ancestries<sup>17</sup>, yielding an effective number of tests for each ancestry. This overall effective number of tests was taken over only samples that had at least 5% of that ancestry represented across their autosomes. Using this framework, together with the spectrum0.ar function implemented in the R package coda v 0.19, p-value thresholds for genome-wide significance at  $p=.05$  for each ancestry were determined: African  $2.54 \times 10^{-3}$ , East Asian  $3.89 \times 10^{-3}$ , Indigenous American  $1.15 \times 10^{-3}$ , Oceanian  $6.93 \times 10^{-3}$ , South Asian  $2.21 \times 10^{-3}$ , and European/West Asian  $1.83 \times 10^{-3}$ .

### *Host HLA Sequencing and Typing*

Host genomic DNA samples ranging from 22-75ng were batched in sets of 46 plus one positive and one negative control. AllType™ FASTplex™ NGS Assay kits (One Lambda, A Thermo Fisher Scientific Brand, Canoga Park, CA) were used to prepare DNA sequencing libraries for 11 classical HLA genes (*HLA-A*, *HLA-C*, *HLA-B*, *HLA-DRB3*, *HLA-DRB4*, *HLA-DRB5*, *HLA-DRB1*, *HLA-DQA1*, *HLA-DQB1*, *HLA-DPA1*, and *HLA-DPB1*). As the success of the DNA sequencing is dependent on the initial target amplification and the subsequent library preparation, the following changes were made to the manufacturer's protocol: 1) increased input DNA volume to 8.6  $\mu$ l while maintaining the manufacturer's recommended multiplex PCR protocol per sample; 2) eluted DNA in 12  $\mu$ l of suspension buffer after the initial amplicon purification, and proceeded to the library preparation without normalization process; 3) increased the number of thermal cycles to 17 in the final DNA library amplification; 4) eluted DNA fragments in 22  $\mu$ l for DNA sequencing. 500  $\mu$ l of 1.3 pM DNA sequencing library was loaded into a MiniSeq Mid Output Kit (300-cycles) (FC-420-1004), and sequenced using MiniSeq DNA sequencer (Illumina Inc., San Diego, CA).

A total of 429 subjects (301 cases, 128 SARS-CoV-2 negative controls) yielded interpretable sequence reads to generate HLA genotypes. We supplemented these samples with sequencing of buffy coat or whole blood collected from high severity COVID19 patients collected from hospitalized patients ( $n=193$ ). Fastq files were automatically imported into the TypeStream Visual NGS Analysis Software Version 2.0 upon the completion of DNA sequencing, and bioinformatically processed for DNA sequence assembly and HLA genotype assignments with IPD IMGT/HLA Database release version 3.39.0.<sup>49</sup> We modified the software setting so that a maximum of 1.5 million sequences or 750,000 paired-end sequences are used for the sequence assembly and HLA allele assignments. We visually inspected the HLA genotype calls by the software, and made corrections as needed. The approved HLA genotype results were exported in Histoimmunogenetics Markup Language (HML) format,<sup>50</sup> and generated comma separated value (CSV) reports for HLA genotypes, HLA serotypes including Bw4 and Bw6, KIR ligands (C1 and C2) and imputed HLA haplotypes.<sup>51,52</sup>

Subjects were grouped in three categories (Negative: SS\_MAX = 0; Mild: SS\_MAX = 1 - 3; Severe: SS\_MAX = 4 - 8), and organized in six broad ancestry groups [European (EUR), Hispanic (HIS), Asian



(ASI), African American (AFA), Native American (NAM) and Native Hawaiian/Pacific Islander (HPI) ] based on self-reported ethnicity in clinical records. When self-reported ethnicity was not available, genetic ancestry calculated from the low pass WGS in this study was used as described above. We converted the genetic ancestry information to self-reported medical record ethnicity format as follows: European and West Asia => EUR; some Indigenous American and Hispanic => HIS; East Asian and South Asian = ASI; African => AFA; fully Indigenous American => NAM; Oceanian => HPI. We compared the distribution of both HLA serotypes and alleles from COVID19 positive individuals with low disease severity (maximum severity score 1-3, n=336) to those with high disease severity (maximum severity score 4-8, n=94). HLA serotype and allele frequencies were calculated in both Mild and Severe groups, and Odds Ratio (OR: Mild vs. Severe) and *p*-values were calculated for each serotype and allele using Bridging ImmunoGenomic Data-Analysis Workflow Gaps (BIGDAWG).<sup>53</sup> Cochran- Mantel-Haenszel (CMH) tests<sup>54</sup> were subsequently performed for all observed HLA-A, -B, -C, -DRB1, -DQB1 and -DPB1 serotypes and alleles across three major ethnic groups (EUR, HIS and ASI) using the “mantelhaen.test” function in stats R package. Subjects with AFA, NAM and HPI ethnic groups were excluded from CMH tests, because we had only 6, 1 and 5 subjects, respectively, that yielded HLA genotypes.

### *Phylogenetic Analysis*

For Bayesian inference of the viral phylogeny, we assumed the Extended Bayesian Skyline Plot<sup>55</sup> prior on the effective population size and coalescent prior on the phylogeny, a fixed molecular clock with a uniform prior distribution centered at  $8 \times 10^{-4}$  substitutions per site per year as done in<sup>56</sup>. We assumed the HKY mutation model<sup>57</sup> with default hyperparameter priors in the BEAST2 software<sup>58</sup>. We ran a Markov chain Monte Carlo chain to approximate the posterior distribution of the model parameters for 20 million iterations and thinned every 5000 iterations. The first 10% of samples were discarded as burn-in. We used Tracer<sup>59</sup> to assess the convergence and confirm that the effective sample size (ESS) was >120 for all parameters (except in 15% of effective population size parameters, estimations not shown). Finally, we used TreeAnnotator<sup>60</sup> to summarize the phylogeny posterior distribution and generated the maximum clade credibility tree of Figure S3. To test the association between clade composition and binary traits, we used the R package treeSeg.<sup>61</sup>

### **Disclosures:**

VNP is a consultant and scientific advisor for Biomarin. CAB is on the SAB of Catamaran Bio and DeepCell, Inc. AJR is a Scientific Advisor of Merck, MTW is a stockholder of Personalis, Inc., GPS is on the SAB of Jumpcode Genomics. CDB is a member of the scientific advisory boards for Liberty Biosecurity, Personalis, 23andMe Roots into the Future, Ancestry.com, IdentifyGenomics, Genomelink, and Etalon and is a founder of CDB Consulting. MAR is on the SAB of 54Gene, Related Sciences and scientific founder of Broadwing Bio and has advised BioMarin, Third Rock Ventures and MazeTx. EAA is a founder of Personalis, Inc and DeepCell, Inc, a founding advisor of Neuvocor, a non-executive director at AstraZeneca, and an advisor of GenomeMedical, Novartis, Medical Excellence Capital, Foresite Capital, Cathay Capital and Third Rock Ventures.

### **Funding:**

NHLBI K08HL143185 (VNP), NIH R01HL144843 (EAA), NIH R01GM121404 (JAP), NIH U01HG009080 (MAR, AGI), (MAR), NIH U24 OD026629 04S1 (EAA, MTW), Stanford Medicine Dean's Postdoctoral Fellowship, American Heart Association Postdoctoral Fellowship, & Arnold O. Beckman Postdoctoral Independence Award (MJS), the John Taylor Babbitt Foundation (VNP), and the Sarnoff Cardiovascular Research Foundation (VNP), the Chan Zuckerberg Biohub, OneLambda, Thermofisher, Illumina, Inc., and Takeda Development Center Americas, Inc.



## REFERENCES

1. Pairo-Castineira, E. *et al.* Genetic mechanisms of critical illness in COVID-19. *Nature* **591**, 92–98 (2021).
2. Genomewide Association Study of Severe Covid-19 with Respiratory Failure. *N. Engl. J. Med.* **383**, 1522–1534 (2020).
3. Initiative, C.-19 H. G. & COVID-19 Host Genetics Initiative. Mapping the human genetic architecture of COVID-19. *Nature* (2021) doi:10.1038/s41586-021-03767-x.
4. Zeberg, H. & Pääbo, S. The major genetic risk factor for severe COVID-19 is inherited from Neanderthals. *Nature* **587**, 610–612 (2020).
5. Price-Haywood, E. G., Burton, J., Fort, D. & Seoane, L. Hospitalization and Mortality among Black Patients and White Patients with Covid-19. *N. Engl. J. Med.* **382**, 2534–2543 (2020).
6. Martinez, D. A. *et al.* SARS-CoV-2 Positivity Rate for Latinos in the Baltimore-Washington, DC Region. *JAMA* **324**, 392–395 (2020).
7. Figueroa, J. F., Wadhera, R. K., Lee, D., Yeh, R. W. & Sommers, B. D. Community-Level Factors Associated With Racial And Ethnic Disparities In COVID-19 Rates In Massachusetts: Study examines community-level factors associated with racial and ethnic disparities in COVID-19 rates in Massachusetts. *Health Aff.* **39**, 1984–1992 (2020).
8. Zhou, F. *et al.* Clinical course and risk factors for mortality of adult inpatients with COVID-19 in Wuhan, China: a retrospective cohort study. *Lancet* **395**, 1054–1062 (2020).
9. Homburger, J. R. *et al.* Low coverage whole genome sequencing enables accurate assessment of common variants and calculation of genome-wide polygenic scores. *Genome Med.* **11**, 74 (2019).
10. Garcia, M. *et al.* Sarek: A portable workflow for whole-genome sequencing analysis of germline and somatic variants. *F1000Res.* **9**, 63 (2020).
11. Poplin, R. *et al.* Scaling accurate genetic variant discovery to tens of thousands of samples.

- bioRxiv 201178. *Publisher Full Text* (2017).
12. Rubinacci, S., Ribeiro, D. M., Hofmeister, R. & Delaneau, O. Efficient phasing and imputation of low-coverage sequencing data using large reference panels. *bioRxiv* 2020.04.14.040329 (2020) doi:10.1101/2020.04.14.040329.
  13. Lowe, H. J., Ferris, T. A., Hernandez, P. M. & Weber, S. C. STRIDE--An integrated standards-based translational research informatics platform. *AMIA Annu. Symp. Proc.* **2009**, 391–395 (2009).
  14. Peng Y. D. *et al.* [Clinical characteristics and outcomes of 112 cardiovascular disease patients infected by 2019-nCoV]. *Zhonghua Xin Xue Guan Bing Za Zhi* **48**, 450–455 (2020).
  15. Livingston, E. & Bucher, K. Coronavirus Disease 2019 (COVID-19) in Italy. *JAMA* **323**, 1335 (2020).
  16. Wang, D. *et al.* Clinical Characteristics of 138 Hospitalized Patients With 2019 Novel Coronavirus-Infected Pneumonia in Wuhan, China. *JAMA* **323**, 1061–1069 (2020).
  17. Shriner, D., Adeyemo, A. & Rotimi, C. N. Joint ancestry and association testing in admixed individuals. *PLoS Comput. Biol.* **7**, e1002325 (2011).
  18. Wang, H. *et al.* Genome-wide interaction analysis of pathological hallmarks in Alzheimer's disease. *Neurobiol. Aging* **93**, 61–68 (2020).
  19. Paraskevopoulou, M. D. *et al.* DIANA-LncBase v2: indexing microRNA targets on non-coding transcripts. *Nucleic Acids Res.* **44**, D231–8 (2016).
  20. Fagerberg, L. *et al.* Analysis of the Human Tissue-specific Expression by Genome-wide Integration of Transcriptomics and Antibody-based Proteomics. *Mol. Cell. Proteomics* **13**, 397–406 (2014).
  21. Hsu, S.-D. *et al.* miRTarBase: a database curates experimentally validated microRNA–target interactions. *Nucleic Acids Research* vol. 39 D163–D169 (2011).
  22. Moazzam-Jazi, M., Lanjanian, H., Maleknia, S., Hedayati, M. & Daneshpour, M. S. Interplay

- between SARS-CoV-2 and human long non-coding RNAs. *J. Cell. Mol. Med.* (2021) doi:10.1111/jcmm.16596.
23. Ginn, L., La Montagna, M., Wu, Q. & Shi, L. Diverse roles of long non-coding RNAs in viral diseases. *Reviews in Medical Virology* (2020) doi:10.1002/rmv.2198.
24. Elbe, S. & Buckland-Merrett, G. Data, disease and diplomacy: GISAID's innovative contribution to global health. *Glob Chall* **1**, 33–46 (2017).
25. Funke, G., Frodl, R. & Sommer, H. First comprehensively documented case of *Paracoccus yeei* infection in a human. *J. Clin. Microbiol.* **42**, 3366–3368 (2004).
26. Schweiger, M. *et al.* Case of *Paracoccus yeei* infection documented in a transplanted heart. *Transpl. Infect. Dis.* **13**, 200–203 (2011).
27. Arias, M. A. & Clark, J. *Paracoccus yeei* as a cause of peritoneal dialysis peritonitis in the United Kingdom. *IDCases* **15**, e00486 (2019).
28. Wallet, F. *et al.* *Paracoccus yeei*: a new unusual opportunistic bacterium in ambulatory peritoneal dialysis. *Int. J. Infect. Dis.* **14**, e173–4 (2010).
29. Ferretti, A. P. *et al.* Unbiased Screens Show CD8 T Cells of COVID-19 Patients Recognize Shared Epitopes in SARS-CoV-2 that Largely Reside outside the Spike Protein. *Immunity* vol. 53 1095–1107.e3 (2020).
30. Wang, S.-F. *et al.* Human-Leukocyte Antigen Class I Cw 1502 and Class II DR 0301 Genotypes Are Associated with Resistance to Severe Acute Respiratory Syndrome (SARS) Infection. *Viral Immunology* vol. 24 421–426 (2011).
31. Moura, R. R. de *et al.* Immunoinformatic approach to assess SARS-CoV-2 protein S epitopes recognised by the most frequent MHC-I alleles in the Brazilian population. *Journal of Clinical Pathology* jclinpath–2020 (2020) doi:10.1136/jclinpath-2020-206946.
32. Reitsma, M. B. *et al.* Racial/Ethnic Disparities In COVID-19 Exposure Risk, Testing, And Cases At The Subcounty Level In California. *Health Aff.* **40**, 870–878 (2021).



33. Jack A. Kosmicki Julie E. Horowitz Nilanjana Banerjee Rouel Lanche Anthony Marcketta Evan Maxwell Xiaodong Bai Dylan. Sun Joshua D. Backman Deepika Sharma Fabricio S. P. Kury Hyun M. Kang Colm O'Dushlaine Ashish Yadav Adam J. Mansfield Alexander H. Li Kyoko Watanabe Lauren Gurski Shane E. McCarthy Adam E. Locke Shareef Khalid Sean O'Keeffe Joelle Mbatchou Olympe Chazara Yunfeng Huang Erika Kvikstad Amanda O'Neill Paul Nioi Meg M. Parker Slavé Petrovski Heiko Runz Jospeh D. Szustakowski Quanli Wang Emily Wong Aldo Cordova-Palomera Erin N. Smith Sandor Szalma Xiuwen Zheng Sahar Esmaeeli Justin W. Davis Yi-Pin Lai Xing Chen Ann E. Justice Joseph B. Leader Tooraj Mirshahi David J. Carey Anurag Verma Giorgio Sirugo Marylyn D. Ritchie Daniel J. Rader Gundula Povysil David B. Goldstein Krzysztof Kiryluk Erola Pairo-Castineira Konrad Rawlik Dorota Pasko Susan Walker Alison Meynert Athanasios Kousathanas Loukas Moutsianas Albert Tenesa Mark Caulfield Richard Scott James F. Wilson J. Kenneth Baillie Guillaume Butler-Laporte Tomoko Nakanishi Mark Lathrop J. Brent Richards Regeneron Genetics Center\* UKB Exome Sequencing Consortium\* Marcus Jones Suganthi Balasubramanian William Salerno Alan R. Shuldiner Jonathan Marchini John D. Overton Lukas Habegger Michael N. Cantor Jeffrey G. Reid Aris Baras Goncalo R. Abecasis Manuel A. R. Ferreira. Pan-ancestry exome-wide association analyses of COVID-19 outcomes in 586,157 individuals. *Am. J. Hum. Genet.* (2021) doi:10.1016/j.ajhg.2021.05.017.
34. Hogan, C. A., Sahoo, M. K. & Pinsky, B. A. Sample Pooling as a Strategy to Detect Community Transmission of SARS-CoV-2. *JAMA* (2020) doi:10.1001/jama.2020.5445.
35. Deng, X. *et al.* Metagenomic sequencing with spiked primer enrichment for viral diagnostics and genomic surveillance. *Nat Microbiol* **5**, 443–454 (2020).
36. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).
37. Wood, D. E., Lu, J. & Langmead, B. Improved metagenomic analysis with Kraken 2. *Genome Biol.* **20**, 257 (2019).

38. Grubaugh, N. D. *et al.* An amplicon-based sequencing framework for accurately measuring intrahost virus diversity using PrimalSeq and iVar. *Genome Biol.* **20**, 8 (2019).
39. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* vol. 25 2078–2079 (2009).
40. Li, H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* **27**, 2987–2993 (2011).
41. Bergström, A. *et al.* Insights into human genetic variation and population history from 929 diverse genomes. *Science* **367**, (2020).
42. Mallick, S. *et al.* The Simons Genome Diversity Project: 300 genomes from 142 diverse populations. *Nature* **538**, 201–206 (2016).
43. Chang, C. C. *et al.* Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* **4**, 7 (2015).
44. Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).
45. Manichaikul, A. *et al.* Robust relationship inference in genome-wide association studies. *Bioinformatics* **26**, 2867–2873 (2010).
46. Maples, B. K., Gravel, S., Kenny, E. E. & Bustamante, C. D. RFMix: a discriminative modeling approach for rapid and robust local-ancestry inference. *Am. J. Hum. Genet.* **93**, 278–288 (2013).
47. Alexander, D. H. & Lange, K. Enhancements to the ADMIXTURE algorithm for individual ancestry estimation. *BMC Bioinformatics* **12**, 246 (2011).
48. Chi, C. *et al.* Admixture mapping reveals evidence of differential multiple sclerosis risk by genetic ancestry. *PLoS Genet.* **15**, e1007808 (2019).
49. Robinson, J. *et al.* The IPD and IMGT/HLA database: allele variant databases. *Nucleic Acids Res.* **43**, D423–31 (2015).

50. Milius, R. P. *et al.* Histoimmunogenetics Markup Language 1.0: Reporting next generation sequencing-based HLA and KIR genotyping. *Hum. Immunol.* **76**, 963–974 (2015).
51. Gragert, L., Madbouly, A., Freeman, J. & Maiers, M. Six-locus high resolution HLA haplotype frequencies derived from mixed-resolution DNA typing for the entire US donor registry. *Hum. Immunol.* **74**, 1313–1320 (2013).
52. Osoegawa, K. *et al.* HLA Haplotype Validator for quality assessments of HLA typing. *Hum. Immunol.* **77**, 273–282 (2016).
53. Pappas, D. J., Marin, W., Hollenbach, J. A. & Mack, S. J. Bridging ImmunoGenomic Data Analysis Workflow Gaps (BIGDAWG): An integrated case-control analysis pipeline. *Hum. Immunol.* **77**, 283–287 (2016).
54. Mantel, N. & Haenszel, W. Statistical aspects of the analysis of data from retrospective studies of disease. *J. Natl. Cancer Inst.* **22**, 719–748 (1959).
55. Heled, J. & Drummond, A. J. Bayesian inference of population size history from multiple loci. *BMC Evol. Biol.* **8**, 289 (2008).
56. Nadeau, S. A., Vaughan, T. G., Scire, J., Huisman, J. S. & Stadler, T. The origin and early spread of SARS-CoV-2 in Europe. *Proc. Natl. Acad. Sci. U. S. A.* **118**, (2021).
57. Hasegawa, M., Kishino, H. & Yano, T. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *J. Mol. Evol.* **22**, 160–174 (1985).
58. Bouckaert, R. *et al.* BEAST 2: a software platform for Bayesian evolutionary analysis. *PLoS Comput. Biol.* **10**, e1003537 (2014).
59. Rambaut, A., Drummond, A. J., Xie, D., Baele, G. & Suchard, M. A. Posterior Summarization in Bayesian Phylogenetics Using Tracer 1.7. *Syst. Biol.* **67**, 901–904 (2018).
60. Rambaut, A. & Drummond, A. J. TreeAnnotator v1. 8.2. *MCMC Output analysis* (2015).
61. Behr, M., Ansari, M. A., Munk, A. & Holmes, C. Testing for dependence on tree structures. *Proc. Natl. Acad. Sci. U. S. A.* **117**, 9787–9792 (2020).

62. Camacho-Bydume, C. *et al.* Specific Class I HLA Supertypes but Not HLA Zygosity or Expression Are Associated with Outcomes following HLA-Matched Allogeneic Hematopoietic Cell Transplant: HLA Supertypes Impact Allogeneic HCT Outcomes. *Transplant Cell Ther* **27**, 142.e1–142.e11 (2021).

## Supplementary Material

### **Deconvoluting complex correlates of COVID19 severity with local ancestry inference and viral phylodynamics: Results of a multiomic pandemic tracking strategy**

Victoria N. Parikh, MD\*; Alexander G. Ioannidis, PhD\*; David Jimenez-Morales, PhD; John E. Gorzynski, DVM; Hannah N. De Jong, BS; Xiran Liu, MS; Jonasel Roque, BS; Victoria P. Cepeda-Espinoza, PhD; Kazutoyo Osoegawa, PhD; Chris Hughes, BS; Shirley C. Sutton, MS; Nathan Youlton, BS; Ruchi Joshi, MS; David Amar, PhD; Yosuke Tanigawa, PhD; Douglas Russo, MS; Justin Wong, MS; Jessie T. Lauzon, MS; Jacob Edelson, MS; Daniel Mas Montserrat, PhD; Yongchan Kwon, PhD; Simone Rubinacci, PhD; Olivier Delaneau, PhD; Lorenzo Cappello, PhD; Jaehee Kim, PhD; Massa J. Shoura, PhD; Archana N. Raja, PhD; Nathaniel Watson; Nathan Hammond; Elizabeth Spiteri PhD; Kalyan C. Mallemapati, Gonzalo Montero-Martín, PhD; Jeffrey Christle, PhD; Jennifer Kim, BS; Anna Kirillova, BS; Kinya Seo, PhD; Yong Huang, MD; Chunli Zhao, PhD; Sonia Moreno-Grau PhD; Steven G. Hershman, MS, MSE, PhD; Karen P. Dalton; Jimmy Zhen; Jack Kamm, PhD, Karan D. Bhatt, MS; Alina Isakova, PhD; Maurizio Morri, PhD; Thanmayi Ranganath, MS; Catherine A. Blish MD PhD; Angela J. Rogers, MD; Kari Nadeau, MD PhD; Samuel Yang, MD; Andra Blomkalns, MD; Ruth O'Hara, MD PhD; Norma F. Neff, PhD; Christopher DeBoever, PhD, Sándor Szalma, Matthew T. Wheeler MD PhD; Kyle Farh, MD PhD; Gary P. Schroth, PhD, Phil Febbo, MD, Francis deSouza, Marcelo Fernandez-Vina, PhD; Amy Kistler, PhD; Julia Palacios, PhD; Benjamin A. Pinsky, MD PhD; Carlos D. Bustamante, PhD\*; Manuel A Rivas, PhD\*; Euan A. Ashley, MD PhD\*



## **Supplementary Text**

### *Additional detail regarding HLA-Severity associations*

The *HLA-C\*15:02* allele is common in some Native American tribes from North, Central and South America. The alleles of *HLA-C* can be classified into the C2 and C1 groups according to their interaction with the inhibitory receptors KIR2DL1 and KIR2DL2/3 present in NK cells, respectively; the allele *HLA-C\*15:02* belongs to the C2 group. The current study found no significant associations of other *HLA-C* alleles of the C2 group or of other HLA ligands of KIR receptors suggesting that the broad KIR/HLA ligand interactions are not major determinants of disease course. We could not confirm the risk reduction in the prototype allele *HLA-B\*15:01* for the serotype B62, because we observed other B62 alleles (*HLA-B\*15:04*, *-B\*15:25*, *-B\*15:27*, *-B\*15:35*, *-B\*15:48* and *-B\*15:146*). They are mostly HLA alleles for Hispanic and Asian. The *HLA-B\*15:01* allele was recently associated with asymptomatic SARS-CoV-2 infection in European ancestry subjects. In our study subjects only 32% percent of the subjects had European ancestry; we did not find statistical significance for the individual allele *HLA-B\*15:01* that is by far more common than other alleles of the serotype B62 in Caucasians. The alleles *HLA-B\*15:04*, *-B\*15:25*, *-B\*15:27*, *-B\*15:35*, *-B\*15:48* and *-B\*15:146* are found in subjects with Hispanic and Asian. The B\*15 alleles corresponding to the B62 serotype detected in the present study show high structural homology (differing by 1-5 amino acid replacements) and share identity in the B-pocket substructure that determines what type of peptide residue at the second position is anchored. A recent study in outcomes of hematopoietic stem cell transplantation found that the HLA-B62 supertype was associated with decreased transplant-related mortality in the entire patient cohort.<sup>62</sup> It can be speculated that HLA-B62 may play a role in immune response; the associations identified in the present study lend support to the hypothesis that peptide presentation by alleles of the B62 serotype in addition to the *HLA-C\*15:02* allele may contribute to beneficial outcomes in SARS-CoV-2 Disease.

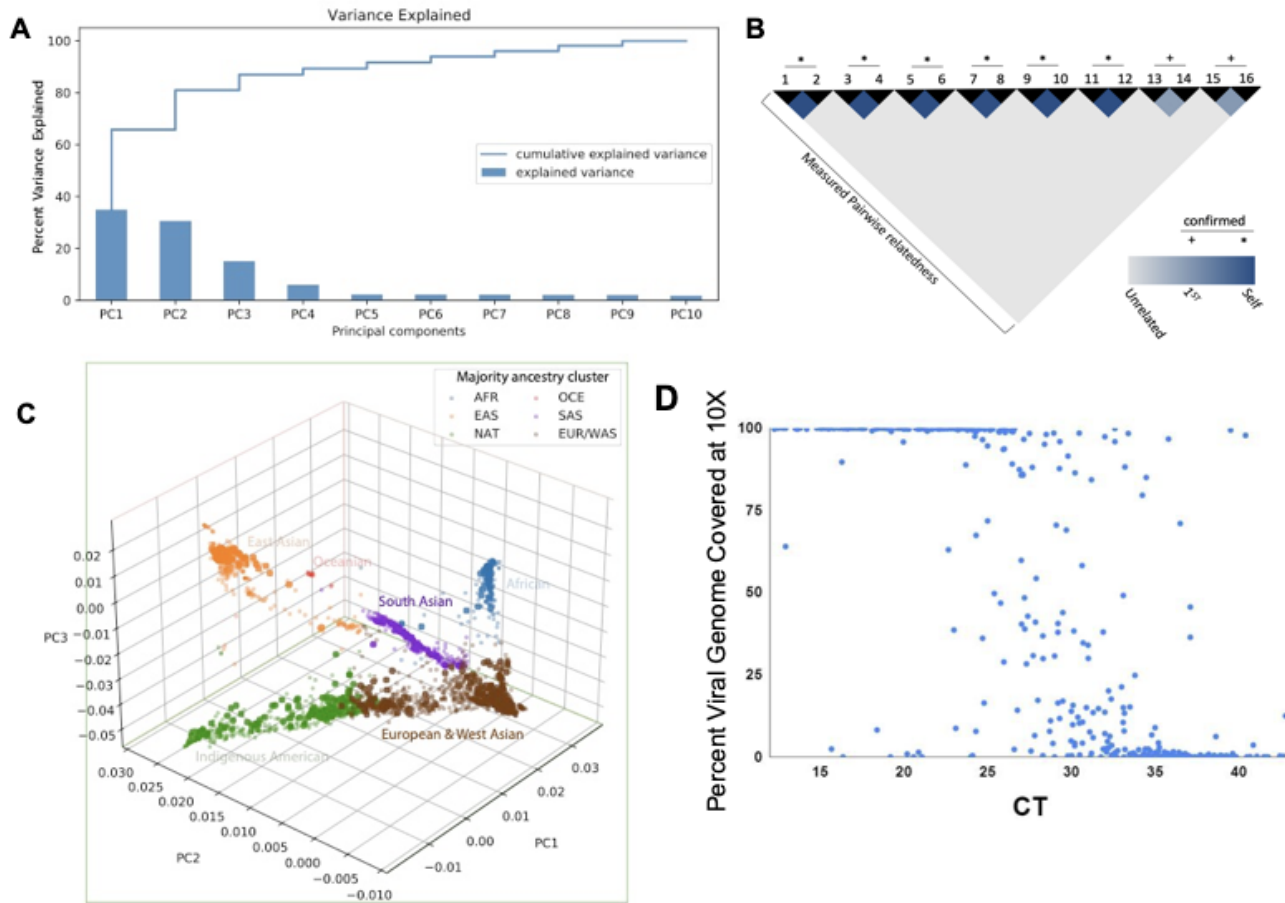
## Supplementary Tables

**Supplementary Table 1.** WHO-based COVID-19 Clinical Severity Scale

Score	Patient Status	Description	EHR annotations
0	Uninfected	No clinical or virological evidence of infection	SARS-CoV-2 negative
1	Ambulatory	Asymptomatic	Non COVID-related symptoms
2		Symptomatic	COVID related symptoms (cough, fever, shortness of breath, chest pain, tachycardia, acute pharyngitis, fever, others)
3	Hospitalized, Mild disease	Hospitalized, no oxygen	COVID related symptoms + hospitalization time (without oxygen support)
4		Hospitalized, oxygen mask	Nasal cannula; face mask; high-oxygen mask; aerosol mask; tracheal collar; others
5	Critical Care, Severe disease	non-invasive ventilation / high flow oxygen	High-flow nasal cannula; non-invasive positive pressure ventilation
6		intubation and mechanical ventilation	Invasive Ventilation; endotracheal tube
7		intubation + organ support	Intubation + dobutamine, dopamine, epinephrine, norepinephrine, phenylephrine, vasopressin
8	Dead	Deceased	Deceased

## Supplementary Figures

### Supplementary Figure 1



**Figure S1. (A)** Variance explained by the top 10 PCs. The first four PCs explain ~90% of variance. **(C)** Genomic data of patients (bold points) projected onto top three PCs created by samples from 1000 Genomes, HGDP and SGDP (light points). **(B)** Genotyping by low pass WGS recapitulates known kinship relationships. **(D)** Viral genome recovery using ARTIC primer capture was improved at higher CT values compared with shotgun RNAseq (Figure 1D).

## Supplementary Figure 2

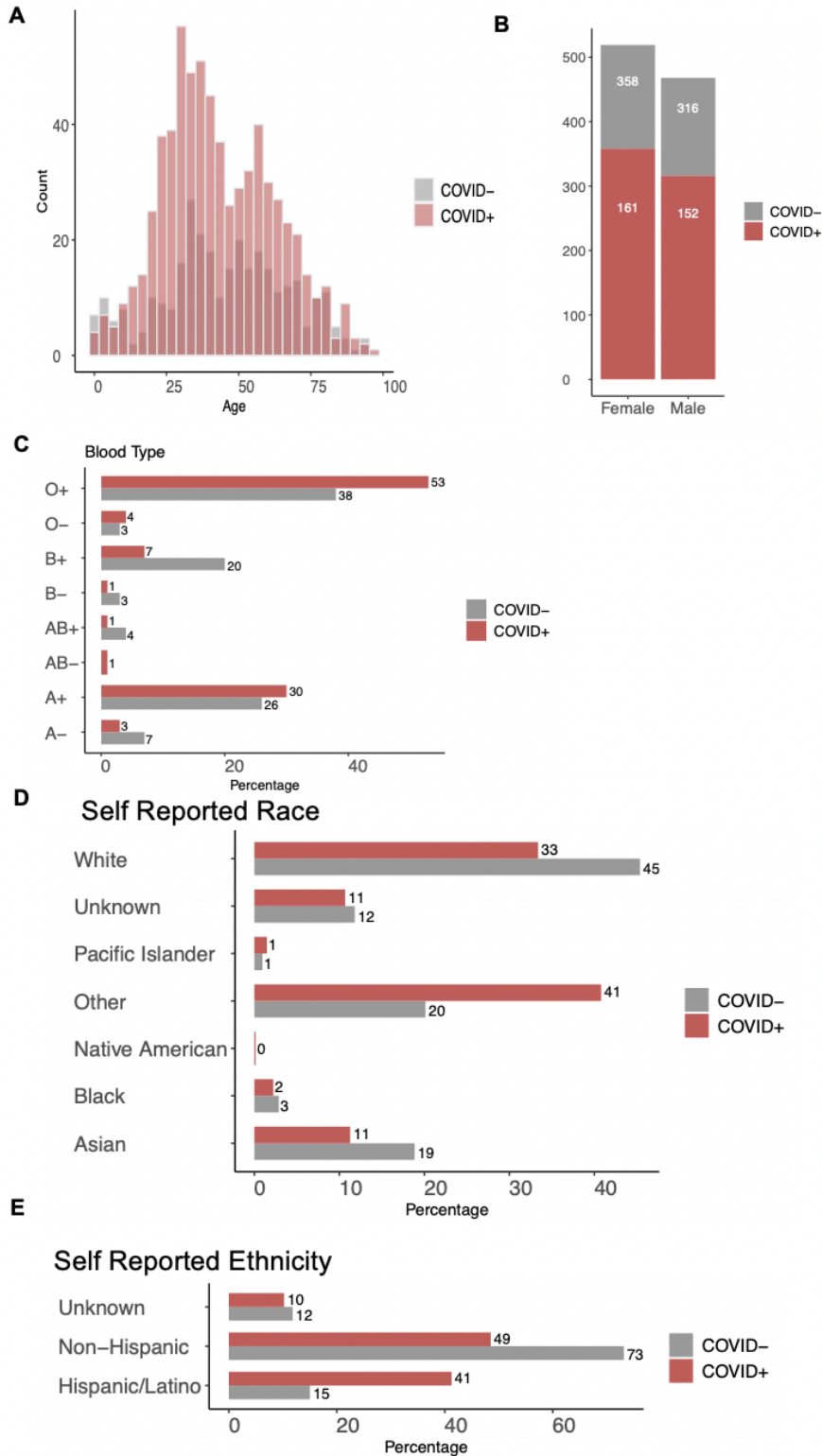
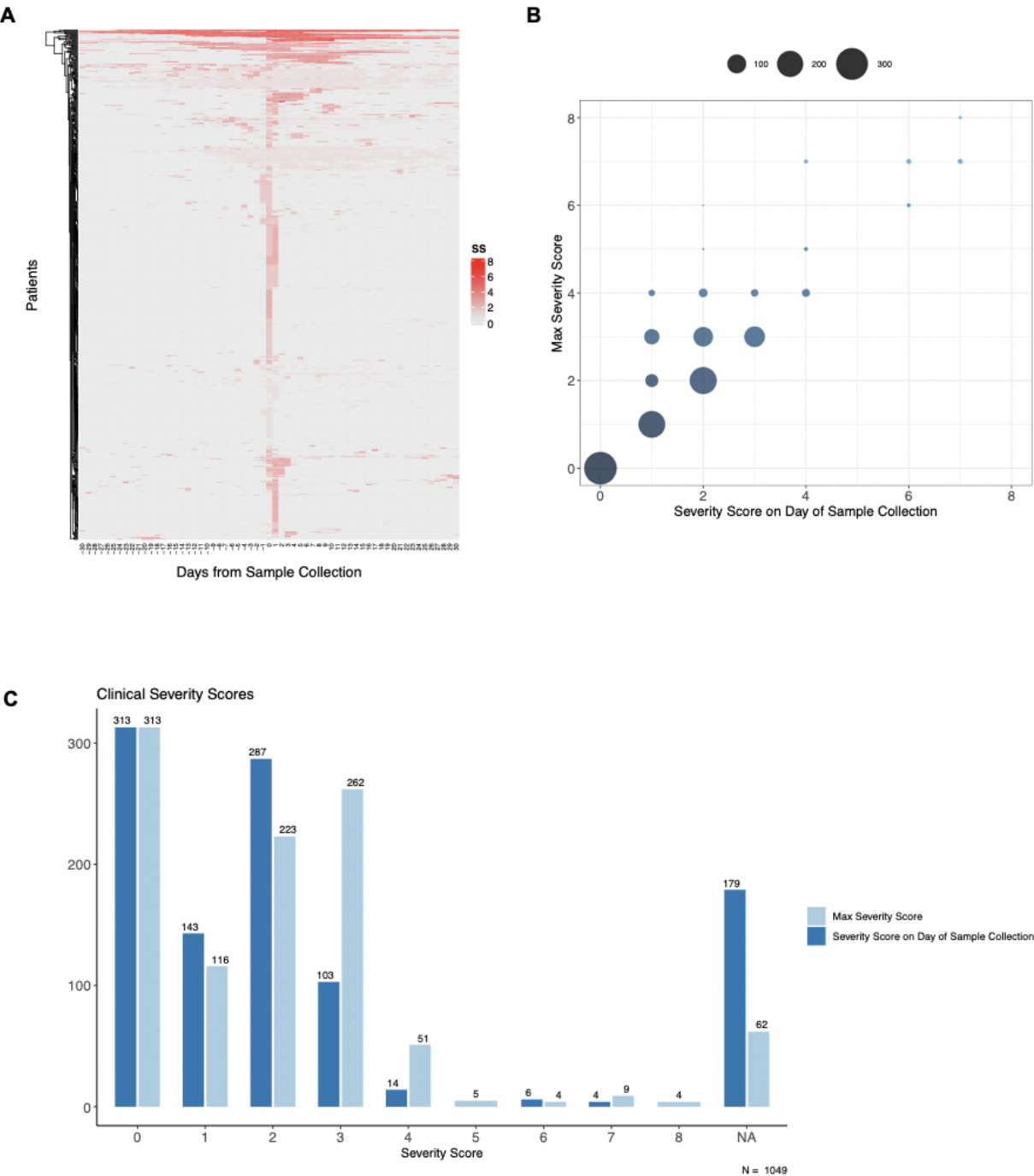


Figure S2. Clinical Characteristics of COVID19 positive individuals versus negative controls.

### Supplementary Figure 3



**Figure S3. Severity score.** (A) Variance in severity score for 30 days before and after NP swab collection for all samples. (B) Correlation between severity on day of sample collection and highest severity (Max Severity Score) in the 30 days before and after swab collection. (C) Redistribution of severity scores in the 30 days before and after collection vs day of collection.

