medRxiv preprint doi: https://doi.org/10.1101/2020.07.27.20163147; this version posted July 29, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

High-throughput SARS-CoV-2 and host genome sequencing from single nasopharyngeal swabs

J. E. Gorzynski[1,2,*], H. N. De Jong[1,2,*], D. Amar[2], C. Hughes[1], A. Ioannidis[3], R. Bierman[3], D. Liu[3], Y. Tanigawa[3], A. L. Kistler[4], J. Kamm[4], J, Kim[5], L. Cappello[6], N. F. Neff[3], S. Rubinacci[7], O. Delaneau[7], M. J. Shoura[1,8], K. Seo[2,], A. Kirillova[2], A. Raja[2], S. Sutton[2], C. Huang[8], M. K. Sahoo[8], K. C. Mallempati[9], G. Montero-Martin[9], K. Osoegawa[9], N. Watson[10], D. Jimenez-Morales[2], N. Hammond[10], R. Joshi[10], M. A. Fernández-Viña[6,8], J. W. Christle[2], M.T. Wheeler[2], P. Febbo[11], K. Farh[11], G. P. Schroth[11], F. DeSouza[11], J. Palacios[3,6], J. Salzman[3], B. A. Pinsky[8,12], M. A. Rivas[3], C.D. Bustamante[1,3], E. A. Ashley[1,2*], V. N. Parikh[2*]

1. Department of Genetics, Stanford University, Stanford CA
2. Department of Medicine, Division of Cardiovascular Medicine, Stanford University School of Medicine, Stanford, CA
3. Biomedical Data Science, Stanford University, Stanford, CA
4. Chan Zuckerburg Biohub, San Francisco, CA
5. Department of Biology, Stanford University
6. Department of Statistics, Stanford University
7. Department of Computational Biology, and Swiss Institute of Bioinformatics (SIB), University of Lausanne, Lausanne, Switzerland
8. Department of Pathology, Stanford University School of Medicine, Stanford, CA
9. Histocompatibility & Immunogenetics Laboratory, Stanford Blood Center, Palo Alto, CA
10. Clinical Genomics Program, Stanford Medicine, Palo Alto, CA
11. Illumina, Inc., San Diego, CA
12. Department of Medicine, Division of Infectious Diseases and Geographic Medicine, Stanford University School of Medicine, Stanford, CA

* These authors contributed equally to this manuscript

Abstract

During COVID19 and other viral pandemics, rapid generation of host and pathogen genomic data is critical to tracking infection and informing therapies. There is an urgent need for efficient approaches to this data generation at scale. We have developed a scalable, high throughput approach to generate high fidelity low pass whole genome and HLA sequencing, viral genomes, and representation of human transcriptome from single nasopharyngeal swabs of COVID19 patients.

NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.

medRxiv preprint doi: https://doi.org/10.1101/2020.07.27.20163147; this version posted July 29, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

Respiratory virus pandemics, most recently COVID-19 caused by SARS-CoV-2, have caused devastating loss of life and crippled health care systems worldwide. Our public health response to and recovery from these catastrophes depends on the speed and agility with which we generate both viral and host sequencing data. In recent pandemics, viral sequencing has been crucial, but limited by low throughput and high expense. Collection of concomitant host genomics can inform familial relationship tracking, examination of underlying genetic risk, and identify ancestries at elevated risk, but has been encumbered by need for multiple sampling.[1] Therefore, there is an urgent need for protocols to produce these data in real time and at scale. Here, we describe a method to achieve simultaneous viral and host sequencing from single SARS-CoV-2 diagnostic nasopharyngeal swab residuals **(Figure 1)**. We use low-pass whole host genome sequencing as an alternative to array-based genotyping to provide rich information for trait mapping at scale[2,3] and demonstrate that our method regularly provides DNA of sufficient quality for host genome, and HLA sequencing. Further, we present a high-throughput RNAseq workflow for sequencing full viral genomes and human transcriptome reads from hundreds of samples concomitantly. Finally, we describe how this method creates a strong multi-omic foundation for data integration and sharing across global institutions.

In this study, residual viral transport media (VTM) from SARS-CoV-2 clinical diagnostic tests were collected. Institutional Review Board approval for anonymous sequencing of host and viral genomics was obtained from the Stanford University School of Medicine IRB. Positive samples are those that had a crossing threshold (CT) of 40 cycles or less on the RT-qPCR diagnostic tests used at Stanford Health Care clinical laboratory.[4] Virus was inactivated by diluting each sample with one-third volume lysis buffer. Using silica membrane/ethanol nucleotide extraction protocols, we recovered RNA (mean mass: 112ng, 0-1100ng) and DNA (mean mass: 200ng, 0-4700ng) from aliquots of VTM. Seventy percent of samples yielded a total DNA mass greater than 100ng, providing enough input material for many high depth whole genome sequencing protocols.

Multiplexed shotgun sequencing libraries were prepared for RNA and DNA. Purified genomic DNA or cDNA was enzymatically fragmented and unique index adapters were added to the ends by PCR. Fluorometric quantification and capillary electrophoresis fragment size analysis determined the molarity of samples. Samples were pooled at equimolar ratios in batches of 160. For DNA libraries, prior to high-throughput sequencing, low pass sequencing determined sample representation within the pool. Samples were re-pooled if necessary to achieve a balanced library. Pools of 160 genomic DNA or cDNA samples underwent paired-end sequencing by synthesis (2x150bp) using an Illumina NovaSeq and S4 (DNA) or S2 (RNA) flow cell (300 cycles). 117 samples from which more than 2ng/ul of DNA were recovered underwent human leukocyte antigen (HLA) sequencing using at least 20 ng of total sample. In brief, libraries were sequenced for 11 classical HLA genes prepared from batched host genomic DNA samples. Sequences were assembled and assigned to HLA genotypes and used to generate reports for HLA genotypes, HLA serotypes, and imputed HLA haplotypes.

medRxiv preprint doi: https://doi.org/10.1101/2020.07.27.20163147; this version posted July 29, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

We aligned and called host genomic variants using methods for low pass sequencing [2,5,6] followed by imputation with GLIMPSE.[3] Alignment revealed a mean of 98±3.6% DNA reads mapped to the host genome (GRCh38 assembly), with an average of 61% of the genome covered at 1X, 52% covered at 2X, and 30% covered at 3X. The autosomal mean of medians was 2.21±0.58X, median 2.16X (**Figure 2A**). Variant calling and imputation yielded high fidelity data, as demonstrated by ancestry classification consistent with reference genomes (**Figure 2B)**. This also enabled confirmation of six blindly duplicated samples and 2 pairings of first-degree relatives through kinship analysis (**Figure 2C**). Together, these analyses demonstrate the robustness and internal reproducibility of host genome sequencing, variant calling and genotype imputation. We recovered interpretable HLA sequencing from 85% of 116 samples attempted. The remaining samples showed either no amplification of HLA sequences or non-specific amplification. To understand the reason for sample failure, we tested the hypothesis that failed samples had DNA concentration too low to amplify. We found that there was no correlation between DNA concentration and detection (p=0.07).

RNA sequence was aligned to human transcriptome Hg19, SARS-CoV-2 (MN908947.3 (GenBank)/NC_045512.2 (RefSeq)) and 40 other human viral pathogens **(Supplemental Table 1)** using STAR, kraken2 plus minimap2, and Illumina BaseSpace DRAGEN Pathogen Detection, respectively. RNA sequencing revealed 12.4±22.7% aligned to the SARS-CoV-2 genome. The DRAGEN pipeline identified none of the 160 samples to have 5X coverage at more than 7% of any other viral pathogen genome, indicating no significant co-infections were observed among these samples. On average, 3.9±10.3% of reads aligned to the human genome.

Fifty-two percent of samples had at least 10X coverage at >95% of the SARS-CoV-2 genome. We show that CT number from the Stanford clinical RT-PCR test correlates with viral genome coverage, as well, with a steep drop off in viral genome coverage at CT values of 26 and higher (R^2=0.61, p=0.0001, **Figure 2D**). Greater than 95% of the genome was covered in 95.3% samples with a diagnostic CT value equal or less than 25. In samples with a CT less than 30, 77% yielded >95% 10X of the viral genome. Three outliers with low CT numbers (high viral load) but low viral genome coverage had low RNA yield from extraction, though overall, total RNA yield did not correlate with viral coverage (R^2=0.004, p=0.39, **Figure 2D**). Aligned human reads were not inversely correlated with viral genome coverage (R^2=0.0001, p=0.9), indicating that drop out of viral genome coverage is likely due to other factors, sample contamination or degradation. However, as only three samples with CT numbers below 25 did not yield full viral sequences, we recommend proceeding to library preparation regardless of RNA yield. Using the consensus sequences derived from the initial cohort reported here as well as samples collected later in March 2020, we created a phylogenetic tree, which allows critical public health phylodynamic tracking (**Figure 2E**). Further, we demonstrate high correlation between CT range and detection of all known SARS-CoV-2 genes (**Figure 2F**, R^2= 0.50, p=5.2e-262). To distinguish reads from the viral genome versus the viral transcriptome (called “sub-genomic” or sgRNA), we also quantified reads containing splicing junctions (which exist solely in sgRNA) and found the same relationship (**Supplemental Figure 1**).

medRxiv preprint doi: https://doi.org/10.1101/2020.07.27.20163147; this version posted July 29, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

Here we demonstrate that a single nasopharyngeal swab can reveal substantial host *and* viral genomic information in a high-throughput manner that will facilitate public health pandemic tracking and research into the mechanisms underlying virus-host interactions. Certainly, nasopharyngeal swabs have previously been used to perform whole viral genome sequencing of respiratory viruses in low throughput.[7,8,9] Our method accelerates this discovery both in terms of time and number of subjects sequenced: compared to these reports, we show a comparable rate of viral genomic coverage with the capability of studying at least 10 times the number of samples in a single sequencing run. Although our initial swab collection did not reveal any viral co-infections, especially as the current pandemic enters the regular flu and cold season, our method allows for acceleration of metagenomics analysis.[9,10] Further, advances in low pass genome calling allow the same nasopharyngeal swab to be used to gather a wealth of human genomics data and in many cases yielded enough DNA for deep sequencing of HLA type, which is a critical component of the host immunomic response. [2,3]

Future infectious disease outbreaks will inevitably occur, and the strategy we describe is applicable for collection of host and viral genomic information from any respiratory virus in laboratories around the world. Perhaps the most critical application of this workflow is that it enables the rapid development of large scale, multicenter/global host and viral multi-omic data repositories. Global data repositories have been critical to advancing research in the current pandemic and prior. For example, just six months into the SARS-CoV-2 pandemic, nearly 70,000 submissions to the Global Initiative for Sharing All Influenza Data (GISAID) allowed unprecedented tracking of viral mutagenesis and outbreaks.[11] With the method we propose here, the same number of viral genomes could be produced by less than 100 sequencing centers within weeks, along with matched host genome, transcriptome and HLA typing. These findings could be easily incorporated with data abstracted from the electronic health care record (as is being accomplished with increasing speed[12,13]), and mobile digital reporting platforms[14] (**Supplemental Figure 2**). The methods described here represent a crucial scaffold for the integration of these complex inputs to centralized data repositories, enabling unprecedented rapidity of the discovery and implementation necessary to overcome these devastating pandemics.

medRxiv preprint doi: https://doi.org/10.1101/2020.07.27.20163147; this version posted July 29, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

Figures

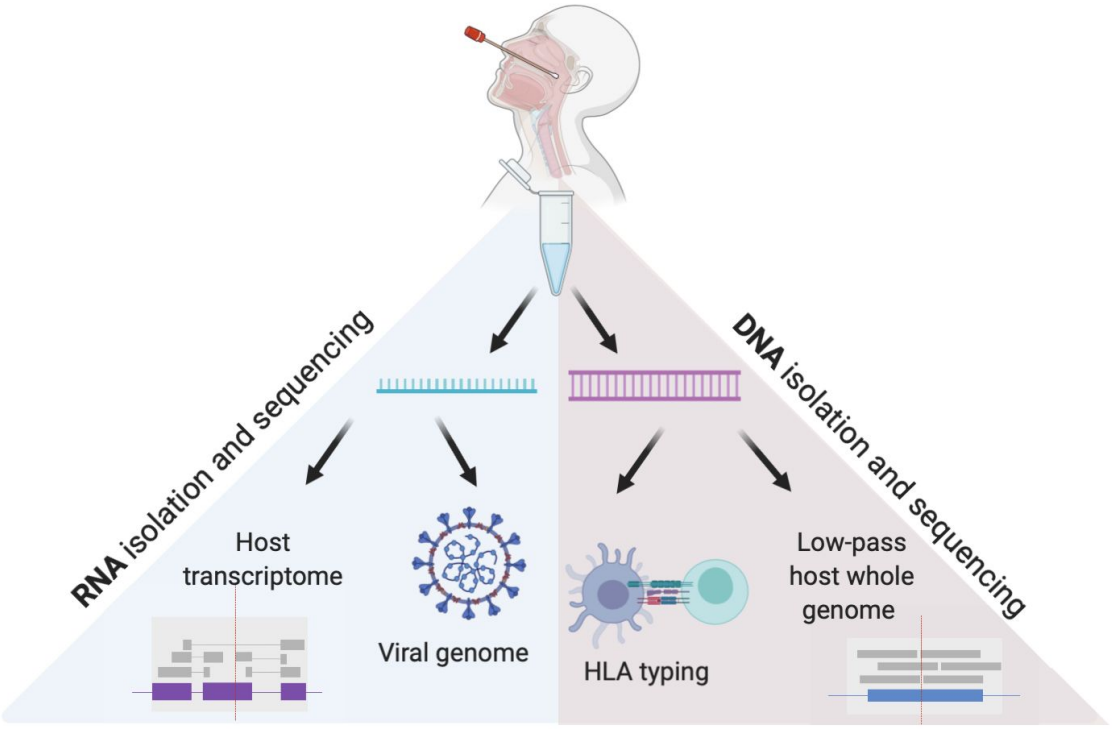


Figure 1. Viral and Host genomes and transcriptomes from a single nasopharyngeal swab. This method allows for independent RNA and DNA isolation from nasopharyngeal swab VTM, enabling viral genome sequencing, detection of host transcriptome, low pass host genome sequencing and HLA sequencing in high throughput.

medRxiv preprint doi: https://doi.org/10.1101/2020.07.27.20163147; this version posted July 29, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

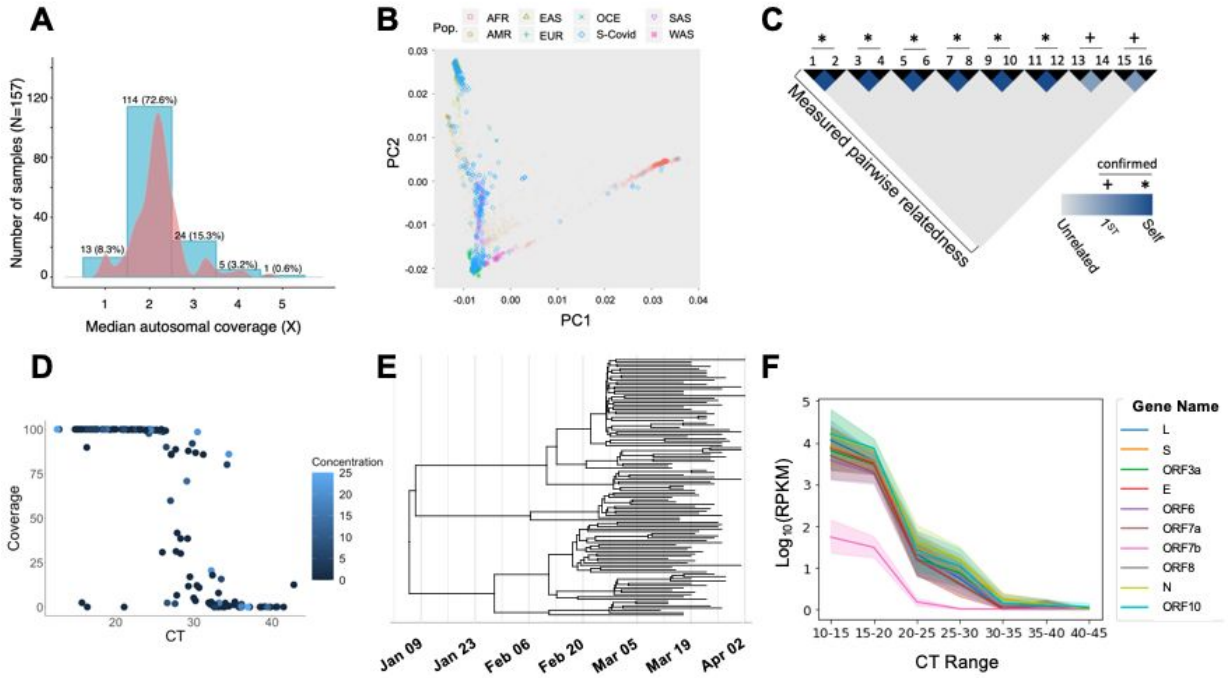


Figure 2. High fidelity, rapid throughput SARS-CoV-2 genome, transcriptome and low pass whole genome sequencing from single nasopharyngeal swabs. (A) For low pass genomes, an average of 61% of the human genome was covered at 1X, 52% covered at 2X, and 30% covered at 3X.The autosomal mean of median coverages was 2.21±0.58. **(B)** Principal components analysis shows low pass variant calling and imputation reliably predict ancestry in COVID19 samples (blue diamonds) as compared to reference genomes (ancestries as indicated in legend: AFR - African, AMR - Native American, EAS- East Asian, EUR - European, OCE - Oceanian, SAS - South Asian, WAS - West Asian). **(C)** Blinded duplicates and first degree relatives are predicted by variant calling and imputation of low pass host genomes. **(D)** Percent reads mapped to Hg19 human transcriptome from each sample demonstrating an average of 3.9±10.3% reads from the host. **(E)** Phylogenetic tree of 114 whole viral genomic consensus sequences sampled in March-April 2020. The estimated phylogenetic tree is the maximum clade credibility tree obtained with BEAST 2[15] using a fixed mutation rate of 1.04×10^{-3} per base per year, the Coalescent Extended Bayesian Skyline prior[16], and the HKY substitution model[17]. **(F)** RPKMs for individual SARS-CoV-2 genes were averaged over samples with similar CT values. All gene expression increases with viral load (CT range), and all have concordant normalized abundances, with the exception of ORF7b.

medRxiv preprint doi: https://doi.org/10.1101/2020.07.27.20163147; this version posted July 29, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

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. 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.

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(Protocol including modifications found in Sup. Material, Google Doc). Total RNA was extracted from 200ul of VTM using a modified Ambion MirVana mRNA kit protocol (Protocol including modifications found in Sup. Material, Google Doc) 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 (Protocol including modifications found in Sup. Material, Google Doc). 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; a modification of Deng et al. (2020)[18]). 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.

medRxiv preprint doi: https://doi.org/10.1101/2020.07.27.20163147; this version posted July 29, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

Viral Genome Alignment

Metatranscriptomic FASTQ sequences were aligned to the SARS-CoV-2 reference genome NC_045512.2 using minimap2[19]. Non-SARS-CoV-2 reads were filtered out with Kraken2[20], 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[21]. Finally, a pileup of the aligned reads was generated with samtools[22], and consensus genomes were called with ivar. The full pipeline used is publicly available on Github (https://github.com/czbiohub/sc2-illumina-pipeline).

For multiple viral genome alignment, FASTQ files were input into the Illumina BaseSpace DRAGEN Pathogen Detection software with the following parameters: Somatic small variant base caller, k-mers generated from reference (SARSCoV2_NC_045512.2), Minimum depth 10, minimum allele frequency 0.5, virus detected threshold 5% of genome at 5X coverage.

Quantification of total Viral RNA and sgRNA

RNA mapping was performed using STAR run against a combined index of grch38, SARS-CoV2, and ERCC spike ins. STAR parameters were chosen to avoid bias towards GTAG eukaryotic splice signatures for both the viral RNA and sgRNA analyses [23]. The STAR parameters used are included in **Supplemental Table 2.** Reads per COVID gene were collected from the ReadsPerGene STAR output file, and the total mappable reads were collected from the Log.final files. Gene lengths were calculated using the annotated start and stop positions in the annotation file (https://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/009/858/895/GCF_009858895.2_ASM985889v3/GCF_009858895.2_ASM985889v3_genomic.gff.gz). Read counts, gene length, and total mappable reads were together used to calculate the RPKM per gene within each sample.

The spliced RNA junction software SICILIAN[24] was used to sensitively and accurately identify leader-body junctions within each sample. Modifications to standard SICILIAN protocol were to the combined index and STAR mapping parameters described above (**Supplemental Table 2**).[23]

Host HLA Sequencing: 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 µl while maintaining the manufacturer's recommended multiplex PCR protocol per sample; 2) eluted DNA in 12 µl 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 µl for DNA sequencing. 500 µl of 1.3 pM DNA sequencing library was loaded into a MiniSeq Mid Output Kit (300-cycles) (FC-420-1004), and

medRxiv preprint doi: https://doi.org/10.1101/2020.07.27.20163147; this version posted July 29, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

sequenced using MiniSeq DNA sequencer (Illumina Inc., San Diego, CA). 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.[25] 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,[26] 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.[27,28]

Host 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: XXX (https://www.protocols.io/private/8CFBD1AD8FE611EA815E0A58A9FEAC2A)

Variant Calling, Imputation, PCA, Kinship: BAM files were used for an initial calling with bcftools v1.9 mpileup[29]. To account for the low-coverage sequencing we used the GLIMPSE algorithm v1.0 for imputation[3]. Briefly, this algorithm uses a reference set of haplotypes (1000 genomes 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 HGDP, and (3) the SGDP. 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)[30,31]. 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)[32].

Acknowledgments

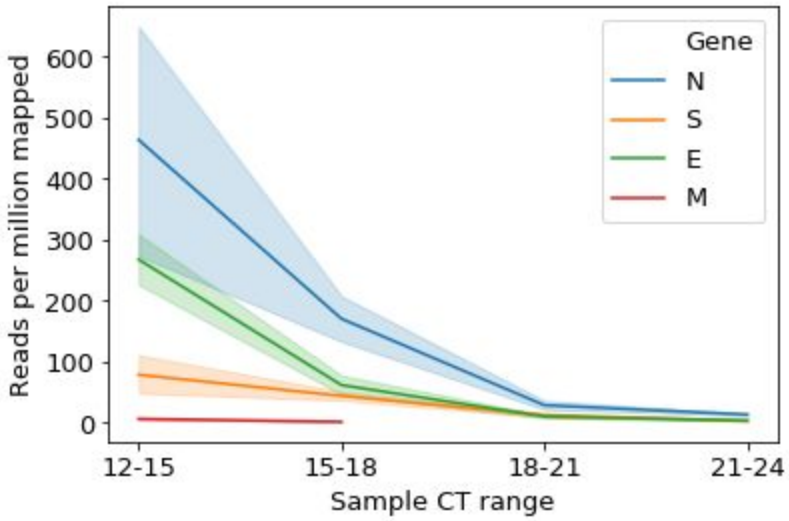
Takeda Pharmaceuticals provided funding support for nucleic acid extraction and library preparation. One Lambda, Inc supplied HLA sequencing reagents and software. The Chan Zuckerberg Initiative and Biohub supported all viral genome sequencing and alignment. This work is supported in part by NIH Common Fund (U24OD026629 to EAA and MTW), Arnold O.

medRxiv preprint doi: https://doi.org/10.1101/2020.07.27.20163147; this version posted July 29, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

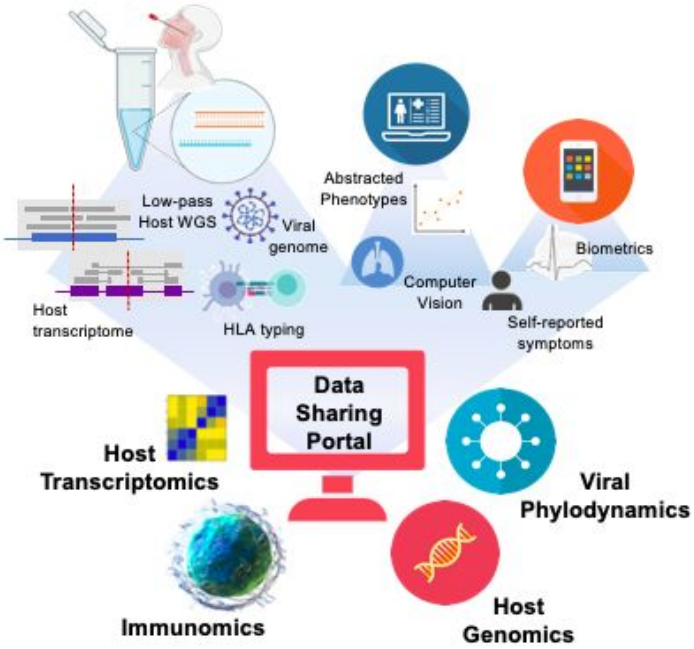
Beckman independence fellowship (MJS) and American Heart Association (MJS, KS), the National Heart Lung and Blood Institute (K08HL143185 to VNP), The John Taylor Babbitt Foundation (VNP) and Sarnoff Cardiovascular Research Foundation (VNP).

medRxiv preprint doi: https://doi.org/10.1101/2020.07.27.20163147; this version posted July 29, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

Supplemental Material



Supplemental Figure 1. sgRNA expression is correlated with CT range. Patient samples with lower CT values had significantly higher sub-genomic RNA (sgRNA) reads per million mapped (RPM) for genes E (R^2= 0.23, p=1.4e-17), M (R^2=0.17 , p=1.0e-03), and N (R^2=0.02 , p=5.1e-03). These results show that our method is capable of sequencing SARS-CoV2 viral transcripts as well as genomes. Of note, proteins M and S were observed only in low-CT value samples.



Supplemental Figure 2. A strong multi-omic foundation for data integration and sharing across global institutions. Using these methods in combination with electronic health record abstraction, and digital medicine, the methods described here builds the foundation for a data repository allowing rapid access to critical data on CoVID19 or any other pandemic via open-source sharing.

medRxiv preprint doi: https://doi.org/10.1101/2020.07.27.20163147; this version posted July 29, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

SARS-CoV-2 coverage
SARS-CoV-2
SARS coronavirus
Influenza B virus (B/Brisbane/60/2008)
Influenza A virus (A/Michigan/45/2015(H1N1))
Influenza A virus (A/Texas/50/2012(H3N2))
Influenza B virus (B/Wisconsin/01/2010)
Influenza B virus (B/Washington/02/2019)
Human coronavirus 229E
Human coronavirus NL63
Human adenovirus B1
Human parainfluenza virus 3
Human coronavirus OC43
KI polyomavirus Stockholm 60
Human adenovirus E4
Respiratory syncytial virus (type A)
Human bocavirus 2c PK isolate PK-5510
Human Respiratory syncytial virus 9320 (type B)
Human adenovirus C2
Human bocavirus 4 NI strain HBoV4-NI-385
Human parainfluenza virus 1
Influenza A virus (A/Hong Kong/1073/99(H9N2))
Human coronavirus HKU1
Human parainfluenza virus 4a
Human rhinovirus A89
Human parechovirus 6
Human bocavirus 1 (Primate bocaparvovirus 1 isolate st2)
Human parainfluenza virus 2
Influenza A virus (A/Puerto Rico/8/1934(H1N1))

medRxiv preprint doi: https://doi.org/10.1101/2020.07.27.20163147; this version posted July 29, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

Human rhinovirus B14
Human enterovirus C104 strain: AK11
WU Polyomavirus
Human bocavirus 3
Influenza B virus (B/Lee/1940)
Human rhinovirus C (strain 024)
Influenza A virus (A/goose/Guangdong/1/1996(H5N1))
Influenza A virus (A/Korea/426/1968(H2N2))
Human enterovirus C109 isolate NICA08-4327
Human metapneumovirus (CAN97-83)
Influenza A virus (A/New York/392/2004(H3N2))
Influenza A virus (A/Zhejiang/DTID-ZJU01/2013(H7N9))
Human parechovirus type 1 PicoBank/HPeV1/a

Supplemental Table 1. List of viral genomes assessed by alignment using DRAGEN in Illumina BaseSpace.

RNA quantification	STAR mapping parameters
Total viral RNA and sgRNA identification	--outFilterMultimapNmax 20 --alignSJoverhangMin 8 --outSJfilterOverhangMin 12 12 12 12 --outSJfilterCountUniqueMin 1 1 1 1 --outSJfilterCountTotalMin 1 1 1 1 --outSJfilterDistToOtherSJmin 0 0 0 0 --outFilterMismatchNmax 999 --outFilterMismatchNoverReadLmax 0.04 --scoreGapNoncan -4 --scoreGapATAC -4 --chimSegmentMin 10 --chimOutType WithinBAM SoftClip Junctions --chimScoreJunctionNonGTAG 0 --alignSJstitchMismatchNmax -1 -1 -1 -1 --alignIntronMin 20 --alignIntronMax 1000000 --alignMatesGapMax 1000000

Supplementary Table 2. STAR mapping parameters used for detection of total viral RNA and sgRNA.

medRxiv preprint doi: https://doi.org/10.1101/2020.07.27.20163147; this version posted July 29, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

References

1. Ciancanelli, M. J., Abel, L., Zhang, S.-Y. & Casanova, J.-L. Host genetics of severe influenza: from mouse Mx1 to human IRF7. *Curr. Opin. Immunol.* **38**, 109–120 (2016).
2. 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).
3. 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.
4. 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.
5. Garcia, M. *et al.* Sarek: A portable workflow for whole-genome sequencing analysis of germline and somatic variants. *F1000Res.* **9**, 63 (2020).
6. Poplin, R. *et al.* Scaling accurate genetic variant discovery to tens of thousands of samples. bioRxiv 201178. *Publisher Full Text* (2017).
7. Cotten, M. *et al.* Transmission and evolution of the Middle East respiratory syndrome coronavirus in Saudi Arabia: a descriptive genomic study. *Lancet* **382**, 1993–2002 (2013).
8. Huang, W. *et al.* Whole-Genome Sequence Analysis Reveals the Enterovirus D68 Isolates during the United States 2014 Outbreak Mainly Belong to a Novel Clade. *Sci. Rep.* **5**, 15223 (2015).
9. Imai, K. *et al.* Whole Genome Sequencing of Influenza A and B Viruses With the MinION Sequencer in the Clinical Setting: A Pilot Study. *Frontiers in Microbiology* vol. 9 (2018).
10. Lopez, S. M. C. *et al.* A method of processing nasopharyngeal swabs to enable multiple

medRxiv preprint doi: https://doi.org/10.1101/2020.07.27.20163147; this version posted July 29, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

testing. *Pediatr. Res.* **86**, 651–654 (2019).

11. GISAID - Initiative. https://www.gisaid.org/.
12. Callahan, A. *et al.* Medical device surveillance with electronic health records. *NPJ Digit Med* **2**, 94 (2019).
13. Ouyang, D. *et al.* Video-based AI for beat-to-beat assessment of cardiac function. *Nature* **580**, 252–256 (2020).
14. Turakhia, M. P., Desai, M., Perez, M. V. & Apple Heart Study Investigators. A Smartwatch to Identify Atrial Fibrillation. Reply. *The New England journal of medicine* vol. 382 975–976 (2020).
15. Bouckaert, R. *et al.* BEAST 2.5: An advanced software platform for Bayesian evolutionary analysis. *PLoS Comput. Biol.* **15**, e1006650 (2019).
16. Heled, J. & Drummond, A. J. Bayesian inference of population size history from multiple loci. *BMC Evol. Biol.* **8**, 289 (2008).
17. 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).
18. Deng, X. *et al.* Metagenomic sequencing with spiked primer enrichment for viral diagnostics and genomic surveillance. *Nat Microbiol* **5**, 443–454 (2020).
19. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).
20. Wood, D. E., Lu, J. & Langmead, B. Improved metagenomic analysis with Kraken 2. *Genome Biol.* **20**, 257 (2019).
21. 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).
22. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* vol. 25

medRxiv preprint doi: https://doi.org/10.1101/2020.07.27.20163147; this version posted July 29, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

2078–2079 (2009).

23. Kim, D. *et al.* The Architecture of SARS-CoV-2 Transcriptome. *Cell* **181**, 914–921.e10 (2020).
24. Dehghannasiri, R., Olivieri, J. E. & Salzman, J. Specific splice junction detection in single cells with SICILIAN. *bioRxiv* 2020.04.14.041905 (2020) doi:10.1101/2020.04.14.041905.
25. Robinson, J. *et al.* The IPD and IMGT/HLA database: allele variant databases. *Nucleic Acids Res.* **43**, D423–31 (2015).
26. 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).
27. 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).
28. Osoegawa, K. *et al.* HLA Haplotype Validator for quality assessments of HLA typing. *Hum. Immunol.* **77**, 273–282 (2016).
29. 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).
30. Chang, C. C. *et al.* Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* **4**, 7 (2015).
31. 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).
32. Manichaikul, A. *et al.* Robust relationship inference in genome-wide association studies. *Bioinformatics* **26**, 2867–2873 (2010).

medRxiv preprint doi: https://doi.org/10.1101/2020.07.27.20163147; this version posted July 29, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

medRxiv preprint doi: https://doi.org/10.1101/2020.07.27.20163147; this version posted July 29, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

Author Contributions	Writing	Data Visualisation	Data Analysis	Exp Design	Sample Processing (extraction,lib prep, seq)	Sample Collection	Leadership	Email
J Gorzynski	XXX	XXX	XXX	X	XXX			jgorz@stanford.edu
H De Jong	XXX	XXX	XXX	X	XXX			hnd@stanford.edu
V Parikh	XXX	XXX	XXX	XXX	X	XX	XXX	vparikh@stanford.edu
A Kistler			X	X	XX			amy.kistler@czbiohub.org
M Shoura	X		X		X			massa86@stanford.edu
R Joshi					X			rjoshi@stanfordhealthcare.org
N Watson					X			https://orcid.org/0000-0003-4547-3285
N Hammond					X			https://orcid.org/0000-0001-9635-2294
C Hughes	X	XX	X					crhughes@stanford.edu
K Seo					X			kseo1@stanford.edu
A Kirillova					X			akirillova00@gmail.com
D Amar	X	XX	X					davidama@stanford.edu
D Jimenez-morales								david.jimenez-morales@stanford.edu
A Raja			X					araja7@stanford.edu
J. W. Christle						X		
M Wheeler						X	X	wheelerm@stanford.edu
C Huang						XX		chhuang@stanford.edu
M Sahoo						XX		mksahoo@stanford.edu
K Mallempati			XX	X	XX			kalyanm@stanford.edu
G			XX	X	XX			gmonter2@stanford.edu

medRxiv preprint doi: https://doi.org/10.1101/2020.07.27.20163147; this version posted July 29, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

Montero-Martin								
K Osoegawa	X		X	X				kazutoyo@stanford.edu
M **Fernández-Viña**			X				X	marcelof@stanford.edu
B Pinsky						XX	X	
M Rivas				X			X	mrivas@stanford.edu
C Bustamante				X			X	cdbustam@stanford.edu
Y tanigawa				X				
E Ahsley				XXX			XXX	euan@stanford.edu
G P Schroth				XXX			X	gschroth@illumina.com
N Neff				X	X			norma.neff@czbiohub.org
J Kamm			X		X			jack.kamm@czbiohub.org