Appendix: SARS-CoV-2 Whole Genome Sequencing Methods

Viral RNA Extraction

Viral RNA was extracted from clinical specimens utilizing the QIAamp 96 Virus QIAcube HT Kit (Qiagen, cat. no. 57731). Clinical testing for SARS-CoV-2 presence was performed by quantitative reverse transcription and PCR (qRT-PCR) with the CDC 2019-nCoV RT-PCR Diagnostic Panel utilizing the N1 probe in SARS-CoV-2 and RP probes for sample quality control as previously described (IDT, cat. no. 10006713). All specimens that failed to amplify the RP housekeeping gene were excluded from this study. All specimens with an N1 probe cycle threshold (Ct) less than or equal to 35 were considered positive and included in this study. RT-PCR was performed on all specimens to validate Ct values obtained by the clinical diagnostic laboratory. Ct values from the N1 probes were used in all subsequent analyses.

cDNA Synthesis and Viral Genome Amplification

cDNA synthesis was performed with SuperScript IV First Strand Synthesis Kit (ThermoFisher, cat. no. 18091050) using 11 µl of extracted viral nucleic acids and random hexamers according to manufacturer's specifications. Direct amplification of the viral genome cDNA was performed in multiplexed PCR reactions to generate ~400 bp amplicons tiled across the genome. The multiplex primer set, comprised of two non-overlapping primer pools, was created using Primal Scheme and provided by the Artic Network (version 3 and 4 releases). PCR amplification was carried out using Q5 Hot Start HF Taq Polymerase (NEB, cat. no. M0493L) with 5 µl of cDNA in a 25 µl reaction volume. A two-step PCR program was used with an initial step of 98 °C for 30 s, then 35 cycles of 98 °C for 15 s followed by five minutes at 64 °C. Separate reactions were carried out for each primer pool and validated by agarose gel electrophoresis alongside negative controls. Each reaction set included positive and negative amplification controls and was performed in a space physically separated for pre- and post-PCR processing steps to reduce contamination. Amplicon sets for each genome were pooled prior to sequencing library preparation.

Sequencing Library Preparation, Illumina Sequencing, and Genome Assembly

Sequencing library preparation of genome amplicon pools was performed using the SeqWell plexWell 384 kit per manufacturer's instructions. Pooled libraries of up to 96 genomes were sequenced on the Illumina MiSeq using the V2 500 cycle kit. Sequencing reads were trimmed to remove adapters and low-quality sequences using Trimmomatic v0.36. Trimmed reads were aligned to the reference genome sequence of SARS-CoV-2 (accession MN908947.3) using bwa v0.7.15. Pileups were generated from the alignment using samtools v1.9 and consensus sequence determined using iVar v1.2.2 with a minimum depth of 10, a minimum base quality score of 20, and a consensus frequency threshold of 0 (i.e. majority base as the consensus).[1] Consensus genome sequences were deposited in the GISAID public database (Table S1).

Bioinformatics Analyses

Genome sequences were aligned using MAFFT v7.453 software and manually edited using MEGA v6.06. We used the Pango classification scheme (https://cov-lineages.org/; PangoLearn version 9/28/2021) to identify the lineages that subsequently were grouped in variants using scorpio (scorpio v0.3.12). To confirm the clustering of the sequences by lineage, we inferred a Maximum Likelihood phylogeny with IQ-Tree v2.0.5 using its ModelFinder function before each analysis to estimate the nucleotide substitution model best-fitted for each dataset by means of Bayesian information criterion (BIC). We assessed the support of the clusters formed by the different lineages both with the Shimodaira–Hasegawa approximate likelihood-ratio test (SH-aLRT) and with ultrafast bootstrap (UFboot) with 1000 replicates each.

Reference

1. Grubaugh ND, Gangavarapu K, Quick J, et al. An amplicon-based sequencing framework for accurately measuring intrahost virus diversity using PrimalSeq and iVar. *Genome biology*. Jan 8 2019;20(1):8. doi:10.1186/s13059-018-1618-7