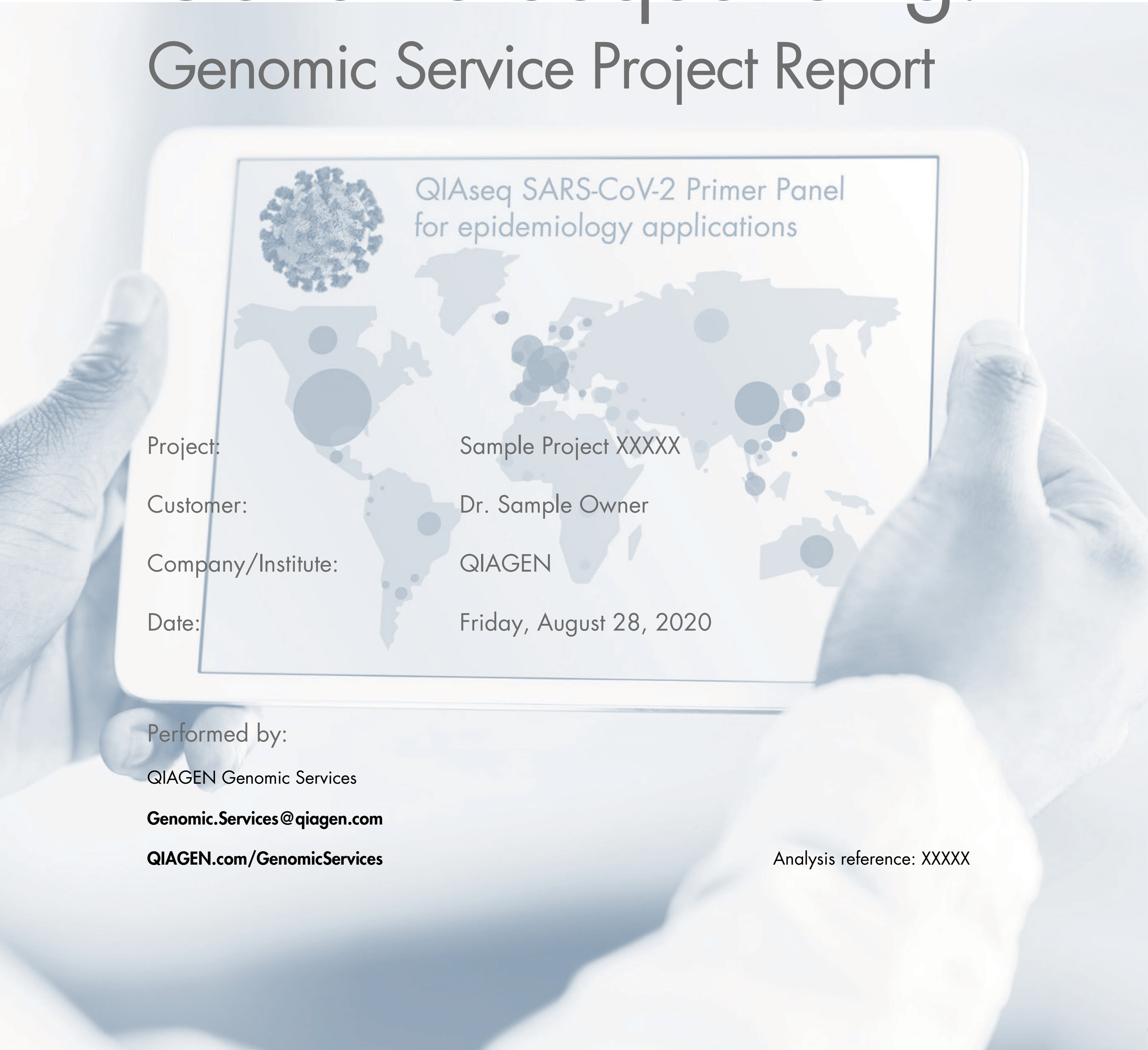
SARS-CoV-2 Whole Genome Sequencing:

Genomic Service Project Report



Project:	Sample Project XXXXX
Customer:	Dr. Sample Owner
Company/Institute:	QIAGEN
Date:	Friday, August 28, 2020

Performed by:

QIAGEN Genomic Services

Genomic.Services@qiagen.com

QIAGEN.com/GenomicServices

Analysis reference: XXXXX

Contents

Project Summary

Sample overview and metadata

Table 1. Example sample table.

Sample ID	Sample name	Condition
99999-001	Sample 1	Infected
99999-002	Sample 2	Infected
99999-003	Sample 3	Healthy control

Main findings and conclusions

Dear Customer,

We have now finalized the Next Generation Sequencing (NGS) analysis for the SARS-CoV-2 samples you have submitted to QIAGEN Genomic Services.

QIAseq® SARS-CoV-2 panel NGS libraries were successfully prepared, quantified, and sequenced for all your samples. The collected reads were subjected to quality control, alignment, and downstream analysis.

The principal structure of your result data is summarized in this report. In addition, the report contains details on the technical background.

We have computed the hypothetical phylogenetic relationship of the virus samples within this project. If the additional evolutionary relationships between the samples from this project and SARS-CoV-2 genomes from other publicly available databases (e.g., GISAID) is of interest, please let us know and we can perform more in-depth custom analyses for the samples. We also support exploring biological pathways that are affected in the host cells with Ingenuity Pathway Analysis.

If you have any questions, please contact your local QIAGEN representative or our Genomic Services lab scientists at **Genomic.Services@qiagen.com**.

Kind regards,

QIAGEN Genomic Services

Data Package Overview

All analysis was carried out using CLC Genomics Workbench (version 20.0.4) and CLC Genomics Server (version 20.0.4).

SARS-CoV-2 genome reference: Wuhan-Hu-1 (GenBank: MN908947.3)

Content	Description
FASTQ Quality control	**[1] Quality Control** QC report and supplementary QC report (per sample)
FASTQ adapter- and quality- trimming	**[2] Trimming** Trimming report (per sample)
Mapping Statistics and Variants	**[3] Mapping and Variants** Mapping report (per sample) Combined summary report (per sample) Excel table and VCF file for unfiltered variants (per sample) Excel table and VCF file for filtered variants (per sample)
Viral sequences	**[4] Consensus viral genome** Consensus viral genome FASTQ file (per sample) Consensus sequence annotation (per sample) Circular cladogram

Quality Control

The QC reports were generated by “QC for Sequencing Reads” tool from CLC Genomics Workbench, and can be found in **[1] Quality Control**, assess, and visualize statistics on: Sequence-read lengths and base-coverages, nucleotide-contributions and base-ambiguities, quality scores as emitted by the base-caller, over-represented sequences and hints suggesting contamination events. This information illustrates the overall sequencing quality, more specifically, an average PHRED score >30 is expected in the case of high-quality sequencing.

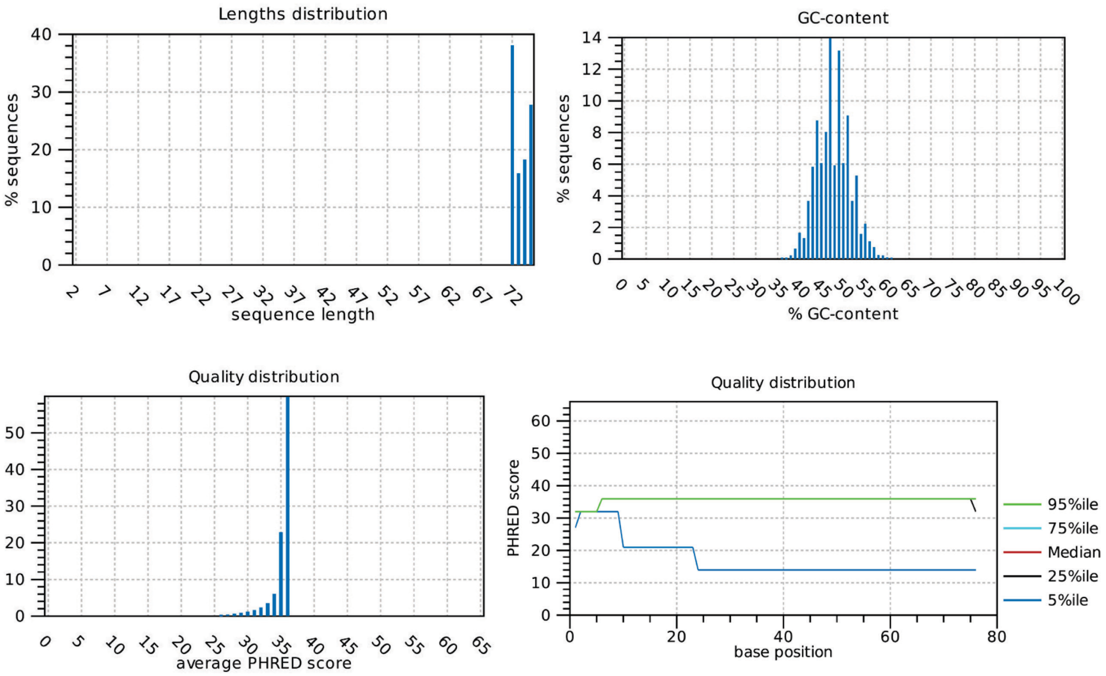


Figure 1. Example quality control plots for sequencing length distribution, GC content, average base qualities for reads, and quality distributions over read length (from top left to bottom right panel).

Adapter and Quality Trimming

Adapter and quality trimmings were done by “Trim Reads” tool from CLC Genomics Workbench. While adapters are often removed directly by the sequencer during base-calling, part of the adapter region may be included in the sequenced reads. Such adapters artefacts were removed by identifying read-through adapter sequences, whereby the 3′ end of one read includes the reverse complement of the adapter from the other read.

Further, reads were trimmed based on quality scores and ambiguous nucleotides, (e.g., due to stretches of Ns.) A maximum of 2 ambiguous nucleotides were allowed in a read. The trimming reports can be found in **[2] Trimming**. For SARS-CoV-2 Whole Genome Sequencing Services, the read length before and after trimming should not vary significantly with 2x150 paired end sequencing for the corresponding ~400 bp amplicons. However, because of cDNA amplicon fragmentation , there will be read-through adapters in some of the reads, meaning the actual DNA fragments (after trimming) are shorter than the original read lengths (before trimming). The auto-detected read-through adapter should start with 5'AGATCGGAAGAG for both reads. Another possible explanation for read lengths being significantly shorter than prior to trimming is low sequencing quality at the read ends, in which case that should be reflected in the QC report under **[1] Quality Control**.

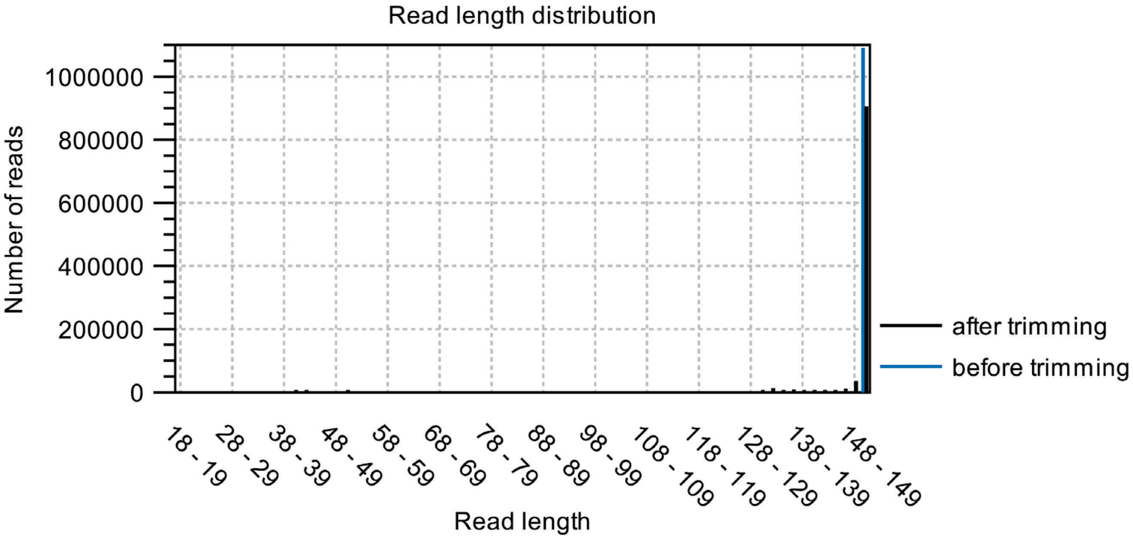


Figure 2. Example read length distribution before and after trimming.

Mapping and Variant Callings

Mapping and variant calling were performed using CLC Genomic Workbench with QIASeq SARS-CoV-2 workflow, which can be download freely from GitHub (**https://github.com/jonathanjacobs/CLC-Genomics**). In this workflow, the adapter- and quality- trimmed reads were aligned to the SARS-CoV-2 reference genome (MN908947.3) using "Map Reads to Reference" tool. Mapping statistics can be found under **[3] Mapping and Variants**. After aligning to the genome reference, sequence segments from the 5' and 3' ends of the read alignment corresponding to primer priming sites are removed to avoid false-negative and false-positive variants contributed by the input primers. The amplicon design scheme is shown in Figure 3 (pool 1 and pool 2) and the primer priming sites are located at the two ends of the amplicons.

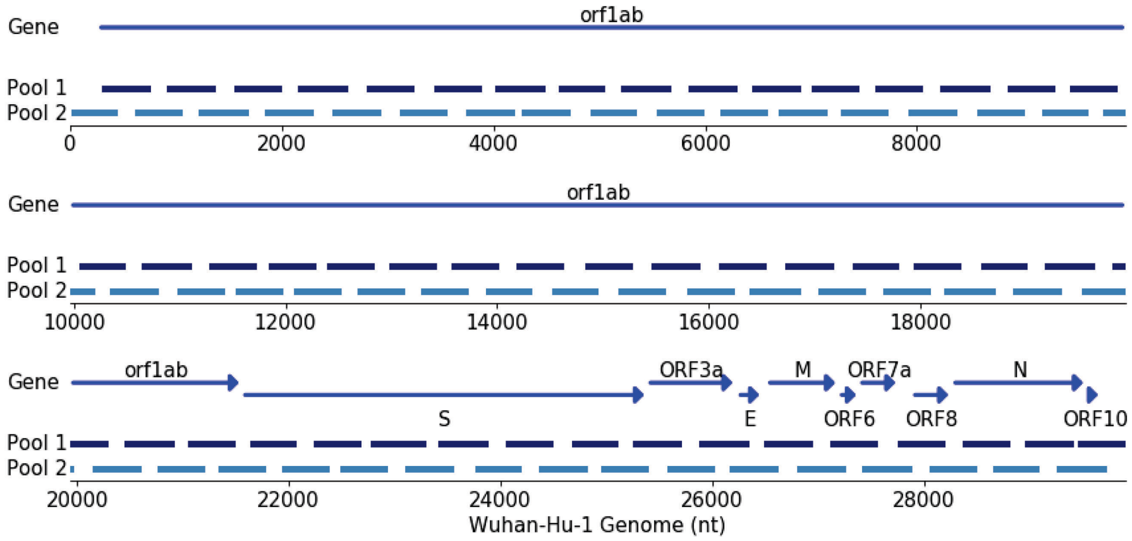


Figure 3. Amplicon scheme.

Variant calling is sensitive to alignment artefacts that can lead to incorrect reporting of variants at genome loci with insertions and deletions, such as homopolymer regions. To improve variant callings at these genomic loci, a local realignment step is performed using the "Local Realignment" tool. Variants were then identified from the refined read alignments with at least 5% variant frequency at positions with at least 100 read coverages. The lists of variants for each sample can be found as VCF files and Excel tables under **[3] Mapping and Variants**.

The variant table contains the following columns describing the details of each variant:

- **Chromosome**: The name of the reference sequence on which the variant is located.
- **Region**: The region on the reference sequence at which the variant is located. The region may be either a 'single position', a 'region' or a 'between position region'.
- **Type**: The type of variant. This can either be SNV (single-nucleotide variant), MNV (multi-nucleotide variant), insertion, deletion, or replacement.
- **Reference**: The reference sequence at the position of the variant.
- **Allele**: The identity of the variant.
- **Reference allele**: Describes whether the variant is identical to the reference.
- **Count**: The number of 'countable' fragments supporting the allele. The 'countable' fragments are those that are used by the variant caller when calling the variant.
- **Coverage**: The fragment coverage at this position.
- **Frequency**: Variant frequency calculated by 'Count' divided by 'Coverage'.
- **Average quality**: The average base quality score of the bases supporting a variant.

- **Read count**: The number of ‘countable’ reads supporting the allele. Note that each read in an overlapping pair contributes a count of 1.
- **Read coverage**: The read coverage at this position.
- **# Unique start positions**: The number of unique start positions for ‘countable’ fragments that support the variant. This value can be important to look at in cases with low coverage. If all reads supporting the variant have the same start position, you could suspect that it is a result of an amplification error.
- **# Unique end positions**: The number of unique end positions for ‘countable’ fragments that support the variant. This value can be important to look at in cases with low coverage. If all reads supporting the variant have the same end position, you could suspect that it is a result of an amplification error.
- **BaseQRankSum**: The BaseQRankSum column contains an evaluation of the quality scores in the reads that has a called variant compared with the quality scores of the reference allele. Variants for which no corresponding reference allele is called does not have a BaseQRankSum value. Likewise, no values are calculated for reference alleles. The score is a Z score, wherein a value of 2.0 means that the observed qualities for the variant two standard deviations below the qualities for the reference allele. The scoring is performed using a Mann-Whitney U for comparing the two sets of quality scores from the reference allele and the variant.
- **Read position test probability**: The test probability for the test of whether the distribution of the read positions variant in the variant carrying reads is different from that of all the reads covering the variant position.
- **Read direction test probability**: The test probability for the test of whether the distribution among forward and reverse reads of the variant carrying reads is different from that of all the reads covering the variant position.
- **Homopolymer**: The column contains ”Yes” if the variant is likely to be a homopolymer error and ”No” if not.
- **Amino acid change**: If the reference sequence of the mapping is annotated with ORF or CDS annotations, the variant caller will also report whether the variant is synonymous or non-synonymous. If the variant changes the amino acid in the protein translation, the new amino acid will be reported.

Consensus sequence

The sequence of the predominant SARS-CoV-2 variant from the sequenced sample was calculated from the locally realigned sequencing reads. The consensus sequence is calculated by evaluating each position at a time along the reference genome. If there’s a conflict at a position, meaning there are at least 2 nucleotides being sequenced, the nucleotide with highest count will be reported (annotated as conflict in Table 2). When there is a tie, the base calling quality scores for each nucleotide are summed, and the nucleotide with the highest total quality score is selected as the consensus. For regions with no read coverage, an ambiguous nucleotide ”N” will be reported (annotated as Low coverage in Table 2). The resultant consensus sequence will have quality scores assigned for each consensus base using the base calling quality scores from all the aligned bases at that position. FASTQ files of the consensus viral genome sequences are provided in **[4] Consensus viral genome** for each sample.

Table 2. An example table of consensus sequence annotations.

Name	Type	Region	Qualifiers
Low coverage	Low coverage	1..17	/Length (low coverage) = 17; /Reference position = 1..17
Low coverage	Low coverage	29837..29856	/Length (low coverage) = 20; /Reference position = 29837..29856

Conflict	Conflict	30	/Conflict resolution = Quality sum vote called 'A';
			/Coverage = 12;
			/A quality-count = 11;
			/C quality-count = 1;
			/Reference position = 30

Conflict	Conflict	32	/Conflict resolution = Quality sum vote called 'C';
			/Coverage= 380;
			/C quality-count = 378;
			/G quality-count = 2;
			/Reference position = 32

2.5 Phylogenetic Tree

A maximum-likelihood-based phylogenetic tree is constructed by the "SNP tree" tool using variants identified from the sequencing read alignments among all samples. This algorithm uses variants that we found from each SARS-CoV-2 sample to measure the relatedness between the samples. The maximum likelihood phylogeny tree is constructed with a General Time Reversible nucleotide substitution model using a transition/transversion ratio of 2. The resultant circular cladogram shows the hypothetical phylogenetic relationship among samples without showing the evolutionary distances in the branch lengths.

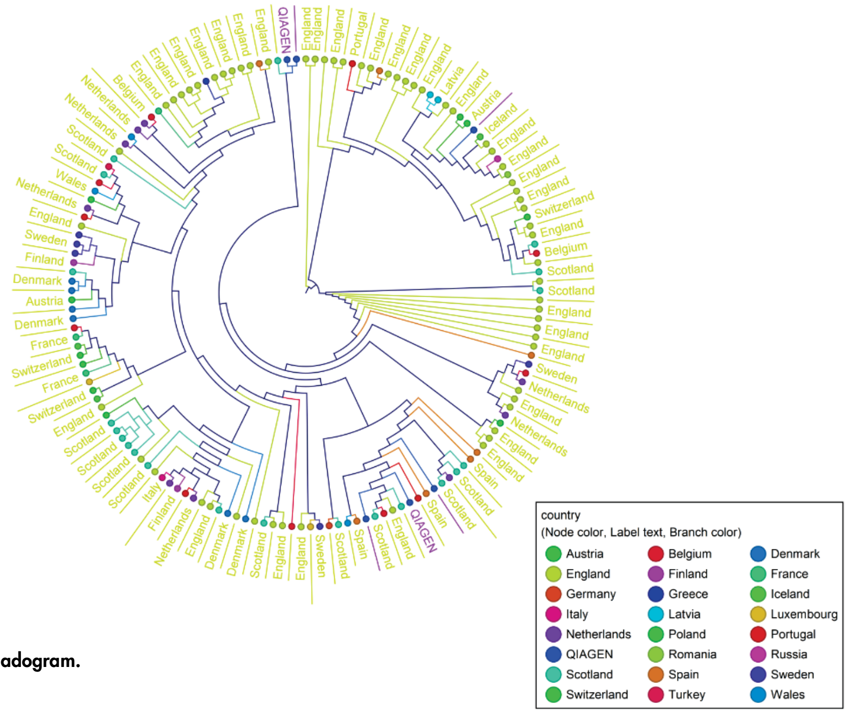


Figure 4. An example circular cladogram.

Material and Methods

RNA isolation

Viral RNA was isolated using the QIAamp® Viral RNA Mini Kit (QIAGEN) according to manufacturer`s instructions.

RNA QC

The viral load was determined using a proprietary qPCR protocol. The primer sequences used were specific to N1/N2 regions of SARS-CoV-2 based on the publication of the US Center of Disease Control and Prevention (CDC).

SARS-CoV-2 library preparation and sequencing

Library preparation was performed by a targeted amplification of viral nucleic acid with the QIAseq SARS-CoV-2 Primer Panel followed by library preparation with the QIAseq FX Kit for DNA library preparation. Briefly, 5 µl of viral RNA input was reverse transcribed by random priming. Viral sequences were enriched by PCR using primer pools based on the ARTICnetwork, yielding amplicons of ~400bp size. 50 ng of DNA was enzymatically fragmented and end-repaired, and the 3′ end was adenylated in a single step. Sequencing adapters were ligated to the 3′ overhangs and the libraries were purified with AMPure® XP beads following the manufacturer's protocol. Adapted-ligated fragments were enriched by 6, 10, and 12 cycles of PCR. After an AMPure bead-based cleanup, the libraries size distribution was validated, and quality inspected on a 2100 Bioanalyzer® or 4200 TapeStation® (Agilent Technologies). High-quality libraries were pooled to obtain equimolar concentrations. The library pool/s were quantified using qPCR for concentration adjustments and subjected to sequencing on a NextSeq® 500/550 instrument (2x150 cycles) according to the manufacturer instructions (Illumina Inc.).

Data Analysis

Data analysis was done using CLC Genomics Workbench (version 20.0.4) and CLC Genomics Server (version 20.0.4). Briefly, the read-through adapter sequences and low-quality bases were trimmed from the sequencing reads, and the high quality trimmed reads were mapped to the SARS-CoV-2 genome (Wuhan Hu 1; GenBank: MN908947.3). To avoid artefacts from variant detection, SARS-CoV-2 specific primer sequences were then trimmed from the aligned sequences, and a local realignment step was then performed to improve the read mappings at sites with insertions or deletions. Variant detection down to 1% was then reported from the read alignments and a consensus sequence was computed from the read alignments.

Notes

Distributed by:

bioNova científica, s.l.

Tel.: 915 515 403

Fax: 914 334 545

e-mail: info@bionova.es

www.bionova.es

Trademarks: QIAGEN®, Sample to Insight®, QIAamp®, QIAseq® (QIAGEN Group); Bioanalyzer®, TapeStation® (Agilent Technologies, Inc.); Ampure® (Beckman Coulter, Inc.); NextSeq® (Illumina Inc.). Registered names, trademarks, etc. used in this document, even when not specifically marked as such, are not to be considered unprotected by law.