CoronaHiT: large scale multiplexing of SARS-CoV-2 genomes using Nanopore sequencing

Dave J. Baker[1], Gemma L. Kay[1], Alp Aydin[1], Thanh Le-Viet[1], Steven Rudder[1], Ana P. Tedim[1,2], Anastasia Kolyva[1,3], Maria Diaz[1], Leonardo de Oliveira Martins[1], Nabil-Fareed Alikhan[1], Lizzie Meadows[1], Andrew Bell[1], Ana Victoria Gutierrez[1], Alexander J. Trotter[1,4], Nicholas M. Thomson[1], Rachel Gilroy[1], Luke Griffith[4], Evelien M. Adriaenssens[1], Rachael Stanley[3], Ian G. Charles[1,4], Ngozi Elumogo[1,3], John Wain[1,4], Reenesh Prakash[3], Emma Meader[3], Alison E. Mather[1,4], Mark A. Webber[1,4], Samir Dervisevic[3], Andrew J. Page[1,*,+] and Justin O'Grady[1,4,*,+]

1 Quadram Institute Bioscience, Norwich Research Park, Norwich, NR4 7UQ, UK.

2 Grupo de Investigación Biomédica en Sepsis - BioSepsis. Hospital Universitario Rio Hortega/Instituto de Investigación Biomédica de Salamanca (IBSAL), Valladolid/Salamanca, Spain.

3 Norfolk and Norwich University Hospital, Colney Lane, Norwich, NR4 7UY, UK.

4 University of East Anglia, Norwich Research Park, Norwich, NR4 7TJ, UK.

*Contributed equally

[+]Corresponding authors: andrew.page@quadram.ac.uk (bioinformatics); justin.ogrady@quadram.ac.uk (sequencing)

Keywords: SARS-CoV-2, Nanopore, Sequencing, NGS, Genome, Genetic, multiplexing, ARTIC

Abstract

The COVID-19 pandemic has spread to almost every country in the world since it started in China in late 2019. Controlling the pandemic requires a multifaceted approach including whole genome sequencing to support public health interventions at local and national levels. One of the most widely used methods for sequencing is the ARTIC protocol, a tiling PCR approach followed by Oxford Nanopore sequencing (ONT) of up to 24 samples at a time. There is a need for a higher throughput method to reduce cost per genome. Here we present CoronaHiT, a method capable of multiplexing up to 95 small genomes on a single Nanopore flowcell, which uses transposase mediated addition of adapters and PCR based addition of symmetric barcodes. We demonstrate the method using 48 and 94 SARS-CoV-2 genomes per flowcell, amplified using the ARTIC protocol, and compare performance with Illumina and ARTIC ONT sequencing. Results demonstrate that all sequencing methods produce inaccurate genomes when the RNA extract from SARS-CoV-2 positive clinical sample has a cycle threshold (Ct) >= 32. Results from set same set of 23 samples with a broad range of Cts show that the consensus genomes have >90% coverage (GISAID criteria) for 78.2% of samples for CoronaHiT-48, 73.9% for CoronaHiT-94, 78.2% for Illumina and 73.9% for ARTIC ONT, and all have the same clustering on a maximum likelihood tree. In conclusion, we demonstrate that CoronaHiT can multiplex up to 94 SARS-CoV-2 genomes per nanopore flowcell without compromising the quality of the resulting genomes while reducing library preparation complexity and significantly reducing cost. This protocol will aid the rapid expansion of SARS-CoV-2 genome sequencing globally, to help control the pandemic.

Introduction

The COVID-19 pandemic caused by the SARS-CoV-2 virus began late 2019 in Wuhan, China and has now spread to virtually every country worldwide with millions of confirmed cases and hundreds of thousands of deaths (Dong, Du, and Gardner 2020). Key to the control of the pandemic is understanding the epidemiological spread of the virus at global, national and local scales (Shu and McCauley 2017). Whole genome sequencing of SARS-CoV-2 is the fastest and most accurate method to study virus epidemiology as it spreads. We are sequencing SARS-CoV-2 as part of the COVID-19 Genomics UK (COG-UK) consortium, a network of academic and public health institutions across the UK brought together to collect, sequence and analyse whole genomes to fully understand the transmission and evolution of this virus (https://www.cogconsortium.uk/). SARS-CoV-2 genome was discovered in China using a metatranscriptomic approach (Wu et al. 2020). This facilitated the design of tiling PCR approaches for genome sequencing, the most widely used of which is the ARTIC Network (https://artic.network) protocol. Consensus genome sequences are typically made publicly available on GISAID (Elbe and Buckland□Merrett 2017). This has enabled real-time public health surveillance of the spread and evolution of the pandemic through interactive tools such as NextStrain (Hadfield et al. 2018). The ARTIC network protocol was designed for Nanopore sequencing, enabling rapid and flexible genome sequencing. However, the method is capable of testing only 24 samples at a time on a flowcell. Currently, a single MinION flowcell can yield ~12 Gbases of data using the ARTIC protocol; given that SARS-CoV-2 genomes are approximately 30,000 bases long, this gives 7,000-10,000X depth of coverage for a single sample. This far exceeds the coverage needed to produce a consensus genome from amplicon

sequencing, leading to needlessly high per sample cost and an excess of data per sample. A high throughput option is required for SARS-CoV-2 sequencing on the MinION to reduce cost and increase throughput. Here we describe a flexible protocol, Coronavirus High Throughput (CoronaHiT), which allows for 1-95 samples to be multiplexed on a single flowcell. We demonstrate CoronaHiT's performance on 48 and 94 SARS-CoV-2 genomes on the MinION and compare data from the same samples sequenced using the standard ARTIC protocol and Illumina.

Methods

Patient samples and RNA extraction

Nasopharyngeal swab samples with suspected SARS-CoV-2 were processed at either the Cytology or Clinical Virology Departments, Norfolk and Norwich University Hospital, Norwich, UK. In the Cytology Department, samples were processed using the Roche Cobas® 8800 SARS-CoV-2 test (https://www.who.int/diagnostics_laboratory/eul_0504-046-00_cobas_sars_cov2_qualitative_assay_ifu.pdf?ua=1) according to the manufacturer's

instructions. RNA was extracted from patient swab samples in the Clinical Virology Department using the QIAsymphony (Qiagen) according to the manufacturer's instructions and the presence of SARS-CoV-2 was determined using the AusDiagnostics SARS-CoV-2, Influenza and RSV 8-well panel (AusDiagnostics 2020).

Excess RNA from positive samples was collected from Virology but positive swabs had to be collected from Cytology as RNA is not retrievable from the Roche machine. In total, 285 excess positive SARS-CoV-2 inactivated swab samples ((200µl viral transport medium from nose and throat swabs inactivated in 200µl Zymo DNA/RNA shield and 800µl Zymo viral DNA/RNA buffer) were collected from Cytology and 94 SARS-CoV-2 positive RNA extracts (~20µl) were collected from Virology as part of the COG-UK Consortium project (PHE Research Ethics and Governance Group R&D ref no NR0195). For inactivated samples, RNA was extracted using the Quick DNA/RNA Viral Magbead kit from step 2 of the DNA/RNA purification protocol (Zymo - https://files.zymoresearch.com/protocols/_r2140_r2141_quick-dna-rna_viral_magbead.pdf).

The lower Ct or take-off value produced by the two SARS-CoV-2 assays in the Roche or AusDiagnostics tests were used to determine if samples needed to be diluted according to the ARTIC protocol (for AusDiagnostics results, 13 was added to the take-off value to generate an approximate Ct value - this is because 15 cycles of PCR are performed before a dilution step and a further 35 cycles of nested PCR (the take-off value is determined in the nested PCR)).

ARTIC SARS-CoV-2 multiplex tiling PCR

cDNA and multiplex PCR reactions were prepared following the ARTIC nCoV-2019 sequencing protocol v2 (Josh Quick 2020). Dilutions of RNA were prepared when required based on Ct values following the guidelines from the ARTIC protocol.

V3 CoV-2 primer scheme (https://github.com/artic-network/artic-ncov2019/tree/master/primer_schemes/nCoV-2019/V3) were used to perform the multiplex PCR for SARS-CoV-2 according to the ARTIC protocol (Josh Quick 2020) with minor changes. Due to variable Ct values all samples were run for 35 cycles in the two multiplex PCRs. Odd and even PCR reactions were pooled and cleaned using a 1x SPRI bead clean using KAPA Pure Beads (Roche Catalogue No. 07983298001) according to manufacturer instructions. Final elution was performed using 30 µl of 10 mM Tris-HCL buffer, pH 7.5. Quantification was performed using QuantiFluor® ONE dsDNA System (Promega, WI, USA).

CoronaHiT library preparation

Libraries were prepared using a novel modified Illumina Nextera low input tagmentation approach (Rowan et al. 2019). Primers with a 3' end compatible with the Nextera transposon insert and a 24bp barcode at the 5' end with a 7 bp spacer were used to PCR barcode the tagmented ARTIC PCR products. The barcode sequences are from the PCR Barcoding Expansion 1-96 kit (EXP-PBC096, Oxford Nanopore Technologies). Symmetrical dual

barcoding was used, i.e. the same barcode added at each end of the PCR product and up to 95 samples could be run together using this approach (Supplementary Table 5).

Tagmentation was performed as follows; 0.9 µl of TD Tagment DNA Buffer (Illumina Catalogue No. 20034197), 0.09 µl TDE1 Tagment DNA Enzyme (Illumina Catalogue No. 20034197) and 4.01 µl PCR grade water was made as a master mix scaled to sample number. On ice, 5 µl of tagmentation mix was added to each well of a chilled 96-well plate. Next, 2 µl of normalised PCR product (normalised to 0.5 ng/µl, 1 ng total) was pipette mixed with the 5 µl tagmentation mix. This plate was sealed and briefly centrifuged before incubation at 55°C for 10 minutes in a thermal cycler.

PCR barcoding was performed using Kapa 2G Robust PCR kit (Sigma Catalogue No. KK5005) as follows: 4 µl Reaction buffer (GC), 0.4 µl dNTP's, 0.08 µl Kapa 2G Robust Polymerase and 7.52 µl PCR grade water per sample were mixed and 12 µl was added to each well in a new 96-well plate. 1 µl of the appropriate barcode pair (Supplementary Table 5) at 10µM was added to each well. Finally, the 7 µl of Tagmentation mix was added. PCR reactions were run at 72°C for 3 minutes, 95°C for 1 minute, followed by 14 cycles of 95°C for 10 seconds, 55°C for 20 seconds and 72°C for 1 minute. Following PCR, libraries were quantified using the QuantiFluor® ONE dsDNA System and from the results pooled together in equal nanograms and SPRI size selected at 0.8X bead volume using KAPA Pure Beads (Roche Catalogue No. 07983298001) with final elution in 50µl PCR grade water. The barcoded pool was quantified and sized on a Tapestation D5000 tape (Agilent).

A nanopore sequencing library was then made with the pooled fractions following the SQK-LSK109 protocol. Briefly, 190 ng of the washed barcoded pool was used in the end-prep reaction, excluding FFPE repair (7 µl Ultra II end prep buffer, 3 µl Ultra II end prep enzyme mix, and barcoded pool plus water up to a final volume of 60 µl). The reaction was incubated at 20°C for 10 mins and 65°C for 5 mins. This was bead-washed (using 70% ethanol) and eluted in 61 µl. The end-prepped DNA was taken forward to the adapter ligation as follows: 60 µl end-prepped pool (~140ng), 25 µl LNB (ONT). 10 µl NEBNext Quick T4 Ligase, 5 µl AMX was mixed and incubated at room temperature for 20 minutes. Finally, after a bead-wash with SFB, ~17.5 ng (~70 fmol) was used to load the MinION flowcell. and sequenced for up to 48 hours.

Illumina library preparation

Normalised PCR products were tagmented and barcoded as described for the CoronaHiT library preparation, however, standard Nextera XT Index Kit primers were used (Sets A to D for up to 384 combinations, Illumina Catalogue No's FC-131-2001, FC-131-2002, FC-131-2003 and FC-131-2004) . The PCR master mix was adjusted and water removed to add 2 µl each of the P7 and P5 primers. Barcoded libraries were quantified using the QuantiFluor® ONE dsDNA System and from the result pooled together in equal nanograms, identically to CoronaHiT method and again SPRI size selected at 0.8X bead volume using KAPA Pure Beads (Roche Catalogue No. 07983298001) with final elution in 50 µl EB, 10mM TrisHCl. The barcoded pool was sized on a Agilent Tapestation D5000 tape and the concentration quantification performed using QuantiFluor® ONE dsDNA System (Promega, WI, USA) and the molarity calculated. The

Illumina library pool was run at a final concentration of 1.5 pM on an Illumina Nextseq500 instrument using a Mid Output Flowcell (NSQ® 500 Mid Output KT v2 (300 CYS) Illumina Catalogue FC-404-2003) following the Illumina recommended denaturation and loading recommendations which included a 1% PhiX spike in (PhiX Control v3 Illumina Catalogue FC-110-3001).

ARTIC protocol Nanopore library preparation

After bead clean-up and quantification of the two amplicon pools, ARTIC Nanopore library preparation was performed using the protocol "PCR tiling of COVID-19 virus" (Oxford Nanopore Technologies), Revision E (released 26th of March 2020 - https://community.Nanoporetech.com/protocols/pcr-tiling-ncov/v/PTC_9096_v109_revE_06Feb2020). Briefly, each sample was end-repaired prior to barcode ligation (1.75 µl Ultra II end prep buffer, 0.75 µl Ultra II end prep enzyme mix, 50 ng of washed amplicon and water up to 15 µl) using a thermal cycler at 20°C for 5 min and 65°C for 5 min. A tenth of this was used directly for native barcode ligation (1.5 µl end-prepped DNA, 5.5 µl nuclease free water, 2.5 µl native barcode, 10 µl NEBNext Ultra II Ligation Master Mix, 0.5 µl NEBNext Ligation Enhancer, final volume 20 µl) in a thermal cycler at 20°C for 20 min and 65°C for 10 min. Amplicons were pooled together and underwent a 0.4X AMPure bead wash with two 700 µl SFB washes and one 80% ethanol wash. DNA was eluted in 35 µl of nuclease free water. Adapter ligation was performed on 50 ng of the pool (5 µl Adapter Mix II (ONT), NEBNext Quick Ligation Reaction Buffer (5X), 5 µl Quick T4 DNA Ligase, 50 ng pooled barcoded amplicons and water up to 50 µl). The ligation reaction was incubated at room

temperature for 20 min and 0.4X bead washed with 125 µl SFB two times. The library was eluted in 15 µl of elution buffer and quantified. 20 ng of the adapted library was used for final loading.

Nanopore sequence analysis

Basecalling was performed using Guppy v.3.6.0 (Oxford Nanopore Technologies) in high accuracy mode (model dna_r9.4.1_450bps_hac), on a private OpenStack cloud at Quadram Institute Bioscience using multiple Ubuntu v18.04 virtual machines running Nvidia T4 GPU.

The CoronaHiT sequencing data were demultiplexed using guppy_barcoder (v3.6.0) with the option 'require_barcodes_both_ends' and a score of 50 at the front, 40 at the end to produce 95 FASTQ files (94 SARS-CoV-2 samples and 1 negative control) and 49 FASTQ files (48 SARS-CoV-2 samples and 1 negative control) respectively. The ARTIC ONT sequencing data were demultiplexed using guppy_barcoder (v3.6.0) with the option 'require_barcodes_both_ends' and a score of 50 at the front, 40 at the end to produce 24 FASTQ files (23 SARS-CoV-2 samples and 1 negative control).

The downstream analysis was performed using the ARTIC pipelines (v1.1.1) as previously described (Loman, Rowe, and Rambaut 2020), which produced a consensus sequence for each sample in FASTA format. Input reads were filtered based on size (ARTIC: 400-700; CoronaHiT: 150-550), and mapped to the Wuhan-Hu-1 reference genome (accession MN908947.3) to produce a consensus sequence. The consensus sequences were uploaded to GISAID and the raw sequence data was uploaded to the European Nucleotide Archive under BioProjects ERP122169

and ERP121228. The accession numbers for each sample are available in Supplementary Table 1, including raw reads and consensus sequences for each sequencing method. The metrics and results of all experiments are available in Supplementary Table 2 and are summarised in Table 1.

Illumina sequence analysis

The raw reads were demultiplexed using bcl2fastq (v2.20) (Illumina Inc.) to produce 384 FASTQ files (380 SARS-CoV-2 samples and 4 negative controls), with only the first 95 analysed in this paper. The reads were used to generate a consensus sequence for each sample using an open source pipeline adapted from https://github.com/connor-lab/ncov2019-artic-nf (https://github.com/quadram-institute-bioscience/ncov2019-artic-nf/tree/qib). Briefly, the reads had adapters trimmed with TrimGalore (https://github.com/FelixKrueger/TrimGalore), were aligned to the Wuhan-Hu-1 reference genome (accession MN908947.3) using BWA-MEM (v0.7.17) (Li 2013), the ARTIC amplicons were trimmed and a consensus built using iVAR (v.1.2) (Grubaugh et al. 2019).

Quality Control

The COG-UK consortium defined a consensus sequence as passing COG-UK quality control if greater than 50% of the genome was covered by confident calls or there was at least 1 contiguous sequence of more than 10,000 bases and with no evidence of contamination. This is regarded as the minimum amount of data to be phylogenetically useful. A confident call was defined as having a minimum of 10X depth of coverage for Illumina data and 20X depth of coverage for Nanopore data. If the coverage fell below these thresholds, the bases were masked with Ns. Low quality variants were also masked with Ns. The QC threshold for inclusion in GISAID was

higher, requiring that greater than 90% of the genome was covered by confident calls and there was no evidence of contamination.

Phylogenetic analysis

For each sample sequenced in 4 separate experiments (n=23, CoronaHiT-48, CoronaHiT-94, ARTIC ONT, Illumina), a phylogeny was generated from all of the consensus genomes (n=76) passing GISAID QC (n=19). A multiple FASTA alignment was created by aligning all samples to the reference genome MN908947.3 with MAFFT v7.467 . A maximum likelihood tree was estimated with IQTREE2 (v2.0.4) (Minh et al. 2020) under the HKY model (Hasegawa, Kishino, and Yano 1985), collapsing branches smaller than 10^{-7} into a polytomy. SNPs in the multiple FASTA alignment were identified using SNP-sites (v2.5.1) (Page et al. 2016) and the tree was visualised with FigTree (v1.4.4) (https://github.com/rambaut/figtree).

Results

A novel library preparation method, CoronaHiT, was developed for SARS-CoV-2 genome sequencing on the MinION, which combines transposase-based introduction of adapters (Illumina Nextera) with symmetric PCR barcoding of up to 95 samples. Nextera adapter complementary primer sequences were added to ONT PCR barcodes and used to barcode ARTIC PCR products (Figure 1) as described in the methods. This approach eliminates the requirement of individual end-prep and ligation reactions per sample.

The CoronaHiT method was tested by multiplexing 94 SARS-CoV-2 samples on a single MinION flowcell. A subset of 48 of these samples were run on a second flowcell to demonstrate the impact of increased coverage on the quality of the genomes generated. The 94 samples were also sequenced using the Illumina NextSeq platform and a subset of 23 samples were multiplexed using the ARTIC ONT protocol on the MinION as comparators for CoronaHiT. The different methods produced different amounts of data, affecting coverage. CoronaHiT-94 yielded 9.2 Gbases of sequence data (43.58 hrs sequencing time) following demultiplexing with 757X average coverage per sample. The CoronaHiT-48 dataset yielded slightly more data at 11.1 Gbases (43.13 hrs sequencing time) and with fewer samples the average coverage of mapping reads increased to 2037X. The ARTIC sequencing produced 9 Gbases of data (23.02 hours sequencing time) for 23 samples giving 7509X coverage. Illumina sequencing yielded 44.1 Gbases (48 hours sequencing time) with an average coverage of 2755X (see Table 1).

Taking the genomes which passed COG-UK QC, the Illumina sequencing produced the shortest mean read size at 141 bases, just short of the maximum 150 bases for the PE 151 chemistry, ARTIC ONT produced the longest mean reads at 507 bases and CoronaHiT-48 and 94 sequencing produced mean read lengths of 407 and 421 bases respectively. The shorter read lengths for CoronaHiT are related to the use of tagmentation for introducing adapters, resulting in the removal of the ends of the ARTIC PCR products.

CoronaHiT demultiplexing was performed as described in the methods section. The demultiplexing steps were different from those used for ARTIC ONT sequencing, removing

reads <100bp and >550bp. The negative control samples contained zero mapping reads to SARS-CoV-2 for all datasets with the exception of Illumina (80 reads).

For the CoronaHiT-48 and 94 datasets, 69.12% and 65.07% of reads respectively were demultiplexed successfully when only reads with a PHRED (quality) score above Q7 are considered as compared with 87.23% for the ARTIC ONT dataset. The rest of the reads were unassigned, due to an inability to detect the barcode sequences at both ends of the reads.

Poor quality consensus genomes were generally associated with a lower SARS-CoV-2 viral load in the clinical samples i.e. higher RT-qPCR Ct values (generally above Ct 32 - Figure 2), and were more likely to fail COG-UK and GISAID quality control thresholds. For all methods the number of Ns increased significantly in samples with a Ct above 32, which equates to approx 100 viral genome copies in the PCR reaction (Figure 2a). Samples sequenced using Illumina and ARTIC ONT had genome coverage (~4000-10,000X) far in excess of the advised 1000X and produced similar results (number of Ns per sample). CoronaHiT had lower coverage and hence the consensus sequence quality was poorer. Figure 2b (and Supplementary Figures 2a and 2b) shows the Ns (missing or masked bases) within the consensus genomes - the three ARTIC PCR primer dropout areas (Benjamin Farr et al. 2020) are clearly visible. Comparing the samples sequenced in all four sequencing runs with a Ct below 32 (n=16), the mean (median) number of Ns was 743 (121) for ARTIC ONT, 912 (46) for Illumina, 1859 (490) for CoronaHiT-94, and 1205 (132) for CoronaHiT-48. If all 23 samples are included (including higher Ct samples) then the number of Ns increases substantially to an average of 5297.65 bases for ARTIC ONT, 4926.60 for Illumina, 4859.21 CoronaHiT-94 and 4235.34 for CoronaHiT-48. Downsampling

the ARTIC ONT data to have the same coverage as CoronaHiT samples showed that results were similar when the coverage was the same (Supplementary Figures 2a and 2b). When all samples with a Ct below 32 were included (Table 2), the mean (median) number of Ns was 157 (121) for ARTIC Nanopore, 384 (40) for Illumina, 881 (369) for CoronaHiT-94, and 409 (128) for CoronaHiT-48.

The pass rate for the COG-UK QC criteria for all methods was in a reasonably tight range of 76% to 82.6%, with some variation accounted for by different samples tested. The stricter GISAID QC criteria reduced the pass rate of samples, ranging from 65% for CoronaHiT-94 to 73.9% for ARTIC ONT. Most of the samples failing GISAID QC had a Ct above 32, with 6/7 (85.7%) for ARTIC ONT, 19/25 (76%) for Illumina, 27/33 (81.1%) for CoronaHiT-94 and 11/13 (84.6%) for CoronaHiT-48. When considering higher viral load samples with a Ct below 32, there was very little difference between the proportion of samples passing both GISAID and COG-UK QC with 16/16 (100%) for ARTIC Nanopore, 230/240 (95.8%) for Illumina, 32/34 (94.11%) for CoronaHiT-48 and 60/65 (92.2%) for CoronaHiT-94. Full details are shown in Table 2. When considering the same 16 samples sequenced on all technologies (Ct below 32), all samples passed GISAID and COG-UK QC for all methods.

To assess the impact of data quality differences on clustering of lineages, we built a maximum likelihood tree with each of the 19 consensus genomes from the ARTIC ONT, CoronaHiT-48, CoronaHiT-94 and Illumina sequencing experiments. When the consensus genomes were placed on a phylogenetic tree, Illumina, ARTIC ONT, CoronaHiT-48 and CoronaHiT-94 showed the exact same clustering for the same samples. All variant differences between the 19 samples are

noted in Supplementary Table 3, with the only other source of variation between the samples being that the Illumina genomes contain IUPAC (IUPAC-IUB Comm. on Biochem. Nomenclature (CBN) 1970) symbols for "partially" ambiguous bases. These data suggest that the higher N count in CoronaHiT compared to Illumina or ARTIC nanopore data had no effect on accurate lineage calling.

The overall variation on average in the number of SNPs between the Wuhan-Hu-1 reference genome and the consensus genomes varied between 8.26 SNPs for CoronaHiT-48 and 9.9 SNPs for Illumina (see Table 2 and Supplementary Table 2). Ambiguous bases in the Illumina dataset were regarded as SNPs in these calculations, resulting in an elevated value. If only samples with a Ct below 32 are considered, this range of SNPs reduced to 8 SNPs for CoronaHiT-94, 10.1 SNPs for Illumina, 8.42 SNPs for Corona-HiT-48 and 8.3 SNPs for ARTIC ONT.

The cost for the standard ARTIC ONT method is approximately £35.75 per sample from extracted RNA to Nanopore sequence data with 24 samples on a flowcell (see Supplementary Table 6). In comparison, when sequencing 95 samples using CoronaHiT, the cost is £12.72 per sample (Supplementary Table 6) giving a 3-fold reduction. 48 samples with CoronaHiT costs £17.12 per sample, giving a 2-fold reduction compared to standard ARTIC ONT.

Discussion

Rapid viral genome sequencing during outbreaks is changing how we study disease epidemiology (Kafetzopoulou et al. 2019; Joshua Quick et al. 2016). The recent SARS-CoV-2

global pandemic has again highlighted the use of sequencing in the control of the spread of the disease. Nanopore sequencing is particularly suited to outbreak sequencing as it is portable, does not require expensive machinery and is accessible throughout the world (Faria et al. 2016). Current Nanopore sequencing-based methods are low throughput, therefore relatively expensive and slow to generate the large numbers of sequences required to accurately study the epidemiology of this pandemic at local and national level. We present a novel high-throughput method, CoronaHiT, for cost effective and low complexity sequencing of SARS-CoV-2 genomes on MinION flowcells.

The ARTIC protocol (Josh Quick 2020) has been widely adopted for SARS-CoV-2 genome sequencing and allows up to 23 samples (plus a negative control) to be sequenced at a time on a MinION. This number is limited, not by sequencing yield, but by barcode chemistry and library preparation complexity - native barcode ligation is required, which is currently limited to 24 individual barcodes. Ligation libraries are expensive, due to the cost of ligase, and do not scale well. CoronaHiT is cheap and scalable, utilising a novel combination of transposase introduction of adapters with PCR based barcoding. This allows the user to increase the number of samples sequenced on a single flowcell from 24 (ligation-based native barcoding method (Bayliss et al. 2017; Josh Quick 2020)) up to 95 samples, at a cost up to 3 times less per sample than ARTIC ONT. The cost saving is related to the number of samples loaded per flowcell and the cost of ligation barcoding versus PCR barcoding. While costs will vary between countries and in different laboratories, the relative saving should remain similar.

Tiling PCR approaches, such as ARTIC, are prone to high genome coverage variation due to variable primer efficiency in multiplex reactions. Some regions of the SARS-CoV-2 genome have hundreds of times higher coverage than adjacent regions using ARTIC, therefore average coverage of at least 1000X is required to obtain at least 20X coverage of the difficult regions of the genome. In our experience, MinION flowcells yield approx. 12Gb using the ARTIC protocol, hence there is sufficient data produced to cover significantly more than 24 samples per flowcell. To determine the optimum number of samples that should be sequenced on the MinION using CoronaHiT, we multiplexed 94 samples (plus one negative control) on one flowcell and a subset of 48 samples (plus one negative control) on a second flowcell. We compared the results to the same 94 samples run on the Illumina NextSeq and to 24 samples run on the MinION using the ARTIC ONT protocol.

Results demonstrate that all methods are unreliable at producing high quality consensus genomes from positive clinical samples with diagnostic RT-qPCR Cts above 32 (approx 100 viral genome copies - Figure 2). Below Ct 32 CoronaHiT-48, ARTIC ONT and Illumina produce very similar results. CoronaHiT-94 shows lower average SNPs compared to the reference against CoronaHiT-48 (8.0 vs 8.42) and higher median number of Ns in the consensus (369 vs 128). Therefore the performance of 48 samples per flowcell was marginally superior to 94 (Table 1). Downsampling of ARTIC ONT coverage and CoronaHiT to similar coverage (700X) shows the mean percentage N's for ARTIC ONT (19.9%) is similar CoronaHiT-48 (17.7%) (Supplementary Figure 2). Hence, fewer than 48 samples would be superior again, however data produced from CoronaHiT-94 was sufficient to provide accurate consensus genomes that result in the same lineages and on the same branches on the phylogenetic tree as Illumina and ARTIC ONT (Figure

3). Therefore, we have demonstrated high quality, multiplexed SARS-CoV-2 genome sequencing of 94 samples on a single flowcell. If the ARTIC ONT method is optimised to correct the three low coverage regions of the genome as expected, less overall coverage will be required per genome and more samples can be multiplexed using the CoronaHiT method.

CoronaHiT reduces library preparation complexity, which is also a major benefit. The method uses a 10 minute tagmentation and 30 minute PCR to barcode up to 95 libraries at a time in just two steps, taking less than 45 minutes to go from normalised pooled ARTIC PCR products to the final sequencing library.

In conclusion, we demonstrate that CoronaHiT can be used to sequence 48 and 95 SARS-CoV-2 samples on a single MinION flowcell, with a marginal increase in N's but without compromising the resulting phylogenetic results. The protocol is simple, fast, 3-folder cheaper and flexible and can be applied for high throughput sequencing of up to 95 small genomes on the MinION. CoronaHiT can help scientists around the world sequence more SARS-CoV-2 genomes faster, thereby increasing our understanding, and reducing the spread, of the pandemic.

Ethical approval

The COVID-19 Genomics UK Consortium has been given approval by Public Health Englands Research Ethics and Governance Group (PHE R&D Ref: NR0195).

Acknowledgements

Thanks to George Taiaroa and Torsten Seemann from the Microbiological Diagnostic Unit Public Health Laboratory at the University of Melbourne for their advice and assistance, to Niamh Tumelty from the University of Cambridge for assistance and to Darren Heavens from the Earlham Institute for his advice on library preparation. Thanks to the COG-UK Consortium Study Group for their contributions.

Funding statement

The authors gratefully acknowledge the support of the Biotechnology and Biological Sciences Research Council (BBSRC); this research was funded by the BBSRC Institute Strategic Programme Microbes in the Food Chain BB/R012504/1 and its constituent projects BBS/E/F/000PR10348, BBS/E/F/000PR10349, BBS/E/F/000PR10351, and BBS/E/F/000PR10352. DJB, NFA, TLV and AJP were supported by the Quadram Institute Bioscience BBSRC funded Core Capability Grant (project number BB/CCG1860/1). EMA was funded by the BBSRC Institute Strategic Programme Gut Microbes and Health BB/R012490/1 and its constituent project(s) BBS/E/F/000PR10353 and BBS/E/F/000PR10356. The sequencing costs were funded by the COVID-19 Genomics UK (COG-UK) Consortium which is supported

by funding from the Medical Research Council (MRC) part of UK Research & Innovation (UKRI), the National Institute of Health Research (NIHR) and Genome Research Limited, operating as the Wellcome Sanger Institute. The author(s) gratefully acknowledge the UKRI Biotechnology and Biological Sciences Research Council's (BBSRC) support of The Norwich Research Park Biorepository. LG was supported by a DART MRC iCASE and Roche Diagnostics. APT was funded by Sara Borrell Research Grant CD018/0123 from ISCIII and co-financed by the European Development Regional Fund (A Way to Achieve Europe program) and APT QIB internship additionally funded by "Ayuda de la SEIMC". The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Financial declaration

LG received a partial support for his PhD from Roche. The use of Roche technology for diagnostics in NNUH is coincidental.

Author contributions

All authors have read this manuscript and consented to its publication. The CoronaHiT method was developed by DJB. The study was designed and conceived by DJB, JOG, AJP. Paper writing was by DJB, AJP, JOG, GLK, AA, APT, TLV, SR, LM. Sequencing and library preparation was performed by DJB, SR, GLK, AA, APT, AB, AJT, NMT, RG, JOG. Bioinformatics analysis and informatics were performed by TLV, LM, NFA, AJP. Clinical diagnostics and extractions were managed by AK, SD, RP, NE, EM. Samples and metadata were collected by LG, AB, AVG, EMA, AK and MD and biobanked by RS, RNA was extracted by AB, AJT. Risk assessments

were by GLK, JW. Project management and oversight was by GLK, JOG, AJP, LM, MW, AEM, JW. Funding for the project was secured by JOG, AJP, IGC.

References

AusDiagnostics. 2020. 'SARS-COV-2, INFLUENZA AND RSV 8-WELL'. 20081. https://www.ausdx.com/qilan/Products/20081-r01.1.pdf.

Bayliss, Sion C., Vicky L. Hunt, Maho Yokoyama, Harry A. Thorpe, and Edward J. Feil. 2017. 'The Use of Oxford Nanopore Native Barcoding for Complete Genome Assembly'. *GigaScience* 6 (3). https://doi.org/10.1093/gigascience/gix001.

Benjamin Farr, Diana Rajan, Emma Betteridge, Lesley Shirley, Michael Quail, Naomi Park, Nicholas Redshaw, et al. 2020. 'COVID-19 ARTIC v3 Illumina Library Construction and Sequencing Protocol', May. https://doi.org/10.17504/protocols.io.bgq3jvyn.

Dong, Ensheng, Hongru Du, and Lauren Gardner. 2020. 'An Interactive Web-Based Dashboard to Track COVID-19 in Real Time'. *The Lancet Infectious Diseases* 0 (0). https://doi.org/10.1016/S1473-3099(20)30120-1.

Elbe, Stefan, and Gemma Buckland□Merrett. 2017. 'Data, Disease and Diplomacy: GISAID's Innovative Contribution to Global Health'. *Global Challenges* 1 (1): 33–46. https://doi.org/10.1002/gch2.1018.

Grubaugh, Nathan D., Karthik Gangavarapu, Joshua Quick, Nathaniel L. Matteson, Jaqueline Goes De Jesus, Bradley J. Main, Amanda L. Tan, et al. 2019. 'An Amplicon-Based Sequencing Framework for Accurately Measuring Intrahost Virus Diversity Using PrimalSeq and IVar'. *Genome Biology* 20 (1): 8. https://doi.org/10.1186/s13059-018-1618-7.

Hasegawa, Masami, Hirohisa Kishino, and Taka-aki Yano. 1985. 'Dating of the Human-Ape Splitting by a Molecular Clock of Mitochondrial DNA'. *Journal of Molecular Evolution* 22 (2): 160–74. https://doi.org/10.1007/BF02101694.

IUPAC-IUB Comm. on Biochem. Nomenclature (CBN). 1970. 'Abbreviations and Symbols for Nucleic Acids, Polynucleotides, and Their Constituents'. *Biochemistry* 9 (20): 4022–27. https://doi.org/10.1021/bi00822a023.

Kafetzopoulou, L. E., S. T. Pullan, P. Lemey, M. A. Suchard, D. U. Ehichioya, M. Pahlmann, A. Thielebein, et al. 2019. 'Metagenomic Sequencing at the Epicenter of the Nigeria 2018 Lassa Fever Outbreak'. *Science (New York, N.Y.)* 363 (6422): 74–77. https://doi.org/10.1126/science.aau9343.

Li, Heng. 2013. 'Aligning Sequence Reads, Clone Sequences and Assembly Contigs with BWA-MEM'. *ArXiv:1303.3997 [q-Bio]*, March. http://arxiv.org/abs/1303.3997.

Loman, Nicholas J., Will Rowe, and Andrew Rambaut. 2020. 'NCoV-2019 Novel Coronavirus Bioinformatics Protocol'. v1.1.0. https://artic.network/ncov-2019/ncov2019-bioinformatics-sop.html.

Minh, Bui Quang, Heiko A. Schmidt, Olga Chernomor, Dominik Schrempf, Michael D. Woodhams, Arndt von Haeseler, and Robert Lanfear. 2020. 'IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era'. *Molecular Biology and Evolution*, February. https://doi.org/10.1093/molbev/msaa015.

Page, Andrew J., Ben Taylor, Aidan J. Delaney, Jorge Soares, Torsten Seemann, Jacqueline A. Keane, and Simon R. Harris. 2016. 'SNP-Sites: Rapid Efficient Extraction of SNPs from Multi-FASTA Alignments'. *Microbial Genomics* 2 (4): e000056.

https://doi.org/10.1099/mgen.0.000056.

Quick, Josh. 2020. 'NCoV-2019 Sequencing Protocol V2', April. https://doi.org/10.17504/protocols.io.bdp7i5rn.

Quick, Joshua, Nicholas J. Loman, Sophie Duraffour, Jared T. Simpson, Ettore Severi, Lauren Cowley, Joseph Akoi Bore, et al. 2016. 'Real-Time, Portable Genome Sequencing for Ebola Surveillance'. *Nature* 530 (7589): 228–32. https://doi.org/10.1038/nature16996.

Shu, Yuelong, and John McCauley. 2017. 'GISAID: Global Initiative on Sharing All Influenza Data – from Vision to Reality'. *Eurosurveillance* 22 (13): 30494. https://doi.org/10.2807/1560-7917.ES.2017.22.13.30494.

Tables

Table 1: Summary statistics for each sequencing experiment. Extended experiment statistics are available in Supplementary Table 4.

	CoronaHiT-48	CoronaHiT-94	ARTIC ONT	Illumina
No. of samples	48	94	23	380
Total yielded bases (Gb)	11.1	9.2	9.0	44.1
Bases deplexed (Gb)	6.9	5.4	7.2	36.8
Run time (h)	43.13	43.58	23.02	48
Reads sequenced (>Q7)	24,488,530	19,297,539	16,299,333	-
Average PHRED score	12.323	12.388	12.941	35
Average coverage (X)	2037X	774X	7509X	2755X
Average read length (bases)	412	442	475	140
Samples passing GISAID QC (<Ct 30)	70.83% (82.35%)	64.2% (88%)	73.9% (82.35%)	67.1% (80.78%)

Table 2: The number of consensus genomes passing and failing the different QC thresholds for each experiment. Extended data are available in Supplementary Tables 2 and 4.

Consensus genome (all)	CoronaHiT-48 (48 samples)	CoronaHiT-94 (94 samples)	ARTIC ONT (23 samples)	Illumina (380 samples)
Passing COG-UK QC	81.25%	80.0%	82.6%	76.05%
Passing GISAID QC	66.67%	65.0%	73.9%	67.10%
Failing COG-UK QC	18.75%	20.0%	17.4%	23.95%

Failing GISAID QC	4.17%	35.0%	26.1%	32.90%
Avg (Median) Ns of COG-UK passed	409 (128)	1859 (490)	743 (121)	912(46)
Avg SNPs of COG-UK passed	8.26	7.95	8.37	10.1
Consensus genome (Ct <=32)	**34 samples**	**65 samples**	**16 samples**	**240 samples**
Passing COG-UK QC	97.05%	98.4%	100%	97.9%
Passing GISAID QC	94.11%	92.2%	100%	95.8%
Failing COG-UK QC	2.95%	1.6%	0%	2.1%
Failing GISAID QC	5.89%	7.8%	0%	4.2%
Avg (Median) Ns of COG-UK passed	409 (128)	881 (369)	157 (121)	384 (40)
Avg SNPs of COG-UK passed	8.42	8.0	8.3	10.1

Figures

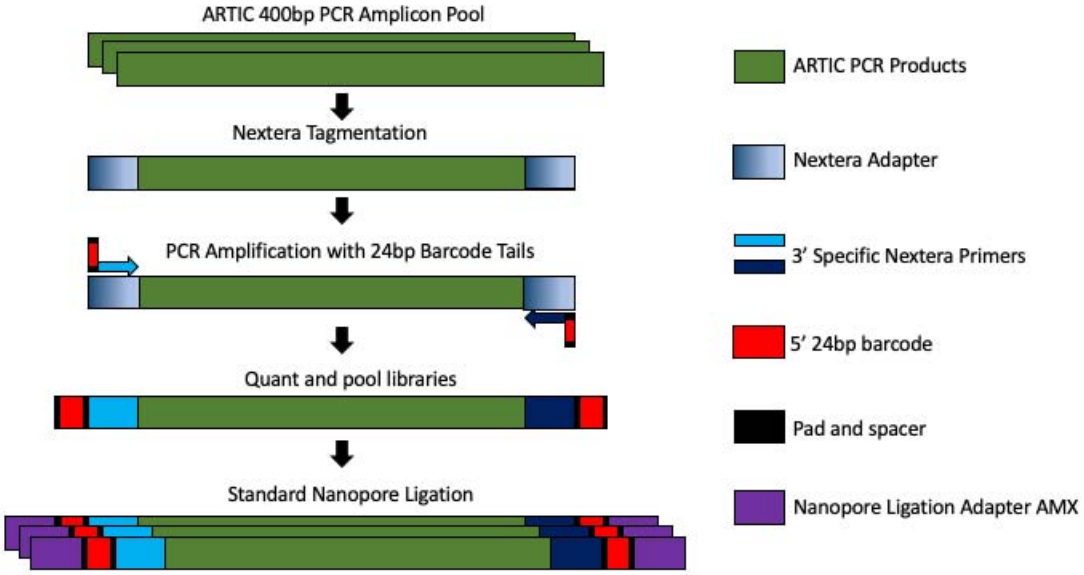
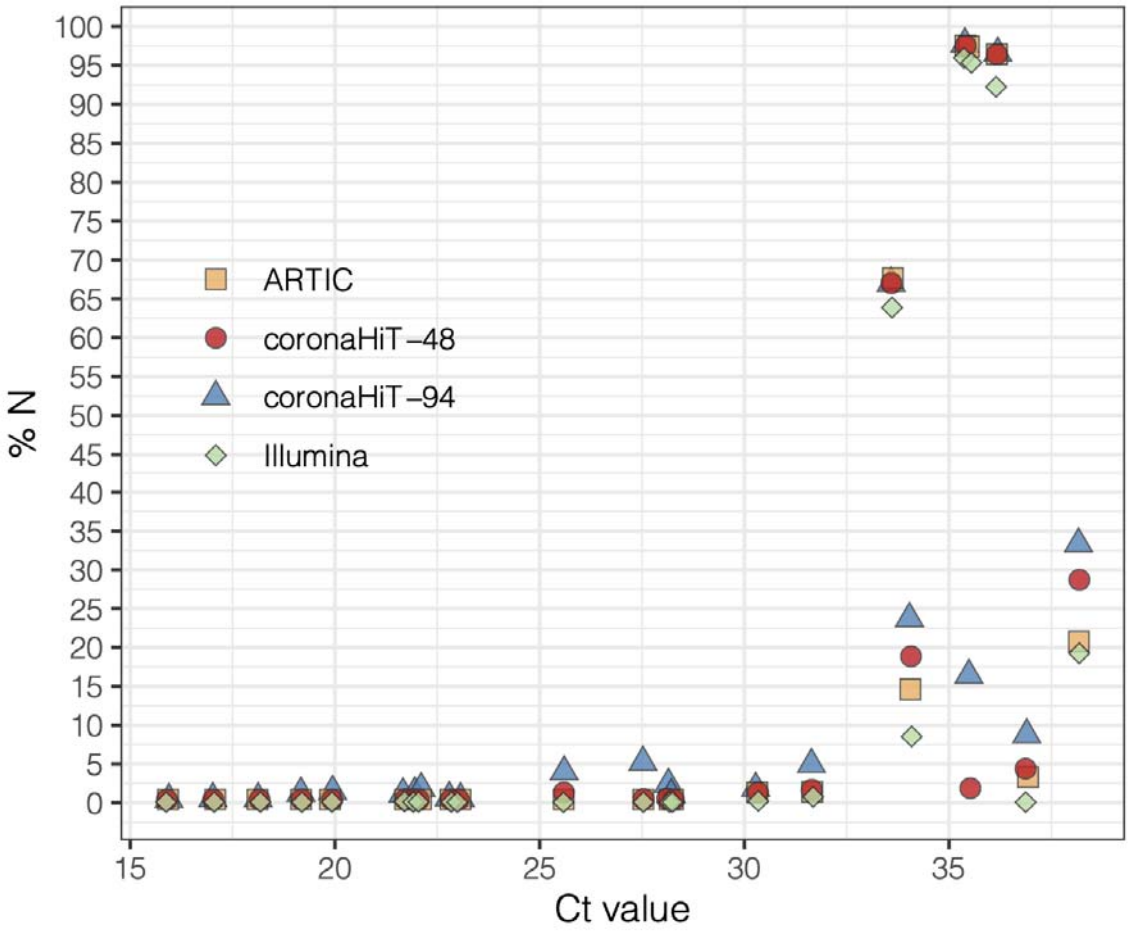


Figure 1: Workflow of CoronaHiT library preparation.

(a)



(b)

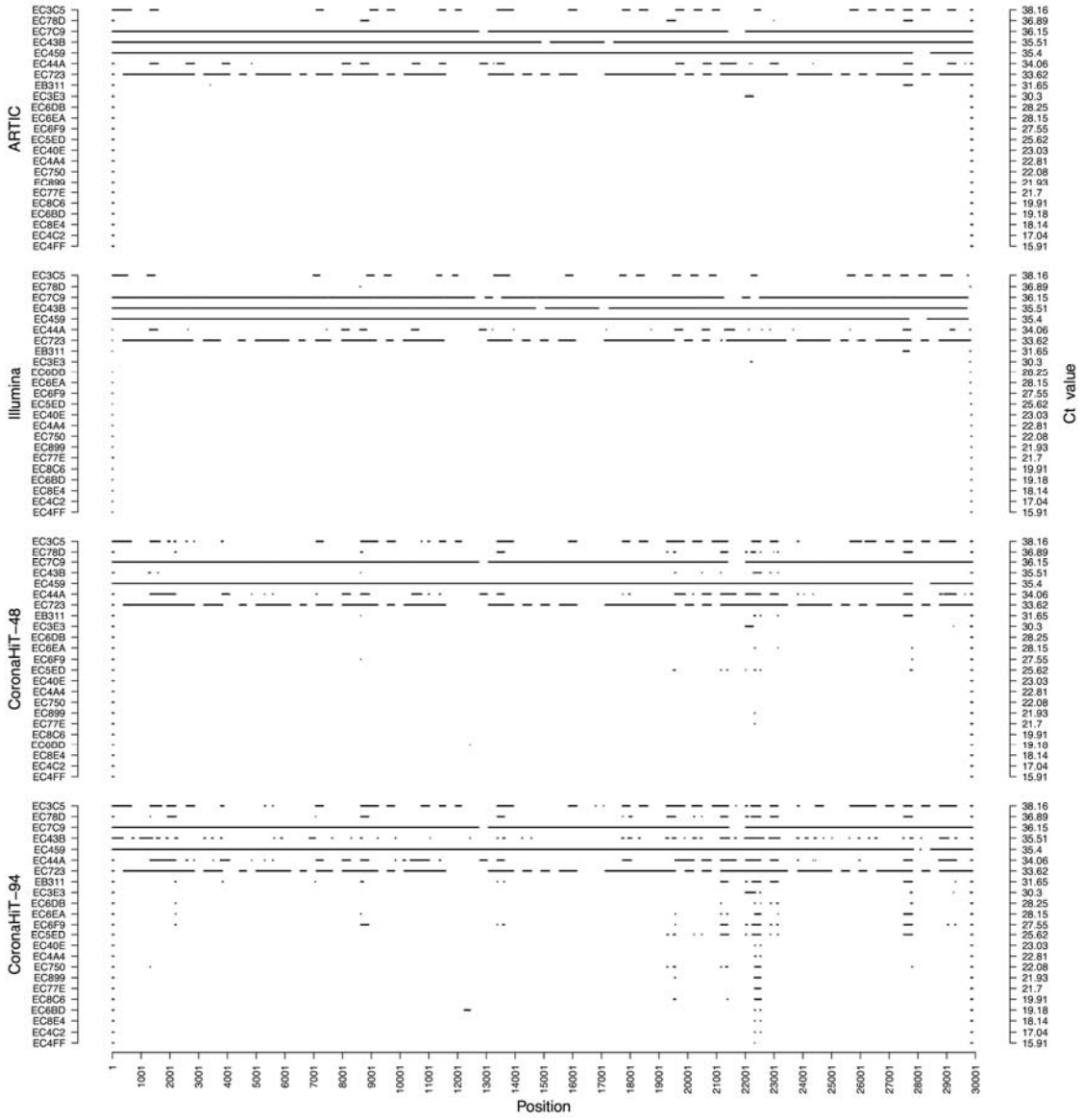


Figure 2: (a) Ct value of the 23 SARS-CoV-2 positive RNA samples sequenced using CoronaHiT, Illumina, and ARTIC ONT methods vs total number of Ns in the consensus sequence (b) Overview of the SARS-CoV-2 genome against the consensus the 23 samples sequenced using the CoronaHiT, Illumina, and ARTIC ONT methods. Lines and dots indicate regions where there is missing data, or where the coverage drops below 20X on Nanopore, and below 10X on Illumina.

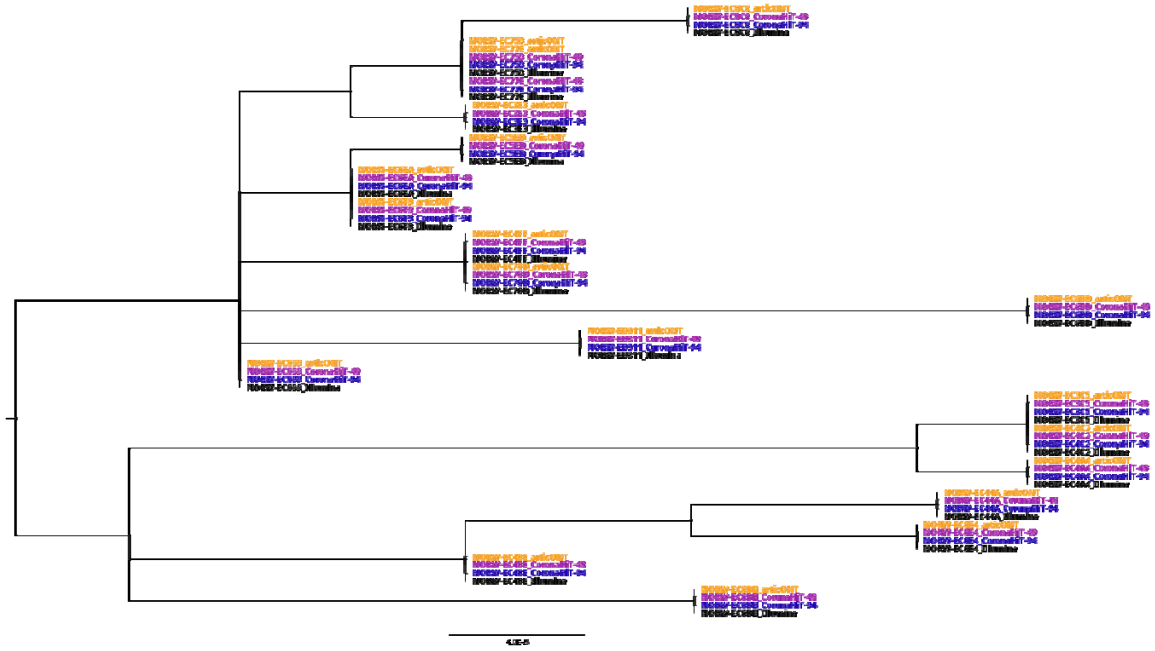


Figure 3: Maximum likelihood tree of the 19 consensus genomes from each sequencing method, showing full agreement between methods.