High-Spatiotemporal-Resolution Nanopore Sequencing of SARS-CoV-2 and Host Cell RNAs

Authors

Dóra Tombácz¹, Ákos Dörmő¹, Gábor Gulyás¹, Zsolt Csabai¹, István Prazsák¹, Balázs Kakuk¹, Ákos Harangozó¹, István Jankovics², Béla Dénes³, Zsolt Boldogkői^{1*}

Affiliations

1. Department of Medical Biology, Albert Szent-Györgyi Medical School, University of Szeged, Szeged, Hungary

2. Complex Medical Center, Budapest, Hungary

3. Veterinary Diagnostic Directorate, National Food Chain Safety Office, Budapest, Hungary
corresponding author(s): Zsolt Boldogkői (boldogkoi.zsolt@med.u-szeged.hu)

Abstract

Recent studies have disclosed the genome, transcriptome and epigenetic compositions of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) and the effect of viral infection on gene expression of the host cells. It has been demonstrated that, besides the major canonical transcripts, the viral genome also codes for non-canonical RNA molecules. While the structural characterizations have revealed a detailed transcriptomic architecture of the virus, the kinetic studies provided poor and often misleading results on the dynamics of both the viral and host transcripts due to the low temporal resolution of the infection event and the low virus/cell ratio (MOI=0.1) applied for the infection. In this study, we used direct cDNA and direct RNA nanopore sequencings for the generation of high-coverage, high-temporal-resolution transcriptomic datasets on SARS-CoV-2 and on primate host cells infected with a high virus titer (MOI=5). Sixteen sampling time points ranging from 1 to 96h with a varying time resolution and three biological replicates were used in the experiment for both the infected and the non-infected cells.

Background & Summary

Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) is a positive-sense single RNA-stranded betacoronavirus and the etiological agent of the current COVID-19 pandemic¹. The replication and the transcription of the RNA genome are interrelated because the same enzyme, an RNA-dependent RNA polymerase (RdRP), carries out both processes². First, negative-sense RNA intermediates are generated to serve as templates for the synthesis of both the genomic RNA (gRNA) and the nested set of subgenomic RNAs (sgRNAs) mapping to the 3'-third of the viral genome³. The gRNA and the sgRNAs have common 3'-termini since the RdRP synthesizes the positive sense RNAs from this end of the genome. Template-switching occurs during the synthesis of the negative-strand of sgRNAs, which is mediated by the transcription-regulating sequences (TRSs) in the genome body (TRS-B) and in the 5'-leader sequence (TRS-L) resulting in the fusion of leader-body sequences^{4,5}. Recent studies have disclosed the transcriptomic architecture of SARS-CoV-2 and the effect of viral infection on the host gene expression⁶. It has been shown that, besides canonical TRS-dependent RNA molecules, the viral genome also codes for non-canonical, TRS-dependent and TRS-independent, transcripts, although in a relative low abundance (altogether <10%). Additionally, investigations of the effect of the viral infection on the transcriptome of various cell types have identified several genes and gene networks⁷.

Nonetheless, the kinetic studies of gene expressions used only a few timepoints for monitoring the infection^{8,9}, which do not provide a comprehensive picture on the temporal dynamics of viral transcriptome. Furthermore, typically a low (0.1) multiplicity of infection (MOI) was applied in the experiments, which may lead to misleading conclusions on the kinetic properties of SARS-CoV-2 transcripts, because after a couple of hours the virus can initiate a new infection cycle within the non-infected cells. Low-MOI-infection makes also difficult to assess the host cell response,

especially in the case of the down-regulated genes. Infections are typically carried out using fresh, rapidly growing cells, and fresh cultures are also used as mock-infected cells. However, gene expression profiles may undergo alterations non-infected cells during the propagation therefore, we cannot decide whether the transcriptional changes in infected are due to the effect of the virus or to the time factor of culturing. This phenomenon is practically never tested in the experiments. An additional problem is the use of short-read sequencing for profiling of the host cell reaction to the viral infection⁷ because this approach has severe limitations for the detection of transcript isoforms, such as splice and length variants, and multigenic transcripts, among others^{10,11,12}.

Long-read sequencing (LRS) opened new avenues for the comprehensive analysis of the transcriptomes, for which the major reason is that these techniques are able to detect full-length RNA molecules and thereby to distinguish between transcript isoforms and transcriptional overlaps. LRS-based studies have revealed a hidden transcriptional complexity in viruses^{13,14}, but this approach has also been used for the analysis of the kinetic properties of viral transcriptomes¹⁵, and the virus-host interaction^{16,17}.

In this study, we applied nanopore sequencing based on direct RNA (dRNA) and direct cDNA (dcDNA) approaches for the generation of transcriptomic datasets from SARS-CoV-2 and primate host (Vero) cells. A mixed time point sample was used for dRNA sequencing, while we used 16 time-point samples within an interval of 1 to 96 h from both infected and non-infected host cells using MOI=5 for the infection. Altogether, we generated almost 64 million long-reads, from which more than 1.8 million reads mapped to the SARS-CoV-2 and almost 48 million to the host reference genome, respectively (**Table 1**). The obtained read count resulted in a very high coverage across the viral genome (**Figure 1**). Detailed data on the read counts, quality of reads including read lengths (**Figure 2**), insertions, deletions, as well as mismatches are summarized in **Supplementary Table 1 and 2**.

Table 1. Summary data of the obtained read counts from dcDNA and dRNA sequencings. Low quality (failed) reads (Q-score <8) were filtered out from the passed reads (Q-score ≥8) by the MinkNOWs (Guppy, ONT) software.

	Quality	Total (infected)	Virus	Host (infected)	Unmapped (infected)	Total (uninfected)	Host (uninfected)	Unmapped (uninfected)
dcDNA	all	32,017,113	1,527,249	23,703,827	6,786,037	29,294,533	22,149,844	7,144,689
dcDNA	≥8	23,607,200	1,280,395	21,246,856	1,079,949	20,360,096	19,008,016	1,352,080
dcDNA	<8	8,409,913	246,854	2,456,971	5,706,088	8,934,437	3,141,828	5,792,609
dRNA	all	2,606,502	281,418	1,999,161	325,923	-	-	-
dRNA	≥8	1,950,595	236,518	1,658,588	55,489	-	-	-
dRNA	<8	655,907	44,900	340,573	270,434	-	-	-

Decoding the transcriptional landscape of SARS-CoV-2 virus is a fundamental step in studying its biology, genetics regulation and molecular pathogenesis. Therefore, in this data descriptor, our aim was to provide a robust, precise, reliable dataset based on LRS approaches for understanding the gene expression and genetic regulation of the causative agent of current pandemic, its effect on differential host gene expression, as well as to provide a rich resource for future functional studies.

Methods

Figure 3 shows the detailed workflow of the study.

Cells

The Vero E6 (African green monkey kidney) cell line was obtained from the American Type Culture Collection (ATCC). The cells were plated at a density of 2×10^6 cells per 75 cm² tissue culture flasks (CELLSTAR® Greiner Bio-One GmbH, Frickenhausen, Germany) in Minimum Essential Medium Eagle culture medium (MEM) with 10% fetal bovine serum (FBS) and 2mM L-glutamine and antibiotic-antimycotic solution (all obtained from Sigma-Aldrich). Vero cells were incubated at 37°C in a humidified 5% CO₂ atmosphere until confluency ($\sim 8 \times 10^6$ cells) was reached. The monolayer was washed once with the serum-free MEM immediately before infection.

Collection, detection and isolation of the virus

The SARS-CoV-2 virus was isolated from the human nasopharyngeal swab of the RT-PCR positive (Ct 22) 77-year-old male patient during the official COVID-19 surveillance program at the Veterinary Diagnostic Directorate of the National Food Chain Safety Office (Budapest, Hungary) with the cooperation of the Complex Medical Center (Budapest) in November 2020 at the second wave of COVID-19 pandemic in Hungary. The patient developed respiratory illness, with fever, cough, and fatigue that quickly progress to pneumonia. The patient was hospitalized, where, unfortunately, he died in a few days. In his story, he did not declare any travel abroad in the last 14 days. At the same time, he traveled relatively frequently within Hungary and had been in close contact with the people with COVID-19.

Detection of SARS-CoV-2 in pharyngeal wash samples was performed using RT-PCR amplification of SARS-CoV-2 N-gene fragments. Two hundred microliters (200 µL) of the pharyngeal washes were first processed for RNA extraction in the Thermo Scientific™ KingFisher™ Flex Purification System (Thermo Fisher Scientific, Waltham, MA USA), using the IndiMag® Pathogen Kit (QIAGEN® GmbH, Hilden, Germany). Subsequently, the detection of N-gene of SARS-CoV-2 was performed by using the 2019-nCoV-2 RUO kit (Integrated DNA Technologies, Inc., Coralville, Iowa, USA) and One-Step RT-PCR Kit (QIAGEN® GmbH) on a Rotor-Gene Q real-time PCR cycler (QIAGEN® GmbH). The amplification protocol consisted of a reverse transcription step at 50°C for 30 minutes, a denaturation step at 95°C for 15 minutes and subsequent 45 cycles at 95°C/56°C/72°C for 30/30/60 seconds, respectively. A positive result was defined as amplification of N-gene in a sample with each cycle threshold value (ct) less than 37.

For the virus isolation, 1 ml of viral transport media from the swab was mixed with 3 ml serum-free MEM culture medium supplemented with 2mM L-glutamine and antibiotic-antimycotic solution and were filtered using Ministar® 0.22µm filter (Sartorius AG, Göttingen, Germany). The filtrate was placed onto cells in a 25 cm² tissue culture flask (Corning®, Corning Inc., New York, USA) of Vero E6 cells, then were incubated at 37°C in a humidified 5% CO₂ chamber for 1 hour. After incubation, two and a half milliliters (2.5 ml) of serum-free MEM culture medium with 2mM L-glutamine and antibiotic-antimycotic solution was added to the tissue culture flask. The inoculated culture was grown in a humidified 37°C incubator with 5% CO₂. Cells were observed daily for cytopathic effect (CPE). On day 4, an 80% CPE was evident, and the cells with supernatant were harvested. This provided the first passage virus. The virus was passaged twice at low MOI in Vero E6 cells to obtain a working stock used in the experiments. The viral titer was determined by plaque assay on Vero E6 cells. Virus stock was stored at -80°C until use.

Propagation of the virus

The virus was passaged twice in Vero cells to obtain a working stock used in all experiments. Viral titer was determined by plaque assay on Vero cells. The virus was diluted into a serum-free MEM. Cells were infected with 5 ml of the SARS-CoV-2 virus with 5 plaque-forming unit (pfu)/cell [multiplicity of infection (MOI=5)], then were incubated at 4°C for 1 h. Non-infected control cultures (mock) were prepared using pure non-supplemented MEM as inoculums. Next, the virus inoculum was removed from the flasks. The monolayer was washed once with the serum-free MEM. Ten milliliters of MEM culture medium supplemented with 3% FBS, 2mM L-glutamine, and antibiotic-antimycotic solution was added to the tissue culture flasks. The cells were incubated at

37°C for 1, 2, 4, 6, 8, 10, 12, 14, 16, 18, 20, 24, 36, 48, 72 and 96 hours in a humidified 5% CO₂ atmosphere. Each time experiment was done in triplicate with a mock-infected control. Mock-infected cells were harvested at the same time points as the infected cells. Following incubation, the medium was removed, and the monolayer was washed once with phosphate-buffered saline (PBS). The tissue culture plates were stored at -80°C until use. Next, the infected cells were treated by lysis buffer, then creped and placed into an Eppendorf Tubes® (Thermo Fisher Scientific Inc.).

RNA purification

Total RNA was extracted from the mock-infected and from the SARS-CoV-2-infected cells at various stages of infection from 1 to 96 hours using the Macherey-Nagel's NucleoSpin RNA Kit according to the manufacturer's protocol. In brief, cells were collected by low-speed centrifugation, then 350µl lysis buffer (RA1 from the Kit), 3.5µl β-Mercapthoethanol (Sigma Aldrich) were added followed by vortexing the samples. Mixtures were loaded onto a NucleoSpin Filter and centrifuged for 1min at 11,000 x g. The filters were discarded and 350µl 70% EtOH was added to the lysate. This was loaded to the NucleoSpin RNA Column and centrifuged at 11,000 x g for 30sec. Membrane was desalted with the addition of 350µl Membrane Desalting Buffer (from the NucleoSpin Kit), then dried with a short centrifugation (11,000 x g). Residual DNA was enzymatically removed [with the usage of 95µl mixture of rDNase:rDNase reaction buffer (1:9 ratio, NucleoSpin Kit) and incubation at room temperature (RT) for 15min]. The rDNase was inactivated with the first washing step, by adding 200µl RAW2 Buffer (NucleoSpin Kit) directly onto the NucleoSpin Filter. After a quick centrifugation (30min, 11,000 x g), the filter was placed in a new tube. Six-hundred µl RAW3 Buffer (NucleoSpin Kit) was added, then spun down as before. This washing step was repeated using 250µl RAW3. Finally, the total RNA bound to the Filter was eluted in 60µl nuclease-free water (NucleoSpin Kit). Samples were quantified by Qubit 4.0 using the Qubit RNA BR Assay Kit (Invitrogen, **Supplementary Table 3a**) and then stored at -80°C until use.

Poly(A) selection

Lexogen's Poly(A) RNA Selection Kit V1.5 was used to isolate polyadenylated RNAs from the total RNA samples. The protocol applies oligo(dT) beads, which capture RNAs with poly(A) stretches (most mRNAs), but RNAs without polyadenylated 3' ends (e.g. 28S and 18S rRNAs and tRNAs) do not hybridize to the beads and therefore, they will be removed during the washing steps. The detailed protocol is as follows: the magnetic beads (part of the Lexogen Kit) was resuspended and 4µl for each RNA samples was measured. Beads were placed in a magnet, they were collected and the supernatant was discarded. Samples were resuspended in 75µl Bead Wash Buffer (Lexogen Kit) and then were placed on the magnet. Supernatant was discarded and this washing step was repeated once. Beads were resuspended in 20µl RNA Hybridization Buffer (part of the Lexogen Kit). Ten ng from the total RNA samples were diluted to 20µl UltraPure™ DNase/RNase-Free Distilled Water (Invitrogen) and then denatured at 60°C for 1min followed by holding them at 25°C. Twenty µl denatured RNA was mixed with 20µl (previously washed and resuspended) beads. The mixtures were incubated at 25°C in a shaker incubator with 1250 rpm agitation. After 20min incubation, sample-containing tubes were placed in a magnetic rack. Supernatant was discarded then the tubes were removed from the magnet. Samples were resuspended in 100µl Bead Wash Buffer (Lexogen Kit) then they were incubated for 5min at 25°C with 1250 rpm agitation. Supernatant was discarded and the washing step was repeated. After the complete removal of the supernatant, beads were resuspended in 12µl UltraPure™ DNase/RNase-Free Distilled Water. Samples were incubated at 70°C for 1min, then the tubes were placed on a magnetic rack. Supernatant, containing the poly(A)+ RNA fraction, was placed to new DNA LoBind (Eppendorf) tubes, the RNA concentration was measured using Qubit RNA HS Assay Kit (Invitrogen, **Supplementary Table 3b**), then samples were stored at -80°C.

ONT – direct cDNA sequencing

For the analysis of the dynamic properties of SARS-CoV-2 RNAs and the effect of viral infection on the host cell transcriptome profile, RNA samples from different time points (1, 2, 4, 6, 8, 10, 12, 14, 16, 18, 20, 24, 36, 48, 72 and 96h p.i., **Supplementary Table 3c**) were used individually for the generation of direct cDNA libraries for nanopore sequencing. The non-amplified cDNA libraries were prepared from sixteen time-points from the mock, and from the coronavirus-infected samples, in three biological replicates using the Direct cDNA Sequencing Kit (SQK-DCS109, ONT) and the appropriate ONT protocol. In short, first-strand cDNAs were generated from the polyA(+) RNAs using the Maxima H Minus Reverse Transcriptase (Thermo Fisher Scientific) with SSP and VN primers (supplied in the kit). The RNase Cocktail Enzyme Mix (Thermo Fisher Scientific) was used to eliminate the potential RNA contamination. Synthesis of the second cDNA strands were carried out with LongAmp Taq Master Mix (New England Biolabs). The double-stranded cDNAs were repaired (NEBNext End repair /dA-tailing Module, New England Biolabs) and adapter ligated (NEB Blunt /TA Ligase Master Mix New England Biolabs). Individual barcode sequences were added to each samples for multiplex sequencing for which the Native Barcoding (12) Kit (ONT) was used as recommended by the manufacturer. The cDNAs and the libraries were washed using AMPure XP beads (Agencourt, Beckman Coulter) after every enzymatic reaction step. The barcode labeled samples were loaded onto MinION R9.4 SpotON Flow Cells (ONT, **Table 2**).

Table 2 List of the sequences of barcodes used for multiplex sequencing. This table also contains the information about the barcoded samples loaded on the same flow cell. A, B, and C represent the biological replicates.

Viral infected				Mock infected				Barcode #	Barcode sequence
Sample #	Flow cell #	Sample #	Flow cell #	Sample #	Flow cell #	Sample #	Flow cell #		
1h/A	1	16h/A	3	1h/A	5	16h/A	7	BC01	AAGAAAGTTGTGCGGTGTCTTTGTG
1h/B		16h/B		1h/B		16h/B		BC02	TCGATTCGGTTTGTAGTCGTCTGT
1h/C		16h/C		1h/C		16h/C		BC03	GAGTCTTGTGTCCAGTTACCAGG
2h/A		18h/A		2h/A		18h/A		BC04	TTCCGATTCTATCGTGTTCCTTA
2h/B		18h/B		2h/B		18h/B		BC05	CTTGTCAGGGTTTGTGTAACTT
2h/C		18h/C		2h/C		18h/C		BC06	TTCTCGCAAAGGCAGAAAGTAGTC
4h/A		20h/A		4h/A		20h/A		BC07	GTGTTACCGTGGGAATGAATCCTT
4h/B		20h/B		4h/B		20h/B		BC08	TTCAGGGAACAAACCAAGTTACGT
4h/C		20h/C		4h/C		20h/C		BC09	AACTAGGCACAGCGAGTCTTGTT
6h/A		24h/A		6h/A		24h/A		BC10	AAGCGTTGAAACCTTTGTCCTCTC
6h/B		24h/B		6h/B		24h/B		BC11	GTTTCATCTATCGGAGGGAATGGA
6h/C		24h/C		6h/C		24h/C		BC12	CAGGTAGAAAAGAAGCAGAACTCGGA
8h/A	2	36h/A	4	8h/A	6	36h/A	8	BC13	AGAACGACTTCCATACTCGTGTGA
8h/B		36h/B		8h/B		36h/B		BC14	AACGAGTCTCTTGGGACCCATAGA
8h/C		36h/C		8h/C		36h/C		BC15	AGGTCTACCTCGTAACACCACTG
10h/A		48h/A		10h/A		48h/A		BC16	CGTCAACTGACAGTGGTTCGTACT
10h/B		48h/B		10h/B		48h/B		BC17	ACCCTCCAGGAAAGTACCTCTGAT
10h/C		48h/C		10h/C		48h/C		BC18	CCAAACCAACAACCTAGATAGGC
12h/A		72h/A		12h/A		72h/A		BC19	GTTCTCTGTCAGTGTCAAGAGAT
12h/B		72h/B		12h/B		72h/B		BC20	TTGCGTCCTGTTACGAGAACTCAT
12h/C		72h/C		12h/C		72h/C		BC21	GAGCCTCTCATTGTCCGTTCTCTA
14h/A		96h/A		14h/A		96h/A		BC22	ACCACTGCCATGTATCAAAGTACG
14h/B		96h/B		14h/B		96h/B		BC23	CTTACTACCCAGTGAACCTCCTCG
14h/C		96h/C		14h/C		96h/C		BC24	GCATAGTCTGCATGATGGGTTAG

208

209 ONT – direct RNA sequencing

210 ONT's Direct RNA sequencing (SQK-RNA002; Version: DRS_9080_v2_revO_14Aug2019, Last
 211 update: 10/06/2021) was used to sequence the native RNA strands from a mixture of polyA(+) RNA fractions (**Supplementary Table 3d**). Five-hundred ng RNA in 9µl nuclease-free water was
 212 mixed with 3µl NEBNext Quick Ligation Reaction Buffer (New England BioLabs), 0.5µl RNA CS (ONT
 213 Kit), 1µl RT Adapter (110nM; ONT Kit) and 1.5µl T4 DNA Ligase (2M U/ml New England BioLabs).
 214 The ligation reaction was carried out for 10min at RT. The synthesis of the first strand cDNA was
 215 conducted using SuperScript III Reverse Transcriptase (Life Technologies), as described in the
 216

217 Direct RNA sequencing (DRS) protocol (ONT). In short, a 50 min incubation at 50°C was followed by
218 the inactivation of the enzyme at 70°C for 10 min. Sequencing adapters from the DRS kit were
219 ligated to the cDNA with the T4 DNA ligase enzyme and NEBNext Quick Ligation Reaction Buffer
220 (New England BioLabs). Ligation was carried out at RT for 10 min. The sample was washed using
221 AMPure XP beads (Agencourt, Beckman Coulter) after every enzymatic reaction. Libraries were
222 sequenced on an R9.4 SpotON Flow Cell.

223 Pre-processing and data analysis

224 The MinION raw data was basecalled using ONT Guppy basecalling software version 5.0.11. using -
225 -qscore_filtering. VirStrain¹⁸ tool was used to identify the closest SARS-CoV-2 strains to our isolate
226 (**Supplementary Table 4**). Then, reads with a Q-score greater or equal to 8 were mapped to the
227 SARS-CoV-2 viral genome (NCBI nucleotide accession: MT560672.1) and to the host
228 (*Chlorocebus_sabaeus* 1.1) genome [GenBank assembly accession: GCA_000409795.2 (latest)] with
229 the Minimap2 aligner¹⁹. The mock-infected samples were also mapped to the host genome. The
230 following programs were also used for analysis: SamTools²⁰ [for generation binary alignment (bam)
231 and indexed (bai) files, as well as to categorize the data into viral-mapped, host-mapped and
232 unmapped], the bamCoverage tool from deepTools²¹ (to generate coverage tracks), GATK²²
233 Picard's DownsampleSam tool which applies a downsampling algorithm to retain only a
234 deterministically random subset of the reads. R package²³ (R script) was used for generate violin
235 plots.

236 Our in-house scripts were used to generate the descriptive statistics of reads:
237 <https://github.com/moldovannorbert/segtools>.

238

239 Data Records

240 All data generated in this study including the unmapped reads as well as reads which do not match
241 our strict criteria (Q-score below 8) can be found in European Nucleotide Archive under the
242 accession number: **PRJEB46891**. Data records and supporting information (e.g. sample IDs,
243 experimental steps performed on each, etc.) are summarized in **Supplementary Table 5**.

244

245 Technical Validation

246 **RNA** The Qubit RNA BR Assay Kit (Invitrogen, Carlsbad, CA, United States) was used to check the
247 amount of total RNA. Qubit RNA HS Assay Kit (Invitrogen, Carlsbad, CA, United States) was used to
248 measure the quantity of the poly(A)+ RNA fractions. The final concentrations of the RNA samples
249 were determined by Qubit[®] 4.

250 **cDNA** The concentrations of the cDNA samples and sequencing ready libraries were measured
251 using the Qubit dsDNA HS Assay Kit (Invitrogen, Carlsbad, CA, United States). The quality of RNA
252 was assessed using the Agilent 2200 TapeStation System. RIN scores ≥ 9.6 were used for
253 sequencing (**Figure 3d**).

254 The cDNAs and the sequencing-ready cDNA libraries were washed using AMPure XP beads
255 (Agencourt, Beckman Coulter) after every enzymatic reaction. The samples for dRNA sequencing
256 were treated with RNAClean XP beads.

257 Three biological replicates were used for each of the 16 time points. To monitor the effect of SARS-
258 CoV-2 infection on the gene expression of the host cells, mock-infected cells were harvested at the
259 same time-points, as the virally-infected cells.

260

261 Usage Notes

262 The datasets provided in this report are useful to investigate the time-course expression of SARS-
263 CoV-2 over a 96 h period of infection which provides a deeper understanding of the molecular
264 biology of the virus (e.g. transcriptional analysis of subgenomic region, analysis of the dynamics of

viral replication, examination of the potential interactions between transcription and replication, as well as to study the potential transcript isoforms of the virus). Our data eliminate the limitations of other SARS-CoV-2 transcriptomic experiments. First, we used a high plaque-forming unit per cell (MOI=5 pfu/cell) for the infection (other studies typically apply 0.1 pfu), therefore the large majority of cells in the culture became infected, and hence the possibility of a second round of infection is excluded. Additionally, due to the high temporal resolution, our data is also useful to precisely measure the alteration of the gene expression of both the virus and the host cell. Third, we provide mock-infected cells, which were harvested in the same time-points, as the virally-infected cells, which allows the identification of gene-network alterations due to the aging of the cell culture and to analyze the temporal changes of gene expression patterns during the cultivation. Virus-host interactions can also be examined. Furthermore, due to the very long reads and high coverage across the viral genome, assembly of this Hungarian isolate and the analysis of potential genome editing events can be achieved from the data. Moreover, the applied direct RNA and direct cDNA sequencing approaches provides independent methods for the validation of novel transcripts. Due to the high coverage, this dataset can also be used for the reconstruction of the viral genome. Finally, this dataset can also be used from various bioinformatics aspects: e.g. the data can be further analyzed or other bioinformatic programs [NanoPack²⁴, SQANTI3²⁵ (<https://github.com/ConesaLab/SQANTI3>), Ira²⁶, LoRTIA²⁷ (<https://github.com/zsolt-balazs/LoRTIA>) or any other programs for LRS data analysis listed in LONG-READ-TOOLS^{28,29}] can be tested on these data. Potential template switching artefacts can be tested using the transcript annotator developed by our group²⁷.

The uploaded binary alignment (BAM) files contain reads already mapped to the SARS-CoV-2 reference genome (MT560672.1), as well as to the host genome (GCA_000409795.2) using Minimap2.

The uploaded reads contain terminal poly(A) sequences as well as the 5' and 3' adapters, which can be used to determine the orientations of the reads.

Code Availability

1. Custom (in-house) scripts: <https://github.com/moldovannorbert/seqtools>
2. deepTools: <https://deeptools.readthedocs.io/en/develop/content/tools/bamCoverage.html?highlight=bamcoverage>
3. GATK: <https://gatk.broadinstitute.org/hc/en-us/articles/4404604771227-DownsampleSam-Picard->
4. Guppy: <https://github.com/nanoporetech/pyguppyclient>
5. Minimap2: <https://github.com/lh3/minimap2>
6. R package: <https://www.r-project.org/>
7. SamTools: <https://github.com/SAMPTools/>
8. VirStrain: <https://github.com/liaoherui/VirStrain>

Acknowledgements

This project was supported by the National Research, Development and Innovation Office (OTKA) K 128247 to ZB, National Research, Development and Innovation Office (OTKA) FK 128252 to DT. This research was also supported by the project "Professional, methodological development of health care system: subproject E – Complex epidemiological and clinical response to Covid-19 pandemic" (EFOP-1.8.0-VEKOP-17-2017-00001) to BD. The APC fee was covered by the University of Szeged Open Access Fund: 5401.

Author contributions

D.T. analyzed the data, took part in Nanopore sequencing, drafted the manuscript and coordinated the project. Á.D. performed Nanopore sequencing and RNA purification. G.G. conducted bioinformatics analysis. Z.C. took part in RNA isolation and sequencing. I.P. participated in RNA isolation and analysis. B.K. carried out bioinformatics. Á.H. participated in sequencing and data analysis. I.J. isolated and propagated the virus. B.D. propagated the virus and the host cells and took part in drafting the manuscript. Z.B. conceived and designed the experiments, supervised the project and wrote the manuscript. All authors read and approved the final paper.

Competing interests

The authors declare no competing interests.

Figure Legends

Figure 1. Whole genome coverage plot using high-quality (Q-score ≥ 8) reads which aligned to the SARS-CoV-2 genome used as references for this study.

Figure 2. Violin plots of read length distribution of the sequencing data derived from infected and uninfected samples. a. Length of reads aligned to the viral **b.** and to the host genome. **c.** Read-length distribution of mock-infected samples mapped to the host genome.

Figure 3. Schematic representation of the workflow applied in this project. **a.** Isolation and detection of a Hungarian isolate of SARS-CoV-2 virus. **b.** Experimental workflow of the study. **c.** Bioinformatics workflow **d.** Quality of RNA samples were detected with a TapeStation 2200 system with RNA ScreenTape. TapeStation gel image shows that intact, high quality RNAs were isolated from the samples and used for sequencing. The image shows the following samples: A1: marker; B1: 8h pi sample C; 12h pi sample A; 16h pi sample A; 18h pi sample B, 20h pi sample C; 36h pi sample A; 48h pi sample A; 96h pi sample B

References

1. Zhou, P. *et al.* A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature* **579**, 270-273 (2020).
2. Woo, P.C., Huang, Y., Lau, S.K. & Yuen, K.Y. Coronavirus genomics and bioinformatics analysis. *Viruses* **2**, 1804–20 (2010).
3. Sola, I., Moreno, J.L., Zúñiga, S., Alonso, S. & Enjuanes, L. Role of nucleotides immediately flanking the transcription-regulating sequence core in coronavirus subgenomic mRNA synthesis. *J. Virol.* **79**, 2506-2516 (2005).
4. Hussain, S. *et al.* Identification of novel subgenomic RNAs and noncanonical transcription initiation signals of severe acute respiratory syndrome coronavirus. *J. Virol.* **79**, 5288-5295 (2005).
5. Sola, I., Almazán, F., Zúñiga, S. & Enjuanes, L. Continuous and Discontinuous RNA Synthesis in Coronaviruses. *Annu. Rev. Virol.* **2**, 265-288 (2015).
6. Kim, D. *et al.* The Architecture of SARS-CoV-2 Transcriptome. *Cell* **181**, 914-921.e10 (2020).
7. Sun, J. *et al.* Comparative Transcriptome Analysis Reveals the Intensive Early Stage Responses of Host Cells to SARS-CoV-2 Infection. *Front. Microbiol.* **11**, 593857 (2020).
8. Chang, J.J.-Y. *et al.* Transcriptional and epi-transcriptional dynamics of SARS-CoV-2 during cellular infection. *Cell Rep.* **35**, 109108 (2021).
9. Wang, D. *et al.* The SARS-CoV-2 subgenome landscape and its novel regulatory features. *Mol. Cell* **81**, 2135–2147 (2021).
10. Steijger, T. *et al.* Assessment of transcript reconstruction methods for RNA-seq. *Nat. Methods* **10**, 1177–1184 (2013).

- 361 11. Tombácz, D. *et al.* Full-length isoform sequencing reveals novel transcripts and substantial
362 transcriptional overlaps in a herpesvirus. *Plos One* **11**, e0162868 (2016).
- 363 12. Prazsák, I. *et al.* Long-read Sequencing Uncovers a Complex Transcriptome Topology in Varicella
364 Zoster Virus. *BMC Genomics* **19**, 873 (2018).
- 365 13. Boldogkői, Z., Moldován, N., Balázs, Z., Snyder, M. & Tombácz, D. Long-read sequencing – a
366 powerful tool in viral transcriptome research. *Trends Microbiol.* **27**, 578-592 (2019).
- 367 14. Tombácz, D., Torma, G., Gulyás, G., Moldován, N., Snyder, M. & Boldogkői, Z. Meta-
368 analytic Approach for Transcriptome Profiling of Herpes Simplex Virus Type 1. *Sci. Data*, **7**,
369 223 (2020).
- 370 15. Tombácz, D. *et al.* Characterization of the Dynamic Transcriptome of a Herpesvirus with
371 Long-read Single Molecule Real-Time Sequencing. *Sci. Rep.* **7**, 43751 (2017).
- 372 16. Maróti, Z. *et al.* Time-course Transcriptome Analysis of Host Cell Response to Poxvirus
373 Infection Using a Dual Long-read Sequencing Approach. *BMC Res. Notes*, **14**, 239 (2021).
- 374 17. Maróti, Z. *et al.* Long-read Time-course Profiling of the Host Cell Response to Herpesvirus
375 Infection Using Nanopore and Synthetic Long-Read Transcriptome Sequencing. *Sci. Rep.*
376 **11**, 14219 (2021).
- 377 18. Liao, H., Cai, D. & Sun, Y. VirStrain: a strain identification tool for RNA viruses. *bioRxiv*,
378 423722 (2020).
- 379 19. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–
380 3100 (2018).
- 381 20. Li, H. *et al.* The Sequence alignment/map (SAM) format and SAMtools. *Bioinformatics* **25**,
382 2078-9 (2009).
- 383 21. Ramírez, F. *et al.* deepTools2: a next generation web server for deep-sequencing data
384 analysis. *Nucleic Acids Res.* **44**, W160-5 (2016).
- 385 22. Van der Auwera, G.A. & O'Connor, B.D. *Genomics in the Cloud: Using Docker, GATK, and*
386 *WDL in Terra* 1st edn (O'Reilly Media, 2020)
- 387 23. R Core Team. R: *A Language and Environment for Statistical Computing*. R Foundation for
388 Statistical Computing, Vienna. <https://www.R-project.org/> (2013).
- 389 24. De Coster, W., D'Hert, S., Schultz, D.T., Cruts, M. & Van Broeckhoven, C. NanoPack:
390 visualizing and processing long-read sequencing data, *Bioinformatics* **34**, 2666–2669
391 (2018).
- 392 25. Tardaguila, M. *et al.* SQANTI: extensive characterization of long-read transcript sequences
393 for quality control in full-length transcriptome identification and quantification. *Genome*
394 *Res.* **28**, 396-411 (2018).
- 395 26. Ren, J. & Chaisson, M.J.P. Ira: A long read aligner for sequences and contigs. *PLoS Comput.*
396 *Biol.* **17**, e1009078 (2021).
- 397 27. Balázs, Z., Tombácz, D., Csabai, Z., Moldován, N., Snyder, M. & Boldogkői, Z. Template-
398 switching artifacts resemble alternative polyadenylation. *BMC Genomics* **20**, 824 (2019).
- 399 28. Amarasinghe, S.L., Ritchie, M.E., & Gouil, Q. long-read-tools.org: an interactive catalogue
400 of analysis methods for long-read sequencing data. *GigaSci.* **10**, giab003 (2021).
- 401 29. Amarasinghe, S.L., Su, S., Dong, X., Zappia, L., Ritchie, M.E., & Gouil, Q. Opportunities and
402 challenges in long-read sequencing data analysis. *Genome Biol.* **21**, 30 (2020).