Direct RNA sequencing and early evolution of SARS-CoV-2

George Taiaroa[1,2], Daniel Rawlinson[1], Leo Featherstone[1], Miranda Pitt[1], Leon Caly[2], Julian Druce[2], Damian Purcell[1], Leigh Harty[1], Thomas Tran[2], Jason Roberts[2], Mike Catton[2], Deborah Williamson[1,3], Lachlan Coin[1•], Sebastian Duchene[1•]

1. Department of Microbiology and Immunology, University of Melbourne at The Peter Doherty Institute for Infection and Immunity, Melbourne, Australia
2. Victorian Infectious Diseases Reference Laboratory, Royal Melbourne Hospital, at the Peter Doherty Institute for Infection and Immunity, Victoria, Australia
3. Department of Microbiology, Royal Melbourne Hospital, Victoria, Australia

- These authors contributed equally to the work

Corresponding author:

George Taiaroa, The Peter Doherty Institute for Infection and Immunity, University of Melbourne and Victorian Infectious Diseases Reference Laboratory, Melbourne, Australia Tel: +61 (0)3 8344 5466. Email: george.taiaroa@unimelb.edu.au

Abstract

The rapid sharing of sequence information as seen throughout the current SARS-CoV-2 epidemic, represents an inflection point for genomic epidemiology. Here we describe aspects of coronavirus evolutionary genetics revealed from these data, and provide the first direct RNA sequence of SARS-CoV-2, detailing coronaviral subgenome-length mRNA architecture.

The ongoing epidemic of 2019 novel coronavirus (now called SARS-CoV-2, causing the disease COVID-19), which originated in Wuhan, China, has been declared a public health emergency of international concern by the World Health Organisation (WHO)[1-4]. SARS-CoV-2 is a positive-sense single-stranded RNA ((+)ssRNA) virus of the Coronaviridae family, with related Betacoronaviruses capable of infecting mammalian and avian hosts, resulting in illness in humans such as Middle East respiratory syndrome (MERS) and the original severe acute respiratory syndrome (SARS)[2, 5-6]. Based on limited sampling of potential reservoir species, SARS-CoV-2 has been found to be most similar to bat coronaviruses at the genomic level, potentially indicating that bats are its natural reservoir [7-8].

Following the emergence of SARS-CoV-2, genomic analyses have played a key role in the public health response by informing the design of appropriate molecular diagnostics and corroborating epidemiological efforts to trace contacts [8-10]. Taken together, publicly available sequence data suggest a recently occurring, point-source outbreak, as described in online sources [10-12]. Aspects of the response make the assumption that the genetics of SARS-CoV-2, including mechanisms of gene expression and molecular evolutionary rates, are comparable with previously characterised coronaviruses [11-12]. It remains highly relevant to validate these assumptions experimentally with SARS-CoV-2-specific data, with the potential to reveal further insights into the biology of this emergent pathogen. To address this, we describe (i) the architecture of the coronaviral subgenome-length mRNAs, and (ii) phylogenetic approaches able to provide robust estimates of coronaviral evolutionary rates and timescales at this early stage of the outbreak.

Characterised coronaviral species produce a nested set of polyadenylated subgenome-length mRNA transcripts through a mechanism termed discontinuous extension of minus strands that yields mRNA transcripts of different length. The discontinuous transcription mechanism repositions the 5′ leader sequence upstream of consecutive viral open-reading frames (ORF) where each translation start site becomes located at the primary position for ribosome scanning (Figure 1a). Subgenome-length mRNAs have a common 5′ leader sequence, near-identical to that located in the 5′-UTR of the viral genome, with the genome-length RNAs also having an mRNA function [13-14]. Discontinuous copy-choice RNA recombination modulates the expression of coronaviral genes [14-15]. Standard sequencing technologies for RNA viruses are unable to produce reads representing (i) RNA viral genomes or (ii) subgenome-length mRNAs, as they generate short read lengths and have a reliance on amplification to generate complementary DNA (cDNA) sequences.

To define the architecture of the coronaviral subgenome-length mRNAs, a recently established direct RNA sequencing approach was used, based on a highly parallel array of nanopores [16]. In brief, nucleic acids were prepared from culture material with high levels of SARS-CoV-2 growth, and sequenced with use of poly(T) adaptors and an R9.4 flowcell on a

GridION platform (Oxford Nanopore Technologies). Through this approach, the electronic current is measured as individual strands of RNA translocate through a nanopore, with derived signal-space data basecalled to infer the corresponding nucleobases (Supplementary Methods). The SARS-CoV-2 sample produced 680,347 reads, comprising 860Mb of sequence information in 40 hours of sequencing (BioProject PRJNA608224). Aligning to the genome of the cultured SARS-CoV-2 isolate (MT007544.1), a subset of reads were attributed to coronaviruses sequence (28.9%), comprising 367Mb of sequence distributed across the 29,893 base genome. Of these, a number had lengths >20,000 bases, capturing the majority of the SARS-CoV-2 genome on a single molecule. Together, direct RNA sequencing provided an average 12,230 fold coverage of the coronaviral genome, biased towards sequences proximal to the polyadenylated 3’ tail; coverage ranged from 34 fold to >160,000 fold (Figure 1b), with the bias reflecting the higher abundance of subgenome-length mRNAs carrying these sequences, as well as the directional sequencing from the polyadenylated 3’ tail.

SARS-CoV-2 features captured in direct RNA sequence data include subgenome-length mRNAs, as well as RNA base modifications. The shared 5’-leader sequence was used as a marker to identify each subgenome-length mRNA (Supplementary Methods). In SARS-CoV-2, we observed eight major viral mRNAs, in addition to the viral genome (Figure 1c). Each annotated gene was observed as a distinct subgenome-length mRNA, positioned consecutively 3’ to the replicase polyprotein (ORF1a and 1ab) encoded by the genomic-length mRNA, with ORF10 being the last predicted coding sequence upstream of the poly-A sequence (Supplementary Figure 1). ORF10 is the shortest of the predicted coding sequences at 117 bases in length but despite having proximity to the poly-A sequence, was not found as a subgenome-length mRNA through our direct RNA sequence technique. ORF10 has no annotated function, and the putative encoded peptide does not have a homolog in the SARS-CoV proteome (Proteome ID: UP000000354). These data suggest that the sequence currently annotated as ORF10 does not have a protein coding function in SARS-CoV-2. The sequence annotated as ORF10 is immediately upstream of the 3’ UTR and, rather than coding, may act itself or as a precursor of other RNAs in the regulation of gene expression, replication or modulating cellular antiviral pathways. A small number of cell culture-derived isolates in public databases demonstrate shared deletion of an area of the 3’ UTR (Supplementary Figure 2), this parallel molecular evolution suggesting the region may have functional roles in vivo.

In addition to methylation at the 5' cap structure and 3' polyadenylation needed for efficient translation of viral coding sequences, other RNA modifications may have functional roles in SARS-CoV-2 [17]. A range of modifications may be identified using direct RNA sequence data [16-17]; our available SARS-CoV-2 direct RNA sequence data providing adequate coverage to confidently call specific modifications. Through analysis of signal-space data, we identified 42 positions with predicted 5-methylcytosine modifications, appearing at consistent positions between subgenome-length mRNAs (Supplementary Figure 3, and Table 1). In other positive ssRNA viruses, RNA methylation can change dynamically during the course of infection [18], influencing host-pathogen interaction and viral replication. Other modifications may become apparent once training datasets are available for direct RNA sequence data, with little known of the epitranscriptomic landscape of coronaviruses [17,19].

As well as investigating the above assumed features of SARS-CoV-2 genetics, sequence data also enable an estimate of the molecular evolutionary rate, through analysis of publicly available SARS-CoV-2 genome sequences. Evolutionary rate estimates from other coronaviruses such as Middle East Respiratory Syndrome (MERS) are not necessarily applicable here, particularly because MERS had multiple independent introductions into humans [20-22]. To estimate the evolutionary rate and time of origin of the SARS-CoV-2 outbreak, we carried out Bayesian phylogenetic analyses using a curated set of high quality publicly available SARS-CoV-2 genome sequences, each having a known collection date (Supplementary Table 2). The sampling times were sufficient to calibrate a molecular clock and infer the evolutionary rate and timescale of the outbreak; the evolutionary rate of SARS-CoV-2 was estimated to be 1.16×10^{-3} substitutions/site/year (95% HPD 6.32×10^{-4} - 1.69×10^{-3}), and the of time of origin to be early December 2019 (95% HPD November 2019, December 2019), which is in agreement with epidemiological evidence and other recent analyses (Figure 2a) [1-4, 23-24]. Our estimate of the evolutionary rate of SARS-CoV-2 is in line with those of other coronaviruses (Figure 2b), and the low genomic diversity and recent timescale of the outbreak support a recently occurring, point-source transfer to humans.

In summary, insights into the molecular biology of SARS-CoV-2 are revealed through the use of direct RNA sequence data, enabling a detailed view of viral subgenome-length mRNA architecture. Other insights are gleaned from publicly available sequence data, including an estimate of the molecular evolutionary rate of SARS-CoV-2, shown to be similar to those of previously studied coronaviruses and providing further information to support the forecasting and modelling of ongoing SARS-CoV-2 transmission and evolution.

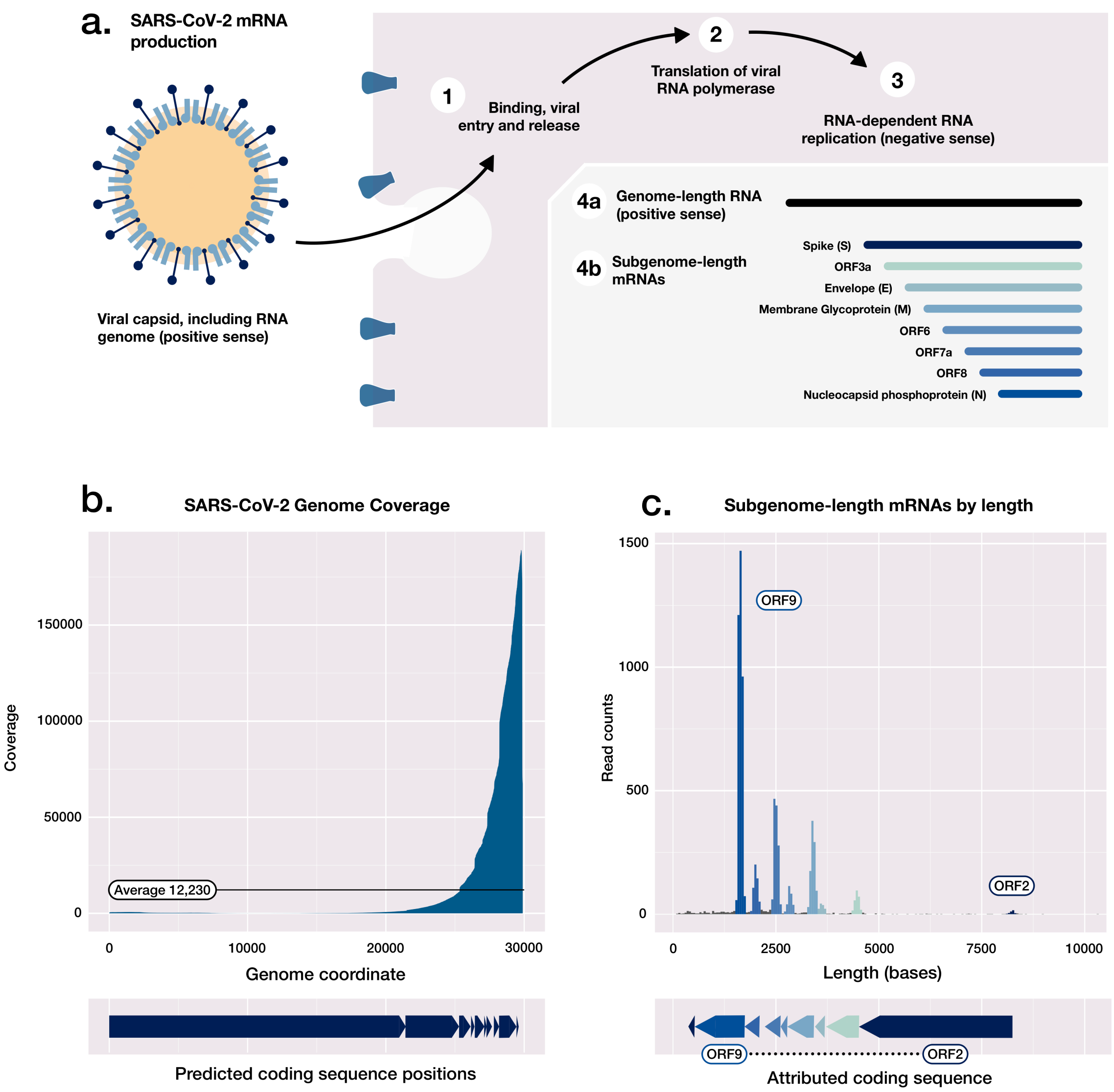
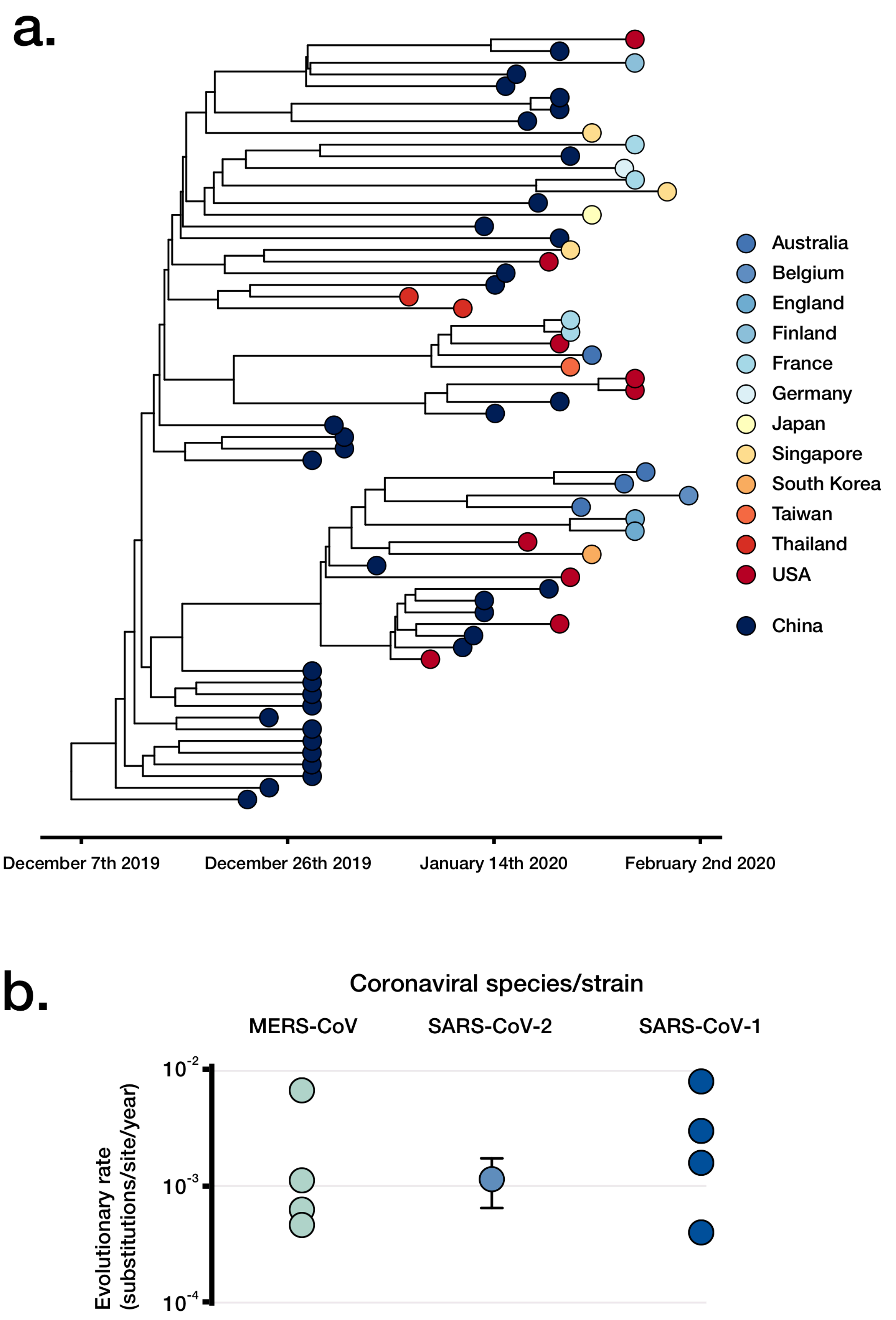


Figure 1. SARS-CoV-2 genetics and subgenome-length mRNA architecture.

A) Schematic of the early stages of SARS-CoV-2 replication, including *in vivo* synthesis of positive sense genome-length RNA molecules, and subgenome-length mRNAs. B) Read coverage of pooled direct RNA reads aligned to the SARS-CoV-2 genome (29,893 bases), showing a bias towards the 3' polyadenylated end. C) Read length histogram, showing subgenome-length mRNAs attributed to coding sequences.



148

149 **Figure 2. Assessment of viral evolutionary rate and outbreak timing with SARS-CoV-2-**
150 **specific data.** A) A timed highest clade-credibility phylogenetic tree of curated SARS-CoV-2
151 genomes as inferred in BEAST. B) Comparison of the SARS-CoV-2 rate estimate and
152 previously published estimates of other coronaviruses.

References

[1] World Health Organization. Pneumonia of unknown cause — China. 2020 (https://www.who.int/csr/don/05-january-2020-pneumonia-of-unkown-cause-china/en/).

[2] World Health Organization. Novel coronavirus — China. 2020 (https://www.who.int/csr/don/12-january-2020-novel-coronavirus-china/en/).

[3] Dong, Ensheng, Hongru Du, and Lauren Gardner. "An interactive web-based dashboard to track COVID-19 in real time." The Lancet Infectious Diseases (2020). DOI: 10.1016/S1473-3099(20)30120-1

[4] World Health Organization. Statement on the second meeting of the International Health Regulations (2005) Emergency Committee regarding the outbreak of novel coronavirus (2019-nCoV). January 30, 2020 (https://www.who.int/news-room/detail/30-01-2020-statement-on-the-second-meeting-of-the-international-health-regulations-(2005)-emergency-committee-regarding-the-outbreak-of-novel-coronavirus-(2019-ncov).

[5] Peiris JS, Yuen KY, Osterhaus AD, Stöhr K. The severe acute respiratory syndrome. New England Journal of Medicine. 2003 Dec 18;349(25):2431-41. DOI: 10.1056/NEJMra032498

[6] Assiri A, McGeer A, Perl TM, Price CS, Al Rabeeah AA, Cummings DA, Alabdullatif ZN, Assad M, Almulhim A, Makhdoom H, Madani H. Hospital outbreak of Middle East respiratory syndrome coronavirus. New England Journal of Medicine. 2013 Aug 1;369(5):407-16. DOI: 10.1056/NEJMoa1306742

[7] Perlman, S. Another decade, another coronavirus. New England Journal of Medicine. 2020 February 20; 382:760-762. DOI: 10.1056/NEJMe2001126

[8] Lu, Roujian, et al. "Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding." The Lancet (2020). DOI: 10.1016/S0140-6736(20)30251-8

[9] Shu Y, McCauley J. GISAID: Global initiative on sharing all influenza data–from vision to reality. Eurosurveillance. 2017 Mar 30;22(13). https://www.gisaid.org/

[10] Hadfield J, Megill C, Bell SM, Huddleston J, Potter B, Callender C, Sagulenko P, Bedford T, Neher RA. Nextstrain: real-time tracking of pathogen evolution. Bioinformatics. 2018 Dec 1;34(23):4121-3. https://nextstrain.org/ncov

[11] Andersen KG, Rambaut A, Lipkin WI, Holmes EC, Garry RF. The Proximal Origin of SARS-CoV-2. Virological, accessed on 27/02/2020. http://virological.org/t/the-proximal-origin-of-sars-cov-2/398

[12] Rambaut A. Phylodynamic Analysis | 129 genomes | 24 Feb 2020. Virological, accessed 27/02/2020. http://virological.org/t/phylodynamic-analysis-129-genomes-24-feb-2020/356

[13] Yount B, Curtis KM, Fritz EA, Hensley LE, Jahrling PB, Prentice E, Denison MR, Geisbert TW, Baric RS. Reverse genetics with a full-length infectious cDNA of severe acute respiratory syndrome coronavirus. Proceedings of the National Academy of Sciences. 2003 Oct 28;100(22):12995-3000. DOI: 10.1073/pnas.1735582100

[14] Brian DA, Baric RS. Coronavirus genome structure and replication. InCoronavirus replication and reverse genetics 2005 (pp. 1-30). Springer, Berlin, Heidelberg.

[15] Chen Y, Cai H, Xiang N, Tien P, Ahola T, Guo D. Functional screen reveals SARS coronavirus nonstructural protein nsp14 as a novel cap N7 methyltransferase. Proceedings of the National Academy of Sciences. 2009 Mar 3;106(9):3484-9. DOI: 10.1073/pnas.0808790106

[16] Garalde DR, Snell EA, Jachimowicz D, Sipos B, Lloyd JH, Bruce M, Pantic N, Admassu T, James P, Warland A, Jordan M. Highly parallel direct RNA sequencing on an array of nanopores. Nature methods. 2018 Mar;15(3):201. DOI: 10.1038/nmeth.4577

[17] Viehweger A, Krautwurst S, Lamkiewicz K, Madhugiri R, Ziebuhr J, Hölzer M, Marz M. Direct RNA nanopore sequencing of full-length coronavirus genomes provides novel insights into structural variants and enables modification analysis. Genome research. 2019 Sep 1;29(9):1545-54. DOI: 10.1101/gr.247064.118

[18] Lichinchi G, Zhao BS, Wu Y, Lu Z, Qin Y, He C, Rana TM. Dynamics of human and viral RNA methylation during Zika virus infection. Cell host & microbe. 2016 Nov 9;20(5):666-73. DOI: 10.1016/j.chom.2016.10.002

[19] Denison MR, Graham RL, Donaldson EF, Eckerle LD, Baric RS. Coronaviruses: an RNA proofreading machine regulates replication fidelity and diversity. RNA biology. 2011 Mar 1;8(2):270-9. DOI: 10.4161/rna.8.2.15013

[20] Chinese SARS Molecular Epidemiology Consortium. Molecular evolution of the SARS coronavirus during the course of the SARS epidemic in China. Science. 2004 Mar 12;303(5664):1666-9. DOI: 10.1126/science.1092002

[21] Dudas G, Carvalho LM, Rambaut A, Bedford T. MERS-CoV spillover at the camel-human interface. Elife. 2018 Jan 16;7:e31257. DOI: 10.7554/eLife.31257

[22] Cotten M, Watson SJ, Kellam P, Al-Rabeeah AA, Makhdoom HQ, Assiri A, Al-Tawfiq JA, Alhakeem RF, Madani H, AlRabiah FA, Al Hajjar S. Transmission and evolution of the Middle East respiratory syndrome coronavirus in Saudi Arabia: a descriptive genomic study. The Lancet. 2013 Dec 14;382(9909):1993-2002. DOI: 10.1016/S0140-6736(13)61887-5

[23] Andersen K. Clock and TMRCA based on 27 genomes. Virological, accessed on 27/02/2020. http://virological.org/t/clock-and-tmrca-based-on-27-genomes/347

[24] Bedford, T. Phylodynamic estimation of incidence and prevalence of novel coronavirus (nCoV) infections through time. Virological, accessed on 27/02/2020. http://virological.org/t/phylodynamic-estimation-of-incidence-and-prevalence-of-novel-coronavirus-ncov-infections-through-time/391

Methods

SARS-CoV-2 virus sample

The SARS-CoV-2 material prepared for this work is of the first Australian case of Covid-2019 (Australia/VIC01/2020), maintained in cell culture. In brief, African green monkey kidney cells expressing the human signalling lymphocytic activation molecule (SLAM; termed Vero/hSLAM cells accordingly) and SARS-CoV-2 were grown at 37°C at 5% CO_2 in media consisting of 10 mL Earle's minimum essential medium, 7% FBS (Bovogen Biologicals, Keilor East, Aus), 2 mM L-Glutamine, 1 mM Sodium pyruvate, 1500 mg/L sodium bicarbonate, 15 mM HEPES and 0.4 mg/ml geneticin in 25cm^2 flasks. The genome of the cultured isolate (MT007544.1) has three single nucleotide variants (T19065C, T22303G, G26144T) relative to the SARS-CoV-2 Wuhan-Hu-1 reference genome (MN908947.3), and a 10 base deletion in the 3' UTR. The T22303G and 3' UTR variants have been confirmed as culture-derived through Sanger sequencing of clinical and culture material.

Nucleic acids were prepared from infected cellular material, following inactivation with linear acrylamide and ethanol. RNA was extracted from a modest cell pellet (~200mg) using manually prepared wide-bore pipette tips and minimal steps to maintain RNA length for long read sequencing, and a QIAamp Viral RNA Mini Kit (Qiagen, Hilden, Germany). Carrier RNA was not added to Buffer AVL, with 1% linear acrylamide (Life Technologies, Carlsbad, CA, USA) added instead. Wash buffer AW1 was omitted from the purification stage, with RNA eluted in 50 µl of nuclease free water, followed by DNase treatment with Turbo DNase (Thermo Fisher Scientific, Waltham, MA, USA) 37°C for 30 min. RNA was cleaned and concentrated to 10 µl using the RNA Clean & Concentrator-5 kit (Zymo Research, Irvine, CA, USA), as per manufacturer's instructions.

Nanopore sequencing of direct RNA

Prepared RNA (~1*µ*g) was carried into a direct RNA sequence library preparation with the Oxford Nanopore DRS protocol (SQK- RNA002, Oxford Nanopore Technologies) following the manufacturer's specifications, although without addition of the Control RNA. The library was loaded on an R9.4 flow cell and sequenced on a GridION device (Oxford Nanopore Technologies), with the sequencing run ending after 40 hours. Signal space data was used to generate nucleobase sequences ('basecalled') using ont-guppy-for-gridion 3.0.6. Both signal space and basecalled read data are available at BioProject PRJNA608224. It should be noted that non-polyadenylated RNAs are not expected to be detected with this approach.

Characterisation of SARS-CoV-2 subgenome-length mRNA architecture

Direct RNA reads passing the above thresholds were aligned to the genome of the cultured Australian SARS-COV-2 isolate (MT007544.1), with parallel and concordant analyses in Geneious Prime (2019.2.1, [M1]) and minimap2 v 2.11 using the "spliced" preset [M2]. Coverage statistics were determined from the resulting read alignments.

To identify intact subgenome-length mRNAs, reads were aligned to a 62 base SARS-COV-2 leader sequence (5'ACCUUCCCAGGUAACAAACCAACCAACUUUCGAUCUCUUGUAGAU CUGUUCUCUAAACGAAC), with reads aligning to the leader sequence being pooled and visualized in a length histogram. Significant peaks were identified visually and confirmed with a smoothed z-score algorithm. Reads captured in this binning-by-length strategy were re-aligned to the reference genome using the above methods and visualized in Tablet [M3]. Subgenome bins were refined to remove reads which did not originate at the 3' poly-A tail as expected for intact subgenome-length mRNAs. Subgenome bins were re-aligned, with coverage calculated in SAMtools [M4], and plotted using ggplot2 [M5] in R [M6]. The IPKnot webserver [M7] was used to predict the RNA secondary structures, and the VARNA visualization applet [M8] used to produce schematics.

Identification of 5mC methylation

Nanopore sequencing preserves *in vivo* base modifications and enables their detection from raw voltage signal information. In brief, the signal space fast5 files corresponding to identified subgenome-length mRNAs were assessed to identify signal changes corresponding to 5mC methylation. These were first retrieved using the fast5_fetcher_multi function in SquiggleKit [M9]. Reads were processed to align raw signal with basecalled sequence data using Tombo v1.5 [https://github.com/nanoporetech/tombo]. Canonical reference sequences were made for each subgenome-length mRNAs, with the binned fast5 files input into the detect_modifications function, with 5mC as the alternate-model parameter. Outputs of which were converted to dampened_fraction wiggle files and exported for visualization and analysis.

SARS-CoV-2 Phylogenetics

In order to estimate the evolutionary rate and time of origin of SARS-CoV-2, we carried out phylogenetic analyses in BEAST v1.101 [M10], with a curated set of 66 complete and high quality SARS-CoV-2 genomes with date of collection data, as available on February 10th from GISAID and GenBank (Supplementary Table 2). Temporal signal was assessed using BETS [M11]. Initially we determined whether the evolutionary signal and time over which the genome data were collected was sufficient to calibrate the molecular clock, allowing for the evolutionary rate and timescale of the outbreak to be inferred. The model selection approach from BETS supported a strict molecular clock model with genome sampling times for calibration and a coalescent exponential tree prior, which posits that the number of infected individuals grows exponentially over time. We used the HKY+Γ substitution model, and set the following priors for key parameters:

- A continuous time Markov chain for the evolutionary rate
- A Laplace distribution with mean of 0 and scale of 100 for the growth rate
- An exponential distribution with mean of 1 for the effective population size.

A Markov chain Monte Carlo of length 10^7 was set, sampling every 10^3 steps, and assessed sufficient sampling by verifying that the effective sample size for all parameters was at least 200 as determined in Tracer [M12], automatically discarding 10% of the burn in. We summarised the posterior distribution of phylogenetic trees by selecting the highest clade credibility tree alongside calculating posterior node probabilities and the distribution of node ages. Comparison to other coronaviral evolutionary rates included studies [M13-20].

[M1] Kearse M, Moir R, Wilson A, Stones-Havas S, Cheung M, Sturrock S, Buxton S, Cooper A, Markowitz S, Duran C, Thierer T. Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. Bioinformatics. 2012 Jun 15;28(12):1647-9.

[M2] Li H. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics. 2018 Sep 15;34(18):3094-100.

[M3] Milne I, Bayer M, Stephen G, Cardle L, Marshall D. Tablet: visualizing next-generation sequence assemblies and mappings. InPlant Bioinformatics 2016 (pp. 253-268). Humana Press, New York, NY.

[M4] Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. The sequence alignment/map format and SAMtools. Bioinformatics. 2009 Aug 15;25(16):2078-9.

[M5] Wickham H. ggplot2: elegant graphics for data analysis. Springer; 2016 Jun 8

[M6] Team RC. R: A language and environment for statistical computing.

[M7] Sato, Kengo, et al. "IPknot: fast and accurate prediction of RNA secondary structures with pseudoknots using integer programming." Bioinformatics 27.13 (2011): i85-i93.

[M8] Darty, Kévin, Alain Denise, and Yann Ponty. "VARNA: Interactive drawing and editing of the RNA secondary structure." Bioinformatics 25.15 (2009): 1974.

[M9] Ferguson JM, Smith MA. SquiggleKit: A toolkit for manipulating nanopore signal data. Bioinformatics. 2019 Dec 15;35(24):5372-3.

[M10] Drummond AJ, Rambaut A. BEAST: Bayesian evolutionary analysis by sampling trees. BMC evolutionary biology. 2007 Dec;7(1):214.

[M11] Duchene S, Stadler T, Ho SY, Duchene DA, Dhanasekaran V, Baele G. Bayesian Evaluation of Temporal Signal in Measurably Evolving Populations. bioRxiv. 2019 Jan 1:810697.

[M12] Rambaut A, Drummond AJ, Xie D, Baele G, Suchard MA. Posterior summarization in Bayesian phylogenetics using Tracer 1.7. Systematic biology. 2018 Sep;67(5):901.

[M13] Cotten M, Watson SJ, Kellam P, Al-Rabeeah AA, Makhdoom HQ, Assiri A, Al-Tawfiq JA, Alhakeem RF, Madani H, AlRabiah FA, Al Hajjar S. Transmission and evolution of the Middle East respiratory syndrome coronavirus in Saudi Arabia: a descriptive genomic study. The Lancet. 2013 Dec 14;382(9909):1993-2002.

[M14] Cotten M, Watson SJ, Zumla AI, Makhdoom HQ, Palser AL, Ong SH, Al Rabeeah AA, Alhakeem RF, Assiri A, Al-Tawfiq JA, Albarrak A. Spread, circulation, and evolution of the Middle East respiratory syndrome coronavirus. MBio. 2014 Feb 28;5(1):e01062-13.

[M15] Dudas G, Carvalho LM, Rambaut A, Bedford T. MERS-CoV spillover at the camel-human interface. Elife. 2018 Jan 16;7:e31257.

[M16] Zhao Z, Li H, Wu X, Zhong Y, Zhang K, Zhang YP, Boerwinkle E, Fu YX. Moderate mutation rate in the SARS coronavirus genome and its implications. BMC evolutionary biology. 2004 Dec 1;4(1):21.

[M17] Salemi M, Fitch WM, Ciccozzi M, Ruiz-Alvarez MJ, Rezza G, Lewis MJ. Severe acute respiratory syndrome coronavirus sequence characteristics and evolutionary rate estimate from maximum likelihood analysis. Journal of virology. 2004 Feb 1;78(3):1602-3.

[M18] Wu SF, Du CJ, Wan P, Chen TG, Li JQ, Li D, Zeng YJ, Zhu YP, He FC. The genome comparison of SARS-CoV and other coronaviruses. Yi chuan= Hereditas. 2003 Jul;25(4):373-82.

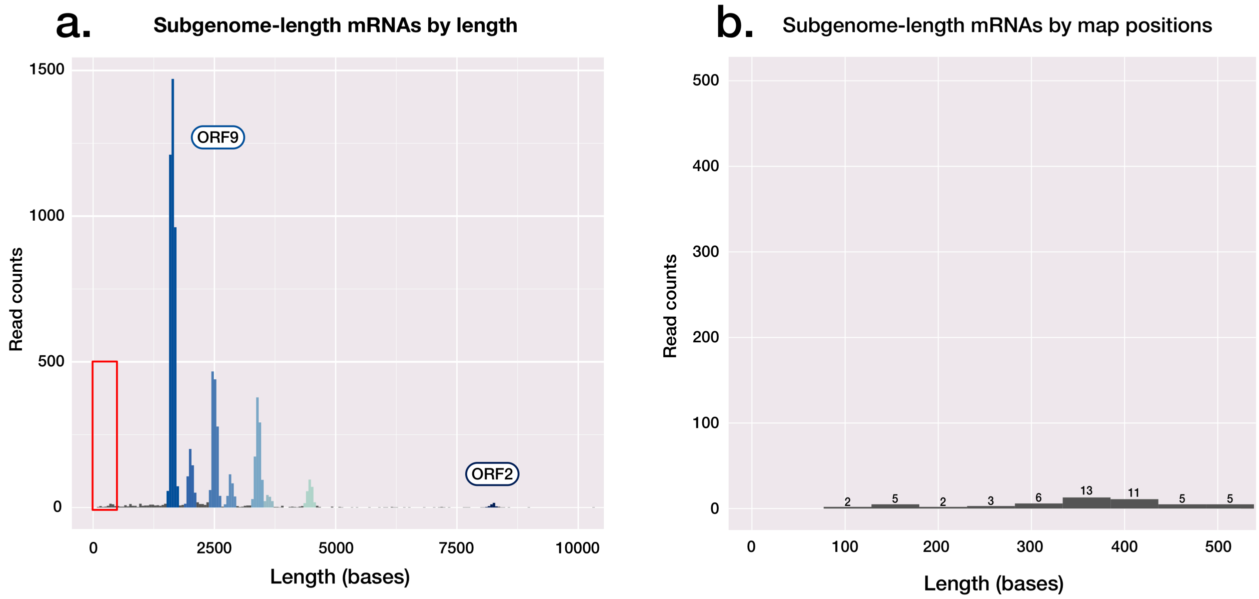
[M19] Chinese SARS Molecular Epidemiology Consortium. Molecular evolution of the SARS coronavirus during the course of the SARS epidemic in China. Science. 2004 Mar 12;303(5664):1666-9.

[M20] Sanjuán R. From molecular genetics to phylodynamics: evolutionary relevance of mutation rates across viruses. PLoS pathogens. 2012 May;8(5).

Acknowledgements

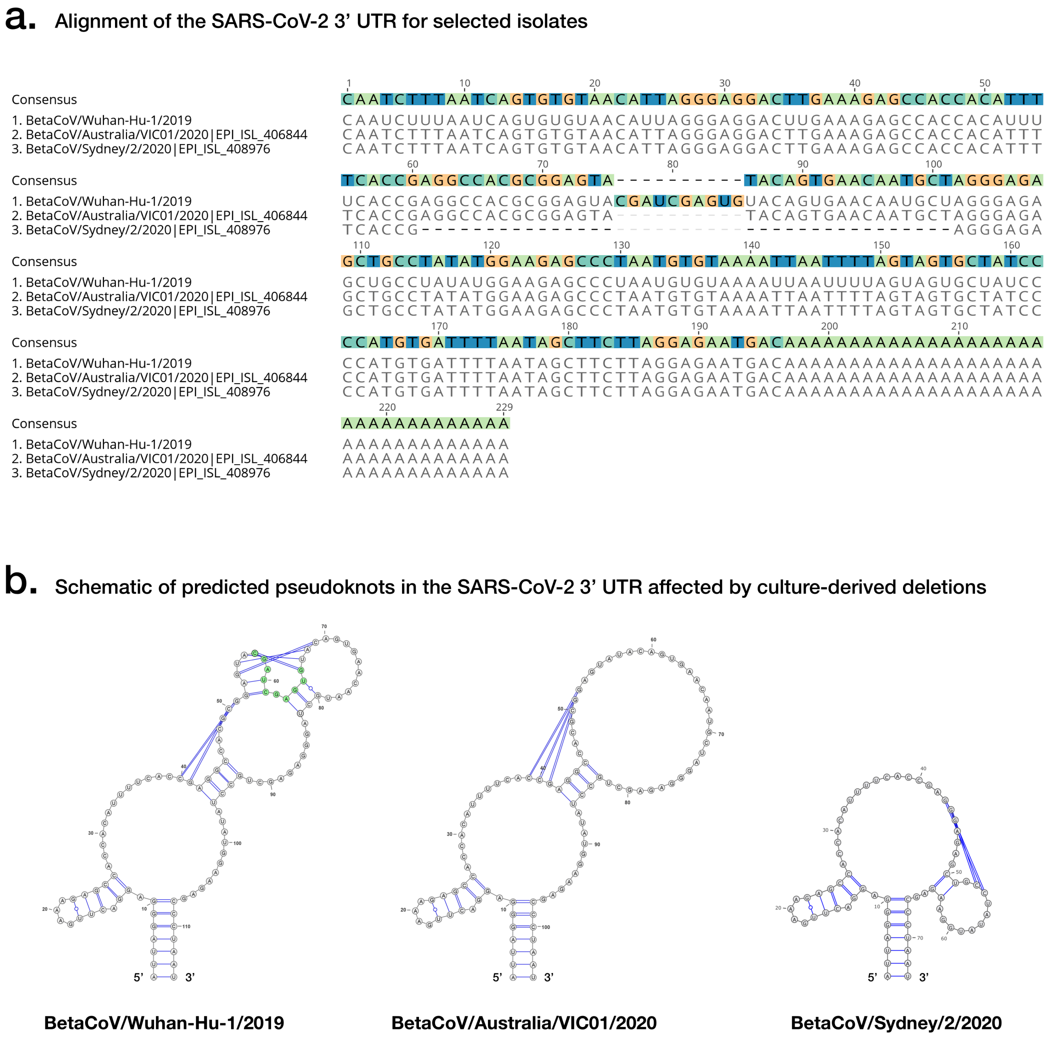
The authors gratefully acknowledge the traditional peoples of the land on which the work was carried out, the Wurundjeri Woi-wurrung people of the Kulin nation.

The authors also acknowledge the broader staff of the Victorian Infectious Disease Reference Laboratory (VIDRL), their public health partners, and VIDRL's major funder, the Victorian Department of Health and Human Services. Lastly, the authors gratefully acknowledge the public health and academic groups contributing to the preparation, release and analysis of SARS-CoV-2 sequence data, including Nextstrain, GISAID and Virological.



Supplementary Figure 1

Absence of observed coding potential for ORF10 in SARS-CoV-2. A) Read length histogram, showing subgenome-length mRNAs attributed to coding sequences, with the area highlighted shown in detail in a second panel. B) Read length histogram, showing read counts of lengths corresponding to those of the ORF10 subgenome-mRNA (~360 bases), if present in the dataset. Of the <500 base reads shown, none align to ORF10.



a. Alignment of the SARS-CoV-2 3' UTR for selected isolates

```
                                            1        10        20        30        40        50
Consensus                                   CAATCTTTAATCAGTGTGTAACATTAGGGAGGACTTGAAAGAGCCACCACATTT
1. BetaCoV/Wuhan-Hu-1/2019                  CAAUCUUUAAUCAGUGUGUAACAUUAGGGAGGACUUGAAAGAGCCACCACAUUU
2. BetaCoV/Australia/VIC01/2020|EPI_ISL_406844 CAATCTTTAATCAGTGTGTAACATTAGGGAGGACTTGAAAGAGCCACCACATTT
3. BetaCoV/Sydney/2/2020|EPI_ISL_408976     CAATCTTTAATCAGTGTGTAACATTAGGGAGGACTTGAAAGAGCCACCACATTT
                                                  60        70        80        90        100
Consensus                                   TCACCGAGGCCACGCGGAGTA----------TACAGTGAACAATGCTAGGGAGA
1. BetaCoV/Wuhan-Hu-1/2019                  UCACCGAGGCCACGCGGAGUACGAUCGAGUGUACAGUGAACAAUGCUAGGGAGA
2. BetaCoV/Australia/VIC01/2020|EPI_ISL_406844 TCACCGAGGCCACGCGGAGTA----------TACAGTGAACAATGCTAGGGAGA
3. BetaCoV/Sydney/2/2020|EPI_ISL_408976     TCACCG-----------------------------------------AGGGAGA
                                              110       120       130       140       150       160
Consensus                                   GCTGCCTATATGGAAGAGCCCTAATGTGTAAAATTAATTTTAGTAGTGCTATCC
1. BetaCoV/Wuhan-Hu-1/2019                  GCUGCCUAUAUGGAAGAGCCCUAAUGUGUAAAAUUAAUUUUAGUAGUGCUAUCC
2. BetaCoV/Australia/VIC01/2020|EPI_ISL_406844 GCTGCCTATATGGAAGAGCCCTAATGTGTAAAATTAATTTTAGTAGTGCTATCC
3. BetaCoV/Sydney/2/2020|EPI_ISL_408976     GCTGCCTATATGGAAGAGCCCTAATGTGTAAAATTAATTTTAGTAGTGCTATCC
                                                    170       180       190       200       210
Consensus                                   CCATGTGATTTTAATAGCTTCTTAGGAGAATGACAAAAAAAAAAAAAAAAAAAA
1. BetaCoV/Wuhan-Hu-1/2019                  CCATGTGATTTTAATAGCTTCTTAGGAGAATGACAAAAAAAAAAAAAAAAAAAA
2. BetaCoV/Australia/VIC01/2020|EPI_ISL_406844 CCATGTGATTTTAATAGCTTCTTAGGAGAATGACAAAAAAAAAAAAAAAAAAAA
3. BetaCoV/Sydney/2/2020|EPI_ISL_408976     CCATGTGATTTTAATAGCTTCTTAGGAGAATGACAAAAAAAAAAAAAAAAAAAA
                                               220      229
Consensus                                   AAAAAAAAAAAAA
1. BetaCoV/Wuhan-Hu-1/2019                  AAAAAAAAAAAAA
2. BetaCoV/Australia/VIC01/2020|EPI_ISL_406844 AAAAAAAAAAAAA
3. BetaCoV/Sydney/2/2020|EPI_ISL_408976     AAAAAAAAAAAAA
```

b. Schematic of predicted pseudoknots in the SARS-CoV-2 3' UTR affected by culture-derived deletions

435

436 Supplementary Figure 2

437 Structured RNAs in the SARS-CoV-2 3' UTR. A) An alignment of SARS-CoV-2 3' UTR

438 sequences, including the original Wuhan-Hu-1 sourced from Wuhan, China and considered

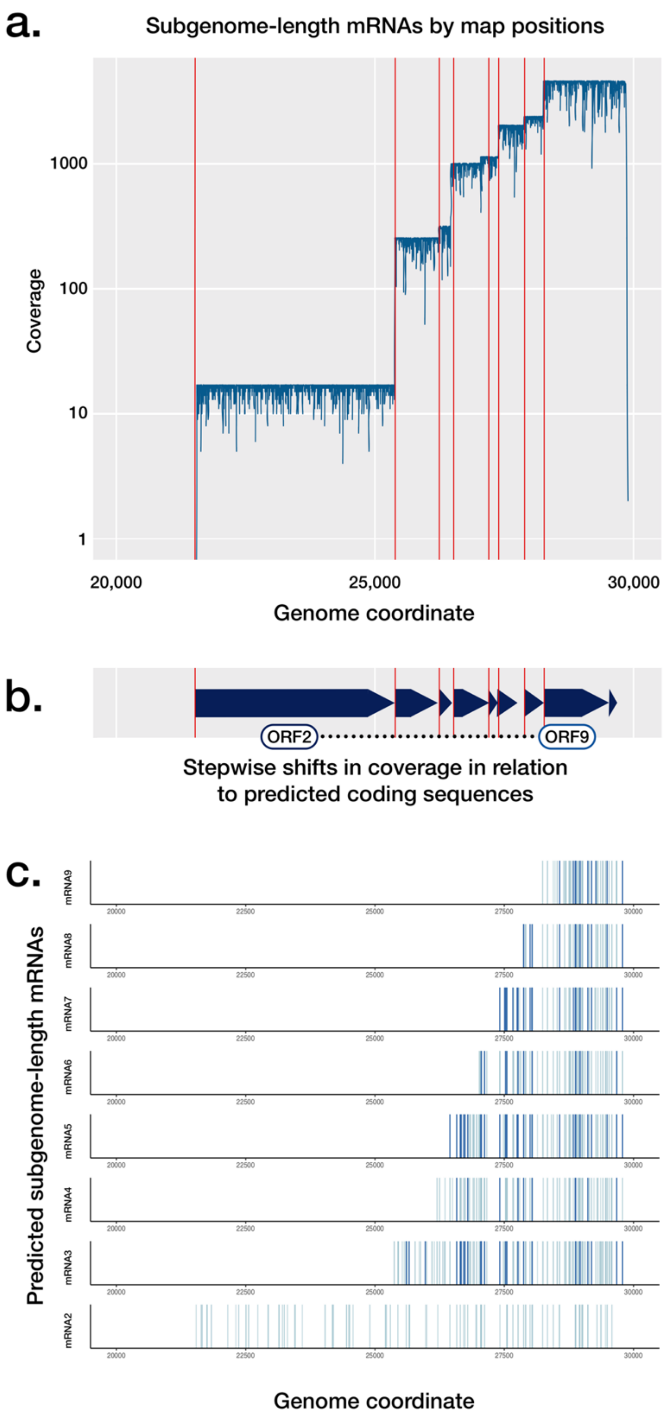
439 the reference genome for the outbreak, and two examples of cultured SARS-CoV-2 isolates

440 exhibiting deletions in a shared 3' UTR region predicted to form a pseudoknot structure. B)

441 Predicted pseudoknot structure of the SARS-CoV-2 3'UTR affected by the above culture-

442 derived deletions.

443



Supplementary Figure 3

Subgenome-length mRNA abundance and predicted sites of modification. A) Coverage of relevant coding sequences achieved by alignment of subgenome-length mRNAs to the SARS-CoV-2 genome (log scale). Red lines indicate the first base of each coding sequence from ORF2-10. B) Schematic of relevant annotated coding sequences. C) Position of predicted m5C positions in subgenome-length mRNAs. Dark blue lines indicate positions predicted to have >90% base modification; light blue lines indicate positions predicted to have between 50% and 90% base modification.