

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

3
4 George Taiaroa^{1,2}, Daniel Rawlinson¹, Leo Featherstone¹, Miranda Pitt¹, Leon Caly²,
5 Julian Druce², Damian Purcell¹, Leigh Harty¹, Thomas Tran², Jason Roberts², Mike Catton²,
6 Deborah Williamson^{1,3}, Lachlan Coin^{1*}, Sebastian Duchene^{1*}

- 7
8 1. Department of Microbiology and Immunology, University of Melbourne at The Peter
9 Doherty Institute for Infection and Immunity, Melbourne, Australia
10 2. Victorian Infectious Diseases Reference Laboratory, Royal Melbourne Hospital, at
11 the Peter Doherty Institute for Infection and Immunity, Victoria, Australia
12 3. Department of Microbiology, Royal Melbourne Hospital, Victoria, Australia
13 • These authors contributed equally to the work

14
15 Corresponding author:

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

21
22 Abstract

23
24 The rapid sharing of sequence information as seen throughout the current SARS-CoV-2
25 epidemic, represents an inflection point for genomic epidemiology. Here we describe
26 aspects of coronavirus evolutionary genetics revealed from these data, and provide the first
27 direct RNA sequence of SARS-CoV-2, detailing coronaviral subgenome-length mRNA
28 architecture.

29
30
31 The ongoing epidemic of 2019 novel coronavirus (now called SARS-CoV-2, causing the
32 disease COVID-19), which originated in Wuhan, China, has been declared a public health
33 emergency of international concern by the World Health Organisation (WHO)[1-4]. SARS-
34 CoV-2 is a positive-sense single-stranded RNA ((+)ssRNA) virus of the Coronaviridae family,
35 with related Betacoronaviruses capable of infecting mammalian and avian hosts, resulting in

36 illness in humans such as Middle East respiratory syndrome (MERS) and the original severe
37 acute respiratory syndrome (SARS)[2, 5-6]. Based on limited sampling of potential reservoir
38 species, SARS-CoV-2 has been found to be most similar to bat coronaviruses at the
39 genomic level, potentially indicating that bats are its natural reservoir [7-8].
40

41 Following the emergence of SARS-CoV-2, genomic analyses have played a key role in the
42 public health response by informing the design of appropriate molecular diagnostics and
43 corroborating epidemiological efforts to trace contacts [8-10]. Taken together, publicly
44 available sequence data suggest a recently occurring, point-source outbreak, as described
45 in online sources [10-12]. Aspects of the response make the assumption that the genetics of
46 SARS-CoV-2, including mechanisms of gene expression and molecular evolutionary rates,
47 are comparable with previously characterised coronaviruses [11-12]. It remains highly
48 relevant to validate these assumptions experimentally with SARS-CoV-2-specific data, with
49 the potential to reveal further insights into the biology of this emergent pathogen. To address
50 this, we describe (i) the architecture of the coronaviral subgenome-length mRNAs, and (ii)
51 phylogenetic approaches able to provide robust estimates of coronaviral evolutionary rates
52 and timescales at this early stage of the outbreak.
53

54 Characterised coronaviral species produce a nested set of polyadenylated subgenome-
55 length mRNA transcripts through a mechanism termed discontinuous extension of minus
56 strands that yields mRNA transcripts of different length. The discontinuous transcription
57 mechanism repositions the 5' leader sequence upstream of consecutive viral open-reading
58 frames (ORF) where each translation start site becomes located at the primary position for
59 ribosome scanning (Figure 1a). Subgenome-length mRNAs have a common 5' leader
60 sequence, near-identical to that located in the 5'-UTR of the viral genome, with the genome-
61 length RNAs also having an mRNA function [13-14]. Discontinuous copy-choice RNA
62 recombination modulates the expression of coronaviral genes [14-15]. Standard sequencing
63 technologies for RNA viruses are unable to produce reads representing (i) RNA viral
64 genomes or (ii) subgenome-length mRNAs, as they generate short read lengths and have a
65 reliance on amplification to generate complementary DNA (cDNA) sequences.
66

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

71 GridION platform (Oxford Nanopore Technologies). Through this approach, the electronic
72 current is measured as individual strands of RNA translocate through a nanopore, with
73 derived signal-space data basecalled to infer the corresponding nucleobases
74 (Supplementary Methods). The SARS-CoV-2 sample produced 680,347 reads, comprising
75 860Mb of sequence information in 40 hours of sequencing (BioProject PRJNA608224).
76 Aligning to the genome of the cultured SARS-CoV-2 isolate (MT007544.1), a subset of reads
77 were attributed to coronaviruses sequence (28.9%), comprising 367Mb of sequence
78 distributed across the 29,893 base genome. Of these, a number had lengths >20,000 bases,
79 capturing the majority of the SARS-CoV-2 genome on a single molecule. Together, direct
80 RNA sequencing provided an average 12,230 fold coverage of the coronaviral genome,
81 biased towards sequences proximal to the polyadenylated 3' tail; coverage ranged from 34
82 fold to >160,000 fold (Figure 1b), with the bias reflecting the higher abundance of
83 subgenome-length mRNAs carrying these sequences, as well as the directional sequencing
84 from the polyadenylated 3' tail.

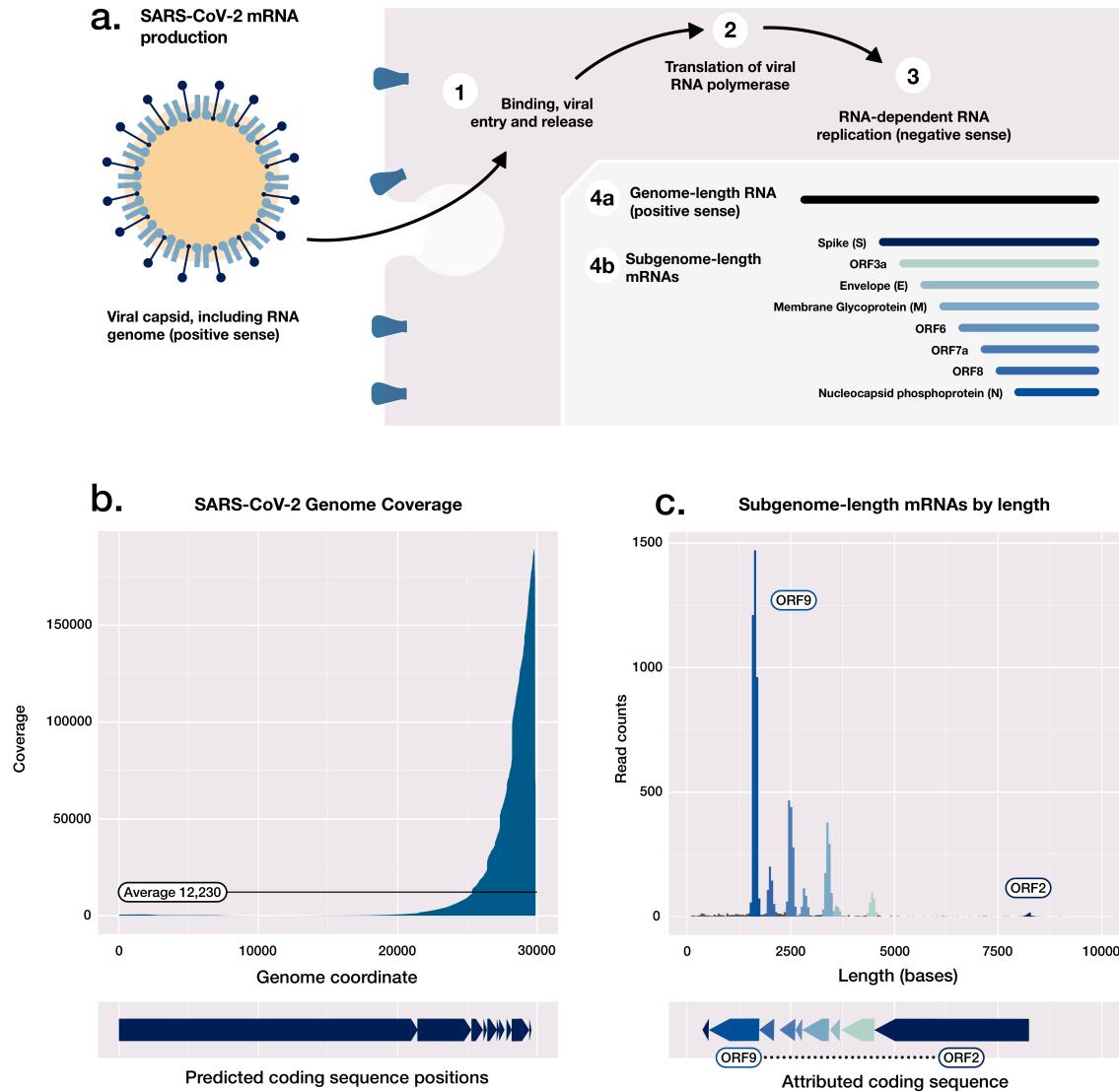
85
86 SARS-CoV-2 features captured in direct RNA sequence data include subgenome-length
87 mRNAs, as well as RNA base modifications. The shared 5'-leader sequence was used as a
88 marker to identify each subgenome-length mRNA (Supplementary Methods). In SARS-CoV-
89 2, we observed eight major viral mRNAs, in addition to the viral genome (Figure 1c). Each
90 annotated gene was observed as a distinct subgenome-length mRNA, positioned
91 consecutively 3' to the replicase polyprotein (ORF1a and 1ab) encoded by the genomic-
92 length mRNA, with ORF10 being the last predicted coding sequence upstream of the poly-A
93 sequence (Supplementary Figure 1). ORF10 is the shortest of the predicted coding
94 sequences at 117 bases in length but despite having proximity to the poly-A sequence, was
95 not found as a subgenome-length mRNA through our direct RNA sequence technique.
96 ORF10 has no annotated function, and the putative encoded peptide does not have a
97 homolog in the SARS-CoV proteome (Proteome ID: UP000000354). These data suggest
98 that the sequence currently annotated as ORF10 does not have a protein coding function in
99 SARS-CoV-2. The sequence annotated as ORF10 is immediately upstream of the 3' UTR
100 and, rather than coding, may act itself or as a precursor of other RNAs in the regulation of
101 gene expression, replication or modulating cellular antiviral pathways. A small number of cell
102 culture-derived isolates in public databases demonstrate shared deletion of an area of the 3'
103 UTR (Supplementary Figure 2), this parallel molecular evolution suggesting the region may
104 have functional roles in vivo.

105

106 In addition to methylation at the 5' cap structure and 3' polyadenylation needed for efficient
107 translation of viral coding sequences, other RNA modifications may have functional roles in
108 SARS-CoV-2 [17]. A range of modifications may be identified using direct RNA sequence
109 data [16-17]; our available SARS-CoV-2 direct RNA sequence data providing adequate
110 coverage to confidently call specific modifications. Through analysis of signal-space data,
111 we identified 42 positions with predicted 5-methylcytosine modifications, appearing at
112 consistent positions between subgenome-length mRNAs (Supplementary Figure 3, and
113 Table 1). In other positive ssRNA viruses, RNA methylation can change dynamically during
114 the course of infection [18], influencing host-pathogen interaction and viral replication. Other
115 modifications may become apparent once training datasets are available for direct RNA
116 sequence data, with little known of the epitranscriptomic landscape of coronaviruses [17,19].
117

118 As well as investigating the above assumed features of SARS-CoV-2 genetics, sequence
119 data also enable an estimate of the molecular evolutionary rate, through analysis of publicly
120 available SARS-CoV-2 genome sequences. Evolutionary rate estimates from other
121 coronaviruses such as Middle East Respiratory Syndrome (MERS) are not necessarily
122 applicable here, particularly because MERS had multiple independent introductions into
123 humans [20-22]. To estimate the evolutionary rate and time of origin of the SARS-CoV-2
124 outbreak, we carried out Bayesian phylogenetic analyses using a curated set of high quality
125 publicly available SARS-CoV-2 genome sequences, each having a known collection date
126 (Supplementary Table 2). The sampling times were sufficient to calibrate a molecular clock
127 and infer the evolutionary rate and timescale of the outbreak; the evolutionary rate of SARS-
128 CoV-2 was estimated to be 1.16×10^{-3} substitutions/site/year (95% HPD 6.32×10^{-4} -
129 1.69×10^{-3}), and the time of origin to be early December 2019 (95% HPD November 2019,
130 December 2019), which is in agreement with epidemiological evidence and other recent
131 analyses (Figure 2a) [1-4, 23-24]. Our estimate of the evolutionary rate of SARS-CoV-2 is in
132 line with those of other coronaviruses (Figure 2b), and the low genomic diversity and recent
133 timescale of the outbreak support a recently occurring, point-source transfer to humans.
134

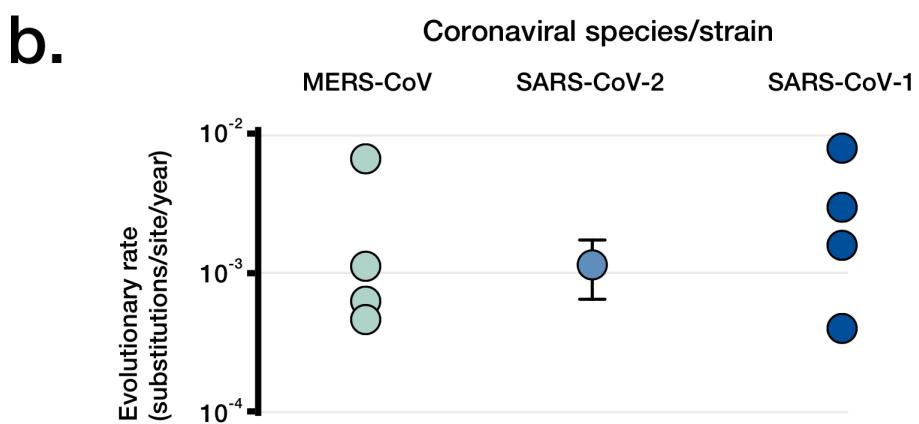
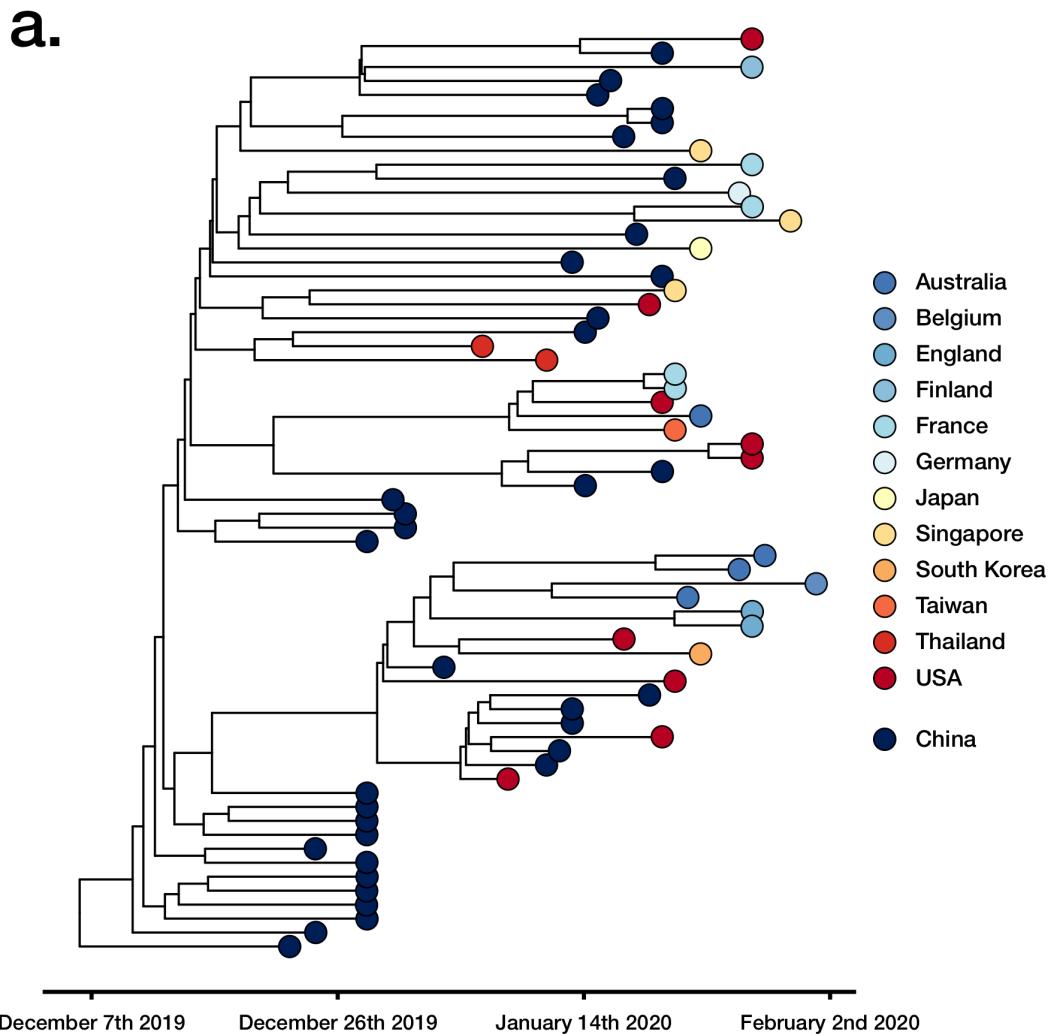
135 In summary, insights into the molecular biology of SARS-CoV-2 are revealed through the
136 use of direct RNA sequence data, enabling a detailed view of viral subgenome-length mRNA
137 architecture. Other insights are gleaned from publicly available sequence data, including an
138 estimate of the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



141

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

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



153 References

154

155 [1] World Health Organization. Pneumonia of unknown cause — China. 2020
156 (<https://www.who.int/csr/don/05-january-2020-pneumonia-of-unkown-cause-china/en/>).
157

158

159 [2] World Health Organization. Novel coronavirus — China. 2020
160 (<https://www.who.int/csr/don/12-january-2020-novel-coronavirus-china/en/>).
161

162

163 [3] Dong, Ensheng, Hongru Du, and Lauren Gardner. "An interactive web-based
164 dashboard to track COVID-19 in real time." *The Lancet Infectious Diseases* (2020).
DOI: 10.1016/S1473-3099(20)30120-1
165

166

167 [4] World Health Organization. Statement on the second meeting of the International
168 Health Regulations (2005) Emergency Committee regarding the outbreak of novel
169 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-\(2019-ncov\)](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))).
170

171

172 [5] Peiris JS, Yuen KY, Osterhaus AD, Stöhr K. The severe acute respiratory syndrome.
173 New England Journal of Medicine. 2003 Dec 18;349(25):2431-41. DOI:
174 10.1056/NEJMra032498
175

176

177 [6] Assiri A, McGeer A, Perl TM, Price CS, Al Rabeeah AA, Cummings DA, Alabdullatif
178 ZN, Assad M, Almulhim A, Makhdoom H, Madani H. Hospital outbreak of Middle East
179 respiratory syndrome coronavirus. *New England Journal of Medicine*. 2013 Aug
180 1;369(5):407-16. DOI: 10.1056/NEJMoa1306742
181

182

183 [7] Perlman, S. Another decade, another coronavirus. *New England Journal of Medicine*.
2020 February 20; 382:760-762. DOI: 10.1056/NEJMe2001126
184

185

186 [8] Lu, Roujian, et al. "Genomic characterisation and epidemiology of 2019 novel
187 coronavirus: implications for virus origins and receptor binding." *The Lancet* (2020).
DOI: 10.1016/S0140-6736(20)30251-8

- 188
- 189 [9] Shu Y, McCauley J. GISAID: Global initiative on sharing all influenza data—from
190 vision to reality. *Eurosurveillance*. 2017 Mar 30;22(13). <https://www.gisaid.org/>
- 191
- 192 [10] Hadfield J, Megill C, Bell SM, Huddleston J, Potter B, Callender C, Sagulenko P,
193 Bedford T, Neher RA. Nextstrain: real-time tracking of pathogen evolution.
194 *Bioinformatics*. 2018 Dec 1;34(23):4121-3. <https://nextstrain.org/ncov>
- 195
- 196 [11] Andersen KG, Rambaut A, Lipkin WI, Holmes EC, Garry RF. The Proximal Origin of
197 SARS-CoV-2. *Virological*, accessed on 27/02/2020. [http://virological.org/t/the-
198 proximal-origin-of-sars-cov-2/398](http://virological.org/t/the-proximal-origin-of-sars-cov-2/398)
- 199
- 200 [12] Rambaut A. Phylodynamic Analysis | 129 genomes | 24 Feb 2020. *Virological*,
201 accessed 27/02/2020. [http://virological.org/t/phylodynamic-analysis-129-genomes-
202 24-feb-2020/356](http://virological.org/t/phylodynamic-analysis-129-genomes-24-feb-2020/356)
- 203
- 204 [13] Yount B, Curtis KM, Fritz EA, Hensley LE, Jahrling PB, Prentice E, Denison MR,
205 Geisbert TW, Baric RS. Reverse genetics with a full-length infectious cDNA of severe
206 acute respiratory syndrome coronavirus. *Proceedings of the National Academy of
207 Sciences*. 2003 Oct 28;100(22):12995-3000. DOI: 10.1073/pnas.1735582100
- 208
- 209 [14] Brian DA, Baric RS. Coronavirus genome structure and replication. In *Coronavirus
210 replication and reverse genetics 2005* (pp. 1-30). Springer, Berlin, Heidelberg.
- 211
- 212 [15] Chen Y, Cai H, Xiang N, Tien P, Ahola T, Guo D. Functional screen reveals SARS
213 coronavirus nonstructural protein nsp14 as a novel cap N7 methyltransferase.
214 *Proceedings of the National Academy of Sciences*. 2009 Mar 3;106(9):3484-9. DOI:
215 10.1073/pnas.0808790106
- 216
- 217 [16] Garalde DR, Snell EA, Jachimowicz D, Sipos B, Lloyd JH, Bruce M, Pantic N,
218 Admassu T, James P, Warland A, Jordan M. Highly parallel direct RNA sequencing
219 on an array of nanopores. *Nature methods*. 2018 Mar;15(3):201. DOI:
220 10.1038/nmeth.4577
- 221

- 222 [17] Viehweger A, Krautwurst S, Lamkiewicz K, Madhugiri R, Ziebuhr J, Hölzer M, Marz
223 M. Direct RNA nanopore sequencing of full-length coronavirus genomes provides
224 novel insights into structural variants and enables modification analysis. *Genome*
225 research. 2019 Sep 1;29(9):1545-54. DOI: 10.1101/gr.247064.118
226
- 227 [18] Lichinchi G, Zhao BS, Wu Y, Lu Z, Qin Y, He C, Rana TM. Dynamics of human and
228 viral RNA methylation during Zika virus infection. *Cell host & microbe*. 2016 Nov
229 9;20(5):666-73. DOI: 10.1016/j.chom.2016.10.002
230
- 231 [19] Denison MR, Graham RL, Donaldson EF, Eckerle LD, Baric RS. Coronaviruses: an
232 RNA proofreading machine regulates replication fidelity and diversity. *RNA biology*.
233 2011 Mar 1;8(2):270-9. DOI: 10.4161/rna.8.2.15013
234
- 235 [20] Chinese SARS Molecular Epidemiology Consortium. Molecular evolution of the
236 SARS coronavirus during the course of the SARS epidemic in China. *Science*. 2004
237 Mar 12;303(5664):1666-9. DOI: 10.1126/science.1092002
238
- 239 [21] Dudas G, Carvalho LM, Rambaut A, Bedford T. MERS-CoV spillover at the camel-
240 human interface. *eLife*. 2018 Jan 16;7:e31257. DOI: 10.7554/eLife.31257
241
- 242 [22] Cotten M, Watson SJ, Kellam P, Al-Rabeeah AA, Makhdoom HQ, Assiri A, Al-Tawfiq
243 JA, Alhakeem RF, Madani H, AlRabiah FA, Al Hajjar S. Transmission and evolution
244 of the Middle East respiratory syndrome coronavirus in Saudi Arabia: a descriptive
245 genomic study. *The Lancet*. 2013 Dec 14;382(9909):1993-2002. DOI:
246 10.1016/S0140-6736(13)61887-5
247
- 248 [23] Andersen K. Clock and TMRCA based on 27 genomes. *Virological*, accessed on
249 27/02/2020. <http://virological.org/t/clock-and-tmrca-based-on-27-genomes/347>
250
- 251 [24] Bedford, T. Phylodynamic estimation of incidence and prevalence of novel
252 coronavirus (nCoV) infections through time. *Virological*, accessed on 27/02/2020.
253 <http://virological.org/t/phylodynamic-estimation-of-incidence-and-prevalence-of-novel-coronavirus-ncov-infections-through-time/391>
254
255
256

257 Methods
258
259 SARS-CoV-2 virus sample
260 The SARS-CoV-2 material prepared for this work is of the first Australian case of Covid-2019
261 (Australia/VIC01/2020), maintained in cell culture. In brief, African green monkey kidney
262 cells expressing the human signalling lymphocytic activation molecule (SLAM; termed
263 Vero/hSLAM cells accordingly) and SARS-CoV-2 were grown at 37°C at 5% CO₂ in media
264 consisting of 10 mL Earle's minimum essential medium, 7% FBS (Bovogen Biologicals,
265 Keilor East, Aus), 2 mM L-Glutamine, 1 mM Sodium pyruvate, 1500 mg/L sodium
266 bicarbonate, 15 mM HEPES and 0.4 mg/ml geneticin in 25cm² flasks. The genome of the
267 cultured isolate (MT007544.1) has three single nucleotide variants (T19065C, T22303G,
268 G26144T) relative to the SARS-CoV-2 Wuhan-Hu-1 reference genome (MN908947.3), and
269 a 10 base deletion in the 3' UTR. The T22303G and 3' UTR variants have been confirmed
270 as culture-derived through Sanger sequencing of clinical and culture material.
271
272 Nucleic acids were prepared from infected cellular material, following inactivation with linear
273 acrylamide and ethanol. RNA was extracted from a modest cell pellet (~200mg) using
274 manually prepared wide-bore pipette tips and minimal steps to maintain RNA length for long
275 read sequencing, and a QIAamp Viral RNA Mini Kit (Qiagen, Hilden, Germany). Carrier
276 RNA was not added to Buffer AVL, with 1% linear acrylamide (Life Technologies, Carlsbad,
277 CA, USA) added instead. Wash buffer AW1 was omitted from the purification stage, with
278 RNA eluted in 50 µl of nuclease free water, followed by DNase treatment with Turbo DNase
279 (Thermo Fisher Scientific, Waltham, MA, USA) 37°C for 30 min. RNA was cleaned and
280 concentrated to 10 µl using the RNA Clean & Concentrator-5 kit (Zymo Research, Irvine,
281 CA, USA), as per manufacturer's instructions.
282
283 Nanopore sequencing of direct RNA
284 Prepared RNA (~1µg) was carried into a direct RNA sequence library preparation with the
285 Oxford Nanopore DRS protocol (SQK- RNA002, Oxford Nanopore Technologies) following
286 the manufacturer's specifications, although without addition of the Control RNA. The library
287 was loaded on an R9.4 flow cell and sequenced on a GridION device (Oxford Nanopore
288 Technologies), with the sequencing run ending after 40 hours. Signal space data was used
289 to generate nucleobase sequences ('basecalled') using ont-guppy-for-gridion 3.0.6. Both
290 signal space and basecalled read data are available at BioProject PRJNA608224. It should
291 be noted that non-polyadenylated RNAs are not expected to be detected with this approach.

292

293 Characterisation of SARS-CoV-2 subgenome-length mRNA architecture
294 Direct RNA reads passing the above thresholds were aligned to the genome of the cultured
295 Australian SARS-CoV-2 isolate (MT007544.1), with parallel and concordant analyses in
296 Geneious Prime (2019.2.1, [M1]) and minimap2 v 2.11 using the “spliced” preset [M2].
297 Coverage statistics were determined from the resulting read alignments.

298

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

310

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

322

323 SARS-CoV-2 Phylogenetics

324 In order to estimate the evolutionary rate and time of origin of SARS-CoV-2, we carried out
325 phylogenetic analyses in BEAST v1.101 [M10], with a curated set of 66 complete and high
326 quality SARS-CoV-2 genomes with date of collection data, as available on February 10th

327 from GISAID and GenBank (Supplementary Table 2). Temporal signal was assessed using
328 BETS [M11]. Initially we determined whether the evolutionary signal and time over which the
329 genome data were collected was sufficient to calibrate the molecular clock, allowing for the
330 evolutionary rate and timescale of the outbreak to be inferred. The model selection approach
331 from BETS supported a strict molecular clock model with genome sampling times for
332 calibration and a coalescent exponential tree prior, which posits that the number of infected
333 individuals grows exponentially over time. We used the HKY+ Γ substitution model, and set
334 the following priors for key parameters:

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

338 A Markov chain Monte Carlo of length 10^7 was set, sampling every 10^3 steps, and assessed
339 sufficient sampling by verifying that the effective sample size for all parameters was at least
340 200 as determined in Tracer [M12], automatically discarding 10% of the burn in. We
341 summarised the posterior distribution of phylogenetic trees by selecting the highest clade
342 credibility tree alongside calculating posterior node probabilities and the distribution of node
343 ages. Comparison to other coronaviral evolutionary rates included studies [M13-20].

344

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

349

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

352

353 [M3] Milne I, Bayer M, Stephen G, Cardle L, Marshall D. Tablet: visualizing next-
354 generation sequence assemblies and mappings. InPlant Bioinformatics 2016 (pp.
355 253-268). Humana Press, New York, NY.

356

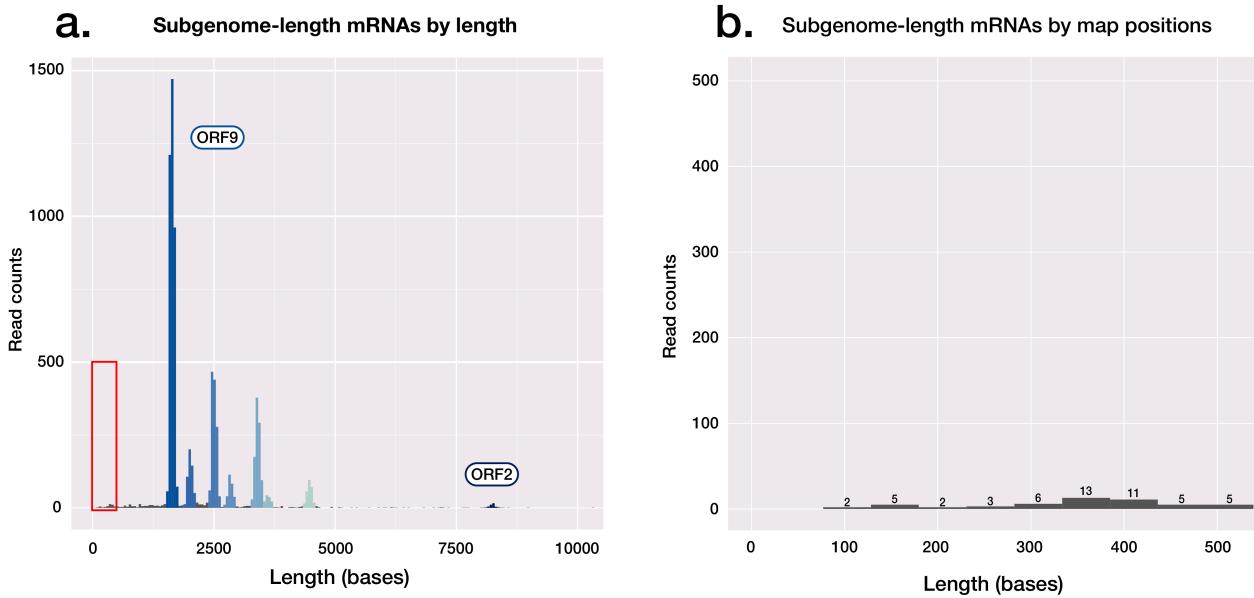
357 [M4] Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G,
358 Durbin R. The sequence alignment/map format and SAMtools. Bioinformatics. 2009
359 Aug 15;25(16):2078-9.

360

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

- 362
- 363 [M6] Team RC. R: A language and environment for statistical computing.
- 364
- 365 [M7] Sato, Kengo, et al. "IPknot: fast and accurate prediction of RNA secondary structures
366 with pseudoknots using integer programming." *Bioinformatics* 27.13 (2011): i85-i93.
- 367
- 368 [M8] Darty, Kévin, Alain Denise, and Yann Ponty. "VARNA: Interactive drawing and
369 editing of the RNA secondary structure." *Bioinformatics* 25.15 (2009): 1974.
- 370
- 371 [M9] Ferguson JM, Smith MA. SquiggleKit: A toolkit for manipulating nanopore signal
372 data. *Bioinformatics*. 2019 Dec 15;35(24):5372-3.
- 373
- 374 [M10] Drummond AJ, Rambaut A. BEAST: Bayesian evolutionary analysis by sampling
375 trees. *BMC evolutionary biology*. 2007 Dec;7(1):214.
- 376
- 377 [M11] Duchene S, Stadler T, Ho SY, Duchene DA, Dhanasekaran V, Baele G. Bayesian
378 Evaluation of Temporal Signal in Measurably Evolving Populations. *bioRxiv*. 2019
379 Jan 1:810697.
- 380
- 381 [M12] Rambaut A, Drummond AJ, Xie D, Baele G, Suchard MA. Posterior summarization in
382 Bayesian phylogenetics using Tracer 1.7. *Systematic biology*. 2018 Sep;67(5):901.
- 383
- 384 [M13] Cotten M, Watson SJ, Kellam P, Al-Rabeeah AA, Makhdoom HQ, Assiri A, Al-Tawfiq
385 JA, Alhakeem RF, Madani H, AlRabiah FA, Al Hajjar S. Transmission and evolution
386 of the Middle East respiratory syndrome coronavirus in Saudi Arabia: a descriptive
387 genomic study. *The Lancet*. 2013 Dec 14;382(9909):1993-2002.
- 388
- 389 [M14] Cotten M, Watson SJ, Zumla AI, Makhdoom HQ, Palser AL, Ong SH, Al Rabeeah
390 AA, Alhakeem RF, Assiri A, Al-Tawfiq JA, Albarak A. Spread, circulation, and
391 evolution of the Middle East respiratory syndrome coronavirus. *MBio*. 2014 Feb
392 28;5(1):e01062-13.
- 393
- 394 [M15] Dudas G, Carvalho LM, Rambaut A, Bedford T. MERS-CoV spillover at the camel-
395 human interface. *Elife*. 2018 Jan 16;7:e31257.
- 396

- 397 [M16] Zhao Z, Li H, Wu X, Zhong Y, Zhang K, Zhang YP, Boerwinkle E, Fu YX. Moderate
398 mutation rate in the SARS coronavirus genome and its implications. BMC
399 evolutionary biology. 2004 Dec 1;4(1):21.
- 400
- 401 [M17] Salemi M, Fitch WM, Ciccozzi M, Ruiz-Alvarez MJ, Rezza G, Lewis MJ. Severe
402 acute respiratory syndrome coronavirus sequence characteristics and evolutionary
403 rate estimate from maximum likelihood analysis. Journal of virology. 2004 Feb
404 1;78(3):1602-3.
- 405
- 406 [M18] Wu SF, Du CJ, Wan P, Chen TG, Li JQ, Li D, Zeng YJ, Zhu YP, He FC. The genome
407 comparison of SARS-CoV and other coronaviruses. Yi chuan= Hereditas. 2003
408 Jul;25(4):373-82.
- 409
- 410 [M19] Chinese SARS Molecular Epidemiology Consortium. Molecular evolution of the
411 SARS coronavirus during the course of the SARS epidemic in China. Science. 2004
412 Mar 12;303(5664):1666-9.
- 413
- 414 [M20] Sanjuán R. From molecular genetics to phylodynamics: evolutionary relevance of
415 mutation rates across viruses. PLoS pathogens. 2012 May;8(5).
- 416
- 417
- 418 Acknowledgements
- 419 The authors gratefully acknowledge the traditional peoples of the land on which the work
420 was carried out, the Wurundjeri Woi-wurrung people of the Kulin nation.
- 421
- 422 The authors also acknowledge the broader staff of the Victorian Infectious Disease
423 Reference Laboratory (VIDRL), their public health partners, and VIDRL's major funder, the
424 Victorian Department of Health and Human Services. Lastly, the authors gratefully
425 acknowledge the public health and academic groups contributing to the preparation, release
426 and analysis of SARS-CoV-2 sequence data, including Nextstrain, GISAID and Virological.



427

428 **Supplementary Figure 1**

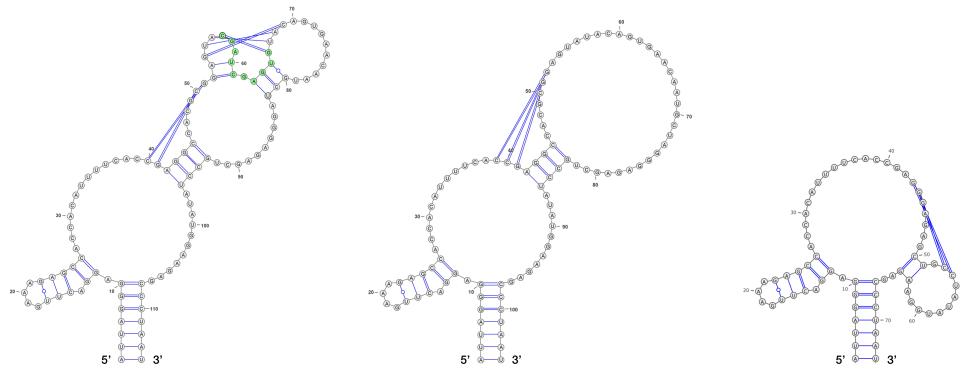
429 Absence of observed coding potential for ORF10 in SARS-CoV-2. A) Read length
430 histogram, showing subgenome-length mRNAs attributed to coding sequences, with the
431 area highlighted shown in detail in a second panel. B) Read length histogram, showing read
432 counts of lengths corresponding to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

434

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



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

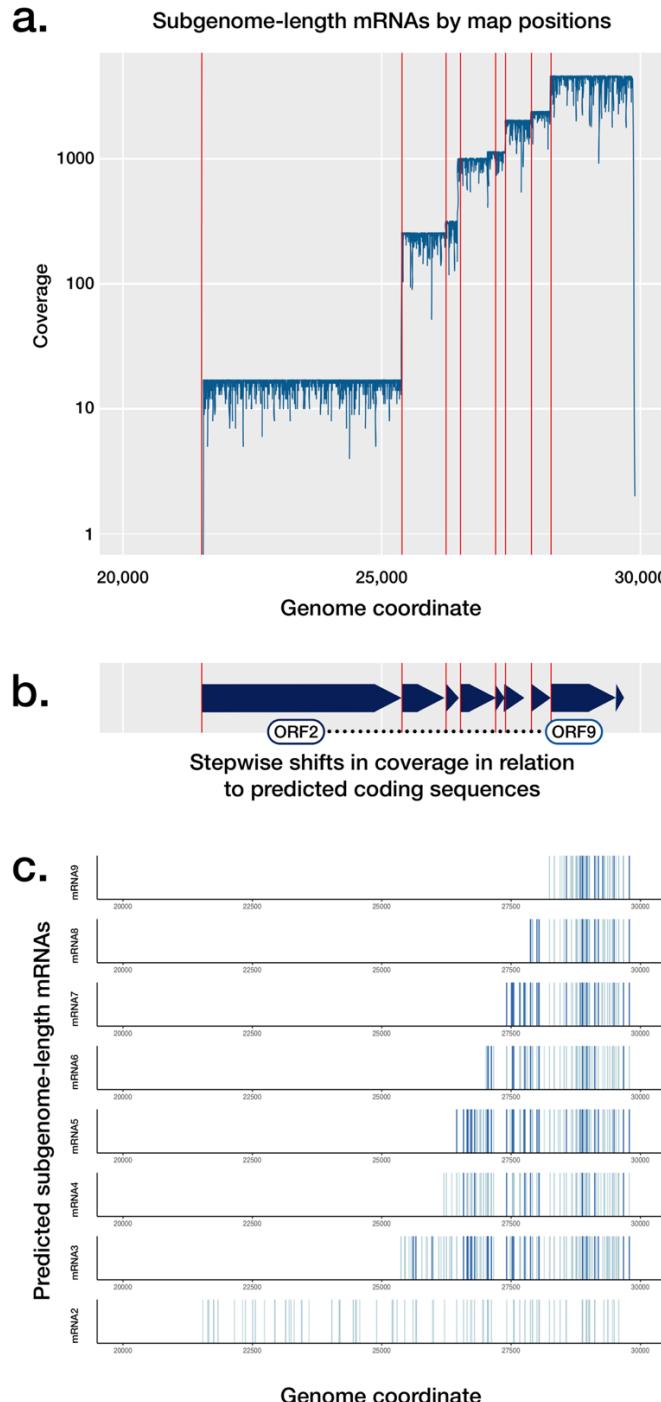


435

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



444

445 Supplementary Figure 3

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