

The coding capacity of SARS-CoV-2

Yaara Finkel^{1,7}, Orel Mizrahi^{1,7}, Aharon Nachshon¹, Shira Weingarten-Gabbay^{2,3}, David Morgenstern⁴, Yfat Yahalom-Ronen⁵, Hadas Tamir⁵, Hagit Achdout⁵, Dana Stein⁶, Ofir Israeli⁶, Adi Beth-Din⁶, Sharon Melamed⁵, Shay Weiss⁵, Tomer Israeli⁵, Nir Paran⁵, Michal Schwartz¹ and Noam Stern-Ginossar^{1*}

¹ Department of Molecular Genetics, Weizmann Institute of Science, Rehovot 76100, Israel.

² Broad Institute of MIT and Harvard, Cambridge, MA 02142, USA.

³ Department of Organismal and Evolutionary Biology, Harvard University, Cambridge, MA 02138, USA

⁴ de Botton Institute for Protein Profiling, The Nancy and Stephen Grand Israel National Center for Personalised Medicine, Weizmann Institute of Science, Rehovot 76100, Israel.

⁵ Department of Infectious Diseases, Israel Institute for Biological Research, Ness Ziona 74100, Israel.

⁶Department of Biochemistry and Molecular Genetics, Israel Institute for Biological Research, Ness Ziona 74100, Israel.

⁷ These authors contributed equally to this work

* To whom correspondence should be addressed: noam.stern-ginossar@weizmann.ac.il

22 **Abstract**

23 Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) is the cause of the ongoing
24 Coronavirus disease 19 (COVID-19) pandemic ^{1,2}. In order to understand SARS-CoV-2
25 pathogenicity and antigenic potential, and to develop diagnostic and therapeutic tools, it is
26 essential to portray the full repertoire of its expressed proteins. The SARS-CoV-2 coding
27 capacity map is currently based on computational predictions and relies on homology to other
28 coronaviruses. Since coronaviruses differ in their protein array, especially in the variety of
29 accessory proteins, it is crucial to characterize the specific collection of SARS-CoV-2 proteins in
30 an unbiased and open-ended manner. Utilizing a suite of ribosome profiling techniques ³⁻⁸, we
31 present a high-resolution map of the SARS-CoV-2 coding regions, allowing us to accurately
32 quantify the expression of canonical viral open reading frames (ORF)s and to identify 23 novel
33 unannotated viral translated ORFs. These ORFs include upstream ORFs (uORFs) that are likely
34 playing a regulatory role, several in-frame internal ORFs lying within existing ORFs, resulting in
35 N-terminally truncated products, as well as internal out-of-frame ORFs, which generate novel
36 polypeptides. We further show that viral mRNAs are not translated more efficiently than host
37 mRNAs; rather, virus translation dominates host translation due to high levels of viral
38 transcripts. Overall, our work reveals the full coding capacity of SARS-CoV-2 genome,
39 providing a rich resource, which will form the basis of future functional studies and diagnostic
40 efforts.

41

42 **Main:**

43 SARS-CoV-2 is an enveloped virus consisting of a positive-sense, single-stranded RNA genome
44 of ~30 kb and shows characteristic features of other coronaviruses. Upon cell entry, two
45 overlapping ORFs are translated from the positive strand genomic RNA, ORF1a and ORF1b.
46 The translation of ORF1b is mediated by a -1 frameshift that allows translation to continue into
47 ORF1b enabling the generation of continuous polypeptides which are cleaved into a total of 16
48 nonstructural proteins (NSPs)⁹⁻¹¹. In addition, the viral RNA-dependent RNA polymerase
49 (RdRP) uses the viral genome to produce negative-strand RNA intermediates which serve as
50 templates for the synthesis of positive-strand genomic RNA and of subgenomic RNAs⁹⁻¹¹. The
51 subgenomic RNAs contain a common 5' leader fused to different segments from the 3' end of the
52 viral genome, and contain a 5'-cap structure and a 3' poly(A) tail^{12,13}. These unique fusions
53 occur during negative-strand synthesis at 6-7 nt core sequences called transcription-regulating
54 sequences (TRS)s that are located at the 3' end of the leader sequence as well as preceding each
55 viral gene. The different subgenomic RNAs encode 4 conserved structural proteins- spike protein
56 (S), envelope protein (E), membrane protein (M), nucleocapsid protein (N)- and several
57 accessory proteins. Based on sequence similarity to other beta coronaviruses and specifically to
58 SARS-CoV, current annotation of SARS-CoV-2 includes predictions of six accessory proteins
59 (3a, 6, 7a, 7b, 8, and 10, NCBI Reference Sequence: NC_045512.2), but not all of these ORFs
60 have been experimentally and reproducibly confirmed in this virus^{14,15}.

61 To capture the full SARS-CoV-2 coding capacity, we initially applied a suite of ribosome
62 profiling approaches to Vero E6 cells infected at MOI=0.2 with SARS-CoV-2
63 (BetaCoV/Germany/BavPat1/2020 EPI_ISL_406862) for 5 or 24 hours (Figure 1A). At 24 hours
64 post infection the vast majority of cells were infected but cells were still intact (Figure S1). For
65 each time point we mapped genome-wide translation events by preparing three different
66 ribosome-profiling libraries (Ribo-seq), each one in two biological replicates. Two Ribo-seq
67 libraries facilitate mapping of translation initiation sites, by treating cells with lactimidomycin
68 (LTM) or harringtonine (Harr), two drugs with distinct mechanisms that inhibit translation
69 initiation by preventing 80S ribosomes formed at translation initiation sites from elongating.
70 These treatments lead to strong accumulation of ribosomes precisely at the sites of translation
71 initiation and depletion of ribosomes over the body of the message (Figure 1A and^{4,6}). The third
72 Ribo-seq library was prepared from cells treated with the translation elongation inhibitor

73 cycloheximide (CHX), and gives a snap-shot of actively translating ribosomes across the body of
74 the translated ORF (Figure 1A). These three complementary approaches provide a powerful tool
75 for the unbiased mapping of translated ORFs. In parallel, RNA-sequencing (RNA-seq) was
76 applied to map viral transcripts. When analyzing the different Ribo-seq libraries across coding
77 regions in cellular genes, the expected distinct profiles are observed in both replicates,
78 confirming the overall quality of the libraries. Ribosome footprints displayed a strong peak at the
79 translation initiation site, which, as expected, is more pronounced in the Harr and LTM libraries,
80 while the CHX library also exhibited a distribution of ribosomes across the entire coding region
81 up to the stop codon, and its mapped footprints were enriched in fragments that align to the
82 translated frame (Figure 1B, Figure S2 and Figure S3). As expected, the RNA-seq reads were
83 uniformly distributed across coding and non-coding regions (Figure 1B and Figure S2). The
84 footprint profiles of viral coding sequences at 5 hours post infection (hpi) fit the expected profile
85 of translation, similar to the profile of cellular genes, both at the meta gene level and at the level
86 of individual genes (Figure 1C, Figure S4 and Figure S5). In addition, the footprint densities at
87 5hpi were highly reproducible between our biological replicates both at the gene level and in
88 single codon resolution on the viral genome (Figure S6). Intriguingly, the footprint profile over
89 the viral genome at 24 hpi, did not fit the expected profile of translating ribosomes and were
90 generally not affected by Harr or LTM treatments (Figure S4). To further examine the quality of
91 footprint measurements we applied a fragment length organization similarity score (FLOSS) that
92 measures the magnitude of disagreement between the footprint distribution on a given transcript
93 and the footprint distribution on canonical CDSs¹⁶. At 5 hpi protected fragments from SARS-
94 CoV-2 transcripts showed the expected size distribution (Figure S7A) and scored well in these
95 matrices and did not differ from well-expressed human transcripts (Figure 1D). However, reads
96 from 24 hpi could be clearly distinguished from cellular annotated coding sequences (Figure 1E
97 and Figure S7B). We conclude that the footprint data from 5hpi constitutes robust and
98 reproducible ribosome footprint information but that the majority of viral protected fragments at
99 24 hpi are likely not generated by ribosome protection and may reflect additional interactions
100 that occur on the viral genome at late time points in infection.

101 A global view of RNA and CHX footprint reads mapping to the viral genome at 5hpi,
102 demonstrate an overall 5' to 3' increase in coverage (Figure 2A). RNA levels are essentially
103 constant across ORFs 1a and 1b, and then steadily increases towards the 3', reflecting the

104 cumulative abundance of these sequences due to the nested transcription of subgenomic RNAs
105 (Figure 2A). Increased coverage is also seen at the 5' UTR reflecting the presence of the 5' leader
106 sequence in all subgenomic RNAs as well as the genomic RNA. Reduction in footprint density
107 between ORF1a and ORF1b reflects the proportion of ribosomes that terminate at the ORF1a
108 stop codon instead of frameshifting into ORF1b (Figure S8). By dividing the footprint density in
109 ORF1b by the density in ORF1a we estimate frameshift efficiency is 57% +/- 12%. This value is
110 comparable to the frameshift efficiency measured based on ribosome profiling of mouse hepatitis
111 virus (MHV, 48%-75%)⁷. On the molecular level this 57% frameshifting rate indicates NSP1-
112 NSP11 are expressed 1.8 +/- 0.4 times more than NSP12-NSP16 and this ratio likely relates to
113 the stoichiometry needed to generate SARS-CoV-2 nonstructural macromolecular complexes ¹⁷.
114 Similarly to what was seen in MHV and avian infectious bronchitis virus (IBV) ^{7,8}, we failed to
115 see noticeable ribosome pausing before or at the frameshift site, but we identified several
116 potential pausing sites within ORF1a and in ORF1b that were reproducible between replicates
117 (Figure S8), however these will require further characterization.

118 Besides ORF1a and ORF1b, all other canonical viral ORFs are translated from subgenomic
119 RNAs. We therefore examined whether the levels of viral gene translation correlate with the
120 levels of the corresponding subgenomic RNAs. Since raw RNA-seq densities represent the
121 cumulative sum of genomic and all subgenomic RNAs, we calculated transcript abundance using
122 two approaches: deconvolution of RNA densities, in which RNA expression of each ORF is
123 calculated by subtracting the RNA read density of cumulative densities upstream to the ORF
124 region; and relative abundances of RNA reads spanning leader-body junctions of each of the
125 canonical subgenomic RNAs. ORF6, ORF7b and ORF10 obtained negative values in the RNA
126 deconvolution, probably due to their short length and relative weaker expression, which make
127 them more sensitive to inaccuracies related to library preparation biases. For ORF10 we also did
128 not detect reads spanning leader-body junctions. For all other canonical ORFs there was high
129 correlation between these two approaches (Pearson's R = 0.897, Figure S9), and in both
130 approaches the N transcript was the most abundant transcript, in agreement with recent studies
131 ^{15,18}. We next compared footprint densities to RNA abundance as calculated by junction
132 abundances for the subgenomic RNA or deconvolution of genomic RNA in the case of ORF1a
133 and ORF1b (Figure 2B). For the majority of viral ORFs, transcript abundance correlated almost
134 perfectly with footprint densities, indicating these viral ORFs are translated in similar

135 efficiencies (probably due to their almost identical 5'UTRs), however three ORFs were outliers.
136 The translation efficiency of ORF1a and ORF1b was significantly lower. This can stem from
137 unique features in their 5'UTR (discussed below) or from under estimation of their true
138 translation efficiency as some of the full-length RNA molecules may serve as template for
139 replication or packaging and are hence not part of the translated mRNA pool. The third outlier is
140 ORF7b for which we identified very few body-leader junctions but exhibited relatively high
141 translation. A probable explanation is that translation of ORF7b arises from ribosome leaky
142 scanning of the ORF7a transcript, as was suggested in SARS-CoV¹⁹.
143 Recently, many transcripts derived from non-canonical junctions were identified for SARS-CoV-
144 2, some of which were abundant and were suggested to affect the viral coding potential^{15,18}.
145 These non-canonical junctions contain either the leader combined with 3' fragments at
146 unexpected sites in the middle of ORFs (leader-dependent noncanonical junction) or fusion
147 between sequences that do not have similarity to the leader (leader-independent junction). We
148 estimated the frequency of these non-canonical junctions in our RNA libraries. We indeed
149 identified many non-canonical junctions and obtained excellent agreement between our RNA-seq
150 replicates for both canonical and non-canonical junctions, demonstrating these junctions are
151 often generated and mostly do not correspond to random amplification artifacts (Figure S10A,
152 S10B and Table S1). The abundance of junction-spanning reads between our data and the data of
153 Kim et al.¹⁸, that was generated from RNA harvested from Vero cells at 24 hpi, showed
154 significant correlation (Pearson's R = 0.816 for 24hpi, Figure S10C and S10D), illustrating many
155 of these are reproducible between experimental systems. However, 111 out of the 361 most
156 abundant leader independent junctions that were mapped by Kim et al., were not supported by
157 any read in our data, illustrating there are also substantial variations. In addition we identified 5
158 abundant leader independent junctions that were not expressed based on Kim et al.¹⁸ (Table S2).
159 We noticed three of these junctions represent short in-frame deletions in the spike protein (5aa,
160 7aa and 10aa long) that overlap deletions that were recently described by other groups^{15,20,21}, in
161 which the furin-like cleavage site is deleted (Figure S11). The re-occurrence of the same
162 genomic deletion strongly supports the conclusion that this deletion is being selected for during
163 passage in Vero cells. In order to examine if any additional non-canonical junctions are derived
164 from genomic deletions we sequenced the genomic RNA of the virus we used in our infections.
165 Indeed, 50% of the genomic RNA contained a 5-10 aa deletion of the furin-like cleavage site in

166 the spike protein. In addition, we identified an 8aa deletion in ORF-E in 2.3% of the genomic
167 RNA, which was also observed in our RNA measurements (Table S2 and Figure 2C). When we
168 compared the frequency of junctions between 5h and 24h time points, the leader dependent
169 junctions (both canonical and non-canonical) and the genomic deletions correlated well but the
170 leader independent junctions were specifically increased at 24 hpi (Figure 2C). Recent kinetic
171 measurements show viral particles already bud out of infected Vero cells at 8 hpi ²¹. Therefore,
172 this time-dependent increase in non-canonical RNA junctions indicates that the leader
173 independent RNA junctions are likely associated with genomic replication. Overall, this data
174 shows a small part of the leader-independent junctions represent genomic deletions that are
175 likely selected for during cell culture passages and a larger subset of leader-independent
176 junctions probably rises during genome replication and therefore less likely to lead to changes in
177 viral transcripts.

178 Our ribosome profiling approach facilitates unbiased assessment of the full range of SARS-CoV-
179 2 translation products. Examination of SARS-CoV-2 translation as reflected by the diverse
180 ribosome footprint libraries, revealed several unannotated translated ORFs. We detected in-frame
181 internal ORFs lying within existing ORFs, resulting in N-terminally truncated product. These
182 included relatively long truncated versions of canonical ORFs, such as the one found in ORF6
183 (Figure 3A and Figure S12A), or very short truncated ORFs that likely serve an upstream ORF
184 (uORF), like truncated ORF7a that might regulate ORF7b translation (Figure 3B, Figure S12B
185 and Figure S12C). We also detected internal out-of-frame translations, that would yield novel
186 polypeptides, such as ORFs within ORF-3a (41aa and 33 aa, Figure 3C and Figure S12D) and
187 within ORF-S (39aa, Figure 3D and Figure S12E) or short ORFs that likely serve as uORFs
188 (Figure 3E and Figure S12F). Additionally, we observed a 13 amino acid extended ORF-M, in
189 addition to the canonical ORF-M, which is predicted to start at the near cognate codon AUA
190 (Figure 3F and Figure S12G and Figure S12H).

191 The presence of the annotated ORF10 was recently put into question as almost no subgenomic
192 reads were found for its corresponding transcript ^{18,22}. Although we also did not detect
193 subgenomic RNA designated for ORF10 translation (Table S1), the ribosome footprint densities
194 indicate translation signal in ORF10 initiation (Figure 3G and Figure S12I). Interestingly, we
195 detected two putative ORFs, an upstream out of frame ORF that overlaps ORF10 initiation and
196 an in-frame internal initiation that leads to a truncated ORF10 product. Further research is

197 needed to delineate how ORFs in this region are translated and whether they have any functional
198 roles.

199 Finally, we detected four distinct initiation sites at SARS-CoV-2 5'UTR. Three of these encode
200 for uORFs that are located just upstream of ORF1a; the first initiating at an AUG (uORF1) and
201 the other two at a near cognate codons (uORF2 and extended uORF2, Figure 3H and Figure
202 S12J). These uORFs are in line with findings in other coronaviruses ^{7,23}. The fourth site is the
203 most prominent peak in the ribosome profiling densities on the SARS-CoV-2 genome and is
204 located on a CUG codon at position 59, just 10 nucleotides upstream the TRS-leader (Figure 3I
205 and Figure S12K). The reads mapped to this site have a tight length distribution characteristic of
206 ribosome protected fragments (Figure S13A). Due to its location upstream of the TRS-leader,
207 footprints mapping to this site can potentially derive from any of the subgenomic as well as the
208 genomic RNAs. Therefore, to view this initiation in its context, we aligned the footprints to the
209 genomic RNA or to the most abundant subgenomic N transcript. On the genome and on ORF-N
210 transcript this initiation results in translation of uORFs, which on the genome will generate an
211 extension of uORF1 (Figure S13B and Figure S13C). In both transcripts, the occupancy at the
212 CUG is higher than the downstream translation signal, implying this peak might reflect
213 ribosomal pausing. Interestingly, ribosome pauses located just upstream of the TRS-leader were
214 also identified in MHV and IBV genomes ^{7,8}. To assess the distribution of footprints at this
215 initiation on the different viral transcripts, viral transcripts were divided into three groups based
216 on their sequence similarity downstream of the leader-junction site (to allow unique footprint
217 alignment, Figure S13D). Interestingly, significantly more footprints were mapped to the group
218 that includes the genomic RNA and the subgenomic E and M transcripts, than would be expected
219 from their relative RNA abundance (Figure S13E). When only footprints that allow unique
220 mapping to genomic RNA or subgenomic M and E transcripts are used (sizes 31-33bp to
221 discriminate M from genome or E transcript, and sizes 32-33bp to discriminate E from the
222 genome) a strong enrichment of footprints that originate from the genome is observed (Figure
223 S13F). This footprint enrichment to genomic RNA suggests ribosome pausing might be more
224 prominent on the genome or that ribosomes engage with genomic RNA differently than with
225 subgenomic transcripts. The proximity of this pause to the leader-TRS, which seem to be
226 conserved in MHV and IBV ^{7,8}, together with the relative enrichment to the viral genome raises
227 the possibility that a ribosome at this position might affect discontinuous transcription either by

228 sterically blocking the TRS-L site or by affecting RNA secondary structure. In addition,
229 ribosomes initiating at the CUG have the potential to generate uORFs or ORF extensions in the
230 different sub-genomic transcripts (Table S3)

231 To systematically define the SARS-CoV-2 translated ORFs we used PRICE and ORF-RATER,
232 two computational methods that rely on a combination of translation features such as LTM and
233 Harr induced peak at the translation initiation site, heightened footprints density and 3-
234 nt periodicity to predict novel translated ORFs from ribosome profiling measurements^{24,25}. After
235 application of a minimal expression cutoff and manual curation on the predictions, these
236 classifiers identified 25 ORFs, these included 10 out of the 11 canonical translation initiations
237 and 15 novel viral ORFs. In addition, ORF-RATER identified three putative ORFs that originate
238 from the CUG initiation and extend to the sub-genomic transcripts of S, M and ORF-6a (Table
239 S3). The majority (85%) of the classifier identified ORFs were independently identified in each
240 of the biological replicate (Table S4 and Figure S14). Visual inspection of the ribosome profiling
241 data suggested additional 8 putative novel ORFs, some of which are presented above (Figure 3A,
242 3B, 3G and Table S4). Overall, we identified 23 putative ORFs, on top of the 12 canonical viral
243 ORFs that are currently annotated in NCBI Reference Sequence and 3 additional potential ORFs
244 that stem from the CUG initiation upstream of the leader.

245
246 To confirm the robustness and relevance of these annotations we extended these experiments to
247 human cells. We first examined the infection efficiency of several human cell lines that were
248 used to study SARS-CoV-2 infection; epithelial lung cancer cell lines, Calu3 and A549, and
249 epithelial colorectal adenocarcinoma cell line, Caco2. Infection of Calu3 was most efficient and
250 infection in the presence of trypsin increased infection efficiency by at least 2-fold (Figure S15).
251 We next infected Calu3 with a SARS-CoV-2 that was isolated from an independent source
252 (BavPat1/2020 Ref-SKU: 026V-03883) and the integrity of the virus that we used for infection
253 was confirmed by sequencing (confirming the virus sequence was intact and there were no
254 abundant genomic deletions). Cells were harvested at 7hpi using the same set of ribosome
255 profiling techniques, each one in two biological replicates and in parallel RNA was harvested for
256 RNA-seq. The different Ribo-seq libraries showed the expected distinct profiles in both
257 replicates, confirming the overall quality of these libraries (Figure S16). We next examined the
258 translation of the new viral ORFs we have annotated using PRICE and ORF-RATER classifiers

259 as well as manual curation. Of the 23 novel ORFs we identified as being translated in Vero cells
260 all showed clear evidence of translation also in Calu3 infected cells, 16 were annotated by
261 PRICE and ORF-RATER (three of which are ORFs that were added manually based on the Vero
262 cells data) and ORF-RATER identified again the same three ORFs that originate from the CUG
263 initiation upstream the leader (Figure S17, Table S3 and Table S4). LTM- induced ribosome
264 accumulation at the canonical and predicted initiation sites were highly reproducible between
265 biological replicates as well as between Calu3 and Vero cells, illustrating the robustness of the
266 translation initiation predictions (Figure S18A-C). Furthermore, ribosome-protected footprints
267 displayed a 3-nt periodicity that was in phase with the predicted start site, in both Vero and
268 Calu3 cells providing further evidence for the active translation of the predicted ORFs (Figure
269 S19). We conclude 23 novel ORFs are reproducibly translated from SARS-CoV-2 independently
270 of the host cell and the exact viral origin and additional ORFs may be translated from the CUG
271 initiation located upstream of the TRS-leader.

272

273 Ribosome density also allows accurate quantification of viral protein production. We first
274 quantified the relative expression levels of canonical viral ORFs. Since many of the ORFs we
275 identified overlap canonical ORFs, the quantification was based on the non-overlapping regions.
276 We found that ORF-N is expressed at the highest level in both Vero and Calu3 cells followed by
277 the rest of the viral ORFs with some differences in the relative expression between the two cell
278 types (Figure 4A). To quantify the expression of out-of-frame internal ORFs we computed the
279 contribution of the internal ORF to the frame periodicity signal relative to the expected
280 contribution of the main ORF. For in-frame internal ORF quantification, we subtracted the
281 coverage of the main ORF in the non-overlapping region. We also utilized ORF-RATER, which
282 uses a regression strategy to calculate relative expression of overlapping ORFs, resulting in
283 largely similar estimates of viral ORF translation levels (Figure S20A and S20B). These
284 measurements show that many of the novel ORFs we annotated are expressed in levels that are
285 comparable to the canonical ORFs (Figure 4B, Figure S20C and Table S4). Furthermore, the
286 relative expression of viral proteins seems to be mostly independent of the host cell origin
287 (Figure 4C).

288

289 Of the novel ORFs we identified 13 are very short (\leq 20 codons) or located in the 5'UTR of the
290 genomic RNA and therefore likely play a regulatory role. Four ORFs are extensions or
291 truncations of canonical ORFs (M, 6, 7a and 10). We examined the properties of the six out-of-
292 frame internal ORFs (iORF)s that are longer than 20aa; one of these is ORF9b and its truncated
293 version (Figure S21A, 97aa and 90aa). ORF9b appears in UniProt annotations and was detected
294 by Bojkova et al.¹⁴ in proteomic measurements, together with our translation measurements this
295 indicates it is a bona fide SARS-CoV-2 protein. In addition we detected an iORF laying at the 5'
296 of ORF-S and its truncated version (Figure 3D, 39 aa and 31 aa), two iORFs within ORF3a
297 (Figure 3C, 41aa and 33 aa) .Mining of recent proteomic measurements of SARS-CoV-2
298 infected cells^{14,15} did not detect peptides that originate from the out-of-frame ORFs we
299 annotated, likely due to challenges in detecting trypsin-digested products from short coding
300 regions²⁴. Indeed, two canonical SARS-CoV-2 proteins, ORF7b (43aa) and ORF-E (75aa) were
301 also not detected by mass-spectrometry^{14,15,26,27}, and our ribosome profiling data is the first to
302 show these SARS-CoV-2 proteins are indeed expressed.

303

304 Using TMHMM, we found S.iORF1 as well as 3a.iORF1 are predicted to contain a
305 transmembrane domain (Figure S22A and S22B). Additionally, using SignalP we found a
306 predicted signal peptide in 3a.iORF2 (Figure S22C). Analysis of the conservation of these out-
307 of-frame iORFs in SARS-CoV and in related viruses (Sarbecoviruses) revealed 3a.iORF1 is
308 highly conserved in Sarbecoviruses (Table S6). This ORF was also identified by three
309 independent comparative genomic studies that demonstrate this ORF has a significant purifying
310 selection signature, implying it is a functional polypeptide²⁸⁻³⁰. In combination with our
311 expression measurements, these findings indicate this internal ORF is a novel and likely
312 functional transmembrane protein, conserved throughout sarbecoviruses and as was suggested by
313 Jungreis et al.³⁰ should be named ORF3c.

314 The second iORF overlapping ORF3a and the iORF overlapping S are not conserved in most
315 other sarbecoviruses (Table S6 and²⁹). The expression of the second iORF overlapping ORF-3a
316 is low (Figure S20) and probably originate from ribosomes that failed to initiate at ORF3a and
317 ORF3c. An extended version of this ORF was pulled-down³¹ and was shown to elicit a strong
318 antibody response³² but we find mainly translation of the short version (Figure S21A). The
319 internal S-ORF is situated just downstream of the ORF-S AUG, suggesting ribosomes might

320 initiate translation via leaky scanning. This region in the S-protein shows extremely-rapid
321 evolution³⁰ but in the SARS-CoV-2 isolates that have been sequenced its coding capacity is not
322 impaired. Future work will have to delineate if this ORF, which is highly expressed (Figure 4B
323 and Figure S20), represents a functional protein. Importantly, translated ORFs that do not act as
324 functional polypeptides could still be an important part of the immunological repertoire of the
325 virus as MHC class I bound peptides are generated at higher efficiency from rapidly degraded
326 polypeptides³³.

327
328 Finally, although we identified two internal out-of-frame ORFs within ORF3a, we did not detect
329 translation of SARS-CoV ORF3b homologue, which contains a premature stop codon in SARS-
330 CoV-2 (Figure S21A). In addition, we did not find evidence of translation of ORF14, which
331 appears in some SARS-CoV-2 annotations²³ (Figure S21B).

332 Translation of viral proteins relies on the cellular translation machinery, and coronaviruses, like
333 many other viruses, are known to cause host shutoff³⁴. In order to quantitatively evaluate if
334 SARS-CoV-2 skews the translation machinery to preferentially translate viral transcripts, we
335 compared the ratio of footprints to mRNAs for virus and host CDSs at 5 hpi and 24 hpi in Vero
336 cells and at 7hpi in Calu3 cells. Since at 24 hpi our ribosome densities were masked by a
337 contaminant signal which do not originate from ribosome protection, for samples from this time
338 point we used the footprints that were mapped to subgenomic RNA junctions (and therefore
339 reflect bona fide transcripts) to estimate the true ribosome densities. In all samples the virus
340 translation efficiencies fall within the low range of most of the host genes (Figure 4D-4F and
341 Figure S23A-C), indicating that viral transcripts are likely not preferentially translated during
342 SARS-CoV-2 infection. Instead, during infection viral transcripts take over the mRNA pool,
343 probably through massive transcription coupled to host induced RNA degradation^{35,36}.

344 In summary, in this study we delineate the translation landscape of SARS-CoV-2.
345 Comprehensive mapping of the expressed ORFs is a prerequisite for the functional investigation
346 of viral proteins and for deciphering viral-host interactions. An in-depth analysis of the ribosome
347 profiling experiments revealed a highly complex landscape of translation products, including
348 translation of 23 novel viral ORFs and illuminating the relative production of all canonical viral
349 proteins. The new ORFs we have identified may serve as novel accessory proteins or as
350 regulatory units controlling the balanced production of different viral proteins. Studies on the

351 functional significance and antigenic potential of these ORFs will deepen our understanding of
352 SARS-CoV-2 and of coronaviruses in general. Overall, our work reveals the coding capacity of
353 SARS-CoV-2 genome and highlights novel features, providing a rich resource for future
354 functional studies.

355

356 **Figure legends:**

357

358 **Figure 1.** Ribosome profiling of SARS-CoV-2 infected cells.

359 **(A)** Vero E6 and Calu3 cells infected with SARS-CoV-2 were harvested at 5, 24 (Vero E6) and 7
360 (Calu3) hours post infection (hpi) for RNA-seq, and for ribosome profiling using lactimidomycin
361 (LTM) and Harringtonine (Harr) treatments for mapping translation initiation or cycloheximide
362 (CHX) treatment to map overall translation. **(B)** Metagene analysis of read densities at the 5' and
363 the 3' regions of cellular protein coding genes as measured by the different ribosome profiling
364 approaches and RNA-seq at 5 hpi (one of two replicates is presented). The X axis shows the
365 nucleotide position relative to the start or the stop codons. The ribosome densities are shown
366 with different colors indicating the three frames relative to the main ORF (red, frame 0; black,
367 frame +1; grey, frame +2). **(C)** Metagene analysis of the 5' region, as described in B, for viral
368 coding genes at 5 hpi **(D and E)** Fragment length organization similarity score (FLOSS) analysis
369 for cellular coding regions and for SARS-CoV-2 canonical ORFs at 5 hpi **(D)** and 24 hpi **(E)**.

370

371 **Figure 2.** Expression level of canonical viral genes.

372 **(A)** RNA-Seq (green) and Ribo-Seq CHX (red) read densities at 5 hpi on the SARS-CoV-2
373 genome. Read densities are plotted on a log scale to cover the wide range in expression across
374 the genome. The lower panel presents SARS-CoV-2 genome organization with the canonical
375 viral ORFs **(B)** Relative abundance of the different viral transcripts relative to the ribosome
376 densities of each SARS-CoV-2 canonical ORF at 5 hpi. Transcript abundance were estimated by
377 counting the reads that span the junctions of the corresponding RNA or for ORF1a and ORF1b
378 the genomic RNA abundance, normalized to junction count. ORF10 is not presented as no
379 junctions designated for its subgenomic RNA were detected. Spearman's R is presented. **(C)**
380 Scatter plot presenting the abundance of viral reads that span canonical leader dependent
381 junctions (red), non-canonical leader dependent junctions (green), non-canonical leader
382 independent junctions (purple) and genomic deletions (cyan) at 5 and 24 hpi. Pearson's R on log
383 transformed values is presented.

384

385 **Figure 3.** Ribosome densities reveal novel viral coding regions.

386 (A-I) Ribosome density profiles of CHX, Harr and LTM samples at 5 hpi. Densities are shown
387 with different colors indicating the three frames relative to the main ORF in each figure (red,
388 frame 0; black, frame +1; grey, frame +2). One out of two replicates is presented. Filled and
389 open rectangles indicate the canonical and novel ORFs, respectively. ORFs starting in a near
390 cognate codon are labeled with stripes. ORFs that stretch beyond the range of the plot are shown
391 as fading out rectangles. (A) In frame internal ORF within ORF6 generating a truncated product,
392 (B) In frame internal initiation within ORF7a (reads marking ORF-7b initiation were cut to fit
393 the scale indicated with black lines), (C) Out of frame internal initiations within ORF-3a, (D)
394 Out of frame internal initiations within ORF-S, (E) Out of frame internal initiation within ORF-
395 M (the predominant junction for ORF6 is upstream of this iORF, outside the range displayed,
396 and is not shown in the figure), (F) an extended version of ORF-M (reads marking ORF-M
397 initiation were cut to fit the scale indicated with black lines), (G) uORF that overlap ORF10
398 initiation and in frame internal initiation generating truncated ORF10 product (H) two uORFs
399 embedded in ORF1a 5'UTR (I) non canonical CUG initiation upstream of the TRS leader.

400

401

402 **Figure 4.** Translation of host and viral genes

403 (A) Relative translation levels of viral coding genes were estimated by counting the ribosome
404 densities on each ORF considering only non-overlapping regions. Data is presented from VeroE6
405 5hpi and Calu3 7hpi. ORFs are ordered based on their genomic location. (B) Viral ORF
406 translation levels as calculated from ribosome densities of infected Vero E6 cells. Data is plotted
407 on a log scale to cover the wide range in expression. Solid fill represents canonical ORFs, and
408 striped fill represents novel ORFs that were annotated. Values were normalized to ORF length
409 and sequencing depth. Points represent the values from the two replicates, except in the case of
410 3a.iORF2 and 1a.uORF2 where there was a missing value in one of the replicates. (C) Scatter
411 plot showing the expression of viral ORFs in infected Vero E6 cells and infected Calu3 cells.
412 Points representing canonical ORFs are outlined in black. Spearman's R is presented. (D -F)
413 Relative transcript abundance versus ribosome densities for each host and viral ORF at 5 hpi (D)
414 and 24 hpi (E) in Vero E6 cells and at 7hpi in Calu3 cells (F). Transcript abundance was
415 estimated by counting the reads that span the corresponding junction (only the most abundant

416 viral transcripts were counted) and footprint densities were calculated from the CHX sample. For
417 ribo-seq viral reads from 24 hpi, only reads that were mapped to junctions were used to avoid
418 non-ribosome footprints. Pearson's R on log transformed values is presented for each sample.

419
420
421

422 **Tables legend:**

423

424 **Table S1.** Junctions sites detected from junction spanning reads.

425 This table lists junction sites that were identified by Kim et al. with more than 100 reads and
426 were also detected in our RNA reads.

427 The genomic coordinates in the “5' site” and “3' site” point to the 3'-most and the 5'-most
428 nucleotides that survive the recombination event, respectively. “Gap” is the size of the deletion.
429 “Leader” true value indicates the junction is TRS-leader dependent. “canonical” true value
430 indicates the junction supports the expression of a canonical ORF, “Kim_count” is the number of
431 the junction-spanning reads that support the recombination event identified by Kim et al. “ORF”
432 the name of an ORF that shares the start codon position with the recombination product based on
433 Kim et al. “mrna_05hr_1”, “mrna_05hr_2”, “mrna_24hr_1” and “mrna_24hr_2” the number of
434 the junction-spanning reads that support the recombination event in each of our RNA samples
435 based on STAR-aligner. “fp_chx_05hr_1”, “fp_chx_05hr_2”, “fp_chx_24hr_1” and
436 “fp_chx_24hr_2” the number of the junction-spanning reads that support the recombination
437 event in each of our footprints CHX samples. “sum_fp” the sum of all footprints counts.
438 “sum_mRNA” the sum of all RNA counts. “star_sum” the sum of number of the junction-
439 spanning reads in all samples

440

441 **Table S2.** Junctions sites uniquely detected in our samples.

442 This table lists junction sites that were identified in our RNA samples with more than 50 reads
443 but were low or unidentified by Kim et al.

444 The genomic coordinates in the “5' site” and “3' site” point to the 3'-most and the 5'-most
445 nucleotides that survive the recombination event, respectively. “Gap” is the size of the deletion.
446 “Leader” true value indicates the junction is TRS-leader dependent. “canonical” true value

447 indicates the junction supports the expression of a canonical ORF, “Kim_count” is the number of
448 the junction-spanning reads that support the recombination event identified by Kim et al. “ORF”
449 the name of an ORF that shares the start codon position with the recombination product based on
450 Kim et al. “mrna_05hr_1”, “mrna_05hr_2”, “mrna_24hr_1” and “mrna_24hr_2” the number of
451 the junction-spanning reads that support the recombination event in each of our RNA samples
452 based on STAR-aligner. “fp_chx_05hr_1”, “fp_chx_05hr_2”, “fp_chx_24hr_1” and
453 “fp_chx_24hr_2” the number of the junction-spanning reads that support the recombination
454 event in each of our footprints CHX samples. “sum_fp” the sum of all footprints counts.
455 “sum_mRNA” the sum of all RNA counts. “star_sum” the sum of number of the junction-
456 spanning reads in all samples.

457

458 **Table S3.** Potential junction spanning SARS-CoV-2 ORFs that can be translated from the
459 CUG initiation upstream the TRS-leader.

460 This table lists junction spanning SARS-CoV-2 ORFs that can be translated from the CUG
461 initiation upstream the TRS-leader at each of the sub-genomic RNAs. “Name” for each ORF,
462 “description”, “supported by PRICE Vero” whether the ORF was predicted by PRICE from the
463 Vero E6 data ,” supported by ORF-RATER Vero” whether the ORF was predicted by ORF-
464 RATER from the Vero E6 data, “supported by PRICE Calu3” and “supported by ORF-RATER
465 Calu3” whether the ORF was predicted by PRICE or ORF-RATER from the Calu3 data, the
466 “start position” and “end position” in SAS-CoV-2 genome, the nature of the “start codon”,
467 “size(aa)”, “sequence” and “junction position (start, end)”.
468

469 **Table S4.** Novel SARS-CoV-2 ORFs that have been identified in our study.

470 This table lists all the SARS-CoV-2 translated ORFs identified in this study. “Name” for each
471 ORF, “description”, , “supported by PRICE Vero” whether the ORF was predicted by PRICE
472 from the Vero E6 data ,” supported by ORF-RATER Vero” whether the ORF was predicted by
473 ORF-RATER from the Vero E6 data, “PRICE replicate detection Vero” and “ORF-RATER
474 replicate detection Vero” whether the ORF was detected by PRICE or ORF-RATER using only
475 one replicate if the data, replicate 1 (rep1) or replicate 2 (rep2), “supported by PRICE Calu3” and
476 “supported by ORF-RATER Calu3” whether the ORF was predicted by PRICE or ORF-RATER

477 from the Calu3 data,, the “start position” and “end position” in SAS-CoV-2 genome, the nature
478 of the “start codon”, “size(aa)”and “sequence”.

479

480 **Table S5.** Translation levels of SARS-CoV-2 ORFs.

481 This table lists all translated SARS-CoV-2 ORFs, canonical and newly identified, and their
482 estimated translation levels based on ribosome profiling. “ORF_ID” and “ORF_name” for
483 each ORF, “type” of ORF including upstream ORFs (uORF), in-frame and out-of-frame
484 internal ORFs (iORF and oof), extended versions of canonical ORF (extension) and canonical
485 ORFs. The genomic region used for calculation of coverage is shown in “Included region”.
486 For in-frame iORFs, the coverage in an upstream region of the main ORF, shown in square
487 brackets, was subtracted from the coverage in the included region to get an approximation of
488 expression. See methods for details. Normalized translation levels in VeroE6 cells and in
489 Calu3 cells are shown for each replicate (“chx_1_tpm” and “chx_2_tpm”) and as average
490 value (“chx_mean_tpm”). Alternative calculation of translation levels using ORF-RATER
491 that utilize regression is also presented (“ORF-RATER_chx_1”, “ORF-RATER_chx_2” and
492 “ORF-RATER_chx_mean”).

493

494 **Table S6.** Multiple sequence alignment for canonical and novel SARS-CoV-2 ORFs in
495 Sarbecoviruses.

496 This table includes links for annotated multiple sequence alignment (MSA) views of all
497 SARS-CoV-2, canonical and novel, described in this study, using the CodAlignView tool ³⁷.
498 The MSA includes SARS-CoV-2, SARS-CoV and 42 bat coronavirus genomes ³⁰.

499

500

501

502

503 **Acknowledgements**

504 We thank Stern-Ginossar lab members, Igor Ulitsky and Schraga Schwartz for providing
505 valuable feedback, to Miri Shnayder, Igor Ulitsky and Noa Gil for technical assistance. We thank
506 Emanuel Wyler for the Calu3 cells. We thank Inbar Cohen-Gihon for sharing sequencing and
507 bioinformatics data. This study was supported by the Ben B. and Joyce E. Eisenberg Foundation.
508 Work in the Stern-Ginossar lab is supported by a European Research Council starting grant (StG-
509 2014-638142) and by the Israel Science Foundation (ISF) grant no. 1526/18. S.W-G. is the
510 recipient of the Human Frontier Science Program fellowship (LT-000396/2018), EMBO non-
511 stipendiary Long-Term Fellowship (ALTF 883-2017), the Gruss-Lipper Postdoctoral Fellowship,
512 the Zuckerman STEM Leadership Program Fellowship and the Rothschild Postdoctoral
513 Fellowship. N.S-G is an incumbent of the Skirball Career Development Chair in New Scientists
514 and is a member of the European Molecular Biology Organization (EMBO) Young Investigator
515 Program. The authors declare no competing interests.

516

517 **Author contributions**

518 Y.F., O.M., N.P and N.S-G. conceptualization. O.M. experiments. Y.F., A.N. and S.W-G. data
519 analysis. Y.Y-R., H.T., H.A., S.M., S.W., I.C-G, D.S, O.I, A. B-D, T.I. and N.P. work with
520 SARS-CoV-2. D.M mined published proteomic data, Y.F., O.M., A.N., M.S. and N.S-G.
521 interpreted data. M.S. and N.S.-G. wrote the manuscript with contribution from all other
522 authors.

523

524 **Material and methods**

525 Cells and viruses

526 Vero C1008 (Vero E6) (ATCC CRL-1586TM) were cultured in T-75 flasks with DMEM
527 supplemented with 10% fetal bovine serum (FBS), MEM non-essential amino acids, 2mM L-
528 Glutamine, 100Units/ml Penicillin, 0.1mg/ml streptomycin, 12.5Units/ml Nystatin (Biological
529 Industries, Israel). Calu3 cells were cultured in 10cm plates with DMEM supplemented with
530 10% fetal bovine serum (FBS), MEM non-essential amino acids, 2mM L-Glutamine,
531 100Units/ml Penicillin, 1% non-essential amino acid and 1% Na-pyrovate. Monolayers were
532 washed once with DMEM (for VeroE6) or RPMI (for Calu3) without FBS and infected with
533 SARS-CoV-2 virus, at a multiplicity of infection (MOI) of 0.2, For calu3 infection 20 ug per ml
534 TPCK trypsin (Thermo scientific) were added. After 1hr infection cells were cultured in their
535 respective medium supplemented with 2% fetal bovine serum, and MEM non-essential amino
536 acids, L glutamine and penicillin-streptomycin-Nystatin at 37°C, 5% CO₂. SARS-CoV-2
537 (GISAID Acc. No. EPI_ISL_406862), was kindly provided by Bundeswehr Institute of
538 Microbiology, Munich, Germany. It was propagated (4 passages) and tittered on Vero E6 cells
539 and then sequenced (details below) before it was used. SARS-CoV-2 BavPat1/2020 Ref-SKU:
540 026V-03883 was kindly provided by Prof. C. Drosten, Charité – Universitätsmedizin Berlin,
541 Germany. It was propagated (5 passages), tittered on Vero E6 and then sequenced before it has
542 been used in experiments. Infected cells were harvested at the indicated times as described
543 below. Handling and working with SARS-CoV-2 virus was conducted in a BSL3 facility in
544 accordance with the biosafety guidelines of the Israel Institute for Biological Research. The
545 Institutional Biosafety Committee of Weizmann Institute approved the protocol used in these
546 studies.

547

548 Preparation of ribosome profiling and RNA sequencing samples

549 For RNA-seq, cells were washed with PBS and then harvested with Tri-Reagent (Sigma-
550 Aldrich), total RNA was extracted, and poly-A selection was performed using Dynabeads
551 mRNA DIRECT Purification Kit (Invitrogen) mRNA sample was subjected to DNaseI
552 treatment and 3' dephosphorylation using FastAP Thermosensitive Alkaline Phosphatase
553 (Thermo Scientific) and T4 PNK (NEB) followed by 3' adaptor ligation using T4 ligase (NEB).

554 The ligated products used for reverse transcription with SSIII (Invitrogen) for first strand cDNA
555 synthesis. The cDNA products were 3' ligated with a second adaptor using T4 ligase and
556 amplified for 8 cycles in a PCR for final library products of 200-300bp. For Ribo-seq libraries,
557 cells were treated with either 50µM lactimidomycin (LTM) for 30 minutes or 2µg/mL
558 Harringtonine (Harr) for 5 minutes, for translation initiation libraries (LTM and Harr libraries
559 correspondingly), or left untreated for the translation elongation libraries (cycloheximide [CHX]
560 library). All three samples were subsequently treated with 100µg/mL CHX for 1 minute. Cells
561 were then placed on ice, washed twice with PBS containing 100µg/mL CHX, scraped from the
562 T-75 flasks (Vero cells) or 10cm plates (Calu3 cells), pelleted and lysed with lysis buffer (1%
563 triton in 20mM Tris 7.5, 150mM NaCl, 5mM MgCl₂, 1mM dithiothreitol supplemented with 10
564 U/ml Turbo DNase and 100µg/ml cycloheximide). After lysis samples stood on ice for 2h and
565 subsequent Ribo-seq library generation was performed as previously described ⁵. Briefly, cell
566 lysate was treated with RNaseI for 45min at room temperature followed by SUPERse-In
567 quenching. Sample was loaded on sucrose solution (34% sucrose, 20mM Tris 7.5, 150mM NaCl,
568 5mM MgCl₂, 1mM dithiothreitol and 100µg/ml cycloheximide) and spun for 1h at 100K RPM
569 using TLA-110 rotor (Beckman) at 4c. Pellet was harvested using TRI reagent and the RNA was
570 collected using chloroform phase separation. For size selection, 15uG of total RNA was loaded
571 into 15% TBE-UREA gel for 65min, and 28-34 footprints were excised using 28 and 34 flanking
572 RNA oligos, followed by RNA extraction and ribo-seq protocol ⁵

573 Virus genomic sequencing

574 RNA from viruses (culture supernatant after removal of cell debris) was extracted using viral
575 RNA kit (Qiagen). The SMARTer Pico RNA V2 Kit (Clontech) was used for library preparation.
576 Genome sequencing was conducted on the Illumina MiSeq platform, in a single read mode 60bp
577 for BetaCoV/Germany/BavPat1/2020 EPI_ISL_406862 and in a paired-end mode 150bp x2 for
578 BavPat1/2020 Ref-SKU: 026V-03883 producing 2,239,263 and 4,332,551 reads
579 correspondingly. Reads were aligned to the viral genome using STAR 2.5.3a aligner. Even
580 coverage along the genome was assessed and the relative abundance junctions (that may reflect
581 genomic deletion) were calculated. For EPI_ISL_406862 passage 4 (that was used for Vero cells
582 infection) the junctions that were found in more than 1% of genomes are listed in Table S2. For
583 BavPat1/2020 Ref-SKU: 026V-03883 passage 5 (that was used to for Calu3 infection) no

584 junctions in abundance of more than 1% of the genomes were detected. All genomic sequencing
585 data was deposited.

586

587 Sequence alignment, normalization and metagene analysis

588 Sequencing reads were aligned as previously described ³⁸. Briefly, linker
589 (CTGTAGGCACCATCAAT) and poly-A sequences were removed and the remaining reads
590 from were aligned to the Chlorocebus sabaeus genome (ENSEMBL release 99) and to the
591 SARS-CoV-2 genomes [Genebank NC_045512.2 with 3 changes to match the used strain
592 (BetaCoV/Germany/BavPat1/2020 EPI_ISL_406862), 241:C->T, 3037:C->T, 23403:A->G].
593 (infection of Vero cells) or to the Hg19 and NC_045512.2 with the same sequence changes
594 (infection of Calu3). Alignment was performed using Bowtie v1.1.2 ³⁹ with maximum two
595 mismatches per read. Reads that were not aligned to the genome were aligned to the
596 transcriptome of Chlorocebus sabaeus (ENSEMBL) and to SARS-CoV-2 junctions that were
597 recently annotated ¹⁸. The aligned position on the genome was determined as the 5' position of
598 RNA-seq reads, and for Ribo-seq reads the p-site of the ribosome was calculated according to
599 reads length using the off-set from the 5' end of the reads that was calculated from canonical
600 cellular ORFs. The offsets used are +12 for reads that were 28-29 bp and +13 for reads that were
601 30-33 bp. Reads that were in different length were discarded. In all figures presenting ribosome
602 densities data all footprint lengths (28-33bp) are presented.

603 Novel junctions were mapped using STAR 2.5.3a aligner ⁴⁰, with running flags as suggested at
604 Kim et. al., to overcome filtering of non-canonical junctions. Reads aligned to multiple locations
605 were discarded. Junctions with 5' break sites mapped to genomic location 55-85 were assigned
606 as leader-dependent junctions. Matching of leader junctions to ORFs, and categorization of
607 junctions as canonical or non-canonical, was adapted from Kim et. al. ¹⁸ Supplementary table 3,
608 or was assigned manually for strong novel junctions that appear only in our data.

609 For the metagene analysis only genes with more than 50 reads were used. For each gene
610 normalization was done to its maximum signal and each position was normalized to the number
611 of genes contributing to the position. In the virus 24hr samples, normalization for each gene was
612 done to its maximum signal within the presented region.

613 For comparing transcript expression level, mRNA and footprint counts from bowtie alignment
614 were normalized to units of RPKM in order to normalize for gene length and for sequencing
615 depth, based on the total number reads for both the host and the virus. The deconvolution of
616 RPKM for RNAs was done by subtracting the RPKM of a gene from the RPKM of the gene
617 located just upstream of it in the genome. The junction counts were based on STAR alignment
618 number of uniquely mapped reads crossing the junction.

619 The estimation of the viral footprint densities from the 24 hpi samples was performed by
620 calculating the ratio of the RPKM of ORF1a to the total number of leader canonical junctions at
621 5hpi. This ratio was used as a factor to calculate a proxy for the “true” viral footprint densities
622 from the number of footprints that were mapped to leader canonical junctions at 24hpi.

623 To quantify the translation levels of novel viral ORFs at 5hpi and 7hpi, many of which are
624 overlapping, three types of calculations were used based on ORF type. For ORFs that have a
625 unique region, with no overlap to any other ORF, bowtie aligned read density was calculated in
626 that region. For out-of-frame internal ORFs, the read density of the internal ORF region was
627 calculated by estimating the expected 3-bp periodicity distribution of footprints based on non-
628 overlapping translated regions in the main ORF. Using linear regression, we calculated the
629 relative contribution of the frames of the main and of the internal ORF to the reads covering the
630 region of the internal ORF. The relative contribution of the internal ORF was then multiplied by
631 the read density in that region to obtain the estimated translation level of the internal out-of-
632 frame ORF. For in-frame internal ORFs the read density of the main overlapping ORF is
633 calculated from a non-overlapping region and then subtracted from the read density in the
634 overlapping internal ORF region to get an estimate of translation levels of the internal ORF. In
635 cases where the unique region used to calculate read density contained the start-codon of the
636 ORF, the first 20% of the codons in the region were excluded from the calculation to avoid bias
637 from initiation peaks, unless the region was very short and trimming it would harm the ability to
638 estimate coverage (ORF 8 and extended ORF M). The exact regions that were used for
639 calculation can be found in Table S5. Finally, read density was normalized to the length of the
640 region used for calculation and to the sum of length normalized reads in each sample to get TPM
641 values. P-values for the relative contribution levels of out-of-frame ORFs were calculated from
642 both replicates using a mixed-effects linear model using the 3-base periodicity distribution as the
643 fixed effect and the replicates as random effect. In parallel, ORF-RATER was used to quantify

644 the translation levels of the viral ORFs (using regression), giving largely similar values
645 (Spearman's R = 0.92 and R = 0.87 in VeroE6 and Calu3, respectively).

646

647

648 Prediction of translation initiation sites and transmembrane domains

649 Translation initiation sites were predicted using PRICE ²⁴ and ORF-RATER ²⁵. To estimate the
650 codons generating the sequencing reads with maximum likelihood, PRICE requires a predefined
651 set of annotated coding sequences from the same experiment. Thus, it does not perform well on
652 reference sequences with a small number of annotated ORFs such as SARS-CoV-2. Since our
653 experiment generated ribosome footprints from both SARS-CoV-2 and host mRNAs, which
654 were exposed to the exact same conditions in the protocol, we used annotated CDSs from the
655 host cells to evaluate the parameters of the experiment. For libraries of infected Vero cells
656 sequencing reads were aligned using Bowtie to a fasta file containing chromosome 20 of
657 Chlorocebus sabaeus (1240 annotated start codons, downloaded from ensembl:
658 ftp://ftp.ensembl.org/pub/release99/fasta/chlorocebus_sabaeus/dna/) and the genomic sequence
659 of SARS-CoV-2 (Refseq NC_045512.2). A gtf file with the annotations of Chlorocebus sabaeus
660 and SARS-CoV-2 genomes was constructed and provided as the annotations file when running
661 PRICE. For technical reasons, the annotation of the first coding sequence (CDS) of the two
662 CDSs in the "ORF1ab" gene was deleted since having two CDSs encoded from a single gene
663 was not permitted by PRICE. For libraries of infected Calu3 cells sequencing reads were mapped
664 to a fasta file containing chromosome 1 of hg19 (2843 annotated start codons) and the genomic
665 sequence of SARS-CoV-2 (Refseq NC_045512.2). A gtf file with the annotations of hg19 and
666 SARS-CoV-2 genomes was constructed and provided as the annotations file when running
667 PRICE. For the data that was generated from infected Vero cells at 5hpi training and ORF
668 prediction by PRICE were done once using the CHX data from both replicates, and again using
669 all Ribo-seq libraries from both replicates, and the resulting predictions were combined. To test
670 reproducibility, the same predictions were performed on each replicate separately. For the data
671 that was generated from infected Calu3 cells at 7hpi training and ORF prediction by PRICE were
672 done using all Ribo-seq libraries from both replicates. The predictions were further filtered to
673 include only ORFs with at least 100 reads at the initiation site in the LTM samples of at least one

674 replicate. ORFs were then defined by extending each initiating codon to the next in-frame stop
675 codon.

676 ORF-RATER was used with the default values besides allowing all start codons with at most one
677 mismatch to ATG. For each cell type, two runs of ORF-RATER were used. One in which ORF-
678 RATER was trained on cellular annotations (chr 20 for the Vero cells, and chr 1 for the Calu
679 cells) and SARS-CoV-2 canonical ORFs (similar to the procedure that was used for running
680 PRICE). In the second run only SARS-CoV-2 canonical ORFs were used for training. In both
681 cases ORF1b and ORF10 were omitted from the training set. BAM files from STAR alignment
682 were used as input. The CHX data from both replicates was used in the first prune step to omit
683 low coverage ORFs. The calculations of the P-site offsets, and the regression were performed for
684 each Ribo-Seq library separately. The final score was calculated based on all three types of
685 libraries. Score of 0.5 was used as cut-off for the final predictions these were further manually
686 curated. Additional ORFs that were not recognized by the trained models (likely due to
687 differences in the features of viral genome compared to cellular genomes) but presented
688 reproducible translation profile in the two cell lines were added manually to the final ORF list
689 (Table S4). ORFs were manually identified as such if they had reproducible initiation peaks in
690 the CHX libraries that were enhanced in the LTM and Harr libraries, and exhibited increased
691 CHX signal in the correct reading-frame along the coding region.

692 Mapping reads to CUG initiation upstream the TRS-leader

693 Reads from ribosome profiling libraries were aligned using bowtie to a single reference that
694 contained the transcripts and the genome allowing no mismatches or gaps. Reads with p-site
695 mapped to position 59 of the viral genome were collected and divided to four groups according
696 to the nucleotide in position +17 of the read (position 76 of the genome). The first group contains
697 reads that are short (28 nucleotides) and do not have any nucleotide at position +17. The other
698 three groups, referred to as T, A and G, correspond to combinations of genomic and subgenomic
699 RNAs based on their sequence, as shown in figure S14. Group T is attributed to the genome or to
700 ORF E and ORF M subgenomic RNAs, group A to the subgenomic RNAs of ORF S, ORF7a,
701 ORF8 and ORF N, and group G to the sub-genomic RNA of ORF 6. Reads mapped uniquely to
702 the subgenomic RNA of ORF3a were excluded from calculation, and the number of reads in
703 each group was summed. Group T, containing genomic reads, was further divided based on the

704 nucleotide at position +18, where reads with A at that position can originate from the
705 subgenomic RNA of ORF M and reads with T at that position can originate from the genome or
706 from the subgenomic RNA of ORF E. Final division of the genomic group was done based on
707 position +19 where T corresponds to genomic reads and A corresponds to ORF E subgenomic
708 reads. RNA values as calculated from junction densities (described above) were summed for the
709 subgenomic and genomic RNAs in each group. The analysis was performed for each ribosome
710 profiling library separately.

711 Mining of proteomics data

712 Data downloaded from Bojkova et al.¹⁴ was searched using Byonic search engine using 10ppm
713 tolerance for MS1 and 20ppm tolerance for MS2, against the concatenated database containing
714 our 26 novel ORFs as well as the human proteome DB (SwissProt Nov2019), and the SARS-
715 CoV-2 proteome. Modifications allowed were fixed carbamidomethylation on C, fixed TMT6 on
716 K and peptide N-terminus, variable K8 and R10 SILAC labeling, variable M oxidation and
717 Variable NQ deamidation. Data downloaded from Davidson et al.¹⁵ was searched using Byonic
718 search engine using 10ppm tolerance for MS1 and 0.6Da tolerance for MS2, against the
719 concatenated database containing our 26 novel ORFs as well as the human proteome DB
720 (SwissProt Nov2019), and the SARS-CoV-2 proteome. Modifications allowed were fixed
721 carbamidomethylation on C, variable N-terminal protein acetylation, M oxidation and NQ
722 deamidation.

723 Immunofluorescence

724 Cells were plated on ibidi slides, fixed in 3% paraformaldehyde for 20 minutes, permeabilized
725 with 0.5% Triton X-100 in PBS for 2 minutes, and then blocked with 2% FBS in PBS for 30
726 minutes. Immunostaining was performed with rabbit anti-SARS-CoV-2 serum. Cells were
727 washed and labeled with anti-rabbit FITC antibody and with DAPI (4,6-diamidino-2-
728 phenylindole). Imaging was performed on a Zeiss AxioObserver Z1 wide-field microscope using
729 a X40 objective and Axiocam 506 mono camera.

730 Data availability

731 All next-generation sequencing data files were deposited in Gene Expression Omnibus under
732 accession number GSE149973.

733 All the RNA-seq and ribosome profiling data generated in this study can be accessed through a
734 UCSC browser session: <http://genome.ucsc.edu/s/aharonn/CoV2%2DTranslation>
735

736 **References**

- 737 1. Zhu, N. *et al.* A novel coronavirus from patients with pneumonia in China, 2019. *N. Engl.
738 J. Med.* **382**, 727–733 (2020).
- 739 2. Zhou, P. *et al.* A pneumonia outbreak associated with a new coronavirus of probable bat
740 origin. *Nature* **579**, 270–273 (2020).
- 741 3. Ingolia, N. T., Lareau, L. F. & Weissman, J. S. Ribosome profiling of mouse embryonic
742 stem cells reveals the complexity and dynamics of mammalian proteomes. *Cell* **147**, 789–
743 802 (2011).
- 744 4. Lee, S. *et al.* Global mapping of translation initiation sites in mammalian cells at single-
745 nucleotide resolution. *Proc. Natl. Acad. Sci. U. S. A.* **109**, E2424–E2432 (2012).
- 746 5. Finkel, Y. *et al.* Comprehensive annotations of human herpesvirus 6A and 6B genomes
747 reveal novel and conserved genomic features. *Elife* **9**, e50960 (2020).
- 748 6. Stern-Ginossar, N. *et al.* Decoding human cytomegalovirus. *Science (80-.).* **338**, 1088–
749 1093 (2012).
- 750 7. Irigoyen, N. *et al.* High-Resolution Analysis of Coronavirus Gene Expression by RNA
751 Sequencing and Ribosome Profiling. *PLoS Pathog.* **12**, 1005473 (2016).
- 752 8. Dinan, A. M. *et al.* Comparative Analysis of Gene Expression in Virulent and Attenuated
753 Strains of Infectious Bronchitis Virus at Subcodon Resolution. *J. Virol.* **93**, 714–733
754 (2019).
- 755 9. Snijder, E. J., Decroly, E. & Ziebuhr, J. The Nonstructural Proteins Directing Coronavirus
756 RNA Synthesis and Processing. in *Advances in Virus Research* **96**, 59–126 (Academic
757 Press Inc., 2016).
- 758 10. Sola, I., Almazán, F., Zúñiga, S. & Enjuanes, L. Continuous and Discontinuous RNA
759 Synthesis in Coronaviruses. *Annu. Rev. Virol.* **2**, 265–288 (2015).
- 760 11. Stadler, K. *et al.* SARS — beginning to understand a new virus. *Nat. Rev. Microbiol.* **1**,
761 209–218 (2003).
- 762 12. Lai, M. M. & Stohlman, S. A. Comparative analysis of RNA genomes of mouse hepatitis
763 viruses. *J. Virol.* **38**, 661–670 (1981).

- 764 13. Yogo, Y., Hirano, N., Hino, S., Shibuta, H. & Matumoto, M. *Polyadenylate in the virion*
765 *RNA of mouse hepatitis virus. Journal of Biochemistry* **82**, (1977).
- 766 14. Bojkova, D. *et al.* Proteomics of SARS-CoV-2-infected host cells reveals therapy targets.
767 (2020). doi:10.1038/s41586-020-2332-7
- 768 15. Davidson, A. D. *et al.* Characterisation of the transcriptome and proteome of SARS-CoV-
769 2 using direct RNA sequencing and tandem mass spectrometry reveals evidence for a cell
770 passage induced in-frame deletion in the spike glycoprotein that removes the furin-like
771 cleavage site. *bioRxiv* 2020.03.22.002204 (2020). doi:10.1101/2020.03.22.002204
- 772 16. Ingolia, N. T. *et al.* Ribosome Profiling Reveals Pervasive Translation Outside of
773 Annotated Protein-Coding Genes. *Cell Rep.* **8**, 1365–1379 (2014).
- 774 17. Plant, E. P., Rakauskaite, R., Taylor, D. R. & Dinman, J. D. Achieving a golden mean:
775 mechanisms by which coronaviruses ensure synthesis of the correct stoichiometric ratios
776 of viral proteins. *J. Virol.* **84**, 4330–40 (2010).
- 777 18. Kim, D. *et al.* The architecture of SARS-CoV-2 transcriptome. *Cell* **S0092-8674**, 30406–2
778 (2020).
- 779 19. Schaecher, S. R., Mackenzie, J. M. & Pekosz, A. The ORF7b protein of severe acute
780 respiratory syndrome coronavirus (SARS-CoV) is expressed in virus-infected cells and
781 incorporated into SARS-CoV particles. *J. Virol.* **81**, 718–31 (2007).
- 782 20. Liu, Z. *et al.* Identification of a common deletion in the spike protein of SARS-CoV-2.
783 *bioRxiv* 2020.03.31.015941 (2020). doi:10.1101/2020.03.31.015941
- 784 21. Ogando, N. S. *et al.* SARS-coronavirus-2 replication in Vero E6 cells: replication kinetics,
785 rapid adaptation and cytopathology. *bioRxiv* 2020.04.20.049924 (2020).
786 doi:10.1101/2020.04.20.049924
- 787 22. Taiaroa, G. *et al.* Direct RNA sequencing and early evolution of SARS-CoV-2. *bioRxiv*
788 2020.03.05.976167 (2020). doi:10.1101/2020.03.05.976167
- 789 23. Wu, A. *et al.* Genome Composition and Divergence of the Novel Coronavirus (2019-
790 nCoV) Originating in China. *Cell Host Microbe* **27**, 325–328 (2020).
- 791 24. Erhard, F. *et al.* Improved Ribo-seq enables identification of cryptic translation events.

- 792 *Nat. Methods* **15**, 363–366 (2018).
- 793 25. Fields, A. P. *et al.* A Regression-Based Analysis of Ribosome-Profiling Data Reveals a
794 Conserved Complexity to Mammalian Translation. *Mol. Cell* **60**, 816–827 (2015).
- 795 26. Bouhaddou, M. *et al.* The Global Phosphorylation Landscape of SARS-CoV-2 Infection.
796 *Cell* (2020). doi:10.1016/j.cell.2020.06.034
- 797 27. Stukalov, A. *et al.* Multi-level proteomics reveals host-perturbation strategies of SARS-
798 CoV-2 and SARS-CoV. *bioRxiv* 2020.06.17.156455 (2020).
799 doi:10.1101/2020.06.17.156455
- 800 28. Cagliani, R., Forni, D., Clerici, M. & Sironi, M. Coding potential and sequence
801 conservation of SARS-CoV-2 and related animal viruses. *Infect. Genet. Evol.* **83**, (2020).
- 802 29. Firth, A. E. A putative new SARS-CoV protein, 3a*, encoded in an ORF overlapping
803 ORF3a. *bioRxiv* 2020.05.12.088088 (2020). doi:10.1101/2020.05.12.088088
- 804 30. Jungreis, I., Sealfon, R. & Kellis, M. Sarbecovirus comparative genomics elucidates gene
805 content of SARS-CoV-2 and functional impact of COVID-19 pandemic mutations.
806 *bioRxiv* 2020.06.02.130955 (2020). doi:10.1101/2020.06.02.130955
- 807 31. Gordon, D. E. *et al.* A SARS-CoV-2 protein interaction map reveals targets for drug
808 repurposing. *Nature* **583**, 459–468 (2020).
- 809 32. Hachim, A. *et al.* Beyond the Spike: identification of viral targets of the antibody response
810 to SARS-CoV-2 in COVID-19 patients. *medRxiv* 2020.04.30.20085670 (2020).
811 doi:10.1101/2020.04.30.20085670
- 812 33. Yewdell, J. W. DRiPs solidify: Progress in understanding endogenous MHC class I
813 antigen processing. *Trends in Immunology* **32**, 548–558 (2011).
- 814 34. Abernathy, E. & Glaunsinger, B. Emerging roles for RNA degradation in viral replication
815 and antiviral defense. *Virology* **479–480**, 600–608 (2015).
- 816 35. Huang, C. *et al.* SARS coronavirus nspl protein induces template-dependent
817 endonucleolytic cleavage of mRNAs: Viral mRNAs are resistant to nspl-induced RNA
818 cleavage. *PLoS Pathog.* **7**, e1002433 (2011).
- 819 36. Kamitani, W., Huang, C., Narayanan, K., Lokugamage, K. G. & Makino, S. A two-

- 820 pronged strategy to suppress host protein synthesis by SARS coronavirus Nsp1 protein.
821 *Nat. Struct. Mol. Biol.* **16**, 1134–1140 (2009).
- 822 37. Jungreis, I. *et al.* Evolutionary Dynamics of Abundant Stop Codon Readthrough. *Mol.*
823 *Biol. Evol.* **33**, 3108–3132 (2016).
- 824 38. Tirosh, O. *et al.* The Transcription and Translation Landscapes during Human
825 Cytomegalovirus Infection Reveal Novel Host-Pathogen Interactions. *PLoS Pathog.* **11**,
826 e1005288 (2015).
- 827 39. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. Ultrafast and memory-efficient
828 alignment of short DNA sequences to the human genome. *Genome Biol.* **10**, R25 (2009).
- 829 40. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21
830 (2013).

831

Figure 1

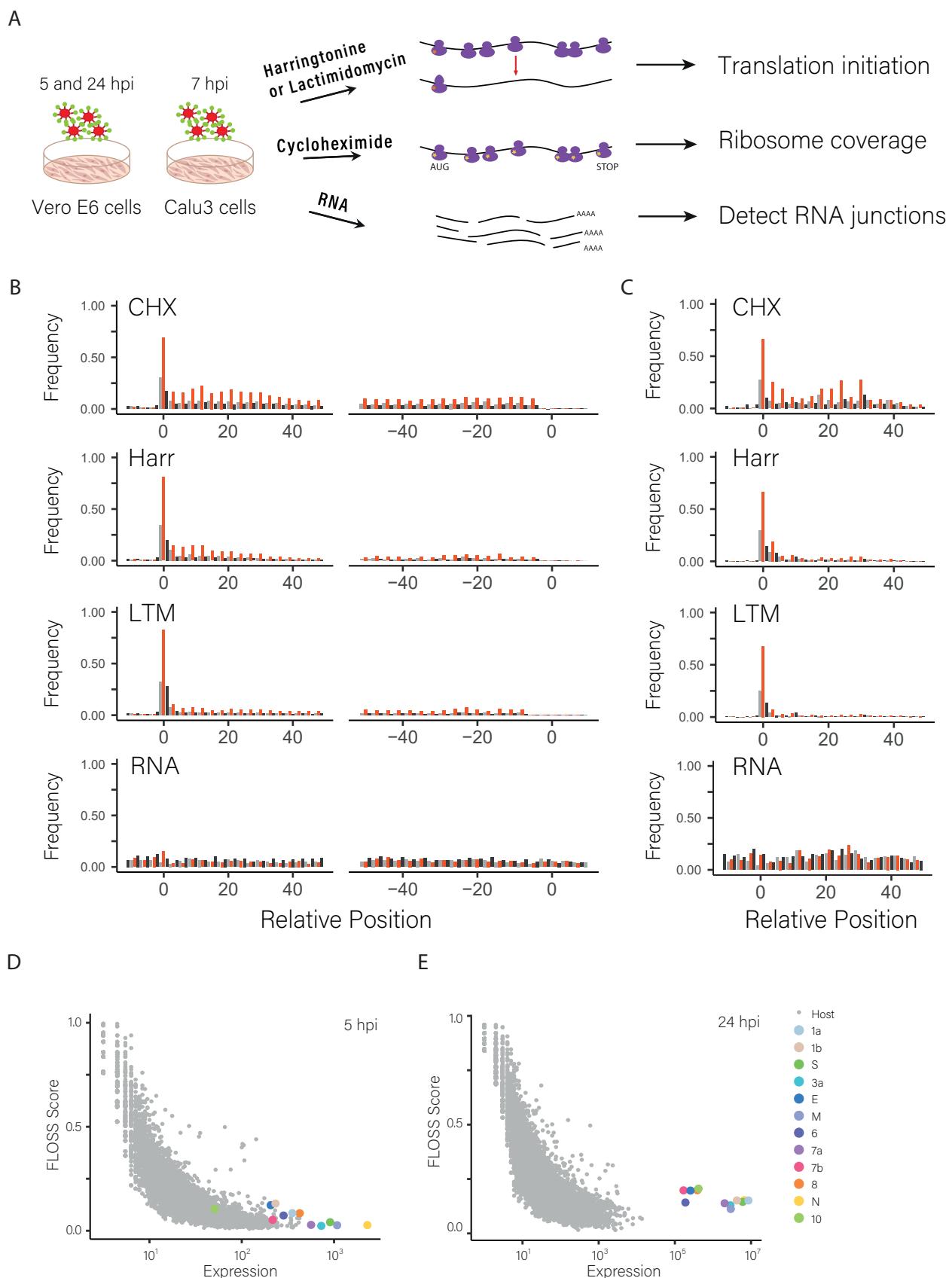


Figure 2

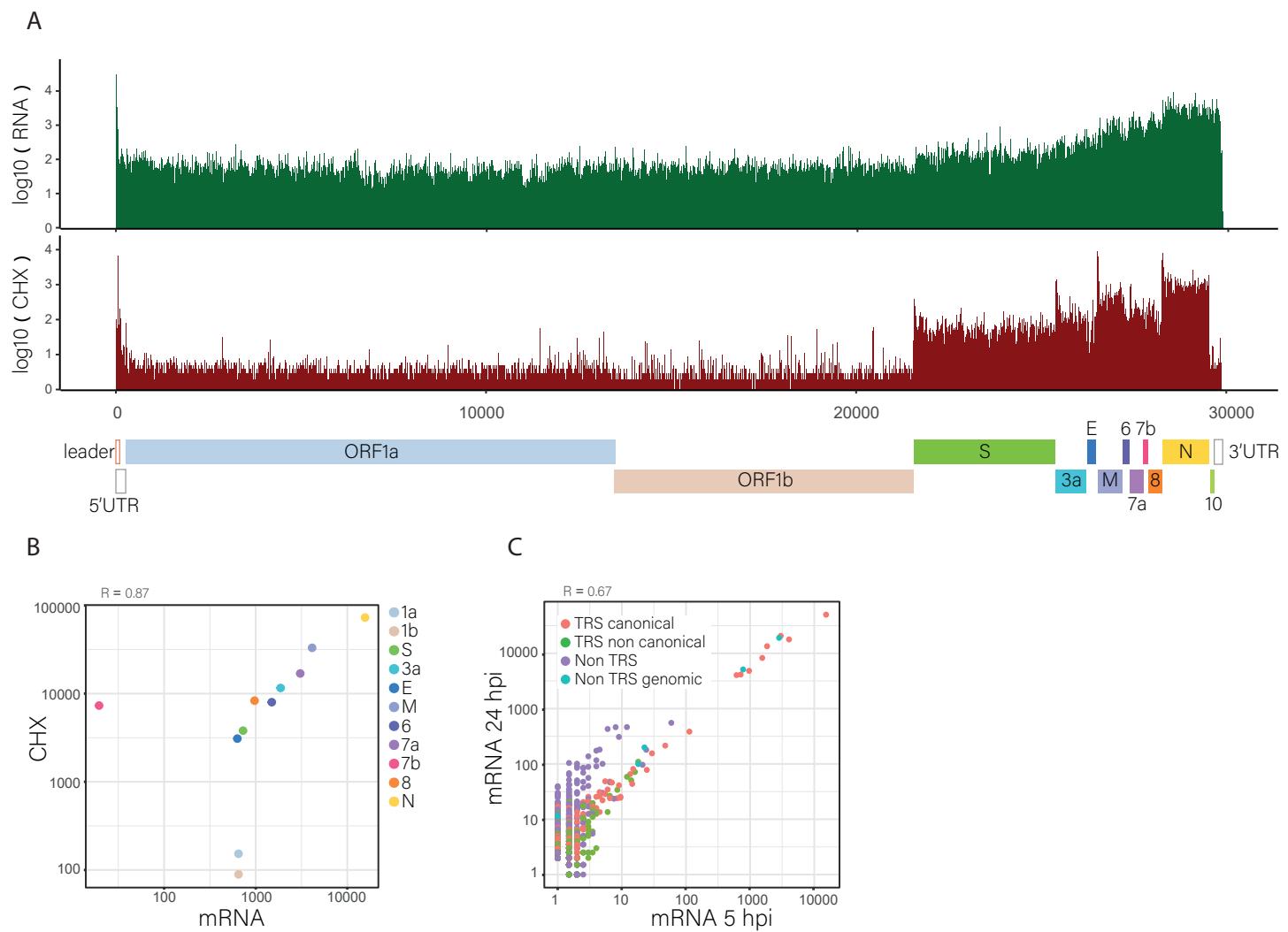
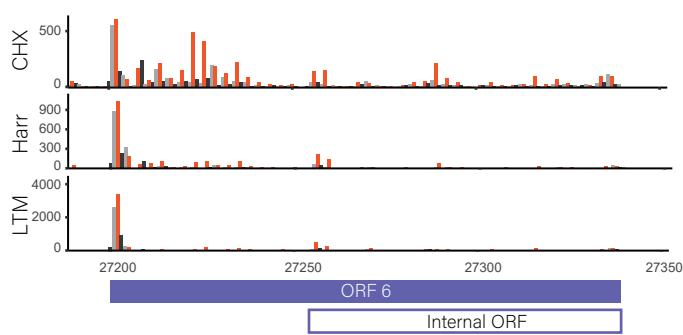
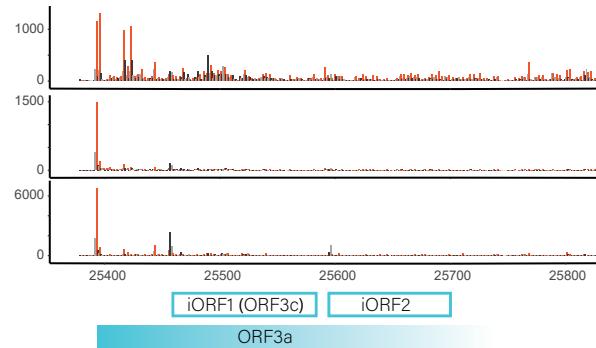


Figure 3

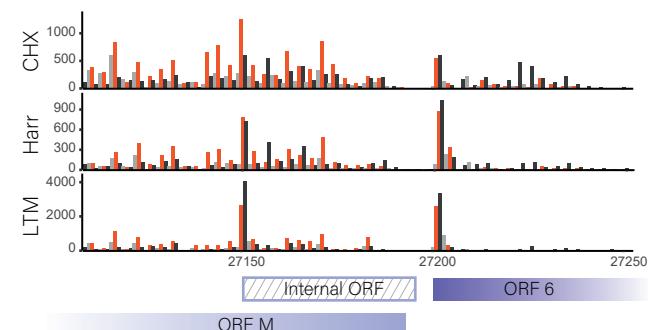
A



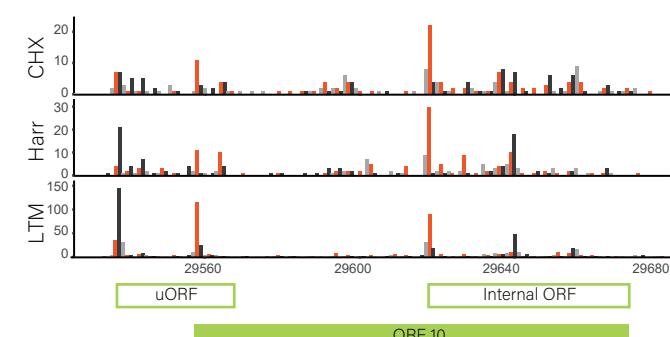
C



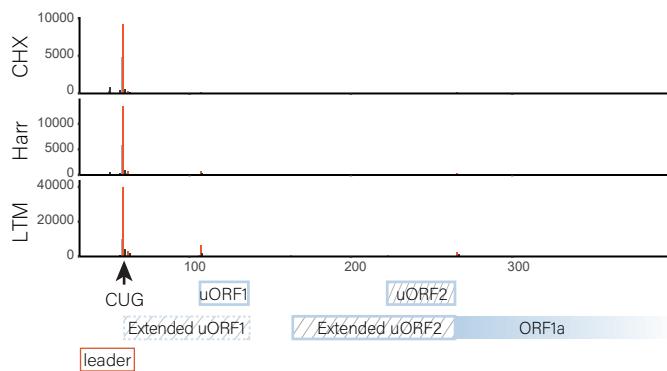
E



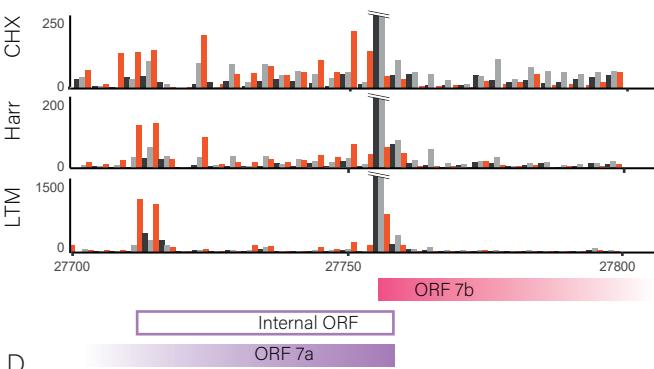
G



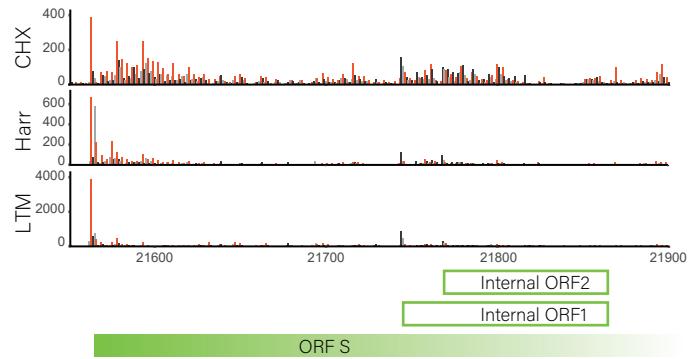
I



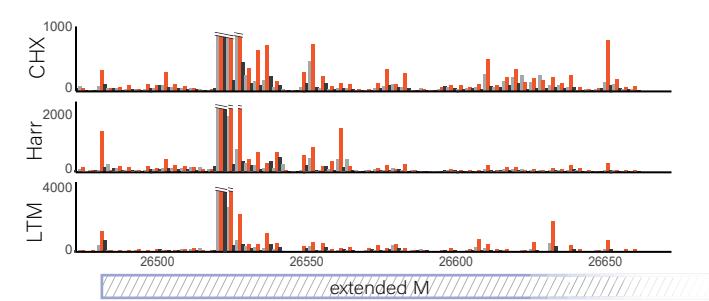
B



D



F



H

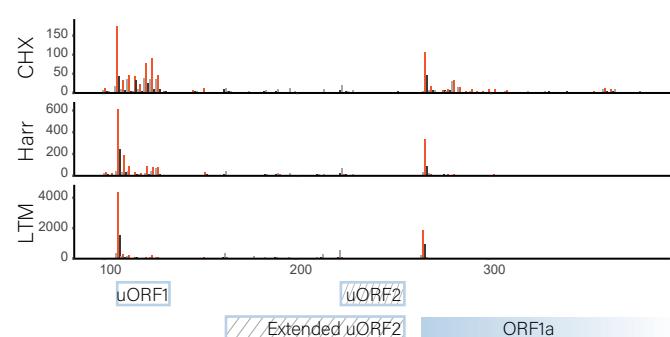


Figure 4

