The coding capacity of SARS-CoV-2

Yaara Finkel[1,5], Orel Mizrahi[1,5], Aharon Nachshon[1], Shira Weingarten-Gabbay[2,3], Yfat Yahalom-Ronen[4], Hadas Tamir[4], Hagit Achdout[4], Sharon Melamed[4], Shay Weiss[4], Tomer Israely[4], Nir Paran[4], Michal Schwartz[1] and Noam Stern-Ginossar[1,6 *]

[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] Department of Infectious Diseases, Israel Institute for Biological Research, Ness Ziona 76800, Israel.

[5] These authors contributed equally to this work

[6] Lead Contact

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

Abstract

SARS-CoV-2 is a coronavirus responsible for the COVID-19 pandemic. In order to understand its pathogenicity, antigenic potential and to develop diagnostic and therapeutic tools, it is essential to portray the full repertoire of its expressed proteins. The SARS-CoV-2 coding capacity map is currently based on computational predictions and relies on homology to other coronaviruses. Since coronaviruses differ in their protein array, especially in the variety of accessory proteins, it is crucial to characterize the specific collection of SARS-CoV-2 translated open reading frames (ORF)s in an unbiased and open-ended manner. Utilizing a suit of ribosome profiling techniques, we present a high-resolution map of the SARS-CoV-2 coding regions, allowing us to accurately quantify the expression of canonical viral ORFs and to identify 23 novel unannotated viral ORFs. These ORFs include several in-frame internal ORFs lying within existing ORFs, resulting in N-terminally truncated products, as well as internal out-of-frame ORFs, which generate novel polypeptides, such as a 97 amino acid (aa) poly-peptide that is translated from the ORF-N transcript. Finally, we detected a prominent initiation at a CUG codon located in the 5'UTR. Although this codon is shared by all SARS-CoV-2 transcripts, the initiation was specific to the genomic RNA, indicating that the genomic RNA harbors unique features that may affect ribosome engagement. Overall, our work reveals the full coding capacity of SARS-CoV-2 genome, providing a rich resource, which will form the basis of future functional studies and diagnostic efforts.

Introduction:

Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) is the cause of the ongoing Coronavirus disease 19 (COVID-19) pandemic (Zhou et al., 2020; Zhu et al., 2020). SARS-CoV-2 is an enveloped virus consisting of a positive-sense, single-stranded RNA genome of ~30 kb. The genome shares ~80% sequence identity with SARS-CoV and shows characteristic features of other coronaviruses (CoVs). Upon cell entry, two overlapping Open Reading Frames (ORFs) are translated from the positive strand genomic RNA, ORF1a (pp1a) and ORF1b (pp1ab). The translation of ORF1b is mediated by a -1 frameshift enabling read through of the stop codon. ORF1a and ORF1b encode continuous polypeptides which are cleaved into a total of 16 nonstructural proteins (NSPs, Snijder et al., 2016; Sola et al., 2015; Stadler et al., 2003). In addition, the viral genome is used by the viral RNA-dependent RNA polymerase (RdRP) to produce negative-strand RNA intermediates which serve as templates for the synthesis of positive-strand genomic RNA and of subgenomic RNAs (Snijder et al., 2016; Sola et al., 2015; Stadler et al., 2003). The subgenomic transcripts contain a common 5' leader fused to different segments from the 3′ end of the viral genome, and contain a 5′-cap structure and a 3′ poly(A) tail (Lai and Stohlman, 1981; Yogo et al., 1977). These unique fusions occur during negative-strand synthesis at 6-7 nt core sequences called transcription-regulating sequences (TRSs) that are located at the 3' end of the leader sequence as well as preceding each viral gene. The different subgenomic RNAs encode 4 conserved structural proteins- spike protein (S), envelope protein (E), membrane protein (M), nucleocapsid protein (N)- and several accessory proteins. Based on sequence similarity to other beta coronaviruses and specifically to SARS-CoV, current annotation of SARS-CoV-2 includes predictions of six accessory proteins (3a, 6, 7a, 7b, 8, and 10, NCBI Reference Sequence: NC_045512.2), but not all of these ORFs have been experimentally and reproducibly confirmed in this virus (Bojkova et al., 2020; Davidson et al., 2020).

Recently, two complementary deep-sequencing approaches were used to map the subgenomic RNAs and TRSs of SARS-CoV-2. In addition to the canonical subgenomic RNAs, numerous discontinuous transcription events were identified, including fusions of the 5′ leader to unexpected 3’ sites, long-distance fusions of sequences that are independent of the 5′ leader sequence, and local fusions yielding transcripts with small deletions (Kim et al., 2020). These results, together with additional direct RNA sequencing studies of SARS-CoV-2 (Davidson et

al., 2020; Taiaroa et al., 2020), suggest a highly complex transcriptome, however the effect of these non-canonical transcripts on SARS-CoV-2 coding capacity remains unclear. In addition, the presence of the annotated ORF10 was put into question (Kim et al., 2020; Taiaroa et al., 2020). The full annotation of the SARS-CoV-2 ORFs is a prerequisite for the understanding of the biology and pathogenesis of this virus and for efficient development of diagnostic and therapeutic tools.

Ribosome profiling (Ribo-seq), a high-throughput method that infers translated regions from deep sequencing of ribosome protected fragments, is a powerful genome-scale approach to accurately delineate translated sequences as well as to quantify gene expression. The robustness, scale and accuracy of this method dramatically increases our ability to define molecular events underlying viral infection with unprecedented depth. Application of ribosome profiling to cells infected with diverse viruses including *Murine coronavirus* (MHV-A59), revealed unanticipated complexity in their coding capacity, characterized the host response on the translation level and elucidated mechanisms of host shutoff (Bercovich-Kinori et al., 2016; Finkel et al., 2020; Irigoyen et al., 2016; Stern-Ginossar and Ingolia, 2015; Stern-Ginossar et al., 2012; Tirosh et al., 2015; Whisnant et al., 2020).

Here we combined Ribo-seq and RNA sequencing to systematically demarcate the landscape of translated ORFs of SARS-CoV-2 and their expression during infection. We found a high correlation between RNA levels and translation of most of the canonical ORFs. Moreover, we characterized 23 novel ORFs including upstream ORFs (uORFs) that are likely playing a role in regulating viral gene expression and out of frame internal ORFs (iORFs) that generate novel polypeptides as well as extensions and truncations of the known ORFs. We further show that viral mRNAs are not translated more efficiently than host mRNAs; rather, virus translation dominates host translation due to high levels of viral transcripts. Our results provide a comprehensive map of the translation landscape of SARS-CoV-2, which is essential for the functional understanding of its pathogenicity.

Results and discussion:

To capture the full complexity of SARS-CoV-2 coding capacity, we applied a suit of ribosome profiling approaches to Vero E6 cells infected with SARS-CoV-2 (BetaCoV/Germany/BavPat1/2020 EPI_ISL_406862) for 5 or 24 hours (Figure 1A). For each time point we mapped genome-wide translation events by preparing three different ribosome-profiling libraries (Ribo-seq), each one in two biological replicates. Two Ribo-seq libraries facilitate mapping of translation initiation sites, by treating cells with lactimidomycin (LTM) or harringtonine (Harr), two drugs with distinct mechanisms that inhibit translation initiation by preventing 80S ribosomes formed at translation initiation sites from elongating. These treatments lead to strong accumulation of ribosomes precisely at the sites of translation initiation and depletion of ribosomes over the body of the message [Figure 1A and (Finkel et al., 2020; Stern-Ginossar et al., 2012)]. The third Ribo-seq library was prepared from cells treated with the translation elongation inhibitor cycloheximide (CHX), and gives a snap-shot of actively translating ribosomes across the body of the translated ORF (Figure 1A). In parallel, we used a tailored RNA-sequencing (RNA-seq) protocol. When analyzing the different Ribo-seq libraries across coding regions in cellular genes, the expected distinct profiles are observed, confirming the overall quality of the libraries. Ribosome footprints displayed a strong peak at the translation initiation site, which, as expected, is more pronounced in the Harr and LTM libraries, while the CHX library also exhibited a distribution of ribosomes across the entire coding region up to the stop codon, and its mapped footprints were enriched in fragments that align to the translated frame (Figure 1B and Figure S1). As expected, the RNA-seq reads were uniformly distributed across the coding regions (Figure 1B and Figure S1).

The footprint profiles of viral coding sequences at 5 hours post infection (hpi) fit the expected profile of translation, similar to the profile of cellular genes, both at the meta gene level and at the level of individual genes (Figure 1C and 1D and Figure S2). Intriguingly, the footprint profile over the viral genome at 24 hpi, was substantially different (Figure S2), and did not fit the expected profile of translating ribosomes, as discussed below in detail.

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

cumulative abundance of these sequences due to the nested transcription of subgenomic RNAs (Figure 2A). Increased coverage is also seen at the 5′ UTR reflecting the presence of the 5′ leader sequence in all subgenomic RNAs as well as the genomic RNA. Ribosome density, as reflected by the footprints profiles, allows accurate quantification of viral gene translation and show the N ORF is expressed at the highest level followed by M, 7a, 3a, 6, 8, 7b, S, E, 10, 1a and 1b. (ORF10 is discussed in detail below) (Figure 2B and 2C). The reduction in footprint density between ORF1a and ORF1b reflects the proportion of ribosomes that terminate at the ORF1a stop codon instead of frameshifting into ORF1b. Frameshift efficiency is 57% +/- 12%, as calculated by dividing the footprint density in ORF1b by the footprint density in ORF1a. This value is comparable to the frameshift efficiency measured based on ribosome profiling of MHV (48%-75%, Irigoyen et al., 2016). On the molecular level this 57% frameshifting rate indicates NSP1-NSP10 are expressed 1.8 +/- 0.4 times more than NSP11-NSP16 and this ratio likely relates to the stoichiometry needed to generate SARS-CoV-2 nonstructural macromolecular complexes (Plant et al., 2010). It will be interesting to examine whether this stoichiometry changes along infection. Similarly to what was seen in MHV (Irigoyen et al., 2016), we failed to see noticeable ribosome pausing before or at the frameshift site, but we identified several potential pausing sites within ORF1a. We also observed few additional potential pausing sites in ORF1b that were reproducible between replicates (Figure S3), but these will require further characterization.

Besides ORF1a and ORF1b, all other canonical viral ORFs are translated from subgenomic RNAs. We therefore examined whether the levels of viral gene translation correlate with the levels of the corresponding subgenomic RNAs. Since raw RNA-Seq densities represent the cumulative sum of genomic and all subgenomic RNAs, we calculated transcript abundance using two approaches: deconvolution of RNA densities, in which RNA densities of each ORF are calculated by subtracting the RNA read density of cumulative densities upstream to the ORF region; and relative abundances of RNA reads spanning leader-body junctions of each of the canonical subgenomic RNAs. ORF6 and ORF7 obtained negative values in the RNA deconvolution, probably due to their short length and inaccuracies introduced by deconvolution and PCR biases. For all other canonical ORFs there was high correlation between these two approaches (R = 0.897, Figure S4), and in both approaches the N transcript was the most abundant transcript, in agreement with recent studies (Davidson et al., 2020; Kim et al., 2020).

We next compared footprint densities to RNA abundance as calculated by junction abundances for the subgenomic RNA or deconvolution of genomic RNA in the case of ORF1a and ORF1b (Figure 2D). Interestingly, for the majority of viral ORFs, transcript abundance correlated almost perfectly with footprint densities, indicating these viral ORFs are translated in similar efficiencies (probably due to their almost identical 5'UTRs), however three ORFs were outliers. The translation efficiency of ORF1a and ORF1b was significantly lower. This can stem from unique features in their 5'UTR (discussed below) or from under estimation of their true translation efficiency as some of the full-length RNA molecules may serve as template for replication or packaging and are hence not part of the translated mRNA pool. The third outlier is ORF7b for which we identified very few body-leader junctions but it exhibited relatively high translation. The translation of ORF7b in SARS-CoV was indeed suggested to arise from ribosome leaky scanning of the ORF7a transcript (Schaecher et al., 2007).

Recently, many transcripts derived from non-canonical junctions were identified for SARS-CoV-2, some of which were abundant and were suggested to affect the viral coding potential (Davidson et al., 2020; Kim et al., 2020). These non-canonical junctions contain either the leader combined with 3' fragments at unexpected sites in the middle of ORFs (leader-dependent noncanonical junction) or fusion between sequences that do not have similarity to the leader (leader-independent junction). We estimated the fusion frequency of these non-canonical junctions in our RNA libraries. We indeed identify many non-canonical junctions and obtained excellent agreement between our RNA-seq replicates for both canonical and non-canonical junctions, demonstrating these junctions are often generated and mostly do not correspond to random amplification artifacts (Figure S5A, S5B and Table S1). The abundance of junction-spanning reads between our data and the data of Kim et al. (Kim et al., 2020) that was generated from RNA harvested at 24 hpi showed significant correlation (R=0.816, Figure S5C and S5D), illustrating many of these are reproducible between experimental systems. However, 111 out of the 361 most abundant leader independent junctions that were mapped by Kim et al., were not supported by any reads in our data, illustrating there are also substantial variations. When we compared the frequency of junctions between 5h and 24h time points, the leader dependent transcripts (both canonical and non-canonical) correlated well but the leader independent transcripts were increased at 24 hpi (Figure 2E). Recent kinetics measurements show viral particles already bud out of infected Vero cells at 8 hpi (Ogando et al., 2020). This time-

dependent increase in non-canonical RNA junctions indicates that part of the leader independent RNA junctions might be associated with genomic replication. To further examine these non-canonical junctions, we also mapped the leader independent junctions in our RNA reads. We identified 5 abundant junctions (Table S2), and remarkably two of these junctions represent short deletions in the spike protein (7aa and 10aa long) that overlap deletions that were recently described by other groups (Davidson et al., 2020; Liu et al., 2020; Ogando et al., 2020), in which the furin-like cleavage site is deleted (Figure 2F) . The re-occurrence of the same deletion strongly supports the conclusion that this deletion is being selected for during passage in Vero cells and has clear implications for the use of Vero cells for propagating viruses (Davidson et al., 2020; Liu et al., 2020; Ogando et al., 2020). Another abundant junction we identified represents a potential 8aa deletion in ORF4b. Overall this data suggests that at least part of the leader-independent junctions may represent genomic deletions and that some of these deletions could be specifically selected for during cell culture passages.

Our ribosome profiling approach also facilitates unbiased assessment of the full range of SARS-CoV-2 translation products. Examination of SARS-CoV-2 translation as reflected by the diverse ribosome footprint libraries, revealed several unannotated novel ORFs. We detected in-frame internal ORFs lying within existing ORFs, resulting in N-terminally truncated product. These included relatively long truncated versions of canonical ORFs such as the one found in ORF6 (Figure 3A and Figure S6A) or very short truncated ORFs that likely serve an upstream ORF (uORF), like truncated ORF7a that might regulate ORF7b translation (Figure 3B, Figure S6B and Figure S6C). We also detected internal out of frame translations, that would yield entirely novel polypeptides, such as novel ORFs within ORF-N (97aa, Figure 3C and Figure S6D) and ORF-S (39aa, Figure 3D and Figure S6E) or short ORFs that likely serve as uORFs (Figure 3E and Figure S6F). Additionally, we observed a 13 amino acid extended ORF-M, in addition to the canonical ORF-M, which is predicted to start at the near cognate codon ATA (Figure 3F and Figure S6G and Figure S6H). Finally, we detected three distinct initiation sites at SARS-CoV-2 5’UTR. Two of these encode for uORFs that are located just upstream of ORF1a; the first initiating at an AUG (uORF1) and the second at a near cognate codon (uORF2, Figure 3G and Figure S6I). The third site is the most prominent peak in the ribosome profiling densities on the SARS-CoV-2 genome and is located on a CUG codon at position 59, just 10 nucleotides upstream the TRS-leader (Figure 3H and Figure S6J). The reads mapped to this site have a tight

length distribution characteristic of true ribosome protected fragments (Figure S7A). Due to its location upstream of the leader, footprints mapping to this site have the potential to be derived from all subgenomic RNAs as well as the genomic transcripts. We therefore analyzed the footprint distribution on genomic RNA and subgenomic junctions by analyzing footprints alignment to the different junctions. Surprisingly, even though the subgenomic RNAs compose the majority of transcripts, 99.5% of footprints that mapped to the CUG codon do not contain a junction and therefore originate from the genomic RNA (Figure S7B). On the genome, this initiation results in an extension of uORF1 (Figure 3H) but it is noteworthy that the occupancy at the CUG is much higher than the downstream translation signal, suggesting this peak might reflect a pause of the ribosome on the genomic RNA. Since translation starts at the 5'-cap and both genomic and subgenomic RNAs contain an identical 5' end, including this initiation site, the specificity of the CUG initiation to the genomic RNA indicates that the genomic RNA is somehow identified or engaged differently by the ribosomes.

The presence of the annotated ORF10 was recently put into question as almost no subgenomic reads were found for its corresponding transcript (Kim et al., 2020; Taiaroa et al., 2020). Although we also did not detect subgenomic RNA designated for ORF10 translation (Table S1), the ribosome footprints densities show ORF10 is translated (Figure 3I and Figure S6K). Translation levels were low but not negligible and according to ribosome densities, its expression is estimated to be 4-fold higher that of ORF1B. Interestingly, we detected two novel ORFs that are more highly expressed in this region, an upstream out of frame ORF that overlaps ORF10 initiation and an in-frame internal initiation that leads to a truncated ORF10 product. Further research is needed to delineate how ORF10, and possibly other ORFs in its vicinity, are translated.

To systematically define the SARS-CoV-2 translated ORFs using the ribosome profiling densities, we used PRICE, a computational method that is designated to predict overlapping ORFs and noncanonical translation initiation from ribosome profiling measurements (Erhard et al., 2018). After applying a minimal expression cutoff on the predictions, this approach identified 17 ORFs, these included 7 out of the 11 canonical translation initiations and 10 new viral ORFs. Visual inspection of the ribosome profiling data confirmed these identified ORFs and suggested additional 12 putative novel ORFs, some of which are presented above (Figure 3A, 3D, 3I and

Table S3). Overall, we identified 23 novel ORFs, on top of the 12 canonical viral ORFs, 56% of which initiate at AUG and the rest at near cognate codons.

The ribosome profiling data from 24 hpi revealed highly reproducible reads that align to SARS-CoV-2 transcripts but did not fit the expected translation profile and were globally not correlated with any sequence feature related to translation (Figure 4A). Importantly, these non-canonical profiles were specific for the virus as cellular transcripts presented the expected profiles, ruling out any technical issues with sample preparation (Figure S1 and Figure S2). Furthermore, in contrast to the footprints that mapped to cellular genes, fragments that mapped to the viral genome at 24 hpi were generally not affected by Harr or LTM treatments (Figure 4A and Figure S2) indicating that they were likely not protected by translating ribosomes. Importantly, when initiation sites of individual canonical viral genes were examined, an enhanced signal at the start codon was still observed and this signal was indeed affected by LTM treatment (Figure 4B), indicating there is also residual signal that reflects bona fide ribosome footprints. To further examine whether the protected fragments at 24 hpi reflect ribosome protection, we applied a fragment length organization similarity score (FLOSS) that measures the magnitude of disagreement between the footprint distribution on a given transcript and the footprints distribution on canonical CDSs (Ingolia et al., 2014). At 5 hpi protected fragments from SARS-CoV-2 transcripts scored well in these matrices and they did not differ from well-expressed human transcripts (Figure 4C). However, reads from 24 hpi could be clearly distinguished from cellular annotated coding sequences (Figure 4D). We conclude that the majority of protected fragments from the viral genome at 24 hpi are not generated by ribosome protection. It is likely that these fragments originate from protection by the N protein that wraps the viral genome into a helical nucleocapsid (Chen et al., 2007). This possibility is supported by the absence of such signal at 5 hpi and the specificity of the signal to the plus strand (99.7% of protected fragments at 24 hpi were mapped to the positive strand). Since the peaks we obtained were very reproducible, we examined their profiles. We identified the prominent peaks and found a consistent distance between adjacent peaks of approximately 33 bps (Figure S8). We reasoned this distance might reflect some periodicity related to the packing of the genome. Motif discovery algorithms failed to find an obvious sequence that defines these protected peaks, nonetheless downstream of the peak we detected a bias in nucleotide content towards higher G+C content.

Translation of viral proteins relies on the cellular translation machinery, and coronaviruses, like many other viruses, are known to cause host shutoff (Abernathy and Glaunsinger, 2015). In order to quantitatively evaluate if SARS-CoV-2 skews the translation machinery to preferentially translate viral transcripts, we compared the ratio of footprints to mRNAs for virus and host CDSs at 5 hpi and 24 hpi. Since at 24 hpi our ribosome densities were masked by a contaminant signal from non-translating genomes, for samples from this time point we used the footprints that were mapped to subgenomic RNA junctions (and therefore reflect bona fide transcripts) to estimate the true ribosome densities. At both 5 hpi and 24 hpi the virus translation efficiencies fall within the low range of most of the host genes (Figure 5A and 5B), indicating that viral transcripts are not preferentially translated during SARS-CoV-2 infection. Instead, during infection viral transcripts take over the mRNA pool, probably through massive transcription coupled to host induced RNA degradation (Huang et al., 2011; Kamitani et al., 2009).

In summary, in this study we delineate the translation landscape of SARS-CoV-2. Comprehensive mapping of the expressed ORFs is a prerequisite for the functional investigation of viral proteins and for deciphering viral-host interactions. An in-depth analysis of the ribosome profiling experiments revealed a highly complex landscape of translation products, including translation of 23 novel viral ORFs and illuminating the relative production of all canonical viral proteins. The new ORFs we have identified may serve as accessory proteins or as regulatory units controlling the balanced production of different viral proteins. Studies on the functional significance of these ORFs will deepen our understanding of SARS-CoV-2 and of coronaviruses in general. Overall, our work reveals the coding capacity of SARS-CoV-2 genome and highlights novel features, providing a rich resource for future functional studies.

Figure legends:

Figure 1. Ribosome profiling of SARS-CoV-2 infected cells**.**
(A) Vero E6 cells infected with SARS-CoV-2 were harvested at 5 or 24 hours post infection (hpi) for RNA-seq, and for ribosome profiling using lactimidomycin (LTM) and Harringtonine (Harr) treatments for mapping translation initiation or cycloheximide (CHX) treatment to map overall translation. **(B)** Metagene analysis of read densities at the 5' and the 3' regions of cellular protein coding genes as measured by the different ribosome profiling approaches and RNA-seq at 5 hpi (one of two replicates is presented). The X axis shows the nucleotide position relative to the start or the stop codons. The ribosome densities are shown with different colors indicating the three frames relative to the main ORF (red, frame 0; black, frame +1; grey, frame +2). **(C)** Metagene analysis of the 5' region, as presented in A, for viral coding genes at 5 hpi (**D**) Ribosome densities are presented for ORF3a. Different colors indicating the three phases relative to the translated frame (red, frame 0; black, frame +1; grey, frame +2).

Figure 2. Expression level of canonical viral genes
(A) RNA-Seq (green) and Ribo-Seq CHX (red) densities at 5 hpi on the SARS-CoV-2 genome. Read densities are plotted on a log scale to cover the wide range in expression across the genome. The lower panel presents SARS-CoV-2 genome organization with the canonical viral ORFs. **(B)** Viral coding genes relative abundance was estimated by counting the ribosome densities on each ORF. ORFs are ordered based on their genomic location **(C)** Viral ORFs expression as calculated from ribosome densities plotted on a log scale to cover the wide range in expression. Values were normalized to ORF length and library size. **(D)** Relative abundance of the different viral transcripts relative to the ribosome densities of each viral ORF at 5 hpi. Transcript abundance were estimated by counting the reads that span the junctions of the corresponding RNA or for ORF1a and ORF1b the genomic RNA abundance, normalized to junctions count. **(E)** Scatter plot presenting the abundance of viral reads that span canonical leader dependent junctions (red), non-canonical leader dependent junctions (green) and non-canonical leader independent junctions (green) at 5 and 24 hpi. **(F)** Ribosome profiling (CHX) and RNA densities over the deletion in the S protein. Lower panels present the sequence in the region and the translation of the WT and of the 7aa deleted version of the S protein.

Figure 3. Ribosome densities reveal novel viral coding regions.

(A-I) Ribosome density profiles of CHX, Harr and LTM samples at 5 hpi. Densities are shown with different colors indicating the three frames relative to the main ORF in each figure (red, frame 0; black, frame +1; grey, frame +2). One out of two replicates is presented. Rectangles indicate the canonical and novel ORFs and ORFs starting in a near cognate codon are labeled with stripes. **(A)** In frame internal initiation within ORF6 generating a truncated product, **(B)** In frame internal initiation within ORF7a, **(C)** Out of frame internal initiations within ORF-N, **(D)** Out of frame internal initiations within ORF-S, **(E)** Out of frame internal initiation within ORF-M, **(F)** an extended version of ORF-M (reads marking ORF-M initiation were cut to fit the scale), **(G)** two uORFs embedded in ORF1a 5'UTR **(H)** non canonical CUG initiation upstream of the TRS leader **(I)** uORF that overlap ORF10 initiation and in frame internal initiation generating truncated ORF10 product.

Figure 4. Protected fragments at 24 hpi likely reflect genome packaging

(A) Metagene analysis of the expression profile as measured by the different ribosome profiling approaches at the 5' region of viral coding genes of samples harvested at 24 hpi (one of two replicates is presented). The X axis shows the nucleotide position relative to the start. The footprints densities are shown with different colors indicating the three frames relative to the main ORF (red, frame 0; black, frame +1; frame, frame +2). **(B)** Footprint densities are presented for ORF3a, different colors indicate the three frames (red, frame 0; black, frame +1; grey, frame +2) **(C and D)** Fragment length organization similarity score (FLOSS) analysis for cellular coding regions and for SARS-CoV2 ORFs at 5 hpi **(C)** and 24 hpi **(D)**.

Figure 5. Comparison of host and virus translation

(A and B) Relative transcript abundance versus ribosome densities for each host and viral ORF at 5 hpi **(A)** and 24 hpi **(B)**. Transcript abundance was estimated by counting the reads that span the corresponding junction (only the most abundant viral transcripts are presented) and footprint densities were calculated from the CHX sample. For ribo-seq viral reads from 24 hpi, only reads that were mapped to junctions were used to avoid non-ribosome footprints.

Tables legend:

Table S1. Junctions sites detected from junction spanning reads.

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

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

Table S2. Junctions sites uniquely detected in our samples

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

The genomic coordinates in the "5′ site" and "3′ site" point to the 3′-most and the 5′-most nucleotides that survive the recombination event, respectively. "Gap" is the size of the deletion. "Leader" true value indicates the junction is TRS-leader dependent. "canonical" true value indicates the junction supports the expression of a canonical ORF, "Kim_count" is the number of the junction-spanning reads that support the recombination event identified by Kim et al. "ORF" the name of an ORF that shares the start codon position with the recombination product based on Kim et al. "mrna_05hr_1", "mrna_05hr_2", "mrna_24hr_1" and "mrna_24hr_2" the number of

the junction-spanning reads that support the recombination event in each of our RNA samples based on STAR-aligner. “fp_chx_05hr_1”, “fp_chx_05hr_2”, “fp_chx_24hr_1” and “fp_chx_24hr_2” the number of the junction-spanning reads that support the recombination event in each of our footprints CHX samples. “sum_fp” the sum of all footprints counts. “sum_mRNA” the sum of all RNA counts. “star_sum” the sum of number of the junction-spanning reads in all samples.

Table S3. Novel SARS-CoV-2 ORFs that have been identified in our study

This table lists all the SARS-CoV2 translated ORFs identified in this study. “Name” for each ORF, “description”, “predicted method”the method that was used to detect the ORF, the “start position” and “ end position” in SAS-CoV-2 genome, the nature of the ‘start codon”, “size(aa)” and “sequence.

Acknowledgements

We thank Stern-Ginossar lab members, Igor Ulitsky and Schraga Schwartz for providing valuable feedback, to Miri shnayder and Igor Ulitsky for technical assistance. This study was supported by the Ben B. and Joyce E. Eisenberg Foundation. Work in the Stern-Ginossar lab is supported by a European Research Council starting grant (StG-2014-638142) and by the Israel Science Foundation (ISF) grant no. 1526/18. S.W-G. is the recipient of the Human Frontier Science Program fellowship (LT-000396/2018), EMBO non-stipendiary Long-Term Fellowship (ALTF 883-2017), the Gruss-Lipper Postdoctoral Fellowship, the Zuckerman STEM Leadership Program Fellowship and the Rothschild Postdoctoral Fellowship. N.S-G is an incumbent of the Skirball Career Development Chair in New Scientists and is a member of the European Molecular Biology Organization (EMBO) Young Investigator Program. The authors declare no competing interests.

Author contributions

Y.F., O.M. and N.S-G. conceptualization. O.M. experiments. Y.F., A.N. and S.W-G. data analysis. Y.Y-R., H.T., H.A., S.M., S.W., T.I. and N.P. work with SAR-CoV-2. Y.F., O.M., A.N., M.S. and N.S-G. interpreted data. M.S. and N.S.-G. wrote the manuscript with contribution from all other authors.

Material and methods

Cells and viruses

Vero C1008 (Vero E6) (ATCC CRL-1586™) were cultured in DMEM supplemented with 10% fetal bovine serum (FBS), MEM non-essential amino acids, 2mM L-Glutamine, 100Units/ml Penicillin, 0.1mg/ml streptomycin, 12.5Units/ml Nystatin (Biological Industries, Israel). Monolayers were washed once with DMEM Eagles medium without FBS and infected with SARS-CoV-2 virus, at a multiplicity of infection (MOI) of 0.2, and cultured in MEM Eagles medium supplemented with 2% fetal bovine serum, and MEM non-essential amino acids, L glutamine and penicillin-streptomycin-Nystatin at 37°C, 5% CO2. SARS-CoV-2 (GISAID Acc. No. EPI_ISL_406862), was kindly provided by Bundeswehr Institute of Microbiology, Munich, Germany. The virus was propagated (4 passages) and tittered on Vero E6 cells. Infected cells were harvested at 5 and 24 hpi as described below. Handling and working with SARS-CoV-2 virus was conducted in a BSL3 facility in accordance with the biosafety guidelines of the Israel Institute for Biological Research. The Institutional Biosafety Committee of Weizmann Institute approved the protocol used in these studies.

Preparation of ribosome profiling and RNA sequencing samples

For RNA-seq, cells were harvested with Tri-Reagent (Sigma-Aldrich), total RNA was extracted, and poly-A selection was performed using Dynabeads mRNA DIRECT Purification Kit (Invitrogen) mRNA sample was subjected to DNAseI treatment and 3' dephpsphorylation using FastAP Thermosensitive Alkaline Phosphatase (Thermo Scientific) and T4 PNK (NEB) followed by 3' adaptor ligation using T4 ligase (NEB). The ligated products used for reverse transcription with SSIII (Invitrogen) for first strand cDNA synthesis. The cDNA products were 3' ligated with a second adaptor using T4 ligase and amplified for 8 cycles in a PCR for final library products of 200-300bp. For Ribo-seq libraries, cells were treated with either 50µM lactimidomycin (LTM) for 30 minutes or 2µg/mL Harringtonine (Harr) for 5 minutes, for translation initiation libraries (LTM and Harr libraries correspondingly), or left untreated for the translation elongation libraries (cycloheximide [CHX] library). All three samples were subsequently treated with 100µg/mL CHX for 1 minute. Cells were then placed on ice, washed twice with PBS containing 100µg/mL CHX, scraped from the flasks, pelleted and lysed with lysis buffer (1% triton in

20mM Tris 7.5, 150mM NaCl, 5mM $MgCl_2$, 1mM dithiothreitol supplemented with 10 U/ml Turbo DNase and 100µg/ml cycloheximide). After lysis samples stood on ice for 2h and subsequent Ribo-seq library generation was performed as previously described (Finkel et al., 2020).

Sequence alignment, normalization and metagene analysis

Sequencing reads were aligned as previously described (Tirosh et al., 2015). Briefly, linker (CTGTAGGCACCATCAAT) and poly-A sequences were removed and the remaining reads were aligned to the Chlorosebus sabaeus genome (ENSEMBL release 99) and to the SARS-Cov-2 genomes [Genebank NC_045512.2 with 3 changes to match the used strain (BetaCoV/Germany/BavPat1/2020 EPI_ISL_406862), 241:C−>T, 3037:C−>T, 23403:A−>G] using Bowtie v1.1.2 (Langmead et al., 2009) with maximum two mismatches per read. Reads that were not aligned to the genome were aligned to the transcriptome of Chlorosebus sabaeus (ENSEMBL) and to SARS-CoV-2 junctions that were recently annotated (Kim et al., 2020). Novel junctions were mapped using STAR 2.5.3a aligner (Dobin et al., 2013), with running flags as suggested at Kim et. al., to overcome filtering of non-canonical junctions. Reads aligned to multiple locations were discarded. Junctions with 5' break sites mapped to genomic location 55-85 were assigned as leader junctions. Matching of leader junctions to ORFs, and categorization of junctions as canonical or non-canonical, was obtained from Kim et. al. Supplementary table 3, or was done manually for strong novel junctions.

For the metagene analysis only genes with more than 50 reads were used. For each gene normalization was done to its maximum signal and each position was normalized to the number of genes contributing to the position. In the 24hr samples, normalization for each gene was done to its maximum signal within the presented region (5' region of the gene).

For comparing transcript expression level, mRNA and footprint counts were normalized to units of RPKM in order to normalize for gene length and for sequencing depth, based on the total number reads for both the host and the virus. The counts of junctions were done according to the STAR number of uniquely mapped reads crossing the junction. All junctions were included regardless of the strand. The de-accumulation of RPKM for the sub genomic RNA molecules was done by subtracting the RPKM of a gene from the RPKM of the gene located just upstream

of it in the genome. For genes that the de-accumulated value was negative (ORF6 and ORF10), these values were ignored.

The estimation of the ribosomal fraction of CHX reads from the 24 hpi samples was done as follows: for all CHX samples, the ratio of the RPKM of ORF1a to the total number of leader canonical junctions was calculated. The ratio between the average of this value for the two 24 hpi replicates and for the two 5hr replicates was taken as the factor to divide the RPKM of all viral genes at 24hr to get an estimation for the portion of the ribosomal reads.

For calculation of the distance between footprints peaks on the viral genome from 24 hpi, peaks were defined as those present in all 6 samples, using an ad hoc script looking at the ratio of the coverage of a sliding window of 9 nucleotides relative to the 2 flanking windows of the same size. In Figure 2D, ORF1a and 1b levels were estimated from the correlation of the junctions to the deaccumlated RPKM shown in Figure S4.

Prediction of translation initiation sites

Translation initiation sites were predicted using PRICE (Erhard et al., 2018). The PRICE algorithm requires annotated ORFs in order to evaluate the noise level of the detected codons, which in turn is being used to compute the p-value for ORF identification. Thus, it does not perform well on reference sequences with small numbers of annotated ORFs such as SARS-CoV-2. In order to provide a comprehensive dataset of annotated ORFs from the same experiment, reads were mapped to a fasta file containing chromosome 20 of Chlorocebus sabaeus (downloaded from ensembl: ftp://ftp.ensembl.org/pub/release-99/fasta/chlorocebus_sabaeus/dna/) and the genomic sequence of SARS-CoV-2 (Refseq NC_045512.2). A gtf file with the annotations of Chlorocebus sabaeus and SARS-CoV-2 genomes was constructed and provided as the annotation file when running PRICE. For technical reasons, the annotation of the first coding sequence (CDS) of the two CDSs in the “ORF1ab” gene was deleted since having two CDSs encoded from a single gene was not permitted by the PRICE. The predictions were further filtered to include only ORFs with at least 100 reads at the initiation site in the LTM samples. ORFs were then defined by extending each initiating codon to the next in-frame stop codon. Additional ORFs that were not recognized by the trained model but presented reproducible translation in manual inspection were added to the final ORF list (Table

S3). ORFs were manually identified as such if they had reproducible initiation peaks in the CHX libraries that were enhanced in the LTM and Harr libraries, and exhibited increased CHX signal in the correct reading-frame along the coding region.

Data availability

All next-generation sequencing data files were deposited in Gene Expression Omnibus under accession number GSE149973.
All the data generated in this study can be accessed through a UCSC browser session: http://genome.ucsc.edu/s/aharonn/CoV2%2DTranslation

References

Abernathy, E., and Glaunsinger, B. (2015). Emerging roles for RNA degradation in viral replication and antiviral defense. Virology *479–480*, 600–608.

Bercovich-Kinori, A., Tai, J., Gelbart, I.A., Shitrit, A., Ben-Moshe, S., Drori, Y., Itzkovitz, S., Mandelboim, M., and Stern-Ginossar, N. (2016). A systematic view on influenza induced host shutoff. Elife *5*, e18311.

Bojkova, D., Klann, K., Koch, B., Widera, M., Krause, D., Ciesek, S., and Cinatl, J. (2020). SARS-CoV-2 infected host cell proteomics reveal potential therapy targets.

Chen, C.Y., Chang, C. ke, Chang, Y.W., Sue, S.C., Bai, H.I., Riang, L., Hsiao, C.D., and Huang, T. huang (2007). Structure of the SARS Coronavirus Nucleocapsid Protein RNA-binding Dimerization Domain Suggests a Mechanism for Helical Packaging of Viral RNA. J. Mol. Biol. *368*, 1075–1086.

Davidson, A.D., Williamson, M.K., Lewis, S., Shoemark, D., Carroll, M.W., Heesom, K., Zambon, M., Ellis, J., Lewis, P.A., Hiscox, J.A., et al. (2020). Characterisation of the transcriptome and proteome of SARS-CoV-2 using direct RNA sequencing and tandem mass spectrometry reveals evidence for a cell passage induced in-frame deletion in the spike glycoprotein that removes the furin-like cleavage site. BioRxiv 2020.03.22.002204.

Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T.R. (2013). STAR: ultrafast universal RNA-seq aligner. Bioinformatics *29*, 15–21.

Erhard, F., Halenius, A., Zimmermann, C., L'Hernault, A., Kowalewski, D.J., Weekes, M.P., Stevanovic, S., Zimmer, R., and Dölken, L. (2018). Improved Ribo-seq enables identification of cryptic translation events. Nat. Methods *15*, 363–366.

Finkel, Y., Schmiedel, D., Tai-Schmiedel, J., Nachshon, A., Winkler, R., Dobesova, M., Schwartz, M., Mandelboim, O., and Stern-Ginossar, N. (2020). Comprehensive annotations of human herpesvirus 6A and 6B genomes reveal novel and conserved genomic features. Elife *9*, e50960.

Huang, C., Lokugamage, K.G., Rozovics, J.M., Narayanan, K., Semler, B.L., and Makino, S. (2011). SARS coronavirus nsp1 protein induces template-dependent endonucleolytic cleavage of

mRNAs: Viral mRNAs are resistant to nsp1-induced RNA cleavage. PLoS Pathog. *7*, e1002433.

Ingolia, N.T., Brar, G.A., Stern-Ginossar, N., Harris, M.S., Talhouarne, G.J.S., Jackson, S.E., Wills, M.R., and Weissman, J.S. (2014). Ribosome Profiling Reveals Pervasive Translation Outside of Annotated Protein-Coding Genes. Cell Rep. *8*, 1365–1379.

Irigoyen, N., Firth, A.E., Jones, J.D., Chung, B.Y.W., Siddell, S.G., and Brierley, I. (2016). High-Resolution Analysis of Coronavirus Gene Expression by RNA Sequencing and Ribosome Profiling. PLoS Pathog. *12*, 1005473.

Kamitani, W., Huang, C., Narayanan, K., Lokugamage, K.G., and Makino, S. (2009). A two-pronged strategy to suppress host protein synthesis by SARS coronavirus Nsp1 protein. Nat. Struct. Mol. Biol. *16*, 1134–1140.

Kim, D., Lee, J.-Y., Yang, J.-S., Kim, J.W., Kim, V.N., and Chang, H. (2020). The architecture of SARS-CoV-2 transcriptome. Cell *S0092-8674*, 30406–2.

Lai, M.M., and Stohlman, S.A. (1981). Comparative analysis of RNA genomes of mouse hepatitis viruses. J. Virol. *38*, 661–670.

Langmead, B., Trapnell, C., Pop, M., and Salzberg, S.L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. *10*, R25.

Liu, Z., Zheng, H., Yuan, R., Li, M., Lin, H., Peng, J., Xiong, Q., Sun, J., Li, B., Wu, J., et al. (2020). Identification of a common deletion in the spike protein of SARS-CoV-2. BioRxiv 2020.03.31.015941.

Ogando, N.S., Dalebout, T.J., Zevenhoven-Dobbe, J.C., Limpens, R.W., Meer, Y. van der, Caly, L., Druce, J., Vries, J.J.C. de, Kikkert, M., Bárcena, M., et al. (2020). SARS-coronavirus-2 replication in Vero E6 cells: replication kinetics, rapid adaptation and cytopathology. BioRxiv 2020.04.20.049924.

Plant, E.P., Rakauskaite, R., Taylor, D.R., and Dinman, J.D. (2010). Achieving a golden mean: mechanisms by which coronaviruses ensure synthesis of the correct stoichiometric ratios of viral proteins. J. Virol. *84*, 4330–4340.

Schaecher, S.R., Mackenzie, J.M., and Pekosz, A. (2007). The ORF7b protein of severe acute respiratory syndrome coronavirus (SARS-CoV) is expressed in virus-infected cells and

incorporated into SARS-CoV particles. J. Virol. *81*, 718–731.

Snijder, E.J., Decroly, E., and Ziebuhr, J. (2016). The Nonstructural Proteins Directing Coronavirus RNA Synthesis and Processing. In Advances in Virus Research, (Academic Press Inc.), pp. 59–126.

Sola, I., Almazán, F., Zúñiga, S., and Enjuanes, L. (2015). Continuous and Discontinuous RNA Synthesis in Coronaviruses. Annu. Rev. Virol. *2*, 265–288.

Stadler, K., Masignani, V., Eickmann, M., Becker, S., Abrignani, S., Klenk, H.D., and Rappuoli, R. (2003). SARS — beginning to understand a new virus. Nat. Rev. Microbiol. *1*, 209–218.

Stern-Ginossar, N., and Ingolia, N.T. (2015). Ribosome Profiling as a Tool to Decipher Viral Complexity. Annu. Rev. Virol. *2*, 335–349.

Stern-Ginossar, N., Weisburd, B., Michalski, A., Le, V.T.K., Hein, M.Y., Huang, S.X., Ma, M., Shen, B., Qian, S.B., Hengel, H., et al. (2012). Decoding human cytomegalovirus. Science (80-.). *338*, 1088–1093.

Taiaroa, G., Rawlinson, D., Featherstone, L., Pitt, M., Caly, L., Druce, J., Purcell, D., Harty, L., Tran, T., Roberts, J., et al. (2020). Direct RNA sequencing and early evolution of SARS-CoV-2. BioRxiv 2020.03.05.976167.

Tirosh, O., Cohen, Y., Shitrit, A., Shani, O., Le-Trilling, V.T.K., Trilling, M., Friedlander, G., Tanenbaum, M., and Stern-Ginossar, N. (2015). The Transcription and Translation Landscapes during Human Cytomegalovirus Infection Reveal Novel Host-Pathogen Interactions. PLoS Pathog. *11*, e1005288.

Whisnant, A.W., Jürges, C.S., Hennig, T., Wyler, E., Prusty, B., Rutkowski, A.J., L'hernault, A., Djakovic, L., Göbel, M., Döring, K., et al. (2020). Integrative functional genomics decodes herpes simplex virus 1. Nat. Commun. *11*, 2038.

Yogo, Y., Hirano, N., Hino, S., Shibuta, H., and Matumoto, M. (1977). Polyadenylate in the virion RNA of mouse hepatitis virus. J. Biochem. *82*, 1103-1108.

Zhou, P., Yang, X. Lou, Wang, X.G., Hu, B., Zhang, L., Zhang, W., Si, H.R., Zhu, Y., Li, B., Huang, C.L., et al. (2020). A pneumonia outbreak associated with a new coronavirus of probable bat origin. Nature *579*, 270–273.

Zhu, N., Zhang, D., Wang, W., Li, X., Yang, B., Song, J., Zhao, X., Huang, B., Shi, W., Lu, R., et al. (2020). A novel coronavirus from patients with pneumonia in China, 2019. N. Engl. J. Med. *382*, 727–733.

Figure 1

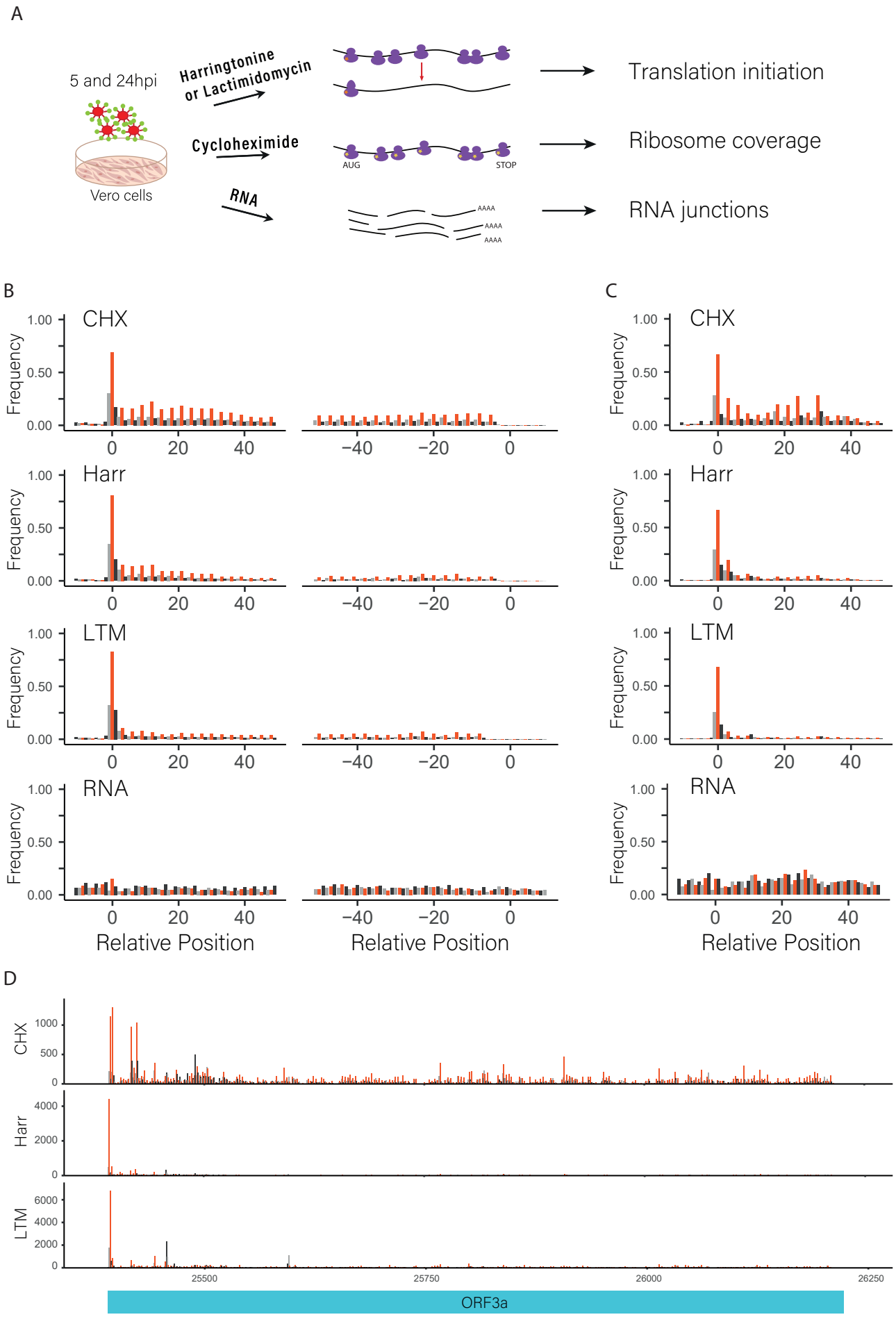
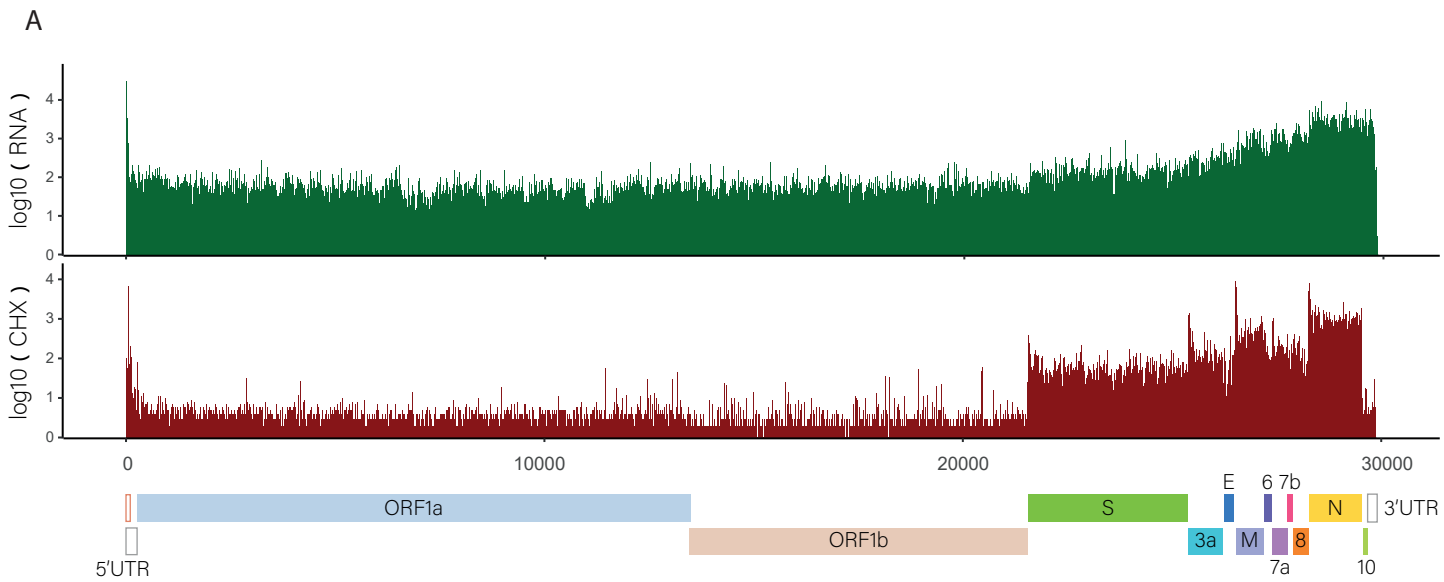
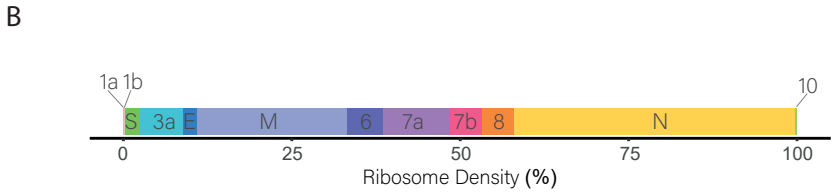
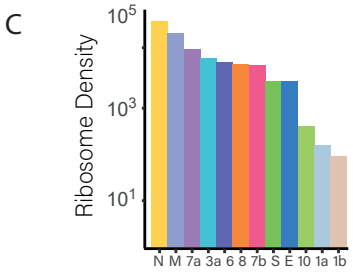
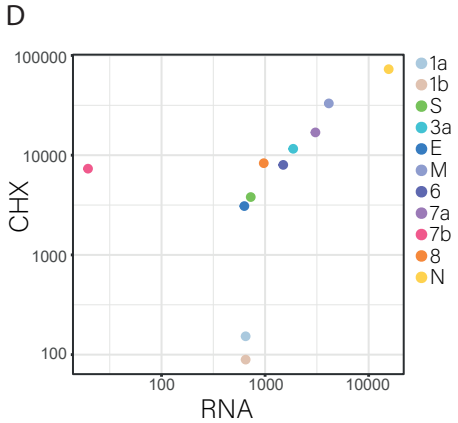

A
5 and 24hpi
Vero cells
Harringtonine or Lactimidomycin
Cycloheximide
RNA
AUG
STOP
AAAA
Translation initiation
Ribosome coverage
RNA junctions
B
C
CHX
Harr
LTM
RNA
Frequency
Relative Position
D
CHX
Harr
LTM
ORF3a

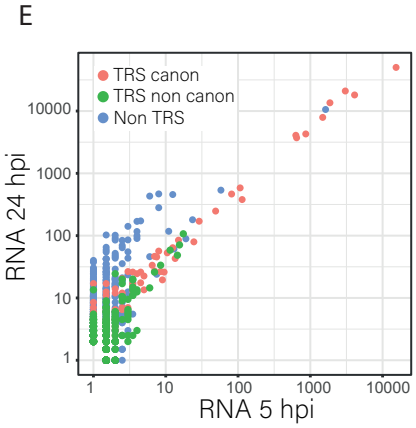
Figure 2


A
log10 (RNA)
log10 (CHX)
0
10000
20000
30000
ORF1a
ORF1b
S
E
6 7b
N
3'UTR
3a
M
8
7a
10
5'UTR


B
1a 1b
S
3a
E
M
6
7a
7b
8
N
10
0
25
50
75
100
Ribosome Density (%)


C
Ribosome Density
N M 7a 3a 6 8 7b S E 10 1a 1b


D
CHX
RNA
1a
1b
S
3a
E
M
6
7a
7b
8
N


E
TRS canon
TRS non canon
Non TRS
RNA 24 hpi
RNA 5 hpi

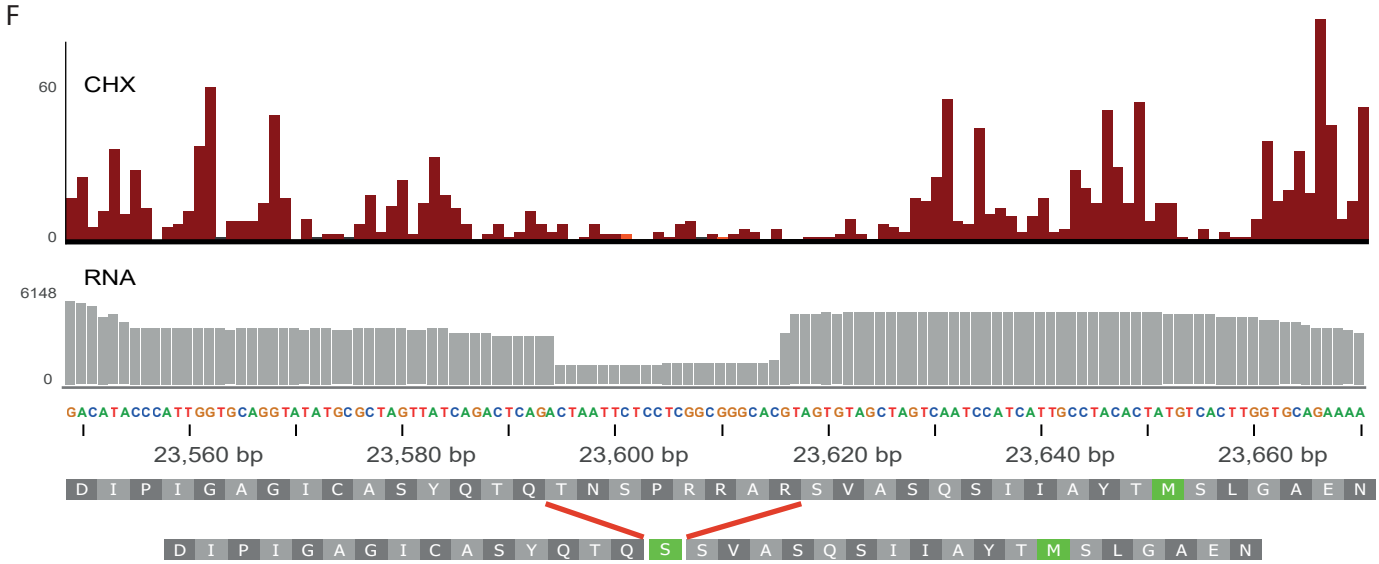

F
CHX
RNA
6148
GACATACCCATTGGTGCAGGTATATGCGCTAGTTATCAGACTCAGACTAATTCTCCTCGGCGGGCACGTAGTGTAGCTAGTCAATCCATCATTGCCTACACTATGTCACTTGGTGCAGAAAA
23,560 bp
23,580 bp
23,600 bp
23,620 bp
23,640 bp
23,660 bp
D I P I G A G I C A S Y Q T Q T N S P R R A R S V A S Q S I I A Y T M S L G A E N
D I P I G A G I C A S Y Q T Q S S V A S Q S I I A Y T M S L G A E N

Figure 3

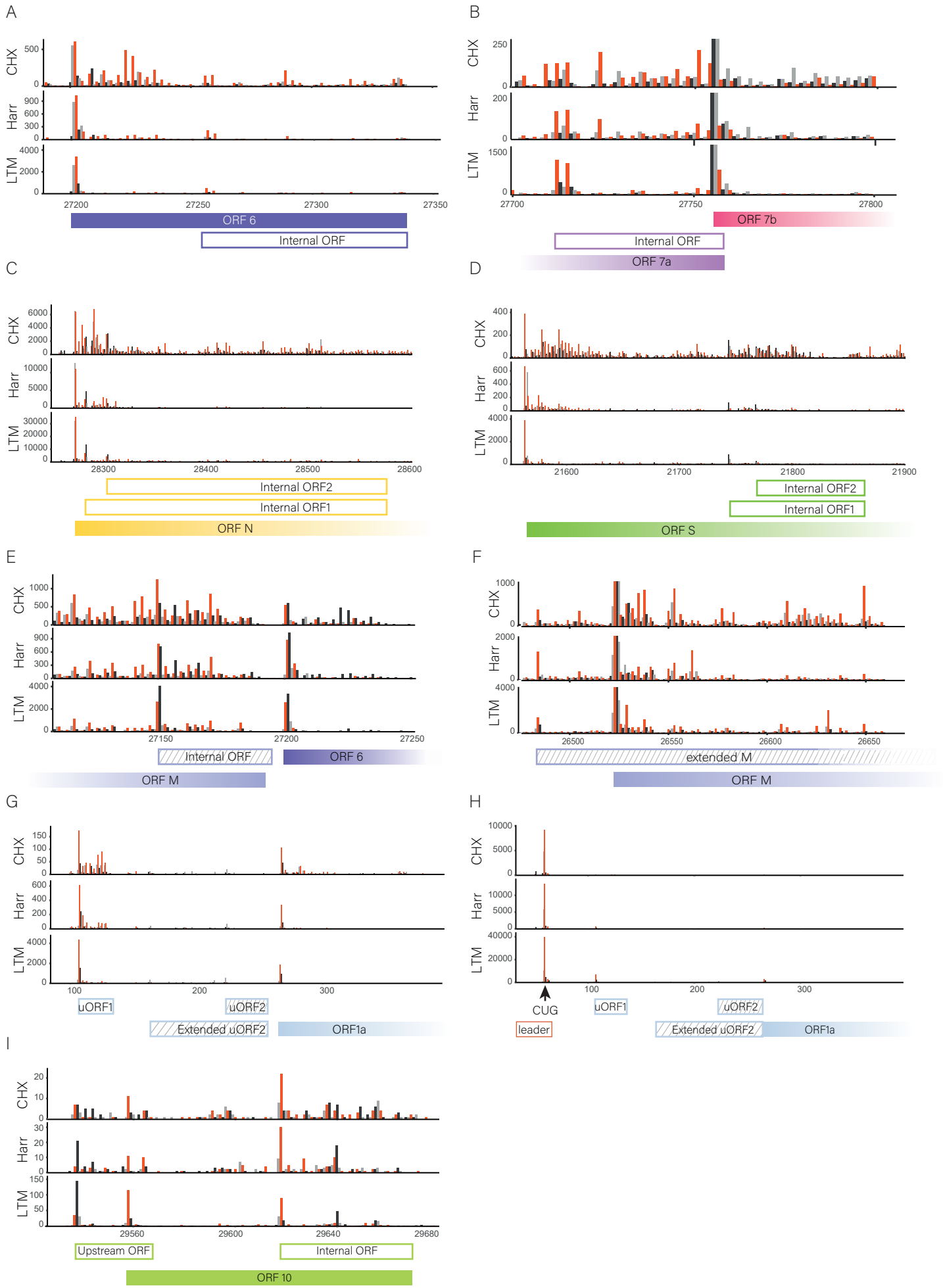

A
CHX
Harr
LTM
ORF 6
Internal ORF
B
ORF 7b
Internal ORF
ORF 7a
C
Internal ORF2
Internal ORF1
ORF N
D
Internal ORF2
Internal ORF1
ORF S
E
Internal ORF
ORF 6
ORF M
F
extended M
ORF M
G
uORF1
uORF2
Extended uORF2
ORF1a
H
CUG
leader
uORF1
uORF2
Extended uORF2
ORF1a
I
Upstream ORF
Internal ORF
ORF 10

Figure 4

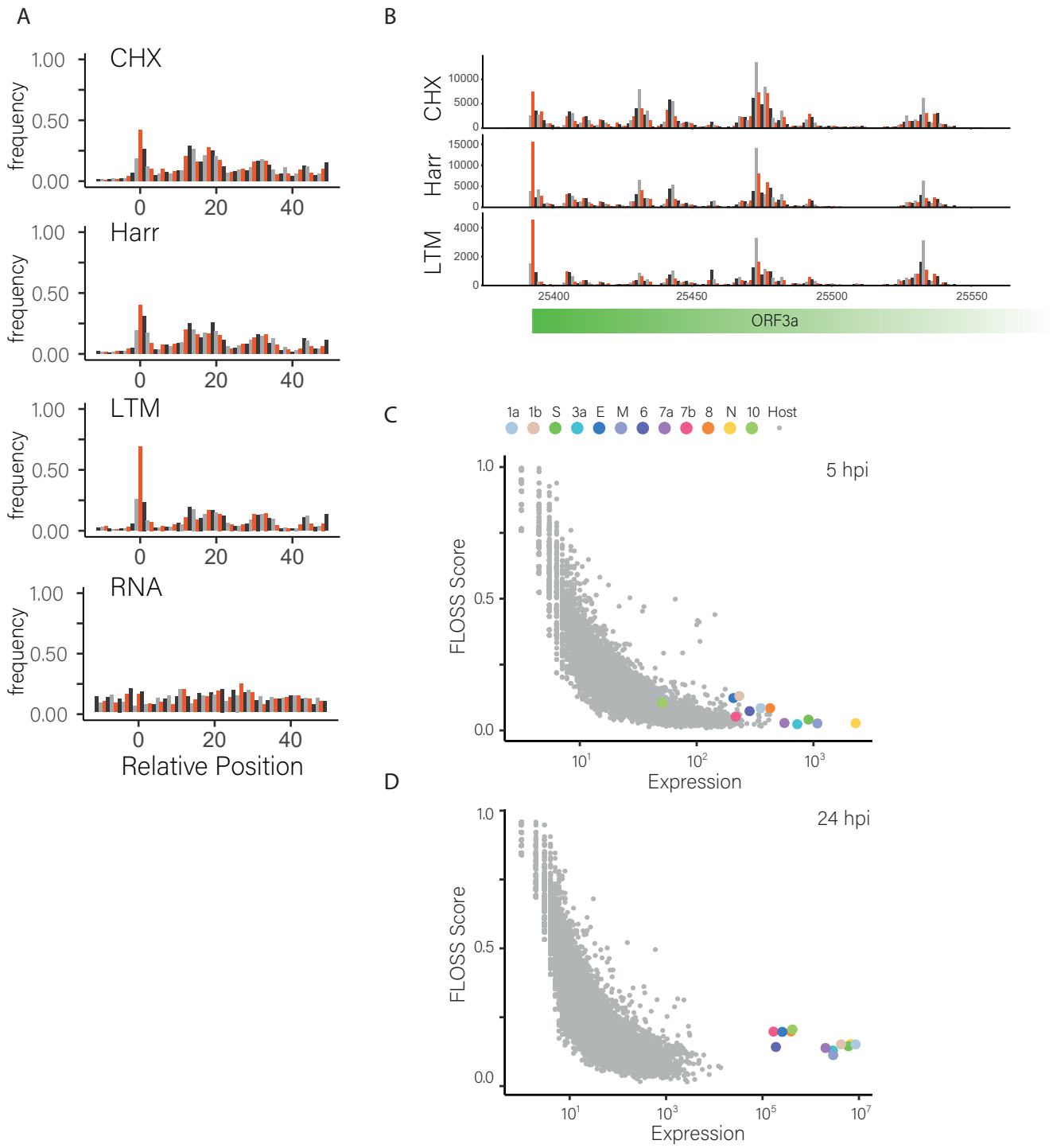
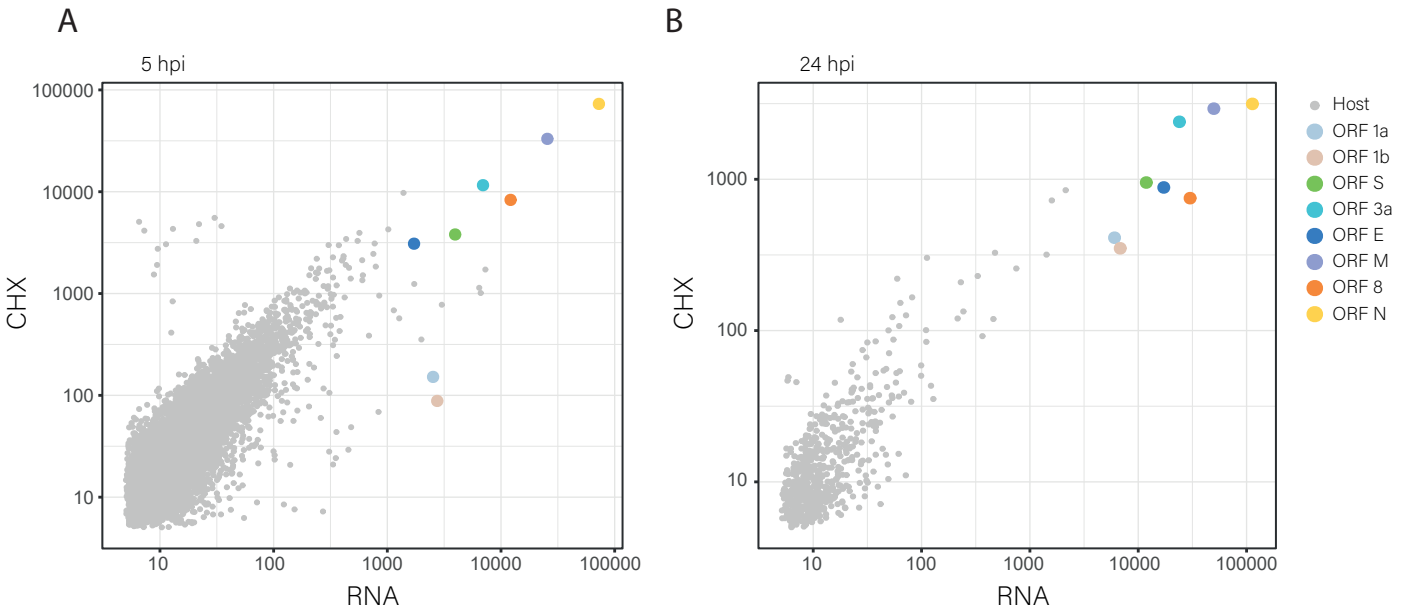

A
CHX
Harr
LTM
RNA
frequency
1.00
0.50
0.00
0
20
40
Relative Position
B
CHX
Harr
LTM
10000
5000
15000
4000
2000
25400
25450
25500
25550
ORF3a
C
1a 1b S 3a E M 6 7a 7b 8 N 10 Host
5 hpi
FLOSS Score
1.0
0.5
0.0
10^1
10^2
10^3
Expression
D
24 hpi
10^5
10^7

Figure 5


A
5 hpi
CHX
RNA
100000
10000
1000
100
10
B
24 hpi
Host
ORF 1a
ORF 1b
ORF S
ORF 3a
ORF E
ORF M
ORF 8
ORF N