RNA-GPS Predicts SARS-CoV-2 RNA Residency to Host Mitochondria and Nucleolus

Kevin E. Wu, Furqan M. Fazal, Kevin R. Parker, James Zou, Howard Y. Chang

Summary

Initial Submission:	Received Apr. 29, 2020 Preprint: https://doi.org/10.1101/2020.04.28.065201 *Deposited on bioRxiv, Apr. 28, 2020* Scientific editor: Quincey Justman, Ph.D.
Editorial Decision after first round of review:	Revision invited May 10, 2020 *Major changes anticipated* Revision received May 20, 2020
Editorial Decision after second round of review:	Accepted Jun. 17, 2020

Editor's View

Choosing which COVID-19 papers to review has been challenging, particularly when they beg the question, “Is this too thin a slice of the salami?” Under more normal circumstances, or if it had focused on a different virus, this manuscript by Wu et al. would have been the starting point of a larger study that included in vivo validation of the computational results presented here. But circumstances are not normal, and this paper is focused in SARS-CoV-2. More importantly, though, I found its computational predictions profoundly interesting.

Profoundly interesting computational predictions demand in vivo follow-up, now or in the future. The COVID pandemic precludes in vivo follow-up now. Under this circumstance, my job as an editor is to make sure that the experimentalists who will do the follow-up experiments in the future will not be wasting their time. That is, we need to be confident in the computational predictions: either they’re likely to be right or likely to be wrong, but wrong for biologically interesting and important reasons that are themselves worthy of study. What we want to filter out is computational predictions that are wrong for trivial reasons: for example, signatures are actually weak no matter how they might look in visualizations, assumptions aren't robust or biologically reasonable, work that's not done to a high technical standard, or importantly in this case, work that's likely to be subject to misinterpretation.

The review process for this paper was expedited, as all COVID-19 papers are, but it was also quite complicated. I will summarize it here, although key aspects of this process took place over Zoom and in confidential emails.

Background. Wu et al., out of Howard Chang's lab, relies heavily on Fazal et al., published in Cell in 2019 https://doi.org/10.1016/j.cell.2019.05.027, also from Howard Chang's lab in collaboration with Alice Ting's lab. I will outline Fazal et al. very briefly here, although my description should not be taken in lieu of reading the paper itself. The Ting lab had previously developed and refined APEX technology; APEX2 (as used in Fazal et al.) is an enzyme engineered to harness ascorbate peroxidase chemistry in a controlled manner. Specifically, APEX2 covalently links biotin to macromolecules in close physical proximity to the APEX2 enzyme within a very narrow window of time, defined by H_2O_2 treatment. The biotin tag can then be used as it is in conventional biochemistry: to pull-down biotin-labeled macromolecules on streptavidin beads for subsequent analysis. Fazal et al. tagged APEX2 with well-characterized sequences known to localize proteins to distinct subcellular compartments, generating handful of constructs, and then used APEX chemistry to label RNAs in HEK293T cells. After labeling, mRNAs which were proximal to the handful of tagged APEX2s could then be pulled down and identified by sequencing. Over 3000 mRNAs were characterized in this way and assigned one of nine subcellular localizations (e.g. mitochondrial matrix, nucleolus, etc) based on which tagged APEX2 construct labeled them. Another paper from the same group demonstrated how a computational approach called RNA-GPS can learn the mRNA sequence determinants of the assigned localizations (doi:10.1261/rna.074161.119) to predict the localization of other mRNAs.

Review process. The present paper, Wu et al., uses RNA-GPS to predict where SARS-CoV-2 mRNAs are likely to reside in the cell. The headline location is the mitochondria. Reviewer 1 rightly understood that this conclusion is exquisitely sensitive to quality of the original Fazal et al. data ("garbage in, garbage out"). When s/he went back to the original Fazal et al. data, s/he found Figure S3H and the following excerpt from the STAR Methods (note highlights):

> **MITO APEX-seq extended analysis**
> For the MITO APEX-seq, we obtained strong enrichment ($\log_2$fold-change > 2.9) of the 13 MT mRNAs and 2 MT rRNAs in the targets (i.e., labeled libraries) relative to unlabeled controls. These 15 genes made up > 50% of reads in the MITO APEX-seq labeled samples. In addition to the 15 expected mitochondrial RNAs, we also recovered ~400 transcripts that were moderately enriched ($\log_2$fold- change > 0.75), some of which are known mitochondrial pseudogenes. To rule out that this labeling was not because the biotin-phenoxy radical, generated during the labeling experiment, was escaping from the mitochondrial matrix we confirmed that OMM APEX-seq enriched transcripts (> 1000-enriched transcripts) showed no enrichment (average $\log_2$fold-change ~0) in the MITO APEX-seq samples. While we do not believe these transcripts to be present in the mitochondrial matrix (Mercer et al., 2011), attempts to confirm the localization by FISH were not successful. We hypothesize two explanations for the observations: (1) Due to the large perturbation introduced by APEX-seq labeling, the DESeq2 analysis does not perform

properly; or (2) There is some small background labeling by Cox4-APEX (i.e., MITO-APEX) as the protein makes it way from the cytosol, where it is translated, to the mitochondrial matrix.

So, from the perspective of gene identity (i.e. not labeled transcript abundance) is it fair to call MITO APEX-seq genes mitochondrial, when apparently more than 400 are not? Reviewer 1, rightly, called foul.

This had two major effects on the paper, which I'll summarize briefly. First, the authors reconsidered their computation and investigated whether the ~400 transcripts dominated the output of RNA-GPS under a variety of circumstances. It did not. Second, we reconsidered the presentation of the paper and thought carefully about Fazal et al. We recognized that "MITO APEX2" is more accurately described as APEX-2 appended to the N-terminus of COX4, a component of mitochondrial electron transport chain that is encoded by the nuclear genome, translated in the cytoplasm, and then transported to the mitochondria. As such, the signal that is called "MITO" actually represents the entire lifecycle of COX4 and its subcellular localization throughout. This renders the original Fazal et al. dataset less simple, but definitely not less interesting. Quite the contrary. To me, appreciating the "MITO" data in this light is actually much richer. It also contains the critically important nuance that is a precondition of successful future experiments *in vivo*. Accordingly, I asked the authors to be "radically transparent" about their conclusions' dependence on COX4 and structure their paper to prevent misconceptions about what the "MITO" label actually means.

The following Transparent Peer Review Record is not systematically proofread, type-set, or edited. Because it reflects the version of paper that was formally accepted by Cell Systems, *before copy editing and approval of proofs, details may vary slightly between it and the published paper. Special characters, formatting, and equations may fail to render properly. Standard procedural text has been deleted for the sake of brevity, but all official correspondence specific to the manuscript has been preserved.*

Editorial decision letter with reviewers' comments, first round of review

Dear Howard,

I'm enclosing the comments that reviewers made on your paper, which I hope you will find useful and constructive. As you'll see, they express interest in the study, but they also have a number of criticisms and suggestions. Based on these comments, it ' premature to proceed with the paper in its current form; however, if it's possible to address the concerns raised with additional experiments and/or analysis, we'd be interested in considering a revised version of the manuscript.

To expedite and guide this revision, I discuss 4 topics here. Also, we appreciate that the COVID-19 pandemic challenges and limits what you and your lab can do, so to make sure we're absolutely on the same page about the feasibility of revisions, let's schedule a Zoom call at our earliest mutual convenience.

1. The "non-mitochondrial" mRNAs that may be mitochondrial after all. This the most important concern raised during the review process, and as it came from Reviewer 1 by email, it's not reflected in the reviewer's comments. It's the concern I emailed you about last week, and my email is excerpted here for convenience:

> *One of my reviewers has raised a concern about how a particular feature of the original APEX-seq data was accounted for in the present work. Specifically, the section called "MITO APEX-seq extended analysis" in the STAR Methods of Fazal et al. says quite explicitly that there's an issue with falsely identifying non-mitochondrial proteins as mitochondrial, and that these false calls (~400) outnumber the real ones (15) dramatically. What steps did you and your collaborators take to ensure that these false mitochondrial calls are not influencing the results in the current paper?*

You'd responded that these 400 genes aren't an artifact but rather reflect real biology, and "they are RNAs labelled by the mitochondrial-targeted APEX protein en route to the mitochondria. We are getting a view of the biogenesis and trafficking of the components that build up the mitochondria from the rest of the cell. The SARS-CoV-2 signal is more similar to these 400 genes, which deepens our interpretation that the mitochondrial localization prediction indicates a relationship between mitochondria biogenesis and double membrane vesicles used for viral replication."

This is great and very interesting! It needs to be demonstrated empirically within this paper. Also, the language used to describe these 400 genes in Fazal et al. is very stark. This is to your great credit, of course, because it accurately reflected the state of your knowledge at the time. However, if the state of your knowledge has changed, that point needs to be made just as starkly within the text of this paper. Please note that if addressing this concern requires additional figures, that's absolutely fine.

2. The reviewer's comments that I've highlighted in yellow are straightforward sanity checks that need to be included within a revision. Please let me know if this poses difficulties.

3. The reviewers' comments highlighted in blue fall into the category of "nice to have." These comments are very constructive and would certainly increase the impact of your paper, but I understand that we are balancing completeness and timeliness given the COVID situation.

4. Being wrong for interesting reasons. Reviewer 2 makes an interesting point that I've highlighted in green. I think this point is spot on, but I don't see that as a problem for your paper. Instead, if indeed the virus is altering the host's cell biology in this way, discordance between your predictions and experimental observations may prompt new studies and new discoveries (in which case, godspeed!). This healthy scientific process is contingent upon having strong confidence in your predictions, though -- one needs to be convinced that discordance reflects interesting biological possibilities, not something technically wrong with what you've done. You may wish to finesse this point in your revision.

As you address these concerns, it's important that you and I stay on the same page. I'm always happy to talk, either over email or by phone, if you'd like feedback about whether your efforts are moving the manuscript in a productive direction. Do note that we generally consider papers through only one major round of revision, so the revised manuscript would be either accepted or rejected based on the next round of comments we receive from the reviewers. If you have any questions or concerns, please let me know. More technical information and advice about resubmission can be found below my signature. Please read it carefully, as it can save substantial time and effort later.

I look forward to seeing your revised manuscript.

All the best,
Quincey

Quincey Justman, Ph.D.
Editor-in-Chief, Cell Systems

Reviewers' comments:

Reviewer #1: Wu et al present an interesting and timely application for the RNA-GPS tool in "RNA-GPS Predicts SARS-CoV-2 RNA Localization to Host Mitochondria and Nucleolus". A few observations are particularly interesting in the context of the publication of the discussed Gordon et al paper regarding the protein-protein interactions of SARS-CoV-2 in human cells. The main discovery, i.e that SARS-CoV-2 RNA Localize to the mitochondrial matrix and nucleolus, is very intriguing. The approach taken by the authors, comparing SARS-CoV-2 RNAs localization to other coronaviruses and human RNAs is a nice twist for a tool previously constructed only for human RNAs. However, it would be nice to have significance tests for these comparisons (Wilcoxon rank test?). I think it would be a necessary control to benchmark RNA-GPS predictions with some of the previously characterized localizations of RNAs from other coronaviruses. Finally, I was wondering if maybe testing whether the sequence properties of the mitochondrial-matrix bound RNAs make sense with the mitochondrial genetic code, given that probably translation would happen there (to my knowledge there is no mRNA imported to mitochondria and translated). If this is not the case, RNA's might be localizing to the outside (or going in and out) as it's a bit of a hotspot for translation and then use the mitochondrial double membrane to form the virion.

Minor comments:

-In the figure a heatmap legend would be welcome.

-I may be missing something but why does Orf9b does not appear anywhere in the manuscript? It is the one supposed to interact with TOMM70 in Gordon et al

- Why did the authors use RNN given that other ML methods were used to validate RNA-GPS in the previous manuscript?

Reviewer #2:
While in silico RNA localization models may be sufficient when applied to the data they were trained on (see initial RNA-GPS paper), I have strong reservations about using a model trained on human RNA localization to predict viral RNA localization. It seems that viral RNA would represent something drastically different from any training data given to the localization model, and thus its output may be unpredictable. ***[From QJ: As requested (and also reiterated below), this caveat needs to be emphasized more strongly.]*** Furthermore, the virus may itself alter the subcellular machinery in ways that a model trained on healthy cells could not predict. ***[From QJ: Discussed in my letter, above.]*** To their credit, the authors are honest about both of these potential limitations, and they do show that an independent model (albeit trained on the same data) produces similar results. While this helps, it does nothing to support the presumption that human RNA-trained models can predict viral RNA localization. Without further experimental results using actual viral RNA, I feel it is difficult to do so. Given the nature of the COVID-19 pandemic and the urgent need for insights, this paper could be published. However, it runs the risk of driving experiments in the wrong direction. If it were to be published, the caveats of applying machine learning models on a different data type than the training data should be more strongly emphasized.

Authors' response to the reviewers' first round comments

Attached.

Revised manuscript

Attached.

Editorial decision letter with reviewers' comments, second round of review

Dear Howard,

Thanks again for talking last week. This email summarizes our conversation and my expectations, as shaped by our conversation. It's also an official invitation to revise your manuscript.

Most essentially, we agreed that the MITO-APEX signal was not simply mitochondrial, per se. Instead, it is a complicated signal that reflects the lifetime and trafficking pattern of COX4, which is encoded by the nuclear genome and must be transported to the mitochondria, where it performs its hallmark function. As such, calling NtermCOX4-APEX-labeled transcripts mitochondrial is a potentially misleading simplification. The revision must make this crystal clear, and in doing so, it must similarly clarify and caveat other classifications brought in from the Cell paper, because as I understand it, all sub-cellular localization calls made by RNA-GPS are based on single(?) landmark -APEX constructs. So how to do this?

First, all of this must be described in detail in the first paragraph of the results section with, as I mentioned on Zoom, "radical transparency." Your text must be candid enough that nobody who reads it will assume that NtermCOX4-APEX labeled transcripts are simply mitochondrial. Within this paragraph, please also list the other -APEX landmark constructs with their dominant locations (as in Fig. S2A of Fazal et al.). After you introduce all of the constructs, say something like "for the sake of simplicity, we will describe the NtermCOX4-APEX2 signal as "mitochondrial matrix" for remainder of this work because these are the most enriched when described by log2fold-change over XXX (see Fazal et. al. Fig. S3H), but note that known mitochondrial COX4 substrates comprise 15 of the 4XX transcripts recovered by APEX-seq." A similar statement and treatment of e.g. the nucleolus signal is also expected.

Second, the word "localization" becomes problematic when the original dataset is described more fully and it should not be used. What you are actually predicting is **dominant subcellular residency,** as defined by the APEX landmark constructs. If you don't like "dominant subcellular residency," you can substitute something else, but it should be consistent throughout the text and it should indicate that there's a reasonable chance that it may transcript may be elsewhere. I prefer "residency" to localization because usage of common terms like "co-localize' suggest an precision/accuracy in time and space that aren't appropriate here.

Third, in Figure 1B, the actual APEX constructs need to be added to the existing labels, e.g. as follows: N-terminal region of COX4 (Mitochondrial Matrix). This can be done graphically or with text, but it must be crystal clear. Nobody should have to go digging into the supplement of the Fazal paper to understand how the APEX data were generated because understanding this is a prerequisite for understanding the present paper. I strongly encourage a reproduction of something like Fig S2A in Fazal et al. within Fig 1 (please feel free to add panels as necessary).

Fourth, the figure legends need to reiterate that the labels with figures reflect predicted dominant subcellular residency, as defined by the APEX landmark constructs.

You're welcome to make other changes to the text as well as long as they have similar clarifying aims. Note that my goal here is not to muddy the waters or to make your paper less interesting. Instead, the goal is as follows: **if someone reads your paper and wants to follow it up with in vivo work, they should have enough precise information available to them in your paper to design excellent experiments and interpret their results accurately.** When viewed in this light, your work in response to Reviewer 1 represents a good computational effort to deconvolute the NtermCOX4-APEX2 signal into the component dominated by COX4's work in the mitochondria and the component(s) that reflect its journey from nucleus to mitochondria (opening up the interesting question: what if these processes were corrupted by SARS-CoV-2?).

Finally, I'm expecting that we'll be in agreement about the next version of the paper, so please pay special attention to our formatting requirements (if you need me to send you the checklist again, just ask). Two important exceptions, though: 1) **If you need your text to be longer than the limits to accommodate this extra clarification, that's fine. Please take the space you need. 2) If you need more supplemental figures/items, you're welcome to include them too.**

I look forward to seeing your revised manuscript.

All the best,
Quincey

Quincey Justman, Ph.D.
Editor-in-Chief, *Cell Systems*

Reviewers' comments:

Reviewer #1: We thank the authors for trying to clarify their approach and taken into account my suggestions.

1) Still my main problem concerns the data used to train their model. Specifically in the Fazal et al paper it is said that:

"In addition to the 15 expected mitochondrial RNAs, we also recovered ~400 transcripts that were moderately enriched (log2fold-change > 0.75), some of which are known mitochondrial pseudogenes. To rule out that this labeling was not because the biotin-phenoxy radical, generated during the labeling experiment, was escaping from the mitochondrial matrix we confirmed that OMM APEX-seq enriched transcripts (> 1000-enriched transcripts) showed no enrichment (average log2fold-change ~0) in the MITO APEX-seq samples. While we do not believe these transcripts to be present in the mitochondrial matrix (Mercer et al., 2011), attempts to confirm the localization by FISH were not successful. We hypothesize two explanations for the observations: (1) Due to the large perturbation introduced by APEX-seq labeling, the DESeq2 analysis does not perform properly; or (2) There is some small background labeling by Cox4-APEX (i.e., MITO-APEX) as the protein makes it way from the cytosol, where it is translated, to the mitochondrial matrix."

So the Fazal et al paper claims that all the other those 400 counts are probably artefacts, not seen by FISH. That is precisely the data used to train their model and now claimed to be above noise. So you cannot have it both ways and training the model with less transcripts does not change that.

2) The authors also claim that since their model also predicts CMV to be WITHIN mitochondria the model

has to be correct. I think at the very least it would be useful to see how some other non-mitochondrially localized viruses behave in the model (negative controls). CMV is a DNA virus 10 times the size of SARS-CoV-2 and makes a lot more proteins. For all we know the model might only localize RNA that are short and intron-less to mitochondria.

3) Coronaviruses are well documented to replicate in membrane compartments that are distinct from mitochondria and no virus that replicates within mitochondria has ever been reported.

I think the prediction of the model regarding the localization of SARS2-COVID is interesting and I appreciate that the addition of the authors from the experimental paper is an attempt to bring credibility to the data, but I am still puzzled that data that was deemed an artefact is now deemed a true positive. However I understand that authors claim that "any followup experiments in this vein that our work inspires could prove meaningful for the field even if they may be contradictory. " and indeed finding a virus in mitochondria would be an interesting discovery. An experimental confirmation with another technique is deemed essential. So I would like that the authors very clearly state that the localization to the mitochondria is still a prediction that needs to be experimentally verified.

Attachments: RNA-GPS Predicts SARS-CoV-2 RNA Residency to Host Mitochondria and Nucleolus

Kevin E. Wu, Furqan M. Fazal, Kevin R. Parker, James Zou, Howard Y. Chang

Author-provided documents included with the Transparent Peer Review Record:

1. Authors' response to the reviewers' first round comments
2. Revised manuscript

Wu et al. point-by-point response to reviewer comments

RNA-GPS Predicts SARS-CoV-2 RNA Localization to Host Mitochondria and Nucleolus
CELL-SYSTEMS-D-20-00231

We thank the reviewers for their thoughtful comments and their positive assessment of the computational analysis presented in our paper. We have improved the manuscript with the following major additions to our results:

- Validation of RNA-GPS's concordance with (limited) experimentally measured viral subcellular transcript localizations, specifically the human cytomegalovirus.
- Wilcoxon rank tests quantifying the significance of the predicted localization at the nucleolus and mitochondrial matrix – the corresponding p-values are all significant.
- Additional analysis of the APEX-seq mitochondrial transcripts used to train RNA-GPS. New analysis shows that the predicted localization of SARS-CoV-2 is robust to potential noise in the APEX-seq data.

Corresponding sections in the discussion and STAR methods have also been updated accordingly. Please find our detailed responses to the reviewer suggestions below in black (reviewer comments have been reproduced in blue).

Reviewer 1 (via email) feedback: One of my reviewers has raised a concern about how a particular feature of the original APEX-seq data was accounted for in the present work. Specifically, the section called "MITO APEX-seq extended analysis" in the STAR Methods of Fazal et al. says quite explicitly that there's an issue with falsely identifying non-mitochondrial proteins as mitochondrial, and that these false calls (~400) outnumber the real ones (15) dramatically. What steps did you and your collaborators take to ensure that these false mitochondrial calls are not influencing the results in the current paper?

We thank the reviewer for their efforts in thoroughly examining our model and the data it was trained on. Three lines of evidence indicate that these non-canonical mitochondrial transcripts do not adversely bias the analyses, and help us to inform our interpretation of SARS-CoV-2 RNA localization:

1. The noncanonical mitochondrial genes are real signals. This required additional analyses of the original APEX-seq data that we have now provided (revised manuscript Figure S3, reproduced below, page 5).
2. A variant of the RNA-GPS model trained without the potentially noisy noncanonical mitochondrial transcripts yields the same prediction of mito matrix localization for SARS-CoV-2 RNA.
3. K-mer analysis showing SARS-CoV-2 location prediction is not biased by the noncanonical signal.

These lines of evidence are detailed below.

1. Within the APEX-seq data, the mitochondrial matrix transcripts can be categorized into the known canonical mitochondrial transcripts, which are transcribed from the mitochondrial genome, and a larger group of “non-canonical” transcripts. These non-canonical RNAs are transcripts from the nuclear genome and the experimental evidence from APEX-seq[1] suggests that they are transported to the mitochondria. The COX4-APEX2 fusion protein used to tag RNAs in the mitochondrial matrix is itself a nuclear encoded protein that is then transported to and imported into the mitochondria. Hence it is expected that the mitochondrial APEX-seq experiment can tag transcripts that originated in the mitochondrial matrix (such as those encoded by the mitochondrial genome) as well as RNAs that share transport or import mechanisms with nuclear-encoded mitochondrial-resident proteins that must use the Tomm70 import pathway. We do not believe that this non-canonical group is dominated by noise or false positives. Figure S3H in the APEX-seq manuscript (reproduced below) demonstrates that these non-canonical transcripts exhibit significant mitochondrial enrichment compared to a physically proximate, but biologically distant control – the outer mitochondrial membrane.

We can also show that the non-canonical transcripts have mitochondrial enrichment that exceeds all other compartments. Figure S3A in our revised manuscript (reproduced below, page 4) shows the logFC enrichment of the 251 nuclear-encoded mitochondrial-enriched genes present in the data we use to develop RNA-GPS, at each localization. This plot shows that these transcripts are exclusively enriched by the mito-APEX experiment but NOT at 7 other subcellular locations examined, and thus are not noise. Furthermore, a gene ontology enrichment analysis of the top 100 most enriched non-canonical mitochondrial transcripts shows coordinate enrichment of multiple transcripts in macromolecular

[1] Fazal, F.M., Han, S., Parker, K.R., Kaewsapsak, P., Xu, J., Boettiger, A.N., Chang, H.Y., and Ting, A.Y. (2019). Atlas of Subcellular RNA Localization Revealed by APEX-Seq. Cell *178*, 473-490.e426.

complexes, and many cytoskeletal and intracellular transport terms (Figure S3B in our revised manuscript, reproduced below).

We have added more discussion and clarification of this point in the revision. One possible explanation is that non-canonical transcripts are indeed imported into the mitochondria; this might suggest that SARS-CoV-2 carry localization signals tailored for double-membrane organelles bearing similarity to mitochondria. Another hypothesis is that these transcripts are co-local to the pathway by which the APEX-COX4 fusion protein (COX4 is fused to APEX to guide mitochondrial localization) itself localizes to the mitochondrial matrix, and are picked up by APEX "en route." This might suggest that SARS-CoV-2 transcripts localize near "highways" of protein transport to double-membrane organelles, perhaps in an attempt to be pulled in the same direction. This hypothesis is also supported by the aforementioned gene ontology enrichment analysis. Both hypotheses suggest that the mitochondrial signal we identify alludes to *active* localization toward double-membrane structures. To this point, we have added the following to our second Discussion section paragraph (*italicized* text is for context and is not new):

> *This suggests a high degree of resemblance between the DMV and mitochondrial localization mechanisms – leading to the hypothesis that our mitochondrial matrix localization predictions are capturing this similarity between the DMV and mitochondria.* Indeed, our observation that many of the transcripts used to "teach" RNA-GPS to recognize mitochondrial matrix localization may actually be picked up as the APEX protein itself localizes to the mitochondria further supports the idea that our localization prediction is alluding to localization mechanisms (perhaps more so than a specific physical destination).

2. To further validate the robustness of the SARS-CoV-2 mitochondrial localization signals, we retrained RNA-GPS on a stricter subset of the original APEX-seq dataset where 20% of the non-canonical RNAs with relatively weaker mitochondrial enrichment are removed. The remaining data likely have a higher signal-to-noise ratio. The retrained RNA-GPS still predicts the mitochondrial matrix and the nucleolus as the two strongest localization targets for SARS-CoV-2 RNA (revised manuscript Figure 2B, reproduced below, page 6). This analysis suggests that our finding is robust to potential noise in non-canonical mitochondrial transcripts identified by APEX-seq and lends more confidence in these results.

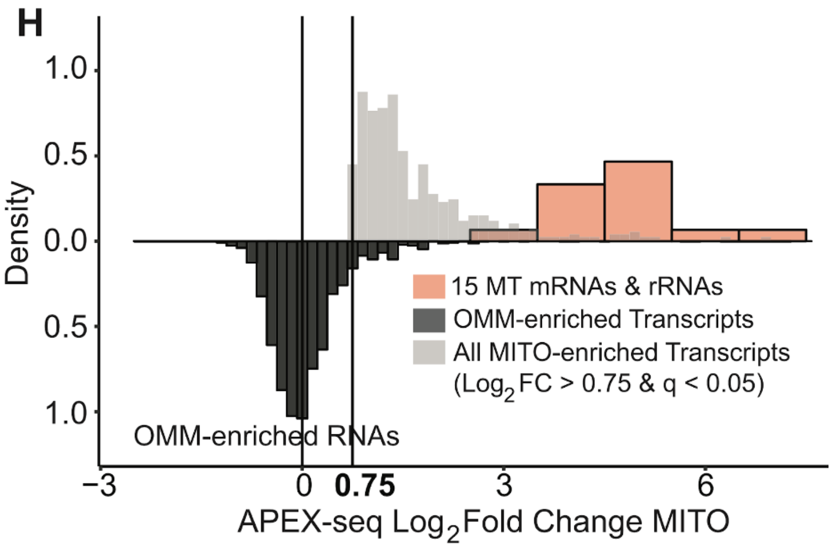
To summarize the above results, we have also added the following text regarding these transcripts, as the final paragraph in our first Results subsection "SARS-CoV-2 RNA subcellular localization patterns":

> In addition to evaluating robustness of our results to modelling strategies, we also evaluated robustness with respect to the APEX-seq data used to train the models. Within the APEX-seq data, many of the transcripts observed to localize to the mitochondrial matrix are actually encoded in the nucleus. Though surprising, these transcripts are unlikely to be artifacts of experimental noise (Figure S3A) and actually enrich for cytoskeletal processes (Figure S3B); we hypothesize that they are picked up as the APEX protein itself localizes to the mitochondria. Nonetheless, in order to ensure that our SARS-CoV-2 localization predictions were not affected by potentially noisy data, we excluded the nuclear-encoded, "non-canonical" mitochondrial matrix transcripts that had relatively low APEX-seq signal, and retrained RNA-GPS. This denoised model recapitulates the same localization enrichment for SARS-CoV-2 towards the

mitochondrial matrix and nucleolus (Figure 2B), suggesting that our predictions are robust to noise in the training data. In summary, our predicted localizations are robust across different modelling strategies, and across variation in the data used to train these models.

3. We performed additional analysis to compare the sequences of SARS-CoV-2 with that of the canonical and non-canonical human mitochondrial RNAs. In terms of k-mer frequency, the SARS-CoV-2 RNAs are about equally distant from the canonical and non-canonical RNAs (Response Table 1, page 6). As this is the feature space that RNA-GPS is built upon, this also suggests that the RNA-GPS predictions for SARS-CoV-2 are not particularly biased by the non-canonical mitochondrial transcripts identified by APEX-seq.

We believe that the three lines of supporting analysis above present a more nuanced, complete picture of our most current understanding of these mitochondrial transcripts, and their impact on our model and predictions. We apologize that the text in the APEX-seq manuscript does not fully reflect this understanding.



Fazal et al. Supplementary Figure S3H (reproduction). The black bars show mitochondrial enrichment for transcripts significantly enriched at the outer mitochondrial membrane (OMM). As the OMM is physically close yet membrane-separated and biologically distinct, the lack of enrichment for these transcripts within the mitochondria itself indicates the specificity of APEX labelling. The orange bars, indicating canonical mitochondrial transcripts, and grey bars indicating non-canonical transcripts, are both distinct from the OMM “control,” suggesting that they are not false positives or noise.

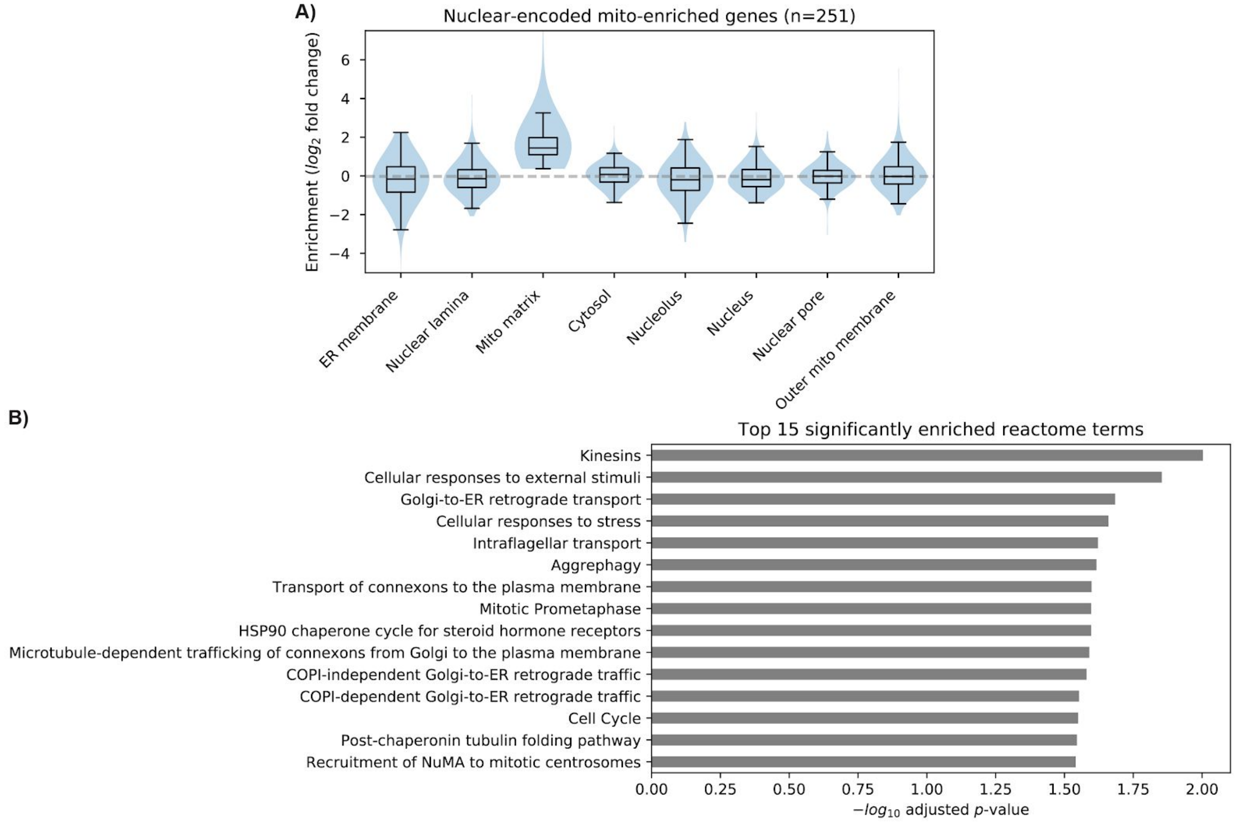


Figure S3 (related to Figure 2): Further analysis of mitochondrial transcripts used to train RNA-GPS. Within the APEX-seq training data, many transcripts localized at the mitochondrial matrix are actually encoded within the nucleus. (A) Shows a plot of enrichment scores at each compartment for these mitochondrial-enriched, nuclear-encoded “non-canonical” transcripts. We see that these transcripts have enrichment centered around 0 for all but the mitochondrial matrix, indicating that while these transcripts are nuclear-encoded, the APEX-seq labelling technology consistently and nonrandomly associates them with the mitochondrial matrix. These transcripts are also biologically meaningful; (B) shows reactome ontology analysis of 100 most enriched (by p-value) non-canonical mitochondrial matrix transcripts. There is a clear emphasis for cytoskeletal and intracellular transport terms (e.g. kinesins, post-chaperonin tubulin folding pathway, recruitment of NuMA to mitotic centrosomes; adjusted $p < 0.05$). This suggests that the non-canonical transcripts might be consistently picked up as the APEX-seq protein is itself trafficked to the mitochondria.

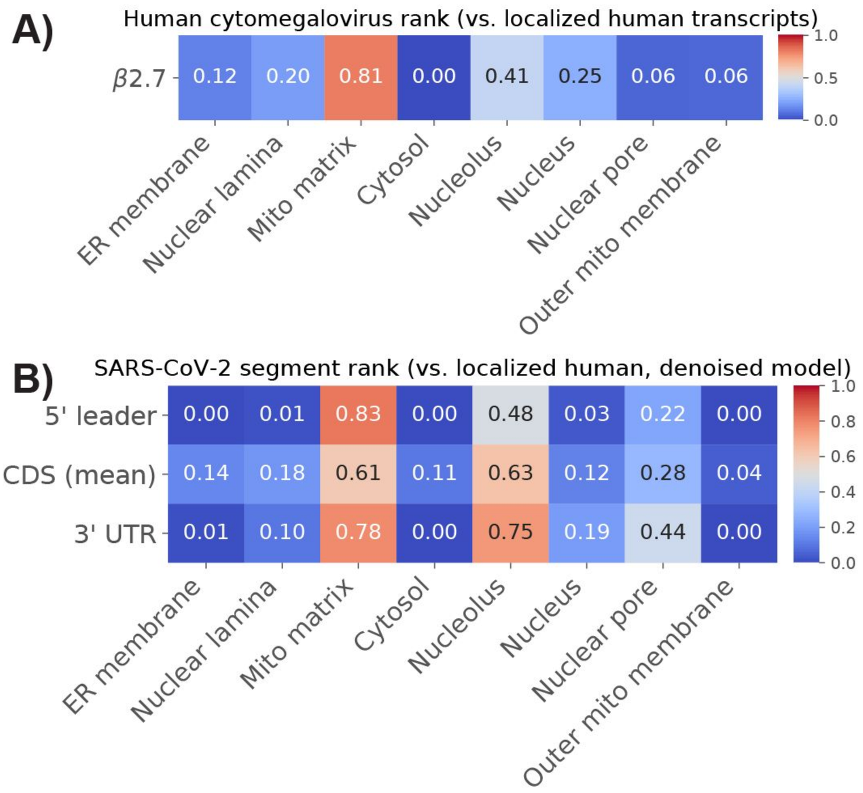


Figure 2: Validation of SARS-CoV-2 localization predictions. (A) RNA-GPS predictions for the human cytomegalovirus β2.7 transcript, which has been shown to localize to the inner mitochondrial membrane. RNA-GPS correctly predicts its localization to the closest compartment it has been trained on – the mitochondrial matrix. This provides support that RNA-GPS can make accurate predictions on viral RNA. (B) To evaluate the effect of the potentially noisy mitochondrial examples in our APEX-seq training set on predicted SARS-CoV-2 localizations, we trained a denoised variant of RNA-GPS on a subsetted dataset that excludes these examples. This denoised model predicts the same localization pattern for the 3 segments of the SARS-CoV-2 sgRNAs (compare to Figure 1E). For additional validation experiments, see Figure S3A/B.

Euclidean distance (feature space)	**SARS-CoV-2**	**Noncanonical**	**Canonical**
SARS-CoV-2	0.166	0.382	0.363
Non-canonical mito matrix	0.382	0.366	0.383
Canonical mito matrix	0.363	0.383	0.276

Response Table 1: Average Euclidean distance in RNA-GPS *k*-mer feature space (4,032 dimensions) between all pairwise comparisons of transcripts in shown categories. Values along the diagonal reflect average pairwise feature distance within each category of transcripts. For SARS-CoV-2 transcripts, we only compare complete genomes (leaving out sgRNAs) for simplicity. SARS-CoV-2 does not appear to be biased in distance towards either the canonical or non-canonical APEX mitochondrial transcripts; we do not expect neither class of APEX transcripts to have an outsize effect on SARS-CoV-2 localization predictions.

Reviewer #1: Wu et al present an interesting and timely application for the RNA-GPS tool in "RNA-GPS Predicts SARS-CoV-2 RNA Localization to Host Mitochondria and Nucleolus". A few observations are particularly interesting in the context of the publication of the discussed Gordon et al paper regarding the protein-protein interactions of SARS-CoV-2 in human cells. The main discovery, i.e that SARS-CoV-2 RNA Localize to the mitochondrial matrix and nucleolus, is very intriguing. The approach taken by the authors, comparing SARS-CoV-2 RNAs localization to other coronaviruses and human RNAs is a nice twist for a tool previously constructed only for human RNAs. However, it would be nice to have significance tests for these comparisons (Wilcoxon rank test?). I think it would be a necessary control to benchmark RNA-GPS predictions with some of the previously characterized localizations of RNAs from other coronaviruses. Finally, I was wondering if maybe testing whether the sequence properties of the mitochondrial-matrix bound RNAs make sense with the mitochondrial genetic code, given that probably translation would happen there (to my knowledge there is no mRNA imported to mitochondria and translated). If this is not the case, RNA's might be localizing to the outside (or going in and out) as it's a bit of a hotspot for translation and then use the mitochondrial double membrane to form the virion.

Thank you for the positive assessment and thoughtful comments.

We agree that having a significance test around the comparisons would be helpful. To that end, we have performed a Wilcoxon rank-sum test on the SARS-CoV-2 predictions for each region, against a distribution of unlocalized transcripts for that corresponding region. We find that all of the mitochondrial matrix and nucleolus predictions significantly exceed the negative, no-localization baseline (revised manuscript Table S2, reproduced below, page 11), indicating that the high rank scores we report for these compartments is highly statistically significant. Thank you for this suggestion that has strengthened the work.

Second, regarding testing RNA-GPS predictions against a known viral RNA benchmark, we were able to do so with a human cytomegalovirus (HCMV) RNA (we are not aware of any transcript localization studies performed on coronaviruses). HCMV's β2.7 mRNA (accession NC_006273.2) has been demonstrated to exhibit mitochondrial localization, specifically to the inner mitochondrial membrane (IMM)[2]. The new manuscript text discussing this result (3rd to last paragraph in Results section "SARS-CoV-2 localization patterns") is excerpted below, and the pertinent Figure 2A is reproduced below on page 10 as well.

> While direct experimental data measuring coronavirus sgRNA transcript localization is not currently available, we sought to validate our predictions on other human viruses with known subcellular localizations. After conducting a systematic literature search, we found one such example: the human cytomegalovirus β2.7 mRNA transcript, which localizes to the inner mitochondrial membrane (IMM). RNA-GPS predicts this transcript to localize to the mitochondrial matrix with a rank score of 0.81; no other compartments have a rank score exceeding 0.5 (Figure 2A). Thus, the algorithm's prediction is in close agreement with experimental evidence for β2.7 mRNA localization. While large-scale comparisons are not

[2] Williamson, C.D., DeBiasi, R.L., and Colberg-Poley, A.M. (2012). Viral product trafficking to mitochondria, mechanisms and roles in pathogenesis. Infect Disord Drug Targets *12*, 18-37.

> currently feasible due to lack of datasets measuring viral transcript localization, this example provides some reassurance that RNA-GPS' predicted localizations are reasonable.

Regarding the sequence similarity with the mitochondrial genetic code – this is certainly an interesting idea. However, our interpretation of our localization results focuses on their relation to double membrane vesicles and protein-protein interactions; we do not suggest that viral transcripts are actually imported into the mitochondria for subsequent translation by mitochondrial mechanisms. Thus, we do not feel that performing such an analysis would not substantially add to readers' comprehension and understanding of our results and methods.

Minor comments:

-In the figure a heatmap legend would be welcome.

Thank you for pointing this out. We have revised the Figure 1 legend to contain a single color scale shared between all heatmaps. All figure panels across primary and supplementary figures now have heatmap legends as well. We specify the identities of the column and rows, and explain in the caption the meaning of the color scale.

-I may be missing something but why does Orf9b does not appear anywhere in the manuscript? It is the one supposed to interact with TOMM70 in Gordon et al

Thank you for this suggestion. ORF9b is an accessory protein encoded by the sgRNA corresponding to the "N" gene, and does not appear to explicitly form a separate sgRNA transcript, which is why we did not call out its localization. We have clarified this point in our revised manuscript. We have also pointed out the important interaction between ORF9b and TOMM70, the mitochondrial import channel, as the reviewer suggested. The relevant updated text is included below:

> … The same study found that the ORF9b protein, produced by the "N" sgRNA, interacts with TOMM70, a mitochondrial import receptor that plays a critical role in modulating interferon response – a key anti-viral cellular defense pathway….

- Why did the authors use RNN given that other ML methods were used to validate RNA-GPS in the previous manuscript?

We use the GRU RNN model because it also achieved relatively good prediction performance on APEX-seq; please see Figure 2a in the RNA-GPS manuscript[3]. In fact, we use the exact same recurrent GRU model to validate RNA-GPS as we do to validate predictions here. Using the GRU model conveys several advantages: (1) it uses a fundamentally different, one-hot-encoding featurization strategy compared to RNA-GPS's k-mer featurization, and is thus "orthogonal" in its modelling approach; (2) compared to other sequence models such as Basset[4], recurrent models can process variable-sized input,

[3] Wu, K.E., Parker, K.R., Fazal, F.M., Chang, H., and Zou, J. (2020). RNA-GPS predicts high-resolution RNA subcellular localization and highlights the role of splicing. RNA.

[4] Kelley, D.R., Snoek, J., and Rinn, J. (2016). Basset: Learning the regulatory code of the accessible genome with deep convolutional neural networks. Genome Research.

which allows it to “see” the entire transcript regardless of its length; and (3) we previously demonstrated in our RNA-GPS manuscript that existing neural network localization models, such as RNATracker[5], suffer from a tendency to overfit and exhibit poor generalization even on human APEX-seq held-out test data, which suggests that they would be even poorer models for generalizing to viral localization.

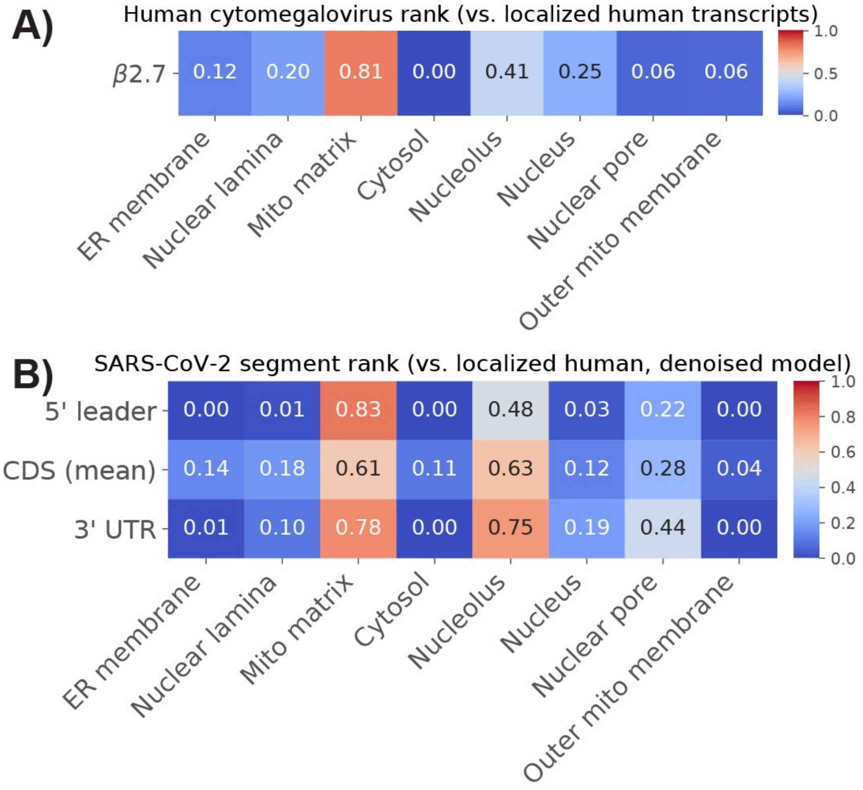


Figure 2: Validation of SARS-CoV-2 localization predictions. (A) RNA-GPS predictions for the human cytomegalovirus β 2.7 transcript, which has been shown to localize to the inner mitochondrial membrane. RNA-GPS correctly predicts its localization to the closest compartment it has been trained on – the mitochondrial matrix. This provides support that RNA-GPS can make accurate predictions on viral RNA. (B) To evaluate the effect of the potentially noisy mitochondrial examples in our APEX-seq training set on predicted SARS-CoV-2 localizations, we trained a denoised variant of RNA-GPS on a subsetted dataset that excludes these examples. This denoised model predicts the same localization pattern for the 3 segments of the SARS-CoV-2 sgRNAs (compare to Figure 1E). For additional validation experiments, see Figure S3A/B.

[5] Yan, Z., Lecuyer, E., and Blanchette, M. (2019). Prediction of mRNA subcellular localization using deep recurrent neural networks. Bioinformatics *35*, i333-i342.

	ER membrane	Nuclear lamina	Mito matrix	Cytosol	Nucleolus	Nucleus	Nuclear pore	Outer mito membrane
orf1ab	**1.38E-05**	1.00	**4.13E-32**	1.00	**4.86E-24**	1.00	1.00	1.00
s	**3.54E-05**	1.00	**8.01E-41**	1.00	**1.04E-37**	1.00	1.00	1.00
orf3a	1.00	1.00	**9.08E-61**	1.00	**5.89E-36**	1.00	1.00	1.00
e	8.91E-01	**2.86E-13**	**3.11E-65**	1.00	**1.38E-22**	1.00	1.00	1.00
m	**1.23E-03**	**2.14E-05**	**8.72E-53**	1.00	**1.73E-21**	1.00	1.00	1.00
orf6	1.00	**7.60E-09**	**1.31E-47**	1.00	**3.47E-25**	1.00	1.00	1.00
orf7a	5.92E-02	1.00	**1.01E-54**	1.00	**4.26E-25**	1.00	1.00	1.00
orf8	2.62E-01	7.38E-01	**2.56E-61**	1.00	**7.50E-26**	1.00	1.00	1.00
n	1.00	1.00	**7.83E-64**	1.00	**1.15E-22**	1.00	6.10E-01	1.00
orf10	1.00	**6.84E-03**	**5.04E-85**	1.00	**5.17E-05**	1.00	**7.26E-08**	1.00
orf7b	1.00	8.94E-01	**6.77E-09**	1.00	**2.85E-05**	1.00	1.00	1.00

Table S2 (related to Figure 1): Wilcoxon rank-sum test p-values comparing SARS-CoV-2 sgRNAs' localization predictions against those of unlocalized transcripts. All p-values are Holm-adjusted. Values that exceed our significance cutoff of 0.05 are in bold. The SARS-CoV-2 sgRNA localization predictions towards the mitochondrial matrix and nucleolus both have consistently significant p-values, indicating that their predictions are significantly higher than that of unlocalized transcripts (for the respective compartment), suggesting significant predicted localization.

Reviewer #2: While in silico RNA localization models may be sufficient when applied to the data they were trained on (see initial RNA-GPS paper), I have strong reservations about using a model trained on human RNA localization to predict viral RNA localization. It seems that viral RNA would represent something drastically different from any training data given to the localization model, and thus its output may be unpredictable. [From QJ: As requested (and also reiterated below), this caveat needs to be emphasized more strongly.] Furthermore, the virus may itself alter the subcellular machinery in ways that a model trained on healthy cells could not predict. [From QJ: Discussed in my letter, above.] To their credit, the authors are honest about both of these potential limitations, and they do show that an independent model (albeit trained on the same data) produces similar results. While this helps, it does nothing to support the presumption that human RNA-trained models can predict viral RNA localization. Without further experimental results using actual viral RNA, I feel it is difficult to do so. Given the nature of the COVID-19 pandemic and the urgent need for insights, this paper could be published. However, it runs the risk of driving experiments in the wrong direction. If it were to be published, the caveats of applying machine learning models on a different data type than the training data should be more strongly emphasized.

Thank you for your encouragement and thoughtful comments. We fully acknowledge the reviewer's reservations regarding using a model trained on "normal" human cells to extrapolate to a viral setting – one that represents not only a different species, but also a setting where many cellular functions might be disrupted or altered. While we touched on these shortcomings, we agree that we could emphasize these more strongly.

We have also newly validated RNA-GPS's predictions on the human cytomegalovirus's $\beta 2.7$ mRNA (accession NC_006273.2), which has been experimentally demonstrated to exhibit mitochondrial localization, specifically to the inner mitochondrial membrane (IMM)[6]. The following text is excerpted from our revised manuscript (Results section, 3rd-to-last paragraph of "SARS-CoV-2 localization patterns"), and explains this result in detail.

> While direct experimental data measuring coronavirus sgRNA transcript localization is not currently available, we sought to validate our predictions on other human viruses with known subcellular localizations. After conducting a systematic literature search, we found one such example: the human cytomegalovirus $\beta 2.7$ mRNA transcript, which localizes to the inner mitochondrial membrane (IMM). RNA-GPS predicts this transcript to localize to the mitochondrial matrix with a rank score of 0.81; no other compartments have a rank score exceeding 0.5 (Figure 2A). Thus, the algorithm's prediction is in close agreement with experimental evidence for $\beta 2.7$ mRNA localization. While large-scale comparisons are not currently feasible due to lack of datasets measuring viral transcript localization, this example provides some reassurance that RNA-GPS' predicted localizations are reasonable.

The mentioned Figure 2A has been reproduced below on page 14, with its caption. We thank you for this piece of feedback that has helped strengthen our work.

[6] Williamson, C.D., DeBiasi, R.L., and Colberg-Poley, A.M. (2012). Viral product trafficking to mitochondria, mechanisms and roles in pathogenesis. Infect Disord Drug Targets *12*, 18-37.

We have also revised the Discussion to highlight the reviewer's concern. We acknowledge that a computational model trained on normal human cells could be wrong about viral RNA localization, and highlight why this is a situation that we can be wrong for interesting reasons. Please see the following revised penultimate paragraph in our Discussion section, reproduced in its entirety:

> A limitation of our work lies in that it applies models trained on human RNA transcript localization data to viral transcripts. It is possible that SARS-CoV-2 infection could alter the host subcellular structures and RNA transport machinery so drastically that our learned localization patterns from human cells no longer hold. If RNA-GPS's predictions turn out to be wrong for this reason, this might suggest that coronavirus infection devastates host cell mRNA trafficking and localization – a previously unrecognized feature of COVID-19 pathobiology. After all, the vast majority of RNA binding proteins in the host cell, which are key drivers of transcript localization, recognize and process RNAs irrespective of whether they are endogenous or foreign, and inability to "properly" localize viral RNAs should mirror a similar breakdown for host cell mRNAs. As we are unable to use existing experimental evidence to thoroughly evaluate and cross-reference the predictions discussed here, future experiments in this vein are clearly necessary. Given the historical scarcity of studies focusing on viral transcript localization, such experiments would likely reveal interesting, crucial insights into viral pathobiology, whether they confirm our specific mitochondrial and nucleolus predictions or not. It is worth pointing out, though, that this is but one of many complex, interconnected viral mechanisms at play.

Nevertheless, we understand the reviewer's concerns regarding the possibility of driving experiments in the wrong direction. We believe that beyond the specific nucleolus and mitochondrial matrix predictions that are discussed, the larger message we wish to convey is that RNA transcript localization is, in general, an interesting facet of coronavirus biology that we feel warrants additional exploration. While we sincerely hope that our predictions here can be confirmed experimentally, we feel that given the relative lack of works currently focusing on RNA localization for coronaviruses, any followup experiments in this vein that our work inspires could prove meaningful for the field even if they may be contradictory. Please see the above reproduced penultimate paragraph, specifically the penultimate sentence.

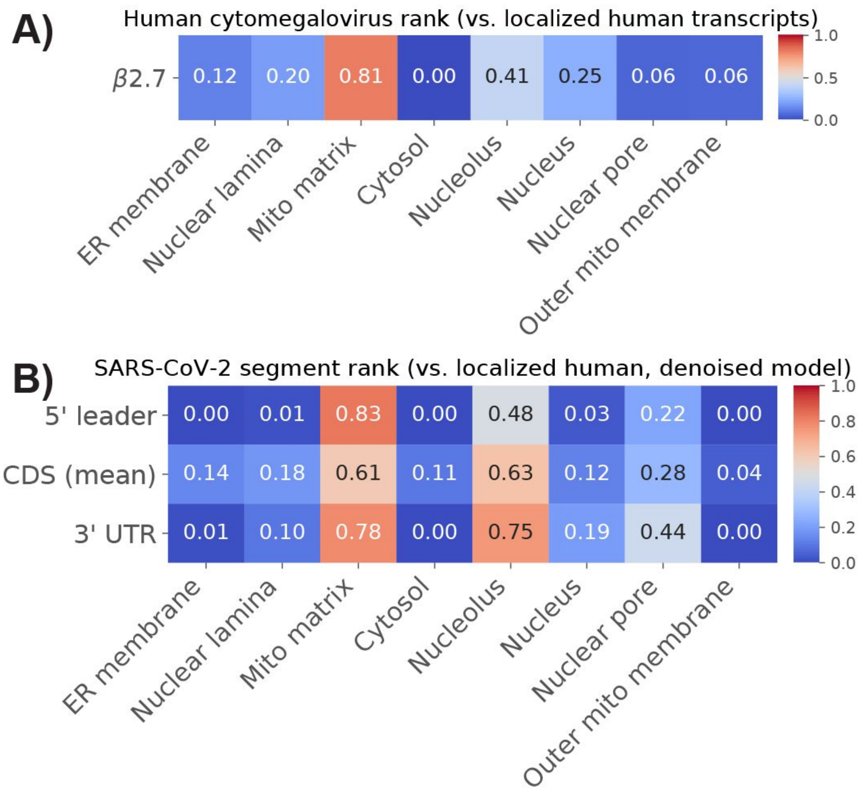
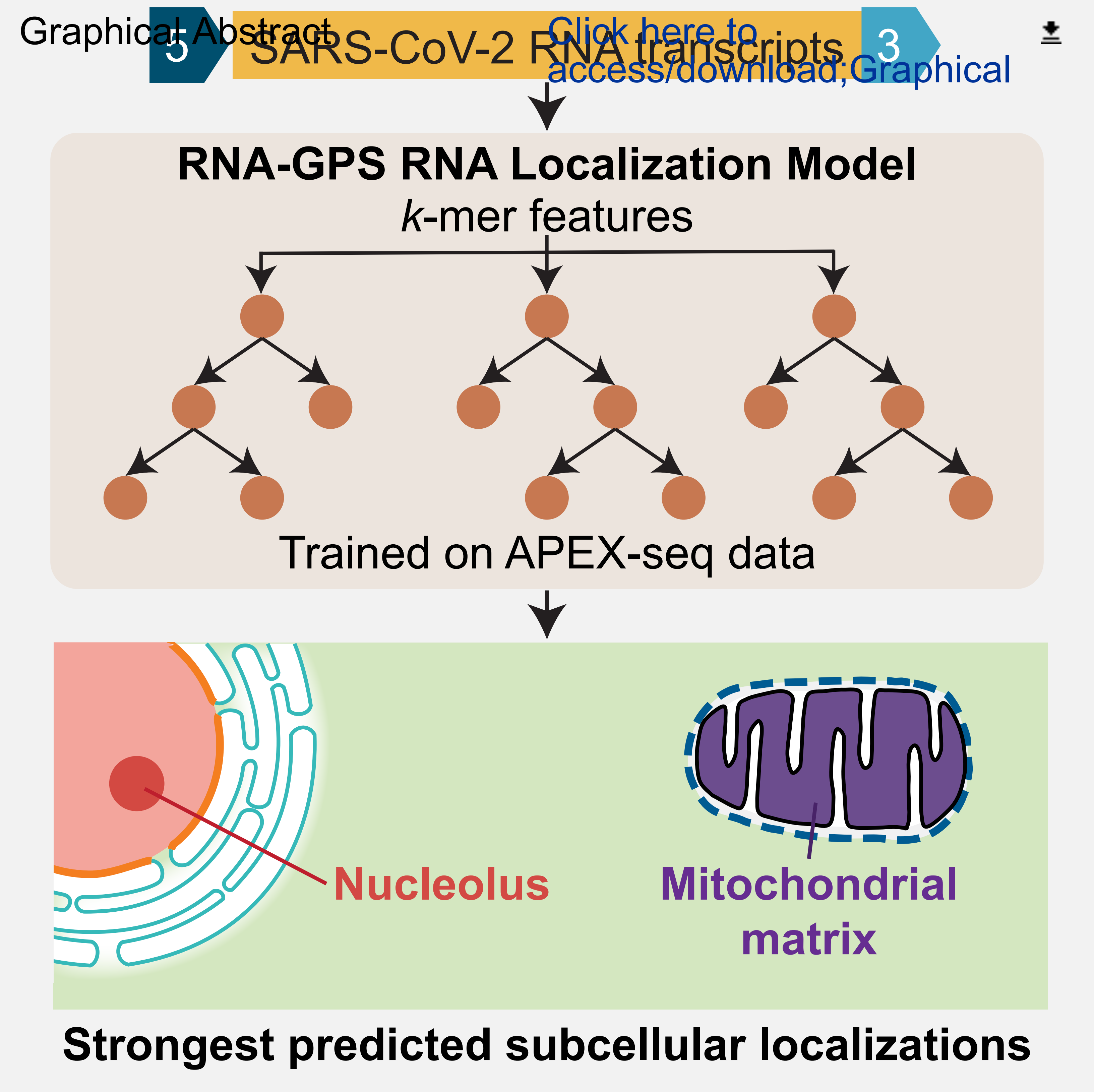


Figure 2: Validation of SARS-CoV-2 localization predictions. (A) RNA-GPS predictions for the human cytomegalovirus β2.7 transcript, which has been shown to localize to the inner mitochondrial membrane. RNA-GPS correctly predicts its localization to the closest compartment it has been trained on – the mitochondrial matrix. This provides support that RNA-GPS can make accurate predictions on viral RNA. (B) To evaluate the effect of the potentially noisy mitochondrial examples in our APEX-seq training set on predicted SARS-CoV-2 localizations, we trained a denoised variant of RNA-GPS on a subsetted dataset that excludes these examples. This denoised model predicts the same localization pattern for the 3 segments of the SARS-CoV-2 sgRNAs (compare to Figure 1E). For additional validation experiments, see Figure S3A/B.


5
SARS-CoV-2 RNA transcripts
3
RNA-GPS RNA Localization Model
k-mer features
Trained on APEX-seq data
Nucleolus
Mitochondrial matrix
Strongest predicted subcellular localizations

RNA-GPS Predicts SARS-CoV-2 RNA Localization to Host Mitochondria and Nucleolus

Kevin E. Wu[1,2,3], Furqan M. Fazal[3], Kevin R. Parker[3], James Zou[1,2,*], Howard Y. Chang[3,4,*]

[1]Department of Computer Science, Stanford University, Stanford, CA 94305, USA
[2]Department of Biomedical Data Science, Stanford University School of Medicine, Stanford, CA 94305, USA
[3]Center for Personal and Dynamic Regulomes, Stanford University School of Medicine, Stanford, CA 94305, USA
[4]Howard Hughes Medical Institute, Stanford University School of Medicine, Stanford, CA 94305, USA
[*]Co-corresponding authors: James Zou (jamesz@stanford.edu) & Howard Y. Chang (howchang@stanford.edu)

Abstract/Summary

The SARS-CoV-2 genomic and subgenomic RNA (sgRNA) transcripts hijack the host cell's machinery. Subcellular localization of its viral RNA could thus play important roles in viral replication and host antiviral immune response. Here we perform computational modeling of SARS-CoV-2 viral RNA localization across eight subcellular neighborhoods. We compare hundreds of SARS-CoV-2 genomes to the human transcriptome and other coronaviruses. We predict that the SARS-CoV-2 RNA genome and all sgRNAs to be enriched towards the host mitochondrial matrix and nucleolus, and that the 5' and 3' viral untranslated regions contain the strongest, most distinct localization signals. We discuss the mitochondrial localization signal in relation to the formation of double-membrane vesicles, a critical stage in the coronavirus life cycle. Our computational analysis serves as a hypothesis generation tool to suggest models for SARS-CoV-2 biology and inform experimental efforts to combat the virus.

Keywords: SARS-CoV-2, COVID-19, coronavirus, viral RNA localization, machine learning, hypothesis generation

Introduction

COVID-19 (coronavirus disease 2019) has become a global pandemic, fueled by the rapid spread of the coronavirus SARS-CoV-2 (severe acute respiratory syndrome coronavirus 2), a positive strand RNA virus (Wu et al., 2020a, Sanche et al., 2020). The scientific community is actively trying to understand SARS-CoV-2's biological mechanisms and effects. Here, we computationally analyze the subcellular localization patterns of SARS-CoV-2 RNA transcripts. Our results suggest potential avenues for experimental validation and follow-up, while providing a template for *in silico* analyses of viral RNA.

RNA subcellular localization is critical to a myriad of cellular processes (Ryder and Lerit, 2018, Chin and Lécuyer, 2017, Buxbaum et al., 2015). Researchers have also discovered that RNA localization plays a significant role in the life cycle of viruses, with functions ranging from regulating sites of virion assembly (Becker and Sherer, 2017) to disrupting host mitochondrial function (Somasundaran et al., 1994). However, subcellular localization for SARS-CoV-2, and for other coronaviruses, is largely unexplored.

Gaining a better understanding of the behavior and localization of SARS-CoV-2 RNA transcripts can lead to a better understanding of its function and pathogenicity, potentially revealing targetable mechanisms.

We perform computational modelling for SARS-CoV-2 subcellular RNA localization. In particular, we build upon our recent work developing RNA-GPS, a state-of-the-art computational model predicting high-resolution RNA localization in human cells (Wu et al., 2020b) that was trained on transcriptome-wide localization patterns of human RNAs across eight subcellular landmarks (Fazal et al., 2019). RNA-GPS's strong performance, coupled with viruses' dependence on hijacking and repurposing existing cell machinery for reproduction, suggests that RNA-GPS could provide insights into SARS-CoV-2's localization behavior and can focus future experimental efforts.

We use RNA-GPS to interrogate the localization patterns of SARS-CoV-2's genome, which spans approximately 30 kilobases of single-stranded positive-sense RNA (Kim et al., 2020) (Figure 1A). RNA-GPS predicts that SARS-CoV-2 is enriched for localization in the nucleolus and the mitochondria. Comparison of SARS-CoV-2's predicted localization with that of other human coronaviruses, including strains causing the common cold, Middle East respiratory syndrome (MERS), and the SARS outbreak of 2003, shows that SARS-CoV-2 exhibits a stronger mitochondrial and nuclear localization signal than a large majority of its coronavirus relatives. We additionally find that this localization signal appears to be driven by the 5' and 3' ends of the viral genome. We conclude by connecting our observations to known RNA and viral biology, proposing possible explanatory mechanisms for previously observed phenomena. Our analysis suggests experimental validation of our predictions and serves as a framework for applying machine learning for principled hypothesis generation in viral biology.

Results

We leverage our recent work developing RNA-GPS, a computational model predicting high-resolution RNA localization in human cells trained with HEK293T APEX-seq data (Wu et al., 2020b). Briefly, RNA-GPS predicts localization of RNA transcripts to eight different subcellular locations: the cytosol, endoplasmic reticulum, mitochondrial matrix, outer mitochondrial membrane, nucleus, nucleolus, nuclear lamina, and nuclear pore (Figure 1B), and has been shown to generalize well to additional cell lines, including HeLa-S3 and K562. Although RNA-GPS is trained on human, not viral, RNA transcripts, its strong test performance combined with the fact that viruses commandeer human cellular machinery suggest that it offers a reasonable hypothesis of viral transcript localization behavior given currently available data.

We consider localization predictions to each compartment averaged across all released and annotated SARS-CoV-2 genomes available as of April 6, 2020 (n = 213) on GenBank (Coordinators, 2018). SARS-CoV-2 is believed to enter the cell as a positive strand genomic RNA, subsequently forming 11 positive strand sub-genomic RNA (sgRNA) transcripts encoding different open reading frames and sharing the same 5' leader sequence and 3' untranslated region (UTR) (Figure 1A). Within each viral genome, we predict the localization of each sgRNA produced from the primary SARS-CoV-2 genome.

To provide more context, we frame these predicted localization probabilities relative to the predictions of other relevant baseline transcript sequences. We consider two such baselines: the distribution of model predictions on transcripts exhibiting significant enrichment within the human HEK293T cell line (n = 366 transcripts) (Fazal et al., 2019), and the distribution of model predictions on transcripts derived from human coronaviruses, excluding SARS-CoV-2 (n = 191 genomes, spanning diseases from the common cold to MERS, Table S1). The human baseline quantifies the strength of localization signals in SARS-CoV-2 relative to naturally occurring human transcripts with well-characterized localization behaviors. The coronavirus baseline focuses on differences in the localization behavior of SARS-CoV-2 relative to similar viral specimens – differences that may help researchers focus on the peculiarities of this virus. For both baselines, we calculate the proportion of the baseline distribution that the SARS-CoV-2 localization prediction exceeds, which we refer to as a rank score. For example, a localization rank score of 0.6 for the nucleolus relative to human transcripts suggests that the particular viral RNA is more likely to localize to the nucleolus compared to 60% of the human RNAs that show some localization to the nucleolus.

SARS-CoV-2 RNA subcellular localization patterns

We find that compared to transcripts with known localizations in human cells, SARS-CoV-2 has a notable localization signal towards the mitochondrial matrix, as well as the nucleolus (Figure 1C). We observe consistent localization predictions across different sgRNAs encoded by the virus (shown in each row, Figure 1C). Prior works have shown that some RNA viruses exhibit transcript localization to mitochondria (Somasundaran et al., 1994), and that the nucleolus plays a prominent role in the viral life cycle, even for viruses that primarily replicate in the cytoplasm as SARS-CoV-2 presumably does (Salvetti and Greco, 2014).

In addition to framing our results with respect to endogenous human transcripts, we also compare predicted localization signals of SARS-CoV-2 sgRNAs to that of other human coronaviruses (Figure 1D). Here, we observe similar overall trends in our localization predictions. Consistent with the comparison to human transcripts, we find the SARS-CoV-2 mitochondrial matrix localization signal is stronger than that of many other coronaviruses. Additionally, we see an overall pattern suggesting that SARS-CoV-2 may have a greater affinity for nuclear localizations (nuclear pore, nucleus, nucleolus, and nuclear lamina) compared to other coronaviruses.

We also compared the overall localization patterns of the coronavirus family (excluding SARS-CoV-2) with human transcripts using RNA-GPS. We found that the most prominent localization signals for general human coronaviruses pointed towards the nucleolus, mitochondrial matrix, and ER membrane (Figure S1). Overall, our computational analysis suggests that SARS-CoV-2's sgRNA transcript localization towards the mitochondrial matrix and nucleolus may be amplifications of localization behaviors that were already present in coronaviruses.

While direct experimental data measuring coronavirus sgRNA transcript localization is not currently available, we sought to validate our predictions on other human viruses with known subcellular localizations. After conducting a systematic literature search, we found one such example: the human cytomegalovirus β2.7 mRNA transcript, which localizes to the inner mitochondrial membrane (IMM) (Williamson et al., 2012). RNA-GPS predicts this transcript to localize to the mitochondrial matrix with a rank score of 0.81; no other compartments have a rank score exceeding 0.5 (Figure 2A). Thus, the algorithm’s prediction is in close agreement with experimental evidence for β2.7 mRNA localization. While large-scale comparisons are not currently feasible due to lack of datasets measuring viral transcript localization, this example provides some reassurance that RNA-GPS’ predicted localizations are reasonable.

To further validate the robustness of these localization results, we also trained a different predictive algorithm (a recurrent neural network, see STAR Methods for additional details) on the APEX-seq data and performed a similar set of experiments, comparing SARS-CoV-2 localization predictions to human and coronavirus baselines. This alternative model also predicts strong mitochondrial matrix and nucleolus localization for SARS-CoV-2 (Figure S2A/B). Since this algorithm uses a very different modeling strategy as RNA-GPS and converges to similar findings, this suggests that the mitochondrial matrix and nucleolus signals are not artifacts of a particular model and increases our confidence in our findings.

In addition to evaluating robustness of our results to modelling strategies, we also evaluated robustness with respect to the APEX-seq data used to train the models. Within the APEX-seq data, many of the transcripts observed to localize to the mitochondrial matrix are actually encoded in the nucleus (Fazal et al., 2019). Though surprising, these transcripts are unlikely to be artifacts of experimental noise (Figure S3A) and actually enrich for cytoskeletal processes (Figure S3B); we hypothesize that they are picked up as the APEX protein itself localizes to the mitochondria. Nonetheless, in order to ensure that our SARS-CoV-2 localization predictions were not affected by potentially noisy data, we excluded the nuclear-encoded, “non-canonical” mitochondrial matrix transcripts that had relatively low APEX-seq signal, and retrained RNA-GPS. This denoised model recapitulates the same localization enrichment for SARS-CoV-2 towards the mitochondrial matrix and nucleolus (Figure 2B), suggesting that our predictions are robust to noise in the training data. In summary, our predicted localizations are robust across different modelling strategies, and across variation in the data used to train these models.

SARS-CoV-2 negative strand RNA also localizes to mitochondria and nucleolus

During their replication life cycle, coronaviruses like SARS-CoV-2 copy their positive strand RNA to create a negative strand RNA that serves as the template for viral “transcription” and production of sgRNAs (Wu and Brian, 2010). We applied RNA-GPS to the negative strand SARS-CoV-2 sgRNA precursors and discovered that they also exhibit localization signal to the mitochondrial matrix and nucleolus (Figure S4). This result suggests that the sequence features driving these localization patterns are independently present in both positive and negative strand RNAs, further boosting the localization capability of SARS-CoV-2 during different stages of its viral cycle.

SARS-CoV-2 5’ and 3’ UTRs contain strong localization signals

In addition to predicting localization, our computational model can also help understand which regions of the transcript may be more responsible for driving these observed localizations. We specifically investigated the potential contribution of the three main regions of the SARS-CoV-2 genome: the shared 5’ leader sequence, the shared 3’ UTR, and the variable “coding” sequence in the middle. We predicted the localization for each of these regions by itself (Figure 1E). The 5’ leader sequence shows the strongest localization signal to the mitochondrial matrix, and relatively low signal for the nucleolus. In contrast, the

3’ UTR has the strongest localization for the nucleolus and also has strong signal for the mitochondrial matrix. The coding sequence (CDS) also shows specific signals for these two compartments. Because the 5’ and 3’ sequences are shared by the different SARS-CoV-2 sgRNAs, this is likely a strong factor behind the consistent localization patterns we find across the different sgRNAs. We also performed further computational ablation studies of RNA binding protein (RBP) motifs in SARS-CoV-2. However, computational deletions of all instances of each RBP motif in SARS-CoV-2 sequence, repeated across all enriched RBPs, did not significantly alter the RNA-GPS predictions. This result suggests that the SARS-CoV-2 localization signal could be abundant in the viral genome and may involve complex interactions not captured by relatively short single RBP binding motifs.

Discussion

In this work, we apply computational models of human RNA transcript localization to better understand the subcellular localization of the SARS-CoV-2 genome and its constituent sgRNAs. This approach builds upon the idea that the virus uses existing human cell machinery to reproduce, and consequently that sequence-based localization signals are likely shared between human and coronavirus transcripts. The strengths of this approach include (1) the potential to understand viral RNA localization without the risk of live viral cultures; (2) the ability to examine hundreds of viral isolates and related coronaviruses and thousands of RBP motif ablations; (3) the ability to examine viral genes, UTRs, and negative strands individually, which may otherwise require the ability to precisely synchronize and arrest the viral life cycle. We find that SARS-CoV-2 appears to harbor strong localization signals towards the mitochondrial matrix and nuclear compartments, comparable to human RNA and more so than other coronaviruses. This intriguing hypothesis suggests future experimental exploration and validation.

These results might appear surprising, as one might expect localization signals to enrich towards regions like the endoplasmic reticulum, where viral translation, viral assembly, and disruption of normal cell activity are commonly known to occur (Fung and Liu, 2014, Minakshi et al., 2009, Nal et al., 2005). However, coronaviruses are known to produce complex double-membrane vesicle (DMV) structures during viral replication, which may serve functions like concealing the virus from cellular defenses (Hagemeijer et al., 2012, Knoops et al., 2008). While these DMVs are generally believed to be formed via viruses manipulating the ER membrane (Blanchard and Roingeard, 2015), the mechanism for importing and packaging proteins and RNA into these miniature organelles is not as clearly understood. One possible mechanism for importing viral RNA involves the virus exploiting endogenous RNA localization mechanisms that the cell already possesses for existing double-membrane organelles: namely, the mitochondria. Indeed, introducing just 2 amino acid point mutations in the murine coronavirus can cause both a significant drop in the number of DMV structures observed, as well as a sharp increase in viral protein localization at the mitochondria (Clementz et al., 2008). This suggests a high degree of resemblance between the DMV and mitochondrial localization mechanisms – leading to the hypothesis that our mitochondrial matrix localization predictions are capturing this similarity between the DMV and mitochondria. Indeed, our observation that many of the transcripts used to “teach” RNA-GPS to recognize mitochondrial matrix localization may actually be picked up as the APEX protein itself localizes to the mitochondria further supports the idea that our localization prediction is alluding to localization mechanisms (perhaps more so than a specific physical destination). Furthermore, DMVs have been shown to contain double-stranded RNA (Hagemeijer et al., 2012), and our strand-agnostic localization predictions are concordant with this evidence and might even encourage formation of such complexes. Under this model, SARS-CoV-2's strong mitochondrial localization signal relative to other coronaviruses may even contribute to its similarly high infectivity by increasing its efficacy in generating and importing materials into these DMV structures.

Another possible interpretation of these localization results is that previously studied viral protein localizations are actually driven by transcript-level localizations, a mechanism that is highly prevalent for proteins in normal human cells (Blower, 2013). Protein-protein interaction studies performed on SARS-CoV-2 have found that its NSP5 (within ORF1a), NSP13 (within ORF1b), ORF6, and ORF10 proteins interact with host proteins that predominantly localize to nuclear compartments (Gordon et al., 2020). The same study found that the ORF9b protein, produced by the "N" sgRNA, interacts with TOMM70, a mitochondrial import receptor that plays a critical role in modulating interferon response – a key anti-viral cellular defense pathway (Liu et al., 2010). In both cases, localized viral transcripts could be driving viral protein localization, enabling more focused protein-protein interactions.

Additional protein localization patterns appear within SARS-CoV-2's more thoroughly-studied relative, the SARS-CoV coronavirus, responsible for the 2003 SARS outbreak (Ksiazek et al., 2003). The nucleocapsid (N) protein has been shown to dynamically localize to the nucleolus (Cawood et al., 2007). The transcribed protein corresponding to ORF3 of SARS-CoV localizes to both the mitochondria (Yuan et al., 2006) and the nucleolus, causing cell cycle arrest and apoptosis (Yuan et al., 2005). Some have suggested a possible mechanism where translocation of this protein between the nucleus and mitochondria influences the cell's interferon response (Freundt et al., 2009). It is plausible that this protein localization behavior is conserved across both SARS-CoV and its contemporary SARS-CoV-2, especially given their relatively high sequence homology (Lu et al., 2020), and is driven by underlying transcript localization. Indeed, these patterns can also be found much more broadly across viral species. For coronaviruses in general, N protein localization to the host nucleolus appears to be a fairly conserved functional attribute and may play a role in disrupting cell division and cytokinesis (Rawlinson and Moseley, 2015, Wurm et al., 2001). Within the influenza A virus, nonstructural protein 1 localizes to the mitochondria (Tsai et al., 2017). For many viruses that primarily replicate within cytoplasmic regions, multiple studies have identified viral proteins that nonetheless localize to the nucleus to aid replication and disrupt host cell functionality (Hiscox, 2003, Weidman et al., 2003).

A limitation of our work lies in that it applies models trained on human RNA transcript localization data to viral transcripts. It is possible that SARS-CoV-2 infection could alter the host subcellular structures and RNA transport machinery so drastically that our learned localization patterns from human cells no longer hold. If RNA-GPS's predictions turn out to be wrong for this reason, this might suggest that coronavirus infection devastates host cell mRNA trafficking and localization – a previously unrecognized feature of COVID-19 pathobiology. After all, the vast majority of RNA binding proteins in the host cell, which are key drivers of transcript localization, recognize and process RNAs irrespective of whether they are endogenous or foreign, and inability to "properly" localize viral RNAs should mirror a similar breakdown for host cell mRNAs. As we are unable to use existing experimental evidence to thoroughly evaluate and cross-reference the predictions discussed here, future experiments in this vein are clearly necessary. Given the historical scarcity of studies focusing on viral transcript localization, such experiments would likely reveal interesting, crucial insights into viral pathobiology, whether they confirm our specific mitochondrial and nucleolus predictions or not. It is worth pointing out, though, that this is but one of many complex, interconnected viral mechanisms at play.

In summary, we build upon recent computational models of RNA subcellular localization to study, *in silico*, the localization properties of SARS-CoV-2 transcripts. Our results suggest that transcript localization patterns, specifically towards the nucleolus and mitochondrial matrix, may be an important, unique characteristic of SARS-CoV-2 that warrants additional study. We connect these observations to known viral biology regarding DMV structures in viral replication, as well as SARS-CoV-2 protein localization patterns. In doing so, we propose potential cellular mechanisms that underpin viral biology – mechanisms that warrant experiments validating their accuracy, and perhaps even their potential as therapeutic targets. More broadly, we hope that our study helps define a framework for applying machine learning models to enable focused hypothesis generation, enabling similar studies that leverage data science to rapidly respond to emerging epidemiological challenges.

Acknowledgements

We thank the members of the Chang and Zou laboratories for helpful discussions. H.Y.C. is supported by RM1-HG007735 and R01- HG004361. H.Y.C. is an Investigator of the Howard Hughes Medical Institute. J.Z. is supported by NSF CCF 1763191, NIH R21 MD012867-01, NIH P30AG059307, NIH U01MH098953 and grants from the Silicon Valley Foundation and the Chan-Zuckerberg Initiative. F.M.F. is supported by an NIH K99/R00 award from NHGRI (HG010910).

Author Contributions

H.Y.C. and J.Z. conceived the idea for this project and supervised its execution. K.E.W. gathered, preprocessed, and analyzed data for this project with input from all authors. F.M.F. and K.R.P. contributed analysis of mitochondrial APEX-seq data with input from all authors. All authors contributed to interpreting localization results in the context of coronavirus biology. K.E.W. wrote the manuscript with input from all authors.

Declaration of Interests

K.R.P. is a consultant for Maze Therapeutics. H.Y.C. is affiliated with Accent Therapeutics, Boundless Bio, 10x Genomics, Arsenal Bio, and Spring Discovery.

Main Figure Title and Legends

Figure 1: Depictions of the SARS-CoV-2 genome (A), the eight compartments that RNA-GPS predicts transcript localization to (B), and the predicted localizations for SARS-CoV-2 sgRNAs (C, D) and its 5'/CDS/3' sequence segments (E). The SARS-CoV-2 genome produces a series of sub-genomic RNAs (sgRNAs), each encoding one or more genes/proteins (A). These sgRNAs share a common leader 5' sequence and a common trailing 3' UTR sequence (arrow blocks). For each sgRNA, RNA-GPS predicts localization to each compartment in (B) (figure reproduced from (Wu et al., 2020b)), the results of which are shown in (C). This heatmap shows rank scores, indicating how strongly each sgRNA (rows) localizes to each compartment (columns), compared to endogenous human transcripts localizing to that compartment. Colors directly correlate with rank scores; color scale is shared across all heatmaps. Most sgRNAs share similar localization patterns, exhibiting statistically significant enrichment towards the mitochondrial matrix and nucleolus (see Table S2). We also computed these rank scores against a baseline of other coronavirus localization signals (D). SARS-CoV-2 exhibits a stronger mitochondrial matrix localization signal than most other coronaviruses, along with greater overall nuclear localization, particularly at the nucleolus. For context, coronaviruses generally localize to the nucleolus, mitochondrial matrix, and ER membrane (see Figure S1). These predictions are also consistent across different models (see Figure S2) and the negative-strand SARS-CoV-2 sgRNA precursors (see Figure S4). (E) Shows the predicted localization rank scores for shared 5' and 3' segments, and an averaged localization rank score for the variable coding segments. Just on their own, the short ~90-250 base pair 5' and 3' segments carry mitochondrial and nucleolar localization signals.

Figure 2: Validation of SARS-CoV-2 localization predictions. (A) RNA-GPS predictions for the human cytomegalovirus β2.7 transcript, which has been shown to localize to the inner mitochondrial membrane. RNA-GPS correctly predicts its localization to the closest compartment it has been trained on – the mitochondrial matrix. This provides support that RNA-GPS can make accurate predictions on viral RNA. (B) To evaluate the effect of the potentially noisy mitochondrial examples in our APEX-seq training set on predicted SARS-CoV-2 localizations, we trained a denoised variant of RNA-GPS on a subsetted dataset that excludes these examples. This denoised model predicts the same localization pattern for the 3 segments of the SARS-CoV-2 sgRNAs (compare to Figure 1E). For additional validation experiments, see Figure S3A/B.

STAR Methods

Key Resources Table

REAGENT or RESOURCE	SOURCE	Identifier
RNA-GPS model and SARS-CoV-2 analysis code	(Wu et al., 2020b)	https://github.com/wukevin/rnagps
Coronavirus (including SARS-CoV-2) genome sequences	NCBI GenBank	Various (see query strings in covid19/baseline.py and covid19/covid19.py source code files)
Human cytomegalovirus genome sequence	NCBI GenBank	NC_006273.2
APEX-seq data	(Fazal et al., 2019)	GSE116008
RNA binding protein motif database	(Ray et al., 2013)	MEME Motif Databases

Resource Availability

Lead Contact

Further information and requests for resources should be directed to and will be fulfilled by the Lead Contact, Howard Chang (howchang@stanford.edu).

Materials Availability

This computational study did not generate or use new reagents.

Data and Code Availability

The data supporting the findings of this study are all available within publicly available repositories as listed in the Key Resources Table. All code required to query and download viral sequences, as well as to reproduce results and figures can be found within the GitHub repository listed in the Key Resources Table. All software dependencies for RNA-GPS and the SARS-CoV-2 analysis described herein are freely available as well. Within the GitHub repository, most code pertaining to SARS-CoV-2 analysis can be found under the "covid19" folder; other folders contain supporting data and source code.

Method Details

Obtaining viral genomes

SARS-CoV-2 viral genomes were programmatically queried from the NCBI GenBank online database using the BioPython library's Entrez module (Cock et al., 2009). The exact query sequence used can be found within the "covid19/covid19.py" file in the GitHub repository. Returned results were then filtered to retain only assemblies that included annotated, named sgRNA "genes." We consider the sgRNAs corresponding to ORF1ab, S, ORF3a, E, M, ORF6, ORF7a, ORF7b, ORF8, N, and ORF10, as these have the most consistent annotations. In cases where the shared 5' leader sequence or the 3' tail were not explicitly annotated, their regions were inferred to be the 5' and 3' trailing bases outside of any coding regions, respectively. As there are many SARS-CoV-2 genome assemblies that fit these criteria, localization predictions are averaged across all genomes.

Viral genomes constituting the coronavirus baseline follow an identical process, save for using a different NCBI GenBank query sequence that specifically fetches matches to the six coronaviruses known to infect humans (excluding SARS-CoV-2): 229E, NL63, OC43, HKU1, MERS-CoV (beta coronavirus that causes Middle East Respiratory Syndrome, or MERS), and SARS-CoV (the beta coronavirus that causes severe acute respiratory syndrome, or SARS) (Su et al., 2016). The exact query sequence used can be found in the "covid19/baseline.py" source file in the GitHub repository. A detailed breakdown of the exact number of genomes we use from each strain is in Table S1.

The human cytomegalovirus was chosen for additional evaluation based on a systematic literature review of viral RNA localization studies. This is the only example we found that associates a specific viral transcript with a consistent experimentally validated localization. The viral sequence for validation of our model predictions was obtained from the NCBI GenBank reference sequence NC_006273.2. Due to lack of standardized 5' and 3' UTR region annotations for this transcript (despite these being referenced in the literature), we manually determined these regions after reviewing literature and the overall genome annotation.

Sequence featurization and predictive models

RNA-GPS uses k-mer featurization with k = 3, 4, 5, applied independently to the 5' untranslated region (UTR), coding sequence (CDS), and 3' UTR parts of the transcript (Wu et al., 2020b). This creates a feature space of $(4^3 + 4^4 + 4^5) \times 3 = 1344 \times 3 = 4032$ dimensions. These features are then consumed by a random forest model (using the scikit-learn Python library) to generate localization predictions. Extending this definition to the coronavirus sgRNA sequences, we consider the shared 5' leader sequence the fixed 5' UTR input to our model, shared 3' UTR sequence the fixed 3' UTR input to our model, and the variable sgRNA sequence the "CDS" input. For sake of consistency with sgRNA transcript mechanisms, this "CDS" sequence includes the current reading frame, along with any 3' downstream bases until the shared 3' UTR region begins. Each sgRNA is individually assigned predicted localizations. RNA-GPS's per-segment featurization also enables the per-segment localization analysis. For this, we selectively provide the model with only features that correspond to a single segment (i.e. the 5' UTR, CDS, or 3' UTR), with zero values for other features.

For the deep recurrent model, we implemented and trained a recurrent neural network that consumes raw bases as input, maps these to a 32-dimensional embedding layer, passes these through two 64-dimensional gated recurrent units (GRU), and finally a fully-connected layer with sigmoid activation producing 8 localization predictions. This flavor of GRU network is popular in sequence modelling and uses "gating" mechanisms to improve learning of longer-range sequence dependencies (Chung et al., 2014). The model was implemented in PyTorch and was trained to minimize a binary cross-entropy loss using the

Adam optimizer (Kingma and Ba, 2014) with a batch size of 1, with early stopping based on validation set area under the receiver operating characteristic (AUROC).

Training data for predictive models

Both RNA-GPS and the GRU model are trained and tuned on the same APEX-seq data, measuring localization within HEK293T cells (Fazal et al., 2019). Localization is expressed as an enrichment score compared to the rest of the cell. We consider transcripts that exhibit significant enrichment (log fold change > 0 and adjusted p-value ≤ 0.05) for at least one of the eight measured compartments (n = 3660). Many transcripts contain more than one significant localization. We use data splits of 80% train (n = 2928), 10% validation (n = 366), and 10% train (n = 366). As is conventional, the validation set was used for hyperparameter tuning and model architecture tuning.

When removing potentially spurious mitochondrial examples, we start with the above dataset and remove all transcripts that localize to the mitochondrial matrix but have a log fold change in the bottom 20th percentile of localized mitochondrial matrix. This removes the bottom 20% of mitochondrial sequences with the lowest enrichment relative to the rest of the cell (n = 61) – this denoised dataset contains 240 mitochondrial matrix transcripts instead of 301, and a total of 3599 transcripts compared to 3660 previously.

RNA binding protein motif identification and ablation

We use a database of 102 RNA binding protein binding motifs (Ray et al., 2013). To identify matches, we use the same methodology as was used in the RNA-GPS manuscript (Wu et al., 2020b). Briefly, we start with the position weight matrix (PWM) that describes the motif, adjust its probabilities to account for the background nucleotide composition of each transcript sequence, define a cutoff score slightly lower than the maximum achievable log-likelihood for that PWM, and identify any subsequences that exceed that cutoff.

When ablating these PWMs, we use the same methodology for identifying hits, and subsequently replace all hits with “N” bases, re-featurizing the ablated sequence as necessary before feeding into the model, thus generating the ablated localization predictions.

Quantification and Statistical Analysis

Baseline construction and rank score

Baseline distributions are constructed by running a set of baseline transcript sequences through a model predicting transcript localization. For each individual model, there is a per-localization baseline derived from human APEX-seq measurements, and one derived from human coronaviruses excluding SARS-CoV-2. For each localization compartment within the human baseline, we consider only transcripts that exhibit significant localization to that compartment, as defined by having a logFC > 0 and adjusted p-value ≤ 0.05 when running differential expression analysis against the remainder of the cell. Additionally, we only use transcripts not used for model training/tuning (i.e. the test data split), as this most closely approximates what the model would predict when presented with novel sequences. For the coronavirus baseline, we do not have systematically measured localization data, so we do not filter as such. However, we make slight adjustments to the process of calculating the rank score (see below). Note that due to these differences, the values produced by these two baselines are not directly comparable.

For these baselines, we define a rank score as the proportion of baseline values that a SARS-CoV-2 sgRNA localization prediction exceeds. A hypothetical value of 0.5 would correspond to a median, 0.25 would correspond to the first quartile, etc.; rank score is thus bound between 0 and 1 (inclusive). Note that this rank is calculated for each individual compartment separately, as the baselines themselves are compartment specific. For the coronavirus baseline, each SARS-CoV-2 sgRNA is also compared only to homologous sgRNAs from other coronaviruses. For example, the spike protein's localization prediction is only compared against localization predictions of other coronavirus spike proteins. This limits our comparison to the set of genes with easily traceable homology across human coronaviruses, namely ORF1ab, spike (S), envelope (E), membrane (M), and nucleocapsid (N) (Woo et al., 2010). As previously discussed, localization predictions are averaged across all valid SARS-CoV-2 genomes prior to calculating rank scores.

Significance test for localization

In addition to computing the rank scores described above, we also evaluate whether these rank scores correspond to significantly increased localization. To do this, we evaluate the underlying predicted localization probabilities (not the rank scores) against a "null" distribution of localization probabilities for human transcripts exhibiting no significant localization using a one-sided Wilcoxon rank-sum test (scipy Python package (Virtanen et al., 2020)). Our data satisfies the Wilcoxon rank-sum test's assumptions of independence, and our localization predictions are naturally ordinal. To counteract the fact that we do multiple comparisons, we use the Holm method (statsmodels Python package (Seabold and Perktold, 2010)) to correct the resultant p-values.

Gene ontology enrichment analysis

To perform gene ontology enrichment analysis, we used the PANTHER tool (Mi et al., 2018) provided by the Gene Ontology Consortium (Ashburner et al., 2000, The Gene Ontology, 2019). Genes were compared in an overrepresentation test against a reference list of all genes in the Homo sapiens database using Fisher's Exact test, with false discovery rate correction. The annotation used was "Reactome version 65."

Plotting

All plots were generated using a combination of seaborn and matplotlib Python packages (Hunter, 2007).

Supplementary Tables & Figure Legends

Strain	**Count**	**Proportion**
Human coronavirus NL63	48	0.25
Human coronavirus 229E	22	0.12
Human coronavirus OC43	82	0.43
Human coronavirus HKU1	3	0.02
MERS coronavirus	20	0.10
SARS coronavirus (2003)	16	0.08
Total	**191**	**1.00**

Table S1 (related to STAR Methods): Viral strains comprising the human coronavirus baseline. The strains NL63, 229E, OC43, and HKU1 historically commonly infect humans worldwide, while the MERS and SARS coronavirus have been recently responsible for more severe outbreaks.

	ER membrane	Nuclear lamina	Mito matrix	Cytosol	Nucleolus	Nucleus	Nuclear pore	Outer mito membrane
orf1ab	**1.38E-05**	1.00	**4.13E-32**	1.00	**4.86E-24**	1.00	1.00	1.00
s	**3.54E-05**	1.00	**8.01E-41**	1.00	**1.04E-37**	1.00	1.00	1.00
orf3a	1.00	1.00	**9.08E-61**	1.00	**5.89E-36**	1.00	1.00	1.00
e	8.91E-01	**2.86E-13**	**3.11E-65**	1.00	**1.38E-22**	1.00	1.00	1.00
m	**1.23E-03**	**2.14E-05**	**8.72E-53**	1.00	**1.73E-21**	1.00	1.00	1.00
orf6	1.00	**7.60E-09**	**1.31E-47**	1.00	**3.47E-25**	1.00	1.00	1.00
orf7a	5.92E-02	1.00	**1.01E-54**	1.00	**4.26E-25**	1.00	1.00	1.00
orf8	2.62E-01	7.38E-01	**2.56E-61**	1.00	**7.50E-26**	1.00	1.00	1.00
n	1.00	1.00	**7.83E-64**	1.00	**1.15E-22**	1.00	6.10E-01	1.00
orf10	1.00	**6.84E-03**	**5.04E-85**	1.00	**5.17E-05**	1.00	**7.26E-08**	1.00
orf7b	1.00	8.94E-01	**6.77E-09**	1.00	**2.85E-05**	1.00	1.00	1.00

Table S2 (related to Figure 1): Wilcoxon rank-sum test p-values comparing SARS-CoV-2 sgRNAs' localization predictions against those of unlocalized transcripts. All p-values are Holm-adjusted. Values that exceed our significance cutoff of 0.05 are in bold. The SARS-CoV-2 sgRNA localization predictions towards the mitochondrial matrix and nucleolus both have consistently significant p-values, indicating that their predictions are significantly higher than that of unlocalized transcripts (for the respective compartment), suggesting significant predicted localization.

Figure S1 (related to Figure 1): Summary of localization patterns aggregated across all transcripts comprising the human coronavirus baseline. We see that coronaviruses in general primarily exhibit localization towards the nucleolus, mitochondrial matrix, and ER membrane – a pattern similar to that seen in SARS-CoV-2's sgRNAs (albeit less dramatic).

Figure S2 (related to Figure 1): Heatmaps of rank scores of SARS-CoV-2 localization predictions, relative to localized human transcripts (A) and other coronavirus genomes (B), according to a deep-learning recurrent model (GRU). This model takes a very different computational approach to predicting localization compared to RNA-GPS, and thus serves as an orthogonal computational support of results covered in our primary figures. (A) Recapitulates that mitochondrial matrix and nucleolus are among the two most prominent localization signals for SARS-CoV-2. (B) Recapitulates that compared to other coronaviruses, SARS-CoV-2 generally exhibits a stronger nuclear localization signal.

Figure S3 (related to Figure 2): Further analysis of mitochondrial transcripts used to train RNA-GPS. Within the APEX-seq training data, many transcripts localized at the mitochondrial matrix are actually encoded within the nucleus. (A) Shows a plot of enrichment scores at each compartment for these mitochondrial-enriched, nuclear-encoded "non-canonical" transcripts. We see that these transcripts have enrichment centered around 0 for all but the mitochondrial matrix, indicating that while these transcripts are nuclear-encoded, the APEX-seq labelling technology consistently and nonrandomly associates them with the mitochondrial matrix. These transcripts are also biologically meaningful; (B) shows reactome ontology analysis of 100 most enriched (by p-value) non-canonical mitochondrial matrix transcripts. There is a clear emphasis for cytoskeletal and intracellular transport terms (e.g. kinesins, post-chaperonin tubulin folding pathway, recruitment of NuMA to mitotic centrosomes; adjusted $p < 0.05$). This suggests that the non-canonical transcripts might be consistently picked up as the APEX-seq protein is itself trafficked to the mitochondria.

Figure S4 (related to Figure 1): Localization of negative strand sgRNA precursors. Figure 1C shows that the positive strand sgRNA transcripts tend to exhibit localization towards the mitochondrial matrix and nucleolus. Here, we look at the negative-strand precursors to those sgRNAs and observe that these transcripts share similar mitochondrial matrix and nucleolus localization patterns. This suggests another layer of conservation of this localization signal.

References

ASHBURNER, M., BALL, C. A., BLAKE, J. A., BOTSTEIN, D., BUTLER, H., CHERRY, J. M., DAVIS, A. P., DOLINSKI, K., DWIGHT, S. S., EPPIG, J. T., HARRIS, M. A., HILL, D. P., ISSEL-TARVER, L., KASARSKIS, A., LEWIS, S., MATESE, J. C., RICHARDSON, J. E., RINGWALD, M., RUBIN, G. M. & SHERLOCK, G. 2000. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet,* 25**,** 25-9.

BECKER, J. T. & SHERER, N. M. 2017. Subcellular Localization of HIV-1 <em>gag-pol</em> mRNAs Regulates Sites of Virion Assembly. *Journal of Virology,* 91**,** e02315-16.

BLANCHARD, E. & ROINGEARD, P. 2015. Virus-induced double-membrane vesicles. *Cell Microbiol,* 17**,** 45-50.

BLOWER, M. D. 2013. Chapter One - Molecular Insights into Intracellular RNA Localization. *In:* JEON, K. W. (ed.) *International Review of Cell and Molecular Biology.* Academic Press.

BUXBAUM, A. R., HAIMOVICH, G. & SINGER, R. H. 2015. In the right place at the right time: visualizing and understanding mRNA localization. *Nat Rev Mol Cell Biol,* 16**,** 95-109.

CAWOOD, R., HARRISON, S. M., DOVE, B. K., REED, M. L. & HISCOX, J. A. 2007. Cell Cycle Dependent Nucleolar Localization of the Coronavirus Nucleocapsid Protein. *Cell Cycle,* 6**,** 863-867.

CHIN, A. & LÉCUYER, E. 2017. RNA localization: Making its way to the center stage. *Biochimica et Biophysica Acta (BBA) - General Subjects,* 1861**,** 2956-2970.

CHUNG, J., GULCEHRE, C., CHO, K. & BENGIO, Y. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling.

CLEMENTZ, M. A., KANJANAHALUETHAI, A., O'BRIEN, T. E. & BAKER, S. C. 2008. Mutation in murine coronavirus replication protein nsp4 alters assembly of double membrane vesicles. *Virology,* 375**,** 118-129.

COCK, P. J. A., ANTAO, T., CHANG, J. T., CHAPMAN, B. A., COX, C. J., DALKE, A., FRIEDBERG, I., HAMELRYCK, T., KAUFF, F., WILCZYNSKI, B. & DE HOON, M. J. L. 2009. Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics,* 25**,** 1422-1423.

COORDINATORS, N. R. 2018. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res,* 46**,** D8-D13.

FAZAL, F. M., HAN, S., PARKER, K. R., KAEWSAPSAK, P., XU, J., BOETTIGER, A. N., CHANG, H. Y. & TING, A. Y. 2019. Atlas of Subcellular RNA Localization Revealed by APEX-Seq. *Cell,* 178**,** 473-490.e26.

FREUNDT, E. C., YU, L., PARK, E., LENARDO, M. J. & XU, X.-N. 2009. Molecular Determinants for Subcellular Localization of the Severe Acute Respiratory Syndrome Coronavirus Open Reading Frame 3b Protein. *Journal of Virology,* 83**,** 6631-6640.

FUNG, T. S. & LIU, D. X. 2014. Coronavirus infection, ER stress, apoptosis and innate immunity. *Frontiers in Microbiology,* 5.

GORDON, D. E., JANG, G. M., BOUHADDOU, M., XU, J., OBERNIER, K., O'MEARA, M. J., GUO, J. Z., SWANEY, D. L., TUMMINO, T. A., HUETTENHAIN, R., KAAKE, R. M., RICHARDS, A. L., TUTUNCUOGLU, B., FOUSSARD, H., BATRA, J., HAAS, K., MODAK, M., KIM, M., HAAS, P., POLACCO, B. J., BRABERG, H., FABIUS, J. M., ECKHARDT, M., SOUCHERAY, M., BENNETT, M. J., CAKIR, M., MCGREGOR, M. J., LI, Q., NAING, Z. Z. C., ZHOU, Y., PENG, S., KIRBY, I. T., MELNYK, J. E., CHORBA, J. S., LOU, K., DAI, S. A., SHEN, W., SHI, Y., ZHANG, Z., BARRIO-HERNANDEZ, I., MEMON, D., HERNANDEZ-ARMENTA, C., MATHY, C. J. P., PERICA, T., PILLA, K. B., GANESAN, S. J., SALTZBERG, D. J., RAMACHANDRAN, R., LIU, X., ROSENTHAL, S. B., CALVIELLO, L., VENKATARAMANAN, S., LIBOY-LUGO, J., LIN, Y., WANKOWICZ, S. A., BOHN, M., SHARP, P. P., TRENKER, R., YOUNG, J. M., CAVERO, D. A., HIATT, J., ROTH, T. L., RATHORE, U., SUBRAMANIAN, A., NOACK, J., HUBERT, M., ROESCH, F., VALLET, T., MEYER, B., WHITE, K. M., MIORIN, L., ROSENBERG, O. S., VERBA, K. A., AGARD, D., OTT, M., EMERMAN, M., RUGGERO, D., GARCÍA-SASTRE, A., JURA, N., VON ZASTROW, M., TAUNTON, J., ASHWORTH, A., SCHWARTZ, O., VIGNUZZI, M., D'ENFERT, C., MUKHERJEE, S., JACOBSON, M., MALIK, H. S., FUJIMORI, D. G., IDEKER, T., CRAIK, C. S., FLOOR, S., FRASER, J. S., GROSS, J., SALI, A., KORTEMME, T., BELTRAO, P., SHOKAT, K., SHOICHET, B. K. & KROGAN, N. J. 2020. A SARS-CoV-2-Human Protein-Protein Interaction Map Reveals Drug Targets and Potential Drug-Repurposing. *bioRxiv***,** 2020.03.22.002386.

HAGEMEIJER, M. C., VONK, A. M., MONASTYRSKA, I., ROTTIER, P. J. & DE HAAN, C. A. 2012. Visualizing coronavirus RNA synthesis in time by using click chemistry. *J Virol,* 86**,** 5808-16.

HISCOX, J. A. 2003. The interaction of animal cytoplasmic RNA viruses with the nucleus to facilitate replication. *Virus Res,* 95**,** 13-22.

HUNTER, J. D. 2007. Matplotlib: A 2D Graphics Environment. *Computing in Science & Engineering,* 9**,** 90-95.

KIM, D., LEE, J. Y., YANG, J. S., KIM, J. W., KIM, V. N. & CHANG, H. 2020. The Architecture of SARS-CoV-2 Transcriptome. *Cell,* 181**,** 914-921.e10.

KINGMA, D. & BA, J. 2014. Adam: A Method for Stochastic Optimization. *International Conference on Learning Representations*.

KNOOPS, K., KIKKERT, M., WORM, S. H. E. V. D., ZEVENHOVEN-DOBBE, J. C., VAN DER MEER, Y., KOSTER, A. J., MOMMAAS, A. M. & SNIJDER, E. J. 2008. SARS-Coronavirus Replication Is Supported by a Reticulovesicular Network of Modified Endoplasmic Reticulum. *PLOS Biology,* 6**,** e226.

KSIAZEK, T. G., ERDMAN, D., GOLDSMITH, C. S., ZAKI, S. R., PERET, T., EMERY, S., TONG, S., URBANI, C., COMER, J. A., LIM, W., ROLLIN, P. E., DOWELL, S. F., LING, A. E., HUMPHREY, C. D., SHIEH, W. J., GUARNER, J., PADDOCK, C. D., ROTA, P., FIELDS, B., DERISI, J., YANG, J. Y., COX, N., HUGHES, J. M., LEDUC, J. W., BELLINI, W. J., ANDERSON, L. J. & GROUP, S. W. 2003. A novel coronavirus associated with severe acute respiratory syndrome. *N Engl J Med,* 348**,** 1953-66.

LIU, X.-Y., WEI, B., SHI, H.-X., SHAN, Y.-F. & WANG, C. 2010. Tom70 mediates activation of interferon regulatory factor 3 on mitochondria. *Cell Research,* 20**,** 994-1011.

LU, R., ZHAO, X., LI, J., NIU, P., YANG, B., WU, H., WANG, W., SONG, H., HUANG, B., ZHU, N., BI, Y., MA, X., ZHAN, F., WANG, L., HU, T., ZHOU, H., HU, Z., ZHOU, W., ZHAO, L., CHEN, J., MENG, Y., WANG, J., LIN, Y., YUAN, J., XIE, Z., MA, J., LIU, W. J., WANG, D., XU, W., HOLMES, E. C., GAO, G. F., WU, G., CHEN, W., SHI, W. & TAN, W. 2020. Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding. *The Lancet,* 395**,** 565-574.

MI, H., MURUGANUJAN, A., EBERT, D., HUANG, X. & THOMAS, P. D. 2018. PANTHER version 14: more genomes, a new PANTHER GO-slim and improvements in enrichment analysis tools. *Nucleic Acids Research,* 47**,** D419-D426.

MINAKSHI, R., PADHAN, K., RANI, M., KHAN, N., AHMAD, F. & JAMEEL, S. 2009. The SARS Coronavirus 3a Protein Causes Endoplasmic Reticulum Stress and Induces Ligand-Independent Downregulation of the Type 1 Interferon Receptor. *PLOS ONE,* 4**,** e8342.

NAL, B., CHAN, C., KIEN, F., SIU, L., TSE, J., CHU, K., KAM, J., STAROPOLI, I., CRESCENZO-CHAIGNE, B., ESCRIOU, N., VAN DER WERF, S., YUEN, K.-Y. & ALTMEYER, R. 2005. Differential maturation and subcellular localization of severe acute respiratory syndrome coronavirus surface proteins S, M and E. *Journal of General Virology,* 86**,** 1423-1434.

RAWLINSON, S. M. & MOSELEY, G. W. 2015. The nucleolar interface of RNA viruses. *Cellular Microbiology,* 17**,** 1108-1120.

RAY, D., KAZAN, H., COOK, K. B., WEIRAUCH, M. T., NAJAFABADI, H. S., LI, X., GUEROUSSOV, S., ALBU, M., ZHENG, H., YANG, A., NA, H., IRIMIA, M., MATZAT, L. H., DALE, R. K., SMITH, S. A., YAROSH, C. A., KELLY, S. M., NABET, B., MECENAS, D., LI, W., LAISHRAM, R. S., QIAO, M., LIPSHITZ, H. D., PIANO, F., CORBETT, A. H., CARSTENS, R. P., FREY, B. J., ANDERSON, R. A., LYNCH, K. W., PENALVA, L. O.

F., LEI, E. P., FRASER, A. G., BLENCOWE, B. J., MORRIS, Q. D. & HUGHES, T. R. 2013. A compendium of RNA-binding motifs for decoding gene regulation. *Nature,* 499**,** 172-177.

RYDER, P. V. & LERIT, D. A. 2018. RNA localization regulates diverse and dynamic cellular processes. *Traffic,* 19**,** 496-502.

SALVETTI, A. & GRECO, A. 2014. Viruses and the nucleolus: The fatal attraction. *Biochimica et Biophysica Acta (BBA) - Molecular Basis of Disease,* 1842**,** 840-847.

SANCHE, S., LIN, Y. T., XU, C., ROMERO-SEVERSON, E., HENGARTNER, N. & KE, R. 2020. The Novel Coronavirus, 2019-nCoV, is Highly Contagious and More Infectious Than Initially Estimated. *medRxiv***,** 2020.02.07.20021154.

SEABOLD, S. & PERKTOLD, J. 2010. Statsmodels: Econometric and Statistical Modeling with Python. *Proceedings of the 9th Python in Science Conference,* 2010.

SOMASUNDARAN, M., ZAPP, M. L., BEATTIE, L. K., PANG, L., BYRON, K. S., BASSELL, G. J., SULLIVAN, J. L. & SINGER, R. H. 1994. Localization of HIV RNA in mitochondria of infected cells: potential role in cytopathogenicity. *Journal of Cell Biology,* 126**,** 1353-1360.

SU, S., WONG, G., SHI, W., LIU, J., LAI, A. C. K., ZHOU, J., LIU, W., BI, Y. & GAO, G. F. 2016. Epidemiology, Genetic Recombination, and Pathogenesis of Coronaviruses. *Trends Microbiol,* 24**,** 490-502.

THE GENE ONTOLOGY, C. 2019. The Gene Ontology Resource: 20 years and still GOing strong. *Nucleic Acids Res,* 47**,** D330-D338.

TSAI, C.-F., LIN, H.-Y., HSU, W.-L. & TSAI, C.-H. 2017. The novel mitochondria localization of influenza A virus NS1 visualized by FlAsH labeling. *FEBS Open Bio,* 7**,** 1960-1971.

VIRTANEN, P., GOMMERS, R., OLIPHANT, T. E., HABERLAND, M., REDDY, T., COURNAPEAU, D., BUROVSKI, E., PETERSON, P., WECKESSER, W., BRIGHT, J., VAN DER WALT, S. J., BRETT, M., WILSON, J., MILLMAN, K. J., MAYOROV, N., NELSON, A. R. J., JONES, E., KERN, R., LARSON, E., CAREY, C. J., POLAT, İ., FENG, Y., MOORE, E. W., VANDERPLAS, J., LAXALDE, D., PERKTOLD, J., CIMRMAN, R., HENRIKSEN, I., QUINTERO, E. A., HARRIS, C. R., ARCHIBALD, A. M., RIBEIRO, A. H., PEDREGOSA, F., VAN MULBREGT, P., VIJAYKUMAR, A., BARDELLI, A. P., ROTHBERG, A., HILBOLL, A., KLOECKNER, A., SCOPATZ, A., LEE, A., ROKEM, A., WOODS, C. N., FULTON, C., MASSON, C., HÄGGSTRÖM, C., FITZGERALD, C., NICHOLSON, D. A., HAGEN, D. R., PASECHNIK, D. V., OLIVETTI, E., MARTIN, E., WIESER, E., SILVA, F., LENDERS, F., WILHELM, F., YOUNG, G., PRICE, G. A., INGOLD, G.-L., ALLEN, G. E., LEE, G. R., AUDREN, H., PROBST, I., DIETRICH, J. P., SILTERRA, J., WEBBER, J. T., SLAVIČ, J., NOTHMAN, J., BUCHNER, J., KULICK, J., SCHÖNBERGER, J. L., DE MIRANDA CARDOSO, J. V., REIMER, J., HARRINGTON, J., RODRÍGUEZ, J. L. C., NUNEZ-IGLESIAS, J., KUCZYNSKI, J., TRITZ, K., THOMA, M., NEWVILLE, M., KÜMMERER, M., BOLINGBROKE, M., TARTRE, M., PAK, M., SMITH, N. J., NOWACZYK, N., SHEBANOV, N., PAVLYK, O., BRODTKORB, P. A., LEE, P., MCGIBBON, R. T., FELDBAUER, R., LEWIS, S., TYGIER, S., SIEVERT, S., VIGNA, S., PETERSON, S., MORE, S., PUDLIK, T., OSHIMA, T., et al. 2020. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nature Methods,* 17**,** 261-272.

WEIDMAN, M. K., SHARMA, R., RAYCHAUDHURI, S., KUNDU, P., TSAI, W. & DASGUPTA, A. 2003. The interaction of cytoplasmic RNA viruses with the nucleus. *Virus Research,* 95**,** 75-85.

WILLIAMSON, C. D., DEBIASI, R. L. & COLBERG-POLEY, A. M. 2012. Viral product trafficking to mitochondria, mechanisms and roles in pathogenesis. *Infect Disord Drug Targets,* 12**,** 18-37.

WOO, P. C., HUANG, Y., LAU, S. K. & YUEN, K. Y. 2010. Coronavirus genomics and bioinformatics analysis. *Viruses,* 2**,** 1804-20.

WU, F., ZHAO, S., YU, B., CHEN, Y. M., WANG, W., SONG, Z. G., HU, Y., TAO, Z. W., TIAN, J. H., PEI, Y. Y., YUAN, M. L., ZHANG, Y. L., DAI, F. H., LIU, Y., WANG, Q. M., ZHENG, J. J., XU, L., HOLMES, E. C. & ZHANG, Y. Z. 2020a. A new coronavirus associated with human respiratory disease in China. *Nature,* 579**,** 265-269.

WU, H.-Y. & BRIAN, D. A. 2010. Subgenomic messenger RNA amplification in coronaviruses. *Proceedings of the National Academy of Sciences,* 107**,** 12257-12262.

WU, K. E., PARKER, K. R., FAZAL, F. M., CHANG, H. & ZOU, J. 2020b. RNA-GPS predicts high-resolution RNA subcellular localization and highlights the role of splicing. *RNA*.

WURM, T., CHEN, H., HODGSON, T., BRITTON, P., BROOKS, G. & HISCOX, J. A. 2001. Localization to the Nucleolus Is a Common Feature of Coronavirus Nucleoproteins, and the Protein May Disrupt Host Cell Division. *Journal of Virology,* 75**,** 9345-9356.

YUAN, X., SHAN, Y., YAO, Z., LI, J., ZHAO, Z., CHEN, J. & CONG, Y. 2006. Mitochondrial location of severe acute respiratory syndrome coronavirus 3b protein. *Mol Cells,* 21**,** 186-91.

YUAN, X., SHAN, Y., ZHAO, Z., CHEN, J. & CONG, Y. 2005. G0/G1 arrest and apoptosis induced by SARS-CoV 3b protein in transfected cells. *Virol J,* 2**,** 66.

Figure 1

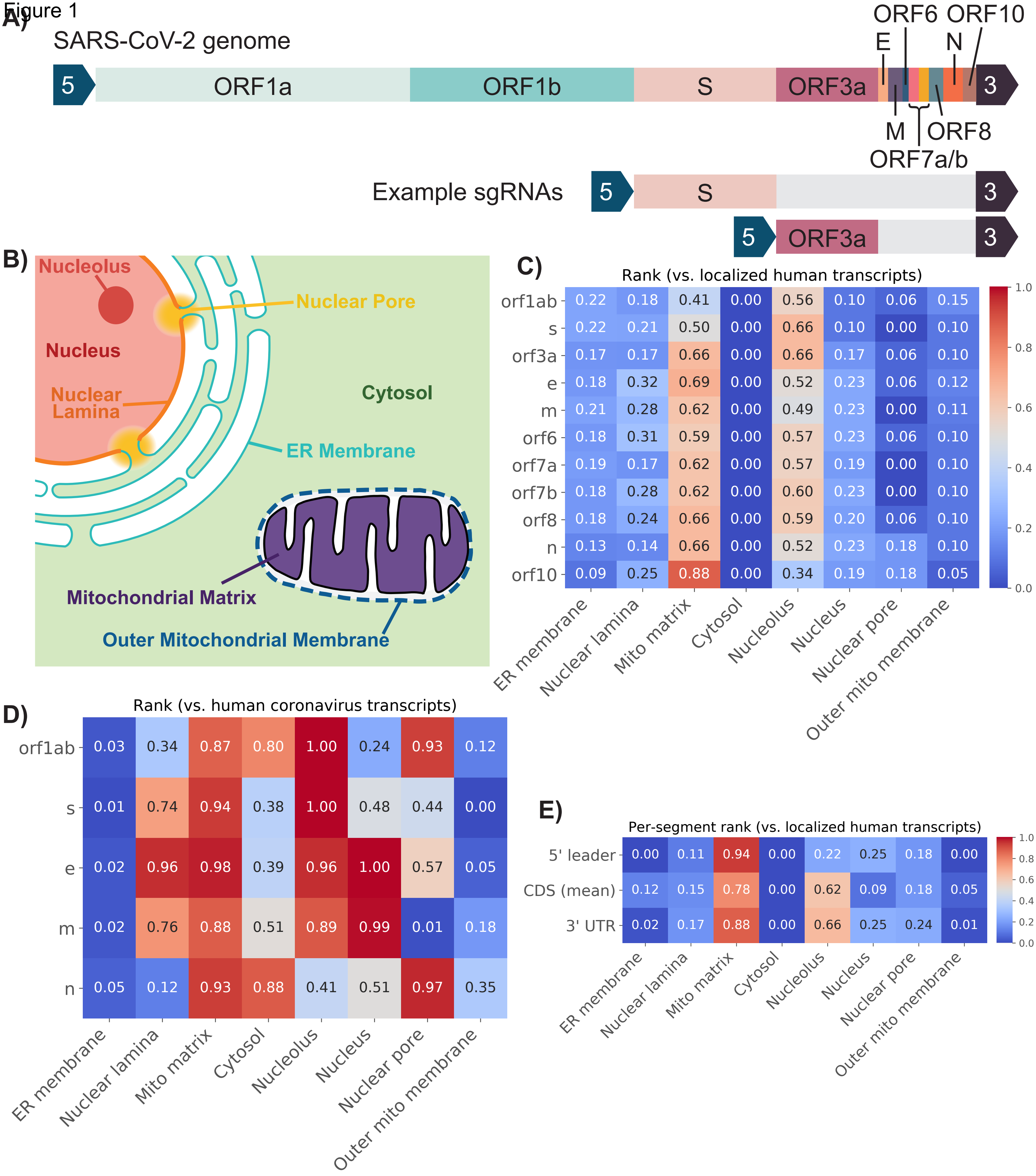
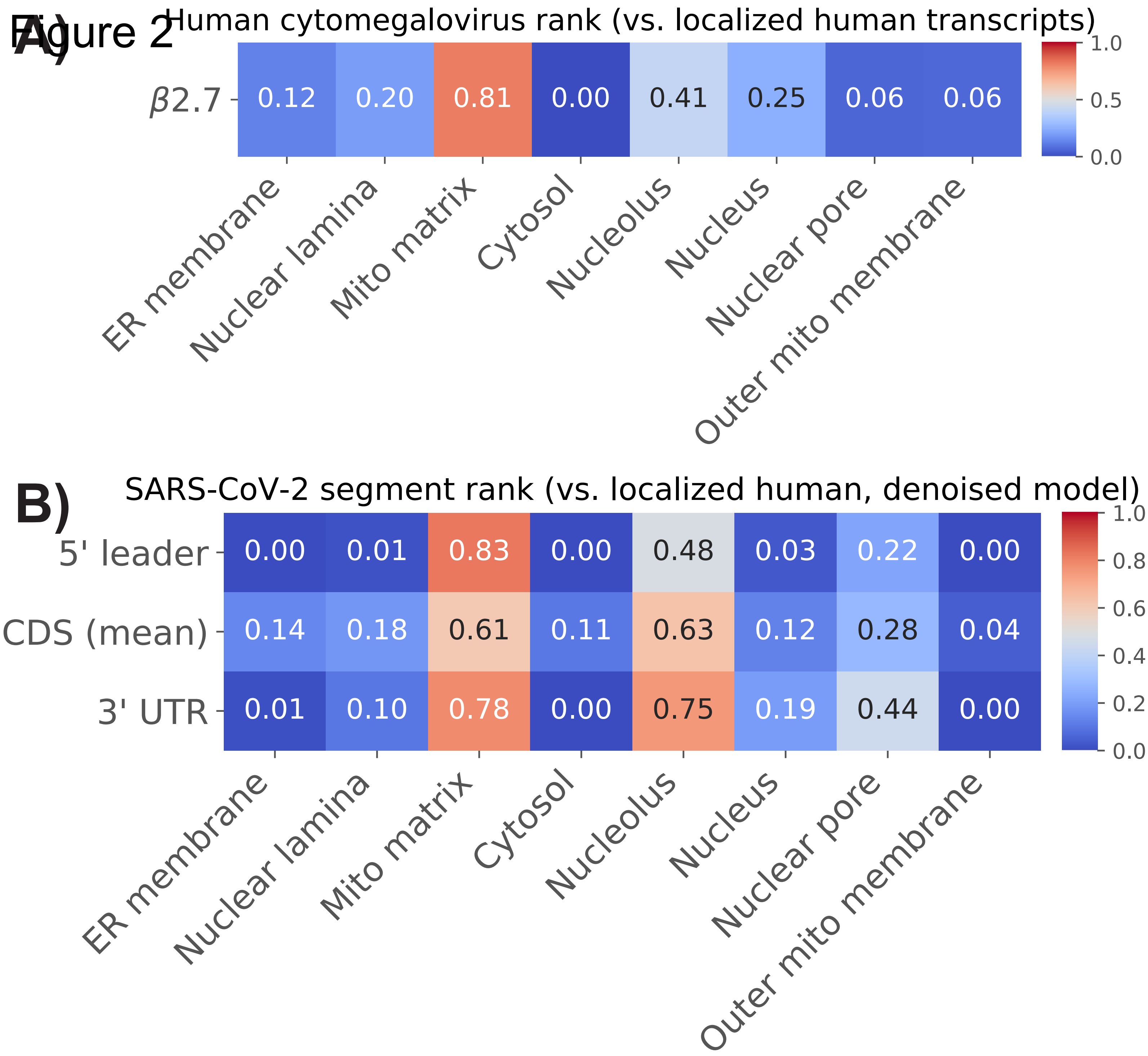

A)
SARS-CoV-2 genome
5
ORF1a
ORF1b
S
ORF3a
E
ORF6
ORF10
N
M
ORF7a/b
ORF8
3
Example sgRNAs
5
S
3
5
ORF3a
3
B)
Nucleolus
Nuclear Pore
Nucleus
Nuclear Lamina
Cytosol
ER Membrane
Mitochondrial Matrix
Outer Mitochondrial Membrane
C)
Rank (vs. localized human transcripts)
ER membrane, Nuclear lamina, Mito matrix, Cytosol, Nucleolus, Nucleus, Nuclear pore, Outer mito membrane
orf1ab 0.22 0.18 0.41 0.00 0.56 0.10 0.06 0.15
s 0.22 0.21 0.50 0.00 0.66 0.10 0.00 0.10
orf3a 0.17 0.17 0.66 0.00 0.66 0.17 0.06 0.10
e 0.18 0.32 0.69 0.00 0.52 0.23 0.06 0.12
m 0.21 0.28 0.62 0.00 0.49 0.23 0.00 0.11
orf6 0.18 0.31 0.59 0.00 0.57 0.23 0.06 0.10
orf7a 0.19 0.17 0.62 0.00 0.57 0.19 0.00 0.10
orf7b 0.18 0.28 0.62 0.00 0.60 0.23 0.00 0.10
orf8 0.18 0.24 0.66 0.00 0.59 0.20 0.06 0.10
n 0.13 0.14 0.66 0.00 0.52 0.23 0.18 0.10
orf10 0.09 0.25 0.88 0.00 0.34 0.19 0.18 0.05
1.0 0.8 0.6 0.4 0.2 0.0
D)
Rank (vs. human coronavirus transcripts)
ER membrane, Nuclear lamina, Mito matrix, Cytosol, Nucleolus, Nucleus, Nuclear pore, Outer mito membrane
orf1ab 0.03 0.34 0.87 0.80 1.00 0.24 0.93 0.12
s 0.01 0.74 0.94 0.38 1.00 0.48 0.44 0.00
e 0.02 0.96 0.98 0.39 0.96 1.00 0.57 0.05
m 0.02 0.76 0.88 0.51 0.89 0.99 0.01 0.18
n 0.05 0.12 0.93 0.88 0.41 0.51 0.97 0.35
E)
Per-segment rank (vs. localized human transcripts)
ER membrane, Nuclear lamina, Mito matrix, Cytosol, Nucleolus, Nucleus, Nuclear pore, Outer mito membrane
5' leader 0.00 0.11 0.94 0.00 0.22 0.25 0.18 0.00
CDS (mean) 0.12 0.15 0.78 0.00 0.62 0.09 0.18 0.05
3' UTR 0.02 0.17 0.88 0.00 0.66 0.25 0.24 0.01
1.0 0.8 0.6 0.4 0.2 0.0

Figure 2


A)
Human cytomegalovirus rank (vs. localized human transcripts)
β2.7
0.12
0.20
0.81
0.00
0.41
0.25
0.06
0.06
1.0
0.5
0.0
ER membrane
Nuclear lamina
Mito matrix
Cytosol
Nucleolus
Nucleus
Nuclear pore
Outer mito membrane
B)
SARS-CoV-2 segment rank (vs. localized human, denoised model)
5' leader
0.00
0.01
0.83
0.00
0.48
0.03
0.22
0.00
CDS (mean)
0.14
0.18
0.61
0.11
0.63
0.12
0.28
0.04
3' UTR
0.01
0.10
0.78
0.00
0.75
0.19
0.44
0.00
1.0
0.8
0.6
0.4
0.2
0.0
ER membrane
Nuclear lamina
Mito matrix
Cytosol
Nucleolus
Nucleus
Nuclear pore
Outer mito membrane

Supplementary Tables & Figures

Strain	**Count**	**Proportion**
Human coronavirus NL63	48	0.25
Human coronavirus 229E	22	0.12
Human coronavirus OC43	82	0.43
Human coronavirus HKU1	3	0.02
MERS coronavirus	20	0.10
SARS coronavirus (2003)	16	0.08
Total	**191**	**1.00**

Table S1 (related to STAR Methods): Viral strains comprising the human coronavirus baseline. The strains NL63, 229E, OC43, and HKU1 historically commonly infect humans worldwide, while the MERS and SARS coronavirus have been recently responsible for more severe outbreaks.

	ER membrane	Nuclear lamina	Mito matrix	Cytosol	Nucleolus	Nucleus	Nuclear pore	Outer mito membrane
orf1ab	**1.38E-05**	1.00	**4.13E-32**	1.00	**4.86E-24**	1.00	1.00	1.00
s	**3.54E-05**	1.00	**8.01E-41**	1.00	**1.04E-37**	1.00	1.00	1.00
orf3a	1.00	1.00	**9.08E-61**	1.00	**5.89E-36**	1.00	1.00	1.00
e	8.91E-01	**2.86E-13**	**3.11E-65**	1.00	**1.38E-22**	1.00	1.00	1.00
m	**1.23E-03**	**2.14E-05**	**8.72E-53**	1.00	**1.73E-21**	1.00	1.00	1.00
orf6	1.00	**7.60E-09**	**1.31E-47**	1.00	**3.47E-25**	1.00	1.00	1.00
orf7a	5.92E-02	1.00	**1.01E-54**	1.00	**4.26E-25**	1.00	1.00	1.00
orf8	2.62E-01	7.38E-01	**2.56E-61**	1.00	**7.50E-26**	1.00	1.00	1.00
n	1.00	1.00	**7.83E-64**	1.00	**1.15E-22**	1.00	6.10E-01	1.00
orf10	1.00	**6.84E-03**	**5.04E-85**	1.00	**5.17E-05**	1.00	**7.26E-08**	1.00
orf7b	1.00	8.94E-01	**6.77E-09**	1.00	**2.85E-05**	1.00	1.00	1.00

Table S2 (related to Figure 1): Wilcoxon rank-sum test p-values comparing SARS-CoV-2 sgRNAs' localization predictions against those of unlocalized transcripts. All p-values are Holm-adjusted. Values that exceed our significance cutoff of 0.05 are in bold. The SARS-CoV-2 sgRNA localization predictions towards the mitochondrial matrix and nucleolus both have consistently significant p-values, indicating that their predictions are significantly higher than that of unlocalized transcripts (for the respective compartment), suggesting significant predicted localization.

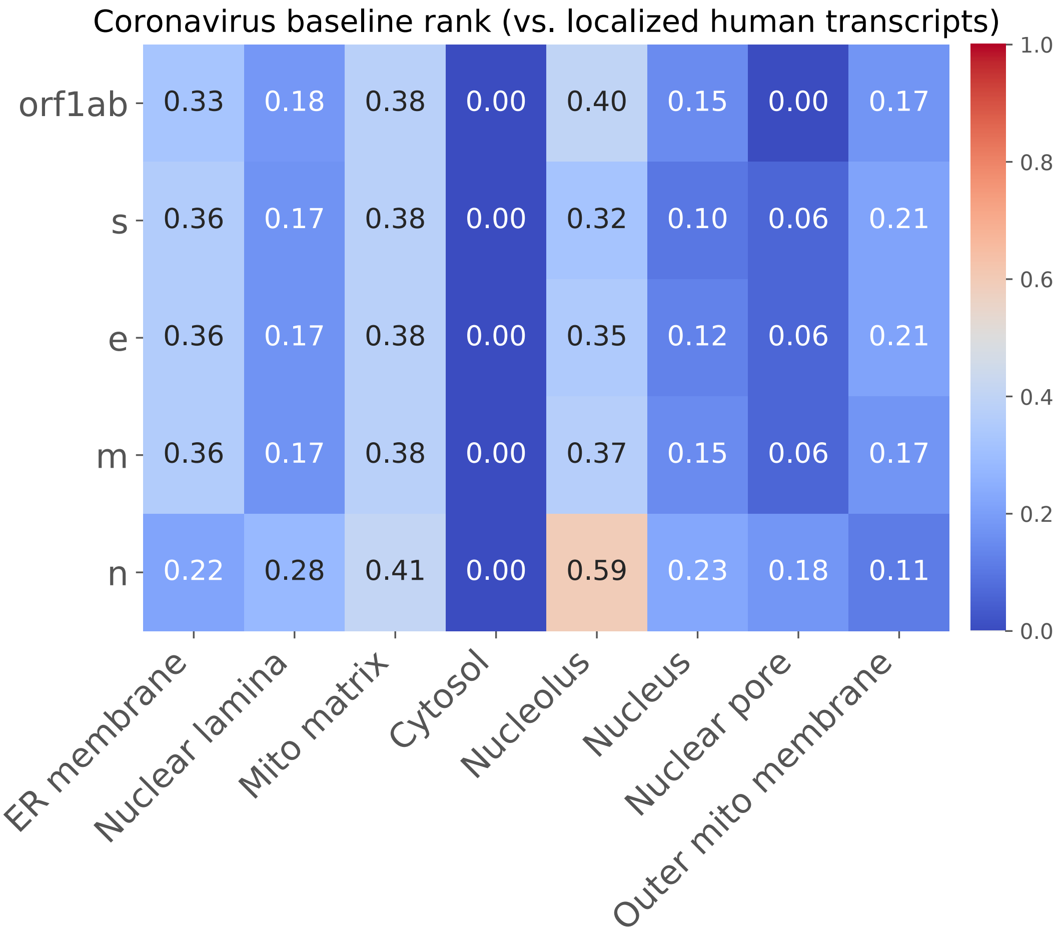


Figure S1 (related to Figure 1): Summary of localization patterns aggregated across all transcripts comprising the human coronavirus baseline. We see that coronaviruses in general primarily exhibit localization towards the nucleolus, mitochondrial matrix, and ER membrane – a pattern similar to that seen in SARS-CoV-2's sgRNAs (albeit less dramatic).

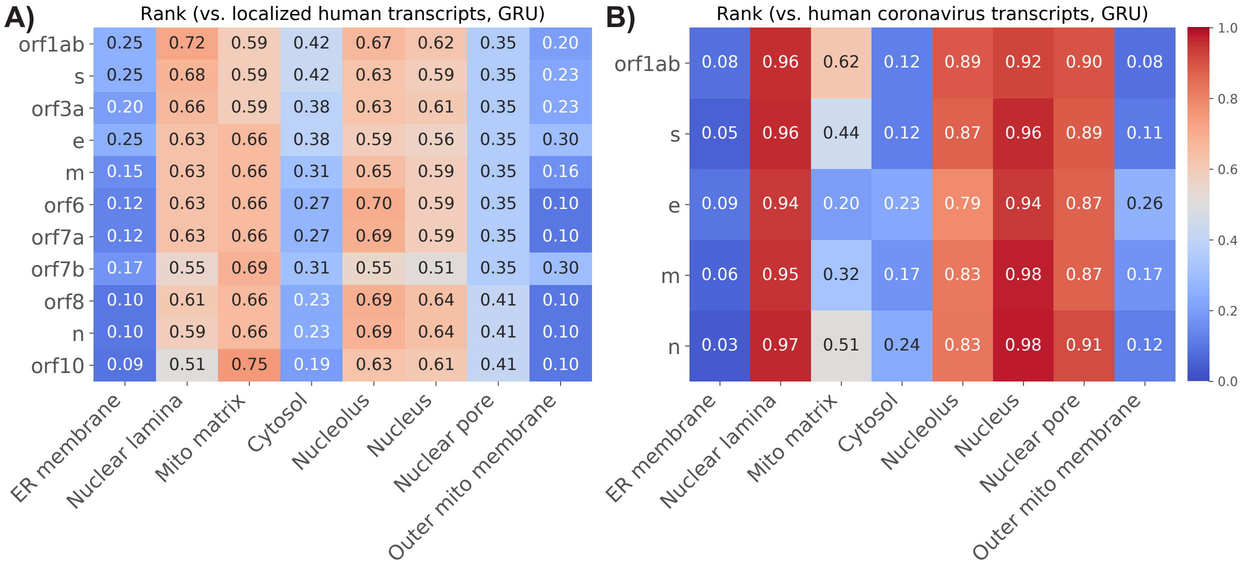


Figure S2 (related to Figure 1): Heatmaps of rank scores of SARS-CoV-2 localization predictions, relative to localized human transcripts (A) and other coronavirus genomes (B), according to a deep-learning recurrent model (GRU). This model takes a very different computational approach to predicting localization compared to RNA-GPS, and thus serves as an orthogonal computational support of results covered in our primary figures. (A) Recapitulates that mitochondrial matrix and nucleolus are among the two most prominent localization signals for SARS-CoV-2. (B) Recapitulates that compared to other coronaviruses, SARS-CoV-2 generally exhibits a stronger nuclear localization signal.

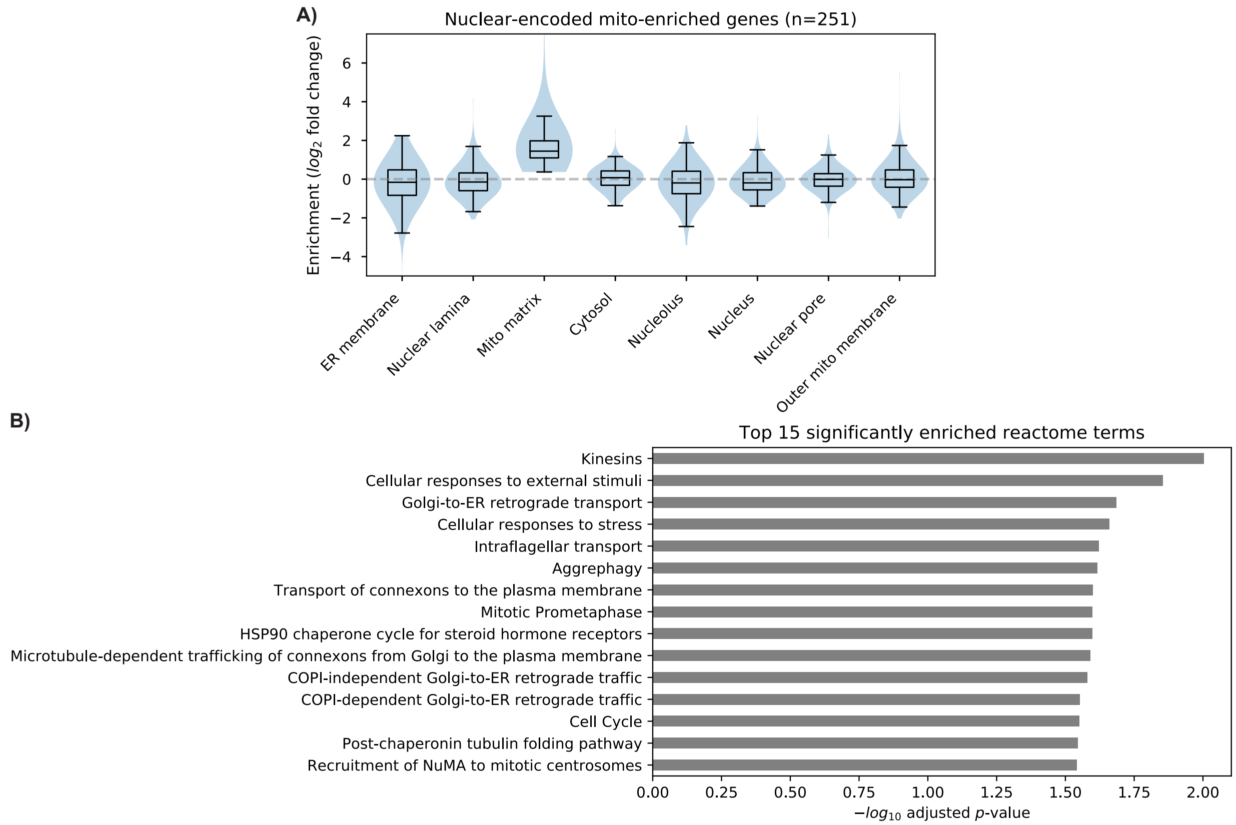


Figure S3 (related to Figure 2): Further analysis of mitochondrial transcripts used to train RNA-GPS. Within the APEX-seq training data, many transcripts localized at the mitochondrial matrix are actually encoded within the nucleus. (A) Shows a plot of enrichment scores at each compartment for these mitochondrial-enriched, nuclear-encoded “non-canonical” transcripts. We see that these transcripts have enrichment centered around 0 for all but the mitochondrial matrix, indicating that while these transcripts are nuclear-encoded, the APEX-seq labelling technology consistently and nonrandomly associates them with the mitochondrial matrix. These transcripts are also biologically meaningful; (B) shows reactome ontology analysis of 100 most enriched (by p-value) non-canonical mitochondrial matrix transcripts. There is a clear emphasis for cytoskeletal and intracellular transport terms (e.g. kinesins, post-chaperonin tubulin folding pathway, recruitment of NuMA to mitotic centrosomes; adjusted $p < 0.05$). This suggests that the non-canonical transcripts might be consistently picked up as the APEX-seq protein is itself trafficked to the mitochondria.

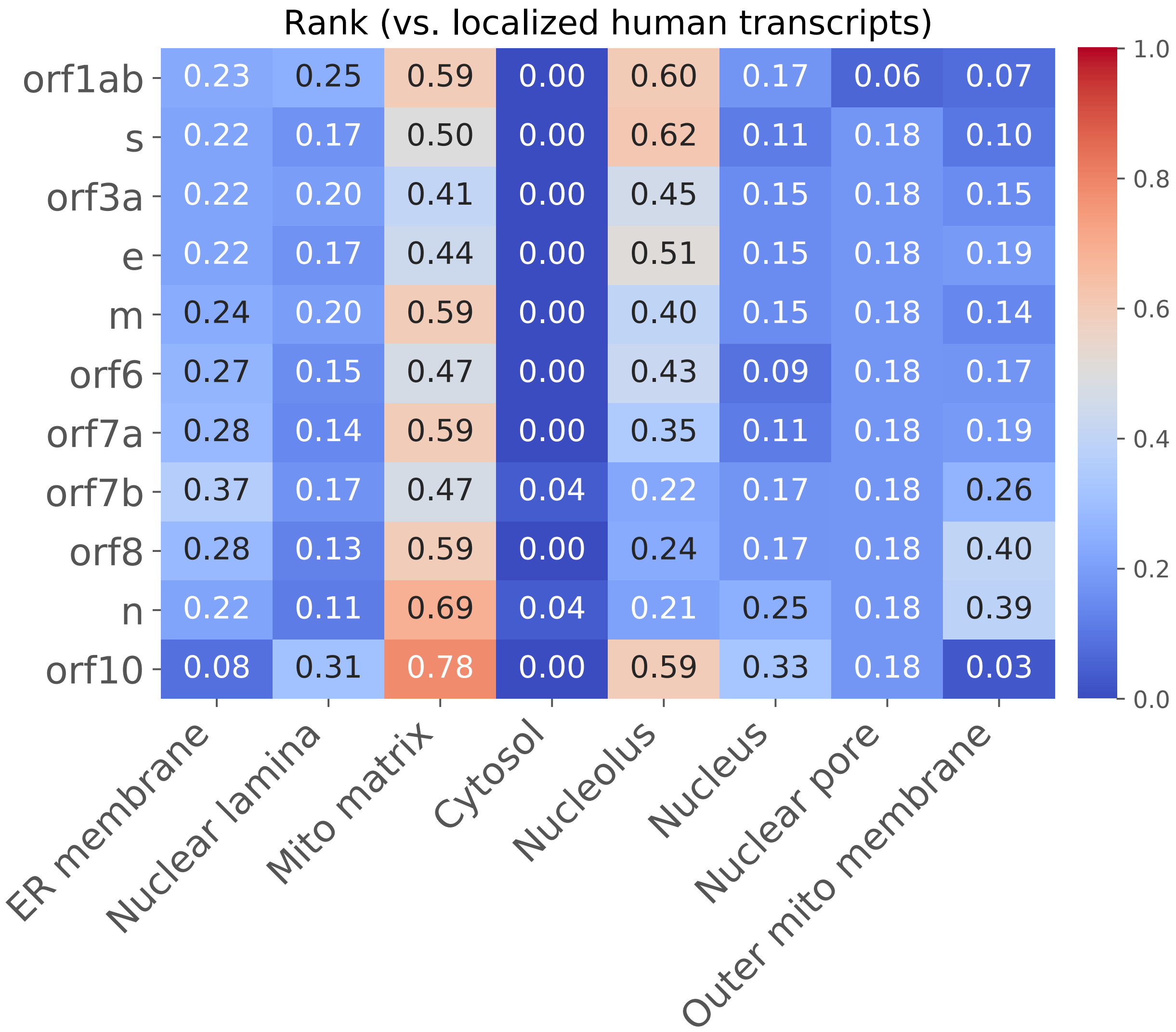


Figure S4 (related to Figure 1): Localization of negative strand sgRNA precursors. Figure 1C shows that the positive strand sgRNA transcripts tend to exhibit localization towards the mitochondrial matrix and nucleolus. Here, we look at the negative-strand precursors to those sgRNAs and observe that these transcripts share similar mitochondrial matrix and nucleolus localization patterns. This suggests another layer of conservation of this localization signal.