PLOS COMPLEX SYSTEMS

RESEARCH ARTICLE

System and transcript dynamics of cells infected with severe acute respiratory syndrome virus 2 (SARS-CoV-2)

João M. F. Silva[1]*, **Jose Á. Oteo**[1,2], **Carlos P. Garay**[1,3], **Santiago F. Elena**[1,4]*

1 Instituto de Biología Integrativa de Sistemas (CSIC-Universitat de València), Paterna, València, Spain, **2** Department de Física Teòrica, Universitat de València, Burjassot, València, Spain, **3** ICTS Laboratorio Subterráneo de Canfranc, Canfranc Estación, Huesca, Spain, **4** Santa Fe Institute, Santa Fe, New Mexico, United States of America

* jm.fagundes.silva@csic.es (JMFS); santiago.elena@csic.es (SFE)

OPEN ACCESS

Citation: Silva JMF, Oteo JÁ, Garay CP, Elena SF (2024) System and transcript dynamics of cells infected with severe acute respiratory syndrome virus 2 (SARS-CoV-2). PLOS Complex Syst 1(3): e0000016. https://doi.org/10.1371/journal.pcsy.0000016

Editor: Chris Gaiteri, SUNY Upstate Medical University Norton College of Medicine, UNITED STATES OF AMERICA

Received: March 25, 2024

Accepted: September 2, 2024

Published: November 5, 2024

Peer Review History: PLOS recognizes the benefits of transparency in the peer review process; therefore, we enable the publication of all of the content of peer review and author responses alongside final, published articles. The editorial history of this article is available here: https://doi.org/10.1371/journal.pcsy.0000016

Copyright: © 2024 Silva et al. This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data Availability Statement: All the R code used to generate these results are available at https://github.com/jmfagundes/sarscov2scrna.

Abstract

Statistical laws arise in many complex systems and can be explored to gain insights into their structure and behavior. Here, we investigate the dynamics of cells infected with severe acute respiratory syndrome virus 2 (SARS-CoV-2) at the system and individual gene levels; and demonstrate that the statistical frameworks used here are robust in spite of the technical noise associated with single-cell RNA sequencing (scRNA-seq) data. A biphasic fit to Taylor's power law was observed, and it is likely associated with the larger sampling noise inherent to the measure of less expressed genes. The type of the distribution of the system, as assessed by Taylor's parameters, varies along the course of infection in a cell type-dependent manner, but also sampling noise had a significant influence on Taylor's parameters. At the individual gene level, we found that genes that displayed signals of punctual rank stability and/or long-range dependence behavior, as measured by Hurst exponents, were associated with translation, cellular respiration, apoptosis, protein-folding, virus processes, and immune response. Those genes were analyzed in the context of a protein-protein interaction network to find possible therapeutic targets.

Author summary

Viruses replicate themselves inside susceptible cells by exploiting the cellular machinery. In turn, cells must counteract by mounting a defense against the virus and to alert other cells, in particular immune cells. This arms-race causes cells to significantly alter its gene expression pattern throughout the infection course. Here, we apply statistical laws from complex systems theory to gene expression data from single cells infected with severe acute respiratory syndrome virus 2 (SARS-CoV-2) to uncover how viral infection affects the cells at the system and individual gene levels. Our main findings are that viral infection influences the type of distribution of the gene transcripts along the infection in a cell type-dependent manner; and that, as infection progresses, several genes related to crucial cellular processes and immunity display signal of punctual instability and/or autocorrelation,

Funding: This work was supported by the CSIC Global Health Platform and European Union NextGenerationEU/PRTR (SGL2021-03-009 and SGL2021-03-052 to SFE), Ministerio de Ciencia, Innovación y Universidades-Agencia Estatal de Investigación (PID2022-136912NB-I00 to SFE; PID2019-109592GB-I00 to JAO) and Generalitat Valenciana (CIPROM/2022/59 to SFE; CIAICO/2021/180 to JAO). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

meaning that they are responding to viral infection at some point during the course of infection.

Introduction

Transcriptomics analyses commonly rely on linear models to test whether the mean expression of any set of genes is altered in response to a treatment or condition which are usually treated as factors in the model. Although changes in gene expression level are of the utmost importance in biology, aspects about the whole system behavior are not captured by these models. Another limitation of linear models is that by treating conditions as factors might lead to the loss of important information about the time course variation of transcripts. For instance, in single-cell RNA-sequencing (scRNA-seq) data, cells from the same cell type can be at distinct differentiation stages during sample preparation. Thus, the inference of a continuous pseudo-time trajectory of the transition from one cell type/stage to another, where each cell is assigned a value based on its relative position along it, can provide a continuous covariate for statistical models with higher sensitivity than factors to identify differentially expressed genes [1,2]. Accordingly, it has been recently shown that along the progression of cellular infection, the response to severe acute respiratory syndrome virus 2 (SARS-CoV-2) is triphasic, and that treating infected cells as one factor in an infected *vs*. uninfected linear model will lead to biases in identifying differentially expressed genes [3]. Nonetheless, even the addition of a pseudo-time measure as a covariate in linear models might fail to detect some complex behavior of gene expression.

Several statistical models and frameworks have been applied to transcriptomics data to model its structure and disentangle useful biological information from sampling noise and/or intrinsic stochastic biological variation. A recent study identified various emerging statistical laws from complex compartment systems on scRNA-seq data [4]. While the negative binomial distribution is often used to model both scRNA-seq and bulk RNA-seq count data, scRNA-seq data present some unique characteristics. The low capture rate of transcripts in scRNA-seq experiments makes that only about 10–20% of the transcripts from each cell are sequenced [5,6]. Due to this phenomenon, known as dropout, where a gene is expressed in a cell but its transcripts are not captured, gene count matrices from this type of experiments are sparse. Protocols for the preparation of scRNA-seq data also often rely on unique molecular identifier (UMI) tags that are added to the transcripts during RT-PCR to drastically reduce amplification bias [5,7–9].

The dynamics of various cellular processes can be explored with statistical models from complex systems. For instance, power law relationships arise naturally in many complex systems, including in scRNA-seq data [4]. In particular, an empirical law known as Taylor's law states that there is a power relationship between the mean of an element x and its standard deviation in the form of $\sigma = V\langle x\rangle^{\beta}$ [10]. If $\beta = 0.5$, then the system dynamics follows a Poisson distribution, and if $\beta = 1$, then the system fits to an exponential distribution, meaning that its elements are aggregated [10–12]. In the special case of time series data, Taylor's parameter V has been used as a proxy of system stability through time for data from the human microbiome [12]. In log-log scale, V is the intercept term. Whenever V is large, the standard deviation of each element in the system will also be large, a fact that is associated with system instability. In practice, the parameters and β are obtained as the intercept and the slope, respectively, of the linear fit to the pairs of data ($\langle x\rangle$,σ) represented in log-log scales.

The long-range dependence of a time series is a feature that has been thoroughly studied with the so-called Hurst's rescaled range analysis [13–15]. Records in time are associated to an index H, known as Hurst exponent, that runs between zero and one and, importantly, has an interesting interpretation. Values of $H > 0.5$ (irrespective, $H < 0.5$) convey that the sequence presents persistence (antipersistence). Persistence is a kind of bias which means that the future variations tend to be similar to the past ones in the sequence. Antipersistence is defined the other way around. Hurst found empirically that a large number of natural processes studied with the rescaled range yield H values close to 0.7, which is termed *Hurst phenomenon* in the literature. This analysis has been applied to temporal transcriptomic data from *Saccharomyces cerevisiae* and *Escherichia coli*, where it was shown that most genes exhibited $H > 0.5$ values, which are indicative of persistent long-range dependence [16]. Also, in a recent study, the rescaled range analysis shows the persistent character of the distribution of mutations along human chromosomes [17].

Here, we aim to gain insights into the structure and system behavior of cells infected with SARS-CoV-2 along the course of infection. We first hypothesize that the rank dynamics of transcripts and system behavior of scRNA-seq data from infected cells can be explored by fitting gene abundances to Taylor's law. We find that, in all cases, fluctuations grow with mean value on a biphasic Taylor's law, consisting in a Poisson and an exponential laws separated by a breaking point. Both, progression of infection and sampling noise have a significant impact on the estimation of Taylor's parameters.

The rank dynamics of a gene gives us information about its relative importance in the system. For each gene, we investigate about its stability along the course of infection and calculate their associated Hurst exponent. The robustness of these methods was further assessed by the use of simulated control datasets that mimics the technical noise associated with the accumulation of viral RNA inside the cells. These simulated datasets help us discerning whether the observed systemic and gene-level changes in stability are due to the infection process itself or simply due to sampling noise.

We will make use of one statistic introduced in [12], *RSI*, and new statistic derived from it, *PSI*. The former measures the gene rank stability and the latter tell us about the punctual stability of a gene (see Methods). Essentially, *RSI* computes the observed rank hops along pseudotime normalized to the number of genes and cells. *PSI* is the ratio of the gene *RSI* to the mean *RSI* of replicates. Thus, values $PSI > 1$ are interpreted as punctual stability. Overall, we found evidence of several genes exhibiting punctual rank stability and/or persistent behavior that are related to viral processes or immune responses that could serve as potential pharmaceutical targets for the treatment of COVID-19.

Finally, we framed our results in the context of a protein-protein interaction network, commonly known as an interactome, to further assess the relevance of these genes as therapeutic targets.

Results

Acquisition of scRNA-seq datasets of SARS-CoV-2-infected cells

Processed transformed counts matrices from human bronchial epithelial cells (hBECs) [18] and human intestinal epithelial cells (hIECs) [19] infected with SARS-CoV-2 were obtained from [3], which used a common pre-processing strategy across all samples. All data gathered here was generated using 10× Genomics Chromium scRNA-seq protocols that add UMIs to the RNA transcripts, which are used to count captured RNA molecules mitigating amplification bias. The hBECs dataset is composed by four samples: one mock-infected and three that were sequenced at 1-, 2- and 3-days post-infection (dpi), containing a total of 16,531, 9345, 12,472 and 17,820 cells, respectively. The hIECs dataset, on the other hand, is composed by six samples: three from colon organoids and three from ileum organoids, with one mock-infected sample and two samples sequenced at 12- and 24-hours post-infection (hpi) for each organoid. The total number of cells in the mock-infected, 12 hpi and 24 hpi colon cells samples is 4024, 3992 and 4407, respectively; whereas for the ileum cells samples it is, respectively, 3457, 4115 and 3629.

Detection of SARS-CoV-2 genome and identification of infected cells

Viral RNA was detected in all samples, although the detection of SARS-CoV-2 RNA in 28 mock-infected hBECs, 11 colon cells and 4 ileum cells is likely due to misalignments. To differentiate between infected cells supporting active viral activity from droplets that contained viral RNA from attached viral particles or ambient viral particles or RNA, we sought to estimate the mean SARS-CoV-2 UMI count from empty droplets to set a threshold for annotating infected cells. However, no viral RNA was detected in the empty droplets, and thus, a threshold of 10 viral UMIs was set. With this strategy, 1%, 8.5% and 11.5% of hBECs were infected at 1, 2 and 3 dpi, respectively; 11.5% and 96.3% colon cells were infected at 12 and 24 hpi, respectively; and 23.9% and 95% ileum cells were infected at 12 and 24 hpi, respectively. A high proportion of infected cells was observed, in particular for hIECs. This is in contrast with the number of infected cells in these datasets reported in [19], where the infection rate was estimated to be lower than 10%. Despite this, we decided to follow with our strategy due to the following two reasons. Firstly, in the original work with the hIECs datasets, the proportion of infected cells was estimated based on immunofluorescence staining of dsRNA and SARS-CoV-2 N protein [19]. It is likely that, at the beginning of infection, viral replication is low. Therefore, dsRNA might not be readily detected and N protein translation might also be low or even not yet synthesized. This means that many infected cells at the beginning of infection might be missed by immunofluorescence staining. Secondly, keeping uninfected cells in our analyses should not compromise our results. Instead, it would just make that the infection trajectory starts with uninfected cells, irrelevant for the application itself of statistical laws from complex systems to pseudotemporal scRNA-seq data.

Transcript abundances fit a biphasic Taylor's law

Gene abundances data of SARS-CoV-2-infected cells are represented in the Taylor plot of Fig 1A. The distribution of points suggests a biphasic, or segmented, linear regime. The outcomes of the linear regressions are shown in Table 1. According to the partial F-tests in Table 1, the data from all three cell types fit better to the biphasic model than to an unsegmented model with no breakpoint. (V_1, β_1) and (V_2, β_2) are the fit parameters to the left and right of the crossover, respectively.

In the biphasic model, two V and two β parameters are estimated, where V_1 and β_1 are the ones estimated for the data points before the breakpoint and V_2 and β_2 the ones estimated for the data points after the breakpoint. For large abundances, we find a slope $\beta \approx 1$, characteristic of the exponential distribution. For small abundances, we find a slope $\beta \approx 0.5$, characteristic of the Poisson distribution. As an example, we show the distribution of three genes in hBECs: one that follows a Poisson distribution (*ESD*), one that is close to the breakpoint (*HSP90AB1*) and one that follows an exponential distribution (*RPL13A*). Whereas *ESD* seems to fit to a Poisson distribution with a small λ parameter, *RPL13A* seems to better fit a negative binomial rather than an exponential distribution (Fig 1B). Indeed, the Poisson and negative binomial distribution are suitable to model counts from scRNA-seq data [20]. Note that given that we are fitting abundance data and not counts to Taylor's law, we should not observe true Poisson/exponential distribution, but we opted to keep the terms associated to the distributions in Taylor's law given their interpretation to system dynamics. A visual inspection of the distributions suggests that the transition from a Poisson to an exponential distribution is associated to the

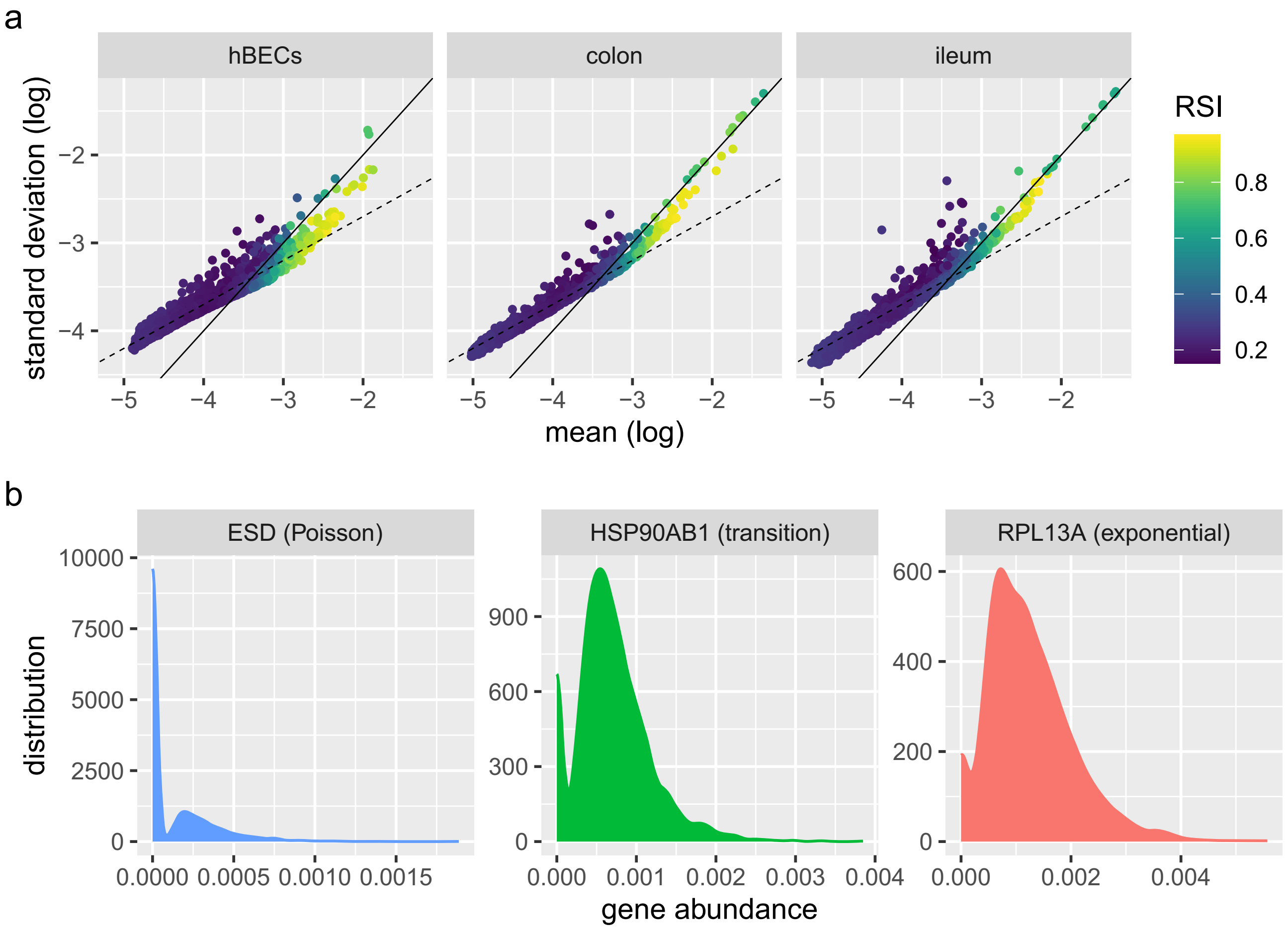

Fig 1. Taylor's law plots. (a) Mean and standard deviation, along pseudotime, are plotted in log-log scales for each gene. For illustrative purposes, solid lines correspond to the exponential distribution ($\beta = 1$ and $\log(V) = 0$), and dashed lines to the Poisson distribution ($\beta = 0.5$ and $\log(V) = -1.7$), where log (V) was chosen to be −1.7 for better visualization. The *RSI* of each gene is color coded. 3207, 4703 and 4433 cells; and 7978, 7567 and 8333 genes were present in each matrix for hBECs, colon and ileum cells, respectively. (b) Distributions of genes in hBECs, as a function of TPT counts, that are to the left of the breakpoint (Poisson, left panel), near the breakpoint (transition, central panel) and to the right of the breakpoint (exponential, right panel), intended to illustrate the transition Poisson to exponential distribution.

https://doi.org/10.1371/journal.pcsy.0000016.g001

peak at zero counts (Fig 1B). The observed biphasic behavior in the Taylor plot is most likely related to sampling noise for the low capture rate of transcripts in scRNA-seq protocols [20]. Additionally, the rank stability index (*RSI*; see Methods) was calculated for each gene. This index varies between zero and one and measures how much the rank of a gene is changing from one point to another along the pseudotime. An *RSI* value close to zero means that the gene is always alternating between last and first, whereas an *RSI* value close to unit means that the gene always maintains the same rank. Higher rank stability seems to be associated to highly expressed genes that follow an exponential distribution (Fig 1A).

The simulated control datasets mimic the increase in sampling noise seen in infected cells

In order to conduct a more thorough analysis of the progression of infection, infected cells were divided into bins with cells showing a similar viral load. We found that 30 bins

Table 1. Parameters of the biphasic fit to Taylor's law for infected cells. Matrix sparsity (proportion of zeros), number of genes and number of cells are also shown.

		hBECs	colon	ileum
Matrix properties	sparsity	0.77	0.76	0.72
	Number of genes	7978	7567	8333
	Number of cells	3207	4703	4433
Partial *F*-test[1]	*F*	0.82	0.53	0.76
	P	< 0.0001	< 0.0001	< 0.0001
Biphasic fit	V_1	0.02	0.03	0.04
	β_1	0.48	0.53	0.57
	V_2	0.25	1.47	1.28
	β_2	0.9	1.12	1.07
	breakpoint (log(mean))	-3.16	-2.99	-3.07

[1] One-tailed *F*-tests of the residuals of the biphasic *vs*. unsegmented fits.

https://doi.org/10.1371/journal.pcsy.0000016.t001

represented a good compromise between number of bins and number of cells in each bin. As infection progresses and cells accumulate more viral RNA, the sampling of cellular transcripts become compromised and a higher incidence of zeros in the matrix due to dropout is seen. Therefore, to better understand the effect of dropout in our analyses, we simulated the expected increase in sampling noise as infection progresses by down-sampling UMIs from uninfected cells to produce simulated datasets. Given that these simulated datasets are basically uninfected cells with an added Poisson process to capture transcripts, we expect that, as less transcripts are sampled, any observed change in the systemic or gene-level stability will be strictly associated with sampling noise. On the other hand, in infected cells, those changes might be associated with either sampling noise or response to the infection. Therefore, these datasets serve as controls for our analyses. The proportion of zeros per gene per bin of the simulated control datasets follows the same trend as true infected cells (S1A Fig). To investigate whether the simulated dataset retains the same transcriptional profile from the uninfected cells, we performed standard clustering analysis with cells from the simulated dataset and their "matching" uninfected cell. Cells from the simulated datasets clustered together with uninfected cells confirming that they still retain the same transcriptional profile (S1B Fig).

Analysis of infection progression reveals signals of varying system dynamics

To further investigate system stability throughout infection, Taylor's parameters were estimated for each bin (see section above) in the three cell types. An increase-decrease-increase pattern in both V and β was observed in hIECs (Fig 2A). In these cells, V and β increase with sampling noise for the simulated dataset, in contrast to the pattern seen in infected cells (Fig 2A). On the contrary, a similar pattern between the infected and simulated datasets was seen for hBECs, with an initial oscillation and a sharp increase of parameters V and β at the end of the infection. Mitochondrial expressed genes show an exponential distribution, even for lower abundances (Fig 2B; bins 27 and 28 for colon and ileum cells, respectively). This suggests that they are not responding to infection, but rather their rank is shifting due to differences in the expression of other genes. Detailed inspection of the plots shows that, for hBECs, some nuclear genes, most notably *LCN2*, *S100A2*, *S100A9*, *SCGB1A1*, *SCGB3A1*, *SERPINB3*, *SLPI*, and *WFDC2*, generally follow an exponential distribution regardless of their rank, suggesting that, like mitochondrial genes, they are aggregating in most bins and may not be responding to infection. We can observe how the thickness in the distribution of points in the Taylor's plot

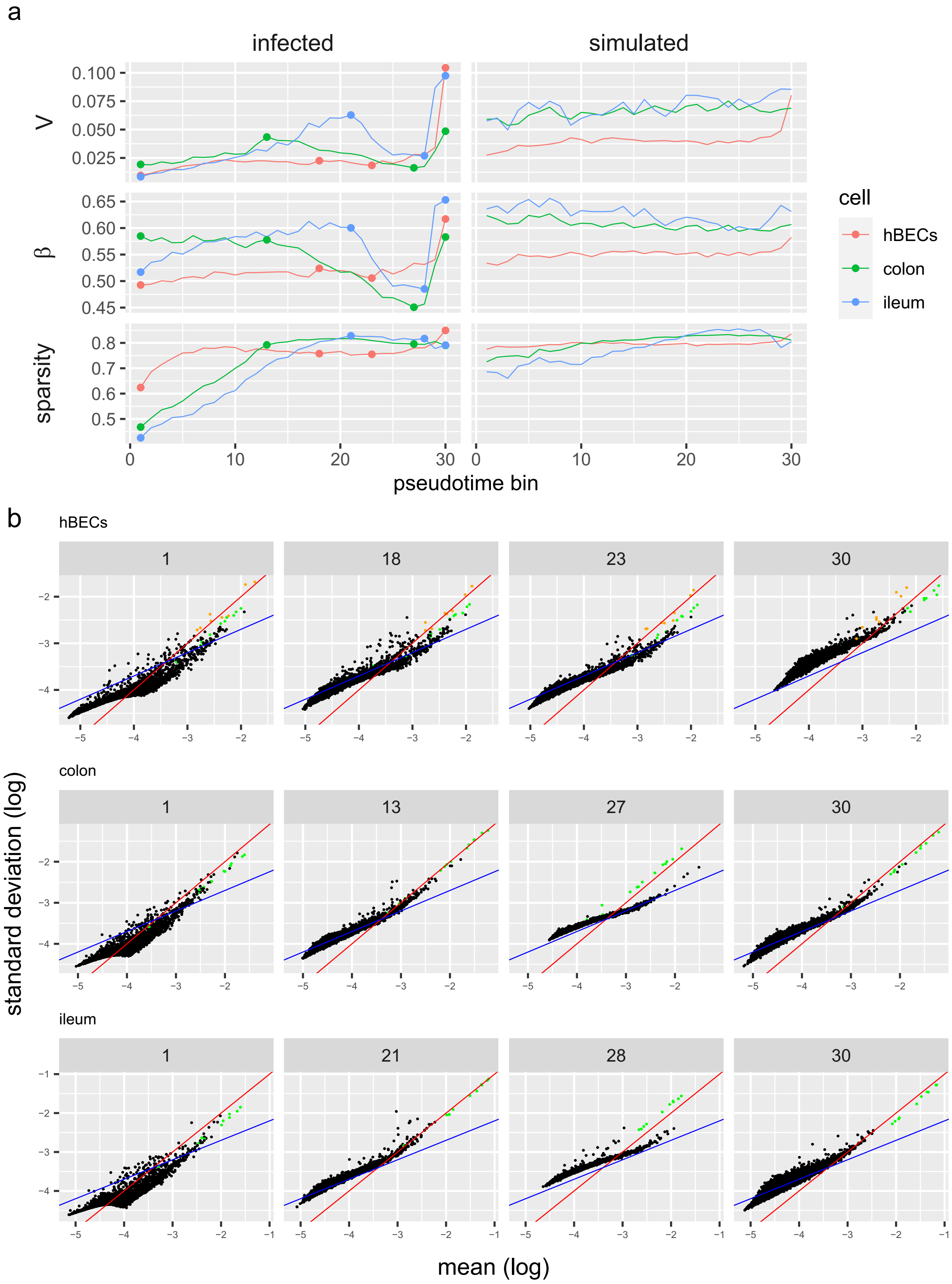
a
infected
simulated
V
0.100
0.075
0.050
0.025
β
0.65
0.60
0.55
0.50
0.45
sparsity
0.8
0.7
0.6
0.5
0
10
20
30
pseudotime bin
cell
hBECs
colon
ileum
b
hBECs
1
18
23
30
colon
1
13
27
30
ileum
1
21
28
30
−1
−2
−3
−4
−5
standard deviation (log)
mean (log)

Fig 2. Evolution of Taylor's law parameters along the course of infection. (a) Taylor's parameters, V and β, estimated in pseudotime intervals (bins) for infected cells and for the simulated datasets. (b) Taylor plots for the points in panel (a). Mitochondrial genes are shown in green. Selected nuclear genes in hBECs that generally followed an exponential distribution regardless of rank are shown in orange. Red lines correspond to an exponential distribution ($\beta = 1$ and $\log(V) = 0$), and blue lines to a Poisson distribution ($\beta = 0.5$ and $\log(V) = -1.7$), where $\log(V) = -1.7$ was chosen for better visualization.

https://doi.org/10.1371/journal.pcsy.0000016.g002

changes with the infection process (Fig 2B). This effect is due to the sparsity (*i.e.*, proportion of zeros) in the matrix of the gene counts, which grows with infection (S1A Fig). If the gene counts matrix contains an even number of zero and nonzero counts, a bell shape distribution of bins is observed (beginning of infection). Otherwise, if the gene counts matrix contains a dominant number of zero counts, the shape distribution of bins is much thinner.

Next, we performed a segmented fit to Taylor's law for each bin to estimate Taylor's parameters in the biphasic regime. Here, due to the noise associated with the smaller number of cells in each bin, we noticed that if genes with a high proportion of zeros are not removed, a triphasic fit to Taylor's law could appear in some bins, where genes that have a number of non-zero counts close to one appear to behave like an exponential distribution (S2A Fig). Hence, genes exhibiting more than 70% of zeros were removed from the binned data. Biphasic Taylor's parameters V_1 and β_1, that fit to gene abundances with a Poisson behavior, exhibited a similar pattern to the unsegmented fit parameters V and β (S2B Fig). We also noted that in this case sparsity also had a similar behavior, albeit it varied less (between ~35% and ~50%) than in the unsegmented fit. Notably, β_1 was lower than β at the beginning of infection, although V_1 and β_1 exhibited the same increase-decrease-increase behavior of V and β for hIECs. As infection progresses, the breakpoint increases for all three cell types (S2B Fig).

To further ascertain that the observed changes in Taylor's parameters are not due to technical noise, we performed ANCOVA tests for the effect of infection progression (herein pseudotime; where each bin corresponds to a different point throughout the infection), cell type and their interaction on each Taylor parameter, while also adding the number of genes and matrix sparsity as covariates to control for increasing sampling noise in the system. The rationale of adding these covariates is that as less cellular transcripts are captured due to increasing viral RNA accumulation a higher proportion of zeros will be observed and less genes will have their transcripts captured. All explanatory variables, with the exception of sparsity for β, had a significant effect on parameters V and β. The largest effect sizes (partial η^2) were estimated for cell type and the interaction between pseudotime and cell type for parameter V and cell type, number of genes and the interaction between pseudotime and cell type for β (Table 2). Additionally, we performed these analyses on the simulated control datasets. Whereas sampling noise had a significant effect on Taylor's parameters for the simulated datasets, its interaction with cell type was not significant for V and, although significant for parameter β, its effect size was lower than that of the infected dataset (Table 2). From these analyses we conclude that

Table 2. *P*-values from ANCOVA analysis of Taylor's parameters for cells infected with SARS-CoV-2 and the simulated dataset. Partial η^2 values are shown in parenthesis (partial $\eta^2 \geq 0.15$ are conventionally taken as large effects).

	V		β	
Effect	infected	simulated	infected	simulated
Pseudotime	< 0.0001 (0.33)	< 0.0001 (0.44)	0.0096 (0.08)	< 0.0001 (0.19)
Cell	0.0003 (0.18)	< 0.0001 (0.85)	< 0.0001 (0.46)	< 0.0001 (0.94)
Sparsity	0.0006 (0.13)	0.0187 (0.07)	0.87962 (2.81×10^{-4})	< 0.0001 (0.46)
Number of genes	0.0016 (0.11)	0.8539 (4.16×10^{-4})	< 0.0001 (0.52)	0.0319 (0.05)
Pseudotime-by-cell	0.0002 (0.19)	0.1033 (0.05)	< 0.0001 (0.45)	0.0014 (0.15)

https://doi.org/10.1371/journal.pcsy.0000016.t002

noise induced by dropout has a uniform effect on Taylor's parameter *V* and a cell type-dependent effect on parameter *β*, and that infection with SARS-CoV-2 induces changes in the distribution properties of the system that is also dependent on cell type.

Genes that display punctual rank stability are, most notably, related to translation, protein folding, and apoptosis

Genes exhibiting punctual rank stability were found by evaluating whether its *RSI* value (calculated from its mean expression at each bin; see Methods) is higher than expected by chance, irrespective of whether its *RSI* was high or low. In short, genes that display signal of punctual rank stability are those which have undergone a significant change in their stability throughout the course of infection, meaning that they are either consistently gaining or losing ranks, or have gone through an abrupt change in their stability. In contrast, genes that do not display punctual rank stability will always present the same stability throughout the infection. The *RSI* and punctual stability index (*PSI*; see Methods) of each gene is shown in Fig 3A. In total, 380, 2840 and 4230 genes were found to exhibit punctual rank stability in hBECs, colon and ileum cells, respectively (S1 File). To ascertain that these results are robust, we also applied this approach to the simulated control datasets and uninfected cells, where four, eight and six false-positives were detected in hBECs, colon and ileum cells simulated datasets, respectively; and 15, 10 and 11 false-positives were detected in uninfected hBECs, colon and ileum cells, respectively (S1 File). The low false-positive rate of this analysis indicates our results are robust, and that the higher number of genes displaying punctual stability in hIECs might be related to intrinsic differences in the response to viral infection between hBECs and hIECs.

Punctual rank stability in all three cell types was found for 172 genes. GO terms enrichment analysis of these genes revealed an enrichment in those related to cytoplasmic translation (GO:0002181), regulation of apoptotic signaling pathway (GO:2001233), regulation of endoplasmic reticulum unfolded protein response (GO:1900101), and a few terms related to innate immunity such as response to lipopolysaccharide (GO:0032496), among others (Fig 3B; S1 File). Those genes that showed signal of punctual rank stability in all three cell types were generally stable, as shown by their median *RSI* (Fig 3B; S1 File). Among these, genes associated with translational processes and gene expression, such as translational elongation (GO:0006414), ribosome assembly (GO:0042255), maturation of SSU-rRNA (GO:0030490) and ncRNA processing (GO:0034470) were the most stable; and those associated with protein-DNA complex subunit organization (GO:0071824), detoxification (GO:0098754), epithelial cell apoptotic process (GO:1904019) and processes associated with immune response, such as response to lipopolysaccharide (GO:0032496), response to molecule of bacterial origin (GO:0002237) and myeloid leukocyte migration (GO:0097529) were the least stable (Fig 3B; S1 File).

Several genes related to translation, cellular respiration, and viral processes, show evidence of persistent rank behavior

Next, we examined the presence of persistent behavior of gene rank along the course of infection, which indicates whether a gene has a tendency to maintain its rank once it changes. The robustness of the estimation of *H* from rank data was assessed by performing the analyses on a set of control datasets that included the simulated data, random matrices, uninfected cells, and shuffled infected cells that are not ordered according to viral RNA accumulation. The Hurst exponents calculated from these control datasets seemed to follow a normal distribution with a mean close to ~0.5 which is expected for data with no temporal correlation, with the exception of the simulated ileum dataset that showed a small deviation towards higher *H* values (Fig 4A).

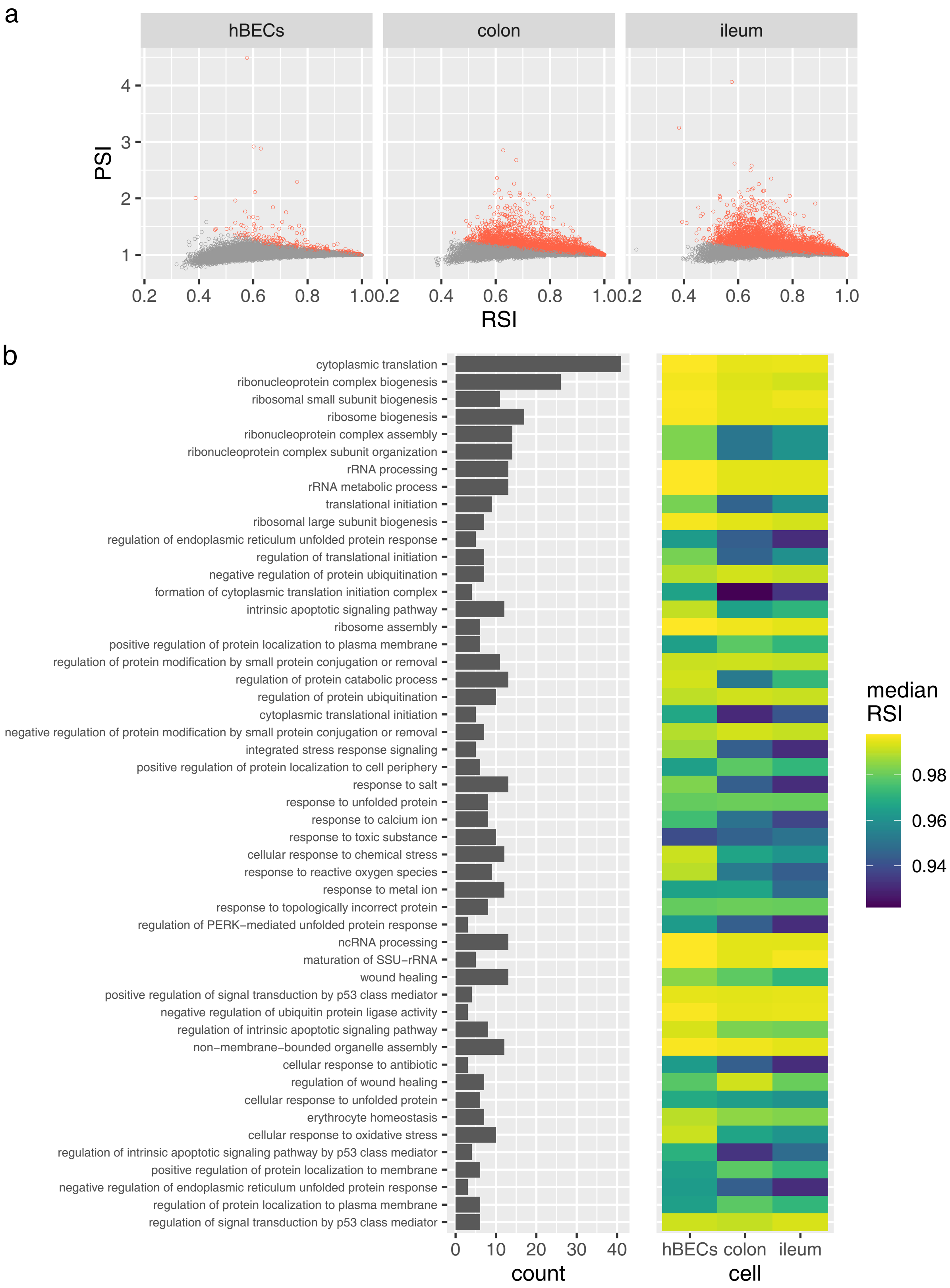
a
hBECs
colon
ileum
PSI
4
3
2
1
0.2
0.4
0.6
0.8
1.0
RSI
b
cytoplasmic translation
ribonucleoprotein complex biogenesis
ribosomal small subunit biogenesis
ribosome biogenesis
ribonucleoprotein complex assembly
ribonucleoprotein complex subunit organization
rRNA processing
rRNA metabolic process
translational initiation
ribosomal large subunit biogenesis
regulation of endoplasmic reticulum unfolded protein response
regulation of translational initiation
negative regulation of protein ubiquitination
formation of cytoplasmic translation initiation complex
intrinsic apoptotic signaling pathway
ribosome assembly
positive regulation of protein localization to plasma membrane
regulation of protein modification by small protein conjugation or removal
regulation of protein catabolic process
regulation of protein ubiquitination
cytoplasmic translational initiation
negative regulation of protein modification by small protein conjugation or removal
integrated stress response signaling
positive regulation of protein localization to cell periphery
response to salt
response to unfolded protein
response to calcium ion
response to toxic substance
cellular response to chemical stress
response to reactive oxygen species
response to metal ion
response to topologically incorrect protein
regulation of PERK−mediated unfolded protein response
ncRNA processing
maturation of SSU−rRNA
wound healing
positive regulation of signal transduction by p53 class mediator
negative regulation of ubiquitin protein ligase activity
regulation of intrinsic apoptotic signaling pathway
non−membrane−bounded organelle assembly
cellular response to antibiotic
regulation of wound healing
cellular response to unfolded protein
erythrocyte homeostasis
cellular response to oxidative stress
regulation of intrinsic apoptotic signaling pathway by p53 class mediator
positive regulation of protein localization to membrane
negative regulation of endoplasmic reticulum unfolded protein response
regulation of protein localization to plasma membrane
regulation of signal transduction by p53 class mediator
0
10
20
30
40
count
hBECs
colon
ileum
cell
median RSI
0.98
0.96
0.94

Fig 3. Gene rank stability dynamics. (a) Comparison of *RSI* with *PSI* of all genes for hBECs, colon and ileum cells. Each point represents one gene. Genes with significant punctual stability (false discovery rate $FDR < 0.05$) are shown in red. (b) Top 50 enriched GO terms, ranked by *P*-value, for genes that displayed signal of punctual rank stability in all three cell types. The median *RSI* of the genes in each GO term is shown for each cell type.

https://doi.org/10.1371/journal.pcsy.0000016.g003

Infected cells ordered according to viral RNA accumulation displayed a broad distribution of *H* values whose mean were nonetheless visibly higher than those from the control datasets (Fig 4A). By analyzing *H* values inside the Taylor's law plot for each dataset of infected cells, we found that low expressed genes tended to exhibit slightly higher *H* exponents, most noticeably for ileum cells (Fig 4C). Given that gene rank was randomized in the case of ties, and that low expressed genes will be tied when their expression is zero, the higher *H* values of these low expressed genes is most likely artificially inflated. However, it was also noted the most highly expressed genes also show a general decrease in their *H* exponents (Fig 4C). In this case, it is likely that the most expressed genes are constantly expressed throughout the infection course, and thus, will tend to exhibit lower *H* values, sometimes near 0.5.

Based on the observations above, genes exhibiting strong persistent rank behavior were determined by comparing their *H* exponents from infected cells *vs*. the ones calculated from the simulated dataset. Overall, the *H* exponents from infected cells tended to present higher values than their corresponding exponents from the simulated dataset (Fig 5A). We observed that 610, 657 and 1569 genes, out of 9910, 7582 and 8333, showed evidence of strong persistent rank behavior ($H \geq 0.7$ in infected cells and, concomitantly, $H < 0.7$ in simulated dataset) in hBECs, colon and ileum cells, respectively. As discussed above, a higher false-positive rate in ileum cells is expected given that low expressed genes showed slightly higher *H* values. In total, 297 genes displayed evidence of strong persistent rank behavior concomitantly in all three cell types. Amongst those, we found an enrichment of genes related to cytoplasmic translation (GO:0002181), cellular respiration (GO:0045333) and processes related to viral infection, such as internal ribosome entry site (IRES)-dependent viral translational initiation (GO:0075522), viral process (GO:0016032), viral translation (GO:0019081), and viral life cycle (GO:0019058), among others (Fig 5B; S2 File). An enrichment of genes related to IRES-dependent viral translation is unexpected since SARS-CoV-2 is not known to contain an IRES [21]. The median *H* of the genes in these GO categories varied little between 0.7 and ~0.8, and were overall higher in ileum cells and lower in colon cells (Fig 5B; S2 File). Accordingly, *H* values were significantly different across cell types (pairwise Wilcoxon rank sum tests; $FDR < 0.0001$ in all cases). In general, processes related to cellular respiration, such as mitochondrial electron transport, cytochrome c to oxygen (GO:0006123), mitochondrial electron transport, ubiquinol to cytochrome c (GO:0006122) and aerobic electron transport chain (GO:0019646) exhibited higher median *H* values, together with other processes such as mRNA stabilization (GO:0048255), regulation of substrate adhesion-dependent cell spreading (GO:1900024) and negative regulation of oxidative stress-induced intrinsic apoptotic signaling pathway (GO:1902176).

Mapping genes with statistical relevance to the human interactome reveals possible targets for the treatment of COVID-19

We sought to investigate whether genes that display punctual rank stability or persistent behavior are of any particular importance in the human interactome. For this purpose, we have obtained a high quality human protein interactome built from experimental protein-protein interaction screenings [22] to unravel which ones are hubs and/or bottlenecks. Whereas genes that constitute hubs have a high number of interactions, those that constitute bottlenecks

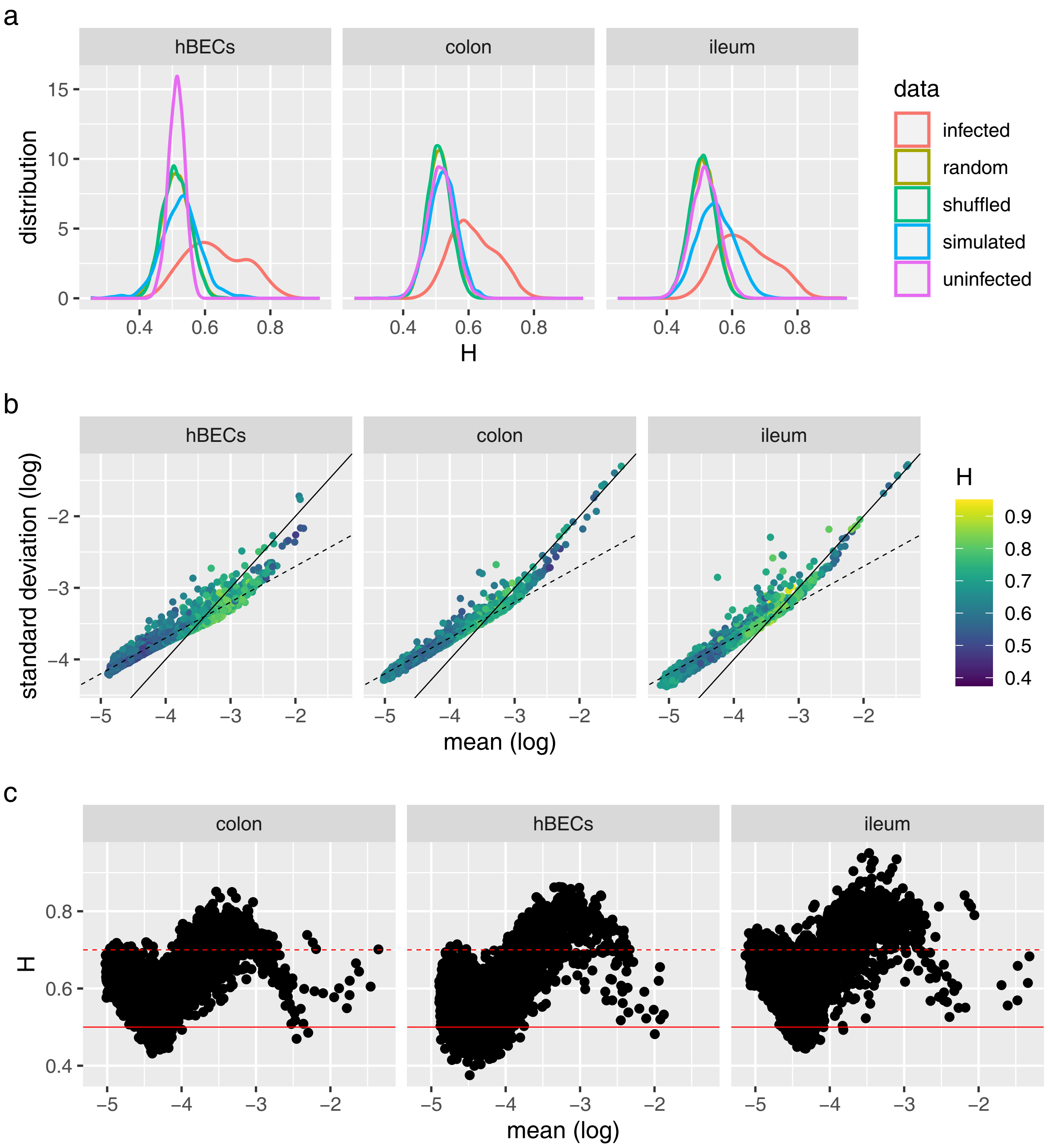

Fig 4. Estimation of *H* exponents from rank data. (a) Distribution of *H* exponents estimated from gene rank data from infected cells and control datasets for hBECs, colon and ileum cells. (b) Taylor's law plots showing the *H* value (color coded) of each gene for each cell type. Solid lines correspond to the exponential distribution ($\beta = 1$ and $\log(V) = 0$), and dashed lines to the Poisson distribution ($\beta = 0.5$ and $\log(V) = -1.7$, chosen for better visualization). (c) Plots of the *H* exponents calculated from infected cells *vs* the mean expression of genes present in the Taylor's plots in panel (b). A dashed red line shows indicates the value $H = 0.7$ (Hurst phenomenon). $H > 0.7$ corresponds to strong persistent behavior. A solid red line indicates $H = 0.5$. Remind that $H > 0.5$ corresponds to persistent behavior.

https://doi.org/10.1371/journal.pcsy.0000016.g004

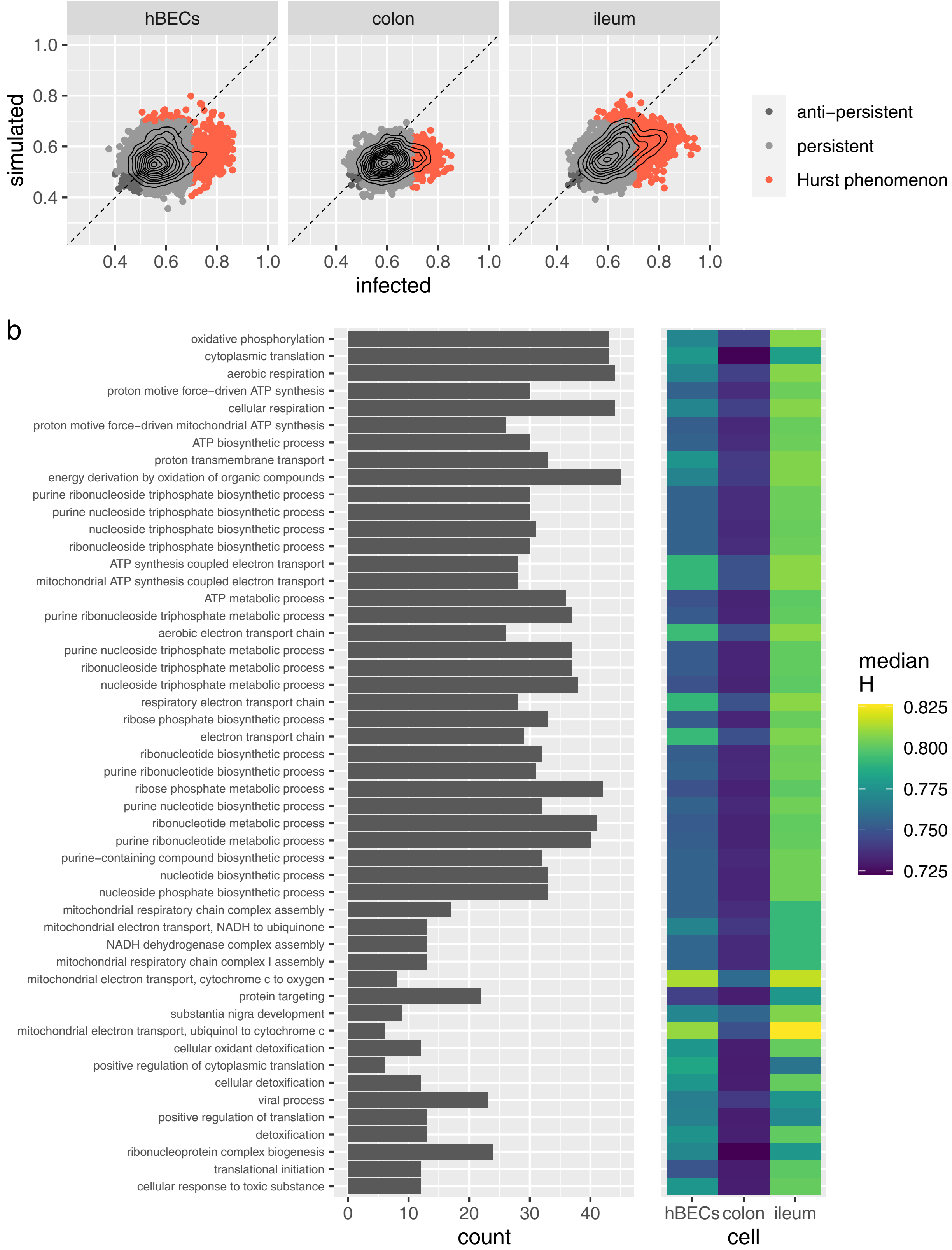
hBECs
colon
ileum
simulated
1.0
0.8
0.6
0.4
0.4
0.6
0.8
1.0
infected
● anti−persistent
● persistent
● Hurst phenomenon
b
oxidative phosphorylation
cytoplasmic translation
aerobic respiration
proton motive force−driven ATP synthesis
cellular respiration
proton motive force−driven mitochondrial ATP synthesis
ATP biosynthetic process
proton transmembrane transport
energy derivation by oxidation of organic compounds
purine ribonucleoside triphosphate biosynthetic process
purine nucleoside triphosphate biosynthetic process
nucleoside triphosphate biosynthetic process
ribonucleoside triphosphate biosynthetic process
ATP synthesis coupled electron transport
mitochondrial ATP synthesis coupled electron transport
ATP metabolic process
purine ribonucleoside triphosphate metabolic process
aerobic electron transport chain
purine nucleoside triphosphate metabolic process
ribonucleoside triphosphate metabolic process
nucleoside triphosphate metabolic process
respiratory electron transport chain
ribose phosphate biosynthetic process
electron transport chain
ribonucleotide biosynthetic process
purine ribonucleotide biosynthetic process
ribose phosphate metabolic process
purine nucleotide biosynthetic process
ribonucleotide metabolic process
purine ribonucleotide metabolic process
purine−containing compound biosynthetic process
nucleotide biosynthetic process
nucleoside phosphate biosynthetic process
mitochondrial respiratory chain complex assembly
mitochondrial electron transport, NADH to ubiquinone
NADH dehydrogenase complex assembly
mitochondrial respiratory chain complex I assembly
mitochondrial electron transport, cytochrome c to oxygen
protein targeting
substantia nigra development
mitochondrial electron transport, ubiquinol to cytochrome c
cellular oxidant detoxification
positive regulation of cytoplasmic translation
cellular detoxification
viral process
positive regulation of translation
detoxification
ribonucleoprotein complex biogenesis
translational initiation
cellular response to toxic substance
0
10
20
30
40
count
hBECs
colon
ileum
cell
median
H
0.825
0.800
0.775
0.750
0.725

Fig 5. Functional analysis of genes exhibiting strong persistent behavior. (a) Comparison of the H exponents from infected and simulated data for hBECs, colon and ileum cells. Each point represents one gene, which are colored depending on whether they show anti-persistent ($H < 0.5$, black) or persistent ($0.7 > H > 0.5$, gray) behavior or if Hurst phenomenon ($H > 0.7$, red) is seen. The dashed lines are bisecting lines. Black lines represent kernel density of the data. (b) Top 100 enriched GO terms, ranked by P-value, for genes exhibiting evidence of persistent rank behavior in all three cell types. The median H of the genes in each GO term is shown for each cell type.

https://doi.org/10.1371/journal.pcsy.0000016.g005

(articulation points) connect different modules of the interactome. Given that disruption of hubs and bottlenecks have a great impact on the connectivity and flow of information in the network, respectively, these genes are suitable candidates for drug targets.

Of the genes that simultaneously showed signal of punctual rank stability in all cell types, *KRT19*, *SNRPB* and *TMEM14C* represent hubs in the human interactome, whereas *ARPC3*, *BTF3*, *IFIT2*, *S100P*, *SET*, and *UBC* represent bottlenecks. Four genes, namely *AGR2*, *EIF3F*, *GLRX3*, and *REL*, are both hubs and bottlenecks (S3 File). For genes that showed strong persistence behavior in all cell types, 14 are hubs in the human interactome, 24 are bottlenecks and 12 are both, including *EIF3F* (S3 File).

Novel therapeutic targets for the treatment of COVID-19

Lastly, we analyze 29 genes with punctual stability and/or persistent behavior in all cell types but that were not previously detected as being differentially expressed in the datasets that we analyzed in their original studies [18,19] (S4 File). Of those genes, we selected the ones that represents hubs and/or bottlenecks in the human interactome as possible novel therapeutic targets, where *HNRNPF* and *RNASEK* are hubs, *PCBP1* is a bottleneck, and *EIF3F* and *GLRX3* are both hubs and bottlenecks (S4 File).

Discussion

The underlying characteristics and dynamics of complex systems can be captured by several simple statistical laws. Here, we focus on a dynamical complex system of cells infected with SARS-CoV-2 to uncover how the system behaves as a function of infection progression. First, we fitted transcript abundance data to Taylor's law to study system level dynamics as previously done with data from the human gut microbiome [12]. A biphasic fit to Taylor's law was observed, where the most expressed genes followed an exponential distribution, and the remaining genes followed a Poisson distribution. A biphasic behavior in scRNA-seq has been previously identified and was mainly attributed to the sampling process. For instance, shallow sequencing can mask the evidence of overdispersion which results in low expressing genes fitting to a Poisson distribution [20]. Interestingly, Lazzardi *et al*. found a triphasic Zipf's law behavior in scRNA-seq data [4]. Here, a triphasic fit to Taylor's law was only observed in some bins when no filtering of genes with a high proportion of zeros was performed (S2A Fig).

Overall, the infection course evolution of Taylor's parameters between infected cells and the control simulated dataset was similar for hBECs, although infected hIECs present a distinct increase-decrease-increase behavior in comparison to the simulated datasets. This suggests that the progression of infection had a significant impact on the system dynamics of hIECs cells, whereas for hBECs, Taylor's parameters were mostly influenced by sampling noise. Taylor's parameter V has been used as a proxy to system stability in data from the human gut microbiota [12]. If β is constant across different samples, then changes in V correspond to variations to the standard deviation of all elements of the systems equally. If all elements display large standard deviation, we can assume that their rank is unstable. Here, however, both parameters V and β varied simultaneously along the course of infection, which might compromise the relationship between V and system stability. Nevertheless, our results show that infection with SARS-CoV-2 has a systemic effect on the properties of the distribution of transcripts at the cell level.

Whether there is a direct or indirect relationship between infection progression and Taylor's parameters is inconclusive. One possibility is that in the absence of dropout (*i.e.*, all transcripts in a cell are sequenced), the whole system will better fit to an exponential distribution. In this case, it is likely that the observed change in the breakpoint as infection progresses is due to increase in noise in the system (here, not technical/sampling noise), meaning that the relationship is indirect. Another possibility is that the Poisson to exponential transition dynamics might arise from the interplay between RNA transcription bursts and RNA degradation, or as previously suggested, a suppression in the export of newly transcribed RNA out of the nucleus that will be latter degraded [3], which is affected by viral infection. The *RSI* of most genes was low, which is consistent with constant rank hopping along the course of infection due to transcriptional bursts and high rates of RNA degradation. The most expressed genes, however, displayed high *RSI* values, suggesting higher stability and lower RNA degradation rates. These genes followed the exponential distribution, which can be interpreted as aggregation behavior (Fig 1A). Nevertheless, our results suggest that, at least for hIECs cells, the switch from a Poisson to exponential distribution and Taylor's parameters are not only influenced by sampling noise but also by the progression of the disease, revealing that the whole system dynamics of transcripts at the cellular level is affected, directly or indirectly, by viral infection.

While it is likely that noise influences the detection of genes showing relevant statistical properties, we focused our analysis on the subset of genes that showed significant statistical relevance in all three cell types to find the ones most relevant to the infection process. Several ribosomal proteins and some genes related to cellular respiration, protein folding, apoptosis and immune response showed signatures of punctual rank stability and/or persistent behavior in all cell types. Mitochondria-encoded genes are likely not responding to infection given that, along the progression of infection, they always followed an exponential distribution even when their expression decreased (Fig 2B). However, some nuclear genes related to cellular respiration indeed showed signals of punctual stability and/or persistent behavior in all cell types. The ORF9b of both SARS-CoV-1 and -2 localizes to the mitochondria and interacts with the translocase of outer membrane (TOM) protein 70 (TOM70), a receptor involved in mitochondrial antiviral signaling and apoptosis [23], to suppress the cellular immune defense [24]. In line with this, the chaperone HSP90AA1, that interacts with TOM70 to induce apoptosis [24], showed signature of punctual stability in all cell types; while its paralog, HSP90AB1, showed signature of persistent rank behavior. Additionally, other proteins that are part of the TOM complex in the mitochondria, such as TOM5, TOM6, TOM7, and TOM20 displayed evidence of persistent rank behavior in all cell types, and TOM6 was also found to be a bottleneck in the human interactome. Comparing our results with those from the original studies that used more conventional approaches, TOM5, TOM7 and TOM20 were not found to be differentially expressed on hBECs, whereas TOM5 and TOM6 were not found to be differentially expressed on hIECs.

Next, we further examined genes related to apoptosis that exhibited signal of punctual rank stability and/or persistent behavior in all cell types. The expression of 14 genes related to immunogenic cell death (ICD), which is characterized by cellular death due to stress that promotes an immune response against cell death antigens, have been previously shown to be altered in COVID-19 patients [25], including *HSP90AA1* that was discussed above. We found that *HMGB1*, a gene related to apoptosis and ICD [26] showed signature of persistent rank behavior in all cell types. Six genes related to apoptosis that showed signals of persistent rank behavior were shown to be hubs and/or bottlenecks in the human interactome: *ARL6IP1* and *FIS1* are hubs; *LAMTOR5*, *RHOA* and *VPS35* are bottlenecks and *TMBIM6* is both a hub and a bottleneck. *FIS1* was shown to be down-regulated in young COVID-19 patients with idiopathic pulmonary fibrosis compared to young COVID-19 patients with chronic obstructive pulmonary diseases [27]. The SARS-CoV-2 spike protein was shown to activate RhoA, a protein that regulates endothelial cytoskeleton, causing a disruption of the blood-brain barrier function [28]. *TMBIM6* was previously identified as a potential target for the treatment of COVID-19 [29].

Focusing on some genes that are known to be associated with COVID-19, we found that the C-X-C motif ligand chemokine genes *CXCL1* and *CXCL3*; the interferon stimulated gene *IFIT2*; the transcription factor *IRF1*; the AP-1 transcription factor proteins JUN and JUND; and the NF-κB inhibitor genes *NFKBIA*, *NFKBIZ* and *TNFAIP3*, showed evidence of punctual rank stability. Both *CXCL1* and *CXCL3* were found to be upregulated in response to SARS--CoV-2 infection [30]. *IFIT2* showed a bimodal expression pattern in immune cell types from patients with severe COVID-19 [31], and was found to be a bottleneck in the human interactome in our analysis. Interestingly, the bimodal expression of *IFIT2* should resemble aggregation behavior in these datasets. IRF1 regulates the expression of MHC class I, and was shown to be inhibited by SARS-CoV-2 ORF6 [32]. JUN was found to be a hub in the SARS-CoV-2-host interactome [33], and both *JUN* and *JUND* showed signal of abnormal behavior in the same datasets used in the present study [3]. Lastly, the NF-κB signaling pathway is activated upon infection with SARS-CoV-2 and triggers inflammation and the production of cytokines [34,35]. Higher expression levels of *NFKBIA* and *TNFAIP3* in basal, ciliated and T cells were associated with the severity of COVID-19 [31]; and an insertion homozygosis of the *NFKBIZ* gene is associated with higher mortality by COVID-19 [36]. It is important to note, however, that in the datasets used in our study, it is likely that some infected cells were already responding to interferon and other immune signaling proteins from other previously infected cells. Thus, the punctual stability of some genes related to immune response may not be due to the cellular infection itself, but rather due to response to other infected cells. Additionally, given the higher-than-expected number of infected cells detected here, it is likely that some uninfected bystander cells are present at the beginning of the infection, meaning that the infection progression analyzed here starts at a point prior to infection.

Abnormal dynamics of ribosomal proteins and a few genes related to immune response in SARS-CoV-2-infected cells was previously detected in the same datasets used in this study [3]. Recently, an inverse relationship between inflammation and ribosome level was found, and furthermore, an increase in inflammation and decrease in ribosome level was associated with the severity of COVID-19 symptoms [37]. However, it remains unclear whether the ribosome content-inflammation interplay along the course of cellular infection bears any relevance to the dysregulated immune response associated with COVID-19 severity. Nevertheless, ribosomal proteins are tantalizing therapeutic targets due to their importance to viral translation as it has been recently shown that two ribosome inactivating proteins can inhibit SARS-CoV-2 replication in human lung epithelial cells (A549) [38]. In addition to ribosomal proteins, some translation initiation factors also showed evidence of punctual stability and/or persistent behavior. *EIF3A* and *EIF3F*, which are involved in the IRES-dependent translation of hepatitis C virus [39,40], showed, respectively, signature of persistent rank behavior and evidence of punctual stability and strong persistent rank behavior; *EIF3E* showed evidence of punctual stability and persistent rank behavior; and several other translation initiation factors, namely *EIF1*, *EIF2AK2*, *EIF3K*, *EIF4G2*, and *EIF5*, showed evidence of persistent rank behavior. The RNA-binding activity of several components of the eIF3 complex is inhibited by SARS-CoV-2, which is in agreement with the role of SARS-CoV-2 NSP1 in inhibiting the recruitment of 40S to cellular mRNAs [41].

Lastly, we focus on those genes that exhibited punctual rank stability and/or persistent behavior in all three cell types that are either hubs and/or bottlenecks and were not previously found to be differentially expressed in the same datasets analyzed here. We found that, of those genes, *HNRNPF* and *RNASEK* are hubs, *PCBP1* is a bottleneck, and *EIF3F* and *GLRX3* are both, and thus, should represent the best candidates for pharmaceutical targets. As discussed above, *EIF3F* is involved in viral IRES-dependent translation [39,40]. However, given that SARS-CoV-2 is not known to have an IRES, and that eIF3F was not shown to be enriched in viral ribonucleoproteins [41], it is more likely that its observed rank dynamics during SARS-CoV-2 infection is related to the inhibition of recruitment of 40S to cellular mRNAs. If so, targeting *EIF3F* might be an efficient strategy to counteract the inhibition of translation of cellular transcripts. *GLRX3* encodes for a protein known as PICOT that inhibits the transcription factors AP-1 and NF-κB [42], and could be targeted to regulate the exacerbated inflammatory response to SARS-CoV-2.

Conclusion

Here, we successfully applied statistical frameworks from complex systems to scRNA-seq data to investigate the dynamics of cells infected with SARS-CoV-2 at the system and individual gene levels. Our results suggest a cell type-dependent systemic instability in response to SARS-CoV-2 infection. In hIECs, SARS-CoV-2 infection led to an increase, decrease and final increase in system stability (Fig 2A). In contrast, for hBECs, infection and sampling noise seemingly had the same effect on systemic instability (Fig 2A). Despite this systemic cell type-dependent response, several genes involved in translation, cellular respiration, apoptosis, protein-folding, and immune response showed evidence of deterministic behavior in all three cell types along the course of infection in the form of punctual rank stability or persistent rank behavior.

Methods

Data collection

Processed scRNA-seq data of human bronchial epithelial cells (hBECs) [18] and human intestinal epithelial cells (hIECs) from colon and ileum intestinal organoids [19] were obtained from [3]. The obtained processed gene frequencies matrices were previously generated by transforming UMI counts to transcript abundances. Briefly, UMI counts were modeled under a Poisson distribution, where transcript abundances were represented as the weighted average of transcript frequencies based on a normalized likelihood function [3]. Cells with at least 10 uncorrected viral UMIs were considered to be infected, and cells from mock data were considered to be uninfected. Infected cells were ordered based on their percentage of viral RNA, which is used here as a proxy of infection time and thus provide a measure of pseudotime of infection progression. Viral RNA counts were removed from the count matrices before downstream analyses, meaning that gene abundances were calculated using only cellular transcripts.

Fit to Taylor's law

To analyze the progression of Taylor's parameters through infection, cells were first ordered based on the accumulation of viral RNA then separated in 30 bins containing a similar number of cells (~105, ~147 and ~124 cells for hBECs, colon and ileum cells, respectively) with a similar viral load. Genes exhibiting more than 95% of zeros were filtered out. The mean expression and standard deviation of each gene were calculated over the 30-bins based on their abundances in each cell. Then, Taylor's parameters were estimated by fitting the log of means and standard deviations to a linear regression. The segmented R package v1.6–4 [43] was used to fit the log-transformed data to a segmented linear regression with one breakpoint. When fitting binned data to a segmented (biphasic) regression, mitochondrial genes and some selected nuclear genes were removed given that they always fit to an exponential distribution regardless of their mean expression, and therefore, they are likely not responding to infection and could influence the estimation of the parameters at some specific bins. Additionally, for binned data only, genes with more than 70% of zeros were filtered out when fitting the data to a biphasic model with one breakpoint in order to avoid that genes with a high number of zeros dominate the fit of genes below the breakpoint. A simple schematic of the structure of the data used to estimate Taylor's parameters is shown in S3A Fig, and S3B Fig shows a representation of the binned data used to investigate the progression of Taylor's parameters along the course of infection.

Simulation of increasing technical noise in uninfected cells

A down-sampled dataset was created for each cell type to simulate the expected increase in cellular transcript dropout due to viral RNA accumulation. To create a simulated cell, the transcriptional profile of an uninfected cell was used to randomly sample *n* transcripts, where *n* corresponds to the total number of cellular UMIs from an infected cell, and the probability of sampling a transcript from a given gene is its abundance in the uninfected cell. Sampling transcripts based on the gene abundances of an uninfected cell and the number of cellular UMIs from an infected one will create a simulated cell that will inherit the transcriptional profile of the uninfected cell and the sampling noise of the infected cell. Simulated cells are ordered based on the viral RNA accumulation of the infected cells that were used to simulate their sampling noise, and for each cell type, there are as many simulated cells as there are infected cells. The Seurat package v4.3.0 [44] was used for downstream analyses of simulated and uninfected cells. Counts were log-normalized, and a standard clustering analysis was performed, where the top ten principal components (PCs) were used for clustering and uniform manifold approximation and projection (UMAP) dimensional reduction. When dividing the simulated data into 30 bins, genes with more than 95% of zeros were filtered out before fits to Taylor's law. For this simulated data, the progression through pseudotime should reflect increase in sampling noise.

Gene rank stability

The rank stability index (*RSI*) of one gene is defined from its rank, determined from abundances matrices of cells ordered according to their viral load. Due to the high prevalence of zeros due to dropout, ties were resolved by randomization to avoid overestimation of stability of the less expressed genes. The *RSI* of each gene was computed based on its observed rank hops, *D*, (*i.e*., the sum of the absolute number of rank differences between ordered adjacent cells) divided by the number of total possible rank hops as $RSI = \left(1 - \frac{D}{(N-1)(t-1)}\right)^k$, where *N* is the number of genes (rows), *t* is the number of cells (columns), and the power index $k = 4$ is arbitrarily chosen to increase the resolution of stable elements according to [12]. S3C Fig shows a representation of a matrix containing the rank of each gene that was used to calculate their associated *RSI*.

Estimation of punctual rank stability

To investigate signals of punctual rank stability, *RSI* values were calculated from the mean gene abundance at each bin instead of individual cells. Only genes that were expressed in at least one cell in every bin were further analyzed. Genes that displayed punctual stability, *i.e.*, that presented higher stability at some point along the infection, were determined based on a resampling strategy with 1000 replicates. In addition, for each replicate, an *RSI* was calculated from a matrix where the order of the bins was shuffled, with the exception of the first and last ones. The probability of finding an *RSI* value at least as high as the observed *RSI* of a given gene was calculated by applying a survival function (1 –empirical cumulative distribution function) estimated from the *RSI* values calculated from shuffling. Genes with *RSI* values with a false discovery rate (FDR) < 0.05 were considered to have undergone through a change in their rank stability at some point throughout the course of infection. A punctual stability index (PSI) was calculated by dividing the gene *RSI* by the mean *RSI* of the replicates, where $PSI > 1$ is indicative of punctual stability. A schematic representation of the data transformation that was employed to estimate the punctual rank stability of each gene is shown in S3D Fig.

Persistent behavior of gene rank

Long-range dependence and persistent behavior along the course of infection was investigated by estimating the Hurst exponent H for each gene separately for each cell type. A detailed explanation of the rescaled range analysis is available in S1 Appendix. Here, gene rank (see *Gene rank stability*; S3C Fig) was used to estimate H with the R package pracma v2.4.2 [45]. The robustness of this analysis was assessed by also estimating H for a set of control datasets that included the simulated datasets, uninfected cells, infected cells where cells were shuffled (and thus not ordered according to viral RNA accumulation) and a random matrix with the same number of rows (genes) and columns (cells) as the infected matrix where each value was drawn from a uniform distribution within the range [–1, 1]. A minimum window size of 50 was used when estimating H using the gene rank data. Genes with persistent behavior that simultaneously showed an $H \geq 0.7$ in the infected dataset and $H < 0.7$ in its respective simulated dataset were further investigated.

Gene ontology (GO) analyses

All gene set enrichment analyses were performed with the R packages clusterProfiler v4.8.2 [46] and org.Hs.eg.db v3.17.0 [47].

Network analyses

A reference human protein interactome was obtained from [22]. Network analyses were conducted in R using the igraph package v1.5.1 [48]. Genes were mapped to the interactome by converting the gene symbols to ENSEMBL IDs with org.Hs.eg.db v3.17.0 [47]. Hubs were determined by the 1.5×IQR rule of the degree of each node, and articulation points were considered to represent bottlenecks.

Differential gene expression of hIECs

In order to compare our results to the ones obtained by [18,19], differential gene expression analysis was conducted as described with MAST v1.26 [49] on the Seurat R object [44] that was made available by the authors. Briefly, pairwise comparisons of infected cells at 24 hpi, bystander cells at 24 hpi and mock-infected for all cells, as well as for the subset of immature enterocyte 2 cells, were performed on the "RNA" assay of the Seurat object. Of note, a list of differentially expressed genes in hBECs was made available by the authors [18], and as such, differential expression analysis on this dataset was not necessary.

Supporting information

S1 Fig. Characteristics of simulated datasets. (a) For each cell type, boxplots of the proportion of zeros of each gene for each bin for the infected and simulated dataset. (b) UMAP projections of the infected and simulated datasets. Each point represent a cell.
(EPS)

S2 Fig. Biphasic fit for binned data per cell. (a) Two selected bins from colon cells with no filtering of genes with a high proportion of zeros. A triphasic fit to Taylor's law can be observed in some bins when genes are not filtered out, as exemplified by the left panel. (b) Taylor's parameters of the biphasic fit. Breakpoint, V_1 and β_1 (*V* and β for genes with mean expression below the breakpoint), V_2 and β_2 (*V* and β for genes with mean expression above the breakpoint) and the matrix sparsity of each pseudotime bin.
(EPS)

S3 Fig. Schematic representation of the data structure used in each analysis. (a) Representation of a gene abundance matrix (left) that was log-transformed and fitted to a linear regression (right) to estimate Taylor's parameters. (b) Pseudotime binned data (left). Infected cells were sorted into bins so that the viral load of any cell in bin *i* is lower than the viral load of any cell in bin $i + 1$. Taylor's parameters were estimated for each bin (right). (c) Representation of a gene rank matrix used for the calculation of *RSI* shown in Fig 1A and for the estimation of the Hurst exponent of each gene. (d) Mean gene abundances of each bin (left) were used to generate a rank matrix (right) from which punctual rank stability analyses were conducted.
(EPS)

S1 File. Results of the rank stability dynamics analyses. The *RSI*, mean *RSI* of a 1000 replicates, *P*-value and adjusted *P*-value (*FDR*) and *PSI* of each gene for each dataset is shown in separate sheets. The last sheet corresponds to the GO enrichment analysis of the genes that exhibited signal of punctual rank stability concomitantly in all three cell types.
(XLSX)

S2 File. Results of the R/S analyses. The empirical *H* exponent of each gene for each dataset is shown in separate sheets for each cell type. The last sheet corresponds to the GO enrichment analysis of the genes that exhibited signal of strong persistent behavior concomitantly in all three cell types.
(XLSX)

S3 File. Hubs and bottlenecks in the human interactome. Sheets containing the ENSEMBL ID and gene symbol of genes identified as hubs and bottlenecks. Also includes hubs and bottlenecks that showed signal of punctual rank stability in all three cell types and strong persistence behavior.
(XLSX)

S4 File. Genes found to respond to infection that were missed by their original studies. Genes that showed signals of punctual rank stability and/or strong persistent behavior simultaneously in all cell types that were not found to be differentially expressed in their original studies. Also included are sheets for those genes that are hubs and for those that are bottlenecks in the human interactome.
(XLSX)

S1 Appendix. Detailed explanation of the rescaled range analysis.
(PDF)

Author Contributions

Conceptualization: João M. F. Silva, Jose Á. Oteo, Carlos P. Garay, Santiago F. Elena.

Data curation: João M. F. Silva.

Formal analysis: João M. F. Silva, Jose Á. Oteo, Carlos P. Garay.

Funding acquisition: Santiago F. Elena.

Investigation: João M. F. Silva, Jose Á. Oteo, Carlos P. Garay, Santiago F. Elena.

Methodology: João M. F. Silva.

Project administration: Santiago F. Elena.

Software: João M. F. Silva.

Supervision: Jose Á. Oteo, Carlos P. Garay, Santiago F. Elena.

Visualization: João M. F. Silva.

Writing – original draft: João M. F. Silva, Jose Á. Oteo.

Writing – review & editing: Carlos P. Garay, Santiago F. Elena.

References

1. Trapnell C, Cacchiarelli D, Grimsby J, Pokharel P, Li S, Morse M, et al. Pseudo-temporal ordering of individual cells reveals dynamics and regulators of cell fate decisions. Nat Biotechnol. 2014; 32: 381–386. https://doi.org/10.1038/nbt.2859 PMID: 24658644
2. Song D, Li JJ. PseudotimeDE: inference of differential gene expression along cell pseudotime with well-calibrated p-values from single-cell RNA sequencing data. Genome Biol. 2021; 22: 124. https://doi.org/10.1186/s13059-021-02341-y PMID: 33926517
3. Gutiérrez PA, Elena SF. Single-cell RNA-sequencing data analysis reveals a highly correlated triphasic transcriptional response to SARS-CoV-2 infection. Commun Biol. 2022; 5: 1302. https://doi.org/10.1038/s42003-022-04253-4 PMID: 36435849
4. Lazzardi S, Valle F, Mazzolini A, Scialdone A, Caselle M, Osella M. Emergent statistical laws in single-cell transcriptomic data. Phys Rev E. 2023; 107: 044403. https://doi.org/10.1103/PhysRevE.107.044403 PMID: 37198814
5. Islam S, Zeisel A, Joost S, La Manno G, Zajac P, Kasper M, et al. Quantitative single-cell RNA-seq with unique molecular identifiers. Nat Methods. 2014; 11:163–166. https://doi.org/10.1038/nmeth.2772 PMID: 24363023
6. Rostom R, Svensson V, Teichmann SA, Kar G. Computational approaches for interpreting scRNA-seq data. FEBS Lett. 2017; 591: 2213–2225. https://doi.org/10.1002/1873-3468.12684 PMID: 28524227
7. Zheng GX, Terry JM, Belgrader P, Ryvkin P, Bent ZW, Wilson R, et al. Massively parallel digital transcriptional profiling of single cells. Nat Commun. 2017; 8: 14049. https://doi.org/10.1038/ncomms14049 PMID: 28091601
8. Macosko EZ, Basu A, Satija R, Nemesh J, Shekhar K, Goldman M, et al. Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. Cell. 2015; 161: 1202–1214. https://doi.org/10.1016/j.cell.2015.05.002 PMID: 26000488
9. Han X, Wang R, Zhou Y, Fei L, Sun H, Lai S, et al. Mapping the mouse cell atlas by microwell-seq. Cell. 2018; 172: 1091–1107. https://doi.org/10.1016/j.cell.2018.02.001 PMID: 29474909
10. Taylor LR. Aggregation, variance and the mean. Nature. 1961; 189: 732–735. https://doi.org/10.1038/189732a0
11. Zhang Z, Geng J, Tang X, Fan H, Xu J, Wen X, et al. Spatial heterogeneity and co-occurrence patterns of human mucosal-associated intestinal microbiota. ISME J. 2014; 8: 881–893. https://doi.org/10.1038/ismej.2013.185 PMID: 24132077
12. Martí JM, Martinez-Martinez D, Rubio T, Gracia C, Pena M, Latorre A, et al. Health and disease imprinted in the time variability of the human microbiome. mSystems. 2017; 2: e00144–16. https://doi.org/10.1128/mSystems.00144-16 PMID: 28345059

13. Hurst HE. Long-term storage capacity of reservoirs. Trans Am Soc Civ Eng. 1951; 116: 770–799. https://doi.org/10.1061/TACEAT.0006518

14. Mandelbrot BB, Wallis JR. Robustness of the rescaled range R/S in the measurement of noncyclic long run statistical dependence. Water Resour Res. 1969; 5: 967–988. https://doi.org/10.1029/WR005i005p00967

15. Feder J. Fractals New York: Plenum Press; 1988

16. Ghorbani M, Jonckheere EA, Bogdan P. Gene expression is not random: scaling, long-range cross-dependence, and fractal characteristics of gene regulatory networks. Front Physiol. 2018; 9: 1446. https://doi.org/10.3389/fphys.2018.01446 PMID: 30459629

17. Oteo Araco JÁ, Oteo-García G. Mutations along human chromosomes: How randomly scattered are they? Phys Rev E. 2022; 106: 064404. https://doi.org/10.1103/PhysRevE.106.064404 PMID: 36671182

18. Ravindra NG, Alfajaro MM, Gasque V, Huston NC, Wan H, Szigeti-Buck K, et al. Single-cell longitudinal analysis of SARS-CoV-2 infection in human airway epithelium identifies target cells, alterations in gene expression, and cell state changes. PLoS Biol. 2021; 19: e3001143. https://doi.org/10.1371/journal.pbio.3001143 PMID: 33730024

19. Triana S, Metz-Zumaran C, Ramirez C, Kee C, Doldan P, Shahraz M, et al. Single-cell analyses reveal SARS-CoV-2 interference with intrinsic immune response in the human gut. Mol Sys Biol. 2021; 17: e10232. https://doi.org/10.15252/msb.202110232 PMID: 33904651

20. Choudhary S, Satija R. Comparison and evaluation of statistical error models for scRNA-seq. Genome Biol. 2022; 23: 27. https://doi.org/10.1186/s13059-021-02584-9 PMID: 35042561

21. Brant AC, Tian W, Majerciak V, Yang W, Zheng ZM. SARS-CoV-2: from its discovery to genome structure, transcription, and replication. Cell Biosci. 2021; 11: 136. https://doi.org/10.1186/s13578-021-00643-z PMID: 34281608

22. Luck K, Kim DK, Lambourne L, Spirohn K, Begg BE, Bian W, et al. A reference map of the human binary protein interactome. Nature. 2020; 580: 402–408. https://doi.org/10.1038/s41586-020-2188-x PMID: 32296183

23. Gordon DE, Hiatt J, Bouhaddou M, Rezelj VV, Ulferts S, Braberg H, et al. Comparative host-coronavirus protein interaction networks reveal pan-viral disease mechanisms. Science. 2020; 370: eabe9403. https://doi.org/10.1126/science.abe9403 PMID: 33060197

24. Shi CS, Qi HY, Boularan C, Huang NN, Abu-Asab M, Shelhamer JH, et al. SARS-coronavirus open reading frame-9b suppresses innate immunity by targeting mitochondria and the MAVS/TRAF3/TRAF6 signalosome. J. Immunol. 2014; 193: 3080–3089. https://doi.org/10.4049/jimmunol.1303196 PMID: 25135833

25. Li C, Wu K, Yang R, Liao M, Li J, Zhu Q, et al. Comprehensive analysis of immunogenic cell death-related gene and construction of prediction model based on WGCNA and multiple machine learning in severe COVID-19. Sci Rep. 2024; 14: 8450. https://doi.org/10.1038/s41598-024-59117-0 PMID: 38600309

26. Garg AD, De Ruysscher D, Agostinis P. Immunological metagene signatures derived from immunogenic cancer cell death associate with improved survival of patients with lung, breast or ovarian malignancies: A large-scale meta-analysis. Oncoimmunology. 2016; 5: e1069938. https://doi.org/10.1080/2162402X.2015.1069938 PMID: 27057433

27. Maremanda KP, Sundar IK, Li D, Rahman I. Age-dependent assessment of genes involved in cellular senescence, telomere, and mitochondrial pathways in human lung tissue of smokers, COPD, and IPF: associations with SARS-CoV-2 COVID-19 ACE2-TMPRSS2-Furin-DPP4 Axis. Front Pharmacol. 2020; 11: 584637. https://doi.org/10.3389/fphar.2020.584637 PMID: 33013423

28. DeOre BJ, Tran KA, Andrews AM, Ramirez SH, Galie PA. SARS-CoV-2 spike protein disrupts blood–brain barrier integrity via RhoA activation. J Neuroimmune Pharmacol. 2021; 16: 722–728. https://doi.org/10.1007/s11481-021-10029-0 PMID: 34687399

29. Han Q, Wang J, Luo H, Li L, Lu X, Liu A, et al. TMBIM6, a potential virus target protein identified by integrated multiomics data analysis in SARS-CoV-2-infected host cells. Aging. 2021; 13:9160. https://doi.org/10.18632/aging.202718 PMID: 33744846

30. Hasan MZ, Islam S, Matsumoto K, Kawai T. SARS-CoV-2 infection initiates interleukin-17-enriched transcriptional response in different cells from multiple organs. Sci Rep. 2021; 11: 16814. https://doi.org/10.1038/s41598-021-96110-3 PMID: 34413339

31. Li Y, Duche A, Sayer MR, Roosan D, Khalafalla FG, Ostrom RS, et al. SARS-CoV-2 early infection signature identified potential key infection mechanisms and drug targets. BMC Genom. 2021; 22: 125. https://doi.org/10.1186/s12864-021-07433-4 PMID: 33602138

32. Yoo JS, Sasaki M, Cho SX, Kasuga Y, Zhu B, Ouda R, et al. SARS-CoV-2 inhibits induction of the MHC class I pathway by targeting the STAT1-IRF1-NLRC5 axis. Nat Commun. 2021; 12: 6602. https://doi.org/10.1038/s41467-021-26910-8 PMID: 34782627

33. Zhou Y, Hou Y, Shen J, Huang Y, Martin W, Cheng F. Network-based drug repurposing for novel coronavirus 2019-nCoV/SARS-CoV-2. Cell Discov. 2020; 6: 14. https://doi.org/10.1038/s41421-020-0153-3 PMID: 32194980

34. Khan S, Shafiei MS, Longoria C, Schoggins JW, Savani RC, Zaki H. SARS-CoV-2 spike protein induces inflammation via TLR2-dependent activation of the NF-κB pathway. eLife. 2021; 10: e68563. https://doi.org/10.7554/eLife.68563 PMID: 34866574

35. Attiq A, Yao LJ, Afzal S, Khan MA. The triumvirate of NF-κB, inflammation and cytokine storm in COVID-19. Int Immunopharmacol. 2021; 101: 108255. https://doi.org/10.1016/j.intimp.2021.108255 PMID: 34688149

36. Camblor DG, Miranda D, Albaiceta GM, Amado-Rodríguez L, Cuesta-Llavona E, Vázquez-Coto D, et al. Genetic variants in the NF-κB signaling pathway (NFKB1, NFKBIA, NFKBIZ) and risk of critical outcome among COVID-19 patients. Hum Immunol. 2022; 83: 613–617. https://doi.org/10.1016/j.humimm.2022.06.002 PMID: 35777990

37. Zhu H, Chen J, Liu K, Gao L, Wu H, Ma L, et al. Human PBMC scRNA-seq–based aging clocks reveal ribosome to inflammation balance as a single-cell aging hallmark and super longevity. Sci Adv. 2023; 9: eabq7599. https://doi.org/10.1126/sciadv.abq7599 PMID: 37379396

38. Watts NR, Eren E, Palmer I, Huang PL, Huang PL, Shoemaker RH, et al. The ribosome-inactivating proteins MAP30 and Momordin inhibit SARS-CoV-2. PLoS One. 2023; 18: e0286370. https://doi.org/10.1371/journal.pone.0286370 PMID: 37384752

39. Sun C, Querol-Audí J, Mortimer SA, Arias-Palomo E, Doudna JA, Nogales E, et al. Two RNA-binding motifs in eIF3 direct HCV IRES-dependent translation. Nucleic Acids Res. 2013; 4: 7512–7521. https://doi.org/10.1093/nar/gkt510 PMID: 23766293

40. Sizova DV, Kolupaeva VG, Pestova TV, Shatsky IN, Hellen CU. Specific interaction of eukaryotic translation initiation factor 3 with the 5′ nontranslated regions of hepatitis C virus and classical swine fever virus RNAs. Journal of virology. 1998; 72: 477547–82. https://doi.org/10.1128/JVI.72.6.4775-4782.1998 PMID: 9573242

41. Kamel W, Noerenberg M, Cerikan B, Chen H, Järvelin AI, Kammoun M, et al. Global analysis of protein-RNA interactions in SARS-CoV-2-infected cells reveals key regulators of infection. Mol Cell. 2021; 81: 2851–2867. https://doi.org/10.1016/j.molcel.2021.05.023 PMID: 34118193

42. Witte S, Villalba M, Bi K, Liu Y, Isakov N, Altman A. Inhibition of the c-Jun N-terminal kinase/AP-1 and NF-κB pathways by PICOT, a novel protein kinase C-interacting protein with a thioredoxin homology domain. J Biol Chem. 2000; 275: 1902–1909. https://doi.org/10.1074/jbc.275.3.1902 PMID: 10636891

43. Muggeo VM. segmented: an R Package to Fit Regression Models with Broken-Line Relationships. R News. 2008; 8: 20–25. URL https://cran.r-project.org/doc/Rnews/

44. Hao Y, Hao S, Andersen-Nissen E, Mauck WM, Zheng S, Butler A, et al. Integrated analysis of multimodal single-cell data. Cell. 2021; 184: 3573–3587. https://doi.org/10.1016/j.cell.2021.04.048 PMID: 34062119

45. Borchers H. pracma: Practical Numerical Math Functions. Version 2.4.2 [R package]. 2022. Available from: https://CRAN.R-project.org/package=pracma

46. Wu T, Hu E, Xu S, Chen M, Guo P, Dai Z, et al. clusterProfiler 4.0: A universal enrichment tool for interpreting omics data. Innovation. 2021; 2: 100141. https://doi.org/10.1016/j.xinn.2021.100141 PMID: 34557778

47. Carlson M. org.Hs.eg.db: Genome wide annotation for Human. Version 3.17.0 [R package]. 2023. Available from: https://bioconductor.org/packages/release/data/annotation/html/org.Hs.eg.db.html

48. Csárdi G, Nepusz T, Traag V, Horvát S, Zanini F, Noom D, et al. igraph: Network Analysis and Visualization in R. Version 2.0.3 [R package]. Available from https://CRAN.R-project.org/package=igraph. https://doi.org/10.5281/zenodo.7682609

49. Finak G, McDavid A, Yajima M, Deng J, Gersuk V, Shalek AK, et al. MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. Genome Biol. 2015; 16: 278. https://doi.org/10.1186/s13059-015-0844-5 PMID: 26653891