



RESEARCH ARTICLE

**REVISED** A proposed molecular mechanism for pathogenesis of severe RNA-viral pulmonary infections [version 2; peer review: 4 approved]

Peter K. Rogan <sup>1,2</sup>, Eliseos J. Mucaki<sup>1</sup>, Ben C. Shirley<sup>2</sup>

<sup>1</sup>Biochemistry, University of Western Ontario, London, Ontario, N6A 2C8, Canada

<sup>2</sup>CytoGnomix Inc, London, Ontario, N5X 3X5, Canada

**V2** First published: 07 Aug 2020, 9:943  
<https://doi.org/10.12688/f1000research.25390.1>

Latest published: 06 Jan 2021, 9:943  
<https://doi.org/10.12688/f1000research.25390.2>

**Abstract**

**Background:** Certain riboviruses can cause severe pulmonary complications leading to death in some infected patients. We propose that DNA damage induced-apoptosis accelerates viral release, triggered by depletion of host RNA binding proteins (RBPs) from nuclear RNA bound to replicating viral sequences.

**Methods:** Information theory-based analysis of interactions between RBPs and individual sequences in the Severe Acute Respiratory Syndrome CoronaVirus 2 (SARS-CoV-2), Influenza A (H3N2), HIV-1, and Dengue genomes identifies strong RBP binding sites in these viral genomes. Replication and expression of viral sequences is expected to increasingly sequester RBPs - SRSF1 and RNPS1. Ordinarily, RBPs bound to nascent host transcripts prevents their annealing to complementary DNA. Their depletion induces destabilizing R-loops. Chromosomal breakage occurs when an excess of unresolved R-loops collide with incoming replication forks, overwhelming the DNA repair machinery. We estimated stoichiometry of inhibition of RBPs in host nuclear RNA by counting competing binding sites in replicating viral genomes and host RNA.

**Results:** Host RBP binding sites are frequent and conserved among different strains of RNA viral genomes. Similar binding motifs of SRSF1 and RNPS1 explain why DNA damage resulting from SRSF1 depletion is complemented by expression of RNPS1. Clustering of strong RBP binding sites coincides with the distribution of RNA-DNA hybridization sites across the genome. SARS-CoV-2 replication is estimated to require 32.5-41.8 hours to effectively compete for binding of an equal proportion of SRSF1 binding sites in host encoded nuclear RNAs. Significant changes in expression of transcripts encoding DNA repair and apoptotic proteins were found in an analysis of influenza A and Dengue-infected cells in some individuals.

**Conclusions:** R-loop-induced apoptosis indirectly resulting from viral replication could release significant quantities of membrane-

**Open Peer Review**

Reviewer Status

	Invited Reviewers			
	1	2	3	4
<b>version 2</b> (revision) 06 Jan 2021			 report	 report
<b>version 1</b> 07 Aug 2020	 report	 report	 report	 report

- Maurizio Romano** , University of Trieste, Trieste, Italy
- Mansi Srivastava**, Indiana University School of Medicine, Indianapolis, USA
- Gregory Fonseca**, McGill University, Montreal, Canada
- Ian Eperon**, University of Leicester, Leicester, UK

Any reports and responses or comments on the article can be found at the end of the article.

associated virions into neighboring alveoli. These could infect adjacent pneumocytes and other tissues, rapidly compromising lung function, causing multiorgan system failure and other described symptoms.

### Keywords

SARS-CoV-2, Influenza A, HIV-1, Dengue Virus, Apoptosis, R-loop, DNA damage, RNA binding protein

**Corresponding author:** Peter K. Rogan ([progan@uwo.ca](mailto:progan@uwo.ca))

**Author roles:** **Rogan PK:** Conceptualization, Formal Analysis, Funding Acquisition, Investigation, Methodology, Project Administration, Supervision, Validation, Writing – Original Draft Preparation, Writing – Review & Editing; **Mucaki EJ:** Data Curation, Formal Analysis, Investigation, Software, Validation, Writing – Original Draft Preparation, Writing – Review & Editing; **Shirley BC:** Data Curation, Methodology, Software, Visualization, Writing – Review & Editing

**Competing interests:** PKR cofounded and BCS is an employee of CytoGnomix Inc.

**Grant information:** PKR acknowledges Compute Canada for a special allocation of computing resources dedicated to COVID-19 research. He has been supported previously by the Canada Foundation for Innovation and Canada Research Chairs. *The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.*

**Copyright:** © 2021 Rogan PK *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 work is properly cited.

**How to cite this article:** Rogan PK, Mucaki EJ and Shirley BC. **A proposed molecular mechanism for pathogenesis of severe RNA-viral pulmonary infections [version 2; peer review: 4 approved]** F1000Research 2021, 9:943 <https://doi.org/10.12688/f1000research.25390.2>

**First published:** 07 Aug 2020, 9:943 <https://doi.org/10.12688/f1000research.25390.1>

**REVISED Amendments from Version 1**

This revision includes changes requested by the reviewers, including a comparative analysis of information weight matrices of RNA binding proteins with experimental mock controls, a revised Scatchard analysis of inhibitory viral SRSF1 binding sites vs host transcriptome (Figure 6), elimination of a paragraph in the Discussion that was tangential to the proposed mechanism, and new literature citations (Ref. 27, 54 and 57). We have also updated the Zenodo archive associated with this study (Ref. 39).

**Any further responses from the reviewers can be found at the end of the article**

## Introduction

### Background

RNA viruses have long been known as an important source of zoonotic disease transmission<sup>1</sup>. In these infections, a key question that needs to be answered is which infected individuals will progress from mild to severe symptoms that require intensive care? While complex underlying conditions increase susceptibility, severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) and Influenza A can lead to severe or lethal outcomes regardless of the age or health status in certain individuals. The Chinese and the initial US patients with SARS-CoV-2 showed that higher viral replication and multiplicity of infection are evident in severely ill individuals<sup>2-4</sup>. Textbook depictions of viral release and infection indicate budding from the cell membrane. This explanation might not adequately explain the rapid onset of symptoms and transmissibility seen in some individuals infected with these agents. We suggest that these factors can be explained by a cytopathology of induced lytic events, releasing high titers of virus. Programmed cell death (apoptosis), which has been suggested to occur in RNA viral conditions such as Influenza, is activated through innate immunity, with concomitant inflammatory responses. Viral RNA has been suggested to signal Toll-Like receptors and type I interferon expression, which binds to its receptor, IFNAR, and stimulates induction of PCD genes such as FasL or TRAIL<sup>5</sup>.

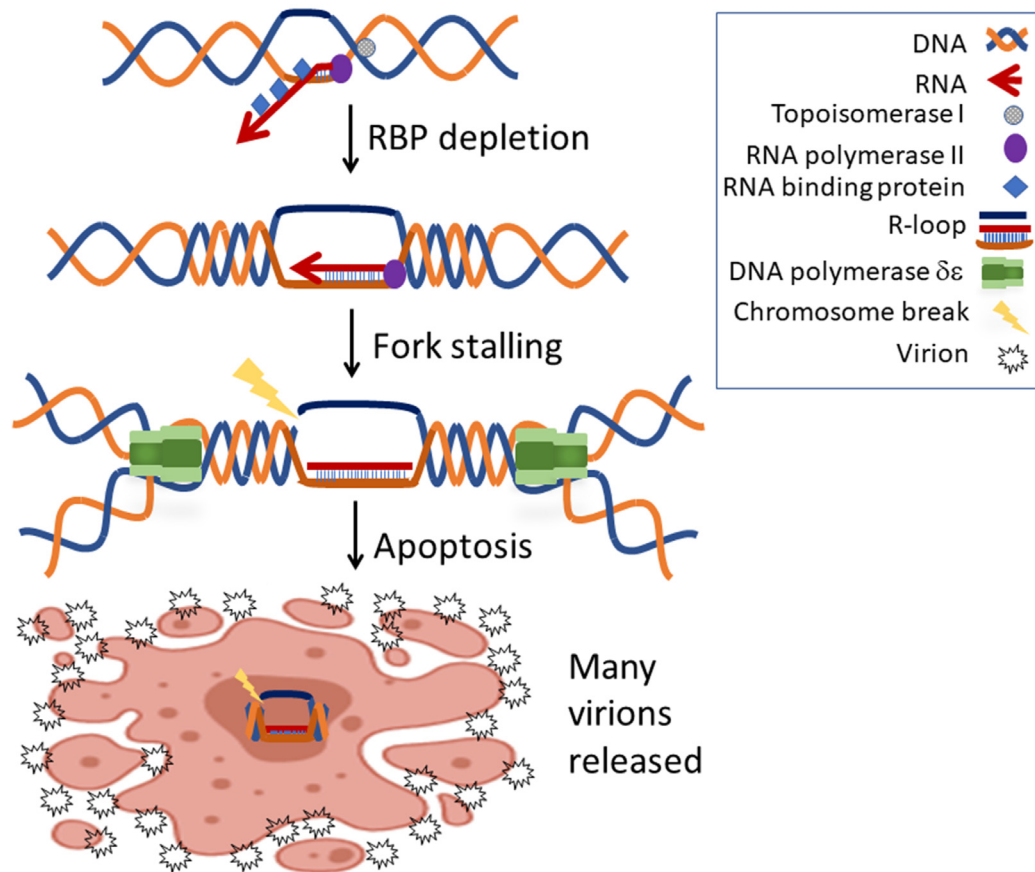
We propose an alternative mechanism in which infection of RNA virus triggers unrepaired sites of chromosomal breakage, causing apoptosis and consequentially, high-titer viral release (Figure 1). This is precipitated by the binding of RNA binding proteins (RBPs) to viral genomes and transcripts instead of nuclear transcripts, to prevent destabilization of chromosome structure. This study identifies the sequences, locations and abundance of these binding sites and presents evidence for specific expression changes in DNA damage genes in Influenza and Dengue infections and evidence of expression changes consistent with induction of apoptosis. The damage is thought to arise as the result of replication forks colliding with R-loops formed by host transcripts. Ordinarily these structures are mitigated through formation of stable interactions with frequently bound endogenous RBPs<sup>6</sup>.

The SR protein family consists of RNA binding proteins that play significant roles in the regulation of mRNA splicing<sup>7</sup>. SRSF1

(formerly ASF/SF2) is an exonic splicing enhancer (ESE) that has been shown to interact with the U1 snRNP and recruit the protein to the donor (5') splice site<sup>8,9</sup>. However, binding of SRSF1 to nascent transcripts has also been shown to play a significant role in genome stability, first described in reference 10, whereby the presence of SRSF1 bound to pre-mRNA repressed the formation of DNA:RNA hybrids, which led to R-loops, double-stranded breaks, and a hypermutation phenotype. This phenotype could be corrected not only by increasing RNase H expression (to eliminate DNA:RNA hybrids), but with the overexpression of the RNA binding protein RNPS1<sup>11</sup>. RNPS1, part of the apoptosis-and splicing-associated protein (ASAP) complex, can directly interact with SRSF1<sup>12</sup> and could possibly help recruit SRSF1 to ESE sites<sup>13</sup>. Other RNA binding proteins have been shown to increase genome instability when depleted, including *THOC1*<sup>14</sup>, *MFAP1*<sup>15</sup>, and *FIP1L1*<sup>16</sup>.

Binding sites for these RBPs are identified using information theory (IT)-based sequence analysis, which has proven both theoretically and in numerous practical examples to be an accurate approach for predicting binding affinities of nucleic acid sequences recognized by particular DNA or RNA binding proteins<sup>17</sup>. IT can be used to identify binding sites, and to evaluate the impact a sequence variant may have on binding site strength<sup>18</sup>. IT has been applied in studies which involved mRNA splicing<sup>19,20</sup>, splicing regulatory factors (SRFs<sup>21,22</sup>), other RNA binding proteins<sup>23</sup> and transcription factor binding sites (TFBS<sup>24,25</sup>), and has been used to accurately predicted level of gene expression and identify causative mutations in a wide spectrum of diseases<sup>17</sup>. IT-based analysis has the distinct advantage to other bioinformatic approaches as the predicted information content (known as  $R_i$ ; measured in bits) can be quantified as binding site affinity as it is related to thermodynamic entropy<sup>26</sup>. The binding affinity of a sequence predicted by IT has been shown experimentally to directly relate to the observed binding quantity of said sequence<sup>26</sup>. IT-based models are generated from a series of annotated binding sites for a particular RBP. The average strength of the sites used to generate said is referred to as its  $R_{sequence}$ . IT-based models can also be derived from high-throughput binding site identification techniques such as ChIP-seq (e.g. the derivation of TFBS models in 24). Information density-based clustering (IDBC) analysis, where groups of closely situated binding sites are evaluated based on their combined strength (their "information density") and intersite distances, has been applied along with these TFBS models in both the identification of TFBS-dense clusters, and accurate prediction of gene expression patterns<sup>25</sup>.

We and others have suggested that the viral genome binds to these RBPs (e.g. SRSF1 is enriched among SARS-CoV-2 RNA-protein interactions<sup>27</sup>) as well, and we define the locations of likely strong binding sites across the genomes of various RNA viruses. We propose that the replicating viral genome and transcriptome binds and sequesters these proteins, preventing their reimportation into the nucleus where they are normally needed for essential post-transcriptional activities. We theorize that incremental replication and transcription of viral RNAs in the cytoplasm creates a sink for these proteins, starving the host



**Figure 1. Proposed mechanism of high multiplicity of RNA viral infections.** Newly synthesized host RNA binding proteins (SRSF1, RNPS1) are required to stabilize nascent transcripts throughout the nucleus. During influenza or other viral infections, these proteins can be bound to viral genomes and transcriptomes. As viral replication and transcription proceeds, these nucleic acids containing strong binding sites for these RBPs in the cytoplasm (SARS-CoV-2) and nucleus (Influenza) that compete with host RNAs and deplete these proteins from the nucleus. This enables nascent transcripts to reanneal with transcription templates, and R-loops are formed. If not removed by RNase H or other helicases, unresolved R-loops at numerous genomic loci triggers genomic instability. Their frequency and density of unrepaired chromosome damage would be expected to overwhelm DNA repair components (BRCA1/2, FANCD1, and XPC), inducing multiple chromosomal strand breaks in each cell<sup>10</sup>. These breakage events initiate apoptosis, releasing a high multiplicity of infectious viral particles.

nucleus, and initiating a series of events that release viral particles into the lumen, enabling rapid infection of neighboring lung epithelial cells (Figure 1). An infographic has been created to provide a detailed step-by-step guide to the proposed mechanism, from the initial viral infection to spread of infection to the lungs and other major organs, leading to lowered blood oxygen levels, and multi-system organ failure<sup>28</sup>.

#### Proposed molecular pathogenetic mechanism of RNA-viral infection

RNA viral genomes of Influenza viruses replicate in the nucleus and are processed by host RNA spliceosomes. For example, the M and NS segments of the Influenza genome are processed using the host splicing mechanism<sup>29</sup>. Viral RNAs, like host transcripts, are capable of sequence specific binding to RBPs. This can conceivably deplete RBPs from host encoded RNAs, where they ordinarily function. These unbound RNAs are capable

of hybridizing to the non-template derived strand of the chromosome<sup>30</sup>. RNA naturally forms a stronger bond to DNA than DNA does to itself, especially rG:dC hybrids<sup>10</sup>. As a result, mRNAs would replace DNA by hybridizing complementary bases, resulting in R-loop formation, and can lead to DNA damage.

The RNA spliceosome regulator SRSF1 acts on exonic splicing enhancer sequences in pre-mRNA and forms RNP complexes with nascent mRNA precursors. Aside from its established role in enhancing exon recognition<sup>9</sup>, binding of SRSF1 to these transcripts is required to prevent or destabilize the formation of R-loops<sup>10</sup>. R-loops are derived from RNA transcripts that anneal to the chromosomal strand complementary to the transcription template stand. If not eliminated, these structures pose a threat to genomic integrity as targets for DNA damage. The structure of R-loops consists of two duplex-single strand junctions which are

recognized by nucleases that cleave the DNA<sup>30</sup>. DNA fragmentation causes a G2 phase cell cycle arrest which can potentially lead to cell death<sup>11</sup>. R-loops that are not targeted by nucleases are nonetheless still non-functional and thus, inflict damage on the cell<sup>10</sup>. As RNA viruses enter the cell and replicate, the nucleic acid sequences they encode divert RBPs such as SRSF1 away from binding to nuclear RNA transcripts, thus promoting the creation of R-loops.

RNPS1 is a pre-mRNA splicing activator protein that functions together with SRSF1 to form RNP complexes on nascent transcripts<sup>13,31</sup>, but also has a role in preventing transcriptional R-loop formation<sup>11</sup>. RNPS1 also suppresses high molecular weight DNA fragmentation at high expression levels. These two proteins work together but have independent mechanisms as RNPS1 cannot compensate for SRSF1 splicing function in its absence and vice versa<sup>11</sup>.

In Dengue virus, the protein called NS5 binds to host spliceosome complexes and modulates endogenous splicing to change mRNA isoform abundance of antiviral factors. By also interacting with U5 snRNP particles, it reduces the efficiency of pre-mRNA processing, hence resulting in a less restrictive environment for viral replication. It has also been shown that NS5 interacts with the host protein, RNPS1, which disrupts normal nuclear RNA binding processes<sup>32</sup>.

Viral infections interfere with post-transcriptional processing of host pre-mRNA including splicing, capping, and translation during viral invasion. Since SRSF1 binds and interacts with pre-mRNA during the earliest stages of splicing, diversion of SRSF1 and other spliceosomes to other RNA sequences depletes the cell's resources. Normally, cellular mRNA is 7-methylguanosine cap is added to the 5' end to protect the sequence from degradation. However, Influenza carries proteins that has "cap-snatching" abilities<sup>33</sup>. Influenza snatches the 5' cap by cleaving the mRNA 10 to 15 nucleotides away from the guanosine and this cap is used to prime transcription of the virus. Finally, during viral infections, all RNA processing mechanisms are now being shared between two genomes. Ultimately, as transcriptional and translation mechanisms fail to facilitate the mRNA, they will create R-loops with DNA, cause DNA damage, and induce higher expression of DNA repair genes (such as *DDB2*; see Results).

Unrepaired damaged DNA that encounters a replication fork leads to unresolved double strand breaks, triggering apoptosis. The quantity of virus that escapes into tissues, blood and other conduits (e.g. lymphatic), and other systems would likely dwarf the amount that is released by conventional viral budding from the cell membrane. This viral load will likely overwhelm the immune system in individuals who are already immune deficient and might provoke a systemic inflammatory response (like a cytokine storm). However, the high titer of virus is likely to infect neighboring cells and other tissues. The extent of the apoptotic response may be the distinguishing finding which separates the patients who survive the infection from those who end up in intensive care, develop pulmonary insufficiency and multi-system failure.

The deficiency in SRSF1 and other RBPs in the nuclei of Influenza, Dengue or SARS-CoV-2 infected cells does not require any specialized mechanism. Assuming that the virus is replicating freely in the cytoplasm (or nucleus, in the case of Influenza), the significant excess of unpackaged, replicated viral RNA acts as a sponge to sequester newly synthesized, folded RBPs. Based on mass action, the quantity of RBPs that would be transported into the nucleus for host mRNA processing would have a much-diminished nuclear stoichiometry in comparison with normal, uninfected cells.

## Results

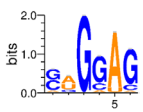
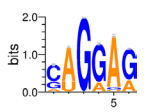
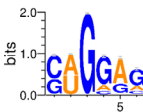
### Derivation of CLIP-based SRSF1 and RNPS1 information theory-based models

Cells depleted of SRSF1 has been shown to have unstable genomes which can be corrected by overexpression of RNPS1<sup>11</sup>. In order to investigate the significance of SRSF1 and RNPS1 binding in viral genomes, we first developed information theory-based models for the recognition sequences for each of these proteins using binding site datasets derived from transcriptome-wide RNA binding protein datasets of CLIP sequencing data. We then scanned multiple RNA viral genomes, as well as the human transcriptome, with these derived models to identify and predict the strength of individual binding sites.

An Information Weight Matrix (IWM) for SRSF1 has been previously derived<sup>21</sup>, however, it was only based on very small set of manually curated binding sites (N=28). We therefore derived new SRSF1 IWMs using publicly available eCLIP data (two separate replicates from 34). Multiple SRSF1 models exhibited very similar binding motifs, however, their differences justified our analyses using the two most divergent IWMs in this study. These models are referred to as SRSF1 "Replicate 1" and "Replicate 2" models, as they are models derived from two separate eCLIP experimental replicates from the same study. SRSF1 "Replicate 1" is derived from a larger number of eCLIP peaks (50,000) compared to 5,000 for "Replicate 2". Since SRSF1 "Replicate 1" was derived from a greater number of sites, it therefore may be more accurate for detection of weaker SRSF1 binding sites.

A distinct IWM was derived by iCLIP data from transcriptome-wide, protein crosslinking to sequences recognized by RNPS1<sup>31</sup>. It was evident that the RNPS1 IWM and the newly derived SRSF1 models exhibited a similar pattern of nucleotide conservation based on comparison of their respective sequence logos (Table 1). STAMP, a software program which analyzes position weight matrices of nucleic acid (or protein) motifs, was used to compare these models based on their e-values<sup>35</sup>. The SRSF1 "Replicate 1" and "Replicate 2" models were both highly similar (motif alignment e-value < 0.01) to the RNPS1 IWM (Table 1), implying that individual binding sites recognized by these two factors are similar. Indeed, the motif similarity between these two factors has been described<sup>13</sup>. We suggest that this overlap in their respective binding affinities may account for why RNPS1 overexpression can enable SRSF1-deficient cells to overcome their inherent genomic instability phenotype.

**Table 1. Comparison of Derived SRSF1 and RNPS1 Information Models and Binding Sites in Genomes.**

Factor	SRSF1 [Rep1] <sup>1,2</sup>	SRSF1 [Rep1] / RNPS1 Model Comparison		RNPS1 <sup>1</sup>	SRSF1 [Rep2] / RNPS1 Model Comparison		SRSF1 [Rep2] <sup>1,2</sup>				
Sequence Logo		-			-						
$R_{sequence}$ (bits)	6.7 ± 2.1	-		7.8 ± 1.9	-		6.4 ± 2.1				
Motif Similarity (E-value) <sup>3</sup>	-	5.0e-09		-	1.1e-09		-				
No. of Expressed Binding Sites (A549; ≥ 0 bits) <sup>4</sup>	1.3e08	5.4e07 (57%) <sup>7</sup>		9.3e07	6.4e07 (69%)		1.5e08				
No. of Expressed Binding Sites (Pneumocytes; ≥ 0 bits) <sup>5</sup>	6.8e07	2.9e07 (58%)		5.0e07	3.4e07 (69%)		7.9e07				
No. of Sites (SARS-CoV-2; + - strand)	≥ 0 bits	860	732	435 (72%)	305 (65%)	608	466	363 (60%)	273 (59%)	810	772
	≥ 1/2 $R_{seq}$	311	232	131 (51%)	86 (44%)	256	196	155 (61%)	115 (59%)	376	358
	≥ $R_{seq}$	31	42	16 (46%)	10 (40%)	35	25	35 (100%)	25 (100%)	60	33
No. of Sites (Influenza A; + - strand) <sup>6</sup>	≥ 0 bits	697	339	289 (61%)	118 (63%)	475	188	268 (56%)	129 (69%)	616	388
	≥ 1/2 $R_{seq}$	263	118	122 (49%)	47 (55%)	248	85	162 (65%)	65 (76%)	373	188
	≥ $R_{seq}$	50	23	24 (53%)	12 (75%)	45	16	45 (100%)	16 (100%)	84	35

<sup>1</sup> RNPS1 model derived from publicly available iCLIP data (E-MTAB-4215; ArrayExpress), while SRSF1 models were derived from eCLIP data (ENCSR456FVU; ENCODE Data Coordination Center); <sup>2</sup> SRSF1 [Rep1] and [Rep2] were derived from eCLIP dataset replicate 1 [50,000 peaks] and replicate 2 [5,000 peaks], respectively; <sup>3</sup> RNA binding motifs were compared using STAMP<sup>35</sup> using the Pearson Correlation Coefficient distance metric<sup>36</sup>; <sup>4</sup> A549 cell line expression from GSE141171 dataset; <sup>5</sup> Primary type II pneumocyte expression from GSE86618 dataset; <sup>6</sup> Influenza A virus H3N2 strain (Ontario/104-25/2012).

<sup>7</sup> RNPS1 sites used as denominator for all percentages.

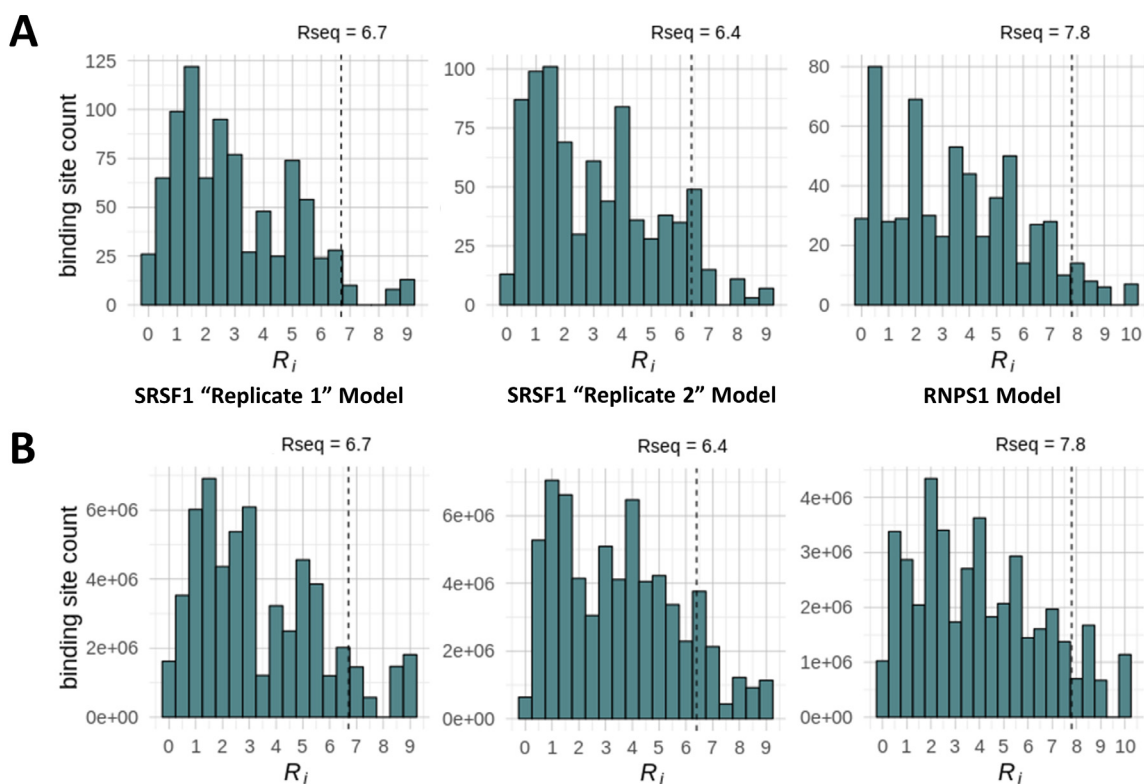
We also derived IWMs from negative controls in the eCLIP and iCLIP resources which consisted of sequence libraries constructed from crosslinking studies of mock substrates<sup>31,34</sup>. The resultant IWMs did not resemble those obtained from crosslinking RNPS1 and SRSF1 to their cognate binding sites. The LOD scores (logarithm of the odds ratio of the respective e-values of RBP motifs relative to different mock sequence motifs) ranged from 3.8 to 6.1 for RNPS1 and from 4.4 to 8.8 for SRSF1.

### RBP binding sites in RNA viral genomes

The newly derived SRSF1 and RNPS1 models (as well as an hnRNP A1 model to act as a positive control [its derivation described in 22], as the RBP has been shown to regulate transcription of beta coronaviral genes<sup>37</sup>) were used to scan the genomes of multiple RNA viruses: Dengue (Type 3), HIV (Strain B and C), Influenza A (H3N2; two separate strains), and SARS-CoV-2 (NC\_045512.2). In coronaviruses, the infectious particle contains the positive strand, but the negative strand copy of the RNA is generated for protein translation<sup>38</sup> and may be available to bind RBPs. Therefore, both the positive and negative strands of the viral genomes were scanned for SRSF1, RNPS1 and hnRNP A1 binding, regardless of the replication mechanism of the virus.

The SARS-CoV-2 genome was determined to contain >600 SRSF1 (with either SRSF1 model) and RNPS1 binding sites

(Table 1). However, histograms which illustrate the distribution of the strengths of all SRSF1 and RNPS1 binding sites in SARS-CoV-2 (Figure 2A) reveal that the majority of these are weak sites (where  $R_i < R_{sequence}$ ) that may not be used. We therefore focused downstream analysis on strong binding sites (where  $R_i \geq R_{sequence}$ ) of each IWM ( $R_{sequence}$ : 6.7 bits for the SRSF1 “Replicate 1” model; 6.4 bits for the SRSF1 “Replicate 2” model; 7.8 bits for the RNPS1 model; and 4.6 bits for the hnRNP A1 model). There are only 35 RNPS1 and between 31-60 SRSF1 binding sites (depending on SRSF1 model) on the positive strand of the SARS-CoV-2 genome that meet this  $R_{sequence}$  threshold (Table 1). The total number of SRSF1 binding sites within all other viral genomes tested are provided in Table 2, while RNPS1 and hnRNP A1 binding site counts are available within a Zenodo repository for this study (extended data<sup>39</sup> Section 1 – Table 1). The hnRNP A1 model consistently predicts more strong binding sites than the SRSF1 and RNPS1 models across all the RNA viral genomes tested, as well as in the human gene controls. This is likely partially due to its relatively low  $R_{sequence}$  threshold compared to the other models used. Interestingly, we observed significantly more SRSF1 and RNPS1 binding sites on the positive strand compared to the negative strand for all tested RNA viral genomes (exception: sites in SARS-CoV-2 predicted by SRSF1 “Replicate 1” model). This phenomenon was observed in both positive-strand and negative-strand RNA viruses (e.g. both Influenza A strains tested). This imbalance was not observed in the human genes tested (Table 2).



**Figure 2.**  $R_i$  of SRSF1 and RNPS1 Binding Sites in the SARS-CoV-2 and Human Genomes. Histograms display the distribution of  $R_i$  values for SRSF1 ["Replicate 1" and "Replicate 2" models] and RNPS1 binding sites strengths identified in **A**) the SARS-CoV-2 viral genome, and **B**) all transcribed regions in the human genome.

Previously, tightly organized groups of transcription factor binding sites (TFBS) were identified using information dense clustering<sup>25,40</sup>. We applied this method to identify regions of the viral genomes with large concentrations of binding sites (extended data<sup>39</sup> Section 1 – Table 2). Clusters of weak SRSF1 and RNPS1 sites are common (e.g. there are 5 SRSF1 clusters on the positive strand of SARS-CoV-2; extended data<sup>39</sup> Section 1 – Tables 2A and 2B); however, clusters made up exclusively of strong binding sites ( $R_i \geq R_{sequence}$ ) are extremely rare in the viral genomes tested.

We observed that all strong RNPS1 sites were also predicted to be strong ( $R_i \geq R_{sequence}$ ) by the SRSF1 "Replicate 2" model. This is not surprising, as the two models were found to have significantly similar binding motifs (Table 1). This overlap, as well as the location and strength of all other strong SRSF1 ("Replicate 2" model only) and RNPS1 binding sites, can be observed in Figure 3 where sites were mapped across the SARS-CoV-2 and Influenza A genomes. This was not observed, however, for SRSF1 "Replicate 1" despite its similarity to the RNPS1 model. For this SRSF1 model, nearly half of all strong RNPS1 sites were predicted to be weak ( $R_i$  below the  $R_{sequence}$  threshold).

Despite its low mutation rate, over 220 SARS-CoV-2 strains have already been identified, with potential mutational

hot spots of different geographic origins<sup>41</sup>. If the proposed mechanism does play a role in the severity of infection, then it is expected that various strains of SARS-CoV-2 would not significantly differ in numbers of binding sites, as no particular strain of SARS-CoV-2 has yet been proven to affect disease recovery (indeed, more transmissible strains have been identified but none more pathogenic<sup>42,43</sup>). To test this theory, genomes of 8 SARS-CoV-2 strains were downloaded from the [Global Initiative on Sharing All Influenza Data](#) (GISAID) database and analyzed using the IWMs for SRSF1, RNPS1 and hnRNP A1 (Table 3 for positive strand analysis; extended data<sup>39</sup> Section 1 – Table 3 for analysis of both strands). The particular strains that were selected were those that showed maximum divergence from one other based on analyses by [NextStrain](#) (which tracks the genomic epidemiology of SARS-CoV-2<sup>44</sup>). Binding site counts of different strains were within 90% across all strains, except for MT198652.1 (Spain), which contains an undetermined sequence where binding site differences are mapped. A strong consistency between binding site counts and strengths was noted, despite maximizing in the divergence between the selected SARS-CoV-2 strains. For RBPs binding, it was therefore not significant as to which SARS-CoV-2 sequence was selected for the subsequent analyses.

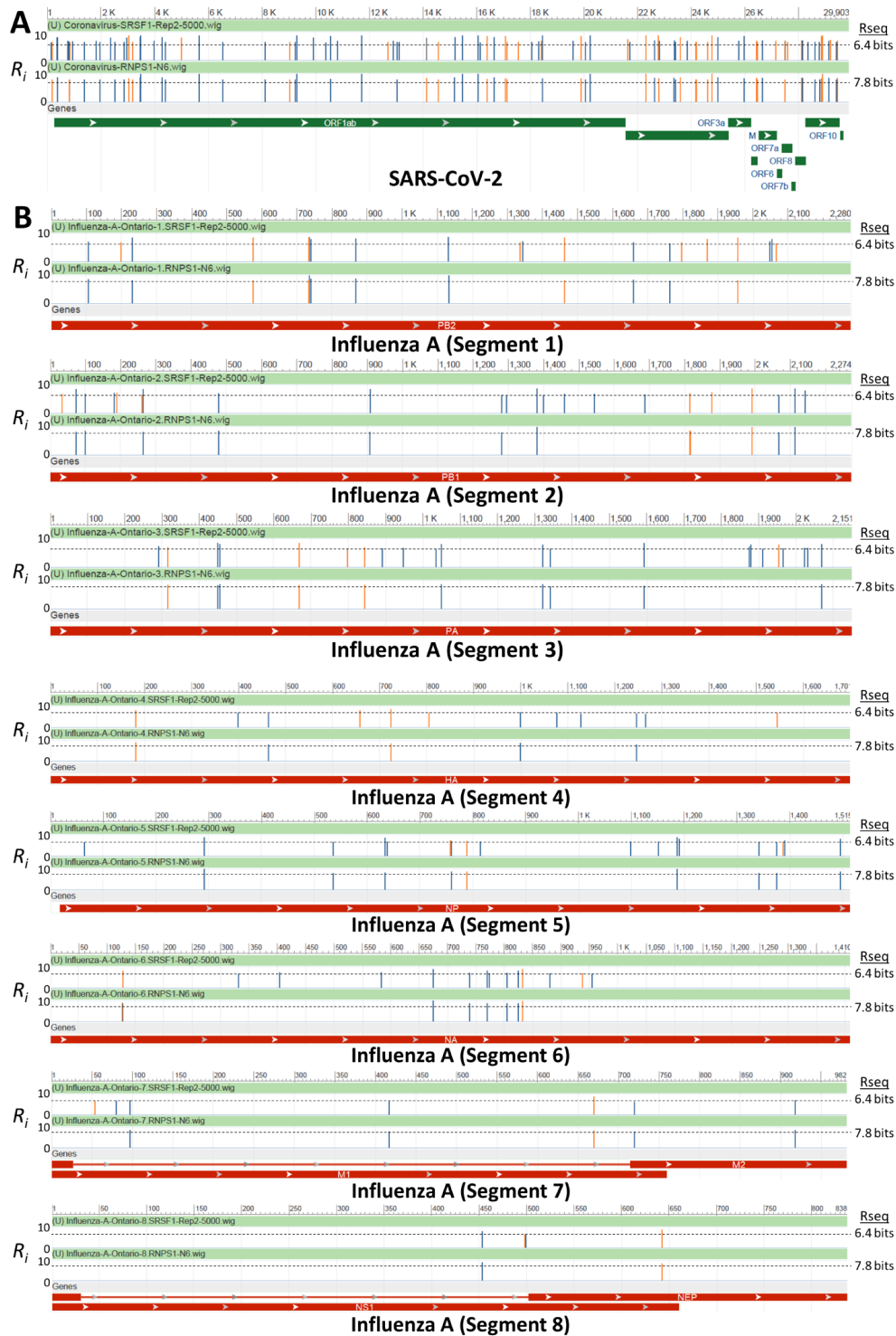
The absence of severe symptoms associated with the SARS-CoV-2 Singaporean strain (which features a deletion in ORF8

**Table 2.** SRSF1 Binding Site and Information-Dense Cluster Counts in RNA Viral Genomes.

Virus (Strain)	Length (nt)	SRSF1 Sites (Replicate 1; $\geq$ Rseq [6.7 bits])			SRSF1 Clusters (Replicate 1; $\geq$ 0.1 bits)			Strongest SRSF1 Cluster (bits; Replicate 1)	SRSF1 Sites (Replicate 2; $\geq$ Rseq [6.4 bits])			SRSF1 Clusters (Replicate 2; $\geq$ 0.1 bits)			Strongest SRSF1 Cluster (bits; Replicate 2)
		Both	+	-	Both	+	-		Both	+	-	Both	+	-	
Dengue Virus (Type 3)	10,707	65	47	18	28	26	2	162.1	107	85	22	26	24	2	107.1
HIV-1 (Strain B)	9,719	55	44	11	20	20	0	152.2	106	87	19	24	23	1	115.7
HIV-1 (Strain C)	9,031	58	48	10	21	18	3	142.8	103	81	22	16	14	2	143.8
Influenza A (Ontario)	13,151	73	50	23	23	21	2	100.4	119	84	35	24	21	3	111.9
Influenza A (Shanghai)	11,863	79	63	16	25	24	1	165.9	121	91	30	32	30	2	127.3
SARS-CoV-2 (NC_045512.2)	30,899	73	31	42	10	5	5	69.9	93	60	33	7	5	2	66.7
<i>IKBKB</i> (Human Gene)	61,352	615	305	310	244	124	120	288.2	803	410	393	288	143	145	577.1
<i>SIRT1</i> (Human Gene)	33,721	225	109	116	72	30	42	526.0	283	125	158	80	37	43	152.6
<i>WDR4</i> (Human Gene)	30,358	351	176	175	129	64	65	266.3	519	265	254	176	88	88	324.4

Columns labeled as "Both" indicate the number of binding sites or clusters on both strands of the viral genome.





**Figure 3. Distribution of SRSF1 and RNPS1 Binding Sites Across SARS-CoV-2 and Influenza A.** The viral genomes of **A**) SARS-CoV-2 (NCBI Reference Sequence: NC\_045512.2) and **B**) Influenza A virus (A/swine/Ontario/104-25/2012[H3N2]) were scanned for strong pre-existing binding sites for the RBP RNPS1 and SRSF1 (newly derived “Replicate 2” model). Custom wiggle tracks which contained those RBP of  $R_i \geq R_{sequence}$  were generated and visualized by NCBI Nucleotide. Track images were manually adjusted to indicate the strand in which the binding site was identified (blue vertical lines indicate sites on the positive strand, orange on the negative strand). The majority of sites predicted by the RNPS1 model were simultaneously predicted by the SRSF1 model, however the SRSF1 model identifies additional unique binding sites.

**Table 3. Binding Site Counts in Genome Sequences of Multiple Coronavirus Strains (Positive Strand only).**

Coronavirus Strain	Model			
	SRSF1 (Replicate 1)	SRSF1 (Replicate 2)	RNPS1	hnRNP A1
MT007544.1 (Australia)	31	60	35	573
MT066176.1 (Taiwan)	31	60	35	573
MT121215.1 (China)	31	60	35	573
MT163718.1 (USA)	31	60	35	573
MT188339.1 (USA)	31	60	35	572
MT198652.1 (Spain) <sup>a</sup>	28	57	33	532
MT198653.1 (Spain) <sup>a</sup>	31	60	35	558
NC_045512.2 (China)	31	60	35	573

<sup>a</sup> Sequences contains a small stretch of undefined nucleotides, which is likely contributing to the lower number of binding sites found.

[pos. 27,848 to 28,229]<sup>45</sup>), however, is not related to a significant loss of strong SRSF1 and RNPS1 binding sites. The SRSF1 “Replicate 1” model does not identify any binding sites ( $\geq R_{sequence}$ ) in this region. The SRSF1 “Replicate 2” model predicts 2 strong binding sites in this region, as does the RNPS1 model. There are 17 hnRNP A1 binding sites in this region, however there are 1,168 sites in total across the coronavirus genome; therefore, the missing hnRNP A1 sites account for only 1.4% of the total detectable hnRNP A1 binding sites.

Given the high Influenza A mutation rate, we evaluated the variability in RBP site count and affinities between strains, that is, whether these binding sites might be under selection for conservation of RBP binding. Four Influenza A strains (H3N2) from four separate clades (analogous to the SARS-CoV-2 strain selection procedure using NextStrain<sup>44</sup>; A/Denmark/316/2020; A/England/323/2019; A/Singapore/TT0333/2019; and A/Sydney/1017/2018) along with the two Influenza A strains previously selected (A/swine/Ontario/104-25/2012 and A/Duck/Shanghai/C84/2009) were analyzed and their genomes scanned for the presence of strong RNPS1, SRSF1 and hnRNP A1 binding sites (extended data<sup>39</sup> Section 1 – Table 4). Depending on the strain, 13 to 16 RNPS1 and 30 to 35 SRSF1 (“Replicate 2” model) binding sites were identified on the negative strand of Influenza A (a range of 16–23 binding sites for SRSF1 “Replicate 1”, and 221 to 241 strong hnRNP A1 binding sites). Thus, it appears as though the overall number of binding sites remains relatively consistent between each Influenza A strain, despite their divergent genomic sequences.

The locations of all predicted binding sites and information-dense clusters within the genome of each RNA virus tested has been made available within the extended data archive (Section 2<sup>39</sup>). This data is provided in the form of ‘bedgraph’ genome browser tracks. The locations of binding site clusters are also provided as lollipop plots within the archive (Section 3), as are the IWMs used to evaluate each site (Section 4).

### Human transcriptome analysis of RNA binding sites

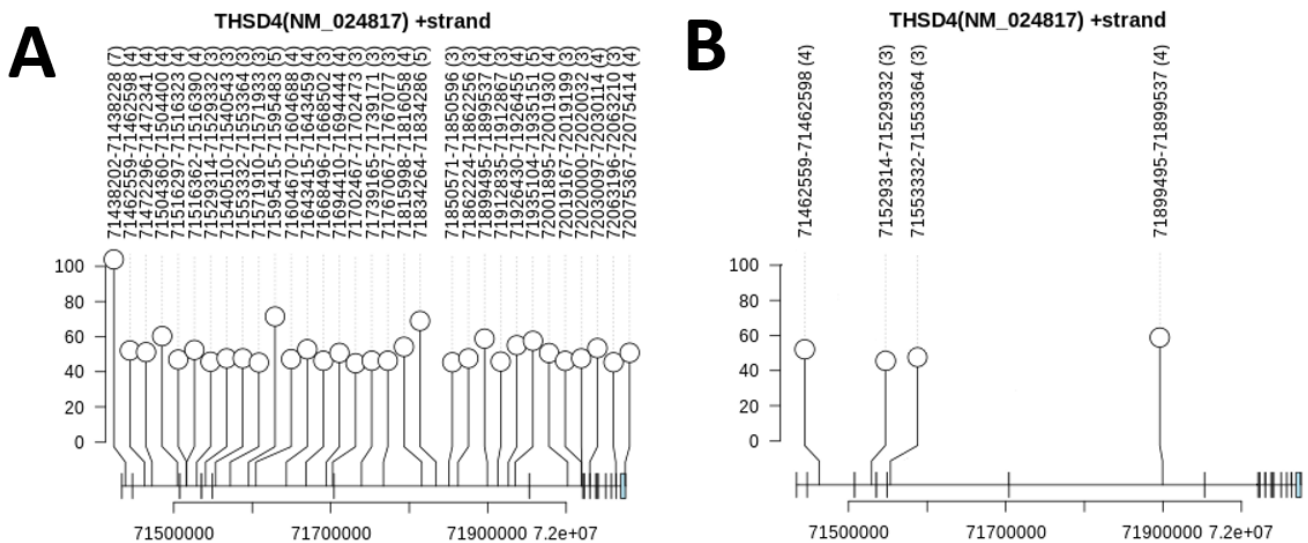
Each of these RNA viral genomes contain multiple strong RNA binding sites. The frequency of RBP binding in human transcriptomes was determined to relate the relative abundance of these proteins bound to viral RNAs compared to their normal reservoir in host nuclear RNA of infected cells. Expressed host gene sequences were scanned with IWMs for SRSF1, RNPS1 and hnRNP A1 to locate all potential binding sites throughout transcribed regions of the human genome, then partitioned among these genes based on their abundance in relevant cell types. These were compared with binding sites within 300nt of a known exon, as many of these RBPs have critical functions in exon recognition and maturation of mRNA splice isoforms (provided as bedgraph tracks in the Zenodo archive [Section 2]<sup>39</sup>). While the majority of these binding sites are considered weak ( $R_i < R_{sequence}$ ; Figure 2B), the numbers of strong (binding sites with  $R_i > R_{sequence}$ ) residing within transcribed regions are substantial (SRSF1 “Replicate 1” Model: 5,543,429; SRSF1 “Replicate 2” Model: 8,275,472; RNPS1: 4,368,943; hnRNP A1: 44,885,381). The intersite distance (the average distance between binding sites) appears to be inversely related to the overall number of binding sites, as the mean intersite distance between strong hnRNP A1 binding sites was considerably shorter than the distance between strong SRSF1 and RNPS1 binding sites (hnRNP A1: 24±40 nt; RNPS1: 149±248 nt; SRSF1 [“Replicate 1” model]: 105±241 nt; SRSF1 [“Replicate 2” model]: 89±197 nt; analysis using a maximum intersite distance threshold of 1,000nt). Regardless of these differences, however, this analysis illustrates that many strong binding sites are separated by < 200nt and highlights how densely arrayed these sites are in the human transcriptome.

The number of strong SRSF1, RNPS1 and hnRNP A1 binding sites ( $R_i \geq R_{sequence}$ ) were enumerated by gene (extended data<sup>39</sup> Section 1 – Table 5 [A–D]); genes without any strong binding sites are not listed). Similar tables were created which count the

number of information-dense clusters located within each gene (extended data<sup>39</sup> Section 1 - Table 5 [E–H]). In general, there were more hnRNP A1 clusters identified than SRSF1 and RNPS1 clusters (SRSF1 “Replicate 1” Model: 112,955; SRSF1 “Replicate 2” Model: 98,872; RNPS1: 39,285; hnRNP A1: 709,226), which is likely due to the higher frequency of strong hnRNP A1 sites and significantly lower hnRNP A1 intersite distance. Table 5 (from extended data<sup>39</sup> Section 1) also provides type II pneumocytes (from single-cell [sc] RNAseq data) and the A549 (human alveolar adenocarcinoma) cell line (RNAseq) expression values for each gene listed (in Transcripts Per Million [TPM]). Genes that are both highly expressed in lung cells and contain a high frequency of SRSF1 and/or RNPS1 information-dense binding site clusters would be considered strong candidate genes for R-loop formation in cells infected by an RNA virus. The gene *PTPRN2* has the highest total number of SRSF1 clusters (N=116 to 138 depending on the SRSF1 model used) but has relatively low level expression in pneumocytes (TPM = 0.052). The *THSD4* gene, however, has 35-36 high-density SRSF1 clusters (N=2,236-3,475 individual strong SRSF1 binding sites) and is expressed ( $\geq 1$  TPM) in both lung cell expression data sets tested (Figure 4A; extended data<sup>39</sup> Section 1 - Table 5 [E and F]). Overall, there are 1,225 genes with  $\geq 10$  SRSF1 and 127 genes with  $\geq 10$  RNPS1 information-dense clusters which are also expressed (TPM  $\geq 1$ ) in the expression datasets tested.

DRIP (DNA-RNA immunoprecipitation) sequencing is a high-throughput method of identifying regions of the genome where R-loops can form. DRIPc sequencing is an improvement which

provides higher resolution mapping data in a strand-specific manner<sup>46</sup>. To determine to what degree these DRIP-seq (GSE68845 [IMR90 cells]) and DRIPc-seq intervals (GSE70189 [NTERA2 cells]) overlapped RNPS1 and SRSF1 binding sites in uninfected cells, we performed an intersection between the two datasets and information dense clusters (extended data<sup>39</sup> Section 1 – Table 6 [A and B]) or individual binding sites (extended data<sup>39</sup> Section 1 – Table 6 [C and D]). It was uncommon for strong binding site clusters to overlap a DRIP-seq interval (0.4 – 1.7% of all transcriptome-wide clusters overlap a DRIP-seq interval). Despite an additional level of filtering (where the strand of the clusters and DRIPc-seq intervals must match), the frequency of overlap between binding site clusters and DRIPc-seq was much higher compared to the frequency of overlap to the DRIP-seq dataset (~15-17% overlap depending on IWM; extended data<sup>39</sup> Section 1 – Table 6A). In all test cases, limiting analysis to only those genes that are expressed in A549 cells ( $\geq 1$  TPM) increased the percent overlap of clusters and both DRIP- and DRIPc-seq data sets (e.g. we find a 15.3% of RNPS1 clusters/DRIPc-seq overlap among all genes, but 20.2% overlap when considering expressed genes in the A549 cell line only). When this analysis was repeated but limited to only those clusters near an exon (within 300nt), this also showed a significant increase in the fraction of clusters overlapping DRIP-seq intervals (extended data<sup>39</sup> Section 1 – Table 6B). These observations remain consistent when considering individual binding sites, rather than binding site clusters (extended data<sup>39</sup> Section 1 – Table 6C and 6D). It therefore seems that the vast majority of individual binding sites and information-dense binding site clusters do not overlap these DRIP- and DRIPc-seq regions. For example, only 5 of 36 clusters



**Figure 4. SRSF1 information dense clusters in the *THSD4* gene. A** Lollipop plot of information density of clusters annotated by coordinate range and number of sites comprising that cluster using the SRSF1 “Replicate 2” information-based weight models (all  $R_i \geq R_{sequence}$ ) for the NM\_024817 mRNA splice form of *THSD4* (some clusters counted in Section 1 – Table 5 (extended data<sup>39</sup>) are found in other *THSD4* splice forms which span beyond the range of this particular mRNA). **B** Information dense SRSF1 clusters within *THSD4* that overlap a DRIPc-seq interval (GSE70189 DRIPc-seq dataset). One additional overlapping cluster is not displayed, as is located immediately upstream of the 5' untranslated region of the NM\_024817.2 splice form. No intervals from the GSE68845 DRIP-seq dataset overlap this gene.

within *THSD4* overlap the DRIPc-seq dataset (Figure 4B; extended data<sup>39</sup> Section 1 – Table 5F).

Interestingly, the computed intersite distances for RNPS1, SRSF1 and hnRNP A1 binding sites that overlap DRIPc-seq intervals were shorter compared to the intersite distances of sites across the entire transcriptome (mean intersite distances: hnRNP A1: 22±45nt; RNPS1: 120±228nt; SRSF1 [“Replicate 1” model]: 76±205nt; SRSF1 [“Replicate 2” model]: 69±170nt; maximum intersite distance of 1,000nt). The general distributions of intersite distances between these two analyses were also found to be quite similar (extended data<sup>39</sup> Section 5). As we are limiting this analysis to sites that are within a few, often short DRIPc-seq intervals, the distances between pairs of sites are likely to be tightly grouped. We also computed the average number of all binding sites and clusters, and only those which overlap the DRIPc-seq dataset, for each individual gene (sites and clusters per 100nt of gene length; extended data<sup>39</sup> Section 1 - Table 5). Binding site densities within specific genes are reduced for sites overlapping DRIPc-seq intervals (e.g. *THSD4* SRSF1 cluster density reduces from 5.2E-03 to 7.0E-04 clusters per 100nt).

#### DNA damage response by RNA viral infection

We have previously described a machine learning (ML) based approach for developing gene signatures for expression various environmental exposures to cells, initially focusing on prediction of chemotherapy effects<sup>47</sup>. This method was applied to ionizing radiation data, from which accurate gene signatures were derived that could differentiate levels of radiation exposures. In particular, low exposures were distinguished from higher radiation levels that cause Acute Radiation Syndrome (ARS<sup>48</sup>). ARS is characterized by vomiting, diarrhea, fever, low white blood cell count and fatigue. Physicians might not consider ARS in the differential diagnosis when presented with a patient exhibiting these symptoms, since Influenza and Dengue (viral) infections also present with vomiting, diarrhea, lymphopenia (especially Influenza H1N1<sup>49</sup>) and fatigue, and are more common. Like ARS, these conditions lead to death in some cases. While Influenza A has a worldwide distribution, Dengue is more prevalent in Southeast Asia, the Americas and the Western Pacific where it presents typically with severe manifestations including hemorrhagic fever and shock. We have considered how the life cycle of these viruses might be related to the corresponding cellular responses.

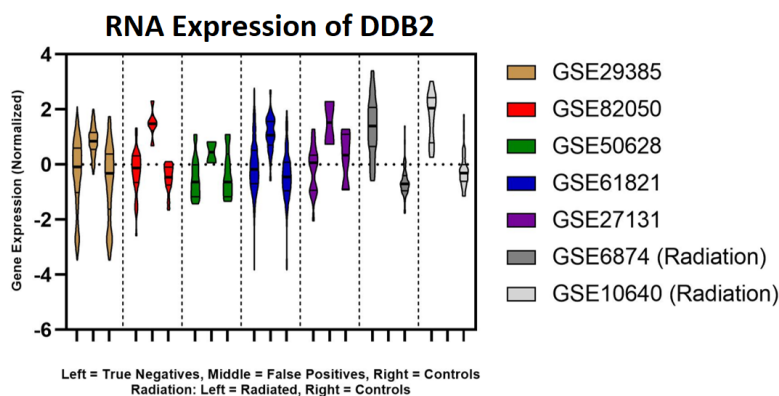
Expression data from irradiated blood samples were used to derive the human radiation gene signatures reported in Zhao *et al.*<sup>48</sup>. While it was assumed that these ML models were specific for diagnosing ARS, the models were further tested to determine if they could distinguish ARS from other conditions that share similar clinical presentation (e.g. vomiting, diarrhea). Four human ML radiation signatures from Zhao *et al.* (assessed by traditional validation; denoted as ML models “M1”, “M2”, “M3” and M4” which are described in extended data<sup>39</sup> Section 1 – Table 7) were used to evaluate 11 gene expression studies of patients infected with: Influenza (N=5, includes Influenza A [H3N2], swine flu [H1N1] and Influenza B viral infection

data sets), Dengue virus (N=4) and aplastic anemia (N=2). On average, the ML models misclassified 26.4% of Influenza and 22.4% of Dengue patients as irradiated (Section 1 – Table 7). Approximately 15% of aplastic anemia patients were also misclassified. The model “M1” showed the lowest misclassification rate against Influenza patients (9–29% of patients misclassified), models “M2” best classified Dengue-infected patients (7–33% misclassified), while models “M1” and “M3” performed well with patients with aplastic anemia (5–20% misclassified for “M1” and 0–14% misclassified for “M3”). In nearly every instance, the inclusion of normal controls from the Influenza and Dengue studies improved overall accuracy of all four ML models (17.4% and 18.1% average misclassification of Influenza and Dengue-infected patients, respectively). This phenomenon was not observed in the aplastic anemia dataset tested. The observation that normal controls are more often correctly classified indicates that these models are not so much incorrectly classifying infected patients, as they are identifying gene expression differences that may be a response to or caused by the viral infection itself.

The four radiation gene signatures assessed from Zhao *et al.*<sup>48</sup> consist of 32 unique genes. When performing feature removal analysis (where model accuracy is reassessed after each gene is individually removed from it), 10 genes were identified that greatly contribute to patient misclassification: *DDB2*, *PCNA*, *GTF3A*, *PRKCH*, *CDKN1A*, *GADD45A*, *BCL2*, *MOAP1*, *TRIM22* and *TALDO1* (extended data<sup>39</sup> Section 1 – Table 8). *DDB2* is a DNA damage binding protein that is present in all four ML models. *DDB2* expression levels were elevated in irradiated patients, which is likely due a cellular response to radiation exposure, as this gene participates in nucleotide excision repair (it ubiquitinates histones H3 and H4 to increase accessibility of nucleosomes, exposing DNA and enabling access to XPC [xeroderma pigmentosum group C-complementing protein], which performs NER<sup>50,51</sup>). *DDB2* shared a similar pattern of expression between irradiated samples as well as infected patients that were misdiagnosed as irradiated (elevated *DDB2* expression in misclassified Influenza and Dengue patients; Figure 5). The activation of *DDB2* would be consistent with the proposed mechanism, whereby high levels of RNA viral genome increase the formation of abnormal, unresolved R-loops which in turn activate a DNA damage response. Expression of *DDB2* between those correctly classified and those misclassified as irradiated was deemed significant by the Mann-Whitney test (p-value = 0.0001). Other genes with significant differences in expression included *GTF3A*, *PRKCH* and *PCNA* (which also has a role in the DNA damage response; extended data<sup>39</sup> Section 1 – Table 8).

#### Biochemical kinetics of depleted RNA binding proteins in the human transcriptome

In the mechanism proposed (Figure 1), the fraction of SRSF1 and RNPS1 bound to host RNA decreases as the fraction of SARS-CoV-2 genome increases as it replicates in the cell, causing RNA:DNA hybrids which result in R-loops. We therefore estimate the quantity of viral genomes and extent of viral replication required for viral binding site counts to approach, match, and exceed the number of host RNA sites available. These are



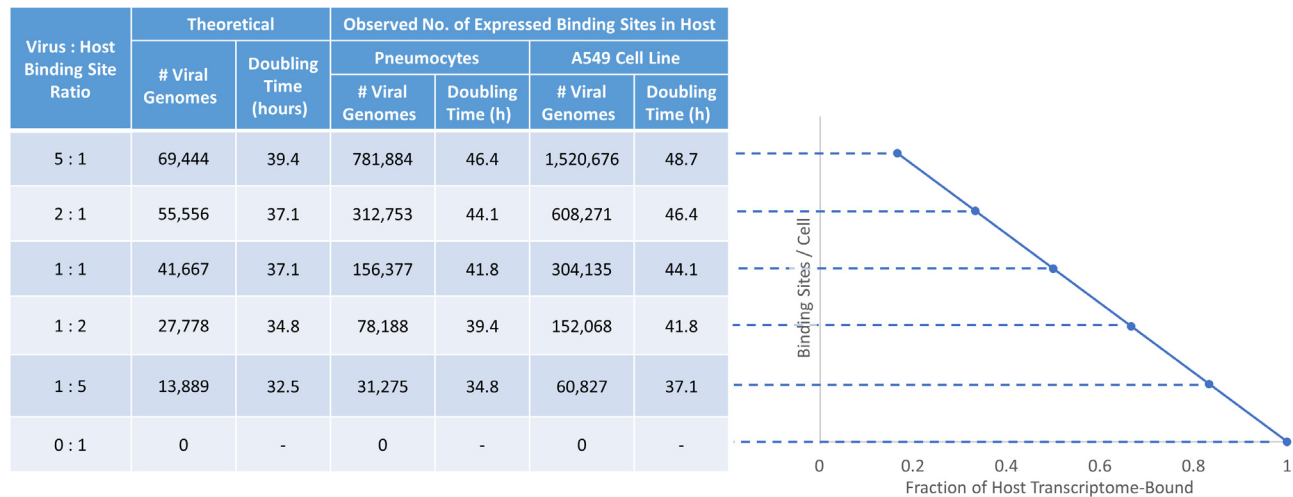
**Figure 5. Violin Plots of *DDB2* Expression in Influenza- and Radiation-Exposed Patients.** *DDB2* expression for Model 1 including Influenza patients, controls and radiation patients plotted using GraphPad. Each colour represents a different dataset. The left distribution of the radiation data (shaded grey) represents the expression of the radiated patients and the right distribution represents unirradiated controls. For all Influenza datasets (coloured), the left-most distribution represents the true negatives, the middle distribution represents the false positives, and the right-most distribution represents uninfected controls.

derived from the number of SRSF1 and RNPS1 sites expressed in either a single A549 cell or a type II primary pneumocyte (cells were not infected; note that infection would be expected to alter the expression profile, which could affect expressed binding site estimates). The overall expression of each host gene was normalized by dividing by total expression of the given dataset, then by multiplying the number of all binding sites within a gene to its normalized gene expression value, and finally by multiplying the sum of all expression-adjusted binding site counts by the expected number of mature RNAs in a cell. We estimate a total of 80,000 RNAs per single cell (as determined by Marinov *et al.*<sup>52</sup>), which is comparable with totals determined in other studies (e.g. Xia *et al.*<sup>53</sup> determined that a single osteosarcoma cell contains  $92,000 \pm 32,000$  mature RNAs).

Based on this approach, the total number of expressed binding sites (of any strength) was computed for SRSF1 and RNPS1 (Table 1). However, this estimate includes sites expected to be weakly binding. When taking only strong binding sites into account, we estimate 12.7 to 18.2 million expressed SRSF1 (“Replicate 1” and “Replicate 2” SRSF1 models, respectively) and 9.9 million expressed RNPS1 binding sites in a single A549 cell. In a single primary pneumocyte, we estimated 6.6 to 9.4 million expressed SRSF1 sites (“Replicate 1” and “Replicate 2” models, respectively), as well as 5.2 million expressed RNPS1 binding sites. These estimates are based on expression levels in normal cells and may differ in infected cells. While the dissociation constant for RNPS1 is unknown, the dissociation constant of SRSF1 ( $K_d$ ) bound to the RNA sequence 5'-UCAGAGGA-3' has been experimentally measured as  $0.2 \mu\text{M}$ <sup>54</sup>. With the  $K_d$ , a Scatchard plot for SRSF1 binding was derived where host binding sites are substrates and viral binding sites are considered to be inhibitors of host RNA binding. We assumed no free RNA binding protein (that the vast majority of SRSF1 is bound to either host or viral binding sites) as the concentration of free RBPs is likely to be low due to sequestration of RBPs by the excess of

viral sequences present in infected cells (~60% of all RNA<sup>55</sup>). This assumption is reasonable for strong binding sites (where  $R_i \geq R_{\text{sequence}}$ ). We use  $K_d$  to compute the theoretical number of viral genomes required to satisfy various viral genome to host binding site ratios (Figure 6 [Table left]). This calculation is also carried out without reference to  $K_d$ , by instead computing the number of viral genomes required to achieve binding site ratios in viral to host-bound RBP from a direct analysis of primary pneumocyte and A549 transcriptomes. The number of strong SRSF1 binding sites in a single viral genome multiplied by the level of viral replication is compared with the estimated number of expressed SRSF1 sites in the host nucleus (in a pneumocyte or an A549 cell; Figure 6 [Table right]). The data presented in Figure 6 uses the number of sites predicted by SRSF1 “Replicate 2” model, and only considers the positive strand of SARS-CoV-2. Despite their similarities, the SRSF1 “Replicate 2” model predicts far more binding sites on the positive strand of SARS-CoV-2 compared to the “Replicate 1” model (N=60 and 31, respectively). This leads to small differences in the estimated doubling time, when only the positive strand of the virus (extended data<sup>39</sup> Section 1 – Table 9A) is considered. An examination of potential binding sites on both strands of SARS-CoV-2 does not appreciably alter the estimated doubling time for both SRSF1 IWMs (extended data<sup>39</sup> Section 1 – Table 9B).

The doubling times required for infection initiated by a single virion were computed for varying numbers of viral genomes, as replication increases the overall counts of viral RBP binding sites. The processivity rate of genome replication for SARS-CoV-2 is currently unknown, so a value was estimated based on a polymerization rate of 3.7 nt/s for a different RNA-dependent viral RNA polymerase, that of Vesicular Stomatitis Virus (VSV)<sup>56</sup>. The doubling time was then adjusted to 2.31 hours per replication event, based on the increased length of the SARS-CoV-2 genome (L=30,899nt) compared



**Figure 6. Inhibition of Host SRSF1 Binding by Viral Genome Replication.** As the fraction of SARS-CoV-2 genomes increase in the host cell, the fraction of SRSF1 bound to the host transcriptome versus the viral genome decreases, resulting in R-loops. Strong SRSF1 binding sites ("Replicate 2" model) were identified in both SARS-CoV-2 (N=60 on the positive strand) and in the human transcriptome. A Scatchard plot (right) was created and used to determine the theoretical number of viral SRSF1 binding sites expected at different viral genome (inhibitor) to host (substrate) ratios (left).

to VSV. The doubling time is estimated to be between 37.1 to 44.1 hours to achieve a level of SARS-CoV-2 binding that depletes RBP from an equal number of expressed host nuclear RNA sites (1:1 ratio). However, fewer replication events and shorter doubling times are computed using the published  $K_d$  of SRSF1 (between 5-9 hours less). The number of replication events required for viral genome binding sites to overtake host RNA binding was less in primary pneumocytes compared to A549 cells (~2.3 hours or 1 doubling of the SARS-CoV-2 genome). This was anticipated, since the total number of expressed SRSF1 (and RNPS1) sites are lower in primary pneumocytes than the immortalized cell line due to lower overall gene expression levels.

## Discussion

We propose a previously undescribed putative mechanism of RNA viral infection-induced apoptosis, supported RNA binding events determined by information theoretic analysis. In the mechanism, viral release is enhanced by viral genome replication, which sequesters RBPs, thereby depleting native binding of RBPs to and stabilization of host-encoded transcripts. This process can occur in either the cytoplasm or the nucleus of the host cell, depending on specific replication requirements of different viral families. In SARS-CoV-2, this is expected to substantially reduce import of RBPs into the nucleus. Reduced availability of nuclear RBPs promotes R-loops through formation of complementary duplexes between nascent transcripts and chromosomal sequences. High densities of R-loops at a late stage of infection would be expected to overwhelm cellular DNA repair mechanisms that ordinarily remove these structures and eliminate DNA breakage. DNA damage markers *DDB2* and *PCNA* are increased in both Influenza and Dengue infections, respectively. Unrepaired, persistent chromosome

double strand breaks are unstable and induce apoptosis, which would be expected to release high viral titers.

We utilized a well-established information theory-based approach to demonstrate the validity of this proposed mechanism<sup>17-24</sup>. IT-based models of RBP binding sites was used to scan viral RNA genomes (Influenza, SARS-CoV-2 and Dengue) and host transcriptomes. IT models derived from thousands of validated RBP binding sites delineated numerous strong SRSF1, RNPS1 (and hnRNP A1) RNA binding sites within these viral genomes. The derived SRSF1 and RNPS1 binding motifs were shown to be highly similar, consistent with previous published studies demonstrating that RNPS1 could partially complement genomic instability due to SRSF1 deficiency. Indeed, both models detected many of the same RNA binding sites in the host transcriptome and all strong RNPS1 binding sites detected in the SARS-CoV-2 genome were simultaneously detected by at least one SRSF1 information model. In divergent strains of both SARS-CoV-2 and Influenza A (H3N2), the frequencies and strengths of these binding sites are highly consistent. Finally, we estimate that the quantity of replicated viral genomes necessary to meet or exceed the number of binding sites expressed within a lung can exceed the site counts in the host genome, and the doubling time required to deplete these RBPs which is consistent with the observed time course of severe infections.

The estimated doubling times were based on the assumption that the RNA polymerization rate of SARS-CoV-2 was similar of that of VSV. However, the replication rate of RNA dependent RNA polymerase of the original SARS-CoV virus is considerably faster (600-700 nt/s<sup>57</sup>). The similarity between these coronaviral sequences implies that the SARS-CoV-2

genome might replicate in under a 1 minute. If the viral replication rate in our study from VSV is an underestimate, the corrected processivity of this enzyme would be expected to accelerate sequestration of RBPs, as well as the proposed R-loop formation and apoptosis. However, the polymerization rate measured *in vitro* may not be sustainable due to reaction *in vivo* constraints (e.g. nucleotide pool depletion and subcellular compartmentalization).

Functional analyses will be needed to prove that this mechanism plays a role in viral pathogenicity. Such studies should further investigate how infections of SARS-CoV-2 (and other RNA viruses) cause increased DNA damage. RNAseq and protein expression analysis of *DDB2*, *RAD17*, *PRKDC*, *PCNA* and other ATR pathway markers of infected cells accompanied by time course studies of nascent double stranded chromosomal breaks (i.e. H2AX antibody staining due to viral infection) would provide such evidence. Increased R-loop formation upon infection will be required, with particular attention to host encoded transcripts enriched in SRSF1 and RNPS1 binding site clusters. Although the genomic coordinates where R-loops form can be anticipated from information dense clustering, the strand- and gene specific techniques used to detect these, i.e. DRIPc-Seq, cannot measure RNA-DNA hybrids of lengths shorter than 70bp<sup>66</sup>. Sequence-based chromatin immunoprecipitation with antibodies to H2AX, 53BP1 or other markers of DNA damage should be consistent with the sites of R-loop formation. Changes in the expression of apoptotic markers (e.g. *BCL2*, *BCL2L2*, *BAX*, and *TNFRSF10B*) would also be expected in infected cells with high levels of replication. Direct interaction between RBPs and viral genomes must also be demonstrated, possibly by immunoprecipitation or copackaging in viral capsids. It should also be possible to evaluate the possibility that inhibitors of viral replication, such as remdesivir (and any other nucleoside analogs), can reduce DNA damage, R-loop formation, and apoptosis of infected cells.

SARS-CoV-2 efficiently infects multiple species of mammals<sup>58</sup>, and possesses an RNA polymerase with proofreading capability, which enables it to faithfully and accurately replicate and transcribe its genome. In this study, we suggest that effects of SARS-CoV-2 infection are mild in most individuals because most of us mount robust immune responses and eventually clear the virus. The mechanism that we propose (Figure 1), which may be a contributing factor of a variety of different RNA viruses, has the potential to overwhelm that response through jackpot replication coupled to apoptotic events caused by loss of chromosome integrity stemming from depletion of essential RBPs. This results in high multiplicities of infection of cells in the most vulnerable cells. This could cause a rapid onset of loss of viable pneumocytes, and compromising oxygen transport, to a point where it is insufficient to maintain blood pO<sub>2</sub> levels to support organ functions. Systemic inhibition of viral replication and transcription of viral proteins will be essential to prevent or mitigate this pathological mechanism.

Other coronaviruses such as MERS and SARS have been shown to induce apoptosis<sup>59</sup>. The polyphenol Resveratrol has

been shown to downregulate apoptosis *in vitro*<sup>59,60</sup>, possibly by overexpressing sirtuins (a family of signalling proteins). However, this is ultimately not a practical solution to infection, as the drug will only delay an eventual high multiplicity infection event. In order to inhibit the viral mechanism proposed in this study, a drug must inhibit the viral machinery that sequesters spliceosomal components, leading to R-loops and DNA damage. This may explain, in part, why remdesivir (Gilead) improves the recovery of SARS-CoV-2 patients. The drug, which was originally developed for treatment of Ebola virus by inhibiting its RNA dependent RNA polymerase, also inhibits viral replication of SARS-CoV-2. Other potential therapies include those targeting expression of genes encoded by the viral genome, which use a common 5' leader sequence of all transcripts. The promoter sequence for these genes binds to the host encoded hnRNP A1, which regulates transcription of beta coronaviral genes (of which SARS is a member of that family). While hnRNP A1 could be a potential drug target for therapy (there are small molecules that have been shown to inhibit hnRNP A1 RNA splicing activity<sup>61</sup>), there would be concerns that this may cause inadvertent side effects due to its impact on normal mRNA splicing.

Regardless of whether apoptosis releases large quantities of mature infectious virus, the proposed mechanism will still likely impact pneumocyte function. Should high multiplicities of infection be the result of apoptotic release of virions, then the proposed RBP depletion mechanism would be expected to kill both the original infected cell and neighboring infected pneumocytes. The severe symptoms might be the result of rapid, overwhelming lysis of cells responsible for oxygen transport, rather than by a cytokine storm. Autopsies of infected individuals from Wuhan China have shown evidence of inflammation, but not necessarily macrophage invasion and pulmonary edema<sup>62</sup>. Furthermore, apoptosis has been demonstrated in lung epithelial cells in Macaques infected with Influenza virus<sup>63</sup>. This could explain why physicians and other health professionals in repeated contact with multiple infected patients do not seem to have time to develop immunity to the virus, regardless of their age. Type II pneumocytes which produce surfactant, required at high levels in newborns, decrease with age<sup>64</sup> and are particularly diminished in individuals with respiratory disease like COPD (Chronic obstructive pulmonary disease) and ARDS (Acute respiratory distress syndrome). If the multiplicity of infection (MOI) of virus damages this population of cells, then individuals with fewer cells might be more susceptible to exhibiting insufficient pulmonary function due to the high MOI released by the mechanism proposed. These patients would be at greater risk for severe complications requiring assisted ventilation. It is also possible that the deficiency of functional pneumocytes in such individuals cannot be compensated for by extracorporeal membrane oxygenation to rescue multiple organ failure.

Humans have high numbers of type II pneumocyte cells at birth to fulfill demands for surfactant to rapidly expand lung volume. Synthetic surfactant is an essential treatment for premature birth, since these cells mature late in gestation. Age-related loss of these cells has been measured and the mechanism leading

to it was described<sup>64</sup>. Loss of functional pneumocytes is particularly evident in individuals with ARDS, who exhibit significant lung fibrosis, which is also seen in patients with SARS-CoV-2 infections. Older individuals (or those with pre-existing respiratory conditions) are more susceptible to the loss of the remaining cells by apoptosis or autophagia. Decreased pneumocyte counts affect O<sub>2</sub> transport efficiency, which lowers blood pO<sub>2</sub>, and extant tissues and organs. The proposed mechanism implies that jackpot viral replication events, regardless of age of the infected individual, enhances viral release through apoptosis and infection. Such events are more likely in cells infected by coronaviruses like SARS-CoV-2, which are capable of repressing the innate immune response, i.e. induction of interferon response to viral double stranded RNA (unlike Influenza)<sup>55,65,66</sup>. Repression of innate immunity enables the virus to replicate unabated in these cells, which would be expected to delay their recognition by regulatory T cells and killing by macrophages.

Viral infections significantly alter the transcriptional profiles of host genes in infected cells. Recent studies of Zika virus (an RNA virus) have revealed that infection not only impacts transcription, but affects alternative mRNA splicing as well<sup>67</sup>. Both RNA and DNA viral infections encode factors that directly<sup>68</sup> or indirectly<sup>67</sup> alter host RNA processing, resembling alternative mRNA isoforms. We suggest that the mRNA splicing changes observed subsequent to infection of an RNA virus could be a consequence of replicated viral genome binding to RBPs, thus changing the nuclear stoichiometry of splicing proteins (such as SRSF1). This would effectively reduce the concentration of available splicing factors, which could be responsible for the observed alternative splicing events of other splicing factors (such as SRSF2 and SRSF3) reported by Bonenfant *et al.*<sup>67</sup>. Thus, the mechanism proposed in this study may not only impact genome stability by the introduction of R-loops, but may simultaneously alter the global alternative splicing landscape in infected host cells.

RNA-based vaccines based upon synthetic SARS-CoV-2 transcripts containing modified nucleosides that have been dephosphorylated to escape innate immunity are being tested<sup>69</sup>. These candidates exploit host protein synthesis machinery to transiently express viral antigens that activate B and T-cell immunity. However, these synthetic RNAs would also be available for RBP binding. A transcript encoding the SARS-CoV-2 spike glycoprotein 'S' gene, for example, would contain 7 strong RNPS1 and between 6 to 8 strong SRSF1 binding sites (depending on SRSF1 model). If the levels of expression produced from these transcription templates cannot be carefully controlled, excess production of these RNAs could potentially elicit undesirable side effects through sequestration of critical host RNA binding proteins required to inhibit R-loop formation.

Localization of viral replication to the cytoplasm does not obviate the fact that there is still a competition between the host and viral genomes for these RNA binding proteins. While the binding site stoichiometry calculations are unchanged, compartmentalization of the viral and host genomes does have

implications for preventing R-loops during host transcription. Since coronavirus replicates in the cytoplasm, binding of newly synthesized RBPs occurs there. This makes less protein available to be imported into the nucleus for binding to nascent transcripts to prevent R-loops from forming. The viral genome may have an advantage in this competition for binding to RBPs relative to nuclear transcripts, due to the proximity of the viral genome to nascent RBPs in the cytoplasm, which may limit transport and impede their import into the nucleus. RBPs are often highly expressed, including SRSF1 and RNPS1, and are abundant in the lung (where SARS-CoV-2 infection is most prominent). Thus, the cytoplasmic concentration of viral genome necessary to prevent the localization of RBPs into the nucleus is likely to vary between different tissues.

The proposed mechanism of RNA virally-induced apoptosis is supported by extensive bioinformatic analyses indicating that strong RNA binding sites of host RBPs are common in RNA viral genomes, and that the frequencies of such binding sites are relatively consistent between divergent strains in both Influenza A and SARS-CoV-2. Future efforts should elucidate details of the mechanism with functional analysis of infected cells, including demonstration of increased R-loop formation, induction of relevant apoptotic or DNA repair responses, and direct interaction between viral genomes and host RBPs. This would justify further investigations into binding of specific RBPs to viral sequences in infected patients. The potentially prognostic significance of such data could be useful in differentiating among drug therapies that target RNA viral genome replication and/or expression.

## Methods

### Information theory-based RNA binding site analysis

The IWMs for the RBPs investigated in this study (SRSF1, RNPS1 and hnRNP A1) were either obtained for previously published analyses or derived in this study. The hnRNP A1 IWM used in this study was previously derived in Peterlongo *et al.* (using PoWeMaGen software [v1]<sup>22</sup>) using an hnRNP A1 CLIP-seq dataset<sup>70</sup>. The functionality provided by PoWeMaGen is also available in *Delila* software, which is open source. IWMs can also be derived with the 'Ri' program, and RBP binding sites can be localized with the 'Scan' program of the *Delila* package. Individual binding site strengths ( $R_i$  values) of these IWMs can also be determined using the 'Scan' program.

A previously described IWM for SRSF1<sup>21</sup> was based on only 28 manually curated and validated and aligned binding sites<sup>71</sup>. To update this IWM, we derived new SRSF1 models from high-throughput eCLIP datasets containing thousands of validated binding sites of 150 different RBPs<sup>34</sup>. Narrow peak files from two separate SRSF1 eCLIP replicates (*ENCFF179SCM* and *ENCFF184TBM*), as well as two non-target, negative control replicates (*ENCFF241ORF* and *ENCFF773PUP*) were retrieved from the ENCODE Data Coordination Center (ID: *ENCSCR456FVU*)<sup>34</sup>. The new SRSF1 and negative control SRSF1 IWMs were generated using *Maskminent* v1.0.2 (24; <https://doi.org/10.5281/zenodo.49234>). Both PoWeMaGen and *Maskminent* utilize the Bipad algorithm to align binding sites<sup>72</sup>. Similarly, RNPS1 and



GFP-control IWMs were derived from publicly available iCLIP data (31; E-MTAB-4215). However, this iCLIP dataset was only available in FASTQ file format, which required further processing to identify CLIPseq peaks. Thus, the available RNPS1 iCLIP data was first aligned to the human genome (GRCh37) with TopHat v2.1.1, and then converted to peaks using Piranha v.1.2.1 (a CLIP- and RIP-seq peak caller) under default settings.

IWMs for SRSF1 and RNPS1 were derived from eCLIP and iCLIP-seq datasets (respectively) using Maskminent under varying model length conditions (6-10nt long; 1,000 Monte Carlo cycles). As experimental noise has been found to contribute to non-specific IWMs<sup>24</sup>, we limited model derivation to only the to the 5,000 or 50,000 iCLIP peaks with the highest signal value (SRSF1) or the lowest p-values (RNPS1; computed by Piranha). In practice, the derived models remained similar regardless of the size of peak subset used. As many intervals from the SRSF1 and RNPS1 datasets were short (<20nt), peak lengths were extended on either direction by the sequence length (e.g. a 10nt interval becomes 30nt long). We found that both RNPS1 and SRSF1 models derived at lengths of 6nt to be most informative with similar  $R_i$  densities, although they differed slightly (Table 1). Both the RNPS1 model and the SRSF1 model derived from the second replicate (SRSF1 “Replicate 2”) selected was generated from 5,000 CLIP-seq peaks, while the SRSF1 “Replicate 1” model was derived from 50,000 peaks.

To evaluate the similarity between these IWMs, the RNPS1 and SRSF1 motifs were compared using the STAMP web server<sup>35</sup>, which performs a pairwise alignment between each motif (ungapped Smith-Waterman alignment method) and compared using a Pearson correlation coefficient distance metric, and outputs results as e-values. Statistical significant differences between the e-values of IWMs for RBPs were compared with their corresponding negative control motifs. These were quantified as log<sub>10</sub> likelihood ratios determined from the pairwise RBP motif comparison relative to the same RBP with its negative control motif e-values according to:

$$\text{LOD score} = \log_{10} \left( \frac{e - \text{value} (\text{RBP vs. RBP IWM})}{e - \text{value} (\text{RBP vs. neg. control IWM})} \right)$$

### Transcriptome, exome and viral genome RBP scans

The human reference genome (GRCh37; Genbank Acc. GCA\_000001405.1; downloaded from UCSC [<https://hgdownload.soe.ucsc.edu/goldenPath/hg19/bigZips/>]) and viral genomes (Dengue virus 3 [GenBank accession: NC\_001475.2]; human immunodeficiency virus type 1 [HIV-1] HXB2 [Genbank: K03455.1] and subtype C [Genbank: U46016.1]; Influenza A H3N2 strains [Ontario/104-25/2012; Genbank (segments 1-8): KJ413878.1, KJ413896.1, KF840477.1, KJ413897.1, KJ413880.1, KJ413864.1, KJ413915.1, KJ413925.1] and [Shanghai/C84/2009 Genbank (segments 1-8): JX286598.1, JX286597.1, JX286596.1, JX308801.1, JX286594.1, JX286593.1, JX286592.1, JX286595.1]; and SARS-CoV-2 [Genbank: NC\_045512.2]) were scanned with each IWM (SRSF1 [two

separate models], RNPS1 and hnRNP A1). Human genome scans were then filtered so that only those predicted binding sites found in transcribed regions (using the Ensemble Genes database [release 99]) would be considered. Only sites exceeding  $R_{\text{sequence}}$ , the average information content of the binding site model, were retained in subsequent analyses as these consist of mean binding affinity or higher and are likely to more effectively compete for binding to these proteins<sup>24,25</sup>. The  $R_{\text{sequence}}$  values of each model are: 6.7 bits (SRSF1 “Replicate 1” model), 6.4 bits (SRSF1 “Replicate 2” model), 7.8 bits (RNPS1 model), and 4.6 bits (hnRNP A1 model).

Besides those previously indicated, viral genomes of multiple other SARS-CoV-2 and Influenza A (H3N2) strains were scanned using the IWMs for SRSF1, RNPS1 and hnRNP A1 to evaluate whether divergent strains of these viruses carry significantly different strong binding sites counts. NextStrain (which provides real-time tracking of the SARS-CoV-2 and Influenza A) was utilized to choose divergent strains of either virus by selecting strains from separate clades, i.e. different monophyletic groups. The viral genome sequences of selected SARS-CoV-2 strains (Genbank accessions: MT007544.1 [Australia], MT066176.1 [Taiwan], MT121215.1 [China], MT163718.1 [USA], MT188339.1 [USA], MT198652.1 [Spain], and MT198653.1 [Spain]) and Influenza A H3N2 (GISAID accessions: EPI1676017-EPI1676024 [Denmark], EPI1635542-EPI1635549 [England], EPI1594883-EPI1594890 [Singapore], EPI1614613-EPI1614620 [Sydney]) were downloaded from the GISAID database. Each of these genome sequences were evaluated for strong SRSF1, RNPS1 and hnRNP A1 binding sites. All binding sites (with  $R_i \geq R_{\text{sequence}}$ ) are provided in extended data<sup>39</sup>, Section 1 – Tables 3 and 4.

### Expressed RNA binding sites in lung cells

Publicly-available expression datasets were downloaded from the Gene Expression Omnibus for A549 cell lines (GSE141171; RNAseq) and primary type II pneumocytes (GSE86618; scRNAseq). Normal expression for each cell type was computed by taking the average of all control samples from each dataset (N=3 control samples in GSE141171; N=215 control samples in GSE86618). We then use this information to estimate the total number of binding sites present in a single pneumocyte or A549 cell. First, the program “ScanDataSummaryProgram.pl” (available within underlying data<sup>39</sup> Section 6) was used to compute the total number of binding sites ( $\geq R_{\text{sequence}}$ ) in each cell type for each expressed gene (TPM >0; underlying data<sup>39</sup> Section 1 - Table 5). The overall expression of each gene was then normalized using the program “TotalBindingSitePerCellCalculator.pl” (underlying data<sup>39</sup> Section 6), which divides expression by the sum of all TPM values in the cell, multiplied by the estimated number of mature RNAs in a cell at any given timepoint (80,000 RNAs per lymphoblastoid cell<sup>52</sup>). It then multiplies this normalized gene expression value with its binding site total to determine the overall contribution of binding sites from that gene in a single cell. The sum of this value across all expressed genes gives the total number of RNA binding sites expected to be available in a cell at any given time (Figure 6).

## Information-dense clustering of RBPs across viral genomes and human transcriptome

Information dense clustering has previously been applied to the human genome to identify clusters of organized TFBSs<sup>25,40</sup>. The clustering software (v1; described in reference 25; software provided in a Zenodo archive - <https://doi.org/10.5281/zenodo.1707423>) was used in this study to identify clusters of low-affinity ( $R_i > 0$  bits), moderate-affinity ( $\geq \frac{1}{2}R_{sequence}$ ) and high-affinity ( $\geq R_{sequence}$ ) RBP sites in both the viral genomes investigated in this study, and across the entire human transcriptome. To be considered a cluster, each set of component sites was required to occur  $\leq 25$ nt from one other, and the total information of all sites within the cluster equalled or exceeded  $\geq 50$  bits. In its original design, the clustering algorithm considered binding sites on both strands in forming clusters. To maintain strand specificity, we separated input by strand. Due to the high memory demands of the clustering algorithm, transcriptome scan input was separated into segments of  $\sim 200,000$  sites per run, which was then subsequently combined. To avoid the inadvertent separation of a binding site cluster, input was split only when two sequential binding sites were  $> 1,000$ nt apart.

## Identification of RBP sites and clusters within DRIP-seq intervals

All binding sites and information-dense clusters identified in the human genome were intersected with DRIP-seq and DRIPc-seq intervals, which indicate where there is evidence of R-loop formation in the human genome (performed by “ClusterToDRIPseqAnalysisProgram.pl”; underlying data<sup>39</sup> Section 6). The DRIP-seq dataset (GSE68845; IMR90 cells) is not strand specific, thus binding sites and clusters from either strand are considered when intersected against these intervals. DRIPc-seq data (GSE70189; NTERA2 cells), however, is strand specific which has been taken into account (e.g. positive strand clusters found in positive strand DRIPc-seq intervals reported). We then computed the gene density of sites and clusters that are found within these intervals (underlying data<sup>39</sup> Section 1 - Table 5) using the script “ClusterToDRIPseqAnalysisProgram.GeneDensityFinder.pl” (underlying data<sup>39</sup> Section 6) to determine if there is a correlation between the presence of binding sites and R-loop formation.

## Lollipop plots and intersite distance histograms

Lollipop plots which indicate the location of information-dense clusters for all viral genomes described in this study and for all genes in the human transcriptome (with  $\geq 1$  cluster) were generated in R (version 3.6.3) using the Bioconductor package “trackViewer” (v.1.20.373). The lollipop plots presenting human genes contain intron and exon boundary information which was generated using the RefSeq database (release 60). Multiple lollipop plots were generated for multi-segmented viral genomes (one image per segment). The height of each “lollipop” corresponds to the information density of a cluster, and its location in the genome is indicated (GRCh37) along with the number of sites which comprise the cluster.

Histograms which illustrate the distribution of binding site  $R_i$  values and the frequency of the distance between RBPs

(“intersite distances”) were generated using the R package ‘ggplot2’ (v3.1.1<sup>72</sup>). Intersite distance frequency was determined by first grouping all RBP by gene, followed by determining the distance between each site in sequential order. Distance thresholds of 500nt or 1,000nt were assigned for all intersite distance histograms. Rare instances of distances greater than these thresholds were excluded from the histogram, as their inclusion led to plots too wide to be informative.

## Radiation gene expression signatures and viral infection

Gene expressions for individuals with the diseases above were collected from Gene Expression Omnibus (GEO), which consisted of 5 Influenza studies (GSE29385, GSE82050, GSE50628, GSE61821, GSE27131), 4 Dengue studies (GSE97861, GSE97862, GSE51808, GSE58278) and 2 studies involving Aplastic Anemia patients (GSE16334, GSE33812). We also collected expression data from two studies with radiation-exposed samples (GSE6874 and GSE10640). The best performing human signatures (assessed by traditional validation; described in Table 7 [underlying data<sup>39</sup> Section 1]) from Zhao *et al.*<sup>48</sup> were then used to test the gene expression datasets in order to determine if these models would misclassify infected patients as irradiated (with and without control patients). Models were tested using the MatLab script used to perform “traditional validation” in the Zhao *et al.* study (“regularValidation\_multiclassSVM.m”, <https://zenodo.org/record/1170572>), which first normalizes gene expression values by quantile normalization before applying the radiation model to the infected patient data to predict outcome. The script then compares prediction of radiation exposure to the clinical data provided. MatLab scripts are compatible with GNU Octave.

To better understand why the radiation models are predicting certain Influenza- and Dengue-infected patients as irradiated, violin plots were generated using GraphPad Prism v8 to visually illustrate differences in gene expression between infected individuals correctly classified and those misclassified by each radiation model (Figure 5). When inspecting violin plots of the 32 genes which make up the 4 radiation models tested, 10 genes were identified to have contributed towards false positives predictions as they shared a similar pattern of expression in those that were radiated in two gene expression datasets of irradiated individuals (GSE6874 and GSE10640). The 10 genes are: *DDB2*, *PCNA*, *GTF3A*, *PRKCH*, *CDKN1A*, *GADD45A*, *BCL2*, *MOAP1*, *TRIM22* and *TALDO1*. Mann-Whitney tests were used to compare the expression of these genes in false negative and true positive patients. Four genes (*DDB2*, *PCNA*, *GTF3A* and *PRKCH*) were consistently found significant in most of the studies tested.

## Association kinetic analysis

The dissociation constant of SRSF1 bound to the RNA sequence 5'-UCAGAGGA-3' was experimentally determined to be  $0.2 \mu\text{M}$ <sup>34</sup>. This information allowed for the derivation of a theoretical Scatchard plot for SRSF1 binding by varying the relative proportions of viral to host binding sites bound (where viral binding sites are considered inhibitors, and host binding

sites as substrate). We can compute the theoretical number of viral genomes necessary to reach these relative proportions according to:

$$\frac{v}{[L]} = \frac{n}{K_d} - \frac{v}{K_d}$$

Where  $K_d$  is the SRSF1 dissociation constant,  $n$  is the number of sites (or sequences) that a single protein can bind ( $n=1$ ),  $[L]$  is the concentration of free SRSF1, and  $v$  is the amount of SRSF1 bound to the viral genome relative to host. Upon infection and viral replication, it is assumed there is no free RNA binding protein (all RBP is assumed to be bound to either viral or host RNA). These proportions were converted to numbers of viral genomes per infected host cell (determined using the above formula in an MS- Excel spreadsheet), adjusted for the computed number of viral genomes per cell by the number of SRSF1 binding sites in a single viral genome (described earlier). We also computed the number of viral genomes necessary to reach these proportions by taking A549 or pneumocyte host cell binding site expression (computed previously) into account. We then used the published processivity rate of 3.7 nucleotides/sec for VSV RNA dependent RNA polymerase<sup>56</sup> to estimate the doubling time required.

### Statistical analysis

The average distances between adjacent binding sites of SRSF1, RNPS1 and hnRNP A1 were determined within both expressed human genes and RNA viral genomes (Dengue, HIV-1 strains B and C, Influenza A and SARS-CoV-2). A program script “calculateIntersiteDistance.pl” (underlying data<sup>39</sup> Section 6) takes a set of binding site coordinates and their associated genes as input and determines the pairwise distances between all consecutive binding sites in the same gene. Subsequently, “removeOutliersHigherThanN.pl” is used to discard extreme outlier distances exceeding a specified threshold (thresholds of 500nt and 1,000nt were evaluated). Finally, “getStatisticsOnCol.pl” evaluates a given set of intersite distances and computes the count, geometric mean, median, arithmetic mean and their standard deviation. The program was used to evaluate intersite distances at multiple  $R_i$  thresholds (low- [ $R_i > 0$  bits], moderate- [ $\geq \frac{1}{2} R_{sequence}$ ] and high-affinity [ $\geq R_{sequence}$ ] binding sites). We also examined binding sites which intersect DRIPc-seq intervals in the human genome using this procedure. Output from this analysis are provided as histograms in extended data<sup>39</sup> Section 5, as described earlier.

### Data availability

A data repository titled “Characteristics of human and viral RNA binding sites and site clusters recognized by SRSF1 and RNPS1” has been deposited as a Zenodo archive (DOI: [10.5281/zenodo.3737089](https://doi.org/10.5281/zenodo.3737089)<sup>39</sup>). The archive contains the following underlying and extended data, organized across 6 sections. Section 1 primarily consists of extended data, and Sections 2–6 contains the underlying data presented in the paper.

### Extended data

Zenodo: Characteristics of human and viral RNA binding sites and site clusters recognized by SRSF1 and RNPS1. [http://doi.org/10.5281/zenodo.3737089](https://doi.org/10.5281/zenodo.3737089)<sup>39</sup>

This project contains the following extended data:

Section 1 – The nine additional tables described in this study (“Section 1 - Tables 1–9”), which provide SRSF1, RNPS1 and hnRNP A1 binding site and information-dense cluster counts across various RNA viral genomes [including multiple SARS-CoV-2 and Influenza strains] and the human transcriptome, the estimated SARS-CoV-2 doubling time necessary for viral genome SRSF1 binding site availability to exceed sites within the host transcriptome, and an analysis of Influenza, Dengue, and aplastic anemia patients misdiagnosed as irradiated by established radiation gene signatures.

### Underlying data

Zenodo: Characteristics of human and viral RNA binding sites and site clusters recognized by SRSF1 and RNPS1. [http://doi.org/10.5281/zenodo.3737089](https://doi.org/10.5281/zenodo.3737089)<sup>39</sup>

Section 2. All SRSF1, RNPS1 and hnRNP A1 binding site genome browser tracks for human and all viral genomes analyzed in this study (GRCh37).

Section 3. The full set of lollipop plots (indicating the location of SRSF1, RNPS1 and hnRNP A1 information-dense clusters) in all human genes and in each of the viral genomes analyzed.

Section 4. The  $R_i(b,l)$  matrices or IWMs for all RBPs analyzed (SRSF1, hnRNP A1 and RNPS1).

Section 5. The full set of histograms which display the distribution of  $R_i$  strength and intersite distance between the binding sites for each RBP [across all transcribed regions or within known DRIPc-seq intervals].

Section 6. A set of 7 Perl scripts created specifically for this study, with instructions for their use: A) “ClusterToDRIPseqAnalysisProgram.pl” – reports which information-dense clusters are located within DRIPc- and/or DRIP-seq intervals (individually and by gene); B) “ClusterToDRIPseqAnalysisProgram.GeneDensityFinder.pl” – uses the output from script “A” to determine the number and the density of information-dense clusters within a gene (total clusters within the gene and those within DRIPc-seq intervals); C) “calculateIntersiteDistance.pl” – determines the distance between all binding sites in the same gene from a list of genomic coordinates; D) “removeOutliersHigherThanN.pl” – discards intersite distances computed by script “C” that are greater than a specified threshold; E) “getStatisticsOnCol.pl” – calculates the count, geometric mean, median, arithmetic mean, and standard deviation of values from script “D”; F) “ScanDataSummaryProgram.pl” – determines the number of binding sites (above a specified  $R_i$  threshold) found within known genes (the program also reports the total expression of those genes using external A549 and pneumocyte expression datasets) from binding site coordinate data; G) “TotalBindingSitePerCellCalculator.pl” – estimates the number of binding sites expressed in a single A549 or pneumocyte cell at any given time.

Data are available under the terms of the [Creative Commons Attribution 4.0 International license](https://creativecommons.org/licenses/by/4.0/) (CC-BY 4.0).

## References

1. Carrasco-Hernandez R, Jácome R, López Vidal Y, et al.: **Are RNA Viruses Candidate Agents for the Next Global Pandemic? A Review.** *ILAR J.* 2017; **58**(3): 343–58.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
2. Liu Y, Yan LM, Wan L, et al.: **Viral dynamics in mild and severe cases of COVID-19.** *Lancet Infect Dis.* 2020; **20**(6): 656–657.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
3. Zheng S, Fan J, Yu F, et al.: **Viral load dynamics and disease severity in patients infected with SARS-CoV-2 in Zhejiang province, China, January–March 2020: retrospective cohort study.** *BMJ.* 2020; **369**: m1443.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
4. Holshue ML, DeBolt C, Lindquist S, et al.: **First Case of 2019 Novel Coronavirus in the United States.** *N Engl J Med.* 2020; **382**(10): 929–36.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
5. Fujikura D, Miyazaki T: **Programmed Cell Death in the Pathogenesis of Influenza.** *Int J Mol Sci.* 2018; **19**(7): 2065.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
6. Aguilera A, García-Muse T: **R loops: from transcription byproducts to threats to genome stability.** *Mol Cell.* 2012; **46**(2): 115–24.  
[PubMed Abstract](#) | [Publisher Full Text](#)
7. Manley JL, Tacke R: **SR proteins and splicing control.** *Genes Dev.* 1996; **10**(13): 1569–79.  
[PubMed Abstract](#) | [Publisher Full Text](#)
8. Krainer AR, Conway GC, Kozak D: **The essential pre-mRNA splicing factor SF2 influences 5' splice site selection by activating proximal sites.** *Cell.* 1990; **62**(1): 35–42.  
[PubMed Abstract](#) | [Publisher Full Text](#)
9. Black DL: **Mechanisms of alternative pre-messenger RNA splicing.** *Annu Rev Biochem.* 2003; **72**: 291–336.  
[PubMed Abstract](#) | [Publisher Full Text](#)
10. Li X, Manley JL: **Inactivation of the SR protein splicing factor ASF/SF2 results in genomic instability.** *Cell.* 2005; **122**(3): 365–78.  
[PubMed Abstract](#) | [Publisher Full Text](#)
11. Li X, Niu T, Manley JL: **The RNA binding protein RNPS1 alleviates ASF/SF2 depletion-induced genomic instability.** *RNA.* 2007; **13**(12): 2108–15.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
12. Ewing RM, Chu P, Elisma F, et al.: **Large-scale mapping of human protein-protein interactions by mass spectrometry.** *Mol Syst Biol.* 2007; **3**: 89.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
13. Deka B, Singh KK: **Multifaceted Regulation of Gene Expression by the Apoptosis- and Splicing-Associated Protein Complex and Its Components.** *Int J Biol Sci.* 2017; **13**(5): 545–60.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
14. Gómez-González B, García-Rubio M, Bermejo R, et al.: **Genome-wide function of THO/TREX in active genes prevents R-loop-dependent replication obstacles.** *EMBO J.* 2011; **30**(15): 3106–19.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
15. Salas-Armenteros I, Barroso SI, Rondón AG, et al.: **Depletion of the MFAP1/SPP381 Splicing Factor Causes R-Loop-Independent Genome Instability.** *Cell Rep.* 2019; **28**(6): 1551–1563.e7.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
16. Stirling PC, Chan YA, Minaker SW, et al.: **R-loop-mediated genome instability in mRNA cleavage and polyadenylation mutants.** *Genes Dev.* 2012; **26**(2): 163–75.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
17. Caminsky N, Mucaki EJ, Rogan PK: **Interpretation of mRNA splicing mutations in genetic disease: review of the literature and guidelines for information-theoretical analysis [version 1; peer review: 2 approved].** *F1000Res.* 2014; **3**: 282.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
18. Rogan PK, Schneider TD: **Using information content and base frequencies to distinguish mutations from genetic polymorphisms in splice junction recognition sites.** *Hum Mutat.* 1995; **6**(1): 74–6.  
[PubMed Abstract](#) | [Publisher Full Text](#)
19. Rogan PK, Faux BM, Schneider TD: **Information analysis of human splice site mutations.** *Hum Mutat.* 1998; **12**(3): 153–71.  
[PubMed Abstract](#) | [Publisher Full Text](#)
20. Mucaki EJ, Ainsworth P, Rogan PK: **Comprehensive prediction of mRNA splicing effects of BRCA1 and BRCA2 variants.** *Hum Mutat.* 2011; **32**(7): 735–42.  
[PubMed Abstract](#) | [Publisher Full Text](#)
21. Mucaki EJ, Shirley BC, Rogan PK: **Prediction of mutant mRNA splice isoforms by information theory-based exon definition.** *Hum Mutat.* 2013; **34**(4): 557–65.  
[PubMed Abstract](#) | [Publisher Full Text](#)
22. Peterlongo P, Catucci I, Colombo M, et al.: **FANCM c.5791C>T nonsense mutation (rs144567652) induces exon skipping, affects DNA repair activity and is a familial breast cancer risk factor.** *Hum Mol Genet.* 2015; **24**(18): 5345–55.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
23. Mucaki EJ, Caminsky NG, Perri AM, et al.: **A unified analytic framework for prioritization of non-coding variants of uncertain significance in heritable breast and ovarian cancer.** *BMC Med Genomics.* 2016; **9**: 19.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
24. Lu R, Mucaki EJ, Rogan PK: **Discovery and validation of information theory-based transcription factor and cofactor binding site motifs.** *Nucleic Acids Res.* 2017; **45**(5): e27.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
25. Lu R, Rogan PK: **Transcription factor binding site clusters identify target genes with similar tissue-wide expression and buffer against mutations [version 2; peer review: 2 approved].** *F1000Res.* 2018; **7**: 1933.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
26. Schneider TD: **Information content of individual genetic sequences.** *J Theor Biol.* 1997; **189**(4): 427–41.  
[PubMed Abstract](#) | [Publisher Full Text](#)
27. Schmidt N, Lareau CA, Keshishian H, et al.: **A direct RNA-protein interaction atlas of the SARS-CoV-2 RNA in infected human cells.** *bioRxiv.* 2020.  
[Publisher Full Text](#)
28. Rogan PK, Klesic R, Mucaki EJ, et al.: **Proposed mechanism of SARS-CoV-2 severe infection.** *Figshare.* 2020.  
[Publisher Full Text](#)
29. Dubois J, Terrier O, Rosa-Calatrava M: **Influenza viruses and mRNA splicing: doing more with less.** *mBio.* 2014; **5**(3): e00070–00014.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
30. Li X, Manley JL: **Cotranscriptional processes and their influence on genome stability.** *Genes Dev.* 2006; **20**(14): 1838–47.  
[PubMed Abstract](#) | [Publisher Full Text](#)
31. Hauer C, Sieber J, Schwarzl T, et al.: **Exon Junction Complexes Show a Distributional Bias toward Alternatively Spliced mRNAs and against mRNAs Coding for Ribosomal Proteins.** *Cell Rep.* 2016; **16**(6): 1588–603.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
32. De Maio FA, Rizzo G, Iglesias NG, et al.: **The Dengue Virus NS5 Protein Intrudes in the Cellular Spliceosome and Modulates Splicing.** *PLoS Pathog.* 2016; **12**(8): e1005841.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
33. Samji T: **Influenza A: understanding the viral life cycle.** *Yale J Biol Med.* 2009; **82**(4): 153–9.  
[PubMed Abstract](#) | [Free Full Text](#)
34. Van Nostrand EL, Freese P, Pratt GA, et al.: **A Large-Scale Binding and Functional Map of Human RNA Binding Proteins.** *Nature.* 2020; **583**(7818): 711–719.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
35. Mahony S, Benos PV: **STAMP: a web tool for exploring DNA-binding motif similarities.** *Nucleic Acids Res.* 2007; **35**(Web Server issue): W253–258.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
36. Pietrokovski S: **Searching databases of conserved sequence regions by aligning protein multiple-alignments.** *Nucleic Acids Res.* 1996; **24**(19): 3836–45.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
37. de Wilde AH, Snijder EJ, Kikkert M, et al.: **Host Factors in Coronavirus Replication.** *Roles Host Gene Non-Coding RNA Expr Virus Infect.* 2017; **419**: 142.  
[Publisher Full Text](#)
38. Perlman S, Netland J: **Coronaviruses post-SARS: update on replication and pathogenesis.** *Nat Rev Microbiol.* 2009; **7**(6): 439–50.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
39. Rogan PK, Mucaki EJ, Shirley BC: **Characteristics of human and viral RNA binding sites and site clusters recognized by SRSF1 and RNPS1.** *Zenodo.* 2020.  
<http://www.doi.org/10.5281/zenodo.3737089>
40. Dinakarpanian D, Raheja V, Mehta S, et al.: **Tandem machine learning for the identification of genes regulated by transcription factors.** *BMC Bioinformatics.* 2005; **6**: 204.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
41. Pachetti M, Marini B, Benedetti F, et al.: **Emerging SARS-CoV-2 mutation hot spots include a novel RNA-dependent-RNA polymerase variant.** *J Transl Med.* 2020; **18**(1): 179.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
42. Tang X, Wu C, Li X, et al.: **On the origin and continuing evolution of SARS-CoV-2.** *Natl Sci Rev.* 2020; [cited 2020 May 22].  
[Publisher Full Text](#)
43. Korber B, Fischer WM, Gnanakaran S, et al.: **Spike mutation pipeline reveals the emergence of a more transmissible form of SARS-CoV-2.** *bioRxiv.* 2020; 2020.04.29.069054.  
[Publisher Full Text](#)
44. Hadfield J, Megill C, Bell SM, et al.: **Nextstrain: real-time tracking of pathogen evolution.** *Bioinformatics.* 2018; **34**(23): 4121–3.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
45. Su YC, Anderson DE, Young BE, et al.: **Discovery of a 382-nt deletion during the early evolution of SARS-CoV-2.** *bioRxiv.* 2020; 2020.03.11.987222.  
[Publisher Full Text](#)

46. Sanz LA, Chédin F: **High-resolution, strand-specific R-loop mapping via S9.6-based DNA-RNA immunoprecipitation and high-throughput sequencing.** *Nat Protoc.* 2019; **14**(6): 1734–55.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
47. Rogan PK: **Multigene signatures of responses to chemotherapy derived by biochemically-inspired machine learning.** *Mol Genet Metab.* 2019; **128**(1–2): 45–52.  
[PubMed Abstract](#) | [Publisher Full Text](#)
48. Zhao JZL, Mucaki EJ, Rogan PK: **Predicting ionizing radiation exposure using biochemically-inspired genomic machine learning [version 2; peer review: 3 approved].** *F1000Res.* 2018; **7**: 233.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
49. Merekoulis G, Alexopoulos EC, Belezos T, et al.: **Lymphocyte to monocyte ratio as a screening tool for influenza.** *PLoS Curr.* 2010; **2**: RRN1154.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
50. Scrima A, Konicková R, Czyzewski BK, et al.: **Structural basis of UV DNA-damage recognition by the DDB1-DDB2 complex.** *Cell.* 2008; **135**(7): 1213–23.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
51. Wang H, Zhai L, Xu J, et al.: **Histone H3 and H4 Ubiquitylation by the CUL4-DDB-ROC1 Ubiquitin Ligase Facilitates Cellular Response to DNA Damage.** *Mol Cell.* 2006; **22**(3): 383–94.  
[PubMed Abstract](#) | [Publisher Full Text](#)
52. Marinov GK, Williams BA, McCue K, et al.: **From single-cell to cell-pool transcriptomes: stochasticity in gene expression and RNA splicing.** *Genome Res.* 2014; **24**(3): 496–510.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
53. Xia C, Fan J, Emanuel G, et al.: **Spatial transcriptome profiling by MERFISH reveals subcellular RNA compartmentalization and cell cycle-dependent gene expression.** *Proc Natl Acad Sci U S A.* 2019; **116**(39): 19490–9.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
54. Anczuków Olga, Akerman M, Cléry A, et al.: **SRSF1-Regulated Alternative Splicing in Breast Cancer.** *Mol Cell.* 2015; **60**(1): 105–17.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
55. Blanco-Melo D, Nilsson-Payant BE, Liu WC, et al.: **Imbalanced Host Response to SARS-CoV-2 Drives Development of COVID-19.** *Cell.* 2020; **181**(5): 1036–1045.e9.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
56. Timm C, Gupta A, Yin J: **Robust kinetics of an RNA virus: Transcription rates are set by genome levels.** *Biotechnol Bioeng.* 2015; **112**(8): 1655–62.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
57. Shannon A, Selisko B, Le NT, et al.: **Rapid incorporation of Favipiravir by the fast and permissive viral RNA polymerase complex results in SARS-CoV-2 lethal mutagenesis.** *Nat Commun.* 2020; **11**(1): 4682.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
58. Shi J, Wen Z, Zhong G, et al.: **Susceptibility of ferrets, cats, dogs, and other domesticated animals to SARS-coronavirus 2.** *Science.* 2020; **368**(6494): 1016–1020.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
59. Lin SC, Ho CT, Chuo WH, et al.: **Effective inhibition of MERS-CoV infection by resveratrol.** *BMC Infect Dis.* 2017; **17**(1): 144.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
60. Zhang L, Guo X, Xie W, et al.: **Resveratrol exerts an anti-apoptotic effect on human bronchial epithelial cells undergoing cigarette smoke exposure.** *Mol Med Rep.* 2015; **11**(3): 1752–8.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
61. Carabet LA, Leblanc E, Lallous N, et al.: **Computer-Aided Discovery of Small Molecules Targeting the RNA Splicing Activity of hnRNP A1 in Castration-Resistant Prostate Cancer.** *Mol Basel Switz.* 2019; **24**(4): 763.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
62. Zhou Y, Fu B, Zheng X, et al.: **Aberrant pathogenic GM-CSF<sup>+</sup> T cells and inflammatory CD14<sup>+</sup>CD16<sup>+</sup> monocytes in severe pulmonary syndrome patients of a new coronavirus.** *bioRxiv.* 2020; 2020.02.12.945576.  
[Publisher Full Text](#)
63. Shinya K, Gao Y, Cilloniz C, et al.: **Integrated Clinical, Pathologic, Virologic, and Transcriptomic Analysis of H5N1 Influenza Virus-Induced Viral Pneumonia in the Rhesus Macaque.** *J Virol.* 2012; **86**(11): 6055–66.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
64. Lehmann M, Hu Q, Hu Y, et al.: **Chronic WNT/ $\beta$ -catenin signaling induces cellular senescence in lung epithelial cells.** *Cell Signal.* 2020; **70**: 109588.  
[PubMed Abstract](#) | [Publisher Full Text](#)
65. Konno Y, Kimura I, Urieu K, et al.: **SARS-CoV-2 ORF3b is a potent interferon antagonist whose activity is further increased by a naturally occurring elongation variant.** *bioRxiv.* 2020; 2020.05.11.088179.  
[Publisher Full Text](#)
66. Goldstein SA, Weiss SR: **Origins and pathogenesis of Middle East respiratory syndrome-associated coronavirus: recent advances [version 1; peer review: 3 approved].** *F1000Res.* 2017; **6**: 1628.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
67. Bonenfant G, Meng R, Shotwell C, et al.: **Asian Zika Virus Isolate Significantly Changes the Transcriptional Profile and Alternative RNA Splicing Events in a Neuroblastoma Cell Line.** *Viruses.* 2020; **12**(5): 510.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
68. Boudreault S, Armero VES, Scott MS, et al.: **The Epstein-Barr virus EBNA1 protein modulates the alternative splicing of cellular genes.** *Virology.* 2019; **16**(1): 29.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
69. Mandal PK, Rossi DJ: **Reprogramming human fibroblasts to pluripotency using modified mRNA.** *Nat Protoc.* 2013; **8**(3): 568–82.  
[PubMed Abstract](#) | [Publisher Full Text](#)
70. Huelga SC, Vu AQ, Arnold JD, et al.: **Integrative genome-wide analysis reveals cooperative regulation of alternative splicing by hnRNP proteins.** *Cell Rep.* 2012; **1**(2): 167–78.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
71. Bi C, Rogan PK: **Bipartite pattern discovery by entropy minimization-based multiple local alignment.** *Nucleic Acids Res.* 2004; **32**(17): 4979–91.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
72. Wickham H: **ggplot2: Elegant Graphics for Data Analysis.** New York: Springer-Verlag; 2009 [cited 2020 May 19]. (Use R!).  
[Reference Source](#)
73. Ou J, Zhu LJ: **trackViewer: a Bioconductor package for interactive and integrative visualization of multi-omics data.** *Nat Methods.* 2019; **16**(6): 453–4.  
[PubMed Abstract](#) | [Publisher Full Text](#)

# Open Peer Review

Current Peer Review Status:    

---

## Version 2

Reviewer Report 07 January 2021

<https://doi.org/10.5256/f1000research.31422.r76718>

© 2021 Fonseca G. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.



### Gregory Fonseca

Division of Experimental Medicine, Department of Medicine, McGill University, Montreal, Canada

The authors have answered my concerns.

**Competing Interests:** No competing interests were disclosed.

**Reviewer Expertise:** Bioinformatics, transcriptomics, genomics. Lung disease.

**I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.**

Reviewer Report 06 January 2021

<https://doi.org/10.5256/f1000research.31422.r76719>

© 2021 Eperon I. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.



### Ian Eperon

Leicester Institute of Structural & Chemical Biology, Department of Molecular & Cell Biology, University of Leicester, Leicester, UK

No further comments.

**Competing Interests:** No competing interests were disclosed.

**Reviewer Expertise:** RNA splicing.

**I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.**

---

**Version 1**

Reviewer Report 26 November 2020

<https://doi.org/10.5256/f1000research.28014.r73886>

© 2020 Eperon I. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.



**Ian Eperon**

Leicester Institute of Structural & Chemical Biology, Department of Molecular & Cell Biology, University of Leicester, Leicester, UK

The purpose of this research is to establish by computational methods whether the genomes of certain RNA viruses could provide enough binding sites for proteins that bind nascent transcripts to reduce their availability in the cell and thereby trigger R-loop formation and ultimately apoptosis. The authors focussed their attentions of SRSF1 and RNPS1, both of which are known to prevent R-loops.

The first part of the work involved a derivation of a new information weight matrix from ENCODE eCLIP datasets. I am unable to comment on the methods used, but it is reassuring that the resultant consensus sequences for SRSF1 matched a coalescence of previous results from SELEX, RNA-seq and structural work. The authors then analysed the occurrence and distribution of all motifs in both viral and genomic transcripts that had a higher information content than the mean of the information content for all the binding sites in the model for each protein. It was assumed, but has not yet been tested, that the sites elected are the stronger binding sites. It is not clear whether this has any biological relevance, i.e., how bound lifetimes (affinity) vary across all the sequences in the model and whether the threshold chosen is likely to reflect the real behaviour of the proteins. Nonetheless, in the context of this heuristic work, this is not an unreasonable choice to make.

The authors analysed in particular the occurrence of clusters of these sites, i.e., where sites were within 25 nts of each other, and they looked for a correspondence between the locations of these clusters and sites located by high-throughput methods at which it was known that R-loops are likely to form. This correspondence was not strong, although the binding sites for RNPS1, SRSF1 and hnRNP A1 were closer than average in these regions. This is followed by a comparison of gene expression in patients with acute radiation sickness with those infected by the RNA viruses, which concluded that certain DNA damage-related proteins were expressed more highly in both types of patient, consistent with the overall hypothesis.

The final part describes an attempt to calculate the effects of viral genomes on the availability of

SRSF1 in cells. The authors assume that the Kd is 0.8  $\mu\text{M}$ . The Kd term they use is taken from assays with just the second RRM domain, but the value with both RRM domains is around 0.2  $\mu\text{M}$  (Anczukow *et al.* (2015))<sup>1</sup>. The native protein is, of course, affected by its RS domain and therefore the state of phosphorylation. The authors also assume that there is no free protein, but this is not supported by any arguments. Taking the Kd to be 0.2  $\mu\text{M}$ , the cellular concentration of the protein to be 3.6  $\mu\text{M}$  (Hein *et al.* (2015))<sup>2</sup> and the concentration of sites to be as described by the authors, this seems reasonable. However, and this is a significant caveat, the protein is not distributed evenly throughout the cell: it is largely nuclear and, within the nucleus, may be sequestered in speckles. Thus, the concentration in the cytoplasm, where it would encounter the viral RNA, might be much lower and thus affect the authors' argument. The authors' model involves competition between the nascent transcripts (nuclear) and the viral RNA (cytoplasmic) for SRSF1 binding, and any difference in the local concentrations and proportions of sites bound would undermine the model. However, none of the values required are known accurately and so, again, for the purpose of developing a model it is reasonable to make these simplifying assumptions.

Overall, this is an interesting piece of work that makes the best use of the limited data available to support a model that proposes new and plausible routes by which RNA viruses could cause widespread apoptosis. It would be improved by a more rigorous discussion of the assumptions made, as noted above.

### References

1. Anczuków O, Akerman M, Cléry A, Wu J, et al.: SRSF1-Regulated Alternative Splicing in Breast Cancer. *Mol Cell*. 2015; **60** (1): 105-17 [PubMed Abstract](#) | [Publisher Full Text](#)
2. Hein MY, Hubner NC, Poser I, Cox J, et al.: A human interactome in three quantitative dimensions organized by stoichiometries and abundances. *Cell*. 2015; **163** (3): 712-23 [PubMed Abstract](#) | [Publisher Full Text](#)

**Is the work clearly and accurately presented and does it cite the current literature?**

Yes

**Is the study design appropriate and is the work technically sound?**

Yes

**Are sufficient details of methods and analysis provided to allow replication by others?**

Yes

**If applicable, is the statistical analysis and its interpretation appropriate?**

Yes

**Are all the source data underlying the results available to ensure full reproducibility?**

Yes

**Are the conclusions drawn adequately supported by the results?**

Yes

**Competing Interests:** No competing interests were disclosed.



**Reviewer Expertise:** RNA splicing.

**I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.**

Author Response 11 Dec 2020

**Peter Rogan**, University of Western Ontario, London, Canada

*It was assumed, but has not yet been tested, that the sites elected are the stronger binding sites. It is not clear whether this has any biological relevance, i.e., how bound lifetimes (affinity) vary across all the sequences in the model and whether the threshold chosen is likely to reflect the real behaviour of the proteins.*

Response: The IWMs derived in this manuscript may not yet been verified in the laboratory, however we have verified binding sites by such approaches previously (Vyhlidal, Rogan et al. *J Biol Chem.* 2004. 279:46779-86). The information theory framework used to derive them has been well validated and the relationships between information content and binding affinity have been rigorously proven (Schneider. *J Theor Biol.* 1997; 189:427-41). While the models used in this study were recently derived, other information theory-based RNA-protein binding site models have been utilized in hundreds of published studies, some involving the verification of phenotypes of mutations that alter splice sites, and others applied to transcription factor binding site recognition (Rogan et al. *Hum Mutat.* 1998; **12**: 153-171; Rogan et al. *Pharmacogenetics.* 2003; **13**: 207-218; Caminsky et al. *F1000Res.* 2014; 3:282; Lu et al. *Nucl. Ac. Res.* 45: e27, 2017). Our assumptions regarding binding site strength are well founded both theoretically and experimentally for many proteins. Quantification of the predicted strengths of these binding sites made by these models is reasonable and may likely reflect actual protein binding events.

*The K<sub>d</sub> term they use is taken from assays with just the second RRM domain, but the value with both RRM domains is around 0.2 μM (Anczukow et al. (2015)). The authors also assume that there is no free protein, but this is not supported by any arguments.*

Response: We are grateful to the reviewer for pointing out the updated dissociation constant of SRSF1 (Anczukow et al. 2015; Ref. 54), based on assays that included both RRM domains. We had inadvertently overlooked this study. The K<sub>d</sub> value used in the initial version of our paper (0.8μM) was based on an earlier publication from the same group (Cléry et al. *Proc Natl Acad Sci U S A.* 2013; 110(30):E2802-2811). The new K<sub>d</sub> value altered Scatchard analysis presented in Figure 6. The number of doublings for replicated virus was significantly increased, which reduced the discrepancy with the number of viral genomes required to compete with host transcriptome binding sites in A549 cell lines. In this revision, we have recomputed all values based on K<sub>d</sub> from Anczukow et al. (2015) and have updated the main text, Figure 6 and Section 1 Tables 9A and 9B (extended data; Ref. 39).

We previously did not include a justification for our assumption that [L] = 0 (no free protein) in our derivation of the Scatchard plot. The proposed mechanism relies on the likelihood that these RBPs are largely sequestered by binding to viral sequences, so that their effective concentration in the nucleus is inadequate to prevent R-loops (thus, [L] ≈ 0). SARS-CoV-2

replication is highly efficient and rapid, leading to levels constituting up to 60% of the total cellular RNA (Blanco-Melo et al. 2020; Ref. 55). Viral replication produces an excess of viral binding sites that will perturb the equilibrium between bound and free RBPs (Le Chatelier's Principle) and drive binding of free RBPs and reduce the pool of free RBPs. The degree of viral replication that depletes nuclear RBP concentrations to a point at which the abundance of these factors becomes insufficient to prevent R-loop formation is not known.

To clarify our assumption that all RBPs are bound to viral (or host), we have added the following to the Results:

"We assumed no free RNA binding protein (that the vast majority of SRSF1 is bound to either host or viral binding sites) as the concentration of free RBPs is likely to be low due to sequestration of RBPs by the excess of viral sequences present in infected cells (~60% of all RNA<sup>55</sup>)"

and to the Methods:

"Upon infection and viral replication, it is assumed there is no free RNA binding protein (all RBP is assumed to be bound to either viral or host RNA)."

*However, and this is a significant caveat, the protein is not distributed evenly throughout the cell: it is largely nuclear and, within the nucleus, may be sequestered in speckles. Thus, the concentration in the cytoplasm, where it would encounter the viral RNA, might be much lower and thus affect the authors' argument. The authors' model involves competition between the nascent transcripts (nuclear) and the viral RNA (cytoplasmic) for SRSF1 binding, and any difference in the local concentrations and proportions of sites bound would undermine the model.*

Response: The reviewer has commented that the difference in local RBP concentrations between the cytoplasm and the nucleus would have an impact on the amount of RBP (such as SRSF1) that could possibly be sequestered in the cytoplasm by viral RNA. We do not state that we assume that these RBPs are uniformly distributed within the cell. We suggest that viral RNA binds to newly synthesized RBPs that have been translated in the cytoplasm, which could result in such an imbalance by limiting their nuclear entry (last paragraph of the "Proposed molecular pathogenetic mechanism of RNA-viral infection" section). It was also illustrated in panels 3 and 4 of the infographic of our proposed mechanism (Ref. 28; (<http://doi.org/10.6084/m9.figshare.12718799.v2>).

**Competing Interests:** PKR cofounded and BCS is an employee of CytoGnomix Inc.

Reviewer Report 24 November 2020

<https://doi.org/10.5256/f1000research.28014.r73887>

© 2020 Fonseca G. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

**Gregory Fonseca**

Division of Experimental Medicine, Department of Medicine, McGill University, Montreal, Canada

In this study, the authors present evidence that R-loops are associated with RNA binding protein binding sites and this may lead to DNA damage associated apoptosis. The authors define IWMs for RNA binding proteins, SRSF1 and RNPS1 based on previously published, high quality data. They then show the occurrence and quality of these IWMs in viral genomes including the relative stability of these IWMs across evolution. The authors then compared the quality and relative quantity of these IWMs in the human transcriptome. They predict the number of viral RNA particles necessary to squelch RNA binding proteins from the human genome.

Overall, this is an extremely interesting paper with a very exciting hypothesis. The paper is well written and well organized and makes use of available datasets very well.

A few notes.

- It should be mentioned whether IWM discovery was compared to background to understand if IWNs are found above random chance.
- If you randomly curate IWMs from 5000 sites of Rep1 or bin 5000 sites do the results compare to Rep2?
- What would the predicted likelihood of IWMs changing by chance compare to observed?
- Is there a correlation of SRSF and RNPS1 sites in the mRNA and gene expression at basal and during infection?

**Is the work clearly and accurately presented and does it cite the current literature?**

Yes

**Is the study design appropriate and is the work technically sound?**

Yes

**Are sufficient details of methods and analysis provided to allow replication by others?**

Yes

**If applicable, is the statistical analysis and its interpretation appropriate?**

I cannot comment. A qualified statistician is required.

**Are all the source data underlying the results available to ensure full reproducibility?**

Yes

**Are the conclusions drawn adequately supported by the results?**

Yes

**Competing Interests:** No competing interests were disclosed.

**Reviewer Expertise:** Bioinformatics, transcriptomics, genomics. Lung disease.

**I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.**

Author Response 11 Dec 2020

**Peter Rogan**, University of Western Ontario, London, Canada

*1. It should be mentioned whether IWM discovery was compared to background to understand if IWNs are found above random chance.*

Response:

We derived IWMs from (mock) control binding studies for SRSF1 and RNPS1. Sequences obtained for 3 negative controls were generated in the same dataset that was used to build IWMs for RNPS1 (E-MTAB-4215; control datasets 'ERR1201436', 'ERR1201437' and 'ERR1201438'). These were samples containing a GFP-tag (lacking the RNPS1 fusion protein for pulldown). IWMs were also derived from two SRSF1 control datasets ('ENCFF241ORF' and 'ENCFF773PUP'; positive and negative strands from mock input sample). From these datasets, we derived IWMs of length 6nt from the top 5,000 peaks from each dataset with Maskminent using the same parameters (see Methods).

The motifs of the newly derived control IWMs were compared to the true SRSF1 and RNPS1 models using STAMP software (described in Methods). The e-value obtained from STAMP is the number of hits expected against a database of the same size (i.e. 5,000 sequences containing random sequences of the same length). The comparison is based on the log likelihood ratio of the e-values of the IWM motifs derived from the RBP bound sequences relative to the sequences obtained from the mock control, which is a modified LOD score. The LOD scores for the RNPS1 model ranged from 3.8 to 6.1 depending on which control IWM was compared, and for SRSF1, these scores ranged from 4.4 to 8.8. We therefore conclude that the motifs obtained for these protein binding sites are *significantly* more robust (and different from a random set of sequences of the same size and composition generated from control samples). Please note that this analysis has been incorporated into the manuscript (the fourth paragraph of the Results and in the second and fourth paragraphs of the Methods).

*2. If you randomly curate IWMs from 5000 sites of Rep1 or bin 5000 sites do the results compare to Rep2?*

Response:

In this study, the two SRSF1 IWMs utilized were derived from two separate replicates from publicly available eCLIP data. The first model "SRSF1 Replicate 1" was based on the 50,000 largest eCLIP peaks from the 'ENCFF179SCM' replicate, while "SRSF1 Replicate 2" was based on the top 5,000 peaks in the 'ENCFF184TBM' replicate. Despite being derived from a far smaller dataset, the models were computed to be quite similar and non-random by STAMP analysis (e-score:  $7.4e-10$ ).

While the SRSF1 "Replicate 1" and "Replicate 2" models were those which were selected to be used in the study, IWMs for SRSF1 were derived utilizing a series of different conditions (i.e. number of peaks, number of Monte Carlo cycles, etc.), however discussion of these

additional derived SRSF1 models was not included in the final manuscript. Most commonly, models derived under these varying conditions were similar to that of the final models. On occasion, the method used here has been reported to identify binding motifs of other factors whose binding site sequences are in close proximity with the factor being crosslinked, as well as IWMs with a noise motifs that can resemble repetitive sequences (Lu et al. Nucleic Acids Res. 2017 Mar 17;45(5):e27). We carefully evaluated each IWM before it was utilized in any downstream analyses. For example, while the SRSF1 model derived from 10,000 peaks from the 'replicate 1' dataset is highly similar to that of the 50,000 replicate 1 peak model (as well as those models derived from replicate 2 data), the 5,000 replicate 1 peak model contained a slight variation of the primary motif, reporting instead an unexpected "G[G/C]AG" sub-motif. The pairwise IWM e-values from comparison of SRSF1 Replicates 1 and 2 are: "SRSF1 Replicate 2" (self comparison): 3.9e-11; "SRSF1 Replicate 1" (top e-CLIP 50,000 peaks): 7.4e-10; "SRSF1 Replicate 1" (top 10,000 peaks): 3.6e-09; and "SRSF1 Replicate 1" (top 5,000 peaks): 5.0e-03. In general, however, modifying the number of binding sites from which models are derived generally leads to IWMs with highly similar binding motifs.

3. *What would the predicted likelihood of IWMs changing by chance compare to observed?*

Response:

See response to Question 1.

4. *Is there a correlation of SRSF and RNPS1 sites in the mRNA and gene expression at basal and during infection?*

Response:

The expression of basal SRSF1 and RNPS1 may be significantly altered in an infected cell. Blanco-Melo et al. (Ref. 55) performed differential gene expression analysis of A549 immortalized cell lines, comparing infection with influenza A- or SARS-CoV-2 with controls and ranking genes by fold change p-values adjusted for multiple testing. Both SRSF1 and RNPS1 exhibited statistically significant lowered expression in SARS-CoV-2 infected A549 cell lines (73% [p=3.7e-16] and 47% [p=3.5e-89] of controls, respectively). Their expression was also significantly decreased in SARS-CoV-2 infected Calu-3 cell lines (74% for SRSF1 and 75% for RNPS1). Expression of SRSF1 and RNPS1 in A549 cells infected with respiratory syncytial virus was also significantly reduced (73% for both). No significant differences were evident in A549 infected with either SARS-CoV-1, MERS nor influenza A. Expression of SRSF1 and RNPS1 is not only altered by viral infection, but the extent of these changes is related to the specific infectious pathogen.

We have added a statement to the manuscript acknowledging that the gene expression datasets utilized in this study were from uninfected cells which may differ in infected cells (new text bolded):

"These are derived from the number of SRSF1 and RNPS1 sites expressed in either a single A549 cell or a type II pneumocyte (**cells were not infected; note that infection would be expected to alter the expression profile, which could affect expressed binding site estimates**)."

**Competing Interests:** 1.PKR cofounded and BCS is an employee of CytoGnomix Inc.

Reviewer Report 18 November 2020

<https://doi.org/10.5256/f1000research.28014.r73885>

© 2020 Srivastava M. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.



**Mansi Srivastava**

Indiana University School of Medicine, Indianapolis, IN, USA

This study provides mechanistic insights into the RNA viral infection that triggers unrepaired sites of chromosomal breakage, causing apoptosis and consequentially, high-titer viral release. The hypothesis suggests that the viral genome binds RNA binding proteins of the host thus, preventing their essential post-transcriptional activities.

In the result section that describes the human transcriptome analysis of RNA binding sites, the authors evaluate the frequency of RBP binding in human transcriptomes to relate the relative abundance of these proteins bound to viral RNAs compared to their normal reservoir in host nuclear RNA of infected cells. To support their observation, authors should include a discussion on the impact of other RNA binding proteins that may bind/regulate the same site on the viral genome.

Authors discuss that the viral genome may have an advantage in the competition for binding to RBPs relative to nuclear transcripts, due to the proximity of the viral genome to nascent RBPs in the cytoplasm, which may limit transport and impede their import into the nucleus. However, the authors should also mention RBPs that shuttle between the nucleus and cytoplasm dynamically and thus account for binding to the nascent transcripts at the basal level.

**Is the work clearly and accurately presented and does it cite the current literature?**

Yes

**Is the study design appropriate and is the work technically sound?**

Yes

**Are sufficient details of methods and analysis provided to allow replication by others?**

Yes

**If applicable, is the statistical analysis and its interpretation appropriate?**

Yes

**Are all the source data underlying the results available to ensure full reproducibility?**

Yes

**Are the conclusions drawn adequately supported by the results?**

Yes

**Competing Interests:** No competing interests were disclosed.

**Reviewer Expertise:** Genomics and Systems Biology of RNA regulatory processes.

**I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.**

Author Response 11 Dec 2020

**Peter Rogan**, University of Western Ontario, London, Canada

Thank you for reviewing our manuscript. Newly synthesized RBPs imported into the nucleus and bound to nascent transcripts do not necessarily have an impact on R-loop formation. Our study put a strong focus on SRSF1 and RNPS1, which have been documented to be antagonistic to R-loop formation. We have not investigated other RNA binding proteins that are known to stabilize nascent transcripts, such as the THO complex, *PCF11*, and the exoribonucleases *EXOSC3* and *EXOSC10* (Santos-Pereira & Aguilera. *Nat Rev Genet.* 2015. 16: 583-597). The CLIP data required to analyze RNA for binding sites is not currently available for many of these RBPs.

Ribosomal RNA (rRNA) constitutes the most abundant RNA in the cytoplasm (indeed, the cell), and would likely be the most likely to interact with SRSF1 and RNPS1. Ribosomal proteins interacting with rRNA-scaffold would likely represent the most abundant competitor to viral RNAs in infected cells. rRNA interactions have a structural basis (e.g. bulged duplexes, hairpin loops) that explains their affinity for ribosomal proteins. This contrasts with the sequence-specific binding by SRSF1 and RNPS1 and other RBPs containing one or more RRM domains (Ciriello et al. *BMC Bioinformatics.* 2010; 11(Suppl 1): S41). The method we describe does not detect or quantify the type of structural interactions seen in rRNA and ribosomal proteins. The present approach cannot determine whether viral RNAs could bind to ribosomal proteins.

**Competing Interests:** PKR cofounded and BCS is an employee of CytoGnomix Inc.

Reviewer Report 13 November 2020

<https://doi.org/10.5256/f1000research.28014.r73888>

© 2020 Romano M. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.



**Maurizio Romano** 

Department of Life Sciences, University of Trieste, Trieste, Italy

In the present manuscript, starting from the observation that riboviruses can cause fatal pulmonary in some infected patients, the Authors propose the interesting hypothesis that depletion of host RNA binding proteins from nuclear RNA bound to replicating viral sequences might be part of the mechanism that trigger apoptosis and viral release.

Information theory-based analysis was used to test interactions between RBPs and individual sequences in different virus, since expression of viral sequences might sequester RBPs (the study is focused on SRSF1 and RNPS1). It is proposed a correlation RBPs depletion / destabilization of R-loops / chromosomal breakage.

The stoichiometry of inhibition of RBPs in host nuclear RNA has been estimated by counting competing binding sites in replicating viral genomes and host RNA.

It is concluded that the RNA virally-induced apoptosis could lead to release significant quantities of membrane-associated virions and cause the fatal pulmonary.

Although functional analyses might be helpful to strengthen the validity of the proposed mechanism, all the steps and conclusions of the study are sufficiently clear to support the hypothesis.

The Discussion might be shortened by taking out aspects that are not directly related to the proposed theory.

**Is the work clearly and accurately presented and does it cite the current literature?**

Yes

**Is the study design appropriate and is the work technically sound?**

Yes

**Are sufficient details of methods and analysis provided to allow replication by others?**

Yes

**If applicable, is the statistical analysis and its interpretation appropriate?**

Yes

**Are all the source data underlying the results available to ensure full reproducibility?**

Yes

**Are the conclusions drawn adequately supported by the results?**

Yes

**Competing Interests:** No competing interests were disclosed.

**Reviewer Expertise:** Splicing; Neuroscience.



**I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.**

Author Response 11 Dec 2020

**Peter Rogan**, University of Western Ontario, London, Canada

Thank you for reviewing our manuscript. We concur that the Discussion section of the manuscript is lengthy. We have therefore removed the paragraph beginning with “The immune system appears to be a witness, rather than a direct participant ...”, as it is only tangentially related from the mechanism being proposed.

**Competing Interests:** PKR cofounded and BCS is an employee of CytoGnomix Inc.

The benefits of publishing with F1000Research:

- Your article is published within days, with no editorial bias
- You can publish traditional articles, null/negative results, case reports, data notes and more
- The peer review process is transparent and collaborative
- Your article is indexed in PubMed after passing peer review
- Dedicated customer support at every stage

For pre-submission enquiries, contact [research@f1000.com](mailto:research@f1000.com)

**F1000Research**