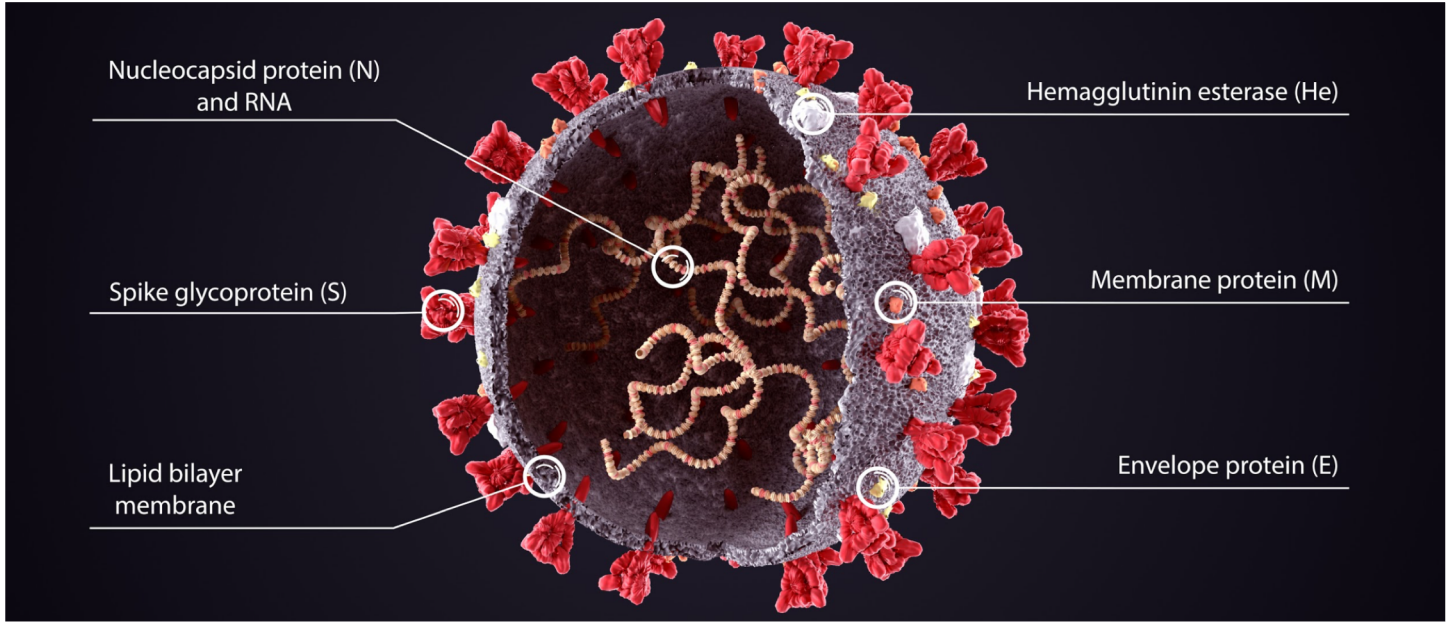
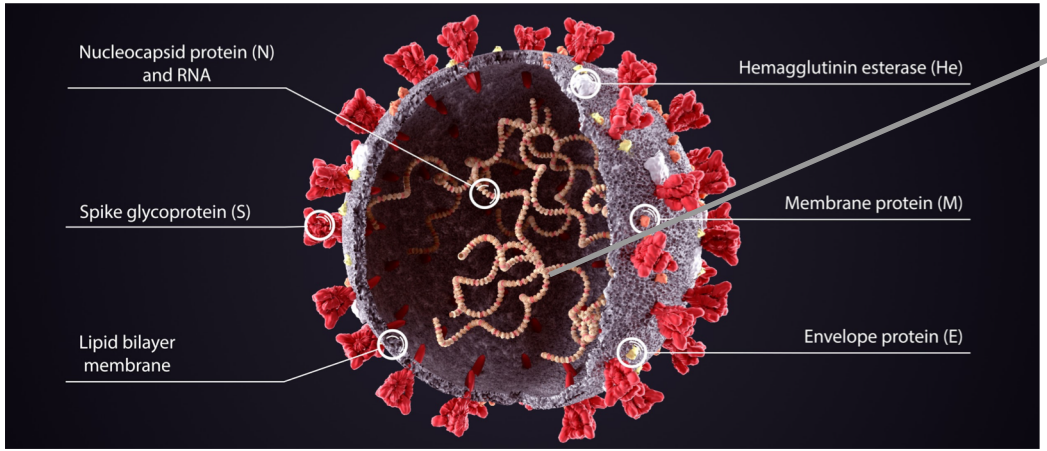
An R framework to predict interactions between the SARS-CoV-2 genome and human proteins

Avantika Lal

SARS-CoV-2 structure



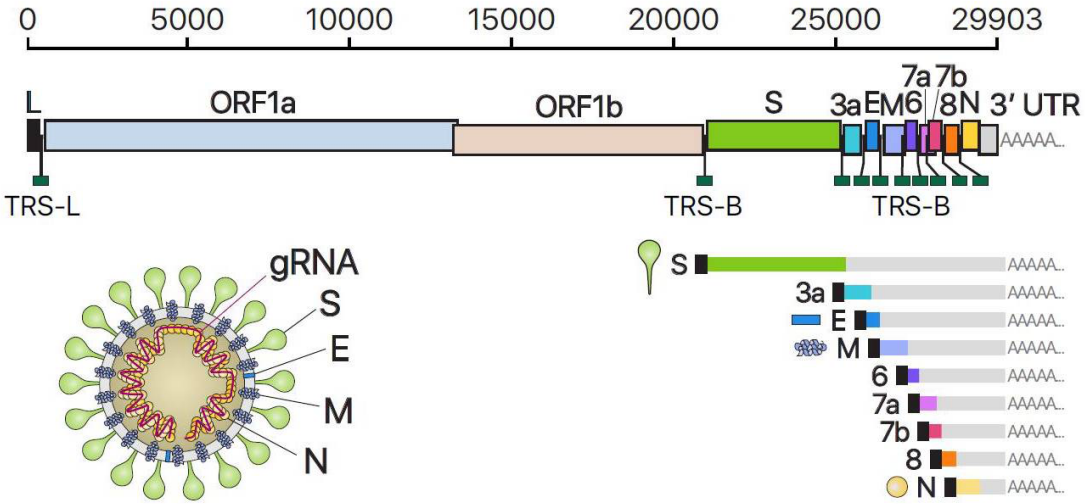
SARS-CoV-2 RNA genome



ATTAAAGGTTTATACCTTCC
CAGGTAACAAACCAACCAA
CTTTCGATCTCTTGTAGATC
TGTTCTCTAAACGAACTTTA
AAATCTGTGTGGCTGTCACT
CGGCTGCATGCTTAGTGCA
CTCACGCAGTATAATTAATA
ACTAATTACTGTCGTTGACA
GGACACGAGTAACTCGTCT
ATCTTCTGCAGGCTGCTTAC
GGTTTCGTCCGTGTTGCAGC
CGATCATCAGCACATCTAG
GTTTCGTCCGGGTGTGACC
GAAAG... (approximately
30,000 letters)

SARS-CoV-2 RNA encodes viral proteins

- 5'UTR
- Genes encoding viral proteins
- 'Intergenic' regions
- 3'UTR

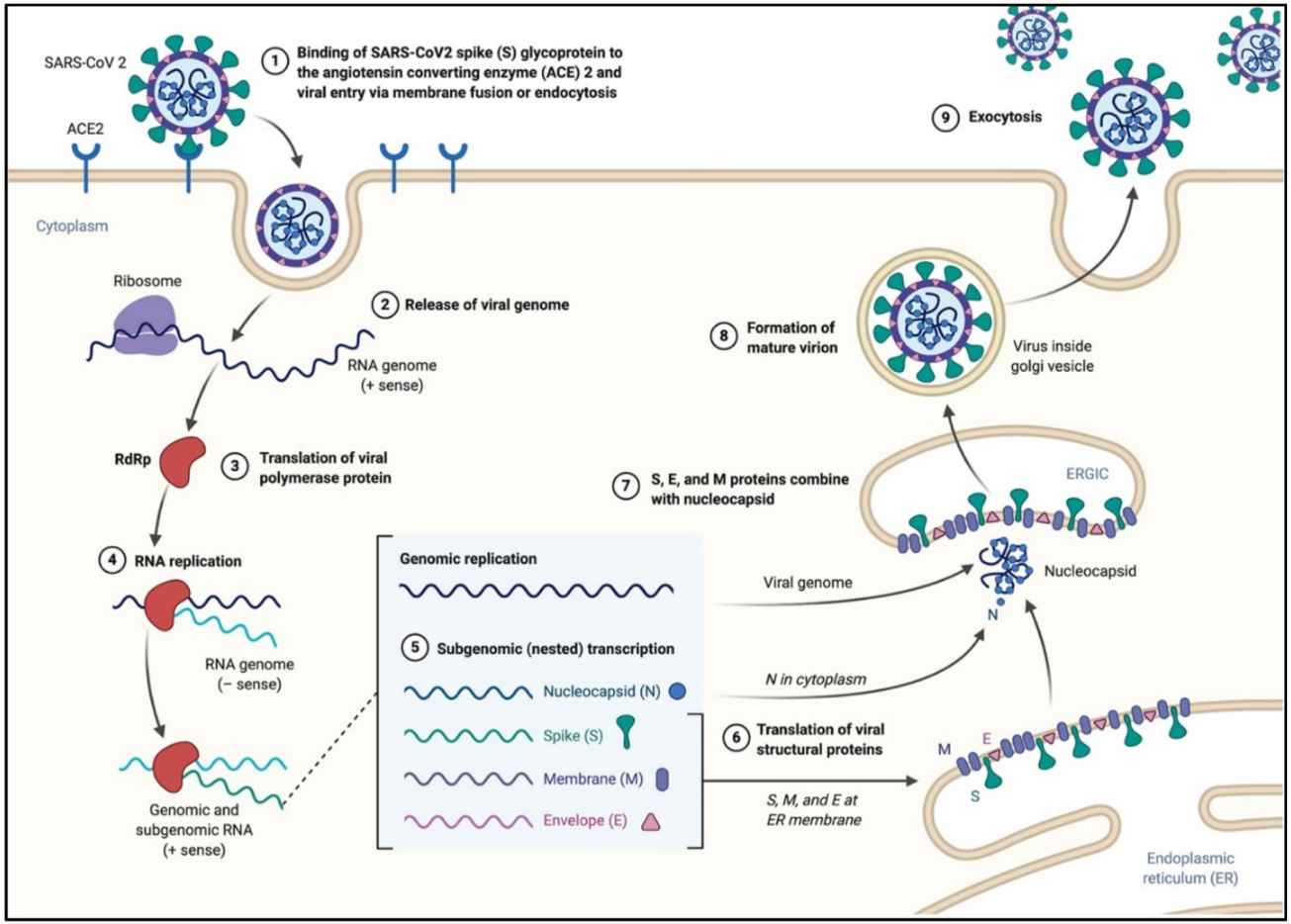


Kim et al. 2020

SARS-CoV-2 life cycle

SARS-CoV-2 needs to interact with many components of the human cell.

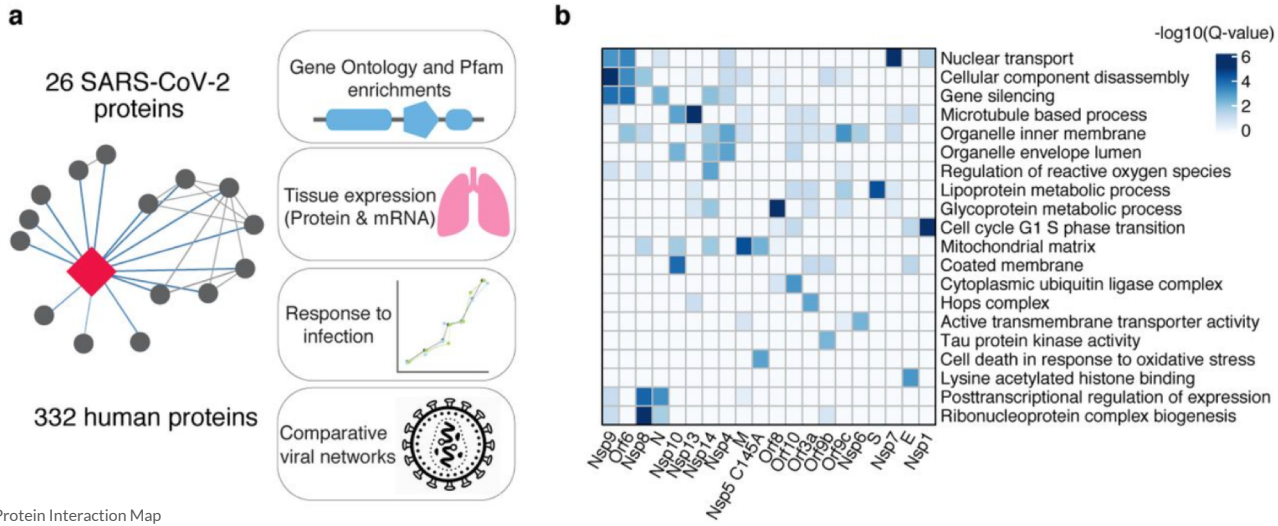
Protein synthesis, RNA replication, stability, transport, assembly.



Interaction map of SARS-CoV-2 proteins and human proteins

What molecules in human cells interact with SARS-CoV-2?

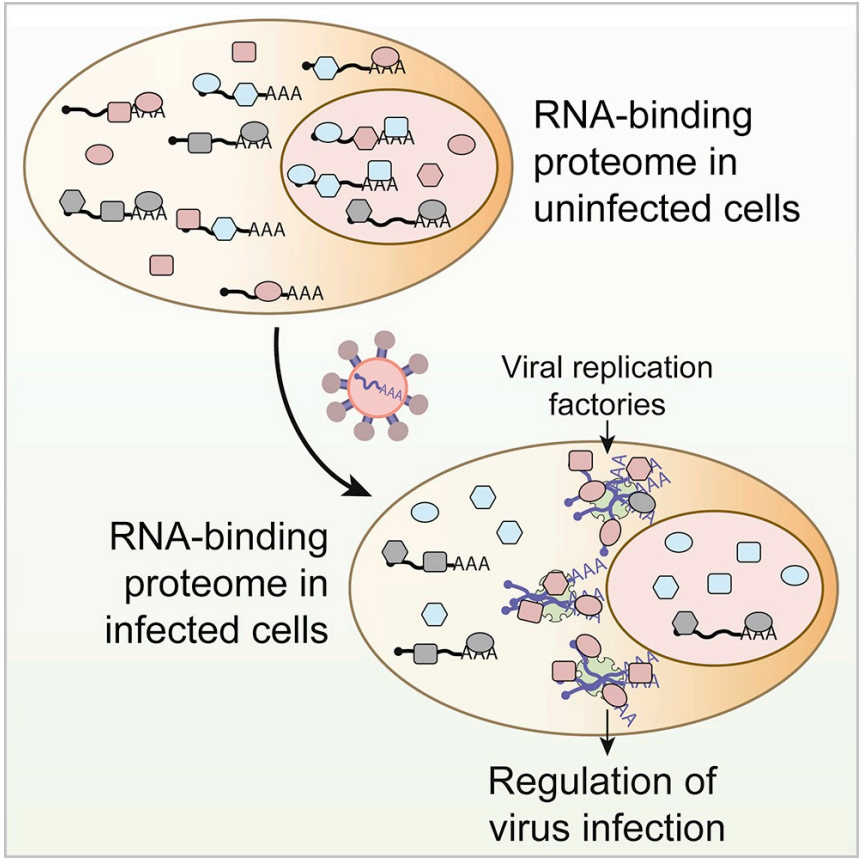
Potential drug targets.



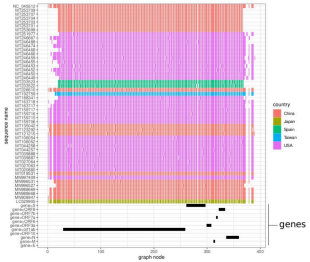
Gordon, David E., et al. "A SARS-CoV-2-Human Protein-Protein Interaction Map Reveals Drug Targets and Potential Drug-Repurposing." *bioRxiv* (2020).

RNA-protein interactions in other viruses

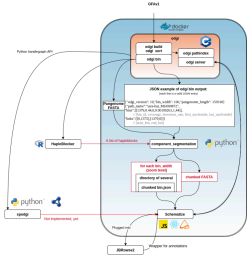
- Replication
- Stability
- Protein synthesis



Virtual Biohackathon on Covid-19



Pangenome

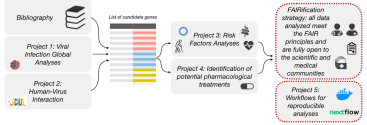


Pangenome browser

Machine learning

nf-core/viralrecon

Workflows

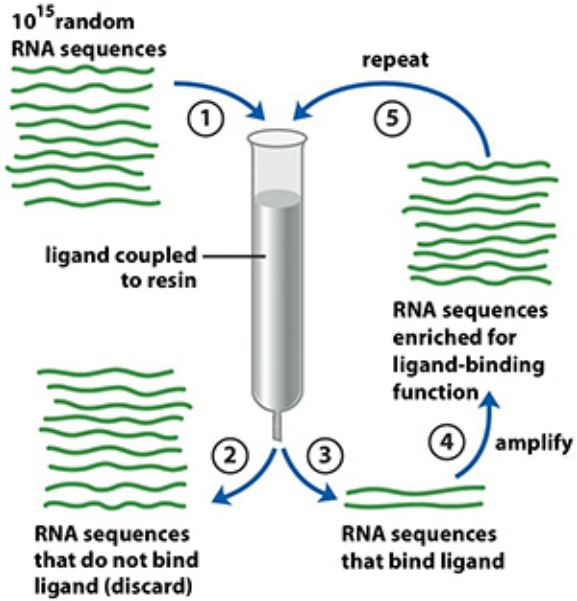


Gene expression

Question (one of many)

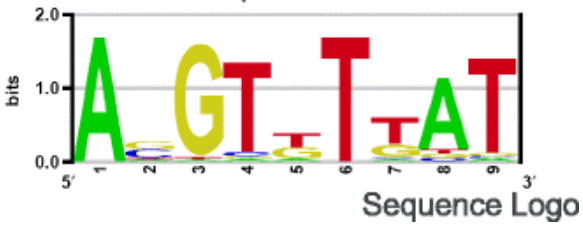
- Can we computationally predict interactions between SARS-CoV-2 RNA and human proteins?

Encoding RNA-protein binding affinities



	1	2	3	4	5	6	7	8	9
A	.97	.10	.02	.03	.10	.01	.05	.85	.03
C	.01	.40	.01	.04	.05	.01	.05	.05	.03
G	.01	.40	.95	.03	.40	.01	.3	.05	.03
T	.01	.10	.02	.90	.45	.97	.6	.05	.91

PWM



Sequence Logo

Scanning the SARS-CoV-2 genome with PWMs

ATtRACT Database for RBP PWMs

RBP	PWM

Binding specificities for 102 human RNA-binding proteins

Obtained by competitive experiments

	1	2	3	4	5	6	7	8	9
A	.97	.10	.02	.03	.10	.01	.05	.85	.03
C	.01	.40	.01	.04	.05	.01	.05	.05	.03
G	.01	.40	.95	.03	.40	.01	.3	.05	.03
T	.01	.10	.02	.90	.45	.97	.6	.05	.91

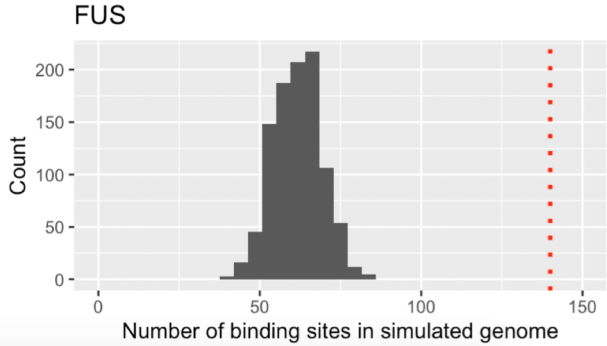
Scan sequence with multiple PWMs in parallel

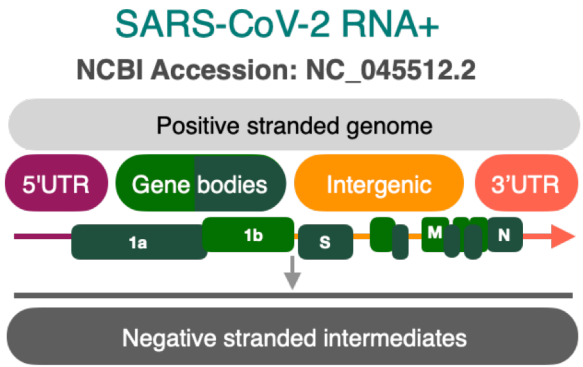
AGCGAT**AGGTATGAT**CTAGTATAGAT

Score potential binding sites using Biostrings package in R

Binding site enrichment

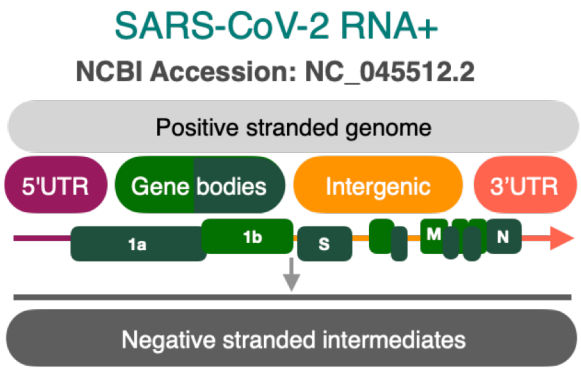
- Shuffle sequence 1000 times
- Perform binding site discovery on each shuffled sequence
- Distribution of number of binding sites per protein
- Calculate z-score relative to background distribution
- Significantly under-represented and over-represented proteins



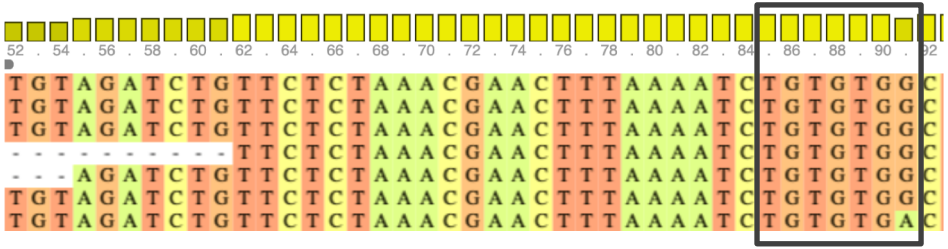


Region	Sites	RBPs
5’UTR	8	3
Intergenic regions	39	8
3’UTR	77	10
Positive Sense Genome	6848	19
Negative sense Intermediates	4616	16

Conservation of binding sites across SARS-CoV-2 isolates



CELF5 binding site

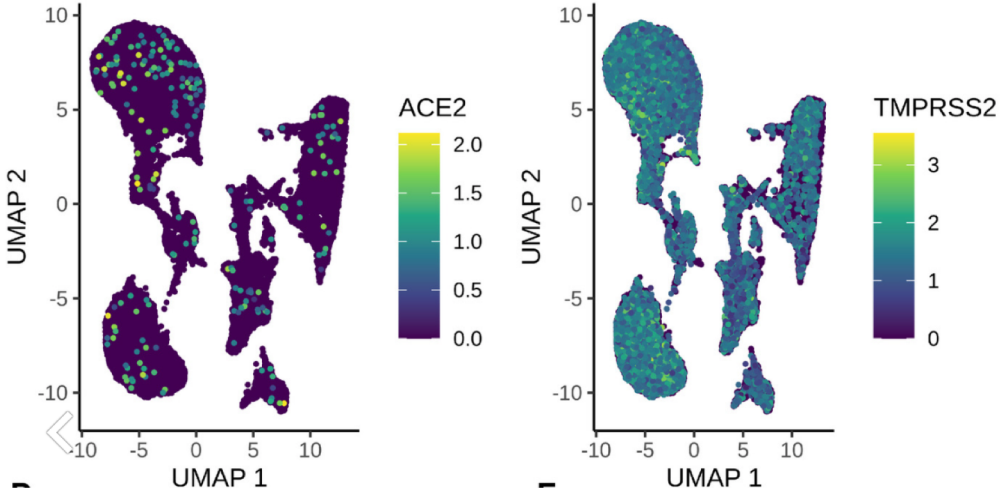


Region	RBPs
5'UTR	CELF5, FMR1, RBM24
3'UTR	HNRNPA1, HNRNPA1L2, HNRNPA2B1, LIN28A, PPIE
Intergenic regions	EIF4B, ELAVL1, ELAVL2, KHDRBS1, PABPC1, PPIE, TIA1, TIAL1

Load and analyze sequence alignment using seqinR package

Which of these proteins are present in susceptible cells?

- Lung cells expressing ACE2 and TMPRSS2 receptors are susceptible to infection.
- Analyzed single-cell gene expression data from human lungs and bronchi.
- Which of our candidate proteins are present in susceptible cells?



Lukassen, Soeren, et al. "SARS-CoV-2 receptor ACE 2 and TMPRSS 2 are primarily expressed in bronchial transient secretory cells." *The EMBO journal* 39.10 (2020): e105114.

Does the level of the candidate proteins change in infected cells?

- Analyzed gene expression data from three human lung/bronchial cell lines, uninfected and infected by SARS-CoV-2
- Do any of our candidate proteins increase/decrease in infection in SARS-CoV-2 infected cells?

- This could indicate:
  - Virus increasing expression of a protein it needs
  - Host cell increasing expression to compensate for virus taking up protein
  - Host cell decreasing expression to repress virus

Biological connections

Five of our top candidates (EIF4B, hnRNPA1, PABPC1, PABPC4, and YBX1) were experimentally shown to bind to SARS-CoV-2 RNA in an infected human liver cell line.

RBP	DE Analysis			Expression in GTex Datasets		PPI Map*2	RBP binding site prediction	
	A549 LogFC	Calu-3 LogFC	SARS-CoV-2 Specific DEG	Lung Tissue (TPM)	scRNA*1	Viral Protein	Enriched	Conserved*3
PABPC1	0.72	0.44	✓	448.025	✓	N	Positive Strand	✓
PABPC4	0.30	-0.28		103.082	✓	N		
HNRNPL	-0.97	-0.30	✓	208.765	✓			
PABPC1	0.72	0.44	✓	448.025	✓	N	3'UTR	✓
PABPC4	0.30	-0.28		103.082	✓	N		
HNRNPA1	0.00	-0.32		331.336	✓			✓
HNRNPA2B1	-1.08	-0.29		539.829	✓			✓
CELF5	0.56	0.00		0.079	✓		5'UTR	✓
RBM24	0.34	0.00		1.412				✓
PABPC1	0.72	0.44	✓	448.025	✓	N	Intergenic	✓
EIF4B	0.53	0.64		170.303	✓			✓
TIA1	0.34	0.41	✓	46.934	✓			✓
TIA1	0.34	0.41	✓	46.934	✓		Negative Strand	✓
HNRNPL	-0.97	-0.30	✓	208.765	✓			

*1 scRNA expression in ACE+ and TMPRSS2+ lung cells, dataset GSE122960;
*2 PPI Map : Gordon et al, 2020, PMID 32353859;
*3 Conserved in SARS-CoV-2 genomes.

Comparisons with related coronaviruses

- Currently: SARS1 and RaTG13
- SARS-CoV-2 genome is uniquely enriched for binding sites of CELF5 in its 5'UTR, PPIE on its 3'UTR, and ELAVL1 in the viral negative-sense RNA molecule.
- These three proteins are important for RNA stability.

In progress

- Adding more proteins to the analysis – import PWMs from different sources and formats
- Add more sophisticated methods (including deep learning) to improve binding site identification
- Package code
    - Download and format PWMs from different sources/experiment types
    - Import viral genomes and annotations
    - Scan genomes to identify binding sites
    - Shuffle genomes with different settings
    - Visualize and calculate enrichment/depletion in selected regions
    - Visualize and calculate binding site conservation
    - Compare species
- New goal: Apply this methodology to all viruses that infect humans.

Preprint

https://www.biorxiv.org/content/10.1101/2020.07.28.225581v2

Genome-wide bioinformatic analyses predict key host and viral factors in SARS-CoV-2 pathogenesis

Mariana G. Ferrarini, Avantika Lal, Rita Rebollo, Andreas Gruber, Andrea Guarracino, Itziar Martinez Gonzalez, Taylor Floyd, Daniel Siqueira de Oliveira, Justin Shanklin, Ethan Beausoleil, Taneli Pusa, Brett E. Pickett, Vanessa Aguiar-Pulido

doi: https://doi.org/10.1101/2020.07.28.225581

This article is a preprint and has not been certified by peer review [what does this mean?].

Abstract Full Text Info/History Metrics Preview PDF

Abstract

The novel betacoronavirus named Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) caused a worldwide pandemic (COVID-19) after initially emerging in Wuhan, China. Here we applied a novel, comprehensive bioinformatic strategy to public RNA sequencing and viral genome sequencing data, to better understand how SARS-CoV-2 interacts with human

Participants

This collaboration started in the context of the virtual biohackathon for COVID-19 (covid-19-bh20) which took place from April 5 to April 11 2020.

INRAE INSA INSTITUT NATIONAL DES SCIENCES APPLIQUÉES LYON UNIVERSITÉ DE LYON

M.G. Ferrarini*
R. Rebollo
D. S. Oliveira

A. Lal*

NVIDIA

Weill Cornell Medicine

V. Aguiar-Pulido#
T. Floyd

B.E. Pickett#

BRIGHAM YOUNG UNIVERSITY FOUNDED BYU 1875 PROVO, UTAH

BIG DATA INSTITUTE UNIVERSITY OF OXFORD

A. Gruber

T. Pusa

Luxembourg Centre for Systems Biomedicine

TOR VERGATA UNIVERSITÀ DEGLI STUDI DI ROMA

A. Guarracino

Amsterdam UMC

I.M. Gonzalez

A. Kanitz

SIB

*Equal contribution
Corresponding authors