Thoughts on repurposing our current bioinformatics research to address COVID-19

M Gerstein & M Rutenberg Schoenberg

Slides freely downloadable from Lectures.GersteinLab.org & "tweetable" (via @MarkGerstein). See last slide for more info.

Importance of molecular data science for COVID-19 research: "Molecule to molecule combat"

You Retweeted

William L. Jorgensen @JorgensenWL · Apr 3

Science and scientists will provide the solution to #COVID19. That ability is based on continued commitment to basic research that in at the start may seem esoteric. In the COVID aftermath there will hopefully be greater appreciation of this and wiser investment for the future.

Andrei Yudin @andrei_yudin · Apr 2

When all is said and done it will be a molecule that will stop this virus.

We'll stare at its structure, imagine how its bonds spin, and lament at why it took so long. It will be a moon landing sort of moment. Some people will get tattoos of that structure, I am telling you.

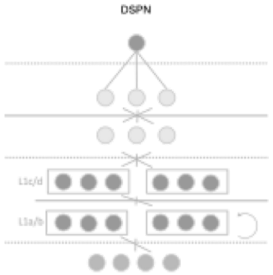
[COVID-HASTE-042120]

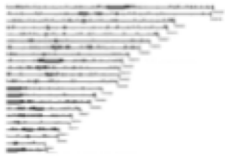
[lectures.gersteinlab.org]

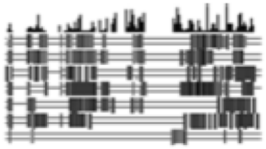
Current Lab Research Areas, which perhaps could be repurposed to address COVID-19

- Studying molecular evolution of the human genome & tumor evolution =>
  Evolutionary analysis of the SARS-CoV-2 genome
- Analysis of macromolecular structures & cryo-EM images =>
  Deep learning evaluation of variants in SARS-CoV-2 host protein structures
- Studying personal genomics & neurogenomics =>
  How human variation relates to SARS-CoV-2 host proteins & host response
- Analysis of RNA-seq, ChIP-seq, Hi-C & Single-cell seq. =>
  Studying the host response to virus with functional genomics
- Analysis of molecular networks & machine learning models for gene interactions =>
  Predicting host-virus interactions
- Studying genomic privacy & security =>
  Developing contact tracing approaches with blockchain & data sanitization

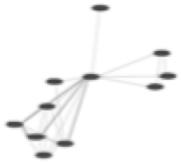
Tool development & large dataset expertise

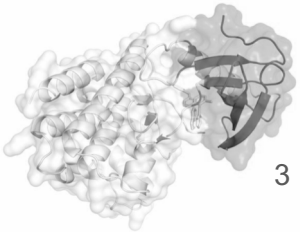






[COVID-HASTE-042120]





[lectures.gersteinlab.org]

People in the Gerstein Lab + Collaborations related to COVID-19

COVID-19 subgroup

Michael Rutenberg Schoenberg
Prashant Emani
Yucheng Yang
Shaoke Lou
Gamze Gursoy
Charlotte Brannon

(Assoc.) Research Scientists

Fabio Navarro
*Jing Zhang
Jinrui Xu
*Joel Rozowsky
Jonathan Warrell
Sushant Kumar

Postdocs

Christopher Cameron
Garrett Ash
Kun Xiong
*Leonidas Salichos
Timur Galeev
Declan Clarke
Donghoon Lee

Grad Students

Hussein Mohsen
Jiahao Gao
Tianxiao Li
William Meyerson
Xiaotong Li
*Jason J. Liu

Collaborators on COVID-19 and related research

*Yong Xiong (Yale)
*Dan Spakowicz (Ohio State)
*Bian Li (Vanderbilt)
*Geoff Chupp Lab (Yale)
*Rob Bjornson (Yale)

Consortia

ENCODE
PsychENCODE
PCAWG
ExRNA
1000 Genomes

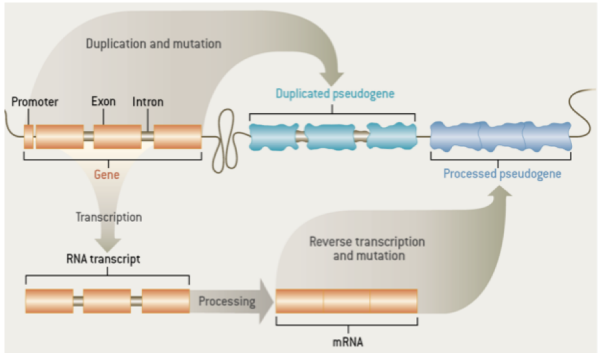
Contributed to today's presentation
* Contributed prior work relevant to COVID-19

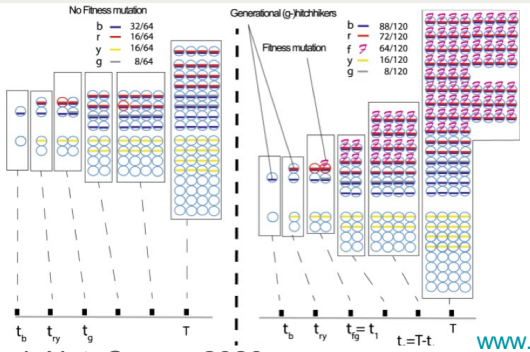
[COVID-HASTE-042120]

[lectures.gersteinlab.org]

Variation & Evolution of Coronavirus Sequences

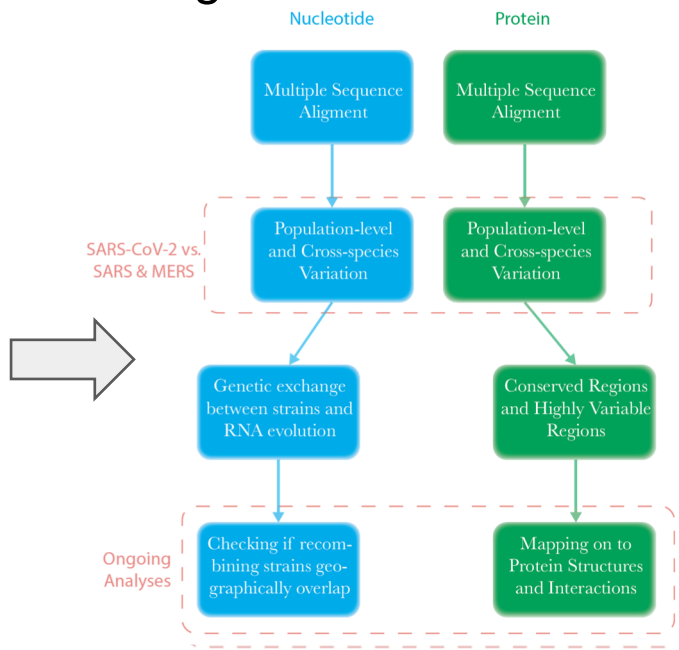
Previous work: Genome & Tumor evolution





Salichos *et al*. *Nat. Comm.* 2020
Sisu *et al. bioRXiv* 2019, Sisu *et al.* 2014 *PNAS*

Investigation of SARS-CoV-2



Generated Resources: Fasta files, VCF file
- Data Sources: GISAID, NCBI VIRUS

www.gisaid.org https://www.ncbi.nlm.nih.gov/labs/virus/vssi/#
https://www.scientificamerican.com/article/the-real-life-of-pseudoge/

Goals for RNA alignment:

a. Establish evolution through mutational "distance" - Phylogenetic Analysis
b. ***Q. Did viral strains recombine when geographically co-located?***

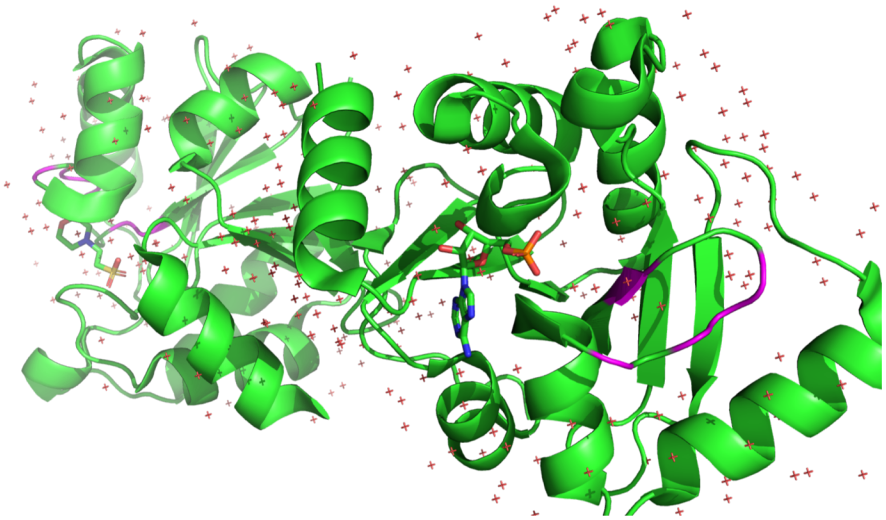
[COVID-HASTE-042120]

[lectures.gersteinlab.org]

Slide: Prashant Emani + Leonidas Salichos

Protein Variation Maps of Coronaviruses

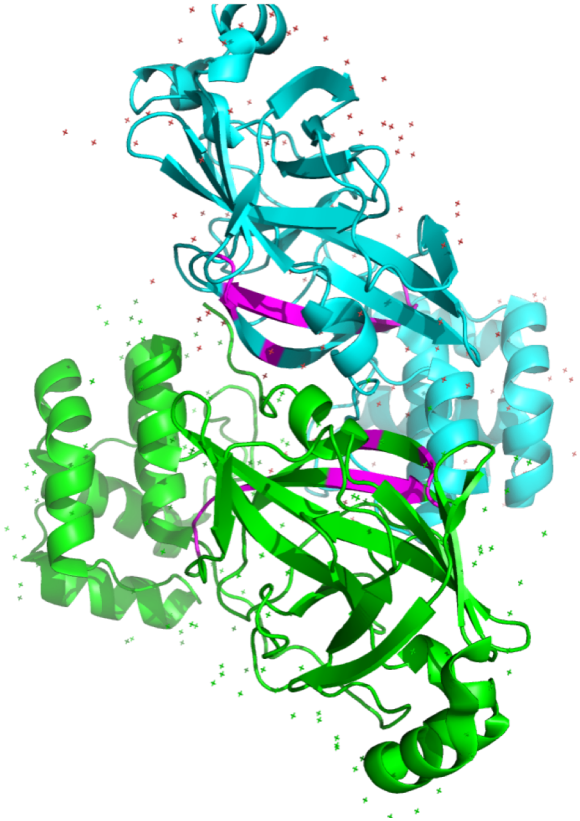
(Magenta: Variable regions)



SARS-CoV-2 ADP Phosphatase in Complex with AMP
(*PDB ID: 6W6Y*)

Key Questions

- ***Are certain regions conserved within SARS-CoV-2, or between SARS-CoV-2 and SARS and MERS?***
- ***Some regions are highly variable: how does the virus cope with frequent variation?***



SARS 3C-like Protease (in dimeric form: green and turquoise monomers) (*PDB ID: 4HI3*)

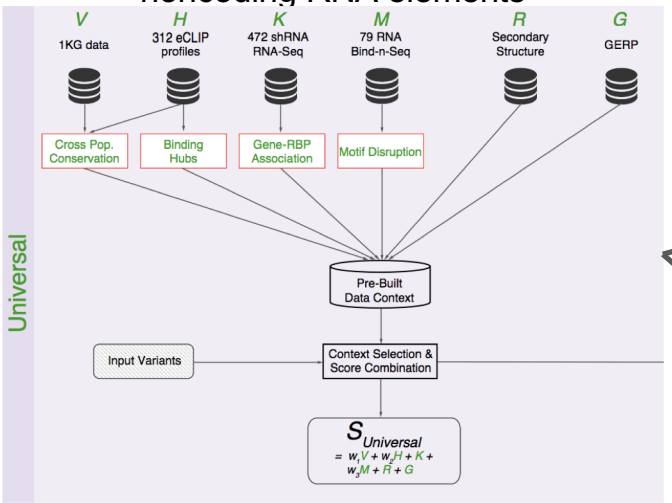
Slide: Prashant Emani

[COVID-HASTE-042120]

[lectures.gersteinlab.org]

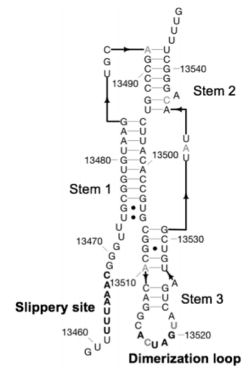
Interpreting variants in SARS-CoV-2 for their non-coding RNA properties

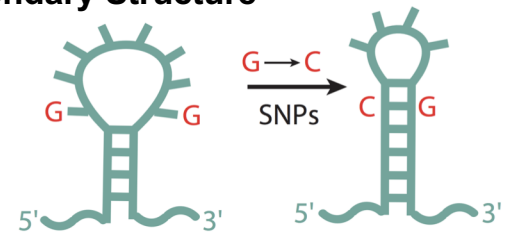
RADAR: Tool to interpret variants in noncoding RNA elements



Zhang J, Liu JL *et al bioRXiV* 2019 (GenomeBiology, 2020, in press)

RNA Secondary Structure





Apply to conserved RNA structure elements in SARS-CoV-2 + search for novel elements

[COVID-HASTE-042120]

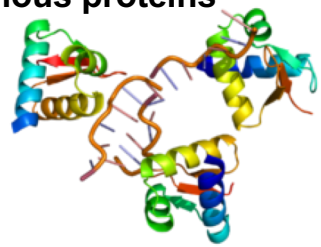
Frameshifting Element

Rangan *et al.* 2020 *bioRXiV*

Predicted binding to endogenous proteins

motif-breaking or gain

CNN

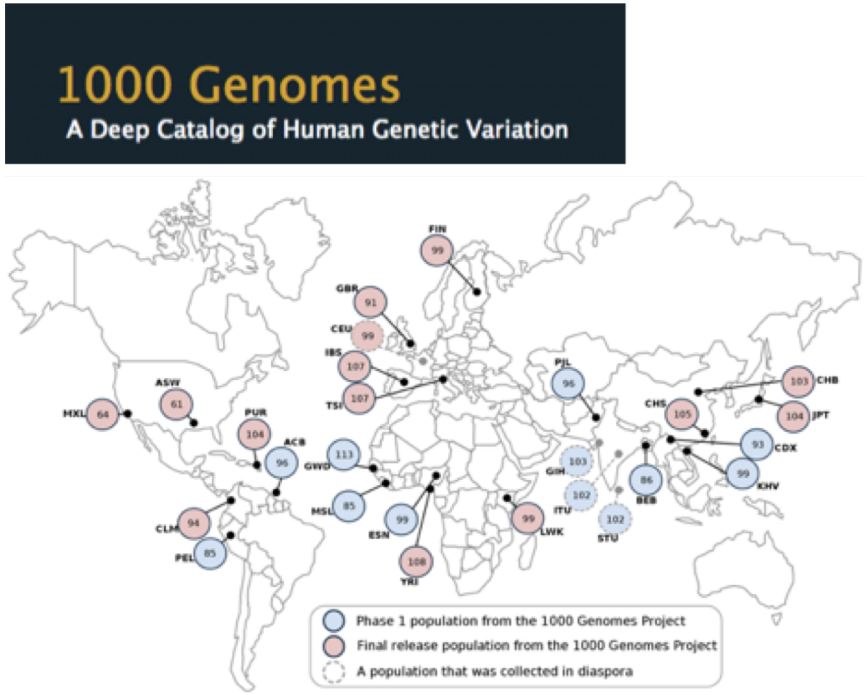


gain/loss of RNA binding

[lectures.gersteinlab.org]

Slide: Michael Rutenberg Schoenberg

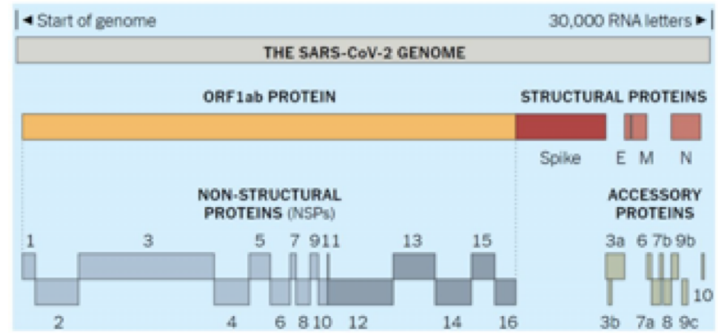
How human population variation relates to cryo-EM structures of host proteins

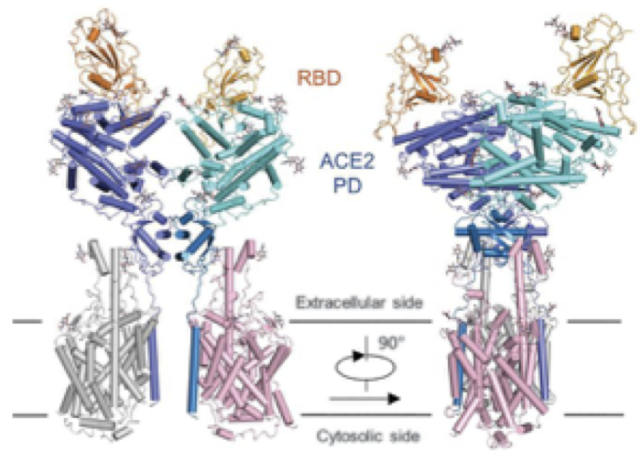


1000 Genome Project Consortium, Nature ('15); Abyzov et al, Nat Commun ('15); Chen et al, Nat Commun ('17); Balasubramanian et al, Nat Commun ('17)

Yan et al. 2020 Science.

SARS-CoV-2 encodes 29 proteins



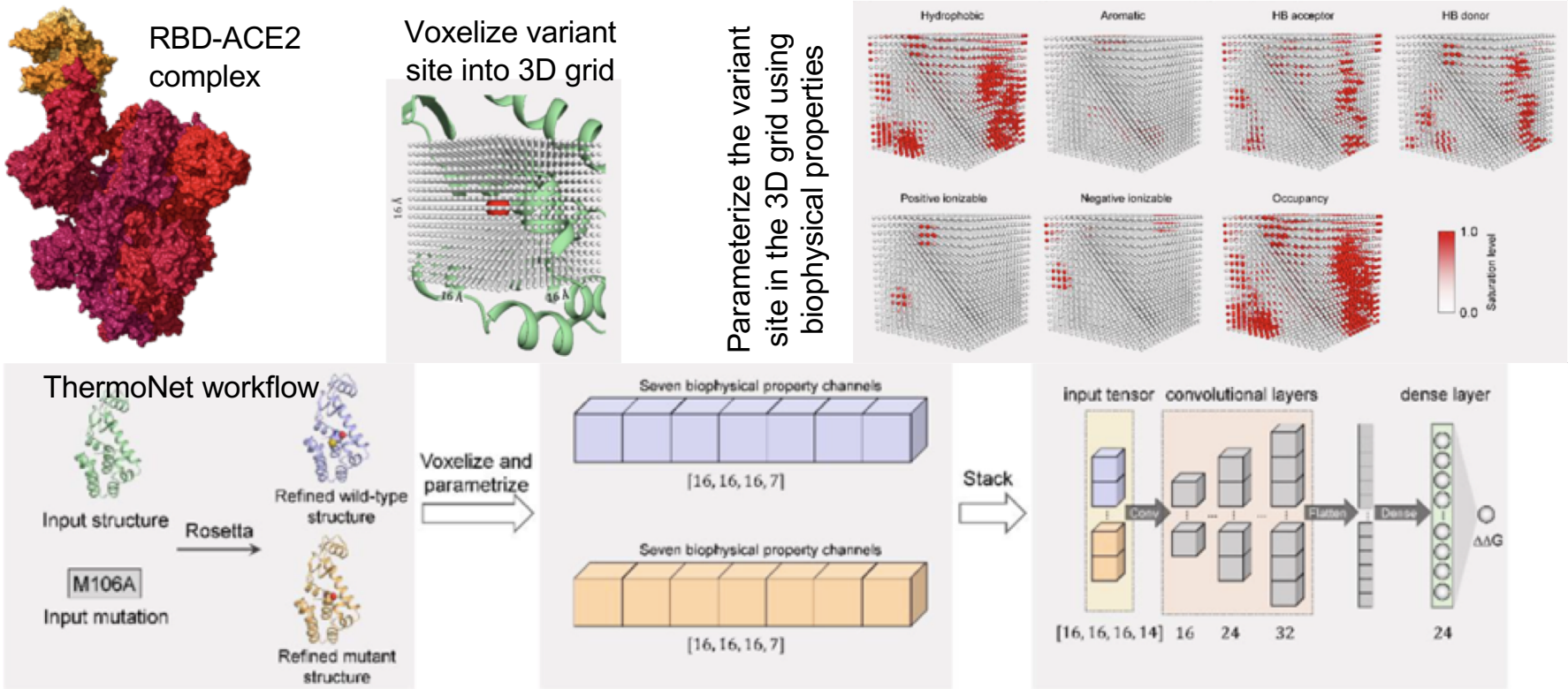


[COVID-HASTE-042120]

[lectures.gersteinlab.org]

Slide: Yucheng Yang

Deep learning model to predict how variants impact the stability of host proteins



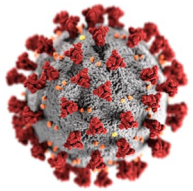
[COVID-HASTE-042120]

[lectures.gersteinlab.org]

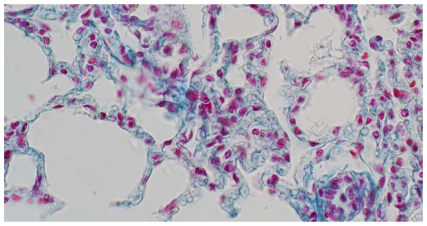
Bian Li *et al. bioRXiV 2020 (PLoS CB under review)*

Slide: Yucheng Yang

Repurposing existing RNA-seq tools for studying host/virus interactions

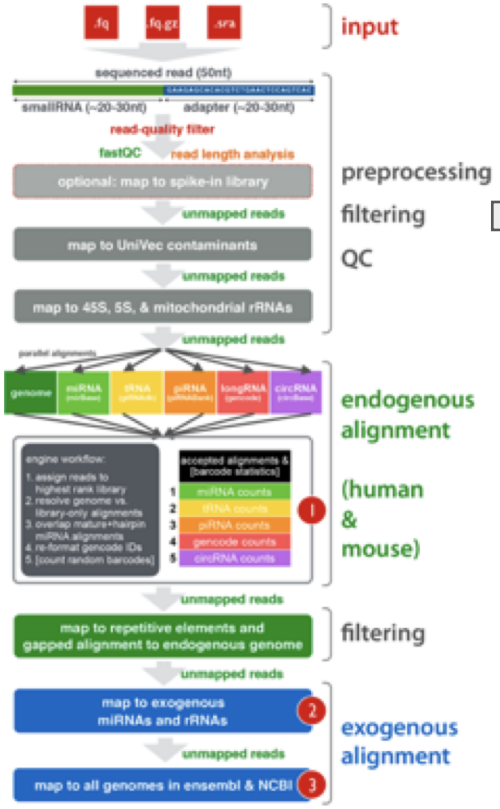


+

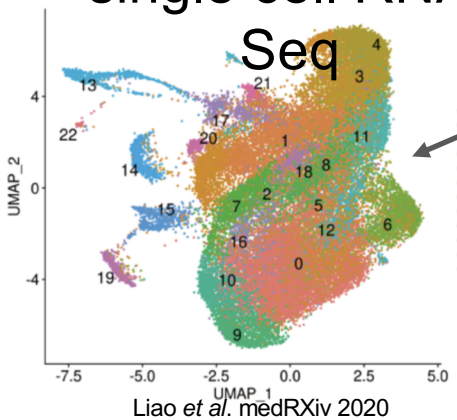


COVID-19 patient sample

exceRpt pipeline



single cell RNA-Seq



Liao *et al*. medRXiv 2020

Which cells have SARS-COV-2?

* Investigate connections between virus + host + other microbes

* exceRpt is from the exRNA consortium

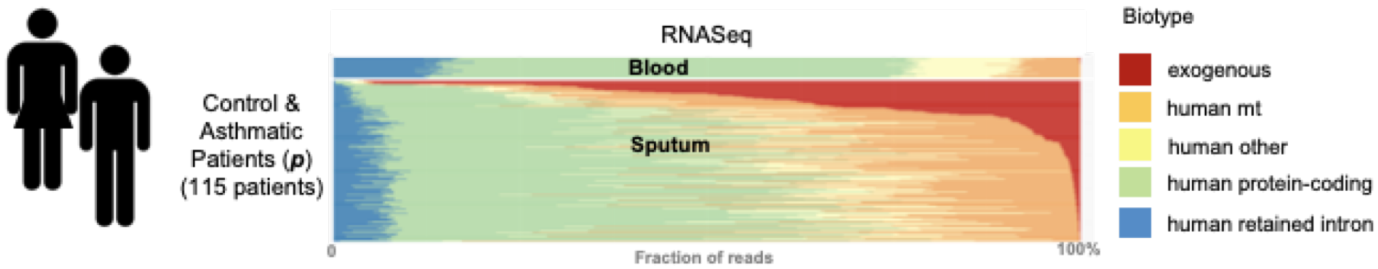
Rozowsky et al. *Cell Systems* 2019, Daniel Spakowicz, Shaoke Lou *et al.* bioRXiV 2019 (*Genome Biology* 2020 *in press)*

[COVID-HASTE-042120]

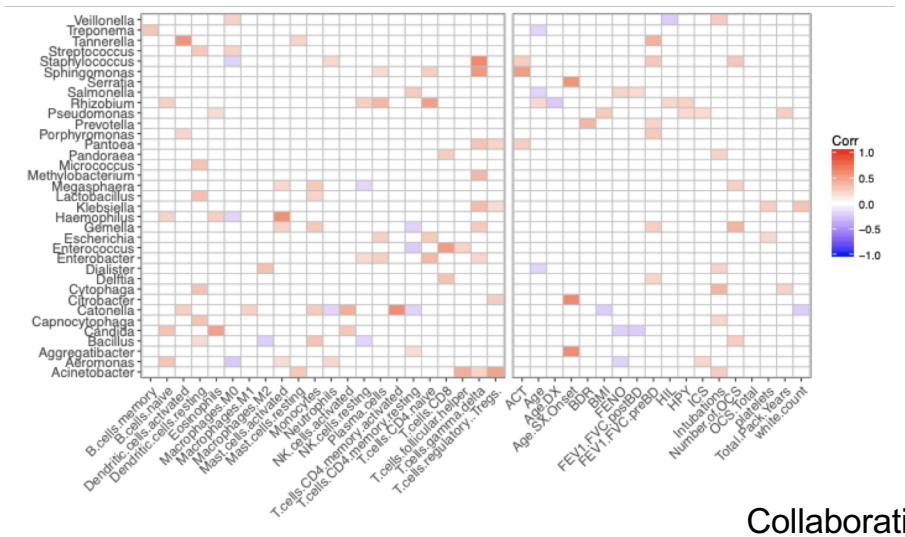
[lectures.gersteinlab.org]

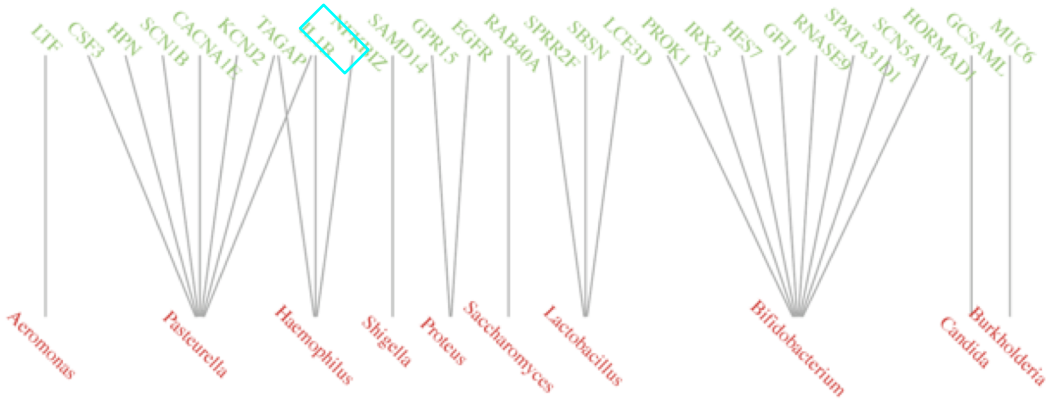
Slide: Michael Rutenberg Schoenberg

Adapting a machine learning framework for linking microbes to genes in asthma to find SARS-CoV-2 - human gene linkages



Use ExceRpt to discover Exogenous RNAs





Microbe-gene links identified by LDA-Link framework

[COVID-HASTE-042120]

[lectures.gersteinlab.org]

Collaboration with Prof. Geoffrey Chupp

Daniel Spakowicz, Shaoke Lou *et al.* bioRXiV 2019 (*Genome Biology* 2020 *in press)*

Slide: Shaoke Lou

Adapting work on biomedical data privacy & security for COVID-19 contact tracing

One of the most effective ways of stopping the spread is via “contact tracing”, but may not work under certain privacy laws

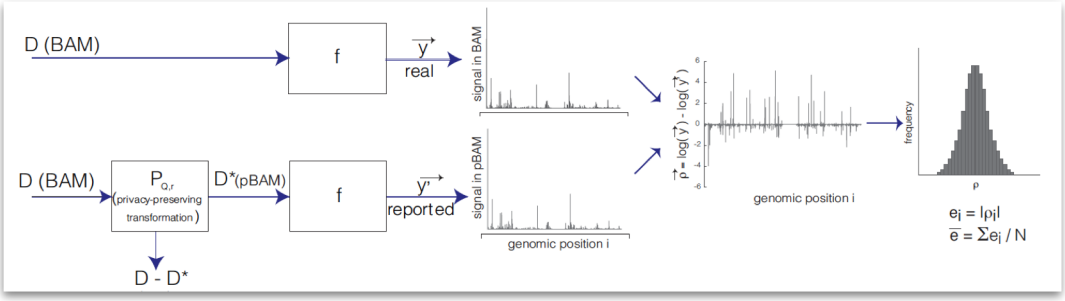
NEWS · 18 MARCH 2020

South Korea is reporting intimate details of COVID-19 cases: has it helped?

Extensive contact tracing has slowed viral spread, but some say publicizing people’s movements raises privacy concerns.

nature

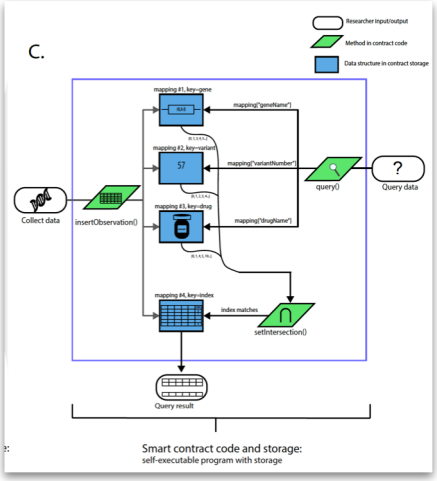
Data **Sanitization** Techniques for Genome Privacy



Decentralized

Blockchain

solutions for robust storage & query

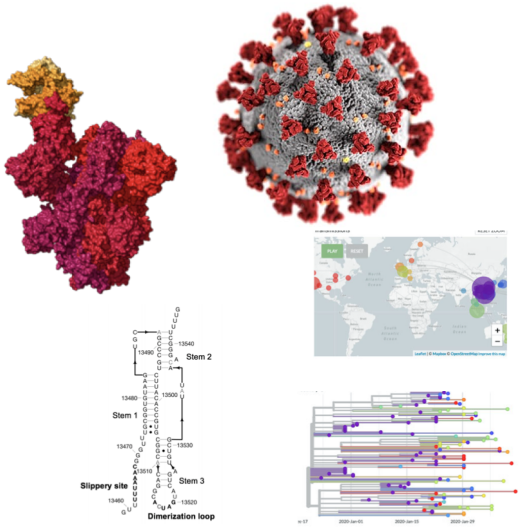


G Gursoy, P Emani, O Jolanki, C Brannon, A Harmanci, S Strattan, A Miranker, M Gerstein, Biorxiv, doi: https://doi.org/10.1101/345074
G Gursoy, R Bjornson, M Green and M Gerstein, BMC Medical Genomics
G Gursoy*, C Brannon*, and M Gerstein, BiorXiv, doi: https://doi.org/10.1101/2019.12.16.878488
G Gursoy, C Brannon, S Wagner, and M Gerstein, BiorXiv, doi: https://doi.org/10.1101/2020.03.03.975334
C Brannon and G Gursoy, https://thegccontent.wordpress.com/2020/04/13/a-practical-ethereum-and-multichain-blockchain-tutorial/

Slide: Gamze Gursoy and Charlotte Brannon

Connect with us

COVID-19 Data



Data analysis + Integration

Biological Expertise

Analysis ideas

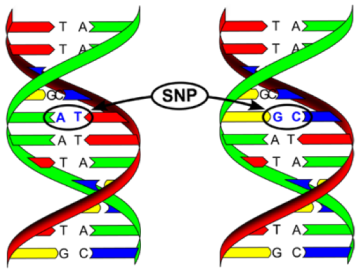
michael.rutenbergschoenberg@yale.edu &
mark@gersteinlab.org

Photo credits:
https://pixabay.com/vectors/cartoon-icon-light-bulb-symbol-1294877/
https://www.massgeneral.org/news/coronavirus/treatment-guidance
https://gisaid.org

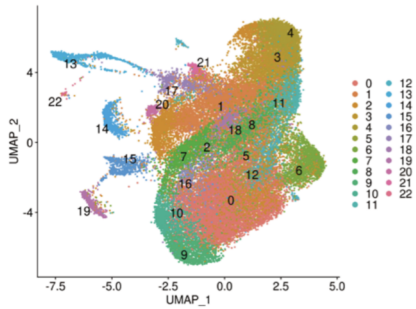
[COVID-HASTE-042120]

[lectures.gersteinlab.org]

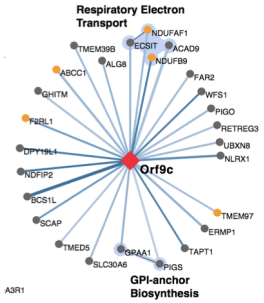
Data we're particularly interested in



Genomes



Functional Genomics:
Single-cell, RNA-Seq, ChIP-Seq, Hi-C, Microbiome, etc.



Proteomics + Protein-protein interactions

Cryo-EM & X-ray Structures

Biosensors

#coronavirus
Twitter mining

Epidemiology/ EHR records

Matched Host + Virus Samples

Liao *et al*. medRXiv 2020 https://twitter.com/who/status/1236281806661586946 https://www.gisaid.org/ Gordon *et al. bioRXiV* 2020

[COVID-HASTE-042120]

[lectures.gersteinlab.org]