Computational approaches for guiding rational vaccine design:

Case studies in HCV, HIV, and COVID-19

Matthew R. McKay

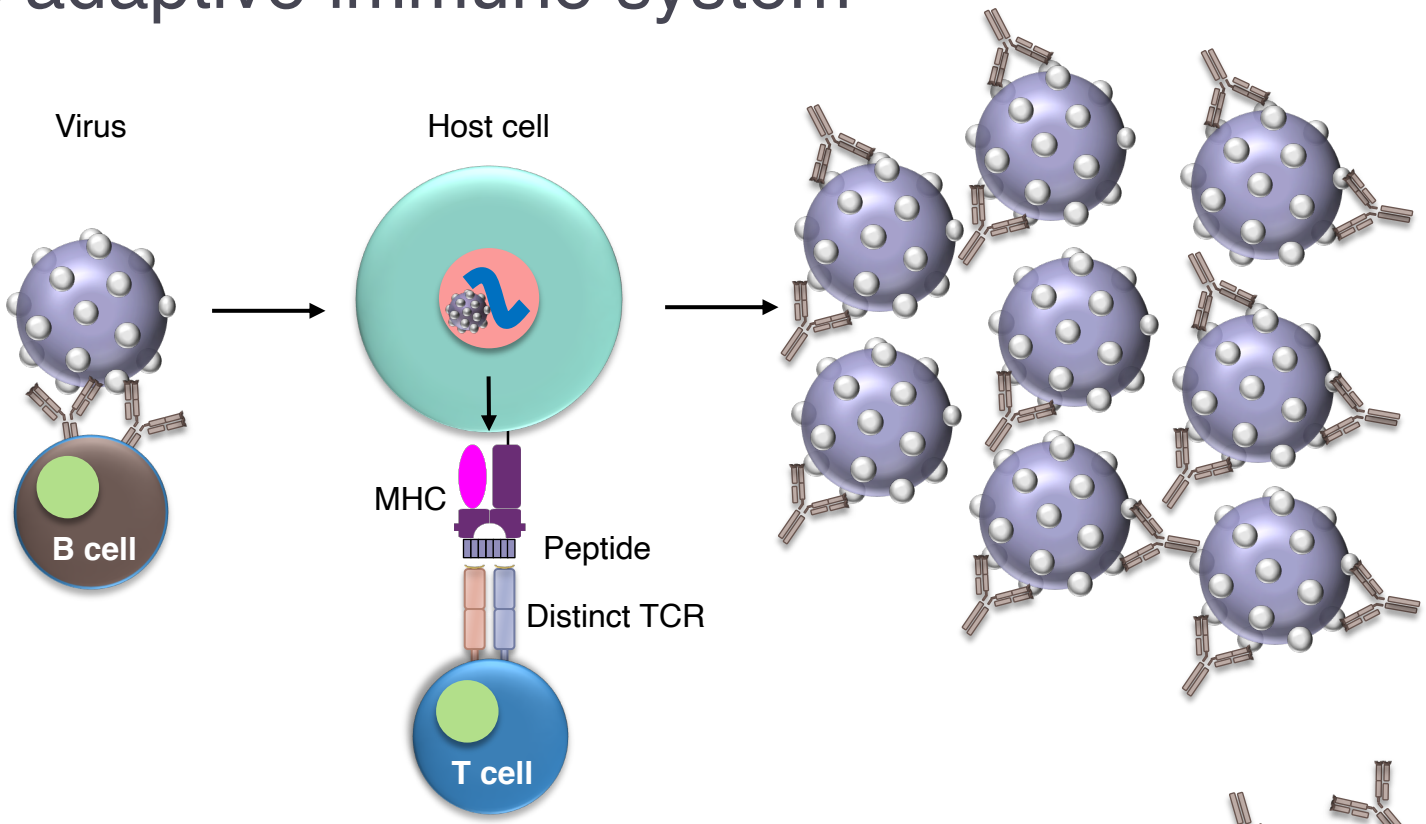
Department of Electronic and Computer Engineering

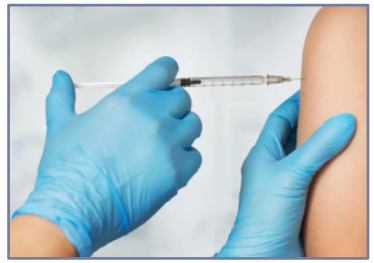
Department of Chemical and Biological Engineering

June 12, 2020, One World Signal Processing Seminar Series

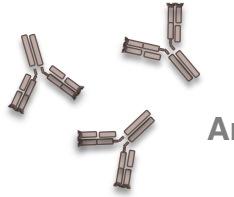
香港科技大學
THE HONG KONG
UNIVERSITY OF SCIENCE
AND TECHNOLOGY

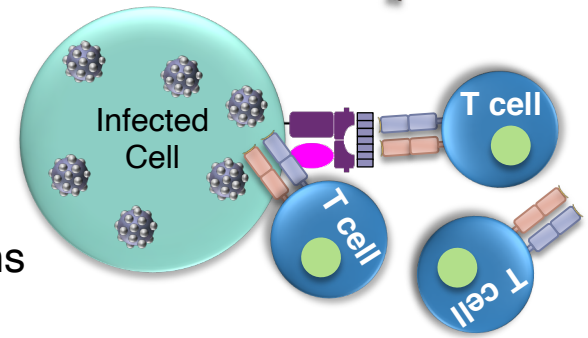
Pathogen specific adaptive immune system





Basis for vaccination





Memory of past infections

Vaccination

- Eradication or near-eradication of diseases such as smallpox and polio
- Still no effective vaccine against many pathogens
- Main focus of today's talk: **Hepatitis C virus (HCV), SARS-CoV-2, and a bit about HIV**

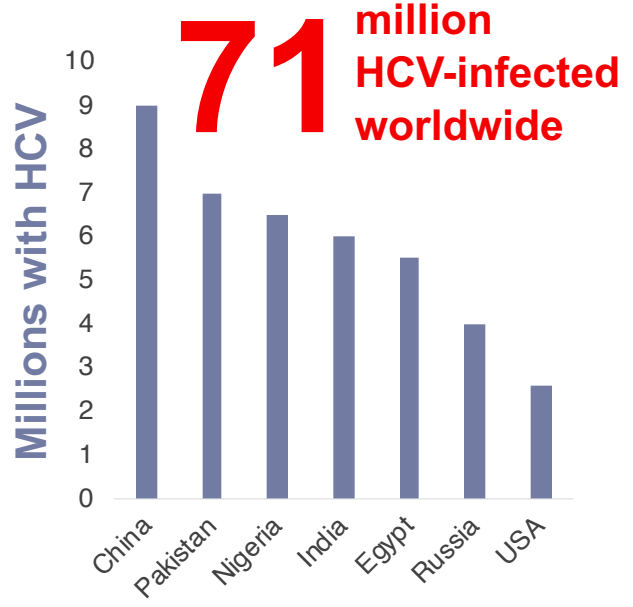
…. and …

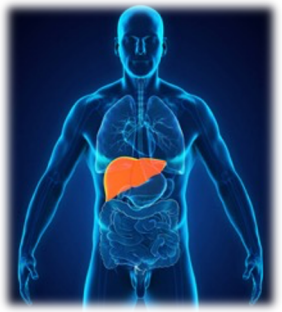
How data analytics, modelling, and statistical inference can help?

PART 1: Identifying escape-resistant antibodies for guiding HCV vaccine design

Hepatitis C virus

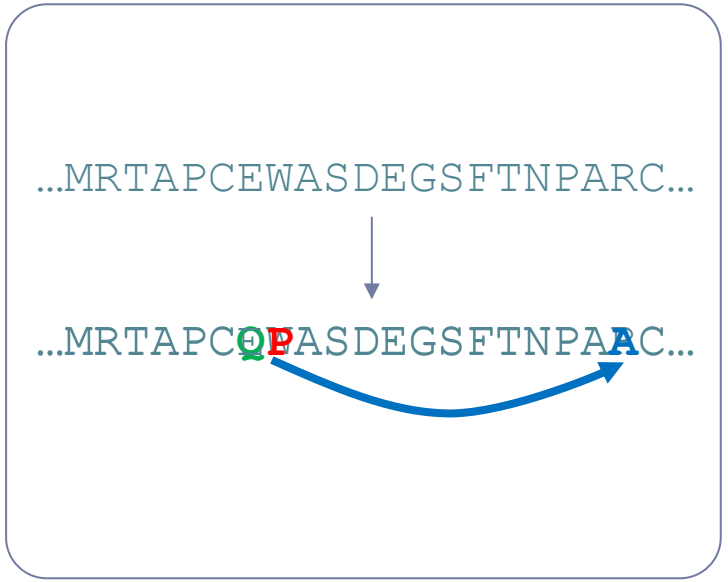
- Global public health problem
- Around 20–30% of infections are asymptomatic and resolve within 6 months
- A leading cause of liver transplants and liver cancer
- Key challenges in HCV vaccine development:
  - High replication rate (~10^{12} copies per day)
  - High mutation rate (~10^{-4} mutations/nucleotide/replication cycle)
- Effective new drugs available, but problem still not fully solved
- Widespread vaccination would play a key role in eradicating HCV

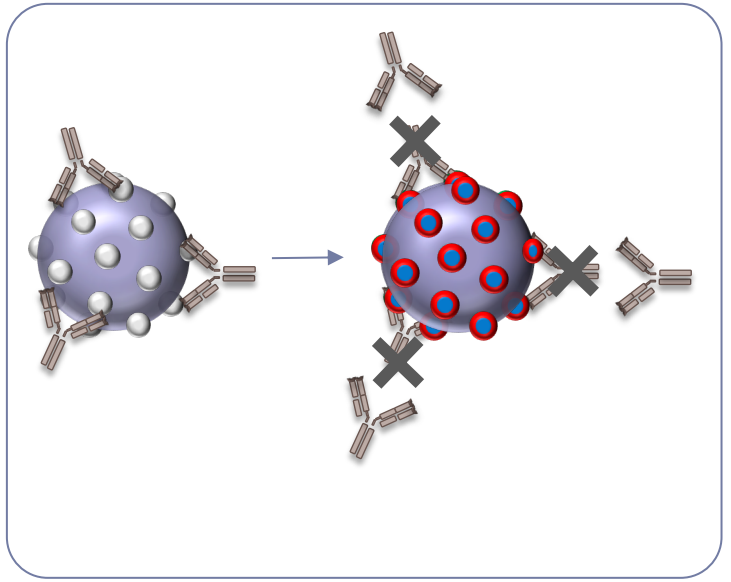


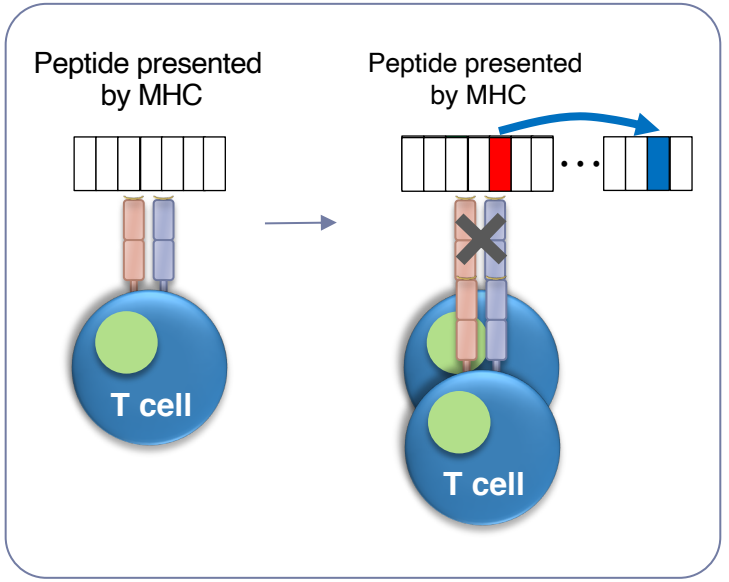


Immune system evasion by HCV

Key Point: Mutations generally do **not** act independently







HCV mutation during replication

High mutation rate of HCV results in immune escape if the mutant virus has a **high fitness**

Additional complication

Compensation of **deleterious** effect of individual mutations

Problem statement

- **E2-specific broadly neutralizing human monoclonal antibodies (HmAbs)** have been identified for HCV
  - Spontaneous clearance associated with their early appearance ✓
  - Escape mutations have been observed experimentally ✗

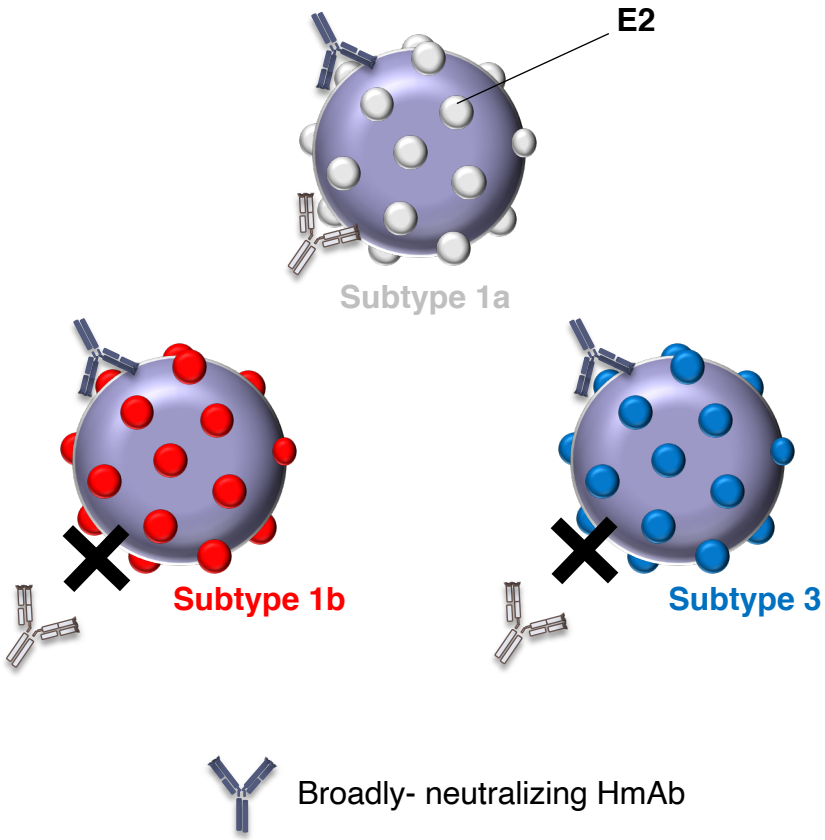
Key open question:

How “broadly-neutralizing” are the identified HmAbs?

Which ones are the most difficult to escape?

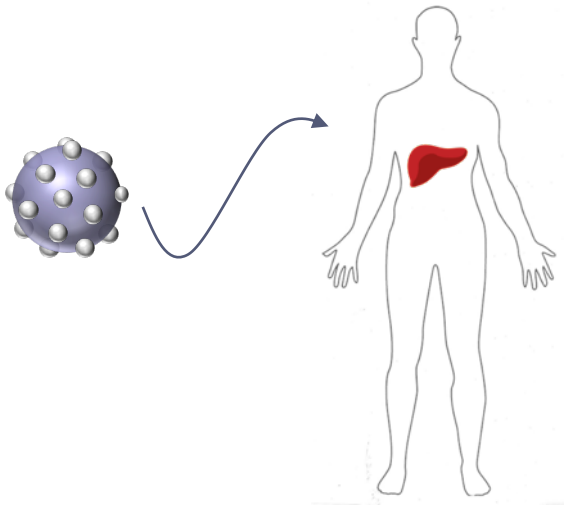
Proposed strategy:

Use **sequence data of E2, statistical modeling and inference** to try to identify escape-resistant HmAbs that can aid HCV vaccine development



Quadeer, Ahmed A. et al. “Identifying immunologically-vulnerable regions of the HCV E2 glycoprotein and broadly neutralizing antibodies that target them”, *Nature Communications* 10 **(2019)**

Within-host viral evolution model



…ASDEGSFTNPARC…
…ASDEGSFTNPARC…
…ASDEGSFTNPARC…
…ASDEGSFTNPARC…
…ASDEGSFTNPARC…
…ASDEGSFTNPARC…
…ASDEGSFTNPARC…
…ASDEGSFTNPARC…
…ASDEGSFTNPARC…
…ASDEGSFTNPARC…
…ASDEGSFTNPARC…
…ASDEGSFTNPARC…
…ASDEGSFTNPARC…
…ASDEGSFTNPARC…

Mutation

…ASHEGSFTNPARC…
…ASDEGSFTNPARC…
…ASDEGSFTNPARC…
…ASDEGSFTNPARC…
…ASDEGSFTNPARC…
…ASDEGSFTNPARC…
…ASDEGSFTNPARC…
…ASDEGSFTNPARC…
…ASDEGSFTNPARC…
…ASDEGSFTNPARC…
…ASDEGSFTNPARC…
…ASDEGSFTNPARC…
…ASDEGSFTNVARC…
…ASDEGSFTNPARC…

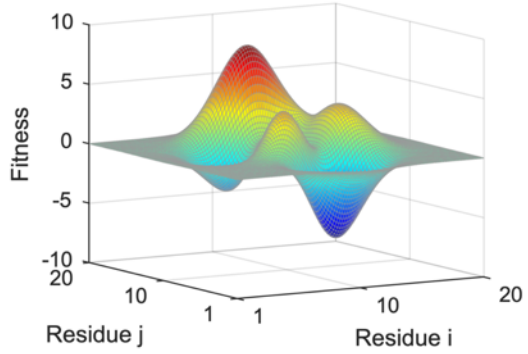
Selection

…ASHEGSFTNPARC…
…ASHEGSFTNPARC…
…ASHEGSFTNPARC…
…ASHEGSFTNPARC…
…ASHEGSFTNPARC…
…ASHEGSFTNPARC…
…ASHEGSFTNPARC…
…ASHEGSFTNPARC…
…ASHEGSFTNPARC…
…ASDEGSFTNPARC…
…ASDEGSFTNPARC…
…ASDEGSFTNPARC…
…ASDEGSFTNPARC…
…ASDEGSFTNVARC…

High fitness
Low fitness

Fitness landscape required

…ASHEGSFTNPARC…
…ASDEGSFTNPARC…
…ASDEGSFTNPARC…
…ASDEGSFTNPARC…
…ASDEGSFTNPARC…
…ASDEGSFTNPARC…
…ASDEGSFTNPARC…
…ASDEGSFTNPARC…
…ASDEGSFTNPARC…
…ASDEGSFTNPARC…
…ASDEGSFTNVARC…



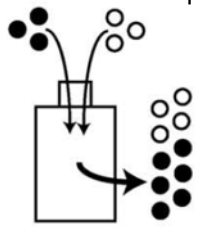
E2 protein sequences

Fitness of each sequence

Experimental approach

Sequence to be measured

Reference sequence



Alternative Solution:

Data-driven computational approach

E2 fitness landscape inference – Unsupervised ML approach

Available E2 data

...ASHEGSFTNPARC...
...ASDEGSFTNPARC...
...ASDEGSFTNPARC...
...ASDEGSFTNPARC...
...ASDEGSFTNPARC...
...ASDEGSFTNPARC...
...ASDEGSFTNPARC...
...ASDEGSFTNPARC...
...ASDEGSFTNPARC...
...ASDEGSFTNVARC...

3,363 sequences of
363 residues long E2 protein extracted
from **1,298** HCV-infected individuals
(subtype 1a)

Statistical model:
Maximum entropy (prevalence) model

Fitness Prevalence

$$f(\mathbf{x}) \sim p(\mathbf{x}) = \frac{\exp[-E(\mathbf{x})]}{Z}$$

$$E(\mathbf{x}) = \sum_{i=1}^{L} h_i(x_i) + \sum_{i=1}^{L} \sum_{j=i+1}^{L} J_{ij}(x_i, x_j), \quad Z = \sum_{\mathbf{x}'} \exp[-E(\mathbf{x}')]$$

Maximum entropy formulation

$$\max S = -\sum_{\mathbf{x}} p(\mathbf{x}) \log p(\mathbf{x})$$

$$\text{s.t.} \sum_{\mathbf{x}} p(\mathbf{x}) = 1$$

$$\sum_{\mathbf{x}} p(\mathbf{x}) \delta(x_i, a) = p_i^{obs}(a)$$

$$\sum_{\mathbf{x}} p(\mathbf{x}) \delta(x_i, a) \delta(x_i, b) = p_{ij}^{obs}(a, b)$$

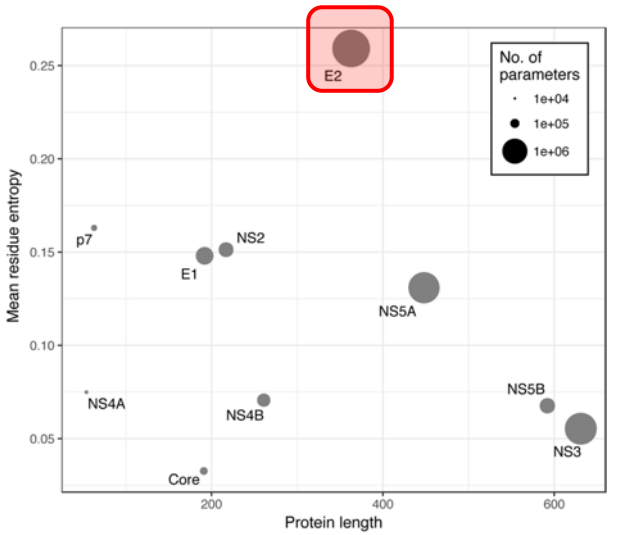
...ASDEGSFT...
...ASDEGSFT...
...ASDEGSFT...
...ASDEGSFT...
...ASDEGSFT...

...ASDEGSFT...
...ASDEGSFT...
...ASDEGSFT...
...ASDEGSFT...
...ASDEGSFT...

Solution:

$$(\mathbf{h}^*, \mathbf{J}^*) = \arg\min_{\mathbf{h},\mathbf{J}} \mathrm{KL}\,(p_0 \,||\, p_{\mathbf{h},\mathbf{J}})$$

Challenge:
Huge number of parameters

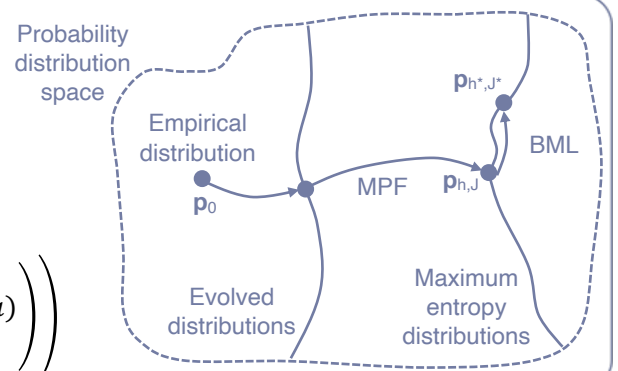


HCV sequence data → Maximum entropy model → Predicted fitness

Framework for model inference

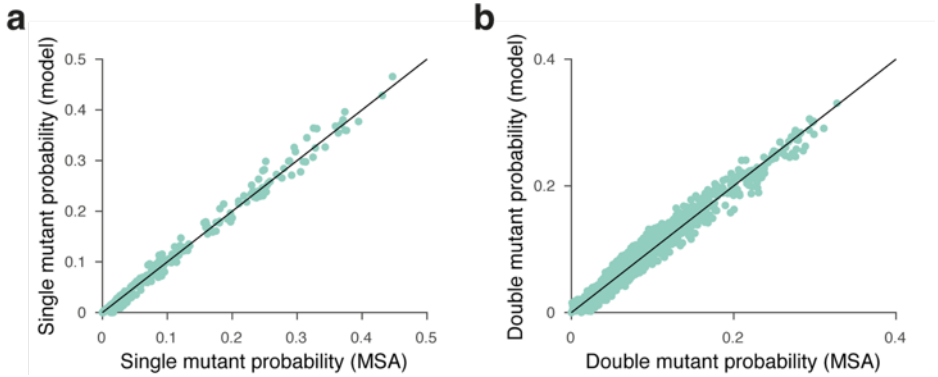
$$(\mathbf{h}_{\mathrm{MPF}}, \mathbf{J}_{\mathrm{MPF}}) = \arg\min_{\mathrm{h,J}} \mathrm{KL}(p_0 || p_{\mathrm{h,J;t}}) + \lambda_1 |\mathbf{J}|_1 + \lambda_2 |\mathbf{J}|_2$$

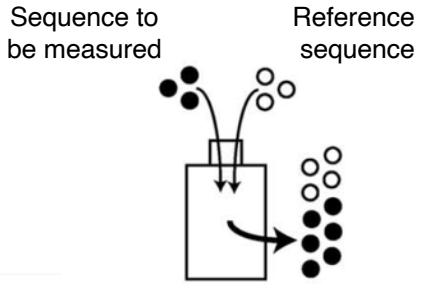
$$\mathrm{KL}(p_0 || p_{\mathrm{h,J;t}}) \approx \sum_{b,i,a} \exp\left(\frac{1}{2}\left(\left(2y_{b,(i-1)L+a} - 1\right)\sum_{j,b} y_{b,(j-1)L+b} J_{ij}(a,b) - h_i(a)\right)\right)$$



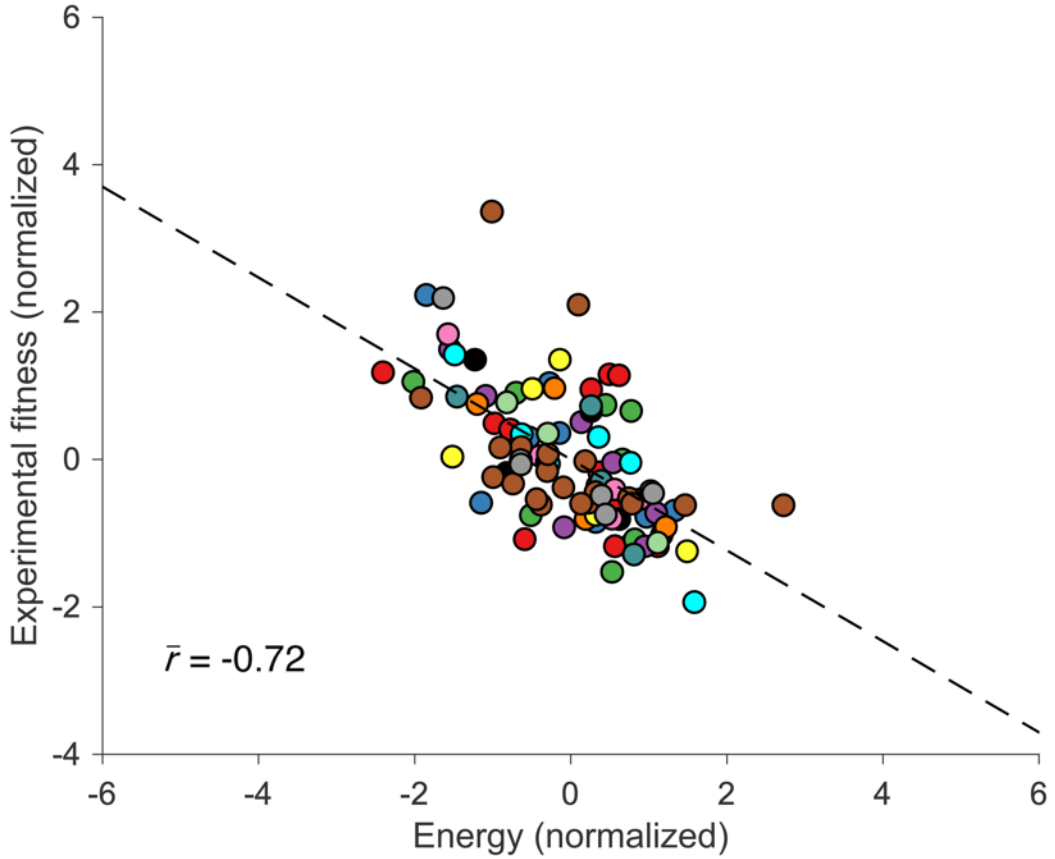
E2 fitness landscape validation

Statistical validation

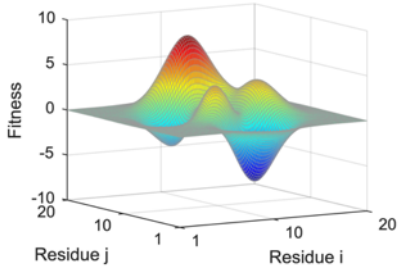




Biological validation



Within-host viral evolution model



Inferred fitness landscape

Antibody pressure at site i

T/F strain

…ASDEGSFTNPARC…

Generation 1

Mutation

Selection

Repeat until **"escape":** $f_i > 0.5$

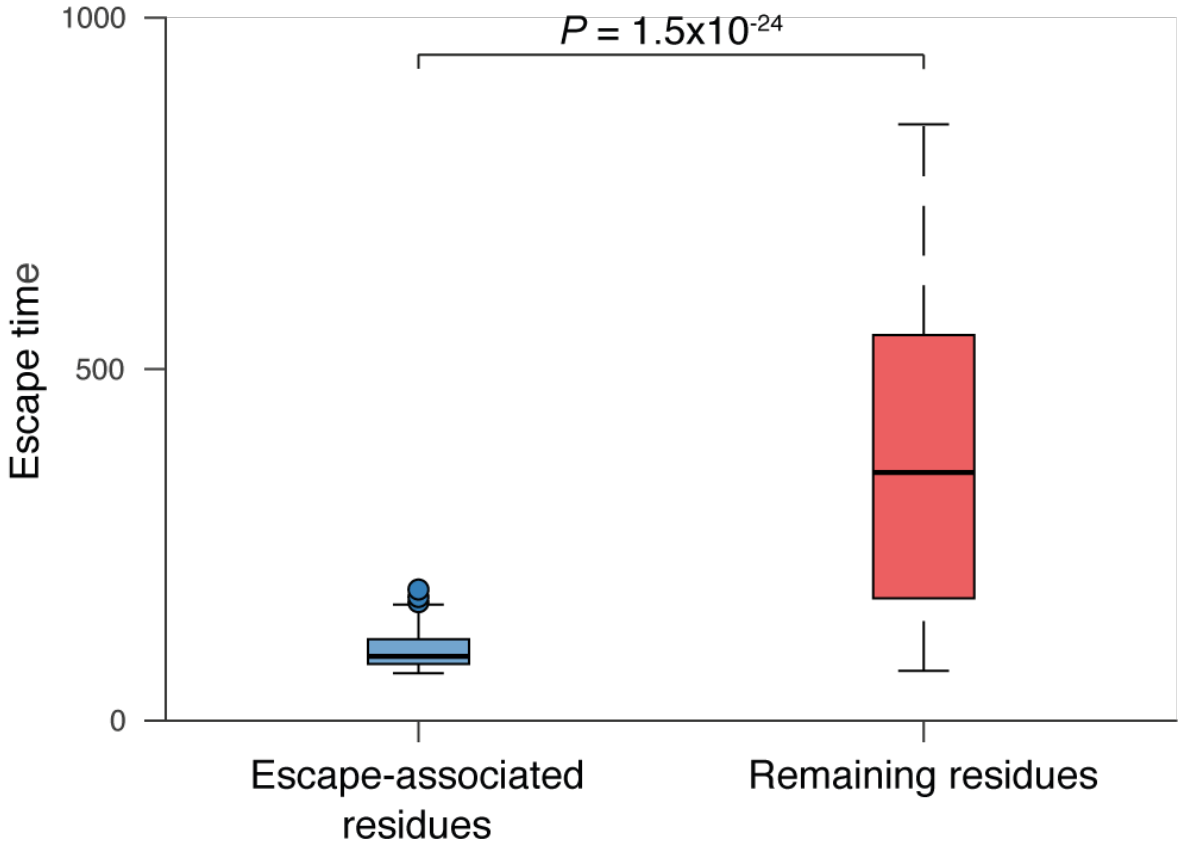
363 residues in E2
50 T/F strains per residue
100 iterations per T/F strain per residue

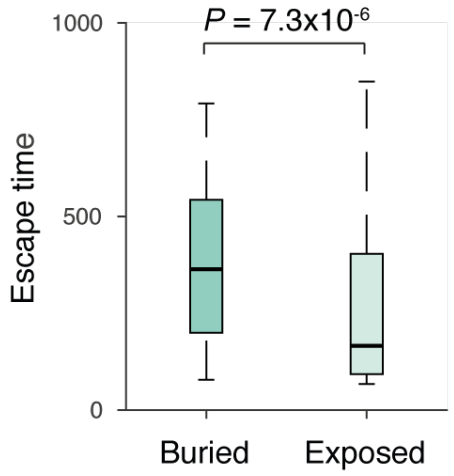
125 nodes

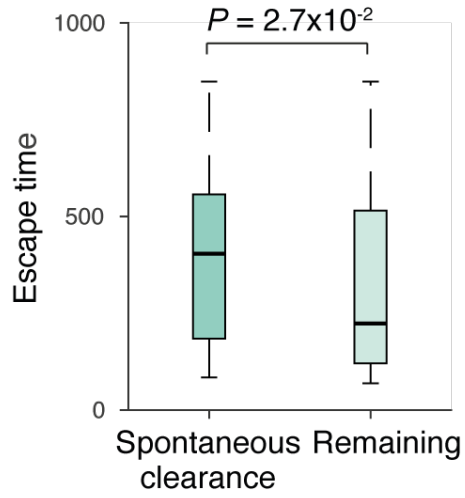
24 cores (2.5 GHz)

Mean escape time per residue

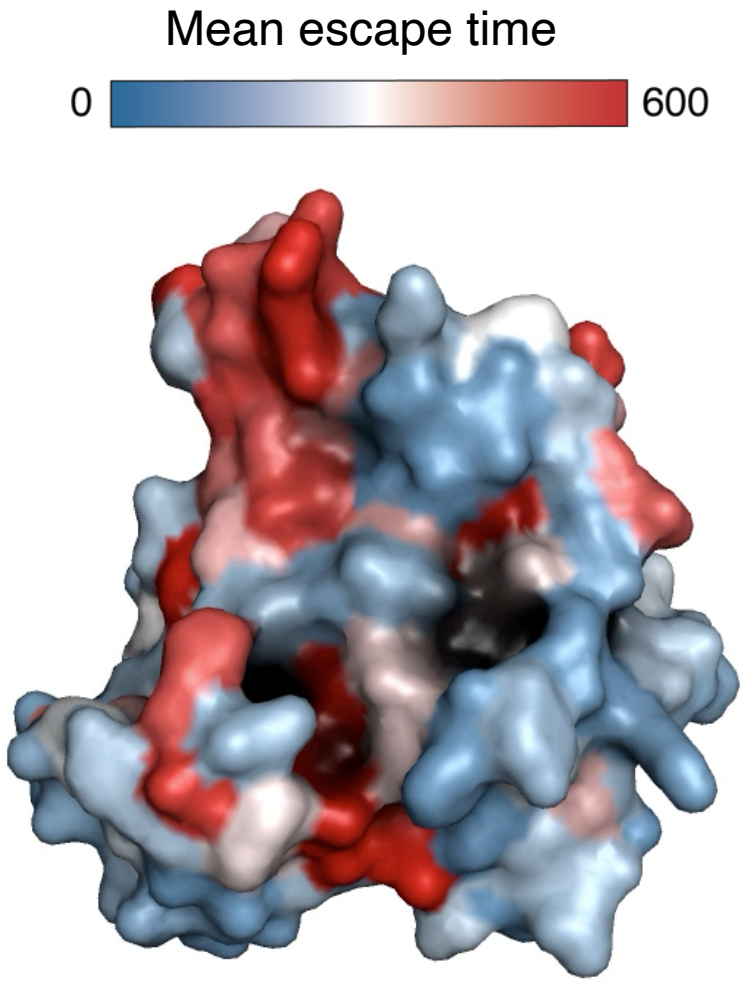
Validation against experimental/clinical data





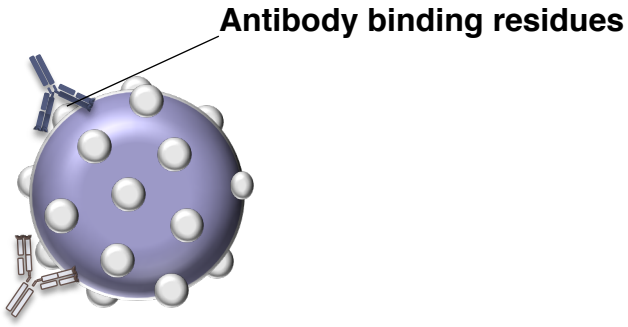


Mapping the predicted escape times on HCV E2 structure

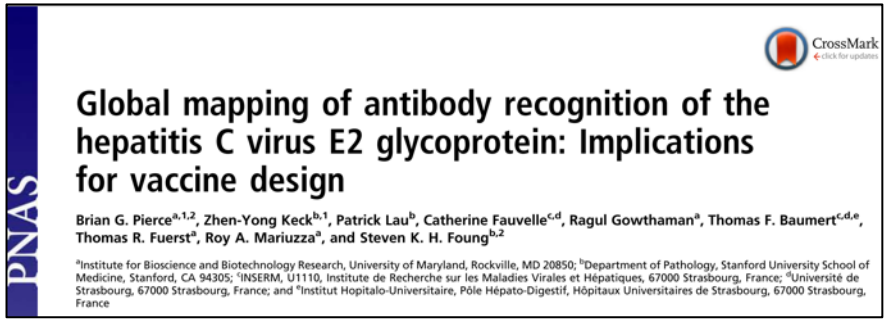


Protein data bank, https://www.rcsb.org/ (PDB ID: **4MWF)**

Escape resistance of HmAbs



Antibody binding residues obtained from the recent extensive study by Pierce et al., 2016



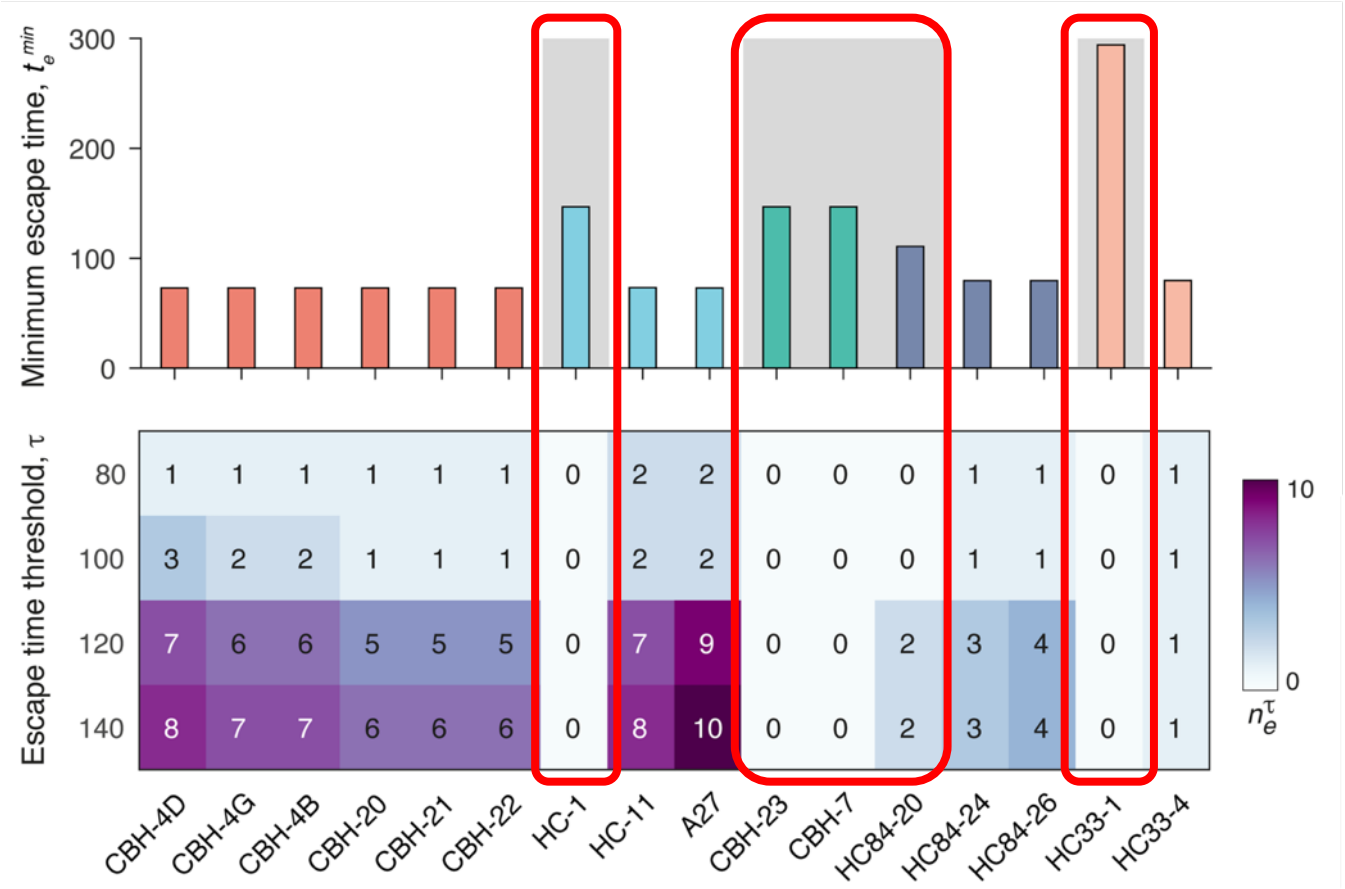
PNAS

CrossMark

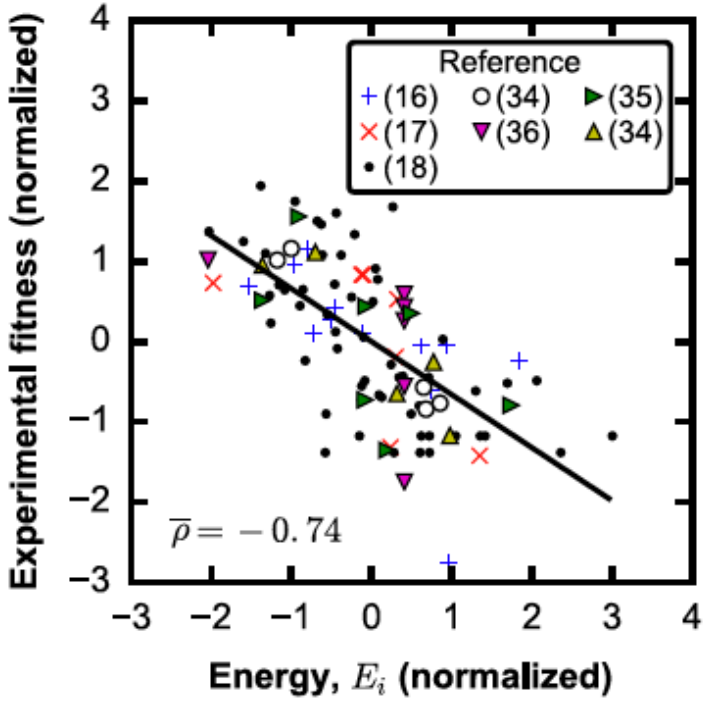
Global mapping of antibody recognition of the hepatitis C virus E2 glycoprotein: Implications for vaccine design

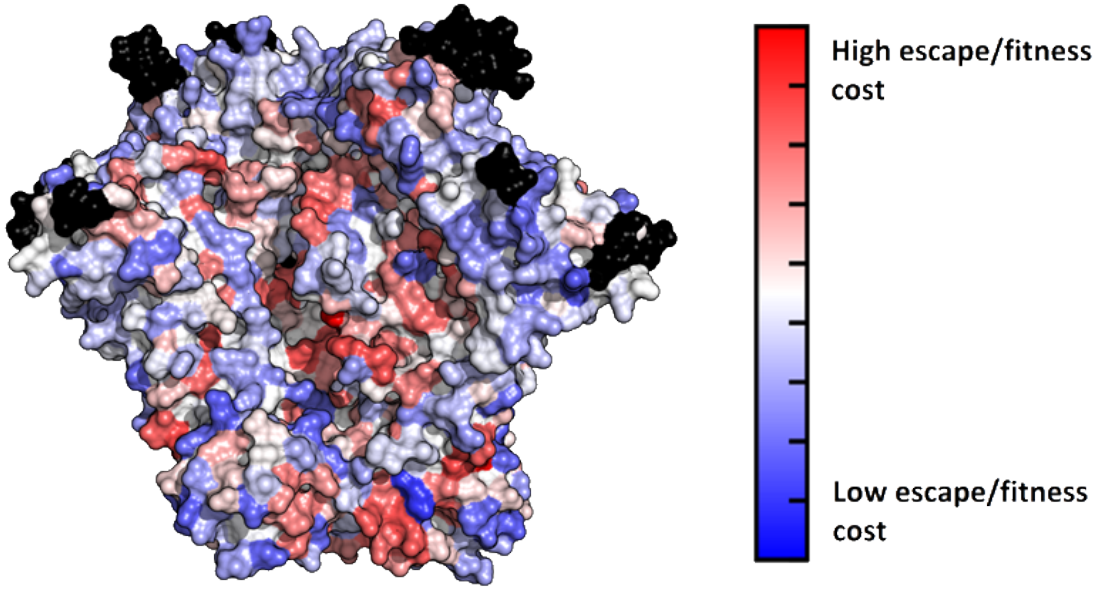
Brian G. Pierce[a,1,2], Zhen-Yong Keck[b,1], Patrick Lau[b], Catherine Fauvelle[c,d], Ragul Gowthaman[a], Thomas F. Baumert[c,d,e], Thomas R. Fuerst[a], Roy A. Mariuzza[a], and Steven K. H. Foung[b,2]

[a]Institute for Bioscience and Biotechnology Research, University of Maryland, Rockville, MD 20850; [b]Department of Pathology, Stanford University School of Medicine, Stanford, CA 94305; [c]INSERM, U1110, Institute de Recherche sur les Maladies Virales et Hépatiques, 67000 Strasbourg, France; [d]Université de Strasbourg, 67000 Strasbourg, France; and [e]Institut Hopitalo-Universitaire, Pôle Hépato-Digestif, Hôpitaux Universitaires de Strasbourg, 67000 Strasbourg, France



HIV gp160 (envelope) fitness landscape





Louie, Raymond H. Y., et al. “Fitness landscape of the human immunodeficiency virus envelope protein that is targeted by antibodies ”, *Proceedings of the National Academy of Sciences of the USA (PNAS),* 115 **(2018)**.

Part 1: Summary

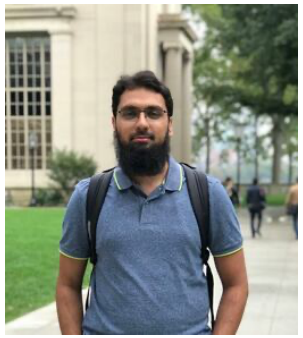
nature COMMUNICATIONS

ARTICLE

https://doi.org/10.1038/s41467-019-09819-1 OPEN

Identifying immunologically-vulnerable regions of the HCV E2 glycoprotein and broadly neutralizing antibodies that target them

Ahmed A. Quadeer [1], Raymond H.Y. Louie [1,2,3,4] & Matthew R. McKay [1,5]



Ahmed

Ray (UNSW)

PART 2:
Finding vaccine targets for COVID-19

SARS-CoV-2 and vaccine design

- Comparison of recent coronavirus infections in humans
  - SARS-CoV (2003 – 2004)[1]
    Infections: 8,098; Deaths: 774; Case-fatality rate: 15%
  - MERS-CoV (2012 –)[2]
    Infections: 2,494; Deaths: 858; Case-fatality rate: 34.4%
  - SARS-CoV-2 (2019 –)[3]
    Infections[4] : **>7,000,000**; Deaths : **>400,000**; Case-fatality rate[5] : **1.4%**

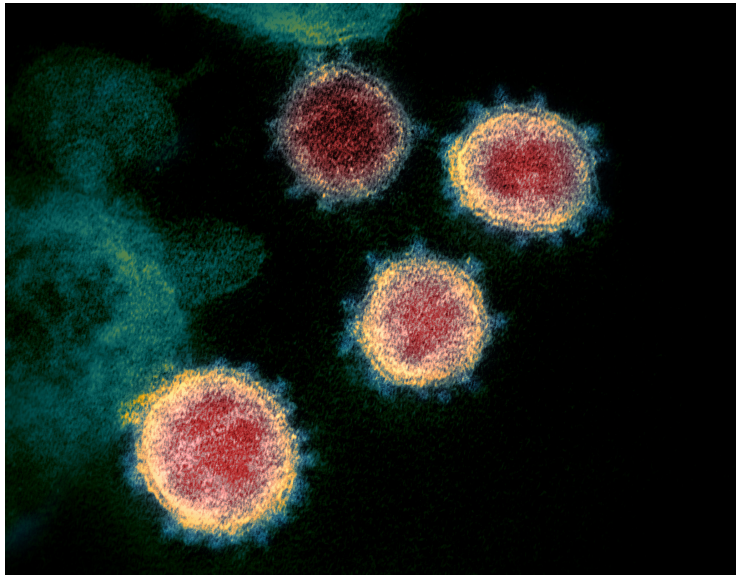
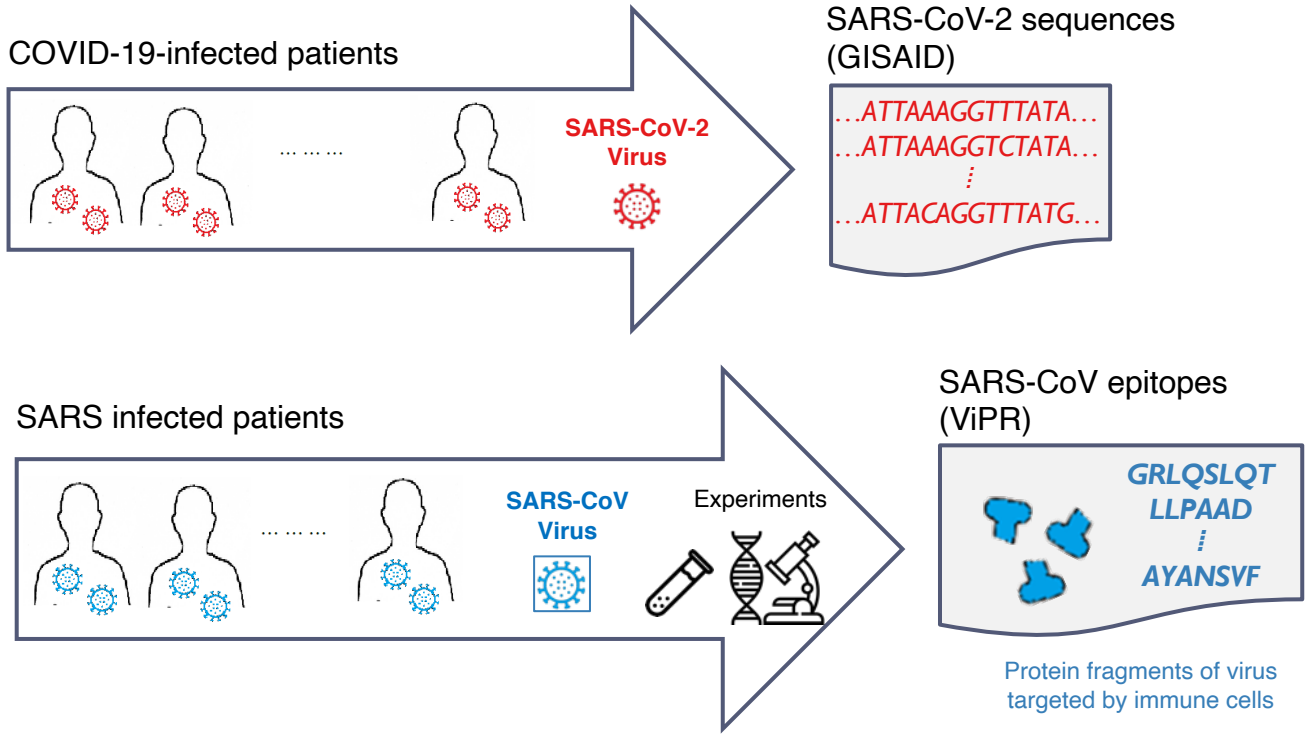


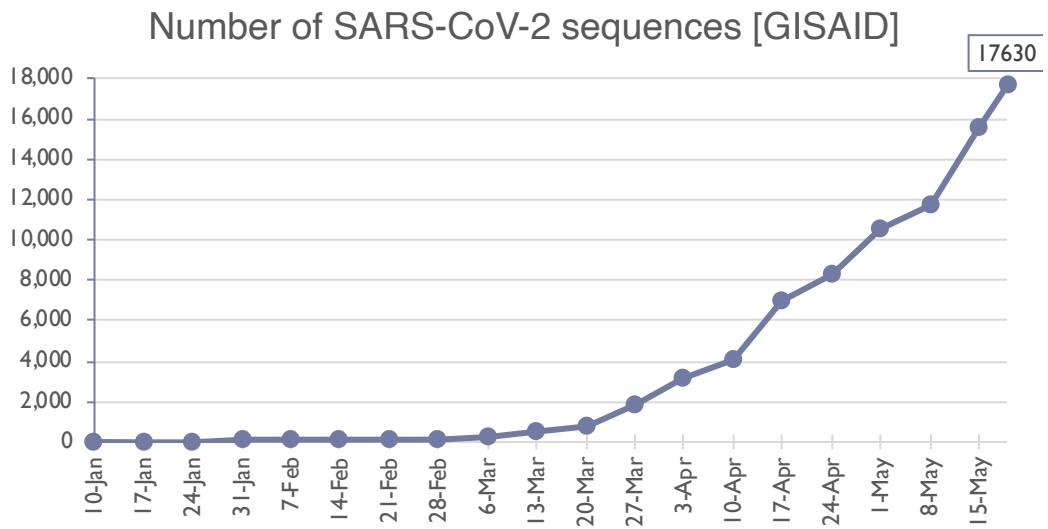
Image credit: NIH US

- Clear need for an effective vaccine
- **Our goal:** help guide vaccine design by presenting early vaccine target recommendations
- Seek to *identify which parts of the virus may elicit a protective immune response?*
- Challenged by a lack of knowledge of SARS-CoV-2

1. https://www.who.int/csr/sars/archive/2003_05_07a/en/
2. https://www.who.int/emergencies/mers-cov/en/
3. https://www.who.int/docs/default-source/coronaviruse/articles/coronavirus-(covid-19)-selected-bibliographic-references-18-02-2020-v1.pdf?sfvrsn=c8b8baa5_0
4. https://www.arcgis.com/apps/opsdashboard/index.html#/bda7594740fd40299423467b48e9ecf6
5. https://www.med.hku.hk/en/covid-19/articles/fatality-rate-of-covid-19

Providing SARS-CoV-2 vaccine target recommendations



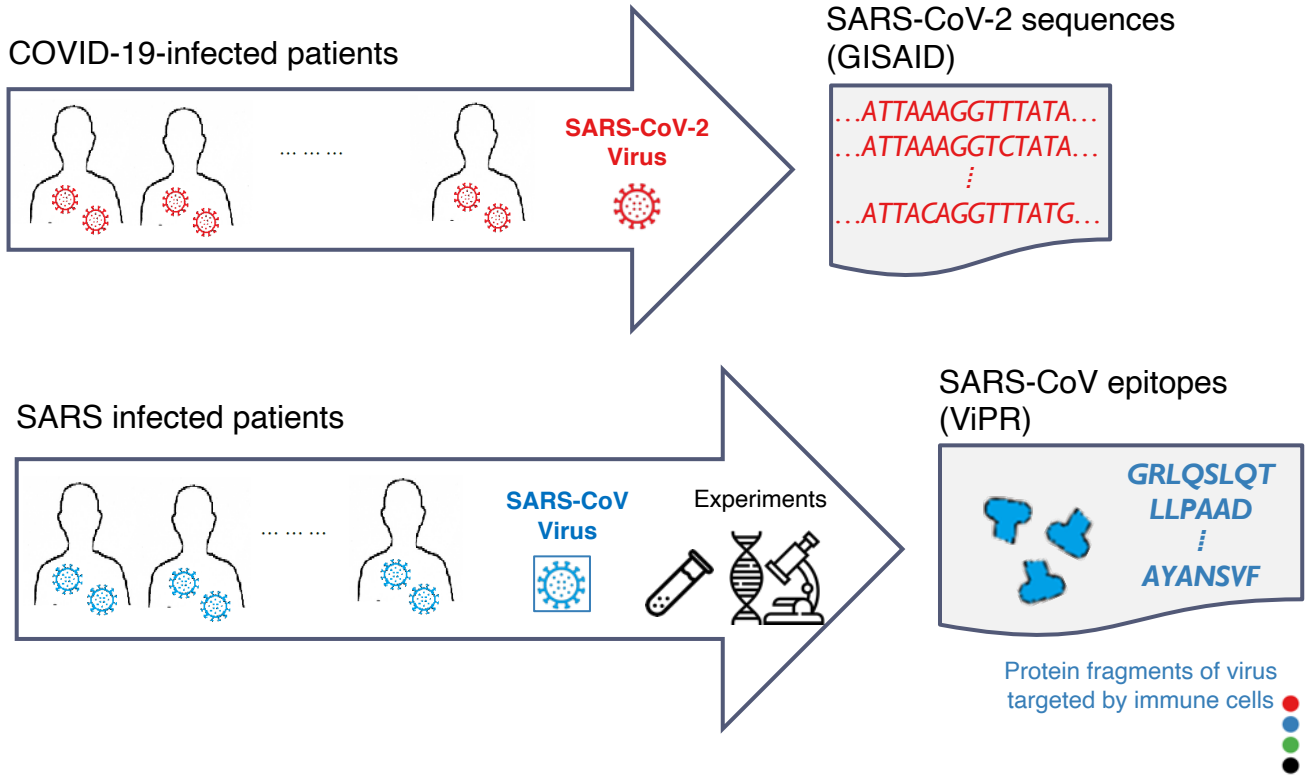


Percentage sequence identity with SARS-CoV-2

	S protein	N protein	M protein	E protein
SARS-CoV	76.0%	90.6%	90.1%	94.7%
MERS-CoV	29.4%	45.9%	39.2%	34.1%

Ahmed, Syed F. et al. "Preliminary identification of potential vaccine targets for the COVID-19 coronavirus (SARS-CoV-2) based on SARS-CoV immunological studies." *Viruses* 12 (**2020**).

Providing SARS-CoV-2 vaccine target recommendations



Idea

Exploit **immunological data for SARS-CoV** to provide vaccine target recommendations for SARS-CoV-2

Approach

Search for SARS-CoV epitopes with a **close genetic match** in SARS-CoV-2 sequences

Output

Potentially **immunogenic** epitopes

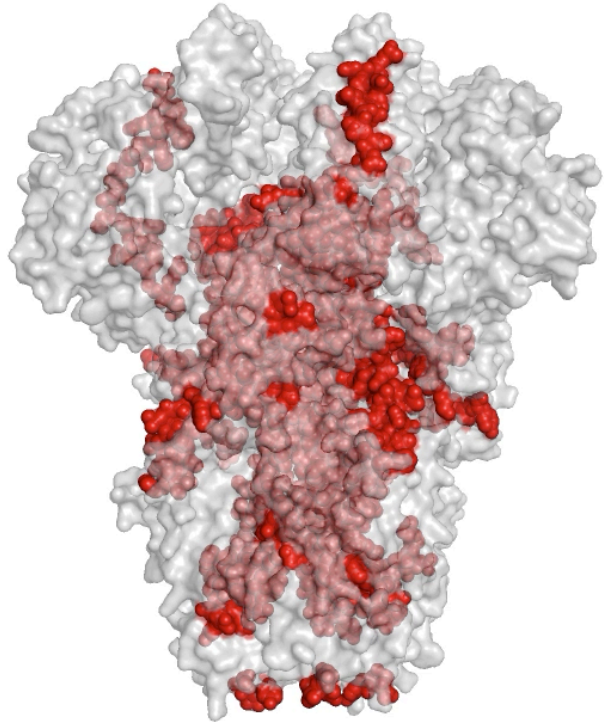
Ahmed, Syed F. et al. “Preliminary identification of potential vaccine targets for the COVID-19 coronavirus (SARS-CoV-2) based on SARS-CoV immunological studies.” *Viruses* 12 (**2020**).

Results summary

SARS-CoV epitopes	Percentage with identical genetic match in SARS-CoV-2
T cell epitopes	**24%**
B cell epitopes	**16%**
overall	**20%**

Identified set of T cell epitopes may provide broad population coverage globally (**96%**) as well as in China (**88%**)

B cell epitopes with an identical genetic match in SARS-CoV-2

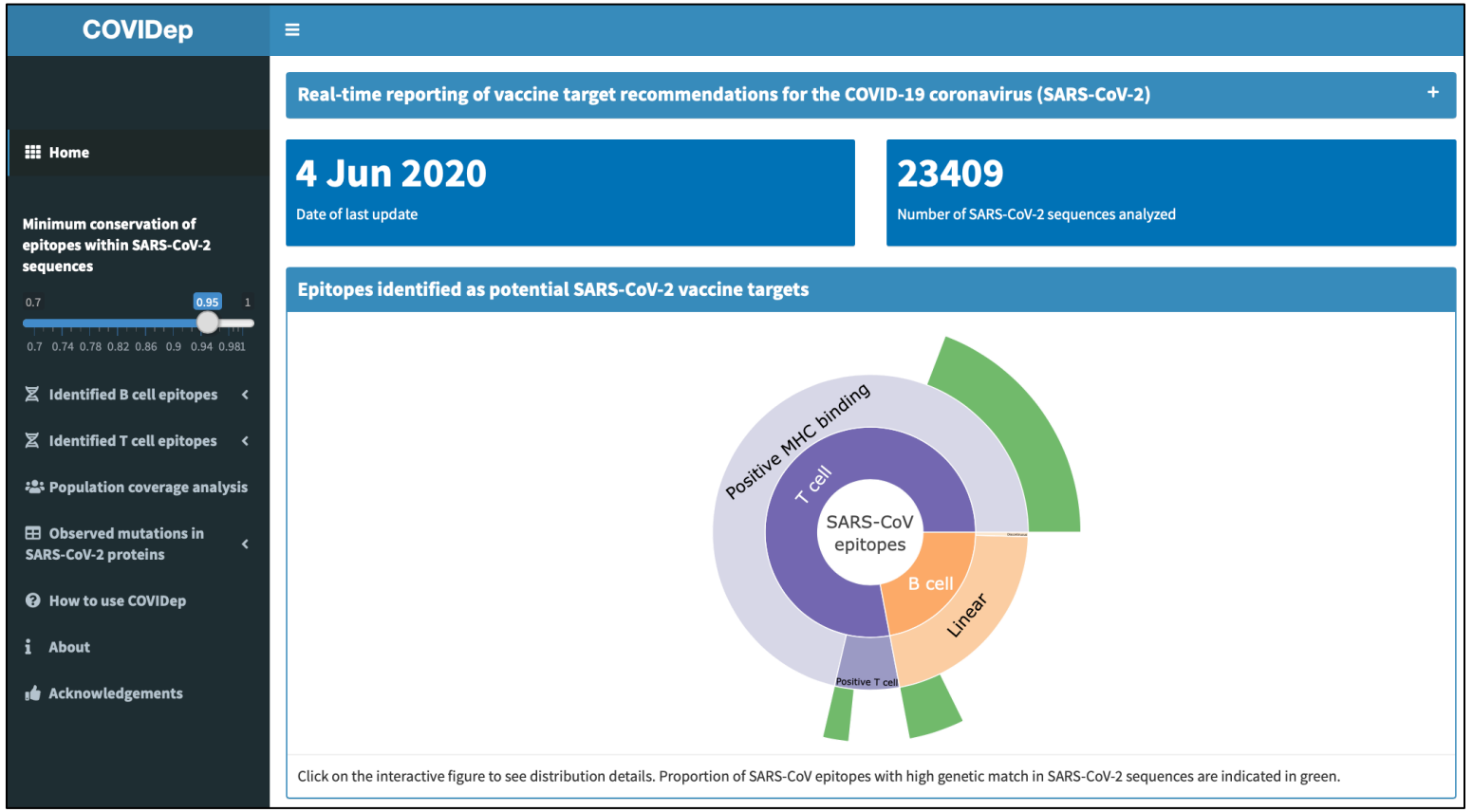


SARS-CoV spike protein
(PDB ID: 5XLR)

Ahmed, Syed F. et al. “Preliminary identification of potential vaccine targets for the COVID-19 coronavirus (SARS-CoV-2) based on SARS-CoV immunological studies.” *Viruses* 12 (**2020**).

Extension: COVIDep https://covidep.ust.hk

- Increase in number of SARS-CoV-2 sequences
  - **120 → now over 25,000**
- COVIDep: A web-based platform for real-time reporting of vaccine target recommendations for SARS-CoV-2
- Features:
  - Identification of SARS-derived B-cell and T-cell epitopes that provide vaccine target recommendations for SARS-CoV-2;
  - For T cell epitopes, it reports estimated population coverage using HLA/MHC statistical information;
  - Up-to-date reporting based on latest sequence data available (from GISAID).



Ahmed, Syed F. et al. "COVIDep: A web-based platform for real-time reporting of vaccine target recommendations for SARS-CoV-2." *to appear in Nature Protocols* (**2020**).

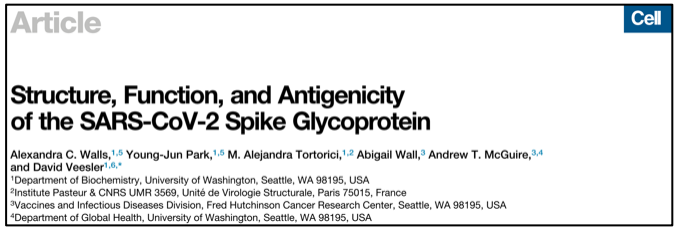
Connections with COVID-19 **B cell** responses and preclinical vaccine trials

- Most (24/29) identified B cell epitopes of the spike protein are in the **S2 subunit**, reported to be a main region targeted by **cross-neutralizing antibodies**

- Overlap with regions in the S1 subunit reported to be targeted by **cross-neutralizing antibodies**

- Overlap with an epitope targeted by neutralizing antibodies in a **preclinical trial of a SARS-CoV-2 vaccine candidate**

- Overlap with regions recognized by neutralizing antibodies in **recovered COVID-19 patients**


Article

Cell

Structure, Function, and Antigenicity of the SARS-CoV-2 Spike Glycoprotein

Alexandra C. Walls,[1,5] Young-Jun Park,[1,5] M. Alejandra Tortorici,[1,2] Abigail Wall,[3] Andrew T. McGuire,[3,4] and David Veesler[1,6,*]

[1]Department of Biochemistry, University of Washington, Seattle, WA 98195, USA
[2]Institute Pasteur & CNRS UMR 3569, Unité de Virologie Structurale, Paris 75015, France
[3]Vaccines and Infectious Diseases Division, Fred Hutchinson Cancer Research Center, Seattle, WA 98195, USA
[4]Department of Global Health, University of Washington, Seattle, WA 98195, USA


nature COMMUNICATIONS

ARTICLE

https://doi.org/10.1038/s41467-020-16256-y OPEN

A human monoclonal antibody blocking SARS-CoV-2 infection


nature

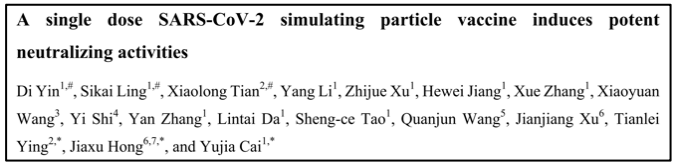
https://doi.org/10.1038/s41586-020-2349-y

Accelerated Article Preview

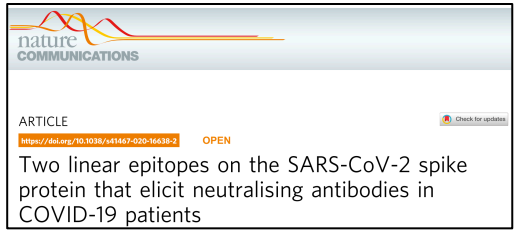
Cross-neutralization of SARS-CoV-2 by a human monoclonal SARS-CoV antibody

Received: 6 April 2020
Accepted: 12 May 2020
Accelerated Article Preview Published online 18 May 2020

Dora Pinto, Young-Jun Park, Martina Beltramello, Alexandra C. Walls, M. Alejandra Tortorici, Siro Bianchi, Stefano Jaconi, Katja Culap, Fabrizia Zatta, Anna De Marco, Alessia Peter, Barbara Guarino, Roberto Spreafico, Elisabetta Cameroni, James Brett Case, Rita E. Chen, Colin Havenar-Daughton, Gyorgy Snell, Amalio Telenti, Herbert W. Virgin, Antonio Lanzavecchia, Michael S. Diamond, Katja Fink, David Veesler & Davide Corti


A single dose SARS-CoV-2 simulating particle vaccine induces potent neutralizing activities

Di Yin[1,#], Sikai Ling[1,#], Xiaolong Tian[2,#], Yang Li[1], Zhijue Xu[1], Hewei Jiang[1], Xue Zhang[1], Xiaoyuan Wang[3], Yi Shi[4], Yan Zhang[1], Lintai Da[1], Sheng-ce Tao[1], Quanjun Wang[5], Jianjiang Xu[6], Tianlei Ying[2,*], Jiaxu Hong[6,7,*], and Yujia Cai[1,*]


nature COMMUNICATIONS

ARTICLE

https://doi.org/10.1038/s41467-020-16638-2 OPEN

Two linear epitopes on the SARS-CoV-2 spike protein that elicit neutralising antibodies in COVID-19 patients

Details of the identified B cell epitopes in the S protein

Show 50 entries

Download csv

Search:

IEDB	Epitope	Length	Start	End	Conservation
30987	KGIYQTSN	8	310	317	0.9646
70719	VRFPNITNLCPFGEVFN	17	327	343	0.9948
15972	FGEVFNAT	8	338	345	0.9976
52020	QQFGRD	6	563	568	0.9999
18594	GAGICASY	8	667	674	0.9993
22321	GSFCTQLN	8	757	764	0.9996
16183	FIEDLLFNKVTLADAGF	17	817	833	0.9952
18515	GAALQIPFAMQMAYRFN	17	891	907	0.9989
47479	PFAMQMAYRFNGIGVTQ	17	897	913	0.9993
3176	AMQMAYRF	8	899	906	0.9994
41177	MAYRFNGIGVTQNVLYE	17	902	918	0.9995
10778	DVVNQNAQALNTLVKQL	17	950	966	0.9988
50311	QALNTLVKQLSSNFGAI	17	957	973	0.999
2092	AISSVLNDILSRLDKVE	17	972	988	0.9994
27357	ILSRLDKVEAEVQIDRL	17	980	996	0.9997
11038	EAEVQIDRLITGRLQSL	17	988	1004	0.9997
54599	RLITGRLQSLQTYVTQQ	17	995	1011	0.9997
59425	SLQTYVTQQLIRAAEIR	17	1003	1019	0.9995
51379	QLIRAAEIRASANLAAT	17	1011	1027	0.9979
53202	RASANLAATKMSECVLG	17	1019	1035	0.998
462	AATKMSECVLGQSKRVD	17	1025	1041	0.9989
67220	TVYDPLQPELDSFKEEL	17	1136	1152	0.9696
32508	KNHTSPDVDLGDISGIN	17	1157	1173	0.9974
9094	DLGDISGINASVVNIQK	17	1165	1181	0.9981
12426	EIDRLNEVAKNLNESLIDLQELGKYEQY	28	1182	1209	0.9974
558417	EIDRLNEVAKNLNESLIDLQELGKYEQY	28	1182	1209	0.9974
14626	EVAKNLNESLIDLQELG	17	1188	1204	0.9979
6476	CKFDEDDSEPVLKGVKLHYT	20	1254	1273	0.9911
7868	DDSEPVLKGVKLHYT	15	1259	1273	0.9914

Showing 1 to 29 of 29 entries

Previous 1 Next

Connections with COVID-19 **T cell** responses and preclinical vaccine trials

- Of the 14 HLA-A*02:01-restricted spike protein epitopes identified by COVIDep, **9 epitopes** overlap with SARS-CoV-2 **immunogenic** epitopes.

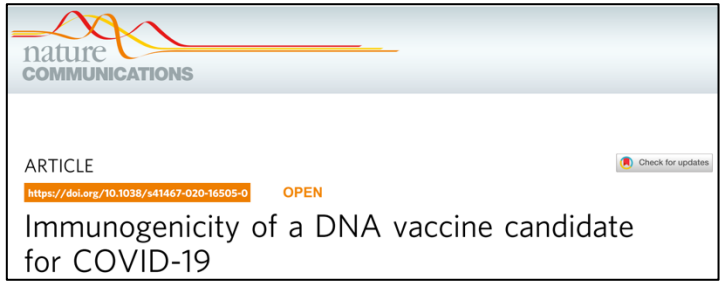
Shared Antigen-specific CD8⁺ T cell Responses Against the SARS-COV-2 Spike Protein in HLA-A*02:01 COVID-19 Participants

William Chour[1,2,3], Alexander M. Xu[1,4], Alphonsus H.C. Ng[1,4], Jongchan Choi[1], Jingyi Xie[1,5], Dan Yuan[1,6], , Diana C. DeLucia[7], Rick A. Edmark[1], Lesley C. Jones[1], Thomas M. Schmitt[8], Mary E. Chaffee[8], Venkata R. Duvvuri[1], Kim M. Murray[1], Songming Peng[9], Julie Wallick[10], Heather A. Algren[10], William R. Berrington[10], D. Shane O'Mahony[10], John K. Lee[7,11], Philip D. Greenberg[8,12], Jason D. Goldman[10,13*], and James R. Heath[1*]

SARS-CoV-2 epitopes are recognized by a public and diverse repertoire of human T-cell receptors

Alina S. Shomuradova[1], Murad S. Vagida[1], Savely A. Sheetikov[1], Ksenia V. Zornikova[1], Dmitriy Kiryukhin[1], Aleksei Titov[1], Iuliia O. Peshkova[1], Alexandra Khmelevskaya[1], Dmitry V. Dianov[1], Maria Malasheva[1], Anton Shmelev[1], Yana Serdyuk[1], Dmitry V. Bagaev[2], Anastasia Pivnyuk[3], Dmitrii S. Shcherbinin[4,5], Alexandra V. Maleeva[1], Naina T. Shakirova[1], Artem Pilunov[1], Dmitry B. Malko[1], Ekaterina G. Khamaganova[1], Bella Biderman[1], Alexander Ivanov[6], Mikhail Shugay[3,4,5*] and Grigory A. Efimov[1**]

- In a **preclinical vaccine trial**, T cell responses have also been recorded against a protein region comprising a COVIDep-identified epitope



Details of the identified T cell epitopes in the S protein

Show 25 entries

Download csv

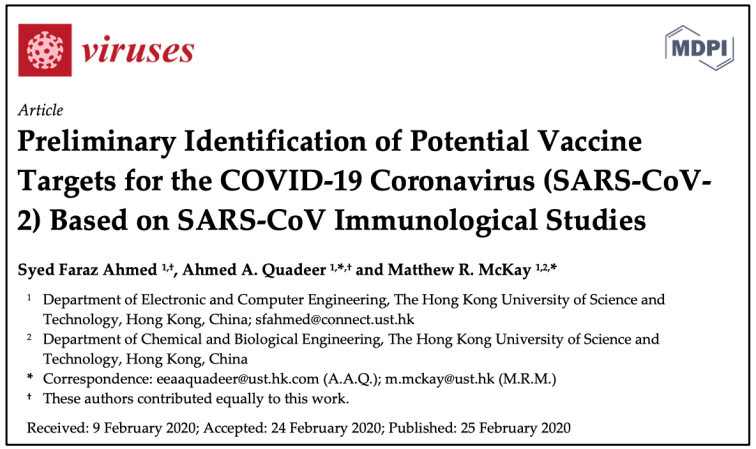
Search: HLA-A*02:01

IEDB	Epitope	Length	Start	End	MHC allele class	MHC allele names	Conservation
36724	LITGRLQSL	9	996	1004	I	HLA-A2/HLA-A*02:01	0.9998
54507	RLDKVEAEV	9	983	991	I	HLA-A*02:01/HLA-A*02:02/HLA-A*02:06/HLA-A*02:03/HLA-A*68:02	0.9998
54725	RLQSLQTYV	9	1000	1008	I	HLA-A*02:01/HLA-A*02:02/HLA-A*02:03/HLA-A*02:06/HLA-A*68:02	0.9998
37544	LLLQYGSFC	9	752	760	I	HLA-A*02:01	0.9997
37724	LLQYGSFCT	9	753	761	I	HLA-A*02:01	0.9997
69657	VLNDILSRL	9	976	984	I	HLA-A*02:01	0.9997
71663	VVFLHVTYV	9	1060	1068	I	HLA-A*02:01/HLA-A*02:02/HLA-A*02:03/HLA-A*02:06/HLA-A*68:02	0.9995
2801	ALNTLVKQL	9	958	966	I	HLA-A*02:01	0.9994
44814	NLNESLIDL	9	1192	1200	I	HLA-A*02:01	0.9993
26710	IITTDNTFV	9	1114	1122	I	HLA-A*02:01	0.9992
54680	RLNEVAKNL	9	1185	1193	I	HLA-A*02:01	0.9992
16156	FIAGLIAIV	9	1220	1228	I	HLA-A*02:01/HLA-A*02:02/HLA-A*02:03/HLA-A*02:06/HLA-A*68:02/HLA-A2	0.9991
20907	GLIAIVMVTI	10	1223	1232	I	HLA-A*02:02/HLA-A*02:03/HLA-A*02:01/HLA-A*02:06/HLA-A*68:02	0.9985
37289	LLFNKVTLA	9	821	829	I	HLA-A*02:01/HLA-A*02:02/HLA-A*02:03/HLA-A*02:06/HLA-A*68:02	0.9976

Showing 1 to 14 of 14 entries (filtered from 75 total entries)

Previous 1 Next

Part 2: Summary



viruses

MDPI

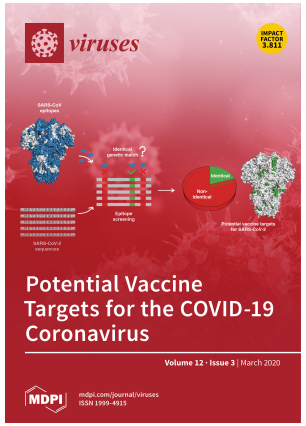
Article

Preliminary Identification of Potential Vaccine Targets for the COVID-19 Coronavirus (SARS-CoV-2) Based on SARS-CoV Immunological Studies

Syed Faraz Ahmed [1,†], Ahmed A. Quadeer [1,*,†] and Matthew R. McKay [1,2,*]

[1] Department of Electronic and Computer Engineering, The Hong Kong University of Science and Technology, Hong Kong, China; sfahmed@connect.ust.hk
[2] Department of Chemical and Biological Engineering, The Hong Kong University of Science and Technology, Hong Kong, China
* Correspondence: eeaaquadeer@ust.hk.com (A.A.Q.); m.mckay@ust.hk (M.R.M.)
† These authors contributed equally to this work.

Received: 9 February 2020; Accepted: 24 February 2020; Published: 25 February 2020



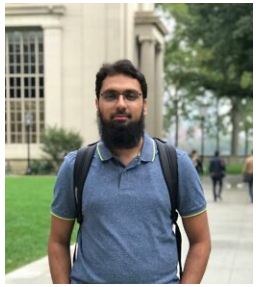
COVIDep platform for real-time reporting of vaccine target recommendations for SARS-CoV-2: Description and connections with COVID-19 immune responses and preclinical vaccine trials

Syed Faraz Ahmed[1], Ahmed A. Quadeer[1,†], Matthew R. McKay[1,2,†]

[1]Department of Electronic and Computer Engineering, The Hong Kong University of Science and Technology, Hong Kong, China
[2]Department of Chemical and Biological Engineering, The Hong Kong University of Science and Technology, Hong Kong, China

[†]Joint corresponding authors: eeaaquadeer@ust.hk and m.mckay@ust.hk

Faraz



Ahmed

Some informative reviews on COVID-19 vaccine development

1. T. T. Le, et al, The COVID-19 vaccine development landscape, *Nature Reviews Drug Discovery*, April 2020
2. F. Amanat, F. Krammer, "SARS-CoV-2 vaccines: Status report, *Immunity*, April 2020
3. E. Callaway, "The race for coronavirus vaccines", *Nature*, April 2020

PART 3:
Other related projects

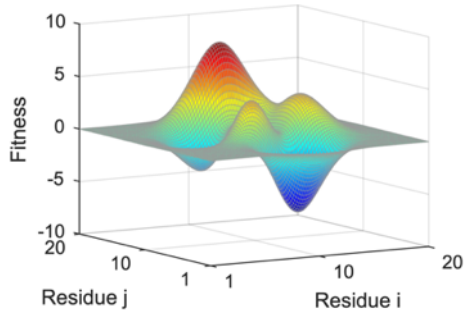
Illustration: istock

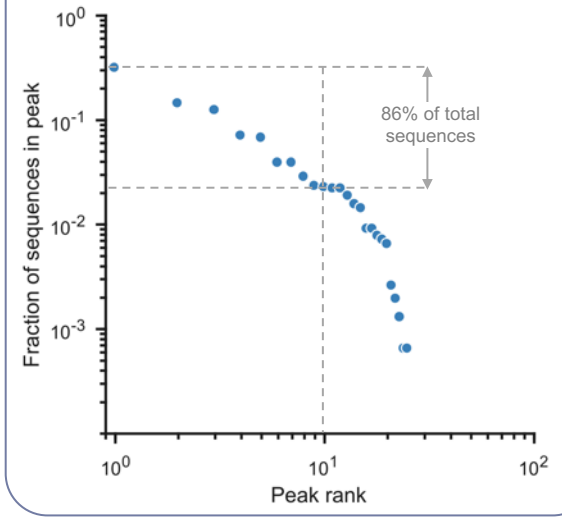
Using maximum entropy model to explore why the polio vaccine is so effective

Fitness × Prevalence

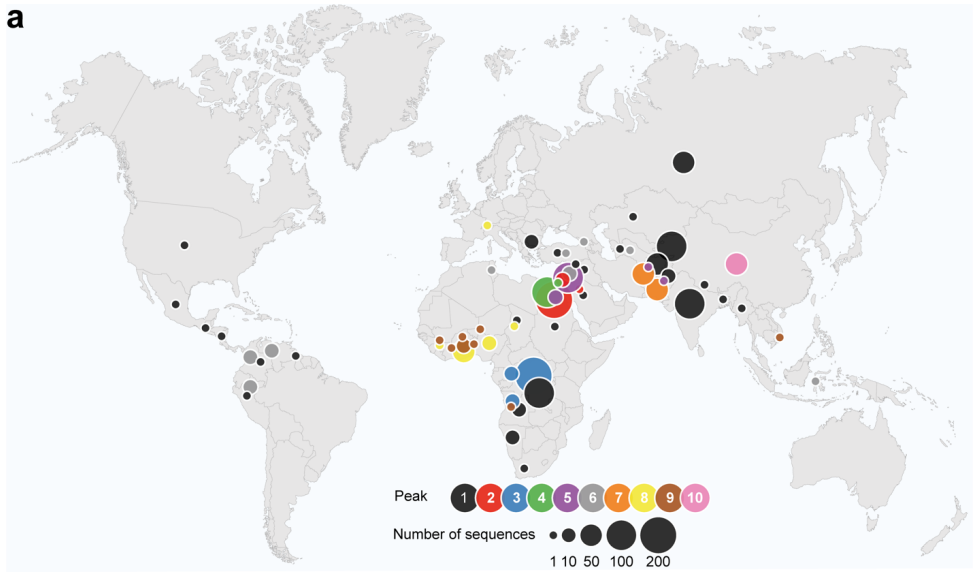
$$f(\mathbf{x}) \sim p(\mathbf{x}) = \frac{\exp[-E(\mathbf{x})]}{Z}$$

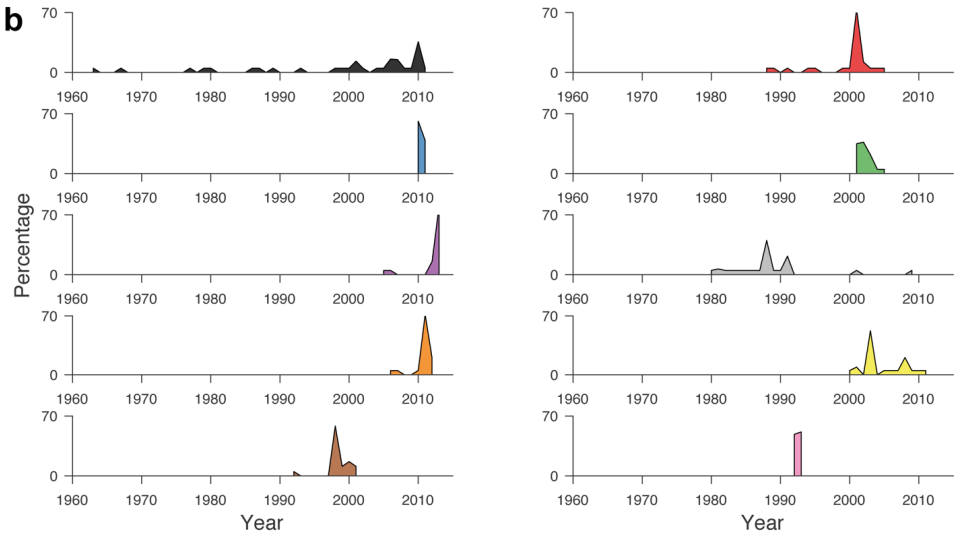
Inferred landscape comprises multiple peaks



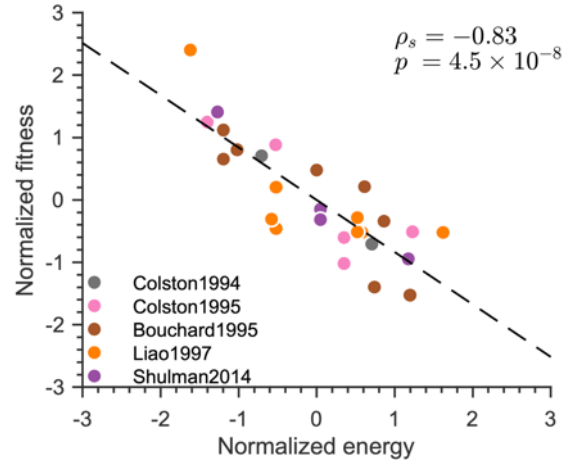


All peaks, except peak 1, reflect geographically and temporally-localized outbreaks

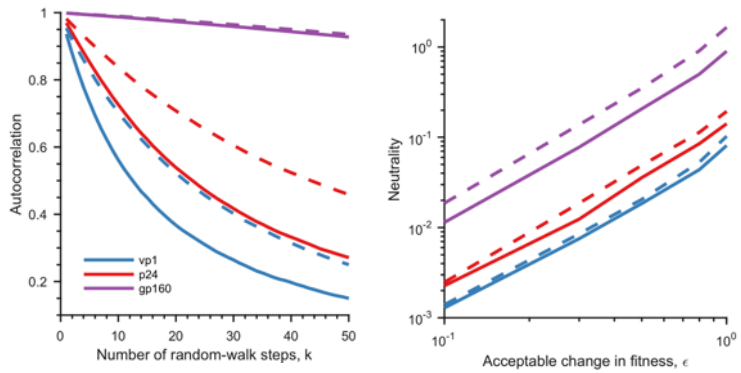




Landscape based on peak 1 reflects intrinsic fitness landscape



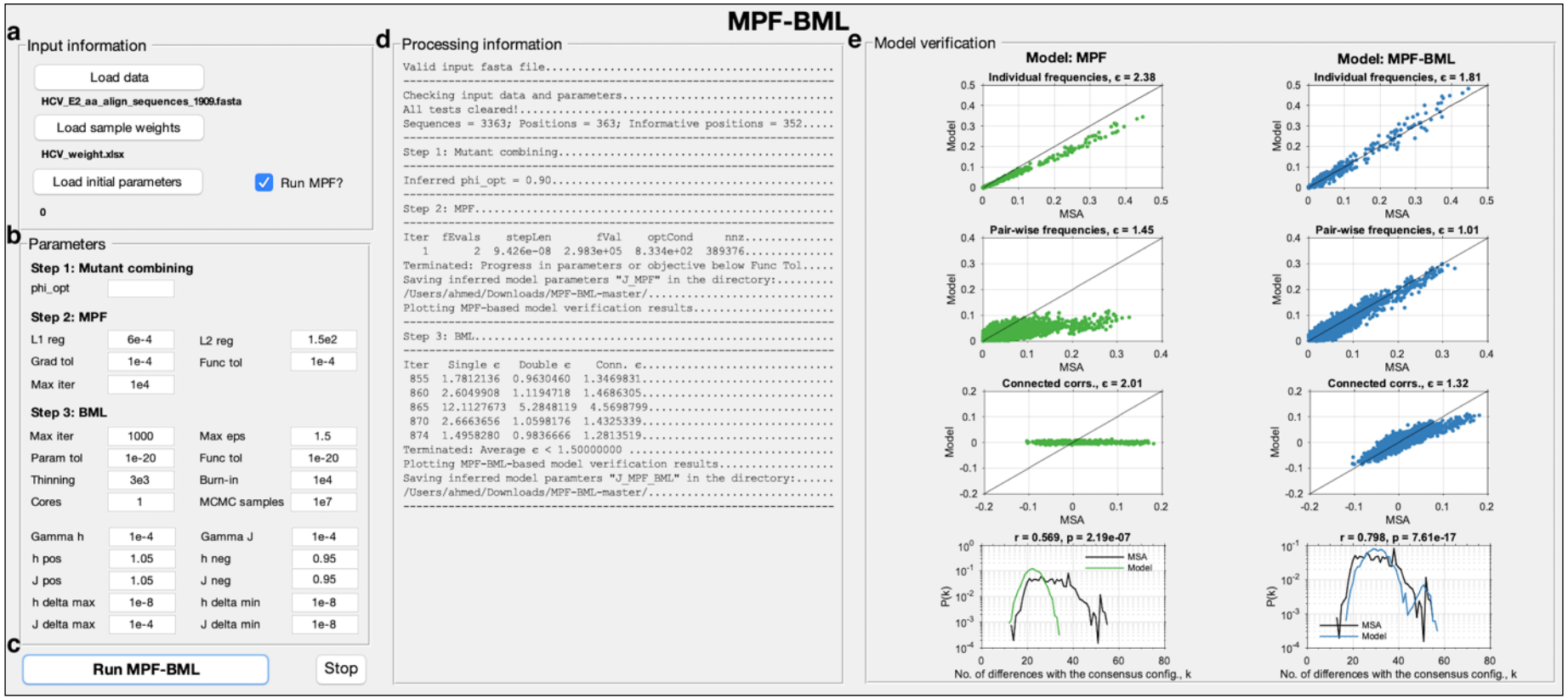
Compared to HIV proteins, PV is subject to more restrictive constraints



Quadeer, Ahmed A. et al. "Deconvolving mutational patterns of poliovirus outbreaks reveals its fitness landscape," *Nature Communications* 11 (**2020**).

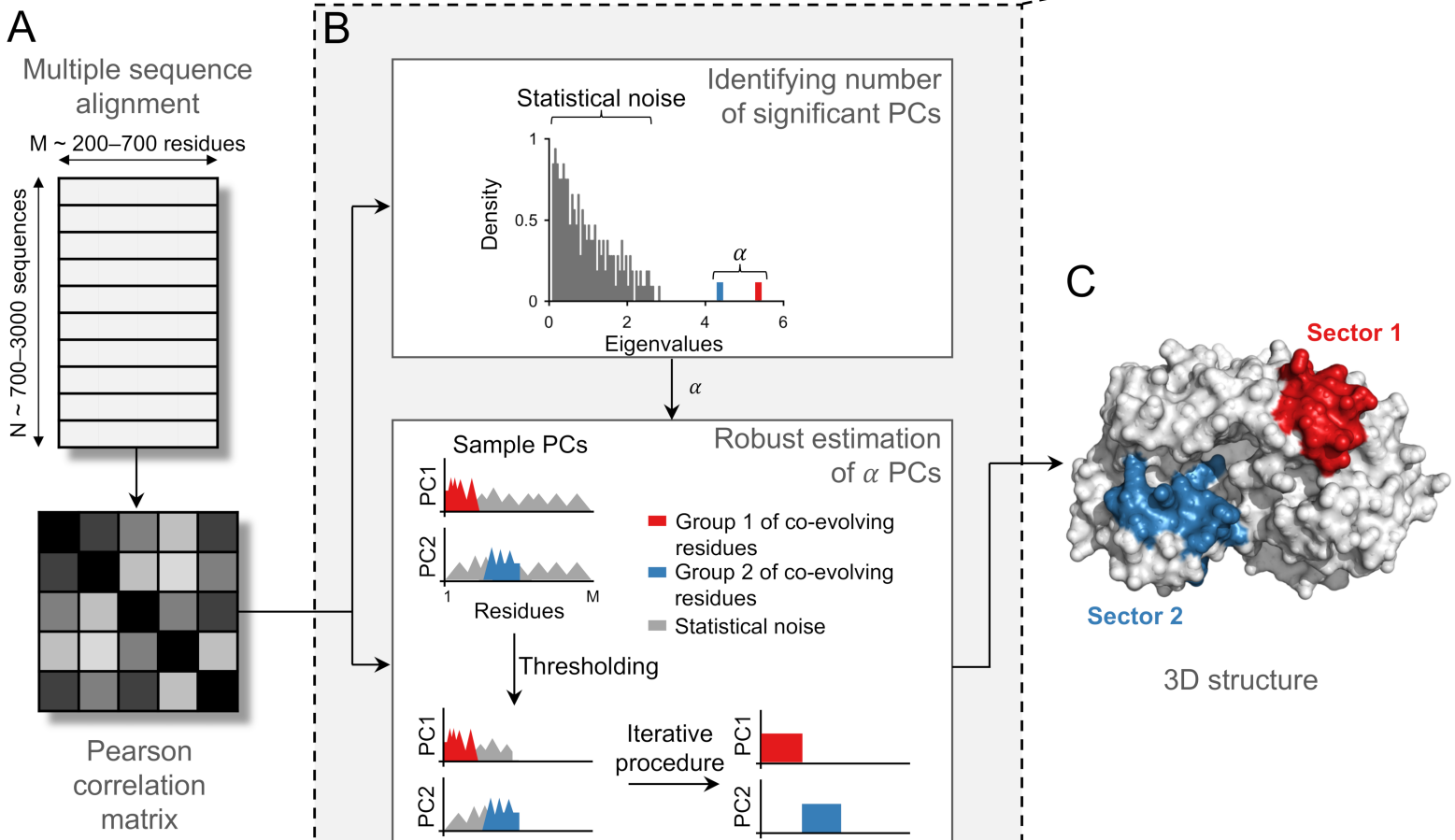
MPF-BML: A standalone GUI-based package for maximum entropy model inference

- **Standalone**—no requirement of any pre-installed application;
- **Cross-platform**—works for Windows, Linux, and mac OS;
- **GUI-based**—no knowledge of any programming language required;
- **Minimum input requirement**—only sample data and sample weights (if available) required;
- **Publication-quality figures output**—all results are saved as vector graphics.



Quadeer, Ahmed A. et al. "MPF-BML: A standalone GUI-based package for maximum entropy model inference", *Bioinformatics* 36 (**2020**).

Robust co-evolutionary analysis (RoCA) of proteins



- RMT-based sparse PCA (SPCA)
  - Pioneer work – Diagonal thresholding [Johnstone 2009]
  - Augmented SPCA [Paul 2012]
  - **Iterative thresholding SPCA (ITSPCA)** [Ma 2013]

Summary

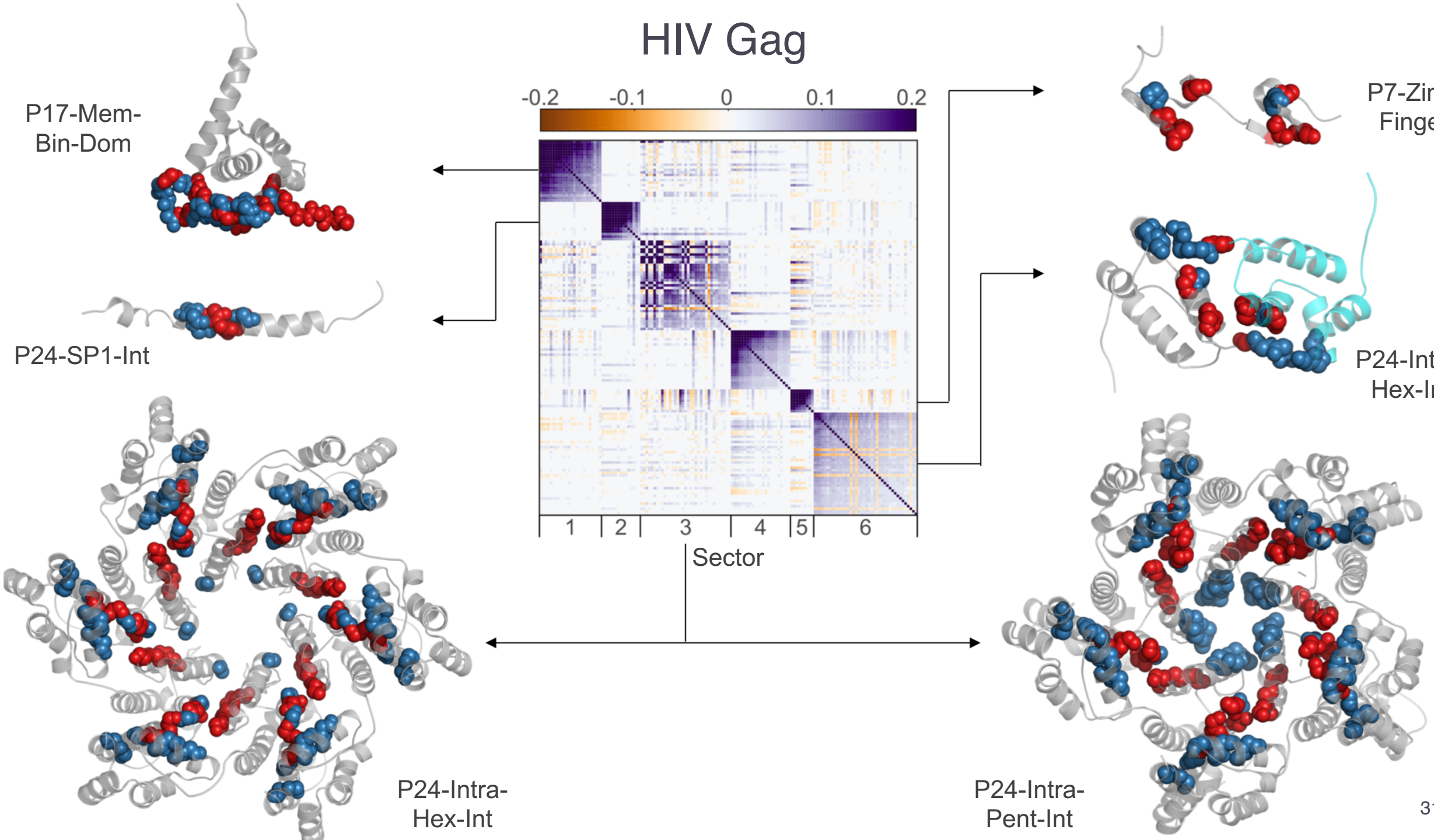
- Based on "orthogonal iteration method"
- Estimates sparse subspace of the leading eigenvectors
- Steps involved:
  - Projecting the correlation matrix on a subspace
  - Thresholding the columns of the resultant matrix using a **thresholding** parameter
  - Repeat until convergence

Robust Co-evolutionary Analysis (RoCA)

- Adapted to work on correlation matrices
- A data-driven **thresholding** parameter designed using ideas from random matrix theory

Quadeer, Ahmed A. et al. "Co-evolution networks of HIV/HCV are modular with direct association to structure and function." *PLoS Computational Biology* 14 (**2018**).

HIV Gag

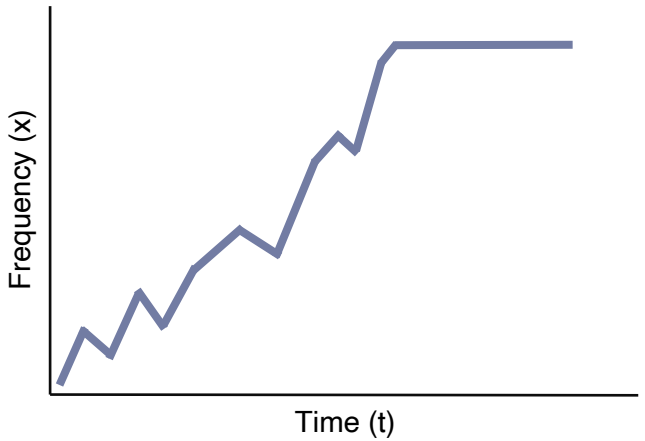

P17-Mem-Bin-Dom
P24-SP1-Int
-0.2
-0.1
0
0.1
0.2
1
2
3
4
5
6
Sector
P7-Zinc-Finger
P24-Inter-Hex-Int
P24-Intra-Hex-Int
P24-Intra-Pent-Int

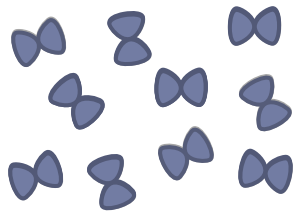
RocaNet: A standalone GUI-based package for robust co-evolutionary analysis of proteins

Quadeer, Ahmed A. et al. “RocaNet: A GUI-based package for robust co-evolutionary analysis of proteins”, *Bioinformatics* 36 (**2020**).

Inferring fitness based on evolutionary histories

Evolutionary history

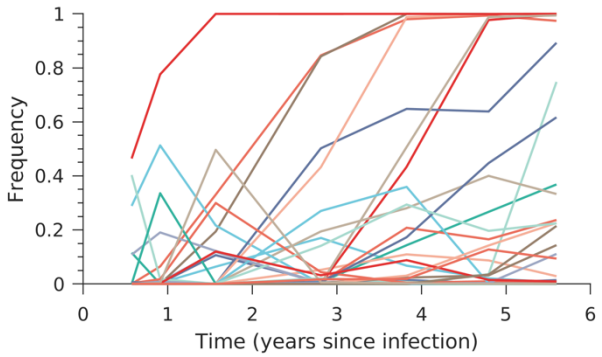




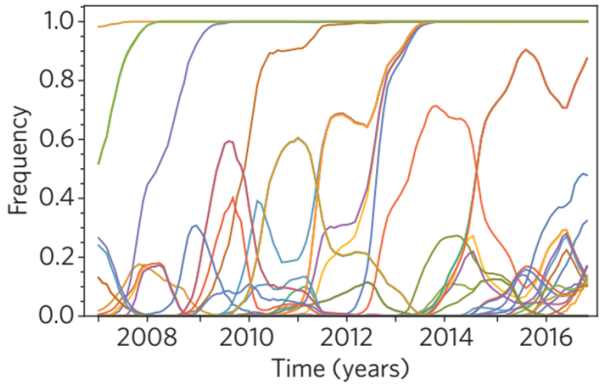
Wild-type ...CAGT... Mutant ...CATT...

Real data

HIV inpatient evolution[1]



Human Influenza global evolution[2]



Model: Wright-Fisher dynamics in the diffusion limit

Assumption: Fitness effects are additive

Evolution of mutant frequencies $\underline{x} = \{x_1, x_2, \ldots, x_L\}$ follows a **Fokker-Planck** (diffusion) equation

Drift term: expected change in mutant frequencies

$$d_i = \underbrace{x_i(1-x_i)s_i}_{\text{Selection}} + \underbrace{\sum_{i \neq j}(x_{ij} - x_i x_j)s_j}_{\text{Linked selection}} + \underbrace{(1-2x_i)\mu}_{\text{Mutation}}$$

Mutant frequency; Selection coefficient; Double mutant frequency; Mutation probability

Diffusion term: characterizes fluctuations due to the stochasticity of replication

$$C_{ij} = \begin{cases} x_i(1-x_i)/N & i = j \\ (x_{ij} - x_i x_j)/N & i \neq j \end{cases}$$

[1]Zanini, Fabio, et al. "Population genomics of intrapatient HIV-1 evolution." *Elife* 4 **(2015)**: e11282.

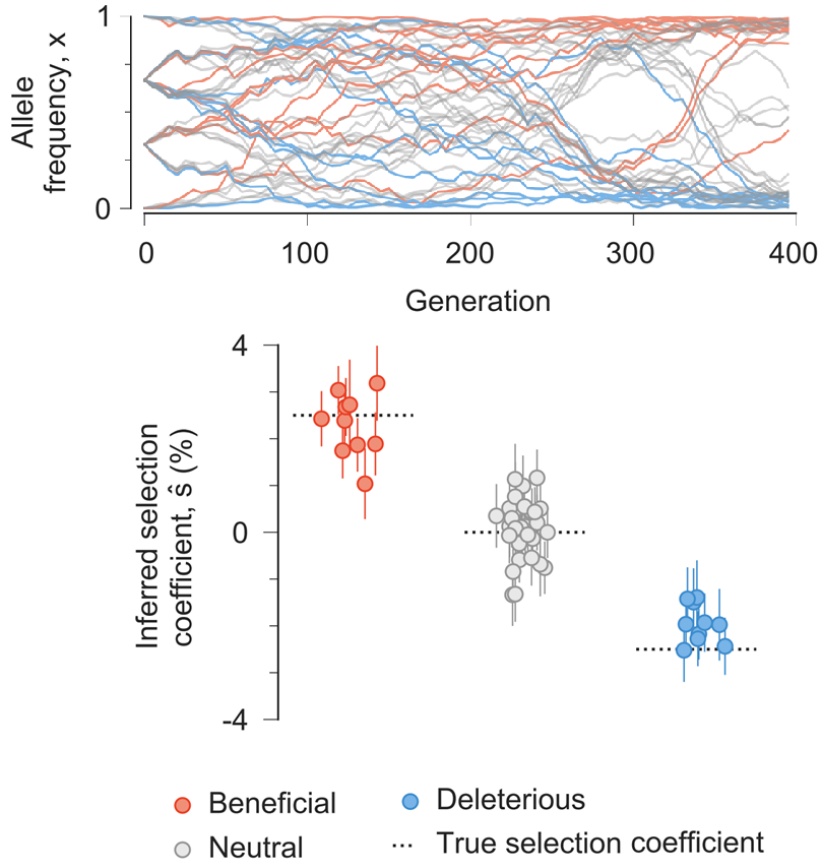
[2]Lässig, Michael, Ville Mustonen, and Aleksandra M. Walczak. "Predicting evolution." *Nature Ecology & Evolution* 1 **(2017)**: 0077.

Marginal path likelihood (MPL) estimate

- Evolutionary path $X = \{\underline{x}(t_0), \underline{x}(t_1), \ldots, \underline{x}(t_K)\}$
- Path integral --- probability of the evolutionary path $P(X|s)$
- Maximum a posteriori solution

$$\widehat{s_i} = \sum_{j=1}^{L} \left[\left(\sum_{k=0}^{K-1} C(t_k)\Delta t \right) + \gamma I \right]_{ij}^{-1} \left[x_j(t_K) - x_j(t_0) - \mu \sum_{k=0}^{K-1} \Delta t (1 - 2x_j(t_k)) \right]$$

Linkage effects | Change in frequency | Mutational flux

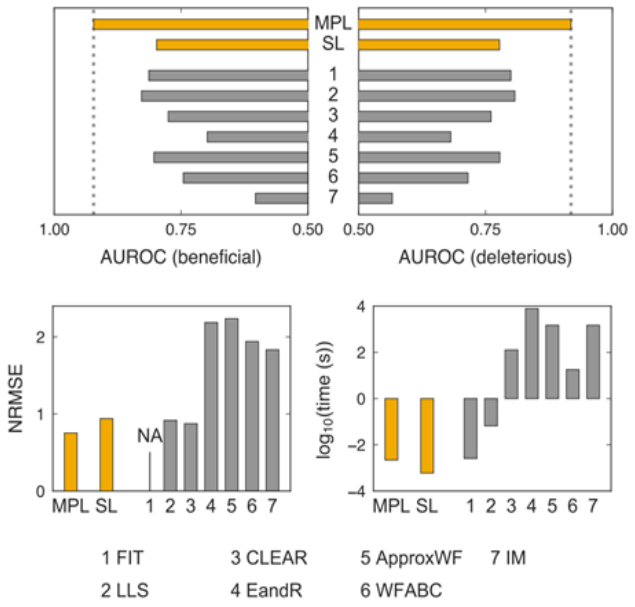


Simulated data: **50-site system**

M. S. Sohail, R. H. Y. Louie, M. R. M., and J. P. Barton, "Resolving genetic linkage reveals patterns of selection in HIV-1 evolution", https://www.biorxiv.org/content/10.1101/711861v1

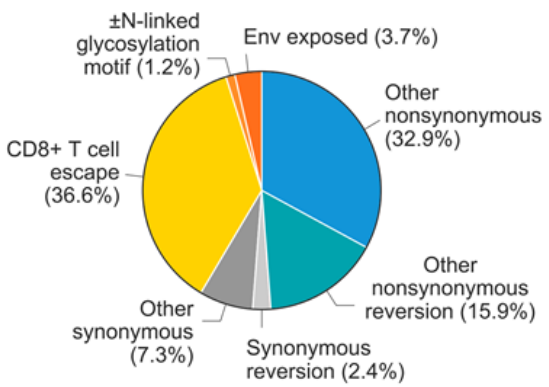
Results

Comparison with other methods

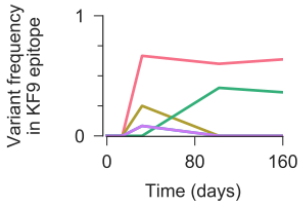


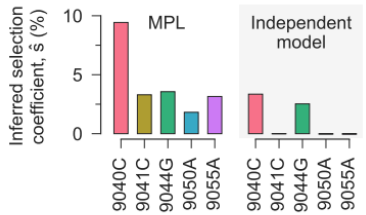
Simulated data: **50-site system**

Analysis of HIV half-genome longitudinal data of 14 individuals

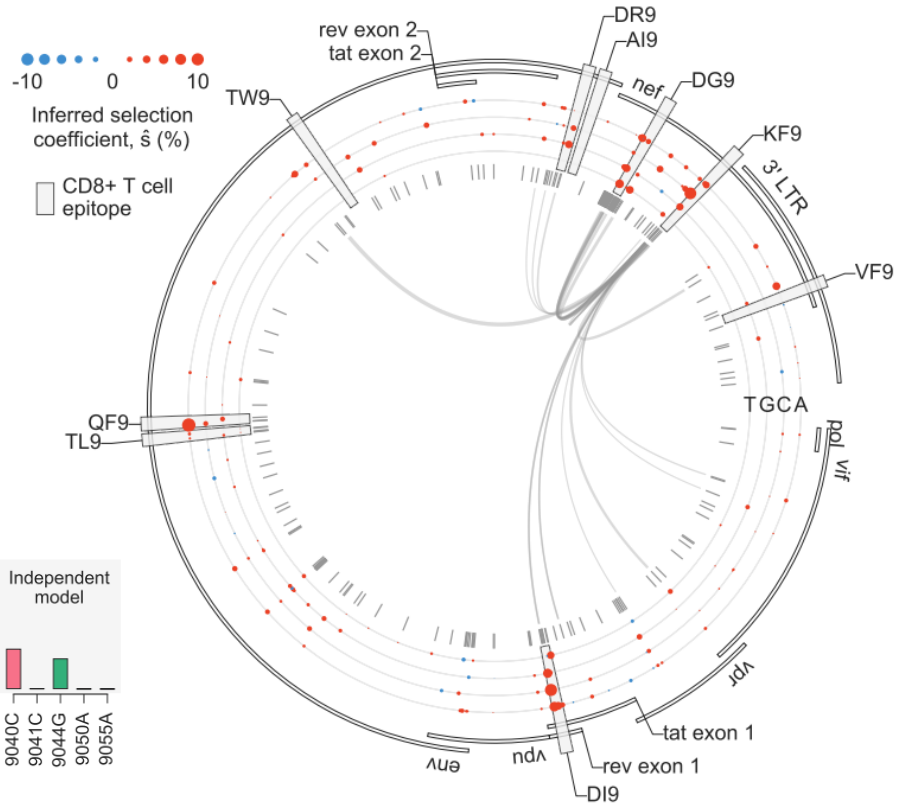


Patterns of **strong selection:**
Top **1%** selection coefficients





Clonal interference in **KF9** epitope of patient **CH77**



Linkage patterns in patient **CH77**

Acknowledgements

HKUST	Ahmed Abdul Quadeer Syed Faraz Ahmed Saqib Sohail
MIT	Arup Chakraborty
UC-Riverside	John Barton
UNSW	Raymond Louie
QUB	David Morales-Jimenez
Stanford	Iain Johnstone, Jeha Yang
Berkeley	Karthik Shekhar

Funding: General Research Fund of Hong Kong RGC

HKUST Signal Processing & Computational Biology Lab, https://www.mckayspcb.com