Dynamic immune landscape of SARS-Cov-2

SoSe 25: SARS-CoV-2 Bioinformatics & Data Science

Project 3+5, present to you by: Hang Mai Anh Vo and Kompal Fayyaz

Outline

Introduction

Methods & Results

- Project 3: Development of a risk score to analyze SARS-CoV-2 genomes
- Project 5: Cross-Immunization

Discussion

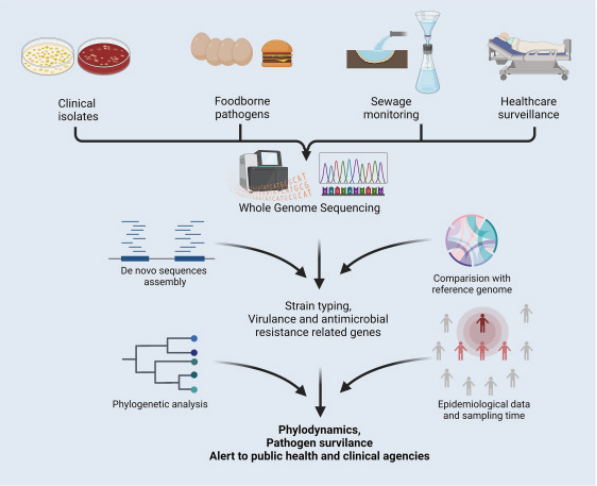
Conclusion

Introduction

Motivation & Background

Background

- Genomic surveillance needs quick triage of sequences (quality + biological signal).
- Typical issues: high ambiguity (N's), frameshifts/stop codons, improbable mutation patterns.



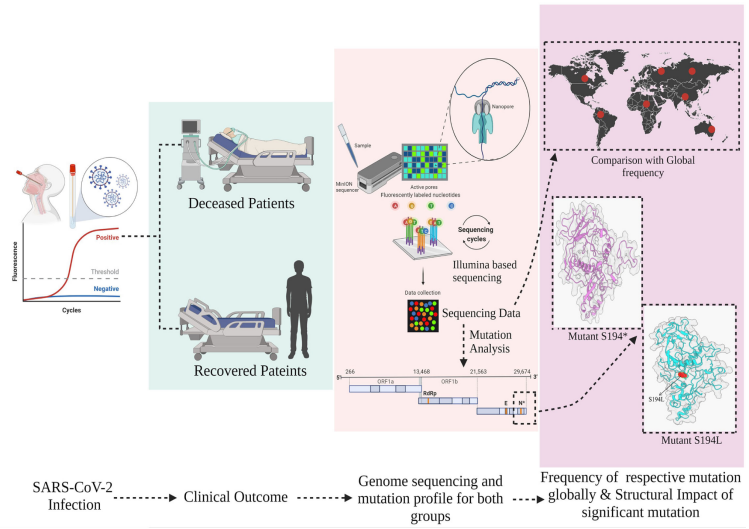
https://www.sciencedirect.com/science/article/abs/pii/B9780443187698000118

immunity is heterogeneous and time-dependent.

Recap: Viral Mutations

Why this matters:

- SARS-CoV-2 keeps evolving. Some mutations are benign; others shift transmissibility, immune escape, or diagnostic performance.
- A transparent risk score helps quickly flag unusual or low-quality genomes and prioritize follow-up.



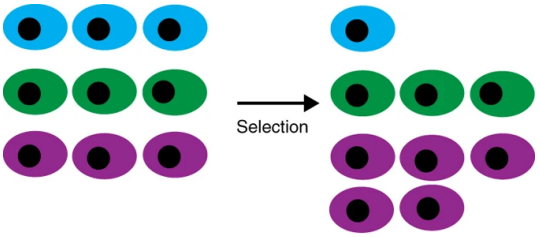
https://www.frontiersin.org/journals/cellular-and-infection-microbiology/articles/10.3389/fcimb.2022.868414/full

Variants with mutations that **evade existing antibodies** gain a relative fitness edge at that moment. When immunity realigns (new waves/boosters), the advantage can fade and a different lineage can rise.

Deep Mutational Scanning (DMS)

also called massively parallel mutagenesis, a high-throughput experiment

→ tests how every possible amino-acid mutation in a protein changes some property — for example, how well antibodies bind to the SARS-CoV-2 spike.



Variant	Mutation	Counts (input)	Counts (selected)	Functional score
Blue	A60P	3	1	0.33
Green	WT	3	3	1
Purple	S36T	3	5	1.67

Philosophy:

"By enabling the impact of mutations to be examined in an unbiased fashion, deep mutational scanning can reveal the unexpected."

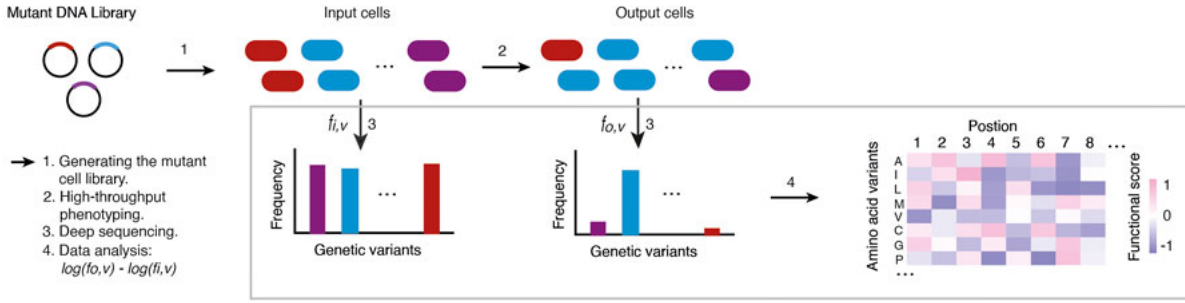
Image and quote from: Fowler, D. M., & Fields, S. (2014). Deep mutational scanning: a new style of protein science. Nature methods, 11(8), 801-807.
https://www.nature.com/articles/nmeth.3027

Large-scale mutational data could empower these computational approaches. First, these data provide a new resource for benchmarking computational approaches. Second, analysis of a modest number of large-scale mutagenesis data sets derived from proteins with diverse structures and functions could enhance our understanding of how, in a general sense, mutations affect protein function. This information should be useful for improving the accuracy of physicochemical models of the impact of mutations. Third, large-scale mutagenesis data in model organisms that are selected for their fitness could even contribute to developing computational models that predict the effects of mutations on a more complex organism.

First experimental implementations: ~2010–2013.

First paper that coined and defined the method "Deep Mutational Scanning" and became the field's reference: Fowler & Fields, 2014, Nat Methods.

Deep Mutational Scanning (DMS)



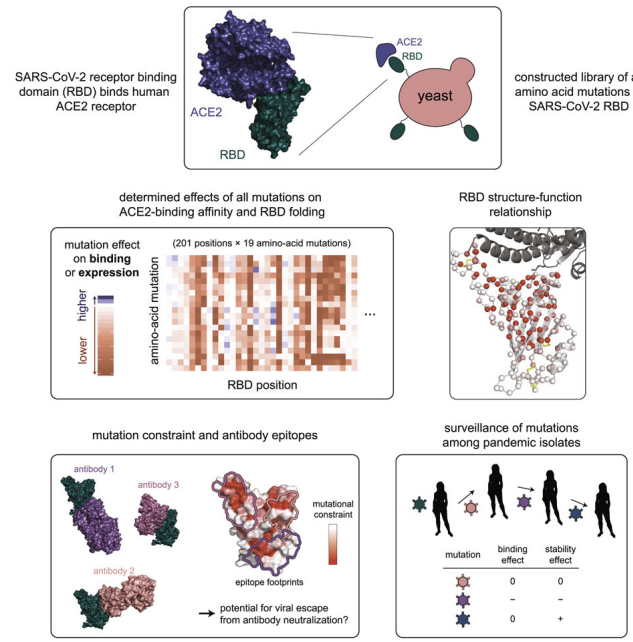
DMS links genotype to phenotype on a massive scale by generating a comprehensive library of gene or protein variants, subjecting them to a functional selection or screen, and using next-generation sequencing (NGS) to count each variant’s frequency before and after selection.

DMS in SAR-CoV-2 research

Focus on which viral mutations escape from antibody binding

→ “Escape fraction”, “escape maps”/”escape mutation maps”

Starr, T. N., Greaney, A. J., Hilton, S. K., Ellis, D., Crawford, K. H., Dingens, A. S., ... & Bloom, J. D. (2020). Deep mutational scanning of SARS-CoV-2 receptor binding domain reveals constraints on folding and ACE2 binding. cell, 182(5), 1295-1310.



yeast-surface-display platform → it can be assayed for ligand-binding affinity or protein expression levels, a close correlate of protein folding efficiency and stability

Starr et al. suggest that purifying selection is the main force acting on RBD mutations observed in human SARS-CoV-2 isolates to date.

Because yeast have protein-folding quality control and glycosylation machinery similar to mammalian cells, they add N-linked glycans at the same RBD sites as human cells (Chen et al., 2014), although these glycans are more mannose rich than mammalian-derived glycans (Hamilton et al., 2003). The yeast-expressed RBD from SARS-CoV-1 has similar antigenic and structural properties to the RBD expressed in mammalian cells (Chen et al., 2014, 2017, 2020a) and binds to ACE2 as expected (Chen et al., 2014).

By examining the concordance of RBD variant sequences for barcodes sampled by multiple PacBio reads, we validated that this process correctly determined the sequence of >99.8% of the variants (Figure S1B). RBD variants contained an average of 2.7 amino acid mutations, with the number of mutations per variant roughly following a Poisson distribution (Figure S1C).

studies of protein evolution demonstrating that epistastic entrenchment causes amino acid preferences to change as proteins diverge
→ So the sentence means: When proteins evolve, earlier mutations change how later ones behave. As a result, the best (most stable or functional) amino acid at a given position can change over time — because the protein’s context has changed.
In simpler words: As proteins evolve, earlier changes make later ones depend on them, so the "preferred" amino acids shift as the protein sequence drifts away from its ancestor.

Task Description

Project 3: Development of a risk score to analyze SARS-CoV-2 genomes

Download and align SARS-CoV-2 sequence to Wuhan outbreak (NCBI: NC_045512.2)

- Find mutant positions in the Spike gene
- Perform basic Statistics
- Number of ambiguous nucleotides per sequence and (SNPs, INDELs)
- Project and cluster sequences (for example: PCA, kmeans)

https://www.ncbi.nlm.nih.gov/nuccore/NC_045512.2

Project 5: Cross Immunization

Mutation profile of five Covid19 virus.

- JN.1
- JN.2
- JN.3
- KP.3
- XBB.1.5

Reproduce the Relative fitness Trends of 5 of the variants

Analyse Results

Paper project 5

- Very nice paper, with clear method.
- Data is abundant (both for calibrate the model and testing the model):
  - Model calibration:
  - Model testing:
    - Germany genomic surveillance data for SAR-CoV-2 (sample of 607,000 sequences, from 2021.07.01 to 2024.07.01) → predict immune dynamics bw 2022.03.01 and 2024.07.01 (2021.07.01 - 2022.03.01: burn-in initial immunological landscape)

Methods & Results

Installation

R:

- make sure you have R
- Install devtools install.packages("devtools")
- Install APIs for outbreak: devtools::install_github("outbreak-info/R-outbreak-info")

Python:

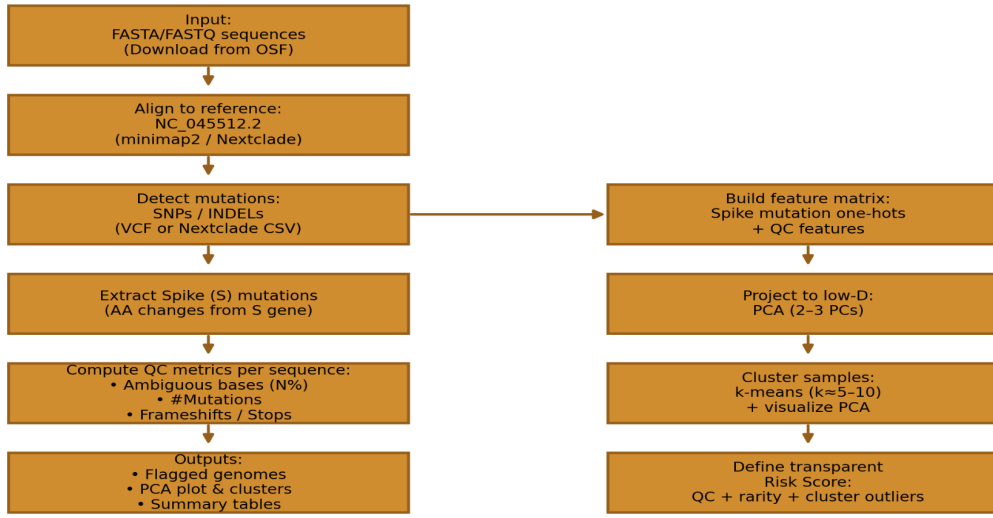
- snakemake
- envs/workshop

The environment setup file is available on our project GitHub repository.

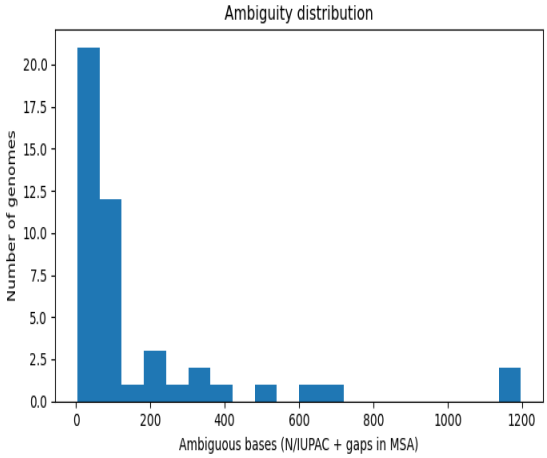
in terminal:
sudo apt-get install -y build-essential libcurl4-openssl-dev libssl-dev libxml2-dev pkg-config
sudo apt-get install -y build-essential pkg-config libfreetype6-dev libfontconfig1-dev libharfbuzz-dev libfribidi-dev libpng-dev libjpeg-dev libtiff5-dev libwebp-dev
pkg-config --modversion freetype2 harfbuzz fribidi libpng libjpeg libtiff-4 libwebp # To check if packages install successfully?
sudo apt-get install -y libudunits2-dev libgdal-dev libgeos-dev libproj-dev
sudo apt-get install -y cmake libabsl-dev libprotobuf-dev protobuf-compiler

in R:
install.packages(c("curl","openssl","credentials","urlchecker","rversions"))
install.packages(c("systemfonts","textshaping","ragg"))
install.packages("pkgdown")
install.packages("devtools") # To check if something successfully installed or not, use: packageVersion("ragg"), or whatever packages you want to check
install.packages(c("units", "s2", "sf"))
devtools::install_github("outbreak-info/R-outbreak-info")

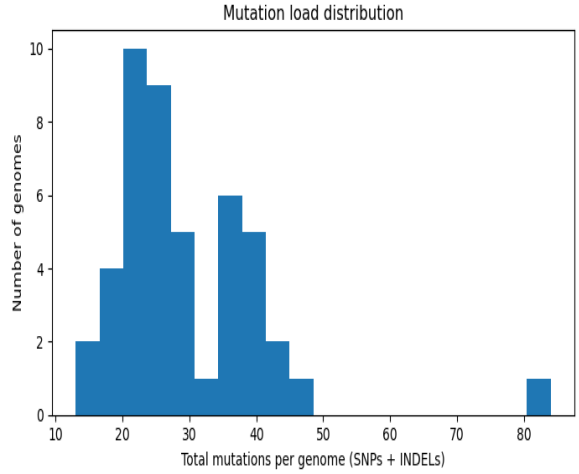
Project 3 Outline



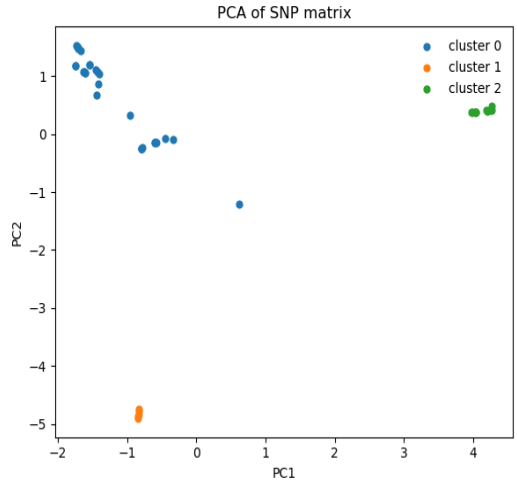
Results

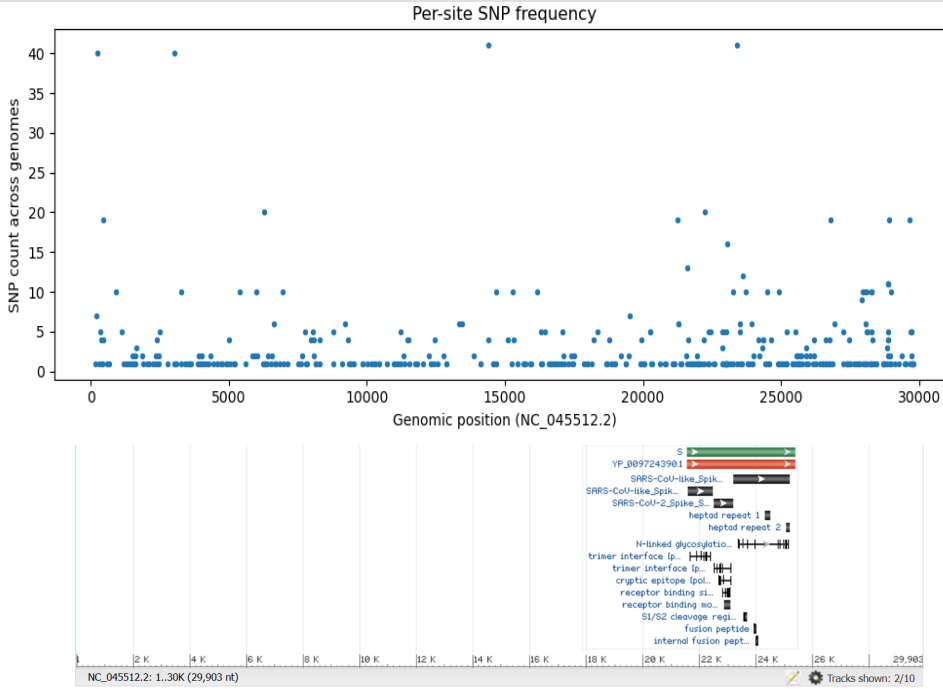

Ambiguity distribution
Number of genomes
Ambiguous bases (N/IUPAC + gaps in MSA)
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
20.0
0
200
400
600
800
1000
1200

t3


Mutation load distribution
Number of genomes
Total mutations per genome (SNPs + INDELs)
0
2
4
6
8
10
10
20
30
40
50
60
70
80

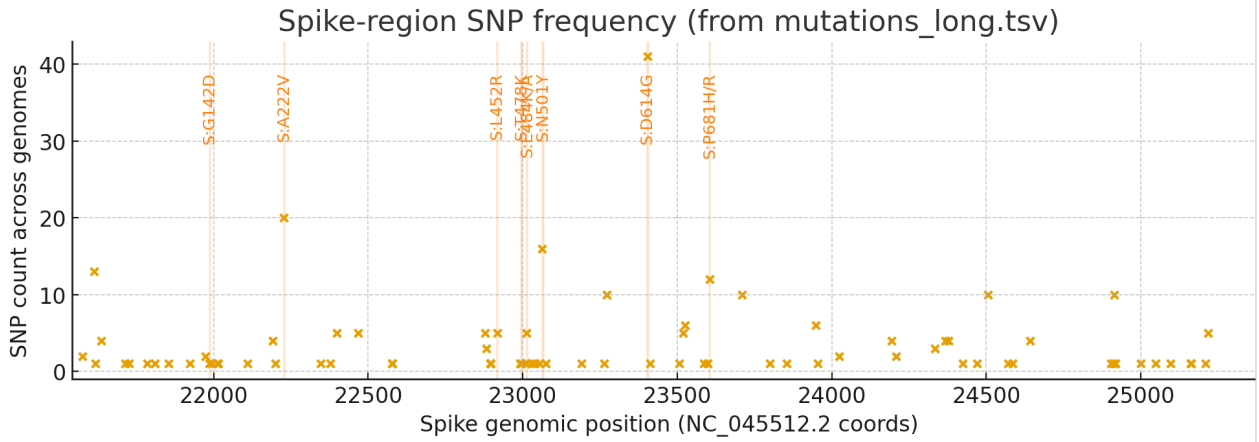
PCA and Clustering




Per-site SNP frequency
SNP count across genomes
Genomic position (NC_045512.2)
0
5000
10000
15000
20000
25000
30000
0
5
10
15
20
25
30
35
40

S
YP_009724390.1
SARS-CoV-like_Spik...
SARS-CoV-like_Spik...
SARS-CoV-2_Spike_S...
heptad repeat 1
heptad repeat 2
N-linked glycosylatio...
trimer interface (p...
trimer interface (p...
cryptic epitope (pol...
receptor binding si...
receptor binding mo...
S1/S2 cleavage regi...
fusion peptide
internal fusion pept...
1
2 K
4 K
6 K
8 K
10 K
12 K
14 K
16 K
18 K
20 K
22 K
24 K
26 K
29,903
NC_045512.2: 1..30K (29,903 nt)
Tracks shown: 2/10

Spike-region SNP



Top spike mutations

D614G- aspartic acid changes to glycine.

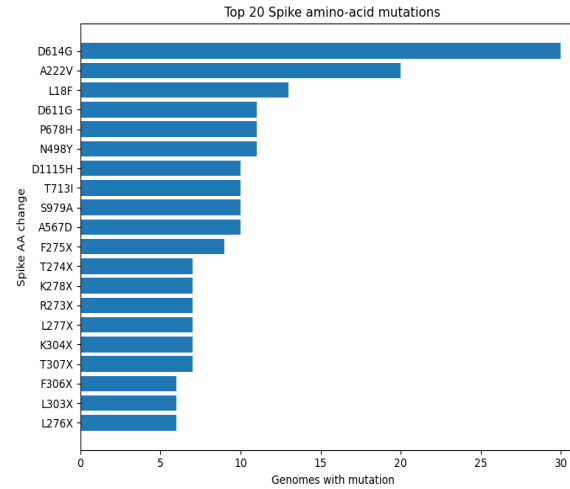
- Dominant background (all VOCs) inherited(positive selection).
  Higher infectivity/transmissibility

A222V- not robust lineage marker

LIBF - NTD antigenic drift

P681H/R (furin site) and RBD mutations.

-



Risk score

risk(sample) = Σ[1 per Spike AA change] + Σ[+1 extra if site in RBD (aa319–541)] + (Bloom escape weight per site, if loaded) − 0.1 × ambiguous bases

WHY? fast flag for genomes carrying clusters of Spike/RBD mutations; not a clinical risk.

Genome-wide SNP burden: median 26.0 (IQR 21.0 32.0)

Spike SNPs: median 5.0 (IQR 3.0 8.0)

Flagged (high risk): 5 genomes (90th percentile)

Project 5

We will investigate 5 variants of SAR-Cov-2:

- ★ JN.1
- ★ JN.2
- ★ JN.3
- ★ KP.3
- ★ XBB.1.5

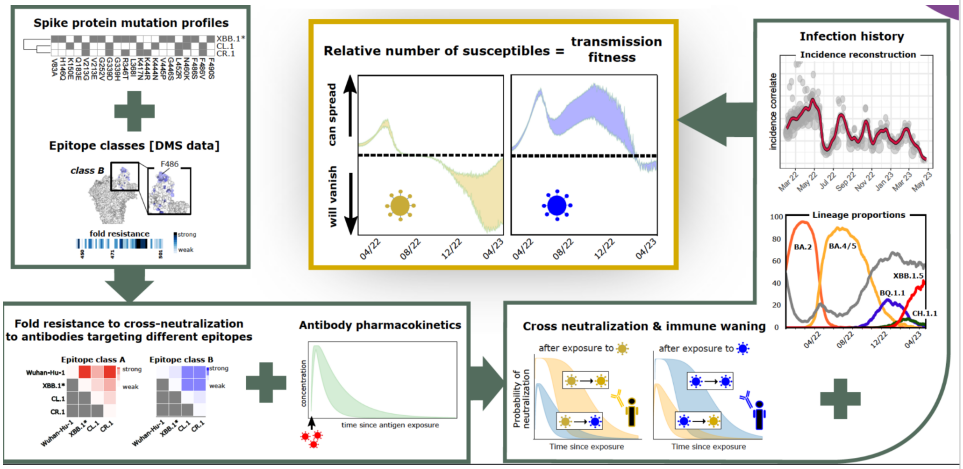


Image from our honorable professor and TAs teaching this course :) from the presentation on Immune Waning (day 3)

Link: here (You need to log in to your Whiteboard account to see it)

Overview on the data

Data Source	Purpose
Deep Mutational Scanning (DMS)	Tells how each spike mutation changes antibody binding (escape fraction).
Antibody pharmacokinetics	Describes how antibody levels rise and decay after infection or vaccination.
Genomic surveillance data	Shows which variants circulated and when (infection history) (Germany only)
Incidence reconstruction (GInPipe + wastewater)	Estimates how many infections occurred at each time point.

mutation profiles → epitope features → cross-reactivity

mutation profiles → epitope features → cross-reactivity

Mathematical models

1) From escape fractions to a per-antibody binding probability

$$b_a(x,y) = \prod_{s \in \Omega(x,y)} \left(1 - ef_{s,a}\right) \qquad b_a(x,y) = \frac{c_a}{\mathrm{FR}_{x,y}(a) \cdot \mathrm{IC}_{50(\mathrm{DMS})}(a) + c_a}$$

2) Convert binding to fold resistance

$$\mathrm{FR}_{x,y}(a) = \frac{c_a}{\mathrm{IC}_{50(\mathrm{DMS})}(a)} \left(\frac{1}{b_a(x,y)} - 1\right)$$

3) Aggregate across epitope classes

$$\mathrm{FR}_{x,y}(\vartheta) = \mathrm{mean}(\mathrm{FR}_{x,y}(a) : a \in \vartheta) \qquad \mathrm{FR}_{x,y}(\mathrm{NTD}) = 10^{|\Omega(x,y)|}$$

Full information in the Method part of Raharinirina et al. (2025) paper are very interesting and worth reading. We just extract some few important part to understand the results we present.

Mathematical models

4) Variant cross-neutralization probability

$$b_\vartheta(t,x,y)=\frac{c_\vartheta(t)}{\mathrm{FR}_{x,y}(\vartheta)\cdot \mathrm{IC}_{50(x)}(\vartheta)+c_\vartheta(t)}$$

$$P_{\mathrm{Neut}}(t,x,y)=1-\prod_{\vartheta\in A_{x,y}}\left(1-b_\vartheta(t,x,y)\right)$$

$$\widehat{\mathrm{IC}}_{50(x)}(\vartheta)=D(\vartheta)\cdot\widehat{\mathrm{IC}}_{50(x)}$$

5) Antibody potency

$$D(\vartheta)=\frac{\widehat{\mathrm{IC}}_{50(\mathrm{DMS})}(\vartheta)}{\frac{1}{|\mathcal{A}|}\sum_{\zeta\in\mathcal{A}}\widehat{\mathrm{IC}}_{50(\mathrm{DMS})}(\zeta)}$$

6) Expected sterilizing immunity against variant y

$$\mathbb{E}[\mathrm{Immune}_y(t)]=\sum_{x\in\mathcal{X}}\int_0^t \pi_x(s)\cdot I(s)\cdot P_{\mathrm{Neut}}(t-s,x,y)\,ds.$$

7) Variant dynamics (Fitness)

$$\gamma_y(t)=\frac{\alpha_y\,\mathbb{E}[S_y(t)]-\sum_{x\in\mathcal{X}}\pi_x(t)\,\alpha_x\,\mathbb{E}[S_x(t)]}{\sum_{x\in\mathcal{X}}\pi_x(t)\,\alpha_x\,\mathbb{E}[S_x(t)]}.$$

Figure 2c

receptor-binding domain (RBD):

10 class of epitopes (from 836 antibodies)

amino-terminal domain (NTD)

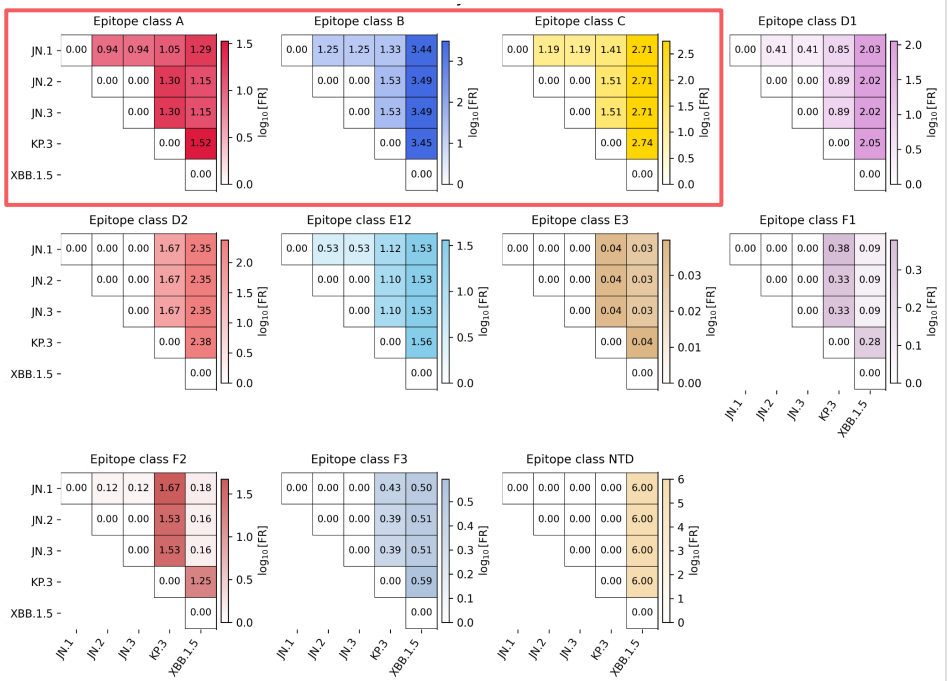


Figure 3A

We ran the simulation to get the predicted relative fitness

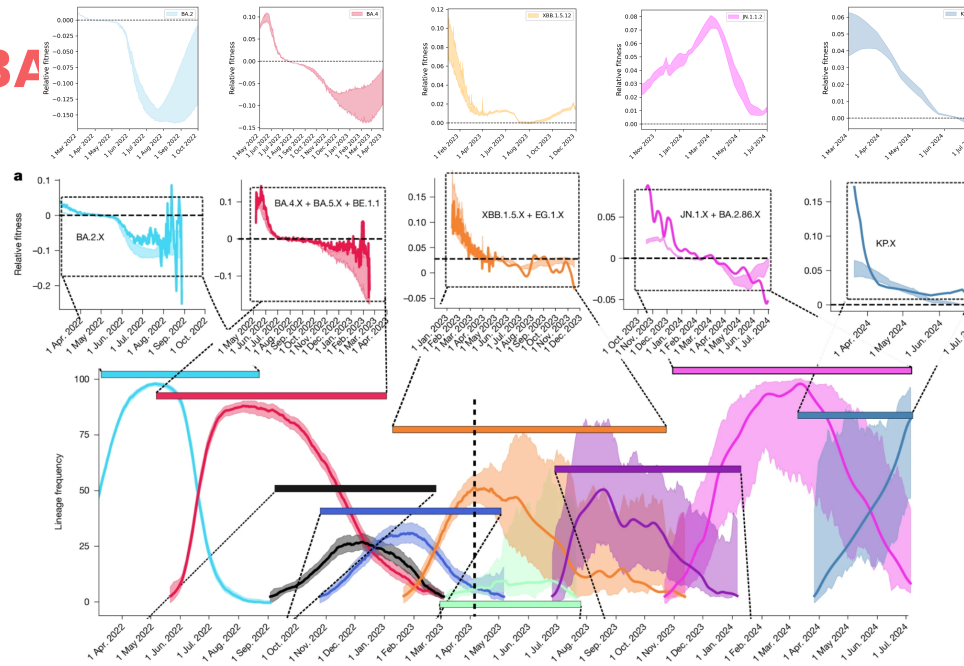


Figure 3A - real data

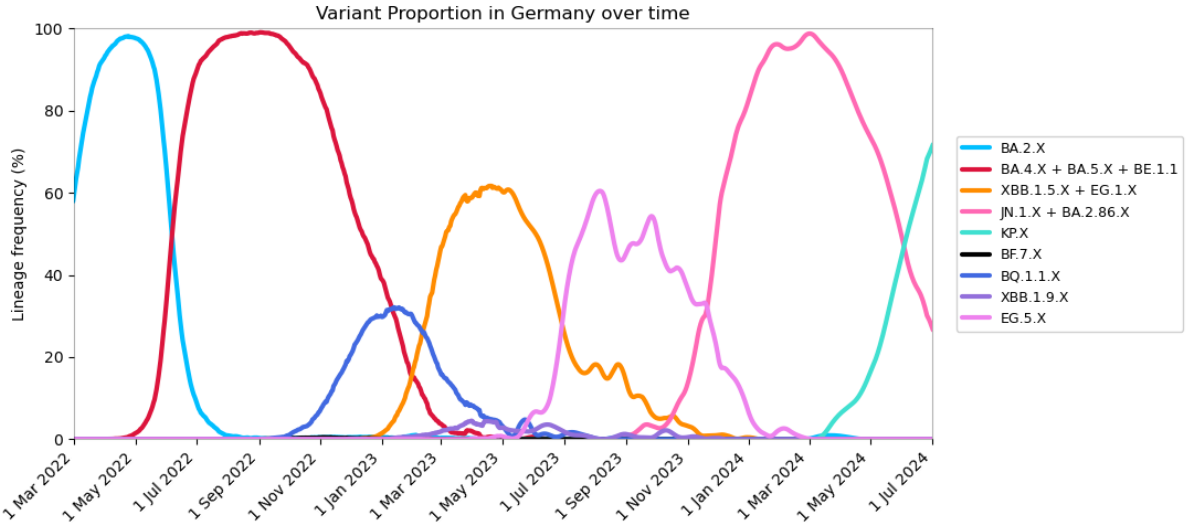


Figure 3B: We need more data

Need another step running via Covsonar.

Discussion

Discussion

Rising “risk” largely reflects variants better escaping population immunity rather than intrinsic virulence.

limits: small cohort and ambiguity (Ns) can bias clusters.

Future direction: Map clusters → Pango lineages using a marker panel for clearer labels.

It means your **risk score flags genomes with mutation patterns likely to spread better in an immune-exposed population** (e.g., RBD changes that reduce antibody binding), not ones that necessarily cause more severe disease. In other words, a higher score suggests **immune escape / transmission advantage**, **not** intrinsic virulence; you’d need clinical/outcome data to assess severity.

What can we know when we understand immunological landscape of SAR-CoV-2?

Early warning: Compute relative fitness (γ) → predict which variants rise (>0) or decline (<0) → Use spike mutations + DMS to estimate cross-neutralization and γ before case surges.

Vaccine guidance: Identify epitopes under pressure and immunity gaps to choose broader antigens.

This reveal a big landscape of virus evolution, here, SAR-CoV-2 in particular, and can assist the forecasting of new variants that have the risk to cause outbreak.

It means your **risk score flags genomes with mutation patterns likely to spread better in an immune-exposed population** (e.g., RBD changes that reduce antibody binding), not ones that necessarily cause more severe disease. In other words, a higher score suggests **immune escape / transmission advantage**, **not** intrinsic virulence; you'd need clinical/outcome data to assess severity.

Conclusion

For project 3:

PCA cleanly separates genomes by Spike mutation patterns; top loadings hit known hot-spots.

For project 5:

Timing: Anticipate inflection points and wave timing from changing susceptibility.

Regional risk: Different infection histories → different local immune landscapes → country-specific success/failure.

Worth-to-mention: Nobel Prize 2025

Nobel Prize in Physiology or Medicine 2025 to

Mary E. Brunkow, Fred Ramsdell, and Shimon Sakaguchi on regulatory T cells:

“for their discoveries concerning peripheral immune tolerance”

https://www.nobelprize.org/prizes/medicine/2025/press-release/. 2025 Oc

Our Codes

GitHub:

https://github.com/vmeomeo/FUB-Covid19DS-Proj35/

References and Resources

Papers

Starr, T. N., Greaney, A. J., Addetia, A., Hannon, W. W., Choudhary, M. C., Dingens, A. S., ... & Bloom, J. D. (2021). Prospective mapping of viral mutations that escape antibodies used to treat COVID-19. Science, 371(6531), 850-854.

Raharinirina, N. A., Gubela, N., Börnigen, D., Smith, M. R., Oh, D. Y., Budt, M., ... & von Kleist, M. (2025). SARS-CoV-2 evolution on a dynamic immune landscape. Nature, 639(8053), 196-204.

Code repository

https://github.com/KleistLab/VASIL

https://github.com/KleistLab/GInPipe

https://github.com/rki-mf1/2025-SC2-Data-Science/

Additional plo

Data smoothing strategy