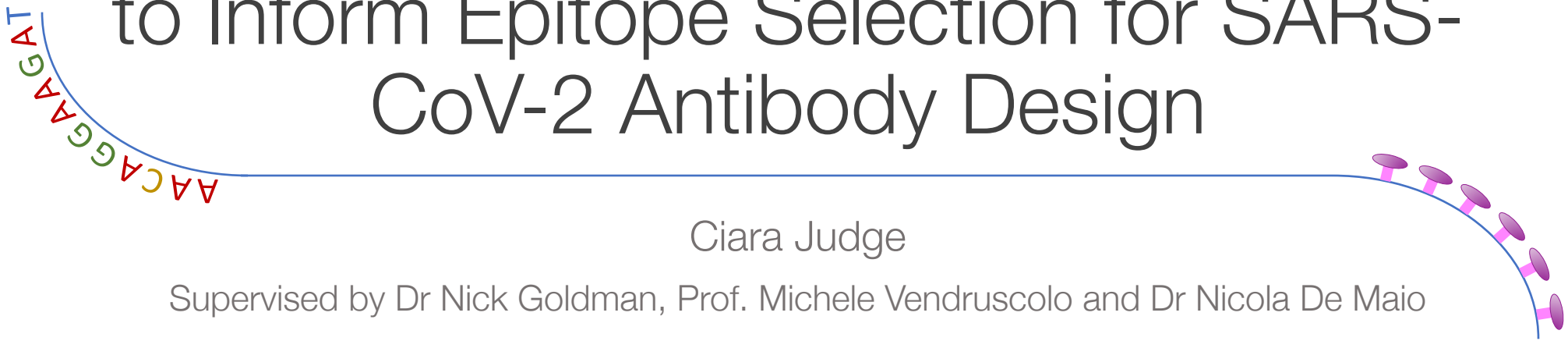
Methods for Using Evolutionary Analyses to Inform Epitope Selection for SARS-CoV-2 Antibody Design



Ciara Judge

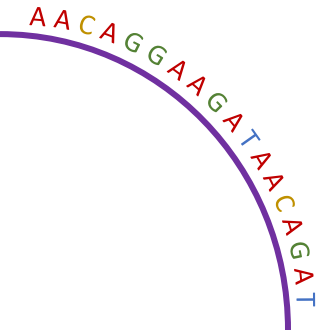
Supervised by Dr Nick Goldman, Prof. Michele Vendruscolo and Dr Nicola De Maio

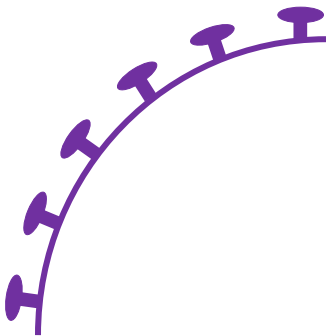
EMBL-EBI

UNIVERSITY OF CAMBRIDGE

The problem / opportunity

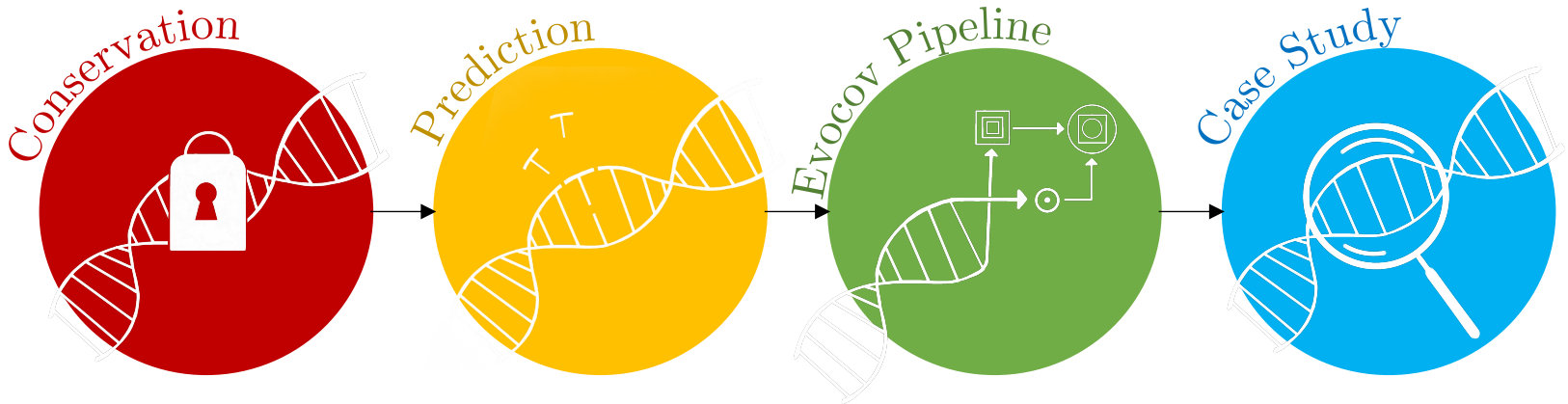
- SARS-CoV-2 evolution creates a challenge for therapeutic development
- Extended timescale from variant identification to therapeutic clinical trials
- Genomic surveillance data is vast and complex





Project Objectives

- Select conserved targets on spike for therapeutic intervention
- Predict the most likely mutations in the target regions
- Create a cohesive pipeline to automate the process

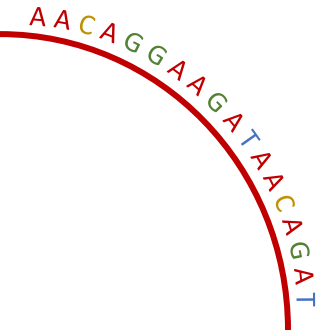


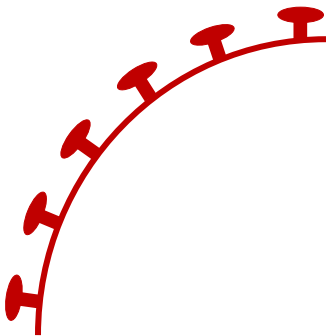
Conservation Analysis

Initial analysis and selection of the measure of conservation

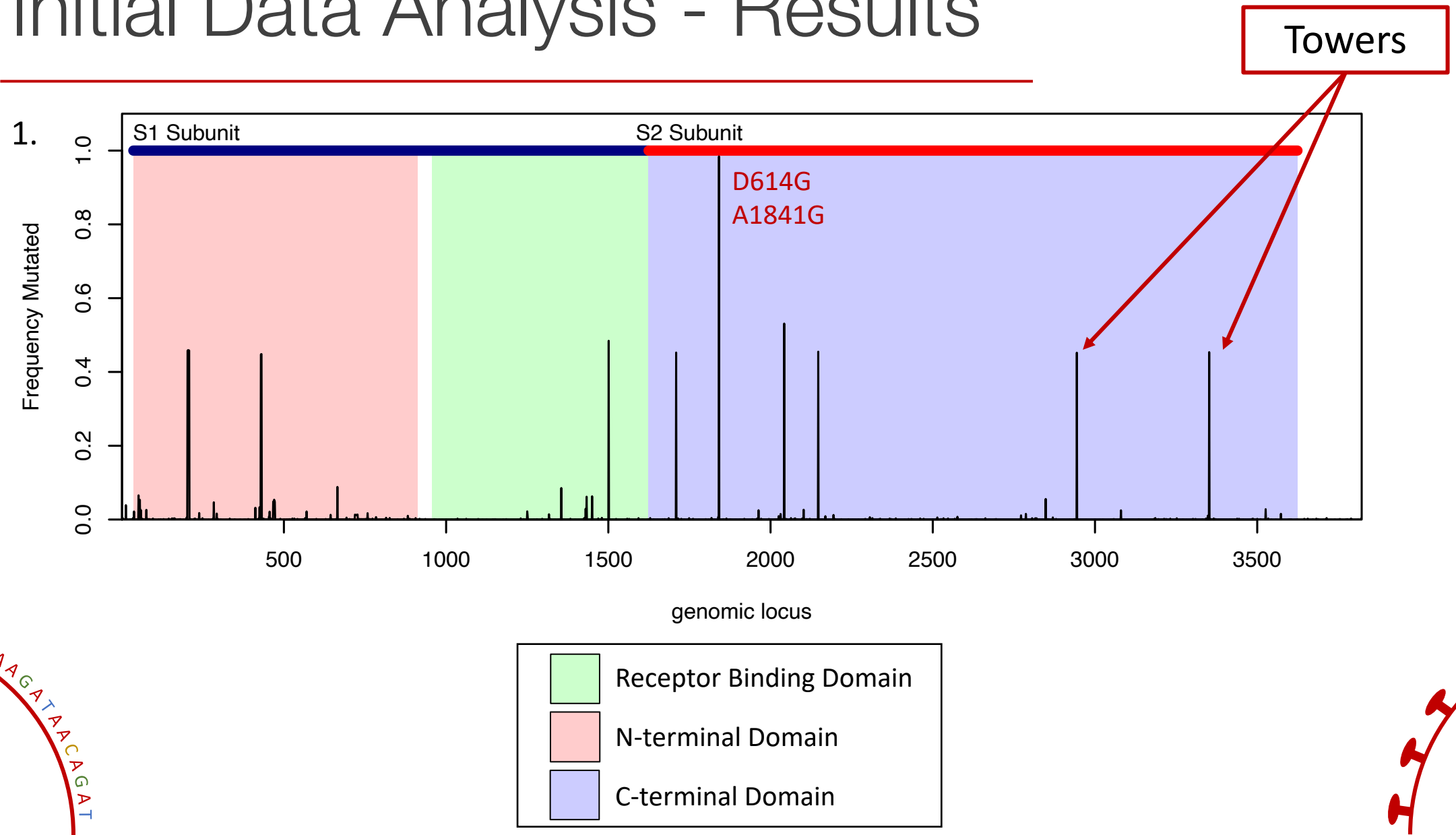
Initial Data Analysis

- 2.6 million SARS-CoV-2 genome sequences obtained from GISAID in fasta format
- Sequences matched to their metadata and differences from the reference were noted in terms of nucleotide and amino acid differences
- Frequency and nature of mutations counted and investigated
- Sequences also separated out based on their metadata (time, variant, country)



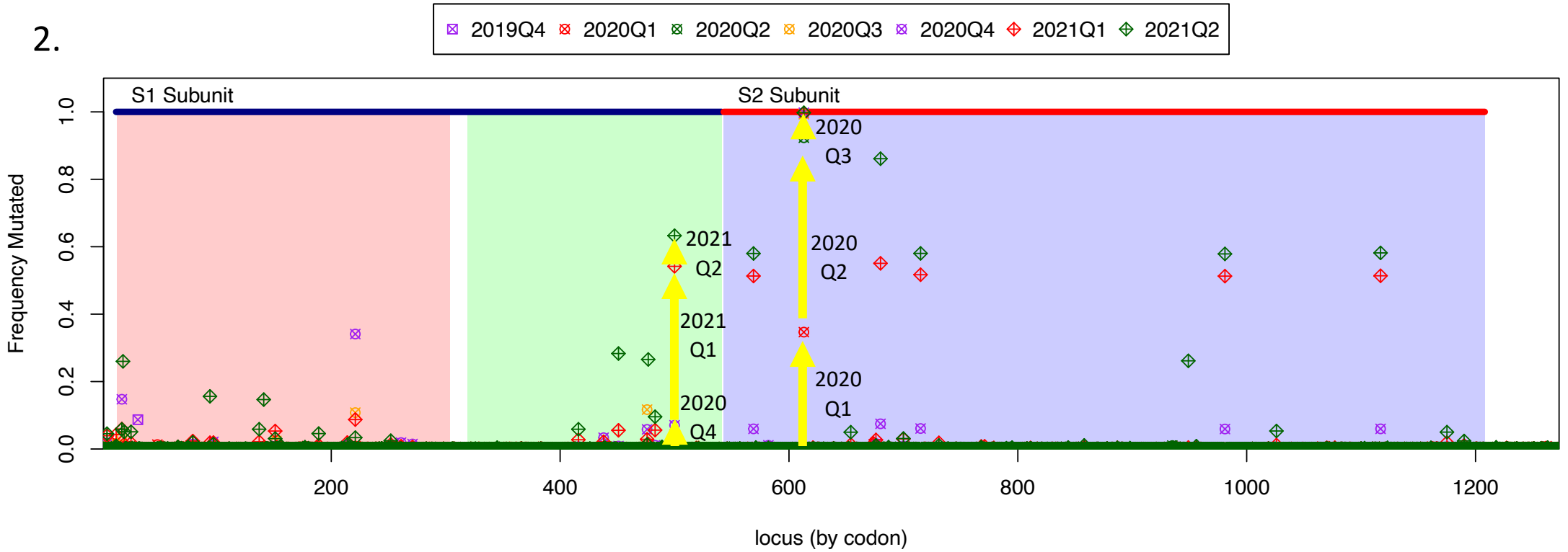


Initial Data Analysis - Results



Initial Data Analysis - Results

2.

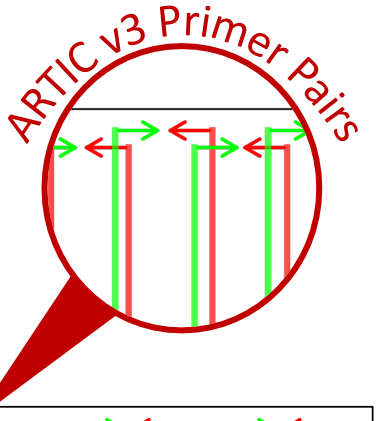


Receptor Binding Domain

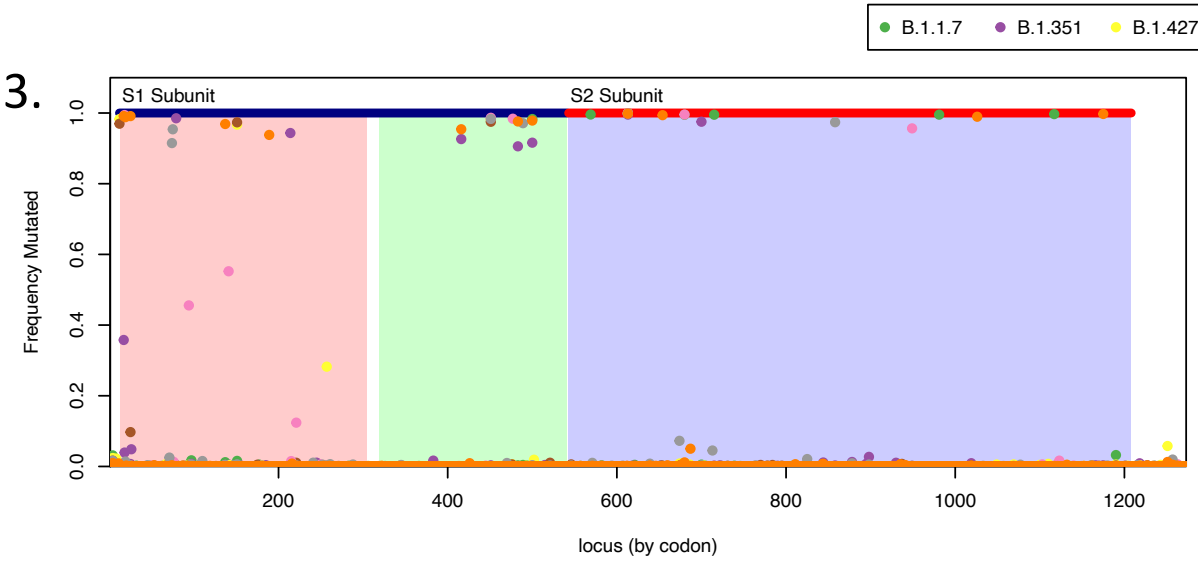
N-terminal Domain

C-terminal Domain

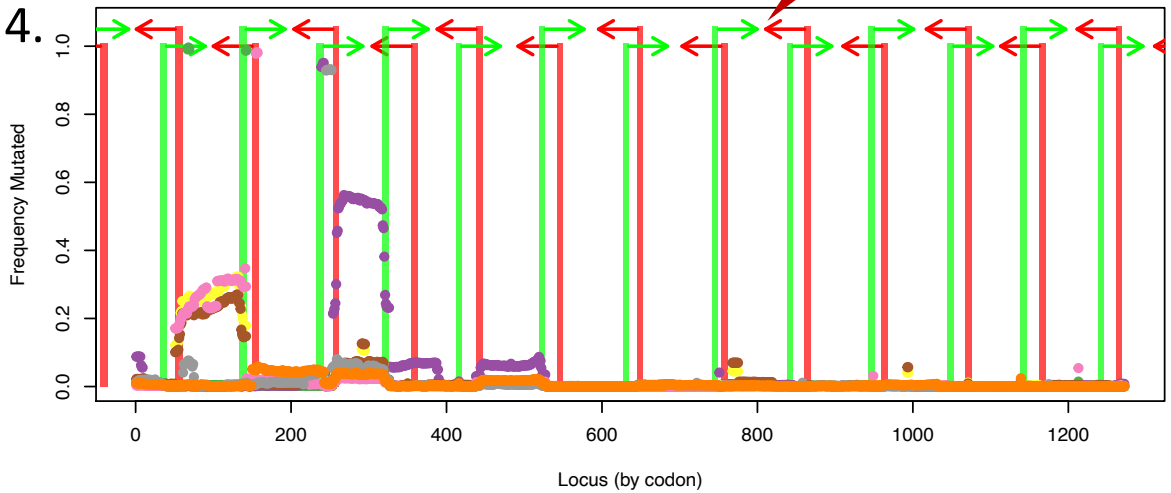
Initial Data Analysis - Results

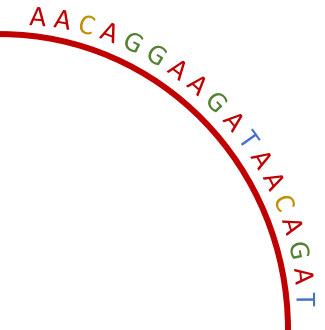


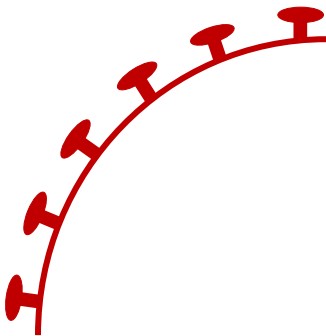
3.



4.

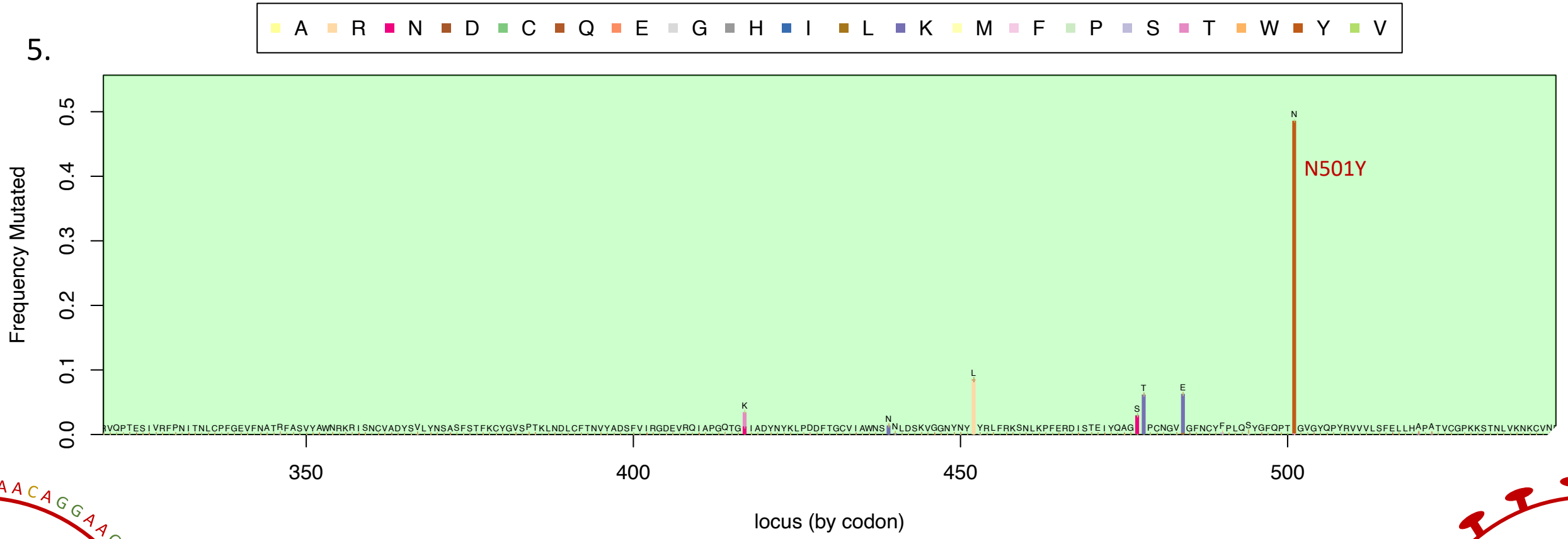






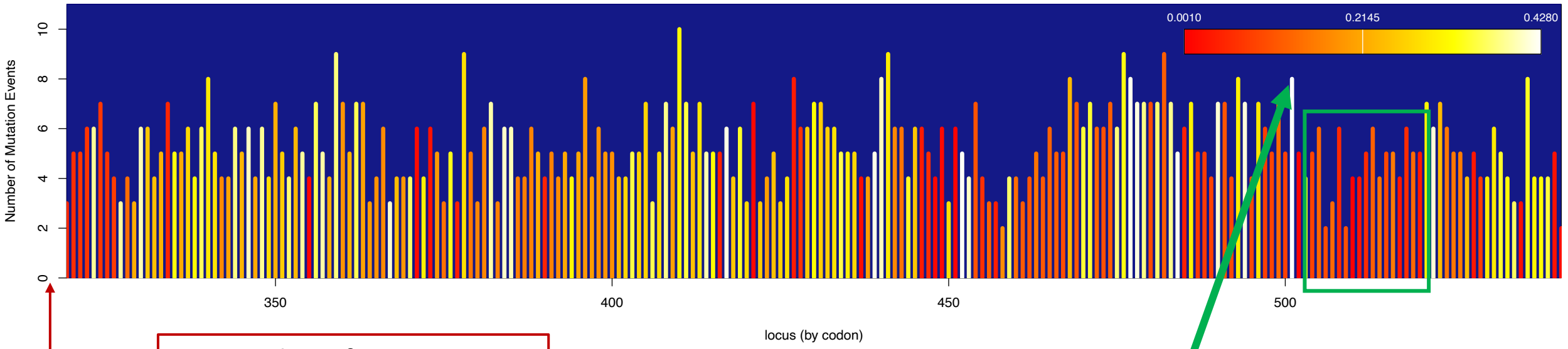
Initial Data Analysis - Results

5.



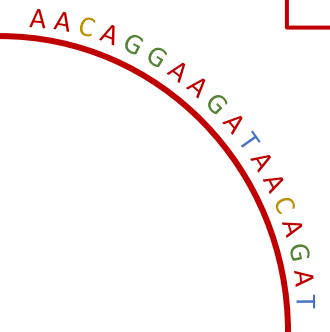
Initial Data Analysis - Results

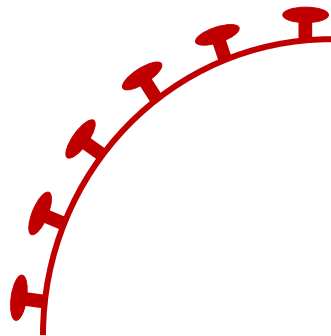
6.



Number of unique AA substitutions at the locus

Example: N501Y frequently different from reference (white colour) and 8 different AAs have been reported at the locus over the pandemic





Entropy

- Site-wise entropy (Equation 1) was used as a numerical measure of sequence conservation and consistency
- This captures two criteria: frequency of differences from the reference, and number of unique transitions at a locus

1

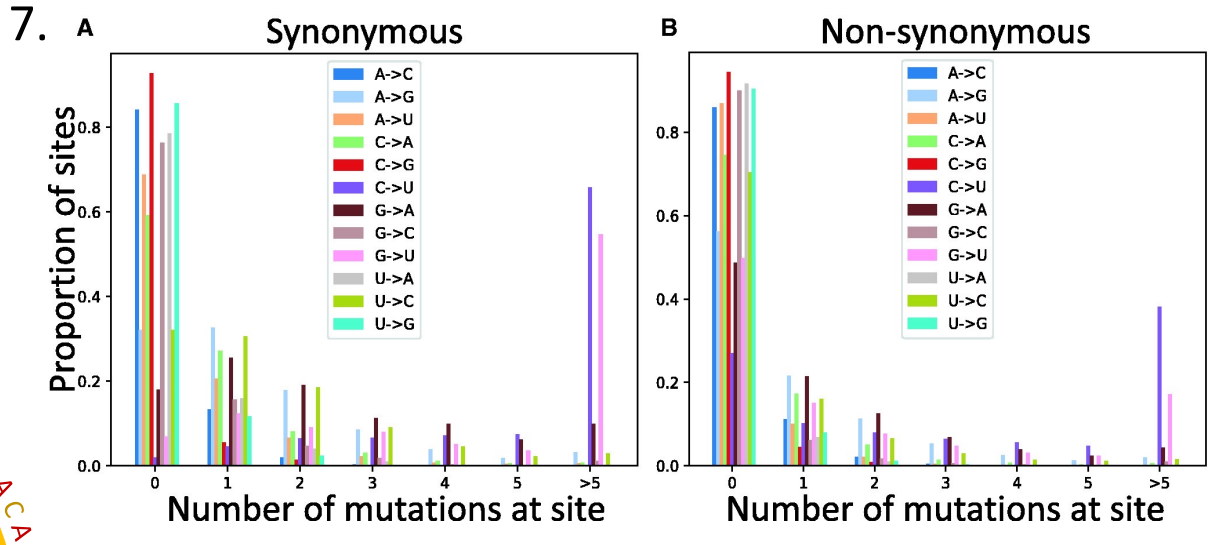
$$H(X) = -\sum_{i=1}^{N} P(x_i) \log P(x_i)$$

Prediction of Mutants

Explanation of the four approaches

Comparison to Neutral Substitution Rates

- Mutation frequencies calculated for all possible nucleotide substitutions at all loci encoding the epitope
- Ratio of these frequencies to corresponding neutral rates [1] calculated
- Highest ratio also encoding a non-synonymous mutation selected (Eqn 2)



2

$$\hat{n}, \hat{j} = \underset{n,j}{\operatorname{argmax}}\left(\frac{f_j^n}{\mu_{ij}}\right)$$

[1] De Maio, N. et. al (2021). Mutation Rates and Selection on Synonymous Mutations in SARS-CoV-2.

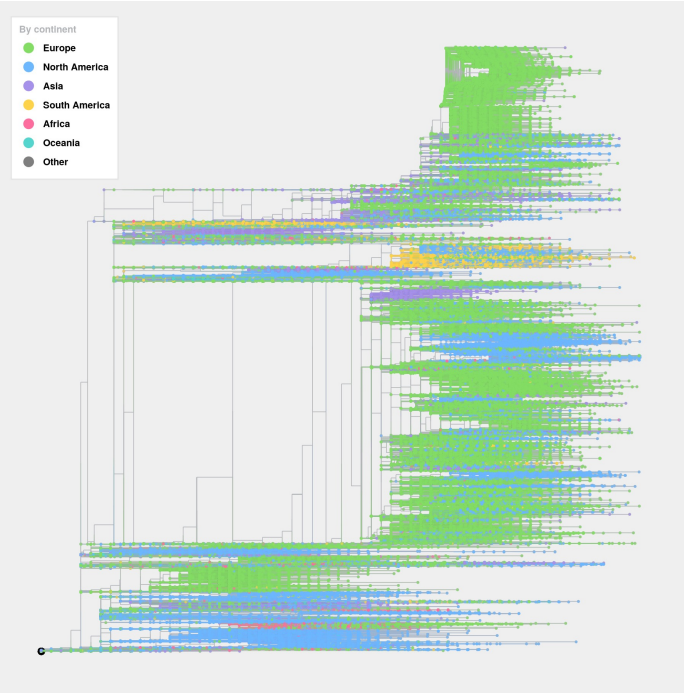
Estimated Mutation Rates

- Site-wise mutation rates calculated using baseml [2] – takes into account structure of phylogenetic tree
- Phylogenetic analysis of 600,000 sequence tree from GISAID

Treecov

1. Sample of 500 sequence labels taken from the large tree
2. Corresponding sequences taken from fasta file
3. Baseml analysis conducted for site-wise rates, which are saved
4. Repeat steps 1-3 100x

8.

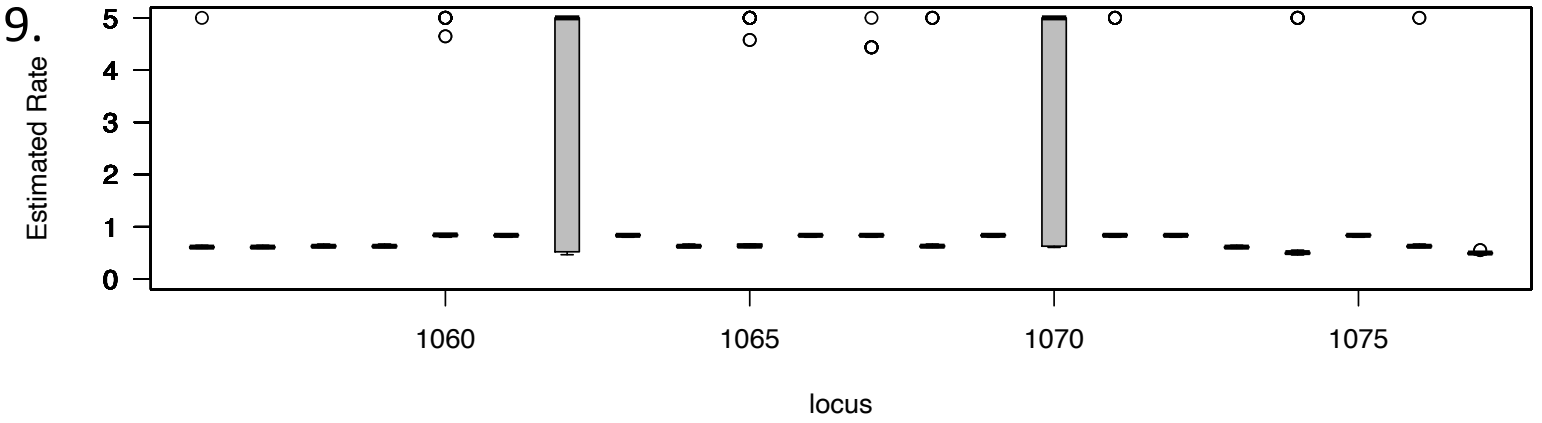


[2] Yang, Z. (2007). PAML 4: Phylogenetic Analysis by Maximum Likelihood..

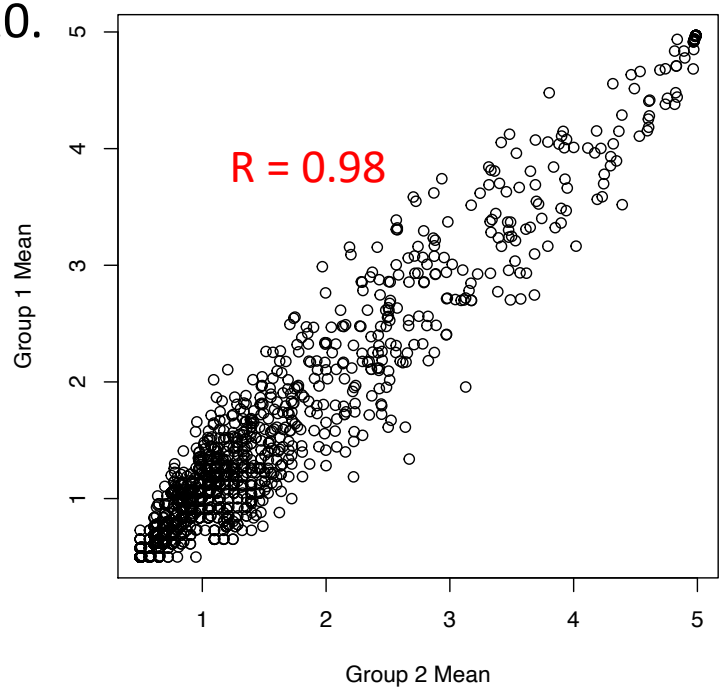
Estimated Mutation Rates

- Confirmed legitimacy of using baseml
- Used mutation rates to select locus, and ratio to neutral to select substitution (Eqn 3)

9.



10.



3

$$\hat{j} = \underset{j}{\operatorname{argmax}}\left(\frac{f_j^n}{\mu_{ij}}\right)$$

Normalised Global Frequency

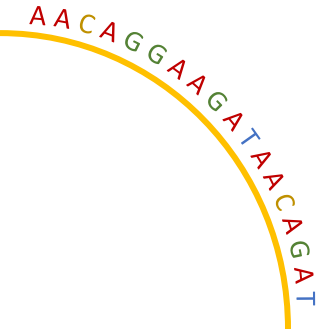
- Frequency of all variations of the epitope nucleotide sequence normalized using country case data obtained from the World Health Organisation

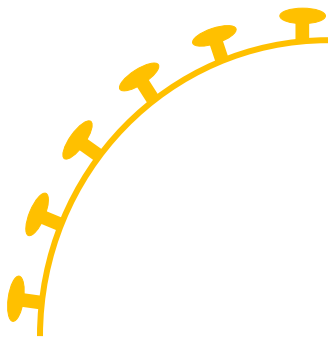
Selection Coefficient

- Equation for selection coefficient derived from a Hardy Weinberg model

4

$$S = \frac{1}{t}\log\left(\frac{p_t q_0}{q_t p_0}\right)$$





Evocov Pipeline

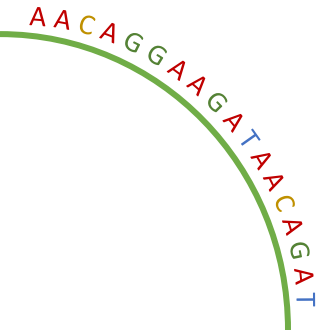
Pipeline to rank candidate epitopes based on conservation and predict new mutations

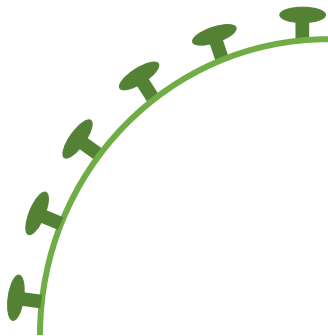
Flow of information

1. Parsing
2. Counting
3. Scoring
4. Prediction
5. Plotting

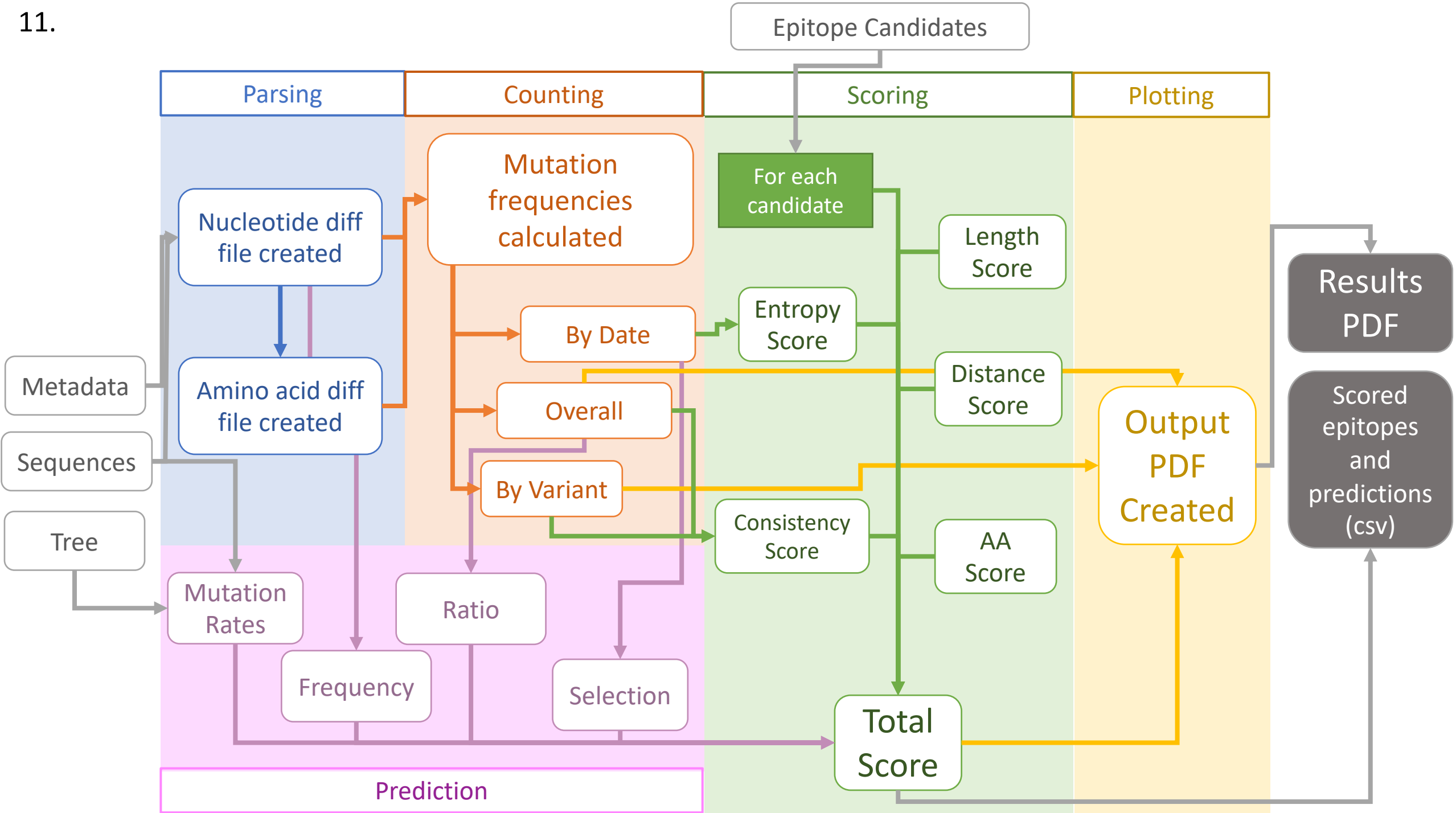
Other variables that contribute to an epitope's score

- Consistency (10pt)
- Entropy (65pt)
- Distance between residues (10pt)
- AA composition (10pt)
- Length (5pt)



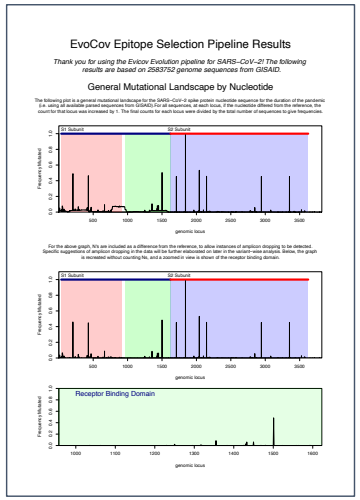


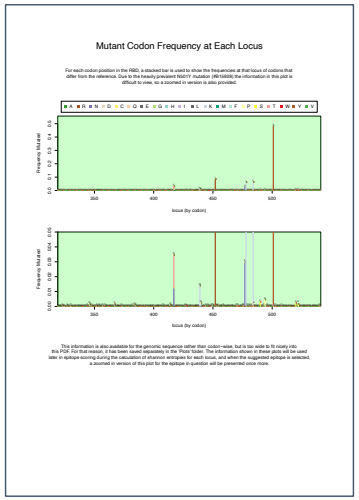
11.

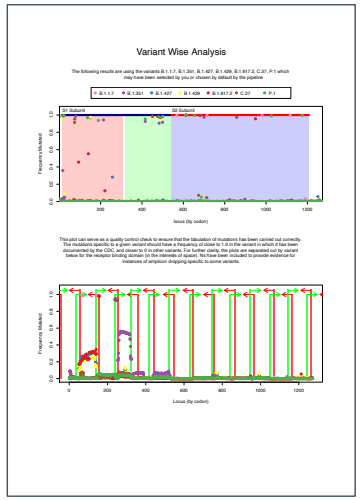

Epitope Candidates
Parsing
Counting
Scoring
Plotting
Nucleotide diff file created
Amino acid diff file created
Mutation frequencies calculated
By Date
Overall
By Variant
For each candidate
Length Score
Entropy Score
Distance Score
Consistency Score
AA Score
Metadata
Sequences
Tree
Mutation Rates
Ratio
Frequency
Selection
Total Score
Output PDF Created
Results PDF
Scored epitopes and predictions (csv)
Prediction

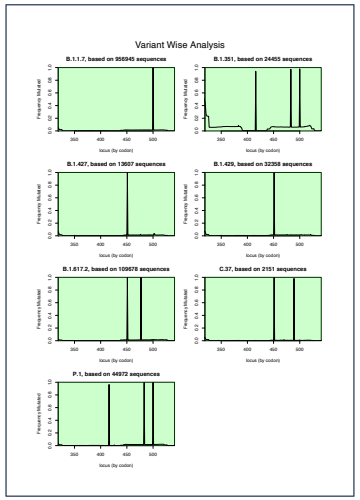
Automated Results File

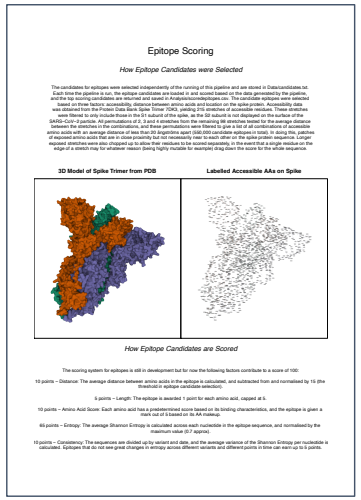
12.

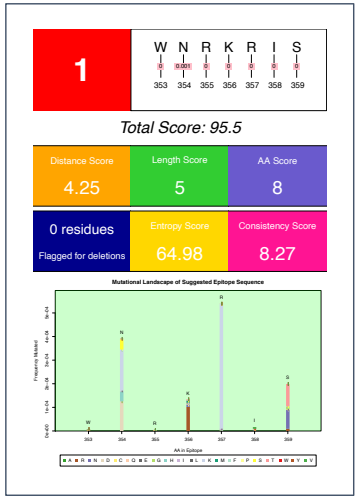


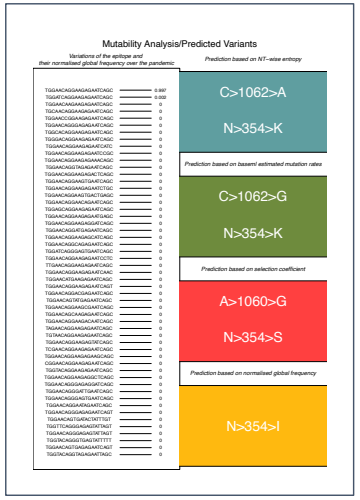


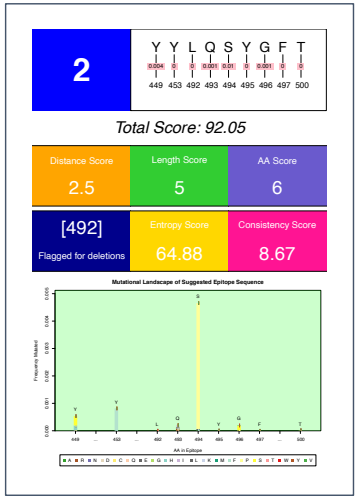


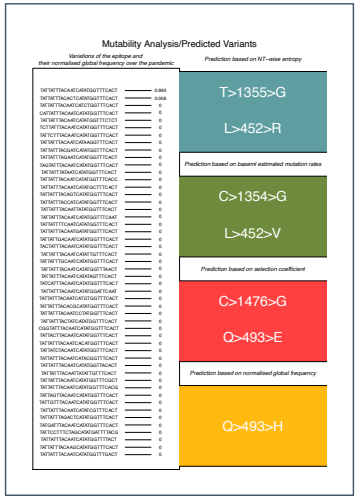






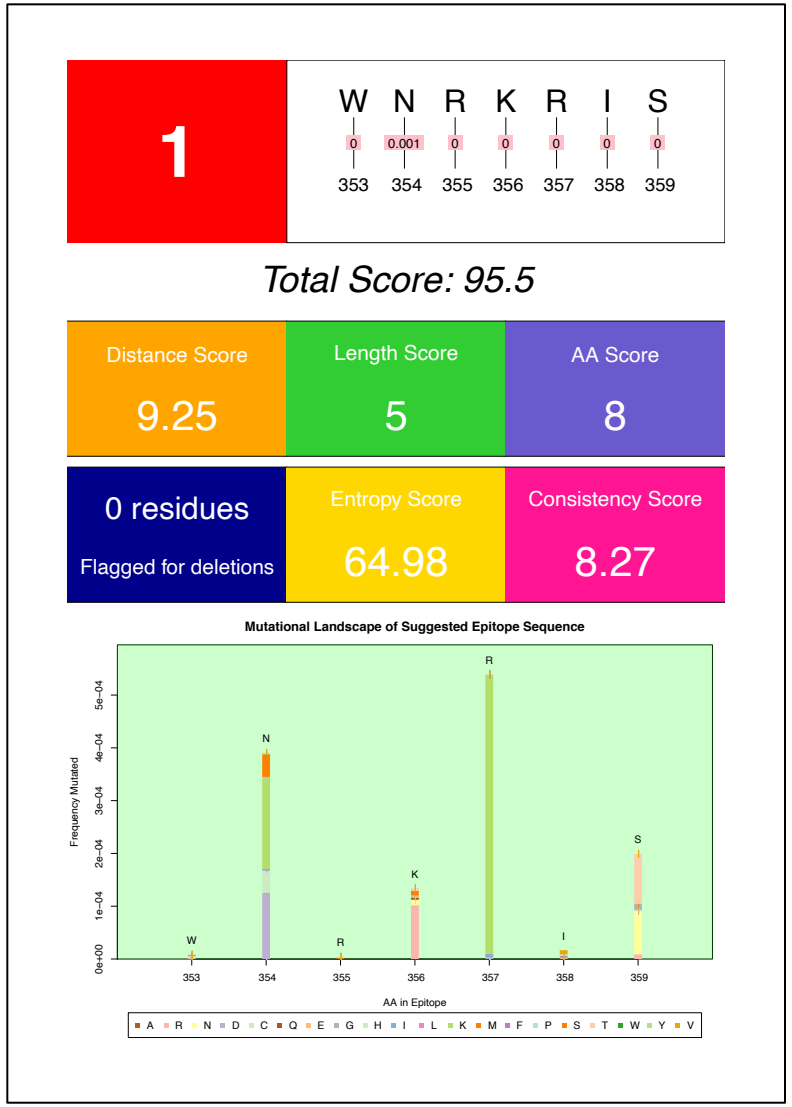


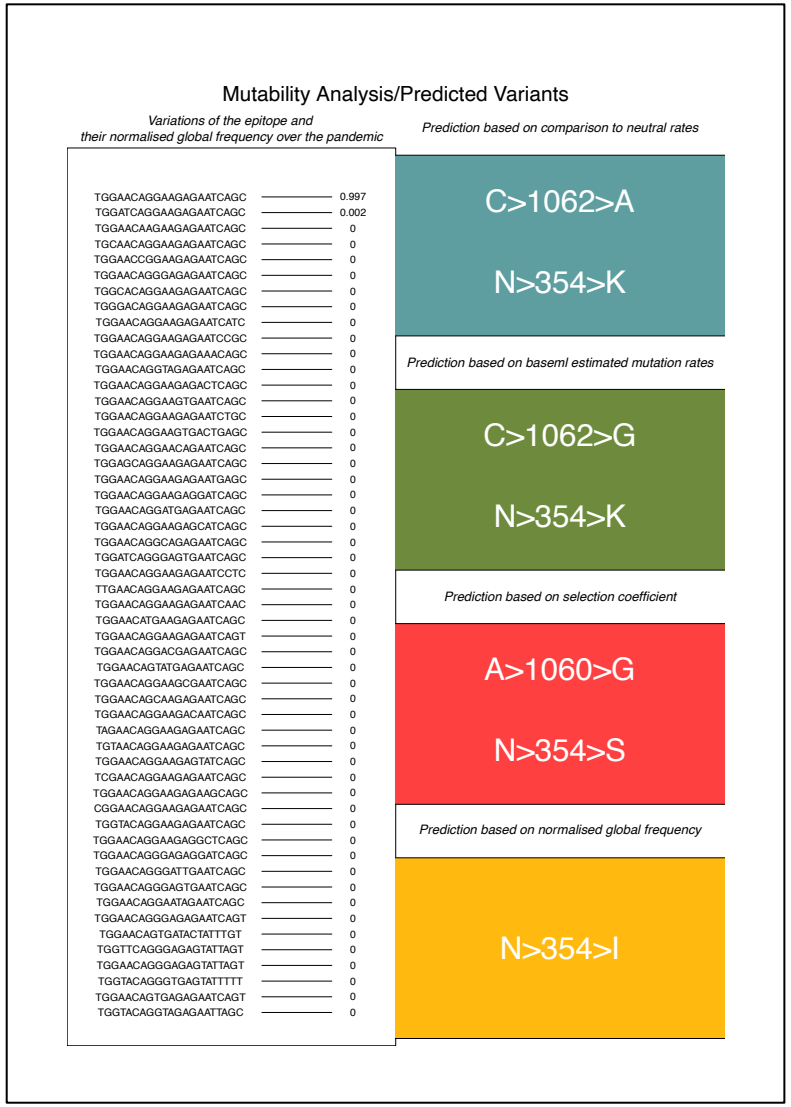




Example Epitope Breakdown

13.





Pipeline Case Study

Use of the pipeline of two pools of epitope candidates

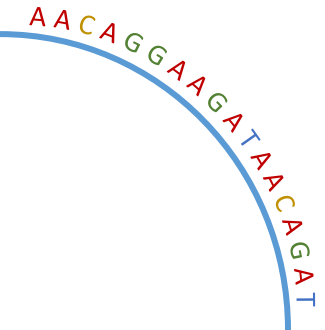
Candidates

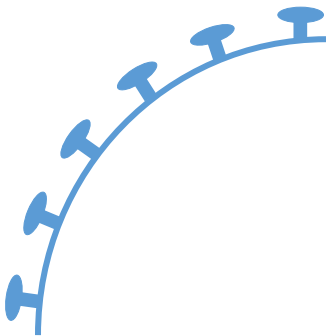
De Novo Candidates

- External residues on spike S1 tagged
- Consecutive stretches kept together
- Residue positions in 3D space obtained
- Permutations of sufficiently proximal exposed and accessible stretches chosen, yielding ‘patches’
- ~550,000 candidate sequences

Experimentally Validated Candidates

- Two experimentally validated epitopes from antibody design program [3]:
  - DesAb-RBD-1: WNRKRIS
  - DesAb-RBD-2: YYLQSYGFT
- Intended for prediction rather than scoring



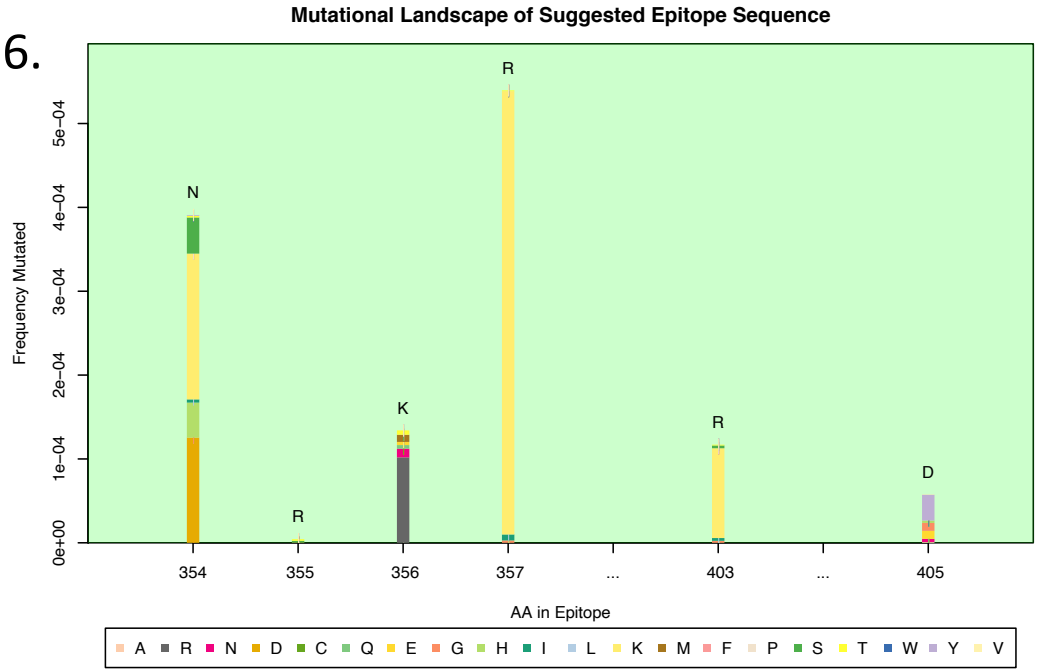


[3] Aguilar Rangel, M., et al (2021). Fragment-based computational design of antibodies targeting structured epitopes.

Epitope Scores

De novo epitope scores		
Epitope	Loci	Total
NRKRR	354,355,356,357,403	97.36
NRKRD	354,355,356,357,405	97.31
NRKRRD	354,355,356,357,403,405	97.13
KPDDE	424,426,427,428,471	96.52
ANRKR	352,354,355,356,357	96.46

16.

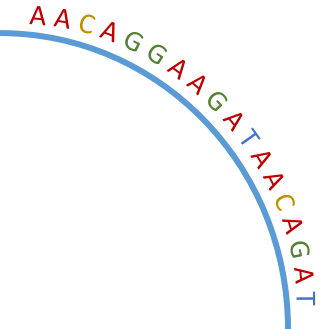


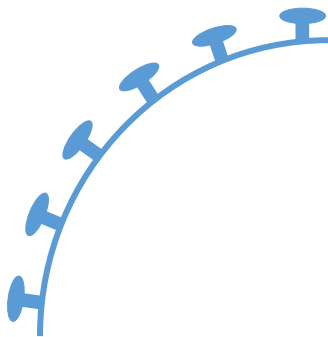
Experimentally validated epitope scores		
Epitope	Loci	Total
WNRKRIS	353,354,355,356,357,358,359	95.5
YYLQYSGFT	449,453,492,493,494,495,496,497,500	92.05

Predicted Mutations

De novo epitope mutation predictions				
Epitope	Frequency	Ratio to Neutral	Mutation Rate	Selection
NRKRR	N354I	N354K	N354K	N354S
NRKRD	N354I	N354K	N354K	N354S
NRKRRD	N354I	N354K	N354K	N354S
KPDDE	P426R	D427N	D427H	D427N
ANRKR	N354I	N354K	N354K	N354S

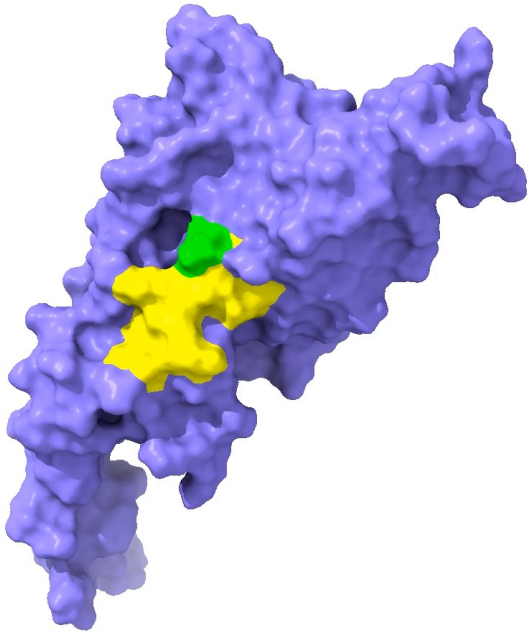
Experimentally validated epitope mutation predictions				
Epitope	Frequency	Ratio to Neutral	Mutation Rate	Selection
WNRKRIS	N354I	N354K	N354K	N354S
YYLQYSGFT	Q493H	L452R	L452V	Q493E



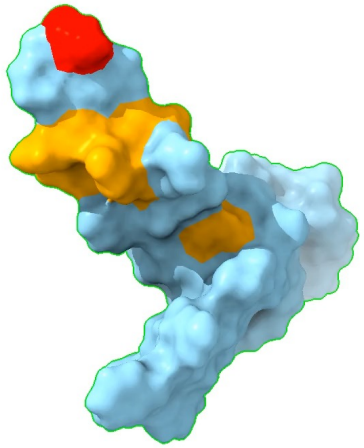


Mutant Models

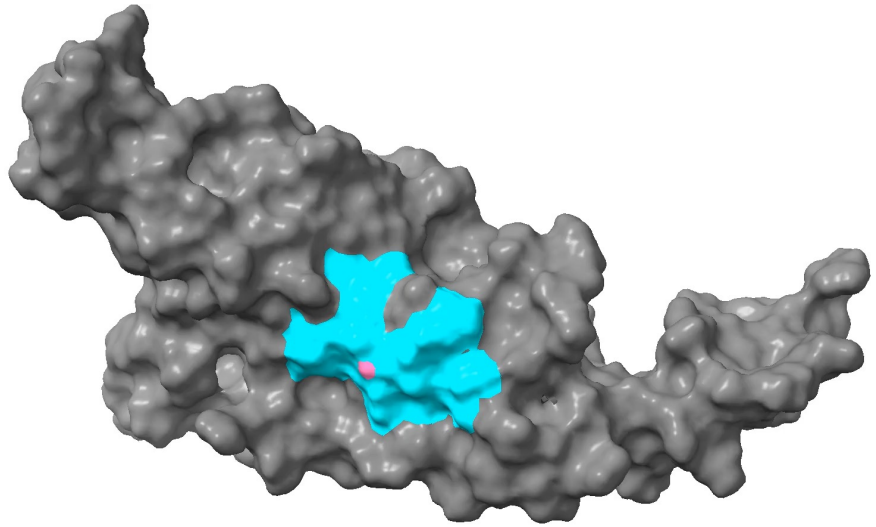
- Mutant models generated by I-TASSER and ChimeraX



N354I



Y449C



K355Q

Conclusions

- Ranking epitopes by conservation is valuable
- Pre-emptively designing therapeutics for COVID-19 may be possible
- Automating analysis and making results more accessible was successfully achieved: 1-2hr weekly update, user-friendly output

Recommendations

- Continue to outfit pipeline with features
- Experimentally validate antibody designs for predicted mutants
- Explore potential of mRNA vaccine applicability
- Apply this concept to other disease agents

Acknowledgements

Dr Nick Goldman & Dr Nicola De Maio

~

Professor Michele Vendruscolo

~

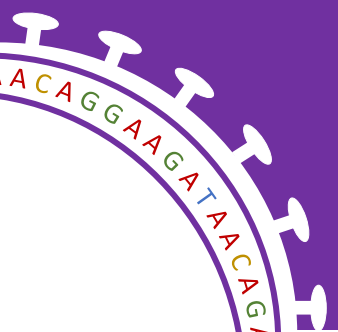
European Bioinformatics Institute

~

University of Cambridge

~

Coffee, red bull and spotify



Thank you for your attention! Any questions?