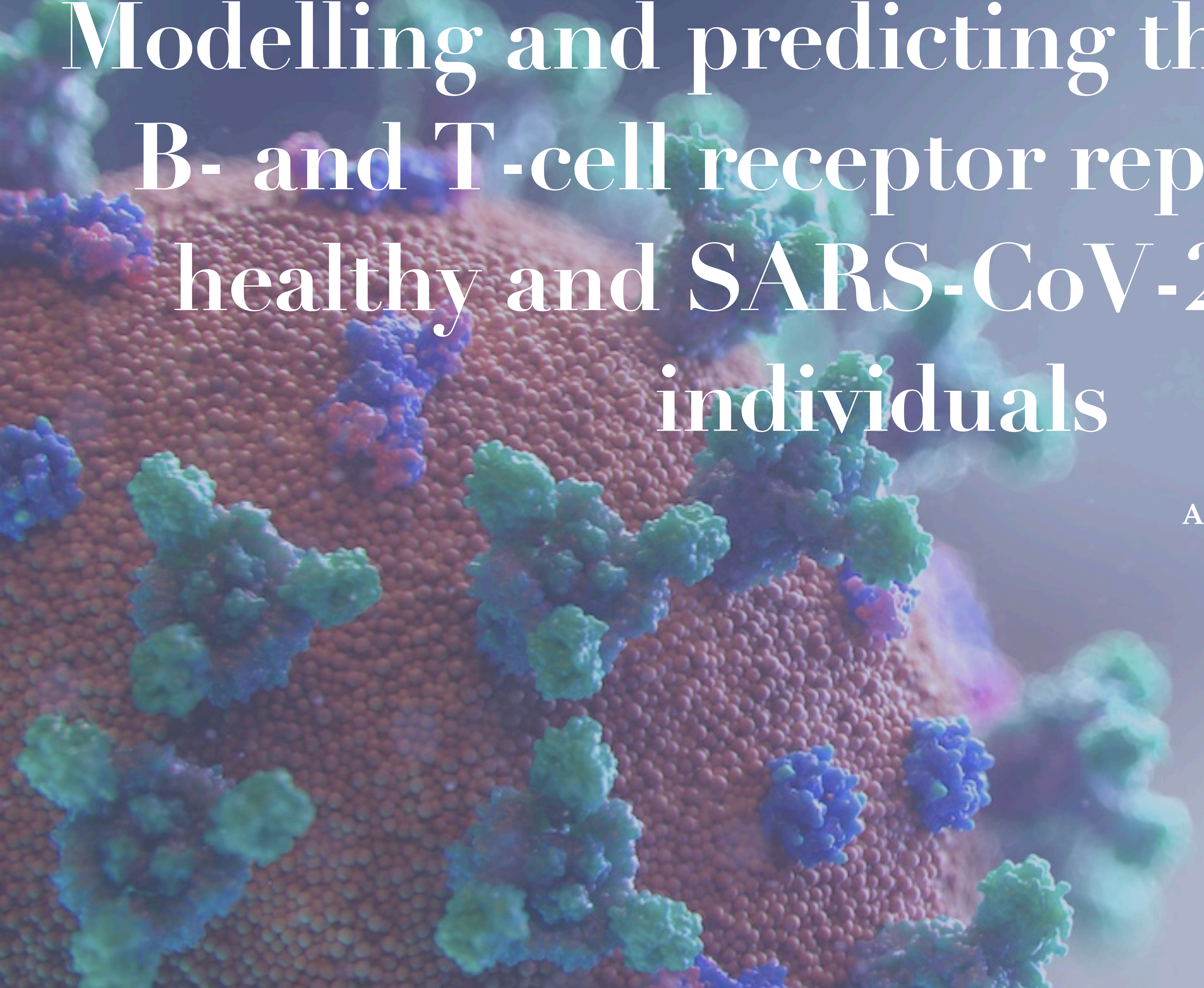


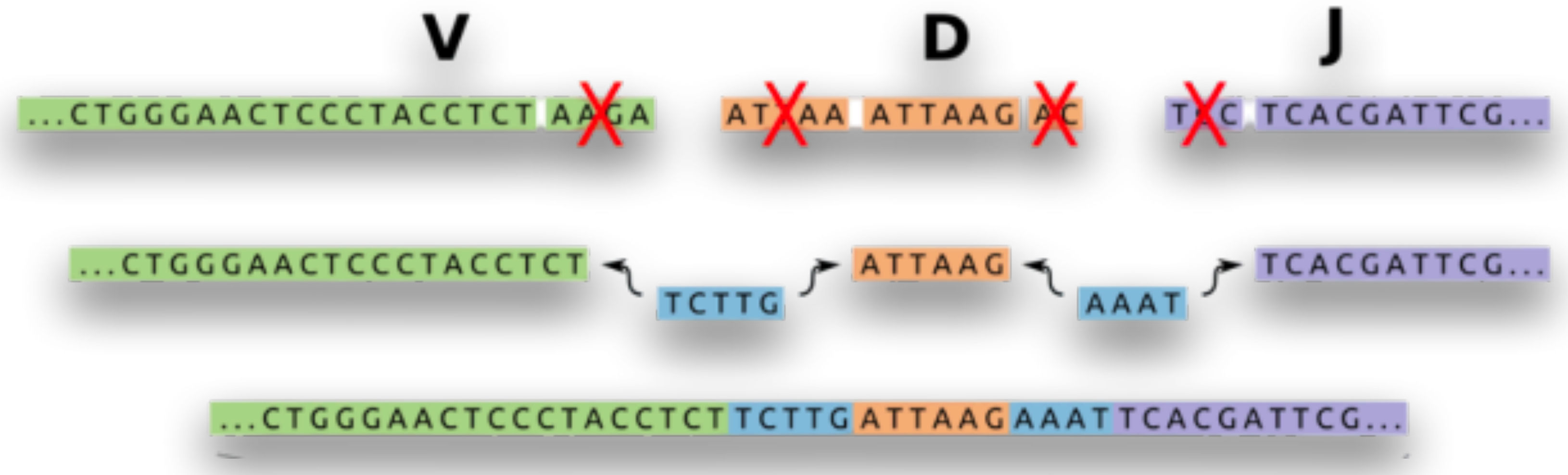
Modelling and predicting the overlap of B- and T-cell receptor repertoires in healthy and SARS-CoV-2 infected individuals

María Ruiz Ortega,

Aleksandra Walczak and Thierry Mora

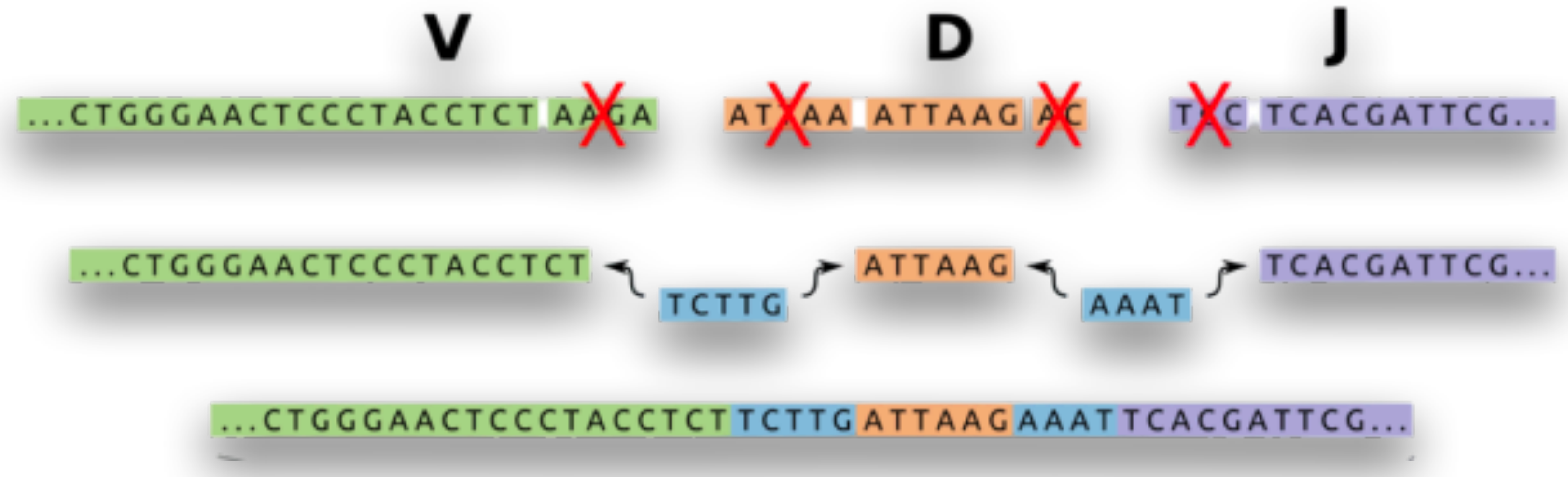
Rencontre des Jeunes Physicien.ne.s 2022

INTRODUCTION



PLOS Computational Biology | DOI:10.1371/journal.pcbi.1004409 January 11, 2016

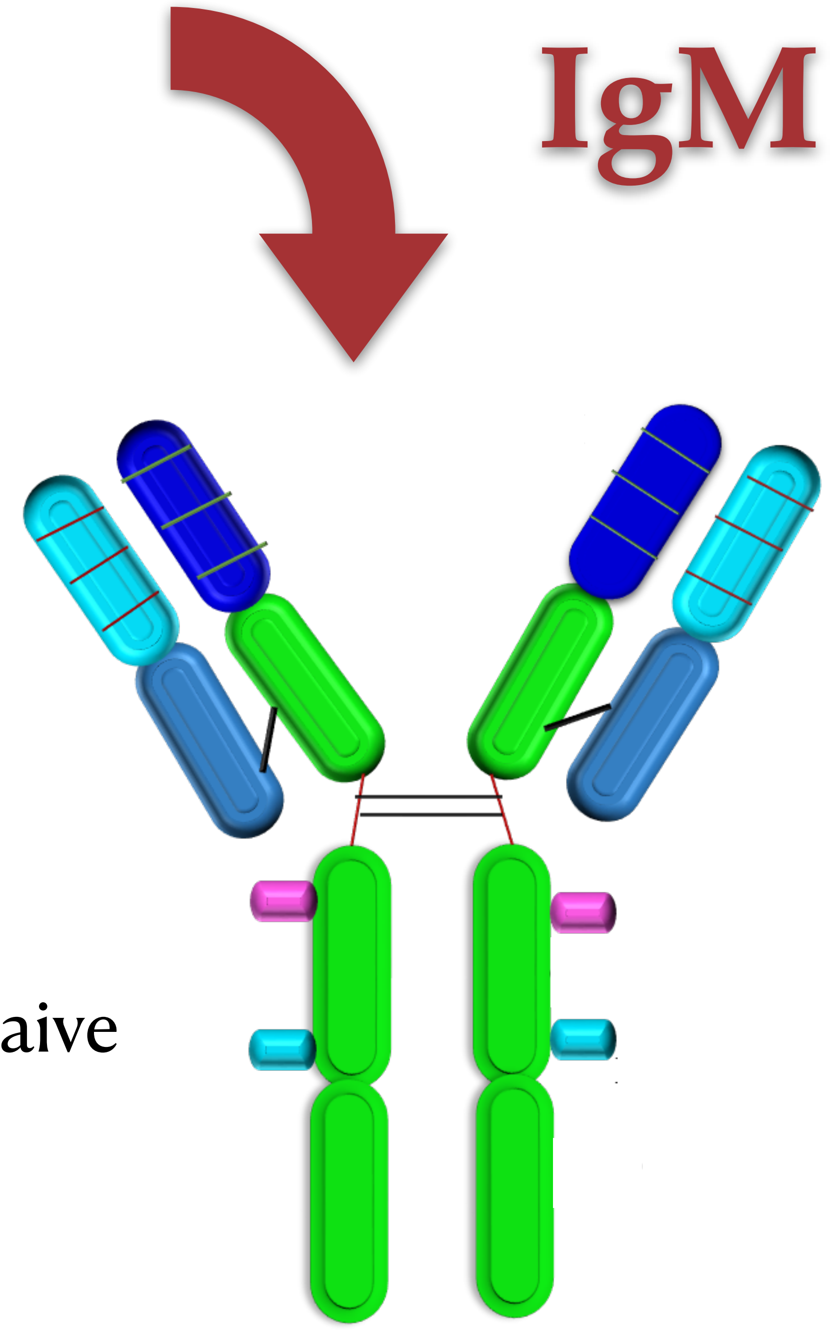
INTRODUCTION



PLOS Computational Biology | DOI:10.1371/journal.pcbi.1004409 January 11, 2016

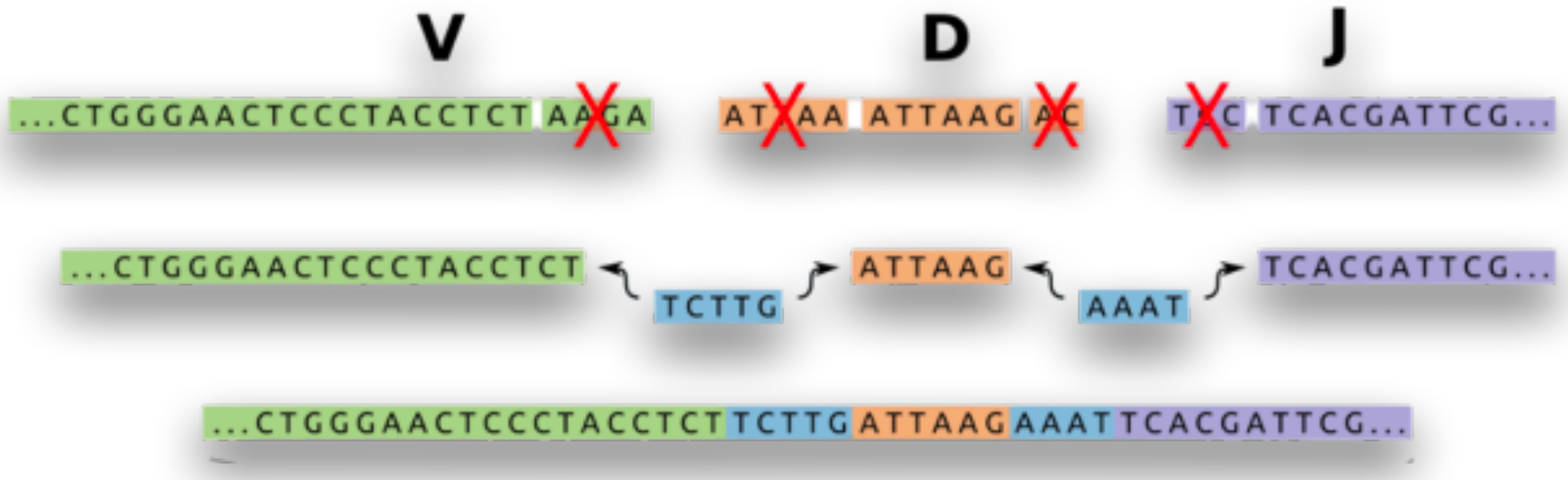
10^{61} different sequences

IgM



10^{12} unique naive sequences

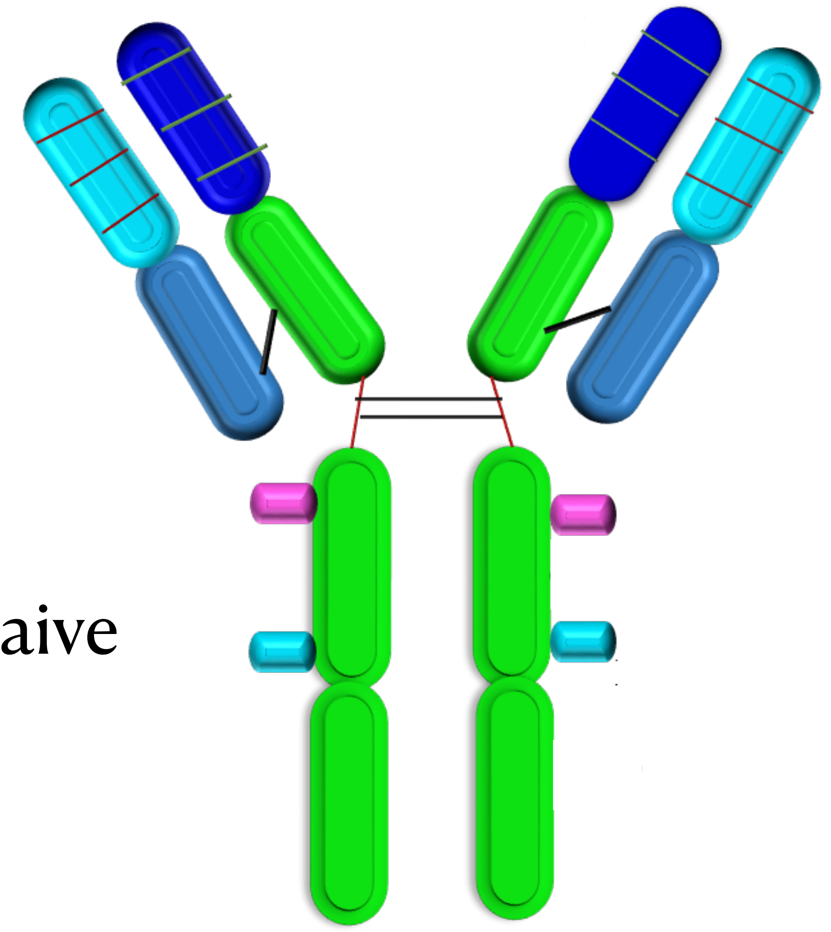
INTRODUCTION



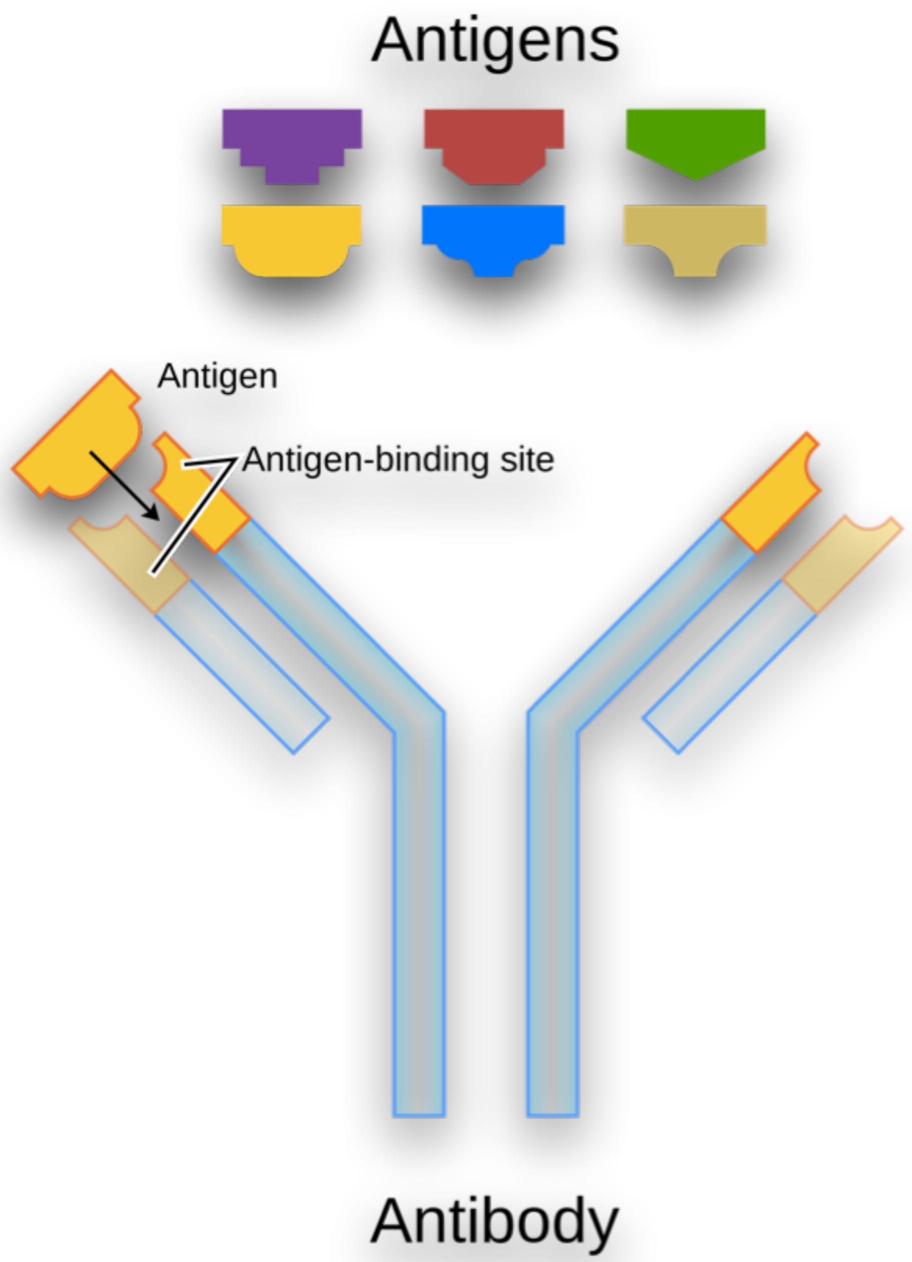
PLOS Computational Biology | DOI:10.1371/journal.pcbi.1004409 January 11, 2016

10^{61} different sequences

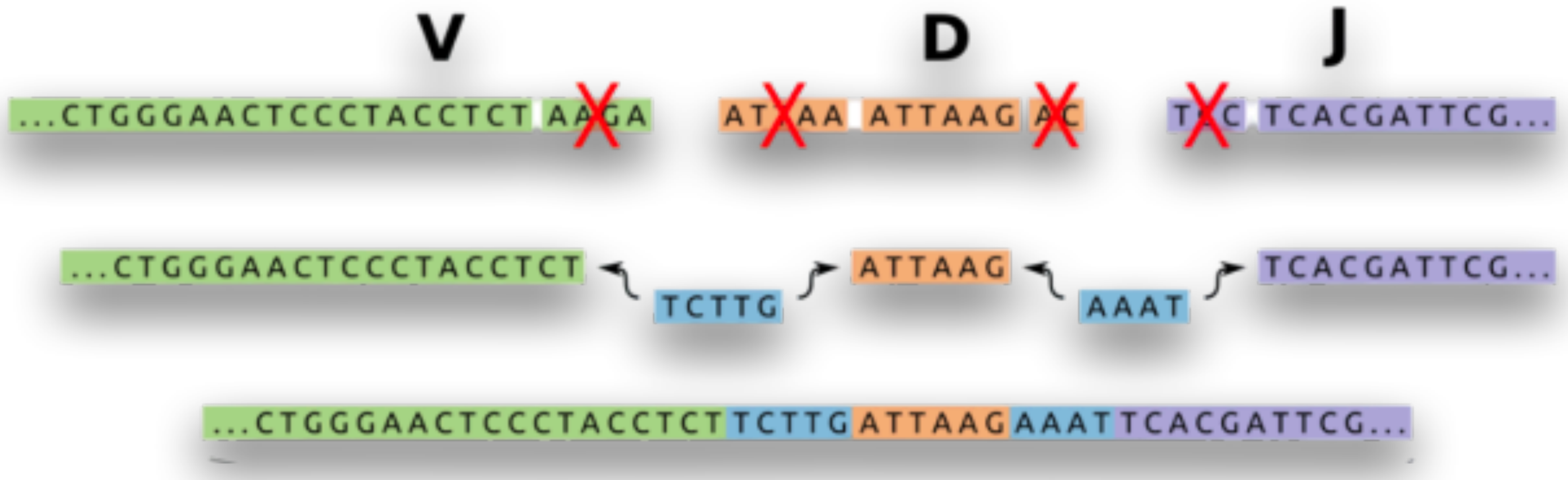
IgM



10^{12} unique naive sequences



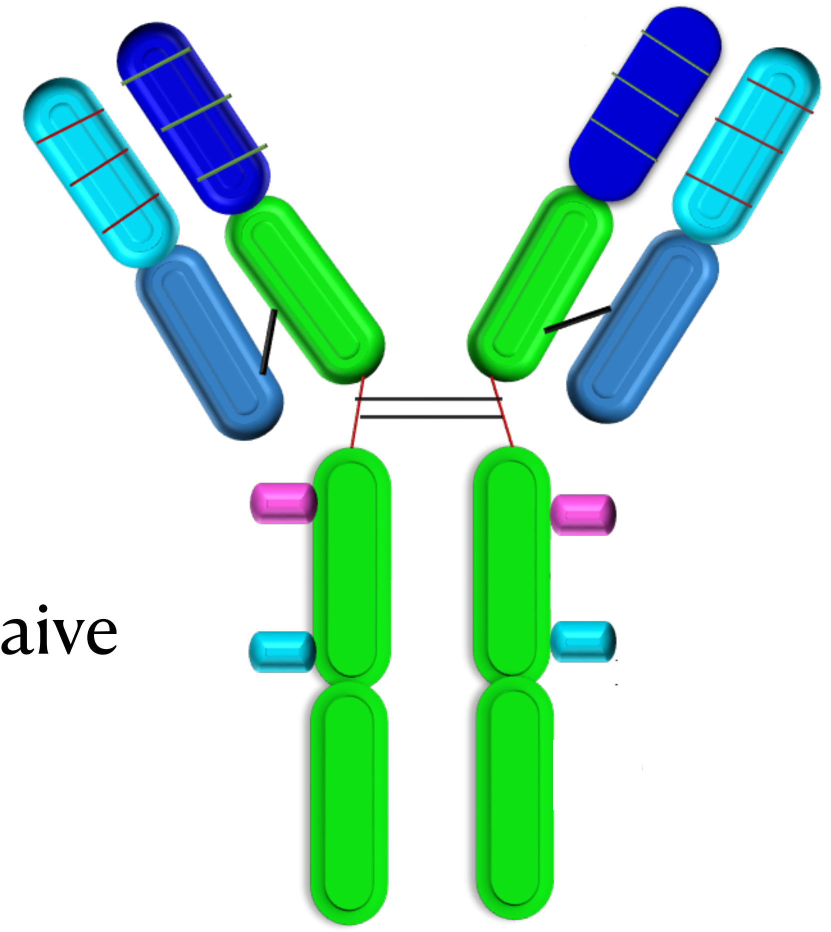
INTRODUCTION



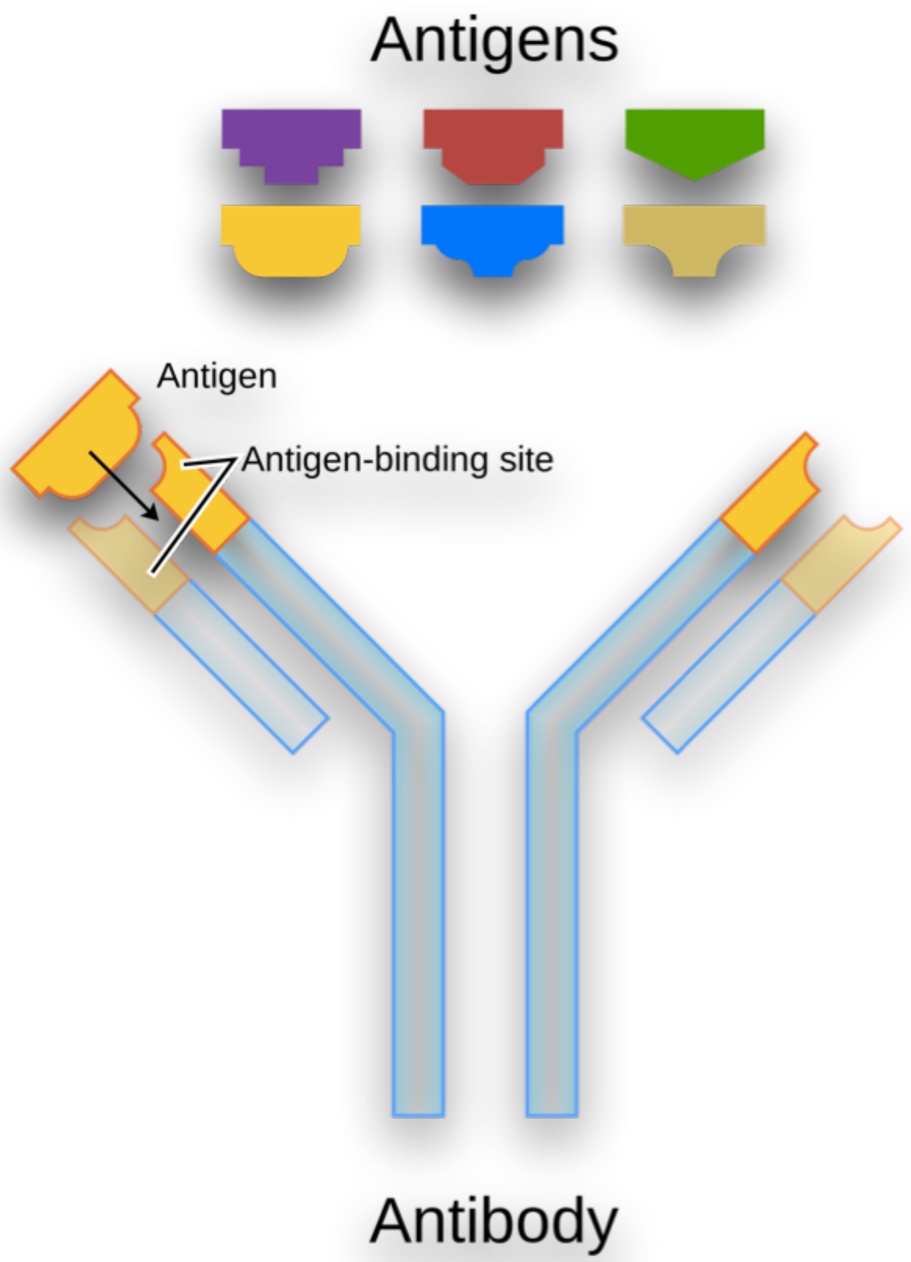
PLOS Computational Biology | DOI:10.1371/journal.pcbi.1004409 January 11, 2016

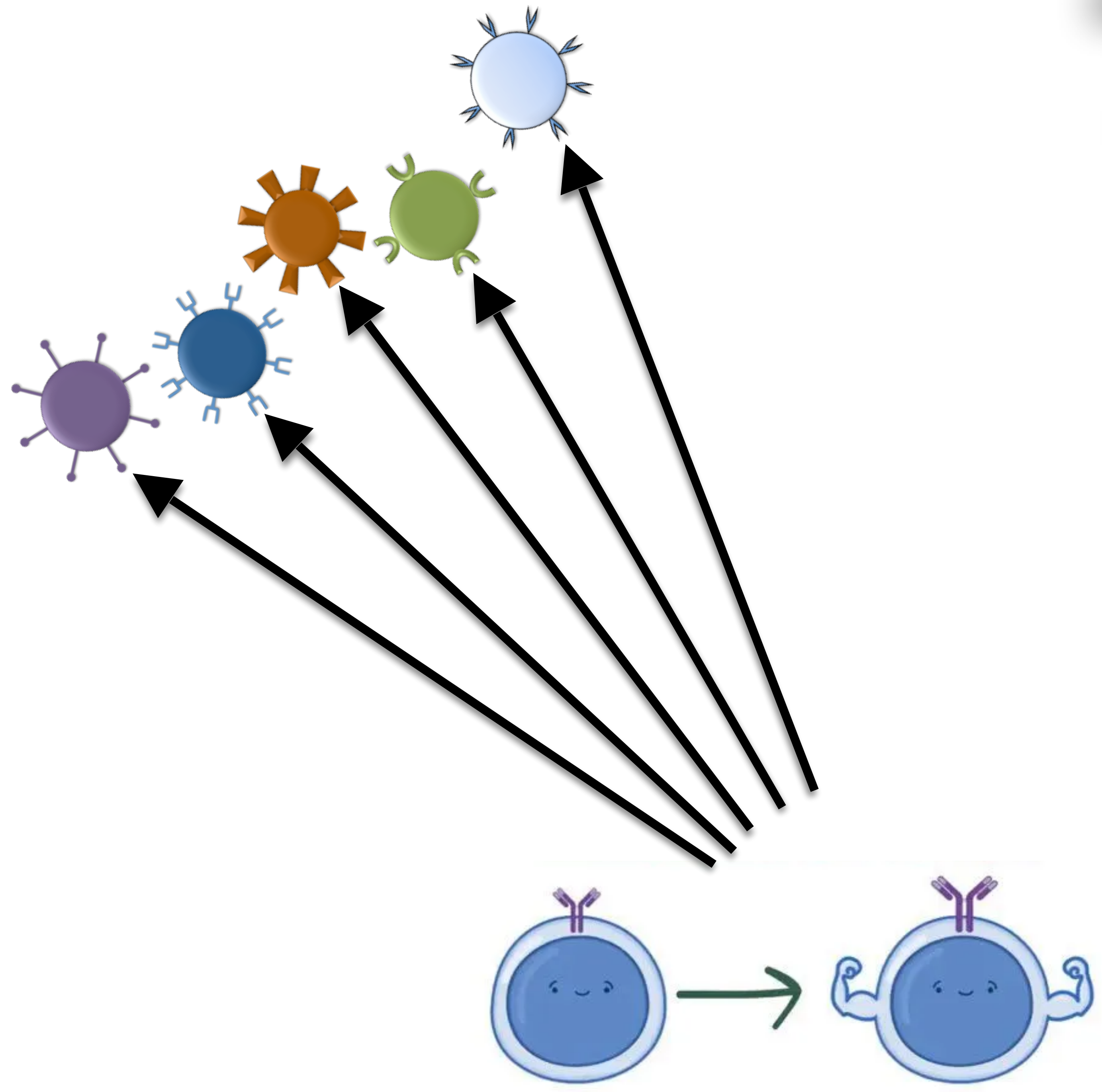
10^{61} different sequences

IgM

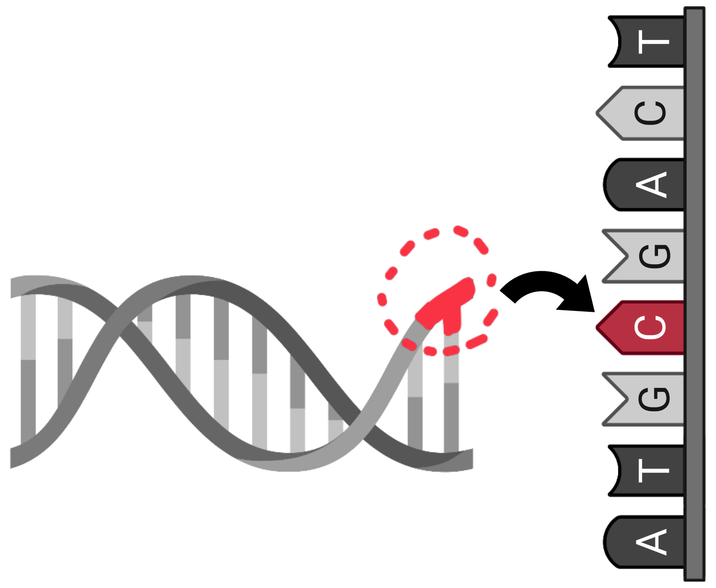


10^{12} unique naive sequences





IgG



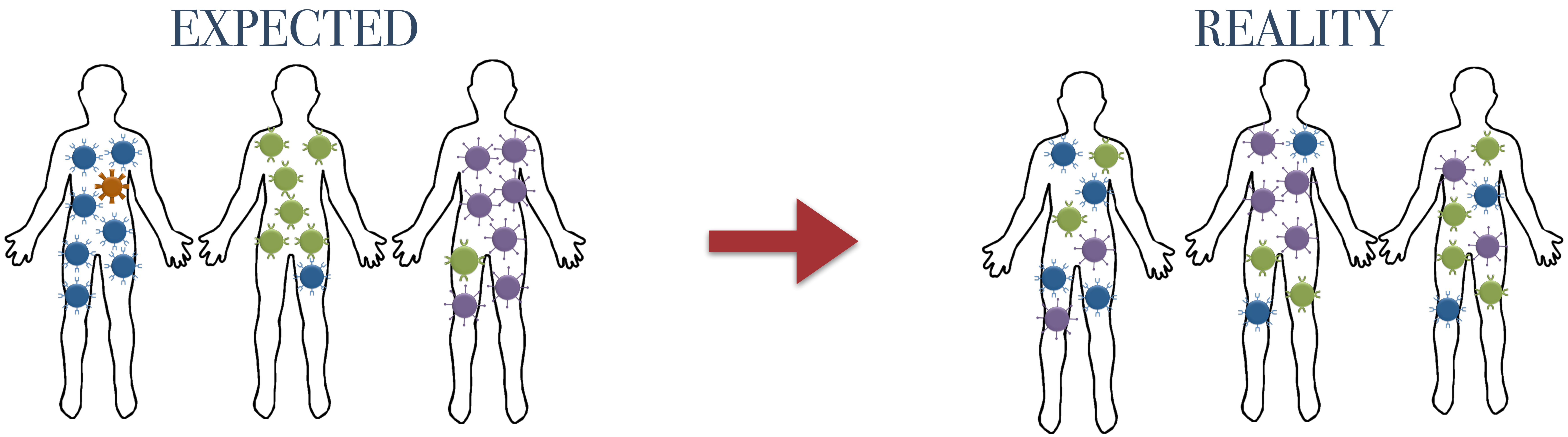
MOTIVATION

- B cell receptor (BCR) repertoires are highly diverse thanks to V(D)J recombination process + somatic hypermutations.

MOTIVATION

- B cell receptor (BCR) repertoires are highly diverse thanks to V(D)J recombination process + somatic hypermutations.

- A higher than expected by chance overlap of receptors is observed when repertoires from different individuals are compared.



EXPECTED

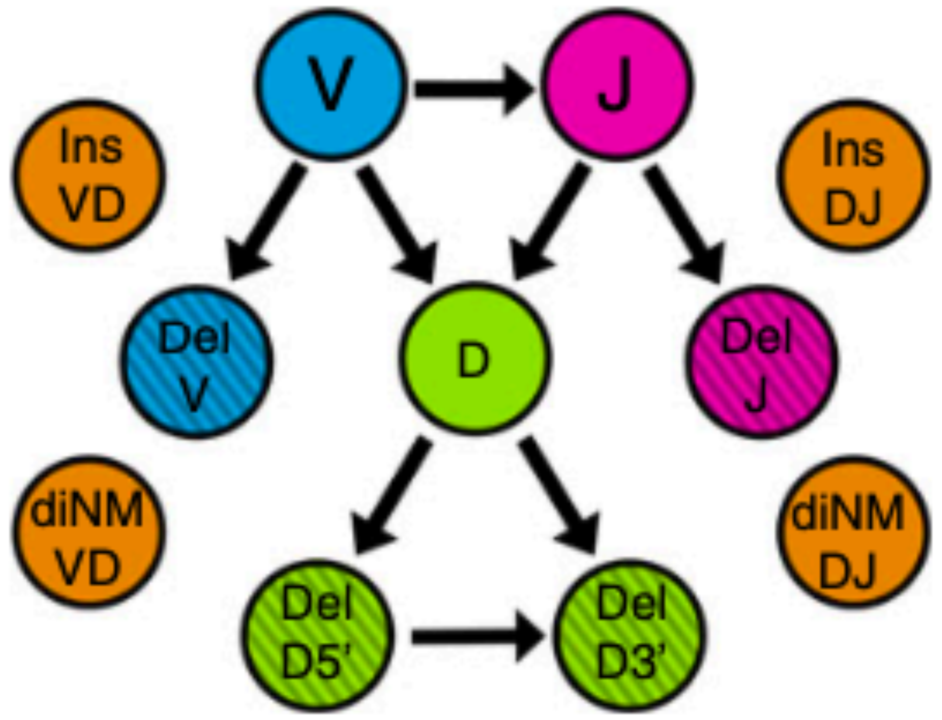
REALITY

- **OBJECTIVE**: design a statistical model that is able to predict the number of sequences that will be shared among individuals.

MODELLING THE SHARING MECHANISMS

- **Recombination biases** → Not all scenarios all equally likely.

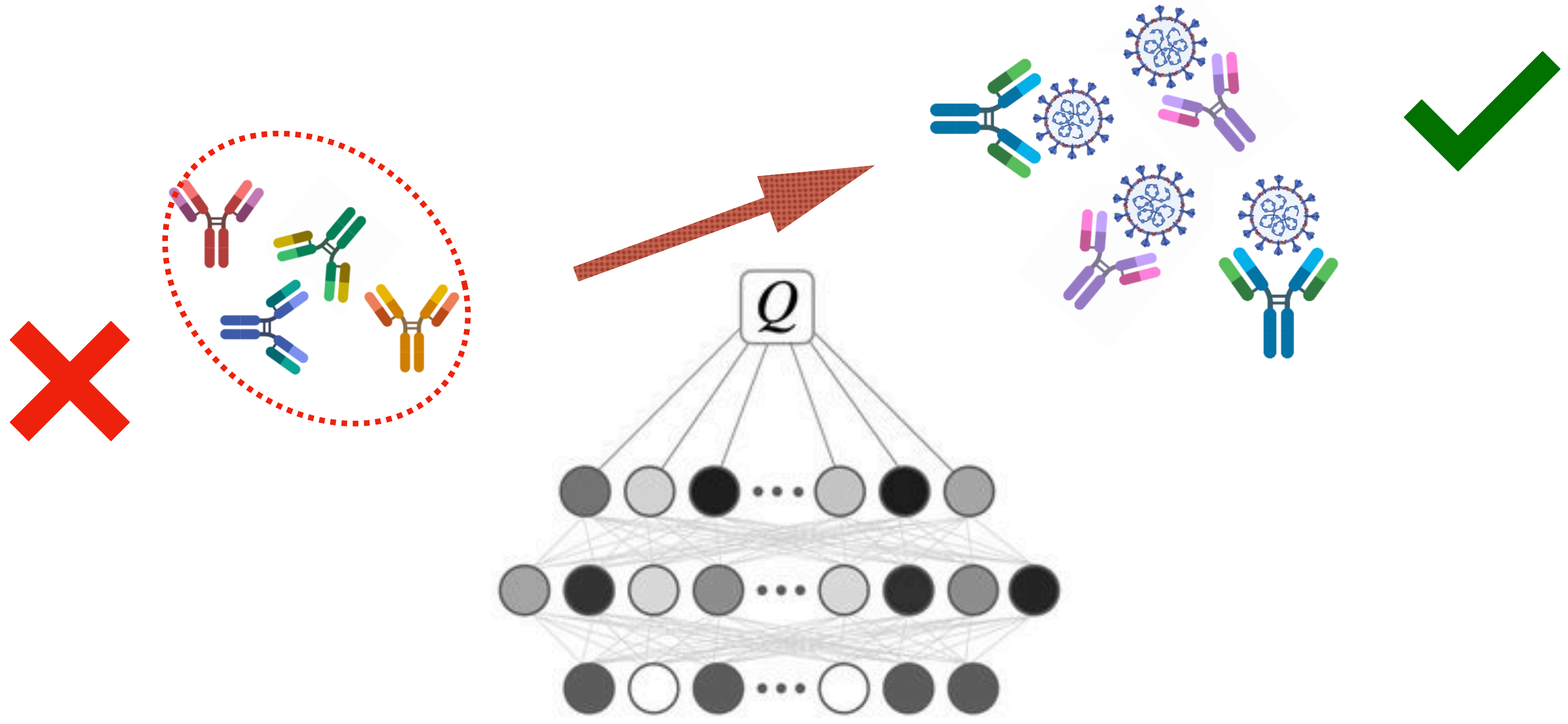
The statistics of V(D)J recombination process are well captured by P_{gen} model.



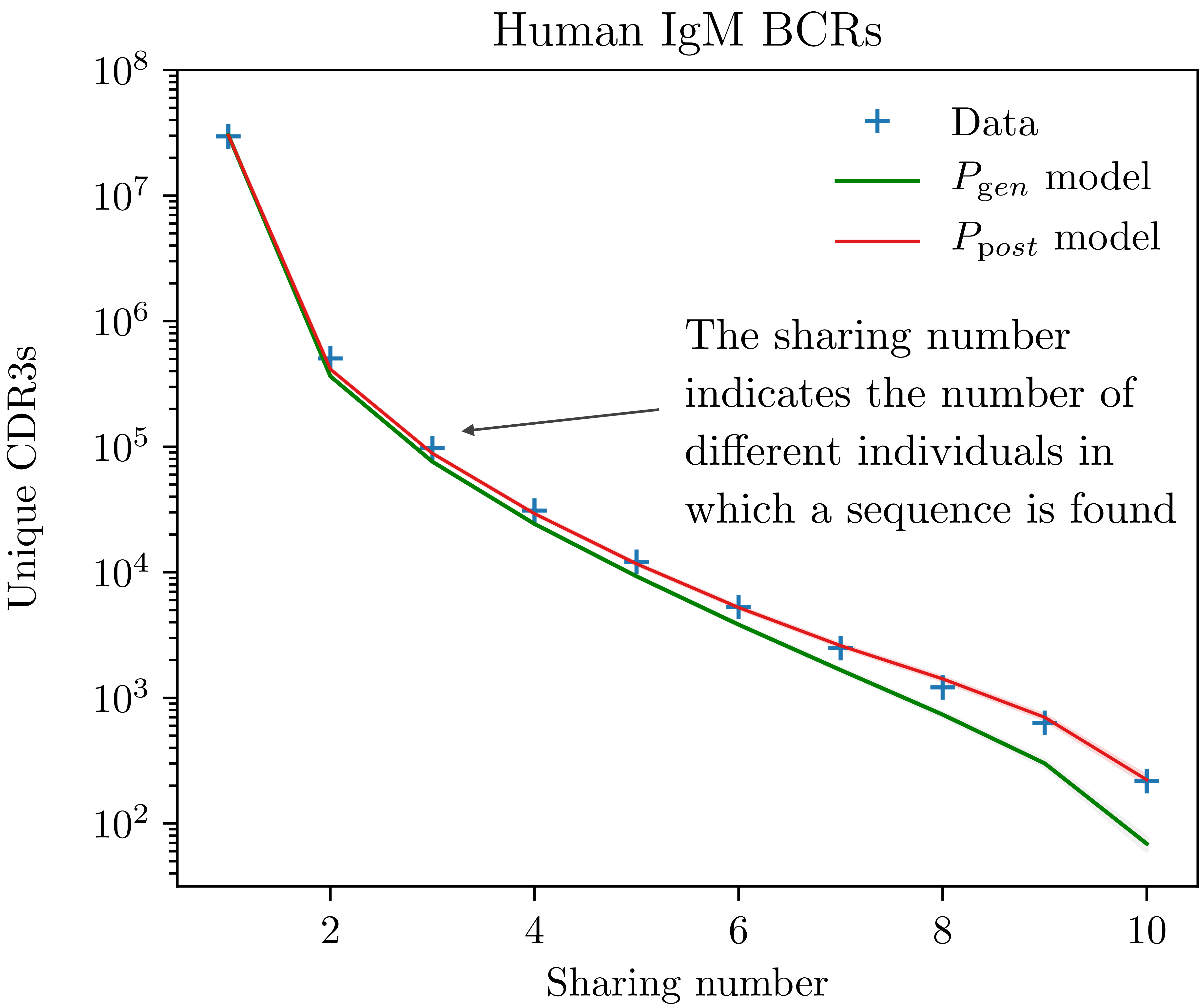
$$\begin{aligned} P(\text{scenario}) = {} & P(V)P(J|V)\,P(D|V,J)P(\text{del}V|V) \\ & \times P(\text{del}J|J)\,P(\text{delD5'}|D)\,P(\text{delD3'}|\text{de} \\ & \times P(\text{insVD}) \prod_i^{\text{InsVD}} P(n_i|n_{i-1}) \\ & \times P(\text{insDJ}) \prod_i^{\text{InsDJ}} P(n_i|n_{i-1}) \end{aligned}$$

- **Selection processes** → Like central tolerance or clonal selection.

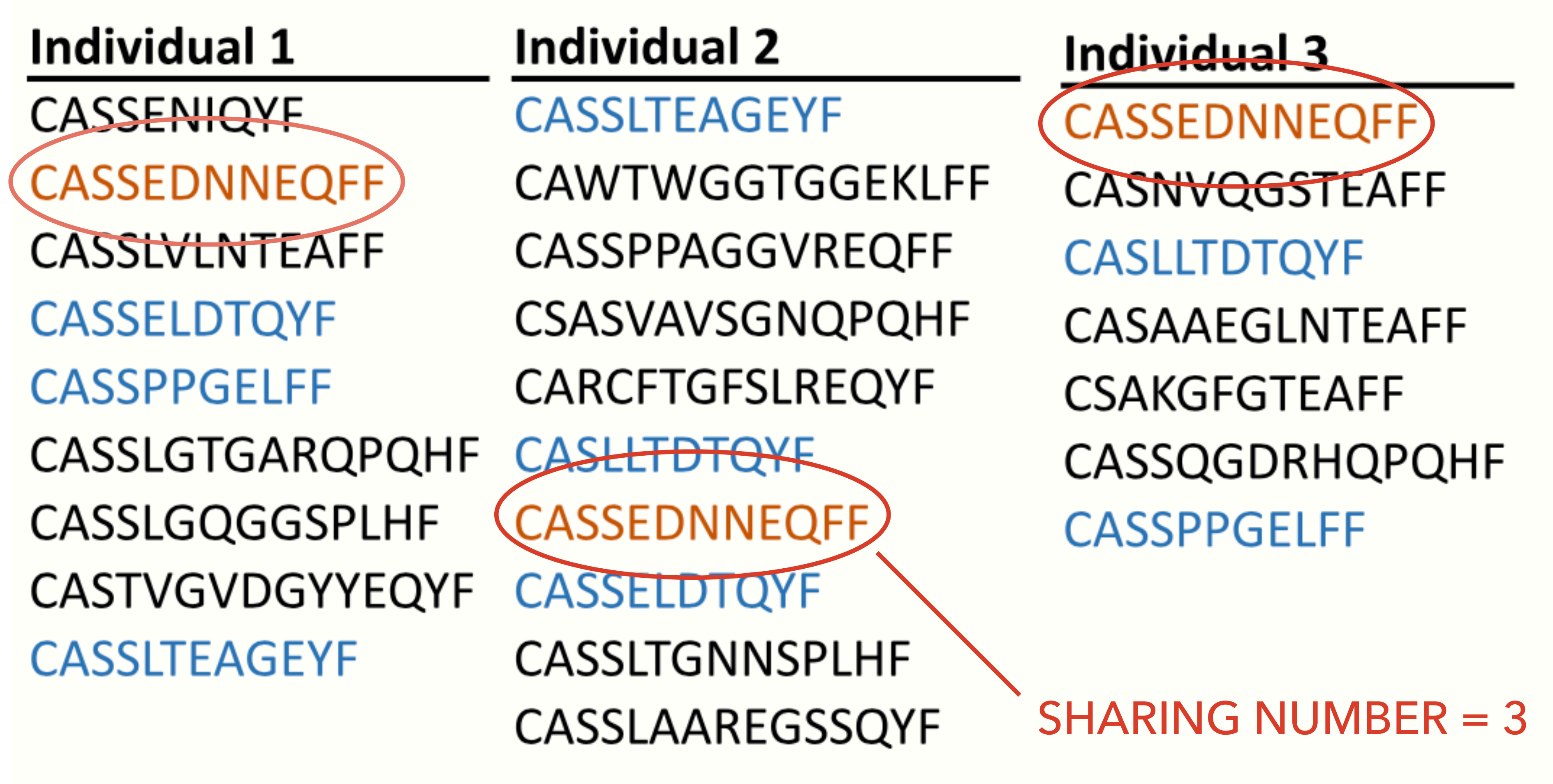
P_{post} is a selection model that identifies sequence features characteristic of functional lymphocyte repertoires



PREDICTING SHARING IN IGM REPERTOIRES



The sharing number indicates the number of different individuals in which a sequence is found



Individual 1	Individual 2	Individual 3
CASSENIQYF	CASSLTEAGEYF	CASSEDNNEQFF
CASSEDNNEQFF	CAWTWGGTGGEKLFF	CASNVQGSTEAFF
CASSLVLNTEAFF	CASSPPAGGVREQFF	CASLLTDTQYF
CASSELDTQYF	CSASVAVSGNQPQHF	CASAAEGLNTEAFF
CASSPPGELFF	CARCFTGFSLREQYF	CSAKGFGTEAFF
CASSLGTGARQPQHF	CASLLTDTQYF	CASSQGDRHQPQHF
CASSLGQGGSPLHF	CASSEDNNEQFF	CASSPPGELFF
CASTVGVDGYYEQYF	CASSELDTQYF	
CASSLTEAGEYF	CASSLTGNNSPLHF	
	CASSLAAREGSSQYF	

SHARING NUMBER = 3

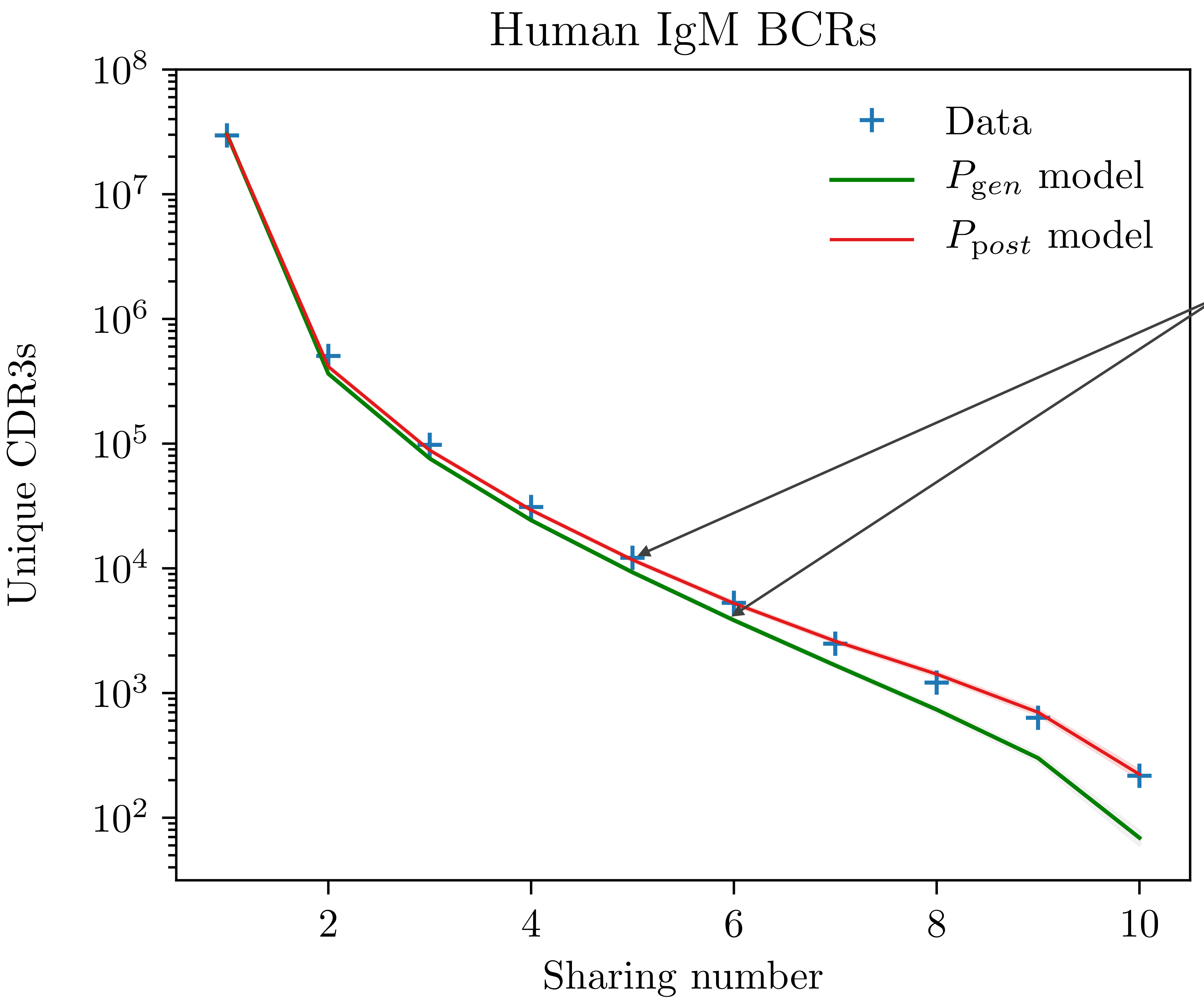
Dataset obtained from:

Briney, B., Inderbitzin, A., Joyce, C. *et al.* Commonality despite exceptional diversity in the baseline human antibody repertoire.

Nature **566,** 393–397 (2019).

https://doi.org/10.1038/s41586-019-0879-

PREDICTING SHARING IN IGM REPERTOIRES



Theoretical prediction using the generation function:

$$G(x, \{N_i\}) = \sum_{m=0}^{n} M_m(N_i)x^m = \sum_{s\in S}\prod_{i=1}^{n}\left[e^{-p(s)N_i} + (1 - e^{-p(s)N_i})x\right]$$

Using $p(s) = P_{gen}$ underestimates how many sequences are shared among individuals. But the model of convergent recombination + selection, $p(s) = P_{post}$, accurately predicts this quantity.

Dataset obtained from:

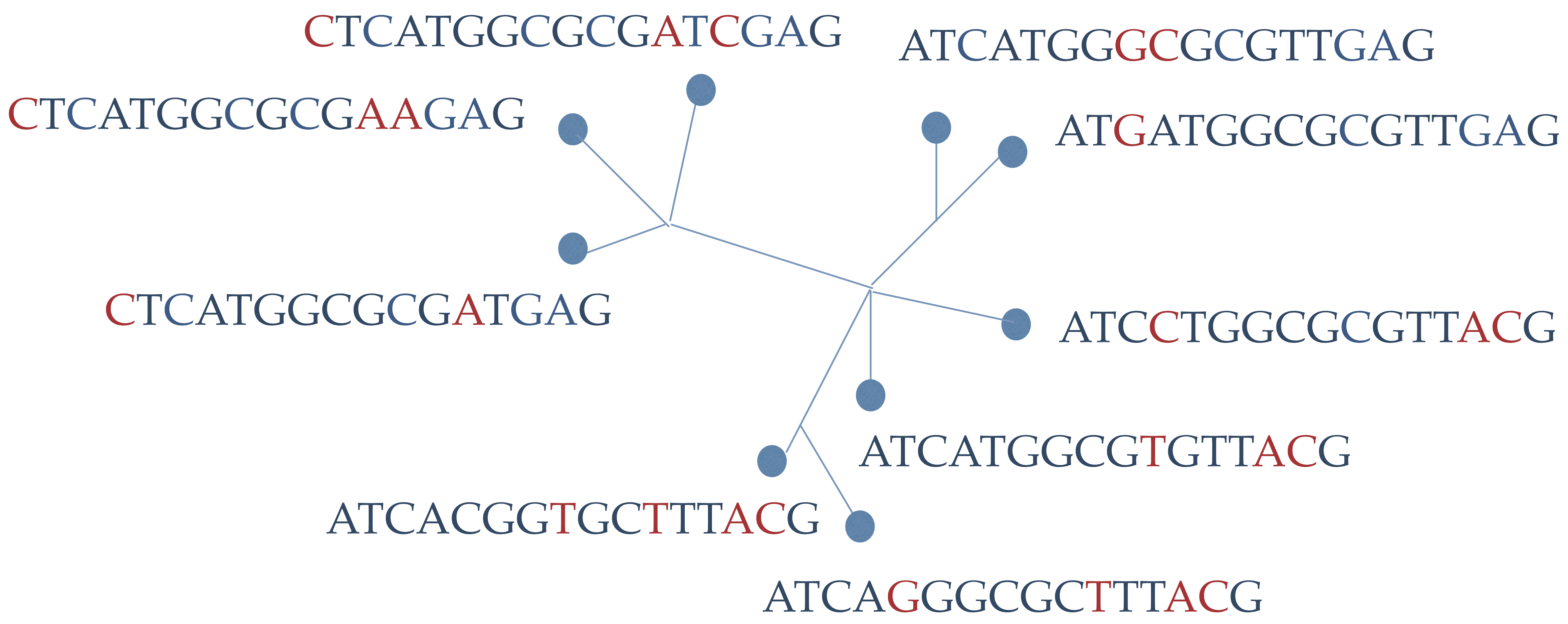
Briney, B., Inderbitzin, A., Joyce, C. *et al.* Commonality despite exceptional diversity in the baseline human antibody repertoire.

Nature **566,** 393–397 (2019).

https://doi.org/10.1038/s41586-019-0879-

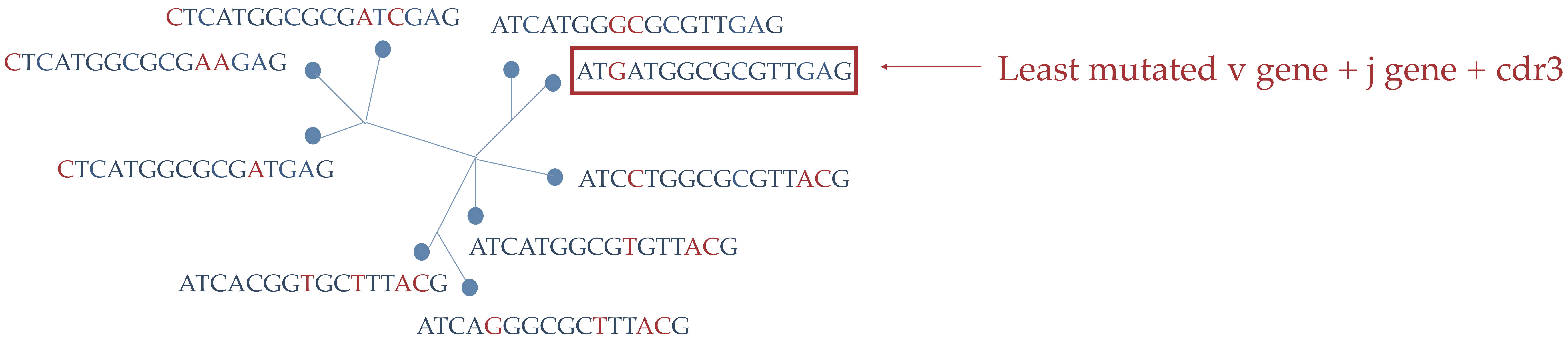
PREDICTING SHARING IN IGG REPERTOIRES

- The same pipeline can be applied to non mutated IgG repertoires. We need to recover the ancestors of clonal families:



PREDICTING SHARING IN IGG REPERTOIRES

- The same pipeline can be applied to non mutated IgG repertoires. We need to recover the ancestors of clonal families:



CTCATGGCGCGATCGAG

ATCATGGGCGCGTTGAG

CTCATGGCGCGAAGAG

ATGATGGCGCGTTGAG

Least mutated v gene + j gene + cdr3

CTCATGGCGCGATGAG

ATCCTGGCGCGTTACG

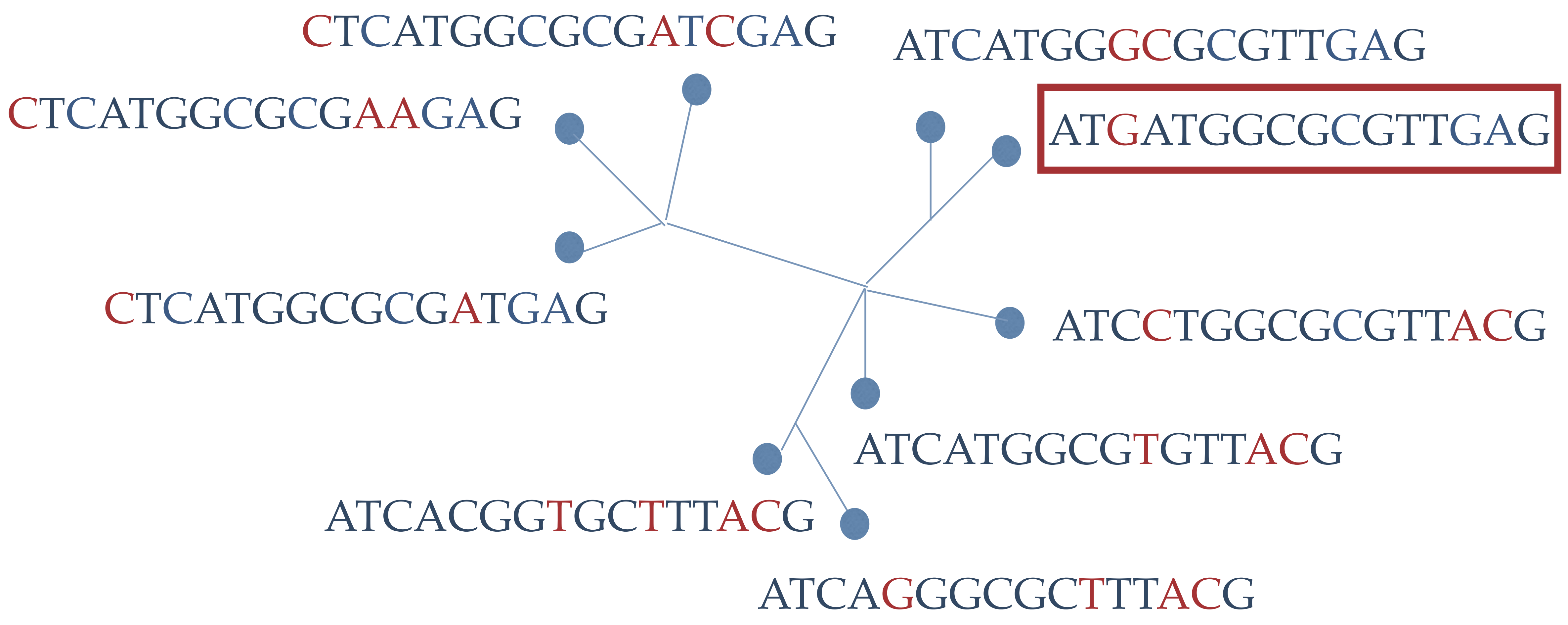
ATCATGGCGTGTTACG

ATCACGGTGCTTTACG

ATCAGGGCGCTTTACG

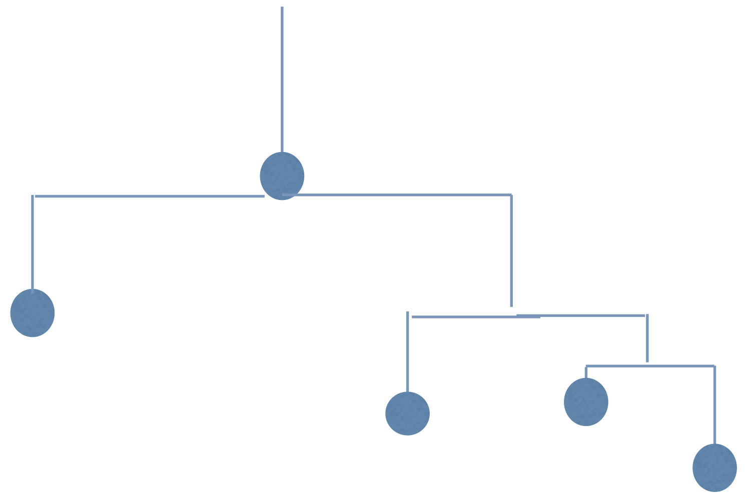
PREDICTING SHARING IN IGG REPERTOIRES

- The same pipeline can be applied to non mutated IgG repertoires. We need to recover the ancestors of clonal families:



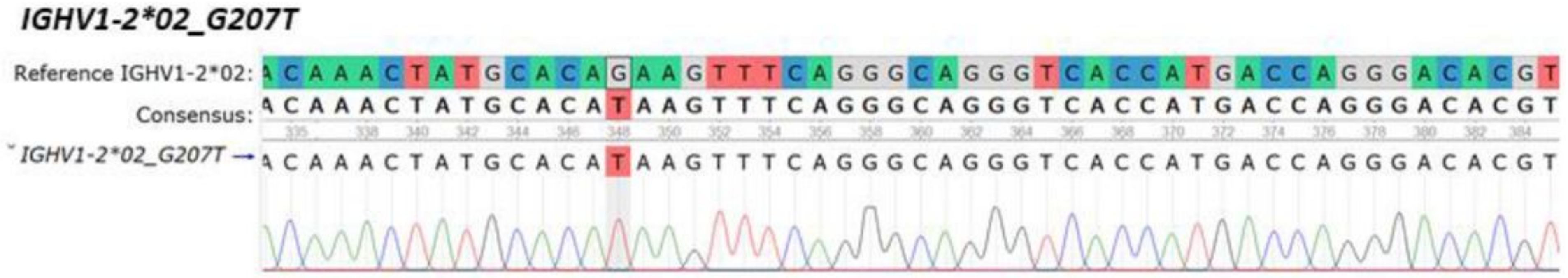
- Naive ancestor

ATCATGGCGCGTTGAG



ImMunoGeneTics Information system®

Remove the mutations comparing with germline templates

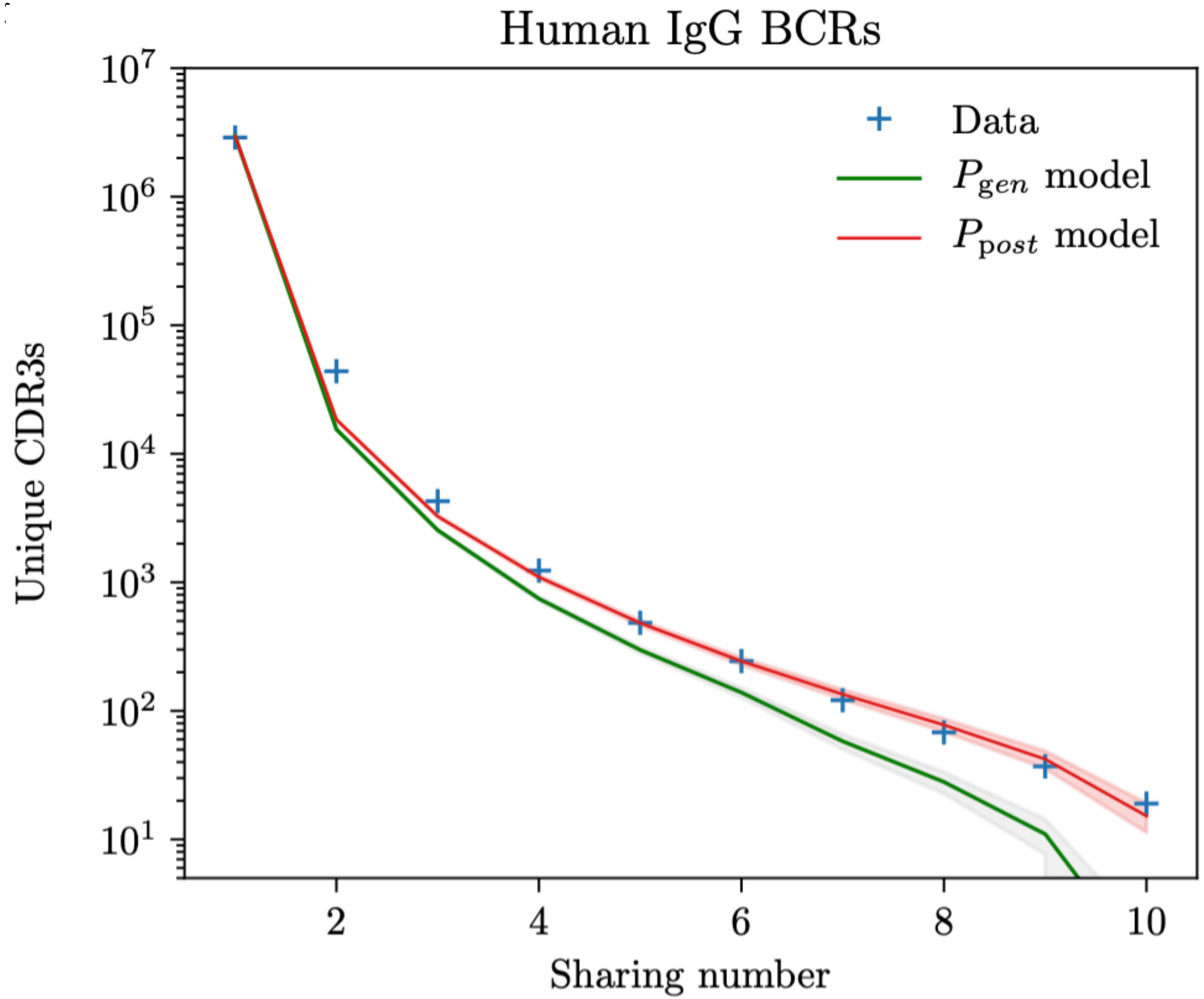


IGHV1-2*02_G207T

```
Reference IGHV1-2*02: ACAAACTATGCACAGAAGTTTCAGGGCAGGGTCACCATGACCAGGGACACGT
          Consensus:  ACAAACTATGCACATAAGTTTCAGGGCAGGGTCACCATGACCAGGGACACGT
                      335  338 340 342 344 346 348 350 352 354 356 358 360 362 364 366 368 370 372 374 376 378 380 382 384
    IGHV1-2*02_G207T →ACAAACTATGCACATAAGTTTCAGGGCAGGGTCACCATGACCAGGGACACGT
```

PREDICTING SHARING IN IGG REPERTOIRES

- The same pipeline can be applied to non mutated IgG repertoires. We need to recover the ancestors of clonal families:



Dataset obtained from:

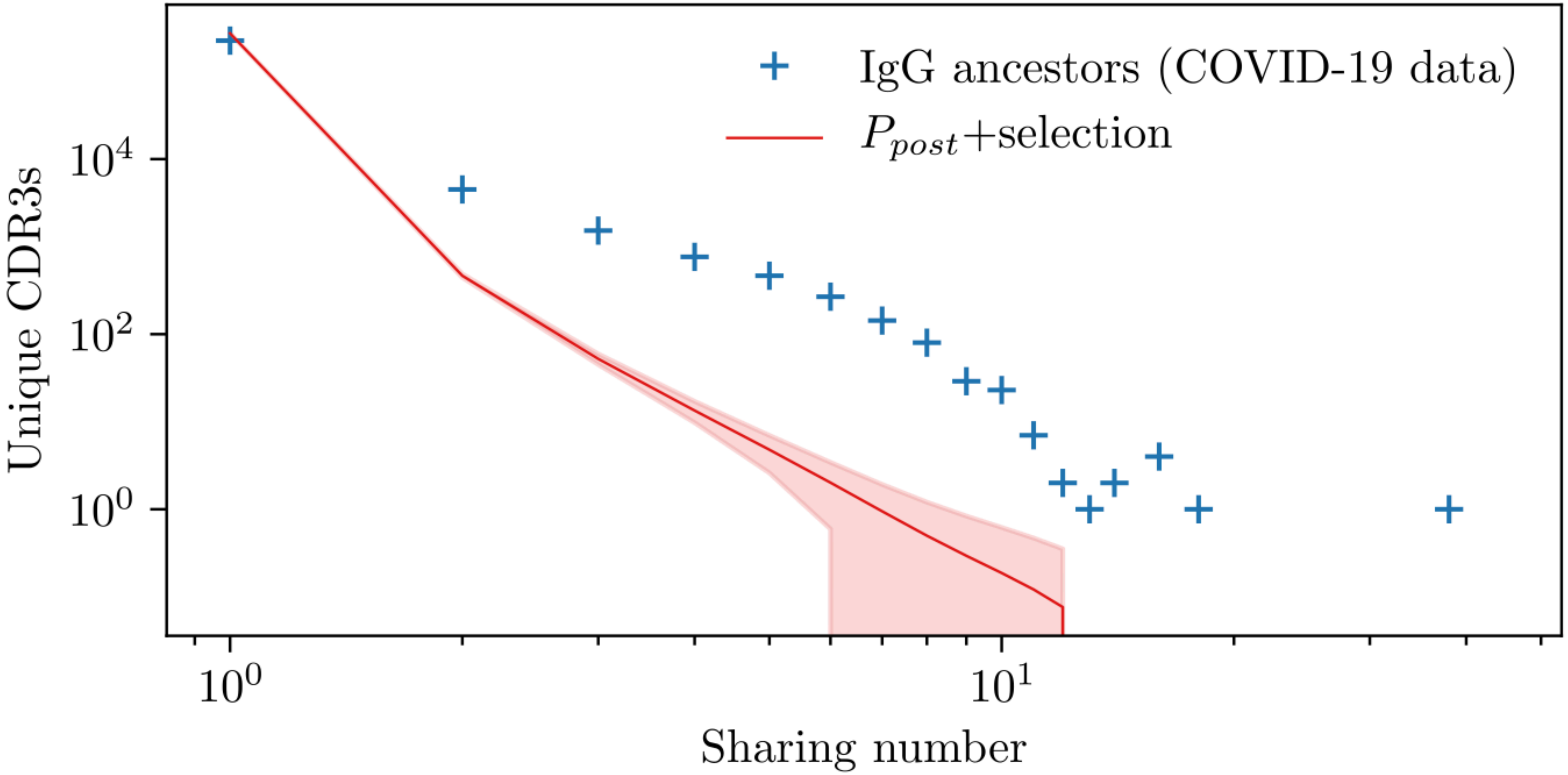
Briney, B., Inderbitzin, A., Joyce, C. *et al.* Commonality despite exceptional diversity in the baseline human antibody repertoire.

Nature **566,** 393–397 (2019).

https://doi.org/10.1038/s41586-019-0879-

PREDICTING SHARING IN COVID-19 REPERTOIRES

- Comparison of sharing number distribution in a cohort of 43 individuals **COVID-19 positive** and predictions from P_{post} :



Datasets obtained from:

Cell Host & Microbe 28, 516–525, October 7, 2020

https://doi.org/10.1016/j.chom.2020.09.002

medRxiv 2020.07.13.20153114;

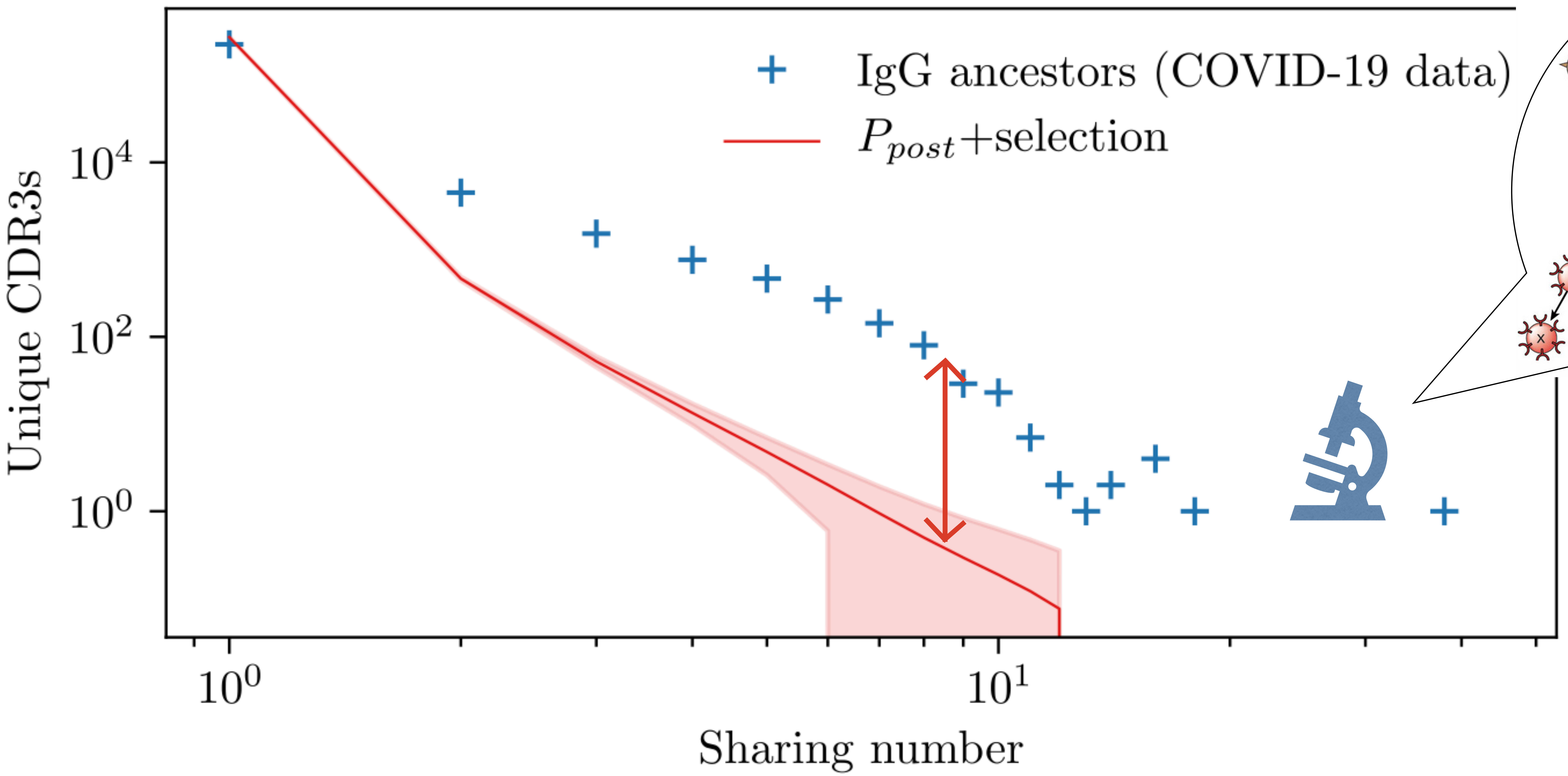
https://doi.org/10.1101/2020.07.13.20153114

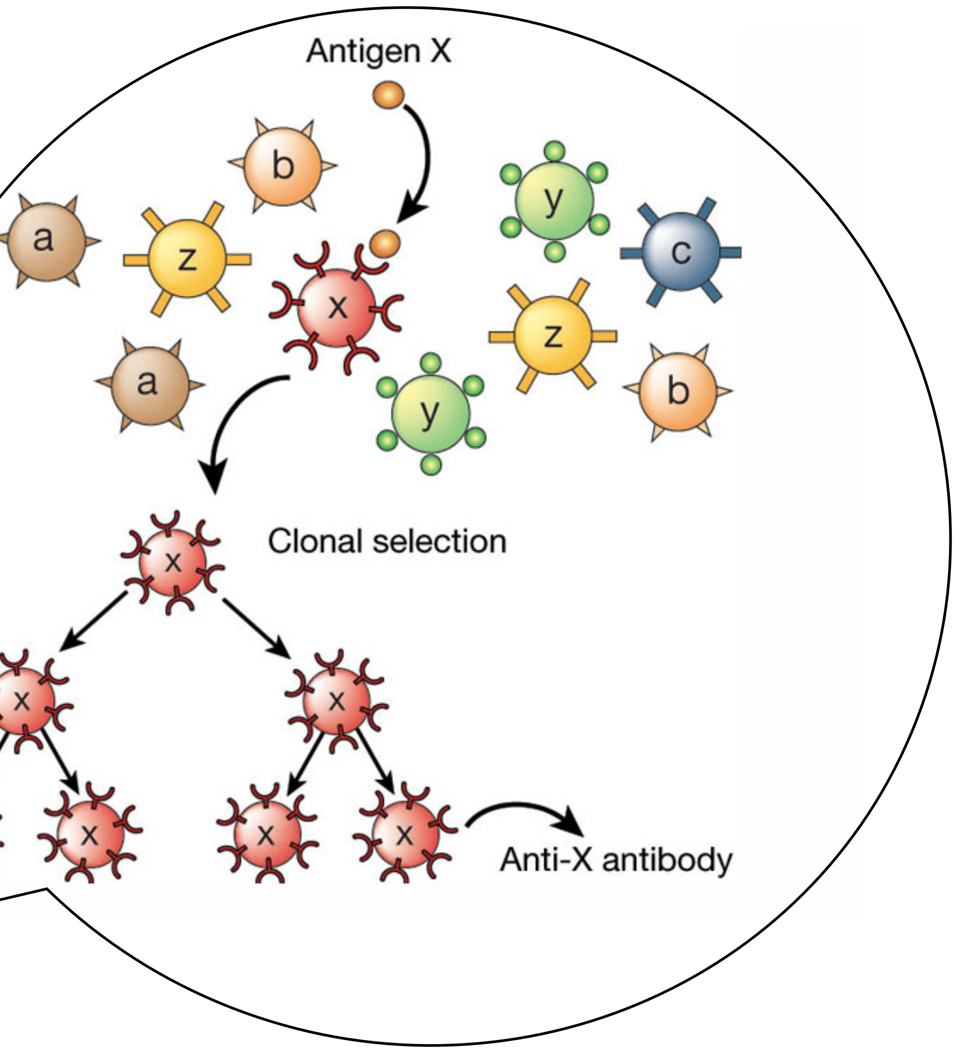
Front. Immunol. 11:605170.

doi: 10.3389/fimmu.2020.605170

PREDICTING SHARING IN COVID-19 REPERTOIRES

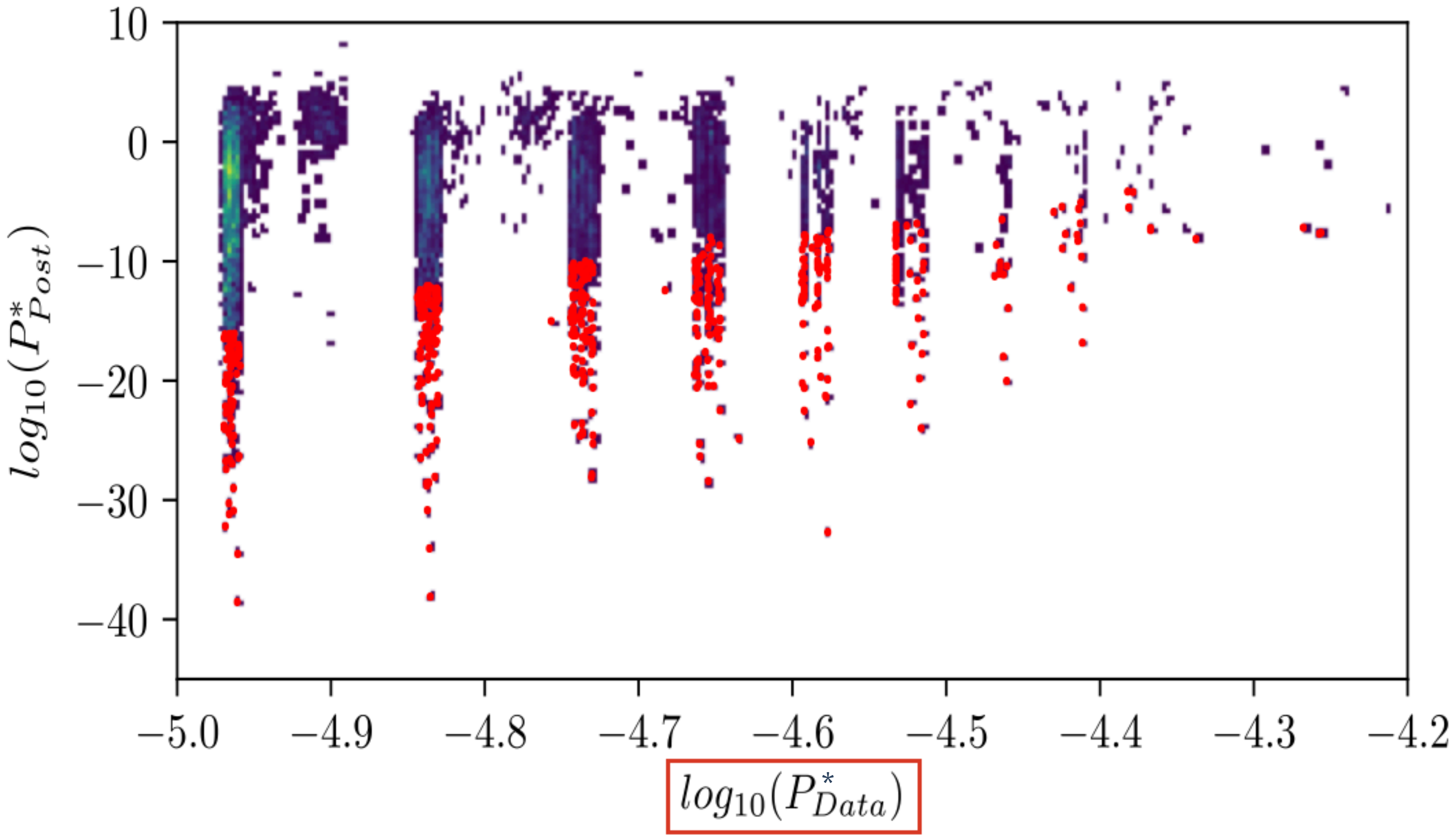
- Comparison of sharing number distribution in a cohort of 43 individuals **COVID-19 positive** and predictions from P_{post} :





Are these overshared sequences indicative of a specific antibody response to an antigen?

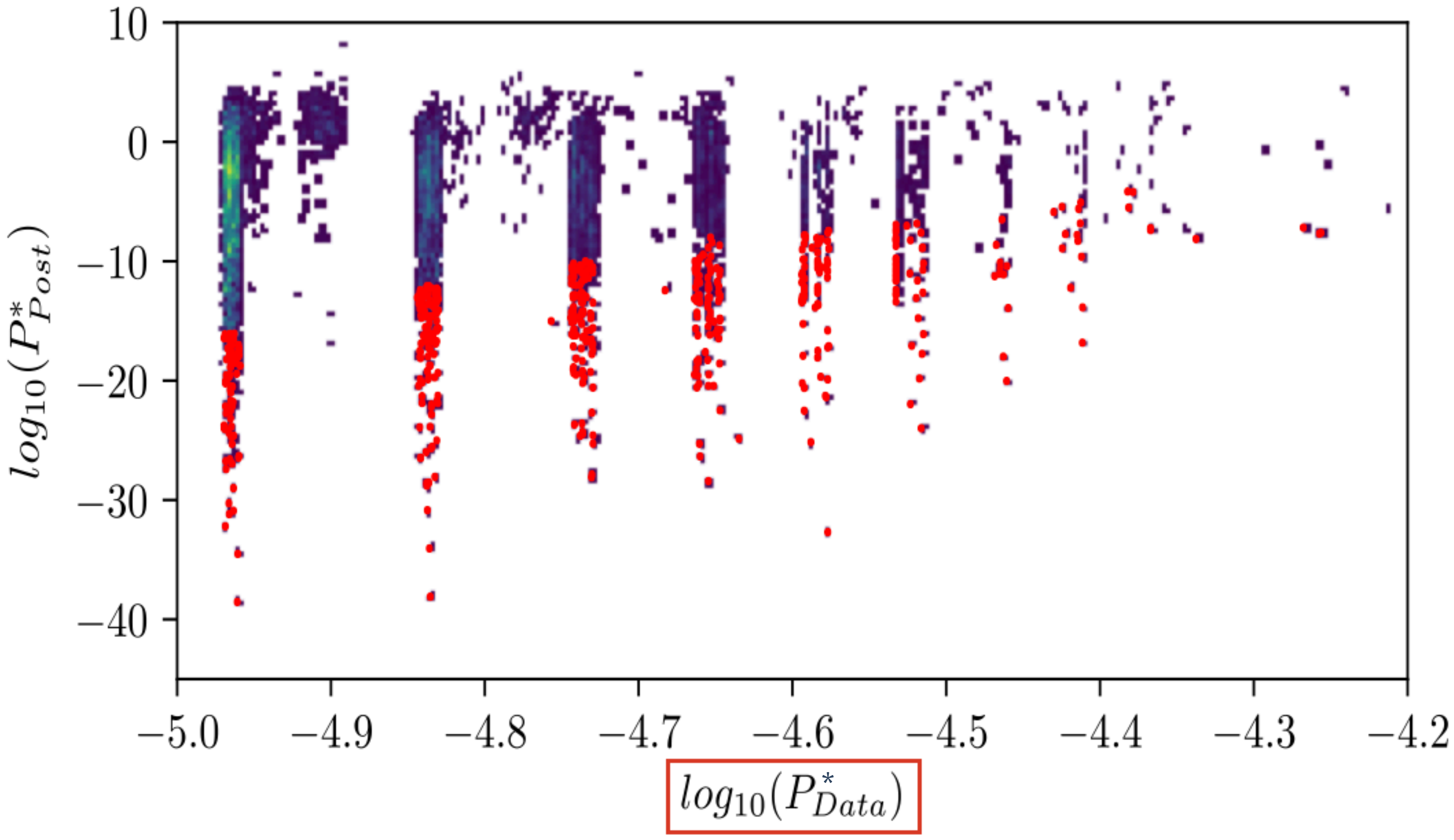
IDENTIFICATION OF ANTIGEN-RESPONDING CLONES



Expected frequency from the sharing pattern

$$P^*_{data} = \text{argmax}_{P_{data}} \mathbb{P}(x_1, \ldots, x_n \mid P_{data})$$

IDENTIFICATION OF ANTIGEN-RESPONDING CLONES



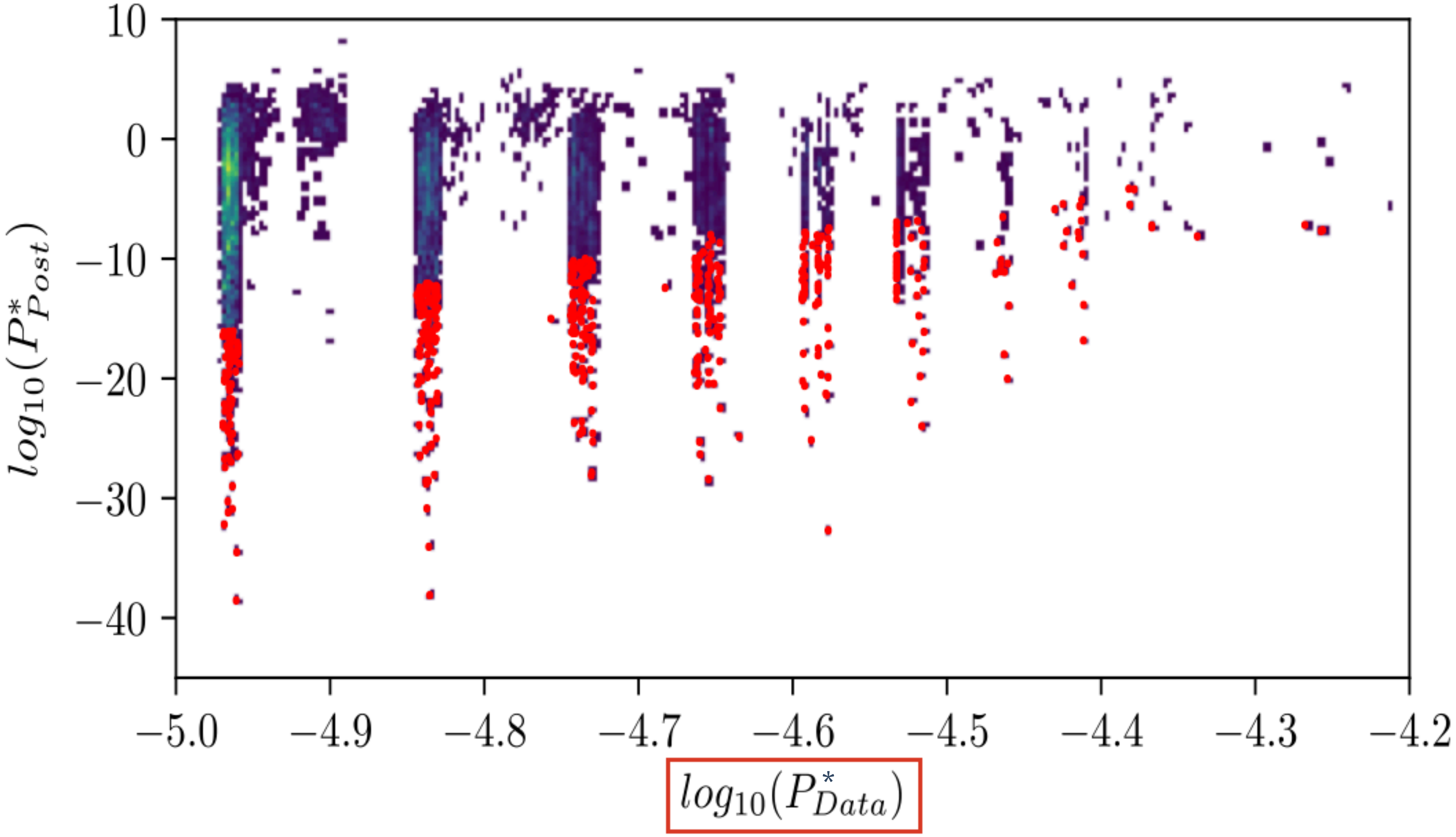
Level of certitude on P_{data} given the observations $\{x_1, x_2, \ldots, x_n\}$:

$$\mathbb{P}(P_{post} > P_{data}) = \int_0^{P_{post}} \frac{\mathbb{P}(x_1, \ldots, x_n \mid P_{data})\rho_{prior}(P_{data})}{\int_0^1 \mathbb{P}(x_1, \ldots, x_n \mid P_{data})\rho_{prior}(P_{data})dP_{data}} dP_{data}$$

Expected frequency from the sharing pattern

$$P^*_{data} = \text{argmax}_{P_{data}} \mathbb{P}(x_1, \ldots, x_n \mid P_{data})$$

IDENTIFICATION OF ANTIGEN-RESPONDING CLONES



Level of certitude on P_{data} given the observations $\{x_1, x_2, \ldots, x_n\}$:

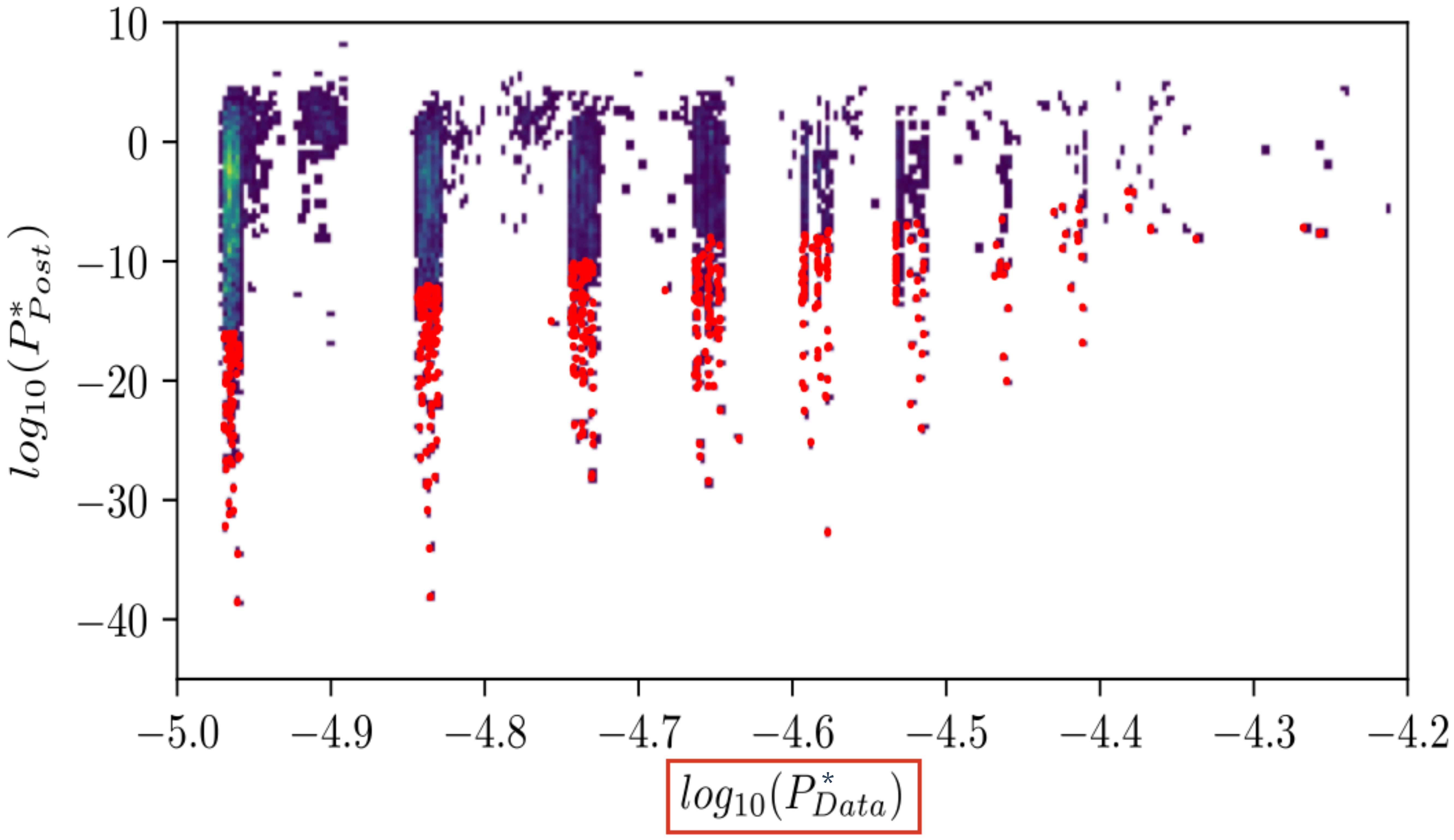
$$\mathbb{P}(P_{post} > P_{data}) = \int_0^{P_{post}} \frac{\mathbb{P}(x_1, \ldots, x_n \mid P_{data})\rho_{prior}(P_{data})}{\int_0^1 \mathbb{P}(x_1, \ldots, x_n \mid P_{data})\rho_{prior}(P_{data})dP_{data}} dP_{data}$$

Bayesian analogous to p-value

Expected frequency from the sharing pattern

$$P^*_{data} = \text{argmax}_{P_{data}} \mathbb{P}(x_1, \ldots, x_n \mid P_{data})$$

IDENTIFICATION OF ANTIGEN-RESPONDING CLONES



Expected frequency from the sharing pattern

$$P^*_{data} = \text{argmax}_{P_{data}} \mathbb{P}(x_1, \ldots, x_n \mid P_{data})$$

Level of certitude on P_{data} given the observations $\{x_1, x_2, \ldots, x_n\}$:

$$\mathbb{P}(P_{post} > P_{data}) = \int_0^{P_{post}} \frac{\mathbb{P}(x_1, \ldots, x_n \mid P_{data})\rho_{prior}(P_{data})}{\int_0^1 \mathbb{P}(x_1, \ldots, x_n \mid P_{data})\rho_{prior}(P_{data})dP_{data}} dP_{data}$$

→ **Their frequency can't be explained by a high recombination likelihood.**

→ Potential candidates of SARS-CoV-2 antibodies.

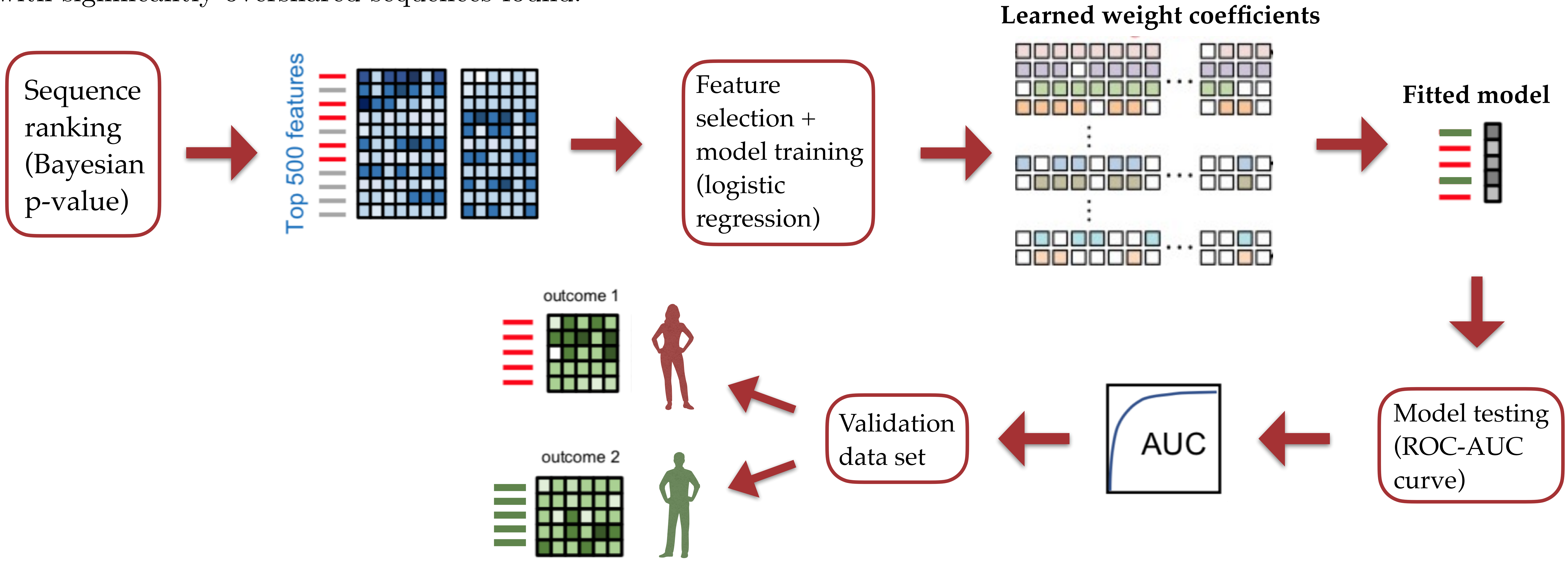
FURTHER ANALYSIS OF OVERSHARED SEQUENCES

(i) **Clustering + annotation** → Overshared IGH sequences are clustered in networks with Levenshtein distance < 2 and are matched against reported SARS-CoV-2 antibodies.

1.
2.
3.
4.
5.
6.
7.
8.
9.
10.
11.
12.
13.
14.
15.
16.
🟩 RBD
🟦 S2
🟥 non – RBD
□ Non determined
🟨 NTD

FURTHER ANALYSIS OF OVERSHARED SEQUENCES

(ii) **Learning of regression models for COVID-19 diagnosis** → Prediction of COVID-19 status based on the overlap with significantly overshared sequences found.



Sequence ranking (Bayesian p-value)

Feature selection + model training (logistic regression)

Learned weight coefficients

Fitted model

Model testing (ROC-AUC curve)

Validation data set

CONCLUSIONS

- The statistical model here presented accurately estimates the probability of observing a productive B cell receptor in a repertoire.

- The model accurately predicts how many sequences will be shared among n healthy individuals but it fails at capturing the selection pressure existing after antigen encounter. The significance of this effect is measured by defining a Bayesian analogous to p-value.

- The sharing analysis here presented might be particularly useful to help designing a vaccine that ellicitates a more transverse immune response since the antibodies that have been isolated have been already produced by a large number of individuals.