Substitution and per-residue selection in B cell affinity maturation

Connor McCoy, Trevor Bedford, Vladimir Minin, Harlan Robins, and Erick Matsen

@ematsen http://matsen.fhcrc.org/

Jenner's 1796 vaccine

Where are we 200 years later?

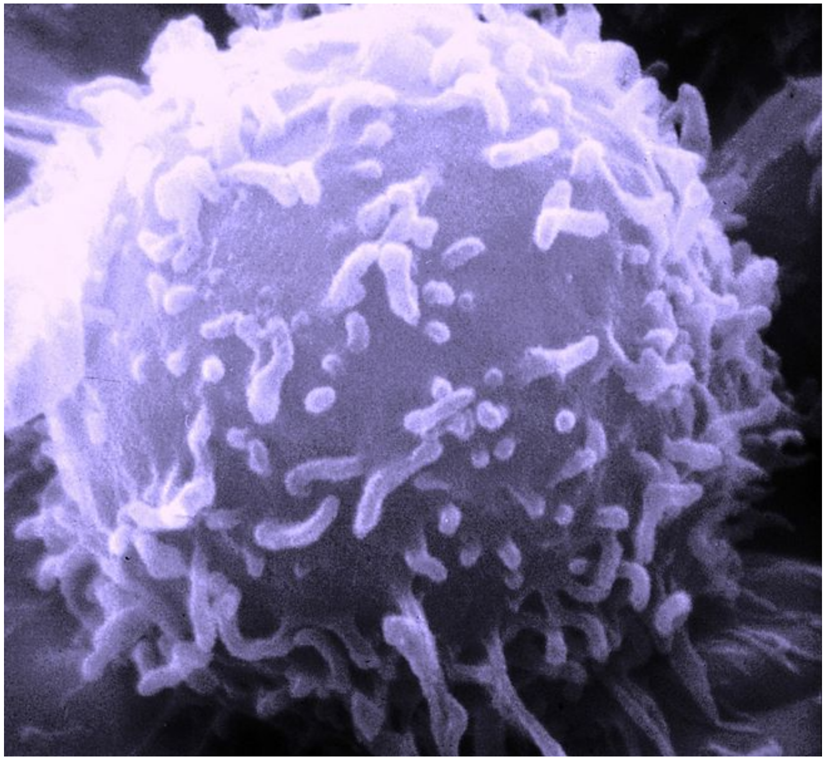
RV144 HIV trial: 2003-2009

- 26,676 volunteers enrolled
- 16,395 volunteers randomized
- 125 infections
- $105,000,000 and 6 years (!!)

Prospective studies are expensive, slow, and entail complex moral issues. This does not lend itself to rapid vaccine development.

How might we guide vaccine development without disease exposure?

Vaccines manipulate the adaptive immune system

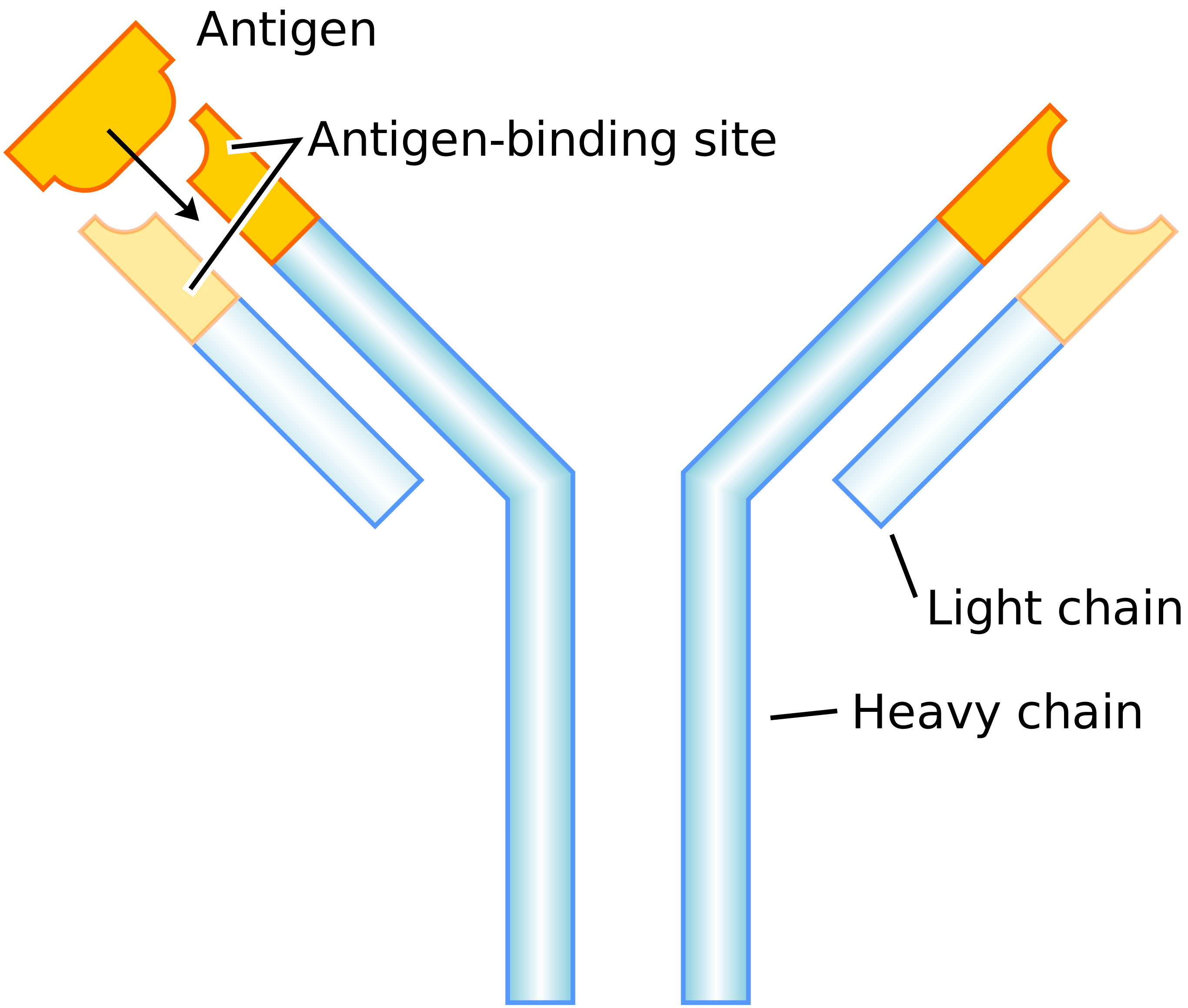


Antibody-making B cells: a key part of adaptive immunity.

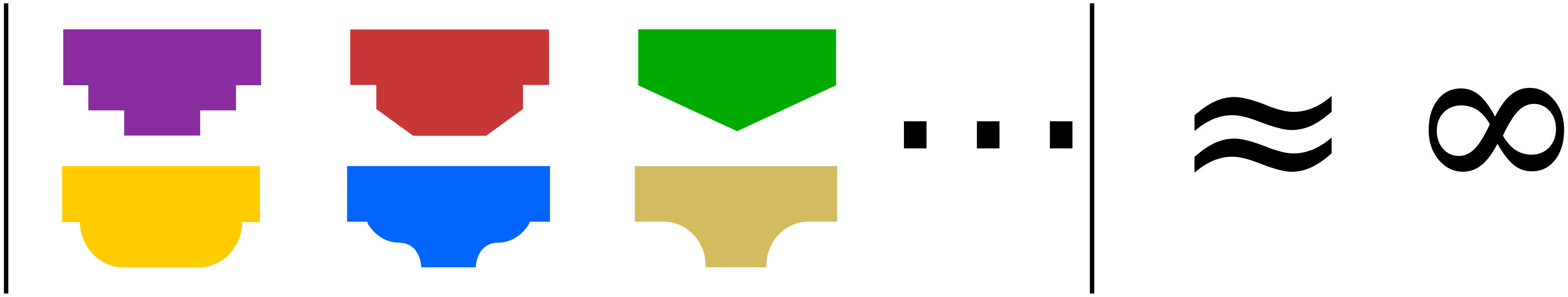
What can we learn from B cells without battle-testing them?

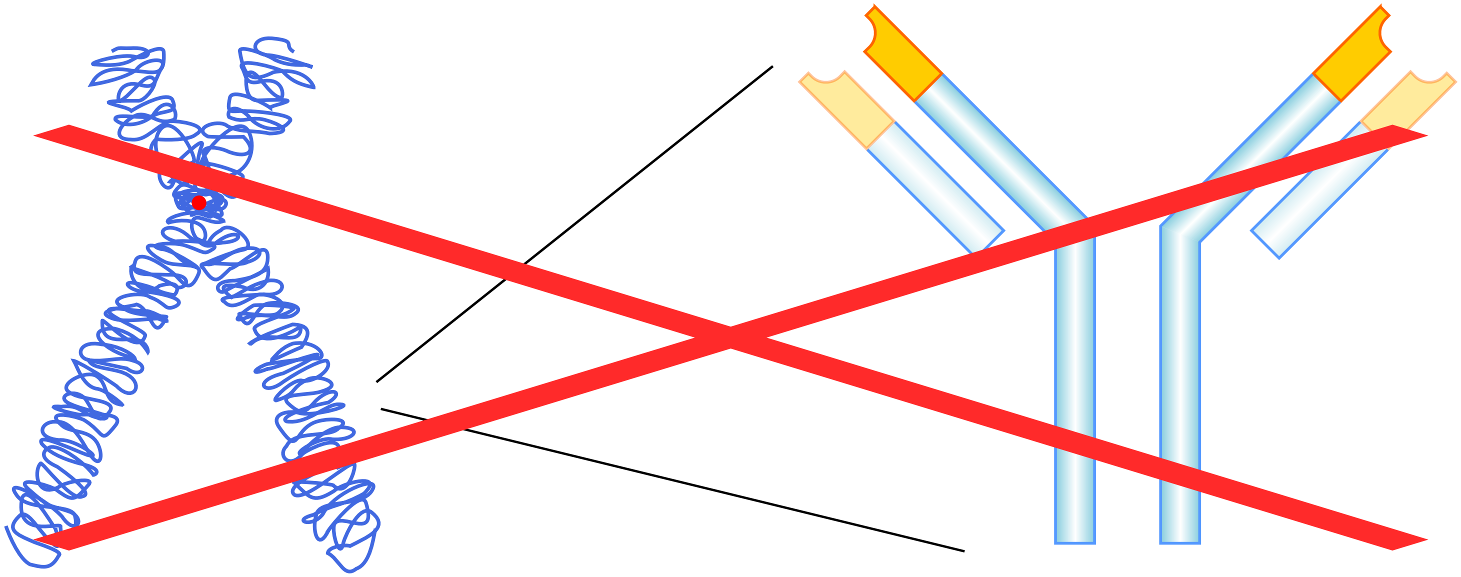
Biological background

Antibodies bind antigens

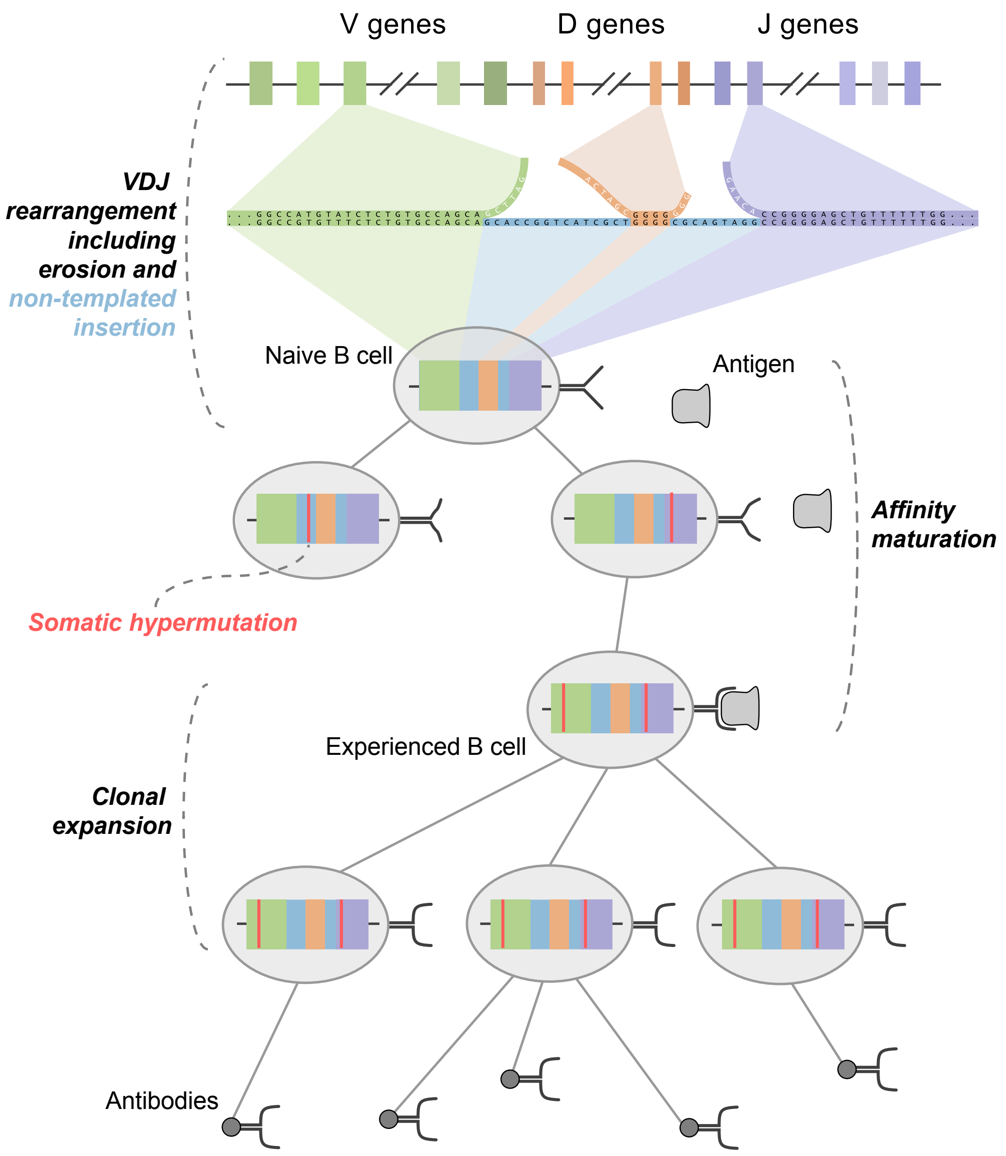


Too many antigens to code for directly

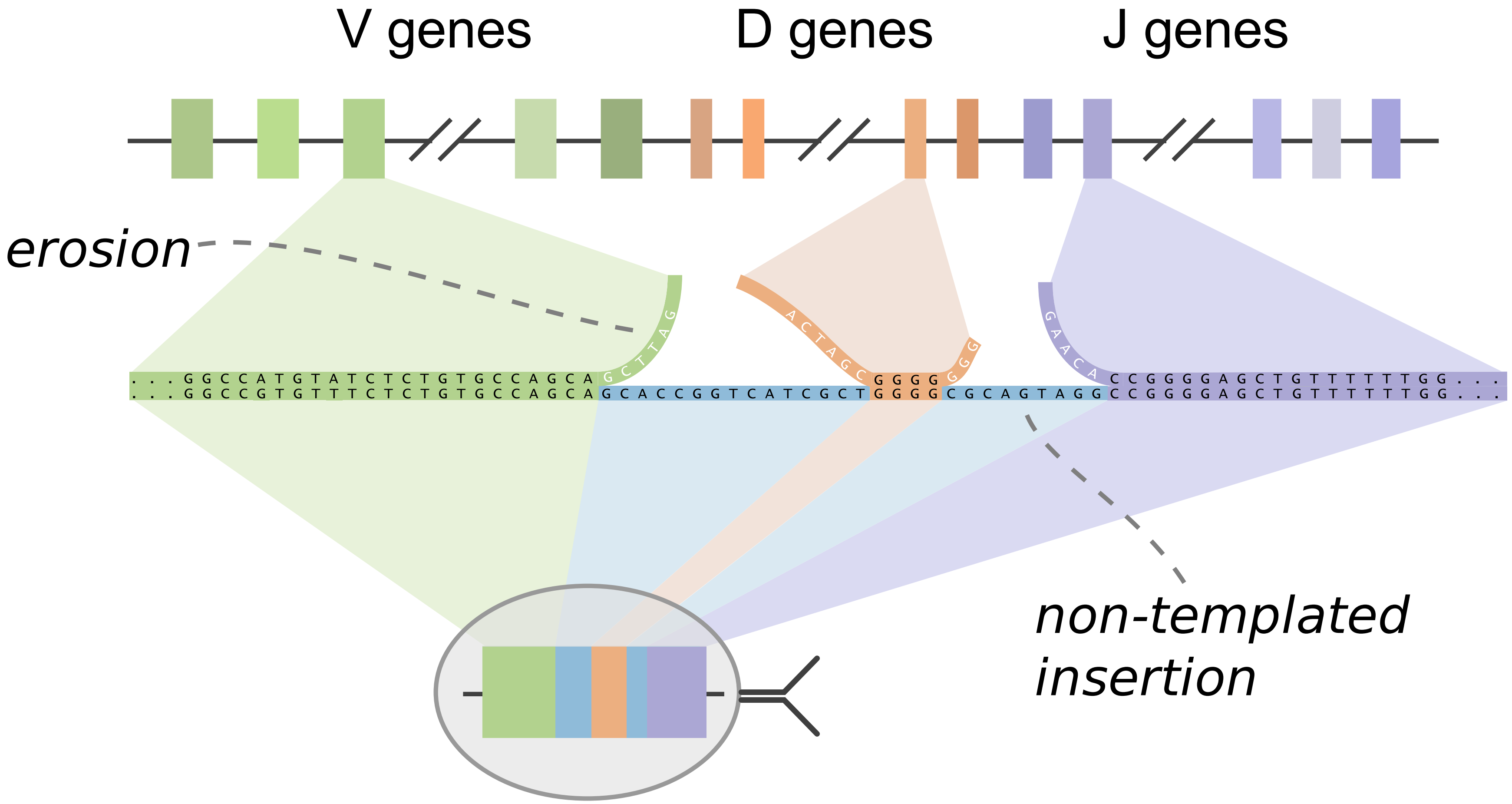




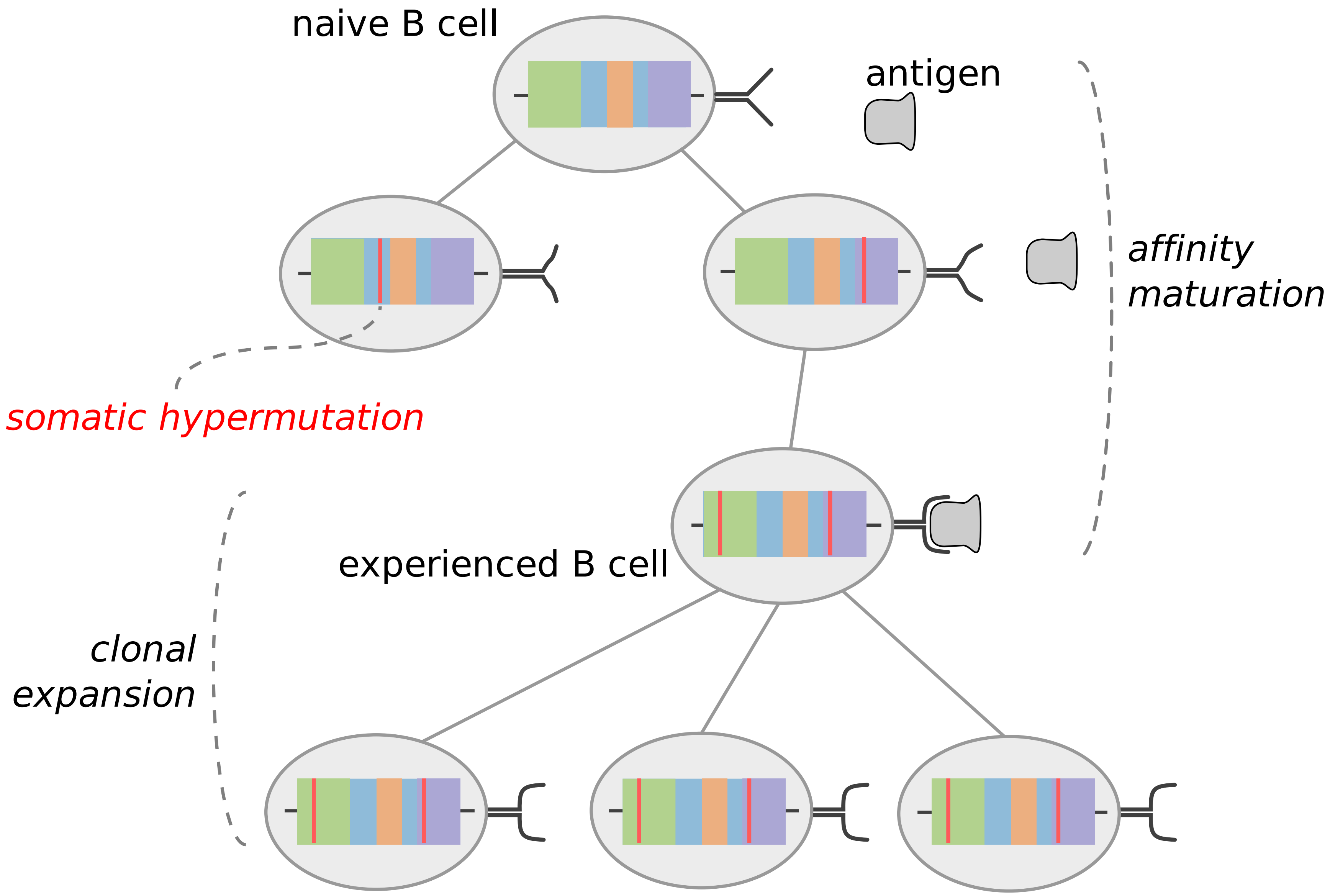
B cell diversification process


V genes
D genes
J genes
VDJ rearrangement including erosion and non-templated insertion
. . . G G C C A T G T A T C T C T G T G C C A G C A
G C T T A G
A C T A G C
G G G G
G G G
G A A C A
C C G G G G A G C T G T T T T T T G G . . .
. . . G G C C G T G T T T C T C T G T G C C A G C A G C A C C G G T C A T C G C T G G G G C G C A G T A G G C C G G G G A G C T G T T T T T T G G . . .
Naive B cell
Antigen
Affinity maturation
Somatic hypermutation
Experienced B cell
Clonal expansion
Antibodies

B cell diversification overview

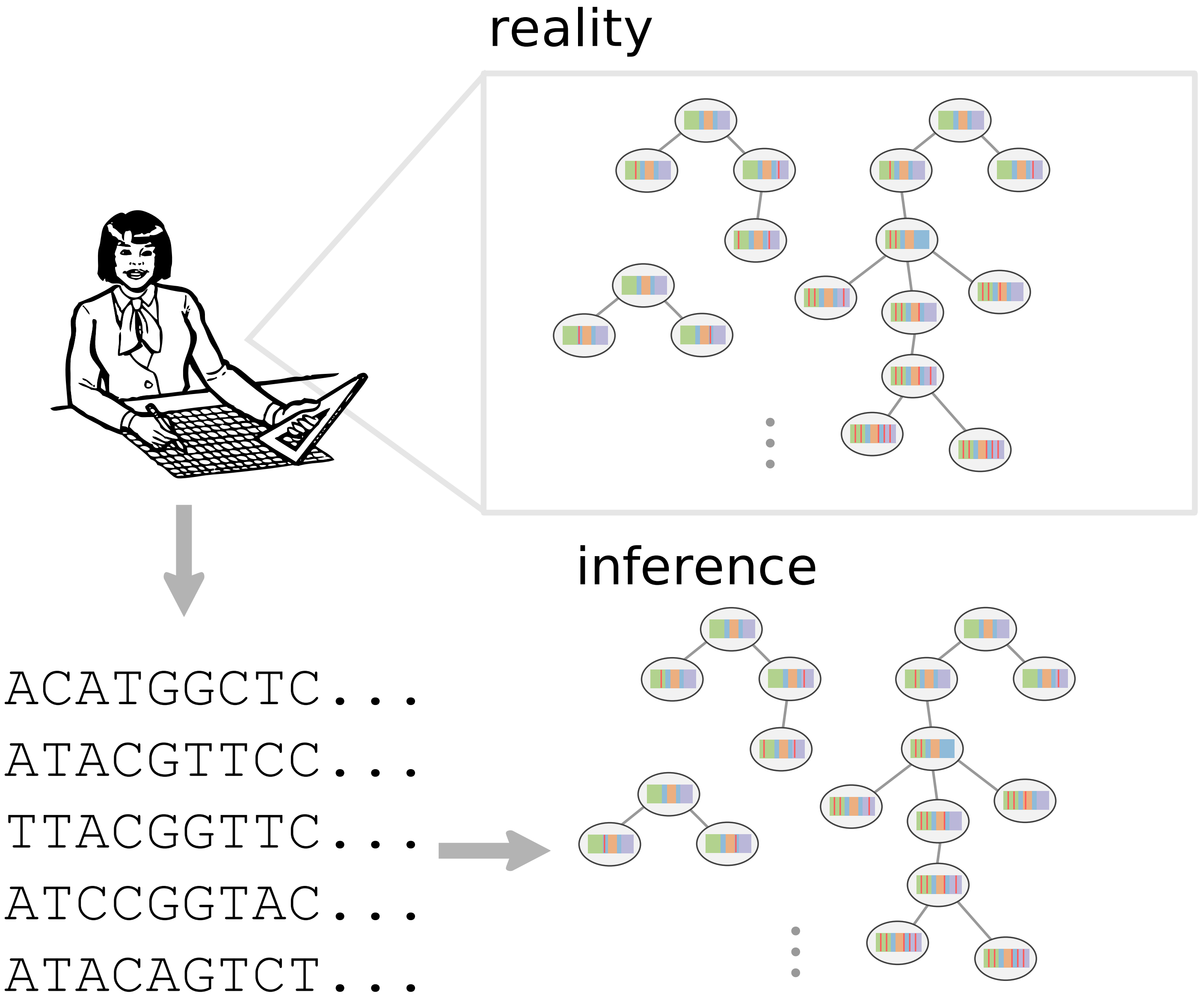

V genes
D genes
J genes
erosion
GCTTAG
ACTAGC
GGGG
GGG
GAACA
...GGCCATGTATCTCTGTGCCAGCA
CCGGGGAGCTGTTTTTTGG...
...GGCCGTGTTTCTCTGTGCCAGCAGCACCGGTCATCGCTGGGGCGCAGTAGGCCGGGGAGCTGTTTTTTGG...
non-templated insertion

B cell diversification overview



Outline

Goal 1: infer immune *history*



reality

inference

ACATGGCTC...

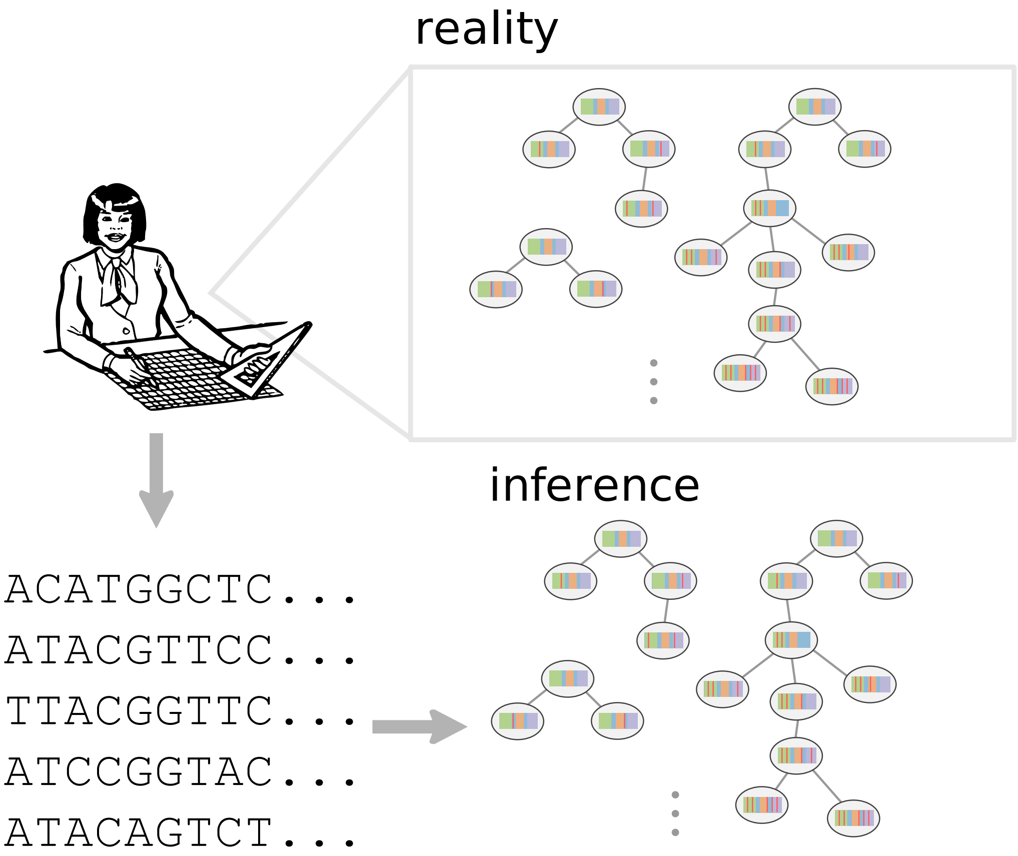
ATACGTTCC...

TTACGGTTC...

ATCCGGTAC...

ATACAGTCT...

Part 1 of talk: find appropriate substitution models



reality

ACATGGCTC...
ATACGTTCC...
TTACGGTTC...
ATCCGGTAC...
ATACAGTCT...

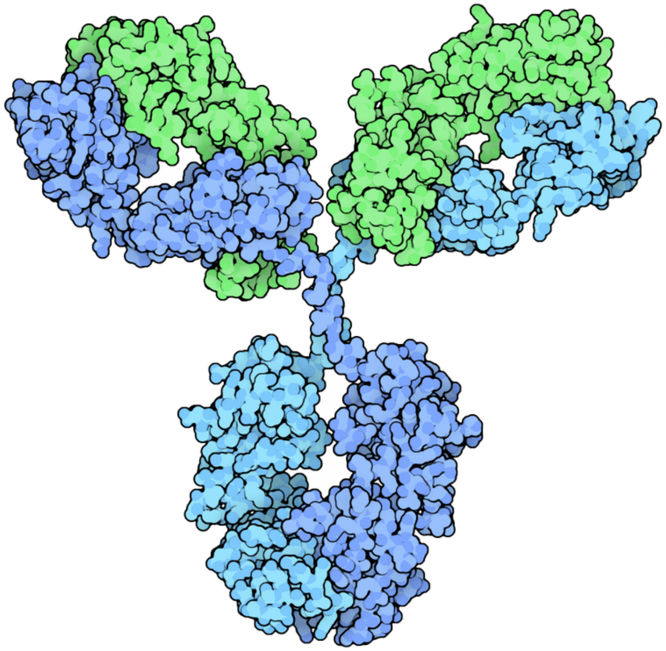
inference

These are needed for likelihood-based phylogenetic inference.

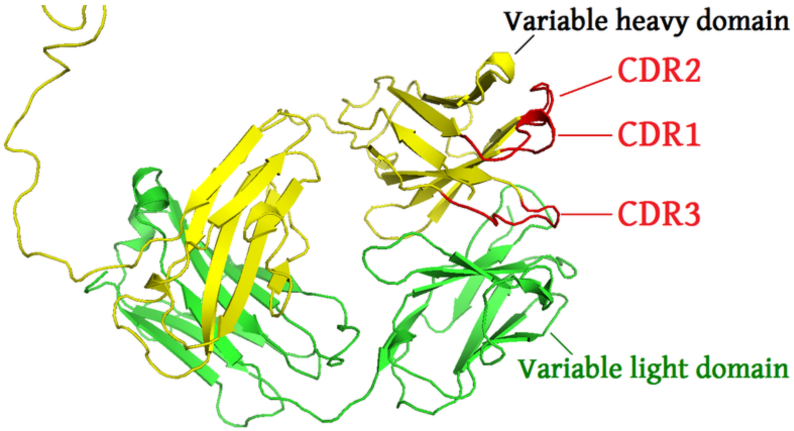
Goal 2: understand how we might *manipulate* immune repertoire with interventions

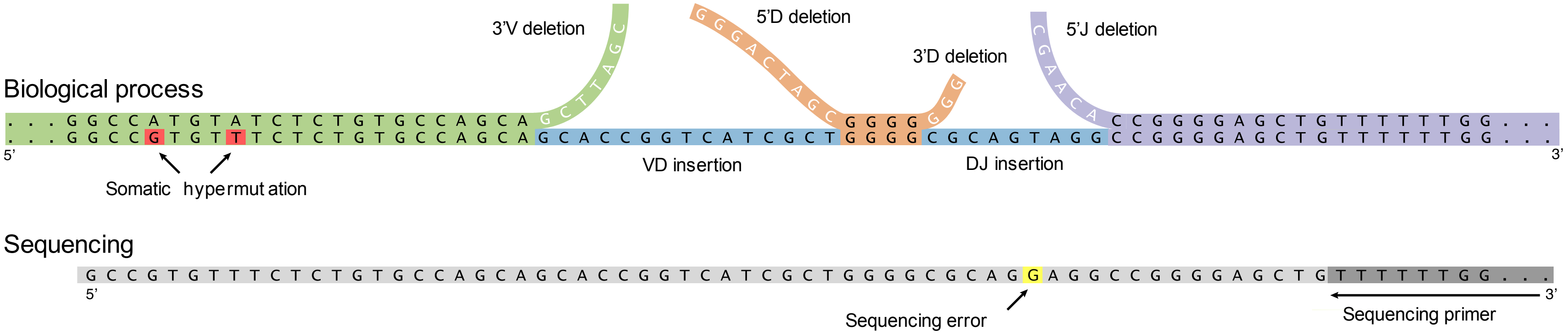
Which sites can be changed?

Part 2 of talk: natural selection inference



The data: sequences from the CDR3 locus



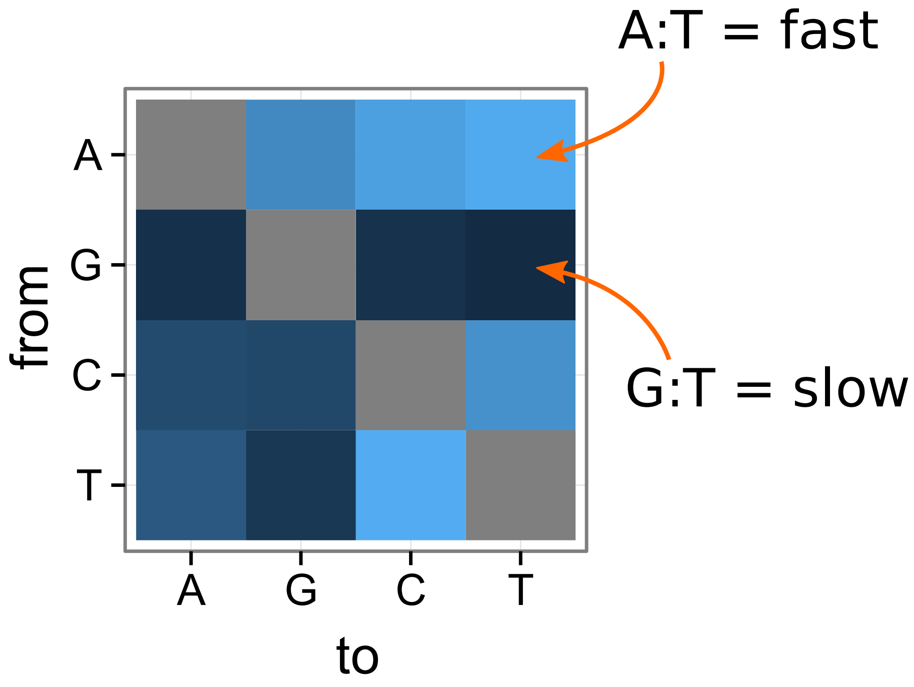


Plenty: a total of about 15 million unique 130nt sequences from memory B cell populations of three healthy individuals A, B, and C.

Part 1

Goal 1: Understand determinants of molecular evolution

Investigate overall mutation patterns of the B cell repertoire.



Models like this are used throughout phylogenetic inference.

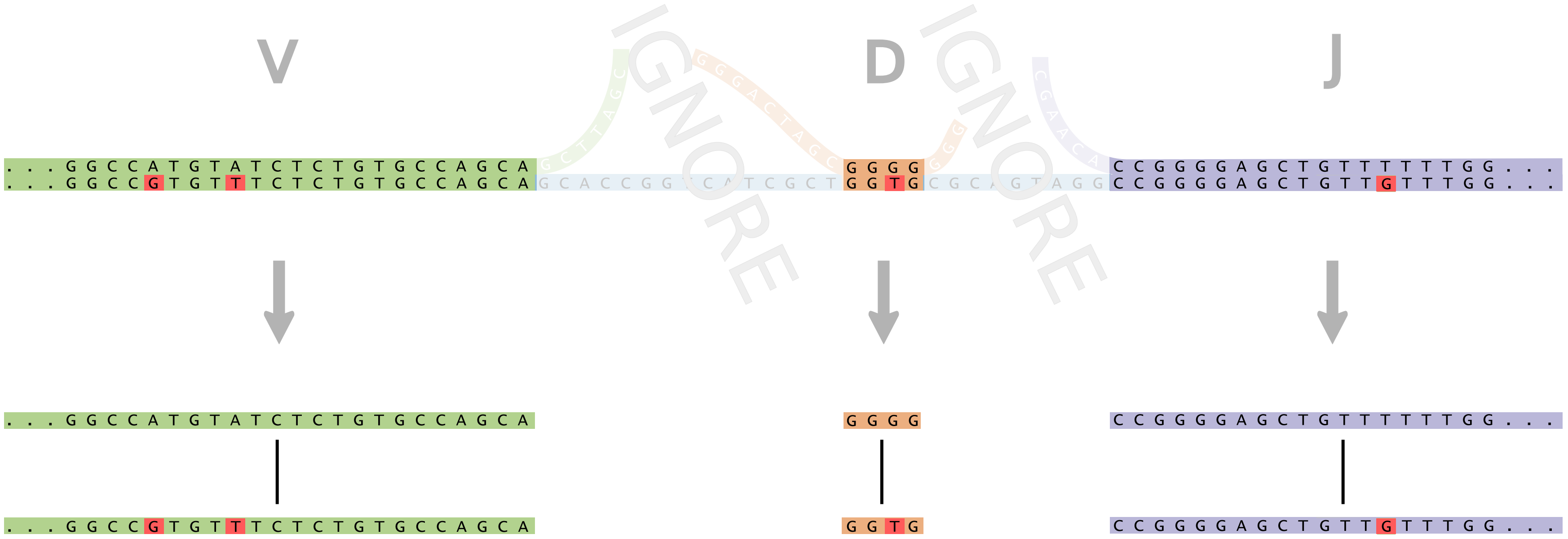
Not all BCRs share ancestry

In fact, most don’t.

This is different from traditional phylogenetics.

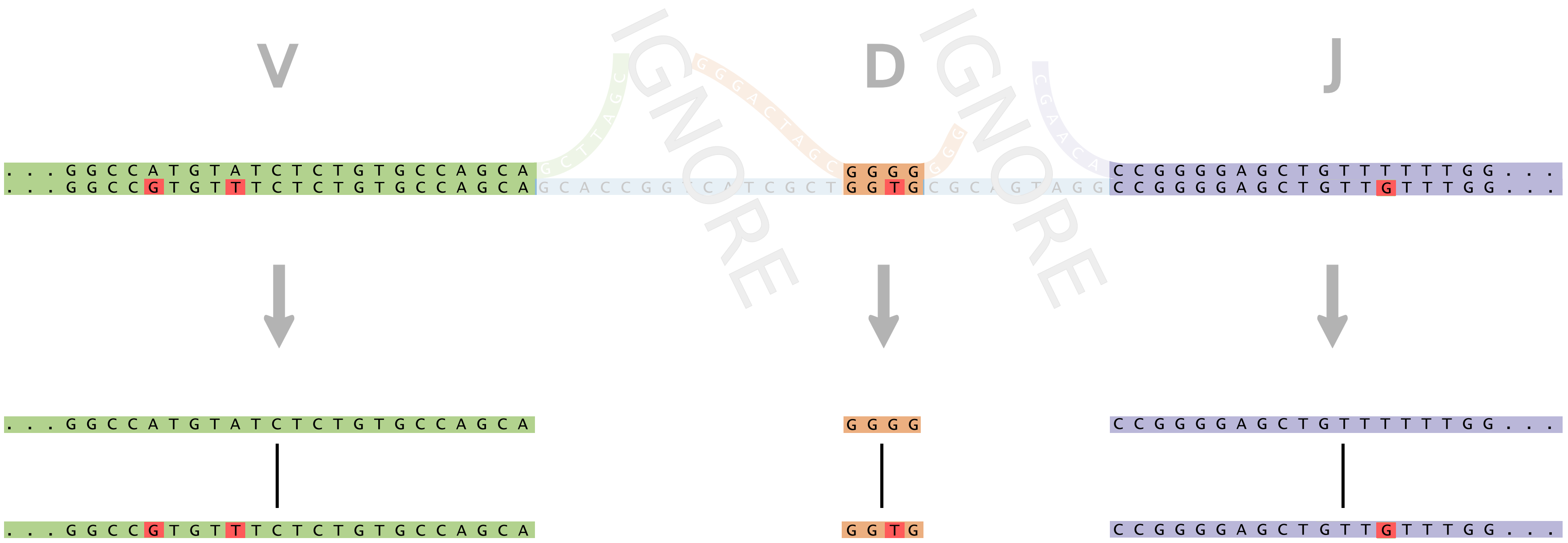
Use two taxon “trees” for model fitting

But: we know ancestral state within V, D, J.


V
D
J
IGNORE
IGNORE
. . . G G C C A T G T A T C T C T G T G C C A G C A
. . . G G C C G T G T T T C T C T G T G C C A G C A
G G G G
G G T G
C C G G G G A G C T G T T T T T T G G . . .
C C G G G G A G C T G T T G T T T G G . . .
. . . G G C C A T G T A T C T C T G T G C C A G C A
. . . G G C C G T G T T T C T C T G T G C C A G C A
G G G G
G G T G
C C G G G G A G C T G T T T T T T G G . . .
C C G G G G A G C T G T T G T T T G G . . .

Our “trees” have an observed read on the bottom and the corresponding “ancestral” germline sequence on top, connected by a branch, representing some amount of divergence.

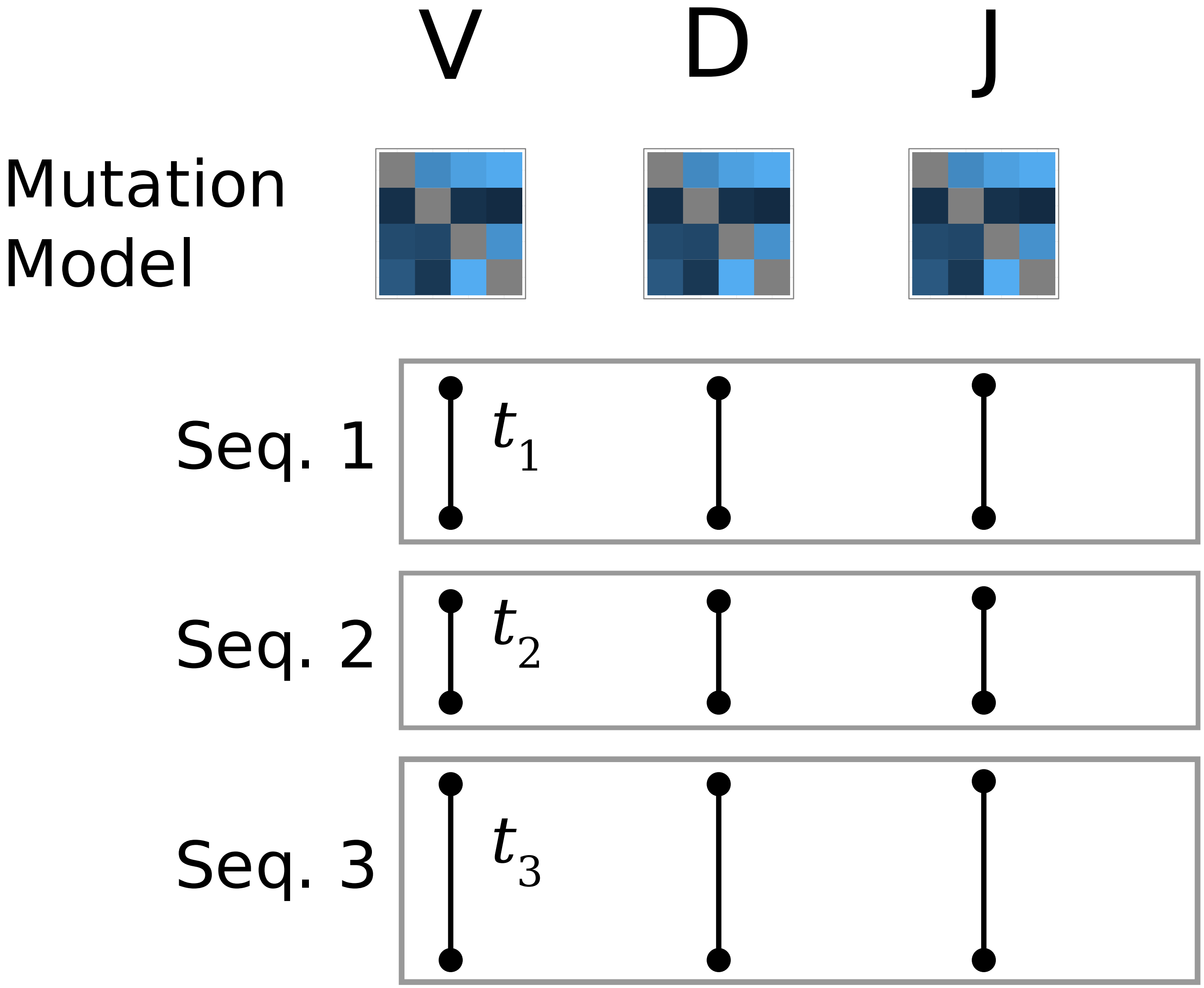
Collection of two taxon "trees" for model fitting


V
D
J
IGNORE
IGNORE
. . . G G C C A T G T A T C T C T G T G C C A G C A
. . . G G C C G T G T T T C T C T G T G C C A G C A
G G G G
G G T G
C C G G G G A G C T G T T T T T T G G . . .
C C G G G G A G C T G T T G T T T G G . . .
. . . G G C C A T G T A T C T C T G T G C C A G C A
. . . G G C C G T G T T T C T C T G T G C C A G C A
G G G G
G G T G
C C G G G G A G C T G T T T T T T G G . . .
C C G G G G A G C T G T T G T T T G G . . .

We will test various models for the V, D, and J segments to select an appropriate evolutionary model for somatic hypermutation to eventually use in phylogenetic inference.

First question: do segments evolve differently?

Simple model



V D J

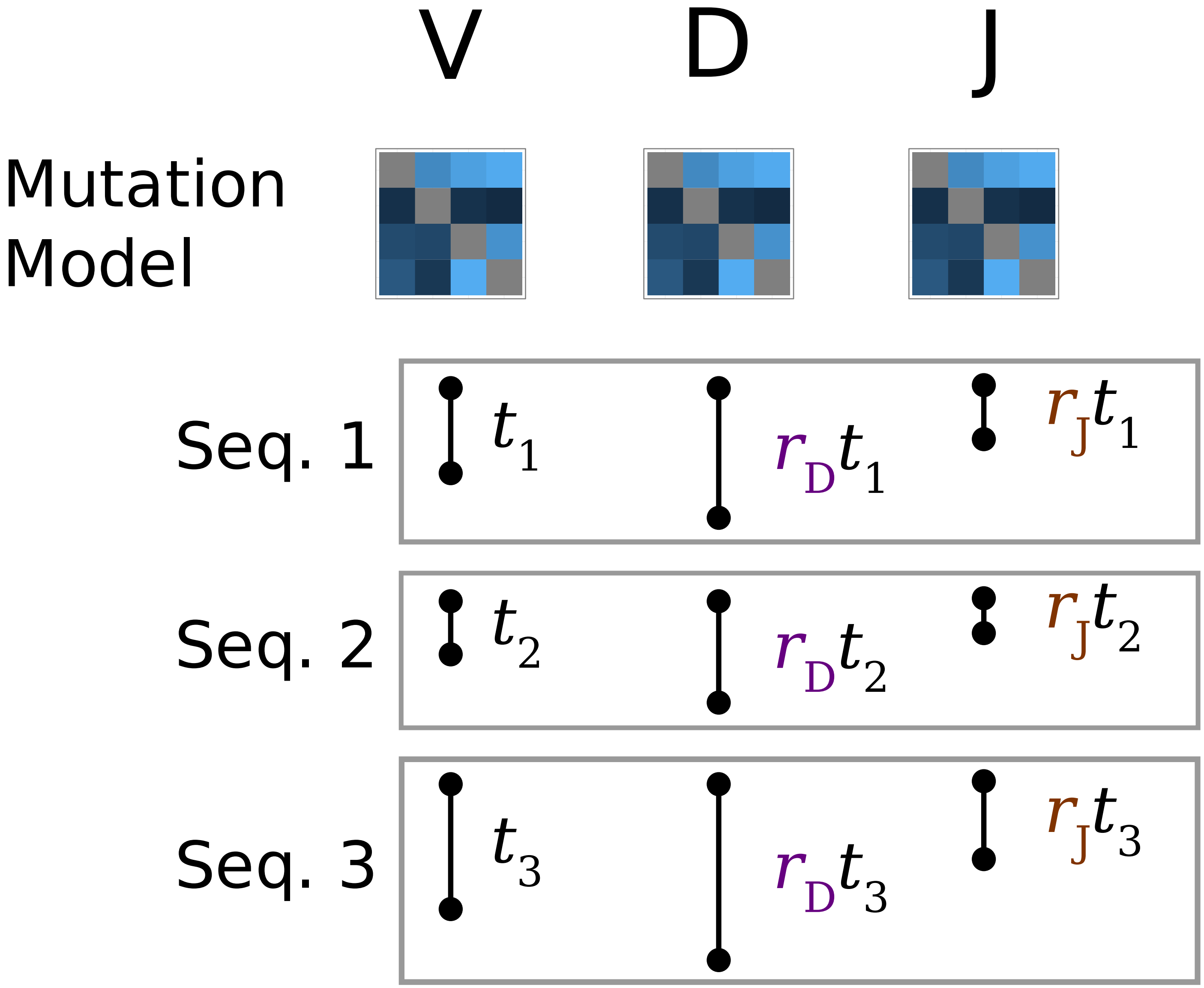
Mutation Model

Seq. 1 t_1

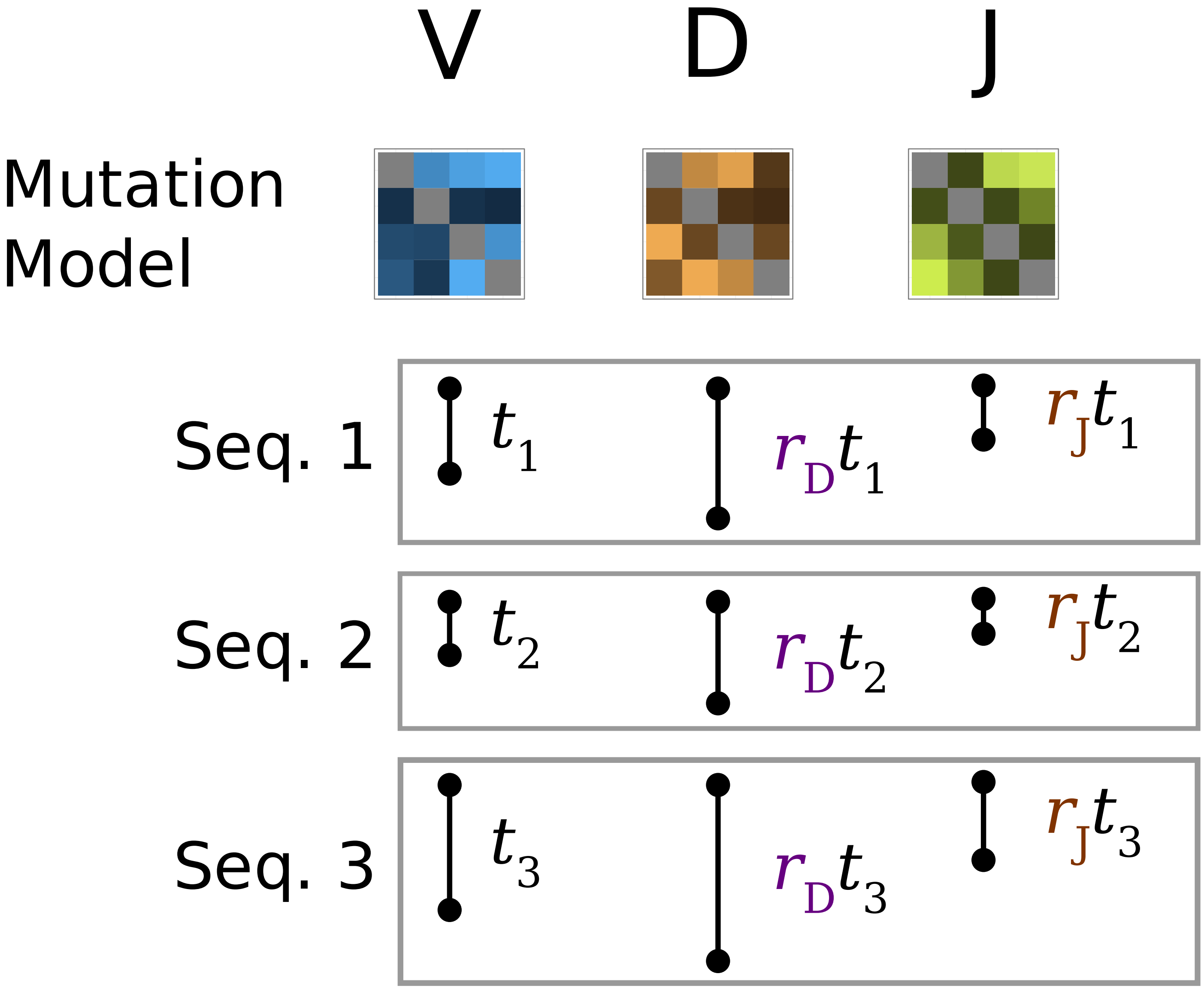
Seq. 2 t_2

Seq. 3 t_3

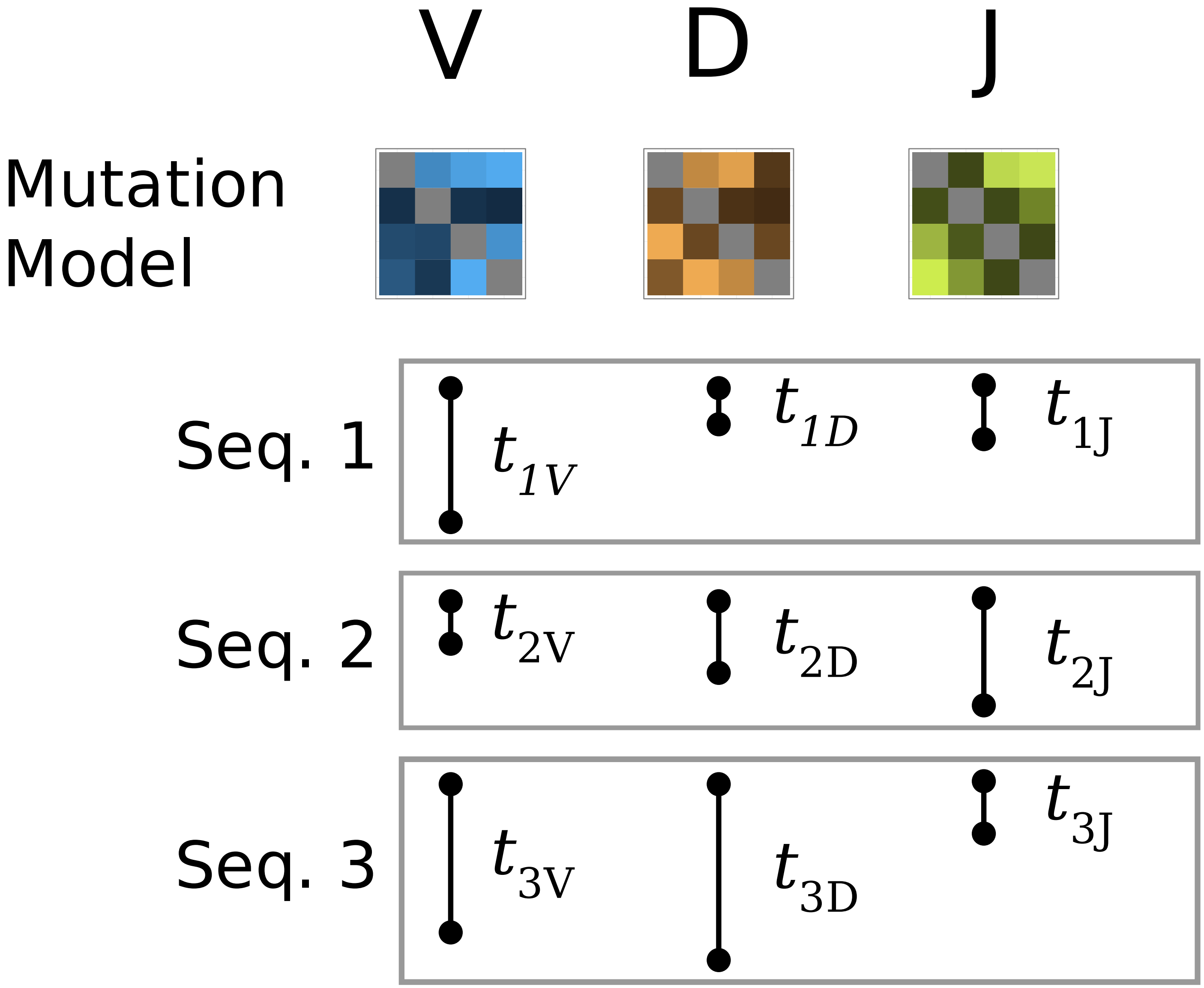
… more complex (i)



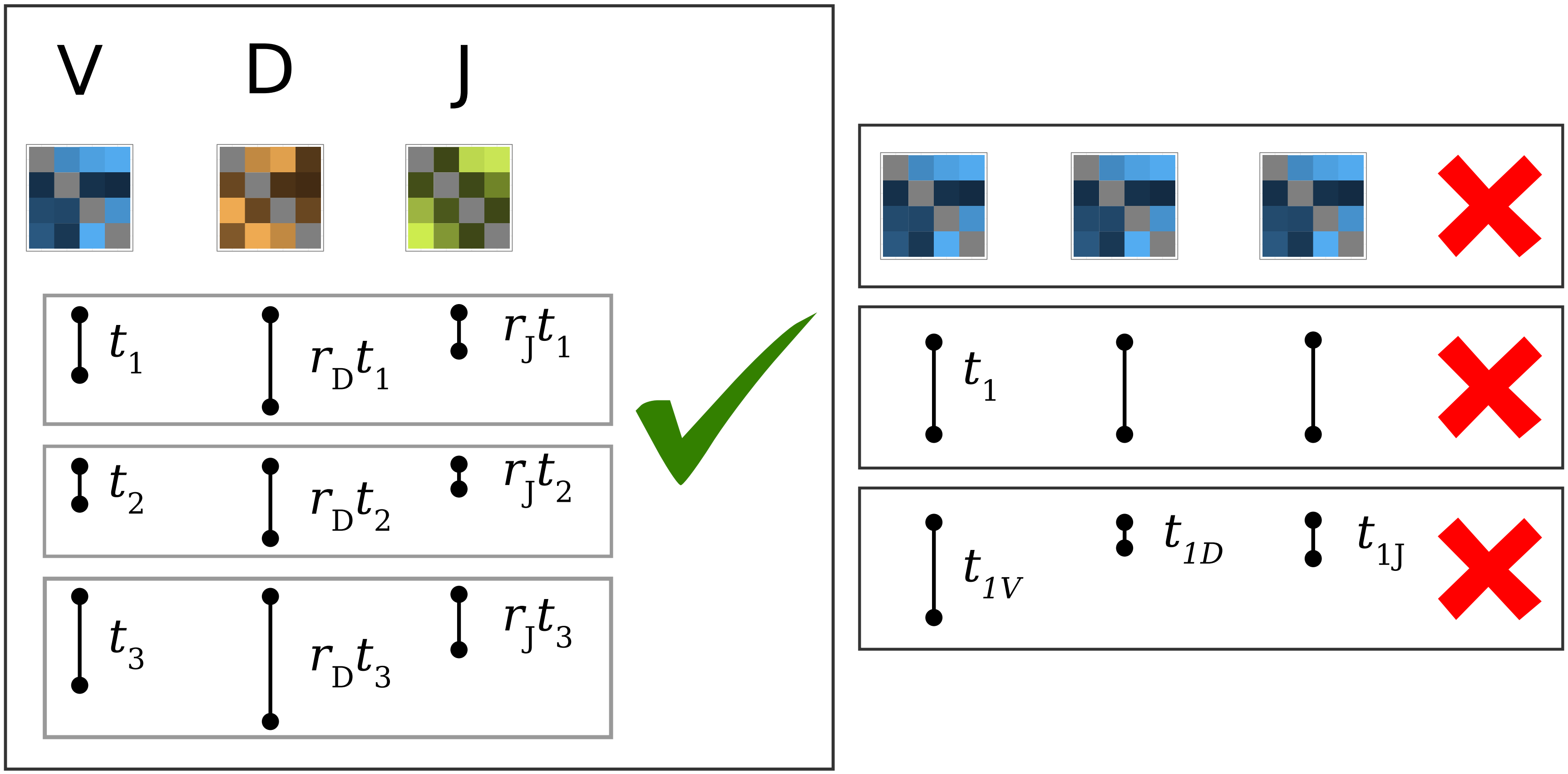
... more complex (ii)



… most complex

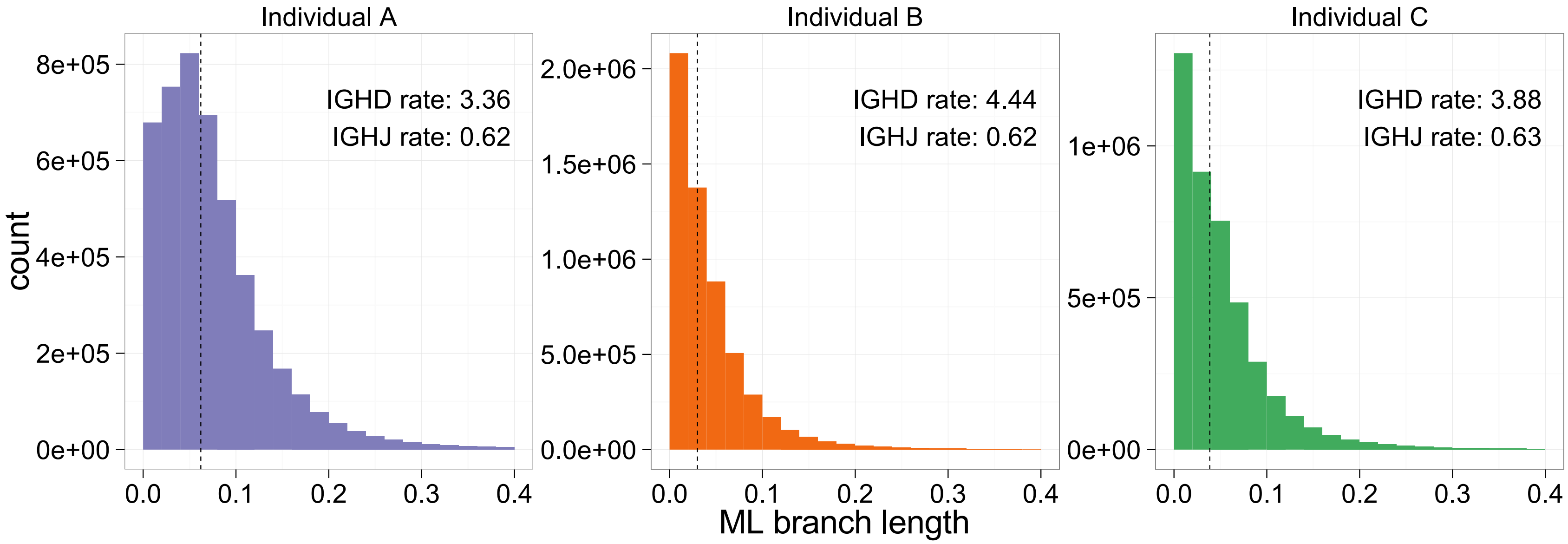


Model testing results



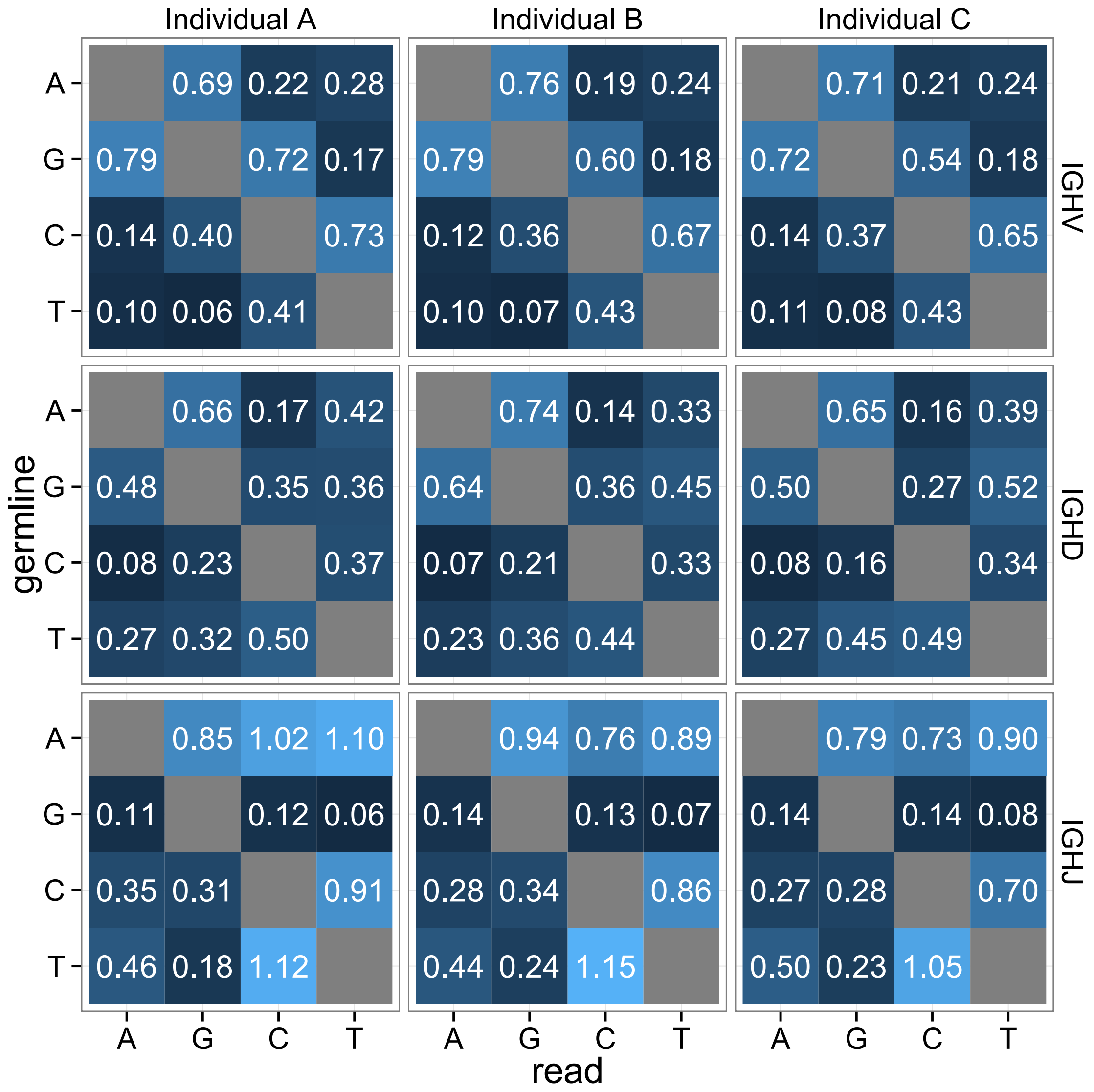
Identical model ranking across individuals (using AIC / BIC).

Branch length distribution under this best model



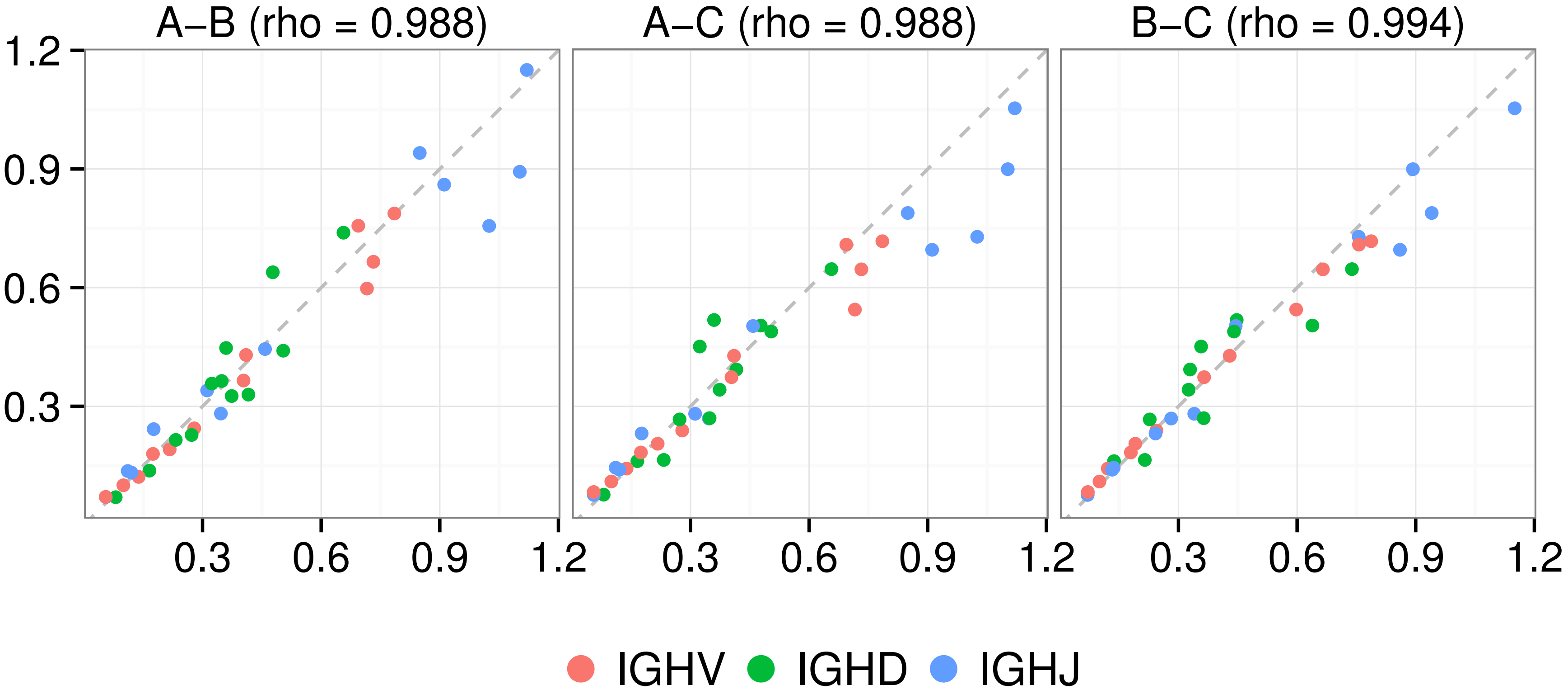
- **D** segments evolve substantially *faster* than **V**
- **J** segments evolve more *slowly* than **V**
- Individual A has a higher mutational load.

Rate matrices for General Time Reversible model



		Individual A				Individual B				Individual C			
		A	G	C	T	A	G	C	T	A	G	C	T
IGHV	A		0.69	0.22	0.28		0.76	0.19	0.24		0.71	0.21	0.24
	G	0.79		0.72	0.17	0.79		0.60	0.18	0.72		0.54	0.18
	C	0.14	0.40		0.73	0.12	0.36		0.67	0.14	0.37		0.65
	T	0.10	0.06	0.41		0.10	0.07	0.43		0.11	0.08	0.43	
IGHD	A		0.66	0.17	0.42		0.74	0.14	0.33		0.65	0.16	0.39
	G	0.48		0.35	0.36	0.64		0.36	0.45	0.50		0.27	0.52
	C	0.08	0.23		0.37	0.07	0.21		0.33	0.08	0.16		0.34
	T	0.27	0.32	0.50		0.23	0.36	0.44		0.27	0.45	0.49	
IGHJ	A		0.85	1.02	1.10		0.94	0.76	0.89		0.79	0.73	0.90
	G	0.11		0.12	0.06	0.14		0.13	0.07	0.14		0.14	0.08
	C	0.35	0.31		0.91	0.28	0.34		0.86	0.27	0.28		0.70
	T	0.46	0.18	1.12		0.44	0.24	1.15		0.50	0.23	1.05	

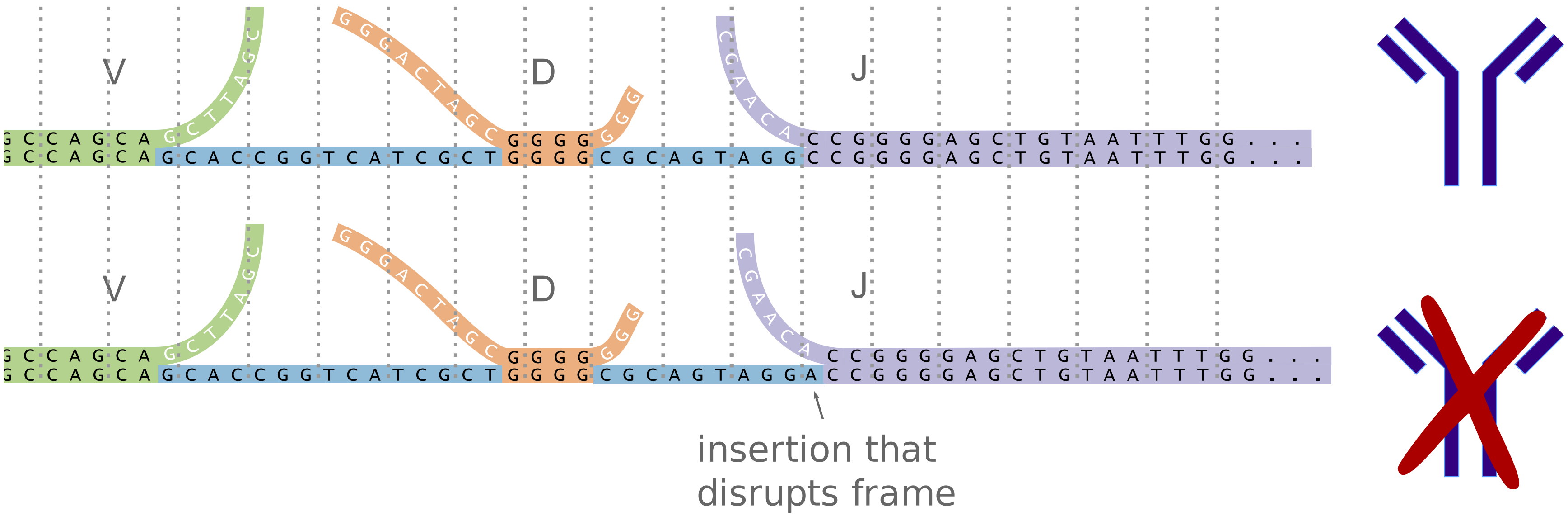
Estimates of the mutational process are quite consistent between individuals



(each point is a single entry for one of the matrices for a pair of individuals.)

(Important) aside: productive versus out-of-frame receptors

Each cell may carry two IGH alleles, but only one is expressed.


V
D
J
G C T T A G C
G G G A C T A G C
G G G
C G A A C A
G C C A G C A G G G G C C G G G G A G C T G T A A T T T G G . . .
G C C A G C A G C A C C G G T C A T C G C T G G G G C G C A G T A G G C C G G G G A G C T G T A A T T T G G . . .
V
D
J
G C T T A G C
G G G A C T A G C
G G G
C G A A C A
G C C A G C A G G G G C C G G G G A G C T G T A A T T T G G . . .
G C C A G C A G C A C C G G T C A T C G C T G G G G C G C A G T A G G A C C G G G G A G C T G T A A T T T G G . . .
insertion that
disrupts frame

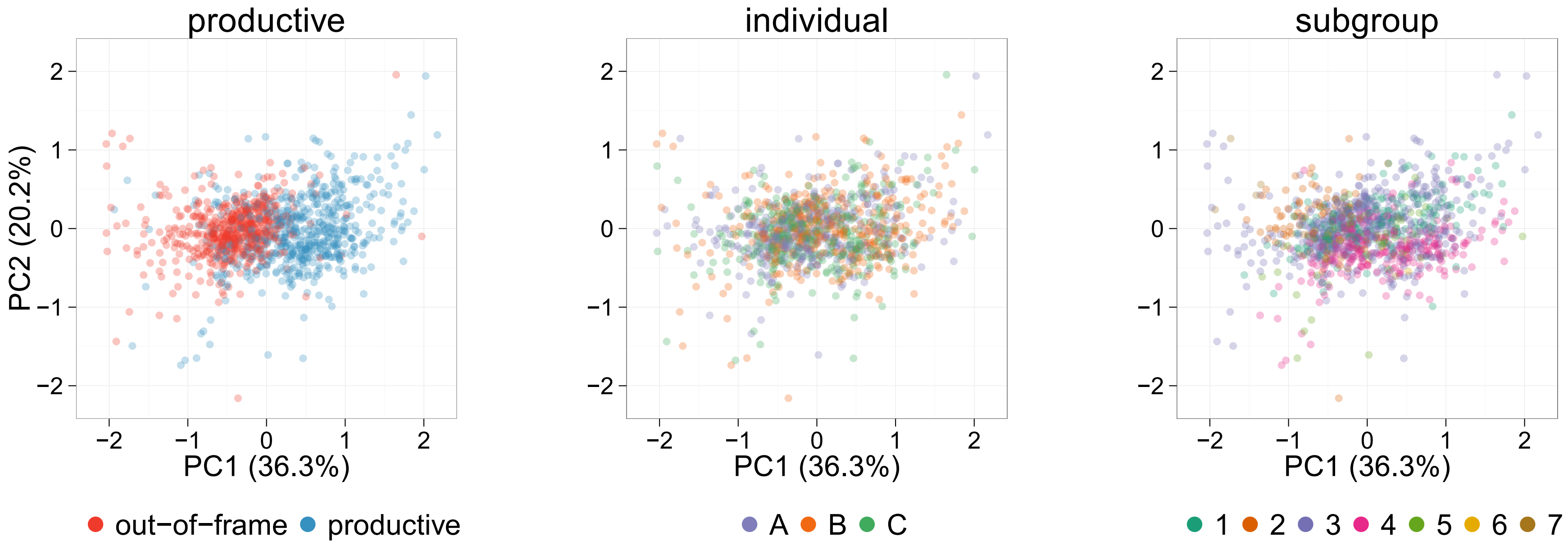
Next: what determines mutational processes of different IGHV genes?

Subdivide V genes:

- by individual (A, B, and C)
- by gene (V1-18, V2-3, etc)
- by productive / unproductive status

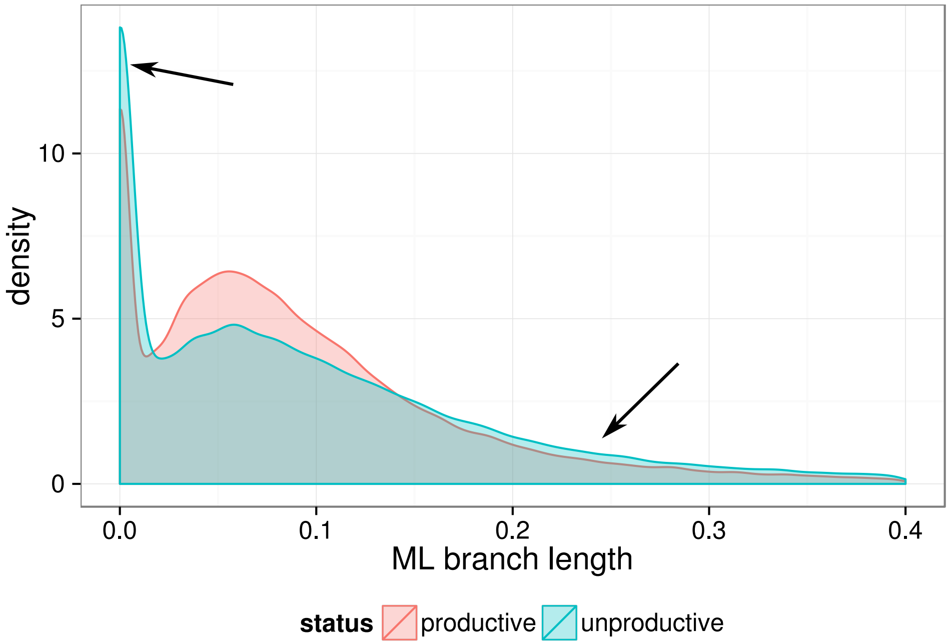
and fit each subset separately.

Principal components analysis of individual IGHV GTR matrices



Inspired by the work of (Kosakovsky Pond *et al*, 2010) on evolutionary "fingerprinting."

Branch length differences between productive, unproductive



Unproductive rearrangements are more likely to be either: unchanged from germline, or more divergent.

Wrap-up of Part 1

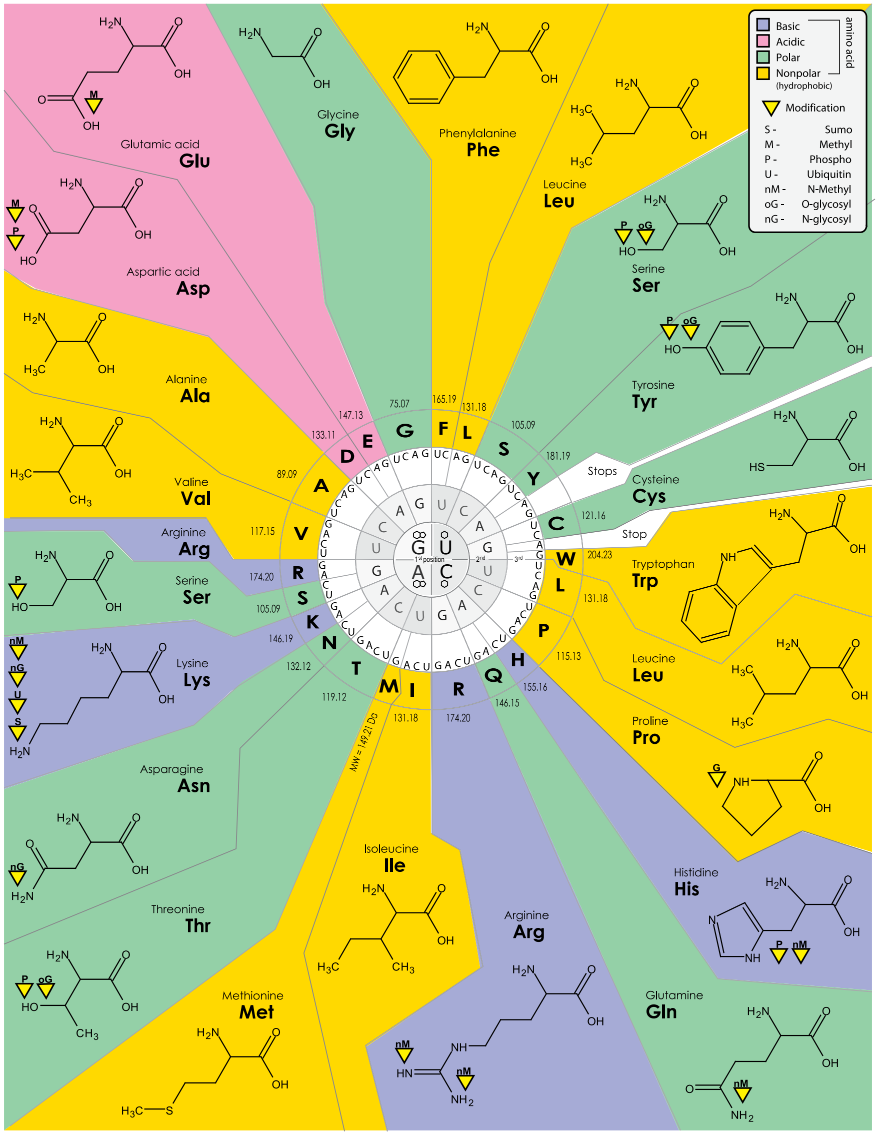
- We find that the data support a moderately complex evolutionary model; similar between individuals
- Mutation process of rearranged IGHV genes primarily varies by in-frame/out-of-frame status, with almost no per-individual signal and a bit of gene group-level signal

Part 2

Goal 2: what if we want to mutate specific residues in an antibody. Is that allowed?

We can’t answer that directly, but we can look across the repertoire at which *sites have tolerated change.*

Genetic code degeneracy is a gift to molecular evolutionists enabling selection inference



Pro → Pro

CCA → CCT

synonymous

Thr → Ile

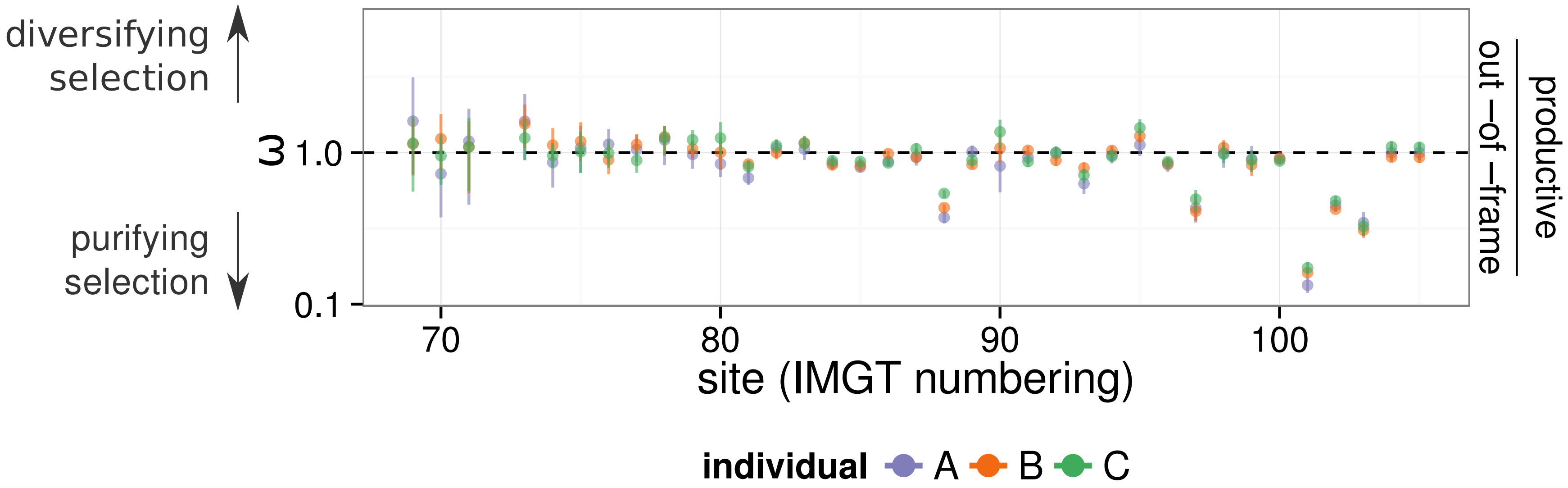
ACC → ATC

nonsynonymous

This is (natural) selection inference

$$\omega \equiv \frac{dN}{dS} \equiv \frac{\text{rate of non-synonymous substitution}}{\text{rate of synonymous substitution}}$$

We want to estimate this value for each site:



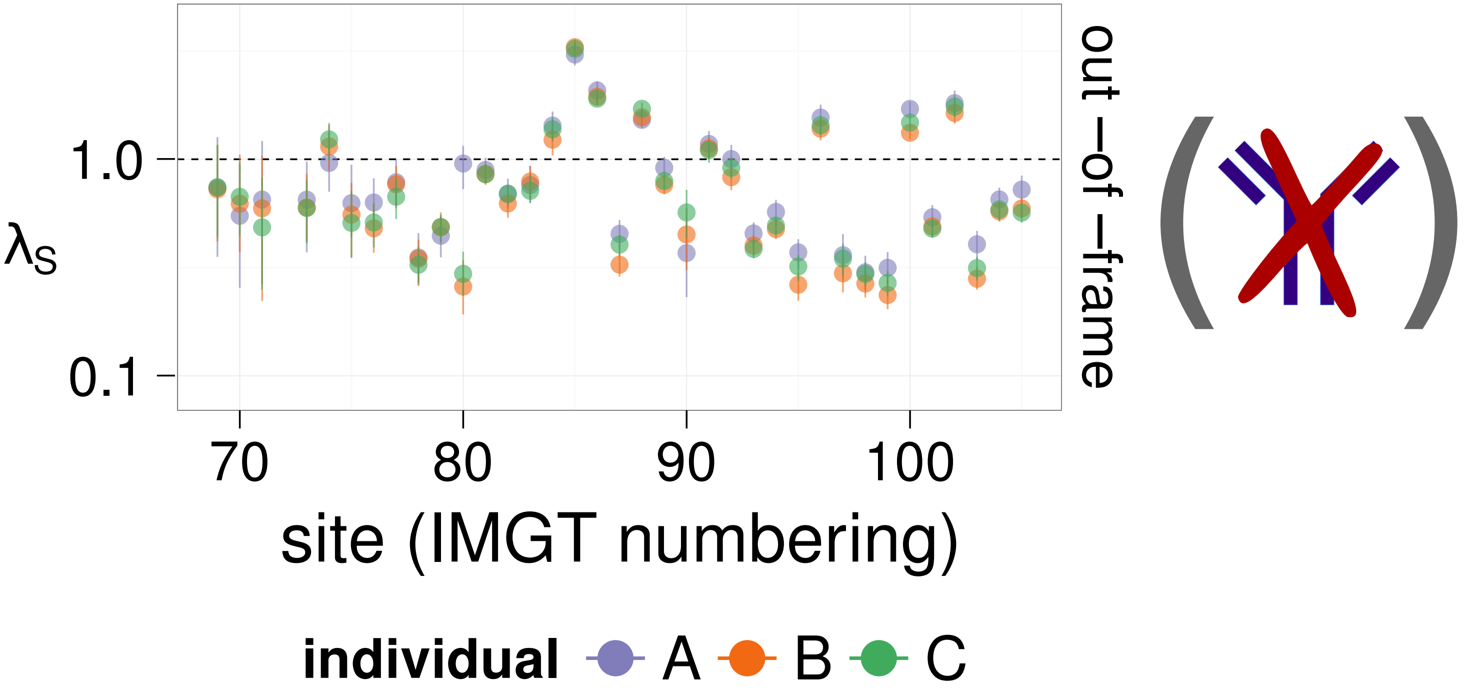
Challenges

- strange mutational process
- millions of unique sequences
  (rules out otherwise lovely tools like PAML, HYPHY):

> *"FUBAR [HYPHY] allows us to analyze larger data sets than other methods: We illustrate this on a large influenza hemagglutinin data set (3,142 sequences)" – Murrell et. al 2013*

Strange mutational process

Per-site inference is made difficult by a complicated mutation process



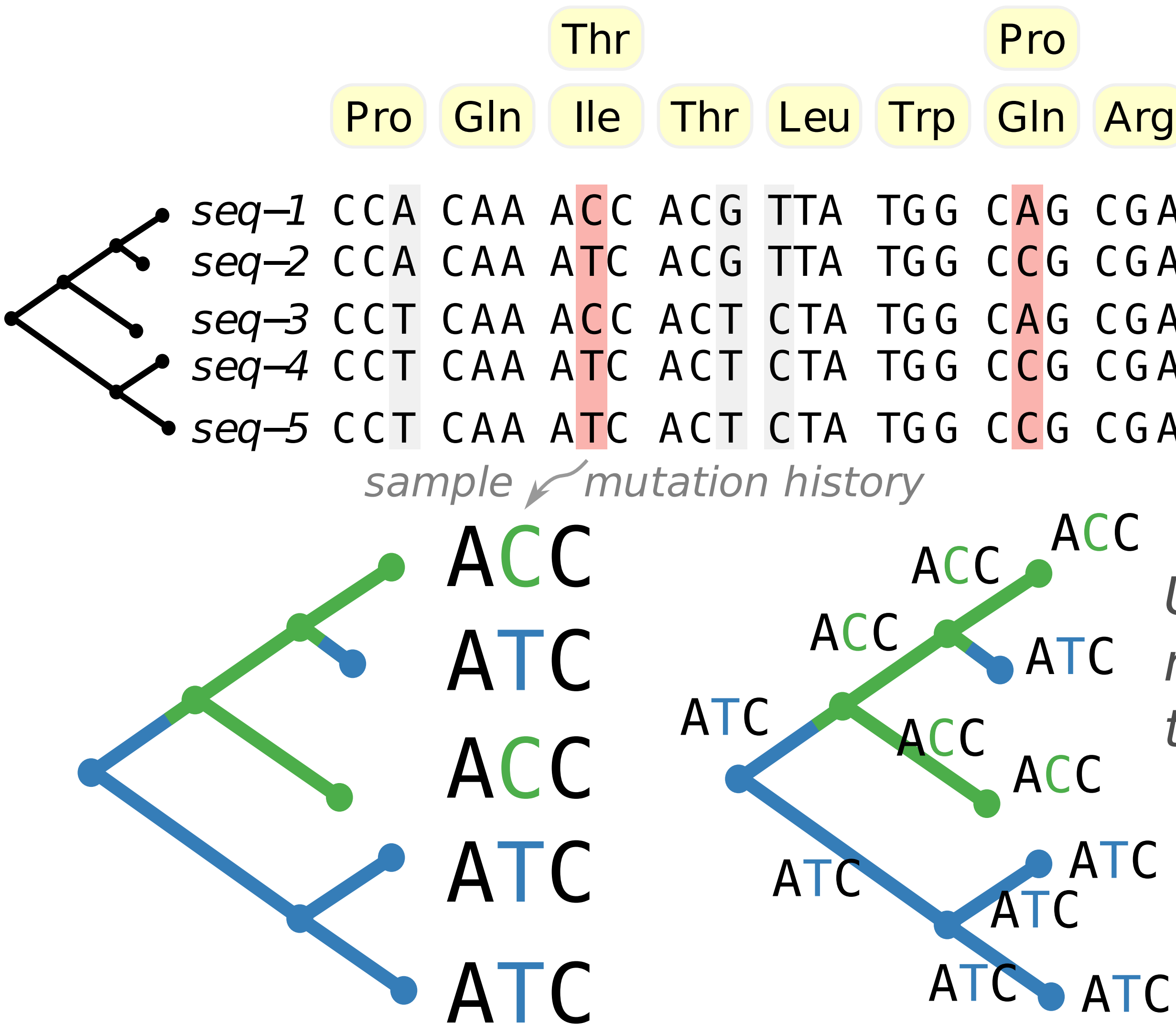
We can use this to tell us about the neutral mutation process.

ω_l is a ratio of rates in terms of observed neutral process

- $\lambda_l^{(N-I)}$: nonsynonymous in-frame rate for site l
- $\lambda_l^{(N-O)}$: nonsynonymous out-of-frame rate for site l
- $\lambda_l^{(S-I)}$: synonymous in-frame rate for site l
- $\lambda_l^{(S-O)}$: synonymous out-of-frame rate for site l

$$\omega_l = \frac{\lambda_l^{(N-I)}/\lambda_l^{(N-O)}}{\lambda_l^{(S-I)}/\lambda_l^{(S-O)}}$$

Renaissance counting! (Lemey, Minin, ... 2012)



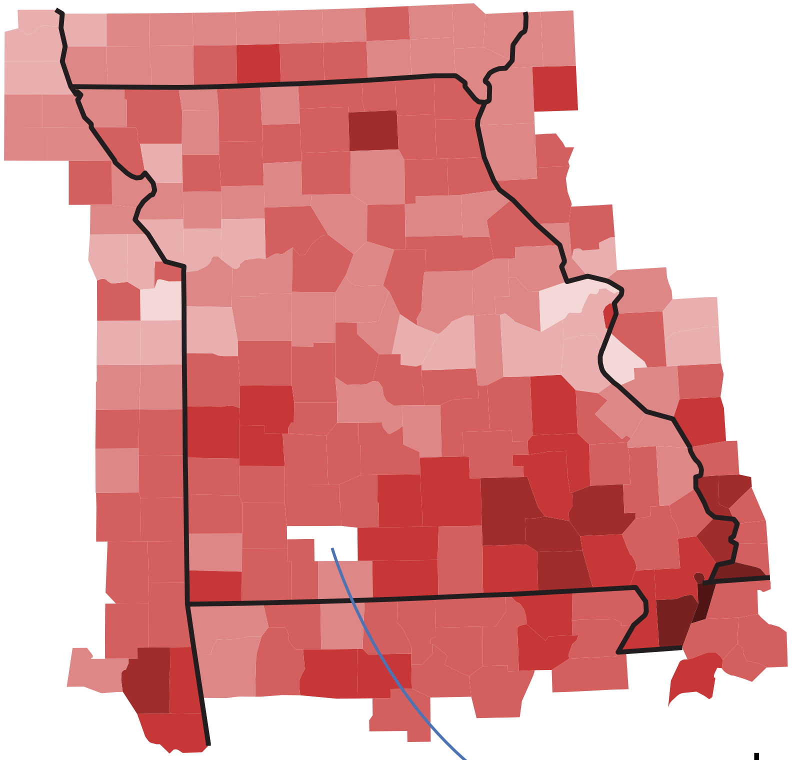
```
seq–1 CCA CAA ACC ACG TTA TGG CAG CGA
seq–2 CCA CAA ATC ACG TTA TGG CCG CGA
seq–3 CCT CAA ACC ACT CTA TGG CAG CGA
seq–4 CCT CAA ATC ACT CTA TGG CCG CGA
seq–5 CCT CAA ATC ACT CTA TGG CCG CGA
```

Use sampled mutation histories to estimate rates...

but such estimates can be unstable.

Stabilize with empirical Bayes regularization

Say we are doing a per-county smoking survey.



zero smokers? Really?

Use all of the data to fit prior distribution of smoking prevalence, then given observations obtain per-county posterior.

Stabilize with empirical Bayes regularization

Assume that λ_l, the substitution rate at site l, comes from a Gamma distribution with shape α and rate β:

$$\lambda_l \sim \mathrm{Gamma}(\alpha, \beta).$$

Model total substitution counts (sampled via stochastic mapping) for a site as Poisson with rate λ_l:

$$C_l \sim \mathrm{Poisson}(\lambda_l),$$

Fit $\hat{\alpha}$ and $\hat{\beta}$ to *all* data, then draw rates λ_l from the posterior:

$$\lambda_l \mid C_l \sim \mathrm{Gamma}(C_l + \hat{\alpha}, 1 + \hat{\beta}).$$

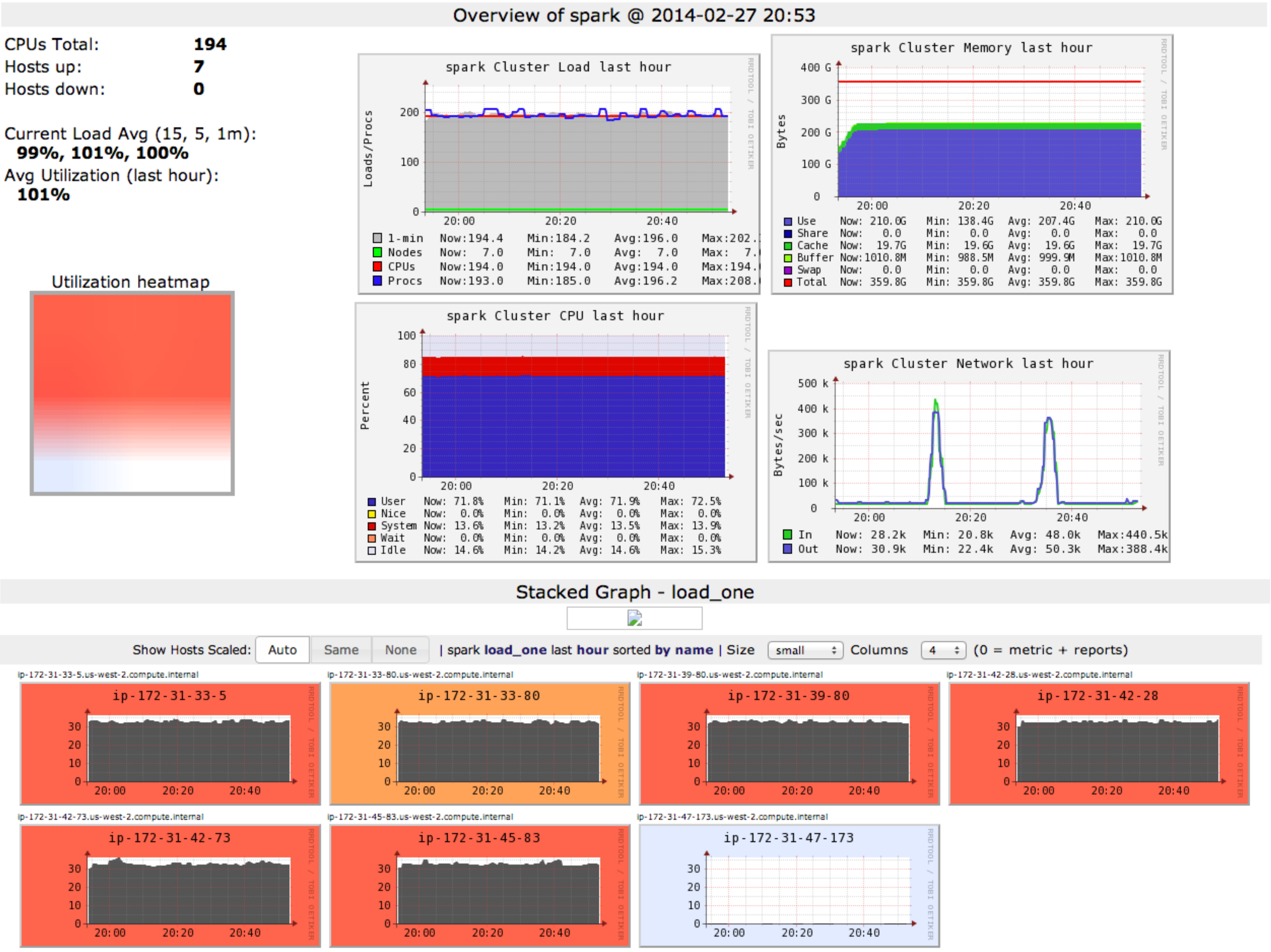
We extended this regularization to case of non-constant coverage.

Estimating selection coefficient ω_l

- $\lambda_l^{(N-I)}$: nonsynonymous in-frame rate for site l
- $\lambda_l^{(N-O)}$: nonsynonymous out-of-frame rate for site l
- $\lambda_l^{(S-I)}$: synonymous in-frame rate for site l
- $\lambda_l^{(S-O)}$: synonymous out-of-frame rate for site l

$$\omega_l = \frac{\lambda_l^{(N-I)}/\lambda_l^{(N-O)}}{\lambda_l^{(S-I)}/\lambda_l^{(S-O)}}$$

Used Spark Map-Reduce engine on EC2



Overview of spark @ 2014-02-27 20:53

CPUs Total: **194**
Hosts up: **7**
Hosts down: **0**

Current Load Avg (15, 5, 1m):
99%, 101%, 100%
Avg Utilization (last hour):
101%

Utilization heatmap

```
spark Cluster Load last hour
1-min   Now:194.4   Min:184.2   Avg:196.0   Max:202.
Nodes   Now:  7.0   Min:  7.0   Avg:  7.0   Max:  7.
CPUs    Now:194.0   Min:194.0   Avg:194.0   Max:194.
Procs   Now:193.0   Min:185.0   Avg:196.2   Max:208.
```

```
spark Cluster Memory last hour
Use    Now: 210.0G   Min: 138.4G   Avg: 207.4G   Max: 210.0G
Share  Now:   0.0    Min:   0.0    Avg:   0.0    Max:   0.0
Cache  Now:  19.7G   Min:  19.6G   Avg:  19.6G   Max:  19.7G
Buffer Now:1010.8M   Min: 988.5M   Avg: 999.9M   Max:1010.8M
Swap   Now:   0.0    Min:   0.0    Avg:   0.0    Max:   0.0
Total  Now: 359.8G   Min: 359.8G   Avg: 359.8G   Max: 359.8G
```

```
spark Cluster CPU last hour
User    Now: 71.8%   Min: 71.1%   Avg: 71.9%   Max: 72.5%
Nice    Now:  0.0%   Min:  0.0%   Avg:  0.0%   Max:  0.0%
System  Now: 13.6%   Min: 13.2%   Avg: 13.5%   Max: 13.9%
Wait    Now:  0.0%   Min:  0.0%   Avg:  0.0%   Max:  0.0%
Idle    Now: 14.6%   Min: 14.2%   Avg: 14.6%   Max: 15.3%
```

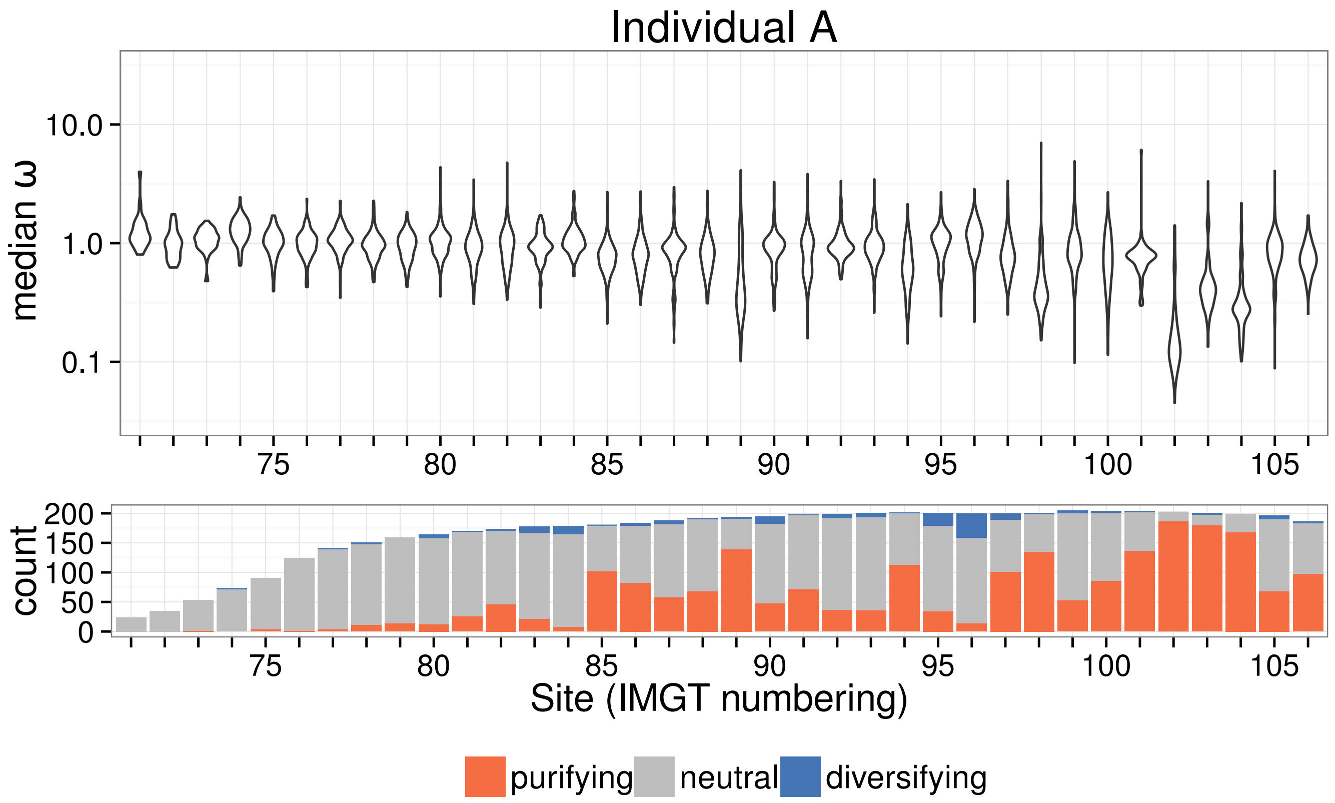
```
spark Cluster Network last hour
In    Now: 28.2k   Min: 20.8k   Avg: 48.0k   Max:440.5k
Out   Now: 30.9k   Min: 22.4k   Avg: 50.3k   Max:388.4k
```

Stacked Graph - load_one

Show Hosts Scaled: Auto | Same | None | spark **load_one** last **hour** sorted **by name** | Size small | Columns 4 (0 = metric + reports)

ip-172-31-33-5.us-west-2.compute.internal — ip-172-31-33-5
ip-172-31-33-80.us-west-2.compute.internal — ip-172-31-33-80
ip-172-31-39-80.us-west-2.compute.internal — ip-172-31-39-80
ip-172-31-42-28.us-west-2.compute.internal — ip-172-31-42-28
ip-172-31-42-73.us-west-2.compute.internal — ip-172-31-42-73
ip-172-31-45-83.us-west-2.compute.internal — ip-172-31-45-83
ip-172-31-47-173.us-west-2.compute.internal — ip-172-31-47-173

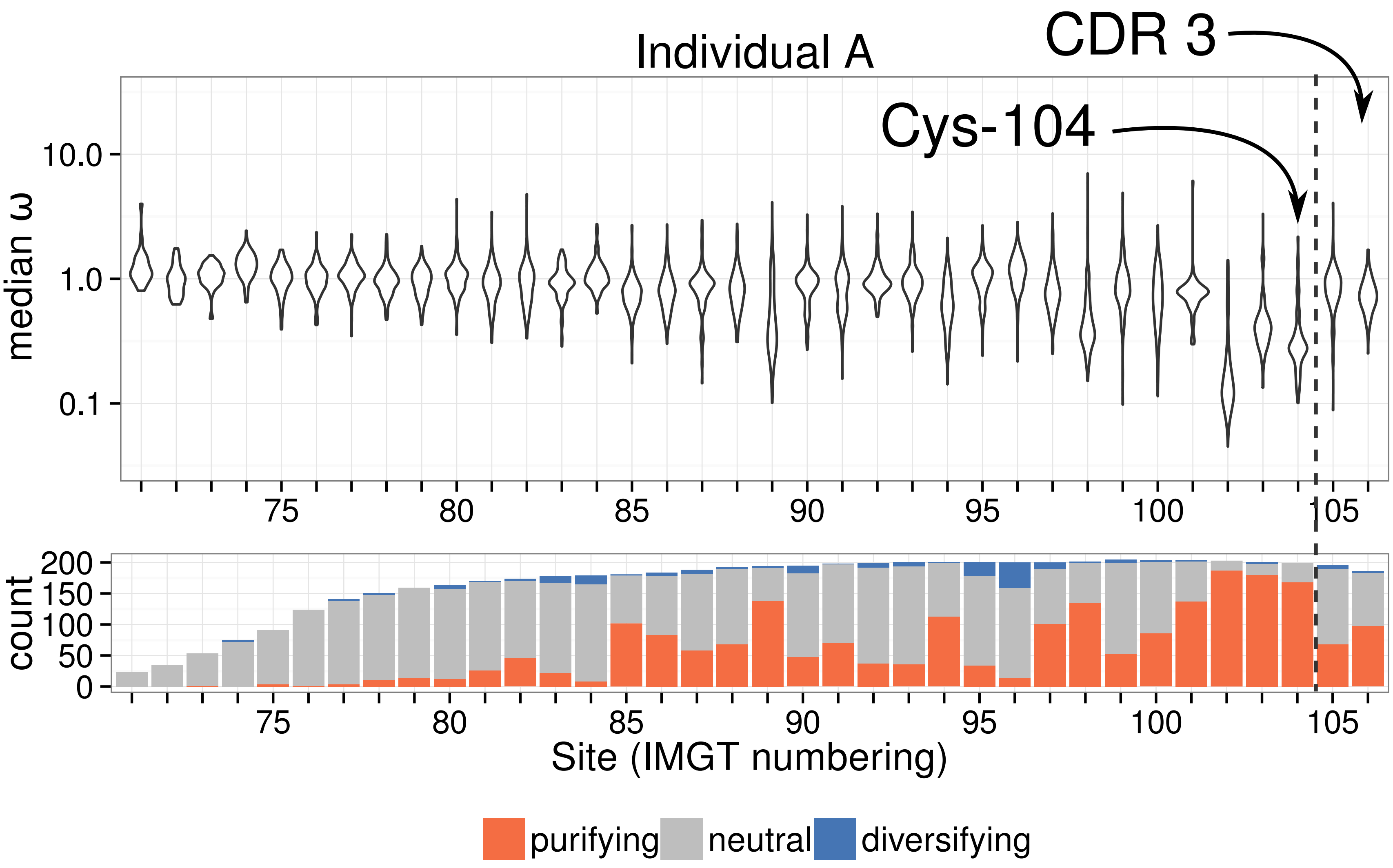
Overall IGHV selection map



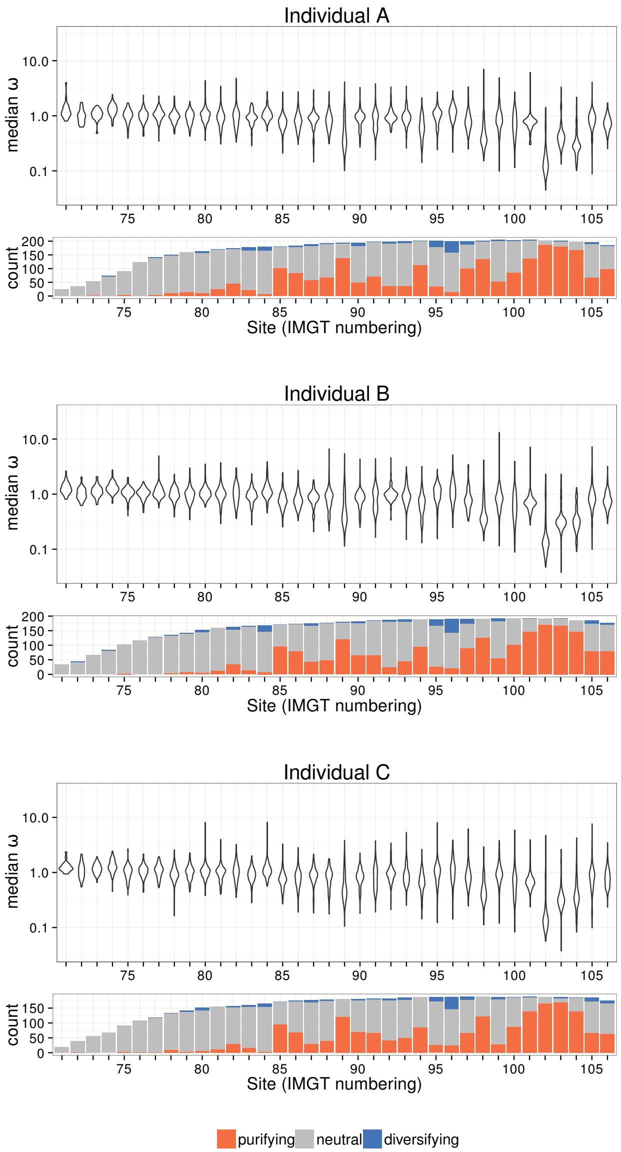
Distribution of median estimates of ω

Distribution of classifications across IGHV genes

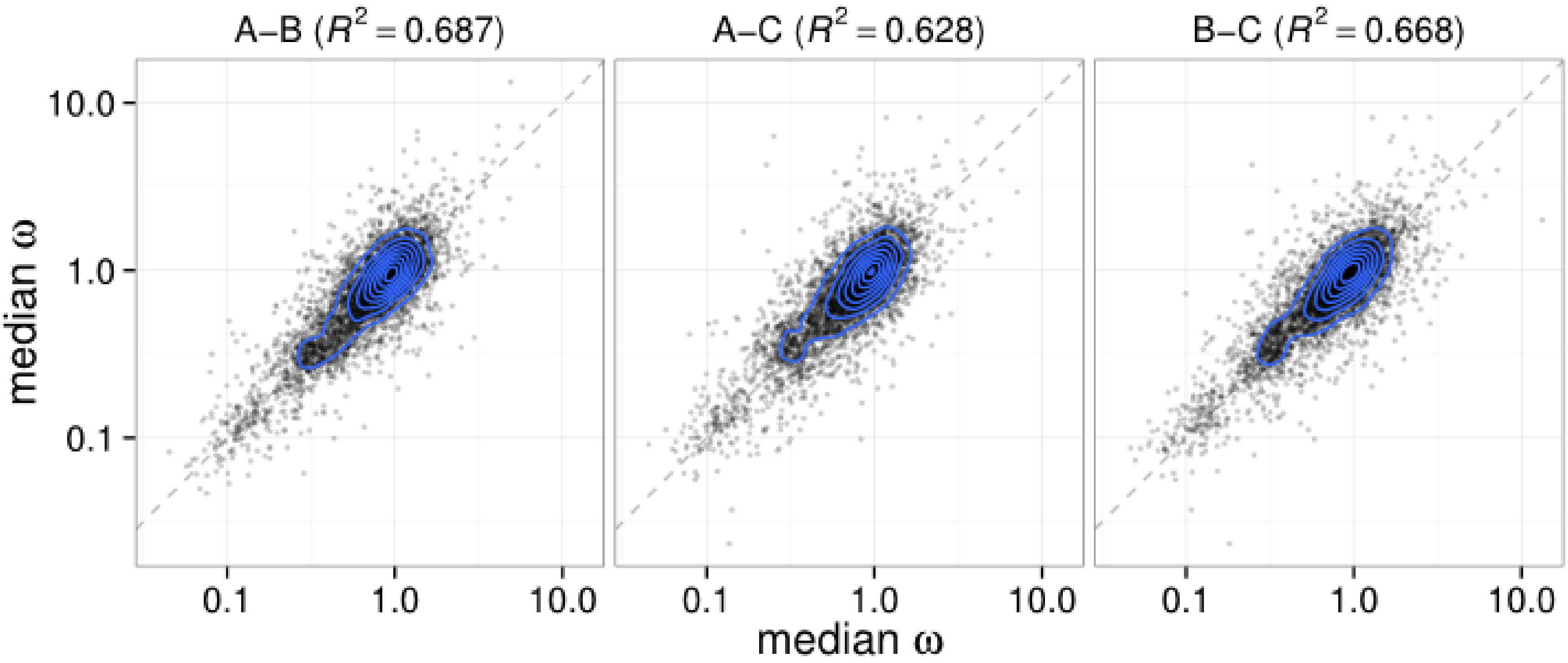
Purifying selection just before CDR3 loop



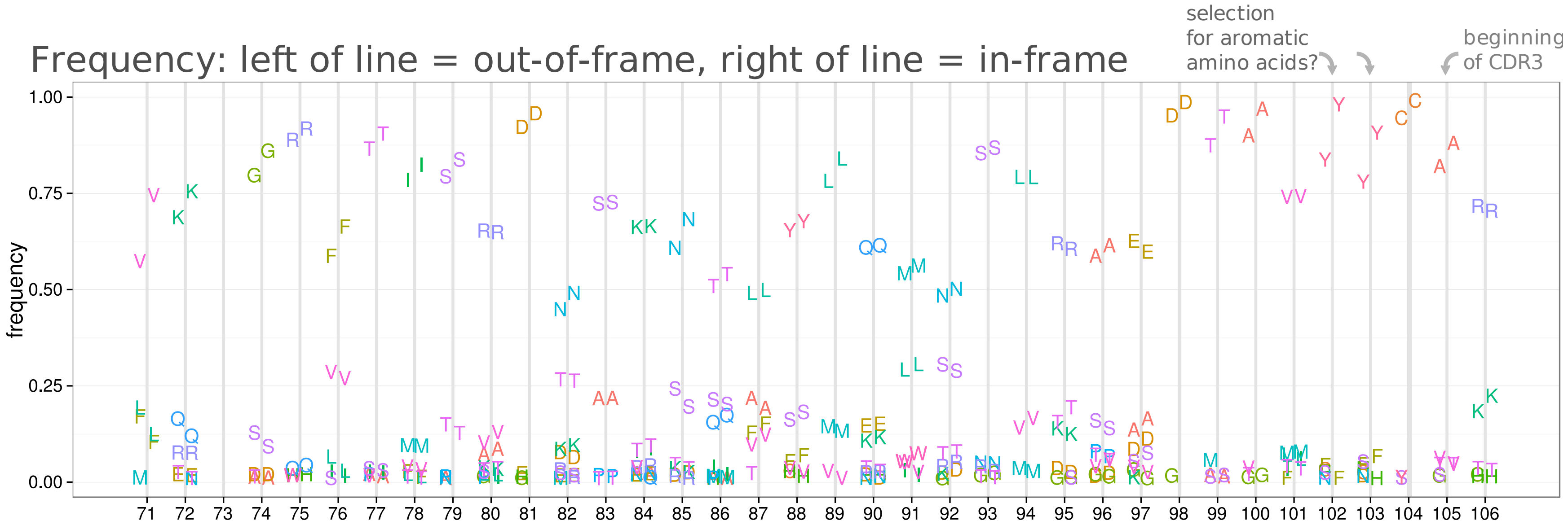
Similar across individuals



Similar across individuals (ii)



Distribution of amino acids



Frequency: left of line = out-of-frame, right of line = in-frame

selection for aromatic amino acids?

beginning of CDR3

Wrap-up of part 2

- We developed a selection inference procedure that can be used for millions of sequences with non-constant coverage
- We used this to derive a per-residue selection map
- We find that sites are generally under purifying selection
- We find especially strong selection near the beginning of the CDR3 corresponding to a preference for aromatic amino acids

For more details, paper is up on arXiv.

Discuss on http://phylobabble.org/

Thank you

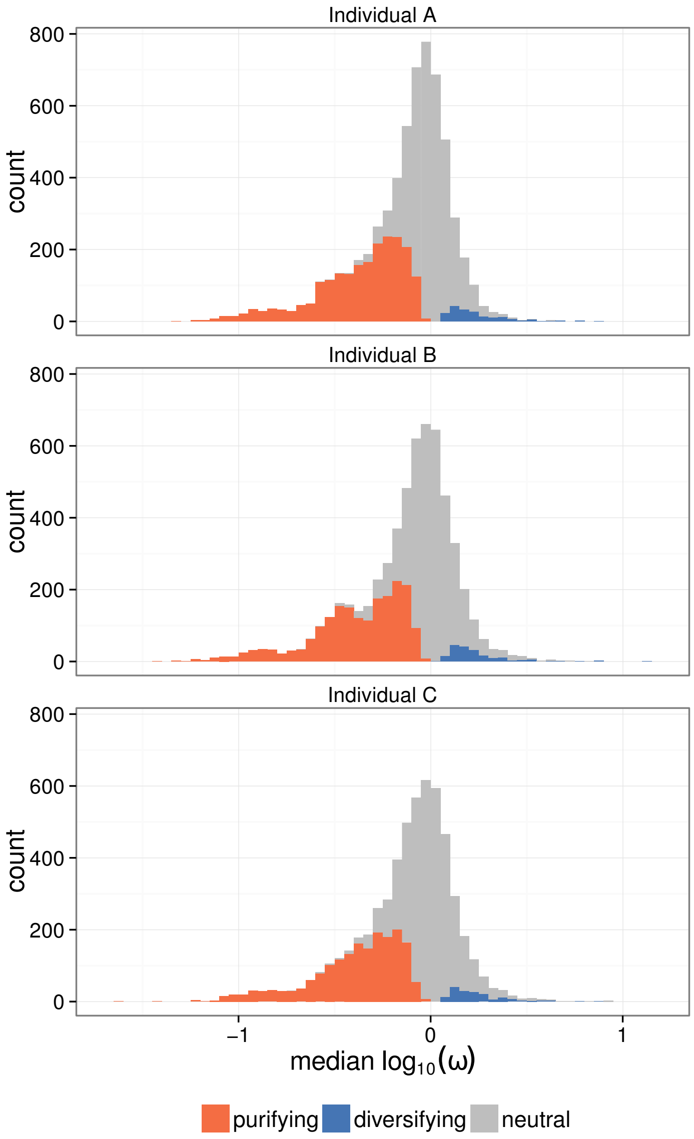
- Connor McCoy, Trevor Bedford, Vladimir Minin, Harlan Robins.
- Molecular work done by Paul Lindau in Phil Greenberg's lab.

- W. M. Keck Foundation
- University of Washington Center for AIDS Research (CFAR)
- University of Washington eScience Institute
- National Science Foundation and National Institute of Health

We have a postdoc opening to work on molecular evolution methods for HIV vaccine experimental design, and probably another for B cell work.

Addenda

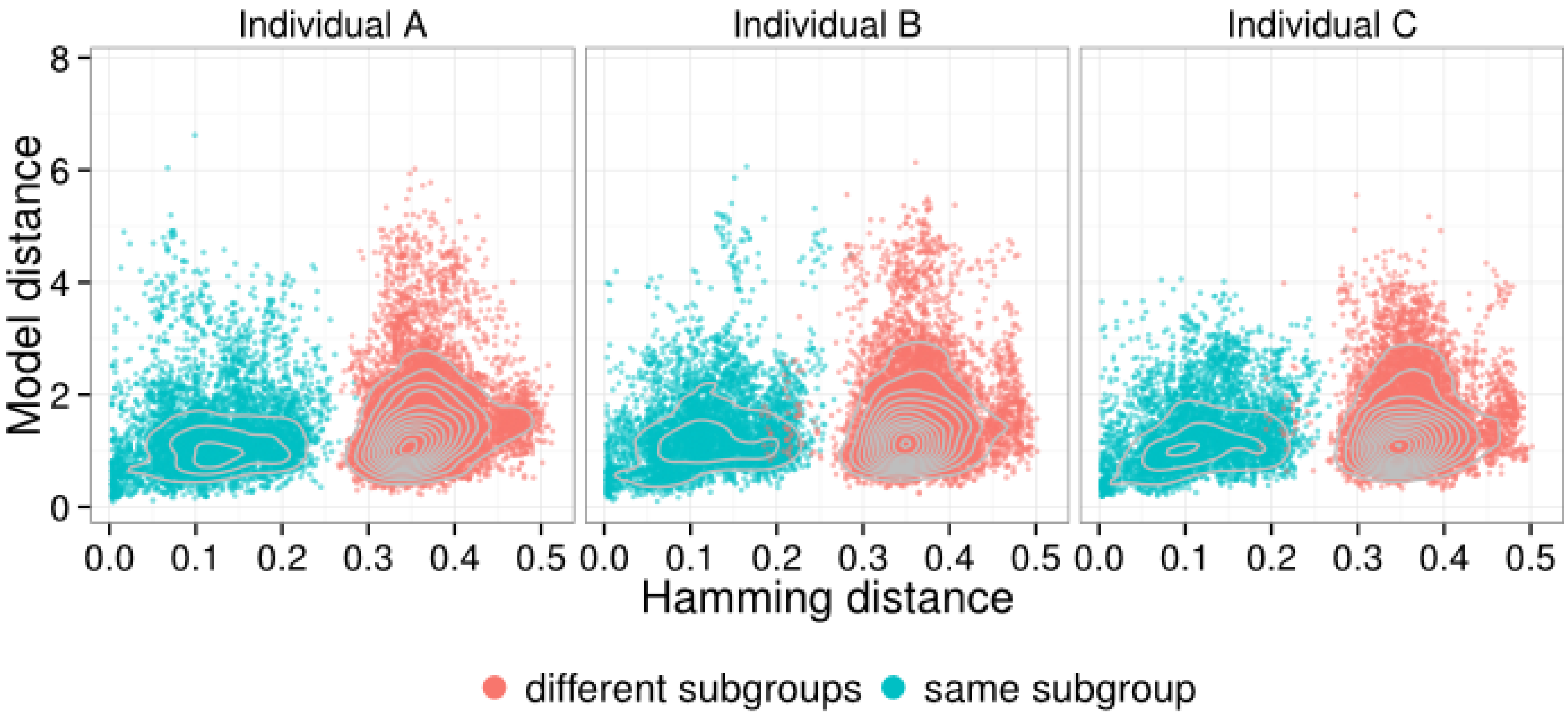
Sites are generally under purifying selection



Sequence counts

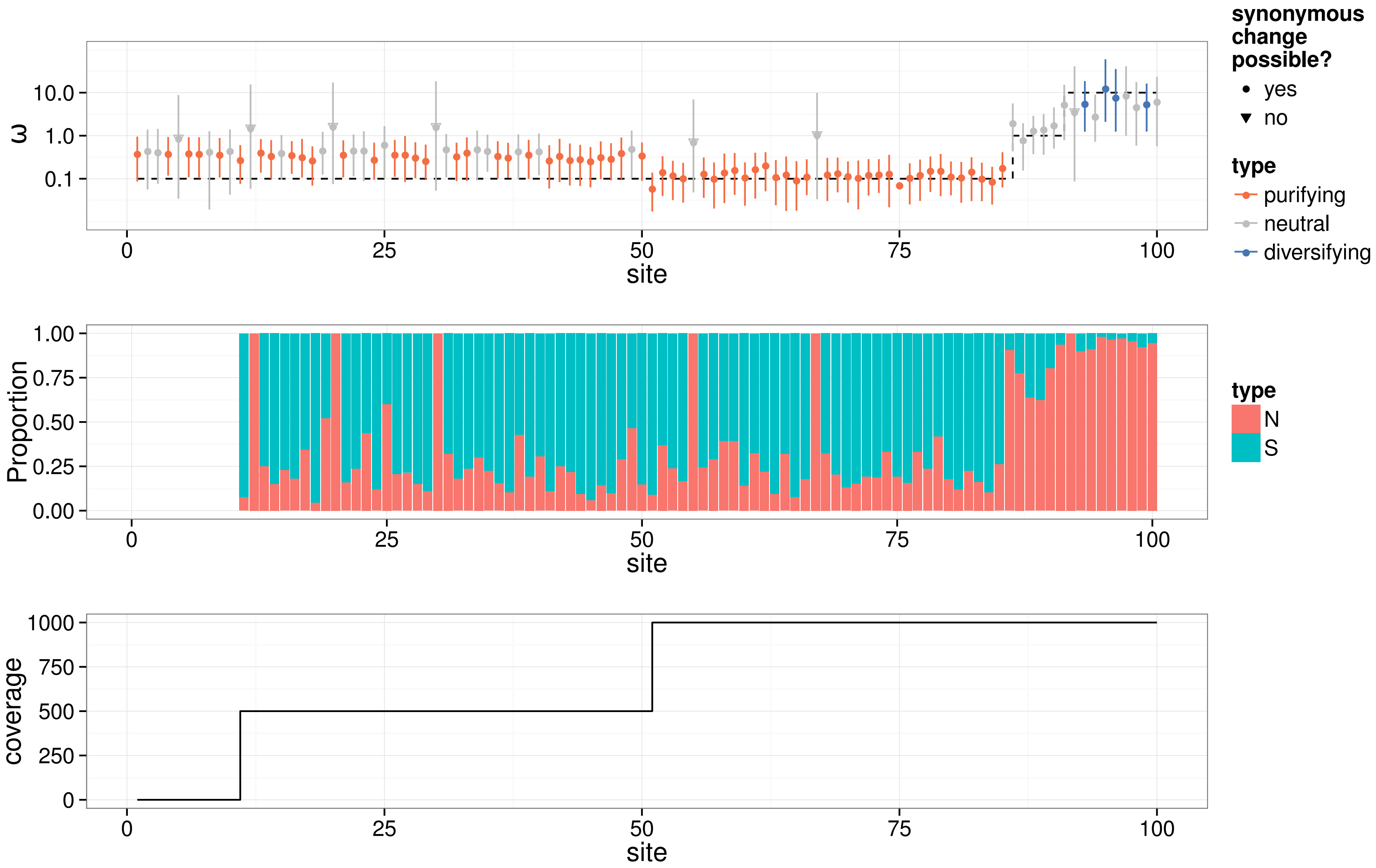
status	**A**	**B**	**C**
functional	4,139,983	4,861,800	3,748,306
out-of-frame	533,919	794,845	558,246
stop	104,525	169,423	112,901

Correlation between sequence and GTR matrix

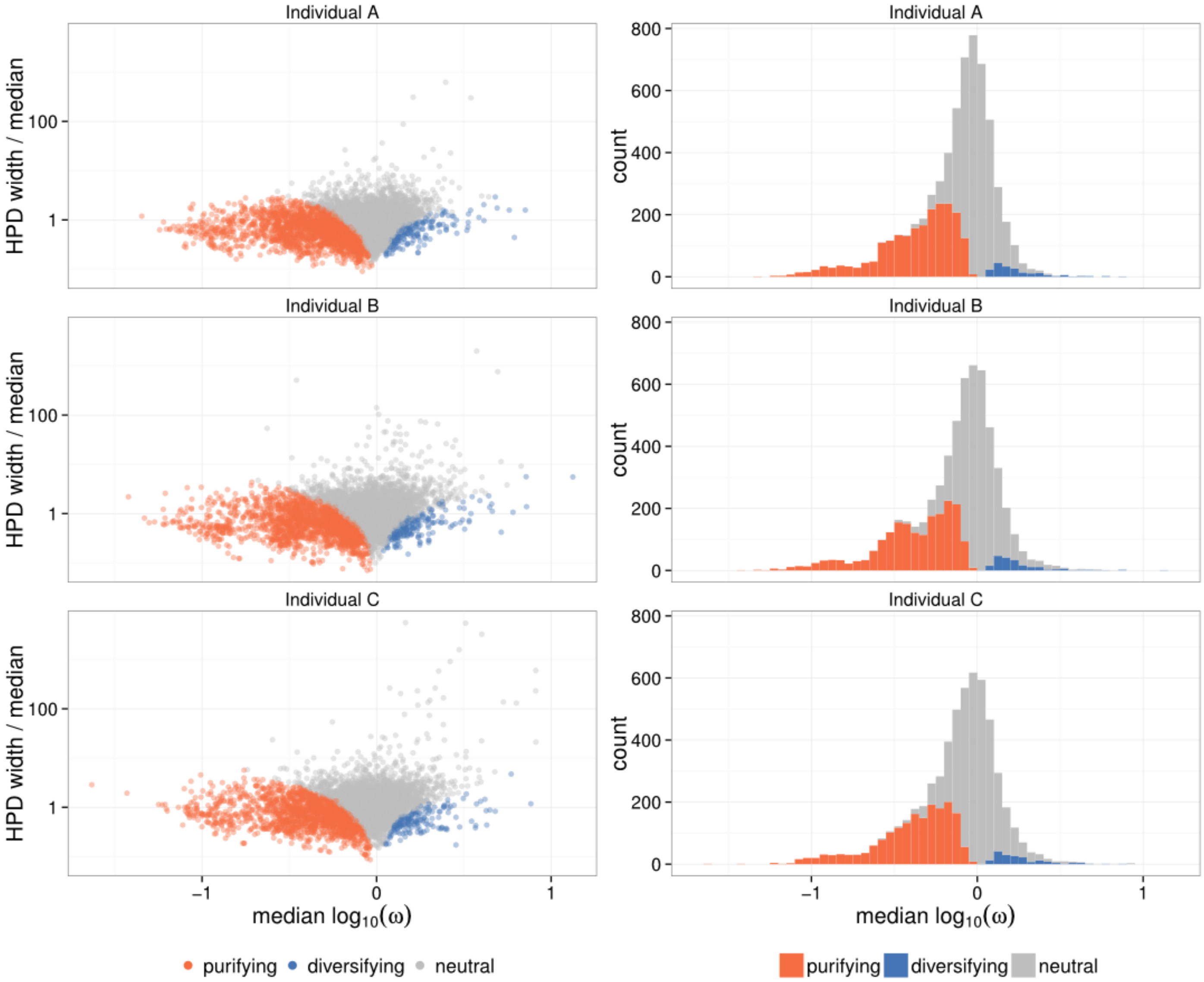


Each dot is a pair of genes.

Simulation results for selection inference



Omega distribution



Random facts

- Mean length of D segment in individual A’s naive repertoire is 16.61.
- Subject A’s naive sequences were 37% CDR3
- Divergence between the various germ-line V genes:

```
> summary(dist.dna(allele_01, pairwise.deletion=TRUE, model='raw'))
Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
0.003846 0.201300 0.344600 0.304700 0.384900 0.539500
```