Scalable likelihood-based methods to infer lineages and estimate selection in B cell repertoires

Frederick “Erick” Matsen
Fred Hutchinson Cancer Research Center
http://matsen.fredhutch.org/
@ematsen

with Trevor Bedford (FH), Vladimir Minin (UW),
***Duncan Ralph** (FH) and David Shaw (FH)*

Philosophy of talk

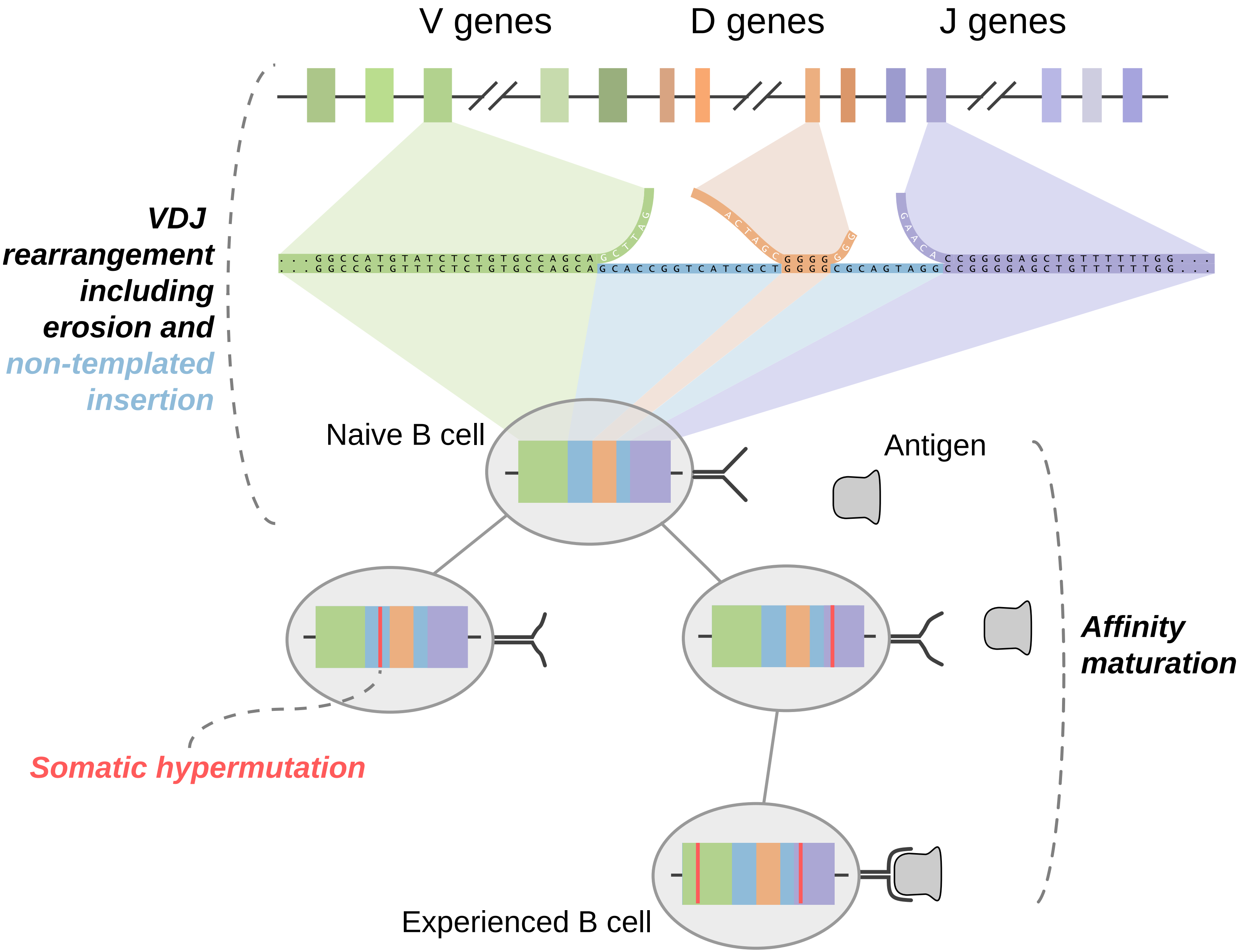
- Model immune cells (in this case B cells) probabilistically
- Infer parameters describing process via likelihoods
- Use these parameters to improve sequence analysis.

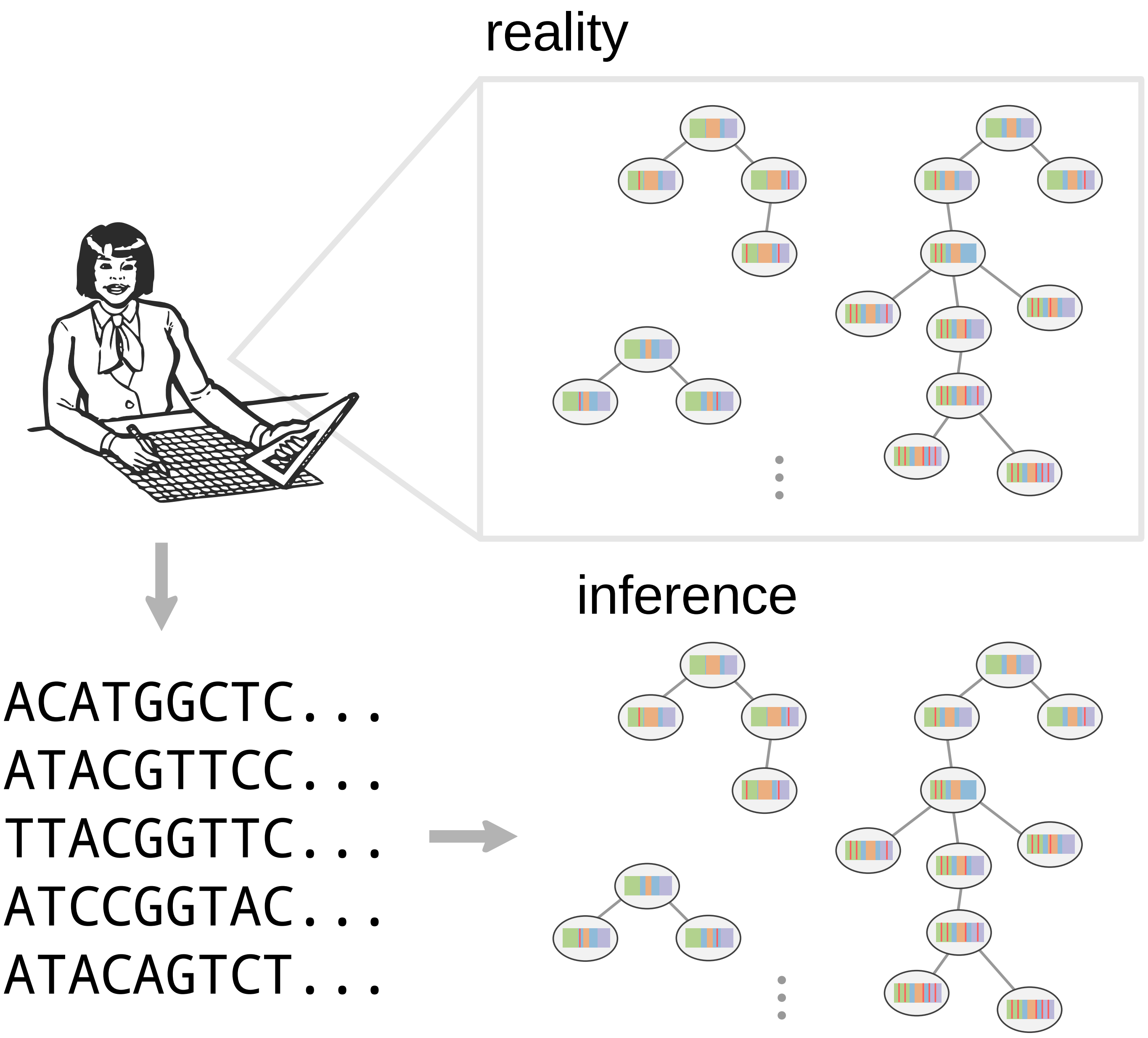
Work in progress.

Statistical phylogenetics

- Develop probabilistic model for sequence evolution
- Write down likelihood function
- Search for the maximum likelihood tree, including optimization heuristics
- Or integrate over trees using MCMC.

Casting phylogenetics as a statistical inference problem provides a solid foundation.


V genes
D genes
J genes
VDJ rearrangement including erosion and non-templated insertion
...GGCCATGTATCTCTGTGCCAGCA
GCTTAG
ACTAGC
GGGG
GG
GAACA
CCGGGGAGCTGTTTTTTGG...
...GGCCGTGTTTCTCTGTGCCAGCAGCACCGGTCATCGCTGGGGCGCAGTAGGCCGGGGAGCTGTTTTTTGG...
Naive B cell
Antigen
Affinity maturation
Somatic hypermutation
Experienced B cell



reality

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

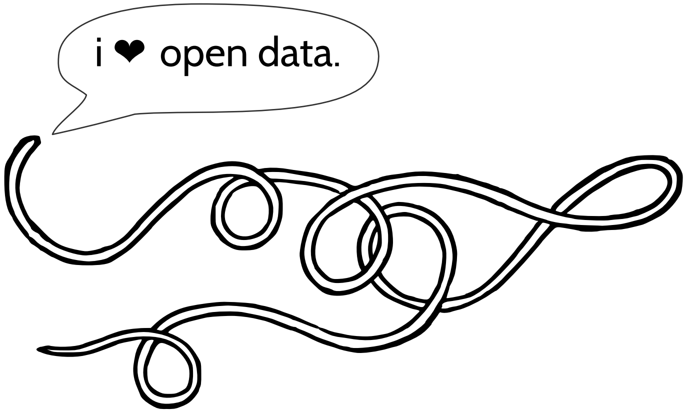
inference

To-do list

0. [Generate high-quality data. Hard!]
1. Annotate BCR sequences
2. Find clonal families
3. Reconstruct BCR phylogenetic trees
4. Infer BCR ancestral sequences
5. Evolutionary selection inference for BCRs

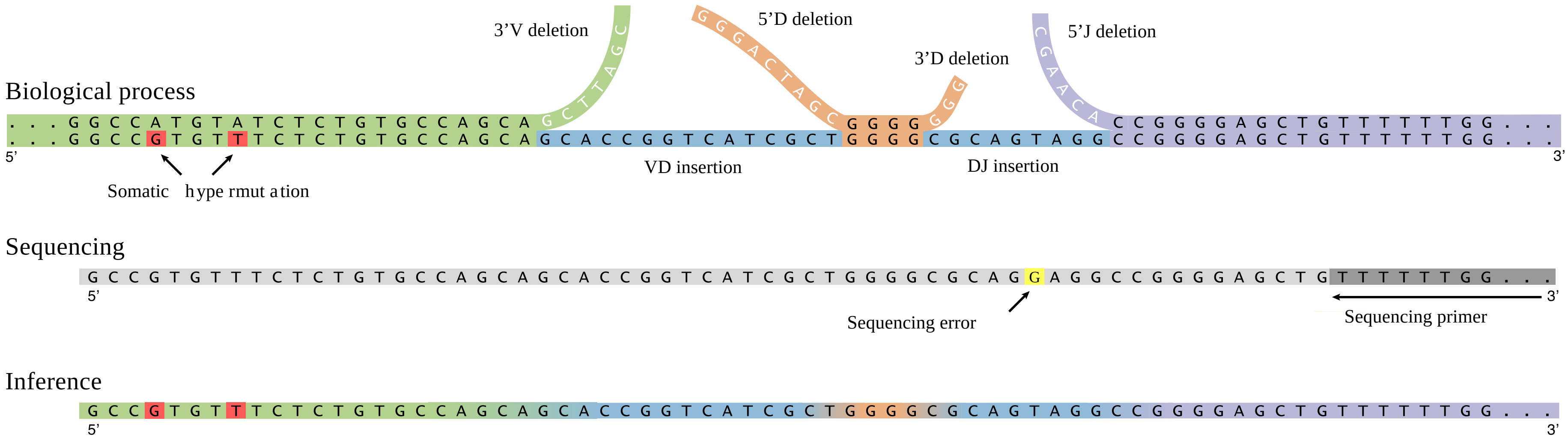
… in a probabilistic framework.

0. Gather data



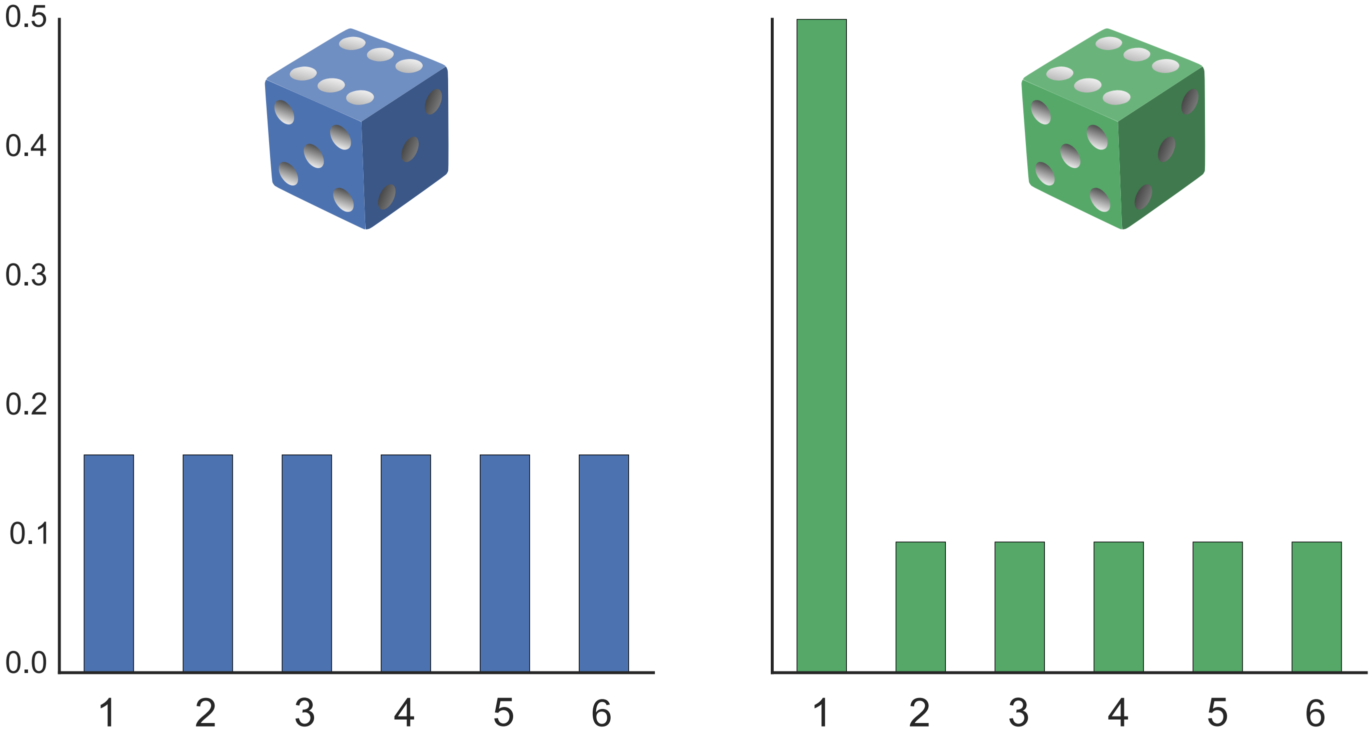
- Data from Adaptive Biotechnologies: 3 healthy individuals, naive/memory sorted, replicate immunosequencing with 188 wells and ~50K cells/well http://adaptivebiotech.com/link/mat2015
- Stern, Yaari, Heiden … O'Connor (2014). B cells populating the multiple sclerosis brain mature in the draining cervical lymph nodes. *Science Translational Medicine*, 6(248), 248ra107.

1. *Annotate BCR sequences*
from where did each nucleotide come?

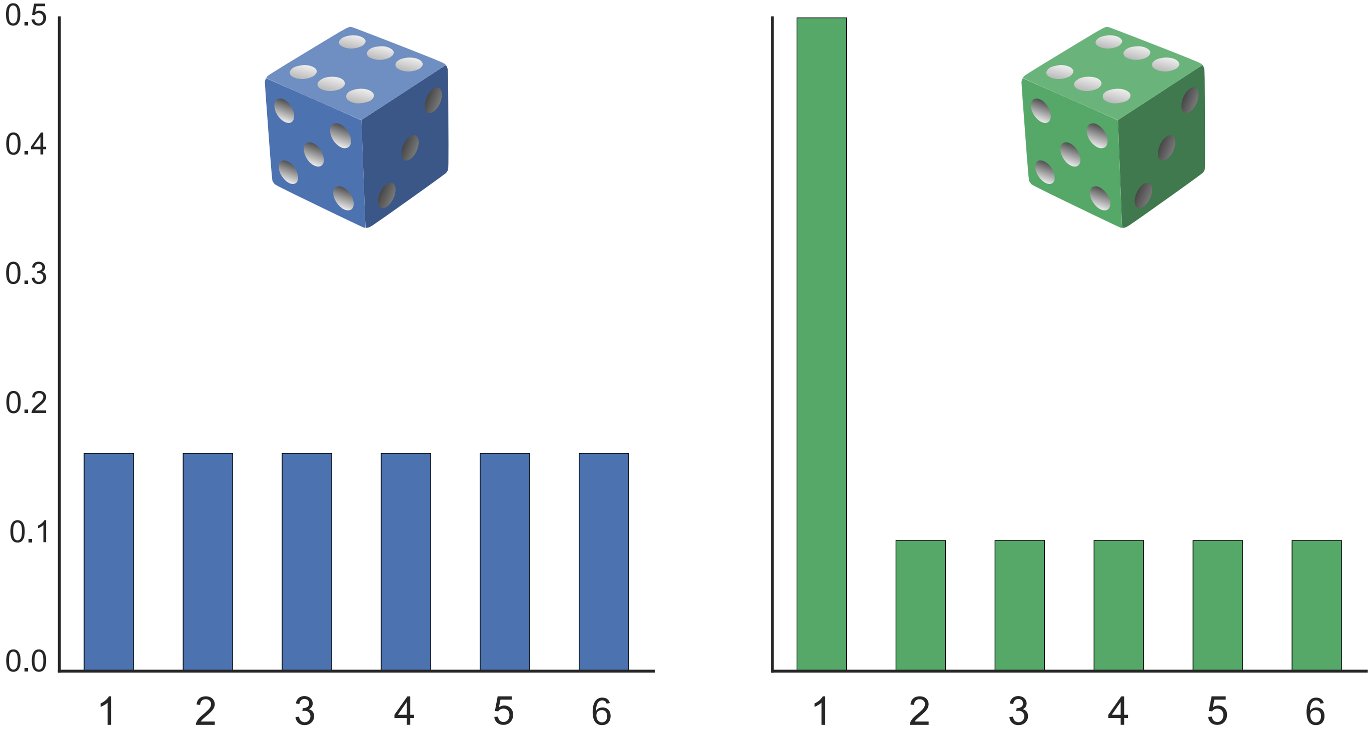

Biological process
3’V deletion
GCTTAGC
5’D deletion
GGGACTAGC
3’D deletion
GGG
5’J deletion
CGAACA
. . . G G C C A T G T A T C T C T G T G C C A G C A
G G G G
C C G G G G A G C T G T T T T T T G G . . .
. . . G G C C G T G T T T C T C T G T G C C A G C A G C A C C G G T C A T C G C T G G G G C G C A G T A G G C C G G G G A G C T G T T T T T T G G . . .
5’
3’
Somatic hype rmut ation
VD insertion
DJ insertion
Sequencing
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 G A G G C C G G G G A G C T G T T T T T T G G . . .
5’
3’
Sequencing error
Sequencing primer
Inference
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 . . .
5’
3’

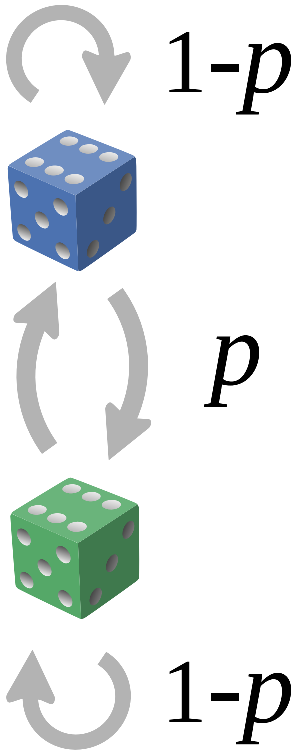
This is a key first step in BCR sequence analysis.

HMM intro: dishonest casino



HMM intro: dishonest casino

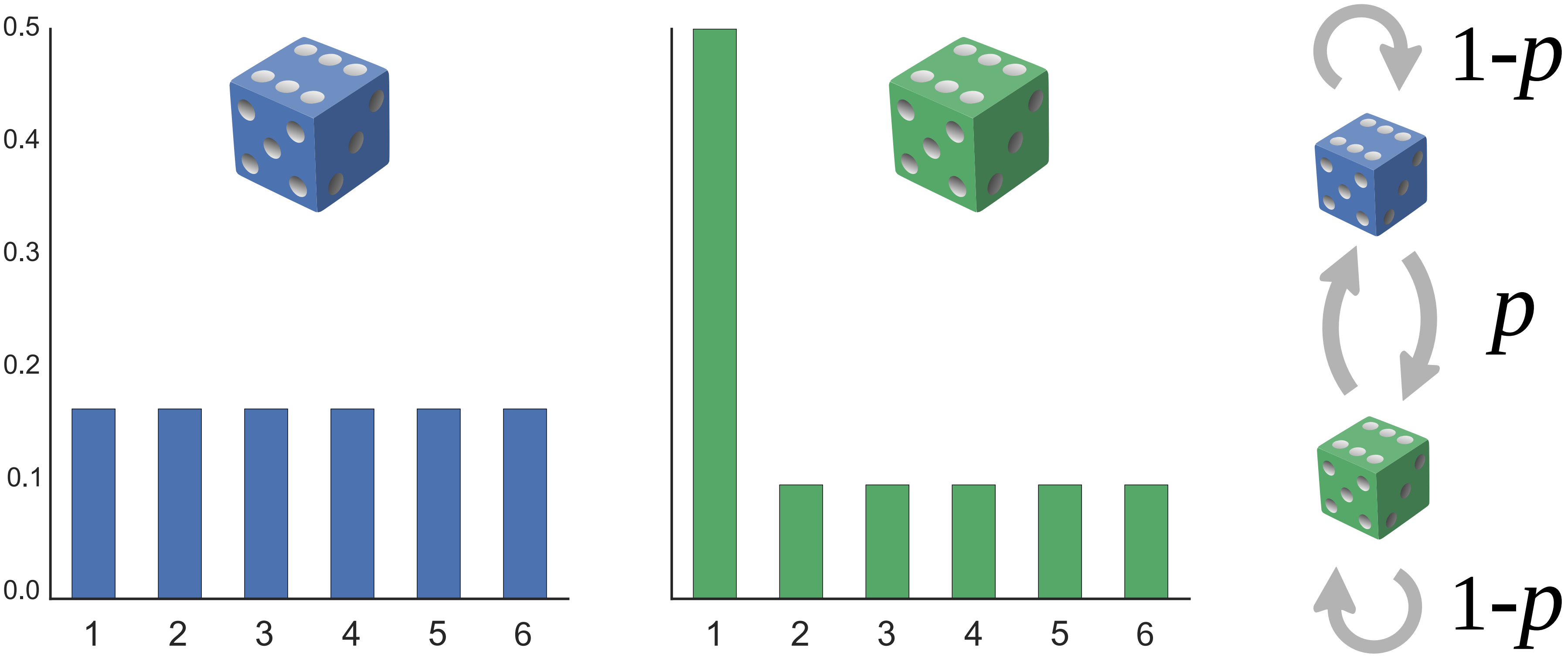


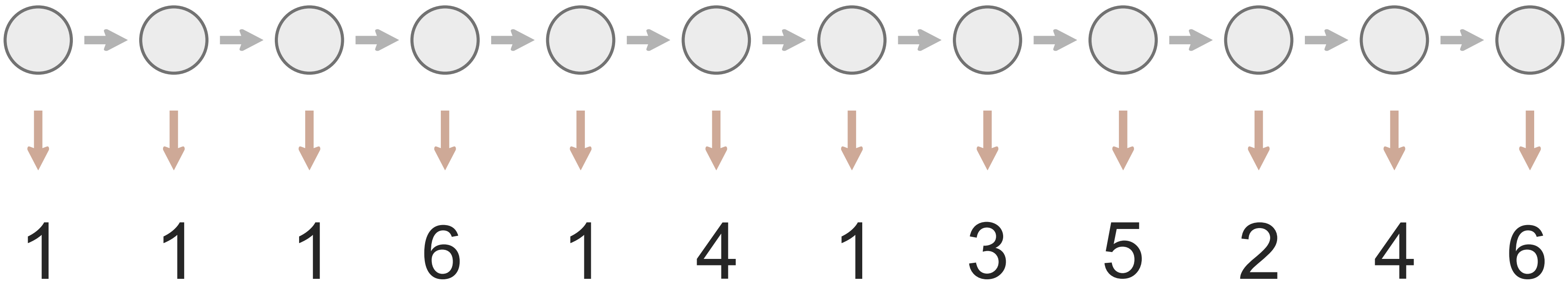

1-p

p

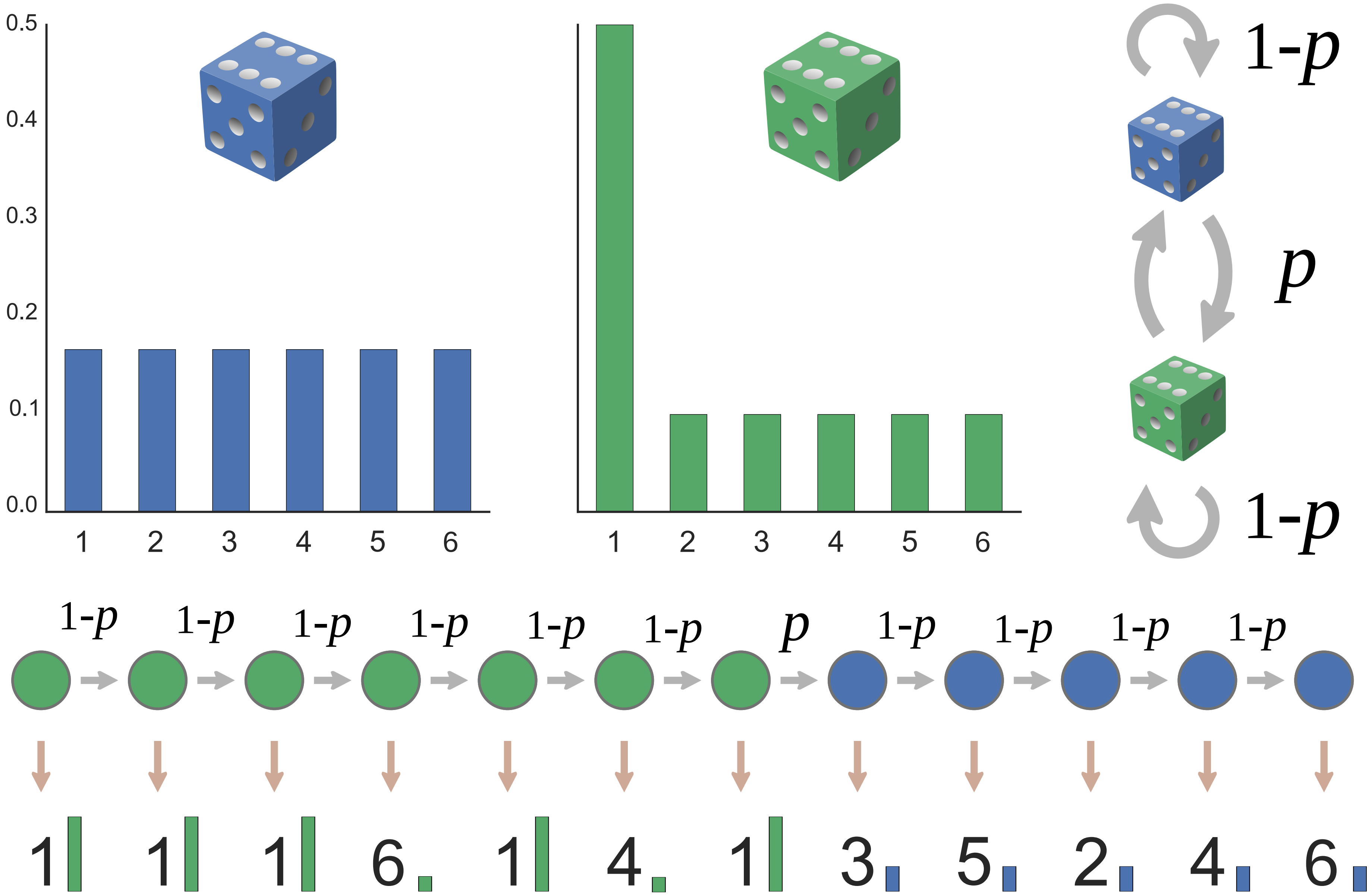
1-p

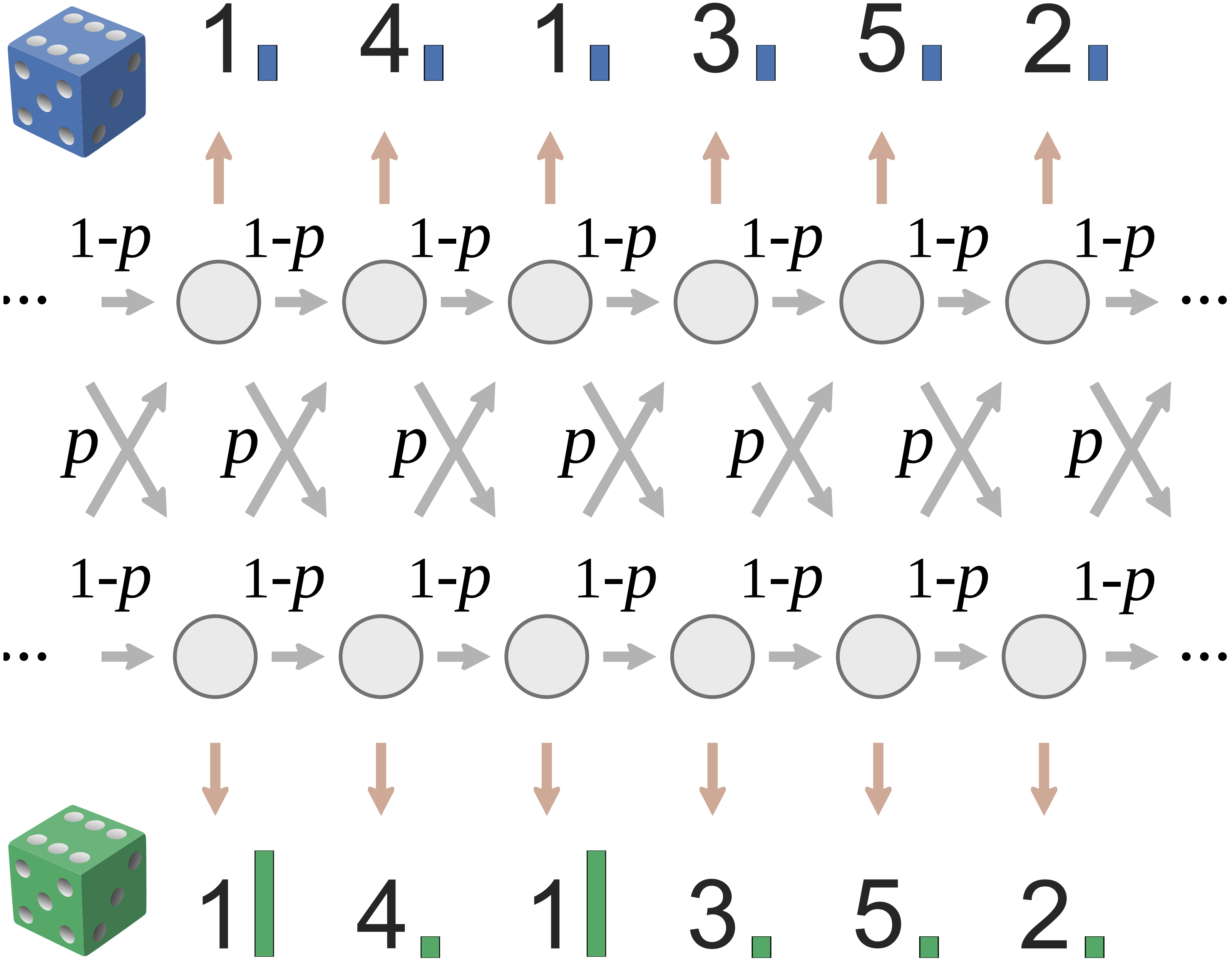
HMM intro: dishonest casino

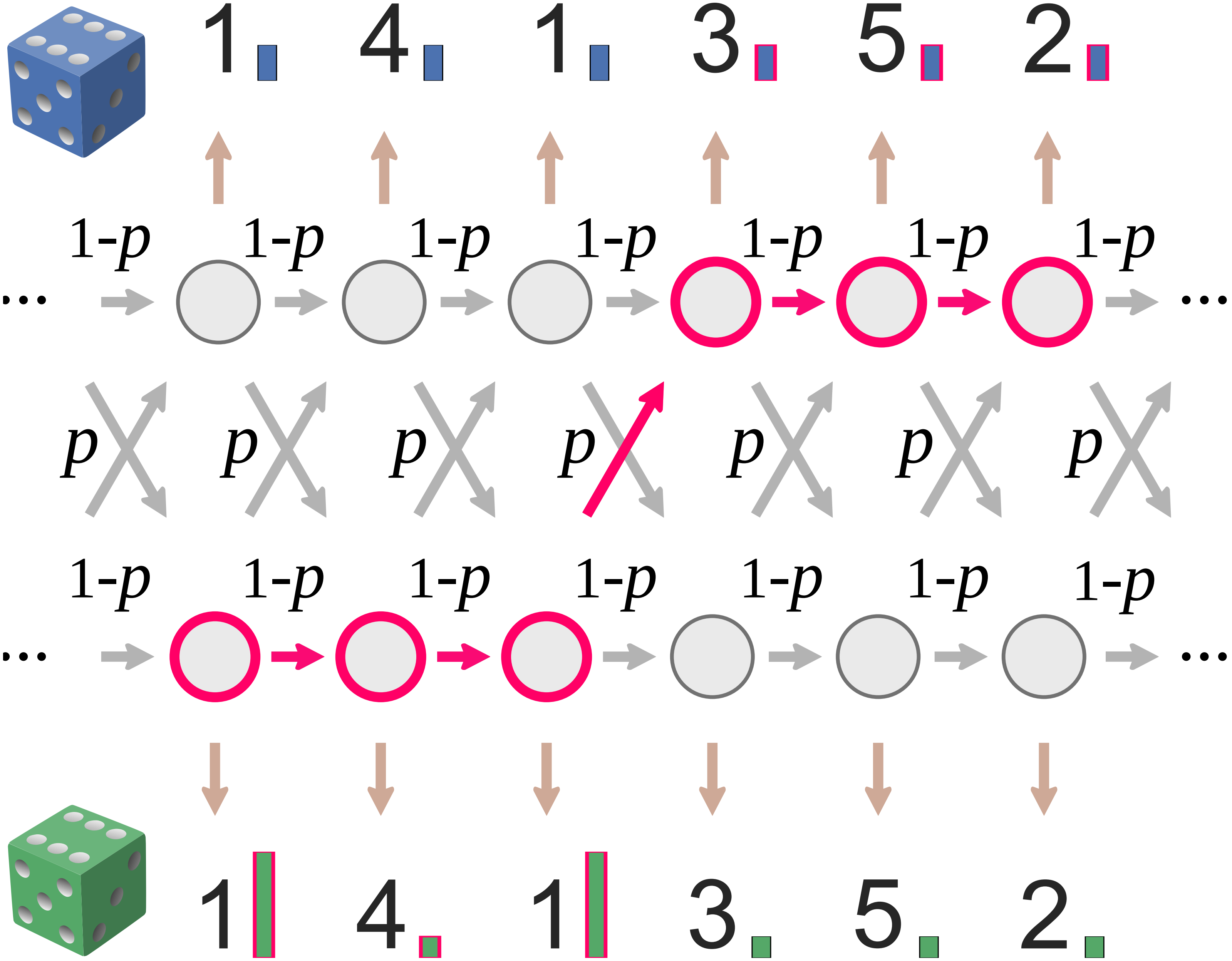




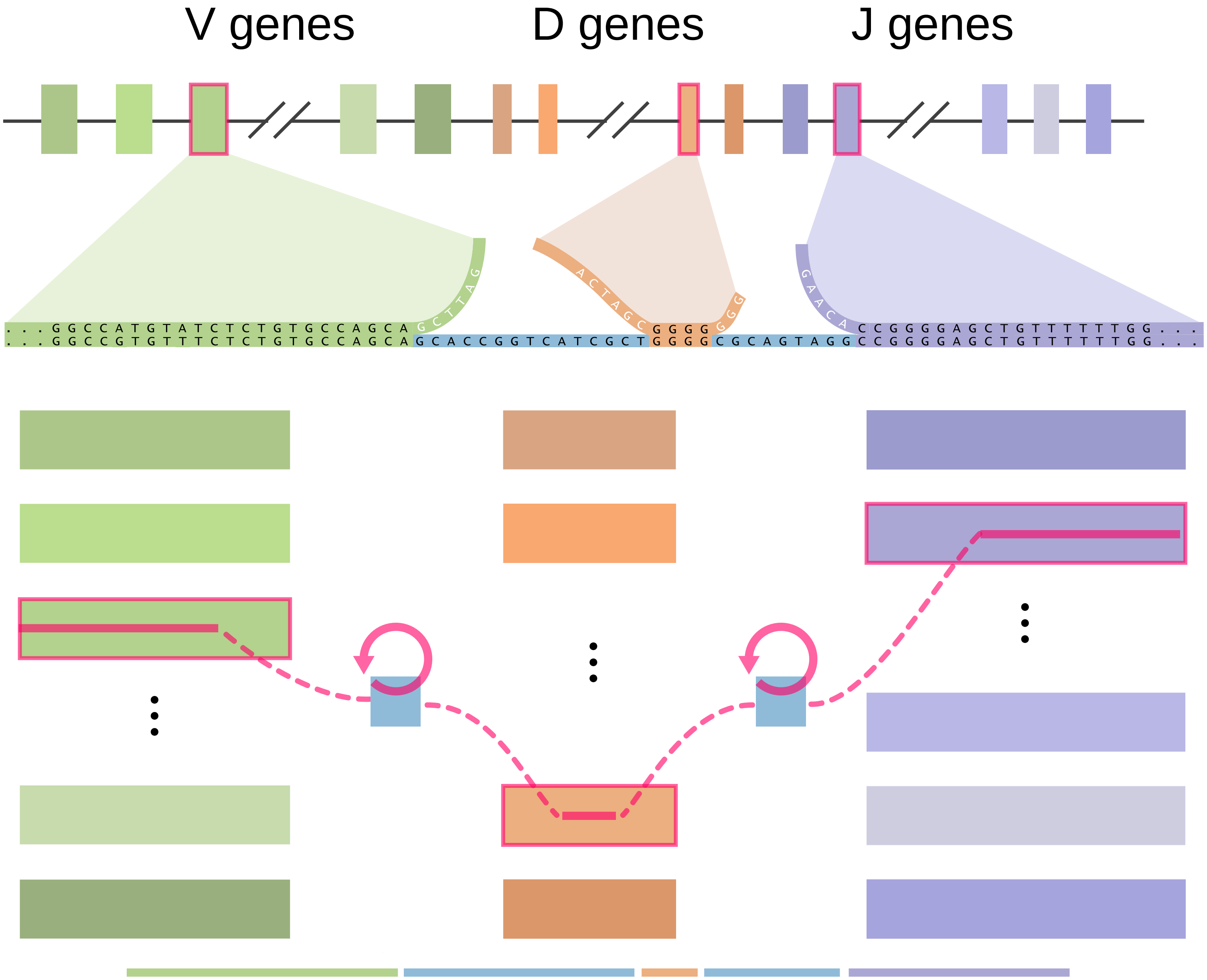
HMM intro: dishonest casino

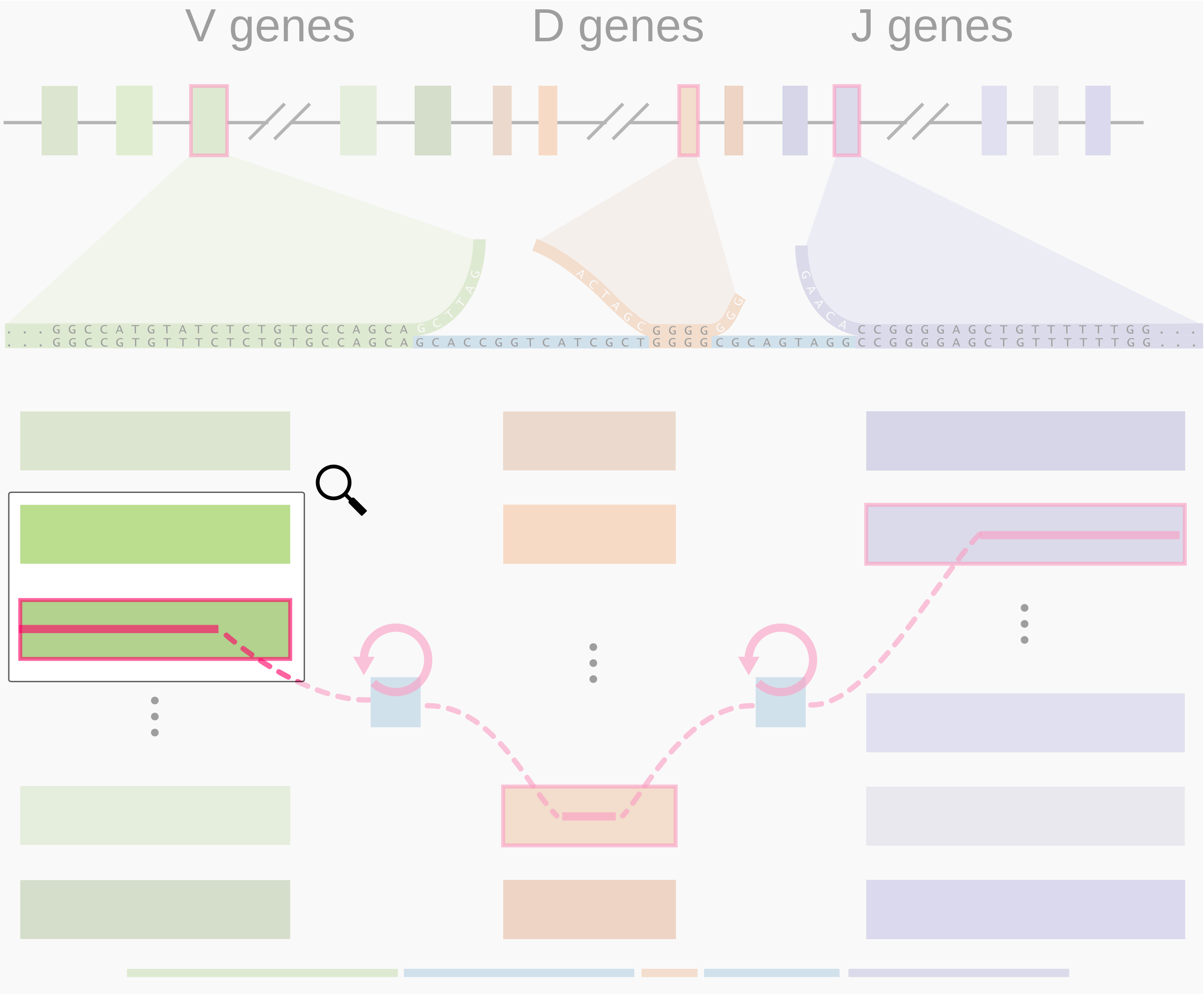


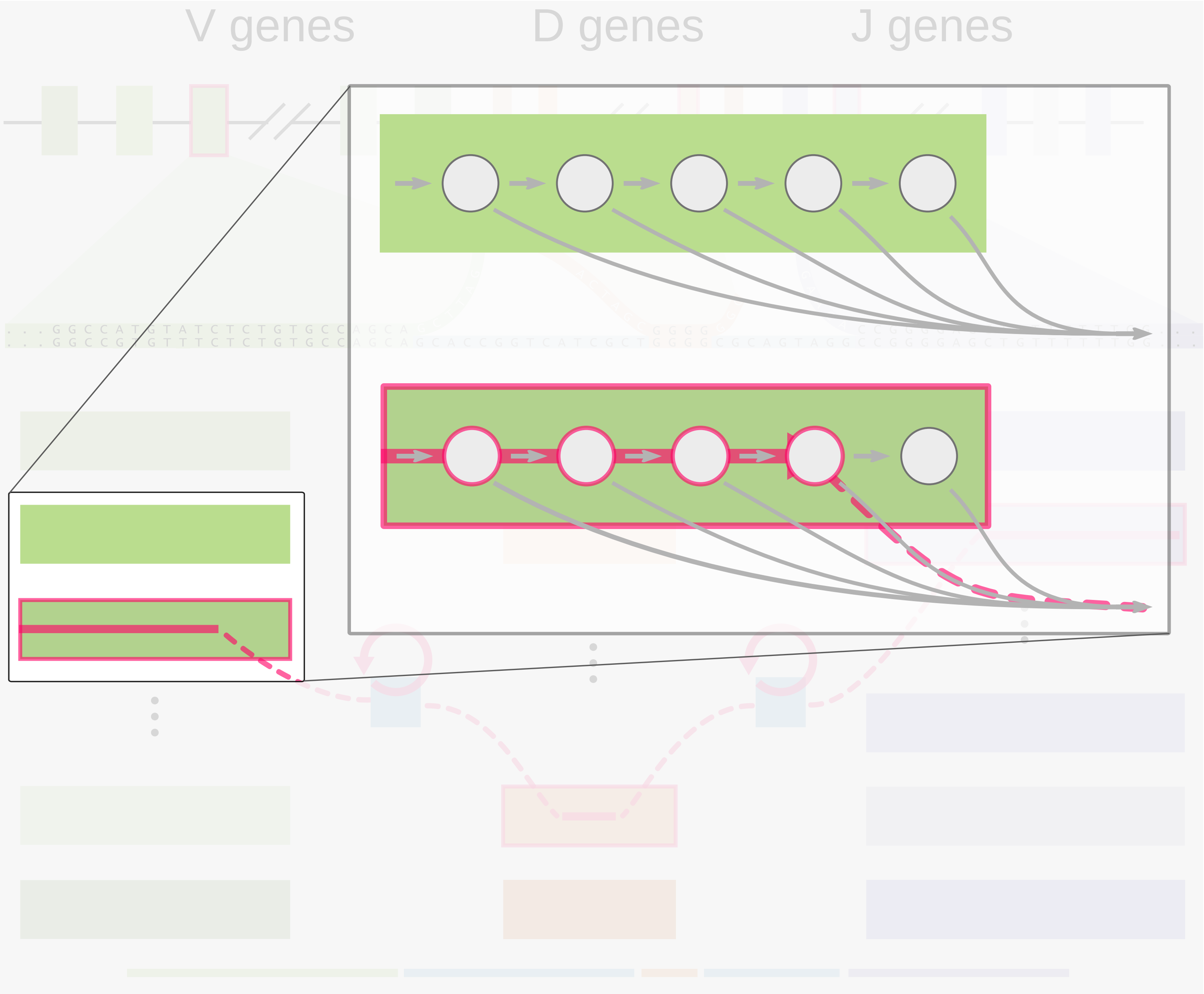

1
4
1
3
5
2
1-p
1-p
1-p
1-p
1-p
1-p
1-p
...
...
p
p
p
p
p
p
p
1-p
1-p
1-p
1-p
1-p
1-p
1-p
...
...
1
4
1
3
5
2


1
4
1
3
5
2
1-p
1-p
1-p
1-p
1-p
1-p
1-p
…
…
p
p
p
p
p
p
p
1-p
1-p
1-p
1-p
1-p
1-p
1-p
…
…
1
4
1
3
5
2

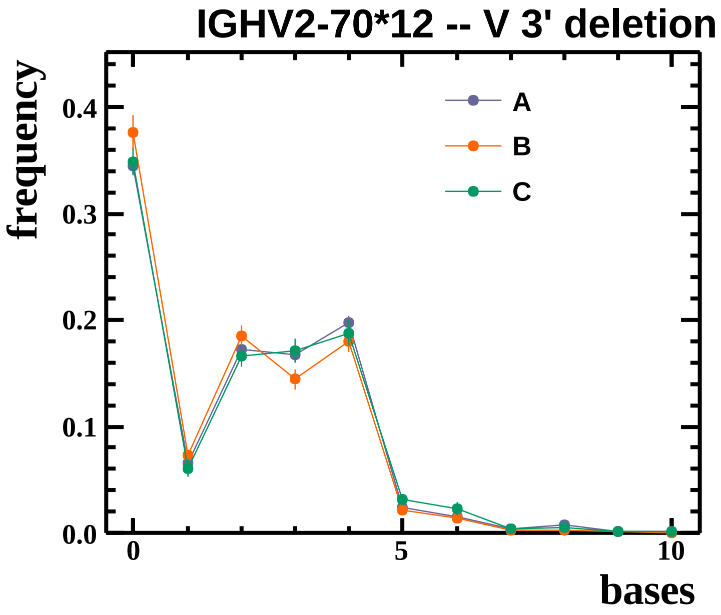
"Thread" reads onto structure

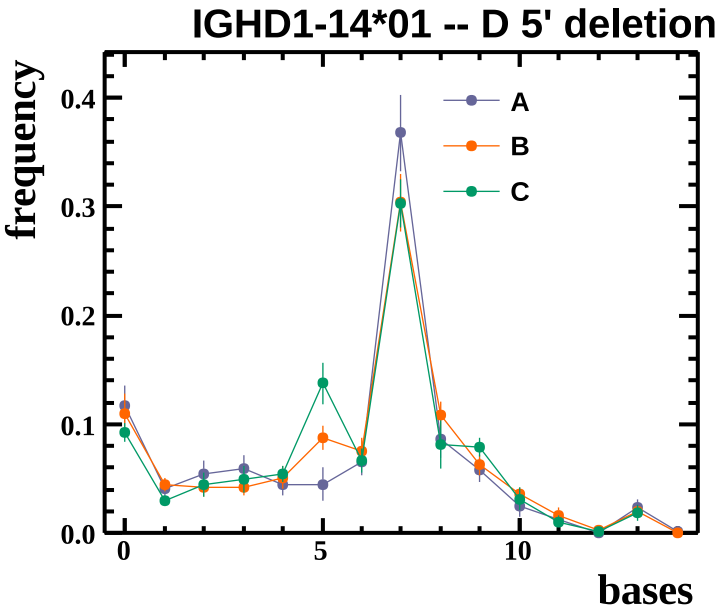

V genes
D genes
J genes
. . . G G C C A T G T A T C T C T G T G C C A G C A G C T T A G
A C T A G C G G G G G G G
G A A C A C C G G G G A G C T G T T T T T T G G . . .
. . . G G C C G T G T T T C T C T G T G C C A G C A G C A C C G G T C A T C G C T G G G G C G C A G T A G G C C G G G G A G C T G T T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

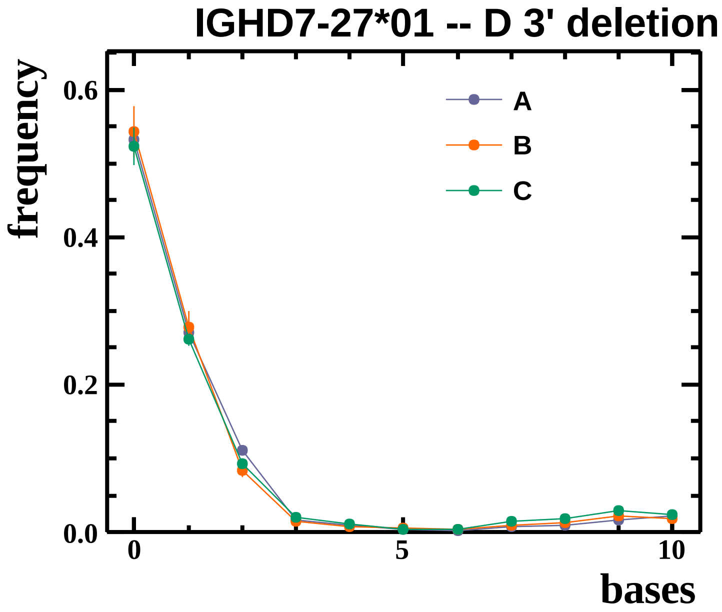

V genes
D genes
J genes
...GGCCATGTATCTCTGTGCCAGCA
GCTTAG
ACTAGC
GGGG
GGG
GAACA
CCGGGGAGCTGTTTTTTGG...
...GGCCGTGTTTCTCTGTGCCAGCAGCACCGGTCATCGCTGGGGCGCAGTAGGCCGGGGAGCTGTTTTTTGG...

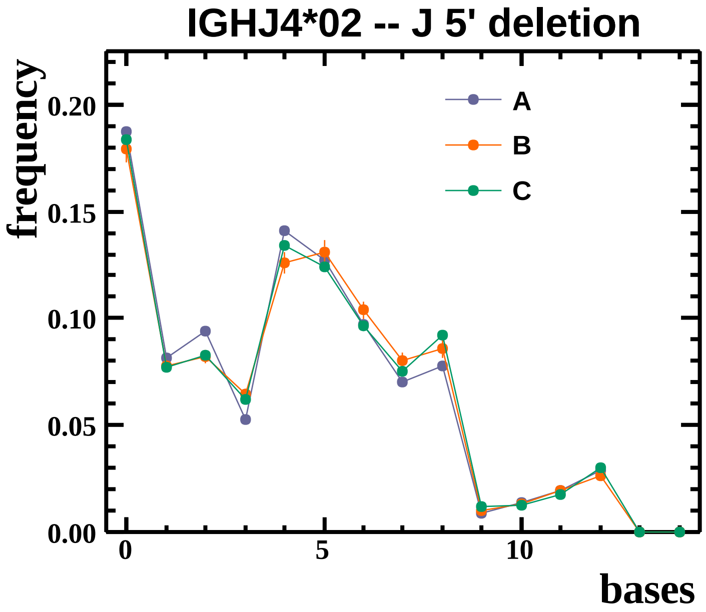

V genes
D genes
J genes
...GGCCATGTATCTCTGTGCCAGCA
...GGCCGTGTTTCTCTGTGCCAGCAGCACCGGTCATCGCTGGGGCGCAGTAGGCCGGGGAGCTGTTTTTGG...

Distributions are reproducibly weird!

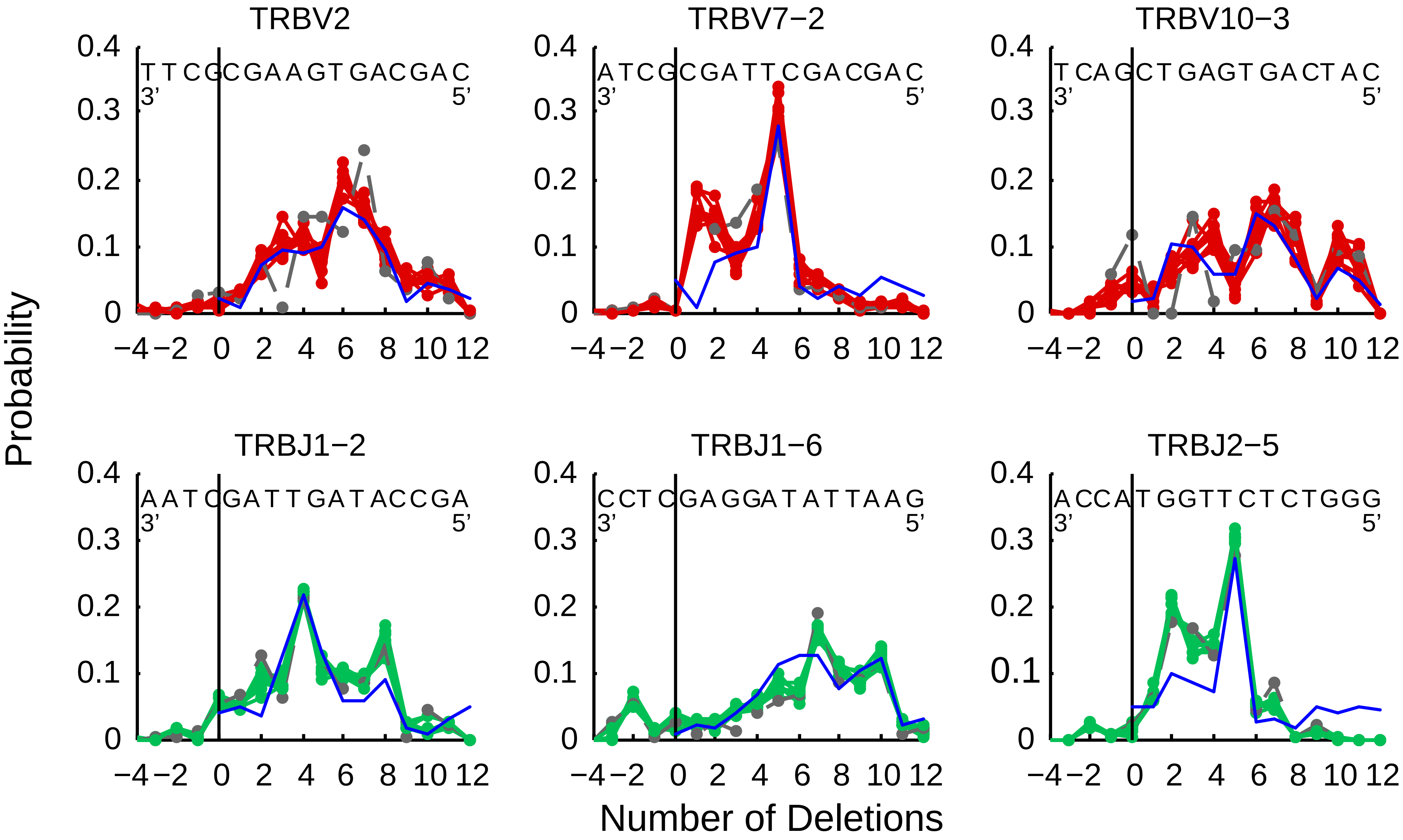

IGHV2-70*12 -- V 3' deletion


IGHD1-14*01 -- D 5' deletion

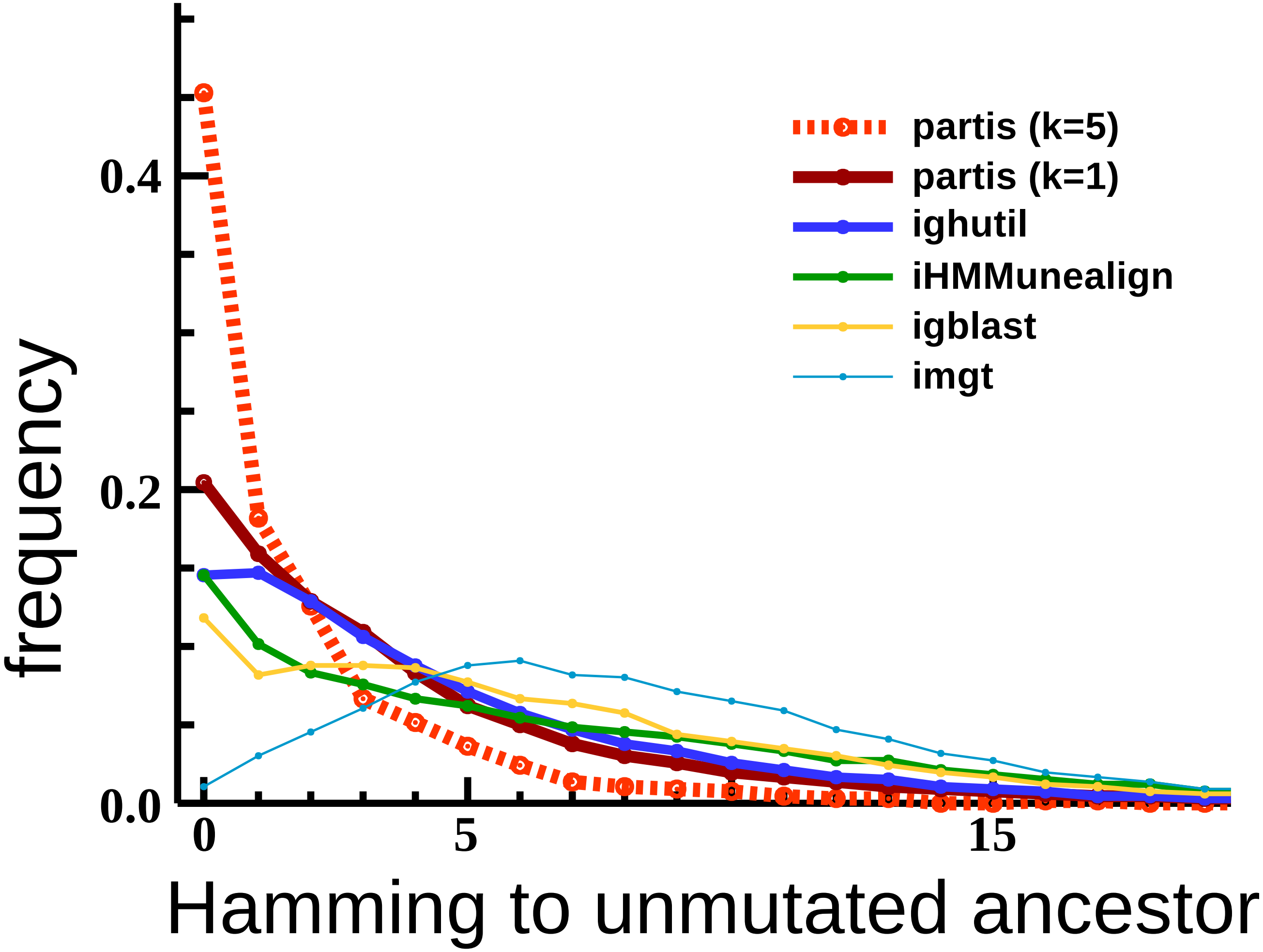

IGHD7-27*01 -- D 3' deletion


IGHJ4*02 -- J 5' deletion

Murugan, Mora, Walczak, Callan (2012)



Incorporating model complexity leads to better inferences



HMMs for BCR annotation

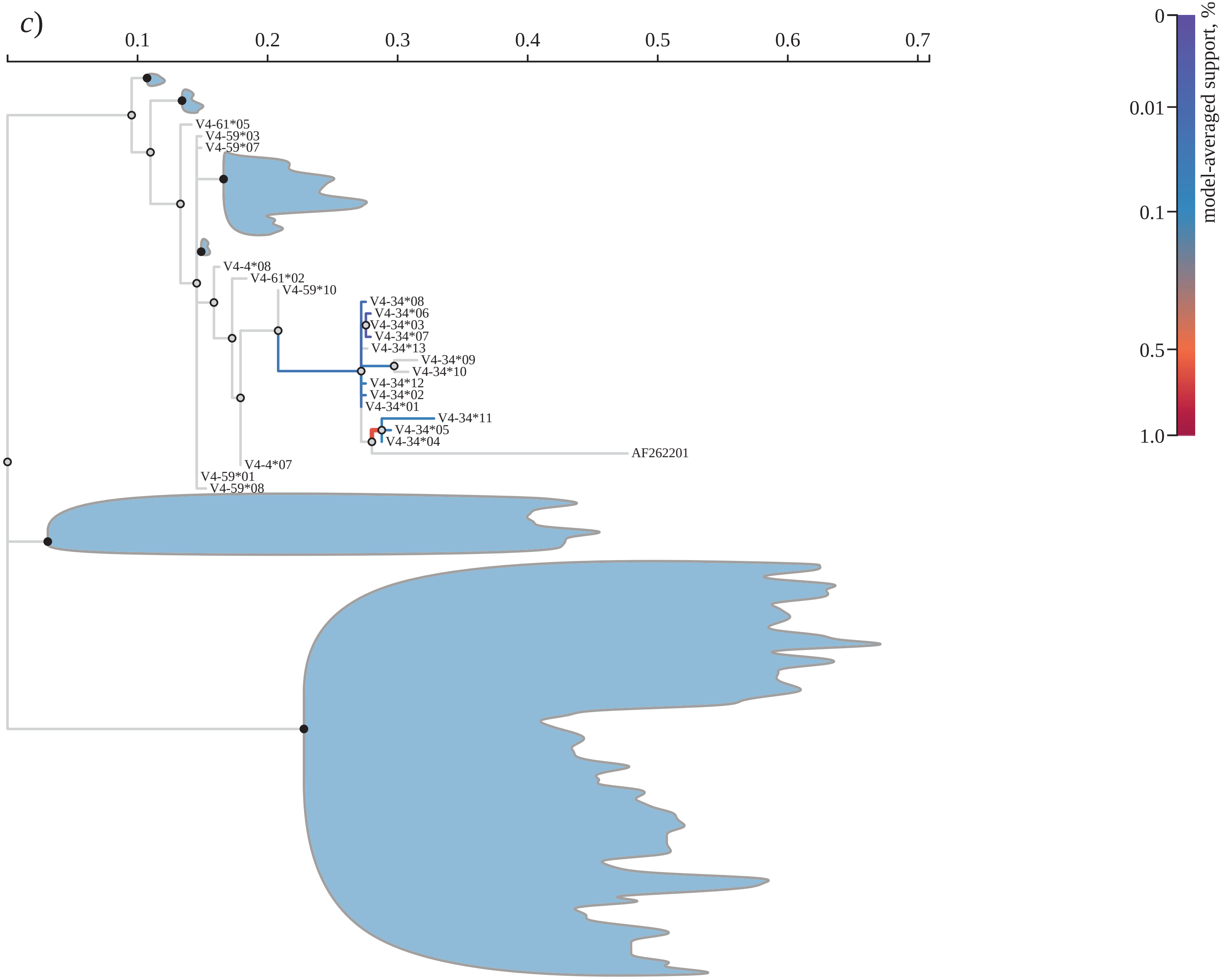
- SoDA: Volpe, Cowell, & Kepler (2005)
- iHMMune-align: Gaëta, Malming, … & Collins (2007)
- SoDA2: Munshaw & Kepler (2010)

These implementations use standard probability distributions for parameters (e.g. deletion lengths).

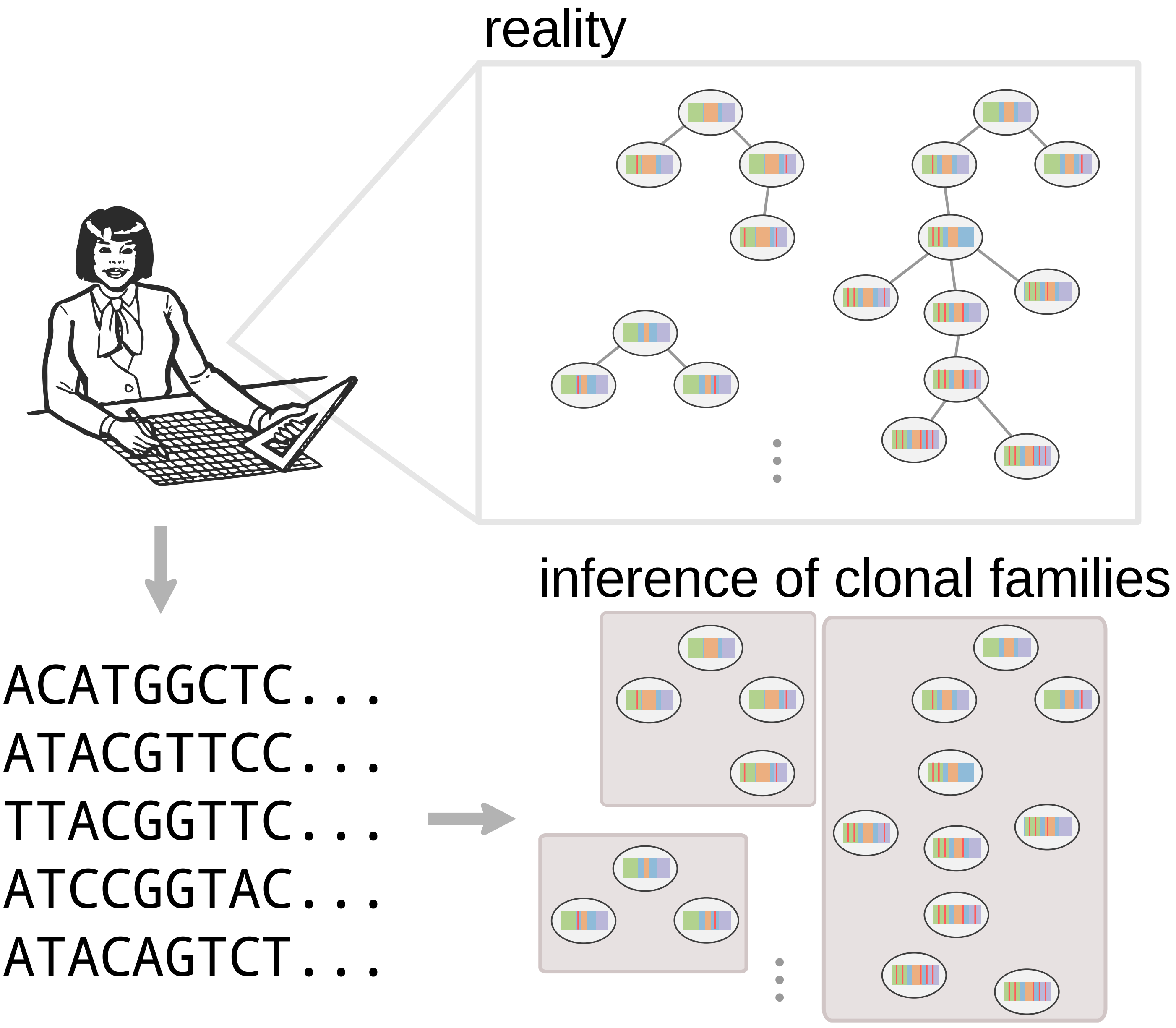
Fit parameter-rich HMMs that are able to capture underlying complexity of the process.

- partis: Ralph & M. (2016)
- repgenHMM: Elhanati, Marcou, Mora & Walczak (2016)

IgSCUEAL: Frost, Murrell, ... K. Pond (2015)



2. Find clonal families



reality

inference of clonal families

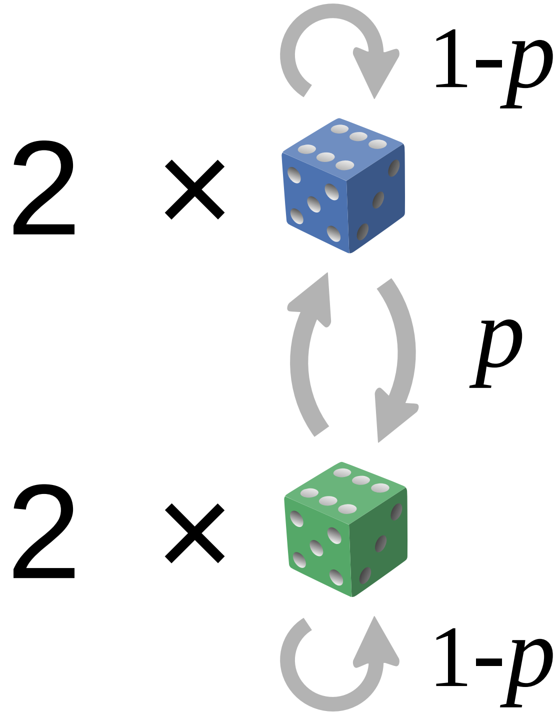
```
ACATGGCTC...
ATACGTTCC...
TTACGGTTC...
ATCCGGTAC...
ATACAGTCT...
```

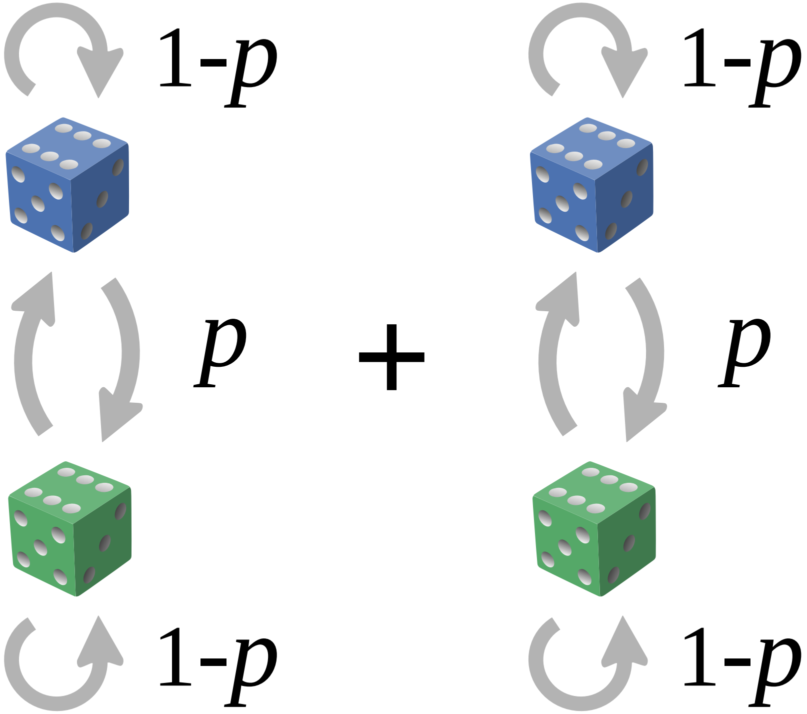
Say we are given *two* sequences

Double roll
of a single die
per turn

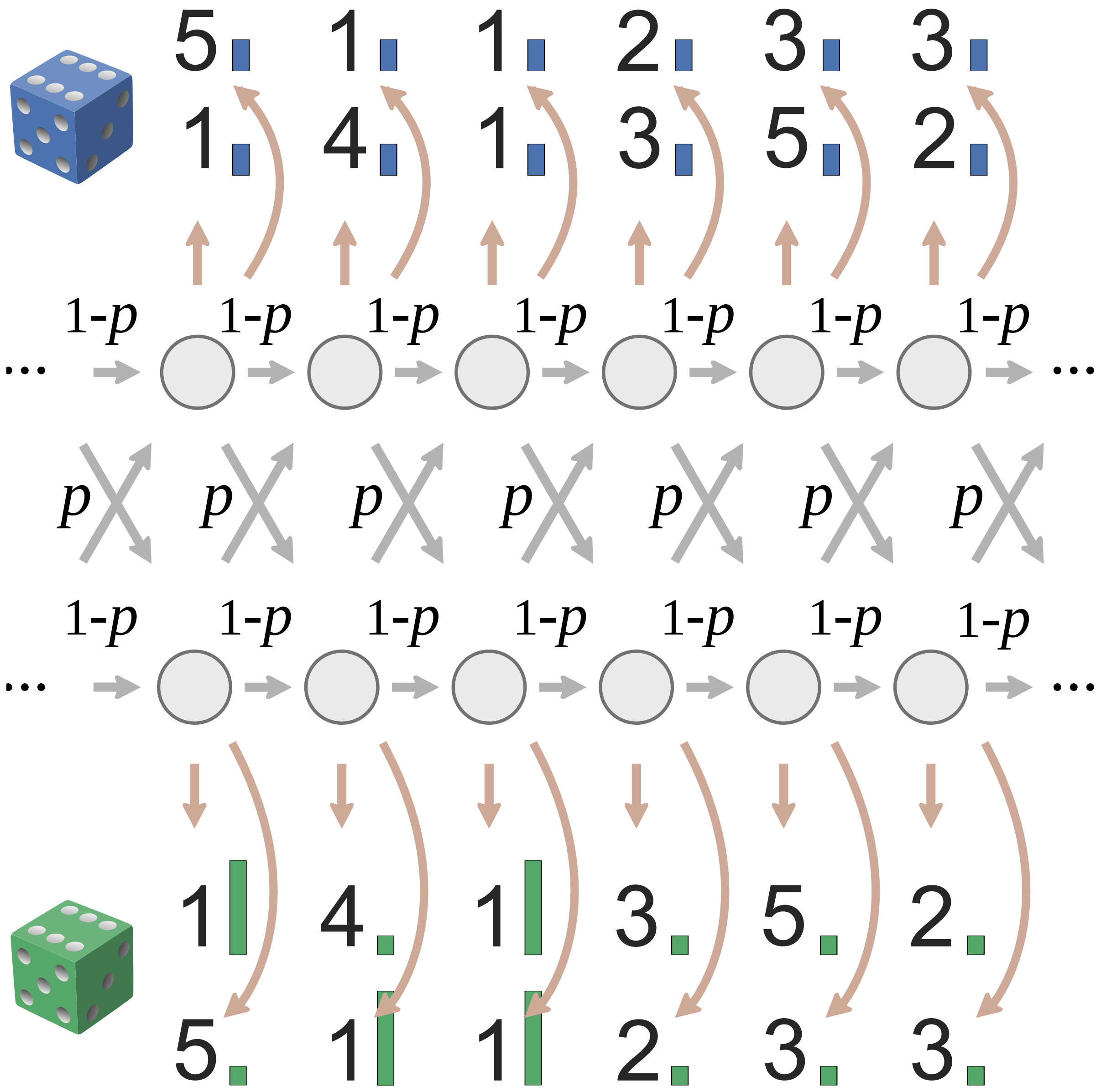
vs.

Two independent
die rolling games





Double roll ↔ Pair HMM



Two sequences from a single *(unknown)* path?

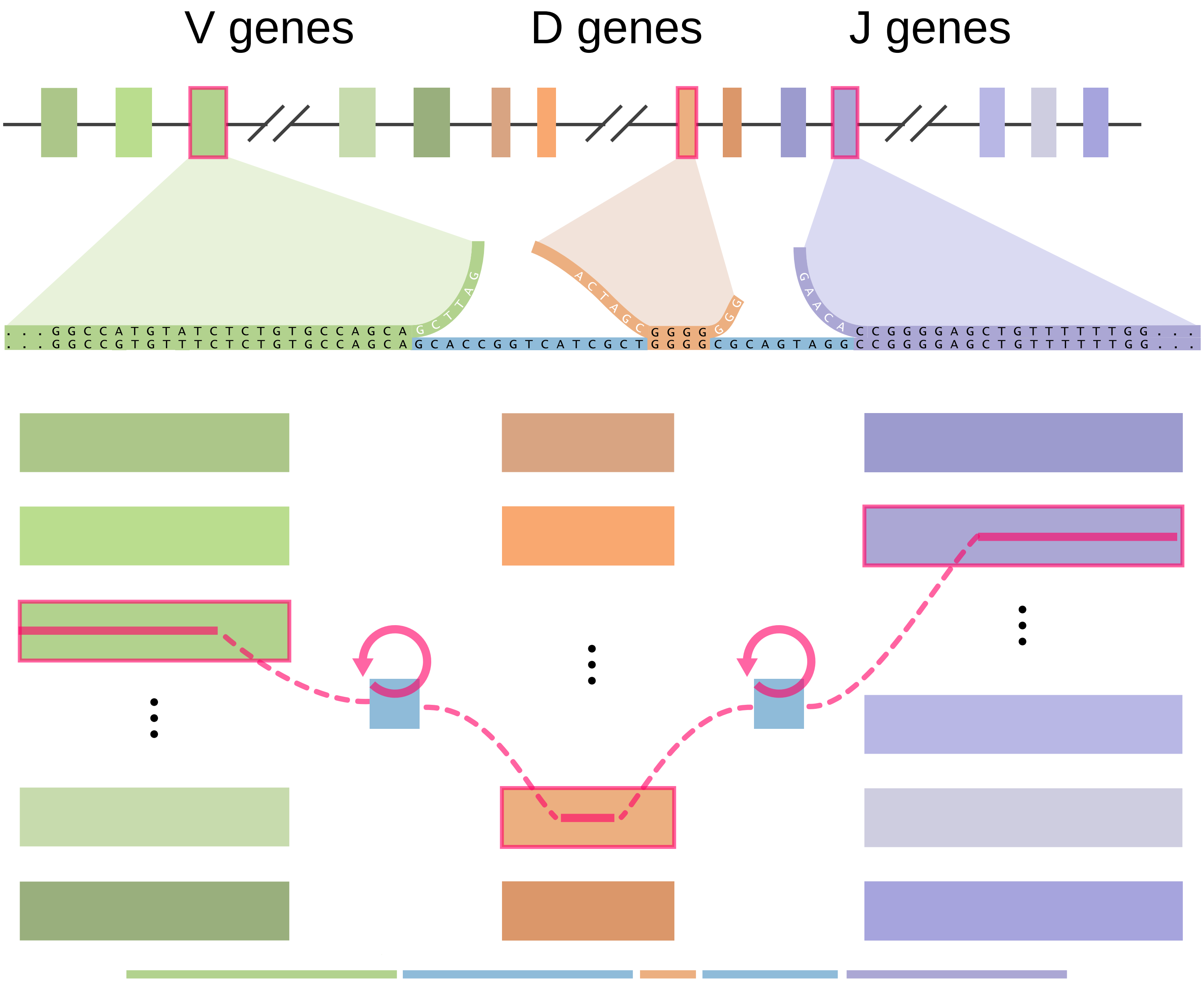
The forward algorithm for HMMs gives probability of generating observed sequence x from a given HMM:

$$\mathbb{P}(x) = \sum_{\text{paths } \sigma} \mathbb{P}(x; \sigma),$$

$$\mathbb{P}(x, y) = \sum_{\text{paths } \sigma} \mathbb{P}(x, y; \sigma),$$

probability of generating two sequences x and y from the same path through the HMM (i.e. from the same rearrangement event).

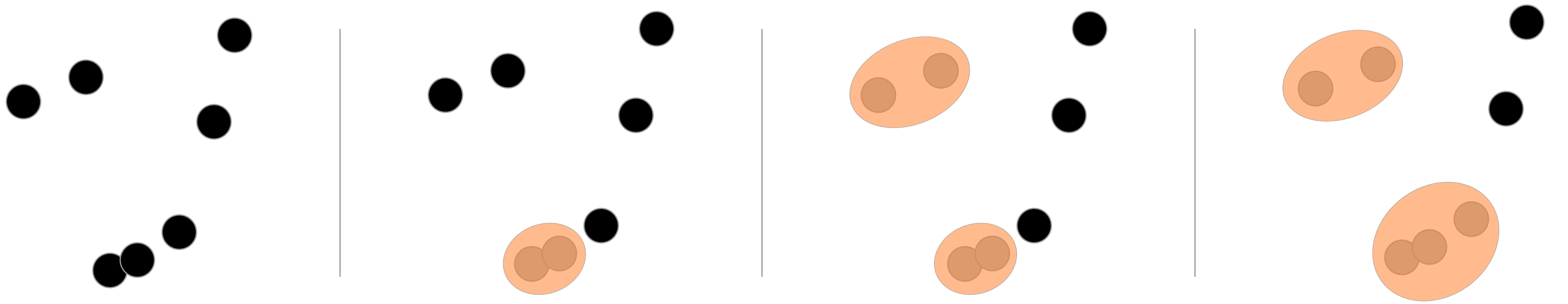
This is obtained by *efficiently* summing across paths.


V genes
D genes
J genes
GCTTAG
ACTAGC
GGGG
GGG
GAACA
... GGCCATGTATCTCTGTGCCAGCA
... GGCCGTGTTTCTCTGTGCCAGCA
GCACCGGTCATCGCT
GGGG
CGCAGTAGG
CCGGGGAGCTGTTTTTTGG ...
CCGGGGAGCTGTTTTTTGG ...

Do sets of sequences come from a single rearrangement event?

$$\frac{\mathbb{P}(A \cup B)}{\mathbb{P}(A)\mathbb{P}(B)} = \frac{\mathbb{P}(A \cup B \mid \text{single rearrangement})}{\mathbb{P}(A, B \mid \text{independent rearrangements})}$$

Use this for agglomerative clustering:



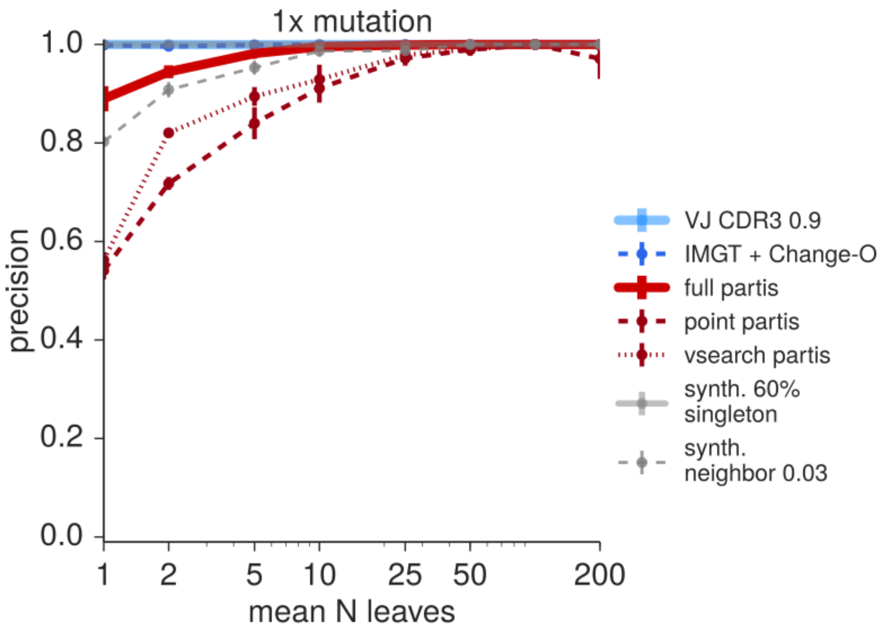
Goal: maximum likelihood clustering

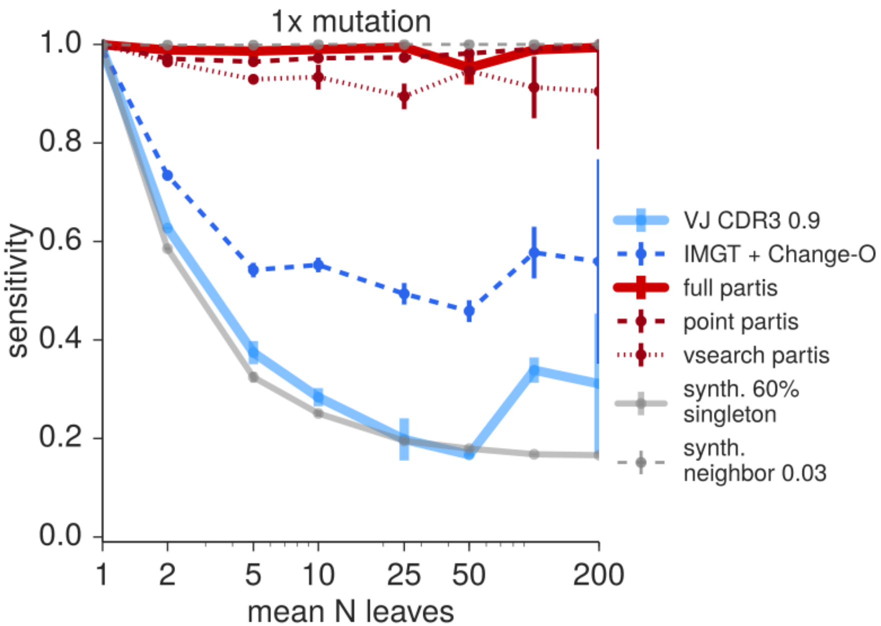
Find the maximum of

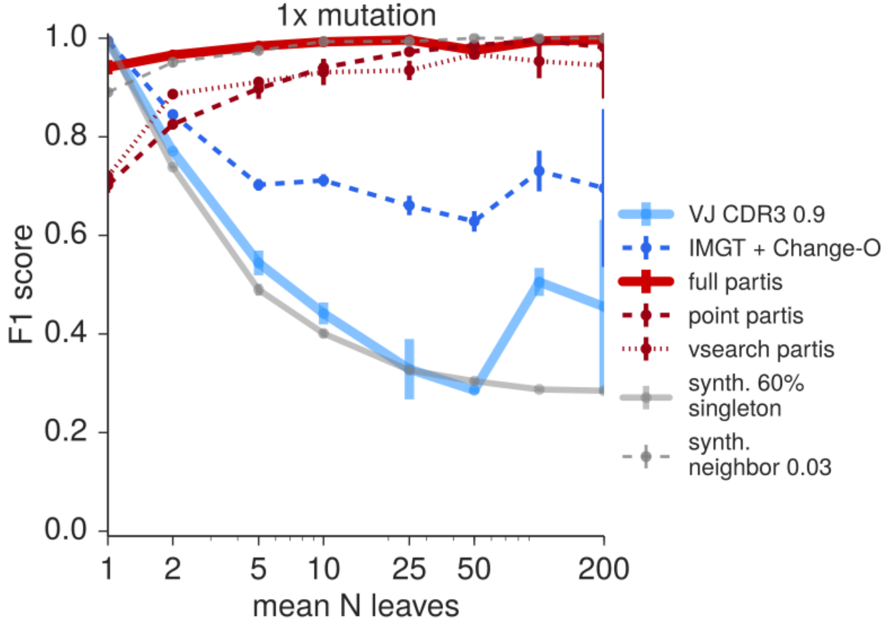
$$L(\{C_i\}_{i=1,...,k}) = \prod_i \mathbb{P}(C_i)$$

across clusterings $\{C_i\}_{i=1,...,k}$ of our sequences.

HMM-based clustering works under simulation







Likelihood-based clustering of clonal families

Phylogenetic empirical Bayes method for inferring unmutated common ancestor (perhaps?):

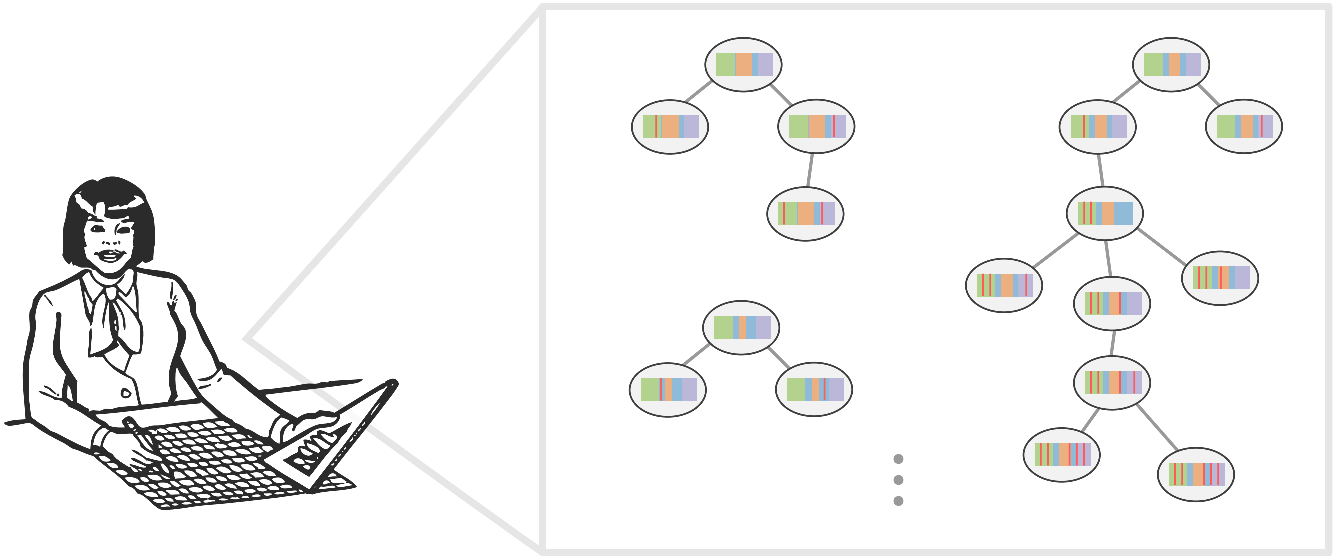
- Clonalyst: Kepler (2014-2015)

Use forward algorithm with parameter-rich HMMs for efficient evaluation of marginal probability.

- partis: Ralph & M. (2016) *in prep.*

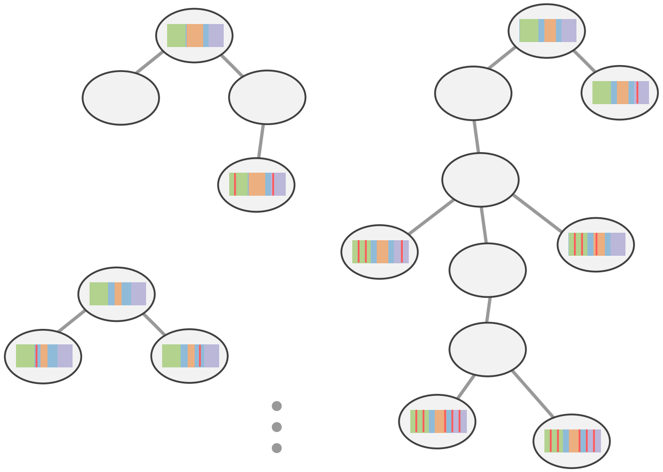
3. Reconstruct BCR phylogenetic trees

reality

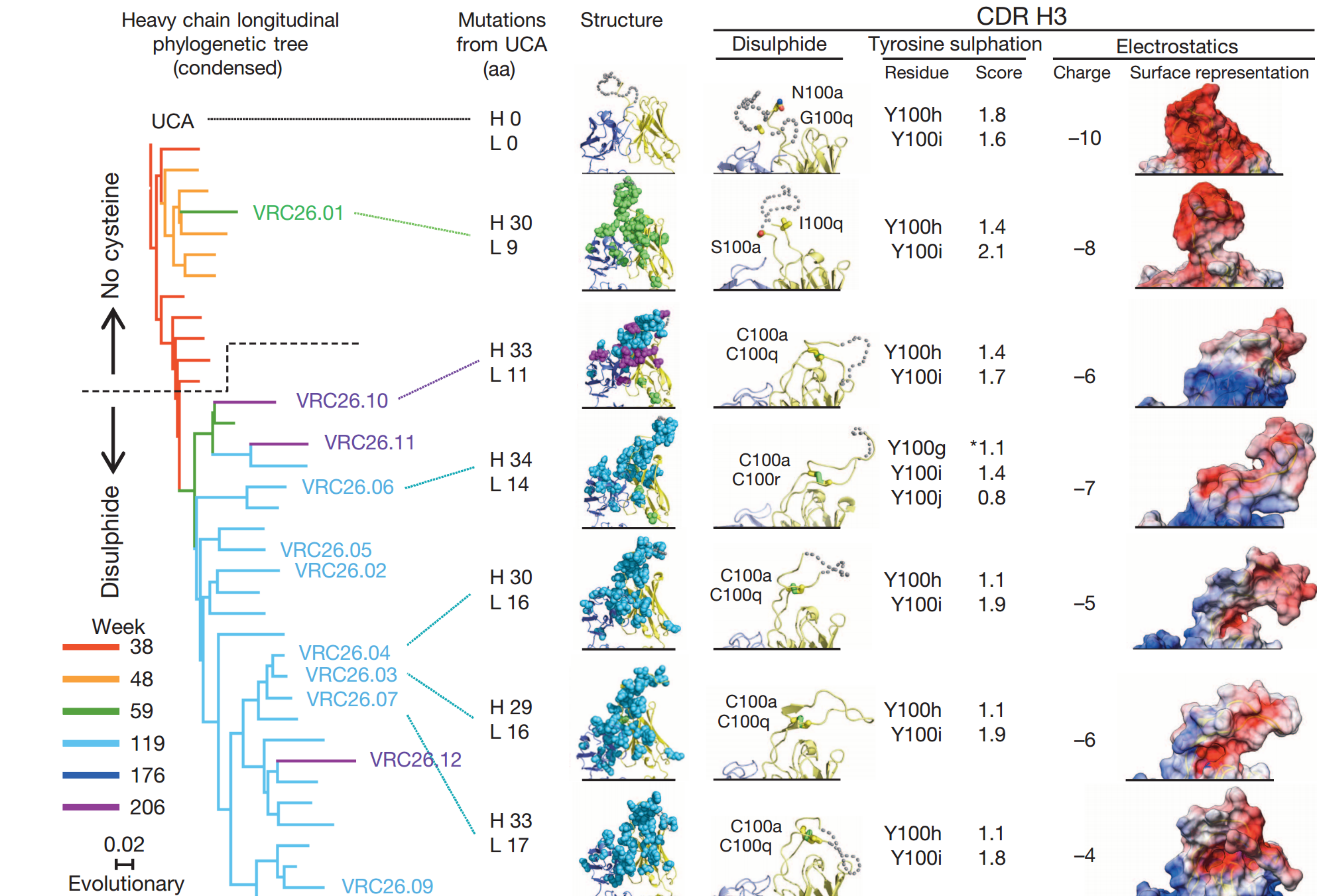


```
ACATGGCTC...
ATACGTTCC...
TTACGGTTC...
ATCCGGTAC...
ATACAGTCT...
```

inference



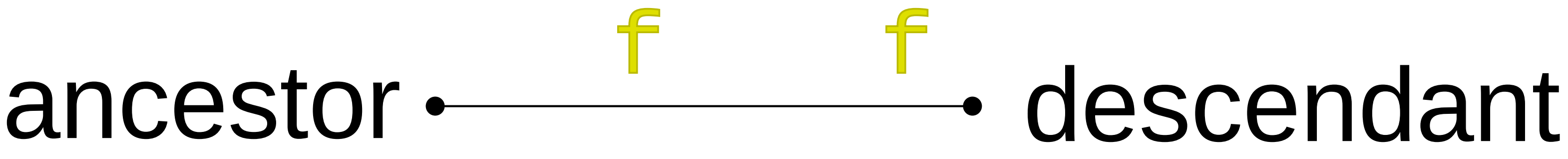
c Structural development of CAP256-VRC26 lineage



Mutations from UCA (aa)	CDR H3			
	Disulphide	Tyrosine sulphation		Electrostatics
		Residue	Score	Charge
H 0 L 0	N100a G100q	Y100h Y100i	1.8 1.6	−10
H 30 L 9	S100a I100q	Y100h Y100i	1.4 2.1	−8
H 33 L 11	C100a C100q	Y100h Y100i	1.4 1.7	−6
H 34 L 14	C100a C100r	Y100g Y100i Y100j	*1.1 1.4 0.8	−7
H 30 L 16	C100a C100q	Y100h Y100i	1.1 1.9	−5
H 29 L 16	C100a C100q	Y100h Y100i	1.1 1.9	−6
H 33 L 17	C100a C100q	Y100h Y100i	1.1 1.8	−4

Likelihood-based phylogenetics

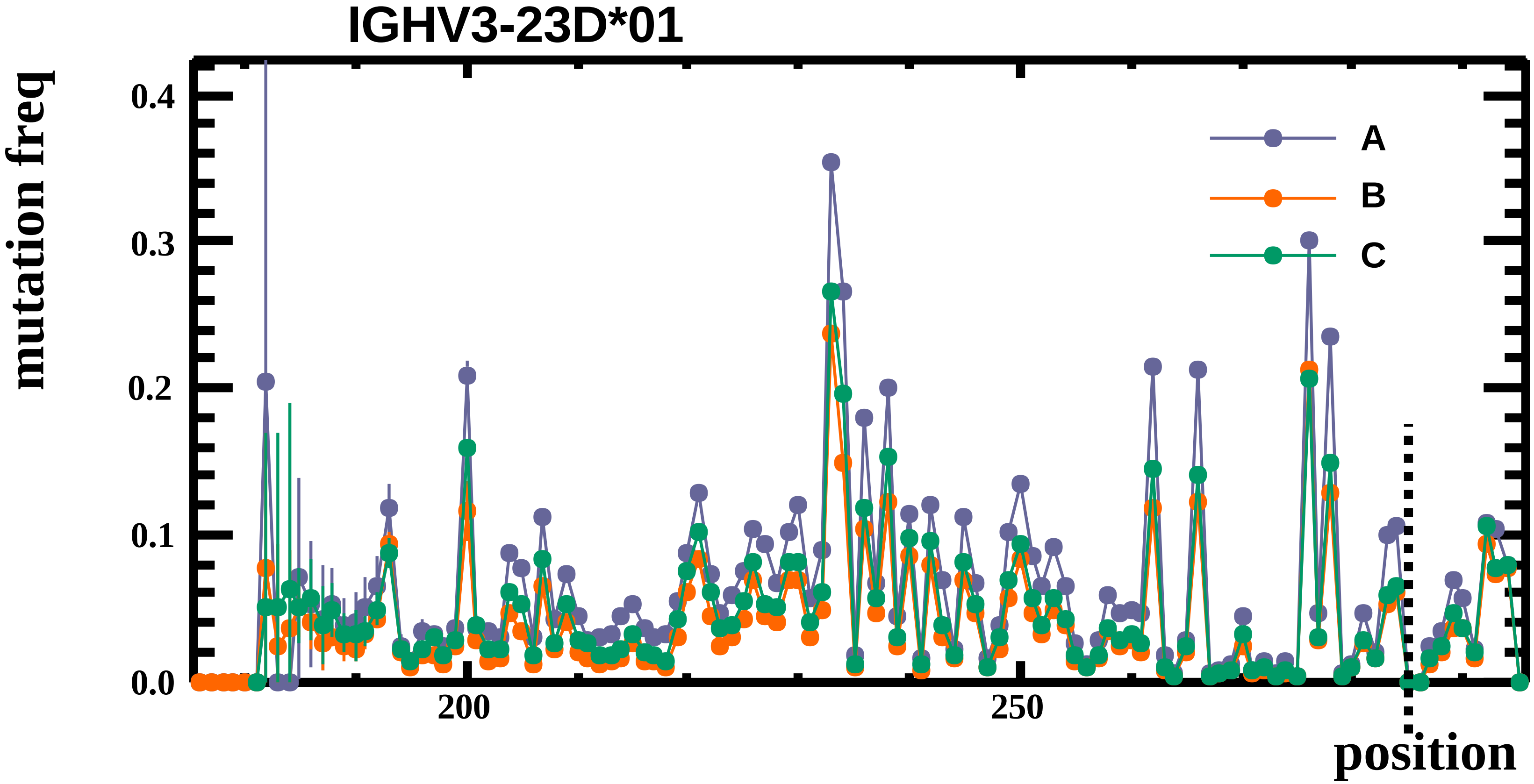
Mutations appear at some rate λ:



Mutations change bases according to substitution matrix:

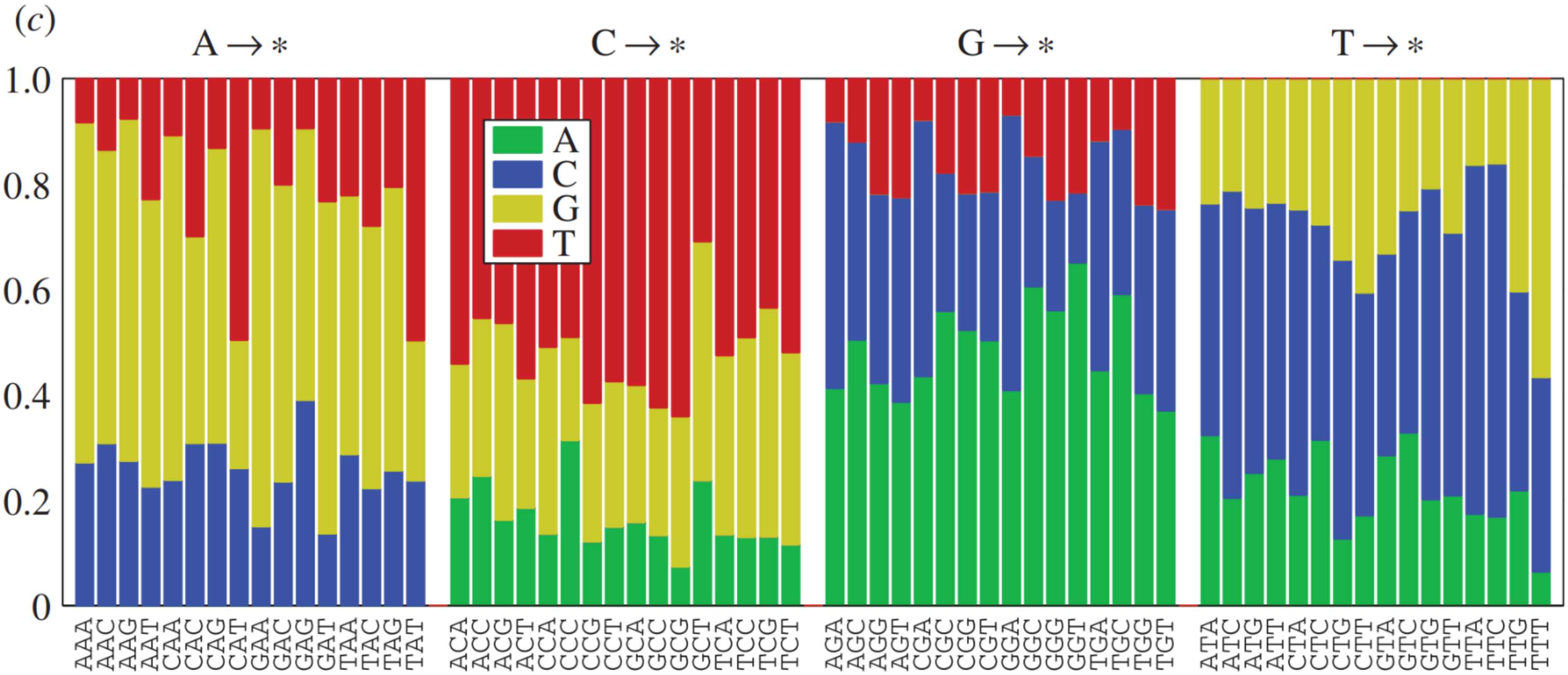
$$\begin{pmatrix} p_{AA} & p_{AG} & p_{AC} & p_{AT} \\ p_{GA} & p_{GG} & p_{GC} & p_{GT} \\ p_{CA} & p_{CG} & p_{CC} & p_{CT} \\ p_{TA} & p_{TG} & p_{TC} & p_{TT} \end{pmatrix}$$

Traditional phylogenetic approaches assume that the same evolutionary process is happening at each site.



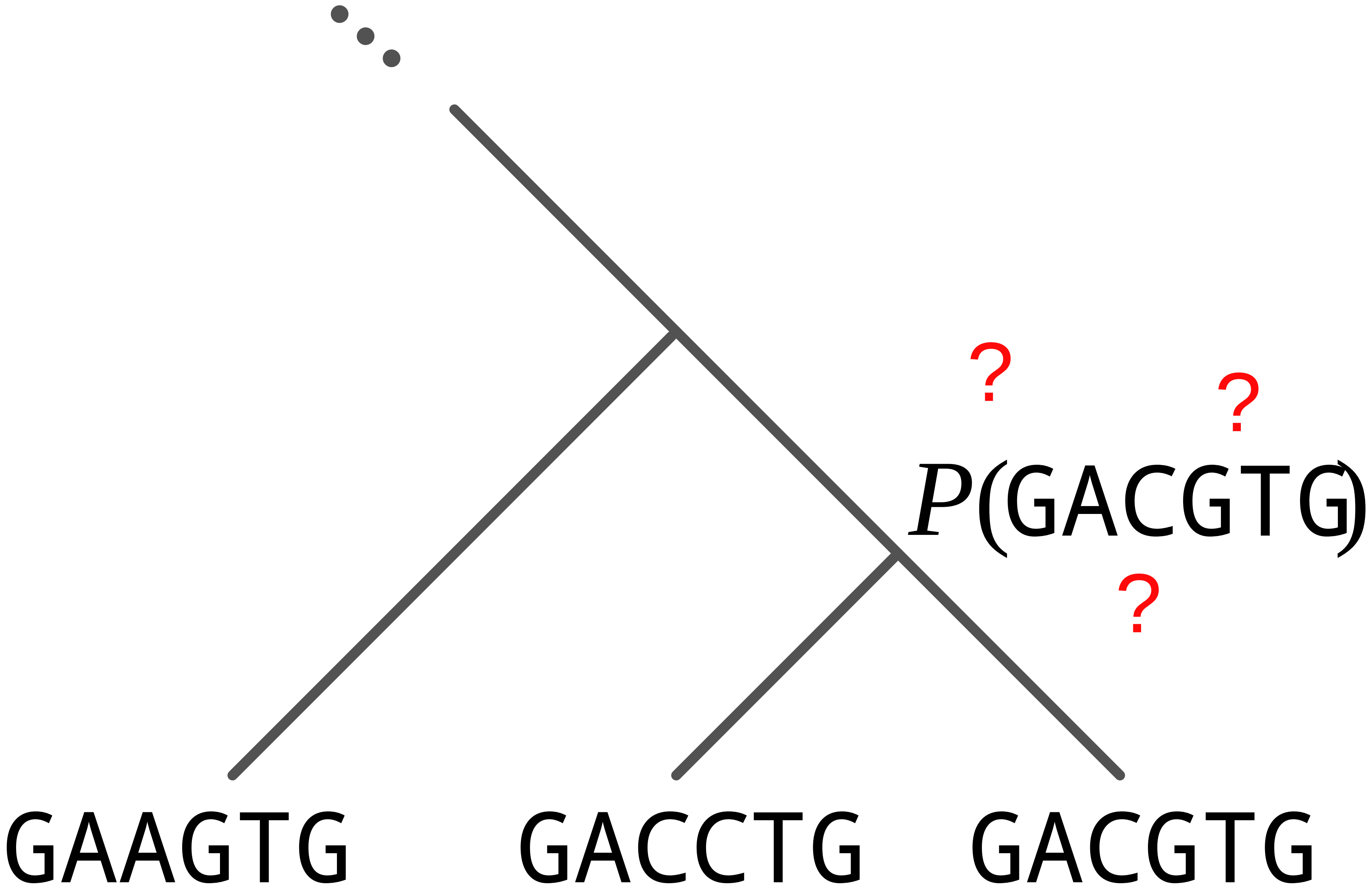
This does not hold for B cell receptor sequences.

Context sensitive substitutions



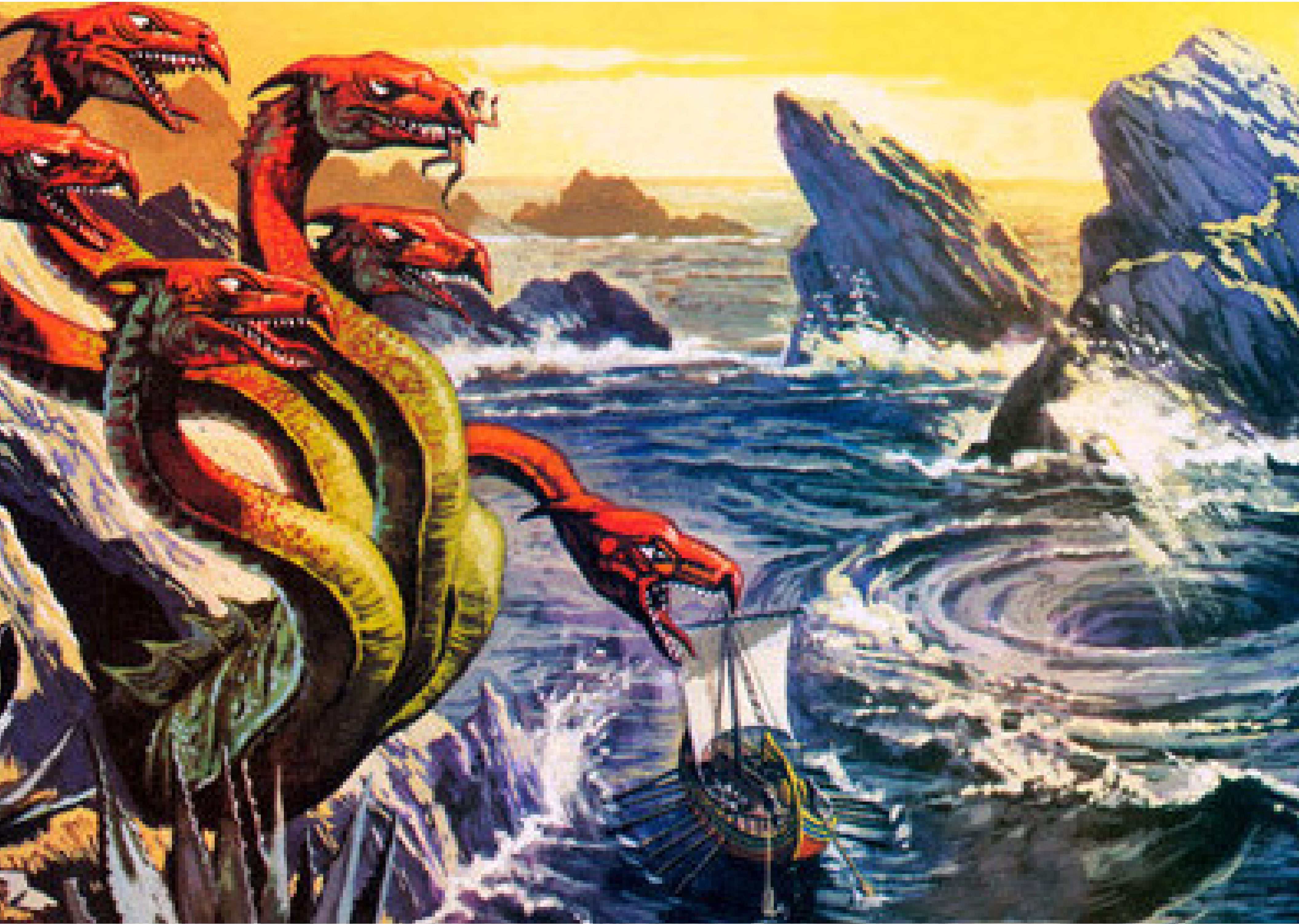
Elhanati et al, 2015

Context sensitive likelihoods are hard



Context sensitive likelihoods are hard

- Siepel & Haussler (MBE 2003); Saunders & Green (MBE 2007): context-sensitive likelihoods via along-sequence Markov cond'n
- Lunter & Hein (Bioinformatics 2004): MCMC approach to estimating likelihoods
- Christensen, Hobolth & Jensen (J Comp Biol 2005): pseudo-likelihood analysis using parsimony-ish inference on flanking bases
- Baele, Van de Peer, & Vansteelandt, (Sys Bio 2008): pseudo-likelihood analysis using context-insensitive likelihood inference on flanking bases
- Bérard & Guéguen (Sys Bio 2012): specific context dependent model enabling independence assumption in many cases
- Peter Ralph (unpublished): approximations using interacting particle systems



Special sauce: per site models

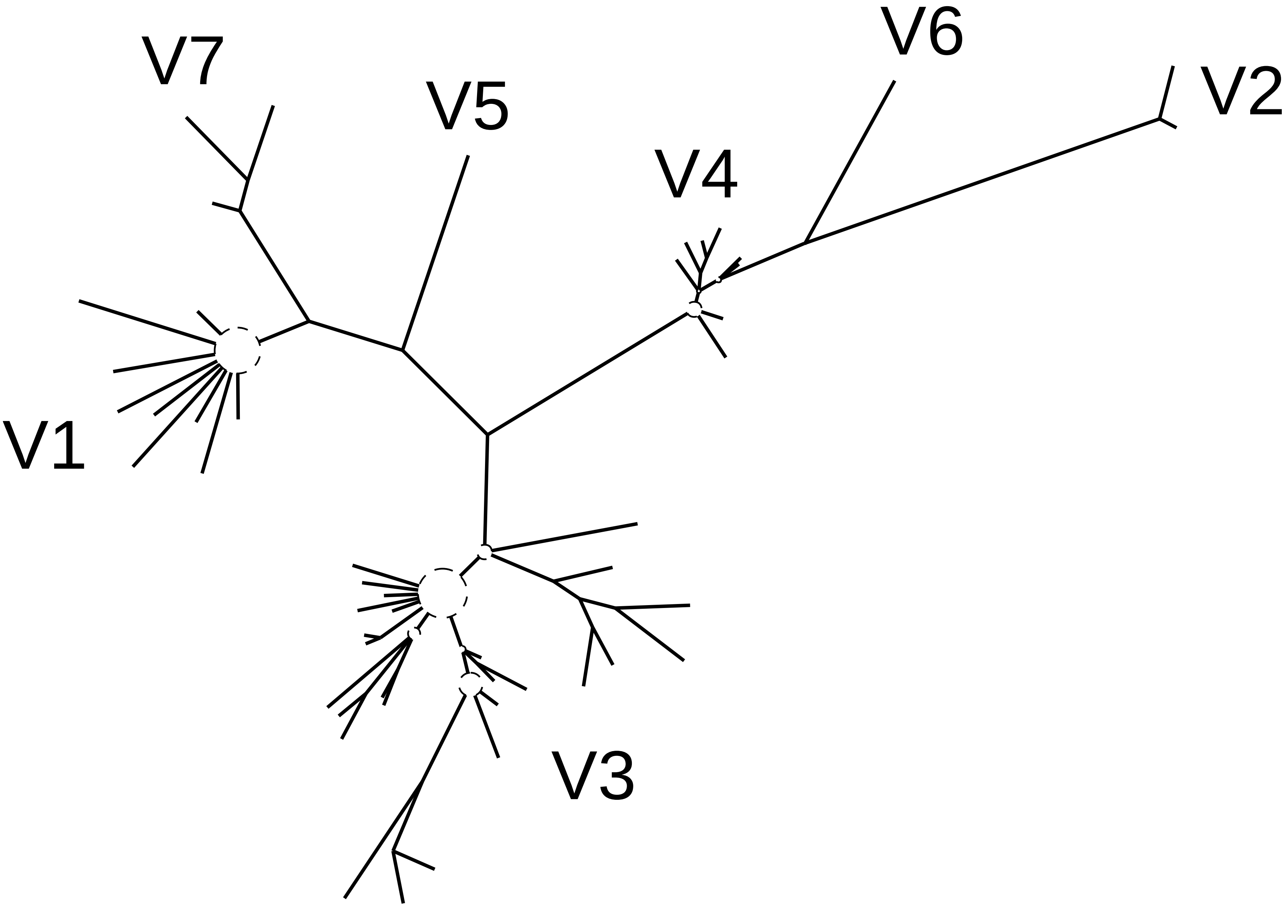
Each site s of every germline gene gets its own substitution rate λ_s and mutation rate matrix:

$$\begin{pmatrix} p_{\cdot A} & p_{\cdot G} & p_{\cdot C} & p_{\cdot T} \\ p_{\cdot A} & p_{\cdot G} & p_{\cdot C} & p_{\cdot T} \\ p_{\cdot A} & p_{\cdot G} & p_{\cdot C} & p_{\cdot T} \\ p_{\cdot A} & p_{\cdot G} & p_{\cdot C} & p_{\cdot T} \end{pmatrix}$$

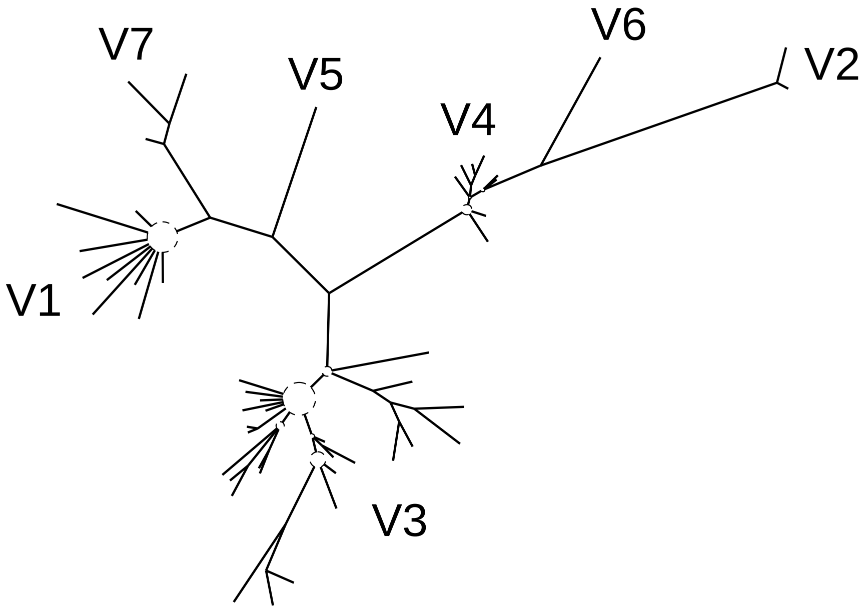
$\approx 350 \times 5 \times 300 = 525,000$ parameters

Ouch! Need to be careful.

B cell germline gene phylogeny



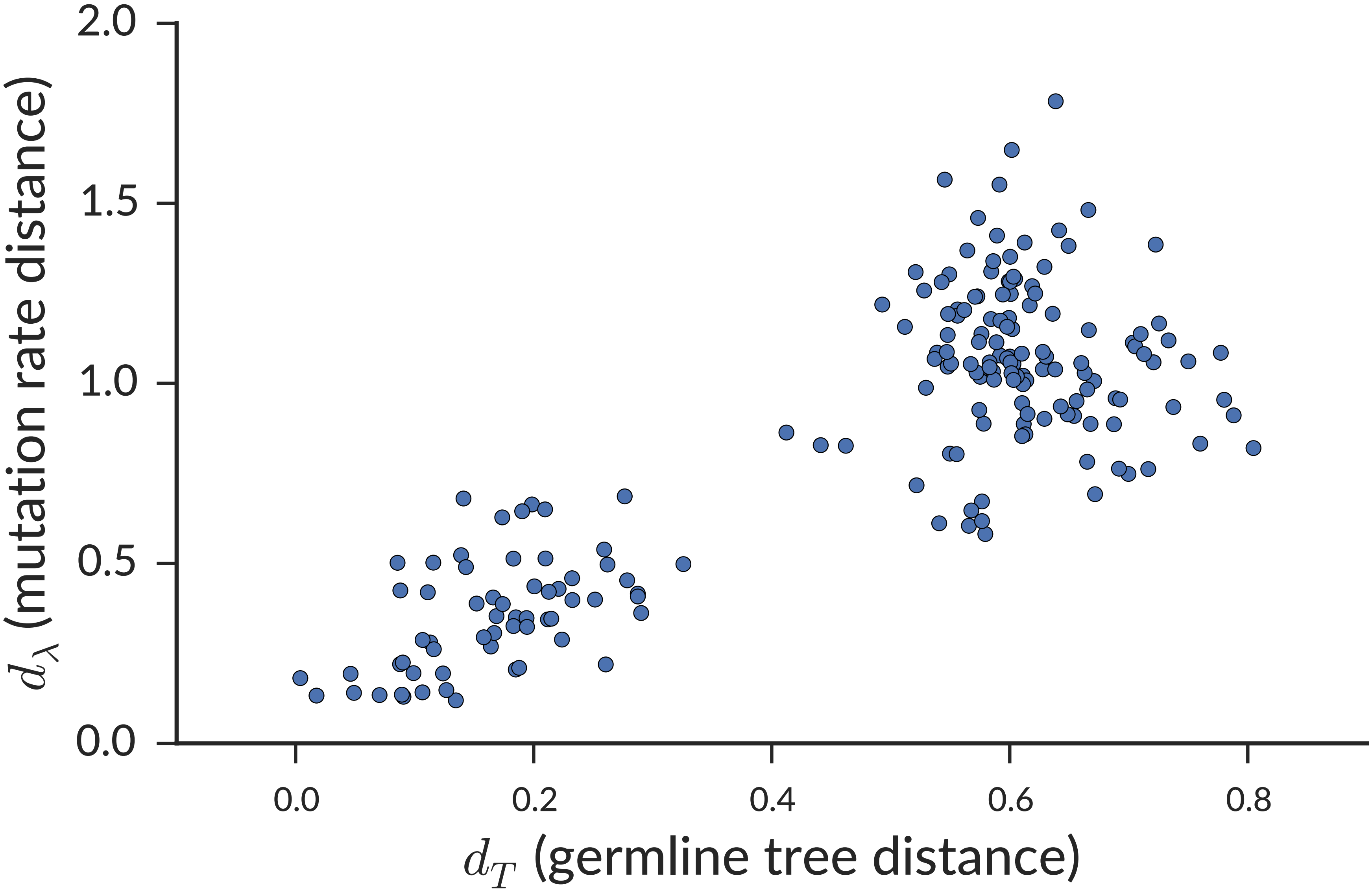
Q: do closely related genes evolve similarly?



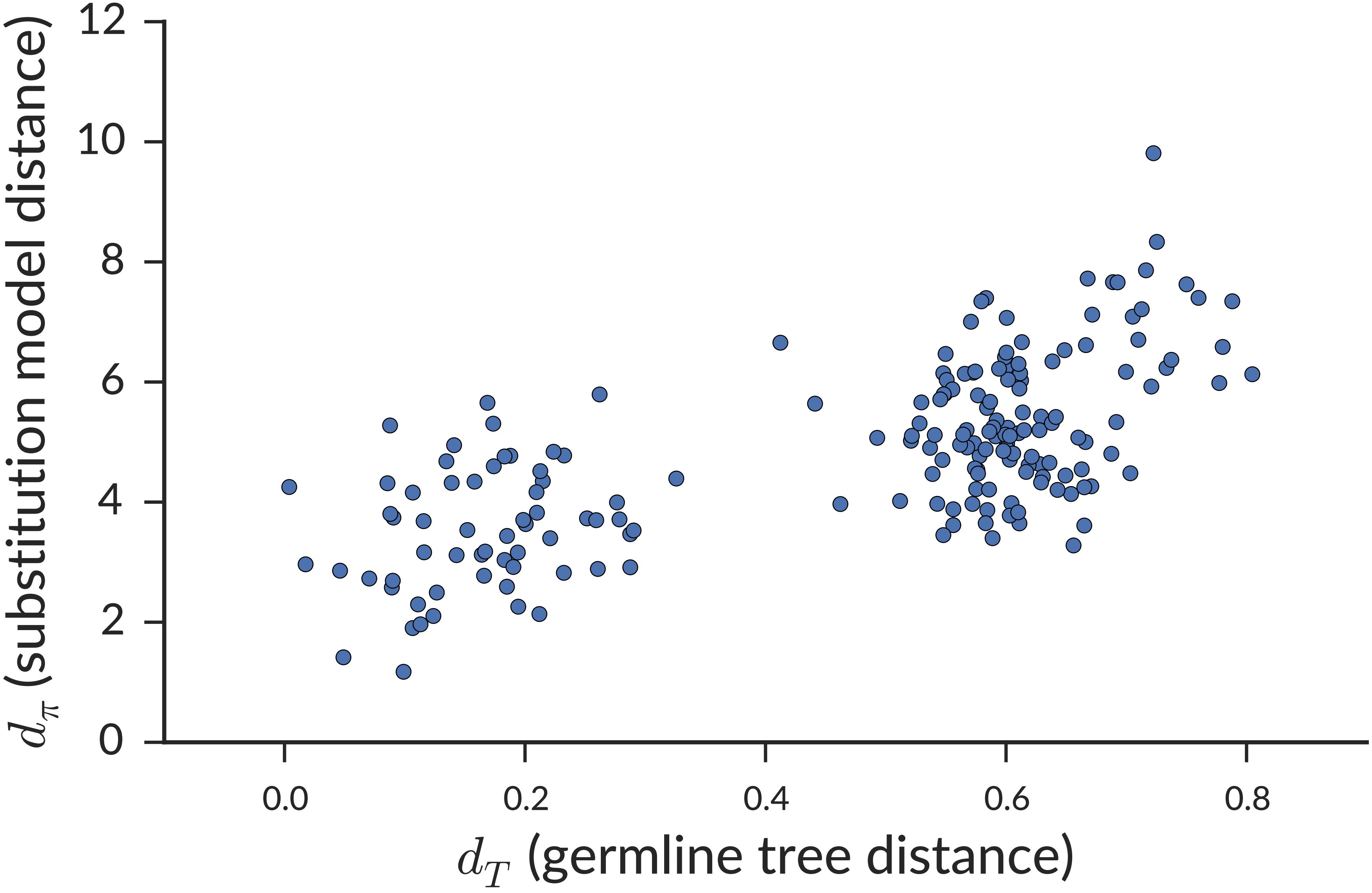
- Fit models for the 20 genes for which we have the most data (these are good estimates)
- Compare parameter fits between these genes
- Compute evolutionary distance between these genes

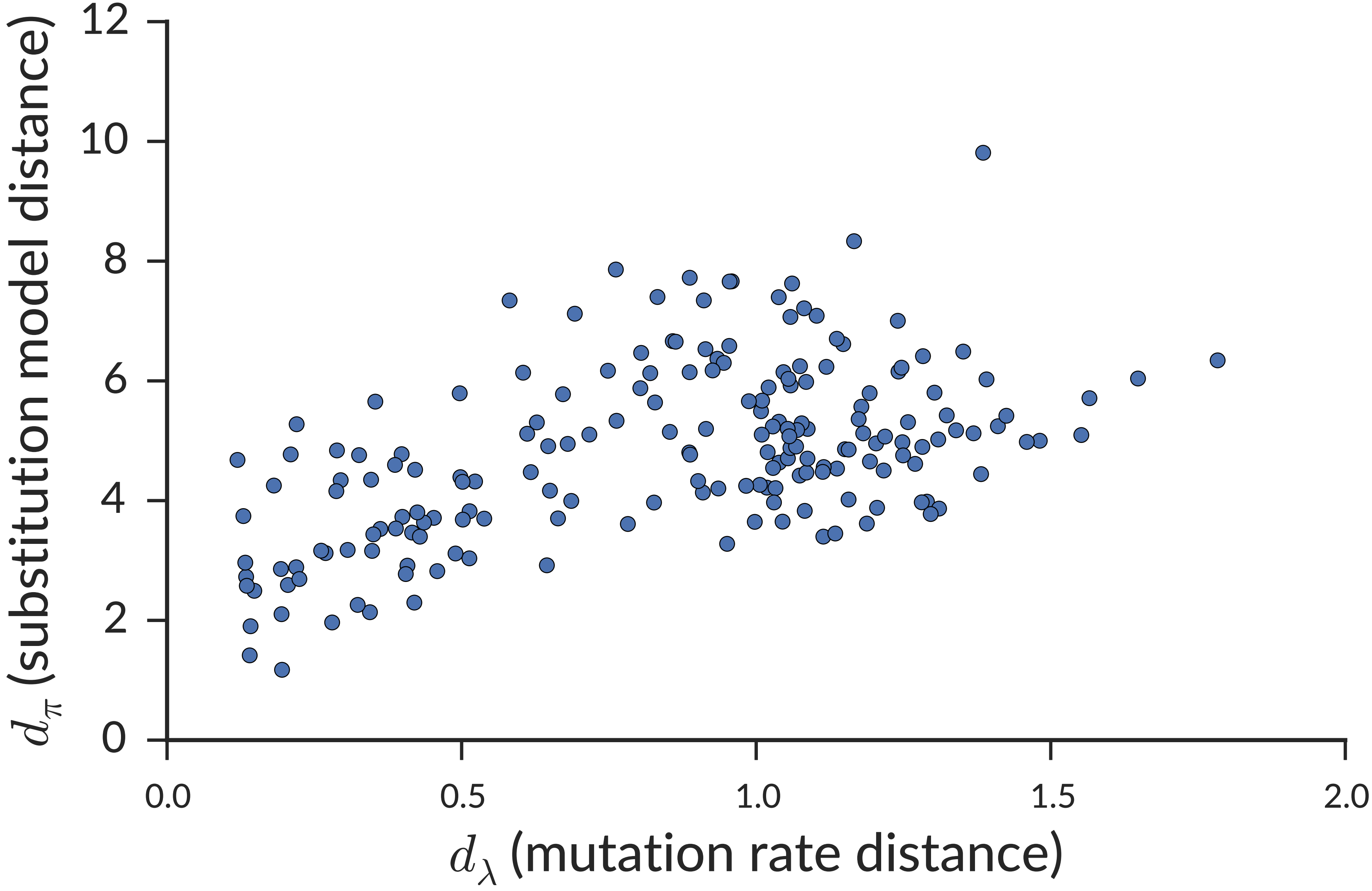
Top 20 genes

```
IGHV1-18*04    34507
IGHV1-2*04     19432
IGHV1-46*02    18453
IGHV1-69D*01   34218
IGHV3-15*07    18789
IGHV3-23D*01   58627
IGHV3-53*02    16552
IGHV3-64*04    38324
IGHV3-69-1*02  22445
IGHV3-7*01     78868
IGHV3-7*02     17992
IGHV3-74*03    18015
IGHV3-9*02     24010
IGHV3-NL1*01   51790
IGHV4-30-4*06  17419
IGHV4-34*13    14089
IGHV4-4*07     20816
IGHV4-61*02    18944
IGHV4-61*08    18644
IGHV5-51*03    25549
```



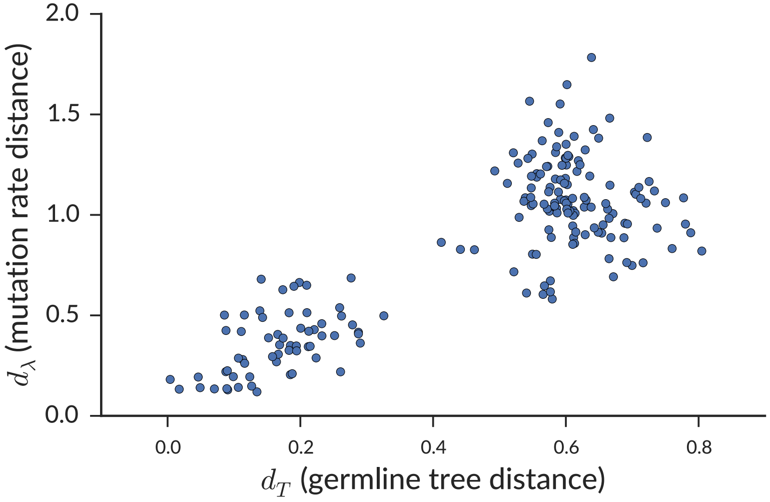
2.0
1.5
1.0
0.5
0.0
d_λ (mutation rate distance)
0.0
0.2
0.4
0.6
0.8
d_T (germline tree distance)

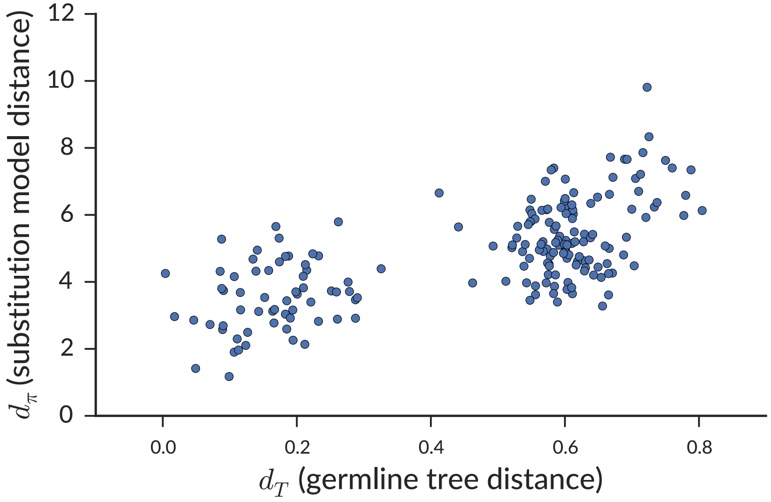

12
10
8
6
4
2
0
d_π (substitution model distance)
0.0
0.2
0.4
0.6
0.8
d_T (germline tree distance)


12
10
8
6
4
2
0
d_π (substitution model distance)
0.0
0.5
1.0
1.5
2.0
d_λ (mutation rate distance)

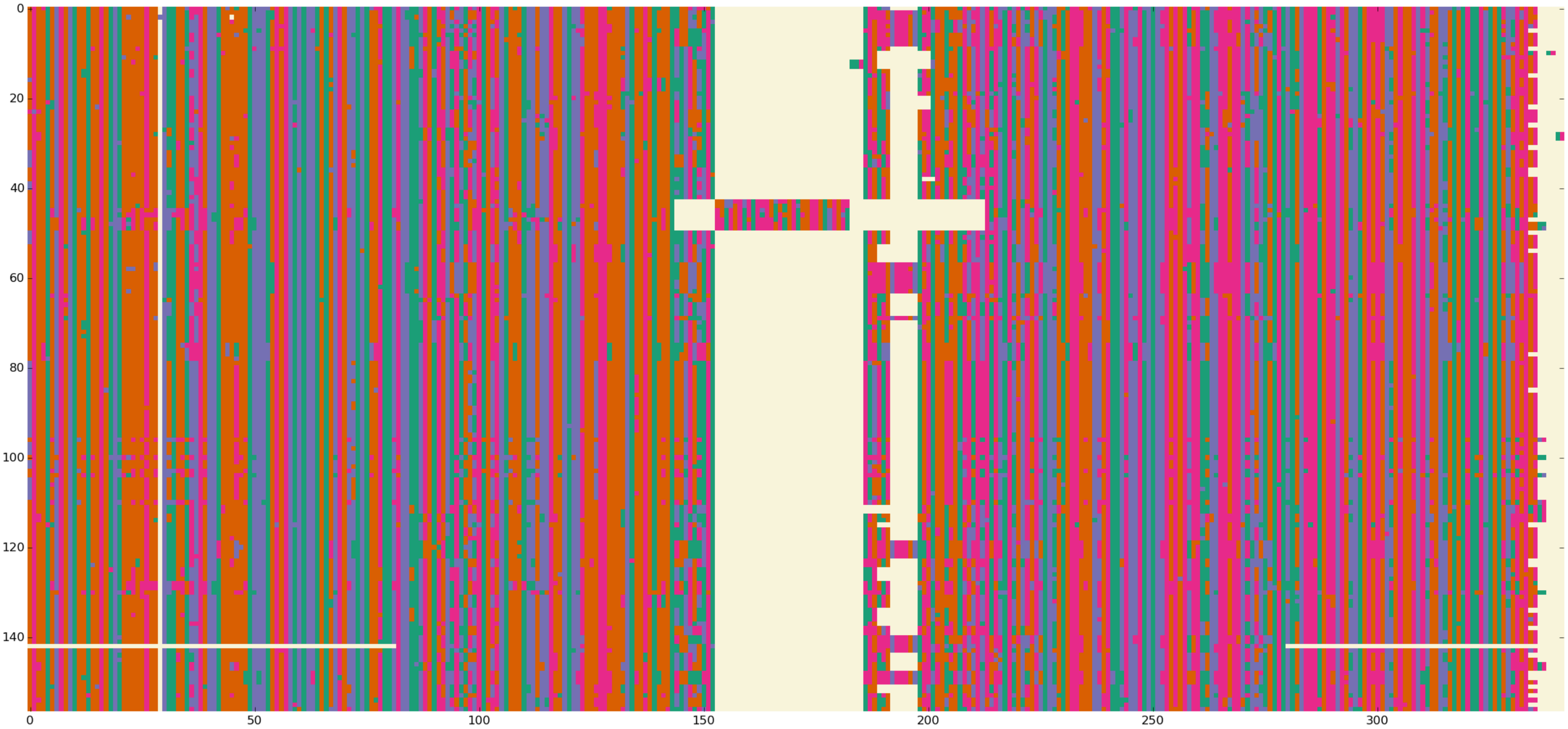
We can use this for model fitting

Use homologous sites to regularize our rate parameters





Multiple sequence alignment of V3 germline



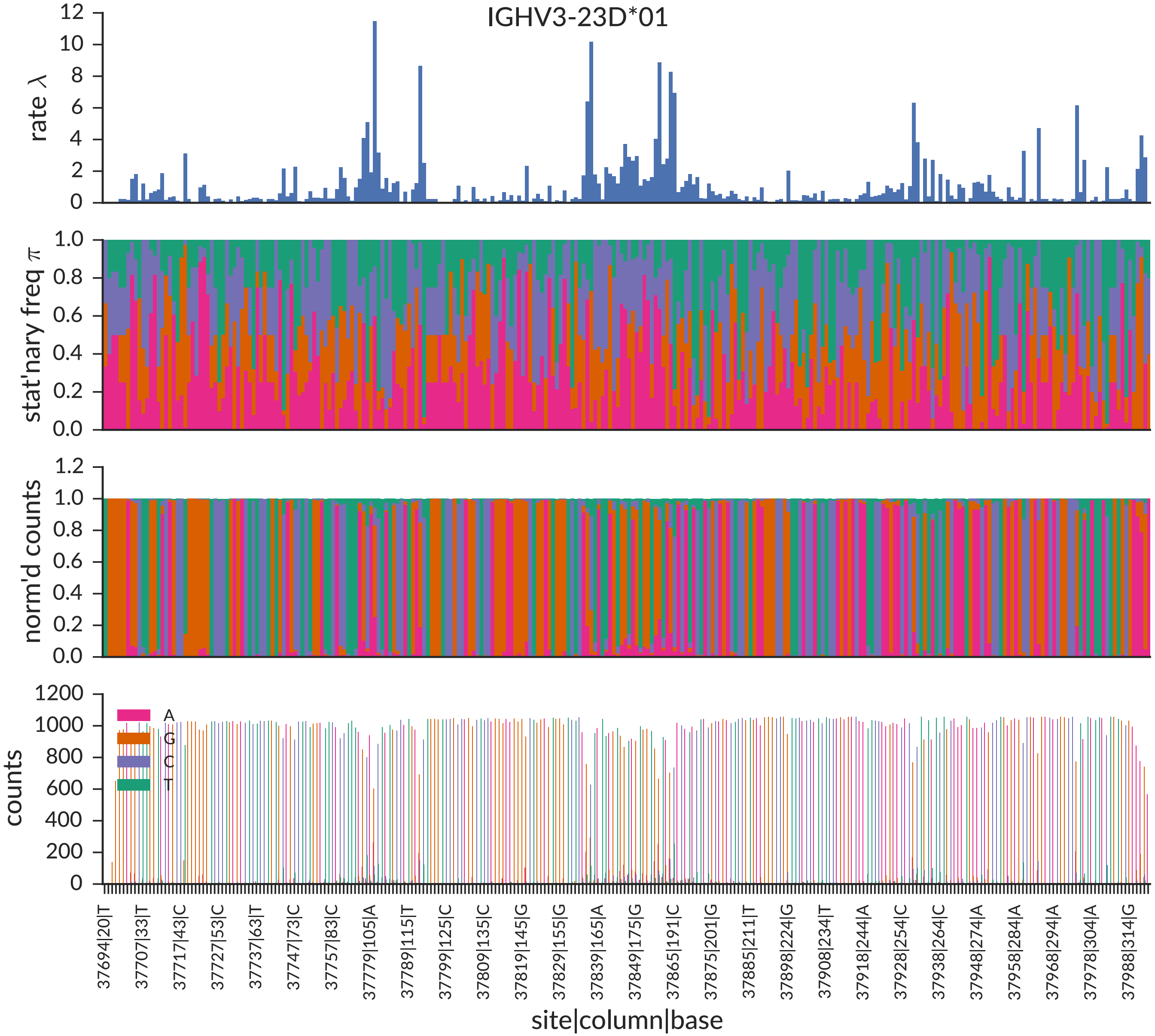
(V3 genes along rows, colors for the four bases)

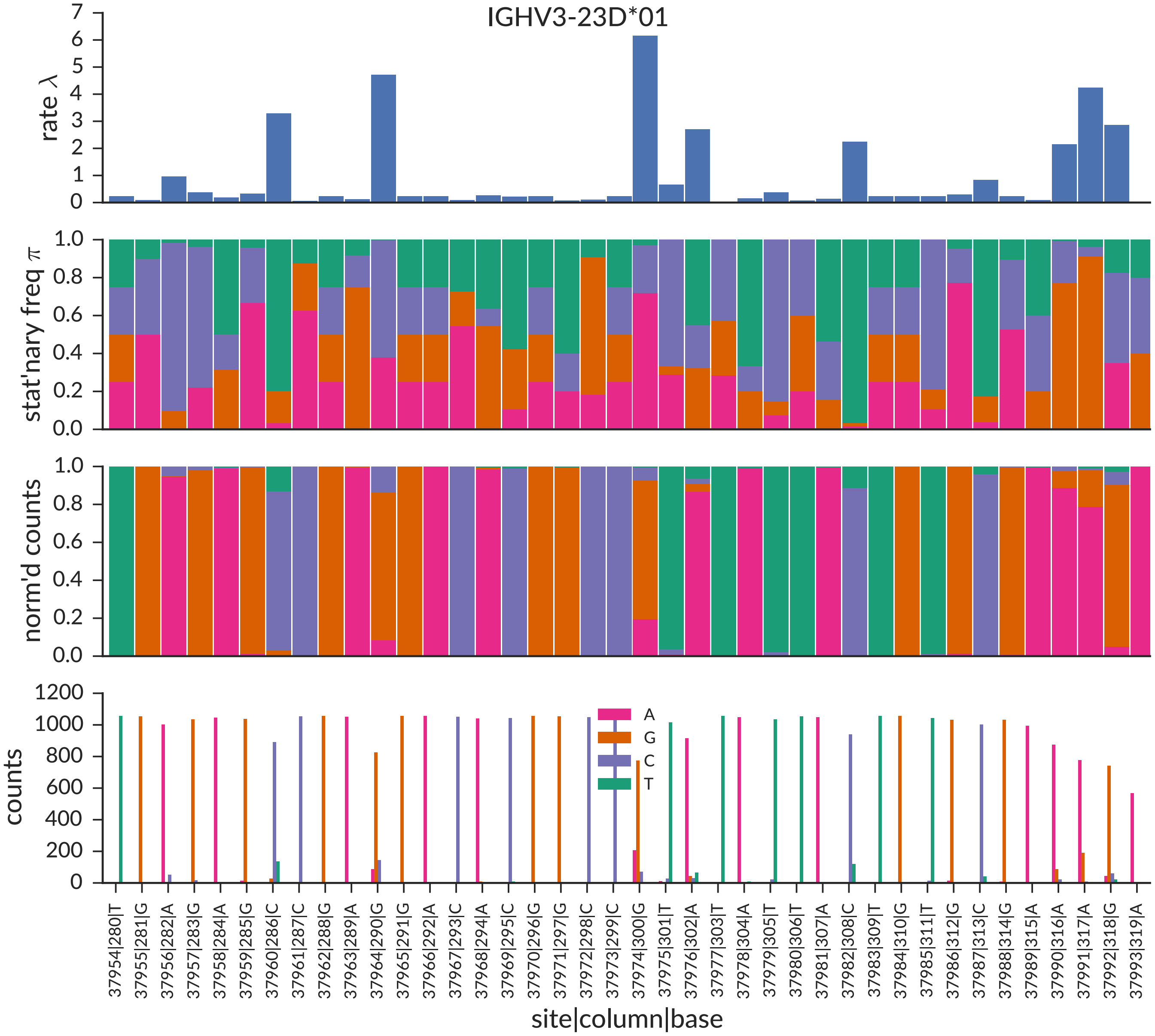
Joint estimation for sites in the same column

Assume that sites in the same column evolve *similarly* within-host.

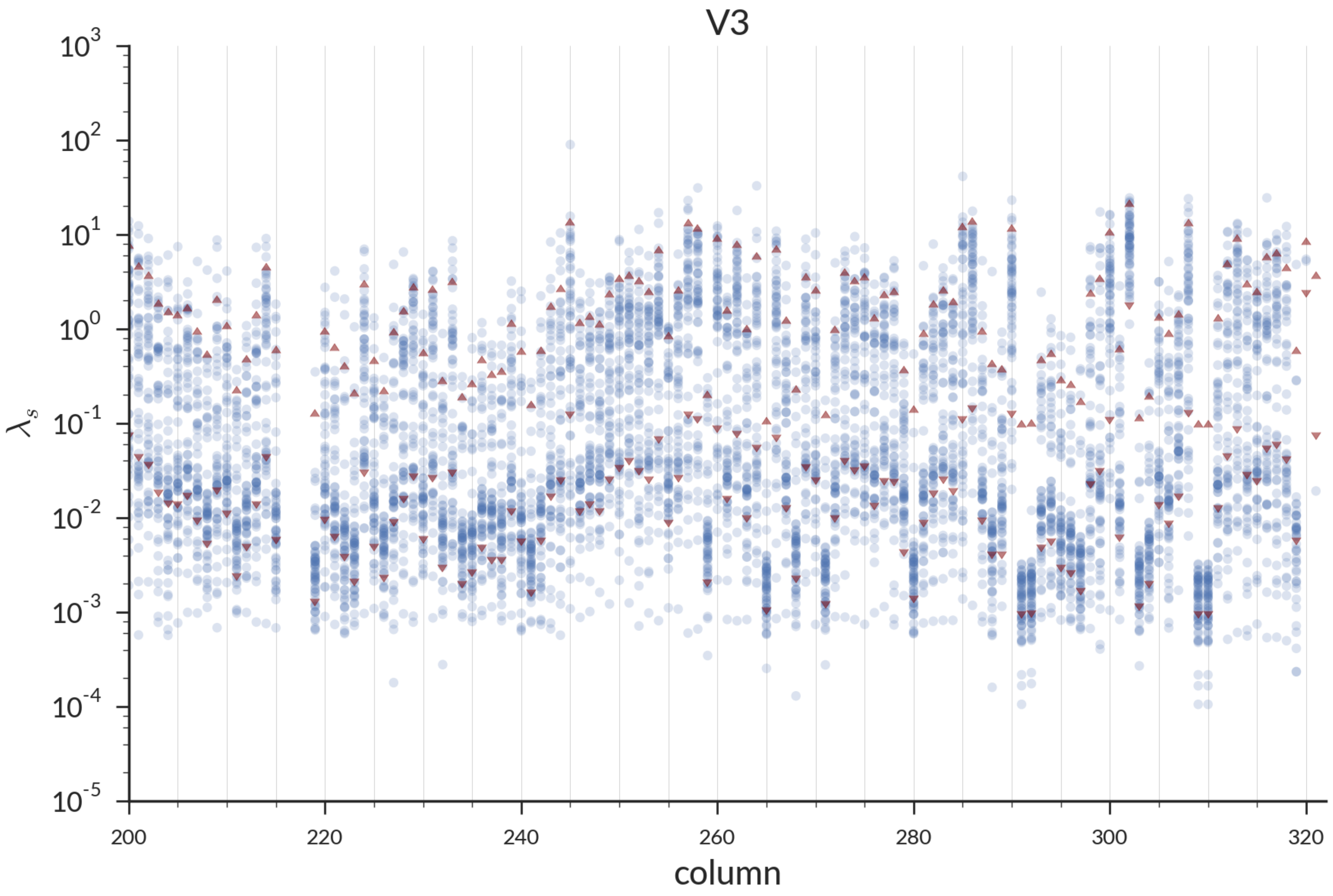
When data is weak, draw estimates back to a per-column average.

- Substitution rate $\lambda_s \sim \mathrm{Gamma}(\omega_c, \theta_c)$
- Gamma mode $\omega_c \sim \text{Log-normal}(1, 1)$
- Gamma dispersion $\theta_c \sim \text{Lévy}(3)$
- Stationary distribution $\pi_s \sim \mathrm{Dirichlet}(3, 3, 3, 3)$
- Branch length $t \sim \mathrm{Exponential}(0.1)$

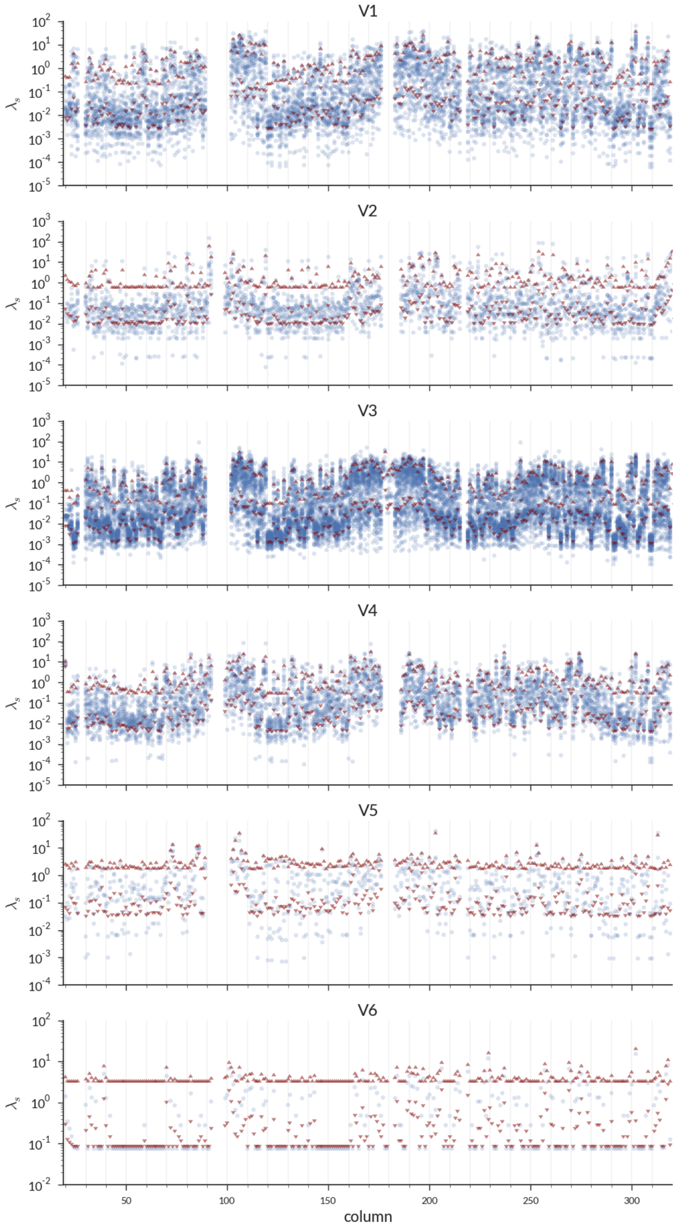

IGHV3-23D*01
rate λ
0
2
4
6
8
10
12
stat'nary freq π
0.0
0.2
0.4
0.6
0.8
1.0
norm'd counts
0.0
0.2
0.4
0.6
0.8
1.0
1.2
counts
0
200
400
600
800
1000
1200
🟥 A
🟧 G
🟪 C
🟩 T
37694|20|T
37707|33|T
37717|43|C
37727|53|C
37737|63|T
37747|73|C
37757|83|C
37779|105|A
37789|115|T
37799|125|C
37809|135|C
37819|145|G
37829|155|G
37839|165|A
37849|175|G
37865|191|C
37875|201|G
37885|211|T
37898|224|G
37908|234|T
37918|244|A
37928|254|C
37938|264|C
37948|274|A
37958|284|A
37968|294|A
37978|304|A
37988|314|G
site|column|base


IGHV3-23D*01
rate λ
0
1
2
3
4
5
6
7
stat'nary freq π
0.0
0.2
0.4
0.6
0.8
1.0
norm'd counts
0.0
0.2
0.4
0.6
0.8
1.0
counts
0
200
400
600
800
1000
1200
A
G
C
T
37954|280|T
37955|281|G
37956|282|A
37957|283|G
37958|284|A
37959|285|G
37960|286|C
37961|287|C
37962|288|G
37963|289|A
37964|290|G
37965|291|G
37966|292|A
37967|293|C
37968|294|A
37969|295|C
37970|296|G
37971|297|G
37972|298|C
37973|299|C
37974|300|G
37975|301|T
37976|302|A
37977|303|T
37978|304|A
37979|305|T
37980|306|T
37981|307|A
37982|308|C
37983|309|T
37984|310|G
37985|311|T
37986|312|G
37987|313|C
37988|314|G
37989|315|A
37990|316|A
37991|317|A
37992|318|G
37993|319|A
site|column|base

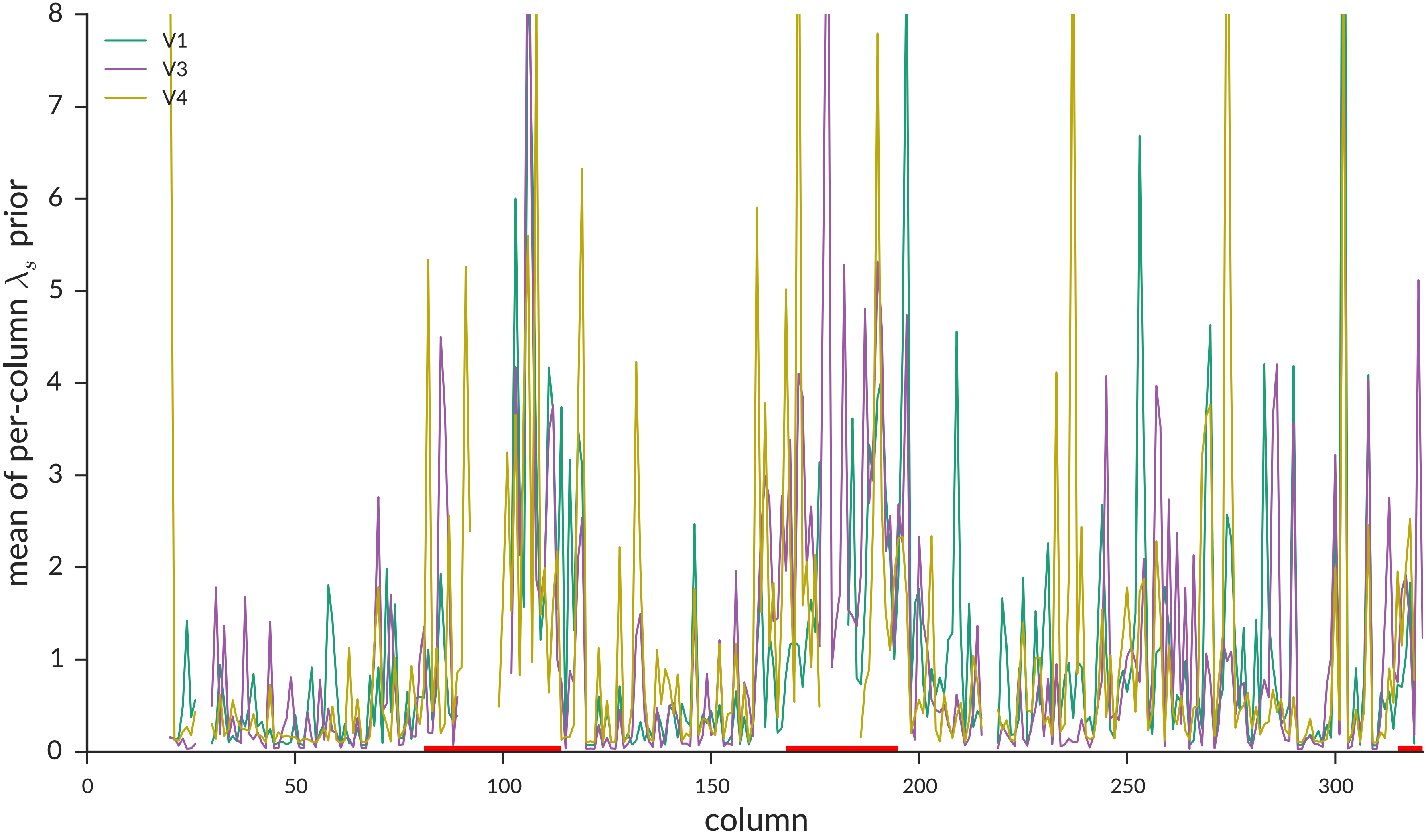
Points = rates; triangles = 95% CI of fit prior



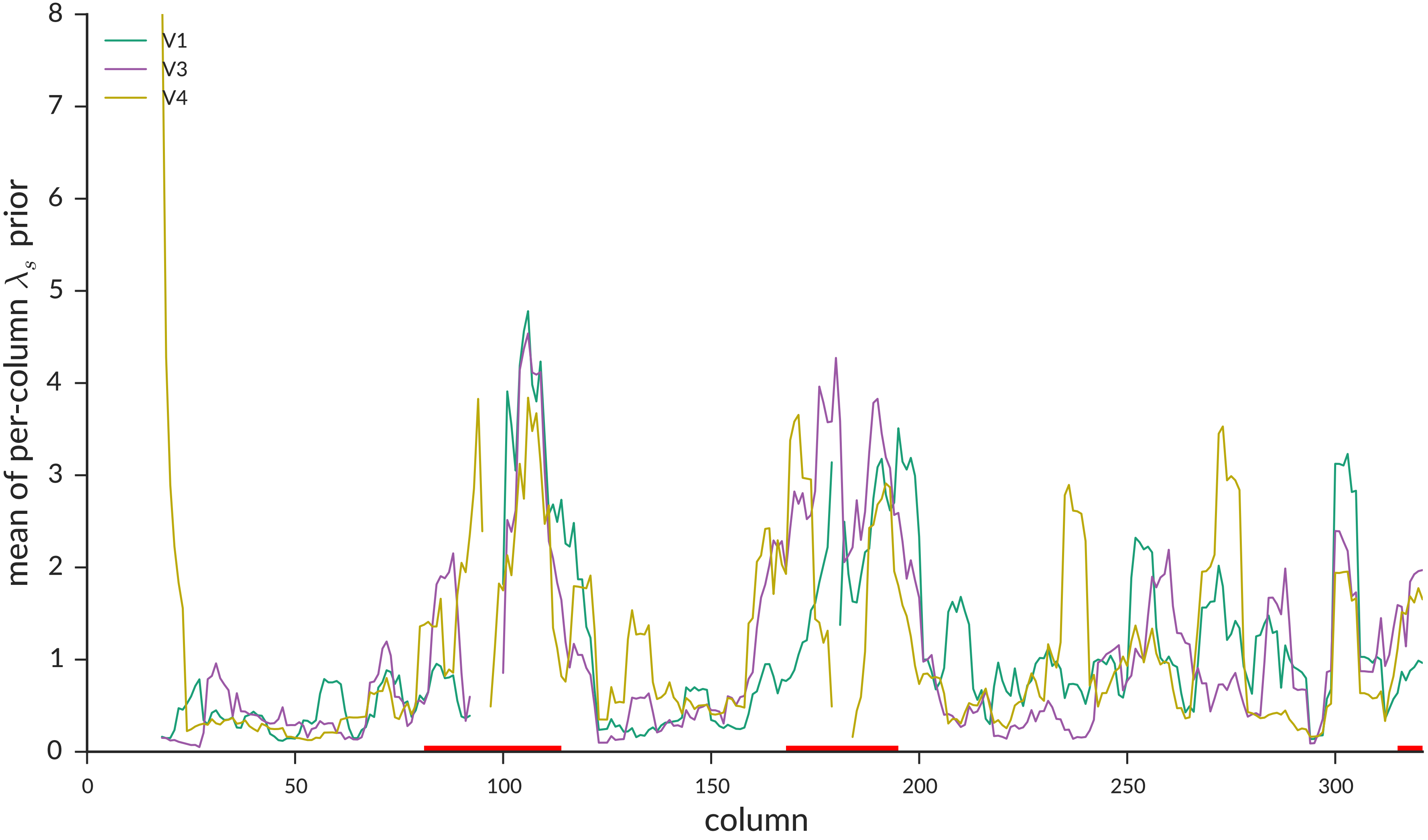
Points = rates; triangles = 95% CI of fit prior

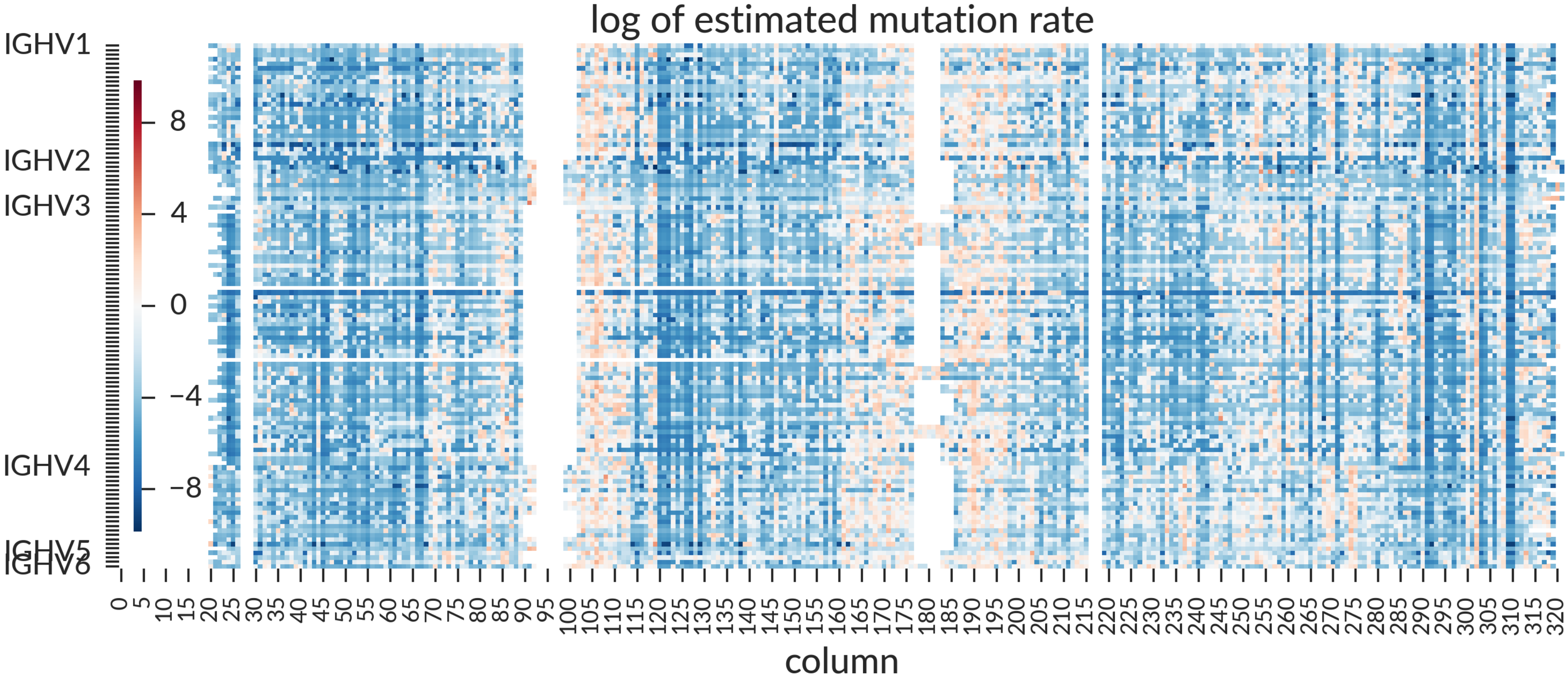


Per-site mutation rate



Per-site mutation rate (smoothed)

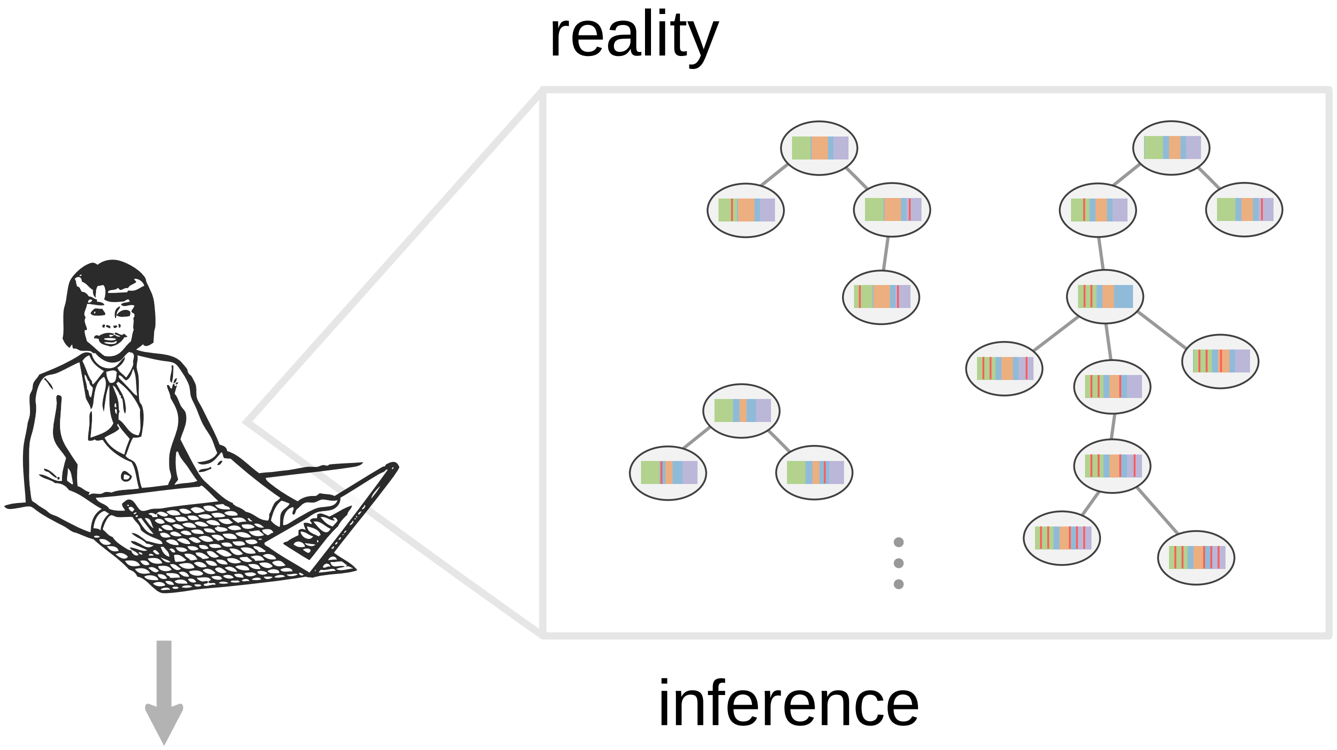



log of estimated mutation rate
IGHV1
IGHV2
IGHV3
IGHV4
IGHV5
IGHV6
8
4
0
−4
−8
0
5
10
15
20
25
30
35
40
45
50
55
60
65
70
75
80
85
90
95
100
105
110
115
120
125
130
135
140
145
150
155
160
165
170
175
180
185
190
195
200
205
210
215
220
225
230
235
240
245
250
255
260
265
270
275
280
285
290
295
300
305
310
315
320
column

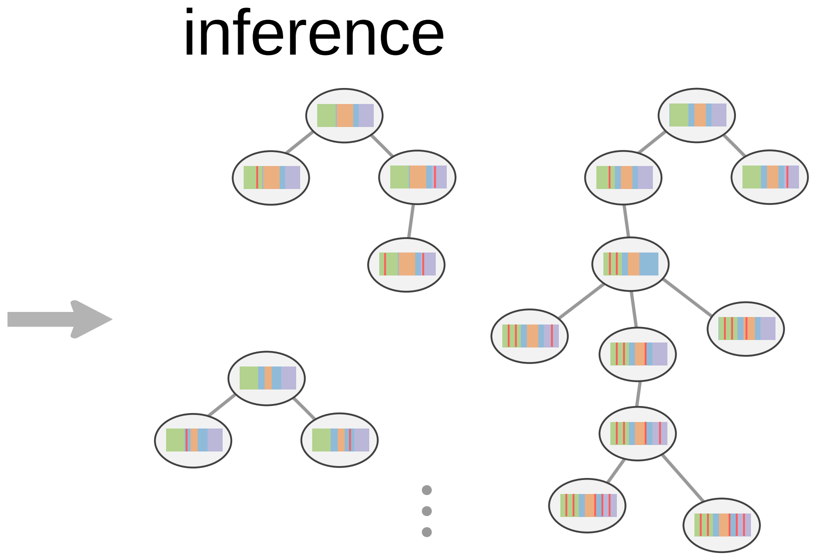
Likelihood-based phylogenetics for B cell receptor sequences

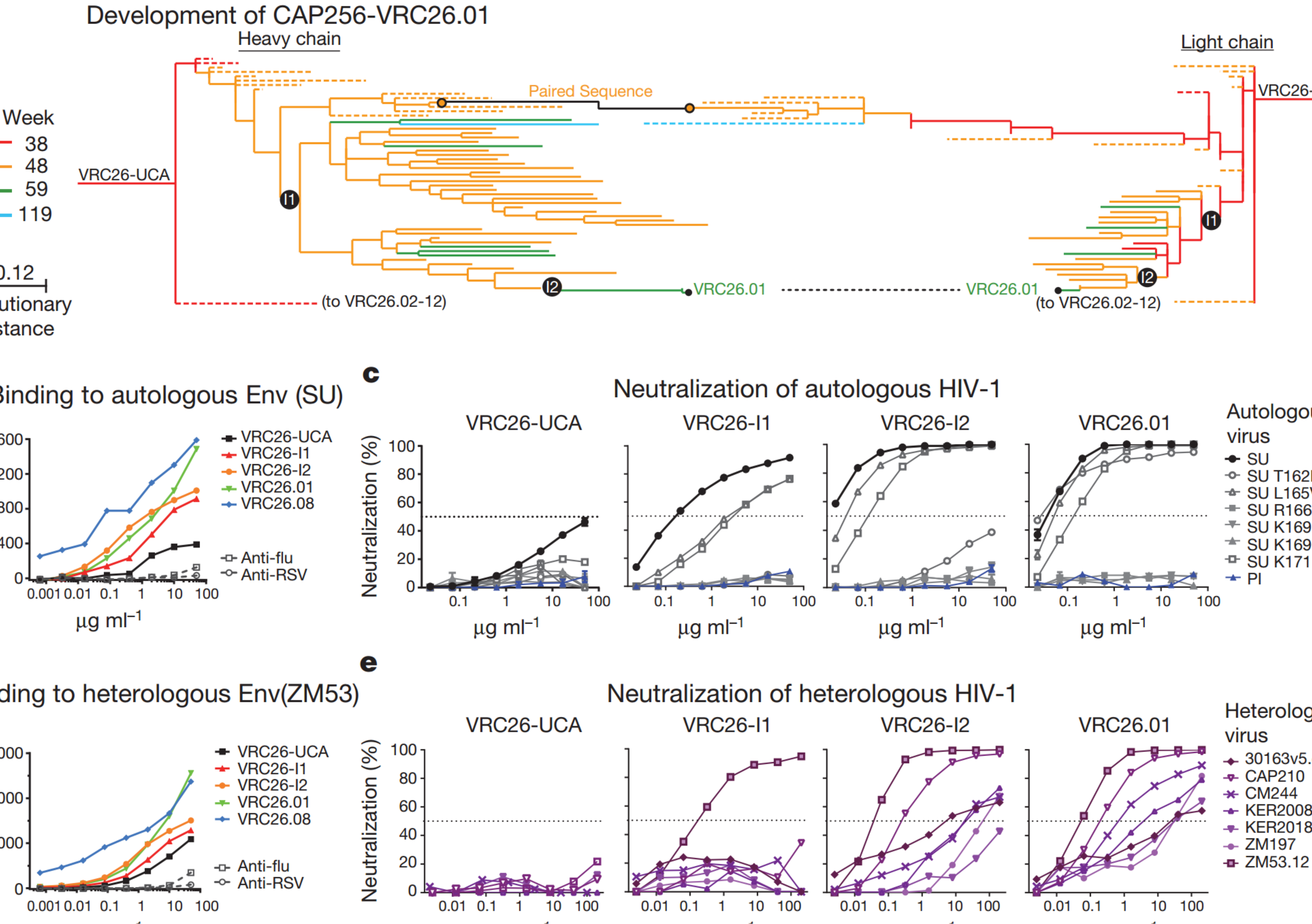
- Many people use likelihood-based phylogenetics in their analysis, but with models that are identical across sites
- Substitution is manifestly *not* identical across sites
- One could work to do phylogenetics with context-sensitive models (hard!) or infer per-site parameters (need regularization!)
- Need to build software that can build trees with these models
- Sampled ancestors also a challenge, but this can be handled in a Bayesian or penalized likelihood framework (in progress).

4. Find BCR ancestral sequences



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





Development of CAP256-VRC26.01
Heavy chain
Light chain
Paired Sequence
Week
38
48
59
119
VRC26-UCA
I1
I2
VRC26.01
(to VRC26.02-12)
VRC26-
0.12
utionary
stance
Binding to autologous Env (SU)
VRC26-UCA
VRC26-I1
VRC26-I2
VRC26.01
VRC26.08
Anti-flu
Anti-RSV
µg ml−1
c
Neutralization of autologous HIV-1
VRC26-UCA
VRC26-I1
VRC26-I2
VRC26.01
Neutralization (%)
µg ml−1
Autologou
virus
SU
SU T162
SU L165
SU R166
SU K169
SU K169
SU K171
PI
e
ding to heterologous Env(ZM53)
Neutralization of heterologous HIV-1
Heterolog
virus
30163v5.
CAP210
CM244
KER2008
KER2018
ZM197
ZM53.12

μg ml^{-1} μg ml^{-1} μg ml^{-1} μg ml^{-1} μg ml^{-1}

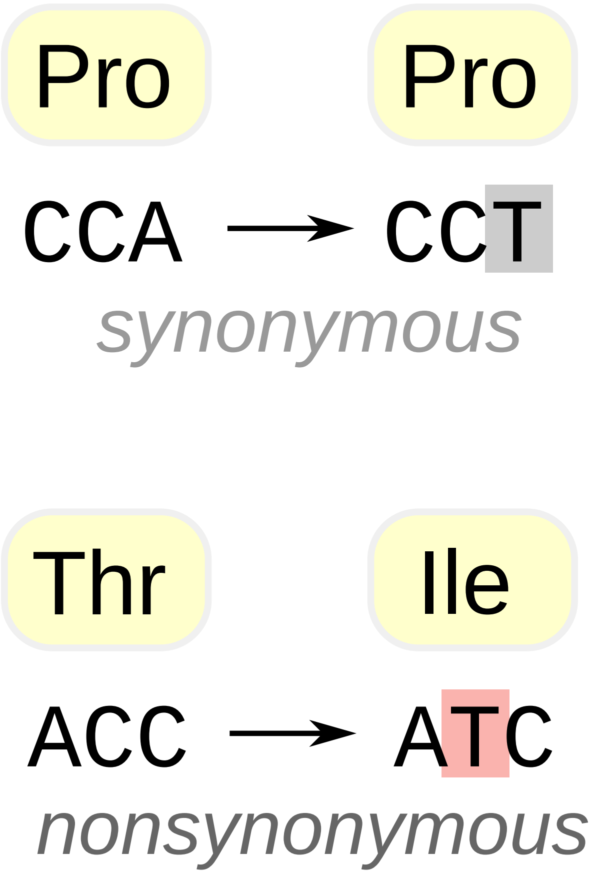
Likelihood-based ancestral sequence reconstruction

Currently being done with identical-across-sites models.

Once we have per-site models, it will be Ś œ Ő .

5. Selection inference for BCRs

For selection



Pro Pro

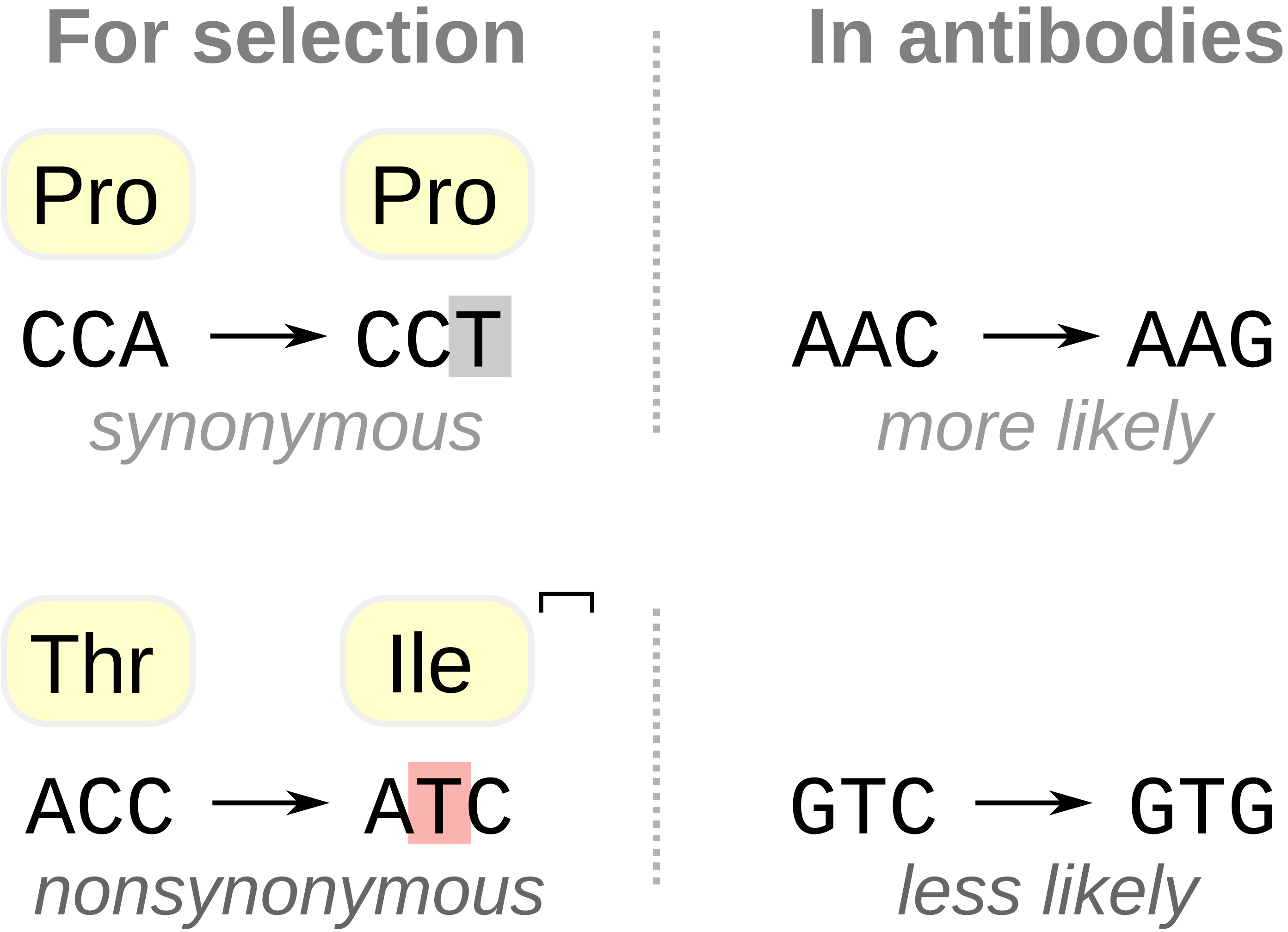
CCA → CCT

synonymous

Thr Ile

ACC → ATC

nonsynonymous



For selection

Pro Pro

CCA → CCT

synonymous

Thr Ile

ACC → ATC

nonsynonymous

In antibodies

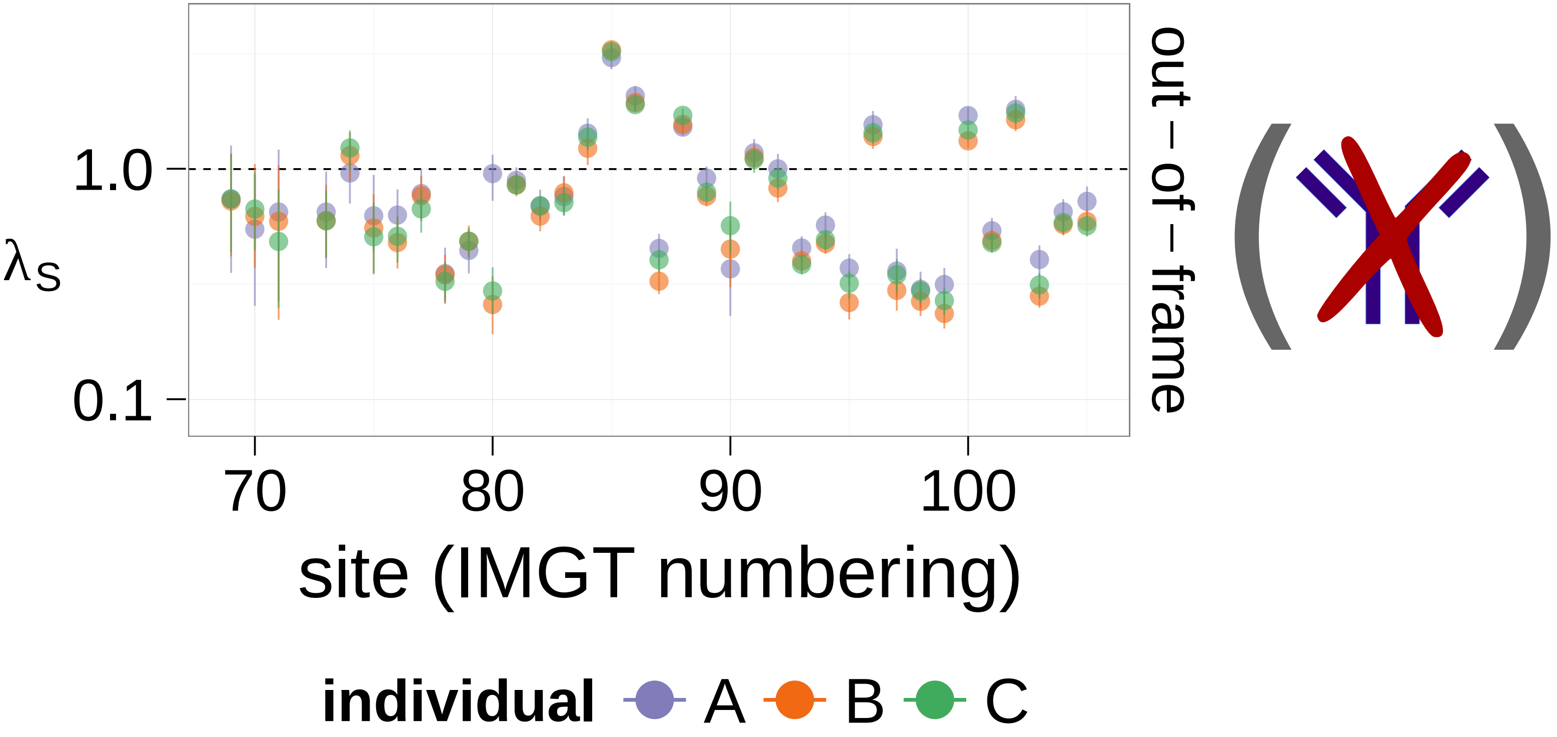
AAC → AAG

more likely

GTC → GTG

less likely

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



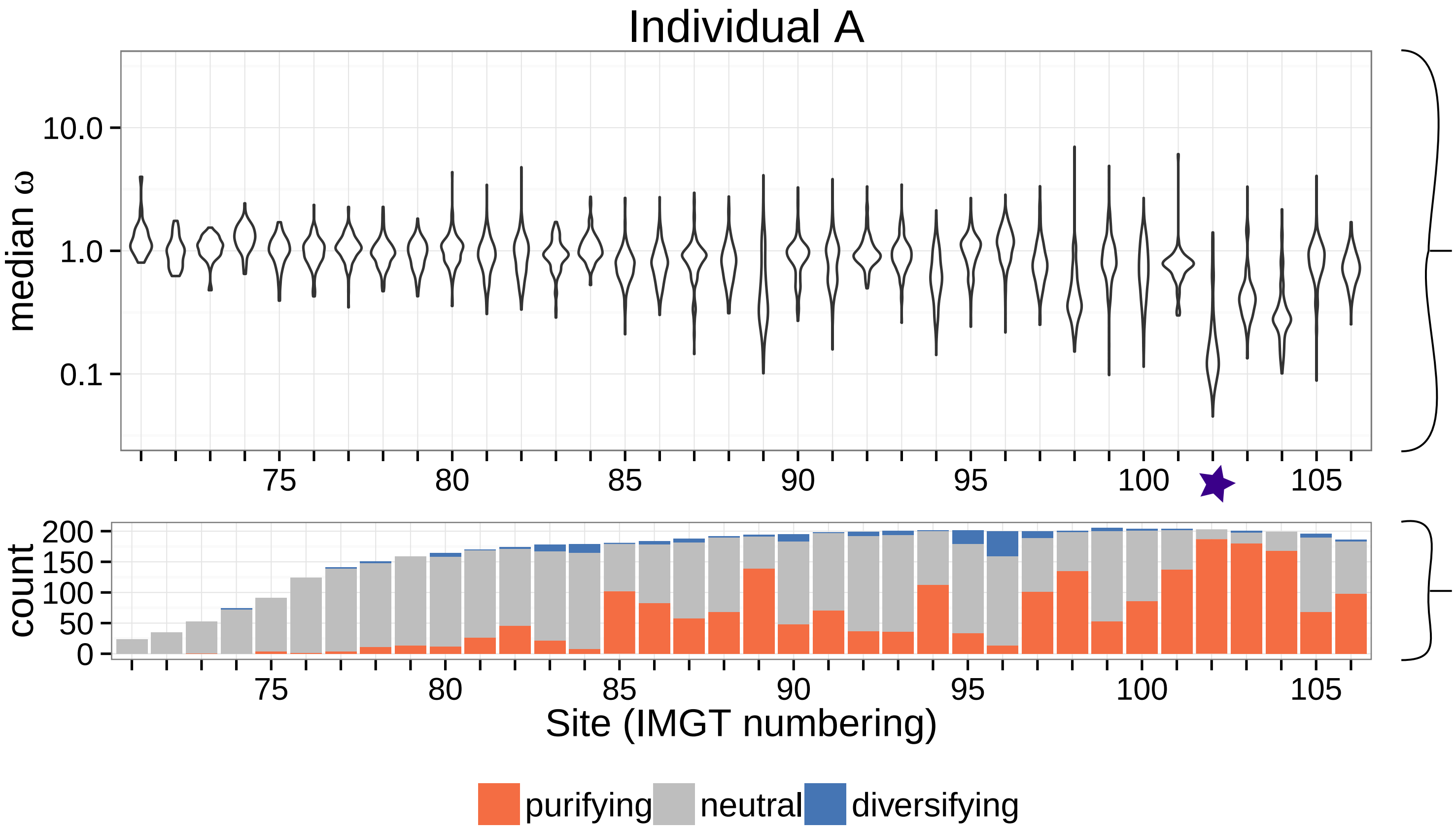
Out-of-frame reads can be used to infer neutral mutation rate.

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

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

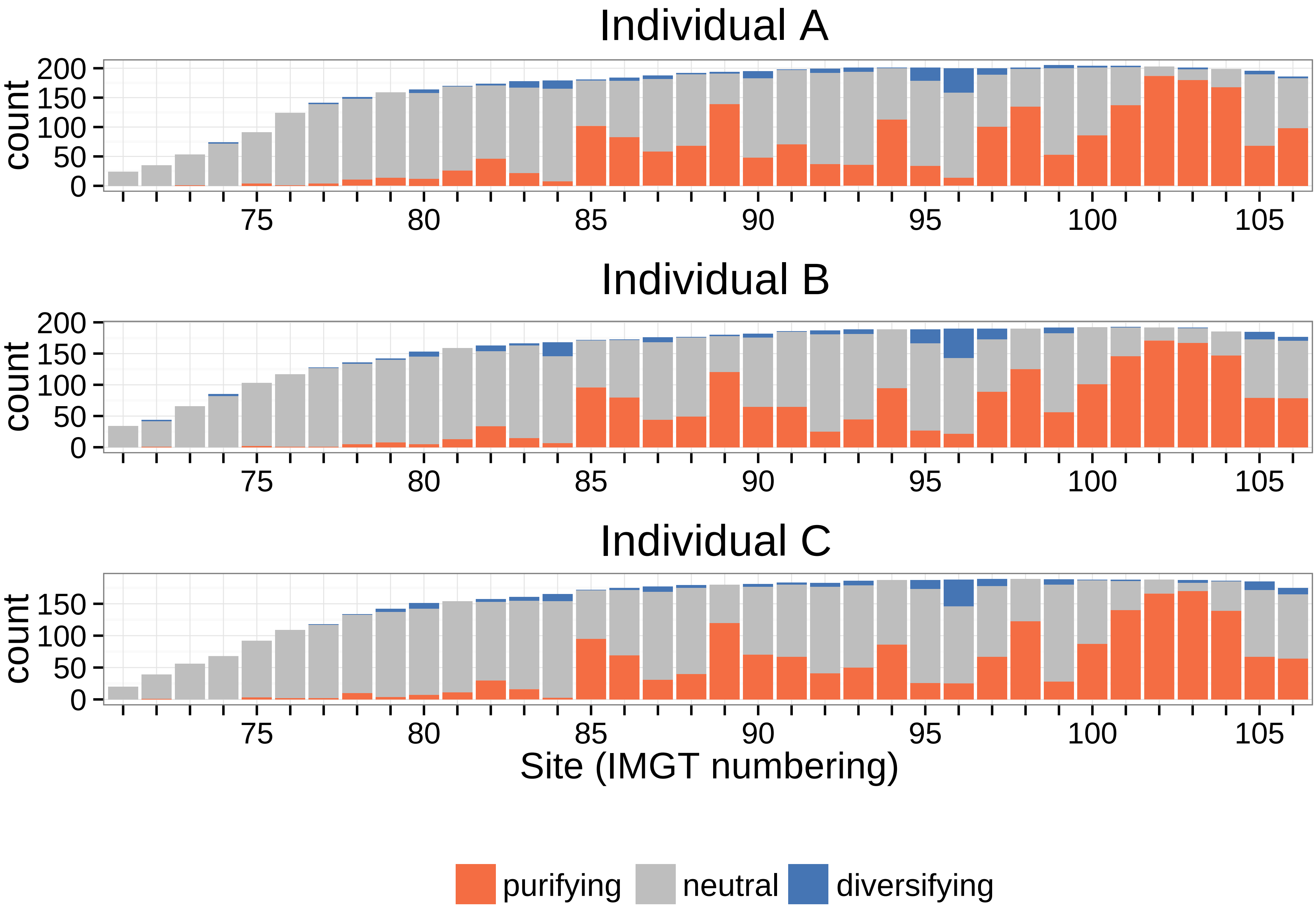
Overall IGHV selection map

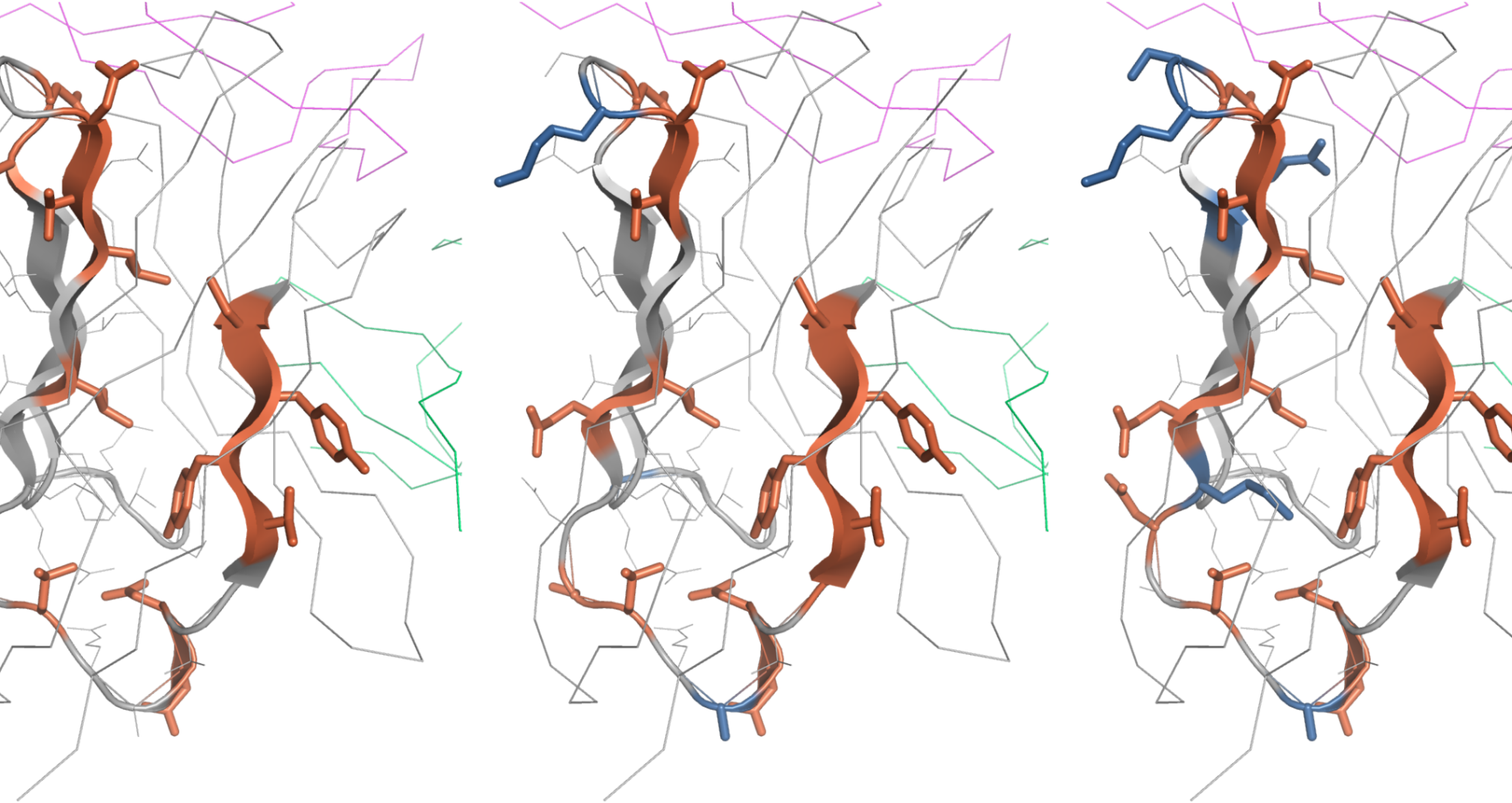


Distribution of median estimates of ω

Distribution of classifications across IGHV genes

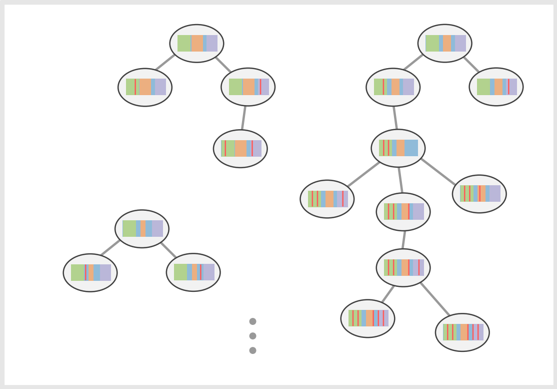
Similar across individuals





Likelihood-based selection inference for BCRs

- Aggregate selection analysis: Yaari, Uduman & Kleinstein (2012). *Nucleic Acids Research*
- Amino acid preferences: Elhanati, Sethna, Marcou, Callan, Mora & Walczak (2015). *Phil Trans Royal Soc B*
- Per-site analysis: McCoy, Bedford, Minin, Bradley, Robins & M. (2015). *Phil Trans Royal Soc B*



$$\sim f(\text{health}, \text{genetics}, \text{age}, \dots)$$

Make this relationship explicit by developing probabilistic models with priors in terms of covariates.

We are approaching this from an abstract statistical perspective rather than via mechanistic models.