Diversity, selection and specificity of immune receptors

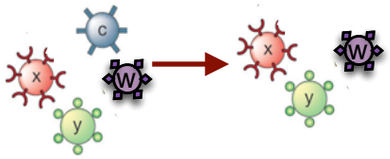
Thierry Mora
Laboratoire de physique statistique
École normale supérieure, Paris
& CNRS

ENS Paris
Rhys Adams
Yuval Elhanati
Quentin Marcou
Aleksandra Walczak

Princeton
Curt Callan
Anand Murugan
Zachary Sethna

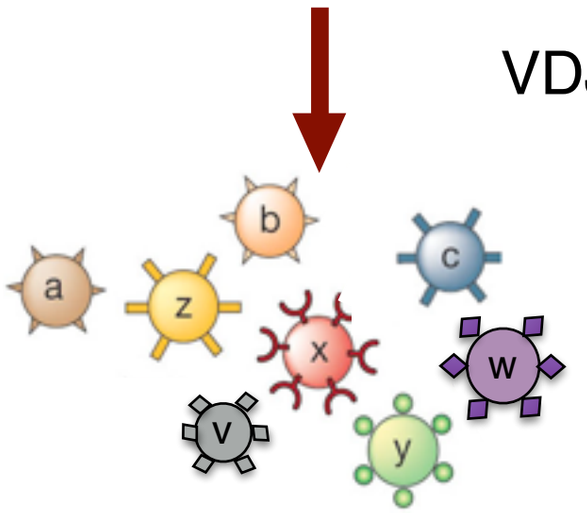
Cold Spring Harbor
Justin Kinney

Repertoire evolution

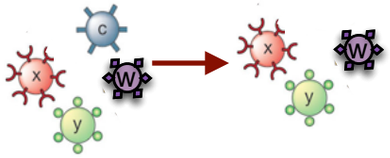


RECEPTOR GENERATION

VDJ recombination

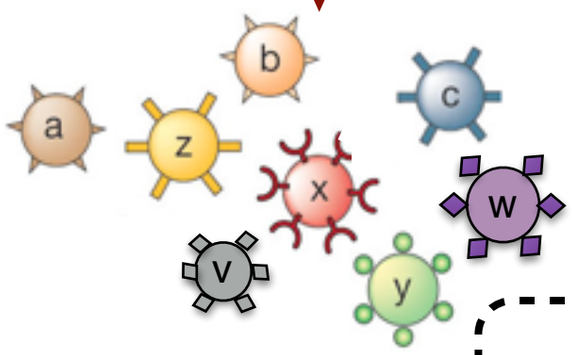


Repertoire evolution

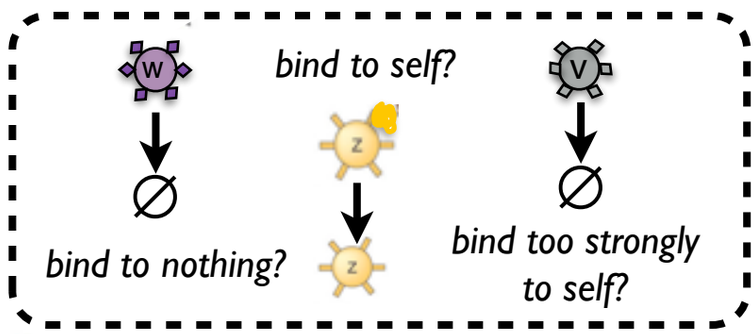


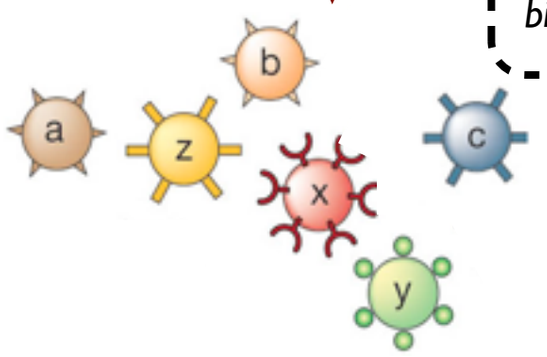
RECEPTOR GENERATION

VDJ recombination

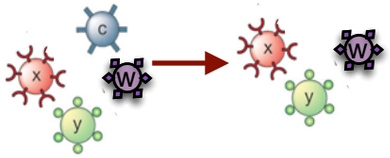


THYMIC SELECTION



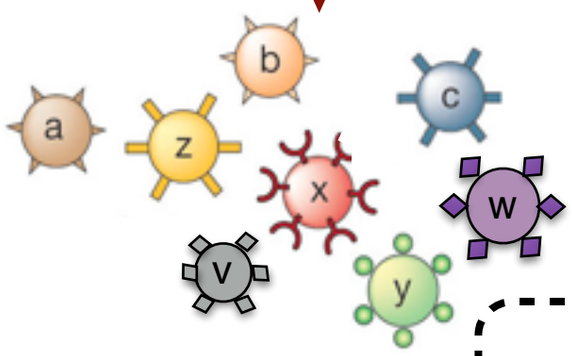


Repertoire evolution

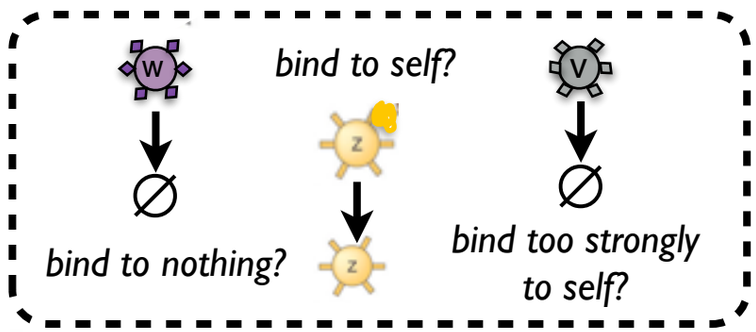


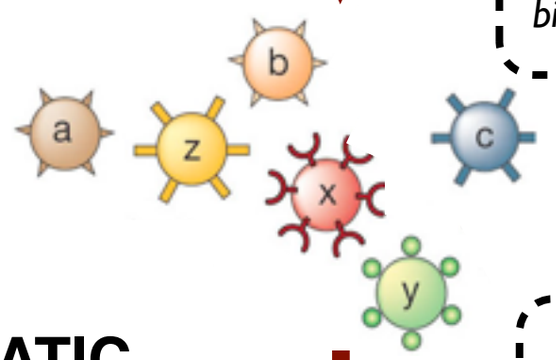
RECEPTOR GENERATION

VDJ recombination

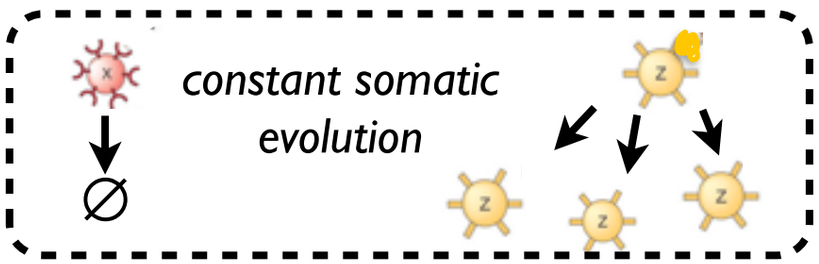


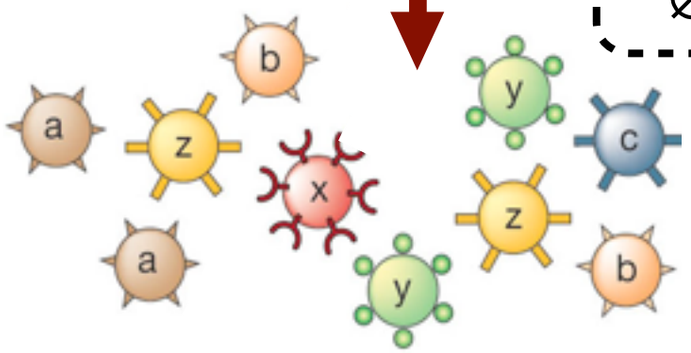
THYMIC SELECTION



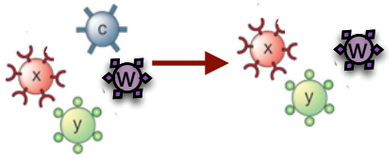


SOMATIC SELECTION



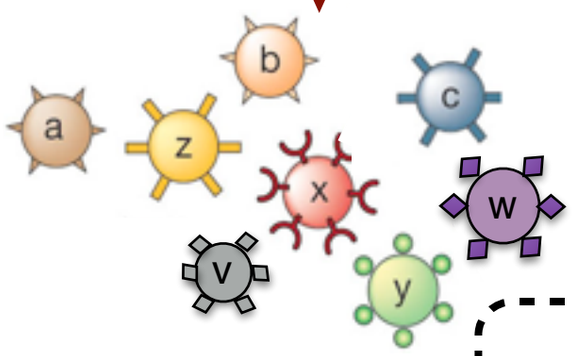


Repertoire evolution

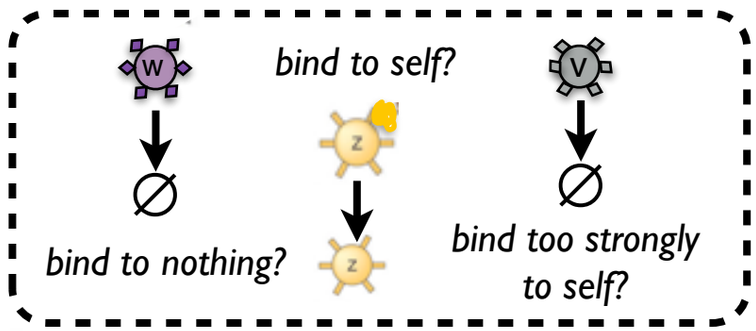


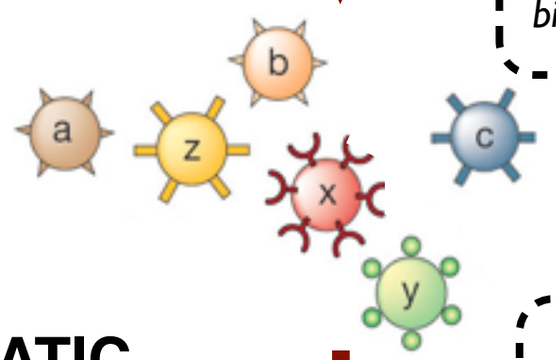
RECEPTOR GENERATION

VDJ recombination

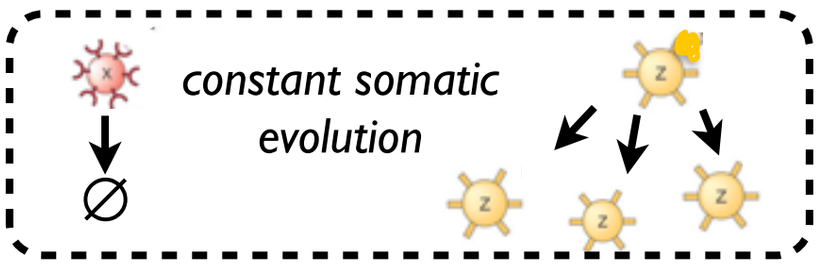


THYMIC SELECTION

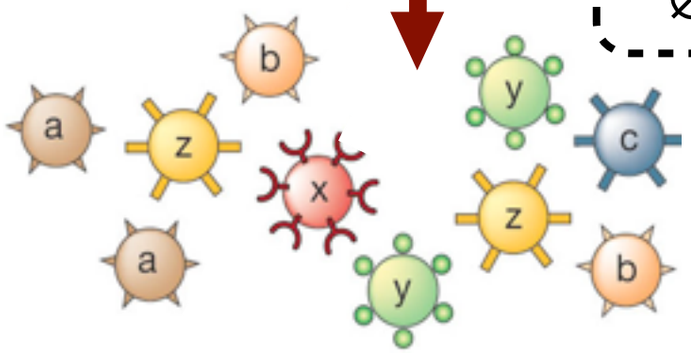




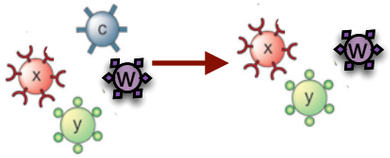
SOMATIC SELECTION



+ Somatic Hypermutations

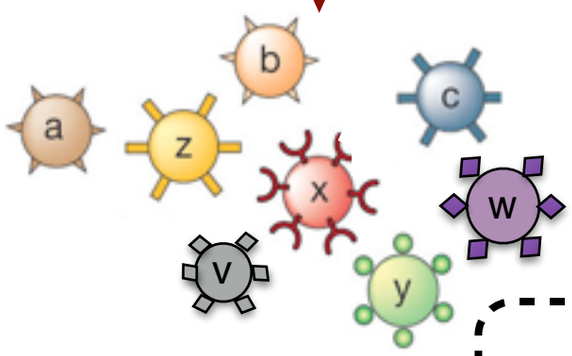


Repertoire evolution

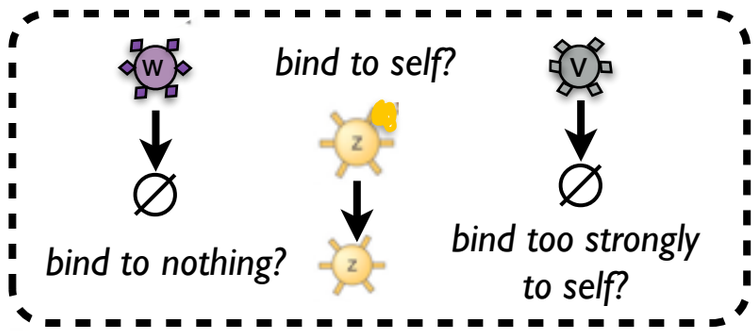


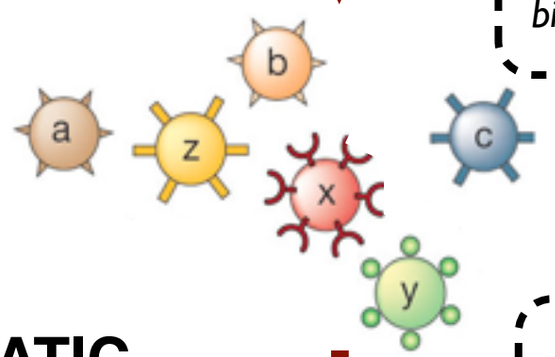
RECEPTOR GENERATION

VDJ recombination

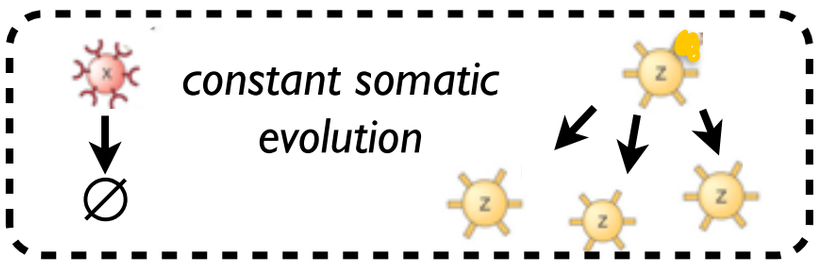


THYMIC SELECTION

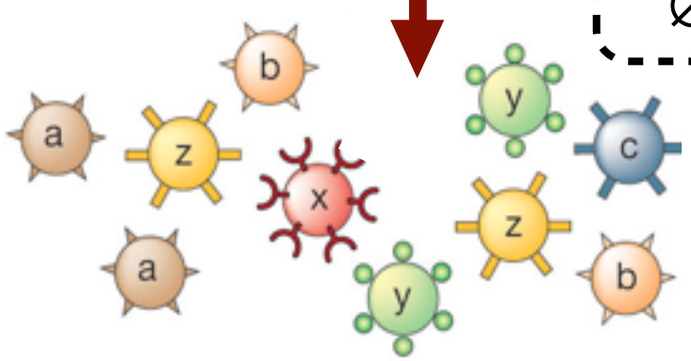




SOMATIC SELECTION

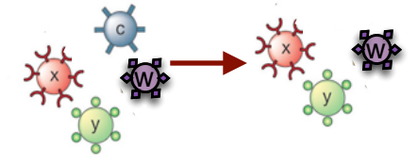


+ Somatic Hypermutations



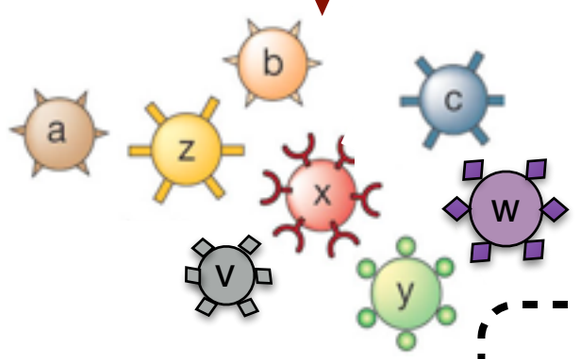
SPECIFICITY

Repertoire evolution

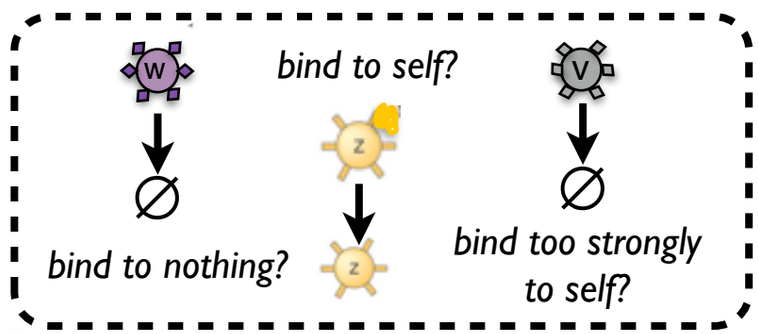


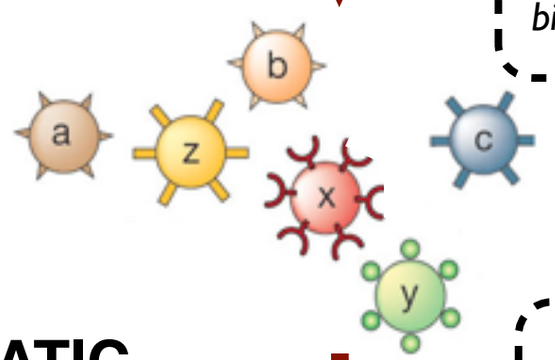
RECEPTOR GENERATION

VDJ recombination

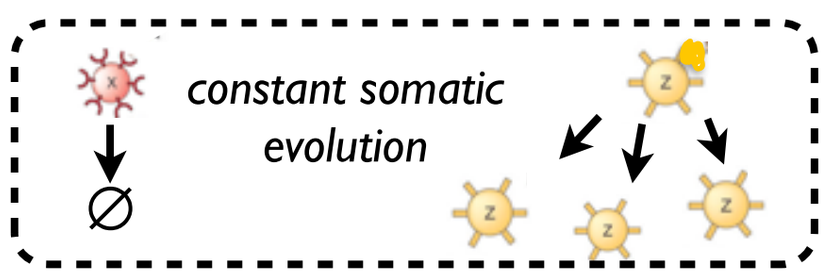


THYMIC SELECTION

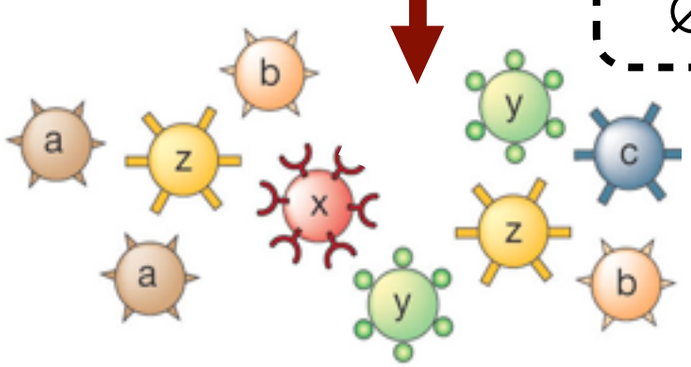




SOMATIC SELECTION

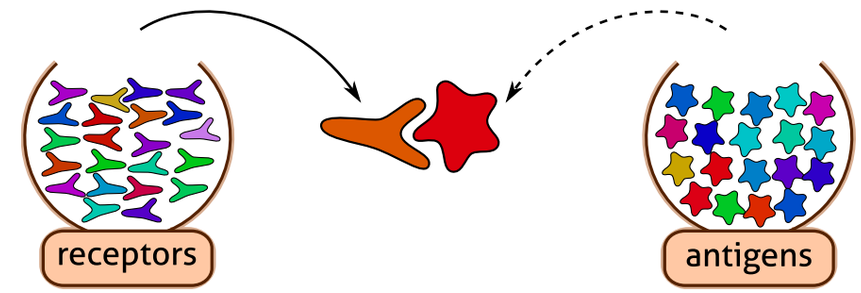


+ Somatic Hypermutations



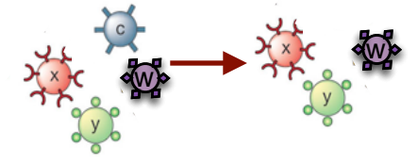
SPECIFICITY

Optimal repertoires



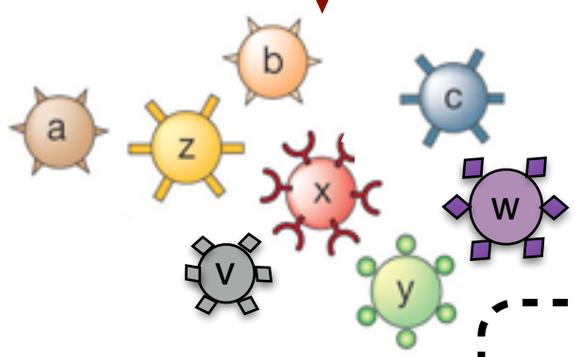
Mayer Balasubramanian Mora Walczak PNAS 2015

Repertoire evolution

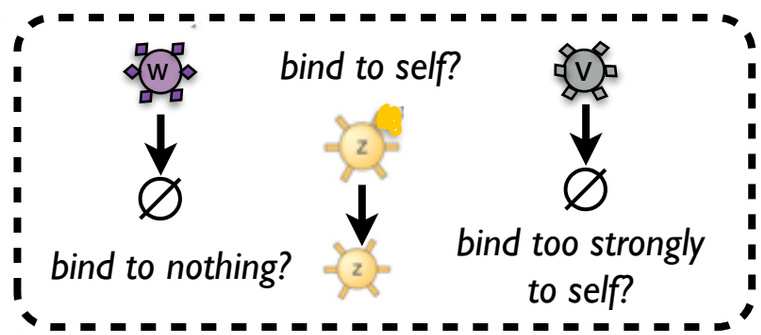


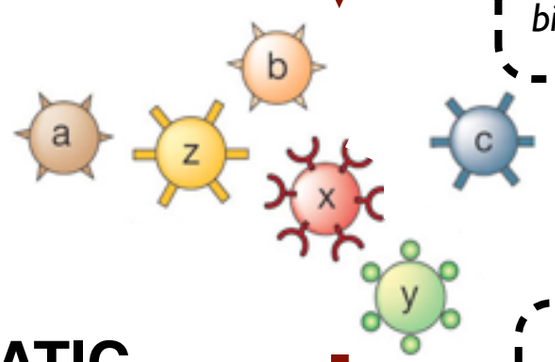
RECEPTOR GENERATION

VDJ recombination



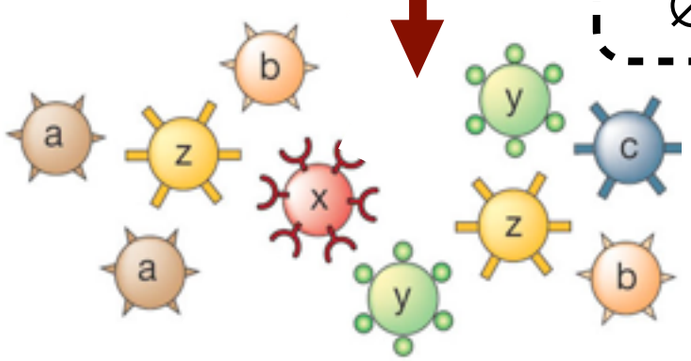
THYMIC SELECTION



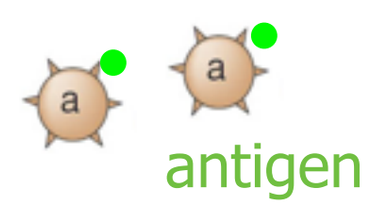


SOMATIC SELECTION

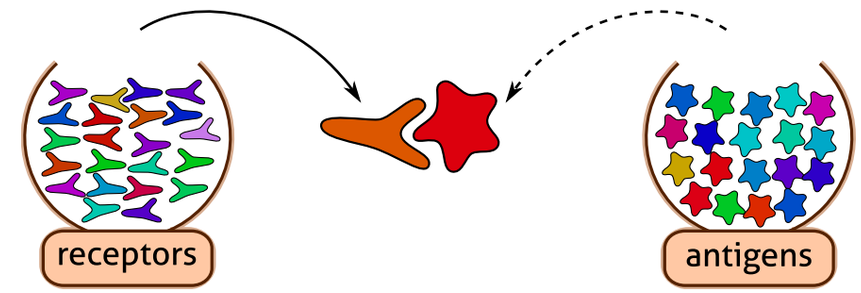
+ Somatic Hypermutations



SPECIFICITY

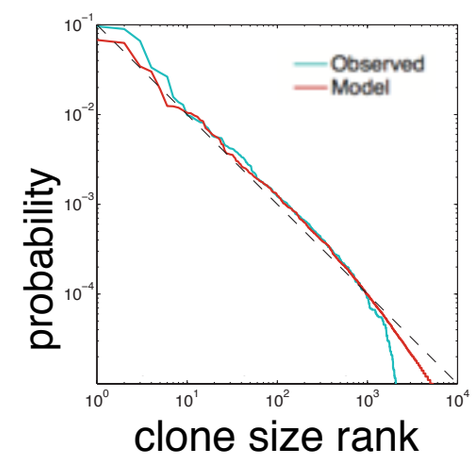


Optimal repertoires



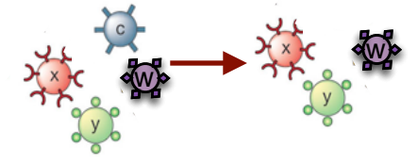
Mayer Balasubramanian Mora Walczak PNAS 2015

Clone-size distribution



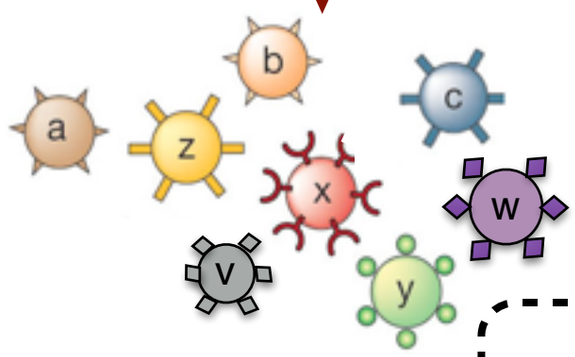
Desponds Mora Walczak PNAS 2016

Repertoire evolution

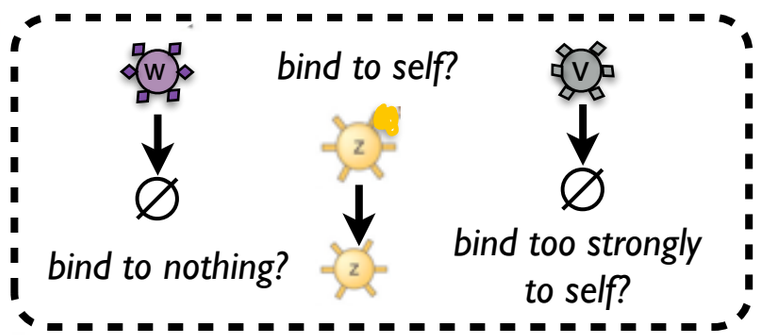


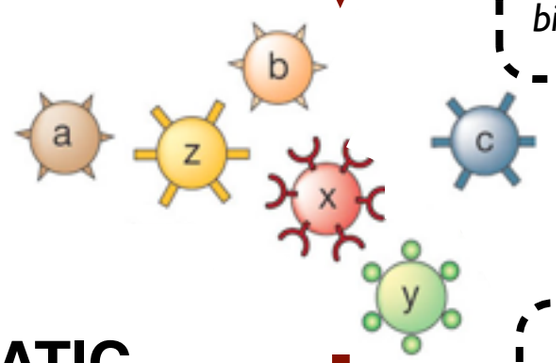
RECEPTOR GENERATION

VDJ recombination



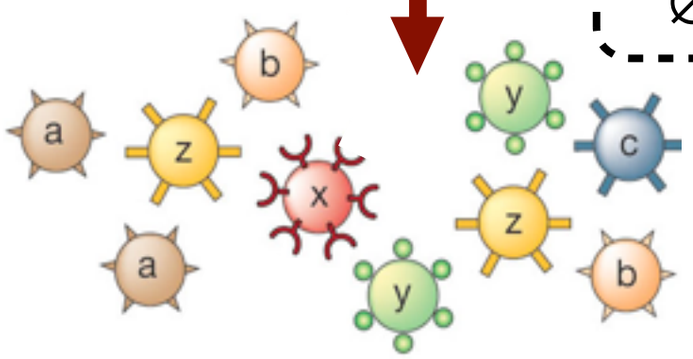
THYMIC SELECTION





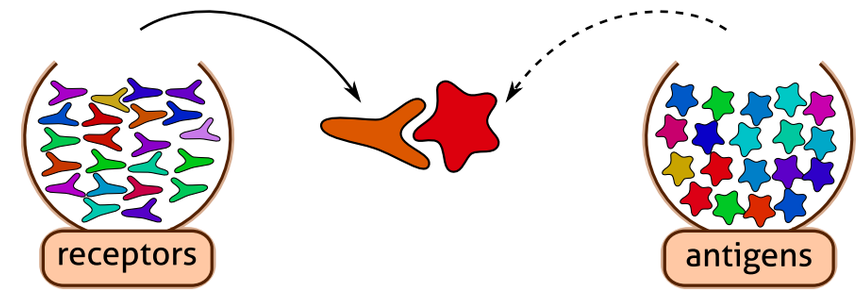
SOMATIC SELECTION

+ Somatic Hypermutations



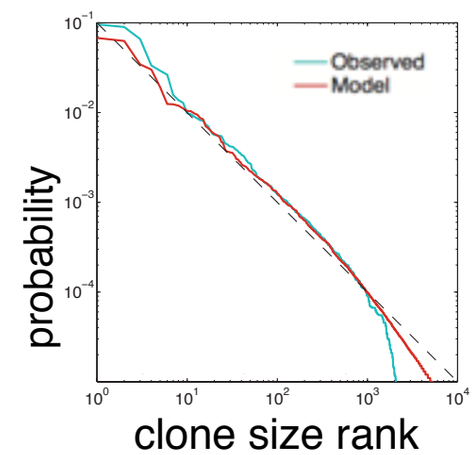
SPECIFICITY

Optimal repertoires



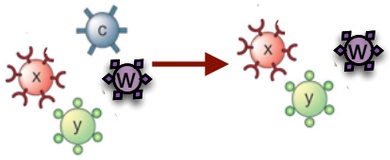
Mayer Balasubramanian Mora Walczak PNAS 2015

Clone-size distribution



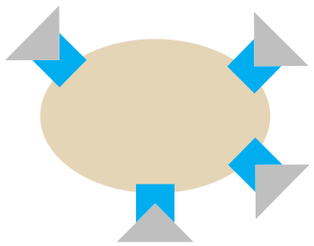
Desponds Mora Walczak PNAS 2016

Sequence-affinity landscape



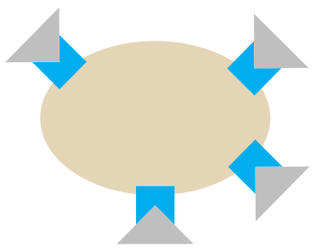
antibody / TCR sequence ⟷ binding affinity

Tite-seq



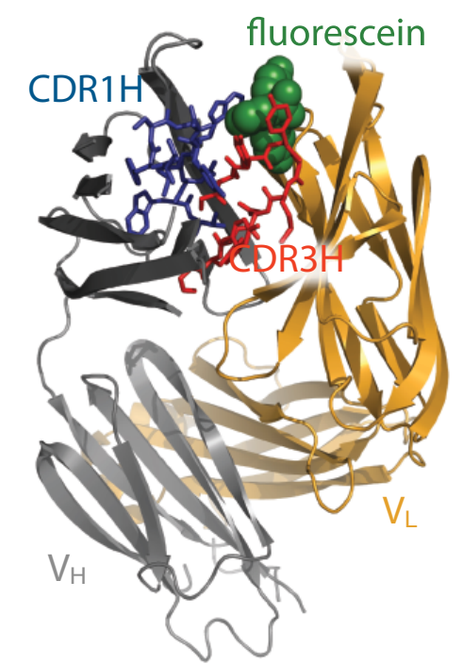
yeast display + FACSort + sequence

Tite-seq

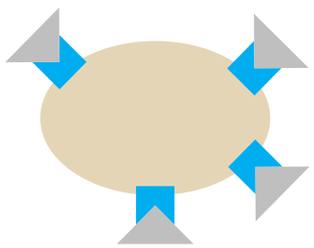


yeast display + FACSort + sequence

- proof of principle: antigen = fluorescein
  antibody = mutagenized scFv

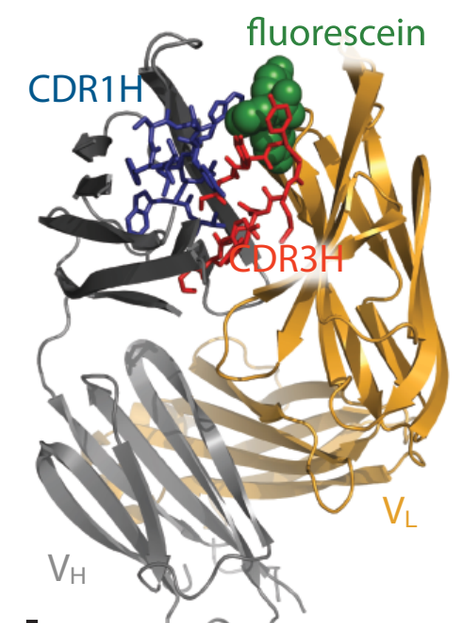


Tite-seq

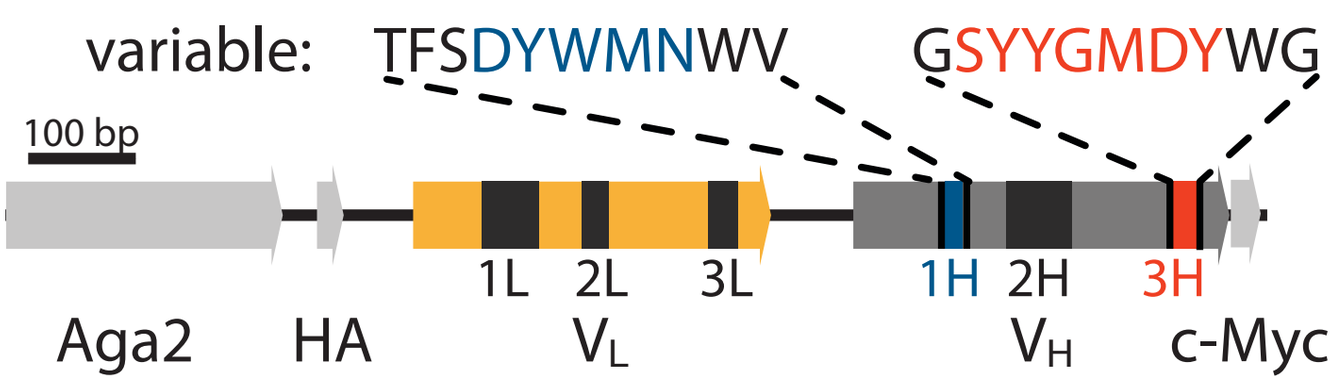


yeast display + FACSort + sequence

- proof of principle: antigen = fluorescein
  antibody = mutagenized scFv

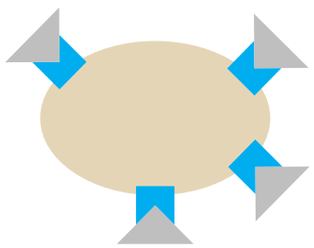


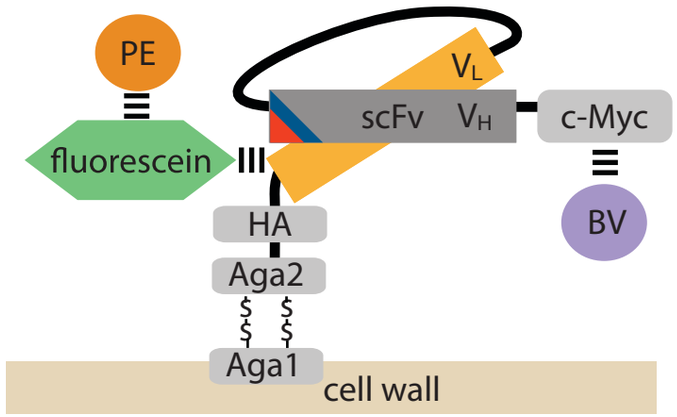
- exhaustive single mutations in CDR1H and CDR3H



variation	1H	3H
1 codon	600	600
2 codon	1100	1100
3 codon	150	150

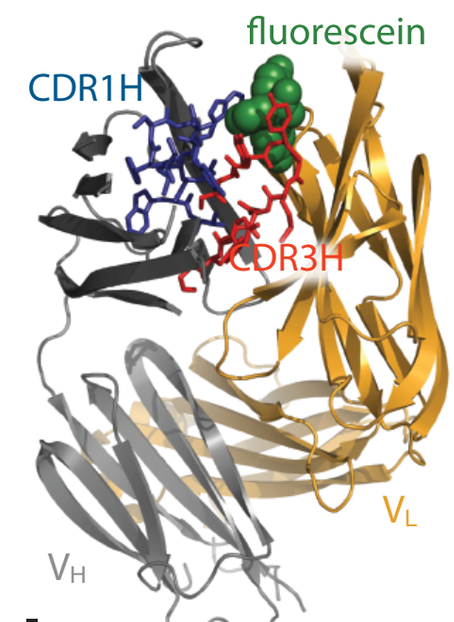
Tite-seq



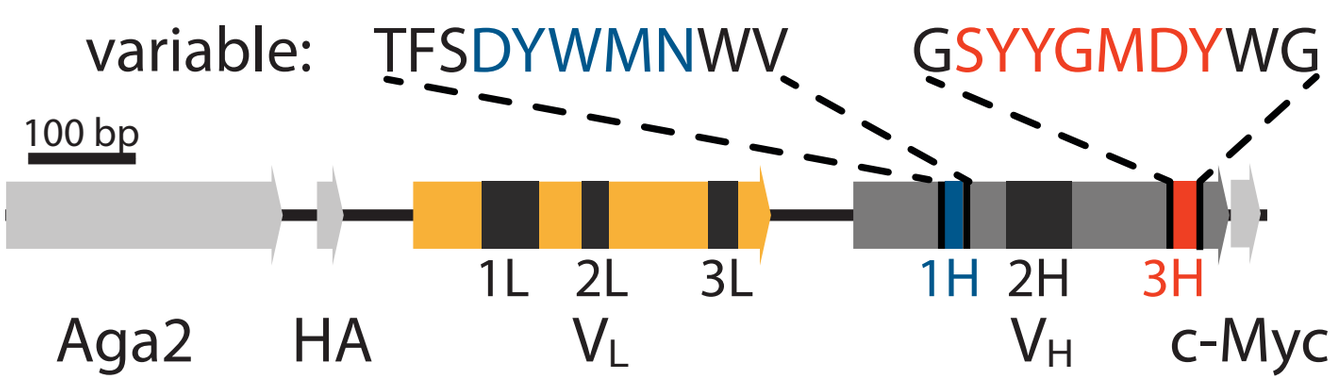


yeast display + FACSort + sequence

- proof of principle: antigen = fluorescein
  antibody = mutagenized scFv

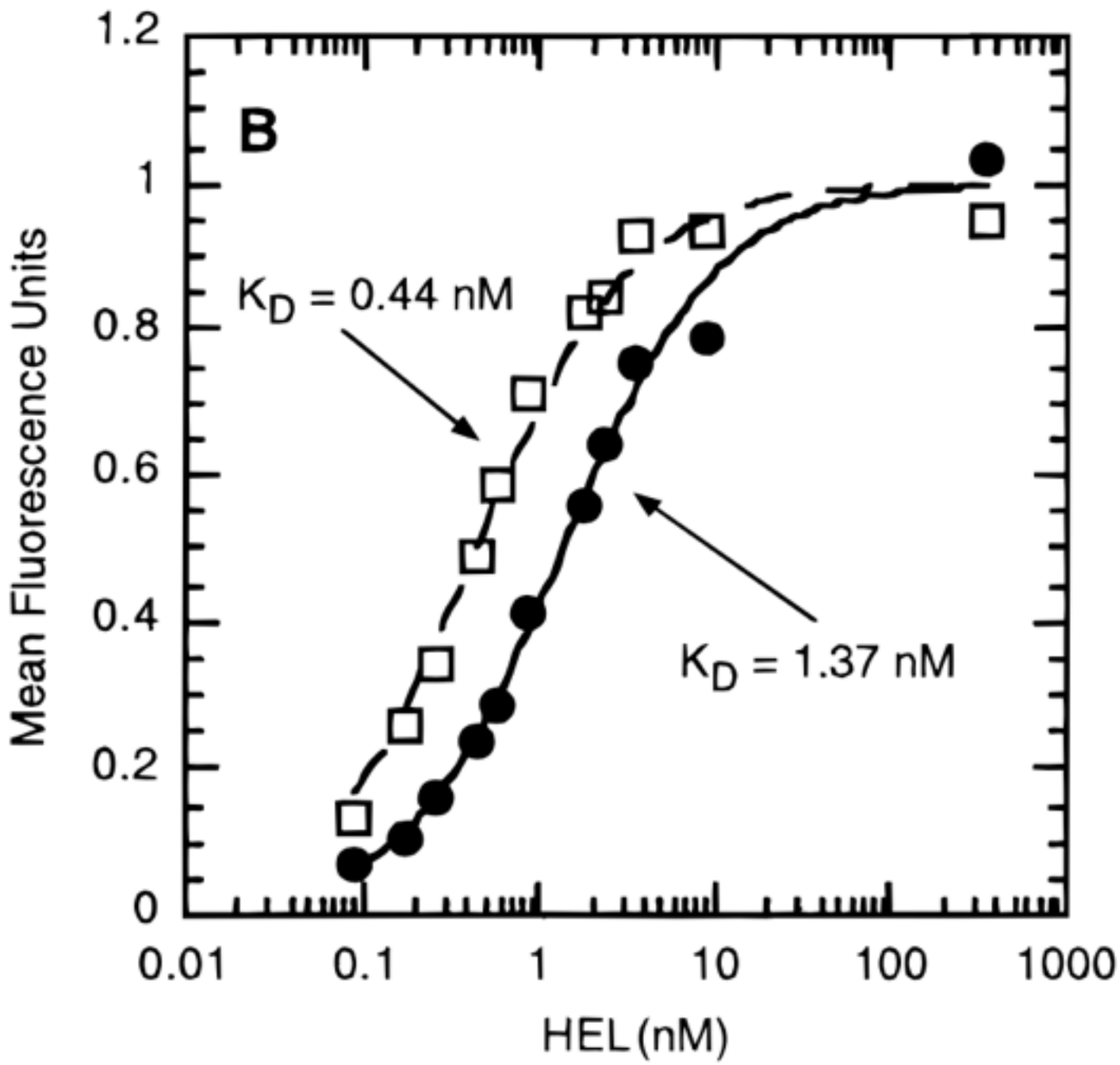


- exhaustive single mutations in CDR1H and CDR3H

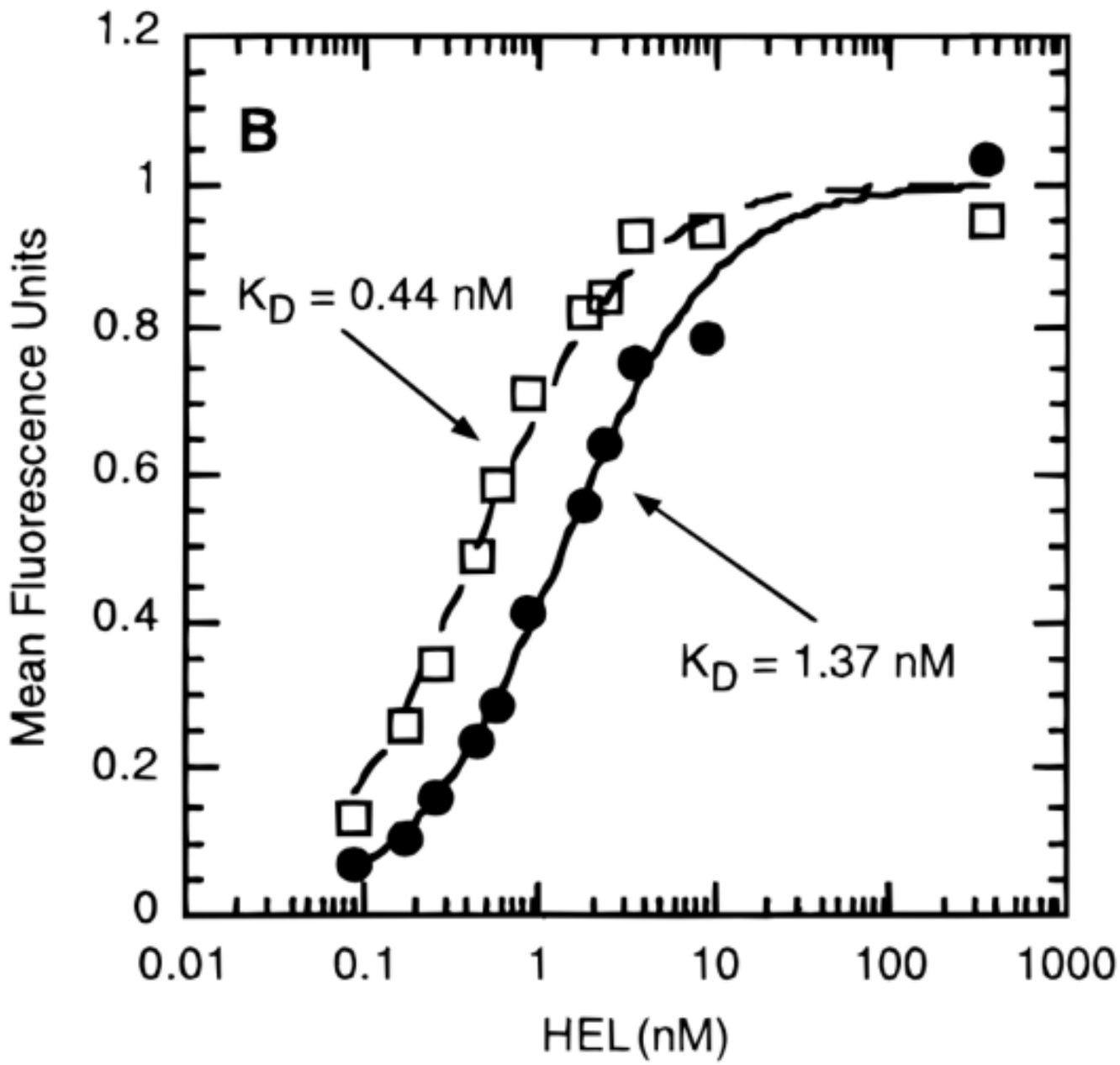


variation	1H	3H
1 codon	600	600
2 codon	1100	1100
3 codon	150	150

Titration by flow cytometry

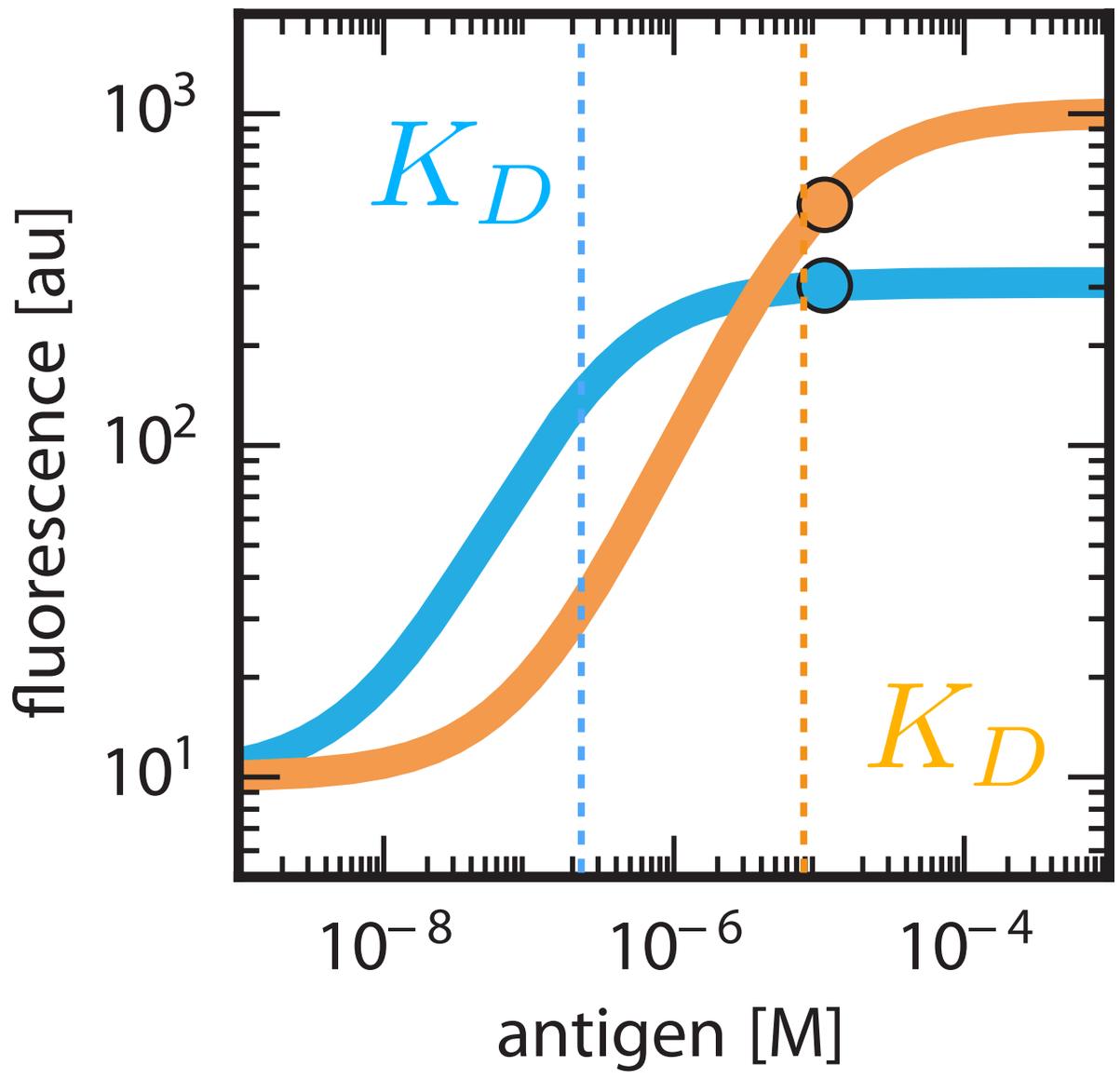


Titration by flow cytometry



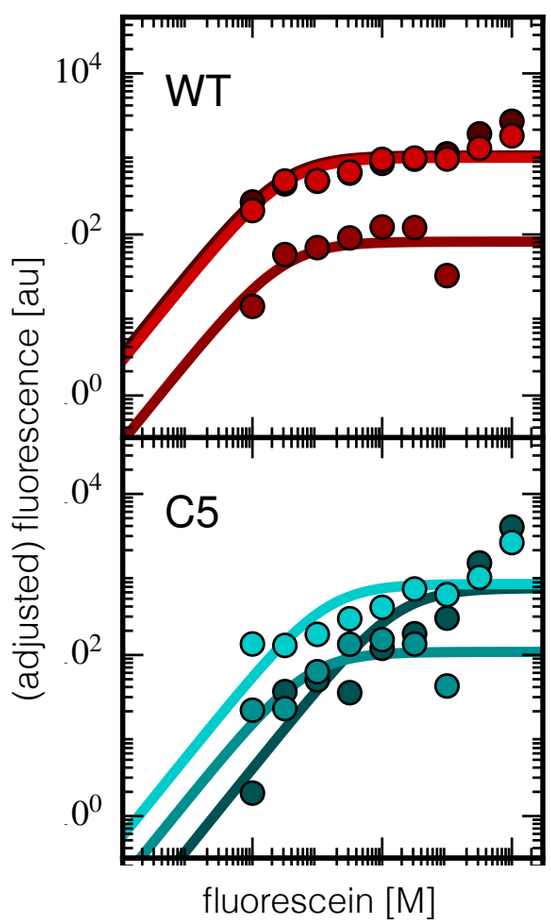
VanAntwerp and Wittrup Biotechnol. Prog. 2000

Titration by flow cytometry

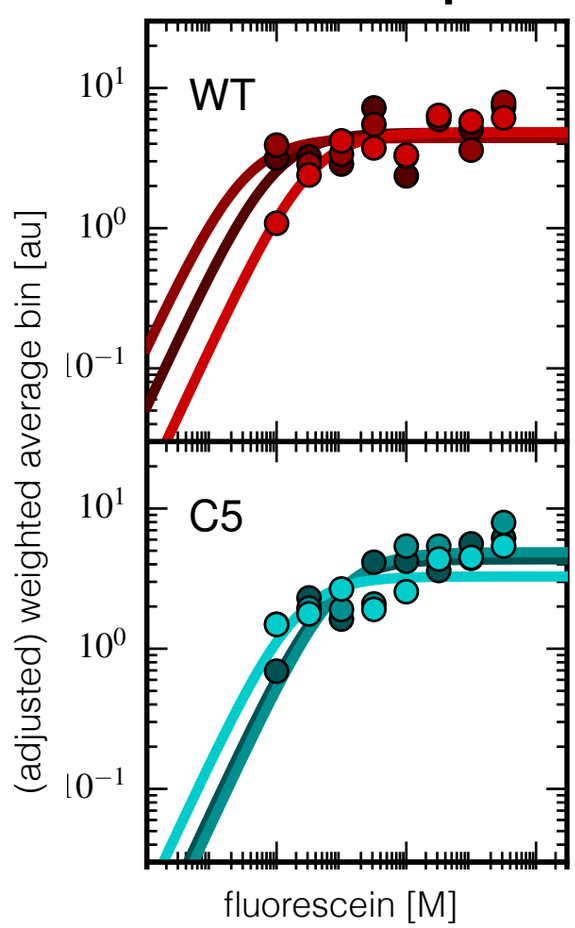


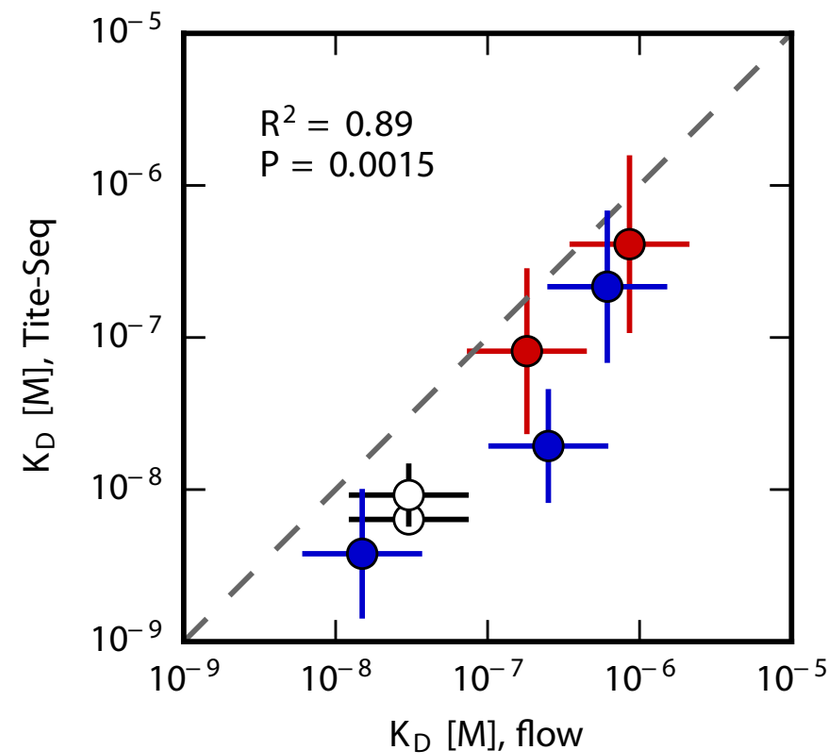
Comparison to direct fluorescence

fluorescence



tite-seq

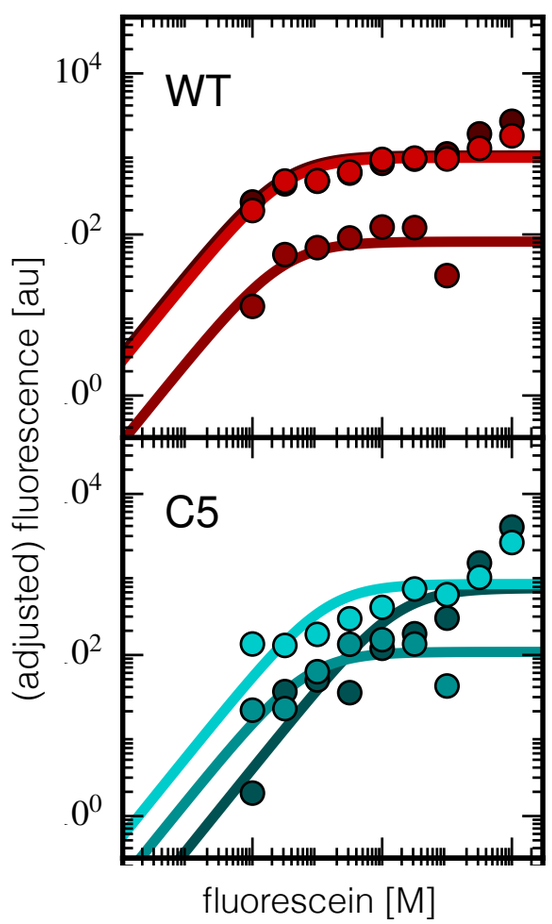




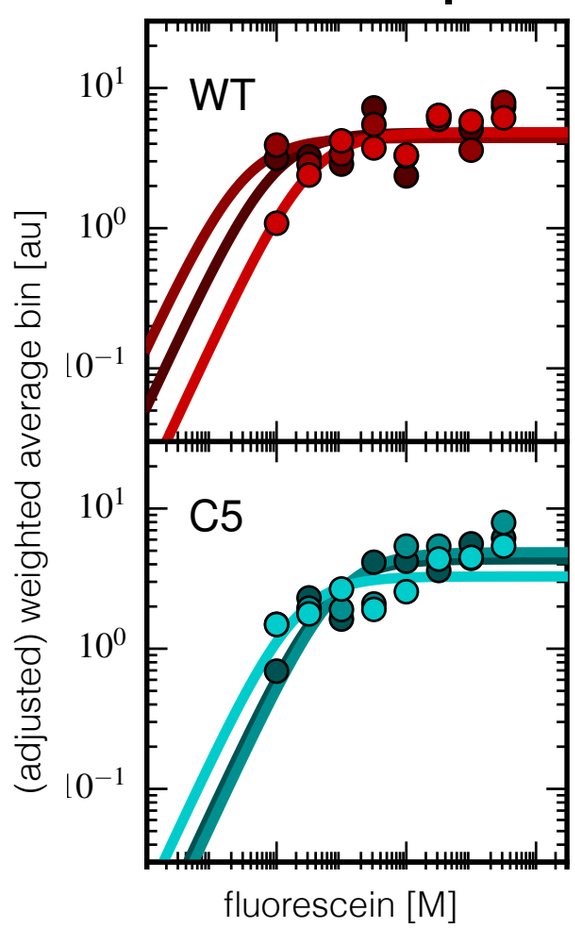
wide range of affinities

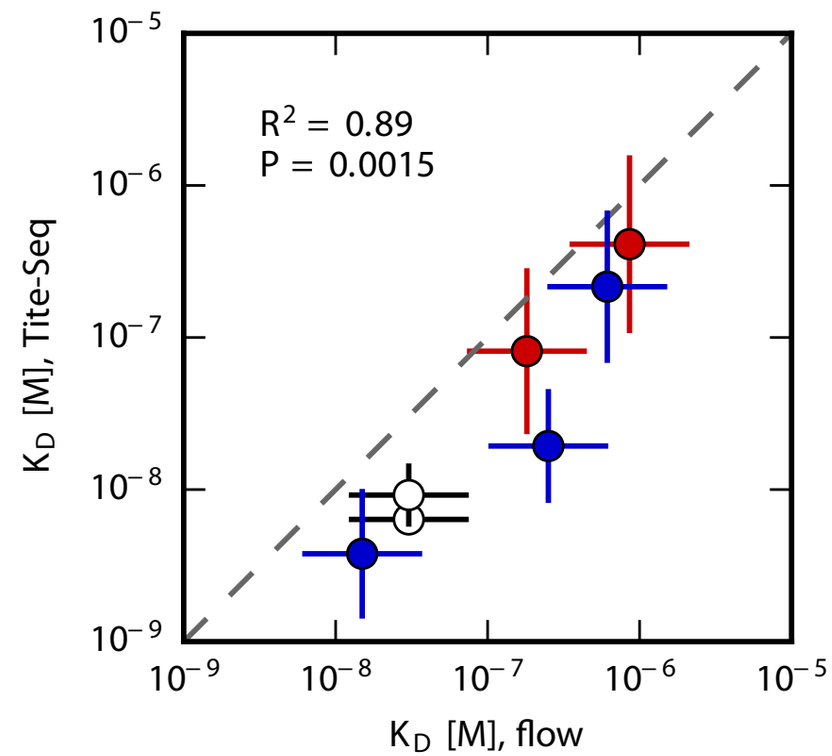
Comparison to direct fluorescence

fluorescence



tite-seq

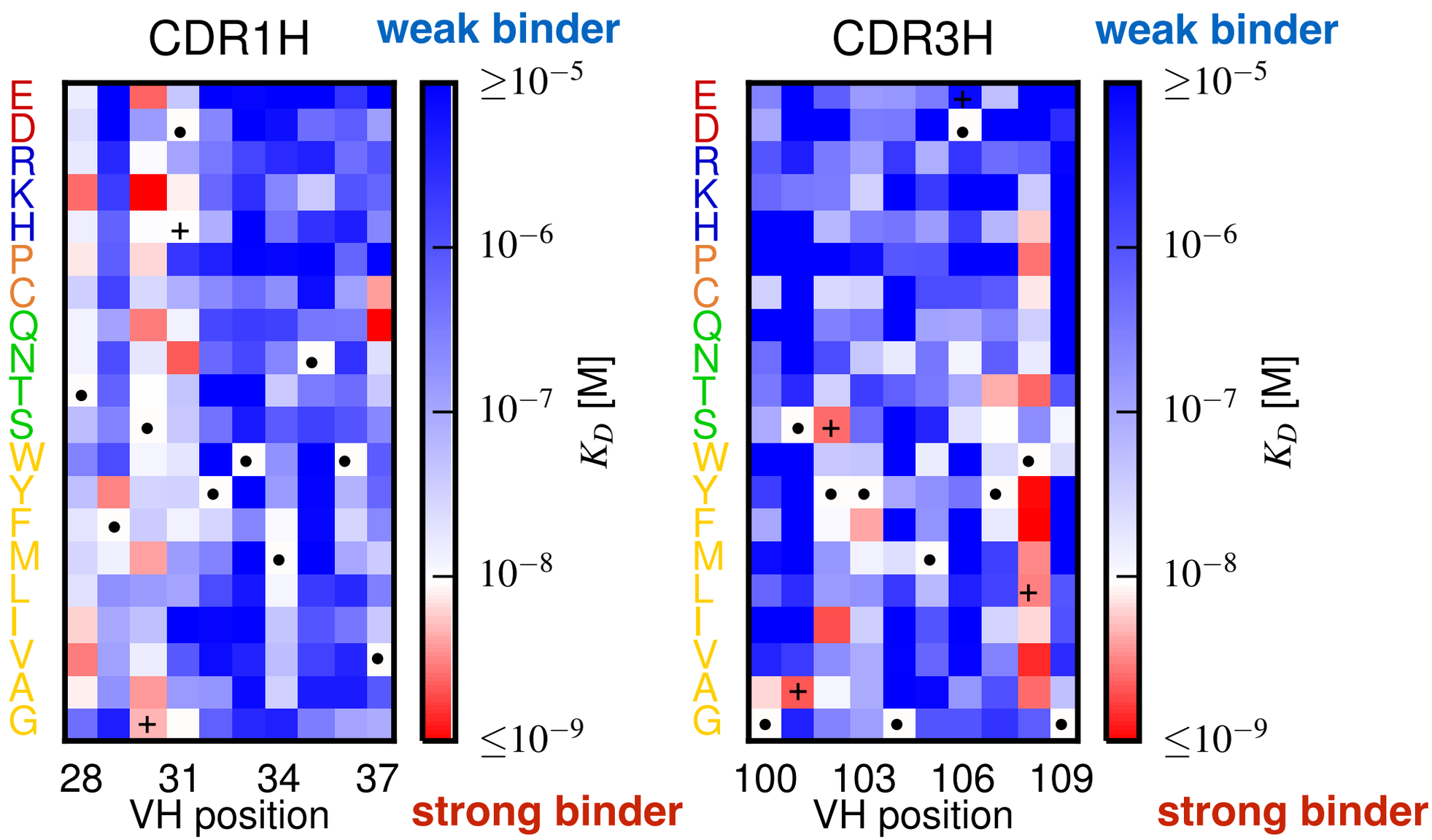




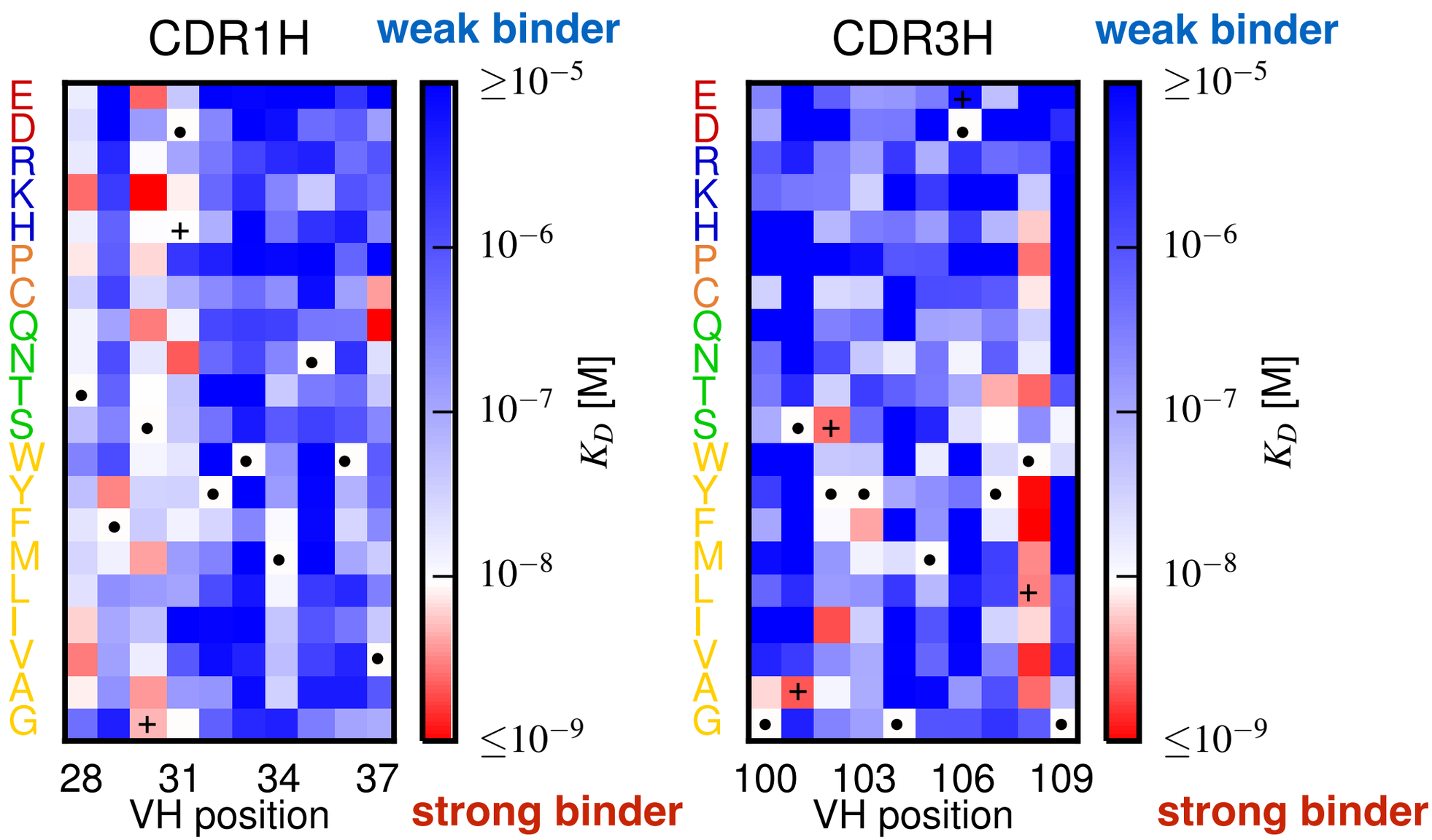
wide range of affinities

Adams Kinney Mora Walczak arXiv 2016

Mutation binding landscape



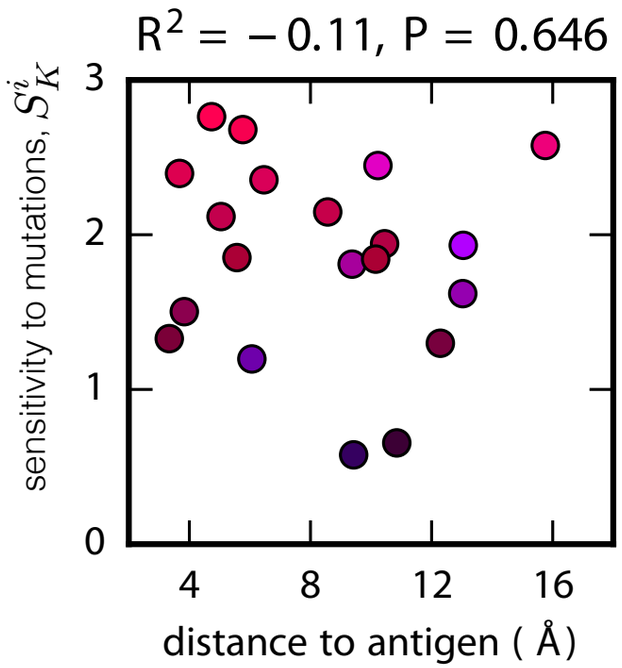
Mutation binding landscape

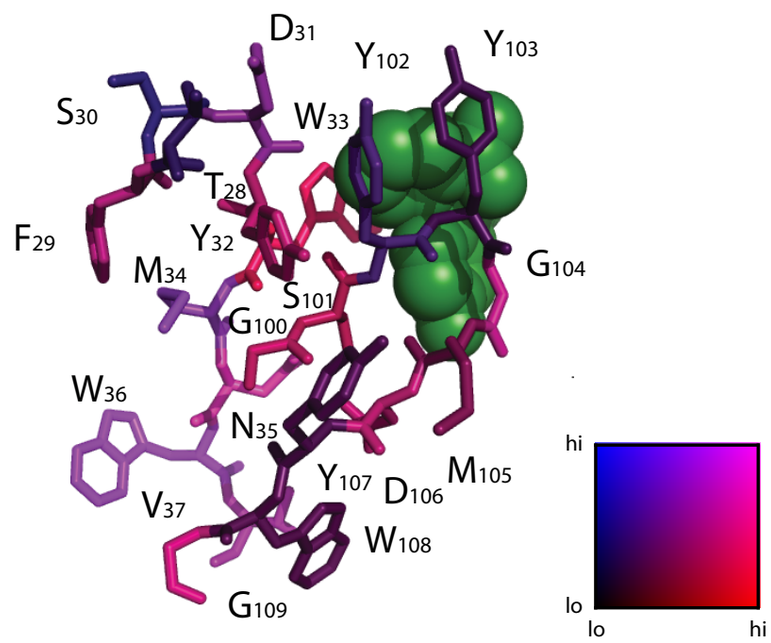


Adams Kinney Mora Walczak arXiv 2016

Effects of mutations

the effect of mutations on affinity: $S_K^i = \sqrt{\left\langle \left(\log_{10} K_D^{ia} - \log_{10} K_D^{\mathrm{WT}}\right)^2 \right\rangle_a}$





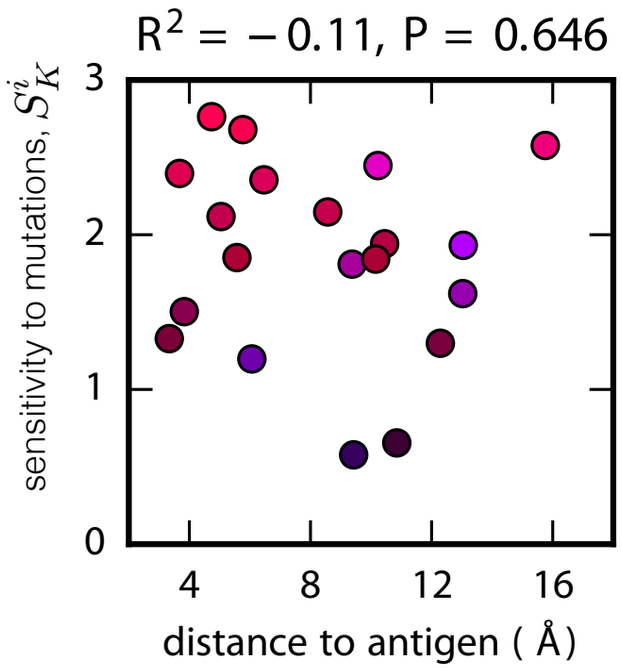
- is independent of distance of mutation to antigen

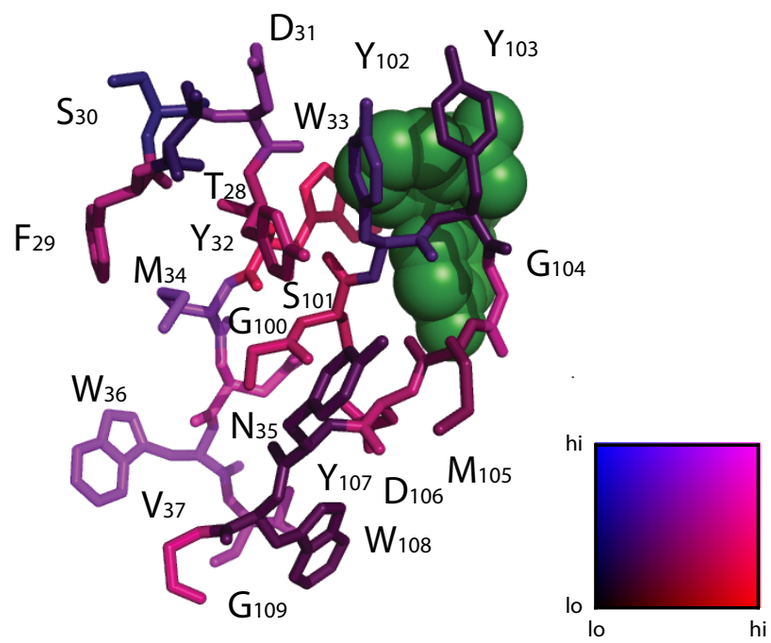
S_E^i

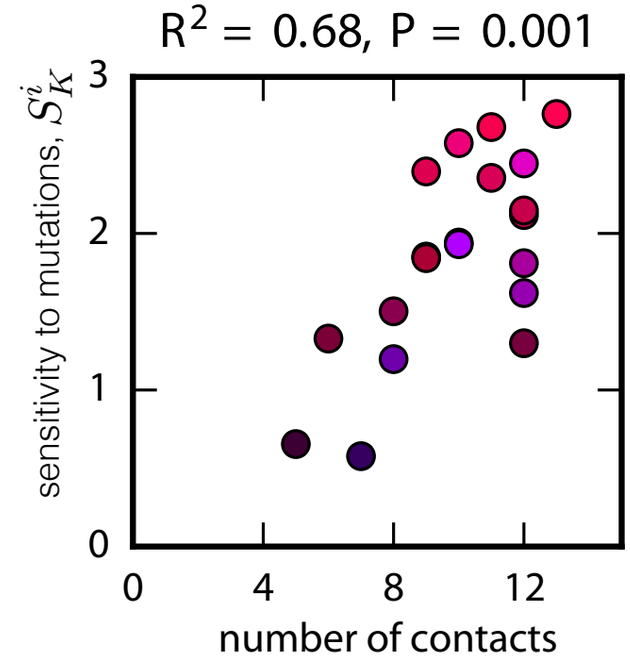
S_K^i

Effects of mutations

the effect of mutations on affinity: $S_K^i = \sqrt{\left\langle \left(\log_{10} K_D^{ia} - \log_{10} K_D^{\mathrm{WT}}\right)^2 \right\rangle_a}$







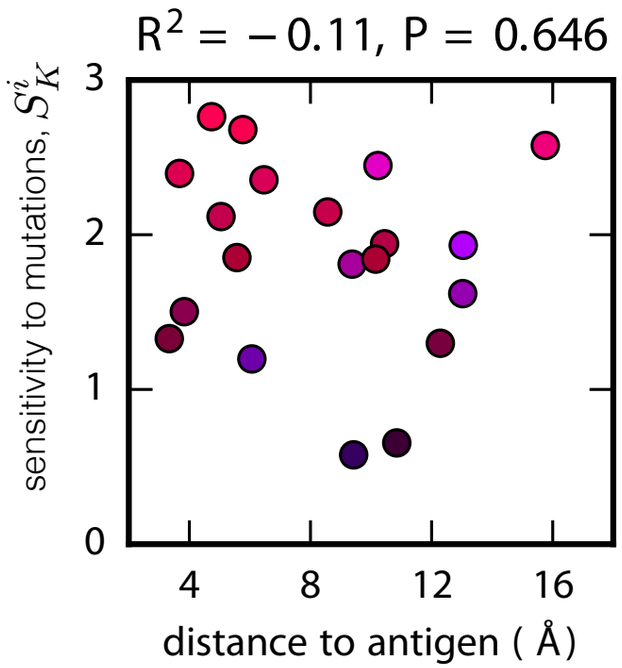
- is independent of distance of mutation to antigen
- depends on the number of contacts a residue makes in the receptor

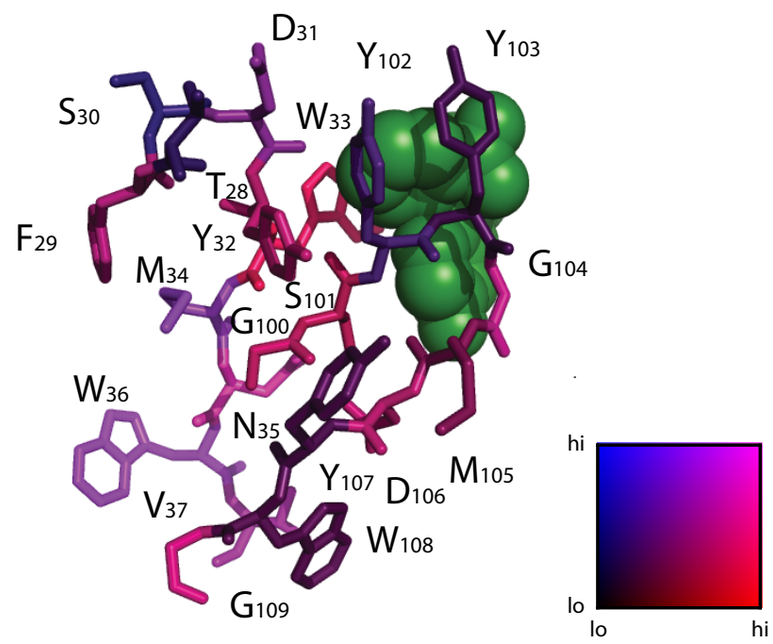
S_E^i

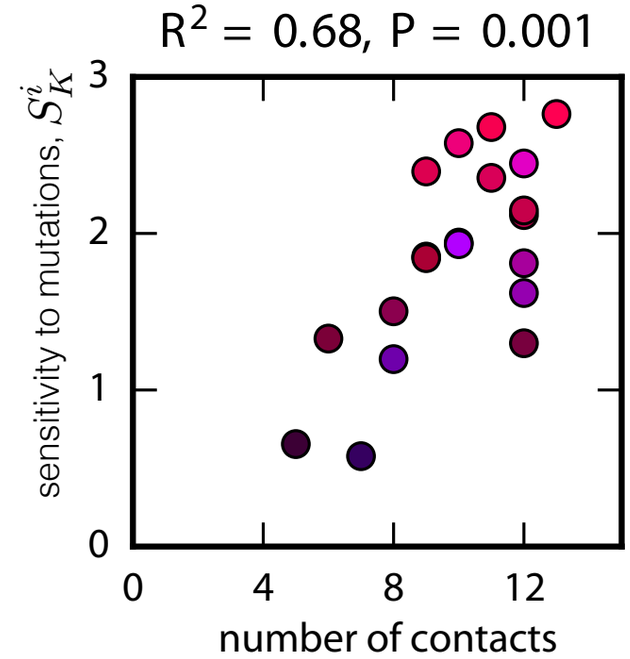
S_K^i

Effects of mutations

the effect of mutations on affinity: $S_K^i = \sqrt{\left\langle \left(\log_{10} K_D^{ia} - \log_{10} K_D^{\mathrm{WT}}\right)^2 \right\rangle_a}$







- is independent of distance of mutation to antigen
- depends on the number of contacts a residue makes in the receptor

- non-local effects

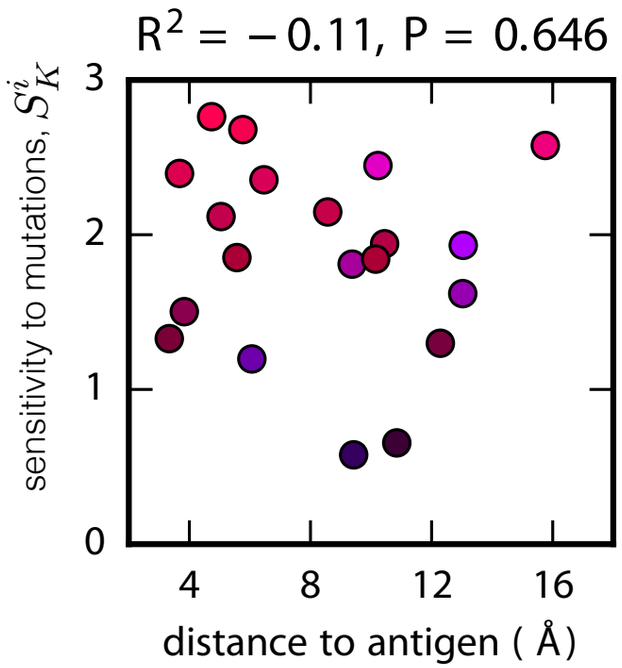
  → effect of interactions between receptor residues

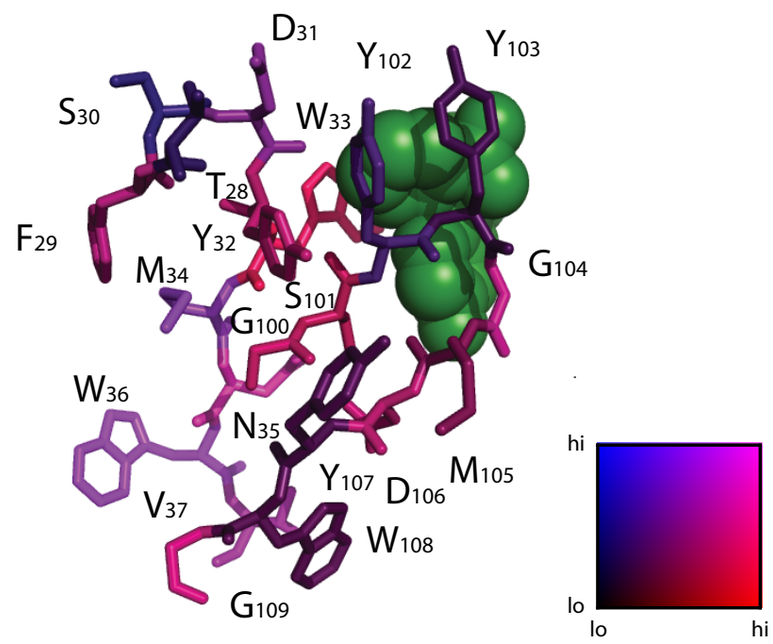
S_E^i

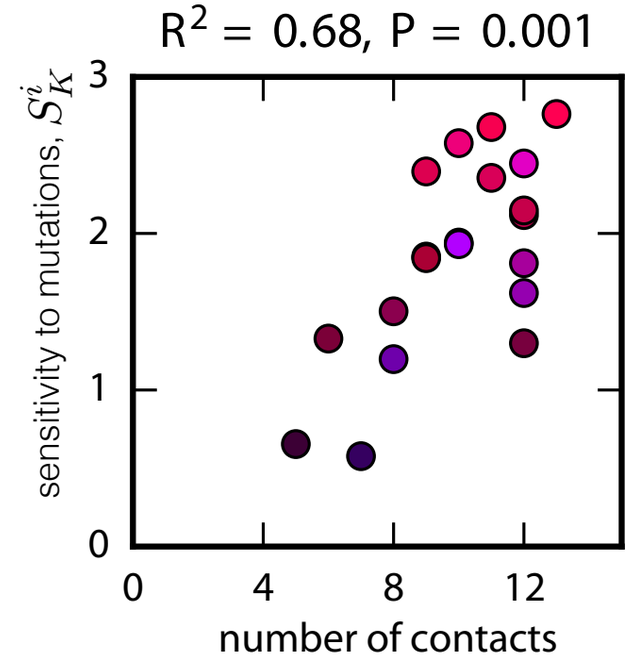
S_K^i

Effects of mutations

the effect of mutations on affinity: $S_K^i = \sqrt{\left\langle \left(\log_{10} K_D^{ia} - \log_{10} K_D^{\mathrm{WT}}\right)^2 \right\rangle_a}$







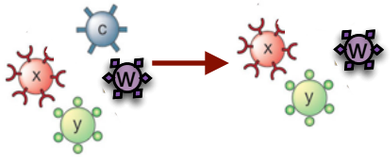
- is independent of distance of mutation to antigen
- depends on the number of contacts a residue makes in the receptor

- non-local effects
  → effect of interactions between receptor residues
- CDR3 mutations have greater effect on affinity
  →more likely to be mutated in functional receptors

S_E^i

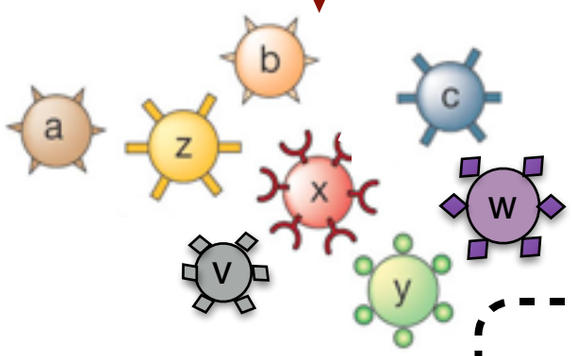
S_K^i

Repertoire evolution

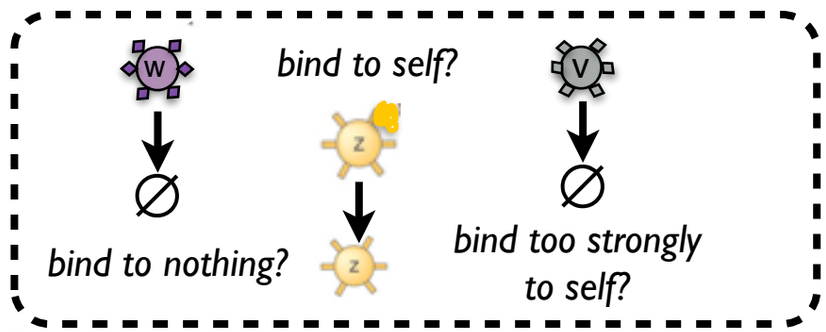


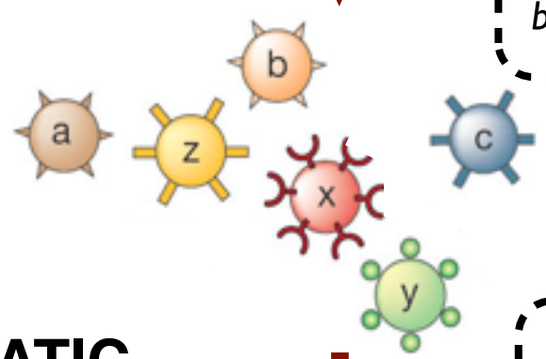
RECEPTOR GENERATION

VDJ recombination

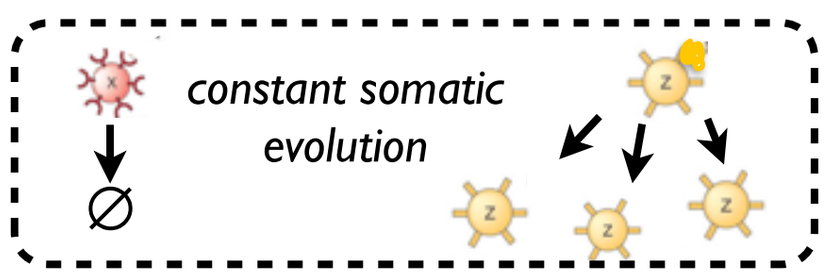


THYMIC SELECTION

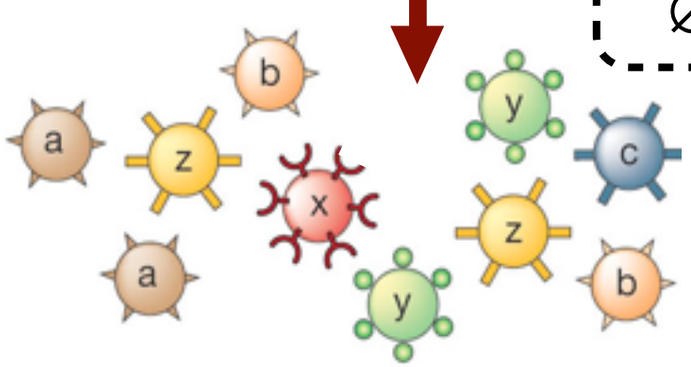




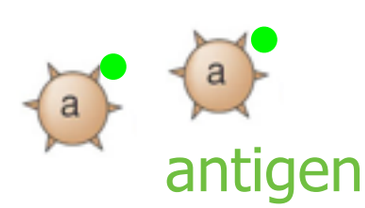
SOMATIC SELECTION



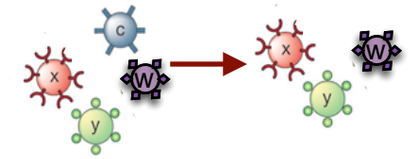
+ Somatic Hypermutations

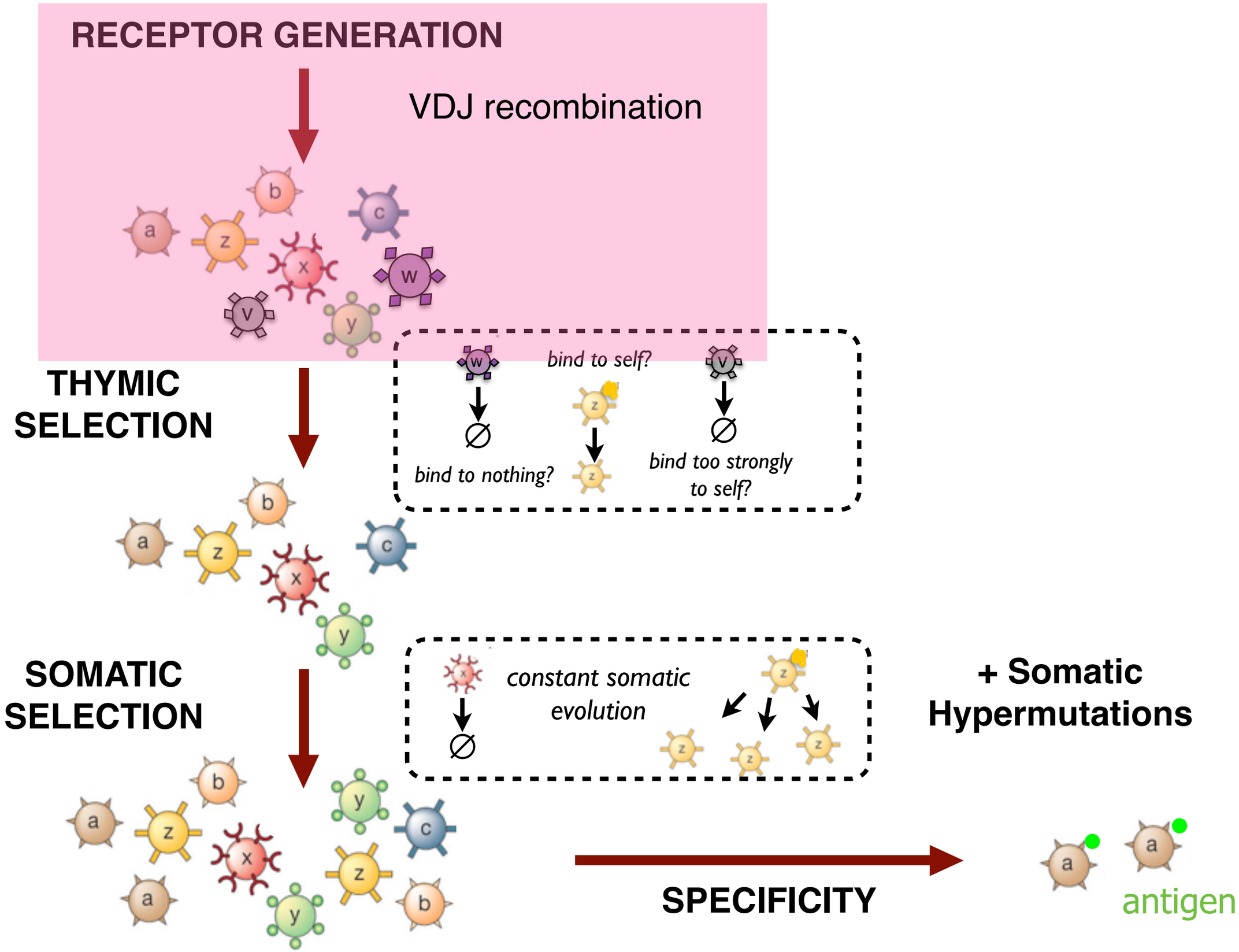


SPECIFICITY

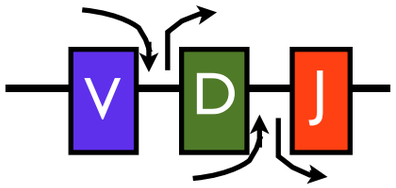


Repertoire evolution



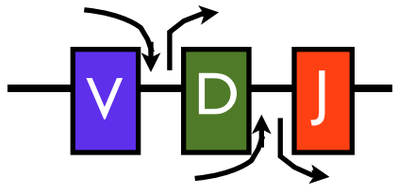


Sequence data

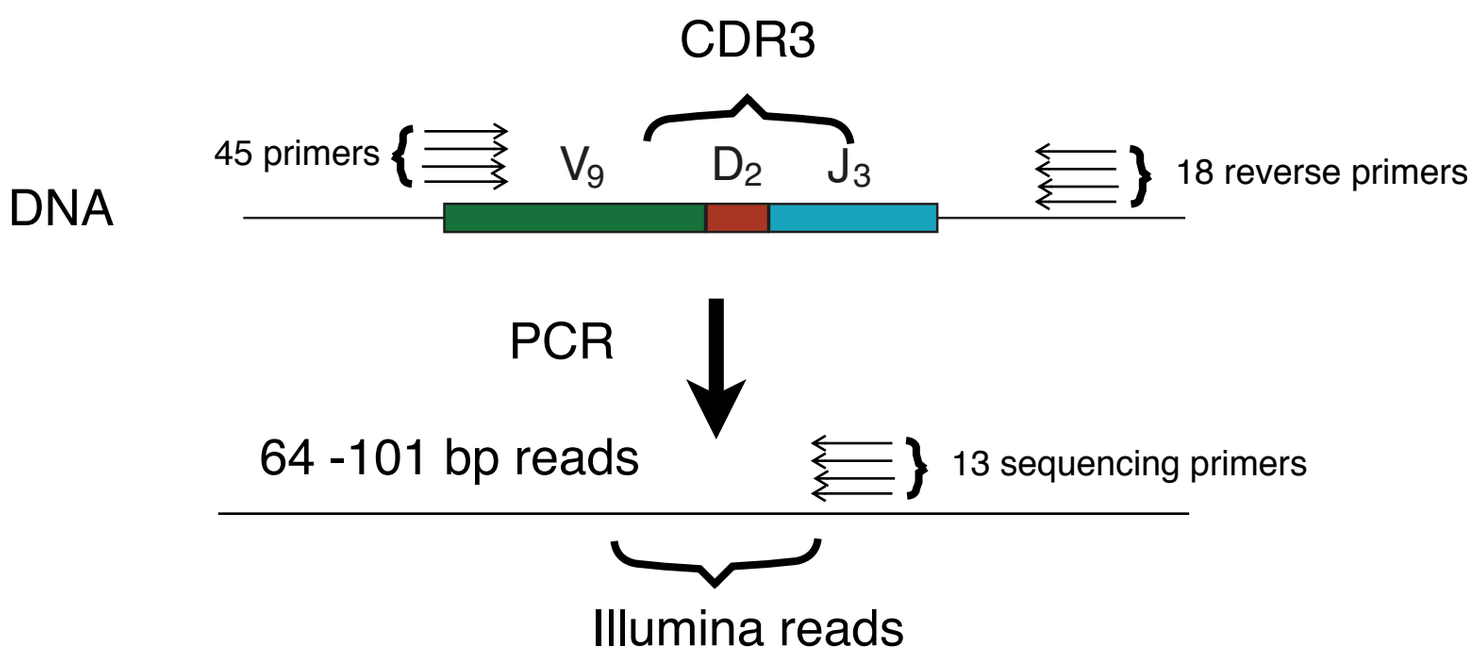


- human T-cell beta chain receptor sequences
- 9 people

Sequence data



- human T-cell beta chain receptor sequences
- 9 people
- CD4+ naive cells
- out of frame reads (~14%) = 35,000 unique reads ⟶ *generation*
- in frame reads (~235,000 unique reads) ⟶ *selection*

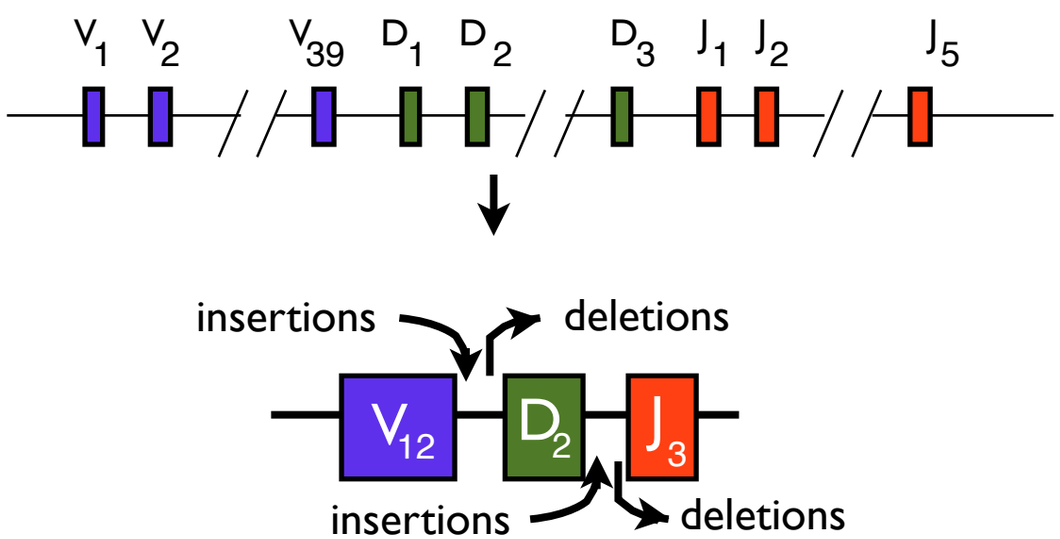


Robins et al, Blood (2009)

data from Robins lab 2009-2012

learning VDJ recombination

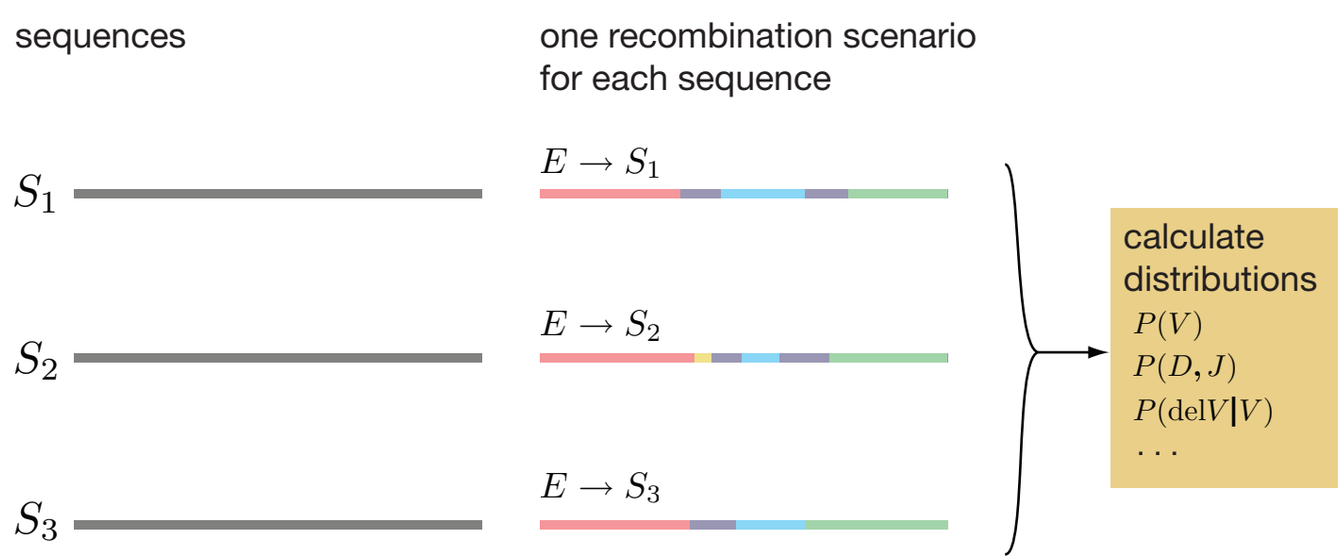
sequence generation: VDJ recombination



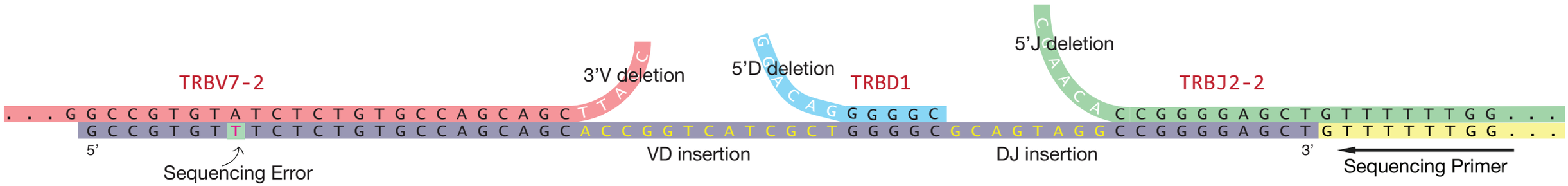

```
TAGGACCTCGGAAACCTCTTCTGCTGGCGCAGAGATACAGGGCTGAGTCTTCTT
TATCTGCCGATGTCGCGAAGGCCCTCCCGCTAAGATCACTGGTGGCACAGAAGT
TAGGGAGAGGTGCCTAACTGCTGGCACAGAAGTACAGAGAGGTCTGGTTGGGGT
TCCGCCGCTAGTCCCTGAAACTACTGGCACAGAGATAGAAAGCTGTCGGGTTCT
CGAAACTGCTGGCACAGAAGTACACAGATGTTTGGGAGGGAGCAGCCGACTCCA
TCTTGGCCGCTAGTCCGAGAAACTGCTGGCACAGAAGTACACAGATGTTTGGGA
TTGTAGGAGCCGGACCGGCCCCCTGTCCCCTTGGCTGCTGGCGCAGAGATACAG
TCATTTAACGTGCGGCCCGCCTGGCACAGAAGTAAAGAGCTGTCTGGTTGTGGT
TAGTAACTCCGCTTCACTGCTGGCACAGAAGTACACAGATGTCTGGGAGGGAGC
TCCCTCCGGTTTGAAGGGTCTGCTGGCACAGAAGTACACAGATGTTTGGGAGGG
CCGGTGTTCGCACAGCCCTGGGGACCCTGGCGCAAACCCCGCTTCCCTCGAGGA
```

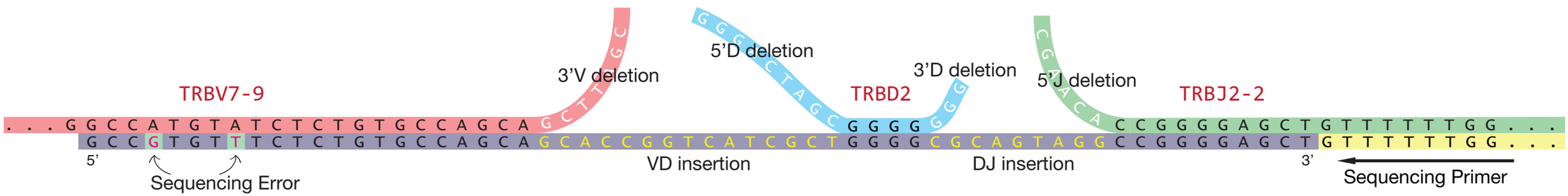
Probability distribution?

- too many possible sequences to sample
- basic approach



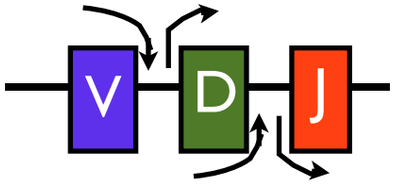
The problem





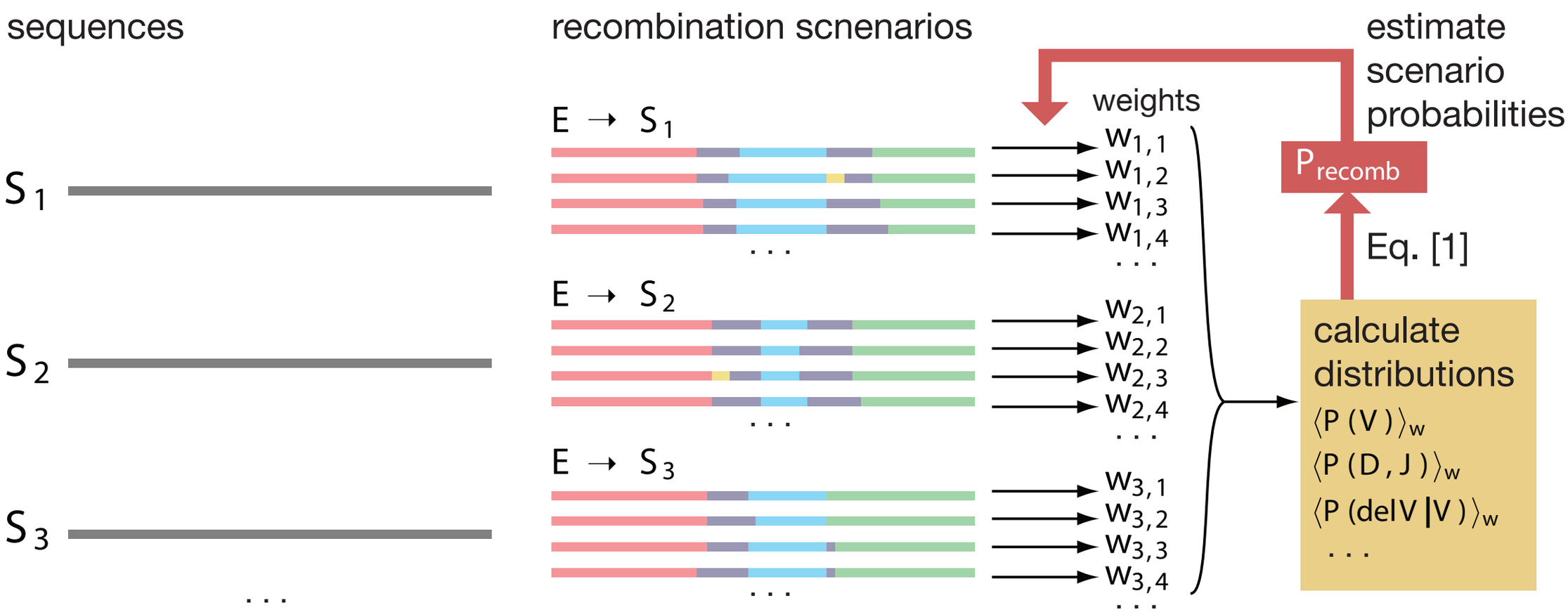
- impossible to reliably assign events (insertions, deletion, ...)
- sequencing errors

Expectation maximization



- probabilistic model for assignment of:
  - genomic VDJ assignment
  - cut position/deletions
  - insertions

$\vec{\sigma}$ - receptor DNA sequence

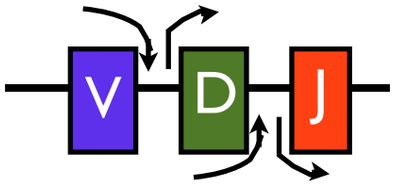


$$\mathrm{P}^{\mathrm{recomb}}(\text{scenario}) = P(V)P(D,J)P(\text{deletions}V|V)P(\text{insertions}DJ)\ldots \qquad [1]$$

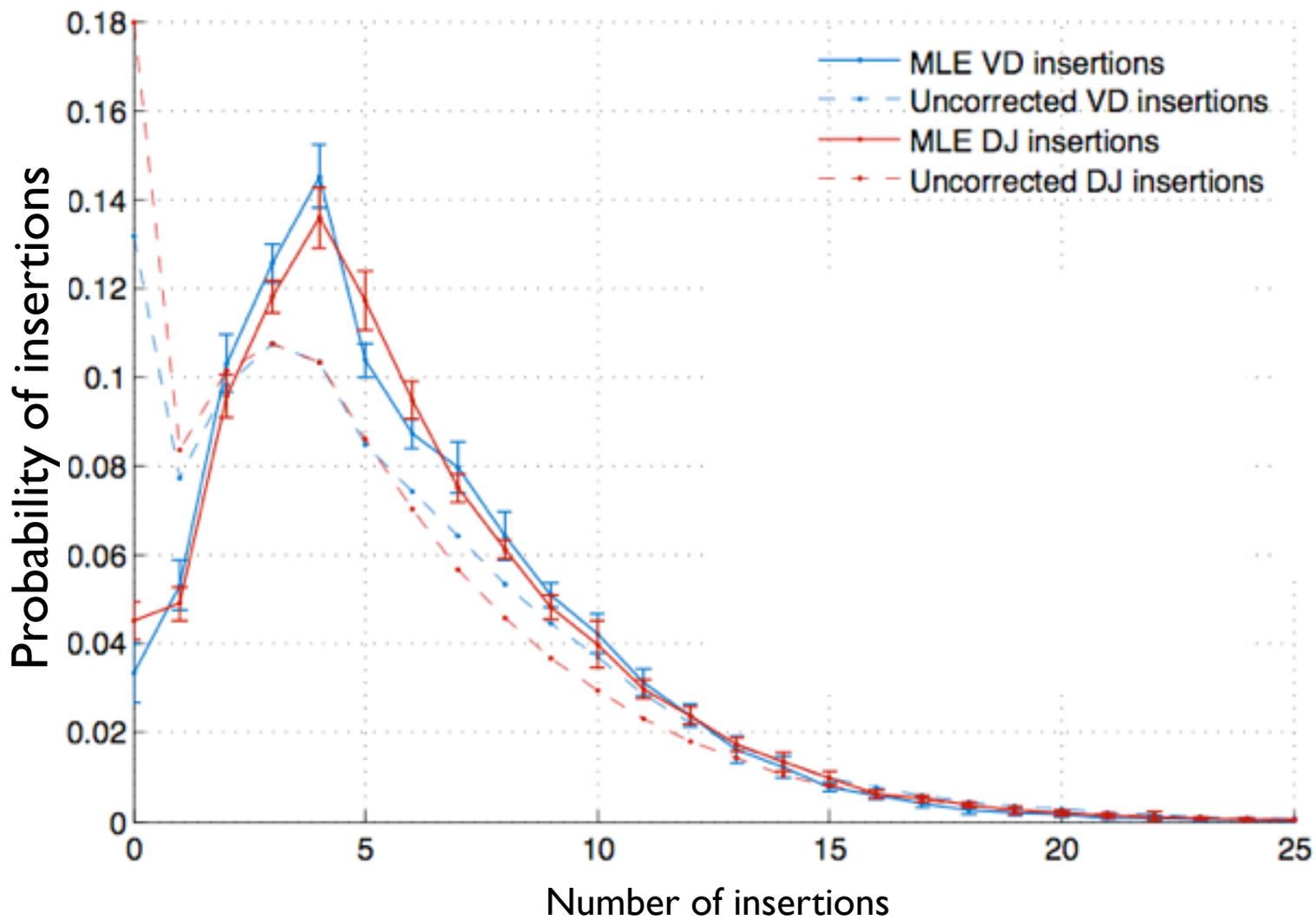
$$P_{gen}(\vec{\sigma}) = \sum_{\substack{\text{scenarios:}\\ V,D,J,\ldots\to\vec{\sigma}}} P^{\mathrm{recomb}}(\text{scenario})$$

Murugan Mora Walczak Callan PNAS (2012)

Universal insertion profiles

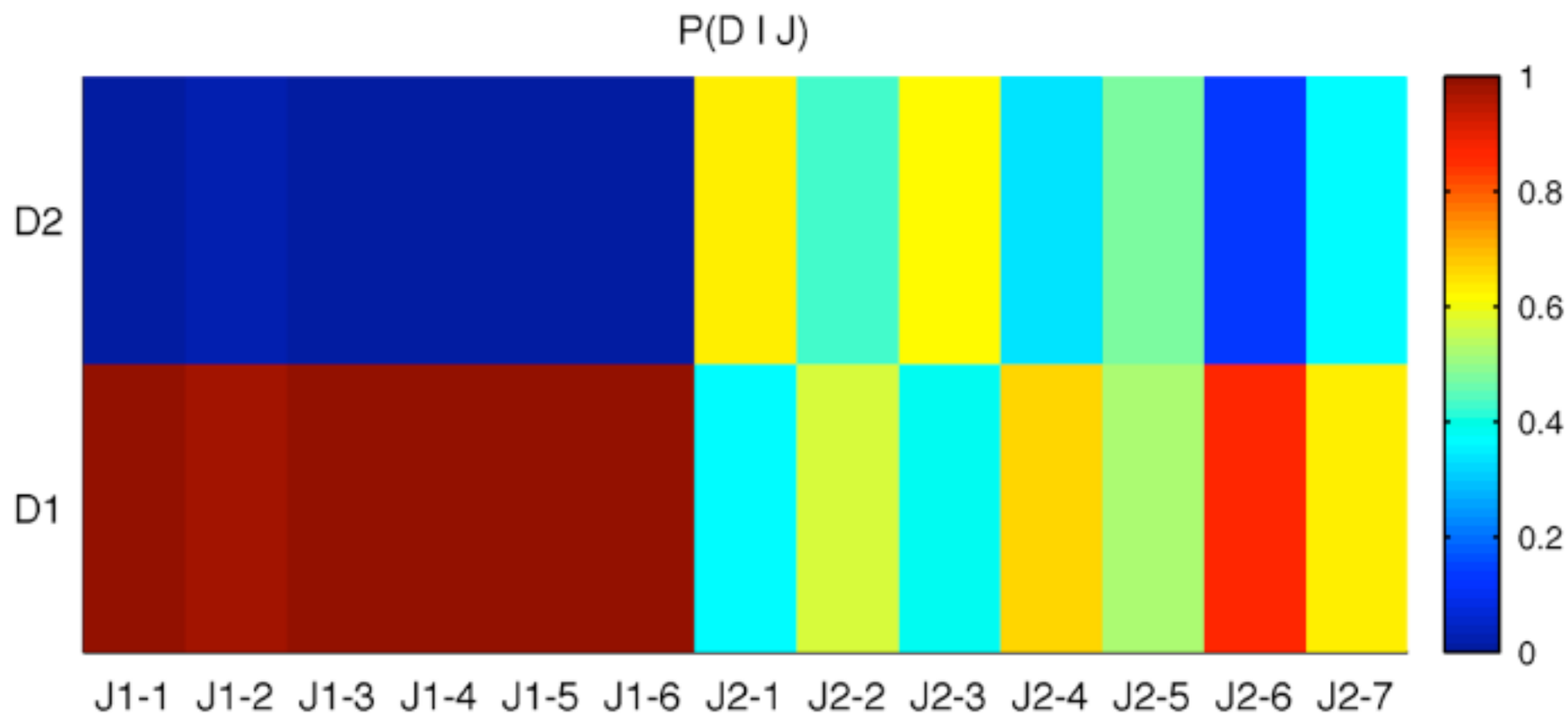


- VD and DJ insertion profiles are identical



Probabilistic is necessary: D and J gene choice

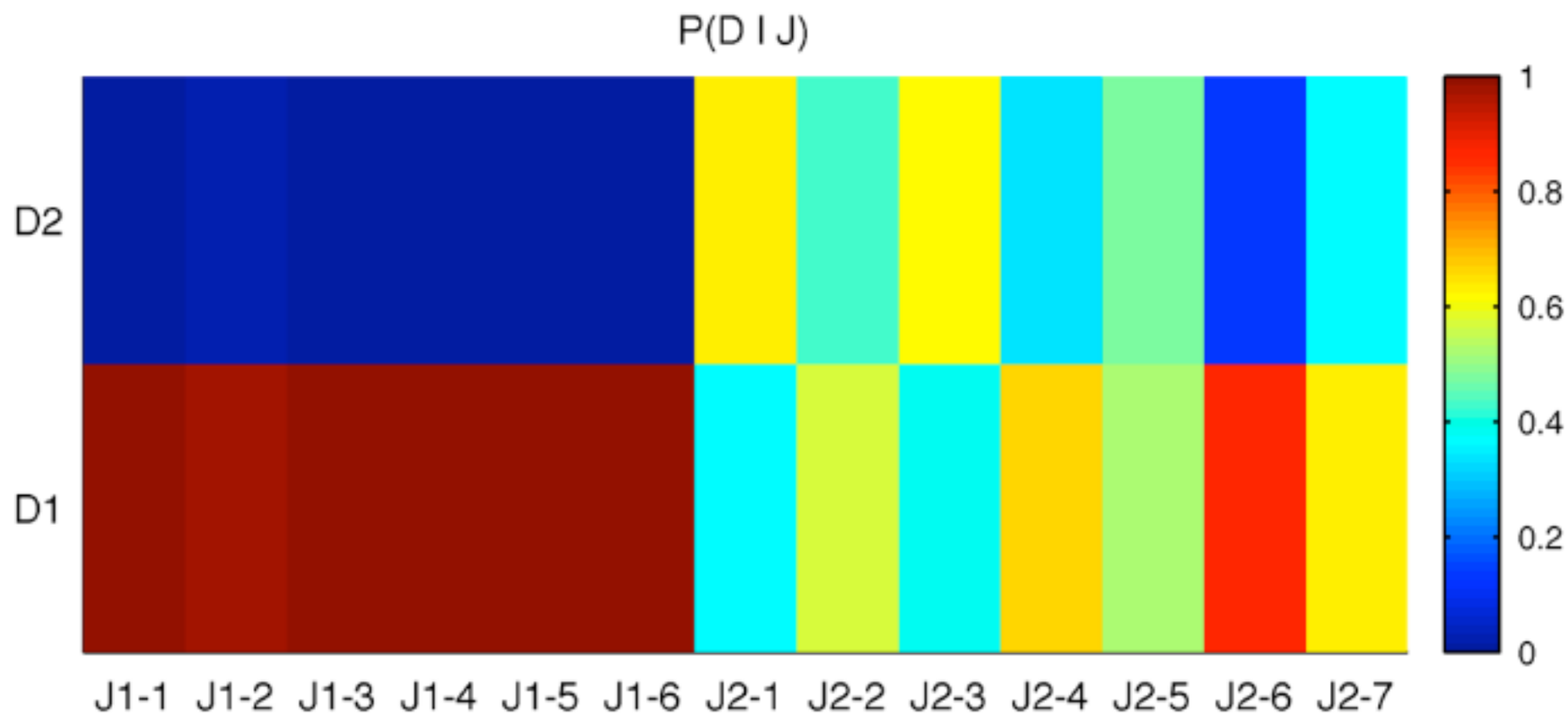
- The choice of J strongly constrains the choice of D



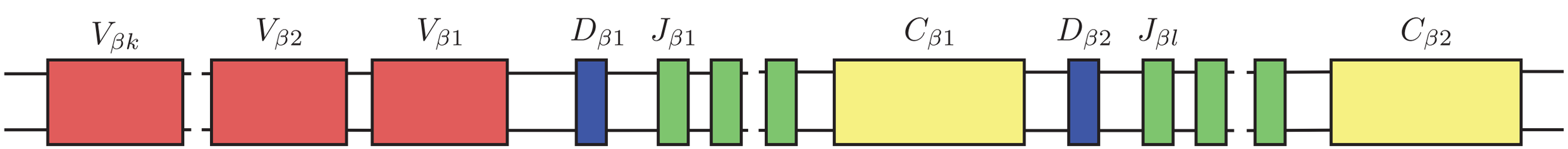
- Not true (20% of forbidden pairings) according to best alignment

Probabilistic is necessary: D and J gene choice

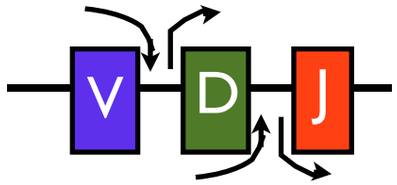
- The choice of J strongly constrains the choice of D



- Not true (20% of forbidden pairings) according to best alignment



Entropy

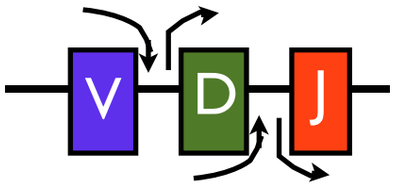


- potential sequence diversity of VDJ recombination

all possible sequences - impossible!

$$S_{\text{gen}} = -\sum_{\sigma} P_{\text{gen}}(\vec{\sigma}) \log P_{\text{gen}}(\vec{\sigma})$$

Entropy



- potential sequence diversity of VDJ recombination

all possible sequences - impossible!

$$S_{\text{gen}} = -\sum_{\sigma} P_{\text{gen}}(\vec{\sigma}) \log P_{\text{gen}}(\vec{\sigma})$$

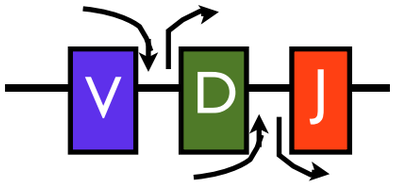
- estimate from

$$S_{\text{gen}} = S_{\text{recomb}} - \langle S(\text{scenario}|\sigma)\rangle_{\sigma}$$

recombination scenario entropy

conditional entropy of recombination events given sequence

Entropy



- potential sequence diversity of VDJ recombination

all possible sequences - impossible!

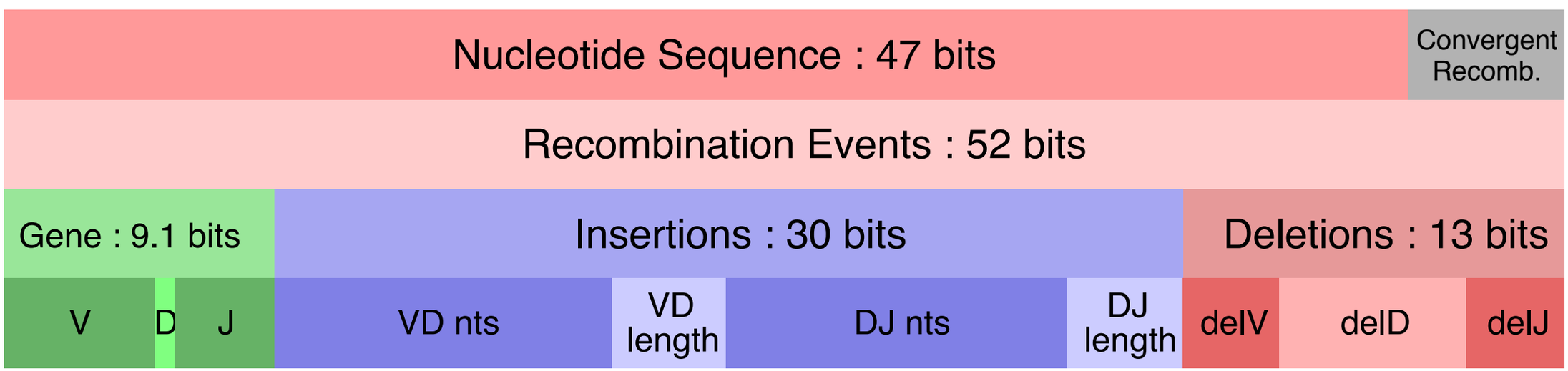
$$S_{\text{gen}} = -\sum_{\sigma} P_{\text{gen}}(\vec{\sigma}) \log P_{\text{gen}}(\vec{\sigma})$$

- estimate from

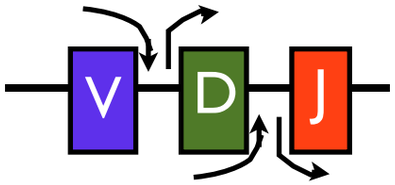
$$S_{\text{gen}} = S_{\text{recomb}} - \langle S(\text{scenario}|\sigma)\rangle_{\sigma}$$

recombination scenario entropy

conditional entropy of recombination events given sequence



Entropy



- potential sequence diversity of VDJ recombination

all possible sequences - impossible!

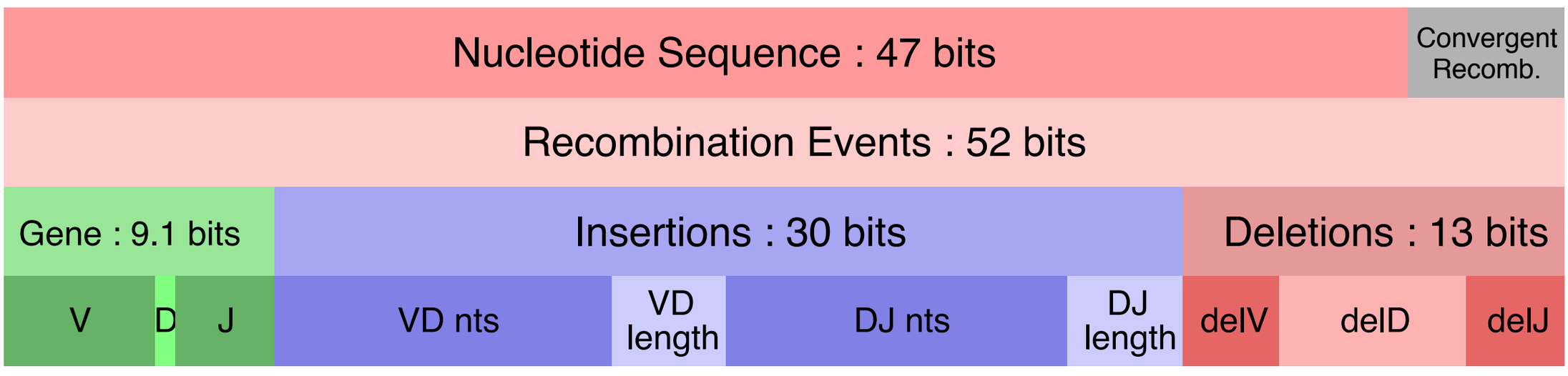
$$S_{\text{gen}} = -\sum_{\sigma} P_{\text{gen}}(\vec{\sigma}) \log P_{\text{gen}}(\vec{\sigma})$$

- estimate from

$$S_{\text{gen}} = S_{\text{recomb}} - \langle S(\text{scenario}|\sigma)\rangle_{\sigma}$$

recombination scenario entropy

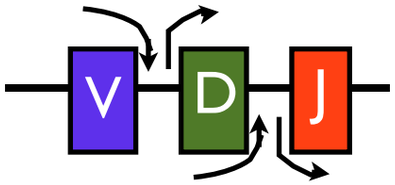
conditional entropy of recombination events given sequence



- 47 bits $\Longrightarrow$ repertoire size 10^{14} sequences > 10^{8}*+ unique seqs in individual
  $3\ 10^{11}$ total T-cells in individual

*Robins et al, Blood (2009)

Entropy



- potential sequence diversity of VDJ recombination

all possible sequences - impossible!

$$S_{\text{gen}} = -\sum_{\sigma} P_{\text{gen}}(\vec{\sigma}) \log P_{\text{gen}}(\vec{\sigma})$$

- estimate from

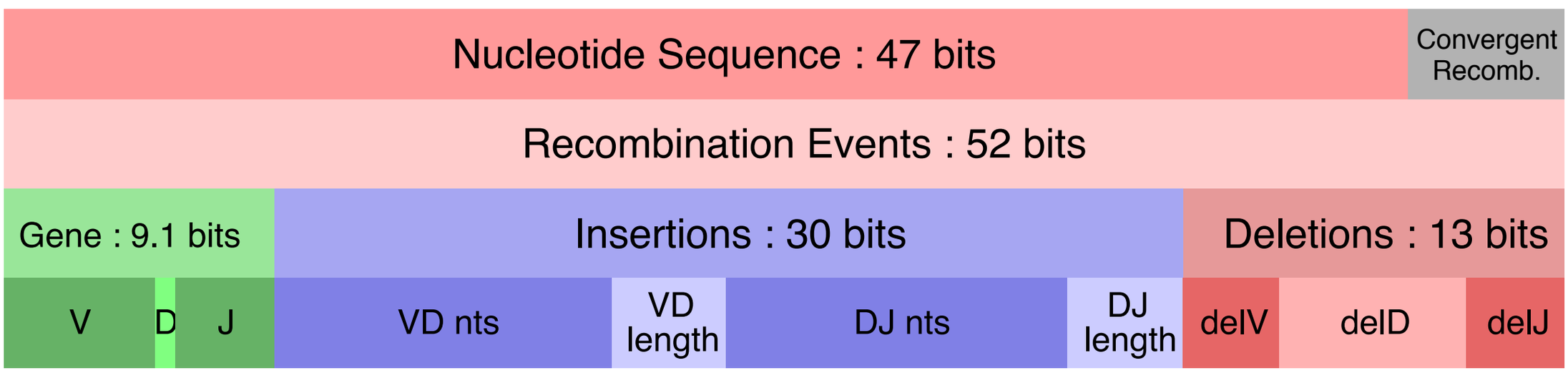
52 bits

5 bits → typical sequence can be generated in 32 different ways

$$S_{\text{gen}} = S_{\text{recomb}} - \langle S(\text{scenario}|\sigma)\rangle_{\sigma}$$

recombination scenario entropy

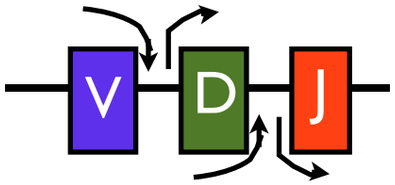
conditional entropy of recombination events given sequence



- 47 bits $\Longrightarrow$ repertoire size 10^{14} sequences > 10^{8}*+ unique seqs in individual
  $3\ 10^{11}$ total T-cells in individual

*Robins et al, Blood (2009)

Entropy



- potential sequence diversity of VDJ recombination

all possible sequences - impossible!

$$S_{\text{gen}} = -\sum_{\sigma} P_{\text{gen}}(\vec{\sigma}) \log P_{\text{gen}}(\vec{\sigma})$$

- estimate from

52 bits

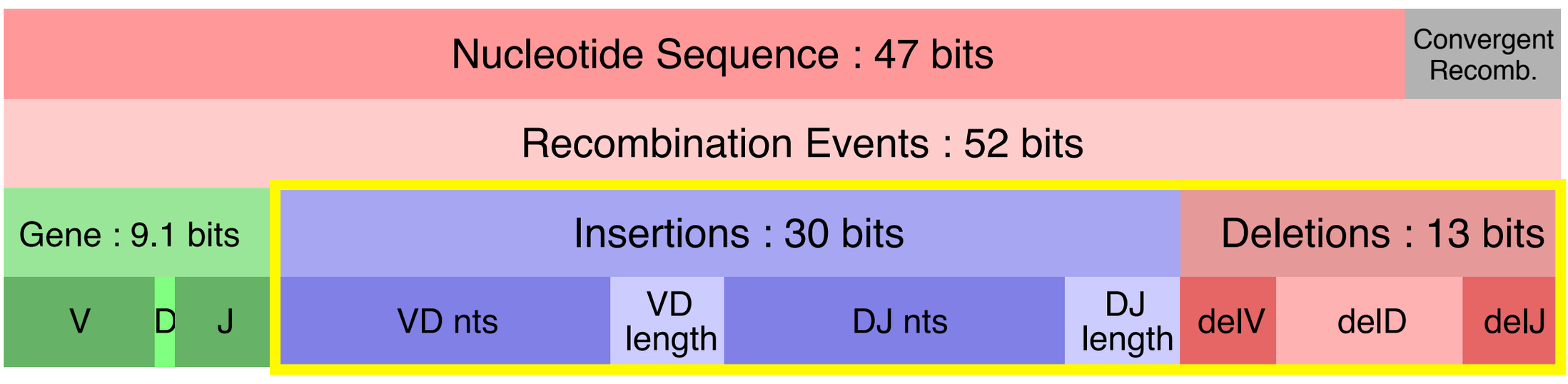
5 bits → typical sequence can be generated in 32 different ways

$$S_{\text{gen}} = S_{\text{recomb}} - \langle S(\text{scenario}|\sigma)\rangle_{\sigma}$$

recombination scenario entropy

conditional entropy of recombination events given sequence

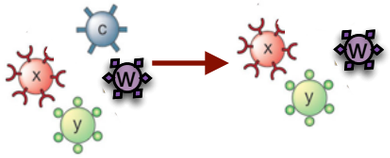
- diversity dominated by junctional diversity



- 47 bits $\Longrightarrow$ repertoire size 10^{14} sequences > 10^{8}*+ unique seqs in individual
  $3\ 10^{11}$ total T-cells in individual

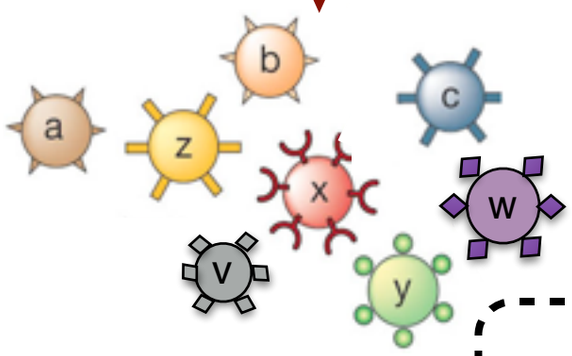
*Robins et al, Blood (2009)

Repertoire evolution

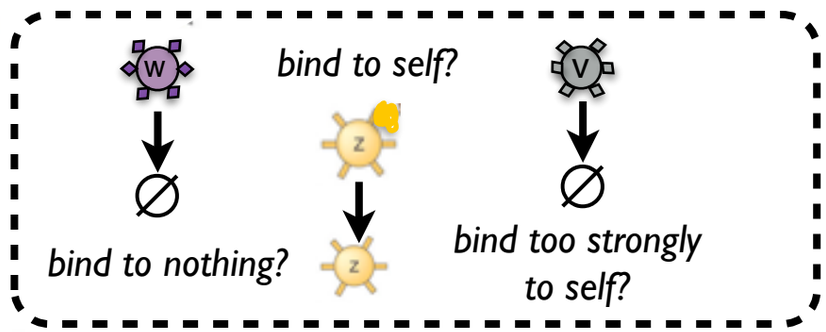


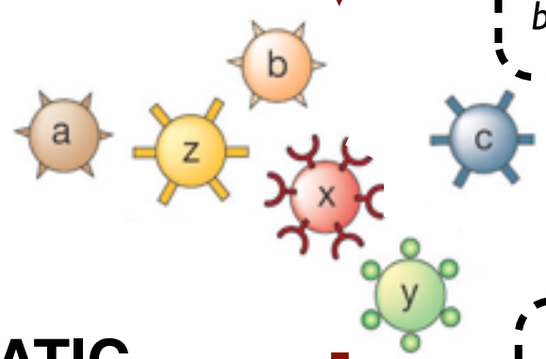
RECEPTOR GENERATION

VDJ recombination

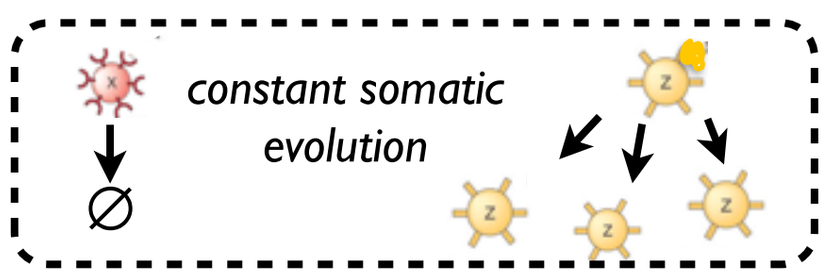


THYMIC SELECTION

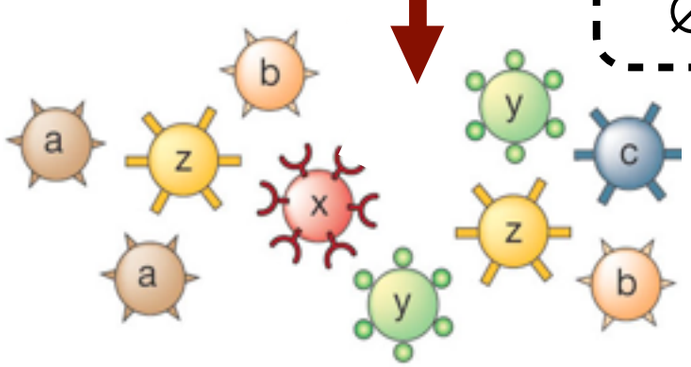




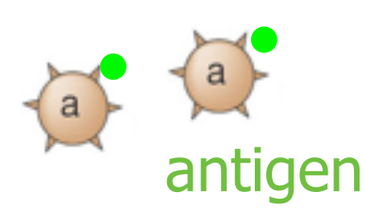
SOMATIC SELECTION



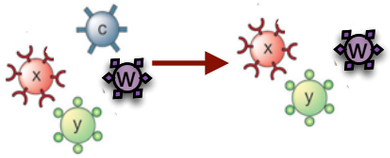
+ Somatic Hypermutations



SPECIFICITY

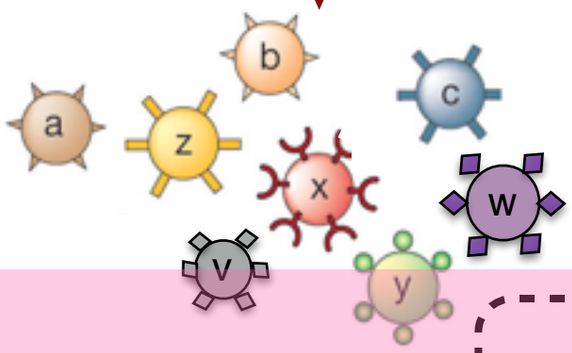


Repertoire evolution

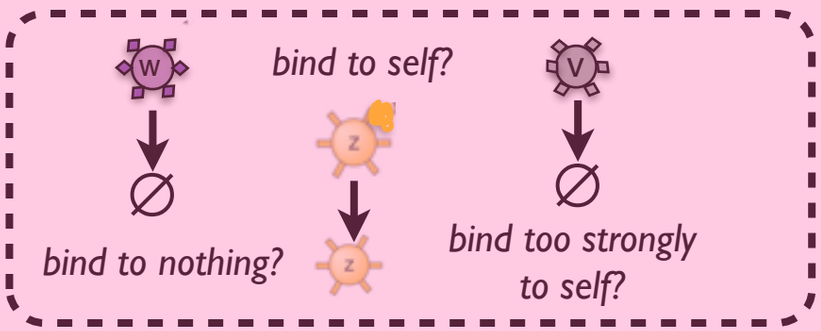


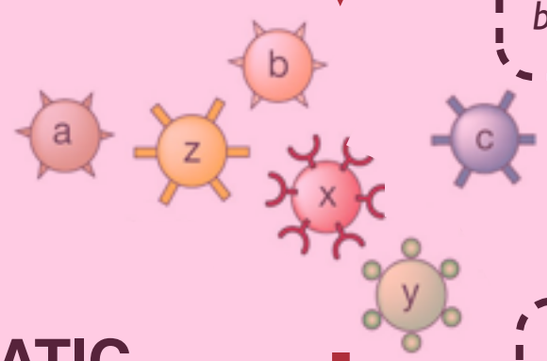
RECEPTOR GENERATION

VDJ recombination

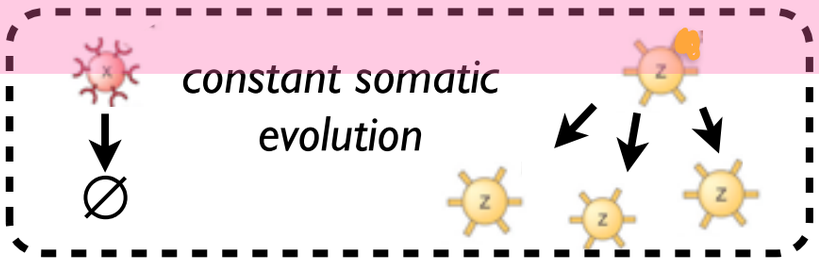


THYMIC SELECTION

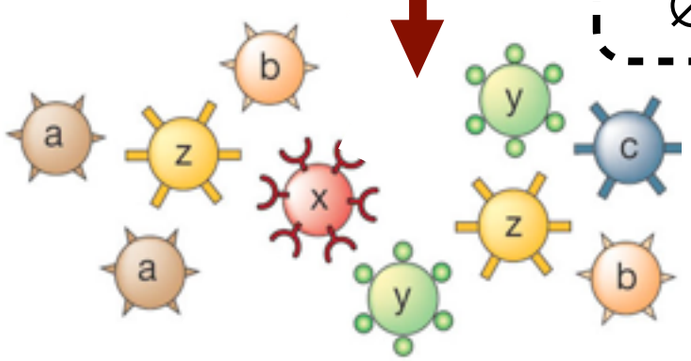




SOMATIC SELECTION

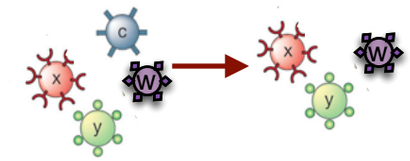


+ Somatic Hypermutations

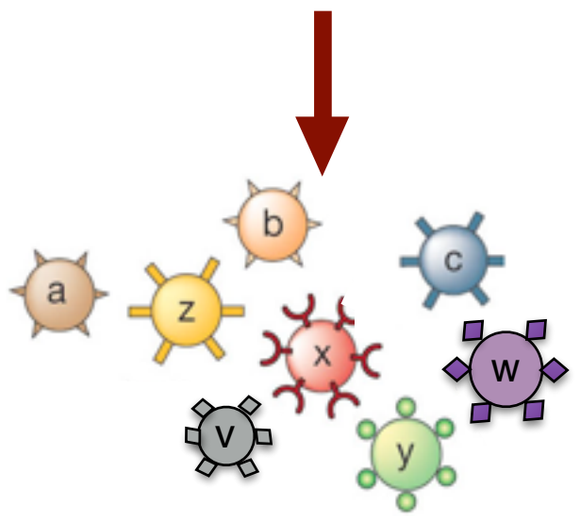


SPECIFICITY

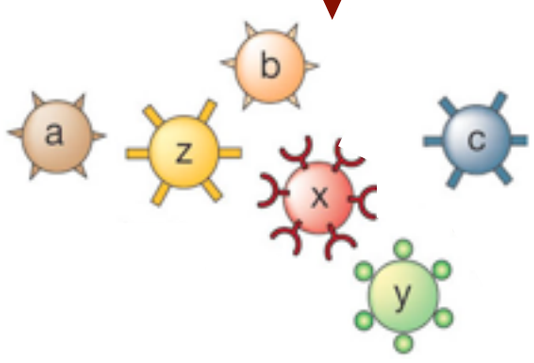
Selection

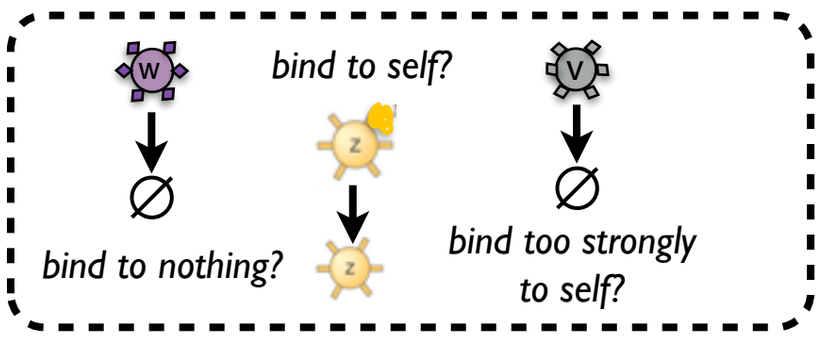


RECEPTOR GENERATION



THYMIC SELECTION

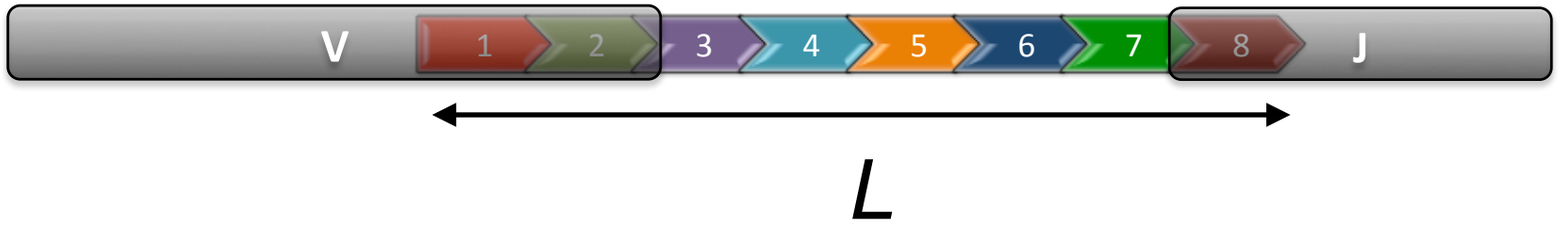




- quantify using selection factors

$$Q(\{\sigma\}) = \frac{P_{\text{post-sel}}(\{\sigma\})}{P_{\text{gen}}(\{\sigma\})}$$

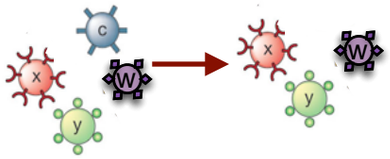
- a model for the observed probabilities

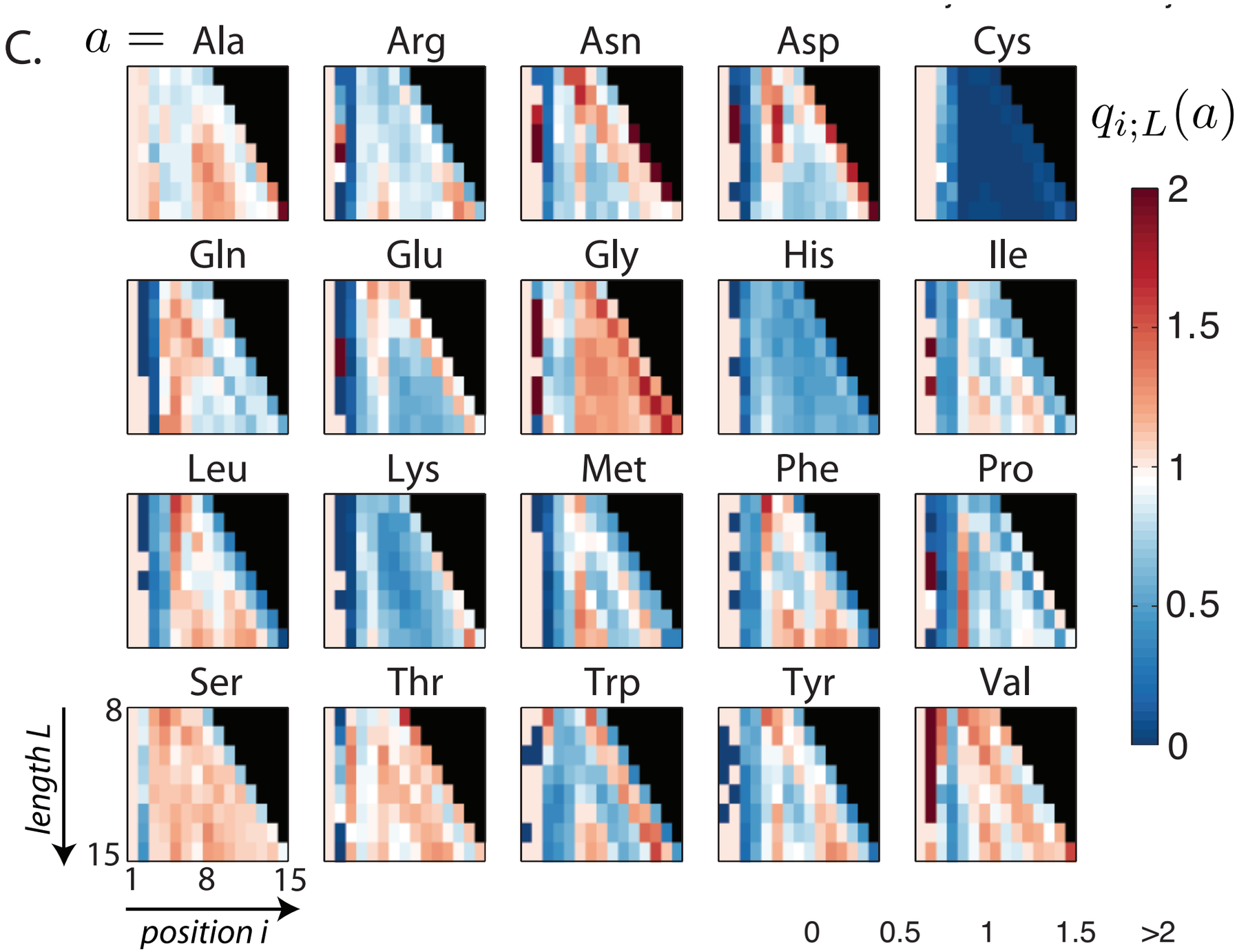


$$Q(\vec{\sigma}, V, J) = \frac{1}{Z}\, q_L\, q_{V,J|L} \prod_{i=1}^{L} q_{i;L}(a_i)$$

Elhanati Callan Mora Walczak PNAS (2014)

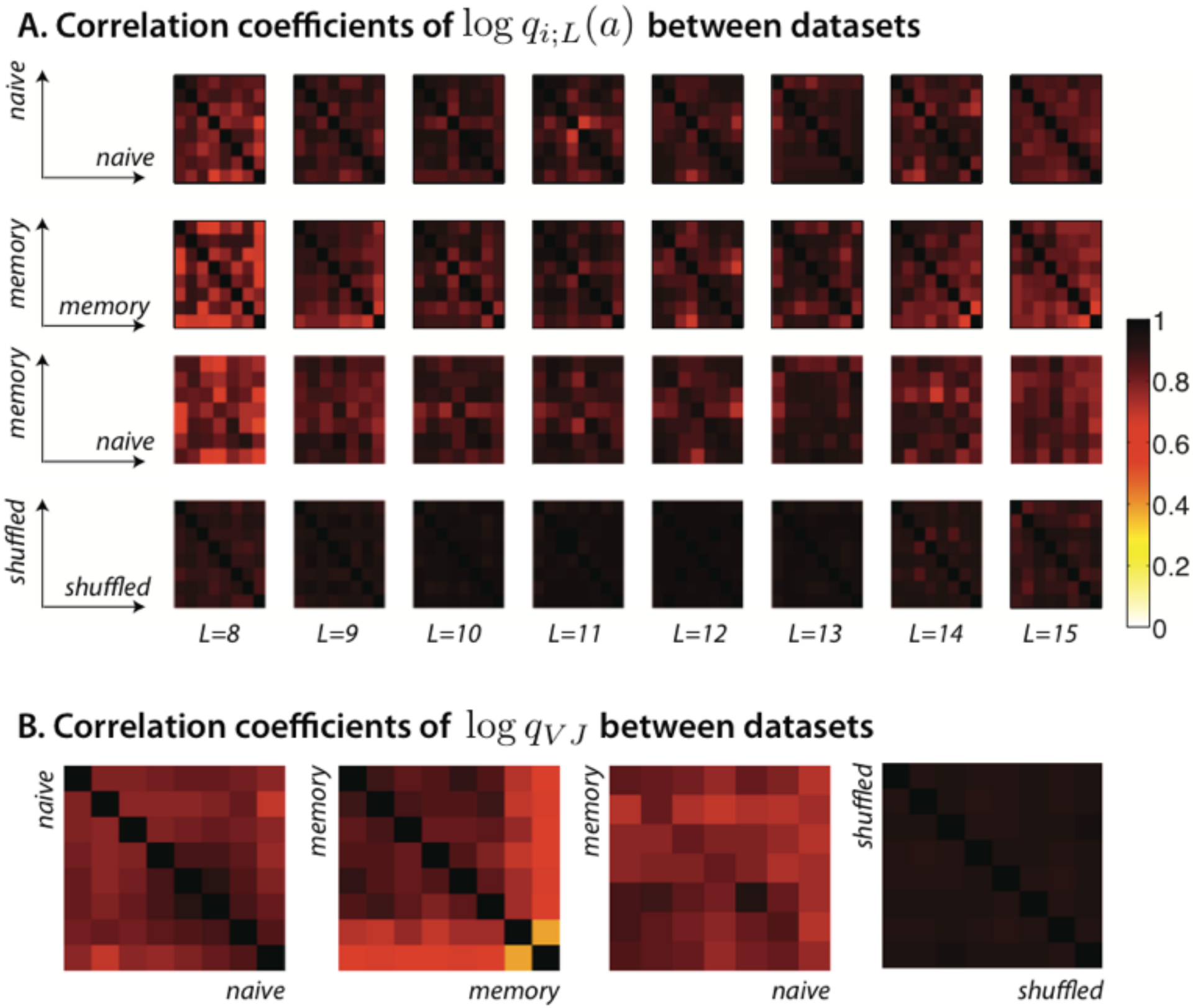
Selection



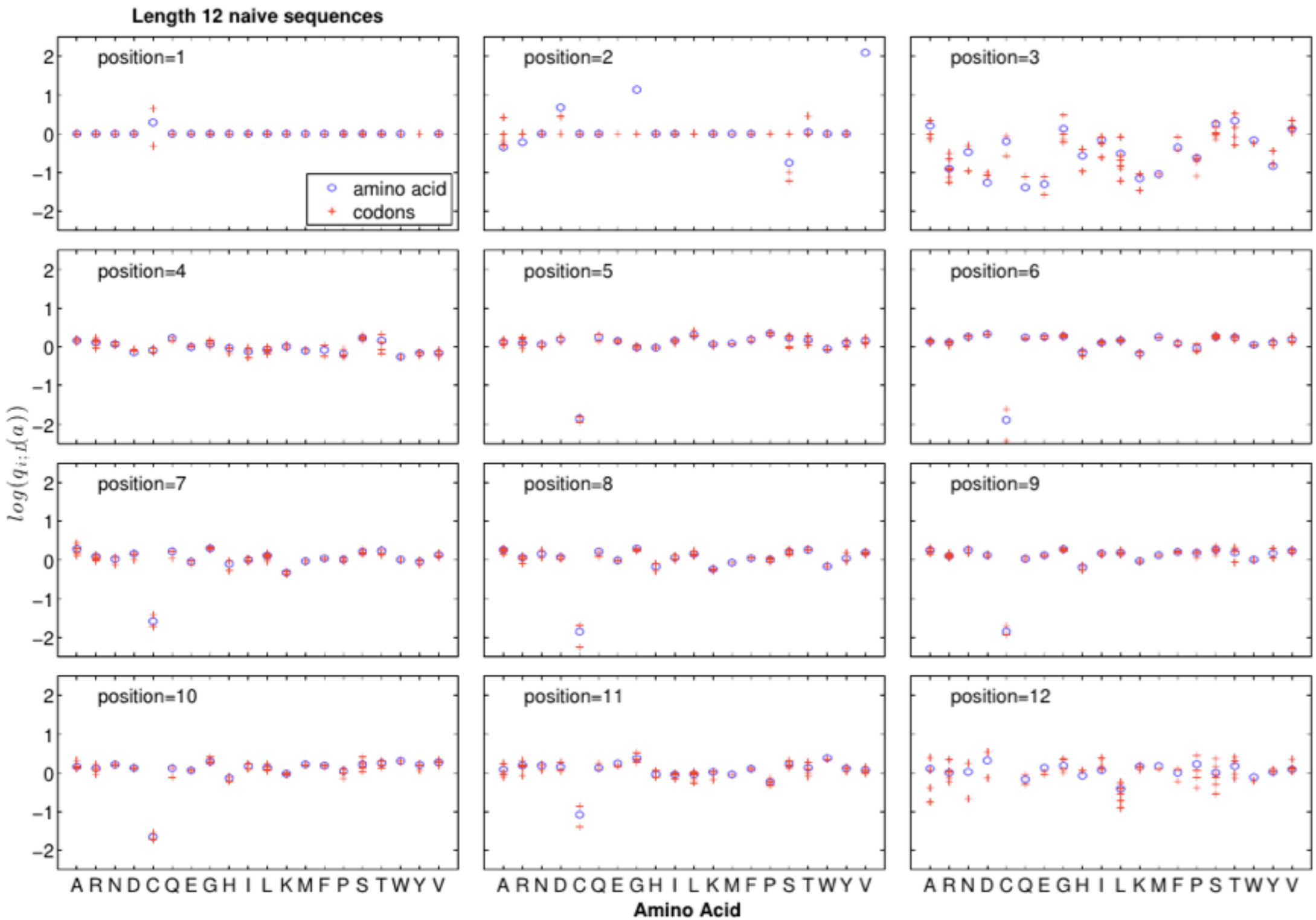


Elhanati Callan Mora Walczak PNAS (2014)

Correlations between individuals

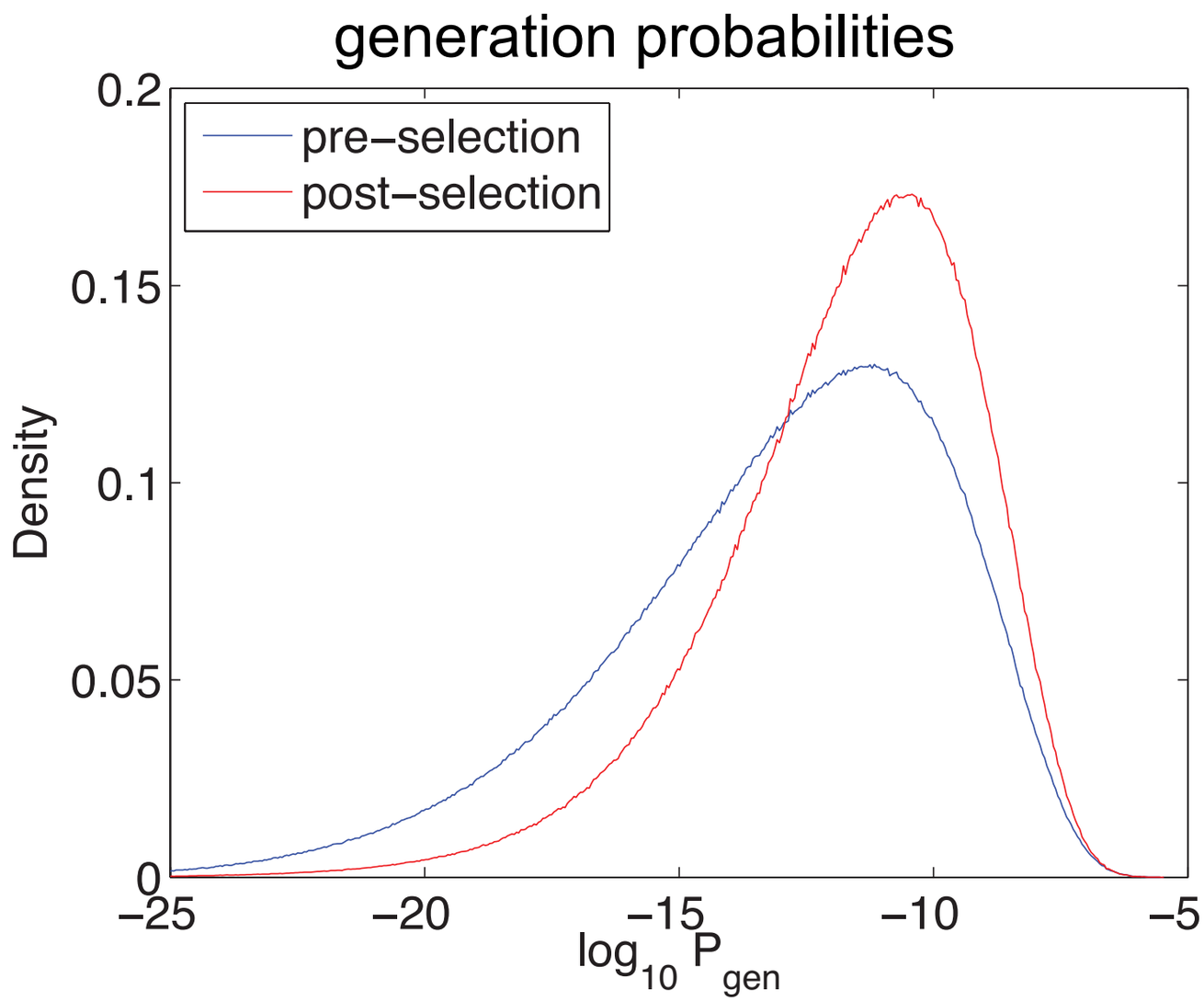


Selection only depends on aa, not codon



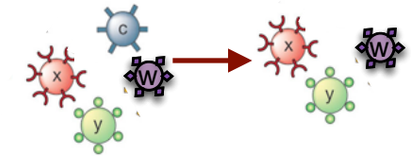
Natural selection anticipates somatic selection

- sequences more likely to be generated → more likely to be selected

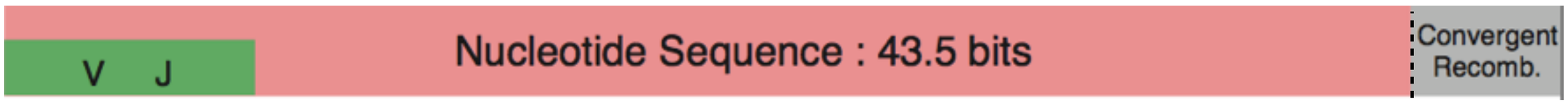


- true for all individuals independently

Entropy, again

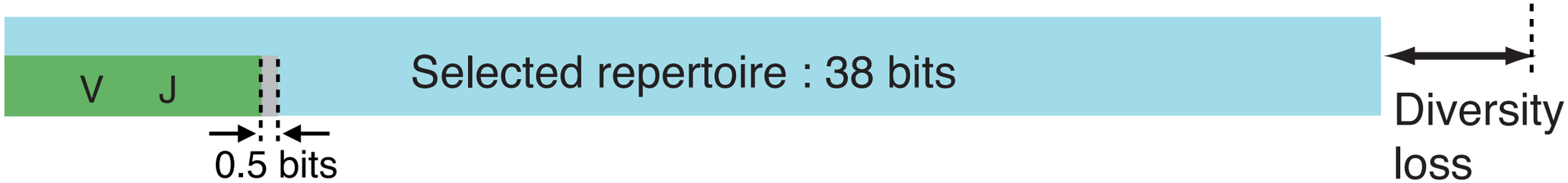


- entropy of generated repertoire



$\Longrightarrow$ repertoire size 10^{13} sequences

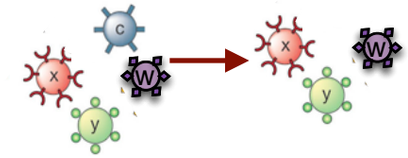
- entropy of post-thymic selection repertoire



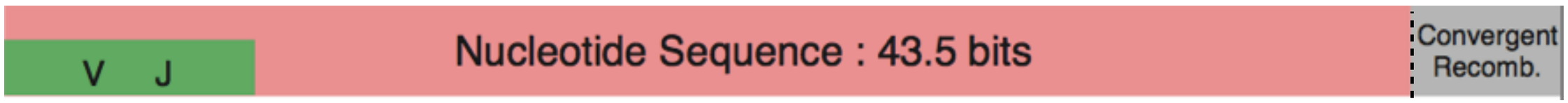
$\Longrightarrow$ repertoire size 10^{11} sequences

$\longrightarrow$ thymic selection gives 50-fold reduction in diversity

Entropy, again

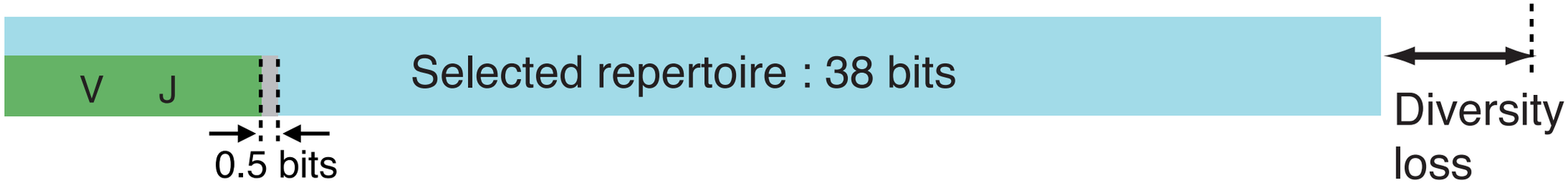


- entropy of generated repertoire



$\Longrightarrow$ repertoire size 10^{13} sequences

- entropy of post-thymic selection repertoire



$\Longrightarrow$ repertoire size 10^{11} sequences

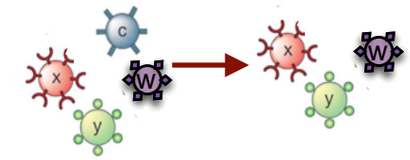
$\longrightarrow$ thymic selection gives 50-fold reduction in diversity

- thymic selection keeps only 2% of diversity

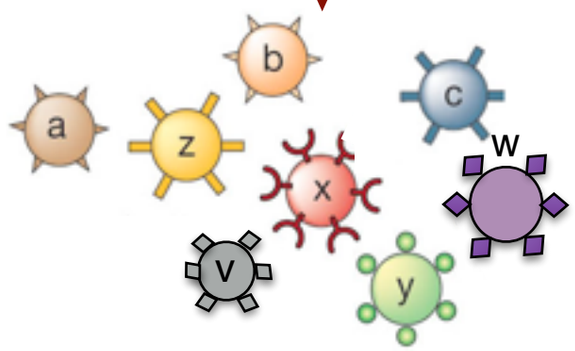
$\longrightarrow$ thymic selection gets rid of rare clones

selection favours clones that are likely to be generated

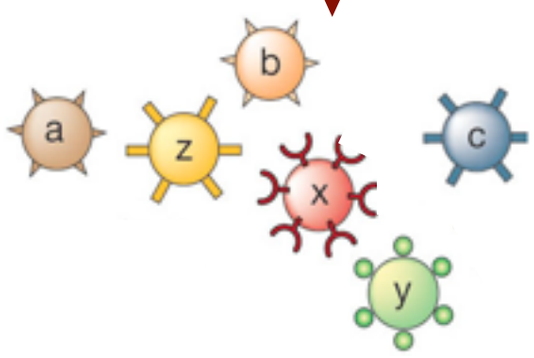
Receptor sharing

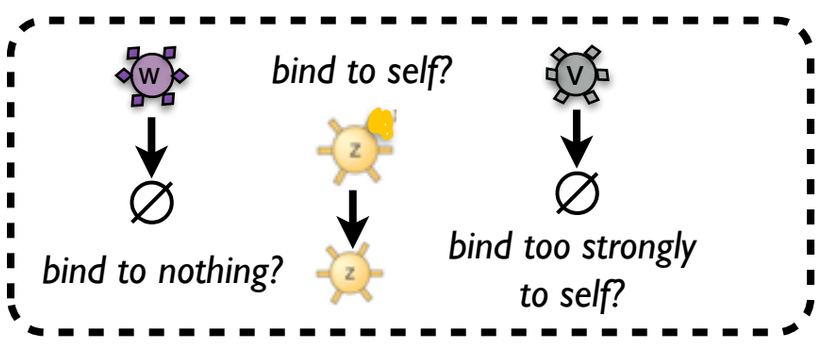


RECEPTOR GENERATION

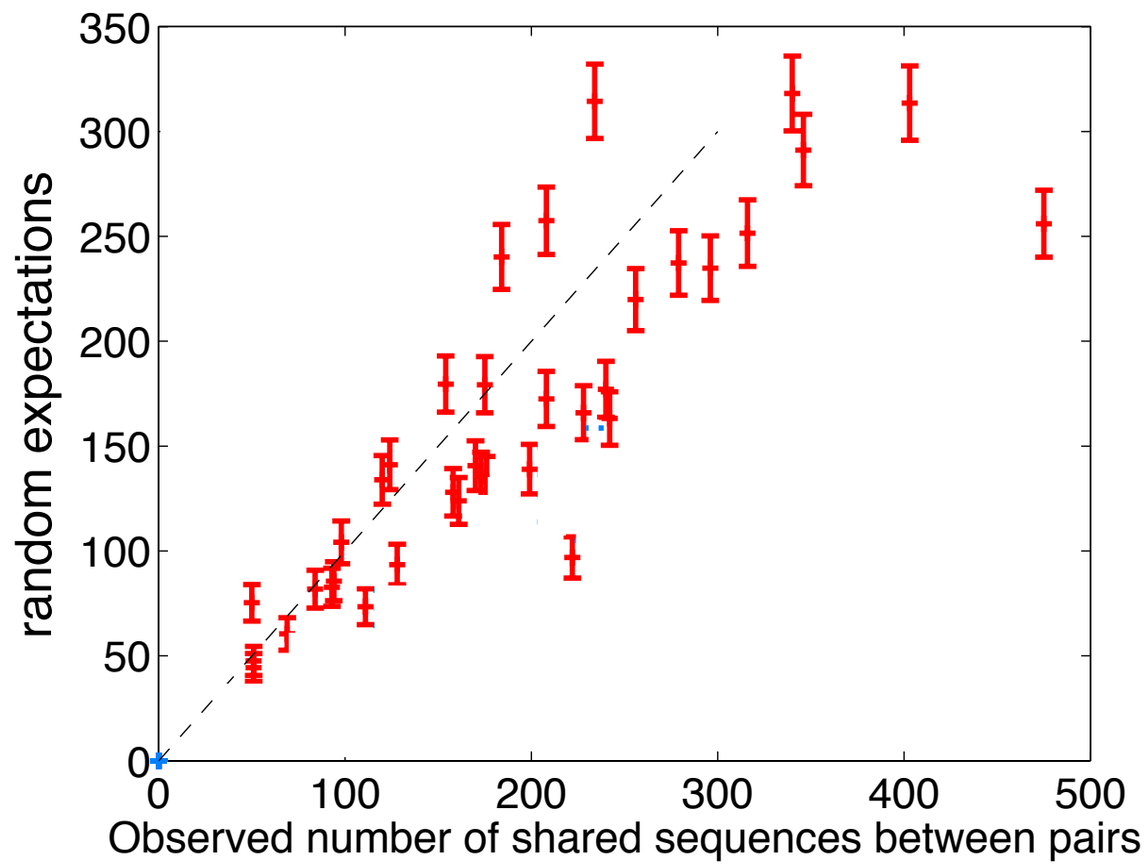


THYMIC SELECTION



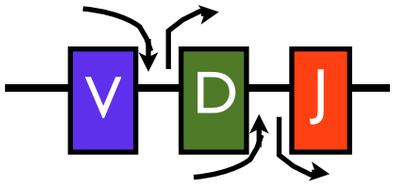


how many shared receptors between 2 people?

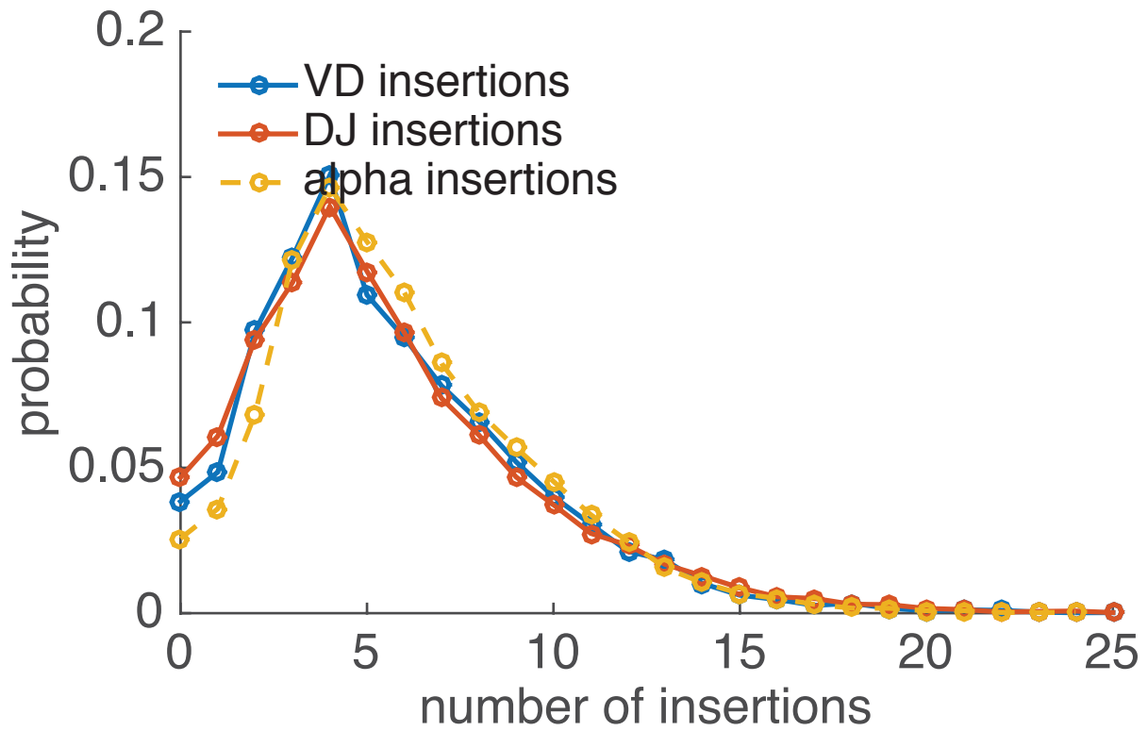


close to random expectations

Other datasets: alpha chains

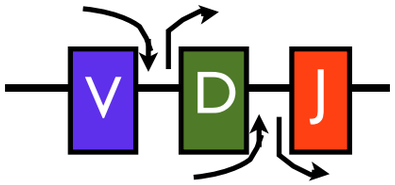


- we can do the same for the alpha chain
- similar insertion profile as beta chain

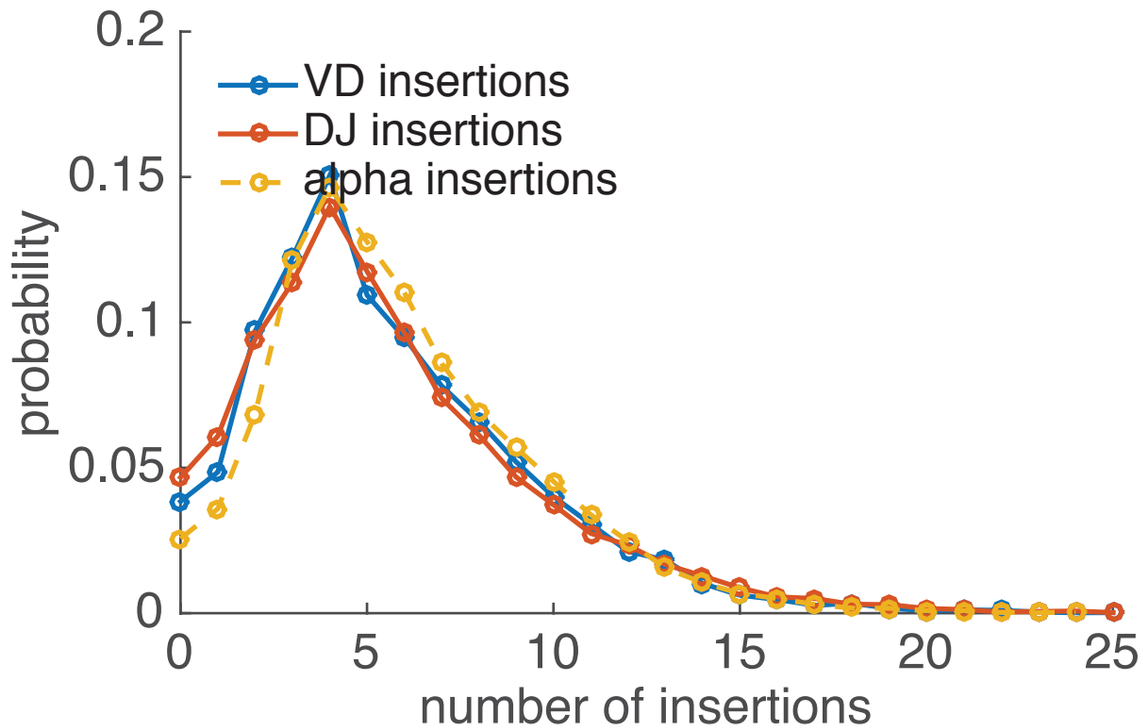


Elhanati Marcou Mora Walczak arXiv 2015

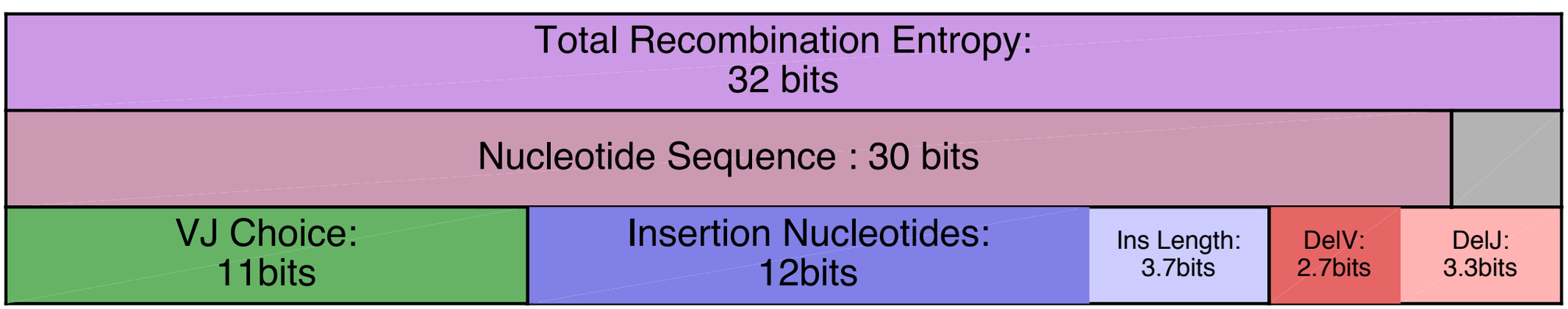
Other datasets: alpha chains



- we can do the same for the alpha chain
- similar insertion profile as beta chain

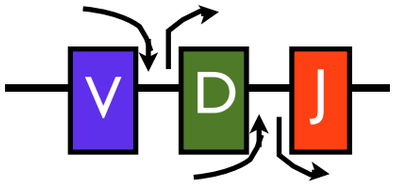


- entropy: 30 (alpha) + 47 (beta) = 77 bits ~ $N = 10^{23}$

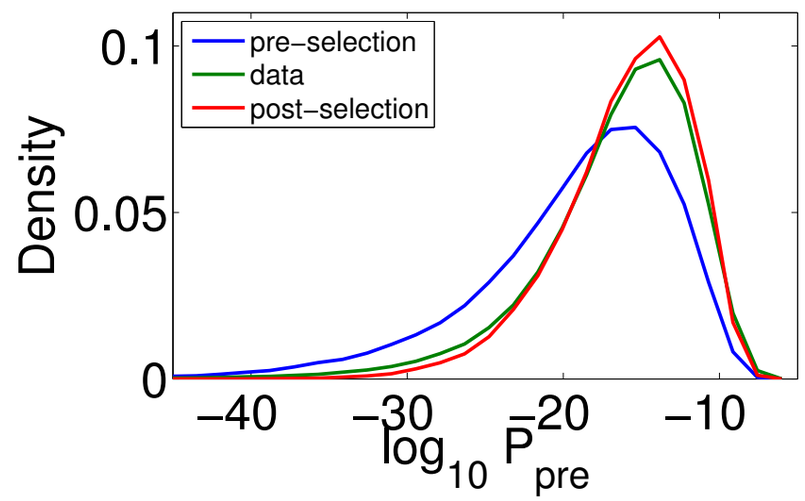


Elhanati Marcou Mora Walczak arXiv 2015

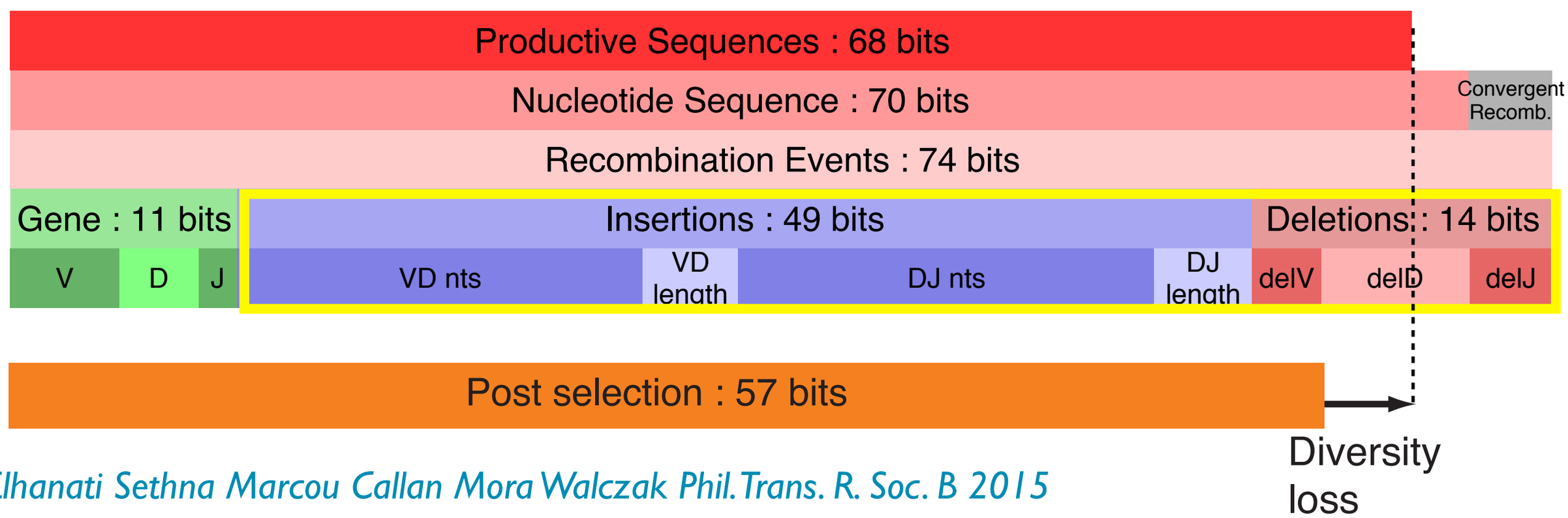
Other datasets: BCR



- we can do the same for B cell receptors: heavy chain
- analyse out-of-frame sequences from naive and memory B cells

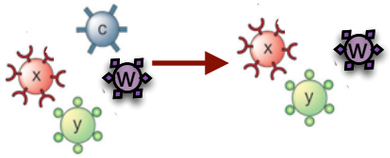


- entropy



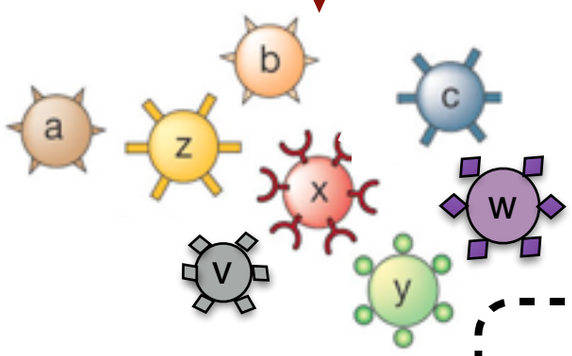
Elhanati Sethna Marcou Callan Mora Walczak Phil. Trans. R. Soc. B 2015

Repertoire evolution

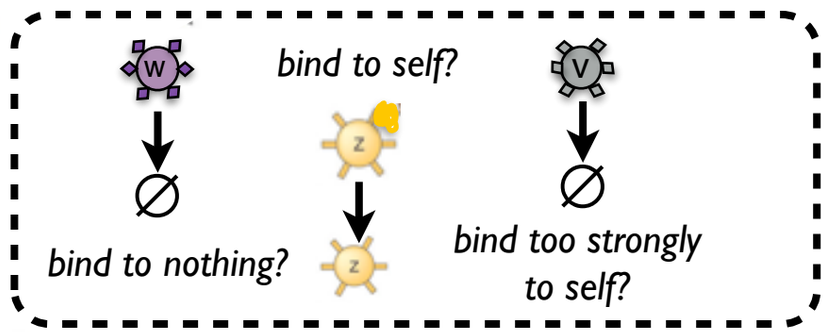


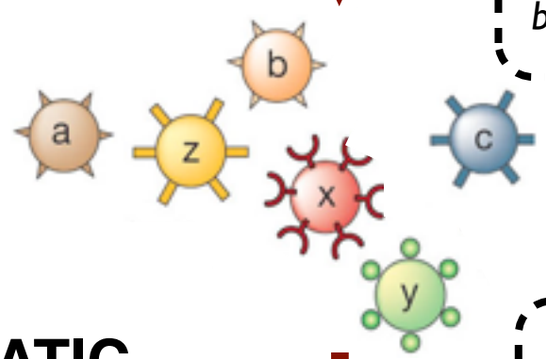
RECEPTOR GENERATION

VDJ recombination

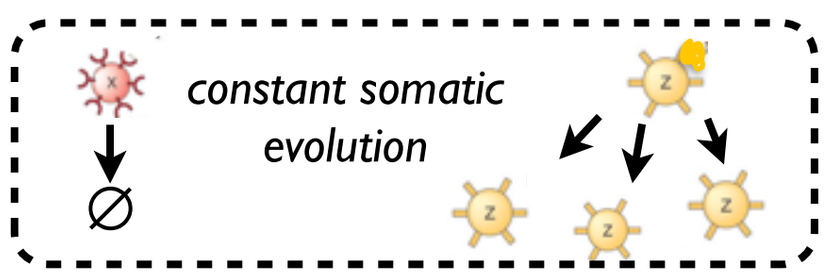


THYMIC SELECTION

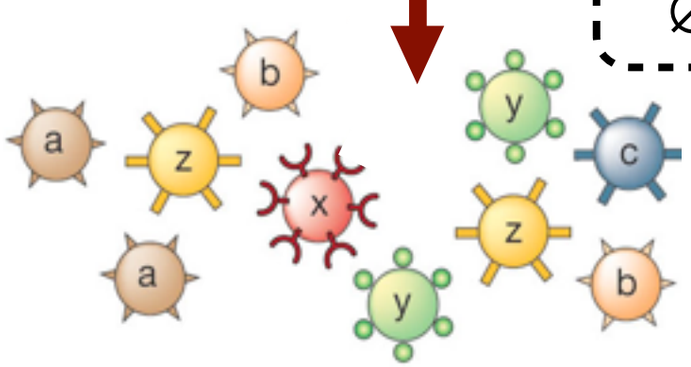




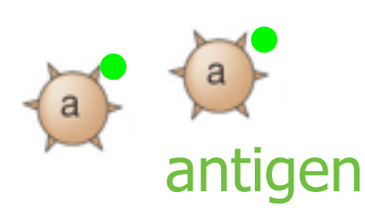
SOMATIC SELECTION



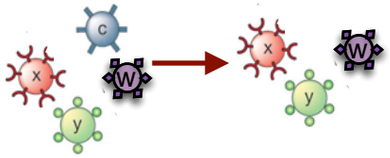
+ Somatic Hypermutations



SPECIFICITY

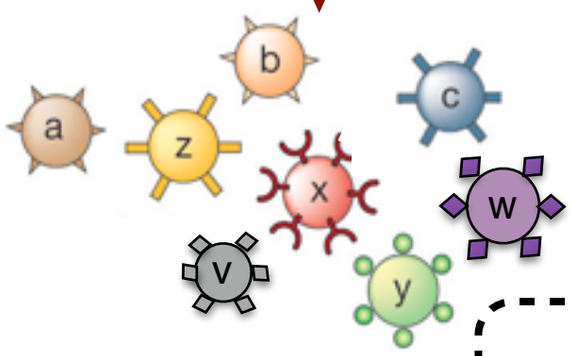


Repertoire evolution

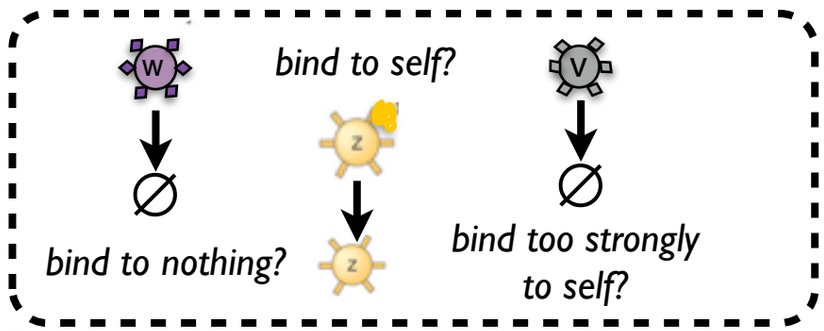


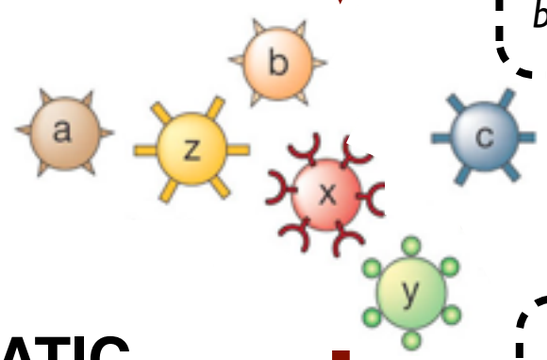
RECEPTOR GENERATION

VDJ recombination

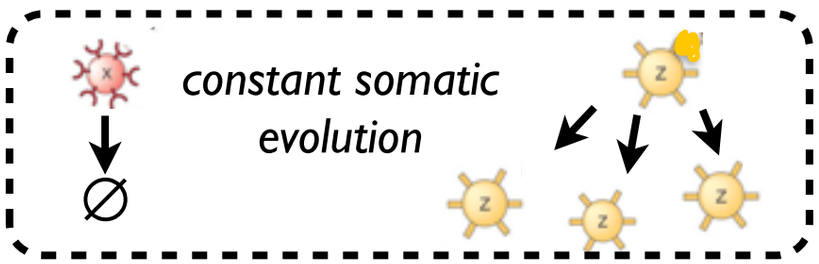


THYMIC SELECTION

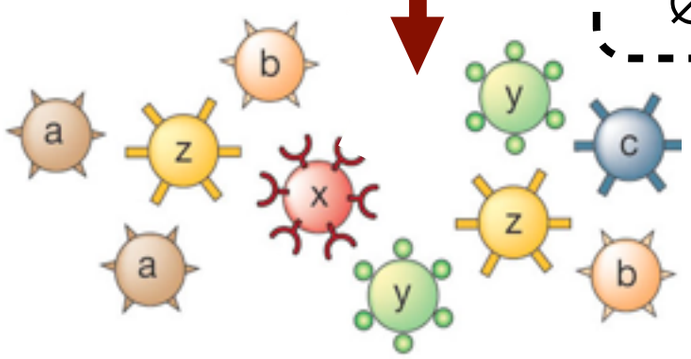




SOMATIC SELECTION

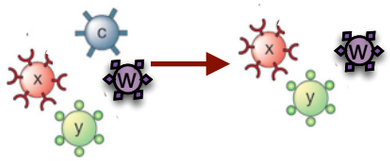


+ Somatic Hypermutations



SPECIFICITY

Somatic hypermutations

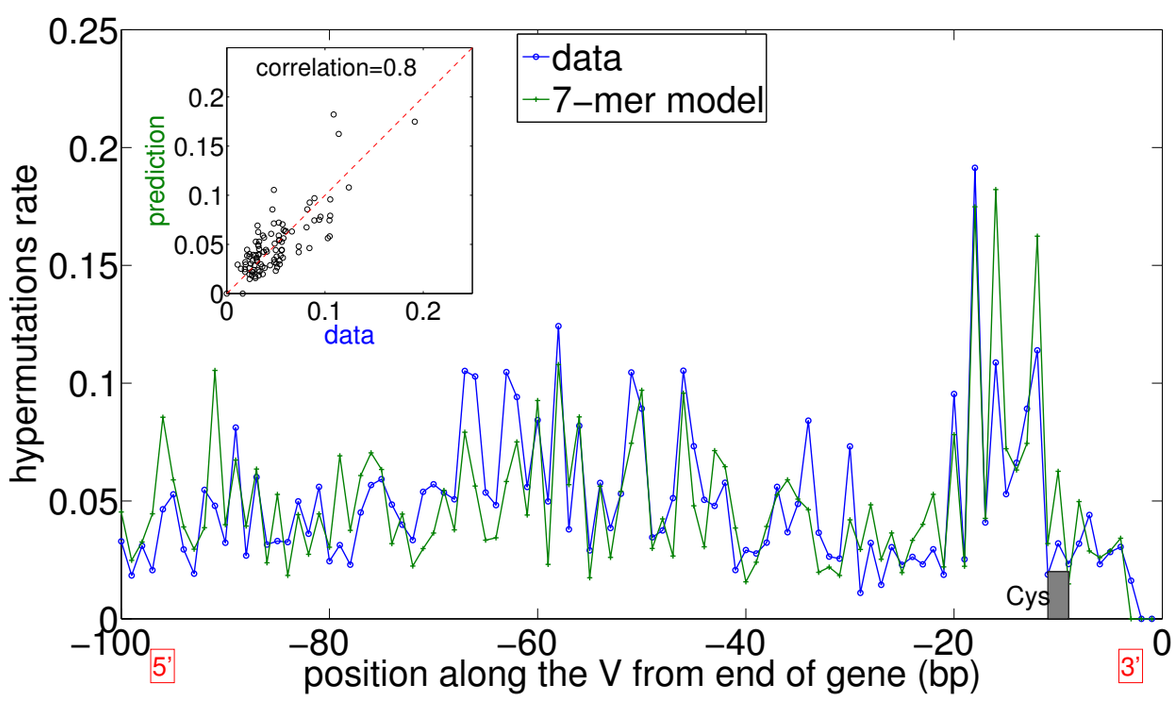


- use out-of-frame sequences from memory B cells
- position-weight matrix model hypermutation hotspots

7-mer

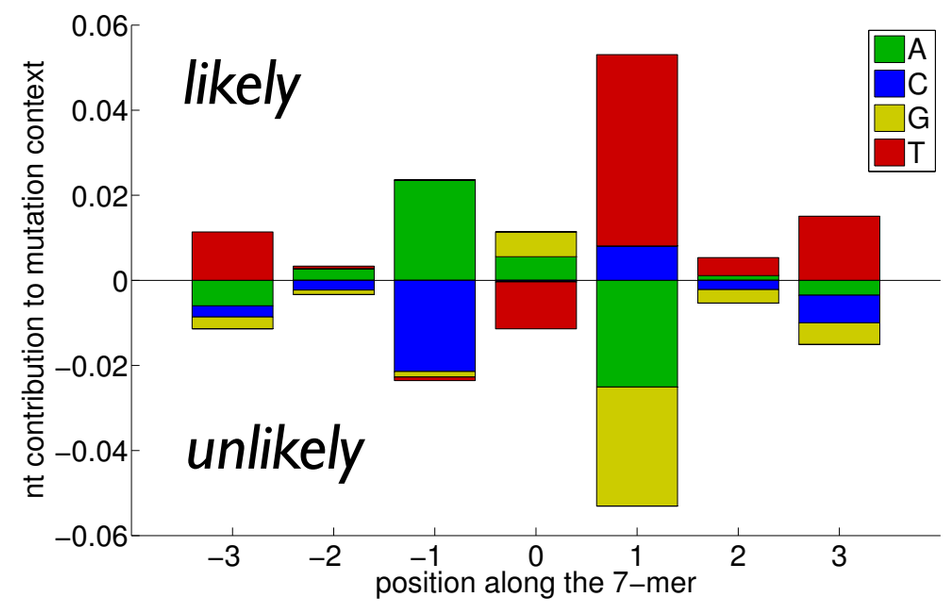
$$p_{\mathrm{SHM}}(\sigma) \propto p_{\mathrm{bg}}(\sigma) \exp[\sum_{i=-3,3} e_i(\sigma_i)]$$

background probability
of that 7-mer



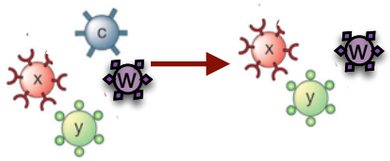
- nt that are likely to hypermutate

hypermutation hotspot 7-mer signature

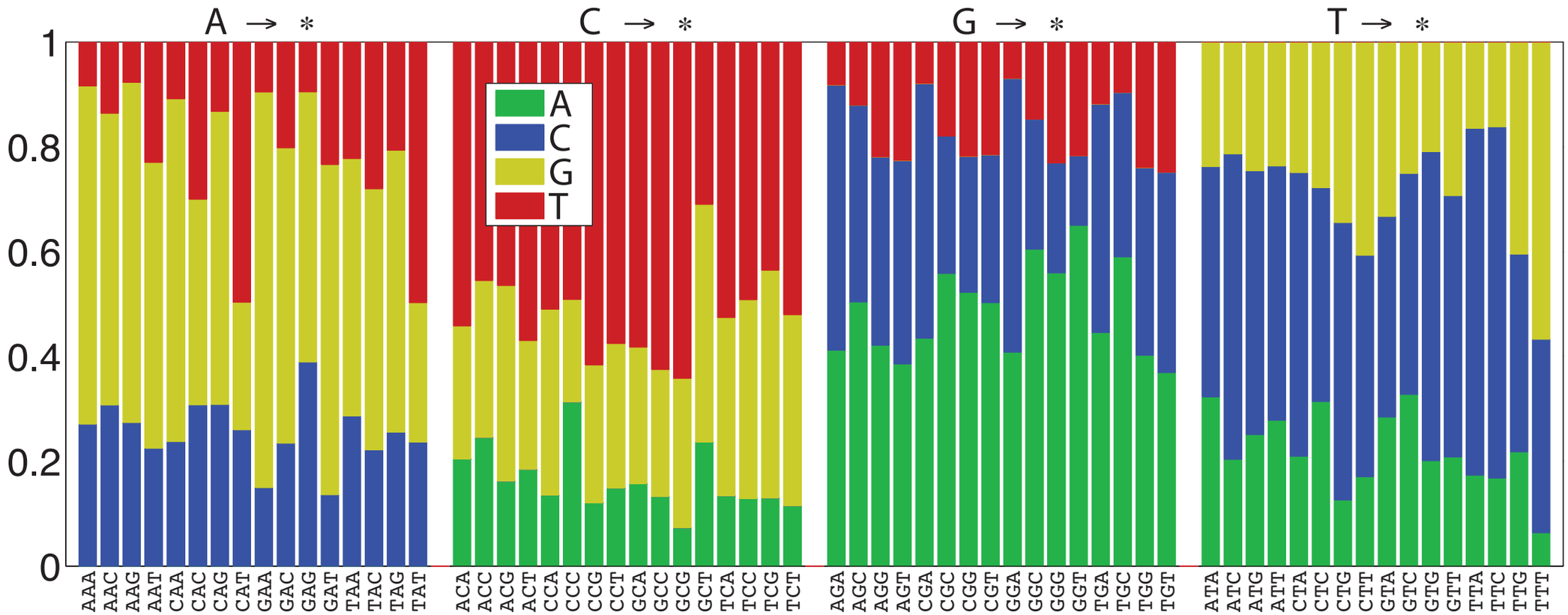


Elhanati Sethna Marcou Callan Mora Walczak Phil. Trans. R. Soc. B 2015

Somatic hypermutations

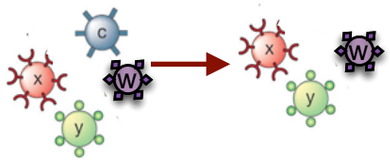


- nt *to which* a nt mutates

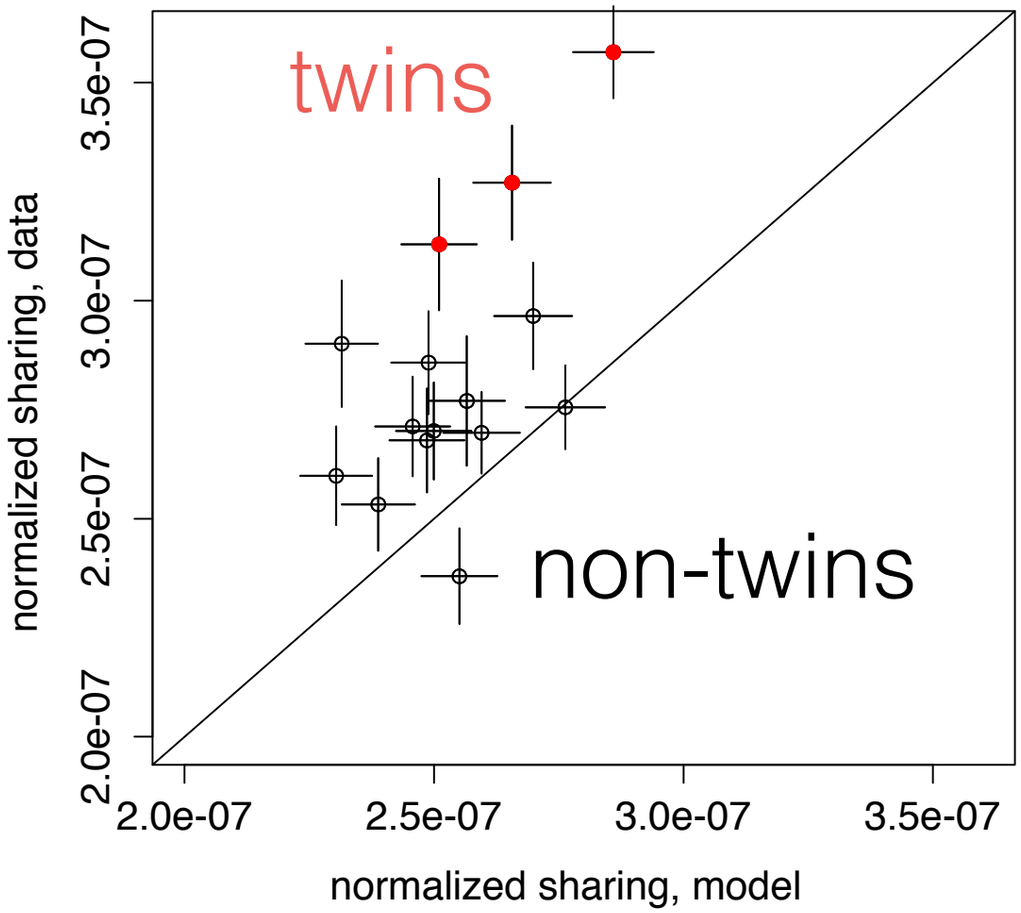


Thank you

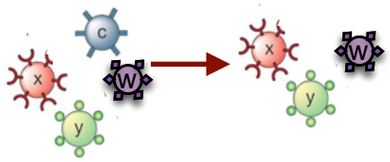
Is everyone the same?



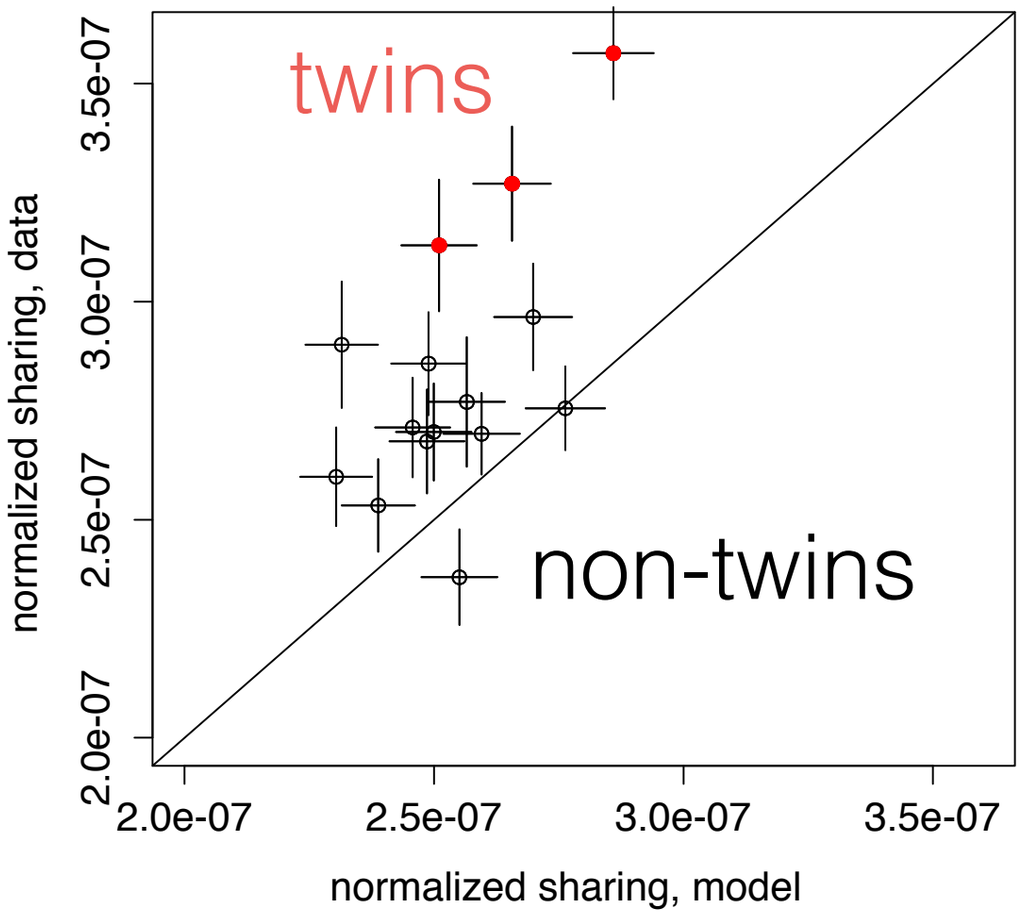
at the level of generation



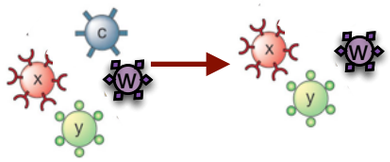
Is everyone the same?



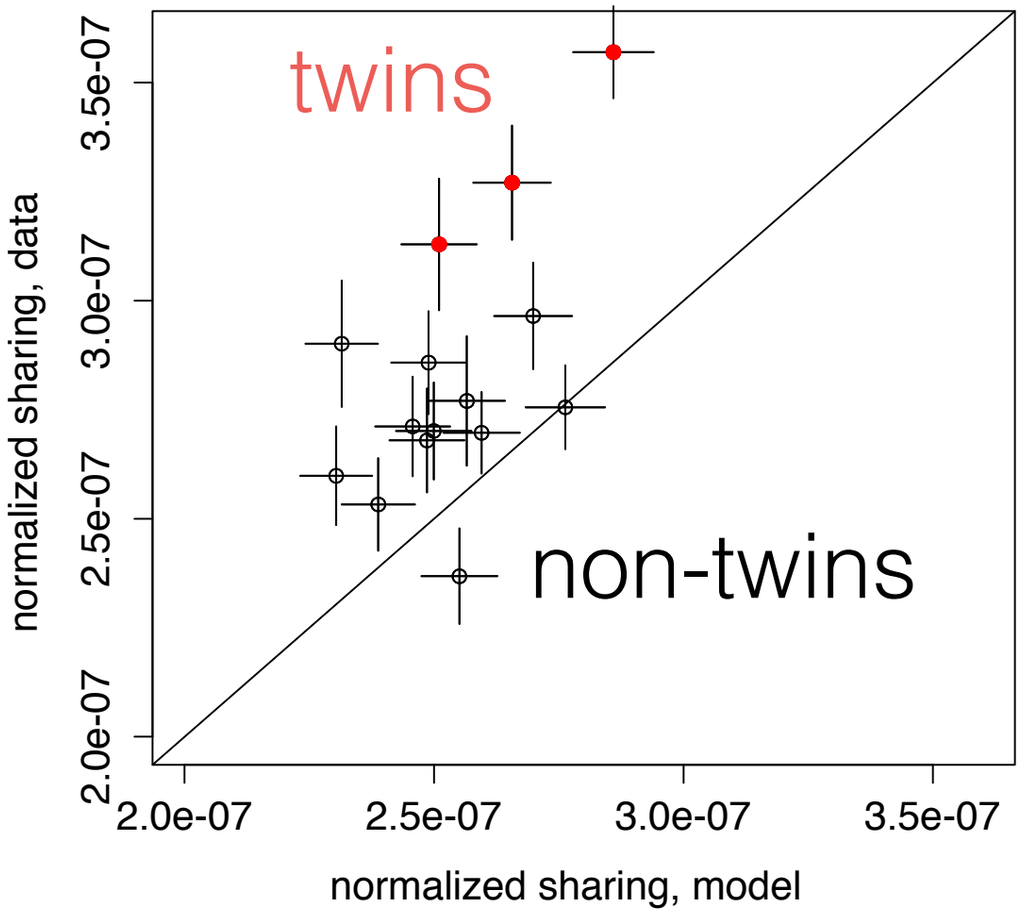
at the level of generation → twins are special

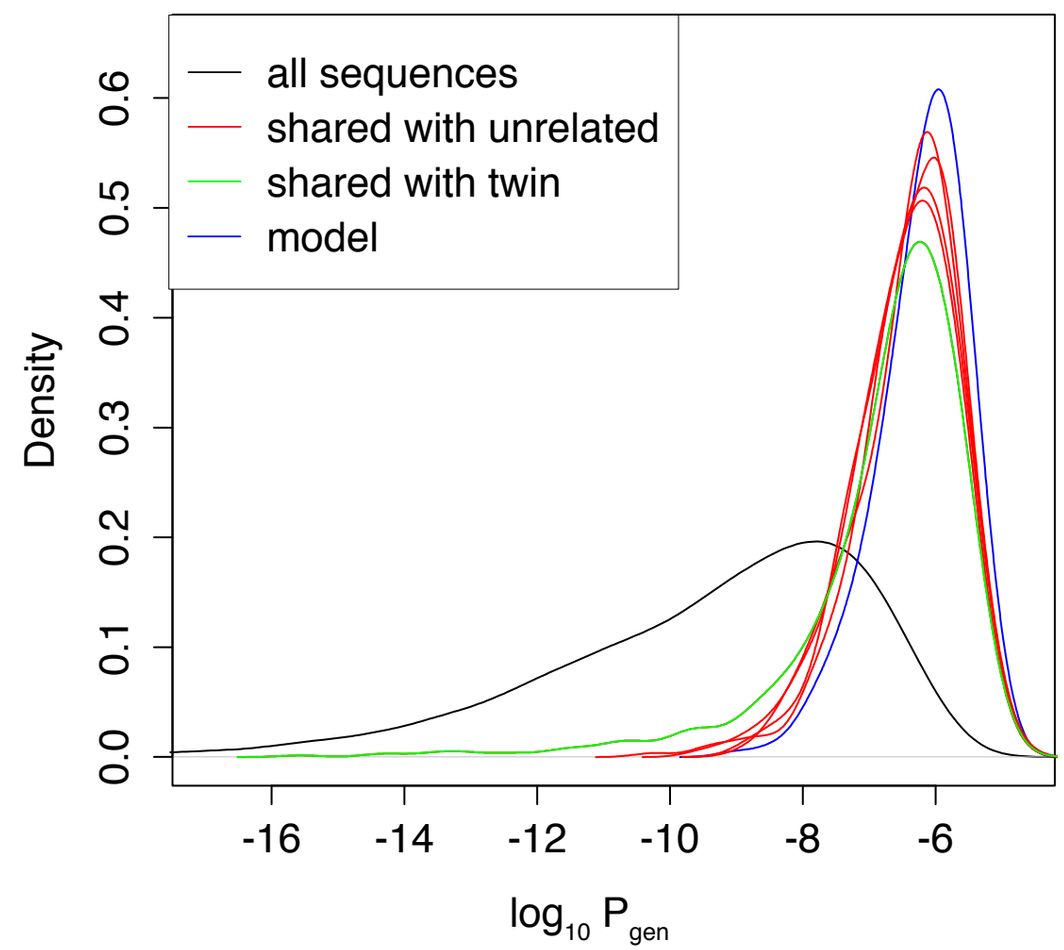


Is everyone the same?



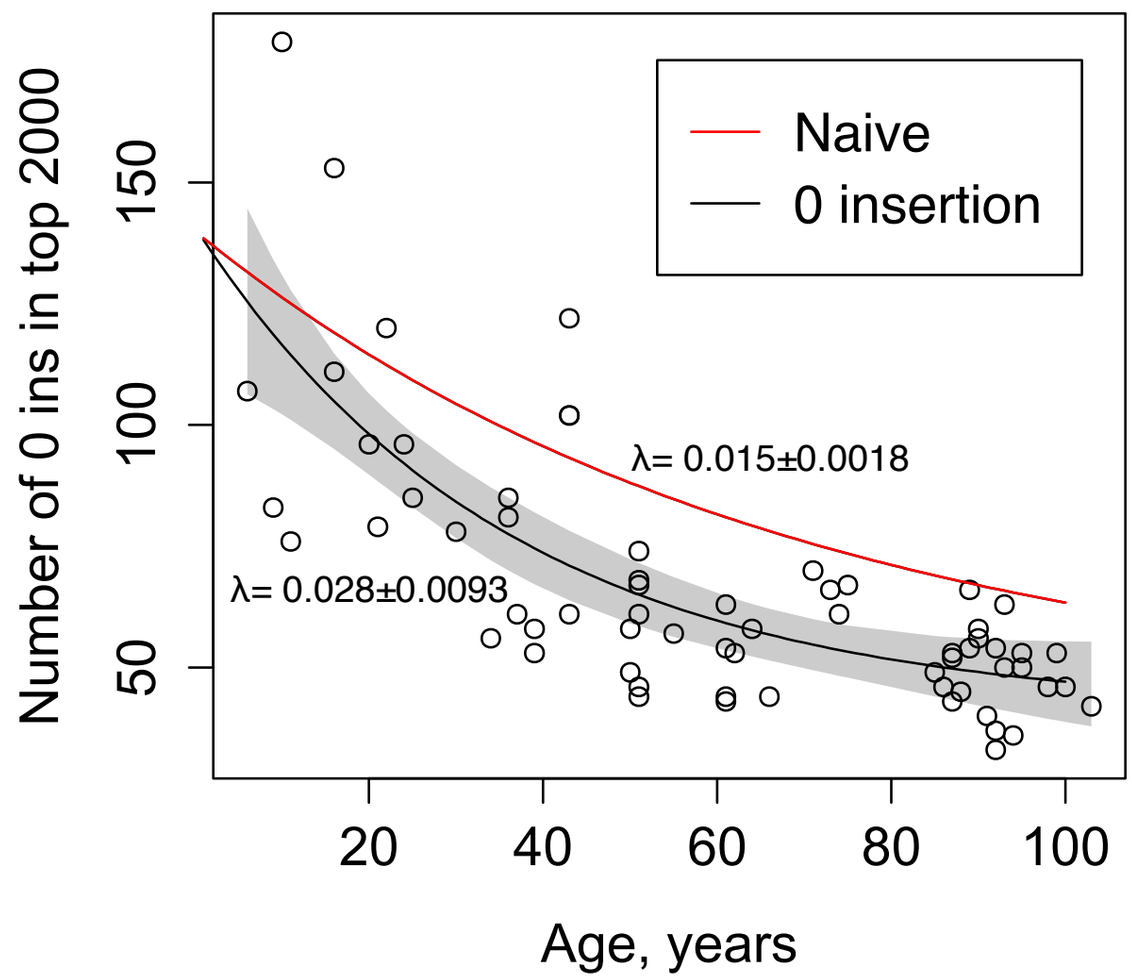
at the level of generation → twins are special



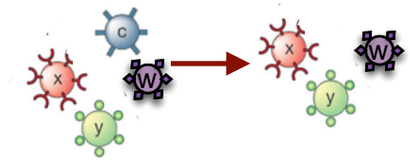


The source: long lived sequences

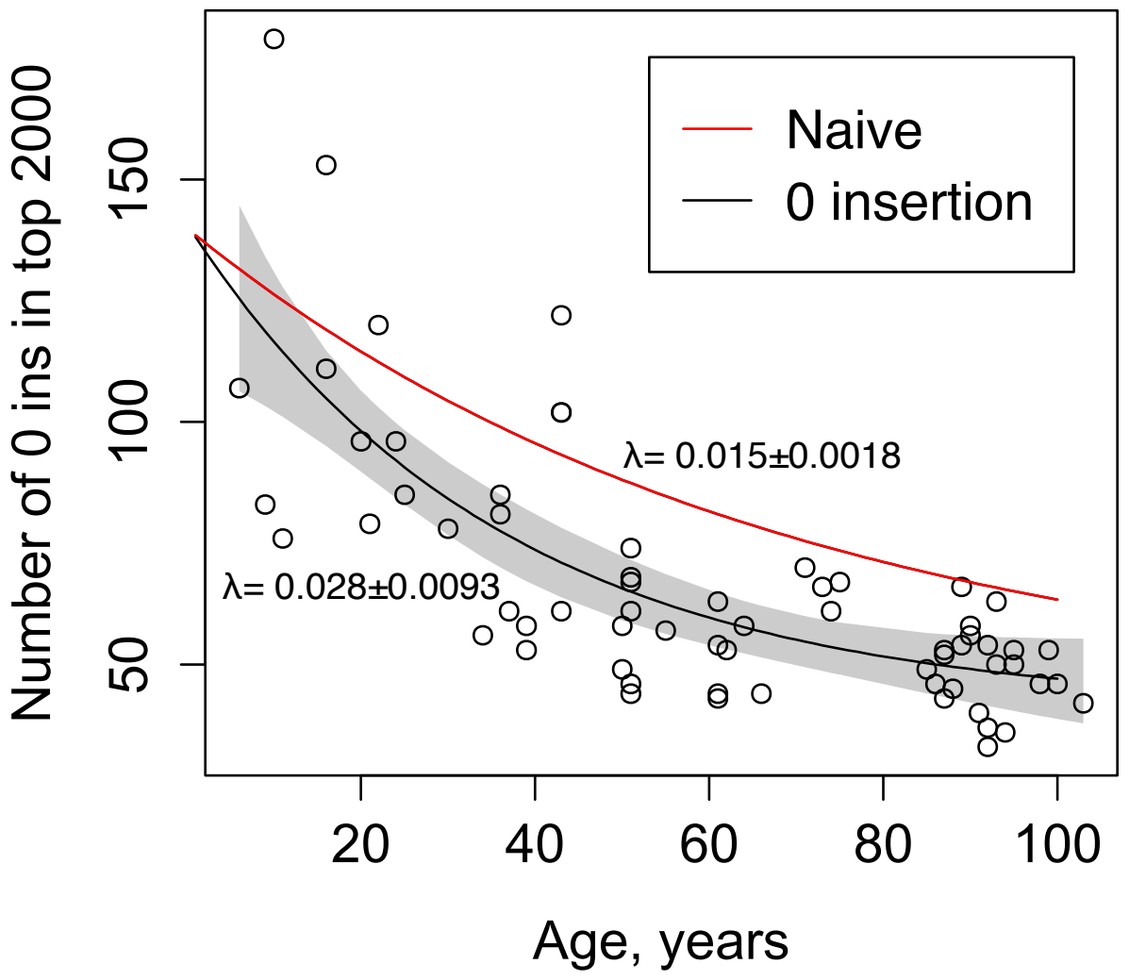
- last time twins shared blood: before birth
- insertions enzyme less active before birth



The source: long lived sequences

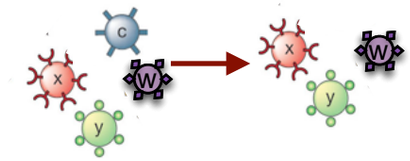


- last time twins shared blood: before birth
- insertions enzyme less active before birth

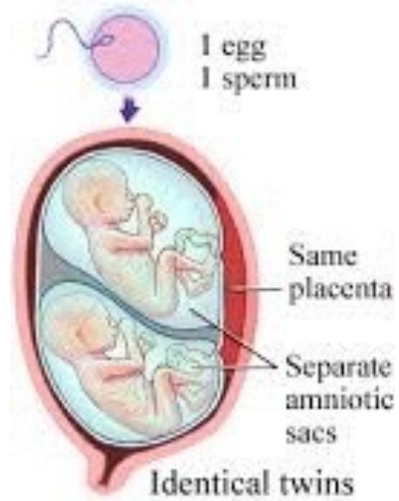


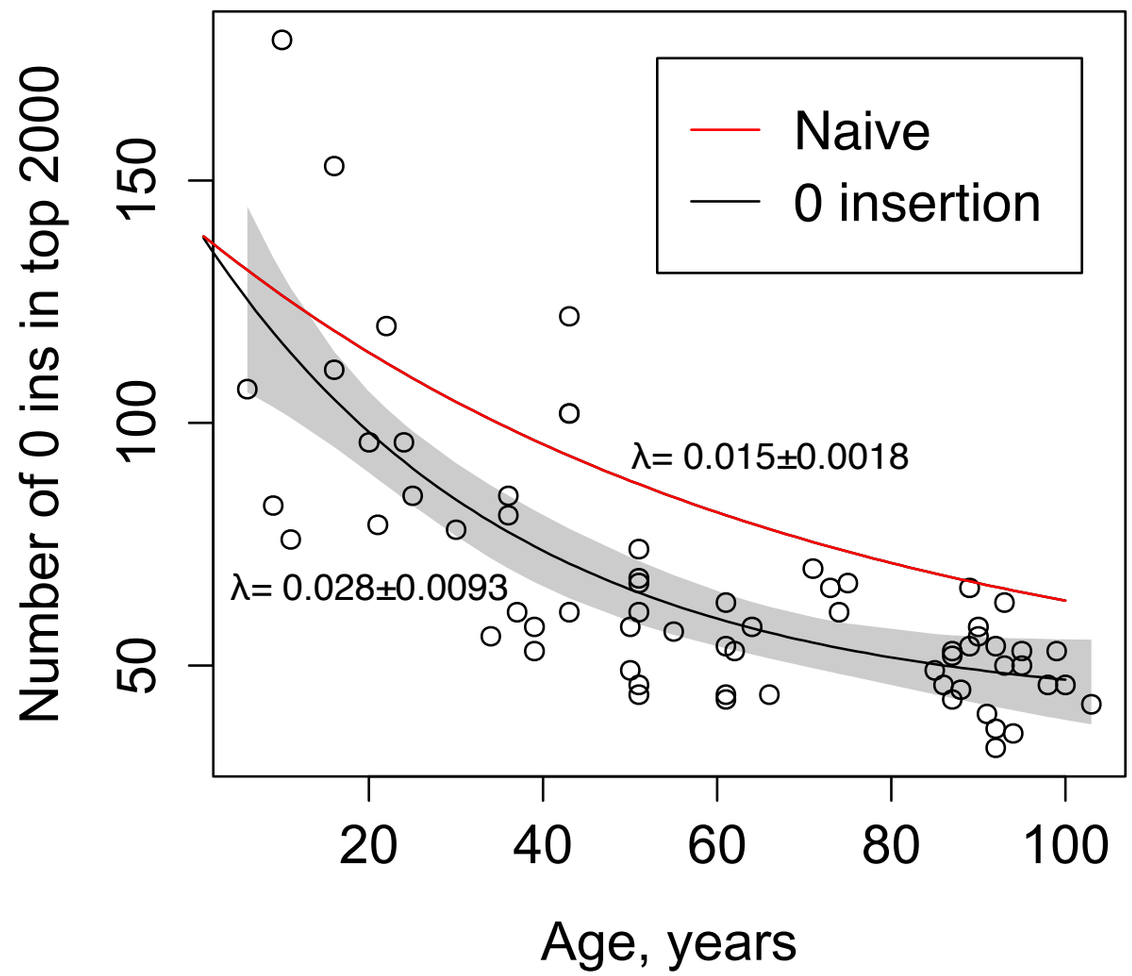
clone lifetime ~ 36 years

The source: long lived sequences

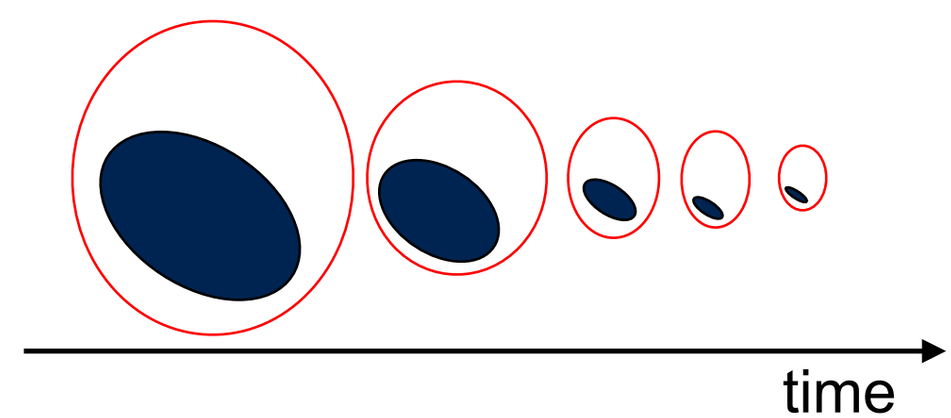


- last time twins shared blood: before birth
- insertions enzyme less active before birth





clone lifetime ~ 36 years



Decay of zero insertion clonotypes:

- zero insertion clonotypes within the naive pool
- size of the total naive pool