

Using sequence data to infer the mechanisms of T-cell diversity generation

Curtis G. Callan, Jr.
Physics Department,
Princeton University

The DNA of the cells of the adaptive immune system have undergone stochastic gene editing to provide the diversity needed to deal with pathogens. Copious data on this diversity are now being provided by high-throughput sequencing. Making sense of this data and putting it to productive use requires a coherent statistical inference framework. I will report on some first steps in this direction, with application to sequence data on T-cells taken from human individuals.

Work with A. Murugan (Stanford), T. Mora (ENS) and A. Walczak (ENS).
Data kindly provided by H. Robins (FHCRC) and A. Levine (IAS)

Learning Probability Distributions

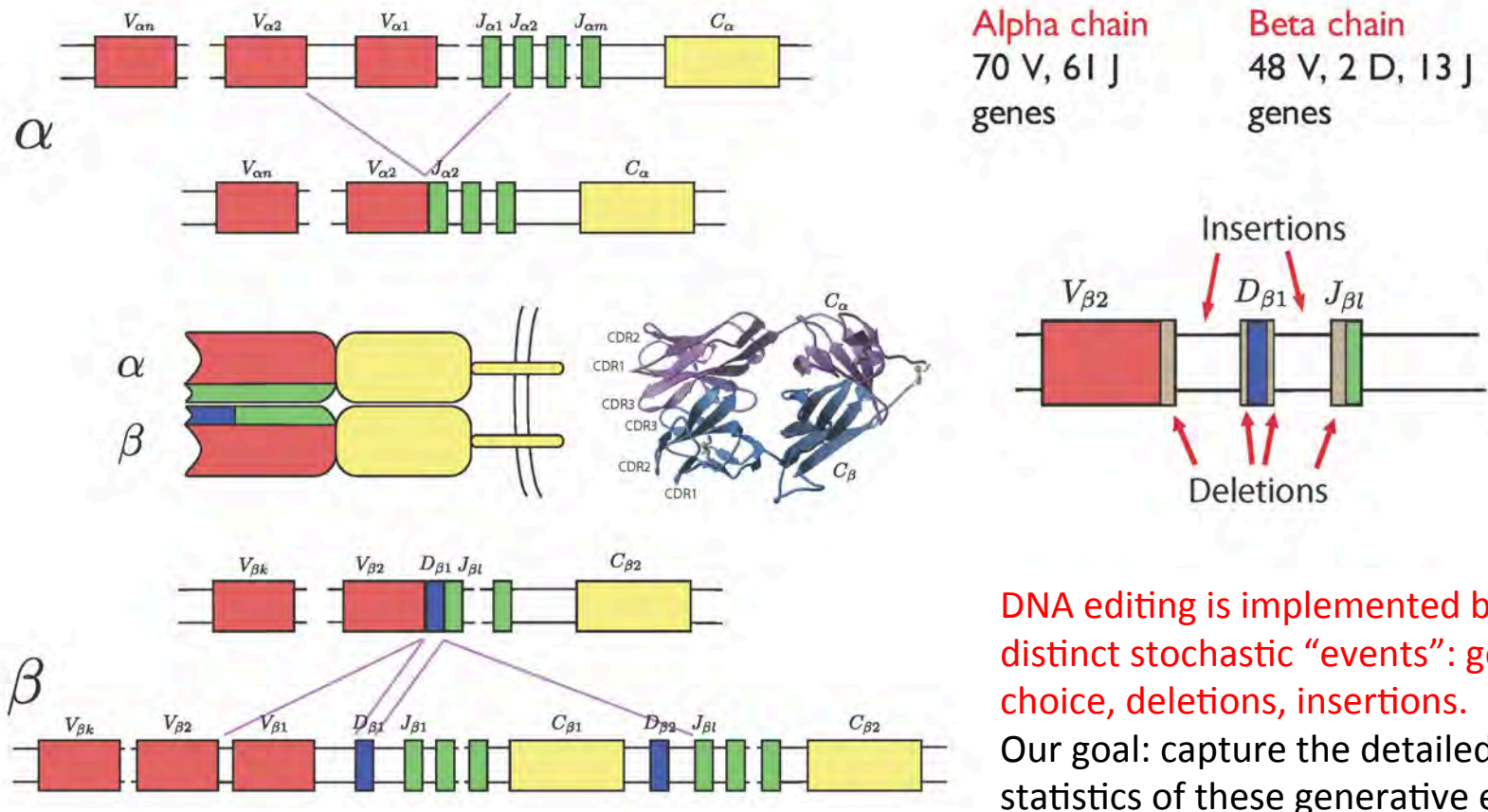
- In biology what matters is often a probability distribution:
 - Response of retinal ganglion cells to repetitions of the same visual scene
 - The amino acid sequence of a family of proteins of similar function
 - The body shape and proportions of an individual of your favorite species
- The distributions are high-dimensional; potentially very high entropy
 - Neurons can spike every 10 ms or so, 10s of ganglions carry info about same “pixel”
 - A functional protein motif can be 100s of aa long, with 20 choices per position
- There is no way to “learn” these distributions by exhaustive sampling
 - Big issue for sensory systems which must distinguish “surprising” from “normal”
 - Can’t do it without some built-in restricted model of the probdist to be learned
- Immune system diversity provides an instructive concrete example
 - Your body contains $\sim 10^7$ different T-cell genotypes (to fight pathogen diversity)
 - The stem cell process potentially can spit out $\sim 10^{14}$ different T-cells (wait for it!)
 - Can we learn the T-cell generation probdist by sequencing a few individuals?
- Yes if we are suitably modest in our ambitions

Overview of the life of a T-cell

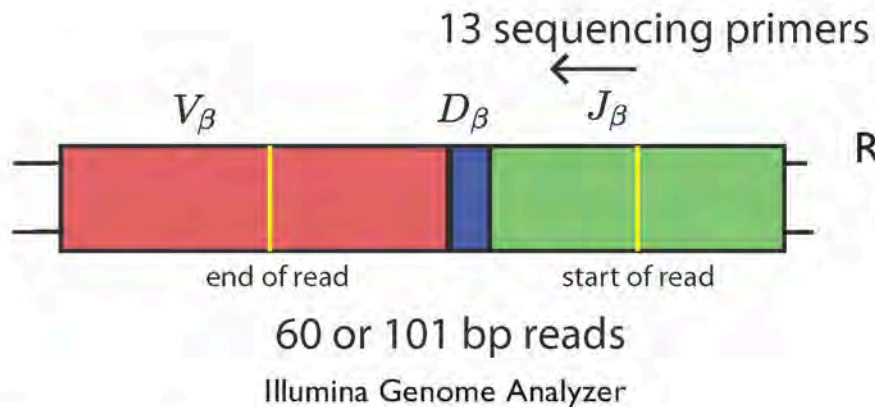
- Key player of the adaptive immune system. Has surface receptors to recognize foreign protein fragments “presented” by other cells.
- Many different T-cell types are in your blood at any one time; T-cells that “see” a pathogen proliferate and orchestrate an attack on the invader.
- New T-cells are made in the bone marrow and germline DNA for the receptor gene is randomly “edited” to create a unique new proto T-cell.
- The T-cell migrates to the thymus where it is tested against “self” protein fragments. Proto T-cell dies if recognition of self is too weak, or too strong.
- The fully vetted T-cell then enters the blood stream and is authorized to respond to foreign protein presentation on the surface of host cells.
- T-cells that recognize pathogen proliferate and, when the crisis is over, leave a small clone of “memory” cells to speed response to re-infection.
- The diversity and randomness of this system means that statistical approaches are needed to understand it. Very much a work in progress.

Receptor Diversity from Stochastic Genome Editing

“VDJ Recombination” of germline DNA produces a unique TCR gene in each new T cell created in the bone marrow. This amazing process is implemented by a suite of DNA repair enzymes.



Output of VDJ rearrangement can be “observed” via high-throughput sequencing technology: (Harlan Robins, FHCRC)



Select T-cells from blood, extract DNA,
amplify relevant CDR3 region, sequence

Robins, H. et al. Comprehensive assessment of T-cell
receptor beta-chain diversity in alpha-beta T cells.
Blood 114, 4099–4107 (2009)

Nine individuals

Naive cells : ~230,000 unique CDR3 sequences per individual

Memory cells : ~140,000 unique CDR3 sequences per individual

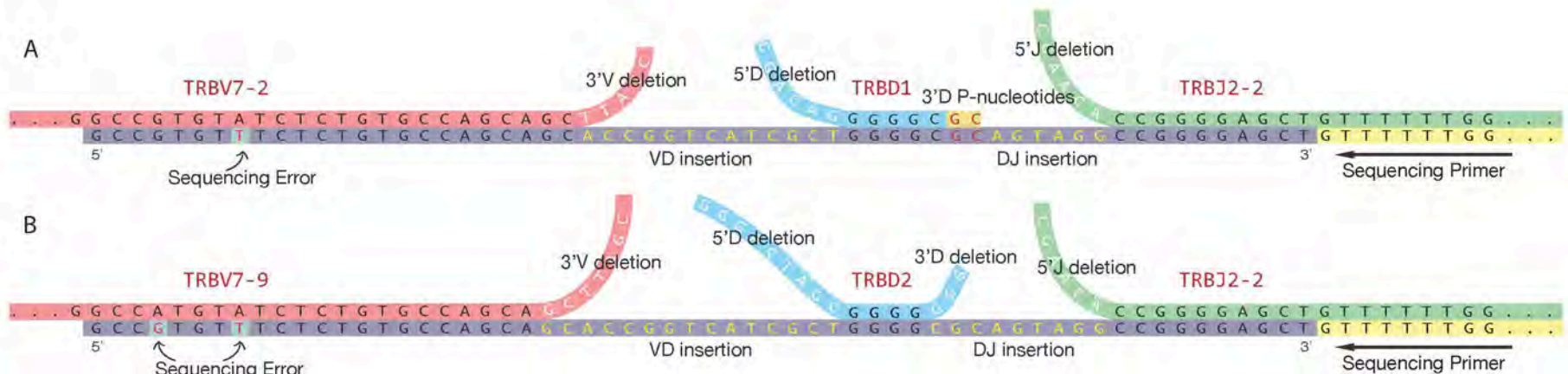
Non-productive sequences : 14%

Naive : ~35,000 per individual

Memory : ~22,000 per individual

These sequences are “fossils” of failed VDJ edits on one chromosome. The system tries again on the other and, if successful, the fossil DNA goes along for the ride, unselected for function. Direct window on the primitive T-cell generative process.

Same TCR sequence can be produced many ways



Goal: infer the distribution of generative event variables (V,D,J, ...) from sequence repertoires.

We can't do this without some hypothesis for the structure of the distribution: too high-dim'l to sample fully. Statistical inference technology to the rescue!

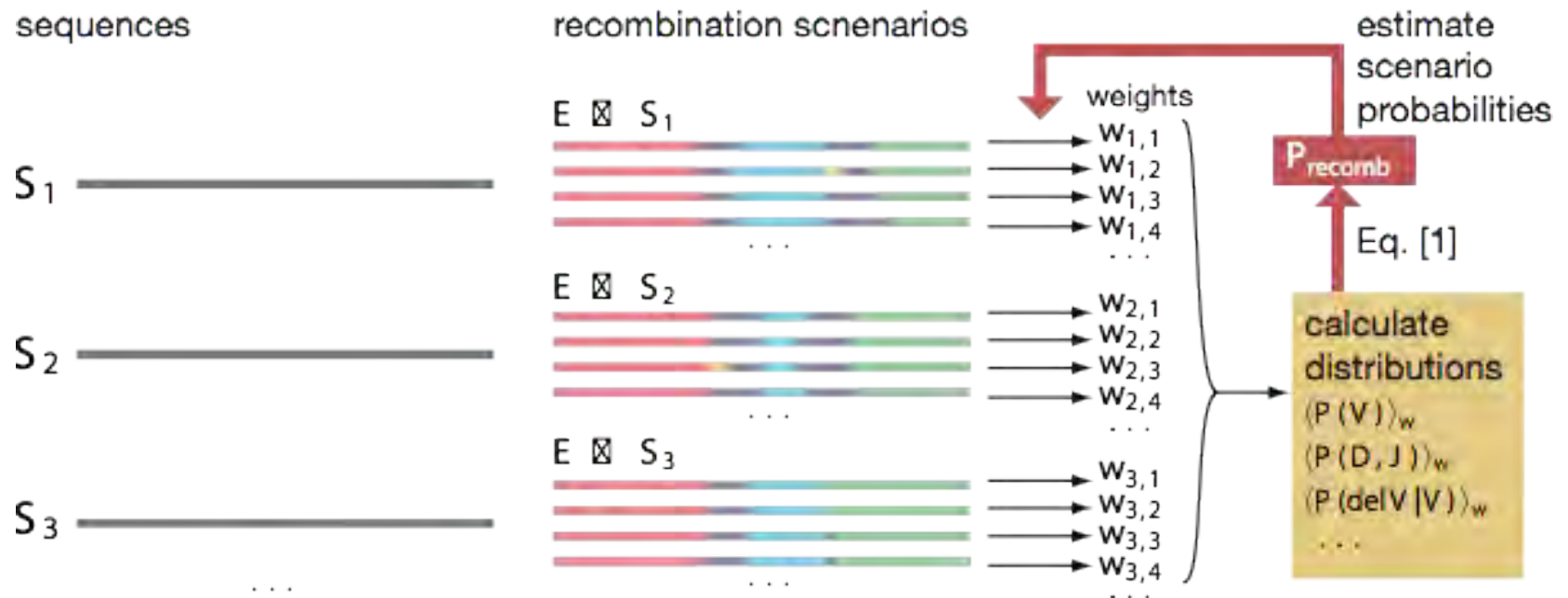
$$E_{CDR3} : \begin{cases} V, D, J \\ \text{del}V, \text{del}J, \text{del}D5, \text{del}D3 \\ \text{pal}V, \text{pal}J, \text{pal}D5, \text{pal}D3 \\ \text{ins}VD, \text{ins}DJ \\ (x_1, \dots, x_{\text{ins}VD}), (y_1, \dots, y_{\text{ins}DJ}) \end{cases}$$

$$P_{\text{gen}}(\sigma) = \sum_{E \in E_{\sigma}} P_{\text{recomb}}(E)$$

Infer the distribution by expectation maximization

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

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



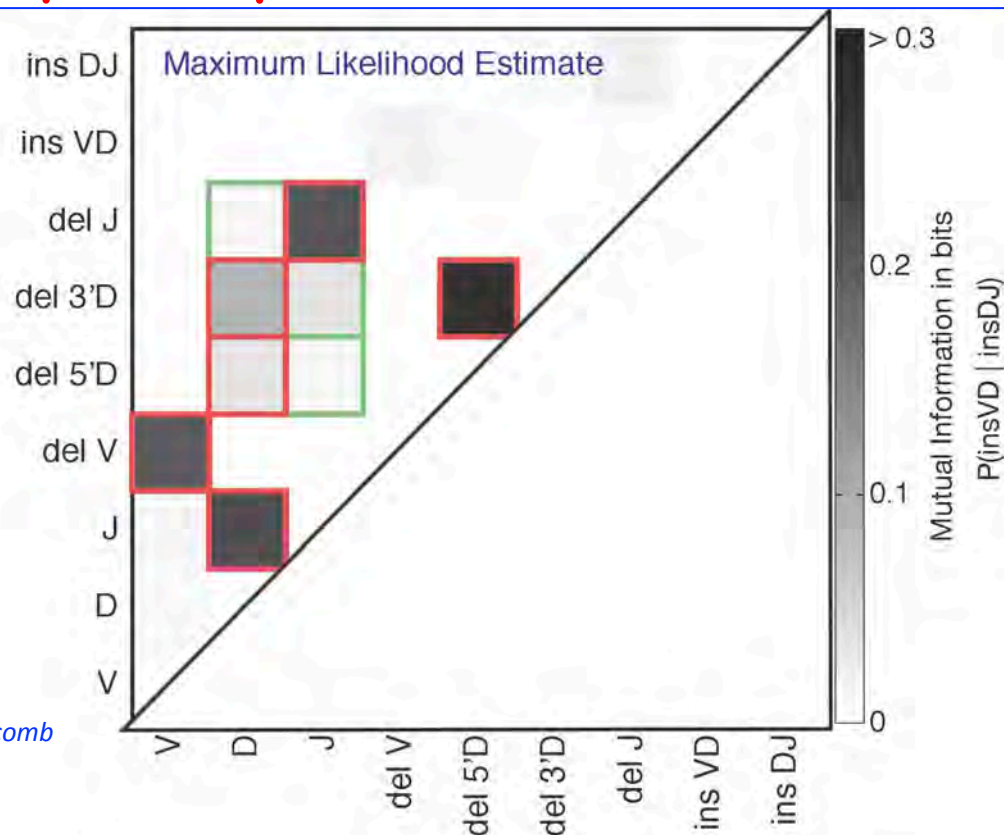
$$P^{\text{recomb}}(\text{scenario}) = P(V)P(D, J)P(\text{deletions } V|V)P(\text{insertions } DJ)\dots \quad [1]$$

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

Check that you capture all correlations in data!

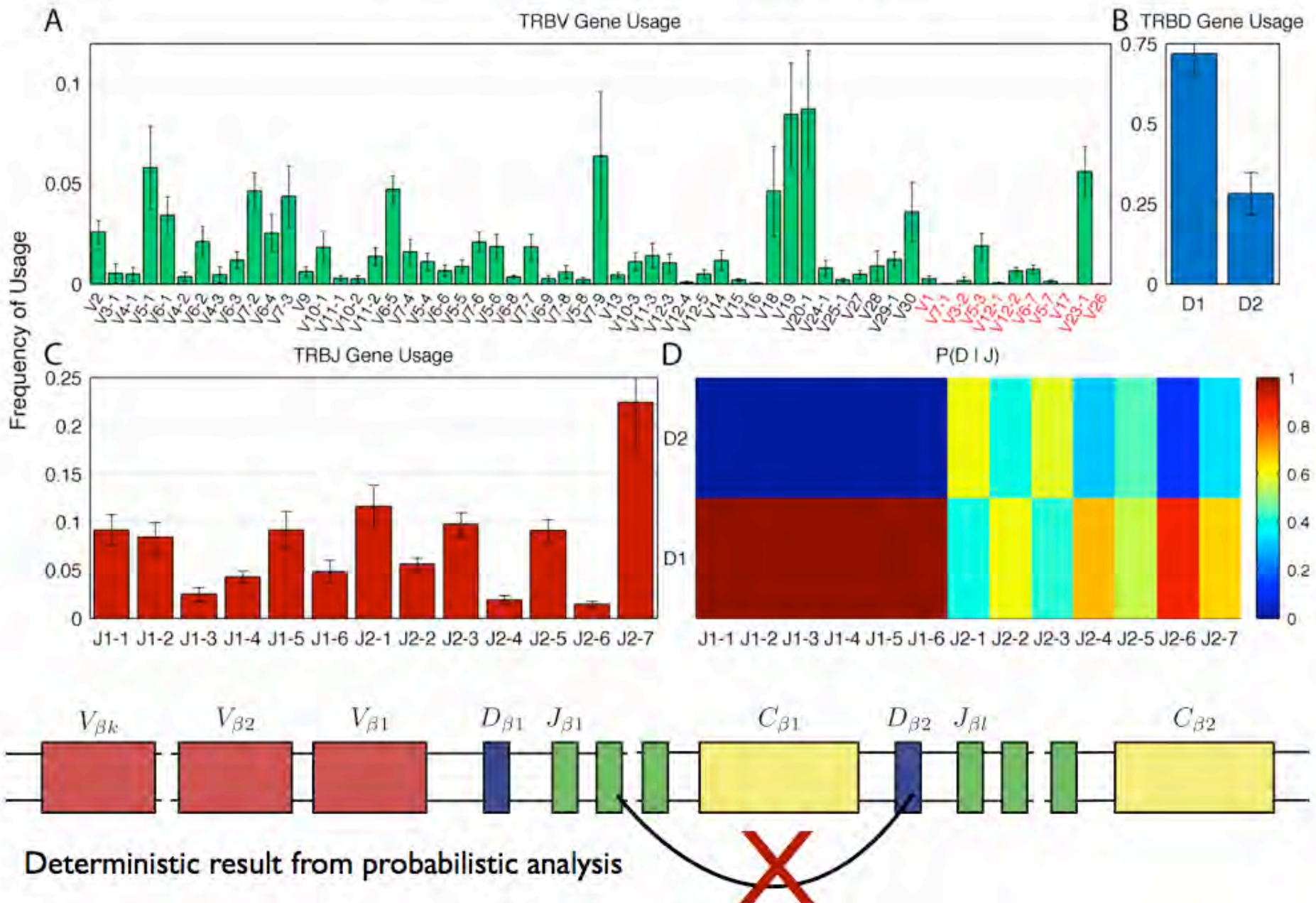
A “simple” model captures ~ all the correlations in the data Data from 9 diff't subjects imply the same model.

Some experimentation was needed to find the proper structure of P_{recomb}

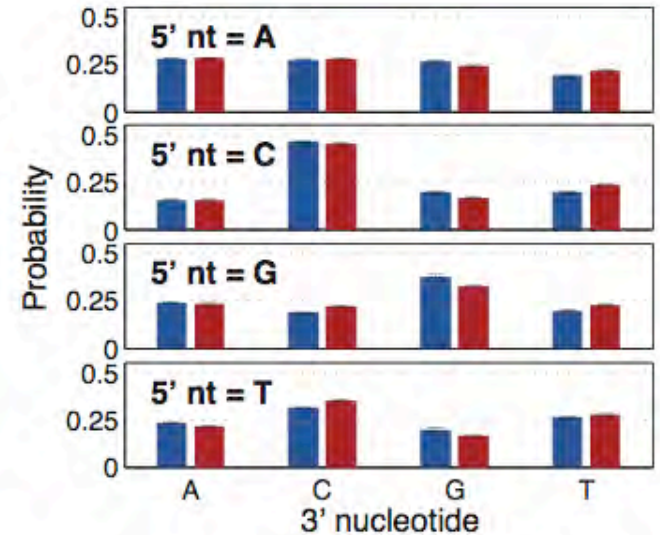
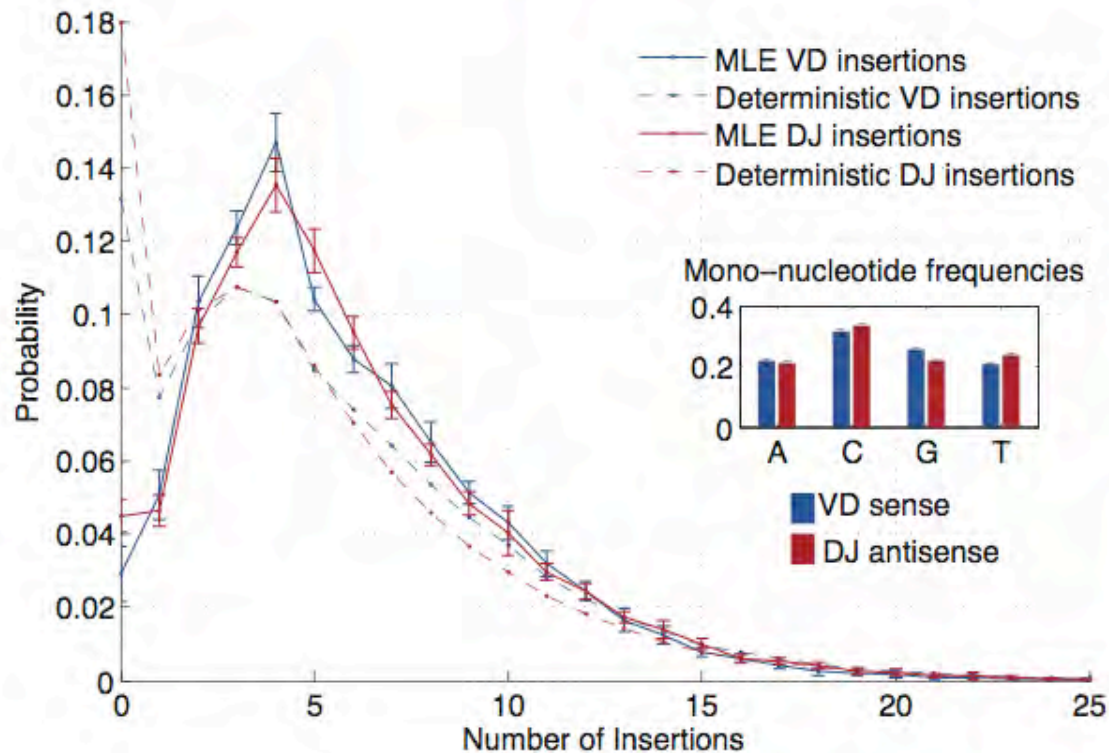


$$\begin{aligned}
 P_{recomb}(E) = & P(V)P(D, J) \times \\
 & P(\text{del}V|V) P(\text{del}J|J) P(\text{del}5'D, \text{del}3'D|D) \times \\
 & P(\text{ins}VD) \prod_{i=1}^{\text{ins}VD} p_{VD}^{(2)}(x_i|x_{i-1}) P(\text{ins}DJ) \prod_{i=1}^{\text{ins}DJ} p_{DJ}^{(2)}(y_i|y_{i-1})
 \end{aligned}$$

Gene Usage Distributions

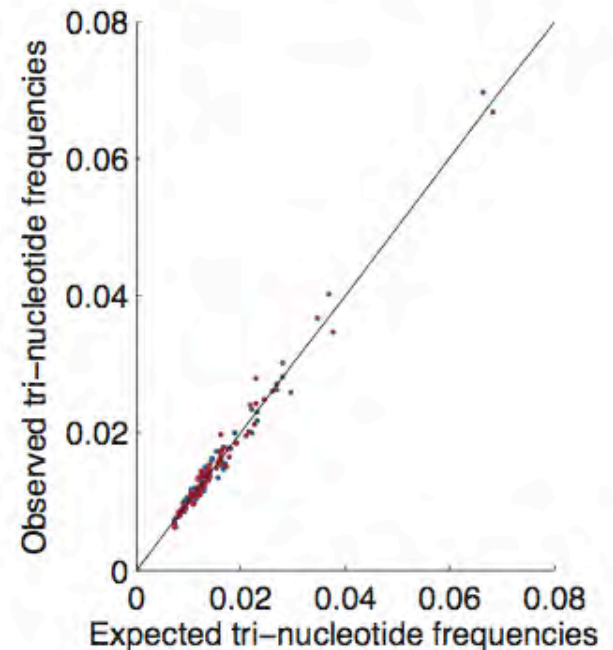


Insertions at VD and DJ junctions are identical and independent

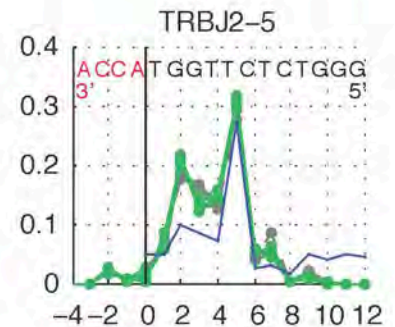
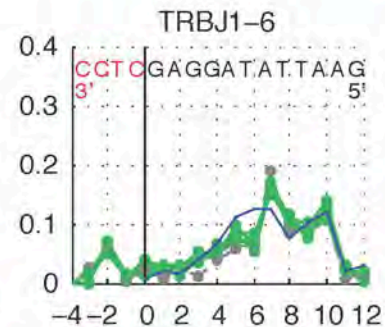
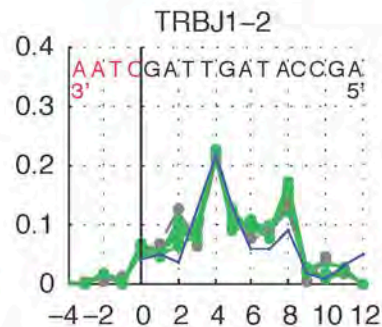
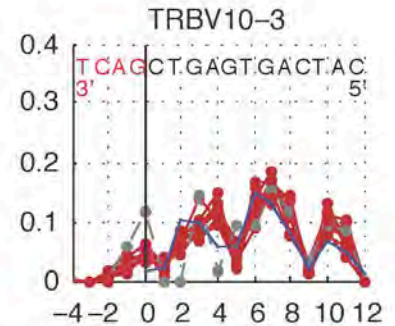
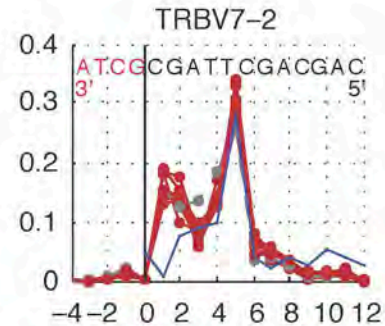
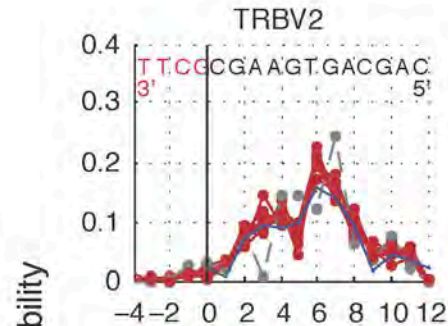
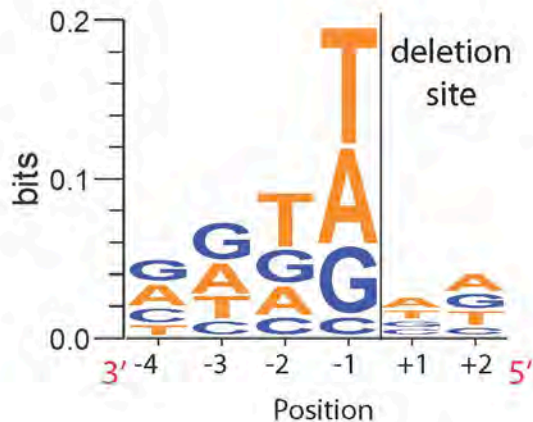
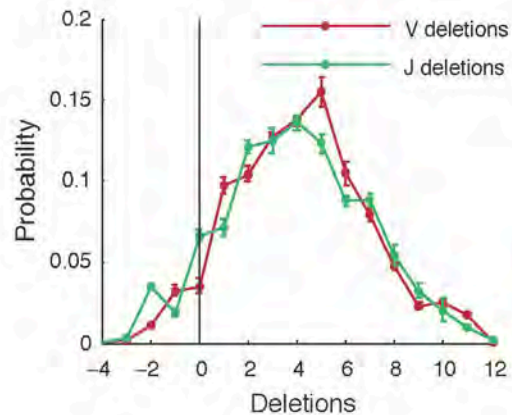


VD and DJ insertions are independent and identically distributed

Nucleotide statistics are captured by dinucleotides and identical on the opposite strands for VD and DJ



Deletions are Sequence Dependent



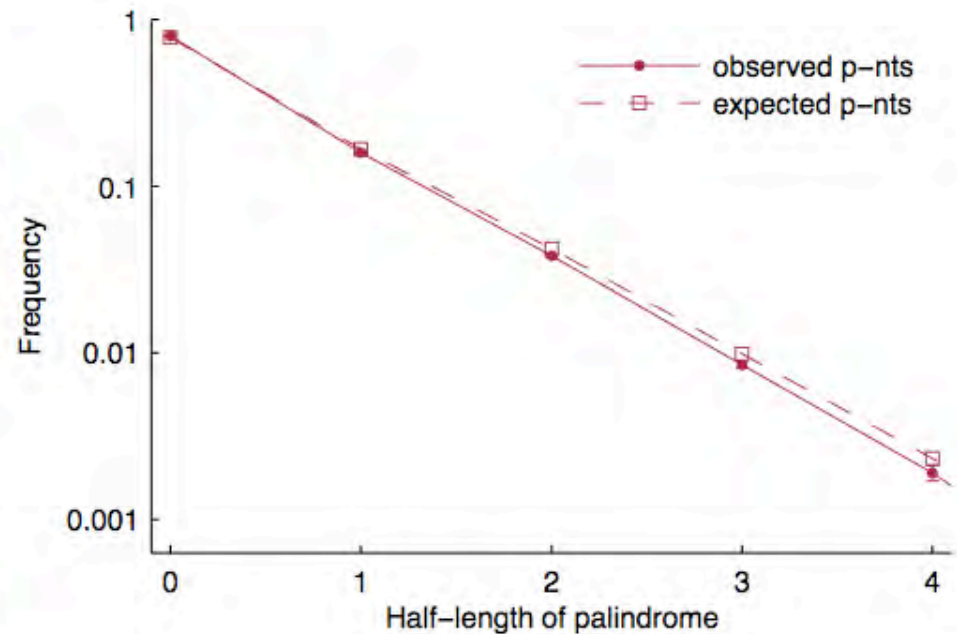
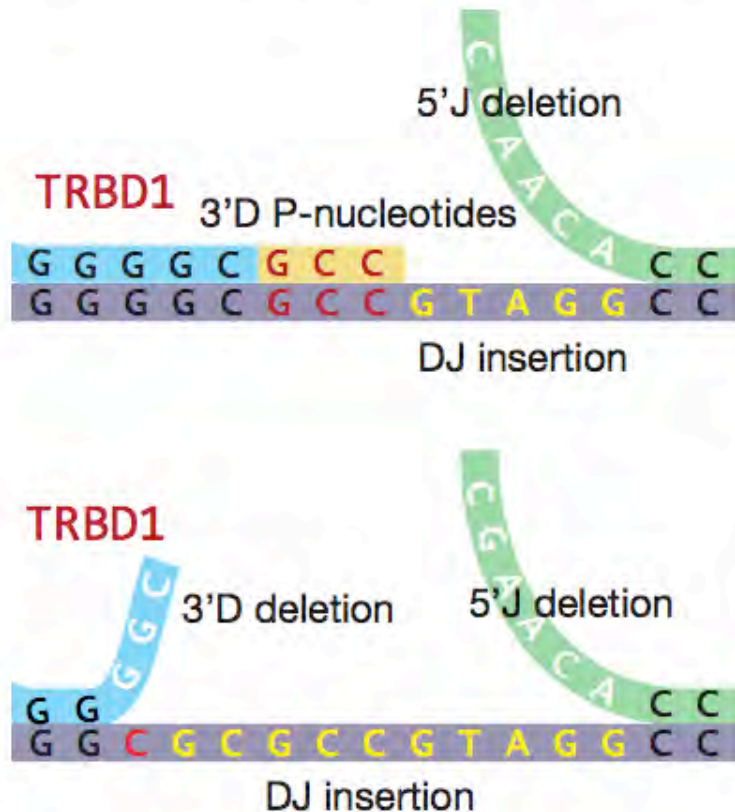
Evidence of sequence dependent nuclease activity

Extremely consistent across individuals

Blue lines: Crude model (no distance effects) explains some of the variation ($r^2=0.7$)

Palindromes co-occur only with zero deletions

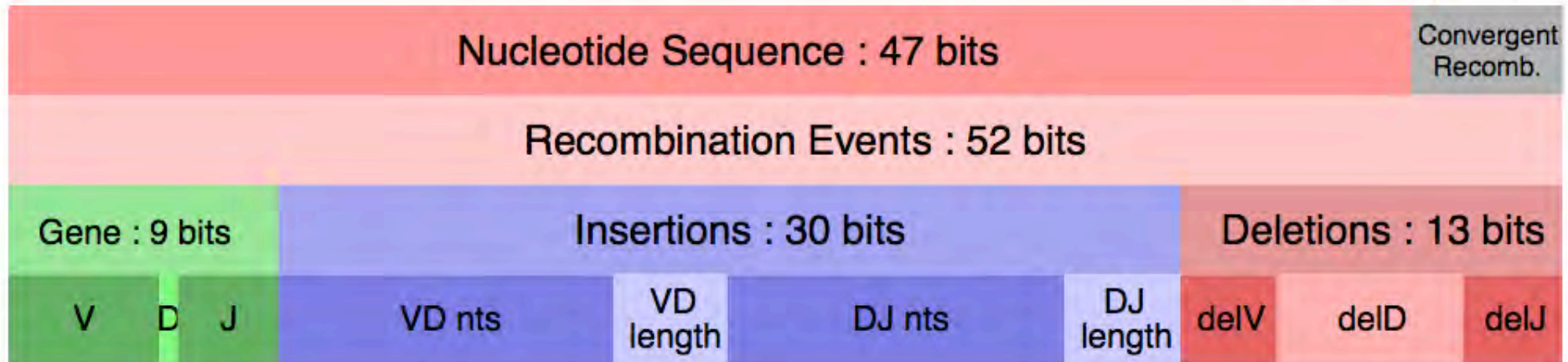
Palindromes with non-zero deletions completely consistent with chance insertions



We label them as 'negative' deletions

Potential Diversity

32 events generate typical sequence



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

$$S_{\text{recomb}} = - \sum_E P_{\text{recomb}}(E) \log P_{\text{recomb}}(E)$$

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

Gene choices : 18%

Nucleotide Deletions : 25%

Nucleotide Insertions : 57%

No. of T-cells in an individual $\sim 10^{12}$

No. of unique TCR β s in an individual $\sim 10^6 - 10^8$

No. of peptide-MHC complexes (with 11 aa) $\sim 10^{13}$

Actual Potential Unique TCR β s $\sim 10^{14}$

Overlap Between Individual Repertoires

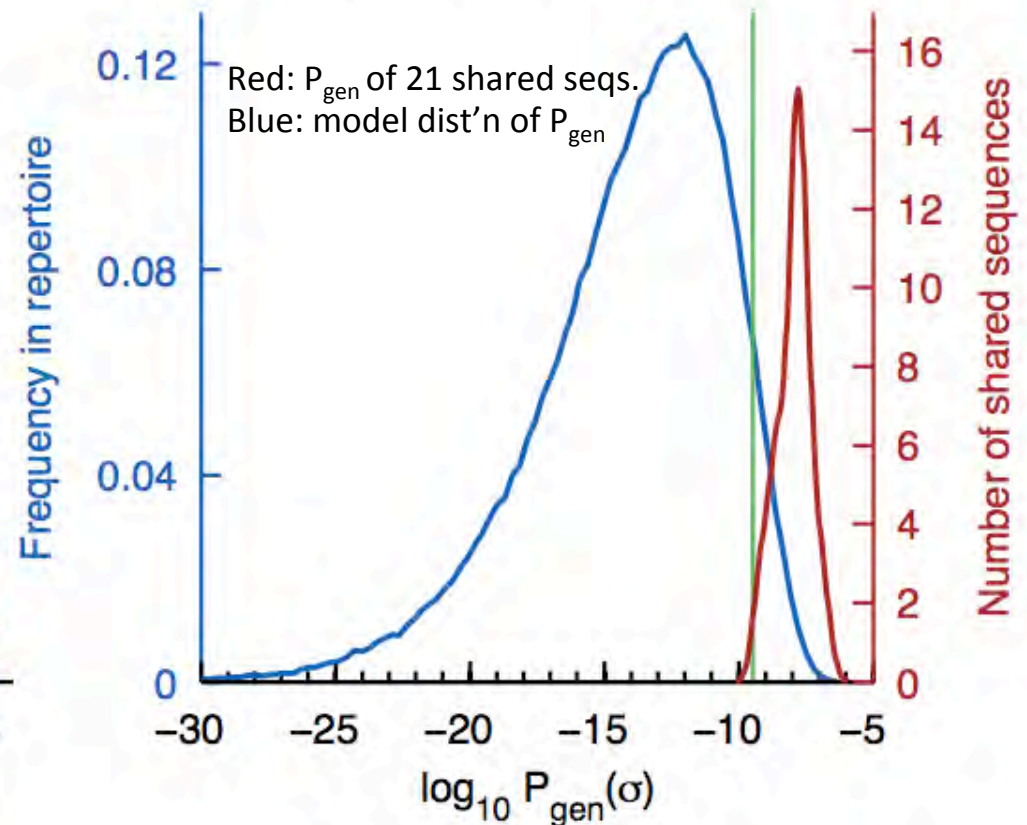
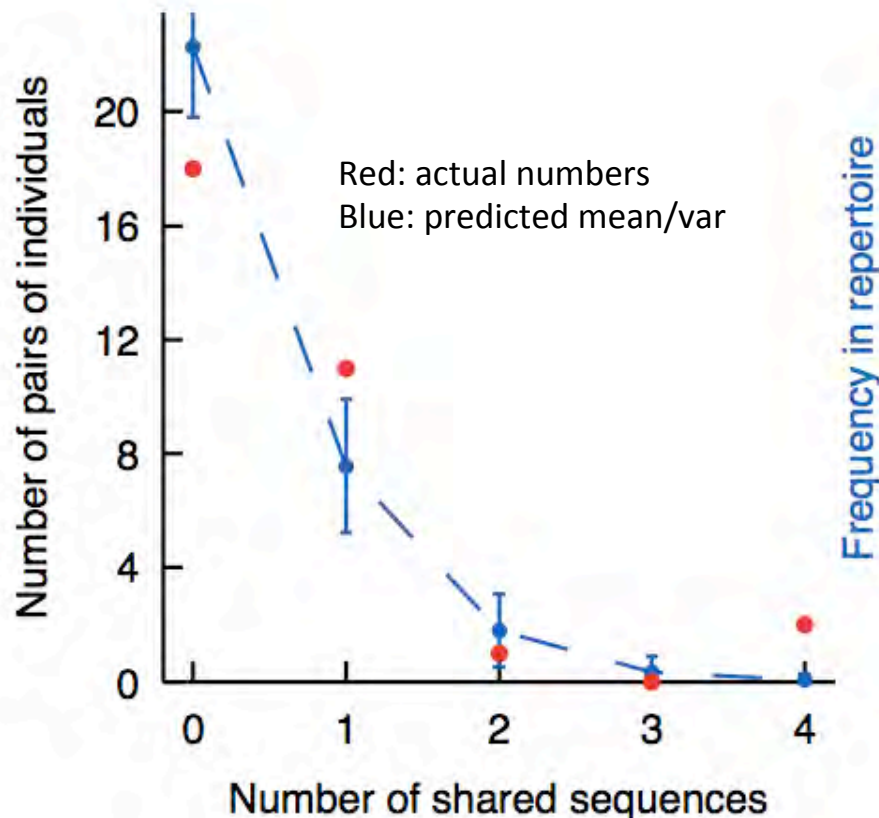
We know the a priori generation probability of any specific sequence!
Hence, can assess the chance likelihood for two individuals to share TCRs.

The number of shared sequences between a pair of individuals should be Poisson distributed.

$$\bar{n} = N_1 N_2 \langle P_{\text{gen}} \rangle_{\sigma}$$

$$\langle P_{\text{gen}} \rangle_{\sigma} = \sum_{\sigma} P_{\text{gen}}^2(\sigma)$$

$$\simeq 3.4 \pm 0.1 \times 10^{-10}$$



Conclusions & Where do we go from here?

- We think we now know the stochastic parameters of VDJ recombination & can assess the generation probability of any specific TCR sequence.
- As expected, VDJ recombination seems to be a universal process within a given species: different individuals have nearly identical distributions.
- The same enzymes edit B-cell receptors; so, with DNA sequence data on BCRs, we can easily check that they obey the same generative statistics.
- The relevant enzymes can be made to function *in vitro*; clever biochemists should be able to directly measure some of these DNA editing statistics.
- Most importantly, the primitive generative statistics provide the baseline for assessing how the statistics are changed by selection for function.
- Selection (in thymus, say) acts on aa sequence; need to model how a TCR binds to *many* peptides presented on *several* MHC complexes. Tough!
- No simple model structure (which we might hope to fit to data) is known for the transfer function between $P_{gen}(\sigma)$ and $P_{naive}(\sigma)$. But we're on it!
- Our point: individual sequences can't tell you much of anything; you have to learn pdf's and then, perhaps, draw functional conclusions from things like $DKL(\text{PDF}_{\text{healthy}}, \text{PDF}_{\text{sick}})$. But this is all very much work in progress.