T cell statistics and the element of surprise in immunology:

Curtis G. Callan, Jr.

Physics Department,

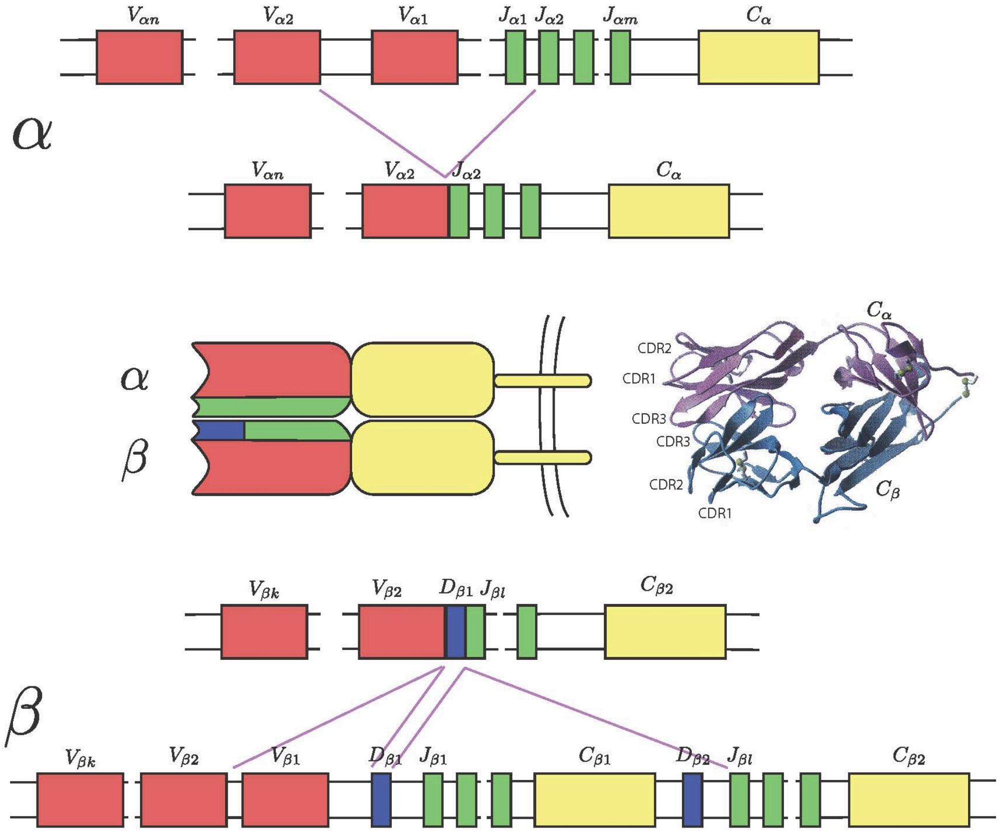
Princeton University

The DNA of the cells of the adaptive immune system undergoes stochastic gene editing to provide the diversity needed to deal with pathogens. Copious data on this diversity are being provided by high-throughput sequencing. Using a statistical inference framework we can quantify how "surprising" it is to see any given T cell sequence in a blood or tissue sample. This approach reveals "hidden variables", such as the generation probability of any T cell sequence, that enable novel modes of analyzing immune function. Emerging cancer data sets provide an ideal field for application of analyses based on these ideas.

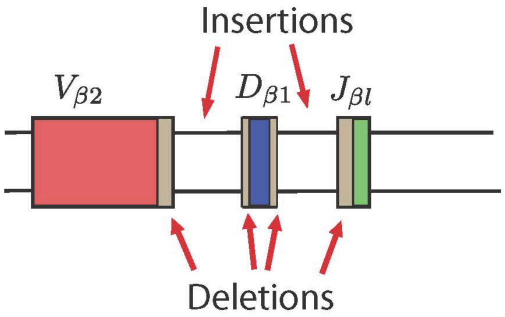
Work with Zach Sethna & Yuval Elhanati (Princeton) , T. Mora (ENS) & A. Walczak (ENS). Data provided by H. Robins (FHCRC), V. Balachandran (MSKCC).

TCR Diversity from Stochastic Genome Editing

"VDJ Recombination" of germline DNA produces a unique TCR (BCR) gene in each new T or B cell created in the bone marrow. This amazing process is effected by the same suite of DNA repair enzymes for T- and B- cells.



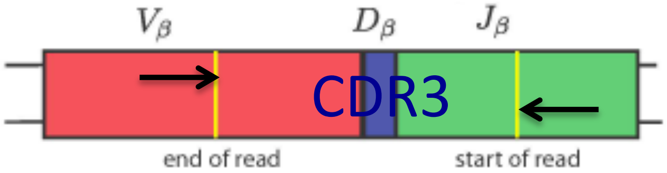
Alpha chain	Beta chain
70 V, 61 J genes	48 V, 2 D, 13 J genes



DNA editing is implemented by distinct stochastic "events": gene choice, deletions, insertions. If result is "out of frame", cell may try again!

Our first goal is to infer the detailed statistics of these generative events

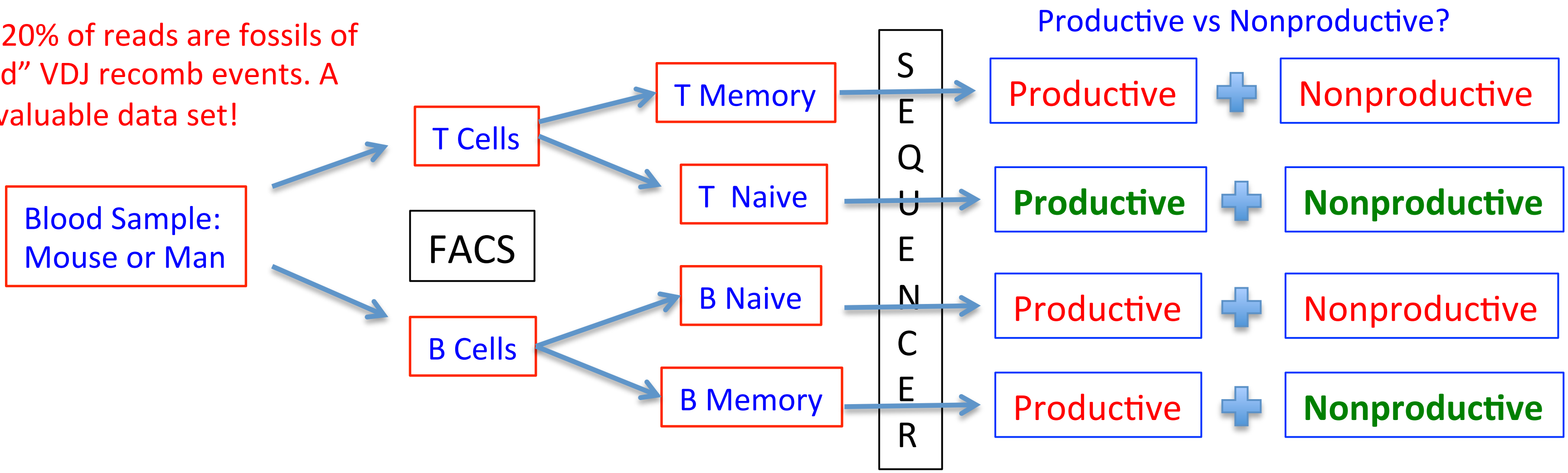
Output of VDJ rearrangement can be "observed" via high-throughput sequencing technology



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

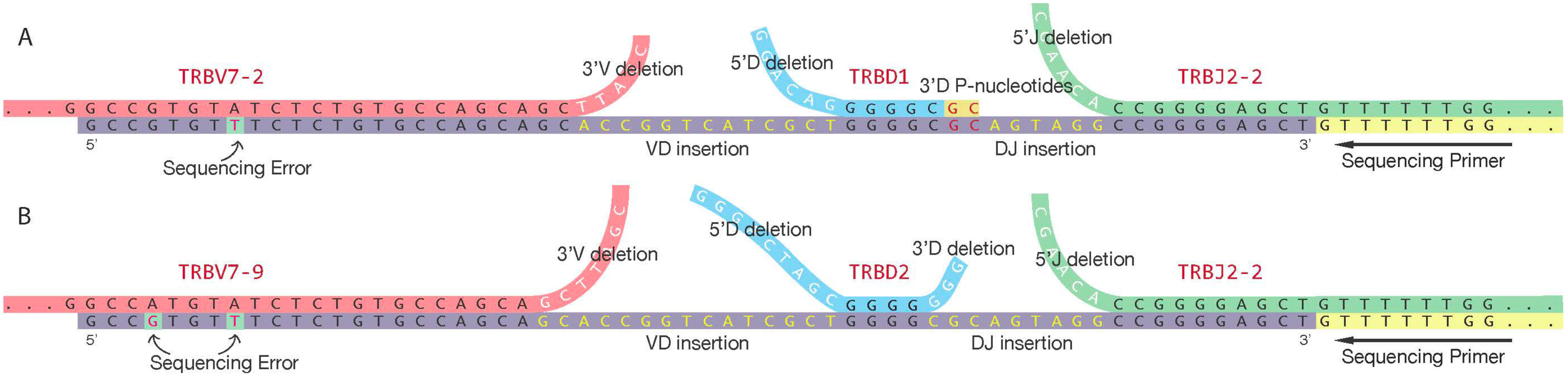
Acquire ~10^6 ~100 bp reads (per blood sample) covering the highly variable CDR3 region of the gene. Data is a list of unique sequences plus occurrence numbers (indicative of clone size of that sequence)

NB: ~20% of reads are fossils of "failed" VDJ recomb events. A very valuable data set!



Our goal is to characterize the statistics of these different sequence repertoires: what is the probability that any given sequence gets made by a stem cell; what is the probability that it is selected to become a mature T cell; how do these statistics differ from individual to individual?

Many generation scenarios for one TCR sequence



Scenario *E* for generating a read σ is a set of values for a set of "actions". Goal is to infer a pdf for the generative scenarios. Need to assume a plausible structure for that pdf:

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

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

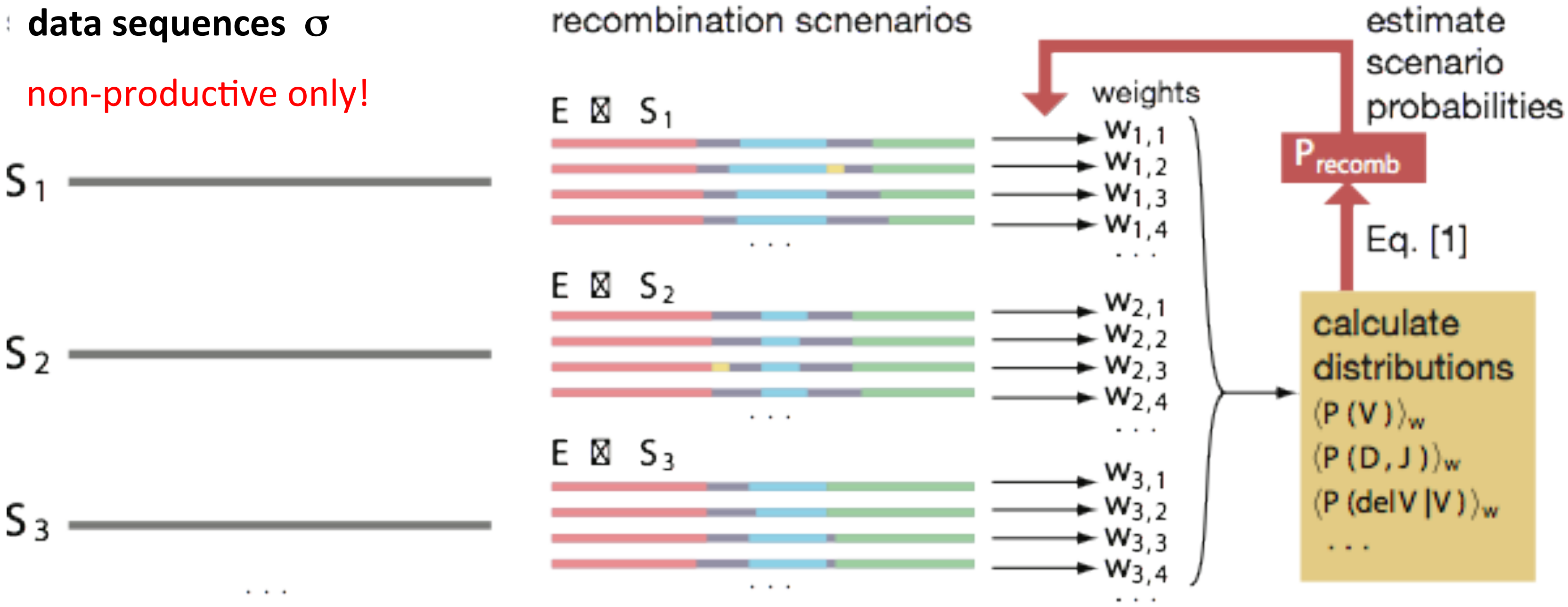
$$P_{gen}(\vec{\sigma}) = \sum_{\substack{\text{scenarios:}\\ V,D,J,\ldots\to\vec{\sigma}}} P^{\mathrm{recomb}}(\mathrm{scenario})$$

N.B. Many "scenarios" can yield the same sequence read σ.

P_{gen}(σ) is the net probability that the read σ is produced in a single stem cell event. It is a measure of the "surprise" value of σ. We want to evaluate it for any sequence.

Infer the "best" generative model, given the data

We use an iterative procedure to find the component pdfs that maximize the likelihood of the observed non-productive sequence repertoire {σ}. This eliminates selection effects and reveals the stochastic machine operating at the stem cell level.



data sequences σ

non-productive only!

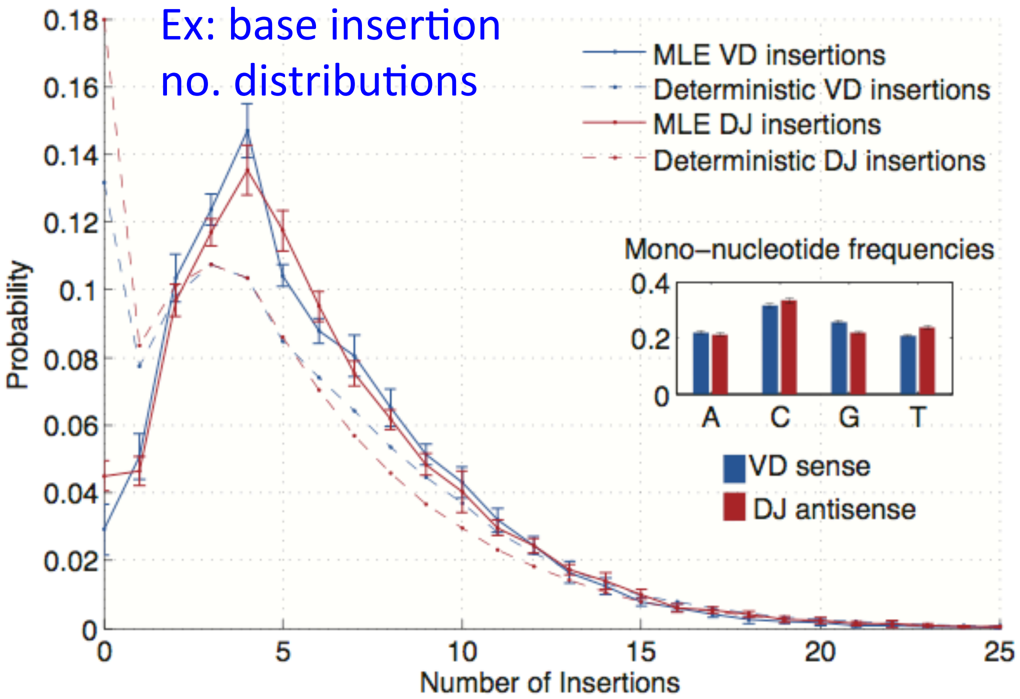
$$P^{\text{recomb}}(\text{scenario}) = P(V)P(D,J)P(\text{deletions}V|V)P(\text{insertions}DJ)\ldots$$

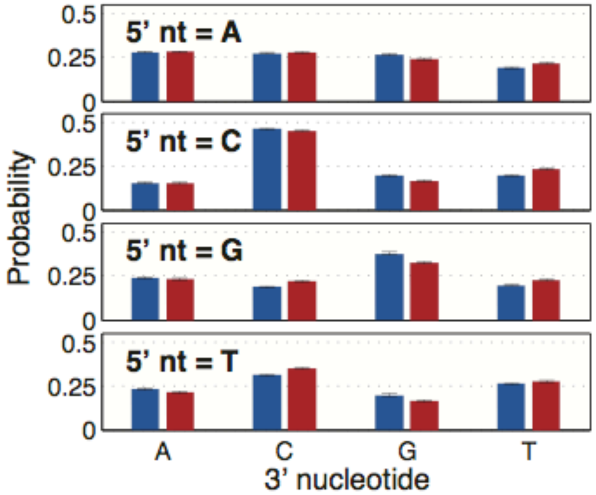
$$P_{gen}(\vec{\sigma}) = \sum_{\substack{\text{scenarios:}\\ V,D,J,\ldots\to\vec{\sigma}}} P^{\text{recomb}}(\text{scenario})$$

Explicit forms for the scenario pdfs → access to P_{gen} of any σ

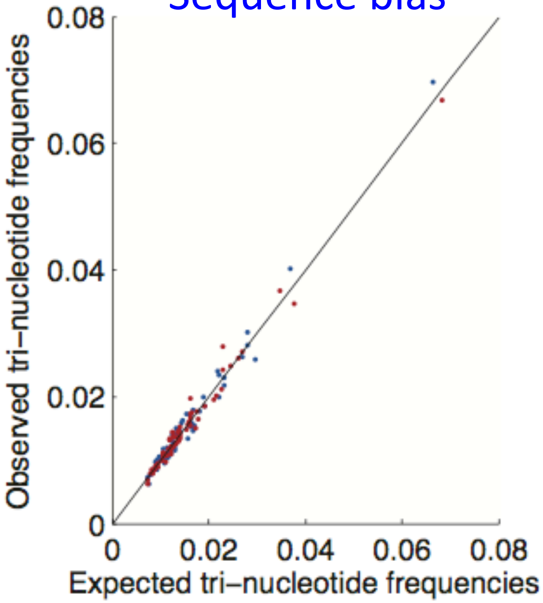
Inferred component pdf’s are quasi-universal

Ex: base insertion no. distributions





Sequence bias

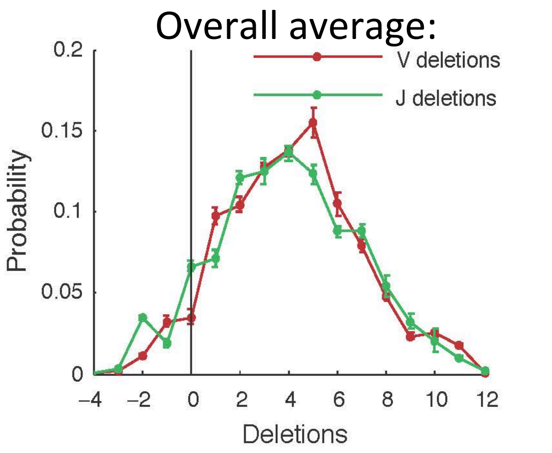


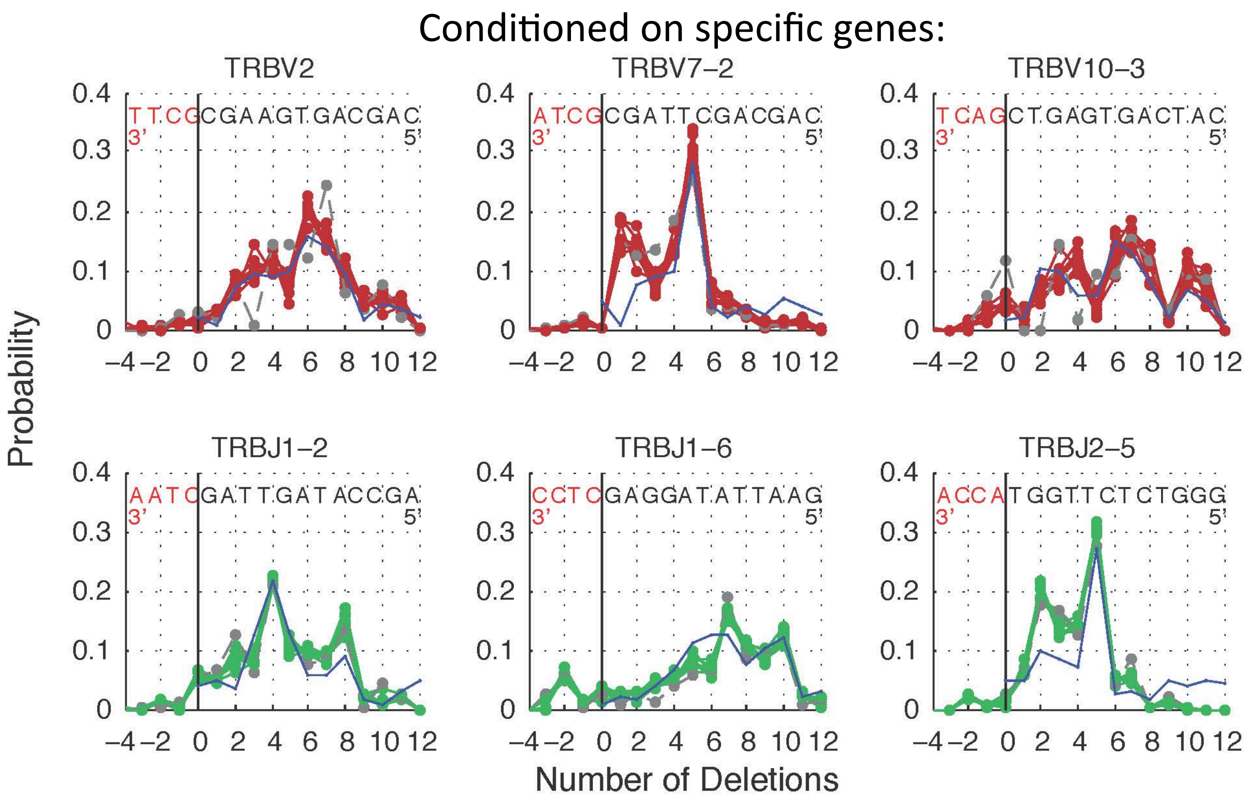
VD and DJ insertions are independent and identically distributed

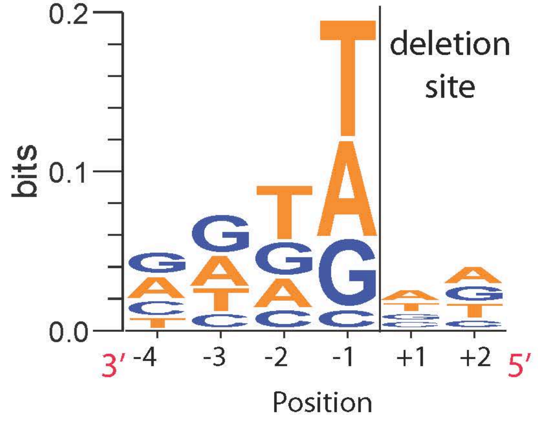
Nucleotide statistics are captured by dinucleotides and identical on the opposite strands for VD and DJ

Peak at 4 insertions may have to do with structure of TdT enzyme that inserts random single bases …..

T Cell Generative Model: Gene Dependent Deletions


Overall average:


Conditioned on specific genes:



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)

Generative Model Quantifies Diversity/Entropy

Shannon entropy is a good measure of sequence repertoire diversity:

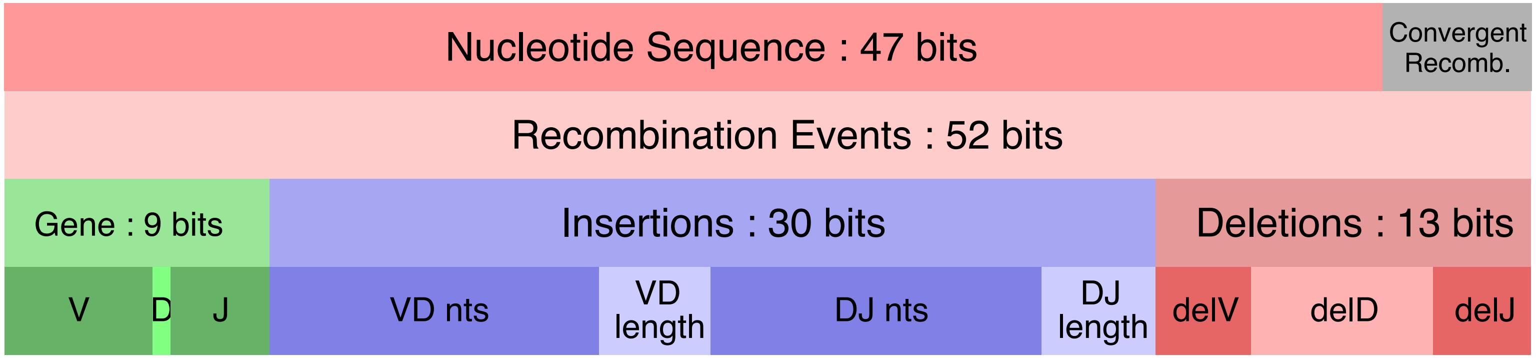
$$S_{\mathrm{seq}} = -\sum_{\sigma} P_{\mathrm{gen}}(\sigma) \log P_{\mathrm{gen}}(\sigma)$$

Can't compute it directly: sequence data only sparsely samples the distribution. Instead we compute the entropy of recombination events, and then correct for convergent recombination: multiple scenarios (E) that give the same sequence:

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

Correction ~5 bits (estimated during inference procedure)

Net of 47 bits can be parsed into contributions from different event types:



Nucleotide Sequence : 47 bits

Convergent Recomb.

Recombination Events : 52 bits

Gene : 9 bits

Insertions : 30 bits

Deletions : 13 bits

V

D

J

VD nts

VD length

DJ nts

DJ length

delV

delD

delJ

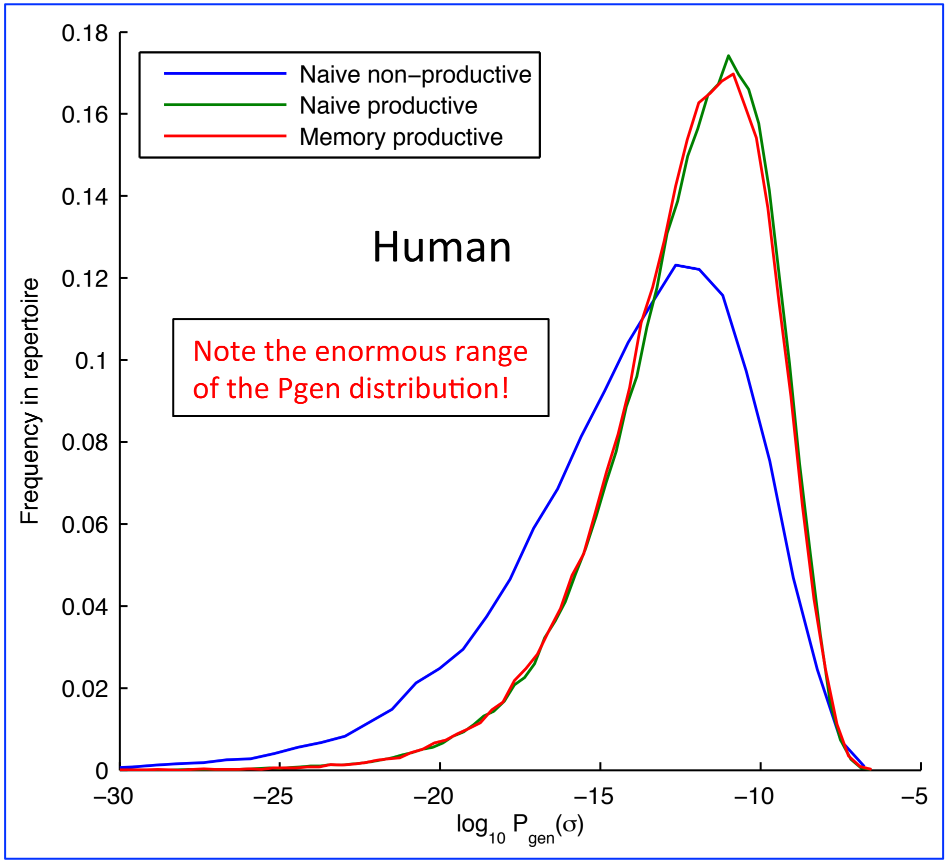
Number of potential unique TCRβs ~ 10^{14} : vastly larger than number of unique T cells in an individual human. Comparison with mouse is interesting …

Some T cells are much more equal than others

Every nt sequence σ has its “surprise value” $P_{gen\_nt}(\sigma)$, or its probability of generation in a one-shot recombination event. We can compute it for any sequence (histogram plots for real T cell repertoires are below)

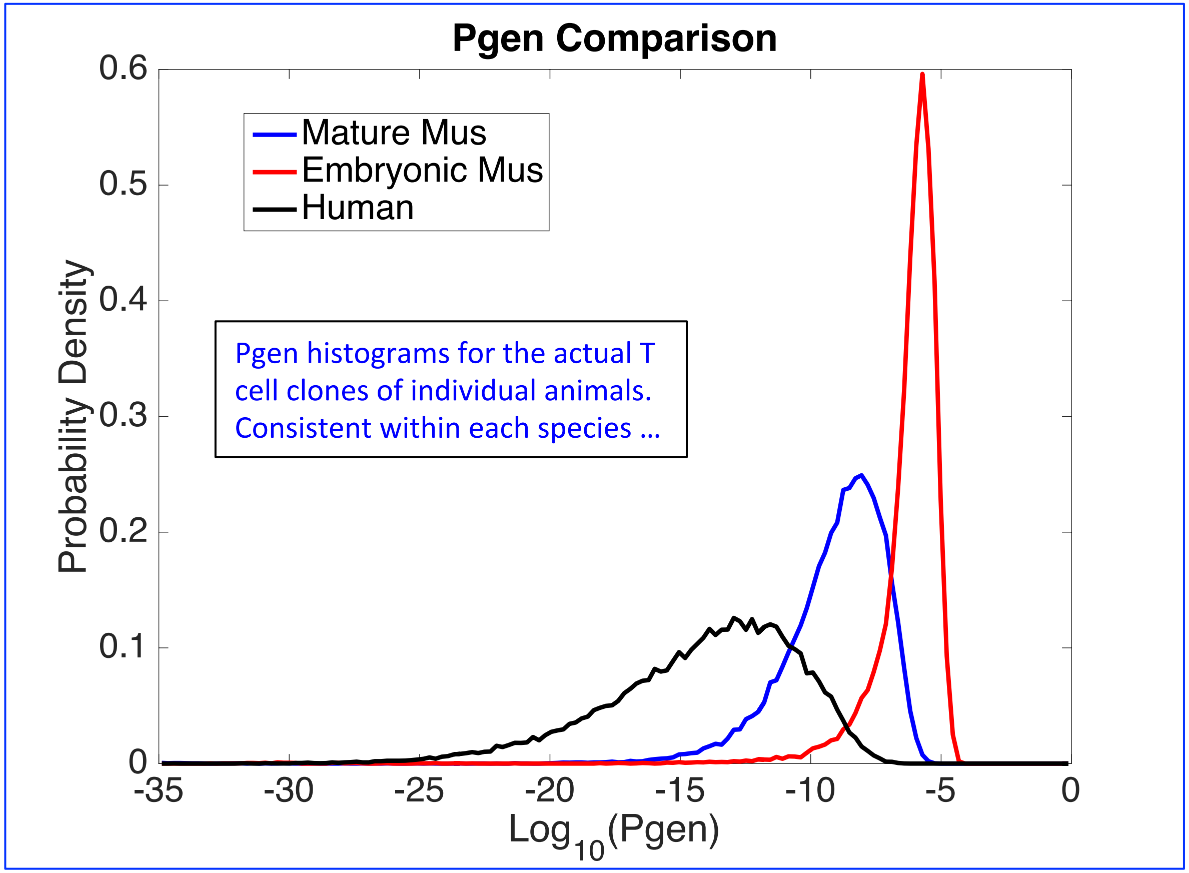
The salient feature of these histograms is the enormous range of the distribution. Some T cells are much more likely than others.

It is worth noting that $N_{clonotype} \sim 10^{10}$ for human. Hence, many T cell seqs have P_{gen} so big that they will be found in every individual! As a corollary, there are many low P_{gen} sequences that will be unique to their owner. Consequences?



Note the enormous range of the Pgen distribution!

$P_{gen\text{-}nt}$ distributions for man vs. mouse



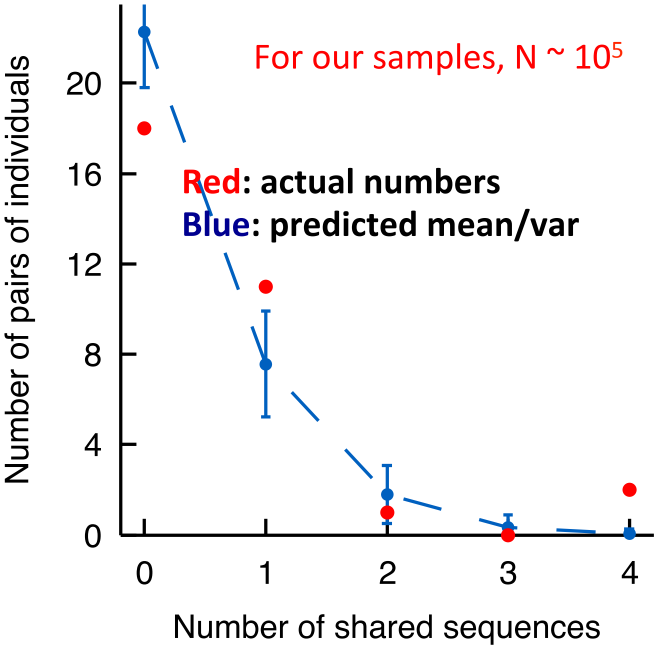
Pgen histograms for the actual T cell clones of individual animals. Consistent within each species …

Sequences are are more "probable" in mouse than in human. At the same time $N_{clonotype}$ ~10^7 for mouse and ~ 10^{10} for human. This has the serendipitous result that in both cases there are specific T cell sequences with P_{gen} so big that they will be found in every individual!

Shared nt sequences are a test of $P_{gen\_nt}(\sigma)$

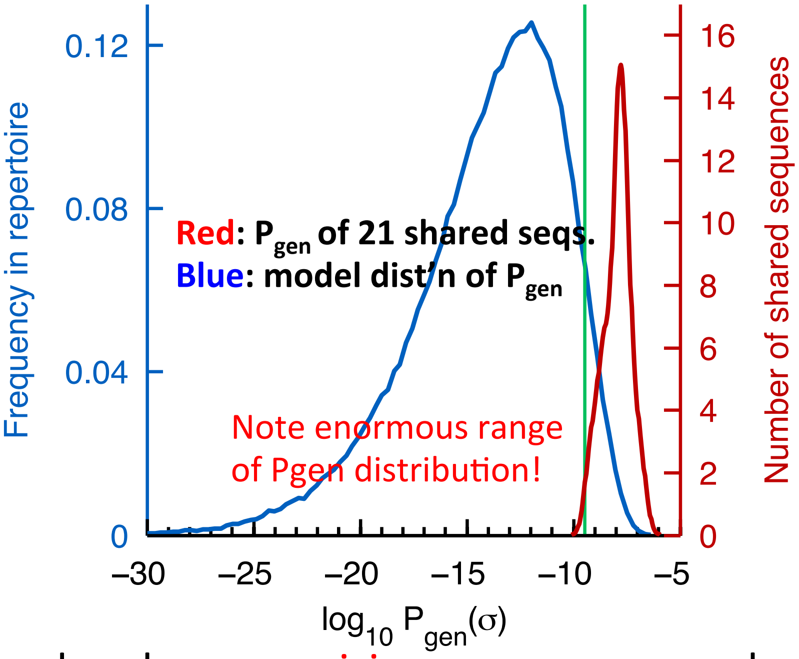
We know the generation probability of any given sequence and can assess the chance likelihood for two individual (human) samples (sizes N_1 and N_2) to share n sequences:

$$\bar{n} = N_1 N_2 \langle P_{\text{gen}} \rangle_\sigma \quad \text{where} \quad \langle P_{\text{gen}} \rangle_\sigma = \sum_\sigma P^2_{\text{gen}}(\sigma) \simeq 3.4 \pm 0.1 \times 10^{-10}$$



For our samples, N ~ 10^5

Red: actual numbers
Blue: predicted mean/var



Red: P_{gen} of 21 shared seqs.
Blue: model dist'n of P_{gen}

Note enormous range of Pgen distribution!

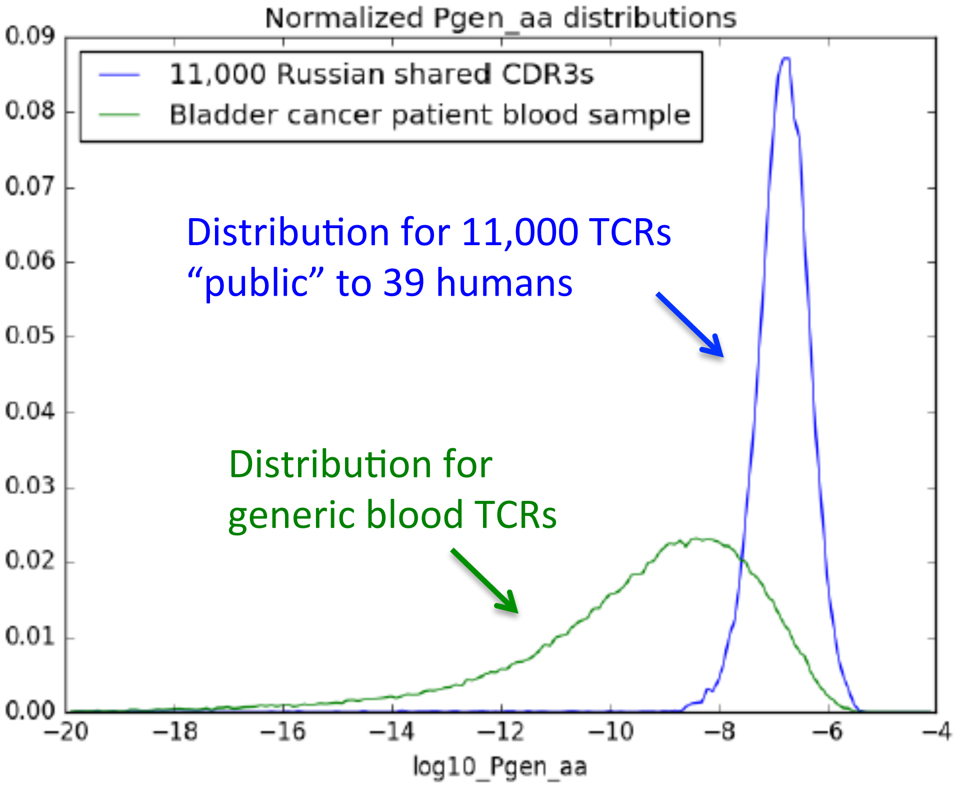
In first approximation, this looks good: only unsurprising sequences are shared.
We take this as evidence that our inference of Pgen is reasonably accurate.
Note that we are talking here about sharing of unproductive, unselected, seqs!

But we should be more interested in $P_{gen\_aa}(\sigma)$!

Functional specificity is actually conferred by the aa, not nt, sequence of the CDR3 region. Codon degeneracy means Pgen_aa > Pgen_nt. By how much?

To get Pgen_aa we must sum Pgen_nt over all nt versions of the same aa seq. Non-trivial, but can be done; the data distribution moves up about 2 powers of ten!

Given the size of human, there are many CDR3 aa sequences that will appear at least once in each and every individual. Thus there will be many “public” CDR3 aa sequences. Is this just randomness at work, or does it have real significance?

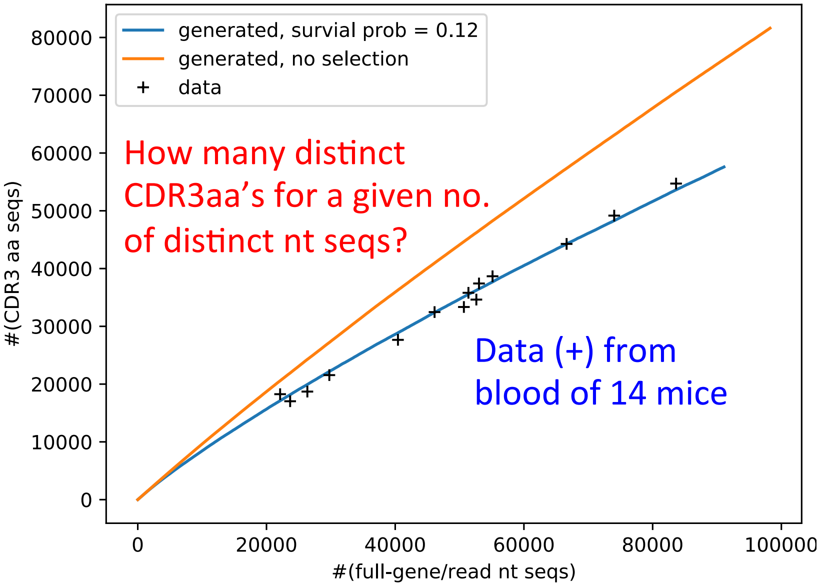


Distribution for 11,000 TCRs “public” to 39 humans

Distribution for generic blood TCRs

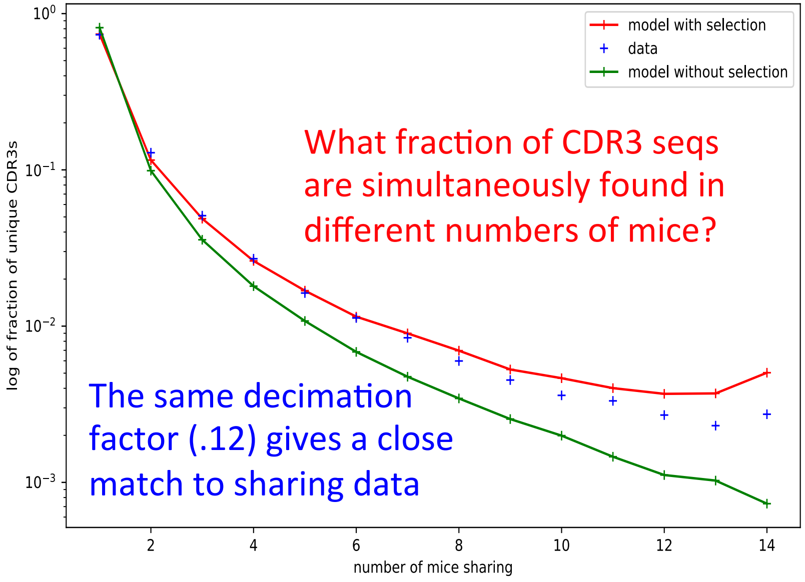
Shared CDR3 aa sequences as $P_{gen\_aa}(\sigma)$ stress test

Our methods allow us to generate T cell repertoires of any size; we can test whether real and synthetic statistics agree. For instance, we can test sharing of CDR3 aa sequences between multiple individuals. We tried it out on data from 14 mice:



How many distinct CDR3aa's for a given no. of distinct nt seqs?

Data (+) from blood of 14 mice



What fraction of CDR3 seqs are simultaneously found in different numbers of mice?

The same decimation factor (.12) gives a close match to sharing data

Now we must deal with the issue of selection! We can generate samples of in-frame CDR3 seqs, but thymic selection will delete many of them on functional grounds. We find that random decimation of generated sequences (keep ~12%) works very well.

Finally, many CDR3's react to the same antigen

These statistically common CDR3 sequences are rich in T cell sequences known to respond to common antigens (Friedman). Look at one particular example in mouse.

Venturi observed that a certain mouse influenza epitope elicited a T cell response that has a restricted motif: CASXGGXNTGQLYF. We can look for such sequences in T cell repertoires from lab mice. They are there even if these mice have never been exposed to this antigen:

Thymus Sequence	Count	Pgen
CASTGGPNTGQLYF	1	1.19E-05
CASVGGANTGQLYF	1	1.19E-05
CASTGGANTGQLYF	1	1.19E-05
CASGGGRNTGQLYF	1	1.19E-05
CASIGGANTGQLYF	1	1.19E-05
out of 82147	5	5.94E-05

Blood Sequence	Count	Pgen
CASSGGPNTGQLYF	1	3.08E-05
CASRGGPNTGQLYF	1	3.08E-05
CASSGGANTGQLYF	1	3.08E-05
out of 32495	3	9.23E-05

Since we know Pgen_aa for any CDR3, we can learn the frequency of random generation of the motif itself. We find that it is ~10^{-4} , consistent with its rate of occurrence in real samples. Since mouse has ~10^{7} clonotypes, this is very public!

What we really need is Pgen(epitope): the probability that a Tcell capable of responding to the epitope of interest can be generated. This requires a deeper understanding of cross-reactivity. Clearly very relevant to immunotherapy.

Connecting to immunotherapy?

A missing link in our understanding of immune function is the existence of a model for the affinity between a T cell receptor and the peptide epitopes it has to scan. Relevant data is accumulating, but we don't have a clear idea of how much is enough.

There is an interesting abstract question here: can you design a function (epitope + T cell -> affinity) that meets the requirements of a working immune system? Recognize all pathogens, no autoimmune problems, statistically realistic T cell repertoires etc ...?

The data being developed by Balachandran *et al* provide a rich resource for studying this problem. With Greenbaum/Luksa, we are using the ideas laid out here to sharpen our ability to identify CDR3 clusters reacting to specific neo-antigens. The goal is to use intelligent statistics to do better than just identifying common T cells between tumor and blood, say. This is a work in progress ... the data is just coming in.

The basic immunotherapy idea has an important statistical question at its core: will a patient's immune repertoire possess T cells that can recognize a typical set of tumor neo-antigens? Always, sometimes, or hardly ever? The goal of our theoretical exercise is to answer this question ... Its certainly worth trying.

Acknowledgement: Work of postdoc Elhanati supported by V Foundation fellowship. I'm supported in part by NSF grant ("Liberating Immune Response to Pancreatic Cancer").