Statistical inference of the generation probability of T-cell receptors from sequence repertoires

Anand Murugan [*], **Thierry Mora** [†], **Aleksandra M. Walczak** [‡] **and Curtis G. Callan, Jr.** [*] [§]

[*]Joseph Henry Laboratories, Princeton University, Princeton, New Jersey 08544 USA,[†]Laboratoire de physique statistique, UMR8550, Centre National de la Recherche Scientifique and École Normale Supérieure, 24, rue Lhomond, 75005 Paris, France,[‡]Laboratoire de physique théorique, UMR8549, Centre National de la Recherche Scientifique and École Normale Supérieure, 24, rue Lhomond, 75005 Paris, France, and [§]Simons Center for Systems Biology, Institue for Advanced Study, Princeton, New Jersey 08544 USA

Submitted to Proceedings of the National Academy of Sciences of the United States of America

Stochastic rearrangement of germline V-, D-, and J-genes to create variable coding sequence for certain cell surface receptors is at the origin of immune system diversity. This process, known as 'VDJ recombination', is implemented via a series of stochastic molecular events involving gene choices and random nucleotide insertions between, and deletions from, genes. We use large sequence repertoires of the variable CDR3 region of human CD4+ T-cell receptor beta chains to infer the statistical properties of these basic biochemical events. Because any given CDR3 sequence can be produced in multiple ways, the probability distribution of hidden recombination events cannot be inferred directly from the observed sequences; we therefore develop a maximum likelihood inference method to achieve this end. To separate the properties of the molecular rearrangement mechanism from the effects of selection, we focus on non-productive CDR3 sequences in T-cell DNA. We infer the joint distribution of the various generative events that occur when a new T-cell receptor gene is created. We find a rich picture of correlation (and absence thereof), providing insight into the molecular mechanisms involved. The generative event statistics are consistent between individuals, suggesting a universal biochemical process. Our distribution predicts the generation probability of any specific CDR3 sequence by the primitive recombination process, allowing us to quantify the potential diversity of the T-cell repertoire and to understand why some sequences are shared between individuals. We argue that the use of formal statistical inference methods, of the kind presented in this paper, will be essential for quantitative understanding of the generation and evolution of diversity in the adaptive immune system.

Introduction

Receptor proteins on the surfaces of B- and T-cells in the immune system interact with pathogens, recognize them and initiate an immune response. The diversity of these receptors is the outcome of a remarkable process in which germline DNA is edited to produce a repertoire of (T- or B-) cells with varied antigen receptor genes [1]. The process is called 'VDJ recombination' because the germline contains multiple versions of so-called V-, D-, and J-genes, particular instances of which are quasi-randomly selected, stochastically edited and joined together to produce a new surface receptor gene each time a new immune system cell is generated.

The statistical distribution of these biochemical events (and the resulting receptor coding sequences) in a population of newly-created receptors is an important quantity: it contains information about the *in vivo* functioning of the biochemical editing mechanism and provides the baseline for a quantitative assessment of the downstream workings of selection in the adaptive immune system. Here, we address the problem of inferring this distribution from the large sequence repertoires that are becoming available via high-throughput sequencing technology [2, 3, 4, 5]. In particular, we focus purely on a subset of receptor sequences that are non-productive, due to a reading frame shift or an accidental stop codon, to isolate the statistics of the molecular mechanism from the effects of selection on the functional repertoires.

In the beta chain of human T-cell receptors (the focus of this work), the germline has 48 different V-genes, 2 D-genes and 13 J-genes. VDJ recombination proceeds by first joining a D-gene with a J-gene, and then a V-gene with the DJ junction. First, the recombination activating gene (RAG) protein complex brings two randomly chosen D- and J-genes together, cuts out the intervening chromosomal DNA, and forms a hairpin loop at the end of each gene [6, 7]. In further steps [8, 9] the hairpin loops are opened, creating overhangs at the end of both genes that may eventually survive as P-nucleotides (short inverted repeats of gene terminal sequence) [10]. This is followed by nucleotide deletions and insertions at the junctions and ends with ligation. The process is then repeated between a random V-gene and the DJ junction. The end product is the so-called CDR3 region of the receptor gene: a short, highly variable region that plays an essential role in determining the antigen specificity of the cell.

Each recombined sequence can thus be thought of as the outcome of a generative event described by several random variables (Fig. 1): V-, D-, and J-gene choices, deletions of variable numbers of nucleotides from the selected genes, insertions of random nucleotides between them, and the possible creation of P-nucleotides (short palindromic nucleotides as in Fig. 1A at the $3'$ end of the D-gene). From the set of observed CDR3 sequences, we wish to infer the underlying probability distribution of these generative events.

To date, this inference has been done via a deterministic alignment procedure which assigns a unique event to each sequence [2, 3, 4]. However, because individual CDR3 sequences can arise in multiple ways (Fig. 1), this assignment must be done probabilistically. Deterministic alignment introduces spurious biases and correlations in the statistics of generative events (Fig. 2). Thus, a statistical inference procedure is needed to accurately infer the underlying event probability distribution from the data. In this paper we present such a method, based on likelihood maximization via an iterative expectation-maximization algorithm [12], and apply it to recent data on human T-cell receptor sequences.

Reserved for Publication Footnotes

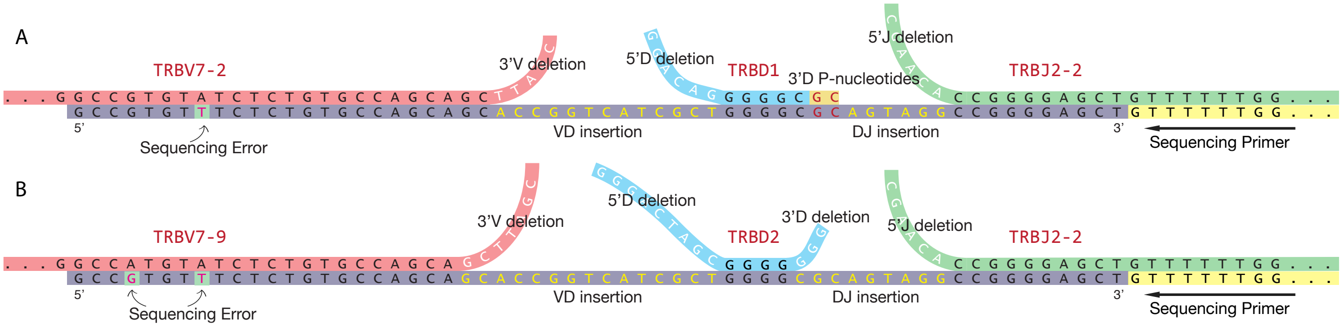
A
TRBV7-2
3'V deletion
5'D deletion
TRBD1
3'D P-nucleotides
5'J deletion
TRBJ2-2
. . . G G C C G T G T A T C T C T G T G C C A G C A G C T T A C
G C C G T G T T T C T C T G T G C C A G C A G C A C C G G T C A T C G C T G G G G C G C A G T A G G C C G G G G A G C T G T T T T T T G G . . .
5'
3'
Sequencing Error
VD insertion
DJ insertion
Sequencing Primer
B
TRBV7-9
3'V deletion
5'D deletion
TRBD2
3'D deletion
5'J deletion
TRBJ2-2
. . . 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 G C
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'
Sequencing Error
VD insertion
DJ insertion
Sequencing Primer

Fig. 1. A 60bp CDR3 read (grey box) can be aligned to different genes (nomenclature follows IMGT conventions [11]) with different deletions (white), insertions (yellow), and P-nucleotides (red). (A) Alignment to specific V-, D-, and J-genes with insVD=13, insDJ=6, delV=5, delJ=6, del5'D=6, del3'D=−2 (in other words, pal3'D=2). (B) Alignment of the same read to different V- and D-genes, and with insVD=15, insDJ=9,delV=7, del5'D=9, del3'D=3 (no P-nucleotides). Note that the alignment to the V-gene is not maximal in this case. A few heavily penalized mismatches are allowed (in the V-gene in this example) in order to accommodate a small sequencing error rate. The location of the sequencing primer is indicated: it is chosen to uniquely identify the start of the CDR3 read within each J-gene.

Analysis

We work with sequence data on CD4+ T-cell beta chain CDR3 regions obtained from nine human subjects by methods described in [4, 5] (see Acknowledgements). In these experiments, T-cells are collected from a blood sample and sorted into 'naïve' (CD45RO-) and 'memory' (CD45RO+) compartments, DNA is extracted, and sequence reads long enough to capture a 5′ piece of the J gene, a 3′ piece of the V gene and the variable sequence lying in between, are obtained. Each sequence is read multiple times, and a clustering algorithm is used to correct for sequencing error. This process produces a data set consisting of an average of 232,000 (140,000) unique CDR3 sequences from the naïve (memory) compartments for each individual subject. Each unique sequence comes with a multiplicity reflecting the prevalence of that particular cell type in the blood sample.

Roughly 14% of the unique CDR3 sequences are 'non-productive': either their J genes have been shifted out of the correct reading frame or the CDR3 sequences have a premature stop codon. They arise from a recombination event on one of a cell's two chromosomes that failed to make a functional receptor, followed by a successful recombination on the other chromosome. Such sequences should not be subject to functional selection [5], and their statistics should reflect only the VDJ recombination process (see *SI Appendix* Sec. 10 for evidence that the non-productive constraint introduces no bias). Because this is our primary concern, we focus on the non-productive CDR3 sequences, of which there are an average of 35,000 (22,000) in the naïve (memory) compartments for each individual subject. We analyze the naïve and memory data sets separately to verify the absence of selection effects. See *SI Appendix* Secs. 1,2 for online access to our data sets.

Structure of recombination event distributions. Each CDR3 generating recombination event can be fully characterized by a set E of discrete variables comprising: the identities of the V-, D- and J-genes selected for recombination[1] (V,D,J); the numbers of bases deleted from the 3' end of the V-gene (del$V$), the 5' end of the J-gene (del$J$), and both ends of the D-gene (del5$'D$ and del3$'D$ for the 5' and 3' ends, respectively); the number of palindromic nucleotides at each of the gene ends (pal$V$, pal$J$, pal5$'D$, pal3$'D$); the specific sequence $(x_1, \ldots, x_{\text{ins}VD})$ of length insVD inserted at the VD junction, and the specific sequence, $(y_1, \ldots, y_{\text{ins}DJ})$ of length insDJ inserted at the DJ junction (see Fig. 1). We choose a convention in which both sequences are read in the 5' to 3' direction, but the VD (DJ) inserted sequence is read from the sense (antisense) strand.

We seek a joint distribution over all of these variables containing the minimal set of dependences between the variables that is required to self-consistently capture the observed correlations in the data. We find that the following factorized form for the probability of a recombination event E (defined by specific values for all the event variables) successfully captures all the significant correlations between sequence features that are present in the data (see Fig. 2):

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

The various factors are normalized joint or conditional distributions on their respective arguments. $P(V)$ and $P(D,J)$ account for the fact that the various genes have different usage probabilities (and that D- and J-gene usage is correlated). The factors $P(\text{del}V|V)$, etc., are distributions on the number of nucleotide deletions, conditioned on the gene being deleted (deletion profiles turn out to be very gene-dependent). $P(\text{ins}VD)$ and $P(\text{ins}DJ)$ give the probabilities of different numbers of nucleotide insertions at each junction. The parameters $p_{VD}^{(2)}$ and $p_{DJ}^{(2)}$ account for possible nucleotide bias in the insertions: they give the conditional probabilities of inserting a specific nucleotide given the identity of the immediately 5' nucleotide, with x_0 referring to the last nucleotide at the 3' end of the truncated V-gene on the sense strand for a VD insertion, or at the end of the truncated J-gene on the antisense strand for a DJ insertion.

P-nucleotides do not appear explicitly in Eqn. 1: we treat them as 'negative' deletions (*i.e.* a palindrome of half-length 2, as in Fig. 1A, is counted as a deletion of value −2). This is possible because we find that when the number of nucleotide deletions is greater than zero, occurrences of palindromic nucleotides at the end of the gene segment are completely explained by chance insertions of the corresponding nucleotides (*SI Appendix* Sec. 11 and Fig. S10). Thus, true P-nucleotides,

[1] Here we distinguish only the genes, not their various alleles. The gene list includes germline pseudogenes: they cannot produce functioning receptor proteins but, because we work with non-coding VDJ rearrangements, pseudogene sequences can appear in the data.

not attributable to chance insertions, only occur in association with zero nucleotide deletions and it is consistent to label them as 'negative' deletions.

The factors in our equation for $P_{\mathrm{recomb}}(E)$ (Eqn. 1) are probability distributions on event variables that take on a finite number of values. Specifying this joint distribution requires a total of 2865 probabilities (more than 90% of which are needed for the deletion length probabilities of the individual V-, D- and J-genes). Despite the large number of probabilities to be inferred, we are able to determine them accurately and without overfitting. We emphasize that our goal is to obtain an accurate description of recombination event statistics, and not (yet) to explain those statistics mechanistically.

Generation probability and likelihood of observed sequences. The probability $P_{\mathrm{gen}}(\sigma)$ of generating a specific CDR3 sequence σ is the sum of the probabilities of all recombination events E_σ that produce σ:

$$P_{\mathrm{gen}}(\sigma) = \sum_{E \in E_\sigma} P_{\mathrm{recomb}}(E). \qquad [2]$$

The likelihood $L(\sigma)$ of observing a specific CDR3 sequence read σ, however, must take into account residual sequencing error as well as allelic variation, and is given by a sum over a larger set of recombination events $\widetilde{E}_\sigma$ that generate sequences close to σ:

$$L(\sigma) = \sum_{E \in \widetilde{E}_\sigma} P(E, \sigma) \qquad \text{where} \qquad [3]$$

$$P(E, \sigma) = P_{\mathrm{recomb}}(E) \times \frac{1}{(1+R)^L}$$
$$\times \sum_{\text{alleles } a} P(V_{\mathrm{a}}|V_E) P(J_{\mathrm{a}}|J_E) P(D_{\mathrm{a}}|D_E) \left(\frac{R}{3}\right)^{n_{\mathrm{err}}(\sigma_E^a, \sigma)}. \qquad [4]$$

In the latter equation, n_{err} is the number of mismatches between the observed read σ and the CDR3 sequence σ_E^a that would be produced by the recombination event E with allele choices a. L is the length of the sequence read. The mismatch rate R is determined in the inference with the rest of the distribution parameters and reflects both sequencing error as well as unknown allelic variation. In practice, we only consider recombination events $\widetilde{E}_\sigma$ that lead to CDR3 sequences with at most a few mismatches from σ. The sum over alleles[2] arises because we do not know *a priori* which alleles are present and reads may not go deep enough into the gene sequence to clearly distinguish alleles from each other [13]. The probabilities of the different alleles, given a gene, are also inferred and are expected to differ from individual to individual.

The likelihood of the whole data set $\mathcal{D}$ is then the product of the individual sequence likelihoods: $\mathcal{L}(\mathcal{D}) = \prod_{\sigma \in \mathcal{D}} L(\sigma)$. This expression depends implicitly on the parameters defining the generative probability distribution (along with the allele distributions and the sequencing error parameter), and we infer their correct values by maximizing $\mathcal{L}(\mathcal{D})$ using an expectation maximization algorithm [12, 14] (see *SI Appendix* for algorithmic details). In order to identify universal features of the diversity generation machinery, we perform this inference separately for each subject. A complete analysis software package is available online (see *SI Appendix* for details).

Results

In what follows, we present results of our analysis of naïve, non-productive, CDR3 sequence repertoires of nine individuals (see *SI Appendix* for a parallel analysis of memory sequence repertoires). All of our analysis results are available as downloadable data files (see *SI Appendix* for links).

Correlations between event variables. It is important to verify that correlations not present in the assumed structure of the probability distribution (Eqn. 1) are in fact not present in the data. To perform this self-consistency check, we use the inferred generative distribution to compute the probability-weighted counts distribution of recombination event variables in the data, and then use this distribution to calculate the mutual information of all pairs of event variables. The matrix of mutual information values is shown in the upper-triangular part of Fig. 2A, where the entries outlined in red are dependences accounted for by individual factors in our assumed form of $P_{\mathrm{recomb}}(E)$ (Eqn. 1), entries outlined in green are indirect dependences that can be induced by these factors, and the rest would vanish if the data were perfectly described by the assumed structure of $P_{\mathrm{recomb}}(E)$. There are a few detectable correlations that are not consistent with the assumed structure: (insVD, del$V$), (insDJ, del$J$) and (V, D). They are, however, all so weak (mutual information < 0.02 bits) that we do not model them explicitly (indeed, they might arise from subtle biases in our inference procedure).

For comparison, in the lower-triangular part of Fig. 2A we show the mutual information values of all pairs of variables, but now calculated from a deterministic assignment of events to sequences based on maximal alignments. The resulting distributions exhibit spurious correlations that are absent from the corrected, maximum likelihood estimate (MLE) of the distributions. For instance, the number of insertions at the two junctions are found to be independent in our analysis while the uncorrected estimate shows a dependence (Fig. 2B,C).

Gene usage distributions. The inferred frequencies of V- and J-genes vary significantly from gene to gene, a phenomenon for which no mechanistic explanation has yet been given. In particular, linear location on the chromosome does not explain the pattern of either V- or J-gene usage (see *SI Appendix* Fig. S4A, C). The usage frequencies are consistent between individuals, though of all the inferred parameters in

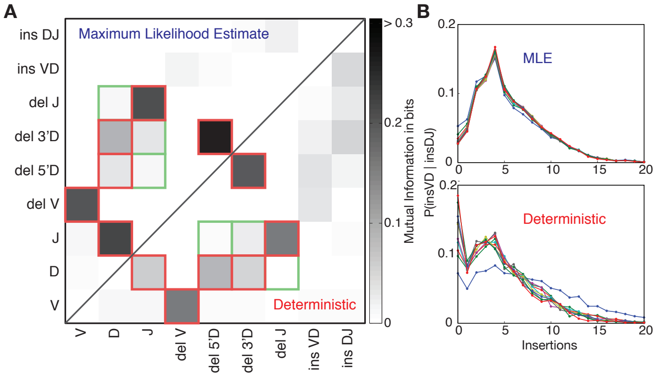

Fig. 2. (A) Data-derived correlations between sequence features: each entry is the mutual information $I(X,Y)$ of a feature pair over the naïve non-productive repertoire. The outlined elements are correlations expected from the form of $P_{\mathrm{recomb}}(E)$: red identifies a direct effect of a factor in Eqn. 1 (e.g. D ↔ J) and green indirect effects (e.g. D ↔ J ↔ delJ). The top-left half of the matrix shows results from the maximum likelihood estimate (MLE), while the bottom-right half corresponds to a deterministic maximum-alignment based identification of recombination events. (B) Probability distribution of the number of VD insertions conditioned on the number of DJ insertions for MLE (top) and deterministic (bottom) analysis. Each curve corresponds to a different value of insDJ, ranging from 0 (blue) to 10. The curves collapse for MLE indicating independence.

[2] We use the known alleles for each gene listed in the IMGT data base [11] augmented by a few additional variants observed in the data (see SI Appendix for details).

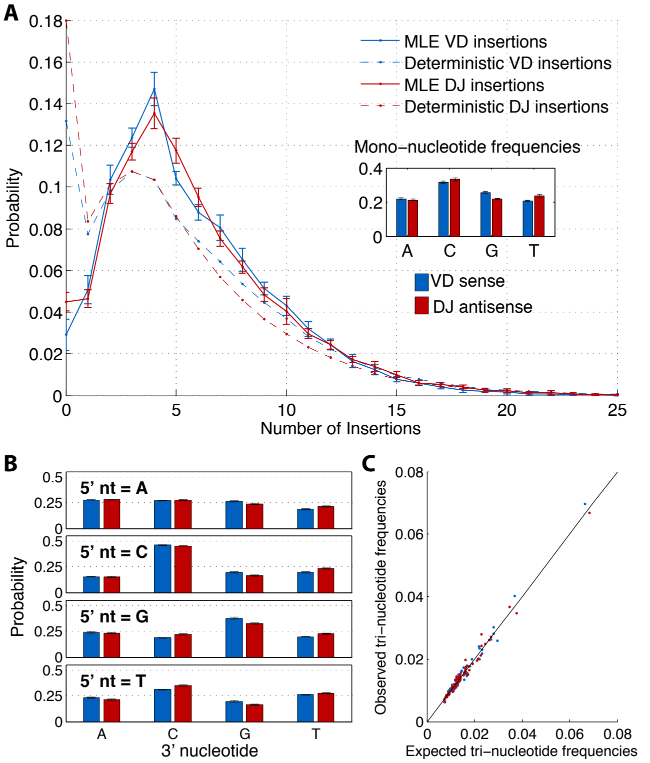

Fig. 3. Statistics of VD and DJ insertions. (A) Insertion length profiles: maximum likelihood estimate (deterministic estimate) displayed as solid (dashed) lines; error bars show variation across the nine individuals. The distribution tail is accurately exponential. The deterministic estimate greatly overestimates the frequency of zero insertions. Inset: mono-nucleotide utilization bias. (B) Dinucleotide utilization in insertions; the bias in DJ insertions is very accurately the reverse complement of the VD insertion bias.(C) Higher-order nucleotide bias in VD (blue) and DJ (red) insertions is completely accounted for by dinucleotide statistics.

$P_{\rm recomb}$, these usage patterns show the most relative variation between individuals.

The pattern of D-gene use conditioned on J-gene choice (Fig. S4D) reveals the known mechanistic constraint prohibiting utilization of D-genes that lie $3'$ of the chosen J-gene [1, 5]. The inferred distribution assigns a total probability of less than 0.1% for joining events using TRBD2 and any TRBJ1 gene. We note that such a determination is impossible without probabilistic analysis due to the uncertainty in identifying genes in specific sequences. The dependence between V gene choice and D or J gene choice is very weak to non-existent (with mutual information less than 0.01 bits). Thus, we believe that previously reported correlations in the use of these genes [15] reflect the effects of selection rather than VDJ recombination. Finally, we note the presence of pseudo V-genes, which occur in almost 10% of the non-productive CDR3s (see *SI Appendix* for details).

Nucleotide insertions. In Fig. 3 we show the factors related to insertions in the inferred distribution $P_{\rm recomb}(E)$. The VD and DJ insertions are uncorrelated (Fig. 2) and their length distributions are nearly identical, with exponential tails (Fig. 3A). The nucleotide frequencies in the inserted segments are not uniform and are well explained by a di-nucleotide Markov model where the probability of inserting A, C, G, or T depends on the immediately 5' nucleotide (see Fig. 3B). The VD inserted segment, on the sense strand, and the DJ inserted segment, on the antisense strand, show a preference for Cs. The frequencies of tri-nucleotides are almost perfectly accounted for by the di-nucleotide preferences (Fig. 3C), suggesting that the sequence statistics are fully captured by dinucleotide statistics. Additionally, the VD insertion dinucleotide bias, taken on the sense strand in the 5'-3' direction, is virtually identical to the DJ insertion di-nucleotide bias, taken on the antisense strand in the 5'-3' direction. This suggests that the mechanism of junctional nucleotide insertions is strand specific and occurs on opposite strands for the VD and DJ junctions. The molecular mechanistic basis of these features is not evident.

Nucleotide deletions. Because there is a strong correlation between number of deletions and gene identity (see the entries for $I({\rm del}V, V)$ and $I({\rm del}J, J)$ in Fig. 2), we allow for gene-dependent deletion profiles in $P_{\rm recomb}(E)$ (Eqn. 1). The results for a few genes are shown in Fig. 4A (see Figs. S12-S16 for all the profiles). P-nucleotides are counted as negative deletions as they occur only in association with zero nucleotide deletions (Fig. S10). The profiles have substantial variation from gene to gene, suggestive of a nuclease activity that depends on sequence context, but they are highly consistent between individuals. We have modeled this context dependence using a position weight matrix summing independent

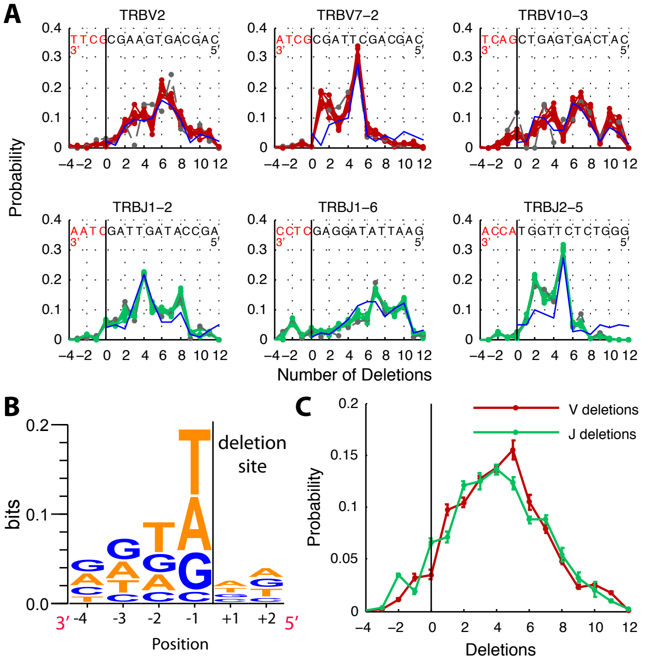

Fig. 4. (A) Gene-specific deletion profiles for selected V (red) and J (green) genes: the profiles vary widely from gene to gene, but are nearly identical across individuals (all nine are plotted; one in grey from an individual with significantly smaller sample size). The blue curves in all panels show the predictions of a simple model for the sequence context dependence of deletion probabilities using a position weight matrix (PWM), fit to the V deletion profiles (see *SI Appendix* Sec. 12 for details). The model ignores P-nucleotide generation and lacks any effects of distance from the gene end but performs reasonably well ($r^2 = 0.7$). (B) Sequence logo of the context dependence of deletion probability, from the PWM fit to the V deletion profiles. Only positions $3'$ of the deletion site have strong effects on the probability. (C) Cumulative deletion profiles for V-genes and J-genes. Error bars indicate variation across individuals.

[3] Recall that this estimate is for the β-chain only. The α-chain will yet add more diversity to this estimate.

contributions from the bases in a 6 nucleotide window (four $3'$ and two $5'$) around the cutting point to the log probability of deletion (see Fig. 4B and Fig. S11 for details). We find that only bases $3'$ of the deletion site have a strong effect on the probability, with T and A nucleotides having the greatest contribution, consistent with previous observations [16]. This simple model, which ignores both the P-nucleotides as well as the effects of distance from the end of the gene, does reasonably well in explaining the variation in deletion probabilities ($r^2 = 0.7$). This modeling is simply meant to suggest that the complexity of the observed deletion distributions may ultimately be explained by a parsimonious mechanistic model that reflects the underlying biochemistry of the deletion process.

Consistency of distributions across individuals. The insertion profiles, and the many different gene-dependent deletion profiles, are very consistent between individuals (Figs. 3, 4 and Figs. S12-16), suggesting the action of a universal molecular mechanism of rearrangement and providing convincing evidence against overfitting. We note that finite sample size statistics accounts for less than 50% of the observed inter-individual variance (indicated by the error bars) in some of our plots, possibly reflecting biological variation.

Potential diversity of repertoire. Our inferred distribution of recombination events (Eqn. 1) implies a probability distribution $P_{\mathrm{gen}}(\sigma)$ on the space of all CDR3 sequences (Eqn. 2) whose entropy $S_{\mathrm{seq}} = -\sum_\sigma P_{\mathrm{gen}}(\sigma) \log P_{\mathrm{gen}}(\sigma)$ is a measure of the potential sequence diversity of VDJ recombination. Since multiple recombination events can lead to the same sequence, we cannot calculate S_{seq} directly. We do, however, have an explicit description of P_{recomb}, the entropy of which we can calculate: $S_{\mathrm{recomb}} = 52$ bits; in addition, we can show that sequence entropy and recombination event entropy are related by

$$S_{\mathrm{seq}} = S_{\mathrm{recomb}} - \langle S(E|\sigma)\rangle_\sigma \simeq 47 \text{ bits}, \qquad [5]$$

where the correction term, $\langle S(E|\sigma)\rangle_\sigma \simeq 5$ bits, is the entropy of recombination events that give the same sequence (which we know for sequences in the repertoire as a byproduct of the inference), averaged over sequences. This means that CDR3 sequences can be generated in ~ 32 different ways, on average, by VDJ recombination; this is the fundamental reason why we must resort to probabilistic inference methods. The total sequence diversity of 47 bits corresponds to a potential CDR3 repertoire size of $\sim 10^{14}$ sequences[3]. This is to be compared with the estimated 4×10^6 unique CDR3 sequences in an individual [4, 17] , the $\sim 10^{11}$ T-cells in the blood of an individual [18] and the $\sim 10^{13}$ potential peptide-MHC complexes [19]. While convergent recombination means that the sequence entropy cannot be neatly partitioned into contributions from gene choice, deletions and insertions, the entropy of recombination events S_{recomb} can be so partitioned (Fig. 5A). We note that the bulk (60%) of the recombination entropy comes from the nucleotide insertions, and little from gene choice (5 bits from V and 4 bits from D and J) consistent with previous estimates [20]. For comparison, uniform usage of the genes would result in an entropy of 5.9 bits for V and 4.7 bits for D and J gene choices.

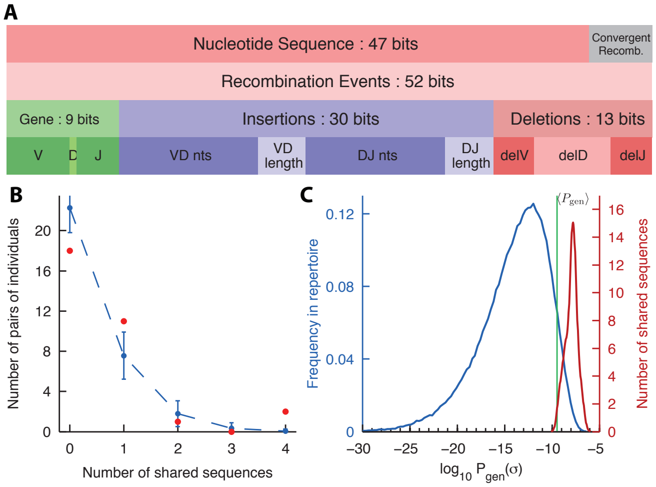

Fig. 5. (A) Entropy decomposition. Top bars: sequence entropy is smaller than recombination entropy by 5 bits because of convergent recombination; Bottom bars: recombination event entropy decomposed into contributions from gene choice, insertions, and deletions. (B) Statistics of the 21 CDR3 sequences shared between pairs of individuals: actual (red) vs. expected on the basis of the inferred $P_{\mathrm{gen}}(\sigma)$ (blue). (C) Histogram of $P_{\mathrm{gen}}(\sigma)$ for all sequences (blue) and for the 21 shared sequences (red, kernel density estimate); $\langle P_{\mathrm{gen}}\rangle$ for the full repertoire is indicated by the vertical green line.

Overlap of repertoires between individuals. Some sequences appear in the repertoires of more than one individual, and we can ask whether their number and specific identities are consistent with chance on the basis of our generative distribution $P_{\mathrm{gen}}(\sigma)$. We see evidence of inter-sample contamination in some of our data leading to a large number of shared sequences between specific individuals. Eliminating such questionable cases (see *SI Appendix* for details), we are left with 21 sequences that occur in the non-productive repertoires of two individuals and none that occur in more than two.

The total number of shared sequences between the repertoire samples of any pair of individuals with sample sizes N_1 and N_2 is expected to be Poisson distributed with mean $\bar{n} = N_1 N_2 \langle P_{\mathrm{gen}}\rangle_\sigma$ where $\langle P_{\mathrm{gen}}\rangle_\sigma = \sum_\sigma P^2_{\mathrm{gen}}(\sigma)$. Note that while the specific shared sequences are likely to have high probabilities of generation, the number of shared sequences, without regard to their identities, is determined by $\langle P_{\mathrm{gen}}\rangle_\sigma$ which is the average value of P_{gen} over the potential repertoire. We estimate this quantity to be $\langle P_{\mathrm{gen}}\rangle_\sigma \simeq 3.4 \pm 0.1 \times 10^{-10}$, by taking the mean of P_{gen} over the observed repertoire.

In Fig. 5B, we compare the expected number of pairs of individuals with a certain number of shared sequences (calculated as a sum of Poisson distributions over the pairs) to the observed number of such pairs, showing excellent agreement. The specific shared sequences have particularly high generation probabilities according to our distribution, with a median value of $\sim 10^{-8}$ compared to the repertoire median of $\sim 10^{-14}$ (Fig. 5C). Because the generative distribution is trained on individual repertoires, and is highly consistent between individuals, its success in accounting for recurring sequences between individuals is a non-trivial test of its validity. We find similar results for the shared sequences among the memory repertoires (Fig. S6).

Convergent recombination has been proposed as an explanation for the occurrence of 'public' T-cell receptors [21, 22, 23]. However, the recombination entropy $S(E|\sigma)$ is only weakly correlated with the generation probability $P_{\mathrm{gen}}(\sigma)$ (correlation coefficient 0.13, see Fig. S7), and we find that the shared non-productive sequences in our data do not have higher recombination entropies than other sequences.

Results from other repertoires. Inference of $P_{\mathrm{recomb}}(E)$ from the non-productive memory repertoires of the same nine individuals leads to results identical with those reported above for the naïve non-productive repertoires (Figs. S5,6). The consistency of the inferred generative distribution between these repertoires as well as between the nine individuals is strong evidence that the non-productive CDR3 sequence statistics, memory or naïve, reflect only the basic recombination process and not selection. In Fig. S8, we show the distributions of generation probabilities of CDR3 sequences from the productive repertoires. While it is tempting to apply our algorithm to the productive sequence repertoires, it would be inconsistent to do so: these sequences have passed selection filters, thymic and adaptive, and we have no analog of Eqn. 1 to parametrize the probability of such success. This is an important subject for future investigation.

Discussion

We have presented a method for inferring the statistics of VDJ recombination events from the large T-cell receptor sequence repertoires that are made available by high-throughput sequencing. We emphasize the crucial importance of using a probabilistic approach: the typical CDR3 sequence can be produced by about 32 different recombination events, and using a deterministic assignment of events to each sequence results in systematic biases and spurious correlations. Our general approach allows us to cope with not-yet-indexed alleles [13] and, most importantly, with sequencing errors, an essential task given the rapid growth of high-throughput but error-prone sequencing technologies.

Since we focus on non-productive sequences, our results describe the probability distribution over CDR3 sequences produced by the recombination machinery *before any functional selection has occurred.* Its remarkable reproducibility across individuals and repertoires (naïve and memory) provides compelling evidence for the consistency and accuracy of our method. The obtained distribution is a central feature of the adaptive immune system and serves as a baseline (or, in evolutionary terms, a neutral model) for analyzing the subsequent processes of the immune system. By calculating the entropy of the generative distribution, we can estimate the potential diversity of the CDR3 sequences ($\sim 10^{14}$ sequences) and the contributions of insertions, deletions and gene choices to this entropy. We find that insertions contribute most (60%) of the diversity.

We are able to evaluate the probability of generating *any* specific CDR3 sequence (including as yet unobserved ones). This probability could be used to estimate the strength of selection on a sequence or group of sequences, or the likelihood that a sequence is shared between individuals or repertoires. Thus, it could help better characterize the significance of shared or 'public' T-cell receptor sequences [23]. We have verified that the sequences that are shared between the non-productive repertoires of different individuals in our data are consistent with the predictions of the inferred probability distribution (Fig. 5B,C), a very stringent test of its accuracy.

The recombination event distributions also provide insight into the molecular mechanism of recombination, and should serve as a starting point for detailed mechanistic models of recombination. We find that the recombination processes at the two junctions are essentially independent of each other, and that insertion events are independent of gene choice and deletions. The inferred distribution confirms that a D-gene can only recombine with downstream J-genes. We derive a precise model for the composition of inserted nucleotides, based solely on frequencies of di-nucleotides. We also show that a relatively crude model of sequence-specific nuclease activity can account for the deletion probabilities reasonably well. Our observed distribution, which is specified by a large number of probabilities, should be reproduced by parsimonious, but more realistic, mechanistic models.

We have focused on characterizing the molecular generation of nucleotide sequences that code for T-cell receptors. The functional receptor repertoire is first shaped by this molecular process and then by thymic selection and adaptation to pathogens. Quantitative models of the latter processes are needed for understanding the adaptive immune system. While the underlying biochemistry conveniently served to parametrize our sequence distributions, finding an analogous functionally relevant parametrization of amino-acid sequences to model the effects of selection is much more challenging [24]. Statistical analysis of the productive receptor repertoires, with our precise characterization of the unselected repertoire in hand, will hopefully aid in this effort.

ACKNOWLEDGMENTS. We are grateful to H. Robins and collaborators for making the datasets on which this work is based available to us. The work of CGC was supported in part by National Science Foundation grant PHY-0957573 and by US Department of Energy grant DE-FG02-91ER40671. The work of AM was supported in part by the NSF Physics of Living Systems program (PHY-1022140). CGC thanks the Institute for Advanced Study for hospitality during the performance of part of this work. The authors declare no conflicts of interest.

1. Murphy KP, Travers P, Walport M, Janeway C (2008) Janeway's immunobiology (Garland, New York).
2. Freeman JD, Warren RL, Webb JR, Nelson BH, Holt RA (2009) Profiling the T-cell receptor beta-chain repertoire by massively parallel sequencing. Genome research 19:1817–1824.
3. Weinstein JA, et al. (2009) High-throughput sequencing of the zebrafish antibody repertoire. Science 324:807–810.
4. Robins HS, et al. (2009) Comprehensive assessment of T-cell receptor beta-chain diversity in alphabeta T cells. Blood 114:4099–4107.
5. Robins HS, et al. (2010) Overlap and effective size of the human CD8+ T cell receptor repertoire. Science translational medicine 2:47ra64–47ra64.
6. Schatz DG, Swanson PC (2011) V (D) J recombination: mechanisms of initiation. Annual review of genetics 45:167–202.
7. Verkaik N, et al. (2002) Different types of V(D)J recombination and end-joining defects in DNA double-strand break repair mutant mammalian cells. European Journal of Immunology 32:701–709.
8. Lieber MR (2010) The mechanism of double-strand DNA break repair by the nonhomologous DNA end-joining pathway. Annual Review of Biochemistry 79:181–211.
9. Lieber MR, Wilson TE (2010) SnapShot: Nonhomologous DNA end joining (NHEJ). Cell 142:496–496.e1.
10. Lafaille JJ, DeCloux A, Bonneville M, Takagaki Y, Tonegawa S (1989) Junctional sequences of T cell receptor gamma delta genes: implications for gamma delta T cell lineages and for a novel intermediate of V-(D)-J joining. Cell 59:859–870.
11. Monod MY, Giudicelli V, Chaume D, Lefranc MP (2004) IMGT/JunctionAnalysis: the first tool for the analysis of the immunoglobulin and T cell receptor complex V-J and V-D-J Junctions. Bioinformatics 20:i379–i385.
12. McLachlan GJ, Krishnan T (2008) The EM Algorithm and Extensions (Wiley Series in Probability and Statistics) (Wiley-Interscience, Hoboken), 2nd edition.
13. Wang Y, et al. (2011) Genomic screening by 454 pyrosequencing identifies a new human IGHV gene and sixteen other new IGHV allelic variants. Immunogenetics 63:259–265.
14. Dempster A, Laird N, Rubin D (1977) Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society, Series B 39:1–38.
15. Wallace ME, et al. (2000) Junctional biases in the naive TCR repertoire control the CTL response to an immunodominant determinant of HSV-1. Immunity 12:547–556.
16. Gauss GH, Lieber MR (1996) Mechanistic constraints on diversity in human V(D)J recombination. Molecular and cellular biology 16:258–269.
17. Arstila TP, et al. (1999) A direct estimate of the human alphabeta T cell receptor diversity. Science (New York, N.Y.) 286:958–961.
18. Blum K, Pabst R (2007) Lymphocyte numbers and subsets in the human blood: Do they mirror the situation in all organs? Immunology Letters 108:45–51.
19. Mason D (1998) A very high level of crossreactivity is an essential feature of the T-cell receptor. Immunology today 19:395–404.
20. Cabaniols JP, Fazilleau N, Casrouge A, Kourilsky P, Kanellopoulos JM (2001) Most alpha/beta T cell receptor diversity is due to terminal deoxynucleotidyl transferase. The Journal of experimental medicine 194:1385–1390.

21. Quigley MF, et al. (2010) Convergent recombination shapes the clonotypic landscape of the naive T-cell repertoire. Proceedings of the National Academy of Sciences of the United States of America 107:19414–19419.
22. Venturi V, et al. (2011) A mechanism for TCR sharing between T cell subsets and individuals revealed by pyrosequencing. Journal of immunology (Baltimore, Md. : 1950) 186:4285–4294.
23. Venturi V, Price DA, Douek DC, Davenport MP (2008) The molecular basis for public T-cell responses? Nature reviews. Immunology 8:231–238.
24. Mora T, Walczak AM, Bialek W, Callan CG (2010) Maximum entropy models for antibody diversity. Proceedings of the National Academy of Sciences of the United States of America 107:5405–5410.