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

Anand Murugan[a], Thierry Mora[b], Aleksandra M. Walczak[c], and Curtis G. Callan, Jr.[a,d,1]

[a]Joseph Henry Laboratories, Princeton University, Princeton, NJ 08544; [b]Laboratoire de Physique Statistique, UMR8550, Centre National de la Recherche Scientifique and École Normale Supérieure, 24 rue Lhomond, 75005 Paris, France; [c]Laboratoire de Physique Théorique, UMR8549, Centre National de la Recherche Scientifique and École Normale Supérieure, 24 rue Lhomond, 75005 Paris, France; and [d]Simons Center for Systems Biology, Institue for Advanced Study, Princeton, NJ 08544

Contributed by Curtis G. Callan, Jr., July 27, 2012 (sent for review June 19, 2012)

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 nonproductive 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 probabilistic model 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.

convergent recombination | expectation maximization | palindromic nucleotides | insertion/deletion profiles

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 T-cell sequence repertoires that are becoming available via high-throughput sequencing technology (2–5). In particular, we focus purely on a subset of receptor sequences that are nonproductive, 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. 1*A* 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 that assigns a unique event to each sequence (2–4). However, because individual CDR3 sequences can arise in multiple ways (see 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 (11) and apply it to recent data on human T-cell receptor sequences.

Analysis

We work with sequence data on CD4+ T-cell beta chain CDR3 regions obtained from nine human subjects by methods described in refs. 4 and 5 (see Acknowledgments). 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′

Author contributions: A.M., T.M., A.M.W., and C.G.C. designed research, performed research, analyzed data, and wrote the paper.

The authors declare no conflict of interest.

[1]To whom correspondence should be addressed. E-mail: ccallan@princeton.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1212755109/-/DCSupplemental.

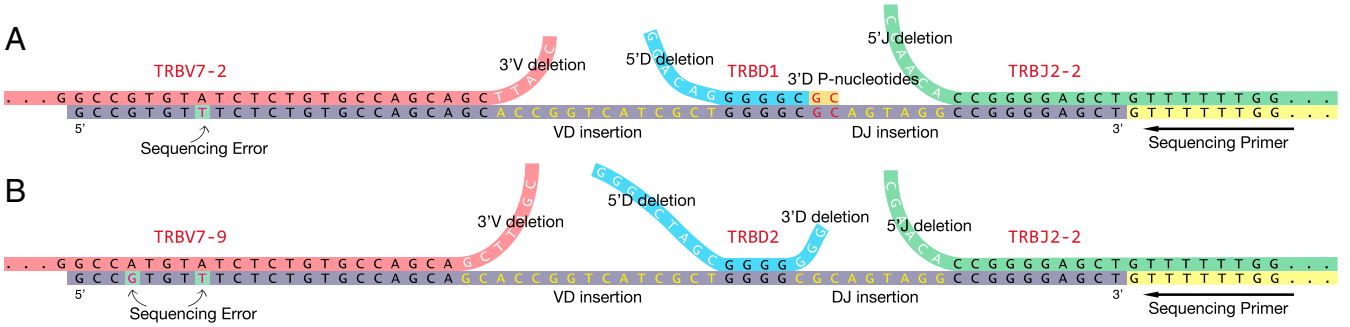

Fig. 1. A 60 bp CDR3 read (gray box) can be aligned to different genes [nomenclature follows IMGT conventions (24)] with different deletions (white), insertions (yellow), and P-nucleotides (red). (*A*) Alignment to specific V-, D-, and J-genes with insVD = 13, insDJ = 6, delV = 5, delJ = 6, del5′D = 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.

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 (4, 5). This process produces a dataset 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 "nonproductive," i.e., 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*, section 10 for evidence that the non-productive constraint introduces no bias). Because this is our primary concern, we focus our analysis on the nonproductive 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 be able to verify the absence of selection effects. Our data sets are available online (see *SI Appendix*, sections 1 and 2 for details).

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* (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):

$$\begin{aligned} P_{\text{recomb}}(E) &= P(V)P(D,J) \\ &\times P(\text{del}V|V)P(\text{del}J|J)P(\text{del5}'D, \text{del3}'D|D) \\ &\times P(\text{ins}VD)\prod_{i=1}^{\text{ins}VD} p_{VD}^{(2)}(x_i|x_{i-1})P(\text{ins}DJ)\prod_{i=1}^{\text{ins}DJ} p_{DJ}^{(2)}(y_i|y_{i-1}). \quad [1] \end{aligned}$$

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 Eq. **1**: we treat them as "negative" deletions (i.e., a palindrome of half-length 2, as in Fig. 1*A*, 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 (see *SI Appendix*, section 11 and Fig. S10). Thus, true P-nucleotides, 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 $P_{\text{recomb}}(E)$ [Eq. **1**] are probability distributions on event variables that take on a finite number of values. Specifying this joint distribution requires a total of 2,865 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.

*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.

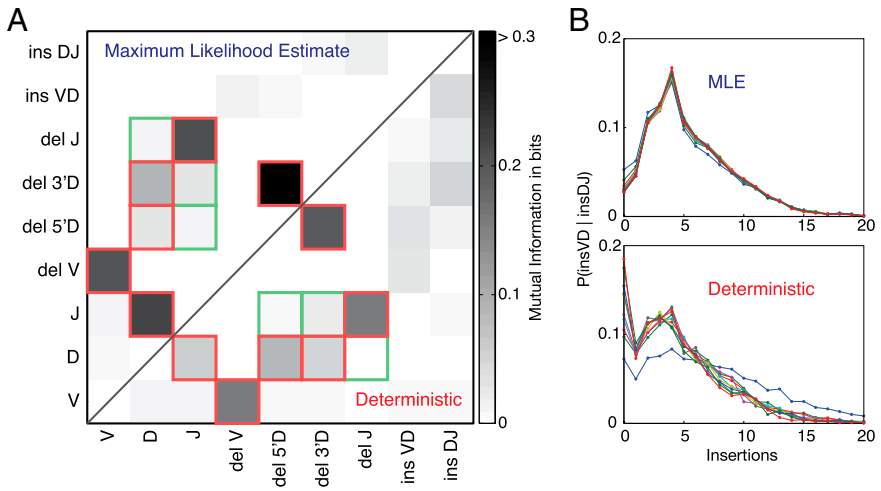

Fig. 2. (*A*) Data-derived correlations between sequence features: Each entry is the mutual information $I(X, Y)$ of a feature pair over the naïve nonproductive repertoire. The outlined elements are correlations expected from the form of $P_{\text{recomb}}(E)$: Red identifies a direct effect of a factor in Eq. **1** (e.g., $D \leftrightarrow J$) and green indirect effects (e.g., $D \leftrightarrow J \leftrightarrow$ delJ). The top-left half of the matrix shows results from the 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 (*Upper*) and deterministic (*Lower*) analysis. Each curve corresponds to a different value of insDJ, ranging from 0 (blue) to 10. The curves collapse for MLE indicating independence.

Generation Probability and Likelihood of Observed Sequences. The probability $P_{\text{gen}}(\sigma)$ of generating a specific CDR3 sequence σ is the sum of the probabilities of all recombination events E_σ that produce σ:

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

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:

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

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

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[†] 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 (12). 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 dataset $\mathcal{D}$ is then the product over 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 (11, 13) (see *SI Appendix* for algorithmic details). In order to identify universal features of the diversity generation machinery, we perform this inference separately for each individual subject. Our analysis software is available online (see *SI Appendix* for details).

Results

In what follows, we present results of our analysis of naïve, nonproductive, CDR3 sequence repertoires of nine individuals (see *SI Appendix* for a parallel analysis of memory sequence repertoires). Selected results data files are available online (see *SI Appendix* for details).

Correlations Between Event Variables. It is important to verify that correlations not present in the assumed structure of the probability distribution [Eq. **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. 2*A*, where the entries outlined in red are dependences accounted for by individual factors in our assumed form of $P_{\text{recomb}}(E)$ [Eq. **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 $P_{\text{recomb}}(E)$. There are a few detectable correlations that are not consistent with the assumed structure: $(\text{ins}VD, \text{del}V)$, $(\text{ins}DJ, \text{del}J)$, 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. 2*A* 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

[†]We use the known alleles for each gene listed in the IMGT database (24) augmented by a few additional variants observed in the data (see *SI Appendix* for details).

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. 2 *B* and *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. S4 *A* and *C*). The usage frequencies are consistent between individuals, though of all the inferred parameters in P_{recomb}, these usage patterns show the most relative variation between individuals.

The pattern of D-gene use conditioned on J-gene choice (*SI Appendix*, Fig. S4*D*) 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 nonexistent (with mutual information less than 0.01 bits). Thus, we believe that previously reported correlations in the use of these genes (14) reflect the effects of selection rather than VDJ recombination. Finally, we note the presence of pseudo V-genes that occur in almost 10% of the nonproductive CDR3s (see *SI Appendix* for more details).

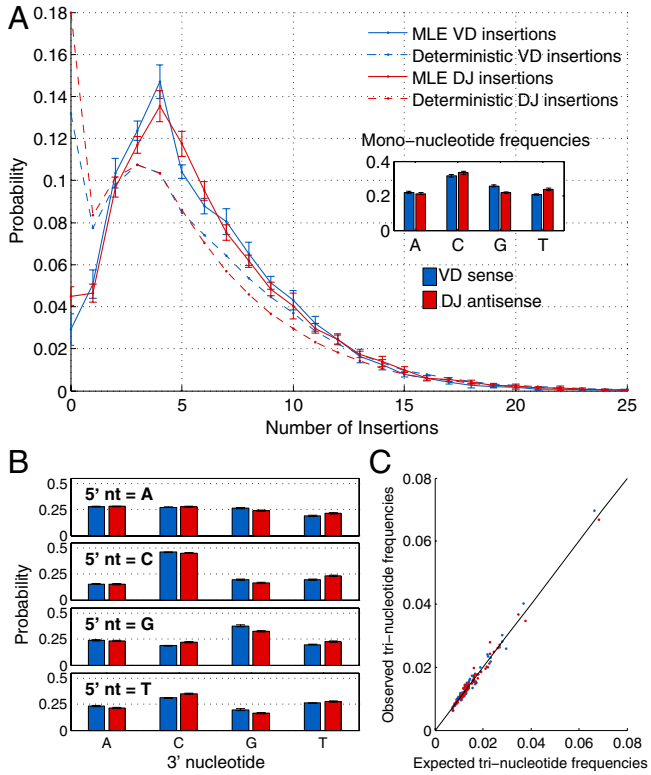

Fig. 3. Statistics of VD and DJ insertions. (*A*) Insertion length profiles: maximum likelihood estimate (deterministic estimate) displayed as solid (dashed) lines; error bars show variation across the nine individuals. The distribution tail is accurately exponential. The deterministic estimate greatly overestimates the frequency of zero insertions. *Inset*: mononucleotide 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.

Nucleotide Insertions. In Fig. 3 we show the factors related to insertions in the inferred distribution $P_{\text{recomb}}(E)$. The VD and DJ insertions are uncorrelated (Fig. 2) and their length distributions are nearly identical, with exponential tails (Fig. 3*A*). The nucleotide frequencies in the inserted segments are not uniform and are well explained by a dinucleotide Markov model where the probability of inserting A, C, G, or T depends on the immediately 5′ nucleotide (see Fig. 3*B*). 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 trinucleotides are almost perfectly accounted for by the dinucleotide preferences (Fig. 3*C*), 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 dinucleotide 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 $I(\text{del}V, V)$ and $I(\text{del}J, J)$ in Fig. 2), we allow for gene-dependent deletion profiles in $P_{\text{recomb}}(E)$ [Eq. **1**]. The results for a few genes are shown in Fig. 4*A* (see *SI Appendix*, Figs. S12–S16 for all the profiles). P-nucleotides are counted as negative deletions as they occur only in association with zero nucleotide deletions (see *SI Appendix*, 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

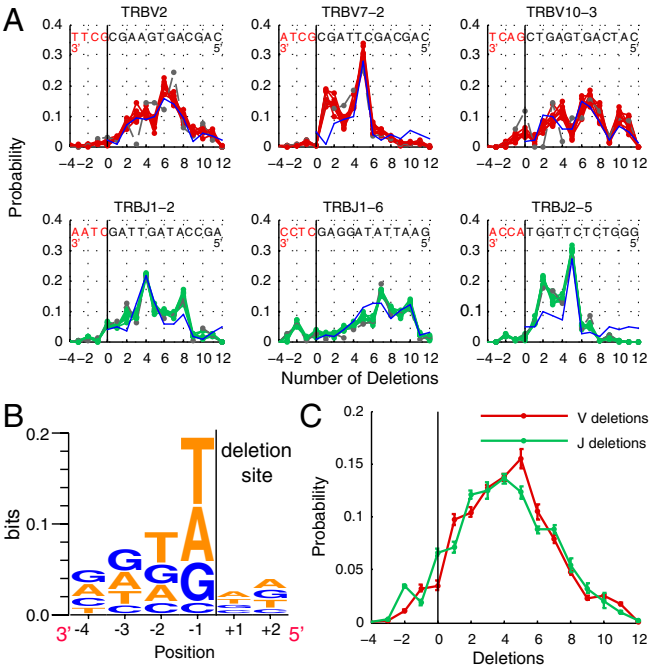

Fig. 4. (*A*) Gene-specific deletion profiles for selected V (red) and J (green) genes: The profiles vary widely from gene to gene but are nearly identical across individuals (all nine are plotted; one in gray from an individual with significantly smaller sample size). The blue curves in all panels show 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* 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.

individuals. We have modeled this context dependence using a position weight matrix summing independent contributions from the bases in a six nucleotide window (four 3′ and two 5′) around the cutting point to the log probability of deletion (see Fig. 4*B* and *SI Appendix*, 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 (15). 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 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 and 4 and *SI Appendix*), suggesting the action of a universal molecular mechanism of rearrangement and providing convincing evidence against overfitting. We note that finite sample size statistics account for less than 50% of the observed interindividual 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 [Eq. **1**] implies a probability distribution $P_{\text{gen}}(\sigma)$ on the space of all CDR3 sequences [Eq. **4**] whose entropy $S_{\text{seq}} = -\sum_\sigma P_{\text{gen}}(\sigma) \log P_{\text{gen}}(\sigma)$ is a measure of the potential sequence diversity of VDJ recombination. Because 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: $S_{\text{recomb}} = 52$ bits; in addition, we can show that sequence entropy and recombination event entropy are related by

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

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 our inference), averaged over sequences. This means that CDR3 sequences can be generated in approximately 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 approximately 10^{14} sequences[‡]. This is to be compared with the estimated 4×10^6 unique CDR3 sequences in an individual (4, 16) , the approximately 10^{11} T cells in the blood of an individual (17) and the approximately 10^{13} potential peptide-MHC complexes (18). Although 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. 5*A*). 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 (19). 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.

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 $P_{\text{gen}}(\sigma)$. Some shared sequences appear simultaneously in too many repertoires to be valid and are probably due to intersample contamination (see *SI Appendix* for details). Eliminating clearly identifiable questionable cases, we are left with 21 sequences that occur in the nonproductive repertoires of two individuals and none that occur in more than two.

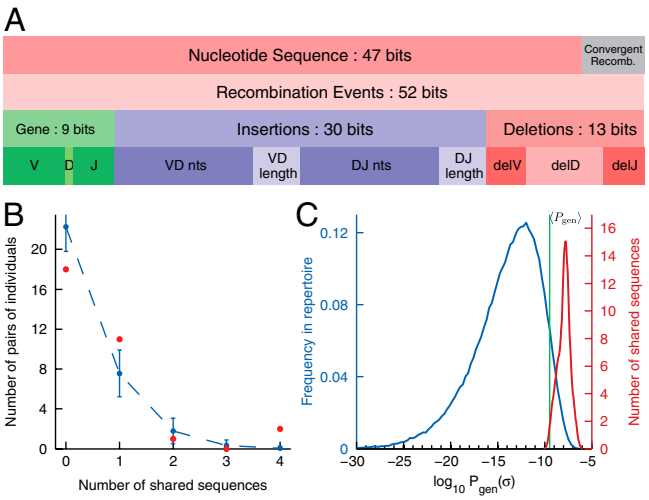

Fig. 5. (*A*) Entropy decomposition. Top bars: Sequence entropy is smaller than recombination entropy by 5 bits because of convergent recombination; Bottom bars: Recombination event entropy decomposed into contributions from gene choice, insertions, and deletions. (*B*) Statistics of the 21 CDR3 sequences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

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 $\bar{n} = N_1 N_2 \langle P_{\text{gen}}\rangle_\sigma$ where $\langle P_{\text{gen}}\rangle_\sigma = \sum_\sigma P_{\text{gen}}^2(\sigma)$. Note that although 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 $\langle P_{\text{gen}}\rangle_\sigma$, which is the average value of P_{gen} over the potential repertoire. We estimate this quantity to be $\langle P_{\text{gen}}\rangle_\sigma \simeq 3.4 \pm 0.1 \times 10^{-10}$ by taking the mean of P_{gen} over the observed repertoire.

In Fig. 5*B*, 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 approximately 10^{-8} compared to the repertoire median of approximately 10^{-14} (Fig. 5*C*). 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 nontrivial test of its validity. We find similar results for the shared sequences among the memory repertoires (see *SI Appendix*, Fig. S6).

Convergent recombination has been proposed as an explanation for the occurrence of "public" T-cell receptors (20–22). However, the recombination entropy $S(E|\sigma)$ is only weakly correlated with the generation probability $P_{\text{gen}}(\sigma)$ (correlation coefficient 0.13, see *SI Appendix*, Fig. S7), and we find that the shared nonproductive sequences in our data do not have higher recombination entropies than other sequences.

Results from Other Repertoires. Inference of $P_{\text{recomb}}(E)$ from the nonproductive memory repertoires of the same nine individuals leads to results identical with those reported above for the naïve

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

nonproductive repertoires (see *SI Appendix*, Figs. S5 and S6). The consistency of the inferred generative distribution between these repertoires as well as between the nine individuals is strong evidence that the nonproductive CDR3 sequence statistics, memory or naïve, reflect only the basic recombination process and not selection. In *SI Appendix*, Fig. S8 we show the distribution of generation probabilities of CDR3 sequences from the productive repertoires. Although it is tempting to apply our approach 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 Eq. **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 (12) and, most importantly, with sequencing errors, an essential task given the rapid growth of high-throughput but error-prone sequencing technologies.

Because we focus on nonproductive 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 (approximately 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 (22). We have verified that the sequences that are shared between the nonproductive repertoires of different individuals in our data are consistent with the predictions of the inferred probability distribution (Fig. 5 *B* and *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 dinucleotides. 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. Whereas 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 (23). 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 C.G.C. was supported in part by National Science Foundation (NSF) Grant PHY-0957573 and by US Department of Energy Grant DE-FG02-91ER40671. The work of A.M. was supported in part by the NSF Physics of Living Systems program (PHY-1022140). C.G.C. thanks the Institute for Advanced Study for hospitality during the performance of part of this work.

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 Res* 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 repertoire. *Sci Transl Med* 2:47ra64.
6. Schatz DG, Swanson PC (2011) V (D) J recombination: Mechanisms of initiation. *Annu Rev Genet* 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. *Eur J Immunol* 32:701–709.
8. Lieber MR (2010) The mechanism of double-strand DNA break repair by the nonhomologous DNA end-joining pathway. *Annu Rev Biochem* 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. McLachlan GJ, Krishnan T (2008) *The EM Algorithm and Extensions* (Wiley-Interscience, Hoboken), 2nd Ed.
12. 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.
13. Dempster A, Laird N, Rubin D (1977) Maximum likelihood from incomplete data via the EM algorithm. *J R Stat Soc Series B Stat Methodol* 39:1–38.
14. 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.
15. Gauss GH, Lieber MR (1996) Mechanistic constraints on diversity in human V(D)J recombination. *Mol Cell Biol* 16:258–269.
16. Arstila TP, et al. (1999) A direct estimate of the human alphabeta T cell receptor diversity. *Science* 286:958–961.
17. Blum K, Pabst R (2007) Lymphocyte numbers and subsets in the human blood: Do they mirror the situation in all organs? *Immunol Lett* 108:45–51.
18. Mason D (1998) A very high level of crossreactivity is an essential feature of the T-cell receptor. *Immunol Today* 19:395–404.
19. Cabaniols JP, Fazilleau N, Casrouge A, Kourilsky P, Kanellopoulos JM (2001) Most α/β T cell receptor diversity is due to terminal deoxynucleotidyl transferase. *J Exp Med* 194:1385–1390.
20. Quigley MF, et al. (2010) Convergent recombination shapes the clonotypic landscape of the naive T-cell repertoire. *Proc Natl Acad Sci USA* 107:19414–19419.
21. Venturi V, et al. (2011) A mechanism for TCR sharing between T cell subsets and individuals revealed by pyrosequencing. *J Immunol* 186:4285–4294.
22. Venturi V, Price DA, Douek DC, Davenport MP (2008) The molecular basis for public T-cell responses? *Nat Rev Immunol* 8:231–238.
23. Mora T, Walczak AM, Bialek W, Callan CG (2010) Maximum entropy models for antibody diversity. *Proc Natl Acad Sci USA* 107:5405–5410.
24. Monod MY, Giudicelli V, Chaume D, Lefranc MP (2004) IMGT/Junction analysis: 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.

Statistical inference of the mechanisms of T-cell receptor diversity generation from sequence repertoires: Supporting Information

Anand Murugan, Thierry Mora, Aleksandra M. Walczak and Curtis G. Callan, Jr.

1 Sequences of V, D, and J-genes and their alleles

Accurate knowledge of the sequences of germ line V-, D-, and J-genes and their allelic variants is essential to minimize errors and bias in our analysis. There are 2 D-genes, 13 J-genes, and 48 V-genes, not counting alleles. There are in addition 19 'pseudo' V-genes on the same germline chromosome: they participate in the recombination process and, though they cannot lead to a functioning receptor, they can appear in the non-productive sequence data sets, provided that a sequencing primer (or an approximate one) is present, which in our case is true for 11 pseudo V-genes.

We curated a list of known and discovered allelic variants of the V-genes by combining those found in the public IMGT database [1] with variants that we discovered with high confidence during our analysis. Not all the sequence reads listed in IMGT are true variants since many of them are from rearranged DNA with variation at the junctional end. Such 'variants' were removed from our list, unless the variation was deeper in the sequence, far from the edited end. In addition, we have found three instances of allelic variants in our data that are not listed in IMGT. The discovered variants of genes TRBV7-7 and TRBV10-1 can actually be found by BLAST in the NCBI database of human sequences; the variant of gene TRBV7-2 is not found by BLAST and appears to be completely novel. Undiscovered variants have rather small impact on overall recombination event statistics, but they can cause systematic errors in the inference of gene-specific deletion profiles.

Complete lists of the genes and alleles used in our analysis are available online[1]. For completeness, we also list the primers used by Robins et. al. [2, 3] in acquiring the data we analyze.

2 CDR3 sequence data files and formats

The CDR3 sequences used in our analysis come from naïve or memory CD4+ T-cells of 9 human individuals, and are further segregated into 'in-frame' and 'non-productive' sequences. The sequences are 60bp in length for 6 of the subjects, and 101bp in length for the remaining three. The reads of different length differ only in how far the sequencing window goes into the V gene: both types are anchored on the same conserved phenylalanine in the J-gene and have the same read depth into the J-gene.

Processed sequence data was made available to us by H. Robins. As described in [2, 3] each sequence is read multiple times and the multiple reads are used to estimate the multiplicity of each specific TCR receptor in its respective compartment. In addition, multiple reads are used to correct for sequencing errors by clustering reads that differ at a small number of positions [2]. In our data files, the effective sequence multiplicity is recorded along with the

[1] `physics.princeton.edu/~ccallan/TCRPaper/genes`

error-corrected sequence (although we do not use multiplicity in our current analysis). The data files used in our analysis are available online[2]. The file names in the repository clearly indicate the category to which the included data belongs.

3 Overall description of the analysis pipeline and software

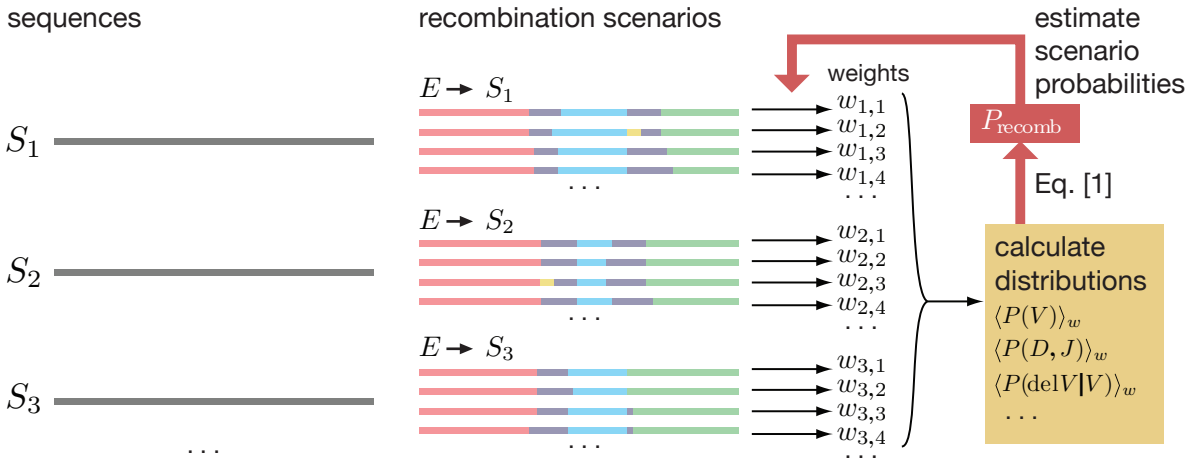

Fig. S1: Flow chart of the analysis pipeline.

There are two major steps in the analysis pipeline that leads from a list of CDR3 sequences to a final estimate of the probability distribution $P_{\text{recomb}}(E)$ of generative recombination events. The first is an 'alignment' step in which, for each read σ, we create a comprehensive list of recombination 'scenarios' $\{E_\sigma\}$ that could plausibly have produced that read. A 'scenario' is a particular set of values for the event variables (gene identities, VD insertions, etc.) that generates a recombined sequence nearly identical to the read in question (with possibly a small number of mismatches). The second major step is an iterative procedure (summarized in the flow chart of Fig. S1) for finding the generative distribution that maximizes the likelihood of the observed data given the functional form of the generative distribution (as expressed in main text Eqn. 2).

The algorithms we have developed to execute these two steps are described in greater detail in the following two subsections. Software to implement these procedures was written in Matlab using the Parallel Computing toolbox and run on a Linux cluster. Compiling key routines into C++ using Matlab Coder greatly improved processing speed, allowing model inference on an individual data set to be completed in about 20 hours running on 8 processors. Our Matlab code, along with summary instructions on how to run it, is available online[3]

[2] `physics.princeton.edu/~ccallan/TCRPaper/data`
[3] `physics.princeton.edu/~ccallan/TCRPaper/scripts`

3.1 Initial parsing of sequence reads by alignment

The first step in our inference procedure is to align each CDR3 read with specific alleles of V, D, and J genes by sequence matching. The goal is to generate a set of plausible recombination events that could produce the read to serve as a starting point for subsequent probabilistic refinement. This preliminary alignment procedure produces, for each read, a finite number of V, D, and J alleles, the maximal length alignments of these alleles to the read, the corresponding minimum nucleotide deletions from the genomic sequences, with possible P-nucleotides identified, and with the unmatched parts of the read identified as VD or DJ insertions. Mismatch information is also stored.

Certain thresholds are imposed on the alignments – gene alignment lengths must be sufficiently long; gene deletions must not be too large; errors are allowed in the alignments (no gaps), but the number of errors must be small. The alignment score (using an appropriate mismatch penalty) is used to rank order alignments, and a threshold on the score relative to the score of the best alignment is also imposed. Specific values for these various parameters are chosen in the light of computational experience to achieve fast and accurate convergence of the overall model-fitting algorithm.

The procedure for finding J matches is simplest. The CDR3 reads all begin at the $3'$ end (sense strand) from a primer in a known position in each J gene. Thus for each candidate J gene, we simply look for exect matches of the end of the sequence read with the portion of the gene just $5'$ of the primer. Proceeding in this way, and imposing the various thresholds mentioned, we find an average of 2-3 J alignments per read.

For the V-gene, the position of alignment to the read is not fixed. So for a given V-gene, we align the $5'$ end of the read to the m-th base from the $3'$ end of the V-gene, and note the best-scoring match at this positioning (this time allowing some mismatches, and penalizing them in the score). We step through the values of m and record the best-scoring match over all positionings. Repeating this process for all the V-genes, and imposing the earlier mentioned thresholds, we are left with a limited set of possible V-gene identifications, together with their specific alignments to the read. Proceeding in this way, we find an average of ~ 15 V alignments per read.

After identifying the plausible alignments to V- and J- genes, we turn to the problem of identifying D-gene matches. This is a more difficult problem because the D-genes are short, and deletions (occurring on both ends) often leave residual sequences which are hard to identify as a D-gene fragment. We therefore put very loose constraints on the D-gene alignments, relying on the probabilistic refinement to narrow them down. Specifically, we consider the read sequence segment lying between the end of the highest-scoring V-gene and the end of the highest-scoring J-gene, and include 10 nucleotides of flanking sequence on either side, to allow for ambiguous origin of these bases. We identify as a possible D-gene match every maximal non-overlapping alignment to this segment of the three D-gene alleles. These D-gene matches are scored by their length and the top 200 are selected as possible D-gene alignments.

Alignment files are available online[4]: the files are in Matlab format and record the outcome of the above alignment strategy for a subset of our data. Inspection of the alignment data for individual sequences should provide instructive illustrations of the above-described procedure. The various thresholds and parameters used in the procedure are found in the files as well.

[4] `physics.princeton.edu/~ccallan/TCRPaper/results/alignments`

The full set of alignment files used in our analysis can be generated using routines provided in our online software repository.

We note that one could generate a unique assignment of sequence features to a given read by selecting from the alignment ensembles just described the V, D, and J assignments with the highest score (i.e. having the longest effective alignment with the read). We will call the occurrence distribution of gene assignments, insertions, and deletions produced in this way as the 'deterministic ' estimate of the sequence feature probability distribution. It corresponds to standard practice in the literature for inferring feature statistics from sequence data, and will be used as a benchmark for comparison and contrast with our more accurate probabilistically inferred distribution.

3.2 The expectation maximization algorithm

As described in the main text, we wish to find model parameters that maximize the likelihood of the data. We use an iterative Expectation-Maximization algorithm to do this. Given a current guess for the model parameters that describe $P_{\rm recomb}(E)$, we update it by calculating the probability-weighted counts of events over the data set and then using those counts to re-estimate the marginal distributions ($P(V)$, $P(D,J)$, $P(\text{ins}VD)$, and so on) that appear as factors in the general functional form of $P_{\rm recomb}(E)$ (main text Eqn. 2).

As indicated in main text Eqns. 2-4, the joint likelihood of a recombination event E and sequence σ is the product of two factors: the probability of the generative event (given by $P_{\rm recomb}(E)$), and the sum over allele choices a of the probability of those allele choices multiplied by the probability of the number of mismatches between σ and the sequence σ_E^a implied by E and a. In other words, in addition to the recombination event probability $P_{\rm recomb}(E)$, likelihood involves the sequencing error rate R and the allele probabilities $P(V_a|V)$, etc. We emphasize that we carry out this exercise independently for the data sets derived from different individuals. While we expect (and find) that $P_{\rm recomb}(E)$ is consistent between individuals, we of course expect different individuals to have different allele probabilities.

In the expectation maximization procedure, we start from a prior in which each factor in main text Eqn. 2 for $P_{\rm recomb}(E)$ is uniform in its variables, the sequencing error rate R is set to a small value (typically 10^{-4}), and the allele probabilities are uniform over all the alleles of each gene. Using main text Eqn. 4, for each CDR3 sequence read σ, we exhaustively compute the likelihoods of all recombination events E given σ, starting from maximal alignments for each sequence identified in the initial parsing of the read (previous section), and looping over the other scenarios, involving extra deletions compensated by chance re-insertions of identical nucleotides, that could also 'explain' the read. We also loop over the number of true P-nucleotides in the cases where they are present.

Normalizing these likelihoods yields the relative weights that observing the sequence σ assigns to different recombination events E, given the current model parameters. Summing these weighted occurrences over all the sequences in the data set gives a new, data-conditioned, estimate of the various factors that enter into the assumed general form of $P_{\rm recomb}(E)$ (as well as a new estimate of the sequencing error probability and allele occurrence frequencies). The formal statement of the update rule is as follows; for each parameter in the model that describes the probability of a specific recombination event feature X (say a particular V-gene choice) we update it to the probability weighted counts over the whole data set of that event.

In other words, the $(k+1)$-th iteration of the model parameters are given by

$$P^{(k+1)}(X) = \sum_{\sigma\in\mathcal{D}} \sum_{E} \delta_{X_E,X}\, P^{(k)}(E|\sigma)$$
$$= \sum_{\sigma\in\mathcal{D}} \sum_{E} \delta_{X_E,X}\, \frac{P^{(k)}(E,\sigma)}{L^{(k)}(\sigma)} \tag{1}$$

where $\delta_{X_E,X}$ is one if X is true in the recombination event E and zero otherwise. This procedure is used to update all the factors entering into the likelihood calculation and the process is repeated until convergence to a stable end point is achieved. Since all sequences in the data set are looped over in the calculation, we can record 'on the fly' the likelihood $L(\sigma)$ (main text Eqn. 4), the generation probability $P_{\rm gen}(\sigma)$ of that sequence (a conceptually different quantity), as well as the conditional entropy of events $S(E|\sigma)$ for each sequence quantifying the multiplicity of recombination events that could have produced the given CDR3 sequence). The product of $L(\sigma)$ over all sequences is the current overall likelihood of the data set, a measure of convergence of the procedure. The generation probabilities $P_{\rm gen}(\sigma)$ have a direct physical significance, reflecting the probability of generation of the sequence by the molecular machinery.

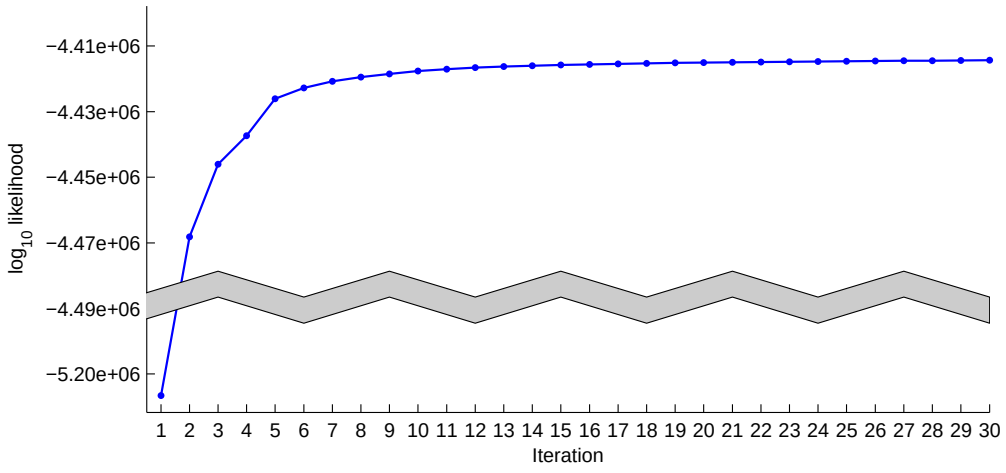

Fig. S 2: Convergence of the total likelihood of all data sets with iterations of the EM algorithm.

Iterating this process is guaranteed, by general expectation maximization arguments, to maximize the overall likelihood of the data set locally. We have found that rapid and direct convergence to a likelihood maximum is the norm for the data sets we work with (see Fig. S2). The models for the probability distribution of generative events inferred in this way from the different data sets are available online[5]. The distribution is also described in a Microsoft Excel file.

[5] `physics.princeton.edu/~ccallan/TCRPaper/results/models`

4 Sequencing error rate

The sequence mismatch rate in our model reflects both uncorrected sequencing error as well as unknown allelic variation. Our model assumes that this mismatch rate R is independent of position along the sequence read. As is well-known, accuracy of the sequencing procedure becomes worse at the end of the sequence read (the $5'$, or V-gene, end of our CDR3 sequence) so, in assaying error rates, we ignore the last 15 nucleotides (at the $5'$ end) for the 101 bp reads, where we can afford to do this. Our alignment procedure also disallows mismatches in the J- and D-gene alignment because of the shortness of these segments and the expected low error rate at this end (more accurately, the beginning) of the sequence read. In assessing position dependence of sequence error rates, therefore, we only need concern ourselves with mismatches to V gene assignments. Summing all such mismatches for the three individuals for which we have 101 bp reads, and plotting them against read position, we obtain the results plotted in Fig. S3. We find that R converges in the mean to a value of order 3×10^{-4} per base pair, two orders of magnitude smaller than the raw instrumental sequencing error rate. There are, however, a few sharp peaks at specific positions along the read; since they appear at the same position for different individuals, they presumably reflect some anomaly in the functioning of the sequencing machine. This shortcoming of the error rate model does not greatly influence the results of the inference because the overall error rate is rather low.

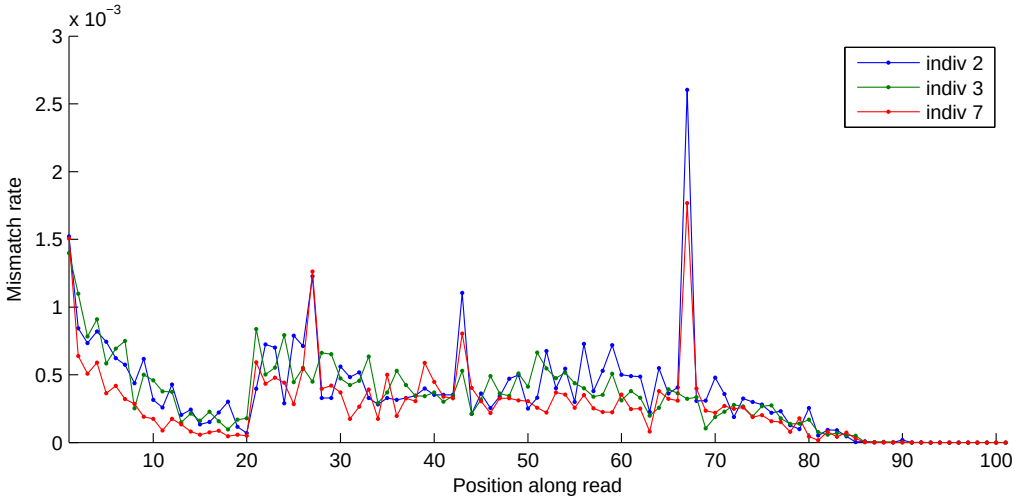

Fig. S3: Position-dependent error profile for the three individuals with read length 101 base pairs. The sequencing read proceeds from the right (101 to 1) where the J gene sequencing primer binds. The spikes in the error rate at specific positions (67, 43 and 27) are true sequencing error spikes and not the result of unknown allelic variants. Positions 1-15 show the characteristic increase in error rate with read length. The overall decreased error rate in positions 10-20 reflect our requirement of a minimum alignment length of 20 nucleotides to a V gene with an upper bound on the allowed errors in the alignment. Since we do not allow any errors in the J and D genes, the error rate is zero in this region.

5 Gene and pseudogene usage

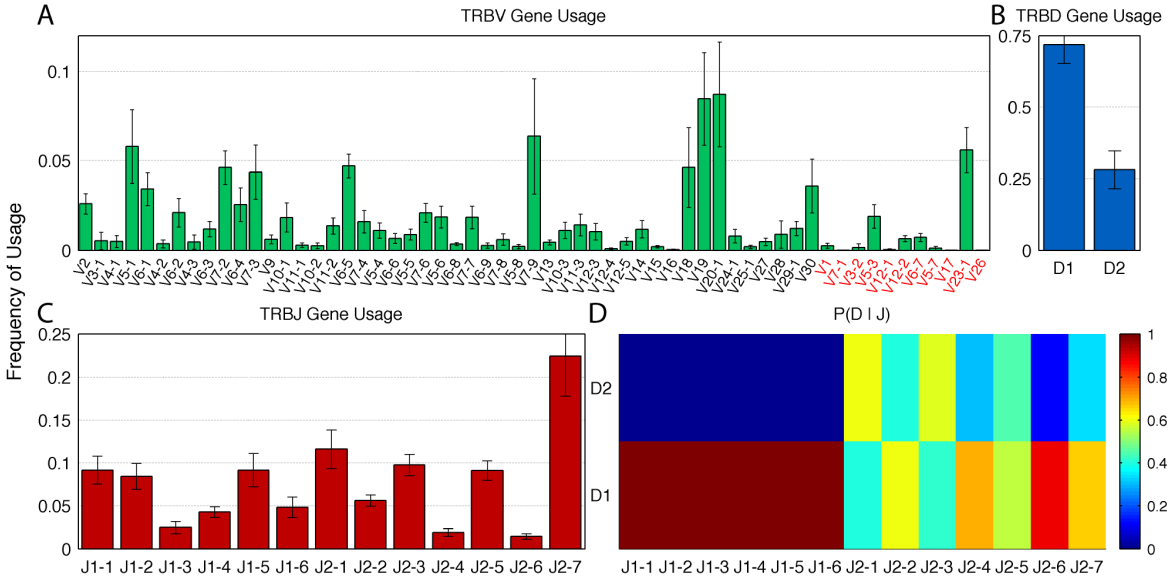

Fig. S4: Statistical aspects of gene usage. (A) Usage frequencies of V-genes, ordered by position on the chromosome, with the exception of pseudogenes (red legend). (B) Usage frequencies of the two D-genes. (C) Same for the 13 J-genes. (D) D-gene usage frequencies, conditioned on J-gene choice. As expected from the mechanistic constraint, TRBD2 has essentially zero probability ($< 0.1\%$) of recombining with any TRBJ1 gene. Error bars indicate variation across the nine individuals.

In Fig. S4, we show the inferred gene usage frequencies. As described in the main text, Fig. S4D reveals the mechanistic constraint prohibiting the recombination of the TRBD2 gene with any upstream TRBJ1 gene. We include pseudo V-genes in our analysis. These pseudogenes cannot produce a functional receptor but they can participate in the recombination process and produce a non-productive rearranged CDR3 sequence which can be transmitted into the naïve or memory compartments just like any other non-productive rearrangement. The set of V gene sequencing primers used by Robins et. al. [2, 3] either exactly or approximately match 11 pseudogenes. Of these, TRBV23-1, TRBV5-3, TRBV12-2 and TRBV6-7 show significant usage, together accounting for almost 10% of CDR3 sequence reads.

6 Memory T-cell non-productive repertoire

We performed the same analysis on both the naive and memory T-cell repertoires. The non-productive CDR3 sequences in both of these compartments should not be subject to selection, and a comparison of inferences from the two provides a test of this important assumption. Results from the larger naïve non-productive compartment (containing an average of 35,000 unique sequences per individual) were reported in the main text. Here we report the results from the smaller memory non-productive compartment (containing an average of 22,000

unique sequences per individual). In Fig. S5, we compare the naive and memory insertions and deletions distributions. In Fig. S6 we show that the occurence of shared sequences between the individual non-productive repertoires is consistent with our generative model for the memory compartments as well. The plots show that the models inferred from the naïve and memory T-cells are identical in all respects, in confirmation of the expectation that non-productive sequences are not subject to selection effects.

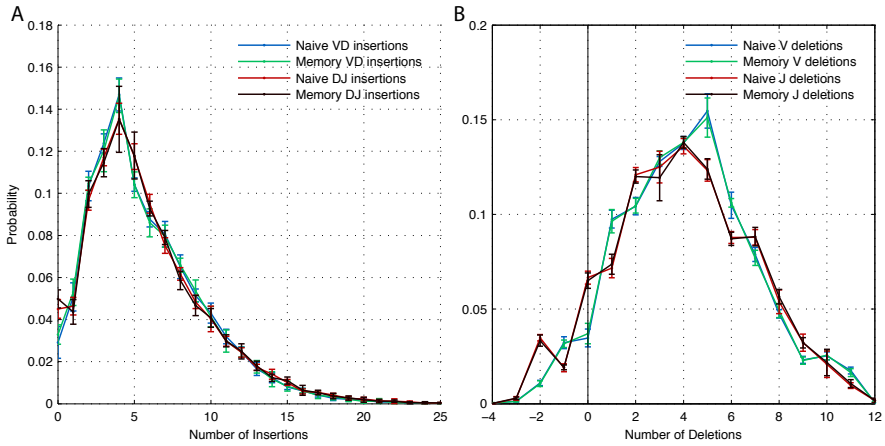

Fig. S5: Comparison of insertions (A) and deletions (B) distributions for the naive and memory T-cell repertoires. We find that the inferred models from the two compartments are statistically identical in all respects. Error bars indicate variation across the nine individuals.

7 Spurious shared sequences between repertoires

Of the 9 individuals, we find three specific pairs of individuals – (2,3), (2,7) and (5,6) – who have an unusually large number of sequences in common, in both the naive and memory compartments. While all other pairs of individuals have between 0 and 4 sequences in common, these three pairs have 15 to 90 shared sequences. Additionally, many of these shared sequences occur in both the naive and memory compartments of the individuals. We suspect that these anomalies are the result of inter-sample contamination.

Hence, for our analysis of the distribution of shared sequences between individuals, we discard from consideration the four pairs of individuals (2,3), (2,7), (3,7) and (5,6). This leaves 32 pairs of individuals for our analysis. We also discard three specific additional sequences that occur in the naive and memory compartments of one individual and also in another individual.

8 Convergent recombination and generation probability

As discussed in the main text, a typical CDR3 sequence can be produced by ≈ 32 different recombination events, corresponding to an entropy of 5 bits per CDR3 sequence. In Fig. S7, we

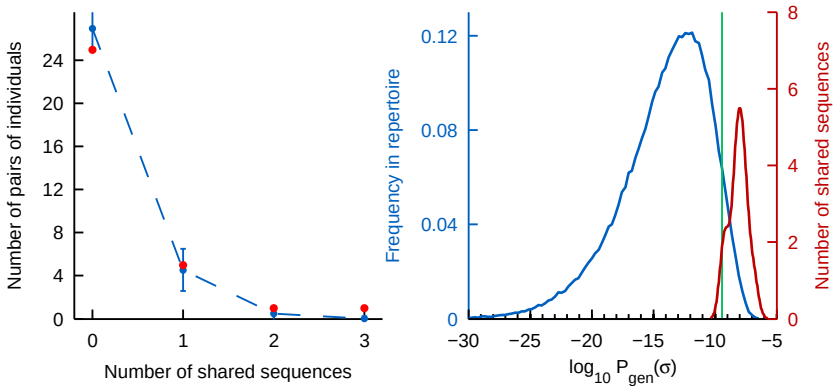

Fig. S6: Shared sequences in memory T-cell non-productive CDR3 sequence repertoires. A) Distribution of number of shared sequences between the 9 individuals. B) Distribution of $P_{\text{gen}}(\sigma)$ for the entire repertoire (blue) and for the recurring sequences (red). $\langle P_{\text{gen}} \rangle$ is indicated by the green vertical line.

show the 2D histogram of the recombination entropy $S(E|\sigma)$ and the generation probability $P_{\text{gen}}(\sigma)$. As expected, sequences with higher recombination entropy tend to have higher total generation probability, with a correlation of 0.13. Note also that while the shared sequences between individuals (red dots) all have high $P_{\text{gen}}(\sigma)$, they are widely distributed with respect to the recombination entropy, since only $P_{\text{gen}}(\sigma)$ determines the recurrence probability of a sequence.

9 Generation probabilities of productive sequences

The probability distribution of recombination events that we infer enables us to calculate the generation probability of any given TCRβ CDR3 sequence. We calculate $P_{\text{gen}}(\sigma)$ for all the sequences in the naive and memory productive repertoires. The distributions of these generation probabilities are shown in Fig. S8. The productive repertoires have systematically higher generation probabilities, implying that sequences that are more likely to be generated are also more likely to pass selection filters and survive in the blood. This is, in part, due to systematically fewer insertions in the productive repertoires, which have exponentially higher generation probabilities.

10 The nonproductive sequence constraint does not bias recombination event statistics

As noted in the main text, we infer the probability distribution of generative events from nonproductive sequences only. One might worry that using such a non-random subset of all the sequences produced by VDJ recombination could introduce an uncontrolled bias into the inference. To look at this in more detail, we note that the condition for a rearranged CDR3 sequence to be out of frame involves the sum of six variables that our analysis has shown to

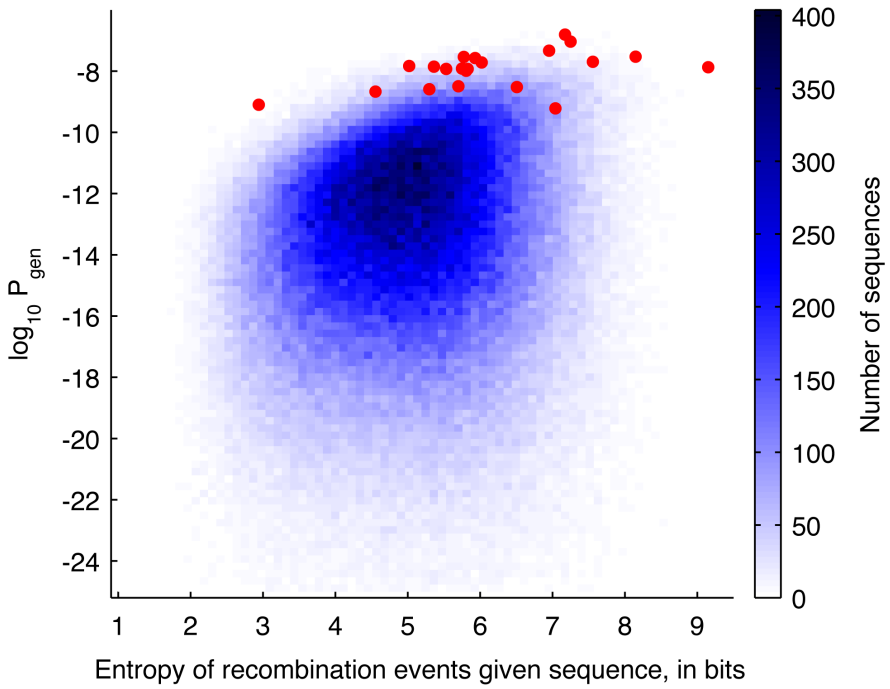

Fig. S7: A 2D histogram of conditional entropy of recombination events given the sequence and $P_{\text{gen}}(\sigma)$. Convergent recombination (as measureed by the recombination event entropy) is a contributing factor to $P_{\text{gen}}(\sigma)$, with correlation coefficient 0.13. The shared sequences in the naive non-productive repertoires are shown in red.

be uncorrelated:

$$[-\text{del}V + \text{ins}VD - \text{del}5'D + \text{length}(D) - \text{del}3'D + \text{ins}DJ - \text{del}J] \bmod 3 > 0.$$

Since a large number of uncorrelated variables are involved, it is a priori unlikely that this constraint would significantly affect the evaluation of the pairwise correlations that define our generative model. We can test this quantitatively by generating a simulated sequence repertoire from our recombination event distribution, running our inference algorithm on the out-of-frame subset of these sequences, and then comparing the inferred and the "actual" event distributions. The result of carrying out this program on a simulated repertoire of 10^5 sequences (two-thirds of which were out-of-frame) is displayed in Fig. S9. It is clear that the initial and the inferred generative distributions are identical to each other, confirming that the condition of being out-of-frame does not bias the statistics of recombination events and does not interfere with our ability to correctly infer the probability distribution of these events. We thank W. Bialek for suggesting this test of our analysis method.

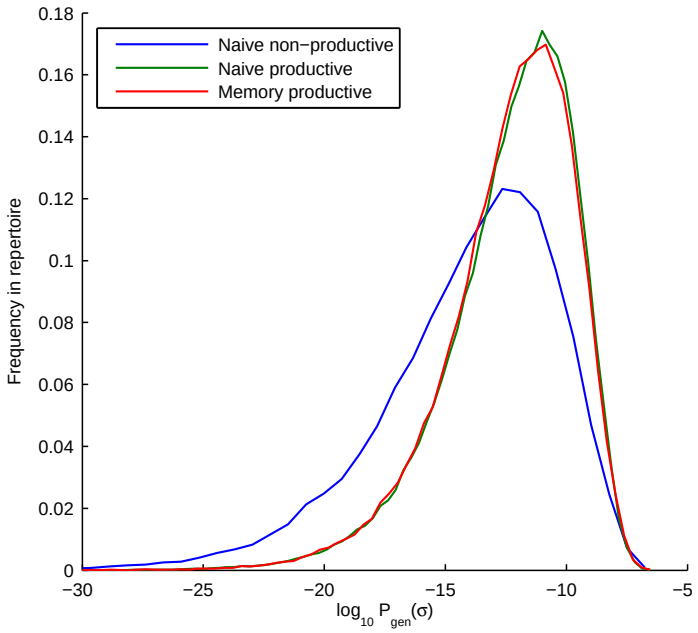

Fig. S8: Generation probabilities of all the CDR3 sequences in the naive and memory productive repertoires were computed using our inferred distribution. The above panel shows the distribution of the logarithm of these probabilities for the three repertoires for one individual. The productive repertoires have systematically higher generation probabilities.

11 Occurrence of palindromic nucleotides with non-zero deletions

To show that the occurrence of palindromic nucleotides with non-zero nucleotide deletions from the ends of the genes is consistent with chance insertions, we keep track of the (model probability weighted) joint frequencies of lengths of observed palindromes conditioned on the number of deletions and on gene choice. Keeping track of this detail is necessary because of the strong dependence of deletion probabilities on gene choice. After we obtain our converged model, we calculate the frequencies of chance palindromic nucleotides of different lengths co-occurring with non-zero deletions (taking into account all the structure of $P_{\text{recomb}}(E)$, including the nucleotide bias in insertions). The plot in Fig. S10 shows that the observed frequencies of palindromic nucleotides co-occurring with non-zero deletions are completely consistent with those expected by chance insertions.

12 Sequence dependence of nucleotide deletion probabilities

Since the sequence at the $3'$ end of the V gene varies between genes, we fit a simple model to the gene dependent deletions profiles to explain the variation in these distributions. The precise mechanism of the generation of P-nucleotides and their relationship to deletions is unclear. Hence, we take only the probabilities of deletions greater than or equal to two

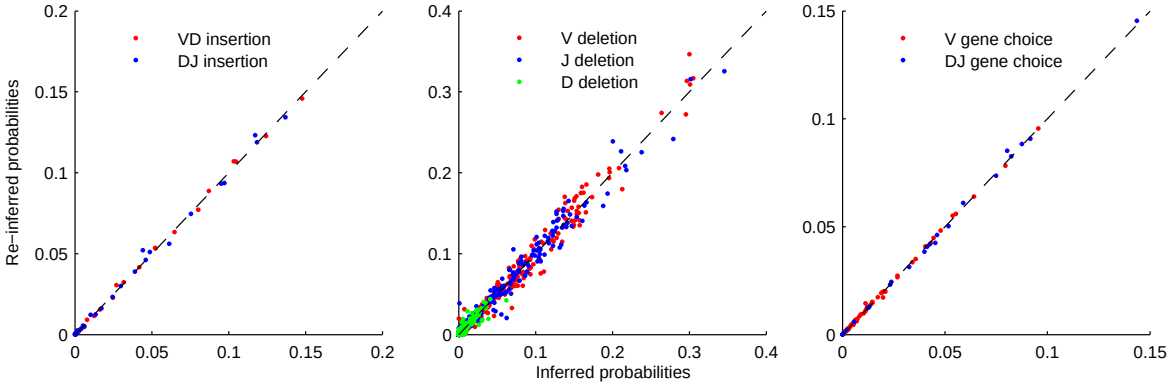

Fig. S9: Probabilities of recombination event variables were re-inferred by simulating sequences from our final distributions, discarding all in-frame sequences, and running the expectation-maximization algorithm on the out-of-frame subset. The above scatter plots show that the original probabilities are obtained. This provides evidence that the use of just the non-productive TCR sequences does not bias the statistics of recombination events.

nucleotides and consider the nucleotide sequence context (four bases $3'$ and two bases $5'$ of the deletion position) as a predictor of the deletion probability. We use a function of the form

$$P(n\,\text{deletions}|\sigma\,\&\,n \geq 2) = \frac{\exp\left(\sum_{k=1}^{6}\epsilon(k,\sigma(n-4+k)\right)}{Z(\sigma)} \tag{2}$$

$$Z(\sigma) = \sum_{n=2}^{12}\exp\left(\sum_{k=1}^{6}\epsilon(k,\sigma(n-4+k)\right) \tag{3}$$

where ϵ is a 6×4 matrix containing the contribution of each possible nucleotide at each of the positions, analogous to a (log) Position Weight Matrix (PWM). We do a least squares fit to determine the elements of ϵ. In Fig. S11, we show ϵ fit to the V deletions. There is a strong preference for T and A, especially in the 2 nucleotides just $5'$ of the position of deletion. Since there are only 13 J-genes, there is less sequence variation among them that we can utilize.

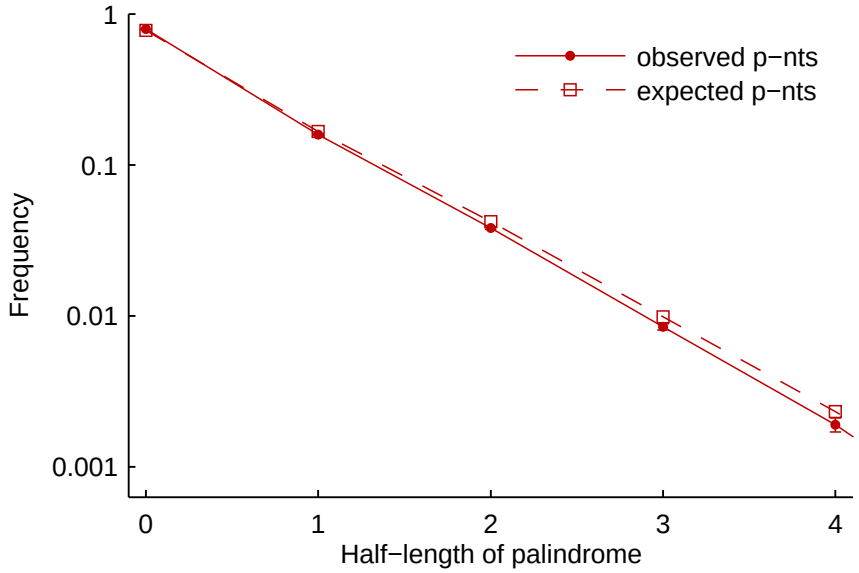

Fig. S10: Occurrence frequency of P-nucleotides for non-zero deletions.

References

[1] 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.

[2] Robins HS, et al. (2010) Overlap and effective size of the human CD8+ T cell receptor repertoire. *Science translational medicine* 2:47ra64.

[3] Robins HS, et al. (2009) Comprehensive assessment of T-cell receptor beta-chain diversity in alphabeta T cells. *Blood* 114:4099–4107.

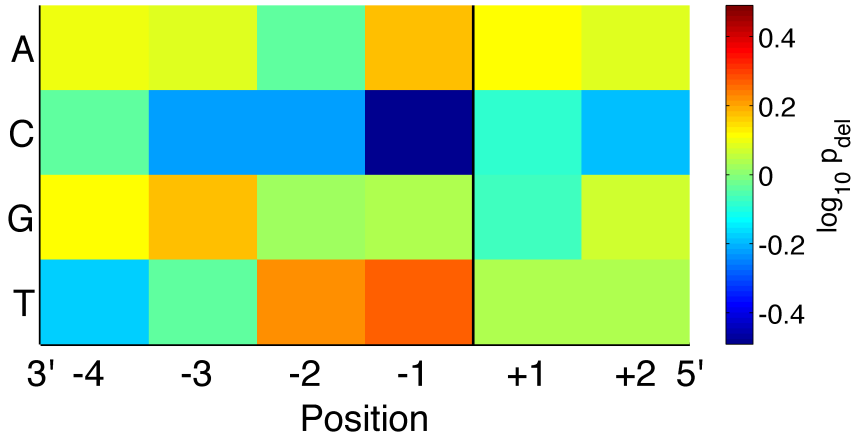

Fig. S11: Position weight matrix for sequence dependence of nucleotide deletion position. The figure shows $\epsilon/\log(10)$ (see SI Appendix section 12 for details) fit to the V gene specific deletions profiles, using four nucleotides $3'$ and two nucleotides $5'$ of the deletion position (black vertical line). The $3'$ nucleotides are the most informative about deletion probability and show a preference for T and A. The sequence logo corresponding to this position weight matrix is shown in the main text Fig. 4B.

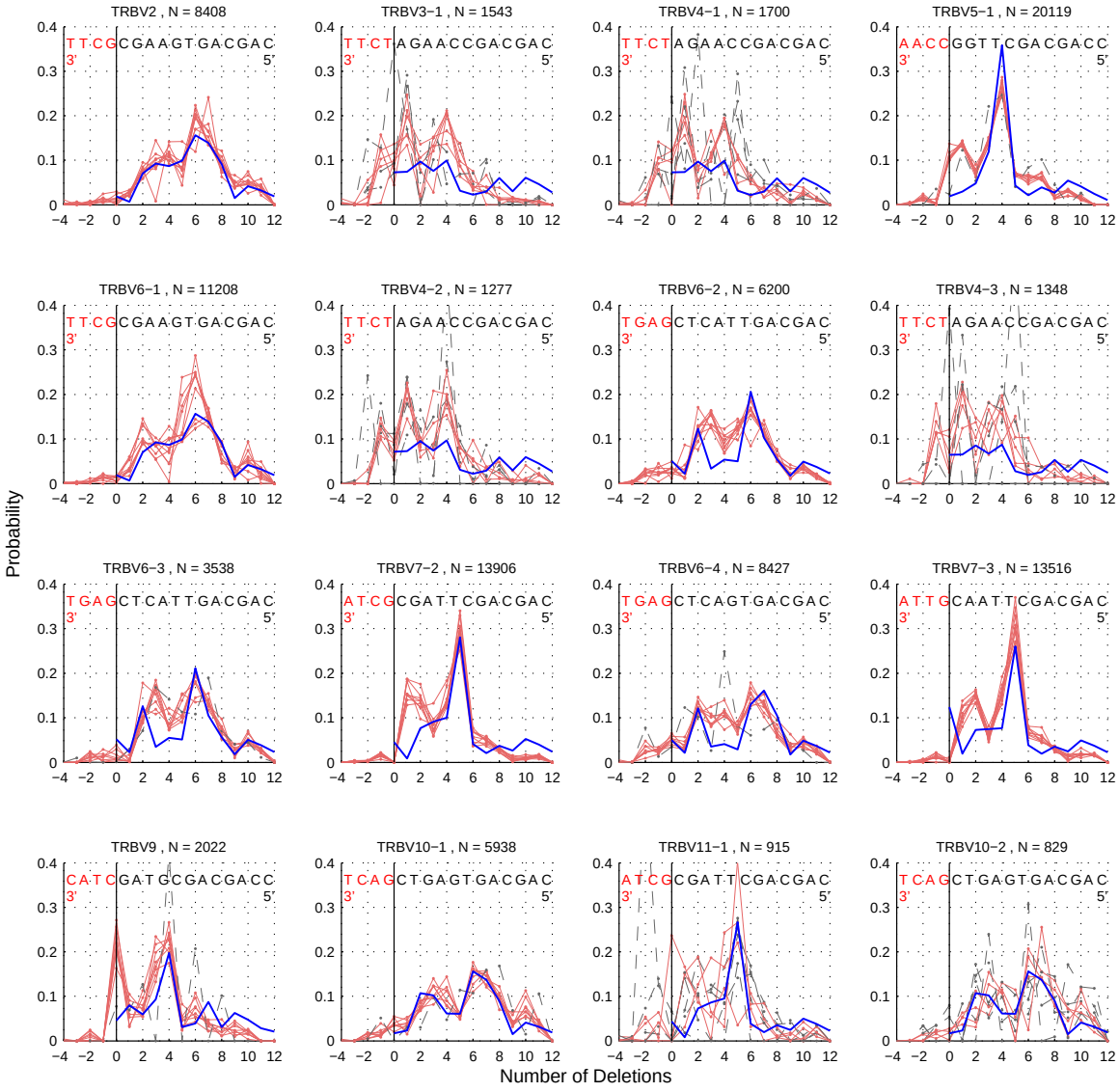

Fig. S12: Deletion profiles for all the V-genes (1 of 3). The title for each panel lists the gene name and total number of counts, across all the individuals studied, of the particular gene in question. Individuals with fewer than 100 counts for a specific gene are plotted in gray dashed lines. The blue lines show the predictions of the position weight matrix based model fit to these curves.

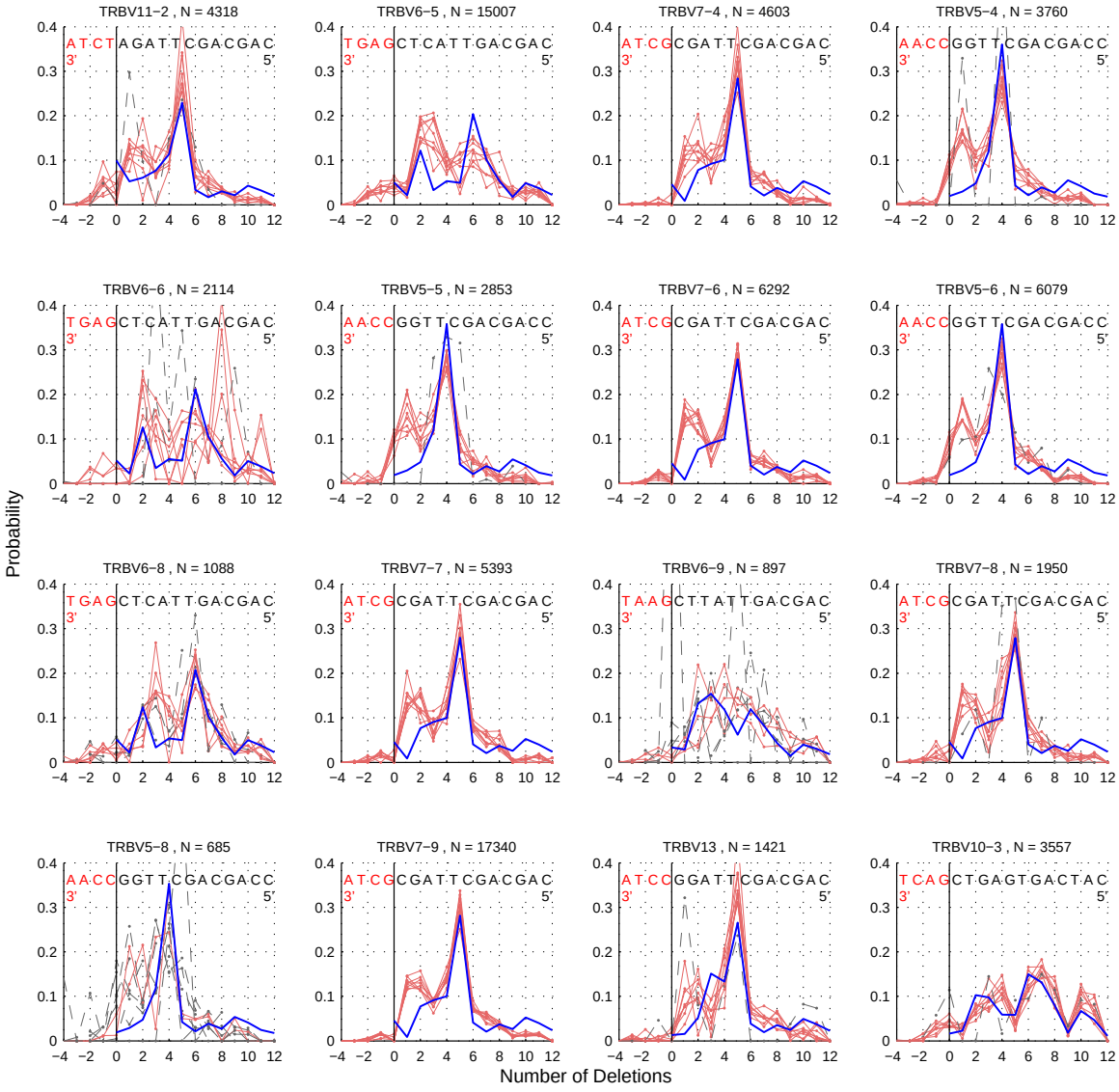

Fig. S13: Deletion profiles for all the V-genes (2 of 3). The title for each panel lists the gene name and total number of counts, across all the individuals studied, of the particular gene in question. Individuals with fewer than 100 counts for a specific gene are plotted in gray dashed lines. The blue lines show the predictions of the position weight matrix based model fit to these curves.

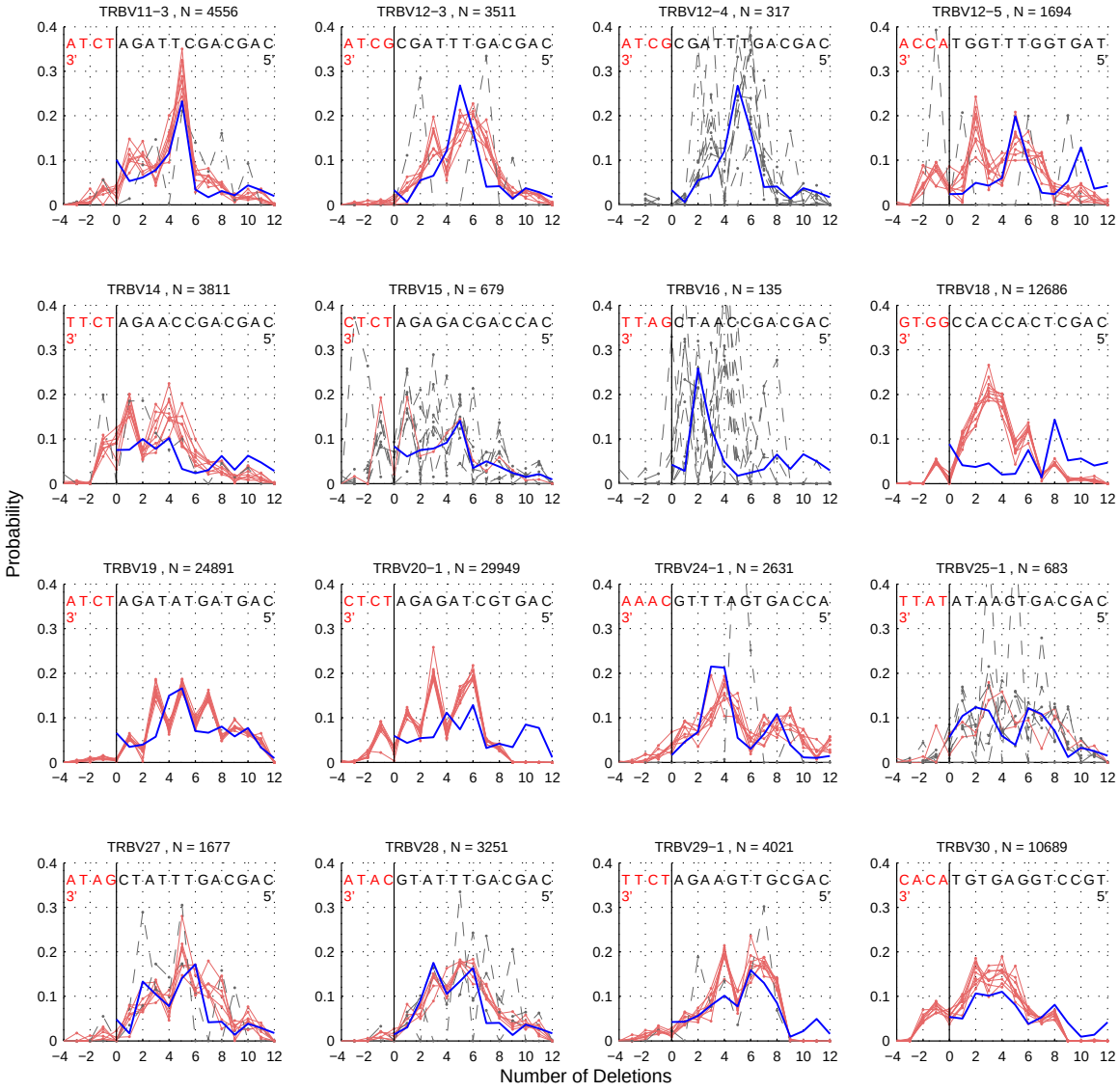

Fig. S14: Deletion profiles for all the V-genes (3 of 3). The title for each panel lists the gene name and total number of counts, across all the individuals studied, of the particular gene in question. Individuals with fewer than 100 counts for a specific gene are plotted in gray dashed lines. The blue lines show the predictions of the position weight matrix based model fit to these curves.

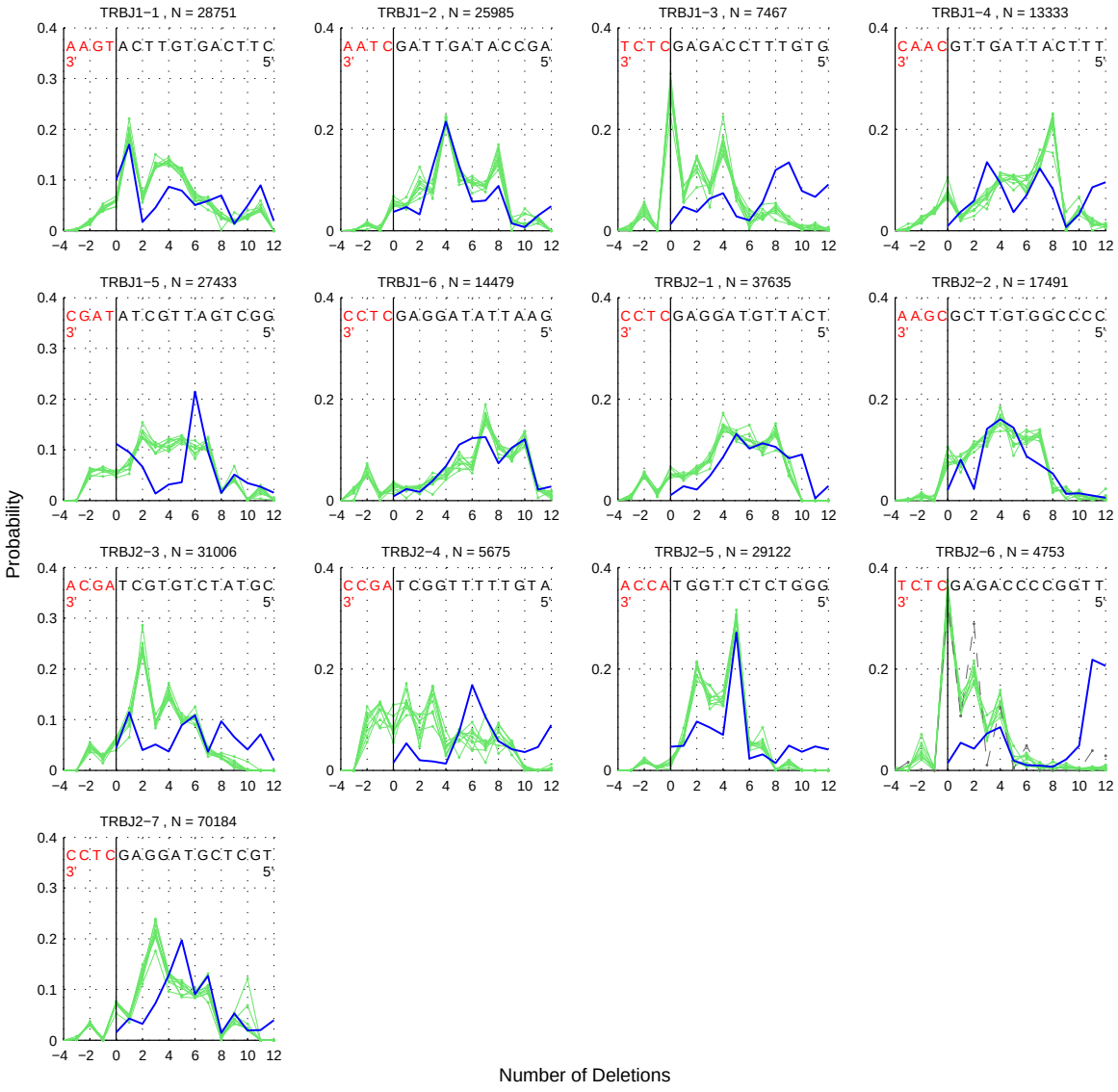

Fig. S15: Deletion profiles for all the J-genes. The title for each panel lists the gene name and total number of counts, across all the individuals studied, of the particular gene in question. Individuals with fewer than 100 counts for a specific gene are plotted in gray dashed lines. The blue lines show the predictions of the position weight matrix based model fit to the V deletions curves, but evaluated on the J gene sequences.

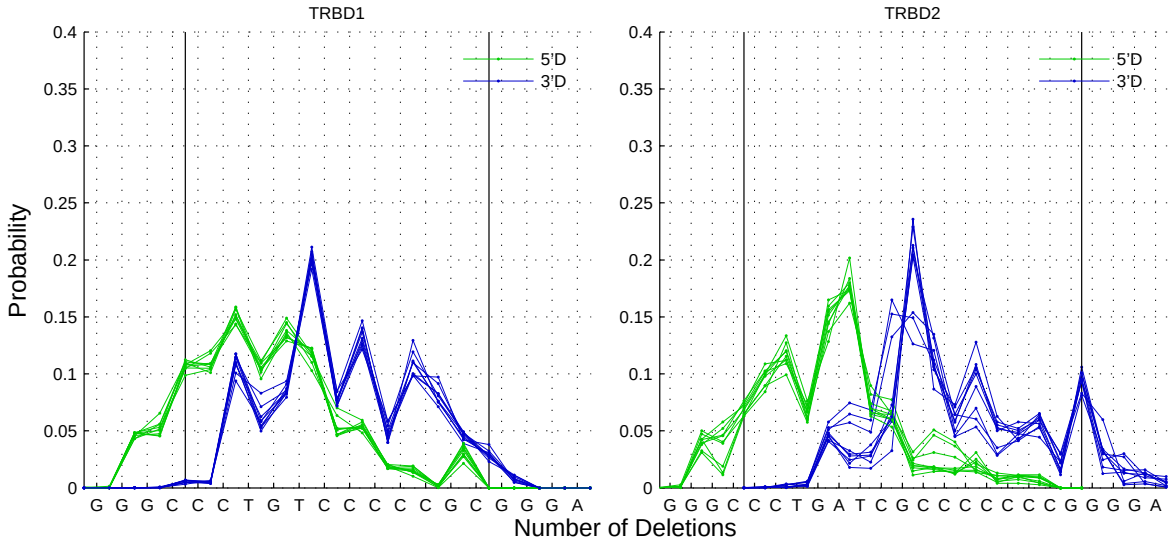

Fig. S16: Marginal deletion probability distributions for the two D-genes. Deletions at the $5'$ end ($3'$ end) of the D gene are shown in green (blue). The x-axis displays the gene sequence from the $5'$ end to the $3'$ end.