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

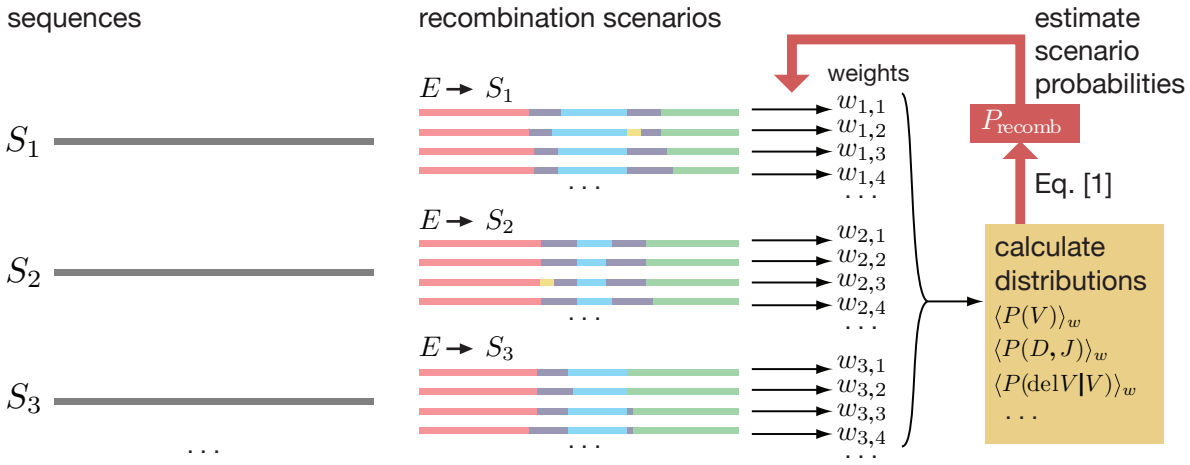

Fig. S1: Flow chart of the analysis pipeline.

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.

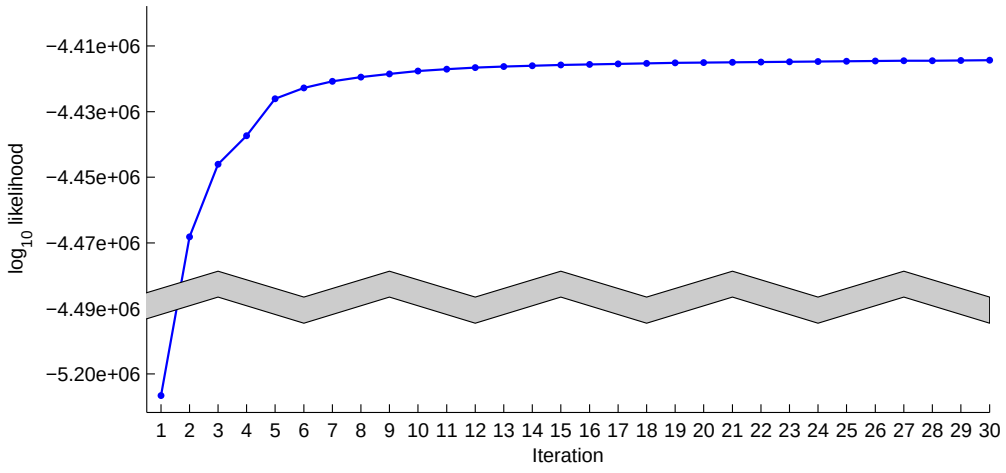

Fig. S 2: Convergence of the total likelihood of all data sets with iterations of the EM algorithm.

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.

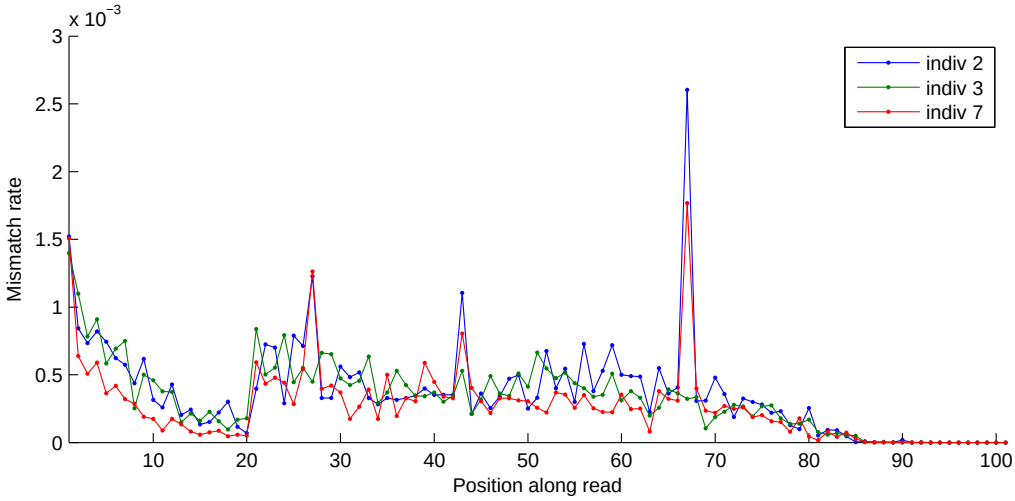

Fig. S3: Position-dependent error profile for the three individuals with read length 101 base pairs. The sequencing read proceeds from the right (101 to 1) where the J gene sequencing primer binds. The spikes in the error rate at specific positions (67, 43 and 27) are true sequencing error spikes and 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

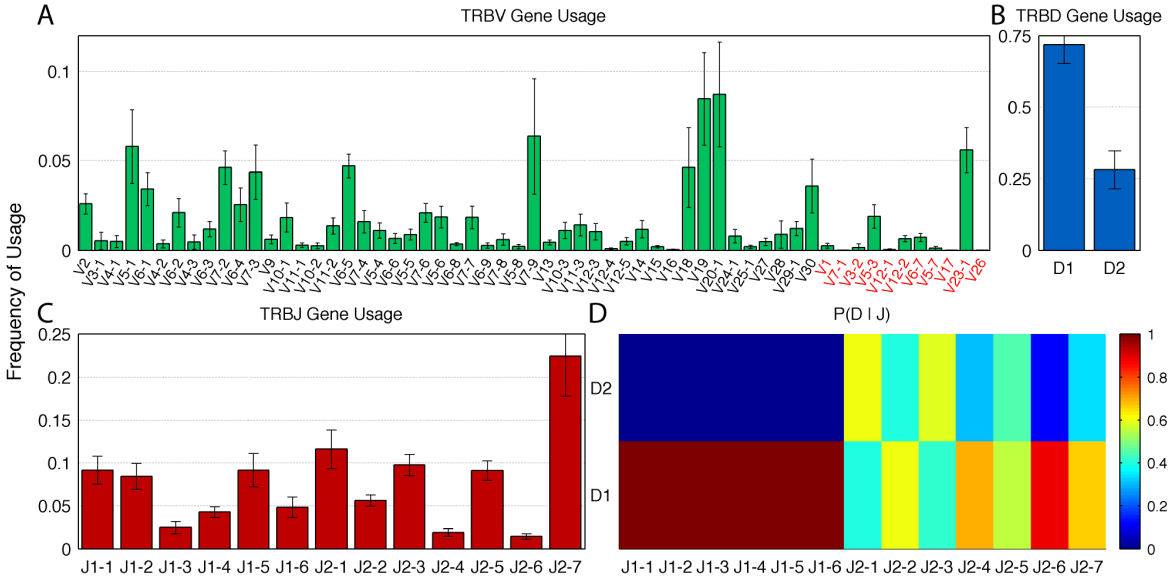

Fig. S4: Statistical aspects of gene usage. (A) Usage frequencies of V-genes, ordered by position on the chromosome, with the exception of pseudogenes (red legend). (B) Usage frequencies of the two D-genes. (C) Same for the 13 J-genes. (D) D-gene usage frequencies, conditioned on J-gene choice. As expected 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.

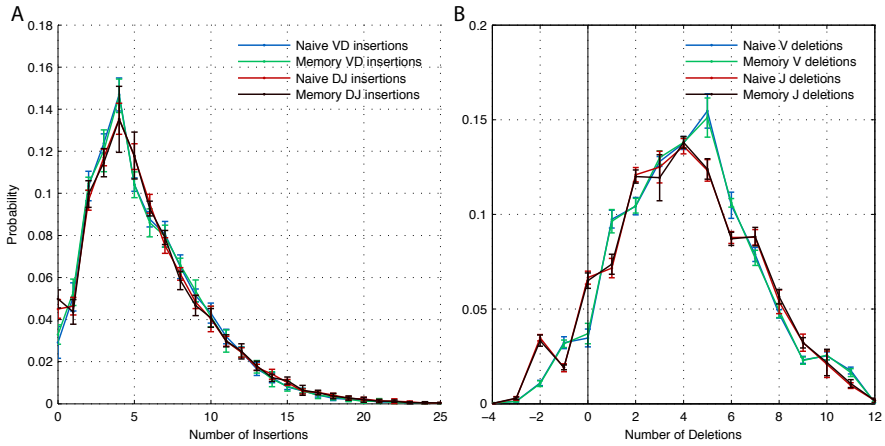

Fig. S5: Comparison of insertions (A) and deletions (B) distributions for the naive and memory T-cell repertoires. We find that the inferred models from the two compartments are statistically identical in all respects. Error bars indicate variation across the nine individuals.

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

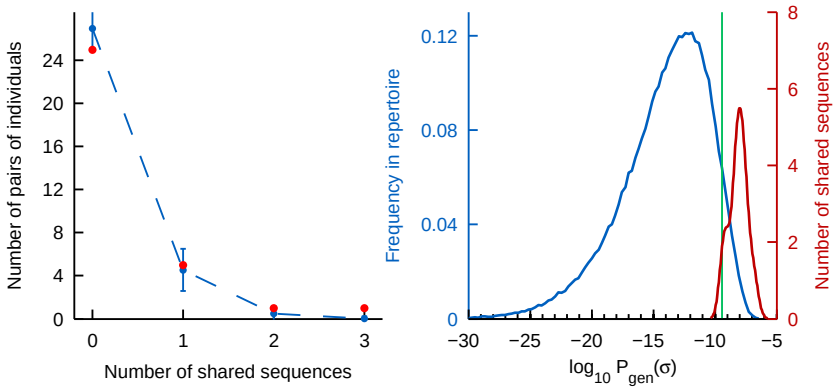

Fig. S6: Shared sequences in memory T-cell non-productive CDR3 sequence repertoires. A) Distribution of number of shared sequences between the 9 individuals. B) Distribution of $P_{\text{gen}}(\sigma)$ for the entire repertoire (blue) and for the recurring sequences (red). $\langle P_{\text{gen}} \rangle$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

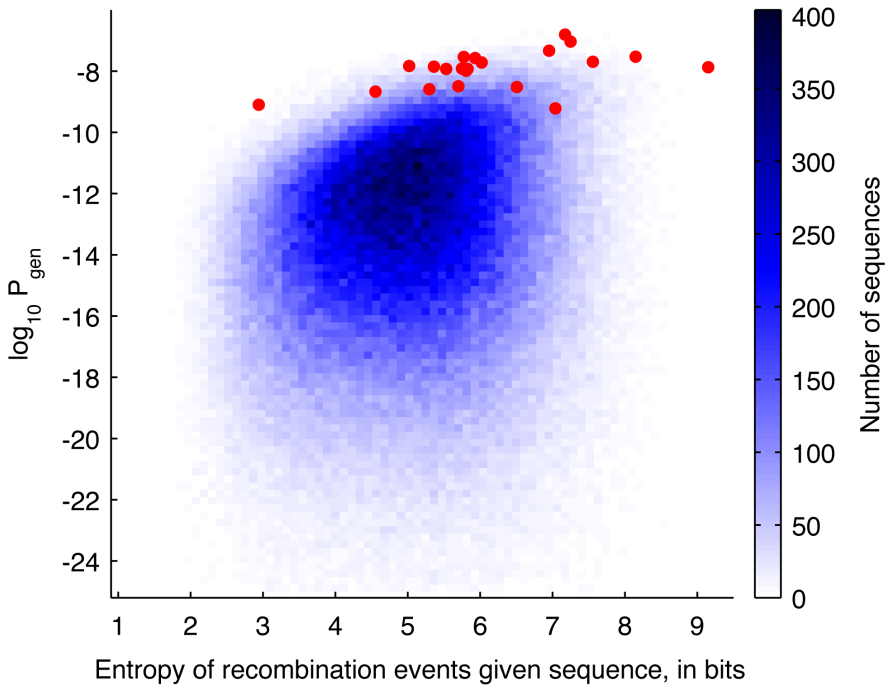

Fig. S7: A 2D histogram of conditional entropy of recombination events given the sequence and $P_{\text{gen}}(\sigma)$. Convergent recombination (as measureed by the recombination event entropy) is a contributing factor to $P_{\text{gen}}(\sigma)$, with correlation coefficient 0.13. The shared sequences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

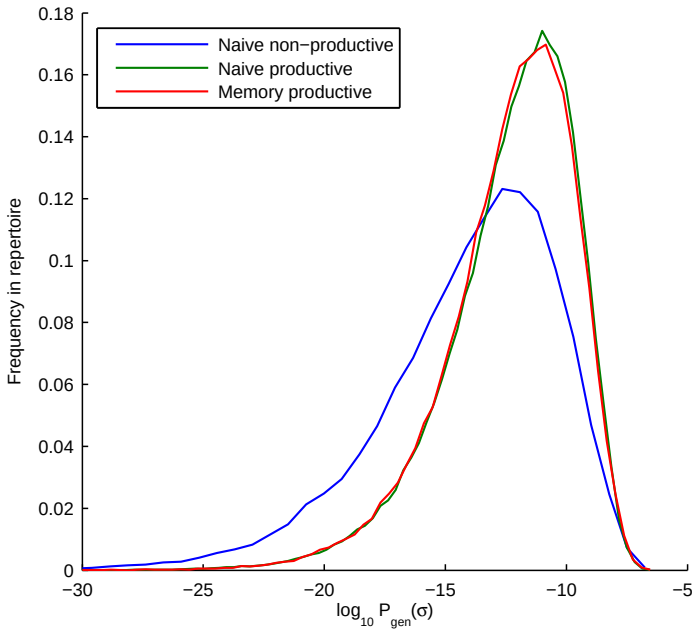

Fig. S8: Generation probabilities of all the CDR3 sequences in the naive and memory productive repertoires were computed using our inferred distribution. The above panel shows the distribution of the logarithm of these probabilities for the three repertoires for one individual. The productive repertoires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

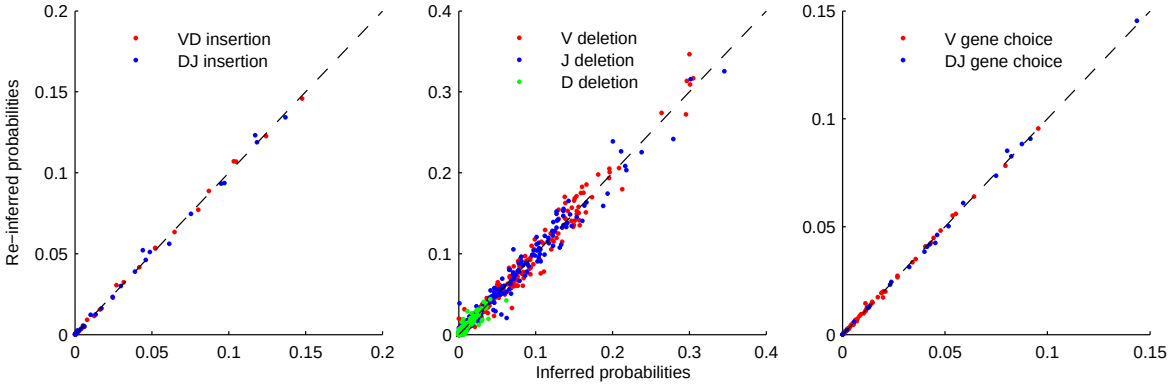

Fig. S9: Probabilities of recombination event variables were re-inferred by simulating sequences from our final distributions, discarding all in-frame sequences, and running the expectation-maximization algorithm on the out-of-frame subset. The above scatter plots show that the original probabilities 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.

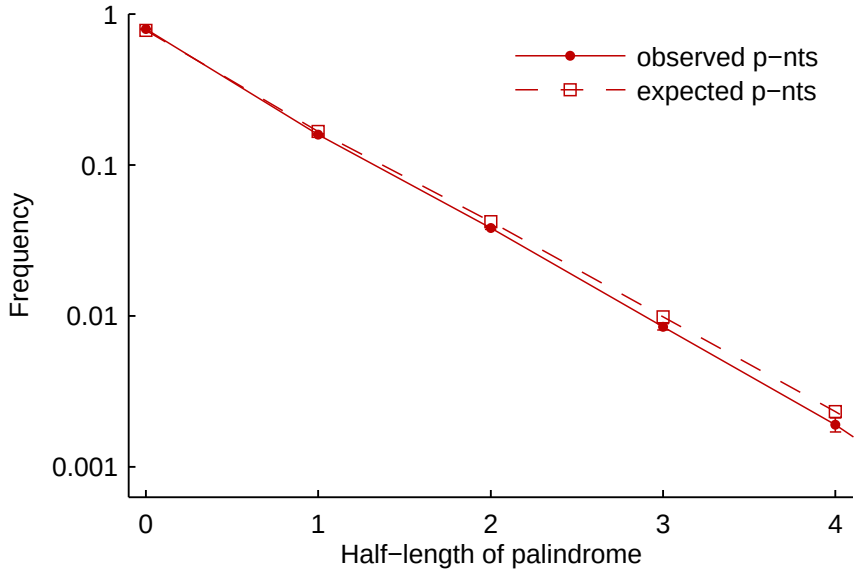

Fig. S10: Occurrence frequency of P-nucleotides for non-zero deletions.

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.

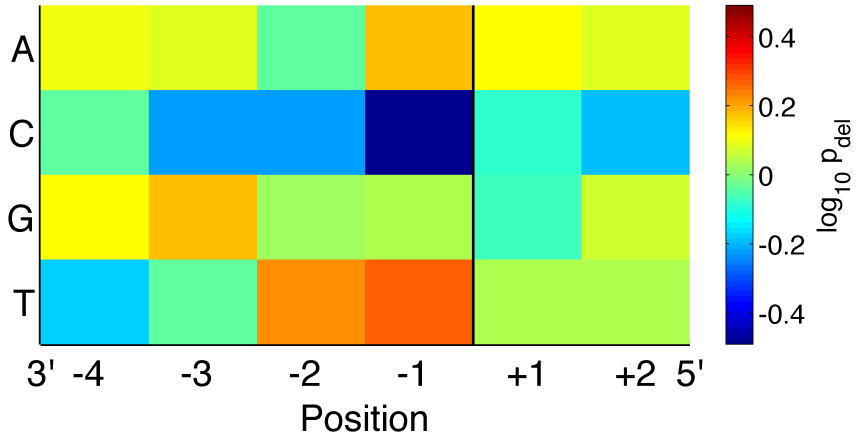

Fig. S11: Position weight matrix for sequence dependence of nucleotide deletion position. The figure shows $\epsilon/\log(10)$ (see SI Appendix section 12 for details) fit to the V gene specific deletions profiles, using four nucleotides $3'$ and two nucleotides $5'$ of the deletion position (black vertical 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.

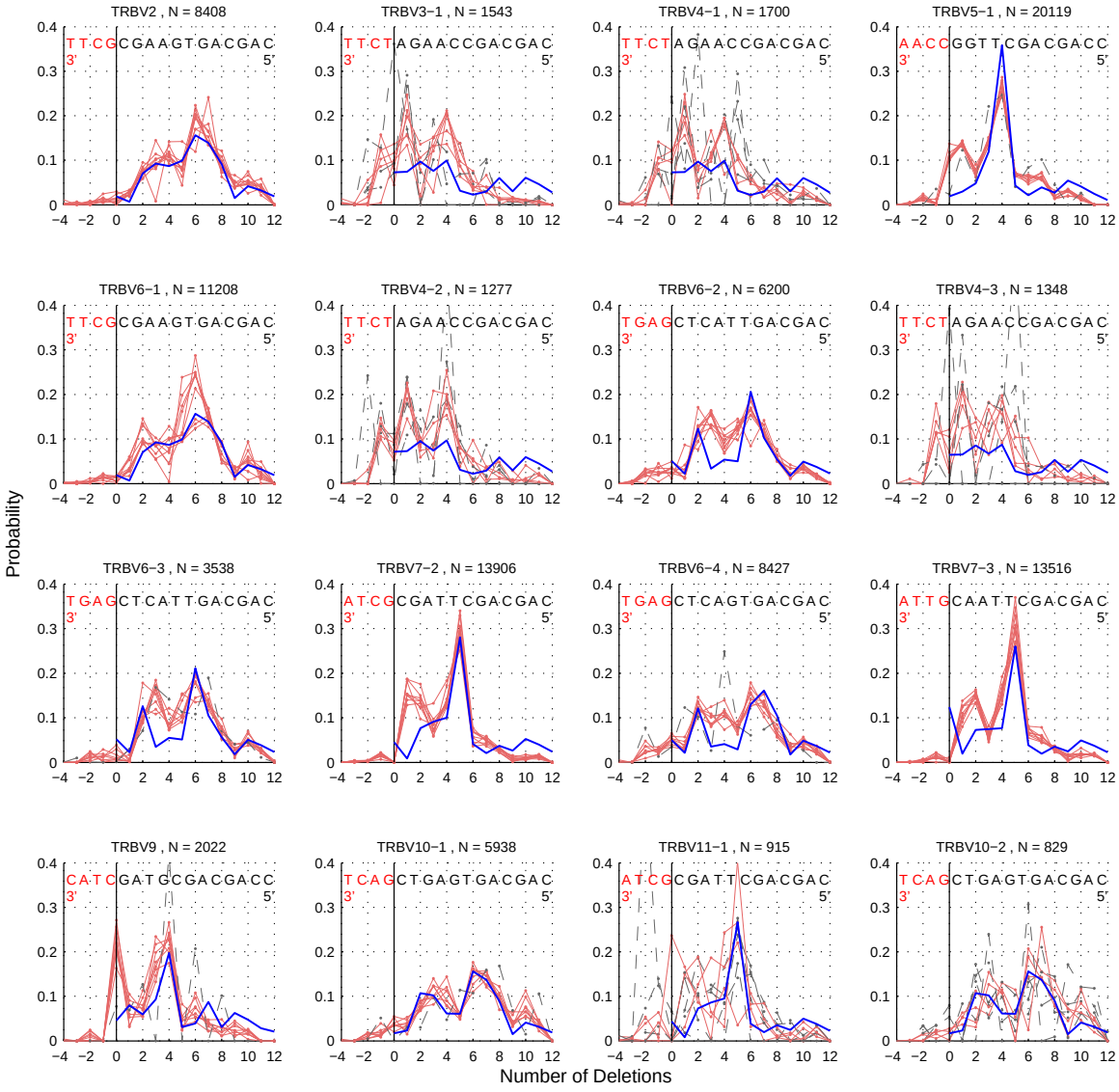

Fig. S12: Deletion profiles for all the V-genes (1 of 3). The title for each panel lists the gene name and total number of counts, across all the individuals studied, of the particular gene in question. Individuals with fewer than 100 counts for a specific gene are plotted in gray dashed lines. The blue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

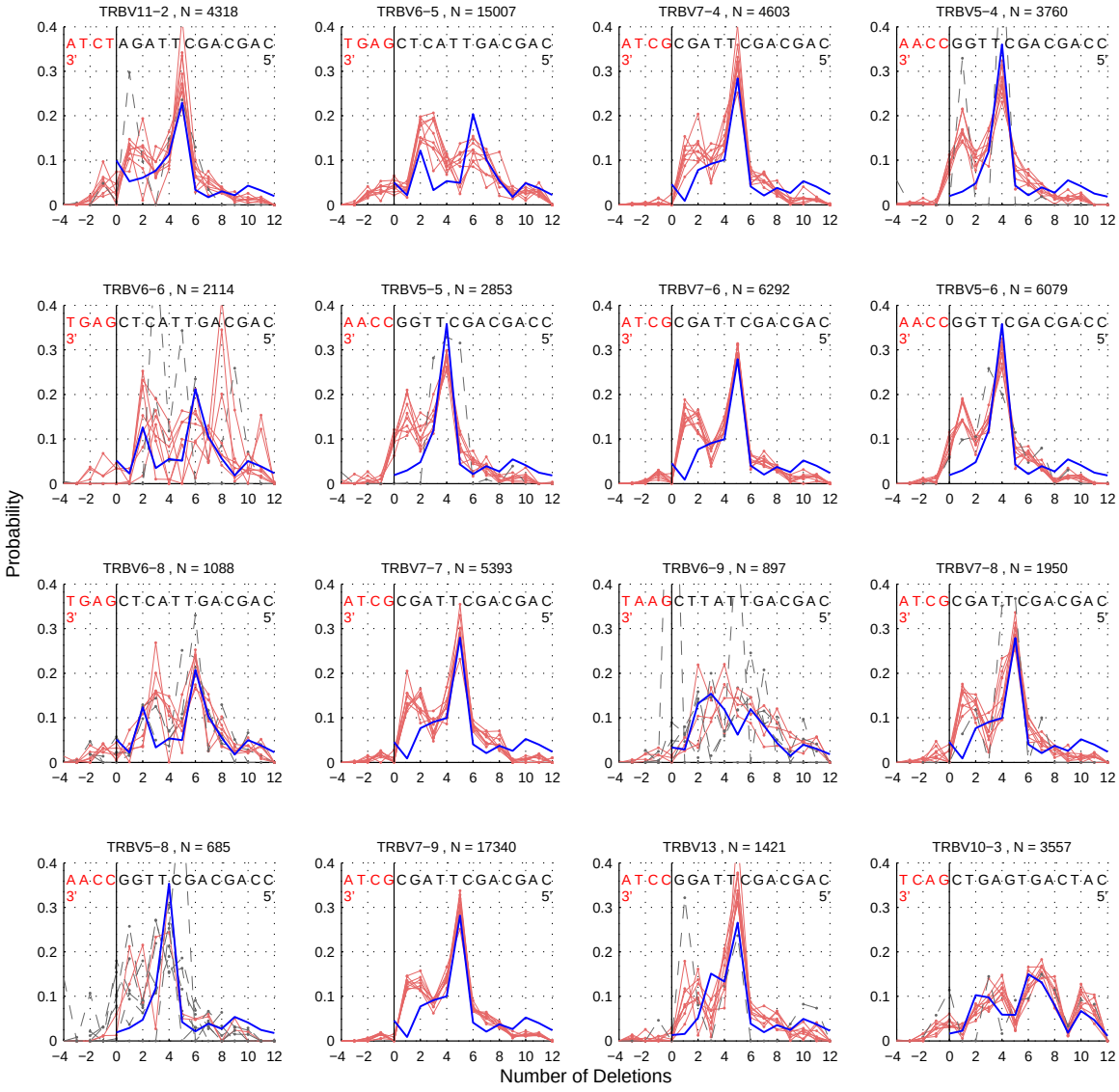

Fig. S13: Deletion profiles for all the V-genes (2 of 3). The title for each panel lists the gene name and total number of counts, across all the individuals studied, of the particular gene in question. Individuals with fewer than 100 counts for a specific gene are plotted in gray dashed lines. The blue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

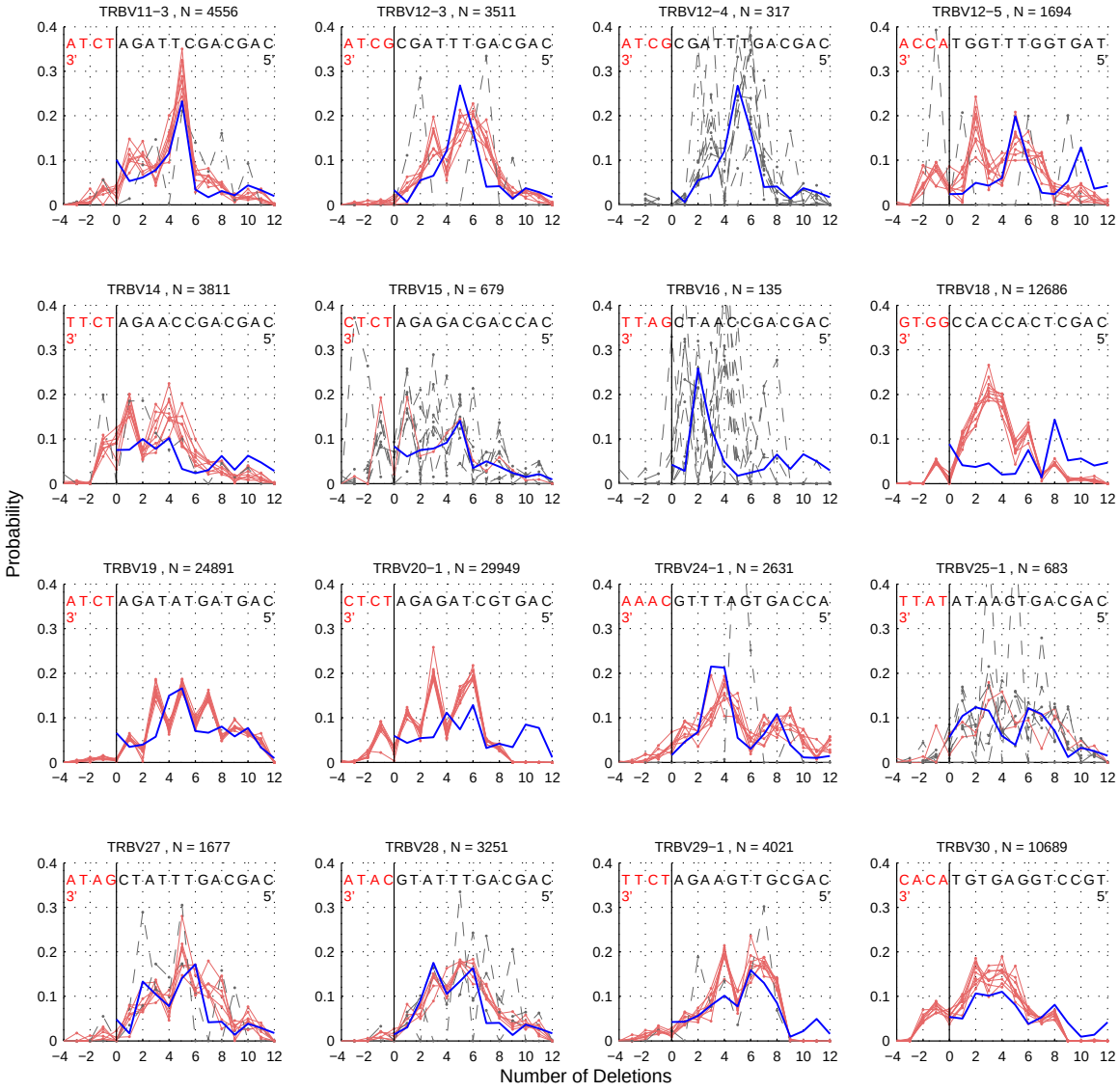

Fig. S14: Deletion profiles for all the V-genes (3 of 3). The title for each panel lists the gene name and total number of counts, across all the individuals studied, of the particular gene in question. Individuals with fewer than 100 counts for a specific gene are plotted in gray dashed lines. The blue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

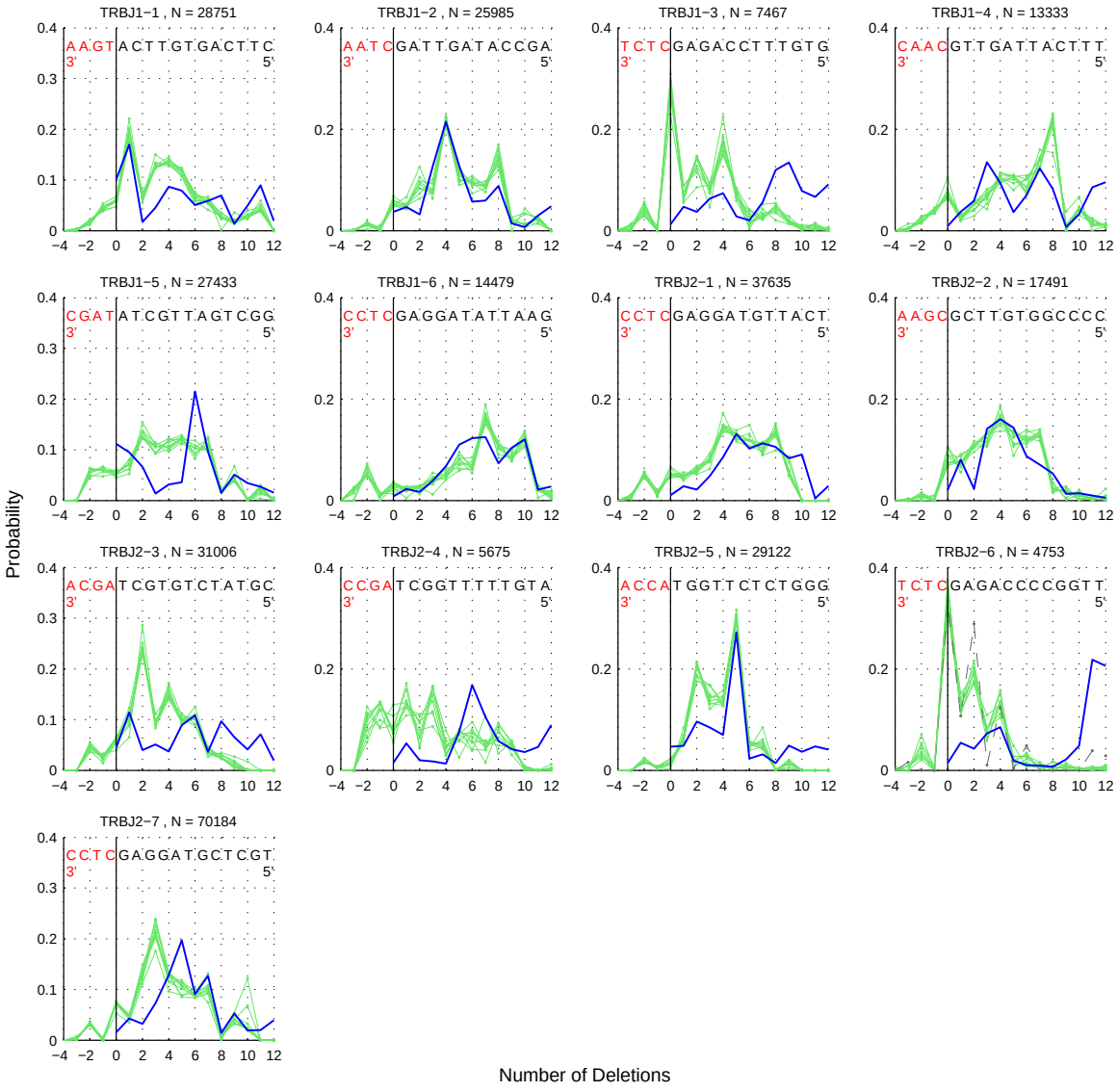

Fig. S15: Deletion profiles for all the J-genes. The title for each panel lists the gene name and total number of counts, across all the individuals studied, of the particular gene in question. Individuals with fewer than 100 counts for a specific gene are plotted in gray dashed lines. The blue lines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

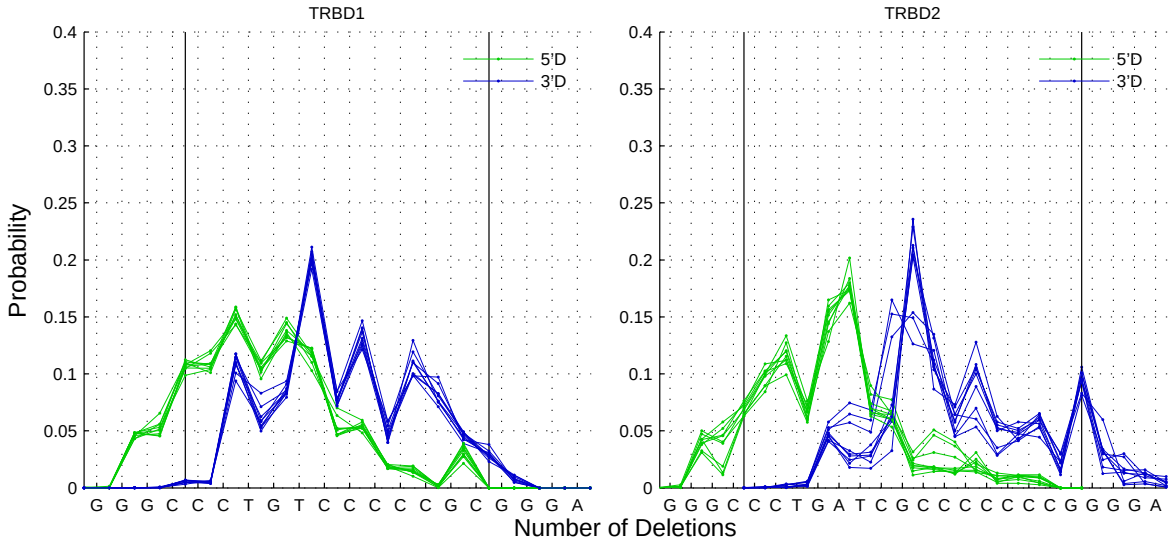

Fig. S16: Marginal deletion probability distributions for the two D-genes. Deletions at the $5'$ end ($3'$ end) of the D gene are shown in green (blue). The x-axis displays the gene sequence from the $5'$ end to the $3'$ end.