

Quantifying selection in immune receptor repertoires

Yuval Elhanati^a, Anand Murugan^b, Curtis G. Callan, Jr.^{c,1}, Thierry Mora^d, and Aleksandra M. Walczak^a

^aLaboratoire de Physique Théorique, Unité Mixte de Recherche 8549 and ^dLaboratoire de Physique Statistique, Unité Mixte de Recherche 8550, Centre National de la Recherche Scientifique and École Normale Supérieure, 75005 Paris, France; ^bDepartment of Applied Physics, Stanford University, Stanford, CA 94305; and ^cJoseph Henry Laboratories, Princeton University, Princeton, NJ 08544

Contributed by Curtis G. Callan, Jr., May 22, 2014 (sent for review April 15, 2014)

The efficient recognition of pathogens by the adaptive immune system relies on the diversity of receptors displayed at the surface of immune cells. T-cell receptor diversity results from an initial random DNA editing process, called VDJ recombination, followed by functional selection of cells according to the interaction of their surface receptors with self and foreign antigenic peptides. Using high-throughput sequence data from the β -chain of human T-cell receptors, we infer factors that quantify the overall effect of selection on the elements of receptor sequence composition: the V and J gene choice and the length and amino acid composition of the variable region. We find a significant correlation between biases induced by VDJ recombination and our inferred selection factors together with a reduction of diversity during selection. Both effects suggest that natural selection acting on the recombination process has anticipated the selection pressures experienced during somatic evolution. The inferred selection factors differ little between donors or between naive and memory repertoires. The number of sequences shared between donors is well-predicted by our model, indicating a stochastic origin of such public sequences. Our approach is based on a probabilistic maximum likelihood method, which is necessary to disentangle the effects of selection from biases inherent in the recombination process.

thymic selection | statistical inference | public repertoire | T cell

The T-cell response of the adaptive immune system begins when receptor proteins on the surface of these cells recognize a pathogen peptide displayed by an antigen-presenting cell. The immune cell repertoire of a given individual is comprised of many clones, each with a distinct surface receptor. This diversity, which is central to the ability of the immune system to defeat pathogens, is initially created by a stochastic process of germline DNA editing (called VDJ recombination) that gives each new immune cell a unique surface receptor gene. This initial repertoire is subsequently modified by selective forces, including nonpathogen-related thymic selection against excessive (or insufficient) recognition of self proteins, which are also stochastic in nature. Because of this stochasticity and the large T-cell diversity, these repertoires are best described by probability distributions. In this paper, we apply a probabilistic approach to sequence data to obtain quantitative measures of the overall (not necessarily pathogenic) selection pressures that shape T-cell receptor repertoires.

New receptor genes are formed by randomly choosing alleles from a set of genomic templates for the subregions (V, D, and J) of the complete gene. Insertion and deletion of nucleotides in the junctional regions between the V and D and D and J genes greatly enhance diversity beyond pure VDJ combinatorics (1). The most variable region of the gene is between the last amino acids of the V segment and the beginning of the J segment; it codes for the Complementarity Determining Region 3 (CDR3) loop of the receptor protein, a region known to be functionally important in recognition (2). Previous studies have shown that immune cell receptors are not uniform in terms of VDJ gene segment use (3–6) or probability of generation (1) and that certain receptors are more likely than others to be shared by different individuals (4, 7). The statistical properties of the

immune repertoire are, thus, rather complex, and their accurate determination requires sophisticated methods.

Recent advances in sequencing technology have made it possible to sample the T-cell receptor diversity of individual subjects in great depth (8). The availability of such data has, in turn, led to the development of sequence statistics-based approaches to the study of immune cell diversity (9, 10). In particular, we recently quantitatively characterized the preselection diversity of the human T-cell repertoire by learning the probabilistic rules of VDJ recombination from out-of-frame DNA sequences that cannot be subject to functional selection and whose statistics therefore reflect only the recombination process (1). After generation, T cells undergo a somatic selection process in the thymus (11) and later in the periphery (12). Cells that pass thymic selection enter the peripheral repertoire as naive T cells, and the subset of naive cells that eventually engage in an immune response will survive as a long-lived memory pool. Although we now understand the statistical properties of the initial repertoire of immune receptors (1) and despite some theoretical studies of thymic selection at the molecular level (13, 14), a quantitative understanding of how selection modifies those statistics to produce the naive and memory repertoires is lacking.

In this paper, we build on our understanding of the preselection distribution of T-cell receptors to derive a statistical method for identifying and quantifying selection pressures in the adaptive immune system. We apply this method to naive and memory DNA sequences of human T-cell β -chains obtained from peripheral blood samples of nine healthy individuals. Our goal is to characterize the likelihood that any given sequence, after it is generated, will survive selection for the ensemble of properties needed to pass into the peripheral repertoire(s). Our analysis reveals strong and reproducible signatures of selection on specific amino acids in the CDR3 sequence and on the usage

Significance

The immune system defends against pathogens through a diverse population of T cells that display different antigen recognition surface receptor proteins. Receptor diversity is produced by an initial random gene recombination process followed by selection for a desirable range of peptide binding. Although recombination is well-understood, selection has not been quantitatively characterized. By combining high-throughput sequencing data with modeling, we quantify the selection pressure that shapes functional repertoires. Selection is found to vary little between individuals or between naive and memory repertoires. It reinforces the biases of the recombination process, meaning that sequences more likely to be produced are also more likely to pass selection. The model accounts for public sequences shared between individuals as resulting from pure chance.

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

The authors declare no conflict of interest.

Freely available online through the PNAS open access option.

¹To 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.1409572111/-DCSupplemental.

of V and J genes. Most strikingly, we find significant correlation between the generation probability of a sequence and the probability that it will pass selection. This correlation suggests that natural selection, which acts on very long timescales to shape the generation mechanism itself, may have tuned it to anticipate somatic selection, which acts on single cells throughout the lifetime of an individual. The quantitative features of selection inferred from our model vary very little between donors, indicating that these features are universal. In addition, our measures of selection pressure on the memory and naive repertoires are statistically indistinguishable, consistent with the hypothesis that the memory pool is a random subsample of the naive pool.

Analysis

We analyzed human CD4⁺ T-cell β -chain DNA sequence reads (60- or 101-nucleotide long) centered around the CDR3 region. T cells were obtained from nine individuals and sorted into naive (CD45RO⁻) and memory (CD45RO⁺) subsets, yielding datasets of $\sim 200,000$ unique naive and $\sim 120,000$ unique memory sequences per individual on average. The datasets are the same as those used in ref. 1 and were obtained by previously described methods (15, 16).

In ref. 1, we used the out-of-frame sequences to characterize the receptor generation process. That analysis yielded an accurate model for the probability $P_{\text{pre}}(\bar{\sigma})$ that a VDJ recombination event will produce a β -chain gene consistent with the sequence read $\bar{\sigma}$ (for any $\bar{\sigma}$). In this study, we focus instead on the in-frame sequences free of stop codons, with the goal of quantifying how their probability of occurrence, $P_{\text{post}}(\bar{\sigma})$, differs from the preselection distribution $P_{\text{pre}}(\bar{\sigma})$. (We only consider the presence or absence of a sequence $\bar{\sigma}$ and not the size of its clone.) Here, we distinguish between the read $\bar{\sigma}$ and the entire β -chain sequence, which is characterized uniquely by the V and J gene choices (denoted by V and J) as well as the CDR3 region $\bar{\tau}$; the latter is defined to run from a conserved Cys near the end of the V segment to the last amino acid of the read (we note that the last amino acid in the read is separated from a conserved Phe in the J gene by two variable amino acids). The CDR3 sequence $\bar{\tau}$ can be uniquely read off from each sequence read; by contrast, the V and J may not be uniquely identifiable (because of the relatively short read length). Because V and J may play a role in selection outside the read $\bar{\sigma}$, we must consider selection in terms of the full β -chain $(\bar{\tau}, V, J)$ rather than the incomplete $\bar{\sigma}$.

For each β -chain sequence $(\bar{\tau}, V, J)$, we define a selection factor $Q = P_{\text{post}}/P_{\text{pre}}$ that quantifies whether selection (thymic selection or subsequent selection in the periphery) has enriched or impoverished the frequency of that sequence compared with the preselection ensemble. Because P_{pre} varies over many orders of magnitude, such a relative enhancement factor is the only way to define selection strength. Our goal is to find a model for Q , such that the distribution $P_{\text{post}}(\bar{\tau}, V, J) = Q(\bar{\tau}, V, J)P_{\text{pre}}(\bar{\tau}, V, J)$ gives a good account of an observed set of selected sequences. We cannot directly estimate $P_{\text{post}}(\bar{\tau}, V, J)$ from the data, but as we outline in Fig. 1A, we can use a reduced complexity model for Q to infer it (and therefore P_{post}) from the data. Specifically, we will show that the following factorized model for Q captures the main features of selection:

$$Q(\bar{\tau}, V, J) = \frac{P_{\text{post}}(\bar{\tau}, V, J)}{P_{\text{pre}}(\bar{\tau}, V, J)} = \frac{1}{Z} q_L q_{VJ} \prod_{i=1}^L q_{i:L}(a_i), \quad [1]$$

where (a_1, \dots, a_L) is the amino acid sequence of the CDR3 (i.e., the translation of $\bar{\tau}$), and L is its length. The factors q_L , $q_{i:L}(a)$, and q_{VJ} denote selective pressures on the CDR3 length, its composition, and the associated VJ identities, respectively. Note that the D segment is entirely included in this junctional region, and therefore,

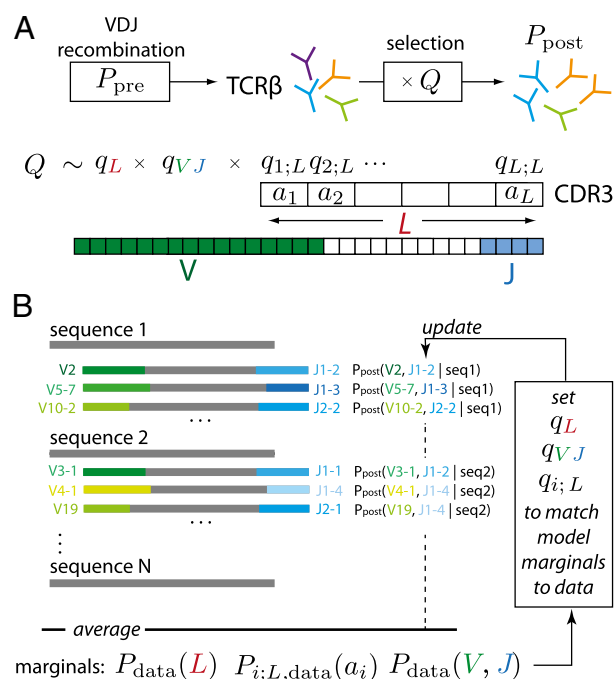


Fig. 1. Graphical representation of our method. (A) T-cell receptor β -chain sequences are formed during VDJ recombination. Sequences from this probability distribution, described by P_{pre} , are then selected with a factor Q defined for each sequence, resulting in the observed P_{post} distribution of receptor sequences. Selection is assumed to act independently on the V and J genes, the length of the CDR3 region, and each of the amino acids, a_i , therein. (B) A schematic of the fitting procedure: the parameters are set so that P_{post} fits the marginal frequencies of amino acids at each position, the distribution of CDR3 lengths, and VJ gene choices. Because the latter is not known unambiguously from the observed sequences, it is estimated probabilistically using the model itself in an iterative procedure.

selection acting on it is encoded in the $q_{i:L}$ factors. Z enforces the model normalization condition $\sum_{\bar{\tau}, V, J} Q(\bar{\tau}, V, J) P_{\text{pre}}(\bar{\tau}, V, J) = 1$.

Because V and J cannot always be inferred deterministically from the read $\bar{\sigma}$, the V and J assignments of any given read will have to be treated as probabilistically defined hidden variables. In addition, because of correlations in P_{pre} , the q factors cannot be identified with marginal enrichment factors [therefore, for example, $P_{i:L, \text{data}}(a_i)/P_{i:L, \text{pre}}(a_i)$ cannot be set equal to $q_{i:L}(a_i)$]. For these reasons, we must use a maximum likelihood procedure to learn the q_L , $q_{i:L}$, and q_{VJ} factors of Eq. 1. We use an expectation maximization algorithm that iteratively modifies the q values until the observed marginal frequencies—CDR3 length distribution, amino acid usage as a function of CDR3 position, and VJ usage—in the data match those implied by the model distribution in Eq. 1, with the preselection distribution P_{pre} being taken as a fixed, known input. The procedure is schematically depicted in Fig. 1B (full details are in *SI Appendix*).

Our model for the selection factor Q assumes factorization on a small set of sequence features with no interactions between these features. This choice is in the same spirit as the classic position weight matrix method for identifying transcription factor binding sites (17). We have verified that such interactions are, in fact, not necessary to describe the data: Fig. 2B plots the covariances of amino acid pairs as predicted by P_{post} vs. the observed values in the data, and *SI Appendix*, Fig. S10 displays a similar comparison of the covariances of (V, J) with L on the one hand and the (V, J) identity with amino acid choice on the other hand. All of the pairwise correlations in the data are well-predicted by the model, although Q does not model them directly. Nonzero pairwise correlations are, in fact, inherited from

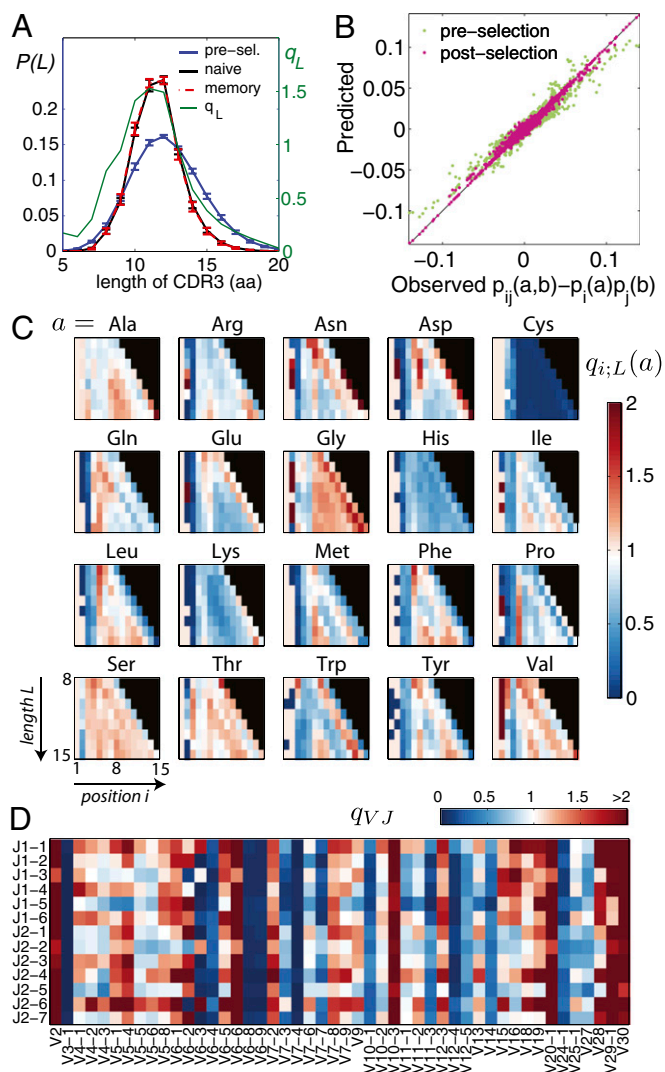


Fig. 2. Characteristics of selection. (A) CDR3 length distributions pre- and postselection and the length selection factor q_L (green). Selection makes the length distribution of CDR3 regions in the preselection repertoire more peaked for the naive and memory repertoires (overlapping). Error bars show standard deviation over nine individuals. (B) Comparison between data and the model of the connected pairwise correlation functions, which were not fitted by our model. The excellent agreement validates the inference procedure. As a control, the prediction from the preselection model (green) does not agree with the data as well. (C) Values of the inferred amino acid selection factors for each amino acid, ordered by length of the CDR3 region (ordinate) and position in the region (abscissa). (D) Values of the VJ gene selection factors.

the preselection distribution, which has correlations of its own (shown by the green points in Fig. 2B).

Another assumption of our model is that selection acts at the level of the amino acid sequence, regardless of the underlying codons. To test this, we learned more general models, where a represented one of 61 possible codons instead of one of 20 aa. We found that codons coding for the same residue had similar selection factors (SI Appendix, Fig. S2), except near the edges of the CDR3, where amino acids may actually come from genomic V and J segments and reflect their codon biases.

To compare the different donors, we learned a distinct model for each donor and cell type (memory or naive) as well as a universal model for all sequences of a given type from all donors taken together (details are in SI Appendix). We also learned

models from random subsets of the sequence dataset to assess the effects of low-number statistical noise.

Results

Characteristics of Selection and Repertoire Diversity. The length, single-residue, and VJ selection factors, learned from the naive datasets of all donors taken together, are presented in Fig. 2A, C, and D. The q_L factor (Fig. 2A) simply reflects the substantial reduction in variance in CDR3 lengths between the preselection ensemble and the observed sequence datasets. The q_{VJ} factor shows that the different V and J genes are subject to a wide range of selection factors (note that these factors act in addition to the quite varied gene segment use probabilities in P_{pre}). The position-dependent amino acid selection factors $q_{i:L}(a)$ are also quite variable but have striking systematic features, such as uniform suppression (or enhancement) away from the CDR3 region boundaries. We looked for correlations between the $q_{i:L}(a)$ and a variety of amino acid biochemical properties (18): hydrophobicity, charge, pH, polarity, volume, and propensities to be found in α - or β -structures in turns at the surface of a binding interface, on the rim, or in the core (19) (details in SI Appendix). We found no significant correlations, except for a negative correlation with amino acid volume and α -helix association as well as a positive correlation with the propensities to be in turns or the core of an interacting complex (SI Appendix, Fig. S7).

To estimate differences between datasets, we calculated the correlation coefficients between the logs of the q_{VJ} and $q_{i:L}(a)$ selection factors (SI Appendix, Fig. S4). Comparing naive vs. naive, memory vs. memory, or naive vs. memory between donors (Fig. 3A–C shows an example for $q_{i:L}$, and SI Appendix, Fig. S3 shows an example for q_{VJ}) gave correlation coefficients of ~ 0.9 in $\log q_{i:L}$, whereas the naive vs. memory repertoires of the same donor gave 0.95. To get a lower bound on small-number statistical noise, we also compared the factors inferred from artificial datasets obtained by randomly shuffling sequences between donors (SI Appendix), yielding an average correlation coefficient of 0.98. Repeating the analysis for $\log q_{VJ}$, we found correlation coefficients of ~ 0.8 between datasets of different donors and 0.84 for the naive and memory dataset of the same donor, all of which must be compared with 0.94, which was obtained between shuffled datasets. We also calculated Jensen–Shannon divergences (SI Appendix) between the P_{post} distributions of all donors and found them to be small—0.07 bits on average. Thus, the observed differences between donors of $q_{i:L}$ and q_{VJ} are small and consistent with their expected statistical variability.

We use Shannon entropy, $S = -\sum_{\vec{r}, V, J} P_{post}(\vec{r}, V, J) \cdot \log_2 P_{post}(\vec{r}, V, J)$, to quantify the diversity of the naive and memory distributions. Entropy is a diversity measure that accounts for nonuniformity of the distribution, and it is additive in independent components. Because $S = \log_2 \Omega$ when there are Ω equally likely outcomes, the diversity index 2^S can be viewed as an effective number of states. The entropy of the naive repertoire according to the model is 38 bits (corresponding to a diversity of $\sim 2.7 \cdot 10^{11}$), which is down from 43.5 bits in the preselection repertoire (Fig. 3D). The majority of this 5.5-bits (or 50-fold) reduction in diversity comes from insertions and deletions, which accounted for most of the diversity in the preselection repertoire. The entropies of the memory and naive repertoires are the same, indicating that selection in the periphery does not further reduce diversity.

Knowing the postselection distribution of sequences, we can ask how different features of the recombination scenario fare in the face of selection. We do not mean to imply that somatic selection acts on the scenarios themselves—it acts on the final product—but it is an a posteriori assessment of the fitness of particular rearrangements. For example, the distributions of insertions at VD and DJ junctions in the postselection ensemble have shorter tails

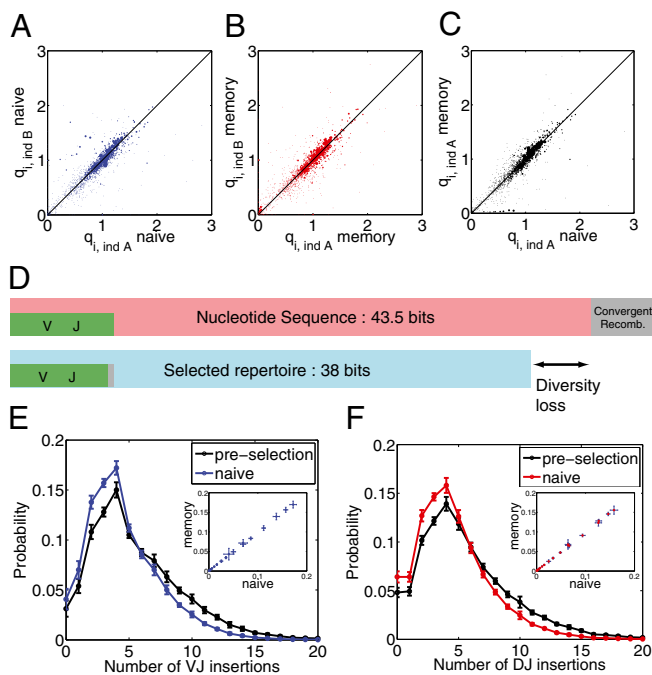


Fig. 3. Repertoire diversity. (A–C) Variability between repertoires. The scatter between $q_{i,ind}$ selection factors between two sample individuals A and B for (A) naive and (B) memory repertoires compared with that of (C) memory and naive repertoires for the same individual shows great similarity between them (SI Appendix, Fig. S4). (D) The entropy of the preselection repertoire (Upper) is reduced in the postselection repertoire (Lower). (E and F) Distribution of (E) VJ and (F) DJ insertions in the preselection and naive repertoires shows elimination of long insertions. Error bars show standard deviations over nine donors. The insertion distributions for the memory repertoire are the same as for the naive repertoire (scatter plots in Insets).

(Fig. 3 E and F), whereas the distribution of deletions at the junctions seems little affected by selection (SI Appendix, Fig. S5), although large numbers of deletions are selected against.

Selection Factor Q as a Measure of Fitness. The selection factor Q is a proxy for the probability of an in-frame sequence after it is generated by recombination to survive the different forms of selection to which it is subjected: the proper folding of the T-cell receptor (TCR) protein, appropriate binding to self peptides, etc. One can think of Q as an intrinsic physical property of the β -chain, and it is instructive to compare Q distributions of the various sequence repertoires of interest: preselection model P_{pre} , postselection model P_{post} , and postselection observed sequences; for each of these repertoires, we assign a Q value to each sequence using the inferred model and create Q -value histograms, denoted by $P_{pre}(Q)$, $P_{post}(Q)$, and $P_{data}(Q)$, respectively (SI Appendix, Eqs. S34–S37 shows details of the calculations).

We observe that the data sequences are enriched in large Q values compared with preselection sequences (Fig. 4 A, Inset and B, Inset), consistent with the interpretation of Q as a selection factor. Furthermore, because the definition of P_{post} implies that $P_{post}(Q) = QP_{pre}(Q)$, we expect $P_{data}(Q)/P_{pre}(Q) = Q$ if the selection model accurately describes the data. This ratio is plotted in Fig. 4, and we see that, for $Q \leq 5$ (accounting for more than 91% of the data sequences), this ratio is, indeed, equal to Q , whereas for $Q > Q_{max} \sim 7$ (accounting for less than 3% of the data sequences), the ratio plateaus. Thus, only the small population of high- Q (fittest) data sequences fails to satisfy this stringent model prediction.

The approach of projecting genotypes onto a single phenotypic variable and using the distribution of that variable to

identify selection effects has previously been used to characterize the fitness landscape of transcription factor binding sites (20, 21). Although in that problem, the phenotypic variable, equivalent to our $\log Q$, is simply the binding affinity of the sequence to the transcription factor, we have (so far) not been able to identify a simple physical quantity linked to Q .

The high- Q plateau suggests that sequences with $Q > Q_{max}$ all have the same selective advantage within the resolution of the model. We can use this line of reasoning to put bounds on the probability for rearranged TCR sequences to pass selection. If we assume that Q is proportional to the probability for sequence $(\vec{\tau}, V, J)$ to be selected, then $P_{sel}(\vec{\tau}, V, J) = \alpha Q(\vec{\tau}, V, J)$. Because P_{sel} cannot exceed unity, Q cannot exceed α^{-1} or $\alpha < Q_{max}^{-1}$. The mean probability that a sequence produced by VDJ rearrangement will pass selection is $\sum_{\vec{\tau}, V, J} P_{pre}(\vec{\tau}, V, J) P_{sel}(\vec{\tau}, V, J) = \alpha$ (as follows from the normalization condition on P_{post}). Thus, an upper limit on the average fraction of rearranged TCRs to pass selection is $\alpha < Q_{max}^{-1} \simeq 15\%$. This limit is consistent with existing estimates (2) for passing positive and negative thymic selection: 10–30% for positive selection only and $\sim 5\%$ for both together. Our analysis only includes the β -chain, and including the α -chain could further reduce our estimate.

The saturation phenomenon indicates that our model is too coarse-grained to describe the very fit (high- Q) sequences. Because of its factorized structure, our model can only account for the coarse features of selection and may not capture very individual-specific traits, such as avoidance of self (corresponding to $Q \ll 1$ in localized regions of the sequence space) or response to pathogens ($Q \gg 1$ for particular sequences). This individual-dependent ruggedness of the fitness landscape Q , schematized in Fig. 4C, is probably ignored by our description and may be hard to model in general. To check that the saturation does not affect our inference procedure, we relearned our model parameters from simulated data, where sequences were generated from P_{pre} and then selected with probability $\min(Q/Q_{max}, 1)$ (details in SI Appendix). We found that essentially the same model was recovered (SI Appendix, Fig. S6).

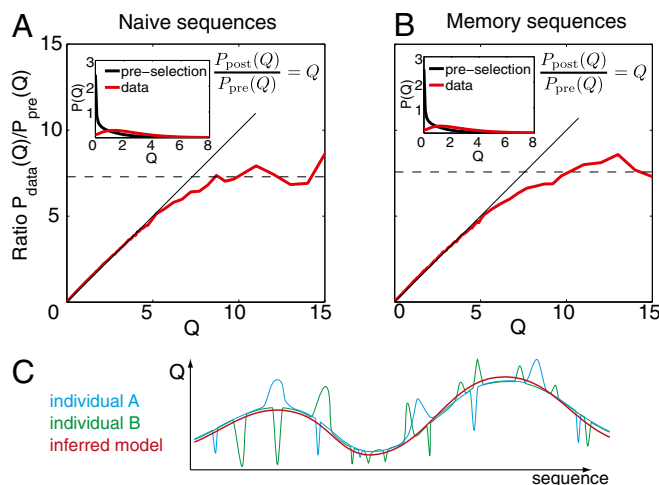


Fig. 4. Probability of passing selection. (A and B) Ratio of the distributions of sequence-wide selection factors Q between the observed sequences and the preselection ensemble (red line), plotted as a function of Q for (A) naive and (B) memory repertoires. The model prediction $P_{post}(Q)/P_{pre}(Q) = Q$ is shown in black, and the preselection and observed distributions of Q are shown in Insets. The selection ratio saturates around approximately seven, which may be interpreted as the maximum probability of being selected. Naive and memory repertoires show similar behaviors. (C) A cartoon of the effective selection landscape captured by our model (red line). Our method does not capture localized selection pressures (such as avoiding self) specific to each individual but captures general global properties.

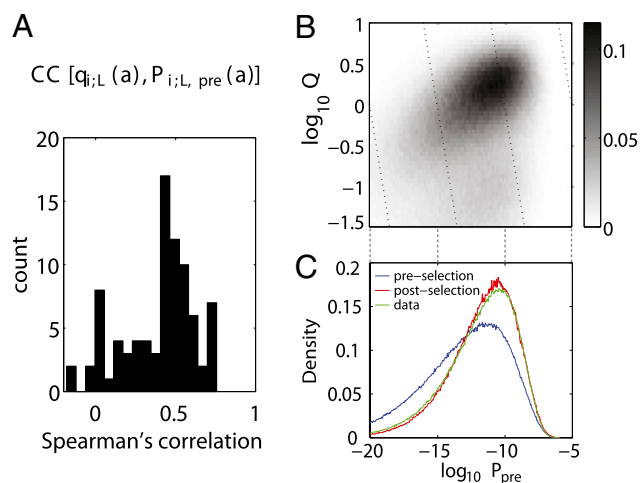


Fig. 5. Correlations between the pre- and postselection repertoires. (A) A histogram of Spearman correlation coefficient (CC) values between the $q_{i,L}$ selection factors in the CDR3 region and their generation probabilities $P_{i,L,pre}(a)$ for all i, L shows an abundance of positive correlations. (B) Heat map of the joint distribution of the preselection probability distribution P_{pre} and selection factors Q for each sequence shows that the two quantities are correlated. (C) Sequences in the observed selected repertoire (green line) had a higher probability to have been generated by recombination than unselected sequences (blue line). Agreement between the postselection model (red line) and data distribution (green line) is a validation of the model.

Natural Selection Anticipates Somatic Selection. Comparing the pre- and postselection length distributions in Fig. 2A shows that the CDR3 lengths that were the most probable to be produced by recombination are also more likely to be selected. Formally, the Spearman rank correlation coefficient between $P_{pre}(L)$ and q_L is 0.76, showing good correlation between the probability of a CDR3 length and the corresponding selection factor. We asked whether this correlation was also present in the other sequence features. The histogram of Spearman correlations between the selection factors $q_{i,L}(a)$ and the preselection amino acid use $P_{i,L,pre}(a)$ for different lengths and positions (i, L) (Fig. 5A) shows a clear majority of positive correlations. Likewise, the selection factors q_{VJ} are positively correlated with the preselection VJ use $P_{VJ,pre}$ (Spearman rank correlation = 0.3, $P < 2 \cdot 10^{-20}$).

The correlations observed for each particular feature of the sequence (CDR3 length, amino acid composition, and VJ use) combine to create a global correlation between the probability $P_{pre}(\vec{r}, V, J)$ that a sequence \vec{r}, V, J was generated by recombination and its propensity $Q(\vec{r}, V, J)$ to be selected (Spearman rank correlation = 0.4, $P = 0$) (Fig. 5B). Consistent with this observation, the postselection repertoire is enriched in sequences that have a high probability to be produced by recombination (Fig. 5C). This enrichment is well-predicted by the model, providing another validation of its predictions at the sequence-wide level.

Taken together, these results suggest that the mechanism of VDJ recombination has evolved to preferentially produce sequences that are more likely to be selected by thymic or peripheral selection.

Shared Sequences Between Individuals. The observation of unique sequences that are shared between different donors has suggested that these sequences make up a public repertoire common to many individuals that is formed through convergent evolution or a common source. However, it is also possible that these common sequences are just statistically more frequent (6) and likely to be randomly recombined in two individuals independently, as discussed by Venturi et al. (7, 22). In other

words, public sequences could just be chance events. Here, we revisit this question by asking whether the number of observed shared sequences between individuals is consistent with random choice from our inferred sequence distribution P_{post} .

We estimated the expected number of shared sequences between groups of donors in two ways: (i) by assuming that each donor had its own private model learned from his own sequences or (ii) by assuming that sequences are drawn from a universal model learned from all sequences together (details on how these estimates are obtained from the models are in *SI Appendix*). Although the latter ignores small but perhaps, significant differences between the donors, the former may exaggerate them where statistics are poor. In Fig. 6A, we plot, for each pair of donors, the expected number of shared nucleotide sequences in their naive repertoires under assumptions i and ii vs. the observed number. The number is well-predicted under both assumptions: the universal model assumption gives a slight overestimate, and the private model gives a slight underestimate. We repeat the analysis for sequences that are observed to be common to at least three or four donors (Fig. 6B and C). The universal model predicts their number better than the private models, although it still slightly overestimates it.

These results suggest that shared sequences are, indeed, the result of pure chance. If that is so, shared sequences should have a higher occurrence probability than average; specifically, the model predicts that the sequences that are shared between at least two donors are distributed according to P_{post}^2 (*SI Appendix*). We test this prediction by plotting the distribution of P_{post} for regular sequences as well as pairwise-shared sequences according to the model and in the naive datasets (Fig. 6D), and we find excellent agreement. In general, sequences that are shared

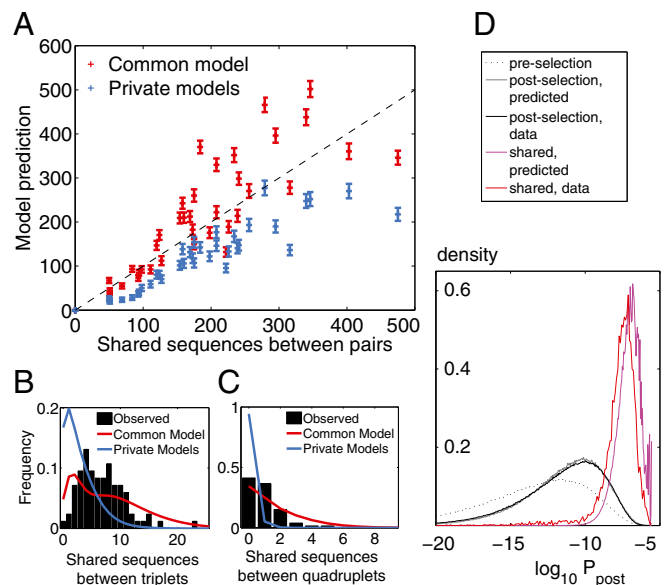


Fig. 6. Shared sequences between individuals. (A) The mean number of shared sequences between any pair of individuals compared with the number expected by chance (model prediction) for one common model for all individuals (red crosses) and private models learned independently for each individual (blue crosses). Error bars are standard deviations from distributions over pairs. The distribution of shared sequences between (B) triplets and (C) quadruplets of individuals for the data (black histogram) from common (red line) and private (blue line) models. (D) The shared sequences are most likely to be generated and selected: comparison of the P_{post} postselection distribution for sequences from the preselection (dotted line) and post-selection repertoires (according to the model in gray and the data in black) as well as the sequences shared by at least two donors (model prediction in magenta and data in red).

between at least n individuals by chance should be distributed according to P_{post}^n . For triplets and quadruplets, this model prediction is not as well-verified (SI Appendix, Fig. S8). This discrepancy may be explained by the fact that such sequences are outliers with very high occurrence probabilities and may not be well-captured by the model, which was learned on typical sequences.

We repeated these analyses for sequences shared between the memory repertoires of different individuals with very similar conclusions, except for donors 2 and 3 and donors 2 and 7, who shared many more sequences than expected by chance (SI Appendix, Fig. S9). We conclude that the vast majority of shared sequences occurs by chance and is well-predicted by our model of random recombination and selection.

Discussion

We have introduced and calculated a selection factor $Q(\bar{\sigma})$ that serves as a measure of selection acting on a given receptor sequence $\bar{\sigma}$ in the somatic evolution of the immune repertoire. Using this measure, we show that the observed repertoires have undergone significant selection starting from the initial repertoire produced by VDJ recombination.

We find little difference between the naive and memory repertoires, which is in agreement with recent findings showing no correlation between TCR sequence and T-cell fate (23). We also find little difference between the repertoires of different donors, which is perhaps surprising, because the donors have distinct HLA types and could, therefore, experience markedly different selective pressures. Also, memory sequences have undergone additional selection compared with the naive ones—pathogen recognition—and could show different signatures of selection. A possible interpretation of both findings is that our model only captures coarse and universal features of selection related to the general fitness of receptors and not fine-grained, individual-specific selective pressures, such as avoidance of self, or recognition of particular pathogen epitopes, as illustrated schematically in Fig. 4C. A strategy for incorporating these highly specific effects in our analysis has yet to be defined. In other words, our selection factors may smooth out the complex landscapes of specific repertoires and fail to capture individual-specific tall peaks or deep

valleys in the landscape of selection factors. To really probe these fine-grained individual-specific details, we need to develop methods based on accurate sequence counts. Another interesting future direction would be to see whether, at this global level, the signatures of selection are similar between (relatively) isolated populations. Lastly, comparing data from different species (mice and fish), particularly where inbred individuals with the same HLA type can be compared, would be an interesting avenue for addressing these issues.

Our results suggest that natural selection has refined the VDJ recombination process over evolutionary timescales to produce a preselection repertoire that anticipates the downstream actions of somatic selection: sequences that are likely to fail selection are not very likely to be produced in the first place. Because of this rich become richer effect, selection reduces the diversity of the repertoire by a factor of 50 in terms of diversity index. This reduction in diversity does not mean that only 2% of the sequences pass selection: our results are consistent with an acceptance ratio as large as 15%. This paradoxical result is possible because selection, by preferentially keeping clones that were more likely to be generated, gets rid of the many rare clones that are responsible for the large initial sequence diversity. We do not have a mechanistic understanding of how the VDJ recombination process has evolved to achieve this result. Exploration of this question would require an analysis of data on multiple species in different environments.

To summarize, our work has provided the first, to our knowledge, quantitative statistical description of the way that thymic selection and later, peripheral selection modify the TCR sequence repertoire that emerges from VDJ recombination. These results provide a detailed characterization of the background against which one would have to work to detect sequence signatures of more subtle selection effects, such as those associated with autoimmunity and pathogen response.

ACKNOWLEDGMENTS. The work of Y.E., T.M., and A.M.W. was supported, in part, by European Research Council Starting Grant 306312. The work of C.G.C. was supported, in part, by National Science Foundation Grants PHY-0957573 and PHY-1305525 and the W. M. Keck Foundation Award (dated December 15, 2009).

- Murugan A, Mora T, Walczak AM, Callan CG, Jr. (2012) Statistical inference of the generation probability of T-cell receptors from sequence repertoires. *Proc Natl Acad Sci USA* 109(40):16161–16166.
- Janeway C (2005) *Immunobiology, the Immune System in Health and Disease* (Garland, New York).
- Weinstein JA, Jiang N, White RA, Fisher DS, Quake SR (2009) High-throughput sequencing of the zebrafish antibody repertoire. *Science* 324(5928):807–810.
- Ndifon W, et al. (2012) Chromatin conformation governs T-cell receptor β gene segment usage. *Proc Natl Acad Sci USA* 109(39):15865–15870.
- Mora T, Walczak AM, Bialek W, Callan CG, Jr. (2010) Maximum entropy models for antibody diversity. *Proc Natl Acad Sci USA* 107(12):5405–5410.
- Quigley MF, et al. (2010) Convergent recombination shapes the clonotypic landscape of the naive T-cell repertoire. *Proc Natl Acad Sci USA* 107(45):19414–19419.
- Venturi V, Price DA, Douek DC, Davenport MP (2008) The molecular basis for public T-cell responses? *Nat Rev Immunol* 8(3):231–238.
- Baum PD, Venturi V, Price DA (2012) Wrestling with the repertoire: The promise and perils of next generation sequencing for antigen receptors. *Eur J Immunol* 42(11):2834–2839.
- Six A, et al. (2013) The past, present, and future of immune repertoire biology – the rise of next-generation repertoire analysis. *Front Immunol* 4:1–16.
- Robins H (2013) Immunosequencing: Applications of immune repertoire deep sequencing. *Curr Opin Immunol* 25(5):646–652.
- Yates AJ (2014) Theories and quantification of thymic selection. *Front Immunol* 5:13.
- Jameson SC (2002) Maintaining the norm: T-cell homeostasis. *Nat Rev Immunol* 2(8):547–556.
- Detours V, Mehr R, Perelson AS (1999) A quantitative theory of affinity-driven T cell repertoire selection. *J Theor Biol* 200(4):389–403.
- Kosmrlj A, Jha AK, Huseby ES, Kardar M, Chakraborty AK (2008) How the thymus designs antigen-specific and self-tolerant T cell receptor sequences. *Proc Natl Acad Sci USA* 105(43):16671–16676.
- Robins HS, et al. (2009) Comprehensive assessment of T-cell receptor beta-chain diversity in alphabeta T cells. *Blood* 114(19):4099–4107.
- Robins H, et al. (2010) Overlap and effective size of the human CD8+ T cell receptor repertoire. *Sci Transl Med* 2(47):47ra64.
- Berg OG, von Hippel PH (1987) Selection of DNA binding sites by regulatory proteins. Statistical-mechanical theory and application to operators and promoters. *J Mol Biol* 193(4):723–750.
- Stryer L, Berg JM, Tymoczko JL (2002) *Biochemistry* (Freeman, New York), 5th Ed.
- Martin J, Lavery R (2012) Arbitrary protein-protein docking targets biologically relevant interfaces. *BMC Biophysics* 1(5):7.
- Mustonen V, Lässig M (2005) Evolutionary population genetics of promoters: Predicting binding sites and functional phylogenies. *Proc Natl Acad Sci USA* 102(44):15936–15941.
- Mustonen V, Kinney J, Callan CG, Jr., Lässig M (2008) Energy-dependent fitness: A quantitative model for the evolution of yeast transcription factor binding sites. *Proc Natl Acad Sci USA* 105(34):12376–12381.
- Venturi V, et al. (2006) Sharing of T cell receptors in antigen-specific responses is driven by convergent recombination. *Proc Natl Acad Sci USA* 103(49):18691–18696.
- Wang C, et al. (2010) High throughput sequencing reveals a complex pattern of dynamic interrelationships among human T cell subsets. *Proc Natl Acad Sci USA* 107(4):1518–1523.