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]

[a]Laboratoire de Physique Théorique, Unité Mixte de Recherche 8549 and [d]Laboratoire de Physique Statistique, Unité Mixte de Recherche 8550, Centre National de la Recherche Scientifique and École Normale Supérieure, 75005 Paris, France; [b]Department of Applied Physics, Stanford University, Stanford, CA 94305; and [c]Joseph 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.

[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.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 ∼200,000 unique naive and ∼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 $P_{\rm pre}(\vec{\sigma})$ that a VDJ recombination event will produce a β-chain gene consistent with the sequence read $\vec{\sigma}$ (for any $\vec{\sigma}$). 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, $P_{\rm post}(\vec{\sigma})$, differs from the preselection distribution $P_{\rm pre}(\vec{\sigma})$. (We only consider the presence or absence of a sequence $\vec{\sigma}$ and not the size of its clone.) Here, we distinguish between the read $\vec{\sigma}$ 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 $\vec{\tau}$; 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 $\vec{\tau}$ 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 $\vec{\sigma}$, we must consider selection in terms of the full β-chain $(\vec{\tau},V,J)$ rather than the incomplete $\vec{\sigma}$.

For each β-chain sequence $(\vec{\tau},V,J)$, we define a selection factor $Q = P_{\rm post}/P_{\rm pre}$ 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 $P_{\rm pre}$ 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 $P_{\rm post}(\vec{\tau},V,J) = Q(\vec{\tau},V,J)P_{\rm pre}(\vec{\tau},V,J)$ gives a good account of an observed set of selected sequences. We cannot directly estimate $P_{\rm post}(\vec{\tau},V,J)$ from the data, but as we outline in Fig. 1*A*, we can use a reduced complexity model for Q to infer it (and therefore $P_{\rm post}$) from the data. Specifically, we will show that the following factorized model for Q captures the main features of selection:

$$Q(\vec{\tau},V,J) = \frac{P_{\rm post}(\vec{\tau},V,J)}{P_{\rm pre}(\vec{\tau},V,J)} = \frac{1}{Z} q_L q_{VJ} \prod_{i=1}^{L} q_{i;L}(a_i), \qquad [1]$$

where $(a_1, \ldots, a_L)$ is the amino acid sequence of the CDR3 (i.e., the translation of $\vec{\tau}$), 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,

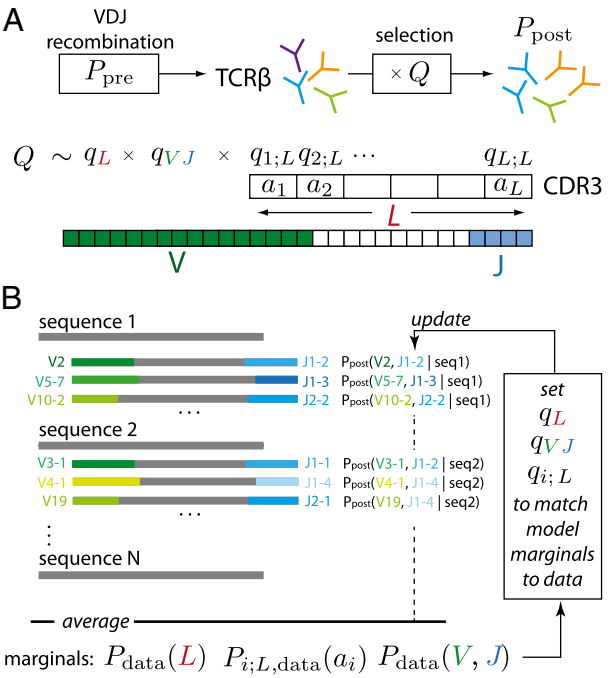

Fig. 1. Graphical representation of our method. (*A*) T-cell receptor β-chain sequences are formed during VDJ recombination. Sequences from this probability distribution, described by $P_{\rm pre}$, are then selected with a factor Q defined for each sequence, resulting in the observed $P_{\rm post}$ 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 $P_{\rm post}$ 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 $\sum_{\vec{\tau},V,J} Q(\vec{\tau},V,J)P_{\rm pre}(\vec{\tau},V,J) = 1$.

Because V and J cannot always be inferred deterministically from the read $\vec{\sigma}$, 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 $P_{\rm pre}$, the q factors cannot be identified with marginal enrichment factors [therefore, for example, $P_{i;L,{\rm data}}(a_i)/P_{i;L,{\rm pre}}(a_i)$ 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 $P_{\rm pre}$ being taken as a fixed, known input. The procedure is schematically depicted in Fig. 1*B* (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. 2*B* plots the covariances of amino acid pairs as predicted by $P_{\rm post}$ 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

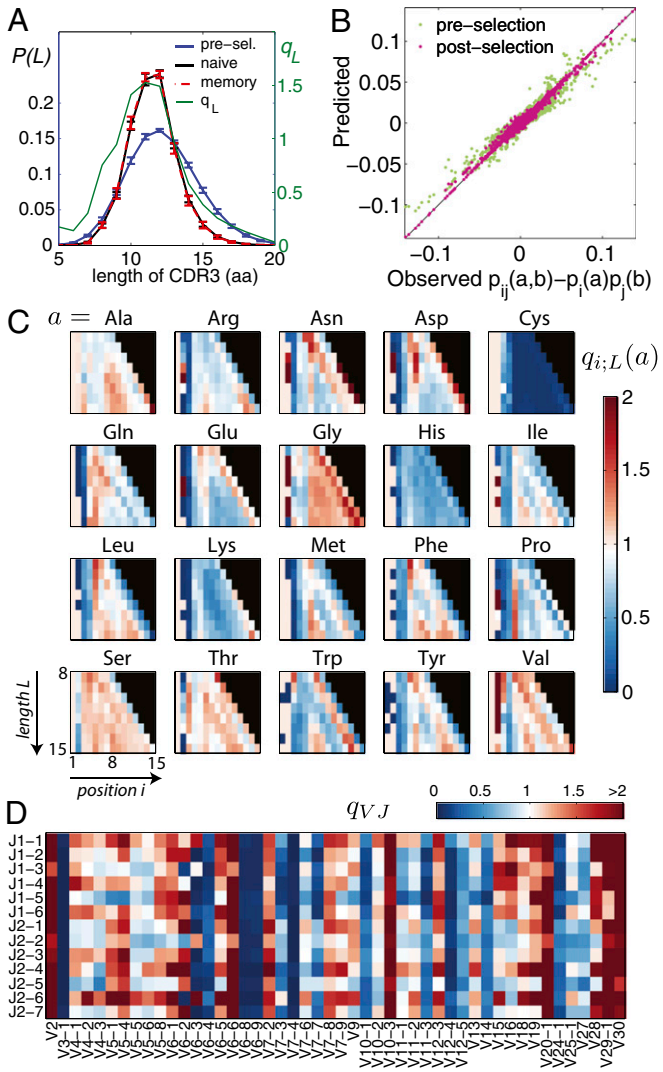

Fig. 2. Characteristics of selection. (*A*) CDR3 length distributions pre- and postselection and the length selection factor q_L (green). Selection makes the length distribution of CDR3 regions in the preselection repertoire more peaked for the naive and memory repertoires (overlapping). Error bars show 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. 2 *A*, *C*, and *D*. The q_L factor (Fig. 2*A*) 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 $P_{\rm pre}$). 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 $q_{i;L}(a)$ factors 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. 3 *A*–*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 $P_{\rm post}$ 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, $S=-\sum_{\vec{\tau},V,J}P_{\rm post}(\vec{\tau},V,J)\cdot\log_2 P_{\rm post}(\vec{\tau},V,J)$, 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. 3*D*). 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

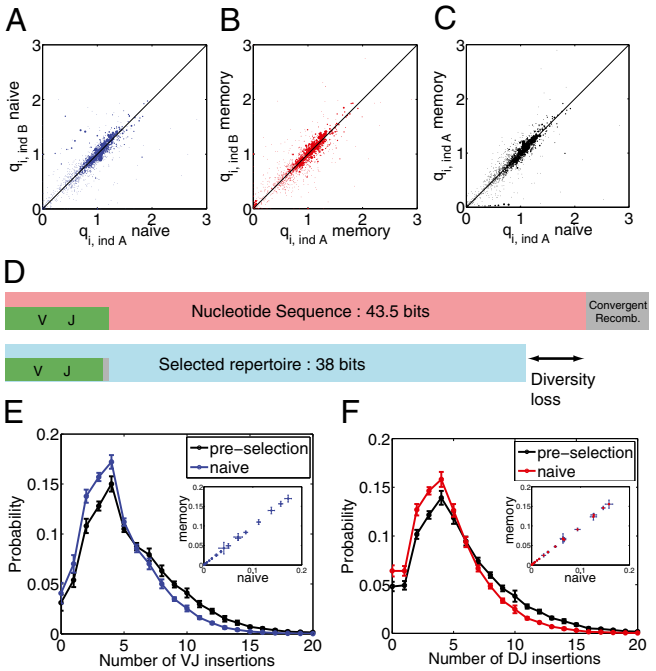

Fig. 3. Repertoire diversity. (*A*–*C*) Variability between repertoires. The scatter between $q_{i;L}$ selection factors between two sample individuals A and B for (*A*) naive and (*B*) memory repertoires compared with that of (*C*) memory and naive repertoires for the same individual shows great similarity 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 $P_{\rm pre}$, postselection model $P_{\rm post}$, 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 $P_{\rm pre}(Q)$, $P_{\rm post}(Q)$, and $P_{\rm data}(Q)$, 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 $P_{\rm post}$ implies that $P_{\rm post}(Q) = QP_{\rm pre}(Q)$, we expect $P_{\rm data}(Q)/P_{\rm pre}(Q) = Q$ if the selection model accurately describes the data. This ratio is plotted in Fig. 4, and we see that, for $Q \le 5$ (accounting for more than 91% of the data sequences), this ratio is, indeed, equal to Q, whereas for $Q > Q_{\rm max} \sim 7$ (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 $Q > Q_{\rm max}$ 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 $P_{\rm sel}(\vec{\tau}, V, J) = \alpha Q(\vec{\tau}, V, J)$. Because $P_{\rm sel}$ 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 $\sum_{\vec{\tau},V,J} P_{\rm pre}(\vec{\tau}, V, J) P_{\rm sel}(\vec{\tau}, V, J) = \alpha$ (as follows from the normalization condition on $P_{\rm post}$). 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 ∼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. 4*C*, 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 $P_{\rm pre}$ and then selected with probability $\min(Q/Q_{\rm max}, 1)$ (details in *SI Appendix*). We found that essentially the same model was recovered (*SI Appendix*, Fig. S6).

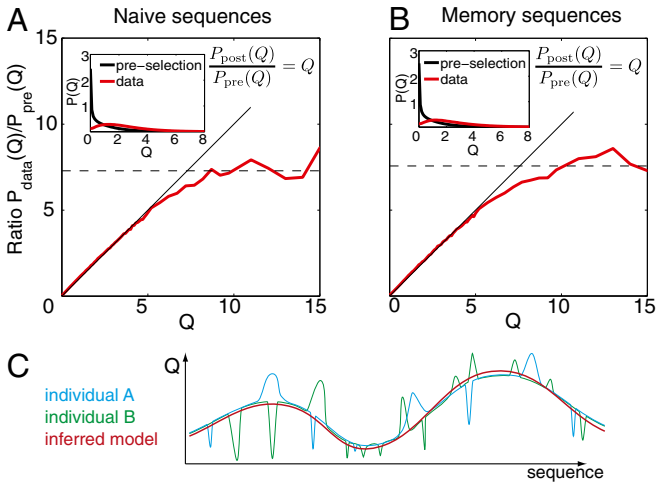

Fig. 4. Probability of passing selection. (*A* and *B*) Ratio of the distributions of sequence-wide selection factors Q between the observed sequences and the preselection ensemble (red line), plotted as a function of Q for (*A*) naive and (*B*) memory repertoires. The model prediction $P_{\rm post}(Q)/P_{\rm pre}(Q) = Q$ 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.

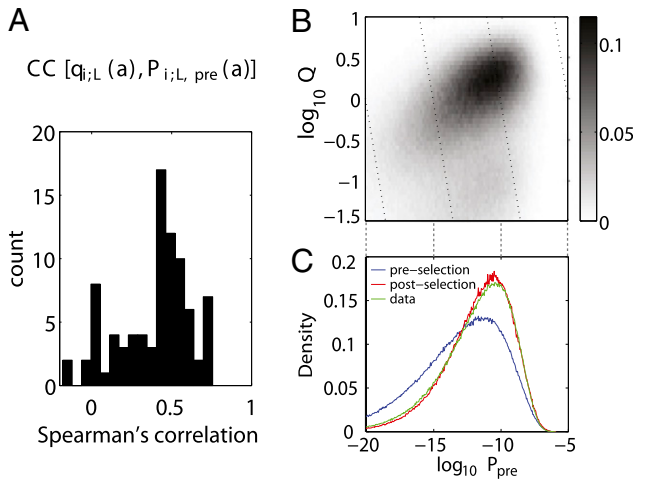

Fig. 5. Correlations between the pre- and postselection repertoires. (*A*) A histogram of Spearman correlation coefficient (CC) values between the $q_{i;L}(a)$ selection factors in the CDR3 region and their generation probabilities $P_{i;L,pre}(a)$ for all i, L shows an abundance of positive correlations. (*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. 2*A* 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. 5*A*) 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 $P_{pre}(\vec{\tau},V,J)$ that a sequence $\vec{\tau},V,J$ was generated by recombination and its propensity $Q(\vec{\tau},V,J)$ to be selected (Spearman rank correlation = 0.4, $P = 0$) (Fig. 5*B*). Consistent with this observation, the postselection repertoire is enriched in sequences that have a high probability to be produced by recombination (Fig. 5*C*). 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. 6*A*, 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. 6 *B* 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. 6*D*), and we find excellent agreement. In general, sequences that are shared

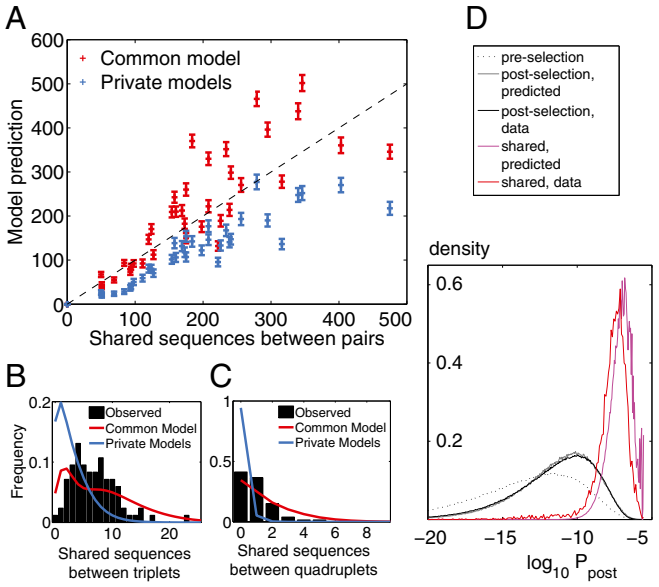

Fig. 6. Shared sequences between individuals. (*A*) The mean number of shared sequences between any pair of individuals compared with the number expected by chance (model prediction) for one common model for all individuals (red crosses) and private models learned independently for each individual (blue 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 postselection 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^n_{post}. 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 $Q(\vec{\sigma})$ that serves as a measure of selection acting on a given receptor sequence $\vec{\sigma}$ 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. 4*C*. 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).

1. 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.
2. Janeway C (2005) *Immunobiology, the Immune System in Health and Disease* (Garland, New York).
3. Weinstein JA, Jiang N, White RA, Fisher DS, Quake SR (2009) High-throughput sequencing of the zebrafish antibody repertoire. *Science* 324(5928):807–810.
4. Ndifon W, et al. (2012) Chromatin conformation governs T-cell receptor Jβ gene segment usage. *Proc Natl Acad Sci USA* 109(39):15865–15870.
5. 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.
6. 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.
7. Venturi V, Price DA, Douek DC, Davenport MP (2008) The molecular basis for public T-cell responses? *Nat Rev Immunol* 8(3):231–238.
8. 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.
9. 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.
10. Robins H (2013) Immunosequencing: Applications of immune repertoire deep sequencing. *Curr Opin Immunol* 25(5):646–652.
11. Yates AJ (2014) Theories and quantification of thymic selection. *Front Immunol* 5:13.
12. Jameson SC (2002) Maintaining the norm: T-cell homeostasis. *Nat Rev Immunol* 2(8):547–556.
13. Detours V, Mehr R, Perelson AS (1999) A quantitative theory of affinity-driven T cell repertoire selection. *J Theor Biol* 200(4):389–403.
14. 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.
15. Robins HS, et al. (2009) Comprehensive assessment of T-cell receptor beta-chain diversity in alphabeta T cells. *Blood* 114(19):4099–4107.
16. Robins H, et al. (2010) Overlap and effective size of the human CD8+ T cell receptor repertoire. *Sci Transl Med* 2(47):47ra64.
17. 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.
18. Stryer L, Berg JM, Tymoczko JL (2002) *Biochemistry* (Freeman, New York), 5th Ed.
19. Martin J, Lavery R (2012) Arbitrary protein-protein docking targets biologically relevant interfaces. *BMC Biophysics* 1(5):7.
20. 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.
21. 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.
22. 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.
23. 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.

Supporting information: Quantifying selection pressures in somatic immune receptor evolution

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

(Dated: May 21, 2014)

I. DATA

The DNA nucleotide data used in our analysis consists of human CD4+ naive (CD45RO-) or memory (CD45RO+) β chain sequences from 9 healthy individuals, sequenced and made available to us by H. Robins and already used in [1]. Reads are 60 base pair long for 6 donors and 101 base pair long for 3 donors (individuals 2, 3 and 7) and contain the CDR3 region and neighboring V and J gene nucleotides. All end at the same position in the J gene, with four nucleotides between this position and the first nucleotide of the conserved phenylalanine. The data were divided into out-of-frame reads (non-coding), used to learn the pre-selection model as described in [1] and in-frame (coding) reads used in the analysis presented in this paper. The sequence data we used are available at `http://princeton.edu/~ccallan/TCRPaper/data/`.

In our study we limit ourselves to unique sequences. The experimental procedure and initial assessment of the quality of the reads were done in the Robins lab following the procedures described in [2, 3]. Each sequence was read multiple times, allowing for the correction of most sequencing errors. The numbers of unique sequences used in each dataset is shown in Table SI.

	Naive	Memory
Donor 1	311917	177744
Donor 2	242254	135567
Donor 3	195007	119906
Donor 4	130958	142017
Donor 5	147848	32468
Donor 6	187245	104119
Donor 7	251335	136419
Donor 8	42326	120527
Donor 9	254349	89830

Table S I: Number of unique coding sequences in each datasets.

The alignment to all possible V and J genes was done using the curated datasets in the IMGT database [4]. There are 48 V genes, 2 D genes and 13 J genes plus a number of pseudo V genes that cannot lead to a functioning receptor due to stop codons. We discarded sequences that were associated to a pseudo-gene as our model only accounts for coding genes. The germline sequences of the genes used in our analysis are the same as were used in [1] to analyze the generative V(D)J recombination process. The complete list of gene sequences can be found at `http://princeton.edu/~ccallan/TCRPaper/genes/`.

II. PRE-SELECTION MODEL

The pre-selection, or generative model, assumes the following structure for the probability distribution of recombination scenarios S [1]:

$$\begin{aligned} P_{\rm pre}(S) =& P(V)P(D,J)P(\text{insVD})P(\text{insDJ}) \\ & P(\text{delV}|V)P(\text{dellD},\text{delrD}|D)P(\text{delJ}|J) \\ & P(s_1)P(s_2|s_1)\cdots P(s_{\text{insVD}}|s_{\text{insVD}-1}) \\ & P(t_1)P(t_2|t_1)\cdots P(t_{\text{insDJ}}|t_{\text{insDJ}-1}), \end{aligned} \tag{1}$$

where a scenario is given by the VDJ choice, the number of insertions insVD, insDJ and the number of deletions (delV,dellD), (delrD,delJ) at each of the two junctions, together with the identities $(s_1,\ldots,s_{\text{insVD}})$,$(t_1,\ldots,t_{\text{insDJ}})$ of the inserted nucleotides. It is worth noting that the insertions are assumed to be independent of the identities of the genes between which insertions are made. By contrast, the deletion probabilities are allowed to depend on the identity of the gene being deleted. The validity of these assumptions is verified *a posteriori*.

III. MODEL FITTING

A. Maximum likelihood formulation

The model probability to observe a given coding nucleotide sequence is:

$$P_{\rm post}(\vec{\tau},V,J) = Q(\vec{\tau},V,J)P_{\rm pre}(\vec{\tau},V,J), \tag{2}$$

where $\vec{\tau}=(\tau_1,\ldots,\tau_{3L})$ is the nucleotide sequence of the CDR3 (defined as running from the conserved cysteine in the V segment up to the last amino acid in the read, leaving two amino acids between the last read amino acid and the conserved phenylalanine in the J segment), L is the length of the CDR3, and V and J index the choice of the germline V and J segments (which completely determine the sequence outside the CDR3 region). The D segment is entirely absorved into $\vec{\tau}$, and is not explicitly

*

tracked in assessing selection. The selection factor Q is assumed to take the following factorized form:

$$Q(\vec{\tau}, V, J) = \frac{1}{Z}\ q_L\ q_{V,J}\ \prod_{i=1}^{L} q_{i;L}(a_i). \tag{3}$$

where $\vec{a} = (a_1, \ldots, a_L)$ is the amino-acid sequence of the CDR3, and Z is a normalization constant that enforces

$$\sum_{\vec{\tau},V,J} P_{\rm post}(\vec{\tau}, V, J) = 1. \tag{4}$$

The probability, $P_{\rm pre}(\vec{\tau}, V, J)$, of generating a specific sequence in a V(D)J recombination event can be obtained from the noncoding sequence reads by the methods explained in [1]. Specifically, the pre-selection model gives the probability $P_{\rm pre}(S)$ of a recombination scenario $S = (V, D, J, \text{insVD}, \text{insDJ}, \text{delV}, \ldots)$ as given by Eq. 1. A scenario S completely determines the sequence $\vec{\tau}$, but the converse is not true. The pre-selection probability for a coding sequence is thus given by

$$P_{\rm pre}(\vec{\tau}, V, J) = \frac{1}{p_{\rm coding}} \sum_{S\to(\vec{\tau},V,J)} P_{\rm pre}(S) \tag{5}$$

where we sum over scenarios resulting in a particular CDR3 sequence $\vec{\tau}$ and a particular V, J pair. The normalization factor $p_{\rm coding} \approx 0.26$ corrects for the fact that a randomly generated sequence is not always productive (*i.e.* in-frame and with no stop codon). From this point on, we regard the initial generation probability of any specific read as known. When we make statements about the pre-selection distribution of CDR3 properties, such as length or amino acid utilization, they are derived from synthetic repertoires drawn from the above pre-selection distribution.

We want to infer the parameters q_L, $q_{V,J}$ and $q_{i;L}(\cdot)$ of the model from the observed coding sequence repertoires. Formally we want to maximize the likelihood of the data given the model. Unfortunately the sequence reads from the data are not long enough to fully specify the V and J segments, so we cannot use $P_{\rm post}(\vec{\tau}, V, J)$ as our raw likelihood. Instead, we need to write the probability of observing a given (truncated) read $\vec{\sigma}$, of length 60 or 101 nucleotides, depending on the donor:

$$P_{\rm post}(\vec{\sigma}) = \sum_{(V,J,\vec{\tau})\to\vec{\sigma}} P_{\rm post}(\vec{\tau}, V, J). \tag{6}$$

where we note again that $(\vec{\tau}, V, J)$ fully specifies $\vec{\sigma}$, while $\vec{\sigma}$ fully specifies $\vec{\tau}$, but not V and J. Given a dataset of N sequences, $\vec{\sigma}^1, \ldots, \vec{\sigma}^N$ (see Fig. S1 for notations), the likelihood reads:

$$\mathcal{L}(Q) = \prod_{a=1}^{N} P_{\rm post}(\vec{\sigma}^a). \tag{7}$$

Our goal is maximize $\mathcal{L}$ with respect to the parameters q_L, $q_{V,J}$, and $q_{i;L}(\cdot)$ (globally refered to as Q).

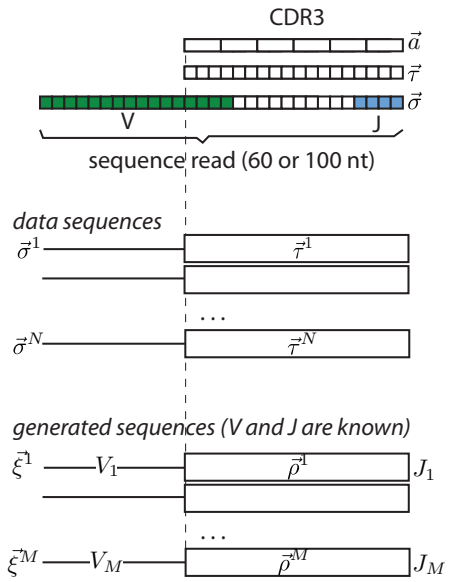

Fig. S 1: Summary of the notations used in this paper for the sequences. The CDR3 region is defined from the conserved cysteine around the end of the V segment to the last amino-acid in the read, leaving two amino acids to the conserved phenylalanine in the J segment. The nucleotides in the read are defined as σ_i, the nucleotides in the CDR3 region as τ_i and the amino acids in the CDR3 region as a_i. The data sequences therefore can be defined in terms of $\vec{\sigma}$, or their V, J genes and $\vec{\tau}$. The generated sequences, with known V and J genes, are defined in terms of $\vec{\xi}$ for the whole sequence or $\vec{\rho}$ for only the CDR3.

B. Expectation maximization

Calculating $P_{\rm post}(\vec{\sigma})$ is computationally intensive. Given the form of the model, it seems more natural to work with $P_{\rm post}(\vec{\tau}, V, J)$, but this likelihood involves the "hidden" variables V and J. To circumvent this problem, we use the expectation maximization algorithm [5, 6]. This algorithm uses an iterative two-step process, with two sets of model parameters Q and Q'. The log-likelihood of the data is calculated using the set of parameters Q'; in the "Expectation" step, this log-likelihood is averaged over the hidden variables with their posterior probabilities, which are calculated using the second set of parameters Q. In the "Maximization" step, this average log-likelihood is maximized over the first set Q', while keeping the second set Q fixed. Then Q is updated to the optimal value of Q', and the two steps are repeated iteratively until convergence.

In practice, starting with a test set of parameters Q, we calculate, for each sequence of the data, the posterior probability of a (V, J) pair:

$$P_{\rm post}(V_a, J_a|\vec{\sigma}^a) = \frac{Q(\vec{\tau}^a, V_a, J_a) P_{\rm pre}(\vec{\tau}^a, V_a, J_a)}{\sum_{V,J} Q(\vec{\tau}^a, V, J) P_{\rm pre}(\vec{\tau}^a, V, J)}. \tag{8}$$

The log-likelihood, expressed in terms of the hidden vari-

ables V and J, is maximized after averaging over V and J using that posterior. Specifically we will maximize:

$$\hat{\mathcal{L}}(Q'|Q) = \sum_{a=1}^{N} \langle \log P_{\rm post}(\vec{\tau}^a, V_a, J_a; Q') \rangle_Q$$
$$\equiv \sum_{a=1}^{N} \sum_{V^a, J^a} P_{\rm post}(V_a, J_a|\vec{\sigma}^a; Q) \log P_{\rm post}(\vec{\tau}^a, V_a, J_a; Q'). \tag{9}$$

Here we have added the Q dependencies explicitly because there are two different parameter sets Q and Q'. The maximization is performed over Q', which parametrizes the log-likelihood itself, while keeping Q, which parametrizes how the average is done over the hidden variables, constant. After each maximization step we substitute:

$$Q \leftarrow \text{argmax}_{Q'} \hat{\mathcal{L}}(Q'|Q), \tag{10}$$

and iterate until convergence. This procedure is guaranteed to find a local maximum of the likelihood $\mathcal{L}(Q)$.

C. Equivalence with fitting marginal probabilities

The expectation-maximization step can be simplified by noting that at the maximum, derivatives vanish:

$$\frac{\partial \hat{\mathcal{L}}(Q'|Q)}{\partial Q'} = 0. \tag{11}$$

Precisely, we take derivatives with each of the parameters, q_L, q_{VJ} *etc.* and set them to zero. Since $P_{\rm post}(\vec{\tau}, V, J)$ is naturally factorized in the Q parameters, we obtain simple expressions, *e.g.* $\partial\hat{\mathcal{L}}/\partial \log q'_L = 0$ gives:

$$\sum_{a=1}^{N} \sum_{V^a, J^a} P_{\rm post}(V_a, J_a|\vec{\sigma}^a; Q) \left(\delta_{L_a, L} - \frac{\partial \log Z}{\partial \log q'_L} \right) = 0, \tag{12}$$

where $\delta_{a,b}$ is Kronecker's delta function. The term in the sum gives the total number of sequences in the data with length L. Besides we have:

$$\frac{\partial \log Z}{\partial \log q'_L} = \sum_{\vec{\tau}, V, J} \delta_{L(\vec{\tau}), L} P_{\rm post}(\vec{\tau}, V, J; Q') = P_{\rm post}(L; Q'). \tag{13}$$

Hence the maximality condition simply becomes:

$$P_{\rm data}(L) = P_{\rm post}(L; Q'), \tag{14}$$

i.e. that the length distribution of the model must be equal to that of the data. Similarly, maximizing with respect to $q_{i;L}(a_i)$ entails that single amino-acid frequencies at a given position are matched between data and model:

$$P_{i;L,\rm data}(a_i) = P_{i;L,\rm post}(a_i; Q'). \tag{15}$$

The condition for q_{VJ} is slightly different, because we do not directly have the frequencies of V and J in the data. This is replaced by their expected frequency under the posterior $P_{\rm post}(V_a, J_a|\vec{\sigma}^a)$ taken with parameters Q:

$$\frac{1}{N} \sum_{a=1}^{N} P_{\rm post}(V, J|\vec{\sigma}^a; Q) = P_{\rm post}(V, J; Q'), \tag{16}$$

where again the left-hand side is the empirical distribution of V and J (indirectly estimated with the help of the model with parameters Q), and the right-hand side is the model distribution of the same quantities (estimated with parameters Q', which are then varied to achieve equality with the data estimate). The approach of iteratively adjusting model parameters to match a corresponding set of data marginals is a conceptually clear and computationally effective implementation of the expectation maximization algorithm.

D. Gauge

As defined above, the model is degenerate: for each i, L, the factors $q_{i;L}(a)$ and Z may be multiplied by a common constant without affecting the model. We need to fix a convention, or gauge, to lift this degeneracy. We impose that, for each i, L:

$$\sum_{a=1}^{20} P_{i;L,\rm pre}(a) q_{i;L}(a) = 1. \tag{17}$$

where $P_{i;L,\rm pre}(a)$ is the probability of having amino-acid a at position i in CDR3s of length L.

E. Numerical implementation

To solve the fitting equations (14)-(16) in practice, we use a gradient descent algorithm:

$$q_L \leftarrow q_L + \epsilon \left[P_{\rm data}(L) - P_{\rm post}(L; Q') \right], \tag{18}$$

and similarly for $q_{i;L}$ and q_{VJ}. To do this, we must be able to calculate the marginals $P_{\rm post}(L; Q')$, $P_{i;L,\rm post}(a_i; Q')$ and $P_{\rm post}(V, J; Q')$ from the model at each step.

This leaves us with the problem of estimating marginals in the model, which we do using importance sampling. Although it is easy to sample sequences from $P_{\rm pre}$ by picking a random recombination scenario, sampling from $P_{\rm post} = QP_{\rm pre}$ is much harder, as the $q_{i;L}$, q_L and q_{VJ} factors introduce complex dependencies between the different features of the recombination scenario. To overcome this issue, we sample a large number M of $(\vec{\tau}, V, J)$ triplets from $P_{\rm pre}(\vec{\tau}, V, J)$, and, when estimating $P_{\rm post}$ expectation values, weight the contribution of each sequence with its $Q(\vec{\tau}, V, J)$ value (this is a particularly simple instance of importance sampling). The generated

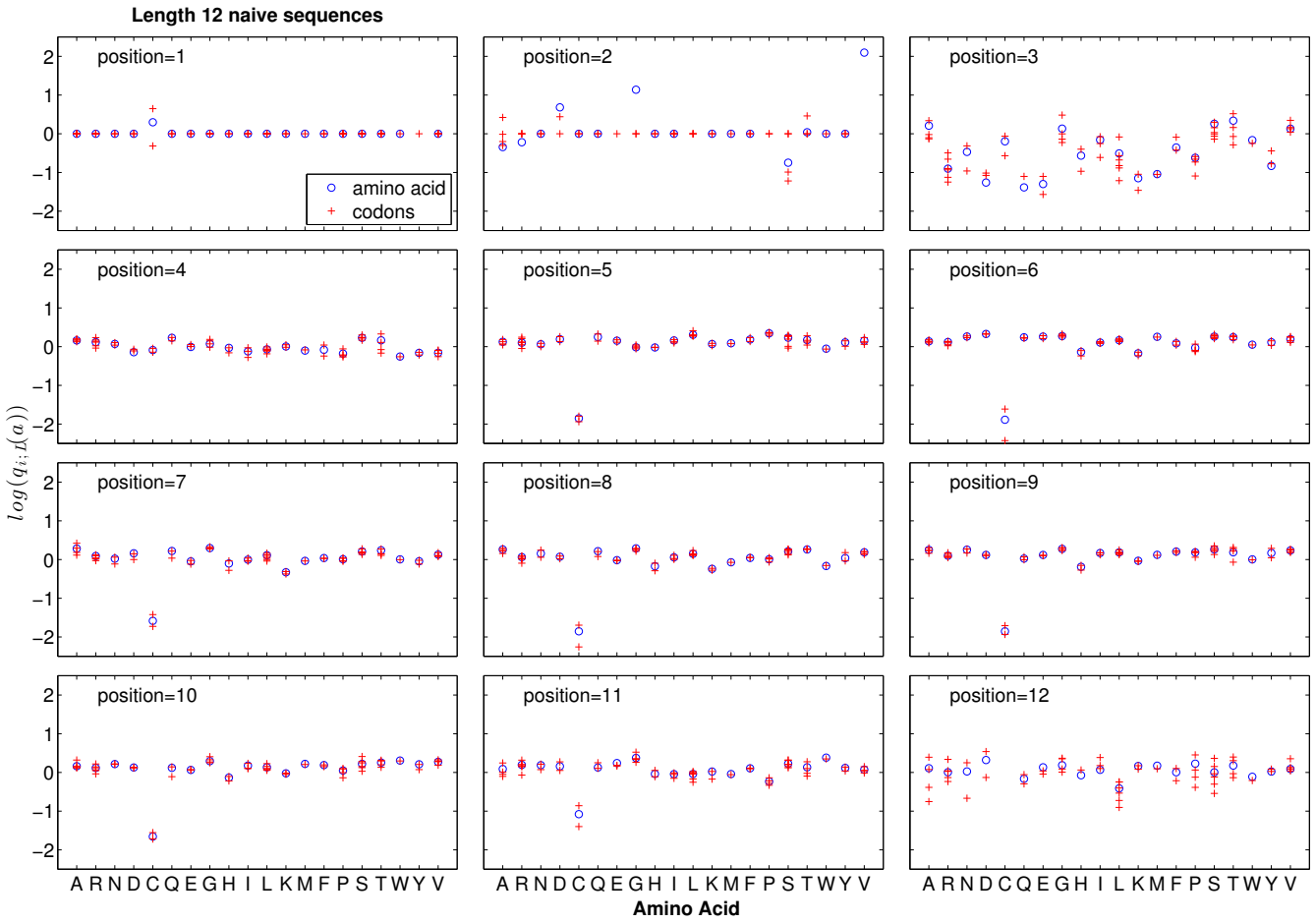

Fig. S 2: The $q_{i;L}(a)$ selection factors learned for codons (red crosses) agree with those learned for amino acids (blue). The $q_{i;L}(a)$ are plotted for each position in the CDR3 region (panels from 1 to 12) for naive CDR3 sequences of length 12, as a function of the amino acids at each position. A given amino acid at a given position can come from different codons, which are marked by multiple crosses at that position. Codons or amino acids for which there was not enough data to infer the selection factors are not represented.

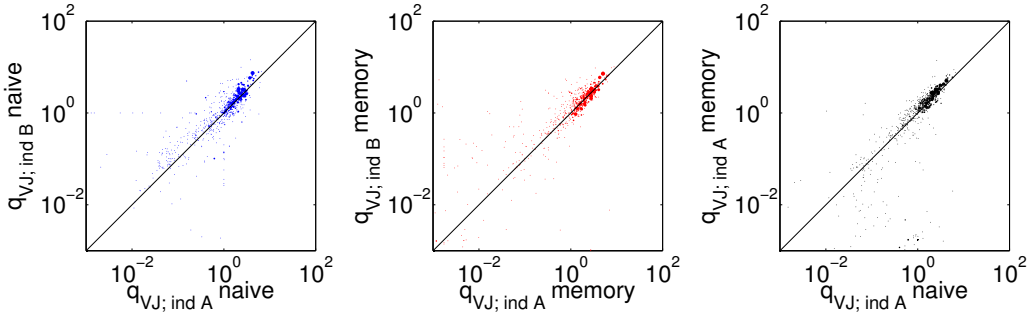

Fig. S 3: The scatter of VJ gene selection factors q_{VJ} between donors A and B for naive (**A**) and memory repertoires (**B**), as well as between the memory and naive repertoires of the same individual (**C**) shows that the memory and naive repertoires are statistically similar to each other and across individuals. See Fig. S4 for the correlation analysis of all individuals and cell types.

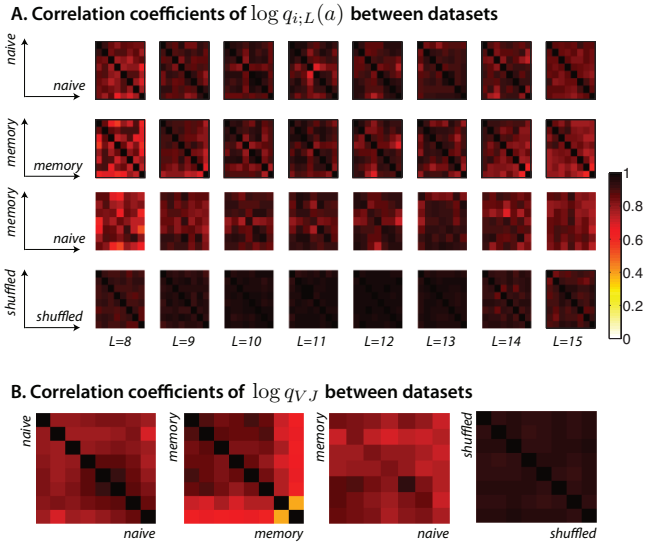

Fig. S 4: Correlation coefficients between selection factors obtained for models learned for different donors and cell type (naive and memory). The compared factors are the amino-acid selection factors $q_{i;L}$ (**A**) and the VJ gene selection factors q_{VJ} (**B**). Each position along the two axes in each plot corresponds to a different individual. The naive dataset of donor 8, and the memory dataset of donor 5 were removed because of too low statistics. In all heat maps, the x and y axes correspond to different donors (1-7;9 for naive, 1-4;6-9 for memory, and 1,2,3,4,6,7,9 for comparison between naive and memory).

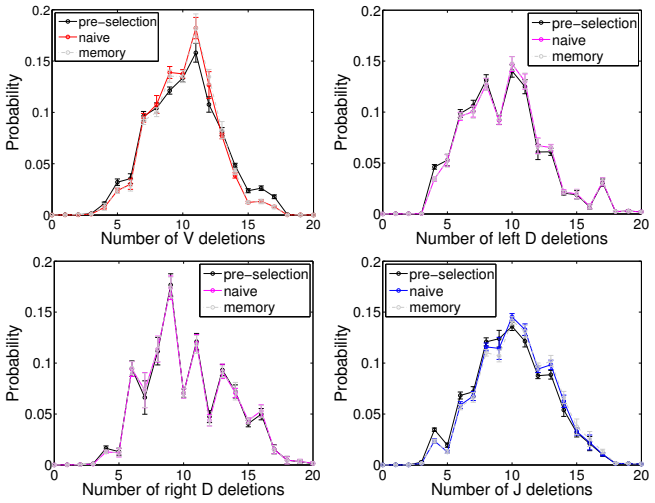

Fig. S 5: The effects of selection on deletion profiles. Distribution of V (**A**), D left-hand side (**B**), D right-hand side (**C**), and J (**D**) deletions in the pre-selected (black lin e), naive (colored line) and memory (gray dashed line) repertoires. Error bars show standard deviation over 9 individuals. Results using 9 separate models learned for each of the individuals. The deletion distributions for the memory repertoire are the same as for the naive repertoire. Selection has a slight effect on favoring distributions with non-extreme deletion values of deletions for V and J deletions, and does not have a significant effect on D deletions.

triplets are denoted by $[(\vec{\rho}^1, V_1, J_1), \ldots, (\vec{\rho}^M, V_M, J_M)]$, and the corresponding reads by $(\vec{\xi}^1, \ldots, \vec{\xi}^M)$ (see Fig. S1 for notations). The marginal probability distribution of lengths, for instance, is estimated by

$$P_{\rm post}(L;Q') \approx \frac{\sum_{b=1}^{M} \delta_{L_b,L} Q'(\vec{\rho}^b, V_b, J_b)}{\sum_{b=1}^{M} Q'(\vec{\rho}^b, V_b, J_b)}. \qquad (19)$$

and similar expressions give estimates of $P_{i;L,\rm post}(a_i;Q')$ and $P_{\rm post}(V,J;Q')$. Since we are optimizing over Q', the sequences $(\vec{\rho}^b, V_b, J_b)$ can be generated once and for all at the beginning of the algorithm. Then the marginal probabilities are updated according to the modified Q' using Eq. 19. Finally, the normalization constant is evaluated by calculating:

$$Z \approx \frac{1}{M} \sum_{b=1}^{M} q_{L_b} q_{V_b J_b} \prod_{i=1}^{L_b} q_{i;L_b}(a_i^b). \qquad (20)$$

so that

$$\sum_{\vec{\tau},V,J} P_{\rm post}(\vec{\tau},V,J) \approx \frac{1}{M} \sum_{b=1}^{M} Q(\vec{\rho}^b, V_b, J_b) = 1. \qquad (21)$$

F. Equivalence with minimum discriminatory information

The principle of minimum discriminatory information is to look for a distribution that reproduces exactly some mean observables of the data, such as position-dependent amino-acid frequencies, while being minimally biased with respect to some background distribution. When the background distribution is uniform, this principle is equivalent to the principle of maximum entropy.

Taking $P_{\rm pre}$ as our background distribution, assume we are looking for the distribution $P_{\rm post}$ that satisfies Eqs. (14)-(16) while minimizing the divergence or relative entropy with respect to $P_{\rm pre}$, defined as:

$$D_{\rm KL}(P_{\rm post} \| P_{\rm pre}) = \sum_{\vec{\tau},V,J} P_{\rm post}(\vec{\tau},V,J) \log \frac{P_{\rm post}(\vec{\tau},V,J)}{P_{\rm pre}(\vec{\tau},V,J)}. \qquad (22)$$

Solving this problem is mathematically equivalent to solving the maximum likelihood problem described above.

We present the values of these minimized $D_{\rm KL}$ divergences for each donor in Table II.

	D_{KL}
Donor 1	0.9646
Donor 2	0.9598
Donor 3	0.9945
Donor 4	0.9664
Donor 5	0.9402
Donor 6	0.9999
Donor 7	1.0195
Donor 8	1.1730
Donor 9	1.0831
Universal Donor	0.9175

Table S II: Kullback-Leibler divergence between the pre and post-selection distributions (see Eq. 22).

IV. INDIVIDUAL, UNIVERSAL AND SHUFFLED DONORS

We partition the data in three different ways to learn the model. First, we learn a distinct model for each donor, and for each of the naive and memory pools. For each donor, we have a distinct $P_{\rm pre}$ learned from the out-of-frame sequences of that donor (although in fact they differ little from donor to donor as discussed in [1]). Second, we pool all the sequences of a given type (naive or memory) from all nine donors together, and learn a "universal" or average model. For this we use a mean $P_{\rm pre}$ averaged over all nine donors, and then learn Q using all sequences. Third, to assess the effect of finite-size sampling in the universal model, we partition the data from all donors into nine random subsamples of equal sizes. This way we can estimate how much variability one should expect from just sampling noise.

V. SELF-CONSISTENCY OF THE MODEL

We check the self-consistency of the assumption that Q has a factorized form by calculating the covariances between the different sequence features (V, J), L and $(a_1, \ldots, a_L)$. We plot the model predictions for these covariances against the same quantities calculated from the data (Fig. 2B of the main text and Fig. S10). We observe a very good agreement, which validates the factorization assumption.

VI. ENTROPY, DISTRIBUTIONS OF $P_{\rm pre}$, $P_{\rm post}$ AND Q

To estimate global statistics, such as entropy, from the model, we draw a large set of sequences $(\vec{\xi}^1, \ldots, \ldots, \vec{\xi}^M)$ from $P_{\rm pre}$, and weight them according to the inferred (normalized) Q values. Specifically, for each generated sequence, we estimate its primitive generation probability by summing over all the possible scenarios that could have given rise to it:

$$P_{\rm pre}(\vec{\xi}^b) = \frac{1}{p_{\rm coding}} \sum_{S \to \xi^b} P_{\rm pre}(S) \tag{23}$$

where $\vec{\xi}^b$ is the full nucleotide sequence, including the CDR3 $\vec{\rho}^b$ as well as the V_b and J_b segments. The entropy (in bits) of the selected sequence repertoire is defined as

$$H[P_{\rm post}] = -\sum_{\vec{\sigma}} P_{\rm post}(\vec{\sigma}) \log_2 P_{\rm post}(\vec{\sigma}) \tag{24}$$

and, to include selection effects, we estimate it by

$$H[P_{\rm post}] \approx -\frac{1}{M} \sum_{b=1}^{M} Q(\vec{\rho}^b, V_b, J_b) \log \left[Q(\vec{\rho}^b, V_b, J_b) P_{\rm pre}(\vec{\xi}^b) \right]. \tag{25}$$

The difference in the entropies of the pre- and post-selection repertoires for each donor (~ 5.5 bits) can be linked to this Kullback-Leibler divergence by the following relation:

$$S_{\rm pre} - S_{\rm post} = D_{\rm KL}(P_{\rm post} \| P_{\rm pre}) + \langle (Q-1) \log_2 P_{\rm pre} \rangle_{\rm pre},$$

where $\langle \cdots \rangle_{\rm pre}$ denotes an average over the pre-selection ensemble $P_{\rm pre}$, approximated by $((\vec{\rho}^1, V_1, J_1), \ldots, (\vec{\rho}^M, V_M, J_M))$.

The Kullback-Leibler divergence (≈ 1 bit, see Table SII) is much smaller than the difference of entropies between the distributions (≈ 4.5 bits, see main text). Eq. 26 allows us to interpret that the main reduction in entropy can be attributed to the fact that selection simply amplifies the characteristics of the pre-selection distribution (as discussed in the "Natural selection anticipates somatic selection" section in the main text). This is evidenced by the strong correlation between Q and $P_{\rm pre}$ (Fig. 5B of the main text) which results in the second term in Eq. 26 being the main contribution to entropy reduction.

The distributions of $P_{\rm pre}$, $P_{\rm post}$ and Q over the selected sequences are determined from the same draw of M sequences from $P_{\rm pre}$, weighted by the normalized selection factors Q. For example the distribution of $\log P_{\rm pre}$ is:

$$\mathbb{P}(\log P_{\rm pre}) \approx \frac{1}{M} \sum_{b=1}^{M} Q(\vec{\rho}^b, V_b, J_b) \delta \left[\log P_{\rm pre} - \log P_{\rm pre}(\vec{\xi}^b) \right]. \tag{26}$$

Marginal distributions over pairs of amino-acids (a_i, a_j) at two positions i and j can also be calculated using the $\vec{\rho}^b$ sequences and weighting them with Q. This can be generalized to arbitrary marginals or statistics.

VII. SHARED SEQUENCES

The number of shared sequences in a subset of donors is counted based on the nucleotide sequences. This empirical number can then be compared to two kinds of

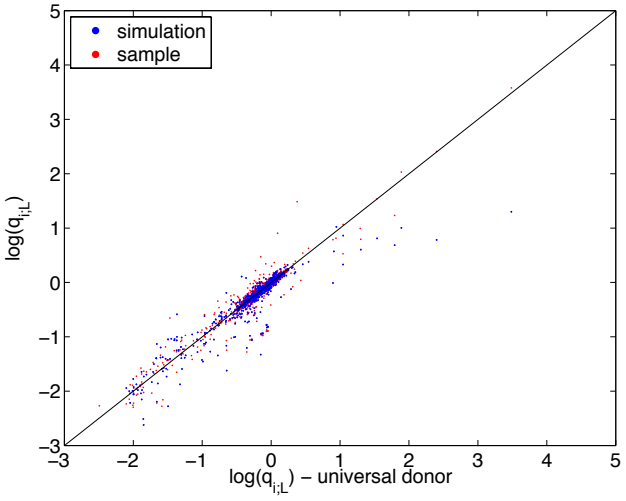

Fig. S 6: The saturation of the $P_{data}(Q)/P_{pre}(Q)$ ratio does not affect the inference of the model. We simulated a dataset from $P_{\rm pre}$ and selected sequences with probability $\min[Q(\vec{\sigma})/7, 1]$. The plot compares the $q_{i;L}(a)$ selection factors directly inferred from data (ordinate) to values inferred from such simulated data (blue dots: simulation). The scatter in these points is compared to the scatter obtained from learning the selection factors using a random subset of the data (red dots: sample). The size of the points denotes the probability $P_{i;l,\rm data}(a)$ in the data repertoire.

theoretical predictions. Either by assuming that the sequences of each donor were generated and selected by a "private" model $P^{(\alpha)}_{\rm post}$, where α denotes the donor, *i.e.* a model inferred from the sequences of donor α; or by assuming that sequences were generated and selected by a "common" or universal model $P^{(u)}_{\rm post}$ inferred from all sequences together. The latter is justified by the fact that differences between private models are small, and could reflect spurious noise that would exaggerate differences between individuals.

If we assume private models, the expected number of shared sequences between donors α and β is:

$$N_\alpha N_\beta \sum_{\vec{\sigma}} P^{(\alpha)}_{\rm post}(\vec{\sigma}) P^{(\beta)}_{\rm post}(\vec{\sigma}), \tag{27}$$

where N_α and N_β are the numbers of sequences in each donor dataset. To estimate that number, we collect sequences that are shared between the generated datasets $\{\vec{\xi}^a\}$ of two (or more) donors, and reweight them by Q:

$$\frac{N_\alpha N_\beta}{M_\alpha M_\beta} \sum_{(\vec{\rho},V,J)\in\alpha\cap\beta} Q^{(\alpha)}(\vec{\rho},V,J) Q^{(\beta)}(\vec{\rho},V,J), \tag{28}$$

where M_α and M_β are the number of generated sequences for each donor model, and where the sum is over the sequences found in the $\{\vec{\xi}^a\}$ dataset of both donors. Similar equations are used for comparing more than two donors.

If we assume a common model, the expected number of shared sequences reads:

$$N_\alpha N_\beta \sum_{\vec{\sigma}} [P^{(u)}_{\rm post}(\vec{\sigma})]^2. \tag{29}$$

This can be estimated by:

$$\frac{N_\alpha N_\beta}{M} \sum_{b=1}^{M} P^{(u)}_{\rm pre}(\vec{\xi}^b) [Q^{(u)}(\vec{\rho}^b, V_b, J_b)]^2, \tag{30}$$

where $\{\vec{\xi}^a\}$ are sequences generated from the mean VDJ recombination model $P^{(u)}_{\rm pre}$. Similarly, the number of shared sequences between a triplet of donors α, β, γ is:

$$\frac{N_\alpha N_\beta N_\gamma}{M} \sum_{b=1}^{M} [P^{(u)}_{\rm pre}(\vec{\xi}^b)]^2 [Q^{(u)}(\vec{\rho}^b, V_b, J_b)]^3, \tag{31}$$

and likewise for quadruplets and more.

The expected numbers of shared sequences calculated above are averages. Their distribution is given by a Poisson distribution of the same mean. We use these Poisson distribution to estimate the error bars in Fig. 6A of the main text and S9A, as well as the distributions in Fig. 6B-C and S9B-C.

If we assume a common model, sequences that are shared between at least n individuals are distributed according to $\propto [P^{(u)}_{\rm post}]^n$. To explore the statistics of these sequences, we take our $\vec{\rho}^b$ sequences generated from $P^{(u)}_{\rm pre}$ and weigh them with $[P^{(u)}_{\rm pre}(\vec{\rho}^b)]^{n-1}[Q^{(u)}(\vec{\rho}^b)]^n$. For example, to estimate the distribution of $\log P_{\rm post}$ in shared sequences as in Fig. 6D of the main text (for pairs), and Fig. S8 (for triplets and quadruplets), we calculate:

$$\mathbb{P}(\log P_{\rm post}) \approx \frac{1}{M} \sum_{b=1}^{M} [P^{(u)}_{\rm pre}(\vec{\xi}^b)]^{n-1} [Q^{(u)}(\vec{\rho}^b, V_b, J_b)]^n \times \delta\left[\log P_{\rm post} - \log P^{(u)}_{\rm post}(\vec{\xi}^b)\right]. \tag{32}$$

Sampling from shared sequences is equivalent to sampling from the high-probability, large deviation regime of the distribution. This statement can be made more physically intuitive by rewriting $P_{\rm post}$ as a Boltzmann distribution $e^{-E/T}$ with $T = 1$ and $E = -\log P_{\rm post}$. Considering sequences observed in at least n donors, is equivalent to sampling from $(1/Z(n))e^{-nE}$ (where $Z(n)$ is a normalisation constant), *i.e.* the Boltzmann distribution with

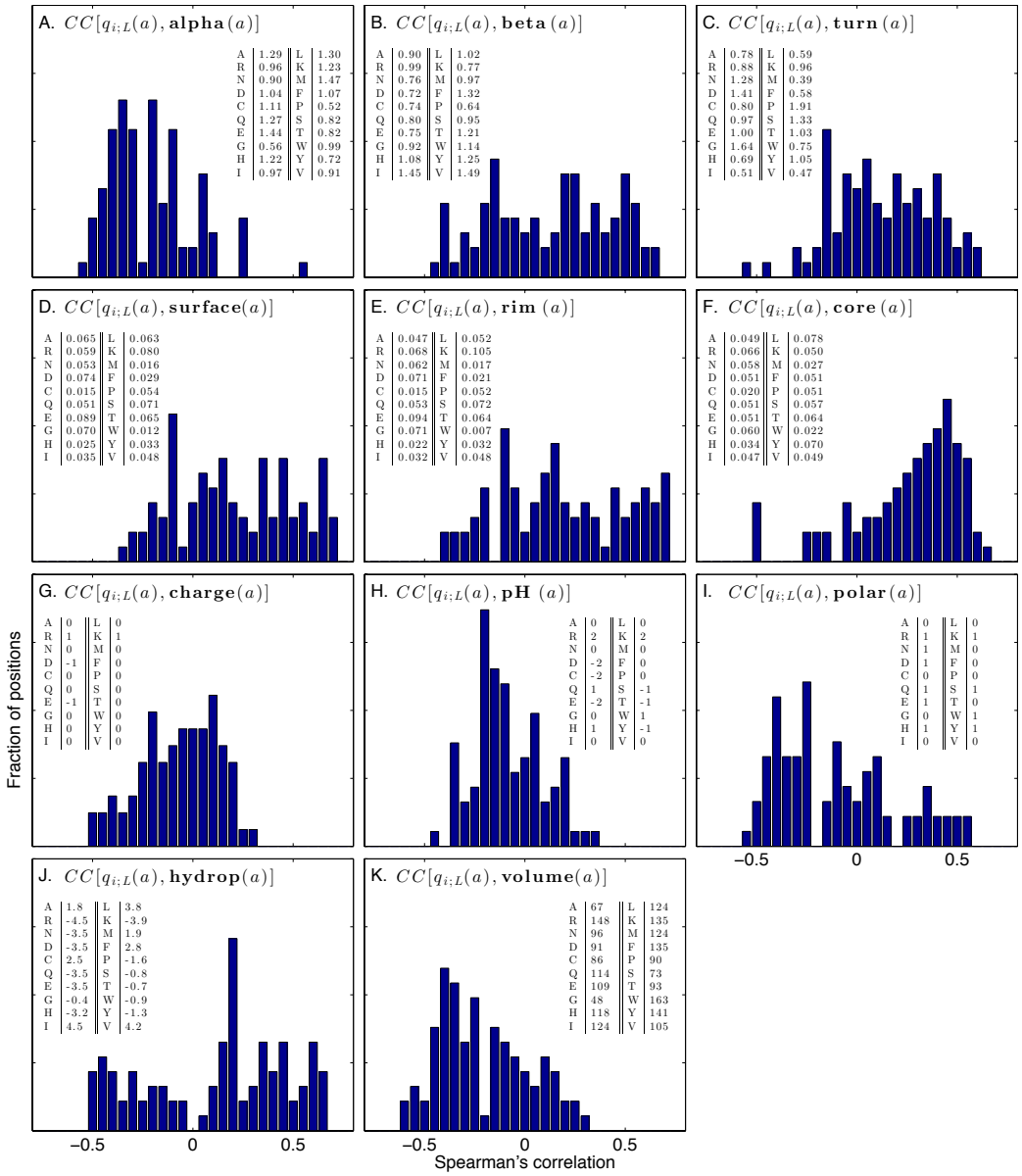

Fig. S 7: Correlation of the $q_{i;L}$ selection factors with several biochemical properties. Each panel shows the histogram, over all positions and lengths, of Spearman's correlation coefficient between the $q_{i;L}(a)$ values for a given amino acid and the biochemical properties of that amino acid. The following biochemical properties are considered (from left to right, top to bottom): preference to appear in alpha helices (**A**), beta sheets (**B**), turns (**C**) (source for (**A-C**): Table 3.3 [7]). Residues that are exposed to solvent in protein-protein complexes (following definitions and data from [8], specifically Fig. S6 in the SI) are divided intothree groups: surface (interface) residues that have unchanged accessibility area when the interaction partner is present (**D**), rim (interface) residues that have changed accessibility area, but no atoms with zero accessibility in the complex (**E**) and core (interface) residues that have changed accessibility area and at least one atom with zero accessibility in the complex (**F**). Rim residues roughly correspond to the periphery of the interface region, and core residues correspond to the center. Finally we plot the basic biochemical amino acid properties (source: `http://en.wikipedia.org/wiki/Amino_acid` and `http://en.wikipedia.org/wiki/Proteinogenic_amino_acid`): charge (**G**), pH (**H**), polarity (**I**), hydrophobicity (**J**) and volume (**K**). For all properties the actual numerical values used to calculate the correlations are listed in the inset tables. We see a positive correlation trend with turns and core residues and a negative correlation trend with the preference of amino acids to appear in alpha helices and volume.

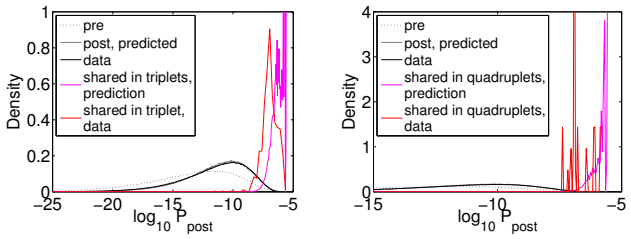

Fig. S 8: Model prediction (magenta) and observed (red) distributions of $P_{\rm post}$ in the naive sequences that are shared between at least three (left) or four (right) donors. The model discrepancy may be attributed to its failure to capture the very highly probable sequences.

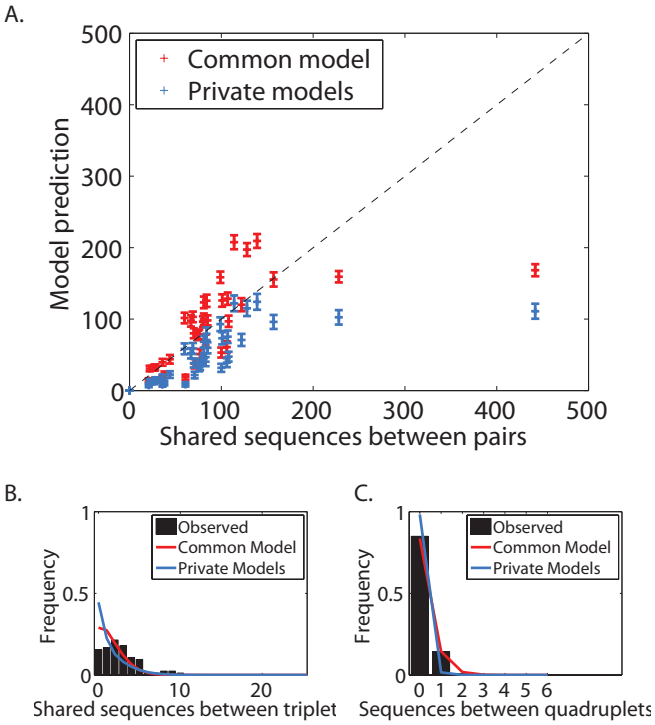

Fig. S 9: Comparison between data and model for the number of shared sequences in the *memory* repertoires, in pairs (**A**), triplets (**B**) and quadruplets (**C**) of individuals.

$T = 1/n$. Sequences shared between more and more individuals correspond to lower and lower temperatures, and thus lower energies and higher probabilities. In the low temperature regime, the roughness of the landscape depicted in Fig. 4C of the main text is starting to become important, and may not be well captured by our model, as suggested by Fig. S8.

VIII. CODON MODEL

It is reasonable to assume that selection acts on the protein structure, at the amino acid level. But each amino acid can be obtained using a number of different codons, which could in principle each have a different selection factor. We checked the robustness of our selection coefficients by learning an alternative model in which selection acts on codons. We present the results of this alternative codon model in Fig. S2 on the example of CDR3 sequences of length 12. We show the $q_{i;L}(a)$ selection factors at each position for each amino acid, and compare them to the selection factors obtained for the codons coding for that amino acid. We see that, especially in the bulk of the CDR3 sequence, selection at the level of codons or amino acids are equivalent, proving the generality of our approach. We observe a very slight correlation between the discrepancies of the selection factors learned for the codon and amino acid models ($\log(q^{\rm codon}_{\rm i;L}(a)) - \log(q^{\rm aa}_{\rm i;L}(a))$) and the G/C content of these codons for amino acids at position 3 from the initial cysteine (correlation coefficient of 0.09 calculated with a p-value of 0.04) and the last position before the J primer (correlation coefficient of 0.1 calculated with a p-value of 0.01).

IX. ADDITIONAL EFFECTS OF SELECTION ON REPERTOIRE PROPERTIES

In the main text we present several repertoire properties, such as insertion profiles and comparisons of the $q_{i;L}(a)$ selection factors between naive and memory repertoires. In Fig. S5 we plot the deletion profiles for V, J and D-lefthand side and D-righthand side deletions, comparing the distributions for the pre-selection, naive and memory repertoires. We note that the deletion profiles for the V and J distributions are more peaked, favoring intermediate deletion values. However the D distributions are little affected by selection. Similarly to the case of insertion distributions shown in the main text in Fig. 3E-F, the naive and memory distributions appear indistinguishable within the error bars.

In Fig. 3A-C of the main text, the selection factors $q_{i;L}(a)$ acting on amino acids are compared between individuals and cell type. Similarly, the selection factors acting on the genes q_{VJ} are statistically indistinguishable between the memory and naive repertoires for one individual, compared to the variability between the naive (or memory) repertoires taken from two sample individuals (see Fig. S3).

To compare the repertoires of individuals as well as the naive and memory repertoires with each other, we consider the correlation coefficients between the selection factors $\log q_{i;L}$, and between the VJ gene selection factor $\log q_{VJ}$, of different individuals (Fig. S4). Correlations between memory and naive repertoires are similar to those between naive-naive or memory-memory repertoires for different individuals; all are a bit smaller than the correlations between the artificial, shuffled sequence datasets, where the discrepancy is entirely attributable to statistical noise. These observations lead us to the conclusion that at this level of description, the selection

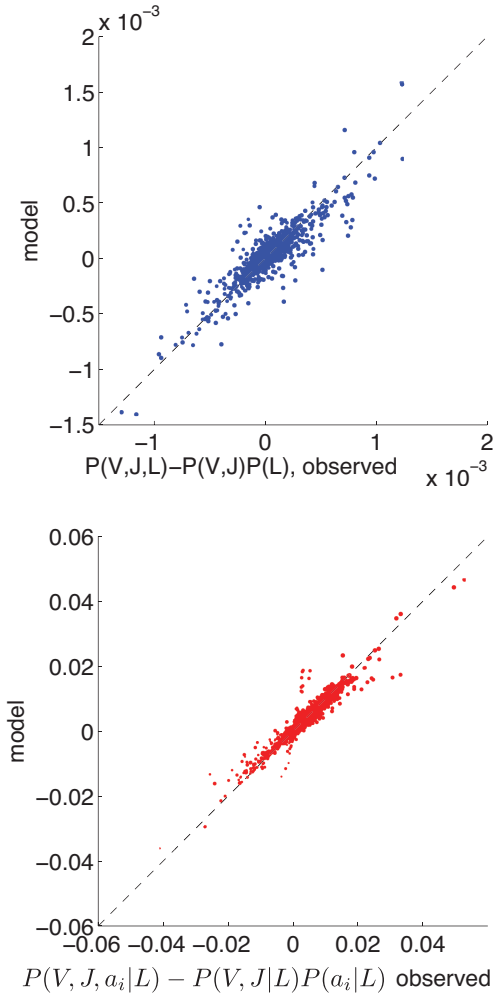

Fig. S 10: Comparison of the covariances between the model and data between (V, J) and L (top) and (V, J) and a_i given L on the other hand (bottom). The model, which assumes that the selection factors factorize, predicts the observed covariances well, thus validating this factorization assumption.

processes that shape the memory and naive repertoires are very similar with each other and between different individuals.

We also calculated the Jensen-Shannon divergence $\mathrm{JS}(P^{(\alpha)}_{\mathrm{post}}, P^{(\beta)}_{\mathrm{post}})$ between individual models, where the JS divergence between two distributions P and Q is defined as:

$$JS(P,Q) = \frac{1}{2}\sum_x P(x)\log\frac{P(x)}{M(x)} + \frac{1}{2}\sum_x Q(x)\log\frac{Q(x)}{M(x)} \tag{33}$$

with $M(x) = \frac{1}{2}[P(x)+Q(x)]$. This measure is preferable to the Kullback-Leibler divergence because it is symmetric. The values of this divergence for all pairs of donors are shown in Table SIII.

	1	2	3	4	5	6	7	8
2	0.02							
3	0.11	0.11						
4	0.03	0.03	0.10					
5	0.07	0.07	0.13	0.07				
6	0.03	0.03	0.10	0.04	0.05			
7	0.03	0.03	0.12	0.03	0.08	0.04		
8	0.08	0.07	0.14	0.07	0.12	0.07	0.08	
9	0.07	0.08	0.15	0.07	0.11	0.07	0.06	0.13

Table S III: Jensen-Shannon divergence between the P_{post} distributions for each donor.

X. SATURATION OF THE SELECTION RATIO

We consider distributions of the selection factor Q in the pre-selection ensemble $P_{\mathrm{pre}}(Q)$, in the post-selection ensemble according to the model $P_{\mathrm{post}}(Q)$, and in the actual data sequences $P_{\mathrm{data}}(Q)$. These three distributions are formally defined as:

$$P_{\mathrm{pre}}(Q) = \frac{1}{M}\sum_{b=1}^{M}\delta\left[Q - Q(\vec{\rho}^{\,b}, V_b, J_b)\right]. \tag{34}$$

$$P_{\mathrm{post}}(Q) = \frac{1}{M}\sum_{b=1}^{M} Q(\vec{\rho}^{\,b}, V_b, J_b)\delta\left[Q - Q(\vec{\rho}^{\,b}, V_b, J_b)\right] \tag{35}$$

$$= QP_{\mathrm{pre}}(Q). \tag{36}$$

$$P_{\mathrm{data}}(Q) = \frac{1}{N}\sum_{a=1}^{N}\sum_{V_a,J_a} P_{\mathrm{post}}(V_a, J_a|\vec{\sigma}^a) \times\delta\left[Q - Q(\vec{\tau}^a, V_a, J_a)\right] \tag{37}$$

As can be seen in Fig. 4 of the main text, the ratio of the distribution of global selection factors $P_{\mathrm{data}}(Q)/P_{\mathrm{pre}}(Q)$ saturates for large values of Q. To make sure that this saturation does not impair our ability to correctly infer the selection factors, we simulated a dataset from P_{pre} and selected sequences with probability $\min[Q(\vec{\sigma})/7, 1]$ to mimic the effects of this plateau. We then inferred the selection coefficients for this artificial dataset. We see that the saturation does not affect our ability to correctly infer the selection coefficients (Fig. S6) and the variability in the inferred $q_{i;L}(a)$ selection factors is of the same order as from using random subsamples of the original data.

We also checked that this saturation did not affect much the prediction for the number of shared sequences, by repeating the procedure replacing Q by $\max(Q, 7)$ in Sec. VII. For example, $\sum_{\vec{\sigma}}[P^{(u)}_{\mathrm{post}}(\vec{\sigma})]^2$, the probability for any two sequences to be the same, only decreased by 2%, $\sum_{\vec{\sigma}}[P^{(u)}_{\mathrm{post}}(\vec{\sigma})]^3$, the probability for any three sequences to be the same, by 6%, and $\sum_{\vec{\sigma}}[P^{(u)}_{\mathrm{post}}(\vec{\sigma})]^4$ by 8%.

XI. BIOCHEMICAL CORRELATIONS

To check for correlations of our inferred $q_{i;L}(a)$ selection factors with known biochemical properties, we calculated Spearman's coefficient between the selection factors and a number of standard quantities (see Fig. S7 for the full list). We find that the selection factors do not correlate well with most standard properties, such as charge, hydrophobicity and polarity. However we do find a trend of positive correlation with amino acids that are likely to appear in turns (Fig. S7 C) and ones that have been identified as those that make the core of the interface in a protein-protein complexes (Fig. S7 F) [8]. We find a trend of negative correlations with amino acids that have large volume (Fig. S7 K) and are likely to appear in alpha helices (Fig. S7 A). These observations are consistent with the fact that structurally CDR3 regions form loops and bulky amino acids as well as stabilizing alpha helix-like interactions would interfere with this structure. Core amino acids are at the center of the interface and are known to be the main contributors to interface recognition and affinity. On the other hand interface rim and non-interface (surface) residues, which are both in touch to various degrees with the solvent and are not crucial interface forming elements, show similar non-distinctive correlation patterns.

[1] Murugan A, Mora T, Walczak AM, Callan CG (2012) Statistical inference of the generation probability of t-cell receptors from sequence repertoires. *Proceedings of the National Academy of Sciences* 109:16161–16166.

[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.

[4] 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.

[5] Dempster A, Laird N, Rubin D (1977) Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B* 39:1–38.

[6] McLachlan GJ, Krishnan T (2008) *The EM Algorithm and Extensions (Wiley Series in Probability and Statistics)* (Wiley-Interscience), 2 edition.

[7] Stryer L, Berg JM, Tymoczko JL (2002) *Biochemistry, 5th edition* (W.H. Freeman & Co Ltd) Vol. 5th edition.

[8] Martin J, Lavery R (2012) Arbitrary protein-protein docking targets biologically relevant interfaces. *BMC Biophysics* 1:7.