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).

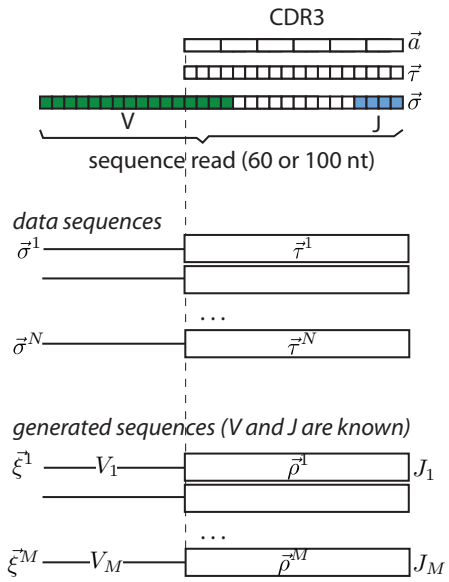

Fig. S 1: Summary of the notations used in this paper for the sequences. The CDR3 region is defined from the conserved cysteine around the end of the V segment to the last amino-acid in the read, leaving two amino acids to the conserved phenylalanine in the J segment. The nucleotides in the read are 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

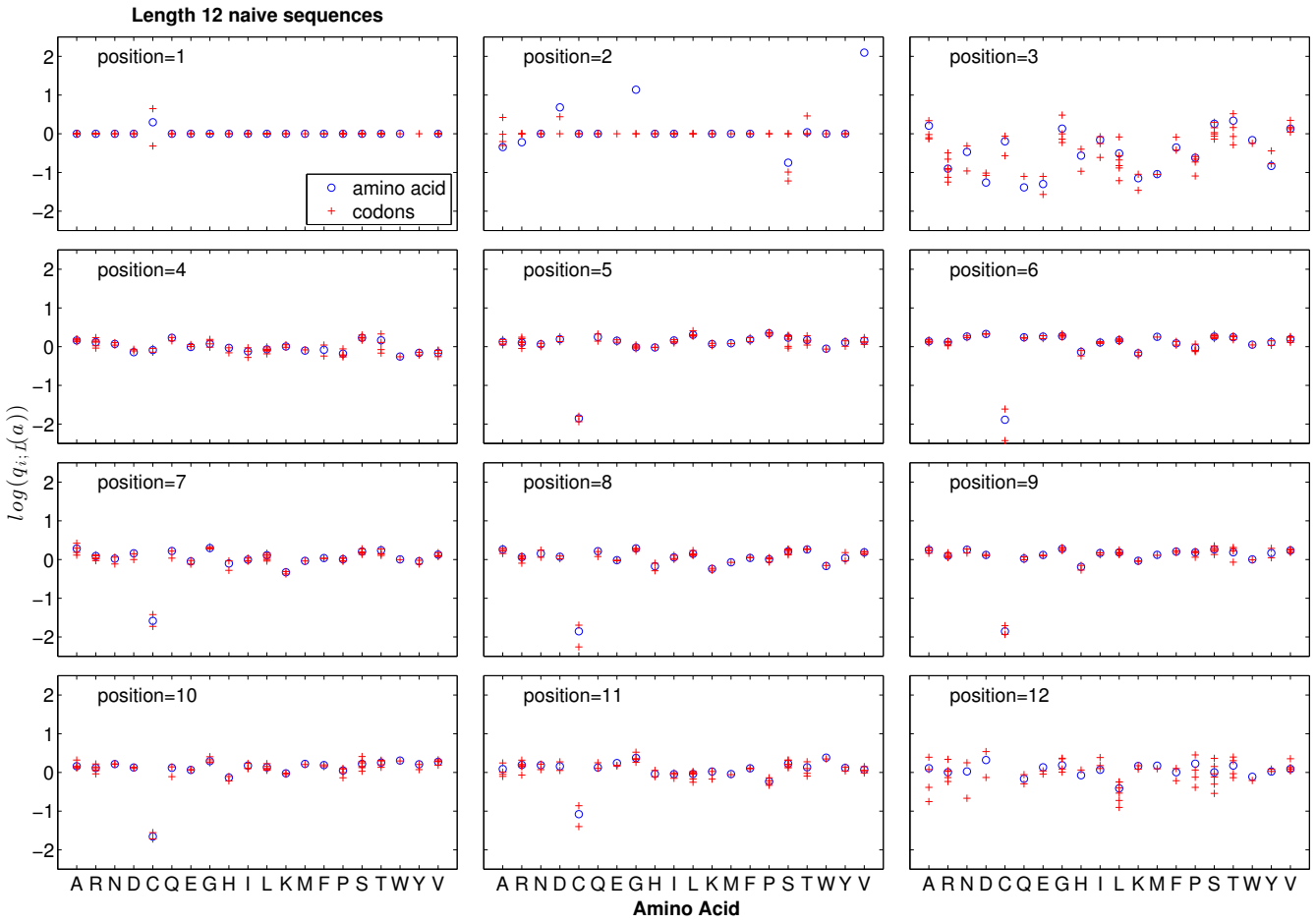

Fig. S 2: The $q_{i;L}(a)$ selection factors learned for codons (red crosses) agree with those learned for amino acids (blue). The $q_{i;L}(a)$ are plotted for each position in the CDR3 region (panels from 1 to 12) for naive CDR3 sequences of length 12, as a function of the amino acids at each position. 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.

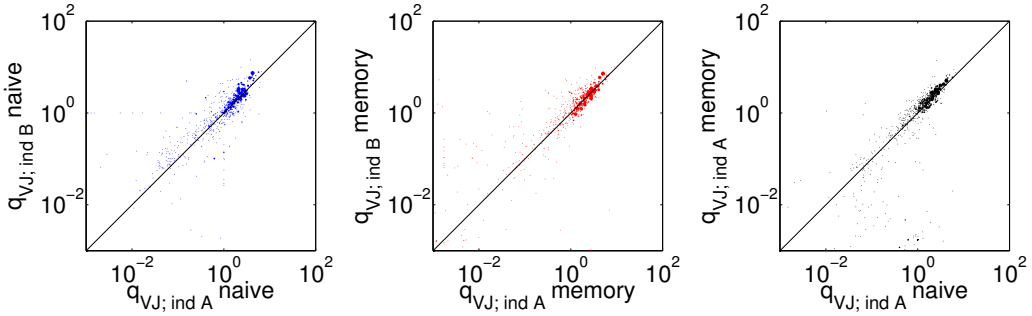

Fig. S 3: The scatter of VJ gene selection factors q_{VJ} between donors A and B for naive (**A**) and memory repertoires (**B**), as well as between the memory and naive repertoires of the same individual (**C**) shows that the memory and naive repertoires are statistically similar to each other and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

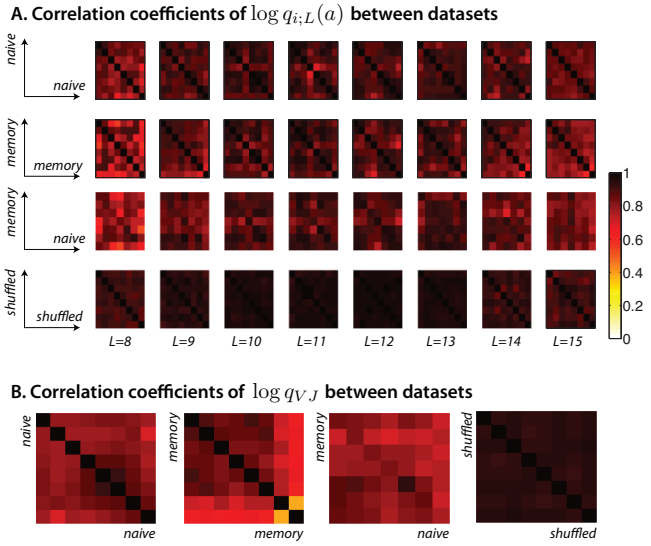

Fig. S 4: Correlation coefficients between selection factors obtained for models learned for different donors and cell type (naive and memory). The compared factors are the amino-acid selection factors $q_{i;L}$ (**A**) and the VJ gene selection factors q_{VJ} (**B**). Each position along the two axes 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).

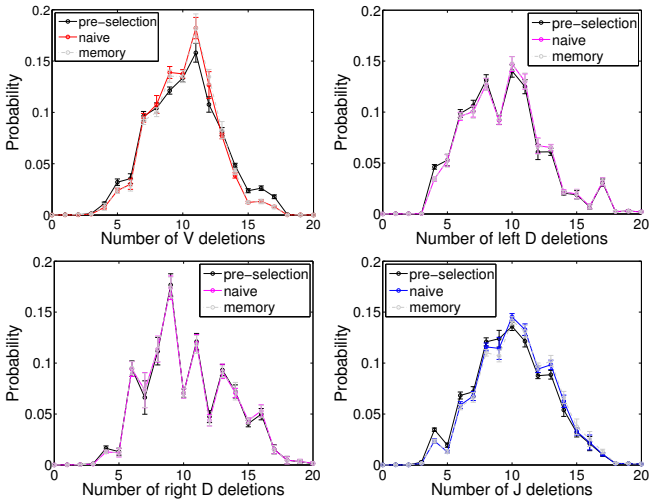

Fig. S 5: The effects of selection on deletion profiles. Distribution of V (**A**), D left-hand side (**B**), D right-hand side (**C**), and J (**D**) deletions in the pre-selected (black lin e), naive (colored line) and memory (gray dashed line) repertoires. Error bars show standard deviation over 9 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

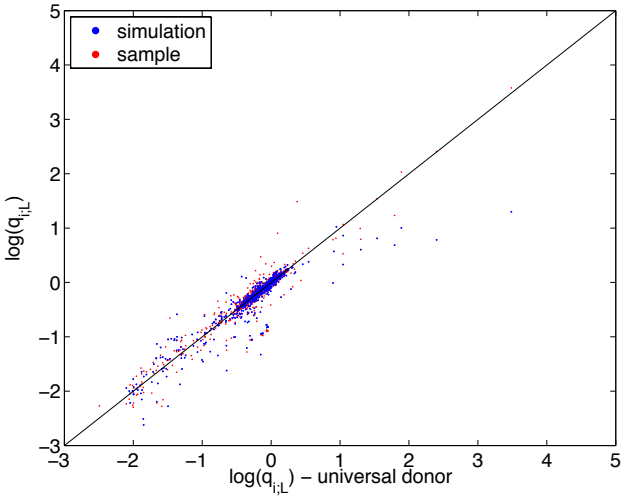

Fig. S 6: The saturation of the $P_{data}(Q)/P_{pre}(Q)$ ratio does not affect the inference of the model. We simulated a dataset from $P_{\rm pre}$ and selected sequences with probability $\min[Q(\vec{\sigma})/7, 1]$. The plot compares the $q_{i;L}(a)$ selection factors directly inferred from data (ordinate) 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

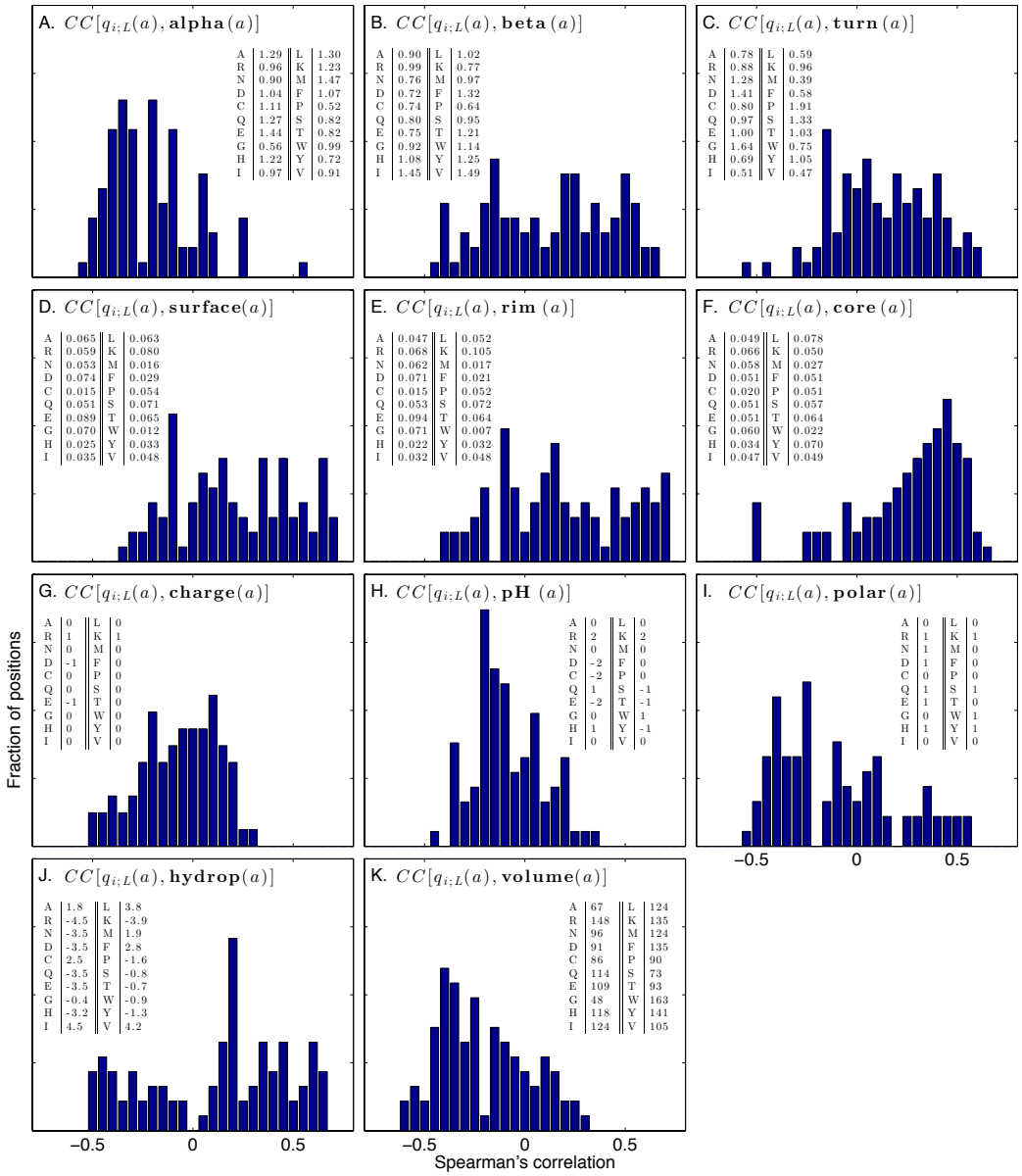

Fig. S 7: Correlation of the $q_{i;L}$ selection factors with several biochemical properties. Each panel shows the histogram, over all positions and lengths, of Spearman's correlation coefficient between the $q_{i;L}(a)$ values for a given amino acid and the biochemical properties of that amino acid. 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.

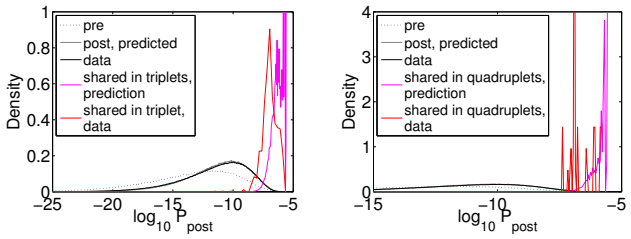

Fig. S 8: Model prediction (magenta) and observed (red) distributions of $P_{\rm post}$ in the naive sequences that are shared between at least three (left) or four (right) donors. The model discrepancy may be attributed to its failure to capture the very highly probable sequences.

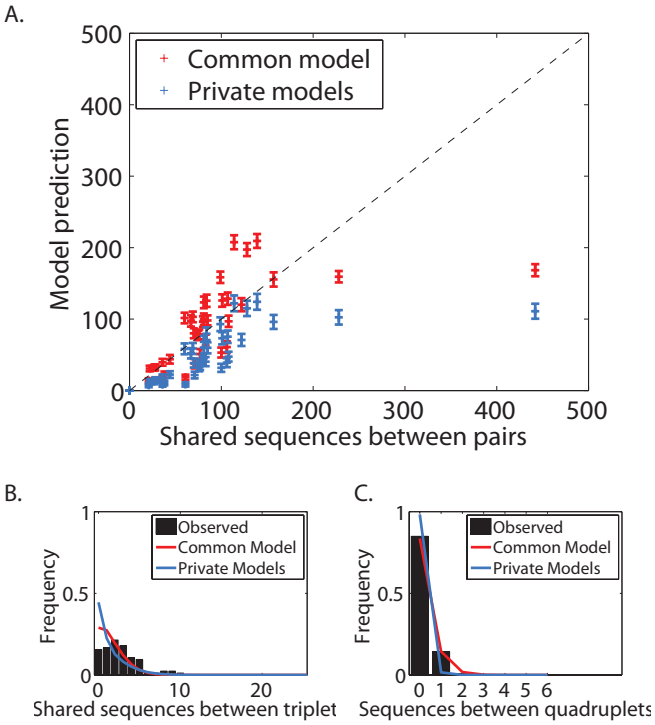

Fig. S 9: Comparison between data and model for the number of shared sequences in the *memory* repertoires, in pairs (**A**), triplets (**B**) and quadruplets (**C**) of individuals.

$T = 1/n$. 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

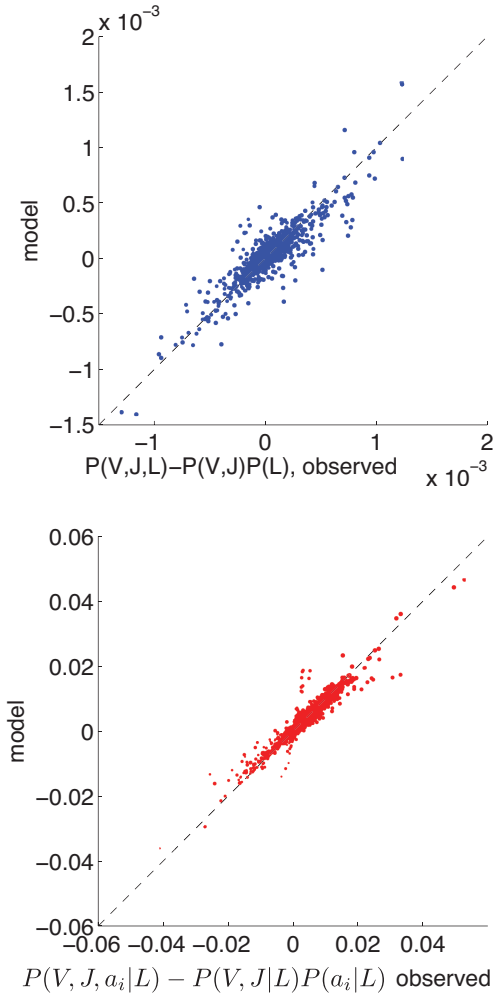

Fig. S 10: Comparison of the covariances between the model and data between (V, J) and L (top) and (V, J) and a_i given L on the other hand (bottom). The model, which assumes that the selection factors factorize, predicts the observed covariances well, thus validating this factorization assumption.

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.