

Conserved variation: identifying patterns of stability and variability in BCR and TCR V genes with different diversity and richness metrics

This article has been downloaded from IOPscience. Please scroll down to see the full text article.

2013 Phys. Biol. 10 035005

(<http://iopscience.iop.org/1478-3975/10/3/035005>)

View [the table of contents for this issue](#), or go to the [journal homepage](#) for more

Download details:

IP Address: 129.25.131.235

The article was downloaded on 11/06/2013 at 05:30

Please note that [terms and conditions apply](#).

Conserved variation: identifying patterns of stability and variability in BCR and TCR V genes with different diversity and richness metrics

Gregory W Schwartz and Uri Hershberg

School of Biomedical Engineering, Science and Health Systems Drexel University, Philadelphia, PA, USA

E-mail: uri.hershberg@drexel.edu

Received 8 October 2012


Accepted for publication 26 March 2013

Published 4 June 2013

Online at stacks.iop.org/PhysBio/10/035005

Abstract

The immune system can detect most invading pathogens. The potential for detection of pathogens is dependent on the somatic diversity of the immune repertoires. While it is known that this somatic diversity is carefully generated, it is unclear how the diversity is distributed in the different genes encoding receptors of immune cells. Utilizing different metrics for richness and diversity at the level of small sequence fragments, we present here an analysis of the entire known human germline repertoire as represented by the sequences from the ImMunoGeneTics database of immune receptors. We have developed a fragment sequence quantification analysis to track variation of repertoires with different degrees of precision. Somatic diversity has previously been functionally characterized mostly by division of the V gene sequences into the more conserved and invariant framework (FR) of the receptor and more varied complementarity determining regions (CDR), that interact with the antigen. We find that CDR and FR can be explicitly identified with our sequence fragment diversity quantification technique. In terms of diversity, CDR and FR are especially distinct in B cell V genes. T cell V genes show less of the CDR/FR periodicity but are more diverse overall. Our analysis further shows that there are other areas of diversity outside the CDR and FR that are found widely dispersed in T cell receptor V genes and more tightly focused in FR1 and FR3 in the B cell receptor V genes. The diversity we observe is not dependent on allelic differences nor is this diversity segregated by individual V gene families. We would thus expect that each individual exhibit a diversity equivalent to that of the entire potential repertoire.

 Online supplementary data available from stacks.iop.org/PhysBio/10/035005/mmedia

1. Introduction

To design vaccines and understand how immune repertoires react to disease, we would like to identify the region of the receptor that interacts with antigens. Even before immune receptors had been structurally characterized, it was hypothesized that areas of greater amino acid sequence variability and diversity could be considered surrogate indicators of antigen interaction points in the receptor [1]. Although this general connection has since been verified [2],

the number of existing structures is still very small, with ~300 B cell receptor (BCR) and ~50 T cell receptor (TCR) structures found in the ImMunoGeneTics (IMGT) 3D structure database [3]. This number is inadequate compared to the 10^{15} possible receptor types immune repertoires can generate [4, 5] and the innumerable number of antigens they can interact with. Therefore, the question of which regions in the receptor interact with antigen is still an open one and clear identification of regions of greater amino acid sequence variability is still of paramount importance.

TCRs and BCRs are each composed of a heavy chain and a light chain. BCRs are made of one heavy chain (V_H) and one light chain (either a V_κ or V_λ) while TCRs are made of one V_β (heavy) chain and one V_α (light) chain. These chains are divided into the complementarity-determining region (CDR) that is thought to comprise most of the positions that interact and bind to the antigen and the framework (FR) region, which is the backbone of the receptor [6]. Analysis of the V genes that encode the different heavy and light chains in BCRs and TCRs has shown that indeed the CDR is more varied than the FR in its amino acid makeup [1, 7, 8]. However, there were limitations both to their methods and to the sequences existing at that time for analysis. Specifically, previously only a small and partial set of germlines V_H , V_κ and the Vs of TCR were identified and could be analyzed [7, 8]. Furthermore, due to the limited amounts of sequence data, previous analysis was done on both germline and mutant data, without verifying that sequences were not clonally related. We here perform a full analysis of the diversity of all heavy and light chain V genes in BCRs and TCRs at the germline level. To do so we measure and compare the diversity of amino acid and short amino acid fragments across the different V gene sequences. Our definition of diversity is what Jost and others, interested in characterizing species diversity, have called ‘true’ diversity [9, 10].

Variability and diversity are not identical terms. Variability is a measure of certainty that things that are always of one type. Diversity is an indication of abundance of types. While the two are related and measures of variability may often be considered indices of diversity they are not equivalent. The original metric, suggested by Wu and Kabat to measure position variability in BCR genes was [1]

$$\frac{\text{Number of different AA found at a specific position}}{\text{Frequency of the most abundant AA}}.$$

This metric is to some degree a hybrid combining influences of both variability and diversity. It is maximal when most varied, but also grows substantially when things are more diverse. Stewart *et al* rightly criticized this metric as being accurate only for identifying the most variable positions due to its unclear distribution of values [8]. To show a clearer picture of a wider range of values they chose to use the Shannon entropy index. However, although Shannon entropy is often termed a diversity index, this entropy measures the ability to predict the identity of a species (in this case an amino acid) from the prediction of the rest of the sample (sequences). This property, although related, is not directly equivalent to measuring diversity.

The confusion regarding the measuring and comparison of diversity and diversity indices is widespread in biology. Recently, several authors have attempted to clarify the issue [9, 10]. They suggest that rather than using diversity indices (which indirectly describe diversity), we should simply measure ‘true’ diversity, which they define as the effective number of species in a sample. Although related to diversity, entropy is a distinct facet of diversity which describes such properties in a less intuitive way that is not conceptually proportional to the samples represented. For instance, we would expect that a pool of equality abundant species be

half as diverse as a pool of twice as many equally abundant species—a linear property not held by entropy [9]. While entropy is informative in sequence analysis, for instance if we wish to gauge the importance of a position to viability, in this case, where function actually depends on diversity, we should simply measure true diversity directly [10]. In measuring diversity, it is important to consider the order of the diversity. When measuring the effective number of species, the order determines the extent to which we are influenced by sample abundances of the different species. Diversity with an order of 0 considers all species equally regardless of their abundance and is equivalent to richness. Diversities with an order lower than 1 give effective diversity values that disproportionately consider rare species while those with a value above 1 disproportionately consider common species. When the order is exactly 1 the effective diversity is calculated without bias [9]. Thus by calculating the effective diversity for a population at several orders, we can determine the effect of common and rare species by the difference between diversities at different orders. This manipulation of the ‘orders’ of diversity is essential in the analysis of over- or underrepresented receptors.

By measuring diversity of a repertoire of germline genes that include all known alleles and gene families, we can better characterize the relative impact of these different levels of genetic similarity on the repertoire’s potential for diversity. We find that, as previously shown in partial data sets, the CDR is more diverse than the FR. The distinction between CDR and FR is clearly apparent in germline BCR V genes but less clear in germline TCR V genes, who exhibit more sequence positions with high diversity across the whole sequence. As a result, even though ranges of amino acid diversity are similar for all the V genes, TCR V genes are more diverse overall. However, contrary to previous findings [8], we still observed the presence of CDR in TCRs, although they were less pronounced than the those within the BCRs. Even as CDR and FR were characterized in the analysis of the first 11 sequenced V_H genes, it was observed that there were highly variable positions in the FR and especially FR3 [7]. We show here, in our more comprehensive analysis, that this is a phenomena found in most BCR V genes. We found that some of these positions use only hydrophilic amino acids to generate their diversity while others are more promiscuous. A bias toward hydrophilicity is indicative of poly-reactivity [11]. We would therefore suggest that some of the diverse positions outside of the CDR are also participating in antigen interaction while others either influence binding indirectly or are simply less rigorously controlled structural positions.

2. Methods and materials

2.1. Sequences analyzed

We analyzed the amino acid sequences of *Homo sapiens* germline BCR V_H , V_κ and V_λ genes as well as TCR V_β and V_α genes. Germline sequences were obtained from the IMGT database [12]. Non-functional, partial, and duplicate sequences were filtered out of the analysis. Finally, between 48 to 155 alleles, from 33 to 49 genes, were studied for every

Table 1. Sequences obtained from the IMGT database for *Homo sapiens* germline V regions.

Repertoire	Genes	Alleles
V_H	49	155
V_κ	38	48
V_λ	33	60
V_α	45	86
V_β	47	96

type of V gene (see table 1). All sequences were numbered according to the IMGT unique numbering system based off of the universal alignment provided by IMGT for over 5 000 sequences defined by CDR and FR positions, structural data, and hypervariable loops [12]. A sliding window was applied to each set of sequences in a repertoire, dividing the gene sequences into fragments defined by a starting position and a window length. IMGT alignment gaps were removed from the fragments, however the IMGT unique position numbering was conserved.

2.2. Diversity measures

For each V gene repertoire, the diversity for the collection of fragments at every position and window length was quantified for three orders of diversity. We will describe here the results of our analysis for window lengths 1 and 3, the maximum window for which we have sufficient data as determined by the rarefaction curve analysis—see below in section 2.3. Both forward and reverse sliding windows were calculated for the amino acid windows longer than 1. A window length of 1 measures the diversity of amino acid at each position. At each window length w and each position p , the number of fragments in the collection is $N_{w,p}$ and the richness of fragments in the collection is $R_{w,p}$.

The measure of diversity used for these collections of fragments was ‘true’ diversity ${}^qD_{w,p}$, where

$${}^qD_{w,p} \equiv \left(\sum_{i=1}^{R_{w,p}} p_{i,w,p}^q \right)^{(1/1-q)} \quad (1)$$

and q is the order of diversity and p_i is the frequency of fragment i [9]. At $q = 1$, (1) does not exist, however the limit as q approaches 1 is

$${}^1D_{w,p} \equiv \exp\left(-\sum_{i=1}^{R_{w,p}} p_{i,w,p} \ln p_{i,w,p}\right). \quad (2)$$

The analysis was run for three orders of diversity: $q = 0$ (resulting in the richness, or the number of different fragments), $q = 1$ (exp(Shannon Entropy [13] with base e)), and $q = 2$ (1/(1 –Gini–Simpson Index [14])). These orders of diversity, sometimes referred to as ‘Hill numbers’ [15], are only dependent on q and the frequency of each fragment [9]. As stated in the introduction, if q , or the order of diversity/Hill number, equals 1, we calculate the effective diversity without giving added weight to rare or abundant species. An order less than 1 gives greater weight to rare species and an order greater than 1 gives greater weight to abundant ones. All three scenarios are necessary to reveal properties of the abundances in each pool of fragments.

In order to compare the CDR, FR, and overall V gene diversities, we calculated the weighted means for diversity over all positions of each chain at a certain window length, where the weights corresponded to the number of fragments at that position. We also did this with respect to CDR and FR defined by a modified IMGT numbering (FR1: 1–24, CDR1: 25–40, FR2: 41–53, CDR2: 54–68, FR3: 69–104, CDR3: 105–111) [16, 12, 17]. A non paired two way t -test was used to identify when genes and regions were significantly more or less diverse than each other.

To refrain from any type of sample bias we excluded from our calculations any position that had less than the minimum number of sequences for a V gene (48, corresponding to the number of sequences in V_κ), while maintaining IMGT numbering so as to be able to compare positions. This left us with 83 positions of which 18 were in CDR. For the weighted means analyses we included all positions with any gene representation and relied on the weighted nature of this test to compensate.

To explore the effects of allelic diversity within a gene, the diversity of each gene was calculated from its alleles. Using this analysis, the average for all genes was found for each position and window length. Additionally, to negate the effects of genes with no diversity at a given position, the diversities of positions were also averaged only for those genes with greater than 1 diversity, meaning that there must be at least two types of fragments at that position.

2.3. Window determination

To refrain from weighting our analysis by any of the different alleles or genes, we limited ourselves to a single copy for each known allele. For this reason, we could envision that our diversity results could be influenced by our sample size. To verify that this was not the case and to analyze only sequence fragment sizes whose diversity was well covered by our sequence sample, we performed a rarefaction curve analysis. Rarefaction curves are often used to verify that sample sizes are big enough to identify different levels of diversity. If at each window length w and each position p , the number of fragments in the collection is $N_{w,p}$ and the richness of in the collection is $R_{w,p}$. Then rarefaction curves [18], for each window and position with incrementing $n \in \{1, \dots, N_{w,p}\}$ subsamples, $E[R_{w,p_n}]$ were generated with

$$E[R_{w,p_n}] = R_{w,p} - \binom{N_{w,p}}{n}^{-1} \sum_{i=1}^{R_{w,p}} \binom{N_{w,p} - N_{w,p_i}}{n}. \quad (3)$$

Sampling is deemed sufficient if the rarefaction curve plateaus and, despite greater sample sizes, no greater abundance is found. In our analysis to determine a plateau, we measured the fraction of the curve that had a ‘horizontal line’, defined by a continuous set of $E[R_{w,p_n}]$ that stays within 95% of the value $E[R_{w,p_n}]$. For instance, if there are 100 sequences represented in a certain pool and the rarefaction curve shows that $E[R_{w,p_n}] \geq 95$ for $80 < n \leq 100$, then we say that the plateau composes 20% of the rarefaction curve. Going by the rarefaction curves in our V gene data, we find that for fragments of amino acids that are 1 and 3 residues long

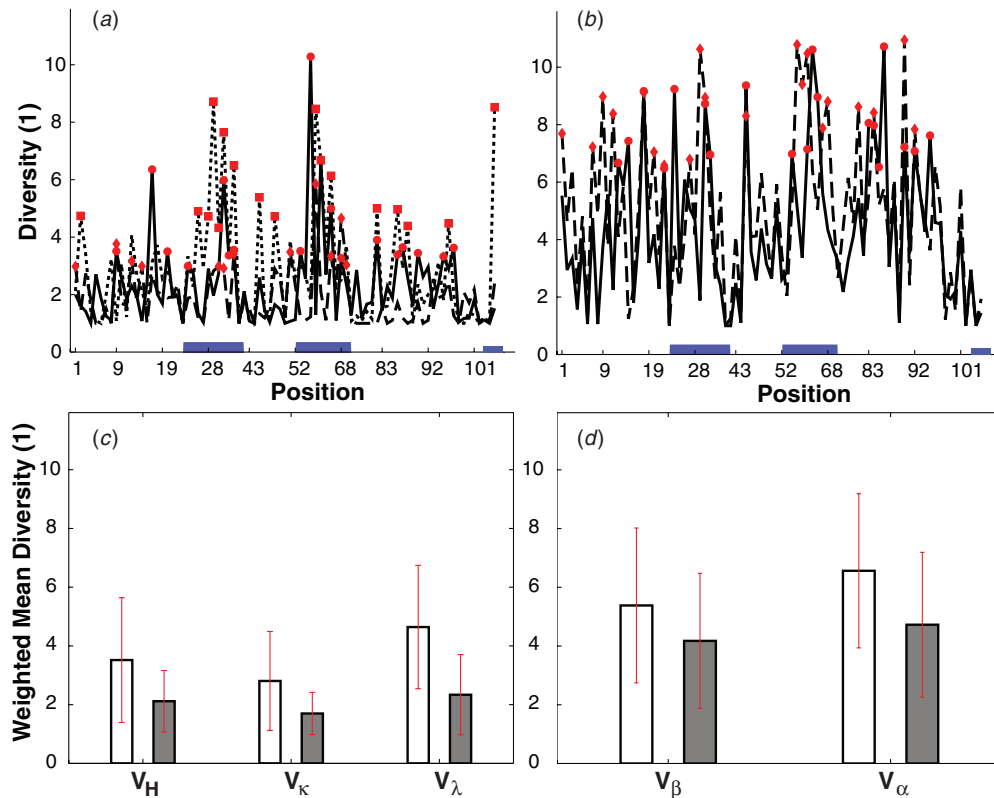


Figure 1. The diversity of order 1 with a window length of 1. (a) The diversity at each residue for V_H (—, \circ), V_κ (---, \diamond), and V_λ (\cdots , \square). The positions in the upper quartile are indicated in red. CDRs are highlighted on the position axis by blue bars, where CDR1 is the first bar, CDR2 is the second bar, and so on. (b) The diversity at each residue for V_β (—, \circ) and V_α (---, \diamond). (c) The overall weighted mean for order 1 diversity of V_H (leftmost set), V_κ (center set), and V_λ (rightmost set). The left bar in each set represents the weighted mean for all CDRs (white) while the right bar represents the weighted mean for the FR region (gray). The error bars represent one weighted standard deviation above and one below the mean. (d) The overall weighted mean for diversities of V_β (left set) and V_α (right set).

at all positions, the curves plateau and describe 95% of the richness in the samples after at most $\sim 86\%$ of the curve for 3 residues (supplementary table 1 and supplementary figures 1–3 (available from stacks.iop.org/PhysBio/10/035005/mmedia)). We therefore feel confident in presenting our analysis of the diversity of single amino acids and fragments of three consecutive amino acids.

2.4. Amino acid usage analysis

Finally, we used our diversity measure to determine if different positions had a bias toward using a specific type of amino acid. The diversity measure returns the effective number of fragments, or the fragments contributing most to that position. Therefore, at a window length of 1, we can extract the amino acids most contributing to the diversity by taking the rounded effective number of most common amino acids at a position. Each amino acid, determined to be relevant to a given position by its effective diversity, was categorized as belonging to one of three categories based on its hydrophobicity and tendency to be buried or on the surface of Ig. These categories are: hydrophobic (IVLFCMW), neutral (AGTSPYH), and hydrophilic (NDQEKR) [6, 16].

We could now characterize, for every type of V gene, each position into one of six types depending on how biased it was to using amino acids from only one of the categories. If a position

used only amino acids from one category, that position was considered to be of that type (i.e. a hydrophobic, a neutral, or a hydrophilic position). If the position had both neutral and one other category of amino acids, that position would be considered a ‘weak’ version of that category (i.e. weak hydrophobic or weak hydrophilic). If there were amino acids in all categories, then that position was considered indeterminate. In all instances, if a position had a single amino acid in one category and three or more in another category, the single amino acid category was ignored (i.e. if V genes were found to express, at a given position 4, neutral amino acids and 1 hydrophilic amino acid, then this position would be considered neutral).

3. Results

3.1. Analysis of diversity at window length 1

To get the cleanest view of amino acid level diversity we started by calculating the diversity of single amino acids without abundance bias ($q = 1$) (figure 1). This starting point is also beneficial as these results can be closely related to existing analyses of the Shannon entropy at the single amino acid level of V_κ , V_β and V_α genes [8]. We found that when comparing heavy chains to heavy chains (V_H/V_β) and light chains to light chains (V_κ/V_α and V_λ/V_α), TCR V genes are always more

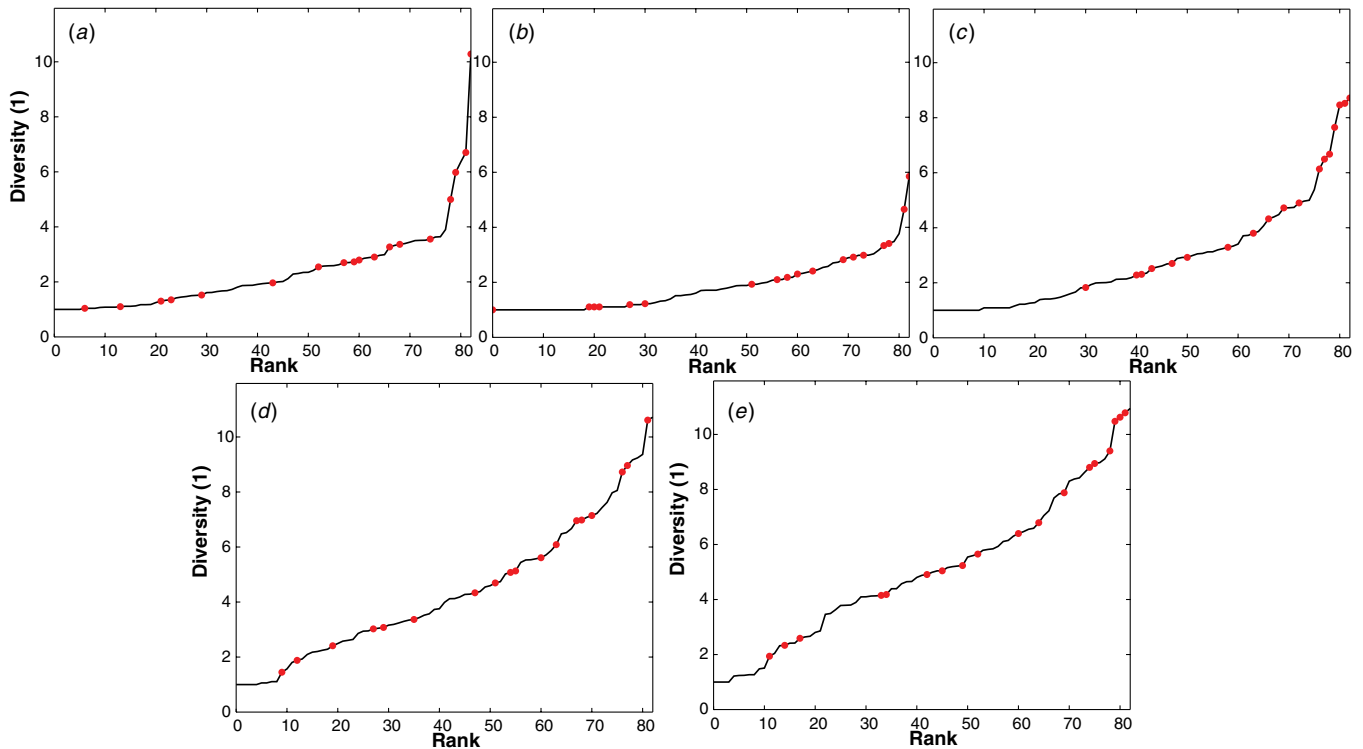


Figure 2. The ranked diversity of order 1 with a window length of 1. (a) The ranked diversity for V_H . The black line represents the diversity at each position sorted by diversity, while the red points (\circ) represents a value found in a CDR. (b) The ranked diversity for V_k . (c) The ranked diversity for V_λ . (d) The ranked diversity for V_β . (e) The ranked diversity for V_α .

Table 2. Quantile ranges for diversity of order 1 at a window length of 1.

Germline	Maximum–Q3	Q3–Q2	Q2–Q1	Q1–Minimum
V_H	10.3–2.99	2.96–2.29	2.12–1.51	1.50–1.00
V_k	5.85–2.92	2.89–2.10	2.10–1.61	1.56–1.00
V_λ	8.72–4.32	4.07–2.92	2.88–1.93	1.83–1.00
V_α	10.9–6.59	6.55–4.99	4.91–3.46	2.86–1.00
V_β	10.7–6.48	6.09–4.18	4.13–2.86	2.64–1.00

diverse than their BCR counterparts $p < 0.01$. In all cases we also found that CDR was more diverse than FR $p < 0.01$ (figures 1(c) and (d)). While most of the 18 CDR positions in all genes are in the top half of the diversity ranking, not all diverse sequence positions were in the CDR.

The ranked positions, sorted by diversity of order 1 with a window length of 1, revealed unique distributions for each germline with similar shapes within both the BCR V genes and the TCR V gene repertoires (figure 2). The BCR curves appear to follow convex functions while the TCR curves seem more linear or concave in distribution. We found the maximums and minimums of diversity to be in the same range in BCR and TCR V gene. However, most of the difference in BCR V gene diversity is in the upper quartile, while the TCR diversity distribution is more uniform, with the greatest difference in diversity being in the bottom quartile (see table 2). To more carefully quantify the relationship between diversity and CDR we next looked at the positions in the upper quartile of the diversity ranking. We found that in all cases the position with highest diversity was in the CDR and roughly half of positions

in CDR were from the upper quartile of diversity. Furthermore, in all V genes except for V_β the CDR is overrepresented with positions from the upper quartile of the diversity distribution. We can thus conclude that while the distinction between CDR and FR holds for BCR genes, it is less pronounced in TCR genes especially in V_β . The BCR has a very prescribed region of diversity, coinciding mostly with the CDR, while TCR are allowed some diversity throughout with only a few positions strongly conserved and invariant for reasons of receptor structure. Therefore, the reason CDR and FR are less distinct in TCR is not because the CDR is not diverse, rather it is because diversity is spread also in the FR.

To determine the extent to which positions of greater diversity and lesser diversity were consistent across genes we looked to see which positions from the upper or lower quartile of diversity were found in all BCR V genes or in all TCR V genes (figure 3(a)). A very clear picture emerges, with most of the positions from the lower quartile being consistent across V genes in BCR and V genes in TCR, 7 of which in all V genes. The diverse positions however seem to be less exactly consistent, although most of those that are consistent across V genes are in the CDR. Interestingly we also find three positions in FR3 (~at positions 80–95) that are diverse in both TCR V genes. It appears, however, that the lack of alignment of the most diverse positions was because they were only slightly mis-aligned. If we look at the ranked diversity of a window of three amino acids, suddenly multiple positions with high diversity are found in the CDR, of which five are in the same position in all V genes (figure 3(b), supplementary figure 4 (available from stacks.iop.org/PhysBio/10/035005/mmedia)).

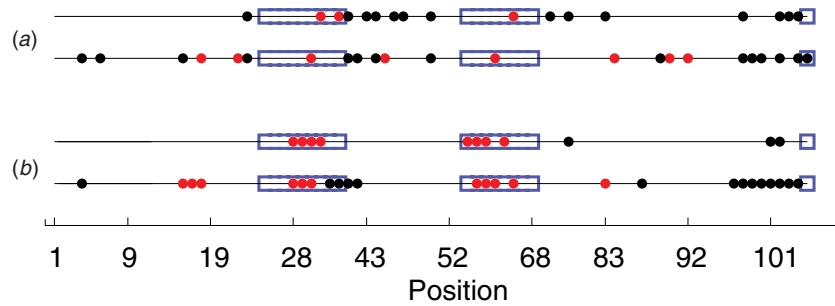


Figure 3. Diversity of order 1 overlapping positions in BCR and TCR germline sequences. (a) Overlapping positions for window length 1. The top line represents the overlapping positions across the BCR sequences, while the bottom line represents the same for TCR sequences. The CDR is highlighted in blue, while the red \circ are positions that are in the upper quartile and the black \circ are positions that are in the lower quartile. (b) Overlapping positions for window length 3.

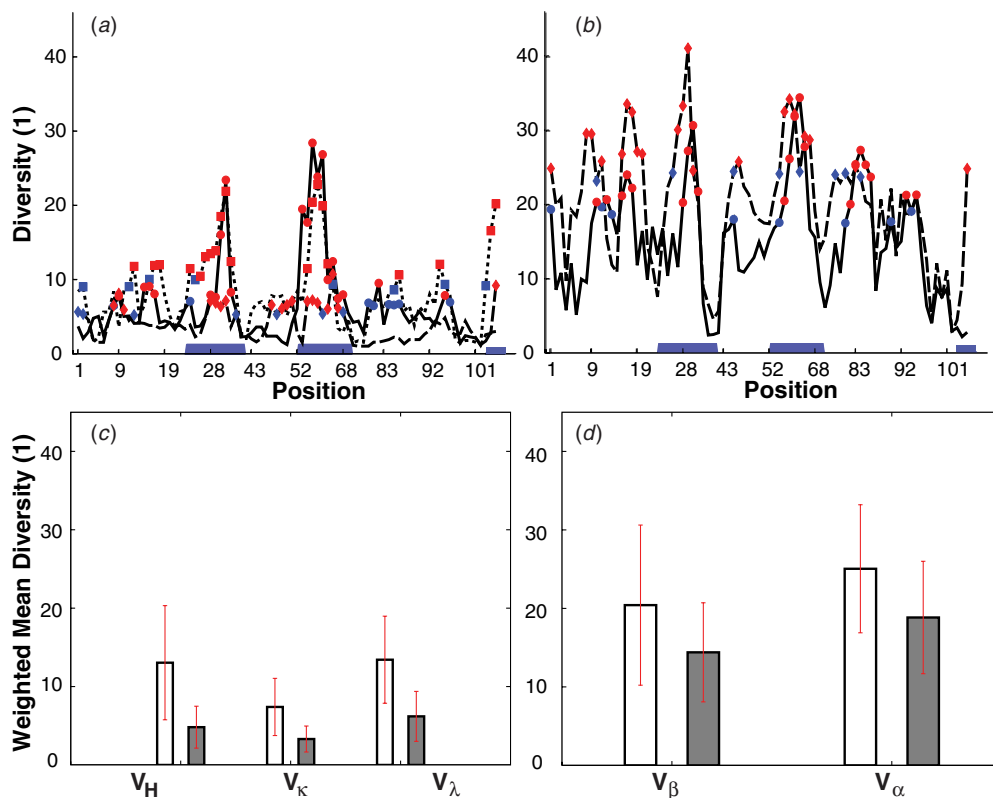


Figure 4. The diversity with a window length of 3. (a) The diversity of order 1 at each residue for V_H (—, \circ), V_κ (---, \diamond), and V_λ (\cdots , \square). (b) The diversity for order 1 at each residue for V_β (—, \circ) and V_α (---, \diamond). The positions in the upper quartile are indicated in red, while the next 10% maximal values below the upper quartile are indicated in blue. (c) The overall weighted mean for order 1 diversity of V_H (leftmost set), V_κ (center set), and V_λ (rightmost set). The left bar in each set represents the weighted mean for all CDRs (white) while the right bar represents the weighted mean for the FR region (gray). The error bars represent one weighted standard deviation above and one below the mean. (d) The overall weighted mean for diversities of V_β (left set) and V_α (right set).

3.2. Window length 3 analysis

To complement our view of diversity at a single position and take dependency of neighboring residues into consideration, we analyzed the sequence at different window lengths. Generally the results were similar at all window lengths with key features being more pronounced at some lengths. We show here the results of increasing the window length to three amino acids (figure 3(b), supplementary figure 4 (available from stacks.iop.org/PhysBio/10/035005/mmedia) and figure 4). Going from the 5' to the 3' end we could see clearly that in both TCRs and BCRs the diverse positions were

aligned in CDR (figures 3(b) and 4(a)). In general, replicating the analysis of the single amino acid level with the three amino acid fragment level revealed a clearer picture of the patterns occurring in different V genes, the similarities of V gene within each receptor type and the differences between them (figure 4). For $q = 1$, TCR repertoires were still significantly greater in their diversity than BCR repertoires and the CDR was more diverse than the FR, similar to the case for a window length of 1. However, due to the longer window length, the difference between CDR and FR in terms of ranges of diversity was greater, allowing us to more clearly identify diverse regions in

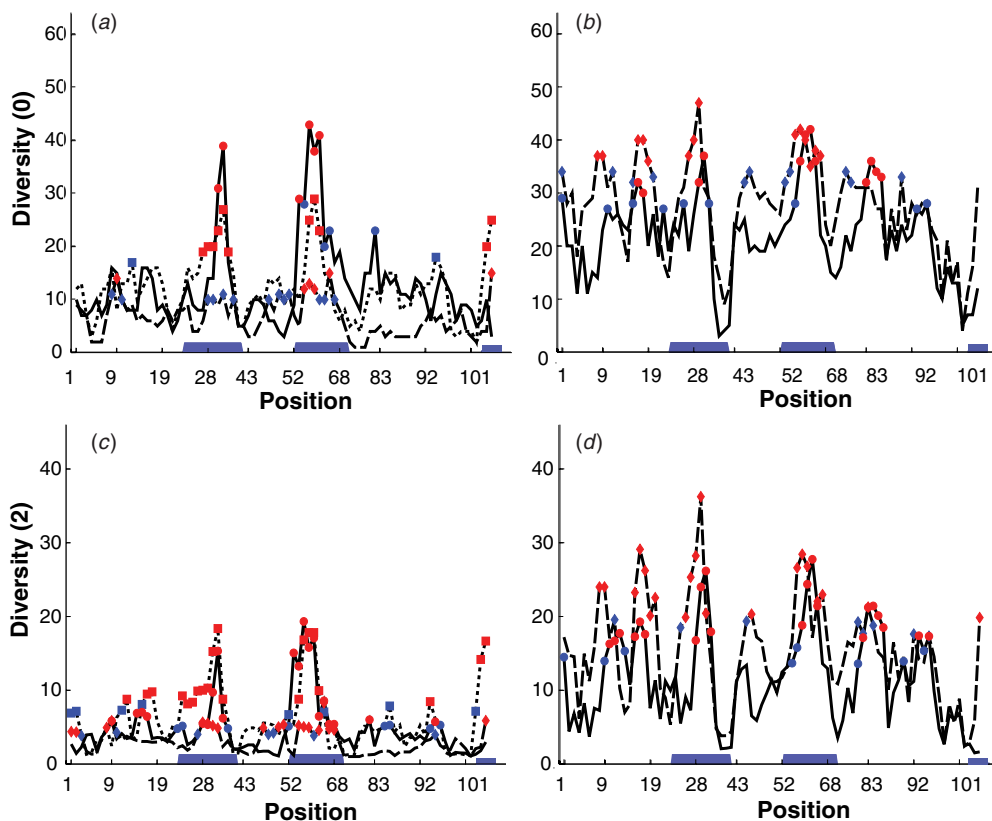


Figure 5. The diversity with a window length of 3 at orders 0 and 2. (a) The diversity of order 0 at each residue for V_H (—, \circ), V_k (---, \diamond), and V_λ (\cdots , \square). (b) The diversity for order 0 at each residue for V_β (—, \circ) and V_α (---, \diamond). (c) The diversity of order 2 at each residue for V_H (—, \circ), V_k (---, \diamond), and V_λ (\cdots , \square). (d) The diversity for order 2 at each residue for V_β (—, \circ) and V_α (---, \diamond).

CDR and invariant ones in the FR. In addition, the novel diverse regions in FR3 (~positions 80–92) and FR1 (~positions 6–19) are even more pronounced at a window length of three amino acids (figure 4). The range of diversity is higher than that of a single amino acid. However, given the ranges of diversity of single amino acids (table 2), even at its highest, it is indicative of the diversification of no more than two positions out of a window of three amino acids. It is interesting to note that when comparing positions both of single amino acids and three amino acid windows, V_k —the only BCR V gene characterized in depth until now—differs from the other BCR V genes, while V_H , and V_λ show the same placements of diversity and invariance (figures 1(a) and 4(a)). This difference is especially clear in the the CDR positions of V_k , which are less diverse, and also in the diverse positions in the middle of FR3, between positions 80 and 92.

3.3. Robustness of diversity analysis at different orders and directions of windowing

To determine if the diversities at each position would be influenced if we considered the number of genes that had a specific diversity at that position we compared our results at $q = 1$ to calculations of true diversity which were either biased toward rare sample amino acids ($q = 0$) (figures 5(a) and (b)) or common amino acids ($q = 2$) (figures 5(c) and (d)). While all positions are more diverse for $q = 0$ and less diverse for

$q = 2$ (figure 5) the range of diversity is at the same scale at all orders and the general relationships of diversity across the sequences remained the same as for $q = 1$. From this we can deduce that by and large diversity is as we described for order 1 and the difference between orders of diversity are sampling noise and not indications of the greater importance of specific amino acid patterns.

Completing our analysis on sequences of length 3 on a sliding window from 3' to 5' on the protein sequences also did not change our results other than allowing us to more clearly see diverse parts of CDR3 on the 3' side (supplementary figure 5 (available from stacks.iop.org/PhysBio/10/035005/mmedia)).

3.4. Using the diversity measure to analyze the tendency to use hydrophobic or hydrophilic amino acids

To illustrate how our diversity metric can be used to study the impact of diversity at different positions, we used diversity to characterize how biased different positions are in terms of their adherence to using hydrophobic or hydrophilic amino acids. We used this analysis to determine if diverse positions are more prone to using non-hydrophobic amino acids or if they are indeterminate in their choice of amino acids. Considering that regions with less hydrophobic residues will be more polyreactive [11], positions skewed against using hydrophobic residues are probably involved in antigen interaction. We

Table 3. The number of positions in the CDR and FR of each V gene type that are from the upper quartile in diversity, have a rounded diversity of 4 or more and belong to one of three categories: non-hydrophobic (+)—i.e. hydrophilic, weak hydrophilic or neutral; hydrophobic (—), i.e. hydrophobic or weak hydrophobic; and indeterminate (\sim).

	V_H^a		V_K^a		V_λ^a		V_α		V_β	
	CDR	FR	CDR	FR	CDR	FR	CDR	FR	CDR	FR
+	4	3	2	2	9	8	5	8	4	10
—	1	2	—	—	2	0	3	2	2	1
\sim	—	2	—	—	0	1	0	3	1	3
Sum	5	7	2	2	11	9	8	13	6	14
$D < 4$	3	5	6	11	0	0	0	0	0	0
Sum	8	12	8	13	11	9	8	13	6	14

^a The number of non-hydrophobic (+) positions were greater than expected compared to the relative size of the CDR and FR (18 versus 65 positions respectively).

looked at all positions in the upper quartile of diversity for each V gene. These highly diverse positions are over represented in CDR as opposed to FR in all V genes except V_β (table 3). However, all V genes also show diverse positions in FR. For the hydrophobicity analysis we only looked at positions whose diversity was 4 or above. We choose this cutoff as for lower diversities it becomes unclear how reliably the amino acids differentiated between three categories. Once we look only at these highly diverse positions we see the following— (1) CDRs of BCR V genes over-express non-hydrophobic positions (χ^2 , $p < 0.05$) while in TCR V genes they are dispersed equally between CDR and FR (compared to number of positions in each region—18 in CDR and 65 in FR) (table 3). (2) Hydrophobic positions are, in general, over expressed in the CDR of all V genes. However, their numbers are too low to make statistical inferences. (3) Indeterminate positions are evenly distributed between CDR and FR. This fits well with the idea that the non-hydrophobic diverse positions are involved in antigen interaction. In BCRs they are more focused in the CDR region, however they are found in the FR of both TCRs and BCRs. The V_K repertoire is simply not diverse enough to exhibit most of the positions we are studying here. This is not to say that V_K does not have positions with these kinds of structural roles. This result only means that looking at the diversity of the germline repertoire to detect these positions is not feasible. Potentially by including mutant sequences we would pin point these types of diverse positions in κ light chains as well.

3.5. Determining the impact of allelic diversity on the overall diversity

The last part of our analysis was to determine if the diversity patterns we see are the result of diversity amongst genes or amongst alleles. The average allelic diversity of order 1 for window length 1 at any position was rarely greater than 2 for both BCR and TCR (supplementary figures S6(a) and (b) (available from stacks.iop.org/PhysBio/10/035005/mmedia)). This average represents only a subset of the genes as many had only one allele type, and most had no diversity amongst their alleles (table 4). In all positions, many genes did not have a diversity over 1, meaning that all of the alleles in that gene at that position had the same sequence of amino acid. The number of genes that did have positions or alleles that differed from each other and expressed some level of diversity was smaller but in these genes the range of allele numbers was the same. Average allelic diversity for window length 3 also showed small fluctuations for some chains, although all repertoires again rarely surpassed a diversity of 2 (supplementary figure S6(c) and (d) (available from stacks.iop.org/PhysBio/10/035005/mmedia)). The number of alleles did not seem to be the limiting factor in determining the existence of diversity as, similarly to the single amino acid case, both positions with diversity of 1 and higher diversities had the same range of allele numbers (table 4). While the number of alleles reached as high as 10 in some genes even at a window size of 6 the diversity of alleles was in the range of 2, reaching 3 only in two cases (data not shown). This implies that by and large the number of positions that have intragenic diversity in their alleles is ~ 1 per 6 amino acid positions. We can see that this is not the case for the total levels of diversity amongst genes, since their diversity is much higher when we look at a sliding window of three amino acids than if we look at the diversity of single amino acids. It is therefore quite clear that most diversity in our analysis did not come from allelic diversity. We cannot conclusively disqualify the idea that better sampling of the human population and a more comprehensive database of allelic diversity, would change this. However, the plateauing of our rarefaction curves (see supplementary table 1 and supplementary figures 1–3 (available from stacks.iop.org/PhysBio/10/035005/mmedia)), the similarity in the numbers of alleles with no change in diversity and those who do change diversity, alongside the similarity in allelic diversity ranges at windows 1 and 3, lead us to conclude that this will probably not be the case. Having said that, it is interesting to note that some positions with

Table 4. Gene and allele counts for allelic diversity analysis for order 1. G_1 is the number of genes with one allele. ($G_{n,w} = 1$) is the range of genes with multiple alleles with a fragment diversity of 1 at window w . ($G_{n,w} > 1$) is the range of genes with multiple alleles with a fragment diversity greater than 1 at window w . (A) signifies the range of average number of alleles for the range of genes in the given column.

Germline	G_1	($G_{n,1} = 1$)/(A)	($G_{n,1} > 1$)/(A)	($G_{n,3} = 1$)/(A)	($G_{n,3} > 1$)/(A)
V_H	14	23–34/3.04–5.00	1–12/2.00–10.00	3–34/3.04–5.00	1–14/2.5–9.00
V_K	29	7–8/2.00–2.14	1–2/2.00–3.00	7–8/2.00–2.14	1–2/2.00–3.00
V_λ	14	15–18/2.00–2.55	1–4/2.00–4.00	12–18/2.00–2.55	1–7/2.00–4.00
V_α	20	20–24/2.50–3.00	1–2/2.00–5.00	9–24/2.50–3.00	1–4/2.00–5.00
V_β	18	26–28/2.45–5.00	1–3/2.00–7.00	21–28/2.00–5.00	1–6/2.00–7.00

Table 5. Division of sequence positions into Shannon entropy H categories as described in Stewart *et al* [8].

H	V_H	V_κ	V_λ	V_α	V_β	$V_{\kappa^{a,b}}$	V_α^a	$V_\beta^{a,b}$
0	45	33	55	12	14	34.03	12.45	11.62
1	33	32	26	17	28	27.39	14.94	15.77
2	5	19	2	54	41	21.58	55.61	55.61

^a Expected number of positions at each H category according to [8]. ^b Significantly different from expected ($p < 0.05$).

three alleles exhibited maximal diversity ($D = 3$). It would be interesting to see if, in actual immune repertoires in the population, V genes of each of these alleles would be used in equal measure or if potentially their effective diversity in such a case would be less than 3. Having verified that diversity is not based on the allelic level, we next checked at the V gene family level. We found that even in the bigger families the patterns of diversity followed those of the population as a whole and no part of the diversity of the repertoire was family specific (data not shown).

4. Conclusions

We have adapted here the methods of Jost and others [9] to use ‘true’ diversity to analyze amino acid usage patterns in V genes. We show that our diversity metric identifies known phenomena of CDR and FR structure while allowing for clearer comparisons and identifications of which amino acids drive the diversity we observe. We use this metric to analyze the entire known human germline repertoire, allowing us to more clearly characterize the source of this diversity as being based on the diversity of the genes and not the influenced by the allelic or family level. We also show here that there are patterns of amino acid diversity in TCR and BCR V genes that can characterize the different parts of these genes. It is important to note that while in the present analysis order of diversity did not have a huge impact on results, this may well not be the case when we use our methods to study an actual immune response. In such a case, we may in fact use the order of analysis to identify the more abundant clones. Potentially, this is for the same reason that gene families did not have much of an imprint on diversity patterns. The germline repertoire is evenly distributed amongst clones and so the different orders changed little in the results.

In BCR V genes we show that V_H and V_λ exhibit similar diversity patterns to those described in V_κ [8]. However, their CDR is more diverse and they appear to have hitherto undescribed areas of high diversity in FR1 and FR3, previously suggested in V_H [7] (figure 4(a)). Part of the reason that diversities seem so high in V_H and V_λ is that the V_κ germline repertoire we analyzed exhibits very low levels of diversity overall and is far less diverse than demonstrated by Stewart *et al*. Replicating their analysis [8], we see that our dataset shows a significantly different distribution into the H categories with significantly less (χ^2 , $p \ll 0.001$) of the highly diverse H2 positions in V_κ (table 5). We believe that this difference is an indication that the original dataset included many mutated sequences. We do not think this lower diversity is an issue of sampling as V_κ plateaued after utilizing

~62% of the data (supplementary table 2 (available from stacks.iop.org/PhysBio/10/035005/mmedia)).

TCR V genes are, as we expected, more diverse than BCR V genes [8]. However, the range of this diversity is similar to that in BCR V genes and it is only the number of positions that have median diversity values that is different (figure 2). Most of the high diversity areas in the CDR (figure 3(b)) and the exact positions of low diversity positions (figure 3(a)) (which presumably coincides with structurally important amino acids) are the same for both TCRs and BCRs. The differences in the numbers of median diversity positions makes the CDR and FR less distinct in the TCR V genes.

All these findings taken together lead us to the conclusion that the definition of CDR as the region of antigen interaction and FR as the region of structural importance, while generally true for V genes, may need reconsideration in some cases. High diversity positions can be found in the FR and some invariant positions are identified in the CDR. This is especially true for TCR V genes that appear from these results to interact with antigens in a much more flexible way. It is not clear why TCR V genes are more diverse than BCR V genes or why V_H and V_λ are more diverse than V_κ . Potentially this difference in diversity has to do with mutation. Unlike T cells, B cells can further diversify through mutation during an immune response to disease [19]. Thus potentially TCR V genes that cannot expect to add mutations to their diversity have evolved to be more diverse. In this context it is interesting that V_κ , whose codon usage makes it the most unstable for mutation (i.e. most diversifying under mutation [16]) is the V gene with the least amino acid diversity. Taken together these two findings seem to imply that perhaps V genes need to reach a certain average level of diversity and that the source of diversity is less important. However, the actual explanation is probably not as simple as this solution, since the positions with the greatest diversity, the CDR, are also the position which are most prone to change (in diversity) under mutation [16].

One final aspect of our analysis is the ability to determine how much more (or less) diverse each position is than other positions. For instance, we can observe that no position has an effective diversity of more than ~10 amino acids at a specific position. Indeed, we see that in general the ratio between maximum and minimum diversity is about the same in all V genes ($D_{\max/\min} \sim 10$). The difference in the mid-range diversities could be because the range of effective diversities we find in the middle quartiles of TCR and BCR V gene positions is very different: 6–3 for TCR and 3–2 for BCR. In other words, in most positions TCRs are twice as diverse, but not in the most diverse positions or the least diverse positions where they are the same as BCRs. The similarity in maximal and minimal diversity indicates to us that some characteristics of antigen interaction are the same for TCR and BCR and have the same limitations in terms of the amino acid usage they imply. At the same time, TCRs are more flexible in the exact contact points by which they interact with antigen, while BCRs need to maintain their structure to position most antigen contacts in the CDR, because of this the mid-range diversities differ between BCR and TCR V genes.

Our findings regarding the amino acid binding profiles of the most diverse positions strengthen this view. We

divided the diverse positions into three categories: anti-hydrophobic, hydrophobic and indeterminate. A bias against using hydrophobic amino acids is linked to poly-reactivity [11] and so this bias could be an indication of antigen interaction at that position. Those positions that have high diversity but are biased to using hydrophobic amino acids could be positions of structural importance that have a flexible role. As such, we note that they are mostly found at the edges of the CDRs, for instance at position 55 in CDR2 (see supplementary figure 7 and supplementary tables 3–7 (available from stacks.iop.org/PhysBio/10/035005/mmedia)). Finally, the high diversity positions that are indeterminate in their amino acid usage may simply be under less stringent selection. TCR V genes express anti-hydrophobic diverse positions evenly in both CDR and FR. BCR V genes also have such diverse positions in both CDR and FR but preferentially express them in the CDR. The more structural diverse positions are evenly distributed in both TCR and BCR V genes, although overall TCR V genes have many more indeterminate positions than do BCR V genes (supplementary figure 7 and supplementary tables 3–7 (available from stacks.iop.org/PhysBio/10/035005/mmedia)). We thus see again that while antigen interaction probably depends on sites in both CDR and FR, this interaction is more focused in the CDR in BCRs. At the same time TCR are more flexible and less constant in what sites will be important for a specific response. We would speculate that this is potentially because of differences between the MHC bound peptide string antigens TCRs interact with and the free floating protein structure antigens BCRs interact with. The simpler TCR antigen has less restriction on ways to interact than the complicated BCR antigen.

References

- [1] Wu T T and Kabat E A 1970 An analysis of the sequences of the variable regions of Bence Jones proteins and myeloma light chains and their implications for antibody complementarity *J. Exp. Med.* **132** 211–50
- [2] MacCallum R M, Martin A C and Thornton J M 1996 Antibody–antigen interactions: contact analysis and binding site topography *J. Mol. Biol.* **262** 732–45
- [3] Ehrenmann F and Lefranc M-P 2011 IMGT/3Dstructure-DB: Querying the IMGT Database for 3D Structures in Immunology and Immunoinformatics (IG or Antibodies, TR, MH, RPI, and FPIA) <http://cshprotocols.cshlp.org>
- [4] Arstila T P 1999 A direct estimate of the human T cell receptor diversity *Science* **286** 958–61
- [5] Boyd S D *et al* 2010 Individual variation in the germline Ig gene repertoire inferred from variable region gene rearrangements *J. Immunol.* **184** 6986–92
- [6] Chothia C, Gelfand I and Kister A 1998 Structural determinants in the sequences of immunoglobulin variable domain *J. Mol. Biol.* **278** 457–79
- [7] Capra J D and Kehoe J M 1974 Variable region sequences of five human immunoglobulin heavy chains of the VHIII subgroup: definitive identification of four heavy chain hypervariable regions *Proc. Natl Acad. Sci. USA* **71** 845–8
- [8] Stewart J J, Lee C Y, Ibrahim S, Watts P, Shlomchik M, Weigert M and Litwin S 1997 A Shannon entropy analysis of immunoglobulin and T cell receptor *Mol. Immunol.* **34** 1067–82
- [9] Jost L 2006 Entropy and diversity *Oikos* **113** 363–75
- [10] Tuomisto H 2010 A consistent terminology for quantifying species diversity? Yes, it does exist *Oecologia* **164** 853–60
- [11] Adib-Conquy M, Gilbert M and Avrameas S 1998 Effect of amino acid substitutions in the heavy chain CDR3 of an autoantibody on its reactivity *Int. Immunol.* **10** 341–6
- [12] Lefranc M-P 2008 IMGT, the international immunogenetics information system for immunoinformatics methods for querying IMGT databases, tools and web resources in the context of immunoinformatics *Mol. Biotechnol.* **40** 101–11
- [13] Shannon C E 1948 A mathematical theory of communication *Bell Syst. Tech. J.* **27** 379–423
- [14] Simpson E H 1949 Measurement of diversity *Nature* **163** 688
- [15] Hill M 1973 Diversity and evenness: a unifying notation and its consequences *Ecology* **54** 427–32
- [16] Hershberg U and Shlomchik M J 2006 Differences in potential for amino acid change following mutation reveals distinct strategies for and light chain variation *Proc. Natl Acad. Sci. USA* **103** 15963–8
- [17] Kabat E A, Wu T T, Reid-Miller M, Perry H and Gottesman K 1987 *Sequences of Proteins of Immunological Interest* (Washington, DC: US Govt Printing Office)
- [18] Heck L K, van Belle G and Simberloff D 1975 Explicit calculation of the rarefaction diversity measurement and the determination of sufficient sample size *Ecology* **56** 1459–61
- [19] McKean D, Huppe K, Bell M, Staudt L, Gerhard W and Weigert M 1984 Pillars article: generation of antibody diversity in the immune response of BALB/c mice to influenza virus hemagglutinin *Proc. Natl Acad. Sci. USA* **81** 3180–4