Molecular Immunology 65 (2015) 157–167

Contents lists available at ScienceDirect

Molecular Immunology

journal homepage: www.elsevier.com/locate/molimm

B cell Variable genes have evolved their codon usage to focus the targeted patterns of somatic mutation on the complementarity determining regions

Jasmine Saini [a], Uri Hershberg [a,b,*]

[a] School of Biomedical Engineering Sciences and Health Systems, Drexel University, Philadelphia, PA 19104, United States
[b] Department of Microbiology and Immunology, College of Medicine, Drexel University, Philadelphia, PA 19104, United States

ARTICLE INFO

Article history:
Received 11 September 2014
Received in revised form
29 November 2014
Accepted 2 January 2015

Keywords:
Somatic hypermutation
Codon usage
Affinity maturation
Evolution
Codon bias
B cells

ABSTRACT

The exceptional ability of B cells to diversify through somatic mutation and improve affinity of the repertoire toward the antigens is the cornerstone of adaptive immunity. Somatic mutation is not evenly distributed and exhibits certain micro-sequence specificities. We show here that the combination of somatic mutation targeting and the codon usage in human B cell receptor (BCR) Variable (V) genes create expected patterns of mutation and post mutation changes that are focused on their complementarity determining regions (CDR). T cell V genes are also skewed in targeting mutations but to a lesser extent and are lacking the codon usage bias observed in BCRs. This suggests that the observed skew in T cell receptors is due to their amino acid usage, which is similar to that of BCRs. The mutation targeting and the codon bias allow B cell CDRs to diversify by specifically accumulating nonconservative changes. We counted the distribution of mutations to CDR in 4 different human datasets. In all four cases we found that the number of actual mutations in the CDR correlated significantly with the V gene mutation biases to the CDR predicted by our models. Finally, it appears that the mutation bias in V genes indeed relates to their long-term survival in actual human repertoires. We observed that resting repertoires of B cells overexpressed V genes that were especially biased toward focused mutation and change in the CDR. This bias in V gene usage was somewhat relaxed at the height of the immune response to a vaccine, presumably because of the need for a wider diversity in a primary response. However, older patients did not retain this flexibility and were biased toward using only highly skewed V genes at all stages of their response.

1. Introduction

The diversity of the immune system's B cell receptor (BCR) repertoires is a key part of its ability to generate protective immunity and respond to most any disease (Burnet, 1959; McHeyzer-Williams and McHeyzer-Williams, 2005). B cells are unique in that their receptor diversity is generated somatically (Weill and Reynaud, 1996; Neuberger and Milstein, 1995; Reynaud et al., 1991). During an immune response, in a process termed affinity maturation, B cells proliferate, mutate their BCR genes, and die resulting in a novel population of BCR mutants with higher affinity of interaction with the antigen (Burnet, 1957). This somatic diversity is generated on an already diversified background since every receptor is a combination of a heavy chain and a light chain (κ or λ)(Tonegawa, 1983). These chains are each constructed through the recombination of specific gene segments – Variable (V), Diverse (D) (in heavy chains only) and Joining (J) genes (Early et al., 1980). This leads to the construction of a highly diverse repertoire whose sequence structure and affinities, while related, are not directly based on the germline genes that constructed it. Thus, although the germline diversification of the V gene repertoire is probably influenced by the need for affinities for specific pervasive pathogens the evolution of the repertoire (Baumgarth et al., 1999; Artavanis-Tsakonas et al., 2003), its germline diversity and eventual somatic diversity are linked in a more complex manner (Kepler, 1997; Hershberg and Shlomchik, 2006). Specifically, unlike any other coding region, we would expect that the BCR genes have evolved their nucleotide structure to withstand high levels of point mutations as part of their normal function. Indeed it has been shown that the V genes encoding the BCR have evolved their nucleotide sequences to maximize the utility of mutation (Neuberger and Milstein, 1995; Kepler, 1997; Hershberg and

* Corresponding author at: School of Biomedical Engineering Sciences and Health Systems, Drexel University, Philadelphia, PA 19104, United States.
Tel.: +1 215 895 1698.
E-mail address: uri.hershberg@drexel.edu (U. Hershberg).

http://dx.doi.org/10.1016/j.molimm.2015.01.001
0161-5890/

Shlomchik, 2006; Wagner et al., 1995; Oprea and Kepler, 1999; Kepler et al., 2014).

At the heart of this statement is the understanding that the redundancy of the genetic code allows for different nucleotide sequences that can encode the same protein structure. Thus, using specific codons (i.e. codon bias) can impact the potential of mutations (Hershberg and Shlomchik, 2006; Hershberg and Petrov, 2008; Plotkin and Kudla, 2011). There are 64 codons translating to 20 amino acids. Amino acids are thus differently redundant in the way that they are encoded. Some have 6 codons encoding them while others have between 1 and 4 codons. The number of mutations needed to traverse between codons encoding different amino acids is also not identical and can range from one to three mutations, depending on the difference in their encoding codons. An immediate impact of the redundancy in the genetic code is that certain point mutations are silent (or synonymous) and the equivalent amino acid is maintained, thereby enhancing the stability of such codons against random point mutations.

To fully understand the practical implications of codon bias on BCR mutation and stability in the face of mutation, we must consider the receptor's structure. The BCR is subdivided into three regions of high variability that are the focus points of antigen binding, called complementarity-determining regions (CDR), and four lower variability framework regions (FR), coding for the receptor structural backbone (Wu and Kabat, 1970). We would thus expect that CDR and FR to have opposing codon biases that enhance stability in the face of mutation in the FR and the generation of further diversity in the CDR.

Indeed, differential codon bias of CDR and FR was observed in the rat B cell repertoire shortly after V genes were first described (Miyata et al., 1979). Similarly, it was suggested that proper models of somatic mutation must consider this codon bias (Chang and Casali, 1994). The possible impact of codon bias for different mutational outcomes in different V gene regions was first described by comparing Serine codon usage in CDR and FR of the heavy and κ light chains in BCR and α and β chains of T cell receptors (TCR) (Wagner et al., 1995). This study showed that CDR of BCRs mostly have the less stable codon encoding Serine (AGY) while the FR utilizes mostly the stable one (TCN). TCR α and β chains exhibited no such bias, which is not surprising, as they are not thought to somatically mutate (Wagner et al., 1995).

To understand how these results can be generalized beyond Serine we must consider also that somatic mutation is engineered and by no means random (Betz et al., 1993a, 1993b). It has been shown that micro-sequence relationships can predict much of the patterns of mutation (Shapiro et al., 1999, 2002; Cowell et al., 1999; Yaari et al., 2013). Thus, the codon usage and nucleotide structure of the sequences in general can both influence the distribution of mutations within the sequence and the likelihood of a mutation resulting in an amino acid change. Following on these possibilities, previous studies have shown that the CDR is indeed more prone to mutation than FR (Kepler, 1997) and that codon usage, even beyond Serine, is biased toward more amino acid change as a result of mutation in CDR and less in FR (Hershberg and Shlomchik, 2006). Finally, it was also shown that there was some synergistic effect between the two biases (Oprea and Kepler, 1999). Interestingly, these last results seemed to suggest that germline bias to accommodate mutation occurs in both TCR and BCR genes, in direct contradiction of the original findings of Neuberger regarding serine codon usage (Wagner et al., 1995). However, in many of these studies codon bias was compared only between CDR and FR directly without using any outside indicator of the scale of differences. This is problematic since FR and CDR are under different selective pressures: CDR is under selection to encourage change upon mutation and FR, to remain stable (Hershberg and Shlomchik, 2006). It is therefore ill advised to use one as control against the other. Instead, they should both be compared against other genes that are not undergoing somatic mutations and thus are not under selection pressures that V genes undergo (Hershberg and Shlomchik, 2006). When this is done we find that while heavy and λ light chains are indeed unstable in the face of mutation in CDR and stable in FR, the CDR and FR of κ are both skewed to be unstable in the face of mutation, albeit the CDR is more so (Hershberg and Shlomchik, 2006). This study further suggested that the skewed bias of mutation outcomes is focused on generating nonconservative amino acid changes in the CDR that were shown to be of meaningful and quantifiable consequence in murine immune response (Hershberg and Shlomchik, 2006). Other light chain homologs that do not undergo somatic mutations, such as CD8, were shown to have a non-skewed codon usage in comparison (Hershberg and Shlomchik, 2006). However, this last study only considered the expected outcomes of mutations and not the effects on targeting of somatic mutations.

In the research presented here, we have taken mixed analysis simulation approach to study the impact of V gene nucleotide structure and codon usage on both mutation targeting and mutation outcome. We found that both TCRs and BCRs show a skew toward targeting and change of amino acids in the CDR, compared to the CDRs relative size in the V gene sequences, and that this skew is more pronounced in BCRs. Furthermore, using the codon usage of the Immunoglobulin Superfamily (SFIg) proteins as a background, we found that only in BCRs is the skew toward change in the CDR dependent on a significant skew in codon usage. We therefore suggest that the bias in TCRs stems from the structural similarities of TCRs and BCRs that are expressed in very similar amino acid usages between the two regions (Supplemental Fig. 1), and that the targeting has evolved to focus on the parts of the genetic code that encode the amino acids of which all CDRs (in BCR and TCR) are made.

Supplemental Fig. 1 related to this article can be found, in the online version, at http://dx.doi.org/10.1016/j.molimm.2015.01.001.

Finally, we compared the predicted bias to change in the CDR for each different V_H gene, to the actual level of mutations observed in clones that use that V_H. We found them to be significantly positively correlated in 4 different human recombined datasets (Wu et al., 2012, 2010; Wang et al., 2011) and across all V_H.

2. Methods

2.1. Sequences analyzed

All germline data was taken from the international ImMunoGeneTics database (IMGT) (Lefranc et al., 2003). To eliminate bias from V gene families, only the first functional alleles were studied. The dataset included BCR V genes (49 V_H, 53 Vκ and 33 Vλ) and TCR V genes (45 Vα and 48 Vβ) for comparisons. As controls we also studied the related pseudo genes for each of the BCR V gene types (51 V_H, 23 Vκ and 8 Vλ) and the V_H gene repertoires of mouse (*Mus musculus*), rat (*Rattus norvegicus*), rabbit (*Oryctolagus cuniculus*) sheep (*Ovis aries*) and zebrafish (*Danio rerio*) (174, 117, 39, 7 and 35 genes respectively). In all cases, the IMGT unique numbering system was used to describe sequence positions across all V gene types. Within this numbering, FR and CDR segment were modified from IMGT definitions similarly to our previous studies of codon bias (FR1 = 1–24; CDR1 = 25–40; FR2 = 41–53; CDR2 = 56–65; FR3 = 66–104; CDR3 = 105–106)(Hershberg and Shlomchik, 2006; Lefranc et al., 2003; Kabat et al., 1991). These definitions slightly enlarge the CDR regions compared to IMGT definitions. The results presented here do not change in their trends or significance if IMGT CDR definitions are used. BCR and TCR V genes are part of SFIg proteins that have the PF00047 Ig-folding domain and are involved in antigen presentation, recognition and binding processes of the cell

(Williams and Barclay, 1988; Bork et al., 1994; Brümmendorf and Rathjen, 1995).

We compared our estimations of mutation frequencies and amino acid changes, based on germline analysis, to four datasets of recombined sequence repertoires taken from the following published experiments:

DATASET1 = Sequences from V_H repertoire following influenza and pneumococcal vaccinations from twelve healthy volunteers from two sets – 6 young from ages 19–45 years and 6 old from ages 70–89 (Wu et al., 2012).
DATASET2 = Sequences from mononuclear cells isolated from 3 healthy volunteers aged 21–26 years (Wu et al., 2010).
DATASET3 = Sequence samples collected from 14 healthy residents, aged 22–53 years, of Papua New Guinea (PNG) region (area of endemic parasitism) (Wang et al., 2011).
DATASET4 = Sequence samples collected from the 14 healthy residents of Sydney Australia (AUZ) (Wang et al., 2011).

The sequences were separated into clones defined by same V gene, J gene and CDR3 length.

2.2. Likelihood to have a targeted point mutation and how likely are these mutations to change amino acid and/or its traits

For each nucleotide position, we calculated its Mutation Likelihood and Changeability. (i) The tendency to mutate was measured by Mscore, Changeability or the likelihood that the mutation would result in an amino acid change was calculated in two ways: (ii) the probability mutation leads to a nonsynonymous change (Rscore) and (iii) the probability of having nonconservative change (Tscore).

(i) Mutation Likelihood – Mutability and Mscore: This calculation was built on a targeting model of pentamers, which assigns a specific mutation targeting score called mutability based on the center of the pentamer at which the nucleotide position is found (Yaari et al., 2013). The mutability value was based on the Yaari et al. pentamer model predictions of somatic mutation (Yaari et al., 2013). We chose this model as it has ten times (10×) more differentiation in it's hot vs. cold spots compared to the older tri-nucleotide model which only showed twofold differences (Yaari et al., 2013). This difference makes the mutability in CDR more accurate compared to the reported effects of hot spots (Yaari et al., 2013). Furthermore, it makes low and high mutabilities more easily distinguishable then if we use the previous tri-nucleotide models (Shapiro et al., 1999) that would have forced us to average mutabilities at each position. We calculated the relative mutability for each nucleotide compared to all other positions in the sequence.

As this targeting model describes the ratio of mutabilities of different positions, Mscore of 1 represents a neutral position with no targeting bias. Mscore ranging from 0.1 to 1 are stable positions, i.e. less likely to be targeted. Unstable positions, i.e., mutational hotspots, have Mscore of >1. It is important to note that these scores are relative. Hence if we consider two positions, one with an Mscore of 0.1 and one that has 0.5, the later position would be 5 times (0.5/0.1) more likely to be targeted by mutation than its neighbor. Thus, to understand the impact of an Mscore, we must know its background in the rest of the sequence. For this reason we divided each positional Mscore by the sum of all Mscores in the sequence of interest. In this way we got an Mscore for each position that was a fraction of the total likelihood of mutation anywhere in the sequence, while maintaining the correct ratio compared to the other positions in the sequence. In this way we get fractions for each position summing to 1 for the whole sequence. To obtain the likelihood of a given mutation occurring in the CDR, we summed the fractional mutation scores only in the CDR positions. Given the area taken by CDRs in the sequence, if mutation was completely unbiased we would expect the likelihood of a mutation to be in the CDRs to be ~ 0.27, as this is the relative fraction of positions that are in the CDR of V genes.

When we discus the relative impact of stable and mutable positions we wished to preserve their symmetry of the impact. We therefore considered the log transformation of the Mscore. To consider the average score of a sequence, or a position across different genes, the summed score of nucleotide log mutability of all the positions represented the sequence Mscore, which was then normalized by the length of the sequence and inverse logged.

$$\text{Mscore}_{(\text{seq})} = 2^{\left(\left[\sum_{i=1}^{n} \text{Log}_2\,\text{Mutability}_{(i)}\right]/n\right)}$$

(ii) Mutational Outcome – Sscore, Rscore and Tscore: For each nucleotide, the probability of having a viable (i.e. not to stop) nonsynonymous change (Rscore) and probability of having a viable nonconservative change (Tscore) were calculated, given the nucleotide specific transition–transversion bias (Yaari et al., 2013). Amino acids were characterized into 3 trait groups based on their hydrophobicity and location on the receptor – hydrophobic/buried (F, L, I, M, V, C, W), hydrophilic/surface (Q, R, N, K, D, E) or neutral/intermediate (S, P, T, A, Y, H, G)(Chothia et al., 1998). This division of amino acids has been shown to be relevant to the process of immune selection (Hershberg and Shlomchik, 2006). All nonconservative mutations were considered equal step changes.

These probabilities range from 0 to 1, where 0 means that a mutation at that position would never lead to an amino acid replacement or trait change, and probability score of 1, would always lead to a replacement or trait change, with a viable amino acid. All these positional scores were then made relative by multiplying them by the fractional Mscores calculated above and dividing them by the total sum of probabilities at all positions where such a mutation is possible. We could then ask again what fraction of amino acid changes (or nonconservative amino acid changes) are expected to occur in the CDR by adding the scores for all positions in the CDR. The expected likelihood of a change following mutation being in the CDR segments, given their length in the sequences, is again ~0.27 if amino acids are evenly distributed.

To calculate the averaged sequence values of the changeability scores we follow the same equation as for mutability (n being the number of positions where such a change is possible):

$$\text{Rscore}_{(\text{seq})} = \sum_{i=1}^{n} \left(\frac{\text{Mutability}_{(i)}}{\sum_{i=1}^{n} \text{Mutability}_{(i)}} \times \text{Rscore}_{(i)} \right)$$

$$\text{Tscore}_{(\text{seq})} = \sum_{i=1}^{n} \left(\frac{\text{Mutability}_{(i)}}{\sum_{i=1}^{n} \text{Mutability}_{(i)}} \times \text{Tscore}_{(i)} \right)$$

2.3. Measuring codon bias

To calculate the effect of codon bias we generated 5000 simulated sequences for every germline V gene. The simulated genes had an identical amino acid composition to their related germline and differed only in their codon usage. For comparison to the human V genes the simulated sequences' codon usage was taken from the

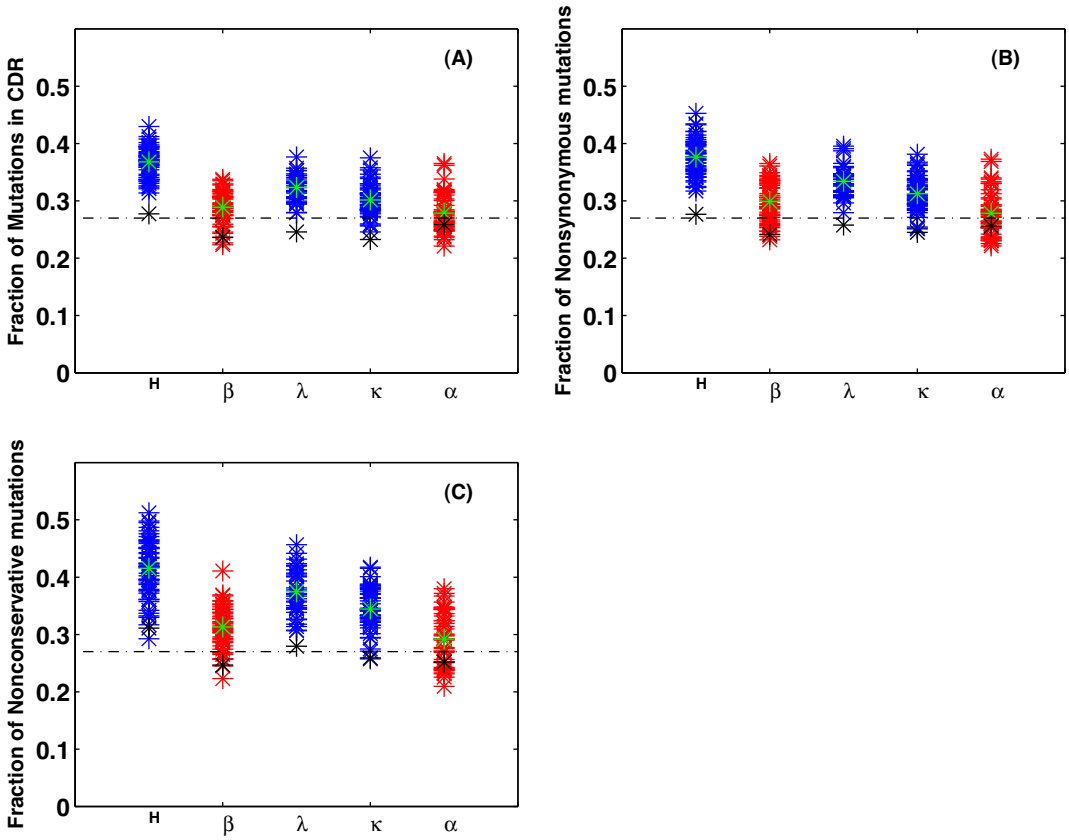

Fig. 1. Expected likelihood of mutation or change of amino acid, following mutation being in the CDR, of BCR VH, Vλ, Vκ, and TCR Vα and Vβ genes, under a targeted model of mutation (Yaari et al., 2013): Expected likelihood that given a mutation occurs it will occur (a) in the CDR of a given V gene (b) cause an amino acid change in the CDR and (c) cause a non-conservative amino acid change in the CDR. Given the length of the V sequences, the fraction of positions in CDR is ~0.27 marked by the dashed line. The black star represents the actual CDR fraction of that family and the green star is the mean fraction of each family. (For interpretation of the references to color in figure legend, the reader is referred to the web version of the article.)

codon distribution of the 2993 SFIg sequences from the Ensemble database (Kersey et al., 2014). We used these genes to represent a codon usage uninfluenced by somatic mutation as the SFIg proteins share many structural characteristics with V genes but do not undergo somatic hypermutation. For comparison to nonhuman V genes we generated the simulated sequences with a codon bias appropriate to each species taken from the codon usage database (Nakamura et al., 2000; National Institute of Genetics). In all cases we used the human pentamer targeting model to target somatic mutations (Yaari et al., 2013). We then calculated for the germline and each sequence in its matching simulated dataset, the mutability and changeability (nonsynonymous and nonconservative) at each position and the likelihood of mutations or changes following the mutation to occur in the CDR (and not in the FR). Graphically, we found that all of these measures were normally distributed across the simulated dataset. We therefore quantified the skew of germline sequences from the simulated distribution by using a *z*-test statistic. The *z*-score value was obtained for each germline sequence, by comparing each actual V gene's likelihood of mutation or change being in the CDR to the mean and standard deviation values of this measure obtained from its associated simulated dataset. The sequence model was written in Perl programming language. All the statistical analysis and graphs were done either using Matlab or R.

2.4. Validation of model

To check the efficacy of our predictions, we counted mutations in the recombined datasets (see above). We then tested the Spearman correlation between our calculated expected fraction of mutations in the CDR of each germline V gene, and the fraction of actual mutation counts observed in CDR of clones from the same V gene. To prevent any selection bias, when correlating the probability to have silent (synonymous) mutations, we considered only the 4 fold redundant 'wobble' positions of amino acids encoded with four codons, where all mutations are synonymous.

3. Results

3.1. The targeting patterns of somatic mutation cause the expected patterns of mutation and the tendency to change upon mutation to be focused in the CDR

We calculated for each V gene the expected fraction of mutations that should occur in the CDR and the expected fraction of mutations that change amino acid or do so in a non-conservative way in the CDR (Fig. 1). Using nonparametric Wilcoxon rank test, we compared the uniform distribution to the expected fractions and found that the V genes of BCR heavy and light chains are all significantly skewed toward mutations and changes of amino acid following mutation occurring in the CDR. All Heavy and λ V genes and most κ V genes have expected fractions of mutation and change above the uniform distribution in the CDR. TCR V genes, which are less skewed, also show significant bias, with some V genes having more mutations targeted to the CDR ($p < 0.05$, Fig. 1a–c). In general heavy chains are more biased in the skew toward mutating than light chains.

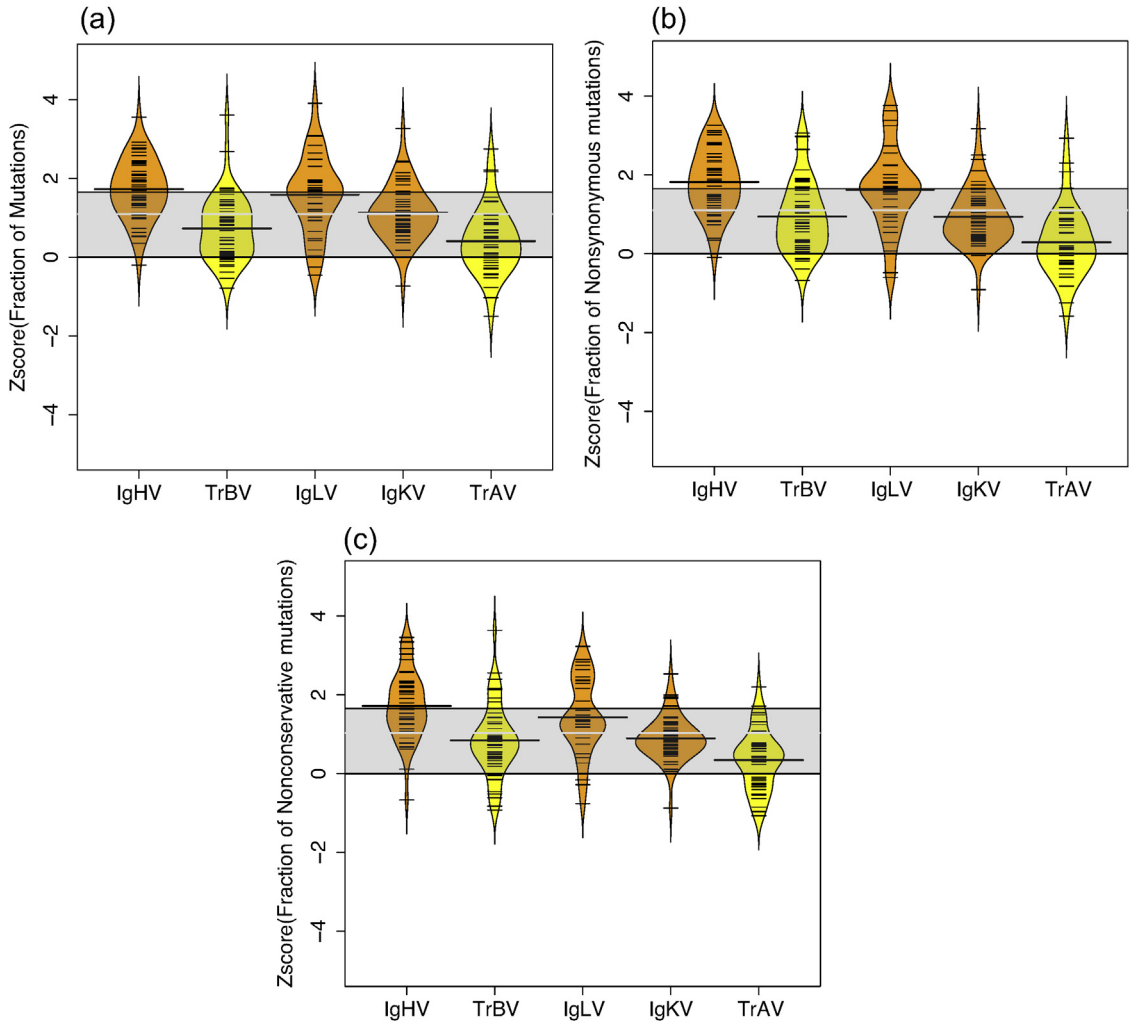

Fig. 2. The distribution of *z*-scores when comparing the original germline CDR fraction of each V gene to the mean and standard deviation of its related simulated dataset (a) comparing *z*-scores of the expected fraction of mutations in CDR; (b) the expected fraction of amino acid change in the CDR; and (c) the expected fraction of nonconservative amino acid changes in the CDR. The distribution means falling in the shaded gray region are not statistically distinct in their codon bias from the simulated distribution.

3.2. The codon bias of BCR V genes enhances the skew toward changes in the CDR

We next tested if the fraction of mutations and changes that occur in the CDR in the actual germline sequences is the result of a codon bias. To test this we calculated *z*-score of each germline sequence and compared it to the mean and standard deviation obtained from the distribution of simulated sequences with identical amino acid compositions but with a codon usage sampled from the SFIg (see Section 2.3). In BCR heavy and λ light chains, we found that the codon usage significantly enhanced the tendency to focus changes in the CDR ($p < 0.05$), this was not the case for κ and TCR V genes (Fig. 2a–c). In fact in BCR light chains it accounted for the entire effect observed (Supplemental Fig. 2). This was true both in terms of the bias to mutate and the bias to change amino acid upon mutation. In TCRs the codon bias has a much less pronounced effect if any (Fig. 2a–c). As for TCRs, the pseudo-genes of BCR heavy and light chain V genes exhibit less codon bias than the their functional equivalents (Mann–Whitney $p < 0.05$ – Supplemental Fig. 3). To further bolster this finding, we also considered the expected fraction of mutations in the CDR of heavy chains in other species and then calculated the influences of codon bias on these skews. In all cases we see that greater skew toward mutating in the CDR regions is accompanied by greater codon bias. Amino acid patterns in these different CDRs stay much the same (Supplemental Fig. 4).

Supplemental Figs. 2–4 related to this article can be found, in the online version, at http://dx.doi.org/10.1016/j.molimm.2015.01.001.

3.3. The V gene of BCRs has nucleotide structure that localizes the targeting of mutations

We had expected the skew in κ light chains would be different from that found in heavy and κ light chains, given that previous findings have shown that the bias in κ is not based on focusing change on CDR compared to FR (Kepler, 1997; Hershberg and Shlomchik, 2006). Rather, κ light chains exhibit codon bias in both CDR and FW (Kepler, 1997; Hershberg and Shlomchik, 2006). This would be hard to see with the test we describe above. To attempt to test if the V genes in κ light chains and other BCR V genes in general have a subset of positions that are more focused to mutated (even if this focus is not always in the CDR) we compared the distribution of expected mutations across the sequence. To do so we averaged the relative fraction of mutation at each position, as predicted by our model (calculated as described in Section 2), across all the sequences of a given V gene type (e.g. we averaged the fraction of mutations at position *n* across all 49 BCR heavy chains, then position $n+1$ and so on). We verified that this averaging had not changed the distribution of fractions by ensuring that the sum of averaged fractions for the V gene type was 1. We then ranked the different V

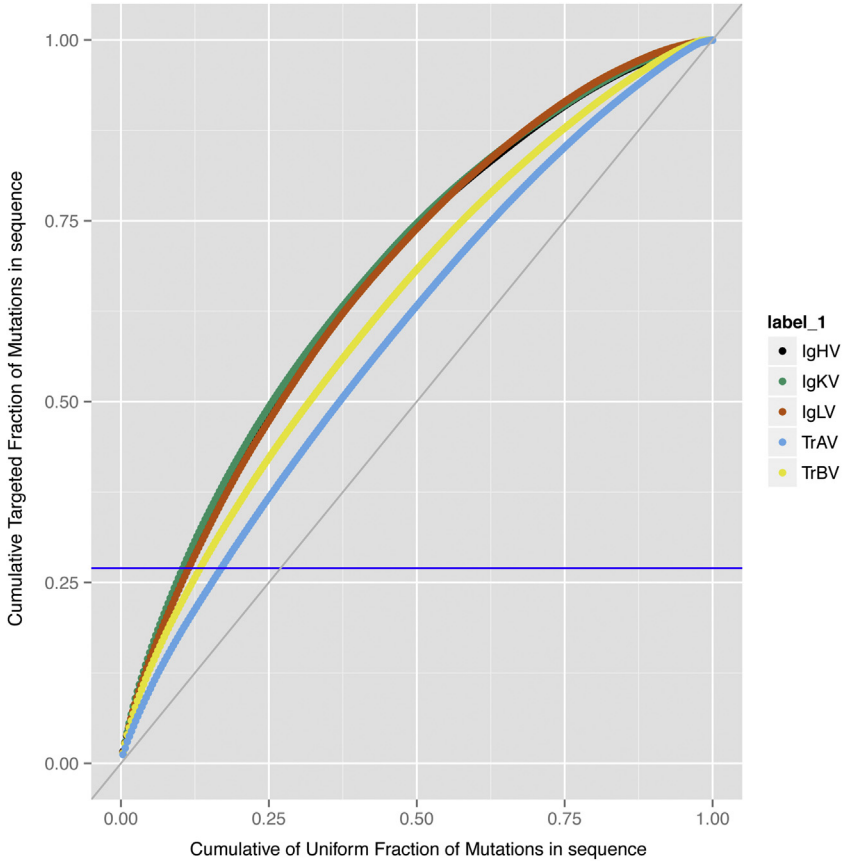

Fig. 3. CDF of the average mutation fraction (see Results Section 3.3) per position compared to a uniform distribution of mutation fractions across the V genes – BCR V_H (black), Vλ (orange), Vκ (green), TCR Vβ (yellow) and Vα (blue). BCR V genes were not distinct from each other but were significantly distinct from TCR Vα. TCR Vβ was intermediate as it was not significantly distinct from either TCR α V or the BCR V genes. (For interpretation of the references to color in figure legend, the reader is referred to the web version of the article.)

gene positions by their fractional potential mutability and plotted their cumulative distribution function (CDF). We did this for every V gene type in TCR and BCR V genes. The distributions were compared using nonparametric Kolmogorov–Smirnov test. We found that all BCR V genes show a nearly identical focusing of the mutability while in TCR's mutability is more evenly distributed across the whole sequence, i.e. closer to the diagonal ($x = y$) line. Interestingly β chains still show some intermediate structure between α and the BCR V genes (Fig. 3).

3.4. Mutations in the CDR are focused on non-conservative changes

We calculated the average sequence Mscore, Rscore, and Tscore for the two regions, FR and CDR, of each V gene. These average scores represent the likelihood that the average position in each region will mutate, change amino acid or do so in a non-conservative way. When we incorporate mutation targeting into our calculations, we find, as we would expect from the results above, that CDRs have significantly more mutable positions and FR have less mutable ones. The distinction between CDR and FR is significant in both B cell and T cell V genes (Mann–Whitney all $p < 0.05$ (Fig. 4a)) It is interesting to note that even in these sequences highly targeted for mutation most positions are actually biased against mutation as the average even in CDR is below the ratio score of 1 (red line in Fig. 4a). This does not contradict any of previous statements, as biased targeting toward CDR depends on the difference between CDR and FR, not on their absolute scores. It does indicate that even in the CDR most positions are biased against mutation.

In terms of the propensity to change upon mutation, when we incorporate mutation targeting, an interesting phenomenon emerges. While FR indeed has positions with a propensity to change that is less than expected, the positions in the CDR are even **less** changeable than those in the FR (Fig. 4b, **all $p < 0.05$**). With respect to non-conservative mutations, BCRs show a higher tendency for nonconservative changes in the CDR than FR. BCR CDRs are thus especially focused on nonconservative mutations at the expense of having amino acid changes of simply any kind. The CDRs of TCR on the other hand continue to show the same skew as they did in general non-synonymous mutations, i.e. the CDR has an average position tendency to change non-conservatively that is less than that observed in the FR. (Fig. 4c). Overall this implies for TCRs that they are biased to mutate in the CDR but then not change amino acid.

3.5. The expected skew toward changes in the CDR can be seen in recombined V gene mutants of the immune repertoire

To test how well our germline-based model of expected mutation predicted actual tendencies toward mutation and change of amino acids in the CDR, we analyzed several human recombined heavy chain repertoires. We observed a significant correlation between the predicted CDR fraction of mutations generated by our germline model and the observed mutation fraction in CDR in the recombined sequence datasets. We found a strong correlation between our predictions and the fraction of nonsynonymous and nonconservative mutation observed in CDR positions (Fig. 5 and Table 1). As expected, silent mutations that do not undergo

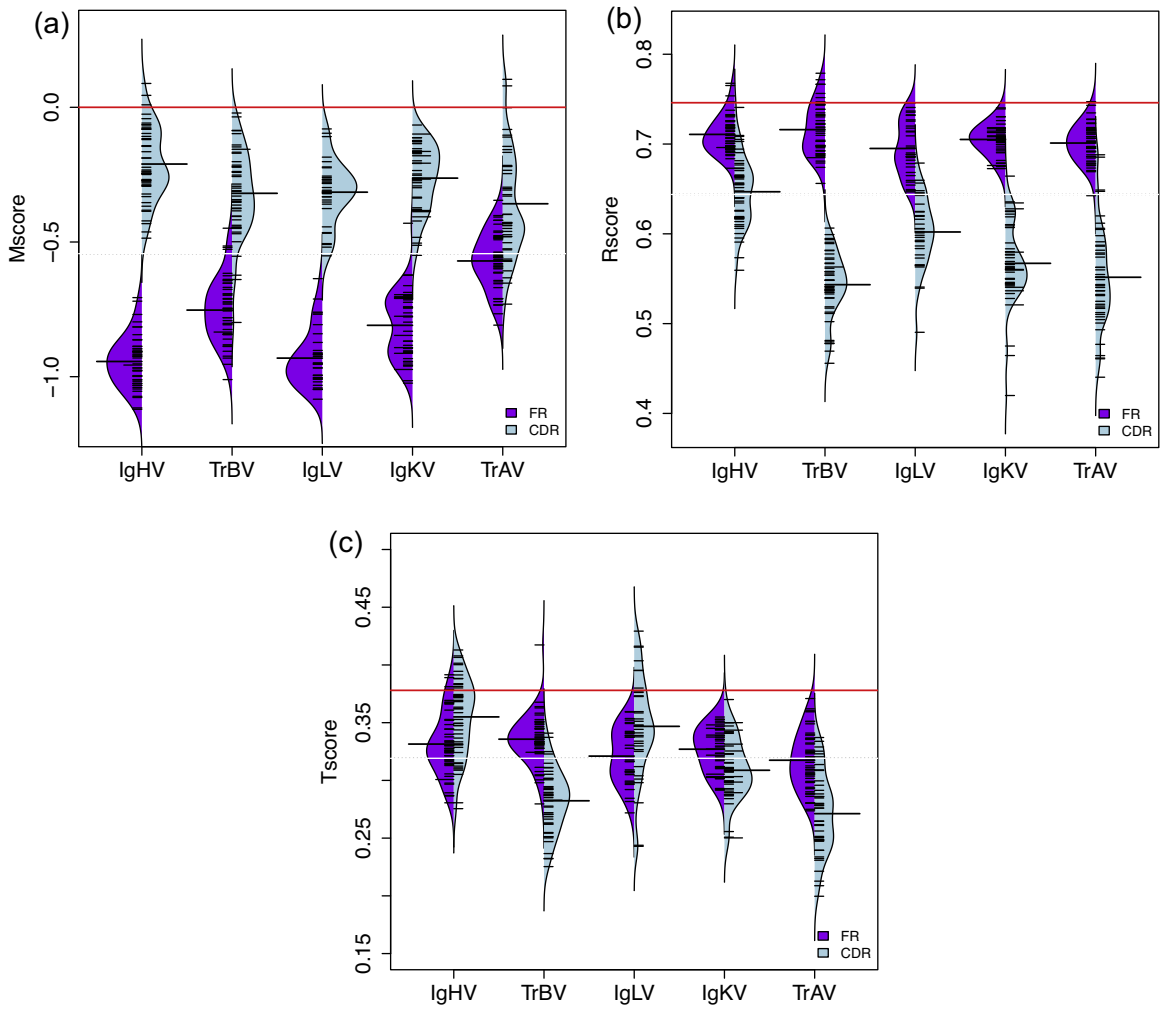

Fig. 4. The average, by positions scores for BCR V_H, Vλ, Vκ, TCR Vβ and Vα in CDR (purple) and FR (blue) for (a) Mscore (b) Rscore and (c) Tscore, under a targeted model of mutation (Yaari et al., 2013). (For interpretation of the references to color in figure legend, the reader is referred to the web version of the article.)

selection were even more strongly correlated to the expected pattern of mutation in CDR, thereby validating the model of mutation we used (Yaari et al., 2013) and showing its efficacy even when analyzing a relatively small number of synonymous mutations. Interestingly, we found that the actual fraction of non-synonymous mutations is always higher than expected. This can be seen in Fig. 5 where all points indicating fraction of amino acid change are above the trend line $x = y$ while the synonymous mutations are both above and below this line. This suggests that targeting is indeed focused toward regions of the CDR that are more likely to improve affinity through mutations and that CDRs regularly have positive selection of non-synonymous mutations raising their numbers above the level predicted by the model of mutation targeting.

3.6. *V gene usage in the repertoire is related to the extent of their ability to focus mutations and amino acid change in the CDR*

Different V_H genes have different biases toward changes in the CDR. We measured if the difference in this bias could be related to skews in V_H usage. Specifically, we looked at the repertoires of

Table 1
Correlations values observed in all datasets.

	CDR fraction	Rscore-R	Tscore-T	Sscore-S
Dataset 1 Wu et al., 2012	Rho	0.526133	0.458156	0.760024
	p-Value	<0.01	<0.01	<0.01
Dataset 2 Wu et al., 2010	Rho	0.390807	0.499808	0.640248
	p-Value	0.048378	0.010958	0.010138
Dataset 3 – PNG Wang et al., 2011	Rho	0.382436	0.440454	0.470864
	p-Value	0.0104	<0.01	<0.01
Dataset 4 – AUZ Wang et al., 2011	Rho	0.384182	0.482616	0.593967
	p-Value	0.01439	<0.01	<0.01

Dataset 1 are recombined V_H sequences from 12 healthy adults post influenza vaccination – 6 young from ages 19 to 45 years and 6 old from ages 70 to 89 (Wu et al., 2012).
Dataset 2 are recombined V_H sequences isolated from 3 healthy volunteers aged 21–26 years (Wu et al., 2010).
Dataset 3 are recombined V_H sequences from 14 healthy residents, aged 22–53 years, of Papua New Guinea region (Wang et al., 2011).
Dataset 4 are recombined V_H sequences from the 14 healthy residents of Sydney Australia (Wang et al., 2011).

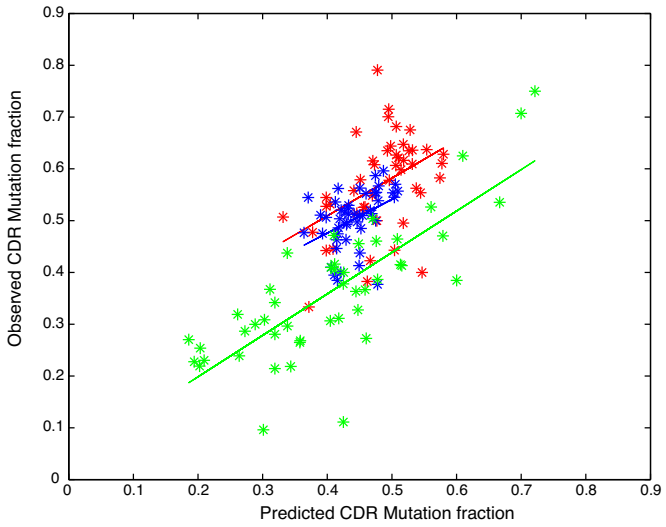

Fig. 5. The germline based model predictions of expected mutation fraction are predictive of the observed mutation in CDR segments of the gene: A plot of the correlation of the predicted CDR fraction of nonconservative changes to observed fraction nonconservative mutations in the CDR, (blue line – $\rho = 0.458$, $p = 1.526e{-}03$). The same for nonsynonymous changes in CDR, (red line – $\rho = 0.526$, $p = 1.817e{-}04$) and for synonymous change in the 4 fold redundant amino acids ["A", "G", "P", "T", "V" where no selection is expected to occur] (green line $\rho = 0.760024$, $p = 5.80e{-}10$). (For interpretation of the references to color in figure legend, the reader is referred to the web version of the article.)

young and elderly adults (see Section 2) at 0, 7 and 28 days post administration of an influenza vaccine (Wu et al., 2012). We found that both in the most expressed V_H genes (Table 2) and in general (Fig. 6), the skew in recombined repertoires is toward V_H genes that have a greater bias in mutation and change in the CDR. So much so, that their average likelihood is higher than we would see if all V_H were used equally in the repertoire (Table 2 and Fig. 6; all individuals, young and old, have repertoires at all days that differ and rank higher than if they used the uniform V gene distribution, Mann–Whitney $p < 0.05$). This finding suggests that the expansion of recombined repertoires depends on bias in mutation and change in the CDR. Visually; we see that in the immune repertoires of the younger cohort, at day 7, the internal repertoire distribution is different from the resting repertoires before (day 0) and after (day 28) the immune response to the vaccination. The skew toward V_H genes that focus mutation on the CDR is relaxed and V_H genes with a wider range of biases are used. This would lead us to predict that at day 7 V_H genes that are less likely to focus change in the CDR and thus would be less likely to survive somatic mutation during affinity maturation are also used to enhance the breadth of the response. Indeed we find that the repertoires at day 28 revert back to the day 0 structure. In older individuals this change for day 7 V_H gene repertoires is much less pronounced (Table 2 and Fig. 6); the repertoires in the older individuals always behave similarly (no consistent trends in comparison between any two time points), and are skewed to use V_H genes with a high bias toward changes in the CDR both at the height of the response and at rest. The fact that this lack of difference is because of a lack of response can be seen in the differences in repertoire sizes we see. In the younger repertoires day 7 is ~1.5 the size in terms of numbers of clones compared to the old repertoires. We did a Wilcoxon signed test to compare across time points to ask if day 0 > day 7 and day 7 < day 28. In all cases the direction is as stated but only day7 <28 is significant. This is not really conclusive as due to our very small n (six individuals) significant difference can only be observed if all but 1 of the individuals behave in the same way. Furthermore, we can not at this n find significant difference between day 0 and 28 (as in this case we have no hypothesis as to trend and must do a two tailed test).

4. Discussion

We show here that somatic mutation targeting together with the nucleotide structure of V genes create an expected pattern of mutations and changes post mutation that are focused on the CDR regions of the V genes (Fig. 1). A significant source of this skew lies in the BCR V genes codon bias. We found this bias to be significantly distinct compared to the codon usage of the SFIg genes, Ig related genes that are not exposed to somatic mutation (Fig. 2). In contrast, we found that while TCR V genes also show skewed expected patters of mutation focused on the CDR, these are far less pronounced than in BCR V genes, and are not based on significant skew of codon usage compared to genes from the SFIg. The lack of codon bias in TCRs leads us to refute previous claims that the bias observed in TCRs is an indication of their using and having evolved to use somatic mutations in their diversification process (Oprea and Kepler, 1999).

The fact that TCR show a skew toward potential mutation in the CDR but lack codon bias suggests to us that somatic targeting evolved to target amino acids that are common to the CDR regions. Specifically the NSDG amino acid transition neighborhood (Hershberg and Shlomchik, 2006) (see Supplemental Fig. 1), many of these amino acids are encoded by the known mutational hotspot RGY (Shapiro et al., 1999). Thus TCRs bias toward the CDR in their expected pattern of mutation may be an artifact of their similarity of structure to BCR. Further strengthening this statement we see that pseudogenes related to the BCR V genes exhibit less codon bias than their related V genes (Supplemental Fig. 3). Finally, the heavy chains of different species all show similar amino acid usage and some skew toward mutation in the CDR. Moreover, those species that show the greatest skew of mutation to the CDR exhibit a complimentary greater bias in codon usage (Supplemental Fig. 4). Interestingly, the greatest codon bias and greatest skew toward

Table 2
CDR fraction of scores top 10 V genes used in Young and Old repertoire (ratio compared to fraction at day 0).

Gene	Mscore CDR fraction[a]	Rscore CDR fraction	Tscore CDR fraction	Old day 0	Old day 7 (ratio)	Old day 28 (ratio)	Young day 0	Young day 7 (ratio)	Young day 28 (ratio)
VH3-74	**0.3978**	**0.4211**	**0.4807**	0.035	0.082 (2.32)	0.077 (2.18)	0.0317	0.057 (1.79)	0.028 (0.87)
VH3-23	**0.3919**	**0.4065**	**0.4611**	0.137	0.174 (1.27)	0.147 (1.07)	0.1705	0.086 (0.50)	0.171 (1)
VH3-30	**0.3900**	**0.4104**	**0.4183**	0.070	0.048 (0.69)	0.069 (0.99)	0.0679	0.056 (0.82)	0.085 (1.26)
VH4-39	**0.3832**	**0.4005**	**0.498**	0.038	0.034 (0.90)	0.047 (1.23)	0.0241	0.028 (1.14)	0.035 (1.44)
VH3-7	**0.3814**	**0.4029**	**0.4231**	0.032	0.111 (3.41)	0.030 (0.91)	0.0267	0.054 (2.03)	0.023 (0.87)
VH4-59	0.3659	0.3602	**0.4412**	0.041	0.085 (2.10)	0.052 (1.29)	0.0448	0.135 (3.01)	0.049 (1.09)
VH4-b	0.3633	0.366	**0.452**	0.004	0.010 (2.40)	0.001 (0.33)	0.0096	0.089 (9.30)	0.018 (1.90)
VH1-69	0.3542	0.3549	0.3421	0.041	0.036 (0.87)	0.022 (0.55)	0.0523	0.088 (1.69)	0.031 (0.60)
VH6-1	0.3363	0.3525	**0.4172**	0.011	0.042 (3.76)	0.003 (0.23)	0.0226	0.094 (4.14)	0.018 (0.81)
VH2-5	0.3212	0.3252	0.3721	0.016	0.006 (0.38)	0.013 (0.81)	0.0347	0.037 (1.07)	0.026 (0.76)

[a] Fractions in bold are of V_H genes whose fraction of expected mutation in the CDR is above the mean of the repertoire if all V genes were used equal.

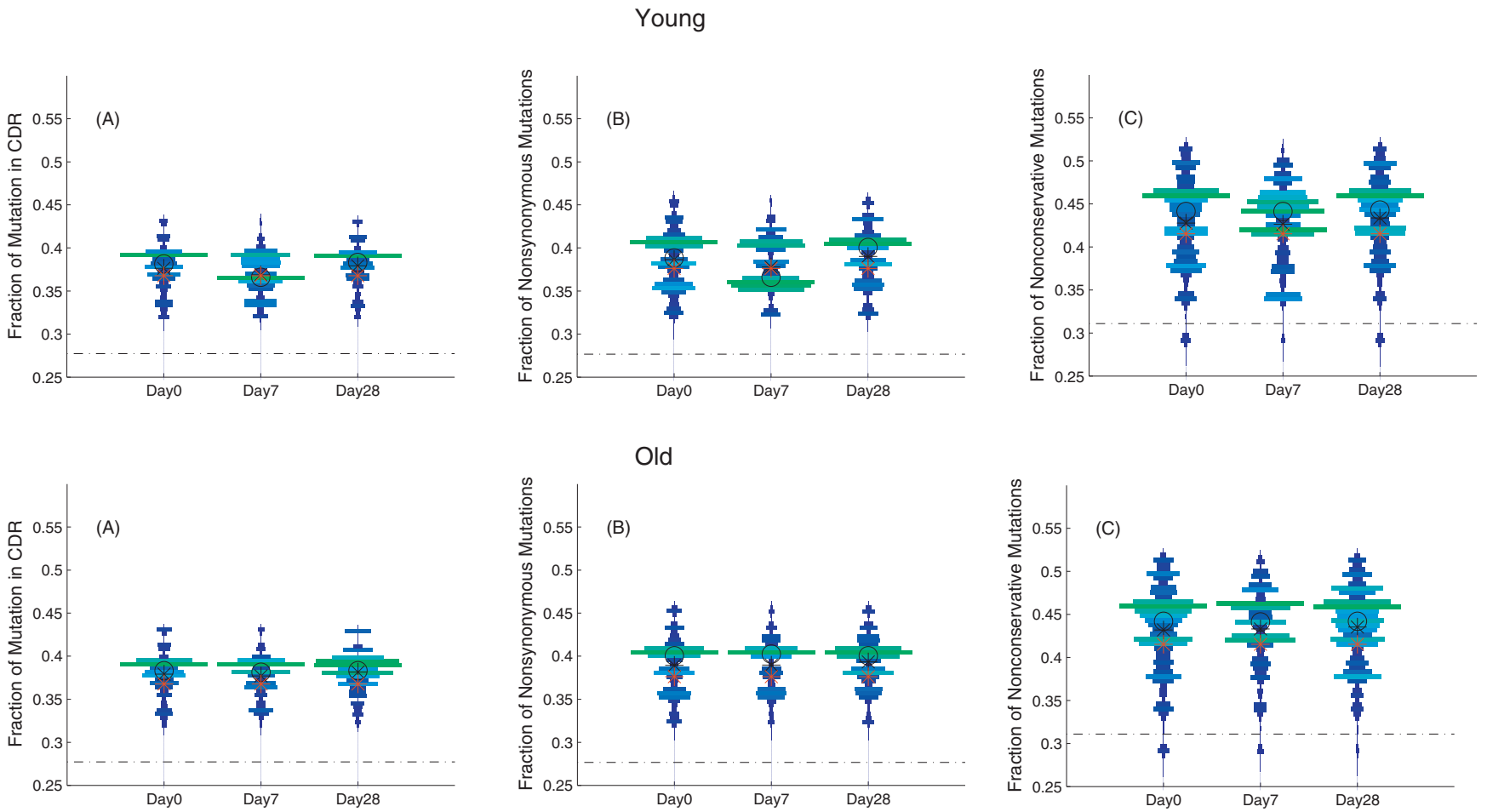

Fig. 6. Comparison of V_H gene usage post immunization and their distribution amongst different levels of bias for change in the CDR. V_H gene distribution is shown in young (top) and old (bottom) populations at day 0, 7 and 28 post immunization. Greater spread represents higher usage of the respective V_H gene with the specific expected fraction of (a) mutations in the CDR, (b) amino acid changes following mutation in the CDR and (c) nonconservative changes in the CDR. The dashed line represents the actual CDR fraction. The red star is the mean if V_H genes were all used equally and black star is the mean when V_H gene usage is taken into account. (For interpretation of the references to color in figure legend, the reader is referred to the web version of the article.)

focusing on the CDR is evidenced in sheep, which are a species that use mutation to generate even their initial naïve repertoires (Jenne et al., 2006; Reynaud et al., 1995).

The methodology we used here depends on contrasts between CDR and FR to describe the impact of codon bias and targeting on the uniformity of changes. It is known that κ light chains lack the distinction of FR vs. CDR, as mutable regions in κ light chains have been observed also in the FR (Kepler, 1997; Hershberg and Shlomchik, 2006). To show that nonetheless κ light chains have hotspot-like structures and a focused region of mutation (and TCRs do not) we compared the CDF of the average positional fractions of mutations (see Section 2). All BCR V genes, including those of κ chains showed a very similar distribution in which a few positions accounted for most of the expected fraction of mutations (0.5 of estimated mutations arose from 0.25 of positions in BCR V genes in contrast to 0.3 of positions in Vβ and 0.38 in Vα – Fig. 3).

We next analyzed the behavior of CDR and FW separately Surprisingly, we found that the CDR of BCRs focused mutations in such a way that total nonsynonymous mutations where actually less abundant than at random but that the fraction of nonconservative mutations was increased. This was not the case in TCRs where all types of nonsynonymous mutations where less than would be expected if all parts of the genetic code were used equally (Fig. 4).

To compliment our analysis of germline genes we validated the expected patterns of skew toward the CDR we had predicted on a diverse set of recombined human heavy chain sequences (Wu et al., 2012, 2010; Wang et al., 2011). As expected, the predicted pattern of mutations in the different V genes had the highest correlations when we considered only the mutations at the 4-fold redundant amino acid positions. This can be considered a further validation of the mutation targeting model we have used in our analysis which despite being created from a large set of mutations taken without any specific bias to area and in a very different study and sequencing methodology, predicts mutation patterns even when looking at a much more limited number of mutations and asking a question about how mutations focus in a specific region. Significant correlations where shown also in the fraction of changes post mutation seen in the CDR. As expected, since these biases are also dependent on selection (Schwartz and Hershberg, 2013) these correlations were less pronounced (Table 1). However, as our hypothesis of evolving germline codon bias to benefit from mutation would predict, our correlations appear to indicate a synergistic effect between the targeting of mutation to CDR and selection. This can be seen by the fact that the expected mutation fractions for non-synonymous and non-conservative mutations are consistently lower than the actual fractions (the corr. line is always above the diagonal and the residual values are skewed toward the positive). In contrast, the expected values for the synonymous mutations at the 4 fold redundant positions, which could not possibly undergo any kind of selection, are sometimes more and sometimes less than their prediction (Fig. 5).

Finally, we considered if the measures suggested here could explain repertoire changes observed post vaccination. It has been observed that in older patients (ages 70–89), the immune repertoire does not change in response to vaccinations as it does in younger people (ages 19–45)(Wu et al., 2012). We found that in both populations at resting (day of vaccination) the repertoire of V genes is skewed such that the BCRV_H genes with the highest skew toward focusing mutations and changes due to mutations in CDR are over represented. However, at day 7 post vaccination, which is the height of the response to the vaccine, the younger patients showed a significant shift toward the use of genes with less of focus toward the CDR while the older patients did not. At day 28 post vaccination, when presumably the response to vaccine was over, both groups of patients returned to the same distribution of V genes and tendency to focus mutations to the CDR as at day 0 (Fig. 6 and Table 2).

The small sample size of this study (Wu et al., 2012) (compared to the total repertoires of these 12 individuals) limits the extent we can consider them to be representative of the entire human population and of all immune response. Furthermore, there could be many other reasons why V gene usage would be skewed. Nonetheless, the differences we see between young and old and at different time points post vaccination, suggests to us that indeed the focusing in the CDR is important for long-term survival of clones into memory. Thus clones using V genes that are more 'adaptive' will expand more and come to dominate the repertoire. Our final observation is not as conclusive, due to the small number of individuals in this study (Wu et al., 2012), which preclude statements on its significance. However, we do see trends that in an acute immune response all V genes, regardless of bias toward mutation in the CDR, are used and needed, however, they do not survive as well into the memory pool. In the older immune system where the generation of novel naïve cells is less and the repertoire is more clonal it is more difficult to muster the cells whose V genes did not focus mutations as strictly to the CDR. For this reason the response at day 7 is not as skewed, is smaller, and potentially not as effective for responses to novel signals.

In combination, this analysis of human germline V genes and the outcomes of mutations in recombined sequence repertoires shows that the nucleotide sequence has evolved to focus the somatic targeting of mutations to the CDR and to create nonconservative changes. We have further shown that this skew in mutation influences the eventual patterns of mutation, potentially in a synergistic way with the forces of selection (Cowell et al., 1999; Schwartz and Hershberg, 2013). Finally, we have suggested one way that the range in the skew toward changes in the CDR, that we find in different BCR V_H genes, may reflect the different role genes play at different stages of an immune response and their efficiency over a single response and a lifetime.

Acknowledgements

The authors wish to thank Deborah Dunn-Walters and Andrew Collins for providing access to their sequences data and aiding in its interpretation and Anat Ninio and the two reviewers for their many helpful and insightful comments, corrections and questions. Research reported in this publication was supported by the National Institute of Allergy and Infectious Diseases of the National Institutes of Health under Award Number P01AI106697. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health

References

Artavanis-Tsakonas, K., Tongren, J.E., Riley, E.M., 2003. The war between the malaria parasite and the immune system: immunity, immunoregulation and immunopathology. Clin. Exp. Immunol. 133, 145–152.

Baumgarth, N., et al., 1999. Innate and acquired humoral immunities to influenza virus are mediated by distinct arms of the immune system. Proc. Natl. Acad. Sci. U. S. A. 96, 2250–2255.

Betz, A.G., Neuberger, M.S., Milstein, C., 1993a. Discriminating intrinsic and antigen-selected mutational hotspots in immunoglobulin V genes. Immunol. Today 8, 405–411.

Betz, A.G., Rada, C., Pannell, R., Milstein, C., Neuberger, M.S., 1993b. Passenger transgenes reveal intrinsic specificity of the antibody hypermutation mechanism: clustering, polarity, and specific hot spots. Proc. Natl. Acad. Sci. U. S. A. 90, 2385–2388.

Bork, P., Holm, L., Sander, C., 1994. The immunoglobulin fold. Structural classification, sequence patterns and common core. J. Mol. Biol. 242, 309–320.

Brümmendorf, T., Rathjen, F.G., 1995. Cell adhesion molecules. 1: Immunoglobulin superfamily. Protein Profile 2, 963.

Burnet, F.M., 1957. Clonal selection theory: a modification of Jerne's theory of antibody production using the concept of clonal selection. Aust. J. Sci. 20, 67–69.

Burnet, F.M., 1959. The Clonal Selection Theory of Acquired Immunity.

Chang, B., Casali, P., 1994. The CDR1 sequences of a major proportion of human germline Ig VH genes are inherently susceptible to amino acid replacement. Immunol. Today 15, 367–373.
Chothia, C., Gelfand, I., Kister, A., 1998. Structural determinants in the sequences of immunoglobulin variable domain. J. Mol. Biol. 278, 457–479.
Cowell, L.G., Kim, H.J., Humaljoki, T., Berek, C., Kepler, T.B., 1999. Enhanced evolvability in immunoglobulin V genes under somatic hypermutation. J. Mol. Evol. 49, 23–26.
Early, P., Huang, H., Davis, M., Calame, K., Hood, L., 1980. An immunoglobulin heavy chain variable region gene is generated from three segments of DNA: V, D and J. Cell 19, 981–992.
Hershberg, R., Petrov, D.A., 2008. Selection on codon bias. Annu. Rev. Genet. 42, 287–299.
Hershberg, U., Shlomchik, M.J., 2006. Differences in potential for amino acid change after mutation reveals distinct strategies for kappa and lambda light-chain variation. Proc. Natl. Acad. Sci. U. S. A. 103, 15963–15968, http://dx.doi.org/10.1073/pnas.0607581103.
Jenne, C.N., Kennedy, L.J., Reynolds, J.D., 2006. Antibody repertoire development in the sheep. Dev. Comp. Immunol. 30, 165–174.
Kabat, E.A., Wu, T.T., Perry, H.M., Gottesman, K.S., Foeller, C., 1991. Sequences of Proteins of Immunological Interest. US Dept. of Health and Human Services, US Government Printing Office.
Kepler, T.B., 1997. Codon bias and plasticity in immunoglobulins. Mol. Biol. Evol. 14, 637–643.
Kepler, T.B., et al., 2014. Reconstructing a B-cell clonal lineage II. Mutation, selection, and affinity maturation. Front. Immunol. 5.
Kersey, P.J., et al., 2014. Ensembl genomes 2013: scaling up access to genome-wide data. Nucleic Acids Res. 42, D546–D552.
Lefranc, M.P., Pommie, C., Ruiz, M., Giudicelli, V., 2003. IMGT unique numbering for immunoglobulin and T cell receptor variable domains and Ig super family V-like domains. Dev. Comp. Immunol. 27, 55–77.
McHeyzer-Williams, L.J., McHeyzer-Williams, M.G., 2005. Antigen-specific memory B cell development. Annu. Rev. Immunol. 23, 487–513.
Miyata, T., Hayashida, H., Yasunaga, T., Hasegawa, M., 1979. The preferential codon usages in variable and constant regions of immunoglobulin genes are quite distinct from each other. Nucleic Acids Res. 7, 2431–2438.
Nakamura, Y., Gojobori, T., Ikemura, T., 2000. Codon usage tabulated from the international DNA sequence databases: status for the year 2000. Nucleic Acids Res. 28, 292.
National Institute of Genetics, Flat File Release 160.0 (15.06.07), http://www.kazusa.or.jp/codon/
Neuberger, M.S., Milstein, C., 1995. Somatic hypermutation. Curr. Opin. Immunol. 7, 248–254.
Oprea, M., Kepler, T.B., 1999. Genetic plasticity of V genes under somatic hypermutation: statistical analyses using a new resampling-based methodology. Genome Res. 9, 1294–1304, http://dx.doi.org/10.1101/gr.9.12.1294.
Plotkin, J.B., Kudla, G., 2011. Synonymous but not the same: the causes and consequences of codon bias. Nat. Rev. Genet. 12, 32–42.
Reynaud, C.A., Mackay, C.R., Müller, R.G., Weill, J.C., 1991. Somatic generation of diversity in a mammalian primary lymphoid organ: the sheep ileal Peyer's patches. Cell 64, 995–1005.
Reynaud, C.-A., Garcia, C., Hein, W.R., Weill, J.-C., 1995. Hypermutation generating the sheep immunoglobulin repertoire is an antigen-independent process. Cell 80, 115–125.
Schwartz, G.W., Hershberg, U., 2013. Germline amino acid diversity in B cell receptors is a good predictor of somatic selection pressures. Front. Immunol. 4, 357, http://dx.doi.org/10.3389/fimmu.2013.00357.
Shapiro, G., Aviszus, K., Ikle, D., Wysocki, L., 1999. Predicting regional mutability in antibody V genes based solely on di- and tri-nucleotide sequence composition. J. Immunol. 163, 259–268.
Shapiro, G.S., Aviszus, K., Murphy, J., Wysocki, L.J., 2002. Evolution of Ig DNA sequence to target specific base positions within codons for somatic hypermutation. J. Immunol. 168 (2302), 2306.
Tonegawa, S., 1983. Somatic generation of antibody diversity. Nature 302, 575–581.
Wagner, S.D., Milstein, C., Neuberger, M.S., 1995. Codon bias targets mutation. Nature 376, 732.
Wang, Y., et al., 2011. IgE sequences in individuals living in an area of endemic parasitism show little mutational evidence of antigen selection. Scand. J. Immunol. 73, 496–504.
Weill, J.C., Reynaud, C.A., 1996. Rearrangement/hypermutation/gene conversion: when, where and why? Immunol. Today 17, 92–97.
Williams, A.F., Barclay, A.N., 1988. The immunoglobulin superfamily-domains for cell surface recognition. Annu. Rev. Immunol. 6, 381–405.
Wu, T.T., 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, 211–250.
Wu, Y.C., et al., 2010. High-throughput immunoglobulin repertoire analysis distinguishes between human IgM memory and switched memory B-cell populations. Blood 116, 1070–1078.
Wu, Y.C., Kipling, D., Dunn-Walters, D.K., 2012. Age-related changes in human peripheral blood IgH repertoire following vaccination. Front. Immunol. 3, 193, http://dx.doi.org/10.3389/fimmu.2012.00193.
Yaari, G., et al., 2013. Models of somatic hypermutation targeting and substitution based on synonymous mutations from high-throughput immunoglobulin sequencing data. Front. Immunol. 4, 358, http://dx.doi.org/10.3389/fimmu.2013.00358.