Antibody repertoire deep sequencing reveals antigen-independent selection in maturing B cells

Joseph Kaplinsky\(^a,b\), Anthony Liu\(^a,b\), Amy Sun\(^c\), Maryaline Coffre\(^c\), Sergei B. Korolov\(^c,1\), and Ramy Arnaout\(^a,b,d,1,2\)

\(^a\)Department of Pathology and \(^b\)Division of Clinical Informatics, Department of Medicine, Beth Israel Deaconess Medical Center, Boston, MA 02215; \(^c\)Department of Systems Biology, Harvard Medical School, Boston, MA 02115; and \(^d\)Department of Pathology, New York University Medical Center, New York, NY 10016

Antibody repertoires are known to be shaped by selection for antigen binding. Unexpectedly, we now show that selection also acts on a non–antigen-binding antibody region: the heavy-chain variable (V\(_H\))–encoded “elbow” between variable and constant domains. By sequencing 2.8 million recombined heavy-chain genes from immature and mature B-cell subsets in mice, we demonstrate a striking gradient in V\(_H\) gene use as pre-B cells mature into follicular and then into marginal zone B cells. Cells whose antibodies use V\(_H\) genes that encode a more flexible elbow are more likely to mature. This effect is distinct from, and exceeds in magnitude, previously described maturation-associated changes in heavy-chain complementarity determining region 3, a key antigen-binding region, which arise from junctional diversity rather than differential V\(_H\) gene use. Thus, deep sequencing reveals a previously unidentified mode of B-cell selection.

The mature antibody repertoire is shaped by selective forces that influence B-cell survival (1). Comparison of immature and mature B-cell subsets has shown that selection acts specifically on complementarity determining region 3 (CDR3) of the antibody heavy-chain molecule, an antigen-binding region that is a key determinant of antigen specificity (2). On average, mature B-cell subsets express antibodies that have shorter and more negatively charged CDR3s, which is the result of selection against autoreactive and polyreactive B cells (3, 4).

Each recombined antibody heavy-chain gene is composed of a variable (V\(_H\)), diversity (D), and joining (J\(_H\)) gene segment. Because the CDR3 region spans the V\(_{1\text{H}}\)-D-J\(_{1\text{H}}\) joint, investigators have asked whether selection might favor B cells whose antibodies use specific V\(_{1\text{H}}\), D, or J\(_{1\text{H}}\) gene segments. Selection in favor of specific gene segments during B-cell maturation might help to explain the observed maturation-associated changes in CDR3 length and charge, and might suggest a preference for “hard-wired” antigen specificities. Evidence against selection would suggest that differences in CDR3 result exclusively from the nontemplated addition and deletion of nucleotides at the V\(_{1\text{H}}\)-D and D-J\(_{1\text{H}}\) junctions, and therefore that the death of B cells with counterselected CDR3s during maturation is simply the evolutionary cost (3) of maintaining this mechanism of generating antibody diversity.

Nearly two decades ago, a low-throughput sequencing study in mice suggested that specific V\(_{1\text{H}}\) gene segments were used at different frequencies by pre-B cells (in which heavy-chain recombination has been completed) and mature B cells in the spleen (5). This observation was interpreted as selection for hard-wired specificities. However, the statistical robustness of this observation was limited by the small number of recombined genes that were sequenced, and although subsequent investigations have detected differences in V\(_{1\text{H}}\) use between pre-B and upstream pro-B cells, they have failed to confirm such differences between pre-B and mature B cells (6).

More recent sequencing studies have looked for differences in V\(_{1\text{H}}\) gene segment use between follicular (FO) B cells, which circulate through the spleen, and marginal zone (MZ) B cells, “innate-like” (7) cells in the spleen that are theorized not to undergo B-cell receptor (BCR)-based selection during maturation (8–11). These studies also found no differences in V\(_{1\text{H}}\) use, although, again, the number of sequences studied was small. In contrast, studies have found differences in D segment use between antibodies with short and long CDR3s, including the appearance of tandem D gene segments in very long CDR3s, suggesting a potential causative relationship (4, 12). However, these studies did not investigate whether these differences were a sign of selection during B-cell maturation.

The advent of high-throughput antibody repertoire sequencing makes it possible to investigate the forces that govern B-cell selection with statistical rigor (13, 14). This is what we set out to do.

**Results**

We sorted pre-B, FO, and MZ cells from the bone marrow and spleen of nine unimmunized 2.5-month-old C57BL/6J mice according to surface expression of CD19, B220, CD21, CD23, and IgM (Fig. S1). We prepared heavy-chain libraries from genomic DNA and sequenced V(D)J-recombined genes for 2.8 million recombination events at 4.3-fold coverage. We annotated reads for V\(_{1\text{H}}\), D, and J\(_{1\text{H}}\) gene segment use; CDR3 length and charge; and V\(_{1\text{H}}\)-D, D-J\(_{1\text{H}}\), and total junctional nucleotides; as well as whether the reads were productive (full-length and free of stop codons) or nonproductive, as previously described (14, 15).

**Significance**

Antibodies play essential roles in vaccination, infection, autoimmunity, aging, and cancer. A key question is how the antibody repertoire achieves its remarkable diversity. Part of the answer is that B cells, which express antibodies on their surface, are selected for survival based on the specific antigens that their antibodies bind, with antigen specificity determined by the protein sequence of antibodies’ antigen-binding regions. Unexpectedly, we find that B cells are also selected based on whether their antibodies have a loose or tight “elbow joint,” independent of the sequence of their antigen-binding regions. This discovery, enabled by sequencing technology and mathematics, adds a surprising new dimension to our understanding of antibody repertoires, and might one day help us shape them ourselves.


*The authors declare no conflict of interest.*

*This Direct Submission article had a prearranged editor.*

*Data deposition: The antibody heavy-chain sequences reported in this paper have been deposited in the Sequence Read Archive, www.ncbi.nlm.nih.gov/sra (submission no. 432110; NCBI BioProject number PRJNA248676).*

*S.B.K. and R.A. contributed equally to this work.*

*To whom correspondence should be addressed. E-mail: rarnaout@gmail.com.*

This article contains supporting information online at www.pnas.org/cgi/doi/10.1073/pnas.1403278111

www.pnas.org/cgi/doi/10.1073/pnas.1403278111

PNAS Early Edition | 1 of 8
Similarity among V_H and D gene segments and junctional diversity at the V_H-D and D-J_H junctions preclude unique annotation of all reads (16); we obtained unique V_H and J_H gene segment annotations in 89% and 100% of reads, respectively, and unique V_J and VDJ combinations in 89% and 63% of reads, respectively (Table 1). The remaining reads were excluded from subsequent analysis of gene segment use. Comparison of 32 pairs of PCR replicates showed excellent reproducibility for frequency of gene segment use and for CDR3 length and charge (Fig. S2). V_H, D, and J_H gene segment use among mature subsets also agreed well with our previous high-throughput study of mature subsets, which used a different primer set (14).

Junctional Diversity and Differential D Gene Segment Use Account for Most of the Trend Toward Shorter CDR3s During B-Cell Maturation. We first measured differences in CDR3 length and charge in pre-B, FO, and MZ subsets, treating productive rearrangements, which contribute to functional BCRs and are thereby subject to selection, separately from nonproductive rearrangements, which are not subject to selection.

CDR3 lengths followed an approximately normal (Gaussian) distribution among both productive and nonproductive rearrangements in all three B-cell subsets (Fig. 1A). Distributions largely overlapped. Among productive rearrangements, pre-B and FO subsets had the longest CDR3s, at 13.3 ± 3.1 aa and 13.4 ± 2.8 aa, respectively (Fig. 1A, Right). Although their means were statistically indistinguishable (P = 0.27), the length distribution in FO subset was significantly narrower (P = 1.8 × 10⁻⁶), suggesting selection against extreme-length CDR3s during maturation from pre-B to FO cells. Meanwhile, MZ CDR3s were, on average, 12.7 ± 2.8 aa long, or about one-half of an amino acid shorter than in either pre-B or FO cells (P = 3.1 × 10⁻⁸ and P = 8.0 × 10⁻¹⁰, respectively), consistent with selection with differences between MZ and FO cells in the rat model (11) and, again, significantly narrower than in pre-B cells (P = 3.8 × 10⁻⁶).

Junctional diversity accounted for 49% of the length difference between MZ and pre-B cells (Fig. 1B). Differential D gene segment use accounted for approximately one-third (36%) of the difference, whereas differential V_H and J_H gene segment use played minor roles at 8% and 7%, respectively.

In contrast, CDR3 lengths for nonproductive rearrangements, which are not under selection, were statistically identical in pre-B, FO, and MZ cells (13.4 ± 4.2 aa, 13.1 ± 3.9 aa, and 13.2 ± 4.4 aa, respectively; Fig. 1A, Left). Interestingly, the mean CDR3 length for nonproductive rearrangements in all three subsets was indistinguishable from the means for productive rearrangements in pre-B and FO cells (P > 0.14 for pairwise comparisons), but the length distributions were wider for nonproductive rearrangements than for productive rearrangements (P = 5.1 × 10⁻⁹ to 6.2 × 10⁻¹¹ for pairwise comparisons between nonproductive and productive rearrangements within each subset; compare Fig. 1A, Right vs. Left), suggesting selection against extreme-length CDR3s even in pre-B cells.

Junctional Diversity and Differential D Gene Segment Use also Account for Most of the Trend Toward More Negatively Charged CDR3s During B-Cell Maturation. CDR3 charge was also approximately normally distributed in all three subsets (Fig. 1C). Again, distributions largely overlapped. The mean charge across animals ranged from -0.18 to -0.48 per CDR3 for productive rearrangements, compared with +0.92 to +0.96 predicted for nonproductive rearrangements (P = 4.6 × 10⁻²⁴ to 5.7 × 10⁻¹⁷), consistent with previous reports (3, 14). Among productive rearrangements, CDR3s from FO and MZ cells were, on average, 0.13-0.15 unit more negative than those from pre-B cells (P = 3.2 × 10⁻⁶ and 8.4 × 10⁻⁹, respectively; Fig. 1C, Right).

As with CDR3 length, the distribution of charges was slightly narrower among productive rearrangements than among nonproductive rearrangements (P = 6.8 × 10⁻⁶ to 3.0 × 10⁻¹⁴), consistent with selection against extremes of charge in addition to selection for more negative charge. Among productive rearrangements, the distribution of charges in MZ and FO cells was narrower than in pre-B cells (P = 2.4 × 10⁻¹¹ and 3.5 × 10⁻⁹, respectively; Fig. 1C). Also, as with CDR3 length, junctional diversity was largely responsible for the charge difference between pre-B and MZ cells, accounting for 51% of the difference, followed by D gene segment use (34%), with V_H accounting for only 4% (Fig. 1D).

Selection for V_H Gene Segment Use, Not D Gene Segment Use, Drives VDJ Differences Between Immature and Mature B-Cell Subsets. Although junctional diversity was the single largest contributor to differences in both CDR3 length and charge between immature and mature B-cell subsets, differential D gene segment use also played a sizeable role, accounting for approximately one-third of the observed differences (Fig. 1B and D; see above). Therefore, we asked whether differential D use might also result in differences in the frequencies of VDJ combinations during B-cell maturation.

We used principal component analysis (PCA) to analyze and visualize the overall variability in VDJ use among productive rearrangements without presupposing that there would be a relationship between pre-B, FO, and MZ cells. In PCA, variability across multiple variables (dimensions), such as the use frequency of many different gene segments, is summarized by combining variables that happen to be correlated with each other into “components” that are then ordered by how much of the total

Table 1. VDJ assignment statistics

<table>
<thead>
<tr>
<th>Subset</th>
<th>Reads</th>
<th>Clusters</th>
<th>Unique V_H</th>
<th>Unique D</th>
<th>Unique J_H</th>
<th>Unique VDJ</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pre-B</td>
<td>Mean</td>
<td>328,636</td>
<td>64,825</td>
<td>89</td>
<td>70</td>
<td>100</td>
</tr>
<tr>
<td></td>
<td>SD</td>
<td>79,160</td>
<td>29,322</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>FO</td>
<td>Mean</td>
<td>658,595</td>
<td>194,862</td>
<td>90</td>
<td>70</td>
<td>100</td>
</tr>
<tr>
<td></td>
<td>SD</td>
<td>282,083</td>
<td>142,414</td>
<td>2</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>MZ</td>
<td>Mean</td>
<td>342,505</td>
<td>46,073</td>
<td>88</td>
<td>67</td>
<td>100</td>
</tr>
<tr>
<td></td>
<td>SD</td>
<td>110,880</td>
<td>25,441</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Total</td>
<td>Mean</td>
<td>11,967,624</td>
<td>2,751,840</td>
<td>89</td>
<td>69</td>
<td>100</td>
</tr>
</tbody>
</table>

A total of 12.0 million reads representing 2.8 million recombination events (clusters) were obtained from pre-B, FO, and MZ B-cell subsets from each of nine mice, of which an average of 89%, 69%, and 100% were able to be annotated with a unique V_H, D, and J_H gene segment, respectively.
pre-B, FO, and MZ cells reflect VDJ rearrangements made at the pro-B stage, and are thus not subject to selection. We therefore grouped nonproductive rearrangements from these three subsets together as a proxy for pro-B cells and asked whether variation in VH, D, JH, and VDJ use in pre-B, FO, and MZ cells and these pooled nonproductive rearrangements is still primarily defined by VH use, and if so, whether nonproductive rearrangements cluster along the same axis, “upstream” of pro-B cells. We found that they do: Nearly three-quarters of the variability in VH gene segment use is explained by the first two principal components, which again cluster samples by subset (Fig. 2F).

The first principal component again separates subsets in order of their lineage relationships, with nonproductive rearrangements clustering upstream (i.e., to the left) of pre-B cells on the previously seen pre-B–MZ gradient. A shuffled-U test confirmed that all four of these subsets differed significantly from each other in VH use ($P \leq 2.0 \times 10^{-8}$). These four subsets also differed in D use, albeit more weakly ($P \leq 8.8 \times 10^{-3}$), but not in JH use.

**B-Cell Maturation Selects for Specific V<sub>H</sub> Gene Segments.** In PCA of VH use, each principal component summarizes the use of multiple different V<sub>H</sub> gene segments. To investigate how the clustering of different B-cell subsets along PC1 might relate to selection for one or more specific gene segments, we next asked how use of specific V<sub>H</sub> gene segments contributed to PC1 of productive rearrangements (Fig. 2C).

**Fig. 1.** CDR3 length and charge distributions for productive and nonproductive rearrangements in pre-B, FO, and MZ cells and source of differences in productive rearrangements. (A) Distributions (histograms) of CDR3 length for each subset in nonproductive (Left) and productive (Right) rearrangements. Arrowheads denote means. The asterisk indicates a statistically significant difference in mean length between MZ and pre-B subsets ($P = 3.1e^{-4}$) and between MZ and FO subsets ($P = 8.0 \times 10^{-10}$). (B) Relative contributions of differences in junctional diversity, V, D, and J use to CDR3 length differences between pre-B and MZ subsets. Arrowheads denote means. The asterisk indicates a statistically significant difference in means between pre-B and FO subsets ($P = 3.2 \times 10^{-4}$) and between pre-B and MZ subsets ($P = 8.4 \times 10^{-6}$). (C) Distributions of calculated CDR3 charge for each subset in nonproductive (Left) and productive (Right) rearrangements. Arrowheads denote means. The asterisk indicates a statistically significant difference in means between pre-B and FO subsets ($P = 7.9e^{-4}$) and between pre-B and MZ subsets ($P = 8.8 \times 10^{-6}$). (D) Relative contributions of differences in junctional diversity and V, D, and J use to CDR3 charge differences between pre-B and MZ subsets.

variability in the data they account for. Thus, the first two principal components, called PC1 and PC2, always account for the largest portion of the variability. Consequently, plotting data against PC1 as the x axis and PC2 as the y axis makes it possible to visualize variability of samples across many variables in a single x-y scatterplot.

PCA of VDJ use showed a clear separation between pre-B, FO, and MZ subsets (Fig. 2A). Surprisingly, the pattern mirrored these subsets’ lineage relationships (Fig. 2B). To test whether this pattern was driven by differences in D gene segment use between the subsets, we next performed PCA separately on VH, D, and JH gene segment use. If driven by D use, PCA of D use would display a pattern at least as pronounced as that for VDJ. Unexpectedly, PCA revealed stark separation between pre-B, FO, and MZ cells in VH use (Fig. 2C) but only weak separation in D use (Fig. 2D) and no separation in JH use (Fig. 2E). PCA of VH use in nonproductive rearrangements, which do not undergo selection, showed no separation (Fig. S3), confirming that separation among productive rearrangements reflects selection for VH gene use.

It was interesting that PC1 of VH gene segment use defined a gradient from pre-B, to FO, to MZ cells (Fig. 2C, left to right), because there was no prior assumption of a relationship between selection at the pre-B–FO and FO–MZ transitions. We therefore asked whether this gradient might extend still earlier in the maturation process, to pro-B cells. Nonproductive rearrangements in
The 10 \( V_H \) gene segments that contributed most to PC1 accounted for nearly one-half (49%) of the variability in \( V_H \) gene segment use; the rest of the variability was spread among the remaining 90 \( V_H \) gene segments. Therefore, we focused our analysis on these “top 10” \( V_H \) gene segments: V1–81, V1–53, V8–8, V1–80, V1–82, V1–5, V1–19, V1–52/1–61/1–69, V1–26/1–34, and V1–75 (Fig. 3; the slashes denote gene segments that could not be distinguished from each other, given our sequencing read lengths; for simplicity, we henceforth refer to these equivalence classes as single “gene segments”). We note that these top 10 segments are all frequently used gene segments: Even though they comprise only 10% of all \( V_H \) gene segments, they account for 31% of all recombination events in pre-B, FO, and MZ cells (Fig. 3).

Use of these top 10 \( V_H \) gene segments differed in pre-B, FO, and MZ cells, with five gene segments—V1–81, V8–8, V1–5, V1–19, and V1–75—being more frequent in pre-B than in MZ cells and the other five—V1–53, V1–80, V1–82, V1–52/1–61/1–69, and V1–26/1–34—being more frequent in MZ than in pre-B cells (Fig. 3B). In every case but one (V1–52/1–61/1–69, in which there was overlap with MZ cells), \( V_H \) gene segment use in FO cells fell between that of pre-B and MZ cells, consistent with the pre-B–FO–MZ ordering of these three subsets along PC1 (Fig. 2C). When we included pooled-nonproductive rearrangements as a proxy for pre-B cells in the analysis, the top 10 \( V_H \) gene segments for PC1 included several of the same gene segments (V1–82, V1–53, and V1–26/1–34; Fig. 3C). For each of these gene segments, as well as for V1–50/1–59, use frequency among nonproductive rearrangements was lower than in pre-B cells, again showing a gradient of increasing use of these \( V_H \) gene segments as B cells mature (Fig. 3D).

Interestingly, the rest of the top 10 \( V_H \) gene segments in this expanded nonproductive-rearrangement–containing analysis—V1–52/61/69, V2–2, V5–4, V1–11, V1–62/2, and V2–6—appeared in 15–25% of nonproductive rearrangements but rarely in productive rearrangements from pre-B, FO, or MZ cells (Fig. 3C). In the starkest example, V2–6 accounted for as much as 4.3% of nonproductive rearrangements but was not used once in productive rearrangements in any subset. Inspection revealed that four of these five gene segments ended in an incomplete codon relative to the first base of the recombination signal sequence. Overall, we found that \( V_H \) gene segments, like V2–6, ending in incomplete codons were 26% less likely to lead to productive rearrangements than \( V_H \) gene segments that end in a complete codon [relative risk (RR) = 0.74 ± 0.08].

**Selection on \( V_H \) Acts on Specific Regions, and Sequences, of Framework 1/2 but Not CDR3.** We next asked what sequence-level features of these specific \( V_H \) gene segments might explain the observed gradient. Although overall \( V_H \) gene segment use correlates with gene segment proximity to the D locus in mice (17), this correlation cannot explain the gradient, because choice of \( V_H \) gene segment occurs before selection. An alternative explanation is that selection of \( V_H \) gene segments is related to how they contribute to CDR3, given that, on average, CDR3 properties differ between pre-B and MZ subsets. However, all but two of the top 10 \( V_H \) gene segments from PC1 of Fig. 2C contribute the same three residues to CDR3 (the initial C-A-R). Thus, the observed pattern does not result from \( V_H \) gene segments’ contribution to CDR3, consistent with our finding that \( V_H \) gene segments do not explain the differences between subsets in CDR3 length and charge.

We therefore tested for systematic differences in coding sequence upstream of CDR3. Indeed, alignment of all \( V_H \) gene segments showed systematic bias in amino acid use in two regions: in framework 1 (FR1) at \( V_H \) amino acid positions 10–13 and just distal to CDR1 at positions 37–39 (Fig. 4A and Fig. S4). Analysis of nucleotide sequence revealed no additional regions. We investigated these two regions in detail. Four of the top five \( V_H \) gene segments that were used more frequently in MZ cells had the amino acid sequence E-L-V-K at positions 10–13, compared with only one of the top five \( V_H \) gene segments that were used more frequently in pre-B cells (Fig. 4B). This was not simply because these gene segments belonged to the VH4 family, because many gene segments in the VH1 family do not encode the ELVK motif (Fig. 4B).

To investigate the structural significance of the ELVK motif, we analyzed crystal structures from the Protein Data Bank (www.rcsb.org). These structures showed that positions 10–13, along with highly conserved serine and threonine residues in distal JH, form the socket of a ball-and-socket joint at a region known as the elbow (18), which connects the variable domain and constant domain 1 (Fig. 5A). An Arg or Glu at residue 13, found more frequently in \( V_H \) gene segments expressed in pre-B cells, helps to form a narrower socket (Fig. 5B), whereas the conserved Lys of the ELVK motif, found frequently in \( V_H \) gene segments in MZ cells, forms a wider socket, presumably allowing more flexibility between these two domains in MZ BCRs (Fig. 5C). Interestingly, in several structures, \( V_H \) gene segments that lack ELVK in the germ line had undergone somatic hypermutations that resulted in the presence of K13 and a wider socket, further suggesting a preference for this feature in mature antibodies, such as those in these crystallographic studies. Together, these results suggest that MZ cells and, to a lesser extent, FO cells are selected for \( V_H \) use based on differences in FR1, independent of CDR3.

In the second region, four of the top five \( V_H \) gene segments that were used more frequently in MZ B cells contained a W-M motif at positions 37–38, with two containing a W-M-N motif, whereas only one of the top five \( V_H \) gene segments that were used more frequently in pre-B cells contained a WM motif, and none of these contained a WMN motif (Fig. 4B). In crystal structures, this region localized deep to the CDR1 binding pocket in the interior of the antibody. Unlike ELVK and the elbow between VH4 and heavy-chain constant-1 (CH1) domains, there were no obvious structural differences between antibodies that contained WM(N) and those that did not.

**B-Cell Maturation also Selects for Shorter D Gene Segments.** Although the most prominent gene segment use differences along the pre-B–FO–MZ gradient were for \( V_H \) gene segments (see above), we also asked whether the weaker overall differences in D gene segment use could be traced to specific D gene segments. We found that differences in D gene segment use along the pre-B–FO–MZ gradient were dominated by D1-1 being more frequent in pre-B cells and D4-1 and D3-1 being more frequent in MZ cells (Fig. 6A and B).

These differences were strongly correlated with the length of these gene segments, with D4-1 and D3-1 being the two shortest D gene segments at 16 and 17 nt, respectively, and D1-1 being the longest at 23 nt (Fig. 6 C–E). These results were consistent with our observation above that D gene segment use contributes to differences in CDR3 length between immature and mature B-cell subsets (Fig. 1B).

**Discussion**

The mature B-cell repertoire is the product of selection. Selection acts, at least in part, on the BCR, which, through V(D)J recombination, is the product of gene segment choice, junctional diversity [resulting from template-dependent and template-independent addition of nucleotides (N/P nucleotides) and terminal deoxynucleotidyl transferase (TdT)-mediated nucleotide removal], and the pairing of heavy and light chains. Here, we have investigated how gene segment choice and junctional diversity in the heavy chain influence selection of B cells for survival and inclusion in the mature naive repertoire, using high-throughput sequencing to study the antibody repertoire at its native level of complexity.
Selection acts, at least in part, against BCRs that are auto-
ractive or polyreactive; these reactivities are, in turn, related to the length and charge of heavy-chain CDR3 in both humans and mice (3, 4). Here, we show that in mice, the differences in length and charge between immature and mature subsets result principally from differences in junctional diversity, as opposed to the choice of VH, D, or JH gene segment. Indeed, we find that junctional diversity accounts for one-half of the differences in both length and charge. Regarding gene segment choice, only the choice of D gene segment plays an appreciable role, specifically driven by selection against D1-1, the longest D gene segment (23 nt), and in favor of D4-1, the shortest D gene segment (16 nt), as B cells mature. In contrast, we show that differential VH and JH gene segment choice contribute only marginally to trends in CDR3 length and charge, consistent with their smaller contributions to the amino acid sequence of heavy-chain CDR3.

In quantifying the relative contributions of junctional diversity and gene segment choice to CDR3 length and charge differences during B-cell maturation, we note that the overall magnitude of these differences is small. The distributions from all three subsets largely overlap (Fig. 1 A and C, Lower), and, on average, the heavy-chain CDR3 of an MZ B cell is only one-half of an amino acid shorter and 0.14 unit more negative than that of a pre-B cell. The statistically significant narrowing of both length and charge diverse segments during maturation, which further reflects a preferential loss of B cells with the longest and most positively charged heavy-chain CDR3s, is also slight. In particular, the length difference between pre-B and MZ cells can be accounted for by an average of just one to two extra nucleotides per recombination event. Thus, although differences in CDR3 length and charge represent a clear and statistically significant trend, other factors must play a more prominent role in determining which B cells survive to enter the mature repertoire. The relative weakness of selection for length would be consistent with why, for example, the longest D gene segment (D1-1) continues to be maintained in the genome: Despite experiencing the strongest counter-selection of any D gene segment during maturation, it is still the single most frequently used D gene segment in MZ cells (Fig. 6A). Thus, the benefit of maintaining longer CDR3s in the mature repertoire must outweigh the cost of having to eliminate many D1-1–using B cells during maturation (4).

Fig. 3. Use of the top 10 VH gene segments from PC1 of Fig. 2C. Percentage of use among productive rearrangements from pre-B, FO, and MZ subsets (A) and difference in percentage relative to the percentage of VH use in the pre-B subset (B). (C) Same as in A, but including VH use in pooled nonproductive rearrangements as a proxy for pre-B cells. (D) Same as in B, but the difference is calculated relative to the percentage of VH use in nonproductive rearrangements. Bars indicate ± 1 SD from the mean over all animals. prod., productive.

We also found striking differences in VDJ repertoire among B-cell subsets. Surprisingly, these were driven by differences not in D gene segment use but in VH gene segment use, which was unexpected because VH gene segments have only a minor influence on CDR3 length or charge (or indeed on CDR3 sequence). These differences were largely driven by VH1 family gene segments, consistent with some of the earliest, low-throughput investigations in this area (19). With deep sequencing, we are now able to show that differences in VH use form a gradient from B-cell precursors to mature B-cell subsets that extends from nonproductive rearrangements not subject to selection pressure, which we used as a proxy for pro-B cells, through pre-B, to FO, and finally to MZ cells in a pattern that mirrors the lineage relationships of these subsets (Figs. 2 and 3). This suggests that MZ B cells, which have been thought not to undergo selection on BCR (8–11), are actually subject to the same selective pressure that drives selection of FO cells, in fact to an even greater degree. We note that nonproductive rearrangements are imperfect proxies for pro-B cells; direct deep sequencing of rearrangements in pro-B cells, as has been recently performed (17), offers the possibility of comparing segment use in this population directly with that in pre-B, FO, and MZ cells as data on the frequency of specific gene segments are made available.

By mapping sequence changes onto antibody crystal structures, we identified a potential cause for the pattern of preferential VH use we observe, based on sequence distant from CDR3, in the four amino acids in FR1 that represent the VH gene segment’s contribution to the socket of the elbow joint between the VH1 and C1H1 domains (18). The elbow is one of a growing number of examples of ball-and-socket joints in which a Phe moiety contributes to the ball (20–23). We found that VH gene segments that contribute to a “loose” elbow, encoded by the four-amino acid motif ELVK, are found more often in mature B cells, both FO and (especially) MZ cells, compared with those found in pro-B and pre-B cells (Figs. 4B and 5).

How could a loose elbow promote survival? B-cell selection and maturation depend on signaling through the antigen receptor (1). Amino acid substitution experiments have shown that disrupting the ball-and-socket joint can markedly decrease signaling while having only a minimal effect on antigen binding (24). Antigen binding is thought to cause the elbow to bend, affecting the initiation of signaling (25, 26). However, precisely how bending relates to signaling remains unclear; indeed, the mechanism of signaling itself remains uncertain, with competing proposals disagreeing over whether aggregation or disaggregation of BCRs at the B-cell surface is required (27–29). Although our findings do not distinguish between these alternatives, it is possible that differences in conformational flexibility afforded by looser or tighter elbow joints could have an impact on signaling by affecting aggregation, disaggregation, or cross-linking. Moreover,
in addition to playing a central role in negative and positive selection during B-cell development in the bone marrow, there is evidence that stronger signaling through the BCR results in repertoire changes, rather than differences in CDR3 length and charge. Thus, these differences in V<sub>H</sub> gene segments vary in their affinity for different light chains and that it is a preference for certain features of light chains that, in turn, causes the differences we see in IgH repertoire. However, it seems unlikely that this could affect the pre-B repertoire, because the mature light chain has not yet been assembled at this stage (although in the context of pre-BCR-dependent signaling, selection for differential binding to a surrogate light chain remains a possibility). Deep-sequencing transitional subsets between pre-B and FO cells, during which a light chain is expressed, and light chain repertoires may be useful ways to investigate this possibility directly. It is also possible that there are other, cryptic patterns in amino acid or nucleotide sequences that affect the antigen specificities and affinities of V<sub>H</sub> gene segments, either through direct contributions to one or more binding loops or through conformational changes arising in the framework region. Similarly, selection of specific V<sub>H</sub> gene segments by either endogenous or exogenous superantigens binding to FR1 (32) may contribute to the bias in V<sub>H</sub> as being under obvious selection, and no others.

Together, these newly reported motifs in FR1 and FR2 result in differences of twofold or more in the frequency of many V<sub>H</sub> gene segments among the B-cell subsets that we examined (Fig. 3A); in all, they have a larger impact on repertoire bias in mice than differences in CDR3 length and charge. Thus, these differences in V<sub>H</sub> gene segment use represent important but hitherto unrecognized features shaping selection of the mature B-cell repertoire, features that, surprisingly, are independent of antigen specificity.

**Methods**

**Animals and Sample Collection.** Organs were harvested from nine 11-wk-old female C57BL/6J mice. Animals were housed and cared for in accordance with protocols approved by New York University Institutional Animal Care and Use Committee. Four of the animals were raised under germ-free conditions (Harvard Medical School Gnotobiotic Core Facility). Five were raised under standard conditions. Flow-sorted bone marrow pre-B cells (B220<sup>hi</sup> IgM<sup>−</sup> c-kit<sup>−</sup> CD25<sup>−</sup>) and splenic FO (CD19<sup>+</sup> B220<sup>−</sup> CD21<sup>+</sup> CD23<sup>−</sup>) and MZ (CD19<sup>+</sup> B220<sup>+</sup> CD21<sup>−</sup> CD23<sup>+</sup>) B cells were obtained from each animal.

**Fig. 4.** Regions of V<sub>H</sub> gene segments under selection. (A) Probability that a four-amino acid motif at the indicated position correlates with PC1 by chance, showing a statistically significant signal of selection at positions 9–13 and 37–39. (B) Amino acid sequence at these two regions in the top 10 gene segments from PC1: the five V<sub>H</sub> gene segments most overrepresented in pre-B cells (Left) and the five most overrepresented in MZ cells (Right), showing selection for ELVK and WMN.

**Fig. 5.** Mapping the ELVK motif to antibody crystal structures. (A) Location of the elbow between the V<sub>H</sub> and C<sub>H1</sub> domains. Key residues from a narrow elbow-containing structure [Protein Data Bank (PDB) ID code 1KEG], which uses V1–5 (B), and a loose-elbow-containing structure (PDB ID code 2VWE), which uses V1–53 (C). Hydrogen bonds are shown in green.
Check sequences and combine paired-end reads into single reads. Where base calls in the region of overlap differed and base call with the higher base call was used. The Mann–Whitney U statistic was then used to test the null hypothesis that intra- and intergroup distances are drawn from the same distribution (i.e., that there is no difference between groups). However, because these distances are not all independent, the U statistic does not follow the standard Mann–Whitney U distribution. Therefore, P values were derived from permutations in which samples were randomly reassigned to the groups (“shuffled-U test”) and by PCA.

**Clustering.** To reduce the effects of bias and error from PCR and sequencing, and the effects of variability in B-cell clone size, reads were clustered, using our replicate samples to determine the degree of clustering that best clustered together reads from single recombination events. Specifically, reads that shared a V,N annotation, a J,N annotation, and CD3 length were clustered until pairwise comparisons of V,N, D, and J,N distributions between replicate samples reached R^2(1−S_{V,N}/S_{total}) ≥ 0.98. This corresponded to a cluster diameter of 6-nt differences—the same as the diameter obtained through a different technique (34)—after which R^2 plateaued. We note that three or fewer cell divisions are expected to have taken place among pre-B and FO cells, limiting the expected number of progeny cells corresponding to each recombination event for these subsets (35). Further supporting this threshold, manual inspection of several dozen clusters revealed no examples in which reads annotated with different V,N was differed by six or fewer nucleotides. Each cluster was assigned a consensus CD3 constructed from the most frequent amino acid at each position within the cluster (or, in case of a tie, one of the most frequent amino acids). Where this consensus CD3 was in-frame and lacked stop codons, the cluster was annotated as a productive rearrangement.

**Comparisons of Gene Segment and Combination Use.** To test for differences between groups of samples, we calculated frequency histograms of gene segment use for individual gene segments (V,N, D, and J,N) and gene segment combinations (VD, V,J,D, and VDJ). Pairwise differences between subsets and between animals were tested in two ways: using a permutation test based on the Mann–Whitney U statistic applied to distances between samples (the “shuffled-U test”) and by PCA.

**Shuffled-U Test.** Distances between each sample histogram were calculated by using R^2 (which is biased toward differences in high-frequency gene segments and combinations) or χ^2 (which weights high- and low-frequency gene segments and combinations equally), although we note that any measure of distance could have been used. The Mann–Whitney U statistic was then used to test the null hypothesis that intra- and intergroup distances are drawn from the same distribution (i.e., that there is no difference between groups). However, because these distances are not all independent, the U statistic does not follow the standard Mann–Whitney U distribution. Therefore, P values were derived from permutations in which samples were randomly reassigned to the groups (“shuffled”). Note that this test is not symmetrical (e.g., just because pre-B samples are more similar to each other than they are to MZ samples does not guarantee that MZ samples will also be more similar to each other than to pre-B samples). Asymmetry indicates differential clustering or the presence of outliers, making it worthwhile to perform this test in conjunction with visualization, which we did.

**PCA.** PCA with singular-value decomposition using Python’s scikit-learn package (36) was performed on V,N, D, and J,N use and on all combinations thereof (e.g., VDJ).

**CD3 Length and Charge.** We calculated the length of each CD3, the number of charged residues, and the net charge per CD3. Distributions for each of these properties were determined for each sample and were found to be approximately normally distributed, justifying the use of SD as a measure of spread (and of the t test, below). Means and SDs were calculated, resulting in nine means and nine spreads for each subset for the set of nine mice, one per mouse. For each property (length, charged residues per CD3, and net charge per CD3), we used the t test to test the null hypotheses that the means and the spreads were equal between pairs of subsets.

**Dependence of Length and Charge on Gene Segment Use.** For each subset, dependence of length or charge on gene segment use was tested by calculating the mean (e.g., length) for each (e.g., V,N) gene segment over all samples, using those means and the V,N gene segment use frequencies in the subset to make a weighted-average predicted mean length for that subset, and using a one-way t test to test the null hypothesis that the prediction agrees with the observed mean length for the subset (i.e., in this example, that V,N gene segment use is sufficient to determine the observed length distribution). This method assumes no correlation between junctional diversity and gene segment use.

**Pooled and Subgroup Comparisons.** The above analyses were performed irrespective of germ-free status and separately comparing germ-free vs. eubiotic animals to confirm no differences between these groups for VDJ. The V,N and J,N gene segment use frequencies in the subset were used to create a predicted mean length for each V,N combination (i.e., V,N–D,J). This method assumes no correlation between junctional diversity and gene segment use.
et al. Each comparison was performed separately using only productive rearrangements, nonproductive rearrangements, and all rearrangements.

**Alignments and Structural Analysis.** We aligned the top five gene segments from PCA of productive joins using ClustalW (37) and looked for positions that were conserved in the gene segments used most commonly in pre-B cells and MZ cells but differed between these two subsets. For the top-contributing gene segment in PC1, we calculated a P-value for each contiguous 4-mer for the null hypothesis that the number of amino acid matches (0, 1, 2, 3, or all 4) is uncorrelated with the weight of the gene segment in the principal component. A conservative significance cutoff was determined by dividing P = 0.05 by the number of independent 4-mers (length of the gene segment, in amino acids, divided by 4) and by the average P-value scaled to 0.5.

We then retrieved crystal structures (PDB ID codes 12E8, 1A3L, 1AST, 1CSB, 1F5K, 1H3P, 1I3G, 1I1G, 1IK1, 1JN6, 1JRH, 1KEG, 1IMNU, 1TMA, 1P9X, 1RQJ, 2C10, 2GSG, 2HKH, 2OTW, 2VLS, 2VWE, 2Z4Q, 3CMO, 3ESV, 3ET9, 3J7V, 3L5W, 3L7E, and 3NN8) that used these V_{H} gene segments from the Protein Data Bank (www.rcsb.org). We displayed them using Chimera (University of California, San Francisco) (38).

**RR Calculations.** Functional V_{H} gene segments whose germ-line sequence ends with an incomplete codon (i.e., whose length in nucleotides was not a multiple of three) were compared with functional V_{H} gene segments whose sequence ends with a complete codon. The goal was to calculate the likelihood (or “risk”) of a V_{H} gene segment that ends an incomplete codon appearing in a nonproductive VDJ rearrangement relative to the risk of any V_{H} gene segment appearing in a nonproductive rearrangement (i.e., the RR). The risk of a given V_{H} gene segment resulting in a nonproductive rearrangement was defined as the ratio of the number of nonproductive VDJ rearrangements containing the gene segment to the sum of productive and nonproductive rearrangements containing the gene segment. The RR was then defined as the ratio of the risk of being nonproductive among sequences using V_{H} gene segments that end in an incomplete codon to the risk of being nonproductive among sequences using all V_{H} gene segments.

**ACKNOWLEDGMENTS.** We thank K. Rajewsky and G. Silverman for critical reading of the manuscript and their insightful suggestions. M.C. and S.B.K. were supported by a Beckman Foundation grant.

Fig. S1. Flow-sort cytograms for isolation of pre-B, follicular B (FO), and marginal zone B (MZ) cells from bone marrow and spleen. Details are provided in Methods.
Fig. S2. Pairwise comparisons of cluster annotation for replicate PCR reactions. Each plot shows the percentage for each of 32 pairs for the indicated parameter [heavy-chain variable (V\textsubscript{H}), diversity (D), heavy-chain joining (J\textsubscript{H}), VDJ, complementarity determining region 3 (CDR3) length, CDR3 charge], one dot per data point per pair. \( R^2 \) values are the mean \( R^2 \) and SD of \( R^2 \) across all pairs. Red lines represent equality (slope of 1).

Fig. S3. \( V_{\text{H}} \) gene segment use among nonproductive rearrangements in pre-B, FO, and MZ subsets. Each circle denotes one animal. Data are plotted on the first two principal components (PC1 and PC2). The percentage denotes the percentage of all variation in the data that is accounted for by PC1 and PC2.
Fig. 54. Alignment and significance of all four amino acid motifs in all $V_H$ gene segments for selection during maturation. The probability that a four-amino acid motif at the indicated position correlates with PC1 by chance for every four-amino acid string in every $V_H$ gene segment is shown. Significant peaks are at positions of the ELVK and WMN motifs, as in Fig. 4A. FR, Framework.