Heavy Chain CDR 3 and Junctional Length Biases in Human Antibody Repertoires Associated with Heavy and Light Chain Germline Utilization

Kannan Sankar, Kam Hon Hoi and Isidro Hötzel

Department of Antibody Engineering, Genentech, South San Francisco, CA 94080, USA

Corresponding author:

Isidro Hötzel, Ph.D.

Department of Antibody Engineering, Genentech, South San Francisco, CA 94080, USA

Phone: 650-225-1000

Email: ihotzel@gene.com

Keywords: Immunoglobulin; VDJ; CDR; recombination; self-reactivity; allele
Abstract

Antibody variable domain sequence diversity is generated by recombination of germline segments. The third complementarity-determining region of the heavy chain (CDR H3) is the region of highest sequence diversity and is formed by the joining of heavy chain V_H, D_H and J_H germline segments combined with random nucleotide trimming and additions between these segments. We show that CDR H3 length distribution is biased in human antibody repertoires as a function of V_H, V_L and J_H germline segment utilization. Most length biases are apparent in the naïve B cell compartment, with a significant bias towards shorter CDR H3 sequences observed in association with a subset of V_H and V_L germlines in the antigen experienced compartment. Similar biases were not observed in nonproductive heavy chain recombination products, indicating selection of the repertoire during B cell maturation as a major driver of the length biases. Some V_H-associated CDR H3 length biases are dependent on utilization of specific J_H germline segments in a manner not directly linked to J_H segment length in the germline, but are rather associated with selection of differentially trimmed J_H segments in the naïve compartment. In addition, D_H segment and N-region random nucleotide insertion lengths within CDR H3 in the naïve compartment were also biased by specific V_H/J_H germline combinations, indicating a complex set of constraints between germline segments selected during repertoire maturation. Our findings reveal biases in the antibody diversity landscape shaped by V_H, V_L, and J_H germline features with implications for mechanisms of naïve and immune repertoire selection.
Introduction

The diversity of sequences in the variable regions of immunoglobulins is the basis for the ability of these molecules to bind a virtually unlimited number of antigenic structures. Sequence diversity in the primary repertoire is created by recombination of germline segments for both the heavy and light chains which results in the formation of full-length immunoglobulin variable region exons (1). The light chain variable region is created by the joining of \( V_L \) and \( J_L \) germlines while the \( V_H \) region is created by recombination of \( V_H \), \( D_H \) (or \( D \)) and \( J_H \) germlines. The process of recombination starts with the heavy chain in progenitor B cells, initiated by \( D/J_H \) recombination followed by \( V_H/DJ_H \) recombination (2, 3). Light chain recombination occurs in pre-B cells after successful completion of the heavy chain recombination. Germline segments in both chains are also trimmed and extended by a variable number of nucleotides by exonucleolytic nibbling of germline segments and random nucleotide incorporation in the N-regions flanking the D germline mediated by terminal deoxynucleotidyl transferase and germline palindromic duplications (3). B cell clones with full-length, in-frame variable regions are further selected to remove or induce receptor editing of self-reactive clones to form the naïve immune repertoire (4, 5).

The third complementarity determining region (CDR) of the heavy chain (CDR H3) is the region of highest overall sequence and length diversity in antibody repertoires (1). CDR H3 length approximates a Gaussian distribution (6). Average CDR H3 length varies as a function of species, age, isotype, B cell development stage and disease state (6-13). CDR H3 amino acid composition is also biased in a CDR H3 length-dependent manner, associated with differential incorporation of D and \( J_H \) germline sequences of different lengths and sequence composition into CDR H3 of different lengths (6). Beyond the germline-associated biases, CDR H3 has been shown to undergo different biases during B cell maturation. In particular, a bias towards shorter average CDR H3 lengths is observed in mature relative to immature B cells (9). This is accompanied by a reduction of positively charged residue content and hydrophobicity within CDR H3 associated with negative selection of self-reactive clones in the repertoire (9, 11, 14, 15). A similar reduction in CDR H3 length occurs in isotype-switched memory B cells relative to naïve to B cells (10, 16).

The analyses of CDR H3 diversity and biases in health and disease have been mostly performed independently of the V regions contributed by \( V_H \) and \( V_L \) germlines (6-11, 17, 18). Except for sequences that are directly incorporated into CDR H3, the impact of V germline segments on CDR H3 properties has not been addressed. In part this is due to the absence of any
expectation for systematic CDR H3 biases as a function of $V_H$ germline, especially in the naïve B cell compartment prior to selection associated with adaptive immune responses. Analysis of the impact of the $V_L$ on CDR H3 has been limited to properties of the third CDR of the light chain, without any evidence of biases (16). Finally, analysis of the impact of $J_H$ germlines on CDR H3 biases has been confined to the expected effects of differential $J_H$ germline length and sequence composition. A recent analysis of a large dataset of isotype-switched human antibody sequences with paired chain information revealed an unexpected preferential pairing of IGHV3-7 ($V_H$3-7) and Vκ2-30 germlines (19) which, in subsequent detailed analysis, was determined to be linked to biases towards shorter CDR H3 lengths associated with both germlines. This prompted us to investigate the extent to which CDR H3 length might be biased as a function of germline use in human immunoglobulin repertoires. Here we describe high-dimensional analyses of CDR H3 sequences from several independently generated human antibody repertoire sequence datasets. Our results uncover biases in CDR H3 and junctional length distributions associated with $V_H$, $V_L$ and $J_H$ germline segment utilization that shape naïve and immune repertoires in unexpected and unpredictable patterns.

Results

Sequence datasets. In the present study we analyzed sequences from four previously described B cell repertoire deep sequencing datasets including 3 donors each and a fifth dataset with 8 donors (16, 19-22). We refer to these datasets as TX, WA, CA, MA and SRI. These datasets were sequenced and bioinformatically parsed using a diversity of methods (Table 1), minimizing the impact of methodological biases. For simplicity we refer to the TX $CD27^{pos}/IgG/IgA$, CA, MA IgG/IgA and WA $CD27^{pos}$ subsets as “antigen-experienced” or “AE”, the TX $CD27^{pos}/IgM$ as “AE IgM” and the TX and WA $CD27^{neg}$ subsets as “ naïve”. The TX and CA datasets include $V_H/V_L$ chain pairing information. The SRI dataset was analyzed separately in this study to avoid over-representation of donors from a single source in pooled data. No antigen-specific selection of B cells was performed for any of the datasets, although the CA and MA datasets include both pre- and post-vaccination samples.

We aimed at identifying properties that are shared among donors and not influenced by clonal expansion related to specific immune responses. To minimize the impact of clonal expansion, the datasets were processed to retain a random sequence from each lineage, or
Clonotype. Clonotype definition varied according to sequencing and parsing method used for each dataset (Table S1). Nonproductive sequences were not grouped by clonotype. Overall distribution of CDR H3 lengths was not significantly affected by removal of redundant sequences in most datasets except for the WA and MA AE compartments, which had subtle but noticeable shifts (Fig. S1A). Germline-specific analyses were performed with germlines having at least 240 counts for all donors in a dataset, which represent 94 to 98% of the repertoire in the CA, TX and MA datasets. Due to ambiguities in V_H germline calls in the WA dataset, germline-specific analyses in this dataset were performed with 16 V_H germlines that had fewer than 10% ambiguous (“in ties”) calls in the dataset, totaling about one third of the repertoire (Table S2). Analyses of whole repertoire data included all sequences filtered by clonotype in specific B cell subsets, regardless of germline classification. The overall AE CDR H3 length distributions are similar among datasets, allowing pooling data from different datasets for the AE B cell subset (Fig. S1B). However, the relative CDR H3 length distributions of the WA and TX naïve B cell subsets differed by an average of 0.9 residues (Fig. S1B) and were thus analyzed separately.

**Average CDR H3 length varies with V_H and V_L germline use.** As a first step we analyzed average CDR H3 length as a function of V_H or V_L germline use. Average CDR H3 length in the AE subset varied by up to 3 amino acid residues as a function of V_H germline use and correlated well for different datasets, with relatively little deviation from length equivalence for different germlines between datasets (Fig. 1A). Average CDR H3 length also varied as a function of V_L germline use by up to 4 amino acid residues in the AE compartment and correlated well between the CA and TX datasets (Fig. 1B). The naïve compartment showed a more limited spread in average CDR H3 lengths relative to the AE compartment (Fig. 1C-F, blue squares). Significant reductions in average CDR H3 length in the AE relative to naïve compartments were associated with a subset of V_H and V_L germlines (Fig. 1C-F). The TX AE IgM subset showed similar trends as the TX AE IgG/IgA subset except that average CDR H3 length was decreased in association with most V_H germlines relative to the naïve compartment (Fig. 1C and E).

**CDR H3 length distribution varies as a function of V_H germline use.** We next determined whether CDR H3 length distribution varies with germline use. Overall CDR H3 length distribution of the respective B cell compartment was used as a relative standard to which germline-specific CDR H3 length distributions were compared. This was done to facilitate comparison of biases between samples and also because useful objective reference distributions are not available to
determine bias types in naïve compartment sequences. Therefore, most biases described here, including all in the naïve compartment, are relative to the average of the repertoire in each B cell compartment. Overall and germline-specific CDR H3 length distributions were determined by averaging the frequency of each CDR H3 length for all donors across the TX, CA, MA and WA datasets, with the SRI dataset analyzed separately. Statistical analysis of biases was performed in the AE compartment by a two-tailed paired (by donor) t-test of length frequencies with a sliding window of two consecutive CDR H3 lengths to minimize the significance of local distribution fluctuations. Observed length distribution biases included overall shifts in average CDR H3 length for sequences with different V_{H} germlines and also obvious and subtle deviations from the overall CDR H3 distribution in discrete ranges of the CDR H3 length spectrum (Fig. 2A, Fig. S2A).

To further discern the CDR H3 length biases quantitatively, we performed a principal component (PC) analysis of the length distributions (lengths 5 to 26) associated with different V_{H} germlines to capture the trends causing highest variations in the observed length distributions. Results from the PC analysis were visualized by projecting each germline onto the most important trends, to obtain the so-called PC scores (Fig. 3A). Interpretation of the PC score plots was aided by a visual analysis of the corresponding distributions. Overall, the most significant component, PC1, corresponded to apparent skewness towards shorter or longer lengths, whereas PC2 corresponded to apparent kurtosis, or relative enrichment or depletion of sequences in the mid-range lengths, of the distributions. Using the PC analysis results in conjunction with visual inspection of V_{H} germline-associated CDR H3 distributions compared to the overall CDR H3 distribution in the AE compartment, germlines were categorized by bias type in discrete groups as “Short”, “Neutral”, “Long”, “Cut” and “Crested” (Fig. 2 and 3A, Fig. S2). These groups have different degrees of shifts towards longer or shorter lengths and kurtosis relative to the overall distribution. Differences between the distributions of member of different groups can be subtle, both visually and in the PC analysis. One example is the difference between the Neutral and Cut groups, the latter showing depletion of sequences limited to a narrow band of short CDR H3 lengths. The magnitude of the biases and the details of distribution shapes within each group varied for different V_{H} germlines. However, these were consistent across datasets for each germline, V_{H}1-69 being a notable exception (Fig. S3). Germlines in the same V_{H} subfamily did not always belong to the same bias groups. The three major V_{H} subfamilies, V_{H}1, V_{H}3 and V_{H}4, are represented in
the Neutral group. The range of germline prevalence in the repertoire was similar for different groups except for the higher prevalence of some germlines in the Crested group (Fig. S4).

We determined whether the observed distribution biases were also present in the naïve B cell subset. The Long, Crested and Cut biases were also observed in the naïve B cell compartment, without apparent differences relative to the distributions in the AE compartments (Fig. 2B and C and Fig. S2B, C, E and F). All the germlines in the Neutral group showed average CDR H3 length distribution in the naïve subset as well. However, distribution biases of the Short group in the naïve compartment were variable (Fig. S2B and C), consistent with the average CDR H3 length analysis (Fig. 1E and F). Short biases in the naïve compartment were mostly limited to the V_{H3}-73 and V_{H3}-15 germlines in the TX and WA datasets. Despite the differences in overall CDR H3 length between the TX and WA naïve datasets, the biases in the naïve compartment had the same trends in both datasets (Fig. 2B and C and Fig. S2B and C).

The data analysis was performed in datasets aggressively filtered for sequences likely to belong to the same lineage. To confirm that biases are not due to pockets of clonal expansion, we performed a repertoire similarity index (RSI) analysis with the CA, TX and MA datasets similar to a recently described method (23). The RSI analysis computes CDR H3 identities within each donor for sequences with the same germline, CDR length and V_{H}/V_{L} pairing (for the paired CA and TX datasets) or V_{H} and J_{H} combination (for the unpaired MA dataset). Clonal expansion would thus be reflected by higher than average RSI values. Overall, no significant increase in RSI scores was associated with regions of positive prevalence biases in different parts of the CDR H3 length spectrum for different bias groups (Fig. S5A), confirming that clonal expansion does not account for the observed CDR H3 length biases.

**CDR H3 length distribution bias as a function of V_{H} germline is not generated by VDJ recombination.** We next determined whether the biases observed in the naïve compartment are a direct consequence of biases in the VDJ recombination process for each germline. For this we analyzed frameshifted, nonproductive V_{H} sequences that were part of the naïve WA dataset. Nonproductive recombination products are not directly subject to selection and therefore provide information about recombination products prior to any selection of the repertoire. As previously reported, the CDR H3 length of nonproductive V_{H} genes is significantly longer than the productively recombined genes in mature B cell subsets (15). However, CDR H3 length for the nonproductive sequences associated with different V_{H} germlines approximated a Gaussian
distribution, with no observable biases associated with different \( V_H \) germlines relative to the overall repertoire except for minor anomalies associated with some germlines (Fig. 2D and Fig. S2D). Therefore, heavy chain recombination mechanisms do not account for the CDR length distribution biases observed in the naïve repertoire.

**CDR H3 length distribution varies as a function of \( V_L \) germline use.** We performed a similar analysis of CDR H3 length distribution as a function of \( V_L \) germline and B cell compartment using PC and visual analysis. Similar to the \( V_H \) germline-associated biases, \( V_L \)-associated biases in the AE compartment could be classified into three groups, named here “Short” (high value of PC1), “Long” (low value of PC1) and “Neutral” (intermediate values of PC1), present in both the CA and TX datasets, each group including a diverse set of \( V_k \) and \( V_l \) germlines (Fig. 3B and 4A, Fig. S6A). PC1 and PC2 for the light chain were also associated with apparent skewness and kurtosis. The \( V_L \) Long bias group has Gaussian CDR H3 length distributions, whereas the \( V_L \) Short bias group includes distribution shapes with significant deviations from Gaussian, including localized frequency spikes in discrete sections in the short range. Only \( V_k \) germlines in the Long group were associated with similar CDR H3 length biases in the TX naïve compartment (Fig. 4B, Fig. S6B and C). The magnitude of the \( V_L \)-associated biases varied for different germlines within each bias group but were consistent between datasets (Fig. S7). As above, the RSI analysis results indicated that clonal expansion does not account for the \( V_L \) germline-associated CDR H3 length biases (Fig. S5B). The prevalence of germlines in the Short group in the repertoire was lower than for germlines of the other two groups (Fig. S4)

**CDR H3 length is biased as a function of \( V_H/J_H \) combination.** \( J_H \) germlines vary in the number of amino acid residues that can be potentially contributed to CDR H3 from 4 in \( J_H \)4 to 9 in \( J_H \)6. We assessed whether differential \( J_H \) germline usage as a function of \( V_H \) and \( V_L \) germline is the basis for V segment-associated CDR H3 length biases. No significant deviations from average \( J_H \) usage were observed in association with most \( V_H \) germlines in the WA unproductive sequences (Fig. S8A). Although deviations in \( J_H \) prevalence linked to some germlines were observed in the naïve compartment of both datasets (e.g., \( V_H \)2-5, \( V_H \)3-9), those deviations do not readily explain CDR H3 distribution biases associated with these \( V_H \) germlines (Fig. S8B, C and D). One exception was a single \( V_L \) germline with a bias for longer CDR H3 lengths, \( V_k \)2-28, which was associated with a higher prevalence of the longer \( J_H \)6 and lower prevalence of the shorter \( J_H \)4.
germline segments (Fig. S8D). Some of these biases were observed in a reciprocal analysis of J\textsubscript{H} germline usage as a function of V\textsubscript{H} germline and CDR H3 length in the naïve compartment but not in nonproductive sequences (Fig. S9). These include not only the generally skewed J\textsubscript{H} usage proportional to J\textsubscript{H} germline length, as expected, but also higher or lower than average prevalence of the J\textsubscript{H}5 germline and different J\textsubscript{H}4/J\textsubscript{H}6 germline usage ratios in the average to longer CDR H3 lengths.

We next analyzed CDR H3 length distributions associated with different V\textsubscript{H}/J\textsubscript{H} germline combinations, comparing these to CDR H3 length distribution of all sequences with the corresponding J\textsubscript{H} germline. As expected, CDR H3 length distributions were generally shifted according to length of J\textsubscript{H} the segment in the germline regardless of V\textsubscript{H} germline (Fig. 5, Fig. S10 and S11). However, a subset of V\textsubscript{H}-associated CDR H3 length biases were impacted by J\textsubscript{H} germline in a manner independent of length of the J\textsubscript{H} segment in the germline, with very similar patterns in the naïve WA and IgM/naïve SRI subsets (Fig. 5, Fig. S10 and S11). These included a short CDR H3 length bias associated with sequences with the V\textsubscript{H}3-72, V\textsubscript{H}3-73 and V\textsubscript{H}3-15 germlines and a long bias associated with V\textsubscript{H}1-2, V\textsubscript{H}1-3, V\textsubscript{H}2-5 and V\textsubscript{H}7-4-1 combined with the J\textsubscript{H}5 germline (Fig. 5, Fig. S10). Other CDR H3 length biases associated with specific V\textsubscript{H}/J\textsubscript{H} germline combinations were also observed, including V\textsubscript{H}3-73/J\textsubscript{H}1, V\textsubscript{H}3-73/J\textsubscript{H}4, V\textsubscript{H}7-4-1/J\textsubscript{H}6 and other minor biases. Two other members of the Crested group, V\textsubscript{H}3-20 and V\textsubscript{H}3-23, showed their characteristic bias when combined with all J\textsubscript{H} germlines except J\textsubscript{H}2 and J\textsubscript{H}6 (Fig. S10), the longest J\textsubscript{H} germlines. The Cut group was apparent only when overall population distribution is shifted toward shorter CDR H3 lengths in association with the short J\textsubscript{H}4 germline, making the reductions in the number of sequences in very short range in this group more apparent (Fig. 5). In contrast, the two V\textsubscript{H} germlines in the Long group that were analyzed, V\textsubscript{H}1-18 and V\textsubscript{H}2-26, were associated with long CDR H3 length biases in the context of most or all J\textsubscript{H} segments. Distributions associated with V\textsubscript{H}3-9 were unique in that the same peak of sequences with length 14 to 17 occurred independently of J\textsubscript{H} segment in both datasets. Our results indicate that CDR H3 length distribution biases are not necessarily uniform for each V\textsubscript{H} germline but may vary as a function of J\textsubscript{H} germline. In addition, the effect of J\textsubscript{H} on CDR H3 length distribution is not necessarily similar within V\textsubscript{H} bias groups, indicating some degree of heterogeneity within bias groups.

**Selection of differentially trimmed J\textsubscript{H} segments associated with different V\textsubscript{H} germlines in the naïve compartment.** The CDR H3 length distribution biases associated with a subset of V\textsubscript{H}/J\textsubscript{H}
germline combinations may be a consequence of biases in J\textsubscript{H} trimming as a function of V\textsubscript{H} germline. J\textsubscript{H} residue occupancy in the last CDR H3 positions of J\textsubscript{H}4 and J\textsubscript{H}5 sequences was used to indirectly determine J\textsubscript{H} trimming. The J\textsubscript{H}1, 2, 3 and 6 germlines were not analyzed due to lack of sufficient data or, in the case of J\textsubscript{H}6, absence of apparent J\textsubscript{H} segment-associated CDR H3 length biases. No apparent biases in J\textsubscript{H} residue occupancy relative to overall repertoire were observed for any of the analyzed V\textsubscript{H}/J\textsubscript{H} combinations in the nonproductive WA sequences (Fig. S12). However, J\textsubscript{H} residue trimming biases were observed for different V\textsubscript{H}/J\textsubscript{H} combinations in the naïve WA compartment (Fig. 6 and Fig. S12). General trends in residue occupancy in J\textsubscript{H}4 were similar in IgM/naïve SRI sequences for the V\textsubscript{H}/J\textsubscript{H}4 germline combinations with sufficient numbers for analysis (Fig. S13). Residue-specific trimming biases were found to be mostly coordinated for different J\textsubscript{H} residues in each analyzed V\textsubscript{H}/J\textsubscript{H} combination, as expected due to the directional nature of trimming. However, closely related V\textsubscript{H} germlines can be associated with distinct trimming biases of different J\textsubscript{H}4 residues. For instance, V\textsubscript{H}2-5/J\textsubscript{H}4 sequences are associated mostly with reduced trimming of IMGT® residues 114 and 115 (Tyr and Phe) whereas in the case of V\textsubscript{H}2-70/J\textsubscript{H}4 strongly reduced trimming of residue 116 (Asp) was also observed. The results indicate a complex set of constraints leading to selection of differentially trimmed J\textsubscript{H} segments in the context of certain V\textsubscript{H} and J\textsubscript{H} germlines during naïve repertoire maturation.

**Biases in D segment and N-region lengths within CDR H3 sequences as a function of V\textsubscript{H} and J\textsubscript{H} germline use.** The observed J\textsubscript{H} segment length biases in CDR H3 sequences of specific lengths could be an indirect consequence of biases elsewhere in CDR H3, including the length of V\textsubscript{H} and D sequences and number of N-region and palindromic nucleotide insertions (NP-region) flanking the D region in CDR H3. Naïve sequences from the 3 donors of the WA dataset and IgM/naïve sequences from 3 of the donors with higher number of sequences in the SRI dataset were parsed for V\textsubscript{H}, J\textsubscript{H}, D and NP-region lengths within CDR H3. In general, no obvious differences on the prevalence of D germlines with different average lengths were observed in association with different V\textsubscript{H} and the J\textsubscript{H}4 and J\textsubscript{H}5 germlines that would account for J\textsubscript{H} segment length biases (Fig. S14). One possible exception is V\textsubscript{H}6-1 which was associated with shorter D germlines. In addition, the number of nucleotides that V\textsubscript{H} can contribute to CDR H3 did not correlate with J\textsubscript{H} length biases (Fig. 7 and S15). However, different classes of biases in the lengths of D segments and NP-regions were observed for different V\textsubscript{H}/J\textsubscript{H} combinations, even for clones with the same V\textsubscript{H} germline (Fig. 7 and S15). Biases had similar trends in the WA and SRI datasets, with differences between
datasets observed mostly in the magnitude of the biases. No similar biases were observed in the nonproductive WA sequences with exception of differences in average $V_{H}$-derived sequence lengths associated with $V_{H}$ germline length and a generally shorter NP-region length in $V_{H}2$ clones (Fig. S15), indicating that the observed D and NP-region length biases are mostly selected in naïve repertoire maturation.

Different classes of junctional biases were observed for different $V_{H}/J_{H}$ combinations (Fig. 7A and B and S15). In several $V_{H}/J_{H}$ combinations, $J_{H}$ length biases were inversely correlated with both D segment and NP-region length biases including, for instance, $V_{H}3-23/J_{H}5$, $V_{H}3-15/J_{H}4$, $V_{H}3-73/J_{H}4/J_{H}5$ and $V_{H}3-73/J_{H}4$ combinations. In these cases, the factor determining junctional biases is likely to be the $J_{H}$ segment, with the other junctional components compensating for the $J_{H}$ biases. In $V_{H}3-74/J_{H}4$, biases in the D sequence are compensated by both NP-region and $J_{H}$ lengths, suggesting the D segments as the bias determinant. In other cases, such as $V_{H}3-15/J_{H}4$, $V_{H}3-7/J_{H}4$ and $V_{H}3-9/J_{H}4$, D segment and NP-region lengths were inversely correlated with each other without correlations with $J_{H}$ segment length across the CDR H3 length spectrum. In $V_{H}2$ sequences, the reduction in NP-region length observed in nonproductive sequences is maintained in several $J_{H}$ germline combinations in the naïve repertoire and compensates for the longer $V_{H}$ sequences. However, as the $J_{H}$ segments of nonproductive $V_{H}2$ sequences do not appear to be biased relative to the overall repertoire, the observed $J_{H}4$ and $J_{H}5$ trimming biases associated with $V_{H}2$ naïve sequences are presumably due to $J_{H}$ trimming rather than NP-region selection. Overall, the results show different classes of biases in D segment, N-region and $J_{H}$ lengths within CDR H3 of naïve sequences that vary among $V_{H}/J_{H}$ germline combinations.

**Discussion**

Understanding antibody CDR H3 diversity generation, a process critical for antigen binding, has long been a goal in the immunology and antibody engineering fields. Numerous reports have described overall CDR H3 length and amino acid composition biases in health and disease as well as in different B cell differentiation and developmental stages and species. Here we describe detailed, high-dimensional analyses of CDR H3 and junctional segment length distributions and show a complex set of biases determined by $V_{H}$, $V_{L}$, and $J_{H}$ germline use and B cell maturation state. Most of the length biases we describe are evident in the naïve B cell compartment but not, with very few exceptions, in the nonproductive subset, indicating a major
role of naïve B cell repertoire maturation in shaping those biases. In addition, only a subset of \( V_H \) or \( V_L \) germlines is associated with biases towards shorter CDR H3 lengths in the antigen-experienced compartment, indicating general germline-specific adaptive immunity selection processes shared among individuals. The \( V_H \) and \( V_L \)-associated CDR H3 length biases parallel observations of patterns of T cell receptor \( \beta \) chain CDR3 length distribution biases with different TRBV germlines in repertoires that arise in the process of T cell maturation (24), indicating that at least the broader biases are a general phenomenon in immune repertoires. Our results extend these observations by showing biases in the length of junctional segments within CDR H3 determined by \( V_H \), and \( J_H \) germline usage in patterns that are not predictable from germline repertoire sequence properties. A recurring theme in the results presented here was that biases observed at one level (e.g., \( V_H \) germline) were only partly explained by biases at higher dimensional levels (e.g., \( V_H/J_H \) combinations), with additional unexpected biases observed in the higher-dimensional levels. It is expected that analyses including other repertoire descriptors not included here will uncover additional biases, one example being germline allelic variation (Fig. S16). Our results provide a framework and baseline for repertoire analyses in disease and immune states beyond the germline usage and average CDR H3 length and composition properties commonly analyzed in repertoire deep sequencing studies.

Special consideration was given to the repeatability and robustness of the findings. The results are based on a total of 12 donors with AE or isotype-switched sequences, 5 of which also had naïve B cell sequences, and confirmed by analysis of 8 additional donors from the SRI dataset. These datasets were obtained and parsed with different sequencing methods and bioinformatic pipelines, which minimizes the impact of technical artifacts. Some of the biases, such as those associated with \( V_L \), \( J_H \) and D germlines and NP-regions, cannot be easily generated by sequencing or parsing artifacts. However, subtle differences were observed between datasets which may be due to technical reasons, such as the slightly longer average CDR H3 length of sequences in the WA AE dataset compared to other datasets and the significant differences in average CDR H3 length between the naïve compartments of the TX and WA datasets (Fig. 1). However, these differences in baseline average CDR H3 length, which were accounted for in our naïve compartment analyses, did not affect the overall results. The \( V \) germline-associated CDR H3 length biases are not related to clonal expansion as sequences were clustered by clonotype. The stringency of clonotype clustering criteria had limited impact on results. This is exemplified by
the WA and SRI datasets, which yielded results comparable with other datasets despite having been clustered by clonotype using a higher CDR H3 sequence identity threshold than other datasets. Limited reliability of D germline classification, especially when sequence identity length is short, is well known. However, any D germline assignment errors would not be expected to be associated with particular V\textsubscript{H}/J\textsubscript{H} germline combinations sequences. In addition, errors in D parsing should still result in similar length D germline sequence matches, minimizing the impact on D and NP-region length assessments. Thus, the observed D segment length biases associated with sequences with different V\textsubscript{H}/J\textsubscript{H} germline combinations, generally similar in the WA and SRI dataset subsets analyzed, are not expected to be a consequence of systematic junction parsing errors. This is further supported by the lack of similar biases in nonproductive sequences of similar lengths parsed by the same method, except for the expected germline-specific V\textsubscript{H} region length biases in CDR H3.

One factor that was not addressed here is haplotype variations within and between donors. Haplotypes could potentially affect CDR H3 length distributions through differences in D germline composition and differential recombination frequencies of D or J\textsubscript{H} germlines of different lengths in different chromosomes (25). This, combined with differential recombination frequencies of V\textsubscript{H} alleles (26-28), may impact CDR H3 distributions associated with certain V\textsubscript{H} germlines. However, heavy chain variable region haplotype differences would not be expected to impact CDR H3 distributions associated with V\textsubscript{L} germlines and the AE compartment-specific short CDR H3 length biases. In addition, the observation of essentially the same CDR H3 length distribution biases in several donors from five different sources and junctional segment length biases in six donors from two of these sources, along with a lack of systematic associations between V\textsubscript{H} and D and J\textsubscript{H} alleles across donors (27, 28), indicates that haplotype variations are unlikely to be a major factor in the CDR H3 and junctional length distribution biases described here.

The analyses shown here use germline information as a proxy for undefined sequence features that ultimately determine the observed biases. Therefore, the selected CDR H3 sequence and structural properties that result in the observed biases and the germline sequence properties that determine those biases remain to be identified. Analysis of V\textsubscript{H} germline residues that can directly encode or bias CDR H3 IMGT\textsuperscript{®} positions 105 to 107 did not reveal clear correlations between the number or type of encoded residues and most CDR H3 bias groups or junctional segment length biases (Fig. S17). One exception may be the Cut group, which includes two V\textsubscript{H}2
germlines with slightly longer extensions in CDR H3. The extended V_H sequences, along with the observed reduced J_H4 trimming associated with these germlines, may contribute to the low frequency of short CDR H3 sequences in the Cut group. In addition, no obvious correlations between J_H trimming biases and variations in V_H germline residues in positions 40 to 42 generally contacting the differentially trimmed J_H residues 115 and 116 were observed. The differentially trimmed residue 116 is located in a region at the base of CDR H3 that can adopt either a “bulged” or “extended” conformation (29, 30). The factors that determine the more common bulged conformation are not clear and appear to depend on the Ig domain, encoded mostly by V_H germlines (30, 31). Whether J_H trimming biases reflect biases in the structure of the CDR H3 base as a function of V_H germline remains to be determined. The biases in D segment and NP-region lengths appear to be, in several cases, secondary to J_H length biases. In other cases, opposite length biases in D segment and NP-region lengths that are independent of J_H trimming biases in the context of specific V_H/J_H combinations are observed, indicating selection of differentially trimmed D germline segments. These may be due to selection of different amino acid compositions associated with D and NP-region sequences. Finally, CDR H3 length-specific selection independent of sequence appears to occur in V_H3-9, in which sequences of lengths 14-17 are enriched regardless of the underlying J_H germline length-associated biases. One challenge in determining how different germline regions interact to give rise to the observed biases is the relatively limited number of non-redundant human antibody structures with different germline combinations, extent of J_H germline trimming and CDR H3 lengths.

The CDR H3 biases described here pose questions about the functional properties that might shape those biases and the functional consequences of these biases for adaptive immunity. The emergence of some biases in the naïve repertoire suggests selection against self-reactivity, selection for structural integrity, expression or a combination of these factors as possible mechanisms. One factor that seems not to contribute significantly to most or all of these biases is chain pairing, except perhaps for the association between Vk2-28 and J_H6 in the naïve compartment. In agreement with previous reports, no significant association between V_H and V_L germlines of similar bias types was observed, with one exception being the previously described preferential pairing of the V_H3-7/Vk2-30 germlines in the Short V_H and V_L bias groups in the AE compartment (19, 32). If related to selection against self-reactivity, the different biases indicate either that features other than CDR H3 charge and hydrophobicity significantly contribute to self-
reactivity or that V segments modulate the self-reactivity mediated by these factors. The bias towards shorter CDR H3 lengths associated with a subset of VH and VL germlines in the AE compartment may be attributable to these same mechanisms or to immune selection. The latter would suggest widespread convergences in human repertoires associated with certain VH and VL germlines or, possibly, some degree of functional specialization in the germline repertoire linked to short CDR H3 sequences, analogous to the association between CDR H3 length and recognition of different antigen classes (33). Our results point to unexpected cross-constraints between VH, VL, JH and other junctional elements selected at different stages of B cell development that significantly shape antibody repertoires.

Materials and methods

Datasets and analysis. Sequences were obtained from the original publications (16, 19, 20) except for the MA dataset. The sequences in the MA dataset were obtained from a re-sequencing by Illumina MiSeq (34) of the same set of samples previously described by Laserson et al. (21). A summary of the samples used here is given in Table S3. Sequencing methods for the MA dataset are described in the experiment design section associated with sample data (see https://www.ncbi.nlm.nih.gov/sra/SRX2251687). Sequences were used as parsed in the original publications except for the MA dataset, where the raw sequencing files were processed and germlines annotated with a custom pipeline. Briefly, paired-end reads were merged using FLASH (35) to reconstruct the full-length variable domain sequences using the following parameters: read length at 300 bps, expected fragment length at 530 bps, standard deviation at 50 bps. The full-length sequences were subsequently processed to identify the frameworks and CDR regions using position-weighted motifs as previously described (36). IgBlast (37) was used to supplement the region parsed data with germline annotation for each sequence, including nucleotide somatic mutations. Isotypes of the sequences were determined by finding the closest matching human CH1 isotypes on the available CH1 sequences. Each sequence was processed and annotated with the frameworks, CDRs, germline use and clonotype grouping (see below). Nonproductive sequences in the WA dataset used for analyses were limited to frameshifted sequences in the naïve compartment to minimize the indirect effect of clonal expansion. CDR H3 length of nonproductive, frameshifted sequences in amino acids was set as the nearest integer of CDR H3 length in nucleotides divided by 3. For naïve compartment sequences of Donor 1 of the WA dataset only the D1a repeat was used for most analyses (20). Parsing of D segments was done using Blast
(38) after removing the sequences corresponding to V<sub>H</sub> and J<sub>H</sub> regions from CDR H3 sequences. An identity of 100% over a span of at least 5 contiguous nucleotides was required for D germline matches. The samples from the D1Nb subset were also included for D segment parsing, removing redundant sequences as described below. All CDR H3 length distributions and germline prevalence analyses were determined using custom scripts and Microsoft Excel 2016. Paired t-tests of CDR H3 length distributions were performed using Microsoft Excel 2016. Mann-Whitney tests for distributions were done using GraphPad Prism version 6. The IMGT® CDR definition and numbering system is used throughout (39).

**Clonotype clustering.** Clonotypes in the CA and TX datasets were defined as sequences from the same donor, V<sub>H</sub> and V<sub>L</sub> germlines and CDR H3 length with a nominal 57% or greater CDR H3 amino acid sequence identity, which better approximates an average 60% CDR H3 sequence identity across the range of CDR H3 lengths. For the TX dataset IgG/IgA, CD27<sup>pos</sup>/IgM and CD27<sup>neg</sup> sequences were segregated prior to clonotype clustering. The minor fraction of sequences without germline information in the TX dataset was not clustered into clonotypes. Clonotypes in the MA dataset were defined as sequences from the same donor, V<sub>H</sub> and J<sub>H</sub> germlines and CDR H3 length with a nominal 57% or greater CDR H3 amino acid sequence identity as above. IgG/IgA and IgM sequences were also segregated prior to clonotype clustering. Clonotypes in the WA dataset were defined as sequences from the same donor with the same V<sub>H</sub> and J<sub>H</sub> germline and same CDR H3 length and sequence. If V<sub>H</sub> germline information was not available then V<sub>H</sub> subfamily information was used in lieu, retaining as a representative for the clonotype a sequence with V<sub>H</sub> germline information if available. If J<sub>H</sub> germline information was not available then this parameter was ignored, also retaining otherwise identical sequences with available J<sub>H</sub> information as representatives for clonotypes, if available. Nonproductive sequences in the WA dataset were not processed for clonotype clustering. Clonotypes in the SRI dataset were defined as sequences from the same donor with the same V<sub>H</sub> and J<sub>H</sub> germline and same CDR H3 length and sequence. Only sequences labeled as “productive” in the SRI dataset were analyzed.

**Repertoire Similarity Index Analysis.** RSI was computed in a manner similar to a previously described method (23). For a given set $S$ of CDR H3 sequences, all of the same length $n$, RSI is measured as follows:
\[ RSI = \left[ 1 - \frac{\text{median}\left\{ LD(S_i, S_j)_{i \neq j} \right\}}{n} \right] \times 100\% \; \forall \; 1 \leq i, j \leq |S| \]

where \( S_j \) and \( S_j \) refer to any two sequences in the set of CDRH3 sequences and \( LD(S_i, S_j) \) refers to the Levenshtein distance function, which measures the number of amino acid changes necessary to convert \( S_i \) to \( S_j \). For a given \( V_H \) germline and CDR H3 length, RSI values were computed for those sequences that shared the same \( V_L \) germline (for the paired CA and TX datasets) or the same \( J_H \) germline (for the unpaired MA datasets) and the same CDR H3 length. Values were computed separately for each donor in the datasets and averaged for each length. Values shown in graphs in Fig. S5 are the averages in each length for different datasets. All calculations were performed using custom scripts in R.

**Principal component analysis of CDR H3 length distributions.** The length distribution of each germline was captured as a vector of length 22 containing the percentage of sequences of length 5 to 26. For \( V_H \), the values for each germline were averaged over all the AE datasets except the WA dataset due to limited germline coverage. For \( V_L \), the values were averaged over the CA and TX datasets. The distributions of each germline were consolidated into a matrix \( X_{n \times 22} \) where \( n \) is the number of germlines considered for analysis (\( n = 39 \) for \( V_H \) and \( n = 35 \) for \( V_L \)). In order to capture the differences in behaviors of the germlines the naïve and AE sets, they were considered separately (if at least 240 sequences were available when pooled from all donors in each set). The variance covariance matrix \( S_{22 \times 22} \) of \( X \) was computed with elements \( S_{ij} \) as

\[ S_{ij} = \langle (X - \langle X_i \rangle)(X - \langle X_j \rangle) \rangle \; \forall \; 1 \leq i, j \leq 22, \]

where \( <> \) refers to average across all germlines. Eigen decomposition of the matrix \( S \) results in 22 eigenvectors, each of which capture a trend in the distribution as a function of the CDR H3 lengths and are sorted in decreasing order of the variance they capture. Each germline was then projected onto these eigenvectors to obtain the PC scores which enabled visualization of the different trends and comparisons among the different germlines.

**Acknowledgments**

We thank Steve Guerrero and Qing Zhang for helpful discussions.
References

1. H. W. Schroeder, Jr., L. Cavacini, Structure and function of immunoglobulins. J. Allergy Clin. Immunol. 125, S41-52 (2010).
2. R. M. Cobb, K. J. Oestreich, O. A. Osipovich, E. M. Oltz, Accessibility control of V(D)J recombination. Adv. Immunol. 91, 45-109 (2006).
3. C. H. Bassing, W. Swat, F. W. Alt, The mechanism and regulation of chromosomal V(D)J recombination. Cell 109 Suppl, S45-55 (2002).
4. K. Rajewsky, Clonal selection and learning in the antibody system. Nature 381, 751-758 (1996).
5. D. Nemazee, Mechanisms of central tolerance for B cells. Nat. Rev. Immunol. 17, 281-294 (2017).
6. M. Zemlin, M. Klinger, J. Link, C. Zemlin, K. Bauer, et al., Expressed murine and human CDR-H3 intervals of equal length exhibit distinct repertoires that differ in their amino acid composition and predicted range of structures. J. Mol. Biol. 334, 733-749 (2003).
7. H. W. Schroeder, Jr., Similarity and divergence in the development and expression of the mouse and human antibody repertoires. Dev. Comp. Immunol. 30, 119-135 (2006).
8. D. Bagnara, M. Squillario, D. Kipling, T. Mora, A. M. Walczak, et al., A Reassessment of IgM Memory Subsets in Humans. J. Immunol. 195, 3716-3724 (2015).
9. H. Wardemann, S. Yurasov, A. Schaefer, J. W. Young, E. Meffre, et al., Predominant autoantibody production by early human B cell precursors. Science 301, 1374-1377 (2003).
10. Y. C. Wu, D. Kipling, H. S. Leong, V. Martin, A. A. Ademokun, et al., High-throughput immunoglobulin repertoire analysis distinguishes between human IgM memory and switched memory B-cell populations. Blood 116, 1070-1078 (2010).
11. K. M. Roskin, N. Simchoni, Y. Liu, J. Y. Lee, K. Seo, et al., IgH sequences in common variable immune deficiency reveal altered B cell development and selection. Sci. Transl. Med. 7, 302ra135 (2015).
12. M. E. Doorenspleet, P. L. Klarenbeek, M. J. de Hair, B. D. van Schaik, R. E. Esveldt, et al., Rheumatoid arthritis synovial tissue harbours dominant B-cell and plasma-cell clones associated with autoreactivity. Ann. Rheum. Dis. 73, 756-762 (2014).
13. R. J. M. Bashford-Rogers, K. G. C. Smith, D. C. Thomas, Antibody repertoire analysis in polygenic autoimmune diseases. Immunology 155, 3-17 (2018).
14. I. I. Ivanov, R. L. Schelonka, Y. Zhuang, G. L. Gartland, M. Zemlin, et al., Development of the expressed Ig CDR-H3 repertoire is marked by focusing of constraints in length, amino acid use, and charge that are first established in early B cell progenitors. J. Immunol. 174, 7773-7780 (2005).
15. K. Larimore, M. W. McCormick, H. S. Robins, P. D. Greenberg, Shaping of human germline IgH repertoires revealed by deep sequencing. J. Immunol. 189, 3221-3230 (2012).
16. B. J. DeKosky, O. I. Lungu, D. Park, E. L. Johnson, W. Charab, et al., Large-scale sequence and structural comparisons of human naive and antigen-experienced antibody repertoires. Proc. Natl. Acad. Sci. U.S.A. 113, E2636-2645 (2016).
17. Y. Elhanati, Z. Sethna, Q. Marcou, C. G. Callan, Jr., T. Mora, et al., Inferring processes underlying B-cell repertoire diversity. Philos. Trans. R. Soc. Lond. B Biol. Sci. 370, (2015).
18. P. Miqueu, M. Guillet, N. Degauque, J. C. Dore, J. P. Souillou, et al., Statistical analysis of CDR3 length distributions for the assessment of T and B cell repertoire biases. Mol. Immunol. 44, 1057-1064 (2007).
19. L. D. Goldstein, Y. J. Chen, J. Wu, S. Chaudhuri, Y. Hsiao, et al., Massively parallel single-cell B-cell receptor sequencing enables rapid discovery of diverse antigen-reactive antibodies. bioRxiv doi: 10.1101/660472, (2018).

20. W. S. DeWitt, P. Lindau, T. M. Snyder, A. M. Sherwood, M. Vignali, et al., A Public Database of Memory and Naive B-Cell Receptor Sequences. PLoS One 11, e0160853 (2016).

21. U. Laserson, F. Vigneault, D. Gadala-Maria, G. Yaari, M. Uduman, et al., High-resolution antibody dynamics of vaccine-induced immune responses. Proc. Natl. Acad. Sci. U.S.A. 111, 4928-4933 (2014).

22. B. Briney, A. Inderbitzin, C. Joyce, D. R. Burton, Commonality despite exceptional diversity in the baseline human antibody repertoire. Nature 566, 393-397 (2019).

23. V. Greiff, U. Menzel, E. Miho, C. Weber, R. Riedel, et al., Systems Analysis Reveals High Genetic and Antigen-Driven Predetermination of Antibody Repertoires throughout B Cell Development. Cell Rep. 19, 1467-1478 (2017).

24. T. Matsutani, T. Ohmori, M. Ogata, H. Soga, S. Kasahara, et al., Comparison of CDR3 length among thymocyte subpopulations: impacts of MHC and BV segment on the CDR3 shortening. Mol. Immunol. 44, 2378-2387 (2007).

25. M. J. Kidd, K. J. Jackson, S. D. Boyd, A. M. Collins, DJ Pairing during VDJ Recombination Shows Positional Biases That Vary among Individuals with Differing IGHD Locus Immunogenotypes. J. Immunol. 196, 1158-1164 (2016).

26. S. D. Boyd, B. A. Gaeta, K. J. Jackson, A. Z. Fire, E. L. Marshall, et al., Individual variation in the germline Ig gene repertoire inferred from variable region gene rearrangements. J. Immunol. 184, 6986-6992 (2010).

27. M. J. Kidd, Z. Chen, Y. Wang, K. J. Jackson, L. Zhang, et al., The inference of phased haplotypes for the immunoglobulin H chain V region gene loci by analysis of VDJ gene rearrangements. J. Immunol. 188, 1333-1340 (2012).

28. M. Gidoni, O. Snir, A. Peres, P. Polak, I. Lindeman, et al., Mosaic deletion patterns of the human antibody heavy chain gene locus shown by Bayesian haplotyping. Nat. Commun. 10, 628 (2019).

29. H. Shirai, A. Kidera, H. Nakamura, H3-rules: identification of CDR-H3 structures in antibodies. FEBS Lett. 455, 188-197 (1999).

30. B. North, A. Lehmann, R. L. Dunbrack, Jr., A new clustering of antibody CDR loop conformations. J. Mol. Biol. 406, 228-256 (2011).

31. B. D. Weitzner, R. L. Dunbrack, Jr., J. J. Gray, The origin of CDR H3 structural diversity. Structure 23, 302-311 (2015).

32. B. J. DeKosky, T. Kojima, A. Rodin, W. Charab, G. C. Ippolito, et al., In-depth determination and analysis of the human paired heavy- and light-chain antibody repertoire. Nat. Med. 21, 86-91 (2015).

33. A. V. Collis, A. P. Brouwer, A. C. Martin, Analysis of the antigen combining site: correlations between length and sequence composition of the hypervariable loops and the nature of the antigen. J. Mol. Biol. 325, 337-354 (2003).

34. C. Busse, Data from "Dynamics of the human antibody repertoire after influenza vaccination." NCBI Sequence Read Archive. Available at https://www.ncbi.nlm.nih.gov/bioproject/PRJNA349143 (deposited 19 October 2016).

35. T. Magoc, S. L. Salzberg, FLASH: fast length adjustment of short reads to improve genome assemblies. Bioinformatics 27, 2957-2963 (2011).
36. Y. Chen, S. H. Kim, Y. Shang, J. Guillory, J. Stinson, et al., Barcoded sequencing workflow for high throughput digitization of hybridoma antibody variable domain sequences. *J. Immunol. Methods* **455**, 88-94 (2018).

37. J. Ye, N. Ma, T. L. Madden, J. M. Ostell, IgBLAST: an immunoglobulin variable domain sequence analysis tool. *Nucleic Acids Res.* **41**, W34-40 (2013).

38. S. F. Altschul, W. Gish, W. Miller, E. W. Myers, D. J. Lipman, Basic local alignment search tool. *J. Mol. Biol.* **215**, 403-410 (1990).

39. M. P. Lefranc, IMGT, the international ImMunoGeneTics database: a high-quality information system for comparative immunogenetics and immunology. *Dev. Comp. Immunol.* **26**, 697-705 (2002).
**Fig. 1.** Average CDR H3 length associated with $V_H$ and $V_L$ segments. Average CDR H3 lengths of TX dataset clonotypes (abscissa) correlated with the CDR H3 of clonotypes with $V_H$ (A) and CA $V_L$ germlines (B) in the AE compartment. Diagonal lines show 1:1 length correspondence between datasets. (C-F) Average CDR H3 length of TX and WA AE and naïve B cell compartments, segregated by isotype for the TX dataset. Solid horizontal bars indicate average CDR H3 length in the naïve B cell compartment. The order of germlines is the same in panels C, D and E. Lengths are shown in amino acid residues according to the IMGT® CDR definition.
Fig. 2. CDR H3 length distribution groups associated with V_H germlines. Characteristic examples of each V_H bias group are shown, averaged for TX, CA, MA and WA donors in the AE (and/or isotype switched) (A), TX naïve (B), WA naïve (C) and WA nonproductive compartments (D). Orange bars are germline-specific CDR H3 length distributions of unique clonotypes. Blue bars are overall CDR H3 length distributions of unique clonotypes. Blue and red lines above the distributions indicate range of CDR H3 lengths statistically significant different between germline-specific and overall length distributions in a paired (within donor) t-test (P < 10^{-4}) with a sliding window of two contiguous CDR H3 lengths, with red and blue indicating relative enrichment and depletion in the germline-specific distributions. Arrows and black bars highlight subtle depletion of CDR H3 sequence lengths consistent across datasets and B cell compartments. Error bars indicate S.E.M. (A and B) or range (C and D). The full set of distributions is shown in Fig. S2.
Fig. 3. Differentiation of CDR H3 length distributions by PCA analysis. PCA analyses of $V_H$ (A) and $V_L$ (B) germline-associated CDR H3 length distributions in the AE compartments are shown. Analysis in (A) excludes the WA dataset due to limited germline coverage. Analysis in (B) includes both CA and TX datasets. Each data point indicates a germline for which minimum count requirements were met for the analysis. Points are color-coded by bias group determined by visual inspection of CDR H3 length distributions (Fig. S2 and S6). Axes are oriented to position distributions skewed towards long lengths and with high apparent kurtosis to the right and top respectively.
Fig. 4. CDR H3 length distribution groups associated with V\textsubscript{L} germlines. Characteristic examples of each group are shown, averaged for all CA and TX donors in the AE and/or isotype switched compartments (A) and the naïve compartment of TX donors (B). Bar and line colors as in Fig. 1. Error bars indicate S.E.M. for data available from more than 2 donors. The full set of distributions is shown in Fig. S6.
Fig. 5. Modulation of $V_H$-associated CDR H3 length biases by $J_H$ segments in the WA naïve compartment. CDR H3 length distributions associated with individual $V_H/J_H$ combinations (orange bars) are compared to overall repertoire distributions associated with $J_H$ segments (blue bars). Arrows highlight biases dependent on $J_H$ segment use. Dotted lines indicate the peak of peaks in the distributions associated with $V_{H3}$ and all $J_H$ germlines in the same part of the length spectrum. The number of CDR H3 amino acid residues potentially encoded by $J_{H1}$-$J_{H6}$ segments are 6, 7, 6, 4, 6 and 9 respectively. Error bars indicate S.E.M. The full set of distributions is shown in Fig. S10.
Fig. 6. Biases in CDR H3 J<sub>H</sub>4 residue occupancy associated with different V<sub>H</sub> germlines and CDR H3 length in WA naïve sequences. Occupancy of J<sub>H</sub> germline-encoded residues in the last CDR H3 positions (IMGT 114 to 117) is shown. Only sequences with J<sub>H</sub>4 are included in the analysis. J<sub>H</sub> residues are color-coded by position as indicated in the top left panel. Solid lines indicate residue occupancy for all sequences in the naïve repertoire with the J<sub>H</sub>4 germline. Dots indicate average residue occupancy with each V<sub>H</sub> germline and CDR H3 length. Bars indicate S.E.M. for three donors except for V<sub>H</sub>3-9 and V<sub>H</sub>7-4-1, which are present in one donor each. Data points with fewer than 60 sequences were excluded.
Fig. 7. Length biases of regions within CDR H3 as a function of V\textsubscript{H} and J\textsubscript{H} germline use and CDR H3 length. (A) Deviations in nucleotides (Δnt) from whole repertoire averages are shown for the V\textsubscript{H}, D, J\textsubscript{H} segments and NP-region (insertions) in CDR H3 in WA naïve sequences grouped by V\textsubscript{H} and J\textsubscript{H} germlines. Colors for each segment are indicated in the top left panel. Values in parentheses indicate the maximum number of V\textsubscript{H} germline nucleotides that can be included in CDR H3. (B) As in panel (A), showing CDR H3 segment length deviations associated with V\textsubscript{H}3-15 sequences and different J\textsubscript{H} germlines in the SRI dataset (donors 326650, 326797 and 327059). Sequences with undefined D segments were omitted. Values shown are averages of deviations from overall repertoire, with deviations calculated within donors. Data points with fewer than 60 counts within donors were excluded. Error bars not shown for clarity. CDR H3 lengths are shown in amino acid residues. The full dataset is shown in Fig. S15.
|                  | Dataset |
|------------------|---------|
|                  | CA      | MA      | TX       | WA      | SRI     |
| Reference        | 19      | 21,33\(^a\) | 16      | 20      | 22      |
| Donors           | 3       | 3       | 3       | 3       | 8       |
| CD27 marker      | NA\(^b\) | NA      | CD27\(^{neg}\) (3) | CD27\(^{neg}\) (3) | NA |
|                  |         |         | CD27\(^{pos}\) (2) | CD27\(^{pos}\) (3) |         |
| Isotypes         | IgG     | IgG/IgA | IgM (CD27\(^{neg}\))\(^c\) | NA     | IgG/IgM |
|                  |         |         | IgG/IgA/IgM (CD27\(^{pos}\)) |         |         |
| Chain            | \(V_H/V_L\)^\(^d\) | \(V_H\) | \(V_H/V_L\)^\(^d\) | \(V_H\) | \(V_H\) |
| V/J region parsing | Absolve | IgBlast | As published | As published | As published |
| V-region coverage | Full-length | Full-length | Partial | Partial | Partial |
| Unproductive sequences | NA | NA | NA | CD27\(^{neg}\), frameshifted\(^e\) | NA |

\(^a\) Illumina re-sequencing data (31) of the original MA samples (19) was used. See Table S3 for details.

\(^b\) Not available

\(^c\) Isotype information ignored for CD27\(^{neg}\) IgM clones in the TX dataset and includes undefined isotype sequences.

\(^d\) Paired \(V_H/V_L\) variable region sequence data

\(^e\) Stop codon unproductive sequences not used to minimize impact of sequencing errors. Unproductive sequences in the CD27\(^{pos}\) subset not used to minimize clonal expansion effects. CD27\(^{neg}\), frameshifted unproductive sequences were not filtered for clonotype.