Dynamics of heavy chain junctional length biases in antibody repertoires

Kannan Sankar [1], Kam Hon Hoi[1,2] & Isidro Hötzel [1✉]

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 and junctional segment length distributions are 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 naive and antigen experienced B cell compartments but not in nonproductive recombination products, indicating B cell selection as a major driver of these biases. Our findings reveal biases in the antibody CDR H3 diversity landscape shaped by V_H, V_L, and J_H germline segment use during naive and antigen-experienced repertoire selection.

[1] Department of Antibody Engineering, South San Francisco, CA 94080, USA. [2] Department of Bioinformatics and Computational Biology, Genentech, South San Francisco, CA 94080, USA. ✉email: ihotzel@gene.com

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 germline segments while the V_H region is created by recombination of V_H, D_H, and J_H germline segments. 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_H germline segment mediated by terminal deoxynucleotidyl transferase and germline segment palindromic duplications[3]. Self-reactive B cell clones with full-length, in-frame variable regions have V_L sequences replaced by receptor editing or are removed from the repertoire by apoptosis[4,5]. Cells passing this self-reactivity checkpoint form the naive B cell repertoire[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] and the 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_H and J_H germline segment sequences of different lengths and sequence composition into CDR H3 of different lengths[6]. Beyond the germline segment biases, CDR H3 sequence biases that reflect underlying selective biases in B cell maturation have also been described. In particular, a bias towards shorter average CDR H3 lengths is observed in mature relative to immature B cells and in isotype-switched memory B cells relative to naive to B cells[9,10,14]. 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[9,11,15,16].

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 germline segments[6–11,17–20]. Except for sequences that are directly incorporated into CDR H3, the impact of V germline segments on CDR H3 properties has neither been addressed nor expected. Analysis of the impact of the V_L on CDR H3 has been limited to properties of the third CDR of the light chain, which is closely associated with CDR H3, without any evidence of biases[14]. Finally, analysis of the impact of J_H germline segments on CDR H3 biases has been confined to the expected effects of differential J_H germline segment length and sequence composition[6]. 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_H3-7) and Vκ2-30 germline segments[21]. This was determined upon further analysis to be linked with short CDR H3 length biases associated with both germline segments. This unexpected finding prompted us to undertake a high-dimensional analysis of CDR H3 sequences from several human antibody datasets to investigate the extent to which CDR H3 length might be biased by germline segment use in human immunoglobulin repertoires. Our results show several biases in CDR H3 and junctional length distributions associated with V_H, V_L, and J_H germline segment utilization that shape naive and antigen-experienced antibody repertoires in unexpected and unpredictable patterns.

Table 1 Summary of datasets used for analysis.

	Dataset				
	CA	**MA**	**TX**	**WA**	**SRI**
Reference	21	22,24[a]	23	25	20
Donors	3	3	3	3	8
CD27 marker	NA[b]	NA	Available[c]	Available	NA
Isotypes	IgG	IgG/IgA	IgG/IgA/IgM[d]	NA	IgG/IgM
Chain	V_H/V_L[e]	V_H	V_H/V_L[e]	V_H	V_H
Unproductive sequences	NA	NA	NA	Yes[f]	NA

[a]Illumina re-sequencing data (24) of the original MA dataset samples (22) was used.
[b]Not available.
[c]Two donors with CD27[pos] and three donors with CD27[neg] B cells.
[d]Isotype information ignored for CD27[neg] IgM clones in the TX dataset.
[e]V_H/V_L variable region sequence paring information available.
[f]CD27[neg], frameshifted sequences used only.

Results

Sequence datasets. We analyzed sequences from four previously described B cell repertoire deep sequencing datasets including three donors each and a fifth dataset with eight donors (used here as an independent test dataset in most analyses to avoid over-representation of donors from a single source in combined data), referred to here as the TX, WA, CA, MA, and SRI datasets[20–25]. These represent the largest publicly available datasets with V_H/V_L pairing (TX and CA) or V_H-only information (MA, WA, SRI). The datasets were previously sequenced and bioinformatically parsed using a diversity of methods, minimizing the impact of methodological biases. A summary of the datasets including the number of donors, CD27 marker and isotype information has been summarized in Table 1 (see Supplementary Table 1 for details). Data was normalized by analyzing each donor individually and averaging within or among datasets to avoid over-representation of sequences from larger datasets. 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[21,22]. For simplicity we refer to the TX CD27[pos] IgG/IgA, CA IgG, MA IgG/IgA, WA CD27[pos], and SRI IgG subsets as TX AE, CA AE, MA AE, WA AE, and SRI AE, respectively (with "antigen-experienced", or "AE", encompassing all five), the TX CD27[pos] IgM sequences as "AE IgM" and the TX and WA CD27[neg] subsets as TX naive and WA naive respectively (with "naive" encompassing both, Supplementary Table 1). SRI IgM sequences with no amino acid somatic mutations between Cys-23 and Cys-104 (International ImMunoGeneTics, IMGT®, numbering system[26] used throughout) including most of the region covered by reads are referred as SRI naive.

We aimed at identifying properties shared among donors not influenced by clonal expansion. To minimize the impact of clonal expansion, all datasets except nonproductive sequences were processed to retain a single sequence from each lineage, or clonotype, by clonotype clustering, according to germline segments as well as CDR H3 length and sequence similarity (see 'Clonotype clustering' under Methods and Supplementary Data 1, 2, and 3 for unique sequence counts after clustering). Overall distribution of CDR H3 lengths was not noticeably affected by removal of redundant sequences in most datasets except for the WA AE and MA AE compartments, which had subtle shifts (Supplementary Fig. 1a-f). The overall AE CDR H3 length distributions are similar among datasets except SRI, allowing pooling data of AE B cell subsets from different datasets (Supplementary Fig. 1g). However, the relative CDR H3 length distributions of the WA and TX naive B cell subsets differed by an

average of 0.9 residues (Supplementary Fig. 1h) and were analyzed separately. Germline segment-specific analyses were performed with germline segments with at least 80 unique clonotypes in a donor, which, in aggregate, include 94–99% of the unique clonotype sequences in the CA, TX, MA and SRI datasets (Supplementary Data 1, 2, and 3). Germline segment-specific analyses in the WA dataset were performed with 16 V_H germline segments that had fewer than 10% ambiguous calls in the naive subset and germline segments prevalent in other datasets, totaling about one third of the entire dataset (Supplementary Data 4).

Average CDR H3 length varies with V_H and V_L germline segment use. We analyzed average CDR H3 length by V_H or V_L germline segment use. Average CDR H3 length in the AE subset varied by up to 3 amino acid residues as a function of V_H germline segment use and correlated well for different datasets when compared to the TX dataset (Spearman's $r = 0.72$–0.84) (Fig. 1a). Average CDR H3 length also varied as a function of V_L germline segment use by up to four amino acid residues in the AE

compartment and correlated well between the CA and TX datasets (Spearman's $r = 0.93$) (Fig. 1e). The naive compartment showed a more limited spread in average CDR H3 lengths relative to the AE compartment (Fig. 1b, c, d and f, blue squares). More pronounced reductions in average CDR H3 length in the AE relative to naive compartments were associated with a subset of V_H and V_L germline segments (Fig. 1b, c, d and f). The TX AE IgM subset showed similar trends as the TX AE subset except that average CDR H3 length was decreased in association with most V_H germline segments relative to the naive compartment (Fig. 1b, pink dots).

CDR H3 length distribution varies with V_H germline segment use. We next determined whether CDR H3 length distribution varies with germline segment use. Overall and germline segment-specific CDR H3 length distributions were performed individual for each donor and the frequency of each CDR H3 length averaged within or across the TX, CA, MA, and WA datasets (11 donors), with the eight donors from the SRI dataset analyzed
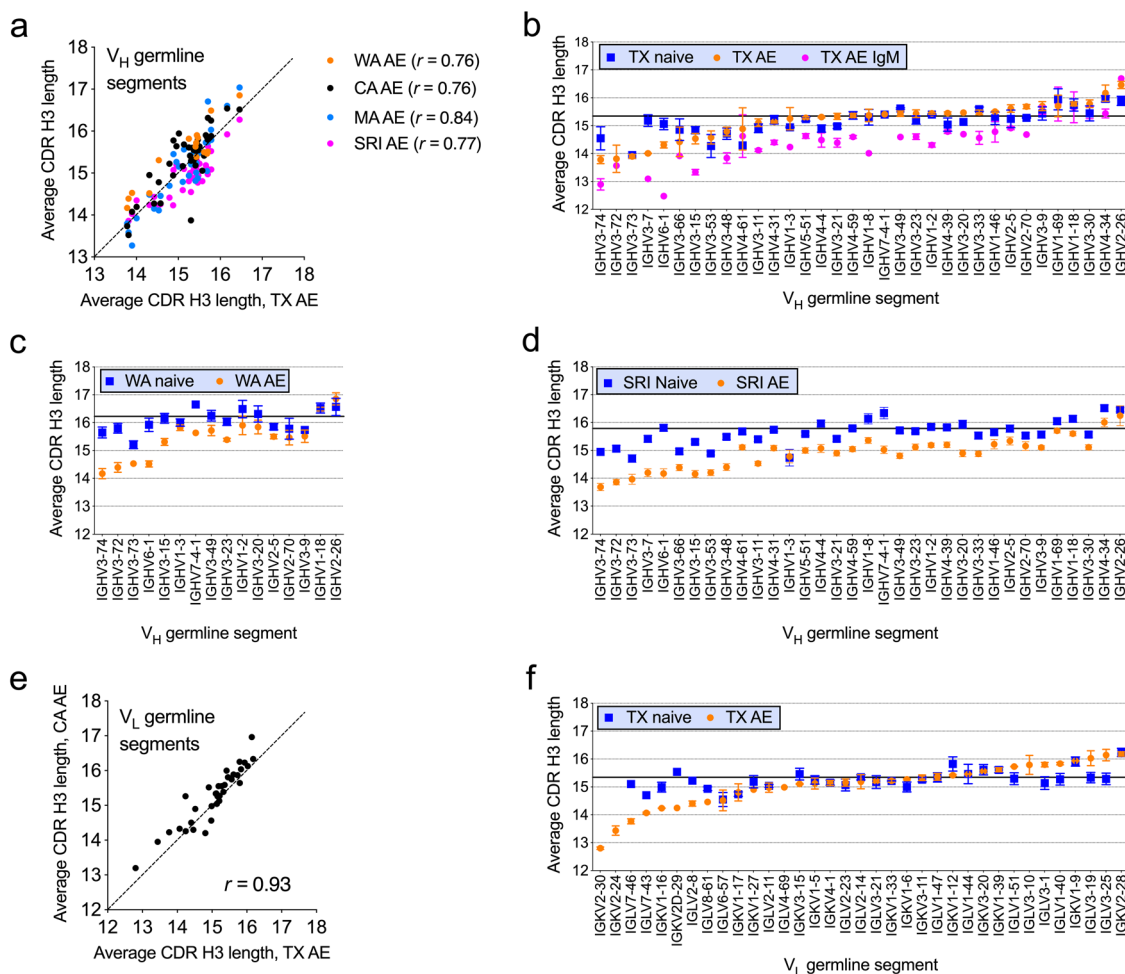

Fig. 1 Average CDR H3 length associated with V_H and V_L germline segments. Average CDR H3 lengths associated with different germline segments in the TX AE dataset (abscissa) correlated with the CDR H3 associated with V_H germline segments of different datasets (**a**) and CA V_L germline segments (**e**) in the AE compartments. Each dot represents the average CDR H3 length for a different V_H or V_L germline segment within each dataset, averaged across donors within each dataset. Diagonal lines show 1:1 length correspondence between datasets, with Spearman's correlation coefficients shown. **b–d** Average CDR H3 length associated with V_H germline segments in the TX AE, TX AE IgM, WA AE, SRI AE and corresponding naive B cell datasets. Solid horizontal bars indicate average CDR H3 length in the naive B cell compartment. The order of germline segments is the same in the panels. **f** Average CDR H3 length associated with V_L germline segments in the TX AE and naive B cell datasets. Solid horizontal bars indicate average CDR H3 length in the naive B cell compartments. Error bars indicate S.E.M. for $n = 3$ donors or range for $n = 2$ donors. Germline segments without error bars have only 1 donor with at least 80 sequences (TX dataset, see Supplementary Data 1, 2 and 3) or sequences mostly represented by 1 donor (WA, see Supplementary Data 1 and 2). Lengths are shown in amino acid residues according to the IMGT® CDR definition[26].

separately. Overall CDR H3 length distribution of the respective B cell compartment, which is influenced by germline segment frequency in the datasets, was used as a relative standard to which germline segment-specific CDR H3 length distributions were compared. This was done to facilitate comparison of biases between germline segments and also because useful objective reference distributions are not available to determine bias types in naive compartment sequences. Therefore, most biases described here, including all in the naive compartment, are relative to the entire set of clonotypes in each B cell compartment. 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 impact of local fluctuations. Observed length distribution biases included overall shifts in average CDR H3 length for sequences with different V_H germline segments and also obvious and subtle deviations from the overall CDR H3 distribution in discrete ranges of the length spectrum (Fig. 2, top row, Supplementary Fig. 2a).

To further discern the CDR H3 length biases quantitatively, we performed a principal component (PC) analysis of the length distributions (lengths 5–26) associated with different V_H germline segments. Results from the PC analysis were visualized by projecting each germline segment onto the most important trends to obtain the so-called PC scores, aided by a visual analysis of the corresponding distributions (Fig. 3a). PC1 and PC2 corresponded to apparent skewness and kurtosis of the distributions, respectively. Using the PC analysis results in conjunction with visual inspection of V_H germline segment-associated CDR H3 distributions in the AE compartment, germline segments were categorized by bias type as "Short", "Neutral" and "Long" (Figs. 2 and 3a, Supplementary Fig. 2). Those germlines that had similar length distributions as the overall distribution were called "Neutral" (located around the center of the PC plot); while those with shifts towards longer or shorter lengths as "Long" (right-skewed, with low values of PC1) and "Short" (left-skewed, with high values of PC1) respectively. Within each of these classes, some germlines also showed varying degrees of kurtosis relative to the overall distribution (extreme values in PC2). Differences between the distributions of members of different groups can be subtle, both visually and in the PC analysis. The magnitude of the biases and the details of distribution shapes varied for different V_H germline segments within each group but were consistent across datasets for each germline segment (Supplementary Fig. 3). Germline segments in the same V_H subfamily did not always have the same biases. The range of germline segment prevalence in the various datasets was similar for the different bias groups (Supplementary Fig. 4a).

We determined whether the observed distribution biases were also present in the naive B cell subset. The biases of the Long
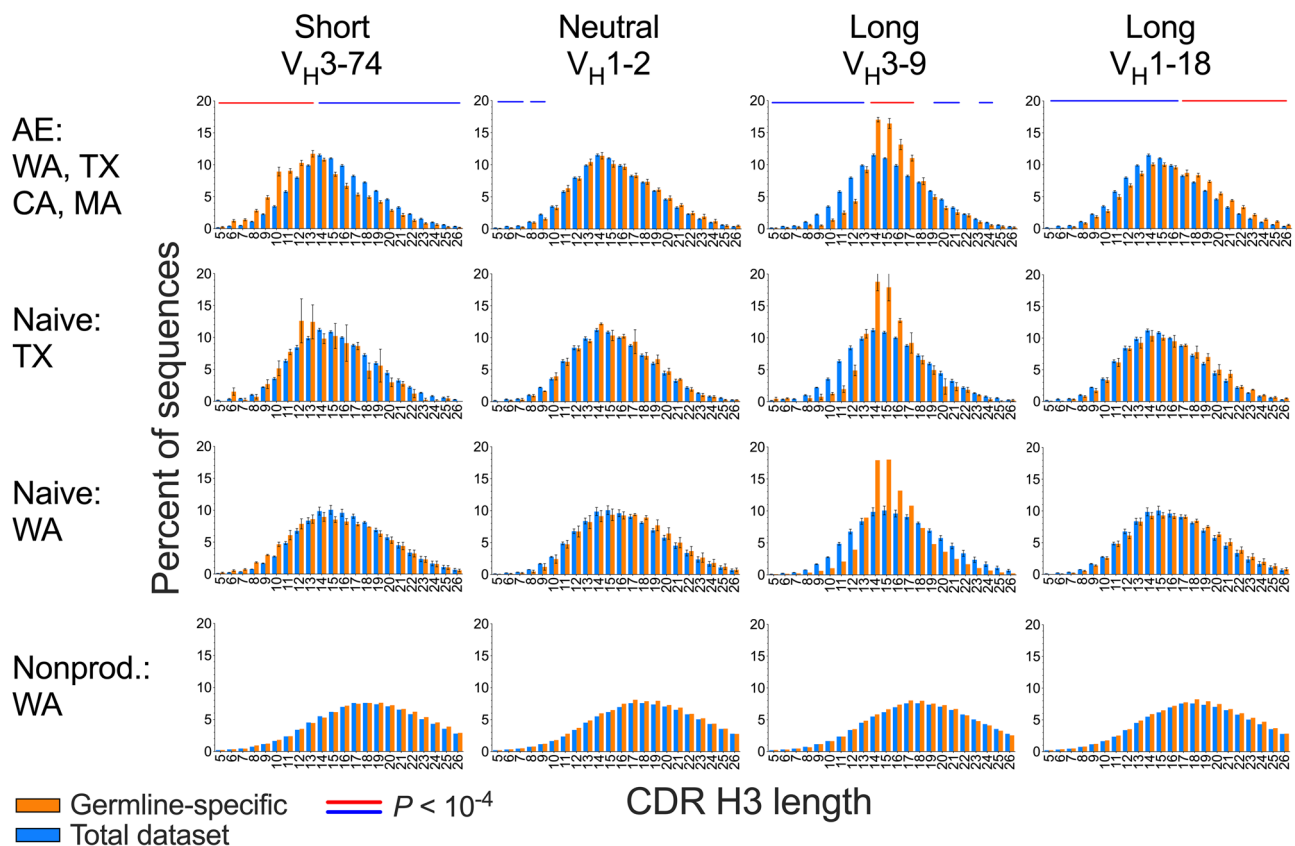

Fig. 2 CDR H3 length distribution groups associated with V_H germline segments. Characteristic examples of each V_H bias group are shown, averaged for TX, CA, MA and WA donors in the AE compartment (top row, $n = 11$ donors), TX naive (second row, $n = 3$ donors), WA naive (third row, $n = 3$ donors) and WA nonproductive compartments (bottom row, three pooled donors). Orange bars are germline segment-specific CDR H3 length distributions of unique clonotypes. Blue bars are overall CDR H3 length distributions of unique clonotypes. Blue and red horizontal lines above the distributions indicate range of CDR H3 lengths statistically significant different between germline segment-specific and overall length distributions in a two-tailed paired *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 segment-specific distributions. Distributions were determined for each donor individually followed by averaging across donors, except for the nonproductive sequences in which donors were combined prior to calculation of distributions. Error bars indicate S.E.M. The full set of distributions is shown in Supplementary Fig. 2.
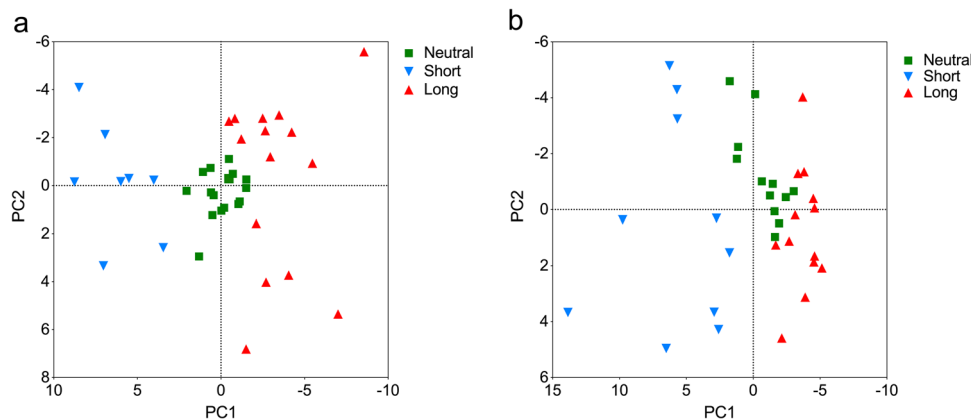
Fig. 3 Differentiation of CDR H3 length distributions by PC analysis. PC analyses of V_H (**a**) and V_L (**b**) germline segment-associated CDR H3 length distributions in the AE compartments are shown. Analysis in **a** excludes the WA dataset due to limited germline segment coverage. Analysis in **b** includes both CA and TX datasets. Each data point indicates a germline segment 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 (Supplementary Fig. 2 and 6). Axes are oriented to position distributions skewed towards long lengths and with high apparent kurtosis to the right and top respectively.

group were also observed in the naive B cell compartment, without apparent differences relative to the distributions in the AE compartments (Fig. 2, Supplementary Fig. 2). All the germline segments in the Neutral group showed average CDR H3 length distribution in the naive subset as well. However, distribution biases of the Short group in the naive compartment were less prominent (Fig. 2 and Supplementary Fig. 2), consistent with the average CDR H3 length analysis (Fig. 1b and c). Short biases in the naive compartment were mostly limited to the V_H3-73 and V_H3-15 germline segments in the TX and WA datasets. Despite the differences in overall CDR H3 length between the TX and WA naive datasets, the biases in the naive compartment had the same trends in both datasets (Fig. 2 and Supplementary Fig. 2).

The data analysis was performed with datasets 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[27] (see Methods for more details). Overall, no apparent increase in RSI scores indicating clonal expansion was associated with regions of positive prevalence biases in different parts of the CDR H3 length spectrum for the different bias groups (Supplementary Fig. 5a), confirming that clonal expansion does not account for the observed CDR H3 length biases.

CDR H3 length distribution biases are not generated by recombination. We next determined whether the biases observed in the naive compartment are a direct consequence of biases in the VDJ recombination process for each germline segment. For this, we analyzed frameshifted, nonproductive V_H sequences that were part of the naive WA dataset. Nonproductive recombination products are not directly subject to selection and therefore provide information about recombination products prior to any repertoire selection. As previously reported[16], the CDR H3 lengths of nonproductive V_H genes are longer than the productively recombined genes in mature B cell subsets (Fig. 2 and Supplementary Fig. 2). However, CDR H3 length for the nonproductive sequences associated with different V_H germline segments approximated a Gaussian distribution, with no observable biases associated with different V_H germline segments relative to the overall dataset, except for minor anomalies associated with some germline segments (Fig. 2 and Supplementary Fig. 2). Therefore, heavy chain recombination mechanisms do not

account for the naive repertoire CDR H3 length distribution biases.

CDR H3 length distribution varies with V_L germline segment use. We performed a similar analysis of CDR H3 length distribution as a function of V_L germline segment and B cell compartment using PC and visual analysis. Similar to the V_H germline segment-associated biases, V_L-associated biases in the AE compartment could be classified into three groups based on the skewness of their distributions, named here "Short" (right-skewed, with high value of PC1), "Long" (left-skewed, with 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 Vκ and V_λ germline segments (Figs. 3b and 4, top row, and Supplementary Fig. 6). 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 marked deviations from Gaussian, including localized frequency spikes in discrete sections in the short range. Only Vκ germline segments in the Long group were associated with similar CDR H3 length biases in the TX naive compartment (Fig. 4, Supplementary Fig. 6). The magnitude of the V_L-associated biases varied for different germline segments within each bias group but were consistent between datasets (Supplementary Fig. 7). As above, the RSI analysis results indicated that clonal expansion does not account for the V_L germline segment-associated CDR H3 length biases (Supplementary Fig. 5b). The prevalence of Short group germline segments in the dataset was lower than for germline segments of the other two groups (Supplementary Fig. 4b).

CDR H3 length is biased as a function of V_H/J_H combination. J_H germline segments vary in the number of amino acid residues that can be potentially contributed to CDR H3 from 4 in J_H4 to 9 in J_H6. We assessed whether differential J_H germline segment usage as a function of V_H and V_L germline segment use is the basis for V segment-associated CDR H3 length biases. No clear deviations from average J_H usage were observed in association with most V_H germline segments in the WA unproductive sequences (Supplementary Fig. 8a and 9, top panel). The observed deviations in J_H prevalence do not readily explain CDR H3 distribution biases associated with V_H and V_L germline segments (Supplementary Fig. 8b, c and d, and Supplementary Fig. 9,
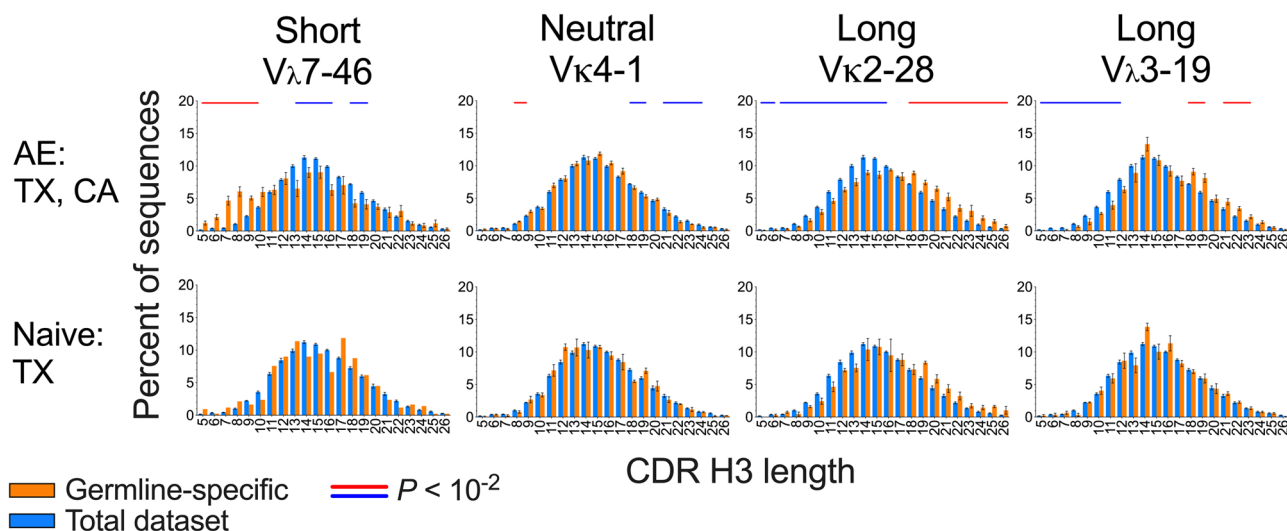
Fig. 4 CDR H3 length distribution groups associated with V_L germline segments. Characteristic examples of each group are shown, averaged for all CA and TX donors in the AE compartments and the naive compartment of TX donors. Blue and red horizontal lines above the distributions indicate range of CDR H3 lengths statistically significant different between germline segment-specific and overall length distributions in a two-tailed paired (within donor, $n = 5$ donors) t-test ($P < 10^{-2}$) with a sliding window of two contiguous CDR H3 lengths, with red and blue indicating relative enrichment and depletion in the germline segment-specific distributions. Number of donors is 5 for TX AE and CA AE combined and 3 for TX naive except for $V_\lambda 7$-46 in TX naive, where $n = 1$. Distributions were determined for each donor individually followed by averaging across donors. Error bars indicate S.E.M. The full set of distributions is shown in Supplementary Fig. 6.

bottom panel) with the exception of V_L germline segment $V\kappa 2$-28 in the Long CDR H3 bias group, which was associated with a higher prevalence of the longer $J_H 6$ and lower prevalence of the shorter $J_H 4$ germline segments in the heavy chain (Supplementary Fig. 8d).

We next analyzed CDR H3 length distributions associated with different V_H/J_H germline segment combinations, comparing these to the CDR H3 length distribution of all sequences with the corresponding J_H germline segment. As expected, CDR H3 length distributions were generally shifted according to length of the J_H segment in the germline regardless of V_H germline segment (Fig. 5, Supplementary Fig. 10 and 11). However, a subset of V_H-associated CDR H3 length biases were impacted by J_H germline segment in a manner independent of length of the J_H segment in the germline, with very similar patterns in the naive WA and naive SRI subsets (Fig. 5, Supplementary Fig. 10 and 11). These included a short CDR H3 length bias associated with sequences with the $V_H 3$-72, $V_H 3$-73, and $V_H 3$-15 germline segments combined with the $J_H 5$ and/or $J_H 4$ germline segments (Fig. 5, Supplementary Fig. 10 and 11). Additional CDR H3 length biases were observed for other V_H/J_H germline segment combinations (Supplementary Fig. 10). Our results indicate that CDR H3 length distribution biases are not necessarily uniform for each V_H germline segment but may vary in association with J_H segment. In addition, the effect of J_H on CDR H3 length distribution is not necessarily similar within V_H bias groups, indicating some degree of heterogeneity within bias groups.

Differentially trimmed J_H segments in the naive compartment. The CDR H3 length distribution biases associated with a subset of V_H/J_H germline segment combinations may be a consequence of biases in J_H trimming as a function of V_H germline segment. J_H residue occupancy in the last CDR H3 positions of $J_H 4$ and $J_H 5$ sequences was used to indirectly determine J_H trimming. The $J_H 1$, 2, 3, and 6 germline segments were not analyzed due to lack of sufficient data or, in the case of $J_H 6$, limited CDR H3 length biases associated with it. No apparent biases in J_H residue occupancy relative to the overall dataset was observed for any of the

analyzed V_H/J_H combinations in the nonproductive WA sequences (Supplementary Fig. 12). However, J_H residue trimming biases were observed for different V_H/J_H combinations in the WA naive compartment (Fig. 6 and Supplementary Fig. 12). General trends in residue occupancy in $J_H 4$ were similar in SRI naive sequences for the $V_H/J_H 4$ germline segment combinations with sufficient numbers for analysis (Supplementary Fig. 13). Residue-specific trimming biases were mostly coordinated for consecutive J_H residues in each analyzed V_H/J_H combination, as expected due to the directional nature of trimming. However, closely related V_H germline segments can be associated with distinct trimming biases of different $J_H 4$ residues. For instance, $V_H 2$-5/$J_H 4$ sequences are associated mostly with reduced trimming of IMGT® residue 115 (Phe) whereas for $V_H 2$-70/$J_H 4$ strongly reduced trimming of residue 116 (Asp) was also observed (Fig. 5 and Supplementary Fig. 12). The results indicate a complex set of constraints leading to selection of differentially trimmed J_H segments in the context of certain V_H and J_H germline segments during naive repertoire maturation.

Discussion
Understanding antibody CDR H3 diversity generation, a process critical for the availability of immune receptors binding a wide range of antigens, 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 and in different B cell developmental stages[6–20]. CDR H3 and junctional segments lengths in the B cell repertoire have been assumed to be independent of V_H and J_H germline segments except for their lengths prior to recombination[17,19,20]. This assumption has been implicitly used, for example, to generate simulated CDR H3 datasets to estimate the significance of observed clonal convergence in repertoires[20]. Analysis of the impact of V_L on CDR H3 length has been limited to CDR L3 length, with negative results[23]. 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 segment use and B cell
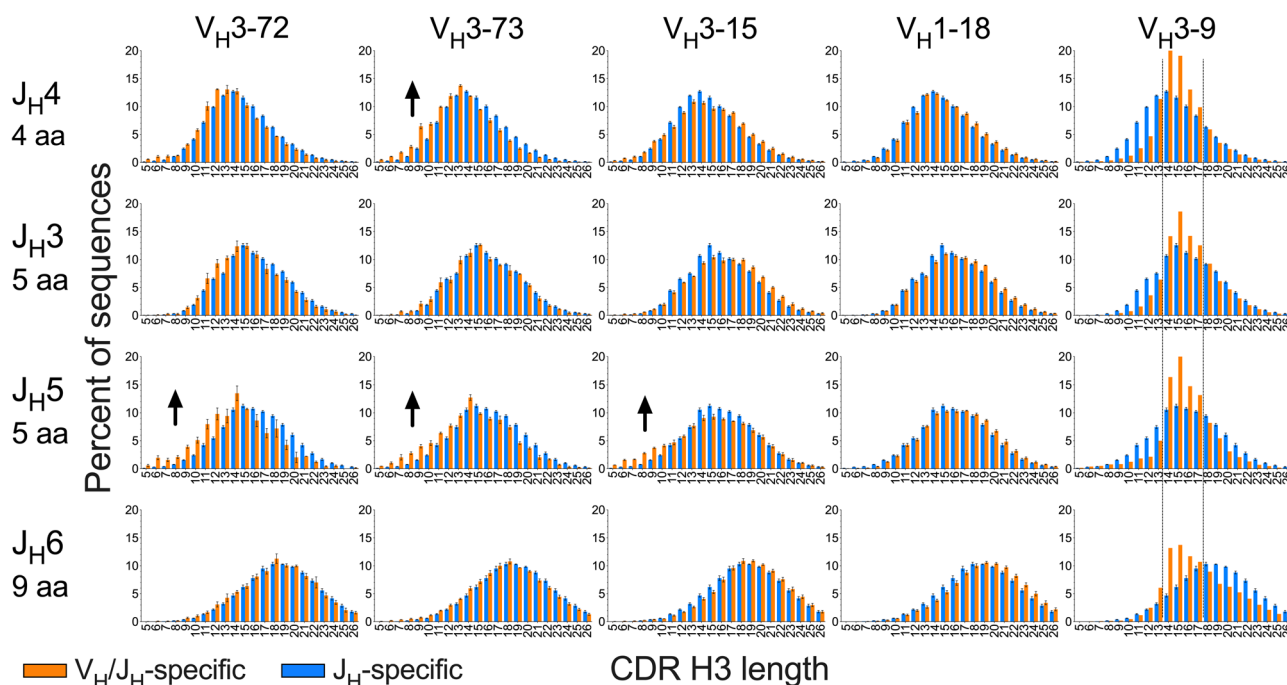
Fig. 5 Modulation of V_H-associated CDR H3 length biases by J_H germline segments in the WA naive compartment. CDR H3 length distributions associated with individual V_H/J_H combinations (orange bars) are compared to overall repertoire distributions associated with J_H germline segments (blue bars). Arrows highlight biases dependent on J_H segment use. Dotted lines indicate the of peaks in the distributions associated with V_H3-9 and all J_H germline segments in the same part of the length spectrum. The number of CDR H3 amino acid (aa) residues potentially encoded by the J_H3 to J_H6 germline segments is indicated in row legends. Distributions were determined for each donor individually followed by averaging across donors. Error bars indicate S.E.M. ($n = 3$). The full set of distributions is shown in Supplementary Fig. 10. The figure panels are ordered vertically by the number of residues the J_H segments can contribute to CDR H3.
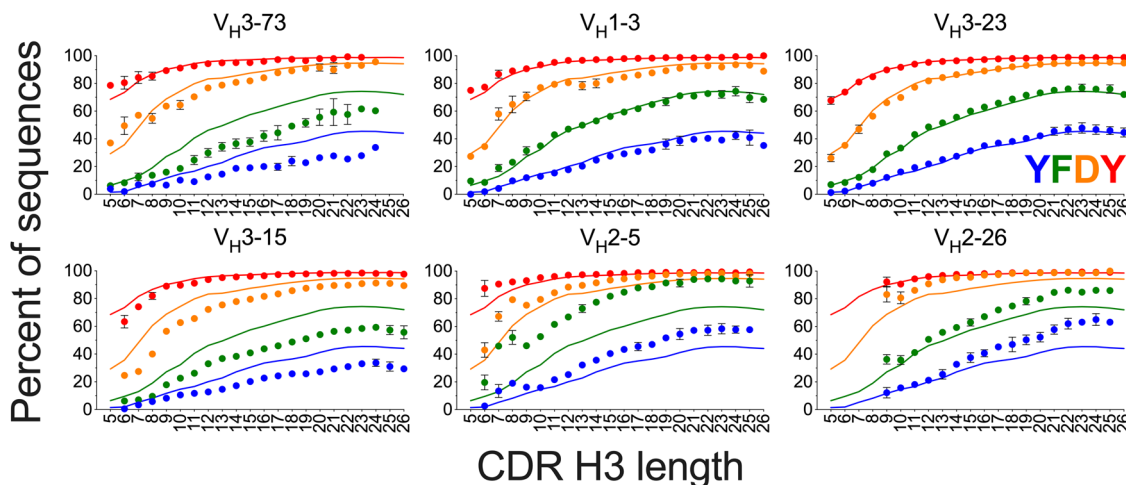

Fig. 6 Biases in J_H4 germline segment residue occupancy in CDR H3 associated with different V_H germline segments and CDR H3 length in WA naive sequences. Occupancy of J_H germline segment-encoded residues in the last CDR H3 positions (IMGT® 114–117) is shown. J_H4 residues are color-coded by position as indicated in the V_H3-23 panel by "YFDY", the four J_H4 residues that can be included in CDR H3 in positions 114–117. Solid lines indicate average residue occupancy for all sequences in the naive repertoire with the J_H4 germline segment. Dots indicate average residue occupancy with each V_H germline segment and CDR H3 length. Bars indicate S.E.M. ($n = 3$ donors). Data points with fewer than 80 sequences were excluded.

maturation state that are not predictable from germline segment lengths and amino acid sequences. Most of the length and junctional biases we describe are evident in the naive B cell compartment but not in the nonproductive subset, indicating a major role of naive B cell repertoire maturation and, by extension, selection against self-reactivity or for structural integrity as likely factors in shaping those biases. In addition, only a subset of V_H or V_L germline segments is associated with biases towards shorter CDR H3 lengths in the antigen-experienced compartment, indicating general germline segment-specific adaptive immunity selection processes shared among individuals. Similar T cell receptor β chain CDR3 length distribution biases with different *TRBV* germline segments in repertoires arise in the process of T cell maturation[28], although selective processes may differ between

B and T cell repertoires due to differences in mechanisms of antigen recognition. Our results provide a detailed view into the dynamics of heavy chain junctional biases in antibody repertoires that complements previously described dynamics of clonal selection and expansion, convergence, sequence diversity and changes in overall CDR H3 properties in B cell maturation[9–11,20,22,25,27,29].

Special consideration was given to the repeatability and robustness of the findings. The results are based on a total of 12 donors in four datasets 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, minimizing the impact of technical artifacts. Some of the biases, such as those associated with V_L, and J_H cannot be easily generated by sequencing or parsing artifacts, especially in a systematic fashion across datasets. The stringency of clonotype clustering criteria had limited impact on results. This is exemplified by the WA and SRI datasets, which yielded CDR H3 length distributions similar to other datasets (Supplementary Fig. 3) despite having been clustered by clonotype using a higher CDR H3 sequence identity threshold than other datasets (Supplementary Table 1).

A recurring theme in the results presented here was that biases observed at one level (e.g., V_H germline segment) 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 higher-dimensional analyses including other repertoire descriptors will uncover additional biases, two examples being D_H junctional length biases (Supplementary Fig. 14, 15, and 16 and Supplementary text) and CDR H3 length biases associated with different V_H germline segment allelic variants (Supplementary Fig. 17). Haplotype variations could also potentially affect CDR H3 length distributions in a V_H allele-dependent manner or through differences in D_H germline segment composition and differential recombination frequencies of D_H or J_H germline segments of different lengths in different chromosomes, combined with differential recombination frequencies of V_H alleles[30–32]. However, the observation of essentially the same CDR H3 length distribution biases in several donors from 5 different sources and junctional segment length biases in 11 donors from 2 of these sources, along with a lack of systematic associations between V_H, D_H, and J_H alleles across donors[31,32], indicates that haplotype variations are unlikely to be a major factor in the CDR H3 and junctional length distribution biases described here. In addition, heavy chain variable region haplotype differences would not be expected to impact CDR H3 distributions associated with V_L germline segments and the AE compartment-specific short CDR H3 length biases.

The analyses shown here use germline segment information as a proxy for undefined sequence features that ultimately determine the observed biases. The selected CDR H3 sequence and structural properties that result in the observed biases and the germline segment sequence properties that determine those biases remain to be identified. Analysis of V_H germline segment residues that can directly encode or bias the first CDR H3 residues in IMGT® positions 105–107 did not reveal clear correlations between these and most CDR H3 bias groups or junctional segment length biases (Supplementary Fig. 18). In addition, no obvious correlations between J_H trimming biases and variations in V_H germline segment residues in positions 40–42 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[33,34]. The bulged conformation appears to depend on the Ig domain, encoded mostly by V_H germline segments[34,35]. Whether V_H germline segment-

dependent J_H trimming biases reflect biases in the structure of the CDR H3 base remains to be determined. One challenge in determining how different germline segment regions determine the observed biases is the relatively limited number of non-redundant human antibody structures with different V_H/J_H combinations or V_L germline segments with different 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 naive repertoire suggests selection against self-reactivity, selection for structural integrity, expression or a combination of these factors as possible mechanisms. If related to selection against self-reactivity, the different biases indicate either that features other than CDR H3 charge and hydrophobicity contribute to self-reactivity or that V germline segments outside CDR H3 modulate the self-reactivity mediated by these factors. The possible role of selection against self-reactivity may allow leveraging the biases observed in CDR H3 in large natural human antibody repertoire deep sequencing datasets to understand CDR H3 properties beyond charge that correlate with antibody polyspecificity and self-reactivity, a subject of considerable interest in therapeutic antibody development and in the understanding of functional consequences of immune disorders with altered CDR H3 sequence profiles[11,36,37]. The bias towards shorter CDR H3 lengths associated with a subset of V_H and V_L germline segments in the AE compartment may be attributable to these same mechanisms or to antigen-driven immune selection. The latter would suggest widespread convergences in human repertoires associated with certain V_H and V_L germline segments or, possibly, some degree of functional specialization in the germline segment repertoire linked to short CDR H3 sequences, analogous to the association between CDR H3 length and recognition of different antigen classes[38].

Our results point to unexpected cross-constraints between V_H, V_L, J_H, and other junctional elements selected at different stages of B cell development that shape CDR H3 and junctional length distributions in antibody repertoires. That is, CDR H3 length distribution in the repertoire is not independent of heavy and light chain V_H, V_L, and V_H/J_H germline segment usage as implicitly assumed[6–11,17–20]. Instead, overall CDR H3 and junctional length distributions in antibody repertoires are aggregates of several sub-repertoires with discrete sets of biases relative to each other that arise in different stages of B cell maturation as a function of V_H, V_L, and J_H germline segment use, shared by normal donors. Thus, proper description of CDR H3 length biases in disease and immune states[11,12,39] requires the context of V_H, V_L, or V_H/J_H germline segment usage in which these biases are observed for meaningful interpretation. The analyses described here provide a high-dimensional CDR H3 analytical framework, in which CDR H3 and junctional length distributions are analyzed in the context of V_H, V_L, and J_H germline segments and combinations of these, and a baseline of these biases with multiple healthy donors for further studies of B cell repertoire maturation and clonal selection in health and disease.

Methods

Datasets and analysis. Sequences were obtained from the original publications[14,21,25] except for the MA dataset. The sequences in the MA dataset were obtained from a re-sequencing by Illumina MiSeq of a set of previously described samples[22], deposited in the Sequence Read Archive (SRA) database[24]. A summary of the samples used here is given in Supplementary Data 5. 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). The SRI dataset donors included in the analyses are 316188, 326650, 326737, 326780, 326797, 326907, 327059, and D103[20]. SRI dataset donors 326713 and 326651 were only used for V_H germline segment allele-specific analyses. Sequences were used as

parsed in the original publications except for sequences of the MA dataset obtained from the SRA database[24], where the raw sequencing files were processed and germline segments annotated with a custom pipeline (available from docker hub repository at https://hub.docker.com/r/kamhonhoi/iganalysis). Briefly, paired-end reads were merged using FLASH[40] 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[41]. IgBlast[42] was used to supplement the region parsed data with germline segment 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 segment use and clonotype grouping (see below). Nonproductive sequences in the WA dataset used for analyses were limited to frameshifted sequences in the naive compartment to minimize the indirect effects of sequencing errors and clonal expansion. CDR H3 length of nonproductive, frameshifted sequence length in amino acid residues was set as the nearest integer of CDR H3 length in nucleotides divided by 3. For naive compartment sequences of Donor 1 of the WA dataset only the D1a repeat was used for most analyses[25]. All CDR H3 length distributions and germline segment prevalence analyses were determined using custom scripts and Microsoft Excel 2016. The IMGT® CDR definition and numbering system are used throughout[26].

Clonotype clustering. Clonotypes in the CA and TX datasets were defined as sequences from the same donor, V_H and V_L germline segments 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. IgG/IgA and IgM sequences were segregated prior to clonotype clustering. For the TX dataset IgG/IgA, CD27^{pos}/IgM and CD27^{neg} sequences were segregated prior to clonotype clustering. The 11% of sequences without isotype information in the TX AE datasets, which included IgG, IgA, and IgM sequences, were excluded from the analyses due to the differences between IgG/IgA and IgM overall distributions in the AE compartment (Fig. 1b). Clonotypes in the MA dataset were defined as sequences from the same donor, V_H and J_H germline segments and CDR H3 length with a nominal 57% (average 60% identity across CDR H3 lengths) or greater CDR H3 amino acid sequence identity as above. Clonotypes in the SRI dataset were defined as sequences from the same donor with the same V_H and J_H germline segment, isotype and same CDR H3 length and sequence. Only sequences labeled as "productive" in the SRI dataset were analyzed. Only a randomly chosen sequence from each clonotype was retained in the dataset for the TX, CA, WA, MA, and SRI datasets. Clonotypes in the WA dataset were defined as sequences from the same donor with the same V_H and J_H germline segment and same CDR H3 length and sequence. If V_H germline segment information was not available then V_H subfamily information was used in lieu, retaining as a representative for the clonotype a sequence with V_H germline segment information if available. If J_H germline segment information was not available then this parameter was ignored, also retaining otherwise identical sequences with available J_H information as representatives for clonotypes, if available. Nonproductive sequences in the WA dataset were not processed for clonotype clustering.

Repertoire similarity index analysis. RSI was computed in a manner similar to a previously described method[27]. For a given set S of CDR H3 sequences, all of the same length n, RSI is measured as follows:

$$RSI = \left[1 - \frac{\text{median}\left\{ LD\left(S_i, S_j\right)_{i \neq j} \right\}}{n} \right] \times 100\% \, \forall \, 1 \leq i, j \leq |S|$$

where S_j and S_j refer to any two sequences in the set of CDR H3 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 segment and CDR H3 length, RSI values were computed for those sequences that shared the same V_L germline segment (for the paired CA and TX datasets) or the same J_H germline segment (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 Supplementary Fig. 5 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 segment was captured as a vector of length 22 containing the percentage of sequences of length 5–26. For V_H, the values for each germline segment were averaged overall the AE datasets except the WA dataset due to limited germline segment coverage. For V_L, the values were averaged over the CA and TX datasets. The distributions of each germline segment were consolidated into a matrix $X_{n \times 22}$ where n is the number of germline segments considered for analysis ($n = 39$ for V_H and $n = 35$ for V_L). 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 $\langle \rangle$ refers to average across all germline segments. 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 segment was then projected onto these eigenvectors to obtain the PC scores which enabled visualization of the different trends and comparisons among the different germline segments. Principal Component Analysis was performed using the 'princomp' function and Repertoire Similarity Index, implemented using the 'sdists' function in the package 'cba' in R version 3.5.0.

Statistics and reproducibility. Samples consist of independently sequenced antibody repertoires from 20 donors from five laboratories under different sequencing conditions, comprising five datasets of with 2–8 donors each with varying number of sequences per donor (see Supplementary Data 1, 2, and 3 for details and sample sizes for each dataset, donor, B cell subset and germline segment). To avoid over-representation of sequences from donors and datasets with higher sequence counts, calculations were performed per donor and then averaged per dataset or across datasets except for Supplementary Fig. 1, where sequences from donors of each dataset were pooled prior to calculations. Germline segments within a donor with fewer than 80 counts were excluded from analyses. Reproducibility was assessed by comparing distributions between datasets, with the exception for nonproductive sequences, which were present only in the WA dataset. Two-tailed paired t-tests of CDR H3 length distributions were performed using Microsoft Excel for the Mac 2016. Data from individual donors comparing germline segment-specific to overall repertoire CDR H3 length prevalence from the CA, TX, MA, and WA datasets ($n = 11$ donors for V_H, $n = 5$ donors for V_L) were used for t-tests of AE compartment samples over a sliding window of two consecutive CDR H3 lengths to minimize local fluctuations. P-value thresholds of 10^{-4} (V_H) and 10^{-2} (V_L) were used to assess significance across the range of CDR H3 lengths. Mann-Whitney tests for distributions in Supplementary Fig. 17 were done using GraphPad Prism version 6.

Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability
Sequence datasets used in analyses have been previously described[20–25]. The subset of the datasets with sequences clustered by clonotype used here are available at https://doi.org/10.5061/dryad.cjsxksn2x [43]. All data supporting this study are available within the article and its Supplementary Information Figures and Supplementary Tables or are available from the corresponding author on reasonable request.

Code availability
Raw NGS sequencing data analyzed with a custom analysis pipeline, available from the docker hub repository at hub.docker.com under kamhonhoi/iganalysis. Principal Component Analysis performed using the 'princomp' function and Repertoire Similarity Index, implemented using the 'sdists' function in the package 'cba' in R version 3.5.0.

References
1. Schroeder, H. W. Jr. & Cavacini, L. Structure and function of immunoglobulins. *J. Allergy Clin. Immunol.* **125**, S41–S52 (2010).
2. Cobb, R. M., Oestreich, K. J., Osipovich, O. A. & Oltz, E. M. Accessibility control of V(D)J recombination. *Adv. Immunol.* **91**, 45–109 (2006).
3. Bassing, C. H., Swat, W. & Alt, F. W. The mechanism and regulation of chromosomal V(D)J recombination. *Cell* **109**(Suppl), S45–S55 (2002).
4. Rajewsky, K. Clonal selection and learning in the antibody system. *Nature* **381**, 751–758 (1996).
5. Nemazee, D. Mechanisms of central tolerance for B cells. *Nat. Rev. Immunol.* **17**, 281–294 (2017).
6. Zemlin, M. 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. Schroeder, H. W. Jr. Similarity and divergence in the development and expression of the mouse and human antibody repertoires. *Dev. Comp. Immunol.* **30**, 119–135 (2006).
8. Bagnara, D. et al. A reassessment of IgM memory subsets in humans. *J. Immunol.* **195**, 3716–3724 (2015).

9. Wardemann, H. et al. Predominant autoantibody production by early human B cell precursors. *Science* **301**, 1374–1377 (2003).
10. Wu, Y. C. et al. High-throughput immunoglobulin repertoire analysis distinguishes between human IgM memory and switched memory B-cell populations. *Blood* **116**, 1070–1078 (2010).
11. Roskin, K. M. et al. IgH sequences in common variable immune deficiency reveal altered B cell development and selection. *Sci. Transl. Med.* **7**, 302ra135 (2015).
12. Doorenspleet, M. E. 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. Bashford-Rogers, R. J. M., Smith, K. G. C. & Thomas, D. C. Antibody repertoire analysis in polygenic autoimmune diseases. *Immunology* **155**, 3–17 (2018).
14. DeKosky, B. J. et al. Large-scale sequence and structural comparisons of human naive and antigen-experienced antibody repertoires. *Proc. Natl Acad. Sci. USA* **113**, E2636–E2645 (2016).
15. Ivanov, I. I. 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).
16. Larimore, K., McCormick, M. W., Robins, H. S. & Greenberg, P. D. Shaping of human germline IgH repertoires revealed by deep sequencing. *J. Immunol.* **189**, 3221–3230 (2012).
17. Elhanati, Y. et al. Inferring processes underlying B-cell repertoire diversity. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **370**, 20140243, https://doi.org/10.1098/rstb.2014.0243 (2015).
18. Miqueu, P. 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. Marcou, Q., Mora, T. & Walczak, A. M. High-throughput immune repertoire analysis with IGoR. *Nat. Commun.* **9**, 561 (2018).
20. Briney, B., Inderbitzin, A., Joyce, C. & Burton, D. R. Commonality despite exceptional diversity in the baseline human antibody repertoire. *Nature* **566**, 393–397 (2019).
21. Goldstein, L. D. et al. Massively parallel single-cell B-cell receptor sequencing enables rapid discovery of diverse antigen-reactive antibodies. *Commun. Biol.* **2**, 304 (2019).
22. Laserson, U. et al. High-resolution antibody dynamics of vaccine-induced immune responses. *Proc. Natl Acad. Sci. USA* **111**, 4928–4933 (2014).
23. DeKosky, B. J. et al. In-depth determination and analysis of the human paired heavy- and light-chain antibody repertoire. *Nat. Med.* **21**, 86–91 (2015).
24. Busse, C. Dynamics of the human antibody repertoire after influenza vaccination. *NCBI BioProject Database*, https://www.ncbi.nlm.nih.gov/bioproject/PRJNA349143 (2016).
25. DeWitt, W. S. et al. A public database of memory and naive B-cell receptor sequences. *PLoS ONE* **11**, e0160853 (2016).
26. Lefranc, M. P. IMGT, the international ImMunoGeneTics database: a high-quality information system for comparative immunogenetics and immunology. *Dev. Comp. Immunol.* **26**, 697–705 (2002).
27. Greiff, V. et al. Systems analysis reveals high genetic and antigen-driven predetermination of antibody repertoires throughout B cell development. *Cell Rep.* **19**, 1467–1478 (2017).
28. Matsutani, T. 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).
29. Miho, E., Roskar, R., Greiff, V. & Reddy, S. T. Large-scale network analysis reveals the sequence space architecture of antibody repertoires. *Nat. Commun.* **10**, 1321 (2019).
30. Boyd, S. D. et al. Individual variation in the germline Ig gene repertoire inferred from variable region gene rearrangements. *J. Immunol.* **184**, 6986–6992 (2010).
31. Kidd, M. J. 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).
32. Gidoni, M. et al. Mosaic deletion patterns of the human antibody heavy chain gene locus shown by Bayesian haplotyping. *Nat. Commun.* **10**, 628 (2019).
33. Shirai, H., Kidera, A. & Nakamura, H. H3-rules: identification of CDR-H3 structures in antibodies. *FEBS Lett.* **455**, 188–197 (1999).
34. North, B., Lehmann, A. & Dunbrack, R. L. Jr. A new clustering of antibody CDR loop conformations. *J. Mol. Biol.* **406**, 228–256 (2011).
35. Weitzner, B. D., Dunbrack, R. L. Jr. & Gray, J. J. The origin of CDR H3 structural diversity. *Structure* **23**, 302–311 (2015).
36. Vander Heiden, J. A. et al. Dysregulation of B cell repertoire formation in Myasthenia Gravis patients revealed through deep sequencing. *J. Immunol.* **198**, 1460–1473 (2017).
37. Jain, T. et al. Biophysical properties of the clinical-stage antibody landscape. *Proc. Natl Acad. Sci. USA* **114**, 944–949 (2017).
38. Collis, A. V., Brouwer, A. P. & Martin, A. C. 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).
39. Breden, F. et al. Comparison of antibody repertoires produced by HIV-1 infection, other chronic and acute infections, and systemic autoimmune disease. *PLoS ONE* **6**, e16857 (2011).
40. Magoc, T. & Salzberg, S. L. FLASH: fast length adjustment of short reads to improve genome assemblies. *Bioinformatics* **27**, 2957–2963 (2011).
41. Chen, Y. et al. Barcoded sequencing workflow for high throughput digitization of hybridoma antibody variable domain sequences. *J. Immunol. Methods* **455**, 88–94 (2018).
42. Ye, J., Ma, N., Madden, T. L. & Ostell, J. M. IgBLAST: an immunoglobulin variable domain sequence analysis tool. *Nucleic Acids Res.* **41**, W34–W40 (2013).
43. Sankar, K., Hoi, K. H., Jr. & Hotzel, I. Dynamics of heavy chain junctional length biases in antibody repertoires. *Dryad Database*, https://doi.org/10.5061/dryad.cjsxksn5062x (2020).

Competing interests
The authors declare no competing non-financial interests but the following competing financial interests: the authors are current or past Genentech employees and may hold shares in Roche.

Additional information
Supplementary information is available for this paper at https://doi.org/10.1038/s42003-020-0931-3.

Correspondence and requests for materials should be addressed to I.Höt.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.