Dynamics of heavy chain junctional length biases in antibody repertoires

Supplementary Information
Supplementary Figures

Supplementary Figure 1. (a-f) CDR H3 length distribution of raw, as published, data and unique lineages (clonotypes) in the AE and naive B cell compartments. (g-h) Comparative CDR H3 length distributions of unique lineages in the AE and naive B cell compartments.
Supplementary Figure 2

$V_H$: Short

Percent of sequences

|   | $V_H$3-7 | $V_H$3-72 | $V_H$3-73 | $V_H$3-74 | $V_H$3-53 | $V_H$3-66 |
|---|----------|----------|----------|----------|----------|----------|
| a | All AE   | All AE   |          |          |          |          |
| b | TX naive | TX naive |          |          |          |          |
| c | WA naive | WA naive |          |          |          |          |
| d | WA unpr. | WA unpr. |          |          |          |          |
| e | TX AE    | TX naive |          |          |          |          |
| f | WA AE    | WA naive |          |          |          |          |

CDR H3 length
Supplementary Figure 2

$V_H$: Short

a
All AE
All AE

b
TX naive
TX naive

c
WA naive
WA naive

d
WA unpr.
WA unpr.

e
TX AE
TX naive

f
WA AE
WA naive

Percent of sequences

CDR H3 length

$V_{H6\cdot1}$ $V_{H3\cdot15}$
Supplementary Figure 2

$V_H$: Neutral

a
All AE
All AE

b
TX naive
TX naive

c
WA naive
WA naive

d
WA unpr.
WA unpr.

e
TX AE
TX naive

f
WA AE
WA naive

Percent of sequences

CDR H3 length
Supplementary Figure 2

**V_{H}: Neutral**

**a**
All AE
All AE

**b**
TX naive
TX naive

**c**
WA naive
WA naive

**d**
WA unpr.
WA unpr.

**e**
TX AE
TX naive

**f**
WA AE
WA naive

Percent of sequences

CDR H3 length
Supplementary Figure 2

V_H: Neutral

a

All AE
All AE

b

TX naive
TX naive

c

WA naive
WA naive

d

WA unpr.
WA unpr.

e

TX AE
TX naive

f

WA AE
WA naive

Percent of sequences

CDR H3 length
Supplementary Figure 2

$V_H$: Long

- **a**
  - All AE
  - All AE

- **b**
  - TX naive
  - TX naive

- **c**
  - WA naive
  - WA naive

- **d**
  - WA unpr.
  - WA unpr.

- **e**
  - TX AE
  - TX naive

- **f**
  - WA AE
  - WA naive

Percent of sequences

CDR H3 length
Supplementary Figure 2

$V_H$: Long

a
All AE
All AE

b
TX naive
TX naive

c
WA naive
WA naive

d
WA unpr.
WA unpr.

e
TX AE
TX naive

f
WA AE
WA naive

Percent of sequences

CDR H3 length
**Supplementary Figure 2.** CDR H3 length distributions associated with $V_H$ germline segments (orange bars) compared to overall distribution in the same samples and compartments (rows a-d, blue bars) or to corresponding naive sequences (rows e and f):

Row (a): AE sequences, all CA, TX, MA and WA donors pooled ($n = 11$)
Row (b) TX naive compartment ($n = 3$)
Row (c) WA naive compartment ($n = 3$)
Row (d) WA $V_H$ unproductive sequences ($n = 3$)
Row (e) TX AE (orange bars, $n = 2$, same as orange bars in Suppl. Figure 3, TX) and naive sequences (blue bars, $n = 3$, same as orange bars in row b)
Row (f) WA AE (orange bars, same as orange bars in Suppl. Figure 3, WA) and naive sequences (blue bars, same as orange bars in row c) ($n = 3$)

Red and blue horizontal bars above the histograms in (a) indicate statistically significant ($P < 10^{-4}$) differences between the germline segment-specific and overall distributions in a paired (within donors, $n = 11$ donors) $t$-test. The “all” label in row (a) refers to datasets CA, TX, MA, and WA. Error bars indicate S.E.M. except orange bars in panel e, orange bars ($n = 2$ donors in TX AE), which indicate range. Arrows indicate subtle enrichment of sequences in the distributions. Figure tags refer to each row of graphs.
Supplementary Figure 3

$V_H$: Short

$V_H^{3-7}$  $V_H^{3-72}$  $V_H^{3-73}$  $V_H^{3-74}$  $V_H^{3-53}$  $V_H^{3-66}$

Percent of sequences

CDR H3 length
**Supplementary Figure 3**

**V<sub>H</sub>: Short**

| Region | V<sub>H</sub>6-1 | V<sub>H</sub>3-15 |
|--------|------------------|------------------|
| TX     | ![TX V<sub>H</sub>6-1](image) | ![TX V<sub>H</sub>3-15](image) |
| CA     | ![CA V<sub>H</sub>6-1](image) | ![CA V<sub>H</sub>3-15](image) |
| MA     | ![MA V<sub>H</sub>6-1](image) | ![MA V<sub>H</sub>3-15](image) |
| WA     | ![WA V<sub>H</sub>6-1](image) | ![WA V<sub>H</sub>3-15](image) |
| SRI    | ![SRI V<sub>H</sub>6-1](image) | ![SRI V<sub>H</sub>3-15](image) |

**CDR H3 length**
Supplementary Figure 3

$V_H$: Neutral

|   | $V_{H1}$-2 | $V_{H1}$-3 | $V_{H1}$-8 | $V_{H1}$-46 | $V_{H3}$-11 | $V_{H3}$-21 |
|---|------------|------------|------------|-------------|-------------|-------------|
| TX | ![Histogram](image1) | ![Histogram](image2) | ![Histogram](image3) | ![Histogram](image4) | ![Histogram](image5) | ![Histogram](image6) |
| CA | ![Histogram](image7) | ![Histogram](image8) | ![Histogram](image9) | ![Histogram](image10) | ![Histogram](image11) | ![Histogram](image12) |
| MA | ![Histogram](image13) | ![Histogram](image14) | ![Histogram](image15) | ![Histogram](image16) | ![Histogram](image17) | ![Histogram](image18) |
| WA | ![Histogram](image19) | ![Histogram](image20) | ![Histogram](image21) | ![Histogram](image22) | ![Histogram](image23) | ![Histogram](image24) |
| SRI| ![Histogram](image25) | ![Histogram](image26) | ![Histogram](image27) | ![Histogram](image28) | ![Histogram](image29) | ![Histogram](image30) |

CDR H3 length
Supplementary Figure 3

$V_H$: Neutral

| State | CDR H3 length |
|-------|---------------|
| TX    |               |
| CA    |               |
| MA    |               |
| WA    |               |
| SRI   |               |
Supplementary Figure 3

$V_H$: Neutral

|       | $V_H4-31$ | $V_H4-39$ | $V_H4-59$ | $V_H7-4-1$ | $V_H2-5$ |
|-------|-----------|-----------|-----------|------------|-----------|
| TX    |           |           |           |            |           |
| CA    |           |           |           |            |           |
| MA    |           |           |           |            |           |
| WA    |           |           |           |            |           |
| SRI   |           |           |           |            |           |

Percent of sequences

CDR H3 length
Supplementary Figure 3

V_H: Long

| TX  | CA  | MA  | WA  | SRI |
|-----|-----|-----|-----|-----|
| V_H2-26 | V_H4-34 | V_H1-18 | V_H1-69 | V_H1-24 | V_H2-70 |

Percent of sequences vs. CDR H3 length
Supplementary Figure 3

$V_H$: Long

|   | $V_H^{3-9}$ | $V_H^{3-20}$ | $V_H^{3-23}$ | $V_H^{3-30}$ | $V_H^{3-30-3}$ | $V_H^{3-33}$ |
|---|-------------|-------------|-------------|-------------|-------------|-------------|
| TX | ![Histogram](image) | ![Histogram](image) | ![Histogram](image) | ![Histogram](image) | ![Histogram](image) | ![Histogram](image) |
| CA | ![Histogram](image) | ![Histogram](image) | ![Histogram](image) | ![Histogram](image) | ![Histogram](image) | ![Histogram](image) |
| MA | ![Histogram](image) | ![Histogram](image) | ![Histogram](image) | ![Histogram](image) | ![Histogram](image) | ![Histogram](image) |
| WA | ![Histogram](image) | ![Histogram](image) | ![Histogram](image) | ![Histogram](image) | ![Histogram](image) | ![Histogram](image) |
| SRI | ![Histogram](image) | ![Histogram](image) | ![Histogram](image) | ![Histogram](image) | ![Histogram](image) | ![Histogram](image) |

Percent of sequences vs. CDR H3 length
Supplementary Figure 3

$V_H$: Long

Supplementary Figure 3. Comparability of $V_H$-associated CDR H3 length biases among AE datasets. $V_H$ germline segment-specific CDR H3 length distributions (orange bars) compared to overall distribution (blue bars) in the AE B cell compartments of the five datasets. Error bars indicate S.E.M. except for TX ($n = 2$ for TX AE donors), which indicates range. Arrows indicate subtle enrichment of sequences in the distributions shared among datasets.
Supplementary Figure 4. Prevalence of V<sub>H</sub> (a) and V<sub>L</sub> (b) germline segments in the AE compartment. Prevalence of germline segments was calculated based on unique clonotypes after pooling of donors within each dataset.
Supplementary Figure 5

Supplementary Figure 5. RSI analysis of V<sub>H</sub> (a) and V<sub>L</sub> (b) germline segment-associated sequences. Points show average and standard deviation of the RSI value for sequences from each donor, germline segment and CDR H3 length. WA and SRI sequences were excluded due to large dataset sizes. Note that RSI values cannot be higher than 60 in the TX, CA and MA datasets due to clonotype definition. Germline segments are noted in the IMGT naming convention.
Supplementary Figure 6

**V_L**: Short

- **a**
  - TX CA AE
  - TX CA AE

- **b**
  - TX naive
  - TX naive

- **c**
  - TX AE
  - TX naive

**Percent of sequences**

**CDR H3 length**
Supplementary Figure 6

**V_L**: Long

**V_k**: 2-28  1-9  1-39  1-12  3-20

**Percent of sequences**

**CDR H3 length**
Supplementary Figure 6

**CDR H3 length**

**Supplementary Figure 6.** CDR H3 length distributions associated with V₅ germline segments (orange bars) compared to overall distribution in the same samples and compartments (rows a and b, blue bars) or to corresponding naive sequences (row c):

Row (a) AE sequences, CA and TX donors pooled (n = 5)
Row (b) TX naive compartment (n = 3)
Row (c) TX AE (orange bars, n = 2, same as orange bars in Suppl. Figure 6 TX) and naive (blue bars, same as orange bars in row b) sequences of the indicated germline segment (n = 3)

Red and blue horizontal bars above the histograms in row (a) indicate statistically significant \( P < 10^{-2} \) differences between the germline segment-specific and overall distributions in a paired (within donors, n = 5 donors) t-test with a rolling window of 2 consecutive CDR H3 lengths (see main text for details). Error bars indicate S.E.M. except for orange bars in row (c), where it indicates range (n = 2).

Arrows indicate subtle deviations in the distributions. Figure tags refer to each row of graphs.
Supplementary Figure 7

$V_L$: Long

| $V_K$ | TX | CA |
|-------|----|----|
| 2-28  |     |     |
| 1-9   |     |     |
| 1-39  |     |     |
| 1-12  |     |     |
| 3-20  |     |     |

| $V_\lambda$ | TX | CA |
|-------------|----|----|
| 3-19        |     |     |
| 3-25        |     |     |
| 3-1         |     |     |
| 1-40        |     |     |
| 1-51        |     |     |

CDR H3 length
Supplementary Figure 7. \( V_L \): Neutral

\[ V_L^{4-69} \]

\( V_L \): Long

\[ V_L^{1-44} \]
\[ V_L^{3-10} \]
\[ V_L^{3-21} \]

\( V_L \) germline segment-specific CDR H3 length distributions (orange bars) compared to overall distribution (blue bars) in the four datasets. Error bars indicate S.E.M. (CA, \( n = 3 \)) and range (TX, \( n = 2 \)). Germline segments are named in the IMGT convention. Asterisks indicate spikes at length 18 in the IGKV1-6 distributions of TX and CA datasets with high prevalence of convergent clones with IGHV3-7, IGHJ3, CDR H3 consensus sequence (A/V)RDX7(L/I/V)(W/Y)YDAFDI and light chain Arg-116 in CDR L3, observed in all CA and TX donors.
Supplementary Figure 8. J_H prevalence as a function of V_H (a-c) and V_L (d) germline segment in WA unproductive sequences (a) and WA (b) and TX (c and d) naive compartment sequences. Dotted lines indicate overall average J_H prevalence. Germline segments are noted in the IMGT naming convention. Values shown for each dataset after pooling of donors.
WA: Nonproductive

Supplementary Figure 9. J<sub>H</sub> germline segment distribution as a function of V<sub>H</sub> germline segment and CDR H3 length in unproductive sequences and in the naive compartments of the WA dataset. Data is averaged for all donors. Germline segments are noted in the IMGT naming convention. A fraction of the sequences does not have an unambiguously annotated J<sub>H</sub> segment, particularly in the short end of the spectrum. Values calculated after pooling of sequences from 3 donors.
Supplementary Figure 10

JH4
4 aa

JH3
5 aa

JH5
5 aa

JH1
6 aa

JH2
6 aa

JH6
9 aa

VH3-72

VH3-73

VH3-74

VH6-1

VH3-15

VH1-2

Percent of sequences

CDR H3 length
Supplementary Figure 10. CDR H3 distribution as a function of V_{H} and J_{H} germline segment in the WA naive compartment. V_{H}/J_{H} combination-specific distributions are shown in orange and JH-specific distributions in blue. V_{H} germline segments are noted in the IMGT naming convention. The number of CDR H3 amino acid residues potentially encoded by J_{H}1-J_{H}6 are shown in row legends. Error bars indicate S.E.M. (n = 3 donors except for IGHV3-9 and IGHV7-4-1, with n = 1 donor each).
Supplementary Figure 11. Modulation of V_{H}-associated CDR H3 length biases by J_{H} segments in the SRI naive compartment. Colors, arrows and panel organization is the same as in Figure 5, adding J_{H}1 and J_{H}2, for ease of comparison. Omitted panels had low sequence counts. The number of CDR H3 amino acid (aa) residues potentially encoded by J_{H}1-J_{H}6 are shown in row legends. Error bars indicate S.E.M. (n = 8 donors except V_{H}3-9, n = 7 donors).
Supplementary Figure 12

IGHJ4: YFDY (IMGT 114-117) WA, nonproductive sequences

IGHJ4: YFDY (IMGT 114-117) WA, naive sequences
IGHJ5: NWFD (IMG 113-117)  WA, Nonproductive sequences

Supplementary Figure 12

IGHJ5: NWFD (IMG 113-117)  WA, Naive sequences
Supplementary Figure 12

**Supplementary Figure 12.** Analysis of J₄ and J₅ sequence trimming as a function of V₇ germline segment use and CDR H3 length in WA unproductive and naive compartment sequences. Occupancy of J₄ germline segment-encoded residues in the last CDR H3 positions are shown. “YFDY” and “NWFDP” refer to the last residues of the J₄ and J₅ germline segments within CDR H3, color coded as in the lines in the figure. Solid lines indicate residue occupancy for all sequences in the unproductive and naive repertoire with a given J₄ segment. Dots indicate average residue occupancy with each V₇/J₄ combination CDR H3 length. Dots above and below the corresponding solid line indicate reduced and increased J₄ trimming relative to the overall repertoire. Note that trimming of J₄ segments in the non-productive sequences is not biased by V₇. Bars indicate S.E.M. for three donors except for V₇3-9 and V₇7-4-1, which are present in one donor each, and the nonproductive sequences, which were pooled before analysis. Data points with fewer than 60 sequences (30 for V₇3-9 and V₇7-4-1) were excluded. Panels are organized in the same order as in Supplementary Figure 9.
IGHJ4: YFDY (IMGT 114-117)  SRI, Naive compartment

Supplementary Figure 13. Analysis of J_{H4} sequence trimming as a function of V_{H} germline segment use and CDR H3 length in SRI naive compartment sequences. Symbols shown as in Figure 6. The panels highlighted with black diamonds indicate germline segments in the same place as in Suppl. Fig. 12. Other panels show different germline segments according to data availability in the dataset. Data points with fewer than 60 sequences were excluded. “YFDY” refers to the last residues of the J_{H4} germline segment within CDR H3, color coded as in the lines in the figure. Error bars indicate S.E.M. (n = 7 except V_{H}2-70, n=5, and V_{H}1-8 and V_{H}6-1, n = 6 donors).
### Supplementary Figure 14

D germline segment prevalence associated with different V<sub>H</sub> and J<sub>H</sub> germline segments in the WA naive compartment. D germline segments are grouped by length. The D germline segment prevalence deviations from averages associated with V<sub>H</sub>6-1 are boxed.
Supplementary Figure 15. Summary of most significant length biases of regions within CDR H3 as a function of V_{H} and J_{H} germline use and CDR H3 length. (a, rows 1 to 4) Deviations in nucleotides (Δnt) from whole repertoire averages are shown for the V_{H}, D_{H}, J_{H} segments and NP-region (insertions) in CDR H3 in WA naive sequences grouped by V_{H} and J_{H} germlines. Colors for each segment are indicated in the top left panel. Values in parentheses indicate the maximum number of V_{H} germline nucleotides that can be included in CDR H3. (b, row 5) As in panels (a), showing CDR H3 segment length deviations associated with V_{H}3-15 sequences and different J_{H} germline segments in 3 donors of the SRI naive 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 Supplementary Fig. 16. See Supplementary text for details.
Supplementary Figure 16

WA naive: J₄₅

SRI IgM/naive: J₄₅

CDR H3 Length

V₃-15 (8)
V₃-9 (10)
V₃-23 (8)
V₃-73 (8)
V₃-12 (8)
V₃-3-9 (8)
V₃-74 (8)
V₃-2-5 (10)
V₃-2-70 (10)
V₃-7-4 (1) (8)
V₄-34 (8)
V₄-15 (8)
V₄-3-9 (8)
V₄-3-7 (8)
V₄-23 (8)
V₄-74 (8)
V₄-1-2 (8)
V₄-4-39 (8)
V₄-1-18 (8)
V₄-1-8 (8)
V₄-2-6 (10)
V₄-1-69 (8)
V₅-6-1 (8)
V₅-3-49 (8)
V₅-5-51 (8)

CDR H3 Length

Δnt

Δnt

Δnt

Δnt

Δnt

Δnt
Supplementary Figure 16. Deviations from average length for V$_H$, D$_H$, J$_H$ and NP-regions (insertions) in clones segregated by V$_H$ and J$_H$ germline segments. Deviations are calculated by subtracting average values for each region from overall repertoire averages within donors. Data points show averages of at least 3 donors from each data set for V$_H$/J$_H$ combinations with at least 60 counts. SRI panels highlighted by blue diamonds indicate V$_H$/J$_H$ combinations also analyzed in the WA naive dataset. Error bars omitted for clarity. Note similarity of trends between WA and SRI naive datasets and lack of trends in WA unproductive sequences except for longer V$_H$ sequences for V$_H$2 and V$_H$3-9, as expected, and a trend for shorter than expected lengths of nucleotide insertions in V$_H$2 sequences. See Supplementary notes for details.
Supplementary Figure 17. CDR H3 length distributions associated with V_{H}1-2 (a, row 1) and V_{H}2-5 (b, rows 2 and 3) allelic variants in the SRI and MA datasets. All panels except the lower right are SRI dataset donors. Only heterozygous donors for these germline segments are shown, indicated in each panel. The 3 MA donors are heterozygous for V_{H}2-5 and were pooled in the lower right panel due to relatively low counts for each allele in individual donors. For SRI donors, only IgM/naive sequences are included. For the MA donors, only IgM sequences with up to 1 amino acid mutation in IMGT® positions 1 to 104 were included. Light blue and orange symbols in donor D326713 distributions indicate haplotypes based on J_{H}6 alleles associated with each V_{H}1-2 and V_{H}2-5 allele for donor D326713. Dark blue and magenta symbols indicate alleles for which haplotype analysis based on J_{H}6 alleles was not possible. Note the opposite biases of V_{H}1-2 and V_{H}2-5 alleles in the same chromosome for donor D326713, indicating that haplotype-associated variations of D germline segments do not directly determine allele-associated CDR H3 length biases. The difference in average CDR H3 length between alleles is shown in each panel. All CDR H3 distribution differences are statistically significant in a Mann-Whitney test (P < 10^{-15}, except V_{H}2-5 alleles of MA dataset, P = 10^{-5}). V_{H}2-5 alleles vary by Asn/Asp-59 and V_{H}1-2 alleles by Arg/Trp-75. Additional variations in the framework region 1 may occur but are not covered in the SRI dataset. Other allelic variants were not analyzed due to lack of sufficient number of heterozygous donors with same alleles or insufficient allele-specific sequence counts.
**Supplementary Figure 18.** CDR H3 residues encoded by V<sub>H</sub> germline segments in the absence of nucleotide trimming. Residues in parentheses show partial codons and favored encoded residue in the absence of nucleotide trimming.
### Supplementary Table

#### Supplementary Table 1. Datasets

|                      | CA       | MA       | TX       | WA       | SRI      |
|----------------------|----------|----------|----------|----------|----------|
| **Number of donors** | 3        | 3        | 2        | 2        | 3        |
| **Cell type**        | IgG<sup>pos</sup> B cells | PBMC | CD3<sup>neg</sup> CD19<sup>pos</sup> CD20<sup>pos</sup> CD2<sup>pos</sup> and CD3<sup>neg</sup> CD19<sup>pos</sup> CD27<sup>pos</sup> and CD3<sup>neg</sup> CD19pos CD27<sup>neg</sup> B cells | CD3<sup>neg</sup> CD19<sup>pos</sup> CD27<sup>pos</sup> and CD3<sup>neg</sup> CD19pos CD27<sup>neg</sup> B cells | PBMC |
| **CD27 marker**      | NA<sup>b</sup> | NA       | CD27<sup>neg</sup> | CD27<sup>pos</sup> | CD27<sup>pos</sup> | NA       |
| **Source**           | cDNA     | cDNA     | cDNA     | genomic DNA | cDNA     |
| **Sequencing platform** | Illumina HiSeq2500 | Illumina MiSeq | Illumina MiSeq | Illumina HiSeq | Illumina HiSeq2500 |
| **Isotype**          | IgG      | IgG, IgA<sup>c</sup> | IgM, Null<sup>d</sup> | IgG, IgA | IgM | NA |
| **B cell compartment in main text** | CA AE | MA AE | TX naive | TX AE | TX AE IgM | WA naive | WA AE |
| **Chain sequences**  | V<sub>H</sub> + V<sub>L</sub> | V<sub>H</sub> | V<sub>H</sub> + V<sub>L</sub> | V<sub>H</sub> | V<sub>H</sub> | V<sub>H</sub> + J<sub>H</sub> |
| **V/J region parsing** | Absolve, as published | IgBlast | IgBlast, As published | IMGT collaboration, as published | abstar, as published |
| **Clonotype clustering** | V<sub>H</sub> + V<sub>L</sub> | V<sub>H</sub> + J<sub>H</sub> | V<sub>H</sub> + V<sub>L</sub> | V<sub>H</sub> + J<sub>H</sub> | V<sub>H</sub> + J<sub>H</sub> |
| **Clonotype clustering CDR H3 aa identity** | 60% | 60% | 60% | 100% | 100% |
| **Raw counts**       | 68,727   | 233,724  | 55,210   | 81,786   | 26,447   | 2.2 x 10<sup>7</sup> | 1.9 x 10<sup>7</sup> | 1.9 x 10<sup>6</sup> |
| **Unique clonotypes** | 63,503   | 63,051   | 52,993   | 67,158   | 22,616   | 1.7 x 10<sup>7</sup> | 8.3 x 10<sup>6</sup> | AE: 6.3 x 10<sup>6</sup> |
| **Unproductive sequences** | NA       | NA       | NA       | 3.1 x 10<sup>6</sup> | Not used | NA |

<sup>a</sup> Dataset with 3 donors, only 2 of which were processed for CD27<sup>pos</sup> B cells.

<sup>b</sup> Not available.

<sup>c</sup> IgM sequences were not included in analyses for the MA dataset.

<sup>d</sup> “Null” isotype sequences in the CD27<sup>neg</sup> compartment were assumed to be IgM.

<sup>e</sup> Sequences with undefined V<sub>H</sub> or J<sub>H</sub> germlines grouped by clonotype as described in methods.

<sup>f</sup> Average 60% amino acid identity across CDR H3 lengths achieved by using a nominal 57% identity threshold.

<sup>g</sup> Number of sequences after filtering for clonotypes as described in methods.

<sup>h</sup> IgM/naive sequences in the SRI dataset defined as IgM sequences without amino acid mutations between IMGT® Cys-23 and Cys-104.

---

52
Supplementary Notes

Biases in $D_H$ segment and N-region lengths within CDR H3 sequences as a function of $V_H$ and $J_H$ germline segment use. Parsing of $D_H$ segments in the WA naive and SRI IgM sequences with no mutations in the $V_H$ region between Cys-23 and Cys-104, as a proxy for naive sequences, was done using Blast after removing the sequences corresponding to $V_H$ and $J_H$ regions from CDR H3 sequences. The samples from the D1Nb subset were also included for $D_H$ segment parsing, removing redundant sequences as described in Methods. An identity of 100% over a span of at least 5 contiguous nucleotides was required for $D_H$ germline segment matches. The identity of the $DH$ segments for each hit were not considered, only the length of the germline segment in the germline. This was done due to parsing uncertainties, especially for the shorter matches. However, each match length represents the longest possible length within the set of $D_H$ segments. In addition, all datasets were parsed the same way, controlling for biases in the parsing procedures.

We determined whether the observed $J_H$ segment length biases in CDR H3 sequences of specific lengths are an indirect consequence of biases elsewhere in CDR H3, such as length of $V_H$ and $D_H$ sequences and number of N-region and palindromic nucleotide insertions (NP-region) flanking the $D_H$ region in CDR H3. Naive sequences from the 3 donors of the WA dataset and SRI IgM/naive sequences from 3 of the donors with higher number of sequences in the SRI dataset were parsed for $V_H$, $J_H$, $D_H$ and NP-region lengths within CDR H3. In general, no obvious differences on the prevalence of $D_H$ germline segments of different lengths within the germline were observed in association with different $V_H$ and the $J_H4$ and $J_H5$ germline segments that would account for $J_H$ length biases (Supplementary Fig. 14). One possible exception is $V_H6-1$ which was associated with shorter $D_H$ germline segments. In addition, the number of nucleotides that $V_H$ can directly contribute to CDR H3 did not correlate with $J_H$ length biases (Supplementary Fig. 15 and 16, red and blue lines and symbols). However, different classes of biases in the lengths of $D_H$
segments and NP-regions were observed for different \( V_\text{H}/J_\text{H} \) combinations, even for clones with the same \( V_\text{H} \) germline segment (Supplementary Fig. 15 and 16, green lines and symbols). 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 \( V_\text{H} \)-derived sequence lengths associated with \( V_\text{H} \) germline segment length and a generally shorter NP-region length in \( V_\text{H}2 \) clones (Supplementary Fig. 16), indicating that the observed \( D_\text{H} \) and NP-region length biases are mostly selected in naive repertoire maturation.

Different classes of junctional biases were observed for different \( V_\text{H}/J_\text{H} \) combinations (Supplementary Fig. 15 and 16). In several \( V_\text{H}/J_\text{H} \) combinations, \( J_\text{H} \) length biases were inversely correlated with both \( D_\text{H} \) segment and NP-region length biases including, for instance, \( V_\text{H}3-23/J_\text{H}5 \), \( V_\text{H}3-15/J_\text{H}4 \) and \( V_\text{H}3-73/J_\text{H}4/J_\text{H}5 \) combinations. In these cases, the factor determining junctional biases is likely to be the \( J_\text{H} \) segment, with the other junctional components compensating for the \( J_\text{H} \) biases. In \( V_\text{H}3-74/J_\text{H}4 \), \( D_\text{H} \) length biases are compensated by both NP-region and \( J_\text{H} \) lengths, suggesting the \( D_\text{H} \) segments as the bias determinant. In other cases, such as \( V_\text{H}3-15/J_\text{H}4 \), \( V_\text{H}3-7/J_\text{H}4 \) and \( V_\text{H}3-9/J_\text{H}4 \), \( D_\text{H} \) and NP-region lengths were inversely correlated with each other without correlations with \( J_\text{H} \) segment length across the CDR H3 length spectrum. In \( V_\text{H}2 \) sequences, the reduction in NP-region length observed in nonproductive sequences is maintained in the naive repertoire and compensates for the longer \( V_\text{H} \) sequences. However, as the \( J_\text{H} \) segments of nonproductive \( V_\text{H}2 \) sequences do not appear to be biased relative to the overall repertoire, the observed \( J_\text{H}4 \) and \( J_\text{H}5 \) trimming biases associated with \( V_\text{H}2 \) naive sequences are presumably due to \( J_\text{H} \) trimming rather than NP-region selection. Overall, the results indicate different classes of biases in \( D_\text{H} \) segment, N-region and \( J_\text{H} \) lengths within CDR H3 of naive sequences that vary among \( V_\text{H}/J_\text{H} \) germline segment combinations.
Supplementary References

1. 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).

2. Busse, C. Dynamics of the human antibody repertoire after influenza vaccination. NCBI BioProject Database, https://www.ncbi.nlm.nih.gov/bioproject/PRJNA349143 (2016).

3. Laserson, U. et al. High-resolution antibody dynamics of vaccine-induced immune responses. *Proc Natl Acad Sci U S A* 111, 4928-4933 (2014).

4. 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).

5. DeWitt, W. S. et al. A Public Database of Memory and Naive B-Cell Receptor Sequences. PLoS One 11, e0160853 (2016).

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