Supporting Appendix

Supplementary Table 1. Primers used in this study.

Name	**Sequence**	**V gene segment**	**V gene segment amplicon (bp)**
ZVH 4-1	TGGTCTCCTCTGCCTTTTGT	5.3	362
ZVH 4-2	AACCATGATCGCCTCATCTC	5.4, 5.8	362,359
ZVH 4-3	GATGGCAACAACATCCTGTG	7.1	314
ZVH 4-4	TGCATTTCAGTTCTGCTGCT	8.2, 8.3, 8.4	346,343,343
ZVH 4-5	ACGAATGCAGGAGTCAGACA	14.1	307
ZVH 4-6	TGTTTCAACTGTTCGTGGTCA	1.1, 1.2	310, 311
ZVH 4-7	TGGAGTTGTGTTGATGATGATT	1.3	326
ZVH 4-8	TTCATATGCACATGGTCAGTCA	1.4	302
ZVH 4-9	TGTGGTGATTGTCTTTCAAGG	2.1, 2.2	349,332
ZVH 4-10	TGGAAAAGGAGTCAAAAAGCAT	2.3	386
ZVH 4-11	GCTTTTGTCATGTTTGCTCTCA	3.2	331
ZVH 4-12	GCTTACTGCTGCTCTCATTCAG	4.3, 4.8, 4.9	339,339,336
ZVH 4-13	TTTCTGCTGCTGTGCTTTAC	4.5, 4.7	343,334
ZVH 4-14	CTGCTGTTTTCATTGGCCTTA	4.1	337
ZVH 4-15	GGTTTATACTGTCAAGGCATGG	4.2	307
ZVH 4-16	CAGCCTCAAGATGAAGAATGC	4.6	350
ZVH 4-17	CTAGTGCTGTTTCTGGCAGT	5.1, 5.7	328,328
ZVH 4-18	CATGATCACCTCATCTCTCTGC	5.2, 5.5	356,359
ZVH 4-19	CATGATTCTGAGCATTTTATCATGT	6.1	329
ZVH 4-20	CAATAATCAACTCACTCCTGCTG	6.2	345
ZVH 4-21	CTGCGTCCAGTGTATATTCCA	8.1	315
ZVH 4-22	TGTATTGACTGTCAGGTTGTGC	9.2, 9.4	304,304
ZVH 4-23	TCTTTCTGCAGTTGGCAG	9.1, 9.3	330,334
ZVH 4-24	TCTCAAAGTTGTTGGTGTCAGA	10.1	313
ZVH 4-25	CTCTCTAAACAAGTGCAAAGGTC	11.1	321
ZVH 4-26	TGGACCTTAAACTTAACTGTCTG	11.2	360
ZVH 4-27	CCATATGTTTCTGGCATCTCCC	13.2	308
C5	TGCACTGAGACAAACCGAAG	**C-μ**	N/A
C6	TCAGAGGCCAGACATCCAAT	**C-ζ**	N/A
VhuCc5	TGATTGACCCATCAGAACCA	**C-μ**	N/A
VhzCc6	GAATGCTGGGTGACGTTTTT	**C-ζ**	N/A

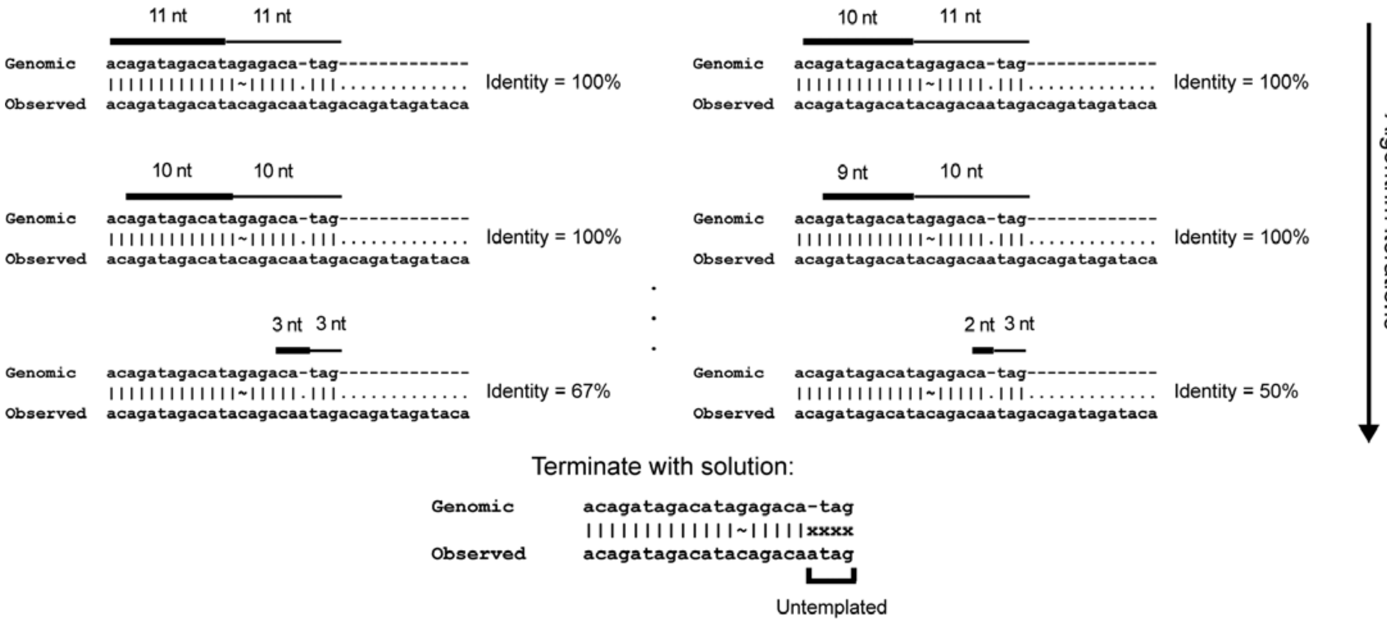
Computational methods and parameters

Pre-processing

Reads entered into the analysis after being identically matched to the corresponding sample's MID and reverse primers (IgM or IgZ). These reads were then filtered for a minimum length of 250 bp. Larger reads were truncated to this length, consistent with the filter applied in (1).

VJ classification and junction determination

All sequences were aligned by the Smith-Waterman algorithm to each genomic V- and J-gene segment with match gain of +3, mismatch cost of -3, gap-open cost of -8, and gap-extend cost of -8. Putative junction boundaries were determined for each alignment by iteratively reducing the size of overhang windows at the end of the alignment, as illustrated in Supplementary Figure 1. Iteration was performed until the largest overhang window containing alignment identity less than a threshold of 51% could be found. At this point, the furthest mismatch or indel within the overhang was counted as the beginning of the junction. Junction boundaries were further extended if fewer than 2 consecutive matches were found at the 5’ end of the J-alignment or fewer than 3 consecutive matches were found at the 3’ end of the V-alignment.



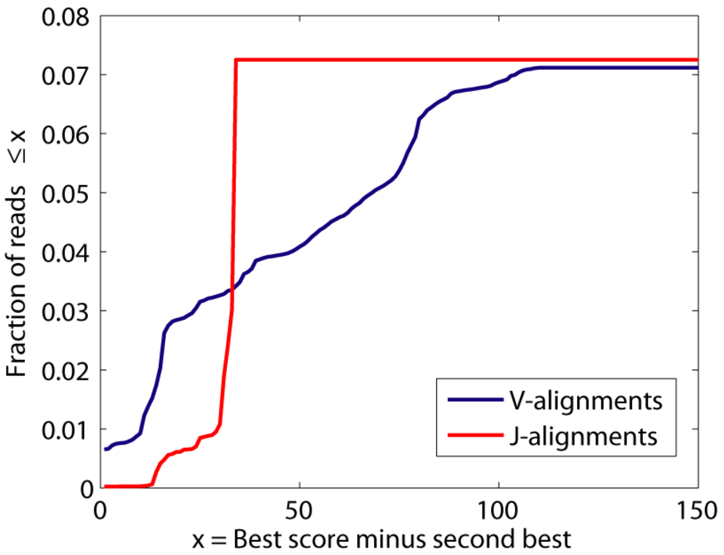
Supplementary Figure 1: Simple illustration of trimming algorithm for delineating junction boundaries at the 3’ end of a sequence.

A sequence was retained (Supplementary Table 2) if it matched at least 75 nucleotides from a V gene-segment and 20 nucleotides from a J gene-segment *after* the junction was excluded from alignment. For those sequences passing this threshold, the same junction-excluded portion of the alignment was scored against all V- and J-matches, with an assignment made to the V and J segments maximizing

$$\text{Score} = \text{Matches} - (\text{Indels} + \text{Mismatches})$$

The degree of ambiguity in assignment based on this score is graphed in Supplementary Figure 2.

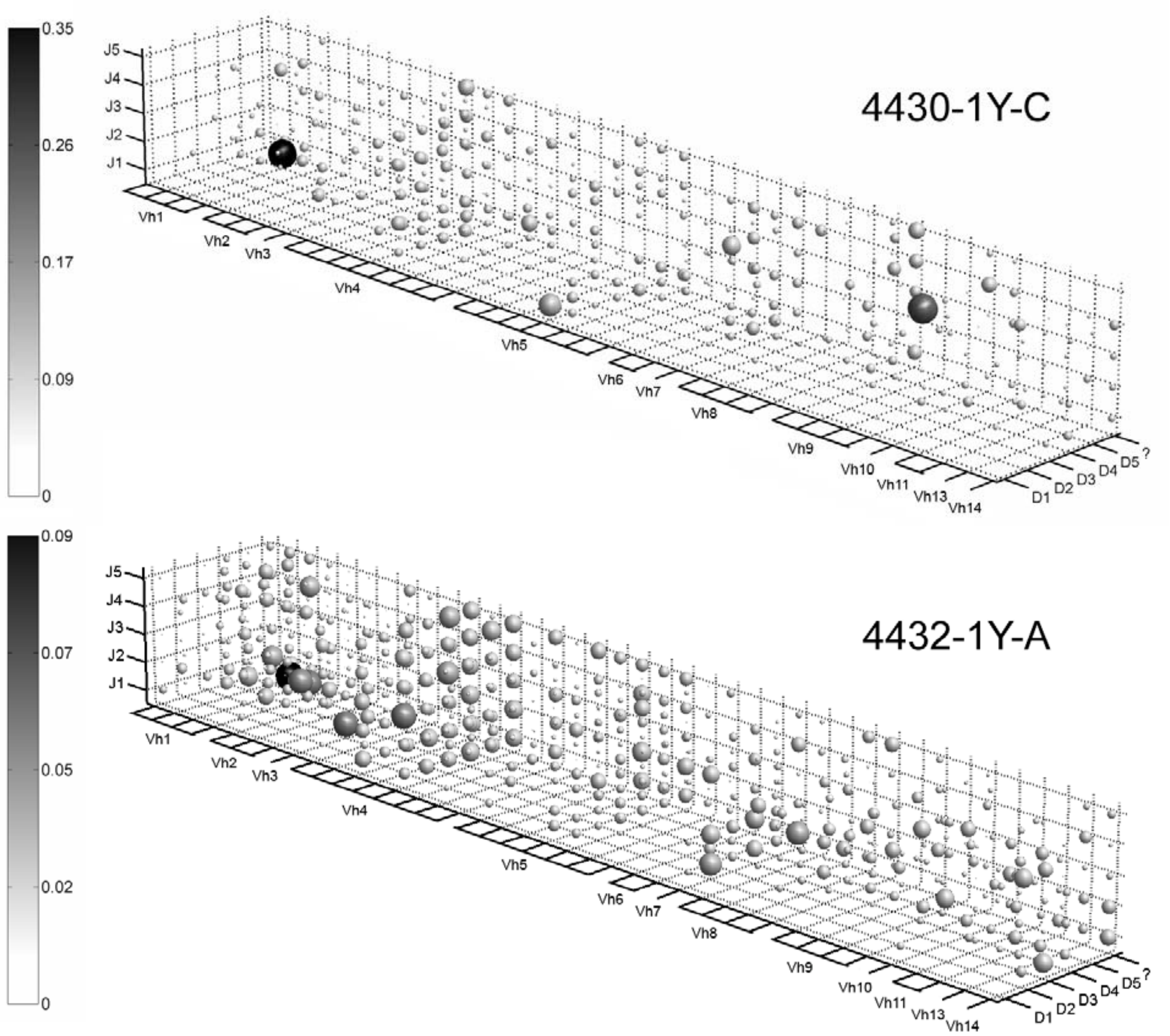
A sequence was labeled “ambiguous” (supplementary table 2) if the score could not distinguish between alignments. In these cases, on average 0.5% of all data, the lower-numbered V or J gene segment was assigned.



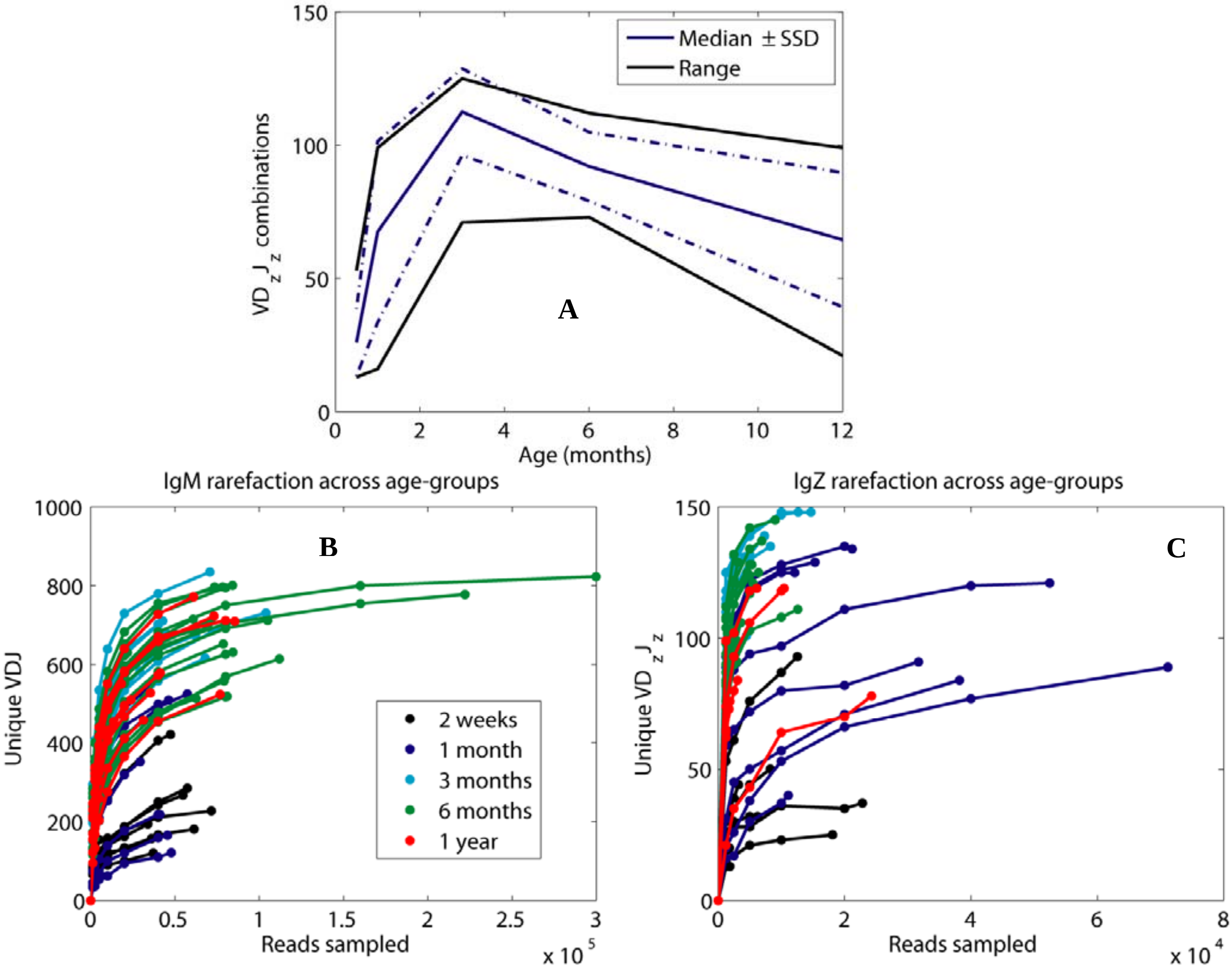
Supplementary Figure 2: Quantification of ambiguity between retained V/J combinations and their next-best alignments. Scores calculated as Matches – (Indels + Mismatches), excluding junction regions delineated as illustrated in Supplementary Figure 1.

Fish	IgM reads	IgM retained	Ambig. IgM	IgZ reads	IgZ retained	Ambig. IgZ	Fraction IgZ
4432-2W-A	47685	47389	250	12738	12526	85	0.209
4432-2W-B	72001	71621	710	23048	22824	462	0.242
4432-2W-C	54922	54800	112	8303	8235	10	0.131
4432-2W-D	33992	33981	110	2086	1791	0	0.05
4433-2W-A	37317	37178	298	2571	1835	0	0.047
4433-2W-B	57446	57347	475	3241	3214	19	0.053
4433-2W-C	61513	61403	458	6347	6226	0	0.092
4433-2W-D	36264	36161	28	18233	18100	11	0.334
4432-1M-A	29535	29415	208	53990	52460	238	0.641
4432-1M-B	57536	57399	160	15559	15334	129	0.211
4432-1M-C	46406	46321	277	12306	12114	129	0.207
4432-1M-D	39821	39707	136	21402	21206	77	0.348
4433-1M-A	41338	41128	63	33560	31704	54	0.435
4433-1M-B	5646	5093	8	81755	71145	62	0.933
4433-1M-C	46592	45606	47	38449	38181	34	0.456
4433-1M-D	48699	47849	120	16554	11104	12	0.188
4432-3M-A	42899	42797	116	2736	2679	21	0.059
4432-3M-B	62566	62228	684	8428	8270	135	0.117
4432-3M-C	29484	29291	286	4344	4196	66	0.125
4432-3M-D	40156	39976	29	7505	7324	33	0.155
4433-3M-A	104320	104006	270	14956	14673	109	0.124
4433-3M-B	44487	44462	229	2317	2304	6	0.049
4433-3M-C	71243	70747	307	12989	12658	71	0.152
4433-3M-D	68103	67909	54	4601	4490	63	0.062
TS-6M-A	649946	649475	186	13288	12603	24	0.019
TS-6M-B	21835	21822	140	194	194	4	0.009
TS-6M-C	105142	105069	81	1405	1400	9	0.013
TS-6M-D	222172	222054	5905	3106	2979	334	0.013
TS-6M-E	62155	62142	198	1710	1705	11	0.027
TS-6M-F	62546	62501	533	1072	1062	11	0.017
4432-6M-A	71527	71403	644	5396	5315	70	0.069
4432-6M-B	84552	84452	291	1519	1466	27	0.017
4432-6M-C	78460	78298	6620	9155	9016	64	0.103
4432-6M-D	81056	81012	375	1369	1340	21	0.016
4432-6M-E	61191	60819	100	4638	4590	21	0.07
4432-6M-F	73674	73571	46	6993	6891	26	0.086
4432-6M-G	84267	84172	48	6444	6395	6	0.071
4433-6M-A	79281	79192	246	3611	3580	4	0.043
4433-6M-B	69497	69442	755	5532	5501	46	0.073
4433-6M-C	78748	78699	109	3167	3129	14	0.038
4433-6M-D	112080	112026	516	2957	2922	7	0.025
4428-1Y-A	13633	13615	54	1725	1709	8	0.112
4428-1Y-B	85591	85489	1042	6209	6094	151	0.067
4428-1Y-C	35558	35409	174	24367	24277	2015	0.407
4430-1Y-A	23999	23977	683	152	148	1	0.006
4430-1Y-B	76876	76831	110	1230	1199	12	0.015
4430-1Y-C	31403	31350	72	443	422	0	0.013
4432-1Y-A	18185	18155	141	1879	1851	9	0.093
4432-1Y-B	41342	41290	70	256	249	1	0.006
4432-1Y-C	73268	73145	154	3113	3083	8	0.04
4432-1Y-D	61205	61084	103	10496	10423	18	0.146
Average	**71356.1**	**71182.5**	**486.9**	**10328.4**	**9885**	**93.1**	**0.138**

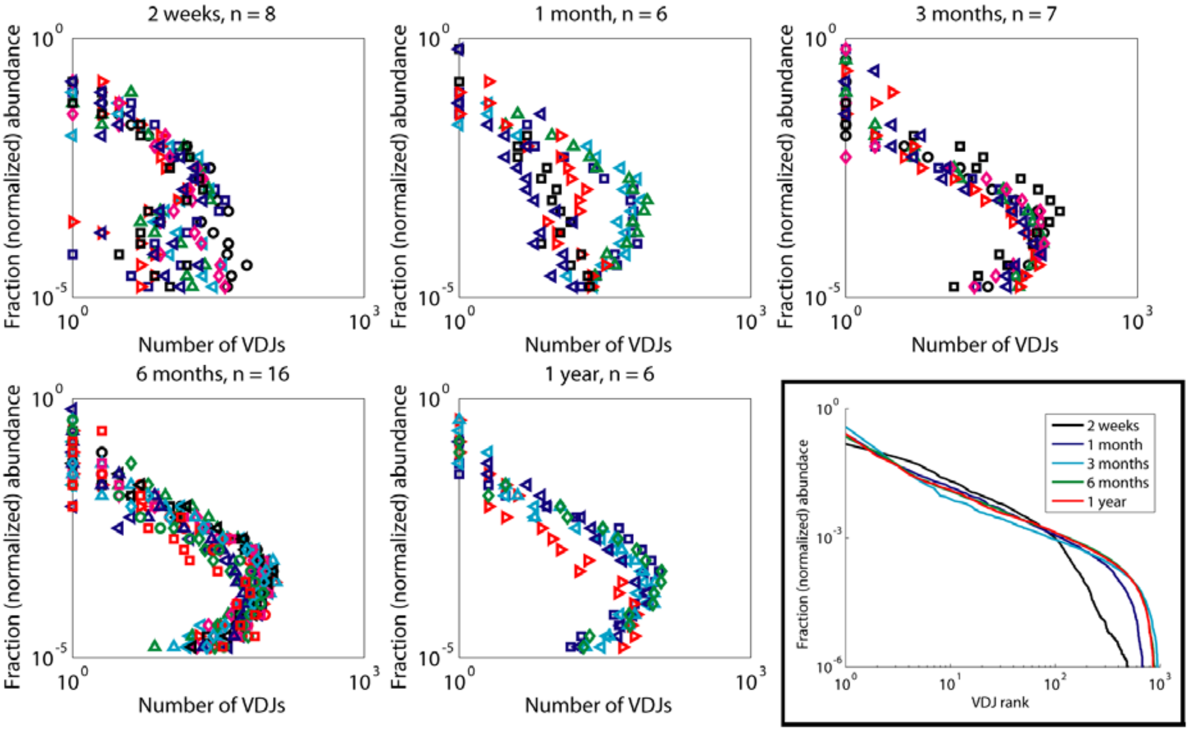
Supplementary Table 2: 4.17 million reads (both IgM and IgZ) were analyzed from 51 samples. Sample name is ordered: family-age-(letter ID). Retained refers to those passing the initial filter, ambiguous (ambig.) refers to sequences that have an equal alignment score between two or more V- or J-gene segments, fraction IgZ refers to the fraction of all reads (IgM + IgZ) that belonged to the IgZ isotype.



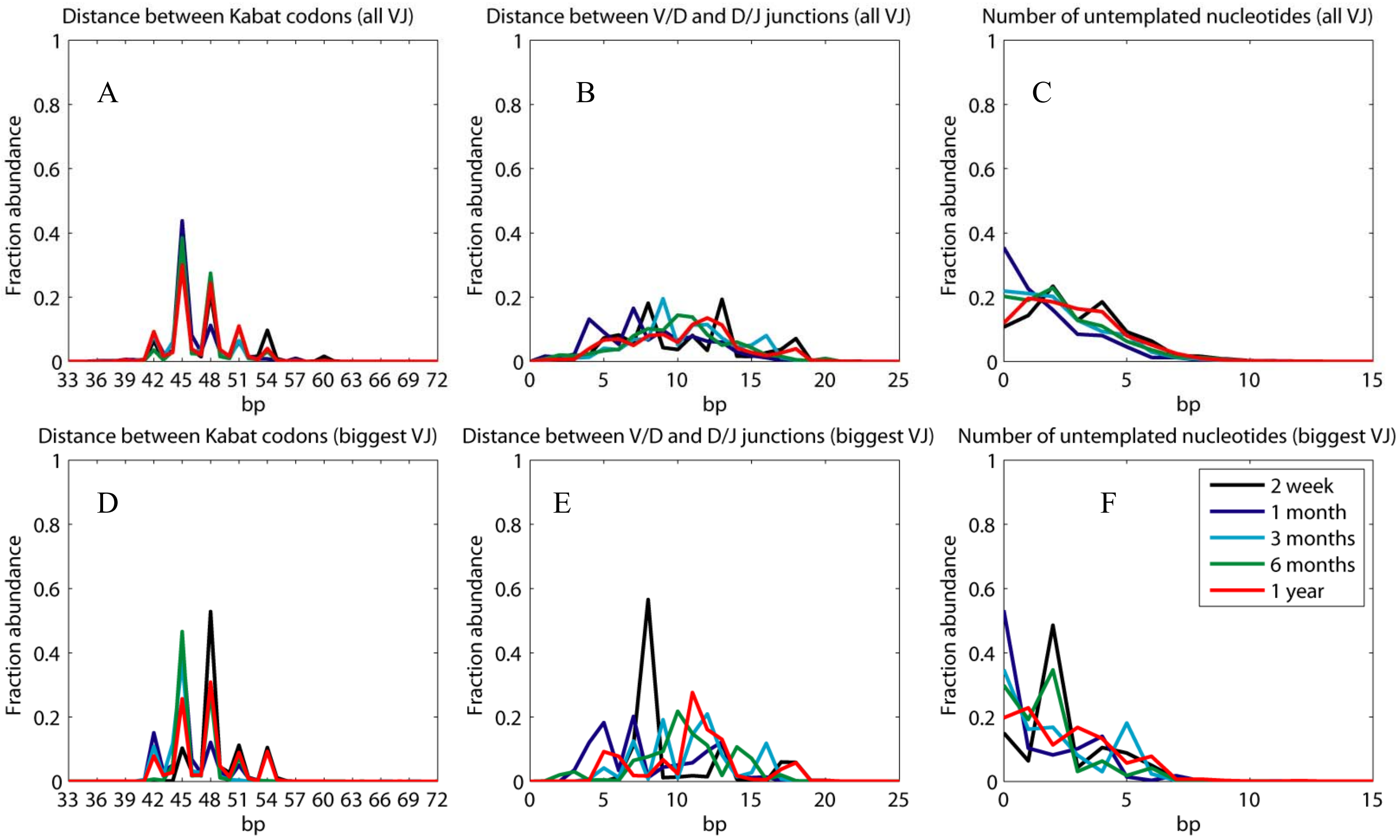
Supplementary Figure 3: VDJ repertoires of a highly-correlated pair (c = 0.67) at age 1 year (sub-sampling is done to 18,000 reads for both individuals). Dot-size scales logarithmically, and color linearly (color-bars indicate fraction of total abundance), with bias-normalized abundance of each VDJ class, as in Figure 1 of the main text. The most abundant VDJ class in both is V1.3-D5-J1.



Supplementary Figure 4: A) Timecourse of IgZ VDJ combinations (of the 39Vx2Dx2J = 156 possible IgZ combinations) sub-sampling 1250 IgZ reads among all fish with at least that many. **B)** Rarefaction of IgM VDJ usage (of the 39Vx5Dx5J = 975 possible combinations), demonstrating variation in VDJ usages both between and within age-groups. **C)** Rarefaction of IgZ VDJ usage, for those fish for which at least 1250 IgZ reads were sequenced. For VDJ assignment was performed after clustering, as described in *Pairwise alignments and full-sequence clustering*.



Supplementary Figure 5: IgM VDJ abundance histograms (rotated) of all age-groups, with sequencing depth fixed at 32,000 reads. VDJ identification was performed after clustering, as described in *Pairwise alignments and full-sequence clustering*. Bottom right, VDJ abundance as a function of rank, averaged between fish within an age group.



Mean lengths (bp)

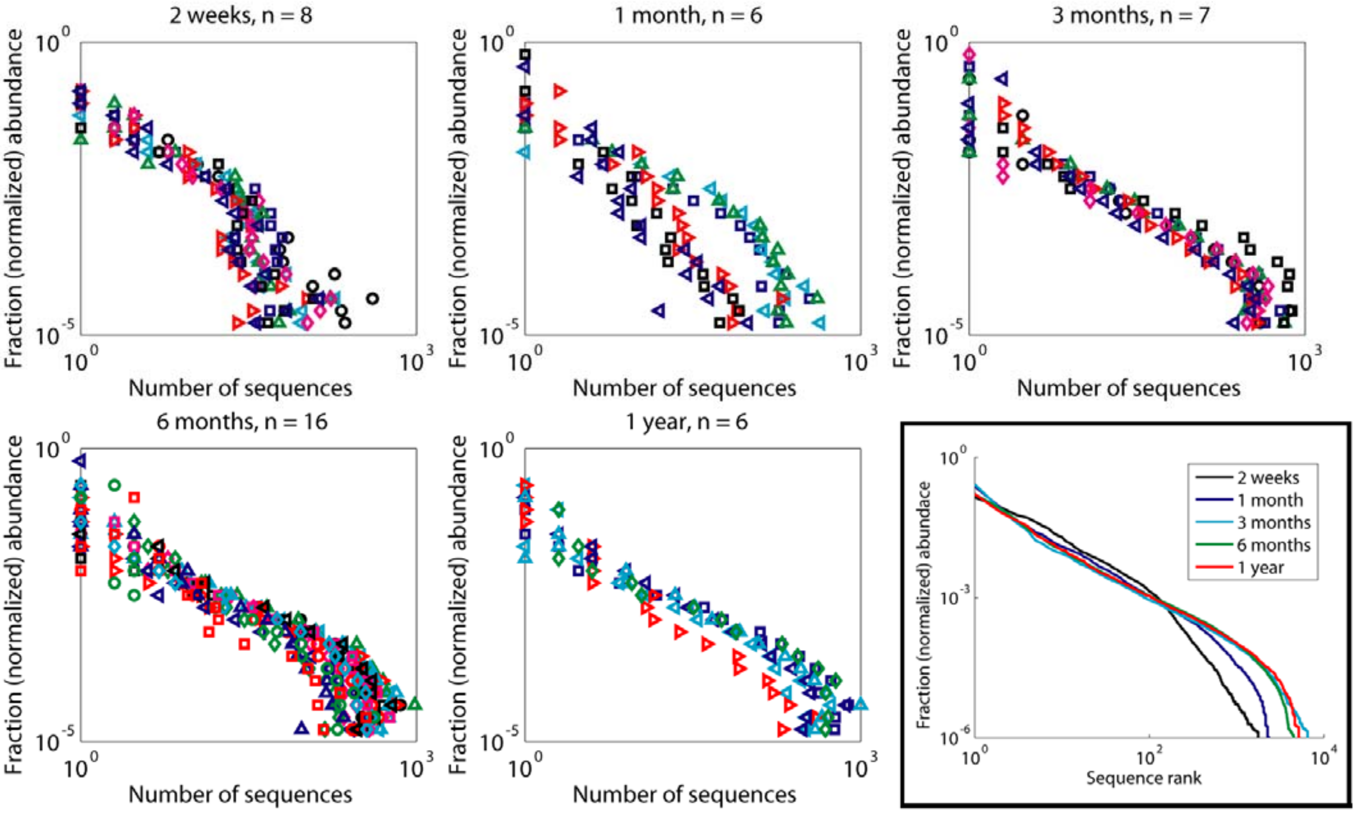
All VJ

Age	Kabat	Junction	Untemp
2 weeks	47.74	10.48	2.75
1 month	45.71	8.22	1.45
3 months	46.65	10.59	2.35
6 months	46.82	10.01	2.13
1 year	47.00	10.40	2.53

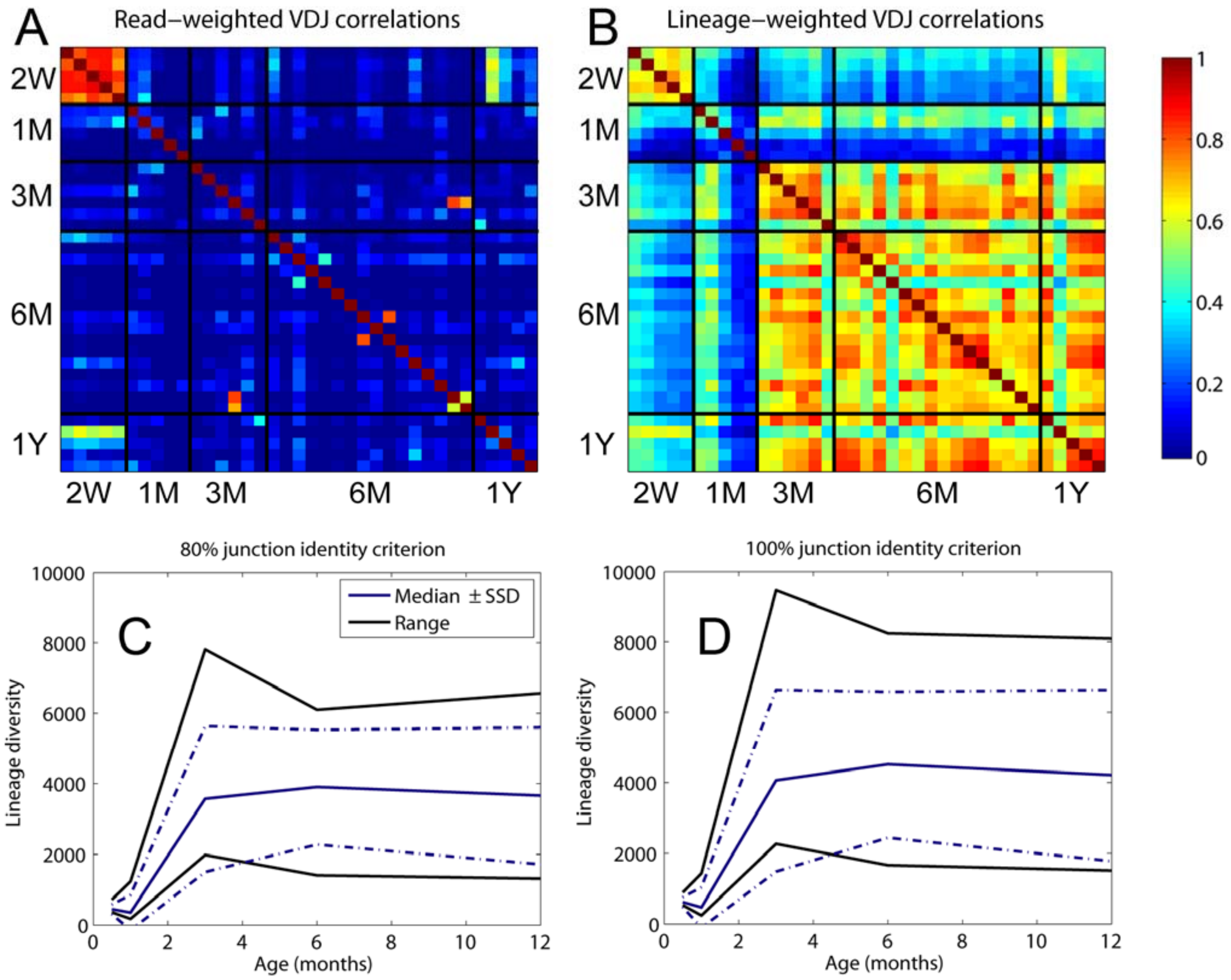
Most abundant VJ

Age	Kabat	Junction	Untemp
2 weeks	48.78	9.88	2.53
1 month	45.30	8.14	1.53
3 months	45.44	10.92	2.39
6 months	45.04	10.55	1.70
1 year	47.43	11.03	2.76

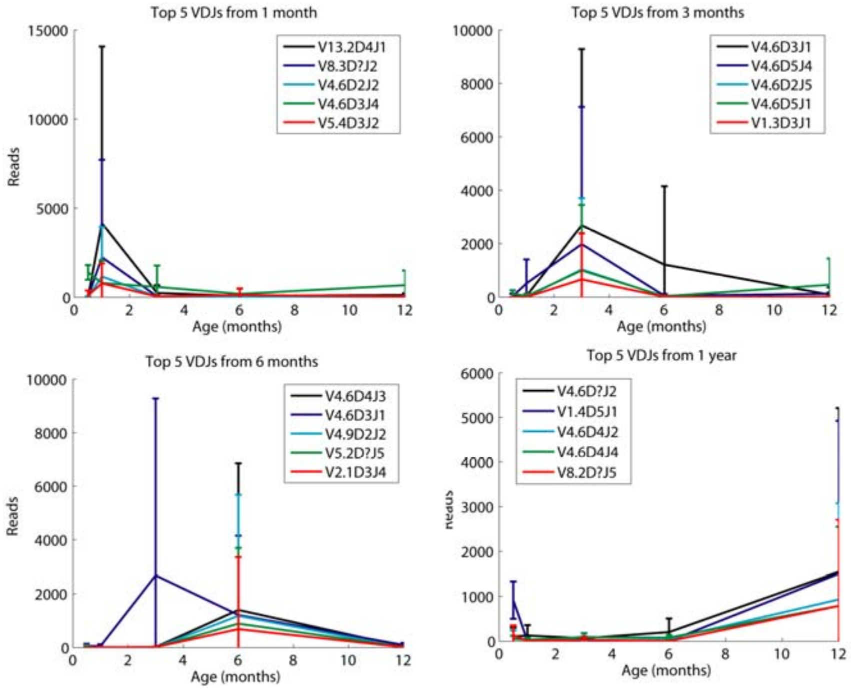
Supplementary Figure 6: The most abundant VJ combinations' CDR3 lengths are histogrammed alongside CDR3 lengths for all data put together. CDR3 length is measured by that between Kabat codons (per Kabat et al.(2), found here by assigning the best un-gapped alignment of the consensus sequence `ta(t/c)t(a/t)(c/t)tg(t/c)` to the 5' bound and the best alignment of `ta(t/c)tgggg` to the 3' bound), between V/D and D/J junctions, and in terms of untemplated (N-) nucleotides at the V/D and D/J junctions. While no clear trend is seen on the average, junction sizes for the most abundant VJ combination at 2 weeks assume 2 localized peaks: a property not seen at later ages.



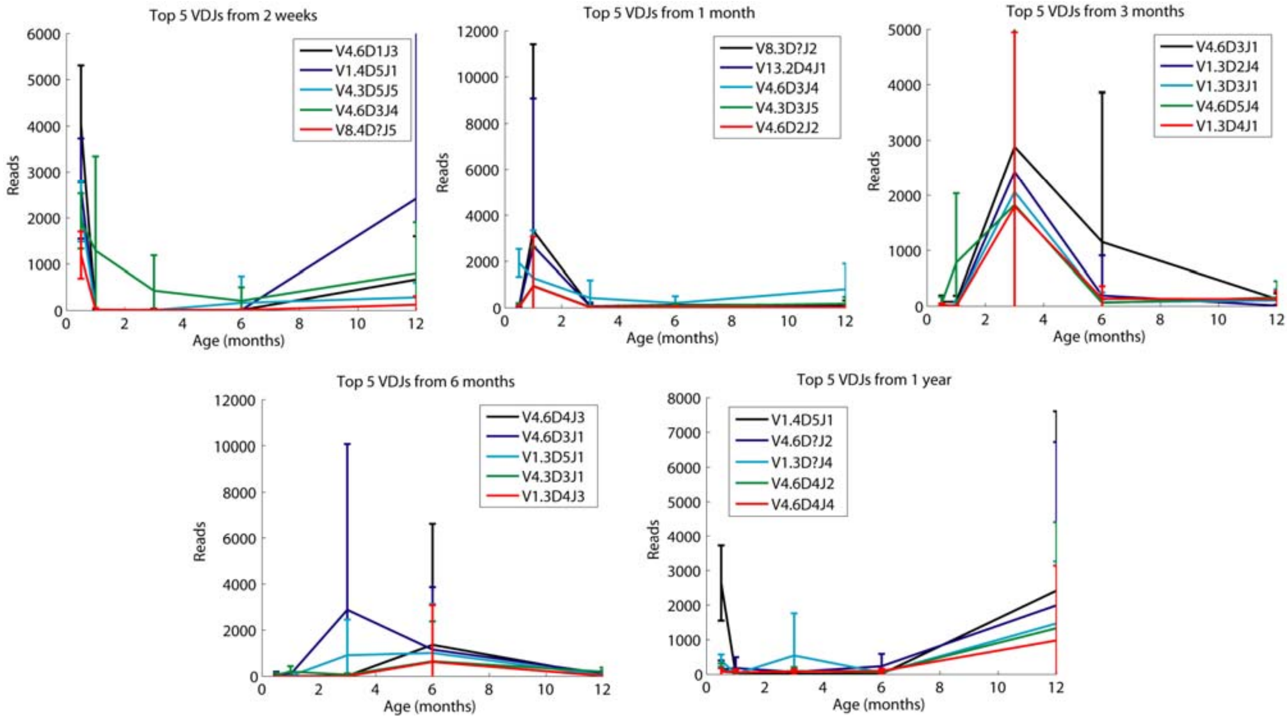
Supplementary Figure 7: Antibody abundance histograms (rotated) of all age-groups, with sequencing depth fixed at 32,000 reads. Antibody counting was performed after clustering, as described in *Pairwise alignments and full-sequence clustering*. Bottom right, antibody abundance as a function of rank, averaged between fish within an age group.



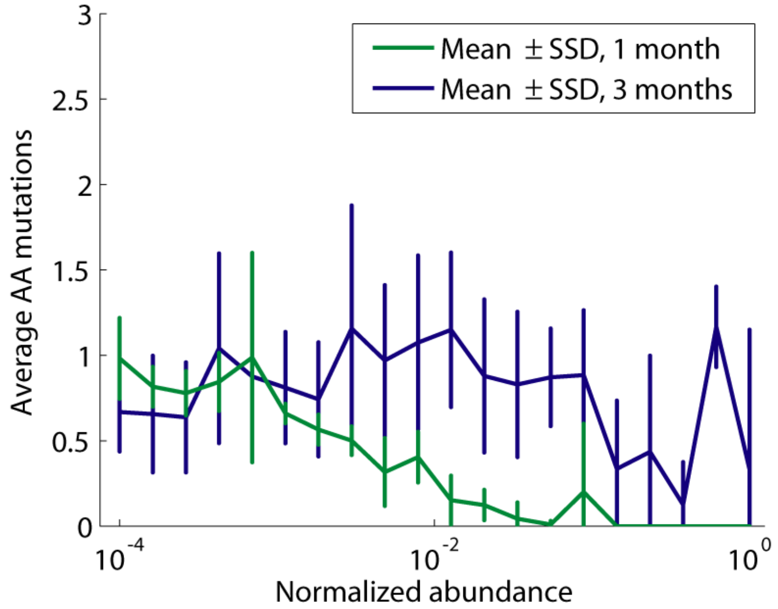
Supplementary Figure 8: A) Read-weighted and **B)** lineage diversity-weighted VDJ correlations between all fish that have been sub-sampled to 40,000 identifiable IgM reads for all fish with at least that many, using *zero-error/mutation* tolerance single-linkage clustering (100% junction-identity, as described under *Lineage Analysis*). **C)** Lineage diversity counted using the 80% junction-identity and **D)** 100% junction-identity criteria described under *Lineage Analysis.*



Supplementary Figure 9: The most abundant VDJ combinations (by raw number of reads) for 1 month, 3 months, 6 months, and 1 year timepoints (2 weeks is shown in the main text Fig. 2C). In contrast to 2 weeks, illustrated in the error bars, no consensus exists for highly represented VDJ combinations within individual timepoints.



Supplementary Figure 10: The most abundant VDJ combinations, ranked by bias-normalized abundance. The results recapitulate the rapid disappearance of VDJ combinations for which a high-abundance consensus (as seen in the error-bars across individuals) exists at the 2 week timepoint.



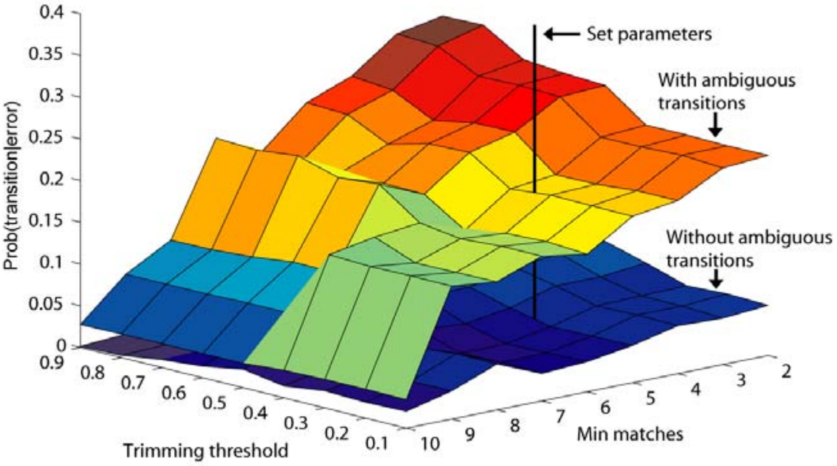
Supplementary Figure 11: The 1 month-to-3 month transition in the amino acid mutation/abundance plot.

Final determination of junction boundaries

The terminal points of germline sequences on VDJ recombinants were determined by using a overhang-window discussed under *VJ classification and Junction Determination*, and illustrated in Supplementary Figure 1. Overhang trimming thresholds of 51% were assigned to V and J alignments and trimming thresholds of 34% were assigned to D-alignments. A simulation in Supplementary Figure 12 applied to the PCR-control data shows the effects of varying the D-trimming parameter along with the minimum number of matches to call a D-assignment unambiguous. If an error (either an indel or mismatch) occurs on the junction, there would be a 5% chance of another unambiguous D-assignment being made. Since the total error rate for a 10-bp stretch (the average junction length from Supplementary Figure 6) is roughly 3%[1], we expect these transitions to affect <0.1% of the data. While the rate at which an indel or mismatch can make a D-assignment ambiguous is higher, the expected total rate remains less than one percent.

Annotations of all reads analyzed can be found at the public URL https://sites.google.com/site/zebrafishdev/files.

[1] From (1) we find that after PCR, 0-error reads of length 170 nt (after the 10 nt MID barcode and 20 nt primer are removed) comprise, conservatively, 60% of the total, giving a Poisson per-bp error rate of ~0.3%.



Supplementary Figure 12: Effect of junction-errors on D-assignments. For each read in the PCR control set (comprising 38 IgM clones), a replacement or indel error was made to the junction itself, and a D-assignment was made. This simulation was iterated over every pair of parameters displayed above. For those transitions to and from ambiguous D

Data sub-sampling

Analyses involving data sub-sampling randomly drew reads (without replacement) from the IgM-retained pool in Supplementary Table 2 (or the IgZ-retained for IgZ analyses). Such analyses were restricted to those individuals that had been sequenced deep enough to make this sub-sampling possible. Lineage analysis, which sub-sampled 40,000 IgM reads, incorporated 5 individuals from the 2 week time point, 5 individuals from 1 month, 6 individuals from 3 months, 16 individuals from 6 months, and 5 individuals from 1 year, for a total of 37 individuals.

Amino acid usage

Amino acid usage (plotted in Figures 4A and 4B of the main text) at fixed loci was determined by fixing sequencing depth, and filtering out reads having insertions and deletions (relative to the reference genome), stop-codons, and ambiguous base-calls. Reads were grouped by perfect identity, and amino acids were tallied from those sequences (including singletons).

Lineage analysis

Preliminary filtering was performed by discarding reads with stop-codons and ambiguous bases. Reads were grouped by perfect nucleotide identity, and VDJ junction boundaries were defined using the method described above. Sequences from the same VJ combination were compared at their VDJ junctions (here defined as the region from the end of the V-templated portion of the sequence to the beginning of the J-templated portion). Any two sequences with junction boundaries varying by at most one nucleotide and having greater than or equal to 80% identity at the VDJ junction were allowed to cluster together. The resulting single-linkage clusters allowed for the chaining of sequences, even if their VDJ junctions differed slightly due to mutation or error. Clustering was performed between sequences of the same VJ combination irrespective of their differences outside the VDJ junction. Sequences retained their identity, but the clusters they formed defined hypothetical lineages. Lineage diversity, whether the total for the individual or for single VDJ combinations, included all such clusters, regardless of size.

Whichever member sequence had the fewest nucleotide differences relative to the reference genome outside the VDJ junction was defined as the naïve sequence of the lineage. Mutations in the mutation-dependence plots (Figure 4C and 4D of the main text and Supplementary Figure 11) were determined by direct comparison to this sequence. For these plots, sequences were excluded if they exhibited insertions or deletions outside of their junction regions (on the V or J segments) or if they exhibited insertions or deletions relative to the designated naïve sequence.

In those cases in which multiple sequences within a lineage had the same number of differences relative to the germline (with their own differences hidden at the junction), the more abundant sequence was chosen as the naïve sequence. Sequences within the lineage that had either stop-codons or frame-shifts relative to the naïve sequence were discarded. Subsequently, all mutations were counted relative to the naïve sequence.

Pearson correlations were performed (main text Figs 2D, 3B, 3C, 3D, and Supp. Fig. 8) either by weighting each VDJ-vector by the read-abundance or lineage diversity. More precisely, the correlation coefficient c_{kl} between fish k and fish l was defined by:

$$c_{kl} = \frac{\Sigma_{VDJ}\left[\left(x^{(k)}_{VDJ} - \left\langle x^{(k)}\right\rangle\right)\left(x^{(l)}_{VDJ} - \left\langle x^{(l)}\right\rangle\right)\right]}{\sqrt{\Sigma_{VDJ}\left(x^{(k)}_{VDJ} - \langle x^{(k)}\rangle\right)^2 \Sigma_{VDJ}\left(x^{(l)}_{VDJ} - \langle x^{(l)}\rangle\right)^2}}$$

with x_{VDJ} representing either read-abundance or lineage-diversity for a particular VDJ class, and $<x>$ representing the read-abundance or lineage-diversity averaged over all VDJ combinations. These VDJ included the ambiguous D-class, and no qualitative difference was found in the correlations by excluding it.

For the correlations in Figure 2D of the main text, bias-parameters (see *Bias calculation*) were applied to VDJ read-numbers to give bias-normalized VDJ-abundances. In order to demonstrate that lineage-diversity and lineage-stereotypy were not being dominated by sequencing depth, raw VDJ read-numbers were used to measure these correlations in Figures 3B and 3C of the main text and Supplementary Figure 8A. These demonstrate that abundance-correlations behave the same regardless of whether VDJ-abundances are bias-normalized, and further demonstrate that the distribution of lineage diversity is not dominated by sequencing depth. This is further expounded upon in the main text discussion (see *VDJ Usage of Antibody Primary Repertoire Is Also Stereotyped in Mature Animals*).

Supplementary Figure 8 demonstrates the robustness of VDJ lineage-diversity correlations to changes in lineage-clustering thresholds. Instead of grouping VDJ junctions if they differ by at least 80% (as described in *Lineage analysis*), VDJ junctions were instead required to be exactly the same. Supplementary Figure 8 demonstrates that this change has a small effect on two-week correlations, which increase slightly – exactly what would be expected from zero-error tolerance making measured diversity much more sensitive to sequencing depth. The increase in lineage-weighted correlations at three months and onward is also slightly affected, however the overall phenomenon remains.

Pairwise alignments and full-sequence clustering

For VDJ- and antibody diversity-counting (Figures 1 and 2 in the main text, and Supplementary Figures 3, 4, 5, 7, 9, and 10), pairwise Smith-Waterman alignments and quality-threshold clustering were performed as described (*1*). In clustering, sequences mapping to V-gene segments whose 75 bp nearest the CDR3 were within twice the clustering-radius of one another were aligned and allowed to cluster. This was done to ensure that sequencing errors and mutations confusing the initial alignment of un-clustered sequences to V-segments would not prevent these sequences from being compared, or even clustered together if they were sufficiently similar. Once clustering was completed, these sequences' consensus sequences – assigned by the most redundant sequence within a cluster (or in the case of zero or equal redundancy, by the read that maximized the kernel Σ_{ij} exp(-d_{ij}), d_{ij} being the pairwise distance between sequences *i* and *j* within the cluster) – were used for V-determination.

Homopolymer errors, the most recognizable of sequence-dependent 454 errors, were followed by implemented a homopolymer error-filter that searched for 3-mers on either side of every gap, with alignments

```
aaaagc                gacaaa
aaa-gc      and       g-caaa
```

de-weighted to one-hundredth that of a normal mismatch. In constructing pairwise distance matrices, Smith-Waterman alignments were averaged so that
d_{ij}= (SW(i,j) + SW(j,i))/2.

Bias calculation

Of the 38 V-gene segments covered by primer set 1, clones from 18 of these V-gene segments were used as templates with and without PCR. Of the 39 V-gene segments covered by primer set 2, 36 were used as templates with and without PCR[2]. Combining information from both sets of clones and the VDJ representations under both primer sets in 6 six month-olds (TS-6M-A through TS-6M-F) were used to approximate a maximum-likelihood estimate of the bias parameters. This was done by weighting discrepancies between both before- and after-PCR time points and fish-fish comparisons with the appropriate binomial error terms. The objective function *H* of bias parameter sets $\{\alpha\}$ and $\{\beta\}$ (corresponding to primer sets 1 and 2, respectively) used summed error terms collected using the 0- and 35-cycle control sets (dealt with separately at first):

$$H(\vec{\alpha},\vec{\beta}) = \frac{1}{2}\left(F_0(\vec{\alpha},\vec{\beta}) + F_{35}(\vec{\alpha},\vec{\beta})\right) + G(\vec{\alpha},\vec{\beta})$$

The function F_C, for C = 0 or 35 PCR cycles, accounted for error in the absolute values of the bias parameters themselves, and was defined by

$$F_C(\vec{\alpha},\vec{\beta}) = \sum_{v=1}^{39}\left[\frac{(\hat{M}_{v,C}(\alpha_v) - M_{v,C})^2}{M_C p_{v,C}(1-p_{v,C})} + \frac{(\hat{N}_{v,C}(\beta_v) - N_{v,C})^2}{N_C q_{v,C}(1-q_{v,C})}\right]$$

where $M_{v,C}$ is the fraction representation of V-gene segment *v* after *C* PCR cycles on the primer 1 control set, $N_{v,C}$ is the same with the primer 2 control set. The following definitions also apply:

[2] These three, V4.7, V5.5, and V9.3, sharing primers with and being nearly identical in sequence and length to V4.5, V5.2, and V9.1, respectively (see Supplementary Table 1), were assigned those V-genes' pre- and post-PCR read-measurements. Subsequent bias-parameter tuning using the described optimization algorithm with real data from samples TS-6M-A through TS-6M-F was applied to these three V-genes exactly like the other 36.

$$\begin{aligned} M_C &= \sum_{v=1}^{39} M_{v,C} & N_C &= \sum_{v=1}^{39} N_{v,C} \\ p_{v,C} &= M_{v,C}/M_C & q_{v,C} &= N_{v,C}/N_C \\ \hat{M}_{v,0}(\alpha_v) &= M_0\left(\frac{M_{v,35}/\alpha_v}{\sum_{v'=1}^{39} M_{v',35}/\alpha_{v'}}\right) & \hat{N}_{v,0}(\beta_v) &= N_0\left(\frac{N_{v,35}/\beta_v}{\sum_{v'=1}^{39} N_{v',35}/\beta_{v'}}\right) \\ \hat{M}_{v,35}(\alpha_v) &= M_{35}\left(\frac{M_{v,0}\alpha_v}{\sum_{v'=1}^{39} M_{v',0}\alpha_{v'}}\right) & \hat{N}_{v,35}(\beta_v) &= N_{35}\left(\frac{N_{v,0}\beta_v}{\sum_{v'=1}^{39} N_{v',0}\beta_{v'}}\right) \end{aligned}$$

The function *G* accounted for error in the *ratios* between bias parameters of the two primer sets, and was defined by

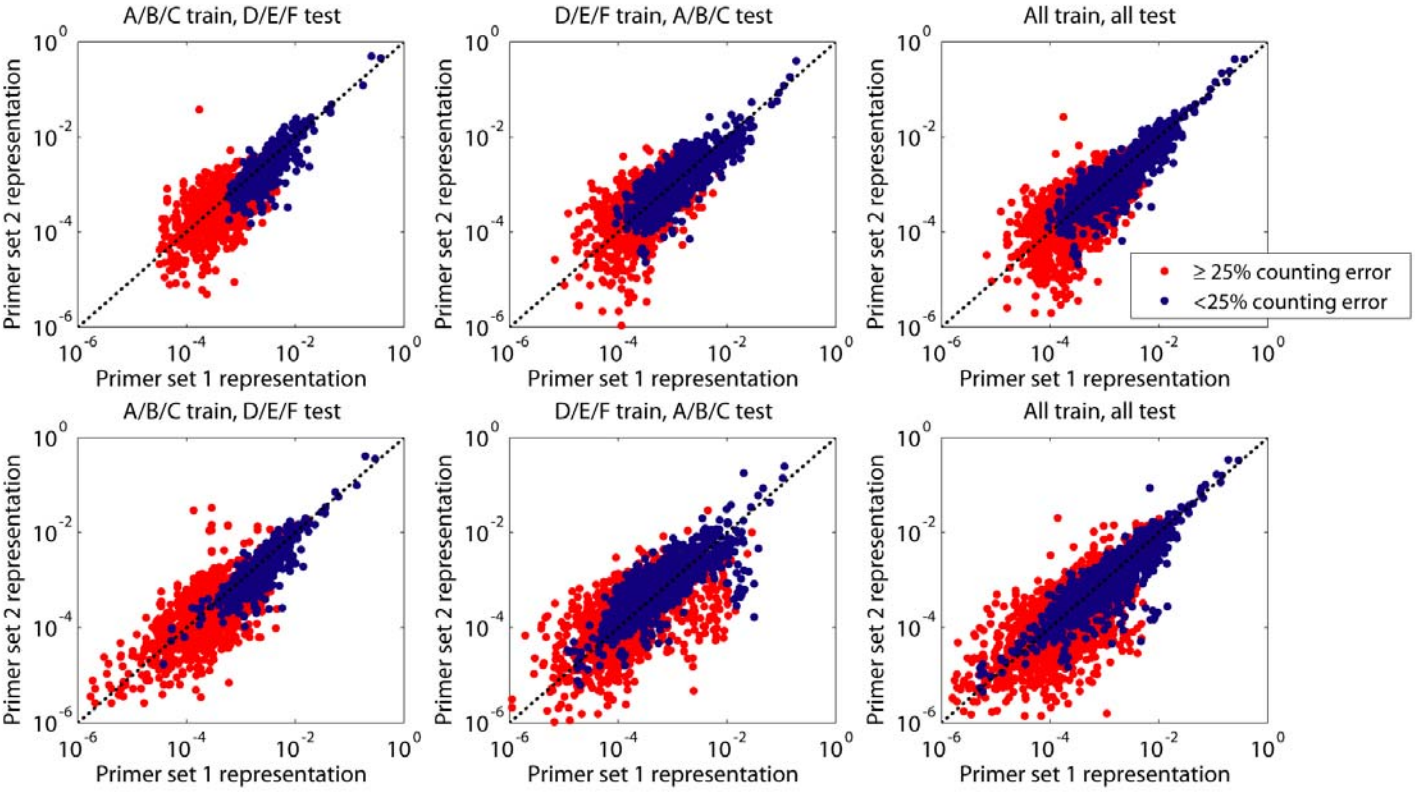
$$G(\vec{\alpha},\vec{\beta}) = \sum_{i}^{\text{all fish}} \sum_{v=1}^{39} \frac{\left(\left(P_{v,i}/\alpha_v\right)\left(\sum_{v'} P_{v',i}/\alpha_{v'}\right)^{-1} - \left(Q_{v,i}/\beta_v\right)\left(\sum_{v'} Q_{v',i}/\beta_{v'}\right)^{-1}\right)^2}{\frac{S_i P_{v,i}(1-P_{v,i})}{S_i^2\alpha_v^2\left(\sum_{v'} P_{v',i}/\alpha_{v'}\right)^2} + \frac{T_i Q_{v,i}(1-Q_{v,i})}{T_i^2\beta_v^2\left(\sum_{v'} Q_{v',i}/\beta_{v'}\right)^2}}$$

where $P_{v,i}$ is the fraction representation of V-gene segment *v* in the *i*th fish with primer set 1 used, $Q_{v,i}$ is the same with primer set 2 used, S_i is the total number of reads collected from fish *i* using primer set 1, and T_i is the same with primer set 2.

MATLAB script performing the above optimization using the `fmincon` function is available upon request, and the bias parameters for primer set 2 (the set used in all other data analysis) is shown in Supplementary Table 3C.

The separation of fish data into “testing” and “training” sets demonstrated the convergence of bias parameters, provided there was sufficient information in the VDJ repertoires of fish subsets. These correlograms are plotted in the first two columns of panels in Supplementary Figure 13. All fish together, with data from all gene segments, are shown in the rightmost top and bottom panels.

Because primer set 1 had before- and after-PCR snapshots covering only 18 out of 39 gene segments, these 18 gene segments are plotted together in the first row of panels in Supp. Fig. 13. The remaining bias parameters for primer set 1 were still estimated, however, using $G(\alpha,\beta)$ which encoded their relationships to bias-parameters from primer set 2. The resulting correlograms are shown in the second row of panels in Supplementary Figure 13, and the resulting Pearson correlations and R^2 coefficients from the use of all fish in training and testing bias-parameters are shown in Supplementary Tables 3A and 3B.



Supplementary Figure 13: Bias normalization using two primer sets establishes strict guideline for replicability. VDJ combinations corresponding to read numbers at greater than 25% binomial counting error are colored red, those with counting error smaller than 25% are colored blue. Train/test-sets examine those V-gene segments for which control clones exist for primer set 1 (top) and for all V-gene segments together (bottom). Clones and fish A-F were the same as from (1), however all controls were re-run using the 454 Titanium platform (and re-analyzed using reads truncated at 250 bp, rather than 200 bp) which generated the data used in this paper.

	<25% counting error	Everything
Correlation	0.954	0.952
R^2	0.911	0.906

Supplementary Table 3A: Primer-set comparisons, data from Supp. Fig. 7.

	A2	B2	C2	D2	E2	F2
A1	**0.958**	0.1	0.152	0.418	0.046	0.037
B1	0.199	**0.936**	0.091	0.299	0.029	0.035
C1	0.141	0.079	**0.895**	0.239	0.044	0.021
D1	0.385	0.155	0.168	**0.868**	0.074	0.227
E1	0.076	0.048	0.043	0.157	**0.943**	0.011
F1	0.037	0.035	0.024	0.22	0.008	**0.991**

Supplementary Table 3B: Cross-correlations of all data in all fish using two primer sets.

Exon	Bias
1.1	2.137
1.2	2.111
1.3	0.169
1.4	0.424
2.1	3.912
2.2	3.228
2.3	0.019
3.2	0.8
4.1	3.554
4.2	7.469
4.3	1.001
4.5	0.272
4.6	0.889
4.7	0.293
4.8	3.763
4.9	2.671
5.1	7.991
5.2	1.463
5.3	0.071
5.4	1.557
5.5	3.523
5.7	3.099
5.8	1.134
6.1	0.095
6.2	0.366
7.1	0.785
8.1	0.35
8.2	1
8.3	0.47
8.4	0.85
9.1	0.07
9.2	6.67
9.3	0.08
9.4	2.65
10.1	0.31
11.1	0.24
11.2	0.25
13.2	2.75
14.1	1.1

Supplementary Table 3C: Bias parameters used (primer set 2), divided by median value. Parameters span two orders of magnitude.

References

1. Weinstein JA, Jiang N, White RA, 3rd, Fisher DS, & Quake SR (2009) High-throughput sequencing of the zebrafish antibody repertoire. *Science* 324(5928):807-810.
2. Kabat EA, Wu, T. T., Perry, H. M., Gottesman, K. S., Foeller, C. (1991) *Sequences of Proteins of Immunological Interest* (Public Health Service, National Institutes of Health, Washington, DC) 5 Ed.