Supplementary Materials for

Lineage Structure of the Human Antibody Repertoire in Response to Influenza Vaccination

Ning Jiang, Jiankui He, Joshua A. Weinstein, Lolita Penland, Sanae Sasaki, Xiao-Song He, Cornelia L. Dekker, Nai-ying Zheng, Min Huang, Meghan Sullivan, Patrick C. Wilson, Harry B. Greenberg, Mark M. Davis, Daniel S. Fisher, and Stephen R. Quake[†]

[†] To whom correspondence should be addressed. Email: quake@stanford.edu

Section 1: data analysis

Primary sequencing reads processing

Reads from Roche 454 entered into the primary analysis; see the flowchart in fig. S1. Only those reads that matched exactly to the corresponding sample's MID code were included for further process. These reads were then filtered for a minimum length of 250 bp, see Supplementary Table 3 for filtered reads number for each sample. Longer reads were truncated to this length. Only these reads were considered for the rest of the analysis. The isotype of each read was identified by the reverse primer at the constant region; see fig. S2 for isotype composition.

Isotype class switch

One important process of affinity maturation is isotype switching. The naïve B cells are mostly expressing IgM and they undergo isotype switching from IgM to IgG, IgA or IgE after antigen stimulation and proliferation. We explicitly verified the presence of isotype switching in the sequence data by taking a closer look at the lineages in the plasmablasts for each subjects. We re-performed the clustering analysis on pooled IgM and IgG sequences and allow both isotypes to join the same lineage as long as their CDR3 region satisfied the predetermined criteria, which is 1 amino acid difference. Here, we consider a lineage is isotype switched, if both IgM and IgG sequences are observed in the lineage. Elderly have less isotype switched lineages than that of children (see fig. S4). We observed only a few cases where the identical variable region can be found in both IgG and IgM isotopes. But for most of the cases, extensive mutations also occurred throughput the variable region as exemplified in fig. S5.

VDJ classification

Human heavy-chain variable gene segment sequences (244 V-exon, 37 D-exon and 13 J-exon) were downloaded from the International ImMunoGeneTics information system database (IMGT, http://www.imgt.org/textes/vquest/refseqh.html) (*27*), excluding pseudogenes. These germline sequence templates of V and J can be grouped by combining alleles into subclasses. In total, 63 V and 7 J subclasses were obtained. D-exons were not grouped because they are highly diversified. Each read was first aligned to V-consensus sequences taken over each V-family. The specific V-variant with a maximum Smith-Waterman score (*15*) was then assigned. J-segments (out of 13 variants) and D-segments (out of 37) were then assigned as described (*15*), with ambiguous D-segments given their own class. Grouping each V, D, and J gene assignment into subclasses gave the total number of possible VDJ subclass assignments of 63 (V) ×37 (D) ×7 (J) = 16317 (VDJ). The mutation from germline sequence was counted as the number of substitutions from the best aligned V, D and J templates. Unaligned region at the VD and DJ junctions was excluded from the mutation count.

Translation from nucleotide to amino acid sequences

Nucleotide sequences were translated into amino acid sequences based on codon translation (see the flowchart in fig. S1). We used the amino acid sequences in the IMGT database as the reference to detect the correct translation frame in the V region. If the reading frame was correct at the constant region, the translation was accepted for further analysis. Since the dominant sequencing error in 454 Roche platform is insertion and deletion, if the translation was out of frame at the constant region, we performed sequence translation rescue. We deleted all insertions and filled the deletions by the corresponding nucleotide base in the germline sequences, and performed a second translation. If the second translation had the correct reading frame at the constant region, the rescue was successful and the protein sequence from the second translation was accepted for further analysis, otherwise, the sequence was discarded. About 12% of reads were discarded. CDR3 region detection was based on two sequence markers. The boundary of CDR3 is defined by amino acid sequence from 'tyr-(tyr/phe)-cys' to 'trp-gly' (*29*).

Protein distance analysis

Protein distance analysis is a way to caption overall relationship of proteins in the entire repertoire. The CDR3 sequence difference d_{ij} was defined as the hamming distance between protein sequence i and protein sequence j in the CDR3 region. We used the minimum distance of sequence i to other sequences in the same sample, formally:

$$\tilde{d}_i = \min\left(d_{ij}, j = 1,2 \dots N\right)$$

The VD junction and DJ junction regions are around 15 nucleotide bases long in total, therefore, two sequences with the same VDJ assignment but from independent VDJ recombination event may differ by 5 amino acids ($d_{ij} \approx 5$). The hypermutation rate is estimated to between 10^{-3} and

10^{-4} per base pair per generation (*30*) and the average CDR3 length is 58bp, thus, two sequences that differ by one amino acid at the CDR3 are likely from the descents of the same naïve B cell.

CDR3 distance distribution

In order to study the relative distance between any two sequence read, we performed the following analysis on amino acid sequences following the scheme described in fig. S1. We first catalog amino acid difference in the CDR3 region between any two sequencing reads, which can be as small as 1 amino acid or as large as 18 amino acids, and grouped sequencing reads according to this difference. If a sequence can be grouped to several groups following this difference, then this sequence will be assigned to a group with minimum amino acid difference. Resulted distribution showed two distinct peaks, one is at 1 amino acid and the other expands 4 to 10 amino acids, as observed in fig. S6. This distribution provided us a guideline to set the threshold when clustering was used as a way to group sequences into lineages.

Clustering Sequences into Clonal lineages

Sequences with similar CDR3 are possible progenies from the same naïve B cell and can be grouped into the same clonal lineage. To detect the lineage structure for the antibody repertoire, we performed single linkage clustering on protein sequences of CDR3 region, using a re-parameterization of the method described in Jiang et al 2011 (*15*), accounting for the larger size of the CDR3 and junction in humans as compared to zebrafish. Protein sequences with the same V and J assignment and with CDR3 region differed by no more than one amino acid were grouped together into a lineage. This is equivalent to a biological clone that is under clonal expansion.

The diversity of a lineage or a sample was defined as the number of unique sequences within the lineage or the sample, after grouping identical reads.

Lineage structure of antibody repertoire

After grouping reads into lineages, we can take a detailed look of inter- and intra-lineage structure for a sample as exemplified in Fig. 2 and 4, and fig. S7-S10.

In fig. S8, the lineage structure of plasmablasts of all subjects in visit 2 were displayed in reference to mutation, diversity and reads of a lineage. In two of the elderly individuals (017-060 and 017-043), we observed dominant lineages that account for 49% of all reads observed in those samples. We performed the same lineage structure analysis for individual 017-060 in visit 1.The dominant lineage of 017-060 at visit 2 was not observed in visit 1.

Lineage structure of naïve B cells

Lineage structure of naïve B cells and plasmablasts from the same subject were studied. The naïve B cells are dominant by single-read clusters without mutations from germline sequences while plasmablasts contain many lineages with varying amount of sequencing reads and mutations (see fig. S7).

Antibody abundance distribution

The majority of the lineages contain only one reads, however, a few lineages are highly abundant. The distribution of lineage size of plasmablasts follows power law with the exponent of power -1.7(fig. S13). There is no highly abundant lineage in naïve B cells and this can be explained by the fact that naïve B cells are not stimulated and have not undergone clonal expansion.

Mutation in V and J region

It is concerned that the mutation in D region is unreliable, because the alignment of D region to germline reference could be ambiguous. Here, we demonstrate that the mutation pattern of different age groups in visit 1 PBMC is similar whether mutation in D region is counted or not (fig. S14). The trend of mutation pattern is the same if one compares fig. S14 to Fig. 3B.

Varying clustering threshold

In the Fig. 2-4 of the main text, the single linkage clustering was performed with the threshold of one amino acid difference in the CDR3 region. Here, we show that the mutation patterns remain the same under different clustering thresholds.

Section 2: Control data

Control library

Control experiments were performed by a mixture of cloned zebra fish immunoglobulin genes that covered all possible V gene segments (*11, 15*). Briefly, 38 immunoglobulin genes for IgM containing different V gene segments were cloned into plasmid and sequenced using Sanger sequencing. To quantitatively measure the PCR bias and sequencing errors, plasmids containing 38 clones were pooled. Part of the pooled material went through the same PCR cycle as human samples and part of the pooled material was not PCR amplified. 454 libraries were made using these two sets of materials and sequenced using 454. The degree of error introduced in the sample amplification and sequencing process was estimated by comparing these two sets of control libraries to the most abundant sequence for each template.

The most abundant sequence from each template was chosen to be the reference sequence. Each 454 read was aligned to these reference sequences. All reads were translated into protein sequence and translation rescue was performed as well. As a result, this corrected most of the insertion and deletions which is the most common error for 454 sequencing. Following analysis is focus on substitution errors from PCR and sequencing.77% sequence reads are error-free in their entire 220 bp nucleotide sequence (with 30 bp from the MID barcode and primer being excluded) and 13.6% sequence reads has one substitution (fig. S16B). The substitution error rate was estimated to be 0.065% per bp (detectable chimeras were excluded). Single lineage clustering is performed on nucleotide sequences. Linear relationship between reads in a lineage and unique sequences in a lineage is observed with a slope of 0.147 in the fig. S16C, in agreement with fraction of one-substitution reads. We also did single lineage clustering in protein level and observed a slop of 0.073 in the fig. S16A. In the human data, the diversity is usually higher than what is estimated from substitution error (fig. S17), if the human data is assumed to have the same error profile as zebrafish control data. In most of the subjects, the real sequences are much more than "artificial sequences" created from substitution errors.

Synthetic control data

To test the reliability of our analysis pipelines, we constructed synthetic control data and processed the synthetic data using the same pipeline that was used on human data. The synthetic control data was generated from a single sequence with 3000 reads. The sequence was a randomly selected IgG heavy chain from one subject. Errors were added to the sequence reads to mimic the substitution errors. Each base of each reads was given a constant probability to be changed to another random nucleotide. Therefore, error rate in the synthetic control data was a predefined parameter and was indicated on the X-axis. The range tested is within the estimated substitution rate of 454 sequencing platform (*11*). The analysis showed that sequences included in the largest cluster linearly decrease with the error rate and the number of unique sequences linearly increase with error rate. The mutation from baseline also linearly depended on the error rate (fig. S18).

Supplementary Figures

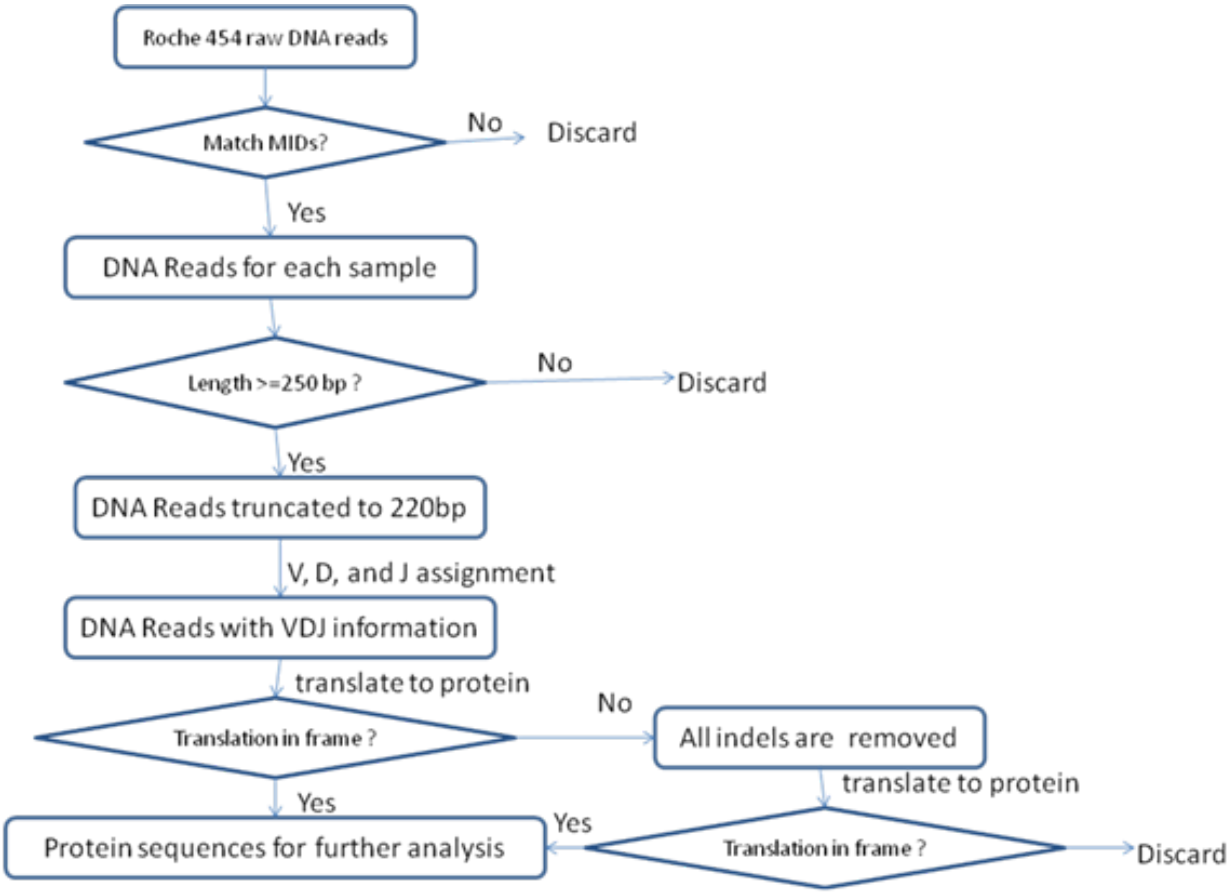
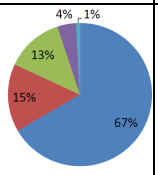
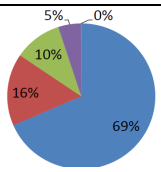
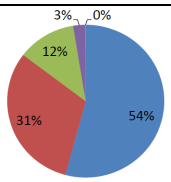
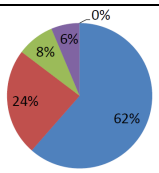
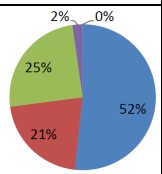
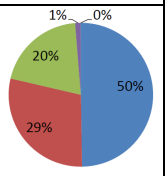
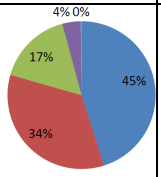
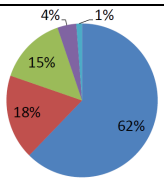
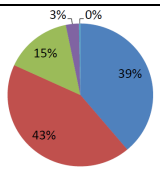
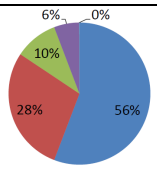
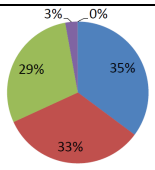
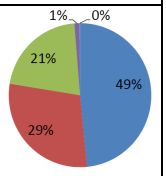
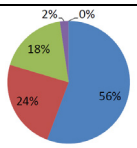
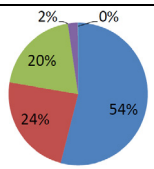
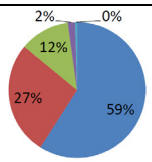
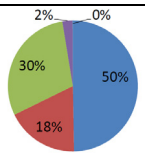
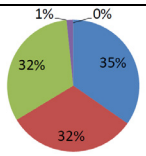
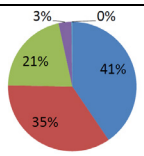
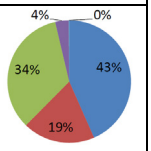
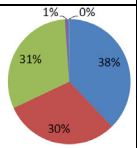
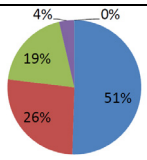
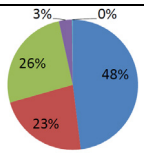
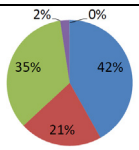
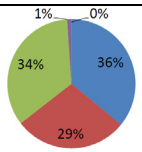
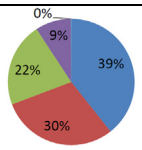
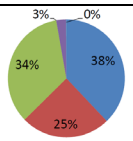
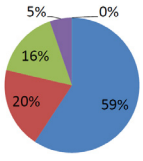
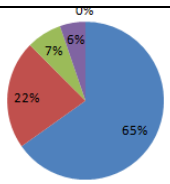
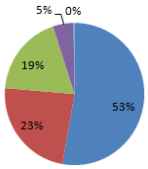
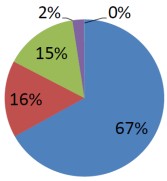


Fig. S1. Flowchart of bioinformatics pipeline. Whether the translated protein was in frame or not was determined by the constant region. The amino acids at the beginning of constant region of correct frame are S-P (IgA), P-T (IgD), S (IgG), S-A-S (IgM). For many samples, IgE accounts for less than 1% of the reads and therefore was not considered further in the analysis.

A

Age 8-17						
Subject	017-051	017-052	017-054	017-053	017-058	017-057
Visit 1						
Visit 3						

Age 18-30							
Subject	017-006	017-011	017-029	017-093	017-134	017-124	017-133
Visit 1							
Visit 3							

Age 70-100				
Subject	017-025	017-043	017-044	017-060
Visit 1				

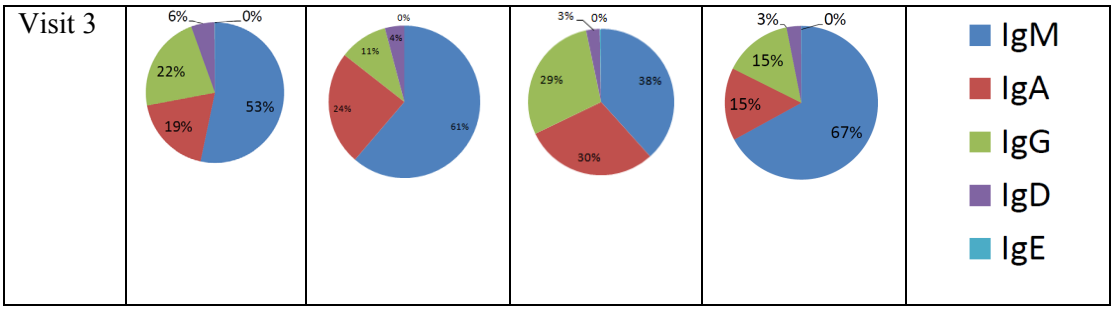


Fig. S2. The composition of five antibody isotypes from PBMCs for each subject at visits 1 and 3. The percentage of an isotype in a subject was calculated using the number of reads within a particular isotype divided by the total number of reads of this sample. Reads from a subset of runs were used. 3000 reads of subsampling was not applied here.

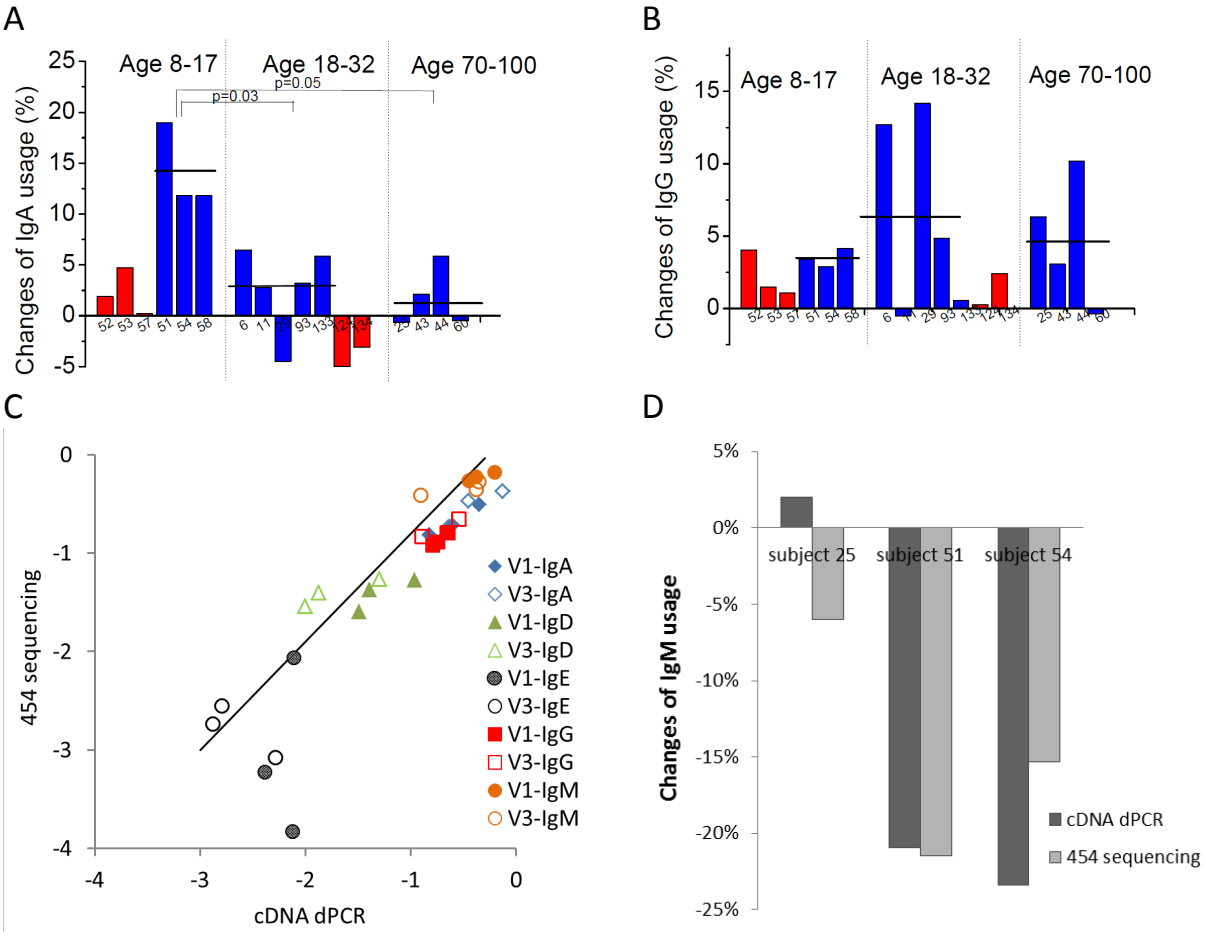


Fig. S3. Isotype changes from visit 1 to visit 3. (A), changes of individual's IgA usage in PBMCs from visit 1 to visit 3. IgA usage increases 14.21% (age 8-17), 2.77% (age 18-32) and 1.71% (age 70-100) at visit 3 on average for TIV receivers (black lines). (B), increase of individual's IgG usage in PBMCs from visit 1 to visit 3. IgG usage increases 3.4% (age 8-17), 6.3% (age 18-32) and 4.8% (age 70-100) at visit 3 on average for TIV receivers (black lines). The subject IDs are labeled on horizontal axis. 3000 reads of subsampling was not applied here, all reads were taken into calculation. Red, LAIV receivers; blue, TIV receivers. Children who received TIV were more likely to have an increased relative IgA usage compared to young adults (p=0.03, Mann-Whitney U test) or elderly (p=0.05, Mann-Whitney U test). Isotype composition was also verified by dPCR (*23, 24*) for three selected subjects. (C), Correlations of 5 antibody isotype compositions (percentage in log scale) between two different measurements (dPCR and 454 sequencing) for three subjects using PBMCs from visits 1 and 3. (D), Changes of relative IgM composition between visits 1 and 3 for these three subjects.

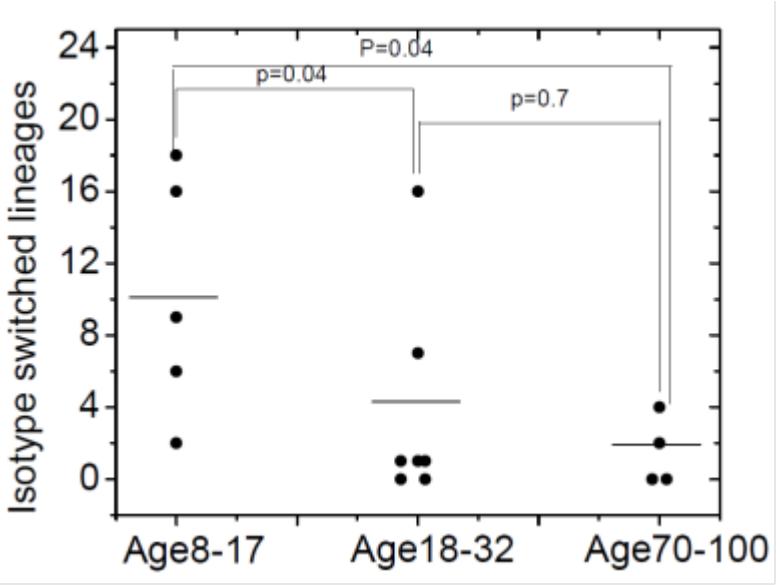


Fig. S4. Young subjects have more lineages that are isotype switched. For each subject, we pooled all IgM and IgG sequences and performed single linkage clustering as defined before. If a lineage is composed of both IgM and IgG sequences, this lineage is defined as an isotype switched lineage. p-values were calculated using Mann-Whitney U test.

Fig. S5. Nucleotide sequence alignment for VDJ region exemplified for one isotype switched lineage. Sequence titles indicate isotype and numbers after N denote number of reads.

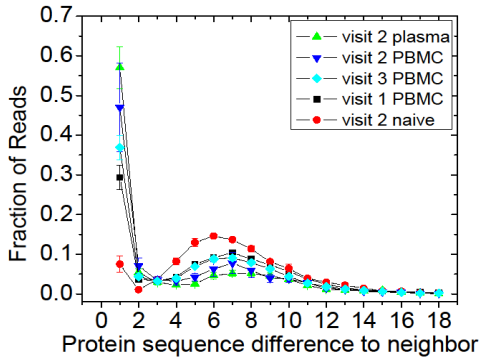
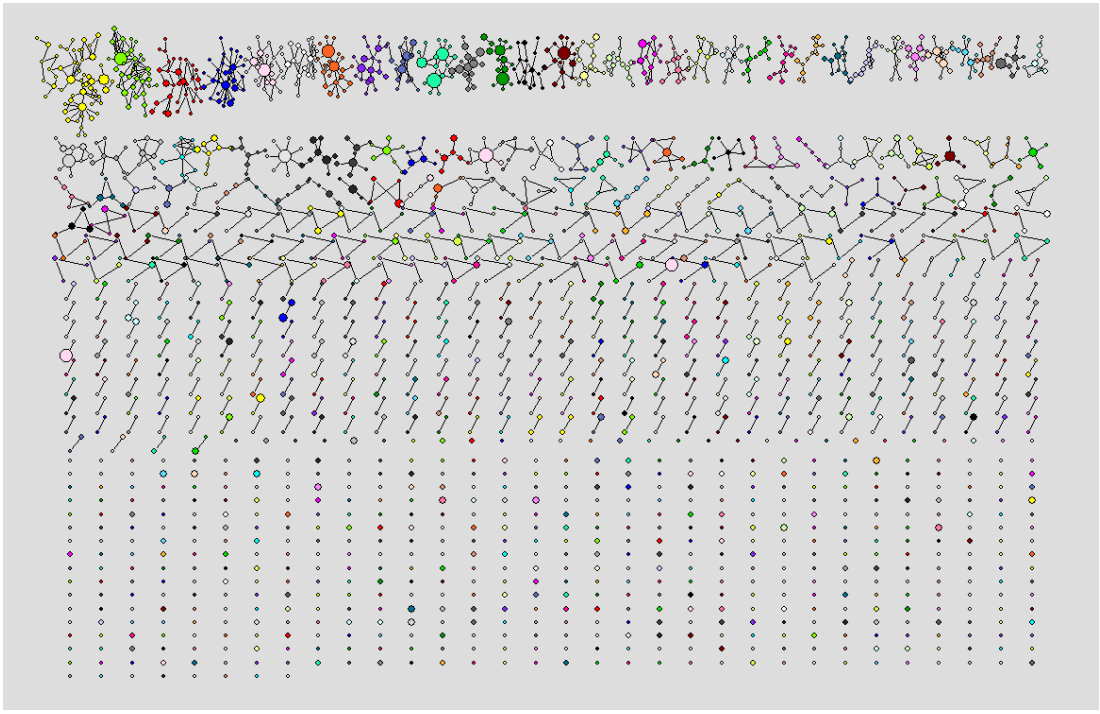
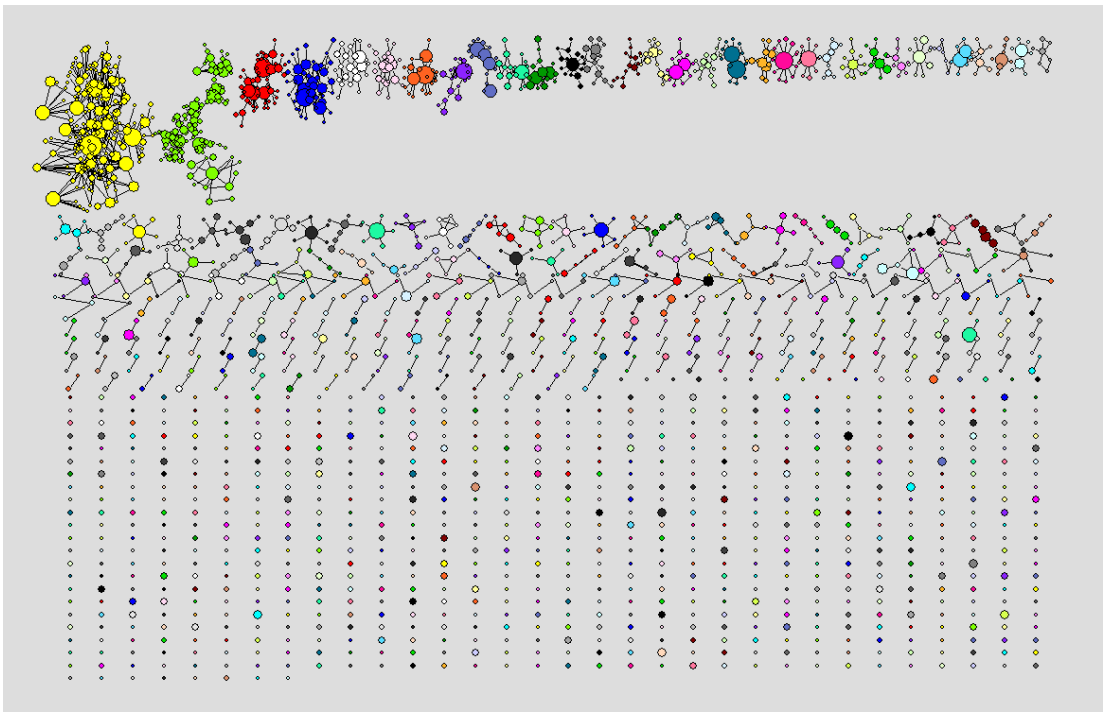


Fig. S6. Reads distribution based on relative sequence distance. For each reads in a sample, we search in the entire repertoire of this sample to find the neighboring read with minimum difference in the CDR3 region at protein sequence level. Distribution of reads with minimum distance to its neighbor were plotted. Sub-sampling of 3000 reads was performed.

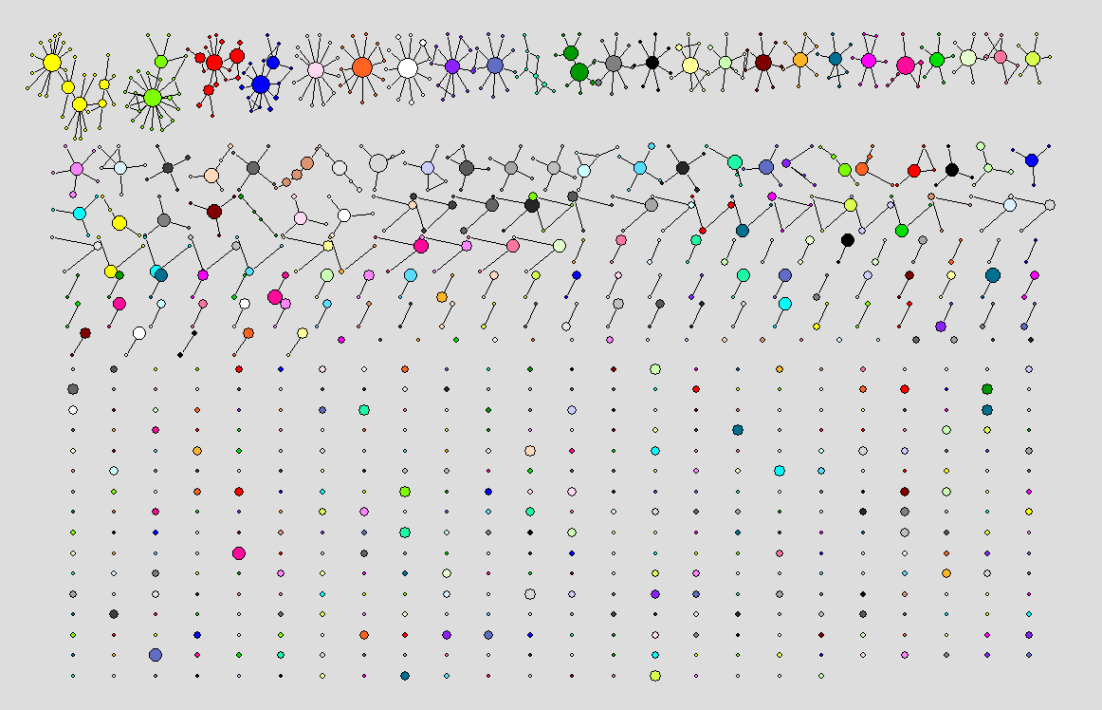
A



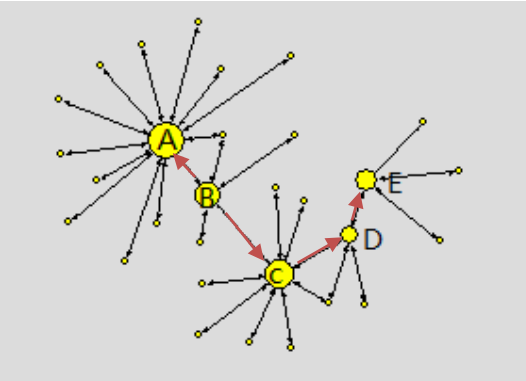
B



C



D



E

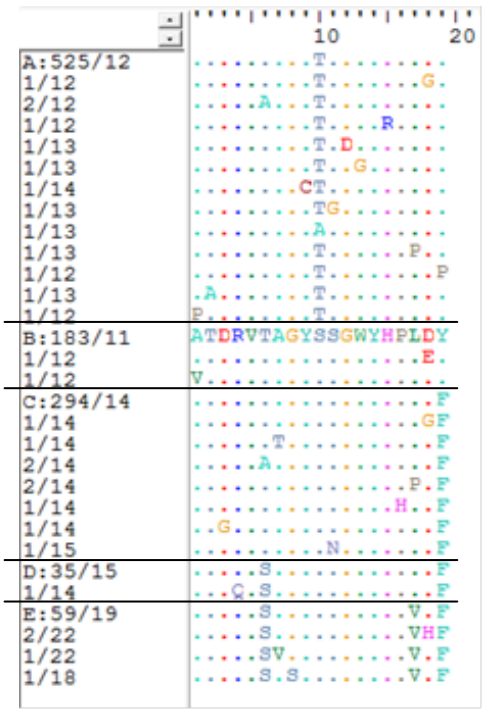
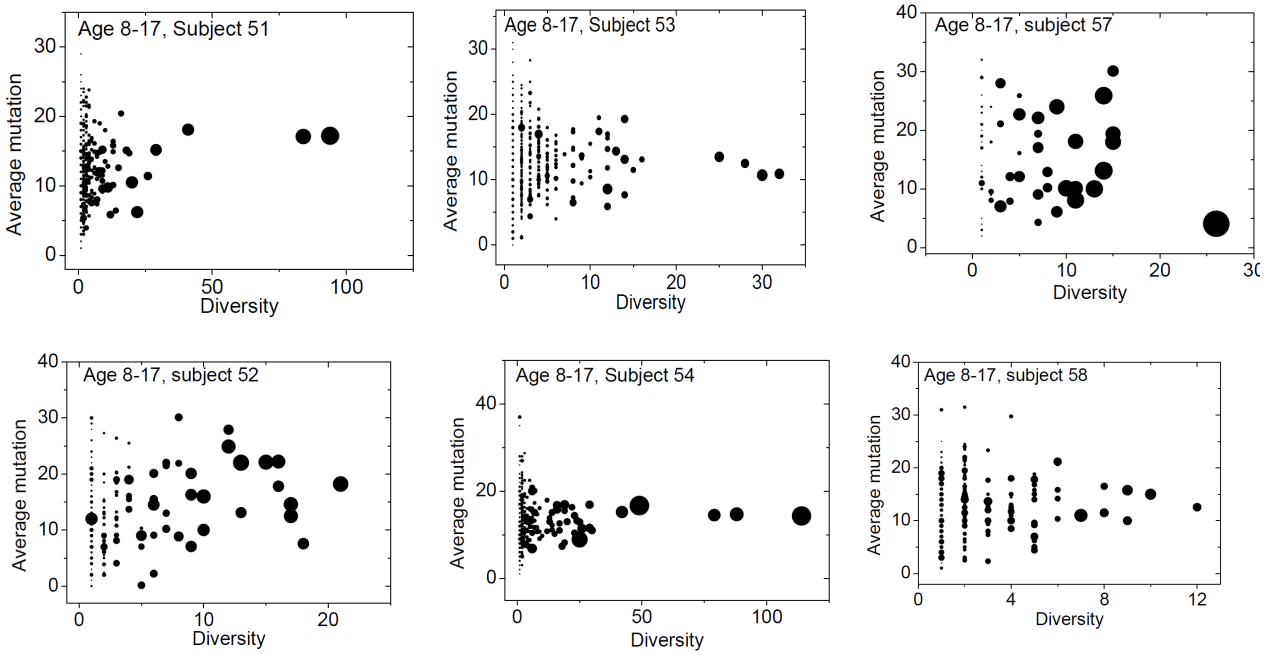
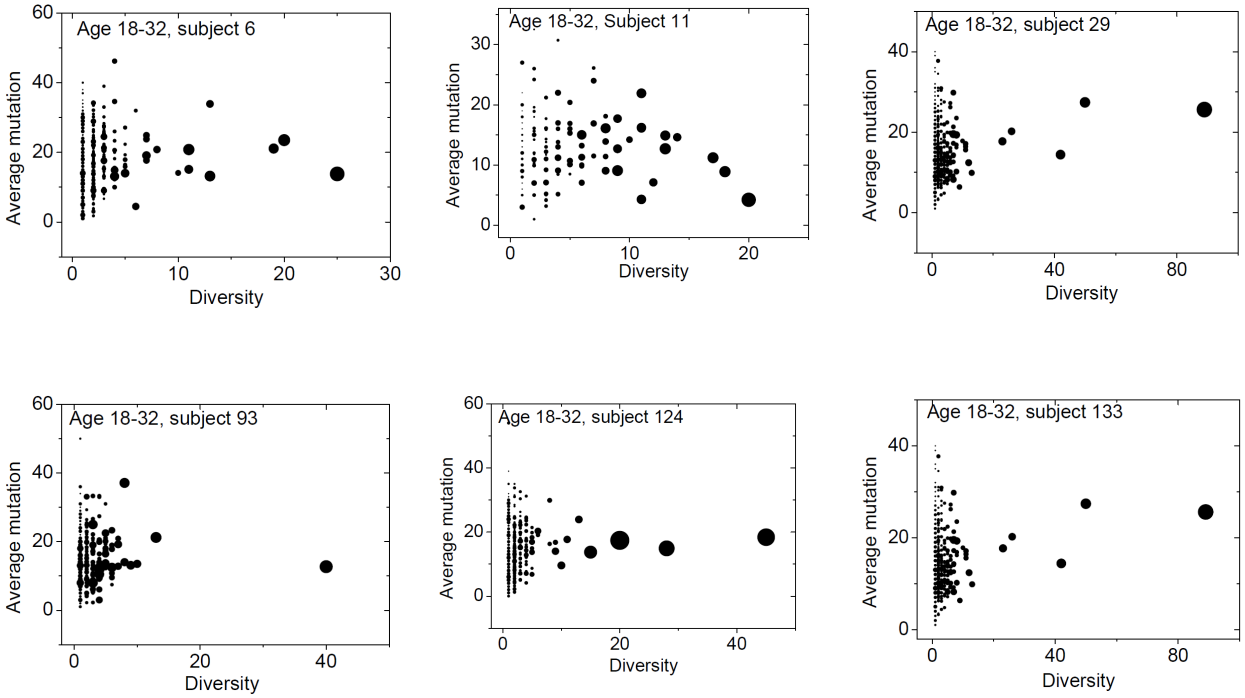


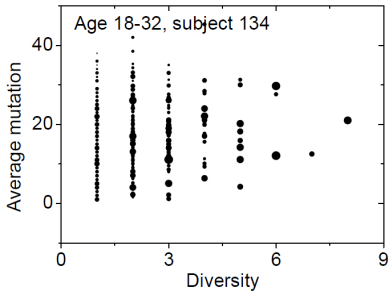
Fig. S7. The inter- and intra-lineage structure of all IgG lineages revealed by sequencing plasmablasts sorted from the visit 2 blood samples for selected subjects. Subject 53 (panel A, 9762 reads), subject 29 (panel B, 28712 reads) and subject 25 (panel C, 27079 reads).In this network representation, each cluster of dots connected by lines stands for a lineage. Different colors were used to distinguish different lineages. Each dot represents a unique CDR3 protein sequence. Two dots are linked if they differ by one amino acid in the CDR3 region as defined by the clustering threshold. The area of a dot is proportional to the number of reads with identical CDR3 protein sequences. Plots were generated using Pajek. (D), graphical presentation of an intra-lineage structure for the top lineage in panel. Arrows indicate direction of mutation increase. (E), AA sequences alignment of the CDR3. Lines separate sequences belong to each node as indicated by letter A, B, C and D in panel D. In the headers of the alignment, X/Y means that the sequence has X identical reads in AA, and the average mutation to germline is Y in nucleotide.

Age 8-17

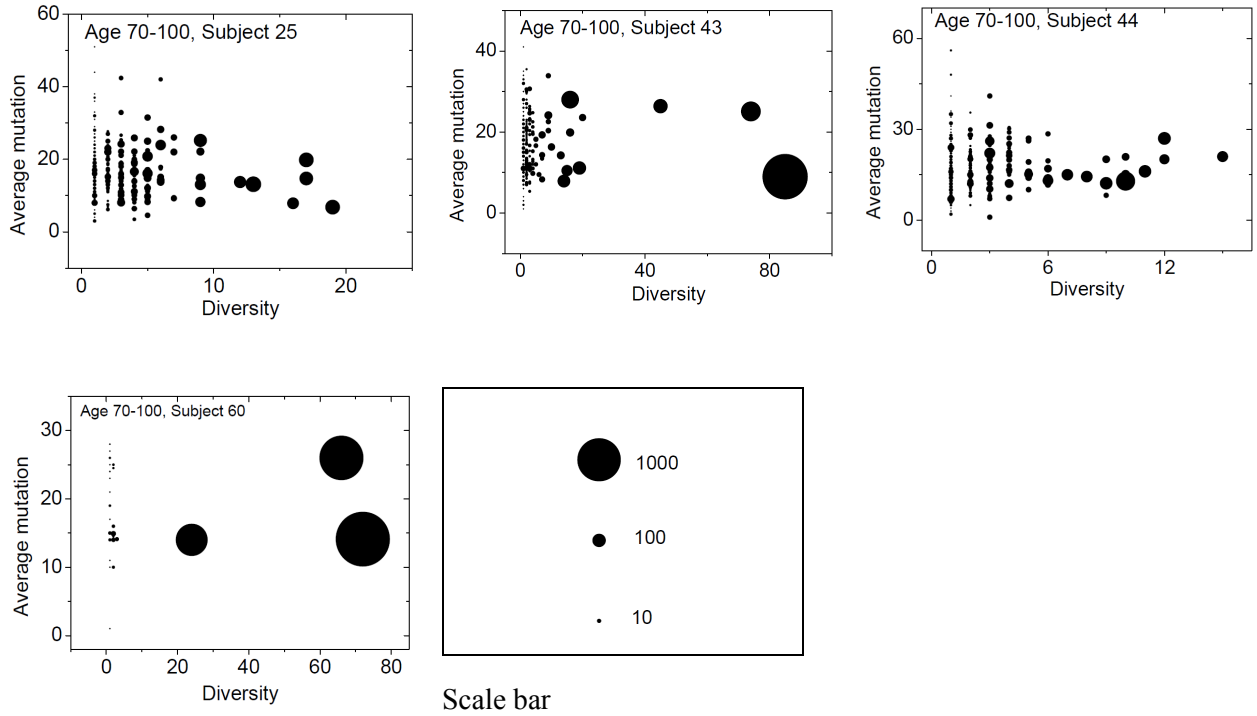


Age 18-32





Age 70-100



Scale bar

Fig. S8. The inter-lineage structure of IgG from plasmablasts sorted for all subjects at visit 2. Each panel represents one volunteer. In each panel, each dot represents a lineage of antibody sequences defined by single linkage clustering with 1 amino acid difference at CDR3 as the threshold. The area of the dot is proportional to the number of reads belongs to this lineage, as indicated in the scale bar in the last panel. X-axis is the diversity of the lineage which measures number of unique protein sequences (full protein sequence, not just the CDR3 region) within the lineage. Y-axis is the number of mutation at nucleotide level of the lineage averaged over reads. Subject 017-051 and subject 017-052, subject 017-053 and subject 017-054, and subject 017-057 and subject 017-058 are twin pairs. 3000 reads of subsampling was applied to this figure.

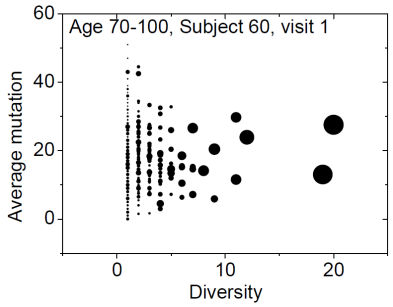


Fig. S9. The inter-lineage structure of IgG from PBMCs purified from subject 017-060 at visit 1. The plot was generated the same way as Supplementary Figure 5. This is in contrast to the lineage structure of plasmablasts sorted for the same subject at visit 2 (Supplementary Figure 5). The dominant lineage of subject 017-060 at visit 2 is not observed in visit 1. 3000 reads of subsampling was applied to this figure.

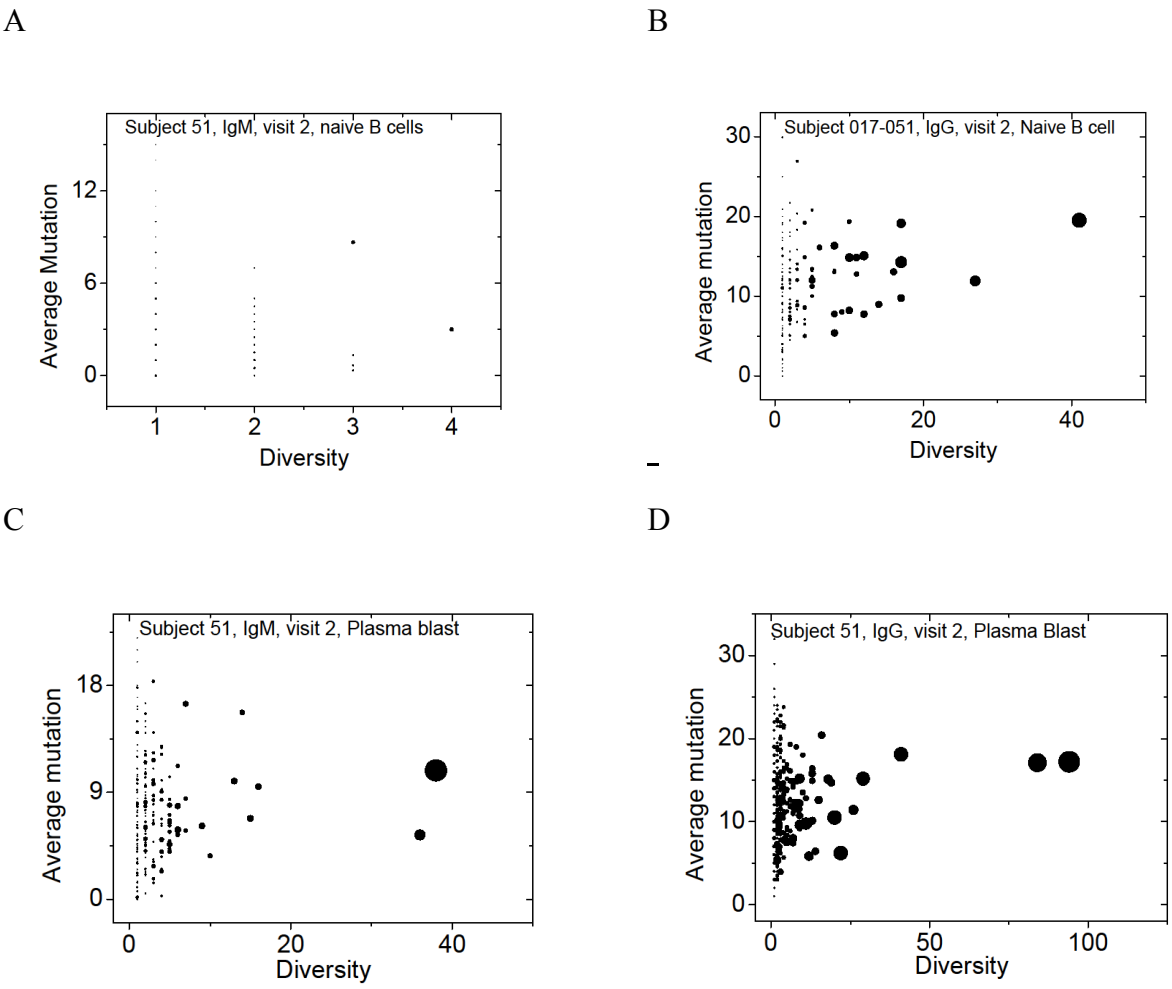
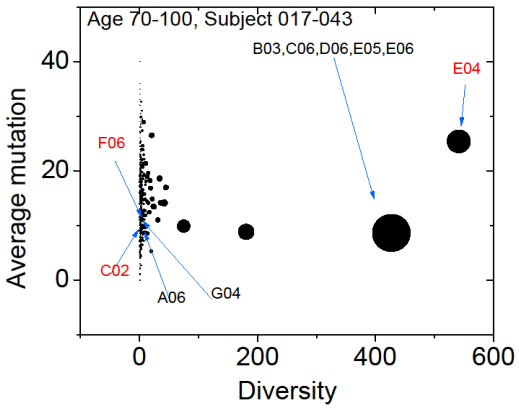
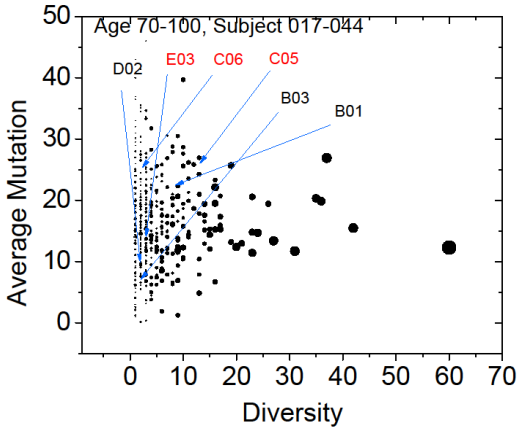


Fig. S10. The inter-lineage structure of IgM and IgG for naïve B cells and plasmablasts from one subject at visit 2. The plot was generated the same way as Supplementary Figure 5. Naïve B cells display minimum intra-lineage diversity (number of unique sequences within a lineage) compared to plasmablasts from the same subject. Most lineages in naïve B cell has only one read with no mutations, especially for the IgM isotype. IgM and IgG account for 60% and 21% of reads in subject 51's naïve B cells. IgM and IgG account for 10% and 63% of reads in subject 51's plasmablasts respectively. The average IgM in naïve B cells of all subjects is 75%. The average IgG in plasmablasts of all subjects is 63%. 3000 reads of subsampling was applied to this figure.

A



B



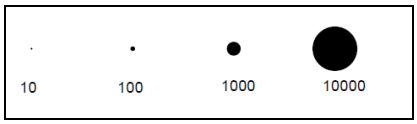


Fig. S11. Overlapping of single-cell cloned antibody sequences with lineages. Single cell cloning and high-throughput sequencing were performed on the same pool of sorted plasmablasts from two volunteers. Sequence overlapping between the two methods was demonstrated. The plot was generated the same way as Supplementary Figure 8. Single cell cloned antibodies are labeled with text. Red text indicates antibodies having a high affinity towards one of the virus strains used in the flu vaccine. Black text indicates antibodies with a low affinity towards one of the virus strains used in the flu vaccine or background level of binding towards all three virus strains. Antibodies G04, A06 in subject 017-043, and B03, C06, and D02 in subject 017-044 were not found in our high-throughput sequencing data. Please note that all reads are included in plotting these two figures, rather than using 3000 reads subsampling. To avoid the overlapping of dots, the size of the dots is scaled down4 fold from number of sequencing reads, see the scale bar at the bottom.

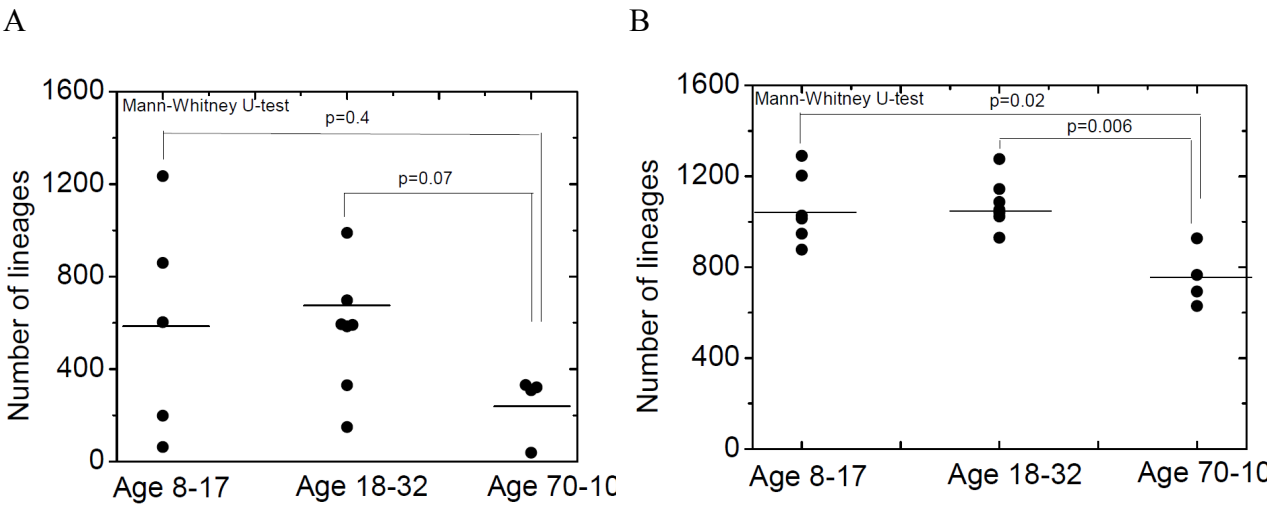


Fig. S12. Repertoire diversity changes with age. Repertoire diversity changes with age as measured by number of lineages in IgG from visit 2 plasmablasts (A) and IgG from visit 3 PBMCs (B). p value was calculated by Mann-Whitney U test. 3,000 reads subsampling was applied. There are 6 samples in the age group 8-17, 7 samples in the age group 18-32, 4 samples in the age group 70-100.

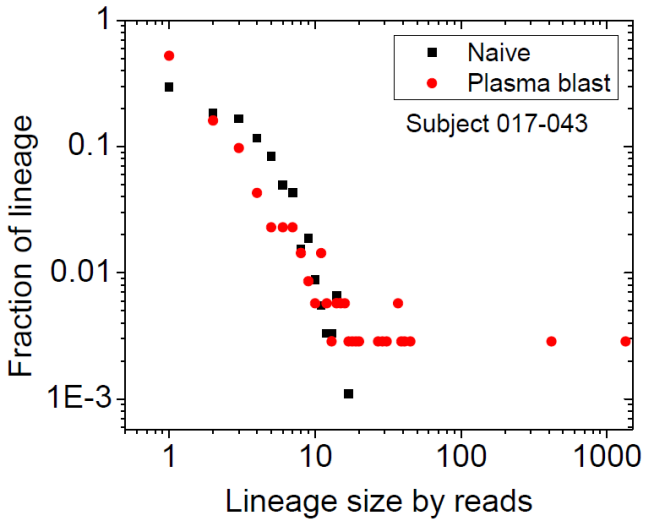


Fig. S13. Distribution of lineage size observes the power-law distribution. IgG lineages defined for the plasmablasts display a power-law distribution with a fat tail. The exponent of the power law is -1.7. The IgM lineage size distribution of naïve B cell does not exhibit this fat tail. 3000 reads of subsampling was applied to this figure.

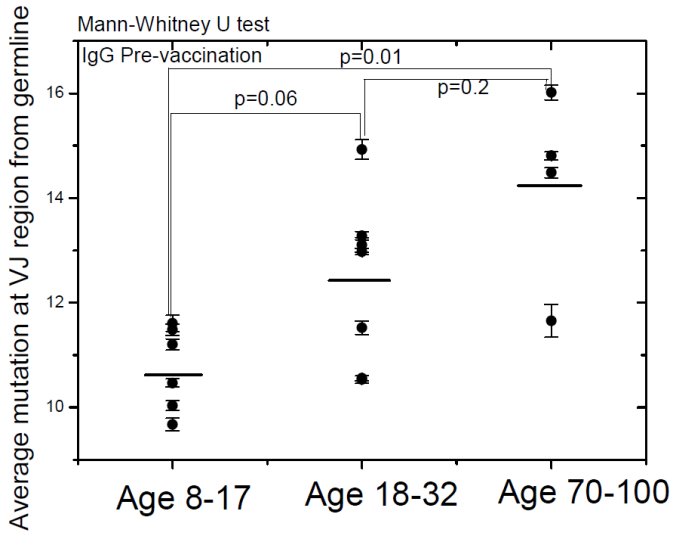


Fig. S14. Mutation pattern of IgG for three age groups in visit 1 PBMCs. The number of mutations for each read was defined as the number of mismatches to germline reference in V and J region. 3000 reads of subsampling was applied to this figure. All error bars are standard error.

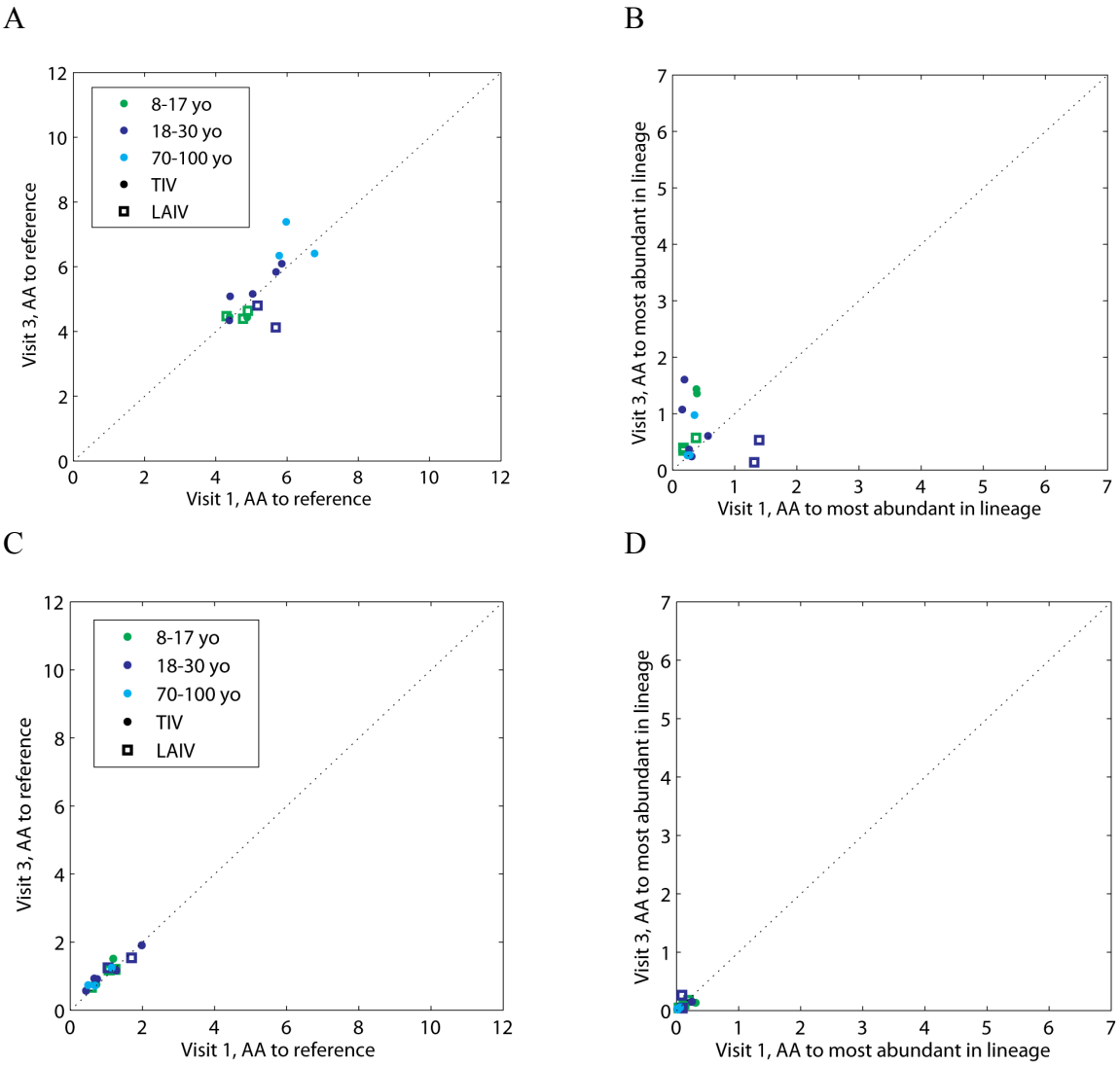


Fig. S15. The mutation patterns for different age groups at threshold of 90% of nucleotide similarity in the CDR3 region. This analysis was applied to IgG in the PBMCs at visit 1 (A, B) and IgM in the PBMCs at visit 1 (C, D). 3000 reads of subsampling was applied to this figure.

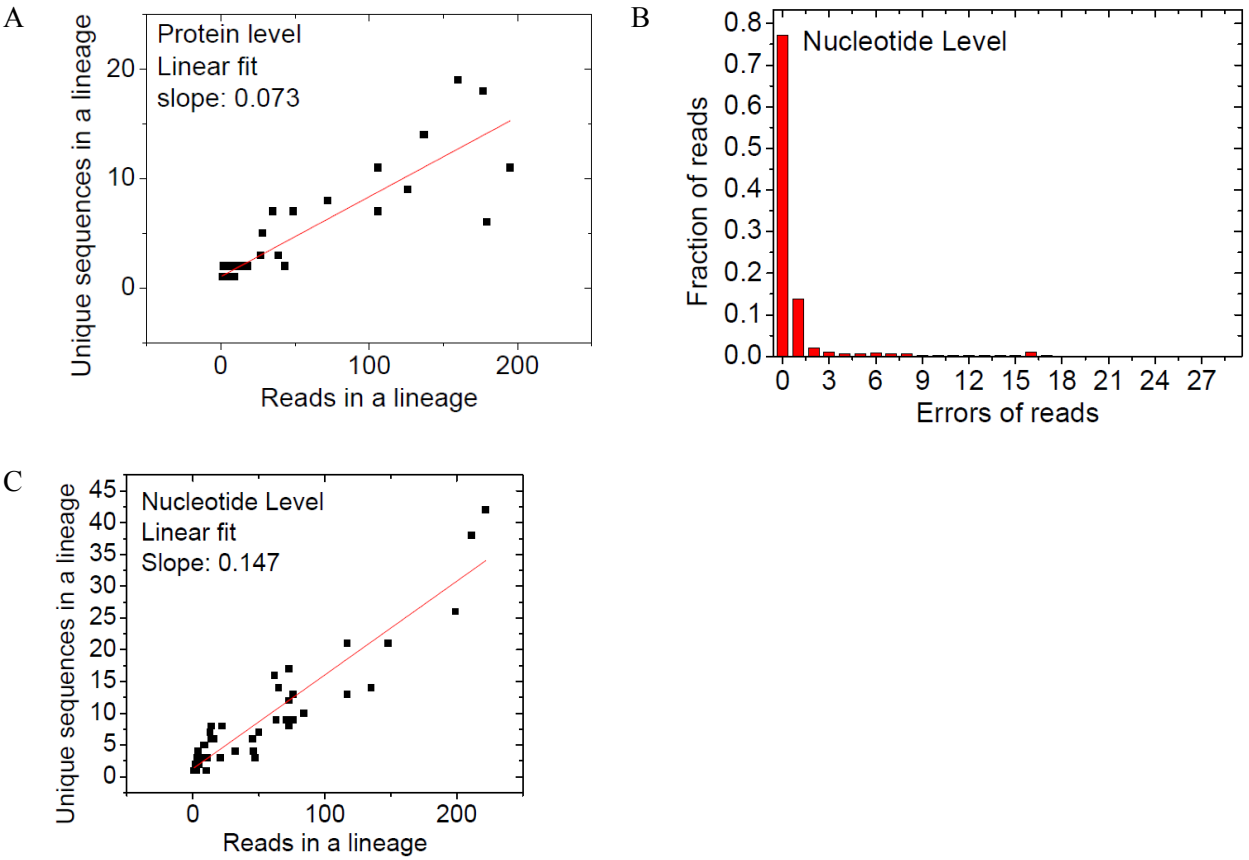


Fig. S16. Zebrafish control data. (A), relationship of number of reads and unique protein sequences in a cluster. Each data point is a cluster. The red line is the linear fit to the data. Clusters with more than 200 reads are excluded from this figure. (B), error rate profile of the control data at nucleotide level. The reference sequences are the most abundant sequence in each template. Indels are excluded (C), relationship of number of reads and unique nucleotide sequences in a cluster. Each data point is a cluster. The red line is the linear fit to the data.

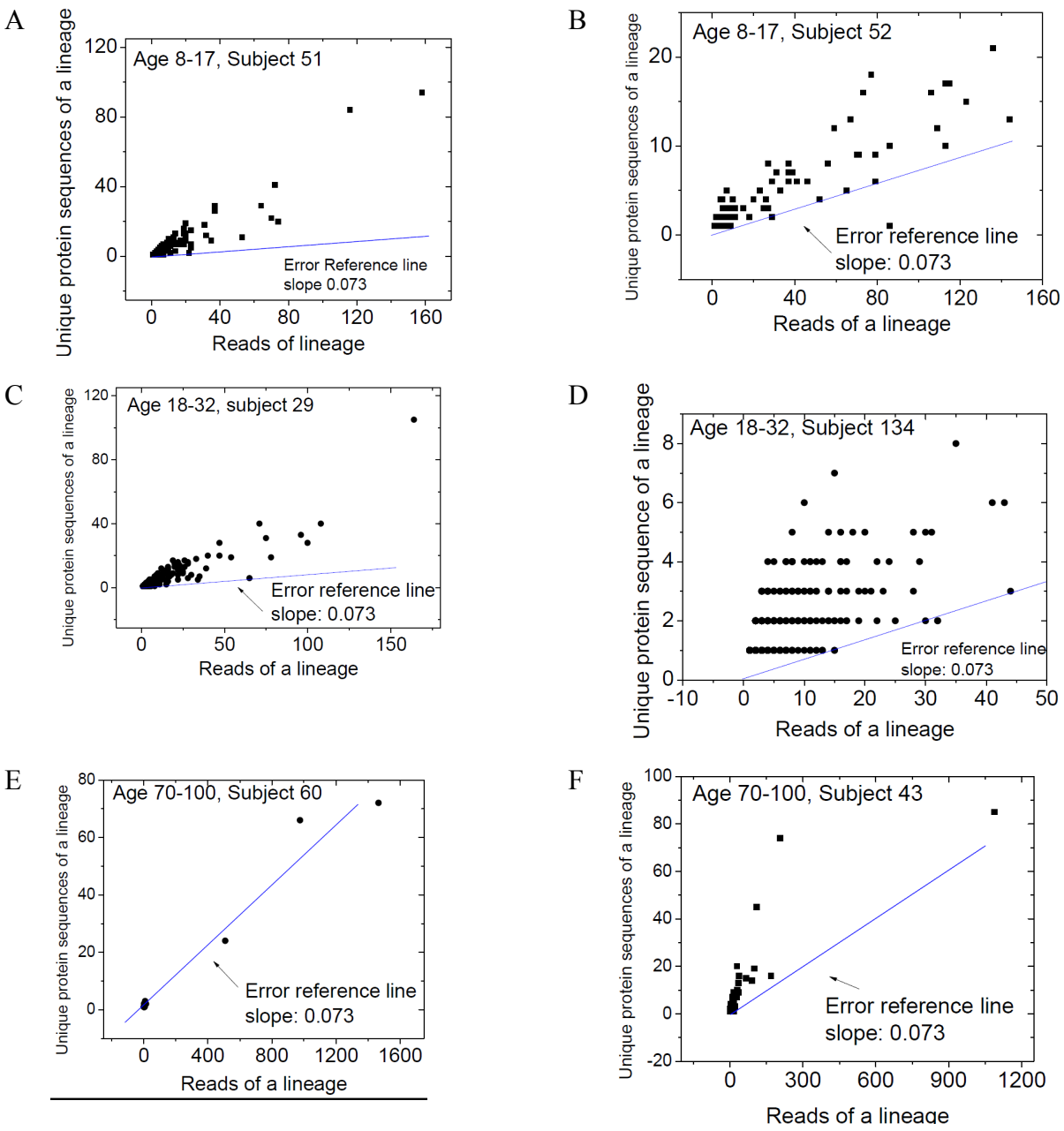


Fig. S17. Diversity and reads of IgG lineages of human plasmablasts at visit 2. Each dot in the figures is a lineage. X- and Y- axis are reads and unique protein sequences of the lineage respectively. The blue line is the expected diversity from zebrafish control data. The slope, 0.073, was estimated from Supplementary Figure 16A.

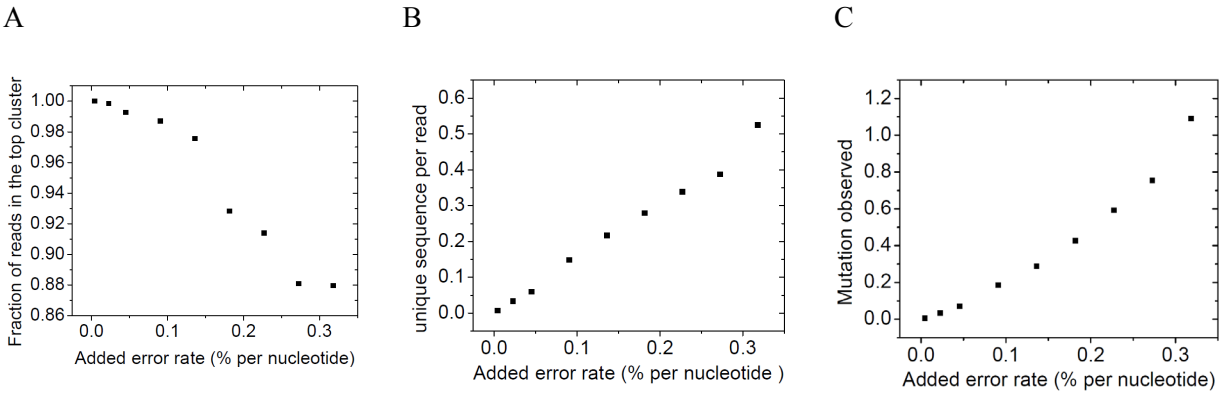


Fig. S18. Synthetic sequence control data. (A), the fraction of reads in the largest cluster is linearly proportional to the added error rate. (B), added errors increase the diversity. (C), added errors increase the mutation from baseline. Each data point is an independent trial.

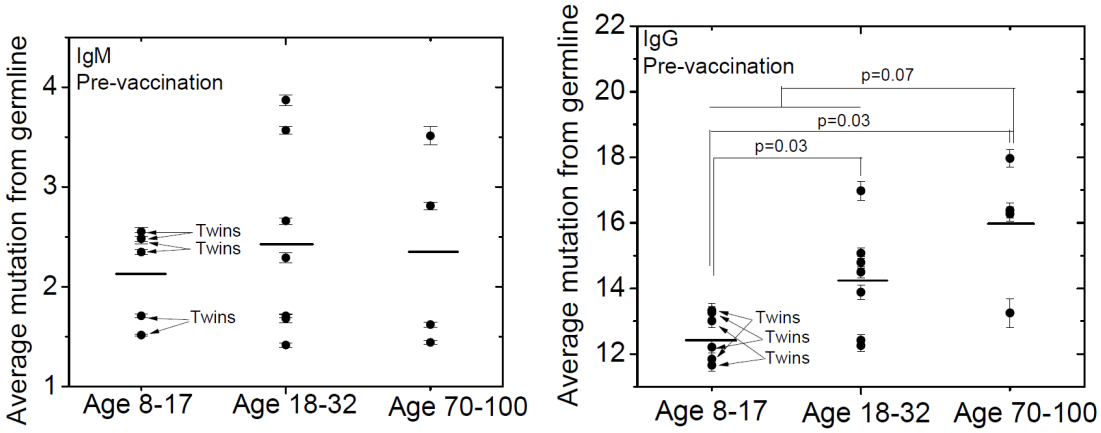


Fig. S19. Figures 3B and 3C, respectively, from the main text with twin status indicated by arrows. p value was calculated by Mann-Whitney U test.

Supplementary Tables

Table S1. Demographical information of human participants.

Patient ID	Age	Vaccine year	Vaccine	twin-status
017-006	18-30	2009	TIV	NA
017-011	18-30	2009	TIV	NA
017-029	18-30	2009	TIV	NA
017-025	70-100	2009	TIV	NA
017-043	70-100	2009	TIV	NA
017-044	70-100	2009	TIV	NA
017-060	70-100	2009	TIV	NA
017-051	8-17	2009	TIV	twin with 052
017-052	8-17	2009	LAIV	twin with 051
017-053	8-17	2009	LAIV	twin with 054
017-054	8-17	2009	TIV	twin with 053
017-057	8-17	2009	LAIV	twin with 058
017-058	8-17	2009	TIV	twin with 057
017-134	18-30	2010	LAIV	NA
017-124	18-30	2010	LAIV	NA
017-133	18-30	2010	TIV	NA
017-093	18-30	2010	TIV	NA

Table S2. Primer sequences.

PCR forward primers	
LR1	CGCAGACCCTCTCACTCAC
LR2	TGGAGCTGAGGTGAAGAAGC
LR3	TGCAATCTGGGTCTGAGTTG
LR4	GGCTCAGGACTGGTGAAGC
LR5	TGGAGCAGAGGTGAAAAAGC
LR6	GGTGCAGCTGTTGGAGTCT
LR7	ACTGTTGAAGCCTTCGGAGA
LR8	AAACCCACACAGACCCTCAC
LR9	AGTCTGGGGCTGAGGTGAAG
LR10	GGCCCAGGACTGGTGAAG
LR11	GGTGCAGCTGGTGGAGTC
PCR reverse primers	
IgG-PCR	AAGACCGATGGGCCCTTG
IgA-PCR	GAAGACCTTGGGGCTGGT
IgM-PCR	GGGAATTCTCACAGGAGACG
IgE-PCR	GAAGACGGATGGGCTCTGT
IgD-PCR	GGGTGTCTGCACCCTGATA
Reverse transcription primers	
IgG-RT	GGGAAGTAGTCCTTGACCAG
IgA-RT	GGGGAAGAAGCCCTGGAC
IgM-RT	GGCCACGCTGCTCGTATC
IgE-RT	AGGGAATGTTTTTGCAGCAG
IgD-RT	CCACAGGGCTGTTATCCTTT

Table S3. Summary of cell numbers and filtered reads for all samples.

Subject ID	Visit 1 PBMC (reads)	Visit 2 Naïve B cells (reads)	Visit2 Naïve B cell #	Visit 2 Plasmablasts (reads)	Visit 2 Plasmablasts #	Visit 3 PBMC (reads)	Visit 2 PBMC (reads)
017-006	18370	251140	88576	25356	12557	23743	N/A
017-011	112250	59494	10467	51575	1847	40451	N/A
017-029	29134	81014	39502	82336	46135	288484	N/A
017-025	66900	42305	8453	83360	3100	84739	N/A
017-043	3949	66050	66472	25290	33455	35149	N/A
017-044	31467	18638	23974	24492	8402	26262	N/A
017-060	33560	26538	29361	14421	3346	26769	N/A
017-051	51442	61240	189294	63736	156019	100499	N/A
017-054	62117	51034	61915	98952	78264	51243	N/A
017-052	95625	57756	86783	56080	6501	64980	N/A
017-053	35488	40826	50815	27124	126634	28840	N/A
017-058	10926	11869	29332	15956	10758	20180	N/A
017-057	28584	25628	28313	28544	13648	23884	N/A
017-093	34364	91	100000	15574	5000	32417	28700
017-134	45987	73627	118000	113360	8600	32296	59247
017-124	16127	35672	124000	13040	27000	27887	27008
017-133	33547	42297	28000	13183	36000	29806	26531
Average	41755	55601		44257		55154	32506

Table S4. Raw reads for five isotypes in each sample. Reads from a subset of runs were used.

subject			IgM	IgA	IgG	IgD	IgE	Isotype identifiable
52	visit 2	Plasmablasts	6862	21998	26422	163	635	56080
53	visit 2	Plasmablasts	3023	5446	12349	3	148	20969
51	visit 2	Plasmablasts	6859	16742	39813	50	272	63736
54	visit 2	Plasmablasts	18005	20555	60055	5	332	98952
57	visit 2	Plasmablasts	2280	12312	13937	15	0	28544
58	visit 2	Plasmablasts	1507	2902	11539	8	0	15956
6	visit 2	Plasmablasts	2687	7749	14890	0	30	25356
11	visit 2	Plasmablasts	4591	17676	28546	499	263	51575
29	visit 2	Plasmablasts	24728	15764	41613	4	227	82336
134	visit 2	Plasmablasts	19983	44492	48333	72	480	113360
93	visit 2	Plasmablasts	2553	4356	8401	58	206	15574
124	visit 2	Plasmablasts	1124	6702	5186	18	10	13040
133	visit 2	Plasmablasts	331	2221	10620	0	11	13183
25	visit 2	Plasmablasts	6343	31777	43851	784	605	83360
43	visit 2	Plasmablasts	1535	126	23615	3	11	25290
44	visit 2	Plasmablasts	1688	526	22268	10	0	24492
60	visit 2	Plasmablasts	150	8244	6022	2	3	14421
6	visit 1	PBMC	10235	4389	3324	415	7	18370
11	visit 1	PBMC	60684	26467	22410	2600	89	112250
25	visit 1	PBMC	39665	12934	10734	3557	10	66900
29	visit 1	PBMC	17189	7875	3440	501	129	29134
43	visit 1	PBMC	2999	70	583	296	1	3949
44	visit 1	PBMC	19177	401	9972	1859	58	31467
52	visit 1	PBMC	65497	15315	10016	4746	51	95625
53	visit 1	PBMC	11456	4443	1558	1177	4	18638
60	visit 1	PBMC	22447	5302	5001	791	19	33560
57	visit 1	PBMC	14223	8256	5744	355	6	28584
93	visit 1	PBMC	17092	6199	10182	875	16	34364
133	visit 1	PBMC	14570	6350	11402	1187	38	33547
134	visit 1	PBMC	15961	14550	14732	725	19	45987
124	visit 1	PBMC	6522	5616	3436	525	28	16127
58	visit 1	PBMC	5652	2315	2714	242	3	10926
51	visit 1	PBMC	34254	7897	6680	2163	448	51442
54	visit 1	PBMC	33666	19360	7492	1562	37	62117
6	visit 3	PBMC	8961	7204	7310	233	35	23743
11	visit 3	PBMC	20428	10671	7870	1465	17	40451
25	visit 3	PBMC	45215	15842	18957	4653	72	84739
29	visit 3	PBMC	138619	65065	74933	9350	517	288484

43	visit 3	PBMC	25903	318	6981	1939	8	35149
44	visit 3	PBMC	12498	327	12349	990	98	26262
52	visit 3	PBMC	40503	11625	9433	2625	794	64980
53	visit 3	PBMC	11589	5926	2046	1193	10	20764
60	visit 3	PBMC	17947	4098	3890	817	17	26769
57	visit 3	PBMC	11575	6951	5058	280	20	23884
93	visit 3	PBMC	13571	6890	11177	762	17	32417
133	visit 3	PBMC	11291	7389	10296	813	17	29806
134	visit 3	PBMC	11613	9220	11120	336	7	32296
124	visit 3	PBMC	10979	8329	6016	2542	21	27887
58	visit 3	PBMC	7098	6657	5850	567	8	20180
51	visit 3	PBMC	45315	34524	16452	4022	186	100499
54	visit 3	PBMC	19909	22037	7655	1496	146	51243
134	visit 2	Naïve B cells	65372	2310	4150	1633	162	73627
93	visit 2	Naïve B cells	90	0	1	0	0	91
124	visit 2	Naïve B cells	32776	737	1328	826	5	35672
133	visit 2	Naïve B cells	37141	1247	2919	952	38	42297
6	visit 2	Naïve B cells	209550	12793	24457	3722	618	251140
11	visit 2	Naïve B cells	48579	1155	2884	5956	920	59494
29	visit 2	Naïve B cells	57741	6148	9659	6028	1438	81014
25	visit 2	Naïve B cells	31518	1885	3631	3199	2072	42305
43	visit 2	Naïve B cells	60009	47	850	5117	27	66050
44	visit 2	Naïve B cells	13576	154	3149	1739	20	18638
60	visit 2	Naïve B cells	24717	91	1129	576	25	26538
51	visit 2	Naïve B cells	35896	5872	13164	4887	1421	61240
52	visit 2	Naïve B cells	46047	2106	3220	5584	799	57756
53	visit 2	Naïve B cells	16385	2127	2410	1208	38	22168
54	visit 2	Naïve B cells	35615	2039	8639	3917	824	51034
57	visit 2	Naïve B cells	23952	455	549	594	78	25628
58	visit 2	Naïve B cells	10281	219	698	665	6	11869

Table S5. Summary of identifiable VJ reads for IgG in visit 2 plasmablasts.

Subject ID	Identifiable IgG VJ reads	Raw IgGreads	Fraction of identifiable VJ reds
51	39389	39813	0.98935
54	59465	60055	0.990176
53	21127	21367	0.988768
58	11528	11539	0.999047
57	13870	13937	0.995193
29	41165	41613	0.989234
6	14772	14890	0.992075
25	42072	43851	0.959431
43	23608	23615	0.999704
44	22222	22268	0.997934
60	6010	6022	0.998007
11	27883	28546	0.976774
52	24693	26422	0.934562
134	48024	48333	0.993607
93	8376	8401	0.997024
124	5173	5186	0.997493
133	10605	10620	0.998588

Table S6. Summary of single cell cloned sequences.

subject 017-043								
antibody	identical AA sequences in visit 2 PB (excluding itself)	unique AA sequences in the cluster (excluding itself) in visit 2 PB	reads of the SCC containing cluster (including itself)in visit 2 PB	Mutation (nt) of SCC antibody	Minimum mutation (nt) of the SCC containing cluster in visit 2 PB	Maximum mutation (nt) of the SCC containing cluster in visit 2 PB	SCC sequence Binding to antigen?	Unique AA sequences in the cluster (excluding itself) in visit 3PBMC
A06	0	0	1	8	8	8	N	0
B03	3764	427	6884	7	3	15	N	1
C02	0	4	38	9	9	12	Y	1
C06	43	427	6884	8	3	15	N	1
D06	5	427	6884	13	3	15	N	1
E04	0	541	2599	39	14	40	Y	0
E05	207	427	6884	8	3	15	N	1
E06	70	427	6884	7	3	15	N	1
F06	1	7	59	14	14	18	Y	17
G04	0	0	1	12	12	12	N	0

subject 017-044								
antibody	identical AA sequences in visit 2 PB (excluding itself)	unique AA sequences in the cluster (excluding itself) in visit 2 PB	reads of the SCC containing cluster (including itself)in visit 2 PB	Mutation (nt) of SCC antibody	Minimum mutation (nt) of the SCC containing cluster in visit 2 PB	Maximum mutation (nt) of the SCC containing cluster in visit 2 PB	SCC sequence Binding to antigen?	Unique AA sequences in the cluster (excluding itself) in

								visit 3 PBMC
B01	0	5	22	22	18	29	N	0
B03	0	0	1	7	7	7	N	0
C05	0	11	103	24	24	28	Y	0
C06	0	0	1	25	25	25	Y	0
D02	0	0	1	10	10	10	N	0
E03	0	3	31	13	11	16	Y	0

Table S7. Control data information.

Zebrafish control library Roche 454 sequencing reads information	
Number of sequence template	38
Total reads in analysis	32453
Total unique sequences	1776
Total reads in top 38 sequences (percentage)	28194 (87%)