Supplementary Information (SI) for

Toward a more accurate view of human B-cell repertoire by next-generation sequencing, unbiased repertoire capture and single-molecule barcoding

Linling He[1], Devin Sok[1], Parisa Azadnia[1], Jessica Hsueh[1], Elise Landais[2], Melissa Simek[3], Wayne C. Koff[3], Pascal Poignard[1,2,3], Dennis Burton[1-5] & Jiang Zhu[1,4,6*]

[1] Department of Immunology and Microbial Science, The Scripps Research Institute, La Jolla, California 92037, USA.
[2] IAVI Neutralizing Antibody Center, The Scripps Research Institute, La Jolla, California 92037, USA.
[3] International AIDS Vaccine Initiative (IAVI), New York, NY 10004, USA.
[4] Center for HIV/AIDS Vaccine Immunology and Immunogen Discovery, The Scripps Research Institute, La Jolla, California 92037, USA.
[5] Ragon Institute of Massachusetts General Hospital, Massachusetts Institute of Technology, and Harvard, Cambridge, MA 02139-3583, USA.
[6] Department of Integrative Structural and Computational Biology, The Scripps Research Institute, La Jolla, California 92037, USA.

[*] Correspondence and requests for materials should be addressed to:
JZ: Phone (858)784-8157; Email- jiang@scripps.edu

This file includes:

Supplementary Tables 1-5

Supplementary Figures 1-6

Table S1. PCR primers used to prepare samples for PGM sequencing (forward direction). [a]

Primer name	Primer sequence (5' → 3')
1. Heavy chain	
5'-end primer with PGM A sequencing adapter	
VRC5'-A-VH1	*CCATCTCATCCCTGCGTGTCTCCGACTCAG* ACAGGTGCCCACTCCCAGGTGCAG
VRC5'-A-VH1#2	*CCATCTCATCCCTGCGTGTCTCCGACTCAG* GCAGCCACAGGTGCCCACTCC
VRC5'-A-VH1-24	*CCATCTCATCCCTGCGTGTCTCCGACTCAG* CAGCAGCTACAGGCACCCACGC
VRC5'-A-VH1-69	*CCATCTCATCCCTGCGTGTCTCCGACTCAG* GGCAGCAGCTACAGGTGTCCAGTCC
VRC5'-A-VH4/6[d]	*CCATCTCATCCCTGCGTGTCTCCGACTCAG* CCCAGATGGGTCCTGTCCCAGGTGCAG
VRC5'-A-VH3/4 #1[d]	*CCATCTCATCCCTGCGTGTCTCCGACTCAG* GTGGCAGCTCCCAGATGGGTCCTGTC
VRC5'-A-VH3/4 #3[d]	*CCATCTCATCCCTGCGTGTCTCCGACTCAG* GTTGCAGTTTTAAAAGGTGTCCAGTG
VRC5'-A-VH5#1[b]	*CCATCTCATCCCTGCGTGTCTCCGACTCAG* GAGTCTGGTGCCGAGGTGCAG
VRC5'-A-VH5#2[b]	*CCATCTCATCCCTGCGTGTCTCCGACTCAG* GGAGTCTGTGCCGAGGTGCAG
3'-end primer with PGM trP1 sequencing adapter	
VRC3'-P1-CγCH1	*CCTCTCTATGGGCAGTCGGTGAT* GGGGAAGACCGATGGGCCCTTGGTGG
VRC3'-P1-CμCH1[e]	*CCTCTCTATGGGCAGTCGGTGAT* GGGAATTCTCACAGGAGACGA
2. Lambda chain	
5'-end primer with PGM A sequencing adapter	
VRC5'-A-VL1/2	*CCATCTCATCCCTGCGTGTCTCCGACTCAG* GCACAGGGTCCTGGGCCCAGTCTG
VRC5'-A-VL3[d]	*CCATCTCATCCCTGCGTGTCTCCGACTCAG* GCTCTGTGACCTCCTATGAGCTG
VRC5'-A-VL4/5	*CCATCTCATCCCTGCGTGTCTCCGACTCAG* GGTCTCTCTCSCAGCYTGTGCTG
VRC5'-A-VL6	*CCATCTCATCCCTGCGTGTCTCCGACTCAG* GTTCTTGGGCCAATTTTATGCTG
VRC5'-A-VL7/8	*CCATCTCATCCCTGCGTGTCTCCGACTCAG* GAGTGGATTCTCAGACTGTGGTG
VRC5'-A-VL1#2	*CCATCTCATCCCTGCGTGTCTCCGACTCAG* GCTCACTGCACAGGGTCCTGGGCC
VRC5'-A-VL3-1[d]	*CCATCTCATCCCTGCGTGTCTCCGACTCAG* GCTTACTGCACAGGATCCGTGGCC
VRC5'-A-VL3-19[d]	*CCATCTCATCCCTGCGTGTCTCCGACTCAG* ACTCTTTGCATAGGTTCTGTGGTT
VRC5'-A-VL3-21[d]	*CCATCTCATCCCTGCGTGTCTCCGACTCAG* TCTCACTGCACAGGCTCTGTGACC
VRC5'-A-VL7-43	*CCATCTCATCCCTGCGTGTCTCCGACTCAG* ACTTGCTGCCCAGGGTCCAATTC
3'-end primer with PGM trP1 sequencing adapter	
VRC3'-P1-Cλ	*CCTCTCTATGGGCAGTCGGTGAT* CACCAGTGTGGCCTTGTTGGCTTG
3. Kappa chain [c]	
5'-end primer with PGM A sequencing adapter	
VRC5'-A-VK1/2	*CCATCTCATCCCTGCGTGTCTCCGACTCAG* ATGAGGSTCCCYGCTCAGCTCCTGGG
VRC5'-A-VK3	*CCATCTCATCCCTGCGTGTCTCCGACTCAG* CTCTTCCTCCTGCTACTCTGGCTCCCAG
VRC5'-A-VK4	*CCATCTCATCCCTGCGTGTCTCCGACTCAG* ATTTCTCTGTTGCTCTGGATCTCTG
3'-end primer with PGM trP1 sequencing adapter	
VRC3'-P1-Cκ	*CCTCTCTATGGGCAGTCGGTGAT* CAGCAGGCACACAACAGAGGCAGTTCC

[a] The forward PGM sequencing protocol (from the beginning of the V gene segment to the end of the J gene segment) was devised in a similar manner to the 454 pyrosequencing protocol described in Refs (11-15). The PGM sequencing adapters are underscored in all the primer sequences.

[b] The two VH5 primers were designed in a similar manner to those described in Refs (11-15).

[c] The kappa chain primers, although not used in current study, are included to complete the primer set.

[d] In this study, we used three VH4-specific primers and four VL3-specific primers to capture the antibodies related to the PGT121 class from donor IAVI 17.

[e] Only Cγ primer was used in current study to capture memory and plasma B-cell transcripts in a donor sample, Cμ primer is provided for completeness.

Table S2. PCR primers used to prepare samples for PGM sequencing (reversed direction). [a]

Primer name	Primer sequence (5' → 3')
1. Heavy chain	
5'-end primer with trP1 sequencing adapter	
VRC5'-P1-VH1	*CCTCTCTATGGGCAGTCGGTGAT* ACAGGTGCCCACTCCCAGGTGCAG
VRC5'-P1-VH1#2	*CCTCTCTATGGGCAGTCGGTGAT* GCAGCCACAGGTGCCCACTCC
VRC5'-P1-VH1-24	*CCTCTCTATGGGCAGTCGGTGAT* CAGCAGCTACAGGCACCCACGC
VRC5'-P1-VH1-69	*CCTCTCTATGGGCAGTCGGTGAT* GGCAGCAGCTACAGGTGTCCAGTCC
VRC5'-P1-VH4/6	*CCTCTCTATGGGCAGTCGGTGAT* CCCAGATGGGTCCTGTCCCAGGTGCAG
VRC5'-P1-VH3/4 #1	*CCTCTCTATGGGCAGTCGGTGAT* GTGGCAGCTCCCAGATGGGTCCTGTC
VRC5'-P1-VH3/4 #3	*CCTCTCTATGGGCAGTCGGTGAT* GTTGCAGTTTTAAAAGGTGTCCAGTG
VRC5'-P1-VH5#1[b]	*CCTCTCTATGGGCAGTCGGTGAT* GAGTCTGGTGCCGAGGTGCAG
VRC5'-P1-VH5#2[b]	*CCTCTCTATGGGCAGTCGGTGAT* GGAGTCTGTGCCGAGGTGCAG
3'-end primer with A sequencing adapter	
VRC3'-A-CγCH1	*CCATCTCATCCCTGCGTGTCTCCGACTCAG* GGGGAAGACCGATGGGCCCTTGGTGG
VRC3'-A-CμCH1[d]	*CCATCTCATCCCTGCGTGTCTCCGACTCAG* GGGAATTCTCACAGGAGACGA
2. Lambda chain	
5'-end primer with trP1 sequencing adapter	
VRC5'-P1-VL1/2	*CCTCTCTATGGGCAGTCGGTGAT* GCACAGGGTCCTGGGCCCAGTCTG
VRC5'-P1-VL3	*CCTCTCTATGGGCAGTCGGTGAT* GCTCTGTGACCTCCTATGAGCTG
VRC5'-P1-VL4/5	*CCTCTCTATGGGCAGTCGGTGAT* GGTCTCTCTCSCAGCYTGTGCTG
VRC5'-P1-VL6	*CCTCTCTATGGGCAGTCGGTGAT* GTTCTTGGGCCAATTTTATGCTG
VRC5'-P1-VL7/8	*CCTCTCTATGGGCAGTCGGTGAT* GAGTGGATTCTCAGACTGTGGTG
VRC5'-P1-VL1#2	*CCTCTCTATGGGCAGTCGGTGAT* GCTCACTGCACAGGGTCCTGGGCC
VRC5'-P1-VL3-1	*CCTCTCTATGGGCAGTCGGTGAT* GCTTACTGCACAGGATCCGTGGCC
VRC5'-P1-VL3-19	*CCTCTCTATGGGCAGTCGGTGAT* ACTCTTTGCATAGGTTCTGTGGTT
VRC5'-P1-VL3-21	*CCTCTCTATGGGCAGTCGGTGAT* TCTCACTGCACAGGCTCTGTGACC
VRC5'-P1-VL7-43	*CCTCTCTATGGGCAGTCGGTGAT* ACTTGCTGCCCAGGGTCCAATTC
3'-end primer with A sequencing adapter	
VRC3'-A-Cλ	*CCATCTCATCCCTGCGTGTCTCCGACTCAG* CACCAGTGTGGCCTTGTTGGCTTG
3. Kappa chain [c]	
5'-end primer with trP1 sequencing adapter	
VRC5'-P1-VK1/2	*CCTCTCTATGGGCAGTCGGTGAT* ATGAGGSTCCCYGCTCAGCTCCTGGG
VRC5'-P1-VK3	*CCTCTCTATGGGCAGTCGGTGAT* CTCTTCCTCCTGCTACTCTGGCTCCCAG
VRC5'-P1-VK4	*CCTCTCTATGGGCAGTCGGTGAT* ATTTCTCTGTTGCTCTGGATCTCTG
3'-end primer with A sequencing adapter	
VRC3'-A-Cκ	*CCATCTCATCCCTGCGTGTCTCCGACTCAG* CAGCAGGCACACAACAGAGGCAGTTCC

[a] The reversed PGM sequencing protocol (from the end of the J gene segment to the beginning of the V gene segment) was devised to minimize sequencing errors, especially homopolymer and indel errors, in the complementarity determining region 3 (CDR3) of the heavy and light chains. The PGM sequencing adapters are underscored in all the primer sequences. Note that in the random barcoding strategy, a stretch of 10 degenerate nucleotides (N_{10}) is inserted between the 3'-end primer and PGM A sequencing adaptor.

[b] The two VH5 primers were designed in a similar manner to those described in Refs (11-15).

[c] The kappa chain primers, although not used in current study, are included to complete the primer set.

[d] Only Cγ primer was used in current study to capture memory and plasma B-cell transcripts in a donor sample, Cμ primer is provided for completeness.

Table S3. Further upstream heavy chain PCR primers used to prepare samples for PGM sequencing (reversed direction). [a]

Primer	Primer sequence (5' → 3')
5'-end primer with trP1 sequencing adapter	
MCN5'-P1-VH1-LEADER-A	*CCTCTCTATGGGCAGTCGGTGAT* ATGGACTGGACCTGGAGGAT
MCN5'-P1-VH1-LEADER-B	*CCTCTCTATGGGCAGTCGGTGAT* ATGGACTGGACCTGGAGCAT
MCN5'-P1-VH1-LEADER-C	*CCTCTCTATGGGCAGTCGGTGAT* ATGGACTGGACCTGGAGAAT
MCN5'-P1-VH1-LEADER-D	*CCTCTCTATGGGCAGTCGGTGAT* GGTTCCTCTTTGTGGTGGC
MCN5'-P1-VH1-LEADER-E	*CCTCTCTATGGGCAGTCGGTGAT* ATGGACTGGACCTGGAGGGT
MCN5'-P1-VH1 LEADER-F	*CCTCTCTATGGGCAGTCGGTGAT* ATGGACTGGATTTGGAGGAT
MCN5'-P1-VH1-LEADER-G	*CCTCTCTATGGGCAGTCGGTGAT* AGGTTCCTCTTTGTGGTGGCAG
MCN5'-P1-VH3-LEADER-A	*CCTCTCTATGGGCAGTCGGTGAT* TAAAAGGTGTCCAGTGT
MCN5'-P1-VH3-LEADER-B	*CCTCTCTATGGGCAGTCGGTGAT* TAAGAGGTGTCCAGTGT
MCN5'-P1-VH3-LEADER-C	*CCTCTCTATGGGCAGTCGGTGAT* TAGAAGGTGTCCAGTGT
MCN5'-P1-VH3-LEADER-D	*CCTCTCTATGGGCAGTCGGTGAT* GCTATTTTTAAAGGTGTCCAGTGT
MCN5'-P1-VH3-LEADER-E	*CCTCTCTATGGGCAGTCGGTGAT* TACAAGGTGTCCAGTGT
MCN5'-P1-VH3-LEADER-F	*CCTCTCTATGGGCAGTCGGTGAT* TTAAAGCTGTCCAGTGT
MCN5'-P1-VH4-LEADER-A	*CCTCTCTATGGGCAGTCGGTGAT* ATGAAACACCTGTGGTTCTTCC
MCN5'-P1-VH4-LEADER-B	*CCTCTCTATGGGCAGTCGGTGAT* ATGAAACACCTGTGGTTCTT
MCN5'-P1-VH4-LEADER-C	*CCTCTCTATGGGCAGTCGGTGAT* ATGAAGCACCTGTGGTTCTT
MCN5'-P1-VH4-LEADER-D	*CCTCTCTATGGGCAGTCGGTGAT* ATGAAACATCTGTGGTTCTT
MCN5'-P1-VH5-LEADER-A	*CCTCTCTATGGGCAGTCGGTGAT* TTCTCCAAGGAGTCTGT
MCN5'-P1-VH5-LEADER-B	*CCTCTCTATGGGCAGTCGGTGAT* CCTCCACAGTGAGAGTCTG
MCN5'-P1-VH6-LEADER-A	*CCTCTCTATGGGCAGTCGGTGAT* ATGTCTGTCTCCTTCCTCATC
MCN5'-P1-VH7-LEADER-A	*CCTCTCTATGGGCAGTCGGTGAT* GGCAGCAGCAACAGGTGCCCA
3'-end primer with A sequencing adapter	
VRC3'-A-CγCH1	*CCATCTCATCCCTGCGTGTCTCCGACTCAG* GGGGAAGACCGATGGGCCCTTGGTGG
VRC3'-A-CμCH1[b]	*CCATCTCATCCCTGCGTGTCTCCGACTCAG* GGGAATTCTCACAGGAGACGA

[a] The reversed PGM sequencing protocol (from end of the J gene segment to the beginning of the V gene segment) was devised to minimize sequencing errors, especially homopolymer and indel errors, in the complementarity determining region 3 (CDR3) of the heavy and light chains. The PGM sequencing adapters are highlighted in all the primer sequences. The 5'-end primers were first described by Schield et al in Ref (36) (optimized for capturing antibody heavy chains with high somatic hypermutation rate) and adapted here for PGM sequencing. The 3'-end primers are adapted from the primers described in Refs (11-15).

[b] Only Cγ primer was used in current study to capture memory and plasma B-cell transcripts in a donor sample, Cμ primer is provided for completeness.

Table S4. 5'-RACE PCR primers used to prepare samples for PGM sequencing (reverse direction). [a]

Primer name	Primer sequence (5' → 3')
1. Heavy chain	
5'-end primer with PGM A sequencing adapter	
trP1/P1-5'-RACE	*CCTCTCTATGGGCAGTCGGTGAT* 5'-RACE adaptor
3'-end primer with PGM P1 sequencing adapter	
VRC3'-A-CγCH1	*CCATCTCATCCCTGCGTGTCTCCGACTCAG* GGGGAAGACCGATGGGCCCTTGGTGG
VRC3'-A-CμCH1[c]	*CCATCTCATCCCTGCGTGTCTCCGACTCAG* GGGAATTCTCACAGGAGACGA
2. Lambda chain	
5'-end primer with PGM A sequencing adapter	
trP1/P1-5'-RACE	*CCTCTCTATGGGCAGTCGGTGAT* 5'-RACE adaptor
3'-end primer with PGM P1 sequencing adapter	
VRC3'-A-Cλ	*CCATCTCATCCCTGCGTGTCTCCGACTCAG* CACCAGTGTGGCCTTGTTGGCTTG
3. Kappa chain [b]	
5'-end primer with PGM A sequencing adapter	
trP1/P1-5'-RACE	*CCTCTCTATGGGCAGTCGGTGAT* 5'-RACE adaptor
3'-end primer with PGM P1 sequencing adapter	
VRC3'-A-Cκ	*CCATCTCATCCCTGCGTGTCTCCGACTCAG* CAGCAGGCACACAACAGAGGCAGTTCC

[a] The 5'-RACE PCR-based PGM sequencing protocol was devised in such a way that the use of a single 3'-end primer can capture all germline V gene families, thus providing an unbiased view of the antibody repertoire. For 5'-RACE PCR products, the PGM sequencing will be done with a reverse direction – from the end of the J gene segment to the beginning of the V gene segment – and end in or past the leader region. The PGM sequencing adapters are underscored in all the primer sequences.
[b] The kappa chain primers, although not used in current study, are added to complete the primer set.
[c] Only Cγ primer was used in current study to capture memory and plasma B-cell transcripts in a donor sample, Cμ primer is provided for completeness.

Table S5. Expression of antibodies with selected IAVI donor 17 heavy and light chains paired with respective native partner chains. [a]

No.	Chain type	Sequence index	Yield (mg/L culture sup)	Neutralization	PGT sequence identity(%)	Amino acid sequence of variable domain
A. Selected heavy and light chain sequences from PGM sequencing with VH4- and VL3-specific primers						
1	H	740459	16.34	Y	121H (97.5)	QVQLQESGPGLVKPSETLSLTCSVSGASISDSYWSWFRRPPGKGLEWIGYVHKSGDTNYSPSLKSRVNLSLDASKNQVSLSLVAATAADSGKYYCARTLHGRRIYGIVAFNEWFTYFYMDVWGNGTQVTVSS
2	H	740466	2.46	Y	121H (97.2)	QVQLQESGPGLVKPSETLSLTCSVSGASISDSYWSWFRRPPGKGLEWIGYVHKSGDTNYSPSLKSRVNLSLDASKKQVSLSLVAATAADSGKYYCARTLHGRRIYGIVAFNEWFTYFYMDVWGNGTQVTVSS
3	H	740522	8.96	Y	121H (97.7)	QMQLQESGPGLVKPSETLSLTCSVSGASISDSYWSWFRRPPGKGLEWIGYVHKSGDTNYSPSLKSRVNLSLDASKNQVSLSLVAATAADSGKYYCARTLHGRRIYGIVAFNEWFTYFYMDVWGNGTQVTVSS
4	H	740527	3.24	N	121H (94.7)	QVQLQESGPGLVKPSETLSLTCSVSGASISDSYWSWFRRPPGKGLEWIGYVHKSGDTNYSPSLKSRVNLSLDASKKQVSLSLVAATAADSGKYYCARTLHGRRIYEYVVFIESVTYFYMDVWGNGTQVTVSS
5	H	682486	4.68	Y	122H (98.0)	QVHLQESGPGLVKPSETLSLTCNVSGTLVRDNYWSWIRQPPGKQPEWIGYVHDSGDTNYNPSLKSRVHLSLDKSKNLVSLRLTGVTAADSAIYYCATTKHGRRIYGVVAFKEWFTYFYVDVWGKGTSVTVSS
6	H	682945	15.42	Y	122H (97.2)	QVHLQESGPGLVKPSETLSLTCNVSGTSVRDNYWSWIRQPPGKQPEWIGYVHDSGDTNYNPSLKSRVHLSLDKSKNLVSLRLTGVTAADSAIYYCATTKHGRRIYGVVAFKEWFTYFYMDVWGKGTSVTVSS
7	H	88610	10.08	Y	122H (99.0)	QVHLQESGPGLVKPSETLSLTCNVSGTLVRDNYWSWIRQPPGKQPEWIGYVHDSGGTNYNPSLKSRVHLSLDKSKNLVSLRLTGVTAADSAIYYCATTKHGRRIYGVVAFKEWFTYFYMDVWGKGTSVTVSS
8	H	750540	14.39	Y	123H (94.2)	QVRLQESGPGLVKPPETLSLTCSVSGASINDAYWSWIRQSPGKRPEWVGYVHHSGDTNYNPSLKRRVTLSLDTAKNEVSLKLVALTAADSAVYFCARALHGKRIYGTVALGELFVYFYMDVWGQGTLVTVSS
9	H	744657	1.72	Y	124H (98.5)	QVQLQESGPGLVRPSETLSVTCIVSGGSISNYYWTWIRQSPGKGLEWIGYICYRETTTYNPSLNSRAVISRDTSKNQLSLQLRSVTTADTAIYFCATARRGQRIYGVVSFGEFFYYYYMDVWGKGTAVTVSS
10	H	679170	6.23	Y	124H (91.2)	QVQLQESGPGLVKPSETLSVTCSVSGDSMNNYYWTWIRQSPGKGLEWIGYVSDRASATYNPSLKSRVVISRDTSKNQLSLKLNSVTLADTAVYYCATARRGQRIYGEVASGEFFYYYSMDVWGKGTAVTVSS
11	H	679309	5.75	Y	124H (91.9)	QVQLQESGPGLVKPSETLSVTCSVSGDSMNNYYWTWIRQSPGKGLEWIGYVSDRASATYNPSLKSRVVISRDTSKNQLSLKLNSVTLADTAVYYCATARRGQRIYGEVPFGEFFYYYSMDVWGKGTAVTVSS
12	H	610643	3.12	Y	133H (96.2)	QVHLQESGPGVVTPSETLSLTCSVSNGSVSGRFWSWIRQPPGRGLEWIGYFSDTDRSEYNPSLRSRLTLSVDKSKNQLSLRLKSVTAADSATYYCARAQQGKRIYGMVSFGEFFYYYYMDAWGKGTPVTVLS
13	H	748882	1.68	Y	133H (97.0)	QVHLQESGPGVVTPSETLSLTCSVSNGSVSGRFWSWIRQPPGRGLEWIGYFFYTDRSEYNPSLRSRLTLSVDKSKNQLSLRLKSVTAADSATYYCARAQQGKRIYGMVSFGEFFYYYYMDAWGKGTPVTVSS
14	H	748844	No expression [b]	N	133H (92.4)	QVHLQESGPGLVTPSETLSLTCSVSNGSVSGRFWSWIRQPPGRGLEWIGYFSDTDRSEYNPSLRSRLTLSVDKSKNQLSLRLRSVTAADSATYYCVETQQGKRIYGMVSFGEFFYYYYMDAWGKGTPSVLAS
15	H	634779	5.44	Y	134H (99.5)	QVHLQESGPGLVTPSETLSLTCTVSNGSVSGRFWSWIRQPPGRGLEWIGYFSDTDRSEYNPSLRSRLTLSVDRSKNQLSLKLKSVTAADSATYYCARAQQGKRIYGIVSFGELFYYYYMDAWGKGTPVTVSS
16	L	2561142	6.24	Y	121L (96.5)	SDISVAPGETARITCGGKSLGSRAVQWYQHRAGQAPLLIIYNNQDRPSGIPERFSGSPDSAFGTTATLTITSVEAGDEADYYCHIWDSRVPTKWVFGGGTTLTVL
17	L	2561163	2.58	Y	121L (96.5)	SDISVAPGETARITCGEKSLGSRAVQWYQHRAGQAPLLIIYNNQDRPSGIPERFSGSPDSAFGTTATLTITSVEAGDEADYYCHIWDSRVPTKWVFGGGTTLTVL
18	L	2561217	No expression [b]	N	121L (93.7)	SDISVAPGETARITCGEKSIGSRAVQWYQQRAGQAPLLIIYNNQDRPSGIPERFSGSPXLPLGTTATLTITSVEAGDEADYYCHIWDSRVATDWVLGGGTTLTVL
19	L	3597305	No expression [b]	N	123L (95.2)	SSMSVSPGETAKISCGKESIGSRAVQCTQARQAPSQPPSLIIYNNQDRPAGVPERFSASPYFRPGTTATLTITNVDAEDEADYYCHIYDARGGTNWVFDRGTTLTVL
20	L	3600527	1.12	Y	123L (99.4)	SSMSVSPGETAKISCGKESIGSRAVQWYQQKPGQPPALIIYNNQDRPAGVPERFSASPDFRPGTTATLTITNVDAEDEADYYCHIYDARGGTNWVFDRGTTLTVL
21	L	1409147	18.25	Y	124L (91.7)	PPVRPLSVALGETASISCGRQALGSRAVQWYQHRPGQAPVLLIYNNQDRPSGIPERFSGTPDINFGTRATLTISGVEAGDEADYYCHMWDSRSGFSWSFGGATRLTVL
22	L	1434766	9.09	Y	124L (92.6)	PPVRPLSVALGETASIPCGRQALGSRAVQWYQHRPGQAPVLLVYNNQDRPSGIPERFSGTPDINFGTRATLTISGVEAGDEADYYCHMWDSRSGFSWSFGGATRLTVL
23	L	1456726	5.26	Y	124L (93.5)	SFVRPLSVALGETASITCGRQALGSRAVQWYQHRPGQAPVLLFYNNQDRPSGIPERFSGTPDINFGTRATLTISGVEAGDEADYYCHMWDSRSGFSWSFGGATRLTVL
B. Selected heavy and light chain sequences from PGM sequencing with 5'-RACE PCR						
24[c]	H	999229	17.40	Y	122H (100)	QVHLQESGPGLVKPSETLSLTCNVSGTLVRDNYWSWIRQPLGKQPEWIGYVHDSGDTNYNPSLKSRVHLSLDKSKNLVSLRLTGVTAADSAIYYCATTKHGRRIYGVVAFKEWFTYFYMDVWGKGTSVTVSS
25	H	2794987	7.58	Y	122H (95.5)	QFHLQESGPGLVKPSETLSLTCNVSGTLVRDNYWSWIRQPLGKQPEWIGYVHDSGDTNYNPSLKSRVHLSLDKSKNLVSLRLAGVTAADSAIYYCATTKHGRRIYGVVAFKEWFTYFYMDVWGKGTSVTVSS
26	H	1404562	11.95	Y	122H (92.7)	QVHLQESGPGLVKPSETLSLTCNVSGTLVRDNYWSWIRQPLGKHPEWIGYVHDSGDTNYNPSLKSRAHLSLDKSKNLVSLRLSAVTAADSAIYYCATTKHSRRIYGIVAFNEWFTYFYMDIWGKGASVTVSS
27	H	2098660	1.25	Y	122H (91.9)	QVHLQESGPGLVNLTETLSLTCNVSGTLVRDNYWSWIRQPSGTHPEWIGYVHDSGDTNYNPSLKSRVHLSLDKSKNLVSLRLTAVTAADSAIYYCATTKHSRRIYGIVAFNEWFTYFYMDIWGKGASVTVSS
28	H	154686	No expression [b]	N	122H (88.6)	QVHLQVSGPGLVNLRETLSLTCNVSGTLVRDNYWSWIRQPSGKHPEWIGYVHDSGDTNYNPSLKSRVHLSLDKSKNLVSLRLSAVTAADSAIYYCATTKHSRRIYGIVAFNEWFTYFYMDILGEGTLVTVSS
29	H	1648082	No expression [b]	N	122H (89.9)	QVHLQESGPGLVKPSETLSLTCNVSRDSVRDNYWSWIRQPLGKHPEWIGYVHDSGDTNYNPSLKSRVHLSLDKSKNLVSLRLTAVTAADSAIYYCATTKHSRRIYGIVAFNEWFTYFYMHIWGKGTSVTVSS
30	L	1009654	10.99	Y	122L (96.2)	SLVSVAPGQTARITCGEESLGSRSVIWYQQRPGQAPSLIIYNNNDRPSGIPERFSGSPGSTFGTTATLTITSVEAGDEADYYCHIWDSRRPTNWVFGEGTTLTVL
31	L	1450263	4.08	Y	122L (94.3)	SLVSVAPGQTARITCGEESLGSRSVIWYQQRPGQAPSLIIYNNNDRPSGIPERFSGSPGSTFGTTATLTIISVEAGDEADYYCHIWDSRRPTNWVFGEGTTLTVL
32	L	2107682	No expression [b]	N	122L (92.1)	DDESVSPGQTGRITCGEESLGSRSVICTQQRPGQALQLIIYNNNDRPSGIPERFSGSPGSTFGTTATLTITSVEAGDEADYYCHIWDSRRPTNWVFGEGTTLTVL
33	L	1031843	3.68	Y	123L (96.2)	SSMSVSPGETAKISCGKESIGGRAVQWYQQKPGQPPSLIIYNNQDRPSGVPERFSASPDFRPGTTATLTITNVEAEDEADYYCHIYDARGGTNWVFDRGTTLTVL
34	L	1117669	0.98	Y	123L (96.2)	SSMSVSPGETAKISCGKESIGSRAVQWYQQKPGQPPQLIIYNNQDRPSGVPERFSASPDFRPGTTATLTITNVEAEDEADYYCHIYDARGGTNWVFDRGTTLTVL

[a] Listed items include sequence number, antibody chain type, sequence index in the PGM data set, protein yield, neutralization, sequence identity to the closest PGT121-class antibody chain and amino acid sequence.

[b] No expression denotes protein yield less than 0.60mg/L. For No.18 (2561217), an undetected stop codon (X) led to no expression.

[c] The neutralization was not tested for No. 24 (999229) since its sequence identity to PGT122 heavy chain is 100% and the reconstituted antibody would be the native PGT122.

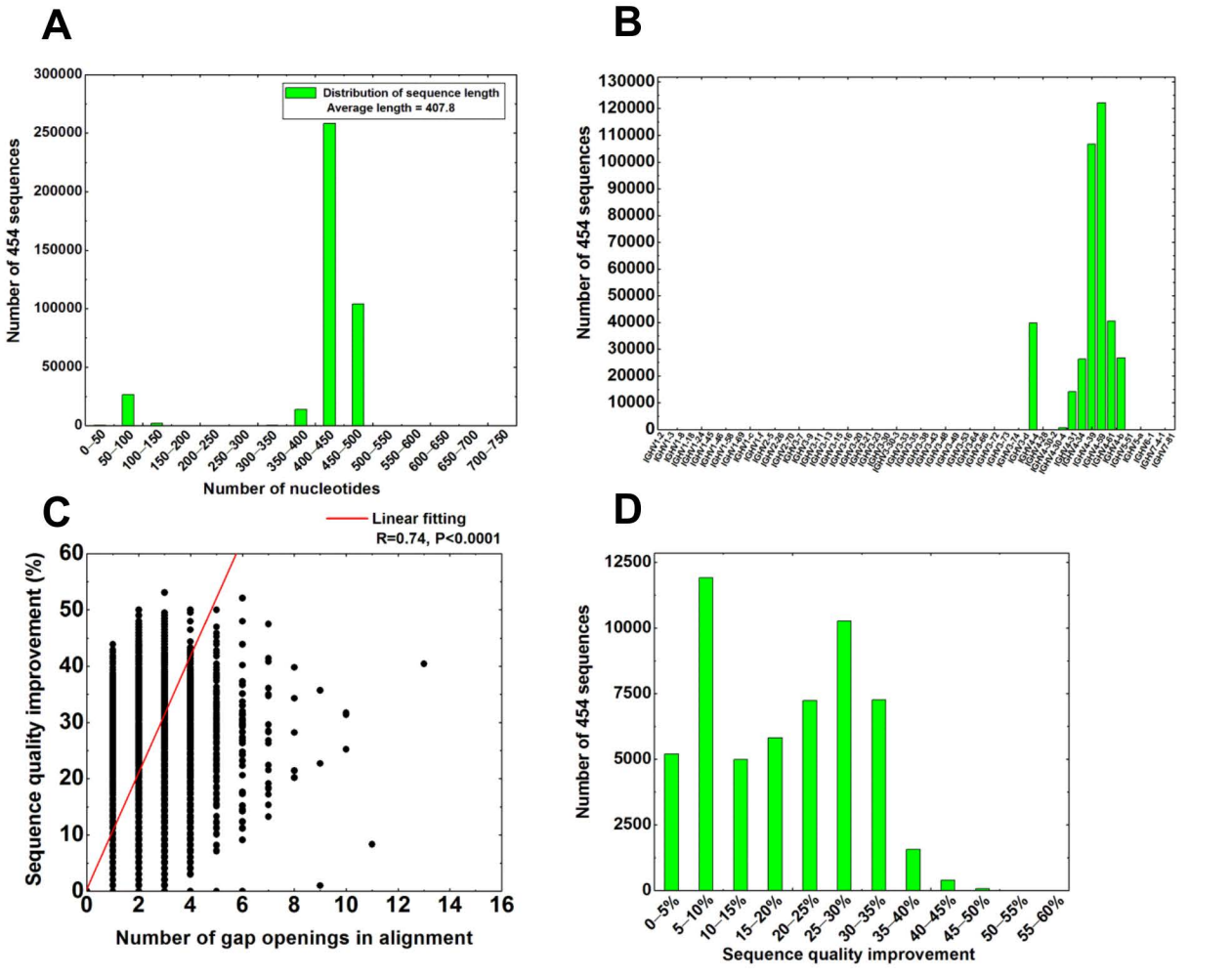

Figure S1. Pipeline processing of IAVI donor 17 heavy-chain 454 sequencing data set reported by Sok et al (ref 15). (A) Read length distribution (step 1). Of 406,063 reads, 362,362 (or 89.2%) are over 400bp. (B) Germline gene distribution (step 2). Note that 122,079 sequences are of IgHV4-59 germline origin. (C) Correlation between number of gaps in V gene alignment and sequence quality improvement measured by the increase of sequence identity to germline gene after correction (step 3). The correlation coefficient is 0.74 and P-value is less than 0.0001. (D) Distribution of sequence quality improvement (step 3), with an average of 18.6%.

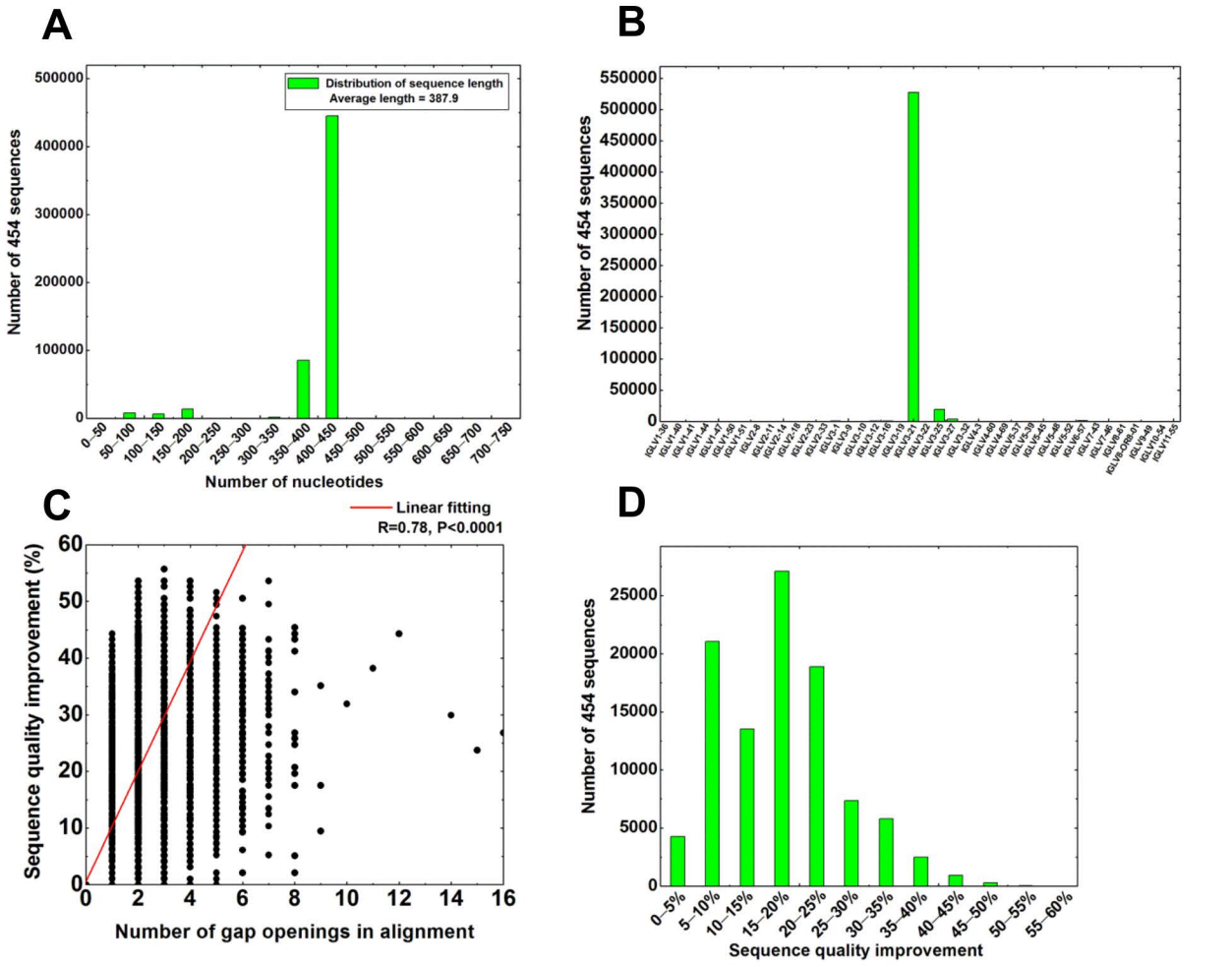

Figure S2. Pipeline processing of IAVI donor 17 light-chain 454 sequencing data set reported by Sok et al (Ref 15). (A) Read length distribution (step 1). Of 560,872 reads, 445,144 (or 79.4%) are over 400bp. (B) Germline gene distribution (step 2). Note that 527,100 sequences are of IgLV3-21 germline origin. (C) Correlation between number of gaps in V-gene alignment and sequence quality improvement measured by the increase of sequence identity to germline gene after correction (step 3). The correlation coefficient is 0.78 and P-value is less than 0.0001. (D) Distribution of sequence quality improvement (step 3), with an average of 17.5%.

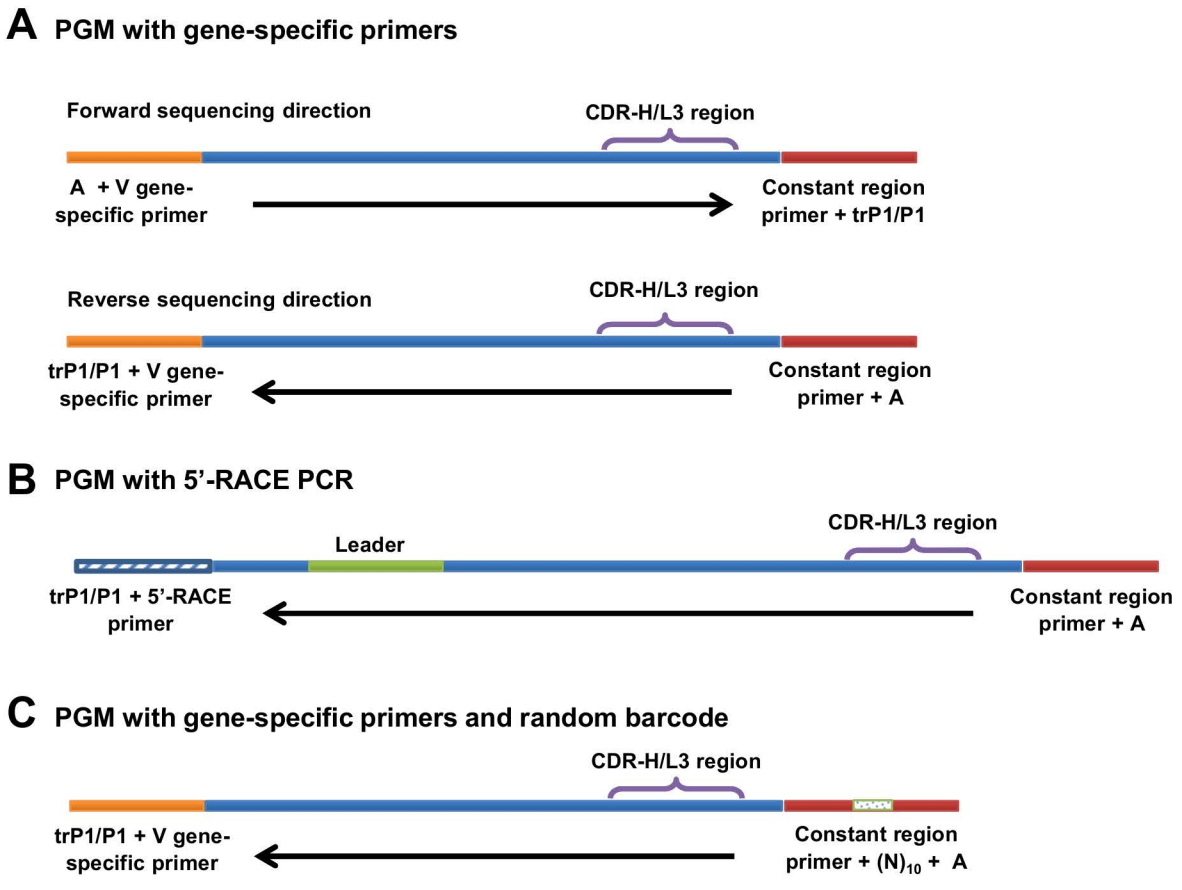

Figure S3. PGM sequencing strategies. (A) PGM sequencing with gene-specific primers in forward (upper panel) and reverse (lower panel) directions. (B) PGM sequencing with 5'-RACE PCR in reverse direction. (C) PGM sequencing with gene-specific primers and a string of 10 degenerate nucleotides – random barcode – inserted between the constant-domain primer and PGM adaptor.

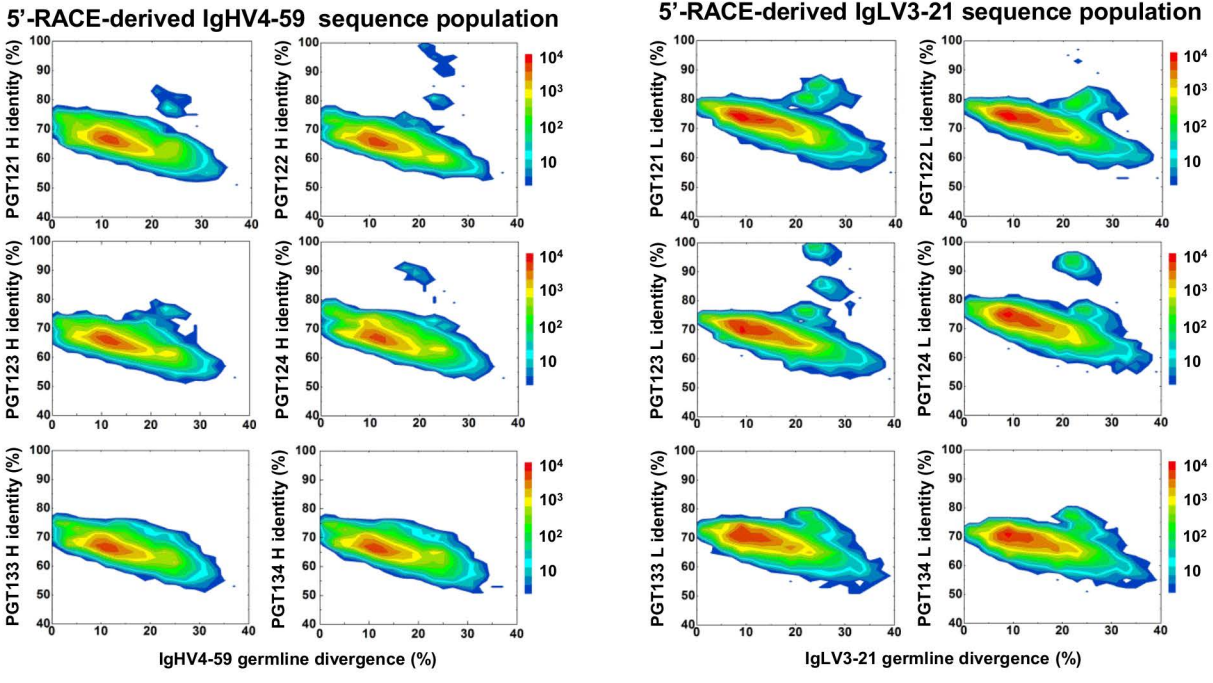

Figure S4. Identity/divergence analysis of 5'-RACE PCR-derived sequence population for IAVI donor 17. Heavy and light chains of the representative PGT121-class antibodies – PGT121-124 and PGT133-134 – are used as template in the sequence identity calculation. The heavy chains of IgHV4-59 origin (left) and the light chains of IgLV3-21 origin (right) are plotted as a function of sequence identity to a template and of sequence divergence from the inferred germline gene.

A Heavy chain

Barcode: GGAGCGCCGG

CDR H3 sequences:

```
>1489 (75nt)
GATCTGCGACATACAGTGGCTCAGATCGACGGAAAGAATTACTATGATAGAAGTGATTATGGTCCCTTTGATATC
>1489 (25aa)
DLRHTVAQIDGKNYYDRSDYGPFDI

>172658 (36nt)
GAGCTACCTTCGACCCACCATGATGTTTTTGATATC
>172658 (12aa)
ELPSTHHDVFDI

>1889503 (30nt)
TCAGGATACAACTACGGGGTCTTTGACATC
>1889503 (10aa)
SGYNYGVFDI
```

B Light chain

Barcode: GTGGCCGGGT

CDR L3 sequences:

```
>1267055 (39nt)
CACCTGTGGGATGCTACTACTGATCATCCGGGTTATGTC
>1267055 (13aa)
HLWDATTDHPGYV

>1541525 (30nt)
CACGTGTGGGATAGAAATAGTGATCCCCTC
>1541525 (10aa)
HVWDRNSDPL
```

Figure S5. Examples of different templates tagged by the same barcode sequence. (A) A group of IgVH4-59-originated heavy chains with the same barcode show three distinct CDR H3 sequences. (B) A group of IgLV3-21-originated light chains with the same barcode show two distinct CDR L3 sequences.

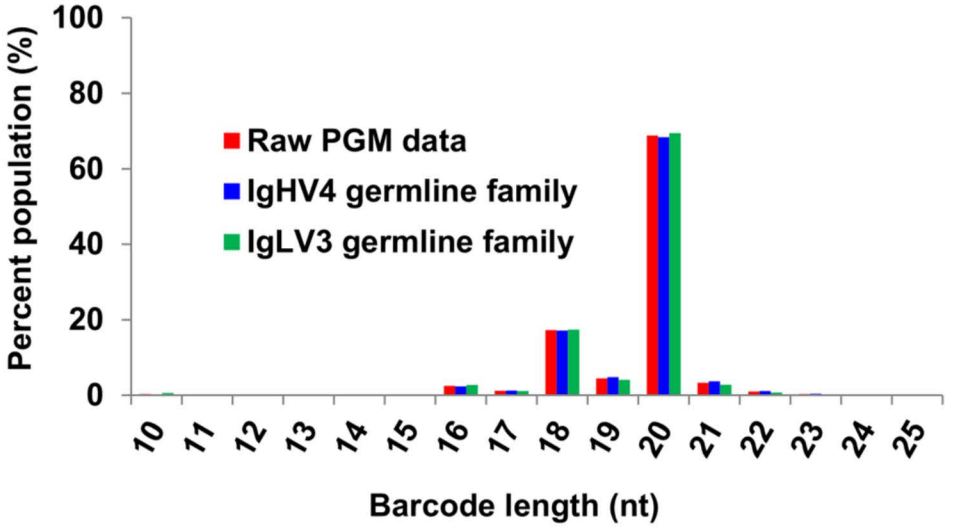

Figure S6. Barcode length distribution for the raw sequencing data (red) and the pipeline-processed IgHV4 germline family (blue) and IgLV3 germline family (green) generated using the same random barcoding strategy (Fig. 5) and a barcode of 20 degenerate nucleotides (nt). The PGM sequencing was performed using an Ion 314 v2 chip and standard settings. Plotted in the distribution are 311,083 raw reads, 186,478 sequences of IgHV4 origin, and 124,605 sequences of IgLV3 origin. Due to the increased homopolymer errors in the barcode region, the distribution of barcode length obtained from a 20-nt barcode is notably broader than that from a 10-nt barcode (Fig. 5), with ~2% of the sequences have a 16-nt barcode, ~17% having a 18-nt barcode, 4-5% having a 19-nt barcode, 68-69% having a 20-nt barcode, and 3-4% having a 21-nt barcode.