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

SCIENTIFIC REPORTS

OPEN

SUBJECT AREAS:
NEXT-GENERATION SEQUENCING
IMMUNOLOGY
DATA PROCESSING

Received
16 June 2014

Accepted
7 October 2014

Published
27 October 2014

Correspondence and requests for materials should be addressed to J.Z. (jiang@scripps.edu)

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[1,2,3,4], Parisa Azadnia[1], Jessica Hsueh[1,2,4], Elise Landais[2], Melissa Simek[3], Wayne C. Koff[3], Pascal Poignard[1,2,3], Dennis R. Burton[1,2,3,4,5] & 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.

B-cell repertoire analysis using next-generation sequencing has become a valuable tool for interrogating the genetic record of humoral response to infection. However, key obstacles such as low throughput, short read length, high error rate, and undetermined bias of multiplex PCR method have hindered broader application of this technology. In this study, we report several technical advances in antibody repertoire sequencing. We first demonstrated the ability to sequence antibody variable domains using the Ion Torrent PGM platform. As a test case, we analyzed the PGT121 class of antibodies from IAVI donor 17, an HIV-1-infected individual. We then obtained "unbiased" antibody repertoires by sequencing the 5′-RACE PCR products of B-cell transcripts from IAVI donor 17 and two HIV-1-uninfected individuals. We also quantified the bias of previously published gene-specific primers by comparing the repertoires generated by 5′-RACE PCR and multiplex PCR. We further developed a single-molecule barcoding strategy to reduce PCR-based amplification noise. Lastly, we evaluated several new PGM technologies in the context of antibody sequencing. We expect that, based upon long-read and high-fidelity next-generation sequencing technologies, the unbiased analysis will provide a more accurate view of the overall antibody repertoire while the barcoding strategy will facilitate high-resolution analysis of individual antibody families.

Next-generation sequencing (NGS) has transformed many areas of biological and translational research[1–4]. Recently, the scope of NGS application has expanded into the analysis of antibody repertoire encoded by B cells[5–7]. Demonstrated with proof-of-concept studies in animal models[8,9], NGS-based antibody repertoire analysis has been applied to examine human samples[10], particularly in the study of human immunodeficiency virus type-1 (HIV-1) infected individuals with broadly neutralizing antibodies (bnAbs)[11–16]. For these bnAbs, special bioinformatics tools have been developed to identify somatic variants and maturation pathways from NGS-derived repertoires[11–16].

Unlike other NGS applications, antibody repertoire analysis faces unique challenges in both sequencing and data analysis due to the complexity of B-cell development, in which antigen-driven affinity maturation selects for somatic mutations throughout variable region of immunoglobulin genes. It is therefore critical to sequence entire antibody variable domains (~450 bp) for a meaningful repertoire analysis and to recover functional antibodies from the NGS data. Long reads are particularly critical for the study of HIV-1 bnAbs, which often show 20–35% sequence divergence compared to their germline precursors[17–20]. As a result, most studies have been carried out using the 454 platform[11–16], as this technology typically has a read length of around 400 bp, but with a relatively low throughput. Another critical factor in NGS-based repertoire analysis is sequencing error, which is platform-specific and thus requires different algorithms for correction[13,21–23]. The 454 and PGM platforms suffer from homopolymer errors, which can be corrected using germline genes as a template[14], whereas the MiSeq platform generates substitution errors, which can be corrected by calculating a consensus[7]. Irrespective of the NGS

platform, experimental details in sample preparation such as polymerase chain reaction (PCR) primers also play a critical role in producing a reliable repertoire[22]. Although 5′-RACE PCR has been proposed as a solution for unbiased repertoire analysis, the long PCR products (~600 bp) pose a significant challenge to current NGS platforms. Recently, a multiplex PCR method with minimized bias was reported for T-cell repertoire analysis[24]. Meanwhile, the library amplification based on PCR will produce redundant cDNA molecules, which when combined with sequencing errors, may lead to artificial antibody clones and diversity in repertoire analysis[22]. Although the basic strategies for antibody repertoire analysis have just been established and not yet optimized, the current research focus has begun to shift from cross-sectional studies to longitudinal analyses[25], which require high-precision dissection of repertoire properties to establish meaningful biological conclusions. Therefore, it remains unclear whether current antibody sequencing technologies will suffice for these new applications.

In this study, we adapted the Ion Torrent Personal Genome Machine (PGM) for high-throughput sequencing of full-length antibody variable domains. We validated the platform with samples from an HIV-1-infected donor (IAVI donor 17), the source of bnAb PGT121 and its siblings[19], and two HIV-1-uninfected donors. The greater depth of PGM sequencing allowed us to identify a more complete somatic population of the PGT121-class antibodies. We then introduced 5′-RACE PCR into template preparation in order to capture antibody repertoires in an unbiased manner. We compared the overall properties of the unbiased repertoires to those obtained using multiplex primers. We also developed a random barcoding strategy to track individual antibody cDNA molecules and to reduce amplification noise and sequencing error. With a side-by-side comparison, we demonstrated that the new template amplification methods and sequencing chemistry could significantly improve the repertoire quality compared to the current methods based on this platform.

Results

PGT121 class of broadly neutralizing antibodies. Identified from African donor 17 of the IAVI Protocol G cohort (IAVI donor 17)[19], the PGT121 class of antibodies was originally described as consisting of six members: PGT121–124 and PGT133–134. These antibodies potently neutralize 65–70% of HIV-1 isolates (median $IC_{50} < 0.05$ μg ml^{-1}) by recognizing a high-mannose patch in the gp120 V3 region[26]. Further structural analysis revealed that PGT121–123 share a similar mode of recognition by recruiting multiple structural elements in the HIV-1 envelope variable regions[27]. In the 454 analysis of IAVI donor 17, a novel phylogenetic method was devised to infer putative intermediates, which potently neutralized diverse strains on a 74-virus panel with half the mutation level of the mature parent antibodies[16]. Recently, Barouch et al demonstrated the therapeutic value of PGT121 in simian-human immunodeficiency virus (SHIV)-infected rhesus monkeys[28]. With the extensive information available for the PGT121 class of antibodies, the sample from IAVI donor 17 provides a unique test case for investigating various aspects of antibody repertoire analysis.

We first reanalyzed the 454 sequencing data of IAVI donor 17[16] with the *Antibodyomics 1.0* pipeline[11–15] (Figs. S1 and S2). Of 966,935 raw reads, 122,079 originated from the IgHV4-59 gene and 527,100 from the IgLV3-21 gene, the heavy and light chain germline precursors of PGT121 class, respectively. Using a sequence identity cutoff of 90%, closely related somatic variants were identified for both PGT121 heavy and light chains, but not for other members of the class (Fig. 1A). Through intra-donor phylogenetic analysis[13,14], we identified 81 heavy chains and 470 light chains that were somatically related to the PGT121 class of antibodies (Fig. 1B).

PGM sequencing of IAVI donor 17 with IgVH4 and IgVL3 primers. We performed deep sequencing of antibody transcripts from memory and plasma B cells using PCR to amplify heavy chains from the IgHV4 family and light chains from the IgLV3 family (Fig. S3A). Briefly, mRNA from an estimated 20 million PBMCs was used for reverse transcription (RT) to produce template cDNA (Table S1). PGM sequencing was performed using an Ion 316 chip and a modified protocol, in which the default 3′-trimming option was turned off in order to obtain longer reads. The *Antibodyomics 1.0* pipeline was used for data processing and sequencing error correction. The pipeline output and the following PGM sequencing experiments are briefly summarized in Table 1.

PGM sequencing provided 3,610,144 raw reads, of which 172,732 sequences were assigned to the IgHV4-59 gene and 1,615,722 to the IgLV3-21 gene, the heavy and light chain germline precursors of PGT121 class, respectively. After pipeline processing, the sequences of IgHV4-59 and IgLV3-21 origins were compared to the PGT121 class of antibodies. The identity/divergence plots revealed over 90% identical sequences for all 6 PGT121-class heavy chains as well as the PGT121 and PGT123 light chains (Fig. 2A). These results indicate that the PGM platform, when used in conjunction with germline gene-specific primers, could effectively capture the closely related somatic variants for this antibody class with greater coverage than the 454 platform (Fig. 1A).

Using the 6 PGT121-class antibodies as a template, iterative intra-donor phylogenetic analysis[13,14] was performed to identify all the heavy and light chains that were somatically related to this class. After 10 iterations, the analysis converged to 1,022 heavy chains and 2,282 light chains (Fig. 2B), which compared to the 454-derived somatic variants show a ~12- and ~5-fold increase in the number of sequences, respectively[16]. The greater number of somatic variants can be attributed to the greater sequencing depth, although some of the new clones might have resulted from errors in RT, PCR and sequencing as stressed in recent reviews[22]. Nevertheless, with more somatic variants identified, the intra-donor phylogenetic trees (Fig. 2C) appeared to be more complete and should be more suitable for inferring intermediate sequences[16].

Our results thus illustrated the importance of sequencing depth in antibody repertoire analysis, especially in the identification of rare bnAb clones and lineage intermediates, which will further advance our understanding of bnAb development and provide new antibody targets for B-cell precursor and lineage-based immunogen design[29,30].

Functional validation of PGM-derived PGT121 somatic variants. We validated the neutralization of newly identified PGT121 variants on a 6 cross-clade virus panel[16]. For each member of the PGT121 class, heavy and light chains alike, we extracted the sequences with an identity of 90% or greater with respect to the template and grouped the sequences using an identity cutoff of 100%. We then manually selected sequences that represented closely related somatic variants (>95%) and have diverged further in the maturation (<95%). This analysis resulted in 15 heavy chains and 8 light chains, which were synthesized and paired with their respective native partner chains (Table S5A). 19 antibodies were expressed and tested in neutralization assays, with PGT121 and PGT133 included for comparison (Table 2A). For HIV-1 isolates 92BR020, 92RW020 and IAVI C22, 70% of the reconstituted antibodies neutralized with an IC_{50} ranging from 0.001 to 1 μg ml^{-1}, which is characteristic of the PGT121 class. The heavy chain somatic variants of PGT121, 122 and 124, when paired with their native light chains, showed comparable if not higher potency than the native antibodies. In the case of PGT123, where the light chain somatic variants were paired with the native heavy chain, neutralizing activity appeared to be similar among the somatic variants and higher than those heavy-chain chimeric antibodies. Taken together, the neutralization results confirmed that PGM can

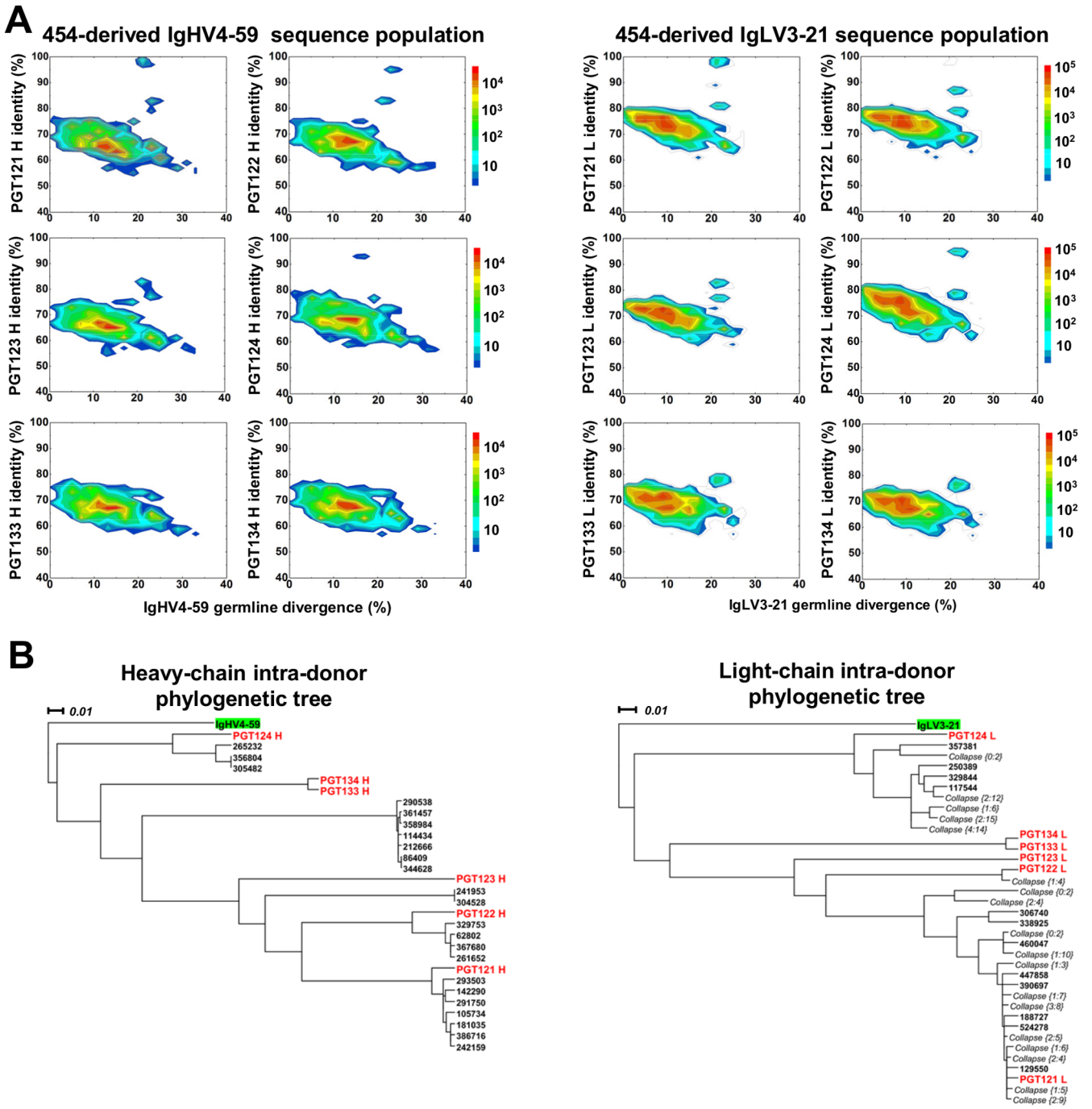

Figure 1 | **Analysis of reported 454 sequencing data of PGT121 class of antibodies.** (A) Identity/divergence analysis of 454-derived sequence population for IAVI donor 17. Heavy and light chains of the representative PGT121-class antibodies (PGT121–124 and PGT133–134) are used as template in the sequence 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 the template and the sequence divergence from the inferred germline gene. (B) Intra-donor phylogenetic trees calculated for heavy (left) and light chains (right). Iterative intra-donor phylogenetic analysis was performed to identify sequences that are somatically related to the PGT121 class of antibodies. After three iterations, the analysis converged to 81 heavy chains and 470 light chains, respectively.

be used to derive functional antibodies with a similar success rate (~70%) to 454[11–16], providing a biologically relevant assessment of the data quality generated by this platform.

Utility of 5′-RACE PCR for an unbiased repertoire analysis. The NGS analysis of HIV-1-infected donors has been focused primarily on specific germline gene families that give rise to the bnAbs of interest[11–16], leaving a majority of the antibody repertoire uncharacterized. Recently, Choi et al reported that 5′-RACE PCR offered an unbiased view of the murine *Igh* repertoire, which allows for an in-depth analysis of the V-gene rearrangement frequency[31]. With the PGM platform, we investigated the utility of 5′-RACE PCR in antibody repertoire analysis for IAVI donor 17 and two HIV-1-uninfected donors. Briefly, we adapted the murine procedure of Choi et al[31] for human samples by designing a new reverse primer that binds just downstream of the variable domain (Fig. S3B). We then sequenced the 5′-RACE PCR products to capture the entire heavy or light chain repertoire from the donor's memory and

Table 1 | Antibody repertoire analysis of an HIV-1-infected individual and two uninfected individuals[a]

Exp. index	Donor	PCR primers	PGM chip	N_{read}	Chain	N_{chain}	<Length>	$Perc_{no\text{-}gap}$	$Perc_{usable}$
A. Validation of PGM for full-length antibody variable domain sequencing									
1	IAVI donor 17	VH4	316	3,610,144	H	900,343	423.7	8.7%	89.8%
		VL3			L	2,675,687	429.8	10.1%	90.8%
B. Unbiased repertoire analysis based on 5′-RACE PCR									
2	IAVI donor 17	5′-RACE	316	3,104,454	H	1,098,334	501.5	15.0%	58.2%
					L	1,424,744	506.2	18.1%	65.5%
3	Uninfected donor #1	5′-RACE	316	3,257,758	H	1,176,731	514.7	13.3%	76.5%
					L	1,454,446	482.0	25.4%	71.5%
4	Uninfected donor #2	5′-RACE	316	3,350,792	H	1,279,994	518.8	8.3%	73.3%
					L	1,535,388	493.5	18.1%	66.5%
C. Comparison of repertoires obtained from 5′-RACE PCR and multiplex PCR									
5	IAVI donor 17	GP-H1	316	4,732,217	H	4,732,217	429.1	30.0%	79.7%
6		GP-H2 & GP-L1	318	6,802,786	H	2,508,220	495.0	7.7%	43.2%
					L	3,449,729	479.2	13.1%	57.9%
D. Validation of random barcoding in combined use with gene-specific primers									
7[b]	IAVI donor 17	VH4	316	3,109,512	H	1,133,165	448.2	18.1%	84.2%
		VL3			L	1,830,635	442.6	21.2%	90.0%
E. Assessment of new PGM technologies in the context of antibody sequencing									
8[c]	Uninfected donor #2	5′-RACE	316	3,902,464	H	2,079,766	565.6	28.4%	86.5%
					L	1,416,722	554.8	33.7%	91.8%
9[d]	Uninfected donor #2	5′-RACE	316	3,888,807	H	1,924,743	435.8	38.8%	65.0%
					L	1,360,189	453.0	42.5%	71.1%
10[e]	Uninfected donor #2	5′-RACE	316	4,205,538	H	1,261,313	560.5	33.3%	85.3%
					L	2,255,705	559.1	33.2%	81.1%
11[f]	Uninfected donor #2	5′-RACE	316	4,183,209	H	1,107,022	461.6	41.5%	73.1%
					L	1,889,100	455.7	40.7%	72.8%

[a]Listed items include the index of sequencing experiment, donor name, PCR primers, PGM sequencing chip, total number of raw reads, antibody chain type, number of antibody chains, average read length, percentage of sequences without indel errors in the V gene, and total number of usable sequences after pipeline processing. Note that the 3′-trimming option was turned off in all PGM sequencing experiments except for #9 and #11 and the *Antibodyomics 1.0* pipeline was used for data processing and error correction.
[b]PGM sequencing using the standard setup and random barcoding technology.
[c–d]PGM sequencing using an improved emulsion-based template preparation method and Hi-Q enzyme (new OT2 + Hi-Q) without 3-trimming (c) and with 3′-trimming (d).
[e–f]PGM sequencing using an emulsion-free template preparation method – isothermal amplification (IA) and Hi-Q enzyme (IA + Hi-Q) without 3′-trimming (e) and with 3′-trimming (f).

plasma B cells using this reverse primer (Table S4). PGM sequencing was performed on an Ion 316 chip without using 3′-trimming in raw data processing to extend the read length.

For IAVI donor 17, the combination of 5′-RACE PCR and PGM sequencing provided 1,098,334 heavy chains and 1,424,744 light chains (Table 1B), of which 71,689 sequences were assigned to IgHV4-59 and 121,652 sequences to IgLV3-21. For the two uninfected donors, 5′-RACE PCR and PGM sequencing provided over 3 million raw reads (Table 1B). Notably, for all three donors the average read length from 5′-RACE PCR was over 500 bp, compared to an average of 420–430 bp from gene-specific primers, highlighting the importance of long-read capability for unbiased antibody repertoire analysis.

For IAVI donor 17, IgHV4-59 accounted for 8.5% of the unbiased heavy chain repertoire (Fig. 3A). Surprisingly, IgHV5-51, which has not been associated with any HIV-1 bnAb, was the most prevalent germline gene family, accounting for 22.3% of the repertoire. The heavy chain repertoires from two uninfected donors presented two extremes (Fig. 3A). Uninfected donor #1 yielded a rather even distribution whereas uninfected donor #2 showed a skewed usage of germline genes IgHV1-69 and IgHV4-34. We then characterized the heavy chain complementarity determining region 3 (CDR H3) (Fig. 3B). IAVI donor 17 and the uninfected donor #1 showed somewhat similar distributions, with IAVI donor 17 having slightly longer CDR H3 regions. Surprisingly, uninfected donor #2 showed a two-peak distribution with the second peak centered at ~22 aa, suggesting that a large portion of antibodies in this repertoire (~26%) possess unusually long CDR H3 loops. We further analyzed the heavy chains with long CDR H3s (22–26aa) and found that 71.2% of the sequences were of IgHV4-34 origin with a preferred usage of J6, the longest J gene segment, suggesting that the long CDR H3s resulted from V-D-J rearrangement rather than sequencing error. We also characterized the distribution of germline divergence to determine the degree of somatic hypermutation (Fig. 3C). IAVI donor 17 showed a higher divergence than the two uninfected donors, which is expected for the prolonged maturation process in HIV-1-infected individuals. The PGT121-class heavy chains showed an average divergence of 22.0%, which is 9% higher than that of the IgHV4-59 family, indicating that these bnAbs require a longer maturation process than non-HIV-1-specific antibodies of the same germline origin. However, uninfected donor #2 showed a substantially lower germline divergence than IAVI donor 17 and uninfected donor #1.

For IAVI donor 17 (Fig. 3D), IgLV3-21 accounted for 9.2% of the unbiased light chain repertoire, whereas IgLV3-1 appeared to be used predominantly (25.7%) in the repertoire. All three donors showed similar CDR L3 length distributions, with a clear preference for a 9–11 aa loop length (Fig. 3E), as opposed to a more spread and diverse CDR H3 length distribution. The light chains also showed a similar divergence pattern to the heavy chains, with more near-germline sequences (with a divergence of 0–1%) in the repertoire from uninfected donor #1 (Fig. 3F). The PGT121-class light chains gave an average germline divergence of 23.2%, compared to a lower value of ~13% calculated for the IgLV3-21 family.

The unbiased analysis revealed intriguing features of IAVI donor 17 repertoire. More specifically, it allowed for the PGT121 class of antibodies to be analyzed in the context of the entire repertoire, which showed that these bnAbs were not the most prevalent family in this donor's repertoire. The large population of near-germline, IgHV4-34-originated antibodies with long CDR H3 loops found in the unbiased repertoire of uninfected donor #2 can be potentially explained by different causes such as an ongoing immune response, an autoimmune condition or a unique genetic background. This finding highlights the potential utility of unbiased repertoire analysis in identifying transient antibody responses and unusual patterns of antibody maturation.

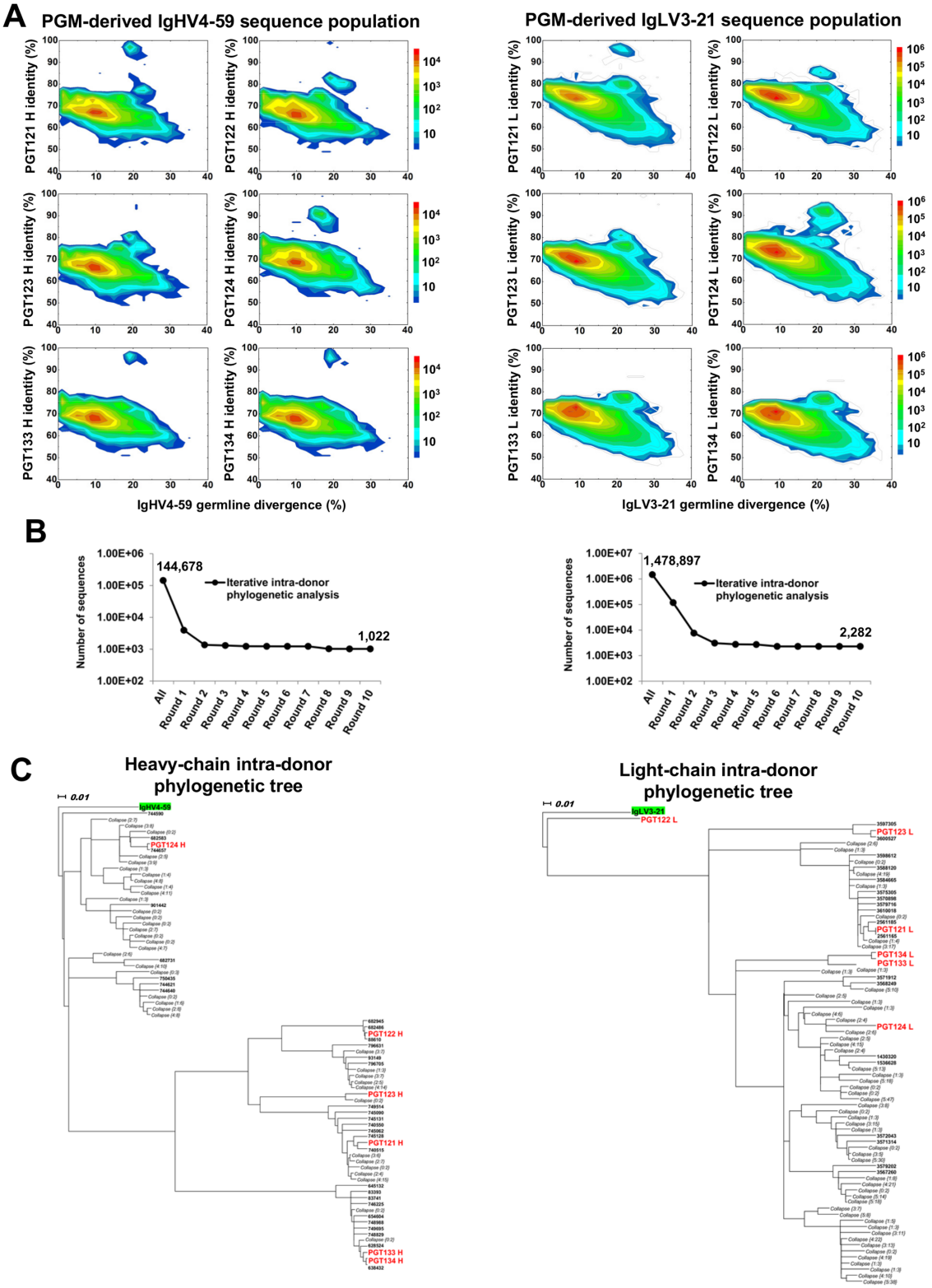

Figure 2 | Analysis of PGM sequencing data generated from IAVI donor 17 using VH4- and VL3-specific primers. (A) Identity/divergence analysis. Heavy chain sequences of the IgHV4-59 origin and light chain sequences of the IgLV3-21 origin are plotted as a function of sequence identity to the sequences of 6 PGT121-class antibodies and of sequence divergence from putative germline genes. (B) Iterative intra-donor phylogenetic analysis. Heavy and light chains with the same germline origin as the PGT121 class are subjected to multiple rounds of intra-donor phylogenetic analysis. In each round, the input sequences are divided into a number of subsets, each with the germline gene and 6 PGT121-class sequences included. After phylogenetic calculation, sequences that are clustered with the PGT121-class antibodies are extracted and used as input for the next round of analysis. The analysis converged to a fixed number of sequences after 10 iterations. (C) Intra-donor phylogenetic trees. Maximum-likelihood trees of V_H sequences of and V_L sequences from IAVI donor 17, along with 6 representative PGT121-class antibody sequences, are rooted by the respective germline gene sequences. Each bar represents a 0.1 change per nucleotide site.

Table 2 | Neutralization titers of 26 chimeric antibodies derived from IAVI donor 17 against 6 HIV-1 Env-pseudoviruses[a]

Antibody	Seq. No.	Neutralization IC_{50} titers (μg/ml)					
		92BR020	92RW020	92TH021	94UG103	IAVI C22	JR-CSF
PGT121	–	0.016	0.009	>50	1.785	0.007	0.042
PGT133	–	0.009	0.003	>50	0.275	0.002	0.027
A. Antibodies reconstituted from somatic variants identified from PGM sequencing with VH4- and VL3-specific primes							
gIAVI-H1$_{d17}$/PGT121L	740459	0.011	0.006	>50	1.844	0.003	0.025
gIAVI-H2$_{d17}$/PGT121L	740466	0.045	0.005	>50	>50	0.006	0.100
gIAVI-H3$_{d17}$/PGT121L	740522	0.045	0.014	>50	>50	0.010	0.165
gIAVI-H4$_{d17}$/PGT122L	682486	0.034	0.015	>50	1.261	0.006	0.106
gIAVI-H5$_{d17}$/PGT122L	682945	0.020	0.010	>50	1.125	0.023	0.063
gIAVI-H6$_{d17}$/PGT122L	88610	0.049	0.007	>50	>50	0.003	0.070
gIAVI-H7$_{d17}$/PGT123L	750540	0.013	0.008	>50	>50	0.006	0.177
gIAVI-H8$_{d17}$/PGT124L	744657	0.004	0.002	>50	0.133	0.001	0.034
gIAVI-H9$_{d17}$/PGT124L	679170	0.135	0.658	>50	>50	0.321	>50
gIAVI-H10$_{d17}$/PGT124L	679309	0.065	0.178	>50	>50	0.107	>50
gIAVI-H11$_{d17}$/PGT133L	610643	0.027	0.005	>50	0.626	0.010	0.051
gIAVI-H12$_{d17}$/PGT133L	748882	0.160	0.041	>50	9.353	0.032	0.271
gIAVI-H13$_{d17}$/PGT134L	634779	0.012	0.002	>50	0.262	0.002	0.022
PGT121H/gIAVI-L1$_{d17}$	2561142	0.006	0.005	>50	0.345	0.004	0.045
PGT121H/gIAVI-L2$_{d17}$	2561163	0.016	0.008	>50	0.331	0.003	0.064
PGT123H/gIAVI-L3$_{d17}$	3600527	0.011	0.001	>50	0.807	0.002	0.079
PGT124H/gIAVI-L4$_{d17}$	1409147	0.007	0.001	>50	0.133	0.001	0.022
PGT124H/gIAVI-L5$_{d17}$	1434766	0.010	0.002	>50	0.150	0.001	0.020
PGT124H/gIAVI-L6$_{d17}$	1456726	0.021	0.005	>50	0.406	0.001	0.070
B. Antibodies reconstituted from somatic variants identified from PGM sequencing with 5′-RACE PCR							
gIAVI-H14$_{d17}$/PGT122L	2794987	0.099	0.042	>50	>50	0.071	0.484
gIAVI-H15$_{d17}$/PGT122L	1404562	0.059	0.067	>50	>50	0.068	0.443
gIAVI-H16$_{d17}$/PGT122L	2098660	1.709	0.346	>50	>50	0.524	>50
PGT122H/gIAVI-L7$_{d17}$	1009654	0.019	0.007	>50	0.793	0.003	0.058
PGT122H/gIAVI-L8$_{d17}$	1450263	0.026	0.013	>50	1.374	0.016	0.103
PGT123H/gIAVI-L9$_{d17}$	1031843	0.007	0.003	>50	>50	0.003	0.090
PGT123H/gIAVI-L10$_{d17}$	1117669	0.004	0.002	>50	0.271	0.001	0.032

[a]The chimeric antibodies were expressed using the 16 heavy chains and 10 light chains derived from PGM sequencing of IAVI donor 17 with their native partner chains from the PGT121 class of antibodies. The wild-type PGT121 and PGT133 were included as a control. The IC_{50} values < 0.01 μg/ml are highlighted in red, 0.01 ~ 1 μg/ml in yellow, and 1 ~ 50 μg/ml in green.

Identification and validation of somatic variants from unbiased repertoire analysis. With 5′-RACE PCR, individual germline gene families were diluted due to non-specific amplification of all germline genes (Figs. 3A and 3D). As a result, the IgVH4-59 and IgLV3-21 families accounted for less than 10% of the total repertoire. The identity/divergence plots identified >90% identical sequence for both PGT122 heavy and light chains and for the PGT123 light chain, but not for other members of this class (Fig. S4). We identified 6 sequences with sequence identities ranging from 88.6% to 100% with respect to the PGT122 heavy chain. We also selected 3 and 2 somatic variants of the PGT122 and PGT123 light chains, respectively. When paired with their native partner chains, 7 chimeric antibodies were expressed (Table S5b) and in most cases neutralized diverse HIV-1 isolates (Table 2B). Taken together, although 5′-RACE PCR can provide a more accurate view of the overall antibody repertoire, gene-specific primers are still more advantageous in the identification of somatic variants and temporal analysis of antibody maturation for an antibody family with a well-defined germline origin[25]. Our analyses of IAVI donor 17 thus highlight the separate advantages of 5′-RACE and standard PCR methods in the current applications of antibody repertoire analysis.

Comparison of repertoires derived from 5′-RACE PCR and multiplex PCR. Due to differential primer efficiencies and primer cross-reactions, multiplex PCR can cause biases in the sequencing library, thus hampering the reliability of antibody repertoire analysis[22]. Primer bias of multiplex PCR has been previously investigated for T cell receptors (TCRs) using a synthetic repertoire[24]. Here, we examined the primer bias in the context of antibody repertoire analysis by comparing the IAVI donor 17 repertoires generated by 5′-RACE PCR and different sets of gene-specific primers (Table 1C). We redesigned the fusion primers such that the sequencing starts from the 3′-end of the variable domain (Tables S2 and S3, and Fig. S3A) similar to 5′-RACE PCR, with the exception that the 5′-end is now anchored by a V_H or V_L primer.

We first tested two sets of heavy chain primers: (1) primers that overlapped the end of the V-gene leader sequence and the start of the V region and (2) upstream primers that annealed to the start of the V-gene leader sequence and have been optimized to capture highly mutated sequences[32]. The first set of primers (GP-H1) were derived from previous 454 studies of HIV-1 bnAbs[11–15] with the addition of two IgHV5 primers designed based on the same principles. The second set of primers (GP-H2) were previously reported by Scheid et al[32]. We postulated that a less biased primer set would better represent the germline gene usage of the unbiased repertoire. Remarkably, GP-H1 produced a close match to the unbiased repertoire, with a correlation coefficient of 0.96 (Fig. 4A), suggesting that GP-H1 can serve as the first-level approximation to 5′-RACE PCR for heavy chain repertoire analysis. However, GP-H2 showed an extremely skewed usage of IgHV1-69 (79.8%) with a correlation coefficient of 0.17 (Fig. 4B), with IgHV4-59 and IgHV5-51 accounting for only 0.039 and 0.001% of the entire repertoire. We then tested a set of forward λ-chain primers (GP-L1), which yielded a biased germline gene usage with a correlation coefficient of 0.37 (Fig. 4C). In this repertoire, IgLV2-14, rather than IgLV3-1, was the most prevalent germline gene and accounted for ~25.7% of the population. IgLV3-21, the germline precursor of PGT121-class light chains, only accounted for 2.0% of the repertoire, as opposed to a 9.2% in the unbiased repertoire (Fig. 3D).

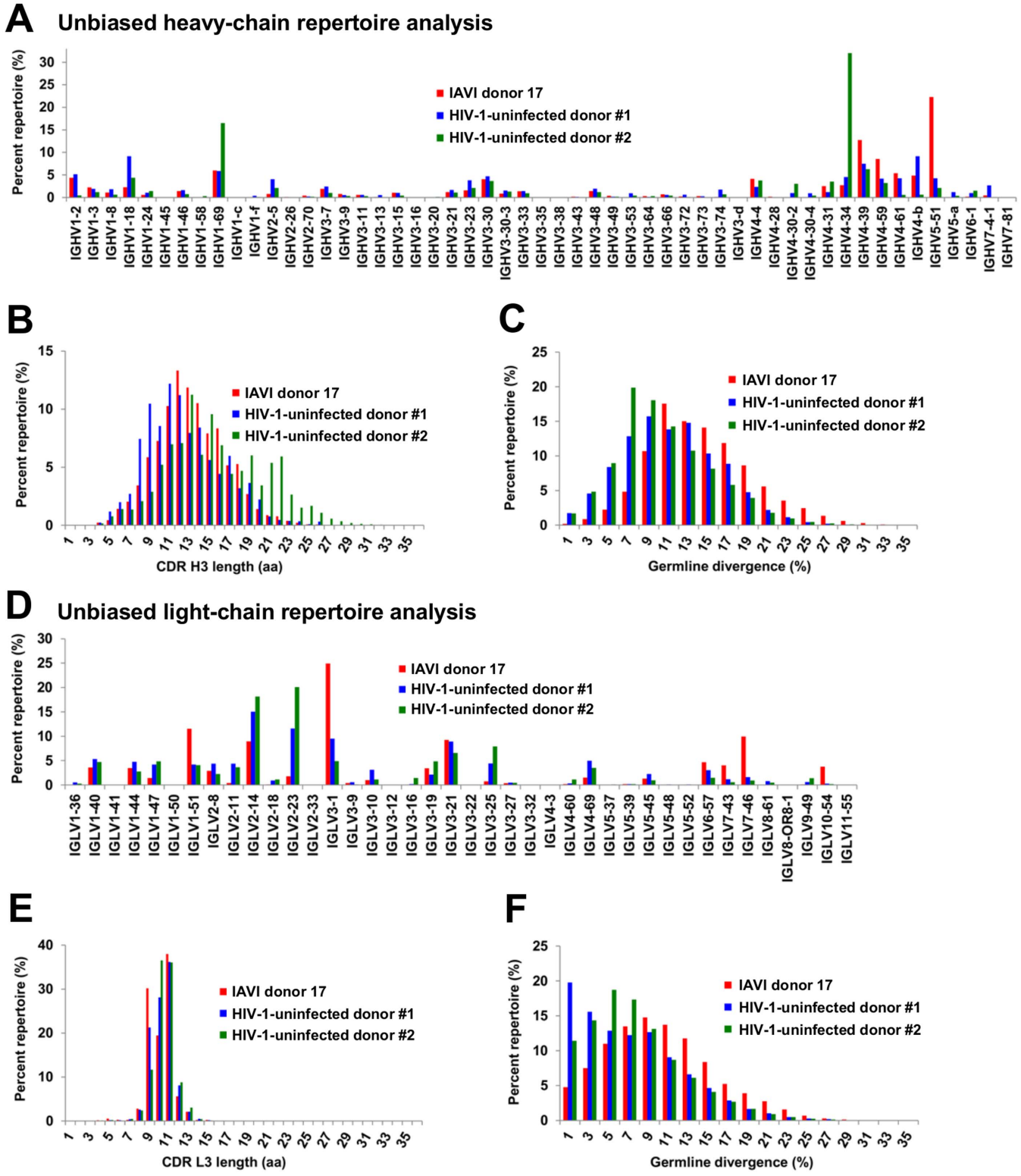

Figure 3 | Unbiased repertoires of IAVI donor 17 and two HIV-1-uninfected donors. Unbiased heavy and light chain repertoires were obtained using 5′-RACE PCR. PGM sequencing was performed using an Ion 316 chip and sequencing data was processed with the *Antibodyomics 1.0* pipeline. The processed sequences were used to calculate heavy (A–C) and light chain (D–F) repertoire properties such as germline gene usage (A and D), complementarity determining region 3 (CDR3) length (B and E), and germline gene divergence (C and F).

The stark difference in germline gene usage between GP-H1 and GP-H2 exemplifies the influence of primer selection upon basic repertoire properties. This comparison further emphasizes the necessity of using 5′-RACE PCR to eliminate primer bias, although there appears to be value in optimizing gene-specific primers and multiplex PCR to minimize bias[24].

Reducing amplification noise and sequencing errors by barcoding cDNA molecules. Antibody repertoire sequencing has been widely used to identify somatic variants and maturation pathways of HIV-1 bnAbs[11–15,25,33]. However, the noise from PCR library amplification combined with sequencing errors can complicate the interpretation of sequence diversity[22] and thus, undermine the reliability of putative intermediates inferred from phylogenetic analysis[11,16]. Recently, two template tagging strategies were proposed to reduce amplification noise in transcriptome[34] and viral RNA[35] sequencing. In this study, we developed a "random barcoding" strategy for antibody sequencing in which 10 degenerate nucleotides (N_{10}) were included in the cDNA synthesis primer such that each template is labeled with a unique identifier (ID) (Fig. S3C). In theory, such random barcodes can create 1,048,576 (4^{10}) distinct sequence IDs, which are comparable to the number of heavy or light chains generated in a typical PGM sequencing run. We sequenced the IgHV4 and IgLV3 gene families of IAVI donor 17 using this random barcoding strategy and observed comparable data quality (Table 1D).

Of ~3.1 million raw reads, 88.1% possessed a random barcode of 10 nucleotides. After pipeline processing, 84.0% of the IgHV4-59 family and 96.3% of the IgLV3-21 family contained the barcodes of correct length (Fig. 5A). 3–12% of the sequences contained an extra nucleotide in the barcode, which was likely caused by errors in cDNA synthesis, PCR amplification or PGM sequencing. We then examined the amplification noise within the germline gene families of the PGT121 class (Fig. 5B). An all-to-all comparison identified 70,235 (42.1%) uniquely barcoded heavy chains from 166,703 IgHV4-59 sequences, with the copy number ranging from 1 to 99. In contrast, only 106,984 (14.2%) out of 754,085 IgLV3-21 sequences were found to be uniquely barcoded, with a maximum copy number of 853. The distribution of identical barcodes did not fit a normal distribution curve (Fig. 5B), suggesting the templates were not amplified equally[35]. In particular, the light chains appeared to be amplified more frequently than the heavy chains, as indicated by a peak population (~19%) of 11–50 copies (Fig. 5B, right panel). Despite the possibility of 4^{10} unique barcodes, different cDNA templates can be labeled by the same barcode sequence. We examined this possibility by using the CDR3 as a secondary sequence ID in the determination of unique templates, namely, two sequences need to have the same barcode and the same CDR3 length (with an error of ±1.5aa) to be considered "identical". Indeed, 1–3% of the sequences were amplified from different templates but assigned with the same barcodes, as indicated by slightly increased single-copy reads (Fig. 5B). This was confirmed by visual inspection of sequences with identical barcodes but different CDR H3/L3 lengths (Fig. S5).

Next we investigated the utility of random barcode to correct sequence errors for PGM-derived PGT121 class of antibodies. 5 iterations of intra-donor phylogenetic analysis[13,14] converged to 2,011 PGT121-class antibody heavy chains. Using random barcodes and the CDR H3 length, 1,105 unique heavy chains were identified. The copy number distribution of the PGT121 heavy chain somatic

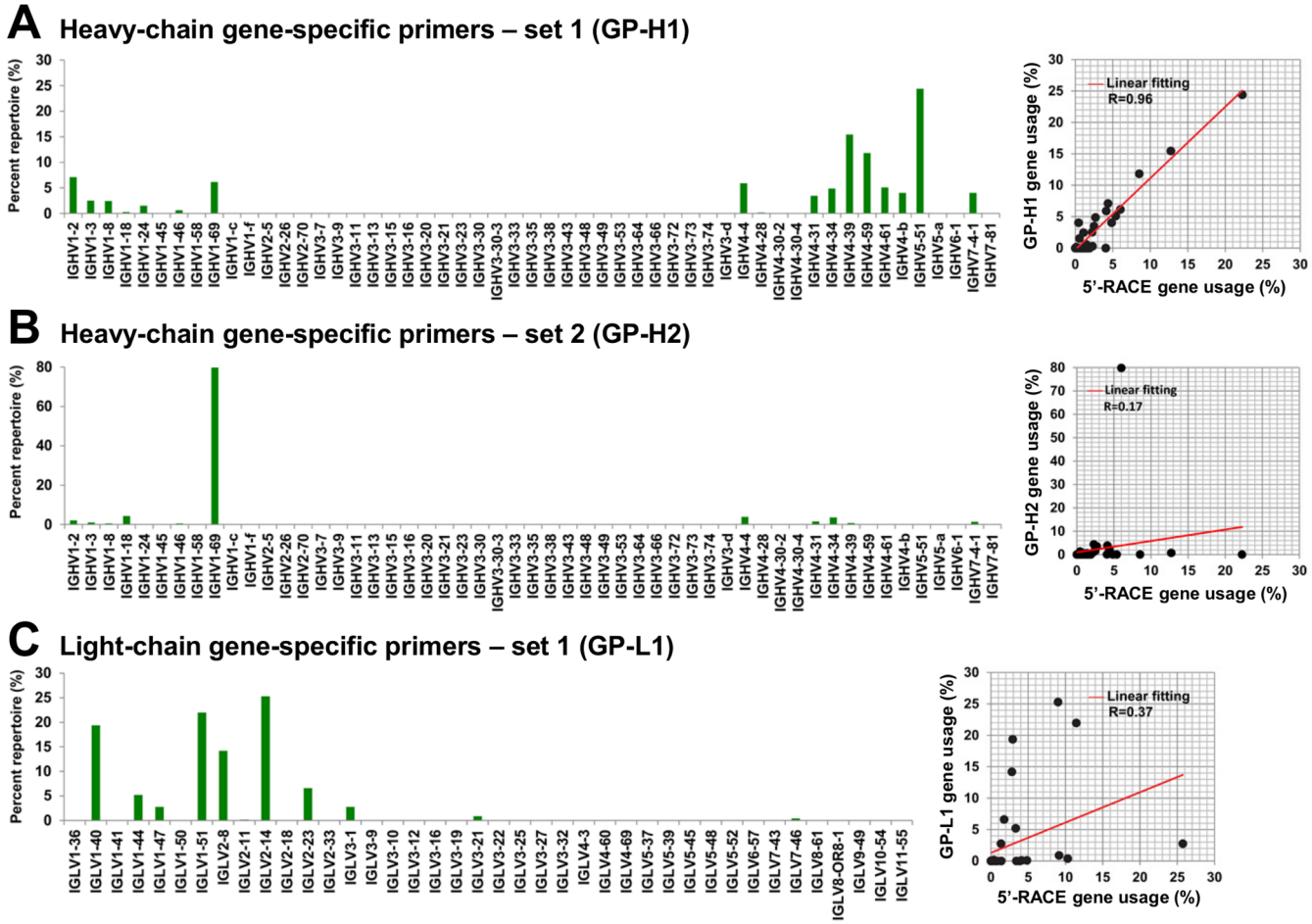

Figure 4 | Comparison of IAVI donor 17 repertoires generated by 5′-RACE PCR and multiplex PCR. (A) Heavy chain germline gene usage (%) from GP-H1 primer set and corresponding correlation with the unbiased repertoire. (B) Heavy chain germline gene usage (%) from GP-H2 primer set and corresponding correlation with the unbiased repertoire. (C) Light chain germline gene usage (%) from GP-L1 primer set and corresponding correlation with the unbiased repertoire. GP-H1 and GP-L1 primers overlap the end of the V-gene leader sequence and the start of the V region, whereas GP-H2 primers anneal more upstream to the start of the V-gene leader sequence and have been optimized to capture highly mutated antibody sequences.

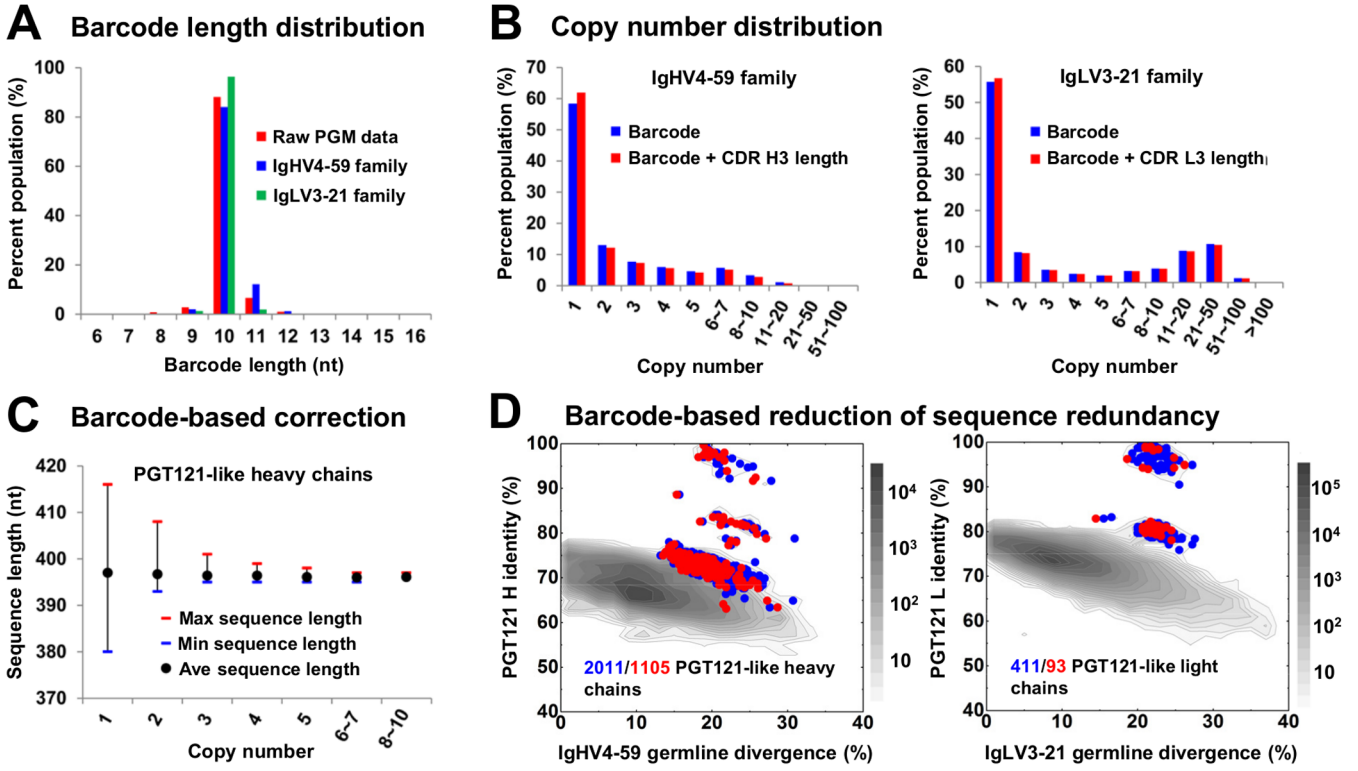

Figure 5 | Random barcoding strategy in the repertoire sequencing of PGT121 class of antibodies. (A) Barcode length distribution for the raw sequencing data (red) and the pipeline-processed IgHV4-59 family (blue) and IgLV3-21 family (green), the germline genes of the PGT121 class of antibodies. Plotted in the distribution are 3,109,512 raw reads, 198,345 IgHV4-59 originated sequences, and 782,720 IgLV3-21 originated sequences. (B) Distribution of copy number for the heavy chains of IgHV4-59 origin (left panel) and light chains of IgLV3-21 origin (right panel) with a correct 10-nt barcode length. Identical cDNA templates were identified using either random barcode alone (blue) or a combination of random barcode and CDR3 length (red). (C) Sequence length variation of PGT121-class heavy chains plotted as a function of copy number. A total of 166,703 heavy chains of IgHV4-59 origin with the correct barcode length were subjected to an iterative intra-donor phylogenetic analysis using 6 PGT121-class heavy chains as a template. After 5 iterations, the analysis converged to 2,011 sequences, which were subjected to further calculation of copy number and length variation. (D) Identity/divergence analysis of PGT121-class antibodies before (blue) and after (red) random barcode-based reduction of sequence redundancy, for heavy (left panel) and light chains (right panel).

variants resembled that of the whole IgHV4-59 family (Fig. 5B, left panel), with 62.9% of the sequences having a single copy. We then calculated the "consensus" sequences for all heavy chains with more than two copies. As expected, the variation of sequence length as a result of PCR or sequencing error decreases significantly as the copy number increases (Fig. 5C). The average sequence length decreased from 397.2 to 396 bp, which is the correct sequence length of PGT121-class heavy chains. The barcode-corrected PGT121-like sequences showed reduced diversity on the identity/divergence plots (Fig. 5D).

Taken together, the random barcoding strategy can quantify the amplification bias in antibody repertoire sequencing and thus provide an effective means to reduce potential artifacts resulting from PCR-based amplification noise and sequencing errors. This strategy is general and can be applied to longer barcodes as demonstrated for a 20-nucleotide barcode (Fig. S6). We also demonstrated that the CDR3, the most conserved antibody signature, can be used as a "natural barcode" to assist in the analysis of sequence redundancy. It should be noted that each B cell can carry multiple copies of mRNA, which can be labeled with different barcodes in RT and treated as non-redundant cDNA molecules. Therefore, the current barcoding strategy can only eliminate the redundancy of the expressed antibody repertoire rather than that of the B-cell repertoire.

Improving antibody repertoire quality with new NGS technologies. Pyrosequencing has not been favored for antibody repertoire analysis primarily due to homopolymer errors[7]. Therefore, it is important to assess whether the recent technical advances for PGM can improve sequencing accuracy. Here, we have evaluated three new PGM technologies that were made available to academic users through the Early Access program. These include two template preparation methods – an improved version of the emulsion-based method and an emulsion-free method called isothermal amplification (IA) – and the Hi-Q sequencing enzyme. We tested various combinations of the template preparation methods, Hi-Q enzyme and data processing methods using the 5′-RACE PCR products from uninfected donor #2 (Table 1E).

The combined use of the improved emulsion-based method and Hi-Q enzyme (new OT2 + Hi-Q) showed a remarkable improvement consisting of a 16% increase in the number of raw reads, a 50–60 bp increase in read length, and a 15–20% increase in sequence population without gaps in V-gene alignment (from 8.3 to 28.4% and 18.1 to 33.7% for the heavy and light chains, respectively) (Table 1E, #8). The change of error profile can be visualized by the distribution of gaps in the V-gene alignment (Fig. 6A). The "error-free" sequences, along with those with only one gap in the V gene, have shifted the repertoire towards a lower error rate. We then investigated whether 3′-trimming, which was turned off in previous PGM sequencing, and bioinformatics filtering can further improve the accuracy. Indeed, 3′-trimming did increase the V-gene error-free population by 10% but with a tradeoff of 20% decrease in sequence reads (Table 1E, #9). After bioinformatics filtering, 40.4% of the heavy chains and 45.8% of the light chains contained no indel errors in the V gene segment, respectively (Fig. 6A).

The combined use of IA and Hi-Q (IA + Hi-Q) showed further improvement on all the metrics examined. The sequencing output was increased by 26% with respect to the standard PGM protocol, with the heavy and light chains showing identical values in read length (560 bp) and sequence quality (33.3% V-gene error-free) (Table 1E, #10). These results confirmed the high fidelity of sequencing templates generated by IA. With 3′-trimming, although there was a 20% decrease in sequence reads, ~41% of the entire repertoire was error-free (Table 1E, #11). Bioinformatics filtering further increased this error-free population to ~47% of the entire repertoire (Fig. 6B).

The significant improvements in sequence quality demonstrate the crucial role of NGS technology in antibody repertoire analysis. The reduced homopolymer errors from the combined use of IA and Hi-Q will further increase the accuracy of antibody lineage analysis and intermediate inference based on the PGM platform.

Discussion

The extraordinary ability of antibodies to recognize the plethora of foreign pathogens relies on their sequence diversity generated by gene rearrangement and affinity maturation[36,37]. NGS-based repertoire analysis is poised to further our understanding of humoral immunity[5] and to accelerate antibody discovery and vaccine design[26,30,38–41]. The promises and challenges in this emerging field have been reviewed in length[5–7,22]. Using samples from a unique HIV-1-infected donor and two uninfected donors, we examined several critical issues in antibody repertoire analysis. (1) Longer reads. Although sequencing the CDR3 may be sufficient in characterizing the antibody response for some pathogens (e.g. human dengue virus[42]), sequencing the entire V(D)J-coding region has become a prerequisite for antibody repertoire analysis. Here we demonstrate that the PGM platform can sequence the entire variable domain with an average read length of 550 bp at an estimated 1% cost of the 454 platform. The sequencing protocols, heavy and light chain primer sets, and bioinformatics pipeline validated in this study provide a set of practical solutions for antibody repertoire analysis based on this platform. (2) Biased vs. unbiased. Gene-specific primers may cause significant bias and thus are not optimal for tracing dynamic antibody responses *de novo* during natural infection or vaccination. We addressed this issue by adopting 5′-RACE PCR in template preparation, which allowed us to analyze HIV-1 bnAbs from a unique patient sample in the context of the entire repertoire. Using the unbiased repertoire as a reference, we quantified the bias generated by various primer sets currently used in antibody repertoire analysis. (3) Artifacts caused by PCR-based amplification. In a recent review, amplification noise was noted as a major problem in repertoire analysis[22]. Molecular tagging has been used to deal with such noise in transcriptome[34] and viral RNA[35] sequencing but not yet extended to immune repertoire sequencing. In this study, we devised a random barcoding strategy to quantify the amplification bias in the analysis of PGT121 class of antibodies. We also examined the utility of this strategy to correct sequencing errors. Such a strategy will benefit the in-depth analysis of HIV-1-infected donor samples with experimentally isolated bnAbs[11–16,33]. Since this strategy can only remove redundancy at the cDNA level, genomic sequencing of the immunoglobulin gene loci after isotype-specific B cell purification by flow

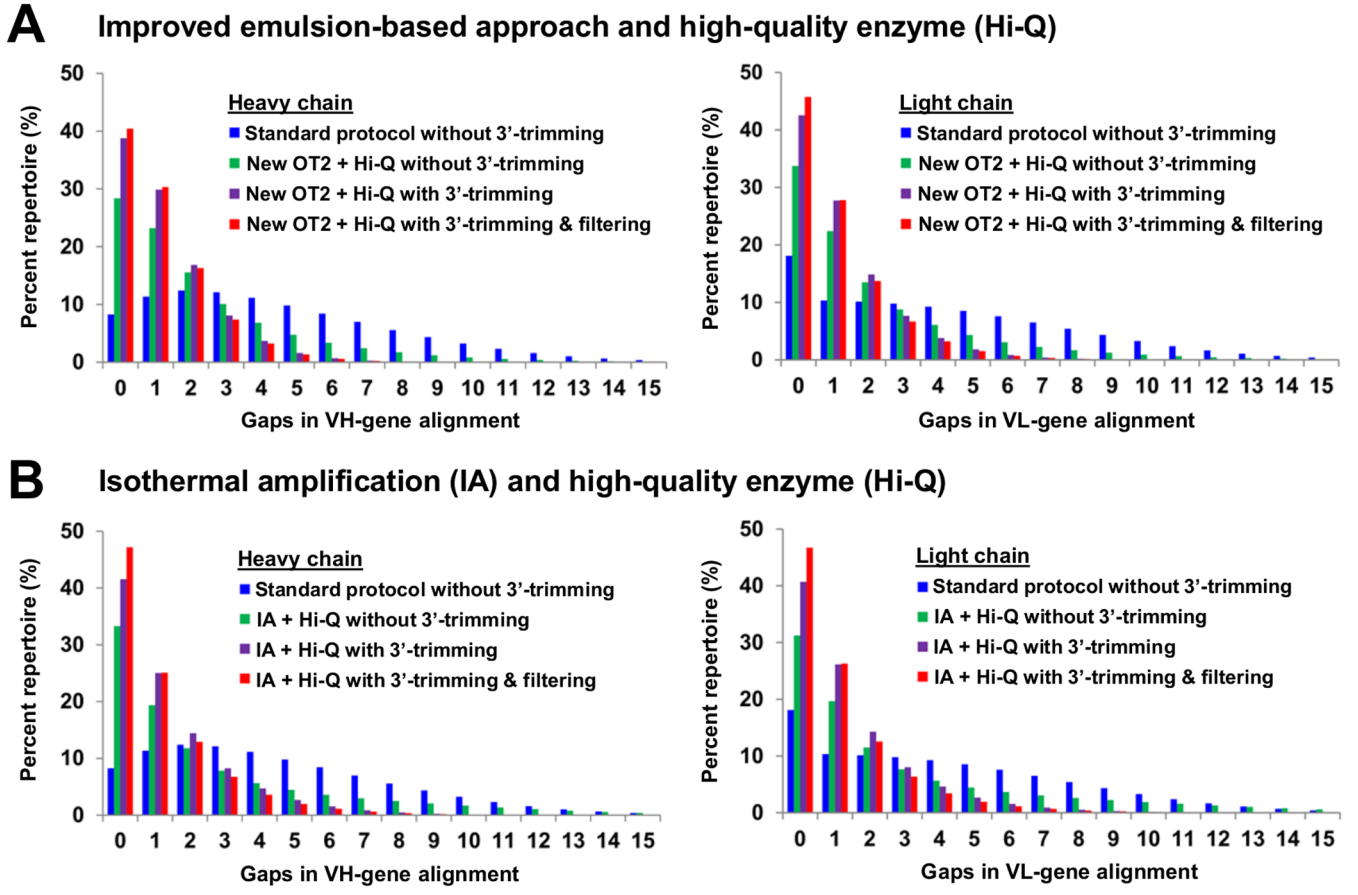

Figure 6 | Improved antibody sequence quality from new PGM technologies. Number of gaps in VH- and VL-gene alignment is plotted for (A) the combined use of improved emulsion-based template preparation method and Hi-Q enzyme (new OT2 + Hi-Q) and (B) the combined use of emulsion-free isothermal amplification (IA) and Hi-Q enzyme (IA + Hi-Q). The template library from uninfected donor #2 was sequenced as a test case. Plotted are the standard PGM protocol without 3′-trimming (blue), the combined use of a new template preparation method and Hi-Q enzyme without 3′-trimming (green), the same combination with 3′-trimming (magenta), and the same combination with 3′-trimming and bioinformatics filtering (red).

cytometry or other approaches may be required to detect redundancy at the mRNA level[10,22]. (4) Improved NGS technologies. As NGS technologies continue to mature, new advances based on the available platforms will likely generate a direct impact on repertoire analysis. For the PGM platform, we have demonstrated improved throughput, read length and sequence quality from the combined use of new template preparation methods and sequencing chemistry. Together, the technology assessment and development described in this study will help establish a more rigorous foundation for antibody repertoire analysis in biomedical research.

Methods

Human specimens. Peripheral blood mononuclear cells (PBMCs) were obtained from donor 17, an HIV-1 infected donor from the IAVI Protocol G cohort[43]. All human samples were collected with written informed consent under clinical protocols approved by the Republic of Rwanda National Ethics Committee, the Emory University Institutional Review Board, the University of Zambia Research Ethics Committee, the Charing Cross Research Ethics Committee, the UVRI Science and Ethics Committee, the University of New South Wales Research Ethics Committee. St. Vincent's Hospital and Eastern Sydney Area Health Service, Kenyatta National Hospital Ethics and Research Committee, University of Cape Town Research Ethics Committee, the International Institutional Review Board, the Mahidol University Ethics Committee, the Walter Reed Army Institute of Research (WRAIR) Institutional Review Board, and the Ivory Coast Comite´ National d'Ethique des Sciences de la Vie et de la Sante´ (CNESVS). The sample from IAVI donor 17 was the source of broadly neutralizing antibodies PGT121–124 and PGT133–134[19]. The PBMCs of two HIV-1-uninfected donors were obtained from the California Blood Bank according to the Institutional Review Board (IRB) at The Scripps Research Institute. The blood samples were collected with written informed consent from the donors.

Sample preparation using gene-specific primers. Total RNA was extracted from 20 million PBMCs into 30 μl of water with TRIzol Reagent (Life Technologies). The reverse transcription (RT) was performed with SuperScript III (Life Technologies) and oligo(dT)12–18. The cDNA was purified and eluted in 20 μl of elution buffer (NucleoSpin PCR Clean-up Kit, Clontech). The immunoglobulin gene-specific PCRs were performed with Platinum *Taq* High-Fidelity DNA Polymerase (Life Technologies) in a total volume of 50 μl, with 5 μl of cDNA as template, 1 μl of gene-specific primers and 1 μl of 10 μM reverse primer. The primers each contained an appropriate adaptor sequence (A or trP1) for subsequent PGM sequencing. Two sequencing directions (Fig. S3A) and their respective primer sets were designed (Tables S1–S3). 25 cycles of PCRs were performed and the expected PCR products (~500 bp) were gel purified (Qiagen).

Sample preparation using 5′-RACE PCR. After total RNA extraction, 5′-RACE was performed with FirstChoice® RLM-RACE Kit (Life Technologies) and oligo(dT)12–18. The immunoglobulin PCRs were set up in a total volume of 50 μl, with 5 μl of cDNA as template, 1 μl of 5′-RACE primer and 1 μl of 10 μM reverse primer. The 5′-RACE primer contained PGM trP1 or P1 adaptor (P1 is required for isothermal amplification [IA]), while the reverse primer contained a PGM A adaptor (Fig. S3B and Table S4). 25 cycles of PCRs were performed and the expected PCR products (~600 bp) were gel purified (Qiagen).

Sample preparation using gene-specific primers with random barcodes. Total RNA extraction was performed using the same protocol as above. A random barcode of ten degenerate nucleotides was inserted between PGM A adaptor and the reverse primer in the constant domain. RT was performed with SuperScript III (Life Technologies) and the barcoded primers. After cDNA purification, the immunoglobulin gene-specific PCRs were set up in a total volume of 50 μl, with 5 μl of cDNA as template, 1 μl of forward gene-specific primers and 1 μl of 10 μM PGM A adaptor. The forward gene-specific primers each contained a PGM trP1 adaptor for subsequent PGM sequencing (Fig. S3C and Table S2). 25 cycles of PCRs were performed and the expected PCR products (~500 bp) were gel purified (Qiagen).

Ion Torrent PGM sequencing of antibody libraries. The antibody heavy- and light-chain libraries were quantitated using Qubit® 2.0 Fluorometer with Qubit® dsDNA HS Assay Kit, and then used at a ratio of 1 : 1 except for the first sequencing experiment, in which a ratio of 1 : 2 was used. The dilution factor required for Ion Torrent PGM template preparation was determined such that the final concentration was 30 pM. The template preparation was performed with either Ion PGM Template OT2 400 Kit on the Ion OneTouch 2 Instrument overnight or the IA Kit. Template enrichment was performed on the Ion OneTouch ES Instrument the following day. Prior to PGM sequencing, quality control of the template was determined by the Qubit® 2.0 Fluorometer with the Ion Sphere™ Quality Control Kit. Sequencing was performed on the Ion PGM System with the Ion PGM™ Sequencing 400 Kit or PGM™ Hi-Q 400 Kit using either an Ion 316 or 318 v2 chip for a total of 850 nucleotide flows (1,100 flows when IA was used). Raw data processing with and without the 3′-end trimming in base calling was compared when evaluating new PGM technologies.

Bioinformatics analysis of antibody sequencing data. The *Antibodyomics 1.0* pipeline described in our previous studies[11–15] was used to process all NGS data. After full-length variable domain sequences were obtained, a new filter was used to detect and remove erroneous sequences that may contain swapped gene segments from PCR errors. Specifically, a full-length read was removed from the data set if the V-gene alignment was less than 250 bp (220 bp in the case of IAVI donor 17 light chains). In this study, a modified procedure for the intra-donor phylogenetic analysis[13,14] was used to analyze IAVI donor 17 sequences. Two changes were made to improve the accuracy and computational efficiency. Firstly, the neighbor-joining (NJ) method used previously was replaced with the maximum likelihood (ML) method. Secondly, the extraction of somatic variants was automated by using a program to recognize evolutionarily related sequences that reside on the same phylogenetic branch as the input template sequences (e.g. PGT121 class of antibodies).

Antibody expression. Antibody production was performed as previously described[11–15]. Briefly, the bioinformatically selected antibody chain sequences were synthesized (GenScript, Inc) and cloned into the CMV/R expression vector containing the constant regions of IgG. The heavy and light chains identified from IAVI donor 17 PGM sequencing data were paired with their respective partner chain DNAs from the PGT121-class antibodies. Full-length IgGs were expressed by transient transfection of 293F cells and purified using a recombinant protein-A column (Pierce). The expression and sequence information of PGM-derived antibodies are summarized in Table S5.

HIV-1 neutralization assays. Neutralization assays were performed on TZM-bl reporter cells using a six-virus panel as previously described[44–46]. A six-virus panel was used in this study. Neutralization curves were fit by a nonlinear regression analysis using a 5-parameter hill slope equation. The 50% inhibitory concentration (IC_{50}) is defined as the antibody concentration required to inhibit HIV-1 infection by 50%.

1. Mardis, E. R. The impact of next-generation sequencing technology on genetics. *Trends Genet.* **24**, 133–141, doi:10.1016/j.tig.2007.12.007 (2008).
2. Mardis, E. R. Next-generation DNA sequencing methods. *Annu. Rev. Genom. Hum. Genet.* **9**, 387–402, doi:10.1146/annurev.genom.9.081307.164359 (2008).
3. Metzker, M. L. Sequencing technologies - the next generation. *Nat. Rev. Genet.* **11**, 31–46, doi:10.1038/nrg2626 (2010).
4. Shendure, J. & Ji, H. Next-generation DNA sequencing. *Nat. Biotechnol.* **26**, 1135–1145, doi:10.1038/nbt1486 (2008).
5. Reddy, S. T. & Georgiou, G. Systems analysis of adaptive immunity by utilization of high-throughput technologies. *Curr. Opin. Biotechnol.* **22**, 584–589, doi:10.1016/j.copbio.2011.04.015 (2011).
6. Fischer, N. Sequencing antibody repertoires The next generation. *Mabs* **3**, 17–20, doi:10.4161/mabs.3.1.14169 (2011).
7. Georgiou, G. *et al.* The promise and challenge of high-throughput sequencing of the antibody repertoire. *Nat. Biotechnol.* **32**, 1–11, doi:10.1038/nbt.2782 (2014).
8. Weinstein, J. A., Jiang, N., White, R. A., III, Fisher, D. S. & Quake, S. R. High-Throughput Sequencing of the Zebrafish Antibody Repertoire. *Science* **324**, 807–810, doi:10.1126/science.1170020 (2009).
9. Reddy, S. T. *et al.* Monoclonal antibodies isolated without screening by analyzing the variable-gene repertoire of plasma cells. *Nat. Biotechnol.* **28**, 965–969, doi:10.1038/nbt.1673 (2010).
10. Boyd, S. D. *et al.* Measurement and Clinical Monitoring of Human Lymphocyte Clonality by Massively Parallel V-D-J Pyrosequencing. *Sci. Transl. Med.* **1**, 12ra23, doi:10.1126/scitranslmed.3000540 (2009).
11. Wu, X. *et al.* Focused Evolution of HIV-1 Neutralizing Antibodies Revealed by Structures and Deep Sequencing. *Science* **333**, 1593–1602, doi:10.1126/science.1207532 (2011).
12. Zhou, T. *et al.* Multidonor analysis reveals structural elements, genetic determinants, and maturation pathway for HIV-1 neutralization by VRC01-class antibodies. *Immunity* **39**, 245–258, doi:10.1016/j.immuni.2013.04.012 (2013).
13. Zhu, J. *et al.* Somatic Populations of PGT135-137 HIV-1-Neutralizing Antibodies Identified by 454 Pyrosequencing and Bioinformatics. *Front. Microbiol.* **3**, 315–315, doi:10.3389/fmicb.2012.00315 (2012).
14. Zhu, J. *et al.* Mining the antibodyome for HIV-1-neutralizing antibodies with next-generation sequencing and phylogenetic pairing of heavy/light chains. *Proc. Natl. Acad. Sci. USA* **110**, 6470–6475, doi:10.1073/pnas.1219320110 (2013).
15. Zhu, J. *et al.* De novo identification of VRC01 class HIV-1-neutralizing antibodies by next-generation sequencing of B-cell transcripts. *Proc. Natl. Acad. Sci. USA* **110**, E4088–E4097, doi:10.1073/pnas.1306262110 (2013).
16. Sok, D. *et al.* The Effects of Somatic Hypermutation on Neutralization and Binding in the PGT121 Family of Broadly Neutralizing HIV Antibodies. *PLoS Pathog.* **9**, e1003754, doi:10.1371/journal.ppat.1003754 (2013).
17. Huang, J. *et al.* Broad and potent neutralization of HIV-1 by a gp41-specific human antibody. *Nature* **491**, 406–412, doi:10.1038/nature11544 (2012).
18. Walker, L. M. *et al.* Broad and Potent Neutralizing Antibodies from an African Donor Reveal a New HIV-1 Vaccine Target. *Science* **326**, 285–289, doi:10.1126/science.1178746 (2009).
19. Walker, L. M. *et al.* Broad neutralization coverage of HIV by multiple highly potent antibodies. *Nature* **477**, 466–470, doi:10.1038/nature10373 (2011).
20. Zhou, T. *et al.* Structural Basis for Broad and Potent Neutralization of HIV-1 by Antibody VRC01. *Science* **329**, 811–817, doi:10.1126/science.1192819 (2010).

21. Deng, W. *et al.* Indel and Carryforward Correction (ICC): a new analysis approach for processing 454 pyrosequencing data. *Bioinformatics* **29**, 2402–2409, doi:10.1093/bioinformatics/btt434 (2013).
22. Baum, P. D., Venturi, V. & Price, D. A. Wrestling with the repertoire: The promise and perils of next generation sequencing for antigen receptors. *Eur. J. Immunol.* **42**, 2834–2839, doi:10.1002/eji.201242999 (2012).
23. Bolotin, D. A. *et al.* Next generation sequencing for TCR repertoire profiling: Platform-specific features and correction algorithms. *Eur. J. Immunol.* **42**, 3073–3083, doi:10.1002/eji.201242517 (2012).
24. Carlson, C. S. *et al.* Using synthetic templates to design an unbiased multiplex PCR assay. *Nat. Commun.* **4**, 2680, doi:10.1038/ncomms3680 (2013).
25. Doria-Rose, N. A. *et al.* Developmental pathway for potent V1V2-directed HIV-neutralizing antibodies. *Nature* **509**, 55–62, doi:10.1038/nature13036 (2014).
26. Kwong, P. D., Mascola, J. R. & Nabel, G. J. Broadly neutralizing antibodies and the search for an HIV-1 vaccine: the end of the beginning. *Nat. Rev. Immunol.* **13**, 693–701, doi:10.1038/nri3516 (2013).
27. Julien, J.-P. *et al.* Broadly Neutralizing Antibody PGT121 Allosterically Modulates CD4 Binding via Recognition of the HIV-1 gp120 V3 Base and Multiple Surrounding Glycans. *PLoS Pathog.* **9**, e1003342, doi:10.1371/journal.ppat.1003342 (2013).
28. Barouch, D. H. *et al.* Therapeutic efficacy of potent neutralizing HIV-1-specific monoclonal antibodies in SHIV-infected rhesus monkeys. *Nature* **503**, 224–228, doi:10.1038/nature12744 (2013).
29. Jardine, J. *et al.* Rational HIV Immunogen Design to Target Specific Germline B Cell Receptors. *Science* **340**, 711–716, doi:10.1126/science.1234150 (2013).
30. Haynes, B. F., Kelsoe, G., Harrison, S. C. & Kepler, T. B. B-cell-lineage immunogen design in vaccine development with HIV-1 as a case study. *Nat. Biotechnol.* **30**, 423–433, doi:10.1038/nbt.2197 (2012).
31. Choi, N. M. *et al.* Deep Sequencing of the Murine Igh Repertoire Reveals Complex Regulation of Nonrandom V Gene Rearrangement Frequencies. *J. Immunol.* **191**, 2393–2402, doi:10.4049/jimmunol.1301279 (2013).
32. Scheid, J. F. *et al.* Sequence and Structural Convergence of Broad and Potent HIV Antibodies That Mimic CD4 Binding. *Science* **333**, 1633–1637, doi:10.1126/science.1207227 (2011).
33. Liao, H.-X. *et al.* Co-evolution of a broadly neutralizing HIV-1 antibody and founder virus. *Nature* **496**, 469–476, doi:10.1038/nature12053 (2013).
34. Shiroguchi, K., Jia, T. Z., Sims, P. A. & Xie, X. S. Digital RNA sequencing minimizes sequence-dependent bias and amplification noise with optimized single-molecule barcodes. *Proc. Natl. Acad. Sci. USA* **109**, 1347–1352, doi:10.1073/pnas.1118018109 (2012).
35. Jabara, C. B., Jones, C. D., Roach, J., Anderson, J. A. & Swanstrom, R. Accurate sampling and deep sequencing of the HIV-1 protease gene using a Primer ID. *Proc. Natl. Acad. Sci. USA* **108**, 20166–20171, doi:10.1073/pnas.1110064108 (2011).
36. Iwasaki, A. & Medzhitov, R. Regulation of Adaptive Immunity by the Innate Immune System. *Science* **327**, 291–295, doi:10.1126/science.1183021 (2010).
37. Litman, G. W., Rast, J. P. & Fugmann, S. D. The origins of vertebrate adaptive immunity. *Nat. Rev. Immunol.* **10**, 543–553, doi:10.1038/nri2807 (2010).
38. Koff, W. C. *et al.* Accelerating Next-Generation Vaccine Development for Global Disease Prevention. *Science* **340**, 1232910, doi:10.1126/science.1232910 (2013).
39. Burton, D. R. *et al.* A Blueprint for HIV Vaccine Discovery. *Cell Host Microbe* **12**, 396–407, doi:10.1016/j.chom.2012.09.008 (2012).
40. Mascola, J. R. & Montefiori, D. C. The Role of Antibodies in HIV Vaccines. *Annu. Rev. Immunol.* **28**, 413–444, doi:10.1146/annurev-immunol-030409-101256 (2010).
41. Klein, F. *et al.* Antibodies in HIV-1 Vaccine Development and Therapy. *Science* **341**, 1199–1204, doi:10.1126/science.1241144 (2013).
42. Parameswaran, P. *et al.* Convergent antibody signatures in human dengue. *Cell Host Microbe* **13**, 691–700, doi:10.1016/j.chom.2013.05.008 (2013).
43. Simek, M. D. *et al.* Human Immunodeficiency Virus Type 1 Elite Neutralizers: Individuals with Broad and Potent Neutralizing Activity Identified by Using a High-Throughput Neutralization Assay together with an Analytical Selection Algorithm. *J. Virol.* **83**, 7337–7348, doi:10.1128/jvi.00110-09 (2009).
44. Li, M. *et al.* Human immunodeficiency virus type 1 env clones from acute and early subtype B infections for standardized assessments of vaccine-elicited neutralizing antibodies. *J. Virol.* **79**, 10108–10125, doi:10.1128/jvi.79.16.10108-10125.2005 (2005).
45. Seaman, M. S. *et al.* Tiered Categorization of a Diverse Panel of HIV-1 Env Pseudoviruses for Assessment of Neutralizing Antibodies. *J. Virol.* **84**, 1439–1452, doi:10.1128/jvi.02108-09 (2010).
46. Wu, X. *et al.* Mechanism of Human Immunodeficiency Virus Type 1 Resistance to Monoclonal Antibody b12 That Effectively Targets the Site of CD4 Attachment. *J. Virol.* **83**, 10892–10907, doi:10.1128/jvi.01142-09 (2009).

Acknowledgments

We would like to thank the members of IAVI who participated or provided donor samples in this project. IAVI's work is made possible by generous support from many donors including: the Bill and Melinda Gates Foundation; the Ministry of Foreign Affairs of Denmark; Irish Aid; the Ministry of Finance of Japan; the Ministry of Foreign Affairs of the Netherlands; the Norwegian Agency for Development Cooperation (NORAD); the United Kingdom Department for International Development (DFID), and the United States Agency for International Development (USAID). The full list of IAVI donors is available at www.iavi.org. This work is made possible by the generous support of the American people through USAID. The contents are the responsibility of the authors and do not necessarily reflect the views of USAID or the United States Government. Support for this work was provided by the grants from the Scripps Center for HIV/AIDS Immunology & Immunogen Discovery (CHAVI-ID UM1 AI 100663).

Author contributions

L.L.H. and J.Z. designed research and analyzed the data; L.L.H. and J.Z. performed the research, with the template preparation and PGM sequencing protocols devised and performed by L.L.H. and bioinformatics analysis devised and performed by L.L.H. and J.Z., reconstitution of synthesized antibodies and purification by P.A. and L.L.H., cDNA, PCR and other sample preparation for IAVI donor 17 and two HIV-1-uninfected donors by L.L.H. and Ion Torrent PGM sequencing by L.L.H., HIV-1 neutralization performed and analyzed by D.S. and J.H., E.L., P.P., M.S., D.R.B. and W.A.K. contributed IAVI donor 17 material, D.S. and D.R.B. contributed HIV-1-uninfected donor materials. L.L.H. and J.Z. wrote the paper, with all principal investigators providing comments or revisions.

Additional information

Supplementary information accompanies this paper at http://www.nature.com/scientificreports

Competing financial interests: The authors declare no competing financial interests.

How to cite this article: He, L. *et al.* Toward a more accurate view of human B-cell repertoire by next-generation sequencing, unbiased repertoire capture and single-molecule barcoding. *Sci. Rep.* **4**, 6778; DOI:10.1038/srep06778 (2014).