mAbs 6:1, 160–172; January/February 2014; © 2014 Landes Bioscience

The antibody mining toolbox

An open source tool for the rapid analysis of antibody repertoires

Sara D'Angelo[1], Jacob Glanville[2], Fortunato Ferrara[1], Leslie Naranjo[3], Cheryl D Gleasner[3], Xiaohong Shen[3], Andrew RM Bradbury[3], and Csaba Kiss[3,*]

[1]New Mexico Consortium; Los Alamos, NM, USA; [2]Stanford Immunology; Stanford University; Stanford, CA USA; [3]B Division; Los Alamos National Laboratory; Los Alamos, NM USA

Keywords: HCDR3, antibody library, deep sequencing, regular expression, AbMining ToolBox

Abbreviations: CDR, complementarity determining regions; VH, heavy chain variable domain; VL, light chain variable domain, scFv, single chain fragment variable; RegEx, regular expression; Ab, antibody; aa, amino acid

In vitro selection has been an essential tool in the development of recombinant antibodies against various antigen targets. Deep sequencing has recently been gaining ground as an alternative and valuable method to analyze such antibody selections. The analysis provides a novel and extremely detailed view of selected antibody populations, and allows the identification of specific antibodies using only sequencing data, potentially eliminating the need for expensive and laborious low-throughput screening methods such as enzyme-linked immunosorbant assay. The high cost and the need for bioinformatics experts and powerful computer clusters, however, have limited the general use of deep sequencing in antibody selections. Here, we describe the AbMining ToolBox, an open source software package for the straightforward analysis of antibody libraries sequenced by the three main next generation sequencing platforms (454, Ion Torrent, MiSeq). The ToolBox is able to identify heavy chain CDR3s as effectively as more computationally intense software, and can be easily adapted to analyze other portions of antibody variable genes, as well as the selection outputs of libraries based on different scaffolds. The software runs on all common operating systems (Microsoft Windows, Mac OS X, Linux), on standard personal computers, and sequence analysis of 1–2 million reads can be accomplished in 10–15 min, a fraction of the time of competing software. Use of the ToolBox will allow the average researcher to incorporate deep sequence analysis into routine selections from antibody display libraries.

Introduction

The selection of antibodies using in vitro methods, including phage,[1] yeast[2] and ribosome[3] display has transformed the generation of therapeutic antibodies,[4] and promises to do the same for research-quality antibodies.[5,6] In particular, the ability to improve affinity,[7,8] and select antibodies lacking cross-reactivity to closely related proteins[5,6] can be performed relatively easily using in vitro methods, but requires extensive screening when traditional methods are used to generate monoclonal antibodies.

Until recently, the analysis of such antibody display libraries has been performed in a relatively blind fashion, with a moderately small number (96–384) of randomly picked clones being analyzed by enzyme-linked immunosorbant assay after the selection is complete, to identify binders for the target of interest. In phage and ribosome display, this is the only point at which concrete information on antibody activity can be obtained during a selection, and is the last step of the selection.

Antibodies are best characterized by full sequencing of the VH and VL domains. In the single chain fragment variable (scFv) format, this requires reads of at least 800 base pair (bp), which is only obtainable with high quality Sanger sequencing.[9] The complementarity-determining regions (CDRs) of an antibody are the hypervariable loops responsible for binding to antigen, of which the heavy chain CDR3 (HCDR3) is the most diverse, and widely used as a surrogate for VH and scFv identity.[10-12] HCDR3s are generated by the random combination of germline V, D and J genes,[13,14] with additional junctional diversity created by nucleotide addition or loss (for a review see ref. 15–17), and subsequent targeted somatic hypermutation.[18,19] As opposed to full-length scFv, the identification of specific HCDR3s requires far shorter reads, and provides a minimum assessment of diversity, in that VH domains with the same HCDR3 may contain additional differences elsewhere in the VH, or they may be paired with different light chains. In general, it is the HCDR3 that provides antibodies with their primary specificity.[11,20]

*Correspondence to: Csaba Kiss; Email: csaba.kiss@lanl.gov
Submitted: 11/01/2013; Revised: 11/04/2013; Accepted: 11/06/2013
http://dx.doi.org/10.4161/mabs.27105

Table 1. List of all primers used for sequencing

Primer ID	Platform	Sequence
454-for	454	CGTATCGCCTCCCTCGCGCCATCAGATGTATACTATACGAAGTTATCCTCGAG
454-MID1-rev	454	CTATGCGCCTTGCCAGCCCGCTCAGACGAGTGCGTGCAGTGGGTTTGGGATTGGTTTGCC
Ion_fw3.vh1	Ion Torrent	CCTCTCTATGGGCAGTCGGTGATTCTACAGACACAGCCTACATGGAGC
Ion_fw3.vh1b	Ion Torrent	CCTCTCTATGGGCAGTCGGTGATACGAGCACAGCCTACATGGAGC
Ion_fw3.vh1c	Ion Torrent	CCTCTCTATGGGCAGTCGGTGATTACATGGAGCTGAGCAGCCTGAG
Ion_fw3.vh2	Ion Torrent	CCTCTCTATGGGCAGTCGGTGATATGACCAACATGGACCCTGTGGAC
Ion_fw3.vh3	Ion Torrent	CCTCTCTATGGGCAGTCGGTGATCCAGAGACAATTCCAAGAACACGC
Ion_fw3.vh3b	Ion Torrent	CCTCTCTATGGGCAGTCGGTGATTGCAAATGAACAGCCTGAAAACCGAGG
Ion_fw3.vh4	Ion Torrent	CCTCTCTATGGGCAGTCGGTGATAACCAGTTCTCCCTGAAGCTGAGC
Ion_fw3.vh5	Ion Torrent	CCTCTCTATGGGCAGTCGGTGATAGTGGAGCAGCCTGAAGGCC
Ion_fw3.vh3c	Ion Torrent	CCTCTCTATGGGCAGTCGGTGATATCTGCAAATGAACAGYCTGAGAGC
Ion_fw3.vh3d	Ion Torrent	CCTCTCTATGGGCAGTCGGTGATAGAGACAATTCCAGGAACWYCCTG
Ion_fw3.vh7	Ion Torrent	CCTCTCTATGGGCAGTCGGTGATCCWTGGACACCTCTGYCAGC
IGHV1–2	Ion Torrent	CCTCTCTATGGGCAGTCGGTGATATCAGCACAGCCTACATGGAGCTG
Ion_IGHV1–68	Ion Torrent	CCTCTCTATGGGCAGTCGGTGATTGAGGACAGCCTACATAGAGCTGAG
Ion_IGHV3–13	Ion Torrent	CCTCTCTATGGGCAGTCGGTGATTCAAATGAACAGCCTGAGAGCCGG
Ion_IGHV3–43	Ion Torrent	CCTCTCTATGGGCAGTCGGTGATAACAGTCTGAGAACTGAGGACACCG
Ion_IGHV3–47	Ion Torrent	CCTCTCTATGGGCAGTCGGTGATAGAGACAACGCCAAGAAGTCCTTG
Ion_IGHV3–49	Ion Torrent	CCTCTCTATGGGCAGTCGGTGATTCGCCTATCTGCAAATGAACAGCC
Ion_IGHV6–1	Ion Torrent	CCTCTCTATGGGCAGTCGGTGATACCCAGACACATCCAAGAACCAG
Ion_MID_SV5_Rev	Ion Torrent	TTCCATCTCATCCCTGCGTGTCTCCGACTCAGACGTGTGCAGTGGGTTTGGGATTGGTTTGCC
Mi_fw3.vh1	MiSeq	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTTCTACAGACACAGCCTACATGGAGC
Mi_fw3.vh1b	MiSeq	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTACGAGCACAGCCTACATGGAGC
Mi_fw3.vh1c	MiSeq	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTTACATGGAGCTGAGCAGCCTGAG
Mi_fw3.vh2	MiSeq	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTATGACCAACATGGACCCTGTGGAC
Mi_fw3.vh3	MiSeq	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTCCAGAGACAATTCCAAGAACACGC
Mi_fw3.vh3b	MiSeq	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTTGCAAATGAACAGCCTGAAAACCGAGG
Mi_fw3.vh4	MiSeq	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTAACCAGTTCTCCCTGAAGCTGAGC
Mi_fw3.vh5	MiSeq	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTAGTGGAGCAGCCTGAAGGCC
Mi_fw3.vh3c	MiSeq	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTATCTGCAAATGAACAGYCTGAGAGC
Mi_fw3.vh3d	MiSeq	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTAGAGACAATTCCAGGAACWYCCTG
Mi_fw3.vh7	MiSeq	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTCCWTGGACACCTCTGYCAGC
Mi_IGHV1–2	MiSeq	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTATCAGCACAGCCTACATGGAGCTG
Mi_IGHV1–68	MiSeq	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTTGAGGACAGCCTACATAGAGCTGAG
Mi_IGHV3–13	MiSeq	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTTCAAATGAACAGCCTGAGAGCCGG
Mi_IGHV3–43	MiSeq	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTAACAGTCTGAGAACTGAGGACACCG
Mi_IGHV3–47	MiSeq	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTAGAGACAACGCCAAGAAGTCCTTG
Mi_IGHV3–49	MiSeq	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTTCGCCTATCTGCAAATGAACAGCC
Mi_IGHV6–1	MiSeq	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTACCCAGACACATCCAAGAACCAG
Mi_MID1_ SV5_Rev	MiSeq	CAAGCAGAAGACGGCATACGAGATCGTGATGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCGCAGTGGGTTTGGGATTGG TTTGCC

Deep sequencing[21-23] refers to sequencing methods producing orders of magnitude more reads than traditional Sanger sequencing. Until recently, these technologies were dominated by systems that were expensive to purchase and operate, and required extensive preparation time before results could be obtained. They have been widely applied to the sequencing and analysis of genomes, and more recently to the investigation of diverse library selections,[24-29] including the analysis of both in vitro antibody libraries[24,26] and in vivo antibody repertoires,[12,25,30-32] where HCDR3 is usually used as an antibody identifier. The results obtained from the analysis of library selections indicate that when only 96 or 384 clones are screened, many abundant, and potentially

Table 2. Sequence Statistics for 454, Ion Torrent and MiSeq data sets of the library

	454		Ion 1	Ion 2	Ion 2.2	MiSeq
Raw reads	1,417,344		2,151,956	3,895,583	3,909,701	5,697,883
Filtered reads	1,296,818		817,468	1,644,295	1,570,152	5,612,344
# of CDR3s	**VDJ**	**Regex**	426,894	1,049,297	797,613	5,046,749
	553,376	613,513				
# of unique CDR3s	363,620	396,183	240,209	604,107	487,428	2,022,431

Table 3. Regex validation by an independent data set of human VH antibody sequences

Filtered reads	**1,976,330**	
	VDJFasta	**Regex**
# ofCDR3s	1,101,812	1,213,417
# of unique CDR3s	165,903	178,055

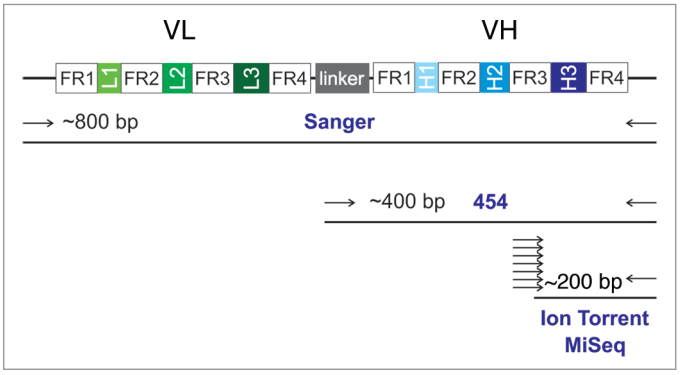

Figure 1. PCR priming scheme for the different sequencing platforms.

valuable clones, are lost,[24,27] a result confirmed with peptide libraries,[28,33] whereas if deep sequencing is applied to selection outputs, the most abundant clones can be unambiguously identified and isolated using specific primers. This also allows access to a far greater diversity of positive clones than the number obtained by random screening.[34]

To enable the use of deep sequencing methods more broadly in selections, the cost of sequencing and the downstream processes need to be streamlined. "Bench-top sequencers" (for review see ref. 35), are laser-printer sized, inexpensive to purchase and run and provide results in a matter of hours, rather than days, making them of great potential utility in this field. Sequence analysis is also challenging and generally performed by experts using specialized computer clusters. In this paper, we compare three different sequencing platforms (454, MiSeq and Ion Torrent PGM) and describe their straightforward implementation to both the analysis of a well-characterized naïve antibody library[36] and selections from it. We provide the necessary HCDR3 primer sequences and easy-to-use open source informatics tools to make deep sequencing routinely available for antibody selection analysis (http://sourceforge.net/projects/abmining/).

Results

The development and validation of RegEx

The identification of HCDR3s is inherently difficult because of their extreme diversity: authentic HCDR3s may have features that render them atypical, even when functional. VDJFasta[26] is a successful algorithm that uses a Hidden Markov Model to statistically analyze sequences upstream and downstream of putative HCDR3s. Although effective on 454 data, because of the read length, VDJFasta is unsuitable for shorter MiSeq and Ion Torrent reads. We developed a new HCDR3 recognition software package based on regular expression (RegEx) pattern, in which nucleic acid sequences encoding critical amino acids (aa) characteristic of HCDR3s and flanking sequences are used as identifiers. A naïve antibody library[36] was sequenced using 454, MiSeq and Ion Torrent: a schematic representation of the primers mapping on the scFv is shown in **Figure 1**. The primers used are shown in **Table 1**, with a summary of the complete sequencing results reported in **Table 2**. The methods used to sequence using MiSeq and Ion Torrent are reported below. HCDR3s were identified in the 454 data set using either RegEx or VDJFasta. RegEx analysis was ~1 000 times faster than VDJFasta, and could be performed on a single personal computer, rather than a computer cluster. RegEx accuracy was shown to be comparable to VDJFasta by comparing the HCDR3s identified by the two algorithms. 84% of HCDR3s were recognized by both algorithms (**Fig. 2A and 2B**), the cumulative total of identified HCDR3s ranked by the corresponding number of occurrences was identical for both (**Fig. 2C**), as was the length distribution of HCDR3s identified using RegEx or VDJFasta[37] (**Fig. 2D**). Furthermore, the aa distribution at each position for all HCDR3s was essentially identical for HCDR3s recognized by either, or both, algorithms (**Fig. 3A**). Finally, we observed that the number of unique HCDR3s identified by Regex in the 454 data set was ~9% higher than the number identified by VDJFasta (**Table 2**; **Fig. 2B**), and that for any specific HCDR3 in this data set, RegEx identified ~10% more clones than VDJFasta. These data indicate that the VDJFasta identification parameters were occasionally too stringent, and appeared to exclude HCDR3s that otherwise appeared to be valid. Although there may be slight differences between the HCDR3s identified by the two algorithms, reflecting the innate difficulty of identifying HCDR3s, the majority are identified by both programs, making RegEx a valid, and extremely rapid, alternative to VDJFasta.

As the naïve antibody library described above was used to train the RegEx algorithm, we used an independent data set of

human VH antibody sequences,[38] to validate its functionality. Both RegEx and VDJFasta were used to identify HCDR3s from the combined data set containing 1 976 330 reads: the sequencing and analysis results are reported in **Table 3**, where RegEx again consistently identified ~10% more of the common HCDR3 sequences and significantly increased the number of unique HCDR3s recognized compared with VDJFasta (**Fig. 2B**). This result validates the regular expression as a universal recognition pattern for the analysis of human antibody libraries. The inherent speed of the regular expression search enabled us to create the AbMining ToolBox, a complete HCDR3 analysis package for antibody deep sequencing outputs using the popular next generation platforms. This software package is freely available at http://sourceforge.net/projects/abmining/ with instructions for the installation of the necessary packages for Windows, Mac and Linux operating systems. A detailed user guide for all the scripts is included in the ToolBox. These include frequency determination, barcode analysis, clustering and Hamming distance calculations, among others. We used the AbMining ToolBox to characterize the antibody library itself and selections using different sequencing platforms.

Comparing the different sequencing platforms using AbMining ToolBox

In order to sequence the antibody library by MiSeq and Ion Torrent, the HCDR3s of the antibody library were amplified by a set of 18 primers mapping upstream of HCDR3 in framework 3 and a downstream vector primer (**Table 1**; **Fig. 1**) designed to cover the entire VH diversity. The MiSeq and Ion Torrent sequences obtained from these amplifications were analyzed using the AbMining ToolBox, identifying and clustering the HCDR3s. The obtained data were compared with the 454 dataset.

Unlike the previous comparison, where the algorithms were assessed on the same data set, these sequencings represent independent samplings of the same extremely large population. When diversity greatly exceeds the number of sequencing reads, most sequences obtained from two independent samples will be different[25,32] and only abundant HCDR3s are expected to be found in both populations. This is observed in **Figures 4A-C**, where the greatest number of sequences is unique for each data set. Similar results are obtained when two independent Ion Torrent runs are compared (**Fig. 4D**). Sequence distributions are broadest when 454 HCDR3s are compared with Ion Torrent or MiSeq (**Fig. 4A and C**) and tightest when comparing MiSeq to Ion Torrent (**Fig. 4B**), or resequencing (**Fig. 4D**), probably reflecting the use of similar primers in MiSeq and Ion Torrent, and different primers for 454. This makes it more difficult to compare the different sequencing methods at the individual HCDR3 level. However, aggregate properties, such as HCDR3 length distribution (**Fig. 2D**) and aa distributions at each HCDR3 position for all HCDR3 lengths, with the three sequencing platforms can be compared, and are essentially identical for the three platforms (**Fig. 3B**).

One possible concern of these deep sequencing platforms is that their error rates[35] will overestimate the number of HCDR3s. To assess this, each individual HCDR3 of a defined length (4–21 aa, Kabat numbering) was compared with all other HCDR3s of the same length and the minimal Hamming distance for the closest HCDR3 determined for each. **Figure 5A** show the percentage of HCDR3s with the minimum calculated Hamming distance for aa sequences. 8–11% of HCDR3s were 1–2 Hamming aa distances away from at least one other HCDR3, with 454 having slightly higher values than MiSeq and Ion Torrent indicating that, within the context used here, error rates are similar for all platforms.

Table 4. Quality trimming optimization on all three sequencing platform outputs. The optimization of average quality value and step value on an Ion Torrent, 454, and MiSeq sequencing output

		Step 1	Step 3	Step 5	Step 10
Q 9	Time	16 min	8 min	7 min	6 min
	CDR3	1305694	1305695	1305696	1305696
	CDRX	56092	56096	56096	56096
	% CDRX	4.2963%	4.2963%	4.2963%	4.2963%
Q 12	Time	13 min	8 min	6:30 min	6 min
	CDR3	1228662	1231206	1233520	1238795
	CDRX	32853	33514	34098	35390
	% CDRX	2.674%	2.722%	2.764%	2.857%
Q 15	Time	11 min	7 min	6 min	5:30 min
	CDR3	1145112	1147791	1150310	1156599
	CDRX	14732	15010	15283	15936
	% CDRX	1.2866%	1.3077%	1.3286%	1.3778%
Q 18	Time	11 min	7 min	6 min	5 min
	CDR3	1088986	1092005	1094978	1102442
	CDRX	9072	9182	9307	9595
	% CDRX	0.833%	0.841%	0.850%	0.870%
Q 21	Time	10 min	7 min	6 min	5 min
	CDR3	1026139	1029917	1033471	1061683
	CDRX	6655	6718	6779	6964
	% CDRX	0.649%	0.652%	0.656%	0.656%
Q 24	Time	9 min	6 min	5:30 min	5 min
	CDR3	921544	926888	931401	942917
	CDRX	5220	5268	5300	5422
	% CDRX	0.566%	0.568%	0.569%	0.575%
Q 27	Time	8 min	N/D	N/D	N/D
	CDR3	732920	N/D	N/D	N/D
	CDRX	3800	N/D	N/D	N/D
	% CDRX	0.52%	N/D	N/D	N/D
Q 30	Time	7:30 min	N/D	N/D	N/D
	CDR3	377137	N/D	N/D	N/D
	CDRX	1819	N/D	N/D	N/D
	% CDRX	0.48%	N/D	N/D	N/D
Q 33	Time	6:30 min	N/D	N/D	N/D
	CDR3	13330	N/D	N/D	N/D
	CDRX	56	N/D	N/D	N/D
	% CDRX	0.042%	N/D	N/D	N/D

Application of AbMining ToolBox to naïve antibody library analysis

As the total combined number of reads obtained with all three platforms (7.9×10^6) exceeds 10% of the maximum potential VH diversity of this library, as measured by the number of transformants (7×10^7), we pooled all the HCDR3s identified using

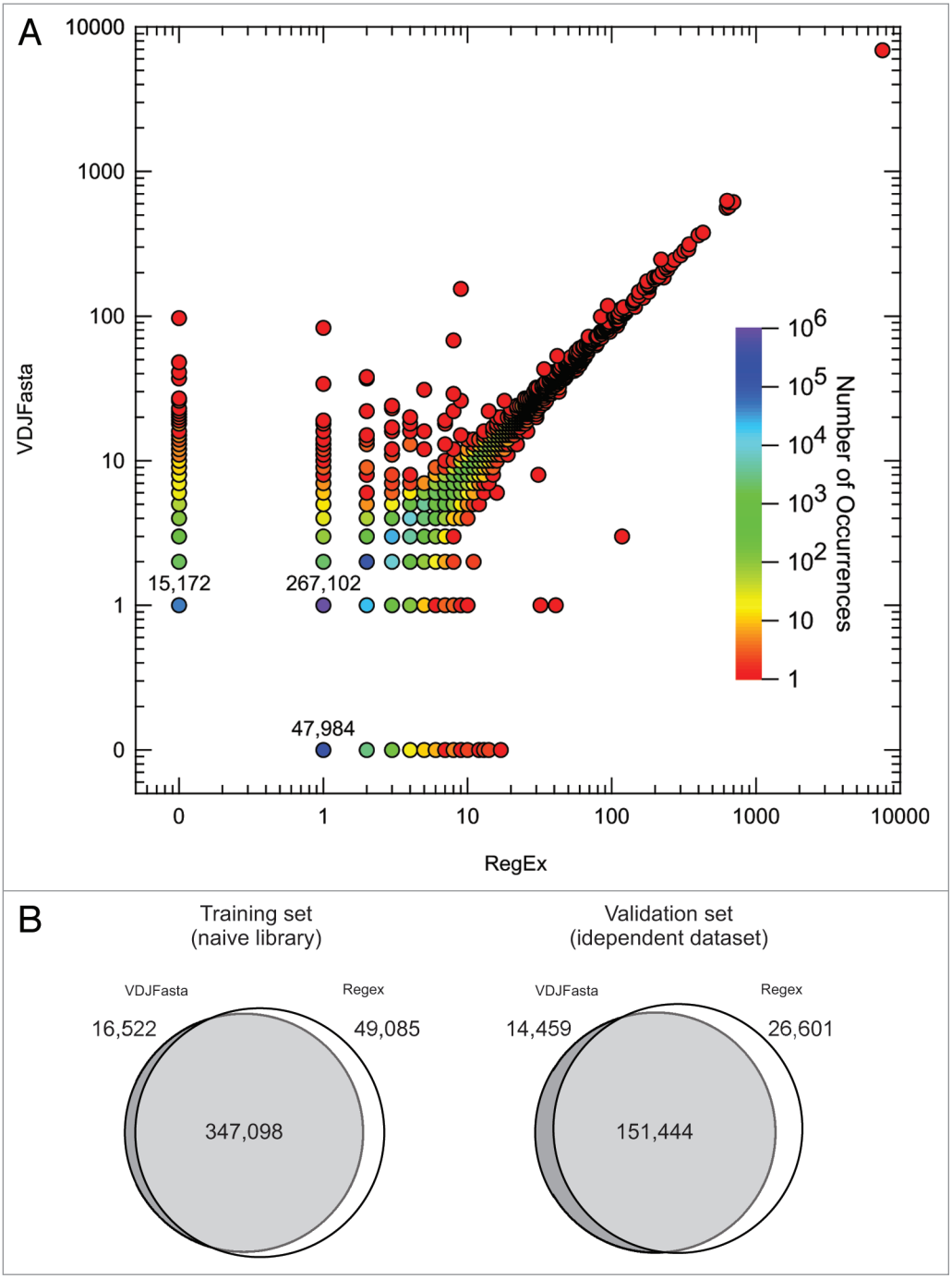

Figure 2. RegEx validation. (**A**) Comparison frequency of HCDR3s identified by RegEx and VDJFastA on the same 454 data set. The numbers of HCDR3s identified at each frequency are color coded with the numbers of HCDR3s recognized by either RegEx, VDJFasta, or both indicated. (**B**) Proportional VENN diagram of the identified unique HCDR3s by RegEx and VDJFasta on the naïve library and an independent data set. The sizes and the intersections of the circles are proportional to the number of HCDR3s.

the AbMining ToolBox from all the different sequencing platforms and plotted the unique HCDR3s against the total number of reads (**Fig. 5B**). This provided a plot of unique HCDR3 accumulation, vs. number of reads, and reached a total of ~3.3 × 10^6 unique HCDR3s for the 7.9 × 10^6 reads. This number of unique HCDR3s includes those that differ by only one or two aa (**Fig. 5A**), which may be a consequence of sequencing errors or somatic hypermutation. The presence of these similar clones will tend to overestimate the functional HCDR3 diversity in this library; however, this reduction in functional diversity will be compensated for by additional diversity in HCDR1 and HCDR2, as well as VL recombination,[26] which will link each identified HCDR3 with different numbers of VL chains.

Selection of antibodies against Ag85

In a final set of experiments, we selected antibodies against Ag85, a tuberculosis antigen, using a combination of phage and yeast display,[34] and identified the 15 most abundant HCDR3 clones by analyzing Ion Torrent sequencing with the AbMining ToolBox. The frequencies of the most abundant binders identified by deep sequencing within the selected population range from 1.68% for the most abundant clone, to 0.32% for the 15th ranked clone. All clones bound the target specifically (**Fig. 6**), with no correlation between abundance rank and binding efficacy. In fact, the clone giving the third strongest signal was ranked 14th in abundance. This confirms the utility of deep sequencing and abundance analysis to identify positive clones that may otherwise be missed,[24] especially when even the most abundant clones have relatively low frequencies, as observed in this particular selection.

Discussion

We have demonstrated here that deep sequencing combined with the AbMining ToolBox package can be extremely effective in the analysis of antibody library diversity and selections. As HCDR3s are well-established antibody diversity surrogates,[11,20] this allows the direct assessment of minimum antibody diversity

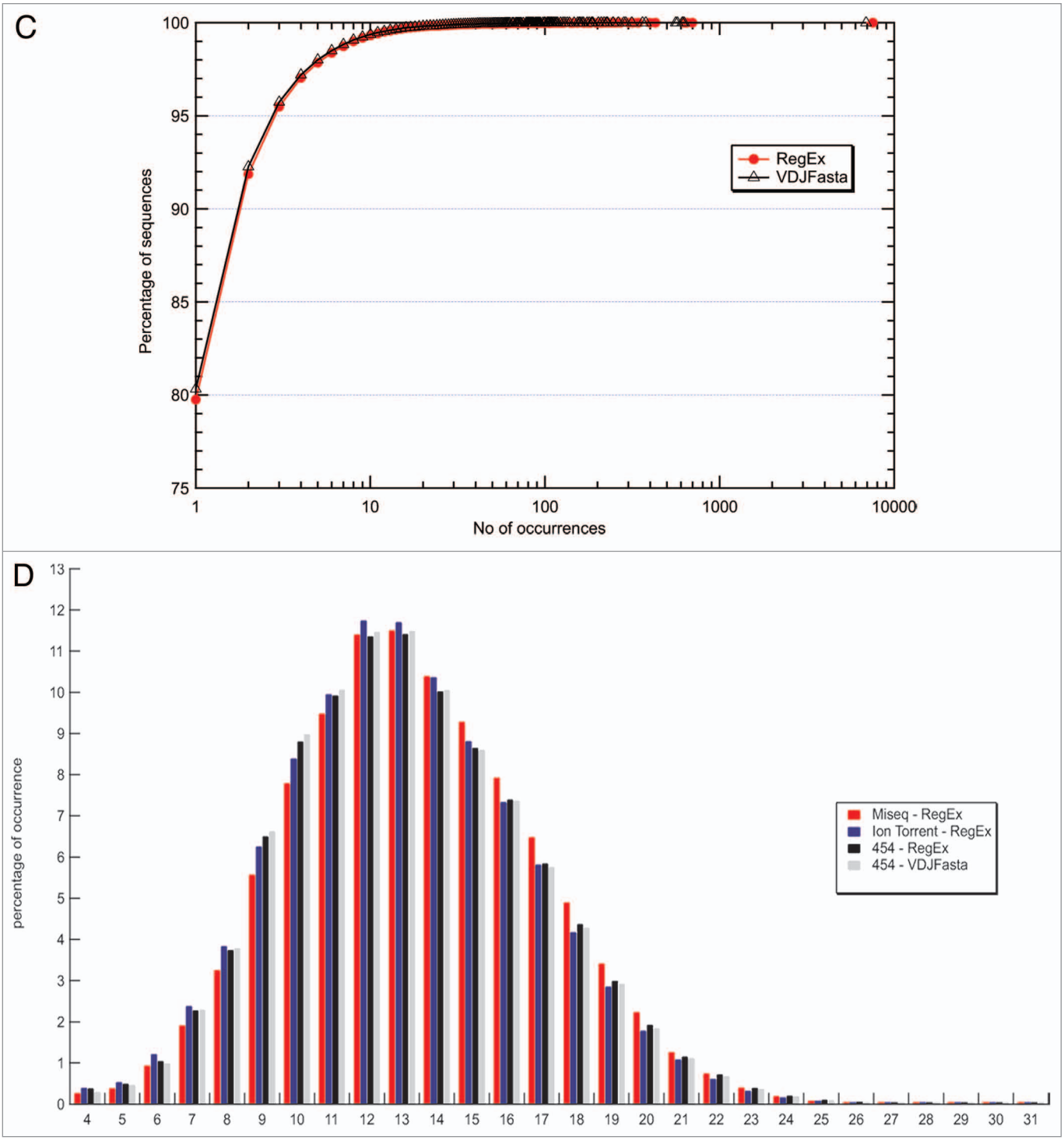

Figure 2. (**C**) The accumulation of unique HCDR3s identified by RegEx or VDJFasta in the 454 data set. (**D**) HCDR3 length distribution determined for all three sequencing platforms by RegEx, and for 454 sequencing using either VDJFasta or RegEx.

in an antibody population, naïve or selected. Additional diversity in HCDR1 and HCDR2 are double that in HCDR3,[26] and recombination pairs most HCDR3s with different VLs, further increasing library diversity estimates. Improvements in deep sequencing capabilities will increase the usable length of sequences, eventually allowing the sequencing of full VH/VL domains, which will also be easily identifiable using modified RegEx patterns.

Compared with other deep sequencing methods, the low cost and sequencing depth of Ion Torrent and MiSeq make them particularly useful in antibody selection, with Ion Torrent having the advantage of greater speed, and MiSeq the advantage of the

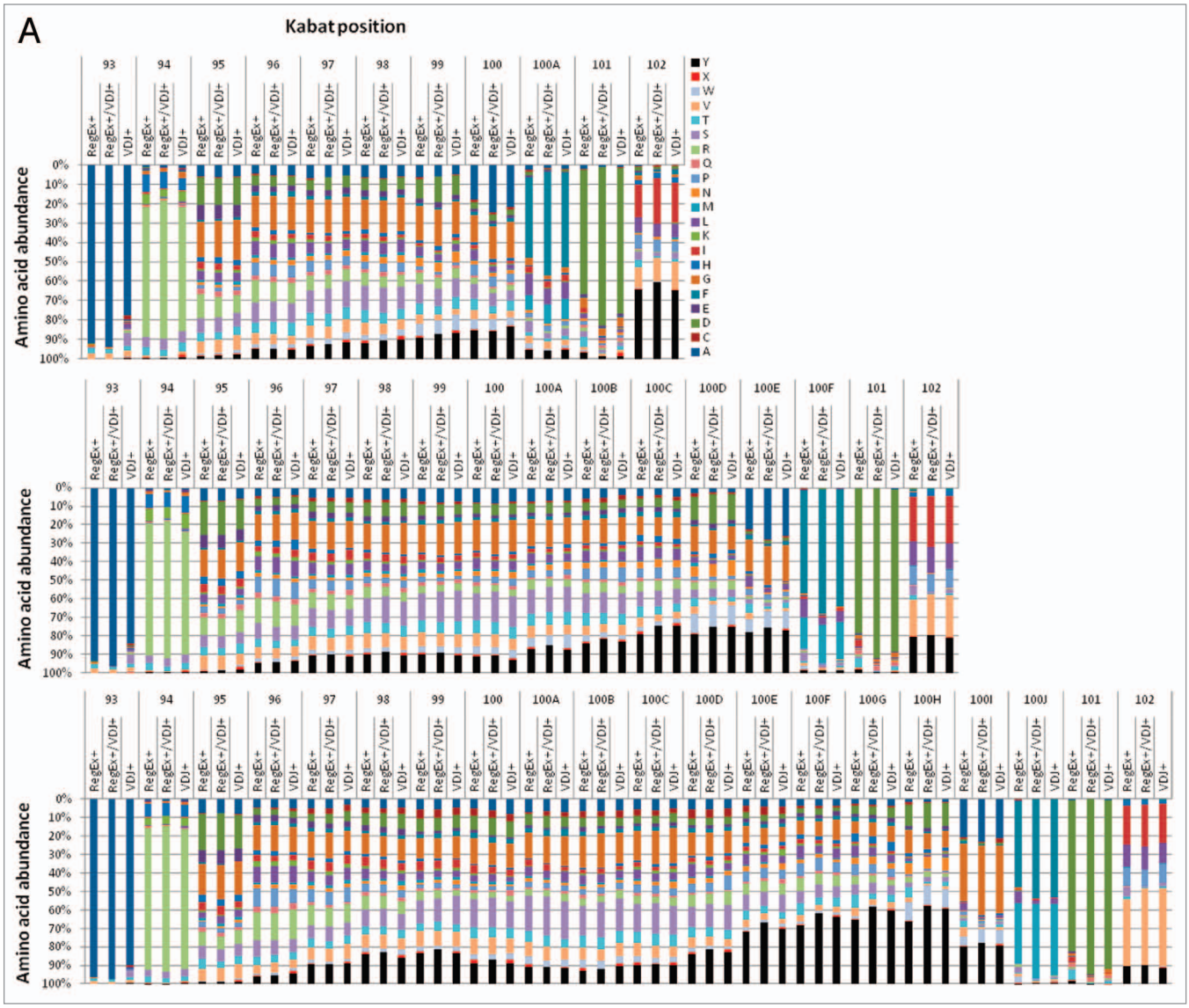

Figure 3. (**A**) The amino acid distribution at each HCDR3 position identified exclusively by RegEx (RegEx+), VDJFasta (VDJFasta+), or by both methods (RegEx+/VDJFasta+) using the 454 sequence data set.

greater number of reads. The output after a single round of phage antibody selection is usually $10^{5,6}$ clones, representing the maximum subsequent attainable diversity. This is matched by present Ion Torrent and MiSeq capacities, making the identification of every clone in a selection output, ranked by abundance, now feasible in only five hours after PCR amplification (30 h for MiSeq). Analyses performed on a standard personal computer will allow sequencing information to directly influence selection outcome, and effectively democratize the use of deep sequencing in antibody selections.[39]

Although to date the application of deep sequencing to the analysis of selections from antibody and other libraries has been limited, it has already been proposed that deep sequencing after a single round of phage peptide library selection is sufficient to identify positive clones.[28] We anticipate this will also become possible for antibody selections, as sequencing costs continue their downward trend, and the number, quality and lengths of reads increases. However, we expect the power of deep sequencing to go well beyond the identification of positive clones in early selection rounds. As more experience is obtained, it is likely that classes of antibodies with particular molecular (e.g., stability, biochemical liabilities in CDRs) and binding (e.g., hapten, protein, peptide) properties may be identifiable by their sequences, as will antibodies with undesirable properties (e.g., plastic or biotin binders[40]) that can be discarded. Furthermore, it may be possible to identify antibodies binding to one target, but not a closely related one, merely on the basis of antibody sequences obtained during selection, or antibodies binding to two different targets (e.g., murine and human versions of the same protein) by identifying common sequences in selections. We expect the

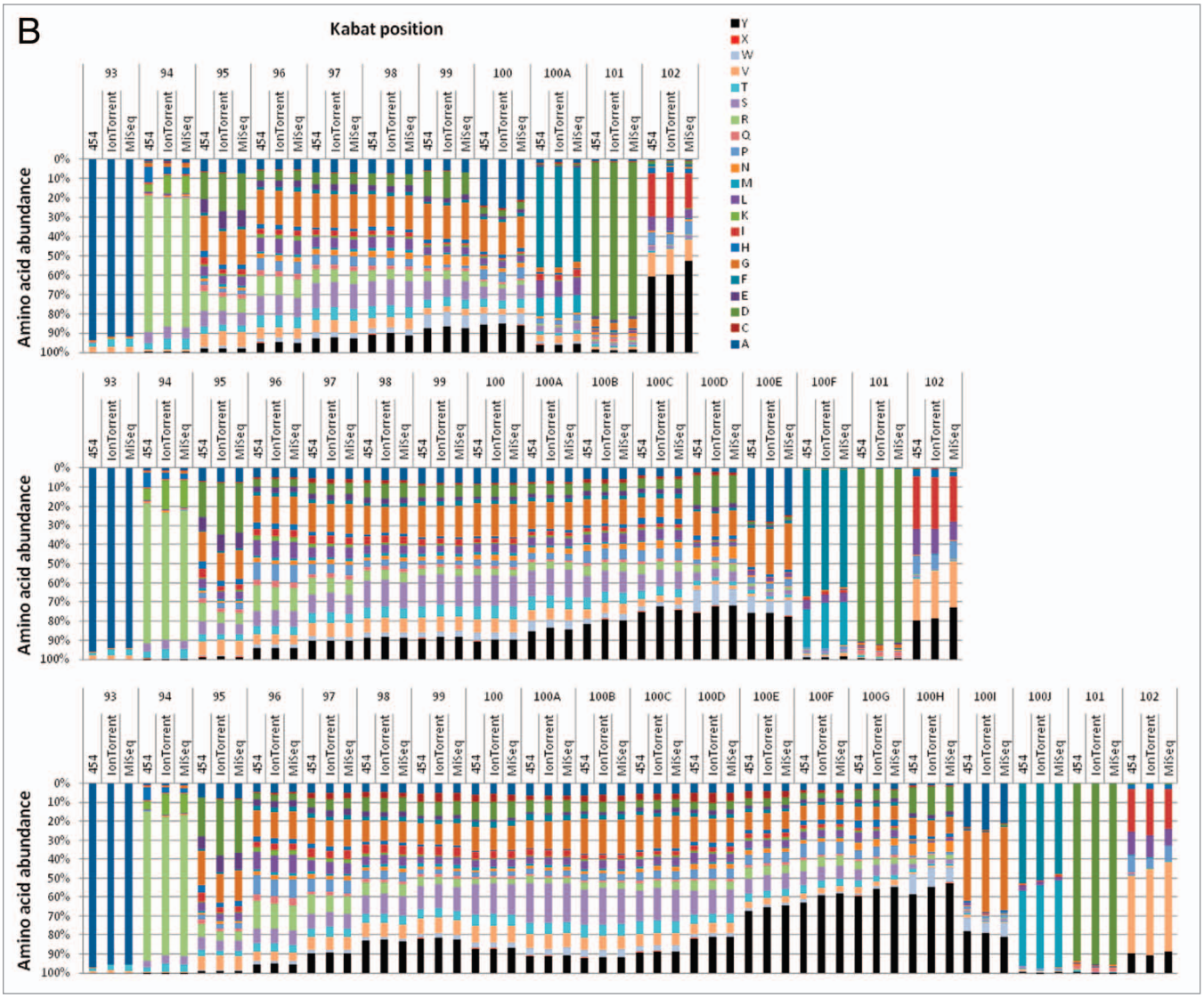

Figure 3. (**B**) for each sequencing platform using RegEx, for three different HCDR3 lengths (9, 14, and 18).

deep sequencing of antibody selections to become an essential and integral part of the selection process as systems such as Ion Torrent and MiSeq become more widely available.

Although the methods described here were applied to HCDR3s in antibody libraries, it is clear that with modifications, the approach taken can also be used in the analysis of selections of other CDRs or other binding scaffolds, by simply modifying the RegEx pattern for the recognition of scaffold boundary sequences.

Materials and Methods

Sequencing primer design

A specific set of primers was designed for the different sequencing platforms (**Table 1**). For 454 sequencing, 2 primers mapping to the pDAN5 vector upstream and downstream of the VH genes were designed. These contain the 454 specific sequencing adaptors.

For IonTorrent and MiSeq, a set of 18 forward primers mapping to the VH framework just upstream the HCDR3 were designed. They maximize the coverage of human framework 3 VH in multiplex reactions with a minimal set of perfect-match primers against germline V-segments. Primers were optimized for a common annealing temperature, GC content, minimal self-annealing or cross-annealing to other primers, and all contained a GC-clamp at the 3' end. Coverage of a curated subset of the 454 data set showed that ~94% of antibody genes were matched, if up to 4 mismatches were permitted outside the 3′ GC-clamp region.

As reverse primer, a primer mapping to the pDAN5 vector just downstream of the VH gene was designed. Sequencing specific adaptors were introduced in both forward and reverse primers.

Sample preparation

The scFv library analyzed here has been previously characterized.[36] Briefly, a 7 × 10^7 primary library of assembled VL and VH domains was created from cDNA derived from the PBMC of

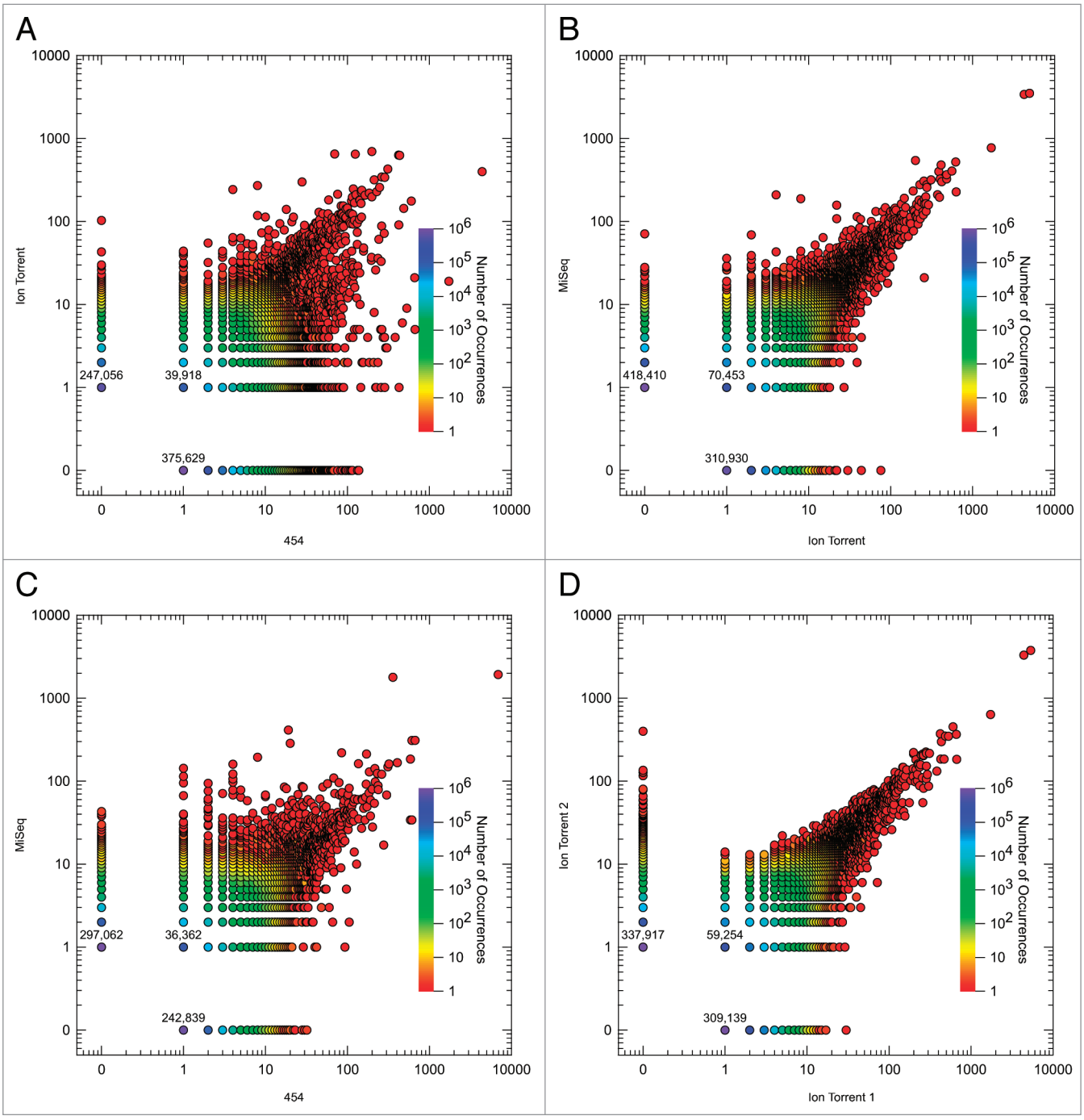

Figure 4. HCDR3 analysis of different data sets. For each panel, HCDR3s were identified using AbMining ToolBox from each indicated data set and then plotted, as described in **Figure 1A**. Comparisons of (**A**) 454 and Ion Torrent. (**B**) MiSeq and Ion Torrent. (**C**) 454 and Miseq. (**D**) Two independent Ion Torrent sequencing runs.

40 healthy donors and cloned into the pDAN5 phagemid vector. Plasmid DNA from this library was obtained and 0.3 fmol used as a template to prepare the amplicon samples for sequencing.

After PCR amplification, the amplicon was gel purified and quantified (Qbit, HS kit, Invitrogen). The sample was prepared for GS FLX Titanium Series Lib-A Chemistry (Roche) bi-directional amplicon sequencing according to the manufacturer's instructions and sequenced on a 2 regions pico titer plate.

For Ion Torrent and MiSeq, the 18 forward primers (**Table 1**) were mixed in equimolar amounts and used for the PCR with Phusion High-Fidelity DNA polymerase (NEB). The ~240 bp amplicon was purified as previously described. The Ion Xpress Amplicon library protocol was used to prepare the sample for sequencing on the Ion 316 chips (Life Technologies). The MiSeq amplicon was prepared with a MiSeq reagent kit and run on a PE151 run.

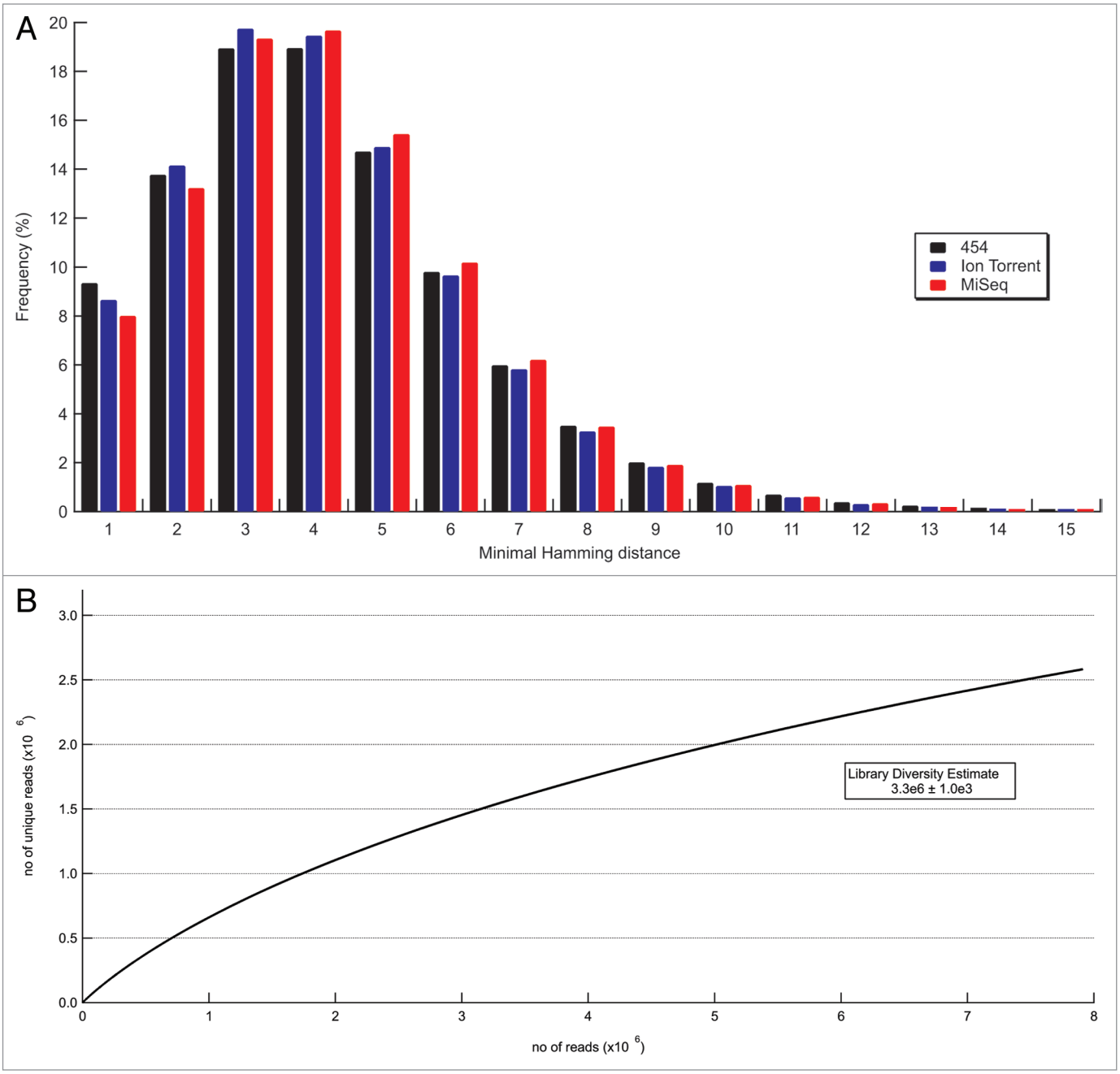

Figure 5. (**A**) Minimal amino acid Hamming distance distribution for the three sequencing platforms for all HCDR3 lengths of the naïve library. (**B**) Library diversity estimate by accumulation using the pooled unique sequences of all three sequencing platforms.

Sequence analysis: VDJFasta

The quality trimmed 454 sequencing reads were split into files containing 10 000 sequences and used in VDJFasta as described in Glanville et al.[26]

Sequence analysis: RegEx construction

The HCDR3 recognizing regular expression (RegEx) pattern used in this article was refined iteratively using the VDJFasta CDR3 data set obtained from the 454 sequences. Once a RegEx pattern was defined, it was used to identify HCDR3s from the 454 data set. The two CDR3 data sets were compared and the VDJFasta exclusive CDR3s were analyzed. The RegEx pattern was modified to include the VDJFasta exclusive CDR3s as well; the process was repeated until the RegEx was sufficiently inclusive and sensitive, with the final RegEx pattern being:

TA[CT](TT[CT]|TA[TC]|CA[TC]|GT[AGCT])TG[TC][GA][AGCT]([ACGT]{3}){5,32}[AGCT]TGGG[GCT][GCT]

The pattern represents a balance between including as many CDR3s as possible, while minimizing the number of false positive sequences.

The AbMining ToolBox developed for this article is freely available at Sourceforge (http://sourceforge.net/projects/abmining/). The required software installation guide provides

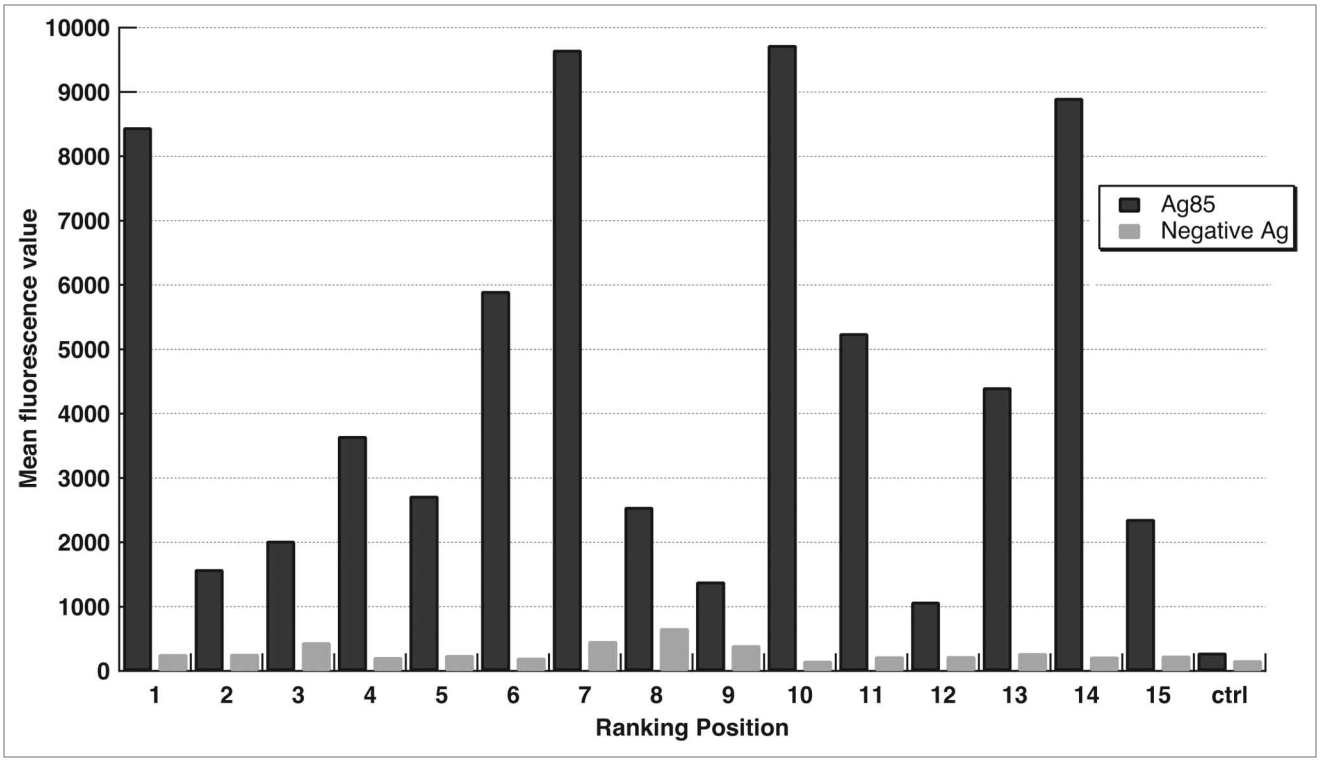

Figure 6. Binding specificity assessment of the 15 most abundant HCDR3 clones by flow cytometry against Ag85 and a negative antigen.

Table 5. Quality trimming optimization on all three sequencing platform outputs. The optimization of average quality value and step value on 454

	Q0	Q10	Q15	Q16	Q18	Q20	Q22	Q25
# CDR3	611536	611520	602941	594041	561389	510993	450001	356249
# CDRX	7605	7605	7105	6682	5367	3950	2907	1962
% of CDRX	1.24%	1.24%	1.18%	1.12%	0.96%	0.77%	0.65%	0.55%

installation information for the necessary software packages, and the user guide contains detailed information how to use the toolbox's scripts.

The raw data of the three platforms were used for optimizing the quality trimming parameters by means of AbMining ToolBox. **Table 4** shows the detailed optimization of an Ion Torrent data set. Two parameters were tested: the quality average value (Q) and the window step value (step). The quality average value influences the overall quality of trimmed DNA reads. Low Q setting would allow too many sequencing errors to slip through; high Q setting would eliminate too many good sequences. The balance between the number of CDR3s identified and the number of CDR3s containing STOP codons (CDRX) was used to determine the optimal Q value.

For the input data, the filtering of the raw sequences was performed and optimized for all 3 platforms' outputs. **Tables 4, 5, and 6** show the quality trimming analysis for Ion Torrent, 454 and MiSeq data sets, respectively. For the Ion Torrent, the optimal Q value was 21. The step setting can be used to speed up the quality trimming. A bigger step value could result in significant time savings with a modest decrease in output quality (**Table 4**). For 454, Q20 was the best compromise average quality value (**Table 5**), while for MiSeq the Q value did not show any significant effect. A Q value of 21 was chosen for all sequence analysis (**Table 6**).

Selection of antibodies against Ag85

Phage display selection and yeast display sorting were performed as described by Ferrara et al.[34] The naïve phage antibody library was used to select Ag85 antibodies: biotinylated Ag85 was used at 50 nM concentration in the first round of phage selection, and 5 nM in the second. After two rounds of phage selection, DNA encoding the selected scFv antibodies was recovered and used as template for PCR amplification and recloned into a yeast display vector. The obtained yeast library was further enriched by

Table 6. Quality trimming optimization on all three sequencing platform outputs. The optimization of average quality value and step value on MiSeq sequencing output

	Q9	Q12	Q15	Q18	Q21	Q24
# CDR3	5067895	5067888	5067821	5066130	5046749	4983417
# CDRX	25530	25530	25522	25435	25035	24446
% of CDRX	0.503%	0.504%	0.504%	0.502%	0.496%	0.490%

one round of sorting using flow cytometry (FACSAria, BD). The scFvs displayed on yeast cells showing both antigen binding and scFv display were sorted. Plasmid DNA was recovered from the sorted yeast and sequenced by Ion Torrent. The unique HCDR3s were identified and ranked by abundance using the ToolBox. The clones corresponding to the 15 most abundant HCDR3s found by Ion Torrent were identified by Sanger sequencing and tested for binding specificity by flow cytometry.

Disclosure of Potential Conflicts of Interest

No potential conflicts of interest were disclosed.

Acknowledgments

This work was supported by the National Institutes of Health [5U54DK093500–02 to ARMB]; and Los Alamos National Laboratory Directed Research Development Directed Research [20120029DR] funds.

References

1. Marks JD, Hoogenboom HR, Bonnert TP, McCafferty J, Griffiths AD, Winter G. By-passing immunization. Human antibodies from V-gene libraries displayed on phage. J Mol Biol 1991; 222:581-97; PMID:1748994; http://dx.doi.org/10.1016/0022-2836(91)90498-U
2. Boder ET, Wittrup KD. Yeast surface display for screening combinatorial polypeptide libraries. Nat Biotechnol 1997; 15:553-7; PMID:9181578; http://dx.doi.org/10.1038/nbt0697-553
3. Hanes J, Plückthun A. In vitro selection and evolution of functional proteins by using ribosome display. Proc Natl Acad Sci U S A 1997; 94:4937-42; PMID:9144168; http://dx.doi.org/10.1073/pnas.94.10.4937
4. Bradbury ARM, Sidhu S, Dübel S, McCafferty J. Beyond natural antibodies: the power of in vitro display technologies. Nat Biotechnol 2011; 29:245-54; PMID:21390033; http://dx.doi.org/10.1038/nbt.1791
5. Colwill K, Gräslund S, Jarvik NE, Wyrzucki A, Wojcik J, Koide A, Kossiakoff AA, Koide S, Sidhu S, Dyson MR, et al.; Renewable Protein Binder Working Group. A roadmap to generate renewable protein binders to the human proteome. Nat Methods 2011; 8:551-8; PMID:21572409; http://dx.doi.org/10.1038/nmeth.1607
6. Pershad K, Pavlovic JD, Gräslund S, Nilsson P, Colwill K, Karatt-Vellatt A, Schofield DJ, Dyson MR, Pawson T, Kay BK, et al. Generating a panel of highly specific antibodies to 20 human SH2 domains by phage display. Protein Eng Des Sel 2010; 23:279-88; PMID:20164216; http://dx.doi.org/10.1093/protein/gzq003
7. Schier R, Bye J, Apell G, McCall A, Adams GP, Malmqvist M, Weiner LM, Marks JD. Isolation of high-affinity monomeric human anti-c-erbB-2 single chain Fv using affinity-driven selection. J Mol Biol 1996; 255:28-43; PMID:8568873; http://dx.doi.org/10.1006/jmbi.1996.0004
8. Boder ET, Midelfort KS, Wittrup KD. Directed evolution of antibody fragments with monovalent femtomolar antigen-binding affinity. Proc Natl Acad Sci U S A 2000; 97:10701-5; PMID:10984501; http://dx.doi.org/10.1073/pnas.170297297
9. Sanger F, Nicklen S, Coulson AR. DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci U S A 1977; 74:5463-7; PMID:271968; http://dx.doi.org/10.1073/pnas.74.12.5463
10. Nicaise M, Valerio-Lepiniec M, Minard P, Desmadril M. Affinity transfer by CDR grafting on a nonimmunoglobulin scaffold. Protein Sci 2004; 13:1882-91; PMID:15169956; http://dx.doi.org/10.1110/ps.03540504
11. Xu JL, Davis MM. Diversity in the CDR3 region of V(H) is sufficient for most antibody specificities. Immunity 2000; 13:37-45; PMID:10933393; http://dx.doi.org/10.1016/S1074-7613(00)00006-6
12. Larimore K, McCormick MW, Robins HS, Greenberg PD. Shaping of human germline IgH repertoires revealed by deep sequencing. J Immunol 2012; 189:3221-30; PMID:22865917; http://dx.doi.org/10.4049/jimmunol.1201303
13. Early P, Huang H, Davis M, Calame K, Hood L. An immunoglobulin heavy chain variable region gene is generated from three segments of DNA: VH, D and JH. Cell 1980; 19:981-92; PMID:6769593; http://dx.doi.org/10.1016/0092-8674(80)90089-6
14. Tonegawa S. Somatic generation of antibody diversity. Nature 1983; 302:575-81; PMID:6300689; http://dx.doi.org/10.1038/302575a0
15. Nezlin R. Combinatorial events in generation of antibody diversity. Comb Chem High Throughput Screen 2001; 4:377-83; PMID:11472226; http://dx.doi.org/10.2174/1386207013330977
16. Silverstein AM. Splitting the difference: the germline-somatic mutation debate on generating antibody diversity. Nat Immunol 2003; 4:829-33; PMID:12942083; http://dx.doi.org/10.1038/ni0903-829
17. Schatz DG, Oettinger MA, Schlissel MS. V(D)J recombination: molecular biology and regulation. Annu Rev Immunol 1992; 10:359-83; PMID:1590991; http://dx.doi.org/10.1146/annurev.iy.10.040192.002043
18. Goyenechea B, Milstein C. Modifying the sequence of an immunoglobulin V-gene alters the resulting pattern of hypermutation. Proc Natl Acad Sci U S A 1996; 93:13979-84; PMID:8943046; http://dx.doi.org/10.1073/pnas.93.24.13979
19. Wagner SD, Milstein C, Neuberger MS. Codon bias targets mutation. Nature 1995; 376:732; PMID:7651532; http://dx.doi.org/10.1038/376732a0
20. Kabat EA, Wu TT. Identical V region amino acid sequences and segments of sequences in antibodies of different specificities. Relative contributions of VH and VL genes, minigenes, and complementarity-determining regions to binding of antibody-combining sites. J Immunol 1991; 147:1709-19; PMID:1908882
21. Niedringhaus TP, Milanova D, Kerby MB, Snyder MP, Barron AE. Landscape of next-generation sequencing technologies. Anal Chem 2011; 83:4327-41; PMID:21612267; http://dx.doi.org/10.1021/ac2010857
22. Pareek CS, Smoczynski R, Tretyn A. Sequencing technologies and genome sequencing. J Appl Genet 2011; 52:413-35; PMID:21698376; http://dx.doi.org/10.1007/s13353-011-0057-x
23. Metzker ML. Sequencing technologies – the next generation. Nat Rev Genet 2010; 11:31-46; PMID:19997069; http://dx.doi.org/10.1038/nrg2626
24. Ravn U, Gueneau F, Baerlocher L, Osteras M, Desmurs M, Malinge P, Magistrelli G, Farinelli L, Kosco-Vilbois MH, Fischer N. By-passing in vitro screening--next generation sequencing technologies applied to antibody display and in silico candidate selection. Nucleic Acids Res 2010; 38:e193; PMID:20846958; http://dx.doi.org/10.1093/nar/gkq789
25. Glanville J, Kuo TC, von Büdingen HC, Guey L, Berka J, Sundar PD, Huerta G, Mehta GR, Oksenberg JR, Hauser SL, et al. Naive antibody gene-segment frequencies are heritable and unaltered by chronic lymphocyte ablation. Proc Natl Acad Sci U S A 2011; 108:20066-71; PMID:22123975; http://dx.doi.org/10.1073/pnas.1107498108
26. Glanville J, Zhai W, Berka J, Telman D, Huerta G, Mehta GR, Ni I, Mei L, Sundar PD, Day GM, et al. Precise determination of the diversity of a combinatorial antibody library gives insight into the human immunoglobulin repertoire. Proc Natl Acad Sci U S A 2009; 106:20216-21; PMID:19875695; http://dx.doi.org/10.1073/pnas.0909775106
27. Di Niro R, Sulic A-M, Mignone F, D'Angelo S, Bordoni R, Iacono M, Marzari R, Gaiotto T, Lavric M, Bradbury ARM, et al. Rapid interactome profiling by massive sequencing. Nucleic Acids Res 2010; 38:e110; PMID:20144949; http://dx.doi.org/10.1093/nar/gkq052
28. 't Hoen PA, Jirka SM, Ten Broeke BR, Schultes EA, Aguilera B, Pang KH, Heemskerk H, Aartsma-Rus A, van Ommen GJ, den Dunnen JT. Phage display screening without repetitious selection rounds. Anal Biochem 2012; 421:622-31; PMID:22178910; http://dx.doi.org/10.1016/j.ab.2011.11.005
29. Yu H, Tardivo L, Tam S, Weiner E, Gebreab F, Fan C, Svrzikapa N, Hirozane-Kishikawa T, Rietman E, Yang X, et al. Next-generation sequencing to generate interactome datasets. Nat Methods 2011; 8:478-80; PMID:21516116; http://dx.doi.org/10.1038/nmeth.1597

30. Reddy ST, Ge X, Miklos AE, Hughes RA, Kang SH, Hoi KH, Chrysostomou C, Hunicke-Smith SP, Iverson BL, Tucker PW, et al. Monoclonal antibodies isolated without screening by analyzing the variable-gene repertoire of plasma cells. Nat Biotechnol 2010; 28:965-9; PMID:20802495; http://dx.doi.org/10.1038/nbt.1673
31. Arnaout R, Lee W, Cahill P, Honan T, Sparrow T, Weiand M, Nusbaum C, Rajewsky K, Koralov SB. High-resolution description of antibody heavy-chain repertoires in humans. PLoS One 2011; 6:e22365; PMID:21829618; http://dx.doi.org/10.1371/journal.pone.0022365
32. Weinstein JA, Jiang N, White RA 3rd, Fisher DS, Quake SR. High-throughput sequencing of the zebrafish antibody repertoire. Science 2009; 324:807-10; PMID:19423829; http://dx.doi.org/10.1126/science.1170020
33. Vodnik M, Zager U, Strukelj B, Lunder M. Phage display: selecting straws instead of a needle from a haystack. Molecules 2011; 16:790-817; PMID:21248664; http://dx.doi.org/10.3390/molecules16010790
34. Ferrara F, Naranjo LA, Kumar S, Gaiotto T, Mukundan H, Swanson B, Bradbury AR. Using phage and yeast display to select hundreds of monoclonal antibodies: application to antigen 85, a tuberculosis biomarker. PLoS One 2012; 7:e49535; PMID:23166701; http://dx.doi.org/10.1371/journal.pone.0049535
35. Loman NJ, Misra RV, Dallman TJ, Constantinidou C, Gharbia SE, Wain J, Pallen MJ. Performance comparison of benchtop high-throughput sequencing platforms. Nat Biotechnol 2012; 30:434-9; PMID:22522955; http://dx.doi.org/10.1038/nbt.2198
36. Sblattero D, Bradbury A. Exploiting recombination in single bacteria to make large phage antibody libraries. Nat Biotechnol 2000; 18:75-80; PMID:10625396; http://dx.doi.org/10.1038/71958
37. Zemlin M, Klinger M, Link J, Zemlin C, Bauer K, Engler JA, Schroeder HW Jr., Kirkham PM. Expressed murine and human CDR-H3 intervals of equal length exhibit distinct repertoires that differ in their amino acid composition and predicted range of structures. J Mol Biol 2003; 334:733-49; PMID:14636599; http://dx.doi.org/10.1016/j.jmb.2003.10.007
38. Vollmers C, Sit RV, Weinstein JA, Dekker CL, Quake SR. Genetic measurement of memory B-cell recall using antibody repertoire sequencing. Proc Natl Acad Sci U S A 2013; 110:13463-8; PMID:23898164; http://dx.doi.org/10.1073/pnas.1312146110
39. Nekrutenko A, Taylor J. Next-generation sequencing data interpretation: enhancing reproducibility and accessibility. Nat Rev Genet 2012; 13:667-72; PMID:22898652; http://dx.doi.org/10.1038/nrg3305
40. Ferrara F, Naranjo LA, D'Angelo S, Kiss C, Bradbury AR. Specific binder for Lightning-Link® biotinylated proteins from an antibody phage library. J Immunol Methods 2013; 395:83-7; PMID:23850993; http://dx.doi.org/10.1016/j.jim.2013.06.010