Published online 15 September 2010

Nucleic Acids Research, 2010, Vol. 38, No. 21 ***e193***
doi:10.1093/nar/gkq789

By-passing *in vitro* screening—next generation sequencing technologies applied to antibody display and *in silico* candidate selection

U. Ravn[1], F. Gueneau[1], L. Baerlocher[2], M. Osteras[2], M. Desmurs[1], P. Malinge[1], G. Magistrelli[1], L. Farinelli[2], M.H. Kosco-Vilbois[1] and N. Fischer[1,*]

[1]NovImmune SA, Ch. des Aulx 14 and [2]Fasteris SA, Ch. du Pont-du-Centenaire 109, 1228 Plan-les-Ouates, Switzerland

Received April 23, 2010; Revised August 17, 2010; Accepted August 23, 2010

ABSTRACT

In recent years, unprecedented DNA sequencing capacity provided by next generation sequencing (NGS) has revolutionized genomic research. Combining the Illumina sequencing platform and a scFv library designed to confine diversity to both CDR3, $>1.9 \times 10^7$ sequences have been generated. This approach allowed for in depth analysis of the library's diversity, provided sequence information on virtually all scFv during selection for binding to two targets and a global view of these enrichment processes. Using the most frequent heavy chain CDR3 sequences, primers were designed to rescue scFv from the third selection round. Identification, based on sequence frequency, retrieved the most potent scFv and valuable candidates that were missed using classical *in vitro* screening. Thus, by combining NGS with display technologies, laborious and time consuming upfront screening can be by-passed or complemented and valuable insights into the selection process can be obtained to improve library design and understanding of antibody repertoires.

INTRODUCTION

In vitro display technologies have provided a powerful means for generating and evolving proteins with novel properties. In most cases, the aim of such approaches has been to isolate novel binding specificities from peptides, antibody fragments or alternative scaffold protein repertoires (1–5). For the past two decades, phage display has been an invaluable technology for *in vitro* evolution and identification of peptides and proteins (6–9). Despite the development of alternative display technologies, such as bacterial display, yeast display and ribosome display, the robustness of filamentous bacteriophage M13 makes it one of the most widely used approaches for academic centers as well as the biopharmaceutical industry. For instance, libraries of antibody fragments displayed on phage have delivered several fully human monoclonal antibodies that are currently in clinical trials, proving the significant contribution of phage display to the success of this class of therapeutic molecules (6,10–13).

In vitro display and selection approaches involve three main steps: (i) the generation of a large collection of variants (a library); (ii) multiple rounds of enrichment of variants having the desired properties via the genotype–phenotype linkage provided by the display system used; and (iii) functional screening and characterization of selected variants using appropriate assays. At each of these steps, analysis of variants via Sanger sequencing is commonly used to control the process and identify sequences of interest. In recent years, the development of next generation sequencing (NGS) technologies has revolutionized multiple aspects of biological research (14–16). These sequencing platforms also have the potential to profoundly impact the display and selection process of proteins with desired properties as follows.

At the library generation stage, it is crucial to cover as much sequence and structural diversity as possible to increase the likelihood of including protein variants with desired properties. The diversity of phage display libraries typically lies between 10^7 and 10^{11} (17). Sequencing several hundred members from the library is usually performed to evaluate the number of clones that are different and in the correct reading frame, reflecting the diversity and quality of the library. A major limitation using Sanger sequencing is that only a minute fraction of the library is actually sampled (a few hundred at best, out of $>10^7$ clones). The large number of sequencing reads delivered by NGS

*To whom correspondence should be addressed. Tel: +41 22 5935184; Fax: +41 22 8397154; Email: nfischer@novimmune.com

The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

technologies (i.e. $>10^6$) is in principle ideally suited for the analysis of vast numbers of library members and a much more extensive evaluation of library diversity and quality. Similarly, during enrichment via multiple selection rounds, sequencing of a limited number of clones only provides a glimpse into the enrichment process and is only used to determine which selection round should be used for the screening step. The capacity to obtain sequence information on far more if not all clones at each round would offer a virtually comprehensive analysis of the selection process, potentially rendering the screening step unnecessary.

NGS has recently been applied to analyze of the immunoglobulin repertoires of zebrafish and humans (18–20). The sequence diversity of immunoglobulins captured from natural repertoires is spread across the six complementary determining regions (CDR) of the heavy and light chain variable domain. Therefore, relatively long reads (i.e. >300 bases) are needed to cover the entire sequence of an immunoglobulin variable domain and require the use of pyrosequencing. However, although producing longer reads, pyrosequencing is currently limited to 10^6 reads per run while other technologies can deliver >10-fold more reads albeit of much shorter lengths (30–100 bases) (15). Therefore, in this study, we applied the Illumina sequencing platform to a specially designed scFv library. Our approach allowed for the in-depth analysis of the library, extensive coverage of sequences at each selection round and ability to follow enrichment during two independent selection processes. Based solely on sequence information, we isolated target specific antibody fragments including some that were missed. Taken together our approach demonstrates a powerful combination which can completely by-pass the need for a primary screening step.

MATERIALS AND METHODS

Library construction

Human VH and VL germlines were amplified from human genomes of Jurkat, HeLa and HEK 293 cells from NEB by PCR. The fragments were extended by PCR (Supplementary Data) to introduce a CDR3 stuffer, a human FR4 sequences and the restriction sites NcoI/XhoI for VH and SalI/NotI for VL. These restriction sites were used for cloning into the phagemid vector pNDS1. The pNDS1 vector that was derived from pHEN1 phagemid vector (21). The cloning site contains a $(Gly_4Ser)_3$ linker sequence flanked by the restriction enzyme sites NcoI/XhoI for VH cloning and SalI/NotI for VL cloning. The NotI site is followed by a hexahistidine tag, a c-myc tag and an amber stop codon. The CDR3 stuffers introduce a frameshift impairing the expression of any functional scFv. The stuffers do also contain two BsmBI restriction sites for the cloning of diversified CDR3 sequences. The synthetic CDR3 diversity was generated by PCR assembly using degenerated oligonucleotides (see Supplementary Data) with codons NNS, DVK, NVT or DVT depending on the CDR3 length. The CDR3 were randomized on 4–10 residues. The resulting cassettes create synthetic diversity both in CDR3 sequence and length. As the oligos were biotinylated (Microsynth) the digested inserts were purified using StreptaBeads (Dynal). After a phenol/chloroform extraction step, they were precipitated with ethanol and resuspended in H_2O. Using BsmBI (NEB) restriction enzyme Type IIS the synthetic diversity was introduced into the respective VH and VL acceptor vectors without any change in the framework sequences. The VH and VL sublibraries were recombined using XhoI/NotI restriction sites. Recombinant pNDS1 was electroporated into *Escherichia coli* TG1 cells.

Phage selections

TG1 cells were grown at 37°C (240 rpm) in 2xTYAG (100 μg/ml ampicillin, 2% glucose) medium. At $OD_{600} = 0.4–0.5$ the AE1 library was rescued by super-infection with M13K07 helper phage for 1 h at 37°C (100 rpm). Culture medium was then changed for 2xTYAK (100 μg/ml ampicillin, 50 μg/ml kanamycin) and TG1 were grown o/n at 30°C (280 rpm). Phage were purified and concentrated from the culture supernatant by two precipitations with one-third v/v of 20% PEG-8000/2.5 M NaCl (Sigma) and resuspended in TE buffer, dialyzed against TE buffer and titrated by infecting TG1 cells. Phage (10^{12} pfu) were blocked with phosphate buffered saline (PBS) containing 3% (w/v) skimmed milk and deselected on immunotubes (Nunc) coated with a rat IgG2b isotype antibody for selections against 5E3 and on streptavidin coated magnetic beads (Invitrogen) for selection against human interferon γ (hIFNγ). Deselected phage were incubated for 2 h (RT) in either immunotubes coated with 10 μg/ml of the anti-mouse TLR4 rat monoclonal antibody 5E3 or with Streptabeads (Invitrogen) coated with 200 nM biotinylated hIFNγ. Non-specific phage was eliminated by five washes with PBS/0.1% Tween-20 and two washes with PBS. Bound phage were eluted with 10 mM triethylamine TEA (Sigma), neutralized with 1M Tris–HCl pH 7.4 (Sigma). The eluate was added to 10 ml of exponentially growing *E.coli* TG1 cells and incubated for 1 h at 37°C (100 rpm). An aliquot of the infected TG1 was serial diluted to titer the selection output. The remaining infected TG1 were spread on 2xTYAG agar Bioassay plates. After overnight incubation at 30°C, bacteria were scraped off with 2xTY medium and aliquots were stored at −80°C in 17% glycerol. For subsequent rounds of panning 10^{10} pfu were used.

Sequencing using the Illumina platform

For the VH, we used FR4 specific (5′-CCTGGCCCCAAT AATC-3′) and M13Rev primers (5′-AACAGCTATGAC CATG-3′) to prepare initial PCR products on which the Genome Analyzer adapters were added in two 5 cycles PCR steps using the Phusion polymerase (Finnzymes): The first amplification added the Illumina Genomic Sequencing primer sequence (SBS) and a four bases bar-code to the 5′-side of FR4 and the Illumina P7

sequence to the 5′-side of M13Rev, while the second amplification completed the library construction with a primer adding the Illumina P5 sequence 5′ to the SBS end and a P7 primer. A similar approach was used for the VL, either on the original libraries or on the same FR4/M13Rev initial PCR products. The two step addition of the Genome Analyzer P5-SBS or P7 sequences occurred on the 5′-side of nested primers 5′-ATGATGATGTGCGGC-3′ or 5′-TTAGATTATTGGGGCCAGG-3′, respectively. Cloning a 1 µl aliquot of the purified final products into a pCR-TOPO-Blunt plasmid and capillary sequencing eight clones controlled the quality of the Genome Analyzer-ready libraries.

The VH bar-coded libraries were sequenced on a Genome Analyzer GAII instrument following Illumina's standard procedures, with cluster generation kit v. 2.0 and sequencing kits v.3.0, multiplexed in one single-reads channel for 76 cycles. The base-calling was performed using the GAPipeline 1.4.0. The VL bar-coded libraries were sequenced on the same instrument, but with cluster generation kit v. 4.0 and sequencing kits v. 4.0, multiplexed in one single-reads channel for 76 cycles. Base-calling was performed using the GAPipeline 1.5.1. In all cases, the run quality is monitored using a standard procedure from Illumina, which is to analyze a control DNA supplied by the manufacturer and determine the error rate of the resulting sequences.

Screening ELISA

Individual clones were grown in 2xTYAG medium in 96-well plates. scFv expression was induced with IPTG (1 mM) overnight at 30°C (150 rpm). The supernatants containing scFv were used in ELISA to evaluate their binding specificity and affinity on 5E3. The 96-well MaxiSorp plates (Nunc) were coated overnight (4°C) with 50 ng/well of 5E3 or a rat isotype antibody. The supernatants and the assay plates were blocked with PBS/3% milk for 1 h (RT). After washing the assay plates three times with PBS/0.05% Tween-20, 50 µl of the blocked supernatants containing scFv were transferred to the wells and incubated 2 h at RT. Binding scFv were detected with mouse anti-cmyc and anti-mouse IgG Fcγ-HRP antibodies. The assay was developed with TMB substrate (Sigma) and the reaction stopped with H_2SO_4 2N and the absorbance read at 450 nm.

VH_{CDR3} based rescue

Sense and anti-sense primers specific for the VH_{CDR3} were used in combination with M13Fwd and M13Rev primers on 15 ng of template DNA from TG1 cells obtained after Round 3. The two resulting DNA fragments were assembled by PCR, digested with NcoI and NotI, ligated into pNDS1 and transformed into TG1 cells. Minipreps of the resulting transformation were sequenced by Sanger sequencing. Their sequences were aligned using the Sequencher vs.4.8 (Gene Codes) software.

RESULTS

Generation of a scFv library with diversity restricted to CDR3

The two complementary determining regions 3 (CDR3) of an immunoglobulin molecule form the center of the antigen combining site and, although the other complementary determining regions (CDR) often contribute to the binding energy, diversification of the CDR3 regions is sufficient to generate antigen specific antibodies (22–24). Thus we have generated a library of human scFv based on a limited number of human immunoglobulin variable genes in which diversity was introduced only in CDR3 of the heavy and light chains. Seven human germline genes were selected for both the heavy and light chain variable regions (VH1-2, VH1-18, VH1-69, VH3-30.3, VH3-48, VH3-23, VH5-51, VK1-33, VK1-39, VK3-11, VK3-15, VK3-20, VL1-44, VL1-51; nomenclature according to the IMGT database) based on their frequency in natural human repertoire, their stability as well as having different CDR1 and CDR2 sequences in order to include some germline encoded diversity in these regions. At the 3′-end of the variable genes, a DNA sequence that contains two BsmBI restriction sites was introduced, changing the reading frame so that the resulting construct cannot encode a functional immunoglobulin variable region. These heavy and light chain 'acceptor frameworks' were used to introduce diversified sequences via Type IIS restriction cloning, restoring the correct reading frame of an antibody variable region (Figure 1A). Diversified sequences encoding heavy chain CDR3 of 9–15 amino acids in length (according to the IMGT definition of CDR) were introduced into the seven VH acceptor frameworks (25). Similarly CDR3 of 8–11 amino acids were cloned into the VL acceptor frameworks. A final library of 7×10^9 transformants, named AE1, was obtained by combining the diversified heavy and light chain variable regions into the phagemid vector pNDS1, allowing for scFv expression and display at the surface of M13 filamentous bacteriophage (7).

Next generation sequencing of the antibody library and selection rounds

The AE1 library was used for the identification of anti-idiotype scFv fragments directed against the rat anti-mouse TLR4 monoclonal antibody, 5E3 (26). Three rounds of selection were performed using a classical panning approach against 5E3 immobilized on immunotubes. At each round, the input phage was first deselected against another immobilized rat antibody of the same isotype (IgG2b) in order to drive the selection towards the variable regions of 5E3.

Currently, the Illumina sequencing platform generates reads of about 76 bp (15). The oligonucleotide primers in our study were specific to a region common to all clones in order to avoid biases. For the heavy chain, the primer was specific for FR4 and the sequencing read covered the whole CDR3 and in most cases provided enough FR3 sequence information to identify the VH family (Figure 1A). As the λ and κ light chains have different

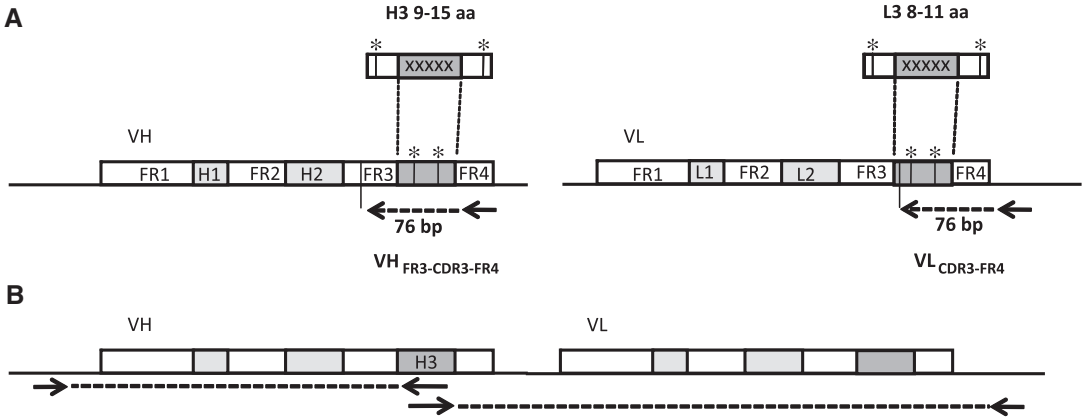


Figure 1. Schematic representation of immunoglobulin heavy and light chain variable regions and CDR3 diversification strategy. (**A**) Framework regions (FR1 to FR4) and CDR regions (H1 to H3 and L1 to L3) are indicated. Stars indicate the location of Type IIS restriction sites in the stuffer fragment located between FR3 and FR4 and in the flanking regions of the diversified VH_{CDR3} and L3. The sizes of the designed VH_{CDR3} and L3 are 9–15 and 8–11 amino acids, respectively. The arrows indicate the location of the primers used for next generation sequencing and the dashed arrows indicate the region covered by the 76 bp reads. The heavy chain primer, which is located in the FR4 region, permits partial sequencing of the FR3 region ($VH_{FR4\text{-}CDR3\text{-}FR3}$). In contrast, as the light chain primer is located downstream of the FR4 region, the sequencing read does not cover the full CDR3 sequence ($VL_{CDR3\text{-}FR4}$). (**B**) ScFv rescue strategy based on VH_{CDR3} sequences. Arrows indicate complementary primers corresponding to a VH_{CDR3} sequence as well as primers flanking the scFv coding region. These primers were used to amplify and assemble the scFv sequence from the pool of clones following the third selection round.

Table 1. Summary of NGS results for the heavy and light chains

	AE1 library (%)	Round 1 (%)	Round 2 (%)	Round 3 (%)
Selection on 5E3				
$VH_{(FR3\text{-}CDR3\text{-}FR4)}$				
Total number of sequences	5 078 705	352 778	561 296	642 878
Unique sequences	5 007 022 (99)	124 909 (35)	130 382 (23)	105 011 (16)
Single occurrence sequences[a]	4 938 237 (99)	89 523 (72)	103 009 (79)	82 525 (79)
Repeated sequences[b]	68 785 (1)	35 386 (28)	27 373 (21)	22 486 (21)
Highest frequency	42 (0)	1439 (0.4)	9870 (2)	37 017 (6)
Identified VH family[c]	4 680 882 (92)	332 201 (94)	538 706 (96)	619 707 (96)
In frame inserts[d]	4 237 321 (91)	322 273 (97)	527 932 (98)	612 130 (99)
$VL_{(CDR3\text{-}FR4)}$				
Total number of sequences	4 412 636	1 531 261	1 302 154	1 649 977
Unique sequences	3 612 120 (82)	733 282 (48)	493 490 (21)	576 062 (19)
Identified VL group[e]	4 051 197 (92)	1 470 871 (96)	1 269 674 (98)	1 597 046 (97)
Selection on IFNγ				
$VH_{(FR3\text{-}CDR3\text{-}FR4)}$				
Total number of sequences		1 176 998	1 484 395	1 247 375
Unique sequences		1 075 049 (91)	167 900 (11)	67 936 (5)
Single occurrence sequences[a]		1 008 532 (91)	109 445 (65)	50 841 (75)
Repeated sequences[b]		66 517 (6)	58 455 (35)	17 095 (25)
Highest frequency		13 139 (1)	19 571 (1)	112 452 (9)
Identified VH family[c]		1 133 917 (96)	1 426 351 (96)	1 234 160 (99)
In frame inserts[d]		1 039 842 (92)	1 314 621 (92)	1 233 561 (100)

[a]Single occurrence sequences: number of sequences occurring a single time in the set of unique sequences.
[b]Repeated sequences: number of sequence occurring a multiple times in the set of unique sequences.
[c]Identified VH family: number of single occurrence sequences for which FR4 information allowed unambiguous VH family identification.
[d]In frame inserts: number of sequences with an identified VH family that contain an in frame insertion.
[e]Identified VL group: number of sequence that could be unambiguously identified as κ or λ based on FR4 sequence.

FR4 regions, the sequencing primer had to be located further downstream in the vector sequence. Therefore, the sequences unambiguously identified κ from λ light chains but only partially covered the CDR3 sequence (Figure 1A). Phagemid DNA was isolated from the AE1 library and bacteria recovered after each round of selection. The scFv coding regions were then amplified using primers introducing a 4-nt sequence tag allowing for simultaneous solid phase sequencing of DNA fragments from the library as well as different selection rounds with the same Illumina channel. The heavy and light chain CDR3 regions were sequenced in two independent runs and samples from the library and each round were mixed in a 7:1:1:1 ratio in order to obtain a maximum number of reads from the more diverse library while covering the diversity of each selection round. A total of 6 635 657 and 8 896 028 reads were obtained for heavy and light chain runs, respectively, representing more than one billion bases sequenced (Table 1).

Quality control of the antibody library

Next, the efficiency of our CDR3 diversification strategy was evaluated by analyzing over five million reads covering the $VH_{FR3\text{-}CDR3\text{-}FR4}$ region of the scFv (Table 1). Over 98.5% of the sequences were unique and 98.6% of those were a single copy in the library. A far smaller number of sequences were found in multiple copies and the most frequent sequence was found 42 times. A total of 92% reads could be attributed to a VH germline family and revealed that the VH1, VH3 and VH5 families were relatively equally represented (Figure 2A). From the sequences in which a sufficient length of the FR4 sequence could be obtained (allowing for a VH family assignment and an accurate reading frame determination) 91% of inserts were inframe (Table 1). Furthermore, the CDR3 lengths and distribution corresponded to the library's design although CDRs of 13 amino acids were slightly underrepresented (Figure 3A).

The analysis of the reads covering the $VL_{CDR3\text{-}FR4}$ region confirmed the expected 5:2 κ/λ ratio and the high diversity of the library with at least 82% representing unique sequences (Figure 2B, Table 1). This number underestimates the diversity as a fraction of the CDR3 sequences were not fully covered by the 76-bp reads (Figure 1A). Thus, high throughput sequence analysis demonstrated that our library construction strategy generated a high percentage of diverse and in frame coding sequences. As the high throughput sequencing data probed a limited portion of the scFv sequence, we confirmed these findings by full length Sanger sequencing of 132 randomly picked library members (data not shown).

Sequence evolution during phage selection against 5E3

During the selection of phage binding to the target, 5E3, the number of phage particles recovered after the first, second and third rounds was 1.8×10^5, 3.4×10^5 and 1.1×10^6, respectively. Therefore, as the maximal number of different sequences present at later selection rounds is limited by the output of the first round and in this case lies in the 10^5 range. Using NGS, between 3.5×10^5 and 6.4×10^5 reads for the VH and $>10^6$ reads for the VL were obtained for each selection round (Table 1). As such, this very large fraction of clones

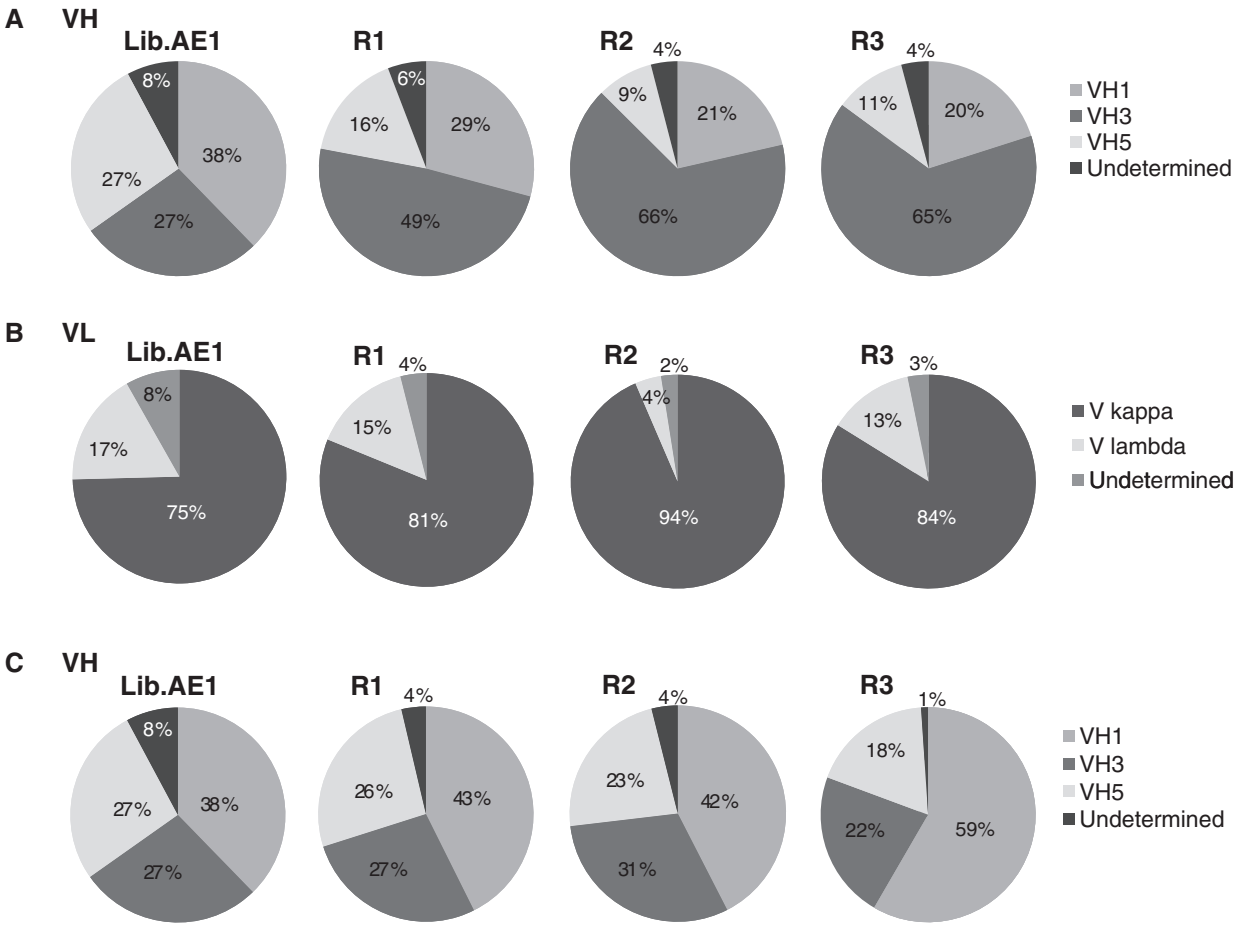


Figure 2. Germline gene family analysis. (**A**) Frequency of heavy chain variable gene families identified and (**B**) proportion of V κ and V λ light chains in the AE1 library and after each selection round (R1–R3) against the target, 5E3. (**C**) Frequency of heavy chain variable gene families identified in the AE1 library and after each selection round (R1–R3) against the target, hIFNγ. Sequences were considered as undetermined if they did not match exactly the signature sequence used for family assignment.

sequenced at each round provided an unprecedented view of the frequency and evolution of individual VH and VL sequences throughout the process.

Next, assessing VH germline genes, VH3 significantly increased from 27% (in the AE1 library) to 65% (after the third round of selection, Figure 2A). This result was not unexpected as VH3 genes encoded antibody fragments have been shown to be well tolerated in display settings and are often used for the generation of semi-synthetic libraries (23,24,27). The κ to λ ratio remained relatively unchanged throughout the selection process although a reduced frequency of λ chains was observed after the second round (Figure 2B). The distribution of VH_{CDR3} lengths changed during the selection process with an enrichment of CDRs of 11 amino acids or more and a marked reduction in CDRs of 9 and 10 residues (Figure 3). Clones containing out of frame VH_{CDR3} and therefore encoding non-functional scFv, were lost during selection (Table 1, Figure 3). Interestingly, VH_{CDR3} of eight amino acids were enriched after the first selection round and reached 10% after the third round (Figure 3D). As the shortest VH_{CDR3} design in the library was nine amino acids in length, these sequences were not theoretically included in the library. Therefore they are probably generated due to errors occurring during oligonucleotide synthesis or are cloning artifacts (~9%, Table 1).

We also followed the enrichment profile of the 10 most frequent $VH_{FR3\text{-}CDR3\text{-}FR4}$ sequences at Round 3 which were present in a range between 37 017 and 2815 times (Figure 4A). In most cases, the major enrichment was observed after the third round of selection except for sequence #3 which was more gradually enriched and already the most abundant after Rounds 1 and 2 (Figure 4A). These 10 top sequences accounted for >17% of all sequences after Round 3 and the three most abundant represented 14% suggesting that these should be readily identified during screening for binding to the target.

Hit identification by classical primary target binding screening

As classically performed in a phage selection program, 96 individual clones from the third round of selection were picked and scFv secretion into the medium induced in order to screen by ELISA for specific binding to the target, 5E3. Positive clones were sequenced and grown in larger volumes in order to purify the scFv and confirm their specificity in dose response experiments. Six scFv were identified having a binding EC_{50} in the nanomolar range for 5E3 versus a control protein (Figure 5). We next determined the frequency of the $VH_{FR3\text{-}CDR3\text{-}FR4}$ and $VL_{CDR3\text{-}FR4}$ sequences of these clones in the AE1 library and selection rounds. As expected, these sequences were enriched during selection and maximal frequencies of 2–6% were observed after the third rounds for the scFv D11 and D6 (Figure 5). These two clones also had the highest apparent binding affinity in the dose response ELISA indicating that the selection process, based on binding to a target, indeed enriched for phage having a high affinity for the target (Figure 5). The parallel increase in frequency of VH_{CDR3} and VL_{CDR3} of each clone suggests that it is due to the enrichment of the selected clone carrying both CDR3 sequences (Figure 5). However, as the light and heavy chain sequence information was generated independently, we cannot exclude a contribution of clones carrying only one of the CDR3 (i.e. VH_{CDR3} or VL_{CDR3}) in combination with a different sequence.

ScFv rescue via VH_{CDR3} amplification

The $VH_{FR3\text{-}CDR3\text{-}FR4}$ sequences of the six scFv identified by screening could all be identified in the sequencing data and their respective frequencies after the third round of selection determined (Figure 5). The sequences corresponding to clone D11, D6 and D4 were found among the 10 most frequent sequences and D11 that has the highest apparent affinity was also the most enriched during selection (Figure 4). However, several sequences, frequently found after Round 3, were not identified during the ELISA screening. In particular, clone #2 representing the second most abundant sequence with close to 6% of all sequences at Round 3, had not been identified (Figure 4). We wanted to understand whether these clones had been missed because of the limited number of scFv that were tested or if there were some characteristics that prevent their identification by an ELISA screening approach. We therefore aimed at rescuing scFv bearing the six most frequent VH_{CDR3} sequences by PCR amplification. These included three scFv identified by screening (5E3R-1, 5E3R-3 and 5E3R-6) as well as three scFv that had not been identified (5E3R-2, 5E3R-4 and 5E3R-5, Figure 4A) Complementary pairs of oligonucleotides primers specific for these six VH_{CDR3} were designed and used in combination with primers located upstream and downstream of the scFv coding region in order to recover from the output of Round 3 the complete scFv sequence, or sequences, bearing these frequent VH_{CDR3} (Figure 1B). The amplification products were cloned into the pNDS1 vector and 10 independent clones for each rescued scFv were sequenced. All clones contained the same pair of VH_{CDR3} and VL_{CDR3} further suggesting that the observed parallel enrichment of two VH_{CDR3} and VL_{CDR3} sequences was mainly due to the enrichment of a single clone and not of various scFv bearing a VH_{CDR3} in combination with a variety of VL_{CDR3} sequences.

These six rescued scFv were expressed and tested for binding to the target, 5E3 using ELISA but with different formats. We first performed an ELISA using phage displaying the scFv at its surface to test the binding in the same context as during the selection process. Binding of secreted scFv was also tested using bacterial supernatant or periplasmic fractions, the latter preparation providing a more concentrated source of crude scFv. Finally, each clone was expressed at a larger scale and purified from the periplasmic space using immobilized metal ion chromatography, quantified and verified on SDS–PAGE for integrity and purity. The purified scFv were then used in dose-response experiments to determine their apparent binding affinity. The results from these experiments using different formats and sources are summarized in

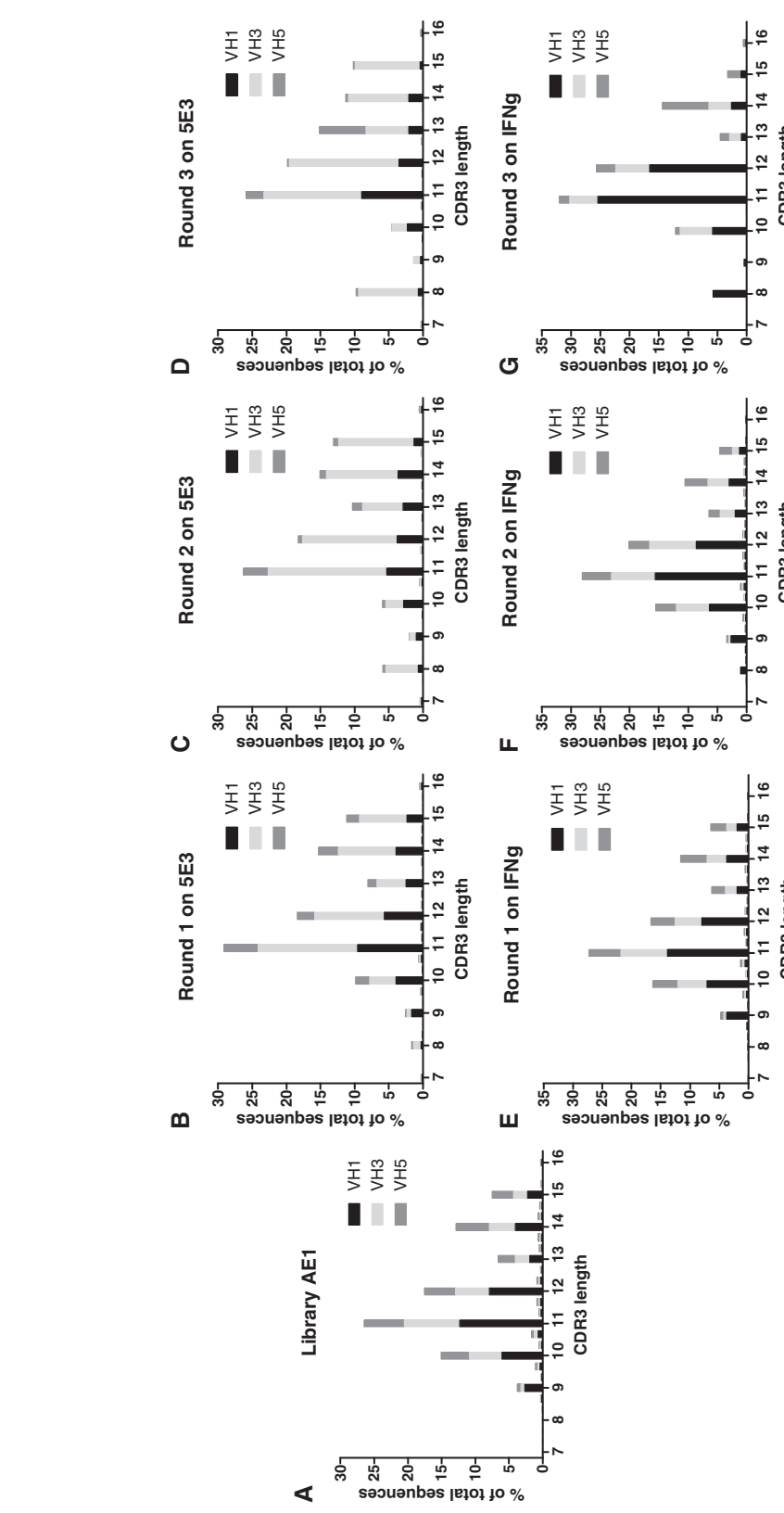
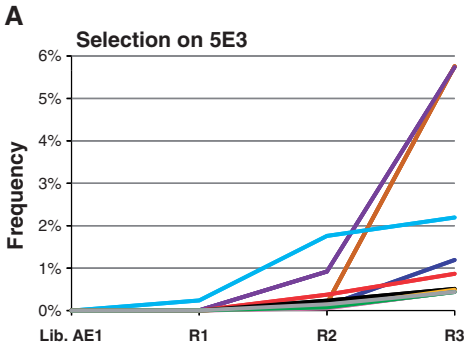
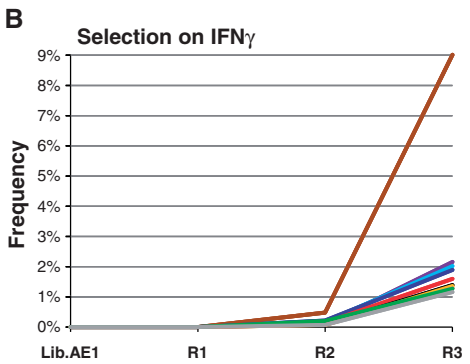


Figure 3. Frequency of VH_{CDR3} lengths and distribution within the three VH families included in the AE1 library and after each selection round against different targets. VH_{CDR3} lengths are expressed as amino acids. VH_{CDR3} lengths of 9–15 amino acids were included in the library design. Bars appearing between these defined lengths correspond to non-functional VH_{CDR3} that are out of frame due to errors in oligonucleotide synthesis or cloning artifacts. (**A**) library, (**C–D**) selection rounds against 5E3, (**E–G**) selection rounds against hIFNγ.



Clone #	VH CDR3 Sequence (IMGT)	Length	Family	Round 3 frequency	Rescued clones	Corresponding screening hit
1	ARGNRRAYMDY	11	VH1	37017	5E3R-1	D11
2	ARGTYANTTSMDY	13	VH5	36901	5E3R-2	
3	ARGDSNTDDADAMDY	15	VH3	14096	5E3R-3	D6
4	ARGRKKEKGMDY	12	VH3	7661	5E3R-4	
5	ARGRPSLFDY	10	VH1	5573	5E3R-5	
6	ARGRRTGGMDY	11	VH5	3290	5E3R-6	D4
7	ARSAGMDY	8	VH3	3159		
8	ARDSRDGTPMDY	12	VH3	2867		
9	ARRKKMDY	8	VH3	2818		
10	ARDAPTGDPFDY	12	VH1	2815		



Clone #	VH CDR3 Sequence (IMGT)	Length	Family	Round 3 frequency	Rescued clones
1	ARETDSWDTFDY	12	VH1	112452	IFNR-1
2	ARGGWATKFDY	11	VH1	26872	IFNR-2
3	ARYTYHPDGGDMDY	14	VH5	25199	IFNR-3
4	ARGSFSLMDY	10	VH3	23608	IFNR-4
5	ARYGYTASNSGAMDY	15	VH5	19892	IFNR-5
6	ARGSGYYTAGSFDY	14	VH5	17325	IFNR-6
7	ARGYWSASFDY	11	VH1	17004	IFNR-7
8	ARGWWSSNFDY	11	VH1	16001	IFNR-8
9	ARGHDRRSGDFDY	13	VH5	15896	IFNR-9
10	ARGGWGASFDY	11	VH1	14459	IFNR-10

Figure 4. Frequency and evolution of top 10 sequences. Frequency of the 10 VH_{CDR3} sequences that were the most abundant after the third selection round against 5E3 (**A**) and against hIFNγ. (**B**) Amino acid sequence, length of the VH_{CDR3} according to the IMGT nomenclature and VH family are shown. Sequences that were also identified during screening by ELISA as well as clones that were rescued based on their VH_{CDR3} sequence and frequency are indicated.

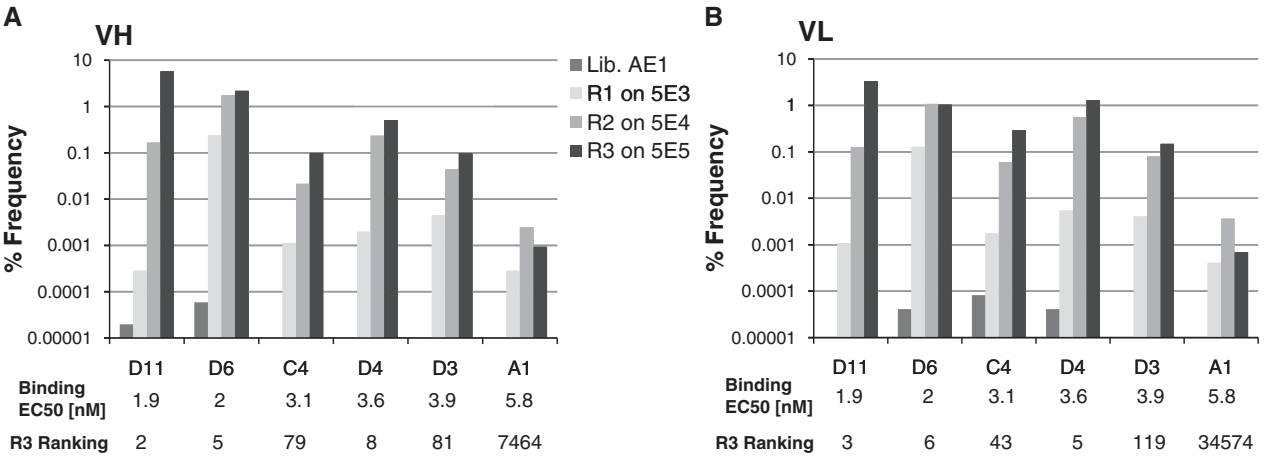


Figure 5. Frequency and evolution of CDR3 sequences identified by classical binding screening against the target 5E3. Frequency evolution of $VH_{FR4\text{-}CDR3\text{-}FR3}$ (**A**) and $VL_{CDR3\text{-}FR4}$ (**B**) sequences corresponding to six scFv binding specifically to the monoclonal 5E3. The EC_{50} for binding of the soluble scFv to the target in ELISA and the frequency ranking after the third round of selection are indicated below each clone.

Table 2. All scFv could bind to 5E3 when displayed on phage in agreement with their enrichment during the selection process. In contrast, only the 5E3R-1, 5E3R-3 and 5E3R-6 scFv gave a positive signal using crude supernatants or periplasmic preparations providing an explanation to why the 5E3R-2, 5E3R-4 and 5E3R-5 clones were not identified during the screening step. These three scFv were probably not well expressed or secreted as reflected by the lower yield obtained after large scale production (Table 2). Interestingly, 5E3R-4 and 5E3R-5 bound 5E3 in scFv format with EC_{50} of 2.3 and 140 nM, respectively. As a soluble scFv, 5E3R-2 did not give any signal, suggesting that it requires the phage context for binding to the target. These results indicate that by using a rescue approach based on VH_{CDR3} sequence frequency, we retrieved the best scFv candidates identified by ELISA approach and,

Table 2. Binding experiments for scFv displayed on the surface of phage or expressed in different soluble formats

scFv	Phage	scFv Supernatant	scFv Periplasmic fraction	Purified scFv EC_{50} (nM)	scFv-yield (mg/l)
5E3R-1	+	+	+	1.9	5
5E3R-2	+	–	–	–	0.25
5E3R-3	+	+	+	2	2.2
5E3R-4	+	–	–	2.3	0.34
5E3R-5	+	–	–	140	0.21
5E3R-6	+	+	+	3.6	0.4
IFNR-1	+	–	–	37.9	5
IFNR-2	+	–	–	41.0	13.4
IFNR-3	+	+	+	1.3	1.7
IFNR-4	+	+	+	0.16	10.6
IFNR-5	+	+	+	1.4	0.2
IFNR-6	+	+	+	0.75	6.3
IFNR-7	+	+	+	0.36	2.9
IFNR-8	+	–	–	139.3	16
IFNR-9	+	–	–	4.7	7.5
IFNR-10	+	+	+	0.1	23.8

more importantly, we could obtain two additional candidates that were missed using a classical screening approach.

Bypassing primary screening

In order to further validate the approach, we applied the same procedure against another target. We performed three rounds of phage selection using the AE1 library against soluble biotinylated hIFNγ. The output of each round was sequenced and over 1 million reads covering the $VH_{FR3\text{-}CDR3\text{-}FR4}$ region were obtained for each output. As expected, the number of unique sequences diminished and the frequency of repeated sequences increased to reach 25% after the third selection round indicating that selection had occurred (Table 1). The VH families could be identified for >95% of the sequences and, in this case, a clear enrichment for VH1 was observed (Figure 2C). Similarly, the VH_{CDR3} lengths distribution was different compared to that observed during selection against 5E3, thus indicating that different CDR lengths were preferentially enriched in a target dependent manner (Figure 3E–G). We then rescued the 10 most frequent clones after the third round by overlapping PCR using primers matching their respective VH_{CDR3} sequences (Figure 4B). These rescued scFv were expressed in soluble form or displayed at the surface of filamentous bacteriophage to characterize their binding properties. All the candidates were able to bind hIFNγ in a specific manner using phage or purified scFv (Table 2). However, four candidates (IFNR-1, 2, 8 and 9) did not give any signal when supernatants or periplasmic preparations from 96-well plates were tested in ELISA, indicating that these candidates would not have been identified in a primary screening step. This second example further demonstrated that antibody candidates can be retrieved without upfront screening of randomly picked clones. In addition, a significant proportion of the most enriched clones (4 out of 10), that would have been repeatedly screened given their high frequency but missed by a standard screening approach, could be identified by this approach.

DISCUSSION

In only a few years, the emergence of NSG technologies has profoundly modified the landscape of whole genome analysis as well as gene expression profiling and will certainly impact many other areas of research (15). The large amount of sequencing data delivered by these platforms is ideally suited for extensive analysis of complex collections of diversified gene segments such as antibody or peptide libraries. The currently available platforms that are capable of providing several million reads per run, generate only short reads of up to 100 bp (15). We therefore combined a synthetic scFv library, containing diversity confined to VH_{CDR3} and VL_{CDR3}, with the Illumina NGS platform. We generated over 9 million reads covering the diversified CDR3 and, although only 0.22% of the library members were sampled, it is the most extensive sequence analysis of an antibody library reported to date. Furthermore, we demonstrated the effectiveness of our library building strategy via a Type IIS restriction enzyme cloning, resulting in >90% of in frame inserts, which is superior to several described library construction methods (28–31).

For the first time, we illustrate that it is possible to follow the evolution of virtually all VH_{CDR3} and VL_{CDR3} sequences during a phage display selection process. By comparing the sequences of hits identified by ELISA screening with VH_{CDR3} and VL_{CDR3} frequencies, we found that apparent binding affinity and enrichment correlated for several clones. However, it was clear that some highly enriched clones were missed during primary screening and the antibodies encoded by these 'lost' clones were valuable candidates.

In recent years, much effort has been spent in order to improve *in vitro* evolution approaches by optimizing or simplifying each step (i.e. library generation, selection and screening). For instance, it has been shown that selection rounds can be drastically reduced—and potentially even skipped—using high-throughput antibody array screening (32). Here, using the capacity of a NGS platform, it is feasible to completely by-pass primary screening, focusing on the most frequent sequences to proceed directly to in depth characterization in more relevant assays. Furthermore, target specific scFv that were not identified via a classical ELISA screening could be identified based on their frequency. All the scFv described in this study are capable of binding to the target when displayed on phage and give positive signals in phage ELISA (Table 2). This is expected as they were enriched during the phage selection process which is driven by binding to the target. However, a significant proportion of clones were negative in classical soluble scFv ELISA screening because the suboptimal growth conditions of a microtiter plate format do not support sufficient expression of these scFv. In contrast, we showed that all the clones rescued, based on sequencing

and frequency analysis, can be expressed and purified as soluble scFv from larger culture volumes with more optimal aeration and growth conditions. Furthermore, the large majority of these purified scFv were positive in dose response ELISA with IC_{50} in the nanomolar range (Table 2). The capacity to identify more candidates from a selection process is significant as it is important to recover a wide diversity of scFv to maximize epitope coverage on the target protein in order to bind functionally relevant epitopes, regardless of the initial affinity of the scFv.

This approach also significantly streamlines projects by removing the need for developing robust screening assays that must be designed and optimized for each target whereas the procedure for NGS will be the same for any target being considered. Laboratories tend to increase their screening capacity to be able to evaluate a maximum number of clones. This involves acquisition and time consuming implementation of expensive robotic and liquid handling equipment (33). It is obvious that random picking of clones for screening leads to the repetitive testing of scFv that have been the most enriched and thus a significant waste in terms of screening capacity. In contrast, comprehensive sequencing after enrichment directly provides information on all potential candidates. Here we have used frequency as a very simple method to select clones for rescue and characterization, but more elaborate sequence analysis including CDR length and amino acid composition should allow clustering of sequences and characterization of candidates based on different criteria. In addition, this approach is of particular interest for difficult or low abundance targets, such as proteins isolated from gels in the course of proteomic projects, that are not readily available in sufficient quantities to support screening campaigns (17,34).

A current limitation worth discussing, is that regions of the VH and VL chains were sequenced independently preventing a direct analysis of individual clones that are defined by a combination of VH_{CDR3} and VL_{CDR3} sequences along with the frameworks in which they have been inserted. The observed parallel increase in frequency of certain VH_{CDR3} and VL_{CDR3} provides indirect evidence for enrichment of clones bearing both CDRs. We demonstrated that the VH_{CDR3} sequence information is sufficient to rescue clones of interest, however, the approach could be further improved by using paired-end sequencing to reconcile VH_{CDR3} and VL_{CDR3} information (35). In addition, new sequencing reagents allowing for longer reads will also support a better coverage of the VL_{CDR3} sequences and allow VL family assignment.

It is clear that the selection of antibody CDR sequences is target dependent. However it remains to be fully elucidated whether there is a preferential usage by the immune system of certain variable gene families, CDR lengths and canonical structures for defined target classes (haptens, carbohydrates, peptides or proteins) (24,36,37). In this study, we observed that VH families were differentially enriched during selection against two protein targets. Comprehensive sequence analysis of selection rounds against different target classes may provide new insights into the composition of immune repertoires and allow for better and potentially more focused library designs (38).

The implications of NGS for *in vitro* display and selection of polypeptide variants will be many and can be widely applied to the study of molecular interactions (39). In the context of antibody discovery, they range from library quality control and improvement to target specific sequence evolution and better understanding of immunoglobulin repertoires. Among these, the possibility of by-passing upfront screening will streamline *in vitro* selection approaches, significantly increasing relevant output and thereby facilitating both fundamental and applied research.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

FUNDING

Funding for open access charge: NovImmune SA.

Conflict of interest statement. U.R., F.G., M.D., P.M., G.M., M.K.V. and N.F. are employees of NovImmune SA. L.B., M.O. and L.F. are employees of Fasteris SA.

REFERENCES

1. Amstutz,P., Forrer,P., Zahnd,C. and Plückthun,A. (2001) In vitro display technologies: novel developments and applications. *Curr. Opin. Biotechnol.*, **12**, 400–405.
2. Leemhuis,H., Stein,V., Griffiths,A.D. and Hollfelder,F. (2005) New genotype-phenotype linkages for directed evolution of functional proteins. *Curr. Opin. Struct. Biol.*, **15**, 472–478.
3. Gebauer,M. and Skerra,A. (2009) Engineered protein scaffolds as next-generation antibody therapeutics. *Curr. Opin. Chem. Biol.*, **13**, 245–255.
4. Binz,H.K., Amstutz,P. and Plückthun,A. (2005) Engineering novel binding proteins from nonimmunoglobulin domains. *Nat. Biotechnol.*, **23**, 1257–1268.
5. Hoogenboom,H.R. (2005) Selecting and screening recombinant antibody libraries. *Nat. Biotechnol.*, **23**, 1105–1116.
6. Bradbury,A.R. and Marks,J.D. (2004) Antibodies from phage antibody libraries. *J. Immunol. Methods*, **290**, 29–49.
7. Carmen,S. and Jermutus,L. (2002) Concepts in antibody phage display. *Brief. Funct. Genomic Proteomic.*, **1**, 189–203.
8. Smith,G.P. and Scott,J.K. (1993) Libraries of peptides and proteins displayed on filamentous phage. *Methods Enzymol.*, **217**, 228–257.
9. Jestin,J.L. (2008) Functional cloning by phage display. *Biochimie*, **90**, 1273–1278.
10. Hanes,J., Jermutus,L. and Plückthun,A. (2000) Selecting and evolving functional proteins in vitro by ribosome display. *Methods Enzymol.*, **328**, 404–430.
11. Reichert,J.M., Rosensweig,C.J., Faden,L.B. and Dewitz,M.C. (2005) Monoclonal antibody successes in the clinic. *Nat. Biotechnol.*, **23**, 1073–1078.
12. Reichert,J.M. (2009) Global antibody development trends. *MAbs.*, **1**, 86–87.
13. Thie,H., Meyer,T., Schirrmann,T., Hust,M. and Dübel,S. (2008) Phage display derived therapeutic antibodies. *Curr. Pharm. Biotechnol.*, **9**, 439–446.
14. Metzker,M.L. (2005) Emerging technologies in DNA sequencing. *Genome Res.*, **15**, 1767–1776.
15. Metzker,M.L. (2010) Sequencing technologies—the next generation. *Nat. Rev. Genet.*, **11**, 31–46.

16. McPherson,J.D. (2009) Next-generation gap. *Nat. Methods*, **6**, S2–S5.
17. Hust,M. and Dübel,S. (2004) Mating antibody phage display with proteomics. *Trends Biotechnol.*, **22**, 8–14.
18. Dias-Neto,E., Nunes,D.N., Giordano,R.J., Sun,J., Botz,G.H., Yang,K., Setubal,J.C., Pasqualini,R. and Arap,W. (2009) Next-generation phage display: integrating and comparing available molecular tools to enable cost-effective high-throughput analysis. *PLoS ONE*, **4**, e8338.
19. Weinstein,J.A., Jiang,N., White,R.A. III, Fisher,D.S. and Quake,S.R. (2009) High-throughput sequencing of the zebrafish antibody repertoire. *Science*, **324**, 807–810.
20. Glanville,J., Zhai,W., Berka,J., Telman,D., Huerta,G., Mehta,G.R., Ni,I., Mei,L., Sundar,P.D., Day,G.M. *et al.* (2009) Precise determination of the diversity of a combinatorial antibody library gives insight into the human immunoglobulin repertoire. *Proc. Natl Acad. Sci. USA*, **106**, 20216–20221.
21. Hoogenboom,H.R., Griffiths,A.D., Johnson,K.S., Chiswell,D.J., Hudson,P. and Winter,G. (1991) Multi-subunit proteins on the surface of filamentous phage: methodologies for displaying antibody (Fab) heavy and light chains. *Nucleic Acids Res.*, **19**, 4133–4137.
22. de Kruif,J., Boel,E. and Logtenberg,T. (1995) Selection and application of human single chain Fv antibody fragments from a semi-synthetic phage antibody display library with designed CDR3 regions. *J. Mol. Biol.*, **248**, 97–105.
23. Pini,A., Viti,F., Santucci,A., Carnemolla,B., Zardi,L., Neri,P. and Neri,D. (1998) Design and use of a phage display library. Human antibodies with subnanomolar affinity against a marker of angiogenesis eluted from a two-dimensional gel. *J. Biol. Chem.*, **273**, 21769–21776.
24. Wilson,I.A., Stanfield,R.L., Rini,J.M., Arevalo,J.H., Schulze-Gahmen,U., Fremont,D.H. and Stura,E.A. (1991) Structural aspects of antibodies and antibody-antigen complexes. *Ciba Found. Symp.*, **159**, 13–28.
25. Lefranc,M.P., Giudicelli,V., Ginestoux,C., Bodmer,J., Müller,W., Bontrop,R., Lemaitre,M., Malik,A., Barbie,V. and Chaume,D. (1999) IMGT, the international ImMunoGeneTics database. *Nucleic Acids Res.*, **27**, 209–212.
26. Daubeuf,B., Mathison,J., Spiller,S., Hugues,S., Herren,S., Ferlin,W., Kosco-Vilbois,M., Wagner,H., Kirschning,C.J., Ulevitch,R. *et al.* (2007) TLR4/MD-2 monoclonal antibody therapy affords protection in experimental models of septic shock. *J. Immunol.*, **179**, 6107–6114.
27. Griffiths,A.D., Williams,S.C., Hartley,O., Tomlinson,I.M., Waterhouse,P., Crosby,W.L., Kontermann,R.E., Jones,P.T., Low,N.M., Allison,T.J. *et al.* (1994) Isolation of high affinity human antibodies directly from large synthetic repertoires. *EMBO J.*, **13**, 3245–3260.
28. Persson,M.A., Caothien,R.H. and Burton,D.R. (1991) Generation of diverse high-affinity human monoclonal antibodies by repertoire cloning. *Proc. Natl Acad. Sci. USA*, **88**, 2432–2436.
29. Söderlind,E., Strandberg,L., Jirholt,P., Kobayashi,N., Alexeiva,V., Aberg,A.M., Nilsson,A., Jansson,B., Ohlin,M., Wingren,C. *et al.* (2000) Recombining germline-derived CDR sequences for creating diverse single-framework antibody libraries. *Nat. Biotechnol.*, **18**, 852–856.
30. Schoonbroodt,S., Frans,N., DeSouza,M., Eren,R., Priel,S., Brosh,N., Ben-Porath,J., Zauberman,A., Ilan,E., Dagan,S. *et al.* (2005) Oligonucleotide-assisted cleavage and ligation: a novel directional DNA cloning technology to capture cDNAs. Application in the construction of a human immune antibody phage-display library. *Nucleic Acids Res.*, **33**, e81.
31. Rothe,C., Urlinger,S., Löhning,C., Prassler,J., Stark,Y., Jäger,U., Hubner,B., Bardroff,M., Pradel,I., Boss,M. *et al.* (2008) The human combinatorial antibody library HuCAL GOLD combines diversification of all six CDRs according to the natural immune system with a novel display method for efficient selection of high-affinity antibodies. *J. Mol. Biol.*, **376**, 1182–1200.
32. de Wildt,R.M., Mundy,C.R., Gorick,B.D. and Tomlinson,I.M. (2000) Antibody arrays for high-throughput screening of antibody-antigen interactions. *Nat. Biotechnol.*, **18**, 989–994.
33. Bradbury,A., Velappan,N., Verzillo,V., Ovecka,M., Chasteen,L., Sblattero,D., Marzari,R., Lou,J., Siegel,R. and Pavlik,P. (2003) Antibodies in proteomics II: screening, high-throughput characterization and downstream applications. *Trends Biotechnol.*, **21**, 312–317.
34. Bradbury,A., Velappan,N., Verzillo,V., Ovecka,M., Chasteen,L., Sblattero,D., Marzari,R., Lou,J., Siegel,R. and Pavlik,P. (2003) Antibodies in proteomics I: generating antibodies. *Trends Biotechnol.*, **21**, 275–281.
35. Fullwood,M.J., Wei,C.L., Liu,E.T. and Ruan,Y. (2009) Next-generation DNA sequencing of paired-end tags (PET) for transcriptome and genome analyses. *Genome Res.*, **19**, 521–532.
36. Almagro,J.C. (2004) Identification of differences in the specificity-determining residues of antibodies that recognize antigens of different size: implications for the rational design of antibody repertoires. *J. Mol. Recognit.*, **17**, 132–143.
37. Wilson,I.A. and Stanfield,R.L. (1994) Antibody-antigen interactions: new structures and new conformational changes. *Curr. Opin. Struct. Biol.*, **4**, 857–867.
38. Persson,H., Lantto,J. and Ohlin,M. (2006) A focused antibody library for improved hapten recognition. *J. Mol. Biol.*, **357**, 607–620.
39. Di Niro,R., Sulic,A.M., Mignone,F., D'Angelo,S., Bordoni,R., Iacono,M., Marzari,R., Gaiotto,T., Lavric,M., Bradbury,A.R. *et al.* (2010) Rapid interactome profiling by massive sequencing. *Nucleic Acids Res.*, **38**, e110.