HHS Public Access

Author manuscript

N Biotechnol. Author manuscript; available in PMC 2019 October 25.

Published in final edited form as:

N Biotechnol. 2018 October 25; 45: 89–97. doi:10.1016/j.nbt.2017.12.005.

Rational library design by functional CDR resampling

Qi Zhao[†], **Diane Buhr**, **Courtney Gunter**, **Jenny Frenette**, **Mary Ferguson**, **Eric Sanford**, **Erika Holland**, **Chitra Rajagopal**, **Melissa Batonick**, **Margaret M. Kiss**, and **Michael P. Weiner**

Abcam plc. 688 E. Main Street, Branford, CT 06405, USA

Abstract

Successful antibody discovery relies on diversified libraries, where two aspects are implied, namely the absolute number of unique clones and the percentage of functional clones. Instead of pursuing the absolute quantity thresholded by current display technology, we have sought to maximize the effective diversity by improving functional clone percentage. With the combined effort of bioinformatics, structural biology, molecular immunology and phage display technology, we devised a bioinformatic pipeline to construct and validate libraries via combinatorial assembly of sequences from a database of experimentally validated antibodies. Furthermore, we showed that the libraries constructed as such yielded a significantly increased success rate against different antigen types and generated over 20-fold more unique hits per targets compared with libraries based on traditional degenerate nucleotide methods. Our study indicated that predefined CDR sequences with optimized CDR-framework compatibility could be a productive direction of functional library construction for in vitro antibody development.

Introduction

Directed evolution methods dominate many protein engineering applications, e.g. enhancing enzyme activity [1], de novo antibody discovery [2-4], improving antibody affinity [5-7] and improving antibody properties [8-9]. Starting from an ensemble or library of proteins with carefully randomized regions, unique clones with desirable phenotypes including superior affinity/activity can be identified through multiple cycles of selection and screening assays. Successful screening relies on the pre-existence of the superior clones in the screened library. To accommodate the vast spectrum of different target sequences and structures, the library must be adequately diversified to guarantee a high success rate. The absolute number

[†]Correspondence: Qi Zhao, Ph.D., Abcam Branford CT., 688 E. Main St., Branford, CT 06405, Office: (203)208-1918, qi.zhao@abcam.com.

Author Contributions.

Q.Z. and M.P.W. conceived and developed the idea of resampling functional CDR ensembles; Q.Z. wrote the bioinformatic pipeline. Q.Z., M.P.W, M.B. and M.K. designed experiments, D.L.B. and E.H. performed library construction and cloning. J.F. performed discovery screening. C.G. and E.S performed protein purifications and ELISA validation experiments; C.R. performed Western Blot experiment. Q.Z, M.P.W., M.K. and M.B wrote and edited the paper,

The authors declared no conflict of interests.

of unique genotypes and the ratio of all genotypes that result in a functional phenotype determine the functional diversity of the library.

To improve library quality, the most straightforward method would be to increase the absolute number of unique sequences. For antibody discovery, in particular, different display methods predetermine the achievable absolute diversity [10,11]. Since one *E. coli* electroporation typically yields 10^9 CFU, by using a thousand transformations, libraries of up to 10^{12} single-chain fragment variable regions (scFv) of IgG can be routinely generated. Similarly, limited by the ribosome concentration used in cell-free systems, up to 10^{15} ribosomes [12] linking unique proteins with their coding RNA can be generated for one ribosome display library. Above these technical limitations, construction costs rise steeply for a marginal gain in diversity.

Besides increasing the overall number of unique clones, various lines of effort have been applied to enhance the percentage of functional clones. One common practice is to limit randomization regions to selected regions out of 6 available CDRs, i.e. CDR3 regions only [13]. Another example is the use of degenerate nucleotides (i.e. “NMY” or “NNK”), excluding stop codons, at naturally diversified amino acid sites within selected complementarity determining regions (CDRs) [14]. More advanced methods using defined CDR sequences [15] or triplet codon synthesis to bias towards non-rare codons, as well as to imitate the natural distribution of amino acids at multiple sites of the CDR, have also been used [16,17]. Moreover, randomizing the combination of naïve light chain and heavy chain variable domains has been tested ti diversify a library at a level higher than primary sequences [18]. Given the extensive effort on codon optimization, two limitations still exist. Firstly, the *a priori* knowledge of natural amino acid distribution stems from the antibody database with diversified variable domain frameworks. This inevitably raises concerns of compatibility with the scFv framework. Secondly, amino acid distribution omits the contextual information of sequence, i.e. sequence correlations between different sites. Certain amino acids at a particular site might preclude another amino acid from incorporating functionally at another site. Therefore, sequences agreeing with a given amino acid distribution are very likely to include self-conflicting amino acid pairs.

To resolve both sources of incompatibility, which would potentially reduce the functional clone percentage, here we propose an experimental-data based computational method to generate library diversity. In essence, we started from a legacy database of experimentally validated antibodies with almost identical framework sequences. Each sequence was computationally annotated and a new database constructed recording unique CDR amino acid sequences. These were back translated and barcoded into a database of DNA sequences using python software. Using combinatorial assembly of these DNA sequences with the identical framework sequence from which the CDRs were extracted, we constructed a library (sequence diversity $>10^{10}$) with predefined CDR sequences (“PDC” library). Furthermore, we demonstrated that this resultant PDC library yielded a significantly increased success rate against different antigen types, and also more than 20-fold unique hits per targets compared with libraries generated by traditional degenerate nucleotide methods. Our studies indicate that predefined CDR sequences could be a productive direction of functional library construction for *in vitro* antibody development.

Materials and Methods

List of materials

All antigens were human targets. Biotinylated peptides antigens were designed in-house and purchased from Biopeptik with 90% purity (peptides: XIAP1B (Biotin-SGSPVSASTLARAGFLYTGE, Human), XIAP2B (Biotin-THADYLLRTGQVVDISDTIY, Human), GRAP2B (Biotin-SLNKLVDYYRTNSISRQKQI, Human), GRAP3B (Biotin-TDPVQLQAAGRVRWARALYD, Human), LMNA3B (Biotin-DEYQELLDIKLALDMEIHAYRK, Human), TDP_42_NLS (Biotin- PKD NKRKMDETDASSAVKVKRA), XIAP3B (Biotin-AEAVDKCPMCYTVITFKQK, Human), LMNA2B (Biotin-RIDSLSAQLSQLQKQLAAKE, Human) purified proteins were provided internally by Abcam (proteins: hIL-12p70(P29459/P29460, Human), FGF basic protein (P09038, Human), Dtk-Fc(Q06418, Human), BCMA-Fc(A7KBT3, Human), Nogo Fc(Q9NQC3, Human), IL-13(P35225, Human), myelin basic protein(P02686, Human), hLeptin (P41159), IFN alpha2(P01563, Human), BDNF(P23560,Human), IL-1 beta(P01584, Human), IL3 beta(P08700, Human). E. coli maltose binding protein (MBP) was produced in house. Turbo Competent E. coli (High Efficiency) cells (C2984I) and SOC media (B9020S) were purchased from New England Biolabs. B-Per Bacterial Protein Extraction Reagent (78248) was purchased from Pierce. Nonfat dry milk powder (M7409) to make 5% MPBS were purchased from Sigma-Aldridge. Goat pAb anti-myc HRP Antibody (ab1261, GR301729-5), mouse mAb anti-FLAG HRP Antibody (ab49763, lot# GR3187436-1), and goat pAb anti-FLAG HRP Antibody A8592 lot # 087k6011) were purchased from Sigma. Efficacy of all antibodies were closely monitored with positive controls (Nter-FLAG-tagged or Cter- Myc-tagged ScFv proteins produced in house) implemented alongside the daily site operation. BioRad Clarity Western ECL Substrate (170-5061) was purchased from BioRad. Precision Plus Protein Standards (161-0373) were purchased from BioRad.1-Step Ultra TMB (34028) was purchased from Thermo Scientific. HisPur Cobalt Resin (89966) was purchased from Pierce. 96-well microplates (222-8030-01F) were purchased from Evergreen. BioRad precast gels, (10-well (4561093), 15-well (4561096) or 26-well (5671095)) were purchased from BioRad. 14 mL Polypropylene conical tube (SS_PH15R) and 50 mL polypropylene conical tubes(SS_PH15R) were purchased from Phenix. Tape pads (19570) were purchased from Qiagen.

DNA array synthesis and storage

Based on the sequences generated from computational pipeline (Figure2), 4000 purified DNA oligos were synthesized by Integrated DNA Technologies Inc.; each oligo was PEG purified to over 95% purity as 10 pmol lyophilized powder pellets. All oligos were dissolved in 20uL double distilled water and stored in sealed 96 well plates in −20 °C freezer prior to use.

Phage library screening

The phage display discovery screening was carried out following previously published protocols [16]. Three rounds of enrichment were performed. 200 individual clones per library per target were analyzed by ELISA against the target (+) and target (−) plate. Clones

with an ELISA signal and >10 fold over background (FOB) for automated phage display screening method were sequenced by an in-house bioinformatics pipeline.

Protein expression and purification

A single colony was added to each tube (14 ml polypropylene conical tubes) and cultures grown overnight at 37 °C in 2YT media supplemented with ampicillin and glucose. The overnight culture was added to each of 3 2L flasks containing 400 mL of Superior Broth supplemented with ampicillin and glucose and cultures were grown at 37 °C with shaking at 250 rpm until an OD600 of 0.9 was reached. A 1:1000 dilution of 1M IPTG was added directly into the flask to induce protein expression. Cultures were grown at 30°C overnight for a total of 18 h with shaking at 250 rpm. Cells were pelleted by centrifugation, re-suspended in lysis buffer and vortexed. Pellets were re-suspended on a platform shaker at room temperature (RT) for 25 min. The cell lysates were transferred to high-speed centrifuge tubes and spun at 4 °C in a high-speed floor Beckman Avanti J-26S centrifuge (rotor SA-600) at 15,000×g. The lysate supernatant was added to an equilibrated HisPur Cobalt Resin column. After the lysates passed through by gravity flow, columns were washed with 10 mL of wash buffer (PBS supplemented with 10 mM imidazole). Each column was eluted with 1.5 mL of elution buffer (PBS supplemented with 400mM imidazole). Eluted proteins were analyzed by Coomassie Plus assay and SDS-PAGE. Samples were stored at −80 °C.

Enzyme linked Immunosorbent Assay (ELISA)

For titration ELISA, Maxisorp 96-well plates were coated with 100 μL/well of NeutrAvidin at a final concentration of 10 μg/mL. The NeutrAvidin-coated plates were washed three times with PBS and blocked with 3% BSA/PBS for 1 h. The plates were incubated at RT for 1 h and then washed 3× with PBS. Seven titrations (1 μg, 0.5μg, 0.25μg, 0.125μg, 0.0625μg, 0.03125μg, 0.015μg) of biotinylated antigens were applied to the plates and incubated for 1h; protein antigens were directly coated to the NeutrAvidin-free plate for 1h. The plates were incubated at RT for 1 h and then washed 3× with PBS. To remove unspecific binding to hydrophobic antigens, the plates were further blocked and incubated at RT for 1 h. The biotinylated-antigen-coated plates were washed 3×, and coated with 100 μL of scFv, (at 1 μg/mL) to each well. The plates were incubated at RT for 1 h and washed 4× with PBS/0.1% Tween 20. Goat pAb anti-myc HRP was diluted 1:5000 in 3% BSA/PBS per well and added to all ELISA plates. The plates were washed 3× with 250 μL/well of PBST. Ultra TMB reagent was added and developed for 2 m at RT and the reaction was stopped by adding 50 μL per well of 2 M H2SO4. ELISA signal (absorbance at 450 nm) was measured by BioTek plate reader.

Western Blot

50 ng of antigen was loaded onto a precast SDS-PAGE gel. To test for MBP (E. coli maltose binding protein, home-made) background binding, 50 ng of MBP was loaded into one well per Western blot. The proteins in the gel were transferred to a nitrocellulose membrane, (BioRad mini or midi format, depending on the size of the gel) using the BioRad Trans-Blot Turbo System as per manufacturer's protocol. When the transfer was complete, the membranes were cut between the molecular weight standard lanes, and placed in a tray

containing 5% MPBST (PBS supplemented with 5% w/v skim milk powder and 0.1% Tween), to fully cover the membrane. The membranes were blocked for 1 h at RT on a platform shaker. Affinity matured scFvs were added at a concentration of 1μg/mL in 3 mL 5% MPBST. The membranes were incubated for 1 h at RT on a platform shaker and then washed 3 times each for 10 m with PBST. Secondary antibody (anti-FLAG-HRP) was added, at a 1:5000 dilution in 5% MPBST. Membranes were incubated for 1 h at RT on a platform shaker, followed by 3×15 m washing with PBST. The blots were developed with ECL Reagent.

Results and Discussion

Concept and Antibody database clean-up

Antibody-antigen affinity is conferred by the collaboration of six CDRs as structured loop regions between the β-strands that form the Immunoglobin (Ig) fold [19-21] of the variable (V) domain (Figure1a). Further investigation into a typical antibody structure reveals the minimal molecular interactions between CDRs L2, L3, H2 and H3 (Figure1a). The apparent independence or modularity of these four CDRs supports the likelihood that switching different CDR sequences at one CDR does not conflict with the existing sequences at other CDRs. Therefore, a new library made from reshuffling CDR sequences (Figure1b) of known functional antibodies at different CDR positions will be likely to achieve the maximum foldability. Hence the significant part of the library should be both soluble and functional. Based on this hypothesis, such a library was designed, constructed and validated. To achieve maximum compatibility, we utilized a legacy database consisting of ~2000 sequences of scFv domains, discovered in a single-framework library generated via a traditional degenerate method, where “NNK” codons were substituted at ~15 different positions on a single scFv template; the parental scFv was developed by linking the VL and VH domains of a human anti-IFNα antibody with a 12 amino acid glycine linker [16].

Following a computational workflow (Figure 2), we successfully reduced the raw antibody database to a set of DNA sequences uniquely encoding the CDR sequences. All 2000 sequences were expressed, purified and validated to bind specifically to their respective targets. The initial sequence data were stored as manually assembled csv data files with sporadically missing data blocks, duplicate inputs and framework sequences from other sources. A database clean-up module was implemented to extract database entries with unique full amino acid sequences. The sequences were homologous to the original framework (at least 85% identity in sequence of the framework region), where sequence homology was calculated by the pairwise alignment method against a reference sequence in Biopython [22,23].

Variable domain extraction, sequence annotation and CDR ensemble generation

After initial database clean-up, 1500 unique scFv sequences were stored as a dictionary of dictionaries, i.e.

{‘sequence_id1’: {‘FL’: ‘… . ’; ‘Heavy’: ‘…’ ‘Light’: ‘…’}; ‘sequence_id2’: {‘FL’: ‘… . ’; (1)
‘Heavy’: ‘…’ ‘Light’: ‘…’}; …}

where sequence ID and relevant amino acid sequences were specified as the key/value pair, and the amino acid sequences of scFv were stored as the first item. Heavy chain and light chain V regions were extracted by aligning to the corresponding chains from a reference sequence and stored into the same value corresponding to the sequence ID. To achieve extendability and compatibility with all possible frameworks, after splitting the scFv sequence into heavy and light chains, the Chothia numbering scheme [24] was used to annotate each sequence. More specifically, a set of regular expressions was used to scan the amino acid sequences to determine the key Chothia positions of all CDR residues. Other Chothia numbering [24] ids were assigned sequentially to fill the gaps between CDRs. The data structure of this annotation step consists of two extra key/value pairs per unique ScFv sequence, or

$$\text{‘L\_chothia’}:\{\text{‘1’}:\text{‘…’};\text{‘2’}:\text{‘…’};\ldots\};\ \text{‘H\_chothia’}:\{\text{‘1’}:\text{‘…’};\text{‘2’}:\text{‘…’};\ldots\}; \quad (2)$$

which were further updated into the original dictionary (1).

Based on the fully annotated protein sequence, the specific CDR sequences were extracted following the Chothia scheme: CDRL1: L24-L34; CDRL2: L50-L56; CDRL3: L89-L97; CDRH1: H26-H32/34; CDRH2: H52-H56; CDRH3: H95-H102. The numbering scheme was supported by three-dimensional alignment among the experimentally determined antibody structures. To generate the CDR sequence ensemble, CDR amino acid sequences from scFv entries were extracted from (2) and pooled as six new dictionaries (3) for all six CDRs, e.g. L1 data would appear as:

$$\{\text{‘id1\_L1’}:\text{‘…’};\ \text{‘id2\_L1’}:\text{‘…’};\ldots\} \quad (3)$$

Starting from the six dictionaries containing all parsed CDR sequences, amino-acid level sequence filtering was performed to remove undesirable entries, which included but were not limited to redundant sequences, cysteines [25], suppressed stop codons, patented sequences, glycosylation motifs, phosphorylation motifs, etc. After data-reduction, approximately 1000 entries survived out of the initial 2000 raw sequence inputs.

Reverse translation: oligo DNA barcoding and filtering

To encode the CDR sequences ensemble using ~ 4000 synthesized DNA oligos, CDR amino acid sequences need to be back-translated in silico, i.e. from amino acid sequence to DNA sequences (Figure2). A customized codon table was generated to focus on preferred codons in both E. coli K2 and human, aiming to optimize the folding and expression quality of proteins in both bacteriophage form and human IgGs.

For downstream production and analysis, we proposed to implement multiple considerations into an integrated back-translation process. First, we aimed to eliminate common type II restriction enzyme sites to accommodate downstream cloning processes. Secondly, we aimed to employ NGS sequencing to over-sample the CDR regions for high-throughput

analysis, preferably by the Illumina HiSeq which can generate 10^9 reads. Therefore, the DNA sequences were uniquely barcoded.

Rolling Circle Amplification (RCA)-based library generation and quality control

DNA sequences encoding CDRs obtained from the steps outlined above were assembled with corresponding flanking sequences and synthesized as purified oligo arrays (10pmol each oligo) (Integrated DNA Technologies, Inc). These oligos were mixed in equimolar ratios as mutagenic oligos for library construction (Figure3). During construction, Eco29KI cleavage sites were encoded in the CDRs targeted for mutagenesis, after which Kunkel-based site directed mutagenesis was used to substitute the Eco29KI sites with the synthesized mutagenic oligos and RCA to amplify the recombined library DNA. The Eco29KI site was used to remove the parental strand as well as partially recombined clones both *in vitro* and *in vivo*. RCA amplified DNA was electroporated into TG1 competent cells and the library of scFvs were displayed on the surface of bacteriophage M13 as a genetic fusion to the gpIII coat protein. Library diversity was calculated by counting serially diluted CFUs directly after library transformation. For comparison with the effect of absolute diversity per CDR, two sets of oligos were used to create two libraries individually: a small library (L2: 300, L3: 100, H2:100, and H3:100) and large library (L2:300, L3:1000, H2:1000 and H3:1000). The experimental diversity was ~2×10^{10} for each library, while the theoretical DNA diversities were 3×10^8 and 3×10^{11} for small and large libraries respectively. Thus the large library was under-sampled and the small library was over-sampled.

Colonies of each library were picked and colony PCR reactions were performed to extract the fragment of each scFv for sequencing. A dataset of DNA sequences of ~800 nt in length covering both the light and heavy chain V domains was collected and analysed by a customized python/R module for both quality control and downstream hits analysis (Figure 4). Frequencies of each CDR of the sequence set were calculated to provide CDR-specific enrichment. Moreover, the CDR quartet of each individual sequence was also identified to evaluate sequences for further validation. Sequencing errors, including single nucleotide deletion/insertions were tolerated during CDR identification for both general statistics and individual sequence analysis. The CDR frequencies showed that each CDR was evenly distributed among the QC sequences of each library. This indicated that both libraries have unbiased CDR incorporation.

Bio panning on antigen test-set, hits validation and analysis

De novo antibody discovery against a set of unknown targets is a critical standard for assessing the quality of regenerated diversities. Therefore, we picked a set of 12 folded proteins and 8 synthetic peptides (proteins: hIL-12p70, FGF basic protein, Dtk-Fc, BCMA-Fc, Nogo Fc, IL-13, myelin basic protein, hLeptin, IFN-α, BDNF, IL-1 beta, IL3-β;,peptides: XIAP1B, XIAP2B, GRAP2B, GRAP3B, LMNA3B, TDP_42_NLS, XIAP3B, LMNA2B) as the initial set to screen with the two PDC generated libraries. The phage display discovery screening was carried out following previously published protocols [16]. Hit sequences were analyzed by a bioinformatics pipeline written in Python/R, where each hit sequence can be uniquely identified by four integers as the entry ID stored in our CDR sequence database (Figure 5). Four clones with enriched CDRs were expressed in *E. coli* and

purified by metal affinity chromatography. Purified proteins were further applied to ELISA and western blot for validation (Figures 5, 7). Success rates were calculated by number of targets yielding validated hits divided by the total number of targets. Indeed, both libraries yielded remarkable screening success (>90% for all targets). More importantly, for various metrics including protein expression level, solubility and pure protein ELISA-positive rate, the two libraries were essentially the same (Figure5). The large library yielded more consensus CDRs than the small library. Thus, PDC libraries are excellent in target spectrum and antibody yields/solubility, and are highly active in ELISA and Western Blot. Moreover, the majority of unique hits generated from the screen on the same antigen clustered into subgroups bearing the common CDR, indicating shared binding regions. The enrichment of CDR sequences rather than full V domain sequences, also supports the assumed independence and modularity of CDR regions of a typical antibody.

CDR enrichment analysis identifies CDRs correlating with nonspecific binding

In addition to the abundant unique hits for the test set of folded proteins and synthesized peptides (Figure 6), data mining of nonspecific hit sequences provided a possible route to optimization of the initial library. Based on the sequence data from 5 independent screens against different antigens, similar phage supernatant ELISAs were performed on 88 monoclonal hits per screen. Nonspecific hits (OD450nm >1 for both negative control plate and target coated plated) were pooled and sequenced. Raw DNA sequences were filtered and processed through the PDC library validation pipeline. A digital ID per CDR sequence was assigned. One typical clone was unexpectedly enriched across different screens (L2: 48, L3: 747, H2: 229 and H3: 591) (Figure 6), indicating that this combination of CDRs might yield nonspecific affinity as well as good growth advantage. Moreover, the H2: 229 was further enriched in many nonspecific clones combined with different CDRs, suggesting that H2: 229 itself confers undesirable affinity to the base material of the negative plate (Figure6). Therefore, re-constructing the PDC library omitting the specific H2 CDR or substituting it with a new CDR sequence was likely to reduce the number of nonspecific binders. This should reduce the cost of hit validation. Iteratively updating the CDR pool by removing CDRs discovered in nonspecific hits and introducing new CDR sequences would be an effective way to maintain and evolve existing libraries built following predefined CDR strategy.

Conclusion

Sufficient sequence diversity, especially within the CDR DNA sequences, is the premise of successful screening. Although quite under-sampled from the theoretical diversity, actual diversity is still considerably larger than the antigen diversity. Therefore, the requirement of antibody diversity might be overestimated. We could further determine a minimal diversity of a functional library using a similar process to that described in this study. Starting from the small library, we could determine a minimal diversity set library via a binary search algorithm that has been widely employed for many optimization problems in engineering. Essentially, a half size library (~150 L2. 50 L3, 50 H2, and 50 H3 sequences) can first be constructed and validated for success rate, If the success rate stays ~90%, we will further reduce the CDR pool size in half (75 L2, 25 L3,25 H2, and 25 H3), and validate again. If the

success rate is significantly lower than the initial success rate of 90%, the CDR set will be increased to 75% of the small set library (120 L2, 75 L3, 75 H2, and 75 H3). Through iterations, a CDR set size will be arrived at that keeps library success rate ~90%, but where the left and right intervals i.e. CDR pool size in previous generation and the generation before that, could converge to less than ~10 CDRs, which is about three generations of library making, from an estimation that $100*(0.5)^3 = 12.5$. Estimation from binary search theory indicates that the minimal size of CDR set will be determined within a reasonable cost of library construction and validation.

Acknowledgments

We thank Dong Wang, Gregory Mirando, and Dawn Alderman for insightful discussions. The work is supported by NIH R44GM105080.

References

1. da Silva MC, Del Sarto RP, Lucena WA, Rigden DJ, Teixeira FR, Bezerra Cde A, et al. Employing in vitro directed molecular evolution for the selection of alpha-amylase variant inhibitors with activity toward cotton boll weevil enzyme. J Biotechnol. 2013; 167:377–385. [PubMed: 23892157]
2. Boder ET, Midelfort KS, Wittrup KD. Directed evolution of antibody fragments with monovalent femtomolar antigen-binding affinity. Proc Natl Acad Sci USA. 2000; 97:10701–10705. [PubMed: 10984501]
3. Riano-Umbarila L, Juarez-Gonzalez VR, Olamendi-Portugal T, Ortiz-Leon M, Possani LD, Becerril B. A strategy for the generation of specific human antibodies by directed evolution and phage display, An example of a single-chain antibody fragment that neutralizes a major component of scorpion venom. FEBS J. 2005; 272:2591–2601. [PubMed: 15885107]
4. Xu L, Aha P, Gu K, Kuimelis RG, Kurz M, Lam T, et al. Directed evolution of high-affinity antibody mimics using mRNA display. Chem Biol. 2002; 9:933–942. [PubMed: 12204693]
5. Juarez-Gonzalez VR, Riano-Umbarila L, Quintero-Hernandez V, Olamendi-Portugal T, Ortiz-Leon M, Ortiz E, et al. Directed evolution, phage display and combination of evolved mutants: a strategy to recover the neutralization properties of the scFv version of BCF2 a neutralizing monoclonal antibody specific to scorpion toxin Cn2. J Mol Biol. 2005; 346:1287–1297. [PubMed: 15713481]
6. Gupta A, Chaudhary VK, Bhat R. Directed evolution of an anti-human red blood cell antibody. MAbs. 2009; 1:268–280. [PubMed: 20069755]
7. Fujii I. Directed evolution of antibody molecules in phage-displayed combinatorial libraries. Yakugaku Zasshi. 2007; 127:91–99. [PubMed: 17202788]
8. Shusta EV, Holler PD, Kieke MC, Kranz DM, Wittrup KD. Directed evolution of a stable scaffold for T-cell receptor engineering. Nat Biotechnol. 2000; 18:754–759. [PubMed: 10888844]
9. Famm K, Hansen L, Christ D, Winter G. Thermodynamically stable aggregation-resistant antibody domains through directed evolution. J Mol Biol. 2008; 376:926–931. [PubMed: 18199455]
10. Yang HY, Kang KJ, Chung JE, Shim H. Construction of a large synthetic human scFv library with six diversified CDRs and high functional diversity. Mol Cells. 2009; 27:225–235. [PubMed: 19277506]
11. Bai X, Shim H. Construction of a scFv Library with Synthetic, Non-combinatorial CDR Diversity. Methods Mol Biol. 2017; 1575:15–29. [PubMed: 28255872]
12. Kim JM, Shin HJ, Kim K, Lee MS. A pseudoknot improves selection efficiency in ribosome display. Mol Biotechnol. 2007; 36:32–37. [PubMed: 17827535]
13. Pini A, Viti F, Santucci A, Carnemolla B, Zardi L, Neri P, et al. 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. 1998; 273:21769–21776. [PubMed: 9705314]

14. Sidhu SS, Li B, Chen Y, Fellouse FA, Eigenbrot C, Fuh G. Phage-displayed antibody libraries of synthetic heavy chain complementarity determining regions. J Mol Biol. 2004; 338:299–310. [PubMed: 15066433]
15. Rothberg JWM. USPTO. , editorMethods of generating antibody diversity in vitro (US20060160178). US Patent Publication. 2006.
16. Batonick M, Holland EG, Busygina V, Alderman D, Kay BK, Weiner MP, et al. Platform for high-throughput antibody selection using synthetically-designed antibody libraries. N Biotechnol. 2016; 33:565–573. [PubMed: 26607994]
17. Zhai W, Glanville J, Fuhrmann M, Mei L, Ni I, Sundar PD, et al. Synthetic antibodies designed on natural sequence landscapes. J Mol Biol. 2011; 412:55–71. [PubMed: 21787786]
18. Sblattero D, Bradbury A. Exploiting recombination in single bacteria to make large phage antibody libraries. Nat Biotechnol. 2000; 18:75–80. [PubMed: 10625396]
19. Carmichael JA, Power BE, Garrett TP, Yazaki PJ, Shively JE, Raubischek AA, et al. The crystal structure of an anti-CEA scFv diabody assembled from T84.66 scFvs in V(L)-to-V(H) orientation: implications for diabody flexibility. J Mol Biol. 2003; 326:341–351. [PubMed: 12559905]
20. Park SY, Lee WR, Lee SC, Kwon MH, Kim YS, Kim JS. Crystal structure of single-domain VL of an anti-DNA binding antibody 3D8 scFv and its active site revealed by complex structures of a small molecule and metals. Proteins. 2008; 71:2091–2096. [PubMed: 18338383]
21. Clark KR, Walsh ST. Crystal structure of a 3B3 variant–a broadly neutralizing HIV-1 scFv antibody. Protein Sci. 2009; 18:2429–2441. [PubMed: 19785005]
22. Cock PJ, Antao T, Chang JT, Chapman BA, Cox CJ, Dalke A, et al. Biopython: freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics. 2009; 25:1422–1423. [PubMed: 19304878]
23. Talevich E, Invergo BM, Cock PJ, Chapman BA. Bio.Phylo: a unified toolkit for processing, analyzing and visualizing phylogenetic trees in Biopython. BMC Bioinformatics. 2012; 13:209. [PubMed: 22909249]
24. Chothia C, Lesk AM. Canonical structures for the hypervariable regions of immunoglobulins. J Mol Biol. 1987; 196:901–917. [PubMed: 3681981]
25. Brych SR, Gokarn YR, Hultgen H, Stevenson RJ, Rajan R, Matsumura M. Characterization of antibody aggregation: role of buried, unpaired cysteines in particle formation. J Pharm Sci. 2010; 99:764–781. [PubMed: 19691118]
26. Marcatili P, Rosi A, Tramontano A. PIGS: automatic prediction of antibody structures. Bioinformatics. 2008; 24:1953–1954. [PubMed: 18641403]

Highlights

- A synthetic antibody fragment library resampling the functional CDRs was constructed.
- The resultant library yielded *de novo* specific hits against a wide spectrum of targets.

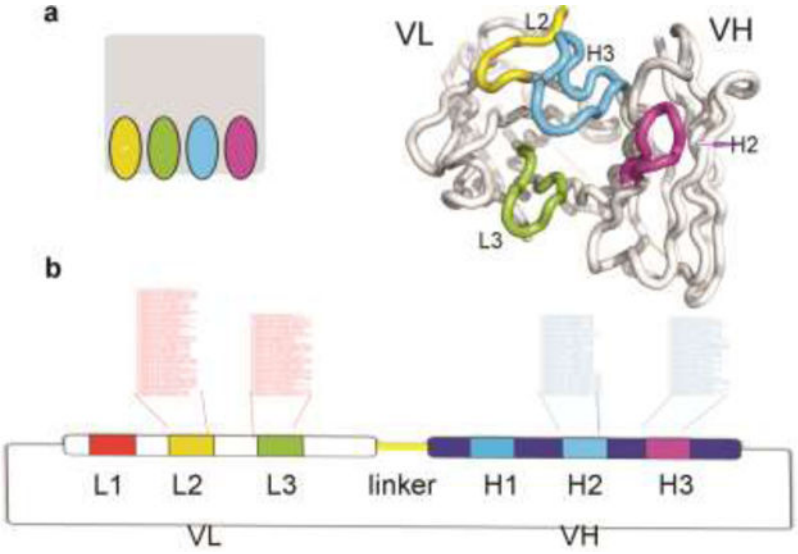

Figure 1. Schematic view of the PDC library strategy

In a library constructed from tens of unique CDR sequences of the L2, L3, H2 and H3 regions (four different ovals) of antibody V domains (gray box). The structural model was built based on the parental scFv sequence of PDC library using homological model building tool *PIGS* [26]. **b**,) CDR sets and the DNA arrangement of the small-set library and large set libraries. The scale of CDR sets for two different libraries are labelled above. CDRs L1, L2, and L3 are in red and CDRs H1, H2 and H3 in cyan. The Light chain framework is shown in white, and the heavy chain framework in dark blue.

Figure 2. Bioinformatic pipeline of DNA array design from the database of validated antibodies

Macro steps including database integration, sequencing filtering, antibody alignment as well as DNA sequence generation are shown in blue; routine steps that do not require extra input are highlighted in brown; sub-steps that can be customized to adapt to different purposes are in red. The entire bioinformatic pipeline was realized via Python/BioPython module.

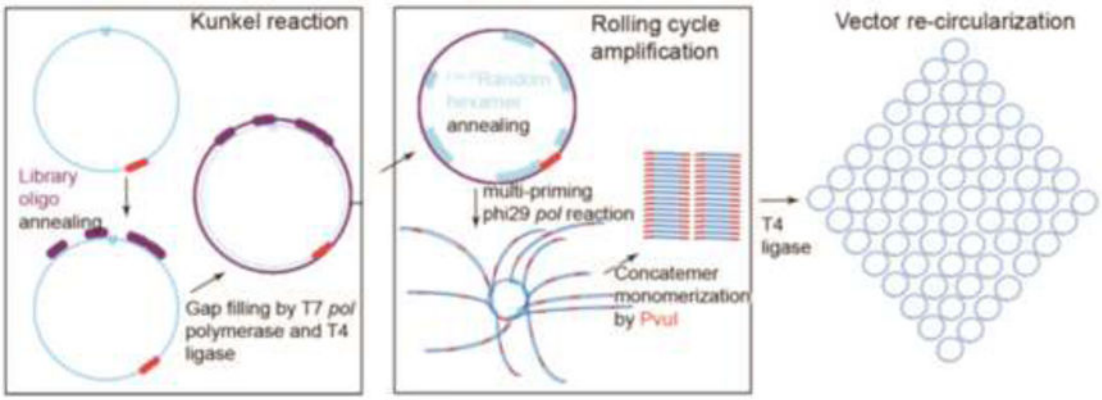

Figure 3. Design, construction and validation of defined CDR libraries

Left panel: CDR sets and the DNA arrangement of small-set and large set libraries. The scale of CDR sets for two different libraries are labelled above. CDRs L1, L2, and L3 are in red and H1, H2 and H3 in cyan. Light chain framework is white, and heavy chain framework is dark blue. **Right panel:** the libraries are generated by Kunkel mutagenesis, where oligos are first annealed to ssDNA, and heteroduplex DNA is generated by filling in the gaps between different mutagenic oligos by T7 DNA polymerase and T4 ligase; the fully recombinant HD DNA will then be selectively amplified ~100-fold using Rolling Circle Amplification (RCA), linearization and recircularization.

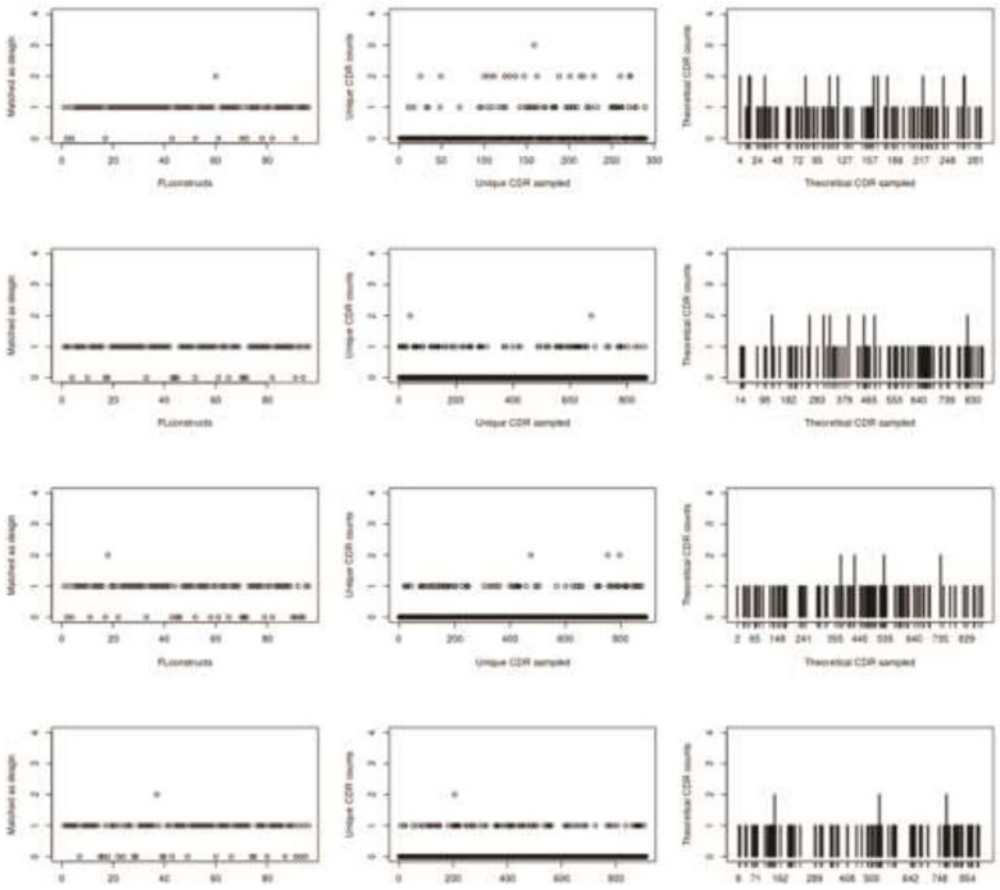

Figure 4. Typical quality control data of PDC library by sequencing

Library quality control analysis for four CDRs **(top row to bottom row**: L2, L3, H2, H3); the left panel of each row shows whether the CDR sequence in the scFv was found in the predefined CDR database, where over 80% of CDRs can be identified as designed (y: CDR frequencies, x: all 96 scFv sequences by Sanger method). The middle panel is the experimental frequencies of any CDR observed in all 96 scFv sequences (y: CDR frequencies, x: all possible CDR sequences in the predefined database.), where an unbiased distribution of CDR frequencies found in 96 scFv sequences (represented by each circle points) should be comparable to the theoretically calculated distribution (represented by the bar Hight) shown in the right panel.

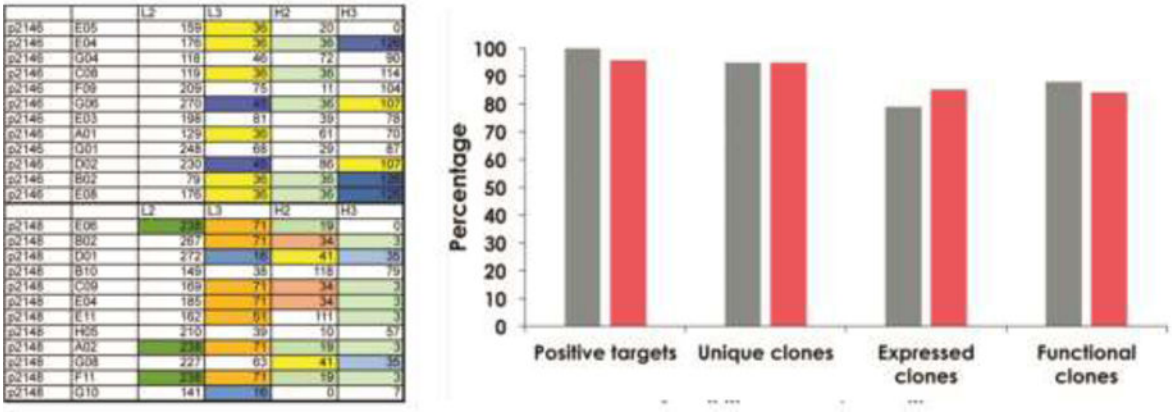

Figure 5. Typical dissection of CDRs for hits discovered from two PDC libraries, and the performance data of these libraries at different stages

Left panel: phage hits CDR dissection for the large library (p2146) and small library (p2148), both screens targeted on the same antigen (XIAP3B). Each sequence is compared with the predefined CDR database; an identifiable CDR ID is recorded as an integer; the same CDR IDs from the same screen but shared among different hits are highlighted in the same colour. **Right panel**, the percentage success rate across four key steps of antibody discovery are shown. Parallel comparison between small library (red bar) and large library (grey bar) are included. Zeros mean failure to match any designed sequences, probably due to sequencing error or mutations introduced in the library construction process.

Screen id	Hits id	L2	L3	H2	H3
2086	A10	178	434	229	54
2086	A8	48	747	229	591
2086	B3	76		229	
2086	A2	232	344	229	125
2086	A1	241	292	171	281
2086	B5	116	85	807	265
2086	B2	209	764	229	570
2086	A9	110	61	287	771
2087	A3	48	747	229	591
2087	A8	48	747	229	591
2087	A9	112	88	229	421
2087	A4	232	814	452	430
2087	A6	277	386	229	66
2087	A7	172	591	229	421
2087	A5	26	92	229	642
2087	A2	185	68	284	285
2088	A10			824	862
2088	A9	173	52	809	296
2088	C3	262	100	80	94
2088	B9	230	643	148	94
2088	B8				
2088	B1	199	671	229	375
2088	B4	223	702	450	466
2089	A6	167	771	229	421
2089	B10	26	651	658	583
2089	B4	203	344	229	463
2089	A9	199	671	229	375
2089	A1	113	2	229	667
2089	B11	235	723	229	663
2089	B1_	120	81	658	594
2092	D5		213	314	
2092	D2	236	581	226	321
2092	A1	256	836	446	890
2092	E1				
2092	B5	234	819	229	520
2092	B8				
2092	B11	178	389	145	131

Figure 6. Identification of CDRs for nonspecific binders

Based on the sequence data from five independent screens against different antigens, similar phage supernatant ELISAs were performed on 88 monoclonal hits per screen. Nonspecific hits (OD450nm >1 for both negative control plate and target coated plated) were pooled and sequenced. Raw DNA sequences were filtered and processed through the PDC library validation pipeline. A digital id per CDR sequence was assigned; one typical clone was unexpectedly enriched across different screens (L2:48, L3: 747, H2: 229 and H3: 591), indicating this combination of CDRs might yield nonspecific affinity as well as good growth advantage. Moreover, the H2:229 was further enriched in the majority of nonspecific clones combined with other CDRs, suggesting H2:229 itself conferring undesirable affinity the base material of the negative plate.

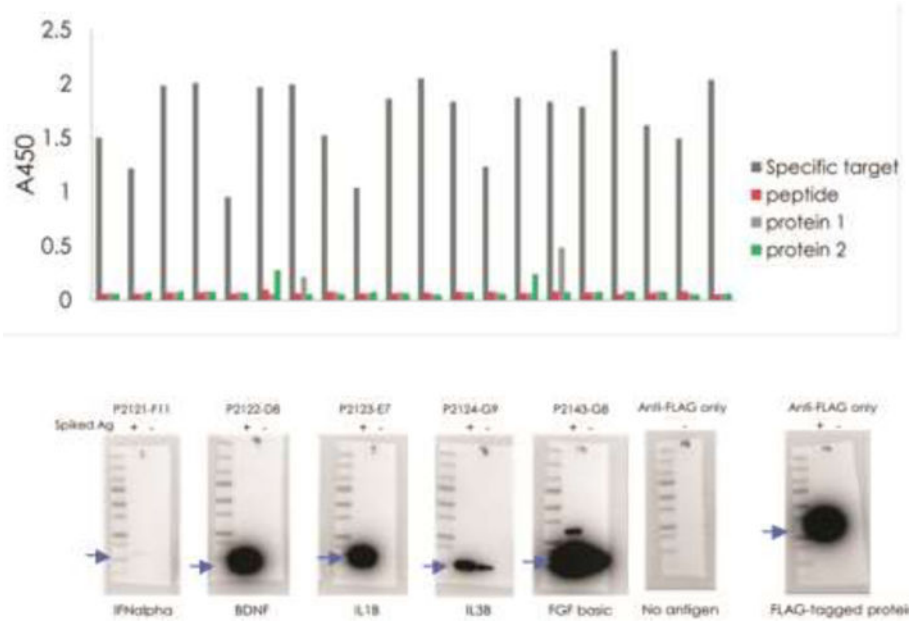

Figure 7. ELISA and western blot validation of discovery screen hits

Top panel, validating ELISA results for scFv: 20 typical scFv proteins were checked against their cognate target (dark grey bar), as well as three other non-target antigens (read, light grey and green); affinity is quantitated as OD absorption at 450nm. **Bottom panel**, scFv hits that were developed using full-length proteins were checked against human cell lysate (HEK293FS) spiked with its specific target protein (lane +) and lysate only (lane −) to validate their affinity as well as specificity in Western blot conditions. For ELISA, the specific targets are (peptides: XIAP1B, XIAP2B, GRAP2B, GRAP3B, LMNA3B, TDP_42_NLS, XIAP3B, LMNA2B, proteins: hIL-12p70, FGF basic protein, Dtk-Fc, BCMA-Fc, Nogo Fc, IL-13, myelin basic protein, hLeptin, IFN alpha, BDNF, IL-1 beta, IL3 beta) For western blot-targets, the specific targets are labelled respectively.