Computational and Structural Biotechnology Journal 20 (2022) 2723–2727

COMPUTATIONAL AND STRUCTURAL BIOTECHNOLOGY JOURNAL

journal homepage: www.elsevier.com/locate/csbj

Reconstruction of full antibody sequences in NGS datasets and accurate V_L:V_H coupling by cluster coordinate matching of non-overlapping reads

Jorge Moura-Sampaio [a,b], André F. Faustino [a], Remi Boeuf [c], Miguel A. Antunes [a], Stefan Ewert [c], Ana P. Batista [a,*]

[a] iBET, Instituto de Biologia Experimental e Tecnológica, Apartado 12, 2781-901 Oeiras, Portugal
[b] Instituto de Tecnologia Química e Biológica António Xavier, Universidade Nova de Lisboa, Av. da República, 2780-157 Oeiras, Portugal
[c] Novartis Institutes for BioMedical Research, Basel, Switzerland

ARTICLE INFO

Article history:
Received 30 March 2022
Received in revised form 27 May 2022
Accepted 27 May 2022
Available online 31 May 2022

Keywords:
Next-generation Sequencing
Phage-display
Synthetic Libraries
Randomization
Diversity
CDR
Fab

ABSTRACT

Next-generation sequencing (NGS) is an indispensable tool in antibody discovery projects. However, the limits on NGS read length make it difficult to reconstruct full antibody sequences from the sequencing runs, especially if the six CDRs are randomized. To overcome that, we took advantage of Illumina's cluster mapping capabilities to pair non-overlapping reads and reconstruct full Fab sequences with accurate V_L:V_H pairings. The method relies on *in silico* cluster coordinate information, and not on extensive *in vitro* manipulation, making the protocol easily deployable and less prone to PCR-derived errors. This work maintains the throughput necessary for antibody discovery campaigns, and a high degree of fidelity, which potentiates not only phage-display and synthetic library-based discovery methods, but also the NGS-driven analysis of naïve and immune libraries.

1. Introduction

Antibody discovery has been potentiated using *in vitro* display methods – such as yeast- and phage-display – which answer to the limitations of *in vivo* antibody discovery and provide great experimental control over antigen presentation. Whereas previously the immunoglobulin repertoires of selection outputs were analyzed by colony picking and Sanger sequencing, we can now employ Next-generation Sequencing (NGS) to analyze the outcome of *in vitro* selection outputs.[1–4] The former allowed the identification of the most dominant clones, but provided a small snapshot of the selection output and did not guarantee the discovery of the best candidates. The latter allows a much deeper inspection of candidate pools after the final round of panning, by retrieving up to 10^7 sequences.[1,5,6] The bigger depth of NGS analysis allows for the selection of a more diverse candidate dataset by providing alternatives to the dominant "top clones", which arise from careful inspection of patterns across conditions and controls. Additionally, NGS has also proven to be a useful tool for the quality control of antibody libraries, of both natural and synthetic origin, in terms of their CDR length distribution, germline frequencies and clone redundancy.[7–9]

The number of CDRs diversified on a library will impact the NGS strategy employed. Although HCDR3 is widely recognized as the most important CDR for antigen binding, the interplay of all six CDRs is necessary for the antibody-antigen interaction.[10–12] Primary library strategies have increasingly relied on the randomization of other CDRs to discover antibodies against antigens. Ylanthia (randomized in LCDR3 and HCDR3)[9] and HuCAL PLATINUM® (all six CDRs)[13] constitute clear examples of state-of-the-art synthetic libraries that employ such strategies.

However, while all NGS platforms can sequence the HCDR3 in its entirety (≈12–105 bp), most won't be able to capture the full CDR diversity of some libraries. This is especially true if diversity is simultaneously located on the separate variable heavy (V_H) and variable light (V_L) chains and will be exacerbated if a Fab library is used, due to the additional C_L domain in between. PacBio has the longest read lengths available and can effectively capture V_L and V_H diversity, but can only do so at the expense of throughput, which may turn it unsuitable for most library-based methods.[14] Currently, the longest read length available with reasonable throughput is provided by Illumina, which allows the paired-end sequencing of 600 bp reads yielding around 10^7 sequences. Still, 600 bp does not allow to recover information on

* Corresponding author.
E-mail address: abatista@ibet.pt (A.P. Batista).

https://doi.org/10.1016/j.csbj.2022.05.054

full V_L and V_H segments simultaneously on a single read, nor does it allow the overlapping of paired-end reads from those segments in Fab amplicons. The lack of clear alternatives to sequence V_L and V_H fragments means that researchers need to opt between randomizing less CDRs, or to perform analysis of V_L and V_H separately and match them afterwards based on frequency alone.

On this work, we take advantage of Illumina's cluster mapping capabilities to match non-overlapping reads from Fab amplicons and effectively capture diversity from far-apart CDRs that would not otherwise be sequenced simultaneously using most NGS applications. Here, paired-end sequencing of V_L (forward read) and V_H (reverse read) is performed after amplifying the whole region of interest. Then, information on cluster coordinates is used to match the forward and reverse reads *in silico*, providing a trustworthy V_L:V_H pairing methodology that is less prone to mismatches than clone frequency-based approaches.

2. Results

2.1. Cluster coordinates allow the interrogation of V_L and V_H simultaneously

Currently, the longest reads produced on an Illumina platform is accomplished by paired-end sequencing on the MiSeq Reagent Kit v3 (600-cycle), which produces a forward read (R1) and a reverse read (R2) whose lengths sum up to a total of 600 bp. 300 bp is enough to cover the distance between the first residue of CDR1 and the last residue of CDR3, for both V_L and V_H, even when considering an extra 12–15 bp in both 5′ and 3′ ends (used for primer annealing during the amplification step and for query purposes during the data analysis step). Thus, if R1 and R2 are correctly assigned to each other, a full Fab fragment can be reconstructed with the information on all six CDRs. Illumina systems make use of Bridge-PCR to replicate the desired amplicons into a cluster of identical amplicons, as a way to increase the signal.[15] If conditions are optimized correctly, each cluster should be perfectly distinguished from each other as to provide a confident signal to the sequencing equipment. On that regard, we tested two different DNA loading concentrations, and saw that 7.2 pM generated an adequate cluster quality without compromising the total number of reads (Table S1). On each image taken, all clusters will have a specific ID that indicate their lane, tile and XY coordinates inside the composite image generated by the imaging software. As such, we have developed a simple methodology which confidently pairs and concatenates non-overlapping R1 and R2 reads that share the same cluster ID (Fig. 1).

3. Coordinate matching reveals hidden V_L:V_H pairs and avoids mispairings that arise from clone frequency-based analysis.

We applied the coordinate matching method to analyze the outputs of 23 different affinity maturation projects, diversified in LCDR1, LCDR3, HCDR1 and HCDR2. Because these libraries did not have diversity in the HCDR3, we opted for the shorter MiSeq V2 500 bp PE kit instead of the longer MiSeq V3 600 bp PE kit, since the method itself is directly transferable in any Illumina's method that retains cluster coordinates. We compared the results against the traditional frequency-based analysis, which tries to infer V_L:V_H pairs by matching sequences from the R1 and R2 reads (which retrieves information on V_L and V_H, respectively), in descending order, based on their occurrences (i.e. the most frequent clone of R1 is matched with the most frequent clone of R2 and so forth. Table S3). The poor similarity between both approaches highlights the relevance of correctly matching V_L:V_H pairs (Fig. 2, Tables S2 and S3). Only 5% of the Top 100 frequency-inferred clones corresponded to the concatenated sequences, with the remaining clones being artificial sequences that are not found originally in the sample. This effect improved only slightly when the Top50, Top25 and Top10 were compared (Fig. 2, Table S2). The mispairing effects can also be seen when looking exclusively at the top clone of each dataset. We searched the top hits of each dataset within the frequency-inferred dataset and found that in 13 of the projects, the Top1 candidate could have not been found if sequence coordinate matching had not been performed. In the remaining projects, there was a match between the top clones in 9 of the datasets (Top1) and, in another project, there was a match of the top clone of the concatenated dataset with the fifth clone of the inferred dataset (Table S4).

4. Discussion

Capturing information on V_L and V_H fragments is instrumental for reconstructing full antibody sequences after an antibody discovery campaign, in both *in vivo* and *in vitro* settings. While the widespread use of single-cell technologies has allowed V_L:V_H pairings to be retrieved from *in vivo* campaigns,[16,17] the issue has only been partially tackled for the commonly-used phage-display technique. Here, researchers are required to chose between read length and sequencing depth. On one hand, a long-read NGS technology such as PacBio will allow the sequencing of both variable chains, but with a relatively low number of reads (up to $\sim 8 \times 10^4$)[18]. Despite being sufficient to sample high frequency (and likely, high-affinity) clones near the top of the dataset, low-depth sequencing will not allow the capturing of less-frequent rarer clones, which often show equivalent or sometimes higher affinities.[19] This is also incompatible with the deep inspection of motifs and patterns within the dataset.[2] On the other hand, the high-depth NGS platforms available do not have the read length necessary to cover both variable chains, which means the analysis of phage-displayed Fab/scFv libraries or selection outputs is usually restricted to V_L or V_H in those cases.

In this regard, Barreto *et al.* achieved high depths and V_L:V_H pairing by successfully coupling Ion Torrent sequencing with Kunkel mutagenesis to physically concatenate CDRs *in vitro* and remove unwanted segments between CDRs.[19] While effective, this approach may increase the accumulation of errors that stem from extensive PCR-based manipulations and increases the experimental load as higher amounts of DNA are required to fulfill the PCR, electrophoresis and purification steps. Both these drawbacks are exacerbated when working in antibody library and phage-display settings, where diversity, throughput and quality-control are adamant.[20,21] It is also likely that this strategy works best in defined-single frameworks, and that more complex samples with multiple V-families may difficult the site-directed mutagenesis process even more.

Inversely, our approach leverages existing data related to cluster mapping from Illumina's sequencing files to match forward and reverse reads and concatenate V_L and V_H *in silico*, which does not require further processing of output DNA. Additionally, our approach is easily compatible with multiple frameworks via primer-barcoding. Special attention must be given nonetheless to the total amplicon length. Cluster quality will tend to decrease with increasing amplicon lengths, since longer amplicons will lead to clusters with larger diameters, with a higher probability to overlap. Our approach generates 1200 bp amplicons which are near the upper limit of MiSeq's capabilities (≈1500 bp) and as such, an optimization of the concentration of loaded DNA was warranted. We found that a loading concentration of 7.2 pM was sufficient to yield good overall cluster quality without compromising the total number of reads and their fidelity. A total of 1.68×10^7 reads were obtained, with 81.7% having a Phred score (hereinafter referred

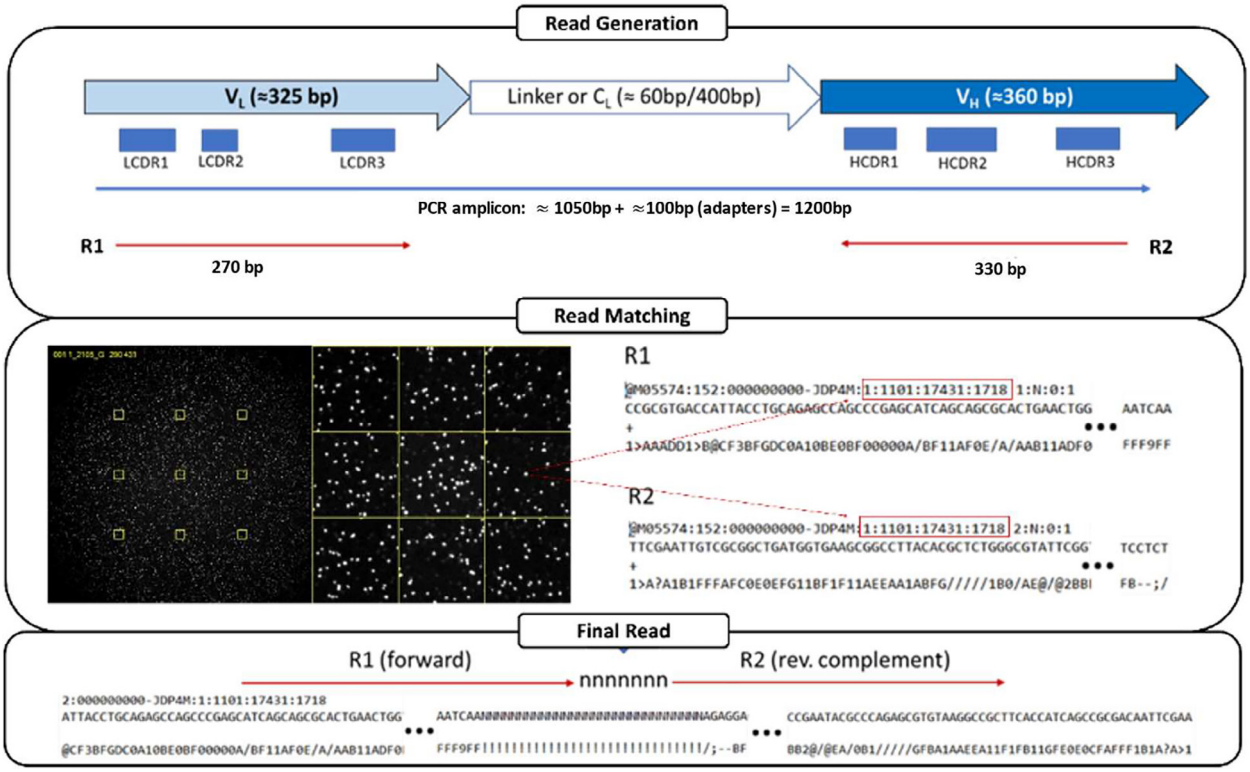

Fig. 1. Schematic representation of paired-end reading and sequence coordinate matching to retrieve correlated V_L:V_H pairs. Read Generation: Big Fab amplicons comprise both V_L and V_H, with C_L in between. In the case of scFv sequences, the total length of the amplicon decreases to about 900 bp. The R1 and R2 reads add up to a maximum of 600 bp (using Illumina's MiSeq system), with R1 shorter than R2 in this case due to the bigger HCDR2 and HCDR3 loops than LCDRs. **Read Matching**: During a sequencing run, clusters are identified on the MiSeq flow cell images and their coordinates are stored on the first row of each information block on the FASTQ raw data files, as follows: <lane>:<tile>:<x-pos>:<y-pos >. The second line of the information block identifies the nucleotide sequence and the third line (after "+") indicates the Q-score of each sequenced nucleotide. **Final Read:** The raw outputs from R1 and R2 runs are matched according to their < lane>:<tile>:<x-pos>:<y-pos > coordinates and used to generate the R1 + R2 reads, which are composed by R1 and the reverse complement of R2. These are united by a string of 30 N nucleotides, to provide a clear separation between the last nucleotides of R1 and the first nucleotides of R2.

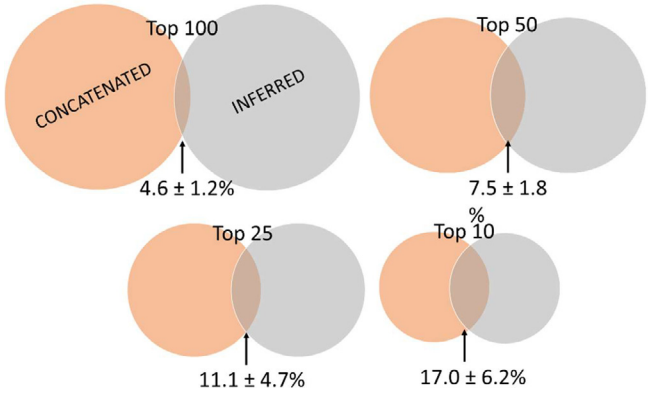

Fig. 2. Intersecting sequences between concatenated and frequency-inferred datasets. Each of the 23 affinity maturation projects was analyzed and their VL and VH reads matched by the sequence coordinate matching method and by the frequency-based method. Independently analyzed R1 and R2 were combined to generate V_L:V_H pairs based on the frequency of each clone (inferred dataset – grey circles). The concatenated and frequency-based datasets were ordered by the occurrences of each sequence and, compared to discover identical sequences on the Top10, 25, 50 and 100, for each affinity maturation project (see also Tables S2 and S3).

to as Q-score) above 30 (Table S1). This method allowed us to extract information on diversified positions (in this case, 8 different amino acid positions across 4 CDRs), with a Q-score > 30 for each nucleotide inspected (24 nt), for at least 100.000 reads per library.

In this paper we have also shown how frequency-based V_L:V_H matching fails to reconstruct trustworthy sequences, regardless of their relative rank within the dataset (Fig. 2, Table S3/S4), which highlights the relevance of our approach to analyze both high- and low-frequency clones. While very high frequency clones may have a chance to have their V_L and V_H correctly matched by frequency-based inference, it is very likely that such a method does not retrieve correct sequences throughout the entire dataset. This is especially relevant considering that the more sophisticated NGS studies require a deep inspection of datasets, such as when looking for motifs and clusters within the dataset[22], when looking for mutations of interest that were predicted by *in silico* tools[2], or when looking for rare clones that were enriched throughout the selection rounds.[19] However, it is also important to note that this analysis and its results are inextricably linked to the types of libraries/sequences assessed, which have 3 randomized positions in R1 and 5 randomized positions in R2. Strong enrichment of certain aminoacids in low diversity reads (e.g. R1) can reduce the ability for a given read to be informative about a clone's identity – i.e. a very enriched sequence in R1 will be shared among many clones in R2 and correctly assigned to only one clone in R2, and incorrectly for the remainder of the clones. It is likely that sequences with more diversified positions will lead to a better outcome when matched based on frequency. In that line of thought, frequency-based matching is expected to decrease as selections become more stringent and certain sequences start dominating the datasets, which further advocates for the use of alternative V_L:V_H matching methods.

Despite being ideal for synthetic Fab and scFv libraries, this work also provides a suitable V_L:V_H pairing methodology for other extensively diversified immunoglobulin repertoires, as long as the aforementioned amplicon length limitation is respected and the correct primer-barcoding strategy is employed. Importantly, the proposed approach tackles several challenges of previous V_L:V_H pairing approaches, on the most widely used NGS platform – Illumina's MiSeq –, with high sequencing depth (>10^7 reads), high read fidelity, and without increasing the experimental burden.

5. Methods

DNA preparation and NGS: Plasmid DNA was isolated directly from the phage-infected cells from the selection round of interest using the GeneJET Plasmid Miniprep Kit (Thermo Scientific™, K0502). Isolated dsDNA was quantified on the Qubit 3.0 fluorometer using the Qubit® dsDNA HS kit (Invitrogen™ Q32851). The generation of V_L:V_H amplicons for sequencing was performed through two PCRs. To amplify the region of interest and to insert the adapter regions for the NGS, the initial PCR utilized a forward primer specific to the vector leader sequence prior to LCDR1 and, since we did not need HCDR3 information in our affinity maturation setting, a reverse primer downstream of HCDR2 was employed. The second PCR inserted the TruSeq universal adapter and the indexes, which are used to distinguish between different samples (i.e. libraries). Samples were quantified in Qubit 3.0, pooled in equimolar proportions, and ran on an electrophoresis gel. Bands with the appropriate size were excised, purified using the Wizard SV Gel and PCR Clean Up System (Promega, A9281) and quantified on Qubit 3.0. The pool was diluted to a final concentration of 4 nM, spiked with 20% PhiX (Illumina; FC-110–3001), denatured for 5 min in 0.1 N of NaOH (5 μL of DNA + PhiX at 4 nM mixed with 5 μL 0.2 N of NaOH), diluted in HT buffer (provided on the NGS kit; kit details, ahead) to 7.2 pM and sequenced on the Illumina MiSeq platform using the 500 cycle v2 kit (Illumina; MS-102–2003). The forward read was 270 bp in length while the reverse read was 230 bp. R1 retrieves information on LCDR1, LCDR2 (non-diversified), and LCDR3. R2 retrieves information on HCDR1 and HCDR2. Note: we performed a sequencing between LCDR1 and HCDR2 with a 500 cycle v2 kit, but the procedure is directly transferrable (to our best knowledge) to a sequencing between LCDR1 and HCDR3 with a 600 cycle v3 kit (Illumina, MS-102–3003) by sequencing 270 + 330 bp (as explained on Fig. 1). An R2 of 330 bp starting at the end of the VH domain will allow to sequence the entirety of HFR4 and up to the last residues of HFR1, with the coverage on HFR1 depending on framework and HCDR3 length essentially. Longer HCDR3 lengths will require that R2 covers less residues of HFR1 or that HFR4 is less sequenced, which can be achieved by designing primers for that specific purpose.

NGS data analysis: The data analysis of the NGS FastQ output files was performed as described previously.[3] For the panning output of each library, 10^5 sequences were analyzed using the fixed-by-design flanking sequences on the boundary of the diversified positions as template to locate and segment out mutations. We queried a total of 8 amino acid positions, specifically: 1aa in LCDR1, 2aa in LCDR3, 2aa in HCDR1 and 3aa in HCDR2. Sequences that contained at least one nucleotide with a Q-score < 30 on the positions of interest were discarded, allowing for 10^5 sequences for each library with a high degree of quality and confidence. Full CDR sequences were reconstructed by coupling the regions fixed-by-design with the information of diversified regions.

To reconstruct a full Fab sequence with V_L and V_H information, we compared two methodologies (as explained on the second section of Results). The first methodology generates a dataset containing correlated V_L:V_H information after concatenation of R1 and R2 using their sequence's coordinates (i.e. the new method proposed here), as described in the Results section. The second methodology tries to infer V_L:V_H pairs from sequence's frequency counts and relative position within the dataset (i.e. the most frequent clone of R1 is matched with the most frequent clone of R2 and so forth). Sequences in both datasets that only had one occurrence were removed from the analysis.

Three additional Methods' sections can be found on the Supplementary Information, which are important for the generation of the libraries, but non-essential for the purpose of this manuscript.

6. Additional information

6.1. Data Availability:

The data used for this work is generated after the paired-end sequencing of amplicons, produced from the amplification of Fab sequences rescued from phage-display selections. In total, 23 fastq files (1 per library) were generated for the forward R1 reads – which sequenced V_L regions –, plus 23 fastq files (1 per library) were generated for the reverse R2 reads – which sequenced V_H regions. Additional 23 fastq files were originated from the merging of forward and reverse reads. The data from these files were analyzed in 23 different Excel spreadsheets to generate the data found on Fig. 2 and Table S2 and Table S3. All of this data are available upon request, including the V_L:V_H pair matching script.

Author contributions

JMS and AFF conceived and designed the experiments. JMS performed the experiments, analyzed the results, and drafted the manuscript. AFF critically discussed the data and helped to draft the manuscript. RB devised the in-house NGS analysis tool used for CDR determination. MAA programmed the sequence matching protocol. APB and SE coordinated the study, critically discussed the data, and helped to draft the manuscript. All authors read and approved the final manuscript.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

We acknowledge Inês A. Isidro for all the insightful comments and discussions. This work was supported by iNOVA4Health – UIDB/04462/2020 and UIDP/04462/2020, a program financially supported by Fundação para a Ciência e Tecnologia / Ministério da Ciência, Tecnologia e Ensino Superior, through Portuguese national funds. JMS personally acknowledges FCT-MEC fellowship PD/BD/128321/2017.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.csbj.2022.05.054.

References

[1] Rouet R, Jackson KJL, Langley DB, Christ D. Next-Generation Sequencing of Antibody Display Repertoires. Front Immunol 2018;9:118.

[2] Campbell SM et al. Combining random mutagenesis, structure-guided design and next-generation sequencing to mitigate polyreactivity of an anti-IL-21R antibody. mAbs 2021;13:1883239.
[3] Liu G et al. Antibody complementarity determining region design using high-capacity machine learning. Bioinformatics 2020;36:2126–33.
[4] Senatore A et al. Protective anti-prion antibodies in human immunoglobulin repertoires. EMBO Mol Med 2020;12:e12739.
[5] Yang W et al. Next-generation sequencing enables the discovery of more diverse positive clones from a phage-displayed antibody library. Exp Mol Med 2017;49:e308.
[6] Noh J et al. High-throughput retrieval of physical DNA for NGS-identifiable clones in phage display library. mAbs 2019;11:532–45.
[7] Zhai W et al. Synthetic Antibodies Designed on Natural Sequence Landscapes. J Mol Biol 2011;412:55–71.
[8] Ravn U et al. Deep sequencing of phage display libraries to support antibody discovery. Methods 2013;60:99–110.
[9] Tiller T et al. A fully synthetic human Fab antibody library based on fixed VH/VL framework pairings with favorable biophysical properties. mAbs 2013;5:445–70.
[10] Schroeder HW, Cavacini L. Structure and function of immunoglobulins. Journal of Allergy and Clinical Immunology 2010;125:S41–52.
[11] North B, Lehmann A, Dunbrack RL. A New Clustering of Antibody CDR Loop Conformations. J Mol Biol 2011;406:228–56.
[12] Sela-Culang I, Kunik V, Ofran Y. The Structural Basis of Antibody-Antigen Recognition. Front Immunol 2013;4.
[13] Prassler J et al. HuCAL PLATINUM, a Synthetic Fab Library Optimized for Sequence Diversity and Superior Performance in Mammalian Expression Systems. J Mol Biol 2011;413:261–78.
[14] Goodwin S, McPherson JD, McCombie WR. Coming of age: ten years of next-generation sequencing technologies. Nat Rev Genet 2016;17:333–51.
[15] Slatko BE, Gardner AF, Ausubel FM. Overview of Next Generation Sequencing Technologies. Curr Protoc Mol Biol 2018;122:e59.
[16] Tanno H et al. A facile technology for the high-throughput sequencing of the paired VH:VL and TCRβ:TCRα repertoires. Sci Adv 2020;6:eaay9093.
[17] DeKosky BJ et al. High-throughput sequencing of the paired human immunoglobulin heavy and light chain repertoire. Nat Biotechnol 2013;31:166–9.
[18] Nannini F et al. Combining phage display with SMRTbell next-generation sequencing for the rapid discovery of functional scFv fragments. mAbs 2021;13:1864084.
[19] Barreto K et al. Next-generation sequencing-guided identification and reconstruction of antibody CDR combinations from phage selection outputs. Nucleic Acids Res 2019;47:e50.
[20] Kanagawa T. Bias and artifacts in multitemplate polymerase chain reactions (PCR). J Biosci Bioeng 2003;96:317–23.
[21] Fox EJ, Reid-Bayliss KS, Emond MJ, Loeb LA. Accuracy of Next Generation Sequencing Platforms. Next Gener Seq Appl 2014;1.
[22] Norman RA et al. Computational approaches to therapeutic antibody design: established methods and emerging trends. Briefings Bioinf 2020;21:1549–67.