MABS
2024, VOL. 16, NO. 1, 2408344
https://doi.org/10.1080/19420862.2024.2408344

Taylor & Francis
Taylor & Francis Group

REPORT

Seq2scFv: a toolkit for the comprehensive analysis of display libraries from long-read sequencing platforms

Marianne Bachmann Salvy, Luca Santuari, Emanuel Schmid-Siegert, Nikolaos Lykoskoufis, Ioannis Xenarios, and Bulak Arpat

NGS-AI Division, JSR Life Sciences, Epalinges, Switzerland

ABSTRACT

Antibodies have emerged as the leading class of biotherapeutics, yet traditional screening methods face significant time and resource challenges in identifying lead candidates. Integrating high-throughput sequencing with computational approaches marks a pivotal advancement in antibody discovery, expanding the antibody space to explore. In this context, a major breakthrough has been the full-length sequencing of single-chain variable fragments (scFvs) used in *in vitro* display libraries. However, few tools address the task of annotating the paired heavy and light chain variable domains (VH and VL), which is the primary advantage of full-scFv sequencing. To address this methodological gap, we introduce Seq2scFv, a novel open-source toolkit designed for analyzing *in vitro* display libraries from long-read sequencing platforms. Seq2scFv facilitates the identification and thorough characterization of V(D)J recombination in both VH and VL regions. In addition to providing annotated scFvs, translated sequences and numbered chains, Seq2scFv enables linker inference and characterization, sequence encoding with unique identifiers and quantification of identical sequences across selection rounds, thereby simplifying enrichment identification. With its versatile and standalone functionality, we anticipate that the implementation of Seq2scFv tools in antibody discovery pipelines will efficiently expedite the full characterization of display libraries and potentially facilitate the identification of high-affinity antibody candidates.

ARTICLE HISTORY
Received 4 July 2024
Revised 15 September 2024
Accepted 19 September 2024

KEYWORDS
Antibody discovery; long-read sequencing; PacBio; phage display; scFvs

1. Introduction

Due to their high target specificity and strong binding affinity, antibodies have become the leading class of biotherapeutics, with their market share continuously expanding.[1] However, conventional wet-lab techniques for antibody screening require significant time, effort, and technical expertise.[2] Additionally, these methods are limited to analyzing only a small fraction of the antibodies generated.[3] Relying solely on experimental approaches makes it challenging to identify binders that target different epitopes and exhibit a range of biophysical characteristics,[3] including those critical for developability.[2,4]

Incorporating computational approaches and Next Generation Sequencing (NGS) into antibody development pipelines for rational design can effectively overcome these limitations. Indeed, NGS-based computational methods can track the frequency and progressive enrichment of clones in *in vitro* display experiments,[5] guiding the selection of lead antibody candidates with higher sequence and epitope diversity.[6,7] Furthermore, computational techniques and novel Artificial Intelligence (AI) methods prove valuable in later stages of antibody development, including improvement of antibody binding affinity, humanization, homology modeling and many others, by leveraging structural bioinformatics or biophysical property characterization as reviewed in.[2,8–10]

Unlocking the full potential of computational and AI methods in immunogenetics relies on accurately identifying and thoroughly characterizing antibody sequences.[11] This process typically begins with mapping antibody DNA sequence reads, such as those derived from NGS, against germline gene databases to identify V(D)J rearrangements in each V-domain. Tools such as IgBLAST,[12] IMGT/HighV QUEST,[13] MiXCR,[14] and others, as extensively reviewed by Norman et al.[10] and Smakaj et al.,[11] are commonly used in this process. These tools generally also identify the correct translation frame of each variable domain (V-domains) and delimit their V, D and J regions.[11] Further characterization involves antibody numbering, which assigns numerical identifiers to individual amino acid residues within the V-domains. This contextualizes each position within the antibody structure according to established schemes such as Chothia,[15] Kabat,[16] IMGT[17] and several others reviewed in Dondelinger et al.[18] Tools such as ANARCI,[19] AbRSA,[20] AbNum[21] or AntPack[22] have been developed for antibody numbering.

In parallel to advancements in bioinformatics methods, the innovative application of Pacific Biosciences' (PacBio) long-read HiFi sequencing to *in vitro* display libraries has enabled the high-throughput full-length sequencing of recombinant antibodies.[23] Previously, reading both the heavy (VH) and light (VL) variable domains that constitute single chain variable fragments (scFvs) was a challenge due to the limitation of

CONTACT Bulak Arpat barpat@jsrlifesciences.com NGS-AI Division, JSR Life Sciences, Phenyl Building, Route de la Corniche 3, Epalinges 1066, Switzerland
Supplemental data for this article can be accessed online at https://doi.org/10.1080/19420862.2024.2408344

most NGS platforms to fragments of 500 bp, while scFvs typically span around 850 base pairs (bp).[24] Fully sequencing both chains is particularly valuable. Indeed, while the complementarity determining regions (CDRs) are known to play a significant role in binding,[25] with the third CDR of the heavy chain (HCDR3) being crucial for antibody specificity,[26] evidence suggests that other regions also contribute to binding.[27] Furthermore, the light chain plays a role in specificity,[28] with specific regions influencing the structure of the HCDR3 loop.[29] Therefore, sequencing the entire scFv provides crucial information for identifying better binders.[30] The advent of Single Molecule Real-Time (SMRT) sequencing on PacBio platforms has already enabled the identification of antibodies with a diverse range of affinities, epitopes and biophysical characteristics, by leveraging full-length scFvs information in *in vitro* display libraries.[3,31]

Currently, there is a gap in the landscape of antibody annotation methods as not many tools address the annotation of VH and VL in a single read, a requisite for the analysis of full length scFvs.[32] Though IMGT/HighV-Quest[13] does provide a scFv analysis option, it is only provided as a web server and is not open source. Additionally, it limits the V(D)J alignment to IMGT sequences, as it does not provide the option of using custom germline gene databases.

Here, we introduce Seq2scFv, a comprehensive collection of integrated approaches and scripts tailored for analyzing full-length scFvs derived from *in vitro* display libraries and long-read sequencing. Open-source for academic use under a noncommercial license, while subject to a separate license for any other purposes, this toolkit streamlines widely used bioinformatic tools with a suite of custom algorithms to process sequences into fully delimited and annotated scFvs. Scripts are provided for sequence cataloging and encoding, characterization and numbering of both V-domains, scFv delimitation and conformation validation. Additionally, a framework for the inference of the linker sequence from the data is provided, together with scripts for evaluating the most frequent linker sequences and their length distribution. The comparison between the consensus or reference linker sequence (if known beforehand) and the linker sequence in each individual read can be used as an additional quality parameter for selecting well-annotated scFvs. A script for flagging scFvs on other quality criteria is also provided. Importantly, Seq2scFv can quantify representative sequences across successive selection rounds, facilitating downstream enrichment analysis for frequency-based lead antibody candidate selection. Outputs from the different analysis steps in the framework are primed for easy analysis, sub-selection, or exportation to human-readable formats. Emphasizing flexibility and consistency, this collection of tools accommodates user-specified parameters, including IMGT and Kabat annotation and numbering schemes, as well as user-provided germline gene databases.

In addition to introducing Seq2scFv, we demonstrate its real-world application using openly accessible data from Nannini and colleagues, published in 2020.[3] By processing this example dataset, we showcase the use of Seq2scFv in analyzing *in vitro* display libraries derived from long-read sequencing platforms. Our aim is to provide researchers with a practical toolkit for comprehensive antibody characterization. It is outside the scope of this article to select potential binding candidates based on Seq2scFv annotation and experimentally validate them. For validation of the NGS-guided selection as a frequency-based approach to identify potential binders, we refer to the original publication by Nannini et al.[3] and the growing literature in the field.[5–7,31,33] Beyond its application in antibody discovery, characterized scFvs can be compiled in large datasets to leverage machine learning applications for antibody discovery and engineering.[30]

2. Methods

The main steps for analyzing long-read display libraries with Seq2scFv consist in tagging sequences with unique IDs, library merging, V (D)J alignment, scFv delimitation, linker detection and scoring, flagging and counting. Figure 1 provides an overview of the analysis framework, and the sections below describe in detail the tools developed and implemented, as well as the type and origin of the publicly available dataset.

2.1. Input, library merging and SEGUID cataloguing

FASTA files, each representing a distinct library sample (or "panning round"), are provided as input. It is expected that demultiplexing, adapter trimming, and reverse-complementation (to ensure all sequences are in the same orientation) have already been performed. Additional quality preprocessing can also include filtering out sequences falling outside of the expected insert size interval. Though preprocessing is not part of the Seq2scFv tools, to allow users the flexibility to choose their preferred software, the methods used to prepare the example dataset are detailed in subsection 2.7.2, and code examples are provided in Supplementary Information 1. This also includes guidance on converting from uBAM, the usual output of PacBio, to FASTQ and subsequently to FASTA.

To facilitate the tracking of clone amplification through panning rounds, all sequences from the different panning rounds are pooled into a single FASTA file before antibody identification and characterization. This process involves assigning a Sequence Globally Unique Identifier (SEGUID[34]) to each sequence and maintaining a correspondence file to track its library origin. Using seqkit v2.7.0,[35] only one representative among identical sequences encoded by the same SEGUID is retained in the combined FASTA file, effectively removing redundancy and optimizing the computational workload for downstream antibody characterization. Despite this, the correspondence file preserves the occurrence of sequences across different panning rounds, ensuring that individual clone amplification monitoring is not compromised.

2.2. Alignment of sequences to germline V, D and J genes

The identification of VH and VL in the query read employs a method resembling the IMGT/HighV-QUEST algorithm for scFv sequence analysis.[32] In IMGT/HighV-QUEST, a V-domain is initially identified and characterized, followed by the search and characterization of a second V-domain in

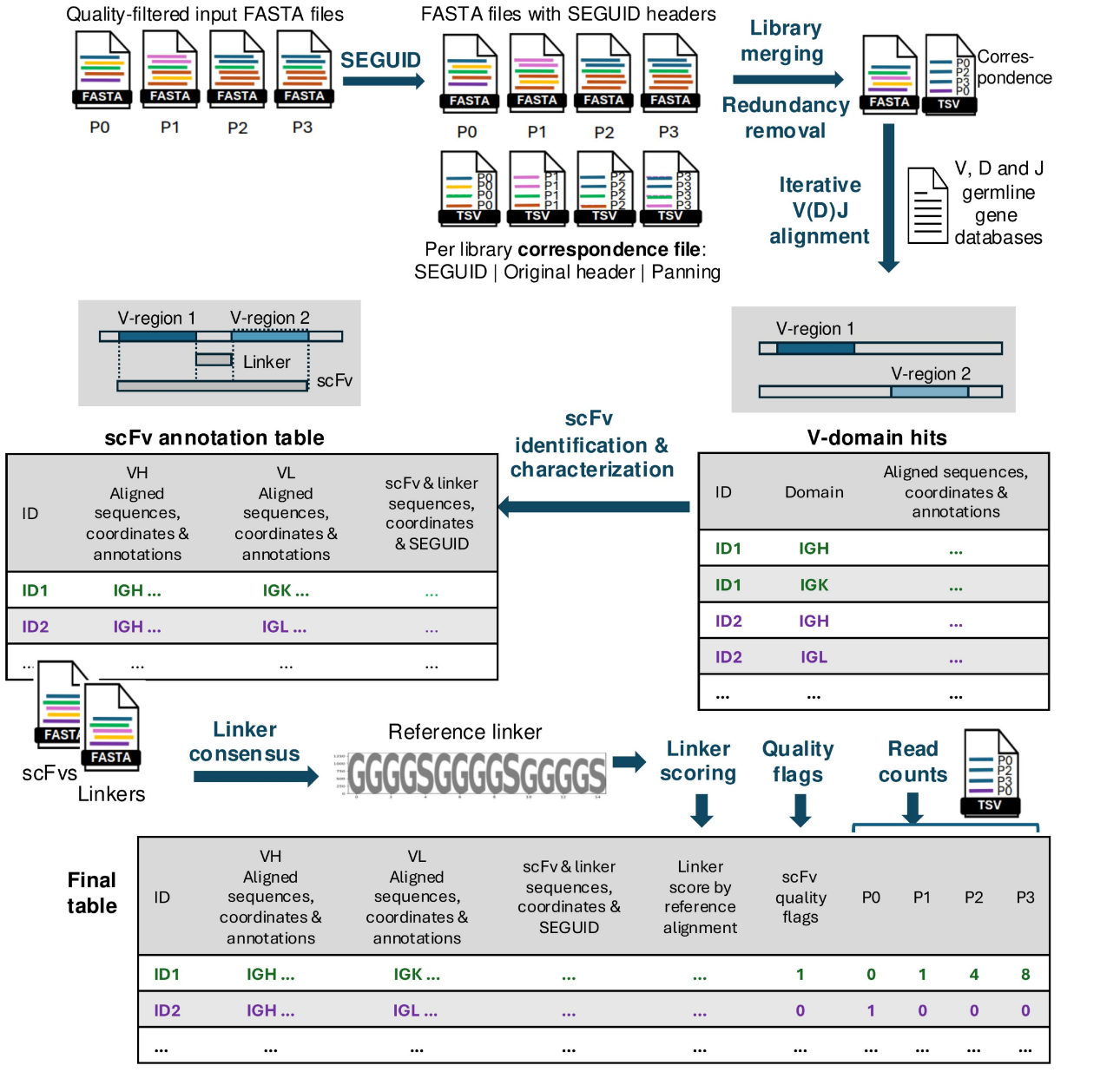

Figure 1. The main processes are indicated with blue arrows and text. Input FASTA files from different panning rounds (P0, P1, P2, P3) undergo renaming with SEGUIDs. They are then merged, cataloged, and deduplicated to remove sequence redundancy. Unique reads are aligned to V, D, and J regions, resulting in a list of VH and VL hits. Reads with one VH and one VL are selected for further scFv delineation, linker sequence inference, and evaluation. The process of identifying scFvs from the alignment hits is illustrated in the grey boxes. The final output includes quality flags and read count data across panning rounds.

the flanking regions, provided they exceed a length threshold of 150 bp. Then, coordinates of the two V-domains are used to delimit the position of the linker sequence.

In this approach, we employ IgBLAST,[12] a specialized alignment tool for immunoglobulins, to identify V-domains. By mapping the nucleotide sequence against a user-provided germline gene database, IgBLAST delineates the V, (D) and J regions that constitute a V-domain, generating a tabular file conforming to the Adaptive Immune Receptor Repertoire (AIRR) Rearrangement format.[36] The table encompasses matched germline genes, nucleotide and amino acid alignment sequences, coordinates and scores, and framework region (FWR) and CDR sequences and coordinates (further details provided in Table S2). IgBLAST v. 1.20.0 is used in an iterative process in a Python script: once a V-domain is identified, adjacent regions are split, and if they meet a user-defined length threshold, they are submitted to a new IgBLAST search. This iterative process annotates all possible V-regions within a read until no additional splitting is possible. Finally, the script reunites all V-domain hits for all sequences into a single tab-separated file, with one hit per row, and adjusts the alignment coordinates to reflect the original sequence.

2.3. scFv delimitation and characterization

To identify single-chain variable fragments (scFvs), a custom set of Python scripts are used. Initially, reads with precisely one VH (IGH locus) and one VL (IGK or IGL locus) hit reported by IgBLAST are selected, while sequences with a different number of hits are excluded. Columns are annotated with

either VH or VL prefixes according to the identified locus. Subsequently, the VH and VL information is merged, consolidating each row in the table to represent a putative scFv rather than individual hits (similar to the format for paired antibodies used in the Observed Antibody Space database[37]). Additionally, sequences containing a stop codon in either VH or VL domain, as determined by IgBLAST's sequence translation, are removed from further consideration.

Next, the translation frame of the full read is determined, ensuring that both VH and VL amino acid sequence alignments are contained within the same frame. Sequences failing to meet this criterion are designated as "out of frame". Upon identifying the read's translation frame, the scFv is delineated at both the amino acid and nucleotide levels leveraging IgBLAST's alignment coordinates. The start and end of the scFv are defined by the 5' end of the first V-domain and the 3' end of the second V-domain, respectively, while the linker region is bounded by the 3' end of the first V-domain and the 5' end of the second V-domain.

Based on these delimitations, sequences and positions of the scFv and linker at both the nucleotide and amino acid levels are incorporated into the table. Additionally, scFv sequences are encoded with a SEGUID to facilitate clone cross-library amplification and cross-experiment comparison. Furthermore, scFv characterization by Seq2scFv tools involves numbering of both V-domains using AntPack, which employs a global alignment approach with a custom scoring matrix.[22] Alongside a list of numbered positions and other AntPack annotations, fields reporting the consistency between alignment and numbering approaches are included (further details provided in Table S2).

The final steps of scFv characterization with Seq2scFv tools involve annotating CDR and FWR regions. While IgBLAST reports sequences at both nucleotide and amino acid levels, positions are only provided at the nucleotide level and are relative to the full read. To simplify integration with downstream analyses, amino acid and nucleotide coordinates are added and updated, respectively, to be relative to the corresponding delimited chain.

After completing the annotation process, output tables and FASTA files are generated. The FASTA files contain nucleotide and amino acid sequences of scFvs and linkers, while the tables report the annotations. An 'in-frame', paired, and delimited table consolidates all correctly annotated scFvs, while three separate tables compile unpaired V-domains, scFvs with non-standard amino acids, and out-of-frame scFvs.

2.4. Linker sequence evaluation

The characterization of scFvs by Seq2scFv can include a systematic and thorough assessment of the inferred linker sequences, which may deviate from the designed linker due to mutations or annotation issues. For instance, if an uncommon V or J gene is present in one of the V-domains, or somatic hypermutation (SMH) introduced many mutations, the alignment might fail to accurately pinpoint its position in the read and infer the correct start and end coordinates of the linker sequence. Reporting linker sequence discrepancies enables a more comprehensive assessment of annotation quality.

Identifying linker sequence divergences requires an amino acid reference linker, which can either be supplied by the user or inferred from the data. To derive a consensus sequence from the dataset, nucleotide and amino acid linkers files are randomly subsampled from their respective FASTA files, with the user having the option to select the proportion of sequences sampled, and aligned with Clustal Omega v.1.2.4.[38] Subsequently, sequence logos are generated using LogoMaker,[39] along with tables reporting the weight (fraction of non-gapped symbols) of each position in the alignments. This weight data enables the determination of the most prevalent nucleotides or amino acids in significant positions, defined by exceeding a predetermined weight threshold. In instances where no user-provided reference linker is available, the consensus linker is considered as reference. Additionally, to provide further characterization of the linkers, tools in Seq2scFv can be used to generate sequence logos and a report containing the top ten linkers ranked by frequency, along with frequency tables and histograms depicting linker lengths.

Furthermore, the scFv annotation table is revisited to align each reported linker with the reference linker. This alignment process, incorporating customized penalties to enhance alignment score interpretability (for more details, refer to Seq2scFv documentation), is conducted using the Smith-Waterman algorithm. Three columns are added to provide insights into linker annotation and additional quality filters for post-processing: the percentage of aligned amino acids relative to the length of the reference linker sequence, the number of mismatched positions, and the count of nonaligned amino acids in the inferred linkers, referred to as "overhang."

2.5. scFv quality flags

Quality flags can be added with tools in Seq2scFv tools to facilitate downstream analysis and filtering based on predefined criteria. These flags utilize a binary annotation (0/1) to indicate whether a given scFv meets specific quality thresholds. Criteria assessed include the correct order of VH and VL domains as well as the linker annotation quality, defined by the user by providing thresholds for the alignment score, the number of mismatches and the length of the linker overhang. Additionally, flags are assigned based on whether each variable domain exceeds a minimum amino acid sequence length and meets IgBLAST criteria for being classified as "productive" (i.e., the V(D)J rearrangement frame is in frame, no stop codon is present, and there are no internal frame shifts in the V gene) and "complete" (i.e., the sequence alignment spans the entire V(D)J region from the first V gene codon to the last complete codon of the J gene).

2.6. Read counts

The final process in the framework proposed by Seq2scFv generates count data for each read listed in the scFv annotation table. This process retrieves entries from the correspondence file, where all identical reads were encoded with a SEGUID, while also recording their originating libraries. Thus, for each SEGUID-encoded read in the annotation table, one column

per panning round is appended. These columns denote the frequency with which identical read was observed in each library. Leveraging the comprehensive annotations in the scFv annotation table, these counts can be aggregated according to various regions of interest, such as full amino acid or nucleotide scFv sequences, VH or VL domains, CDRs, among others.

2.7. Example dataset

2.7.1. Phage libraries

The phage libraries utilized to showcase the application of Seq2scFv were sourced from the publication by Nannini and colleagues (2020).[3] In their study, phage display libraries were generated through immunization of Wistar rats with the CD160 antigen, followed by three iterative biopanning rounds against the targeted gene (further details regarding this study can be found in their publication). The initial phage library (P0), along with the subsequent panning rounds (P1, P2, P3), were sequenced using a Pacific Biosciences RS sequencer. The site https://rdr.ucl.ac.uk/ndownloader/files/28624629https://rdr.ucl.ac.uk/ndownloader/files/28624629 was last accessed in October 2022 to download the published FASTQ files.

2.7.2. Sequence preprocessing

Basseto v.0.3.7[40] was used to selectively retain reads with an expected accuracy of 99%. Next, primer sequences used for phage library amplification prior to sequencing were trimmed using cutadapt v.4.1,[41] while ensuring orientation consistency through reverse-complementation. Only sequences flanked by the primers detailed in Nannini et al.[3]) were kept (5' primer GTCGTCTTTCCAGACGTTAGT and 3' primer CAGGAAACAGCTATGAC). Sequences were then further filtered with cutadapt to select those falling within the expected insert size interval of 900–1200 bp. Processed sequences were then transformed into FASTA format using basseto v.0.3.7. to conform to Seq2scFv input requirements.

2.7.3. Germline gene database

Sequences from Nannini and colleague's publication (2020)[3] were searched against *Rattus norvegicus* immunoglobulin germline gene sequences. These were obtained from the IMGT reference directory;[42] available at https://www.imgt.org/vquest/refseqh.htmlhttps://www.imgt.org/vquest/refseqh.html.) and formatted according to IgBLAST web instructions (available at https://ncbi.github.io/igblast/cook/How-to-set-up.htmlhttps://ncbi.github.io/igblast/cook/How-to-set-up.html). The auxiliary file, containing additional information for each J gene to facilitate CDR3 annotation, was obtained from IgBLAST directory, as they provide this file for all IMGT and NCBI databases.

2.7.4. Parameters

For the V(D)J alignment, parameters included the specification of "rat" as the organism, the path to the germline gene databases, a minimum length of 150 bp for the iterative IgBLAST search and a minimum E-score of 0.01 to validate a hit. Subsequently, during the flag addition step, criteria such as a 90% percent identity score and a maximum of 2 mismatches between inferred linkers for each scFv and the consensus linker derived from the data were incorporated. Also, a minimum domain length of 80 base pairs for both VH and VL domains was imposed. Since information regarding the reference linker sequence was absent, the linker parameter was marked as "undefined". Additionally, for the identification of the consensus linker sequence from the scFvs with the use of Clustal Omega v.1.2.4[38] and LogoMaker,[39] a subsampling of 10% of the linker sequences was chosen.

3. Results

Full scFv characterization of Nannini *et al.* phage display libraries (2020)[3] was swiftly conducted using the framework and comprehensive collection of tools Seq2scFv. The raw sequence counts from panning rounds 0, 1, 2, and 3 were 15,395; 10868; 30607; and 47,013 reads, respectively. After quality filtering and redundancy removal, the 56,953 reads led to the identification of a total of 14,764 unique scFv across all panning rounds after alignment against the *R. norvegicus* immunoglobulin germline genes, scFv delimitation and characterization.

3.1. Preprocessing statistics

Records of the number of reads per panning library across preprocessing stages are presented in Table 1. In the case of Nannini et al.,[3] publicly available sequences from panning rounds 0, 1, and 2 already exceeded the threshold of a read quality higher than 0.99, while 1,775 sequences from panning round 3 were discarded due to lower quality (3.78%). Between 87.54% and 99.01% of the quality-filtered sequences were correctly flanked by the provided sequences used for PCR amplification. However, due to size filtering, a significant number of reads from panning rounds 0 and 3 were removed, with only 58.16% and 33.67% of the high-quality reads respectively meeting the criteria of correct adapter flanking and an insert size between 900 and 1200 bp. In contrast, the filtering was less stringent for panning

Table 1. Read counts across preprocessing steps.

Panning reads	Raw	Quality trimmed reads	Quality filtered reads (% of raw reads)	Adapter-trimmed reads	Reads with adapter (% of quality-filtered reads)	Size filtered reads	Reads with adapter and selected size (% of quality-filtered reads)	Unique reads	Unique reads (% of quality, adapter and size filtered reads)
0	15,395	15,395	100.00	15,243	99.01	8,953	58.16	8,953	58.16
1	10,868	10,868	100.00	10,470	96.34	9,924	91.31	9,889	90.99
2	30,607	30,607	100.00	29,137	95.20	26,283	85.87	23,574	77.02
3	47,013	45,238	96.22	39,600	87.54	15,232	33.67	14,893	32.92

Read counts across preprocessing steps for panning rounds 0 to 3, with percentages of sequences kept relative to the previous processing point.

rounds 1 and 2 (91.31% and 85.57%, respectively). Regarding sequence redundancy, the unselected library (panning round 0) exhibited the highest diversity, with no repeated sequences, while the proportion of identical sequences increased for panning rounds 1 and 2. However, it decreased again for panning round 3.

Additional quality assessment can be achieved by examining the distribution of read lengths. In Figure 2a, the presence of sequences falling outside the 900–1200 bp range may indicate contamination or unpaired VH or VL. Furthermore, narrowing the focus to a specific interval of expected read lenghts (Figure 2b) allows the identification of an accumulation of sequences at particular lengths concurrent with an increase in selective pressure (i.e., with successive rounds of panning), implying an amplification of clones with specific lengths and, ultimately, binding affinities.

3.2. scFv characterization

Preprocessing, redundancy removal, SEGUID encoding and pooling of libraries, resulted in a single file containing 56,953 unique sequences. These sequences underwent V-domain searches using IgBLAST, which identified a total of 112,259 V-domains. However, 1,602 V-domains were excluded from further analysis due to hits with E-scores lower than 0.01 at the V or J gene alignments, or because of the absence of a second V-domain within the same read. Additionally 40,565 sequences were discarded due to being out of frame, either from a discrepancy in frame between the VH and the VL (3,486) or the presence of a stop codon in the linker or any of the V-domains (37,079). The frame discrepancy of the VH-VL affected 5.91% out of the total reads. However, none of the sequences presented nonstandard amino acids within the sequence.

A comprehensive characterization of each domain was compiled into a table format for 14,764 in-frame, paired scFvs. This entailed IgBLAST analysis for each VH and VL, detailing V(D)J gene alignment specifics such as top germline gene match, alignment positions, translations, and FWR and CDR delimitation. Notably, FWR and CDR information was configured to indicate nucleotide and amino acid coordinates relative to each V-domain, using a 1-based system, rather than relative to the full read. Further, V-domain annotation included the generation of ungapped versions of VH and VL nucleotide sequences, along with IMGT numbering for each amino acid chain, and assessing alignment versus numbering consistency.

In addition to V-domain annotation, the table also compiled scFv-level and read-level information. This encompassed scFv and linker coordinates relative to the read, as well as their full sequences and SEGUIDs encoding scFv nucleotide and amino acid sequences. Simultaneously, 13,966 unique nucleotide scFv sequences and 13,598 unique amino acid scFv sequences were written to FASTA files, with SEGUIDs serving as headers. Additional columns described the alignment quality of each detected linker against the inferred consensus, referring to the percentage alignment, the mismatched positions, and the alignment overhang. Quality across several user-specified parameters (minimum V-domain lengths, linker annotation quality, scFv conformation) were summarized by different flags, included in the annotation table. Finally, read occurrences across the different panning libraries were reported in the last columns, with one column per panning round.

3.3. Linker inference and evaluation

A comprehensive evaluation of inferred linker sequences was undertaken with Seq2scFv. Since no user-supplied linker sequence was available, the consensus amino acid linker sequence $(GGGGS)_3$ was used as reference for alignment against all other linkers. This consensus linker was derived through alignment and consensus determination from randomly subsampled amino acid linker sequences using Clustal Omega v. 1.2.4 and LogoMaker, respectively. The generated amino acid sequence logo (Figure 3) provides a succint overview of the conservation within the inferred linker sequence, while Table 2 presents the top 10 most frequent amino acid linker sequences identified in the dataset. Furthermore, the distribution of amino acid linker lengths is depicted in Figure 4, offering a visual representation of the variability in inferred linker lengths across the dataset. Most sequences show lengths consistent with the consensus linker sequence (15 amino acids). Nevertheless, 1,834 scFvs report a linker length of 10 amino acids, warranting special attention and flagging. Other linker lengths display negligible frequencies within the dataset. Nucleotide linker inference is presented in Supplementary Information (Figure S1, Table S1, Figure S2).

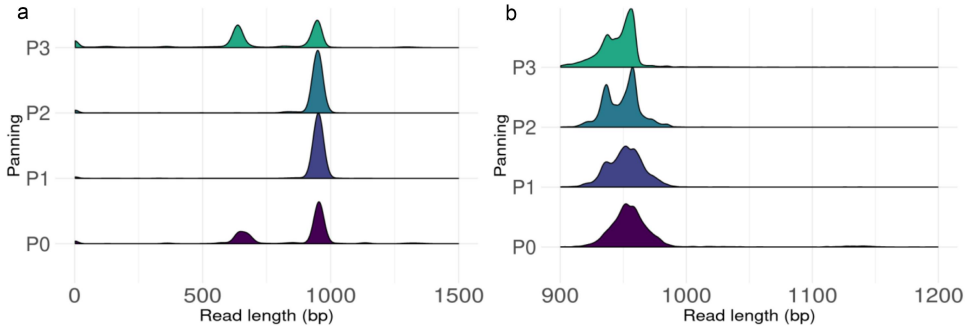

Figure 2. Read length distribution of sequences flanked by the PCR primer sequences. Panel a) displays the distribution without limits, while panel b) exhibits the distribution within the user-provided interval.

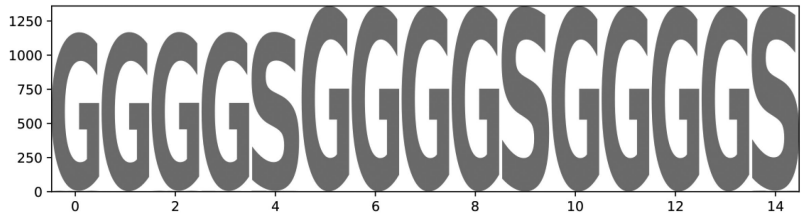

Figure 3. Sequence logo illustrating the amino acid composition at each position of the linker sequence.

Table 2. Linker amino acid sequence.

Sequence	Count	Frequency (%)
GGGGSGGGGSGGGGS	12438	84.24
GGGGSGGGGS	1944	13.17
GGGGSGGGGGS	37	0.25
GGGDSGGGGSGGGGS	13	0.09
GGGGPGGGGSGGGGS	13	0.09
GGGSGGGGSGGGGS	13	0.09
GGGGSGGGDSGGGGS	9	0.06
GGGGSGGGGPGGGGS	9	0.06
GGGGSGGGGSGSGGS	8	0.05
GGGGSDGGGSGGGGS	7	0.05

Top 10 most frequent amino acid linker sequences are identified within the dataset. Each row corresponds to a unique linker sequence, ranked by its frequency of occurrence. The linker sequence selected as reference is highlighted in bold.

3.4. Germline gene usage

One of the advantages of long-read sequencing of scFvs is the ability to evaluate the paired VH-VL germline gene usage comprehensively, which is why we present the chord diagrams generated with R circlize package[43] in Figure 5. These figures provide chord diagram representations of IGHV and IGKV frequency and associations within phage display libraries (no IGLV genes were identified in the dataset). Panel A illustrates the overall number of clones across four panning rounds (0, 1, 2, and 3), where the thickness of the chords reflects the total count of clones associated with each IGHV-IGKV combination. In contrast, Panel B focuses on gene family associations, irrespective of clone counts, providing insight into the distinct relationships between IGHV and IGKV gene families. Each chord diagram offers a visual depiction of the distribution of clone counts and gene family associations, shedding light on the dynamics of antibody repertoire diversity within the phage display libraries across multiple panning rounds.

We compared our results with Nannini *et al.* and observed similar trends (Table 3). The percentage of unique heavy and light chain combinations decreased from 68.07% in P0 to 8.27% in P3 in our study, closely mirroring the published dataset which showed a decrease from 67.74% to 16%. The most frequent heavy-light chain combination in P0 was IGHV1–43-IGHJ3 paired with

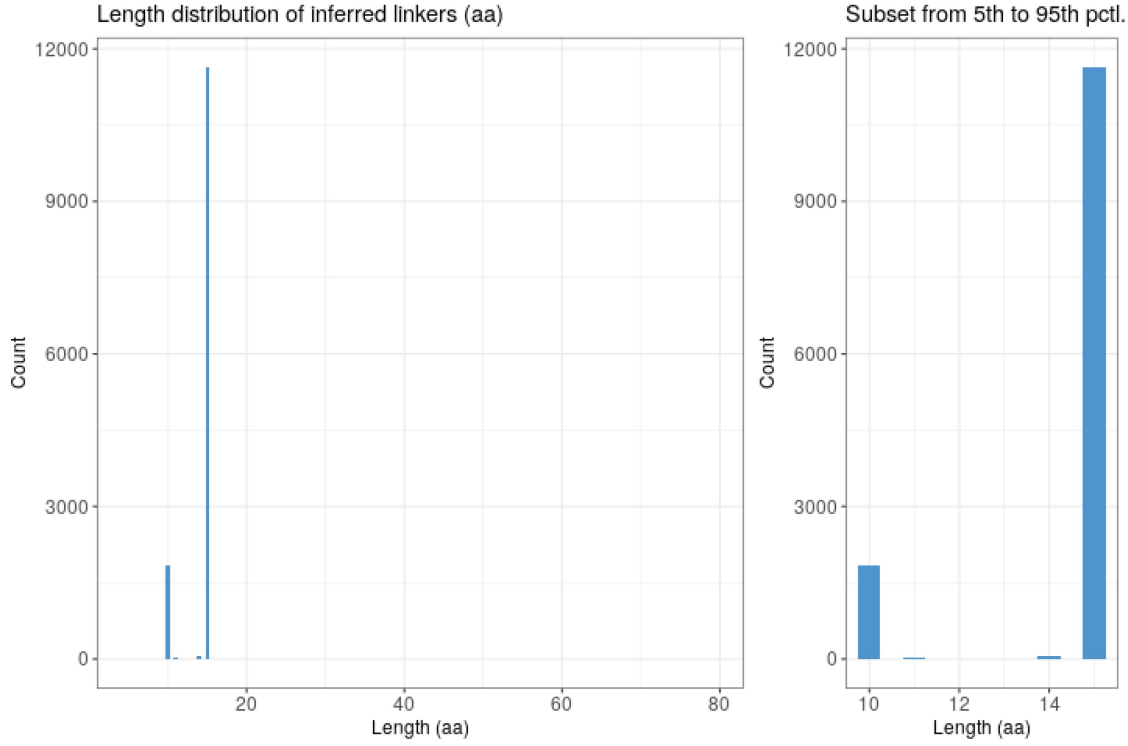

Figure 4. Distribution of inferred amino acid linker lengths within the dataset. Panel a illustrates the distribution across all detected linker lengths, while panel B provides a focused view of the distribution, presenting a histogram within the 5th and 95th percentile interval of linker lengths.

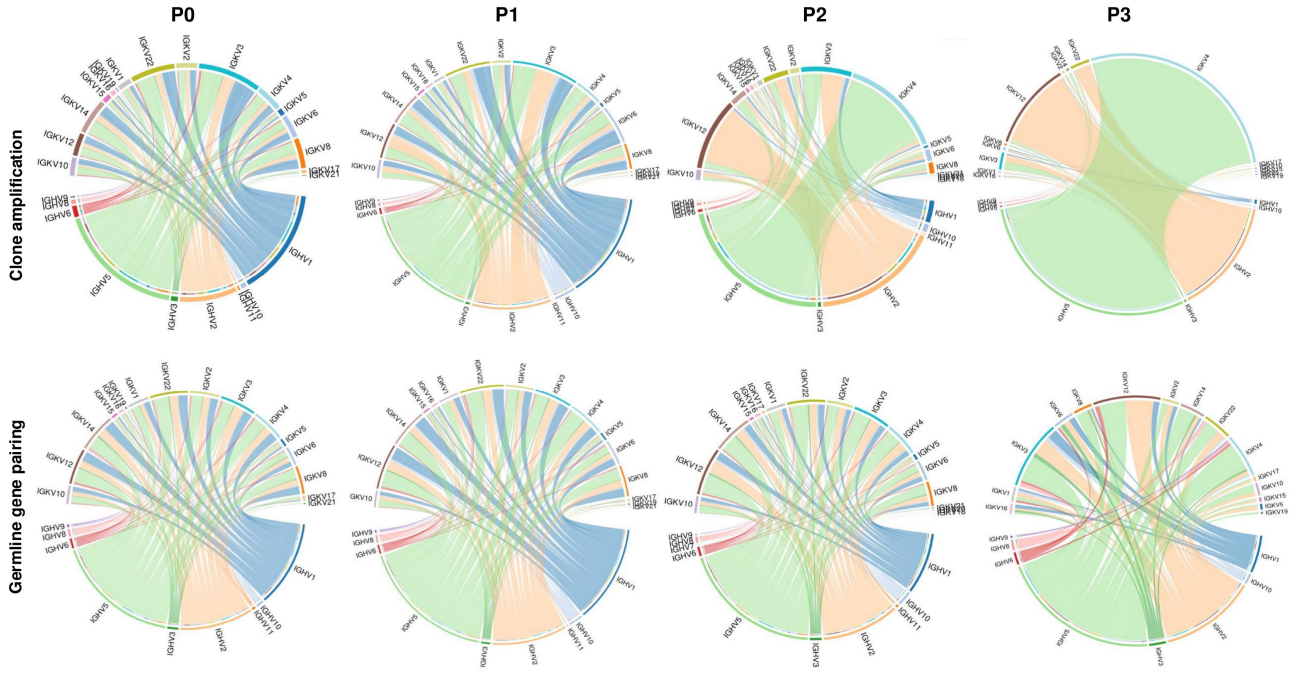

Figure 5. Chord diagram representation of IGHV and IGKV gene family associations.Associations.The top pannel depicts overall clone counts across four panning rounds, detailing total occurrences of IGHV-IGKV associations. The bottom pannel illustrates germline gene associations across the same panning rounds, independent of clone counts, showcasing the distinct associations between IGHV and IGKV gene families.

Table 3. Heavy and light chain pairing across panning rounds.

Panning round	Unique VH-VL (% of library)	Most frequent VH-VL		
		VH	VL	% of library
P0	68.07	IGHV1-43-IGHJ3	IGKV3S10-IGKJ5	0.55
P1	58.59	IGHV2-70-IGHJ3	IGKV3S10-IGKJ5	2.73
P2	20.81	IGHV5S13-IGHJ2	IGKV4S9-IGKJ2	20.58
P3	8.27	IGHV5S13-IGHJ2	IGKV4S9-IGKJ2	47.43

IGKV3S10-IGKJ5, representing 0.55% of the library, while in P3, IGHV5S13-IGHJ2 paired with IGKV4S9-IGKJ2 dominated at 47.43%. Similarly, the published dataset reported the most frequent combination in P0 as IGHV1–43-IGHJ3 with IGKV3S10-IGKJ5 at 0.36%, and in P3, IGHV5S13-IGHJ2 with IGKV4S9-IGKJ2 at 19.5%. These findings align with the published dataset, leading to the same general conclusion about the selective enrichment of specific scFv clones.

The differences between Seq2scFv results and those of Nannini *et al.* are expected given the variations in methodology. While Nannini *et al.* used IMGT/HighV-QUEST for their analysis, Seq2scFv employed IgBLAST. Despite utilizing the same germline gene database, the differences in algorithms can lead to variations in gene call assignments. Moreover, the scFv characterization tool in Seq2scFv applies stringent quality control measures: any sequence with poor alignment or containing a stop codons is discarded, regardless of whether these issues occur within the variable domains or the linker sequence. This rigorous approach ensures the high accuracy and reliability Seq2scFv annotations, resulting in a robust characterization of scFvs. Consequently, while the overall trends are consistent with those observed by Nannini et al., Seq2scFv's stringent criteria may account for the minor discrepancies observed between the two datasets.

This table summarizes the percentage of unique heavy and light chain combinations in the scFv library across the panning rounds (P0, P1, P2, P3) and identifies the most frequent heavy-light chain combinations observed at each round. The percentages represent the proportion of the library constituted by these combinations.

4. Discussion

NGS guided selection of lead antibody candidates generated by *in vitro* display methods has proven its great potential for accelerating the process of antibody discovery.[5–7] Moreover, with the progress in high-throughput long-read sequencing, it is now possible to interrogate display libraries at an unparalleled depth coupled with unprecedented levels of detail. Indeed, sequencing of full-length scFv provides crucial VH-VL pairing information, facilitating a comprehensive evaluation of each candidate's biophysical properties, germline gene usage, and structure.[3] While it's very unlikely to sequence 10,000 unique clones through clone picking alone, due to the overrepresentation of dominant clones,[31] NGS mining with Seq2scFv of phage display libraries generated by Nannini and colleagues (2020)[3] yielded more than 14,756 high-quality, fully annotated and validated unique scFv sequences. Notably, these libraries

were screened using a PacBio RS2 instrument, and new generation PacBio sequencers, such as the Sequel II, offer even greater accuracy and higher throughput, providing larger pools of scFvs to mine.

As the yield of full-length paired antibody data continues to increase, it is crucial for computational tools to keep pace and provide solutions for large-scale comprehensive characterization of scFvs. Seq2scFv addresses the challenges of characterizing both V-domains in a single read, by combining existing tools for the analysis of single domains (IgBLAST and AntPack) with custom scripts specifically developed to delimit and validate scFvs. Seq2scFv identified, characterized and validated 17,479 scFvs (representing 14,764 unique scFvs). The results, compiled in a single table, encapsulate the extensive characterization of single-chain variable fragments (scFvs) with a wide array of attributes across several categories. This includes identifiers (such as amino acid and nucleotide SEGUIDs), sequence information (paired VH-VL sequences, scFv and linker start and end positions, translation frames, linker annotations), linker sequence evaluation, quality flags (indicating the validity and productivity of the sequences), and specific annotations for heavy and light chains (covering gene loci, stop codons, frameshifts, and germline alignments). The inference of a consensus linker sequence, especially when the reference linker is unavailable, further completes the experimental picture and enhances the characterization process. Assuming the most frequent linker represents the linker sequence used in the experiment, the probability of randomly sampling this linker is higher. However, this process remains stochastic, and to balance accuracy with the computational demands of large datasets, sampling around 10%, or approximately 1,000–1,500 sequences, is recommended as an efficient threshold. Lastly, the table also reports key experimental data, including the counts of identical sequences across different panning rounds (P0, P1, P2, P3), enabling the tracking of selection dynamics.

Despite the significant loss of sequences during annotation and scFv validation (74%), it is crucial to emphasize that the retained sequences represent the highest quality subset of scFvs. Indeed, the validated scFv meet significant E-score thresholds for V(D)J alignment and do not present stop codons nor translation frame incongruences between the VH, the VL or the linker sequences. The principal cause of the sequence drop out, accounting for 91% of the discarded sequences, was the presence of stop codons. It is important to consider that the completeness of the germline gene database will impact the quality and quantity of hits obtained during annotation. This completeness will also aid in identifying the exact coordinates of the V(D)J regions with great accuracy, influencing both the delineation of scFvs and the inference of the linker from the data. Because of this, it is important to continually improve and refine these databases. This is also the reason why the Seq2scFv toolkit allows users to choose their own database or leverage public databases such as OGRDB[44] (available at https://ogrdb.airr-community.org/.), NCBI (https://ftp.ncbi.nih.gov/blast/executables/igblast/release/database/https://ftp.ncbi.nih.gov/blast/executables/igblast/release/database/.), or, albeit with licensing restrictions, IMGT ([42]; available at https://www.imgt.org/vquest/refseqh.htmlhttps://www.imgt.org/vquest/refseqh.html). Additionally, Seq2scFv does not remove the scFvs not meeting the quality standards but keeps them in a separate file, allowing users to review and revisit them.

Furthermore, Seq2scFv simplifies downstream analyses such as NGS-guided selection of lead antibody candidates based on frequency. By maintaining a catalog of identical sequences present in each of the sequenced panning rounds, users can track the progressive enrichment of target-binding sequences, which are expected to increase in frequency at later stages of panning. These counts are appended to the final table, enabling users to aggregate the counts at different levels of information. For example, it is possible to identify the enrichment of the scFv at the nucleotide or amino acid levels, or cluster on either the VH or VL, or even any of the CDR or FWR regions. Examples illustrating these capabilities are provided in the supplementary material. However, in the present article, demonstrating NGS-guided selection of antibody candidates is out of scope.

Seq2scFv's encoding of sequences also offers a useful solution for future inter-study comparisons. Input reads, nucleotide and amino acid scFvs, are all encoded with a SEGUID, which enables the efficient comparison and identification of identical sequences across panning rounds within the same experiment, or even across different experiments, databases and platforms. Overall, it is a robust feature for tracking and managing antibody data processed with Seq2scFv.

Apart from evaluating annotated scFvs, it is also of great interest to evaluate preprocessing statistics to consider the overall experiment quality. The preprocessing statistics acquired prior to Seq2scFv analyses highlight the number of reads discarded at various filtering stages, including quality, adapter removal, and size filtering, enabling researchers to assess the accurate construction of the libraries. Additionally, by examining read length distributions, researchers can identify sequences at unexpected lengths, uncovering potential issues during library preparation, amplification, or the presence of adapter sequences. Moreover, detecting an increase in the frequency of specific length intervals in later panning rounds implies an effective reduction in antibody diversity and a potential enhancement in specificity as selection progresses, validating the panning process. Sharing feedback with the laboratory based on these observations can enhance experimental practices and improve data quality in subsequent experiments.

In conclusion, Seq2scFv stands as a versatile and accessible toolkit for the analysis of full-length scFvs sequences obtained through long-read sequencing methods. Its open-source nature, free for academic use but subject to a commercial license for commercial purposes, ensures widespread accessibility while addressing proprietary concerns. Notably, the flexibility inherent in Seq2scFv tools enables researchers to customize key parameters such as germline gene databases, numbering schemes, and quality thresholds, thereby tailoring the analytical process to suit specific research objectives. In theory, Seq2scFv should be able to handle scFv sequences generated by any long-read sequencing platform, including newer generations of PacBio sequencers and Oxford Nanopore Technologies, as the framework is independent of pre-

processing and only requires paired VH-VL sequences in FASTA format. Although such data is not yet publicly available, future benchmarks may become possible as interest in the field grows, leading to more datasets and potentially demonstrating the broader applicability of Seq2scFv across diverse experimental setups. Lastly, the standalone functionality of Seq2scFv provides researchers with the reassurance of data security and privacy, eliminating the need to share sequences online and mitigating potential intellectual property issues. With its emphasis on rapid processing, robust analysis, and comprehensive annotation, we anticipate that Seq2scFv will accelerate the identification and characterization of lead antibody candidates with enhanced efficiency and accuracy.

Disclosure statement

All authors were employees of JSR Life Sciences at the time the toolkit was developed. The authors declare no other competing financial interests. However, the authors affirm that this potential conflict of interest has not influenced the scientific integrity or objectivity of the research presented in this manuscript.

Funding

The author(s) reported there is no funding associated with the work featured in this article.

ORCID

Marianne Bachmann Salvy http://orcid.org/0009-0006-1528-8168
Luca Santuari http://orcid.org/0000-0001-8784-2507
Emanuel Schmid-Siegert http://orcid.org/0000-0003-1339-5120
Nikolaos Lykoskoufis http://orcid.org/0000-0001-6751-7010
Ioannis Xenarios http://orcid.org/0000-0002-3413-6841
Bulak Arpat http://orcid.org/0000-0002-7749-3793

Author contributions

MBS developed the toolkit, wrote the manuscript and analyzed the publicly available dataset. LS provided valuable insights on the toolkit's annotation in relation to its use in downstream analysis. ESS and NL contributed to the processing of the reads used to develop the toolkit. ESS also adapted *basseto* to the specific requirements of the project. IX provided scientific insights crucial for the design and application of the toolkit. BA played a pivotal role in the development of the tools, provided a general overview of the project, and contributed valuable scientific insights throughout the process. Each author reviewed and approved the final version of the manuscript.

Data availability statement

Seq2scFv code and Docker file are available on GitHub (https://github.com/ngs-ai-org/seq2scfvhttps://github.com/ngs-ai-org/seq2scfv). A copy of the code, Docker file, annotated tables, figures and FASTA files are deposited in the Zenodo entry associated to this article https://zenodo.org/records/13747528https://zenodo.org/records/13747528.
Supplementary figures S1 and S2, as well as table S1, are available under the "Supplementary information 1" file, which also includes the code used to analyze the presented dataset and some post-processing examples. Together with Table S2 (in its own file, "Supplementary table S2"), all supplementary materials are available in the online version of the article.

References

1. Lu R-M, Hwang Y-C, Ju Liu I-, Lee C-C, Tsai H-Z, Li H-J, Wu H-C. Development of therapeutic antibodies for the treatment of diseases. J Biomed Sci. 2020;27(1):1–30. doi:10.1186/s12929-019-0592-z.
2. Kim J, McFee M, Fang Q, Abdin O, Kim PM. Computational and artificial intelligence-based methods for antibody development. Trends Pharmacol Sci. 2023;44(3):175–189. doi:10.1016/j.tips.2022.12.005.
3. Nannini F, Senicar L, Parekh F, Kong KJ, Kinna A, Bughda R, Sillibourne J, Hu X, Ma B, Bai Y, et al. Combining phage display with smrtbell next-generation sequencing for the rapid discovery of functional scfv fragments. MAbs. 2021;13(1):1864084. doi:10.1080/19420862.2020.1864084.
4. Fernández-Quintero ML, Ljungars A, Waibl F, Greiff V, Terje Andersen J, Gjølberg TT, Jenkins TP, Gunnar Voldborg B, Marie Grav L, Kumar S, et al. Assessing developability early in the discovery process for novel biologics. MAbs. 2023;15:2171248.
5. Barreto K, Maruthachalam BV, Hill W, Hogan D, Sutherland AR, Kusalik A, Fonge H, DeCoteau JF, Geyer CR. Next-generation sequencing-guided identification and reconstruction of antibody cdr combinations from phage selection outputs. Nucleic Acids Res. 2019;47(9):e50–e50. doi:10.1093/nar/gkz131.
6. Ravn U, Gueneau F, Baerlocher L, Osteras M, Desmurs M, Malinge P, Magistrelli G, Farinelli L, Kosco-Vilbois MH, Fischer N. By-passing in vitro screening—next generation sequencing technologies applied to antibody display and in silico candidate selection. Nucleic Acids Res. 2010;38(21):e193–e193. doi:10.1093/nar/gkq789.
7. Ravn U, Didelot G, Venet S, Ng K-T, Gueneau F, Rousseau F, Calloud S, Kosco-Vilbois M, Fischer N. Deep sequencing of phage display libraries to support antibody discovery. Methods. 2013;60(1):99–110. doi:10.1016/j.ymeth.2013.03.001.
8. Csepregi L, Ehling RA, Wagner B, Reddy ST. Immune literacy: reading, writing, and editing adaptive immunity. Iscience. 2020;23(9):101519. doi:10.1016/j.isci.2020.101519.
9. Daberdaku S, Ferrari C, Valencia A. Antibody interface prediction with 3d zernike descriptors and svm. Bioinformatics. 2019;35(11):1870–1876. doi:10.1093/bioinformatics/bty918.
10. Norman RA, Ambrosetti F, Bonvin AM, Colwell LJ, Kelm S, Kumar S, Krawczyk K. Computational approaches to therapeutic antibody design: established methods and emerging trends. Briefings Bioinf. 2020;21(5):1549–1567. doi:10.1093/bib/bbz095.
11. Smakaj E, Babrak L, Ohlin M, Shugay M, Briney B, Tosoni D, Galli C, Grobelsek V, D'Angelo I, Olson B, et al. Benchmarking immunoinformatic tools for the analysis of antibody repertoire sequences. Bioinformatics. 2020;36(6):1731–1739. doi:10.1093/bioinformatics/btz845.
12. Ye J, Ma N, Madden TL, Ostell JM. Igblast: an immunoglobulin variable domain sequence analysis tool. Nucleic Acids Res. 2013;41(W1):W34–W40. doi:10.1093/nar/gkt382.
13. Alamyar E, Giudicelli V, Duroux P, Lefranc M-P. Imgt/highv-quest: a high-throughput system and web portal for the analysis of rearranged nucleotide sequences of antigen receptors-high-throughput version of imgt/v-quest. Journées Ouvertes de Biologie, Informatique et Mathématiques. 2010;60. https://www.imgt.org/IMGTindex/IMGTHighV-QUEST.php.
14. Bolotin DA, Poslavsky S, Mitrophanov I, Shugay M, Mamedov IZ, Putintseva EV, Chudakov DM. Mixcr: software for comprehensive adaptive immunity profiling. Nat Methods. 2015;12(5):380–381. doi:10.1038/nmeth.3364.
15. Chothia C, Lesk AM. Canonical structures for the hypervariable regions of immunoglobulins. J Mol Biol. 1987;196(4):901–917. doi:10.1016/0022-2836(87)90412-8.
16. Abraham Kabat E. Sequences of proteins of immunological interest. Number 91 In National Institutes Of Health. US Department Of Health And Human Services, Public Health Service. 1991.

17. Lefranc M-P, Pommié C, Ruiz M, Giudicelli V, Foulquier E, Truong L, Thouvenin-Contet V, Lefranc G. Imgt unique numbering for immunoglobulin and t cell receptor variable domains and ig superfamily v-like domains. Dev & Comp Immunol. 2003;27 (1):55–77. doi:10.1016/S0145-305X(02)00039-3.
18. Dondelinger M, Filée P, Sauvage E, Quinting B, Muyldermans S, Galleni M, Vandevenne MS. Understanding the significance and implications of antibody numbering and antigen-binding surface/ residue definition. Front Immunol. 2018;9:2278. doi:10.3389/ fimmu.2018.02278.
19. Dunbar J, Deane CM. Anarci: antigen receptor numbering and receptor classification. Bioinformatics. 2016;32(2):298–300. doi:10. 1093/bioinformatics/btv552.
20. Li L, Chen S, Miao Z, Liu Y, Liu X, Xiao Z-X, Cao Y. Abrsa: a robust tool for antibody numbering. Protein Sci. 2019;28 (8):1524–1531. doi:10.1002/pro.3633.
21. Abhinandan KR, Martin AC. Analysis and improvements to kabat and structurally correct numbering of antibody variable domains. Mol Immunol. 2008;45(14):3832–3839. doi:10.1016/j.molimm. 2008.05.022.
22. Parkinson J, Wang W. For antibody sequence generative modeling, mixture models may be all you need. Bioinformatics. 2024;40 (5):btae278. doi:10.1093/bioinformatics/btae278.
23. Hemadou A, Giudicelli V, Smith ML, Lefranc M-P, Duroux P, Kossida S, Heiner C, Hepler NL, Kuijpers J, Groppi A, et al. Pacific biosciences sequencing and imgt/highv-quest analysis of full-length single chain fragment variable from an in vivo selected phage-display combinatorial library. Front Immunol. 2017;8:1796. doi:10.3389/fimmu.2017.01796.
24. Glanville J, D'Angelo S, Khan TA, Reddy ST, Naranjo L, Ferrara F, Bradbury AR. Deep sequencing in library selection projects: what insight does it bring? Curr Opin Struct Biol. 2015;33:146–160. doi:10.1016/j.sbi.2015.09.001.
25. Akbar R, Robert PA, Pavlović M, Jeliazkov JR, Snapkov I, Slabodkin A, Weber CR, Scheffer L, Miho E, Haff IH, et al. A compact vocabulary of paratope-epitope interactions enables predictability of antibody-antigen binding. Cell Rep. 2021;34 (11):108856. doi:10.1016/j.celrep.2021.108856.
26. Xu JL, Davis MM. Diversity in the cdr3 region of vh is sufficient for most antibody specificities. Immunity. 2000;13(1):37–45. doi:10. 1016/S1074-7613(00)00006-6.
27. Kunik V, Ashkenazi S, Ofran Y. Paratome: an online tool for systematic identification of antigen-binding regions in antibodies based on sequence or structure. Nucleic Acids Res. 2012;40(W1): W521–W524. doi:10.1093/nar/gks480.
28. D'Angelo S, Ferrara F, Naranjo L, Erasmus MF, Hraber P, Bradbury AR. Many routes to an antibody heavy-chain cdr3: necessary, yet insufficient, for specific binding. Front Immunol. 2018;9:336672. doi:10.3389/fimmu.2018.00395.
29. Guloglu B, Deane CM. Specific attributes of the vl domain influence both the structure and structural variability of cdr-h3 through steric effects. Front Immunol. 2023;14:1223802. doi:10.3389/ fimmu.2023.1223802.
30. Levin I, Štrajbl M, Fastman Y, Baran D, Twito S, Mioduser J, Keren A, Fischman S, Zhenin M, Nimrod G, et al. Accurate profiling of full-length fv in highly homologous antibody libraries using umi tagged short reads. Nucleic Acids Res. 2023;51(11):e61–e61. doi:10.1093/nar/gkad235.
31. Erasmus MF, Ferrara F, D'Angelo S, Spector L, Leal-Lopes C, Teixeira AA, Sørensen J, Nagpal S, Perea-Schmittle K, Choudhary A, et al. Insights into next generation sequencing guided antibody selection strategies. Sci Rep. 2023;13(1):18370. doi:10.1038/s41598-023-45538-w.
32. Giudicelli V, Duroux P, Kossida S, Lefranc M-P. Ig and tr single chain fragment variable (scfv) sequence analysis: a new advanced functionality of imgt/v-quest and imgt/highv-quest. Bmc Immunol. 2017;18(1):1–13. doi:10.1186/s12865-017-0218-8.
33. Mejias-Gomez O, Braghetto M, Sørensen MKD, Madsen AV, Guiu LS, Kristensen P, Pedersen LE, Goletz S. Deep mining of antibody phage-display selections using oxford nanopore technologies and dual unique molecular identifiers. New Biotechnol. 2024;80:56–68. doi:10.1016/j.nbt.2024.02.001.
34. Babnigg G, Giometti CS. A database of unique protein sequence identifiers for proteome studies. Proteomics. 2006;6 (16):4514–4522. doi:10.1002/pmic.200600032.
35. Shen W, Le S, Li Y, Hu F, Zou Q. Seqkit: a cross-platform and ultrafast toolkit for fasta/q file manipulation. PLOS ONE. 2016;11 (10):e0163962. doi:10.1371/journal.pone.0163962.
36. Vander Heiden JA, Marquez S, Marthandan N, Bukhari SAC, Busse CE, Corrie B, Hershberg U, Kleinstein SH, Matsen FA IV, Ralph DK, et al. Airr community standardized representations for annotated immune repertoires. Front Immunol. 2018;9:2206. doi:10.3389/fimmu.2018.02206.
37. Olsen TH, Boyles F, Deane CM. Observed antibody space: a diverse database of cleaned, annotated, and translated unpaired and paired antibody sequences. Protein Sci. 2022;31(1):141–146. doi:10.1002/pro.4205.
38. Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li W, Lopez R, McWilliam H, Remmert M, Söding J, et al. Fast, scalable generation of high-quality protein multiple sequence alignments using clustal omega. Mol Syst Biol. 2011;7(1):539. doi:10.1038/ msb.2011.75.
39. Tareen A, Kinney JB, Valencia A. Logomaker: beautiful sequence logos in python. Bioinformatics. 2020;36(7):2272–2274. doi:10. 1093/bioinformatics/btz921.
40. NGS AI Org. Basseto. 2024 [Accessed 2024 06 27].
41. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet J. 2011;17(1):10–12. doi:10.14806/ej.17.1.200.
42. Lefranc M-P, Lefranc G. The immunoglobulin factsbook. London, UK: Academic press; 2001.
43. Gu Z, Gu L, Eils R, Schlesner M, Brors B. Circlize implements and enhances circular visualization in r. Bioinformatics. 2014;30 (19):2811–2812. doi:10.1093/bioinformatics/btu393.
44. Lees W, Busse CE, Corcoran M, Ohlin M, Scheepers C, Matsen FA IV, Yaari G, Watson CT, Community AIRR, Collins A, et al. Ogrdb: a reference database of inferred immune receptor genes. Nucleic Acids Res. 2020;48(D1):D964–D970. doi:10.1093/nar/ gkz822.