PLOS | ONE

RESEARCH ARTICLE

Assessment of antibody library diversity through next generation sequencing and technical error compensation

Marco Fantini[1☯], Luca Pandolfini[1☯¤], Simonetta Lisi[1], Michele Chirichella[1], Ivan Arisi[2], Marco Terrigno[1], Martina Goracci[1], Federico Cremisi[1], Antonino Cattaneo[1,2]*

1 Bio@SNS Laboratory, Scuola Normale Superiore, Pisa, Italy, **2** European Brain Research Institute, Roma, Italy

☯ These authors contributed equally to this work.
¤ Current address: The Wellcome Trust/CRUK Gurdon Institute, University of Cambridge, Cambridge, United Kingdom
* antonino.cattaneo@sns.it

Check for updates

OPEN ACCESS

Citation: Fantini M, Pandolfini L, Lisi S, Chirichella M, Arisi I, Terrigno M, et al. (2017) Assessment of antibody library diversity through next generation sequencing and technical error compensation. PLoS ONE 12(5): e0177574. https://doi.org/10.1371/journal.pone.0177574

Editor: Hikmet Budak, Montana State University Bozeman, UNITED STATES

Received: November 10, 2016

Accepted: April 28, 2017

Published: May 15, 2017

Copyright: © 2017 Fantini et al. This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data Availability Statement: All relevant data are within the paper and its Supporting Information files.

Funding: Funded by European Union Seventh Framework Program [grant No. 604102 A.C.] (Human Brain Project). https://www.humanbrainproject.eu/. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

Abstract

Antibody libraries are important resources to derive antibodies to be used for a wide range of applications, from structural and functional studies to intracellular protein interference studies to developing new diagnostics and therapeutics. Whatever the goal, the key parameter for an antibody library is its complexity (also known as diversity), i.e. the number of distinct elements in the collection, which directly reflects the probability of finding in the library an antibody against a given antigen, of sufficiently high affinity. Quantitative evaluation of antibody library complexity and quality has been for a long time inadequately addressed, due to the high similarity and length of the sequences of the library. Complexity was usually inferred by the transformation efficiency and tested either by fingerprinting and/or sequencing of a few hundred random library elements. Inferring complexity from such a small sampling is, however, very rudimental and gives limited information about the real diversity, because complexity does not scale linearly with sample size. Next-generation sequencing (NGS) has opened new ways to tackle the antibody library complexity quality assessment. However, much remains to be done to fully exploit the potential of NGS for the quantitative analysis of antibody repertoires and to overcome current limitations. To obtain a more reliable antibody library complexity estimate here we show a new, PCR-free, NGS approach to sequence antibody libraries on Illumina platform, coupled to a new bioinformatic analysis and software (Diversity Estimator of Antibody Library, DEAL) that allows to reliably estimate the complexity, taking in consideration the sequencing error.

Introduction

Antibody repertoires have been used in conjunction with display or selection technologies [1–7] and many libraries and antibody formats were created to satisfy the high demand for the different applications of recombinant antibodies [6,8–17].

10x Genomics EX1030

The key parameter for an antibody library is its complexity [18] (also known as diversity), an estimate of the number of distinct elements in that collection. The amount of different functional species is directly related to the probability of that library to contain a functional antibody against a given antigen [19]. Despite the simplicity and the importance of this concept, until recently, measuring the diversity of antibody repertoires in a reliable and quantitative way was not possible and was approximated to the transformation efficiency of bacteria used to amplify the library [18,20,21]. To corroborate this estimate, so far the standard procedure in the literature consisted in testing the fingerprint pattern or the sequencing data of a few hundred library members for the presence of duplicates [14,22]. However, finding no identical clones in a random sample of a few hundred clones gives only a superficial evaluation of the library complexity and cannot be used to derive an estimate of the library complexity, which is expected to be from 10^4 to 10^6 times higher. The final complexity is then calculated by multiplying the estimated frequency of unique elements in the sample by the transformation efficiency. This calculation, albeit intuitive, is not correct since the complexity does not scale linearly with the number of elements in the fingerprinted sample. The probability of finding a duplicated element grows as a function of the number of elements analyzed [23]. Indeed, the 10000th element does not have the same probability to be unique as the 100th element.

Next generation sequencing has transformed functional genomics, and its application to the characterization of natural and synthetic antibody repertoires is growing [18,20,24]. The diversity parameter is however still hard to quantify precisely. The first widely used method in antibody library sequencing has been the Roche 454 pyrosequencing [18,25,26], that provides read lengths in the 300–400 bp range, suitable for antibody variable domains, but associated to a higher error rate (~0.5% per base [27,28]) and a lower throughput (10^4–10^5) [24] than other platforms. Other sequencing platforms, such as PacBio, can sequence up to 8500 bp but have a very low throughput (~10^4) and a very high error rate [20,29]. High error rate is not a critical issue when there is an appropriate coverage and a suitable genome reference that allows errors to be corrected. In the study of antibody repertoires a full coverage is not yet feasible [30] and a genome reference is lacking by definition, because antibodies undergo imperfect genomic V(D)J rearrangement [20]. Indeed, if a discrepancy is found when comparing the library sequences, it is impossible to discriminate whether it has a biological origin or it is due to errors occurring in the processing of the sample (technical error) [31]. This problem is well known in the literature and different groups use different methods to address this issue [20,31]. DeKosky [32] uses a 96% percentage similarity criterion in sequence clustering, while Glanville [25] requires the sequence to have at least 2 amino acid mutations in at least one chain to be considered truly unique. Recently new methods [33,34] based on unique molecular identifiers, barcodes used in alignment to correct errors in clusters, started to become a popular choice. These and others [24,30] consider as reliable only sequences found at least twice or thrice in the sequencing step. We believe that such criteria are too strict to define the complexity, because a great amount of the sequencing data, that include the naturally occurring genomic modification, are eliminated. Furthermore most of these methods are focused on CDR3 complexity disregarding the diversity originating from the other CDRs and framework, which is important for recognition, stability and folding [35]. Our approach is aimed to highlight mutations in any position of the coding region not focusing only on the most variable stretch of the antibody. Moreover those methods neglect the error rate information derived from the sequencing data that can be used to resolve the discrepancy found.

In order to obtain a more reliable complexity estimate, we set up a ligation-based Illumina sequencing strategy that, unlike previously described methods, is PCR-free, to avoid the PCR errors in the sample preparation. We then developed a software (DEAL (Diversity Estimator of Antibody Library)), which relies on base quality to solve the error rate problem. DEAL

allows to get more accurate complexity boundaries increasing the sequence pool for the inferential estimate of the library complexity. We believe that the PCR-free sequencing and DEAL analysis on the whole antibody coding sequence could establish new standards for a more reliable and quantitative estimate of antibody library complexity and define quality criteria for newly created antibody libraries.

Materials and methods

Construction of human SPLINT libraries

We constructed two SPLINT (Single Pot Library of Intracellular Antibodies) scFv antibody libraries from human RNA isolated from peripheral blood lymphocytes extracted from four anonymous voluntary donors which signed written informed consent (see Ethic statement below), following a protocol modified from Marks and Bradbury [16] (see Supporting Information, S1 Fig). IgM cDNA (μ heavy and κ and λ light chains) was used as a template to amplify VH and VL regions. Primers are designed to anneal to the external framework regions of the V genes. In the first library (hscFv1), each VH and VL subclass was first individually amplified, using every possible combination of the 5' and 3' primers available for VH and VL chains. The amplified products were then combined at equimolar ratios, so that each VH and VL subclass was equally represented in the library. In the hscFv2 library, instead, all VH and VL subclasses were amplified together in a single reaction (one for the heavy and one for the light V region) using a mix of the 5' and 3' primers available. A third single domain "nanobody" library (hVH) was created using only the VH domains, amplified in a single reaction. At the end of each library construction, the assembled VH-VL DNA products for hscFv1 and hscFv2, or the VH products for hVH, were ligated in the pLinker220 vector [36] for yeast expression in the SPLINT format, using restriction sites BssHII/NheI (NEB). Ligation of each library (~1μg) was transformed by electroporation into Max Efficiency E.coli DH5α cells (Invitrogen). Transformation efficiency for each library was assessed by colony count. See Supporting Information for details.

Ethic statement. The blood donors provided their written informed consent, and the samples were obtained from the Division of Transfusion Medicine and Transplant Biology, Pisana University Hospital, Pisa, Italy. The samples were received in an anonymous and completely de-identified format, and were not collected by ourselves. No request for approval of this study was presented to our Institutional Ethics Review Board.

Library DNA fingerprinting

100 individual bacterial colonies for each library transformation were picked and plasmid DNA extracted. Each scFv or VH present in the plasmid DNA was amplified by PCR and then digested with BstNI (NEB) at 37˚C for two hours. DNA fragments were resolved on a 8% acrylamide gel and stained with ethidium bromide. Pattern analysis was performed using Gel Quest software.

Real time PCR

Real time PCR was performed on the VH domain of the cDNA of each RNA sample used for library construction. All forward primers classes (BssHII-HuVHXaBACK, see Supporting Information, primers for VH) were tested against the most and the least abundant reverse primer class (HuJH4–5FOR and HuJH1–2FOR respectively). For each reaction, 4ng of cDNA sample were amplified following iTaq Universal SYBR Green Supermix (Bio-Rad) protocol.

Real time PCR was performed on a Rotor-Gene Q Platform and the result analyzed following Nordgård et al. 2006 [37].

Sequencing sample preparation

To attach sequencing adapters to the scFv sequences, a ligation-based approach was designed. DNA adapters were synthesized harbouring overhangs complementary to the cleavage product of the restriction enzymes used for excising the scFv fragment from the plasmid, namely BssHII and NheI.

The forward and reverse strands of the adapters are synthesized independently and annealed in vitro (1:1 ratio, 95˚C 5min, 95 → 25˚C in 5˚C steps 1min/step). Before annealing the reverse strand was phosphorylated (0.2nmol Oligos, 10U PNK (NEB) 37˚C 1h, 65˚C 20min) to allow the ligation.

The scFvs were excised from the library plasmid (~8μg of the library were digested 3h 37˚C with 4U of NheI (NEB), 3h 50˚C with 4U of BssHII (NEB)) and ligated to the adapters (forward adapter: scFv: reverse adapter in 10:1:10 ratio, ~200–250 ng library 400U T4 ligase (NEB), O/N 16˚C). The ligation was run on agarose gel and the band corresponding to the single insert with 5' and 3' adapters was resolved and purified with MinElute Gel Extraction Kit (Qiagen). An example of ligation efficiency is shown in Supporting Information, S2 Fig).

NGS library processing

Libraries were quantified by Qubit dsDNA HS Assay Kit (ThermoFisher Scientific), diluted to 4nM, denatured with 0.1N NaOH (5' at RT), neutralized and diluted again in buffer HT-1 (Illumina) to a final concentration of 12.5 pM. Equimolar denatured Phi-X Control V3 DNA (Illumina) was spiked-in (20% of the volume) as an internal quality control and to increase the sample diversity according to Illumina guidelines. Sequencing was performed on MiSeq system with Reagent Kit v3–600 cycles (Illumina), with a cycle number of 350 and 250 for forward and reverse reads respectively.

Raw data were demultiplexed from.*bcl* files into separate.*fastq* files with *bcl2fastq-1.8.4* (Illumina), using the following barcodes as indexes: i1 = TCAGCG, i2 = GATCAC, i3 = CTGAGA, i4 = AGCTTT. In order to take into account the different length of shifter sequences introduced with the sequencing adapters, a specific number of nucleotides was discarded from the start of the reads (R1 index i1 = 0, i2 = 1, i3 = 7, i4 = 8; R2 index i1 = 13, i2 = 12, i3 = 11, i4 = 10). Reads were purged from adapter dimers, quality-filtered (Phred Score $\geq$ 32) and trimmed in sequences of the same length (R1: 320bp; R2: 220bp) with *trimmomatic-0.32* [38]. All the sequences whose forward and reverse reads both survived from the previous step were selected, taking advantage of the perl script *fastq-remove-orphans.pl*, which is part of *fastq-factory* suite (https://github.com/phe-bioinformatics/fastq-factory).

Forward and reverse-complemented reverse reads were then combined into 540bp-long pseudo-reads with a custom python script, and pseudo-reads from the 4 different indexes were pooled together. The hVH nanobody llibrary reads were merged using PEAR [39], a pair-end read merger available at http://sco.h-its.org/exelixis/web/software/pear/.

Software development

DEAL was developed as a standalone C++ program, without the need of external libraries. To manage the amount of data, a x64 compiler must be used during compilation. Speed optimization (MSVC: /O2 or /Ox, GCC: -O2 or -O3) should also be applied during compilation, to overcome tail-end recursion and to avoid long computing times. The program is tested to compile under Microsoft Visual Studio environment and work in Microsoft Windows 64-bit

operating systems. DEAL was written in a modular structure to allow an easy future parallelization, even if not yet implemented. DEAL is available at http://laboratoriobiologia.sns.it/DEAL.

Small complementary python scripts were created for format conversion, reads joining and primer recognitions.

R scripts were used for graphs generation and statistical analysis.

Pattern recognition for fingerprint analysis is implemented as a small Excel VBA script.

Statistical analysis

The theoretical complexity of the hscFv1, hscFv2 and hVH nanobody libraries were estimated by the truncated Negative Binomial distribution (NBp,s, where p and s are the probability and size parameters) to fit number of sequences Nseq as a function of cluster cardinality x (see S3 Fig). Assuming Nseq(x) ~ C*NBp,s(x), it is possible to calculate the coefficient C, that represent the sought complexity to be estimated from the data (see details in Supporting Information).

Results

Library sequencing

Three sample libraries, 2 scFv libraries (hscFv1 and hscFv2) and a single domain library (hVH), were created from cDNA derived from human lymphocytes RNAs and amplified in bacteria (see Materials and Methods). The first two libraries reflected two different methods used to amplify V regions for scFv libraries construction, and were sequenced to find, if present, the advantage of one method over the other. The hVH instead was sequenced to calculate the complexity of a single domain library and to demonstrate the advantage of single domain sequencing (where the reads overlap).

Assuming that each transformed bacterium takes one copy of plasmid DNA, we can define the first hard cap of the library complexity as the number of total transformants obtained determined through CFU count. For the three libraries, hscFv1, hscFv2 and for the hVH nanobody library, this complexity upper bound is 15.8, 14.0 and 6.0 million elements respectively (Table 1). Standard fingerprint analysis on 100 random library clones for each library was performed and no duplicate was found (data not shown).

PCR-free sample preparation

In order to avoid any PCR-amplification step to attach sequencing adapters to the library to be sequenced which could insert mutations in the sequences, a ligation-based approach was

Table 1. Libraries complexity results.

Library	Total reads[a]	Maximal complexity[b]	DEAL complexity[c]	Minimal complexity[d]	Theoretical Complexity[e]
hscFv1	6.02×10^6	15.8×10^6	3.45×10^6	0.6×10^6	8.96×10^6
hscFv2	4.09×10^6	14.0×10^6	3.60×10^6	0.2×10^6	8.38×10^6
hVH	14.9×10^6	6.00×10^6	4.88×10^6	1.4×10^6	6.41×10^6

[a]Number of reads passing the first quality trimming analyzed by DEAL.

[b]Theoretical maximal complexity derived from the number of bacterial transformants.

[c]Number of DEAL clusters.

[d]Number of DEAL clusters minus the cluster with cardinality = 1.

[e]Complexity estimated by truncated negative binomial fit.

https://doi.org/10.1371/journal.pone.0177574.t001

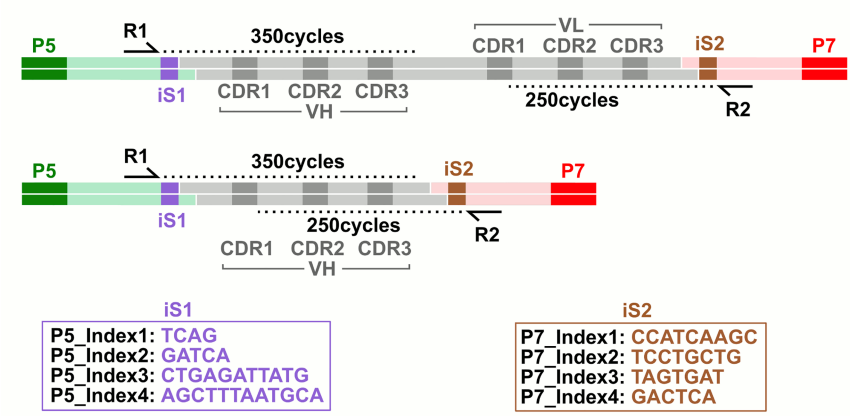


Fig 1. Diagram of sequenced adaptor-antibody-adaptor constructs. A) scFv library (gray), comprising heavy (VH) and light chain (VL) Complementary Determining Regions (CDR), was ligated to adapters (light green and pink) harbouring Illumina P5 and P7 flowcell hybridization sequences (green and red). B) VH nanobody library (gray), comprising heavy chain (VH) Complementary Determining Regions (CDR), was ligated to adapters (light green and pink) harbouring Illumina P5 and P7 flowcell hybridization sequences (green and red). The forward read (R1) uses SBS3 sequencing primer (Illumina), while the reverse read (R2) uses SBS12 primer (Illumina). iS1 and iS2 = index/shifter sequences.

https://doi.org/10.1371/journal.pone.0177574.g001

designed (see Materials and Methods). As shown in Fig 1, the forward read (R1) uses the P5 flowcell adapter (Illumina) and SBS3 sequencing primer (Illumina), while the reverse read (R2) uses the P7 adapter (Illumina) and SBS12 primer (Illumina).

Since sequences outside the Complementary Determining Regions (CDR), including the initial and distal framework regions, are rather constant, this poses serious technical limitations due to the sequencing by synthesis technology (SBS), which results in bad cluster recognition and low cluster count. To circumvent the problem, 4 barcoding index sequences (iS1 and iS2) were inserted just after each Illumina sequencing primer site. Using these barcodes allows to balance the read base content of the first 6 sequencing cycles. Moreover, inserting a different number of shifter bases before scFv fragment allowed to spread the eventual error in different positions of a sequence in the case of a error prone cycle, thus enabling correction during analysis.

Sequencing

Libraries underwent SBS sequencing on Illumina MiSeq, executing 350bp forward (R1) and 250bp reverse (R2) reads after ligation of the adapters to the scFv fragments, excised by restriction enzyme digestion from the library vector (Fig 1; see Materials and Methods). 10–15 million of raw reads were produced for each experiment. The presence of the index/shifter sequence (iS) in the adapters successfully balanced the base composition of the first sequencing cycles, which are critical for cluster detection and run parameter estimation (S4A–S4C Fig). Median Phred score (Q-score) is an empirical measure of the confidence of base identification and remained greater than or equal to 30 (99.9% base accuracy) beyond the 300th and 200th cycle for R1 and R2 respectively (S4A and S4B Fig). Phred score is an estimate of the probability of error for any given base. The error rate distribution for hscFv1 library is shown in Fig 2A. On the other hand, the presence of a Phi-X phage DNA spike-in, as an internal control, allowed to assess the impact of sequencing errors on the library sequence analysis, by comparing Phi-X reads to the reference sequence, getting a real measure of per tile and per cycle error rates (Fig 2B and S4D Fig). Notably both the scale and the shape of these two distributions differ, leading

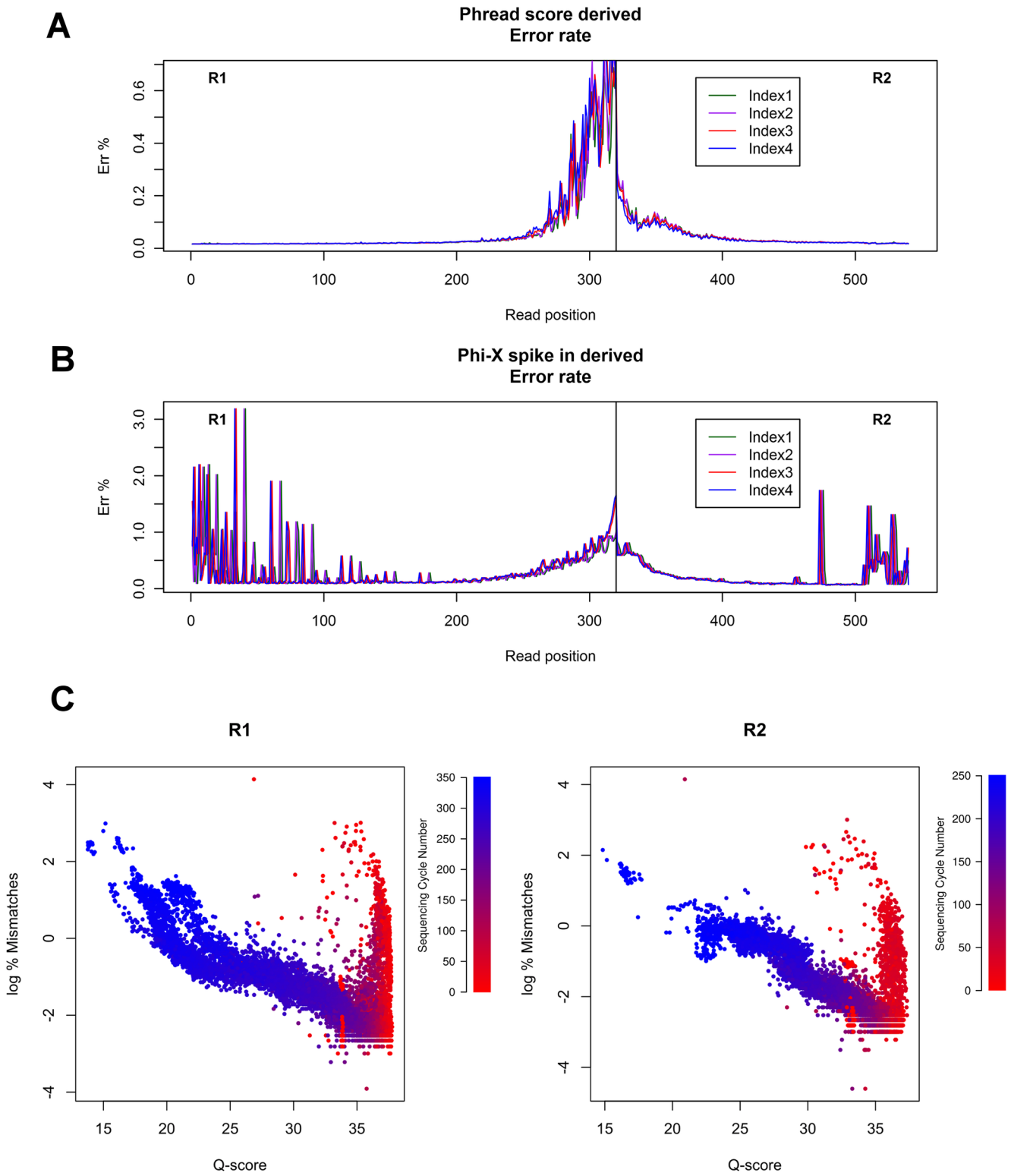


Fig 2. Phi-X derived and Phred score derived error rate distribution. A) Phred score error rate distribution for the hscFv1 library of the merged reads. Error rate increases with sequencing cycles. B) Control Phi-X derived error rate distribution for the hscFv1 library of the merged reads. Error rate is more prominent in the early sequencing cycles (spikes), with a small increase at the end of each read. The error distribution does not match the Phred score distribution and the shape differs as well. C) Scatter plot of the correlation of Q-score and $\log_2$(% Mismatches) in Phi-x control spike-in library. Each point represents the mean value from a single flow cell tile at a given sequencing read number, encoded by

colour (red to blue: R1 cycle 1 to 350; R2 cycle 1 to 250; colour flex point is set at cycle 38). The Q score in the first 40 reads fails to be predictive of mismatch rate. Similar results were obtained for hscFv2 and hVH libraries.

https://doi.org/10.1371/journal.pone.0177574.g002

to the conclusion that neither can be used alone for error rate estimation. While the median error rate was generally low (0.34±0.12%), both the presence of error rate "spikes" (Fig 2B and S4D Fig) and the lack of homogeneous correlation between Phred score and the measured error rate (Fig 2C) revealed that particular attention must be spent, to distinguish the real base variants from background technical noise. The Phred-score-derived error rate and the Phi-X control error rate were similar in all the three libraries analyzed (data not shown). This data analysis revealed that the sequencing run was successfully performed. However, intrinsic technical limitations of the NGS, namely the presence of two unrelated error estimates requires an additional analysis. To this purpose a new software (DEAL) was developed.

The Diversity Estimator of Antibody Library (DEAL) program

DEAL is a software tool to minimize the possible confusion between the real base sequence variants (biological diversity) from the background technical noise (technical misreading), taking into account both Phred derived error rate and Phi-X derived error rate.

DEAL is based on sequences identity collapse designed to ignore the error prone bases. The number of the collapsed sequences is an estimation of the complexity of the analyzed library.

DEAL is divided in two main steps: in the first step, the sequences are clustered by identity, using a "seed" of a 10–20 bp stretch in the CDR3s; in the second step, each element of the cluster is analyzed by binary comparison (Fig 3).

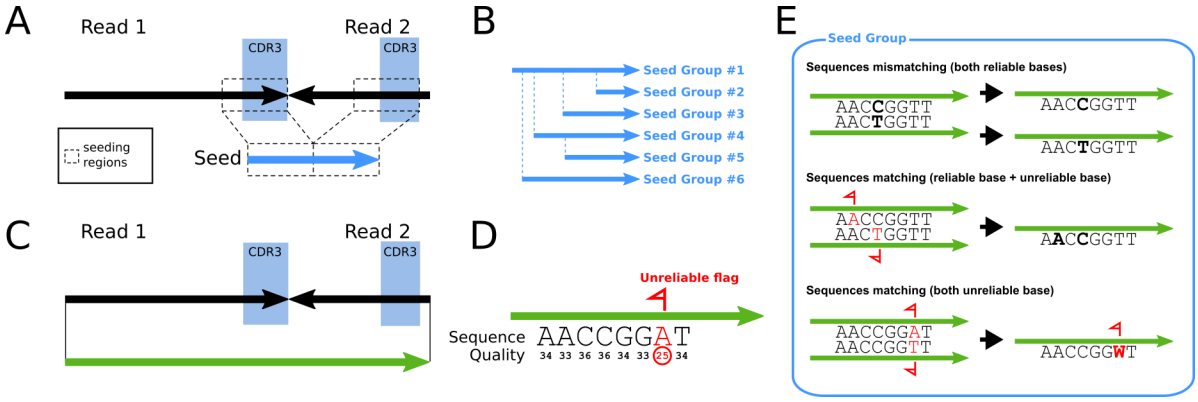


Fig 3. Diagram of DEAL workflow. A) Diagram of the seed creation process. In the figure, the black arrows represent the combined reads of the scFv library after the trimming. The seed is created combining the two seeding regions. The seeding regions are placed in the CDR3s to maximize the number of different seeds: the higher the number, the faster the program will run. B) Binary tree of the seeds. The program uses a binary tree approach to group identical seeds. During the comparison, if one sequence does not match any other sequences seen so far, a new branch of the tree is created in the mismatching position. C) The input of the binary comparison step. While the seeding step takes only into account the diversity of the seeding regions, the binary comparison analyzes the whole length of the combined reads. D) Flagging process. If some positions of the sequence are unreliable due to being associated to a low Phred quality score (as shown in the figure) or to a poor quality cycle (from Phi-X errors, not shown in the figure), the program flag them for correction. E) The three different scenarios that can occur during binary comparison among the sequences in the same seeding group. Mismatching (top): if two compared sequences differ in even only one position (bold) where none of the alternatives are flagged, the program recognize them as different sequences and does not group them. Matching sequences with a position having one flagged nucleotide (middle): the program recognizes the two sequences as identical and groups them together. All the positions where one of the sequences has a flag is resolved, during merging, as the not flagged nucleotide on the other sequence. Matching sequences with a position having both alternative nucleotides flagged (bottom): the program recognizes the two sequences as identical and groups them together. All the positions where both sequences have a flag are resolved using the IUPAC nucleobases ambiguity codes. The resulting merged sequence is flagged in that position.

https://doi.org/10.1371/journal.pone.0177574.g003

The first "seeding" step is necessary to reduce the calculation to local independent subproblems that can be solved by binary comparison. The seeding regions could be placed in any part of the sequence but since the CDR3s are the most variable regions, seeding in these regions would make the program run faster. The more the seed is variable, the bigger the number of seed groups become, lowering the mean number of sequences per group and thus further reducing the computational time required for the next step.

Moreover, this analysis is based on a binary tree approach which has the great advantage of being time and memory saving. The program starts by creating a small sequence stretch (seed) from the seeding region provided (Fig 3A) building a tree of identical sequences. While comparing the seeds, in the presence of a mismatch, a new branch is created (Fig 3B). The first step ends when all sequences are analyzed and all the identical seeds form individual clusters. The number of these clusters is the seed complexity.

In the second step, the sequences within each seed cluster are compared along the whole sequence length (Fig 3C). The process is a binary comparison, so each sequence is compared to every previous analyzed sequence in its seed group. In this step, the possibility of sequencing errors is taken into account. The program flags uncertain base read positions as unreliable, by checking both for a low Phred quality score associated to the base considered (Fig 3D) and for a high error rate in the sequencing cycles, that is retrieved from the error rate of control phage DNA (Phi-X). The two flagging descriptors (Phred quality and cycle quality) relate to two different checkpoints. If one or both the quality checks are not passed, the base is considered unreliable. Fine tuning can be done on the thresholds of these checks which can be defined as command line arguments. At this point, three scenarios (Fig 3E) can occur: i) if the sequences do not match, two subgroups are created; ii) if two sequences match and only in one an unreliable-flagged base is present, the base in that position will be assigned as the other sequence's reliable base, and the two sequences are then merged and resolved as the same one; iii) if in two matched sequences at a specific position both bases are unreliable-flagged the resulting base in the merged sequence will be an unreliable-flagged base. In this case it is possible that two different bases have to be merged. To keep track of all the possibilities in unreliable positions of merged reads we applied the IUPAC ambiguity code for nucleobases to the stored sequences.

The result of the computation is the creation of many small groups of matching sequences, whose number represents the sample complexity.

Information about the grouping steps of an analysis, as well as the final group distribution and all the sequences associated to the groups is also available using DEAL. It is possible to customize some parameters such as the seed positions and the optional computation of complexity of the deduced aminoacid sequence. In conclusion, DEAL software is able to ignore error prone bases during sequence identity collapse, leading to a reliable estimate of the library complexity.

Upper and lower limit estimate of library complexity and outlier determination

The complexity of the three libraries was analyzed by NGS followed by DEAL. Data are summarized in Table 1. The sequencing cluster count (12.27, 10.98 and 15.52 million for human scFv libraries 1 and 2 and VH nanobody library respectively) was of the same order of magnitude as the upper cap complexity of the libraries, determined by transformants count (15.8, 14.0 and 6.0 million respectively). To be noticed, the VH nanobody library sequencing cluster count exceeded almost 3 times the upper cap limit, indicating an almost complete coverage of the library. Since it is crucial to have good quality sequencing data, the reads underwent a very

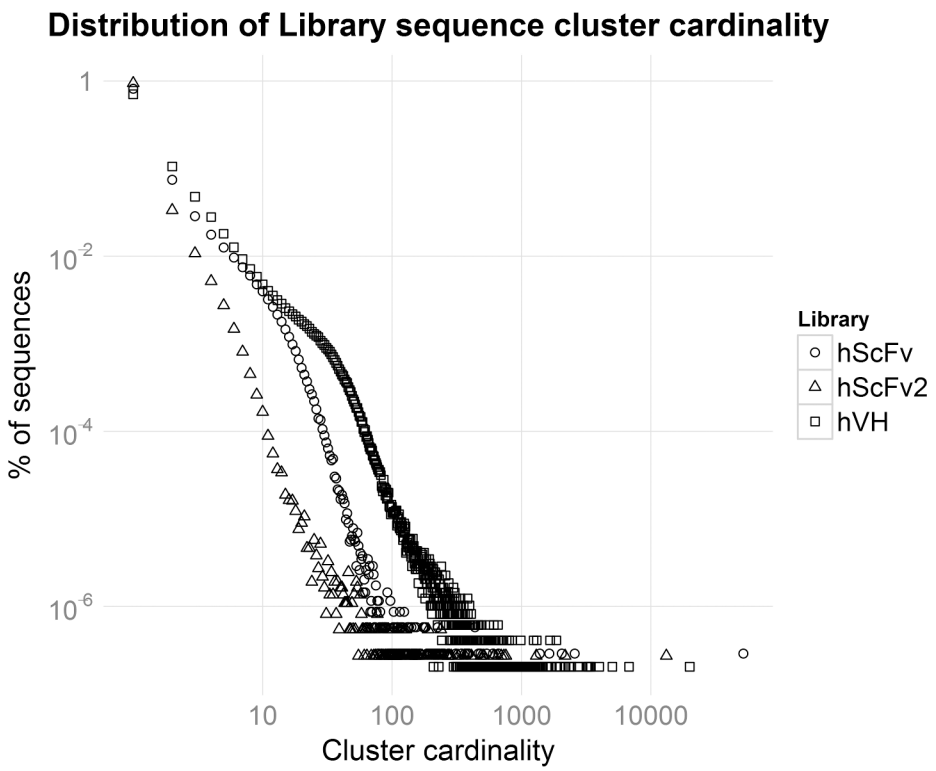


Fig 4. Distribution of library sequence cluster cardinality. Distribution of library sequence cluster cardinality. The more the curve is skewed towards high cardinality clusters, the lower the complexity of the library is expected to be.

https://doi.org/10.1371/journal.pone.0177574.g004

strict quality trimming process. Only the reads which had a median Phred score of at least 32 (base call accuracy > 99.937%) survived the filter. The sequence count after trimming was 6.02 and 4.09 million for hscFv1 and hscFv2 respectively. A higher trimming survival count (14.9 million) was obtained for the hVH nanobody library, due to both the shorter length and the overlap of the two reads, which improves the median quality of the sequences.

DEAL was then applied to the quality trimmed data, using for the 2 human scFv libraries the default parameters: i) unreliable flag set when either error threshold for the Phi-X in the position was over 1% or quality in position is less than Phred 32; ii) seed position was located in the CDR3 of both VH and VL fragment (position 280–300 for read 1 and 470–490 for read 2). For the VH nanobody library, DEAL was set to allow a variable length in the input, the seed was placed in the 280–300 region only. Moreover the VH nanobody library protein complexity was also calculated.

The resulting distribution of clusters by cardinality is shown in Fig 4. A crowding in the first dozen groups can be seen, meaning that clusters with few elements are a clear majority. At the other end of the cardinality plot a clear single high cardinality element can be observed. This outlier element was present in all the analyzed libraries and was shown to originate from the backbone of the library plasmid vector, due to an incomplete digestion of the parental vector in the library construction process. In general the number of outliers is indicative of problems occurred during library construction, due for example to an overrepresentation of few specific clones. Indeed a high number of outliers would be indicative of an unbalance in the library lowering the chances of successful selections.

Another kind of outlier is represented by the single element cluster (group with cardinality 1). This group contains, in addition to the real biological singletons (sequences with no other sequence in the same cluster), all the sequences where errors were not associated with a quality drop (and thus unable to be corrected by DEAL), that are unlikely to cluster with any other sequence in the sequencing run. For these reasons, the single element group is not reliable; therefore it should not be considered in complexity modeling to avoid an overestimation of the complexity. The number of DEAL clusters with cardinality two or greater represent the insurmountable lower limit of the library complexity (Table 1). The three libraries show respectively 0.6, 0.2 and 1.4 million unique clusters that satisfy the requisite, defining their lower limit complexity. However, if the coverage is not sufficient (i.e. for hscFv1 and hscFv2), raw data can not directly provide a measure of the total complexity. Thus a theoretical complexity calculation is also needed. To this purpose, an estimate of the theoretical complexity for each library was obtained using the truncated Negative Binomial distribution (Table 1 and S3 Fig).

Library analysis: Primer class distribution and VH-VL independent assortment validation

From the sequencing data, the chain independent assortment is another useful parameter to measure the skewness of a library. The NGS sequences were parsed with a custom python script for the primers used in the construction of the libraries. The distribution of the libraries by primer class is shown in Fig 5. A large portion of these are unclassifiable, probably due to the error spikes present in the primer regions, that impair the primer recognition, as discussed above (Fig 2B). The independent assortment of the VH-VL chains of the scFv libraries is shown in Fig 5A and 5B. No difference is observed between the frequency of forward and reverse primer pairs, compared to the theoretical combinatorial model, showing a good combinatorial assortment. Moreover, the frequencies of primers pairs assortment of the VH nanobody library matches the theoretical VDJ combinatorial model showing that very little amplification bias occurred during library construction (Fig 5C).

The primers distribution, for all the three libraries, showed a very low percentage of reads for the class corresponding to the forward primer VH5 (HuVH5aBACK). To verify whether this reflected a problem related to NGS or to the real biological distribution, Real-Time PCR on the libraries and their corresponding cDNA was performed. Results showed that the same class distribution and abundance was present in the original cDNA, showing that the distribution bias of the VH5 primer class was not due to a sequencing issue (S5 Fig).

Usually, two different methods are used to amplify V regions for scFv libraries construction. In one, each VH and VL subclass is first amplified individually and then the products are combined at equimolar ratios [40,41]. Alternatively, all the VH and VL subclasses are amplified together in a single reaction (one for each type of V region) [16].

Therefore we compared the deduced complexity of two human scFv libraries, constructed by the two methods. To verify this issue, we constructed two human scFv libraries, using both protocols. The primer distribution obtained from NGS clearly reflects the two different approaches in library construction. hscFv1 shows a more poised distribution of all the V subclasses compared to hscFv2, in which, instead, there is a clear imbalance in classes representation (Fig 5). However, the Real Time data indicate that the protocol with the single common amplification reaction generates a library that faithfully reproduces the real distribution present in the natural repertoire. Thus, the first method guarantees an equal representation of all the different V subclasses, although it might not reflect the natural distribution.

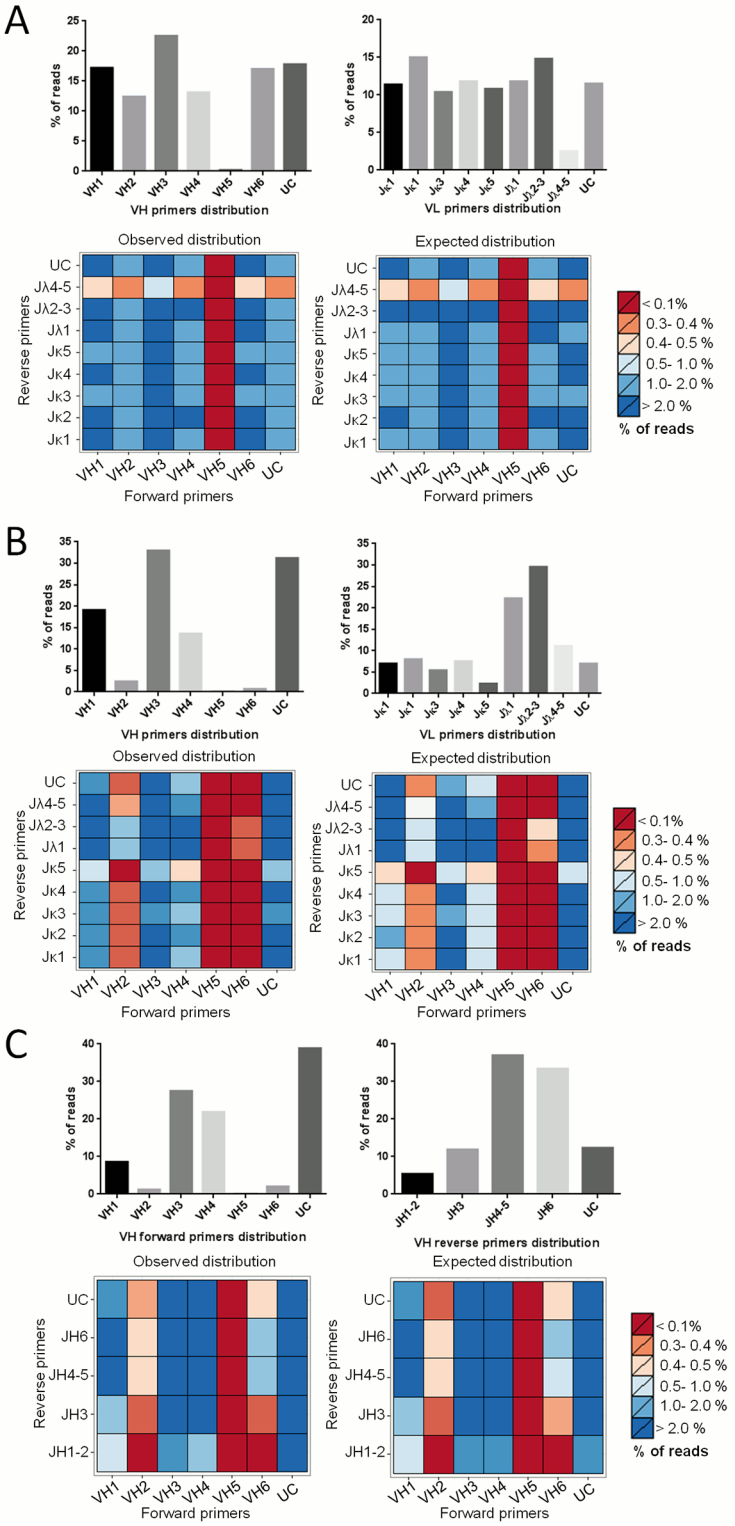


Fig 5. Chain/VDJ assortment independence of libraries. A) hscFv1. B) hscFv2. C) hVH. Top panels: barplots of forward and reverse primer distributions. Bottom panels: heatmaps of library primers distributions. Observed distribution is the primer pair proportion found after sequencing. Expected distribution is the multiplication of the two primers proportion (expected distribution given the independence between chains for the scFv libraries or given a balanced VDJ recombination for hVH). UC = unclassified. This category includes

all the sequences that do not match any primer. The name of the primers is a shorter version of the original name listed in Supporting Information (Primer used for library construction).

https://doi.org/10.1371/journal.pone.0177574.g005

Determination of the protein diversity of a VH library through in silico translation of NGS reads

The quality of an antibody domain library is ultimately determined not only by its nucleic acid sequence complexity but also by the proportion of the sequences in the library that code for full length antibody domain protein. As the quality of a library resides in the protein conformational diversity, its most relevant estimator is protein complexity, which represents the true and relevant complexity of a library. While for scFv libraries the NGS reads do not yet allow to unambiguously deduce the full protein coding sequence by in silico translation, due to sequence length limitations, the shorter VH length allows the full coverage of the entire VH sequences, since R1 and R2 reads overlap in the central DNA stretch. Considering that the central region is read twice, if a sequencing error occurs at a given position of the forward read, it can be corrected comparing the base at the same position of the reverse read and vice versa. Thus, assigning that position to the base with the best quality between the discording pair, this correction lowers the global error rate and guarantees the best quality read in the critical CDRs regions and allows the correct frame to be determined, for amino acid sequence deduction. This is still not possible for scFv libraries, due to the undefined number of nucleotides present in the unsequenced gap between the VH and VL.

The protein diversity of the VH nanobody library was determined by in silico translation of DEAL clusters, where all synonym sequences were aggregated and both the frameshift and nonsense mutation were filtered out. The length distribution of the sequenced VH nanobodies shows that the correct frame is maintained in 86.57% of the analyzed sequences (Table 2, Fig 6). Since, in Illumina sequencing, the proportion of indel mutations is negligible, the frameshifts are mostly due to non-functional VDJ recombination. The majority of the sequences are around 370 nucleotides in length and the in frame sequences are the most abundant. The minority of out of frame (+1 or +2 base pair) sequences shows an identical bell shaped distribution. Sequences with stop codons are also very rare: 10.3% considering the univocal stop codons in error-free sequences, 16% considering possible stop codons in error-flagged sequences. Indeed, filtered data shows that 80% of the hVH library sequences code for functional nanobodies (Table 2). The total protein diversity derived by DEAL clusters count is 4.73 million, which is reduced to 3.2 million by filtering the nonsense and frameshift incomplete peptides. Therefore, ignoring the protein group of cardinality 1 (due to presence of possible technical errors), the lower bound of the library complexity is 1.43 million proteins, of which 1.2 million are functional.

Discussion

A quantitative evaluation of the sequence complexity of antibody libraries is critical to verify their quality and eligibility for screening purposes. Next generation sequencing is dramatically changing how antibody quality and complexity are estimated. We present here a PCR-free NGS sequencing strategy associated with the DEAL software, to compensate for technical errors, which allows to determine the minimal functional complexity of an antibody library. The PCR-free library preparation avoids PCR-induced errors, sacrificing the convenience to selectively amplify the regions of interest in such manner as to minimize sequencing error on those key regions albeit losing all the information on the other areas of the antibody. The main advantage of the proposed method over current approaches is the definition of a "unique" biologically relevant sequence, filtering out as best as possible technical errors from

Table 2. hVH library protein complexity.

Total reads[a]	% in frame sequences[b]	% sequences without stop codon in CDS[c]	% sequences with full length CDS[d]	Total DEAL protein complexity[e]	Total DEAL functional complexity[f]	Minimal DEAL protein complexity[g]	Minimal DEAL functional complexity[h]
14.9×10^6	86.57%	89.7–84.0%	83.4–77.9%	4.73×10^6	3.25–3.19 $\times 10^6$	1.43×10^6	1.25–1.20 $\times 10^6$

[a]Number of reads passing the first quality trimming analyzed by DEAL.
[b]Percentage of functional in frame sequences.
[c]Percentage of sequences which do not contain stop codons or do not contain possible stop codon in a error flagged sequence.
[d]Percentage of sequences in frame and without stop or possible stop codons.
[e]Protein complexity from DEAL protein clustering.
[f]Number of DEAL protein clusters with full length CDS.
[g]DEAL protein complexity minus the protein cluster with cardinality = 1.
[h]DEAL protein complexity with full length CDS minus the protein cluster with cardinality = 1.

https://doi.org/10.1371/journal.pone.0177574.t002

the sequencing data. In fact, we found that Phred quality score generated by the sequencer alone is not sufficient to identify an error. Indeed, when comparing Phred quality to the error rate data extracted from the control phage DNA, it is clear that both parameters have to be taken into consideration.

In fact, Phred score is a pinpoint (both single sequence and single base) quality assessment, while the control phage DNA error rate can only be a per tile sequence quality assessment. Thus, it is not possible to substitute the Phred score with the error rate and both are required for the analysis.

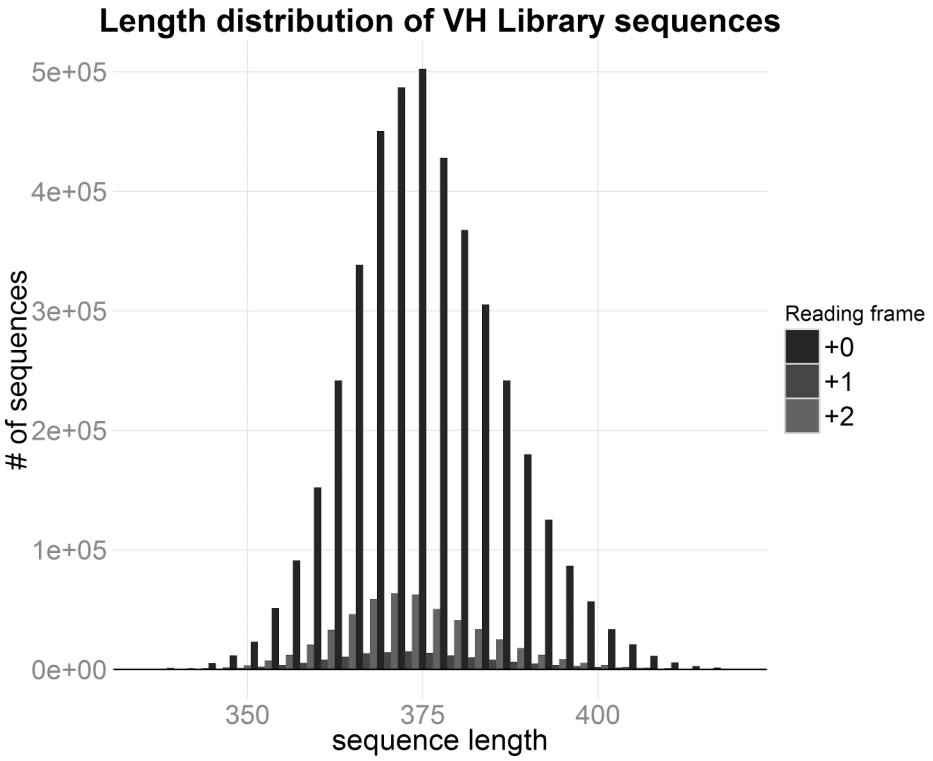


Fig 6. Length distribution of human VH nanobody library sequences. Barplot of the length distribution of human VH nanobody library sequences coloured by reading frame.

https://doi.org/10.1371/journal.pone.0177574.g006

In addition, to define the origin of an error we need to discriminate between pre-sequencing and sequencing errors. While sequencing errors need to be identified and removed (or limited), dealing with pre-sequencing errors is more complicated. In particular, some pre-sequencing errors are introduced by the nucleic acid handling step during library construction (cDNA, V regions amplification). These errors are acceptable, sometimes deliberate [42], since they improve the diversity of the library. Instead, mutations introduced during the sequencing sample preparation (mainly PCR-derived) result in a misinterpretation of the sequences present in the library and thus have to be avoided. Therefore a PCR-free approach in the sequencing sample preparation is certainly to be preferred, since it reduces the number of errors (not associated to a quality drop) that DEAL cannot resolve. It should be noted that this method could also be used to calculate the diversity of the natural repertoire from biological samples, such as immune or cancer cells, however libraries from such samples have to be amplified by PCR in order to obtain the quantity needed for the ligation.

The deduced amino acid sequence of the VH single domain library showed an unexpected length distribution, with more than 10.3% of non-functional out of frame sequences. Because in MiSeq indels are very rare, the length is most likely a natural feature rather than a sequencing artifact. As of today, the molecular mechanism whereby lymphocytes discard nonsense or out of frame variable domains and assemble only functional chains remains somewhat obscure. Interestingly, we observed in the hVH library a 7:1 ratio between functional and non-functional chains, which reflect the IgM RNA composition of PBLs. Indeed, this observation is in accordance with previous reports in mouse models [43,44]. In addition, the observed ratio appears to be remarkably high, even hypothesizing DNA silencing mechanism or RNA nonsense mediated decay taking place after transcription. In fact, if all the non-functional sequences were translated, there would be a considerable amount of useless transcripts and protein products. Moreover since this ratio refers only to a single chain, and an assembled antibody must have four correct chains to be functional, the ratio of functional assembled antibody would be even lower. Thus, it is clear that some kind of mechanism prevents the translation of these non-functional RNAs in vivo. Regarding the implications for antibody library screening, assuming the worst-case scenario where the light chains have the same functional proportion of the heavy chains, we could tentatively conclude that more than 80% of single chain library elements and more than 64% of scFv library elements encode a correct protein product.

Concluding, besides deep sequencing, the only other available method to determine the quality of an antibody library is a functional assay: the actual selection against an antigen. This is a valid practical strategy, although it is time consuming and can only give information on whether a library is adequate for screening purposes. The best general prediction of the capability of a library to undergo successful screenings is the estimate of its complexity in the most accurate possible way. Indeed, the complexity of a library can be directly linked to the probability to find a given binder of adequate affinity [19]. The method presented provides an advance towards this goal, by eliminating PCR steps and by compensating the technical errors. This method was validated for SPLINT intrabody libraries, but it is readily extendable to any antibody library of standard size. Moreover, DEAL software is independent of the sequencing platform and can be even more reliable with the greater number and more precise sequences that will be hopefully available in the future with the advancement of NGS technology.

Supporting information

S1 Fig. Diagram of human scFv library construction. Step1: Amplification of VHs, Vκs and Vλs from human cDNA. Step2: Construction of a linker $(G4S)_3$ with 5' specific for VHs, 3' blunt, and 2 linkers $(G4S)_3$ with 3' specific for Vκs and Vλs, 5' blunt. Step3: construction of

VHs, Vκs and Vλs "blocks" with $(G4S)_3$ linkers. Step4: Pullthrough of "blocks" and insertion of restriction sites for BssHII at the 5' end, and NheI at the 3' end. Step5: Ligation of BssHII/NheI digested pullthroughs to vector pLinker220.
(TIF)

S2 Fig. scFvs-Adapter ligation. Ligation products of hscFv2 with the four Illumina adaptors were resolved on a 0.7% agarose gel. Lane 1: the BssHII/NheI digested hscFv2 library (input of each of the four ligation) band at ~700-800bp. Lane 2: ligation of hscFv2 library with Index 1 adaptors (the correct ligation product is the band at PM~800-900bp). Lane 3: ligation of hscFv2 library with Index 2 adaptors. Lane 4: ligation of hscFv2 library with Index 3 adaptors. Lane 5: ligation of hscFv2 library with Index 2 adaptors. M = 100bp molecular marker.
(TIF)

S3 Fig. Fit of distribution of library sequence cluster cardinality. Distribution of library sequence cluster cardinality and regression curve. The three library A) hscFv1, B) hscFv2 and C) hVH (in black) are plotted with the corresponding Negative Binomial regression fit (in red).
(TIF)

S4 Fig. Quality score in hscFv1 library sequencing run. A) Upper panel: box and whisker plot of R1 Phred quality score per sequencing cycle. Median Phred score remained greater than 30 beyond the 300th cycle. Bottom panel: base composition per cycle. The first dozen bases, critical for cluster detection, are balanced due to index presence. B) Upper panel: box and whisker plot of R2 Phred quality score per sequencing cycle. Median Phred score remained greater than 30 beyond the 200th cycle. Bottom panel: base composition per cycle. The first dozen bases are balanced due to index presence. C) Upper panel: box and whisker plot of joined R1-R2 after index and end trimming Phred quality score of hscFv1. Median Phred score remained greater than 30 in all considered position. Bottom panel: base composition in the considered positions. After index trimming the first and last hundreds bases appear well conserved (belonging to the constant region of variable fragment). D) Phi-X technical error rate per sequencing cycle. Green represent the region after trimming. Upper panel: barplot of the mean %mismatches among sequencing tiles. Bottom panel: box and whisker plot of %mismatches. Error rate is more prominent in the beginning sequencing cycles (spikes), with a small increase at the end of each read. Similar results were obtained for hscFv2 and hVH libraries.
(TIF)

S5 Fig. Real time PCR for VH primers distribution on cDNA used for library construction. Relative expression using RHuJH4-5 as reverse primer, values were normalized on the maximum (VH3), N = 4, errors are expressed as SEM. Similar results were obtained using other reverse primers and different batches of cDNA.
(TIF)

S1 File. Supporting information. Detailed protocol for library construction and statistical analysis.
(DOCX)

Author Contributions

Conceptualization: AC MF LP MC.

Data curation: MF LP IA.

Formal analysis: MF LP IA.

Funding acquisition: AC.

Investigation: MF LP SL MC MT.

Methodology: LP MF MC.

Project administration: AC.

Resources: SL FC AC.

Software: MF.

Supervision: AC.

Validation: SL MG MT.

Visualization: MF LP SL.

Writing – original draft: MF LP SL AC.

Writing – review & editing: MF LP SL FC AC.

References

1. Winter G, Milstein C. Man-made antibodies. Nature. 1991; 349: 293–299. https://doi.org/10.1038/349293a0 PMID: 1987490
2. Marks JD, Hoogenboom HR, Bonnert TP, McCafferty J, Griffiths AD, Winter G. By-passing immunization. Human antibodies from V-gene libraries displayed on phage. J Mol Biol. 1991; 222: 581–597. Available: http://www.ncbi.nlm.nih.gov/pubmed/1748994 PMID: 1748994
3. Hanes J, Pluckthun A. In vitro selection and evolution of functional proteins by using ribosome display. Proc Natl Acad Sci U S A. 1997; 94: 4937–4942. Available: http://www.ncbi.nlm.nih.gov/pubmed/9144168 PMID: 9144168
4. He M, Taussig MJ. Antibody-ribosome-mRNA (ARM) complexes as efficient selection particles for in vitro display and evolution of antibody combining sites. Nucleic Acids Res. 1997; 25: 5132–5134. Available: http://www.ncbi.nlm.nih.gov/pubmed/9396828 PMID: 9396828
5. Visintin M, Tse E, Axelson H, Rabbitts TH, Cattaneo A. Selection of antibodies for intracellular function using a two-hybrid in vivo system. Proc Natl Acad Sci U S A. 1999; 96: 11723–11728. Available: http://www.ncbi.nlm.nih.gov/pubmed/10518517 PMID: 10518517
6. Visintin M, Meli GA, Cannistraci I, Cattaneo A. Intracellular antibodies for proteomics. J Immunol Methods. 2004; 290: 135–153. https://doi.org/10.1016/j.jim.2004.04.014 PMID: 15261577
7. Boder ET, Wittrup KD. Yeast surface display for screening combinatorial polypeptide libraries. Nat Biotechnol. 1997; 15: 553–557. https://doi.org/10.1038/nbt0697-553 PMID: 9181578
8. Hoogenboom HR, de Bruine AP, Hufton SE, Hoet RM, Arends JW, Roovers RC. Antibody phage display technology and its applications. Immunotechnology. 1998; 4: 1–20. PMID: 9661810
9. Kettleborough CA, Ansell KH, Allen RW, Rosell-Vives E, Gussow DH, Bendig MM. Isolation of tumor cell-specific single-chain Fv from immunized mice using phage-antibody libraries and the re-construction of whole antibodies from these antibody fragments. Eur J Immunol. 1994; 24: 952–958. https://doi.org/10.1002/eji.1830240426 PMID: 8149964
10. Cai X, Garen A. A melanoma-specific VH antibody cloned from a fusion phage library of a vaccinated melanoma patient. Proc Natl Acad Sci U S A. 1996; 93: 6280–6285. PMID: 8692806
11. Davies EL, Smith JS, Birkett CR, Manser JM, Anderson-Dear D V, Young JR. Selection of specific phage-display antibodies using libraries derived from chicken immunoglobulin genes. J Immunol Methods. 1995; 186: 125–135. PMID: 7561141
12. Lang IM, Barbas CF 3rd, Schleef RR. Recombinant rabbit Fab with binding activity to type-1 plasminogen activator inhibitor derived from a phage-display library against human alpha-granules. Gene. 1996; 172: 295–298. PMID: 8682320
13. Arbabi Ghahroudi M, Desmyter A, Wyns L, Hamers R, Muyldermans S. Selection and identification of single domain antibody fragments from camel heavy-chain antibodies. FEBS Lett. 1997; 414: 521–526. PMID: 9323027

14. Meli G, Visintin M, Cannistraci I, Cattaneo A. Direct in vivo intracellular selection of conformation-sensitive antibody domains targeting Alzheimer's amyloid-beta oligomers. J Mol Biol. 2009; 387: 584–606. https://doi.org/10.1016/j.jmb.2009.01.061 PMID: 19361429

15. DeKosky BJ, Kojima T, Rodin A, Charab W, Ippolito GC, Ellington AD, et al. In-depth determination and analysis of the human paired heavy- and light-chain antibody repertoire. Nat Med. 2015; 21: 86–91. https://doi.org/10.1038/nm.3743 PMID: 25501908

16. Marks JD, Bradbury A. PCR cloning of human immunoglobulin genes. Methods Mol Biol. 2004; 248: 117–134. PMID: 14970493

17. Chirichella M, Lisi S, Fantini M, Goracci M, Calvello M, Brandi R, et al. Post-translational selective intracellular silencing of acetylated proteins with de novo selected intrabodies. Nat Methods. Nature Publishing Group; 2017; 14: 279–282. https://doi.org/10.1038/nmeth.4144 PMID: 28092690

18. Fischer N. Sequencing antibody repertoires: the next generation. MAbs. 2011; 3: 17–20. Available: http://www.ncbi.nlm.nih.gov/pubmed/21099370 https://doi.org/10.4161/mabs.3.1.14169 PMID: 21099370

19. Perelson AS, Oster GF. Theoretical studies of clonal selection: minimal antibody repertoire size and reliability of self-non-self discrimination. J Theor Biol. 1979; 81: 645–670. PMID: 94141

20. Glanville J, D'Angelo S, Khan T, Reddy S, Naranjo L, Ferrara F, et al. Deep sequencing in library selection projects: what insight does it bring? Curr Opin Struct Biol. 2015; 33: 146–160. https://doi.org/10.1016/j.sbi.2015.09.001 PMID: 26451649

21. Tanaka T, Lobato MN, Rabbitts TH. Single domain intracellular antibodies: a minimal fragment for direct in vivo selection of antigen-specific intrabodies. J Mol Biol. 2003; 331: 1109–1120. PMID: 12927545

22. Tanaka T, Rabbitts TH. Protocol for the selection of single-domain antibody fragments by third generation intracellular antibody capture. Nat Protoc. 2010; 5: 67–92. https://doi.org/10.1038/nprot.2009.199 PMID: 20057382

23. Chao A, Lee S-M. Estimating the Number of Classes via Sample Coverage. J Am Stat Assoc. 1992; 87: 210–217.

24. Weinstein JA, Jiang N, White RA 3rd, Fisher DS, Quake SR. High-throughput sequencing of the zebrafish antibody repertoire. Science (80-). 2009; 324: 807–810. https://doi.org/10.1126/science.1170020 PMID: 19423829

25. Glanville J, Zhai W, Berka J, Telman D, Huerta G, Mehta GR, et al. Precise determination of the diversity of a combinatorial antibody library gives insight into the human immunoglobulin repertoire. Proc Natl Acad Sci U S A. 2009; 106: 20216–20221. https://doi.org/10.1073/pnas.0909775106 PMID: 19875695

26. Logan AC, Gao H, Wang C, Sahaf B, Jones CD, Marshall EL, et al. High-throughput VDJ sequencing for quantification of minimal residual disease in chronic lymphocytic leukemia and immune reconstitution assessment. Proc Natl Acad Sci U S A. 2011; 108: 21194–21199. https://doi.org/10.1073/pnas.1118357109 PMID: 22160699

27. Gilles A, Meglecz E, Pech N, Ferreira S, Malausa T, Martin JF. Accuracy and quality assessment of 454 GS-FLX Titanium pyrosequencing. BMC Genomics. 2011; 12: 245. https://doi.org/10.1186/1471-2164-12-245 PMID: 21592414

28. Luo C, Tsementzi D, Kyrpides N, Read T, Konstantinidis KT. Direct comparisons of Illumina vs. Roche 454 sequencing technologies on the same microbial community DNA sample. PLoS One. 2012; 7: e30087. https://doi.org/10.1371/journal.pone.0030087 PMID: 22347999

29. Rhoads A, Au KF. PacBio Sequencing and Its Applications. Genomics Proteomics Bioinforma. 2015; 13: 278–289.

30. Greiff V, Menzel U, Haessler U, Cook SC, Friedensohn S, Khan TA, et al. Quantitative assessment of the robustness of next-generation sequencing of antibody variable gene repertoires from immunized mice. BMC Immunol. 2014; 15: 40. https://doi.org/10.1186/s12865-014-0040-5 PMID: 25318652

31. Georgiou G, Ippolito GC, Beausang J, Busse CE, Wardemann H, Quake SR. The promise and challenge of high-throughput sequencing of the antibody repertoire. Nat Biotechnol. 2014; 32: 158–168. https://doi.org/10.1038/nbt.2782 PMID: 24441474

32. DeKosky BJ, Ippolito GC, Deschner RP, Lavinder JJ, Wine Y, Rawlings BM, et al. High-throughput sequencing of the paired human immunoglobulin heavy and light chain repertoire. Nat Biotechnol. 2013; 31: 166–169. https://doi.org/10.1038/nbt.2492 PMID: 23334449

33. Vollmers C, Sit R V, Weinstein JA, Dekker CL, Quake SR. Genetic measurement of memory B-cell recall using antibody repertoire sequencing. Proc Natl Acad Sci U S A. 2013; 110: 13463–8. https://doi.org/10.1073/pnas.1312146110 PMID: 23898164

34. Shugay M, Britanova O V, Merzlyak EM, Turchaninova MA, Mamedov IZ, Tuganbaev TR, et al. Towards error-free profiling of immune repertoires. Nat Methods. 2014; 11: 653–5. https://doi.org/10.1038/nmeth.2960 PMID: 24793455

35. Milstein C. The Croonian Lecture, 1989: Antibodies: A Paradigm for the Biology of Molecular Recognition. Proc R Soc London B Biol Sci. 1990; 239. Available: http://rspb.royalsocietypublishing.org/content/239/1294/1

36. Visintin M, Quondam M, Cattaneo A. The intracellular antibody capture technology: towards the high-throughput selection of functional intracellular antibodies for target validation. Methods. 2004; 34: 200–214. https://doi.org/10.1016/j.ymeth.2004.04.008 PMID: 15312673

37. Nordgard O, Kvaloy JT, Farmen RK, Heikkila R. Error propagation in relative real-time reverse transcription polymerase chain reaction quantification models: the balance between accuracy and precision. Anal Biochem. 2006; 356: 182–193. https://doi.org/10.1016/j.ab.2006.06.020 PMID: 16899212

38. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014; 30: 2114–2120. https://doi.org/10.1093/bioinformatics/btu170 PMID: 24695404

39. Zhang J, Kobert K, Flouri T, Stamatakis A. PEAR: a fast and accurate Illumina Paired-End reAd mergeR. Bioinformatics. Oxford University Press; 2014; 30: 614–20. https://doi.org/10.1093/bioinformatics/btt593 PMID: 24142950

40. Koohapitagtam M, Rungpragayphan S, Hongprayoon R, Kositratana W, Sirinarumitr T. Efficient amplification of light and heavy chain variable regions and construction of a non-immune phage scFv library. Mol Biol Rep. 2010; 37: 1677–1683. https://doi.org/10.1007/s11033-009-9583-6 PMID: 19554473

41. Andris-Widhopf J, Steinberger P, Fuller R, Rader C, Barbas CF 3rd. Generation of human scFv antibody libraries: PCR amplification and assembly of light- and heavy-chain coding sequences. Cold Spring Harb Protoc. 2011; 2011.

42. Tanaka T, Chung GT, Forster A, Lobato MN, Rabbitts TH. De novo production of diverse intracellular antibody libraries. Nucleic Acids Res. 2003; 31: e23. PMID: 12595572

43. Vettermann C, Schlissel MS. Allelic exclusion of immunoglobulin genes: models and mechanisms. Immunol Rev. 2010; 237: 22–42. https://doi.org/10.1111/j.1600-065X.2010.00935.x PMID: 20727027

44. Daly J, Licence S, Nanou A, Morgan G, Martensson IL. Transcription of productive and nonproductive VDJ-recombined alleles after IgH allelic exclusion. EMBO J. 2007; 26: 4273–4282. https://doi.org/10.1038/sj.emboj.7601846 PMID: 17805345

< Back to Article

Table 1.

Libraries complexity results.

More »

Library	Total reads[a]	Maximal complexity[b]	DEAL complexity[c]	Minimal complexity[d]	Theoretical Complexity[e]
hscFv1	6.02×10^6	15.8×10^6	3.45×10^6	0.6×10^6	8.96×10^6
hscFv2	4.09×10^6	14.0×10^6	3.60×10^6	0.2×10^6	8.38×10^6
hVH	14.9×10^6	6.00×10^6	4.88×10^6	1.4×10^6	6.41×10^6

[a]Number of reads passing the first quality trimming analyzed by DEAL.
[b]Theoretical maximal complexity derived from the number of bacterial transformants.
[c]Number of DEAL clusters.
[d]Number of DEAL clusters minus the cluster with cardinality = 1.
[e]Complexity estimated by truncated negative binomial fit.

https://doi.org/10.1371/journal.pone.0177574.t001

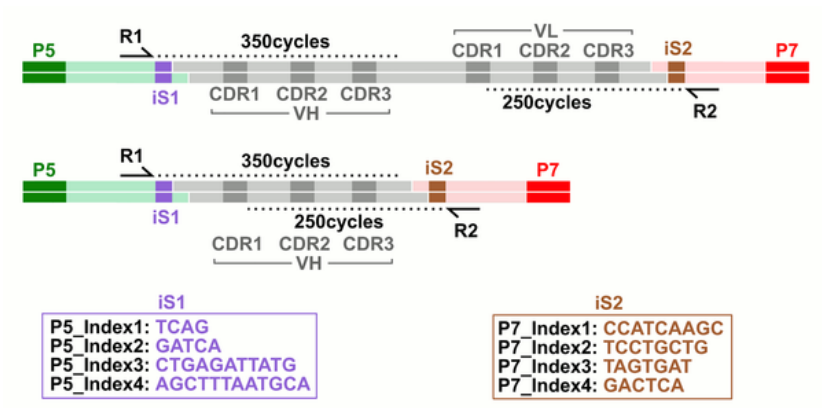
Expand

Fig 1.

Diagram of sequenced adaptor-antibody-adaptor constructs.

A) scFv library (gray), comprising heavy (VH) and light chain (VL) Complementary Determining Regions (CDR), was ligated to adapters (light green and pink) harbouring Illumina P5 and P7 flowcell hybridization sequences (green and red). B) VH nanobody library (gray), comprising heavy chain (VH) Complementary Determining Regions (CDR), was ligated to adapters (light green and pink) harbouring Illumina P5 and P7 flowcell hybridization sequences (green and red). The forward read (R1) uses SBS3 sequencing primer (Illumina), while the reverse read (R2) uses SBS12 primer (Illumina). iS1 and iS2 = index/shifter sequences.

More »



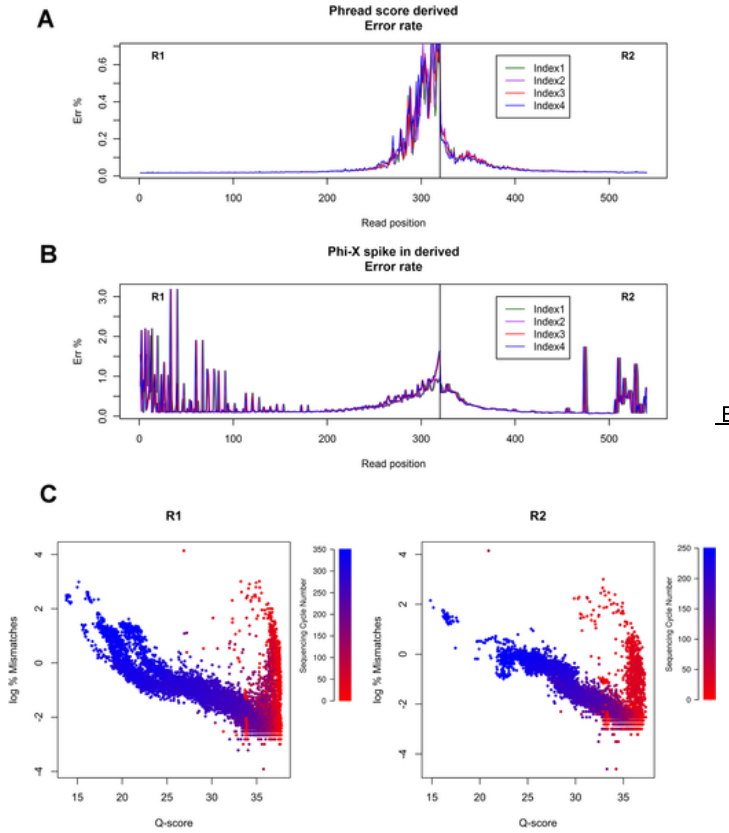
Expand

Fig 2.

Phi-X derived and Phred score derived error rate distribution.

A) Phred score error rate distribution for the hscFv1 library of the merged reads. Error rate increases with sequencing cycles. B) Control Phi-X derived error rate distribution for the hscFv1 library of the merged reads. Error rate is more prominent in the early sequencing cycles (spikes), with a small increase at the end of each read. The error distribution does not match the Phred score distribution and the shape differs as well. C) Scatter plot of the correlation of Q-score and $\log_2$(% Mismatches) in Phi-x control spike-in library. Each point represents the mean value from a single flow cell tile at a given sequencing read number, encoded by colour (red to blue: R1 cycle 1 to 350; R2 cycle 1 to 250; colour flex point is set at cycle 38). The Q score in the first 40 reads fails to be predictive of mismatch rate. Similar results were obtained for hscFv2 and hVH libraries.

More »



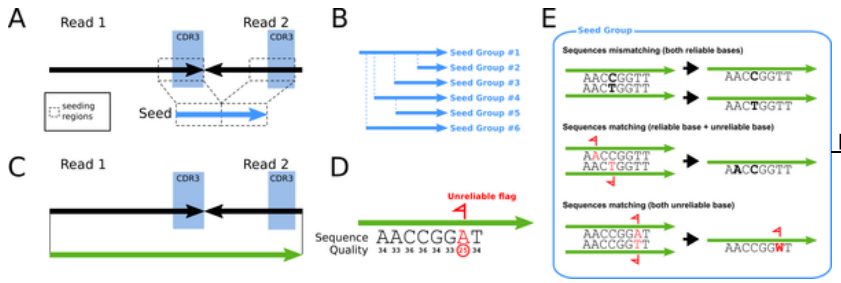
Expand

Fig 3.

Diagram of DEAL workflow.

A) Diagram of the seed creation process. In the figure, the black arrows represent the combined reads of the scFv library after the trimming. The seed is created combining the two seeding regions. The seeding regions are placed in the CDR3s to maximize the number of different seeds: the higher the number, the faster the program will run. B) Binary tree of the seeds. The program uses a binary tree approach to group identical seeds. During the comparison, if one sequence does not match any other sequences seen so far, a new branch of the tree is created in the mismatching position. C) The input of the binary comparison step. While the seeding step takes only into account the diversity of the seeding regions, the binary comparison analyzes the whole length of the combined reads. D) Flagging process. If some positions of the sequence are unreliable due to being associated to a low Phred quality score (as shown in the figure) or to a poor quality cycle (from Phi-X errors, not shown in the figure), the program flag them for correction. E) The three different scenarios that can occur during binary comparison among the sequences in the same seeding group. Mismatching (top): if two compared sequences differ in even only one position (bold) where none of the alternatives are flagged, the program recognize them as different sequences and does not group them. Matching sequences with a position having one flagged nucleotide (middle): the program recognizes the two sequences as identical and groups them together. All the positions where one of the sequences has a flag is resolved, during merging, as the not flagged nucleotide on the other sequence. Matching sequences with a position having both alternative nucleotides flagged (bottom): the program recognizes the two sequences as identical and groups them together. All the positions where both sequences have a flag are resolved using the IUPAC nucleobases ambiguity codes. The resulting merged sequence is flagged in that position.

More »



Expand

Fig 4.

Distribution of library sequence cluster cardinality.

Distribution of library sequence cluster cardinality. The more the curve is skewed towards high cardinality clusters, the lower the complexity of the library is expected to be.

More »

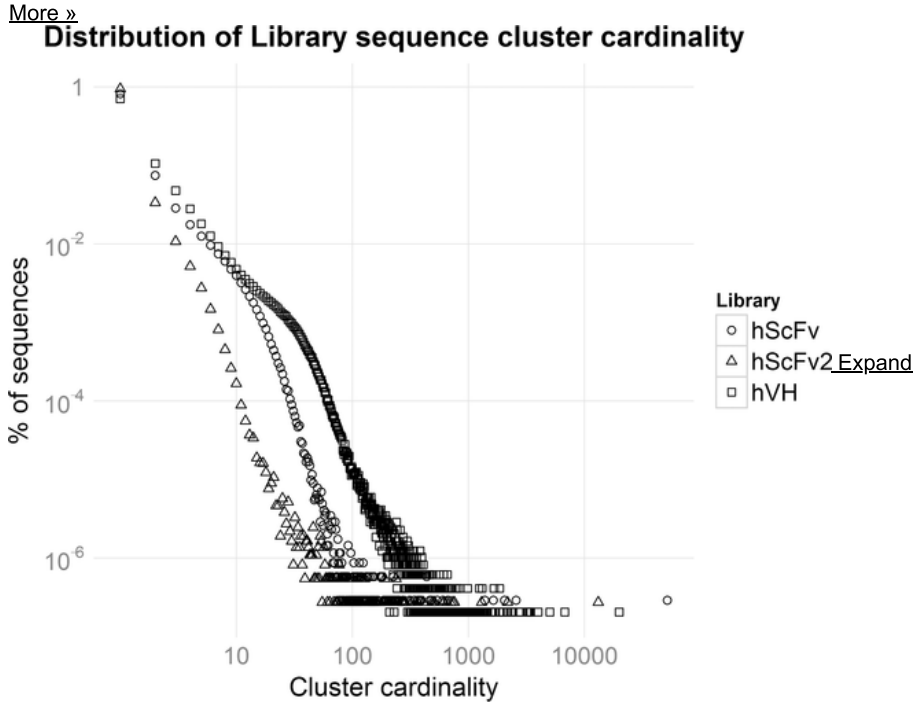
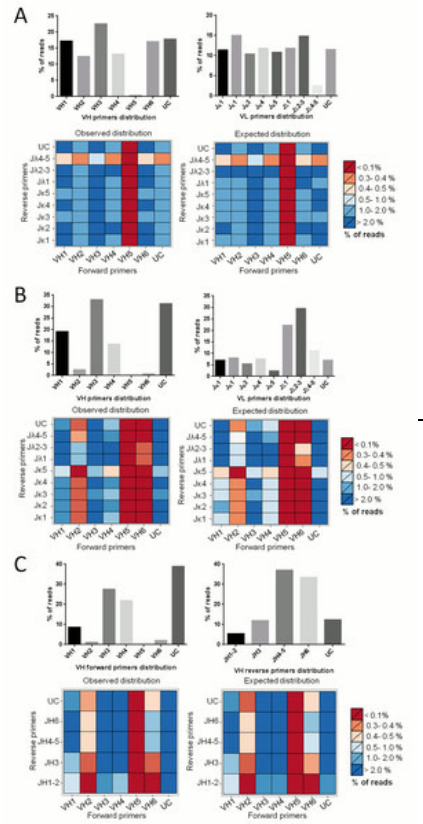


Fig 5.

Chain/VDJ assortment independence of libraries.

A) hscFv1. B) hscFv2. C) hVH. Top panels: barplots of forward and reverse primer distributions. Bottom panels: heatmaps of library primers distributions. Observed distribution is the primer pair proportion found after sequencing. Expected distribution is the multiplication of the two primers proportion (expected distribution given the independence between chains for the scFv libraries or given a balanced VDJ recombination for hVH). UC = unclassified. This category includes all the sequences that do not match any primer. The name of the primers is a shorter version of the original name listed in Supporting Information (Primer used for library construction).

More »



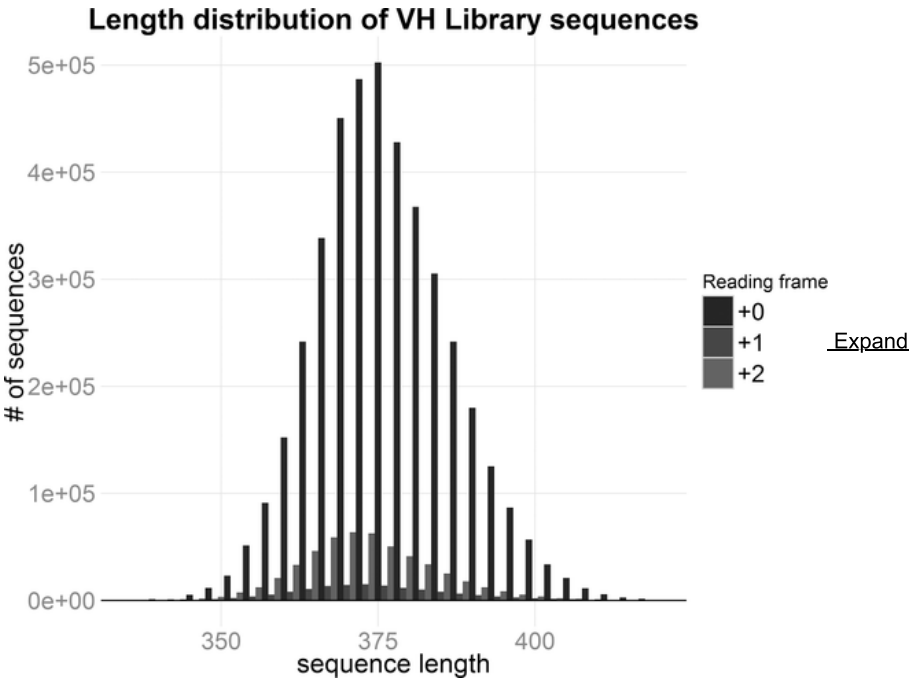
Expand

Fig 6.

Length distribution of human VH nanobody library sequences.

Barplot of the length distribution of human VH nanobody library sequences coloured by reading frame.

More »



Expand

Table 2.

hVH library protein complexity.

More »

Total reads[a]	% in frame sequences[b]	% sequences without stop codon in CDS[c]	% sequences with full length CDS[d]	Total DEAL protein complexity[e]	Total DEAL functional complexity[f]	Minimal DEAL protein complexity[g]	Minimal DEAL functional complexity[h]
14.9×10^6	86.57%	89.7–84.0%	83.4–77.9%	4.73×10^6	$3.25–3.19 \times 10^6$	1.43×10^6	$1.25–1.20 \times 10^6$

[a]Number of reads passing the first quality trimming analyzed by DEAL.
[b]Percentage of functional in frame sequences.
[c]Percentage of sequences which do not contain stop codons or do not contain possible stop codon in a error flagged sequence.
[d]Percentage of sequences in frame and without stop or possible stop codons.
[e]Protein complexity from DEAL protein clustering.
[f]Number of DEAL protein clusters with full length CDS.
[g]DEAL protein complexity minus the protein cluster with cardinality = 1.
[h]DEAL protein complexity with full length CDS minus the protein cluster with cardinality = 1.

Expand

https://doi.org/10.1371/journal.pone.0177574.t002