The pipeline repertoire of Ig-Seq analysis

Laura López-Santibáñez-Jácome,[1,2] Selma Eréndira Avendaño-Vázquez,[1] Carlos Fabián Flores-Jasso[1]*

Affiliations

[1] Consorcio de Metabolismo de RNA, Instituto Nacional de Medicina Genómica. Mexico City, Mexico.

[2] Maestría en Ciencia de Datos, Instituto Tecnológico Autónomo de México. Mexico City, Mexico.

* Correspondence:

Carlos Fabián Flores-Jasso

Periférico Sur 4809, Mexico City, 14610, Mexico

Email address: cfflores@inmegen.gob.mx

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27444v1 | CC BY 4.0 Open Access | rec: 22 Dec 2018, publ: 22 Dec 2018

Abstract

With the advent of high-throughput sequencing of immunoglobulins (Ig-Seq), the understanding of antibody repertoires and its dynamics among individuals and populations has become and exiting area of research. There are an increasing number of computational tools that aid in every step of the immune repertoire characterization. However, since not all tools function identically, every pipeline has its unique rationale and capabilities, creating a rich blend of useful features that may appear intimidating for newcomer laboratories with the desire to plunge into immune repertoire analysis to expand and improve their research; hence, all pipeline strengths and differences may not seem evident. In this review we provide an organized list of the current set of computational tools, focusing on their most attractive features and differences in order to carry out the characterization of antibody repertoires so that the reader better decides a strategic approach for the experimental design, and computational analyses of immune repertoires.

.

Key words: Ig-Seq, Antibody repertoire, Pipeline, V(D)J alignment, Pre-processing.

Background

The study of antibody repertoires by high-throughput sequencing prompted many groups to develop computational pipelines that aid in the processing of large amounts of sequencing data in order to categorize and understand the diversity and dynamics of repertoires in individuals (1,2). Practically every maturation step can be followed experimentally by high-throughput sequencing, giving us the opportunity to analyze how a diverse exposure to antigens has a distinctive effect on a myriad of individual B cells, either at transcriptomic, or genomic level (3–5). As new discoveries arise in the immunology field, novel computational tools have emerged to adapt their algorithms to provide more accurate and statistically robust analyses (6,7). Furthermore, computational pipelines have also helped to unveil details previously unknown about the antibody repertoire; exhibiting the intertwined relationship that exists between modern antibody repertoire analysis and computational immunology (4,8–15). Since the study of antibody repertoires can be addressed from many biological aspects, there is a concomitant diverse set of computational algorithms tailored to many purposes. Whereas high-throughput sequencing has become more available for most laboratories, there is a lag in the expertise required to plunge into the current computational pipelines developed for immunoglobulin sequencing (or Ig-Seq). With the large amount of software devoted for a specific (or all) processing step(s), the analysis of antibody repertoires may seem intimidating for newcomer laboratories; as the necessary processing steps to fulfill a specific type of analysis, or the reason for using a specific tool may not be as evident.

In this review, we focus on the current repertoire of some of the most widely used computational pipelines for Ig-Seq and provide a comparison of all the specific processes they perform. We begin by briefly explaining the basic rationale for each computational step required for Ig-Seq, to then describe the strengths and differences of each pipeline, emphasizing where the computational pipelines may converge and diverge to explore a repertoire biogenesis process. The computational steps discussed in this review are: pre-processing, V(D)J germline assignment, clonal grouping, mutation analysis, evolution and convergence of antibody repertoires. We also provide a table for all the pipelines discussed that displays their differences, and it could serve as a starting point and reference for Ig-Seq analysis, or project design. Based on our analysis, we organized the current pipeline repertoire into three main groups defined by the type of analyses each performs: Broad spectrum, Modular and Specialized pipelines. The table is portrayed so that its printouts can be revisited frequently, and the pipeline's similarities and differences are spotted easily.

While this review focuses only on the pipelines for antibody and B cell repertoire, many pipelines manage T-cell receptors (TCRs) as well. For analysis of TCR repertoire analysis, sample and library preparation, or the mathematical basis for the statistics employed by some of the pipelines discussed here, the reader may refer to other thorough reviews published recently (16–20).

1. Pre-processing

The goal of the data pre-processing step is to transform Ig-Seq raw reads into error corrected sequences. Although results are not significantly different between methodologies, preprocessing steps may vary depending on the amplification and sequencing methods employed (21–23). Due to the large extent of variability that gives rise to all B-cell clones, the identification of antibody repertoires and their germline assignment is intricate and largely compromised by biases and errors introduced during library preparation and sequencing; unlike the computational analysis commonly done for other types of high-throughput sequencing of nucleic acids (1). Because B cells undergo VDJ recombination and somatic hypermutation (Box 1 and 2), the sequences of interest can only be mapped to the reference genome partially. Therefore, errors introduced during the library preparation can be falsely identified as part of the true sequence of the antibody. The first approach to identify and minimize errors introduced during the library preparation was the establishment of *Unique Molecular Identifiers* (UMI) before the PCR amplification (19,20,24). This approach consists in the use of sequence barcodes introduced during the retro-transcription that allow to distinguish if polymerization errors were introduced early in the PCR, or whether they rather reflect a biological change in a given sequence. Along with UMIs, the accuracy of the repertoire reproduction can also be improved significantly by implementing molecular amplification fingerprinting (MAF), which uses UMI tagging before and during multiplex PCR amplification (25). Importantly, the use and identity of UMIs and MAFs should be decided before library preparation because further processing steps require to specify if they were used or not.

The typical steps of data pre-processing are:

Quality control and read annotation. Since BCR sequences could differ theoretically from one another by a single nucleotide, keeping high quality reads is of utmost importance (26). The output obtained from NGS is a FASTQ file that contains each read sequenced and information about its quality per base; referred to as *Phred quality score*. Regularly, reads with a Phred score of 20–30 are considered acceptable, whereas reads below 20 are discarded. After quality score analysis, the information introduced to each sample during library preparation must be annotated and masked or removed for each read —for example, annotation of average quality and UMIs/adaptors used.

Building consensus sequences. The goal of this step is to cluster all reads by UMI and to build a consensus sequence that has minimal amplification errors. Each UMI cluster results in a single consensus sequence with the most reliable base calls. Using UMIs ensures that all the reads coming from the same mRNA will have a unique barcode introduced during the retro-transcription. The amplification of errors and biases are then corrected by clustering the UMIs.

***Assembly of paired-end reads*.** When performing paired-end sequencing, the reads must be assembled into one read. In paired-end sequencing, the nucleic acid fragment sizes

are selected so that the sequences read from both ends (5′ and 3′), overlap with each other to some extent. Assembling the two cognate reads into a single sequence can be done by scoring different possible overlaps and by selecting the most statistically significant.

2. V(D)J germline alignment

The V(D)J germline assignment is one of the most important steps in the processing of Ig-Seq data. The goal of this step is to infer the correct germline alleles that recombined to produce each BCR/antibody. A good germline inference is critical to identify correctly somatic mutations for each read, to cluster into clonal groups, and to have a fair diversity approximation (27). Most commonly, this inference is done by applying an algorithm to choose the best match among a set of potential germline segments from a database of known segment alleles.

2.1 Assessing germline alignment of new alleles

Currently, most germline alignment and assignment tools compare the reads to existing databases of known alleles. Since all existing databases are incomplete, if novel polymorphisms appear in a sequencing study of Ig repertoire, these are difficult to differentiate from the frequent occurrence of somatic hypermutations in antibody sequences. Thus, the alignment of reads to only known genes may yield inaccurate results for sequences with previously undiscovered alleles —the read will be aligned to the closest germline gene, and the new allele variance will be incorrectly identified as a result of somatic hypermutation. To address this limitation, a distinct collection of tools has been developed to identify novel alleles, aimed to generate personalized germline databases containing the specific sets of alleles carried by individuals.

3. Clonal grouping/ clonotyping

The goal of this stage is to group antibodies/BCRs to facilitate lineage reconstruction and diversity analysis. In clonotype grouping, lineage trees are often constructed since they offer a visual display of the relationships among antibodies, and may also be helpful to understand temporal aspects of affinity maturation. There are alternative definitions for clonotypes that govern the type of analysis executed depending on the grouping criteria; one widely used, for example, is by grouping all clonotypes that descend from the same naive B cell but differ only because of their SHM process (13). Clonal grouping, under this example, is defined at the nucleotide level (due to somatic hypermutation occurs on the DNA): reads with the same V and J alleles and a threshold nucleotide difference at the junction region (CDR3) are grouped to form a clonotype. Other pipelines define clonotypes differently, for example, those that consider each clone as a unique antibody, or those who group them by amino acid sequences, meaning every read in the group will react to the same antigen (27–36). A strategy

aimed to provide a more robust clonotyping process, independent of any definition, is single linkage clustering (a statistical method for hierarchical clustering), which defines the distance between groups as the minimum distance between all pairs of points from the given groups (37,38).

4. Repertoire characterization and analysis

4.1 Diversity

Antibody/BCR diversity is associated with an effective immune response against pathogens and bacterial and viral infections(39). Despite the maximum theoretical amino acid diversity of ~10^{140} antibodies/BCRs, the effective repertoire is attributable to a set only ~10^{11} in humans due to the V D and J genes(1). This enormous diversity has led to the development of an increasing number of computational tools designed to tackle the concomitant complexity of immune repertoires. Typically, two complementary modules constitute diversity analyses:

4.1.1 Diversity quantification and analysis of the sequenced sample

Diversity quantification refers to a basic characterization and statistics of the repertoire that may include some or all of the following: mean clonotype sizes and their read counts, number of non-functional clonotypes, CDR3 region characterization, identification of the most used V, D and J alleles in the repertoire, and the most frequent VJ combinations. The latter, can be visualized either through heatmaps, histograms or pie charts. Furthermore, if a 3′ primer was set to cover the C region during the library construction, it is possible to perform an analysis to identify the most abundant isotype; this is especially relevant when conducting protocols for "before-and-after" vaccination and their respective repertoire analysis(3,9,12,14,40–42). The statistical quantification and comparison of clonotype diversity is primarily calculated using the *generalized diversity index* —a general mathematical formulation commonly used in ecology to assess diversity, developed by Hill in 1973(43).

Other indices used for repertoire diversity quantification that derive from Hill's capture a different clonal subset of the clonal frequency distribution(2). These include the species richness index, the Shannon Weiner index, the inverse Simpson index, the Berger Parker index (also referred to as the *reciprocal abundance of the largest clone*) the Gini index, and Chao1 index(44,45).

4.1.2 Estimation of total diversity

It is estimated that from the total number of theoretically possible B cell clones in an individual, ~28% is present in lymph nodes, ~23% in spleen, ~19% in intestinal mucosa, ~17% in bone marrow, and only ~2% in peripheral blood; in practical terms, these percentages account for the ~10^{11} BCR/antibody clones that are the product of the adaptive immune response(46). With the current sequencing technology, the maximum

number of reads a platform can achieve in a single run is $2x10^9$ —four orders of magnitude below the CDRs repertoire attributable to recombination and SHM. This implies that only a fraction of the total diversity repertoire can be identified by Ig-Seq. Therefore, the total diversity analysis must include the estimation of the undetected clones —mostly done using a rarefaction-based method(2). Rarefaction is a type of analysis (also commonly employed in ecology) that allows the calculation of the total species richness, and it is commonly portrayed as a "rarefaction curve" that plots the expected number of species as a function of the number of samples(47,48). Species (or in this case, read counts) are averaged over multiple resamples of the data to obtain the expected number of species as a function of the number of individuals(49). All the computational tools discussed here that support the estimation of total diversity do so by using a rarefaction-based method.

4.2 Mutation Analysis

High affinity BCRs/antibodies are the product of mutational events that accumulate during B cell maturation. The purpose of mutation analysis is to gain insight of the maturation process that B cells underwent during the course of an immune response and their encounter with antigens/foreign epitopes(4,26,50). As mutations accumulate in the CDRs, it is possible to identify hotspots that may give rise to a certain clonal population, providing even more understanding of the lineage and evolution process that took place at specific time points, or immunological events. The common features that build a mutation analysis enclose: mutation frequencies, mutations by position and hotspots identification, mutation types (*i.e.* number of synonymous and non-synonymous mutations which may indicate potential lineages under antigen-driven selection) and selection pressure —selection pressure is calculated by comparing the observed frequency of non-synonymous mutations and the expected number which takes into account hot- and cold-spots, and nucleotide substitution bias(51). An increased frequency of replacements indicates positive selection whereas decreased frequency indicates negative selection —CDRs are expected to be under positive selection, whereas framework regions under negative selection.

4.3 Evolution of repertoire/ Clonal dynamics

Immune repertoires are highly dynamic(40,42). When the immune system encounters an antigen, memory cells may be activated to produce already existent antibodies or naive B cells may hypermutate the variable regions of the antibody thus forming a B cell lineage with new plasma cells producing new antibodies(3). Analyzing how the antibody repertoire evolves throughout time provides an insight into how pathogens, vaccines, and even self-epitopes shape our humoral response(3,9,12,14,40–42,52,53). Although the evolution of repertoires can be studied at one single time point with phylogenetic trees that portray clonal dynamics, multiple time points help to generate

more accurate and robust ontogenic analysis and these can be portrayed as stream graphs, longitudinal phylogenetic "birthday" trees, stack-plots or clonotype tracking heatmaps.

In order to infer the antigen driven evolution of antibody repertoires, a phylogenetic method (such as neighbor joining (NJ), maximum parsimony (MP), maximum likelihood (ML) or Bayesian inference) is applied to the set of Ig reads(5,54), which results in the compartmentalization of all sequence reads into clades. Each clade contains all the sequences that share a common ancestor (i.e., that derive from the same naive B-cell). The phylogenetic tree is constructed using the resulting clades. Although the phylogenetic methods currently used are a fair approximation, most rely on assumptions that may be true for species evolution, but might be invalid for antigen driven evolution of antibodies; this inexorably would decrease the accuracy of the clade prediction.

In order to track clonotypes throughout multiple time points, a comparison between the repertoires at the different time points is performed(5,34,55) . This comparison allows the identification of antibodies that are present in more than one time point.

4.4 Repertoire convergence

Convergence analysis refers to a phenomenon that occurs when identical (or highly similar) immune receptor sequences are shared by two or more individuals in the context of infection or vaccination, and provides evidence that some antigenic stimuli can provoke relatively predictable responses, which is expected to occur in genetically similar populations(10,12,41,56). For example, it is been shown that the similarity of gene segments in productive IgH repertoires of twin brothers is greater in non-mutated than mutated Ig genes(12). This suggests that the process governing naive B-cell repertoire generation is more similar in related individuals, but that the subsequent antigen-driven evolution might be less genetically controlled. However, it is also conceivable that heritable biases in the naïve repertoire may affect the likelihood of clones with specific recombination becoming activated and transiting to the memory compartment(15). Repertoire convergence analysis may be of substantial importance for the prediction and manipulation of adaptive immunity as well as vaccination and gene therapy.

One way to determine the level of repertoire convergence is by finding the clonotype overlap across individuals (either at nucleotide or amino acid level), and express it as a percentage —normalized by the clonal size of the samples used(2). Hence, an accurate clonal clustering is of utmost importance in providing robustness to convergence studies. Here, the definition of clonotype employed usually refers to single sequences, and therefore two samples containing the same clonotype have a convergent sequence. In the case when clonotypes are treated not as single-, but clusters of sequences, two samples will share a cluster if at least one sequence is shared between them. In

convergence studies, it is common to use indices that provide additional information by integrating the clonal frequency of the compared clones; for example the Morisita-Horn index(2). Another index, the Repertoire Dissimilarity Index (RDI), enables the quantification of the average variation among repertoires(57). The RDI is a non-parametric method for directly comparing repertoires, with the goal of rigorously quantifying differences in V, D, and J gene segment utilization. Visualization of repertoire convergence can be achieved through Venn Diagrams, abundance plots, overlap circos-plots, or hierarchical clustering dendrograms. For a more thorough overview of the preprocessing steps and types of analysis of Ig-Seq refer to (2,23,42).

5. The pipeline repertoire

The identification of adaptive immune response receptors has yielded a number of diverse pipelines that perform part, or all the segments discussed above among others specifically tailored to cover specific needs and questions. As new discoveries arise in the immunology field, new tools are generated to manage the concomitant changes and implications in the antibody repertoire, and *vice versa*. The current section compiles the most widely used pipelines, and a brief description of the key features they offer to allow research at the cutting-edge of antibody repertoire analysis. The full list of features and capabilities is summarized in Table 1.

The pipelines that can handle most of the computational analyses discussed here are referred to as *Broad Spectrum* Pipelines. This comprises IgReC, ImmunediveRsity, the Immcantation Framework, IGGalaxy, and VDJServer. Computational tools that perform V(D)J Alignment and Clonal Grouping analyses, either as stand-alone or by wrapping other tools, are referred to as *Modular* Pipelines. Applications that do not perform V(D)J Alignment, and rather focus on specific computational steps of repertoire characterization (such as clonal dynamics, evolution, and convergence) are referred to as *Specialized* Pipelines. Lastly, the tools that only perform VDJ alignment are grouped under the *VDJ Alignment* category.The combination of the Broad Spectrum, Modular, Specialized and VDJ Alignment pipelines allows constructing an interrelated analysis tailored for each specific need, and facilitates the study of the fairly complex immune response.

5.1 Broad Spectrum Pipelines

5.1.1 IgReC

Preprocessing. *IgReC* is part of the Y-tools framework. It is an algorithm for constructing antibody repertoires from high-throughput sequencing datasets. It takes as an input both single and paired-end reads(28), and it provides the option to correct errors using UMIs. In case no UMIs were added to the libraries, error correction is performed by clustering the reads using the Hamming graph —whose vertices represent unique reads that are then used to build a consensus sequence.

V(D)J Germline Assignment. IgReC aligns first all reads to the database of Ig germline genes and discards the unaligned ones(28). To improve its efficiency, IgReC labels the V and J segments based on a fast algorithm for finding the longest subsequence of k-mers between reads and germline segments; the remaining reads are then realigned. This process bypasses the time consuming computation of extended Hamming distances.

Clonal Grouping/Clonotyping. IgReC tackles clonotyping completely different compared to the strategies described above: it builds a Hamming graph to identify similar reads to then identify dense subgraphs that become a clonotype of highly related antibodies. The visualization is done through a lineage tree or hamming graphics.

Diversity and Mutation Analysis. These steps are performed by IgDiversityAnalyzer, a complementary tool for annotation, diversity analysis and mutational analysis of full-length adaptive immune repertoires(28). The tool is capable of creating summary tables with pertinent information as well as plots for visualization.

5.1.2 ImmunediveRsity

Preprocessing. ImmunediveRsity is a tool primarily based in R programming for the integral analysis of B cell repertoire data. Although it is similar to other contemporary developed tools like MiGEC, MiXCR and pRESTO, it was the first to offer a beginning to end analysis, including ready to publish plots, within the same tool(29). ImmunediveRsity supports both single-, and paired-end formats and was originally designed for libraries prepared by RACE amplification. For the pre-processing of data, it performs the quality filtering and designed imm-illumina, a tool for the paired-end read assembly. One important feature to note is that ImmunediveRsity does not perform amplification correction based on UMIs itself; instead, it calls *Acacia*, an error-correction tool, after the V(D)J alignment.

V(D)J Germline Assignment. To assign the V(D) J segments, ImmunediveRsity uses IgBLAST by aligning each read to the current IMGT database. Moreover, the pipeline provides a file with the CDR3 of each read.

Clonal Grouping/Clonotyping. ImmunediveRsity's definition of a clonotype is a group of identical reads. Therefore each clonotype is a unique antibody. ImmunediveRsity uses the clonotyping to correct library preparation errors. Additionally, the pipeline provides the clonal abundance and a visual representation of the clonotype lineages. Instead of a lineage tree, the group visualization is a graphical network of the clonotype and their lineages; which could be helpful for a population dynamics approach.

Diversity. The ImmunediveRsity pipeline performs CDR3 identification and characterization, VJ usage heatmaps, diversity quantification (capable of performing Shannon Weiner index, Shannon Weiner normalized or weighted index and the Gini coefficient) and can plot the rarefaction curves for total diversity estimation.

Mutation Analysis. It reports the synonymous and non-synonymous mutations.

5.1.3 Immcantation Framework

Preprocessing. The Immcantation Framework includes the tool pRESTO which is capable of performing all the stages of pre-processing, from raw sequences up until the paired-end assembly if required. In order to fulfill all of its capacities, pRESTO is composed of a set of stand-alone tools that can be combined to construct commands specific to individual protocols(58). It supports multiplexed and RACE samples and is also capable of performing de-multiplexing if this was not done by the sequencing facility. pRESTO supports single-end sequencing, and it can assemble reads that do, and do not, overlap for the paired-end format. Reads processing can be carried out with or without UMIs, making it a very flexible, adaptable and suitable tool for many existing protocols.

V(D)J Germline and New Allele Assignment. The Tool for the immunoglobulin genotype elucidation (TIgER), identifies novel VJ segment alleles, and constructs a personalized germline database(59). This information is then used to improve the initial V segment assignments from existing tools, like IMGT/HighV-QUEST. This means one has to perform the alignment beforehand and then TIgER will correct the misinterpreted new alleles and create a personalized germline database. The required input for this tool is a germline database in IMGT-gapped fasta format, and a table of reads with the preliminary V and J alleles, and length junction. The table can easily be created with the output of Change-O (also part of the Immcantation portal), or IMGT, V-QUEST and IgBLAST.

Clonal Grouping/Clonotyping. For this step, the portal is assisted by the package Change-O for standardizing the output of alignment software such as IMGT HIGH V-QUEST or IgBLAST, clonal grouping and germline reconstruction(60). Change-O allows processing of reads that contain a premature stop codon, and would be nonfunctional. It groups clonotypes by V and J allele and the nucleotide Hamming distance; it also provides the option to choose which substitution model to use for calculating distance between sequences. The different substitution models will lead to the different definitions of clonotype. Available models are: nucleotide hamming distance, amino acid hamming distance, human specific single nucleotide model and 5-mer content model. It also enables to choose between single, average or complete linkage for the type of hierarchical clustering. The Immcantation framework also includes Alakazam, which is an R package that serves as interface for interacting with the output of Change-O and pRESTO. Alakazam takes the clonotypes grouped by Change-O and plots the lineage tree of the repertoire.

Diversity. This step is also performed by Alakazam, which performs basic repertoire characterization, diversity quantification and total diversity quantification (60). The tool calculates V(D)J allele, gene or family usage, as well as physicochemical properties of the amino acid sequences. For diversity quantification, one can calculate the species richness, the Shannon Weiner index, the inverse Simpson index and the Berger Parker

Index. When inferring the complete clonal abundance, Alakazam uses the Chao to estimate the number of seen clones, and then applies the relative abundance correction and unseen clone frequencies described in Chao et al, 2015 (11,48,61). Furthermore, the tool provides the rarefaction curve.

Mutation Analysis. Its uses the R package SHazaM to quantify the mutational load and it includes tools to build the SHM targeting models from the data. Moreover, the package includes a tool to analyze the selection pressure (60).

Convergence. The RDI package is part of the Immcantation analysis framework and provides methods for visualizing and calculating the Repertoire Dissimilarity Index (57).

5.1.4 IGGalaxy

IGGalaxy is a web-based application capable of use on and individual PC and on a server, which uses Galaxy's Graphical User Interface. Furthermore, the application provides an *ExperimentalDesign* tool that allows samples to be merged with an experimental design structure (i.e. name samples and replicates) (62).

Preprocessing. The preprocessing of the raw reads can be performed either by existing tools in Galaxy and IGGalaxy or can be performed with another tool before uploading the data into the application.

V(D)J Germline Assignment. One can perform the germline alignment using either the IgBlast wrapper or the IMGT HighV-QUEST wrapper provided by IGGalaxy.

Clonal Grouping/Clonotyping. IGGalaxy provides clusters the definition by the given definition of unique sequence (either VJ CDR-AA or VJ CDR nucleotide).

Diversity. The *Report* tool summarizes the frequency of V, D, and J chains as bar charts as well as the combination V-D, V-J and D-J heatmaps based on the definition of unique sequence. IGGalaxy *Report* tool provides further analysis using existing Galaxy genomic and statistical analysis functionality.

5.1.5 VDJServer

VDJServer is the first cloud-based analysis portal for immune repertoire sequence data that provides access to a number of tools for a start to finish analysis workflow. The portal is accessible through a standard web browser via a user-friendly graphical user interface, which facilitates use by research groups that lack bioinformatics expertise. Moreover, VDJServer provides free access to High Performance Computing (HPC) at the Texas Advanced Computing Center (63).

Preprocessing. The portal provides access to pRESTO or VDJPipe. VDJServer automatically calculates base composition statistics and read quality statistics before and after preprocessing and provides comparative visualization for user assessment.

V(D)J Germline and New Allele Assignment. Uses IgBlast to perform the alignment against the IMGT database.

Clonal Grouping/Clonotyping. For clonal grouping and annotation, the portal uses either Change-O or RepSum.

Diversity. The portal provides access to RepCalc and Alakazam.

Mutation Analysis. The portal provides access to SHazaM.

5.2 Modular Pipelines

5.2.1 MiGEC

Preprocessing. It was introduced as one of the first pipelines that processed UMI-tagged reads. MiGEC performs all pre-processing steps, including the de-multiplexing, but differs on the specific order mentioned above. For example, assuming that de-multiplexing is done, the first step in MiGEC is the clustering by UMI, creating molecular identifier groups accompanied by the size distribution and statistics for each group. Once the groups are created, the average quality score of the whole group is calculated. Low quality groups are discarded as they are assumed to contain errors at the early stages of PCR. Furthermore, MiGEC performs two stages for error correcting and the building of the consensus sequence. The first correction stage identifies the dominant sequence variant and corrects minor sequence variants within each group. The second step, builds the consensus sequence and eliminates the variants produced by hotspot PCR errors. The two step correction process minimizes error introduced by amplification and eliminates almost all artificial diversity (7). In terms of amplification, MiGEC supports both multiplexing and RACE amplification. It only supports paired-end sequencing with UMIs.

V(D)J Assignment. MiGEC is capable of mapping the V, D and J segment as well as the extraction of the CDR3 region by using the IMGT database.

Clonal Grouping/Clonotyping. In MiGEC each consensus sequence assembled is a clonotype after the amplification and sequencing correction with UMIs, which implies that every sequence is a unique antibody/BCR. Each clonotype is specified by count, fraction, V, D and J segment identifier list and CDR3 nucleotide and amino acid sequence.

5.2.2 IMSEQ

Preprocessing. IMSEQ is a tool that derives clonotype repertoires from NGS data and introduces a new routine for handling errors produced during the library preparation (29). It supports both single and paired-end formats and it performs all steps of the data preprocessing stage. The quality filtering process allows the filtering by read phred-like score or by cluster phred-like score. Furthermore, IMSEQ allows for amplification error correction without the need to use UMIs. This step is done by adding a second quality-filtering step after clonal grouping. Errors introduced during the amplification will produce new clonotypes that are highly similar to the true clonotype. In the post processing error correction step, IMSEQ checks for every identified clonotype cluster whether it is likely or not to be erroneously derived from another clonotype cluster. In the case that it is indeed likely to be derived from another clonotype cluster, said clonotype cluster is attributed to amplification errors and eliminated (29).

V(D)J Assignment. To identify efficiently the V and J reference genes that yield the best scoring overlap alignments against each read, IMSEQ initially matches a set of short segment substrings, denoted as segment core fragments (SCFs), against each read. After the VJ assignment, the CDR3 region is determined.

Clonal Grouping/Clonotyping. In terms of clonotyping, IMSEQ groups by unique antibody/BCR. In other words, each clone is a unique antibody. Additionally, when CDR3 reads are out of frame, or contain a stop codon inside its region, the read is rejected and considered non-functional. An interesting feature of IMSEQ is that the clustering is also used to provide a framework for PCR and sequencing error correction without the need for UMIs (see V(D)J germline alignment above).

5.2.3 MiXCR

Preprocessing. *MiXCR* is a very simple, yet flexible tool that handles paired- and single-end reads, supports both partial (only variable region) and full-length (full heavy chain) profiling, considers sequence quality and corrects PCR and sequencing errors (30). Since MiXCR does not support libraries prepared with UMIs, error correction is done by assembling the clonotypes with a heuristic multi-layer clustering. Furthermore, MiXCR supports RACE amplification and RNA-Seq methods.

V(D)J Germline Assignment. *MiXCR* employs built-in library of reference V, D, J and C gene sequences based on corresponding loci from the GenBank database. The pipeline also offers the option to use external libraries such as IMGT. One advantage MiXCR offers is that it is a well-rounded tool as it also performs the data pre-processing.

Clonal Grouping/Clonotyping. MiXCR groups clonotypes by their CDR3s by default, and therefore assembles clones by unique antibody/BCR sequences —a feature that could be modified if desired. A key aspect of MiXCR is that it offers capability to choose the gene regions (V, D, J, CDR3 and C) to be used for the assembling of the clonotypes.

5.2.4 LymAnalyzer

LymAnalyzer is a software that receives FASTAQ files and starts its processing at the alignment of the V(D)J genes to the reference alleles. It provides both command line and GUI versions.

V(D)J Germline Assignment. The software uses an alignment algorithm based on fast-tag-searching to map the input sequence to the reference V and J segments, and uses the Hamming distance to choose the for best match (64). Since is shorter than VJ, the alignment of the D segment is done by removing the V and J regions before applying the alignment algorithm. LymAnalyzer has better performance in its accuracy compared to MiXCR. By using the default settings, all reference genes are derived from the most recent update of IMGT database. LymAnalyzer also offers the option to import your own reference database and perform the CDR3 extraction.

Clonal Grouping/Clonotyping. The sequences are grouped when they contain the same V(D)J gene and have identical CDR3 nucleotide sequence.

Mutation Analysis. LymAnalyzer generates the mutation trees using a method that aims to reveal the minimal steps that could have led to the observed sequence. Fruthermore, the tool provides the hypermutation tree for IgSeq data.

5.2.5 Partis

V(D)J Germline Assignment. Partis is a fast, flexible, and open source framework based on the Hidden Markov Model (HMM) to analyze BCR sequences (27). Using a novel HMM "factorization" strategy, Partis performs V(D)J alignment using a database reference of choice. The germline reference database must be downloaded separately.

Clonal Grouping/Clonotyping. The framework clusters the sequences by lineage based on a multi-hidden Markov Model (65) .

Mutation Analysis. Partis reports mutation frequencies as well as of nucleotides corresponding to the non-templated insertions between the V and D segments and D and J segments.

5.2.6 ImmuneDB

V(D)J Germline Assignment. ImmuneDB is a system for analyzing vast amounts of heavy chain variable region sequences and exploring the resulting data. It uses MySQL as a database and accepts pre-annotated sequences in Change-O format as input. ImmuneDB implements a gene anchoring method for V and J identification (34). This package requires that V and J germlines be downloaded separately and specified in two separate FASTA files; each must follow IMGT formats. The tool assigns each sequence a V and J gene, but it also calculates statistics such as how well the sequence matches the

germline, if there is a probable insertion or deletion, and how far into the CDR3 the V and J likely extend.

Clonal Grouping/Clonotyping. It clusters the reads based on CDR-3 amino acid similarity. Therefore, all the reads in one group react against the same antigen. It provides the clonal lineage tree for visualization of the grouping.

Diversity. The tool performs V and CDR length distribution as well as V and J gene usage.

5.2.7 Vidjil

Vidjil is an open source platform for the inspection, analysis and the tracking of clones along the time. The unique feature of Vidjil is that it also provides a web application linked to a patient database where one can keep records of all patients alongside their Ig-Seq data information (32). The web application can visualize data processed by the Vidjil algorithm or by other V(D)J analysis pipelines, and enables to further explore the immune repertoire (66).

V(D)J Germline assignment and Clonal Grouping/Clonotyping. The platform uses as default the algorithm Vidjil to perform a fast alignment and clustering of the clonotypes. This process is based on a seed heuristic and is fast as no alignment is performed with germline database sequences in the first phase (67). Nonetheless, Vidjil also supports data processed by other clonal grouping tools such as IMGT-HighV-QUEST, IgBlast, MiXCR, IMSEQ among others .

Diversity. Provides the Shannon-Wiener and Inverse Simpson Diversity indexes.

Evolution. The Vidjil application allows for the tracking and visualizations of clones along time. It is currently being used in hospitals for the diagnosis of patients suffering Acute Lymphoblastic Leukemia (ALL) (68).

5.3 Specialized pipelines

The computational tools that perform clonal grouping/clonotyping and at least one feature for repertoire characterization are SONAR, TRigS, IMEX, Vidjil, IRProfiler and VDJtools. Those whose main features focus solely on the repertoire characterization are, DIVE, BASELINe, and AbSim.

5.3.1 SONAR

The Software for the Ontogenic aNalysis of Antibody Repertoires (SONAR) was specifically designed for analyzing the development of antibody lineages across time (33).

Preprocessing. The tool supports both single-end and pair end-reads. It performs quality control and annotation on Ig-Seq data.

Clonal Grouping/Clonotyping. SONAR first clusters the reads based on assigned V and J genes. The transcripts in each group are then clustered based on their CDR3 nucleotide identity. Therefore, a clonal group contains all IG reads that share a common ancestor. The tool provides the option for seeded or unseeded lineage assignment.

Diversity. It provides VJ usage and CDR length distribution plots.

Evolution. SONAR is capable of tracking the development of specific antibody lineages across time. The visualization of the evolution is portrayed as longitudinal phylogenetic "birthday" trees. In the case of longitudinal phylogenetic birthday trees, SONAR identifies the sequences that appear at multiple time points and assigns a birthday based on the first observation of the read (33)

5.3.2 TRigS

Clonal Grouping/Clonotyping. By using the built-in tool *ClusterSeqs*, TRigS groups reads by their common ancestor based on single linkage clustering (66). Clonotyping of CDR3s can be defined at the nucleotide or amino acid levels by an identity threshold (Hamming distance divided by the read length). Furthermore, the output from the clonotyping can be easily graphed and translated into annotated lineage trees, showing the positions at which nucleotide/amino acid substitutions occur.

Diversity. The built-in tool "PlotGermline", creates histograms showing germline usage. It is also capable of plotting the relative usage of a requested germline gene. In the case of diversity indexes, TRigS is able to calculate the Gini index for each cluster created.

5.3.3 IMEX

IMEX is a software framework (desktop application) for the detailed characterization and visualization of the state of human IG and TR repertoires.

Clonal Grouping/Clonotyping. In IMEX, the calculation of clonality can be defined by the user by choosing the amino acid or the nucleotide sequence or the V-(D)-J rearranged regions (28). The software enables the calculation of clonality based on the three CDRs (CDR1–3). Total numbers and relative frequencies of the clonotypes are given in tabular view.

Diversity. IMEX calculates sequence diversity using a more elaborated data mining approach based on the most variable region, the CDR3 (28). Furthermore, the software provides several different graphical representation possibilities to visualize the total gene and allele frequencies such as frequency histograms, heat maps, or bubble charts.

Convergence. It is capable of obtaining a list of unique CDR3 clonotypes for a data sample and searching for them in another sample. It also contains a visualization and tabular view to compare overlapping multiple data samples according to CDR3.

5.3.4 IRProfiler

Immune Repertoire Profiler (IRProfiler), runs in the Galaxy environment and delivers a variety of core immune repertoire quantification and comparison functionalities on high-throughput BCR sequencing data (36). IRProfiler receives annotated IMGT HIGH-V Quest reads and other annotated high-throughput dataset that incorporates the same fields as in the Summary Report provided by IMGT.

Pre-processing IRProfiler provides 11 different quality-filtering criteria.

Clonal Grouping/Clonotyping. The pipeline contains five different definitions of clonotype to suit different analysis purposes. All clonotype definitions are in terms of aminoacid sequences. IRProfiler starts with the IMGT definition of clonotype, which includes the same V, D and J gene and the same amino acid sequence at the CDR3, and gradually transitions towards a less detailed definition.

Diversity. IRProfiler provides a summary of the clonotype quantification. This includes the dominant clonotype and its frequency, the total number of clonotypes and the total number of expanding clonotypes. Furthermore, it is capable of performing V and J gene usage.

Convergence. The "public clonotypes" tool implemented by IRProfiler, allows the identification of clonotypes present in more than one repertoire. It receives the list of clonotypes for each repertoire and outputs a file containing the shared clonotypes accompanied by their frequencies in each input repertoire and repertoire counts. Moreover, multiple V or J gene repertoires can be compared with respect to the gene usages.

5.3.5 VDJtools

VDJtools is an open-source framework which computes a wide set of statistics and is able to perform various forms of cross-sample analysis (55). VDJtools provides both tabular output and publication-ready plots.

Preprocessing. VDJtools takes as input already aligned and clonotyped Ig-Seq data. Besides supporting data processed with MiGEC, IgBlast, IMGT HIGH-V QUEST, Vidjil, MiXCR or IMSEQ, the framework also VDJtools provides further preprocessing features. The built-in tool "Correct" performs frequency-based correction to eliminate erroneous clonotypes. Furthermore, the framework provides a variety of filtering options; *e.g.* filtering non-functional clonotypes, filtering out all clonotypes found in another sample, filtering by frequency, and filtering V/D/J segments that match a specified segment set.

Diversity. For each sample, VDJtools calculates basic statistics of read counts, mean clonotype size, and number of non-functional clonotypes. It determines VJ gene usage

and spectra-typing (the distribution of clonotype abundance by CDR3 sequence length). As for diversity indexes, the framework provides a wide arrange of options such as Chao1 or Efron-Thisted estimate for lower bound diversity and Shannon-Wiener index, Normalized Shannon-Wiener index and Inverse Simpson index for diversity.

Evolution. VDJtools performs an all-vs-all intersection between an ordered list of samples for clonotype tracking. Results are visualized as clonotype tracking stackplots or heatmaps.

Convergence. The framework performs a comprehensive analysis of clonotype sharing for a pair of samples. Data can be visualized as scatterplots of overlapping clonotype abundance, abundance plots, hierarchical clustering dendrograms and pairwise overlap circos plots. A unique feature provided by VDJtools is the built in tool "CalcPairwiseDistances" which performs an all-versus-all pairwise overlap for a list of samples and computes a set of repertoire similarity measures (such as Pearson correlation).

5.3.6 DivE

Diversity. DivE is an R package specifically designed to estimate the total diversity of a sample through rarefaction curves. DivE fits various mathematical models to multiple nested subsamples of individual-based rarefaction curves and choses the best performing model to create the final rarefaction curve and diversity estimation (48,49). The result is a new species richness estimator named DivE estimator. This new method was compared with some of the most widely used non-parametric total diversity estimators (Chao 1, abundance-based coverage estimator(ACE), Bootstrap and Good-Turing estimator) and resulted to be more accurate than these.

5.3.7 BASELINe

Mutation Analysis. The statistical framework for Bayesian estimation of Antigen-driven SELectIoN (BASELINe) specializes on the analysis of somatic mutation patterns (51).The tool identifies the type (silent or point) and location (FWR or CDR) of the mutations, it calculates the Bayesian estimation of replacement frequency for every read and positive and negative selection pressure. Moreover, the tool is capable of performing a comparative analysis between groups of sequences derived from different germline V(D)J segments.

5.3.8 AbSim

AbSim is an R package designed to create simulations of antibody repertoires. It allows the user to control biologically relevant parameters such as total time for evolution, rate

and method of SHM, number and rate of VDJ recombination events, baseline mutation rate, rate at which new sequences are produced, clonal frequency and VDJ germline gene distribution (5). AbSim is the first repertoire simulation framework that enables the comparison of commonly used phylogenetic methods with regard to their accuracy in inferring antibody evolution.

5.4 VDJ Alignment

Being that the V(D)J alignment is one of the most important and intricate steps in the Ig-Seq data processing, the following are stand alone tools that focus solely on this computational task.

5.4.1 IgBlast

V(D)J Germline Assignment. It derived from the commonly used BLAST algorithm to perform specialized Ig-Seq alignment and similarity searching. It is a web based application or it can be used as a command line tool that aligns Ig reads to the germline reference database of choice. Moreover, it allows to visualize what matches to the germline alleles, the details at the rearranged junctions, and the framework and CDR regions (69), which positioned IgBlast is a gold standard for V(D)J mapping. However, the output is not straightforward to parse and summarize to a readable clonotype abundance table containing CDR3 sequences, segment assignments and list of somatic hyper-mutations. This motivated the development of MIGMAP; a tool that wraps IgBlast and is designed to facilitate analysis immune receptor libraries profiled using high-throughput sequencing by IgBlast (55). MIGMAP extends IgBlast capabilities like assembling of clonotypes, application of various filtering options such as quality filtering for CDR3 N-regions and mutations, among others.

5.4.2 IMGT High V-Quest

V(D)J Germline Assignment. The international ImMunoGeneTics information system (IMGT) is the most complete and used database for germline immune alleles (70). High V-QUEST is a web based standalone tool created by IMGT that allows the alignment of Ig reads to its germline allele database. It can handle up to $5x10^5$ reads simultaneously; larger files can be split using pRESTO. High V-QUEST identifies the closest V, D and J alleles based on a global pairwise alignment of each read with different subsets of the IMGT reference database, followed by an evaluation by similarity. The standard output contains the reads with its respective closest V, D and J allele, and its corresponding identity percentage/score. It also provides the FR and CDR delineations and the three CDR lengths. Furthermore, the "synthesis view" facilitates the visual comparison of

sequences that express the same V gene and allele but differ in mutation locations and junctions.

5.4.3 HTJoinSolver

V(D)J Germline Assignment. It is an application that introduced a dynamic programming approach method that uses conserved immunoglobulin gene motifs to improve the performance of aligning V(D)J segments (71). In order to run, one has to download the database from IMGT (or preferred database) to then upload it to the application. HTJoinSolver provides methods to download and re-format germline genes from IMGT. The results produced by the HTJoinSolver are the V, D and J germline alleles that best align with the reads.

5.4.4 IMPre

The Immune Germline Prediction, IMPre, is a stand-alone tool designed to predict germline V/J genes and alleles derived from BCR repertoire data.

V(D)J Germline Assignment. IMPre mimics the reverse process of VDJ rearrangement and supports the discovery of new alleles (72). IMPre can process the rearranged sequences with or without the C region. When including the C region, it will be identified using previously reported C sequences. A simple fasta file containing the known germline database must be provided for IMPre to work correctly.

5.4.5 IgDiscover

IgDiscover is a stand-alone tool that supports the discovery of new alleles in heavy chains (VH), light kappa chains (VK) and light lambda chains (VL) genes (73). It can receive input data from pair end reads in FASTAQ format or single end reads in either FASTAQ or FASTA format. It requires a starting database of VH, VK or VL genes that are used for primary assignment —done by using IgBLAST.

V(D)J Germline Assignment. The output of IgDiscover is an individualized germline database that includes a dendrogram of the V, D and J sequences, the VDJ assignments and their expression counts. It was proven to identify germline V genes with 100% accuracy.

5.4.6 IGoR

The Inference and Generation Of Repertoires, IGoR, is a comprehensive tool that takes the repertoire reads and quantitatively characterizes the statistics of receptor generation (6).

V(D)J Germline Assignment. IGoR explores all possible recombination scenarios for the read and provides the probabilities of each; a robust feature that in some instances may outperform other pipelines. IGoR can be used to infer recombination models, evaluate sequence statistics and to create synthetic sequences using an already generated recombination model.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27444v1 | CC BY 4.0 Open Access | rec: 22 Dec 2018, publ: 22 Dec 2018

6. Perspectives

The current repertoire of computational pipelines offers an unprecedented opportunity to study the immune response in individuals and populations. The review presented here compiled the most widely used pipelines and the features by which they converge and diverge. The table accompanying this review aims to aid on when and why use a specific pipeline, as well as helping to visualize the best experimental strategy in order to characterize the immune repertoires. We expect this information may evolve as the pipelines adapt to new discoveries and technologic development.

With the advent of high-throughput sequencing it is now possible to understand the dynamics of the immune response diversity and function. Novel techniques that allow sequencing the transcriptome landscape of single cells will also offer a more precise perspective of the metabolic changes associated to antibody production by B cells (74). Also single cell sequencing offers the advantage of determining the genes and transcripts that give rise to both, heavy and light chains of each antibody; making possible to determine more precisely the genetic relationships and dynamics of antibody repertoires (15,56,74). Yet another aspect of the antibody repertoire analysis that we envision will have an impact in the future of precision medicine is the extent at which immune repertoires overlap among individuals and populations (42). As many more studies are available, the robustness for the convergence analysis will allow framing a less blurred picture of our response to foreign epitopes as species. Importantly, with the aid of convergence analysis, the treatment and prevention of autoimmune diseases will become more evident; as the understanding of clonal dynamics of auto-antibodies will permit to better interpret the causes that trigger the immune system to recognize self-epitopes as foreign. It is our hope that this review will probe to be useful in attracting more research groups into this exciting area.

Author Contributions

LLSJ assembled Table 1; LLSJ, SEAV, and CFFJ discussed and wrote the manuscript.

Conflict of Interest

The authors declare no competing of interests.

Acknowledgements

This work was supported by the Instituto Nacional de Medicina Genomica [05/2017/I-321] to CFFJ.

References

1. 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. doi:10.1038/nbt.2782
2. Miho E, Yermanos A, Weber CR, Berger CT, Reddy ST, Greiff V. Computational strategies for dissecting the high-dimensional complexity of adaptive immune repertoires. *Front Immunol* (2018) doi:10.3389/fimmu.2018.00224
3. Jiang N, He J, Weinstein JA, Penland L, Sasaki S, He X-S, Dekker CL, Zheng N-Y, Huang M, Sullivan M, et al. Lineage Structure of the Human Antibody Repertoire in Response to Influenza Vaccination. *Sci Transl Med* (2013) **5**:171ra19-171ra19. doi:10.1126/scitranslmed.3004794
4. Yaari G, Benichou JIC, Vander Heiden JA, Kleinstein SH, Louzoun Y. The mutation patterns in B-cell immunoglobulin receptors reflect the influence of selection acting at multiple time-scales. *Philos Trans R Soc B Biol Sci* (2015) doi:10.1098/rstb.2014.0242
5. Yermanos A, Greiff V, Krautler NJ, Menzel U, Dounas A, Miho E, Oxenius A, Stadler T, Reddy ST. Comparison of methods for phylogenetic B-cell lineage inference using time-resolved antibody repertoire simulations (AbSim). *Bioinformatics* (2017) **33**:3938–3946. doi:10.1093/bioinformatics/btx533
6. Marcou Q, Mora T, Walczak AM. High-throughput immune repertoire analysis with IGoR. *Nat Commun* (2018) **9**: doi:10.1038/s41467-018-02832-w
7. Shugay M, Britanova O V., Merzlyak EM, Turchaninova MA, Mamedov IZ, Tuganbaev TR, Bolotin DA, Staroverov DB, Putintseva E V., Plevova K, et al. Towards error-free profiling of immune repertoires. *Nat Methods* (2014) **11**:653–655. doi:10.1038/nmeth.2960
8. Emerson RO, Sherwood AM, Rieder MJ, Guenthoer J, Williamson DW, Carlson CS, Drescher CW, Tewari M, Bielas JH, Robins HS. High-throughput sequencing of T-cell receptors reveals a homogeneous repertoire of tumour-infiltrating lymphocytes in ovarian cancer. *J Pathol* (2013) doi:10.1002/path.4260
9. Wu Y-CB, Kipling D, Dunn-Walters DK. Age-Related Changes in Human Peripheral Blood IGH Repertoire Following Vaccination. *Front Immunol* (2012) **3**:1–12. doi:10.3389/fimmu.2012.00193
10. Parameswaran P, Liu Y, Roskin KM, Jackson KKL, Dixit VP, Lee JY, Artiles KL, Zompi S, Vargas MJ, Simen BB, et al. Convergent antibody signatures in human dengue. *Cell Host Microbe* (2013) doi:10.1016/j.chom.2013.05.008
11. Stern JNH, Yaari G, Vander Heiden JA, Church G, Donahue WF, Hintzen RQ, Huttner AJ, Laman JD, Nagra RM, Nylander A, et al. B cells populating the multiple sclerosis brain mature in the draining cervical lymph nodes. *Sci Transl Med* (2014) doi:10.1126/scitranslmed.3008879
12. Wang C, Liu Y, Cavanagh MM, Le Saux S, Qi Q, Roskin KM, Looney TJ, Lee J-Y, Dixit V, Dekker CL, et al. B-cell repertoire responses to varicella-zoster vaccination in human identical twins. *Proc Natl Acad Sci* (2015) **112**:500–505.

doi:10.1073/pnas.1415875112
13. Hershberg U, Luning Prak ET. The analysis of clonal expansions in normal and autoimmune B cell repertoires. *Philos Trans R Soc B Biol Sci* (2015) doi:10.1098/rstb.2014.0239
14. Cortina-Ceballos B, Godoy-Lozano EE, Téllez-Sosa J, Ovilla-Muñoz M, Sámano-Sánchez H, Aguilar-Salgado A, Gómez-Barreto RE, Valdovinos-Torres H, López-Martínez I, Aparicio-Antonio R, et al. Longitudinal analysis of the peripheral B cell repertoire reveals unique effects of immunization with a new influenza virus strain. *Genome Med* (2015) doi:10.1186/s13073-015-0239-y
15. Rubelt F, Bolen CR, McGuire HM, Heiden JAV, Gadala-Maria D, Levin M, Euskirchen GM, Mamedov MR, Swan GE, Dekker CL, et al. Individual heritable differences result in unique cell lymphocyte receptor repertoires of naïve and antigen-experienced cells. *Nat Commun* (2016) **7**:1–12. doi:10.1038/ncomms11112
16. Hou XL, Wang L, Ding YL, Xie Q, Diao HY. Current status and recent advances of next generation sequencing techniques in immunological repertoire. *Genes Immun* (2016) **17**:153–164. doi:10.1038/gene.2016.9
17. Rosati E, Dowds CM, Liaskou E, Henriksen EKK, Karlsen TH, Franke A. Overview of methodologies for T-cell receptor repertoire analysis. *BMC Biotechnol* (2017) doi:10.1186/s12896-017-0379-9
18. Ruggiero E, Nicolay JP, Fronza R, Arens A, Paruzynski A, Nowrouzi A, Ürenden G, Lulay C, Schneider S, Goerdt S, et al. High-resolution analysis of the human T-cell receptor repertoire. *Nat Commun* (2015) doi:10.1038/ncomms9081
19. Friedensohn S, Khan TA, Reddy ST. Advanced Methodologies in High-Throughput Sequencing of Immune Repertoires. *Trends Biotechnol* (2017) doi:10.1016/j.tibtech.2016.09.010
20. Turchaninova MA, Davydov A, Britanova O V., Shugay M, Bikos V, Egorov ES, Kirgizova VI, Merzlyak EM, Staroverov DB, Bolotin DA, et al. High-quality full-length immunoglobulin profiling with unique molecular barcoding. *Nat Protoc* (2016) **11**:1599–1616. doi:10.1038/nprot.2016.093
21. Chaudhary N, Wesemann DR. Analyzing immunoglobulin repertoires. *Front Immunol* (2018) doi:10.3389/fimmu.2018.00462
22. Bashford-Rogers RJM, Palser AL, Idris SF, Carter L, Epstein M, Callard RE, Douek DC, Vassiliou GS, Follows GA, Hubank M, et al. Capturing needles in haystacks: A comparison of B-cell receptor sequencing methods. *BMC Immunol* (2014) **15**:1–9. doi:10.1186/s12865-014-0029-0
23. Yaari G, Kleinstein SH. Practical guidelines for B-cell receptor repertoire sequencing analysis. *Genome Med* (2015) **7**:1–14. doi:10.1186/s13073-015-0243-2
24. Egorov ES, Merzlyak EM, Shelenkov AA, Britanova O V., Sharonov G V., Staroverov DB, Bolotin DA, Davydov AN, Barsova E, Lebedev YB, et al. Quantitative Profiling of Immune Repertoires for Minor Lymphocyte Counts Using Unique Molecular Identifiers. *J Immunol* (2015) doi:10.4049/jimmunol.1500215

25. Khan TA, Friedensohn S, De Vries ARG, Straszewski J, Ruscheweyh HJ, Reddy ST. Accurate and predictive antibody repertoire profiling by molecular amplification fingerprinting. *Sci Adv* (2016) **2**: doi:10.1126/sciadv.1501371
26. de Wildt RM, van Venrooij WJ, Winter G, Hoet RM, Tomlinson IM. Somatic insertions and deletions shape the human antibody repertoire. *J Mol Biol* (1999) doi:10.1006/jmbi.1999.3289
27. Ralph DK, Matsen FA. Consistency of VDJ Rearrangement and Substitution Parameters Enables Accurate B Cell Receptor Sequence Annotation. *PLoS Comput Biol* (2016) **12**:1–25. doi:10.1371/journal.pcbi.1004409
28. Shlemov A, Bankevich S, Bzikadze A, Turchaninova MA, Safonova Y, Pevzner PA. Reconstructing Antibody Repertoires from Error-Prone Immunosequencing Reads. *J Immunol* (2017) **199**:3369–3380. doi:10.4049/jimmunol.1700485
29. Cortina-Ceballos B, Godoy-Lozano EE, Sï¿½mano-Sï¿½nchez H, Aguilar-Salgado A, Velasco-Herrera MDC, Vargas-Chï¿½vez C, Velï¿½zquez-Ramï¿½rez D, Romero G, Moreno J, Tï¿½llez-Sosa J, et al. Reconstructing and mining the B cell repertoire with ImmunediveRsity. *MAbs* (2015) doi:10.1080/19420862.2015.1026502
30. Kuchenbecker L, Nienen M, Hecht J, Neumann AU, Babel N, Reinert K, Robinson PN. IMSEQ-A fast and error aware approach to immunogenetic sequence analysis. *Bioinformatics* (2015) **31**:2963–2971. doi:10.1093/bioinformatics/btv309
31. Bolotin DA, Poslavsky S, Mitrophanov I, Shugay M, Mamedov IZ, Putintseva E V., Chudakov DM. MiXCR: Software for comprehensive adaptive immunity profiling. *Nat Methods* (2015) doi:10.1038/nmeth.3364
32. Rosenfeld AM, Meng W, Luning Prak ET, Hershberg U. ImmuneDB: A system for the analysis and exploration of high-throughput adaptive immune receptor sequencing data. *Bioinformatics* (2017) **33**:292–293. doi:10.1093/bioinformatics/btw593
33. Duez M, Giraud M, Herbert R, Rocher T, Salson M, Thonier F. Vidjil: A web platform for analysis of high-Throughput repertoire sequencing. *PLoS One* (2016) **11**:1–12. doi:10.1371/journal.pone.0166126
34. Schramm CA, Sheng Z, Zhang Z, Mascola JR, Kwong PD, Shapiro L. SONAR: A high-throughput pipeline for inferring antibody ontogenies from longitudinal sequencing of B cell transcripts. *Front Immunol* (2016) doi:10.3389/fimmu.2016.00372
35. Schaller S, Weinberger J, Jimenez-Heredia R, Danzer M, Oberbauer R, Gabriel C, Winkler SM. ImmunExplorer (IMEX): A software framework for diversity and clonality analyses of immunoglobulins and T cell receptors on the basis of IMGT/HighV-QUEST preprocessed NGS data. *BMC Bioinformatics* (2015) **16**:1–10. doi:10.1186/s12859-015-0687-9
36. Maramis C, Gkoufas A, Vardi A, Stalika E, Stamatopoulos K, Hatzidimitriou A, Maglaveras N, Chouvarda I. IRProfiler - a software toolbox for high throughput immune receptor profiling. *BMC Bioinformatics* (2018) **19**:1–11.

doi:10.1186/s12859-018-2144-z
37. Jain AK, Murty MN, Flynn PJ. Data clustering: a review. *ACM Comput Surv* (1999) doi:10.1145/331499.331504
38. Gupta NT, Adams KD, Briggs AW, Timberlake SC, Vigneault F, Kleinstein SH. Hierarchical Clustering Can Identify B Cell Clones with High Confidence in Ig Repertoire Sequencing Data. *J Immunol* (2017) **198**:2489–2499. doi:10.4049/jimmunol.1601850
39. Liu Y -J, Zhang J, Lane PJL, Chan EY-., Maclennan ICM. Sites of specific B cell activation in primary and secondary responses to T cell-dependent and T cell-independent antigens. *Eur J Immunol* (1991) doi:10.1002/eji.1830211209
40. Laserson U, Vigneault F, Gadala-Maria D, Yaari G, Uduman M, Vander Heiden JA, Kelton W, Taek Jung S, Liu Y, Laserson J, et al. High-resolution antibody dynamics of vaccine-induced immune responses. *Proc Natl Acad Sci* (2014) **111**:4928–4933. doi:10.1073/pnas.1323862111
41. Jackson KJL, Liu Y, Roskin KM, Glanville J, Hoh RA, Seo K, Marshall EL, Gurley TC, Moody MA, Haynes BF, et al. Human responses to influenza vaccination show seroconversion signatures and convergent antibody rearrangements. *Cell Host Microbe* (2014) **16**:105–114. doi:10.1016/j.chom.2014.05.013
42. Galson JD, Trück J, Fowler A, Clutterbuck EA, Münz M, Cerundolo V, Reinhard C, van der Most R, Pollard AJ, Lunter G, et al. Analysis of B Cell Repertoire Dynamics Following Hepatitis B Vaccination in Humans, and Enrichment of Vaccine-specific Antibody Sequences. *EBioMedicine* (2015) **2**:2070–2079. doi:10.1016/j.ebiom.2015.11.034
43. Hill MO. Diversity and Evenness: A Unifying Notation and Its Consequences. *Ecology* (1973) doi:10.2307/1934352
44. Chao A. Nonparametric Estimation of the Number of Classes in a Population Author. *Scanadinavian J Stat* (1984) doi:10.1214/aoms/1177729949
45. Bashford-Rogers RJM, Palser AL, Huntly BJ, Rance R, Vassiliou GS, Follows GA, Kellam P. Network properties derived from deep sequencing of human B-cell receptor repertoires delineate B-cell populations. *Genome Res* (2013) doi:10.1101/gr.154815.113
46. Apostoaei AJ, Trabalka JR. Review, synthesis and application of information on the human lymphatic system to radiation dosimetry for chronic lymphocitic leukemia. Oak Ridge TN (2010).
47. Chao A, Gotelli NJ, Hsieh TC, Sander EL, Ma KH, Colwell RK, Ellison AM. Rarefaction and extrapolation with Hill numbers: A framework for sampling and estimation in species diversity studies. *Ecol Monogr* (2014) doi:10.1890/13-0133.1
48. Laydon DJ, Melamed A, Sim A, Gillet NA, Sim K, Darko S, Kroll JS, Douek DC, Price DA, Bangham CRM, et al. Quantification of HTLV-1 Clonality and TCR Diversity. *PLoS Comput Biol* (2014) **10**:1–13. doi:10.1371/journal.pcbi.1003646
49. Laydon DJ, Bangham CRM, Asquith B. Estimating T-cell repertoire diversity: Limitations of classical estimators and a new approach. *Philos Trans R Soc B Biol*

Sci (2015) **370**: doi:10.1098/rstb.2014.0291

50. Yaari G, Vander Heiden JA, Uduman M, Gadala-Maria D, Gupta N, Joel JN, O'Connor KC, Hafler DA, Laserson U, Vigneault F, et al. Models of somatic hypermutation targeting and substitution based on synonymous mutations from high-throughput immunoglobulin sequencing data. *Front Immunol* (2013) doi:10.3389/fimmu.2013.00358
51. Yaari G, Uduman M, Kleinstein SH. Quantifying selection in high-throughput Immunoglobulin sequencing data sets. *Nucleic Acids Res* (2012) **40**:10–12. doi:10.1093/nar/gks457
52. Hou D, Chen C, Seely EJ, Chen S, Song Y. High-throughput sequencing-based immune repertoire study during infectious disease. *Front Immunol* (2016) doi:10.3389/fimmu.2016.00336
53. Tan YG, Wang XF, Zhang M, Yan HP, Lin DD, Wang YQ, Zhang HP, Yu XQ, Liao HY, Wang YP, et al. Clonal characteristics of paired infiltrating and circulating B lymphocyte repertoire in patients with primary biliary cholangitis. *Liver Int* (2018) doi:10.1111/liv.13554
54. Yang Z, Rannala B. Molecular phylogenetics: Principles and practice. *Nat Rev Genet* (2012) doi:10.1038/nrg3186
55. Shugay M, Bagaev D V., Turchaninova MA, Bolotin DA, Britanova O V., Putintseva E V., Pogorelyy M V., Nazarov VI, Zvyagin I V., Kirgizova VI, et al. VDJtools: Unifying Post-analysis of T Cell Receptor Repertoires. *PLoS Comput Biol* (2015) **11**:1–16. doi:10.1371/journal.pcbi.1004503
56. DeKosky BJ, Lungu OI, Park D, Johnson EL, Charab W, Chrysostomou C, Kuroda D, Ellington AD, Ippolito GC, Gray JJ, et al. Large-scale sequence and structural comparisons of human naive and antigen-experienced antibody repertoires. *Proc Natl Acad Sci* (2016) doi:10.1073/pnas.1525510113
57. Bolen CR, Rubelt F, Vander Heiden JA, Davis MM. The Repertoire Dissimilarity Index as a method to compare lymphocyte receptor repertoires. *BMC Bioinformatics* (2017) **18**:1–8. doi:10.1186/s12859-017-1556-5
58. Vander Heiden JA, Yaari G, Uduman M, Stern JNH, O'connor KC, Hafler DA, Vigneault F, Kleinstein SH. PRESTO: A toolkit for processing high-throughput sequencing raw reads of lymphocyte receptor repertoires. *Bioinformatics* (2014) doi:10.1093/bioinformatics/btu138
59. Gadala-Maria D, Yaari G, Uduman M, Kleinstein SH. Automated analysis of high-throughput B-cell sequencing data reveals a high frequency of novel immunoglobulin V gene segment alleles. *Proc Natl Acad Sci* (2015) **112**:E862–E870. doi:10.1073/pnas.1417683112
60. Gupta NT, Vander Heiden JA, Uduman M, Gadala-Maria D, Yaari G, Kleinstein SH. Change-O: A toolkit for analyzing large-scale B cell immunoglobulin repertoire sequencing data. *Bioinformatics* (2015) **31**:3356–3358. doi:10.1093/bioinformatics/btv359
61. Chao A, Hsieh TC, Chazdon RL, Colwell RK, Gotelli NJ, Inouye BD. Unveiling

the species-rank abundance distribution by generalizing the Good-Turing sample coverage theory. *Ecology* (2015) doi:10.1890/14-0550.1

62. Moorhouse MJ, van Zessen D, IJspeert H, Hiltemann S, Horsman S, van der Spek PJ, van der Burg M, Stubbs AP. ImmunoGlobulin galaxy (IGGalaxy) for simple determination and quantitation of immunoglobulin heavy chain rearrangements from NGS. *BMC Immunol* (2014) **15**:1–9. doi:10.1186/s12865-014-0059-7
63. Christley S, Scarborough W, Salinas E, Rounds WH, Toby IT, Fonner JM, Levin MK, Kim M, Mock SA, Jordan C, et al. VDJServer: A cloud-based analysis portal and data commons for immune repertoire sequences and rearrangements. *Front Immunol* (2018) **9**:1–12. doi:10.3389/fimmu.2018.00976
64. Yu Y, Ceredig R, Seoighe C. LymAnalyzer: A tool for comprehensive analysis of next generation sequencing data of T cell receptors and immunoglobulins. *Nucleic Acids Res* (2015) **44**: doi:10.1093/nar/gkv1016
65. Ralph DK, Matsen FA. Likelihood-Based Inference of B Cell Clonal Families. *PLoS Comput Biol* (2016) **12**:1–28. doi:10.1371/journal.pcbi.1005086
66. Lees WD, Shepherd AJ. Utilities for High-Throughput Analysis of B-Cell Clonal Lineages. *J Immunol Res* (2015) **2015**: doi:10.1155/2015/323506
67. Giraud M, Salson M, Duez M, Villenet C, Quief S, Caillault A, Grardel N, Roumier C, Preudhomme C, Figeac M. Fast multiclonal clusterization of V(D)J recombinations from high-throughput sequencing. *BMC Genomics* (2014) **15**:1–12. doi:10.1186/1471-2164-15-409
68. Ferret Y, Caillault A, Sebda S, Duez M, Grardel N, Duployez N, Villenet C, Figeac M, Preudhomme C, Salson M, et al. Multi-loci diagnosis of acute lymphoblastic leukaemia with high-throughput sequencing and bioinformatics analysis. *Br J Haematol* (2016) doi:10.1111/bjh.13981
69. Ye J, Ma N, Madden TL, Ostell JM. IgBLAST: an immunoglobulin variable domain sequence analysis tool. *Nucleic Acids Res* (2013) **41**:34–40. doi:10.1093/nar/gkt382
70. Brochet X, Lefranc MP, Giudicelli V. IMGT/V-QUEST: the highly customized and integrated system for IG and TR standardized V-J and V-D-J sequence analysis. *Nucleic Acids Res* (2008) **36**:503–508. doi:10.1093/nar/gkn316
71. Russ DE, Ho KY, Longo NS. HTJoinSolver: Human immunoglobulin VDJ partitioning using approximate dynamic programming constrained by conserved motifs. *BMC Bioinformatics* (2015) **16**:1–11. doi:10.1186/s12859-015-0589-x
72. Zhang W, Wang IM, Wang C, Lin L, Chai X, Wu J, Bett AJ, Dhanasekaran G, Casimiro DR, Liu X. IMPre: An accurate and efficient software for prediction of T- and B-cell receptor germline genes and alleles from rearranged repertoire data. *Front Immunol* (2016) **7**: doi:10.3389/fimmu.2016.00457
73. Corcoran MM, Phad GE, Bernat NV, Stahl-Hennig C, Sumida N, Persson MAA, Martin M, Hedestam GBK. Production of individualized v gene databases reveals high levels of immunoglobulin genetic diversity. *Nat Commun* (2016) **7**: doi:10.1038/ncomms13642

74. Busse CE, Czogiel I, Braun P, Arndt PF, Wardemann H. Single-cell based high-throughput sequencing of full-length immunoglobulin heavy and light chain genes. *Eur J Immunol* (2014) doi:10.1002/eji.201343917
75. Jones JOJPSSP. *Kuby Immunology*. (2013). doi:10.1017/CBO9781107415324.004
76. Li A, Rue M, Zhou J, Wang H, Goldwasser MA, Neuberg D, Dalton V, Zuckerman D, Lyons C, Silverman LB, et al. Utilization of Ig heavy chain variable, diversity, and joining gene segments in children with B-lineage acute lymphoblastic leukemia: Implications for the mechanisms of VDJ recombination and for pathogenesis. *Blood* (2004) doi:10.1182/blood-2003-11-3857

Table Legend

Table 1. The Pipeline repertoire.

The most common features for each pipeline are written into each square.
Check symbols, a feature performed by the corresponding tool but there is scarce or no detailed information; empty squares, the corresponding feature is not included in the specific tool.

Box 1

VDJ recombination

Antibodies are produced by a developmentally ordered series of somatic gene rearrangement events that occur exclusively in developing lymphocytes. Antibodies consist of heavy (μ, alfa, δ, γ, ε) and light chains (κ, λ), each of which contains a variable and a constant domain. Antigen binding occurs in the variable domain, which is comprised of one variable (V), one diversity (D) and one joining (J) gene segment in heavy chains and one variable (V) and one joining (J) segment in light chains. The germline V(D)J genes contain approximately 41 different V segments, 23 different D segments and 6 different J segments (75,76). However, during B cell maturation —more specifically, when the precursor B cell matures into a naive B cell— the segments are reduced to only one V, one D and one J gene segment. This process is called somatic recombination (VDJ recombination) and provides combinatorial diversity to the antibodies. Furthermore, the recombination process often results in non-templated mutations like the addition or deletion of nucleotides at the junctions between ligated gene segments. The site of the VDJ gene segment ligation, also known as the complementarity-determining-region 3 (CDR-H3), is the most diverse component in terms of length and sequence of the antibody heavy chain (1).

Box 2

Somatic Hypermutation (SHM)

When the immune system encounters an antigen for the first time, T-cell helpers will stimulate naive B cells to mature into antigen specific B cells. The maturation process consists of a B cell clonal expansion followed by somatic hypermutation. During SHM, activation-induced cytidine deaminase (AID) introduces non-templated mutations into the variable region of the antibody genes (1). AID also mediates class switch recombination, which generates antibodies bearing different constant regions (75). B cells expressing somatically mutated, high-antigen-affinity BCRs undergo preferential expansion and survival, a processed referred to as *affinity maturation*. As a result, B cells

bearing the highest-affinity antibodies differentiate into plasma cells, or long-lived memory B cells capable of mediating rapid recall responses to the same antigen.

Both, VDJ recombination and SHM, introduce non-templated mutations into the Ig genes, which make the Ig-Seq data pre-process and analysis intricate. Once the Ig-Seq is performed, the resulted reads will not align perfectly to the germline and the misalignments could be due to either the natural maturation process of the antibodies or to PCR and sequencing errors.

Table 1. The pipeline repertoire.

			BROAD SPECTRUM PIPELINES				
			IgRec	ImmunediveRsity	Immcantation Framework	IGGalaxy	VDJServer
BASIC PROCESSING	PREPROCESSING	**Amplification methods supported**		‣ RACE	**pRESTO** ‣ Multiplex PCR ‣ RACE ‣ Can perform demultiplexing		‣ Access to pRESTO or VDJPipe
		Sequencing supported	‣ Single-end ‣ Paired-ends	‣ Single-end ‣ Paired-ends	‣ Single-end ‣ Paired-ends (overlapping or not)		
		Quality filtering		‣ Read quality	‣ Read quality	✓	
		Amplification error correction	‣ UMI based correction ‣ Amplification correction without the need for UMI	‣ Amplification correction without the need for UMI	‣ UMI based correction		
	V(D)J ALIGNMENT	**V(D)J ALignment**	VJ	VJ	**TIgER** ‣ Takes already aligned reads and corrects it detecting new alleles	‣ Wraps IgBlast ‣ Wraps IMGT HighV-QUEST	‣ Wraps IgBlast
	CLONAL GROUPING	**Clonotype definition(s)**	‣ Grouping by hamming graphs	‣ Unique antibodies	**Change-O** ‣ VJ + CDR3 (common ancestor) ‣ VJ + CDR3-AA	‣ VJ CDR-AA ‣ VJ CDR nucleotide	‣ Wraps Change-O and Alakazam
		Removes nonfunctional reads			✓	✓	
		Lineage Tree	‣ Lineage tree and Hamming graph	‣ Network graphical representation	**Alakazam**		
REPERTOIRE CHARACTERIZATION AND ANALYSIS	DIVERSITY	**Basic Characterization**	**IgDiversityAnalyzer** ‣ Relative/absolute frequency of clonotypes		**Alakazam** ‣ Physicochemical properties of AA sequences ‣ V(D)J allele/ gene usage		‣ Wraps Alakazam
		Diversity Index	**IgDiversityAnalyzer** ‣ Simpson Index (SI) ‣ Clonal Simpson Index (CSI) ‣ Ratio SI/CSI	‣ Gini index ‣ Shannon Weiner index ‣ Shannon normalized	‣ Species richness ‣ Shannon Weiner ‣ Inverse simpson ‣ Berger Parker ‣ Total diversity estimation: Chao1		
		PLOTS	**IgDiversityAnalyzer** ‣ CDR3 length distribution ‣ CDR3 nucleotide content per position ‣ CDR3 aminoacid content ‣ VJ usage heatmap	‣ Histogram of CDR3 length distribution ‣ Proportion of AA in CDR3 ‣ AA distribution of CDR3 ‣ VJ use heatmap ‣ Rarefaction curve	‣ Clonal abundance distribution ‣ V(D)J allele/gene usage graphs ‣ Rarefaction curve	‣ VJ VD and DJ gene usage heatmaps	
	MUTATION ANALYSIS	**Mutation Analysis**	**IgDiversityAnalyzer** ‣ Identifies mutation type (substitution, insertion, deletion) ‣ V region mutation and AA change table ‣ Identifies synonym or stop codon mutations	‣ Identifies synonymous and non synonymous mutations	**SHazaM** ‣ Quantification of mutational load ‣ Statistical models of SHM targetting patterns ‣ Tools to build SHM targeting models from data ‣ Analysis of selection pressure using BASELINe		‣ Wraps SHazaM
		PLOTS	**IgDiversityAnalyzer** ‣ SHM vs relative position ‣ Number of reads vs SHM ‣ Distribution of indel SHMs positions ‣ Distribution of indel SHMs lengths ‣ Distribution of synonymous SHMs positions ‣ Heatmaps of AA or nucleotide substitutions		‣ Mutability frequencies ‣ Visualization of SHM targeting model ‣ Probability density functions resulting from selection analysis ‣ Plot summary stastistics resulting from selection analysis		
	EVOLUTION	**Clonal dynamics**					
	CONVERGENCE	**Convergance of Repertoire**		✓	**RDI**		
	OTHER						

Table 1. The pipeline repertoire; continued.

			MODULAR PIPELINES						
			MiGEC	IMSEQ	MiXCR	LymAnalyzer	Partis	ImmuneDB	Vidjil
BASIC PROCESSING	PREPROCESSING	**Amplification methods supported**	‣ Multiplex PCR ‣ Can perform demultiplexing		‣ RACE ‣ RNA seq				
		Sequencing supported	‣ Paired-ends	‣ Single-end ‣ Paired-ends	‣ Single-end ‣ Paired-ends				
		Quality filtering	‣ Cluster average quality	‣ Read quality filtering ‣ Group quality filtering					
		Amplification error correction	‣ UMI based correction	‣ Amplification correction without the need for UMI					
	V(D)J ALIGNMENT	**V(D)J ALignment**		VJ	VDJ + C	VDJ	VDJ	VJ	VJ
	CLONAL GROUPING	**Clonotype definition(s)**	‣ Unique antibodies	‣ Unique antibodies	‣ Unique antibodies ‣ Can group by C		‣ Common ancestor	‣ VJ+ CDR3-AA	‣ Unique antibodies
		Removes nonfunctional reads							
		Lineage Tree							
REPERTOIRE CHARACTERIZATION AND ANALYSIS	DIVERSITY	**Basic Characterization**				‣ CDR3 extraction and characterization		‣ V length distribution ‣ CDR3 length distribution ‣ V, J usage	
		Diversity Index							‣ Shannon-Wiener ‣ Inverse Simpson
		PLOTS							
	MUTATION ANALYSIS	**Mutation Analysis**					‣ Mutation frequencies ‣ Identification of non-templated insertions		
		PLOTS				‣ Mutation tree (for each CDR3 one can visualize the hypermutation process)			
	EVOLUTION	**Clonal dynamics**							
	CONVERGENCE	**Convergance of Repertoire**							
	OTHER								

Table 1. The pipeline repertoire; continued.

			SPECIALIZED PIPELINES						
			TRiGS	IMEX	IRProfiler	VDJtools	DIVE	BASELINe	AbSim
BASIC PROCESSING	PREPROCESSING	Amplification methods supported							
		Sequencing supported							
		Quality filtering			‣ 11 different quality filtering criteria	‣ Frequency based correction ‣ Frequency based filtering			
		Amplification error corrcetion							
	V(DJ ALIGNMENT	V(D)J ALignment							
	CLONAL GROUPING	Clonotype definition(s)	‣ VJ + CDR3 (Common ancestor) ‣ VJ + CDR3-AA	‣ VJ + CDR3 ‣ VJ + CDR3-AA	‣ VDJ + CDR3-AA ‣ VJ+ CDR3-AA ‣ V + CDR3-AA ‣ J + CDR3 ‣ CDR3				
		Removes nonfunctional reads							
		Lineage Tree							
REPERTOIRE CHARACTERIZATION AND ANALYSIS	DIVERSITY	Basic Characterization		‣ Number of productive / unproductive reads ‣ VDJ gene total and relative frequencies	‣ Clonotype quantification ‣ V, J gene usage	‣ Read counts ‣ Mean clonotype sizes ‣ Non-functional reads ‣ Histogram of CDR3 nucleotide length			
		Diversity Index	‣ Gini index	‣ IMEX calculates sequence diversity using a more elaborated data mining approach.		‣ Shannon-Wiener ‣ Inverse Simpson ‣ D50 index, a recently developed immune diversity estimate ‣ Lower bound total diversity: Chao1 or Efron-Thisted estimate	‣ DiVE speciess richness estimate (fits different models and selects the top 5 model to estimate speciess richness)		
		PLOTS	‣ CDR3 length distribution ‣ Histograms for VJ gene usage	‣ VDJ usage (heatmaps or histogram) ‣ Productive/ non productive pie charts ‣ Diversity curve		‣ V, J, VJ usage (heatmap and hierarchical clustering) ‣ Clonality plot (Plots a three-layer donut chart to visualize the repertoire clonality) ‣ Rarefaction plot	‣ Rarefaction plot		
	MUTATION ANALYSIS	Mutation Analysis						‣ Identifies point and silent mutations ‣ Identifies and quantifies positive or negative selection ‣ Germline normalization for direct comparisson	
		PLOTS							
	EVOLUTION	Clonal dynamics							
	CONVERGENCE	Convergance of Repertoire							
	OTHER								Generates simulations of Ig reprrtoires with the desired characteristics