ViCloD, an interactive web tool for visualizing B cell repertoires and analyzing intra-clonal diversities: application to human B-cell tumors

Lucile Jeusset[1], Nika Abdollahi[1], Thibaud Verny[3], Marine Armand[2], Anne Langlois De Septenville[2], Frédéric Davi[2,*] and Juliana S. Bernardes[1,*]

[1]Sorbonne Université, CNRS, UMR 7238, Laboratoire de Biologie Computationnelle et Quantitative, Paris, France

[2]Sorbonne Université, AP-HP, Hôpital Pitié-Salpêtrière, Department of Biological Hematology, Paris, France

[2]Ecole des Mines ParisTech, Paris, France

[*]To whom correspondence should be addressed; E-mail: juliana.silva_bernardes@sorbonne-universite.fr

High throughput sequencing of adaptive immune receptor repertoire (AIRR-seq) has provided numerous human immunoglobulin (IG) sequences allowing specific B cell receptor (BCR) studies such as the antigen-driven evolution of antibodies (IG secreted molecules). AIRR-seq data allows researchers to examine intra-clonal differences caused primarily by affinity maturation and somatic hypermutation processes. Understating the diversity of antibodies and how they evolve could help elucidate the generation of antibodies with high affinity or broadly neutralizing activities. Consequently, retracing their evolutionary history could also help to clarify how vaccines or pathogen exposition drive the humoral immune response. Computational methods are necessary for large-scale evaluation of AIRR-seq properties. Currently, there is no efficient and interactive tool for analyzing intra-clonal diversity, permitting users to explore AIRR-seq data and potentially discover discriminating clonal features. We developed ViCloD, a web server for large-scale visual analysis of repertoire clonality and intra-clonal diversity. ViCloD uses data preprocessed by IMGT and performs evolutionary and statistical analyses, producing a set of valuable plots. The web server presents diverse functionalities, including repertoire navigation, clonal abundance analysis, and intra-clonal evolutionary tree reconstruction. Users can download the analyzed data in different table formats and save the generated plots as image files. ViCloD is a versatile analytical tool that can help researchers and clinicians in their investigations involving clonal diversity analysis. It is user-friendly, simple, accessible, and open to all users. Moreover, its pipeline is optimized to process hundreds of thousands of sequences in a few minutes, allowing intra-clonal diversity analysis in large and complex repertoires.

Availability and implementation: The web server is available at http://www.lcqb.upmc.fr/viclod/. The pipeline is available at github and allows users to run analyses locally.

key words: B-cell immune repertoire, intra-clonal diversity, B cell lineage trees, immunoinformatics

1 Introduction

B cells are critical components of the adaptive immune system, recognizing and eliminating various foreign molecules (antigens). B cells recognize antigens through a specific protein complex, the B cell Receptor (BCR) attached to the cell surface. The BCR is made up of a recognition unit formed by an immunoglobulin (IG) protein and an associated signaling unit, the Igα an Igβ molecules. Besides this membrane-bound form, IG can also be secreted in the blood plasma as circulating antibodies.

The immune system can respond to almost any antigen to which it is exposed, partially due to an extremely diverse set of BCRs. Such diversity lies in the variable region of IG and results from complex genetic mechanisms that occur during B cell ontogeny. This process, called VDJ recombination (*1*), joins different types of genes contributing to the variable region : V (variable), D (diversity) and J (joining). The resulting combinatorial diversity is further enhanced by a junctional diversity due to the imprecise joining of these genes with deletion and insertion of random nucleotides at the recombination sites, thereby creating an extremely diverse zone within the variable region, the complementary determining region 3 (CDR3) (*2*). Further diversity occurs after the antigen encounter, with a process called affinity maturation which occurs in specialized structures of secondary lymphoid organs (*3*). There, point mutations are introduced in the variable region of IG genes at very high rates, about 10^6 times higher than observed for other genes, by a mechanism termed somatic hypermutation (SHM) (*4*), a robust protective response against the antigens in case of infection. This phenomenon is known as clonal expansion and plays a fundamental role in an efficient immune response (*5*). Therefore, the generation and selection of somatic antibody mutants are decisive for increasing antigen affinity, contributing to the B cell immune repertoire diversity.

The adaptive immune receptor repertoire sequencing (AIRR-seq) has greatly contributed to characterize immune repertoires, potentially revealing the high-dimensional complexity of the immune receptor sequence landscape. AIRR-seq data is often mined to address many scientific and clinical questions. An extremely diversified BCR repertoire reflects the expected diversity in healthy individuals. However, such diversity can be significantly affected by different factors such as autoimmune diseases (*6–12*), allergy (*13–15*), cancer (*16, 17*), and aging (*18, 19*). The exploration of AIRR-seq

data is thus an active field of research, mainly to analyse BCR composition and diversification. Analyzing mutational sequence patterns of a given clonal lineage can provide insights into the diversification and selection processes that lead to B cell clonal evolution. The examination of (intra-)clonal compositions can also help understanding the response to infections (*20*), vaccines (*21,22*), the mechanisms of autoimmune diseases (*23*), allergy (*24*), among others. Rapid advances in high-throughput sequencing have opened up new possibilities for studying B cell clonal expansion and (intra-)clonal compositions. With decreasing costs of DNA sequencing technology, numerous IG sequences with inherent complexity, variability, and mutational load are accessible, including in clinical contexts. Thus, the availability of AIRR-seq data has motivated researchers with different backgrounds (biological, computational, and statistical) to investigate and examine the adaptive immune complexity.

Several methods and computational tools have been developed to treat different AIRR-seq analysis steps, producing multiple integrated context-specific software (*25–27*). Most of the tools were designed for V(D)J assignments (*28*), and identification of groups of clonally-related sequences (*29*). Some solutions exist to tackle and determine the heterogeneity of BCR repertoires (*30,31*). Such tools quantify the diversity of a BCR repertoire by estimating clonal group sizes, characterizing CDR3 regions, identifying the most used V, D, and J genes (alleles), and the most frequent VJ combinations. Indeed, existing tools provide a basic characterization and essential statistics of immune repertoires, but an (intra-)clonal diversity/evolution analysis is not direct. Such analyses can potentially reconstruct antibodies' evolutionary (clonal) development and identify mutable positions that correlate with antigen-binding sites. Few methods have been proposed to retracing the antigen-driven evolution of BCR repertoires (*32–34*), and just one recent publication provides interactive displays of B cell evolutionary trees (*35*). However, no large-scale, time-efficient bioinformatic tool allows a deep analysis of (intra-)clonal diversity and evolution.

We propose ViCloD, a web-based interactive tool for visualizing B cell receptor repertoire clonality and analyzing (intra-)clonal diversity in B cell tumors. ViCloD allows entire repertoire analyses and proposes a flexible repertoire topology panorama. Users can navigate among B-cell clonal lineages and explore their intra-diversity. For that purpose, we have developed a set of algorithms able to

analyze millions of IG sequences within a few minutes (*29,36*). Such algorithms are time efficient and compatible with research and clinical environment. Users can download all analyzed data, save the generated plots, perform their analyses locally, and visualize them in the ViCloD webserver. ViCloD provides a user-friendly data visualization to help the research/medical community select helpful information in BCR repertoire analyses. It allows non-computational experts to analyze (intra-)clonal diversities that could boost their research within the immune repertoire investigation. ViCloD provides unique functionalities absent in other webservers dedicated to AIRR data processing, analysis, and visualization. VICloD is a versatile AIRR-seq tool with the potential to help experimental/clinical investigations of immune repertoires. To demonstrate its functionalities and potential, we applied it to repertoires of human B-cell tumors. However, ViCLoD can be used to investigate B-cell clonal expansion and intra-clonal compositions of any repertoire.

2 Material and Methods

2.1 ViCloD workflow

The ViCLoD workflow is presented in Figure 1; it consists of five main steps: user's input, pipeline, repertoire composition, clonal composition, and lineage architecture, described below.

2.1.1 User inputs

ViCloD can only analyze BCR Immunoglobulin Heavy Chain (IGH) repertoires. In practice, IGH is frequently used as a signature for BCR sequences since it is more diversified than IGL chains. IGH sequences contain the CDR3 region that plays a key role in antigen recognition (*2*). After sequencing, IGH sequences should be annotated to determine their IGHV, IGHD, IGHJ genes and localize their CDR3 region. ViCloD accepts data from IMGT/HighV-QUEST (*37*), the international standard web software for parsing BCR sequences, see Figure 1-A. In principle, the output of any V(D)J annotation tool could be used as long as it provides the minimum information necessary for the analysis, see `http://www.lcqb.upmc.fr/viclod/Format`. ViCloD requires AIRR formatted input files (*38*) that facilitate the use of different tools. Users can upload the AIRR file on the main page of the

ViCLoD webserver and provide their email addresses to receive a link to their outputs. The analysis is relatively fast; for instance, it can take around three minutes to analyze a monoclonal repertoire containing around 300K sequences. Indeed, the required time can vary depending on the structure of the sample (the size/amount of clonal lineages and sequence mutations). Alternatively, users can upload ViCloD formatted files, processed and downloaded from a previous analysis. These files were generated either by the web server or by the stand-alone version. Thus, users can store and re-analyze their data without any time constraints.

2.1.2 Pipeline

Once IGH sequences are uploaded, the first step is to group them into clonally-related sequences. We use MobiLLe (*29*), a multi-objective-based clustering for inferring BCR clonal lineages from high-throughput AIRR-seq data. Then, within each BCR clonal lineage, detected by MobiLLe, we group sequences with the same IGHV, IGHJ genes, and identical CDR3 amino acid sequences, forming *sub-clones*. Note that we adopted the same sub-clone definition as IMGT group (*39*). For each clonal BCR lineage, we try to reconstruct its evolutionary history by using the N most abundant sub-clones. We used ClonalTree (*36*), a time-efficient and accurate approach to reconstructing BCR lineage trees that incorporates genotype abundances into minimum spanning tree algorithms. ClonalTree requires a file in FASTA format containing sequences and their abundances as input. Each sequence in the input file represents a sub-clone. Since sequences in a given sub-clone can differ due to SHM, we chose the most abundant sequence to represent it. The sub-clone abundance is the total number of sequences within it. ClonalTree also requires the hypothetical ancestral unmutated sequence representing the root tree. To create it, we take the IGHV and IGHJ germline sequences provided by IMGT/HighV-QUEST. For the junction part, we consider the sequence with the lowest number of mutations on the IGHV gene compared to the germline. Then, we concatenated the three parties (IGHV, IGHJ, and junction region) to obtain the hypothetical ancestral unmutated sequence. The pipeline is illustrated in Figure 1-B.

2.1.3 Repertoire composition

We used nested circles to represent BCR clonal lineages within a repertoire, Figure 1-C. The outer circle represents the entire BCR repertoire, while internal circles correspond to BCR clonal lineages. The size of each circle correlates to the BCR lineages' abundance. A bar plot on the same web page shows the abundance of the ten most abundant sub-clones within a given BCR clonal lineage. Users can choose between a standard or logarithmic scale; they can also select a threshold of sub-clone abundance for analyzing the clonality of a repertoire, for instance, above 5%. When such a threshold is set, only BCR clonal lineages with an abundance greater than this value are retained in the BCR repertoire visualization; all others are hidden. A table at the bottom of the page summarizes some essential properties of each inferred BCR clonal lineage, such as the identifier, the abundance, the IGHV/IGHJ gene annotations, the CDR3 amino acid sequence and the productivity status of the most abundant sub-clone sequence within the BCR lineage.

parei aqui

2.1.4 Clonal composition

Users can explore sub-clone compositions by clicking on one of the circles representing a BCR clonal lineage. The circle expands and other circles, representing sub-clones, appear (see dotted arrow in Figure 1-CD). The circle area correlates with the sub-clone abundance within the BCR clonal lineage; see Figure 1-D. For a given set of sub-clones (the most abundant), we also show a circular B-cell lineage tree generated by ClonalTree, representing the evolutionary relationships among such sub-clones and the ancestral sequence. Several options for tree representation are available. For instance, nodes can be colored by sub-clones, a different color for each one, or by functionality, where green and red nodes represent productive and unproductive rearrangements, respectively. Note that unproductive sequences contain stop codon(s). It is also possible to display the length of tree branches representing distances between sequences. We highlight the hypothetical naive sequence by a triangle and the most abundant sub-clone by a square to facilitate tree interpretation. At the bottom of the page, we displayed additional information about sub-clones, such as their identifiers, abundances, and CDR3

amino acid sequences.

The lineage tree inferred by the ClonalTree algorithm represents (at most) 200 of the most abundant sub-clones. Interpreting the complete lineage tree can be challenging for large groups containing many sub-clones. Thus, we suggest two simplified tree topologies while conserving nodes with high abundance and their evolutive path to the root. We developed two strategies for pruning the tree (*40*) for a more comprehensible visualization, Figure S1. These simplified versions may be more valuable than an entirely accurate description that involves many details in some situations. The first pruning strategy transforms the tree shown in Figure S1-A in the tree of Figure S1-B; it eliminates nodes with no descendant or an abundance lower than a predefined threshold. This approach is more helpful for relatively small trees. Trees with many sub-clones can require the second pruning strategy that eliminates nodes with low abundance if they present high similarity with more abundant sub-clones (Figure S1-BC). For both strategies, we eliminate just leaf nodes and keep the N most abundant sub-clones in the simplified tree (by default, N=30). We try to preserve the tree's evolutionary paths by eliminating only leaves' nodes, which could be destroyed if an internal node is removed. We keep the N largest sub-clones since they contain relevant information to analyze the lineage tree. We apply the first strategy once, and on the resulting tree, we repeat the second one several times until achieving at most 30 nodes. Users can navigate between the complete lineage tree and pruned trees on the website.

2.1.5 Lineage architecture

The last ViCLoD module provides more details about the B-cell lineage inferred tree, Figure 1-E. Users can explore the simplified tree to examine intra-clonal diversities. ViCloD provides four types of representations: an interactive tree, a circular bar chart (mutational profile), a CDR3 logo, and a multiple sequence alignment.

Interactive lineage tree

We kept the most pruned tree with at least ten most abundant sub-clones to better examine the lineage tree. Colored circles represent observed sub-clones, while white circles represent unobserved

nodes (cf Clonaltree algorithm). A triangle represents the hypothetical germline sequence. The branch length represents the number of somatic hypermutations between the representative sequences of connected sub-clones or the hypothetical germline sequence. We identify sub-clones by a number, sorted by their abundance, which means that sub-clone 1 is the most abundant in the B-cell lineage. Nodes representing the five largest sub-clones have a bold border in the tree and are enumerated from 1 to 5. Sub-clone details (identifier, abundance, and functionality) appear when passing the mouse over nodes. The option ”display abundance” displays each sub-clone abundance by increasing the node's diameter. In this visualization mode, nodes are labeled with their abundance (%) rather than with a sequential number. Each sub-clone may contain multiple unique sequences (clonotypes). Choosing ”Display proportion of unique sequence” will turn each node into a pie chart, illustrating the numerical proportion of unique sequences within the subclone.

Circular bar chart - Mutational profile

The circular bar chart represents the distances (number of somatic hypermutations) between the ancestral (the hypothetical germline sequence) and a set of selected sub-clone representative sequences; we also termed this plot a mutational profile. By default, ViCloD displays the five most abundant sub-clones, but any sub-clone can be included and/or removed from the plot; a maximum of eight items are displayed. Each colored section is related to a sub-clone and represents the number of mutations observed between a given sub-clone and its parent in the tree. To highlight the tree path from a sub-clone to the root, users should hover the mouse over their desired sub-clone identifiers; to highlight a branch in the tree, one can hover the mouse over each section of the circular bar; it also displays the number of mutations between a pair of connected sub-clone.

Multiple sequence alignment

To illustrate the conservations/mutations between the representative sequence of each sub-clone and the hypothetical ancestral sequence, we build a multiple sequence alignment with the MUSCLE program (*41*). For each aligned sequence, we display in a table: the identifier, the percentage and number

of reads, the divergence rate (number of mutations) from the hypothetical naive sequence, and the percent deviation of IGHV sequence from the germline. Only mutated nucleotides of sub-clone representative sequence compared to the hypothetical ancestor are shown in the alignment, whereas a dot represents conserved positions. CDR and framework sections are highlighted with different colors, and the IGHD gene is underlined in each sequence; see an example in Figure 5. Users can sort out sequences in the table based on each column. The table can be downloaded into a text file by clicking on the download button. It is also possible to select one, multiple, or all sequences and send them directly to the IMGT/V-quest software.

CDR3 logo

Sequence logos provide a more detailed and precise description of sequence similarities and are frequently used to unravel significant features that are harder to visualize in the multiple sequence alignments. In the Lineage architecture page, users can generate a sequence logo of the CDR3 by clicking on the ”CDR3 Logo” button. The logo is split into two sections; the upper one shows the most frequent residues that differ from the germline sequence, while the lower section shows the residues of the germline. The height of the stack letter represents the frequency of each residue in the alignment, showing how many the germline was mutated in the clonal lineage; see an example in Figure 4.

2.2 Implementation

The server-side pipeline was implemented in python (version 3) and Bash script. We also use the MySql database (https://www.mysql.com/) to control the user's pending jobs. The client-side web interface was implemented using HTML5, CSS, and JavaScript (https://www.javascript.com/) along with several libraries including jQuery (http://jquery.com/), jQuery-UI (https://jqueryui.com/) for user interface, and d3.js (http://d3js.org/). PHP (version 5.3.3, http://php.net/) was used for interactive data processing. Users submit their input files and can provide an email address to receive their results. We first check if the input file contains all the required information. Then, data are processed by the ViCloD pipeline, and all analyses are directly available or by clicking on the link sent by email.

Users' files (inputs and outputs) are available for download for a while and can be retrieved during this time. Users can also download JSON files that can be input into the ViCLoD web server later. It enables data to be stored locally and re-analyzed without time constraints.

2.3 Tutorials

The "Tutorial" section of the website includes a variety of text-based and video tutorials to help users get started with ViCloD. Multiple analyzed data sets are provided on the "Example" page to assist users in following the tutorials step by step. Several sections under the "About tab" provide additional information, such as references, algorithms used, and the server's versions.

3 Results

BCR immune repertoire studies are having a significant impact in a variety of research and clinical areas. Several works have reported intra-clonal diversities when analyzing BCR repertoires of patients with different lymphoproliferative disorders. In the following, we show how ViCLoD can help examine patients' BCR repertoires, providing a set of visualizations for exploratory analysis.

3.1 Clonal heterogeneity of a patient with Follicular Lymphoma suspicion

Follicular Lymphoma (FL) is lymphatic cancer that occurs when white blood cells (B-lymphocytes) cluster together and form lumps in lymph glands or other organs. FL is a clinically heterogeneous disease; some patients have a high tumor burden with widely disseminated disease at diagnosis, while others have a low tumor burden with no symptoms for many years REF-Fred. Treatment response is unstable and related to the disease's biological heterogeneity. Analyzing intra-clonal diversity can provide insight into lymphoma development and may account for FL progression/transformation REF-Fred.

We used the IGH repertoire of a patient with FL suspicion to demonstrate all types of analyses provided by ViCLoD. A total of 186K sequences from the patient were inputted into ViCLoD web-server, but 85K ($\approx$ 45%) were analyzed after quality control, see Section 2.1.1. Figure 2-A shows

the repertoire composition of the patient, where we remarked the presence of abundant clonal lineage that accounts for more than 91% of the repertoire, annotated with IGHV3-23*01 and IGHJ4*02 gene segments. Such clonal proliferation is common in lymphoproliferative disorders. ViCLoD allows users to examine the five most abundant clonal proliferation, labeled in Figure 2-A. The second most abundant clonal lineage accounts for less than 0.3% of the repertoire; it has unproductive IGH sequences annotated with the same IGHV/IGHJ genes as the previous one. The third and fourth clonal lineages also present a weaker abundance (less than 0.2% of the repertoire), they are also annotated with the same gene segments, but their CDR3 regions are very different, see Table S1.

Users can click on a given clonal lineage to inspect its sub-clonal composition. Figure 2-B shows sub-clones for the most abundant clonal lineage C1. We labeled only the seven most abundant sub-clones, but a total of 652 were detected, indicating a high intra-clonal diversity for C1 lineage. Sequences from almost all sub-clones are productive; the only exception is the sub-clone C1-98 (0.026% of the repertoire), which contains a codon stop in the CDR3 region. Sub-clones C1-1 and C1-2 are the most abundant of the C1 lineage, grouping 56% and 20% of sequences, respectively. We also observed intra-clonal diversity for other clonal lineages; more details are available on the Example tab, labeled with Example 1, at ViCLoD website.

ViCLoD also provides lineage architecture details for a given clonal lineage. Figure 3 shows the lineage architecture and mutational profile for the clonal lineage C1. In the tree, Figure 3-A, the reconstructed germline sequence is indicated by a triangle, while the most relevant sub-clones are by circles. The circle area is proportional to the sub-clonal abundance, and the pie charts indicate the sub-clonal composition: the number of unique sequences, also called clonotypes. For instance, the sub-clone C1-1 is composed of 5401 different clonotypes, while C1-2 of 1794; see clonotype distributions for the five most abundant sub-clones in Figure S2. The tree topology confirms the high intra-clonal diversity of lineage C1. The reconstructed tree has several levels, revealing the intricate evolutionary relationships of the most abundant/relevant sub-clones. According to the tree, the most abundant sub-clone C1-1 gave origin to six sub-clones, and the second abundant C1-2 has evolved from a less abundant sub-clone C1-6. The sub-clone C1-19, which accounts for less than 0.18% of

the clonal lineage C1, is the most conserved, with a lower mutation rate (19 nucleotide differences) compared to the reconstructed germline. The longest path in the tree is until the sub-clone C1-193, which groups 0.01% of lineage. Interestingly, the sub-clone C1-4 presents several descendants, but none of them is among the most abundant.

Figure 3-B shows a polar plot with cumulative distances for the seven most abundant/relevant sub-clones of the C1 lineage. Each colored arc represents a sub-clone, and the arc length corresponds to the number of mutations for its direct parent. The number of cumulative mutations of a given sub-clone is the sum of all arc lengths and corresponds to the path length from the root (hypothetical germline) to a specific sub-clone. For instance, we observe 60 cumulative mutations from C1-193 to the germline sequence and only 19 for C1-19, compare arcs' lengths of C1-193 and C1-19 in Figure 3-B. The polar plot (mutational profile) is interactive, allowing users to add/remove sub-clones by clicking on nodes in the tree; a maximum of eight sub-clones can be represented simultaneously. Passing the mouse over a sub-clone on the polar plot shows its evolutionary path until the germline (highlighted in red in the tree).

Mutated positions can be better examined in Figures 4 and 5, which show the multiple sequence alignment (MSA) and a sequence logo centered in the CDR3 region, respectively. MSA is split into sequence regions: fwr1, fwr2, fwr3, fwr4 and cdr1, cdr2, cdr3. Different colors mark the border between each region pair. Conserved positions are indicated by a dot, while the presence of a nucleotide indicates a mutation in that position. The IGHD gene region is underlined in the cdr3 segment. CDR3 is the most diverse of complementarity-determining regions of immunoglobulin variable genes, and the sequence logo helps examine mutations in that region, Figure 4.

3.2 ViCLoD identifies different types of clonal lineage architectures

We selected the clonal lineages of three patients with CLL to demonstrate how ViCLoD can analyze repertoire with different levels of intra-clonal diversity. To illustrate, we selected the major clonal lineage from the repertoire of Patient 1, 2 and 3 that correspond to 96%, 99% and 99% of sequences, respectively. Figure 6 shows B-cell lineage tree topologies and mutational profiles for the

most abundant clonal lineages of selected individuals. When comparing the clonal architecture of the three patients (Figure 6-A,C,E), we observed different tree topologies. The clonal tree of patient 1 is very branched; the most abundant sub-clone C1-1 descends from C1-2, the second most abundant sub-clone. C2-1 presented seven mutations compared to the germline sequence, while C1-1 had 14 differences; we observed 7 point mutations between them. We can also see the high intra-clonal diversity of patient 1 in the mutational profile shown in Figure 6-B. The sub-clone C1-43 is closest to the germline sequence with six nucleotide differences, while C1-60 has the most mutated sequence with 27 differences. The third most abundant sub-clone, C1-3, is also highly mutated with 22 differences.

Patient 2 also presented a certain level of intra-clonal diversity, but with a different tree topology and mutational profile, when compared to patient 1; compare Figure 6-AB and Figure 6-CD. The most abundant sub-clone of patient 2 grouped 61.92% of clonal lineage sequences and differs from the germline by six-point mutations. The sub-clone C1-2, with an abundance of 35.7%, presented five mutations; between C1-1 and C1-2 we observed ten mutations. Both sub-clones share a few point mutations, which is why the ClonalTree algorithm puts them in different tree branches. The intra-clonal diversity is also evident in the mutational profile of patient 2, Figure 6-D shows a minimum of four (C1-49) and a maximum of 19 (C1-55) point mutations to the germline sequence.

On the other hand, patient 3 shows lower intra-clonal diversity, which is clearly observed in its tree topology and mutational profile, Figure 6-EF. This patient's repertoire has a highly abundant sub-clone, representing more than 93% of the clonal lineage and more than 99% of the entire repertoire. C1-1 differs by just one mutation from the germline, and the sub-clone C1-51 is identical to the germline sequence. We also observed a few branches in the clonal tree of Figure 6-E. All sub-clones (except C1-51) descend from the most abundant sub-clone C1-1, most of them with just one mutation. The most mutated sub-clone C1-227 has only six differences from the germline, but it is less abundant, grouped less than 0.01% of clonal lineage sequences. All mutations were observed in the CDR3 regions; the only exception was for C-227, presenting point mutations in regions frw3 and frw4. Interestingly, all point mutations involve nucleotides 'G' or 'C' mostly the 'G' residue.

4 Discussion

High-throughput sequencing analysis has made possible quantitative studies in immunology, where adaptive immune repertoires (AIRR) can contain millions of different receptor gene variants. Individual repertoire analysis is now routinely possible and affordable, resulting in many practical applications in biology and medical research. Several web servers for analyzing AIRR-Seq data obtained via High-throughput sequencing have been recently reported, most of them for pre-processing, V(D)J gene annotation, clonal grouping, repertoire characterization, and mutation analysis. However, few tools permit the analysis of intra-clonal diversities and the investigation of the dynamic/evolution of an individual repertoire. To fill the gap, we propose ViCloD, an interactive web server for visualizing B cell repertoire composition and analyzing intraclonal diversities. To demonstrate ViClod usefulness, we used it to analyze several BCR repertoires from patients with different lymphoproliferative disorders. These different case studies represent typical scenarios of interest when studying immune responses in the case of B-cell tumor suspicion; They demonstrated how ViCloD could be explored for analyzing this kind of data. Additional features will be implemented in the future, such as: (i) add different definitions of sub-clones and clonotypes. (ii) carry out intra-clonal analysis at the amino acid level since the clonal selection is based on the BCR protein sequence. (iii) Provide comparison functionality, which will allow users to analyze repertoires from the same patient at various times or to compare repertoires of different patients. With the growing importance of immune repertoire research, we believe ViCloD will play an important role in identifying specific features within immune repertoires, helping to elucidate the high-dimensional complexity of the adaptive immune system.

5 Conclusion

We have produced a new AIRR bioinformatics tool, ViCloD, which can help process, analyze and visualize B cell receptor repertoire clonality and (intra-)clonal diversity in B cell tumors. ViCloD provides interactive modules for analyzing and visualizing the repertoire structures, clonal compositions, and architectures. ViCloD distinguishes itself from other AIRR bioinformatics tools with several

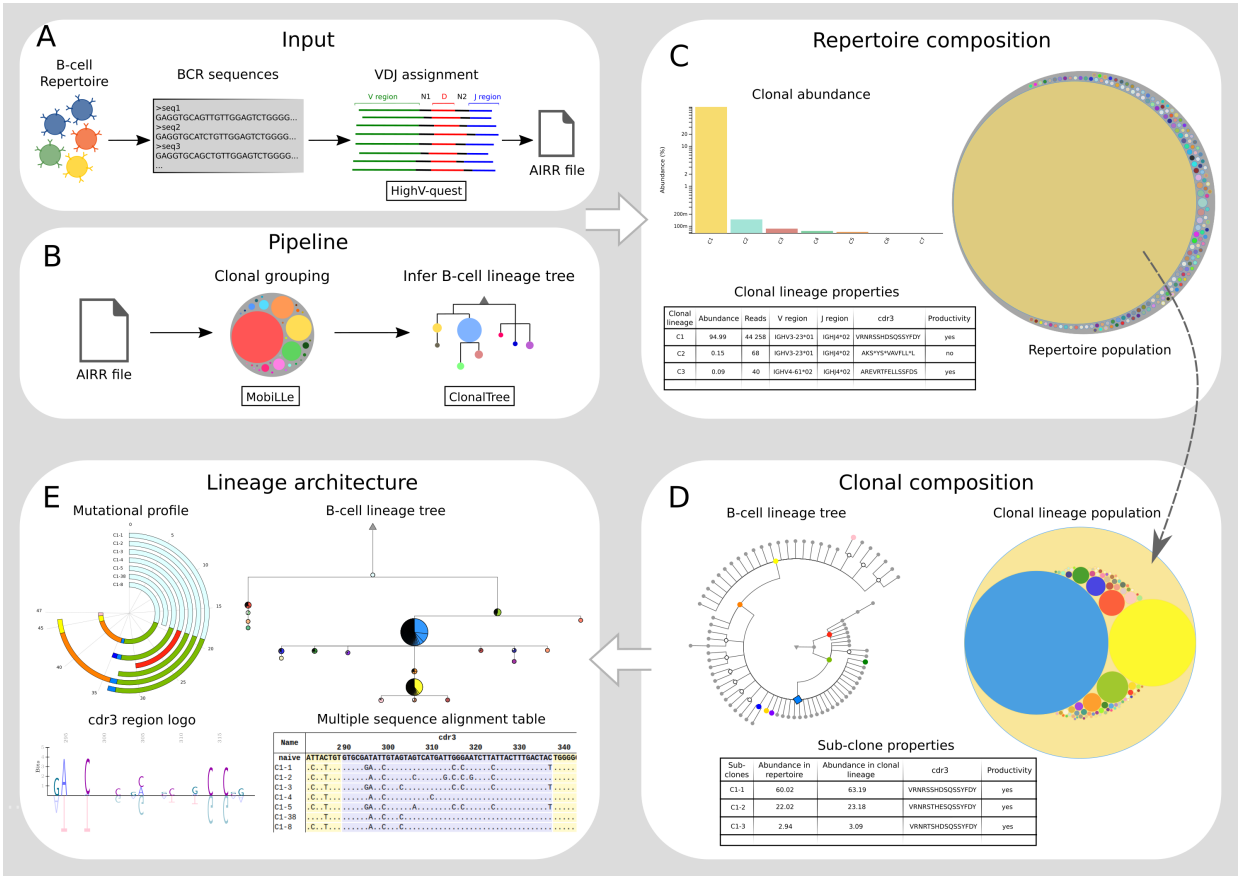

Figure 1: **Flowchart of ViCLoD.**

features. It is flexible, accepting AIRR formatted files generated by the most used tools. ViCloD is user-friendly, providing a graphical web interface that clinicians, immunologists, and researchers without strong computational backgrounds can easily use. It is anywhere accessing tool that does not require any software installation, dependencies, or requirements. ViCloD is exploitable; their plots are interactive and allow users to examine details of their data. Data are exportable; sequences, plots, and figures are easily exported to other applications for further inspection. Finally, ViClod is adaptable to clinical environments, runtime and memory were minimized to be compatible with clinical applications.

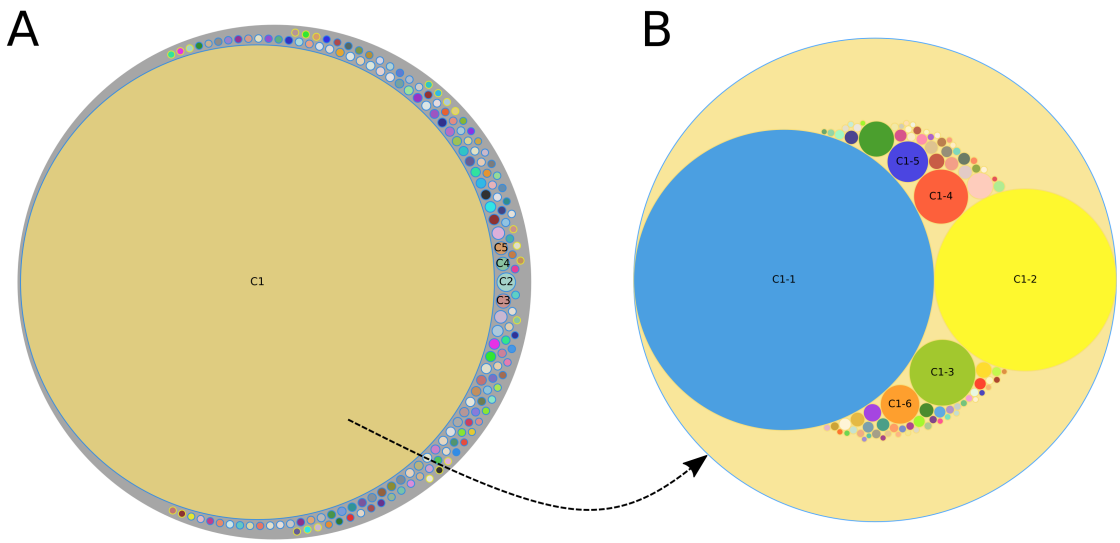

Figure 2: **FL.**oo

References

1. David G Schatz and Yanhong Ji. Recombination centres and the orchestration of v (d) j recombination. *Nature Reviews Immunology*, 11(4):251–263, 2011.

2. Susumu Tonegawa. Somatic generation of antibody diversity. *Nature*, 302(5909):575–581, 1983.

3. Michael S Neuberger, Michael R Ehrenstein, Cristina Rada, Julian Sale, Facundo D Batista, Gareth Williams, and Cesar Milstein. Memory in the b–cell compartment: Antibody affinity maturation. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, 355(1395):357–360, 2000.

4. Valerie H Odegard and David G Schatz. Targeting of somatic hypermutation. *Nature Reviews Immunology*, 6(8):573–583, 2006.

5. Joshy Jacob, Garnett Kelsoe, Klaus Rajewsky, and Ursula Weiss. Intraclonal generation of antibody mutants in germinal centres. *Nature*, 354(6352):389–392, 1991.

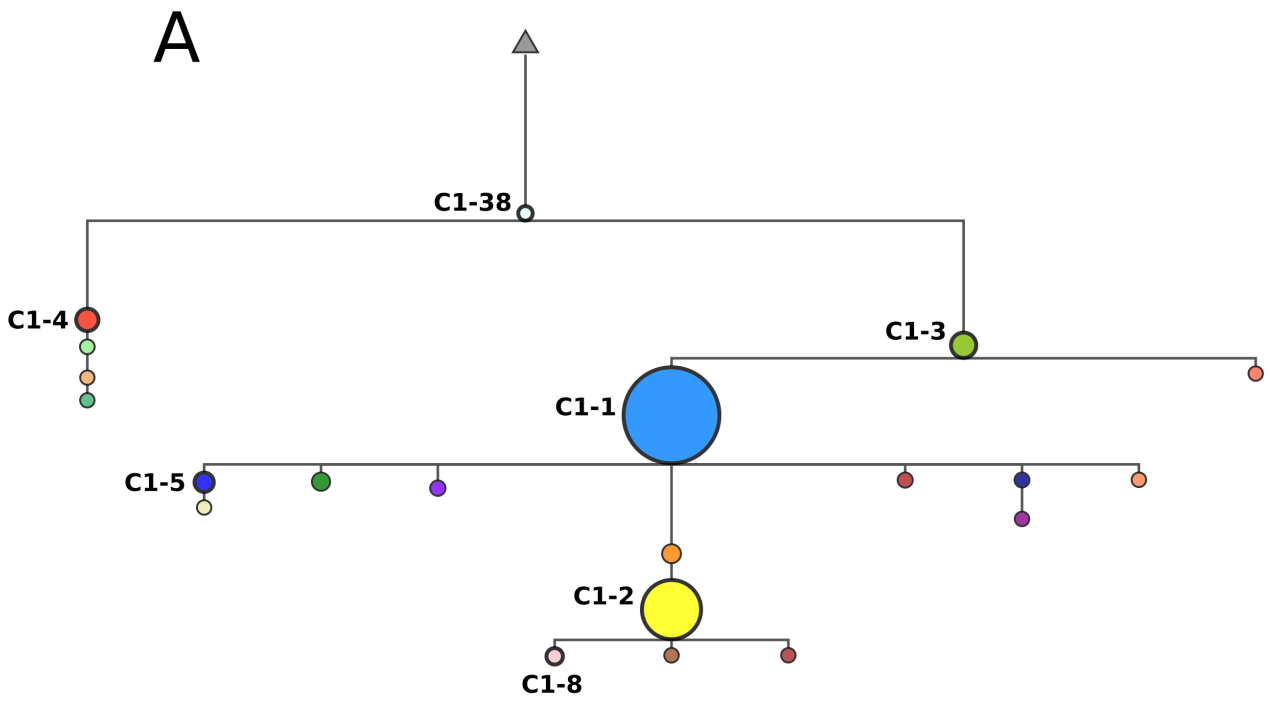

Figure 3: **FL.**oo

6. Elizabeth M. Cameron, Sade Spencer, Jonathan Lazarini, Christopher T. Harp, E. Sally Ward, Mark Burgoon, Gregory P. Owens, Michael K. Racke, Jeffrey L. Bennett, Elliot M. Frohman, and Nancy L. Monson. Potential of a unique antibody gene signature to predict conversion to clinically definite multiple sclerosis. *Journal of Neuroimmunology*, 213(1-2):123–130, aug 2009.

7. Neta S. Zuckerman, Helena Hazanov, Michal Barak, Hanna Edelman, Shira Hess, Hadas Shcolnik, Deborah Dunn-Walters, and Ramit Mehr. Somatic hypermutation and antigen-driven selection of B cells are altered in autoimmune diseases. *Journal of Autoimmunity*, 35(4):325–335, dec 2010.

8. H. Christian Von Büdingen, Tracy C. Kuo, Marina Sirota, Christopher J. Van Belle, Leonard Apeltsin, Jacob Glanville, Bruce A. Cree, Pierre Antoine Gourraud, Amy Schwartzburg, Gabriella Huerta, Dilduz Telman, Purnima D. Sundar, Tyler Casey, David R. Cox, and Stephen L.

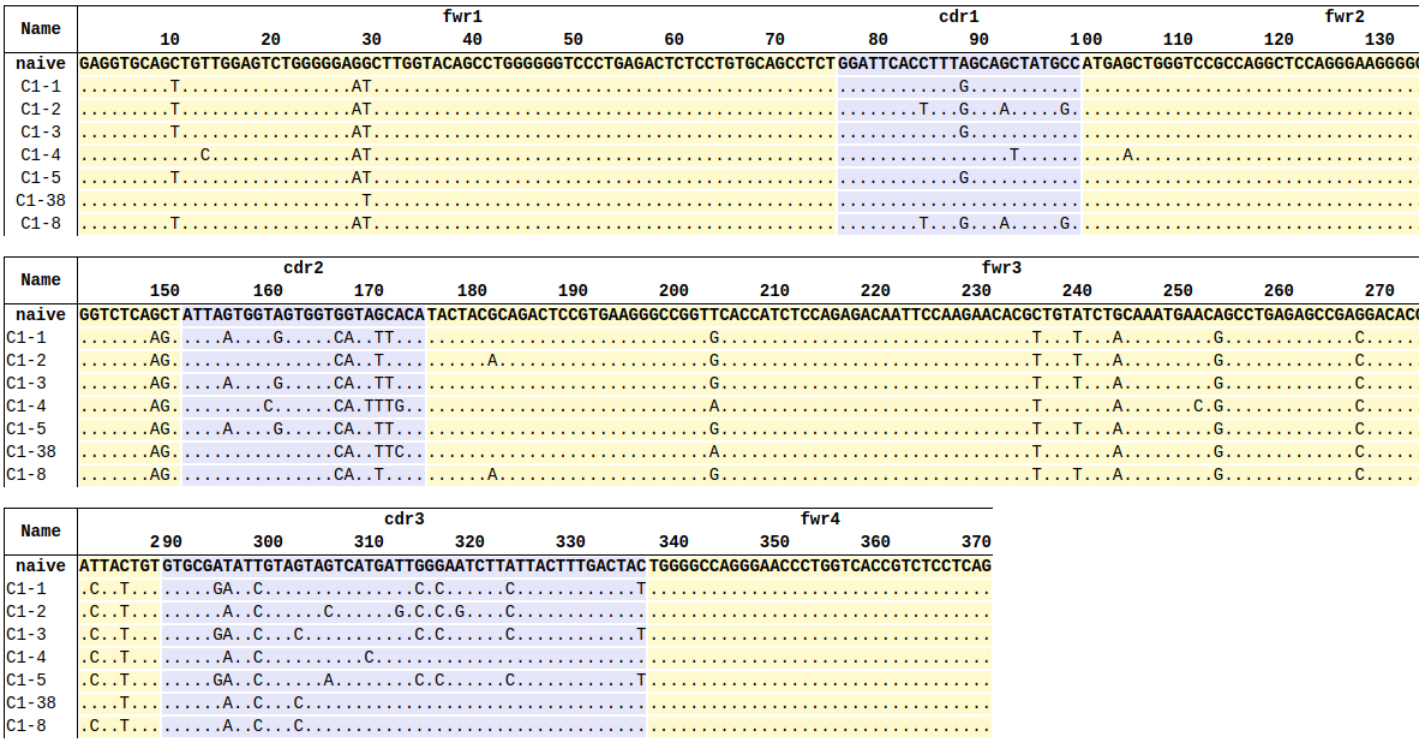

Figure 4: **FL.**oo

Hauser. B cell exchange across the blood-brain barrier in multiple sclerosis. *Journal of Clinical Investigation*, 122(12):4533–4543, dec 2012.

9. Vaibhav Singh, Marcel P. Stoop, Christoph Stingl, Ronald L. Luitwieler, Lennard J. Dekker, Martijn M. Van Duijn, Karim L. Kreft, Theo M. Luider, and Rogier Q. Hintzen. Cerebrospinal-fluid-derived immunoglobulin G of different multiple sclerosis patients shares mutated sequences in complementarity determining regions. *Molecular and Cellular Proteomics*, 12(12):3924–3934, dec 2013.

10. Klaus Lehmann Horn, Helena C. Kronsbein, and Martin S. Weber. Targeting B cells in the treatment of multiple sclerosis: Recent advances and remaining challenges, 2013.

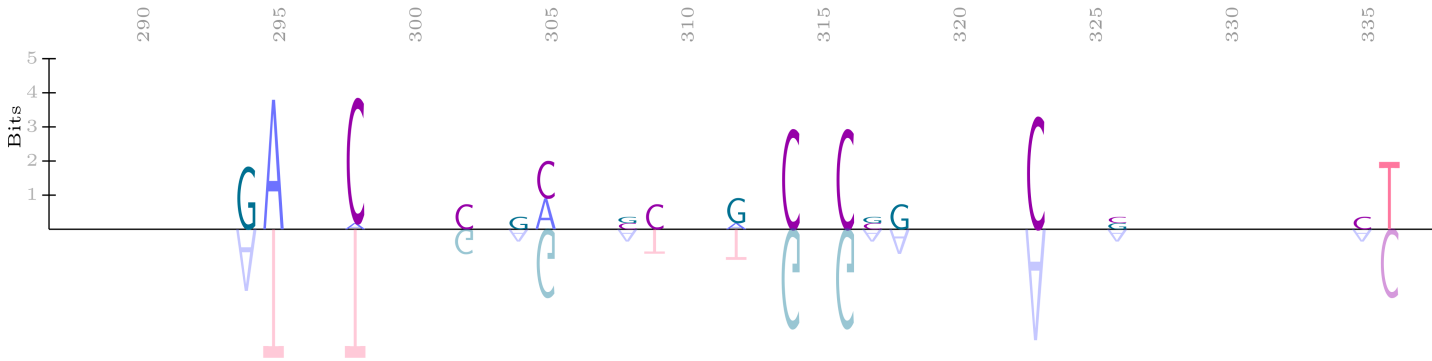

Figure 5: **FL.**oo

11. Joel N.H. Stern, Gur Yaari, Jason A. Vander Heiden, George Church, William F. Donahue, Rogier Q. Hintzen, Anita J. Huttner, Jon D. Laman, Rashed M. Nagra, Alyssa Nylander, David Pitt, Sriram Ramanan, Bilal A. Siddiqui, Francois Vigneault, Steven H. Kleinstein, David A. Hafler, and Kevin C. O'Connor. B cells populating the multiple sclerosis brain mature in the draining cervical lymph nodes. *Science Translational Medicine*, 6(248), aug 2014.

12. Arumugam Palanichamy, Leonard Apeltsin, Tracy C. Kuo, Marina Sirota, Shengzhi Wang, Steven J. Pitts, Purnima D. Sundar, Dilduz Telman, Lora Z. Zhao, Mia Derstine, Aya Abounasr, Stephen L. Hauser, and H. Christian Von Büdingen. Immunoglobulin class-switched B cells form an active immune axis between CNS and periphery in multiple sclerosis. *Science Translational Medicine*, 6(248), aug 2014.

13. Yu Chang B. Wu, Louisa K. James, Jason A. Vander Heiden, Mohamed Uduman, Stephen R. Durham, Steven H. Kleinstein, David Kipling, and Hannah J. Gould. Influence of seasonal exposure to grass pollen on local and peripheral blood IgE repertoires in patients with allergic rhinitis. *Journal of Allergy and Clinical Immunology*, 134(3):604–612, 2014.

14. Sarita U. Patil, Adebola O. Ogunniyi, Agustin Calatroni, Vasisht R. Tadigotla, Bert Ruiter, Alex Ma, James Moon, J. Christopher Love, and Wayne G. Shreffler. Peanut oral immunotherapy

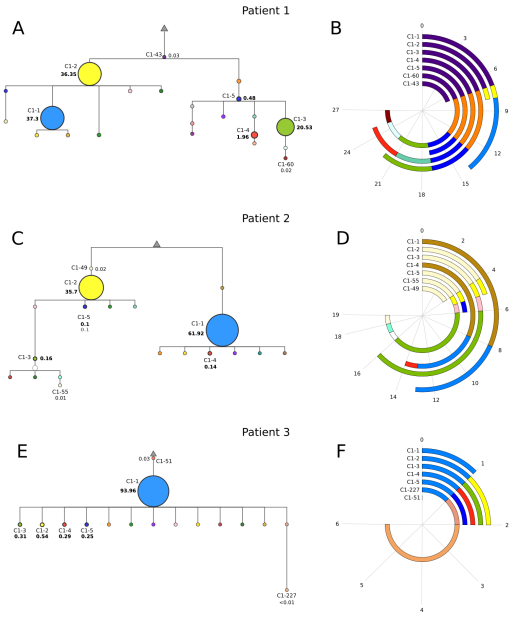

Figure 6: **FL.**oo

transiently expands circulating Ara h 2-specific B cells with a homologous repertoire in unrelated subjects. *Journal of Allergy and Clinical Immunology*, 136(1):125–134.e12, 2015.

15. Ramona A. Hoh, Shilpa A. Joshi, Yi Liu, Chen Wang, Krishna M. Roskin, Ji Yeun Lee, Tho Pham, Tim J. Looney, Katherine J.L. Jackson, Vaishali P. Dixit, Jasmine King, Shu Chen Lyu, Jennifer Jenks, Robert G. Hamilton, Kari C. Nadeau, and Scott D. Boyd. Single B-cell deconvolution of peanut-specific antibody responses in allergic patients. *Journal of Allergy and Clinical Immunology*, 137(1):157–167, jan 2016.

16. Jacob Glanville, Tracy C. Kuo, H. Christian Von Büdingen, Lin Guey, Jan Berka, Purnima D. Sundar, Gabriella Huerta, Gautam R. Mehta, Jorge R. Oksenberg, Stephen L. Hauser, David R. Cox, Arvind Rajpal, and Jaume Pons. Naive antibody gene-segment frequencies are heritable and unaltered by chronic lymphocyte ablation. *Proceedings of the National Academy of Sciences of the United States of America*, 108(50):20066–20071, dec 2011.

17. David M. Kurtz, Michael R. Green, Scott V. Bratman, Florian Scherer, Chih Long Liu, Christian A. Kunder, Kazuhiro Takahashi, Cynthia Glover, Colm Keane, Shingo Kihira, Brendan Visser, Jason Callahan, Katherine A. Kong, Malek Faham, Karen S. Corbelli, David Miklos, Ranjana H. Advani, Ronald Levy, Rodney J. Hicks, Mark Hertzberg, Robert S. Ohgami, Maher K. Gandhi, Maximilian Diehn, and Ash A. Alizadeh. Noninvasive monitoring of diffuse large B-cell lymphoma by immunoglobulin high-throughput sequencing. *Blood*, 125(24):3679–3687, jun 2015.

18. Alexander Ademokun, Yu Chang Wu, Victoria Martin, Rajive Mitra, Ulrich Sack, Helen Baxendale, David Kipling, and Deborah K. Dunn-Walters. Vaccination-induced changes in human B-cell repertoire and pneumococcal IgM and IgA antibody at different ages. *Aging Cell*, 10(6):922–930, dec 2011.

19. Victoria Martin, Yu Chang Wu, David Kipling, and Deborah Dunn-Walters. Ageing of the B-cell repertoire. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 370(1676), sep 2015.

20. Felix Breden and Corey T Watson. Using high-throughput sequencing to characterize the development of the antibody repertoire during infections: A case study of hiv-1. In *Recombinant Antibodies for Infectious Diseases*, pages 245–263. Springer, 2017.

21. Yu-Chang Bryan Wu, David Kipling, and Deborah K Dunn-Walters. Age-related changes in human peripheral blood igh repertoire following vaccination. *Frontiers in immunology*, 3:193, 2012.

22. Alexander Ademokun, Yu-Chang Wu, Victoria Martin, Rajive Mitra, Ulrich Sack, Helen Baxendale, David Kipling, and Deborah K Dunn-Walters. Vaccination-induced changes in human b-cell repertoire and pneumococcal igm and iga antibody at different ages. *Aging cell*, 10(6):922–930, 2011.

23. Jason A Vander Heiden, Panos Stathopoulos, Julian Q Zhou, Luan Chen, Tamara J Gilbert, Christopher R Bolen, Richard J Barohn, Mazen M Dimachkie, Emma Ciafaloni, Teresa J Broering, et al. Dysregulation of b cell repertoire formation in myasthenia gravis patients revealed through deep sequencing. *The Journal of Immunology*, 198(4):1460–1473, 2017.

24. Jin-Shu He, Sharrada Subramaniam, Vipin Narang, Kandhadayar Srinivasan, Sean P Saunders, Daniel Carbajo, Tsao Wen-Shan, Nur Hidayah Hamadee, Josephine Lum, Andrea Lee, et al. Igg1 memory b cells keep the memory of ige responses. *Nature communications*, 8(1):1–12, 2017.

25. Laura López-Santibáñez-Jácome, S. Eréndira Avendaño-Vázquez, and Carlos Fabián Flores-Jasso. The pipeline repertoire for Ig-Seq analysis, 2019.

26. Alexander Yermanos, Victor Greiff, Nike Julia Krautler, Ulrike Menzel, Andreas Dounas, Enkelejda Miho, Annette Oxenius, Tanja Stadler, and Sai T Reddy. Comparison of methods for phylogenetic B-cell lineage inference using time-resolved antibody repertoire simulations (AbSim). *Bioinformatics*, 33(24):3938–3946, December 2017.

27. Gur Yaari and Steven H. Kleinstein. Practical guidelines for B-cell receptor repertoire sequencing analysis. *Genome Medicine*, 7(1):121, 2015.

28. Bernardo Cortina-Ceballos, Elizabeth Ernestina Godoy-Lozano, Hugo Sámano-Sánchez, Andrés Aguilar-Salgado, Martín Del Castillo Velasco-Herrera, Carlos Vargas-Chávez, Daniel Velázquez-Ramírez, Guillermo Romero, José Moreno, Juan Téllez-Sosa, et al. Reconstructing and mining the b cell repertoire with immunediversity. In *MAbs*, volume 7, pages 516–524. Taylor & Francis, 2015.

29. Nika Abdollahi, Lucile Jeusset, Anne Langlois De Septenville, Hugues Ripoche, Frédéric Davi, and Juliana Silva Bernardes. A multi-objective based clustering for inferring bcr clonal lineages from high-throughput b cell repertoire data. *PLoS computational biology*, 18(8):e1010411, 2022.

30. Quentin Marcou, Thierry Mora, and Aleksandra M Walczak. High-throughput immune repertoire analysis with igor. *Nature communications*, 9(1):1–10, 2018.

31. Victor Greiff, Enkelejda Miho, Ulrike Menzel, and Sai T Reddy. Bioinformatic and statistical analysis of adaptive immune repertoires. *Trends in immunology*, 36(11):738–749, 2015.

32. Xingyu Yang, Christopher M Tipton, Matthew C Woodruff, Enlu Zhou, F Lee, Inãki Sanz, and Peng Qiu. Glamst: grow lineages along minimum spanning tree for b cell receptor sequencing data. *BMC genomics*, 21(9):1–11, 2020.

33. Michal Barak, Neta S Zuckerman, Hanna Edelman, Ron Unger, and Ramit Mehr. Igtree©: creating immunoglobulin variable region gene lineage trees. *Journal of immunological methods*, 338(1-2):67–74, 2008.

34. William S DeWitt III, Luka Mesin, Gabriel D Victora, Vladimir N Minin, and Frederick A Matsen IV. Using genotype abundance to improve phylogenetic inference. *Molecular biology and evolution*, 35(5):1253–1265, 2018.

35. Mathilde Foglierini, Leontios Pappas, Antonio Lanzavecchia, Davide Corti, and Laurent Perez. Ancestree: An interactive immunoglobulin lineage tree visualizer. *PLoS computational biology*, 16(7):e1007731, 2020.

36. Nika Abdollahi, Anne de Septenville, Frédéric Davi, and Juliana S Bernardes. Reconstructing the evolutionary history of a bcr lineage with minimum spanning tree and clonotype abundances. *bioRxiv*, 2022.

37. Marie-Paule Lefranc, Patrice Duroux, Shuo Li, Véronique Giudicelli, and Eltaf Alamyar. IMGT/highv-quest: the IMGT web portal for immunoglobulin (ig) or antibody and t cell receptor (tr) analysis from ngs high throughput and deep sequencing. *Immunome Research*, 08(01), 2012.

38. Jason Anthony Vander Heiden, Susanna Marquez, Nishanth Marthandan, Syed Ahmad Chan Bukhari, Christian E Busse, Brian Corrie, Uri Hershberg, Steven H Kleinstein, IV Matsen, A Frederick, et al. Airr community standardized representations for annotated immune repertoires. *Frontiers in immunology*, page 2206, 2018.

39. Eltaf Alamyar, Véronique Giudicelli, Shuo Li, Patrice Duroux, and Marie-Paule Lefranc. Imgt/highv-quest: the imgt web portal for immunoglobulin (ig) or antibody and t cell receptor (tr) analysis from ngs high throughput and deep sequencing. *Immunome research*, 8:26, 01 2012.

40. Manish Mehta, Jorma Rissanen, and Rakesh Agrawal. MDL-based Decision Tree Pruning. Technical report.

41. Robert C. Edgar. MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*, 32(5):1792–1797, 2004.

Supporting information

S1 Table Properties for the five most abundant clonal lineage.

S1 Figure Simplifying tree topology.

S2 Figure Clonotype distribution.