B cell receptor repertoire analysis in clinical context : new approaches for clonal grouping, intra-clonal diversity studies, and repertoire visualization

Nika Abdollahi

B CELL RECEPTOR REPERTOIRE ANALYSIS IN CLINICAL CONTEXT: NEW APPROACHES FOR CLONAL GROUPING, INTRA-CLONAL DIVERSITY STUDIES, AND REPERTOIRE VISUALIZATION

NIKA ABDOLLAHI

supervisors : Pr Martin Weigt and Pr Frédéric Davi

Co-superviser : Dr Juliana Silva Bernardes

Sorbonne Universités, LCQB, Paris

Nika ABDOLLAHI: *B cell receptor repertoire analysis in clinical context: new approaches for clonal grouping, intra-clonal diversity studies, and repertoire visualization*
Mai 2021

ABSTRACT

Next-generation sequencing has enabled researchers to conduct in-depth analyses of the immunological repertoire landscape. However, a significant concern in these studies is the computational cost of analyzing millions of sequences with inherent complexity, variability, and mutational capacity, imposing computational challenges and necessitating the development of efficient methods. This challenge is even more evident in the clinical context that does not always have access to professionals with computing skills or robust computational resources. Thus, the main goal of this thesis is to develop a set of dedicated and integrated tools to be used in the clinical environment, for medical diagnostic and patient care, and in the research environment, to perform large-scale and in-depth repertoire analysis. We have designed and implemented multiple algorithms and gathered them in a user-friendly interactive BCR repertoire visualization pipeline to facilitate the process of integrating BCR repertoire analysis into medical practices.

ACKNOWLEDGEMENTS

First and foremost I am extremely grateful to my supervisor, Dr. Juliana Silva Bernardes for her dedicated support and guidance. She continuously provided encouragement and was always willing and enthusiastic to assist in any way she could throughout the research project. I would also like to thank Pr. Frédéric Davi for his invaluable advice, continuous support, and patience during my Ph.D. study. Their immense knowledge and plentiful experience have encouraged me in all the time of my academic research and daily life.

I would also like to thank Martin Weigt and Alessandra Carbone for their invaluable tutelage, support, and intellectual generosity during the course of my Ph.D. degree. My gratitude extends to Mathieu Giraud and Mikaël Salson from whom's work I've drawn inspiration. Additionally, I would like to express gratitude to Kostas Stamatopoulos and his team, specially Fotis Psomopoulos, for their treasured support which was really influential in shaping my experiment methods and critiquing my results. I also thank Dr. Veronique Giudicelli for assiduously responding to my questions and reading my dissertation. I am particularly grateful for Pr. Michel Cogné and Pr. Thierry Mora who have accepted to take part in my thesis jury.

With many thanks to, Lucile Jeusset, who has performed the daunting task of translating our analyses to images in the best way imaginable. I hope she realizes how brilliant she is, and to Thibaud Verny, who has tremendously helped the progress of this project with his superb energy and creativity. I would very much like to thank Anne Langlois de Septenville, who has taught me a lot about different aspects of conducting research and provided me with constructive views about interpreting analyses, and to Marine Armand, who has gifted me with her insightful opinions and to Hugues Ripoche, who has been a great help during my Ph.D. journey.

I would like to express my appreciation for my friends- lab mates: Elin Teppa, Diego Javier Zea, Chloé Dequeker, Christos Papadopoulos, Leopold Carron, Yaser Mohseni, Ayuna Barlukova, Edoardo Sarti, Kai Shimagaki, Maureen Muscat, and Flavia Corsi for a cherished time spent together in the lab, and in social settings. I sincerely treasure each and every one of you for being by my side through thick and thin and listening to my VDJ rearrangement presentations a thousand times.

My sincere thanks to Laurent David for being the best of friends and the worst of co-workers because speaking with him has been one of the irresistible joys of working at LCQB (Thankfully he has changed his office which has helped me finish my thesis on time).

I wish to express my gratitude to Olivier, whose life has been affected by this project. I hope he would excuse me for taking so much of his mother's time in order to contribute to Bioinformatics' knowledge.

I would like to thank my parents who possess the most beautiful minds and the kindest hearts. There are no words that can express the extent of my gratitude towards them, and my sister, Sara, who has perpetually gifted me with the zest for life. Finally yet importantly, many thanks to my uncle, Majid, who has always supported me and lent me a sympathetic ear.

To Azam, Reza, Sousan, Sara and Asal

CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACRONYMS

BCR B cell receptors

TCR T cell receptors

Rep-Seq Repertoire Sequencing

Ig Immunoglobulin

IgH Immunoglobulin Heavy chain

IgL Immunoglobulin Light chain

SHM Somatic HyperMutation

gDNA genomic DNA

FR Framework region

CDR complementarity determining region

NGS Next-generation sequencing

Part I

INTRODUCTION

1

INTRODUCTION

An effective response to the vast antigenic diversity of the microbial world is the primary physiological function of the immune system. The human immune system has two levels of protection: innate and adaptive immunity. The innate immune system protects us from infection during the first critical hours and days of exposure to a new pathogen; it is not specific to a particular microorganism. The innate immune response depends on a group of cells and proteins that recognize conserved features of pathogens and become rapidly activated to help destroy them. It is quintessential for efficient protection against pathogens, but it cannot target all infectious threats. When the innate immune response is insufficient to control infection, the adaptive immune response interferes by eliminating pathogens or preventing their growth, see Figure 1. Adaptive immune responses are slow to develop on their first encounter with a new pathogen since specific clones of B and T cells have to become activated and then expand; it can therefore take a few days before the responses are functional. However, once established, the adaptive immune system remembers its previous encounters with specific pathogens, and it can destroy them quickly in the case of repeated exposure. This feature is the hallmark of the adaptive immune system; it occurs during an individual's lifetime as an adaptation to infection with many pathogens.

The adaptive immune response is characterized by high specificity and memory, both of which are attributes of T-lymphocytes (Thymus-derived cells) and B lymphocytes (Bone marrow-derived cells) [1]. B-lymphocytes can recognize and directly bind to several pathogen-associated antigens thanks to their cell-surface receptors: the B cell receptors (BCR) [2]. Each B-cell expresses a unique BCR that allows the recognition of a particular antigen. The BCR consists of two elements: the recognition unit, structured by a membrane Immunoglobulin (Ig) protein, and an associated signaling unit.

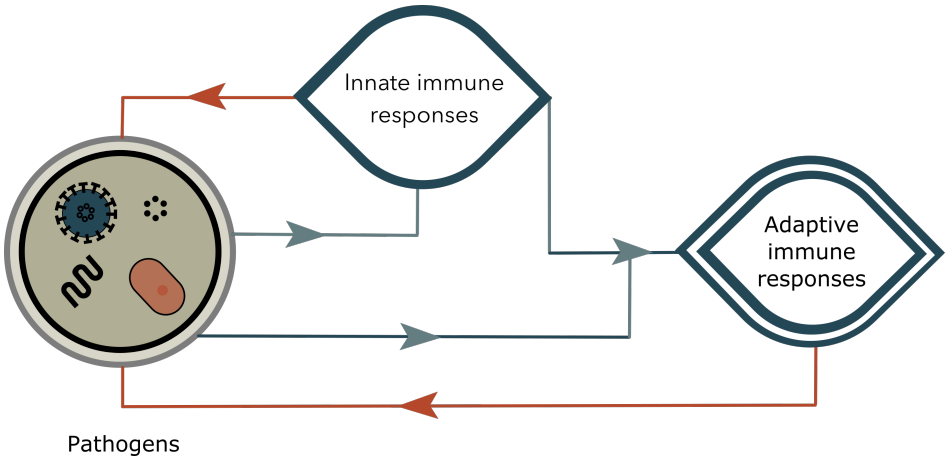

Figure 1: **Innate and adaptive immune responses** Pathogens directly stimulate innate immune responses, which protect all multicellular organisms from infection. Pathogens, together with the innate immune responses they induce adaptive immunological responses in vertebrates, which can then aid in the fight against infection.

The immune system can respond to almost any antigen to which it is exposed due to a multifarious set of BCRs, about 10^{10}, 10^{11} molecules in a human adult, that form the BCR repertoire [3]. This enormous variability is achieved through complex genetic mechanisms for BCR assembly (VDJ gene recombination) during B-cell ontogeny happening before the antigen encounter, and later during affinity maturation that occurs after the antigen encounter. B lymphocytes that encounter their cognate antigen will be activated and therefore proliferate to produce a clone, capable of giving a robust protective response against the pathogen. A clone of B lymphocytes can massively secrete a soluble form of BCRs, called antibodies. This phenomenon is known as clonal expansion and plays a fundamental role in an efficient immune response[2].

Next-generation sequencing (NGS) has transformed the immune system's analysis and shed light on BCR repertoires of healthy individuals and those with various pathologic states [4]. The BCR Repertoire Sequencing (Rep-Seq) studies have crucial theoretical and clinical relevance. 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 [5–11], allergy [12–14], cancer [15, 16], and aging [17, 18]. The study of

BCR repertoire is an active field of research that can provide insights into immunological memory[19], response to infections, mechanisms of vaccines [19], antibody engineering [20, 21], and immunoproliferative diseases [22], among others.

With decreasing costs of DNA sequencing technology, Rep-Seq datasets have become increasingly accessible in clinical contexts leading to a rapid rise in demand for appropriate methods to further investigate and interpret such data. The availability of Rep-Seq data has motivated researchers with different backgrounds (biological, computational, and statistical) to investigate and examine the adaptive immune complexity.

Numerous methods and computational tools have been developed for treating different steps of BCR Rep-Seq analysis, producing multiple integrated context-specific softwares [23–25]. However, only a few of these tools are suitable for the clinical environment [26], hindering their use in the medical context.
At least two factors contribute to this inadequacy :

1. Lack of standard terminology. Not having a common definition for terms like "clone" and "clonotypes" has induced researchers to have different interpretations of the same dataset [27]. Indeed, it complicates the communication among basic, translational, clinical, and technical researchers and decelerates the process of carrying out a meta-analysis.

2. The high complexity of implementing theoretical research-oriented tools for clinicians/immunologists without computational background. BCR Rep-Seq analysis tools often demand high computational resources, calculation time, and software skills, restraining the clinically trained professionals from fully utilizing them. This state has a baneful influence on the Rep-Seq analysis design and its application. The lack of medical practitioners' perpetual feedback can limit extracting the most appropriate information from these analyses to address unmet clinical needs in diagnosis, prognosis, and monitoring of lymphocyte malignancies.Hence, there's a need for developing practical BCR repertoire analysis tools specifically developed for clinical use.

This thesis attempts to address the above-mentioned issues by taking a dynamic multidisciplinary approach to reconciling research-oriented clonal and intra-clonal BCR repertoire analyses with lymphocyte malignancies examination in clinical settings. For this, we have developed several algorithms and assembled them in an interactive visualization pipeline. All algorithms can analyze millions of BCR sequences with inherent complexity, variability, and mutational capacity. They are very efficient, being compatible with the use in a clinical environment. Moreover, we believe that our data visualization can help the medical community by facilitating the selection of helpful information in BCR repertoire analyses and easier patient monitoring. These tools could improve healthcare clinician's work, organizing their tasks, and increase the number of patients treated with a more personalized approach.

1.1 OVERVIEW OF THE STUDY

This dissertation is organized into three sections.

- Section one: Background and problem statement
  - Chapter 2 outlines genetic mechanisms that lead to high BCR repertoire diversity, particularly the VDJ recombination and somatic hypermutations. It also includes practical information about measuring BCR repertoire diversity, such as the sample size and the currently available technologies in Rep-Seq analyses.
  - Chapter 3 presents the principal steps in Rep-Seq analysis and an overview of tools for fundamental analysis of BCR repertoires. We highlight the most critical questions in the literature that guided us towards one tangible area of investigation during my thesis, the BCR intraclonal analysis.
  - The starting point of BCR intra-clonal analysis is precisely defining the clone, as the essential subject of study. As detailed in chapter 4, there is no consensus in the scientific community about how to define clones, and it can cause ambiguity in the interpretation of results produced by different algorithms.

  - Since this study aims to facilitate the usage of Rep-Seq analysis tools for clinically trained professionals, chapter 5 introduces a practical framework for such tools' design and validation processes. This step has helped us clarify the criteria of developing methods to encourage usage of intraclonal analysis in the diagnosis, prognosis, and monitoring of lymphocyte malignancies.
  - After gaining an insight into the current needs in BCR Rep-Seq analysis, chapter 6 presents the problem statement of this research work. It provides the aim and the scope of the study.

- Section two: proposed solutions
  - Given the most appropriate definition for the BCR clone for the intraclonal analysis in the clinical context, we need a clonal grouping tool that is capable of performing this level of clustering granularity. Furthermore, It should be accurate, fast, and easy to use. In chapter 7, we present Agreeable, a tool that has been developed with the aim of responding to these needs.
  - In chapter 8, we evaluate the performance of agreeable by comparing it with some of other BCR clonal grouping tools. This chapter also investigates the impact of different clonal grouping algorithms on results while the clonal definition is the same.
  - Phylogenetic trees can represent evolutionary relations among distinct genotypes in a B cell lineage or clone. For that, phylogenetic reconstruction methods should process B-cell population data derived from experimental sampling, common in clinical routine. Chapter 9 introduces ClonalTree, a fast and accurate algorithm that reconstructs BCR lineage trees using cellular abundances and minimal spanning trees. ClonalTree was designed particularly for analyzing clinical data.
  - Considering the important role of the visualization for a complex multidisciplinary context such as BCR Rep-Seq analysis, we gathered all previously developed programs and created a versatile interactive visualization pipeline called ViCloD. Chapter 10 is dedicated to presenting this tool and its possible clinical usability.

- Section three: Conclusion and Perspectives

– In the last chapter, I will summarize our findings, contributions, and suggestions for future research directions.

Part II

BACKGROUND AND PROBLEM STATEMENT

2

STUDYING IMMUNE REPERTOIRES

This chapter introduces the biological notions used in the following parts; then, we will discuss practical aspects of studying adaptive immune behavior.

2.1 AN OVERVIEW OF THE HUMAN ADAPTIVE IMMUNE SYSTEM

Adaptive immunity response has two major types of immune cells: T and B cells. Such cells, also called lymphocytes, have cell surface antigen receptors, respectively called T cell receptors (TCR) and B cell receptors BCR, capable of recognizing and responding to an unlimited number of pathogens.

2.2 GENERATION AND MATURATION OF LYMPHOCYTES

Both B and T lymphocytes originate in the bone marrow, but only B lymphocytes mature there; T lymphocytes migrate to the thymus to complete their maturation. B and T lymphocytes that have matured but have not yet confronted antigens are known as naive lymphocytes. Such cells circulate continually between the blood and the peripheral lymphoid tissues. If an infection occurs, mature naive lymphocytes with receptors recognizing the infectious agent are held in the lymphoid tissues. These cells are activated and start to divide, giving rise to clones of antigen-specific cells that mediate adaptive immunity to fight the infection. Some of the proliferating B cells differentiate into effector cells generating antibodies, the soluble form of BCR, and some develop into memory B cells, capable of evoking an enhanced response to reinfection. Antibodies, through various mechanisms, help eliminate pathogens and their toxins. Any substance capable of eliciting an adaptive immune response is referred to as an antigen. Since the work presented in this dissertation deals exclusively with B cells and their receptors, they will be discussed in more detail in the following section.

2.3 BASIC STRUCTURE OF B CELL RECEPTOR

BCR sequences determine the B-cell antigen-binding properties. In order to react to a wide variety of pathogens, the immune system needs to generate an equivalent variety of BCRs. However, to have individual genes encoding the number of different types of BCR, the entire human genome should be dedicated to lymphocyte receptor generation. Therefore, the recombination of preexisting genes creates a part of the required diversity of BCRs [28].

The BCR consists of two types of components: the recognition unit, structured by a membrane Immunoglobulin Ig protein, and the transmembrane signal unit formed by the CD79a and CD79b molecules. An Ig is a heterodimer composed of two Immunoglobulin Heavy chain (IgH) and two Immunoglobulin Light chain (IgL) bound by disulfide bridges (Figure 2-A). Each chain has two distinct parts: the variable domain on the N-terminal side responsible for antigen recognition and the constant region on the C-terminal side attached to the cell surface. Three gene groups encode the IgH variable domain: Variable (V), Diversity (D), and Joining (J). They are clustered in loci on human chromosome 14q32, but during early B-cell ontogeny, one gene from each gene group is randomly selected and joined together,by two successive rearrangement events. This leads to the formation of a complete Variable domain encoded by a VDJ-REGION. Joining is imprecise as nucleotides are randomly deleted and inserted in the V-D (N1) and D-J (N2) junctions (Figure 2-B). Altogether such a process is known as VDJ recombination, and it is responsible for the production of highly diversified "naive" BCR repertoire.

As shown in Figure 2-C., the variable domain, after VDJ rearrangement, contains a beta-sheet Framework region (FR), that maintains the structure of the Ig molecule. FR are relatively conserved and support three hypervariable stretches spatially close to each other and form loops that interact directly with antigens. For this reason, they are called complementarity determining region (CDR). The CDR3 is at the junction of the IGHV, IGHD, and IGHJ genes, and has the highest variability, and plays a crucial role in determining antigen properties. The IMGT unique numbering for V-REGION [29, 30] has allowed redefining the limits of

the CDR and FR Regions, known as CDR-IMGT and FR-IMGT. We have used these definitions in this work.

Antigen-activated naive B cells undergo rapid proliferation (clone expansion) and further diversify their BCR by Somatic HyperMutation (SHM), an enzymatically-driven process introducing mainly point substitutions into the Ig locus. In the normal process of SHM, the variable domain, and not the constant region of the expressed heavy chains, are mutated. The mutation rate is estimated to be of the order of 10^3 to 10^4 per base cell per cell generation, and there are hotspots and coldspots of SHM that have been described [31, 32]

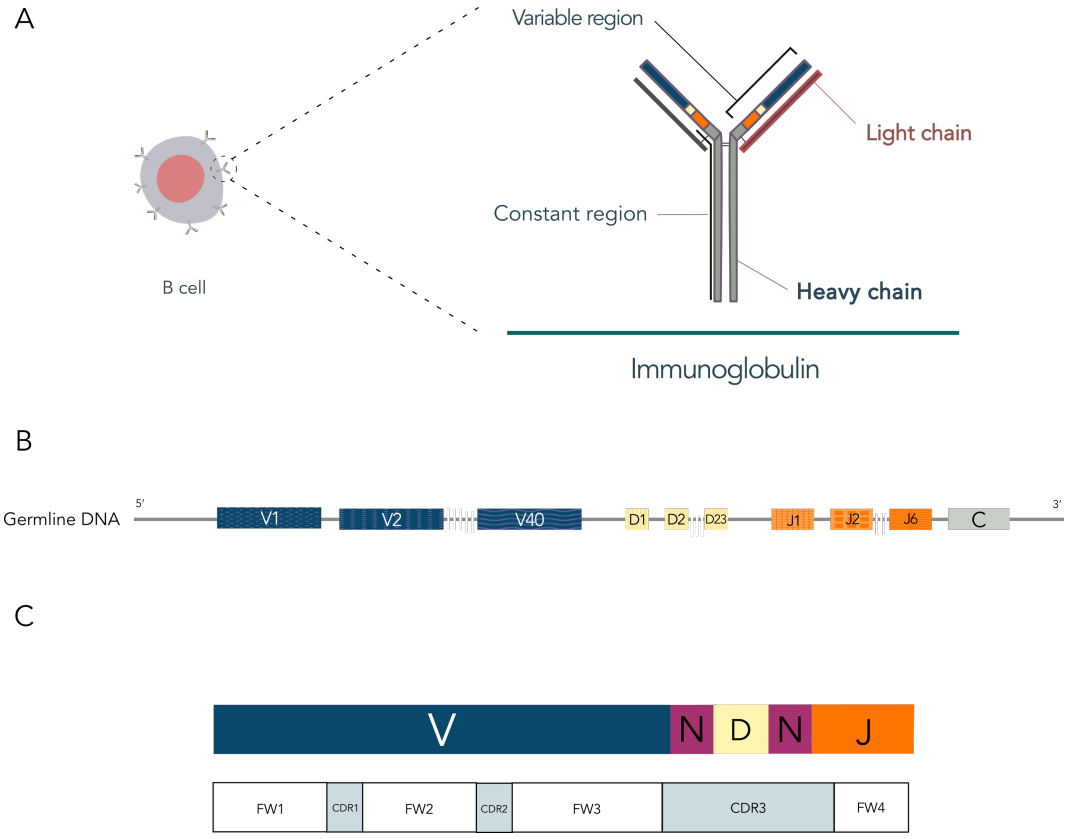

Figure 2: **A**:Representation of BCRs on the surface of B cells and the different parts of immunoglobulins, **B**: Organization of the genes encoding the heavy chains of immunoglobulins, during the rearrangements in the IGH locus, first one of the IGHD genes is joined to one of the IGHJ genes and the intermediary DNA is deleted as an excision loop, then one of the IGHV genes is joined to the partially rearranged DJ gene to generate a completely rearranged IGHV-D-J gene **C**: Schematic representation of the V domain of immunoglobulin after VDJ rearrangement

In summary, there are several sources of Variable domaine of the Ig heavy chain:

1. the combinatorial joining of IGHV, IGHD, and IGHJ genes
2. the junctional diversity derived from exonuclease digestion and the insertion of non-templated bases at the IGHV-D and IGHD-J junctions
3. the pairing of heavy and light chains
4. the intra-clonal diversification provided by somatic hypermutation, causing point mutations and occasionally longer insertions or deletions in VDJ rearrangements in antigen-experienced B cells.

The theoretical estimations of BCR diversity cover a wide range of values (from 10^{14} to 10^{18}). The combinatorial possibilities of genes are approximately 10^4 for human Ig loci. However, the diversity of CDR3 regions encoded by IGHV-D-J gene junctions has the most significant theoretical contribution to repertoire diversities. The only limiting parameter of such diversity is the length of the non-templated base sequences embedded at the junctions.

Thus, gene combinations and junctional diversity could easily exceed 10^{15} different nucleotide compositions for Ig heavy chains. Combined with light chain diversity, one can expect another 10^4 to 10^6 fold of diversity. Lastly, somatic hypermutation could contribute to generating enormous numbers of potential unique BCR sequences. However, not all of the genotype space's points will encode unique amino acid sequences; also, some will be incompatible with Ig protein stability, and some will not be expressed. Concretely, the number of B cells in the human body is estimated to be approximately 10^{11}[3]. Different approaches of estimating BCR diversity from sequencing data suggest that the B cell clone number in a human adult is in the range of 10^8- 10^9 per individual. In practice, Ig heavy chains serve as the signature of each BCR sequence to detect the clonal diversity of an individual's BCR repertoire.

Ig heavy chains are more diverse than Ig light chain sequences; therefore, they are a more appropriate choice for establishing the clonal association between sequences. A high variety of IgH chains can be due to:

1. the presence of the IGHD gene in CDR3, which is lacking in IgL chains. IGHD genes can be read in up to six different reading frames and occasionally undergo D–D fusion [33],

2. the two rearrangement junctions in the CDR3, with a higher junction variability since the enzyme which creates N additions is more active during IgH chain rearrangement [34].

2.4 THE PRACTICAL ASPECTS OF MEASURING BCR REPERTOIRE'S DIVERSITY

To evaluate the BCR repertoire diversity, we should consider two practical aspects: the sample size and the capacity of current sequencing technology.

2.4.1 *The sample size*

Most BCR repertoire analyses in clinical routines consider single blood samples, sometimes repeated in multiple time points. Since the results may be subject to inevitable sampling bias and experimental noise, it is essential to use mathematical and statistical models to estimate the real repertoire diversity from a given sample. In this work, we did not address this point, but one should have in mind the importance of sample significance to better understand the interpretability of the results.

2.4.2 *The capacity of current sequencing instruments*

2.4.2.1 *Sequencing technologies*

In the short history of rapid technological development in Rep-Seq analysis, the Roche (454), Illumina, and Ion Torrent instruments have generated most of the current bulk sequencing data in the literature [35]. These platforms use a sequencing-by-synthesis technology; the Roche (454) and the Ion Torrent platforms use reaction mixtures containing only one of the deoxynucleotide triphosphates (dNTPs) per sequencing cycle, while Illumina uses a mix of all four dNTPs with nucleotide-specific fluorophores in each sequencing cycle. The Roche and Ion torrent plat-

forms produce longer reads than the Illumina; consequently, sequencing errors in homopolymer (single-nucleotide repeat) regions are more likely to occur. Currently, Illumina instruments are the dominant platform in Ig repertoire sequencing for the clinical context. The third-generation DNA sequencing such as Pacific Biosciences' Single Molecule Real Time (SMRT), and nanopore systems allow unprecedented long reads [36]. The generation of long sequencing reads with high accuracy improves the assembly of whole genomes. Such techniques create a valuable opportunity for in-depth Rep-seq analysis; however, their frequent usage in clinical settings seems to be unlikely in the immediate future.

2.4.2.2 *Library construction and technical strategies*

Most sequencing libraries for sequencing platforms have been generated from unpaired Ig heavy chains and light chains in the published literature. The proximity of rearranged V(D)J genes allows for polymerase chain reaction (PCR) amplification of the entire V(D)J-REGION using various gene-specific primer strategies. Most amplification strategies are selected to include the CDR3 containing sufficient information to examine many of the antigen-recognizing features of the receptor molecules.

Libraries can be generated from sorted lymphocyte populations, peripheral blood mononuclear cells (PBMCs), or lymphocyte-containing tissues. An essential technical decision to make before starting a Rep-Seq analysis is the choice of template. Two starting materials can serve as the initial template to sequence Ig repertoires: genomic DNA (gDNA) and mRNA.

1. Using gDNA has the advantage of sampling and analyzing both productive and unproductive V(D)J rearrangements. Unproductive rearrangement happens when the IGHV and IGHJ are not in the same frame. Even though it does not give rise to functional proteins, sequences of unproductive rearrangements can provide helpful information on features like gene rearrangement frequencies, base deletion, and non-templated base addition in junction regions, receptor diversity, and selection in lymphocyte development [36]. Also, the copy number of the gDNA template per cell is consistent (only one productively rearranged heavy and light chain locus per cell). It can be used to evaluate and quantify clonal frequencies and expansions [37].

2. Using mRNA as an initial template requires an additional step to convert mRNA to DNA via reverse transcription. The number of Ig mRNA transcripts can vary widely among different B cell subpopulations, which prevents the reliable quantification of expanded clonal lymphocyte populations while using mRNA as the template. One way of overcoming this inconvenience is to separate cells into different replicate aliquots before isolating the mRNA. Using mRNA as a template, on the other hand, can increase the likelihood of capturing a more exhaustive representation of rare clones due to the existence of multiple Ig transcript copies per cell.

Several approaches for sequencing of lymphocyte receptor repertoires can be taken, depending on the research questions of a particular experiment. The data used in this work were collected during routine diagnostic procedures at Pitié-Salpêtrière hospital in Paris. Sequences were obtained from peripheral blood lymphocytes by performing polymerase chain amplification of IGH-VDJ rearrangements on genomic DNA followed by NGS paired-end sequencing on an Illumina MiSeq platform. Typically 10^5 sequences were obtained per sample.

In the following chapter, we will detail how to analyze the BCR repertoire starting from the raw output of sequencing platforms.

3

BIOINFORMATICS PIPELINES AND REPERTOIRE ANALYSIS

High throughput sequencing methods produce large data sets that demand specialized computational tools for proper analysis and interpretation. This necessity has contributed to the expansion of **Immunoinformatics**. This field combines biological, computational, mathematical, and statistical approaches to investigate the immune system's complexity. Its applications make use of genomic, proteomic, and structural data. This chapter focuses on immunoinformatics tools for decoding B cell receptor repertoires obtained from sequencing studies described in the previous chapter. B cell receptor repertoire analysis can be divided into three main stages: pre-processing, sequence analysis, and clustering of clonally related sequences, see Figure 3.

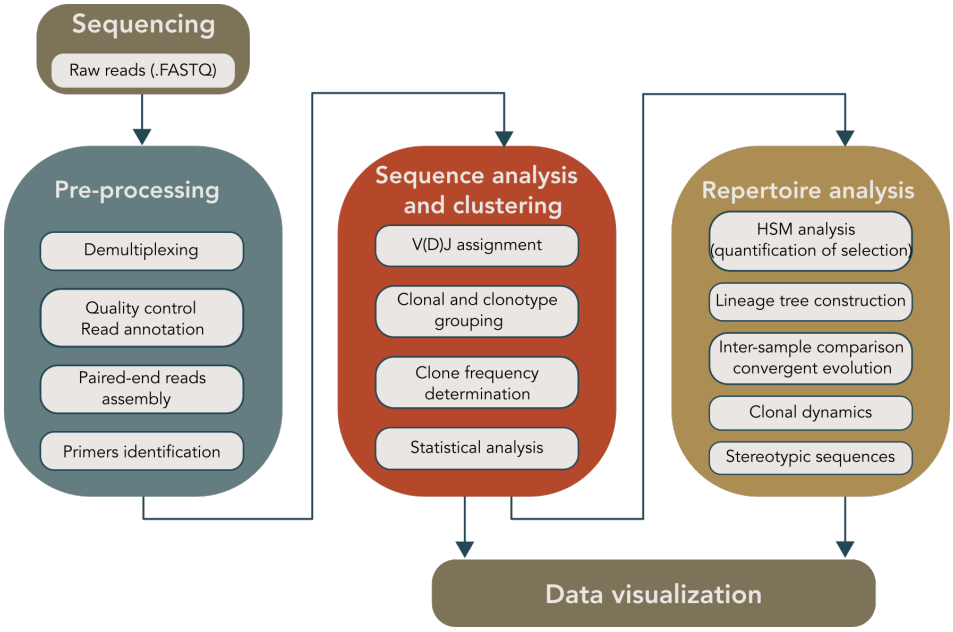

Figure 3: The essential steps in repertoire sequencing analysis, adapted from [38]

3.1 PRE-PROCESSING

The pre-processing stage aims to transform the raw reads produced by sequencing methods into properly curated sequences. The output of high throughput sequencing platforms is a binary file that must be converted to Fasta or Fastq format. Illumina and Roche platform propose integrated scripts (sffinfo-Roche; bcl2fastq-Illumina), but many independent scripts are also available (bamtoFastq, sff-extract). Fasta and Fastq are two standard input formats for most analysis programs. Fasta format consists of a list of sequences with a unique identification tag preceding each sequence. Fastq files also include, in addition to the nucleotide sequence, information about the quality of each residue in the sequence in the form of a Phred score (*Q* score). The *Q* score gives an estimated probability of error for each nucleotide position. Both the sequence letter and quality score are encoded with a single ASCII character, and the quality score can be transformed into integers. Pre-processing includes filtering out low-quality reads, sequence trimming to remove continuous low-quality nucleotides and primer sequences, merging paired-end reads by consensus building, and, if possible, identifying and filtering out PCR artifacts [39].

3.2 SEQUENCE ANALYSIS AND CLUSTERING CLONALLY RELATED SEQUENCES

The two initial steps of B-cell population structure inference named VDJ assignment and clonal grouping (or clone expansion prediction), have a tremendous impact on the success of the following phases. VDJ assignment consists in detecting IGHV, IGHD and, IGHJ germline genes used in the VDJ recombination process, where clonal grouping finds clusters of BCR sequences that might have been derived from the same precursor.

3.2.1 *VDJ germline assignment*

The V(D)J germline assignment is one of the most critical steps when treating Rep-Seq data. This step aims to infer the correct V, D, and J germline genes and alleles that were recombined to produce each BCR sequence. An germline inference is required to correctly identify somatic hypermutations for each sequence,

cluster them into clonal groups, and carry out an appropriate diversity approximation. Frequently, the germline inference applies an algorithm to choose the best match among a set of potential germline genes from a database of known genes and alleles. The current public database for Ig germline genes, the International ImMunoGeneTics information system [40], is the most used reference for an accurate VDJ assignment. It is important to highlight that the inference for D genes is particularly challenging because they tend to be short and modified during the rearrangement.

3.2.2 *Clonal grouping*

Clonal grouping (sometimes referred to as clonotyping) involves clustering a set of BCR sequences into groups that could potentially represent B-cell clones. Clonally related BCR sequences descend from a common ancestor and present the same V(D)J rearrangement, but they may differ due to the accumulation of somatic hypermutations. Consequently, detecting clones from BCR sequences is challenging. Clonal grouping tools are generally restricted to heavy chain sequences due to their high variability compared to the light chain sequences and also because they constitute the first genetic event in the B-cell ontogeny. However, most recently, experimental protocols for determining paired heavy and light chains have been developed [36, 41], these sequences could be combined.

Several clonal grouping methods have been developed to identify clones in B cell populations from IgH patient sequences. Most of them try to identify clones by first performing a VDJ assignment for unraveling V, D, and J genes used in the B-cell rearrangement. Second, sequences with the same V and J genes, and junctions of the same length, are grouped. Finally, clustering algorithms are applied to sequences within each group. For that, sequence-based distance measures are required. Most commonly, the distance measure focuses on nucleotide similarities of the junction regions (CDR3), since they are the most variable part of the BCR with antigen recognition and binding properties. Sequences with a junction similarity above a defined threshold are considered to have been originated from the same rearrangement, composing the same clone. Nevertheless, setting a fixed distance cut-off for clonal definition is inaccurate since it cannot

account for different levels of clonal diversification within a repertoire. Therefore, much effort is put into more advanced and alternative strategies that use different grouping criteria or thresholds to infer clones, including ones based on probabilistic models or spectral clustering with adaptive thresholds [42]. Given that the definition of clones and the assumptions made for each approach are different, interpreting the results of clonal grouping via different tools is laborious. The next chapter is dedicated to further investigating this problem. Most clonal grouping methods use IGHV and IGHJ annotation for detecting clonally related sequences, however, this strategy neglects the potential gene annotation errors which can later on negatively affect the clustering results.

In theory, clonal grouping can also be done before or in parallel with V(D)J assignments. In this case, clonal groups could improve the initial V(D)J allele assignments, as all the sequences in a clone emerge from the same gene rearrangement. Vidjil [43] has put into practice alignment-free clonal grouping methods bypassing initial V and J gene assignments. Kleinstein *et al.* [44] have recently developed an alignment-free clonal identification method that is not restricted to a fixed junction length. They have shown that alignment-free methods can identify clones with multiple V or J gene assignments or junction lengths that are not detectable with the junction-based distance methods. Chapter 8 is dedicated to scrutinizing clonal grouping methods.

3.3 REPERTOIRE CHARACTERIZATION AND ANALYSIS

In order to interpret the Rep-seq results, it is necessary to quantify the repertoire diversity. The quantification of the repertoire diversity helps characterize a repertoire and associate it with an immunological status (e.g., healthy, infected, vaccinated, etc.). Moreover, it can provide features to compare multiple repertoires. The repertoire comparison can be carried out from the same person at different time points or within a group of individuals.

3.3.1 *Diversity Profiles*

Diversity profiles characterize the repertoire's composition and dynamics. There are two main aspects to B cell repertoire diversity analysis: diversity quantification of a sample, and the estimation of total diversity.

1. **Diversity quantification**: it refers to a basic characterization and statistics of a repertoire such as mean clone sizes and their read counts, the number of non-functional clonotypes, CDR3 region characterization, identification of the most used IGHV, IGHD, and IGHJ genes and alleles in the repertoire, and the most frequent VDJ combinations. Other diversity measures, inspired by the quantification of species' diversity in ecology, such as species' richness, Shannon's entropy, and Simpson's index, are also commonly used to evaluate the clonal diversity of the BCR repertoires [23]. The main difference among these diversity measures is their capacity to treat small clones [21]. Depending on the analysis context, one can use measures that account only for abundant clones, or one can investigate rare clones. Hill-based diversity profiles use a continuum of single diversity indices to provide a more exhaustive vision of the diversity [45], which contains, on top of what has been mentioned above, the inverse Simpson index, the Berger Parker index, the Gini index, and the Chao1 index [23].

2. **Estimation of Total Diversity**: the peripheral blood compartment, which is currently the principal source of Rep-Seq analysis mainly for lymphoproliferative diseases, contains only 2.5% [21] (10^9 B cells) of the estimated total number of cells (10^{11}) [46], moreover, we use blood samples to analyze the repertoire which implies that only a fraction of the total diversity repertoire can be identified by Ig Sequencing. Thus, the total diversity analysis must include the estimation of the undetected clones. Mora and colleagues [47] have recently demonstrated that based on existing repertoire data and computational models, no statistical method can overcome the limitations of small sampling. They suggest that combining statistical models with stochastic models of lymphocyte population dynamics could reach a more accurate estimation of immune repertoire diversity.

3.3.2 *Mutation analysis*

High-affinity BCRs are the product of mutational events that have been accumulated during B cell maturation. The purpose of mutation analysis is to gain insight into the maturation process that B cells underwent during the course of an immune response and their encounter with antigens. As mutations accumulate in the CDRs, it is possible to identify specific motifs or punctual mutations as signatures of certain clonal populations. This can provide an enhanced understanding of the lineage and the 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. The selection pressure is measured by comparing the observed frequency of non-synonymous mutations with the expected frequency that considers hot and cold spots and nucleotide replacement bias. A higher replacement frequency implies a positive selection, while a lower frequency indicates a negative selection. A positive selection of the CDRs is expected since it may increase the affinity of the receptor towards the antigen, while a negative selection of the framework regions is necessary to guaranty the maintenance of the Ig functional structure[48].

3.3.3 *Clonal Evolution / Evolution of repertoire /clonal dynamic*

A constantly changing antigenic landscape implies constant modifications of the immune repertoire. Upon antigen recognition, B cells undergo somatic hypermutations, producing BCR sequence variants that share the same specificity but have different affinities. The collection of these sequences is known as the B cell lineage. Phylogenetic trees are often used to reconstruct both ancestral and intermediate relationships between B cell clonal sequences, thereby enabling the tracking of clonal evolution, which can be, inter alia, indicative of the antigen specificity [24, 49], refer to 9 for an in-depth discussion. Under the assumption that B cell maturation drives improved affinity, the study of the tree topology reveals the clonal selection. It allows identification of the clonal sequences or

features correlating with increased affinities for an antigen [36]. Reconstructing the evolutionary path of antigen-reactive B cells is especially relevant in the process of rational vaccine design, for example, in HIV research, where increased somatic hypermutation levels are known to be important features of broadly neutralizing antibodies [50].

Note that clonal grouping is vital for the success of the following phases. Albeit, there is no singular definition of clonal grouping. The definition varies depending on the analysis, available data, and the stage of B cell maturation, which poses a burden when it comes to properly interpreting the results.

In the next chapter, we will highlight different clone definitions and how such definitions impact the study of BCR repertoires.

4

THE DEFINITION OF CLONE

B cell clones are the fundamental selection units of the humoral immune response. A *clone* is a cellular concept that groups all cells having a common progenitor. A major issue in the BCR receptor diversity analysis is clone identification since we are unable to precisely identify clonally related cells. In Rep-Seq analysis, we group BCR sequences into clusters, under the assumption that relatively close sequences have originated from the same clonally-expanded cell. Due to the complex B cell ontogeny and intrinsic variability of BCR genes studied in Rep-Seq, we have a spectrum of molecularly defined clusters that can be associated with the cellular concept of the clone. The main obstacle of this research area is that the term clone addresses multiple representations of biological-related sequences. This causes several misunderstandings and can lead to involuntary misrepresentation of different Rep-Seq analysis methods, and consequently generating inaccurate results.

one solution would be to review the terminologies of this domain and to designate a specific name to groups of clonally related sequences at different levels, which then can potentially improve the interpretability of results. We propose in Figure 4 three different molecular levels to represent BCR clonally related sequences. It is important to highlight that such definitions consider only BCR Ig heavy chains, and other input information can change the molecular meaning of each level. According to suggested definitions, the BCR repertoire can be evaluated in three levels of biological resolutions, which are as follow :

1. **Clone** : a set of sequences that represent a B-cell lineage, that is, all progeny of a given naive B cell. Sequences within a clone should have the same VDJ rearrangement event, but they can vary due to their differences in the SHM rate. This level gathers productive and unproductive sequences.

2. **Clonotype** : Since affinity selection is based on amino acid level, we based our grouping criteria, at least partially, on the amino acid level. We used

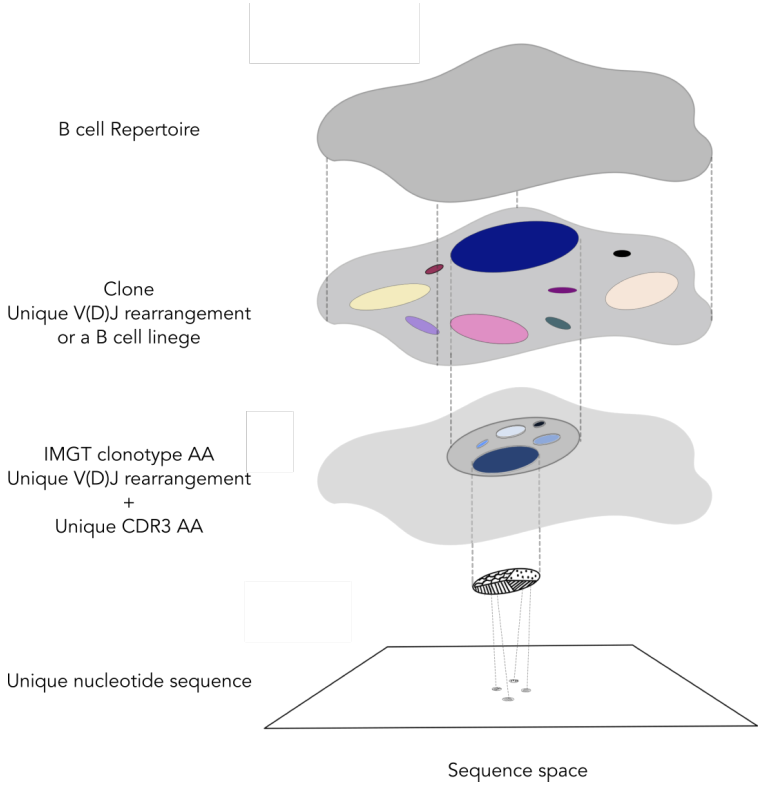

Figure 4: Different levels of grouping related sequences in a BCR repertoire.

the IMGT definition that considers a clonotype as sequences with a unique V-(D)-J rearrangement, conserved anchors (C104, W or F 118), and a unique CDR3-IMGT AA in frame junction [51].

3. **Unique nucleotide sequence** : group of identical sequences within a given clonotype. Knowing that clonotype is defined at the amino acid level, different nucleotide sequences within a clonotype should pinpoint different positions in BCR genotype landscapes.

There are several approaches to grouping sequences within each level. For instance, at the clone level, a commonly used method is to group sequences with the same V and J genes and a certain nucleotide difference at the junction region [36]. Alternatively, other approaches employ single linkage clustering, a statistical method for hierarchical clustering that does not need a pre-defined threshold for junction region [42, 52]. For grouping clonotypes, one can cluster sequences based on the amino acid similarity of the entire sequence or, as the IMGT suggests, cluster sequence with the same CDR3 length and amino acid composition. We have adopted the IMGT clonotype definition because sequences

with identical CDR3 are likely to react to the same antigen. Lastly, grouping sequences at the last level seems trivial, but based on the selected parameters for merging paired-end reads, one can obtain diverse set of unique nucleotide sequences starting from the same set of reads.

Large-scale repertoire analysis of immune receptors has important clinical usages and applications. For that, a clear description of the employed concepts and used terminologies are required. The clinically trained experts need to have a comprehensive view of the algorithmic and parameter choices since they are at the helm of providing feedback on the robustness of implemented solutions, based on their analysis context and research questions. Conclusively, we need a well-defined collaboration model between the developers and users of Immunoinformatics' softwares and pipelines.

In order to increase the usability of Rep-Seq analysis tools in the clinical context, we will outline in the next chapter a communication model for solving multidisciplinary problems observed during the developing phase of Rep-Seq pipelines.

5

A COMMUNICATION MODEL FOR OPTIMIZING REP-SEQ CLINICAL USE

Two major obstacles have hampered the vast clinical application of BCR repertoire analysis. The first one is the lack of gold-standard experimental data, which could analytically and clinically validate the multiples available tools. The second one is the ambiguity of the clones and clonotypes definition among research teams involved in producing and analysing Rep-Seq data. We included a third obstacle that is one of the most significant: the lack of a common language among immunoinformatics experts and clinically trained professionals, which can adversely impact their collaboration and their interpretation of results. The establishment of the Adaptive Immune Receptor Repertoire (AIRR) Community of The Antibody Society [53, 54], was a significant step to face these challenges. The AIRR community has expressed a need for addressing communicational obstacles and therefore is attempting to develop new standards for describing, reporting, storing, and sharing adaptive immune receptor repertoires.

To the best of our knowledge, the widespread model of interdisciplinary cooperation that involves several heterogeneous experts for integrating rep-seq analysis tools in the clinical context is linear, as shown in Figure 5. This linearity neglects the complexity of operating theoretical research-oriented tools for non-specialists in the clinical context. These tools often demand high computational resources, significant calculation time, and software skills. Such inconveniences limit the appropriate usage of such tools and could consequently discourage their integration into the clinical workflow.

A more realistic approach that could help us overcome this predicament is to perpetually receive feedback; on every stage of the development and validation process. Both experts (clinically-trained and immunoinformatics) should provide feedback to ensure the practical clinical use of Rep-seq tools. Feedbacks in this

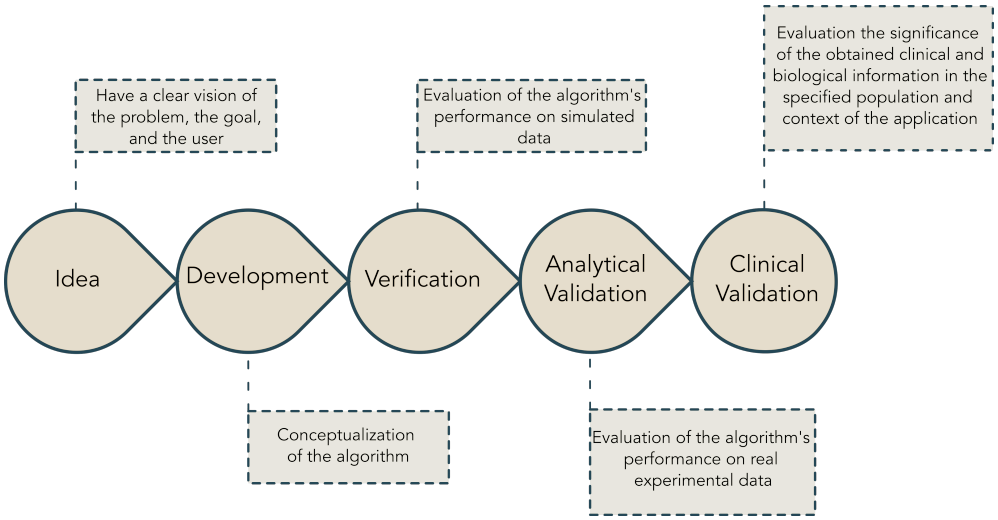

Figure 5: The linear model of interdisciplinary communication to carry out the BCR repertoire analysis

approach can circulate in various ways among different stages, as illustrated in Figure 6.

The V3 model [55] inspires the representation proposed in Figure 6. V3 was initially developed for establishing a reliable evaluation framework for Biometric Monitoring Technologies. This framework has continuous feedback between verification, analytical validation, and clinical validation steps, see solid lines in Figure 6. To adapt the V3 model to our context, we added two extra steps to cover the whole process of developing Rep-Seq tools from the initial idea to clinical validation, see dotted lines in Figure 6.

The communication model proposed in Figure 6 is appealing to all experts involved in this process: mathematical or computational scientists and medical or biological specialists. The nature of Rep-Seq data can be intrinsically interesting to mathematical or computational scientists due to the complexity of the immune system's behaviour and the modelisation complexities. It can also be engaging for medical and biological experts to respond to their research or clinical questions.

According to this model the initial stage is developing an idea for answering a research or clinical question. Based on the answers, during the next stage (the

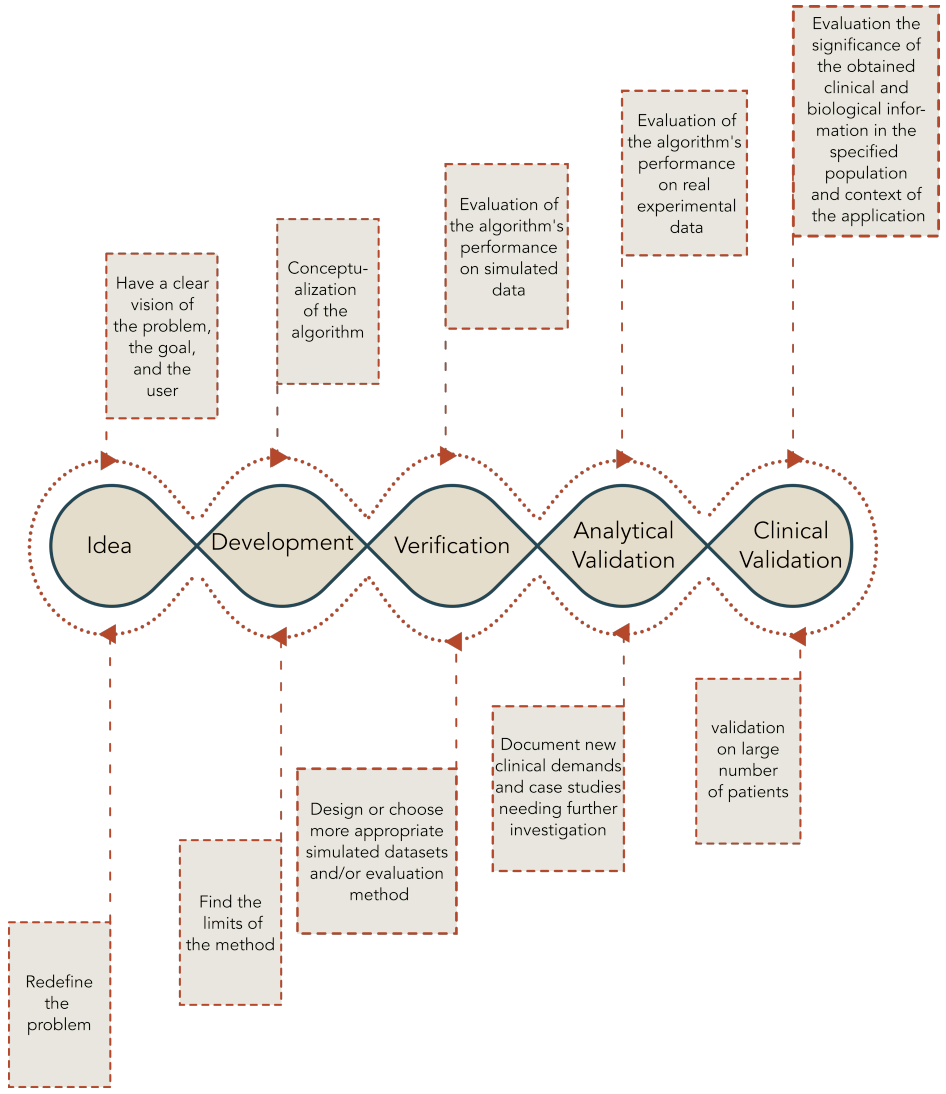

Figure 6: The interactive model of interdisciplinary communication to carry out the BCR repertoire analysis

development stage), an algorithm will be designed and implemented. Since the delivered code should be free of any implementational problems or errors,it will be checked, verified, and corrected during the verification stage. Simulated data sets are often used during this stage because such data are controlled and can reproduce various conditions such as different diseases or treatments. The goal of this stage is to make sure that the developed tool meets the requirements needed to fulfill its intended objective(s).

During the analytical validation stage, the tool's performance will be tested on a set of well-characterized real data sets and provides us with a realistic view of

it's abilities since it has been exposed to real data sets and their complexities. The intent here is to assess the pragmatic feasibility of the analysis and the consequent reliability of the results.

Throughout the clinical validation, we will check whether or not the developed tool is clinically relevant; that is, it identifies, measures, or predicts the clinical, biological, functional state, or experience in the defined context of the use. This type of validation is usually carried out by analysing numerous patients' data sets and their detailed case studies. It is important to note that the original question can be modified during this stage, enabling us to develop a more clinically appropriate method.

The dynamic nature of knowledge organization emphasizes the importance of having transparent interdisciplinary and multidisciplinary communication model throughout the tool development process. In order to further improve this process, have focused on data visualization, a communication tool that provides the maximum clarity by using the least amount of domain-specific jargons which can prevent unnecessary confusion. Data visualization is an extensive field that uses graphic representations to delineate complex quantitative information to facilitate large-scale data-driven applications.

There are an increasing number of tools for visualizing the immune repertoire analyses. For instanceARResT/Interrogate [56]is an interactive web browser-based interface that enables multiple queries on data and metadata, visualizes, provides access to whole sequences, and enables their detailed analysis. In addition, Igrec[57], SONAR[58], ImmuneDB[59], ImmunediveRsity[60] are among the tools that offer one or multiple visualization features in Rep-Seq analysis.

The epitome of visualization in the immune repertoire analysis context ,in my opinion,is put in best light by Vidjil[43]; an open-source platform that uses multiple visualization structures to analyse Rep-Seq data, notably to diagnose patients with Acute Lymphoblastic Leukemia. This visualization tool has taught me how a well-designed and user-friendly interface can easily bring together experts from different fields and help solve complex interdisciplinary problems by facilitating the interpretation of data engendering the acceleration of knowledge

mining. Despite accurate and detailed BCR repertoire analysis methods, some tools like Partis[61, 62] are not commonly employed in the clinical context since they need a high level of informatics skill and computational power. Moreover, the interpretation of their results is not straightforward. Considering these examples, the value of offering a user-friendly interface to gain medical community recognition becomes more evident. This has given us a solid reason to invest in an efficient visualization pipeline that will be explained further in chapter 10 .

Integrating Rep-seq analysis into clinical context is a multidimensional problem that requires multiple stages of development and validation. Such complexity forces us to split it into several tasks and address them subsequently. In the next chapter, we will clarify our research questions and highlight the different tasks.

6

THE PROBLEM STATEMENT

After establishing the bases of immune repertoire analysis, reviewing the current practices/pipelines of BCR immune repertoires in the clinical context, reformulating clone' definitions, and proposing a new communication model for better Rep-seq clinical development, we shed light on the following question : **"How can BCR repertoire organization and intraclonal properties be better explored to produce helpful tools for the medical community?"**.

The answer to this question is particularly important for clinical haematologists, immunologists, and experts in immunoinformatics interested in expanding their understanding of the immune system's behaviour in normal and pathological situations.

The main question can be split into the following sub-questions :

1. Among multiple clone definitions and associated algorithms, how to choose the most appropriate for carrying out a meaningful clonal analysis?
2. Among existing BCR clonal grouping tools using the appropriate clone definition for our research question, is there any that can be used in the clinical context with the aim of intraclonal analysis? If not, how should we design it?
3. What is the most efficient and accurate way to reconstruct the evolution of a B-cell lineage or clone?
4. How, in practice, can we integrate BCR repertoire and intraclonal analyse tools into the clinical context?

In the following chapters, we will discuss each question mentioned above and describe our approaches to addressing them.

Part III

PROPOSED SOLUTIONS

7

AGREEABLE; A BCR REPERTOIRE CLONAL GROUPING METHOD WITH AN APPLICATION FOR INTRA-CLONAL ANALYSIS IN CLINICAL SETTINGS

7.1 INTRODUCTION

Reconstructing clonal families from B cell receptor sequences is an essential step in the Rep-Seq analysis. The accuracy of identifying the clonally-related sequences can significantly impact the reliability and interpretability of all the downstream analyses. Several computational methods for clonal grouping have been developed, which generally employ some clustering algorithms to infer clonal relationships[52, 61, 63]. However, there are limits for their usage while aiming to carry out intraclonal analysis, especially in the clinical context. The principal limitations are the clone definition used for designing the method and the practical usability of the tool. As presented in chapter 4, based on the research question, there are various definitions of a clone that has been used in the BCR repertoire analysis. Therefore, we can obtain different results from the same data set using different clonal grouping algorithms. This difference comes not only from the different premises for designing clustering algorithms but also is influenced by the implementation of the algorithm. Tools that cluster sequences based on their lineage are more appropriate for our project's objective, the intraclonal analysis. The most accurate clonal grouping methods demand high computational resources and are time consuming. Some of them require multiple preprocessing steps; even though each step is not necessarily time-consuming, they often require a high level of computing and software skills.

In this chapter, we present Agreeable, a fast and accurate clustering method for grouping clonally-related Ig heavy chain sequences from bulk sequencing data. Our approach has high scalability, low runtime, and minimal memory requirement. Moreover, it has a few parameter settings that do not require much effort and expertise to be tuned.

7.2 MATERIAL AND METHODS

In this section we will elaborate the Agreeable algorithm and present data sets that have been used to evaluate this method's performance and it's usability. Comparison with other tools are detailed in the chapter 8.

7.2.1 *The algorithm*

We proceed through two main steps: pre-clustering and refinement. Figure 7 shows our method's flowchart and Algorithm 1, the pseudo-code for the refinement step.

7.2.1.1 *Pre-clustering*

The pre-clustering step aims to group similar sequences to form initial clonal groups that can be refined later:

1. Sequences are annotated to identify their IGHV and IGHJ genes (and alleles) and locate their CDR3 regions. For this purpose, we used IMGT/HighV-QUEST [64], but theoretically, any V(D)J annotation software could be used.

2. Sequences with the same IGHV and IGHJ genes and the same CDR3 sequence length are then grouped together.

3. lastly,we seperated sequences with less than t% of CDR3 identity (by default t is 70%), see the "pre-clustering" panel in Figure 7.

7.2.1.2 *Clustering refinement*

In this step, we iteratively refine clonal groups until we reach the minimum values for intra-clonal distances and the maximum values for inter-clonal distances. The algorithm described in Algorithm 1 takes as input the set of initial clones C , generated during the pre-clustering step. For each sequence $i \in C$ it computes two distances: a_i (intraclonal) and b_i (interclonal). Such distances measure the cohesion/separation within detected clones; they were initially introduced to compute the Silhouette [65], a performance measure that helps to interpret and

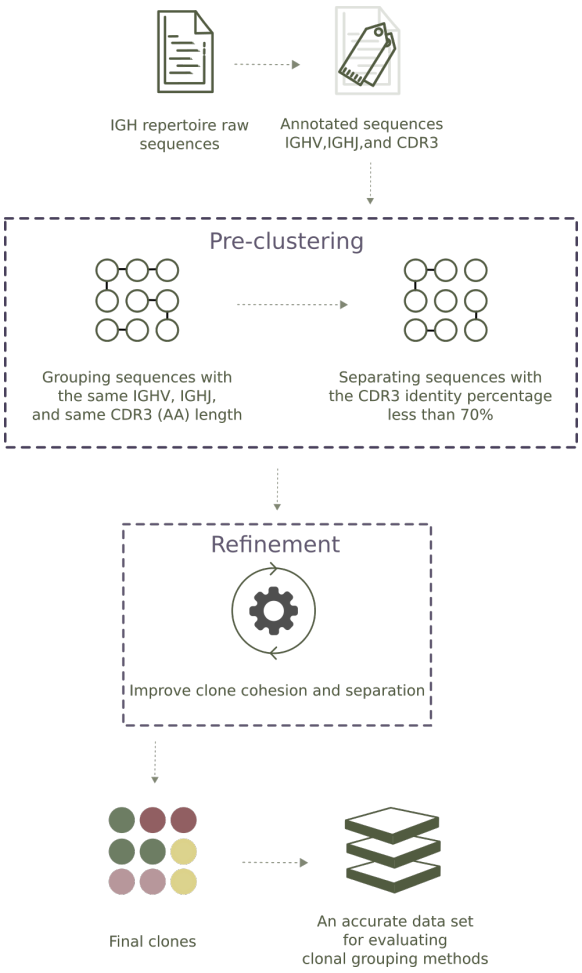

Figure 7: **Flowchart of Agreeable**. This method requires IGH annotated sequences (IGHV, IGHJ, and CDR3 region previously identified). To form initial clusters (pre-clustering step), sequences with the same IGHV, IGHJ, and same CDR3 (AA) length are first grouped together; then, sequences with less than 70% CDR3 identity are separated. During the refinement step, sequences can move amongst different clusters until no improvement is observed in cluster cohesion or separation. The final groups represent clones with low intra-clonal diversity and high inter-clonal diversity

evaluate cluster algorithms when ground truth data are unavailable. a_i is the average distance between the sequence i and any other sequence s in the same clone; b_i is the smallest average distance from i to all sequences in any other clone. In a well-detected cluster, a_i is smaller than b_i , thus, if for a given sequence a_i is higher than b_i , it might indicate that i was placed in the wrong cluster, and it should be moved to the cluster with the smallest average distance. If sequences are moved from a cluster k to a cluster l , then a_i and b_i need to be recomputed for all sequences in both clusters. Consequently, each sequence movement launches a new iteration of the algorithm, and it stops if no movement was observed in the previous iteration. Certainly, the distance metric $d(i,j)$ (between sequences i and j) plays an important role when computing a_i and b_i. Distances based on sequence similarity of the whole sequences can be inaccurate since different IGHV and IGHJ genes can present a considerable similarity. Moreover, CDR3 regions are shorter than IGHV/IGHJ genes, and a normalized distance is more appropriate. For that, we split the sequences into three parts, IGHV, IGHJ and CDR3 region, and compute a different distance of each part, separately. The distance $d(i,j)$ is the arithmetic mean of these three distances and is defined by the equation:

$$d(i,j) = \frac{dV_{ij} + dCDR3_{ij} + dJ_{ij}}{3}, \tag{1}$$

where dV_{ij} is a binary distance based on IMGT/highv-quest gene identification, it is 0 if i and j were annotated with the same IGHV gene or 1 otherwise; $dCDR3_{ij}$ is the normalized Levenshtein distance [66] between i's and j's CDR3 amino acid sequences; dJ_{ij} is the normalized Levenshtein distance between i's and j's IGHJ nucleotide sequences. We recall that the Levenshtein distance computes the minimum number of single-character editions (insertions, deletions or substitutions) required to transform one sequence into the other.

7.2.2 *Data sets*

For verifying any algorithms, a set of data sets are needed that best represent the context in which the algorithm can be used. When real datasets that provide an unbiased evaluation are not accessible, the method has to be tested at hand

Algorithm 1 : Clustering refinement

Require: C {initial groups}
  repeat
    stop $\leftarrow$ true
    for all $k \in C$ **do**
      if $|k| > 1$ **then**
        for all $i \in k$ **do**
          $a_i \leftarrow \frac{1}{|k|-1} \sum_{j \in k} d(i,j)$
          $b_i \leftarrow \min_{l \neq k} \frac{1}{|l|} \sum_{j \in l} d(i,j)$
          $N = \text{argmin}_l \frac{1}{|l|} \sum_{j \in l} d(i,j)$
          if $a_i > b_i$ **then**
            move i to cluster N
            stop $\leftarrow$ false
          end if
        end for
      end if
    end for
  until not stop

on a set of simulated data since they are well-characterized sequences. While the premises to simulate data are unique in many fields, there is no unified and standardized premise for simulating data sets in our field. Consequently, any group that wants to perform an analysis on BCR repertoire needs to create simulated datasets with their own assumptions [42, 62, 67]. Since these assumptions are identical to those used in the analysis process, the achieved results have a certain degree of bias and therefore are not fully reliable.

As far as is known, an independent B-cell repertoire simulator that could produce different types of IGH repertoires (clonal and non-clonal) does not exist. "Independent" is taken to signify stimulators having been designed and implemented separately from a BCR repertoire analysis method to minimize the bias in results. In order to create artificial repertoires, we adapted GCTree [67], a B-cell lineage simulator. To generate one repertoire, we ran GCTree several times to produce independent B-cell lineages that assembled together to simulate a repertoire. To produce a B-cell lineage, GCTree first randomly selects IGHV, IGHD,

and IGHJ germline genes from the IMGT database [40], and then nucleotide(s) can be added to or removed from the IGHV-IGHD and IGHD-IGHJ junction regions. Next, a branching process is performed, and point mutations are included in the descendants. For the branching, GCTree uses an arbitrary offspring distribution that does not require an explicit bounding. Instead, it uses a Poisson distribution with parameter λ to estimate the expected number of offsprings of each node in the lineage tree. SHM is simulated through a sequence-dependent process, where mutations are preferentially introduced within certain hot- and cold-spot different simulations. The clonal size distribution of each repertoire that was tested is shown in Table 1.

Table 1: **Clonal size distribution for three types of simulated repertoires.** Each clone is the result of an IGH rearrangement. We only keep the productive simulated sequences; therefore, the final population size might be different from the total sequence count in this table for different simulated datasets..

Monoclonal		Oligoclonal		Polyclonal	
#Clone	#sequences	#Clone	#sequences	#Clone	#sequences
1	701	1	151	14	51
14	11	1	101	14	11
12	6	10	51	12	6
8	4	14	11	8	4
8	2	12	6	8	2
		8	4		
		8	2		
43	975	54	1036	56	988

7.2.3 *Performance evaluation*

When clonal assignments are known, we can quantitatively assess the ability of clonal grouping algorithms to identify clonally-related sequences. We applied common measures such as precision and recall for comparing the inferred clones

to the true ones. We also computed the F-measure (FM), the harmonic mean of precision and recall, which is an aggregate measure of the inferred cluster's quality. Precision and recall both require three disjoint rates, which are: true positive (TP) rate, false-positive (FP) rate, and false-negative (FN) rate. Then, we have computed precision $p = \frac{TP}{TP+FP}$, recall $r = \frac{TP}{TP+FN}$, and $FM = \frac{2*p*r}{p+r}$. The values of these three metrics are in the interval [0,1], one being the best and 0 the worst performance. Of note, the way TP, FP, and FN are computed will affect the accuracy of precision and recall.

There are at least two ways to compute these values depending on the grouping level considered. The *pairwise* procedure that considers the binary clustering task and focuses on the relationship between each pair of sequences and the *closeness* procedure intends to evaluate the clone compositions and repertoire structure.

7.2.3.1 *Pairwise*

In the pairwise procedure, a pair of sequences is counted as TP, if the sequences are found together in both 'true' and 'inferred' clusters; FP, if the sequences are found separately in the true, but together in the inferred clone; FN, if the pair are found together in the true but separated in the inferred clone, see an example in Figure 8-A.

7.2.3.2 *Closeness*

The closeness procedure first identifies the best correspondence between inferred clones and correct clonal assignments. Then, it associates clone pairs that share the maximum of common sequences. Then, for each pair of clusters/clones, considering 'I' inferred and 'T' true, we compute TP as the intersection between the two sets ($I \cap T$), FP as the difference between inferred and True clusters ($I \setminus T$), and FN as the difference between these sets ($T \setminus I$), see an example in Figure 8-B.

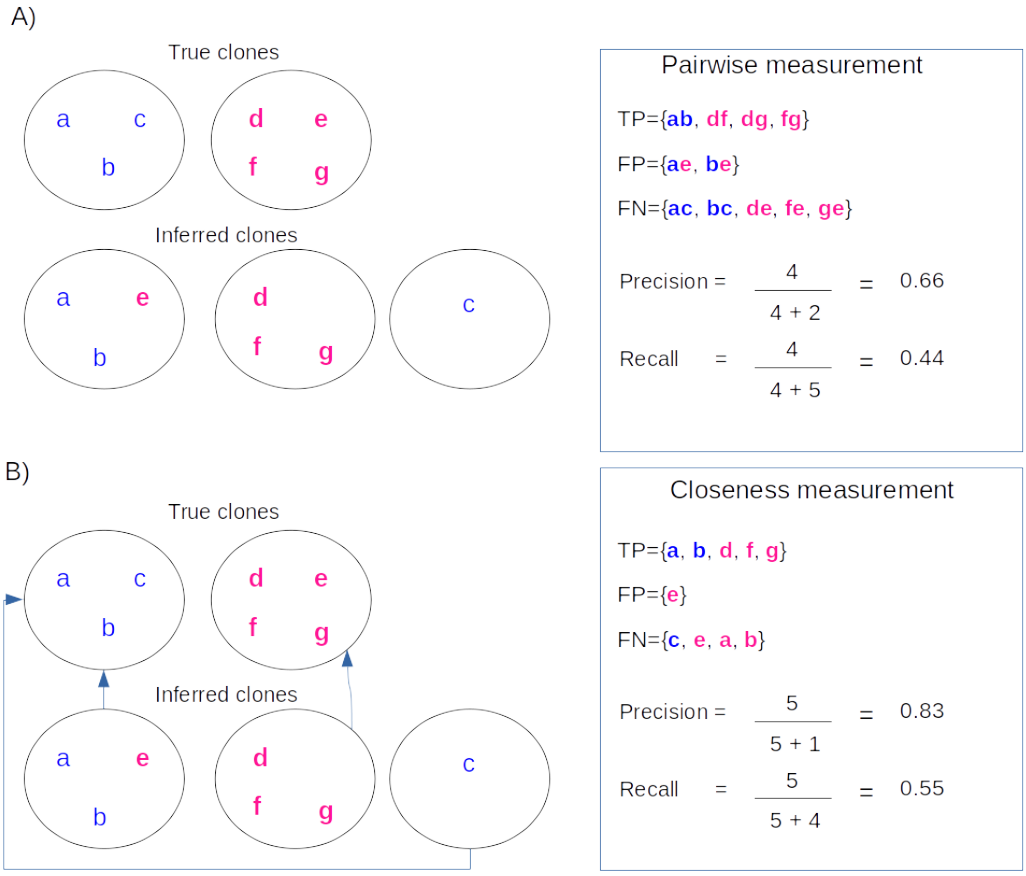

Figure 8: **Clustering performance measures**

7.3 RESULTS

7.3.1 *Reconstruction simulated repertoire's clonal architecture*

Using data presented in Section 7.2.2, we evaluated our method by comparing inferred clones to truly related clonal sequences generated during the construction of each simulated repertoire. We used two complementary approaches to precisely evaluate the clonal grouping's accuracy; pairwise and closeness, see Section 7.2.3. Agreeable achieved high precision, recall, and F-measure across all simulated data sets for both pairwise and closeness performance measures, Table 2. Across all mutation rates, it accurately identified all pairwise relationships and reconstructed all repertoires precisely. Furthermore, the absolute performance measures were remarkably high for both clustering accuracy evaluation approaches, exhibiting a mean recall over 99% and a mean precision/F-measure equal to 1.

7.3.2 *Parameter optimization*

The Agreeable's pre-clustering step adopts these criteria to group clonally related sequences: having the same IGHV gene and allele, the same IGHJ gene, and a CDR3 amino acid identity of at least 70%. IGHV and IGHJ gene annotations are often used for grouping sequences into initial clusters by several tools[23]. The only disputable parameter is the CDR3 amino acid identity threshold. Here we chose a 70% cutoff based on the definition of BCR subgroups with highly similar CDR3 motifs detailed in [68], often referred to as stereotyped BCR. However, other studies [69–71] suggest different cutoffs, varying from 50% to 70%. To obtain the most appropriate threshold, we varied the pre-clustering threshold from 50% to 90% and measured the performance of Agreeable on simulated data sets. The results are shown in Figure 9. For different repertoire types and different mutation rates, a threshold of 70% achieved the best results. Note that a higher CDR3 identity threshold reduces the performance of Agreeable in reconstructing the repertoire. It can be explained by the larger number of singletons generated with a higher threshold. singletons are clusters containing only one sequence. Once singletons are formed, Agreeable cannot merge them into higher

Table 2: **Evaluating the performance of Agreeable on simulated repertoires**. The third, fourth, and fifth columns show the number of sequences, the number of expected clones, and the number of detected clones, respectively. Pre, Rec, and FM are the abbreviations of precision, recall, and F-measure, respectively.

				Agreeable's performance						
λ_0	Clonality	# seq	# exp. clusters	# det. clusters	Pairwise			Closeness		
					Pre	Rec	FM	Pre	Rec	FM
0.16	Monoclonal	958	34	34	1	1	1	1	1	1
	Oligoclonal	1014	43	43	1	1	1	1	1	1
	Polyclonal	968	44	44	1	1	1	1	1	1
0.26	Monoclonal	659	33	33	1	1	1	1	1	1
	Oligoclonal	958	43	43	1	1	1	1	1	1
	Polyclonal	964	44	45	1	0.99	1	1	0.95	0.97
0.36	Monoclonal	924	35	35	1	1	1	1	1	1
	Oligoclonal	991	40	40	1	1	1	1	1	1
	Polyclonal	897	42	43	1	0.99	1	1	1	1
0.46	Monoclonal	952	35	36	1	0.99	1	1	0.99	1
	Oligoclonal	1016	43	43	1	1	1	1	1	1
	Polyclonal	952	43	43	1	1	1	1	1	1

density clusters since its intraclonal distance a_i is zero, and it is smaller than any other interclonal distance. We also observed that closeness performances degraded faster than pairwise when increasing the CDR3 identity threshold (Figure 9-B,D,F), mainly in the monoclonal repertoire with a high mutation rate (Figure 9-B). It is still the effect of a large number of singletons that disturbs the repertoire topology. On the other hand, a lower CDR3 identity threshold does not seem to significantly impact the Agreeable performance. Figures 9-A,B,C,E show threshold levels having no influence, while others (Figure 9-D,F) induce performance perturbations. Agreeable can adapt its performance when using the CDR3 identity threshold in the 50%-70% range, while higher values are not recommended since many singletons are generated.

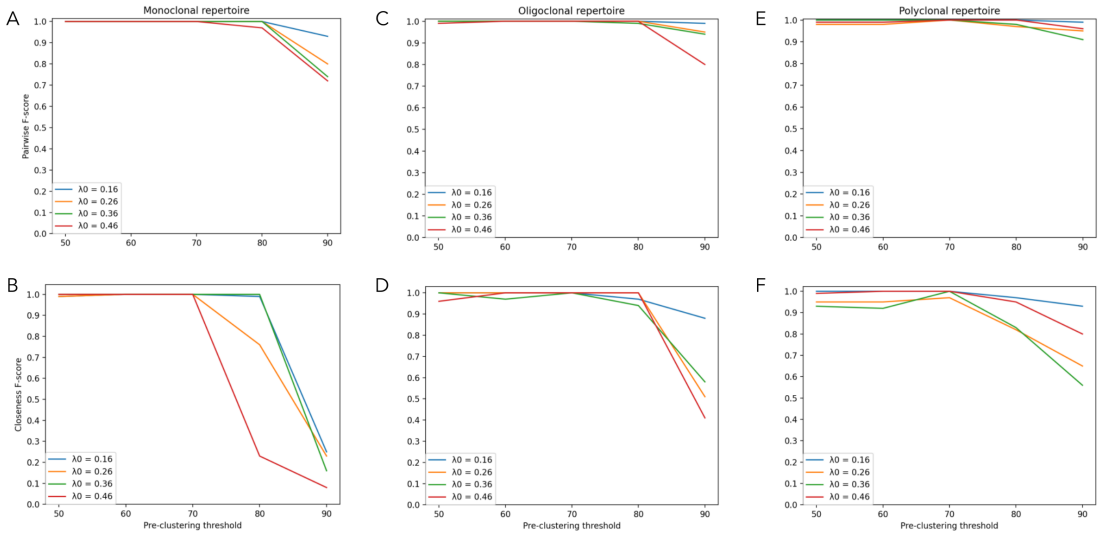

Figure 9: **Effect of pre-clustering threshold on Agreeble's performance**

7.3.3 *Runtime*

Agreeable can process large numbers of BCR sequences within a reasonable amount of time. For example, using a 3.4 GHz Octa-Core processor with 32 GB of memory, Agreeable requires 34 seconds to process a monoclonal repertoire with 33578 sequences and requires 22 seconds to process a polyclonal repertoire with 68133 sequences. Note that the clonal distribution significantly influences the runtime of Agreeable since calculating intraclonal distances in repertoires with a very high abundant clone is more time-consuming. Therefore, for the same amount of sequences but different distribution within clones, Agreeable can present different runtimes.

7.3.4 *Outputs' interpretability*

Agreeable generates multiple output files for a given set of BCR sequences, providing a well-characterized benchmark. The main outputs are :

1. A text file with clonal distribution (clones and their abundance sorted from highest to lowest),

2. A text file with all detected clones, that is, clones obtained after minimizing intraclonal distances and maximizing interclonal distances,

3. A text file with details of each analyzed sequences: the clone identifier, the clonotype identifier (based on clonotype definition of IMGT), functionality, IGHV gene and allele, IGHJ gene and allele, CDR3, and junction, a text file containing sequences that IMGT could not fully annotate.

4. A PNG file (see Figure 10) containing:

   a) Circle representation of the clone abundance. Each circle symbolizes a clone, and its size represents the clone's abundance.

   b) Number of sequences in each clone, all clones are represented, the vertical axis is in logarithmic scale.

   c) Lorenz curve and Gini index [72, 73]. A Lorenz curve shows the graphical representation of clonal inequality. The horizontal axis plots the cumulative fraction of total clones when arranged from the less to the most abundant; on the vertical axis, the cumulative fraction of sequences.

   d) Percentage of the 100 most abundant clones.

More details and examples are presented in the Agreeable Github repository https://github.com/NikaAb/AGREEABLE.

7.3.5 *Usability*

To demonstrate the application of Agreeable on real data, we have selected nine samples of human peripheral blood mononuclear cells collected during routine diagnostic procedures at Pitié-Salpêtrière hospital for this project. Three of these samples contained clonal leukemic cells, and six of them were considered non-clonal (polyclonal) originating from patients devoid of malignancy. Their clonality status was previously established by conventional methods, including PCR amplification of IGH-VDJ rearrangements followed by Genescan analysis [74] (see Figure 11).

DNA sequences were obtained by performing polymerase chain amplification of IGH-VDJ rearrangements followed by NGS paired-end sequencing on an Illumina MiSeq platform. We obtained one "Read 1" and "Read 2" FASTQ files for each sample, which were then merged by the PEAR software [75]. Next, the

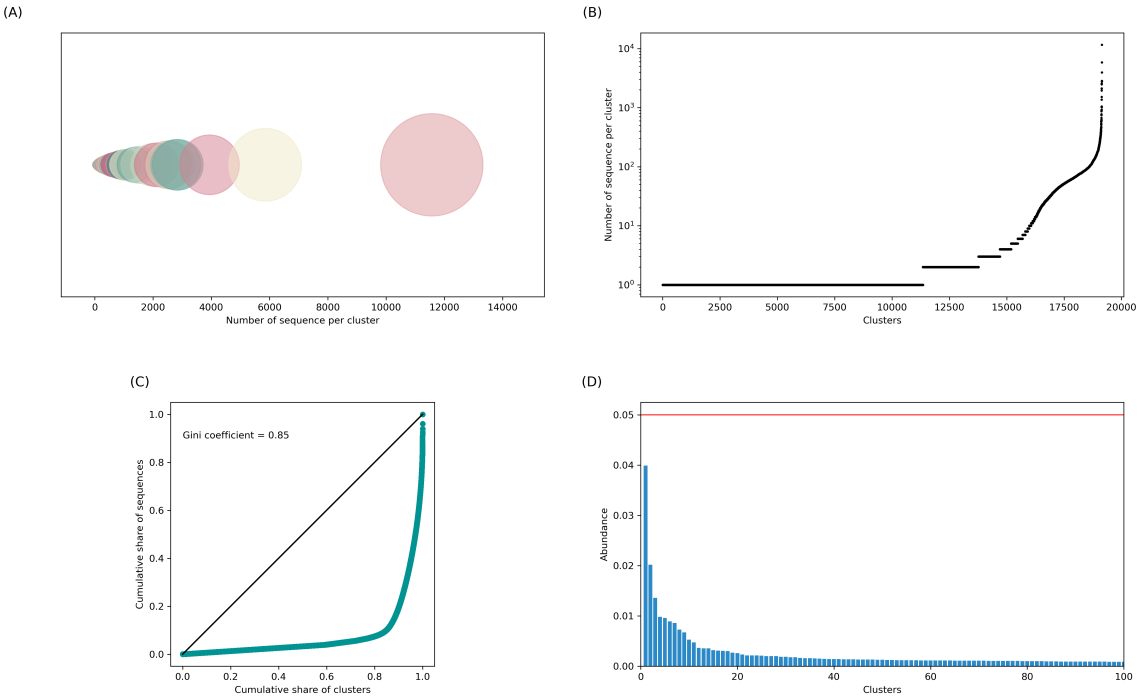

Figure 10: **Agreeable's png file output**. (A) shows the circle representation of the clone abundance. (B) shows the number of sequences in each clone, all clones are represented, the vertical axis is in logarithmic scale. (C) is the Lorenz curve and Gini index.In (D), the horizontal axis plots the cumulative fraction of total clones when arranged from the less to the most abundant; on the vertical axis, the cumulative fraction of sequences.

merged FASTQ files were converted to FASTA format with seqtk (https://github.com/lh3/seqtk). FASTA sequences were then analyzed using IMGT/HighV-QUEST tool [64] to identify the IGHV, IGHD, and IGHJ genes (and alleles) and delimit the junction and CDR3 regions. The first three columns in Table 3 represent the number of reads (sequences), the number of unique sequences, and the clonality status of each repertoire.

Figure 12 shows the clonal distribution for each analyzed repertoire by Agreeable. To measure the disequilibrium of a repertoire, we used the Gini index [76], which reflects the inequalities among values of a frequency distribution; zero indicates perfect equality, while one corresponds to maximal inequality. Clonal repertoires presented the highest Gini index, close to 1 for individuals 1 to 3 (see Figure 12-A,B,C). Repertoires 1 and 3 presented similar clonal distributions, with the presence of a major clone representing the quasi-totality of the repertoire and

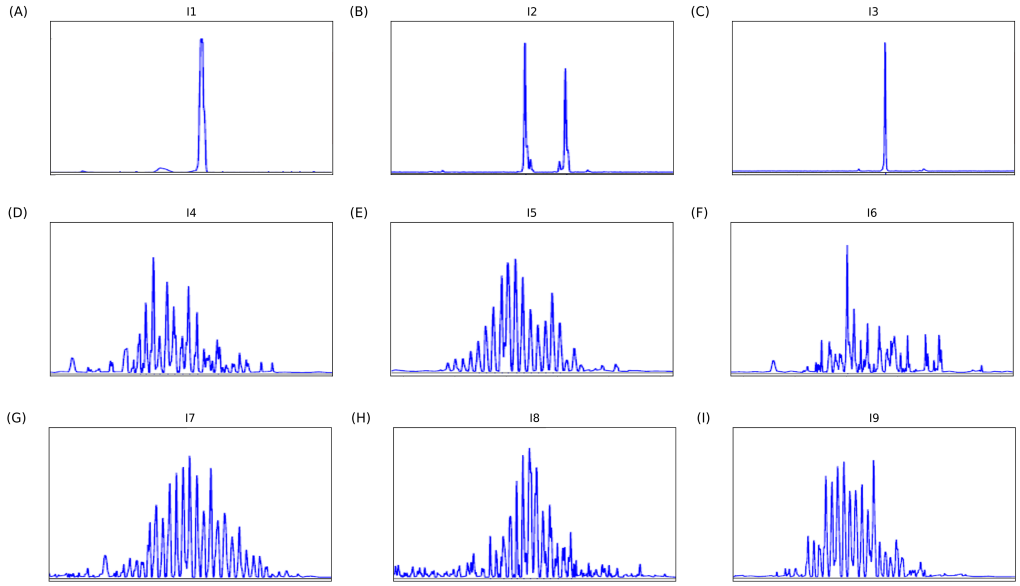

Figure 11: **GeneScan profiles of human peripheral blood samples**. IGH-VDJ rearrangements were amplified using conventional methods and PCR products were further analyzed by capillary electrophoresis. (A-C) Samples from individuals with monoclonal B-cell malignancy: monoallelic profile (A and C) or biallelic profile (B); (D-I) non-malignant samples: regular polyclonal profile (D, E, G, H, I) or irregular polyclonal profile (F).

a small number of minor clones having a low number of sequences (see Figure 12-A,C). Individual 2 presented a different clonal distribution with two major clones, each one accounting for more than 40% of the repertoire, see Figure 12-B. Detailed sequence analysis revealed that the two major clones were composed of a productive and an unproductive IGH-VDJ rearrangement, corresponding to a leukemic cell population with biallelic IGH rearrangements. Allelic exclusion is a process that prevents recombination on the second allele in one cell if the first IGH allele is correctly rearranged [77]. If the first IGH allele is rearranged out-of-frame, the process progresses to the second allele. The lack of allelic exclusion is greater in CLL than in normal cells [78], and these are known as biallelic IGH rearrangements. (see Figure 11-B). The Agreeable's results are compatible with the Genescan analysis; in the next chapter, we will evaluate our tool's performance by comparing it to the other clonal grouping methods.

7.4 DISCUSSION

The ability to obtain millions of antigen receptor sequences using NGS techniques has dramatically changed our possibilities to explore immune repertoires. Clonal relationships can be computationally identified from a large set of IGH sequences. Clonally-related sequences are descending from a common ancestor and present the same V(D)J rearrangement, but they may differ due to the accumulation of somatic hypermutations. Face to the lack of a rapid and accurate tool to group clonally related sequences from bulk sequencing IGH BCR data; we have developed Agreeable. We validated our method on artificial data that simulated three types of immune repertoires (monoclonal, oligoclonal, and polyclonal) with different mutation rates.For that, we used two different evaluation approaches: pairwise and closeness. On unbalanced data sets, both measures have some drawbacks: pairwise tend to bias towards high-density clusters' performance, concealing the performance of less abundant ones; closeness tends to be very sensitive to changes in the repertoire topology and over-penalize singleton detection. Therefore, It is our contention that both strategies should be used to evaluate clonal grouping methods.

Further improvements will be brought to Agreeable by choosing a more appropriate metric to evaluate the quality of inferred clones. The current version does not allow a singleton joining to a multiplet.

Having done the first set of tool validation, the next step is to compare Agreeable performance to other state-of-the-art tools in this domain. The next chapter is dedicated to this study.

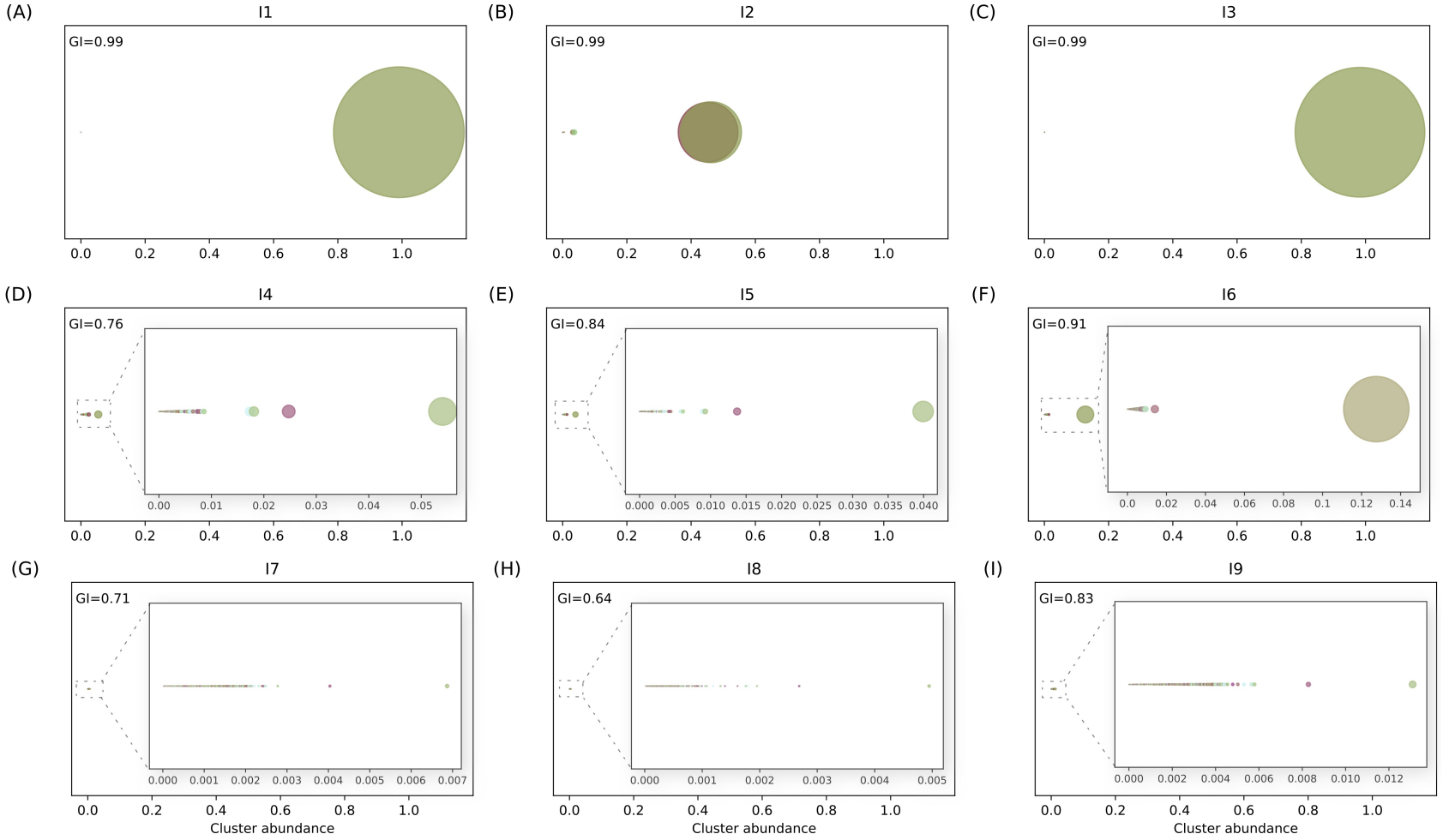

Figure 12: **Clonal distribution in real repertoires.** Each circle represents a clone, and the clone's abundance is displayed through its size.

8

PERFORMANCE EVALUATION OF BCR CLONAL GROUPING ALGORITHMS

8.1 INTRODUCTION

Clonal grouping is at the core of BCR repertoire analysis. All downstream investigations such as repertoire diversity estimation and intraclonal analysis, among others, depend on the correct grouping of BCR sequences. Members of a B-cell clone do not have identical V(D)JGHV sequences due to SHM. Consequently, defining clones based on BCR sequence data is more complex than TCR [79, 80]. In chapter 7 we proposed a new clustering tool Agreeable, that groups sequences representing a B-cell lineage into the clones (see chapter 4). In this chapter, we have chosen four tools that use the same definition of clone to compare with Agreeable. Even though the clone definition is the same, each method has its own set of characteristics concerning the underlying algorithm, prior information, and produced outputs.

8.2 MATERIAL AND METHODS

We start by describing the four freely available BCR clonal grouping methods considered for comparisons (see Table 3), report to chapter 7 to Agreeable methodology. Next, we explain two new data sets used to evaluate those methods. We also presented some metrics to quantify the differences between detected clones of analysed tools.

8.2.1 *Clonal grouping methods*

8.2.1.1 *Brilia*

B-cell Repertoire Inductive Lineage and Immunosequence Annotator (Brilia) builds up lineage tree reconstruction, clonal grouping, and V(D)J annotation into a single algorithm [81]. Since Brilia performs lineage tree reconstruction, it defines a clone as a set of BCR sequences associated with the same cell lineage. From a collection of Igh sequences, Brilia first provides initial V(D)J gene identification. For a given Igh sequence, it first matches the IGHV gene, keeping at least nine nucleotides for matching IGHD and IGHJ genes. It also corrects the indels detected before 104Cys since such indels are probably the result of a sequencing error [82]. Next, it matches the IGHJ gene, preserving at least three nucleotides for IGHD at the right of the IGHV gene. After detecting IGHJ, all remaining nucleotides are used to detect the IGHD gene and N regions. Brilia uses the IMGT database as a reference to annotate V(D)J genes. Brilia proceeds by reconstructing lineage trees that will determine groups of clonally related sequences. It first clusters together sequences with the same IGHV and IGHJ gene subgroups and same CDR3 sequence length. Next, it determines parent-child sequence relationships within each cluster for further reconstructing lineage trees. Evolutionary relationships are based on an adjusted hamming distance called SHM that penalizes dissimilarities in the N region. For each independent tree cluster, Brilia determines the root as the sequence involved in a cyclic dependency having the smallest total SHM distance to all other sequences in that cluster. Finally, a clone is a group of sequences sharing a common root sequence.

8.2.1.2 *Partis*

Partis [62] perform V(D)J assignment before clonal grouping and consider Igh sequences having the same rearrangement event as a clone, level 1 of 4. For gene annotation, Partis uses hidden Markov model (HMM) [61] to represent V(D)J rearrangement events. An HMM is a probabilistic model, where the modelled system is assumed to be a Markov process with hidden states and unknown parameters; HMM is frequently used for modelling biological sequences, where a sequence is modelled as an output of a discrete stochastic process, which progresses through a series of states that are hidden from the observer. Each of

the hidden state emits a symbol representing an elementary unit of the modelled data; for example, in DNA, a symbol represents a nucleotide. In BCR sequences, the hidden states represent either gene positions or N-region (addition or deletion) nucleotides. The HMM states represent nucleotides of each IGHV, IGHD, and IGHJ gene, the emission probabilities incorporate the probability of somatic hypermutation at each nucleotide, and transition probabilities represent the probability of moving from one state to another. The HMM' parameters (emission and transition probabilities) are estimated from a large panel of available sequences. Once the model is trained, BCR sequences are annotated by computing the Viterbi path through the HMM and finding the maximum-likelihood annotation. After V(D)J assignment, Partis applies its clonal grouping strategy. First, it creates initial clones of sequences sharing the same IGHV and IGHJ genes, and the same CDR3 length. Then, it applies an agglomerative clustering algorithm to merge clusters that maximize the likelihood.

8.2.1.3 *SCOPe*

In SCOPe, clone sequences should share a common ancestral, that is, being part of the same B-cell lineage (level 1 of Figure 4). It requires V(D)J annotation before clonal grouping, and tools such as IMGT/HighV-QUEST [64] or IgBlast [83] can be used. To define a clone, SCOPe applies a spectral clustering method with an adaptive threshold to determine the local sequence neighborhood; it means that it does not require a fixed threshold for detecting clonally-related sequences. Given a set of BCR Igh sequences, first SCOPe divides sequences into groups with the same IGHV gene, IGHJ gene, and junction length (VJl groups). Each VJ(l)-group is retrieved for inferring BCR clonal relationships. To do so, SCOPe computes the similarity matrix considering the hamming distance between junction regions of each pair of sequences within the VJl group. Then, it generates a fully connected graph from the data points, and performs local scaling to determine the local neighborhood. Based on the graph, SCOPe builds an adjacency matrix and creates a graph Laplacian. The eigenvalues of such a graph can then be used to find the best number of clusters, and the eigenvectors can be used to find the actual cluster labels. Finally, SCOPe performs k-means clustering on the eigenvectors to get the labels (clone) for each node (sequence).

8.2.1.4 *SONAR*

For the Ontogenic analysis of Antibody Repertoires (SONAR) [58], a clonal group contains all Ig reads that share a common ancestor. This tool focuses further on seeded lineage assignment, where the sequences of one or more known antibodies are used as seeds to find all sequences in the dataset from the same lineage while leaving the rest of the sequences unclassified. In addition, it can perform "unseeded lineage assignment," which consists of classifying sequences into component lineages without any additional information. In order to perform an unseeded lineage assignment, Sonar separates sequences based on their assigned IGHV and IGHJ genes. The sequences in each group are then clustered based on their CDR3 nucleotide identity (by default, 90% of CDR3 sequence), using the UCLUST algorithm in USEARCH [84]. Eventually, each clone is identified as a distinct unseeded lineage.

Table 3: **A few of the general characteristics of the tools that we have compared.**

	Partis	SONAR	Brilia	SCOPe	Agreeable
Year	2016	2016	2017	2018	2021
Number of citations	51	37	12	12	-
Implementation	Python, c++	Python	Matlab	R	python
Required level of computing skills	Advanced	Intermediate	Intermediate	Intermediate	Basic
Approximate prediction of the required time needed for a data set containing 10000 sequences	>10 min	<1 min	<5min	<1min	<1 min

8.2.2 *BCR high throughput sequencing data*

To better understand the distinction among different clonal grouping algorithms and their results, we have used one set of simulated repertoires detailed in Section 7.2.2 and two types of high throughput sequencing data: two artificial monoclonal repertoires and three experimental BCR repertoires.

8.2.2.1 *Artificial monoclonal repertoires*

To evaluate the accuracy of clonal grouping tools, we have constructed artificial monoclonal repertoires by mixing a known clone (from a B cell lineage) and a polyclonal background. For the known clone, we used genomic DNA from a

pure B lymphocyte lineage. These sequences are our ground truth, and accurate clonal grouping algorithms might cluster them together. To make the task more complex, we combine sequences from each pure lineage to sequences obtained from a polyclonal background. Our goal was to determine if clonal grouping methods can separate sequences from these two sources. To form a data set, we consider a total of 10000 sequences, where 10% of them were sampled from the pure lineage and 90% from the polyclonal background. Since we know the truly clonally related sequences in the data set, we can compare the different tools for determining their grouping differences. We created two artificial monoclonal repertoires with two different B cell lineages, each having a specific V(D)J rearrangement. The known clone of the artificial monoclonal data named AMD1 has IGHV1-69*01 and IGHJ6*03 genes, while the known clone of the monoclonal dataset AMD2 is characterized by IGHV3-48*02/IGHJ4*02 rearrangements. For AMD1 and AMD2, the sampling of the main clone has been performed from 83902 and 60522 sequences, respectively. In addition, the polyclonal background was sampled out of 136977 sequences.

8.2.2.2 *Experimental data sets*

To compare the results of Agreeable with other clonal grouping methods on experimental repertoires, we have selected three samples of human peripheral blood mononuclear cells collected during routine diagnostic procedures at Pitié-Salpêtrière hospital. Two samples with clonal leukemic cells (I1 and I2) and one sample (I8) considered non-clonal (polyclonal), taken from patients devoid of malignancy. Their clonality status had been previously established by conventional methods, including PCR amplification of IGH-VDJ rearrangements followed by Genescan analysis [74], see Figure 11. I1 and I2 are monoclonal repertoires, while I8 is a polyclonal repertoire, for more details, see 7.3.5.

8.2.3 *Performance evaluation*

To measure clonal grouping tools' accuracy, we have compared the distribution of sequences within detected clones to the expected ones. For artificial monoclonal repertoires, expected clones are the ground truth (Section 7.2.2), while for experimental repertoires, the expected clones were those obtained by the tool

that achieved the best performance on artificial or simulated data. Once the set of expected clones is defined, we can use the classical measures defined in Section 7.2.3 to evaluate the performance, such as precision, recall, and F-measure.

8.2.3.1 *Clonal comparison*

To compare clones obtained by different tools with expected ones, we have defined four "events" that describe the differences between each pair of clonal distributions. For this, we labelled clusters of a given distribution D1 and compared them with clusters in a distribution D2. These events are represented in Figure 13, and can be interpreted as follows:

1. join: when sequences of different clusters in D1 were joined in the same cluster in D2 (Figure 13-A).
2. identical: clusters in both distributions are identical, they contain the same set of sequences (Figure 13-B)
3. split: when sequences of a cluster in D1 were divided into multiple clusters in D2 (Figure 13-C)
4. Mix: when a mixture of the three above events occurs. For instance, in Figure 13-D, we observe two events, “split” and “join”

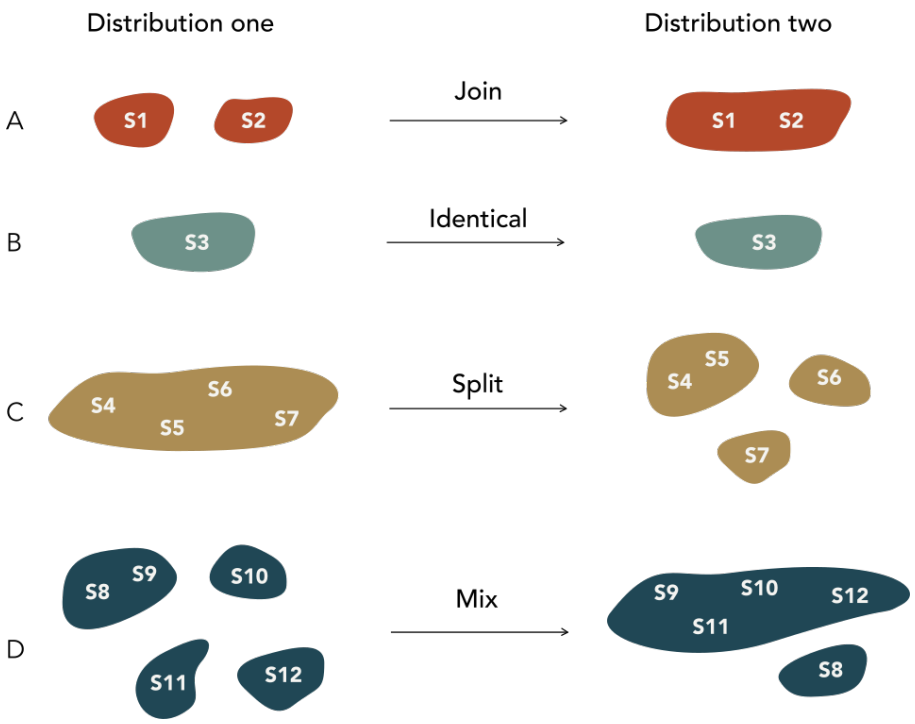

Figure 13: **Four "events" describe the differences between two clonal distributions of the same set of sequences.**

An example of how we have used this labeling system to compare two distributions is illustrated in (Figure 14). We have used pie charts to quantify these events, when comparing each tool and corresponding Agreeable's output. Pie charts are constructed based on the number of sequences in clusters labeled with each event. On the plot legends, besides each event name associated with a color used in the pie chart, there are the number of clusters with this label and number of sequences in these clusters. 'Not found' shows the number of sequences that have been in the analysed tool's output but are not in the Agreeable's output .

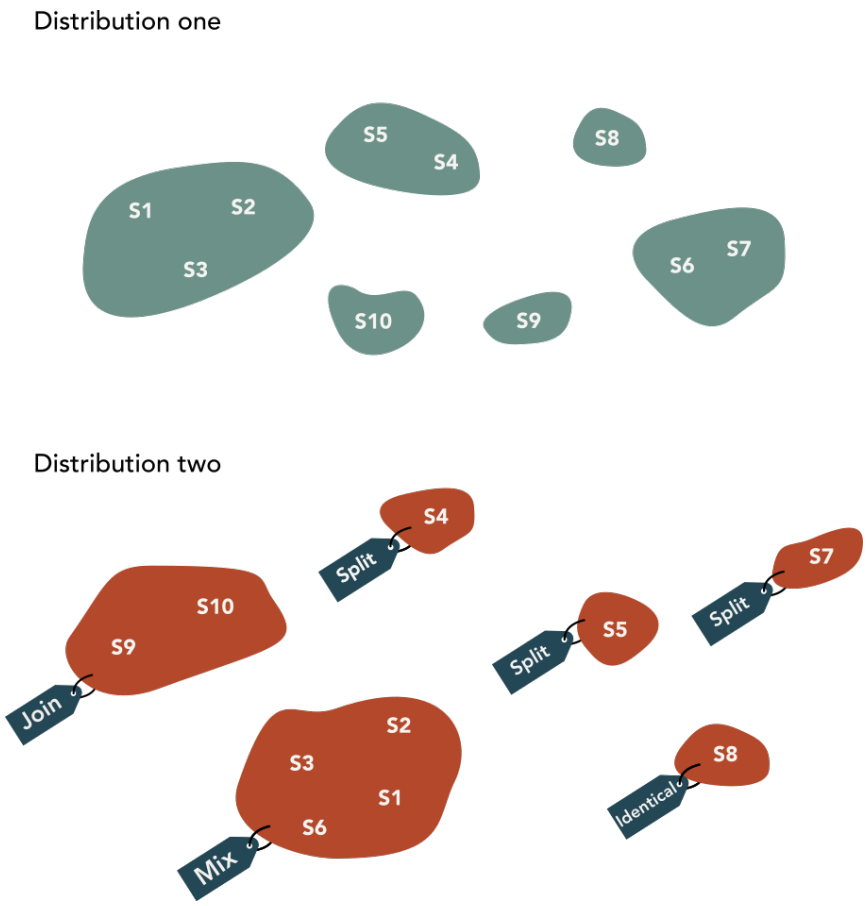

Figure 14: **An example of comparison between two clonal distribution using "4 events" labeling system.**

8.3 RESULTS

8.3.1 *Simulated repertoires*

We created artificial benchmarks that simulate several types of repertoire (clonal and non-clonal) with different SHM rates, see Section 7.2.2. Figure 15 shows the results of five analysed tools for different mutation rates, see also tables 11-22 in appendix.

8.3.1.1 *Pairwise performances*

Agreeable achieved the best pairwise performance across all simulated data sets, see details in Section 7.3.1. Scope and Partis achieved better performances than Sonar and Brilia. All tools achieved a precision close to 1, demonstrating that

few false positives were detected. However, most tools have oversplit clones, detecting many false negatives that considerably decrease their recall and F-score values.

SCOPe achieved high recall and F-measure for simulated data sets with lower mutation rates, see Appendix B (tables with $\lambda_0 = \{0.16, 0.26\}$). Recall and F-measure values were above 0.94 for these six simulated repertoires. For the remaining data sets, produced with higher mutation rates $\lambda_0 = \{0.36, 0.46\}$, we observed lower recalls and F-measures. On the other hand, Partis obtained a good pairwise performance across all simulated data sets independently of mutation rates. The only exception is the monoclonal repertoire produced with $\lambda_0 = 0.36$. For this data set, Partis has detected more clones, decreasing its pairwise recall considerably. Interestingly, for lower mutation rates, Scope outperformed Partis, but we observed the reverse for higher mutation rates for most simulated repertoires. Thus, Partis seems to be more accurate when analysing clonally related sequences with higher divergence.

For oligoclonal and polyclonal repertoires, the different mutation rates seem to influence Sonar performances. Recall and F-score decrease as long as mutation rates increase, especially for the oligoclonal repertoires. For the monoclonal samples with $\lambda_0 = 0.26$, Sonar obtained lower recall and F-score than with $\lambda_0 = 0.36$. We observed that Sonar has oversplit the largest clone of the first repertoire ($\lambda_0 = 0.26$), grouping only 37% of sequences. On the other hand, it has less split the most abundant clone of monoclonal repertoire generated with $\lambda_0 = 0.36$, grouping 62% of sequences. Once splits in large clones contribute more to accuracy decreasing, it can explain the lower performance of Sonar on monoclonal repertoire ($\lambda_0 = 0.26$). For the monoclonal repertoire with $\lambda_0 = 0.46$, Sonar detected four times more clones than expected, obtaining its lowest recall and F-score, 0.03 and 0.06, respectively.

Most of the times Brilia achieved the lowest pairwise performances across all simulated repertoires generated with different mutation rates. Brilia removes sequences that cannot annotate, reducing the original data set, which impacts the accuracy calculation. We also observed that Brilia oversplit clones; it has produced the highest number of clones for most simulated data sets. The best performance was obtained on polyclonal repertoires generated with lower mutation

rates ($\lambda_0 = \{0.16, 0.26\}$) and the lowest performance on monoclonal repertoires with higher mutation rates ($\lambda_0 = \{0.36, 0.46\}$).

8.3.1.2 *Closeness performances*

Agreeable also achieved the best closeness performance across all simulated data sets; see details in Section 7.3.1. As observed for pairwise performances, Scope and Partis outperformed Sonar and Brilia. All tools obtained high precision values but much lower recall and F-score values. We also observed lower closeness performance values for the four tools; the closeness evaluation tends to be more challenging than pairwise since clonal distribution are also evaluated rather than pairwise relationships.

Scope performances were affected by higher mutation rates. In general, it achieved better F-scores on repertoires generated with lower mutation rates, especially for oligoclonal and polyclonal samples, where we observed a notable difference between repertoires generated with $\lambda_0 = \{0.16, 0.26\}$ than those generated with $\lambda_0 = \{0.36, 0.46\}$. Scope obtained higher F-score values (>0.73) on monoclonal repertoires generated with $\lambda_0 = \{0.16, 0.26, 0.36\}$. However, its performance sharply decreased on the monoclonal repertoire with the highest mutation rate, achieving 0.16 and 0.28 for recall and F-score.

Higher mutation rates did not impact the performance of Partis. Its performance was stable on polyclonal repertoires and had some fluctuations on oligoclonal repertoires. Interesting, on monoclonal repertoires, it achieved better performance for highly mutated repertoires, being the best values obtained on sample generated with $\lambda_0 = \{0.46\}$.

Sonar performance was entirely affected by higher mutation rates. We observed a decrease in the performance, especially on oligoclonal and polyclonal repertoires. Independently of mutation rates, Sonar achieved very low recall and F-score values on monoclonal repertoires, smaller than 0.2. It has oversplit the most abundant clones that greatly impact closeness performances. For all repertoires generated with higher mutation rates ($\lambda_0 = \{0.36, 0.46\}$), Sonar achieved a F-score inferior to 0.4.

Brilia achieved the lowest performance for most of the analysed repertoires. The only exception is the oligoclonal sample generated with $\lambda_0 = 0.46$, where it outperformed Sonar. Compared to Sonar, we observed a notable difference in

the polyclonal and oligoclonal repertoires. However, in monoclonal repertoires, Brilia and Sonar achieved an equivalent performance, with very low values for recall and F-score.

8.3.2 *Artificial monoclonal repertoires*

Figures 16 and 17 show the clonal grouping performances for the artificial monoclonal repertoires AMD1 and AMD2, respectively (Section 8.2.2.1). Here we compared Agreeable to the other algorithms. To better interpret the performances, we used an alluvial diagram that represents flows between expected clones (left) and detected ones (right). Blue blocks represent expected clones, and pink or orange detected clones. Pink blocks contain only sequences of the pure B-cell lineage, while the orange blocks sequences from the polyclonal background. Thus, pink blocks represent true positives and orange ones false positives. Block height symbolises the size of a clone, that is, the number of sequences. The split counter (SC) counts the number of splits in the expected clone. The false-positive (FP) gives the number of sequences in the detected clones unrelated to the B-cell pure lineage. The output of tools on AMD1 showed three different profiles of sequence distribution. Agreeable obtained the best separation with zero SC and minimal FP (only 3). Partis and SCOPe also obtained the minimal FP, but higher SC, 4 and 5, respectively. Sonar and Brilia did not find any FP, but both tools carried out a significant number of splits, 90 and 64, respectively. Interestingly, for AMD1, Agreeable has accurately found the whole lineage without detecting separations as observed in Partis and SCOPe results. The sequences separated by other tools possess more than two consecutive Tyrosine (Y) in the CDR3 (ARDRRGEWPPSDYYYYYYMDV, ARDRRGEWPPSDYYYYMDV and ARDRRGEWPPTDYYYMDV). For sequences containing IGHJ6 genes, Agreeable tolerates the change in the number of Tyrosine for similar sequences and considers them as originating from the same lineage. This assumption is based on our observations from many NGS runs during routine analysis; its origin (real biological phenomenon or sequencing artifact) remains unknown. So, naturally, other tools are not tuned to recognize this particularity of the IGHJ6.

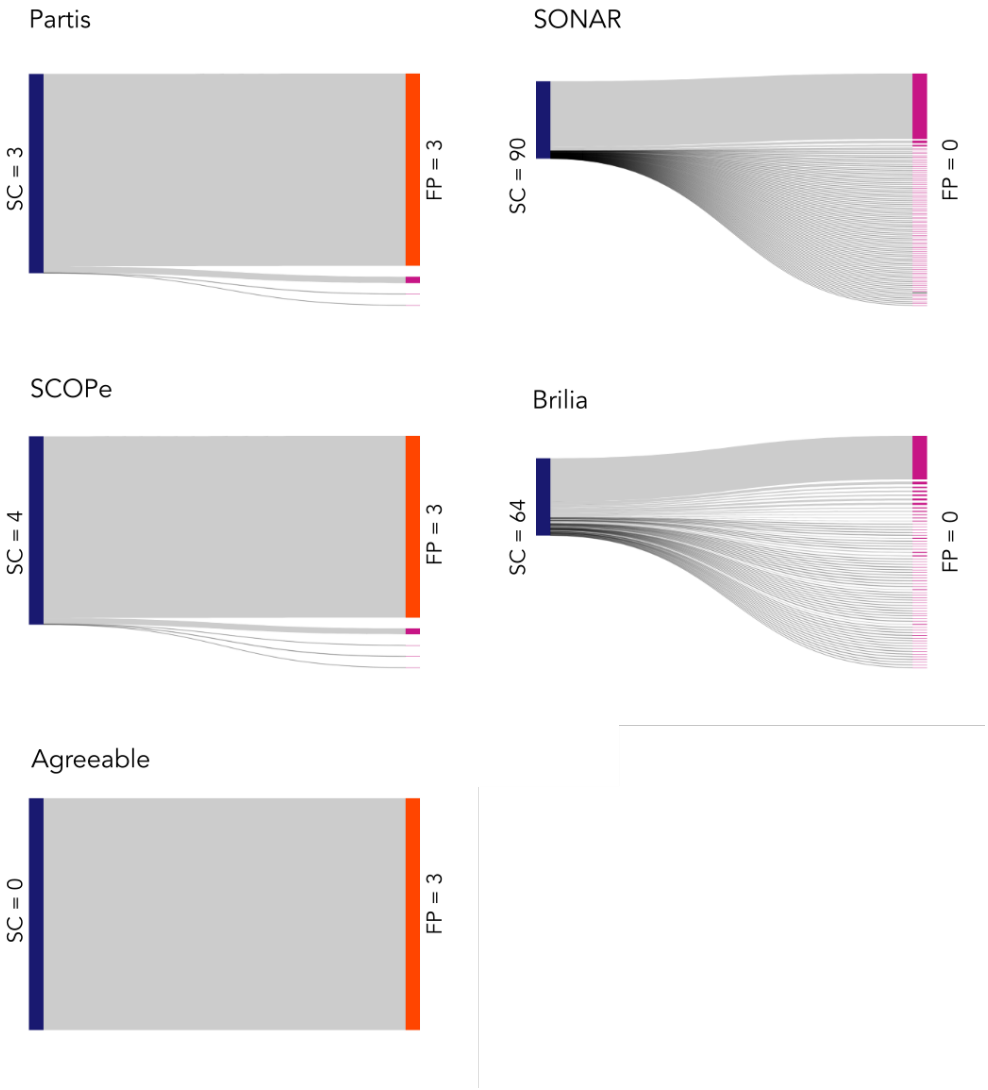

Figure 16: Performance evaluation of four different clonal grouping methods on **AMD1**

As observed for AMD1, Agreeable did not split sequences of the AMD2 B-cell lineage in different clones. Similarly, Partis and SCOPe achieved a SC equal to zero. However, Agreeable achieved the lowest FP number, followed by SCOPe and Partis. Sonar and Brilia still split the sequences of the B-cell lineage considerably. Sonar obtained less FP than Brilia.

For both data sets, we observed the same behaviour in the results of each tool. We can perceive that the granularity of sequence grouping is different. It was surprising to observe different clustering tendencies among tools with a similar clone definition. Partis, SCOPe and Agreeable have grouped most sequences in the pure B-cell lineage, and a few false positives were introduced. Sonar and

Brilia tend to over-split, but they have detected less FPs than Partis, SCOPe and Agreeable.

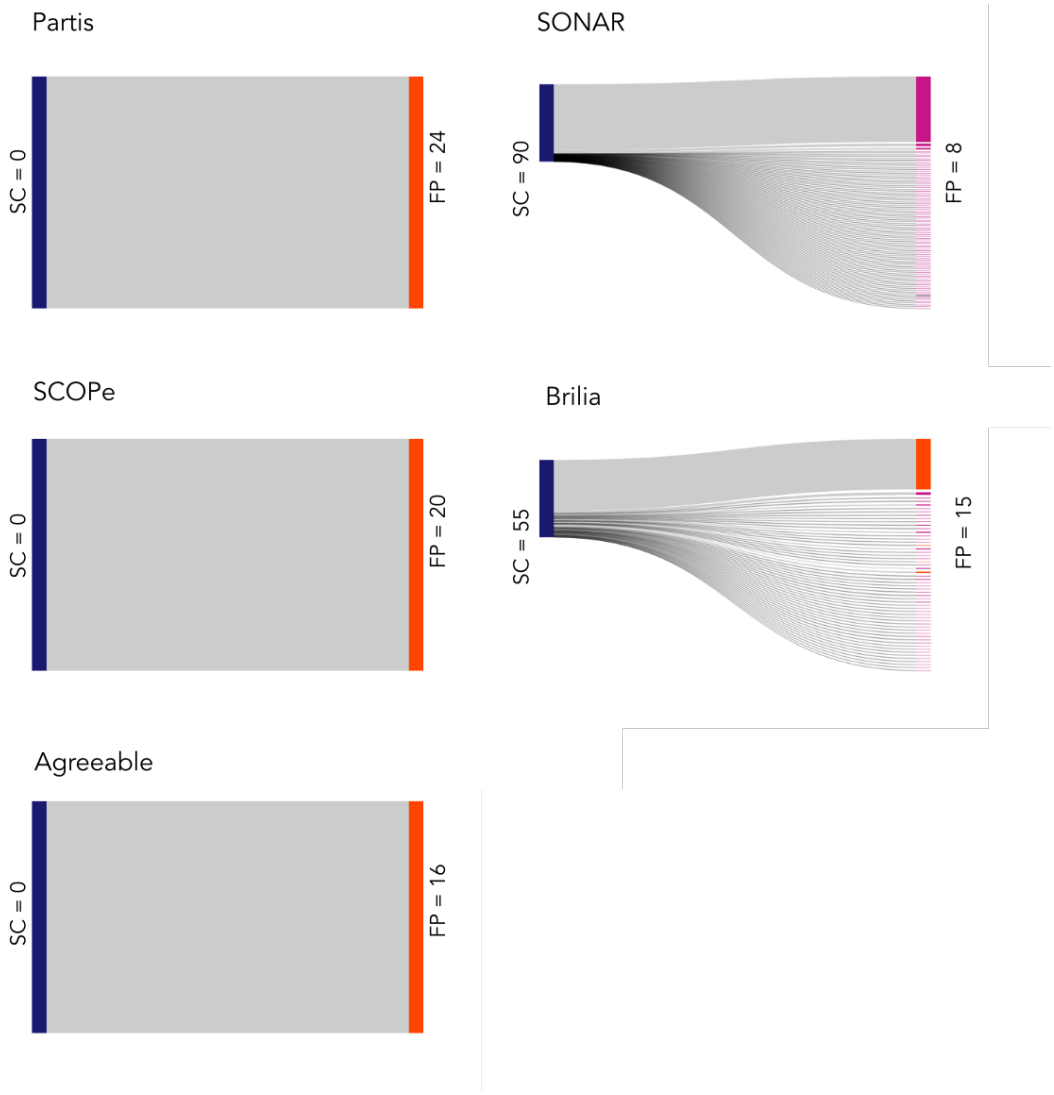

Figure 17: Performance evaluation of four different clonal grouping methods on **AMD2**

8.3.3 *Experimental benchmarks*

We have compared the output of each clonal grouping tool with the Agreeable's clustering results, using the method detailed in 8.2.3.1. Figure 18 and Table 4 show the result of analysing I1, a monoclonal repertoire with a major clone containing 98% of total sequences. As expected, the number of detected clusters by each tool is different. Brilia, and SCOPe have a similar number of clusters, at the order of magnitude of the Agreeable output. Partis has a smaller number of

clusters than Agreeable. On the other hand, Sonar has a significantly higher number of clusters. Considering both pairwise and closeness relationships, the precision values are high for all analysed tools. The overall Pairwise F-score value is greater than the closeness F-score value which was also seen in previous studies discussed in chapter 7. The low recall for the closeness calculation shows that the structure of reconstructed repertoire is different for each tool. Even the best value of recall, which is achieved by using Partis, is equal to 0.14 ,which is particularly low. Sonar has the lowest When counting clustering events: join, identical, split and mix (Section 8.2.3.1 and Figure 13), we observed the prevalence of mix events since compared tools have constructed the major clone differently. We can conclude this difference is in the major clone, because of the number of clusters labeled as Mix and their respective number of sequences.

Figure 19 shows the result of I2, a monoclonal repertoire with a biallelic Igh recombination. Partis has the closest detected repertoire architecture to the Agreeable with a closeness Fscore equal to 0.79(Table 4). It is interesting to note that Brilia, which has the second closest distribution to Agreeable while analysing the first dataset, has a very low overall F-score for this repertoire. One of the reasons is that the number of clustered sequences is significantly less for Brilia, since it only clusters functional sequences, therefore many unproductive sequences of this repertoire have been excluded from the analysis.

The third dataset analyzed in this study is a polyclonal repertoire. The first thing we notice is that all tools achieved a higher F-score compared to the monoclonal repertoire (Table 5). This happens because the task of correctly grouping together small B cell lignages is easier due to the much lower number of sequence variants as compared to an extremely developed lignage. Base on Figure 19 we can also see that Agreeable and all other tools share many identical clusters. Comparing the clustering events between Brilia and SCOPe, we can notice that SCOPe has more splits and joins than Brilia. The latter shared a higher number of identical clusters with Agreeable; however, these differences have translated to comparable F-scores. Sonar has the lowest F-score values due to the high number of splits.

Table 4: **Comparison of Agreeable with four different clonal grouping methods on the I1 data set**. The I1 data set has 33599 sequences distributed into 162 clones by Agreeable

	# analyzed seq	# clones	Pairwise			Closeness		
Tool name			Precision	Recall	F-score	Precision	Recall	F-score
Brilia	33578	161	0.99	0.99	0.99	1	0.06	0.11
Sonar	29335	3542	0.99	0.21	0.36	0.99	0	0
Partis	33585	73	0.99	0.99	0.99	1	0.14	0.25
Scope	33554	176	0.99	0.96	0.98	1	0.01	0.02

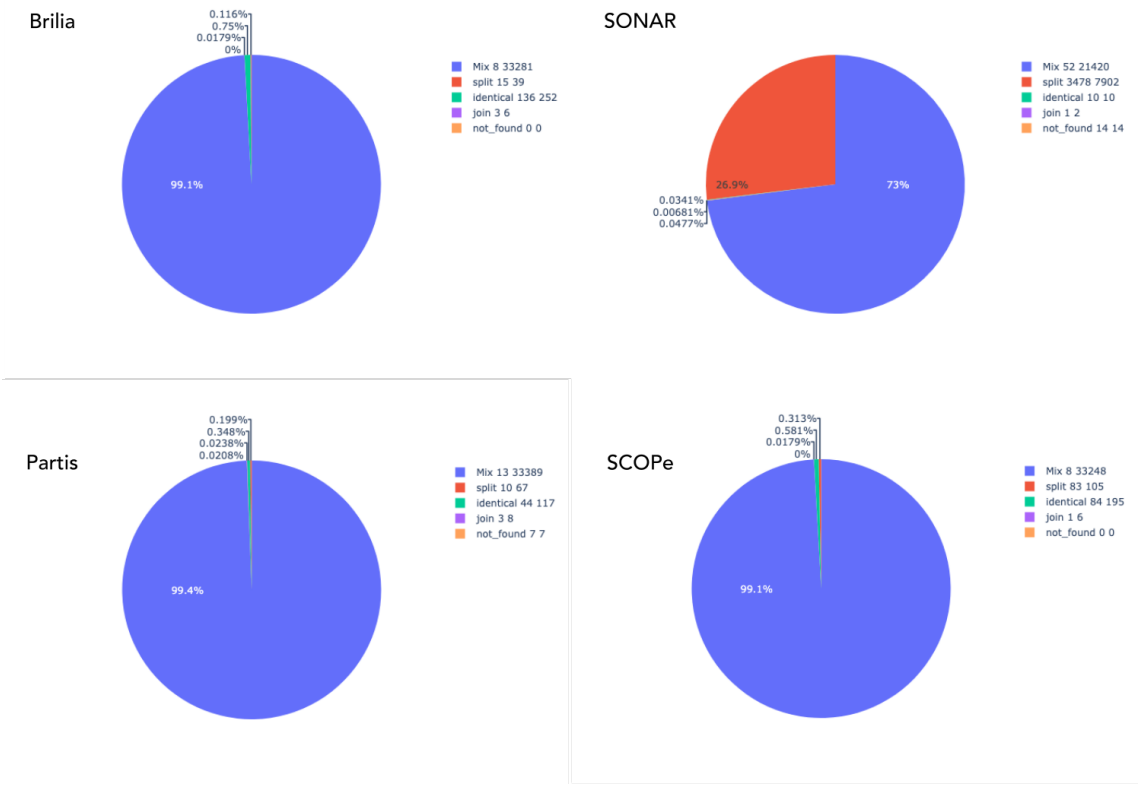

Figure 18: **Performance evaluation of four different clonal grouping methods on I1 dataset using "4 events" labeling system**

Table 5: **Performance evaluation of four different clonal grouping methods on the I2 dataset**. The I2 dataset has 70050 sequences distributed into 2398 clones by Agreeable

	# analyzed seq	# clones	Pairwise			Closeness		
Tool name			Precision	Recall	F-score	Precision	Recall	F-score
Brilia	90236	7234	0.99	0.11	0.2	0.99	0.01	0.01
Sonar	138439	1975	0.99	0.87	0.93	0.99	0.19	0.32
Partis	138688	927	0.99	0.99	1	0.98	0.66	0.79
Scope	138250	1387	0.99	0.98	0.99	0.97	0.5	0.66

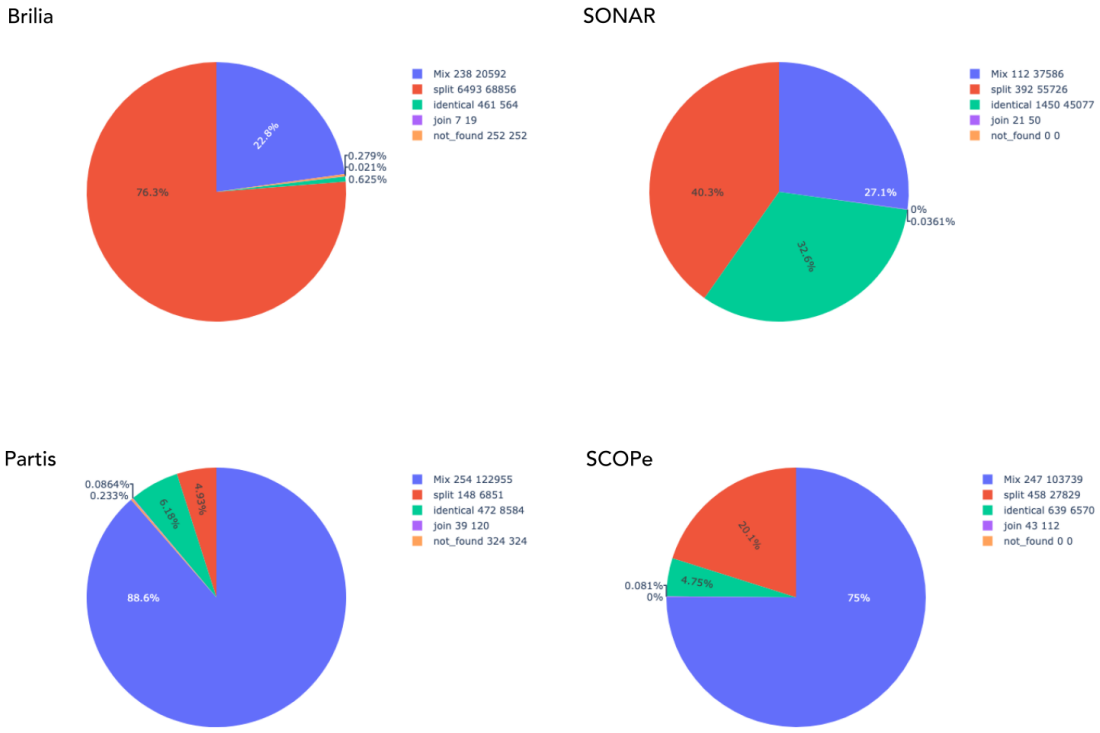

Figure 19: **Performance evaluation of four different clonal grouping methods on I2 dataset using "4 events" labeling system**

Table 6: **Performance evaluation of four different clonal grouping methods on the I8 dataset**. The I8 dataset has 70050 sequences distributed into 10461 clones by Agreeable

	# analyzed seq	# clones	Pairwise			Closeness		
Tool name			Precision	Recall	F-score	Precision	Recall	F-score
Brilia	68133	10461	0.96	0.90	0.93	0.96	0.76	0.85
Sonar	68171	17096	0.57	0.04	0.08	0.43	0.07	0.12
Partis	68327	10095	0.95	0.88	0.92	0.94	0.68	0.79
Scope	67312	7192	0.96	0.89	0.93	0.89	0.77	0.83

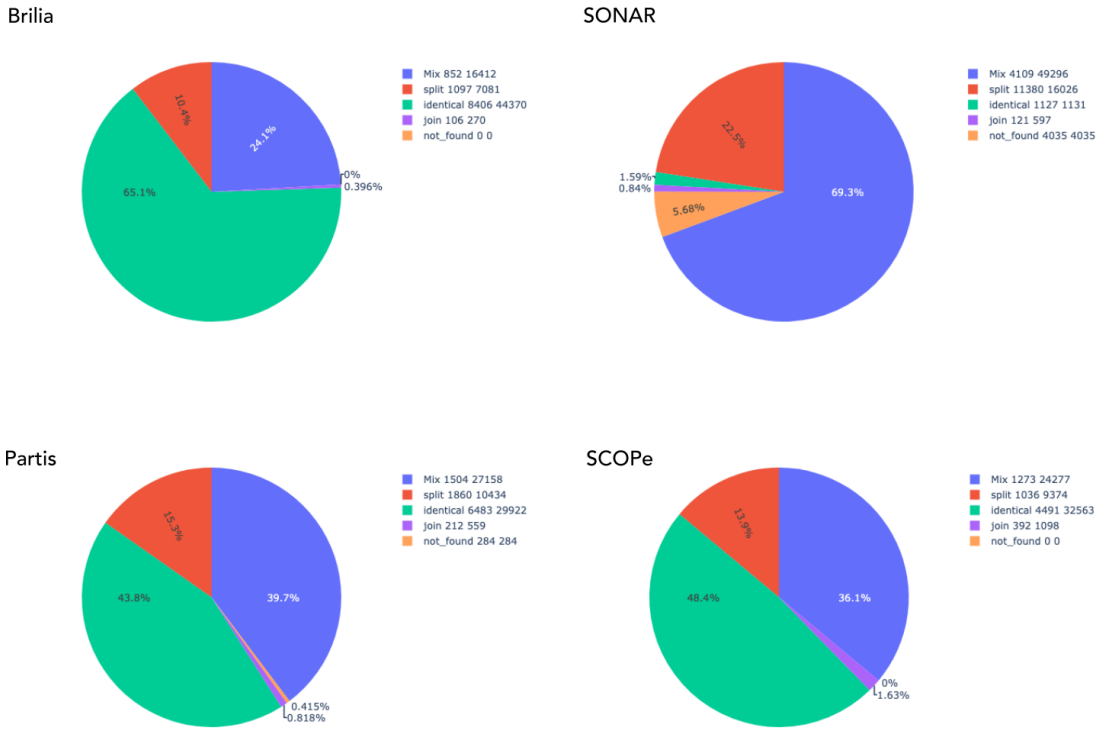

Figure 20: **Performance evaluation of four different clonal grouping methods on I8 dataset using "4 events" labeling system**

8.4 DISCUSSION

We have studied different performance evaluation measures and found out that from the same input data, using different BCR repertoire clustering tools, results in having :

- variable number of clusters
- variable cluster sizes (especially for the major clones of each repertoire)
- various standards to accept a sequence as an input.

The compartmentalization of the mentioned tools is not based on their overall value but their efficiency in a given context. The goal here is not to judge the performance of these tools but instead get a clear understanding of them and their assumptions to use them properly. Note that this is an ongoing study and we would like to expand the number of analyses in order to better comprehend the difference between tools. We have prepared an *in vitro* benchmark to evaluate the clone detection caliber of each tool in a more realistic setting by using data sets constructed by serial dilutions (1%, 0.1%, and 0.01%) of a known clone in a polyclonal background and then sequenced.

We are also aware of the sampling bias impact on the artificial monoclonal repertoires analysis, and it is crucial to repeat the sampling to reduce this effect. Unfortunately, however, running some tools like Partis are highly demanding in time, and we did not have the opportunity to do so yet.

Tools that cluster sequences based on their lineage are more appropriate for our project's objective, intraclonal analysis. Despite comparable output results, Partis, Agreeable, and SCOPe have significant differences in practical usability. Partis needs high computational power and analysing time. SCOPe has multiple preprocessing steps before clustering. Agreeable is relatively fast and does not require complicated preprocessing data preparation. These are the reasons why we have decided to use Agreeable as the clustering tool for this project.

Knowing the clonal architecture of a given repertoire, we will focus on BCR intraclonal analysis in the following chapter.

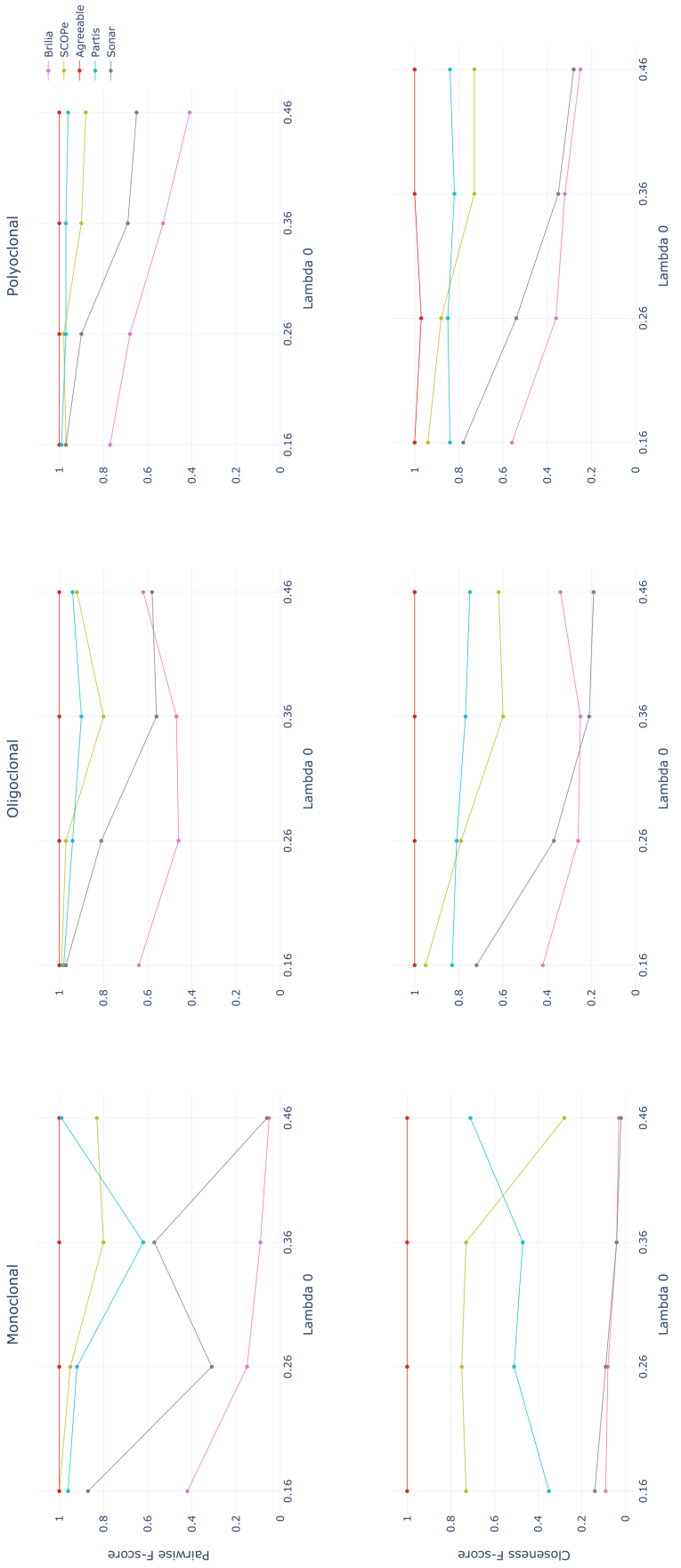

Figure 15: Performance evaluation of five different clonal grouping methods on simulated repertoires

9

RECONSTRUCTING THE EVOLUTIONARY HISTORY OF A BCR LINEAGE USING MINIMUM SPANNING TREE AND CLONOTYPE ABUNDANCES

9.1 INTRODUCTION

When exposed to an antigen, naive B cell's genes that encode Ig undergo multiple rounds of mutations, e.g. somatic hypermutations, and Darwinian antigen selection. This stage of the B-cell development is known as affinity maturation since it progressively increases the Ig's affinity for a pathogen-associated antigen. The naive B cell and its variants (generated during affinity maturation process) form a B cell lineage. The natural selection during affinity maturation permits a variety of Ig-antigen affinities. Among B cells with the same receptor's specificity, those having higher affinity for the antigen can proliferate, and those with lower affinity will be eliminated. Consequently, the number of unique variant sequences and their respective abundance provide a remarkable perspective on the ongoing evolutionary process, and can help to elucidate clonal selection. The BCR intraclonal analysis has several clinical applications such as developing effective vaccines, discovering therapeutic monoclonal antibodies, or diagnostic and prognostic immunoproliferative disease .

Genetic evolution as observed during affinity maturation is often studied through phylogenetic inference, a well-known methodology that describes the evolution of related DNA or protein sequences in various species. Theoretically, phylogenetic inference methods could reconstruct BCR lineage trees if we replace species with sequences having different affinities. However, in a phylogenetic tree, the root is usually unknown, the observed sequences are represented usually only in the leaves, and the inner nodes represent the relationships amongst sequences. Conversely, in a BCR lineage tree, the root or the BCR sequence of the naive B cell giving rise to the lineage is available. Such ancestral sequence is

typically estimated by aligning each sequence with germline genes of a reference genome.

B cells with different BCR affinities can coexist; therefore, the observed BCR sequences can be leaves or internal nodes in the tree. Due to simultaneous divergence, multifurcations are common. Moreover, Ig sequences are under intense selective pressure, and the neutral evolution assumption is invalid, and the SHM process occurring during the affinity maturation is highly nucleotide-context-dependent. In this scenario, conventional phylogenetic tree algorithms are not appropriate for creating BCR evolution lineage trees. The performance of such methods varies substantially in terms of the tree topology and the ancestral sequence.

Some computational tools have been designed specifically for the reconstruction of BCR lineage trees. IgTree [85] employs maximum parsimony criterion to find the minimal set of events that could justify the observed sequences. It first constructs a preliminary tree, which only contains observed sequences, then uses a combined score, based mainly on sequence mutations, to gradually add internal nodes. GCTree [67] employs the maximum parsimony principle, and additionally it incorporates the cellular abundance of a given genotype in phylogenetic inference. This information is used for ranking parsimonious trees, obtained by dnapars [86] with the assumption that the more abundent the parent is,the more likely it is for it to generate mutant decendants. GCtree uses a likelihood function based on the Galton-Watson Branching process [87]. It is an accurate method, but its computational complexity is exorbitant, especially for a high number of sequences. GlaMST [88] is another method for reconstructing BCR lineage trees. It is a minimum spanning tree (MST)-based algorithm [89], and iteratively grows the lineage tree from the root to leaves by adding minimal edge costs. GLaMST is more efficient than GCTree, but it ignores genotype abundance information.

Here we propose ClonalTree, a method to reconstruct BCR lineage trees that combines MST and genotype abundance to infer maximum parsimony trees. ClonalTree starts from the root (the ancestral sequence) and iteratively adds nodes to the tree presenting minimal edge cost and maximum genotype abun-

dance; therefore, it optimises a multi-objective function rather than a single function based only on edge costs as implemented in GLaMST. Using simulated and experimental data, we demonstrate that ClonalTree outperforms GLaMST and achieves a comparable performance to the performance of GCtree. ClonalTree has a lower time complexity, making it more suitable for reconstructing phylogenetic lineage trees from high throughput BCR sequencing data obtained in the clinical applications.

9.2 MATERIAL AND METHODS

We start from a formal description of the BCR lineage tree reconstruction and minimum spanning tree. Next, we describe how we modified Prim's algorithm to incorporate genotype abundance information. Ultimately, we show how trees can be improved by creating intermediate nodes that describe non-observed sequences or editing operations.

9.2.1 *Problem statement*

Given a set of observed BCR sequences and an ancestral sequence (root), we look for a minimum-sized directed tree structure, otherwise known as the maximum parsimony tree, so that, all observed sequences are reachable from the root, vertices (nodes) represent BCR sequences or their relationships, and the weight of edges connecting vertices represents mutation, insertion and deletion operations.

9.2.2 *Minimum spanning Tree*

Given a connected graph (V, E), where V is the vertices, E the weight edges, its minimum spanning tree (MST) is a subset of vertices and edges that form a tree (a graph without cycles/loops) so that the sum of the weights of all the edges is at minimum. For a given connected graph, several MSTs can exist. All trees have the same sum of weights, but their topologies are different. The MST construction is a greedy approximation algorithm in which edges are sorted according to their weights and selected with some criteria. Greedy algorithms

normally find a local optimum solution, which may eventually lead to globally optimised solutions.

9.2.3 *A modified Prim's algorithm*

Prim's [90] and Kruskal's [91] are algorithms for finding the minimum spanning tree of a graph. Both are greedy approaches and present low time complexity. However, Prim's algorithm runs faster than Kruskal in dense graphs. Therefore, we used a modified version of Prim's algorithm to construct BCR lineage trees. We start at the root and add all its neighbours to a priority queue. We then iteratively extract from the priority queue the node with the lowest weight and highest genotype abundance. A node and an edge will be added to the tree if no cycle is formed. When adding a node to a tree, all its neighbours are included in the priority queue. We keep on adding nodes and edges until we cover all nodes. In order to decrease the time complexity of the algorithm, we add each node only once at the priority queue. Prim's algorithm has only one objective function, which minimizes the sum of edge weights. Here we include a second objective function to maximize genotype abundance. If a set of edges have the same weight, we will choose the one that connects nodes with high abundance. Prim's algorithm has a time complexity of $O(|V|2)$ in the worst case, but can be improved up to $O(|E| + log|V|)$ when using data structures based on Fibonacci heaps[92] Figure 21 shows a simple example of the tree construction process.

9.2.4 *Editing the reconstructed lineage tree*

A greedy algorithm never reconsiders its choices. This is the main difference between this algorithm and an optimal algorithm, which is exhaustive and always finds the best solution. A possible amelioration of our algorithm would be to edit the obtained lineage tree. We implemented two strategies: add the unobserved intermediate nodes to the tree and detach/reattach subtrees.

Unobserved internal nodes represent unobserved sequences that were not sampled or disappeared during the affinity maturation process. In those cases, the evolutionary relationships are also missed, but one can try to reconstruct

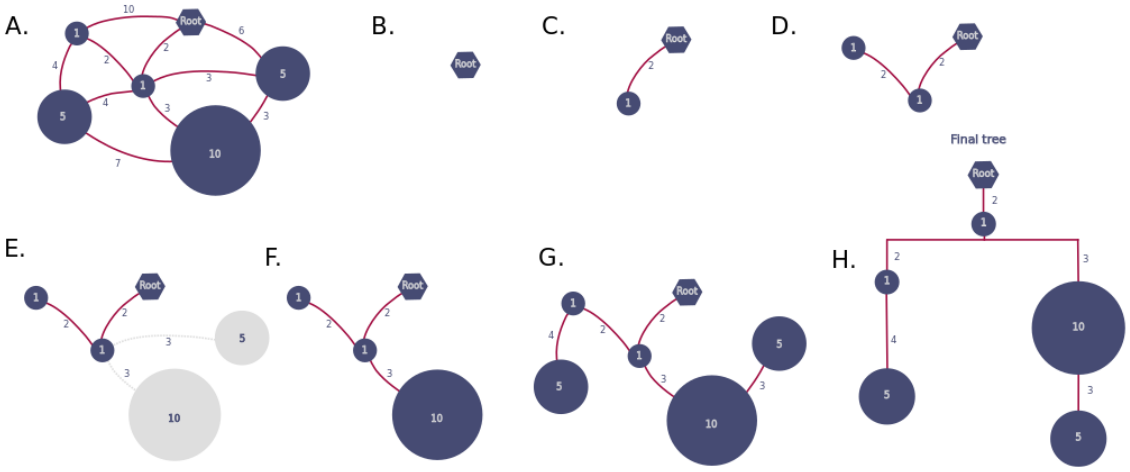

Figure 21: **ClonalTree construction example**. Given a connected weighted graph (A), we start by placing the ancestral sequence or root (B), we iteratively add nodes to the tree with the lowest edge weights and highest genotype abundances (C,D), when edges have the same weight (E) we choose those connected to the node with higher abundance (F), we repeat this process until all nodes were added to the tree (G), the final tree is shown in (H).

them. Unobserved internal nodes represent unobserved sequences that were not sampled or disappeared during the affinity maturation process. In those cases, the evolutionary relationships were also lost. One way to recover them is to analyse the reconstructed tree to identify common ancestors not yet represented. This process is similar to building a phylogenetic tree among species, where unobserved internal nodes represent common ancestors of descendants. In a classical phylogenetic tree, only leaves' nodes are observed, and all internal nodes are unobserved, while in a BCR lineage tree, internal nodes can be observed or unobserved. We add unobserved internal nodes when we detect a common ancestral not represented in the tree. It generally happens when we observe a distance between sister nodes that is smaller or equal to the distance for their parent, see an example in Figure 22-A. Once we find the exact position of an internal node in the tree, we connect it to the observed nodes by direct edges (see Figure 22-B).

We can detach a subtree from an internal node by removing its edge and reattaching it to another internal node or leaf. We perform this editing operation if it can reduce the size of the lineage tree by keeping the overall cost. We consider all branching nodes (i.e. nodes with more than one descendent) for this edition operation. For each node under edition, we try to detach it and reattach it to any

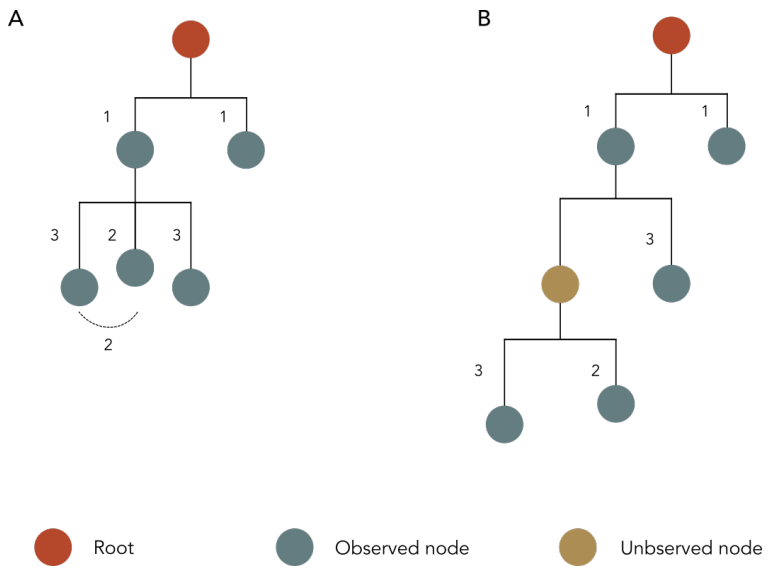

Figure 22: **Editing the reconstructed lineage tree by adding internal nodes**. When the distance between two sister nodes is smaller or equal to the distance to their parent, we add an unobserved internal node as the common ancestor of the two sister nodes.

other node in the lineage tree. If this operation reduces the tree size, we accept it and examine the resulting lineage tree again for additional edition operations that may further reduce the tree size. We repeat this process until no editing operation can reduce the tree size (see an illustration in Figure 23).

9.2.5 *Tools used in the comparisons*

We have selected two tools to compare with Clonaltree. GCTree [67] incorporates genotype abundance information in the parsimony tree inference, and GLaMST [88] uses Minimum Spanning Tree (MST). Since in the publication presenting GCtree, the authors have shown the higher performance of these tools compared to other available methods, we have decided to keep only these two approaches for the evaluation.

9.2.6 *Data sets*

We used two types of data sets to evaluate ClonalTree performance and compare it with other algorithms: artificial and experimental. Artificial data were

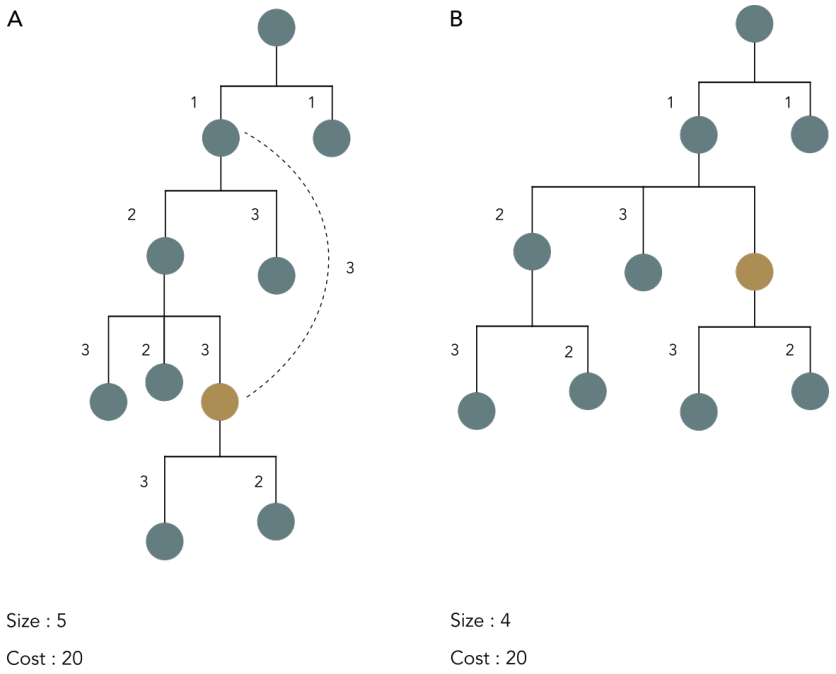

Figure 23: **Editing the reconstructed lineage tree to reduce the size of the tree while keeping its overall cost.** In this example, the total cost of the tree, or the sum of edge weights, remained the same while the size of it has been reduced.edge weights represent the Hamming distance between sequences

produced by GCTree simulator [67], while one of two experimental data sets was created during CLL routine diagnostic procedures at the Pitié Salpêtrière hospital and the other one was collected from public data. In order to create artificial lineage trees, we used the B-cell lineage simulator provided by GCTree. The simulator produces B-cell lineage by randomly selecting IGHV, IGHD and IGHJ germline genes from the IMGT database; then, nucleotide(s) can be added to or removed from the junction region: IGHV-IGHD and IGHD-IGHJ. Next, it performs a branching process, and point mutations can be included in the descendants. Somatic hypermutations are simulated by a sequence-dependent process, where mutations are preferentially introduced within specific hot- and cold-spot motifs [93]. We kept simulator default parameters and generated 92 artificial lineage trees. The sizes of simulated trees ranged from 6 to 99 nodes, being the number of observed sequences between 20 and 200; the degree of roots vary from 1 to 42, and depth trees from 2 to 7, see Table 7.

The public experimental data set contains IGHV gene sequences from 48 germinal B cells of a lineage sorted from a brainbow mouse using multicolor fat

Table 7: **Characteristics of artificial lineage trees**

	Min	Max	Mean
Tree size	6	99	33.82
Observed seq	20	200	81.98
Root degree	1	42	11.54
Tree depth	2	7	4.29

mapping [94]; we label this dataset as "32-IGHV". In the experiment, the authors performed single-cell mRNA sequencing of the IGH and IGL loci of the 48 germinal B cells, resulting in 32 distinct IGH and 26 distinct IGL genotypes, with different SHM mutation rates acquired during affinity maturation. The "32-IGHV" data set, along with IGL sequences obtained from the same set of cells, were used to evaluate the robustness and accuracy of GCtree [67]. In their paper, Masten and colleagues showed that for both loci, parsimony analysis results generated by dnapars [95] did not achieve a consensus, and they demonstrated that GCtree correctly resolved this degeneracy by incorporating abundance information. Therefore, we used the "32-IGHV" data set and compared the inferred trees produced by GLaMST and Clonal tree to GCtree trees considered ground truth. The second experimental data set was generated by sampling sequences from the most abundant clone of a CLL monoclonal repertoire. It contains 20 unique sequences with different abundance, totalizing 3406 sequences; we label this data set as "20-major".

9.2.7 *Tree comparison and evaluation*

To measure the performance of B-cell lineage reconstruction algorithms, we used multiple metrics to compare tree topologies : graph editing distances [96], and ancestral sequence inferences (most recent common ancestor [67], and correctness of ancestral reconstruction [97]).

9.2.7.1 *Graph Editing Distance*

Let G1 and G2 be two graphs; the Graph Editing Distance (GED) finds the best set of graph transformations capable of transforming G1 into G2 through edit

operations on G1. A graph transformation is any operation that modifies the graph : insertion, deletion, and substitutions of vertices and edges. GED is similar to string edit distances such as Levenshtein distance [66] when we replace strings by connected directed acyclic graphs of maximum degree one. We used two versions of GED, one applied to the whole tree (GED tree-based) and the other one applied to each branch separately (GED path-based). The latter version is more stringent than the first one since any difference in the path from each leaf to the root is considered a transformation. Figure 24 illustrates an example of such graph distances.

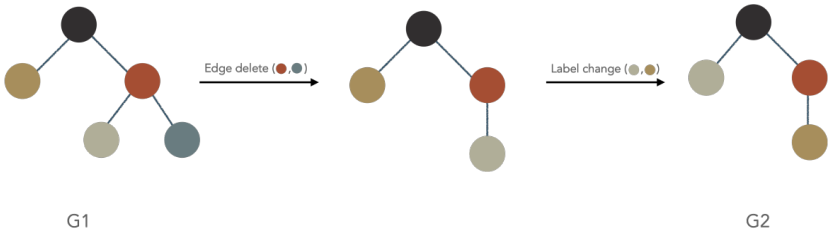

Figure 24: **Example of Graph Edit Distance calculation.**The GED in this example is equal 2.

The problem of computing graph edit distance is NP-complete [98], and there is no optimal solution in a reasonable time. This problem is hard to approximate, and most approximation algorithms have cubic computational time [99, 100]. Here we could use an optimal algorithm implementation since the size of evaluated trees is small.

9.2.7.2 *The Most Recent Common Ancestor (MRCA), and The Correctness Of Ancestral Reconstruction (COAR)*

For a given pair of leaves in the tree, the MRCA distance is the average Hamming distance between the true and the inferred ancestral sequences. COAR distance is a metric that emphasizes the importance of correct ancestral reconstruction and does not penalize the minor topology difference between a true tree and inferred tree when the ancestral reconstruction is accurate. Figure 25 shows for each metric, which are the nodes that have been compared between the true and the inferred tree.

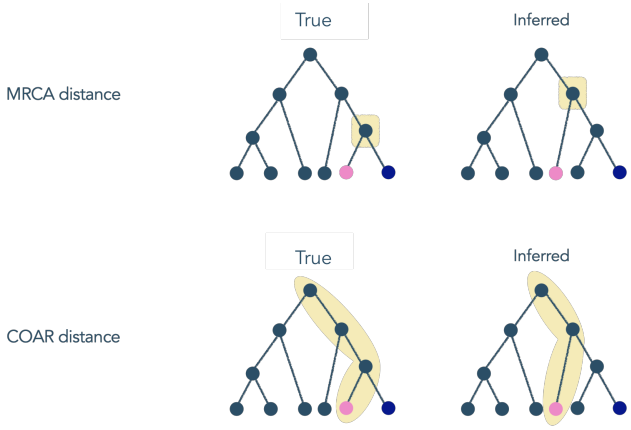

Figure 25: **Nodes compared between two trees with MRCA and COAR metrics.**

9.3 RESULTS

9.3.1 *Reconstructing BCR lineage trees from simulated data*

To evaluate ClonalTree performance and compare it to GCTree and GLaMST, we generated simulated datasets using 92 different simulation settings, varying the root sequence genes and the relative probabilities of mutation, insertion, and deletion (see section7.2.2 and Table 7). The artificial lineage trees served as ground truth that we would like to recover using BCR lineage tree algorithms. Thus, the performance measures how close reconstructed trees are from simulated ground truths. Several approaches can quantify this. Here we have used graph editing distances (GED) that measure the dissimilarity between two graphs/trees and two distances related to the correctness of common ancestral inferences. We have computed two types of GED distances, based on the entire tree (GED tree-based) and in all separate paths (GED path-based), see Section 9.2.7.

Figure 26 shows boxplots of GED distances for each compared method on the 92 simulated lineages trees. We observe that GCtree and ClonalTree had comparable performances. Reconstructed BCR lineage trees of both tools have similar topologies, while trees produced by GLaMST are pretty different. Figure 26-A shows GED tree-based distances, while Figure 26-B shows GED path-based distances. For GED tree-based distances, median values were 0, 2, and 12 and the highest distance 37, 38, and 120 for GCtree, ClonalTree and GLaMST, respec-

tively. GLaMST presented the highest median value and the highest distance. ClonalTree produced 39 correct trees (GED tree-based distance equal to zero), while GLaMST produced only two trees.

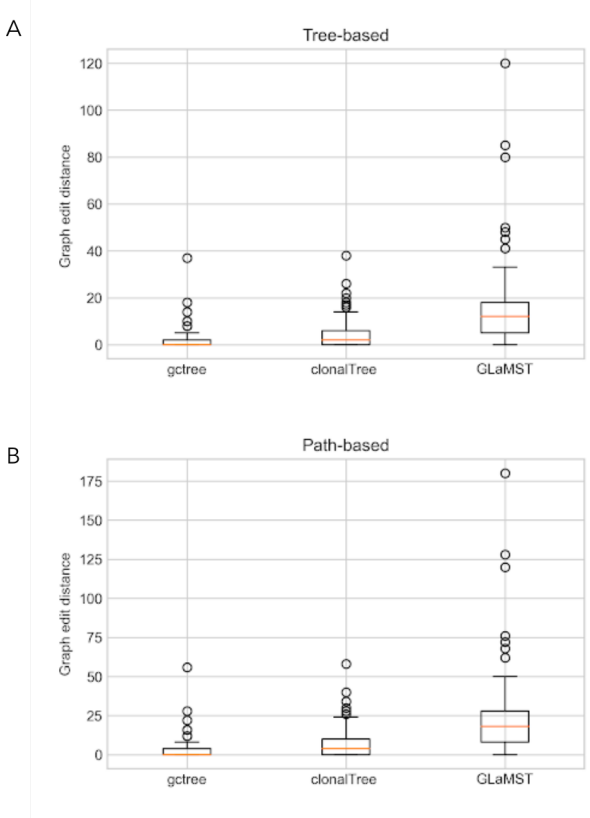

Figure 26: **Performance comparison between GCtree, ClonalTree, and GLaMST using GED distances.**

GED path-based distances compare each tree path, from leaves to the root, between inferred and ground truth trees; therefore, it is more sensitive to topology changes, and the obtained distances were higher than GED tree-based. Figure 26-B confirms that GCtree and ClonalTree reconstruct BCR lineage trees with similar paths. Median values were 0, 4, and 18, the highest distance 56, 58, and 180 for GCtree, ClonalTree and GLaMST, respectively. As observed for GED tree-based, GLaMST presented the highest median value and the highest distance. ClonalTree produced 39 correct paths (GED path-based distance equal to zero), while GLaMST produced only one path.

In order to better evaluate the performances, we devide the trees into three categories according to their number of sequences: small (between 30 and 50),

middle (between 60 and 80), and big (having more than 90 sequences). We observed a slight difference between GCtree and ClonalTree, mainly in small and big groups. On the other hand, we observed that GLaMST had difficulties in all groups; GED distances were increased as the number of sequences grew, see Figure 27.

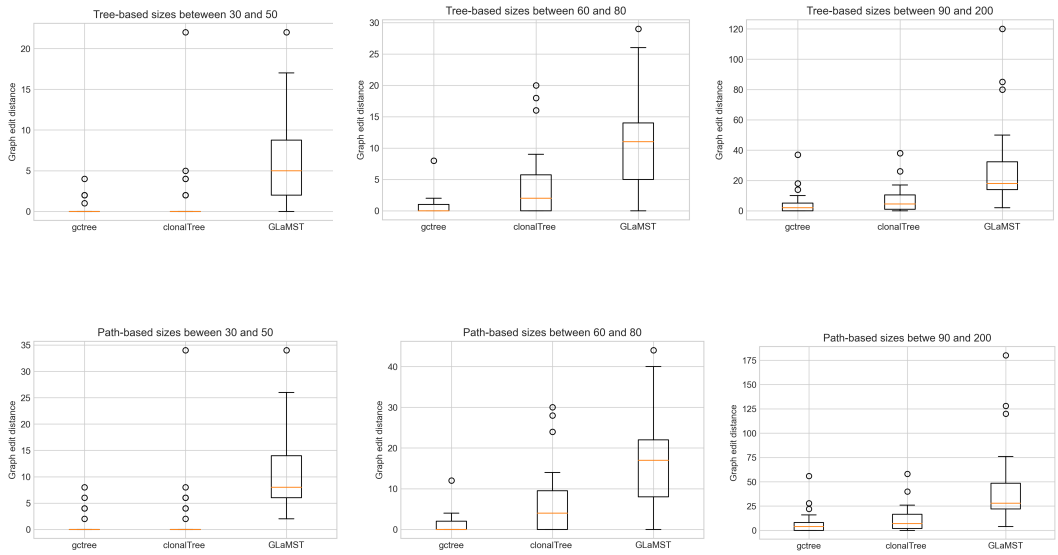

Figure 27: **Performance comparison among GCtree, ClonalTree, and GLaMST using GED distances on three categories of trees.** The categories are based on the tree sizes.

Although GED measures the accuracy of lineage reconstructed trees, it is very dependent on the tree topologies. It should also be an accurate estimator of the correctness of ancestral reconstruction without penalizing minor differences in the tree topologies. For that, we used two metrics: the MRCA and the COAR. MRCA distance focuses on the most recent common ancestor and considers the entire evolutionary lineage. On the other hand, COAR measures the correctness of ancestral reconstruction from the root to any leaf.

We first compared ClonalTree and GLaMST to ground truth trees and then with GCTree. It is important to perform this comparison since it gives us a basis

for evaluating these methods on experimental datasets, where the true lineage evolution is unknown. Figure 28-A shows MRCA distance distributions for ClonalTree and GLaMST when compared to ground truth trees, while Figure 28-B to GCTree. For both plots, we observed better performance for ClonalTree that could reconstruct recent ancestrals more accurately.

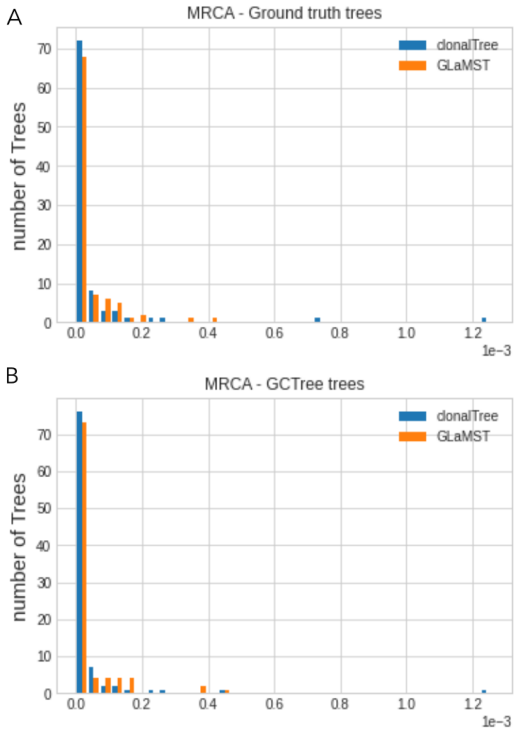

Figure 28: **Performance comparison among GCtree, ClonalTree, and GLaMST using MRCA distance.**

Figure 29-A shows COAR distance distributions for ClonalTree and GLaMST compared to ground truth trees, while Figure 29-B shows comparisons with GCTree trees. As observed for MRCA, ClonalTree outperformed GLaMST for both comparisons. We noted a slight difference between the two plots, the difference between ClonalTree and GLaMST being more significant on the plot of Figure 29-A then on the plot of Figure 29-B.

The number of inferred trees with null COAR distances is significantly larger for ClonalTree than GLaMST, meaning that ClonalTree reconstructs the whole evolutionary path from leaves to the root more accurately.

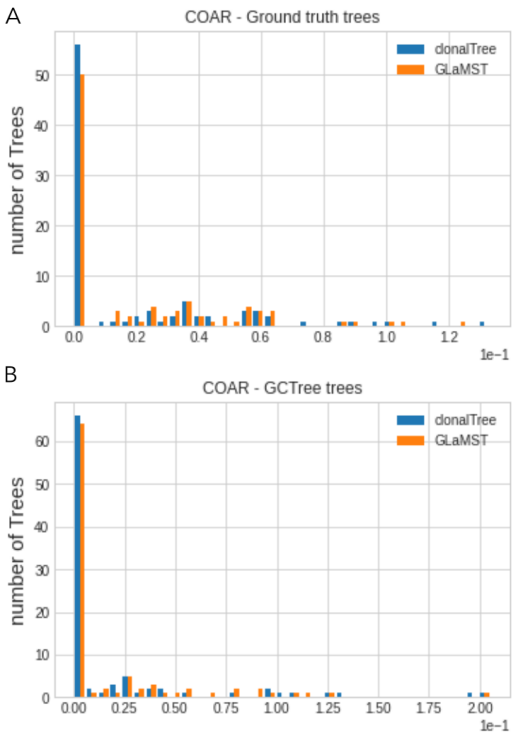

Figure 29: **Performance comparison among GCtree, ClonalTree, and GLaMST using COAR distance.**

9.3.2 *Biological validation using BCR sequencing data*

We then performed a biological validation on two experimental data sets: 32-IGHV and 20-major (see Section 9.2.6). Since ground truth trees are unavailable for these data sets, we compared the inferred trees of ClonalTree and GLaMST with the trees inferred by GCtree. We consider GCTree trees as references for the empirical validation because it achieved the best performance on simulated data sets. Table 8 shows tree distances for the two experimental data sets. We observed that the trees generated by ClonalTree are closer to GCTree trees; all distance metrics are smaller for ClonalTree. We noted a slight difference between ClonalTree and GLaMST MRCA values on both data sets, but a significant difference between COAR values. ClonalTree has reconstructed more accurately the entire evolutionary lineage than GLaMST.

We also compared the shape of trees through a set of extracted features, such as depth, root degree, number of leaves, etc. Table 9 shows six shape features

Table 8: **Performance evaluation of ClonalTree, and GLaMST on real BCR repertoire datasets.**

	32_IGHV		20_major	
	ClonalTree	GlaMST	ClonalTree	GlaMST
GED tree based	10	47	3	36
GED path-based	22	230	18	1616
MRCA	0,000515957	0,000570489	0.00064	0.00075
COAR	0,023255814	0,544186047	0.11759	0.78161

for the three algorithms. All methods infer trees with similar shape features; the main difference is the features related to the total number of nodes.

Table 9: **Comparison of GLAMST and ClonaTree with GCtree using 7 tree features and two metrics based on two real datasets.**

		32_IGHV dataset			20_major dataset		
		Features			Features		
		ClonalTree	GlaMST	Gctree	ClonalTree	GlaMST	Gctree
D-Root	Out-degree of the root of the tree	1	1	1	1	1	1
Depth	Maximum depth of the tree leaves	5	10	6	6	57	7
D-Avg	Average out-degree of all tree nodes	0,98	0,99	0,98	1	1	1
Avg-dist-root	Average distance from root to all nodes	5,24	5,26	5,19	46	47	47
Min-dist-Root	Minimum distance from root to any leaf	4	4	4	44	44	44
Tree Size	Total number of tree nodes	43	79	48	21	74	22
Leaves	Number of leaves	36	35	36	12	13	13

9.4 DISCUSSION

We present ClonalTree, a novel approach to reconstructing BCR lineage evolution that combines the minimum spanning tree algorithm with the genotype abundance of distinct BCRs. Using simulated BCR lineage trees, we demonstrated that such combination improves the accuracy of the inferred tree. ClonalTree outperformed GLaMST, a method based only on the minimum spanning tree algorithm. Genotype abundance seems to be valuable information and allows our method to be a competitor of GCTree, an accurate but time-consuming method.

ClonalTree also achieved acceptable results on experimental BCR sequencing data where the genotype abundance is evident. In some pathological situations, BCR lineage trees can be deep and involve a large set of sequences. In such conditions, GCtree becomes highly time-consuming. For instance, it takes several days to analyze 176000 sequences sampled from the most abundant clone of a monoclonal repertoire. It is impractical to use such a tool in clinical settings, where computational time and accuracy are essential. A way to address this issue is to sample sequences before reconstructing the lineage tree. It could allow researchers to use very precise yet highly time-consuming methods like GCtree. The problem is that less abundant genotypes having an essential contribution to the correct inference of BCR intraclonal evolution trees are likely to be neglected during the sampling process, as illustrated in Figure 30. The tree on the top section was generated with the 30 most abundant genotypes, while the bottom tree was obtained by pruning the complete tree until it achieves 30 nodes (see pruning algorithm details in Section 10.2.3). We observed that tree topologies are significantly different in both scenarios. Although sampling can decrease the computational time, it can cause incorrect interpretations of intraclonal evolutionary events. For instance, the clonotype represented by the dark blue circle has a different evolutionary role in each tree shown in Figure 30.

ClonalTree can be an alternative to GCTree since it can achieve comparable results with a lower runtime. The high performance of our method allows the users to consider all the available sequences when reconstructing lineage trees. It can help researchers to understand B cell receptor affinity maturation lineages, mainly when sequence data from dense quantitative sampling of diversifying loci are available. Integrating ClonalTree into existing BCR sequencing analysis frameworks can lead to significant improvements in the lineage tree reconstruction.

Nonetheless, visual interpretation of such large trees is not trivial, and will be discussed in Chapter 10. In order to put all the above-mentioned information to use, we have been developing a versatile interactive visualization pipeline with the purpose of analyzing BCR repertoires at the clonal and intraclonal level. The detailed description and various functionalities of this tool will be also discussed in Chapter 10.

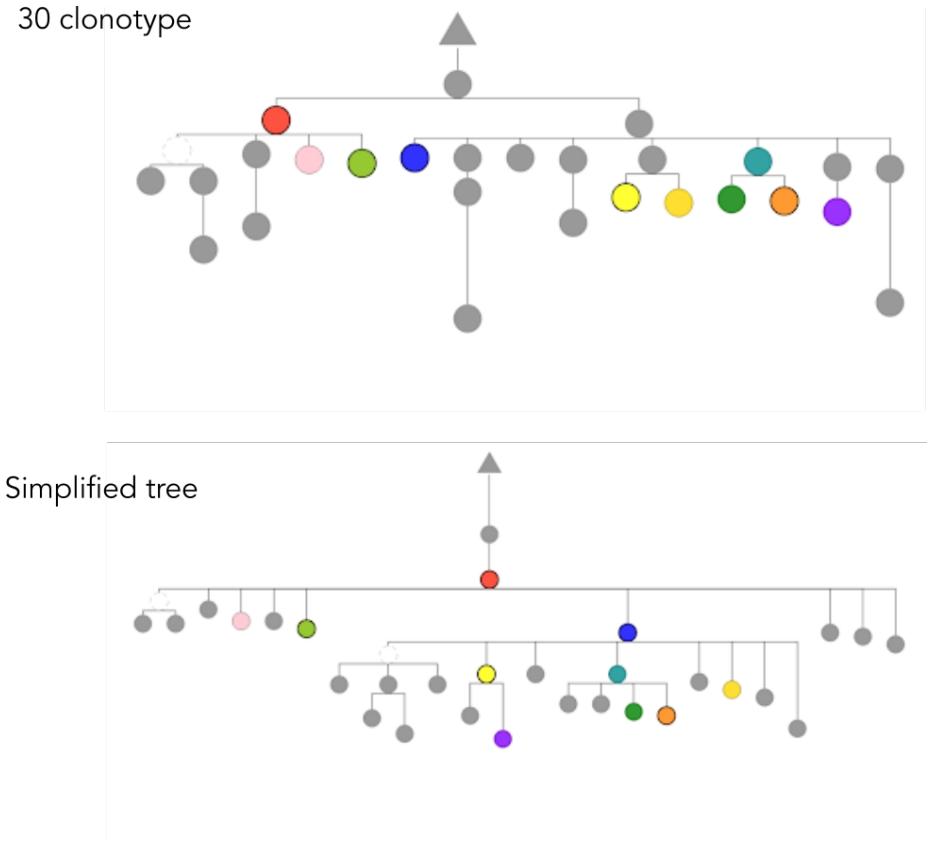

Figure 30: **Two different evolutionary histories of the same B cell lineage.** The tree on top is constructed by using the 30 most abundant clonotypes, while the tree on the bottom is the simplified version of the tree constructed by using the entire collection of clonotypes in the BCR lineage.

10

VICLOD, A TOOL FOR VISUALIZING B CELL REPERTOIRE'S CLONAL AND INTRA-CLONAL DIVERSITIES

BCR repertoire analysis by high throughput sequencing has research and clinical interests, but it is still a challenge to enable immunologists and researchers to explore their data to discover discriminating repertoire features on their particular examinations. As discussed in chapter 5, data visualisation is an efficient communication model, independent from any discipline-specific language, and can provide the maximum clarity and transparency. The flexibility of data visualisation enables users to interactively browse their data and easily interpret their results. To promote repertoire visualisation and complement experimental studies of BCR repertoires, we developed ViCloD, a web-based interactive tool that provides a visual analysis for repertoire clonality and intra-clonal diversity. ViCloD is compatible with clinical applications since it can analyze hundreds of thousands of BCR sequences in a reasonable amount of time. In this chapter, we will explain the general framework of ViCloD and outline how the generated outputs can be used for answering a set of biological questions.

10.1 PIPELINE

The current version of ViCloD is designed to analyze the IGH repertoire. After sequencing, IGH sequences should be annotated to determine their V, D, J genes and CDR3 region. ViCloD accepts data coming directly from IMGT/HighV-QUEST, which is the international standard web software for parsing BCR sequences. However, other annotating tools can be used as long as they provide the minimum information required for the analysis. The input file of the ViCloD pipeline is an AIRR formatted file [26]. Users can upload the AIRR file on the first page of the webserver and provide their email address to receive a link to their outputs. The analysis is relatively fast, for instance, it takes around three minutes to analyze a monoclonal repertoire containing 270 000 sequences. The

required time can vary depending on the structure of the sample (the size of clusters and sequence mutations).

Once sequences are uploaded, the first step is to group clonally-related sequences together. For that, we used Agreeable, detailed in chapter 7, and performed clonotype grouping based on IMGT definition. Then for the five most abundant clones of the repertoire, we inferred the lineage tree using ClonalTree algorithm, demonstrated in chapter 9. ClonalTree requires a file in FASTA format containing sequences and their abundances; we also need to determine which sequence is the hypothetical naive (the root tree). Each sequence in the FASTA file represents a clonotype. Since sequences in a given clonotype can differ due to SHM, we chose the most abundant sequence to represent each clonotype. The clonotype abundance is the total number of sequences. To compose the hypothetical naive, we considered the germline sequences of the corresponding IGHV and IGHJ genes provided by IMGT/HighV-QUEST; for the junction sequences, we took it from those with the lowest number of mutations on the IGHV gene, when compared to the germline determined by IMGT/HighV-QUEST. Eventually, we concatenate the three parties to obtain the hypothetical naive sequence.

10.2 DESCRIPTION OF FUNCTIONALITIES

In figure 31 the ViCloD workflow is illustrated and the list of different functionalities is presented below.

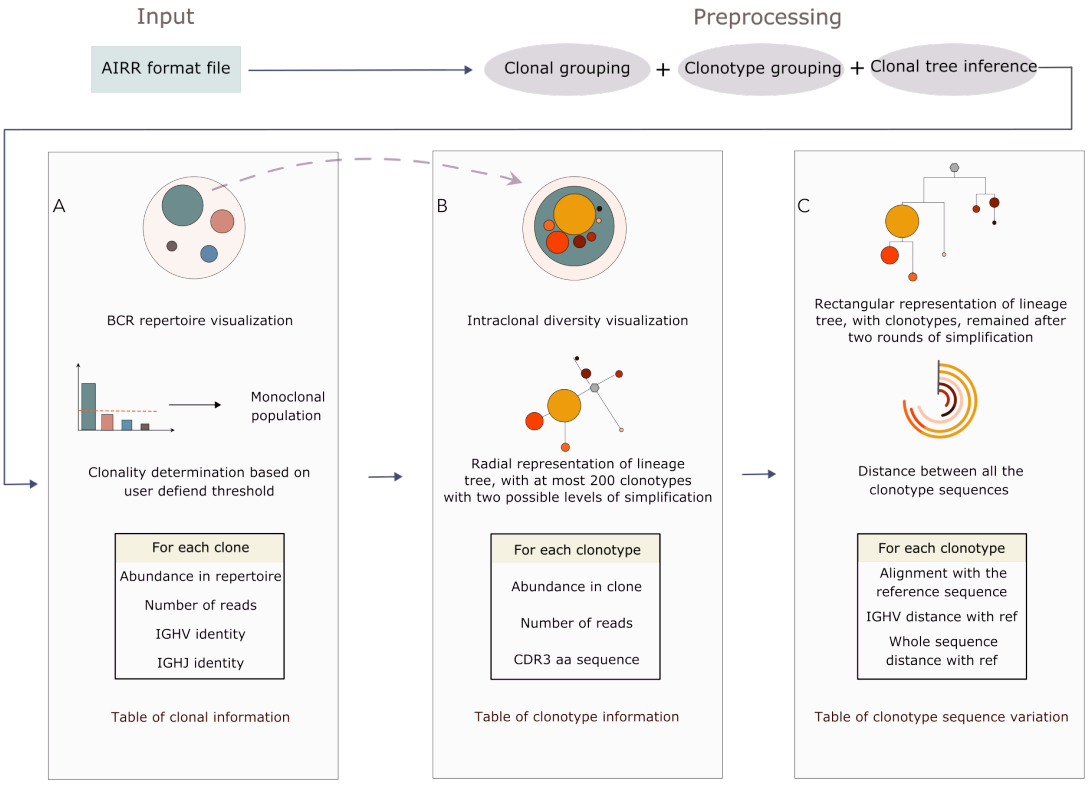

Figure 31: **Overview of ViCloD's workflow.** First, AIRR seq data are devided into groups of clonally related sequences, and clonotypes within each group (clone) are then identified. After that, for the N most abundant clones, lineage trees are inferred. Multiple visualization modules and associated analyses are then available: A) BCR repertoire's clonal analysis, B) intra-clonal diversity analysis and C) and lineage tree study.

10.2.1 *Clonal analysis*

In a repertoire, we presented clones by nested circles. The outer circle represents the entire BCR repertoire, while internal circles correspond to clones (Figure 32). The size of each circle correlates to the clones' abundance. On the same web page, there is a bar plot representing the abundance of each clone (Figure 33). Users can choose between a normal (Figure 33-A) or logarithmic (Figure 33-B) scale; they can also select a threshold of clonotype frequency to analyze the clonality of the repertoire. At the bottom of the page, a table classifies all the information about clones: name, abundance, V gene annotation, J gene annotation, and CDR3 of the most abundant sequence within the clone (Figure 34)

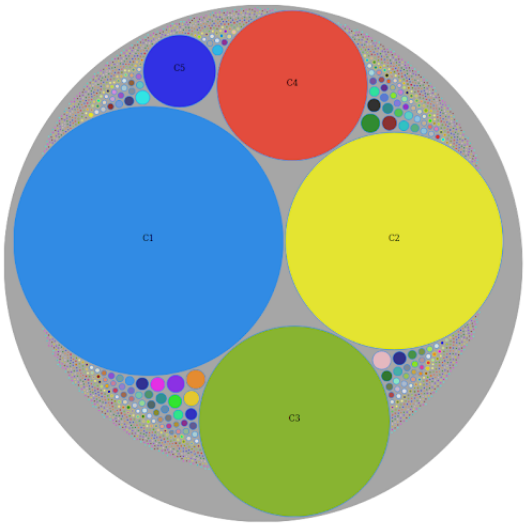

Figure 32: **Repertoire view.** The outer circle (gray) represents the entire repertoire, while inner circles represent clones, their sizes are proportional to their abundance in the repertoires.

Figure 33: **Clone abundance.** Clone abundance is represented by the bars (A) normal scale, and (B) logarithmic scale. In both cases, users can define a threshold for analyzing the clonality of the repertoire, and see which clone bypasses its threshold.

Clone	Abundance(%)	Number of reads	V region	J region	cdr3
C1	42.052	59271	IGHV3-7*03	IGHJ4*02	ARLRGWRDCFDF
C2	27.604	38907	IGHV3-23*01	IGHJ5*02	TKGGIVNYYSGRVWFDP
C3	13.458	18969	IGHV1-69*06	IGHJ6*02	ARISRGDNWNYAQDYDAMDV
C4	12.721	17930	IGHV3-33*01	IGHJ4*02	ARCPRITIFGVVNR**L
C5	1.973	2781	IGHV1-69*01	IGHJ5*02	ALTFYCGGDCYFTENSGWFDP
C6	0.118	167	IGHV3-53*01	IGHJ6*02	ARGPGPGPGDYKVDISPAITIWT
C7	0.116	164	IGHV3-23*01	IGHJ4*02	ARYPDSTTPPDYFDY
C8	0.083	117	IGHV4-38-2*02	IGHJ5*02	ARDLLYYYDSSGYYYSDWFDP
C9	0.06	85	IGHV3-15*01	IGHJ4*02	ATDRRYFFDN

Figure 34: **Clone details** Columns show the clone identifier, abundance in the repertoire (%), number of reads, IGHV gene, IGHJ gene, and CDR3 amino acid sequence. All this information is available for download by clicking on the download button.

10.2.2 *Intra-clonal diversity analysis*

When users click on one of the circles representing clones, it zooms in, and other circles appear (see dotted arrow in Figure 31). These circles correspond to clonotypes, and their size depict their abundance within the clone. Figure 35 shows the clonotypes of clone C3 (the green circle in Figure 32). For a given set of clonotypes, we also show the B cell lineage tree generated by ClonalTree that represents the evolutionary relationships among such clonotypes and the ancestral sequence (Figure 36). Several options of tree representation are available. For instance, nodes can be colored by clonotype (Figure 36-A) or by functionality (Figure 36-B). The green and red nodes represent productive and unproductive rearrangements respectively; unproductive sequences have stop codon(s). It is also possible to display the length of branches representing distances between sequences (Figure 36-B). We chose different geometric forms to represent the hypothetical naive sequence (a triangle) and the most abundant clonotype (a square). At last, we displayed additional information about clonotypes such as identifiers, abundances, CDR3 sequences (Figure 37).

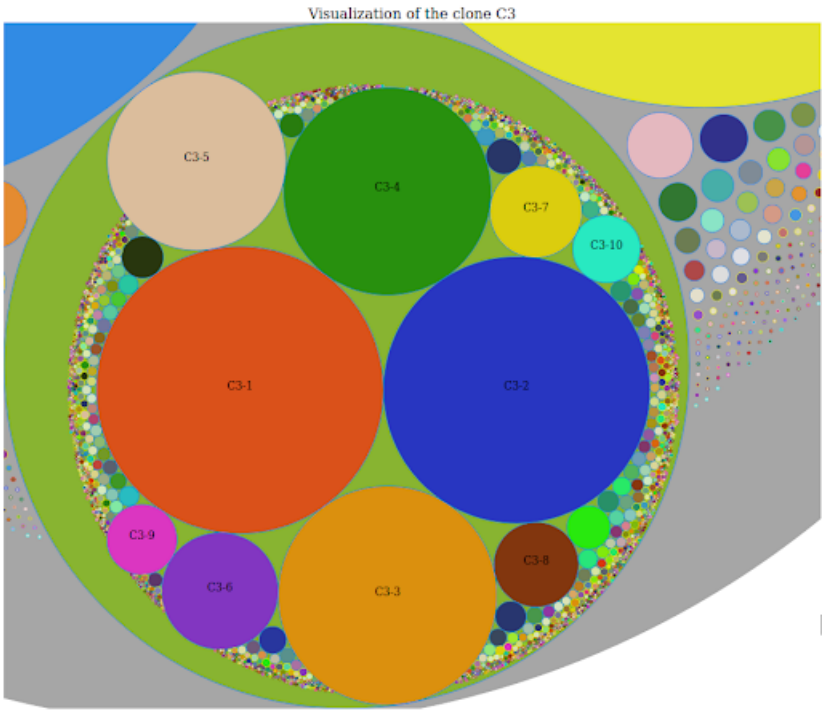

Figure 35: **Clone view**. Each circle represents a clonotype of a selected clone (the light green circle). Circle sizes represent clonotype abundance within the clone.

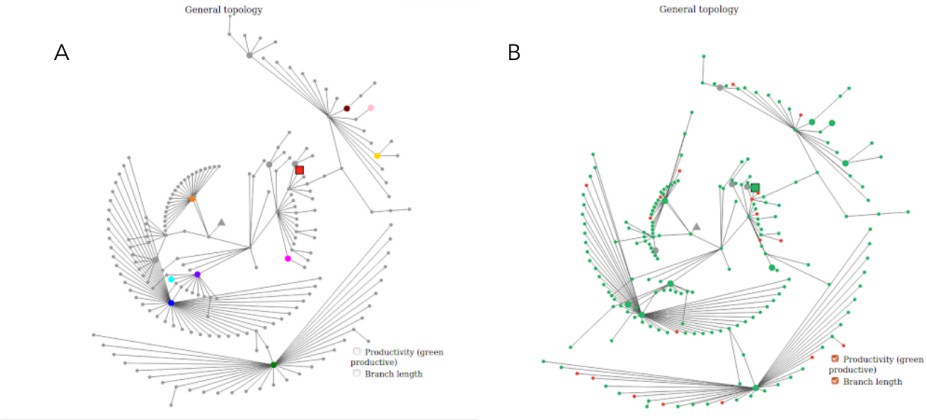

Figure 36: **B-cell lineage trees.** (A) the most abundant clonotypes are colored. (B) nodes are colored according to functionality of their sequence. Green nodes represent productive and red nodes represent unproductive sequences. In both trees, the triangle represents the hypothetical naive sequence, and a square represents the largest clonotype.

Clonotype	Abundance in repertoire (%)	Abundance in clone (%)	cdr3	Productivity
C1-1	26.131	62.14	ARLRGWRDCFDF	yes
C1-2	14.796	35.184	ARLRGWRDCFDY	yes
C1-3	0.134	0.319	AGLRGWRDCFDF	yes
C1-4	0.079	0.187	ARLRGWRDCLDF	yes
C1-5	0.057	0.135	AGLRGWRDCFDY	yes
C1-6	0.051	0.121	ARLRGWRDCFDL	yes
C1-7	0.047	0.111	ARPRGWRDCFDF	yes
C1-8	0.044	0.105	ARLRGWWDCFDF	yes
C1-9	0.043	0.103	VRLRGWRDCFDF	yes

Figure 37: **Clonotype details.** Columns show the clonotype identifiers, abundances in the repertoire (%), abundance in the clone (%), CDR3 amino acid sequence, and functionality of the clonotype representative sequence.

10.2.3 *Pruning trees for a better interpretation*

The lineage tree inferred by the ClonalTree algorithm represents at most the 200 of the most abundant clonotypes of a clone. For large clones containing many clonotypes, interpreting the complete lineage tree can be demanding. After studying multiple tree topologies obtained from different types of repertoires, we decided to simplify the trees while conserving nodes with high abundance and their evolutive path to the root. We developed two strategies for pruning the tree adapted to this context.

Pruning [101] helps achieve a simple, comprehensible, yet approximative description of a tree. In some situations, this simplified version may be more valuable than an entirely accurate description that involves many details. The first pruning strategy, shown in Figure 38-top, eliminates the nodes that do not have any descendant or with an abundance lower than a defined threshold. Figure 39-B shows the simplified tree after applying this first strategy to the tree in Figure 39-A. The first approach is more useful for relatively small trees. Trees with a large number of clonotypes require a second stage of pruning that eliminates nodes with low abundance if they present high similarity with more abundant clonotypes (Figure 38-bottom). Figure 39-C shows the simplified tree after applying this second strategy to the tree in Figure 39-B. For both strategies, we eliminate just leaf nodes, and keep the N most abundant clonotypes in the simplified tree (by default N=10). By eliminating only leaves' nodes, we try to preserve the tree's evolutionary paths, which can be destroyed when an internal node is removed. We keep the N largest clonotypes because they represent 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 we achieve 30 nodes.

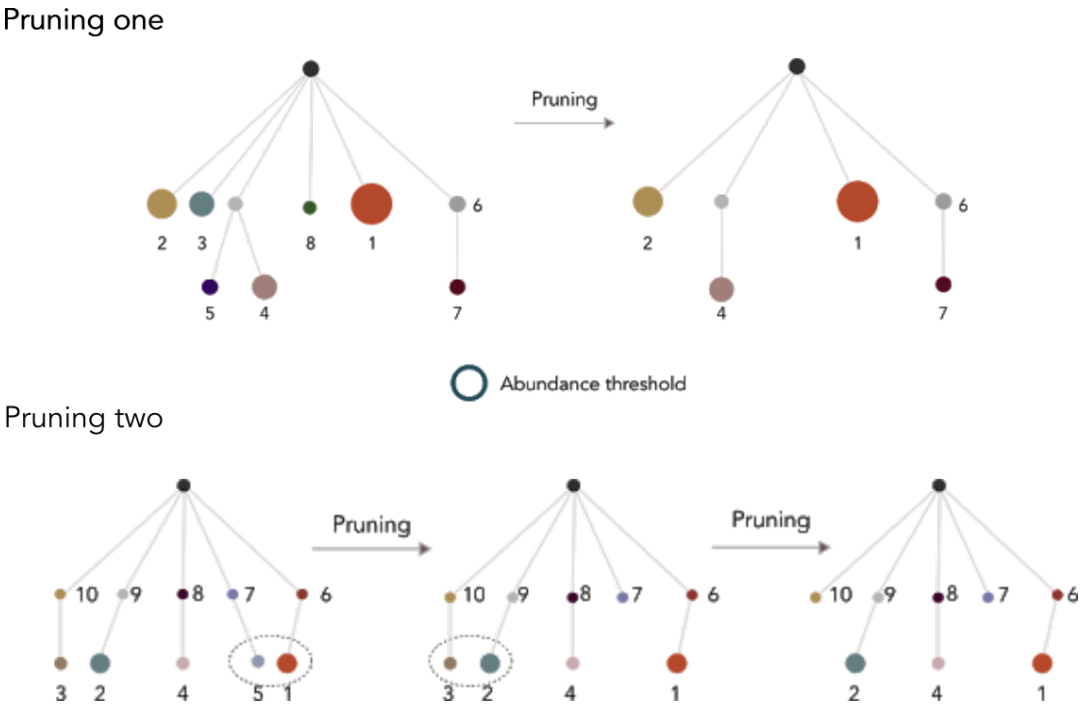

Figure 38: **Strategies for pruning trees.** The first strategy is to eliminate less abundant nodes without descendants. For instance nodes 3, 5, and 8 were eliminated (top). The second strategy eliminates nodes which are highly similar to higher abundant nodes. At first node 5 and then node 3 are removed.

Figure 39: **Simplifying lineage trees** (A) A tree constructed with the 200 most abundant clonotypes. (B) The first simplification (C) The second simplification.

10.2.4 *Intra-clonal diversity analysis*

Users can also explore the simplified lineage tree to examine clonotype diversities. ViCloD provides three types of representations: an interactive tree, a circular bar chart and multiple sequence alignments, see Figures 40, 41 and 42.

10.2.4.1 *Lineage tree*

We kept the most pruned tree with at least the ten most abundant clonotypes in order to better examine the lineage tree (Figure 40-top). Colored circles represent observed clonotypes, while white circles represent unobserved nodes (see Section 9.2.4). The branch length represents the number of somatic hypermutations among connected clonotypes. We identify clonotypes by a number, sorted by their abundance, which means that clonotype **1** is the most abundant in the clone. Nodes that represent the largest clonotypes have a bold border in the tree. When passing the mouse over nodes, clonotype details (identifier, abundance and functionality) appear. By clicking on "display abundance", it is possible to display the abundance of each clonotype by increasing the size of the correspondent node; see an example in Figure 40-bottom, where the most abundant clonotypes were highlighted. Nodes are then labeled with their abundance (%) rather than a sequential number.

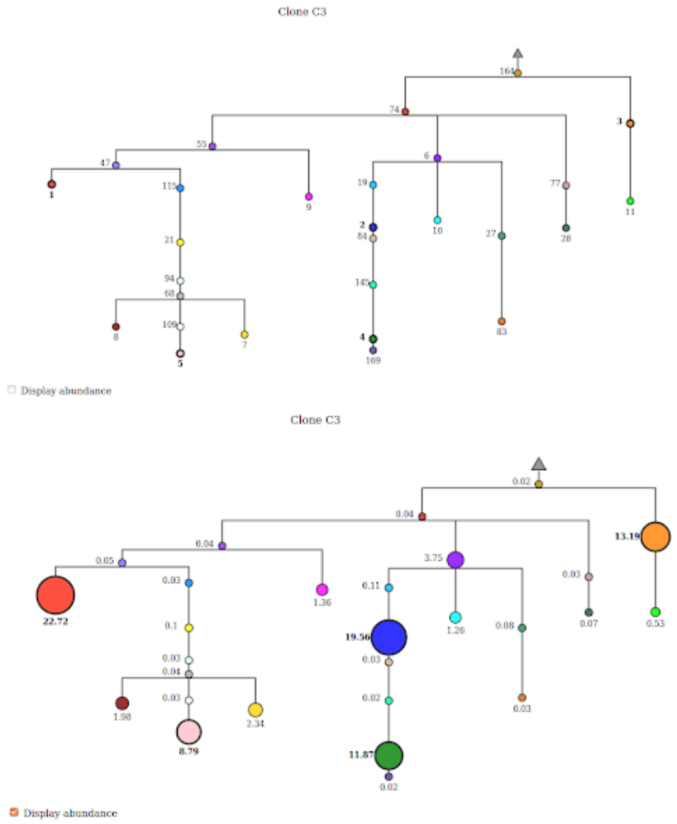

Figure 40: **Lineage tree.** The triangle represents the hypothetical naive sequence, nodes represent clonotypes, and the branch length represents their evolutionary distance. (top) clonotypes are identified by a sequential number, and all have the same size. (bottom) clonotypes are identified by their abundance in the clone (%), and their size is proportional to the clonotype abundance.

10.2.4.2 *Circular bar chart*

The circular bar chart represents the distances (number of somatic hypermutations) between the ancestral and selected clonotypes' sequences (Figure 41). By default, ViCloD displays the five most abundant clonotypes, but any clonotype can be included and/or removed from the plot with a maximum of eight clonotype. Each colored section is related to a clonotype and represents the number of mutations observed between this clonotype and its parent in the tree. To highlight the tree path from a clonotype to the root, users should hover the mouse over their desired clonotype 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 clonotypes.

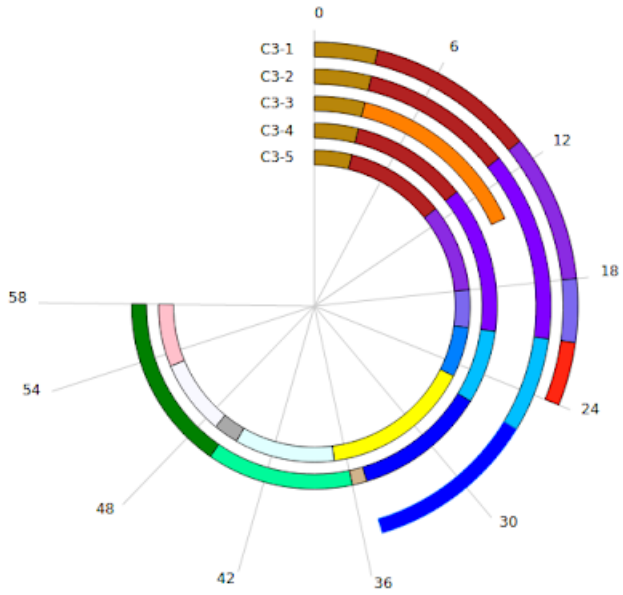

Figure 41: **Circular bar plot.** To display the entire path from a leaf to the root, users should hover the mouse over the clonotype identifiers, for instance, C3-1.

10.2.4.3 *Multiple sequence alignment*

To illustrate the conservation/mutations between representative sequences of each clonotype and ancestral sequences, we build a multiple sequence alignment with MUSCLE program [102] from the biopython "Bio.Align.Applications" package. As shown in Figure 42, for each sequence of this alignment, we have displayed multiple pieces of information separated by columns in the 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 the altered nucleotides of clonotype sequences compared to the hypothetical ancestor are shown in the alignment, whereas a dot represents conserved ones. CDR and framework sections are highlighted with different colors, and the IGHD gene is underlined in each sequence. Users can sort out sequences in the table based on each column. By clicking on the download button, the table will be downloaded into a text file.

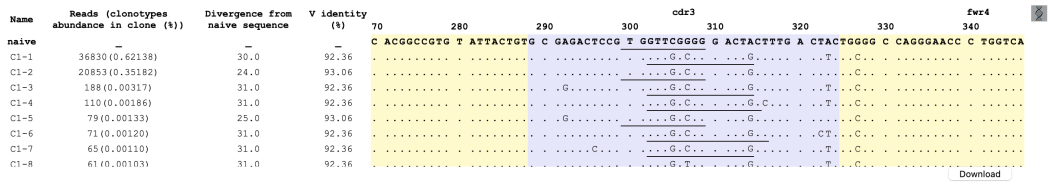

Figure 42: **Intra-clonal multiple sequence alignment in ViCloD.**

10.2.5 *Availability*

ViCloD is a user-friendly web server for visualizing BCR repertoire data, and studying intra-clonal diversities. The web server is available at `http://genome.lcqb.upmc.fr/ViCoD/example.php` For the time-being, a login and a password are required, but ViCloD will be available without registration after the publication.

10.2.6 *Implementation*

The web server uses the PHP language, Java Script (`https://d3js.org/`), bash and Python. It also uses MySql database to control the user pending jobs. Users submit their input file and should provide an email address in order to receive the/their results. For each user we check if the input file contains all required information. Then, users' files (inputs and outputs) are stored in a specific directory; these files are available for download during a certain time. Data analysis is performed by a Bash script that launches the various stages of the pipeline. The pipeline is made of several Python scripts that are executed in an Anaconda environment [103]. When the users' analysis is finished, they will receive an email with the link for their results; we used Exim Internet Mailer [104] as message transfer.

10.2.7 *Downloads*

All the plots and figures are available for download in png format, while tables are available in CSV format. All analyses remain accessible on the server for 7 days after submitting the repertoire's data.

10.3 USE CASE

We demonstrate here how ViCloD could be used to process and analyse BCR high throughput sequencing data.

In order to understand how ViCIoD affects processing and analysing high throughput sequencing data, we analyzed a monoclonal repertoire (obtained from a patient with a circulating population of leukemic B lymphocytes) sequenced during the routine diagnostic analysis at Pitié-Salpêtrière hospital. First, IMGT/HighV-QUEST preprocessed 269206 IGH sequences to generate a series of data annotations, including the assignment of V(D)J-genes and CDR3 delimitations. Then, IMGT's data in AIRR format was uploaded to the ViCloD web server. After being processed by the pipeline, a total of 2646 clones were detected, Figure 43-top left. The major clone, identified as C1 (blue circle), contains 89.5% of the repertoire sequences. Its high abundance can also be seen in the bar plot of Figure 43-top right, which shows the abundance of nine largest clones. C1 has a rearrangement characterised by IGHV4-38-2*01 and IGHJ 5*01 genes, and the CDR3 of its largest clonotype is ARGSADDRNNWFDS (Figure 43-bottom).

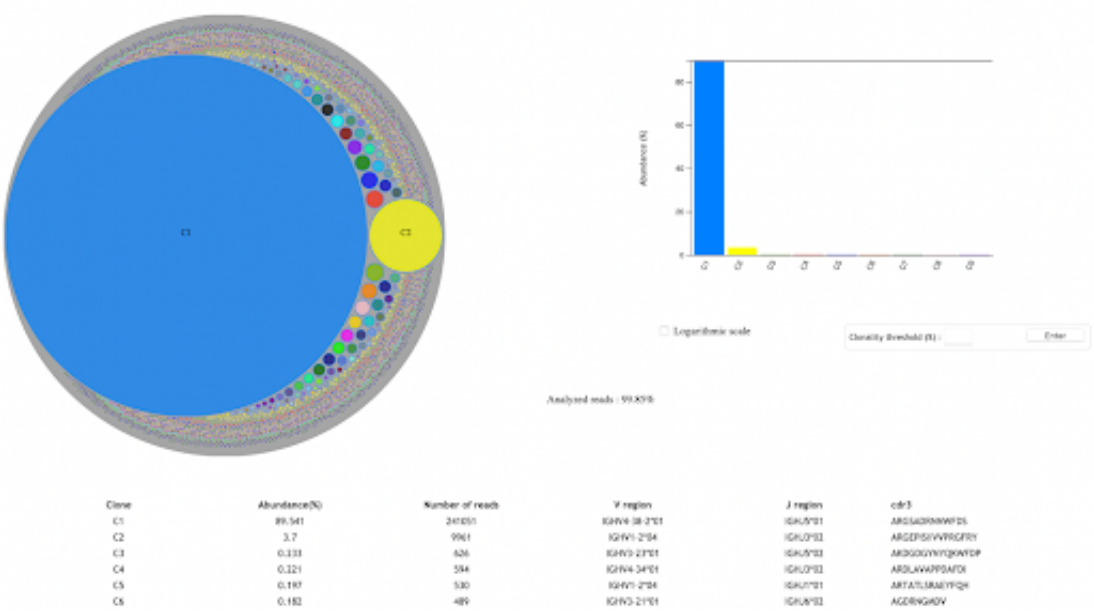

Figure 43: **Example of repertoire visualization using ViCloD.**

Further scrutinization shows that C1 clone is composed by 2608 different clonotypes represented in figure 44-top left; the repartition of the sequences within different clonotypes is not homogeneous, and there is a dominant clonotype, C1-1, with 90% of the C1's sequences, corresponding to 81% of the total sequences of the repertoire. The general tree topology of C1' clonotypes is shown in Figure 44-top right, where the reconstructed naive sequence is represented by a triangle and the most abundant clonotypes by different colors. Note that the largest clonotype C1-1 is highlighted by a square.

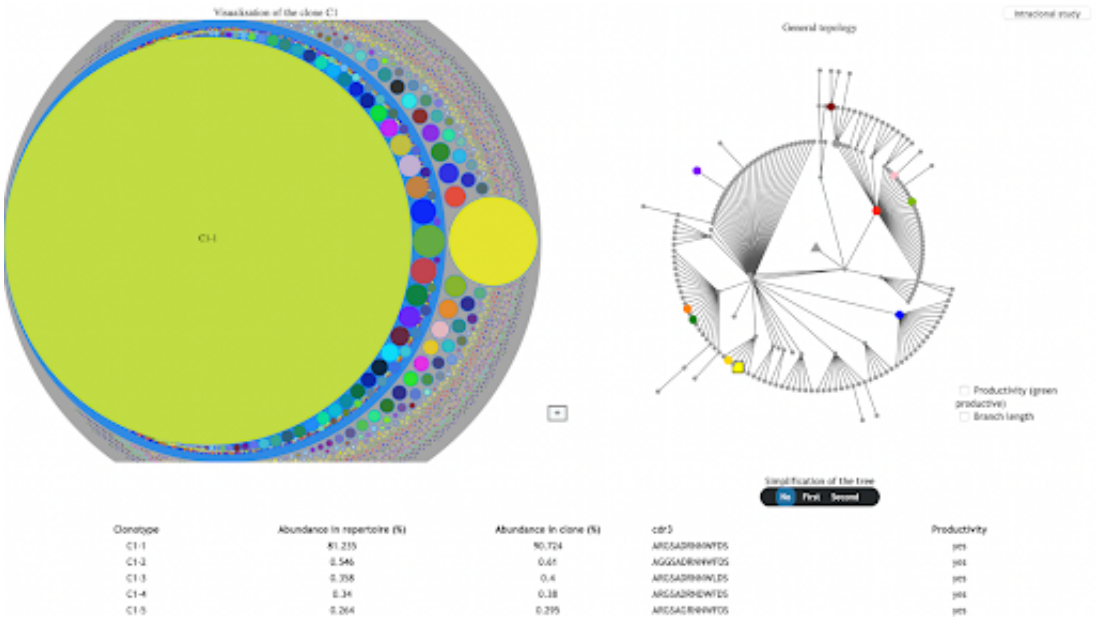

Figure 44: **Example of intra-clonal diversity visualization in ViCloD.**

To see more details of the lineage tree, users can click on the button "Intra-clonal study", which leads to the page shown in Figure 45. The circular elbow tree represents the evolutionary history of at least the 10 most abundant clonotypes. This tree is the result of two consecutive simplifications on the complete lineage tree (Section 10.2.3). The hypothetical naive sequence is represented by a triangle. The dashed circle shows an unobserved node added by ClonalTree for better outlining the evolutionary relations between clonotypes 10 and 33. The most abundant clonotype, identified by the number 1, has three observed ascendentes: 50, 59 et 76; it has 15 nucleotide mutations, when compared to the clonotype 76, which is the closest observed sequence to the hypothetical naive. The clonotype abundances on the tree can also be displayed by selecting the option "display abundance", which changes the tree to the one shown in Figure 45-bottom left. Note that the numbers beside each clonotype in the "display abundance" mode are the abundance percentage. Among the 10-top most abundant clonotypes, 3 (C1-3) is the closest to the hypothetical naive, having the highest IGHV gene identity (Figure 45-bottom).

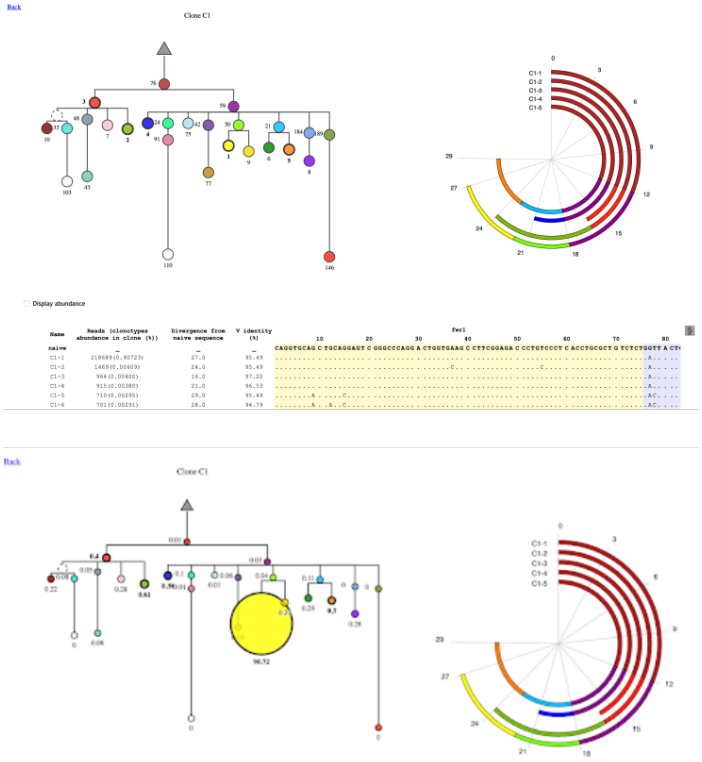

Figure 45: **Example of lineage tree visualization in ViCloD.** without (top) and with (bottom) "display abundance" option.

10.4 CONCLUSION

We have produced a new RepSeq bioinformatics tool, ViCloD, which can be used to process and analyse data obtained from IMGT/HighV-Quest, the international standard web software for parsing adaptive immune receptor sequence data. It is a user-friendly and versatile pipeline, particularly devoted to the analysis of B cell intra-clonal diversity and its visualization. Additional features will be implemented in future such as carrying out intra-clonal analysis at the amino acid level, since the clonal selection is based on the BCR protein sequence. Further progress also includes providing comparison functionality, which will allow users to analyze repertoires from the same patient at various times or to compare repertoires of different patients. These comparative features will be performed at clonal and intra-clonal levels.

Part IV

CONCLUSION AND PERSPECTIVES

11

CONCLUSION AND PERSPECTIVES

11.0.1 *Conclusion*

Next-generation sequencing has enabled researchers to conduct in-depth analyses of the immunological activity and the immune response. However, a significant concern in immune repertoire studies is the computational cost of analyzing millions of sequences with inherent complexity, variability, and mutational capacity, imposing computational challenges and necessitating the development of efficient methods. Such challenge is even more evident in the clinical context that does not necessarily have access to professionals with computing skills or robust computational resources. Thus, the main goal of this thesis was to develop a set of dedicated and integrated tools to be used in the clinical environment, for medical diagnostic and patient care, and in the research environment to perform large-scale and in-depth repertoire analysis.

The first four chapters established the biological foundation for addressing the set of questions posed in chapter 6:

1. Among multiple clone definitions and associated algorithms, how to choose the most appropriate for carrying out a meaningful clonal analysis?
2. Among existing BCR clonal grouping tools using the appropriate clone definition for our research question, is there any that can be used in the clinical context with the aim of intraclonal analysis? If not, how should we design it?
3. What is the most efficient and accurate way to reconstruct the evolution of a B-cell lineage or clone?
4. How, in practice, can we integrate BCR repertoire and intraclonal analyse tools into the clinical context?

In order to respond to them, we started by developing Agreeable, a clustering tool dedicated to clonal grouping, with practical use in the clinical context. The

details of Agreeable's design, execution, and validation were presented in Chapters 7 and 8. In the 8 we also demonstrated the impact of the algorithmic choices on the results of BCR clonal grouping. This comparative study is still ongoing.

In chapter 9, we have presented a new phylogenetic reconstruction method, Clonaltree, that can provide the accuracy and efficiency required to be used for analyzing clinical data. In chapter 10, we have described the process of designed and implemented a user-friendly interactive BCR repertoire visualization pipeline,called ViCloD. We believe that this can make the implementation of the interactive, interdisciplinary communication model possible. This model is described in chapter 4 which can facilitate the process of integrating BCR intraclonal analysis into medical practices.

Each of the demonstrated tools (Agreeable, GCtree, and ViCloD) are objected to be presented in a scientific publication, and their manuscripts are under preparation.

11.0.2 *Direction for future work*

In light of the limitations identified and this study's findings, we plan to analyze multiple repertoires collected from patients with different pathologies to explore the potential relevance of intraclonal diversity to the origin of each pathology and, consequently, suggest a better-adapted treatment to each individual's case. Given that our collaborators are specialized in Chronic Lymphocytic Leukemia (CLL), we would like to start by analyzing CLL datasets. Based on the previous experiences with other forms of B-cell malignancies, the study of intraclonal diversification can improve our understanding of the role of antigen in CLL pathogenesis and potentially to cure them more effectively [105, 106]. BCR intraclonal analysis can assist in defining the original cell of the pathology in a more precise manner [107]. It is prophesied that this analysis can contribute to characterizing the role of intraclonal diversity in prognosis, and response to therapy[108]. We also plan on adding new features to ViCloD which enables the users to compare multiple repertoires at clonal and intraclonal levels.

These new features can compare multiple BCR repertoires of :

- the same person at different time points, which can be particularly useful for studying the effects of a treatment or a vaccine.

- different people for the study of stereotyped BCR immunoglobulins. Several studies have shown that there are subsets of CLL patients expressing highly similar, stereotyped BCR Ig, despite the fact that the chances of two independent B cell clones having identical immunoglobulins is negligible [22].

This visualization pipeline, combined with feature selection methods in machine learning, could permit clinicians/immunologists to discover discriminating features that characterize a repertoire and associate it with an immunological status. (e.g., healthy, infected, vaccinated, etc.[21]). We have also started to examine the possibility of developing a clustering approach that can accurately identify sequences belonging to the same B cell lineage without VDJ annotations since the latter information does not substantially improve the performance of clonal grouping; however, it can slow down the whole process of repertoire analysis. Moreover, clustering into clonally related groups can help annotate more accurately the sequences with a high number of mutations and often have multiple possible genes associated with them.

Part V

APPENDIX

A

AIRR FILE'S REQUIRED FIELDS FOR VICLOD PIPELINE

Table 10: **Required fields of AIRR file in the input of ViCloD.**

Name	Type	Definition
sequence_id	string	Unique query sequence identifier for the Rearrangement.
productive	boolean	True if the V(D)J sequence is predicted to be productive.
v_call	string	V gene with allele.
j_call	string	J gene with allele.
sequence_alignment	string	Aligned portion of query sequence, including any indel corrections or numbering spacers, such as IMGT-gaps. Typically, this will include only the V(D)J region.
germline_alignment	string	Assembled, aligned, full-length inferred germline sequence spanning the same region as the sequence_alignment field (typically the V(D)J region) and including the same set of corrections and spacers (if any).
junction	string	Junction region nucleotide sequence, where the junction is defined as the CDR3 plus the two flanking conserved codons.
np1	string	Nucleotide sequence of the combined N/P region between the V gene and first D gene alignment or between the V gene and J gene alignments.
np2	string	Nucleotide sequence of the combined N/P region between either the first D gene and J gene alignments or the first D gene and second D gene alignments.
cdr1	string	Nucleotide sequence of the aligned CDR1 region.
cdr2	string	Nucleotide sequence of the aligned CDR2 region.
cdr3	string	Nucleotide sequence of the aligned CDR3 region.
fwr1	string	Nucleotide sequence of the aligned FWR1 region.
fwr2	string	Nucleotide sequence of the aligned FWR2 region.
fwr3	string	Nucleotide sequence of the aligned FWR3 region.
fwr4	string	Nucleotide sequence of the aligned FWR4 region.
v_identity	number	Fractional identity for the V gene alignment.
v_germline_alignment	string	Aligned V gene germline sequence spanning the same region as the v_sequence_alignment field and including the same set of corrections and spacers (if any).
d_germline_alignment	string	Aligned D gene germline sequence spanning the same region as the d_sequence_alignment field and including the same set of corrections and spacers (if any).
j_germline_alignment	string	Aligned J gene germline sequence spanning the same region as the j_sequence_alignment field and including the same set of corrections and spacers (if any).

B

COMPARISON OF BCR CLONAL GROUPING TOOLS' PERFORMANCE ON SIMULATED REPERTOIRES

Table 11: Monoclonal repertoire, generated with $\lambda_o = 0.16$, number of sequences is equal to 958 and number of expected clusters is 34.

	# analyzed seq	# clones	Pairwise			Closeness		
Tool name			Precision	Recall	F-score	Precision	Recall	F-score
Brilia	922	68	1	0.26	0.42	1	0.05	0.09
Sonar	958	52	1	0.77	0.87	1	0.07	0.14
Partis	958	43	1	0.92	0.96	1	0.21	0.35
Scope	958	35	1	0.99	1	1	0.58	0.73
Agreeable	958	34	1	1	1	1	1	1

Table 12: Oligoclonal repertoire, generated with $\lambda_o = 0.16$, number of sequences is equal to 1014 and number of expected clusters is 43.

	# analyzed seq	# clones	Pairwise			Closeness		
Tool name			Precision	Recall	F-score	Precision	Recall	F-score
Brilia	658	69	1	0.47	0.64	1	0.26	0.42
Sonar	1014	55	1	0.94	0.97	1	0.56	0.72
Partis	1014	52	1	0.96	0.98	1	0.71	0.83
Scope	1014	46	1	0.97	0.99	1	0.9	0.95
Agreeable	1014	43	1	1	1	1	1	1

Table 13: Polyclonal repertoire, generated with $\lambda_o = 0.16$, number of sequences is equal to 968 and number of expected clusters is 44.

	# analyzed seq	# clones	Pairwise			Closeness		
Tool name			Precision	Recall	F-score	Precision	Recall	F-score
Brilia	947	78	1	0.63	0.77	1	0.39	0.56
Sonar	968	55	1	0.93	0.97	1	0.64	0.78
Partis	968	53	1	0.97	0.99	1	0.72	0.84
Scope	968	46	1	0.94	0.97	1	0.88	0.94
Agreeable	968	44	1	1	1	1	1	1

Table 14: Monoclonal repertoire, generated with $\lambda_o = 0.26$, number of sequences is equal to 659 and number of expected clusters is 33.

	# analyzed seq	# clones	Pairwise			Closeness		
Tool name			Precision	Recall	F-score	Precision	Recall	F-score
Brilia	618	72	1	0.08	0.15	1	0.04	0.08
Sonar	659	72	1	18	0.31	1	0.05	0.09
Partis	659	42	1	0.85	0.92	1	0.34	0.51
Scope	659	36	1	0.9	0.95	1	0.6	0.75
Agreeable	659	33	1	1	1	1	1	1

Table 15: Oligoclonal repertoire, generated with $\lambda_o = 0.26$, number of sequences is equal to 958 and number of expected clusters is 43.

	# analyzed seq	# clones	Pairwise			Closeness		
Tool name			Precision	Recall	F-score	Precision	Recall	F-score
Brilia	745	90	1	0.29	0.46	1	0.15	0.26
Sonar	958	87	1	0.67	0.81	1	0.22	0.37
Partis	958	52	1	0.88	0.94	1	0.69	0.81
Scope	958	50	1	0.94	0.97	1	0.65	0.79
Agreeable	958	43	1	1	1	1	1	1

Table 16: Polyclonal repertoire, generated with $\lambda_o = 0.26$, number of sequences is equal to 964 and number of expected clusters is 44.

	# analyzed seq	# clones	Pairwise			Closeness		
Tool name			Precision	Recall	F-score	Precision	Recall	F-score
Brilia	876	109	1	0.51	0.68	1	0.22	0.36
Sonar	964	79	1	0.82	0.9	1	0.37	0.54
Partis	964	53	1	0.94	0.97	1	0.74	0.85
Scope	964	49	1	0.95	0.98	1	0.78	0.88
Agreeable	964	45	1	0.99	1	1	0.95	0.97

Table 17: Monoclonal repertoire, generated with $\lambda_o = 0.36$, number of sequences is equal to 924 and number of expected clusters is 35.

	# analyzed seq	# clones	Pairwise			Closeness		
Tool name			Precision	Recall	F-score	Precision	Recall	F-score
Brilia	897	116	1	0.04	0.09	1	0.02	0.04
Sonar	924	105	1	0.40	0.57	1	0.02	0.04
Partis	924	44	1	0.45	0.62	1	0.31	0.47
Scope	924	36	1	0.66	0.8	1	0.58	0.73
Agreeable	924	35	1	1	1	1	1	1

Table 18: Oligoclonal repertoire, generated with $\lambda_o = 0.36$, number of sequences is equal to 991 and number of expected clusters is 40.

	# analyzed seq	# clones	Pairwise			Closeness		
Tool name			Precision	Recall	F-score	Precision	Recall	F-score
Brilia	724	102	1	0.3	0.47	1	0.14	0.25
Sonar	991	124	1	0.38	0.56	1	0.12	0.21
Partis	991	49	1	0.81	0.9	1	0.62	0.77
Scope	991	57	0.99	0.66	0.8	1	0.43	0.60
Agreeable	991	40	1	1	1	1	1	1

Table 19: Polyclonal repertoire, generated with $\lambda_o = 0.36$, number of sequences is equal to 897 and number of expected clusters is 42.

	# analyzed seq	# clones	Pairwise			Closeness		
Tool name			Precision	Recall	F-score	Precision	Recall	F-score
Brilia	639	101	1	0.35	0.53	1	0.19	0.32
Sonar	897	120	1	0.53	0.69	1	0.21	0.35
Partis	897	51	1	0.94	0.97	1	0.7	0.82
Scope	897	58	1	0.82	0.9	1	0.57	0.73
Agreeable	897	43	1	0.99	1	1	1	1

Table 20: Monoclonal repertoire, generated with $\lambda_o = 0.46$, number of sequences is equal to 952 and number of expected clusters is 35.

	# analyzed seq	# clones	Pairwise			Closeness		
Tool name			Precision	Recall	F-score	Precision	Recall	F-score
Brilia	926	152	1	0.02	0.05	1	0.01	0.03
Sonar	952	157	1	0.03	0.06	1	0.01	0.02
Partis	952	44	1	0.97	0.99	1	0.55	0.71
Scope	952	46	1	0.73	0.83	1	0.16	0.28
Agreeable	952	36	1	0.99	1	1	0.99	1

Table 21: Oligoclonal repertoire, generated with $\lambda_o = 0.46$, number of sequences is equal to 1016 and number of expected clusters is 43.

	# analyzed seq	# clones	Pairwise			Closeness		
Tool name			Precision	Recall	F-score	Precision	Recall	F-score
Brilia	689	96	1	0.44	0.62	1	0.2	0.34
Sonar	1016	160	1	0.40	0.58	1	0.1	0.19
Partis	1016	53	1	0.89	0.94	1	0.6	0.75
Scope	1016	62	1	0.84	0.92	1	0.45	0.62
Agreeable	1016	43	1	1	1	1	1	1

Table 22: Polyclonal repertoire, generated with $\lambda_o = 0.46$, number of sequences is equal to 952 and number of expected clusters is 43.

	# analyzed seq	# clones	Pairwise			Closeness		
Tool name			Precision	Recall	F-score	Precision	Recall	F-score
Brilia	705	123	1	0.26	0.41	1	0.14	0.25
Sonar	952	147	1	0.48	0.65	1	0.17	0.28
Partis	952	52	1	0.92	0.96	1	0.72	0.84
Scope	952	57	1	0.79	0.88	1	0.57	0.73
Agreeable	952	43	1	1	1	1	1	1

BIBLIOGRAPHY

[1] Marie Paule Lefranc. "Immunoglobulin and T cell receptor genes: IMGT® and the birth and rise of immunoinformatics." In: *Frontiers in Immunology* 5.FEB (2014), p. 22. ISSN: 16643224. DOI: 10.3389/fimmu.2014.00022.

[2] Klaus Rajewsky. *Clonal selection and learning in the antibody system*. 1996. DOI: 10.1038/381751a0. URL: https://pubmed.ncbi.nlm.nih.gov/8657279/.

[3] Vitaly V. Ganusov and Rob J. De Boer. "Do most lymphocytes in humans really reside in the gut?" In: *Trends in Immunology* 28.12 (2007), pp. 514–518. ISSN: 14714906. DOI: 10.1016/j.it.2007.08.009. URL: https://pubmed.ncbi.nlm.nih.gov/17964854/.

[4] Harlan Robins. *Immunosequencing: applications of immune repertoire deep sequencing*. 2013. DOI: 10.1016/j.coi.2013.09.017. URL: https://pubmed.ncbi.nlm.nih.gov/24140071/.

[5] Elizabeth M. Cameron et al. "Potential of a unique antibody gene signature to predict conversion to clinically definite multiple sclerosis." In: *Journal of Neuroimmunology* 213.1-2 (2009), pp. 123–130. ISSN: 01655728. DOI: 10.1016/j.jneuroim.2009.05.014. URL: https://pubmed.ncbi.nlm.nih.gov/19631394/.

[6] 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." In: *Journal of Autoimmunity* 35.4 (2010), pp. 325–335. ISSN: 08968411. DOI: 10.1016/j.jaut.2010.07.004. URL: https://pubmed.ncbi.nlm.nih.gov/20727711/.

[7] H. Christian Von Büdingen et al. "B cell exchange across the blood-brain barrier in multiple sclerosis." In: *Journal of Clinical Investigation* 122.12 (2012), pp. 4533–4543. ISSN: 00219738. DOI: 10.1172/JCI63842. URL: https://pubmed.ncbi.nlm.nih.gov/23160197/.

[8] 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." In: *Molecular and Cellular Proteomics* 12.12 (2013), pp. 3924–3934. ISSN: 15359476. DOI: 10.1074/mcp.M113.030346. URL: /pmc/articles/PMC3861734//pmc/articles/PMC3861734/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC3861734/.

[9] 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. DOI: 10.1177/1756285612474333. URL: https://pubmed.ncbi.nlm.nih.gov/23634189/.

[10] Joel N.H. Stern et al. "B cells populating the multiple sclerosis brain mature in the draining cervical lymph nodes." In: *Science Translational Medicine* 6.248 (2014). ISSN: 19466242. DOI: 10.1126/scitranslmed.3008879. URL: https://pubmed.ncbi.nlm.nih.gov/25100741/.

[11] Arumugam Palanichamy et al. "Immunoglobulin class-switched B cells form an active immune axis between CNS and periphery in multiple sclerosis." In: *Science Translational Medicine* 6.248 (2014). ISSN: 19466242. DOI: 10.1126/scitranslmed.3008930. URL: https://pubmed.ncbi.nlm.nih.gov/25100740/.

[12] 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." In: *Journal of Allergy and Clinical Immunology* 134.3 (2014), pp. 604–612. ISSN: 10976825. DOI: 10.1016/j.jaci.2014.07.010. URL: /pmc/articles/PMC4151999//pmc/articles/PMC4151999/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC4151999/.

[13] 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 transiently expands circulat-

ing Ara h 2-specific B cells with a homologous repertoire in unrelated subjects." In: *Journal of Allergy and Clinical Immunology* 136.1 (2015), 125–134.e12. ISSN: 10976825. DOI: 10.1016/j.jaci.2015.03.026. URL: https://pubmed.ncbi.nlm.nih.gov/25985925/.

[14] Ramona A. Hoh et al. "Single B-cell deconvolution of peanut-specific antibody responses in allergic patients." In: *Journal of Allergy and Clinical Immunology* 137.1 (2016), pp. 157–167. ISSN: 10976825. DOI: 10.1016/j.jaci.2015.05.029. URL: https://pubmed.ncbi.nlm.nih.gov/26152318/.

[15] Jacob Glanville et al. "Naive antibody gene-segment frequencies are heritable and unaltered by chronic lymphocyte ablation." In: *Proceedings of the National Academy of Sciences of the United States of America* 108.50 (2011), pp. 20066–20071. ISSN: 00278424. DOI: 10.1073/pnas.1107498108. URL: https://pubmed.ncbi.nlm.nih.gov/22123975/.

[16] David M. Kurtz et al. "Noninvasive monitoring of diffuse large B-cell lymphoma by immunoglobulin high-throughput sequencing." In: *Blood* 125.24 (2015), pp. 3679–3687. ISSN: 15280020. DOI: 10.1182/blood-2015-03-635169. URL: https://pubmed.ncbi.nlm.nih.gov/25887775/.

[17] 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." In: *Aging Cell* 10.6 (2011), pp. 922–930. ISSN: 14749718. DOI: 10.1111/j.1474-9726.2011.00732.x. URL: https://pubmed.ncbi.nlm.nih.gov/21726404/.

[18] Victoria Martin, Yu Chang Wu, David Kipling, and Deborah Dunn-Walters. "Ageing of the B-cell repertoire." In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 370.1676 (2015). ISSN: 14712970. DOI: 10.1098/rstb.2014.0237. URL: https://pubmed.ncbi.nlm.nih.gov/26194751/.

[19] Ryo Shinnakasu, Takeshi Inoue, Kohei Kometani, Saya Moriyama, Yu Adachi, Manabu Nakayama, Yoshimasa Takahashi, Hidehiro Fukuyama, Takaharu Okada, and Tomohiro Kurosaki. "Regulated selection of germinal-center cells into the memory B cell compartment." eng. In: *Nature Immunology* 17.7 (2016), pp. 861–869. ISSN: 1529-2916. DOI: 10.1038/ni.3460.

[20] Cristina Parola, Daniel Neumeier, and Sai T. Reddy. "Integrating high-throughput screening and sequencing for monoclonal antibody discovery and engineering." eng. In: *Immunology* 153.1 (2018), pp. 31–41. ISSN: 1365-2567. DOI: 10.1111/imm.12838.

[21] Sai T. Reddy et al. "Monoclonal antibodies isolated without screening by analyzing the variable-gene repertoire of plasma cells." eng. In: *Nature Biotechnology* 28.9 (Sept. 2010), pp. 965–969. ISSN: 1546-1696. DOI: 10.1038/nbt.1673.

[22] Andreas Agathangelidis, Richard Rosenquist, Frederic Davi, Paolo Ghia, Chrysoula Belessi, Anastasia Hadzidimitriou, and Kostas Stamatopoulos. "Immunoglobulin Gene Analysis in Chronic Lymphocytic Leukemia." In: *Methods in Molecular Biology*. Vol. 1881. Humana Press Inc., 2019, pp. 51–62. DOI: 10.1007/978-1-4939-8876-1_5. URL: https://pubmed.ncbi.nlm.nih.gov/30350197/.

[23] Laura López-Santibáñez-Jácome, S. Eréndira Avendaño-Vázquez, and Carlos Fabián Flores-Jasso. *The pipeline repertoire for Ig-Seq analysis*. 2019. DOI: 10.3389/fimmu.2019.00899.

[24] 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)." en. In: *Bioinformatics* 33.24 (Dec. 2017). Ed. by Janet Kelso, pp. 3938–3946. ISSN: 1367-4803, 1460-2059. DOI: 10.1093/bioinformatics/btx533. (Visited on 10/08/2018).

[25] Gur Yaari and Steven H. Kleinstein. "Practical guidelines for B-cell receptor repertoire sequencing analysis." In: *Genome Medicine* 7.1 (2015), p. 121. ISSN: 1756-994X. DOI: 10.1186/s13073-015-0243-2. arXiv: arXiv:1408.1149. URL: http://genomemedicine.com/content/7/1/121.

[26] William D. Lees. *Tools for adaptive immune receptor repertoire sequencing*. 2020. DOI: 10.1016/j.coisb.2020.10.003.

[27] Shunsuke Teraguchi et al. *Methods for sequence and structural analysis of B and T cell receptor repertoires*. 2020. DOI: 10.1016/j.csbj.2020.07.008.

[28] Susumu Tonegawa. "Somatic generation of antibody diversity." In: *Nature* 302.5909 (1983), pp. 575–581.

[29] Marie Paule Lefranc. "IMGT unique numbering for the variable (V), constant (C), and groove (G) domains of IG, TR, MH, IgSF, and MhSF." In: *Cold Spring Harbor Protocols* 6.6 (2011), pp. 633–642. ISSN: 15596095. DOI: 10.1101/pdb.ip85.

[30] Marie Paule Lefranc, Christelle Pommié, Manuel Ruiz, Véronique Giudicelli, Elodie Foulquier, Lisa Truong, Valérie Thouvenin-Contet, and Gérard Lefranc. "IMGT unique numbering for immunoglobulin and T cell receptor variable domains and Ig superfamily V-like domains." In: *Developmental and Comparative Immunology* 27.1 (2003), pp. 55–77. ISSN: 0145305X. DOI: 10.1016/S0145-305X(02)00039-3.

[31] Alexander G. Betz, Cristina Rada, Richard Pannell, César Milstein, and Michael S. Neuberger. "Passenger transgenes reveal intrinsic specificity of the antibody hypermutation mechanism: Clustering, polarity, and specific hot spots." In: *Proceedings of the National Academy of Sciences of the United States of America* 90.6 (1993), pp. 2385–2388. ISSN: 00278424. DOI: 10.1073/pnas.90.6.2385. URL: https://www.pnas.org/content/90/6/2385https://www.pnas.org/content/90/6/2385.abstract.

[32] Gary S. Shapiro, Katja Aviszus, David Ikle, and Lawrence J. Wysocki. "Predicting Regional Mutability in Antibody V Genes Based Solely on Di- and Trinucleotide Sequence Composition." In: *The Journal of Immunology* 163.1 (1999), pp. 259–268. ISSN: 0022-1767. eprint: https://www.jimmunol.org/content/163/1/259.full.pdf. URL: https://www.jimmunol.org/content/163/1/259.

[33] Kimberly D. Klonowski, Laura L. Primiano, and Marc Monestier. "Atypical V(H)-D-J(H) rearrangements in newborn autoimmune MRL mice." In: *Journal of Immunology* 162.3 (1999), pp. 1566–1572. ISSN: 0022-1767. URL: https://pubmed.ncbi.nlm.nih.gov/9973414/.

[34] To-Ha Thai and John F Kearney. "Isoforms of terminal deoxynucleotidyltransferase: developmental aspects and function." In: *Advances in immunology* 86 (2005), 113—136. ISSN: 0065-2776. DOI: 10.1016/s0065-2776(04)86003-6. URL: https://doi.org/10.1016/S0065-2776(04)86003-6.

[35] Erand Smakaj et al. "Benchmarking immunoinformatic tools for the analysis of antibody repertoire sequences." en. In: *Bioinformatics* (Dec. 2019). Ed. by Inanc Birols, btz845. ISSN: 1367-4803, 1460-2059. DOI: 10.1093/bioinformatics/btz845. URL: https://academic.oup.com/bioinformatics/advance-article/doi/10.1093/bioinformatics/btz845/5686386 (visited on 03/10/2020).

[36] Sandra C. A. Nielsen and Scott D. Boyd. "Human adaptive immune receptor repertoire analysis-Past, present, and future." en. In: *Immunological Reviews* 284.1 (July 2018), pp. 9–23. ISSN: 01052896. DOI: 10.1111/imr.12667. URL: http://doi.wiley.com/10.1111/imr.12667 (visited on 09/21/2018).

[37] Uri Hershberg and Eline T Luning Prak. "The analysis of clonal expansions in normal and autoimmune B cell repertoires." In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 370.1676 (2015), p. 20140239.

[38] Gur Yaari and Steven H. Kleinstein. "Practical guidelines for B-cell receptor repertoire sequencing analysis." In: *Genome Medicine* 7.1 (2015), p. 121. ISSN: 1756-994X. DOI: 10.1186/s13073-015-0243-2. arXiv: arXiv:1408.1149. URL: http://genomemedicine.com/content/7/1/121.

[39] Néstor Vázquez Bernat, Martin Corcoran, Uta Hardt, Mateusz Kaduk, Ganesh E. Phad, Marcel Martin, and Gunilla B. Karlsson Hedestam. "High-Quality Library Preparation for NGS-Based Immunoglobulin Germline Gene Inference and Repertoire Expression Analysis." In: *Frontiers in Immunology* 10 (2019), p. 660. ISSN: 1664-3224. DOI: 10.3389/fimmu.2019.00660. URL: https://www.frontiersin.org/article/10.3389/fimmu.2019.00660.

[40] Véronique Giudicelli, Denys Chaume, and Marie-Paule Lefranc. "IMGT/GENE-DB: a comprehensive database for human and mouse immunoglobulin and T cell receptor genes." In: *Nucleic Acids Research* 33.Database issue (2004), pp. D256–D261.

[41] Brandon Dekosky, Takaaki Kojima, Alexa Rodin, Wissam Charab, Gregory Ippolito, Andrew Ellington, and George Georgiou. "In-depth determination and analysis of the human paired heavy- and light-chain an-

tibody repertoire." In: *Nature medicine* 21 (Dec. 2014). DOI: 10.1038/nm.3743.

[42] Nima Nouri and Steven H. Kleinstein. "A spectral clustering-based method for identifying clones from high-throughput B cell repertoire sequencing data." In: *Bioinformatics* 34.13 (2018), pp. i341–i349. ISSN: 14602059. DOI: 10.1093/bioinformatics/bty235.

[43] Marc Duez, Mathieu Giraud, Ryan Herbert, Tatiana Rocher, Mikaël Salson, and Florian Thonier. "Vidjil: A Web Platform for Analysis of High-Throughput Repertoire Sequencing." In: *PLOS ONE* 11.11 (2016), e0166126.

[44] Ofir Lindenbaum, Nima Nouri, Yuval Kluger, and Steven H Kleinstein. "Alignment free identification of clones in B cell receptor repertoires." In: *Nucleic Acids Research* 49.4 (Dec. 2020), e21–e21. ISSN: 0305-1048. DOI: 10.1093/nar/gkaa1160. eprint: https://academic.oup.com/nar/article-pdf/49/4/e21/36398505/gkaa1160.pdf. URL: https://doi.org/10.1093/nar/gkaa1160.

[45] M. O. Hill. "Diversity and Evenness: A Unifying Notation and Its Consequences." In: *Ecology* 54.2 (1973), pp. 427–432. DOI: https://doi.org/10.2307/1934352. eprint: https://esajournals.onlinelibrary.wiley.com/doi/pdf/10.2307/1934352. URL: https://esajournals.onlinelibrary.wiley.com/doi/abs/10.2307/1934352.

[46] H. Morbach, E. M. Eichhorn, J. G. Liese, and H. J. Girschick. "Reference values for B cell subpopulations from infancy to adulthood." In: *Clinical & Experimental Immunology* 162.2 (2010), pp. 271–279. DOI: https://doi.org/10.1111/j.1365-2249.2010.04206.x. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1365-2249.2010.04206.x. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1365-2249.2010.04206.x.

[47] Thierry Mora and Aleksandra M. Walczak. "How many different clonotypes do immune repertoires contain?" In: *Current Opinion in Systems Biology* 18 (2019), pp. 104–110. ISSN: 2452-3100. DOI: https://doi.org/10.1016/j.coisb.2019.10.001. URL: https://www.sciencedirect.com/science/article/pii/S2452310019300289.

[48] Donald R. Forsdyke. "Two signal half-century: From negative selection of self-reactivity to positive selection of near-self-reactivity." In: *Scandinavian Journal of Immunology* 89.4 (2019), e12746. DOI: https://doi.org/10.1111/sji.12746. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/sji.12746. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/sji.12746.

[49] Kenneth B. Hoehn, Gerton Lunter, and Oliver G. Pybus. "A Phylogenetic Codon Substitution Model for Antibody Lineages." eng. In: *Genetics* 206.1 (2017), pp. 417–427. ISSN: 1943-2631. DOI: 10.1534/genetics.116.196303.

[50] P Roben, J P Moore, M Thali, J Sodroski, C F Barbas, and D R Burton. "Recognition properties of a panel of human recombinant Fab fragments to the CD4 binding site of gp120 that show differing abilities to neutralize human immunodeficiency virus type 1." In: *Journal of Virology* 68.8 (1994), pp. 4821–4828. ISSN: 0022-538X. DOI: 10.1128/jvi.68.8.4821-4828.1994. URL: https://pubmed.ncbi.nlm.nih.gov/7518527/.

[51] Marie Paule Lefranc and Gérard Lefranc. *Immunoglobulins or antibodies: IMGT® bridging genes, structures and functions*. 2020. DOI: 10.3390/biomedicines8090319. URL: https://pubmed.ncbi.nlm.nih.gov/32878258/.

[52] Nima Nouri and Steven H Kleinstein. "A spectral clustering-based method for identifying clones from high-throughput B cell repertoire sequencing data." In: *Bioinformatics* 34.13 (2018), pp. i341–i349.

[53] Jason Anthony Vander Heiden et al. "AIRR Community Standardized Representations for Annotated Immune Repertoires." In: *Frontiers in immunology* 9 (2018). ISSN: 16643224. DOI: 10.3389/fimmu.2018.02206.

[54] Jamie K. Scott and Felix Breden. *The adaptive immune receptor repertoire community as a model for FAIR stewardship of big immunology data*. 2020. DOI: 10.1016/j.coisb.2020.10.001.

[55] Jennifer C. Goldsack et al. "Verification, analytical validation, and clinical validation (V3): the foundation of determining fit-for-purpose for Biometric Monitoring Technologies (BioMeTs)." In: *npj Digital Medicine* 3.1 (2020), pp. 1–15. ISSN: 23986352. DOI: 10.1038/s41746-020-0260-4. URL: https://doi.org/10.1038/s41746-020-0260-4.

[56] Vojtech Bystry et al. "ARResT/Interrogate: an interactive immunoprofiler for IG/TR NGS data." In: *Bioinformatics (Oxford, England)* 33 (Oct. 2016). DOI: 10.1093/bioinformatics/btw634.

[57] Yana Safonova, Stefano Bonissone, Eugene Kurpilyansky, Ekaterina Starostina, Alla Lapidus, Jeremy Stinson, Laura Depalatis, Wendy Sandoval, Jennie Lill, and Pavel A. Pevzner. "Ig Repertoire Constructor: A novel algorithm for antibody repertoire construction and immunoproteogenomics analysis." In: *Bioinformatics* 31.12 (2015), pp. i53–i61. ISSN: 14602059. DOI: 10.1093/bioinformatics/btv238.

[58] Chaim A. Schramm, Zizhang Sheng, Zhenhai Zhang, John R. Mascola, Peter D. Kwong, and Lawrence Shapiro. "SONAR: A high-throughput pipeline for inferring antibody ontogenies from longitudinal sequencing of B cell transcripts." In: *Frontiers in Immunology* 7.SEP (2016), pp. 1–10. ISSN: 16643224. DOI: 10.3389/fimmu.2016.00372.

[59] Aaron M. Rosenfeld, Wenzhao Meng, Eline T. Luning Prak, and Uri Hershberg. "ImmuneDB, a Novel Tool for the Analysis, Storage, and Dissemination of Immune Repertoire Sequencing Data." In: *Frontiers in Immunology* 9 (2018), p. 2107. ISSN: 1664-3224. DOI: 10.3389/fimmu.2018.02107. URL: https://www.frontiersin.org/article/10.3389/fimmu.2018.02107.

[60] Bernardo Cortina-Ceballos et al. "Reconstructing and mining the B cell repertoire with ImmunediveRsity." In: *mAbs* 7.3 (2015), pp. 516–524. ISSN: 19420870. DOI: 10.1080/19420862.2015.1026502. URL: /pmc/articles/PMC4622655//pmc/articles/PMC4622655/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC4622655/.

[61] Duncan K Ralph and Frederick A Matsen IV. "Likelihood-based inference of B cell clonal families." In: *PLoS computational biology* 12.10 (2016), e1005086.

[62] Duncan K Ralph and Frederick A Matsen IV. "Consistency of VDJ rearrangement and substitution parameters enables accurate B cell receptor sequence annotation." In: *PLoS computational biology* 12.1 (2016), e1004409.

[63] Dmitriy A Bolotin, Stanislav Poslavsky, Igor Mitrophanov, Mikhail Shugay, Ilgar Z Mamedov, Ekaterina V Putintseva, and Dmitriy M Chudakov.

"MiXCR: software for comprehensive adaptive immunity profiling." In: *Nature methods* 12.5 (2015), pp. 380–381.

[64] 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." In: *Immunome Research* 08.01 (2012). (Visited on 07/23/2018).

[65] Peter J Rousseeuw. "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis." In: *Journal of computational and applied mathematics* 20 (1987), pp. 53–65.

[66] Vladimir I Levenshtein. "Binary codes capable of correcting deletions, insertions, and reversals." In: *Soviet physics doklady*. Vol. 10. 8. 1966, pp. 707–710.

[67] William S DeWitt III, Luka Mesin, Gabriel D Victora, Vladimir N Minin, and Frederick A Matsen IV. "Using genotype abundance to improve phylogenetic inference." In: *Molecular biology and evolution* 35.5 (2018), pp. 1253–1265.

[68] George F. Widhopf, Craig J. Goldberg, Traci L. Toy, Laura Z. Rassenti, William G. Wierda, John C. Byrd, Michael J. Keating, John G. Gribben, Kanti R. Rai, and Thomas J. Kipps. "Nonstochastic pairing of immunoglobulin heavy and light chains expressed by chronic lymphocytic leukemia B cells is predicated on the heavy chain CDR3." In: *Blood* 111.6 (2008), pp. 3137–3144. ISSN: 00064971. DOI: 10.1182/blood-2007-02-073130.

[69] Bradley T. Messmer et al. "Multiple distinct sets of stereotyped antigen receptors indicate a role for antigen in promoting chronic lymphocytic leukemia." In: *Journal of Experimental Medicine* 200.4 (2004), pp. 519–525. ISSN: 00221007. DOI: 10.1084/jem.20040544.

[70] Kostas Stamatopoulos et al. "Over 20% of patients with chronic lymphocytic leukemia carry stereotyped receptors: Pathogenetic implications and clinical correlations." In: *Blood* 109.1 (2007), pp. 259–270. ISSN: 00064971. DOI: 10.1182/blood-2006-03-012948.

[71] N. Darzentas et al. "A different ontogenesis for chronic lymphocytic leukemia cases carrying stereotyped antigen receptors: Molecular and computational evidence." In: *Leukemia* 24.1 (2010), pp. 125–132. ISSN: 14765551. DOI: 10.1038/leu.2009.186.

[72] Joseph L. Gastwirth. "A General Definition of the Lorenz Curve." In: *Econometrica* 39.6 (1971), p. 1037. ISSN: 00129682. DOI: 10.2307/1909675.

[73] Thitithep Sitthiyot and Kanyarat Holasut. "A simple method for measuring inequality." In: *Palgrave Communications* 6.1 (2020), pp. 1–9. ISSN: 20551045. DOI: 10.1057/s41599-020-0484-6. URL: https://www.nature.com/articles/s41599-020-0484-6.

[74] E Beillard, N Pallisgaard, VHJ Van der Velden, W Bi, Rob Dee, Ellen van der Schoot, E Delabesse, Elizabeth Macintyre, E Gottardi, G Saglio, et al. "Evaluation of candidate control genes for diagnosis and residual disease detection in leukemic patients using 'real-time'quantitative reverse-transcriptase polymerase chain reaction (RQ-PCR)–a Europe against cancer program." In: *Leukemia* 17.12 (2003), pp. 2474–2486.

[75] J. Zhang, K. Kobert, T. Flouri, and A. Stamatakis. "PEAR: a fast and accurate Illumina Paired-End reAd mergeR." In: *Bioinformatics* 30.5 (2014), pp. 614–620.

[76] Corrado Gini. "Concentration and dependency ratios." In: *Rivista di politica economica* 87 (1997), pp. 769–792.

[77] Kristen Johnson, Cristina Angelin-Duclos, Sinae Park, and Kathryn L. Calame. " Changes in Histone Acetylation Are Associated with Differences in Accessibility of V H Gene Segments to V-DJ Recombination during B-Cell Ontogeny and Development." In: *Molecular and Cellular Biology* 23.7 (2003), pp. 2438–2450. ISSN: 0270-7306. DOI: 10.1128/mcb.23.7.2438-2450.2003. URL: https://journals.asm.org/journal/mcb.

[78] Laura Z. Rassenti and Thomas J. Kipps. "Lack of allelic exclusion in B cell chronic lymphocytic leukemia." In: *Journal of Experimental Medicine* 185.8 (1997), pp. 1435–1445. ISSN: 00221007. DOI: 10.1084/jem.185.8.1435. URL: http://rupress.org/jem/article-pdf/185/8/1435/1111522/5502.pdf.

[79] Zhiliang Chen, Andrew M. Collins, Yan Wang, and Bruno A. Gata. "Clustering-based identification of clonally-related immunoglobulin gene sequence sets." In: *Immunome Research*. Vol. 6. SUPPL. 1. BioMed Central, 2010, S4. DOI: 10.1186/1745-7580-6-S1-S4. URL: /pmc/articles/PMC2946782/ /pmc/articles/PMC2946782/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC2946782/.

[80] Uri Hershberg and Eline T. Luning Prak. *The analysis of clonal expansions in normal and autoimmune B cell repertoires*. 2015. DOI: 10.1098/rstb.2014.0239. URL: http://dx.doi.org/10.1098/rstb.2014.0239.

[81] Donald W. Lee, Ilja V. Khavrutskii, Anders Wallqvist, Sina Bavari, Christopher L. Cooper, and Sidhartha Chaudhury. "BRILIA: Integrated tool for high-throughput annotation and lineage tree assembly of B-cell repertoires." In: *Frontiers in Immunology* 7.JAN (2017), pp. 1–18. ISSN: 16643224. DOI: 10.3389/fimmu.2016.00681.

[82] Bruno A. Gaëta, Harald R. Malming, Katherine J.L. Jackson, Michael E. Bain, Patrick Wilson, and Andrew M. Collins. "iHMMune-align: Hidden Markov model-based alignment and identification of germline genes in rearranged immunoglobulin gene sequences." In: *Bioinformatics*. Vol. 23. 13. Bioinformatics, 2007, pp. 1580–1587. DOI: 10.1093/bioinformatics/btm147. URL: https://pubmed.ncbi.nlm.nih.gov/17463026/.

[83] Jian Ye, Ning Ma, Thomas L. Madden, and James M. Ostell. "IgBLAST: an immunoglobulin variable domain sequence analysis tool." In: *Nucleic acids research* 41.Web Server issue (2013). ISSN: 13624962. DOI: 10.1093/nar/gkt382. URL: https://pubmed.ncbi.nlm.nih.gov/23671333/.

[84] Robert C. Edgar. "Search and clustering orders of magnitude faster than BLAST." In: *Bioinformatics* 26.19 (2010), pp. 2460–2461. ISSN: 13674803. DOI: 10.1093/bioinformatics/btq461. URL: https://pubmed.ncbi.nlm.nih.gov/20709691/.

[85] Michal Barak, Neta S Zuckerman, Hanna Edelman, Ron Unger, and Ramit Mehr. "IgTree©: creating immunoglobulin variable region gene lineage trees." In: *Journal of immunological methods* 338.1-2 (2008), pp. 67–74.

[86] Joseph Felsenstein. *PHYLIP (phylogeny inference package), version 3.5 c*. Joseph Felsenstein., 1993.

[87] Theodore Edward Harris. "The theory of branching process." In: (1964).

[88] Xingyu Yang, Christopher M Tipton, Matthew C Woodruff, Enlu Zhou, F Eun-Hyung Lee, Inãki Sanz, and Peng Qiu. "GLaMST: Grow Lineages along minimum spanning tree for B cell receptor sequencing data." In: *BMC genomics* 21.9 (2020), pp. 1–11.

[89] Joseph B Kruskal. "On the shortest spanning subtree of a graph and the traveling salesman problem." In: *Proceedings of the American Mathematical society* 7.1 (1956), pp. 48–50.

[90] R. C. Prim. "Shortest Connection Networks And Some Generalizations." In: *Bell System Technical Journal* 36.6 (1957), pp. 1389–1401. DOI: https://doi.org/10.1002/j.1538-7305.1957.tb01515.x. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/j.1538-7305.1957.tb01515.x. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/j.1538-7305.1957.tb01515.x.

[91] Joseph B. Kruskal. "On the shortest spanning subtree of a graph and the traveling salesman problem." In: *Proceedings of the American Mathematical Society* 7.1 (1956), pp. 48–48. ISSN: 0002-9939. DOI: 10.1090/s0002-9939-1956-0078686-7. URL: https://www.ams.org/journal-terms-of-use.

[92] Michael L Fredman. *Fibonacci Heaps and Their Uses in Improved Network Optimization Algorithms*. Tech. rep. 3, pp. 596–615.

[93] Gur Yaari, Jason Vander Heiden, Mohamed Uduman, Daniel Gadala-Maria, Namita Gupta, Joel NH Stern, Kevin O'Connor, David Hafler, Uri Laserson, Francois Vigneault, et al. "Models of somatic hypermutation targeting and substitution based on synonymous mutations from high-throughput immunoglobulin sequencing data." In: *Frontiers in immunology* 4 (2013), p. 358.

[94] Jeroen M.J. Tas et al. "Visualizing antibody affinity maturation in germinal centers." In: *Science* 351.6277 (2016), pp. 1048–1054. ISSN: 10959203. DOI: 10.1126/science.aad3439. URL: http://science.sciencemag.org/.

[95] J Felsenstein. "PHYLIP (phylogenetic inference package) version 3.6. Department of Genetics, University of Washington." In: 63 (1993), pp. 188–192. URL: https://ci.nii.ac.jp/naid/10027701621/ (visited on 03/10/2020).

[96] Alberto Sanfeliu, Alberto Sanfeliu, and King Sun Fu. "A Distance Measure Between Attributed Relational Graphs for Pattern Recognition." In: *IEEE Transactions on Systems, Man and Cybernetics* SMC-13.3 (1983), pp. 353–362. ISSN: 21682909. DOI: 10.1109/TSMC.1983.6313167.

[97] Kristian Davidsen and Frederick A. Matsen. "Benchmarking Tree and Ancestral Sequence Inference for B Cell Receptor Sequences." In: *Frontiers in immunology* 9 (2018), p. 2451. ISSN: 16643224. DOI: 10.3389/fimmu.2018.02451. URL: /pmc/articles/PMC6220437//pmc/articles/PMC6220437/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC6220437/.

[98] M.R. Garey and D.S. Johnson. *Computers and Intractability: A Guide to the Theory of NP-completeness*. Mathematical Sciences Series. W. H. Freeman, 1979. ISBN: 9780716710448. URL: https://books.google.com.br/books?id=fjxGAQAAIAAJ.

[99] M. Neuhaus and H. Bunke. *Bridging the Gap Between Graph Edit Distance and Kernel Machines*. Series in machine perception and artificial intelligence. World Scientific, 2007. ISBN: 9789812708175. URL: https://books.google.lu/books?id=r_1oDQAAQBAJ.

[100] K. Riesen. *Structural Pattern Recognition with Graph Edit Distance: Approximation Algorithms and Applications*. Advances in Computer Vision and Pattern Recognition. Springer International Publishing, 2016. ISBN: 9783319272511. URL: https://books.google.fr/books?id=EeG5jgEACAAJ.

[101] Manish Mehta, Jorma Rissanen, and Rakesh Agrawal. *MDL-based Decision Tree Pruning*. Tech. rep.

[102] Robert C. Edgar. "MUSCLE: Multiple sequence alignment with high accuracy and high throughput." In: *Nucleic Acids Research* 32.5 (2004), pp. 1792–1797. ISSN: 03051048. DOI: 10.1093/nar/gkh340. URL: /pmc/articles/PMC390337//pmc/articles/PMC390337/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC390337/.

[103] *Anaconda | The World's Most Popular Data Science Platform*. URL: https://www.anaconda.com/ (visited on 06/01/2021).

[104] *UIT Cambridge Ltd. - The Exim SMTP mail server*. URL: https://www.uit.co.uk/the-exim-smtp-mail-server (visited on 06/01/2021).

[105] E. Kostareli, L. A. Sutton, A. Hadzidimitriou, N. Darzentas, A. Kouvatsi, A. Tsaftaris, A. Anagnostopoulos, R. Rosenquist, and K. Stamatopoulos. "Intraclonal diversification of immunoglobulin light chains in a subset of chronic lymphocytic leukemia alludes to antigen-driven clonal evolution." In: *Leukemia* 24.7 (2010), pp. 1317–1324. ISSN: 14765551. DOI: 10.1038/leu.2010.90. URL: www.nature.com/leu.

[106] Alicia D. Volkheimer, J. Brice Weinberg, Bethany E. Beasley, John F. Whitesides, Jon P. Gockerman, Joseph O. Moore, Garnett Kelsoe, Barbara K. Goodman, and Marc C. Levesque. "Progressive immunoglobulin gene mutations in chronic lymphocytic leukemia: Evidence for antigen-driven intraclonal diversification." In: *Blood* 109.4 (2007), pp. 1559–1567. ISSN: 00064971. DOI: 10.1182/blood-2006-05-020644. URL: https://pubmed.ncbi.nlm.nih.gov/17082314/.

[107] Lesley Ann Sutton, Efterpi Kostareli, Anastasia Hadzidimitriou, Nikos Darzentas, Athanasios Tsaftaris, Achilles Anagnostopoulos, Richard Rosenquist, and Kostas Stamatopoulos. "Extensive intraclonal diversification in a subgroup of chronic lymphocytic leukemia patients with stereotyped IGHV4-34 receptors: Implications for ongoing interactions with antigen." In: *Blood* 114.20 (2009), pp. 4460–4468. ISSN: 00064971. DOI: 10.1182/blood-2009-05-221309. URL: https://pubmed.ncbi.nlm.nih.gov/19713457/.

[108] Paul J. Hengeveld, Mark David Levin, P. Martijn Kolijn, and Anton W. Langerak. *Reading the B-cell receptor immunome in chronic lymphocytic leukemia: revelations and applications*. 2021. DOI: 10.1016/j.exphem.2020.09.194. URL: https://pubmed.ncbi.nlm.nih.gov/32976948/.