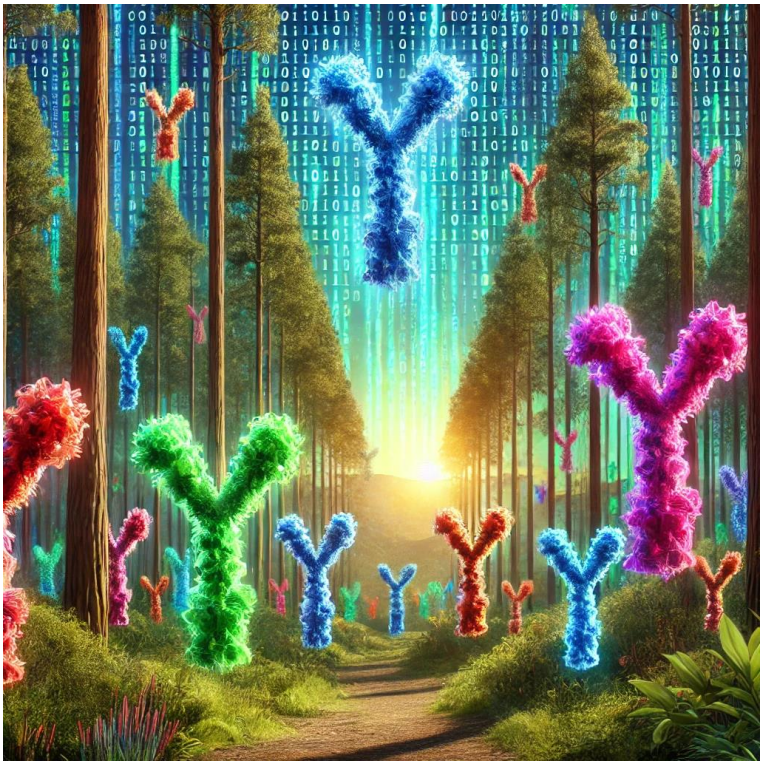
AntibodyForests: delineating B cell evolution at a single-cell resolution

This thesis constitutes the major research project for the Master's in Infection and Immunity at the University of Utrecht.



Created with ChatGTP-4, OpenAI [1].

Student: V.F.W. (Valentijn) Tromp, B.Sc (6644481)

Supervisor: Dr. A.D. (Alex) Yermanos

Daily supervisor: D. (Daphne) van Ginneken, M.Sc

Second assessor: Dr. J. (Julia) Drylewicz

August 23, 2024

UMC Utrecht
Center for
Translational Immunology

Universiteit
Utrecht

Table of contents

Footnote: For this project, ChatGPT-4 was utilized to translate text and provide feedback on the writing style and grammar, enhancing the readability of this thesis [1].

Abstract

B cells are central to the adaptive humoral immune response, producing antibodies that recognize and neutralize pathogens. This thesis introduces AntibodyForests, an R package developed for analyzing single-cell B cell repertoire sequencing data and study B cell evolution. The package facilitates the import of single-cell sequencing output files and the inference of lineage trees for clonotypes, allowing researchers to explore B cell evolutionary trajectories. AntibodyForests offers the flexibility to infer lineage trees using multiple algorithms and includes unique pruning options to remove internal nodes, catering to diverse research needs. This versatility makes AntibodyForests a powerful tool for investigating B cell evolution, a feature that is currently lacking in most other tools. To address undersampling in single-cell sequencing experiments, a computational approach was developed to integrate spleen-derived bulk transcriptomic data to complement bone marrow-derived single plasma cell data. Applied to a dataset from five mice immunized with ovalbumin, the study found significant congruence between single-cell and bulk repertoire sequencing data. Specifically, many clones contained identical antibody sequences in both datasets, and the number of cells in the single-cell dataset was positively correlated with the number of IgM isotype transcripts in the bulk RNA. This integration enhances the resolution with which B cell evolution can currently be tracked and analyzed. Future work will focus on expanding the package's capabilities to integrate phenotypic data and computational predictions, enabling deeper insights into antibody evolution and immune responses.

Layman summary

The human adaptive immune system plays a vital role in defending the body against harmful pathogens. Central to this defense are B cells, which produce antibodies that recognize and neutralize invaders like viruses and bacteria. Each B cell is capable of producing a unique antibody, enabling the human body to recognize a wide variety of foreign substances. The set of B cells in the human body is unique and dynamic in its nature: it is constantly changing over time, evolving to adapt to new challenges. Understanding how these B cells develop and evolve to produce effective antibodies is crucial for improving vaccines and developing new treatments for diseases.

This project focuses on studying B cells and their antibodies using advanced techniques that analyze them at the level of individual cells. Traditionally, scientists have studied B cells by examining large groups of them at once, which provides an overview but misses the finer details of how individual B cells change and evolve. However, newer methods allow researchers to examine individual B cells one at a time, providing much more detailed information about how each cell's antibody develops and changes in response to infections.

To study these cells, a specialized user-friendly computer program called AntibodyForests was developed. This program is used to analyze data from a technology known as single-cell sequencing, which captures detailed information from individual B cells. By tracking these cells over time, AntibodyForests helps researchers organize the data and build "family trees" of B cells, showing how different cells are related and how they have evolved to produce different antibodies, adapting to better fight off pathogens.

However, single-cell sequencing only captures a small portion of all the B cells in the body, which can lead to gaps in the tracking of B cell evolution. To address this, the single-cell data was combined with data from a different method called bulk sequencing, which looks at B cells from a broader perspective. By integrating these two types of data, the study aimed to get a more complete picture of how B cells evolve in response to an infection.

This study applied this approach to data from mice that had been exposed to a harmless protein similar to what might be found in vaccines. By analyzing the B cells from these mice, it was found that combining single-cell and bulk data provided a better understanding of B cell evolution, although some challenges remain. For example, not the complete set of B cells was captured, and the controlled environment in which the mice were kept might not fully represent real-world conditions where organisms are exposed to many different pathogens over their lifetimes.

In the future, improving AntibodyForests to include even more types of data, such as information about how well antibodies bind to their targets, could provide deeper insights into how B cells develop and respond to diseases. This could help scientists design better vaccines and antibody-based treatments for a range of infections and immune-related conditions.

Glossary

AID	activation-induced cytidine deaminase
BCR	B cell receptor
C (gene)	constant gene
C (region)	constant region
CDR	complementarity-determining region
D (gene)	diversity gene
FDC	follicular dendritic cell
FWR	framework region
GC	germinal center
H (chain)	heavy chain
HSC	hematopoietic stem cell
Ig	immunoglobulin
J (gene)	joining gene
L (chain)	light chain
MHC(I//II)	major histocompatibility complex (I//II)
ML	maximum likelihood
MP	maximum parsimony
MST	minimum spanning tree
NJ	neighbor-joining
OVA	ovalbumin
RAG	recombination-activation gene
RSS	recombination signal sequence
SHM	somatic hypermutation
SPF	specific pathogen-free
TCR	T cell receptor
Tfh	T follicular helper cells
V (gene)	variable gene/segment
V (region)	variable region (encoded by rearranged VDJ exon)

Introduction

B cells: essential for pathogen defense, vaccination, and therapeutic antibodies

Humoral immunity is a crucial component of the adaptive immune system, primarily mediated by antibody-producing plasma cells that arise after naïve mature B cells recognize their target antigen and get activated. B cells play an essential role in recognizing pathogens, neutralizing toxins, and facilitating the destruction of infected cells. The human body harbors an estimated number of 10^{11} B cells, each capable of producing a unique antibody [2]. This diversity is achieved through the process of V(D)J recombination, somatic hypermutation (SHM), and class switch recombination (CSR), collectively generating a vast B cell repertoire. This repertoire is the entire collection of B cell receptors (BCRs), allowing the immune system to recognize and respond to an immense variety of antigens. Understanding the B cell repertoire is essential for various fields, including infectious disease research, vaccine development, and immunotherapy. By studying the composition and dynamics of the B cell repertoire, researchers can gain insights into how the immune system adapts to pathogens, the mechanisms behind successful vaccinations, and the identification of antibody candidates for therapeutic use.

A historical appreciation on landmark studies in the field of humoral immunity

Humoral immunity has been a central focus of immunological research for over a century (*Figure 1*) [3]. The field began in 1890 with the discovery that immunity could be transferred through serum components, which were later identified as antibodies [4]-[5]. Early studies demonstrated the protective role of antibodies, including their maternal transmission and their presence in specific globulin fractions, which provided immunity against diseases such as diphtheria and tetanus [6]-[9]. Advancements in techniques like electrophoresis and ultracentrifugation allowed for the isolation and study of antibodies, leading to the identification of distinct antibody isotypes (such as IgG and IgM) and the understanding of their structure, consisting of heavy and light chains [10]-[20]. The identification of plasma cells as the primary producers of antibodies, along with the development of clonal selection theory, highlighted that the vast B cell repertoire, with each B cell capable of recognizing a specific antigen, enables a targeted humoral immune response through the clonal expansion of these specific B cells upon antigen exposure [21]-[29]. Research in the 1960s and 1970s further elucidated the roles of B cells in humoral immunity, with statistical analyses of light chain sequences revealing the highly variable complementarity-determining regions (CDRs) within antibodies and uncovering the mechanisms of V(D)J recombination and SHM that generate antibody diversity [30]-[44]. The delineation of B cells and T cells into their respective roles in humoral and cellular immunity was also clarified during this time. The development of recombinant DNA technology in the early 1970s revolutionized the study of B cells and immunoglobulins, allowing for detailed exploration of gene rearrangement and providing deeper insights into the molecular mechanisms of antibody diversity, solidifying our understanding of B cell evolution [45]-[56].

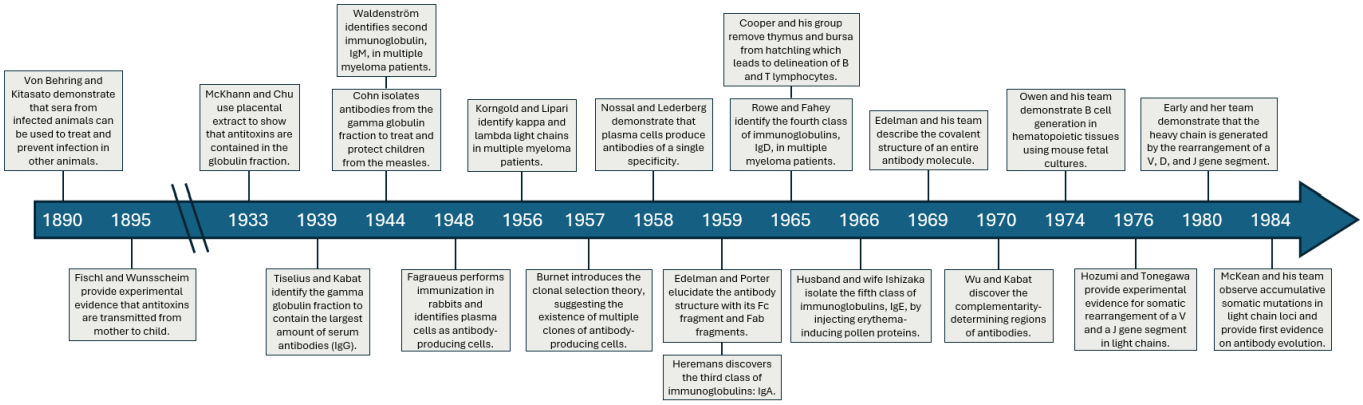


Fig. 1: A historical timeline of the research into antibodies and B cells
The timeline spans from 1890, when Emil on Behring discovered the potential of serum therapy, to 1984, when David McKean and his team discovered the process of somatic hypermutation. This timeline highlights key breakthroughs and important theories that have culminated in our current understanding of B cells and their role in humoral immunity.
Adapted adapted from Cooper [3].

V(D)J recombination and junctional diversity underly initial antibody diversification

B cells produce antibodies that are the main actors in humoral immunity. These molecules consist of two main fragments: the Fab (antigen-binding) fragment and the Fc (crystallizable) fragment [57]-[58]. Structurally, they are composed of two heavy and two light chains, linked by disulfide bonds (*Figure 2A*). The heavy and light chains contain constant and variable regions, with the variable regions forming the antigen-binding sites. The variable regions can be divided into three complementarity-determining regions (CDR1-3), showing the highest variability, which lie in between four conversed framework regions (FWR1-4). Antibodies are encoded by immunoglobulin (Ig) genes. In the human genome, these Ig genes are organized into three loci: the *IGH* locus on chromosome 14 encodes the heavy chain, while the *IGK* and *IGL* loci on chromosomes 2 and 22 encode the light chain [59]-[60].

Each locus contains clusters of gene segments, where one segment from each cluster is selected during V(D)J recombination to encode the antibody's variable region. At the 5' end, there are variable (V) gene segments, with approximately 50 in the *IGH* locus, *35* in the IGK locus, and *30* in the IGL locus. These segments are preceded by leader exons that encode signal peptides that guide the antibody to the endoplasmic reticulum membrane during translation. Downstream of the V genes are the joining (J) segments, typically 30 to 50 base pairs long. In the *IGH* locus, there are additional diversity (D) segments situated between the V and J segments. These D segments are absent in the light chain loci. The number of J and D segments also varies among the different loci. The constant (C) region genes are located 3' of the J segments. Each Ig locus has a unique arrangement and number of C region genes (*Figure 2B*).

During B cell development, one segment from each cluster is selected and recombined to form a functional VDJ exon encoding the variable region of the antibody [57]-[61]. In the *IGH* locus, this involves two sequential recombination events: D-to-J recombination followed by V-to-DJ recombination (*Figure 2C*). In contrast, the light chain loci (*IGK* and *IGL*) lack D segments, so a single rearrangement event joins a V segment directly to a J segment. Once the VDJ (or VJ) exon is generated, it is transcribed along with the first two C genes. Alternative splicing then determines whether the μ or δ heavy chain proteins are expressed.

V(D)J recombination is regulated by recombination signal sequences (RSSs), which are specific sequences located adjacent to each gene segment: 3′ of V segments, 5′ of J segments, and flanking both sides of D segments [57]-[61]. RSSs consist of a conserved heptamer, a non-conserved spacer of either 12 or 23 nucleotides, and a conserved nonamer. The 12/23 rule dictates that recombination occurs only between gene segments with differing spacer

lengths, ensuring the correct assembly of V, D, and J segments. In the *IGH* locus, the V and J segments are flanked by 23-nucleotide spacers, preventing them from directly recombining. Instead, the D segments, flanked by 12-nucleotide spacers, first undergo D-to-J recombination, followed by V-to-DJ recombination, leading to the formation of a functional antigen receptor.

The recombination process consists of four steps: synapsis, cleavage, hairpin opening and end processing, and joining (*Figure 2D*) [57]-[61]. During synapsis, the gene segments and their RSSs are brought together through chromosomal looping. Cleavage is mediated by recombination-activating gene (RAG) proteins, which introduce double-stranded breaks at the RSS-coding sequence junctions. The enzyme Artemis then opens the hairpin structure created by the cleavage, allowing for the removal of some germline nucleotides. Terminal deoxynucleotidyl transferase further diversifies the DNA by adding nucleotides to the broken ends. These ends are finally joined by nonhomologous end joining. The addition and removal of nucleotides, known as junctional diversity, primarily contributes to the extensive variability observed in the junction regions. The CDR3 region is located at the junction of the V, D, and J gene segments; this explains why it exhibits the highest diversity.

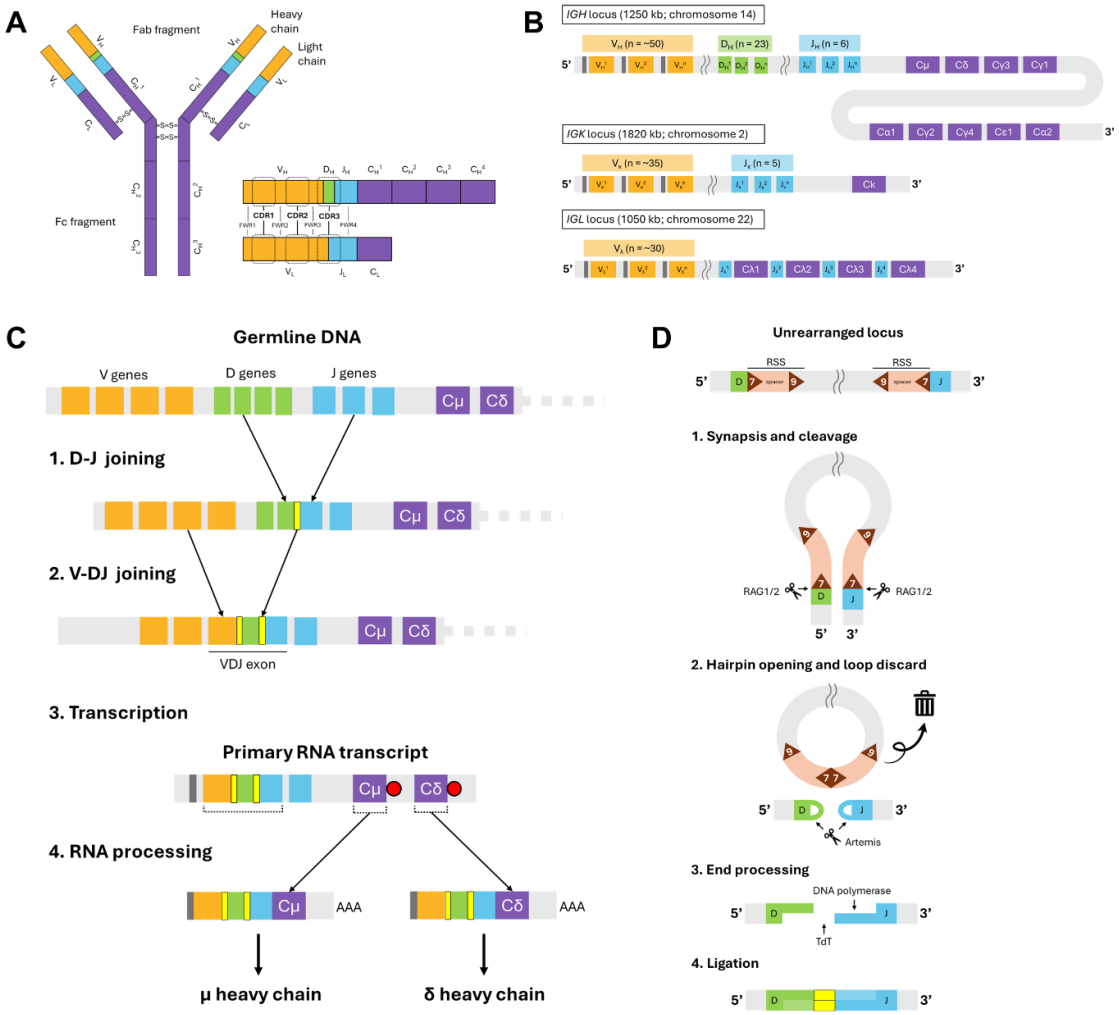


Fig. 2 Antibody structure and molecular mechanisms of antibody generation and diversification
(A) Antibodies are composed of two light and two heavy chains, each with constant and variable regions. The variable regions, responsible for antigen binding, contain three complementarity-determining regions (CDR1-3) that show the highest sequence variability, flanked by four conserved framework regions (FWR1-4).
(B) The human immunoglobulin (Ig) loci: the *IGH* locus on chromosome 14 contains V, D, J, and C gene segments, while the *IGK* and *IGL* loci on chromosomes 2 and 22 contain V, J, and C segments. Leader sequences, shown as dark grey boxes, upstream of V segments guide antibodies to the endoplasmic reticulum.

(C) V(D)J recombination at the *IGH* locus involves sequential D-to-J and V-to-DJ rearrangements, forming the variable region of the heavy chain. This process introduces nucleotide additions and deletions at the junction sites, shown as yellow boxes, contributing to antibody diversity. Alternative splicing results in the dual expression of both μ and δ heavy chains.
(D) The molecular mechanism of V(D)J recombination involves multiple steps: synapsis brings gene segments together, recombination-activating gene 1 and 2 (RAG1/RAG2) complex mediates cleavage, Artemis opens hairpins, and terminal deoxynucleotidyl transferase (TdT) adds nucleotides at the junctions, followed by end joining, generating junctional diversity.
Adapted from Abbas *et al*. [57].

Germinal center reaction and class switching enhance antibody diversity

The process of V(D)J recombination occurs during B cell development, which starts with the differentiation of lymphoid cells from hematopoietic stem cells (HSCs) in the human fetal liver and fetal/adult bone marrow [57][62]. Cells that arise from the fetal liver-derived HSCs will give rise to B cell subsets of the B-1 lineage, while cells that develop from bone marrow-derived HSCs will give rise to B cell subsets of the B-2 lineage. Eventually, this process culminates in the majority of mature B cells belonging to the B-2 lineage class of follicular B cells, which express membrane-bound IgM and IgD.

When mature naïve B cells, which reside in the lymphoid follicles, encounter their specific antigen, they undergo activation. This activation typically requires assistance from CD4+ T helper cells [57]. The B cell binds to the antigen via its BCR, internalizes it, and processes it into fragments presented on major histocompatibility molecules class II (MHCII) molecules. Presented antigens are recognized by T helper cells, and the interaction between CD40 on B cells and CD40L on T helper cells initiates B cell proliferation and differentiation [63]. Initially, some activated B cells quickly differentiate outside the follicle into short-lived plasma cells that secrete low-affinity antibodies, providing an immediate, albeit temporary, immune response.

Simultaneously, other activated B cells re-enter the follicle to form a germinal center (GC), where they undergo somatic hypermutation (SHM) and affinity maturation (*Figure 3A*) [57][64]. In the GC's dark zone, B cells proliferate rapidly while expressing the activation-induced cytidine deaminase (AID). This enzyme deaminates cytosine residues, converting them into uracil residues. Following DNA replication, this deamination mostly results in a C>T transversion. In some cases, the uracil residue is excised by the enzyme urical-DNA glycosylase, initiating the base-excision repair, resulting in substitutions with any of the four nucleotides. In rare cases, a mismatch repair pathway is activated, resulting in the deletion of the uracil residue and adjacent nucleotides, possibly leading to insertions and deletion. The AID enzyme does not act across the entire genome, but specifically in the V gene region, where it induces generally a point mutation approximately in every 10^3 base pairs, resulting in about one mutation per cell division given the length of the V gene. While the exact mechanism of AID's specificity towards the V gene region is not fully understood, it is observed that AID preferentially targets sequences called hotspots, while other regions, known as coldspots, are less frequently mutated [65].

SHM can either enhance or diminish the antibody's affinity for its antigen. B cells leaving the dark zone migrate into the GC's light zone, where they encounter antigen-presenting follicular dendritic cells (FDCs) and T follicular helper (Tfh) cells [57]. In this highly competitive environment, only B cells with higher-affinity BCRs can effectively bind the available antigen, receiving the necessary survival signals (*Figure 3B*). B cells with lower affinity or self-reactive BCRs are typically eliminated through apoptosis. This selective process ensures that the immune system refines its response by producing B cells capable of generating high-affinity antibodies (*Figure 3C*).

During the immune response, B cells may undergo class switching, a process that changes the antibody isotype (*e.g.*, from IgM to IgG, IgA, or IgE) while retaining the same antigen specificity [57][66]. This process allows the immune system to tailor its response to different types of pathogens by altering the effector functions of antibodies. Class switching is

predominantly directed by specific cytokines produced by T helper cells. For example, IL-4 and IL-13 promote switching to IgE, crucial for defense against parasitic infections and allergic responses, while TGF-β and IL-5 drive switching to IgA, important for mucosal immunity. Notably, class switching primarily occurs before the GC reaction, often in extrafollicular regions where initial immune responses are mounted [67]. The molecular mechanism of class switching involves recombination at the IGH locus, where the rearranged VDJ exon is brought into proximity with a different constant (C) region gene. This recombination is facilitated by AID-mediated DNA breaks and repair, enabling the B cell to produce antibodies with different effector functions suited to the immune challenge at hand.

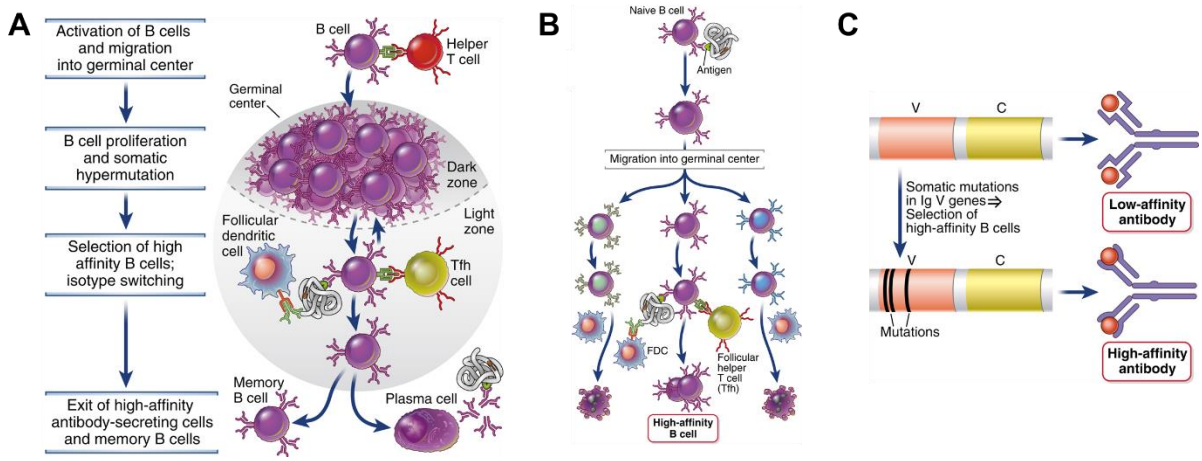


Fig. 3 Germinal center reaction and antibody affinity maturation
(A) In the germinal center (GC), activated B cells proliferate in the dark zone, undergoing somatic hypermutation (SHM) with the help of the enzyme activation-induced cytidine deaminase (AID). B cells move to the light zone, where they interact with antigen-presenting follicular dendritic cells and T follicular helper cells. B cells with high-affinity B cell receptors (BCRs) receive survival signals and return to the dark zone for more proliferation and SHM before differentiating into plasma cells or memory B cells.
(B) SHM in the dark zone alters antibody structure and binding affinity. In the light zone, B cells compete for limited antigen; those with the highest affinity bind antigen, survive and proliferate, increasing overall antibody affinity.
(C) Initially, B cells produce low-affinity antibodies. SHM leads to high-affinity antibodies as B cells mature through the GC reaction. Retrieved from Abbas *et al*. [57].

From Sanger sequencing, to next-generation sequencing, to single-cell sequencing

For over three decades, starting in 1977, Sanger sequencing was the primary method for DNA sequencing [68]. This traditional method provided foundational insights into BCR repertoires by identifying specific V, D, and J gene usage in response to particular antigens, as well as uncovering public sequences shared across individuals in contexts such as influenza, Epstein-Barr virus, and human immunodeficiency virus infections [69]-[72]. However, its low throughput and high costs limited its capacity for comprehensive BCR analysis. The advent of next-generation sequencing (NGS) in the 2000s revolutionized the study of BCR repertoires by enabling high-throughput sequencing of entire BCR sequences at the single nucleotide level [73]-[74]. This advancement allowed for a comprehensive analysis of the B cell repertoire, revealing important aspects such as mutation rates, diversity, and distribution of BCRs across the immune system. Studies in human immunodeficiency virus-infected individuals uncovered that certain V genes were preferentially used in antibodies targeting specific viral proteins, providing crucial insights for vaccine development [75]-[76]. However, despite these advancements, these NGS platforms had limitations, particularly in its inability to capture VDJ-VJ chain pairing and accurately quantify clonal expansion. The correct pairing of VDJ and VJ chains is vital for determining the specificity and affinity of antibodies, and without this information, understanding the full functionality of the BCR repertoire was challenging.

In recent years, single-cell sequencing platforms have gained popularity for their ability to simultaneously capture VDJ-VJ chain pairing and quantify clonal expansion. These advancements have enabled a deeper exploration of the drivers behind clonal selection, expansion, and evolution. Neumeier *et al*. (2022) conducted an in-depth genotypic and phenotypic characterization of antibody repertoires in mice after immunization with the model antigen ovalbumin (OVA), finding that while the most expanded clones showed antigen specificity, the broader clonal repertoire displayed stochastic antigen binding [77]. Additionally, clonal expansion did not necessarily correlate with antigen-binding affinity or SHM. Shehata *et al*. (2019) already explored the relationship between affinity maturation and antibody properties, finding that while affinity maturation was associated with increased specificity, it also led to diminished conformational stability [78]. This supports the notion that SHM contributes not only to enhanced affinity but also to potentially destabilizing mutations. Agrafiotis *et al*. (2023) utilized single-cell sequencing to create a comprehensive atlas of B cell repertoires and transcriptomes following immunization with the model antigen human tumor necrosis factor receptor 2 in mice [79]. This study revealed that antigen-specific, clonally expanded B cells were more prevalent in younger mice, with features like IgG subtype and sequence composition linked to antigen specificity. However, the integration of various repertoire features, such as germline gene usage, did not distinguish antigen-specific from nonspecific B cells, highlighting the complexity of B cell selection and expansion processes.

Clonotyping and lineage tree-inference provide insight into evolutionary pathways

Capturing B cells for single-cell sequencing results in a diverse sample from multiple B cell clones, each originating from a different ancestor. To analyze B cell selection and evolution, these cells must first be assigned to a clonal lineage, a process known as clonotyping. One commonly used clonotyping tool is the 10x Genomics-developed enclone tool [80][81]. Enclone utilizes germline gene annotations, CDR3 sequence identity, and the presence of shared SHMs to accurately group cells into clonotypes. During clonotype grouping, cells are first organized into exact subclonotypes based on identical V(D)J transcripts. These subclonotypes are then iteratively merged into clonotypes by comparing pairs of exact subclonotypes, focusing on shared SHMs, V-J gene identity, and minimal CDR3 mutations.

Once clonotypes are identified, their evolutionary relationships can be inferred by constructing B cell lineage trees. These trees visually represent the evolutionary pathways of B cells, with the germline sequence (the original unmutated ancestor) at the top, branching out as mutations accumulate over time. The branches of the tree represent different descendant B cells, each carrying unique mutations acquired during the immune response. Several methods are available to build these lineage trees, each differing in how they estimate the tree's structure and branch lengths [82]. Distance-based methods, like neighbor joining (NJ) and minimum spanning trees (MST), cluster BCR sequences based on genetic differences, though these methods may sacrifice some accuracy for speed. Parsimony-based methods aim to construct the simplest tree, requiring the fewest mutations to explain the observed sequences. While effective when mutations are sparse, this approach can become biased when mutations are frequent. Likelihood-based methods use models of nucleotide or protein changes to estimate tree topologies and branch lengths. These methods usually perform better in situations where mutations are common, with maximum likelihood offering a good balance between accuracy and computational efficiency [83].

Likelihood-based methods depend on substitution models for inference of evolutionary relationships. There is a great variety of models available, from simple models like Jukes and Cantor's, which assume uniform mutation rates, to more complex models like the General Time-Reversible (GTR) model, which accounts for different rates of nucleotide substitutions

and site-specific variability [84]. Codon substitution models, such as the M0 model, provide a biologically more realistic framework for analyzing protein-coding sequences by accounting for transition/transversion rate bias, codon usage bias, and selective restraints at the protein level through physicochemical distances [85]. However, general protein evolution models may not be suitable for antibody lineages due to their uniquely high mutation rates and the influence of specific mutational hotspots and coldspots driven by the enzyme AID. The IgPhyML model addresses this by incorporating biases from these hotspot and coldspot motifs, making it particularly useful for modeling antibody evolution [86].

Aim of the project

Recent findings that antibody affinity does not necessarily correlate with clonal expansion and SHM have raised fundamental questions about whether antibody evolution follows any deterministic rules. As sequencing technologies continue to advance, the volume of data on BCR repertoires is rapidly increasing. Numerous tools are available for analyzing V(D)J recombination, sequence diversity, and clonal expansion, as reviewed by Zheng *et al* (2022) [87]. However, there is a significant lack of software that enables comprehensive, repertoire-wide analyses both within and across multiple datasets. This gap in tools hinders our ability to use the available datasets to address the question whether clonal expansion and evolution follow deterministic rules, leaving critical questions about B cell phylogenies unanswered.

This project aimed to address this gap by developing and contributing to an R package called AntibodyForests. The software is designed to import single-cell sequencing data from the 10x Genomics platform, infer phylogenetic networks for all observed clonotypes, and perform downstream analyses to track evolutionary patterns within and across entire repertoires. AntibodyForests seeks to provide a user-friendly, comprehensive tool for researchers, facilitating deeper insights into B cell repertoire evolution and improving our understanding of B cell phylogenies.

The second part of this project focused on developing a computational approach to enhance the resolution of single-cell sequencing by integrating bulk sequencing data. Current single-cell platforms capture only a small fraction of the BCR repertoire, potentially biasing the analysis of B cell evolution. By leveraging bulk sequencing, this approach aims to extend the coverage of clonotypes identified in single-cell data, thereby providing a more complete and accurate picture of B cell evolution. This enhancement is crucial for overcoming the limitations of single-cell sequencing and enabling more robust analyses of B cell dynamics.

Results: Analyzing B cell evolution with AntibodyForests

Introducing AntibodyForests: an R package for B cell selection and evolution analysis

The primary objective of this project was to contribute to the development of the R package AntibodyForests, designed to analyze B cell selection and evolution by assessing germline gene usage, clonal expansion, and inferring phylogenetic trajectories. The core pipeline of the AntibodyForests package is depicted in *Figure 4*, with a comprehensive overview of all available functions provided in the package vignette (S1: AntibodyForests vignette). Additionally, the source code for the individual functions is included in the supplementary materials (S2: Functions of AntibodyForests). To address undersampling in single-cell sequencing, we developed a computational approach within AntibodyForests to integrate bulk sequencing data, complementing the single-cell dataset and enhancing the resolution of B cell clonal evolution. Although not yet fully functionalized, this integration aims to provide a more comprehensive view of B cell repertoires. The code used for the integration is also provided in the supplementary materials (S3: R Markdown report). In the following sections, the package will be presented, followed by the demonstration of its utility through the analysis of both a single-cell dataset and a bulk dataset.

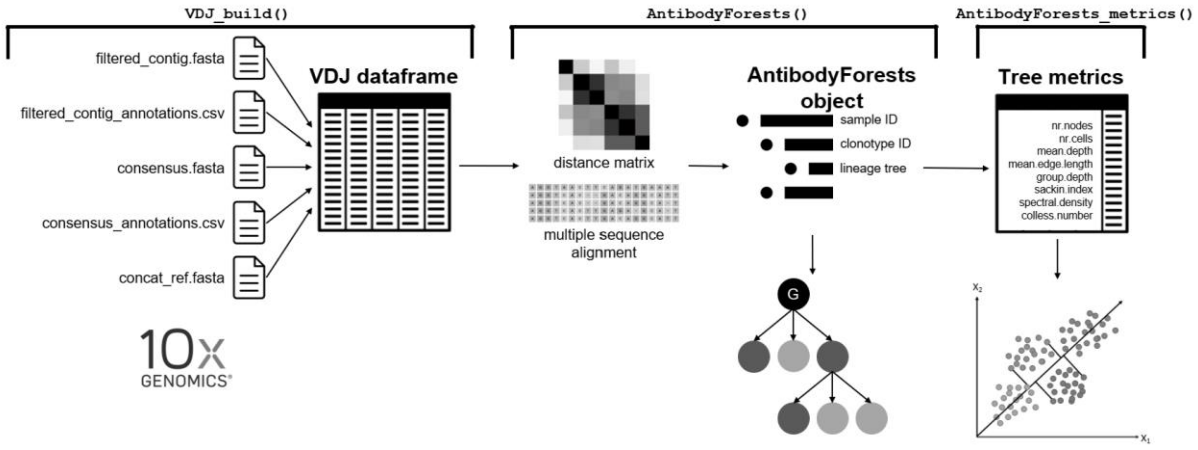


Fig. 4 Schematic outline of the analysis pipeline of AntibodyForests
The AntibodyForests workflow imports 10x Genomics output files into a single R object, consolidating sequence and annotation data files, and infers lineage trees for clonotypes, and tree topology metrics can be visualized to assess repertoire-wide similarity.

Importing 10x Genomics output into VDJ data frame with VDJ_build()

Analyzing B cell repertoires, such as assessing germline gene usage and clonal expansion profiles, and inferring accurate phylogenies from B cell sequencing data, require a streamlined approach to organize the complex data generated by the 10x Genomics platform. The default data processing, including quality checks, contig assembly, and annotation, is handled by Cell Ranger, which is a set of pipelines provided with the 10x Genomics platform [88]-[89]. Cell Ranger outputs multiple files, including sequences of filtered assembled contigs, annotations, consensus sequences, and reference germline sequences. However, these output files are not organized in a way that facilitates easy downstream analysis. To address this, we developed the `VDJ_build()` function to import all these data files into one R object, called the VDJ dataframe. This function not only streamlines data handling but also includes a filtering parameter to exclude cells for which no single VDJ and single VJ transcripts were recovered, ensuring a more focused dataset for analysis.

Additionally, we implemented two key features to enhance the usability of germline sequences within the VDJ dataframe: 1) trimming of germline sequences and 2) replacement

of the CDR3 region. The germline sequences are concatenated V(D)J reference segments detected based on the consensus sequence. These sequences are approximations for the first rearranged B cell, from which all cells of that clonotype are theoretically derived. The trimming option refines these sequences by removing nucleotides that may have been deleted during recombination at junction sites, aligning the germline sequence more closely with the observed data through pairwise alignment with the clonotype's consensus sequence. The CDR3 replacement option substitutes the germline CDR3 region with the most frequently observed sequence in each clonotype. Since the CDR3 region has the highest variability and significantly contributes to antibody specificity, these modified germline sequences can be used in laboratory settings to systematically study the effects of mutations in variable regions on antibody specificity and binding affinities.

Incorporating seven lineage tree-inference algorithms into the AntibodyForests() function

After importing the 10x Genomics output files into the VDJ dataframe, the next step in the AntibodyForests analysis workflow involves the inference of evolutionary trees. To infer lineage trees, we developed the `AntibodyForests()` function, which is a significant outcome of this project. This function utilizes the VDJ and VJ sequences from the VDJ dataframe as input and infers an evolutionary tree that maps B cell evolution at the sequence level across the entire repertoire, generating a distinct tree for each clonotype identified. The `AntibodyForests()` function is designed to streamline and simplify the process of tree construction by collapsing barcodes with identical VDJ and VJ sequences into nodes with unique sequences. To cater to different research needs, the function integrates various tree construction algorithms, combining them in a user-friendly R interface. This versatility allows users to choose between different methods of tree inference, depending on their specific research goals.

One of the key features of the `AntibodyForests()` function is its ability to handle both string distance matrices and multiple sequence alignments (MSA) to infer trees. We wrote the custom “default” algorithm that uses a string distance matrix to position the germline node at the root and iteratively link subsequent nodes based on the minimum distance to existing nodes, ensuring that the resulting trees accurately represent the evolutionary relationships among B cells. In addition to this custom algorithm, we incorporated other tree construction algorithms by importing them from established packages: the minimum spanning tree (MST) and neighbor joining (NJ) algorithms were imported from the ape package, while the maximum parsimony and maximum likelihood algorithms, which are MSA-based methods, were imported from the phangorn package [90]-[91]. An overview of these methods is provided in *Table 1*.

When employing phylogenetic algorithms to infer evolutionary pathways, the resulting trees contain branch points, called internal nodes, which represent divergence events. In a phylogenetic context, these internal nodes denote the common ancestors of all descendant groups from those branch points. In the context of B cell lineage trees, these internal nodes may represent cells not captured during plasma cell isolation. However, it is possible that a significant portion of a clonotype has been captured, making some branching points unnecessary, as one of the descendant cells could be the common ancestor. To address potential overcomplications in the inferred trees, we implemented internal node removal algorithms within the `AntibodyForests()` function. An overview of these algorithms is provided in *Table 2*. Running these algorithms after tree inference with a phylogenetic algorithm results in trees that lack internal nodes and consist solely of sequence-recovered nodes. It is noteworthy that similar pruning methods are not commonly available in other packages, highlighting a unique feature of AntibodyForests.

In addition to the five aforementioned primary algorithms, we extended the functionality of AntibodyForests by implementing the codon substitution model and the antibody lineage-specific codon substitution model [85]-[86]. The general codon substitution model can be employed by specifying the ML algorithm along with the codon substitution model ("M0"). Unfortunately, the IgPhyML tree could not be executed within the R environment directly, so we developed a set of functions to create the input files for the IgPhyML tool. Within the `AntibodyForests()` function, users can specify the path to the IgPhyML output files, allowing the creation of an AntibodyForests object structurally similar to those constructed with other implemented algorithms. When importing trees from IgPhyML output files, the internal nodes can be pruned using one of the internal node removal algorithms.

This comprehensive suite of lineage tree-inferring algorithms within the AntibodyForests package represents a significant advance in the computational tools for the study of B cell evolution, offering researchers a robust and flexible platform for analyzing B cell repertoires.

Table 1: Overview of tree-inference algorithms implemented in the AntibodyForests() function
The "package" column indicates the R package from which the algorithm is derived. The "default" algorithm is self-written, so no R package is employed for it. The "input" column specifies the type of input used by the algorithm: either "dist" for a string distance matrix or "msa" for a multiple sequence alignment.

Algorithm	Package	Input	Description
"default"	-	dist	Positions the germline node above and iteratively links nodes based on minimum distances. Requires specifying how to handle ties, such as prioritizing nodes with the highest expansion or the minimum distance to the germline node.
minimum-spanning tree (MST)	ape	dist	Constructs networks with the minimum sum of edge lengths, similar to the default algorithm, but does not start with positioning the germline node at the top. Requires subsequent reorganization into a germline-rooted lineage tree.
neighbor joining (NJ)	ape	dist	Creates phylogenetic trees by joining pairs of nodes with the minimum distance, resulting in a tree with internal nodes (unrecovered sequences representing common ancestors) and terminal nodes (observed sequences).
maximum parsimony (MP)	phangorn	msa	Builds phylogenetic trees by minimizing the total number of changes needed to explain differences among sequences, using a multiple sequence alignment (MSA) as input and prioritizing the simplest explanations.
maximum likelihood (ML)	phangorn	msa	Builds phylogenetic trees by estimating tree topology and branch lengths that maximize the likelihood of the given sequence data under a specified evolutionary model. By default, it selects the model that returns the highest likelihood.

Table 2: Overview of internal node removal algorithms implemented in the AntibodyForests() function
This table describes the different methods available for removing internal nodes from phylogenetic trees.

Algorithm	Description
"zero.length.edges.only"	Removes internal nodes that only have zero-length edges to terminal nodes. This eliminates internal nodes that do not contribute to the tree structure and are directly connected to terminal nodes.
"connect.to.parent"	First removes internal nodes with zero-length edges to terminal nodes, then connects all terminal nodes directly to the first parental sequence-recovered node higher in the tree. If the number of zero-length edges is minimal, the resulting tree may become germline-directed.
"minimum.length"	Iteratively removes internal nodes by prioritizing the deletion of edges with the minimum length. This simplifies the tree structure by removing internal nodes while preserving the overall topology, ensuring that the most direct connections are retained.
"minimum.cost"	Iteratively removes internal nodes, prioritizing edges that result in the smallest increase in the total edge length when replacing internal nodes with terminal nodes (this increase is referred to as the "cost"). Similar to the "minimum.length" algorithm, it retains the most direct connections, while ensuring the tree has the minimal increase in the sum of all edge lengths.

Integrating bulk sequences into single-cell clones using VJ annotations and CDR3 sequences

Following the inference of lineage trees across the entire B cell repertoire, a key challenge remained: the undersampling inherent in single-cell sequencing. In the dataset from the five mice, an estimated 3-11% of plasma cells are included, which limits the resolution of inferred evolutionary pathways [77]. To address this, we integrated bulk sequencing data to complement the single-cell data. The bulk sequences, obtained from the spleens of the same five BALB/c mice, were reannotated using IgBLAST with the mouse IMGT V, D, and J reference germline sequences [92]-[93]. This reannotation allowed for a more precise matching of bulk sequences with the corresponding single-cell clones.

We developed several approaches to map these bulk sequences to the single-cell dataset. First, we filtered clones by matching their VDJ chain V and J annotations and CDR3 lengths. Subsequently, we refined this integration by selecting clones that either contained at least one sequence with an identical CDR3 sequence to the annotated bulk transcripts or met an 85% sequence identity threshold across all sequences within a clone, based on the enclone clonotyping approach [80]. Further filtering was applied to prioritize clones based on expansion and overall sequence identity, ensuring accurate integration and enhancing the resolution of the B cell repertoire. An overview of these integration methods is illustrated in *Figure 5*, which outlines six distinct approaches using different clone selection options and filtering criteria to integrate bulk transcripts into clones from the single-cell dataset. The integration approach is not (yet) functionalized, but the code is provide in the supplementary materials (S3: R Markdown report).

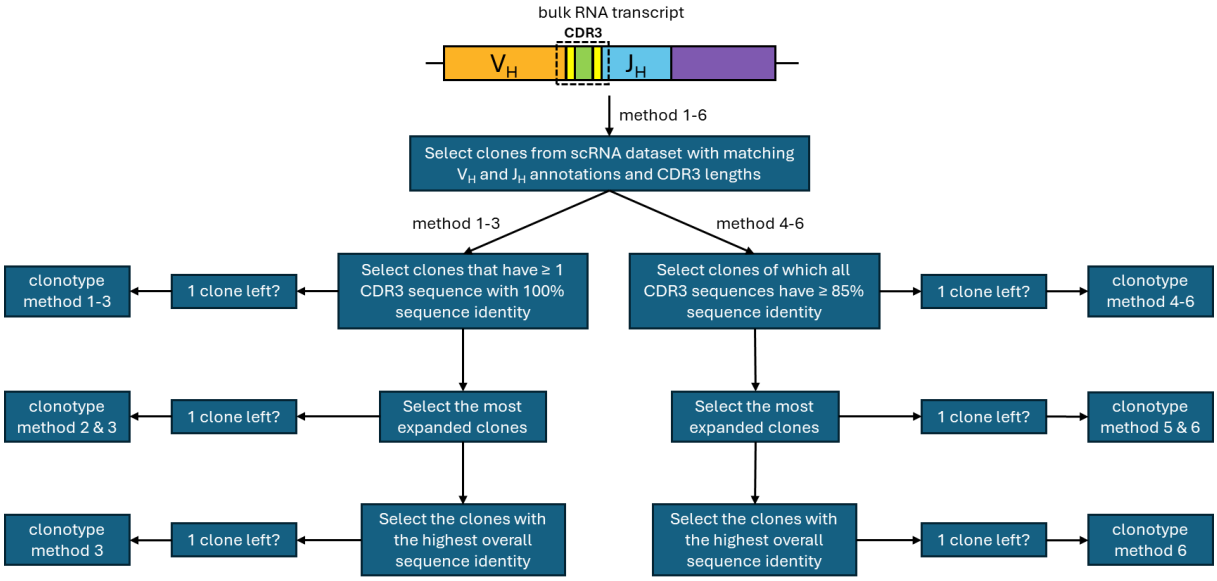


Fig. 5 Integration approaches of transcripts from bulk dataset into clones from single-cell dataset
The diagram illustrates six methods for integrating bulk transcripts into clones observed in the single-cell dataset. The integration process starts with selecting clones from the single-cell dataset with matching VDJ chain V and J gene annotations and CDR3 lengths to the bulk RNA transcript. Methods 1-3 then select clones that have at least one CDR3 sequence with 100% sequence identity to the bulk RNA transcript, whereas method 4-6 select clones for which all CDR3 sequences have at least 85% sequence identity to the bulk RNA transcript. If multiple clones remain, method 2 and 4 select the most expanded clones. If still multiple clones are selected, method 3 and 6 select the clones with the highest overall sequence identity. Each branch of the diagram leads to the selection of clonotypes based on these criteria.

Leveraging single-cell and bulk BCR sequencing datasets to validate utility of AntibodyForests

To demonstrate the utility of the AntibodyForests package and the integrative approach combining a single-cell and bulk BCR sequencing datasets, the analysis pipelines of the AntibodyForests package were applied to a single-cell dataset from five BALB/c mice, as detailed by Neumeier *et al*. (2022) [77]. These mice were repeatedly immunized with monophosphoryl lipid A-adjuvanted ovalbumin, focusing on filtered bone marrow-derived plasma cells from the femurs and tibias. Additionally, the spleens were removed from these same mice to extract VDJ VDJ chain transcripts, which were then sequenced and processed as outlined by Khan *et al*. (2016) [94]. An overview of this combined experimental approach, including mouse immunization and sequencing, is shown in *Figure 6*.

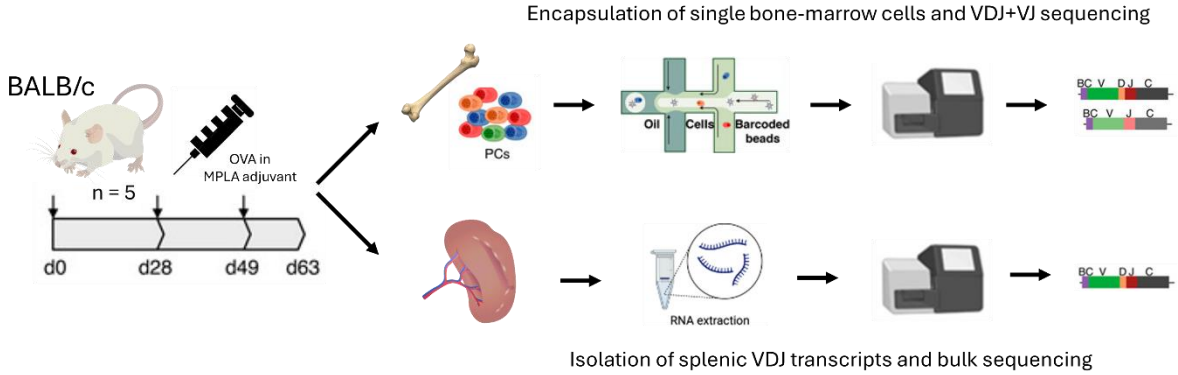


Fig. 6 Outline of mice immunization, followed by single-cell sequencing and bulk sequencing
Five BALB/c mice were subcutaneously immunized with the model antigen ovalbumin (OVA) in a monophosphoryl lipid A (MPLA) adjuvant on days 0, 28, and 49. On day 63, the mice were sacrificed, and femurs and tibias were harvested for bone marrow extraction and plasma cell isolation, while spleens were collected for RNA extraction. Plasma cells underwent single-cell sorting followed by VDJ and VJ sequencing, while VDJ transcripts from the spleen were subjected to bulk sequencing.
Adapted from Neumeier *et al*. [77].

Majority of B cells express a single VDJ and VJ chain and a κ light chain

To accurately characterize the B cell repertoire and validate the AntibodyForests package, it was essential to first evaluate the completeness of VDJ and VJ chain recovery in the single-cell dataset. This step ensures the reliability of downstream analyses, including lineage tree construction and clonal selection studies. The dataset contained a total of 22,425 cells. The number of cells per sample varies significantly, with samples 1 and 5 containing over 6,000 cells, samples 2 and 4 near 4,000 cells, and sample 3 containing only about 2,200 cells (*Figure 7A*). These totals encompass all cells, with varying numbers of VDJ and VJ chains recorded per barcode. Each barcode is unique to a single cell, which should theoretically express one VDJ and one VJ chain. However, VDJ or VJ transcripts might not be captured or amplified sufficiently, or the resulting contigs might not meet quality requirements. Another possibility is that two cells remained attached during the single-cell sorting, leading to doublets. These cells, referred to as divergent or incomplete cells, mostly lacked a VDJ chain or contained two VJ chains (*Figure 7B*). Using the filtering parameters, these cells were excluded, resulting in a total of 17,891 cells expressing a single VDJ and a single VJ chain (*Figure 7C*). Over 90% of these 'complete' cells express κ light chains (*Figure 7D*). This percentage matches the numbers reported in literature for mouse mature B cells [95]. However, it is important to note that in humans, about 60% of mature B cells express κ light chains.

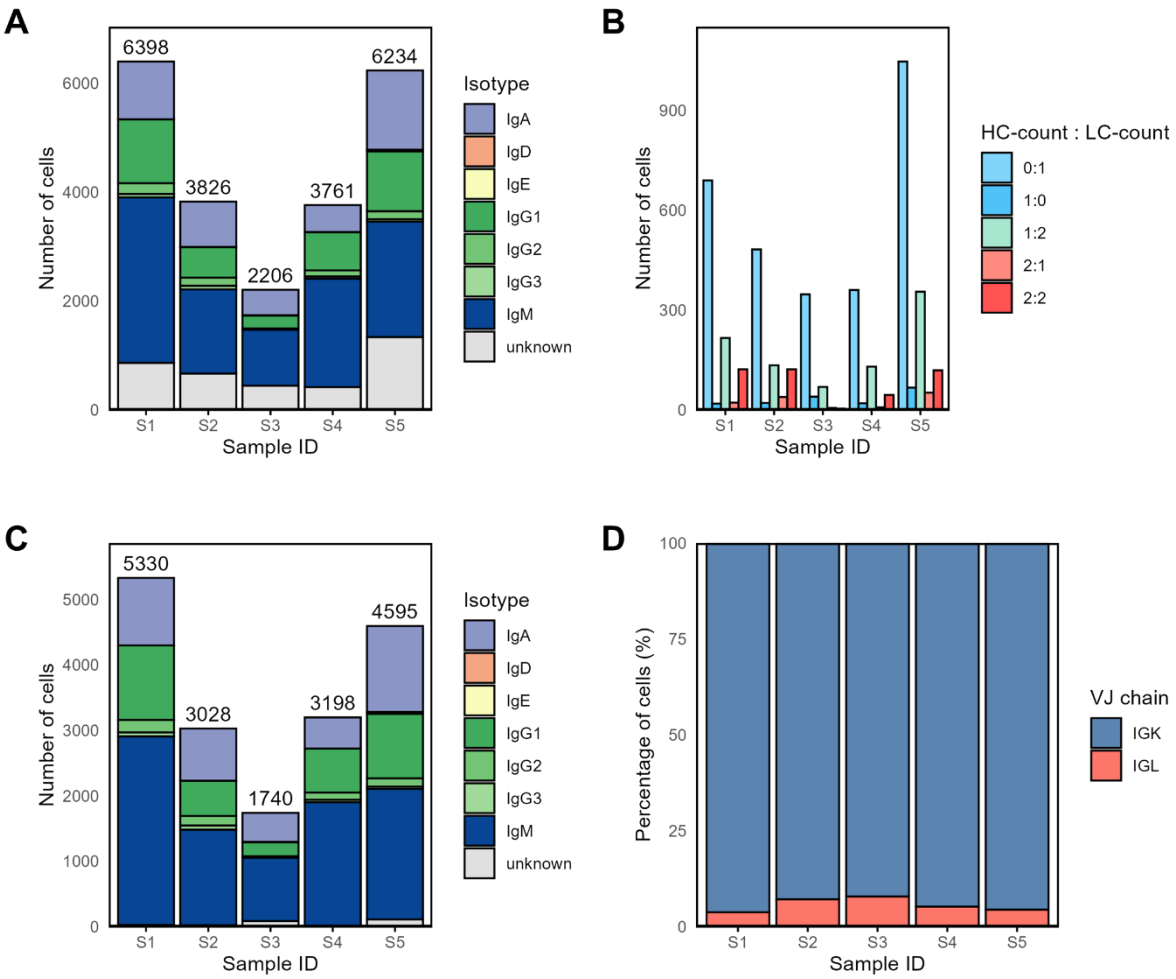


Fig. 7 Number of cells before, during, and after filtering and IGK/IGL usage
(A) Cell count and isotype per sample before filtering, with total cell numbers indicated on the bars.
(B) Cell count lacking a single pair of VDJ and VJ transcripts, indicating cells with no or multiple transcripts.
(C) Cell count per sample after filtering, with totals shown on the bars.
(D) Ratio of cells expressing κ and λ light chains across samples.

The IgM isotype dominates the BCR repertoire, followed by B cells class switched to IgA and IgG

To better understand the isotype distribution and clonal expansion within the BCR repertoire, the number of cells for each isotype across the samples was analyzed. The clonal expansion profiles reveal the distribution of cells per clone, their corresponding isotypes, and the degree of expansion for each clone (*Figure 8*). In all samples, the majority of BCR transcripts belong to the IgM isotype, indicating its dominance within the BCR repertoire. Additionally, a substantial portion of plasma cells have undergone class switching, resulting in transcripts for IgA and IgG, present in approximately equal proportions. As expected, transcripts for IgD and IgE are sparse, reflecting their specialized roles: IgD primarily in its membrane-bound form and IgE in allergic responses or immune defenses against helminths.

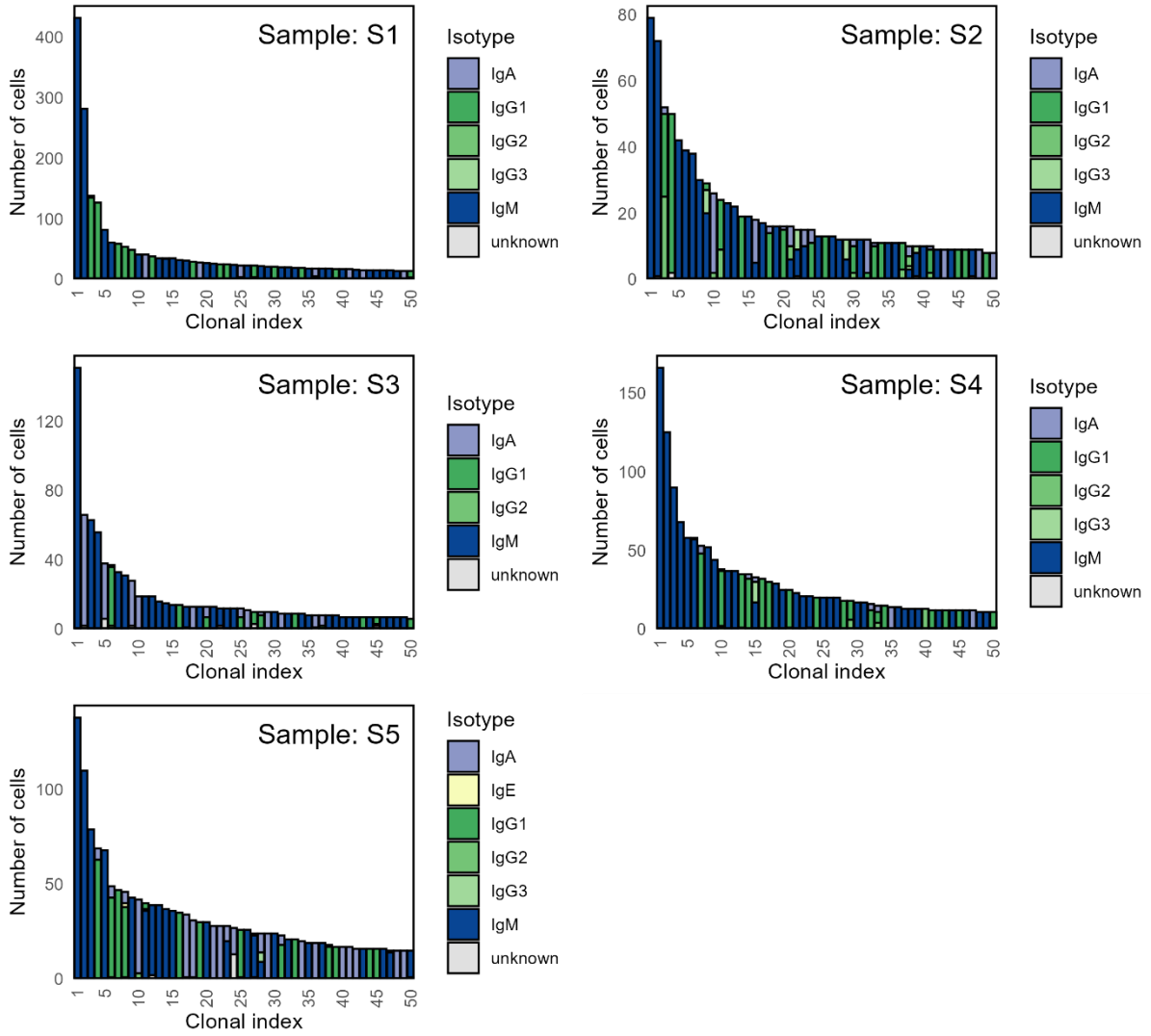


Fig. 8 Clonal expansion profiles
Clonal expansion profiles for the top 50 most expanded clonotypes across all five samples. The horizontal axis represents the clonal index, where a lower clonal index indicates a higher degree of expansion for a clonotype.

IGHV7-1 and IGHV11-2 combine with IGHJ1 in VDJ chains across all samples

Identifying patterns of germline V and J gene usage in BCR repertoires is crucial for understanding the genetic architecture that contribute to effective immune responses. By analyzing these gene pairings, insights can be gained into the genetic characteristics that are selected for in the immune repertoire. The VDJ dataframe generated by the AntibodyForests package allows for exploration of germline V and J gene usage across the B cell repertoire. D genes were excluded from the analysis due to annotation challenges stemming from their short length and high diversity.

In VDJ chains, specific V and J gene pairings, particularly IGHV7-1 and IGHV11-2 with IGHJ1, were consistently observed across all samples (*Figure 9A*). These frequent pairings suggest a selective advantage or structural compatibility, underlining their importance in the humoral immune response of the studied mice. In contrast, the VJ chains displayed greater variability in V-J combinations (*Figure 9B*), reflecting a broader range of pairings that may be necessary for functional flexibility. This diversity indicates different selective pressures or functional requirements for VJ chains compared to the more conserved VDJ chain pairings.

Fig. 9A Germline V and J germline gene usage in VDJ chains
Quantification of observed V-J combinations in the VDJ across the five samples. The red boxes highlight two frequently observed V-J pairings in the VDJ chain.

Fig. 9B Germline V and J germline gene usage in VJ chains
Quantification of observed V-J combinations in the VJ chains across the five samples.

Number of public CDR3 sequences is comparable to number of public VDJ sequences

Beyond examining shared combinations of germline V and J gene usage, it also important to investigate the presence of shared sequences, known as public sequences. These sequences, found across multiple samples, offer deeper insights into conserved immune responses, potentially uncovering common patterns in B cell selection and evolution following specific antigen challenges. First, public VDJ chain CDR3 sequences were quantified (*Figure 10A*). The CDR3 region, being the most variable part of an antibody sequence, plays a critical role in determining antigen specificity. A total of 6 public CDR3 sequences were identified across the samples. The analysis was then extended to quantify shared full-length trimmed VDJ sequences, which incorporate the CDR3 region along with the flanking V, D, and J segments, revealing 5 public trimmed VDJ sequences (*Figure 10B*). This close match between the number of public CDR3 and trimmed VDJ sequences suggests that the antigen-binding regions are tightly linked to the overall VDJ sequence, reflecting a consistent selection of these sequences during the immune response.

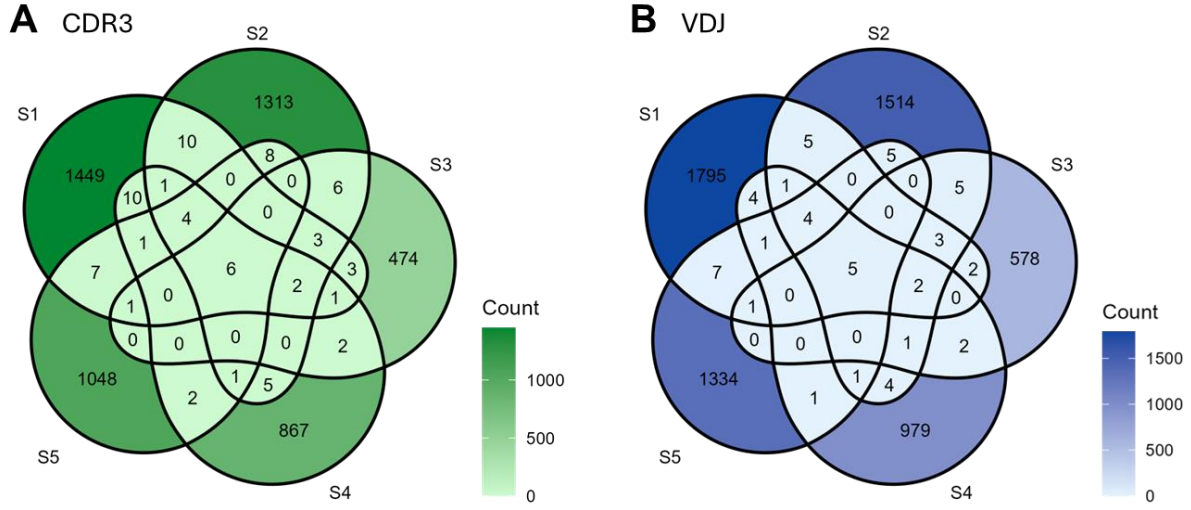


Fig. 10 Number of shared CDR3 and trimmed VDJ sequences
Quantification of shared CDR3 sequences and trimmed VDJ sequences across the five samples.

About 50% of Cell Ranger's VH gene annotations do not match IgBLAST output

In the course of developing the AntibodyForests package, accurate annotation of immunoglobulin sequences was critical for grouping cells into clonotypes, inferring phylogenetic relationships, and identifying SHMs. Given the importance of accurate annotations, we sought to re-annotate the sequences in the VDJ dataframe using the IgBLAST tool with the latest mouse IMGT/V-QUEST Ig reference directory set [96][92]. This step was necessary to align the sequences with the IMGT numbering scheme, which standardizes the positions of conserved amino acids across sequences, ensuring compatibility with codon substitution model specific to antibody lineages (such as IgPhyML) [93].

Upon comparing the re-annotated sequences with the original annotations provided by the Cell Ranger pipeline, a significant discrepancy was observed: approximately 50% of the VDJ chain V gene annotations did not match between the two methods (*Figure 11*). In contrast, the other gene annotations, including those for VDJ J genes, VJ V genes, and VJ J genes, showed much higher concordance. This highlights the importance of choosing an appropriate reference directory set during processing of the sequencing data and annotation, and underscores the critical impact on subsequent BCR repertoire analyses.

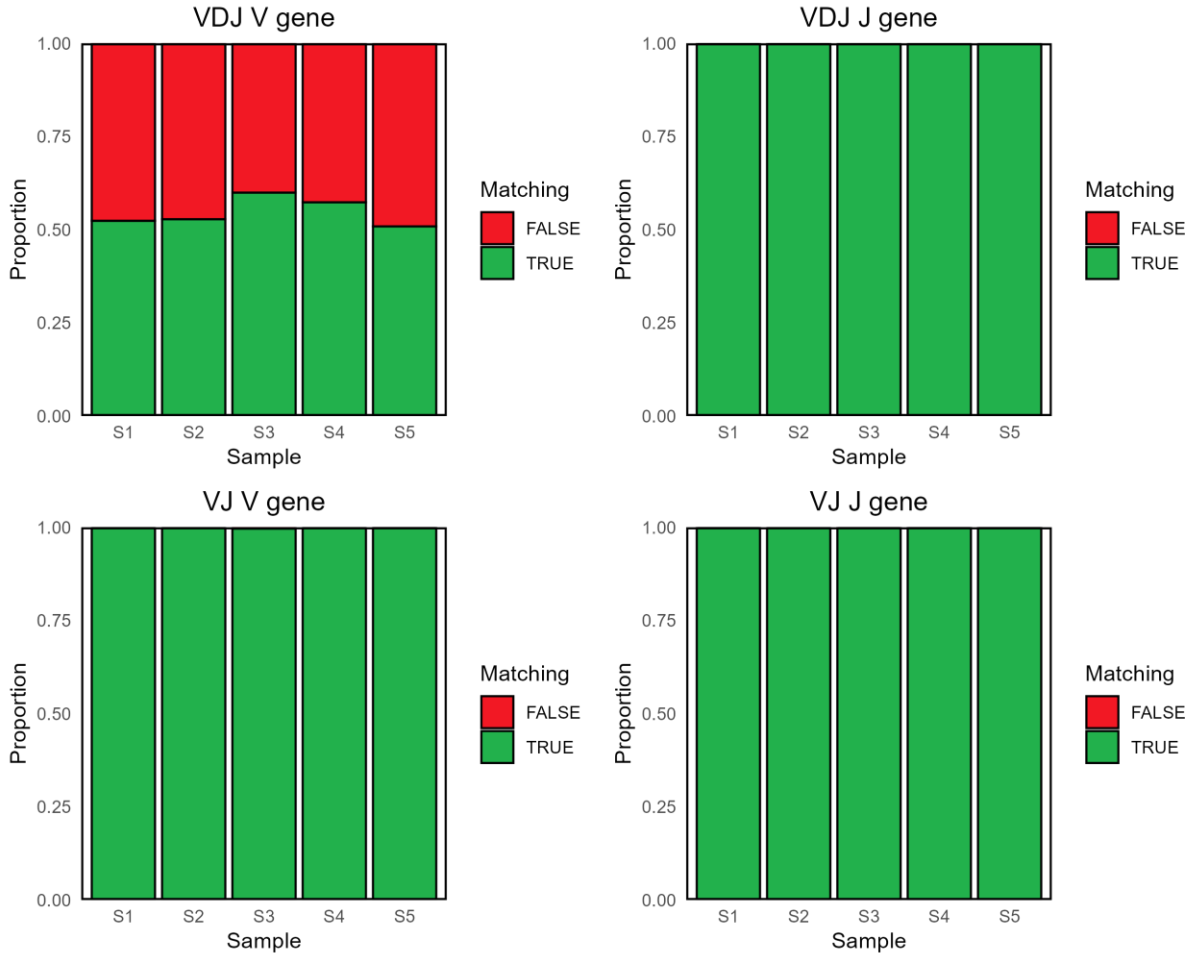


Fig. 11 Matching V and J annotations: Cell Ranger vs IgBLAST
Proportions reflect number of cells in VDJ data frame with matching annotations from Cell Ranger and IgBLAST.

Choice of construction methods impacts inference of evolutionary relationships

After conducting initial repertoire analyses on the VDJ dataframe, we proceeded to apply the `AntibodyForests()` function to infer lineage trees using all available construction methods. This step allows the evaluation how different algorithms impact the inference of B cell evolutionary trajectories. Major discrepancies in tree structure were observed between those created with different construction methods (*Figure 12*). These differences can significantly influence the interpretation of clonal evolution. To quantify these differences, the proportion of shared edges between pairs of trees were examined, whereby a shared edge is defined as a connection between the two same nodes/sequences in both trees. The default, MST, and NJ algorithms show high similarity scores among themselves, as do the MP, ML, and codon substitution model (*Figure 13*). The IgPhyML tool exhibits moderate to low similarity with the other methods. These findings underscore that different algorithms infer different evolutionary relationships between B cells in the lineage trees, emphasizing the importance of substantiating the choice of algorithm, as it could significantly affects subsequent analyses of B cell evolution trajectories.

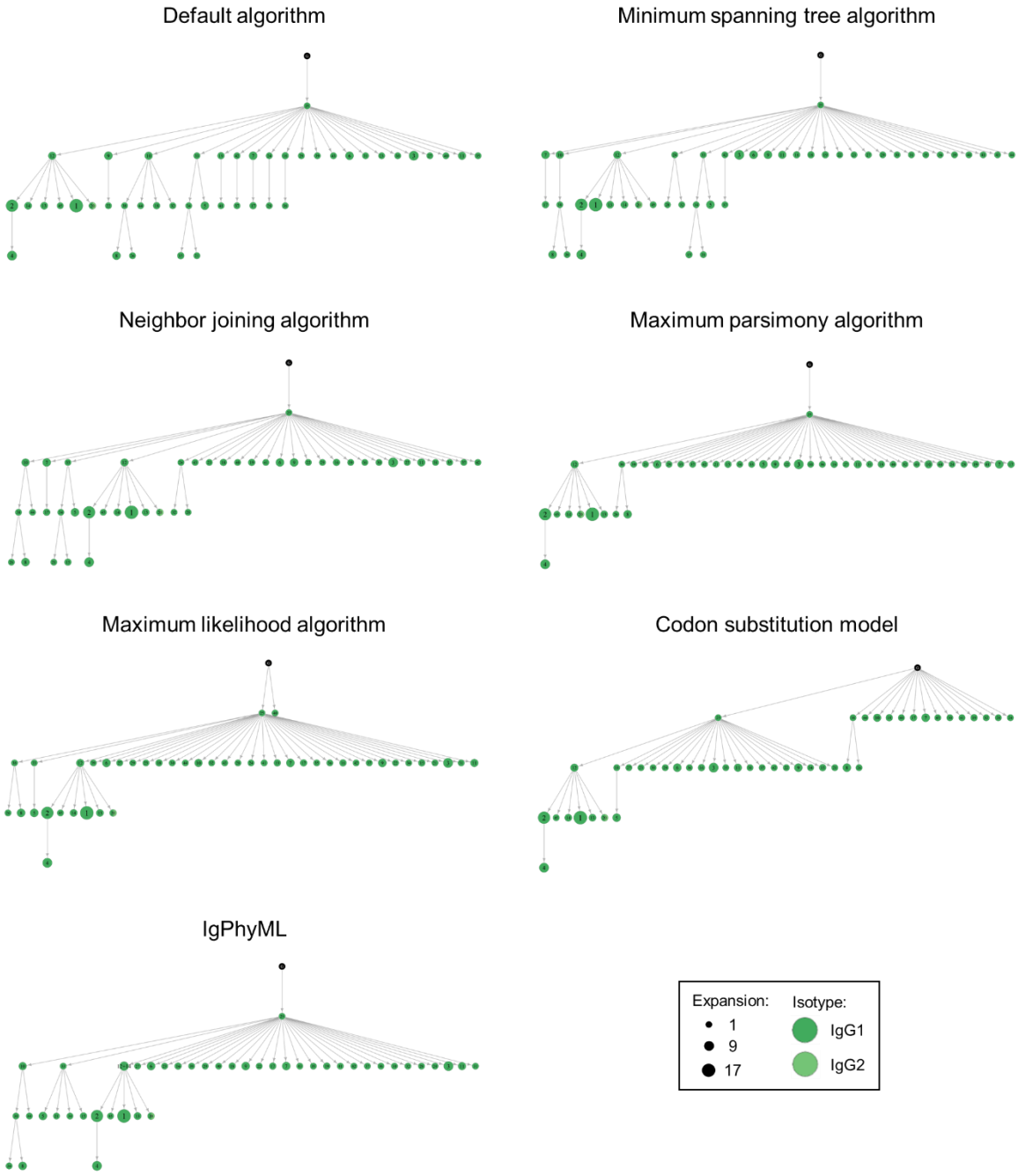


Fig. 12 Example trees from different construction methods in AntibodyForests()
Trees show the evolution of clonotype 4 from sample 1. Trees are inferred using the trimmed VDJ and VJ sequences and their trimmed germline sequences, utilizing the construction methods indicated. Node labels represent the ID from the AntibodyForests object, with lower IDs indicating greater expansion. Node colors indicate isotype, and node size reflects the level of expansion.

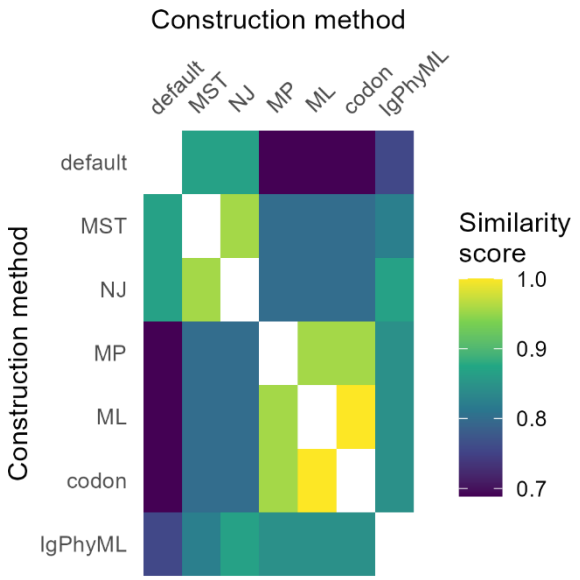


Fig. 13 Similarity of trees of different construction methods Similarity scores are shown of evolutionary trees of clonotype 4 of sample 1, inferred with seven different construction methods as indicated. The similarity scores represent the proportion of shared edges (where two identical nodes, representing identical sequences, are connected) between pairs of trees.

Integration approaches lack comprehensive coverage of all single-cell clones

After inferring the lineage trees for the clones in the single-cell dataset, we applied various integration approaches to incorporate the bulk transcripts into these clones. By complementing the single-cell data with bulk sequencing data, we aimed to enhance the resolution of the inferred evolutionary pathways. The samples contained approximately 500,000 spleen-derived bulk transcripts, with sample 3 notably exceeding this count, containing over 1,000,000 transcripts (*Figure 14A*). Across all samples, the majority of these transcripts were from the IgM and IgG isotypes. To ensure consistency with the single-cell data and enable accurate mapping of bulk sequences to single-cell clones, both the bulk and single-cell sequences were reannotated using the IgBLAST tool with the mouse IMGT V, D, and J reference germline sequences [92][93]. After reannotation, the total number of transcripts for which a single V, a single J, and the CDR3 region could be reliably identified was slightly lower than the initial transcript counts (*Figure 14B*).

Following reannotation, all six integration approaches were employed to incorporate the re-annotated bulk transcripts into the re-annotated single-cell clones. The proportion of transcripts successfully integrated varied between 33.8% and 69.4%, depending on the selection criteria and filtering steps used, with all isotypes being proportionally represented *(Figure 15*). To evaluate the distribution of integrated transcripts across the single-cell clones and to assess the coverage of these clones, the number of integrated bulk transcripts was plotted against the clonal expansion profiles for each integration method (*Figure 16*). For all integration approaches, some single-cell clones did not have any corresponding bulk transcripts integrated. This incomplete integration suggests that the current approaches may lack comprehensive coverage of all single-cell clones, potentially introducing biases in the representation of the B cell repertoire. These findings underscore the limitations of the existing methods and the need for further refinement to achieve more accurate and complete integration of bulk and single-cell datasets.

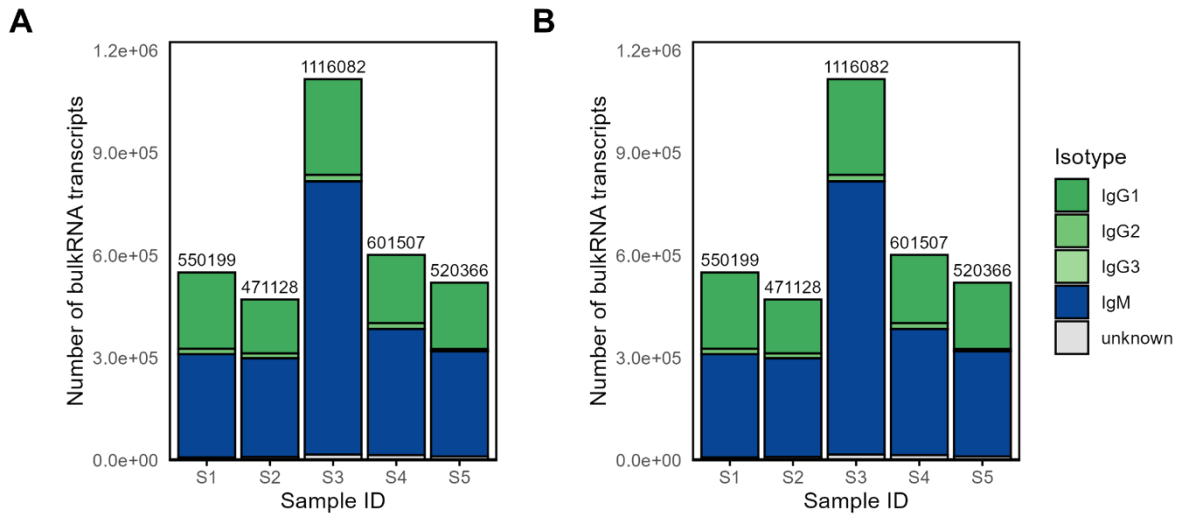


Fig. 14 Number of bulk transcripts before and after annotation with IgBLAST and IMGT reference database
(A) Total number of spleen-derived bulk transcripts and their isotype per sample before annotation. The numbers on top of the bars indicate the total number of bulk transcripts per sample.
(B) Total number of transcripts for which a single V, a single J, and the CDR3 region could be determined after annotation with IgBLAST using the mouse IMGT V, D, J reference germline sequences.

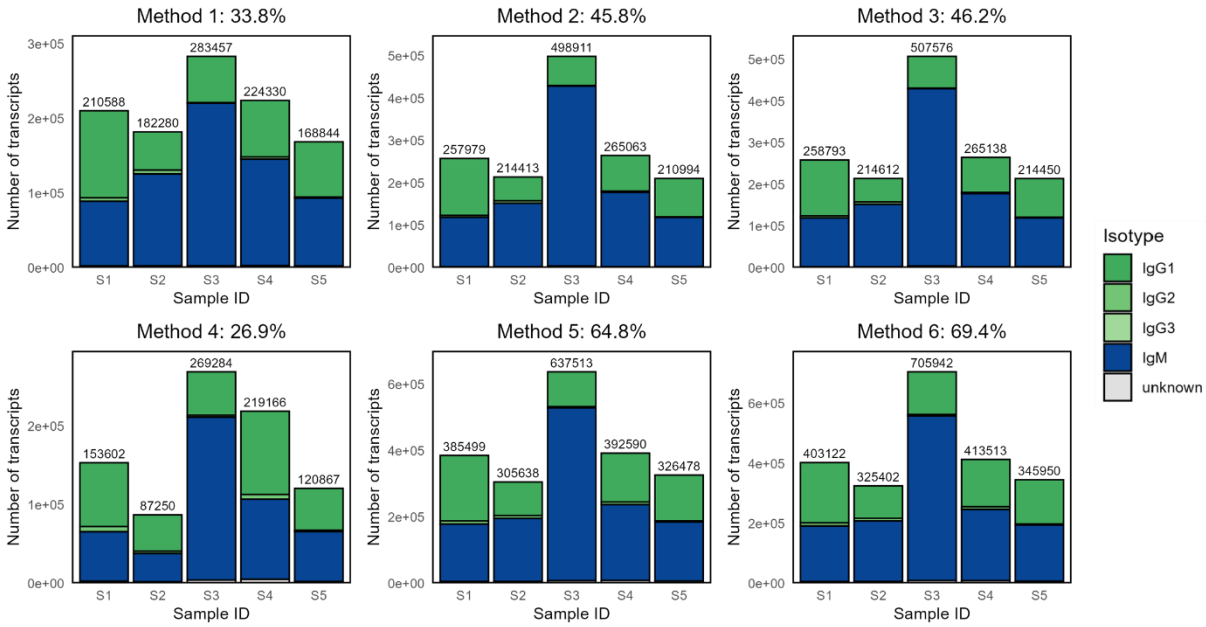


Fig. 15 Proportion of bulk transcripts that could be integrated into single-cell clones per method and per sample
The bar plots show the number of bulk transcripts and their isotypes integrated into single-cell clones across the five samples using the six different integration methods (presented in *Figure 16*). The percentage at the top of each plot indicates the proportion of bulk transcripts successfully integrated into single-cell clones for that specific method.

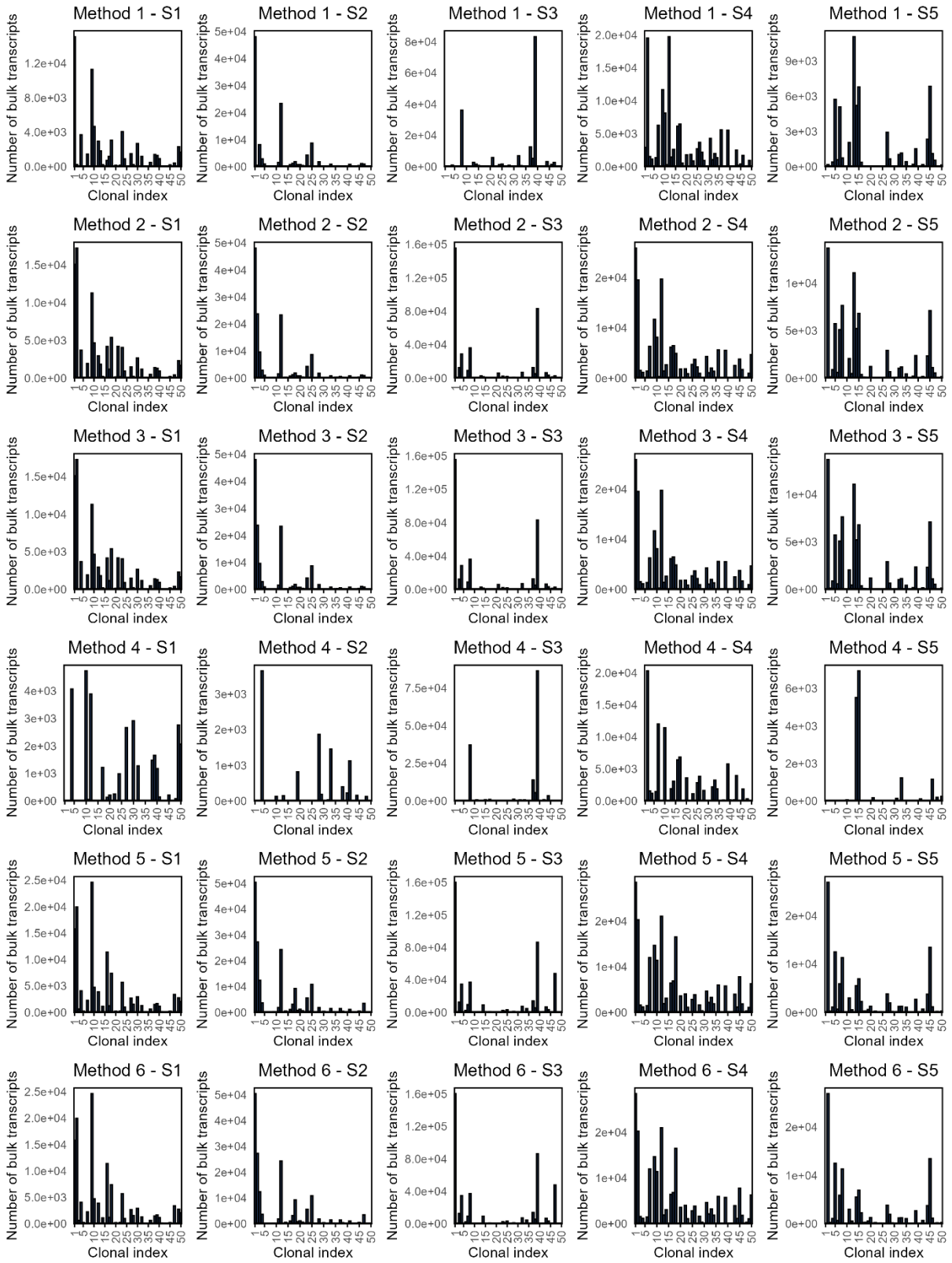


Fig. 16 Number of integrated bulk transcripts per single-cell clone across different samples and methods
The bar plots show the number of bulk transcripts integrated into the single-cell clones for all five samples and all six integration approaches. The clones are ordered by clonal index from the single-cell dataset, with the most expanded clones positioned on the left of each plot. This visualization enables comparison of the effectiveness of each integration method across different samples, highlighting the variability in coverage of single-cell clones.

Approximately 0.6% of bulk transcripts show VDJ sequence overlap with single-cell seqs

To assess the overlap between sequences from the spleen-derived bulk dataset and the bone marrow-derived single-cell dataset, the number of unique CDR3 sequences present in both datasets was identified. The analysis revealed that approximately 1.7% of all unique CDR3 sequences in the bulk dataset were also found in the single-cell dataset, with the number of unique shared CDR3 sequences varying between 267 and 790 across samples (*Figure 17A*). Next, the analysis was extended to determine the number of shared VDJ sequences present in both datasets. Due to the trimming of the FWR1 regions in the bulk dataset, which was necessary because of the choice of sequence adapters, all sequences required uniform trimming. To achieve this, the FWR1 regions from the sequences in integration method 1 across all clones were aligned to establish a consistent trim point, ensuring equal FWR1 lengths across sequences. Following this standardization, approximately 0.6% of all unique trimmed VDJ sequences in the bulk dataset were found to overlap with those in the single-cell dataset, with the number of shared VDJ sequences varying between 127 and 480 across samples, and IgM transcripts being predominantly represented among the shared sequences (*Figure 17B*). These findings suggest that spleen-derived bulk transcripts exhibit higher sequence diversity, and that transcripts from different tissue sources, such as spleen and bone marrow, may not extensively overlap. This indicates the compartmentalization of the B cell repertoire and the importance of integrating multiple data sources for a more complete understanding of BCR diversity and B cell evolution.

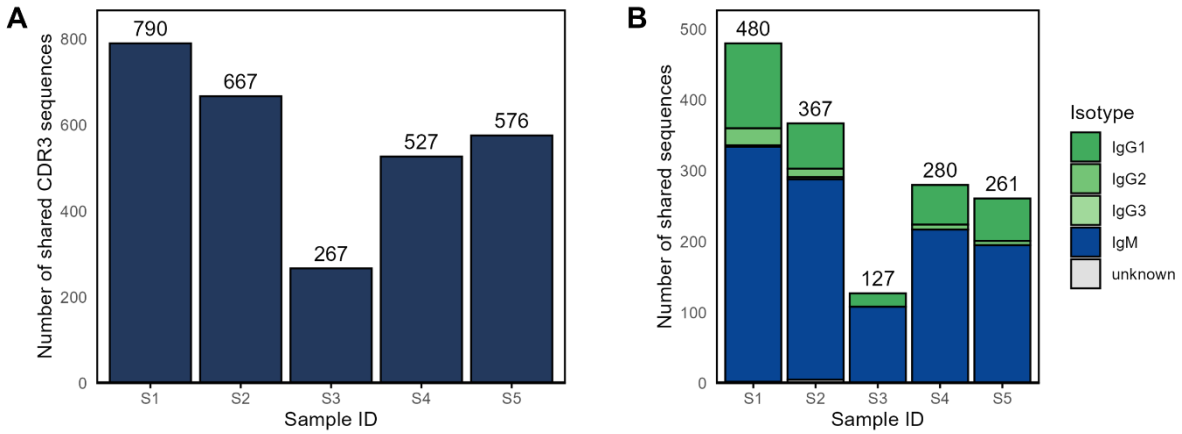


Fig. 17 Number of shared CDR3 and VDJ regions
(A) Number of unique shared CDR3 regions per sample, representing CDR3 sequences found in both the bulk and single-cell datasets. The numbers on the bars indicate the count of shared CDR3 sequences per sample.
(B) Number of unique shared trimmed VDJ sequences per sample, with the total count per sample displayed on top of each bar.

Almost all single-cell clones have CDR3 regions found in bulk dataset

After determining the proportion of CDR3 sequences from the bulk dataset that were also found in the single-cell dataset, the analysis was reversed to see which single-cell clones contained CDR3 sequences that were also present in the bulk dataset. The results reveal that almost all top-expanded single-cell clones contained shared CDR3 sequences (*Figure 18*). This significant overlap suggests that the most prevalent B cell clones in the bone marrow are also well-represented in the spleen-derived bulk dataset. The CDR3 regions are known for their high diversity and serve as key indicators of a clone's specificity. The robust presence of shared CDR3 regions in the top-expanded clones across both datasets underscores their importance as markers for identifying expanded clones across different tissue sources.

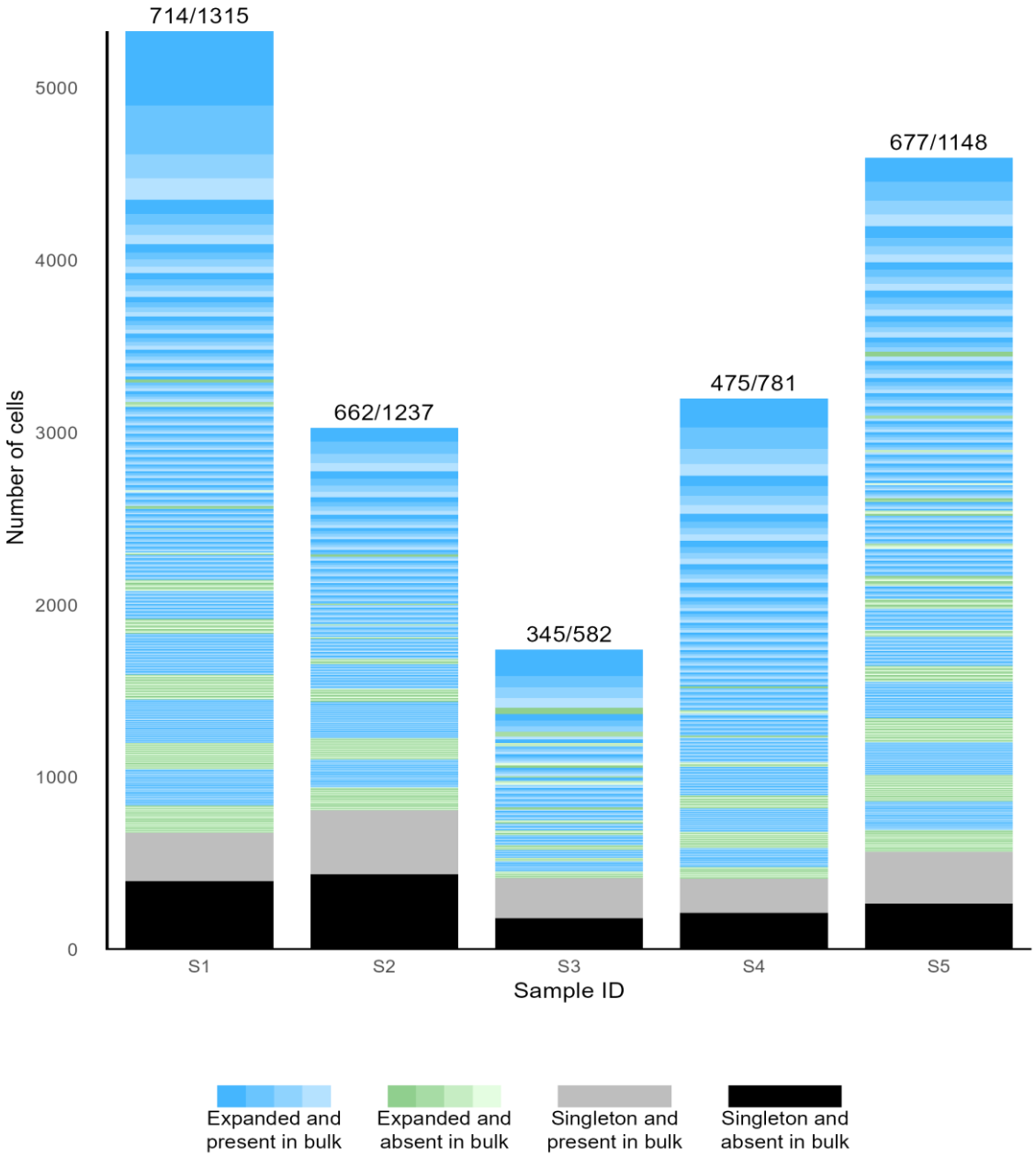


Fig. 18 Single cell clones with shared CDR3 regions
The height of the bars reflects the number of cells per sample from the bone marrow-derived plasma cells of the single-cell dataset. The cells are colored based on whether the single-cell clone contained a CDR3 sequence that was also found in the bulk dataset. Different tints of colors are used to distinguish between different clones. The numbers on top of each bar represent the number of clones with shared CDR3 sequences relative to the total number of clones in that sample.

Bulk transcript and single cell counts with shared sequences show positive correlation

To investigate the congruence of the spleen-derived bulk dataset and the bone marrow-derived single cell dataset, the relationship between the number of cells in the single cell dataset and the number of transcripts in the bulk dataset with shared CDR3 sequences was examined. The number of transcripts from the spleen-derived bulk dataset having a shared CDR3 region was positively correlated with the number of cells having the same shared CDR3 region from the bone marrow-derived single-cell dataset (*Figure 19A*). This positive correlation suggests a significant overlap between the two datasets, indicating that the most abundant clones in the single-cell dataset are also well-represented in the bulk dataset. The analysis was subsequently extended to examine the shared VDJ sequences. It was found that IgM transcripts also demonstrated a positive correlation with the number of cells having the same shared VDJ sequences and the same isotype (*Figure 19B*). However, a similar correlation was not observed for the IgG subisotypes. This finding further supports the congruence between the spleen-derived bulk dataset and the bone marrow-derived single-cell dataset.

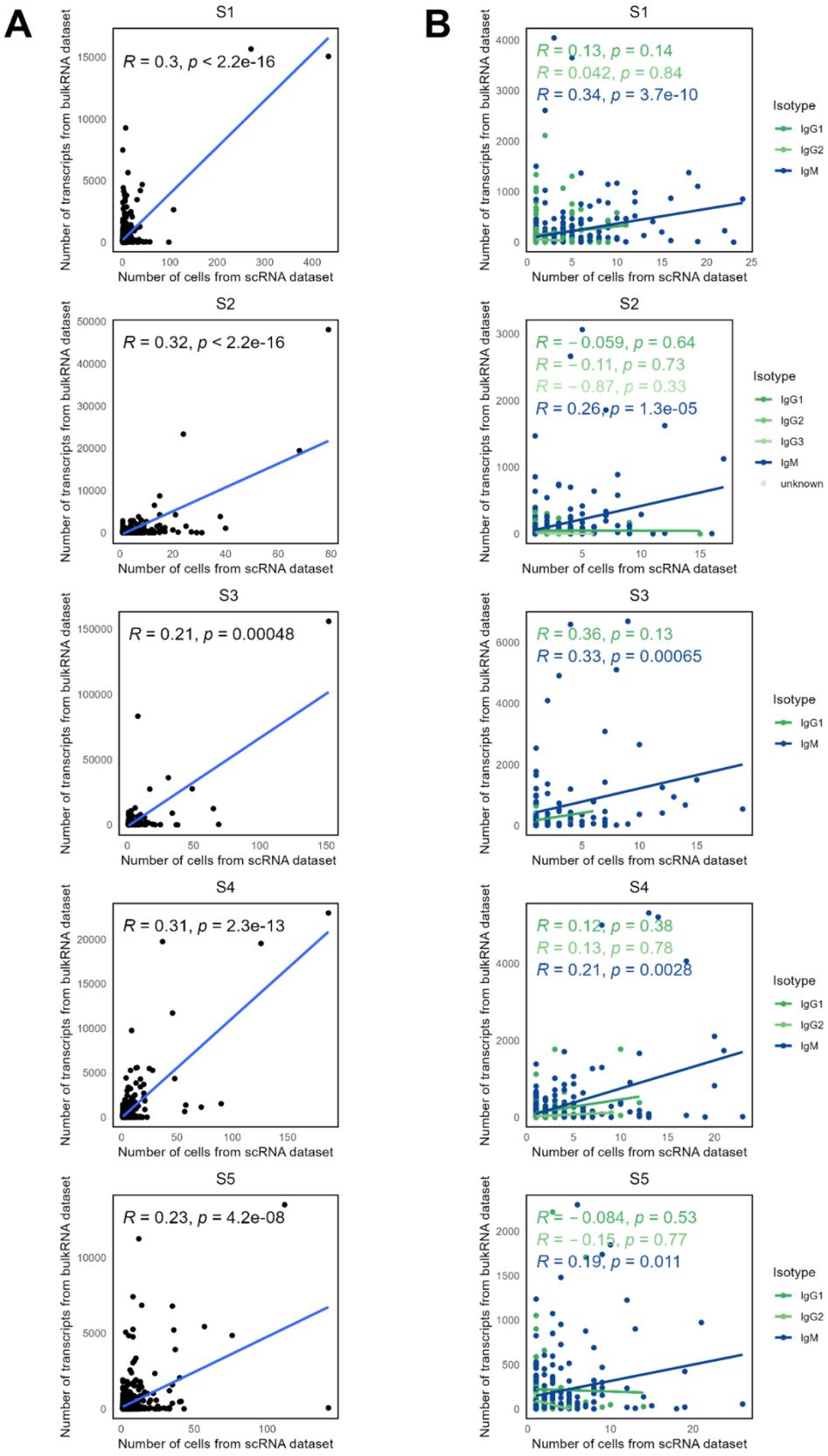


Fig. 19 Relation between number of spleen-derived transcripts and number of bone marrow-derived plasma cells
(A) Correlation between the number of bone marrow-derived plasma cells and spleen-derived transcripts with shared CDR3 sequences, with Spearman's correlation coefficients and p-values for each sample.
(B) Correlation between the number of bone marrow-derived plasma cells and the number of spleen-derived transcripts with shared VDJ sequences, with Spearman's correlation coefficients and p-values provided for each isotype.

Bulk data complements single-cell data, enhancing resolution in B cell clonal evolution

Finally, to assess how the integration of spleen-derived bulk transcripts into clones from bone marrow-derived plasma cells affects tree topology, an example tree was constructed using the default algorithm on clonotype 4 of sample 1 (*Figure 20*). It includes 127 cells from the bone marrow and 3,700 transcripts from the spleen, resulting in a total of 302 nodes (excluding the germline node). Among these, 18 nodes represent sequences that were present in both datasets. When examining the tree topology, it is evident that nodes derived from bulk transcripts and single cells alternate and are evenly distributed throughout the tree. This even distribution and overlap suggest that the integration of bulk transcripts does not disrupt the overall structure of the inferred evolutionary pathways. Instead, it complements the single-cell data, providing a more comprehensive view of the clonal evolution within the B cell population. This indicates that integrating spleen-derived bulk data with bone marrow-derived single-cell data can enhance the resolution and accuracy of B cell lineage tracing, leading to more robust insights into B cell evolution.

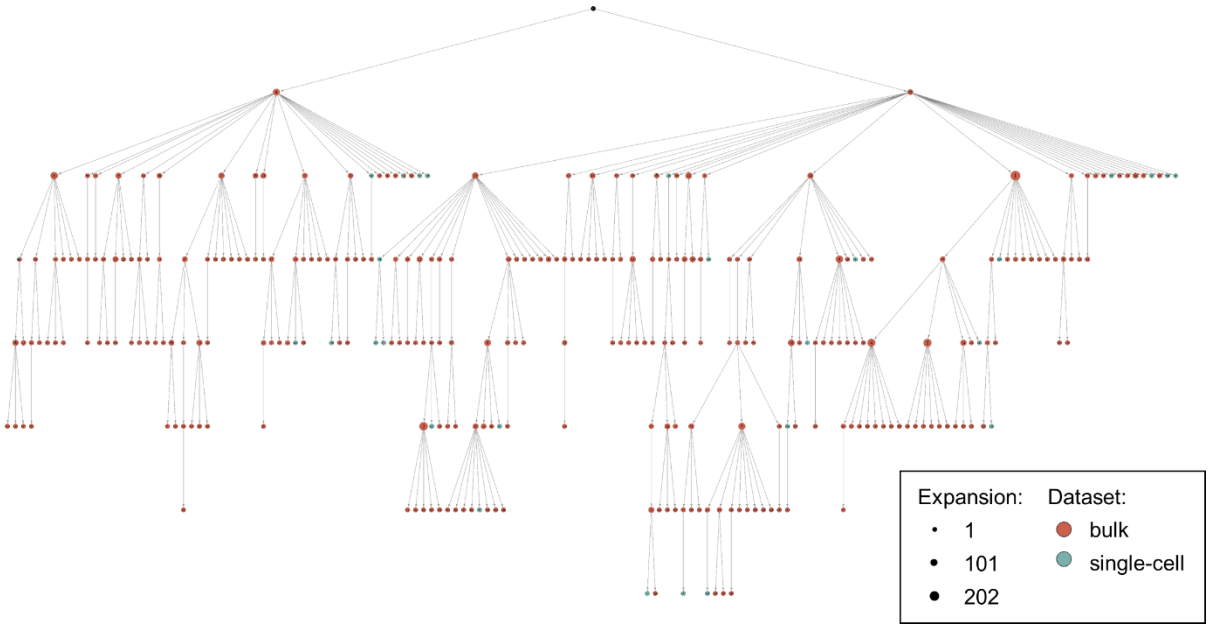


Fig. 20 Lineage tree of clonotype 4 from sample 1 comprising single cell data and bulk data
The lineage tree is constructed with the default construction method and parameter settings and is based on the trimmed VDJ chain sequences. Nodes are colored according to their dataset of origin (bone marrow-derived single cells or spleen-derived transcripts) and sized based on clonal expansion (number of cells or number of transcripts).

Discussion

The primary aim of this project was to develop AntibodyForests, an R package designed to facilitate the study of B cell selection and evolution by integrating single-cell sequencing data from the 10x Genomics platform. The package successfully imports VDJ and VJ sequence data, infers lineage trees for observed clonotypes, and enables downstream analyses that provide insights into the evolutionary trajectories of B cell repertoires. These capabilities were applied to a dataset from five BALB/c mice immunized with the model antigen ovalbumin, focusing on bone marrow-derived plasma cells. A secondary objective was to address the issue of undersampling inherent in single-cell sequencing by integrating bulk transcriptomic data derived from the spleens of the same mice. Although the integration approaches require further refinement to achieve comprehensive coverage of all clones, our findings indicate significant congruence between the single-cell and bulk datasets. This congruence is evidenced by the presence of shared sequences and a positive correlation between the number of cells and the number of transcripts, suggesting that the bulk data effectively complement the single-cell data.

However, the study has several biological limitations. Firstly, the presented analysis of the datasets from the five immunized mice is primarily focused on long-lived plasma cells, which originate from the follicular B cell subset and are derived from bone marrow, specifically from the femurs and tibias. While follicular B cells represent a significant component of the adaptive immune response, they do not encompass the entire spectrum of B cell activity. Other subsets, such as non-activated mature B cells, marginal zone B cells, and memory B cells, were not included in this study. Memory B cells, in particular, play a crucial role in long-term immunity, and their exclusion from the analysis limits our understanding of the full dynamics of B cell evolution. Moreover, the integration of spleen-derived transcripts, while improving coverage, still captures only a small fraction of the total B cell repertoire. This underrepresentation impacts the ability to fully track evolutionary trajectories and identify patterns that may suggest deterministic rules governing B cell selection and expansion.

Secondly, the current mouse model does not account for several factors that significantly influence B cell selection and evolution. One such factor is the role of $CD4^+$ helper T cells, which are essential for B cell activation and differentiation upon exposure to protein antigens. The interaction between B cells and helper T cells is critical for initiating the GC reaction, where B cells undergo SHM and class switching. The diversity of T cell receptors (TCRs) and the specificity of their interactions with antigen-presenting cells further add complexity to the B cell activation process. Since T cell activation precedes and influences B cell activation, the repertoire of TCRs indirectly shapes the B cell response. The lack of consideration for these interactions in the current study represents a significant approximation of the actual processes involved in B cell evolution. Integrating T cell receptor sequencing data would allow the exploration of clonal expansion within both TCR and BCR repertoires, potentially uncovering correlations that clarify how specific antigenic determinants drive the selection and activation of both T and B cell clones.

The antigen presentation process, particularly the role of MHCII molecules, is another factor that influences B cell evolution. B cells must present processed antigen fragments on MHCII molecules to engage helper T cells effectively. Different MHC alleles exhibit varying efficiencies in antigen presentation, which in turn affects the activation and selection of B cells. The variability in MHC alleles across different individuals or strains could lead to different evolutionary outcomes in B cell populations, further complicating our understanding of B cell selection and expansion. Integrating MHC genotyping data could enable the stratification of B cell repertoire analyses based on MHC allele types. This would allow for a more nuanced

understanding of how different MHC alleles influence B cell evolution, providing insights into the differential immune responses observed across diverse genetic backgrounds.

The mice used in this study were housed under specific pathogen-free (SPF) conditions, which minimizes exposure to known pathogens and reduces background immune responses. However, this controlled environment does not reflect the complexity of the immune system in more natural settings, where individuals, including humans and animals, are exposed to a wide array of antigens over time. In such environments, prior exposures to pathogens or vaccinations generate long-lived plasma cells and memory B cells that can influence subsequent immune responses. These pre-existing immune cells can respond more rapidly and robustly to antigens that share structural similarities with previously encountered pathogens, leading to a more complex and dynamic B cell repertoire. The absence of this diverse microbial exposure in SPF conditions likely resulted in a less diverse B cell repertoire in the study, which introduces a layer of complexity to B cell evolution that is not captured in this model. Consequently, the current study may not fully reflect real-life scenarios where the B cell repertoire is shaped by a lifetime of antigenic exposure, underscoring the importance of understanding how environmental factors shape B cell evolution when translating findings from model organisms to humans.

Additionally, while mice serve as valuable models for studying human immune responses, there are notable differences between the two species that must be considered. One such difference is the usage of IGK and IGL light chains. Mice tend to have a higher proportion of B cells expressing IGK light chain with a ratio of about 90:10, whereas the ratio of IGK to IGL light chains in humans is close to 60:40. This difference suggests variations in receptor editing during B cell development, which could impact the final B cell repertoire. If more B cells undergo recombination at the *IGL* locus, it could result in greater diversity in the germline usage of Ig light chain V and J genes. Moreover, the anatomical differences between mice and humans, such as the size and structure of lymphoid organs, and the differences in lifespan, likely influence the generation and maintenance of humoral immune responses. These species-specific differences are critical to consider when extrapolating findings from mouse models to human biology.

Another significant challenge in this study is the lack of a unified reference database with accurate germline sequences specific to BALB/c mice. The initial annotations were performed using a reference database for C57BL/6 mice, another commonly used strain in immunological research. However, there are known differences between C57BL/6 and BALB/c mice, such as the number and type of productive germline alleles per gene cluster [97]. Using a reference database that does not fully match the genetic background of the studied mice can lead to inaccuracies in the identification of SHMs, which are crucial for studying B cell evolution. This mismatch could introduce biases in the analysis, affecting the interpretation of how B cells evolve in response to antigenic stimulation.

When reviewing the software development aspect of this project, AntibodyForests package currently focuses on analyzing sequence data to infer evolutionary relationships within B cell populations. While this provides valuable insights, it does not capture the full picture of B cell evolution, which is also influenced by phenotypic characteristics such as antibody binding affinity and specificity. Integrating phenotypic data, such as affinity binding measurements and epitope binning results, into the inferred lineage trees could enhance our understanding of how B cell clones evolve not just in sequence but also in function. Such integration would allow researchers to track the evolution of antibody specificity and affinity across different stages of the immune response. However, technical limitations currently restrict the ability to phenotypically characterize a large number of plasma cells, limiting the scope of such analyses.

To address these limitations, recent advances in computational biology, particularly in protein language models, offer promising avenues for extending the capabilities of

AntibodyForests. These models, which are trained on extensive databases of protein sequences, can predict evolutionary likelihoods and even estimate protein structures based on sequence data alone. These protein language models have recently been used to predict evolutionary plausible mutation for *in silico* antibody maturation, achieving up to a 16-fold improvement in binding affinity, thereby demonstrating their potential in advancing our understanding of efficient and productive antibody evolution [98]. By incorporating these *in silico* predictions into the AntibodyForests package, it would be possible to predict key phenotypic characteristics, such as binding affinity or specificity, for a broader set of antibodies within a repertoire based solely on their sequence. This capability would allow researchers to estimate how changes at the genetic level translate into functional outcomes, enabling a more comprehensive study of how antibodies evolve in terms of both sequence and function, and providing deeper insights into the mechanisms driving B cell evolution.

While AntibodyForests represents a significant advancement in the analysis tools for B cell evolution, the software itself has several limitations that must be acknowledged. One notable limitation is that not all models can be directly run within the AntibodyForests package. For instance, inferring lineage trees using the IgPhyML tool requires users to generate input files through specific functions, which must then be processed via the IgPhyML command line tool. This adds complexity to the workflow and may hinder accessibility for users who are less familiar with command line operations. Additionally, the various tree inference methods implemented in `AntibodyForests()` function can produce different evolutionary trees, reflecting the inherent variability in how these algorithms interpret sequence data. While this variability underscores the importance of choosing the appropriate method, it also highlights a need for a strategy to integrate these differing outputs into a consensus or summary tree, which would provide a more unified view of the inferred evolutionary relationships.

Moreover, the current repertoire analyses in AntibodyForests are primarily focused on sequence-related features, such as germline gene usage and sequence evolution. Although evolutionary likelihood scores can be plotted on lineage trees, there remains a lack of robust methods to quantify how these scores or predicted structures evolve over time. This limitation restricts the package's ability to provide a comprehensive analysis of B cell evolution, particularly in terms of how functional characteristics like binding affinity might change as a result of SHM. Although some functions are under development to calculate tree topology metrics across all clonotypes and visualize these metrics using multidimensional reduction algorithms, these methods do not yet account for the individual relationships between nodes within the trees. This gap highlights an area where future development could significantly enhance the analytical power of AntibodyForests.

In summary, the development of AntibodyForests represents a significant advancement in the study of B cell selection and evolution. The package provides a valuable tool for analyzing B cell repertoires, particularly when integrated with bulk sequencing data, offering a more complete picture of B cell evolution. However, the study has several limitations, including the focus on a specific subset of B cells, the exclusion of important immunological factors such as T cell interactions and MHC variability, and the challenges associated with species-specific differences and reference biases. Future developments should aim to expand the coverage of B cell subsets, integrate phenotypic data, and refine the methods for integrating bulk and single-cell data. Additionally, incorporating computational predictions of antibody structure and function could provide deeper insights into the evolutionary dynamics of B cell repertoires. Addressing these challenges will enhance our understanding of B cell evolution and improve our ability to track and predict immune responses in both experimental and real-world settings.

Materials and methods

Data acquisition

The data used in this project were provided by the Sai Reddy Lab at the Laboratory for Systems and Synthetic Immunology of ETH Zurich. The datasets utilized for B cell repertoire analysis were derived from previous studies conducted by Neumeier *et al*. (2022) and Khan *et a*l. (2016) [77][94]. These datasets comprised single-cell B cell sequencing data from bone marrow-derived plasma cells of BALB/c mice and bulk transcriptomic data from spleen-derived RNA of the same mice. The single-cell sequencing data was processed with CellRanger version 7.1.0 and annotated with the mouse C57BL/6 V(D)J reference GRCm38_alts_ensembl-7.0.0. The bulk sequencing data was processed by importing it into CLC Genomics Workbench 7, where a custom pipeline was used to correct for sequencing errors and PCR amplification bias through molecular amplification fingerprinting, which is freely available via the "MAF" static repository (https://sissource.ethz.ch/sis/maf_scripts).

Computational environment

The development, testing, and implementation of functions for the AntibodyForests package were conducted using R version 4.3.2 within the RStudio environment (version 2024.04.01+748).

High-performance computing

The computationally intensive scripts, particularly those involving large-scale data processing and phylogenetic tree construction, were executed on the High Performance Computing (HPC) facility of the University Medical Center Utrecht (UMCU). This facility is part of the Utrecht Bioinformatics Center (UBC) and provides the necessary computational resources to handle the extensive datasets and complex analyses required for this project.

IgBLAST and IgPhyML installation and configuration

IgPhyML and IgBLAST were installed on the HPC within a Conda environment created with Python 3.9, which included essential dependencies such as Autoconf, Automake, Change-O, Alakazam, and an OpenMP-enabled C compiler. IgPhyML was cloned and compiled with OpenMP support, following instructions from the official IgPhyML documentation website [99]. IgBLAST version 1.22.0 was downloaded, extracted, and configured within the environment. Additionally, a collection of scripts from the Immcantation repository was employed to download the IMGT reference sequences and build the IgBLAST database, by following the guidelines from the official Change-O documentation website [100].

Data analysis and figure plotting

All data visualizations were performed in the R environment. The clonal expansion profiles were created with the `VDJ_plot_clonal_expansion()` function; the germline gene usage heatmaps were generated with the `VDJ_plot_gene_usage()` function; the Venn diagrams depicting the number of public sequences were plotted with the `VDJ_plot_public_seqs()` function; and lineage trees were visualized with the `AntibodyForests_plot()` function. All other plots were generated using ggplot2 (v3.5.1). Plots without panel labels were assembled with ggpubr (v0.6.0), while plots with panel labels were arranged in Microsoft Powerpoint (Version 2407, build 178030.20138). The code used to generate these figures is available in the supplementary materials (S3: R Markdown report).

Acknowledgements

I would like to thank my supervisor, Dr. Alex Yermanos, for the opportunity to join his group and for his feedback throughout the project. I also appreciate the guidance and contributions of Daphne van Ginneken, M.Sc., during the development of the AntibodyForests package, as well as her feedback on this thesis. The weekly meetings with Alex and Daphne were instrumental in guiding the project and enhancing my understanding of the subject. Additionally, I would like to acknowledge Tudor-Stefan Cotet for his assistance with the coding aspects of this work. This project provided a valuable opportunity to develop my programming and analytical skills. I also recognize the use of GPT-4 in refining the clarity and structure of this thesis. Finally, I extend my thanks to the Yermanos group and the Center for Translational Immunology for providing a welcoming and supportive environment.

Supplementary

S1: AntibodyForests vignette

For a thorough understanding of the AntibodyForests package, including a comprehensive overview of its functions, available parameters, and practical applications, please refer to the official package vignette. This document provides in-depth discussions, examples, and best practices for utilizing the package in BCR repertoire analyses.

https://1drv.ms/u/s!Ahif4RX6I-EniLYRPffLzQBDpHLDSQ?e=9fxcgY

S2: Functions of AntibodyForests

The AntibodyForests package includes a collection of custom functions tailored for the analysis and visualization of antibody sequence data. The complete set of functions is available for download through the link provided below, with each function provided as an individual R script. Each script is thoroughly documented to clarify its workflow, ensuring that users can easily understand and apply the functions in their analyses.

https://1drv.ms/f/s!Ahif4RX6I-Enh5gJpUkBxgONMRQ1iQ?e=4PjTwi

S3: R Markdown report for data analysis and figure creation

The R Markdown report used to generate the figures presented in this thesis is available for reference. This report contains the complete codebase employed for data analysis, visualization, and figure creation, ensuring transparency and reproducibility of the results.

https://1drv.ms/u/s!Ahif4RX6I-EniMlPPbhNqx7VYw4j2w?e=5n02WV

References

[1] OpenAI. ChatGPT [Internet]. Version 4.0. San Francisco: OpenAI; 2024. Available from: https://www.openai.com/chatgpt.

[2] Sender R, Weiss Y, Navon Y, Milo R. The total mass, number, and distribution of immune cells in the human body. *Proc Natl Acad Sci U S A*. 2023 Oct 23; **120**(44): e2308511120. doi: 10.1073/pnas.2308511120.

[3] Cooper MD. The early history of B cells. *Nat Rev Immunol*. 2015 Feb 6; **15**(1): 191-197. doi: 10.1038/nri3801.

[4] Von Behring E, Kitasato S. Über das Zustandekommen der Diphtherie-Immunität und der Tetanus-Immunität bei Thieren. *Dtsch Med Wochenschr*. 1890 Dec 4; **16**(49): 1113-1114. doi: 10.1055/s-0029-1207589.

[5] Von Behring E. Untersuchungen über das Zustandekommen der Diphtherie-Immunität bei Thieren. *Dtsch Med Wochenschrift*. 1890 Dec 11; **16**(50): 1145-1148. doi: 10.1055/s-0029-1207609.

[6] Fischl R, Von Wunschheim G. Über Schutzkörper im Blute des Neugeborenen; das Verhalten des Blutserums des Neugeborenen gegen Diphteriebacillen und Diphtheriegift nebst kritischen Bemerkungen zur humoralen Immunitätstheorie. *Ztschr f Heilk*. 1895; **16**(1): 429.

[7] Polano O. Der Antitoxinubergang von der Mutter auf das Kind. Ein Beitrag zur Physiologie der Placenta. *Ztschr f Geburtsh u Gynäk*. 1904; **53**(1): 456.

[8] Von Groer F, Kassowitz K. Studien uber die normale Diphtherie-immunität des Menschen. *Ztschr. f. Immunitätsforsch. u. exper. Therap*. 1914; **22**(1): 405.

[9] McKhann CF, Chu FT. Antibodies in Placental Extracts. *J Infect Dis*. 1933 Mar 1; **52**(2): 268-277. doi: 10.1093/infdis/52.2.268.

[10] Tiselius A. The moving boundary metho of studying the electrophoresis of proteins. 1st edition. Uppsala: Almqvist & Wiksell; 1930.

[11] Tiselius A, Kabat EA. An electrophoretic study of immune sera and purified antibody preparations. *J Exp Med*. 1939 Jan 1; **69**(1): 119-131. doi: 10.1084/jem.69.1.119.

[12] Cohn EJ, Oncley JL, Strong LE, Hughes WL, Armstrong SH. Chemical, clinical, and immunological studies on the products of human plasma fractionation. I. The characterization of the protein fraction of human plasma. *J Clin Invest*. 1944 Jul 1; **23**(4): 417-432. doi: 10.1172/JCI101508.

[13] Stokes J, Maris EP, Gellis SS. Chemical, clinical, and immunological studies on the products of human plasma fractionation. XI. The use of concentrated normal human serum gamma globulin (human immune serum globulin) in the prophylaxis and treatment of measles. *J Clin Invest*. 1944 Jul 1; 531-540. doi: 10.1172/JCI101518.

[14] Waldenström J. Incipient myelomatosis or “essential” hyperglobulinemia with fibrinogenopenia: a new syndrome? *Acta Med Scand*. 1944 Jan 12; **117**(3): 216-222. doi: 10.1111/j.0954-6820.1944.tb03955.x.

[15] Wallenius G, Trautman R, Kunkel HG, Franklin EC. Ultracentrifugal studies of major non-lipide electrophoretic components of normal human serum. *J Biol Chem*. 1957 Mar 1; **225**(1): 253-267. doi: 10.1016/S0021-9258(18)64927-0.

[16] Heremans JF, Heremans MT, Schultze HE. Isolation and description of a few properties of the beta 2A-globulin of human serum. *Clin Chim Acta*. 1959 Jan 1; **4**(1): 96-102. doi: 10.1016/0009-8981(59)90088-9.

[17] Rowe DS, Fahey JL. A new class of human immunoglobulins. II. Normal serum IgD. *J Exp Med*. 1965 Jan 1; **121**(1): 185-199. doi: 10.1084/jem.121.1.185.

[18] Ishizaka K, Ishizaka T, Hornbrook MM. Physico-chemical properties of human reaginic antibody. IV. Presence of a unique immunoglobulin as a carrier of reaginic activity. *J Immunol*. 1966 Jul 1; **97**(1): 75-85. doi: 10.4049/jimmunol.97.1.75.

[19] Ishizaka K, Ishizaka T, Hornbrook MM. Physico-chemical properties of human reaginic antibody. V. Correlation of reaginic activity with gamma-E-globulin antibody. *J Immunol*. 1966 Dec 1; **97**(6): 840-853. doi: 10.4049/jimmunol.97.6.840.

[20] Korngold L, Lipari R. Multiple-myeloma proteins. III. The antigenic relationship of Bence Jones proteins to normal gammaglobulin and multiple-myeloma serum proteins. *Cancer*. 1956 Mar 1; **9**(2): 262-272. doi: 10.1002/1097-0142(195603/04)9.

[21] Fagraeus A. The plasma cellular reaction and its relation to the formation of antibodies *in vitro*. *J Immunol*. 1948 Jan 1; **58**(1): 1-13. doi: 10.4049/jimmunol.58.1.1.

[22] Waldeyer W. Über Bindegewebszellen. *Arch Mikr Anat*. 1875 Dec 1; **11**(1): 176-194. doi: 10.1007/BF02933794.

[23] Ehrlich P. Zur Kenntnis der Antitoxinwirkung. *Forschr Med*. 1897 Jan 15; **15**(2): 41-43.

[24] Breinl F, Haurowitz F. Chemische Untersuchung des Präzipitates aus Hämoglobin und Anti-Hämoglobin-Serum und Bemerkungen über die Natur der Antikörper. *JBC*. 1930 Aug 4; **192**(1): 45-57. doi: 10.1515/bchm2.1930.192.1-3.45.

[25] Burnet FM, Fenner F. The production of antibodies. *J Immunol*. 1951 Apr 1; **66**(4): 485-486. doi: 10.4049/jimmunol.66.4.485.

[26] Jerne NK. The natural-selection theory of antibody formation. *Proc Natl Acad Sci U S A*. 1955 Nov 15; **41**(11): 849-857. doi: 10.1073/pnas.41.11.849.

[27] Talmage DW. Diversity of antibodies. *J Cell Physiol*. 1957 Dec 1; **50**(51): 229-246. doi: 10.1002/jcp.1030500415.

[28] Burnet FM. A modification of Jerne's theory of antibody production using the concept of clonal selection. *Aust J Sci*. 1957; **20**(3): 67-69.

[29] Nossal GJ, Lederberg J. Antibody production by single cells. *Nature*. 1958 May 17; **181**(4620): 1419-1420. doi: 10.1038/1811419a0.

[30] Porter RR. The hydrolysis of rabbit γ-globulin and antibodies with crystalline papain. *Biochem J*. 1959 Sep 1; **73**(1): 119-127. doi: 10.1042/bj0730119.

[31] Edelman GM. Dissociation of gamma-globulin. *J Am Chem Soc*. 1959 June 20; **81**(12): 3155-3156. doi: 10.1021/ja01521a071.

[32] Hilschmann N, Craig LC. Amino acid sequence studies with Bence-Jones proteins. *Proc Natl Acad Sci U S A*. 1965 Jun 1; **53**(6): 1403-1409. doi: 10.1073/pnas.53.6.1403.

[33] Dreyer WJ, Bennet JC. The molecular basis of antibody formation: a paradox. *Proc Natl Acad Sci U S A*. 1965 Sep 1; **54**(3): 864-969. doi: 10.1073/pnas.54.3.864.

[34] Eisen HN, Siskind GW. Variations in affinities of antibodies during the immune response. *Biochemistry*. 1964 Jul 1; **3**(1): 996-1008. doi: 10.1021/bi00895a027.

[35] Edeman GM, Cunningham BA, Gall WE, Gottlieb PD, Rutishauer U, Waxdal MJ. The covalent structure of an entire gammaG immunoglobulin molecule. *Proc Natl Acad Sci U S A*. 1969 May 1; **63**(1): 78-85. doi: 10.1073/pnas.63.1.78.

[36] Landsteiner K, Chase MW. Experiments on transfer of cutaneous sensitivity to simple compounds. *P Soc Exp Biol Med*. 1942 Apr 1; **49**(4): 688-690. doi: 10.3181/00379727-49-13670.

[37] Landsteiner K, Chase MW. The cellular transfer of cutaneous hypersensitivity to tuberculin. *P Soc Exp Biol Med*. 1945 Jun 1; **59**(2): 134-135. doi: 10.3181/00379727-59-15006P.

[38] Bruton OC. Agammaglobulinemia. *Pediatrics*. 1952 Jun 1; **9**(6): 722-728. doi: 10.1542/peds.9.6.722.

[39] Good RA, Varco RL. A clinical and experimental study of agammaglobulinemia. *Lancet*. 1955 Jun; **75**(6): 245-271.

[40] Glick B, Chang TS, Jaap RG. The bursa of fabricius and antibody production. *Poult Sci*. 1956 Jan 1; **56**(1): 224-225. doi: 10.3382/ps.0350224.

[41] Miller JFAP. Immunological function of the thymus. *Lancet*. 1961 Sep 30; **278**(7205): 748-749. doi: 10.1016/S0140-6736(61)90693-6.

[42] Auerbach R. Experimental analysis of the origin of cell types in the development of the mouse thymus. *Dev Biol*. 1961 Jun 1; **3**(3): 336-354. doi: 10.1016/0012-1606(61)90051-3.

[43] Gowans JL, McGregor DD. The origin of antibody-forming cells. In: Grabar P, Mischier PA, editors. Immunopathology: 3rd International Symposium. Basel: Schwabe; 1963. p. 83-101.

[44] Gowans JL, Knight EJ. The route of re-circulation of lymphocytes in the rat. *R Soc Open Sci*. 1964 Jan 14; **159**(975): 257-282. doi: 10.1098/rspb.1964.0001.

[45] Cooper MD, Chae HP, Lowman JT, Krivit W, Good RA. Wiskott-Aldrich syndrome. An immunologic deficiency disease involving the afferent limb of immunity. *Am J Med*. 1968 Apr 1; **44**(4): 499-513. doi: 10.1016/0002-9343(68)90051-x.

[46] Cooper MD, Peterson RDA, Good RA. Delineation of the thymic and bursal lymphoid systems in the chicken. *Nature*. 1965 Jan 1; **205**(1): 143-146. doi: 10.1038/205143a0.

[47] Cooper MD, Peterson RDA, South MA, Good RA. The functions of the thymus system and the bursa system in the chicken. *J Exp Med*. 1966 Jan 1; **123**(1): 75-102. doi: 10.1084/jem.123.1.75.

[48] Cooper MD, Perey DY, Mckneally MF, Gabrielsen AE, Sutherland DER, Good RA. A mammalian equivalent of the avian bursa of Farbicius. *Lancet*. 1966 Jun 25; **287**(7452): 1388-1391. doi: 10.1016/S0140-6736(66)90300-X.

[49] Owen JJT, Cooper MD, Raff MC. *In vitro* generation of B lymphocytes in mouse foetal liver, a mammalian 'burse equivalent'. *Nature*. 1974 May 24; **249**(1): 361-363. doi: 10.1038/249361a0.

[50] Raff MC, Megson M, Owen JJT, Cooper MD. Early production of intracellular IgM by B-lymphocyte precursors in mouse. *Nature*. 1976 Jan 22; **259**(1): 224-226. doi: 10.1038/259224a0.

[51] Owen JTT, Raff MC, Cooper MD. Studies on the generation of B lymphocytes in the mouse embryo. *Eur J Immunol*. 1976 Jul; **5**(7): 468-473. doi: 10.1002/eji.1830050708.

[52] Ryser JE, Vassalli P. Mouse bone marrow lymphocytes and their differentiation. *J Immunol*. 1974 Sep 1; **113**(3): 719-728. doi: 10.4049/jimmunol.113.3.719.

[53] Wu TT, Kabat EA. An analysis of the sequences of the variable regions of Bence Jones proteins and myeloma light chains and their implications for antibody complementarity. *J Exp Med*. 1970 Aug 1; **132**(2): 211-250. doi: 10.1084/jem.132.2.211.

[54] Hozumi N, Tonegawa S. Evidence for somatic rearrangement of immunoglobulin genes coding for variable and constant regions. *Proc Nacl Acad Sci U S A*. 1976 Oct 1; **73**(10): 3628-3632. doi: 10.1073/pnas.73.10.3628.

[55] Bernard O, Hozumi N, Tonegawa S. Sequences of mouse immunoglobulin light chain genes before and after somatic changes. *Cell*. 1978 Dec 1; **15**(4): 1133-1144. doi: 10.1016/0092-8674(78)90041-7

[56] McKean D, Huppi K, Bell M, Staudt L, Gerhard W, Weigert M. Generation of antibody diversity in the immune response of BALB/c mice to influenza virus hemagglutinin. *Proc Natl Acad Sci U S A*. 1984 May 1; **81**(10): 3180-3184. doi: 10.1073/pnas.81.10.3180.

[57] Abbas AK, Lichtman AH, Pillai S. Cellular and Molecular Immunology. 10th ed. Philadelphia: Elsevier; 2022.

[58] Chiu ML, Goulet DR, Teplyakov A, Gilliland GL. Antibody structure and function: The basis for engineering therapeutics. *Antibodies*. 2019 Dec 1; **8**(4): 55. doi: 10.3390/antib8040055.

[59] Rodriguez O, Safonova Y, Silver CA, Shields K, Gibson WS, Kos JT, *et al*. Genetic variation in the immunoglobulin heavy chain locus shapes the human antibody repertoire. *Nat Commun*. 2023 Jul 21; **14**(1): 4419. doi: 10.1038/s41467-023-40070-x.

[60] Collins AM, Watson CT. Immunoglobulin light chain gene rearrangements, receptor editing and the development of a self-tolerant antibody repertoire. *Front Immunol*. 2018 Oct 8; **9**(1): 2249. doi: 10.3389/fimmu.2018.02249.

[61] Roth DB. V(D)J Recombination: Mechanism, errors, and fidelity. *Micrbiol Spectr*. 2014 Nov 21; **2**(6): 41-51. doi: 10.1128/microbiolspec.mdna3-0041-2014.

[62] Cambier JC, Gauld SB, Merrel KT, Vilen BJ. B-cell anergy: from transgenic models to naturally occurring anergic B cells? *Nat Rev Immunol*. 2007 Jul 20; **7**(1): 633-643. doi: 10.1038/nri2133.

[63] Elgueta R, Benson MJ, De Vries V, Wasiuk A, Guo Y, Noelle RJ. Molecular mechanism and function of CD40/CD40L engagement in the immune system. *Immunol Rev*. 2009 Apr 21; **229**(1): 152-172. 10.1111/j.1600-065X.2009.00782.x.

[64] Maul RW, Gearhart PJ. AID and somatic hypermutation. *Adv Immunol*. 2010; **105**(1): 159-191. doi: 10.1016/S0065-2776(10)05006-6.

[65] Rogozin IB, Kolchanov NA. Somatic hypermutagenesis in immunoglobulin genes: II. Influence of neighbouring base sequences on mutagenesis. *Biochim Biophys Acta*. 1992 Nov 15; **117**(1): 11-18. doi: 10.1016/0167-4781(92)90134-l.

[66] Stavnezer J, Guikema JEJ, Schrader CE. Mechanism and regulation of class switching recombination. *Annu Rev Immunol*. 2008 Apr 1; **26**(1): 261-292. doi: 10.1146/annurev.immunol.26.021607.090248.

[67] Roco J, Mesin L, Binder SC, Nefzger C, Gonzalez-Figueroa P, Canete PF, *et al*. Class-switch recombination occurs infrequently in germinal centers. *Immunity*. 2019 Aug 20; **51**(2): 337-350. doi: 10.1016/j.immuni.2019.07.001.

[68] Sanger F, Nicklen S, Coulson AR. DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci U S A*. 1977 Dec 1; **74**(12): 5463-5467. doi: 10.1073/pnas.74.12.5463.

[69] Moss PA, Moots RJ, Rosenberg WM, Rowland-Jones SJ, McMichael AJ, Bell JI. Extensive conservation of alpha and beta chains of the human T-cell antigen receptor recognizing HLA-A2 and influenza A matrix peptide. *Proc Natl Acad Sci U S A*. 1991 Oct 15; **88**(20): 8987-8990. 10.1073/pnas.88.20.8987.

[70] Price DA, Brenchley JM, Ruff LE, Betts MR, Hill BJ, Roederer M, *et al*. Avidity for antigen shapes clonal dominance in CD8+ T cell populations specific for persistent DNA viruses. *J Exp Med*. 2005 Nov 14; **202**(10): 1349-1361. doi: 10.1084/jem.20051357.

[71] Yu XG, Lichterfeld M, Chetty S, Williams KL, Mui SK, Miura T, *et al*. Mutually exclusive T-cell receptor induction and differential susceptibility to Human Immunodeficiency Virus Type 1 mutational escape associated with a two-amino-acid difference between HLA class I subtypes. *Virol J*. 2007 Feb 15; **81**(4): 1619-1631. doi: 10.1128/jvi.01580-06.

[72] Silverman G, Goodyear C. Confounding B-cell defenses: lessons from a staphylococcal superantigen. *Nat Rev Immunol*. 2006 Jun 1; **6**(1): 465-475. doi: 10.1038/nri1853.

[73] Shendure J, Porreca G, Reppas NB, Lin X, McCutcheon JP, Rosenbaum AM, *et al*. Accurate multiplex polony sequencing of an evolved bacterial genome. *Science*. 2005 Sep 9; **309**(5741): 1728-1732. doi: 10.1126/science.1117389.

[74] Rothberg J, Leamon JH. The development and impact of 454 sequencing. *Nat Biotechnol*. 2008 Oct 9; **26**(10): 1117-1124. doi: 10.1038/nbt1485.

[75] Gorny MK, Wang XH, Williams C, Volsky B, Revesz K, Witover B, *et al*. Preferential use of the VH5-51 gene segment by the human immune response to code for antibodies against the V3 domain of HIV-1. *Mol Immunol*. 2009 Feb 1; **46**(1): 917-926. doi: 10.1016/j.molimm.2008.09.005.

[76] Breden F, Lepik C, Longo NS, Montero M, Lipsky PE, Scott JK. Comparison of antibody repertoires produced by HIV-1 infection, other chronic and acute infections, and systemic autoimmune disease. *PLoS One*. 2011 Mar 30; **6**(3): e16857. doi: 10.1371/journal.pone.0016857.

[77] Neumeier D, Yermanos A, Agrafiotis A, Csepregi L, Chowdhury T, Ehling RA, *et al*. Phenotypic determinism and stochasticity in antibody repertoires of clonally expanded plasma cells. *Proc Natl Acad Sci U S A*. 2022 Apr 29; **119**(18): e2113766119. doi: 10.1073/pnas.2113766119.

[78] Shehata L, Maurer D, Wec AZ, Lilov A, Champney E, Sun T, *et al*. Affinity maturation enhances antibody specificity but compromises conformational stability. *Cell Rep*. 2019 Sep 24; **28**(1): 3300-3308. doi: 10.1016/j.celrep.2019.08.056.

[79] Agrafiotis A, Neumeier D, Hong KL, Chowdhury T, Ehling R, Kuhn R, *et al*. Generation of single-cell B cell atlas of antibody repertoires and transcriptomes to identify signatures associated with antigen specificity. *iScience*. 2023 Jan 25; **26**(3): 106055. doi: 10.1016/j.isci.2023.106055.

[80] Jaffe DB, Shahi P, Adams BA, Chrisman AM, Finnegan PM, Raman N, *et al*. enclone: precision clonotyping and analysis of immune receptors. *bioRxiv*. 2022 July 9. doi: 10.1101/2022.04.21.489084.

[81] 10x Genomics. Enclone: Accurate and user-friendly computational tool for clonal grouping to study the adaptive immune system [Internet]. Available from: https://10xgenomics.github.io/enclone/.

[82] Hoehn KB, Kleinstein SH. B cell phylogenetics in the single cell era. *Trends Immunol*. 2023 Dec 27; **45**(1): 62-74. doi: 10.1016/j.it.2023.11.004.

[83] Felsenstein J. Cases in which parsimony or compatibility methods will be positively misleading. *Syst Biol*. 1978 Dec 1; **27**(4): 401-410. doi: 10.1093/sysbio/27.4.401.

[84] Arenas M. Trends in substitution models of molecular evolution. *Front Genet*. 2015 Oct 26; **6**(1): 319. doi: 10.3389/fgene.2015.00319.

[85] Goldman N, Yang Z. A Codon-based Model of Nucleotide Substitution for Protein-coding DNA Sequences. *Mol Biol Evol*. 1994 Sep; **11**(5): 725-736. doi: 10.1093/oxfordjournals.molbev.a040153.

[86] Hoehn B, Lunter G, Pybus OG. A Phylogenetic Codon Substitution Model for Antibody Lineages. *Genetics*. 2017 May 1; **206**(1): 417-427. doi: 10.1534/genetics.116.196303.

[87] Zheng B, Yang Y, Chen L, Wu M, Zhou S. B-cell receptor repertoire sequencing: Deeper digging into the mechanisms and clinical aspects of immune-mediated diseases. *iScience*. 2022 Oct 21; **25**(1): 105002. doi: 10.1016/j.isci.2022.105002.

[88] 10x Genomics. Chromium Next GEM Single Cell V(D)J Reagent Kits v1.1 User Guide. 2021 Oct. Document Number: CG000207.

[89] 10x Genomics. Getting Started with Cell Ranger v8.0.1 [Internet]. 2024 May 8. Available from: https://www.10xgenomics.com/support/software/cell-ranger/latest/getting-started.

[90] Paradis E, Schliep K. ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics*. 2019 Feb 1; **35**(3): 526-528. doi: 10.1093/bioinformatics/bty633.

[91] Schliep P. phangorn: phylogenetic analysis in R. *Bioinformatics*. 2011 Feb 1; **27**(4): 592-593. doi: 10.1093/bioinformatics/btq706.

[92] Ye J, Ma N, Madden TL, Ostell JM. IgBLAST: an immunoglobulin variable domain sequence analysis tool. *Nucleic Acids Res*. 2013 Jul 1; **41**(1): W34-W40. doi: 10.1093/nar/gkt382.

[93] Lefranc MP, Giudicelli V, Kaas Q, Duprat E, Jabado-Michaloud J, Scaviner D, *et al*. IMGT, the international ImMunoGeneTics information system ®. *Nuecl Acids Res*. 2005 Jan 1; **33**(1): D593-D597. doi: 10.1093/nar/gki065.

[94] Khan T, Friedensohn S, Gorter de Vries AR, Straszewski J, Ruscheweyh HJ, Reddy ST. Accurate and predictive antibody repertoire profiling by molecular amplification fingerprinting. *Sci Adv*. 2016 Mar 1; **2**(3): e1501371. doi: 10.1126/sciadv.1501371.

[95] Richardson E, Binter Š, Kosmac M, Ghraichy M, Von Niederhäusern V, Kovaltstuk A, *et al*. Characterisation of the immune repertoire of a humanised transgenic mouse through immunophenotyping and high-throughput sequencing. *eLife*. 2023 Mar 27; **12**(1): e81629. doi: 10.7554/eLife.81629.

[96] Lefranc MP. Unique database numbering system for immunogenetic analysis. *Immunol Today*. 1997 Nov; **18**(11): 509. doi: 10.1016/s0167-5699(97)01163-8.

[97] Collins AM, Wang Y, Roskin KM, Marquis CP, Jackson KJL. The mouse antibody heavy chain repertoire is germline-focused and highly variable between inbred strains. *Philos Trans R Soc Lond B Biol Sci*. 2015 Sep 5; **370**(1676): 20140236. doi: 10.1098/rstb.2014.0236.

[98] Lie BL, Shanker VR, Xu D, Bruun TUJ, Weidenbacher PA, Tang S, *et al*. Efficient evolution of human antibodies from general protein language models. *Nat Biotechnol*. 2023 Apr 24; **42**(1): 275-283. doi: 10.1038/s41587-023-01763-2.

[99] IgPhyML Documentation. IgPhyML - B cell phylogenetic inference package [Internet]. Available from: https://igphyml.readthedocs.io/en/latest/index.html.

[100] Change-O Documentation. Change-O - Repertoire clonal assignment toolkit [Internet]. Available from: https://changeo.readthedocs.io/en/stable/index.html.