High-throughput single-cell characterization of the genomic and serum antibody repertoire

Khang Lê Quý

Thesis submitted for the degree of

Master of Science in Bioscience

60 credits

Department of Biosciences

Faculty of Mathematics and Natural Sciences

UNIVERSITY OF OSLO

May 2021

Table of contents

Acknowledgements **4**

Abbreviations **5**

Abstract **7**

1. Introduction **8**

1.1 The mammalian immune system **8**

1.2 Innate immunity **8**

1.3 Adaptive immunity **9**

1.4 Lymphocytes **10**

1.4.1 T lymphocytes 10

1.4.1.1 Development of T cells 10

1.4.1.2 Antigen recognition and effector functions of T cells 11

1.4.2 B lymphocytes 12

1.4.2.1 Development of B cells 12

1.4.2.2 Structure and antigen recognition of the antibody 17

1.4.2.3 Class switching and affinity maturation of B-cell receptors 19

1.4.2.3 Classes and effector functions of immunoglobulins 20

1.4.2.4 Diversity of the antibody repertoire 22

1.5 Sequencing of the B-cell receptor repertoire **23**

1.5.1 Current B-cell receptor repertoire sequencing methods 23

1.5.2 Single-cell sequencing of B-cell receptor repertoires 27

1.5.3 Applications of B-cell receptor repertoire analysis 29

1.6 Characterization of serum antibody repertoire by mass spectrometry **30**

1.6.1 Approaches for mass spectrometry-based antibody proteomics 30

1.6.2 Liquid chromatography tandem mass spectrometry in antibody proteomics 31

2. Thesis aims **35**

3. Methods **37**

3.1 Bulk B-cell receptor sequencing **37**

3.1.1 B cell isolation from peripheral blood 38

3.1.2 Cell counting 38

3.1.3 RNA isolation 38

3.1.4 Determination of nucleic acid concentration 38

3.1.5 cDNA synthesis 39

3.1.6 DNA purification 39
3.1.7 Multiplex PCR 40
3.1.8 Gel electrophoresis 40
3.1.9 DNA extraction from agarose gel 41
3.1.10 Adapter extension PCR 41
3.1.11 Capillary Electrophoresis 42
3.2 Single-cell B-cell receptor sequencing 42
3.2.1 Single-cell encapsulation with the Nadia Instrument 44
3.2.2 In-droplet reverse transcription 46
3.2.3 Emulsion breakage and target enrichment PCR 48
3.3 Antibody mass spectrometry 49
3.3.1 Antibody purification from serum 49
3.3.2 GingisKHAN antibody F(ab) fragment collection 49
3.3.3 Enzymatic digestion of antibodies 49
3.3.4 Liquid chromatography and tandem mass spectrometry 50
3.4 Data analysis 51
3.4.1 Quality control of sequencing reads 51
3.4.2 UMI-based error correction 51
3.4.3 Read assembly and clonotyping 52
3.4.4 Immune repertoire analysis 54
3.4.5 Mass spectrometry data analysis 55
4. Results 57
4.1 Bulk B-cell receptor sequencing 57
4.1.1 Assessment of library quality 57
4.1.2 Analysis of B-cell receptor libraries 61
4.2 Single-cell B-cell receptor sequencing 76
4.2.1 B-cell encapsulation 76
4.2.2 Assessment of library quality 78
4.3 Benchmarking of the antibody LC-MS/MS pipeline 79
4.3.1 Proof-of-concept and pilot experiments 79
4.3.1.1 Antibody peptides detection with LC-MS/MS 79
4.3.1.2 Performance comparison between MS settings 81
4.3.1.3 Correlation between intensity ratios and concentration ratios in peptides 84
4.3.1.4 Limit of detection for LC-MS/MS in antibody identification 85
4.3.2 Detection of monoclonal antibodies at different concentrations 86
5. Discussion 92

5.1 Bulk B-cell receptor sequencing: method adaptation and considerations 93
5.2 Single-cell B-cell receptor sequencing: advancements and limitations 96
5.3 Antibody LC-MS/MS: benchmarking of antibody proteomics 98
6. Outlook and future perspectives 102
Appendix 103
References 113

Acknowledgements

The work presented in this master's thesis was carried out at the Department of Immunology, Rikshospitalet, Oslo from January 2020 to March 2021.

First and foremost, I would like to express my sincerest gratitude to my head supervisor Dr Igor Snapkov. Thank you for all your guidance not only in the lab working on cutting-edge research, but also in organizing and planning scientific work, and for continuously supporting me throughout my whole master study. My deepest appreciation also goes to my co-supervisor and head of the research group Associate Professor Victor Greiff. Thank you for giving me the opportunity to work in your lab and for your advice on the whole research process, introducing me to the way research is done, all the best practices, and pitfalls to avoid.

I would also like to thank my internal supervisor Professor Finn-Eirik Johansen. Thank you for your wonderful lectures on immunology. I learned a lot from your lectures and managed to expand my knowledge base beyond my topic of research. For that I am very grateful.

Many thanks to all the members of the Greiff lab. Everyone has been so accommodating and supportive throughout my time in the lab. I feel very privileged working with such a talented and diverse group of co-workers.

Finally, all the love and gratitude to my family for supporting the decision to go halfway around the world pursuing my passion. I would not be writing this without their support and encouragement. My appreciation also goes to my partner, Thu. Thank you for motivating me through all the hardships and for pursuing a future with me.

Oslo, May 2021
Khang Lê Quý

Abbreviations

AID: Activation-induced deaminase

APC: Antigen-presenting cell

BCR: B-cell receptor

CD: Cluster of differentiation

CDR: Complementarity determining region

CID: Collision induced dissociation

CLP: Common lymphoid progenitor

CSR: Class switch recombination

Ct: chymotrypsin

ER: Endoplasmic reticulum

F(ab): Fragment antigen binding

FACS: Fluorescence-activated cell sorting

Fc: Fragment crystallizable

FDC: Follicular dendritic cell

FR: Framework region

GC: Germinal center

HCD: Higher-energy collisional dissociation

HPLC: High-performance liquid chromatography

HTS: High-throughput sequencing

Ig: Immunoglobulin

LC-MS/MS: Liquid chromatography with tandem mass spectrometry

mAb: Monoclonal antibody

MHC: Major Histocompatibility Complex

MIG: Molecular identifier group

MS: Mass spectrometry

MTPX: Multiplex (PCR)

MZ: Marginal zone

NHEJ: Non-homologous end joining

PC: Plasma cell

PTM: Post-translational modification

R1: Forward read (R1)

R2: Reverse read (R2)

RACE: Rapid amplification of cDNA ends

RAG: Recombination activating gene

RSS: Recombination signal sequence

scRNA-seq: Single-cell RNA sequencing

SHM: Somatic hypermutation

SLC: Surrogate light chain

T_c: Cytotoxic T cell

TCR: T-cell receptor

TdT: Terminal deoxynucleotidyl transferase

Tfh: T follicular helper cell

T_H: Helper T cell

Tregs: Regulatory T cell

Tryp: trypsin

UMI: Unique molecular identifier

Abstract

Investigations into the relationship between genomic and phenotypic (serum) diversity of antibodies are of decisive importance for understanding the human adaptive immune response in health and disease. The capability to accurately predict and describe the entire antibody repertoire of the body in detail, including both B-cell receptors (antibody genome) and circulating serum antibodies (antibody phenome) will dramatically alter approaches to vaccine development and disease diagnostics. However, despite recent advances in high-throughput analytical techniques to mine antibody repertoires in great molecular depth, a comprehensive characterization of antibody complexity at the single-cell and single-molecule levels remains elusive due to the fact that most sequencing approaches fail to capture the natural pairing of B-cell receptor (BCR) heavy and light chain variable regions. Additionally, the biological reasons for the abundance difference between the number of B cells with a distinct receptor and the number of circulating antibodies remain unclear. Therefore, extensive proteomic profiling of serum antibodies, as well as coupling of sequencing and proteomics data, are crucial in order to further advance the field of immunology. In this Master's thesis project, we have adapted and improved upon a highly reliable and reproducible experimental protocol for human BCR sequencing in bulk, established a solid foundation for single-cell BCR sequencing, and performed an extensive benchmarking of existing approaches to antibody proteomics. The findings and knowledge obtained in this work will be translated into a Ph.D. project aiming at the development of an open-access systems immunology platform that combines bulk- and single-cell sequencing of BCRs and protein sequencing of serum antibodies in order to characterize a person's B-cell repertoire in great detail.

1. Introduction

1.1 The mammalian immune system

The immune system protects the body from the harmful effects of pathogens, cleans up dead cells, and maintains homeostasis. The immune system has multiple layers built on top of each other as a result of evolution. As pathogens evolve to increase their chances of successfully infecting the host, so too must the host's defenses evolve to deal with them [1]. The mammalian immune system comprises of three layers: (i) mechanical and chemical barriers which operate continuously to limit exposure to pathogens, (ii) innate immunity which responds within minutes to hours once the physical barriers have been breached, and (iii) adaptive immunity which is fully effective within days to weeks after the initial infection to eliminate threats not cleared out by innate immunity [2]. To achieve this level of protection, the immune system organizes a complex system of cells and organs. Primary lymphoid organs, such as bone marrow and thymus, are where immune cells differentiate from lymphoid stem cells and gain their effector functions. Secondary lymphoid organs, such as lymph nodes and spleen, are where antigens are encountered and responded to. However, it is more accurate to consider the immune system not as discrete parts that function independently, but rather as interconnected and complementary processes [3].

1.2 Innate immunity

The innate immune system plays a key role in the early response to pathogens, consisting of cells and molecules that are readily available and spread throughout the body. These cells and molecules can react rapidly with full effectiveness without prior encounters with pathogens. This is achieved through a system of recognition based on common patterns exhibited by pathogens. These include, but are not limited to, lipopolysaccharides in bacterial cell walls, double-stranded RNA produced by replicating viruses, and mannose residues in microbial glycoproteins, all of which are common and essential for many microbial life functions. In addition, molecules created by dying or damaged cells can also be recognized to alert and stimulate the immune response [4]. However, the threat recognition ability of innate immunity is limited by the finite

number of germline-encoded genes and the inability to develop a memory of previous pathogen encounters. In addition, pathogens have long evolved alongside innate immunity and developed strategies to evade and minimize the innate immune response [5].

Cells of the innate immune system include phagocytes (macrophages, neutrophils, and dendritic cells), which engulf and digest pathogens with enzymes; basophils and eosinophils which secrete molecules in response to an infection; and natural killer (NK) cells, which kills infected cells through contact. Phagocytes, especially dendritic cells (DC), also contribute to adaptive immunity by presenting antigens and directing the development of T cells[1].

1.3 Adaptive immunity

The adaptive immune response is highly specific to a wide array of threats, enabled by the expression of antigen-specific receptors on the cell membrane of lymphocytes (T cells and B cells). This specificity exists due to the large number of genes that code for these receptors and the somatic recombination process that greatly expands the variety of gene assemblies. As a receptor binds to an antigen with specificity, the cell expressing that receptor becomes activated and proliferates, creating thousands of clones with the same specificity, a process known as clonal expansion [6,7]. In addition, receptor binding to antigens also creates long-lived memory lymphocytes that can respond more rapidly and strongly to any subsequent exposure to the same antigen. As a result, the body is protected from pathogens that are common and recurring in the environment [8]. This mechanism has been utilized as the basis for vaccination.

There are two types of adaptive immunity: humoral immunity and cellular immunity. Humoral immunity is mediated by B lymphocytes and their secreted glycoproteins, the antibodies, and helps to protect the body from extracellular threats. Cellular immunity, in contrast, is mediated by T lymphocytes and responds to intracellular threats where it is inaccessible to antibodies [4].

[1] Antigen presentation to T cells is explained further in section 1.4.1.2 “Antigen recognition and effector functions of T cells”.

1.4 Lymphocytes

All the cells of the immune system arise from hematopoietic stem cells in the bone marrow or the fetal liver. These stem cells differentiate into precursor cells of the myeloid lineage, which at terminal differentiation creates polymorphonuclear granulocytes, macrophage, and dendritic cells, or the lymphoid lineage, which creates NK cells, T cells and B cells [9].

1.4.1 T lymphocytes

1.4.1.1 Development of T cells

There are two types of T cells, distinguished by their T-cell receptor: αβ T cells and γδ T cells. γδ T cells account for only 5–10% of the T cells population and mainly serve to protect the body's mucosal surfaces. In contrast, αβ T cells account for 90–95% of T cells and are further divided into subpopulations, such as Helper T cells (T_H), Cytotoxic T cells (T_c), and Regulatory T cells (Tregs) [9].

The common lymphoid progenitor (CLP) cells that migrate to the thymus, with the induction of Notch-1 and other transcription factors, commit to the T-cell lineage [10]. T-cell precursors first express neither CD4 nor CD8 ("double negative" cells). It is at this stage that somatic recombination occurs which leads to the expression of the αβ or the γδ T-cell receptors (TCR). First, the TCR β chain undergoes somatic recombination[2] in the thymus subcapsular zone. If the rearrangement is productive, the cell moves to the thymic cortex and undergoes rearrangement for the TCR α chain and expression of the CD3, CD4, and CD8 co-receptors ("double positive" cells) [11].

T cells with productive TCR then undergo positive selection with the Major Histocompatibility Complex (MHC) molecules. Cells that exhibit sufficient affinity to MHC class I (MHC I) lose expression of CD4 and retain expression of CD8, while cells that exhibit sufficient affinity to MHC class II (MHC II) lose expression of CD8 and retain expression of CD4. Failure to recognize self-MHC leads to apoptosis [10,11].

[2] Somatic recombination in T-cell and B-cell receptors follows the same principles and therefore is explained in detail in section 1.4.2.1 "Development of B cells".

T cells that survive positive selection migrate to the thymic medulla and undergo negative selection. This process eliminates cells that recognize self-antigens presented by the MHC molecules. Self-antigens include proteins that are common in tissues and in circulation. In addition, the thymus cells also have a mechanism to exhibit many different types of antigens in different tissues, all to ensure that the T cells do not attack the body's own cells. However, some $CD4^+$ T cells that recognize self-antigens instead differentiate into Tregs cells that work to regulate the immune response [4].

1.4.1.2 Antigen recognition and effector functions of T cells

T cells exit the thymus as mature but "naive" lymphocytes inexperienced with antigens. These cells enter circulation and migrate to the secondary lymphoid organs such as lymph nodes, where antigens can be encountered. The TCR is unable to recognize an antigen in its native form. Instead, the antigen needs to be processed and presented by the MHC [12]. Interaction between the TCR and the MHC with processed antigen, in addition to an array of cell surface molecules, forms a structure collectively termed "the immunological synapse", leading to the activation and execution of T-cell effector functions [13].

MHC class I molecules are ubiquitously expressed on all nucleated cells and bind with peptides processed from proteins produced inside the cell as a result of infection. These proteins are marked by ubiquitin and degraded in proteasomes, the resulting peptides are transported to the endoplasmic reticulum (ER). In the ER, the peptides are packaged in vesicles, exported to the cell surface, and bound with MHC class I peptide-binding groove, which can bind to peptides around 8–10 residues long [11,14]. The binding of the TCR with MHC-peptide complex leads to the destruction of the infected cells, stabilized by the CD8 co-receptor. This is achieved by the release of cytotoxic granules containing perforin and granzymes. Perforin creates pores on the target cell membrane, then granzymes can enter the cell and trigger programmed cell death (apoptosis) [15].

MHC class II is expressed on professional antigen-presenting cells (APC) such as dendritic cells, macrophages, and B cells. However, many other cell types can also be induced to express MHC II with IFN-ɣ [11]. Extracellular antigens such as bacteria, viruses released from infected cells, and proteins are taken up by APCs through phagocytosis, pinocytosis, or receptor-mediated endocytosis [16]. Afterward, these antigens undergo proteolysis, creating peptide fragments that accumulate in lysosomes. In the ER, new MHC II molecules are synthesized and transported to the lysosomes where peptide binding occurs. The peptide-loaded MHC II is then transported to the cell surface where it can display the antigen to $CD4^+$ T cells [17]. $CD4^+$ T cells account for the majority of T cells in the body and are divided into many subsets. T_H1 cells produce IFN-ɣ and IL-2 and play an important role in cell-mediated immunity and production of complement-activating antibodies. T_H2 cells produce IL-4 and mediate primarily the defense against parasites and IgE antibody production. Other T_H subsets include T_H17, T_H22, T_H9, and T follicular helper cells (Tfh) [18]. In particular, Tfh cells play an important role in the formation and maintenance of germinal centers (GC), and B cells are dependent on Tfh cells for survival, proliferation, and differentiation [19].

1.4.2 B lymphocytes

1.4.2.1 Development of B cells

CLP cells that are committed to the B-cell lineage pass through a series of developmental steps to form mature B cells (Figure 1). B-cell development can be divided into two phases: (i) the antigen-independent phase and (ii) the antigen-dependent phase. Antigen-independent development occurs in the bone marrow (Figure 1A) while antigen-dependent development occurs in secondary lymphoid organs such as the spleen and lymph nodes (Figure 1B) [20].

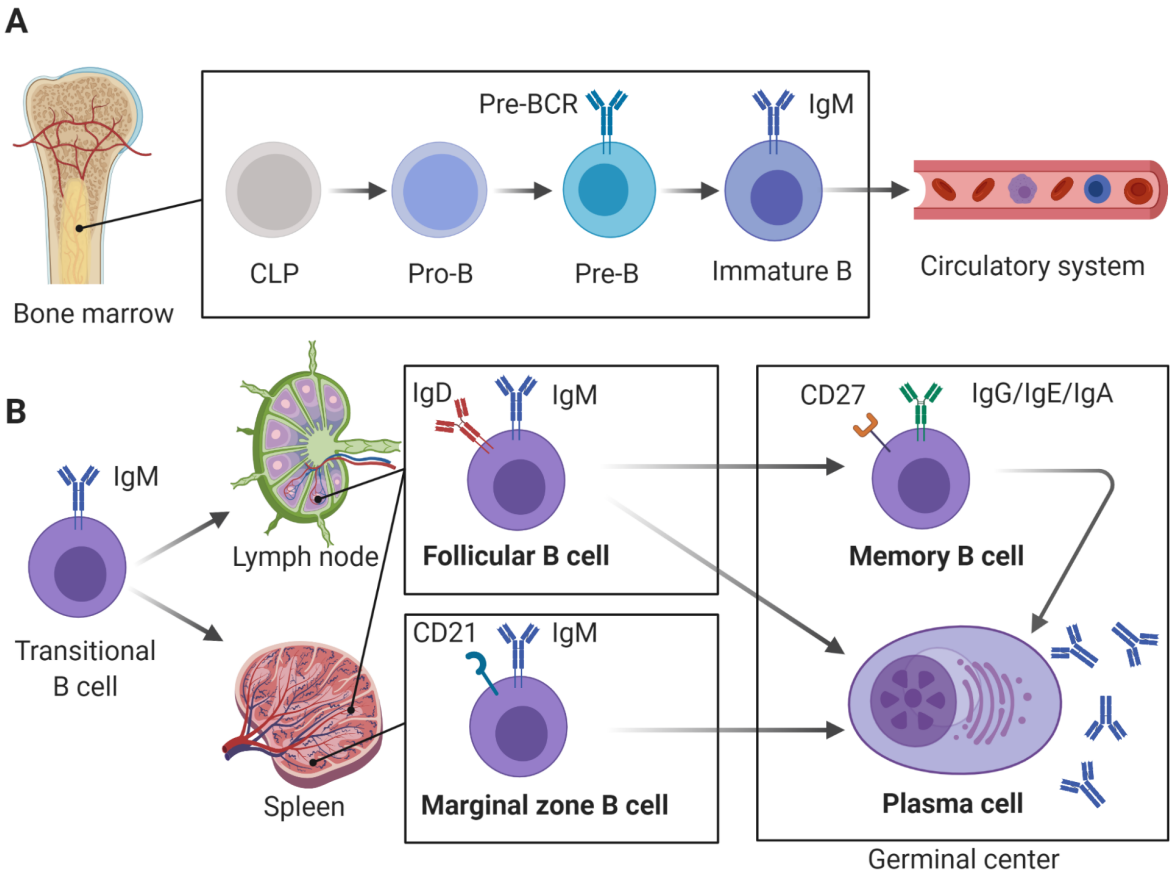


Figure 1: Development of B cells. ***A)*** *Antibody-independent development in the bone marrow. Pro-B cells begin somatic recombination on the heavy chain locus. The heavy chain protein is expressed on the cell surface in association with an invariant surrogate light chain (SLC) to form the pre-BCR complex. Pre-B cells proliferate and expand, initiating light chain recombination. A fully formed BCR marks the immature B cell stage, at which the B cells enter circulation.* ***B)*** *Antibody-dependent development in secondary lymphoid organs. Circulating B cells first enter the spleen to complete the maturation process. The majority of B cells acquire co-expression of IgD on the surface and recirculate between lymphoid organs to encounter antigens and become activated. Activated B cells can differentiate into plasma cells (PC), which secrete antibodies or enter the follicle and form germinal centers. In germinal centers, long-lived memory B cells and plasma cells are created. Alternatively, mature B cells can enter the spleen's marginal zone (MZ) where they can become MZ B cells. This figure is inspired by Bonilla and colleagues [10]. This figure and subsequent figures were created in BioRender.com [21].*

The earliest stage committed to the B-cell lineage is the pro-B cell stage. At this stage, the immunoglobulin (Ig) genes are not yet expressed, and somatic recombination begins on the Ig genes on the heavy chain locus. The Ig V_H domain on chromosome 14 in humans (chromosome 12 in mice [22]) consists of three gene segments: variable (V), diversity (D), and joining (J). In humans, there are 55 V_H genes, 27 D_H genes, and 6 J_H genes, excluding pseudogenes [23]. These genes associate with 9 constant (C) genes to create different effector classes of B cells. Firstly, a D gene segment and a J gene segment are joined together through DNA double-strand break and repair. The pro-B cell then progresses to the pre-B cell stage and a V gene segment is joined to the DJ unit, forming a VDJ exon (Figure 2). The rearranged VDJ exon is transcribed together with the C_μ region and further processed into a complete μ heavy chain protein [4].

V(D)J recombination is mediated by the lymphoid-specific recombination activating gene 1 (RAG1) and RAG2 proteins. The RAG proteins bind to the site-specific recombination signal sequences (RSS) flanking each gene segment. Each RSS consists of a conserved heptamer and a conserved nonamer, separated by a 12bp or 23bp spacer sequence. A 23bp RSS is located 3' of each V gene, 5' of each J gene, and two 12bp RSSs flank both sides of each D gene. This organization ensures the correct joining of gene segments, preventing the binding of V to J directly on the heavy chain locus (the "12/23 rule") [24]. The RAG complex binds to one RSS, forming a signal complex, then binds to the complementary RSS, forming a paired complex. RAG nicks the DNA on a single strand between the heptamer and the coding sequence, resulting in a free 3' OH group. This 3' OH group then attacks the other strand, creating a double-strand break. This creates a hairpin coding end and a blunt signal end [25]. After coupled cleavage, the double-strand breaks are repaired by proteins of the non-homologous end joining (NHEJ) pathway. The signal ends are precisely joined together, creating the signal joint that is lost from the genome. On the coding joint, the hairpin must first be opened, creating palindromic (P) overhangs. Additionally, non-template (N) nucleotides may be added by the enzyme terminal deoxynucleotidyl transferase (TdT) to generate complementary sequences between two coding ends. N nucleotides can also be deleted during this process, resulting in imprecise joining of the coding joint [26], contributing to junctional diversity in V(D)J recombination.

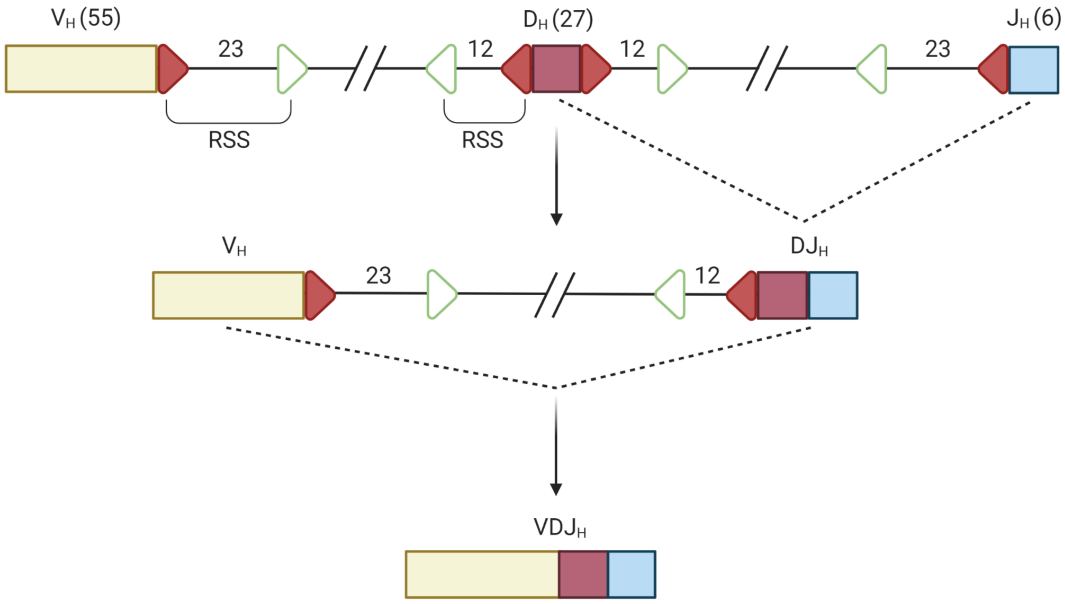


Figure 2: Somatic recombination on the heavy chain locus. *Each V_H, D_H, and J_H gene (Figure 3B) segment is accompanied by a recombination signal sequence (RSS), which consists of a heptamer (dark arrow) and a nonamer (light arrow), separated by a spacer sequence (12 bp or 23 bp). Recombination starts with DNA breakage and ligation between D_H and J_H, followed by V_H to DJ_H. The resulting DNA product is then transcribed into RNA and processed together with a C_H region to create the heavy chain protein. Figure modified from Jung and colleagues [27].*

The addition of nucleotides to the junctions can cause frameshifting, therefore, only ⅓ of V(D)J recombination events are productive. D to J rearrangement occurs on both alleles. If the first V(D)J rearrangement is productive, V(D)J recombination on the other allele is halted at the DJ stage. However, if the first attempt is unsuccessful, V(D)J rearrangement can be completed on the second allele. This phenomenon is termed "allelic exclusion" and it ensures only one receptor sequence can be expressed in a B cell [27]. B cells that fail both V(D)J recombination attempts are eliminated by apoptosis.

Successful expression of the μ heavy chain marks the transition into the pre-B cell stage. At this stage, the μ heavy chain is expressed on the cell surface, together with the invariant surrogate light chain (SLC) and the signal transducers Igα and Igβ to form the pre-B cell receptor (pre-BCR). This serves as the first important checkpoint in the development of B cells [28].

Pre-BCR signaling induces proliferation and expansion of pre-B cells, while at the same time downregulating the expression of the pre-BCR and initiating recombination of the light chain locus [29].

Recombination of the light chain locus follows the same principle as the heavy chain locus. There are two types of light chains: κ and λ. In humans, the κ locus is located on chromosome 2 and consists of 35 V_κ, 5 J_κ, and 1 C_κ functional genes, while the λ locus is located on chromosome 22 with 30 V_λ and 4 J_λ-C_λ functional genes [4]. Expression of the enzyme TdT is lower compared to during heavy chain rearrangement, therefore junctional diversity is also decreased. Recombination first occurs on the κ locus, and if unsuccessful on both alleles, occurs on the λ locus. The ratio between κ-containing and λ-containing antibodies in human serum is around 2:1 but can vary between isotypes and immunological conditions [30]. Successful light chain recombination results in the assembly and expression of the IgM BCR on the B cell surface.

The immature B-cell stage is marked by the expression of IgM on the cell surface. Surface-expressed BCR activates signaling pathways that keep the cell alive and inhibits RAG expression, preventing further rearrangement. If the BCR reacts to autoantigens in the bone marrow, then the B cell can undergo receptor editing or apoptosis in order to prevent autoimmunity [31]. In receptor editing, the light chain or heavy chain can go through RAG-dependent secondary rearrangement in order to eliminate autoreactivity [32]. Immature B cells that are not autoreactive leave the bone marrow and enter circulation, before entering the spleen to complete the maturation process.

In the spleen, B cells are divided into two subtypes: marginal zone (MZ) B cells and follicular B cells. MZ B cells are located at the interface between the white pulp and red pulp of the spleen and are capable of eliciting an antibody response independent of T-cell stimulation. MZ B cells do not recirculate and are distinguished by the high expression of CD21 [33]. The majority of B cells in the spleen, however, are follicular B cells. Due to alternative mRNA splicing, follicular B

cells are capable of expressing surface IgD with the same V(D)J sequence in addition to IgM. These mature B cells recirculate through secondary lymphoid organs such as lymph nodes, spleen, and mucosal-associated lymphoid tissues [10]. Follicular B cells associate with T cells and play an important role in the formation of GCs, which facilitate affinity maturation of B-cell receptors [20].

1.4.2.2 Structure and antigen recognition of the antibody

When B cells are stimulated by antigens, those B cells become activated and differentiate into plasma cells (Figure 1B). Plasma cells secrete a large number of antibodies, a glycoprotein product of alternative mRNA splicing on the Ig domain genes [34]. Antibodies are Y-shaped molecules consisting of 2 heavy chains (HC) and 2 light chains (LC) (Figure 3A). The LCs can be one of the two types: κ or λ, while the HC's isotype is determined by the constant domain genes: μ, ɣ, δ, α, and ε representing IgM, IgG, IgD, IgA, and IgE, respectively [4]. Each chain is made up of independent structures called Ig domains. Each Ig domain consists of 110 to 130 amino acids in an antiparallel β-strands formation, held together by disulfide bonds. A LC has 1 V domain and 1 C domain, whereas a HC has 1 V domain and 3 (in IgG, IgA, IgD) or 4 (in IgM, IgE) C domains [35]. Between the C_H1 and C_H2, there is an unstructured hinge region that is susceptible to enzymatic digestion, separating the antibody into 2 fragments: F(ab) (fragment antigen binding) and Fc (fragment crystallizable).

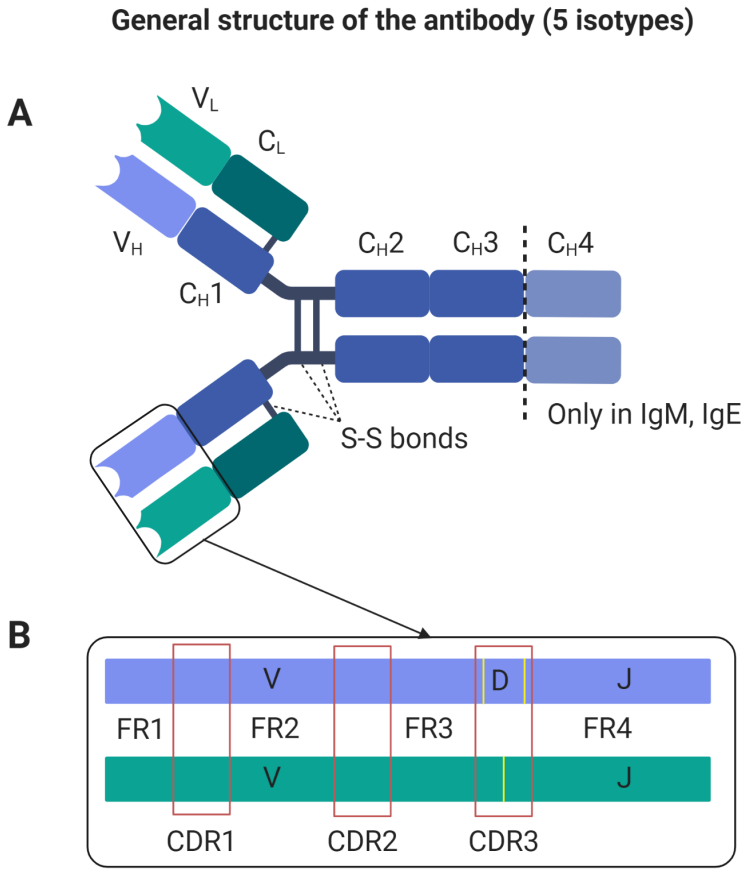


Figure 3: The structure of immunoglobulins. *A) Antibodies of all isotypes consist of two heavy chains (blue) and two light chains (green). Each chain contains a V domain (V_H or V_L) and C domains (C_L or C_H1–$3/4$). The chains are held together by disulfide bonds (S-S). B) Each V domain consists of conserved framework regions (FR) interspersed by hypervariable complementary determining regions (CDR). Red squares indicate the position of the CDRs and yellow lines delineate the borders between the V, D, and J genes. This figure is inspired by the work of Schroeder and colleagues [36].*

The effector function of an antibody molecule varies between isotypes and subtypes. This is mediated by the binding between the Fc region of the antibody (C_H2–$C_H3/4$) and the Fc receptor on different types of cells, such as phagocytes, NK cells, mast cells, and proteins of the complement system [37]. Specificity between an antibody and an antigen is determined by the F(ab) region of the antibody. The F(ab) region comprises the V_H and C_H1 domain on the HC and the V_L and C_L domain on the LC. Each V domain is further divided into 4 framework regions

(FR) interspaced by 3 complementary-determining regions (CDR). CDR1 and CDR2 lie on the V gene on both chains. CDR3 lies on the V-J junction on the LC and on the V-D-J junction on the HC (Figure 3B) [36].

The CDRs have a significantly higher amino acid sequence variation compared to the FRs and are positioned together to form a binding surface with the antigen. This allows antibodies to bind to a wide variety of antigens while retaining a common structure. CDR3 in particular has the highest degree of sequence variation, partly due to junctional diversity [38]. Therefore it has the largest contribution to antigen recognition and is the focus of much research into antibody specificity [39,40]. However, some residues on the FR region can also have a significant effect on antigen binding, evidenced by the loss of binding affinity during the grafting of CDRs from mouse to human antibodies [41].

Antibodies are glycoproteins and each isotype possesses a unique glycosylation pattern. This pattern plays an important role in maintaining structural integrity, modulating effector functions, and can also influence antigen binding [42].

1.4.2.3 Class switching and affinity maturation of B-cell receptors

Activated B cells can either differentiate into PCs and produce antibodies with low affinity or enter follicles of secondary lymphoid organs to form GCs, where class switch recombination (CSR) and somatic hypermutation (SHM) occur (Figure 1B). In CSR, the enzyme activation-induced deaminase (AID) targets the switch sequence 5' of the C gene and converts cytidine to uracil. Enzymes of the base excision pathway remove uracil and create nicks on both strands of the DNA, leading to dsDNA breaks. The broken DNA ends from the 2 switch regions are then joined by NHEJ, bringing the VDJ sequence closer to the new C gene. As a result, the B cells lose the expression of IgM and IgD and gain the expression of any of the other isotypes [43,44]. CSRs are regulated by cytokines and different cytokines induce class switching to different isotypes [45].

At the same time as CSR is occurring, the V domain coding sequence is further diversified through SHM in the dark zone of the GC. In SHM, AID also plays an important role in creating single nucleotide substitution mutations targeting the CDRs, further contributing to sequence variation at those regions [46]. Subsequently, B cells migrate to the light zone of the GC where it is presented with an antigen sequestered by follicular dendritic cells (FDC) and receives necessary survival signals from the Tfh cells. B cells compete with each other to bind antigens and present antigens on MHC II to the Tfh cells [47]. B cells that show low affinity to antigens or autoreactivity are eliminated via apoptosis while those that show high affinity differentiate into PCs or memory B cells and exit the GC. A fraction of light zone B cells with improved affinity is returned to the dark zone for further SHM and affinity maturation [48]. As a result, the average affinity of antibodies to a particular antigen is increased over time.

1.4.2.3 Classes and effector functions of immunoglobulins

The different isotypes of antibodies vary in size, the flexibility of the hinge region, ability to activate the complement system, accessibility through surfaces, and effector functions in response to antigens (Figure 3A). Antibodies are produced and decay at a constant rate, regardless of antigen specificity. Therefore, the serum concentration of antibodies is maintained throughout life [49].

IgM is the first isotype to be expressed in B-cell development in monomeric BCR form. Due to the fact that IgM is expressed early without extensive somatic mutation in response to antigens, IgM antibodies bind to a wide variety of antigens with low affinity [36]. In serum, IgM is secreted as a pentamer linked together by disulfide bonds and a J chain. This gives IgM enhanced avidity by binding to multiple sites, especially to repeating epitopes [50]. IgM acts primarily in the primary immune response by opsonizing antigens and activating the complement system [4].

Similar to IgM, IgD is also expressed early in B-cell development in BCR form and secreted in monomeric antibody form in serum, albeit in low levels. IgD has a long hinge region, giving it

higher flexibility to bind antigens with low-density epitopes while at the same time making it more susceptible to proteolytic cleavage, resulting in a very short serum half-life [51]. Effector functions of IgD mainly involve binding to basophils and mast cells, triggering the production of antimicrobial peptides and induce inflammation [52].

IgG is the most abundant antibody isotype in the serum (70–75% of serum antibody) and is divided into 4 subclasses: IgG1, IgG2, IgG3, and IgG4, in order of decreasing prevalence [9]. The different subclasses of IgG exhibit variation in flexibility, susceptibility to cleavage, and effector functions, with IgG1 and IgG3 responding mainly to protein antigens while IgG2 and IgG4 responding mainly to polysaccharide antigens [36]. All of them are able to cross through the placenta and dominate the secondary immune response. IgG can also bind with the FcRn receptor and diffuse into extravascular sites [53]. IgG antibodies bind to antigens with high affinity, neutralizing the pathogens or toxins, and initiate the complement cascade (with the exception of IgG4) [34].

IgA is the second most common antibody isotype after IgG in circulation but it is the predominant antibody isotype in mucosal surfaces and external secretions. IgA exists as monomers in the serum and as dimers or oligomers, linked together by the J chain in mucosal surfaces [54]. There are 2 subclasses of IgA: IgA1 with a longer hinge region, higher proteolytic cleavage susceptibility and comprise the majority of serum IgA; IgA2 with a shorter hinge region, higher resistance to proteases and predominates mucosal secretions [37]. IgA protects the body's mucosal surfaces from pathogen binding and is the principal component of the colostrum, which provides crucial protection for newborns [55].

IgE is present in the serum at the lowest concentration and shortest half-life out of all the isotypes but with a very potent effector function. IgE binds with extremely high affinity to receptors on mast cells, basophils, and eosinophils. Once bound, IgE can upregulate the expression of these receptors for an extended period of time [56]. As a result, IgE contributes in the defense against parasites and is often a target for therapies in allergies and asthma [57].

1.4.2.4 Diversity of the antibody repertoire

The paradigm for the generation of antibody diversity, established by Tonegawa in 1983 [58], is governed by stochastic mechanisms explained above: somatic recombination, imprecise V-(D)-J joining, insertion or deletion at junctions, and SHM. As a result, the theoretical diversity of the naive antibody repertoire in humans is estimated to be at least 10^{12} [59]. However, with the advent of new technologies and an increase in data throughput, new evidence has been uncovered depicting a more deterministic and biased process: factors such as genetic background [60] and previous exposure to antigens [61] can also have a significant impact on the diversity of the antibody repertoire. All these factors are in a dynamic balance with varying degrees of contribution during different stages of development in B cells [62].

Owing to the massive theoretical diversity of antibody generation, it has long been assumed that an individual's antibody repertoire is overwhelmingly unique (termed "private"). Large-scale analyses revealed that, on the contrary, a notable fraction (>1%) of antibody sequences is shared between individuals (termed "public") [63–66]. An individual's clonal diversity and distribution serve as a fingerprint of their current immunological status and thus contain highly useful information for diagnostics [67]. This further emphasizes the importance of evaluating and quantifying the diversity of the antibody repertoire. By applying bioinformatics methods to analyze immunosequencing data, it is now possible to build an immune repertoire diversity profile consisting of a multitude of singular diversity indices. This can provide valuable insights into people's immunological status at an unprecedented resolution, enabling a systematic, data-driven approach to disease detection and prevention [68].

There are several important characteristics that can be evaluated in repertoire analysis: germline V-gene usage, clonal expansion, clonal diversity, and repertoire size [62]. In particular, clonal diversity is defined as the variation in the amino acid sequence of CDR-H3, calculated using Shannon information entropy. Shannon entropy, designated "H", is a versatile and widely utilized tool to measure sequence variability [69].

1.5 Sequencing of the B-cell receptor repertoire

1.5.1 Current B-cell receptor repertoire sequencing methods

Up until the early 2000s, immunosequencing research was mostly restricted to a throughput of several hundred B cells per run, due to the cost and labor requirement of Sanger sequencing [70–72]. Therefore, it was only possible to sample a minuscule fraction of the BCR repertoire, limiting the conclusions that could be drawn. With the advent of high-throughput sequencing (HTS), Sanger sequencing is now primarily relegated to validating HTS results [73]. HTS offers a more comprehensive look into the diversity of the immune repertoire owing to the high volume of data generated. Multiple sequencing platforms are currently available, such as Illumina, Ion Torrent, PacBio, and Oxford Nanopore. Each platform comes with its own advantages and disadvantages, primarily concerning the read length and error rates [74–76]. The Illumina MiSeq platform has a read length of 300–600 bp for paired-end sequencing. Considering the full-length variable regions being ~350–420 bp long, this restricts the number of options for primer design and library preparation methods [77]. Ion Torrent's platform offers comparable performance with a rapid turnover time, albeit with a higher error rate (1.78% compared to <0.4% in Illumina's platform), making it suitable for clinical settings and less error-sensitive applications [78]. Long-read sequencing platforms, such as PacBio and Oxford Nanopore, can be useful in receptor chain pairing workflows that rely on overlap extension PCR [79], but similarly suffer from lower throughput, higher error rates, and higher cost [78,80]. Overall, Illumina is currently the platform of choice for immunosequencing research because of its high fidelity, high throughput, and comparably low cost [81].

Illumina sequencing belongs to the category of short-read sequencing and sequencing by synthesis, in which a polymerase is utilized and fluorophore-tagged nucleotides provide the signal to identify the incorporation of a base into a DNA template [82]. The first step in the sequencing process is the ligation of common adapters for amplification and sequencing. The adapters contain complementary sequences to the two types of oligos immobilized on the flow cell, index sequences to distinguish between samples, and the binding site for sequencing

polymerase (Figure 4A). Next, each DNA molecule is amplified on the flow cell. The template hybridizes with the first type of oligos, then a polymerase creates a complementary strand of the template. The double-stranded molecule is then denatured and the original template is removed. Then, the adapter on the other end of the newly synthesized sequence interacts with the second type of oligos on the flow cell and the sequence is amplified in a process termed bridge amplification. The bridge structure is subsequently denatured and the process is repeated in order to generate clusters of amplified templates (Figure 4B). The purpose of this process is to ensure the signal is strong enough to be detected during imaging. After enrichment of the template, DNA polymerase, sequencing primers, and modified nucleotides are added into the flow cell. These nucleotides contain a fluorophore specific for each type of base and are modified to terminate the reaction after each incorporation. This ensures that only one nucleotide is added in each cycle. After base incorporation, the unbound nucleotides are washed away and a laser is used to induce the fluorophore at specific wavelengths to identify the incorporated base. Once the base is bound, the fluorophore is cleaved and the 3' terminal is reversed so that the cycle can begin anew. After the first read is completed, the index sequence is sequenced and, in paired-end sequencing, the sequencing starts again from the other end. This process occurs simultaneously in hundreds of millions of clusters on the flow cell (Figure 4C).

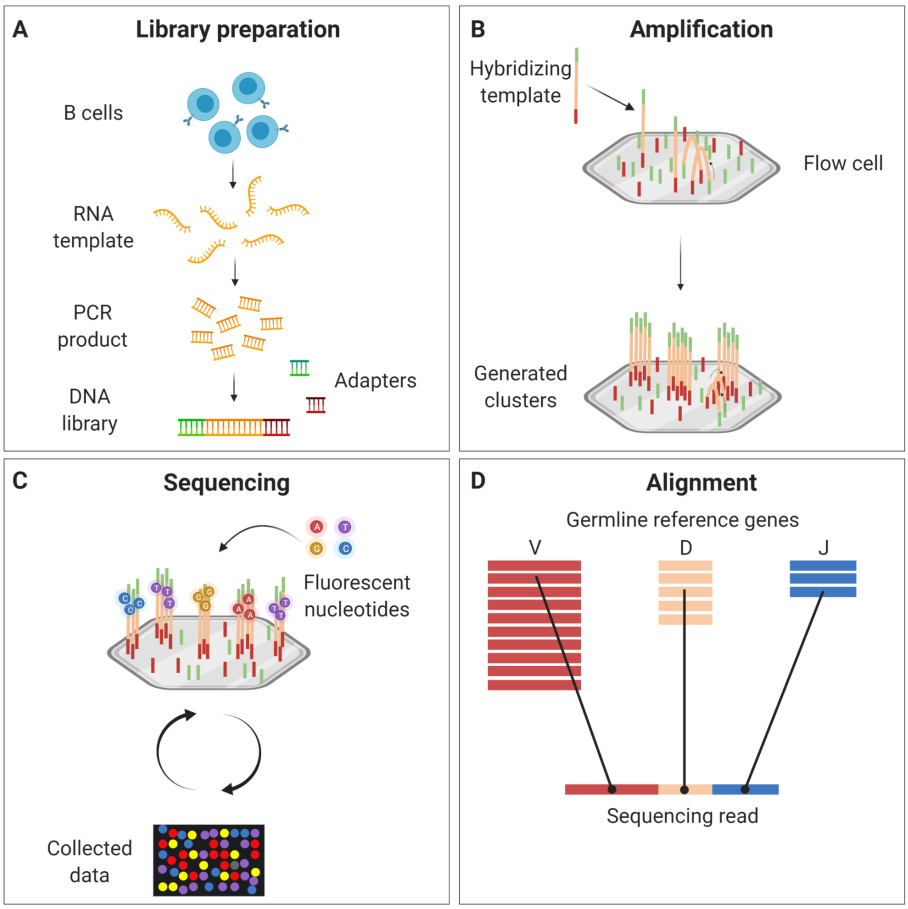


Figure 4: Overview of the Illumina sequencing platform. ***A)*** *The sequencing library is prepared by PCR amplification of the template and ligation of specialized adapters to both ends.* ***B)*** *The library is loaded into a flow cell and hybridizes to the oligos on the surface. Each bound template is then amplified into a cluster through bridge amplification.* ***C)*** *Sequencing reagents and fluorescently labeled nucleotides are added to the flow cell and one base is incorporated. An image of the flow cell is captured and the fluorescence signal is recorded at each cluster. This cycle is repeated a number of times equal to the desired read length.* ***D)*** *Bioinformatic tools align the reads with the reference sequence to construct the complete sequence. In immunosequencing applications, alignment is conducted using the sequence of germline genes as reference. Then, the difference between the reference and the sequenced reads can be analyzed.*

Most Illumina sequencing platforms utilize 4-channel chemistry, in which each nucleotide base is assigned a distinct fluorophore for identification. Newer platforms, such as NextSeq 550 and Miniseq, utilized 2-channel chemistry, where only 2 distinct fluorophores are used. This helps reduce the sequencing time and cost while maintaining similar accuracy. More recently, a

1-channel chemistry system was introduced where only one fluorophore is used but modified differently for each base and having two imaging steps per sequencing cycle [83]. After the sequencing data is obtained, the quality of the base call is assessed. The standard scoring system widely used at the moment is Phred. The Phred score Q is defined as $Q = -10 \times \log_{10}(P)$ where P is the probability of an incorrect base call [84]. A Q score of 30 indicates a 1/1000 probability of incorrect base calling and is commonly used as a benchmark for sequencing quality. Sequences from different samples are distinguished by their respective indices and sequences from the same sample are clustered together. These clusters are then mapped to the reference sequence in a process called sequence alignment. This allows the complete sequence to be reconstructed and any variation to be detected (Figure 4D).

For immune repertoire analysis, sequencing accuracy is critical. However, errors can be introduced during library preparation and during sequencing. Several library preparation strategies depend on targeted enrichment from gDNA or mRNA, going through several amplification steps, such as reverse transcription and multiplex PCR [77]. To minimize errors during amplification, the use of high-fidelity polymerase is often utilized [85]. In addition, various methods have been developed to correct for sequencing errors, such as replicate sequencing, sequence clustering and most notably, unique molecular identifiers (UMI) [77]. UMIs are RNA or DNA molecules with degenerate nucleotide sequences that are incorporated into a gene-specific primer. When reverse transcription takes place, UMI-tagged cDNA molecules are created [86]. After sequencing is performed, reads are grouped together based on their UMIs, and a consensus sequence is built based on the assumption that reads sharing the same UMI originated from the same mRNA molecule. With this, errors originating from sequencing may be significantly reduced and quantification of transcript abundance can also be improved by counting the number of UMIs instead of reads [87]. However, for UMI-based error correction to be effective, the sequencing depth needs to be sufficiently high, which can be challenging for large and highly diverse populations such as B cells [63]. In addition, UMI further increases the sequence length of libraries, which can negatively affect read quality [88] .

Bulk sequencing of B-cell receptors can yield a large amount of useful data and has been applied extensively in immunological research following the work of Weinstein and colleagues in 2009 [89]. However, the native light chain and heavy chain pairing is lost due to the fact that the transcripts come from different chromosomes [4,22]. This poses a challenge in determining antigen specificity because the pairing of heavy and light chains contributes significantly to antibody-antigen interaction [90]. To overcome this limitation, novel methods for single-cell library preparation, such as Smart-seq [91], Drop-seq [92], and commercial platforms such as 10X Genomics Chromium have been developed.

1.5.2 Single-cell sequencing of B-cell receptor repertoires

Single-cell sequencing enables the characterization of gene expression at the individual cell level, revolutionizing transcriptomic studies. Single-cell BCR sequencing offers a solution to the chain-pairing problem unaddressed by bulk sequencing. To isolate single cells, various methods have been developed. Limiting dilution is performed using standard pipetting tools and relies on statistical distribution to achieve single-cell concentrations. Typically, only about 1 in 3 wells will contain cells and require confirmation by a microscope, making this method laborious and low throughput [93]. Laser capture dissection and manual picking by micromanipulation are often utilized on fixed, frozen, or solid tissue samples. Similarly, these methods have very low yield and are uncommon in immunosequencing workflows [94]. Fluorescence-activated cell sorting (FACS) systems rely on fluorescence tagging of cells using specific antibodies. These tagged cells then move through a laser beam with an optical detector and single cells can be separated into tubes or microtiter plates, scalable up to hundreds of cells [93]. Major drawbacks of this method include the use of specific monoclonal antibodies and the inability to scale down the reaction volume below microliter levels, resulting in a higher reagent cost per cell [95]. Microfluidics technology, especially droplet-based microfluidics, has gained popularity in recent years. Each cell flowing through the microfluidics chip is isolated in an aqueous nanoliter-sized droplet containing all the necessary reagents, with the droplets surrounded by oil to prevent mixing [96]. Due to its low-volume nature, droplet-based microfluidics offers distinct

advantages, such as low reagent consumption and low analysis cost, reduced risk of contamination, scalability, and throughput up to thousands of cells per second [93].

At the present time, numerous workflows for single-cell RNA sequencing (scRNA-seq) have been developed [97], with the first method demonstrated by Tang and colleagues in 2009 [98]. Since then, various approaches to single-cell library preparation have been explored, including the use of template-switching reverse transcription (Smart-seq) [91]; capturing of cells in droplets with uniquely barcoded primer beads (Drop-seq) [92]; the pairing of light and heavy chain through emulsion linkage RT-PCR (DeKosky 2013) [99]; linear amplification of mRNA using in vitro transcription (CEL-seq) [100]; full-length capture of the variable region using both molecular and droplet barcodes without the use of beads (Briggs) [101]. Furthermore, novel methods are being actively developed and the majority of immune repertoire data is expected to come from single-cell sequencing data in the coming years [81]. A comprehensive review of currently available single-cell sequencing platforms was conducted by Brown and colleagues of our research group [97].

Due to the low amount of starting materials, scRNA-seq suffers from a higher degree of technical variation compared to traditional bulk sequencing. Factors including low capture efficiency, bias in transcript coverage, and inclusion of dead cells or multiple cells in one droplet can all affect the quality of the sequencing data [102]. Therefore, quality control (QC) is especially important in scRNA-seq. Low-quality reads should be filtered out using software packages such as FastQC or pRESTO [103,104]. Setting a cutoff value to remove clones with low abundance helps to reduce artificial diversity originating from sequencing errors [63]. However, modern sequence alignment tools such as MiXCR can rescue low-quality reads via clustering and reduce data loss from sequencing errors[3]. Additionally, the use of synthetic spike-ins is also increasingly common in scRNA-seq in order to assess sequencing quality and accuracy [105].

[3] Sequence alignment with MiXCR is introduced in section 3.4.3 “Read assembly and clonotyping”.

1.5.3 Applications of B-cell receptor repertoire analysis

Advances in BCR sequencing have greatly expanded the possibilities for immunological research in many different areas, including basic research in adaptive immunity [106], development of novel personalized diagnostics and therapeutics [67,97], and understanding the immune response to infections and diseases [107,108].

HTS data have shown that the immune repertoire is not distributed uniformly and that gaps exist in the recognition ability of the adaptive immune system due to central and peripheral tolerance mechanisms [106]. This information is important in immunoengineering not only for designing antibodies with minimal cross-reactivity to self-antigens but also for the generation of rare neutralizing antibodies that are usually eliminated in vivo [109].

Applications in personalized and precision medicine rely on accurate genotyping of the antigen receptor sequence. The high volume of data generated by HTS makes it suitable for coupling with computational approaches. With machine learning, useful information can be extracted from the receptor sequence data, such as immunological status in the past and present of an individual, leading to a more tailored experience in diagnostics and treatments [97]. In addition, disease progression monitoring, particularly leukemia, can be conducted with much higher sensitivity compared to other methods by identifying receptor rearrangements unique to cancer cells and quantifying the abundance of cancer-specific clones during the course of treatment [67].

BCR sequencing data analysis has also proved to be highly useful in clinical immunology and immune response to infectious diseases. Singh and colleagues discovered that B cells producing autoantibodies exhibit mutations similar to those in lymphoid malignancies which help pathogenic B cells to avoid elimination at immune checkpoints [107]. More recently, Schultheiß et al. reported that in COVID-19 patients, the B-cell response exhibits distinctive signatures and that the degree of SHM can be associated with the severity of the disease [108]. These findings provide important insights into understanding the role of the adaptive immune system in disease pathogenesis and progression.

1.6 Characterization of serum antibody repertoire by mass spectrometry

Despite the fact that BCR sequencing can reveal important information on the nature of the immune system, the immunological protection and memory function of the humoral immune response depends on the circulating antibodies produced by plasma cells. Even with advances in deep sequencing of the BCR repertoire, a sizable fraction of serum antibodies is still missing when compared with proteomic results [110]. In addition, it has been reported that the number of antigen-specific B-cell clones (10^9) does not match the number of distinct circulating antigen-specific antibodies (10^5–10^6) [111]. Therefore, the ability to deconvolve the serum antibody repertoire is essential in many applications. However, the antibody response against antigenic stimulation is both temporal and diverse, making the task more challenging [112]. In order to extract useful information, there is a need to resolve the mixture of antibodies into distinct clonotypes. However, until recently there is no good information about the identification of specific antibodies in serum. This is due to the very high sequence identity between antibodies in the framework region making identification based on peptide mapping difficult [113]. Therefore, studies so far have mostly focused on antigen-specific antibodies after stringent purification and enrichment to reduce sample complexity [114,115].

1.6.1 Approaches for mass spectrometry-based antibody proteomics

Two main approaches to identify and characterize antibodies using mass spectrometry (MS) are the “bottom-up” and the “top-down” approaches. In bottom-up MS, antibodies in a polyclonal mixture or in the monoclonal form are first digested by proteolytic enzymes into peptides, separated by chromatography, and analyzed in a mass spectrometer. By contrast, top-down MS utilizes intact antibodies for analysis without prior digestion. Of the two approaches, bottom-up MS has been more widely utilized to date since it offers several distinct advantages. Digested peptides offer better separation efficiency due to higher solubility, higher throughput, and a straightforward approach compatible with multiple mass spectrometers [116]. In addition, there are multiple commercially available software and data analysis tools compatible with bottom-up MS. However, bottom-up MS cannot achieve full sequence coverage, due to the nature of enzymatic digestion. Additionally, a portion of post-translational modifications (PTM) is lost due

to incomplete sequence coverage [117]. This is noteworthy since PTMs of antibodies have been the subject of study in assessing therapeutic monoclonal antibodies (mAbs) [118].

Top-down MS promises to resolve the shortcomings of bottom-up MS with the potential for a more comprehensive antibody coverage [119]. However, protein solubility remains a challenge, especially for larger biomolecules such as antibodies. Surfactants are commonly utilized to improve solubility although they can interfere with the ionization process, requiring removal or replacement prior to MS analysis [120]. Top-down MS also suffers from lower sensitivity and difficulties in protein identification, requiring higher mass accuracy and resolution to discriminate the ion species correctly [121,122]. Furthermore, since data from top-down MS is incompatible with bottom-up MS workflows, there is a lack of software support for top-down antibody MS, although this is being addressed [123]. Therefore, most top-down applications are limited to the characterization of single mAbs or antibodies in simple mixtures with low throughput since both chemical separation and spectral assignment of different antibodies would prove too complicated otherwise.

In addition to these two approaches, there is also “middle-down” MS, where mAbs are analyzed after digestion of the hinge region [124] or are deglycosylated prior to MS [125]. The approaches have attempted to gain the benefits of top-down and bottom-up MS, albeit with mixed results [121]. While middle down MS provides significantly better coverage, sensitivity, and accuracy, it eschews the benefits of reduced sample processing and native chain pairing information [126].
In short, either approach can be suitable depending on the purpose of the study. Bottom-up MS can be utilized for antibody identification in complex mixtures, while top-down MS is used in PTM studies and tracking of single antibodies, which can provide information on the bioavailability and performance of therapeutic mAbs for quality control purposes [127].

1.6.2 Liquid chromatography tandem mass spectrometry in antibody proteomics

Mass spectrometry, particularly liquid chromatography with tandem mass spectrometry (LC-MS/MS) has allowed researchers to identify and decode a wide variety of biological

molecules and antibodies are no exception. Mass spectrometry is most efficient in very small flow rate (0.05–0.2 mL/min) [128]. Therefore high-performance liquid chromatography (HPLC) is most often the separation method of choice in a LC-MS/MS system (Figure 5A). Reverse-phase chromatography columns made up of C18 alkyl ligands immobilized to silica beads are often utilized due to the absence of salt in the elution process, which if present can interfere with the ionization process. Furthermore, the low elution volume provides better separation and subsequently superior resolution in the mass spectra [129]. Intact antibodies are enzymatically digested in order to create peptide fragments suitable for mass spectrometry analysis. Trypsin is the protease of choice for the vast majority of mass spectrometry workflows since it is very efficient, widely available, and affordable. Other proteases, such as chymotrypsin and LysC, with their own unique characteristics, are also commonly utilized albeit to a lesser extent [130]. The resulting peptides from protein digestion are separated by chromatography and loaded into the inlet of the mass spectrometer, where they are nebulized in a highly charged electric field. These charged droplets are sprayed against a stream of inert dry gas, which rapidly reduces the size of the droplets. Once the charge density reaches the Rayleigh limit, the droplets get torn apart, creating gas-phase ions that are directed into the mass analyzer [131] (Figure 5B). Under high vacuum, the ions travel through the mass analyzer and the mass-to-charge (m/z) ratios are recorded in the detector. Additionally, peptides can be selected to undergo collision induced dissociation (CID), in which ions collide with inert gases in order to create smaller fragment ions, the m/z of which form the basis of peptide sequence determination [128] (Figure 5B).

Deciphering the antibody repertoire in the serum requires reference data gathered by HTS of the BCR repertoire, since circulating antibodies have gone through V(D)J recombination and SHM, therefore unique for each individual. When combined with the paired VH:VL sequencing approach previously described, serum antibody mass spectrometry allows the complete reconstruction of the antibody peptide sequence. This would not only allow researchers to better analyze the humoral immune response against antigens but also aid the discovery and development of antibodies to be used in therapeutics [132]. Although previous research has

performed MS/MS antibody repertoire analysis [113,114,133], no standardized pipeline with evaluations from experimental work to computational analysis exists in this particular field of study. As a result, there remain many outstanding challenges in antibody proteomics. For instance, it is still unclear what is the minimum concentration of an antibody in a sample that can still be reliably identified using mass spectrometry. However, new methods are being developed in order to increase the sensitivity and accuracy of antibody analysis [127]. In addition, due to the low dynamic range of mass spectrometry, the presence of highly abundant antibodies may affect the detection of rare antibodies. Highly similar antibodies are also difficult to distinguish and rely on the identification of unique peptides. Furthermore, integration of BCR sequencing into antibody proteomics can shed light on the variation between genotype and phenotype of the immune repertoire, and possibly open the way to linking antibody sequence and function.

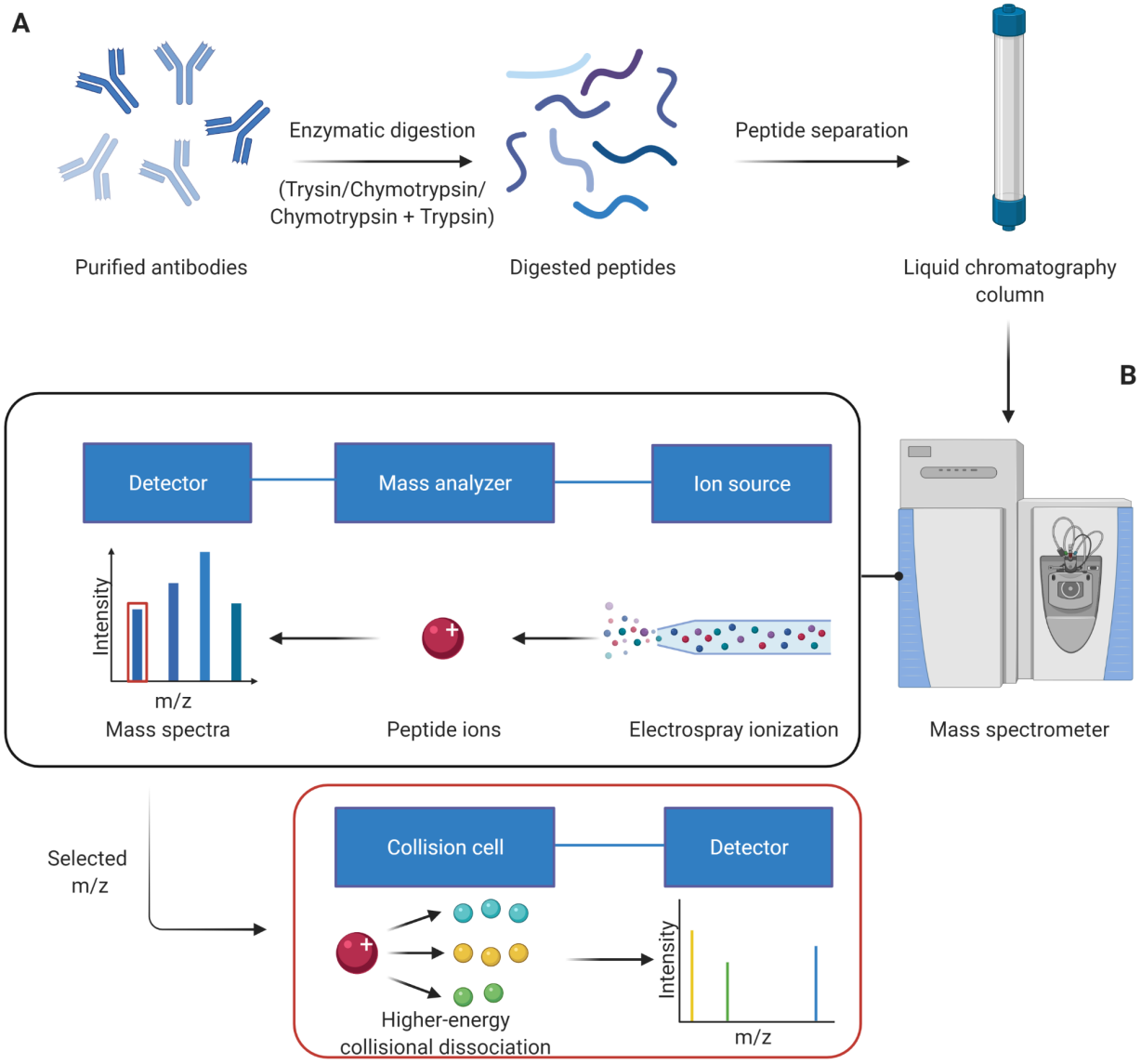


Figure 5: Schematic of a LC-MS/MS workflow in antibody proteomics. ***A)*** *Purified antibodies are digested with enzymes (e.g., Trypsin/Chymotrypsin) to create mass spectrometry-compatible peptides. The resulting peptides are then separated by high-performance liquid chromatography (HPLC) and forwarded into the mass spectrometer.* ***B)*** *In the mass spectrometer, an electric field is applied and the peptides are sprayed into charged droplets. The droplets are converted into charged peptide ions and travel through the mass analyzer, and the mass to charge (m/z) ratio is recorded in the detector. In addition, collision induced dissociation (CID) can be performed on selected peptides to produce secondary ions for antibody sequence determination.*

2. Thesis aims

The BCR repertoire is the product of antigen-driven clonal expansion and selection and represents the genotype of the adaptive immune system. The antibody repertoire, however, is the executor of immune functions regarding infection and diseases and represents the phenotype of the adaptive immune system. Thus, understanding the relationship and connection between these two repertoires is crucial in understanding the function of the immune system. At present, there exists no pipeline that provides a high volume of information on genomic and phenotypic antibody diversity at the single-cell (B cell) and single-molecule (antibody) level. Therefore, our goal for this thesis project is to develop a pipeline that allows high-throughput dissection of antibody repertoire diversity at the genomic and proteomic level (Figure 6). Specifically, this thesis aims to achieve scientific advances in the following areas:

1. Adaptation of a reliable protocol for bulk BCR sequencing that maximizes the coverage of the variable region sequence.
2. Establishment of a workflow for single-cell sequencing that allows the recovery of native VH:VL chain pairing in B cells.
3. Building a foundation for the dissection of the antibody repertoire at the proteomic level by benchmarking existing LC-MS/MS and peptide sequence identification tools.
4. Creating a combined pipeline of experimental work and bioinformatic data analysis.

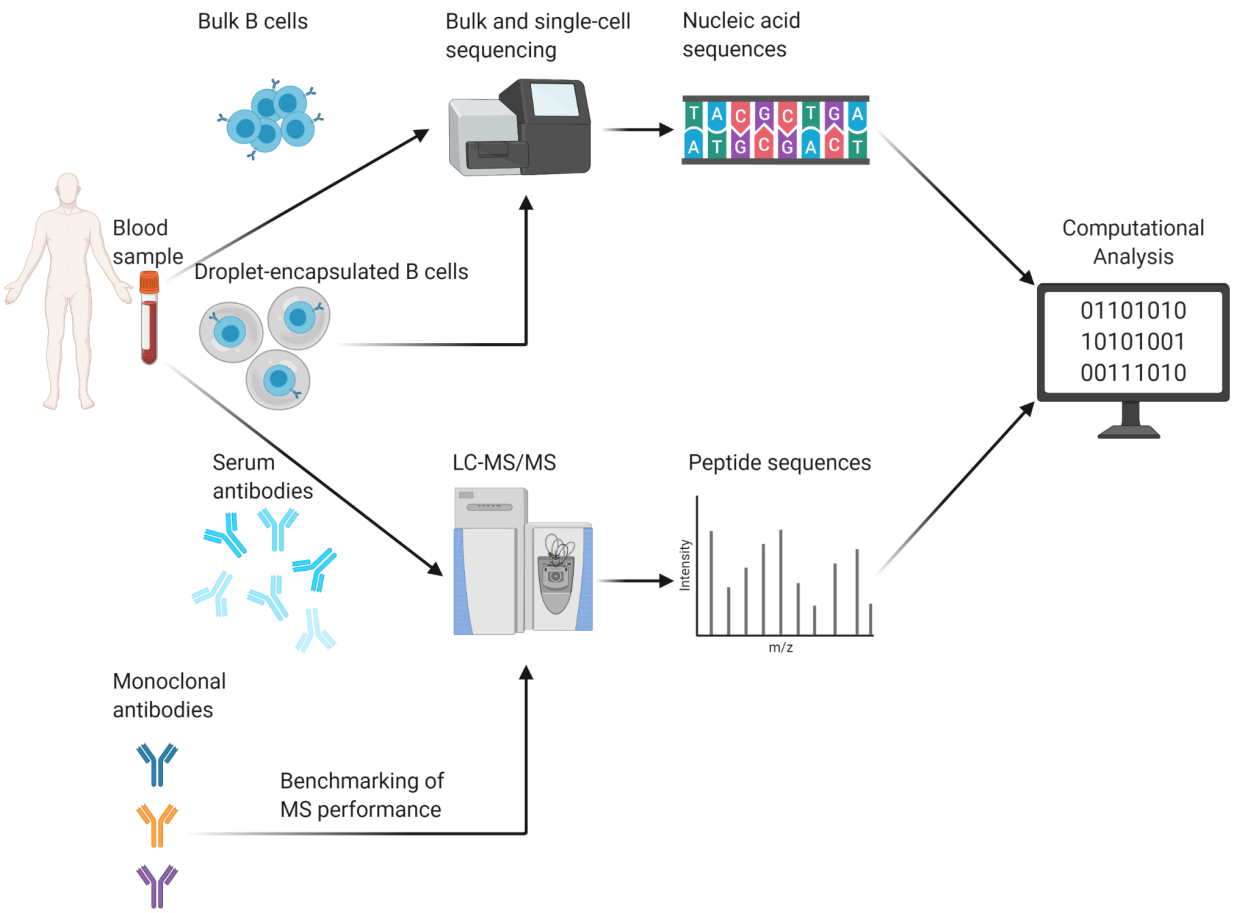


Figure 6: Overview of the thesis aims.

3. Methods

3.1 Bulk B-cell receptor sequencing

Library preparation for bulk BCR sequencing started with B cells isolated from whole blood using magnetic separation. Using RNAs isolated from B cells, cDNAs were synthesized using isotype-specific primers with the addition of UMIs. The cDNAs were then amplified using 5' multiplex (MTPX) primer sets for V genes in the heavy chain and light chain (κ or λ). Finally, Illumina adapters were added with a unique index sequence for each sample (Figure 7).

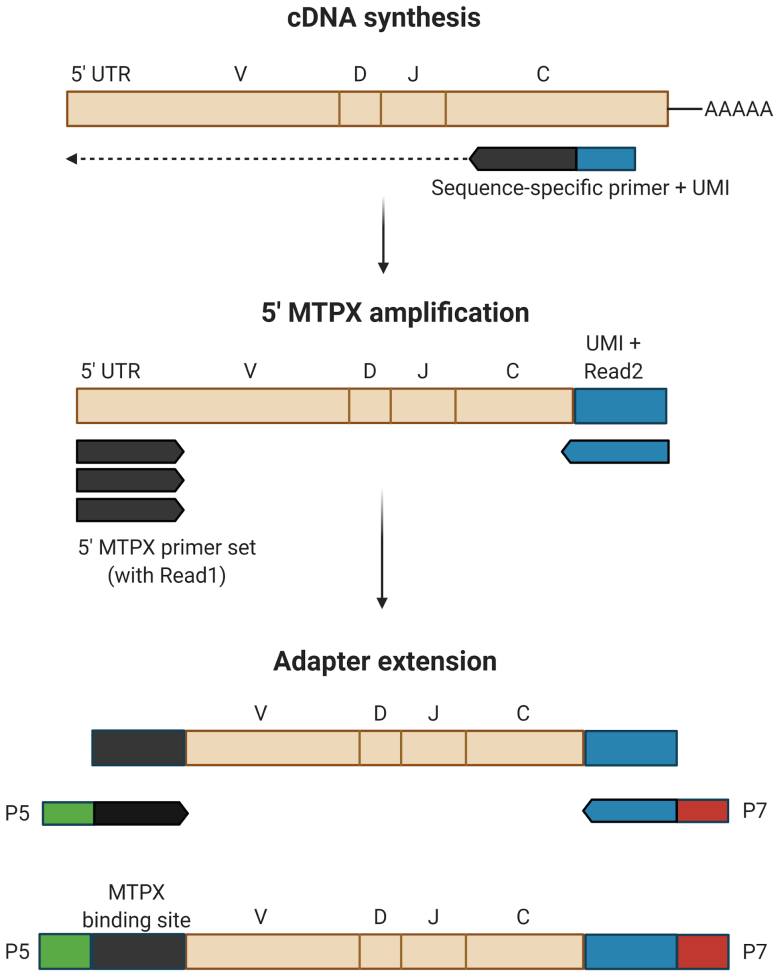


Figure 7: Summary of the bulk B-cell receptor sequencing library preparation workflow. *Isolated RNAs from B cells were reverse-transcribed into cDNAs using isotype-specific primers containing unique molecular identifiers (UMI). The cDNAs were amplified using a set of multiplex (MTPX) PCR primers for heavy and light chain sequences. Finally, Illumina adapters with indices were added in by limited-cycle amplification.*

3.1.1 B cell isolation from peripheral blood

B cells were isolated from whole blood using the MACXpress Whole Blood B Cell Isolation Kit by Miltenyi Biotec according to the manufacturer's instructions. This method removed non-target cells by binding them with magnetic beads and separating them from the solution using strong magnets. The red blood cells were sedimented at the bottom, leaving only B cells in the plasma. The B cells were subsequently recovered via centrifugation, resulting in pure B cell pellets.

3.1.2 Cell counting

The Countess II Automated Cell Counter utilized trypan blue chemistry and bright field microscopy to calculate the concentration and viability of cells in a sample. Cell suspension and trypan blue 0.4% were mixed in a 1:1 ratio and loaded into the counting chamber. Since trypan blue cannot cross the membrane of live cells, only dead cells would be stained blue. The concentration of B cells in the sample was calculated based on the image captured by the cell counter.

3.1.3 RNA isolation

RNAs from isolated B cells were extracted and purified using the RNeasy Plus Mini Kit from Qiagen following the manufacturer's instructions. The cells were lysed and homogenized in a denaturing buffer to inactivate RNases. Then ethanol was added to help bind RNA to the column while contaminants were washed away with washing buffers. Pure RNA was then eluted with an elution buffer and the concentration measured by Nanodrop.

3.1.4 Determination of nucleic acid concentration

The concentration and purity of nucleic acids were measured using the NanoDrop 1000 Spectrophotometer from Thermo Fisher Scientific. The system measured the concentration of nucleic acids using absorbance at 260 nm, and protein at 280 nm, based on Beer-Lambert's law. Sample purity was assessed using the A260/A280 ratio and should be around 1.8 for DNA and

around 2.0 for RNA. This system allows rapid quantification of nucleic acids while minimizing sample loss (only 1 µL of sample is required).

3.1.5 cDNA synthesis

The RNAs isolated from B cells next underwent reverse transcription into cDNA. This was done using isotype-specific reverse transcription primers that targeted the 3' constant region (Figure 7). The primers also included a UMI sequence for error correction and an adapter sequence partially overlapping with Illumina's Read2 sequencing primer (Supplementary Table 1).

Procedure:

1. Measure RNA concentration and use 200 ng as input.
2. Create a reaction mix:
    a. 200 ng RNA (X µL)
    b. 1 µL isotype-specific primer
    c. 1 µL dNTP 10mM
    d. 12.5 - X µL H_2O
3. Incubate for 5 minutes at 65ºC and then place on ice.
4. Add to the sample, in each reaction:
    a. 4 µL 5X Reverse Transcription buffer
    b. 0.5 µL Ribolock
    c. 1 µL Maxima Reverse Transcriptase enzyme
5. Incubate for 30 minutes at 50ºC, 5 minutes at 85ºC, hold at 4ºC.
6. Proceed to DNA purification with Qiagen's MinElute PCR Purification.

3.1.6 DNA purification

The DNA product was purified using Qiagen's MinElute PCR Purification system, which is based on silica membrane purification. The DNA was bound to the silica membrane in the column, while excess reagents, primers, and other contaminants were removed by washing with ethanol. The DNA products were then eluted with 10mM Tris.Cl, pH 8.5, resulting in cleaner, purer DNA for subsequent reactions.

3.1.7 Multiplex PCR

The purified cDNA was amplified in a multiplex PCR reaction with the 3' primer targeting the Read2 sequence ligated during cDNA synthesis and a 5' forward primer mixture (Figure 7). The 5' mixture contained multiple primers targeting the 5' leader sequence of the V genes and a tail overlapping with the Illumina Read1 sequencing primer (Supplementary Table 2). The 5' primers were divided into 3 categories: VH for all the heavy chain isotypes, VK for the κ light chain, and VL for the λ light chain. This was done in order to capture most of the possible V gene sequences present in the sample.

Procedure:

1. Create the 5' forward primer mix: Add 1.0 µL of all the primers in a set (VH, VK, VL) into one microcentrifuge tube, mix well and freeze after use.
2. Create the reaction mix:
    a. 4 µL cDNA template.
    b. 0.5 µL 10mM 3' Read2U primer.
    c. 1.0 µL 5' Forward primer mixture.
    d. 10 µL Kapa HiFi HotStart ReadyMix.
    e. 4.5 µL H_2O .
3. Run the PCR reaction according to the protocol:
    a. 96°C for 5 minutes.
    b. 25 cycles of:
        i. 95°C for 20 seconds.
        ii. 68°C for 20 seconds.
        iii. 72°C for 20 seconds.
    c. 72°C for 5 minutes.
    d. Hold at 8°C.
4. Run the samples on gel electrophoresis.
5. Cut out desired bands and purify the DNA using NEB Monarch Gel DNA extraction kit.

3.1.8 Gel electrophoresis

Gel electrophoresis was performed not only to ensure the DNA libraries were of the correct size but also to separate the DNA libraries from primer-dimers and other non-specific products. The gel matrix was composed of agarose, which created the environment in which DNA migrated. An electric field applied to the gel caused the DNA products to migrate at speed corresponding

to their size, causing the DNA products to separate. In order to determine the size of the DNA bands, a 1Kb+ DNA ladder from Invitrogen was utilized. GelRed was added into the gel matrix, which bound to the DNA and fluoresced under UV light.

3.1.9 DNA extraction from agarose gel

In order to extract DNA bands from the agarose gel, the NEB Monarch DNA Gel Extraction Kit was utilized. The desired DNA bands were excised using a scalpel and dissolved in a chaotropic buffer under heat in order to release the DNA from the gel. The DNA was retained in a silica membrane column and underwent washing with ethanol to remove impurities. Finally, the DNA was eluted using an elution buffer according to the manufacturer's instructions.

3.1.10 Adapter extension PCR

In order to make the DNA libraries compatible with Illumina's MiSeq sequencing platform, adapters and indices were ligated by PCR (Figure 7). The 5' primer targeting the Read1 sequence contained the P5 adapter while the 3' primers targeting the Read2 sequence contained the P7 adapter and the index sequence unique for each sample (Supplementary Table 3).

Procedure:

1. Create the reaction mix:
    a. 3 µL DNA template.
    b. 0.5 µL 10mM Illumina P5 primer.
    c. 0.5 µL 10mM Illumina P7 index primer.
    d. 9 µL H_2O.
    e. 12 µL Kapa HiFi HotStart ReadyMix.
2. Run the PCR reaction according to the protocol:
    a. 96ºC for 5 minutes.
    b. 10 cycles of:
        i. 95ºC for 30 seconds.
        ii. 68ºC for 30 seconds.
        iii. 72ºC for 30 seconds.
    c. 72ºC for 10 minutes.
    d. Hold at 8ºC.
3. Purify the DNA product.

3.1.11 Capillary Electrophoresis

Electrophoresis assays were performed using the Agilent Bioanalyzer 2100 system to evaluate the library before sequencing. The DNA chip was made of glass with multiple wells interconnected by microchannels. When all the wells were filled with DNA samples, marker DNA and intercalating dye mixed with gel matrix, they form a closed electrical circuit. An electric current was then applied to the chip and the DNA separated based on size, with larger fragments moving more slowly and smaller fragments moving more quickly. The samples were then visualized with fluorescence and the results were displayed as an electropherogram or a gel-like image.

3.2 Single-cell B-cell receptor sequencing

Library preparation for single-cell BCR sequencing similarly started with B cells isolated from whole blood. The B cells were then individually encapsulated in oil droplets together with unique barcodes and reagents for reverse transcription (Figure 8A). Two types of barcode sequences were present: droplet barcodes which confer a unique identification sequence to each cell; and molecular barcode which ligates to each mRNA molecule in order to facilitate error correction in the sequence. The concentration of barcodes was calculated and adjusted so that each droplet on average contained one unique droplet barcode and multiple molecular barcodes (Figure 8B). After encapsulation, cell lysis and reverse transcription occurred within each droplet (Figure 9) and the barcode sequences incorporated before the emulsion was broken and library preparation can be implemented.

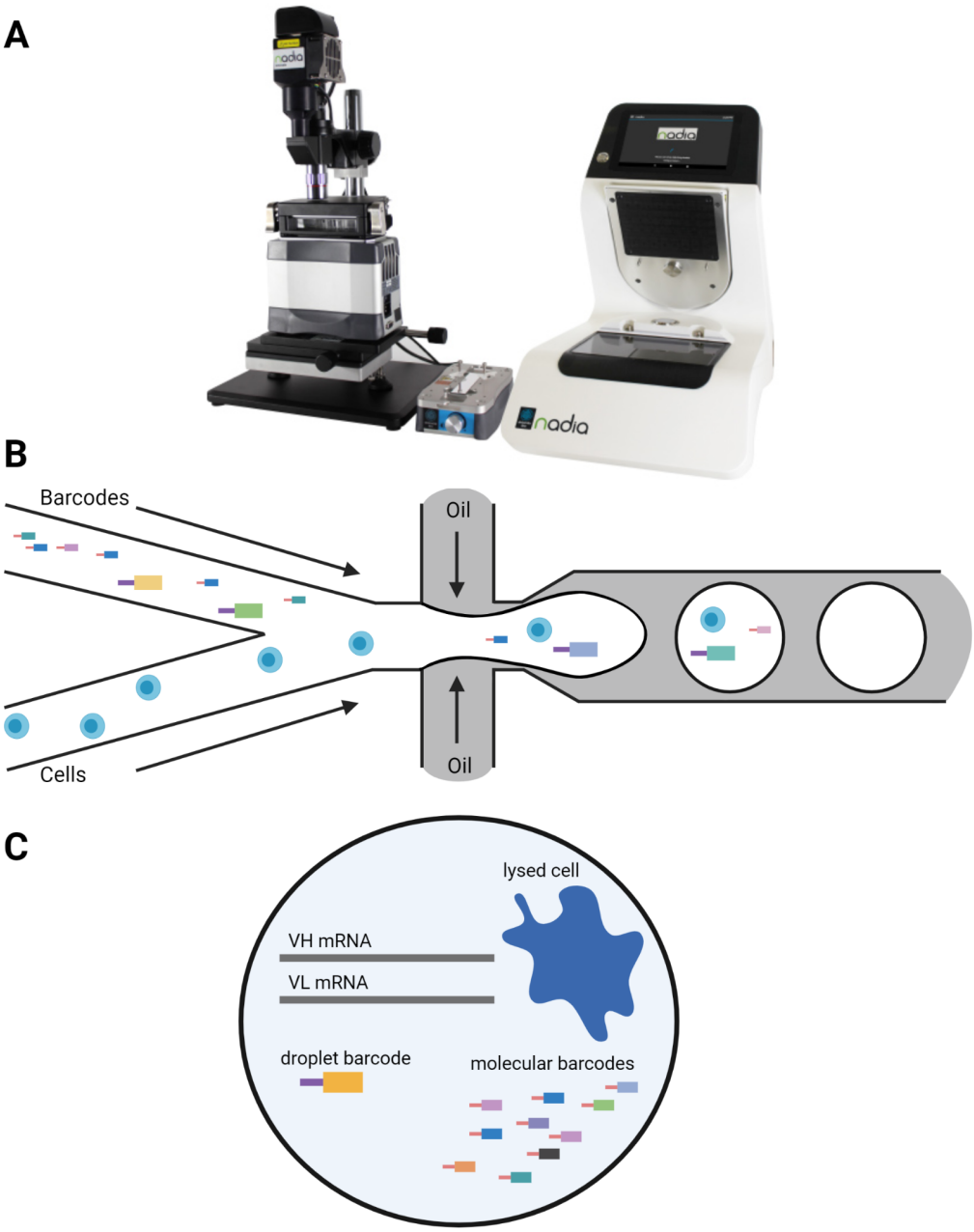


Figure 8: Schematic representation of the droplet encapsulation process. ***A)*** *The Nadia instrument with Innovate module add-on (image by Dolomite Bio [134]).* ***B)*** *Inside each microfluidic chip, B cells and barcodes arriving from separate channels are encapsulated together by the action of oil pressure, creating an emulsion of water droplets surrounded by oil. C) Within each droplet, the B cell is lysed, releasing the heavy and light chain mRNA, which serves as the template for reverse transcription and barcode incorporation.*

3.2.1 Single-cell encapsulation with the Nadia Instrument

The isolated B cells from peripheral blood were encapsulated using the Nadia Instrument from Dolomite Bio with customized parameters shown in Table 1. The pressure profile for each channel during an encapsulation run was shown in Table 2.

Table 1. Droplet encapsulation run parameters.

Sample loading stage		**Pre-run stage**		**Run stage**	
Parameter	Value	Parameter	Value	Parameter	Value
Sample volume	250 µL	Stirring time	30 s	Run temperature	6 ºC
Reagent volume	250 µL	Sample stirrer speed	75 rpm	Initial sample stirrer time	30 s
Oil volume	3 mL	Reagent stirrer speed	200 rpm	Initial sample stirrer speed	75 rpm
Loading temperature	5 ºC			Initial regent stirrer time	30 s
		Post-run stage		Initial regent stirrer speed	200 rpm
		Post-run duration	10 min	Final sample stirrer speed	75 rpm
		Post-run temperature	22 ºC	Final regent stirrer speed	150 rpm

Table 2. Pressure profile of a droplet encapsulation run stage.

Step	Duration (minutes:seconds)	Oil pressure (mbar)	Reagent pressure (mbar)	Sample pressure (mbar)
1	00:01	450	0	0
2	00:01	450	40	40
3	00:01	450	70	60
4	00:04	450	200	100
5	00:01	450	200	100
6	00:01	450	40	40
7	00:01	450	80	400
8	00:01	450	80	400
9	00:06	450	80	133
10	00:01	450	80	133
11	00:03	450	140	600
12	00:01	450	140	600
13	19:26	450	100	133
14	00:01	450	100	133
15	00:01	450	0	0
16	00:01	0	0	0

Procedure:

1. Power up the instrument and insert the microfluidic chip.
2. Load 3 mL of emulsion oil (QX200™ Droplet Generation Oil for EvaGreen) into the oil reservoir.
3. Load 250 µL of Cell lysis buffer (165 µL) and PCR reagents (85µL) (Supplementary Table 4) into the reagent reservoir.
4. Load 250 µL of B-cell suspension into the cell reservoir.
5. Close the lid of the instrument and start the encapsulation process, followed by 10 minutes of incubation.
6. Carefully remove excess oil from the reservoir and collect the cell droplets.

3.2.2 In-droplet reverse transcription

The droplets containing lysed B cells underwent template-switching reverse transcription and addition of droplet and molecular barcodes. Due to the activity of MuMLV reverse transcriptase, a non-templated overhang was created during cDNA synthesis. The molecular barcode acted as the template-switching oligo, annealing to the handle sequence. Within the droplet, the droplet barcode containing the partial Illumina P7 sequence was amplified and annealed to the molecular barcode via the universal adapter sequence (Figure 9).

Procedure:

1. Aliquot the droplets into PCR tubes, with each tube not exceeding 50 μL of droplets
2. Run the PCR according to the protocol:
    a. 65°C for 5 minutes.
    b. 37°C for 50 minutes.
    c. 70°C for 15 minutes.
    d. 95°C for 2 minutes.
    e. 40 cycles of:
        i. 95°C for 15 seconds.
        ii. 65°C for 20 seconds.
        iii. 72°C for 30 seconds.
    f. 72°C for 3 minutes.
    g. Hold at 8°C.

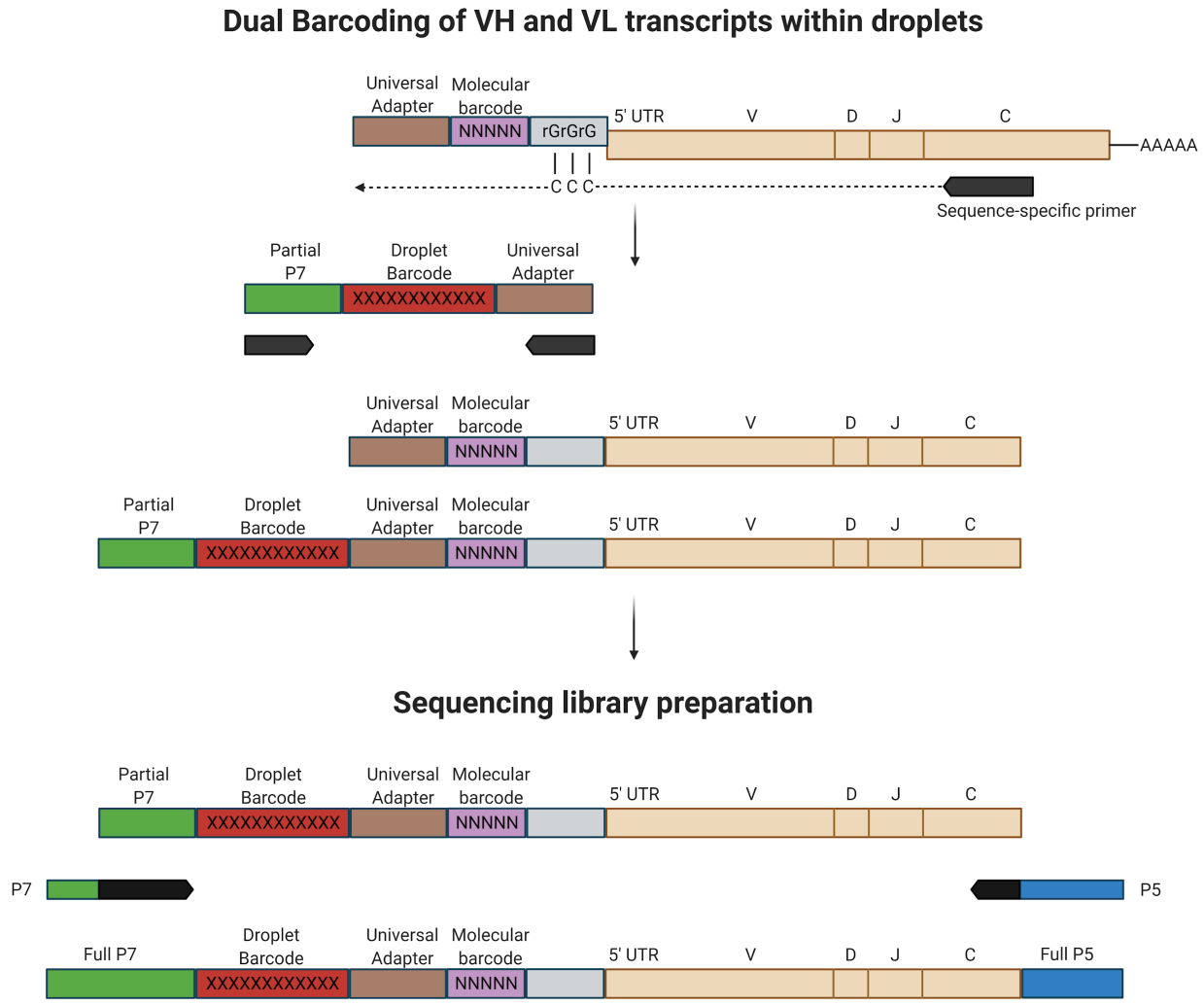


Figure 9: Summary of the single-cell B-cell receptor sequencing library preparation workflow. *Inside each droplet, cDNA synthesis occurred with isotype-specific primers by MuMLV reverse transcriptase, creating an overhang sequence for molecular barcodes to attach. Droplet barcode containing a partial Illumina adapter was amplified and attached to the cDNA via the universal adapter sequence. Afterwards, the droplets were broken and the full adapter sequences were added during limited-cycle amplification.*

3.2.3 Emulsion breakage and target enrichment PCR

After reverse transcription and addition of barcode sequences within the droplets (Figure 8), the emulsion was broken and then target enrichment PCR could occur, which ligated the complete P7 and P5 Illumina adapters to the sequence (Figure 9).

Procedure:

1. Add 40 ul emulsion breaking solution (1:1 FC-40:perfluorooctanol) together with 15ul lysate clearing solution (12.5 µL Qiagen Protease, 2.5 µL 0.5 M Na-EDTA, pH 8.0).
2. Invert 10 times to break the emulsion and incubate the mixture for 15 minutes at 50 °C and 3 minutes at 95 °C to inactivate the protease.
3. Centrifuge at 15,000 g for 1 minute and collect the supernatant.
4. Clean the products with Qiagen's MinElute PCR Purification kit.
5. Determine the nucleic acid concentration of the product.
6. Create the reaction mix (Supplementary Table 5):
    a. 5 µL 5x RT buffer.
    b. 0.5 µL dNTP.
    c. 1.25 µL primer IgG_nest_fullP5.
    d. 1.25 µL primer IgM_nest_fullP5.
    e. 1.25 µL primer fP7.
    f. 30 ng of DNA product.
    g. 0.25 µL Q5 polymerase.
    h. H_2O up to 25 µL.
7. Run target enrichment PCR according to the protocol:
    a. 98°C for 30 seconds.
    b. 20 cycles:
        i. 98°C for 5 seconds.
        ii. 72°C for 45 seconds.
        iii. 72°C for 2 minutes.
    c. Hold at 8°C.
8. Analyze the library products on the BioAnalyzer system.

3.3 Antibody mass spectrometry

3.3.1 Antibody purification from serum

Antibodies were purified from serum using the Nab Protein A/G Spin kit from Thermo Scientific following the protocol provided. This is a column-based affinity chromatography method suitable for small scale antibody purification in a short amount of time using Protein A and protein G immobilized in the column, which have high binding affinity for IgG.

3.3.2 GingisKHAN antibody F(ab) fragment collection

After antibodies were isolated from serum, F(ab) fragments were generated using the GingisKHAN F(ab) kit from Genovis according to the manufacturer's instructions. GingisKHAN is an IgG1 specific protease targeting a single site above the hinge region of antibodies. The reaction occurred at 37^{o}C and pH 8.0. After cleavage, the resulting F(ab) fragments were purified using affinity chromatography that is specific for the CH1 domain of the heavy chain of IgG.

3.3.3 Enzymatic digestion of antibodies

Purified intact antibodies were enzymatically digested prior to chromatographic separation. Three digestion strategies were employed: digestion with trypsin (tryp) only (Trypsin Gold, Mass Spectrometry Grade, Promega Cat 5280), chymotrypsin (ct) only (Chymotrypsin MS grade, Thermo Fisher Cat 90056), and chymotrypsin digestion followed by additional trypsin digestion (ct + tryp). This was done to diversify the resulting peptide sequences and increase the sequence coverage of the antibody peptides (Figure 10).

Procedure:

1. Divide each sample into 3 (20µl in each tube).
2. Add 2µl 100mM Dithiothreitol (final 10mM) and incubate at 56°C for 30 minutes.
3. Add 5.5µl 55mM Indole-3-acetic acid (15mM) and incubate in the dark for 30 minutes.
4. Add 1µl 0.5µg/µl chymotrypsin into two of the replicates and incubate at 37°C for 4 hours.
5. Add 1µl 0.5 µg/µl trypsin into one of the chymotrypsin treated samples and one of the untreated samples. Incubate at 37°C overnight.
6. Save the chymotrypsin-treated samples in the freezer until the next day.

7. Total volumes are 28.5µl (chymotrypsin, trypsin), 29.5µl (chymotrypsin + trypsin).
8. Perform Evotip C18 purification according to standard EVOSEP instructions. Load 7.125µl of trypsin sample, 7.125µl of chymotrypsin sample, and 7.325 of chymotrypsin+trypsin sample.

A

Enzyme name	P2	P1	P1'
Chymotrypsin	-	F or Y	not P
	-	W	not M or P
Trypsin	-	K or R	not P
	W	K	P
	M	R	P

B

Enzyme name	P2	P1	P1'
Trypsin	C or D	K	D
	C	K	H or Y
	C	R	K
	R	R	H or R

C

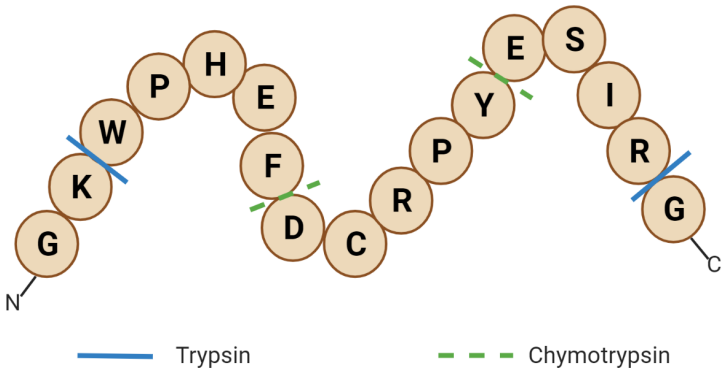


Figure 10: Specificity of commonly used proteolysis enzymes. ***A)*** *Cleavage rules for Chymotrypsin and Trypsin, with regards to the peptide composition of the target site.* ***B)*** *Exception rules for Trypsin detailing specific peptide compositions in which cleavage does not occur.* ***C)*** *Representation of effect of enzymatic cleavage on peptide sequence composition. Enzyme cleavage rules were extracted from the ExPASy Peptide cutter tool [135].*

3.3.4 Liquid chromatography and tandem mass spectrometry

All mass spectrometry experiments were performed by The Proteomics Core Facility at Oslo University Hospital on an EVOSEP liquid chromatography system connected to a quadrupole – Orbitrap (QExactive HF) mass spectrometer (ThermoElectron, Bremen, Germany) equipped with a nanoelectrospray ion source (EasySpray/Thermo). For liquid chromatography separation

an 8 cm C18 column was used (Column details: Dr Maisch C18 AQ, 3um beads, 100um ID, 8cm long.EVOSEP). The standard EVOSEP method 100 samples/day was used.

The mass spectrometer was operated in data-dependent mode to automatically switch between MS and MS/MS acquisition. Survey full scan MS spectra (from m/z 375 to 1,500) were acquired in the Orbitrap with resolution R = 60,000 at m/z 200 (after accumulation to a target of 3,000,000 ions in the quadruple). The method used allowed for sequential isolation of the most intense multiply-charged ions, up to twelve, depending on signal intensity, for fragmentation on the higher-energy collisional dissociation (HCD) cell using high-energy collision dissociation at a target value of 100,000 charges or maximum acquisition time of 50 ms. MS/MS scans were collected at 30,000 resolution at the Orbitrap cell. Target ions already selected for MS/MS were dynamically excluded for 30 seconds. General mass spectrometry conditions were: electrospray voltage, 2.0 kV; no sheath and auxiliary gas flow, heated capillary temperature of 250ºC, normalized HCD collision energy 28%.

3.4 Data analysis

3.4.1 Quality control of sequencing reads

Quality checking of the sequencing libraries was performed by the sequencing facilities using FastQC [103]. The quality control report provided information such as the total number of reads in a sequencing sample, the median quality value of each base across a read, the presence of adapter sequences in a read, and the proportion of overrepresented sequences in a library. Quality reports from all samples in a sequencing run were combined into a single report in MultiQC [136].

3.4.2 UMI-based error correction

The utilization of UMI in the sequencing workflow for error correction necessitated a software pipeline to identify, extract UMI sequences and group reads together by molecular identifiers. MIGEC, developed by Shugay and colleagues processed sequencing data in 3 steps [137]. First, the UMI tags were extracted from the read and added onto the header section by scanning and

comparing with previously provided barcode sequences. Next, the quality of the extracted UMIs was evaluated. UMIs were grouped together into molecular identifier groups (MIG) and each group's over-sequencing level and size were calculated and displayed as a histogram. These values helped to exclude erroneous MIGs and kept the amplified and eligible MIGs for reliable read assembly. In the last step the reads in a MIG were assembled into a consensus read, excluding all the MIGs that fall below the desired thresholds. This helped to correct PCR errors in the reads and excluded MIGs that contained too few reads (usually errors in the UMI sequence). The resulting outputs were then forwarded to assembly mapping and clonotyping. The MIGEC version 1.2.9 workflow was executed in command line using the following script:

```
#Running UMIs search by migec and putting the results in the folder
/checkout
java -jar ../migec-1.2.9.jar CheckoutBatch -c -u --skip-undef
barcodes.txt ./checkout

#Generating statistics for the UMIs found and storing the results in
wd/checkout/histo
java -jar ../../migec-1.2.9.jar Histogram ./ ./histo

#Running R script to create histogram of MIGs
Rscript Histogram.r

#Setting "checkout" as the working directory and data assembly with
output folder wd/output
cd checkout
java -jar ../../migec-1.2.9.jar AssembleBatch -c
--force-collision-filter . ./histo/ ./output/
```

3.4.3 Read assembly and clonotyping

B-cell receptor sequencing data was processed by MiXCR version 3.0.13, which is highly efficient at handling immune repertoire sequencing data [138]. UMI-corrected sequencing data

was first aligned to reference V,D,J and C germline sequences in the reference database and the paired-end reads were overlapped. Subsequently the aligned reads were assembled into clonotypes using the desired assembling features and clonotypes with high similarity were clustered together in a hierarchical manner. Via multi-layered clustering (assembling consensus CDR3 sequences using homologous and identical CDR3s), MiXCR also allowed for correction of both PCR and sequencing errors by reducing artificial clonal diversity, while at the same time rescuing reads that are low quality by matching and aggregating these reads to already constructed clonotypes. Thus the clonal diversity of the repertoire may be preserved [62]. Finally, the resulting clonotypes were exported in a human-readable format containing information such as clonotype count, amino acid and nucleic acid sequences for each feature, best match for germline gene family, mutation in the sequences, among others. MiXCR was executed in command line using the following script:

```
#Read alignment
java -Xmx4g -Xms3g -jar PATH align --species hs --report
alignmentReport.log $(ls|grep R1)  $(ls|grep R2) alignments.vdjca

#Assemble clonotypes
java -Xmx4g -Xms3g -jar PATH assemble
-OassemblingFeatures="[VDJRegion]" -OaddReadsCountOnClustering=TRUE
-OcloneClusteringParameters.clusteringFilter.specificMutationProbabil
ity=1E-4 -OseparateByC=true --report assembleReport.log
alignments.vdjca clones.clns

#Export clonotypes
java -Xmx4g -Xms3g -jar PATH exportClones -vGene -jGene -nMutations
VRegion -aaMutations VRegion -nMutations JRegion -aaMutations JRegion
-nMutations DRegion -aaMutations DRegion --preset full -nFeature CDR3
-nFeature VDJRegion -aaFeature VDJRegion --filter-out-of-frames
--filter-stops clones.clns clones.txt
```

3.4.4 Immune repertoire analysis

The MiXCR output file of each sample provided information on identified features of the V region, including the name and family of the V, D, and J genes, and the identified isotype (C region). In addition, the nucleotide and amino acid sequence of each feature, including the CDRs and FRs, were also provided, along with any identified mutation compared with the germline reference database. A clonotype was defined as sequences with the same identified V and J genes, and containing CDR3 sequences of the same length [139]. Clonotype data from MiXCR was analyzed using the R package Immunarch [140], which performed the main analysis on immune repertoire data, including clonotype abundance, CDR3 length distribution, overlap between repertoires, and V-gene usage.

The CDR3 region is the most important determining factor of antigen recognition [141]. The length of the CDR3 region varies due to somatic recombination, with SHM also contributing the sequence length. By looking at the shape of a CDR3 length distribution, it is possible to identify the presence of clonal expansion in a repertoire [142]. CDR3 overlapping is a common approach to study the overlap of repertoires, defined as the presence of identical immune receptor sequences between samples, by quantifying the proportion of shared clonal sequences. This can yield important information since exposure to antigen can affect clonal selection and increases repertoire similarity [62]. There exist multiple methods to evaluate repertoire overlap, the simplest of which is calculating the overlap coefficient. By regarding the CDR3 amino acid sequence as strings of text, two strings can be a match if one is a subset of the other, and the coefficient is calculated as the size of the intersection divided by the size of the smaller repertoire of the two. If one repertoire is a subset of the other, the overlap value is 1 [143].

The IGHV gene locus contains 7 V gene families, with each family having a myriad of V genes. However, not all V genes have the same likelihood of expressing in the B cell repertoire due to negative selection pressure during B-cell development. Nonetheless, specific V-gene usage patterns have been associated with the development of autoimmune diseases and can fluctuate in response to therapy [144,145]. In order to evaluate consistency in V-gene usage patterns between

repertoires, pairwise Pearson correlation coefficients were calculated and aggregated into a heatmap using R's pheatmap function [146]. Pearson correlation coefficient is a normalized measurement of covariance, a number between -1 and 1 that measures the linear relationship between two sets of data.

While sequence-dependent analyses can provide useful information about an immune repertoire, there exists little overlap in clonal sequence between repertoires, making comparison restricted to the subset of public clones [66]. Repertoire diversity analysis, derived from but not dependent on clonal sequence, can yield important information regarding the immunological status of the subject. This is especially important in studying the immune response to diseases and vaccines [147]. There are many indices to measure diversity, such as species richness [148], Shannon entropy [149], the inverse Simpson index [150], and Berger-Parker index [151]. However, different indices may provide different conclusions regarding repertoire diversity [152]. Therefore, Greiff and colleagues developed the Diversity index profile which united the aforementioned indices and painted a more complete picture of the immune repertoire structure [68]. Alpha diversity was calculated as $^{\alpha}D(f) = \left(\sum_{i=1}^{n} f_i^{\alpha}\right)^{\frac{1}{1-\alpha}}$, where f is the clonal frequency distribution, f_i is the clonal frequency of each clonotype, and n the total number of clonotypes. The Diversity index profile calculations were performed using the R package vegan [153] with clonotype frequencies as the input. By extension, Evenness profiles, defined as the alpha Diversity divided by the species richness, were calculated in order to measure the distance between the clonal frequency distribution and the uniform clonal frequency distribution. Pairwise Pearson correlation of Evenness profiles between repertoires were calculated and visualized using R's pheatmap function [146].

3.4.5 Mass spectrometry data analysis

Proteomics data from LC-MS/MS were processed by MaxQuant version 1.6.7.0, developed by Cox and colleagues [154]. MaxQuant allowed for higher peptide mass accuracy, improved peptide and protein identification, and integrated all steps needed in a computational proteomics workflow in a single package. Thus, MaxQuant database search was performed by The

Proteomics Core Facility at Oslo University Hospital. Data from identified peptide sequences and their respective matching databases were further processed in R. For the antibody LC-MS/MS experiment, the parameters utilized were described in detail in Supplementary Table 6.

Briefly, in addition to the antibodies' own sequences, the experimental data was searched against 12 databases (up to 2 miscleavages allowed, minimal detected peptide length 7aa, false discovery rate 1%):

- UniProt – uniprot-reviewed_Homo_sapiens_Sept_2018 (1 database).
- IGoR igh – 10000 randomly generated naive sequences using the IGoR software suite (1 database) [155].
- IMGT genes – igh (v, d, j genes), igk (v, j genes), igl (v, j, genes) (3 + 2 + 2 databases) [156].
- ImmuneSIM igh, igk, igl – 10000 naive sequences randomly generated with ImmuneSIM (3 databases) [157].
- Antibody-specific databases (Supplementary Table 7).

MaxQuant reports sequences of detected peptides with their intensities and matching databases (one peptide might be detected in multiple databases, e.g. peptide from a V gene or constant region). Each detected peptide was overlapped with the CDR3 sequence of the correspondent antibody. The intensity ratio of peptides that uniquely correspond to an antibody was compared with the concentration ratio.

4. Results

4.1 Bulk B-cell receptor sequencing

4.1.1 Assessment of library quality

The PCR protocol for the amplification of BCR sequence was inspired by Bernat et al. [88], where they have generated 5' multiplex primers for the capture of IgG, IgM heavy chain and κ, λ light chain sequences. Here we expanded upon the framework by designing isotype-specific primers to capture also IgD, IgA, and IgE sequences. Thus we managed to conduct a comprehensive analysis on all Ig isotypes from a blood sample (Table 3).

Table 3. Sample description of bulk B-cell receptor sequencing libraries.

Isotype	Illumina index		
	Replicate 1	Replicate 2	Replicate 3
Heavy chain			
IgG	**i1**	**i2**	i3
IgM	**i3**	i1	i2
IgD	**i4**	**i5**	i6
IgA	**i6**	i4	i5
IgE	**i7**	**i8**	i9
Light chain			
IgK	**i9**	i7	i8
IgΛ	**i10**	**i11**	i12

Bold text represents samples in the first sequencing batch.
Unbolded text represents samples in the second sequencing batch.

The Bioanalyzer results showed that the protocol successfully generated libraries for all the isotypes, which was between 500-600 bp in size. For IgG libraries, there were by-products at around 200 bp that were removed during post-PCR cleanup. In IgE libraries, aside from the expected libraries at 600 bp, there existed also unwanted products in the 400–500 bp range,

suggesting that the isotype-specific IgE cDNA primers needed further adjustments and optimization (Figure 11). Accounting for the relative abundance of the libraries in different isotypes, the proportion of samples used when pooling was adjusted accordingly to ensure sufficient read coverage and depth in all samples before sequencing.

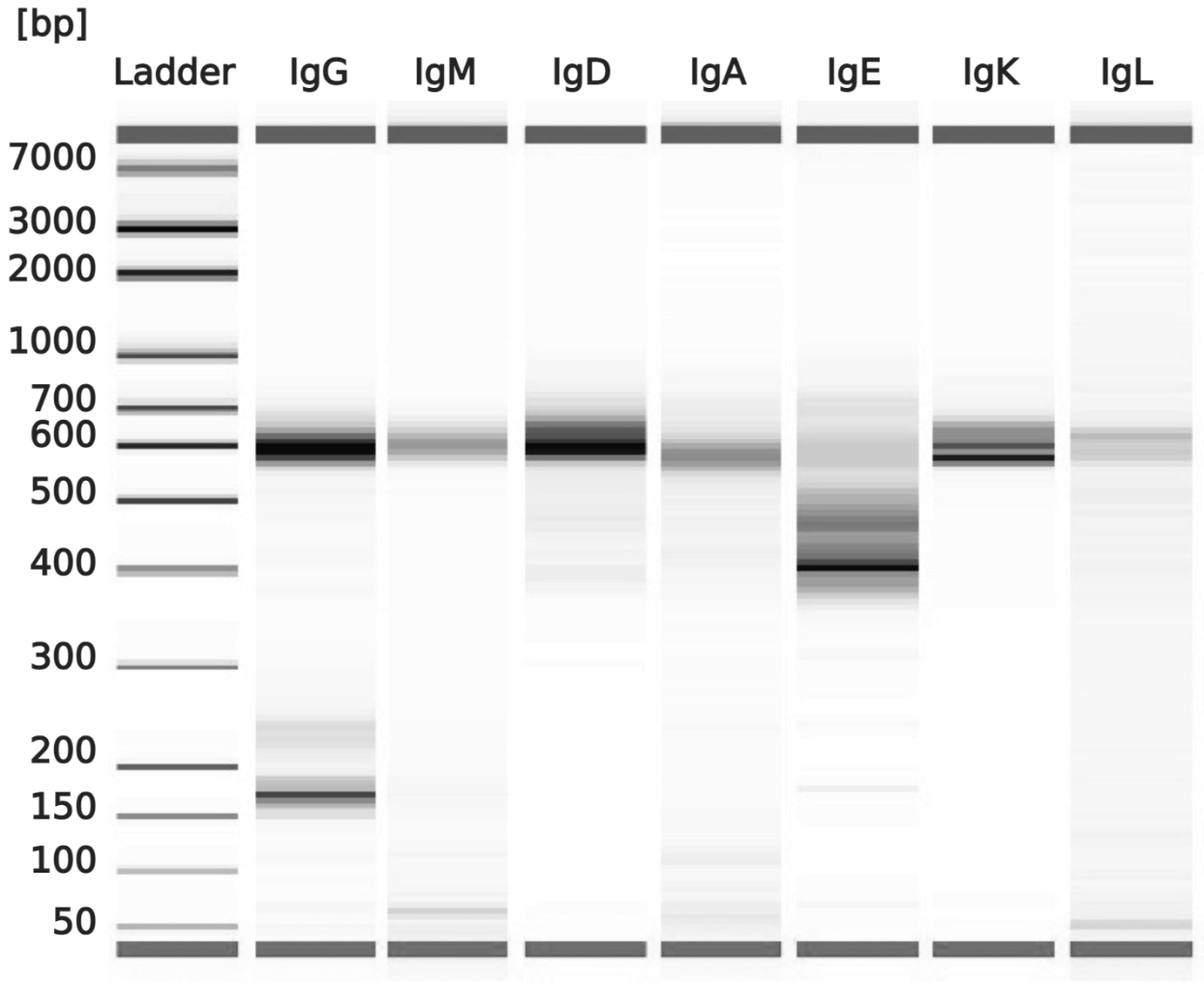


Figure 11: Representative example of bulk B-cell receptor libraries, with each sample tagged with individual Illumina index adapters. The expected size of the libraries is between 500–600 bp. *Before sequencing, the libraries were purified again using AMPure XP Beads to remove non-specific products and pooled into a single microcentrifuge tube.*

The data retrieved from the sequencing facility was evaluated by FastQC and displayed as an aggregate in MultiQC. In each sequencing batch, the sequences for each sample were divided into the forward read (R1) and the reverse read (R2). For the first sequencing run (Table 3, Figure 12A), out of the 22 sample reads available (paired-end reads of 11 samples in each sequencing run), 8 sample reads were of good quality while 14 sample reads were of lower quality. Specifically, all of the R2s fell into the low quality category. Meanwhile, the sequencing quality of the second sequencing batch (Table 3, Figure 12B) is generally better, with 10 R1s having a higher quality score and 10 R2s having a lower quality score. Overall, the quality score for the R2s were lower than the R1s. This is a common pattern in Illumina sequencing, especially in long-read sequencing, and was discussed by Tan and colleagues when they evaluated the quality of paired-end sequencing in different Illumina platforms, overlining the limitations of paired-end sequencing [158]. Although the quality of the bases toward the end of the read was lower than recommended by FastQC, we decided to keep these reads for further analysis since the process of alignment and mapping by MiXCR can potentially rescue these reads when the R1 and R2 of each sample are merged and by clustering reads of lower quality into a larger clonotype during clonotyping, retaining as much of the sequence data as possible.

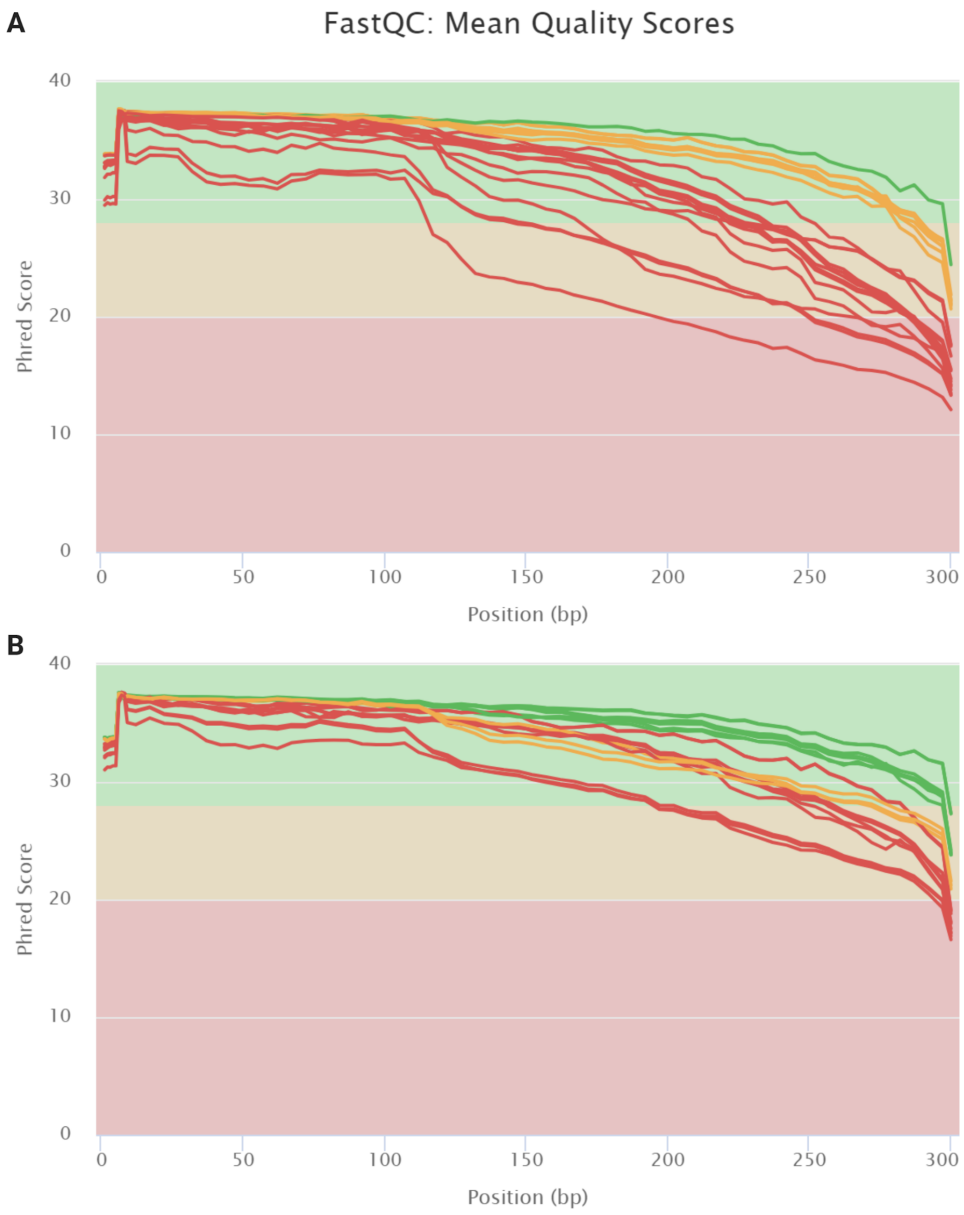


Figure 12: Overview of the bulk B-cell sequencing library quality. *Quality is represented as the mean Phred score of all positions in a sequencing read. The Phred score Q is defined as Q = -10 x log10(P) where P is the probability of an incorrect base call. The quality score distribution of each library is represented as a color-coded line according to FastQC's assessment of library quality and categorized as high quality (green), acceptable quality (yellow), or low quality (red).* ***A)*** *Quality score of libraries in the 1st sequencing run.* ***B)*** *Quality score of libraries in the 2nd sequencing run.*

4.1.2 Analysis of B-cell receptor libraries

Before alignment and clonotyping by MiXCR could occur, the raw sequencing data went through UMI processing by MIGEC and the resulting MIG distributions were visualized by isotype (Figure 13). The vast majority of MIGs (> 80%) in all samples contained less than 3 reads, which is the minimum size necessary to construct consensus sequences. In particular, in very abundant samples such as IgM or κ light chains, more than 95% of MIGs had only one or two reads. If UMI error correction was employed, the amount of useful information retrieved from the data would be reduced more than 5-fold and up to 10-fold. Since the goal was to capture as much diversity in the data as possible, we decided that UMI error correction would not be utilized in this experiment and error correction would depend mainly on MiXCR instead[4].

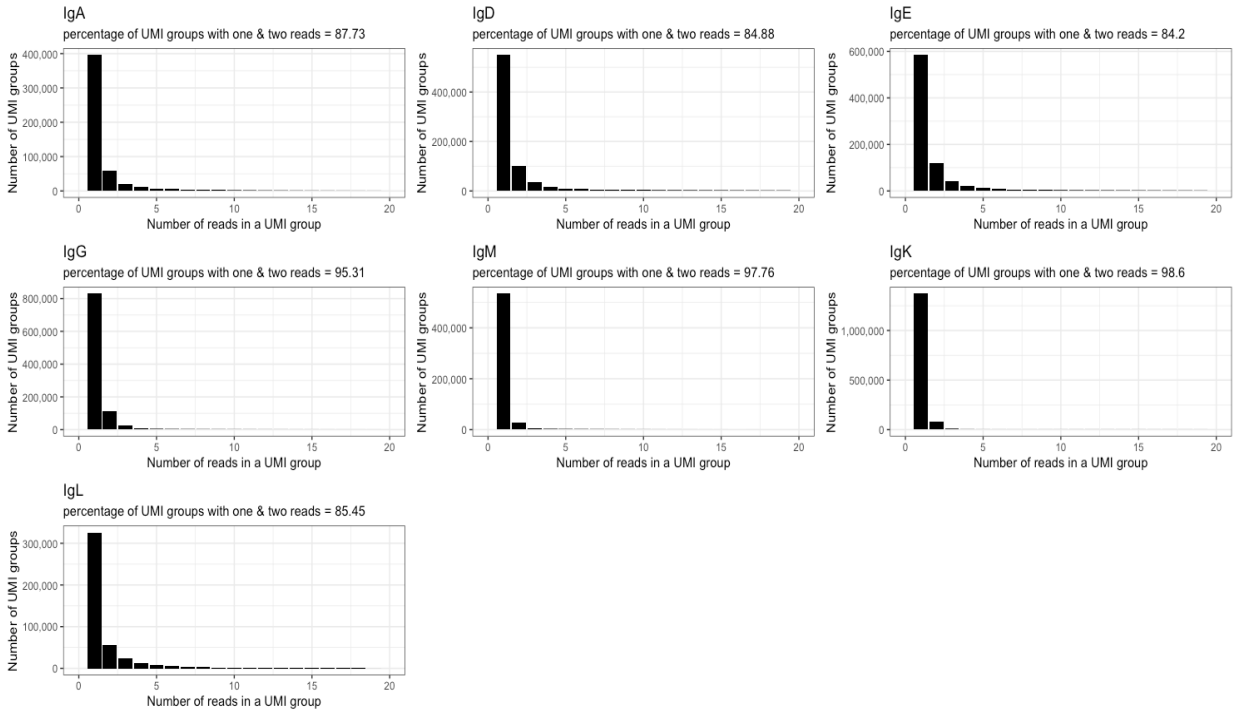


Figure 13: UMI group distribution across isotypes. In all isotypes, including the light chain samples, more than 80% of MIGs contained only one or two reads (exact proportions given above each plot). *Identical UMI sequences were assembled into molecular identifier groups (MIGs). MIGs were categorized based on the number of reads contained. Consensus sequence building requires a MIG size ≥ 3.*

[4] The MiXCR error correction mechanism was introduced in section 3.4.3 "Read assembly and clonotyping".

Before analysis of the immune repertoire can be conducted, the output file of the MiXCR pipeline of each sample was examined. Based on the identification of the C gene (determinant of isotype), it was shown that a sizable number of assembled clonotypes have unidentified C region or C region identified as another isotype. Therefore, these non-specific clonotypes were removed before subsequent analyses were conducted. In order to evaluate the impact of this removal, clonotype counting was performed on both the original data and the isotype-corrected data. Comparing the two datasets showed that in the IgG and IgM samples, around half (54.7% and 43.8%, respectively) of the identified clonotypes had unidentified isotype or an isotype different from the input sample. In the IgA samples, 72.3% of unique clonotypes were removed after correction. IgE samples in particular saw the largest reduction in clonotype count with 92.3% of unique clonotypes removed after correction (Figure 14). However, this is expected due to the presence of some non-specific products in the electropherogram of the IgE lane and the low intensity of the band at the expected product size (Figure 11). In contrast, the IgD samples and the light chain samples saw little reduction in clonotype count after correction, suggesting the presence of highly specific products.

Across all isotypes, IgD had the highest number of unique clonotypes, followed by IgM and IgG, in agreement with the fact that the majority of B cells in the blood are naive $IgM^{+}IgD^{+}$ B cells [159]. Meanwhile, IgE showed the lowest clonotype count, at <1000 unique clonotypes per replicate. This is expected both from the low abundance of IgE B cells in peripheral blood and from the performance of the IgE-specific reverse transcription primer (Figure 11), rendering it difficult to produce only specific products. For the light chain samples, the clonotype count for the κ chain sequences was the highest averaging around 196768 unique clonotypes per sample, owing to the fact that ⅔ B cells in the blood had the κ light chain. Accordingly, the λ chain sequences averaged around 34049 unique clonotypes (Figure 14B). In particular, the high variation of clonotype count in the κ light chain replicates indicated that the true abundance of the sample was higher than what was displayed and sub-sampling might have affected the results.

Clonotypes that contained only one read on average made up around 50–60% of those identified in the samples (IgA, IgE, IgG, and the light chain sequences). However, for IgD samples, the single-read clonotypes accounted for only around 33% of all identified clonotypes. Conversely, for IgM samples, single-read clonotypes comprised 70% of identified clonotypes (Figure 14C). This discrepancy between isotypes can partly be explained by the sequencing depth of each sample. A very high percentage of single-read clonotypes might signify the need for more sequencing depth since the large number of cDNA molecules in such samples might not be sufficiently covered with multiple reads in sequencing. In contrast, low single-read clonotypes samples had a higher degree of saturation regarding the number of reads per cDNA molecule. In short, the number of clonotypes identified and the clonotype abundance in each sample can be affected both by the prevalence in the blood as well as the allocation of sequencing depth for each sample. Thus there is room for optimization in these areas.

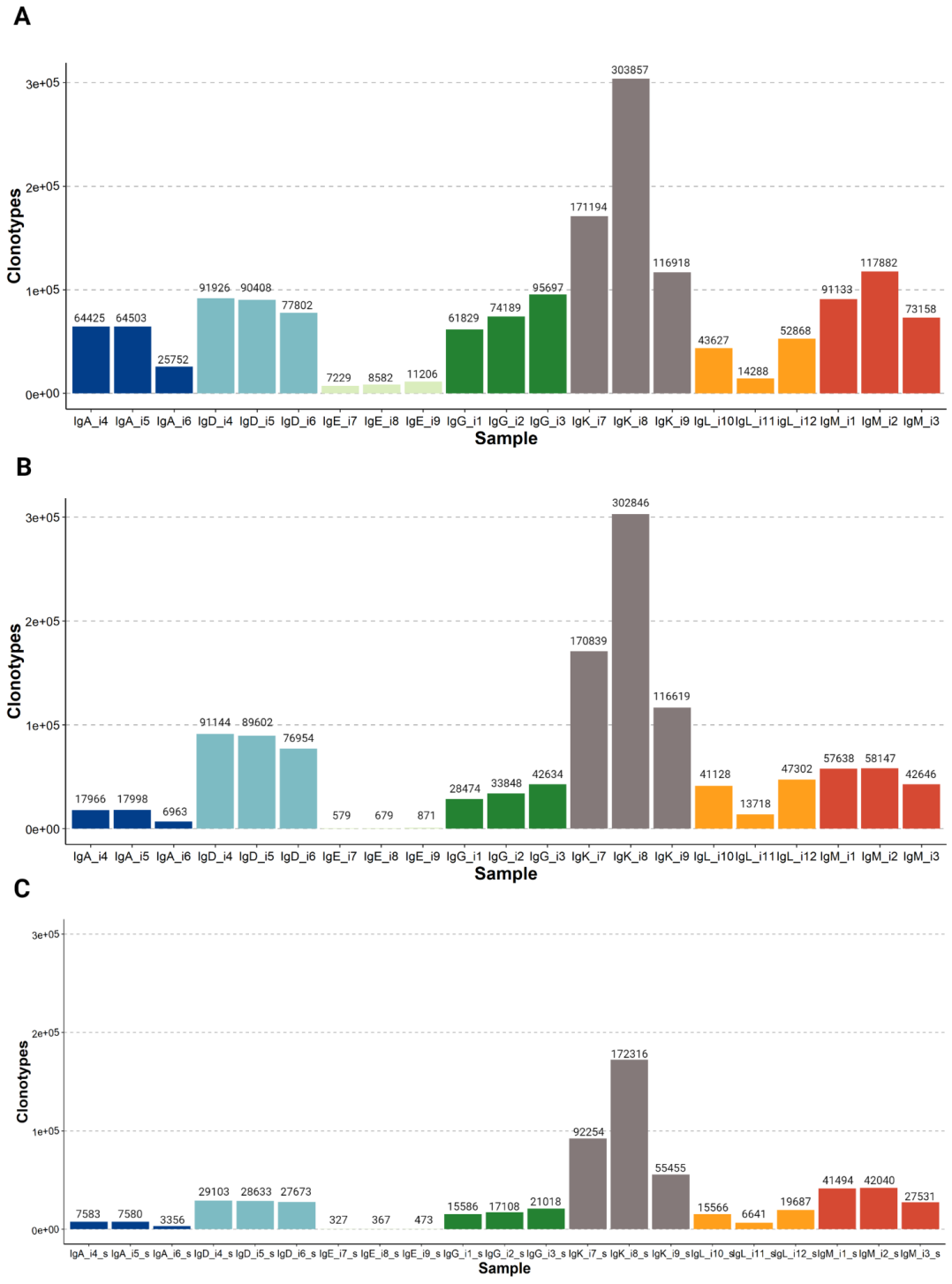
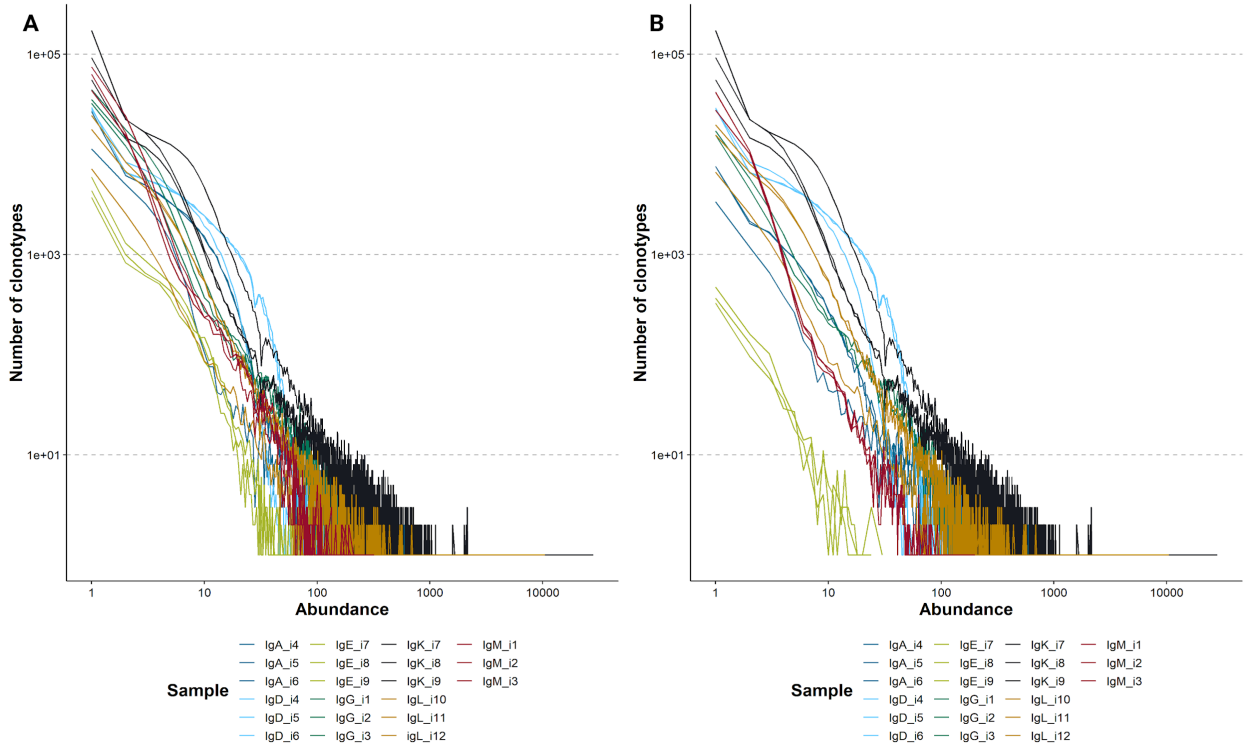


Figure 14: Number of unique clonotypes identified by MiXCR. *A clonotype was defined as sequences with the same identified V and J genes, and containing CDR3 sequences of the same amino acid length.* ***A)*** *The number of clonotypes in each sample before isotype correction.* ***B)*** *The number of clonotypes in each sample after isotype correction.* ***C)*** *The number of clonotypes in each sample after selecting for clonotypes that contained only one read.*

Despite the large impact on clonotype count, the isotype correction primarily affected clonotypes with low clonotype count, mostly in the range of ≤ 10 reads, as evidenced by the clonal abundance distribution (Figure 15). The change in clonotype abundance was most evident in the IgE samples (light green lines), where the number of clonotypes with abundance ≤ 10 was on average around 8455 clonotypes (Figure 15A). After filtering out non-isotype-specific clonotypes, the abundance of clonotypes with ≤ 10 reads reduced to an average of 689 clonotypes (Figure 15B). Similar patterns can be observed in the IgM samples (red lines), where the abundance of clonotypes with ≤ 10 reads was around 91547 clonotypes before correction and around 52318 after isotype correction (Figure 15B). Meanwhile, the most abundant clonotypes in each sample remained relatively unaffected by the correction process, since there were only several clonotypes that were highly expanded in each sample. The distribution of clonotype abundance was as expected from an antigen-inexperienced blood sample, with the vast majority of clonotypes possessing a very small size and a very small fraction of clonotypes that were expanded.



(See figure on the previous page)

Figure 15: Clonotype abundance of each sample, with abundance defined as the number of reads assembled into a clonotype. The bulk of clonotype identified in each sample had an abundance of ≤ 10 reads. Filtering of misidentified and unidentified isotypes reads from samples primarily affected IgE samples and to a lesser extent IgM samples. ***A)*** *Clonotype abundance before isotype correction.* ***B)*** *Clonotype abundance after isotype correction.*

In addition to investigating the number of unique clonotypes in each sample, in order to explore the diversity of each repertoire, the sequence contents of the clonotypes in each repertoire were analyzed, specifically the CDR3 region. Within each isotype, the CDR3 length frequency stayed consistent between technical replicates, with the exception of IgE, which showed slight variation between replicates. This could be attributed to the smaller sample size of IgE, which could amplify the differences between replicates. Regarding the CDRH3 length distribution across different isotypes, IgD had the longest average CDRH3 sequence at 18 aa and a flatter distribution pattern; while IgA, IgE, IgG, and IgM had an average CDRH3 length of 17 aa, which were consistent with other publications [64,160] (Figure 16A). In addition, the pattern of peak distribution in all isotypes followed the typical peak profile which resembled a Gaussian distribution, representative of healthy volunteers, matching what we expected [142].

Regarding the light chain, the average length of the CDRL3 sequence of the λ and κ light chain were 11 aa and 12 aa, respectively. Length variation of the CDRL3 sequences was limited possibly due to the CDRL3 being comprised of only the V-J gene junction, rather than the V-D-J gene junctions in CDRH3, limiting the possibilities for variation in length due to non-template additions and removals during recombination.

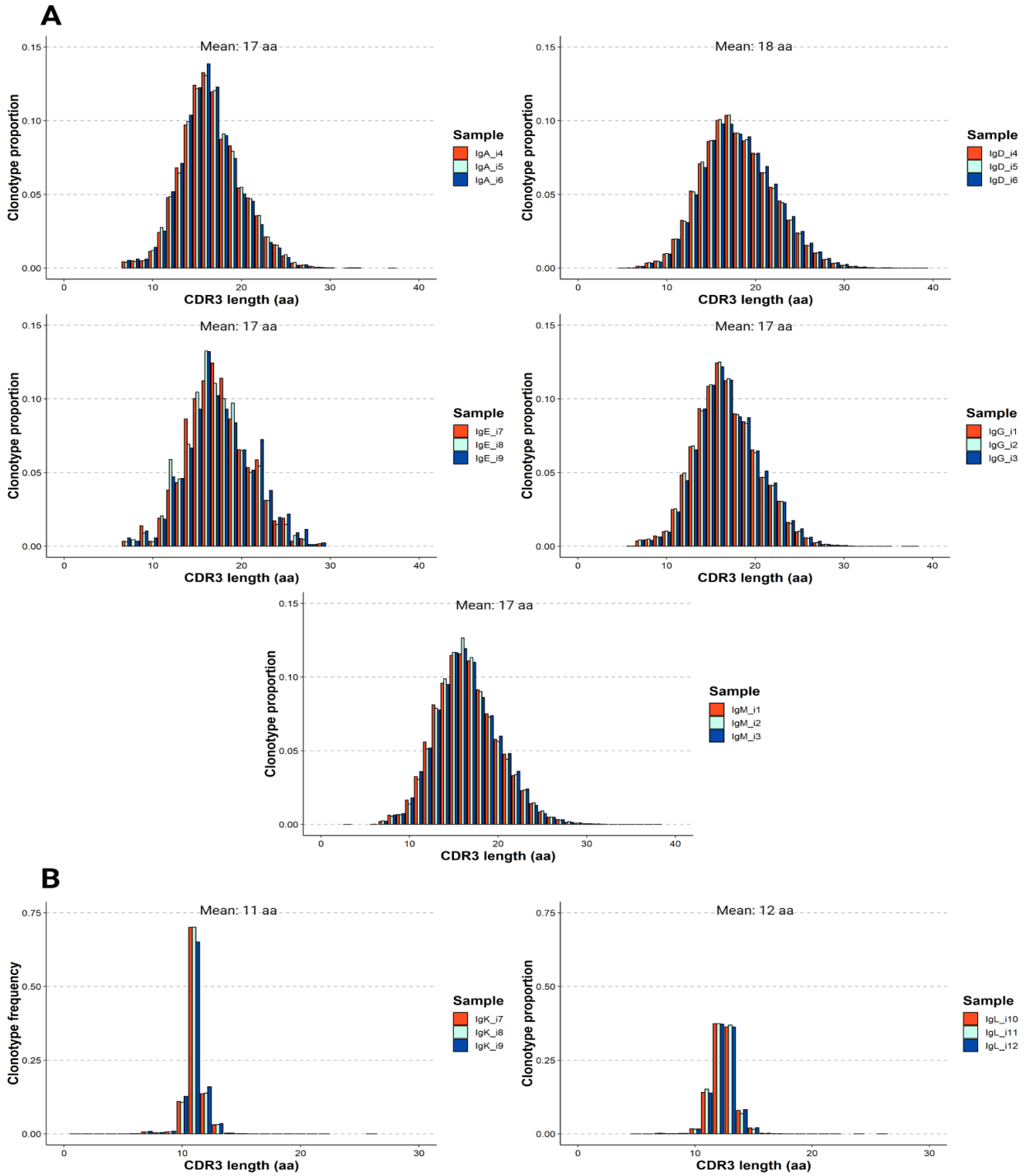


Figure 16: CDR3 length in nucleotides and the frequency of each CDR3 length in a repertoire, adjusted for clonotype count in each sample. CDR3 length distribution in each heavy chain isotype resembled a normal distribution. The mean length of the CDR3 sequence is 17 aa in the heavy chain and 11 aa in the light chain. ***A)*** *CDR3 length distribution of different Ig isotypes: IgA, IgD, IgE, IgG, and IgM,.* ***B)*** *CDR3 length distribution of the light chain sequences: κ and λ.*

Subsequent to clonotype counting, the CDR3 overlap between samples was examined (Figure 17). As expected, CDR3 amino acid sequence similarities were higher between replicates of the same isotype and drastically lower between different isotypes (<0.1000). In particular, the overlap values were highest within the isotypes IgM, IgA, and IgD, at around 0.6000. IgG, IgE and λ chain sequences showed medium overlap values while κ chain sequences showed the lowest overlap between replicates. Interestingly, IgE samples had a higher degree of overlap with IgG samples than between replicates of the same isotype.

After selecting for only single-read clonotypes in each sample, the CDR3 overlap values exhibited a similar pattern between replicates and across isotypes, albeit with lowered values (Figure 18). An exception to this was IgM, which retained a similar average overlap value even after removing expanded clonotypes. This contrasted the pattern exhibited by other isotypes such as IgD, in which the average overlap value reduced by 0.4041 after removing expanded clonotypes. This phenomenon might be explained by the clonotype counts of the isotypes, in which IgD only had around 33% single-read clonotypes whereas IgM had as high as 70% single-read clonotypes (Figure 14). It is possible that between technical replicates, the same clone that contained 1 read might contain > 1 read in other replicates and thus be removed by filtering, which would lower the overlap value. The higher overlap value when single-read clonotypes were included demonstrated that many of the single-read clonotypes were correct and not due to sequencing errors, thus a higher sequencing depth would have rescued them.

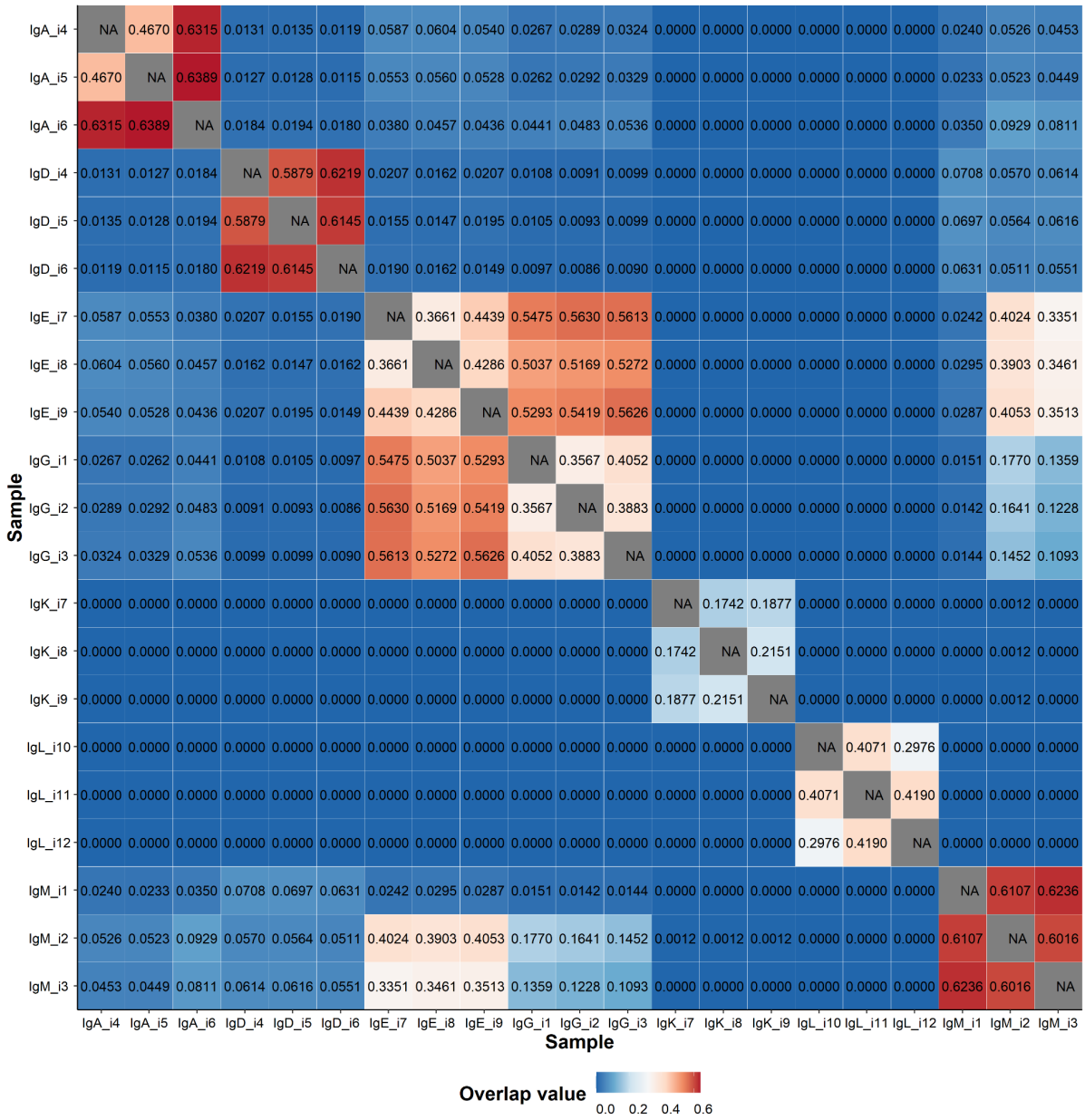


Figure 17: Repertoire overlap between samples. Average CDR3 overlap values within each isotype ranged from 0.1923 to 0.6120. CDR3 overlap between different isotypes generally stayed below 0.100, except for IgE. *Overlap value was measured based on CDR3 amino acid sequence overlap, calculated as the size of the intersection divided by the size of the smaller of the two repertoires.*

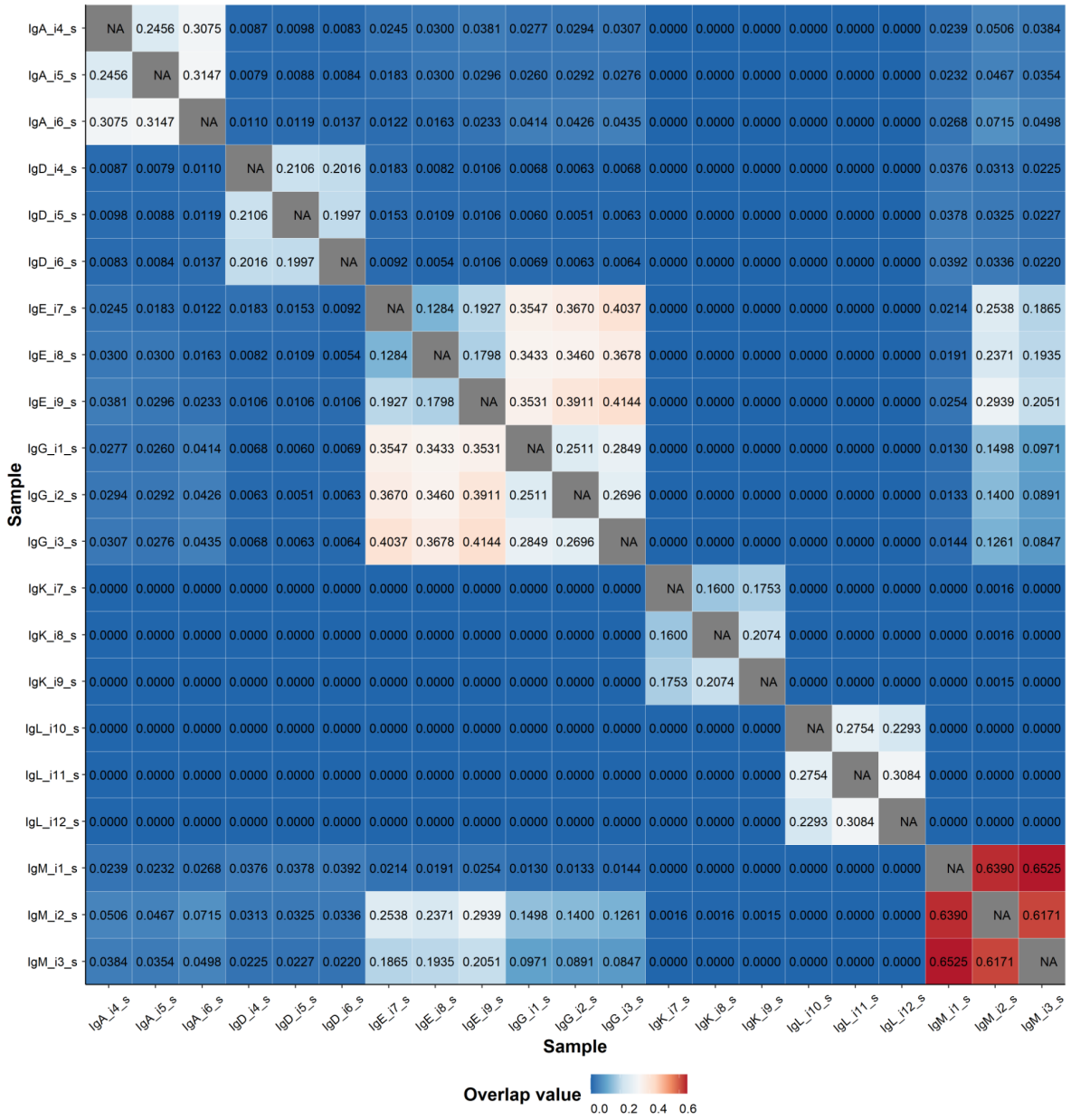


Figure 18: Repertoire overlap between samples after selecting for single-read clonotypes. For the majority of isotypes, the average CDR3 overlap values between technical replicates ranged between 0.1670 and 0.2893, except for IgM samples (0.6362). *Overlap value was measured based on CDR3 amino acid sequence overlap, calculated as the size of the intersection divided by the size of the smaller of the two repertoires.*

Based on the identified V gene in each clonotype, the V gene usage distribution of each isotype was constructed (Figure 19). In general, IGHV3 was the most ubiquitous gene family in IgA (0.4300), IgD (0.3383), and IgE (0.4443), whereas IGHV4 was the most frequent family in IgG (0.3305) and IgM (0.3468). These two gene families combined accounted for over half of all V genes identified in every isotype, which is generally consistent with previous publications [60,64]. Particularly, the most used V genes were IGHV3-30 in IgA (0.1082), IgE (0.1676), and IgG (0.1022) and IGHV4-34 in IgM (0.1062) and IgD (0.1248). However, in the IgG samples, the frequency of usage between IGHV3-30 and IGHV4-34 differed by only a small margin. In addition, between the IgM technical replicates, IgM_i3 (blue) showed a different pattern of V gene usage compared to the two remaining replicates while in other isotypes, V gene usage stayed relatively consistent between replicates. Once again this phenomenon could be explained by the lack of sequencing depth in the IgM samples, thus each technical replicate might have captured a different section of the true sample.

In order to assess the similarities in V gene usage as a whole across samples, pairwise Pearson correlation[5] between V gene frequencies and distance-based clustering between samples were performed (Figure 20A). As expected, V gene frequencies showed perfect correlation within an isotype for the vast majority of samples, except for IgM_i3 as previously observed. The clustering process divided the isotypes into 3 groups: IgA–IgE, IgD–IgM, and IgG which stood in the middle of the other two groups. Between IgA–IgE and IgD–IgM, the correlation coefficient ranged from 0.71 to 0.83 while the pairwise correlation between IgG samples and other isotypes varied from 0.87 to 0.94.

[5] Pearson correlation was introduced in section 3.4.4 “Immune repertoire analysis”.

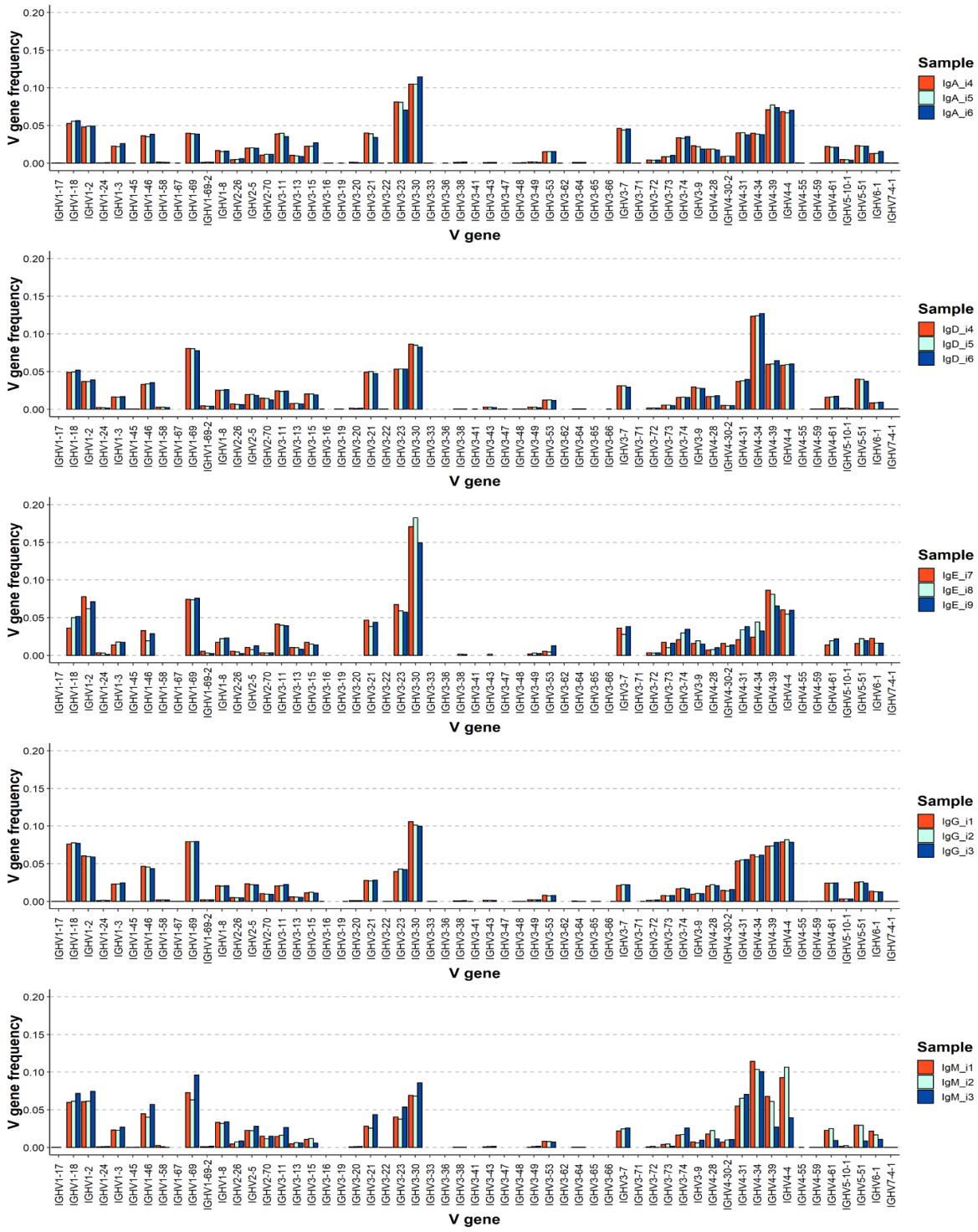


Figure 19: V gene usage pattern across different isotypes. IGHV3 was the most ubiquitous gene family in IgA, IgD, and IgE, while IGHV4 was most ubiquitous in IgG and IgM. IGHV3-30 was most frequently used in IgA, IgE, and IgG, while IGHV4-34 was most frequently used in IgM and IgD. *Relative proportion of each of the heavy chain V genes in different families was adjusted for clonotype count in each sample.*

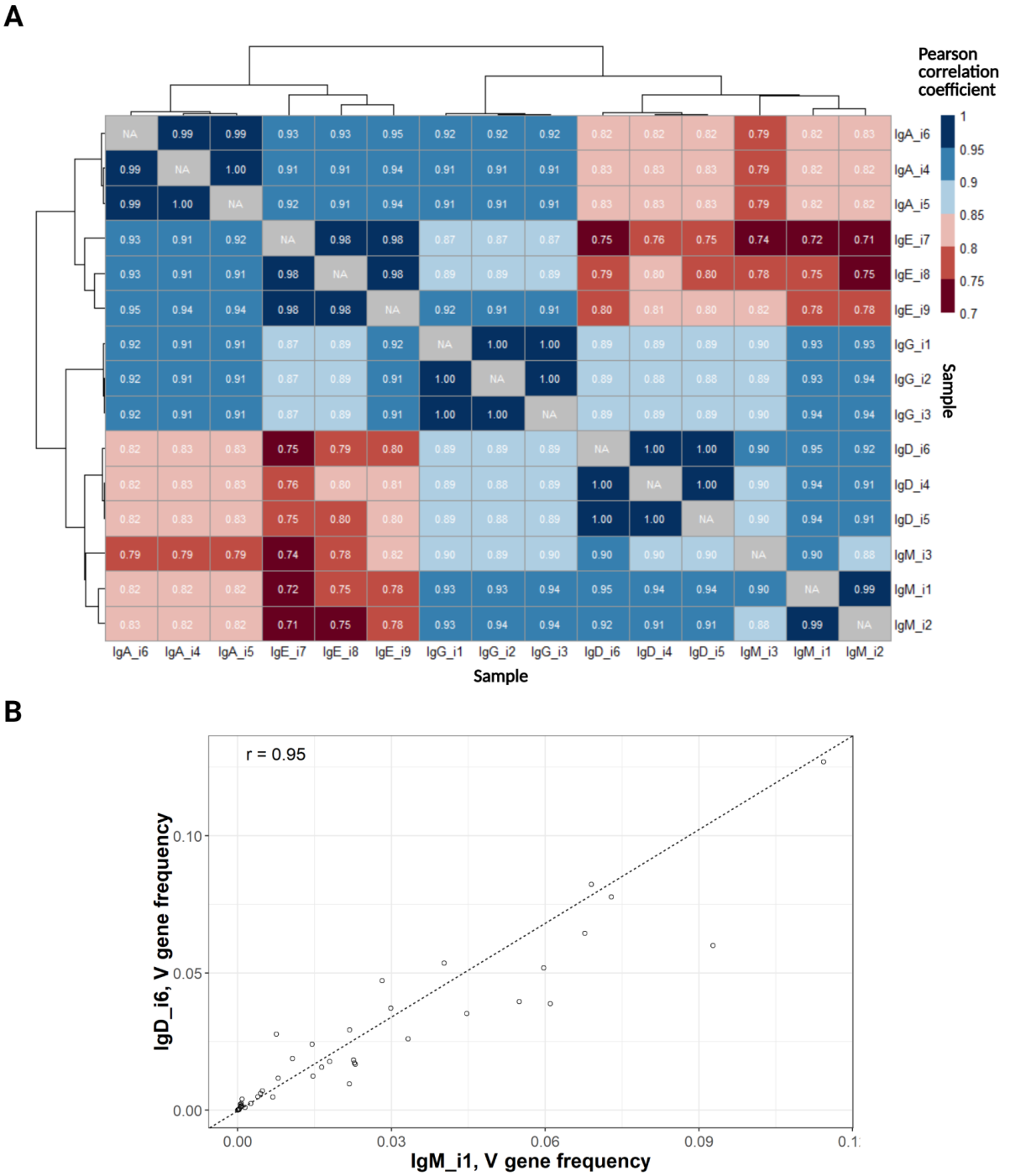


Figure 20: Pearson correlation and hierarchical clustering of heavy chain V gene usage. Samples were clustered into three groups: IgA–IgE, IgD–IgM, and IgG. ***A)*** *Each tile in the heatmap represented the pairwise Pearson correlation of two V-gene repertoires. Color gradient indicated the magnitude of correlation. Clustering categorized the samples into 3 groups: IgA–IgE, IgD–IgM, and IgG.* ***B)*** *An example of Pearson correlation of V gene repertoires for two samples.*

Aside from V gene usages in the heavy chain sequences, the V gene usage distribution was also measured in the light chain samples (Figure 21). For the κ chain, IGKV1 was the most frequently utilized gene family (0.6628), followed by IGKV3 (0.1981). IGKV1-39 (0.2100) and IGKV1-5 (1419) were the 1st and 2nd most frequent V genes, respectively. For the λ chain, IGLV2 was the most frequently utilized gene family (0.4885), followed by IGLV3 (0.2762). IGLV2-11 (0.2788) and IGLV2-14 (0.1979) were the 1st and 2nd most frequent V genes, respectively.

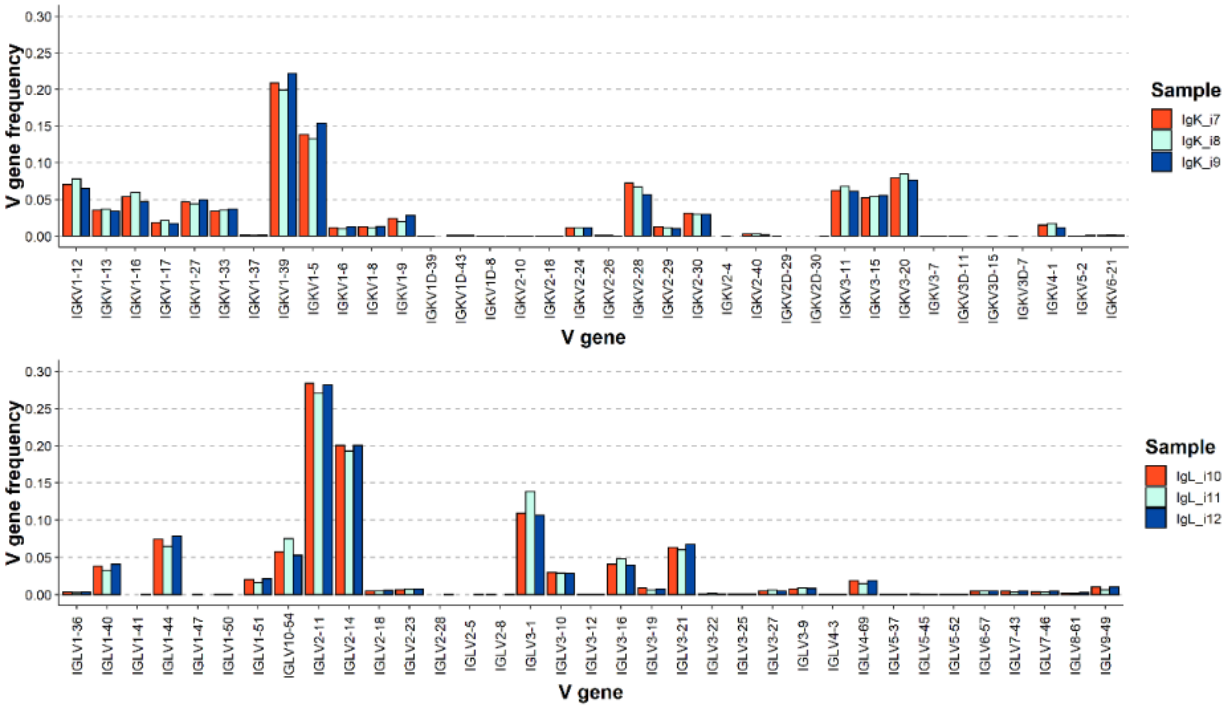


Figure 21: V gene usage pattern in the light chain sequences: κ and λ. The IGKV1 gene family comprised the majority of V genes used in the κ chain, and IGLV2 was the most utilized V gene family in the λ chain. *Relative proportion of each of the V genes in different families was adjusted for clonotype count in each sample.*

In addition to analyses such as CDR3 length distribution, CDR3 overlap, and V gene usage frequency that are dependent on the clonotype's sequence, another sequence-independent characteristics such as clonotype frequencies can also provide important information about a repertoire, especially the clonal diversity and state of clonal expansion. An alpha Diversity profile of each sample was constructed based on clonal frequencies and normalized into an

Evenness profile [6]. Since Evenness describes the extent in which a clonal frequency distribution is distanced from a uniform clonal frequency distribution (no clonal expansion), a steeper curve would signify a higher degree of clonal expansion. Of particular note was the intersection of the profiles for the IgM and the IgD/IgE samples, demonstrating the usefulness of looking at diversity and clonal expansion profiles instead of deciding on a single diversity index, since different indices would provide contradicting conclusions depending on the sub-clonal expansion in a repertoire. The samples' Evenness profiles were divided into 2 groups: IgM–IgE–IgD with a lower degree of clonal expansion and IgA–IgG–light chain samples with a higher degree of clonal expansion (Figure 22). Pairwise Pearson correlation between samples' Evenness profile was also calculated and clustering of results confirmed previous observations in the Evenness profiles grouping the samples according to clonal expansion status (Figure 23).

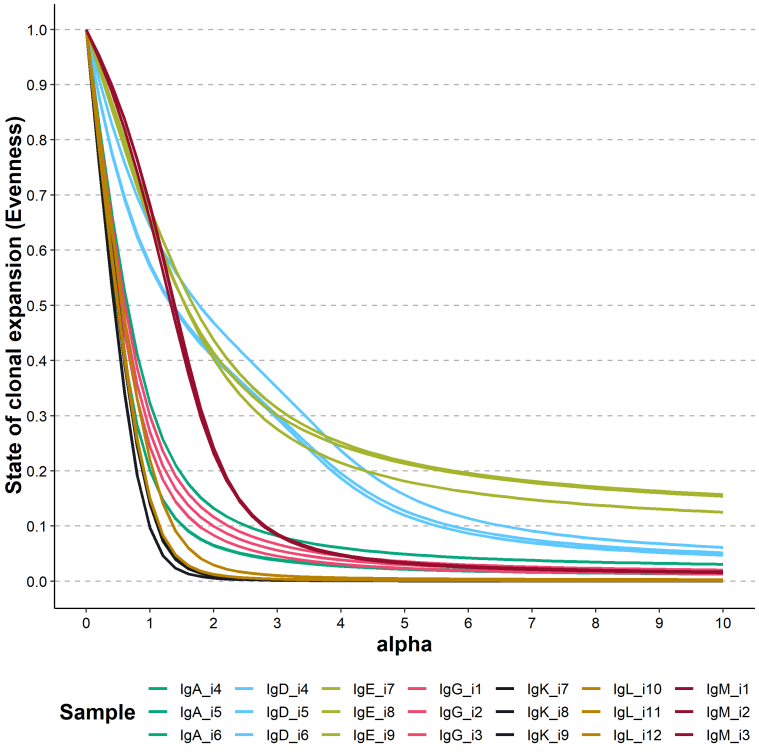


Figure 22: Evenness profiles across isotypes. IgM, IgE and IgD samples showed a high degree of clonal expansion while IgA, IgG, and light chain samples showed a high degree of clonal expansion. *Evenness profiles were constructed as a collection of Evenness values at a range of alpha from 0 to 10 (step size 0.2) and colored according to their respective isotypes.*

[6] Diversity indices, Diversity profiles and Evenness profiles and their applications were introduced in section 3.4.4 "Immune repertoire analysis".

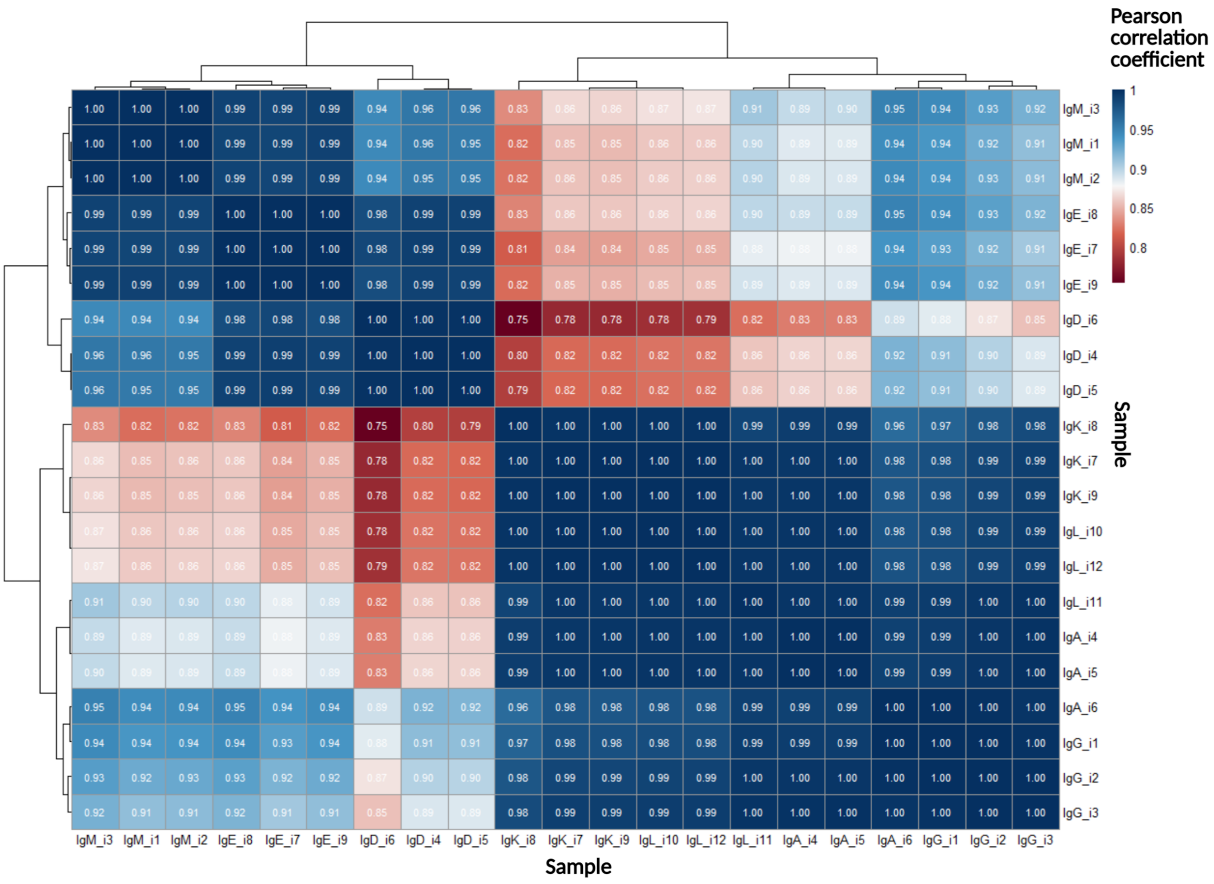


Figure 23: Pearson correlation and hierarchical clustering of Evenness profiles. Two main clusters were identified based on their respective Evenness profile values: IgM–IgE–IgD and IgA–IgG–light chain samples. *Each tile in the heatmap represented the pairwise Pearson correlation between two Evenness profiles. The color gradient indicated the magnitude of correlation.*

4.2 Single-cell B-cell receptor sequencing

4.2.1 B-cell encapsulation

In order to evaluate the efficiency of the cell encapsulation process, we formulated an experiment in which samples varied in the number of starting B-cell input, from 100000 to 1000000 cells. The B cells were stained prior to encapsulation with Hoechst 33342 Solution (20 mM) in order to be visualized under fluorescence microscopy (Figure 24).

The Dolomite Nadia system provided a good performance for input up to 500000 cells, where the fraction of droplets containing multiple cells stayed below 5%. However, at the highest cell input (1000000 cells), the multiplet's fraction sharply increased (15.3%), making this input level unsuitable for downstream applications (Figure 25). Based on the results obtained and balancing between throughput and reliability, we decided to set the input to 500000 B cells per microfluidics chip for our library preparation.

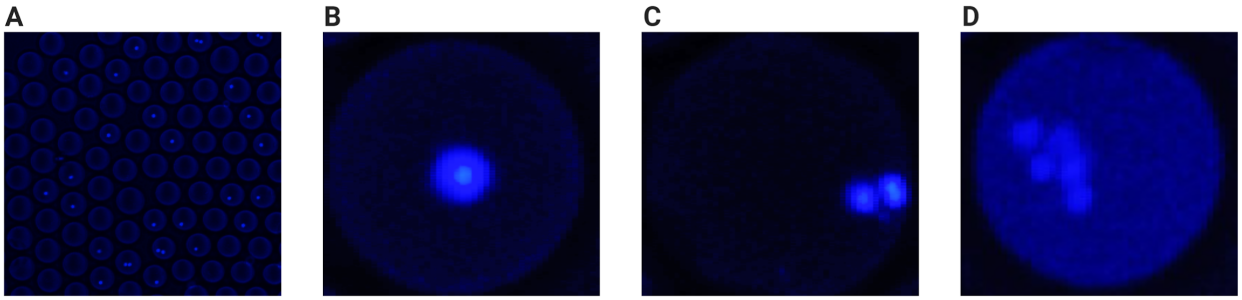


Figure 24: Droplet visualization under fluorescence microscopy, with B cells stained blue. ***(A)*** *Representative image of products from the droplet encapsulation process. Droplets were categorized as either single-cell droplets* ***(B)*** *or multiple-cell droplets, which contained either two* ***(C)*** *or more than two* ***(D)*** *B cells in an oil droplet.*

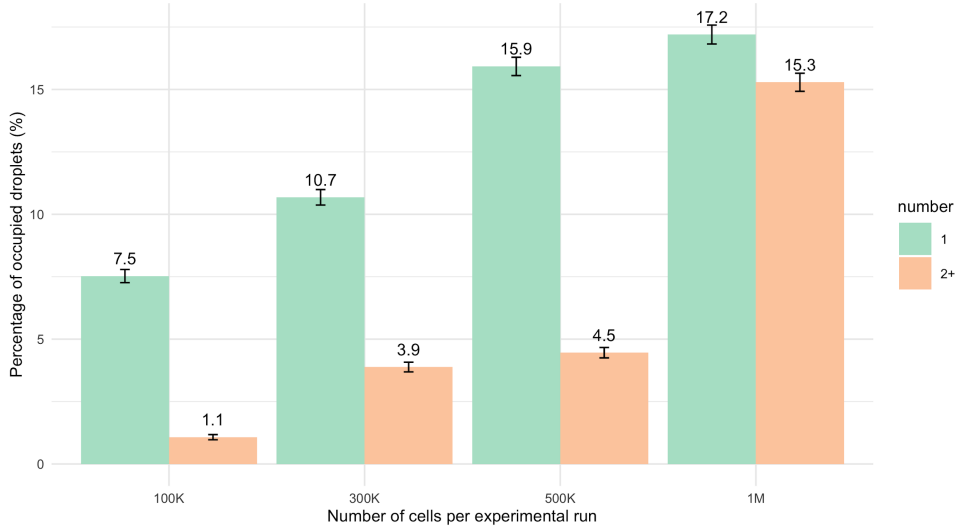


Figure 25: B-cell encapsulation efficiency. *The droplets were classified as either single-cell droplets (green) or multi-cell droplets (orange). Capture efficiency was measured at a range of cell input from 100 thousand to 1 million B cells per microfluidic chip. The percentage is calculated as the number of droplets containing cells over the total number of droplets in a field of view, averaged over 30 fields of view.*

4.2.2 Assessment of library quality

After the captured B cells were lysed and the released mRNAs were reverse transcribed into cDNAs, subsequent steps in the library preparation workflow were performed. The resulting products after emulsion breakage were divided into 3 fractions: fraction 1 contained cDNA products prior to any cleanup, fraction 2 contained cDNA products after purification with AMPure XP Beads, and fraction 3 contained cDNA products after purification with MinElute purification kit. Each fraction was accompanied by a no polymerase control (no Q5) and a no template control (NTC).

The resulting libraries analysed on BioAnalyzer did not show any product at the expected size (~600 bp). Moreover, there were unexpected product bands at around 150–200 bp (Figure 26). Thus, it is likely that the in-droplet reverse transcription process was not successful and did not generate the required cDNA template for subsequent reactions. There are several possible explanations for this result. Firstly, the reverse transcriptase enzyme might not work properly under a very low template concentration [161]. Secondly, the template-switching reaction had a low efficiency. Thus, the universal adapter sequence could not anneal to the 5' end of the cDNA and subsequent reactions could not amplify the products (Figure 9). Finally, the presence of non-specific product bands at around 150–200 bp suggested that primer dimers formed during amplification, requiring a re-evaluated of the design and reaction conditions. In short, the library preparation workflow for single-cell B-cell receptor sequencing did not lead to a viable product and therefore requires further evaluations and redesigns in the future.

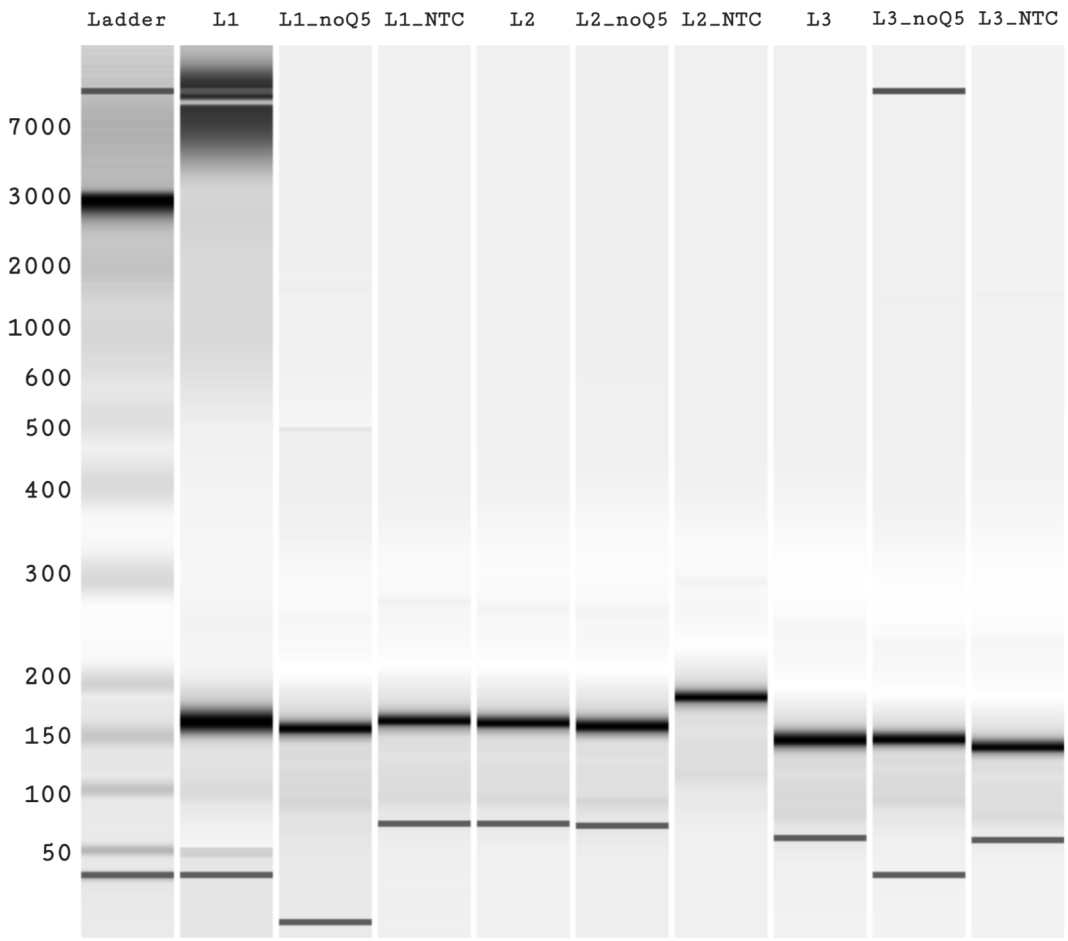


Figure 26: Representative example of single-cell B-cell receptor libraries. *Each library consists of 3 fractions, each containing a no polymerase control (no Q5) and a no template control (NTC): products from emulsion breakage (L1), products from emulsion breakage purified with AMPure XP Beads (L2), and products from emulsion breakage purified with MinElute purification kit (L3).*

4.3 Benchmarking of the antibody LC-MS/MS pipeline

4.3.1 Proof-of-concept and pilot experiments

4.3.1.1 Antibody peptides detection with LC-MS/MS

As a proof-of-concept, we wanted to test the capability of LC-MS/MS in detecting antibody related peptides. Two different treatments were conceived: one where the antibodies were digested with GingisKHAN, leaving only the Fab fragment, and the other where intact antibodies were used. Each treatment contained 4 samples: 14.2 µg of blood-isolated IgG1, 14.2 µg of PGDM1400 (Supplementary Table 7), 14.2 µg of IgG1 + 14.2 µg of PGDM1400, and 14.2 µg of

IgG1 + 142 ng of PGDM1400. All samples were cleaved with chymotrypsin, followed by trypsin prior to mass spectrometry.

In the first 4 samples where the antibody was treated with GingisKHAN, no antibody-related peptide was detected. Sample 6 (contained only PDGM1400) contains only the antibody-related matches. Sample 7 (higher PDGM1400 concentration, mixed with IgG1) contained more antibody-related matches than sample 8 (lower PDGM1400 concentration, mixed with IgG1) (Figure 27). Most of the antibody-related peptides matched with the constant region (which can be similar to other antibodies), only a small fraction matched with the CDR3 region (Figure 28).

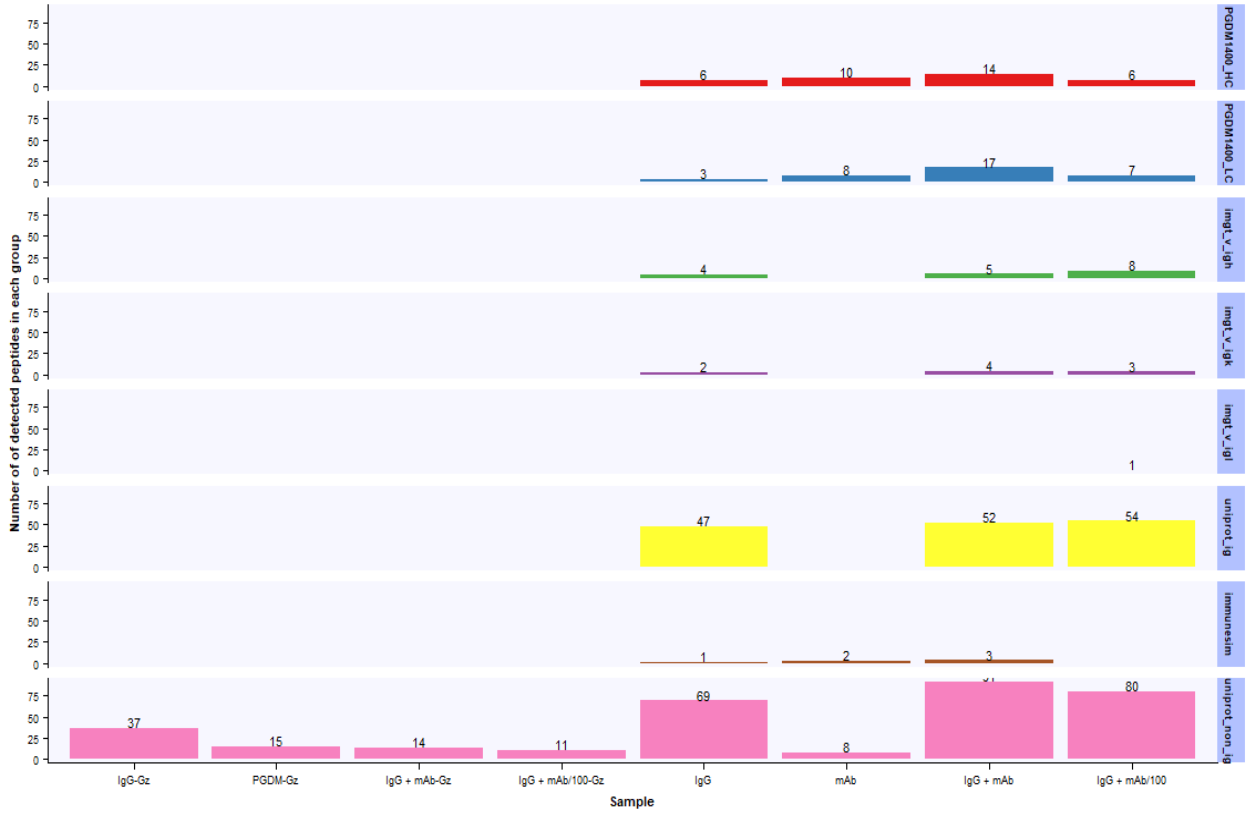


Figure 27: Detection of antibody-related peptides using LC-MS/MS. No antibody-related peptides were detected in samples treated with GingisKHAN (Gz), while only PGDM1400-specific peptides were detected in the PGDM1400 sample (mAb). *Samples 1–4 were treated with GingisKHAN, and samples 5–8 utilized intact antibodies. All samples were digested with chymotrypsin + trypsin.*

Further examining the peptide fragments that matched with either the LC or HC of PDGM1400 detected 1 peptide fragment that had significant overlap (≥ 3 aa) with the CDRL3 region of PDG1400 in samples 6 and 7, suggesting the feasibility of identifying known antibodies from complex mixtures (Figure 28).

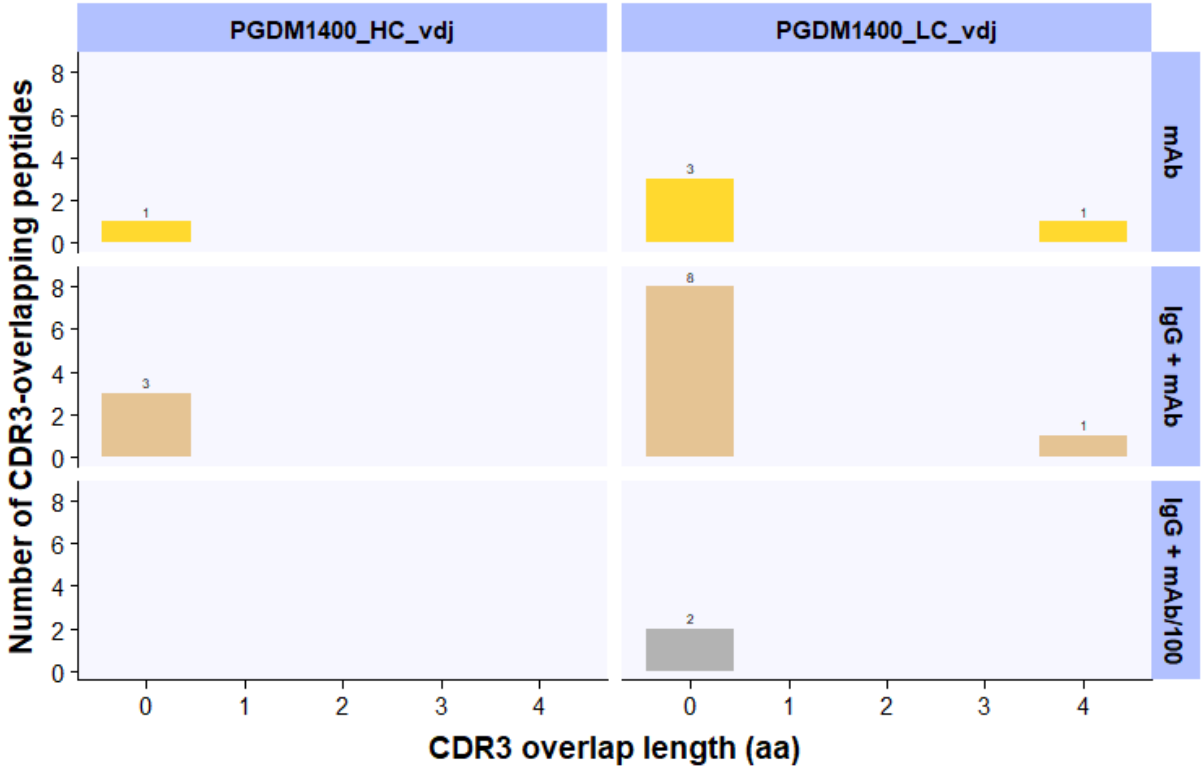


Figure 28: CDR3 overlap distribution of detected peptides. CDR3 overlapping peptides were detected only on the light chain, not on the heavy chain. *Each detected peptide was compared to the known CDR3 sequence (HC and LC) of the PDGM1400 mAb and overlap length was calculated. An overlap is considered significant if the overlap length ≥ 3 aa.*

4.3.1.2 Performance comparison between MS settings

Two mass spectrometry settings available at the Proteomics Core Facility at Oslo University Hospital with different throughputs were compared, which can either process up to 100 samples per day or up to 60 samples per day (100 s/d and 60 s/d). Comparing the same starting sample (14.2 µg of blood-isolated IgG1 + 14.2 µg of PGDM1400) digested by the same enzymes in two different settings showed that more peptides overall were detected in the lower-throughput setting (60 s/d). However, the number of antibody-related peptides detected showed no

noticeable differences, especially those that mapped to the PDGM1400 sequences. Most of the extra peptides detected in the 60 s/d setting were unrelated to the antibody sequences (Figure 29).

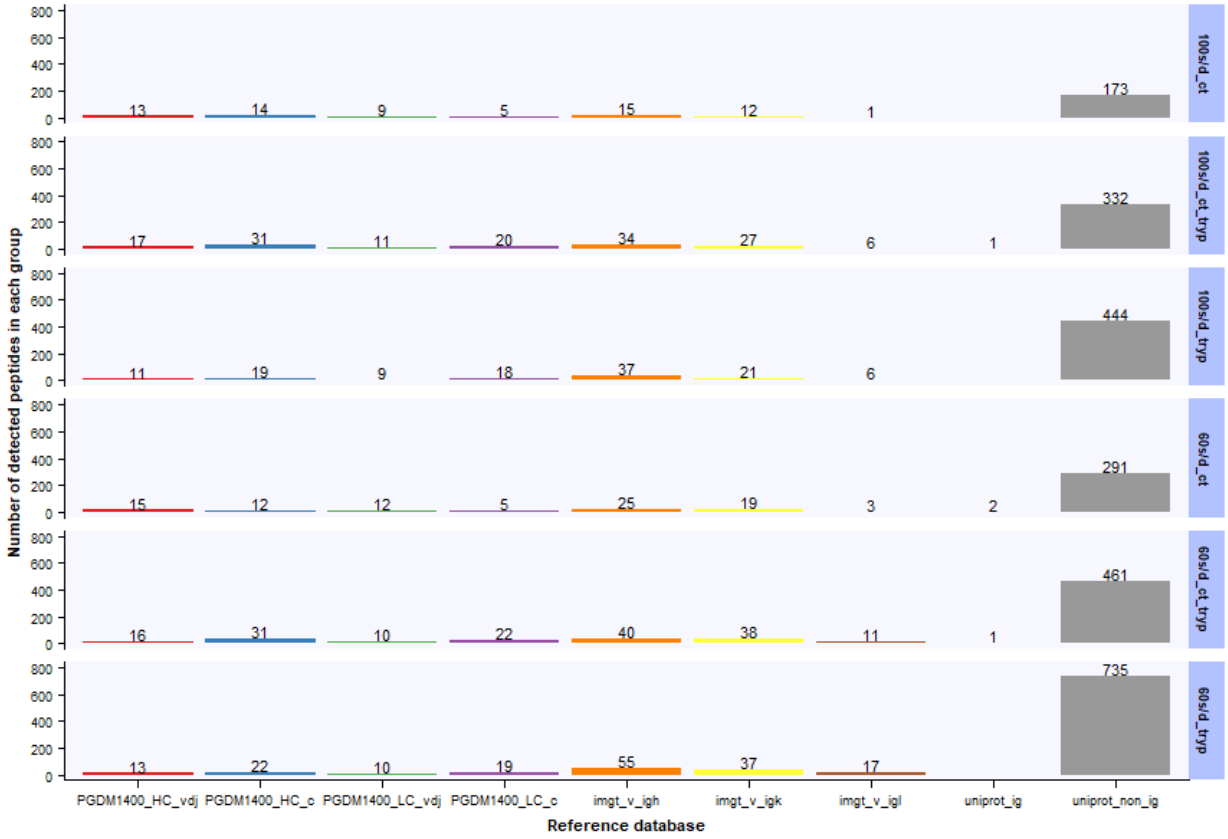


Figure 29: Comparison of LC-MS/MS pipelines on detection of antibody-related peptides. Similar numbers of antibody-related peptides were detected between two mass spectrometry settings. *Each sample (14.2 µg of blood-isolated IgG1 + 14.2 µg of PGDM1400) was digested with either chymotrypsin, trypsin, or chymotrypsin + trypsin and underwent LC-MS/MS with a throughput of 100 samples per day or 60 samples per day, respectively.*

Closer examination of CDR3-related peptides revealed that the number of detected peptides were almost identical in both workflows, except in the chymotrypsin-treated samples where there were more peptides detected in the lower-throughput setting (Figure 30). Overall, the lower-throughput setting did not seem to provide any additional advantages regarding antibody identification and thus we decided to use the 100 s/d mass spectrometry setting for all subsequent experiments.

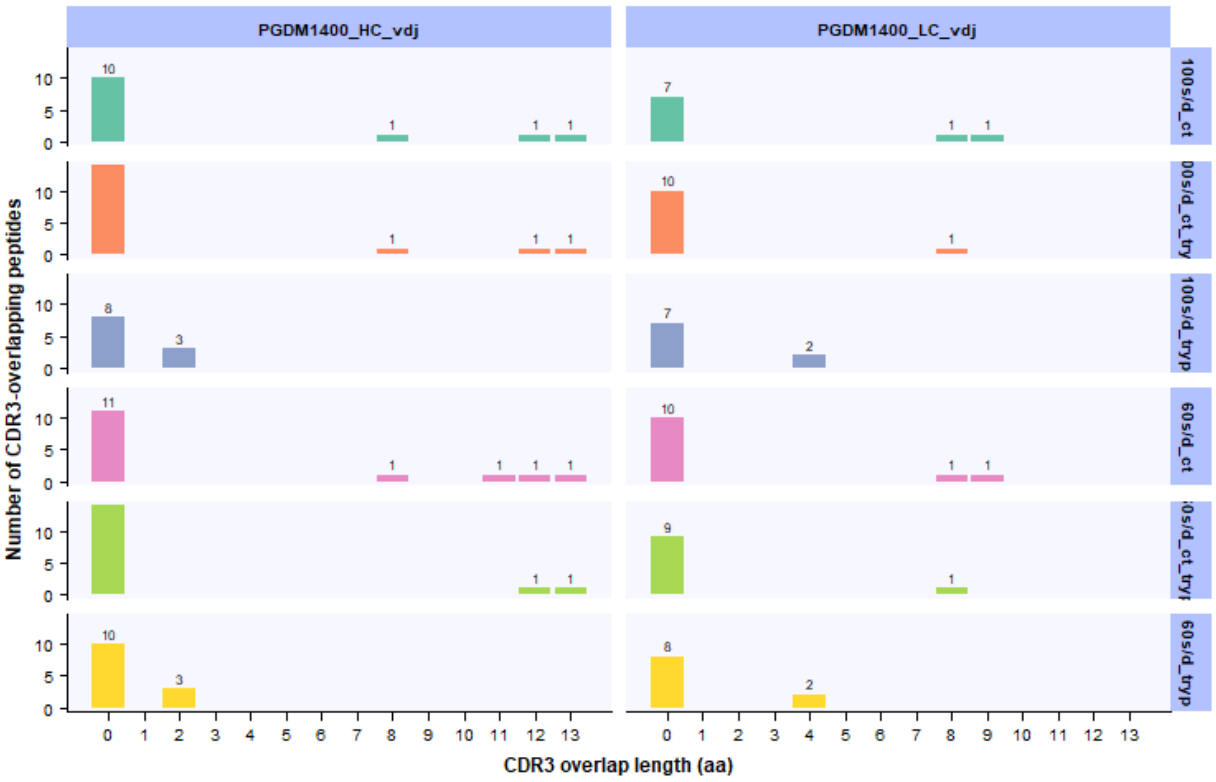


Figure 30: Comparison of LC-MS/MS pipelines on detection of CDR3-related peptides. The number of CDR3-overlapping peptides detected was almost identical between two different mass spectrometry settings. *Each detected peptide was compared to the known CDR3 sequence of the PDGM1400 mAb and overlap length was calculated. An overlap is considered significant if the overlap length ≥ 3 aa.*

For the same detected peptide fragments at the same antibody input (14.2 µg of blood-isolated IgG1 + 14.2 µg of PGDM1400), in two different mass spectrometry settings (100s/d and 60s/d), and digested in three different conditions (chymotrypsin, trypsin, and chymotrypsin + trypsin), the signal intensity ratio between the two is close to 1, with only a few outliers (Figure 31).

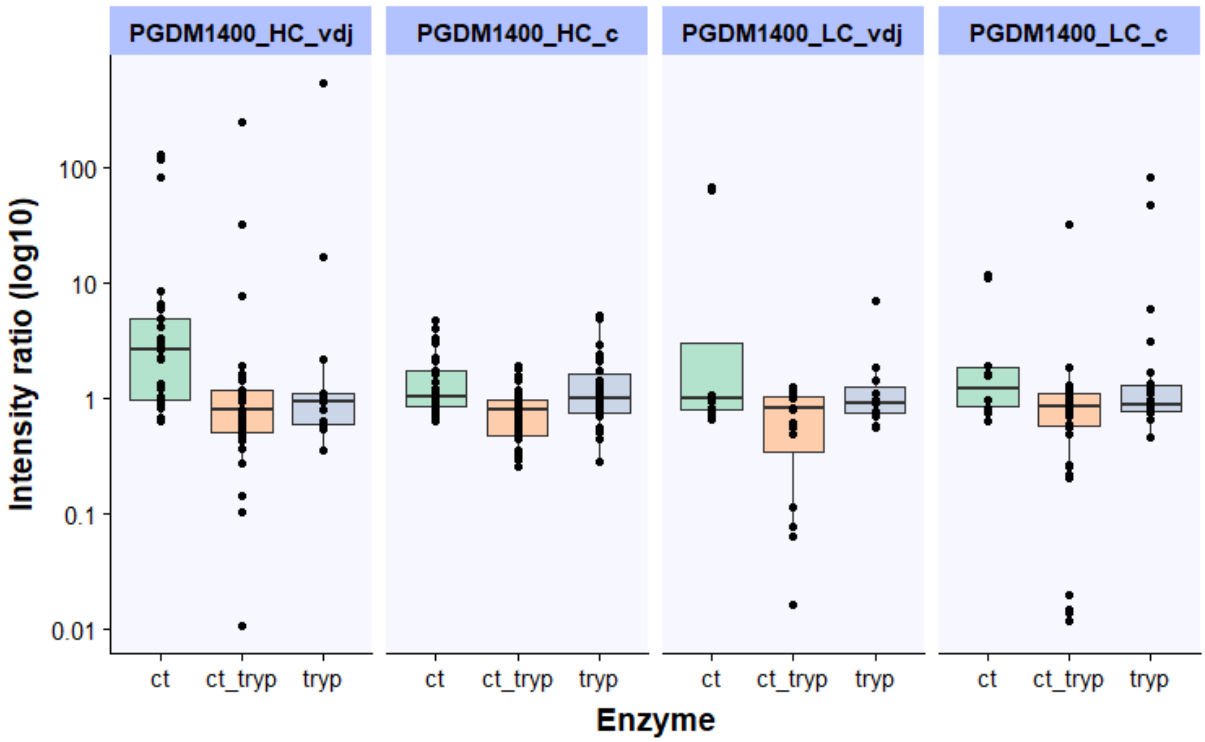


Figure 31: Peptide signal intensity comparison. Most of the identical peptides detected in different samples possessed similar signal intensities. *The intensity ratio is calculated by comparing signal intensity between identical peptides cleaved by the same enzyme and detected in both mass spectrometry settings.*

4.3.1.3 Correlation between intensity ratios and concentration ratios in peptides

With the results from the comparison between MS settings, we set out to investigate whether MS/MS signal intensity ratios correlate and reflect mAb input concentration ratios. For that purpose Trypsin-digested BSA MS standard (CAM-modified) (New England Biolabs #P8108S) at 5 pmol, 500 fmol, 50 fmol, and 5 fmol, with the addition of 100 ng of HeLa Protein digest standard (Pierce #88328) in all BSA samples were used. The results showed a high correlation between signal intensity ratios and peptide concentration ratios (Figure 32), which suggests that signal intensity can be a good indicator for peptide concentration in samples with known sequence.

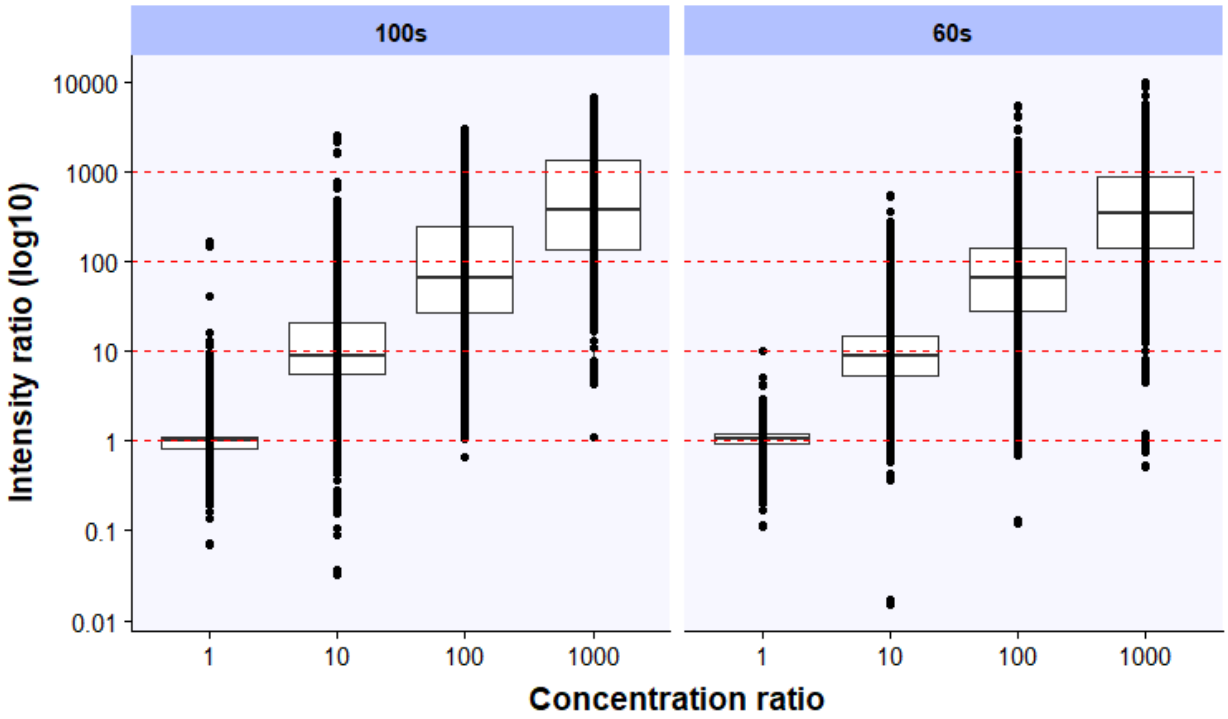


Figure 32: Peptide signal intensity comparison in BSA + HeLa digests. In both settings, the signal intensity ratios of the detected tryptic peptides followed closely with their respective concentration ratios. *Intensity ratios and concentration ratios of detected BSA and HeLa peptides are calculated between identical detected peptides at different amounts of inputs.*

4.3.1.4 Limit of detection for LC-MS/MS in antibody identification

In order to test the limit of LC-MS/MS in antibody identification, we devised an experimental setup with samples in varying levels of antibody input at 1 µg, 100 ng and 10 ng (6.67 pmol, 667 fmol, and 66.7 fmol, respectively). The antibodies were either blood-isolated (sample 1–3) or from the PDGM1400 mAb (sample 4–6) and digested with chymotrypsin + trypsin. The number of peptides detected tended to decrease with lowering concentration, as expected. At 10 ng input (66.7 fmol), no PDGM1400-related peptides were detected, suggesting the lower limit of input for antibody identification with chymotrypsin + trypsin digestion (Figure 33).

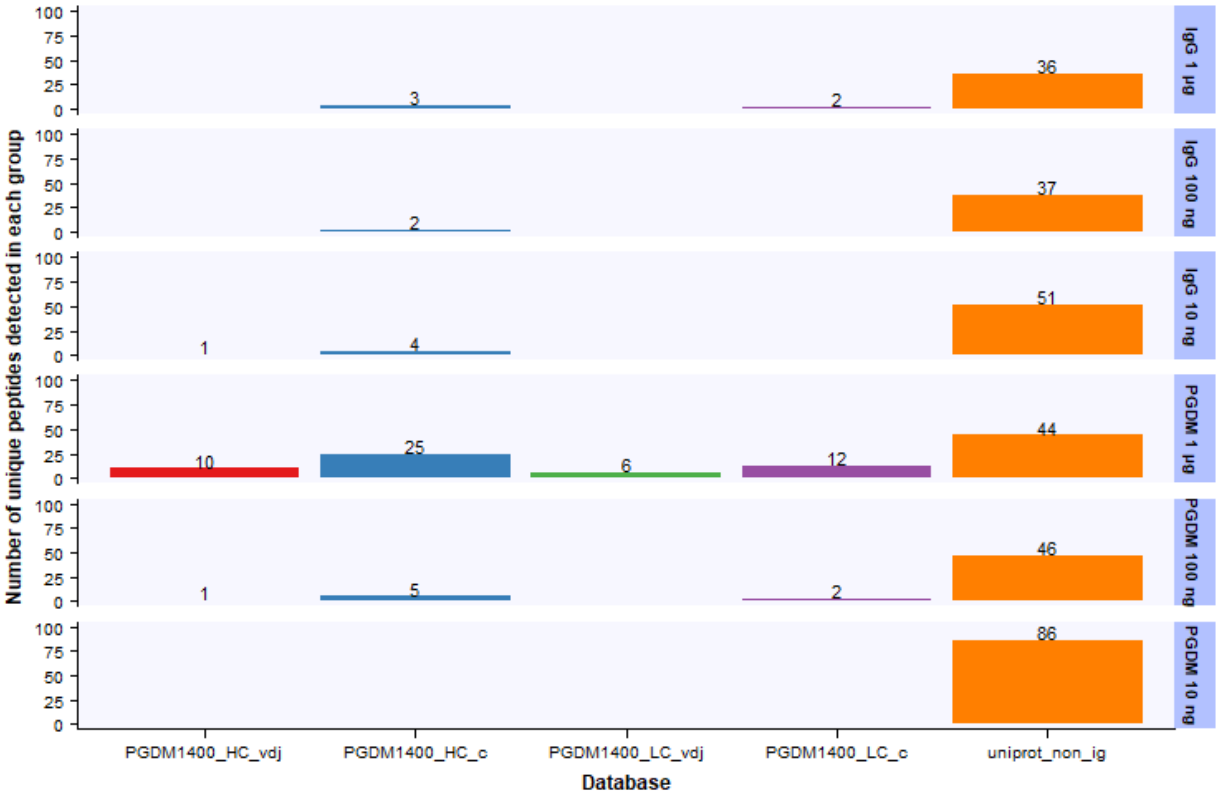


Figure 33: Comparison of monoclonal antibody concentration on peptide detection.The number of detected peptides decreased with declining concentrations. No antibody-related peptides were detected at an input of 10 ng (66.7 fmol). *Each antibody sample was cleaved by chymotrypsin + trypsin with input of 1 µg, 100 ng and 10 ng, respectively.*

4.3.2 Detection of monoclonal antibodies at different concentrations

The pilot experiments helped establish the foundations for further benchmarking of antibody LC-MS/MS. As a result, we decided to expand further into other mAbs in order to investigate whether the detection limit of antibodies would be impacted by the interaction between antibody sequences and the choice of digestion enzymes. An experiment setup was devised where 3 different mAbs were utilized: Briakinumab, PGDM1400, and PGT121 (Supplementary Table 7). The input of each antibody varied from 1 µg, 100 ng, 10 ng, to 1 ng (6.67 pmol, 667 fmol, 66.7 fmol, and 6.67 fmol, respectively). Each sample was further split into 3 different enzymatic treatments: chymotrypsin, trypsin, and chymotrypsin + trypsin. Every sample was performed in 3 technical replicates (each digested antibody sample went through the LC-MS/MS pipeline three times and the resulting data was combined together).

Prior to the experiment, mAb sequences were digested in-silico according to ExPASy cleavage rules, with 0 miscleavage allowed [135]. Subsequently, the peptides identified by MaxQuant were matched with the theoretical peptides, separated by either heavy or light chain sequence and by the enzyme utilized (Figure 34). In general, the in-silico digestion process generated peptides mostly clustered below 30 aa in length, with some exceptions. However, the number of theoretical peptides actually detected in the experimental samples was very low, with the majority of detected peptides being on the light chain instead of the heavy chain sequences. Particularly, in the case of Briakinumab heavy chain, only 1 peptide was detected with trypsin digestion and none with other enzymes. The results showed that only a small fraction of peptides detected followed exactly the rules devised and the majority would contain miscleavages or were not detected altogether.

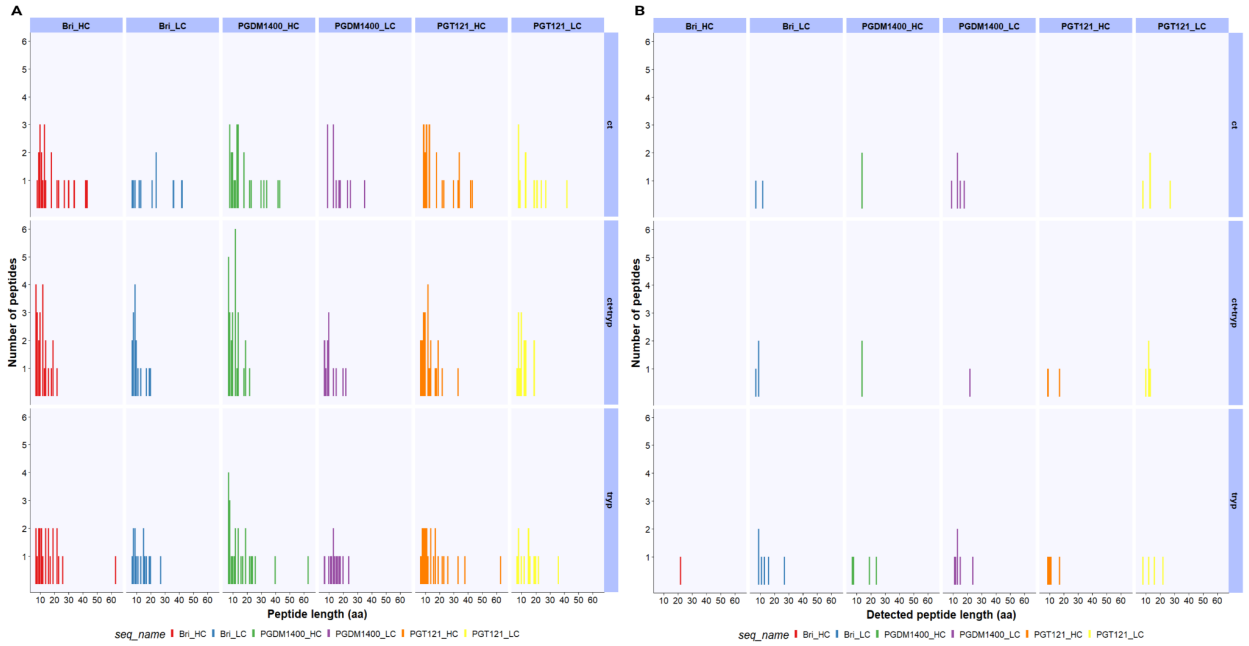


Figure 34: Distribution of enzyme-digested peptides in mAbs. Most detected peptides resided on the light chain. In Briakinumab, only 1 peptide was detected and was derived from trypsin. ***A)*** *Theoretical digested peptide distribution of mAbs, with 0 miscleavage.* ***B)*** *Detected peptide distribution of mAbs that matched theoretical sequences.*

All MS peptides detected by MaxQuant were next overlapped with the CDR3 regions of the mAbs and the number of significant (overlap length ≥ 3 aa) CDR3-overlapping peptides were counted (Figure 35). With Briakinumab, the number of peptides detected as well as the lower detection limit differed greatly between enzymes. Chymotrypsin-digested peptides only mapped to the light chain and only at 1000 ng (6.67 pmol) input, with no peptides at lower inputs. Chymotrypsin + trypsin performed better with a detection limit of 10 ng (66.7 fmol) yet still inferior to trypsin where peptides were identified at only 1 ng (6.67 fmol) of input. Peptides from PGDM1400 and PGT121 samples, on the other hand, only mapped to their respective CDRL3 region and not CDRH3 regions. In PGDM1400, chymotrypsin again performed worse compared to other enzymes when it came to the limit of detection, with a steep drop in detected peptides from 1000 ng (6.67 pmol) to 100 ng (667 fmol), and no detected peptides at 10 ng (66.7 fmol) and lower. On the other hand, peptide detection in PGT121 was noticeably different than the other two mAbs, with similar performance across all three digestion strategies but only at 1000 ng (6.67 pmol). At inputs of 100 ng (667 fmol) and lower, no CDR3-overlapping peptides were detected in PGT121 samples.

Examining the relationship between concentration ratios and signal intensity ratios of the detected peptides revealed similar results to the previous experiments (Figure 31, 32), where the two ratios roughly follow one another, albeit to a different degree with different enzymes (Figure 36). The two ratios matched up best in trypsin-digested peptides, slightly less so in chymotrypsin + trypsin digestion, and worst in chymotrypsin peptides. This could be explained partly by the low digestion efficiency of chymotrypsin, leading to incomplete digestion and skewed peptide concentrations. Nevertheless, the results showed that signal intensity ratios could be useful as an indicator when comparing the amount of Abs in a sample.

In short, the ability to identify mAbs based on CDR3-overlapping peptides and the limit of detection varied dramatically depending on the sequence of the mAbs themselves and the digestion strategy employed.

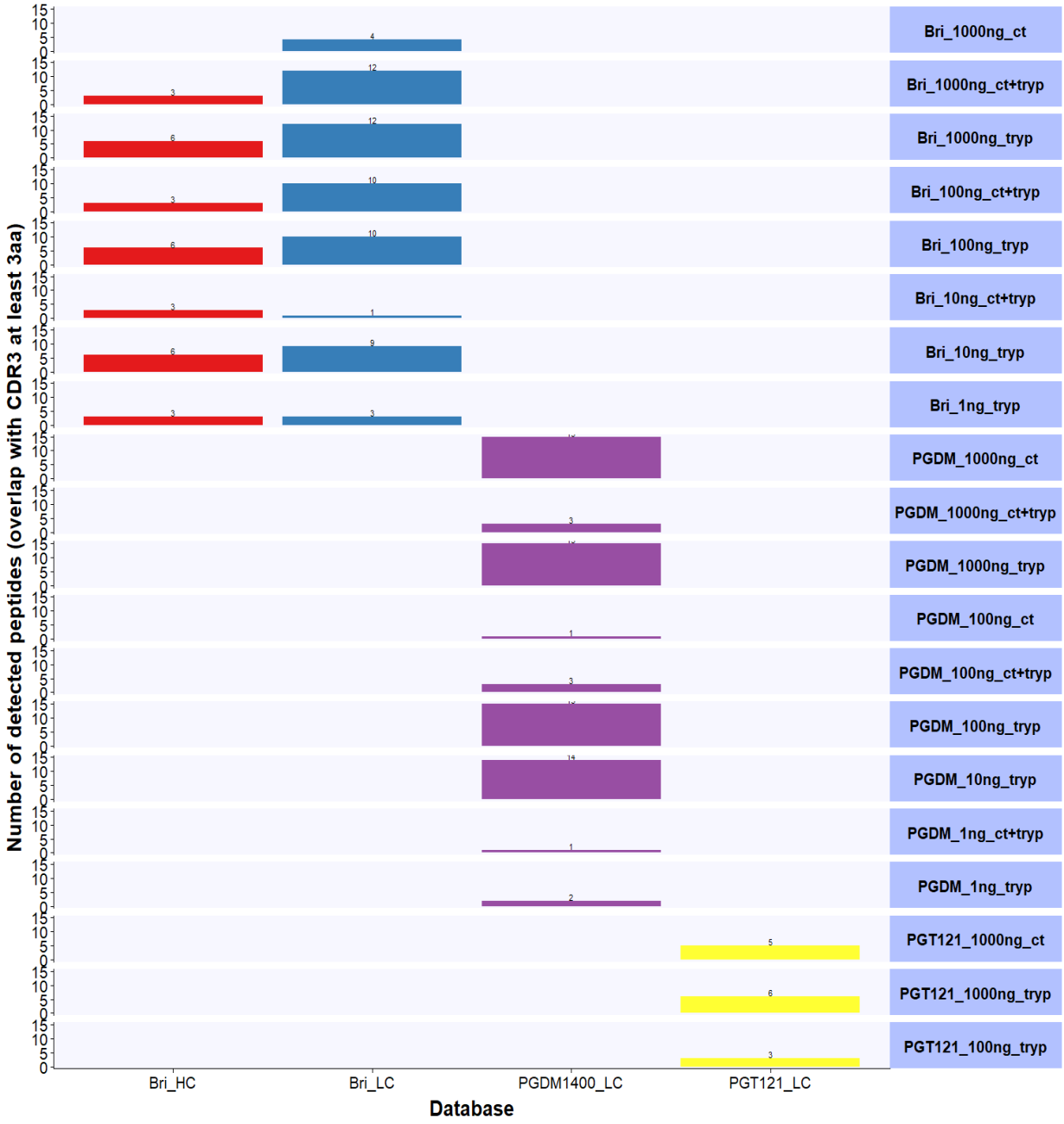


Figure 35: CDR3-overlapping peptides with overlap length ≥ 3 aa detected in different mAbs. Trypsin performed best regarding the number of peptides detected and limit of detection. No peptides were mapped to the CDRH3 region of PGDM1400 and PGT121.

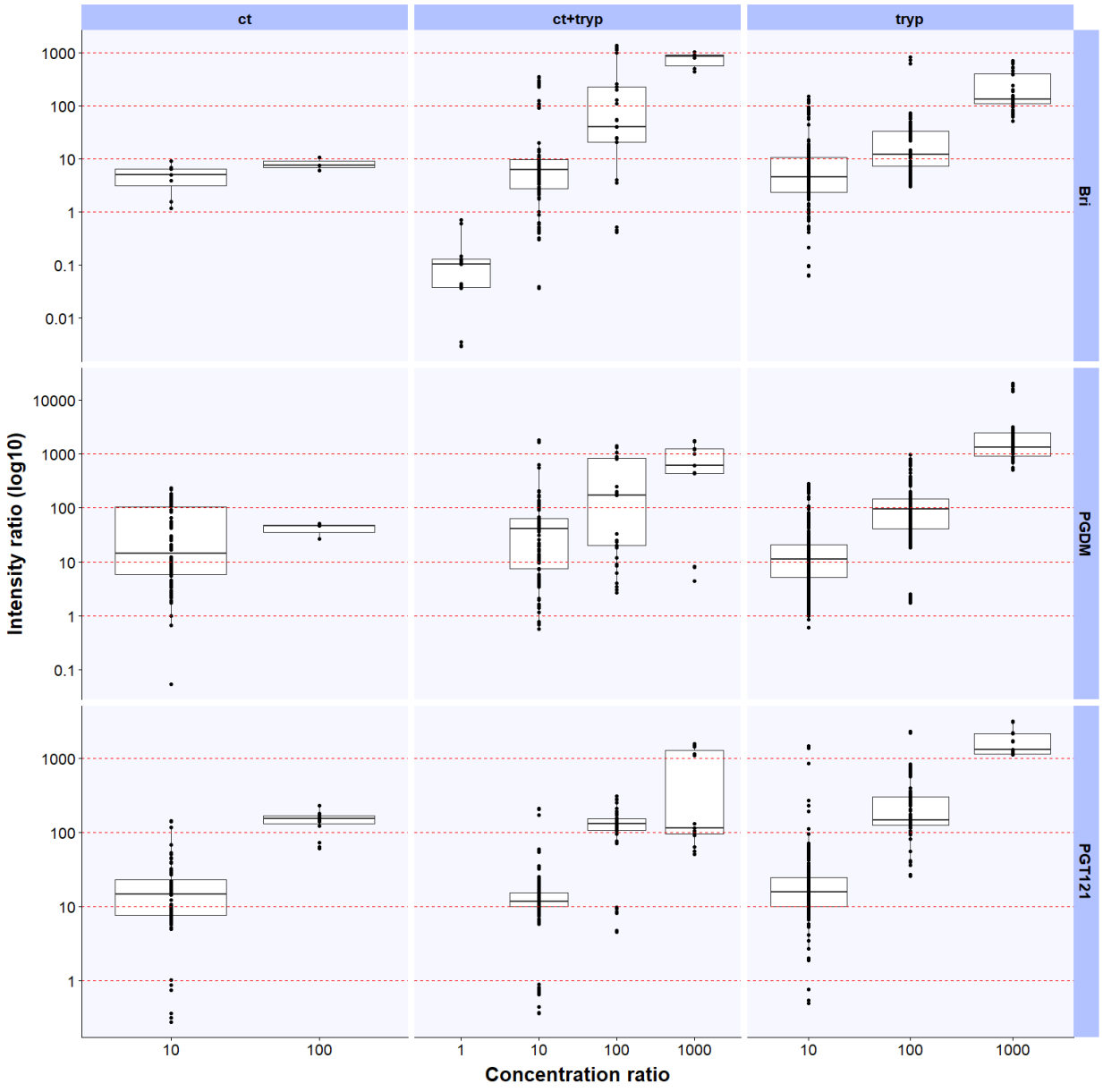


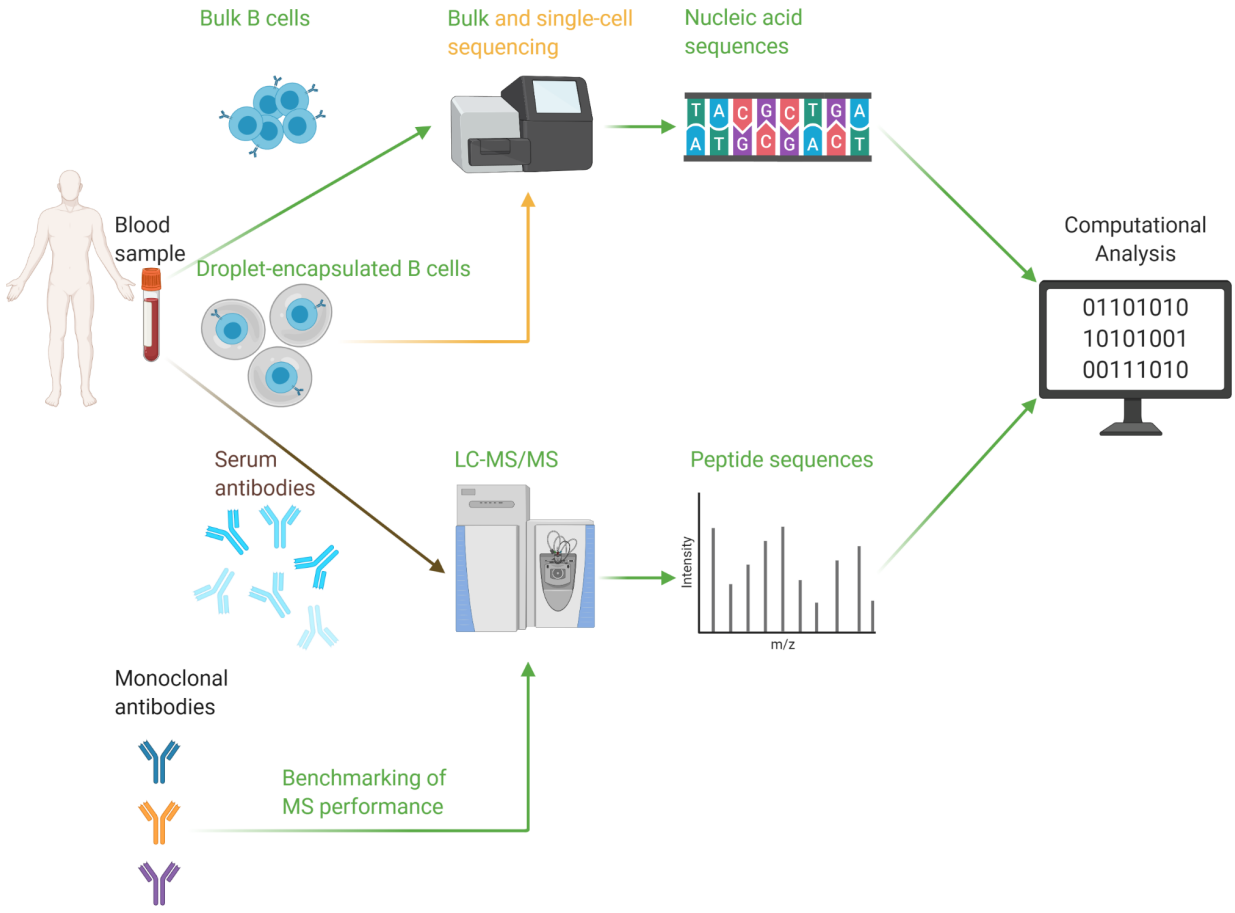
Figure 36: Peptide signal intensity comparison in monoclonal antibody samples. Equivalence between concentration ratios and signal intensity ratios was highest in trypsin-digested peptides and lowest in chymotrypsin-digested peptides. *Intensity ratios and concentration ratios are calculated between identical detected peptides at different amounts of inputs.*

In summary, several insights were gained after conducting the pilot and benchmarking experiments:

- It is possible to confirm the presence of an antibody with known sequence by mapping the MS peptides to the CDR3 region of the antibody sequence (Figure 28).
- Increasing runtime for the LC-MS/MS pipeline does not result in better sensitivity regarding peptide identification (Figure 29).
- Between identical peptides detected in different samples, the signal intensity ratios reflect the concentration ratios of the samples (Figure 31, 32, 36).
- The number of antibody-related peptides decreases with respect to declining sample input (Figure 33).
- Only a small fraction of theoretical peptides can be detected by MS (Figure 34).
- Limit of detection varies significantly depending on the antibody sequence and digestion enzyme utilized (Figure 35).

5. Discussion

In this project, we aimed to establish a multi-pronged approach for examining the composition of the B-cell receptor (genomics) and antibody repertoire (proteomics), combining bulk BCR sequencing with developing novel single-cell BCR sequencing methods, and additionally comparing with data generated from proteomics analysis of serum antibodies. The reason for this is the lack of studies examining the overlap between expressed BCR sequences and secreted antibodies. In addition, the use of single-cell sequencing for the characterization of BCRs remains cost-prohibitive and low-throughput. Furthermore, while there have been studies utilizing mass spectrometry as a tool to deconvolve the serum antibody repertoire, so far there exists no consensus regarding the experimental details such as the exclusion of constant regions, the choice of enzyme for antibody digestion, and limit of detection for antibodies in complex mixtures such as the serum.



(See figure on the previous page)

Figure 37: Summary of thesis results. *Green text and arrows denote completed goals; brown text and arrows denote tasks planned for future projects; yellow text and arrows denote where further work is needed. For detailed presentation of results, section 4.1 concerned bulk B-cell receptor sequencing, section 4.2 concerned single-cell B-cell receptor sequencing, and section 4.3 contained results from the benchmarking experiment for antibody mass spectrometry.*

Of the stated goals previously mentioned in the thesis project's aims, the majority of the tasks have been completed (Figure 37). Specifically, we successfully adapted a reliable protocol for bulk BCR sequencing and expanded the scope of the workflow to include all isotypes in a single sequencing batch. The workflow proved to be reproducible and provided abundant information for immune repertoire analyses. With regards to single-cell BCR sequencing, we accomplished to develop a method for B-cell encapsulation that allowed for much higher throughput while maintaining comparable efficiency compared to other currently available methods. Yet, challenges remain in designing and optimizing a suitable library preparation protocol for single-cell BCR sequencing, requiring additional time and efforts that were beyond the scope of this thesis. In addition, due to the needs arising during the research process, we also conducted a benchmarking experiment detailing the technical aspects of antibody LC-MS/MS and derived from it valuable experiences for further studies in this field. Due to the fact that what we have done is very novel and untested method development, not all aspects of the project turned out as expected. However, we learned a lot from it and managed to build strong foundations for future studies in which these methods can be applied to provide useful biological insights regarding the adaptive immune system. The insights gained, difficulties faced, and future considerations for each method are discussed in further details in their respective sections.

5.1 Bulk B-cell receptor sequencing: method adaptation and considerations

At present the field of BCR sequencing is dominated by two main approaches: 5' rapid amplification of cDNA ends (5' RACE) and multiplex PCR (MTPX), each with their own advantages and drawbacks. Modern 5' RACE protocols are combined with a template-switching reaction and are based on the premise that only a single universal primer is necessary to capture all mRNA sequences while the template-switch oligos capture the 3' mRNA end irrespective of

the 3’ RNA sequence and subsequently synthesize cDNAs [162]. This simplifies the primer design step and reduces amplification bias due to differences in binding efficiencies of primers and is ultilized in numerous BCR sequencing workflows [86]. However, there are certain disadvantages to this approach. Firstly, the efficiency of the template-switching reaction is low and batch-dependent, requiring a higher number of PCR cycles to generate sufficient amounts of cDNA for library preparation [88]. Secondly, owing to the fact that 5’ RACE captures the full-length mRNA sequence, the 5’ UTR sequence is also incorporated into the library as well. The length of the UTR sequence varies between V gene families and can negatively affect the recovery of the full V(D)J sequence due to the length constraints in paired-end sequencing [105]. As a consequence, 5’ RACE protocols may exhibit biases against longer sequences in downstream analyses. In contrast, MTPX protocols require a more elaborate design in order to cover every V-gene family and to ensure that the melting profile of all the multiplex primers remains consistent [163]. Compared to 5’ RACE, MTPX protocols display amplification bias towards specific V-gene families [86]. However, with careful design and optimization, this can be minimized to an acceptable degree that does not hinder further analysis, as shown in the work of Wu and colleagues in optimizing a multiplex PCR system for TCRβ sequencing, which employed similar principles [164]. In return for those drawbacks, 5’ MTPX protocols possess significant benefits. First and foremost, by targeting the leader sequence of the V gene sequence on the 5’ end, the library length is significantly reduced and unaffected by variations in 5’ UTR length, allowing for complete coverage of the V(D)J sequence [88]. Furthermore, 5’ MTPX utilizes C gene-specific sequences for cDNA synthesis, circumventing the need for a semi-nested PCR step to enrich the V(D)J sequences prior to library preparation. In our project, we intended to build a reference database for reconstruction of the antibody sequence in combination with proteomics, therefore the full length sequence of the whole V region is important. Thus, for our BCR sequencing workflow in humans where the reference germline sequence is available, MTPX is a better fit after considering all the characteristics. However, in the field of immune receptor sequencing as a whole, so far no standard protocol exists that is widely adopted.

One of the goals we achieved in our project was expanding the bulk BCR sequencing protocol to all the different B-cell isotypes in the blood, including the κ and λ light chains, contrary to most currently available protocols where only the heavy chain was targeted and within the heavy chain, only IgG or IgM were recovered. We adapted a bulk BCR sequencing protocol from Bernat and colleagues [88] and designed our own in-house primers for the capture of IgA, IgD, and IgE mRNA sequences, coupled with extensive work on validation to improve reliability and reproducibility (Supplementary table 1). Additionally, in order to facilitate error correction in the sequencing data, UMIs were employed in the library preparation process. However, during our implementation of UMI-based error correction, several shortcomings of this approach were revealed. The vast majority of UMI groups (> 80%) in our sequencing data contained only one or two reads per UMI sequence, therefore building consensus reads from the data would require the majority of the sequencing data to be discarded (Figure 13). In order to be a viable option, the sequencing depth of each sample would need to be increased by at least 3–5-fold and up to 10-fold, which would either greatly increase the cost or decrease the number of cells surveyed in each sequencing run. This trade-off between repertoire diversity and accuracy was also briefly discussed by Barennes and colleagues in comparing different library preparation methods, supporting the point that including UMI for error correction depends on the specific goals of the study and not applicable under all circumstances [165]. Since the main goal of our sequencing workflow was to capture the diversity of the B-cell repertoire, it would be necessary to perform over-sequencing to a degree that would severely restrict the number of samples being surveyed. As a result, it was decided that UMI-based error correction would be excluded from the analysis pipeline.

Data from bulk BCR sequencing revealed several areas that require further optimizations. Even though the library preparation workflow proved reproducible, IgE libraries still contained non-specific products (Figure 11) and low clonotype count (Figure 14B). A redesign and performance benchmarking of the isotype-specific IgE primer would be necessary in order to improve yield and specificity. Secondly, clonotype count numbers (Figure 14) and CDR3 overlap values (Figure 17, 18) suggested that sequencing depth is an important consideration when

characterizing different isotype libraries. Due to differences in abundance in the blood, the concentration for resulting BCR isotype libraries can vary significantly. Thus, careful consideration of concentration when pooling the libraries prior to sequencing is necessary to avoid undersampling very abundant samples. Finally, by performing various routine analyses on the BCR repertoires, we demonstrated that our bulk BCR sequencing workflow is robust and applicable to multiple applications. Nevertheless, native chain pairing information can not be recovered by bulk BCR sequencing. Thus combining this workflow with single-cell BCR sequencing would prove highly beneficial.

5.2 Single-cell B-cell receptor sequencing: advancements and limitations

We set out to develop a novel protocol for single-cell BCR sequencing due to the limitations of currently available methods. Of the tools available for single-cell isolation, droplet-based microfluidics offers the biggest advantages in terms of throughput and cost per cell [93]. However, droplet encapsulation necessitates the use of beads for sequence capture and cellular barcoding. Beads can vary in characteristics between batches and comprise the largest cost in a single-cell sequencing experiment. Therefore, our approach attempted to circumvent this problem by utilizing bead-free cellular barcoding, inspired by the work of Briggs and colleagues [101]. However, instead of using custom-made apparatuses, we adapted commercially available instruments such as Dolomite's Nadia system and customized its parameters in order to render it more accessible and less labor-intensive.

Prior to library preparation, we examined the encapsulation efficiency of our system using fluorescent microscopy and adjusted our encapsulation parameters accordingly to achieve a higher encapsulation rate, which is usually not available in other single-cell workflows and commercial kits. Specifically, parameters during the encapsulation process such as the flow rate and pressure of each channel within the microfluidics chip, mixing speed, volume and concentration of B cells, barcodes, and other reagents were meticulously evaluated and adjusted in order to improve efficiency. With this approach, from an input of 500000 B cells, we managed to capture the vast majority of cells since only around 20.4% of droplets are occupied with cells,

and within those droplets, 71.7% are single-cell droplets (Figure 25). This is comparable with most currently available methods, albeit at a much higher cell input. For instance, plate-based protocols such as Smart-seq2 can only handle up to several hundred cells per experiment [166] while commercial platforms such as 10X Chromium allow up to 20000 cells as input per sample well, with only around 10000 of which are recovered [167] at a substantially higher cost per cell (10X: ≈1 USD per 8 cells; Nadia systems: ≈1 USD per 500 cells, based on our estimations). With more refining of the parameters and cell input, it is possible to achieve even higher efficiency.

However, a major limitation encountered in our project was the lack of precedent regarding this specific approach in single-cell sequencing. The majority of primers used in our single-cell library preparation workflow were designed and validated in-house. This would require extensive testing and optimization to find a good balance between specificity and yield. More specifically, the delivery of a single droplet barcode sequence into a droplet remains challenging (Figure 26). The main approach to attain single-molecule dilution depends on mathematical modeling and probabilities. The incorporation of droplet barcodes into droplets follows a Poisson distribution, based on several assumptions. Firstly, the average rate of occurrence is known since the number of barcode molecules and the number of cells is determined. Secondly, incorporation of one barcode does not affect another barcode. Finally, since droplets are formed separately, an encapsulation event does not overlap with another at a single time period. With this distribution, it is possible to estimate a barcode-to-cell ratio so that the vast majority will have one droplet barcode encapsulated with one B cell. However, in practice, the lack of visual confirmation in the form of beads hinders the ability to truly assess barcode distribution before analyzing sequencing data. In addition, since our approach utilized isotope-specific primers for reverse transcription of BCR V(D)J sequences, it is only applicable for immune receptor analysis and not for transcriptome analysis. Nevertheless, we believe that if more work is done on optimizing the library preparation, this approach could lead to major cost reduction by taking advantage of the high throughput of droplet encapsulation and at the same time avoid problems and costs associated with the use of beads. Specifically, a systematic approach reevaluating every section of the workflow is warranted (Figure 9), starting from the use of template-switching mechanism

for cDNA synthesis, choice of reverse transcriptase enzyme and operating conditions, to the distribution of molecular barcodes and droplet barcodes with respect to B-cell input quantity would certainly lead to identification and resolution of problems encountered during the library preparation process.

5.3 Antibody LC-MS/MS: benchmarking of antibody proteomics

The vast diversity of the antibody repertoire, both in terms of variety and magnitude, necessitates a method that is both robust and with a high dynamic range in order to capture useful information. Therefore MS, particularly LC-MS/MS has emerged as a valuable tool to study antibodies in great detail. However, so far there exists no large-scale benchmarking studies that could provide a definitive answer to what extent a given antibody at a certain concentration can be reliably identified using bottom-up MS. Previous efforts so far have focused on either top-down or middle-down MS for mAbs identifications [126]. The sensitivity of MS instruments is constantly improving and in order to take full advantage of those benefits, other factors in the MS workflow from start to finish (from sample preparation to MS settings to data analysis) would need to be improved as well. Therefore, we decided to pursue this benchmarking experiment in order to gain valuable information and experience that would become useful in future studies.

Starting from the sample preparation, the first consideration when conducting antibody LC-MS/MS is isolating and purifying antibodies from samples. When it comes to serum antibodies, the most common approach is utilizing immobilized reagents that bind antibodies. Proteins A, G, L, a plant lectin Jacalin, and mannose-binding lectin are used for class-specific isolation. Although these proteins allow enrichment of multiple immunoglobulin isotypes, the current selection does not allow for the capture of all possible antibody isotypes in the serum. In addition, the isotype-specific immobilizing reagents exhibit their affinity to specific subclasses in an isotype, which hampers the ability to encompass the entirety of the serum antibody repertoire, thus reducing the capacity of mass spectrometry techniques for repertoire characterization. The second consideration regarding sample preparation is whether to cleave

the antibodies into F(ab) or F(ab')2, and Fc fragments. In principle, excluding the constant region, which is highly similar in all antibodies, would yield a massive reduction of the volume of data to be processed and improve the sensitivity and ability to identify peptides of interest in the V region [168]. However, the magnitude of impact of antibody fragmentation on the final experimental output has yet to be measured. In our experiment, we compared antibodies samples with GingisKHAN treatment and intact antibodies for peptide detection (Figure 27). In both serum antibodies samples and mAbs samples, the GingisKHAN treatment resulted in the inability to detect any antibody-related peptides. One possible reason for this could be due to the design of the kit itself, which is designed to digest up to 2 mg of antibodies in each purification run. However, since the goal of our experiments was to test the lower limit of detection for antibody MS, the antibody input was in the range of micrograms to as low as 1 ng (6.67 fmol). Hence, the digestion process would not be able to proceed properly and resulted in poor yield. Therefore, we decided that, for our research purpose, partial digestion of antibodies prior to LC-MS/MS would not be utilized.

The benchmarking experiments revealed several important insights regarding the performance of LC-MS/MS in antibody proteomics. First and foremost, only a small fraction of peptides sequenced by MS/MS could be mapped to a known antibody sequence, with most being contaminants and unrelated peptides (Figure 27, 29, 33). Furthermore, of the mapped peptides, an even smaller fraction could be overlapped with the CDR3 region, which is crucial for discrimination between antibodies (Figure 28, 30, 35). Another factor to consider was the MS-derived peptides contained miscleavages in addition to correctly cleaved peptides, evidenced by the small overlap between theoretical digestion and detected peptides (Figure 34). As a result, antibody identification would have to rely on a very small subset of peptides from MS, which could vary depending on the antibodies examined. Secondly, the limit of detection in terms of identifying CDR3-overlapping peptides were greatly affected by the choice of digestion enzymes (Figure 35). Within the mAbs utilized so far in the benchmarking experiment, the lower detection limit was around 1 ng (6.67 fmol), depending on the enzyme and mAb used, which was in line with other published studies on antibody MS [169]. This phenomenon was expected due

to the interaction between antibody sequence and the enzymes, the same antibody would have different detection limits with different enzymes. However, one unexpected factor was that in some mAbs (PGDM1400 and PGT121), CDR3 peptides were not detected on the heavy chain at any concentration, which meant the variety of digestion enzymes was not high enough to cover all possible variations between antibodies. The last important insight gained was the strong relationship between concentration ratios and signal intensity ratios between identical peptides. Not only identical peptides at the same concentration shared similar signal intensity values (Figure 31), but also identical peptides at certain concentration ratios would also possess similar ratios in signal intensity as well (Figure 32, 36). This finding would be helpful in estimating the native concentration of an antibody in complex samples such as human serum.

These findings represented only the first steps toward a more comprehensive benchmarking of antibody LC-MS/MS. Going forward, further experiments will be conducted at a larger scope, with additional mAbs tested and in different combinations. In addition, another digestion strategy will be implemented, for instance Asp-N, which has specific cleavage activity for Aspartate and Glutamate, and has shown potential in uncovering more CDR3 peptides in samples unique from other enzymes (Figure 38). Our latest preliminary experimental data revealed that Asp-N was able to generate CDR3-overlapping peptides on the heavy chain of PGDM1400, which was not previously detected in other enzymatic treatments [unpublished data]. Therefore, expanding the benchmarking experiment further could alleviate some of the problems encountered during our experiments.

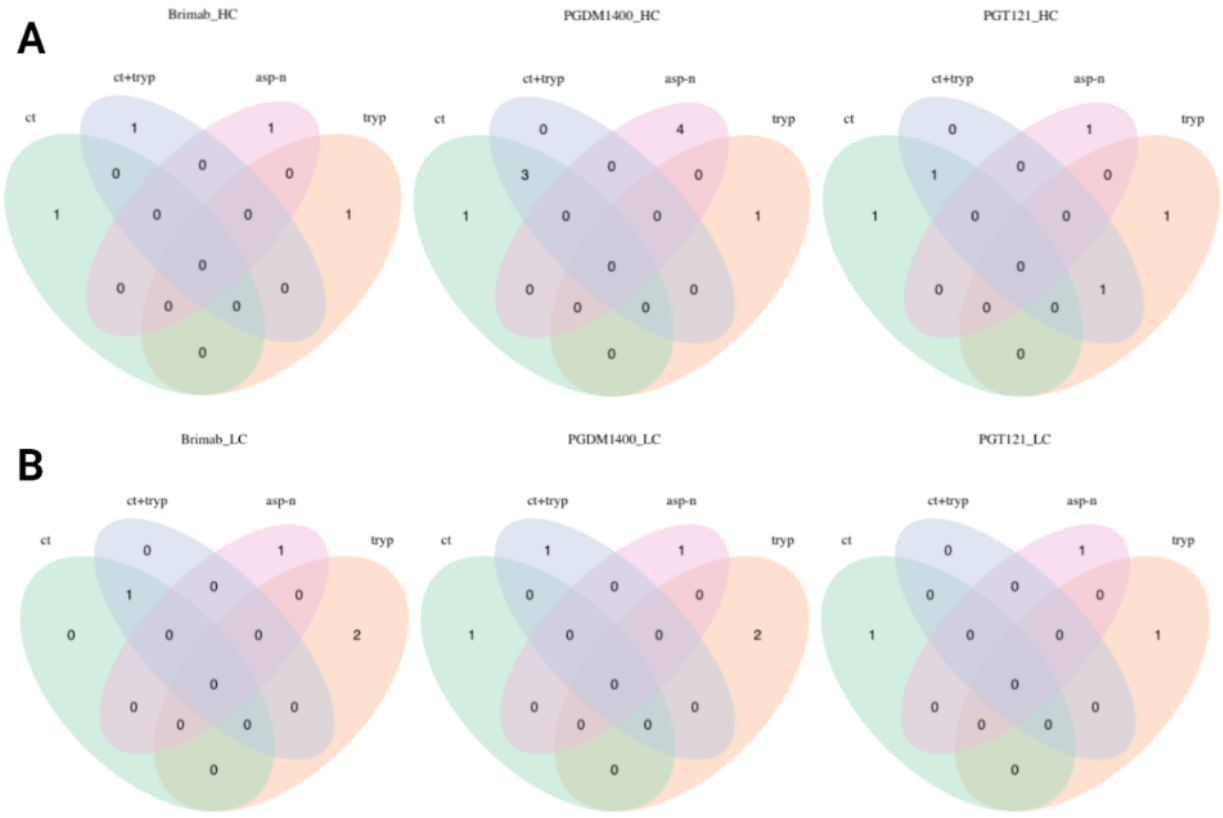


Figure 38: Theoretical cuttings of CDR3-related peptides in Briakinumab, PGDM1400, and PGT121. Asp-N digestion in particular resulted in unique CDR3-related peptides in every mAb used. *Each mAb was digested in-silico by trypsin, chymotrypsin, chymotrypsin + trypsin, and Asp-N.* ***A)*** *Number of CDR3 peptides detected on the heavy chain.* ***B)*** *Number of peptides detected on the light chain.*

6. Outlook and future perspectives

Although we have achieved advancements in bulk BCR sequencing, single-cell BCR sequencing, and antibody LC-MS/MS, further technology and method improvements are required. The single-cell sequencing workflow will need further adjustments and modifications to ensure higher yield and better library quality. The MS benchmarking experiments will be continued to contribute to a more standardized approach to antibody proteomics. Once completed, MS experiments regarding human serum antibodies could be performed in order to investigate the intersection between antibody genome and phenome. In addition, single-cell sequencing data would provide crucial information about native chain pairing in antibodies, helping to reconstruct a complete antibody structure from MS data. By combining all of these workflows together, a comprehensive description of an individual's immune repertoire can be made, which would prove invaluable to multiple applications all across the field of immunology.

Appendix

Supplementary table 1. Overview of the cDNA synthesis primers. Letters in bold indicated the UMI sequence, while letters underlined indicated overlap with Illumina Read2 sequence.

Primer name	Sequence	Ref
3' primers		
Hu_IgG	GGAGTTCAGACGTGTGCTCTTCCGATCT**HHHHHHACAHHHHHHACAHHHH**GCCAG GGGGAAGACCGATGGG	[88]
Hu_IgM	GGAGTTCAGACGTGTGCTCTTCCGATCT**HHHHHHACAHHHHHHACAHHHH**NHCCG ACGGGGAATTCTCACAGGAGACGAGGGGGAAAAG	[88]
Hu_IgA	GGAGTTCAGACGTGTGCTCTTCCGATCT**HHHHHHACAHHHHHHACAHHHH**GAAGA CCTTGGGGCTGGT	In-house
Hu_IgD	GGAGTTCAGACGTGTGCTCTTCCGATCT**HHHHHHACAHHHHHHACAHHHH**GGGTGT CTGCACCCTGATA	In-house
Hu_IgE	GGAGTTCAGACGTGTGCTCTTCCGATCT**HHHHHHACAHHHHHHACAHHHH**GAAGA CGGATGGGCTCTGT	In-house
Hu_IgK	GGAGTTCAGACGTGTGCTCTTCCGATCT**HHHHHHACAHHHHHHACAHHHH**NGGGAT AGAAGTTATTCAGCAGGCACACAACAGAG	[88]
Hu_IgL	GGAGTTCAGACGTGTGCTCTTCCGATCT**HHHHHHACAHHHHHHACAHHHH**TGGCTT GRAGCTCCTCAGAGGAGG	[88]

Supplementary table 2. Overview of the Multiplex PCR primers.

Primer name	Sequence	Ref
3' primers		
Read2U	GGAGTTCAGACGTGTGCTCTTCCGATCT	[88]
5' primers		
Hu_VH_MTPX_1	CCTACACGACGCTCTTCCGATCTGGTGGCAGCAGTCACAGATGCCTACTC	[88]
Hu_VH_MTPX_2	CCTACACGACGCTCTTCCGATCTGGTGGCAGCAGCCACAGGTGCCCACTC	[88]
Hu_VH_MTPX_3	CCTACACGACGCTCTTCCGATCTGGTGGCAGCAGCTACAGGTGTCCAGTC	[88]
Hu_VH_MTPX_4	CCTACACGACGCTCTTCCGATCTGGTGGSAGCAGCAACARGWGCCCACTC	[88]

Hu_VH_MTPX_5	CCTACACGACGCTCTTCCGATCTGCTGGCTGTAGCTCCAGGTGCTCACTC	[88]
Hu_VH_MTPX_6	CCTACACGACGCTCTTCCGATCTCCTGCTGCTGACCAYCCCTTCMTGGGTCTTGTC	[88]
Hu_VH_MTPX_7	CCTACACGACGCTCTTCCGATCTCCTGCTACTGACTGTCCCGTCCTGGGTCTTATC	[88]
Hu_VH_MTPX_8	CCTACACGACGCTCTTCCGATCTGGGTTTTCCTCGTTGCTCTTTTAAGAGGTGTCCAGTG	[88]
Hu_VH_MTPX_9	CCTACACGACGCTCTTCCGATCTGGGTTTTCCTTGTTGCTATTTTAAAAGGTGTCCARTG	[88]
Hu_VH_MTPX_10	CCTACACGACGCTCTTCCGATCTGGATTTTCCTTGCTGCTATTTTAAAAGGTGTCCAGTG	[88]
Hu_VH_MTPX_11	CCTACACGACGCTCTTCCGATCTGGGTTTTCCTTKTKGCTATWTTAGAAGGTGTCCAGTG	[88]
Hu_VH_MTPX_12	CCTACACGACGCTCTTCCGATCTGGTGGCRGCTCCCAGATGGGTCCTGTC	[88]
Hu_VH_MTPX_13	CCTACACGACGCTCTTCCGATCTCTGGCTGTTCTCCAAGGAGTCTGTG	[88]
Hu_VH_MTPX_14	CCTACACGACGCTCTTCCGATCTGGCCTCCCATGGGGTGTCCTGTC	[88]
Hu_VH_MTPX_15	CCTACACGACGCTCTTCCGATCTGGTGGCAGCAGCAACAGGTGCCCACT	[88]
Hu_VH_MTPX_16	CACTCTTTCCCTACACGACGCTCTTCCGATCTATGGAACTGGGGCTCCGCTGGGTTTTCC	[88]
Hu_VH_MTPX_17	CACTCTTTCCCTACACGACGCTCTTCCGATCTATGGACTGCACCTGGAGGATCCTCCTC	[88]
Hu_VH_MTPX_18	CACTCTTTCCCTACACGACGCTCTTCCGATCTTGGCTGAGCTGGGTTTYCCTTGTTGC	[88]
Hu_VH_MTPX_19	CACTCTTTCCCTACACGACGCTCTTCCGATCTGGAGTTKGGGCTGMGCTGGGTTTTCC	[88]
Hu_VH_MTPX_20	CACTCTTTCCCTACACGACGCTCTTCCGATCTGCACCTGTGGTTTTTCCTCCTGCTGGTG	[88]
Hu_VH_MTPX_21	CACTCTTTCCCTACACGACGCTCTTCCGATCTCACCTGTGGTTCTTCCTCCTSCTGG	[88]
Hu_VH_MTPX_22	CACTCTTTCCCTACACGACGCTCTTCCGATCTCCAGGATGGGGTCAACCGCCATCCTC	[88]
Hu_VH_MTPX_23	CTCTTTCCCTACACGACGCTCTTCCGATCTCAGAGGACTCACCATGGAGTTTGGGCTGAG	[88]
Hu_VH_MTPX_24	CCTACACGACGCTCTTCCGATCTGGACTCACCATGGAGTTGGGACTGAGC	[88]
Hu_VH_MTPX_25	CCTACACGACGCTCTTCCGATCTGGGCTGAGCTGGCTTTTTCTTGTGGC	[88]
Hu_VK_MTPX_1	CTACACTCTTTCCCTACACGACGCTCTTCCGATCTATGTTGCCATCACAACTCATTGGGTTTCTG	[88]

Hu_VK_MTPX_2	CTACACTCTTTCCCTACACGACGCTCTTCCGATCTATGGAARCCCCAGCGCAG CTTCTCTTCC	[88]
Hu_VK_MTPX_3	CTACACTCTTTCCCTACACGACGCTCTTCCGATCTATGAGGCTCCCTGCTCAGC TCTTGGGGCT	[88]
Hu_VK_MTPX_4	CTACACTCTTTCCCTACACGACGCTCTTCCGATCTATGAGGCTCCCTGCTCAGC TCCTGGGGCT	[88]
Hu_VK_MTPX_5	CTACACTCTTTCCCTACACGACGCTCTTCCGATCTATGGACATGAGGGTCCCTG CTCAGC	[88]
Hu_VK_MTPX_6	CTACACTCTTTCCCTACACGACGCTCTTCCGATCTATGGACATGAGRGTCCTCG CTCAGC	[88]
Hu_VK_MTPX_7	CTACACTCTTTCCCTACACGACGCTCTTCCGATCTATGGAAGCCCCAGCACAG CTTCTTCTTCC	[88]
Hu_VK_MTPX_8	CTACACTCTTTCCCTACACGACGCTCTTCCGATCTATGAGGCTCCTTGCTCAGC TTCTGGGGCT	[88]
Hu_VK_MTPX_9	CTACACTCTTTCCCTACACGACGCTCTTCCGATCTATGGAAGCCCCAGCTCAGC TTCTCTTCC	[88]
Hu_VK_MTPX_10	CTACACTCTTTCCCTACACGACGCTCTTCCGATCTATGGACATGAGGGTCCCCG CTCAGC	[88]
Hu_VK_MTPX_11	CTACACTCTTTCCCTACACGACGCTCTTCCGATCTATGGGGTCCCAGGTTCACC TCCTCAG	[88]
Hu_VK_MTPX_12	CTACACTCTTTCCCTACACGACGCTCTTCCGATCTATGGTGTTGCAGACCCAGG TCTTCATTTC	[88]
Hu_VK_MTPX_13	CTACACTCTTTCCCTACACGACGCTCTTCCGATCTATGGACATGAGGGTGCCCG CTCAGC	[88]
Hu_VK_MTPX_14	CTCTTTCCCTACACGACGCTCTTCCGATCTCAGGAAGATGTYGCCATCACAAC TCATTGG	[88]
Hu_VK_MTPX_15	CACTCTTTCCCTACACGACGCTCTTCCGATCTCTCRCAATGAGGCTCCCTGCTC AGCTC	[88]
Hu_VK_MTPX_16	CACTCTTTCCCTACACGACGCTCTTCCGATCTCCTGCTCAGCTCYTGGGGCTG CTAATGC	[88]
Hu_VK_MTPX_17	CACTCTTTCCCTACACGACGCTCTTCCGATCTATGGACATGAGGGTGCCCGCTC AGCGCC	[88]
Hu_VK_MTPX_18	CACTCTTTCCCTACACGACGCTCTTCCGATCTATGGACATGAGGGTSCCYGCTC AGCKCC	[88]
Hu_VK_MTPX_19	CACTCTTTCCCTACACGACGCTCTTCCGATCTGCTCCTGGGGCTGCTAATGCTC TGG	[88]

Hu_VK_MTPX_20	CACTCTTTCCCTACACGACGCTCTTCCGATCTGGGGCTCCTGCTGCTCTGGCTCC	[88]
Hu_VK_MTPX_21	CACTCTTTCCCTACACGACGCTCTTCCGATCTGGACATGAGGGTCCCCGCTCAGCTCC	[88]
Hu_VL_MTPX_1	CTACACTCTTTCCCTACACGACGCTCTTCCGATCTATGGCCTGGGCTCCACTACTTCTCACCCTCC	[88]
Hu_VL_MTPX_2	CTACACTCTTTCCCTACACGACGCTCTTCCGATCTATGGCCTGGTCCCCTCTCTTCCTCACCCT	[88]
Hu_VL_MTPX_3	CTACACTCTTTCCCTACACGACGCTCTTCCGATCTATGGCCTGGGCTCTGCTCCTCCTCACCCT	[88]
Hu_VL_MTPX_4	CTACACTCTTTCCCTACACGACGCTCTTCCGATCTATGGCCTGGAYCCCTCTCCTGCTCCCCCTC	[88]
Hu_VL_MTPX_5	CTACACTCTTTCCCTACACGACGCTCTTCCGATCTATGGCCTGGGCTCTGCTGCTCCTCACTCT	[88]
Hu_VL_MTPX_6	CTACACTCTTTCCCTACACGACGCTCTTCCGATCTATGGCATGGATCCCTCTCTTCCTCGGCGTC	[88]
Hu_VL_MTPX_7	CTACACTCTTTCCCTACACGACGCTCTTCCGATCTATGGCATGGGCCACACTCCTGCTCCCACTC	[88]
Hu_VL_MTPX_8	CTACACTCTTTCCCTACACGACGCTCTTCCGATCTATGGCCTGGGTCTCCTTCTACCTACTGCCCT	[88]
Hu_VL_MTPX_9	CTACACTCTTTCCCTACACGACGCTCTTCCGATCTATGGCCTGGACTCCTCTTCTTCTCTTGCTCCT	[88]
Hu_VL_MTPX_10	CTACACTCTTTCCCTACACGACGCTCTTCCGATCTATGGCCTGGACTCCTCTCCTCCTCCTGYTCC	[88]
Hu_VL_MTPX_11	CTACACTCTTTCCCTACACGACGCTCTTCCGATCTATGAGTGTCCCCACCATGGCCTGGATGATGC	[88]
Hu_VL_MTPX_12	CTACACTCTTTCCCTACACGACGCTCTTCCGATCTATGGCCTGGGCTCCTCTGCTCCTCACCCTCC	[88]
Hu_VL_MTPX_13	CTACACTCTTTCCCTACACGACGCTCTTCCGATCTATGRCCDGCTTCCCTCTCCTCCTCACCCT	[88]
Hu_VL_MTPX_14	CTACACTCTTTCCCTACACGACGCTCTTCCGATCTATGGCCTGGACCCCACTCCTCCTCCTCTTCC	[88]
Hu_VL_MTPX_15	CTACACTCTTTCCCTACACGACGCTCTTCCGATCTATGGCCTGGGCTCTGCTSCTCCTCASCCT	[88]
Hu_VL_MTPX_16	CTACACTCTTTCCCTACACGACGCTCTTCCGATCTATGGCCTGGATCCCTCTACTTCTCCCCCTC	[88]

Hu_VL_MTPX_17	CTACACTCTTTCCCTACACGACGCTCTTCCGATCTATGGCCTGGACCSCTCTCC TCCTCRGCCTC	[88]
Hu_VL_MTPX_18	CTACACTCTTTCCCTACACGACGCTCTTCCGATCTATGGCCTGGACTCTTCTCC TTCTCGTGCTCC	[88]
Hu_VL_MTPX_19	CTACACTCTTTCCCTACACGACGCTCTTCCGATCTATGGCCTGGTCTCCTCTCC TCCTCACTCT	[88]
Hu_VL_MTPX_20	CTACACTCTTTCCCTACACGACGCTCTTCCGATCTATGCCCTGGGCTCTGCTCC TCCTGACCCT	[88]
Hu_VL_MTPX_21	CTACACTCTTTCCCTACACGACGCTCTTCCGATCTATGGCCTGGACCCCTCTCT GGCTCACTCTC	[88]
Hu_VL_MTPX_22	CTACACTCTTTCCCTACACGACGCTCTTCCGATCTATGGCCTGGACCGCTCTCC TTCTGAGCCTC	[88]
Hu_VL_MTPX_23	CTACACTCTTTCCCTACACGACGCTCTTCCGATCTATGGCTTGGACCCCACTCC TCTTCCTCACC	[88]
Hu_VL_MTPX_24	CTACACTCTTTCCCTACACGACGCTCTTCCGATCTATGGCCTGGACTCCTCTCT TTCTGTTCCTCC	[88]
Hu_VL_MTPX_25	CACTCTTTCCCTACACGACGCTCTTCCGATCTATGGCCTGGACTCTTCTCCTTC TCGTG	[88]
Hu_VL_MTPX_26	CACTCTTTCCCTACACGACGCTCTTCCGATCTATGGCCTGGACTCCTCTYCTYC TCYTG	[88]
Hu_VL_MTPX_27	CACTCTTTCCCTACACGACGCTCTTCCGATCTATGGCCTGGACCCCACTCCTCC TC	[88]
Hu_VL_MTPX_28	CACTCTTTCCCTACACGACGCTCTTCCGATCTATGGCCTGGGTCTCCTTCTACC TACTGC	[88]
Hu_VL_MTPX_29	CACTCTTTCCCTACACGACGCTCTTCCGATCTGCAGCATCGGAGGTGCCTCAG CCATG	[88]
Hu_VL_MTPX_30	CACTCTTTCCCTACACGACGCTCTTCCGATCTGGCAGAACTCTGGGTGTCTCA CCATG	[88]
Hu_VL_MTPX_31	CACTCTTTCCCTACACGACGCTCTTCCGATCTGCAGCACTGGTGGTGCCTCAG CCATG	[88]
Hu_VL_MTPX_32	CACTCTTTCCCTACACGACGCTCTTCCGATCTGGGCTCTGCTSCTCCTCACYCT CCT	[88]
Hu_VL_MTPX_33	CACTCTTTCCCTACACGACGCTCTTCCGATCTGGGCTCTGCTCCTCCTGACCCT C	[88]

Supplementary table 3. Overview of the adapter extension PCR primers. Letters in bold indicated the index sequence, while letters underlined indicated the Illumina P5/P7 adapter sequence.

Primer name	Sequence	Ref
5' primers		
P5_R1	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT	[88]
3' primers		
P7_R2_I1	CAAGCAGAAGACGGCATACGAGAT**CGTGAT**GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT	[88]
P7_R2_I2	CAAGCAGAAGACGGCATACGAGAT**ACATCG**GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT	[88]
P7_R2_I3	CAAGCAGAAGACGGCATACGAGAT**GCCTAA**GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT	[88]
P7_R2_I4	CAAGCAGAAGACGGCATACGAGAT**TGGTCT**GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT	[88]
P7_R2_I5	CAAGCAGAAGACGGCATACGAGAT**CACTGT**GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT	[88]
P7_R2_I6	CAAGCAGAAGACGGCATACGAGAT**ATTGGC**GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT	[88]
P7_R2_I7	CAAGCAGAAGACGGCATACGAGAT**GATCTG**GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT	[88]
P7_R2_I8	CAAGCAGAAGACGGCATACGAGAT**TCAAGT**GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT	[88]
P7_R2_I9	CAAGCAGAAGACGGCATACGAGAT**CTGATC**GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT	[88]
P7_R2_I10	CAAGCAGAAGACGGCATACGAGAT**AAGCTA**GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT	[88]
P7_R2_I11	CAAGCAGAAGACGGCATACGAGAT**GTAGCC**GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT	[88]
P7_R2_I12	CAAGCAGAAGACGGCATACGAGAT**TACAAG**GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT	[88]

Supplementary table 4. Components of the Cell lysis buffer and PCR reagents.

Component	Volume (µL)	Component	Volume (µL)	Component	Volume (µL)
Cell lysis buffer		**Molecular and droplet barcodes**		**PCR reagents**	
Nuclease-free H_2O	110	LNA_MBC_UA_TS	1.2	IgM-biotin primer	1.2
20 % w/v Ficoll PM-400	82.5	DBC_pP7_UA	0.1	IgG-biotin primer	1.2
20 % v/v Sarkosyl	2.75	DBC_FP	1.2	Herculase II PCR polymerase	2.5
0.5 M EDTA	11	DBC_RP	1.2	MuMLV reverse transcriptase	6
1 M Tris pH 7.5	55			dNTP	22
1 M DTT	13.75			5x RT buffer	50
				RiboLock RNase inhibitor	6

Supplementary table 5. Overview of the single-cell library preparation primers. Letters in bold indicated the index/UMI sequence, while letters underlined indicated the Illumina P5/P7 adapter sequence.

Primer name	Sequence	Ref
IgG-biotin	AGGACAGCCGGGAAGGTGT-biotin	[101]
IgM-biotin	TGTGAGGTGGCTGCGTACTTG-biotin	[101]
LNA_MBC_UA_TS	AAGCAGTGGTATCAACGCAGAGT**NNNNNN**TCTT{GGG}	In-house
DBC_pP7_UA	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT**NNNNNNNNNNNN**AAG CAGTGGTATCAACGCAGAGT	In-house
DBC_FP	GTGACTGGAGTTCAGACGTGTG	In-house
DBC_RP	ACTCTGCGTTGATACCACTGC	In-house
IgG_nest_fullP5	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTT CCGATCTCCAGGGGGAAGACSGATG	[101]
IgM_nest_fullP5	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTT CCGATCTGAGACGAGGTGGAAAAGGGTTG	[101]
fP7-I12	CAAGCAGAAGACGGCATACGAGAT**TACAAG**GTGACTGGAGTTCAGACGT GTGCTCTTCCGATCT	In-house

Supplementary table 6. Parameters utilized in MaxQuant.

Parameter	Value	Parameter	Value	Parameter	Value
Include contaminants	TRUE	Advanced site intensities	TRUE	Da interval. (ITMS)	100
PSM FDR	0.01	Write msScans table	FALSE	MS/MS deisotoping (ITMS)	FALSE
PSM FDR Crosslink	0.01	Write msmsScans table	TRUE	MS/MS deisotoping tolerance (ITMS)	0.15
Protein FDR	0.01	Write ms3Scans table	TRUE	MS/MS deisotoping tolerance unit (ITMS)	Da
Site FDR	0.01	Write allPeptides table	TRUE	MS/MS higher charges (ITMS)	TRUE
Use Normalized Ratios For Occupancy	TRUE	Write mzRange table	TRUE	MS/MS water loss (ITMS)	TRUE
Min. peptide Length	7	Write DIA fragments table	FALSE	MS/MS ammonia loss (ITMS)	TRUE
Min. score for unmodified peptides	0	Write pasef Msms Scans table	TRUE	MS/MS dependent losses (ITMS)	TRUE
Min. score for modified peptides	40	Write accumulated Pasef Msms Scans table	FALSE	MS/MS recalibration (ITMS)	FALSE
Min. delta score for unmodified peptides	0	Max. peptide mass [Da]	4600	MS/MS tol. (TOF)	40 ppm
Min. delta score for modified peptides	6	Min. peptide length for unspecific search	8	Top MS/MS peaks per Da interval. (TOF)	10
Min. unique peptides	1	Max. peptide length for unspecific search	25	Da interval. (TOF)	100
Min. razor peptides	1	Razor protein FDR	TRUE	MS/MS deisotoping (TOF)	TRUE
Min. peptides	1	Disable MD5	FALSE	MS/MS deisotoping tolerance (TOF)	0.01
Use only unmodified peptides and	TRUE	Max mods in site table	3	MS/MS deisotoping tolerance unit (TOF)	Da
Modifications included in protein quantification	Oxidation (M);Acetyl (Protein N-term)	Match unidentified features	FALSE	MS/MS higher charges (TOF)	TRUE
Peptides used for protein quantification	Razor	Epsilon score for mutations	None	MS/MS water loss (TOF)	TRUE
Discard unmodified counterpart peptides	TRUE	Evaluate variant peptides separately	TRUE	MS/MS ammonia loss (TOF)	TRUE

Label min. ratio count	2	Variation mode	None	MS/MS dependent losses (TOF)	TRUE
Use delta score	FALSE	MS/MS tol. (FTMS)	20 ppm	MS/MS recalibration (TOF)	FALSE
iBAQ	FALSE	Top MS/MS peaks per Da interval. (FTMS)	12	MS/MS tol. (Unknown)	20 ppm
iBAQ log fit	FALSE	Da interval. (FTMS)	100	Top MS/MS peaks per Da interval. (Unknown)	12
Match between runs	FALSE	MS/MS deisotoping (FTMS)	TRUE	Da interval. (Unknown)	100
Find dependent peptides	FALSE	MS/MS deisotoping tolerance (FTMS)	7	MS/MS deisotoping (Unknown)	TRUE
Decoy mode	revert	MS/MS deisotoping tolerance unit (FTMS)	ppm	MS/MS deisotoping tolerance (Unknown)	7
Include contaminants	TRUE	MS/MS higher charges (FTMS)	TRUE	MS/MS deisotoping tolerance unit (Unknown)	ppm
Advanced ratios	TRUE	MS/MS water loss (FTMS)	TRUE	MS/MS higher charges (Unknown)	TRUE
Second peptides	TRUE	MS/MS ammonia loss (FTMS)	TRUE	MS/MS water loss (Unknown)	TRUE
Stabilize large LFQ ratios	TRUE	MS/MS dependent losses (FTMS)	TRUE	MS/MS ammonia loss (Unknown)	TRUE
Separate LFQ in parameter groups	FALSE	MS/MS recalibration (FTMS)	FALSE	MS/MS dependent losses (Unknown)	TRUE
Require MS/MS for LFQ comparisons	TRUE	MS/MS tol. (ITMS)	0.5 Da	MS/MS recalibration (Unknown)	FALSE
Calculate peak properties	FALSE	Top MS/MS peaks per Da interval. (ITMS)	8	Main search max. combinations	200

Supplementary table 7. Monoclonal antibodies used in peptide identification

Antibody	Sequence (V-region only)
PGDM1400_HC	QAQLVQSGPEVRKPGTSVKVSCKAPGNTLKTYDLHWVRSVPGQGLQWMGWISHEGDKKVI VERFKAKVTIDWDRSTNTAYLQLSGLTSGDTAVYYCAKGSKHRLRDYALYDDDGALNWAVD VDYLSNLEFWGQGTAVTVSS
PGDM1400_LC	DFVLTQSPHSLSVTPGESASISCKSSHSLIHGDRNNYLAWYVQKPGRSPQLLIYLASSRASGVP DRFSGSGSDKDFTLKISRVETEDVGTYYCMQGRESPWTFFGQGTKVDIK
PGT121_HC	QMQLQESGPGLVKPSETLSLTCSVSGASISDSYWSWIRRSPGKGLEWIGYVHKSGDTNYSPSL KSRVNLSLDTSKNQVSLSLVAATAADSGKYYCARTLHGRRIYGIVAFNEWFTYFYMDVWGN GTQVTVSS
PGT121_LC	SDISVAPGETARISCGEKSLGSRAVQWYQHRAGQAPSLIIYNNQDRPSGIPERFSGSPDSPFGTT ATLTITSVEAGDEADYYCHIWDSRVPTKWVFGGGTTLTVL
Briakinumab_HC	QVQLVESGGGVVQPGRSLRLSCAASGFTFSSYGMHWVRQAPGKGLEWVAFIRYDGSNKYYA DSVKGRFTISRDNSKNTLYLQMNSLRAEDTAVYYCKTHGSHDNWGQGTMVTVSS
Briakiumab_LC	QSVLTQPPSVSGAPGQRVTISCSGSRSNIGSNTVKWYQQLPGTAPKLLIYYNDQRPSGVPDRFS GSKSGTSASLAITGLQAEDEADYYCQSYDRYTHPALLFGTGTKVTVL

References

1. Flajnik MF, Kasahara M. Origin and evolution of the adaptive immune system: genetic events and selective pressures. Nature Reviews Genetics. 2010 Jan;11(1):47–59.
2. Lodish H. Molecular cell biology. 8th ed. New York: Freeman; 2016.
3. Hoebe K, Janssen E, Beutler B. The interface between innate and adaptive immunity. Nature Immunology. 2004 Oct;5(10):971–4.
4. Abbas AK, Lichtman AH, Pillai S, Baker DL, Baker A. Cellular and molecular immunology. Ninth edition. Philadelphia, PA: Elsevier; 2018. 565 p.
5. Reddick LE, Alto NM. Bacteria Fighting Back: How Pathogens Target and Subvert the Host Innate Immune System. Molecular Cell. 2014 Apr;54(2):321–8.
6. S. J. Martin, Dennis R. Burton, Peter J. Delves, Ivan M. Roitt. Roitt's essential immunology. 13th ed. Chichester: Wiley-Blackwell; 2017. xv+556. (Essentials).
7. Burnet FM. The Clonal selection theory of acquired immunity. Cambridge; 1959. 208 p. (The Abraham Flexner lectures of Vanderbilt University 1958).
8. Zinkernagel RM, Bachmann MF, Kündig TM, Oehen S, Pirchet H, Hengartner H. On Immunological Memory. Annual Review of Immunology. 1996;14(1):333–67.
9. David K. . Male. Immunology. 8th ed. S.l.: Elsevier Saunders; 2012. x+472.
10. Bonilla FA, Oettgen HC. Adaptive immunity. Journal of Allergy and Clinical Immunology. 2010 Feb;125(2):S33–40.
11. Chaplin DD. Overview of the immune response. Journal of Allergy and Clinical Immunology. 2010 Feb;125(2):S3–23.
12. Wieczorek M, Abualrous ET, Sticht J, Álvaro-Benito M, Stolzenberg S, Noé F, et al. Major Histocompatibility Complex (MHC) Class I and MHC Class II Proteins: Conformational Plasticity in Antigen Presentation. Frontiers in immunology. 2017;8:292-.
13. Huppa JB, Davis MM. T-cell-antigen recognition and the immunological synapse. Nature Reviews Immunology. 2003 Dec;3(12):973–83.
14. Huston DP. The Biology of the Immune System. JAMA. 1997 Dec 10;278(22):1804–14.
15. Halle S, Halle O, Förster R. Mechanisms and Dynamics of T Cell-Mediated Cytotoxicity In Vivo. Trends in Immunology. 2017;38(6):432–43.
16. Roche PA, Furuta K. The ins and outs of MHC class II-mediated antigen processing and presentation. Nature reviews Immunology. 2015;15(4):203–16.
17. Shannon J. Turley, Kayo Inaba, Wendy S. Garrett, Melanie Ebersold, Julia Unternaehrer, Ralph M. Steinman, et al. Transport of Peptide: MHC Class II Complexes in Developing Dendritic Cells. Science. 2000;288(5465):522–7.
18. Cosmi L, Maggi L, Santarlasci V, Liotta F, Annunziato F. T helper cells plasticity in inflammation. Cytometry Part A. 2014;85(1):36–42.
19. Varricchi G, Harker J, Borriello F, Marone G, Durham SR, Shamji MH. T follicular helper (Tfh) cells in normal immune responses and in allergic disorders. Allergy. 2016;71(8):1086–94.
20. Vale AM, Kearney JF, Nobrega A, Schroeder HW. Chapter 7 - Development and Function of B Cell Subsets. In: Molecular Biology of B Cells. Second Edition. Elsevier Ltd; 2015. p. 99–119.
21. BioRender [Internet]. BioRender - Create Professional Science Figures in Minutes. [cited 2020 Jul 17]. Available from: https://biorender.com/

22. Johnston CM, Wood AL, Bolland DJ, Corcoran AE. Complete Sequence Assembly and Characterization of the C57BL/6 Mouse Ig Heavy Chain V Region. The Journal of immunology (1950). 2006;176(7):4221–34.
23. Wardemann H, Busse CE. Novel Approaches to Analyze Immunoglobulin Repertoires. Trends in immunology. 2017;38(7):471–82.
24. Bassing CH, Swat W, Alt FW. The Mechanism and Regulation of Chromosomal V(D)J Recombination. Cell. 2002 Apr 19;109(2):S45–55.
25. Schatz DG, Ji Y. Recombination centres and the orchestration of V(D)J recombination. Nature reviews Immunology. 2011;11(4):251–63.
26. Little AJ, Matthews A, Oettinger M, Roth DB, Schatz DG. Chapter 2 - The Mechanism of V(D)J Recombination. In: Molecular Biology of B Cells. Second Edition. Elsevier Ltd; 2015. p. 13–34.
27. Jung D, Giallourakis C, Mostoslavsky R, Alt FW. MECHANISM AND CONTROL OF V(D)J RECOMBINATION AT THE IMMUNOGLOBULIN HEAVY CHAIN LOCUS. Annual Review of Immunology. 2006 Apr;24(1):541–70.
28. Melchers F. Checkpoints that control B cell development. The Journal of clinical investigation. 2015;125(6):2203–10.
29. Herzog S, Reth M, Jumaa H. Regulation of B-cell proliferation and differentiation by pre-B-cell receptor signalling. Nature Reviews Immunology. 2009 Mar;9(3):195–205.
30. Haraldsson Á, Kock-Jansen MJH, Jaminon M, v Eck-Arts PBJM, de Boo T, Weemaes CMR, et al. Determination of Kappa and Lambda Light Chains in Serum Immunoglobulins G, A and M. Ann Clin Biochem. 1991 Sep 1;28(5):461–6.
31. Ottens K, Hinman RM, Barrios E, Skaug B, Davis LS, Li Q-Z, et al. Foxo3 Promotes Apoptosis of B Cell Receptor-Stimulated Immature B Cells, Thus Limiting the Window for Receptor Editing. The Journal of immunology (1950). 2018;201(3):940–9.
32. Luning Prak ET, Monestier M, Eisenberg RA. B cell receptor editing in tolerance and autoimmunity. Annals of the New York Academy of Sciences. 2011;1217(1):96–121.
33. Hauser AE, Höpken UE. Chapter 12 - B Cell Localization and Migration in Health and Disease. In: Molecular Biology of B Cells. Second Edition. Elsevier Ltd; 2015. p. 187–214.
34. Parham P. The immune system. 4th ed. New York: Garland Science; 2015. xxi+532.
35. Williams AF, Barclay AN. The Immunoglobulin Superfamily-Domains for Cell Surface Recognition. Annual review of immunology. 1988;6(1):381–405.
36. Schroeder HW, Cavacini L. Structure and function of immunoglobulins. Journal of allergy and clinical immunology. 2010;125(2):S41–52.
37. Lu LL, Suscovich TJ, Fortune SM, Alter G. Beyond binding: antibody effector functions in infectious diseases. Nature reviews Immunology. 2017;18(1):46–61.
38. Sela-Culang I, Kunik V, Ofran Y. The Structural Basis of Antibody-Antigen Recognition. Frontiers in immunology. 2013;4:302-.
39. Chen D, Zhang Z, Yang Y, Hong Q, Li W, Zhuo L. High-throughput sequencing analysis of genes encoding the B-lymphocyte receptor heavy-chain CDR3 in renal and peripheral blood of IgA nephropathy. Bioscience reports [Internet]. 2019 [cited 2020 Jul 27];39(10). Available from: http://dx.doi.org/10.1042/BSR20190482
40. Wang L, Dai Y, Liu S, Lai L, Yan Q, Chen H, et al. Assessment of variation in B-cell receptor heavy chain repertoire in patients with end-stage renal disease by high-throughput sequencing. Renal failure. 2019;41(1):1–13.

41. Kettleborough CA, Saldanha J, Heath VJ, Morrison CJ, Bendig MM. Humanization of a mouse monoclonal antibody by CDR–grafting: the importance of framework residues on loop conformation. Protein engineering, design and selection. 1991;4(7):773–83.
42. Arnold JN, Wormald MR, Sim RB, Rudd PM, Dwek RA. The Impact of Glycosylation on the Biological Function and Structure of Human Immunoglobulins. Annual review of immunology. 2007;25(1):21–50.
43. Bednarski JJ, Sleckman BP. At the intersection of DNA damage and immune responses. Nature reviews Immunology. 2019;19(4):231–42.
44. Kenter AL. Class Switch Recombination: An Emerging Mechanism. In: Singh H, Grosschedl R, editors. Molecular Analysis of B Lymphocyte Development and Activation [Internet]. Berlin, Heidelberg: Springer; 2005 [cited 2020 Jul 15]. p. 171–99. (Current Topics in Microbiology and Immunology). Available from: https://doi.org/10.1007/3-540-26363-2_8
45. Moens L, Tangye SG. Cytokine-Mediated Regulation of Plasma Cell Generation: IL-21 Takes Center Stage. Frontiers in immunology. 2014;5:65-.
46. Franklin A, Blanden RV. On the molecular mechanism of somatic hypermutation of rearranged immunoglobulin genes. Immunology and Cell Biology. 2004;82(6):557–67.
47. Victora GD, Nussenzweig MC. Germinal Centers. Annu Rev Immunol. 2012 Mar 26;30(1):429–57.
48. Mesin L, Ersching J, Victora GD. Germinal Center B Cell Dynamics. Immunity (Cambridge, Mass). 2016;45(3):471–82.
49. Manz RA, Hauser AE, Hiepe F, Radbruch A. Maintenance of serum antibody levels. Annual review of immunology. 2005;23(1):367–86.
50. K. Onoue, A. L. Grossberg, Y. Yagi, D. Pressman. Immunoglobulin M Antibodies with Ten Combining Sites. Science (American Association for the Advancement of Science). 1968;162(3853):574–6.
51. Sun Z, Almogren A, Furtado PB, Chowdhury B, Kerr MA, Perkins SJ. Semi-extended Solution Structure of Human Myeloma Immunoglobulin D Determined by Constrained X-ray Scattering. Journal of Molecular Biology. 2005;353(1):155–73.
52. Chen K, Xu W, Wilson M, He B, Miller NW, Bengtén E, et al. Immunoglobulin D enhances immune surveillance by activating antimicrobial, proinflammatory and B cell–stimulating programs in basophils. Nature immunology. 2009;10(8):889–98.
53. Tzaban S, Massol RH, Yen E, Hamman W, Frank SR, Lapierre LA, et al. The recycling and transcytotic pathways for IgG transport by FcRn are distinct and display an inherent polarity. The Journal of cell biology. 2009;185(4):673–84.
54. Woof JM, Mestecky J. Mucosal immunoglobulins. Immunological reviews. 2005;206(1):64–82.
55. Munblit D, Abrol P, Sheth S, Chow L, Khaleva E, Asmanov A, et al. Levels of Growth Factors and IgA in the Colostrum of Women from Burundi and Italy. Nutrients. 2018;10(9):1216-.
56. Kubo S, Nakayama T, Matsuoka K, Yonekawa H, Karasuyama H. Long Term Maintenance of IgE-Mediated Memory in Mast Cells in the Absence of Detectable Serum IgE. The Journal of Immunology. 2003;170(2):775–80.
57. Chang TW, Wu PC, Hsu CL, Hung AF. Anti‐IgE Antibodies for the Treatment of IgE‐Mediated Allergic Diseases. In: Advances in Immunology. United States: Elsevier Science & Technology; 2007. p. 63–119.

58. Tonegawa S. Somatic generation of antibody diversity. Nature (London). 1983;302(5909):575–81.
59. Bruce. Alberts, John H. Wilson, Hunt T. Molecular biology of the cell. 6th ed. New York: Garland Science; 2015. xxxiv+1342.
60. Rubelt F, Bolen CR, McGuire HM, Heiden JAV, Gadala-Maria D, Levin M, et al. Individual heritable differences result in unique cell lymphocyte receptor repertoires of naïve and antigen-experienced cells. Nature communications. 2016;7(1):11112-.
61. Mina MJ, Kula T, Leng Y, Li M, de Vries RD, Knip M, et al. Measles virus infection diminishes preexisting antibodies that offer protection from other pathogens. Science (American Association for the Advancement of Science). 2019;366(6465):599–606.
62. Greiff V, Menzel U, Miho E, Weber C, Riedel R, Cook S, et al. Systems Analysis Reveals High Genetic and Antigen-Driven Predetermination of Antibody Repertoires throughout B Cell Development. Cell reports (Cambridge). 2017;19(7):1467–78.
63. Greiff V, Miho E, Menzel U, Reddy ST. Bioinformatic and Statistical Analysis of Adaptive Immune Repertoires. Trends in Immunology. 2015;36(11):738–49.
64. DeWitt WS, Lindau P, Snyder TM, Sherwood AM, Vignali M, Carlson CS, et al. A Public Database of Memory and Naive B-Cell Receptor Sequences. PloS one. 2016;11(8):e0160853-.
65. Miho E, Roškar R, Greiff V, Reddy ST. Large-scale network analysis reveals the sequence space architecture of antibody repertoires. Nature communications. 2019;10(1):1321–11.
66. Briney B, Inderbitzin A, Joyce C, Burton DR. Commonality despite exceptional diversity in the baseline human antibody repertoire. Nature (London). 2019;566(7744):393–7.
67. Robins H. Immunosequencing: applications of immune repertoire deep sequencing. Current Opinion in Immunology. 2013;25(5):646–52.
68. Greiff V, Bhat P, Cook SC, Menzel U, Kang W, Reddy ST. A bioinformatic framework for immune repertoire diversity profiling enables detection of immunological status. Genome medicine. 2015;7(1):49-.
69. Stewart JJ, Lee CY, Ibrahim S, Watts P, Shlomchik M, Weigert M, et al. A Shannon entropy analysis of immunoglobulin and T cell receptor. Molecular immunology. 1997;34(15):1067–82.
70. Küppers R, Zhao M, Hansmann ML, Rajewsky K. Tracing B cell development in human germinal centres by molecular analysis of single cells picked from histological sections. The EMBO journal. 1993;12(13):4955–67.
71. Ehlich A, Martin V, Müller W, Rajewsky K. Analysis of the B-cell progenitor compartment at the level of single cells. Current Biology. 1994;4(7):573–83.
72. Klein U, Rajewsky K, Küppers R. Human Immunoglobulin (Ig)M+IgD+ Peripheral Blood B Cells Expressing the CD27 Cell Surface Antigen Carry Somatically Mutated Variable Region Genes: CD27 as a General Marker for Somatically Mutated (Memory) B Cells. The Journal of experimental medicine. 1998;188(9):1679–89.
73. Beck TF, Mullikin JC, Biesecker LG. Systematic Evaluation of Sanger Validation of Next-Generation Sequencing Variants. Clinical chemistry (Baltimore, Md). 2016;62(4):647–54.
74. Loman NJ, Misra RV, Dallman TJ, Constantinidou C, Gharbia SE, Wain J, et al. Performance comparison of benchtop high-throughput sequencing platforms. Nature biotechnology. 2012;30(5):434–9.

75. Archer J, Baillie G, Watson SJ, Kellam P, Rambaut A, Robertson DL. Analysis of high-depth sequence data for studying viral diversity: a comparison of next generation sequencing platforms using Segminator II. BMC bioinformatics. 2012;13(1):47–47.
76. Frey KG, Herrera-Galeano J, Redden CL, Luu TV, Servetas SL, Mateczun AJ, et al. Comparison of three next-generation sequencing platforms for metagenomic sequencing and identification of pathogens in blood. BMC genomics. 2014;15(1):96-.
77. Friedensohn S, Khan TA, Reddy ST. Advanced Methodologies in High-Throughput Sequencing of Immune Repertoires. Trends in Biotechnology. 2017;35(3):203–14.
78. Quail M, Smith ME, Coupland P, Otto TD, Harris SR, Connor TR, et al. A tale of three next generation sequencing platforms: comparison of Ion torrent, pacific biosciences and illumina MiSeq sequencers. BMC genomics. 2012;13(1):341-.
79. DeKosky BJ, Kojima T, Rodin A, Charab W, Ippolito GC, Ellington AD, et al. In-depth determination and analysis of the human paired heavy- and light-chain antibody repertoire. Nature medicine. 2014;21(1):86–91.
80. Oikonomopoulos S, Wang YC, Djambazian H, Badescu D, Ragoussis J. Benchmarking of the Oxford Nanopore MinION sequencing for quantitative and qualitative assessment of cDNA populations. Scientific reports. 2016;6(1):31602-.
81. Georgiou G, Ippolito GC, Beausang J, Busse CE, Wardemann H, Quake SR. The promise and challenge of high-throughput sequencing of the antibody repertoire. Nature biotechnology. 2014;32(2):158–68.
82. Goodwin S, McPherson JD, McCombie WR. Coming of age: ten years of next-generation sequencing technologies. Nature reviews Genetics. 2016;17(6):333–51.
83. Illumina CMOS Chip and One-Channel SBS Chemistry [Internet]. Available from: https://www.illumina.com/content/dam/illumina-marketing/documents/products/techspotlights/cmos-tech-note-770-2013-054.pdf
84. Ewing B, Green P. Base-Calling of Automated Sequencer Traces Using Phred. II. Error Probabilities. Genome research. 1998;8(3):186–94.
85. Ellefson JW, Gollihar J, Shroff R, Shivram H, Iyer VR, Ellington AD. Synthetic evolutionary origin of a proofreading reverse transcriptase. Science (American Association for the Advancement of Science). 2016;352(6293):1590–3.
86. He L, Sok D, Azadnia P, Hsueh J, Landais E, Simek M, et al. Toward a more accurate view of human B-cell repertoire by next-generation sequencing, unbiased repertoire capture and single-molecule barcoding. Scientific reports. 2014;4(1):6778-.
87. Kivioja T, Vähärautio A, Karlsson K, Bonke M, Enge M, Linnarsson S, et al. Counting absolute numbers of molecules using unique molecular identifiers. Nature methods. 2011;9(1):72–4.
88. Vázquez Bernat N, Corcoran M, Hardt U, Kaduk M, Phad GE, Martin M, et al. High-Quality Library Preparation for NGS-Based Immunoglobulin Germline Gene Inference and Repertoire Expression Analysis. Frontiers in immunology. 2019;10:660-.
89. Weinstein JA, Jiang N, White RA, Fisher DS, Quake SR. High-Throughput Sequencing of the Zebrafish Antibody Repertoire. Science (American Association for the Advancement of Science). 2009;324(5928):807–10.
90. Ohlin M, Owman H, Mach M, Borrebaeck CAK. Light chain shuffling of a high affinity antibody results in a drift in epitope recognition. Molecular Immunology. 1996;33(1):47–56.
91. Ramsköld D, Luo S, Wang Y-C, Li R, Deng Q, Faridani OR, et al. Full-length mRNA-Seq

from single-cell levels of RNA and individual circulating tumor cells. Nature biotechnology. 2012;30(8):777–82.
92. Macosko EZ, Basu A, Satija R, Nemesh J, Shekhar K, Goldman M, et al. Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. Cell (Cambridge). 2015;161(5):1202–14.
93. Gross A, Schoendube J, Zimmermann S, Steeb M, Zengerle R, Koltay P. Technologies for Single-Cell Isolation. International journal of molecular sciences. 2015;16(8):16897–919.
94. Espina V, Heiby M, Pierobon M, Liotta LA. Laser capture microdissection technology. Expert Review of Molecular Diagnostics. 2007;7(5):647–57.
95. Kolodziejczyk AA, Kim JK, Svensson V, Marioni JC, Teichmann SA. The Technology and Biology of Single-Cell RNA Sequencing. Molecular cell. 2015;58(4):610–20.
96. Brouzes E, Medkova M, Savenelli N, Marran D, Twardowski M, Hutchison JB, et al. Droplet microfluidic technology for single-cell high-throughput screening. Proceedings of the National Academy of Sciences - PNAS. 2009;106(34):14195–200.
97. Brown AJ, Snapkov I, Akbar R, Pavlović M, Miho E, Sandve GK, et al. Augmenting adaptive immunity: progress and challenges in the quantitative engineering and analysis of adaptive immune receptor repertoires. Molecular Systems Design & Engineering. 2019;4(4):701–36.
98. Tang F, Barbacioru C, Wang Y, Nordman E, Lee C, Xu N, et al. mRNA-Seq whole-transcriptome analysis of a single cell. Nature methods. 2009;6(5):377–82.
99. DeKosky BJ, Ippolito GC, Deschner RP, Lavinder JJ, Wine Y, Rawlings BM, et al. High-throughput sequencing of the paired human immunoglobulin heavy and light chain repertoire. Nature biotechnology. 2013;31(2):166–9.
100. Hashimshony T, Wagner F, Sher N, Yanai I. CEL-Seq: Single-Cell RNA-Seq by Multiplexed Linear Amplification. Cell reports (Cambridge). 2012;2(3):666–73.
101. Briggs AW, Goldfless SJ, Timberlake S, Belmont BJ, Clouser CR, Koppstein D, et al. Tumor-infiltrating immune repertoires captured by single-cell barcoding in emulsion. bioRxiv. 2017 Jan 1;134841.
102. Chen G, Ning B, Shi T. Single-Cell RNA-Seq Technologies and Related Computational Data Analysis. Frontiers in genetics. 2019;10.
103. Simon A. Babraham Bioinformatics - FastQC A Quality Control tool for High Throughput Sequence Data [Internet]. [cited 2020 Aug 24]. Available from: https://www.bioinformatics.babraham.ac.uk/projects/fastqc/
104. Vander Heiden JA, Yaari G, Uduman M, Stern JNH, O'Connor KC, Hafler DA, et al. pRESTO: a toolkit for processing high-throughput sequencing raw reads of lymphocyte receptor repertoires. Bioinformatics (Oxford, England). 2014;30(13):1930–2.
105. Khan TA, Friedensohn S, Gorter de Vries AR, Straszewski J, Ruscheweyh H-J, Reddy ST. Accurate and predictive antibody repertoire profiling by molecular amplification fingerprinting. Science advances. 2016;2(3):e1501371-.
106. Miho E, Yermanos A, Weber CR, Berger CT, Reddy ST, Greiff V. Computational Strategies for Dissecting the High-Dimensional Complexity of Adaptive Immune Repertoires. Frontiers in immunology. 2018;9:224-.
107. Singh M, Jackson KJL, Wang JJ, Schofield P, Field MA, Koppstein D, et al. Lymphoma Driver Mutations in the Pathogenic Evolution of an Iconic Human Autoantibody. Cell. 2020;180(5):878-894.e19.

108. Schultheiß C, Paschold L, Simnica D, Mohme M, Willscher E, von Wenserski L, et al. Next-Generation Sequencing of T and B Cell Receptor Repertoires from COVID-19 Patients Showed Signatures Associated with Severity of Disease. Immunity (Cambridge, Mass). 2020;53(2):442-455.e4.
109. Nemazee D. Mechanisms of central tolerance for B cells. Nature reviews Immunology. 2017;17(5):281–94.
110. Bonissone SR, Lima T, Harris K, Davison L, Avanzino B, Trinklein N, et al. Serum proteomics expands on high-affinity antibodies in immunized rabbits than deep B-cell repertoire sequencing alone. bioRxiv. 2020 Jan 1;833871.
111. Lavinder JJ, Horton AP, Georgiou G, Ippolito GC. Next-generation sequencing and protein mass spectrometry for the comprehensive analysis of human cellular and serum antibody repertoires. Current Opinion in Chemical Biology. 2015 Feb;24:112–20.
112. Luo S, Perelson AS. The challenges of modelling antibody repertoire dynamics in HIV infection. Philosophical transactions Biological sciences. 2015;370(1676):20140247-.
113. Yariv Wine, Daniel R. Boutz, Jason J. Lavinder, Aleksandr E. Miklos, Randall A. Hughes, Kam Hon Hoi, et al. Molecular deconvolution of the monoclonal antibodies that comprise the polyclonal serum response. Proceedings of the National Academy of Sciences - PNAS. 2013;110(8):2993–8.
114. Cheung WC, Beausoleil SA, Zhang X, Sato S, Schieferl SM, Wieler JS, et al. A proteomics approach for the identification and cloning of monoclonal antibodies from serum. Nature biotechnology. 2012;30(5):447–52.
115. Lee J, Paparoditis P, Horton AP, Frühwirth A, McDaniel JR, Jung J, et al. Persistent Antibody Clonotypes Dominate the Serum Response to Influenza over Multiple Years and Repeated Vaccinations. Cell host & microbe. 2019;25(3):367-376.e5.
116. Gregorich ZR, Chang Y-H, Ge Y. Proteomics in heart failure: top-down or bottom-up? Pflügers Archiv. 2014;466(6):1199–209.
117. Yates JR, Ruse CI, Nakorchevsky A. Proteomics by mass spectrometry: approaches, advances, and applications. Annual review of biomedical engineering. 2009;11(1):49–79.
118. Wine Y, Horton AP, Ippolito GC, Georgiou G. Serology in the 21st century: the molecular-level analysis of the serum antibody repertoire. Current Opinion in Immunology. 2015;35:89–97.
119. Catherman AD, Skinner OS, Kelleher NL. Top Down proteomics: Facts and perspectives. Biochemical and biophysical research communications. 2014;445(4):683–93.
120. Kachuk C, Doucette AA. The benefits (and misfortunes) of SDS in top-down proteomics. Journal of proteomics. 2018;175:75–86.
121. Vasicek LA, Zhu X, Spellman DS, Bateman KP. Direct quantitation of therapeutic antibodies for pharmacokinetic studies using immuno-purification and intact mass analysis. Bioanalysis. 2019;11(3):203–13.
122. Vinh J. CHAPTER 17 Proteomics and proteoforms: Bottom-up or top-down, how to use high-resolution mass spectrometry to reach the Grail. In: Fundamentals and Applications of Fourier Transform Mass Spectrometry [Internet]. Elsevier; 2019 [cited 2021 Mar 15]. p. 529–67. Available from: https://hal.archives-ouvertes.fr/hal-02358384
123. Chen B, Brown KA, Lin Z, Ge Y. Top-Down Proteomics: Ready for Prime Time? Analytical chemistry (Washington). 2018;90(1):110–27.
124. Kellie JF, Kehler JR, Mencken TJ, Snell RJ, Hottenstein CS. A whole-molecule

immunocapture LC–MS approach for the in vivo quantitation of biotherapeutics. Bioanalysis. 2016;8(20):2103–14.
125. Lanshoeft C, Cianférani S, Heudi O. Generic Hybrid Ligand Binding Assay Liquid Chromatography High-Resolution Mass Spectrometry-Based Workflow for Multiplexed Human Immunoglobulin G1 Quantification at the Intact Protein Level: Application to Preclinical Pharmacokinetic Studies. Analytical chemistry (Washington). 2017;89(4):2628–35.
126. Srzentić K, Fornelli L, Tsybin YO, Loo JA, Seckler H, Agar JN, et al. Interlaboratory Study for Characterizing Monoclonal Antibodies by Top-Down and Middle-Down Mass Spectrometry. J Am Soc Mass Spectrom. 2020 Sep 2;31(9):1783–802.
127. An B, Zhang M, Qu J. Toward sensitive and accurate analysis of antibody biotherapeutics by liquid chromatography coupled with mass spectrometry. Drug metabolism and disposition. 2014;42(11):1858–66.
128. Pitt JJ. Principles and applications of liquid chromatography-mass spectrometry in clinical biochemistry. Clinical biochemist reviews. 2009;30(1):19–34.
129. Chalkley R. Instrumentation for LC-MS/MS in proteomics. Methods in molecular biology (Clifton, NJ). 2010;658:47-.
130. Giansanti P, Tsiatsiani L, Low TY, Heck AJR. Six alternative proteases for mass spectrometry–based proteomics beyond trypsin. Nature protocols. 2016;11(5):993–1006.
131. Suhara Y, Kamao M, Tsugawa N, Okano T. Method for the Determination of Vitamin K Homologues in Human Plasma Using High-Performance Liquid Chromatography-Tandem Mass Spectrometry. Analytical chemistry (Washington). 2005;77(3):757–63.
132. Mathonet P, Ullman CG. The application of next generation sequencing to the understanding of antibody repertoires. Frontiers in immunology. 2013;4:265-.
133. Iversen R, Snir O, Stensland M, Kroll JE, Steinsbø Ø, Korponay-Szabó IR, et al. Strong Clonal Relatedness between Serum and Gut IgA despite Different Plasma Cell Origins. Cell reports (Cambridge). 2017;20(10):2357–67.
134. Bio D. Nadia Innovate | scRNA-Seq | Dolomite Bio Technology [Internet]. Dolomite Bio. [cited 2021 May 10]. Available from: https://www.dolomite-bio.com/product/nadia-innovate/
135. ExPASy PeptideCutter tool: available enzymes [Internet]. [cited 2021 Apr 19]. Available from: https://web.expasy.org/peptide_cutter/peptidecutter_enzymes.html
136. Ewels P, Magnusson M, Lundin S, Käller M. MultiQC: summarize analysis results for multiple tools and samples in a single report. Bioinformatics (Oxford, England). 2016;32(19):3047–8.
137. Shugay M, Britanova OV, Merzlyak EM, Turchaninova MA, Mamedov IZ, Tuganbaev TR, et al. Towards error-free profiling of immune repertoires. Nature methods. 2014;11(6):653–5.
138. Bolotin DA, Poslavsky S, Mitrophanov I, Shugay M, Mamedov IZ, Putintseva EV, et al. MiXCR: software for comprehensive adaptive immunity profiling. Nature methods. 2015;12(5):380–1.
139. Hershberg U, Luning Prak ET. The analysis of clonal expansions in normal and autoimmune B cell repertoires. Philosophical transactions Biological sciences. 2015;370(1676):20140239-.
140. Vadim Nazarov, immunarch.bot, Eugene Rumynskiy. immunomind/immunarch: 0.6.5: Basic single-cell support [Internet]. Zenodo; 2020 [cited 2021 Feb 17]. Available from:

https://zenodo.org/record/3893991
141. Xu JL, Davis MM. Diversity in the CDR3 region of V(H) is sufficient for most antibody specificities. Immunity (Cambridge, Mass). 2000;13(1):37–45.
142. Miqueu P, Guillet M, Degauque N, Doré J-C, Soulillou J-P, Brouard S. Statistical analysis of CDR3 length distributions for the assessment of T and B cell repertoire biases. Molecular immunology. 2007;44(6):1057–64.
143. M.K V, K K. A Survey on Similarity Measures in Text Mining. MLAIJ. 2016 Mar 30;3(1):19–28.
144. Bashford-Rogers RJM, Bergamaschi L, McKinney EF, Pombal DC, Mescia F, Lee JC, et al. Analysis of the B cell receptor repertoire in six immune-mediated diseases. Nature (London). 2019;574(7776):122–6.
145. Schickel J-N, Glauzy S, Ng Y-S, Chamberlain N, Massad C, Isnardi I, et al. Self-reactive VH4-34-expressing IgG B cells recognize commensal bacteria. The Journal of experimental medicine. 2017;214(7):1991–2003.
146. Kolde R. Pheatmap: pretty heatmaps.
147. Chaussabel D. Assessment of immune status using blood transcriptomics and potential implications for global health. Seminars in immunology. 2015;27(1):58–66.
148. Gotelli NJ, Colwell RK. Quantifying biodiversity: procedures and pitfalls in the measurement and comparison of species richness. Ecology letters. 2001;4(4):379–91.
149. Shannon C. A mathematical theory of communication. Mobile computing and communications review. 2001;5(1):3–55.
150. SIMPSON EH. Measurement of Diversity. Nature (London). 1949;163(4148):688–688.
151. Wolfgang H Berger, Frances L Parker. Diversity of Planktonic Foraminifera in Deep-Sea Sediments. Science (American Association for the Advancement of Science). 1970;168(3937):1345–7.
152. Béla Tóthmérész. Comparison of Different Methods for Diversity Ordering. Journal of vegetation science. 1995;6(2):283–90.
153. Philip Dixon. VEGAN, a package of R functions for community ecology. Journal of Vegetation Science. 2003;14(6):927–30.
154. Cox J, Mann M. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. Nature biotechnology. 2008;26(12):1367–72.
155. Marcou Q, Mora T, Walczak AM. High-throughput immune repertoire analysis with IGoR. Nature communications. 2018;9(1):561–561.
156. Lefranc M-P, Giudicelli V, Ginestoux C, Jabado-Michaloud J, Folch G, Bellahcene F, et al. IMGT®, the international ImMunoGeneTics information system. Nucleic acids research. 2009;37(suppl_1):D1006–12.
157. Weber CR, Akbar R, Yermanos A, Pavlović M, Snapkov I, Sandve GK, et al. immuneSIM: tunable multi-feature simulation of B- and T-cell receptor repertoires for immunoinformatics benchmarking. Bioinformatics. 2020;36(11):3594–6.
158. Tan G, Opitz L, Schlapbach R, Rehrauer H. Long fragments achieve lower base quality in Illumina paired-end sequencing. Scientific reports. 2019;9(1):2856–2856.
159. EuroFlow PID group, Blanco E. Age-associated distribution of normal B-cell and plasma cell subsets in peripheral blood. Journal of allergy and clinical immunology. 2018 Jun;141(6):2208-2219.e16.

160. Larimore K, McCormick MW, Robins HS, Greenberg PD. Shaping of human germline IgH repertoires revealed by deep sequencing. The Journal of immunology (1950). 2012;189(6):3221–30.
161. Darrell P Chandler, Christina A Wagnon, Harvey Bolton Jr. Reverse Transcriptase (RT) Inhibition of PCR at Low Concentrations of Template and Its Implications for Quantitative RT-PCR. Applied and Environmental Microbiology. 1998;64(2):669–77.
162. Wulf MG, Maguire S, Humbert P, Dai N, Bei Y, Nichols NM, et al. Non-templated addition and template switching by Moloney murine leukemia virus (MMLV)-based reverse transcriptases co-occur and compete with each other. The Journal of biological chemistry. 2019;294(48):18220–31.
163. Menzel U, Greiff V, Khan TA, Haessler U, Hellmann I, Friedensohn S, et al. Comprehensive evaluation and optimization of amplicon library preparation methods for high-throughput antibody sequencing. PloS one. 2014;9(5):e96727–e96727.
164. Wu J, Wang X, Lin L, Li X, Liu S, Zhang W, et al. Developing an Unbiased Multiplex PCR System to Enrich the TRB Repertoire Toward Accurate Detection in Leukemia. Frontiers in immunology. 2020;11:1631–1631.
165. Barennes P, Quiniou V, Shugay M, Egorov ES, Davydov AN, Chudakov DM, et al. Benchmarking of T cell receptor repertoire profiling methods reveals large systematic biases. Nature biotechnology. 2021;39(2):236–45.
166. Picelli S, Faridani OR, Björklund AK, Winberg G, Sagasser S, Sandberg R. Full-length RNA-seq from single cells using Smart-seq2. Nature protocols. 2014;9(1):171–81.
167. Zheng GXY, Terry JM, Belgrader P, Ryvkin P, Bent ZW, Wilson R, et al. Massively parallel digital transcriptional profiling of single cells. Nature communications. 2017;8(1):14049–14049.
168. Sato S, Beausoleil SA, Popova L, Beaudet JG, Ramenani RK, Zhang X, et al. Proteomics-directed cloning of circulating antiviral human monoclonal antibodies. Nature Biotechnology. 2012;30(11):1039–43.
169. Bondt A, Hoek M, Tamara S, de Graaf B, Peng W, Schulte D, et al. Human Plasma IgG1 Repertoires are Simple, Unique, and Dynamic [Internet]. Rochester, NY: Social Science Research Network; 2020 Dec [cited 2021 May 9]. Report No.: ID 3749694. Available from: https://papers.ssrn.com/abstract=3749694