Skoltech

Skolkovo Institute of Science and Technology

DYNAMICS OF IMMUNOGLOBULIN REPERTOIRES IN MEMORY AND ANTIBODY-SECRETING B CELL SUBSETS IN HEALTH AND DISEASE

Doctoral Thesis

by

ARTEM MIKELOV

DOCTORAL PROGRAM IN LIFE SCIENCES

Supervisor
Associate Professor Dmitriy Chudakov

Moscow - 2023

I hereby declare that the work presented in this thesis was carried out by myself at Skolkovo Institute of Science and Technology, Moscow, except where due acknowledgement is made, and has not been submitted for any other degree.

Candidate (Artem Mikelov)

Supervisor (As. Prof. Dmitriy Chudakov)

Abstract

B cells of late stages of differentiation - memory B-cells (Bmem), along with antibody-secreting cells (ASCs) - plasmablasts (PBL) and plasma cells (PL) are crucial for mounting an efficient immune response. B-cell receptor (BCR) / produced antibody repertoires of these subsets hold a spectrum of antigen specificities of humoral immune response. In this study we aimed to characterize and compare immunoglobulin heavy chain (IGH) repertoires of three cell subsets (Bmem, PBL and PL) in human peripheral blood over the course of one year.

Memory B cells (Bmem; $CD19^{+}$ $CD20^{+}$ $CD27^{+}$ $CD138^{-}$, plasmablasts (PBL; $CD20^{-}$ $CD19^{Low/+}$ $CD27^{++}$ $CD138^{-}$) and plasma cells (PL; $CD20^{-}$ $CD19^{Low/+}$ $CD27^{++}$ $CD138^{+}$) were isolated from peripheral blood of 6 volunteers in 3 time points using fluorescence activated cell sorting. For each cell sample full-length IGH cDNA libraries for next generation sequencing were obtained, incorporating molecular barcodes for error correction and data normalization. IGH repertoires were recovered using MiXCR. As a background for comparison we used a set of IGH repertoires of naive B-cell subset (Gidoni et al. 2019) processed with the same protocol as the original data. Distribution of isotypes within the repertoires of cell subsets was strikingly different. IgM represented more than a half of the repertoire in Bmem, while IgA was dominant in PBL and PL. The average CDR3 length and the level of somatic hypermutation (SHM) were significantly greater in ASCs than in Bmem. Despite these

differences, IGHV gene segment usage in all three studied subsets was similar, however diverged greatly from that of the naive B-cell subset. Moreover, we clustered IGHV genes based on their sequence similarity and observed that most IGHV segments in each of the four major clusters behaved concordantly with other IGHV segments in that cluster. Analysis of clonal overlaps between repertoires from different time points and individuals revealed large degree of privacy in repertoires of all subsets as well as stability of Bmem repertoires - there was no significant difference in clonal overlap between Bmem samples collected at the same day, 1 month or 1 year, which was not the case for ASCs. Finally we observe an excess of shared clonotypes between Bmem repertoires of unrelated donors compared to naive and pre-immune *in silico* generated repertoires. Our longitudinal analysis of IGH repertoires of Bmem, PBL and PL highlights distinct features of these subsets. The repertoire of Bmem subset is more stable over time, have more inter-individually shared clonotypes, and is dominated by IgM clonotypes. ASC subsets IGH bear signs of more active SHM, are private and transient in peripheral blood with high prevalence of IgA. The all three studied subsets concordantly differ from naive subset by their IGHV-gene segment usage, suggesting an important role of germline-encoded parts of the BCR for the initial recruitment of the B-cells to immune response.

Finally, we developed a novel approach for inference of allelic variants of V- and J-genes from adaptive immune receptor repertoire sequencing data, which showed superior flexibility and sensitivity to the existing tools.

List of publications

Publications

1. **Mikelov, A.**, Alekseeva, E. I., Komech, E. A., Staroverov, D. B., Turchaninova, M. A., Shugay, M., Chudakov, D. M., Bazykin, G. A., & Zvyagin, I. V. (2022). Memory persistence and differentiation into antibody-secreting cells accompanied by positive selection in longitudinal BCR repertoires. *eLife*, *11*, e79254. https://doi.org/10.7554/eLife.79254

2. Lomakin, Y. A., Zvyagin, I. V., Ovchinnikova, L. A., Kabilov, M. R., Staroverov, D. B., **Mikelov, A.**, Tupikin, A. E., Zakharova, M. Y., Bykova, N. A., Mukhina, V. S., Favorov, A. V., Ivanova, M., Simaniv, T., Rubtsov, Y. P., Chudakov, D. M., Zakharova, M. N., Illarioshkin, S. N., Belogurov, A. A., Jr, & Gabibov, A. G. (2022). Deconvolution of B cell receptor repertoire in multiple sclerosis patients revealed a delay in tBreg maturation. *Frontiers in immunology*, *13*, 803229. https://doi.org/10.3389/fimmu.2022.803229

3. **Mikelov, A. I.**, Staroverov, D. B., Komech, E. A., Lebedev, Y. B., Chudakov, D. M., & Zvyagin, I. V. (2019). Correlated dynamics of serum IGE and IGE+ clonotype count with allergen air level in seasonal allergic rhinitis. Bulletin of Russian State Medical University, 9(5), 13-22. https://doi.org/10.24075/brsmu.2019.072

Conference presentations

1. **Mikelov, A.,** Komech, E. A., Staroverov, D. B., Turchaninova, M. A., Shugay, M., Chudakov, Zvyagin, I. V. Common and distinct features of memory B-cell, plasmablast and plasma cell subsets revealed by longitudinal immunoglobulin heavy chain repertoires analysis. 20th EAACI Immunology Winter School, 2022, January 27 – 30. Poster
2. **Mikelov, A.,** Komech, E. A., Staroverov, D. B., Turchaninova, M. A., Shugay, M., Chudakov, Zvyagin, I. V. Longitudinal Profiling Of Immunoglobulin Heavy-chain Repertoires In Memory B-cells, Plasmablasts And Plasma Cells From Peripheral Blood Of Individuals With Birch Pollen Allergy. European Academy of Allergy and Clinical Immunology Annual Congress 2019, Lisbon, Portugal, 2019, June 1-5. https://doi.org/10.1111/all.13959. Poster

Author contribution

A. Mikelov designed the study, recruited donors, participated in sample collection and FACS sorting, optimized protocol for cDNA libraries preparation, prepared cDNA libraries for all of the immune repertoire sequencing in the study, analyzed the data from raw sequencing reads to final plots, developed the algorithm for novel alleles inference and tested its performance.

Acknowledgments

I would like to appreciate great people who helped me during the preparation of this Ph.D Thesis.

First of all, I would like to thank Dr. Ivan Zvyagin who provided support, guidance and advice during the early stages of this Ph.D. research. You have introduced me to the world of immunology, your enthusiasm and passion ignited my interest in adaptive immunity, while your scientific rigor always guarded me from wishful thinking and from dubious conclusions. I am immensely grateful for that.

I am also grateful to my scientific advisor Dr. Dmitriy Chudakov for the guidance, discussion and framing of this Ph.D. thesis. Your support during the hardest times was crucial, this work would never see the light without your guidance at the late stages of the research.

I would also like to thank former and current members of Ivan Zvyagin's lab and dept. of Adaptive Immunity at Institute of Bioorganic Chemistry RAS: Ekaterina Komech, Viktoria Fomchenkova, Anastasia Barinova, Dmitriy Staroverov and Maria Turchaninova, Yuri Lebedev, Mark Izraelson. I learned a lot from you, your contribution to research presented in this thesis was indispensable.

I dedicate special thanks to MiLaboratories inc. and personally the developers of MiXCR software Dmitriy Bolotin, Stanislav Poslavskiy and Georgiy Nefediev. Your software saved me an incalculable amount of time and effort and allowed me to perform sequencing data processing without anguish and frustration which usually accompanies working with bioinformatic tools.

I greatly appreciate my wonderful collaborators Evgenia Alekseeva and Georgiy Bazykin, who helped me to develop understanding of antibody immune repertoires from an evolutionary point of view. I praise Evgenia's patience and stamina demonstrated during the preparation of our eLife manuscript.

Finally, I would like to thank my wonderful wife Valeriia, my family and friends, who supported me all along the way.

Table of Contents

List of Symbols, Abbreviations

5’ RACE Rapid amplification of cDNA ends

ASC Antibody-secreting cell

BCR B-cell receptor

Bmem Memory B-cell

cDNA Complementory DNA

CDR Complementarity determining region

FACS Fluorescence- activated cell sorting

FC Fold change

FDCs Follicular dendritic cells

FR Framework region

GC Germinal center

IGH Immunoglobulin heavy chain

LLPCs Long-lived plasma cells

LNs Lymph nodes

MHCII Major histocompatibility complex class II

mRNA Messanger RNA

NGS Next-generation DNA sequencing

PBL Plasmablast

PBMC Peripheral blood mononuclear cells

PCR Polymerase chain reaction

PL Plasma cell

pMHC Peptide-MHC complex

RT Reverse transcription

SHM Somatic hypermutations

TCR T-cell receptor

Tfh Follicular helper cells

Th Helper T cells

Treg Regulatory T cells

UMI Unique molecular identifier

List of Figures

List of Tables

Chapter 1. Introduction

B cells play a crucial role in protection from various pathogens and cancer cells as well as regulation of the immune response. The structural diversity of B cell receptors (BCRs) is responsible for the B cell-mediated immune system's capacity to recognize a wide variety of different antigens, and every individual harbors a large pool of naive B cell clones, each with a unique BCR. Antigenic challenge triggers the proliferation and maturation of naive B cells with cognate BCRs, and the resulting progeny comprise a number of cell subsets with differing functions and lifespans. During the affinity maturation process, the initial structure of a given BCR can change at the genomic level as a result of somatic hypermutation (SHM), a process that accompanies B cell proliferation after antigen-specific activation. Cells bearing BCRs with higher affinity to the antigen are favored during the affinity maturation process, and produce signals that stimulate further differentiation and expansion (De Silva and Klein 2015). Another process called class-switch recombination further increases the dimensionality of the BCR space. The five main classes, or isotypes, of antibodies (*i.e.,* IgA, IgD, IgE, IgG, and IgM) have different functions in the immune response (Stavnezer, Guikema, and Schrader 2008; Vidarsson, Dekkers, and Rispens 2014), and isotype switching during clonal proliferation can thereby change the functional capabilities of B cells and the antibodies they produce. As a consequence, antigen challenge yields a population of clonally related cells with different BCRs and functionalities.

Recently developed immune repertoire sequencing techniques

provide valuable insights into the development and structure of B-cell immunity with clonal-level resolution. Studies of B-cell receptor clonal repertoires in health and disease have already provided valuable insights in pathogenesis of various conditions and adaptive immunity organization with important implications for treatment and vaccine development. However, detailed characterization of dynamics of BCR clonal repertoires of differentiated cell subsets of B-cell lineage has been lacking until recently, while being of interest because these subsets, in particular memory B cells, plasmablasts, plasma cells, substantiate crucial part of human immune memory - both at humoral and cellular level.

We have investigated immunoglobulin heavy chain repertoires from memory B cells, plasmablasts, and plasma cells from peripheral blood collected from generally healthy volunteers at three time points over the course of a year. In order to obtain detailed and unbiased repertoire data, we used advanced IgH repertoire profiling technology that provides full-length IgH variable region sequences with isotype annotation. Based on comparative and phylogenetic analysis of the resulting data, we are able to describe the structure, distinctive features, clonal relations, isotype distribution and temporal dynamics of B cell subset repertoires.

We have also developed a novel tool for inferring allelic variants of V- and J-genes from adaptive immune receptor repertoires which has a number of substantial advantages compared to existing tools and most importantly is significantly more sensitive.

Chapter 2. Literature review

2.1 Adaptive immune responses

Human immune system protects our bodies from external and internal threats, such as bacteria, viruses or fungi as well as cancer. Two interdependent systems can respond to a threat - innate and adaptive immunity. Innate immunity is the first line of defense after the external barriers (skin, mucosa etc) were penetrated - it mounts a broad initial response of low specificity via pattern recognition receptors on innate immune cells, such as neutrophils, macrophages, eosinophils and dendritic cells as well as through humoral response predominantly via complement system[1]. Adaptive immunity mounts a more specific and efficient response with T cells and B cells. However, this response is rather slow on first encounter to a specific pathogen. On the other hand, these cell subsets can form immunological memory, granting a long-lasting and reliable protection against already seen pathogens, both humoral and cellular.

T cells may have very distinct functionalities, while the main difference is between CD8+ and CD4+ T-cells. Primary function of CD8+ cytotoxic T cells (killer T cells) is killing of infected or cancer cells. CD4+ T cells have a much more diverse range of functions: these T cells can either activate a particular mode of immune response (helper T cells, Th) or suppress it (regulatory T cells, Treg). For example, Th1 cells elevate intracellular killing mechanisms of macrophages, while Th2 cells can enhance extracellular pathogen

clearing. Follicular helper cells (Tfh) provide help for B-cell differentiation to antibody-secreting cells (ASCs), and B-cell memory formation (Crotty, 2019). ASCs secrete large amounts of immunoglobulins which bind specifically to pathogens and their toxins, resulting in neutralization and activation of the complement system. Some of the ASCs can become long-living plasma cells (PL) and reside in the specialized niches in the bone marrow for many years, ensuring a long-lasting protection against repeated challenge with the same pathogen.

2.2 B-cell differentiation to ASC and memory subsets

The initial diversity of B-cell receptors is immense, currently the estimated lower bound is 10^{16} different variants (Briney et al., 2019; Mora & Walczak 2018). Within this diversity the frequency of the cells with a receptor specific to a particular antigen is extremely low. Even within the memory B-cell compartment blood such frequency hardly reaches 10^{-4} as detected by contemporary methods. Within the naive B-cell subset one can expect even lower frequencies of antigen-specific cells, as they have not undergone antigen-driven clonal expansions. Initiation of the B-cell lineage immune response requires that these rare antigen-specific cells encounter the antigen. Secondary lymphoid tissues, such as lymph nodes (LNs), spleen or Peyer patches, serve as major anatomical compartments for this encounter. Lymphocytes from the blood are constantly recruited to these organs for training in specialized compartments (lymphoid

follicles), guided by a chemokine CXCL13, which is the ligand for chemokine receptor CXCR15. The incoming antigens are also delivered to these follicles - lymphoid tissues continually filter body fluids to capture those antigens. (Cyster, 2010). Various cell types handle the intact antigen to deliver it for the B-cells to encounter; e.g. lymph node subcapsular sinus macrophages are the initial handlers of the antigen in the LNs. The intact antigen is delivered to the centers of lymphoid follicles where specialized stromal cells, called follicular dendritic cells (FDCs) present the intact antigens to B-cells. Antigen opsonization (either by antibodies or by complement) also facilitates presentation of antigen on FDCs.(Allen and Cyster, 2008; Heesters et al., 2014). FDCs along with other stromal cells within the follicle also serve as the source of CXCL13. Besides lymphoid stromal cells produce an important B-cell survival factor BAFF (Cremasco et al., 2014; Rodda et al., 2018). In case the B-cell does not encounter antigen after several hours it exits the lymphoid organ due to sensing sphingosine-1-phosphate (S1P) with a S1PR1 receptor (Cyster and Schwab, 2012) and enters a circulatory fluid, either blood (exiting from spleen) or lymphatic vessels (exiting from LNs or Peyer patches). Then it travels to another lymphoid organ to continue surveillance.

In case the contact between an antigen and a B-cell occurs. The exact mechanism by which BCR binding to antigen initiates downstream signaling is yet to be fully understood. One of the current models propose that the BCR clustering on the membrane may be the initiating event (Liu et al., 2016). Another model

postulates that on the contrary the disruption of pre-existing BCR clusters may have the leading role in signal initiation (Yang and Reth, 2016). The subsequent signaling relies on phosphorylation of IgM- and IgD-associated Igα and Igβ molecules, which contain intracellular ITAM motifs (Kurosaki et al., 2010; Yang and Reth, 2016). Stimulation of some co-receptors, e.g. toll-like receptors, leads to signal amplification and alteration (Suthers and Sarantopoulos, 2017). Another important co-receptor, CD19, associates with IgM and IgD and triggers the PI3 kinase-Akt pathway (Kurosaki et al., 2010). Initial BCR-signaling leads to multiple changes in B-cell transcription program including enhancing antigen-presenting capabilities, increasing chemokine receptors and costimulatory molecules (CD80/86) expression. Next, internalization of BCR with antigen occurs in clathrin-dependent way, also requiring tyrosine phosphorylation in Igα and Igβ subunits by Src-family kinases.(Hoogeboom and Tolar, 2016). Subsequent trafficking and processing of the antigen in endosomes and, then, in lysosomes, leads to antigen peptides presentation in the context of major histocompatibility complex class II (MHC II) for receiving T-cell help. The contact with the T-cell via cognate interaction of T-cell receptor (TCR) and peptide-MHC complex (pMHC) is the requirement for further activation of a B-cell, triggering cell proliferation and differentiation. However, this requirement can be overcome in two ways. First, in type 1 T-independent response signaling via co-receptors (e.g. via TLRs) may serve as the “second signal” necessary for further B-cell activation, proliferation and survival. Second, in type 2 T-independent response binding of multivalent

antigens such as polysaccharides on encapsulated bacteria or polymeric proteins, by multiple BCRs on a B-cell can provide sufficient signal for downstream activation. In case of T-dependent response the “second signal” is provided by a CD4 helper T-cell, which can interact with a B-cell at the edge of the B-cell follicle and the T-cell zone. In order for such a rare event to occur (as both T- and B-cells specific to a particular antigen are rare) B-cells upregulate CCR7 and EBI2 - chemokine receptors guiding the pre-activated B-cell to the interface. T follicular helper cells (Tfh cells), which have been primed by antigen-presenting dendritic cells serve as the major source of costimulatory and survival signals. FDCs together with T- and B-cells form specialized structures within the follicles called germinal centers (GCs). Tfh cells interacts with the pre-activated B-cell by means of cognate recognition of pMHC with cognate TCR, and also through interactions of costimulatory molecules such as ICAM1–LFA1 and SLAM family members (Akkaya et al. 2020,Crotty 2019). The “third signal” necessary for B-cell fate progression are secreted signal molecules cytokines, e.g. IL-21, which is necessary for maintenance of expression of crucial GC B-cells transcription factor Bcl-6 and somatic hypermutations (SHM). Naive B-cells entering the first phase GC reaction (phase 1) appear to be multipotent and can commit to at least three major fates. After differentiation to GC B-cells and acquiring SHMs they can enter phase 2 of GC reaction and subsequently differentiate into memory B-cells (Bmem) or long-lived plasma cells (LLPCs) (Cyster & Allen 2019). Besides the possibility of generating memory B-cells independently of GCs was shown in BCL-6 deficient mice (Toyama

et al. 2002). These cells, however, have limited ability to acquire SHM in healthy donors (in contrast to ones affected to some of the autoimmune disorders) and appear to be unable to become LLPCs (Elsner & Shlomchik, 2020). The third possible fate for a naive B-cell is to become a short-lived plasma cell residing primarily in medullary cords of the red pulp of the spleen, with a lifespan limited to the duration of the infection. In phase 2 of GC response happens upon re-entry of GC B-cells to GCs. Here the B-cells continue to express BCL-6 as well as sphingosine 1-phosphate receptor 2 (S1P2), which ensures their retention in the GC (Green et al. 2011). In the GC dark zone B-cells undergo proliferation and SHM and then enter the light zone, where affinity selection happens. Again, one of three possible fates await a B-cell: becoming either a long-lived plasma cell or a memory B-cell, or dark zone re-entry as a GC B-cell for further rounds of SHM and selection. Fourth possibility, which in fact awaits the majority of the B-cells in GC reaction is undergoing apoptosis (Anderson et al., 2009; Mayer et al., 2017).

It has been shown recently that GC reactions have temporal switches which determine to which fate the B-cells commit. Bmem are predominantly produced in the early GCs, while LLPCs emerge very late in the GC responses (Weisel et al. 2016), which may explain the observed higher levels of SHM in antibody-secreting cells (ASCs) (Phad et al. 2022) and assume less mutated more broadly reactive BCRs in Bmem compared to LLPCs.

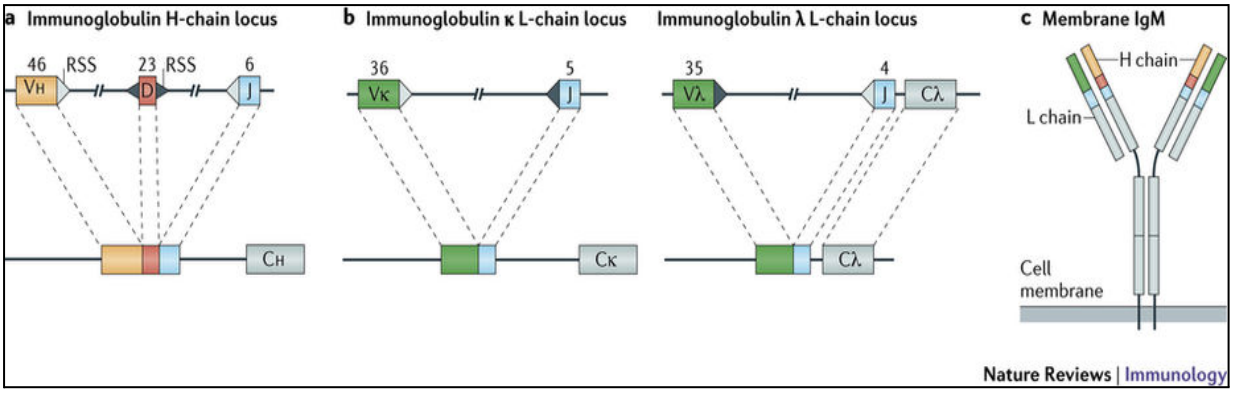
2.3 B-cells recognize antigens with a highly variable B-cell receptor, generated in stochastic, albeit directed manner

Recognition of antigens by B-cells is mediated by the membrane form of an antibody, also called B-cell receptor (BCR). DNA sequence, coding for this receptor, can serve as a defining feature for each clonotype in B-cell lineage.

In every human, there is an enormous diversity of these receptors, and subsequently, of secreted antibodies. This diversity is generated by three processes, complementing each other: V(D)J recombination, somatic hypermutation, and class-switch recombination.

An antibody consists of two heavy (H) and two light (L) chains, each containing a variable domain, derived from DNA rearrangements. In this process, called V(D)J recombination, one of V, D (only for H-chain) and J segments is selected to form a mature BCR gene (Figure 1). Initial combinatorial diversity (e.g. 46 V × 23 D × 6 J for H-chain) is increased further by 'junctional' diversity, generated by V(D)J junction joining (Lefranc & Lefranc, 2020). Hairpin, formed between 5' and 3' ends of rearranging gene by ligation, can leave a 3' palindromic overhang when resolved by nicking. Moreover, each of the rearranging segments can lose several nucleotides during recombination. Finally, terminal deoxynucleotidyl transferase, acting in the process of DNA repair during recombination, can randomly add non-germline coded nucleotides at junction sites (Tonegawa, 1983). Thus, the initial antibody repertoire potential diversity of more than 10^{16} different variants is formed (Mora, 2018).

Figure 1. B-cell receptor generation by DNA segments rearrangement (adapted from Nemazee, 2017)



The second process, contributing to immunoglobulin repertoire diversity is called class-switch recombination. The heavy chain of an antibody contains a constant region, coded by one of the nine consecutively located CH genes, located downstream of the VDJ loci (Fig. 2). Extracellular signals, received upon exposure of B-cell to an antigen, such as cytokines, produced by T-cells, promote excision of one or several CH segments, allowing production of BCRs with different constant part (Fc fragment of Ig). While the variable part of an antibody defines its binding specificity, the Fc fragment sets the isotype and, therefore, the functional characteristics of the resulting immunoglobulin. For instance, the Cε segment is the one which is utilized in the production of IgE antibodies, allowing them to bind to Fcε receptors.

The final mechanism, affecting the diversity of Ig repertoire, is somatic hypermutation (SHM). SHM occurs on the variable domain genes of germinal center B-cells at a rate of approximately 10^{-3} alterations per base pair per cell division. SHM is the major mechanism necessary for affinity maturation of an Ig repertoire in response to the repeated immune challenge.

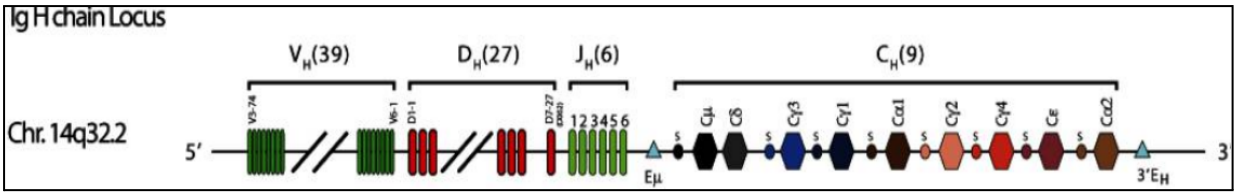


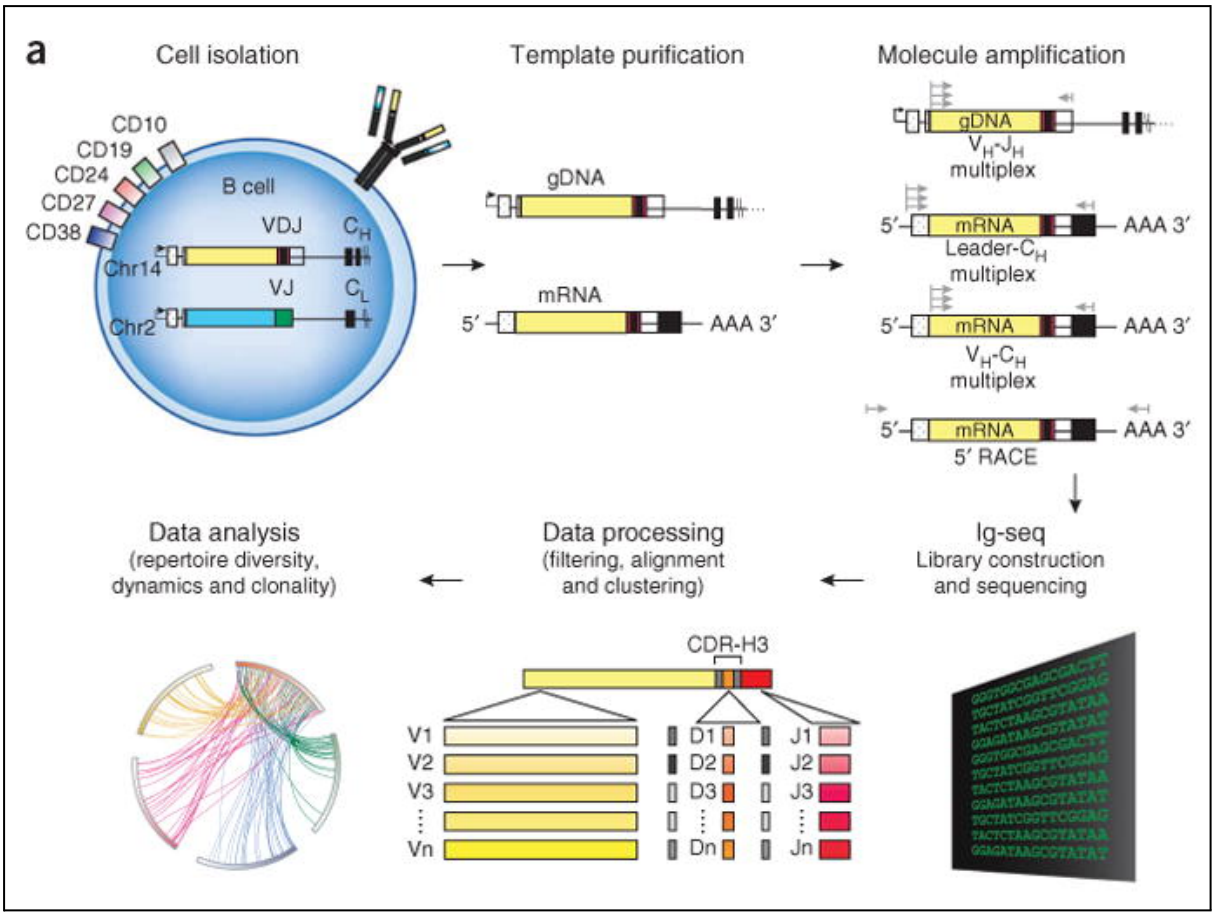
Figure 2. Human IGH locus organization (adapted from Schroeder et al., 2010)

Taken together these three processes yield an astronomical potential diversity of different antibodies. Taking into account the total number of B-cells in the human body, estimated as 1–2 × 10^{11}, it is clear that in a single individual only a small fraction of the potential pool of immunoglobulin variants is present (Mora, 2018).

2.4 Novel ways to study B-cell receptor repertoires were developed with the emergence of next-generation sequencing

Development of next-generation DNA sequencing (NGS) technologies in the last decade made high-throughput genomic studies possible. However, the nature of BCR loci, with its mix of highly variable regions and constant regions, hindered the investigation of BCR repertoires in the early days of NGS. Not until specialized techniques for library preparation for BCR repertoire sequencing (He et al., 2018; Turchaninova et al., 2016) were developed, it became

Figure 3. Conventional workflow for high-throughput sequencing of immunoglobulin repertoire (adapted from Georgio et al., 2014)



possible to fully utilize the power of NGS for research of adaptive immunity.

A typical workflow for immunoglobulin repertoire sequencing (repSeq) is similar to conventional approaches for amplicon sequencing and includes several steps: cell isolation, template purification, amplification, sequencing library preparation and sequencing, data processing and data analysis (Figure 3, Georgio et al., 2014). Nevertheless, several peculiar details should be carefully considered.

First of all, the most readily accessible source of B-cells in humans, the peripheral blood, contains only about 2% of total B-cells in a human body, while in the other body compartments, such as lymph nodes or bone marrow from 17% to 28% of total number of B-cells can be found. Moreover, the typical blood sample size hardly exceeds 8 ml of blood, thus limiting the sampling depth (from 1-2 × 10^9 of peripheral blood B-cells).

Second, heterogeneity of B-cell populations is the critical aspect, which should be taken into account. Both functional and phenotypical characteristics of B-cell subsets vary dramatically. Neglecting this fact may lead to misinterpretation and biased or incorrect conclusions. For instance, Ig-seq library preparation starting from bulk RNA, isolated from peripheral blood mononuclear cells (PBMC), will most likely result in a tangled data set. In such data set clonal sequences from transient plasma cells and plasmablasts will be overrepresented, some portion of clonal sequences, derived from the most abundant memory B-cells will still be present, and little or no sequences from low-frequency memory B-cells and naive B-cells will be present. If replica blood sample is taken, the results will most likely be incoherent between two replicas - in the absence of immune challenge peripheral blood plasma cells are rare and most likely clonally heterogeneous, but still yielding high copy number of Ig mRNA, which is enough to introduce bias into clonal quantification.

This also leads to the third important consideration - whether to use gDNA or mRNA as a template for Ig repSeq library preparation. Genomic DNA may be a good choice, if clonal quantification is the

primary goal of the research, however, there are several critical drawbacks in this approach. Due to the several kB of the intron, present in IGH C segment, it is impossible to infer isotype information, using gDNA as a template - typical NGS read length (max. 310 bp, paired reads for Illumina Hiseq) is not enough to cover such long amplicons. PacBio and Nanopore (the sequencing technologies with read-length long enough to overcome the problem) have unacceptable error rate and very low output. After all, gDNA-based protocols rely on multiplex PCR amplification, with multiple primers, specific to V and J segments. The efficiency of amplification is different for each primer, therefore clonotypes with different VJ combinations are amplified unequally, resulting in biased clonal proportions. Most of these problems are not relevant for the modern mRNA-based protocols, utilizing 5'-RACE amplification with the unique molecular identifier (UMI) and universal amplification adapter incorporation by template switch (e.g. Turchaninova et al., 2014). UMI labeling of initial cDNA molecules before PCR amplification allows unbiased quantification BCR repertoire as well as UMI-guided PCR error correction (Shugay et al., 2014; Ma & He et al., 2018). Despite these advantages of RNA-based approaches, significant variation of Ig mRNA copy number among various B-cell subsets still remains a challenge. At least partially it can be resolved by fluorescence- activated cell sorting (FACS) of cell subpopulations with different Ig mRNA expression levels - PCs, plasmablasts, memory and naive B-cells.

Emerging single-cell sequencing technologies allow to study immune repertoires at several additional layers of complexity. First, pairing of heavy and light antibody chain sequences becomes available at relatively high throughput. Second, the most advanced approaches allow linking BCR sequences with transcriptomic profiles for each particular cell. Besides, variation of Ig mRNA copy number variation between different B-cell subsets which was mentioned as a significant obstacle for clonal structure analysis in bulk RNA-based repertoire sequencing approaches, is no longer a problem for single cell sequencing assay. Despite these advantages, single-cell sequencing is still an emerging technology with significantly lower throughput and often prohibitive costs for studying larger cell populations. These features make studying of particular small cell populations of interest (e.g. tetramer-sorted antigen-specific B-cells) the main target for single cell sequencing approaches at the moment.

Thus, the choice of the particular approach for BCR repertoire sequencing should depend on particular research goals and be aligned with specific research questions.

2.5 B-cell clonal memory in various B-cell subsets through the length of high throughput BCR repertoire sequencing

Development of high-throughput DNA-sequencing methods allowed studying diversity and clonal structure of immune repertoires at unprecedented depth. Human bulk BCR repertoire characteristics, such as clonal diversity, extent of clonal expansions, level of SHM and degree of clonal sharing between were studied by multiple studies (Briney et al. 2019; Soto et al. 2019; Shah et al. 2019; Mandric et al. 2020; Yang et al. 2021). Those repertoire were also studied at the level of different isolated B cell subsets: naïve, marginal zone, switched and plasma cells (Ghraichy et al., 2021). The most recent effort by Phad et al. revealed high degree of clonal memory persistence in memory B-cells as well as clonal relatedness of circulating plasmablasts to persisting memory clonal lineages by means of high-throughput single-cell BCR repertoire sequencing of multiyear serial samples from two healthy adult donors (Phad et al. 2022).

Involvement of B-cell mediated immunity was also studied in patients with different pathologies helping to reveal mechanisms of the diseases (Bashford-Rogers et al. 2019; S. C. A. Nielsen et al. 2020; Gaebler et al. 2021; Sakharkar et al. 2021).

Longitudinal analysis of repertoires at different timepoints has made it possible to study the dynamics of B cell response following antigenic challenge or therapy (Laserson et al. 2014; Davydov et al. 2018; Horns et al. 2019; Nourmohammad et al. 2019; Hoehn et al. 2021). Reconstruction of BCR evolution in B cell clonal lineages and

phylogenetic analysis can reveal which evolutionary forces predominate at different stages of clonal lineage development. De Bourcy *et al.* recently reported on age-related differences in the structure of clonal lineages, somatic hypermutagenesis and affinity maturation processes, and differences in recall response of persisting lineages upon vaccination depending on CMV seropositivity status (de Bourcy et al. 2017). Other studies have described in detail the action of positive selection in the evolution of clonal lineages in vaccination and chronic HIV infection (Bonsignori et al. 2017; Horns et al. 2019; Nourmohammad et al. 2019). Reports have also described persisting clonal lineages which are predominantly represented by cells with IgM/IgD isotypes, and which demonstrate signs of neutral evolution (Horns et al. 2019). Wu *et al.* observed the clonal stability of plasma cells in bone marrow (Wu et al., 2016), representing the largest fraction of ASCs in the human body. Comparison of BCR repertoires between different cell subsets also makes it possible to investigate factors governing the functional assignment of B cells during proliferation, and thereby to understand fundamental aspects of B cell immunity. For example, recent studies have described differences in BCR repertoires of IgM and switched memory B cells as well as the complex interplay between $CD27^{high}$ and $CD27^{low}$ B-cell memory subsets, showing the complex nature of B cell immune memory (Wu et al. 2010; Grimsholm et al. 2020).

2.6 Inference of allele variants of V- and J-genes from adaptive immune receptor repertoires remains a challenge

The allelic diversity of V- and J-genes underlies the overall diversity of immune repertoires, and was shown to have functional impact on immune response in humans. Analysis of the adaptive immune repertoires and most importantly the downstream analysis of immunoglobulin repertoires also heavily depends on accurate allele calling and genotyping, e.g. for somatic hypermutations quantification and lineage trees construction. Adaptive immune repertoire sequencing (AIRR-seq) can be a very powerful technology for getting biological insights about the organization and the dynamics of the adaptive immunity as demonstrated in previous sections. Ability to precisely call known allelic variants and infer novel ones from the same AIRR-seq could add another angle to the analyses and also improve accuracy of many existing downstream approaches. There are several published approaches to the problem of genotyping and allelic inference of V- and J-genes from AIRR-seq data (Table 1, numbers 2-5), however each one of them has some crucial limitations. TIgGER (Gadala-Maria et al., 2015) and Partis (Ralph et al., 2017) are based on the same idea that allelic mutations show a very distinctive pattern over the background of SHMs, therefore requiring SHMs to be present in the dataset for reliable inference. This is not the case for many datasets, most obviously for T-cell receptor repertoire data and naive B-cell repertoires. On the other hand, IgDiscover (Corcoran et al., 2016), a very robust and reliable tool for novel allele inference, requires only data without

hypermutations thus excluding the majority of immunoglobulin repertoire datasets. Another algorithm, ImPre (Zhang et al., 2016), in theory could overcome these obstacles, however it is no longer supported, last code commit in the online repository (github.com/zhangwei2015/IMPre) dates back to 7 years ago; we were not able to run the tool to perform the analysis.

Table 1. Tools for inference of novel alleles and genotyping from AIRR-seq data and their characteristics.

#	Tool name	Year	Supported chain type(s)	Supported gene type(s)	Programming language(s)	Suitable for inference from unmutated repertoires	Suitable for inference from hypermutated repertoires	Works with the minimal starting reference
1	MiStrainer	2023	IGH, IGK, IGL, TRA, TRB	V, J	Java, Kotlin	Yes	Yes	Yes
2	TIgGER	2015	IGH, IGK, IGL	V	R	No	Yes	No
3	IgDiscover	2016	IGH, IGK, IGL, TRA, TRB	V, D, J	Python	Yes	No	No
4	Partis	2019	IGH, IGK, IGL	V	C,C++,Perl, Python	No	Yes	No
5	ImPre	2016	IGH, IGK, IGL, TRA, TRB	V, J	C,Perl	Yes	Yes	No

Another issue common for all of the existing tools is that all of them are quite demanding to the depth of AIRR-seq data for the reliable inference (e.g. IgDiscover recommends at least 750,000 sequencing reads per individual library), while such depth of sequencing creates stiff economical constraints and the most of the publicly available AIRR-seq datasets do not reach such depth. Thus there is an evident need for more sensitive and flexible novel tools addressing the same question of V- and J-gene genotyping from AIRR-seq data. Here we present MiStrainer, a tool developed for this task which lacks the above mentioned limitations - it allows allelic inference and genotyping from both hypermutated and non-hypermutated repertoires and has much milder requirements for the depth of sequencing. Another important aspect is that the algorithm performs well starting with the minimalistic gene reference library with only one allele of each gene present and even sparser. This peculiar feature makes the tool especially useful for studying allelic diversity in novel species where the reference gene libraries are sparse and usually incomplete.

Chapter 3. Methodology

3.1 Study design, cohort, cells, and timepoints

Blood samples from six (4 males and 2 females) young and middle-aged donors (23, 27, 27, 33, 33, and 39 y.o.) without severe inflammatory diseases, chronic or recent acute infectious diseases, or vaccinations were collected at three time points (T1 - 0, T2 - 1 month, T3 - 12 months) (Figure 4); donor details and the number cells collected for each time point and cell subset are provided in Table 2. Four donors suffered allergic rhinitis to pollen, and two also suffered from food allergy. Informed consent was obtained from each donor. The study was approved by the Ethical Committee of Pirogov Russian National Research Medical University, Moscow, Russia. At each time point, 18–22 mL of peripheral blood was collected in BD Vacuette tubes with EDTA. Peripheral blood mononuclear cells were isolated using Ficoll gradient density centrifugation. To isolate subpopulations of interest, cells were stained with anti-CD19-APC, anti-CD20-VioBlue, anti-CD27-VioBright FITC, and anti-CD138-PE-Vio770 (all Miltenyi Biotec) in the presence of FcR Blocking Reagent (Miltenyi Biotec) according to the manufacturer's protocol, and then sorted using fluorescence-activated cell sorting (FACS; BD FacsAria III, BD Biosciences) into the following populations: memory B cells (Bmem; $CD19^{+}$ $CD20^{+}$ $CD27^{+}$ $CD138^{-}$, plasmablasts (PBL; $CD20^{-}$ $CD19^{Low/+}$ $CD27^{++}$ $CD138^{-}$), plasma cells (PL; $CD20^{-}$ $CD19^{Low/+}$ $CD27^{++}$ $CD138^{+}$). For each donor at T1, one replicate sample of each cell subpopulation was collected. At T2 and

T3, two replicate samples were collected (50×10^3 to 100×10^3 Bmem, 1×10^3 to 2×10^3 PBL, 0.5×10^3 to 1×10^3 PL per sample).

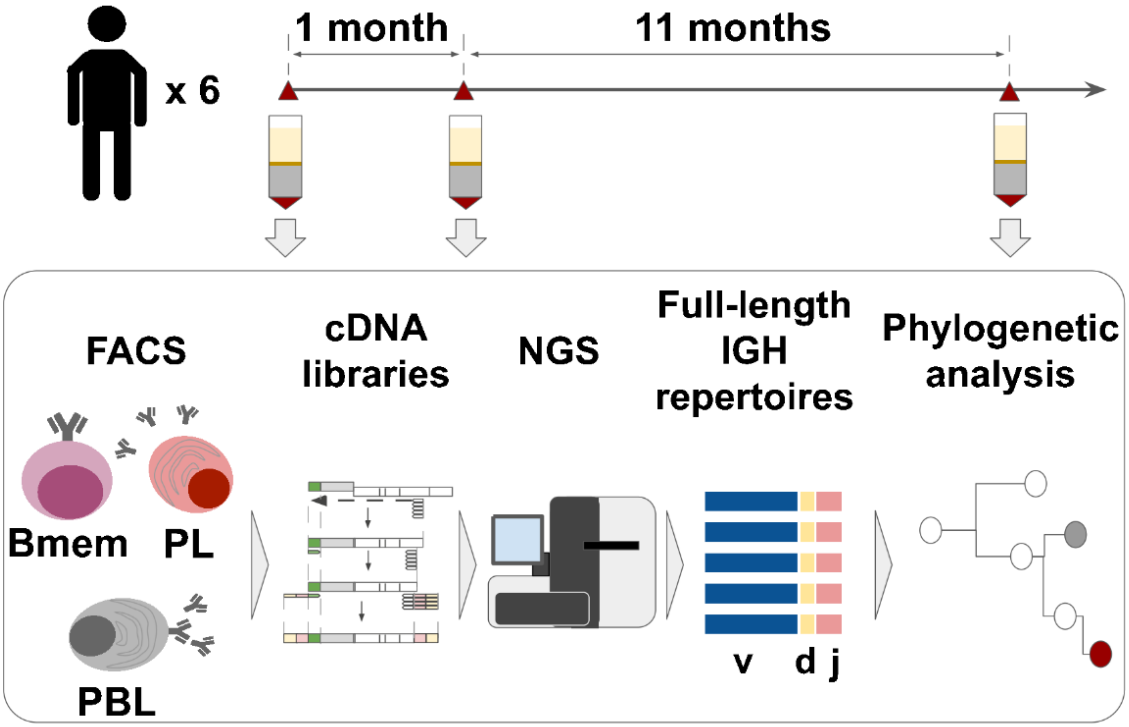


Figure 4. Study design. Peripheral blood from six donors was sampled at three time points: T1 - initial time point, T2 - 1 month and T3 - 12 months after the start of the study. At each time point, we isolated PBMCs and sorted memory B cells (Bmem: $CD19^{+}$ $CD20^{+}$ $CD27^{+}$), plasmablasts (PBL: $CD19^{low/+}$ $CD20^{-}$ $CD27^{high}$ $CD138^{-}$) and plasma cells (PL: $CD19^{low/+}$ $CD20^{-}$ $CD27^{high}$ $CD138^{+}$) in two replicates using FACS. For each cell sample, we obtained IGH clonal repertoires by sequencing respective cDNA libraries covering full-length IGH variable domain.

Table 2. Donor demographics and cell samples size. Several values in table cells separated by a semicolon represent replicates collected for corresponding donor, time point and cellular subset. AR - allergic rhinitis; FA - food allergy; HD - healthy donor.

				Number of cells per sample								
		Time point			T1			T2			T3	
Donor ID	Age	Sex	Status	Bmem	PBL	PL	Bmem	PBL	PL	Bmem	PBL	PL
D01	27	F	AR	n/a	n/a	n/a	50,300; 55,400	2,100; 2,100	1,020; 1,010	50,000; 50,000	1,000; 1,000	500; 500
IM	39	M	AR,FA	186,572	2,200	129	69,900; 68,400	2,000; 2,486	920	50,000; 50,000	2,000; 2,000	1,000; 1,000
MRK	27	M	AR	143,162	5,336	251	51,700; 50,600	2,130; 2,020	1,000; 1,035	50,000; 50,000	1,000; 1,000	400; 200
AT	23	M	AR,FA	101,400	7,200	1,800	50,600; 57,400	2,520	800	50,000; 40,800	1,000; 1,000	400; 400
IZ	33	M	HD	101 800	3,900	850	50,500; 56,300	1,140; 1,840	1,050; 625	50,000; 50,000	2,000; 2,000	200; 200
MT	33	F	HD	n/a	n/a	n/a	n/a	n/a	n/a	50,000; 50,000	1,000; 1,000	400

3.2 IGH cDNA libraries and sequencing

IGH cDNA libraries were prepared as described previously (Turchaninova et al. 2016) with several modifications. Briefly, we used a rapid amplification of cDNA ends (5’ RACE) approach with a template-switch effect to introduce 5’ adaptors during cDNA synthesis. These adaptors contained both unique molecular identifiers (UMIs), allowing error-correction, and sample barcodes (described in Zvyagin et al. 2017), allowing us to rule out potential cross-sample contaminations. In addition to a universal sequence for annealing the forward PCR primer, we also introduced a 5' adaptor during the reverse transcription (RT) reaction, which allowed us to avoid using multiplexed forward primers specific for V segments, thereby reducing PCR amplification biases. Multiplexed C-segment-specific primers were used for RT and PCR, allowing us to preserve isotype information. Prepared libraries were then sequenced with an Illumina HiSeq 2000/2500, (paired-end, 2 x 310 bp).

3.3 Sequencing data pre-processing and repertoire reconstruction

Sample demultiplexing by sample-barcodes introduced in the 5' adapter and UMI-based error-correction were performed using MIGEC v1.2.7 software (Shugay et al. 2014). For further analysis, we used sequences covered by at least three sequencing reads. Alignment of sequences, V-, D-, J-, and C-segment annotation, and

reconstruction of clonal repertoires were accomplished using MiXCR v3.0.10 (Bolotin et al. 2015) with prior removal of the primer-originated component of the C-segment. We defined clonotypes as a unique IGH nucleotide sequence starting from the framework 1 region (FR1) of the V segment to the end of the J segment, and taking into account isotype. Using TIgGER (Gadala-Maria et al. 2015) software, we derived an individual database of V gene alleles for each donor and realigned all sequences for precise detection of hypermutations. For analysis of general repertoire characteristics (isotype frequencies, SHM levels, CDR3 length, IGHV gene usage, and repertoire similarity metrics) we used samples covered by at least 0.1 cDNA molecules per cell for Bmem, and at least 5 cDNA per cell for PBL and PL.

3.4 Repertoire characteristics analysis

Isotype frequencies, rate of SHM, and CDR3 lengths were determined using MiXCR v3.0.10 (Bolotin et al. 2015). For calculation of background IGHV gene segment usage and number of shared clonotypes, we utilized data derived from Gidoni et al. 2019 (European Nucleotide Archive accession number ERP108501) representing naive B cell IGH repertoires, where the IGH cDNA libraries were prepared using 5'RACE-based protocol similar to the protocol used in the current study. We used repertoires containing more than 5,000 clonotypes and processed them in the same way as our data, 20 repertoires in total. IGHV gene frequencies were calculated as the number of unique clonotypes to which a particular IGHV gene was annotated by MiXCR divided by the total number of

clonotypes identified in the sample. To assess IGHV gene segments over- and under-represented in studied subsets, we utilized edgeR package v0.4.4 (Robinson et al., 2010b) with the ‘trended’ dispersion model using trimmed mean of M values method for normalization (Robinson et al., 2010a). To evaluate pairwise similarity between repertoires based on IGHV gene segment frequency distributions, we utilized Jensen-Shannon divergence, calculated using the following formula:

$$JS(P,Q) = \frac{1}{2}\sum_i p_i log_2 p_i + \frac{1}{2}\sum_i q_i log_2 q_i - \sum_i \left(\frac{p_i+q_i}{2} log_2 \left(p_i + q_i \right) \right)$$

where P and Q represent distributions of IGHV gene segment in two repertoires, and p_i and q_i represent frequencies of individual member *i* (IGHV gene segment). *In silico* repertoires used for the calculation of background clonal overlap (each repertoire contained 5000 clonotypes) were generated with OLGA software v1.0.2 (Sethna et al. 2019) under standard settings utilizing the built-in model. For clonal overlap calculation, we downsized repertoires to a fixed number of clonotypes. For **Fig. 12,13 ,** the 14,000 most abundant clonotypes were considered in Bmem, 600 in PBL, and 200 in PL. Cell samples with smaller numbers of clonotypes were excluded from comparisons: IM T1 repl. 1, MRK T1 repl. 1, MRK T3 replicate 2, IZ T3 replicates 1,2; all from PL subset. For **Fig. 14**, we considered 5,000 clonotypes for all cell subsets. Clonotypes with identical CDR3 amino acid sequence and the same IGHV gene segment detected in both analyzed samples were considered

shared. Clonotypes shared between repertoires of at least two individuals were termed as public.

3.5 Statistical analysis and plotting

Statistical analysis was performed utilizing R programming language (R Core Team, 2018) and *ggpubr* v0.4.0. Plots were generated using *ggplot2* package v3.3.4 for R (Wickham, 2016) .

3.6 Allele variants detection algorithm

The algorithm utilizes alignment and clonotype assembly information from the upstream repSeq data processing, specifically mutation calls from reference V- and J-gene reference library for BCR or TCR clonotypes and V- and J-gene annotations, readily available after running ‘analyze’ command of MiXCR software (Bolotin et al. 2015). The clonotype definition for the purpose of allele inference may vary depending on the region covered by sequencing; the approach was tested for clonotypes defined by the unique full length nucleotide sequence of variable region of BCR and TCR. Clonotype definition by a shorter nucleotide sequence spanning from the beginning of Complementarity Determining Region 1 (CDR1) to the end of Framework region 4 (FR4) was also tested with only differences in allelic mutations in the region left out from the full-length region (Framework region 1).

Then allele inference for V- and J- proceeds separately but in the same manner. For simplicity here we demonstrate the further steps of inference for V-genes:

1. Clonotypes are grouped by the annotated V-gene segments.
2. For each mutation from the corresponding V-gene in the reference library within the group, including insertions and deletions, we define a set of clonotypes which contain this mutation.
3. The mutations are filtered using a threshold on the diversity of the combinations of J-genes and CDR3-lengths of clonotypes containing that mutation. The mutations that do not have sufficient diversity of J-genes and CDR3-lengths are then removed from each of the clonotype's mutation sets, thus reducing the noise. However this step is not alone sufficient for separating allelic mutations from SHMs due to presence of both hot-spot hypermutations and PCR and sequencing errors.
4. Then clonotypes are further grouped by these filtered mutation sets, including empty mutation sets, representing alleles already present in the starting library. The diversity of J-genes and CDR3 lengths combinations within each of these groups is determined along with the number of clonotypes containing no mutations in J-gene (after filtering at step 3). Mutation sets are then filtered by thresholds of these two parameters, resulting in a list of allele candidates.

5. Clonotypes are then assigned to the closest allele candidates by comparing their sets of mutations. Candidates are then sorted by the score which is calculated by the number of combinations of different J-genes with the number of different CDR3 lengths and the number of clonotypes containing no mutations in J-gene (after filtering at step 3) in clonotypes assigned to these candidates. Formula for the score:

$$score = N_{Jgene-CDR3len} + 2 \cdot N_{naiveByJgene}$$

Where $N_{Jgene-CDR3len}$ is the number of combinations of different J-genes with the number of different CDR3 lengths; $N_{naiveByJgene}$ - number of clonotypes with no mutations in J-gene after filtering.

6. All candidates that have the score calculated above no less than 35% of the maximum score in that V-gene group are then selected for the resulting subject-specific gene set library.

The same process is then applied to J-gene alleles, but the diversity is calculated for V-genes as well as the number of clonotypes with V-genes containing no mutations is considered at steps at 4 and 5.

3.7 Allele variants detection benchmarking

IGH repertoire sequencing data from Watson et al, 2022 were obtained upon request to authors along with the donor genotype reconstructed using Pacific Biosciences long-read sequencing in the original study. Initial processing of the raw files in fastq format was

performed using MiXCR v4.2.0 (Bolotin et al. 2015) upstream pipeline ‘analyze’ command. Importantly, for the alignment step and V- and J-gene annotation we used a custom minimalistic gene set library with only one allelic variant per V- and J-gene, derived from a custom public genome reference to match the one used for the long-read assembly (Rodriguez et al. 2020) MiXCR built-in capabilities for downsampling capabilities were utilized to retrieve clonotypes derived from a fraction of reads. To test the sensitivity of the tools we downsampled the dataset to 500 000, 100 000, 50 000 and 10 000 sequencing reads aligned to the V-, D- and J-region reference library (command example for downsampling to 10 000 aligned reads: ‘mixcr downsample --downsampling count-reads-fixed-100000 input_file’). For inferring the alleles with the comparison tool, TIgGER (Gadala-Maria et al., 2015) we utilized function MiXCR function ‘exportAirr’ which exports the clonotype sets in format suggested by Adaptive Immune Receptor Repertoire Community (Rubelt et al. 2017). For the full dataset and downsampled ones the algorithm described in the previous section as well as TIgGER v1.0.1 were used for allelic inference and genotyping. The resulting sets of allele sequences were exported from both tools in fasta format and exactly matched with the sequences of the alleles present in the genotype of the donor and the number of matches was determined. Importantly, due to the fact that IGH repertoire sequencing data utilized for comparison was derived using 5’RACE-based technology, inference could be performed only for expressed V- and J-gene alleles. Thus we excluded those alleles from comparison which had less than 10

clonotypes assigned to it, when utilizing the same MiXCR v4.2.0 upstream pipeline, but with allele-resolved V- and J-gene reference library (IGHV4-28*01, IGHV3-38*02, IGHV1-58*02, IGHV4-4*01).

Chapter 4. Results

4.1 IGH repertoire sequencing statistics and analysis depth

We collected peripheral blood from six healthy donors at three time points, where the second sample was collected one month after the first, and the third was collected 11 months after that (**Fig. 4; Table 2**). These samples were subjected to fluorescence-activated cell sorting to isolate memory B cells (Bmem: $CD19^{+}$ $CD20^{+}$ $CD27^{+}$), plasmablasts (PBL: $CD19^{low/+}$ $CD20^{-}$ $CD27^{high}$ $CD138^{-}$) and plasma cells (PL: $CD19^{low/+}$ $CD20^{-}$ $CD27^{high}$ $CD138^{+}$) (**Fig. 5A**). Most of the cell samples were collected and processed in two independent replicates (**Supplementary Data SD1**). For each cell sample, we obtained IGH clonal repertoires using a 5'-RACE-based protocol, which allows unbiased amplification of full-length IGH variable domain cDNA while preserving isotype information, with subsequent unique molecular identifier (UMI)-based sequencing data normalization and error correction (Turchaninova et al. 2016; Shugay et al. 2014). From a total of 83 cell samples, we obtained 1.06×10^{7} unique IGH cDNA molecules, each covered by at least three sequencing reads, representing 8.4×10^{5} unique IGH clonotypes (**Supplementary Data SD1**). An IGH clonotype was defined as a unique nucleotide sequence spanning from the beginning of IGH V gene framework 1 to the 5' end of the C segment, sufficient to determine isotype. The number of unique clonotypes (*i.e.,* species richness) depended on the number of cells per sample (**Fig. 5B**),

even after data normalization by sampling an equal number of unique IGH cDNA sequences. To characterize the number of distinct IGH clonotypes in each cell subset, we selected the samples with the most common number of sorted cells for each sample set. The median number of clonotypes was 20,072 (14,572–32,806, n = 14) per 5 x 10^4 memory B cells, 628 (528–981, n = 8) per 1 x 10^3 plasmablasts, and 800 (623–1,183, n = 9) per 1 x 10^3 plasma cells. Rarefaction analysis in the Bmem subpopulation (**Fig. 5B,** left) revealed an asymptotic increase of species richness that did not reach a plateau, indicating that the averaged species richness can only serve as a lower limit of sample diversity estimation. For all samples of PBL and PL subpopulations, species richness curves plateaued, meaning that we had reached sufficient sequencing depth to evaluate the clonal diversity of the sorted cell samples (**Fig. 5B** center and right).

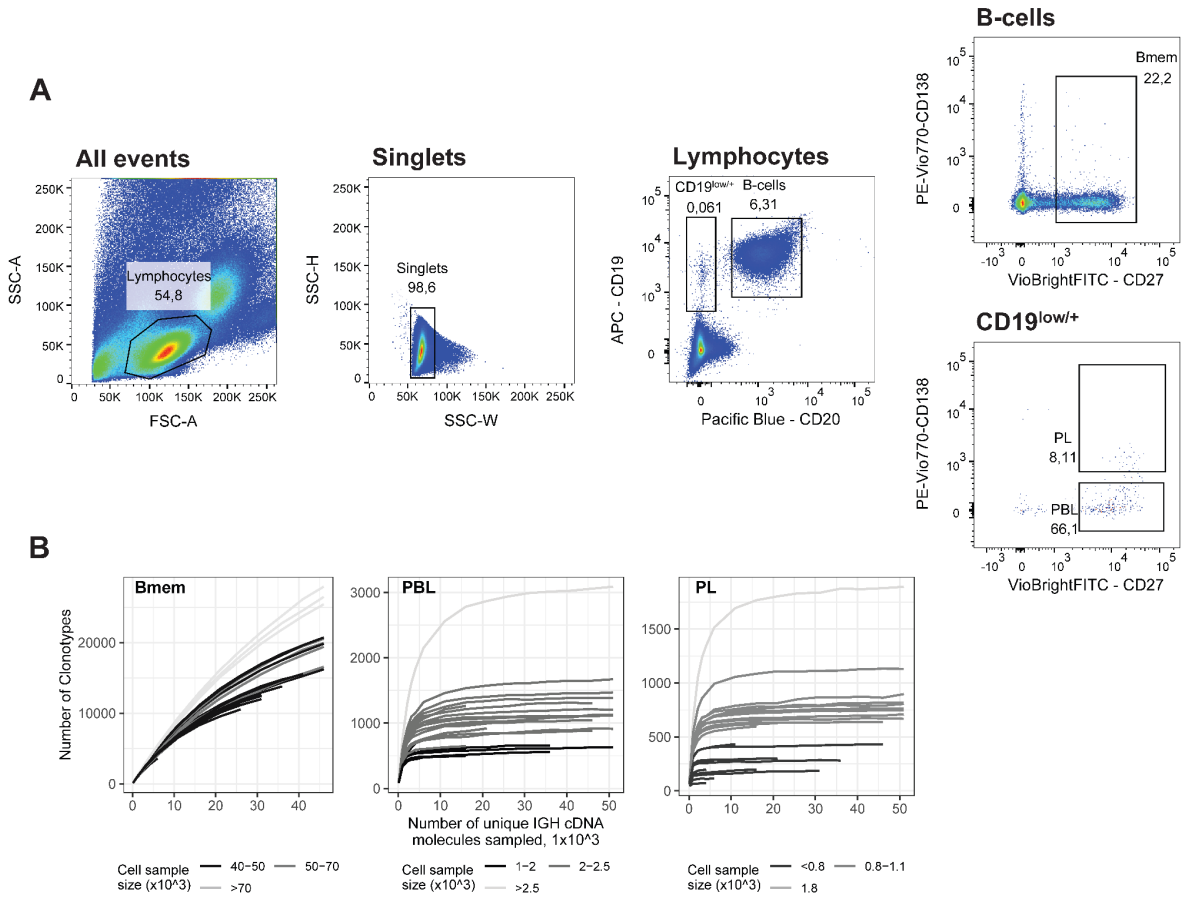


Figure 5. A: FACS gating strategy and the frequencies of studied cell subsets for representative peripheral blood sample (donor IZ time point T3): Memory B-cells (Bmem: $CD19^{+}$ $CD20^{+}$ $CD27^{+}$), plasmablasts (PBL: $CD19^{low/+}$ $CD20^{-}$ $CD27^{high}$ $CD138^{-}$) and plasma cells (PL: $CD19^{low/+}$ $CD20^{-}$ $CD27^{high}$ $CD138^{+}$); **B**: Rarefaction curves by IGH cDNA molecules. From each repertoire a defined number of unique IGH cDNA molecules was sampled and the number of unique IGH clonotypes was determined. Each line represents a single sample. Samples with representative cell number ($5x10^4$ Bmem, $1x10^3$ PBL, $1x10^3$ PL) are shown in black, samples of other sizes - in grey.

4.2 B cell subsets display both divergent and similar characteristics in their IGH repertoires

First, we aimed to characterize features of the IGH repertoires of the Bmem, PBL, and PL subset based on several key properties: usage of germline encoded IGHV segments, clonal distribution by isotypes, rate of SHM in CDR1-2 and FWR1-3, and features of the hypervariable CDR3 region. The proportion of overall clonal diversity occupied by the five major IGH isotypes was strikingly different between Bmem cells and antibody-secreting cells (ASCs; *i.e.,* PBL and PL). IgM represented more than half of the repertoire in Bmem, while IgA was dominant in PBL and PL (**Fig. 6**, **Supplementary Data SD2**). The second most prevalent isotype in ASCs was IgG, which was also less abundant in Bmem compared to IgA. IgD represented a substantial part of the Bmem clonal repertoire, while < 1% clonotypes of ASCs expressed IgD. The proportion of each isotype varied between donors and time points, but IgM and IgA or IgA and IgG consistently remained the most abundant isotypes in Bmem cells or ASCs, respectively (**Fig. 6A, 6B, Supplementary Data SD2**). In all studied subsets, the isotype distribution in terms of number of unique clonotypes roughly mirrored the isotype distribution based on the number of IGH cDNA molecules, indicating absence of large clonal expansions or differences in IGH expression level distorting abundance of isotypes. This could not be determined by sequencing of bulk PBMCs, as higher levels of IGH expression by ASCs can change the isotype proportions and thereby bias the quantification of clonotype abundance (**Supplementary Fig. 6A**).

The obtained IGH isotype distributions based on unique clonotypes roughly correspond to the distribution of IGH isotypes typically detected by flow cytometry of the same subsets (Perez-Andres et al. 2010).

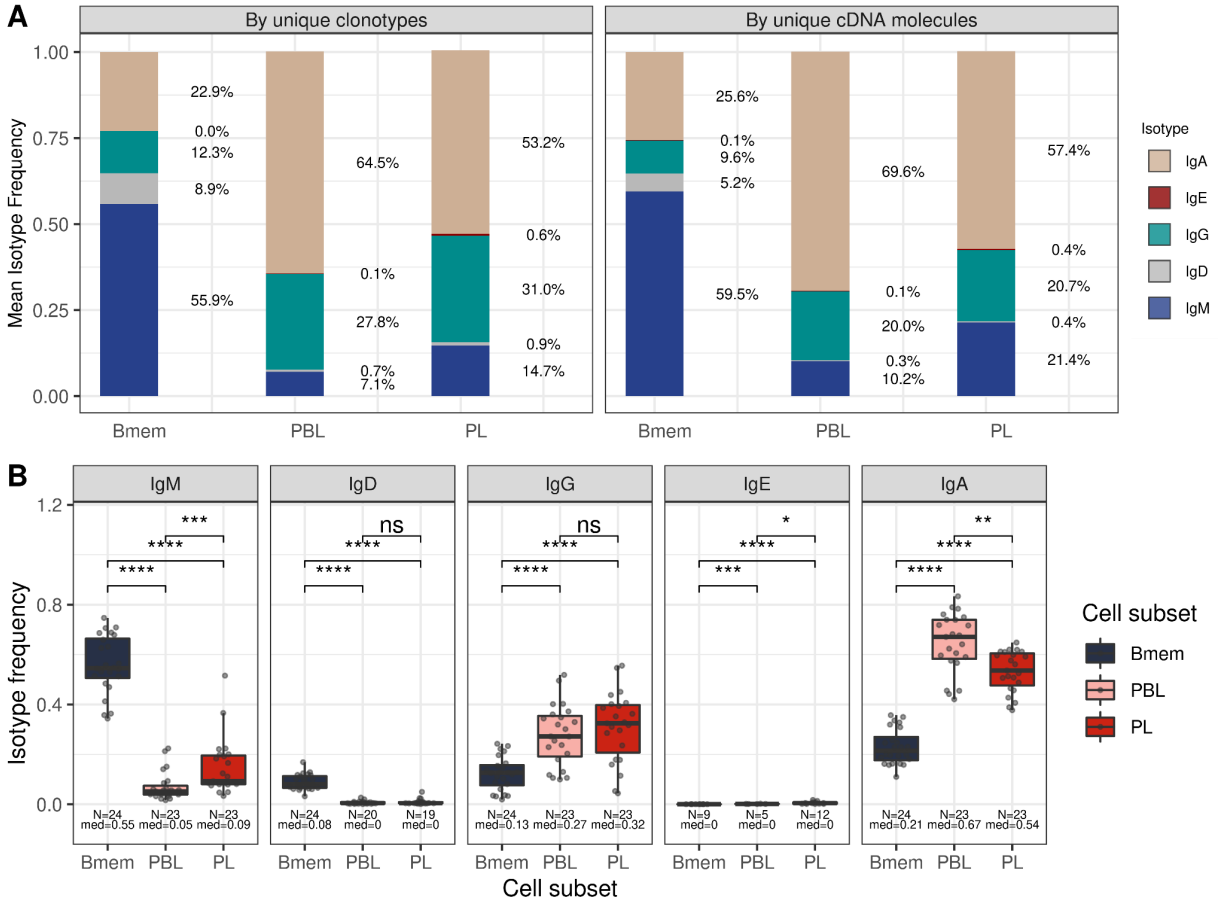


Figure 6. Isotype frequencies in studied subsets. **A:** Isotypes frequencies in studied cell subsets as well as in bulk PBMCs averaged across all obtained samples. Left panel - isotype frequencies calculated as a number of IGH clonotypes (full-length unique nucleotide sequence) with specific isotype divided by total number of clonotypes. Right panel - isotype frequencies calculated as a number of cDNA molecules in isotype divided by total number of cDNA molecules. **B**: Isotype frequencies for by unique clonotypes, comparison of cell subsets. The numbers at the bottom of the plots represent the number of samples in the corresponding group, and the median measurements from each cell type. Comparisons between subsets were performed with two-sided Mann-Whitney U test. * = $p \leq 0.05$, ** = $p \leq 0.01$, *** = $p \leq$ 10-3, **** = $p \leq$ 10-4

The level of SHM was on average significantly higher in ASC subsets, reflecting that PBLs and PLs are enriched for clones that have undergone affinity maturation (**Fig. 7**). The switched isotypes (IgG, IgA) had higher average levels of SHMs in the Bmem subset compared with IgM and IgD isotypes. Interestingly, the SHM level of IgD clonotypes in ASC subsets was significantly higher compared with Bmem. The average number of SHMs for IgE clonotypes did not differ significantly between cell subsets, but was significantly higher compared to the level of SHM detected for IgM and IgD clonotypes in Bmem (**Fig. 7**). Of note, the rate of SHM in PBLs was higher than that in PLs in clonotypes from the three most abundant isotypes (*i.e.,* IgM, IgA and IgG).

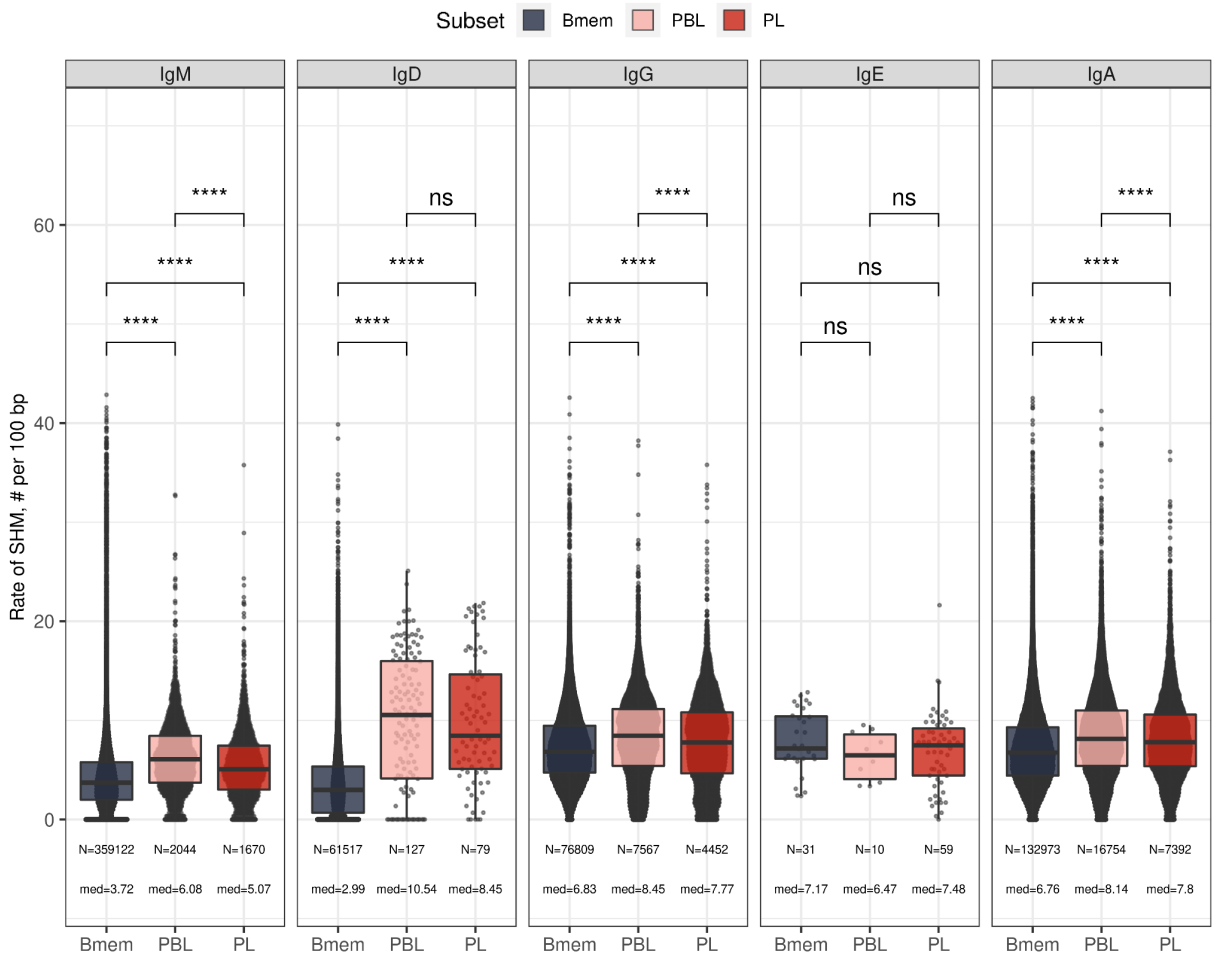


Figure 7. Distribution of the number of somatic hypermutations identified per 100 bp length of IGHV segment for clonotypes within each particular isotype. The numbers at the bottom of the plots represent the number of clonotypes in the corresponding group, pooled from all donors, and the median measurements from each cell type. Comparisons between subsets were performed with two-sided Mann-Whitney U test.
* = $p \leq 0.05$, ** = $p \leq 0.01$, *** = $p \leq 10\text{-}3$, **** = $p \leq 10\text{-}4$.

We also compared the distributions of the lengths of the hypervariable CDR3 region between IGH clonotypes in different cell subsets. PBLs had significantly longer CDR3 regions compared to Bmem cells on average in every isotype except for IgE (**Fig. 8**). Differences for IgE isotype, however, could not be reliably assessed due to low number of clonotypes. Of note, the average CDR3 length

in PL clonotypes was significantly higher compared to Bmem for IgA and IgD, but not for the other isotypes.

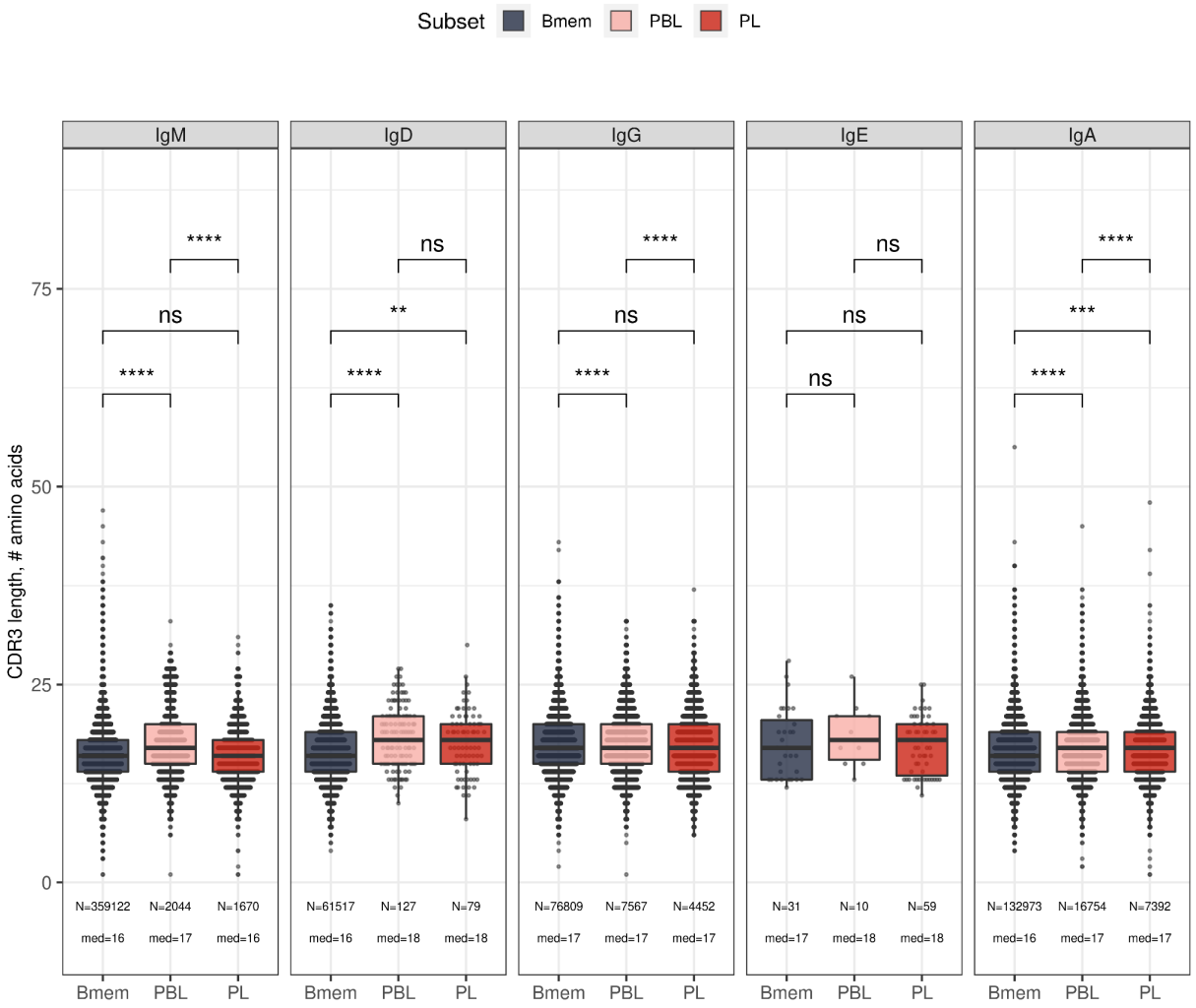


Figure 8. Distribution of CDR3 length of clonotypes in each cell subset by isotype; the numbers at the bottom of the plots represent the number of clonotypes in the corresponding group, pooled from all donors, and the median measurements from each cell type. Comparisons between subsets were performed with two-sided Mann-Whitney U test. * = p ≤ 0.05, ** = p ≤ 0.01, *** = p ≤ 10-3, **** = p ≤ 10-4

IGHV gene segment usage was roughly similar between Bmem, PBL, and PL cells from all donors, indicating generally equal probabilities of memory-to-ASC conversion for B cells carrying BCRs encoded by distinct gene segments (**Fig. 9,10**). This distribution differed significantly between the studied cell subsets and naive B cells (based on data from Gidoni et al. 2019). We observed high concordance in terms of under- or overrepresentation of specific IGHV gene segments in repertoires of all antigen-experienced B cell subsets compared to naive B cells; Pearson correlation coefficients for the fold-change of IGHV gene segment usage frequencies were 0.95 for Bmem and PBL, 0.96 for Bmem and PL, and 0.98 for PBL and PL ($p < 0.01$ for all pairs). Moreover, IGHV gene segment under- or overrepresentation clearly depended on the given gene sequence. We clustered IGHV genes based on their sequence similarity, and observed that most IGHV segments in each of the four major clusters behaved concordantly with other segments in that cluster (**Fig. 9**). This effect was also observed at the level of individual repertoires (**Fig. 10**) with discrepancies that could probably be attributed to genetic polymorphism of the IGH loci of particular donors.

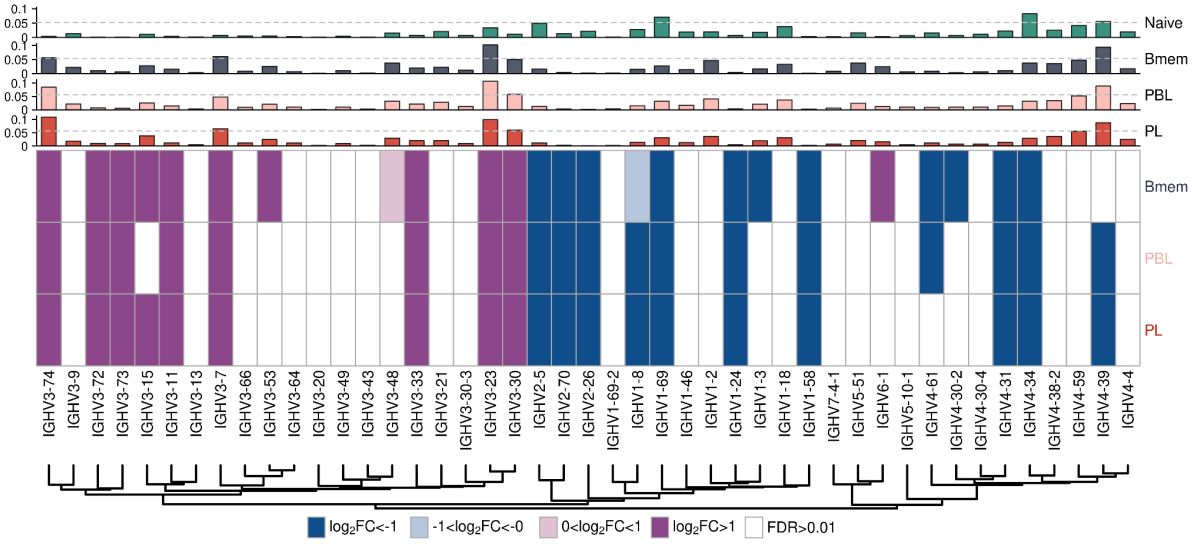


Figure 9. IGHV gene frequencies in studied cell subsets. Distributions of average IGHV gene frequencies based on the number of clonotypes in naive B cells (data from Gidoni et al. 2019), Bmem, PBL, and PL repertoires are shown at the top. Colored squares on heatmap indicate significantly different (false discovery rate, FDR < 0.01) frequencies for IGHV gene segments in corresponding B cell subsets compared to naive B cell repertoires. Color intensity reflects the magnitude of the difference (FC = fold change). Only V genes represented by more than two clonotypes on average are shown, data normalization was performed using trimmed mean of M values method (Robinson, 2010a). IGHV gene segments are clustered based on the similarity of their amino acid sequence, as indicated by the dendrogram at the bottom.

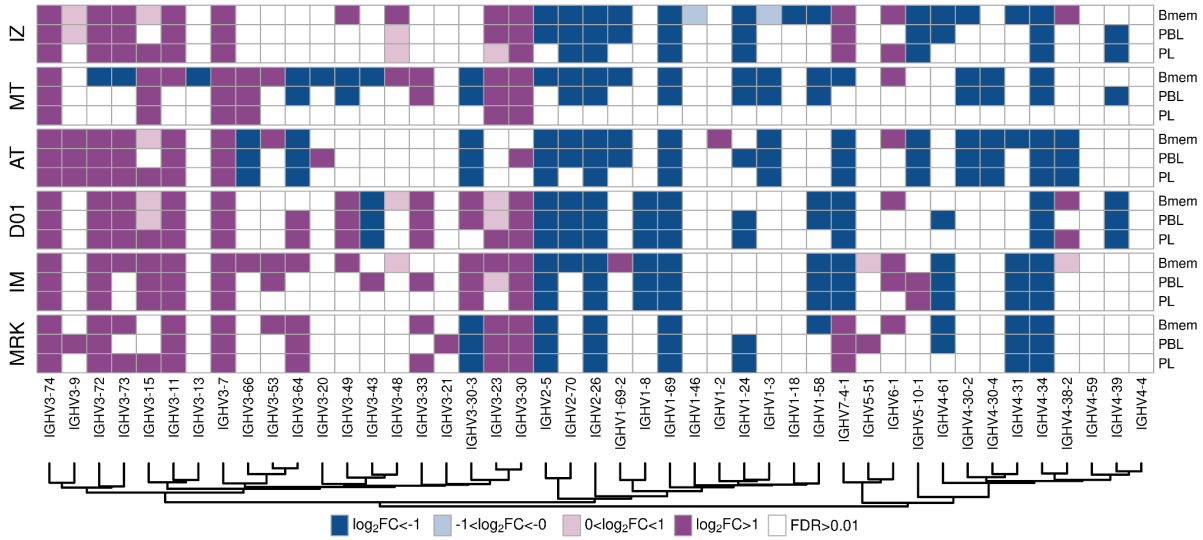


Figure 10. Heatmap of IGHV frequencies enrichment for individual donors. Colored squares on heatmap indicate significantly different (false discovery rate less than 0.01) IGHV-gene segments by their frequency in corresponding B-cell subsets than in publicly available naive B-cell repertoires (Gidoni et al. 2019). Color intensity reflects magnitude difference (FC=fold change). Only V-genes which were represented by more than 2 clonotypes on average are shown. IGHV-gene segments are ordered by the similarity of their amino acid sequence, as indicated by the amino acid similarity dendrogram at the bottom.

These observations highlight the differences in general characteristics of IGH repertoire between the Bmem and ASC subsets, and demonstrate similarity of IGHV gene usage that differs from that in naive B cells.

4.3 Memory B cell repertoires are stable over time and contain a large number of public clonotypes

We further studied the similarity of IGH clonal repertoires of B cell subsets across time points and between individuals, evaluating

repertoire stability (*i.e.,* distance between different time points) and degree of individuality (*i.e.,* distance between repertoires from different donors). We evaluated repertoire similarity at two levels of IGH sequence identity: frequency of clonotypes with identical nucleotide sequence-defined variable regions (FR1–4), and number of clonotypes with identical CDR3 amino acid sequences, IGHV gene segments, and isotypes.

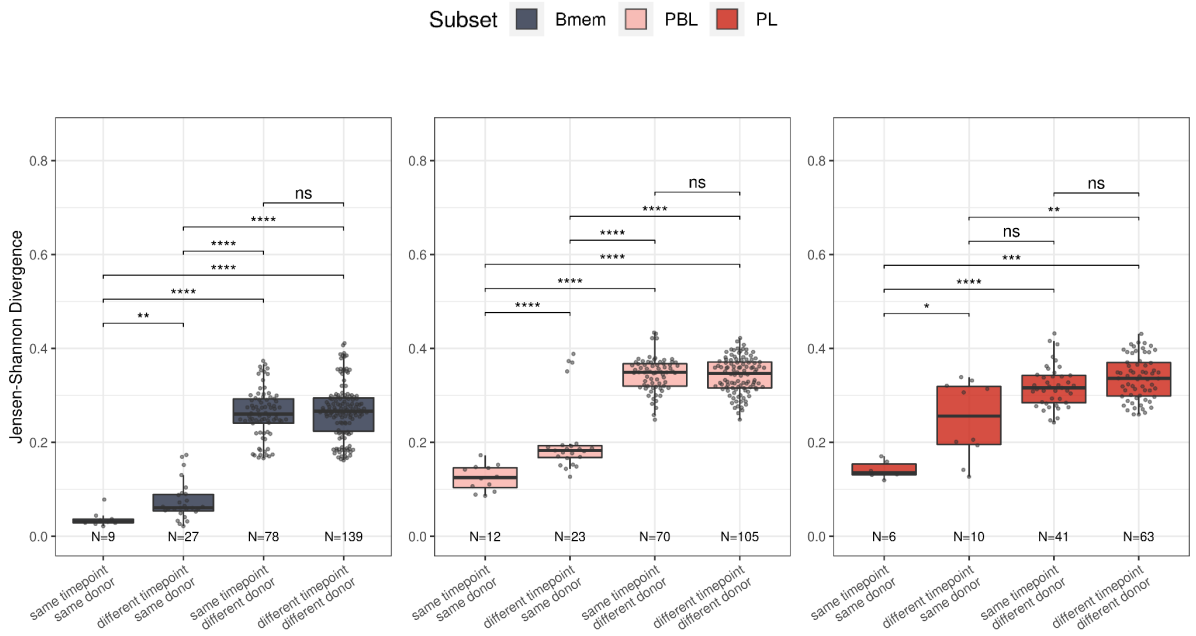


Figure 11. Bmem, PBL, and PL IGH repertoire stability over time by IGHV gene usage. Distance between repertoires obtained at different time points from the same or different donors as calculated by Jensen-Shannon divergence index for IGHV gene frequency distribution. either in repertoires from only one donor (private) or in at least two donors (public). For normalization in Bmem repertoires of 14 000 most abundant clonotypes were considered, in PBL - 600, in PL - 200. Each dot represents a pair of repertoires of the corresponding type; N indicates the number of pairs of repertoires in the group. Comparisons were performed with two-sided Mann-Whitney U test. * = $p \leq 0.05$, ** = $p \leq 0.01$, *** = $p \leq 10^{-3}$, **** = $p \leq 10^{-4}$.

Both metrics showed significantly higher inter-individual differences compared to the divergence of repertoires derived from the same donor, reflecting the fact that IGH repertoires of Bmem, PBL, and PL subsets are private to a large degree (**Fig. 11, 12**). We observed identical clonotypes in the repertoires of PBL and PL collected at different time points, whereas the repertoire similarity was much lower compared to that between replicate samples, reflecting the transient nature of PBL and PL populations in peripheral blood. Notably, we observed lower clonal overlap in PBL and PL for more distant time points (separated by 11 or 12 months) than those that are closer together (1 month) (**Fig. 13**).

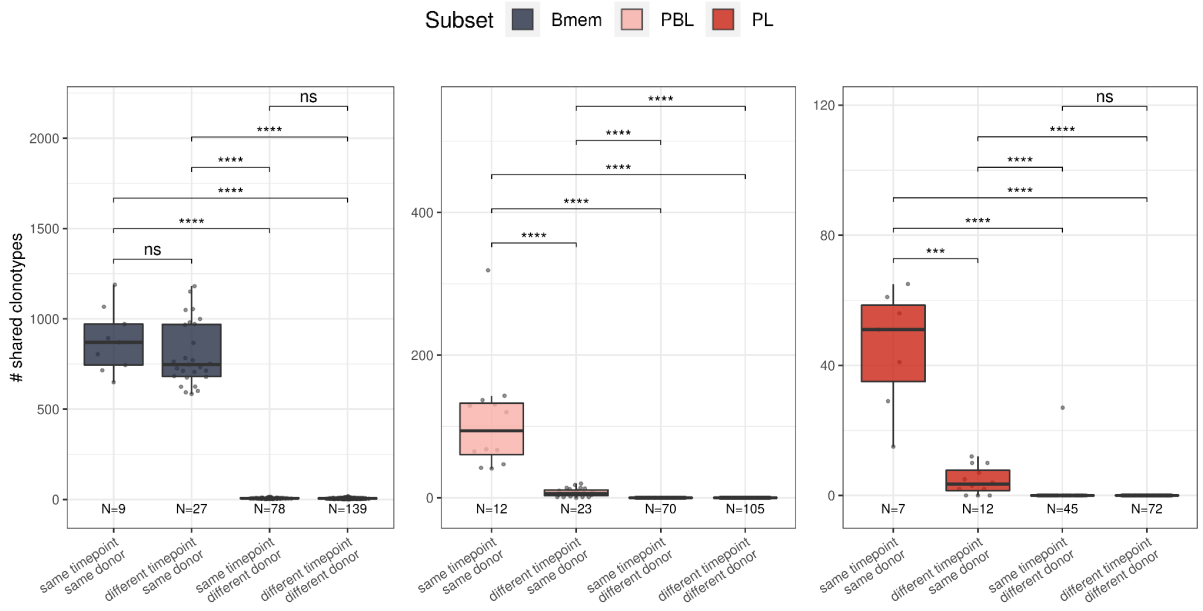


Figure 12. Bmem, PBL, and PL IGH repertoire stability over time by number of shared clonotypes. Number of shared clonotypes between pairs of repertoires from the same or different donors and time points. For data normalization, we assessed the most abundant 14,000 Bmem, 600 PBL, and 200 PL clonotypes. Each dot represents a pair of repertoires of the corresponding type; N indicates the number of pairs of repertoires in the group. Comparisons were performed with two-sided Mann-Whitney U test. * = $p \leq 0.05$, ** = $p \leq 0.01$, *** = $p \leq 10^{-3}$, **** = $p \leq 10^{-4}$.

The dissimilarity between samples collected on the same day versus 1 month or even 1 year later was much lower for Bmem, demonstrating the high stability of the clonal repertoire and long-term persistence of IGH clonotypes in these cells (**Fig. 13**).

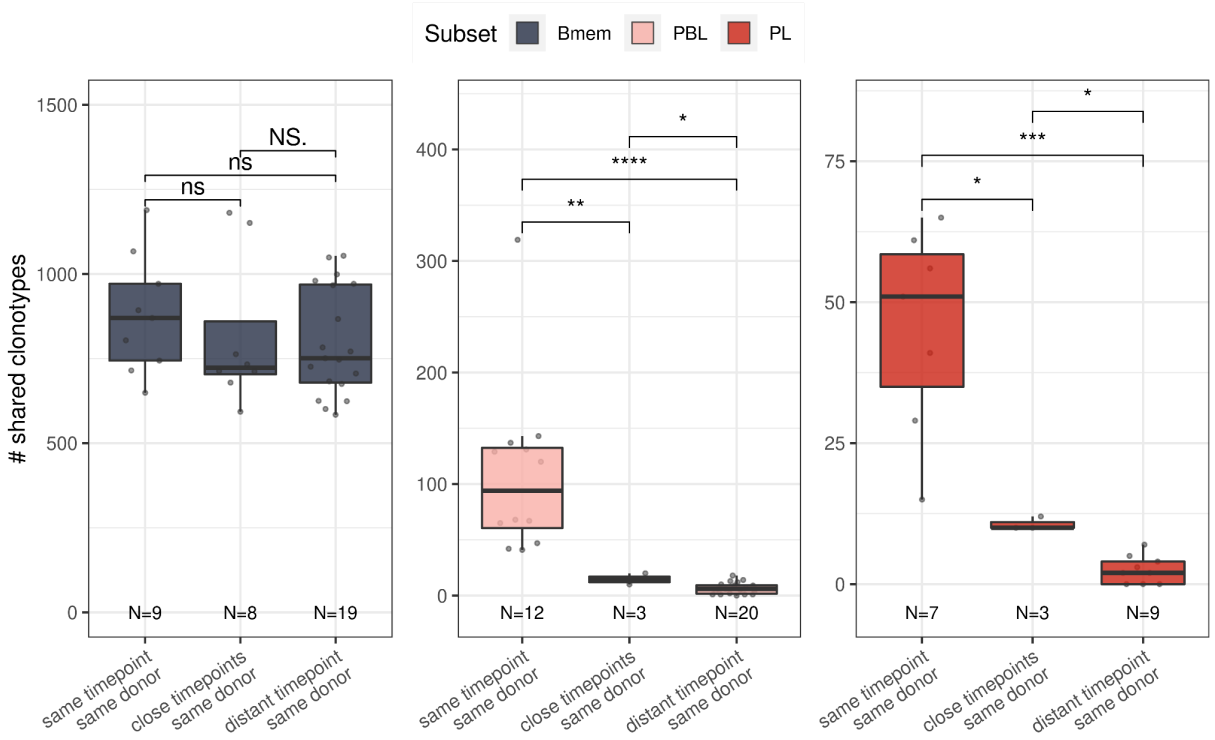


Figure 13. IGH repertoire similarity within subpopulations of B-cell lineage. Number of shared clonotypes between pairs of repertoires from the same donor and same or different time points. “Same time point” represents replicate samples derived from the same blood draw, “close time points” - samples were collected with approx. 1 month interval, and “distant time points” - samples were collected with approx. 1 year interval. either in repertoires from only one donor (private) or in at least two donors (public). For normalization in Bmem repertoires of 14 000 most abundant clonotypes were considered, in PBL - 600, in PL - 200. Each dot represents a pair of repertoires of the corresponding type; N indicates the number of pairs of repertoires in the group. Comparisons were performed with two-sided Mann-Whitney U test. * = $p \leq 0.05$, ** = $p \leq 0.01$, *** = $p \leq 10^{-3}$, **** = $p \leq 10^{-4}$.

To better describe the inter-individual IGH repertoire convergence, we analyzed the number of IGH amino acid clonotypes shared between different donors (*i.e.,* public clonotypes) among 5000 most expanded clonotypes in each Bmem repertoire, assuming that functional convergence could be detected amongst the most abundant clonotypes due to clonal expansions in response to common pathogens. Indeed, the average number of shared clonotypes in Bmem was significantly higher between fractions of the most abundant clonotypes compared to randomly-sampled clonotypes (**Fig. 14**), as well as when compared to the most abundant clonotypes shared by two naive repertoires (from Gidoni et al. 2019) or to pre-immune IGH repertoires obtained by *in silico* generation using OLGA software (Sethna et al. 2019) (**Fig. 14**).

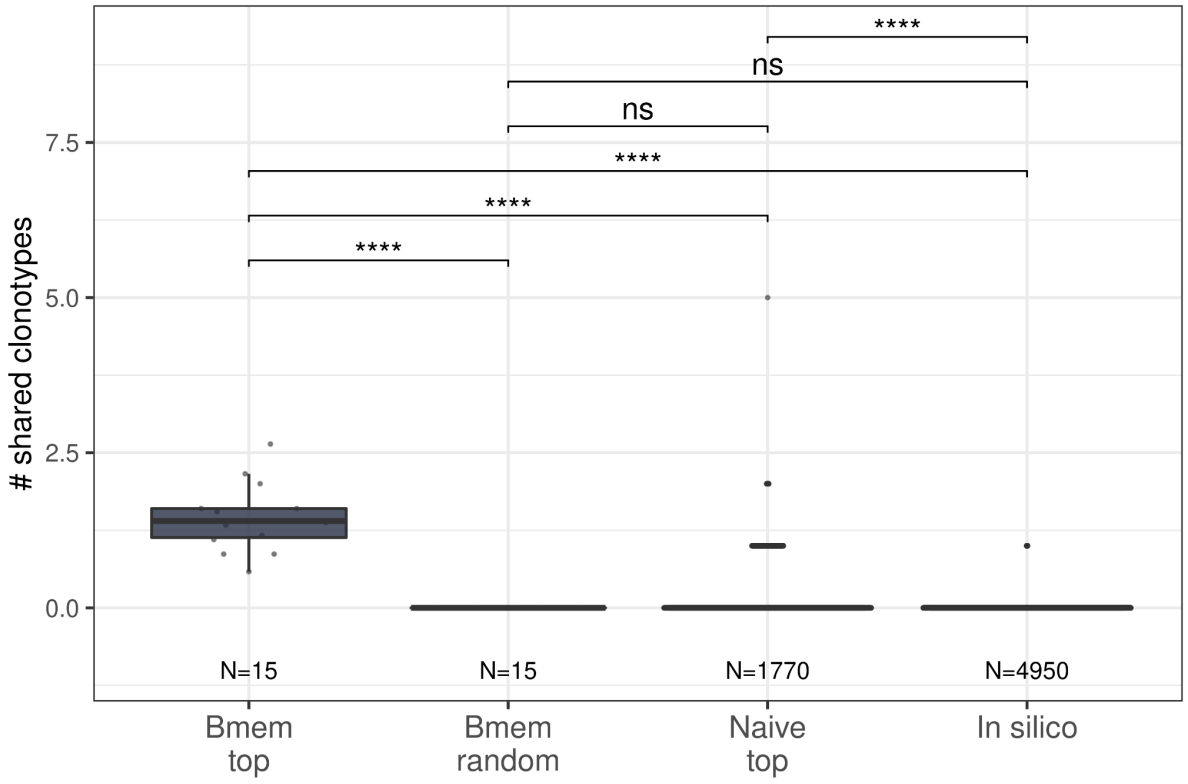


Figure 14. Degree of repertoire sharing between unrelated repertoires. The average number of shared clonotypes between repertoires from pairs of unrelated donors for the most abundant Bmem clonotypes, randomly-selected Bmem clonotypes, most abundant clonotypes from naive repertoires of unrelated donors (from Gidoni et al. 2019), or from synthetic repertoires generated with OLGA software; each repertoire in comparison was represented by a fixed number of clonotypes (5000), either most abundant, randomly selected or generated where indicated. Each dot represents a pair of repertoires of the corresponding type; N indicates the number of pairs of repertoires in the group. Comparisons were performed with two-sided Mann-Whitney U test. * = $p \leq 0.05$, ** = $p \leq 0.01$, *** = $p \leq 10^{-3}$, **** = $p \leq 10^{-4}$.

Public clonotypes were also hypermutated, although the rate of SHM was slightly lower compared to that in clonotypes specific to one donor (private) (**Fig. 15**). These observations indicate functional convergence in Bmem repertoires, which is presumably driven by exposure to common pathogens. Of note, the extent of clonal overlap was significantly higher between naive repertoires than for *in silico*-generated repertoires, indicating functional convergence even in pre-immune repertoires.

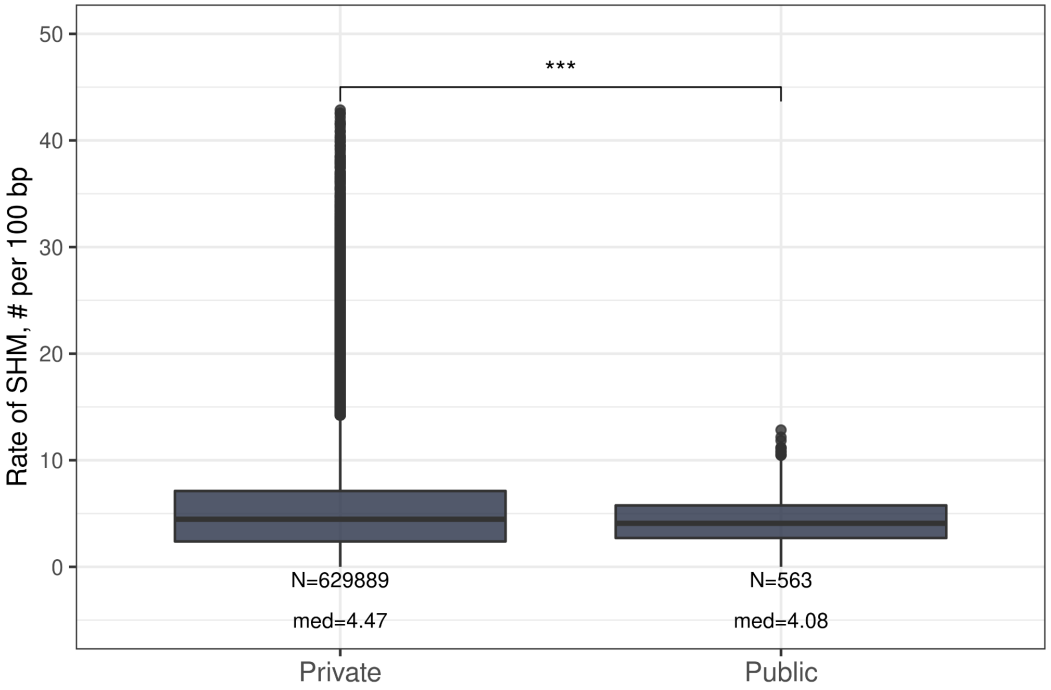


Figure 15. Rate of SHM in private and public clonotypes. Distribution of the number of somatic hypermutations identified per 100 bp length of IGHV-segment for clonotypes detected either in repertoires from only one donor (private) or in at least two donors (public). For normalization in Bmem repertoires of 14 000 most abundant clonotypes were considered, in PBL - 600, in PL - 300. Each dot in each plot represents a pair of repertoires of corresponding type, numbers below each box indicate the number of pairs of repertoires in the group. Comparisons were performed with Mann-Whitney test, notation of the level of significance is the following: * - $p \leqslant 0.05$, ** - $p \leqslant 0.01$, *** - $p \leqslant 10^{-3}$, **** - $p \leqslant 10^{-4}$.

Furthermore, the distance between V segment usage distributions in Bmem repertoires was not significantly different compared to that in naive B cells repertoires. That indicates that the higher clonotype sharing seen in Bmem cells cannot be attributed to lower diversity in IGHV germline usage (**Fig. 16**). The same analysis in PBL and PL subpopulations for the 600 and 200 most abundant clonotypes respectively yielded no shared clonotypes between repertoires of different donors, demonstrating no detectable convergence at this sampling depth.

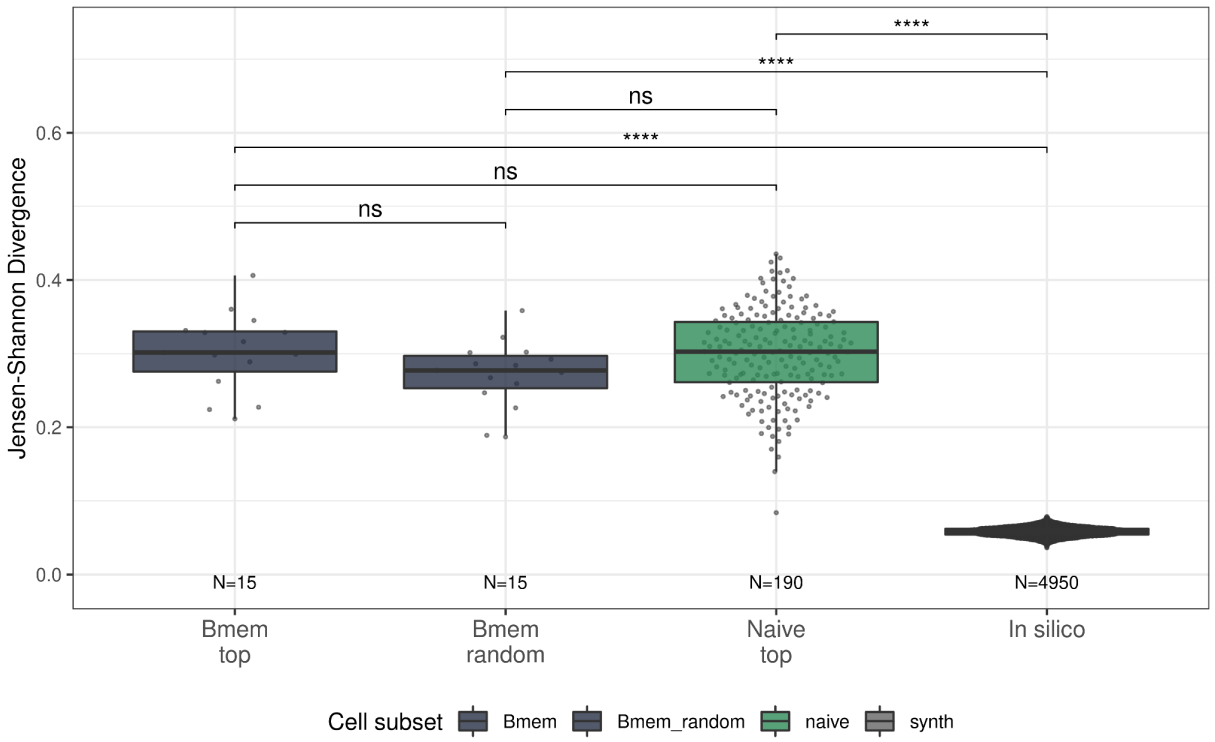


Figure 16. Similarity of repertoires between unrelated repertoires by IGHV gene usage. Inter-individual distance between distributions of V genes in repertoires, calculated as Jensen-Shannon divergence indices for the pairs of repertoires depicted in Fig 15. Each dot represents a pair of repertoires of the corresponding type; N indicates the number of pairs of repertoires in the group. Comparisons were performed with two-sided Mann-Whitney U test. * = $p \leq 0.05$, ** = $p \leq 0.01$, *** = $p \leq$ 10-3, **** = $p \leq$ 10-4.

Finally, we found that public clonotypes were more likely to be detected than private ones in samples collected at different time points (**Fig. 17**), again suggesting persistent memory to common antigens. Thus the results demonstrate the level of stability of memory B-cell receptor repertoires and extent of clonal sharing in repertoires of unrelated donors, which might be attributed to exposure to common antigens.

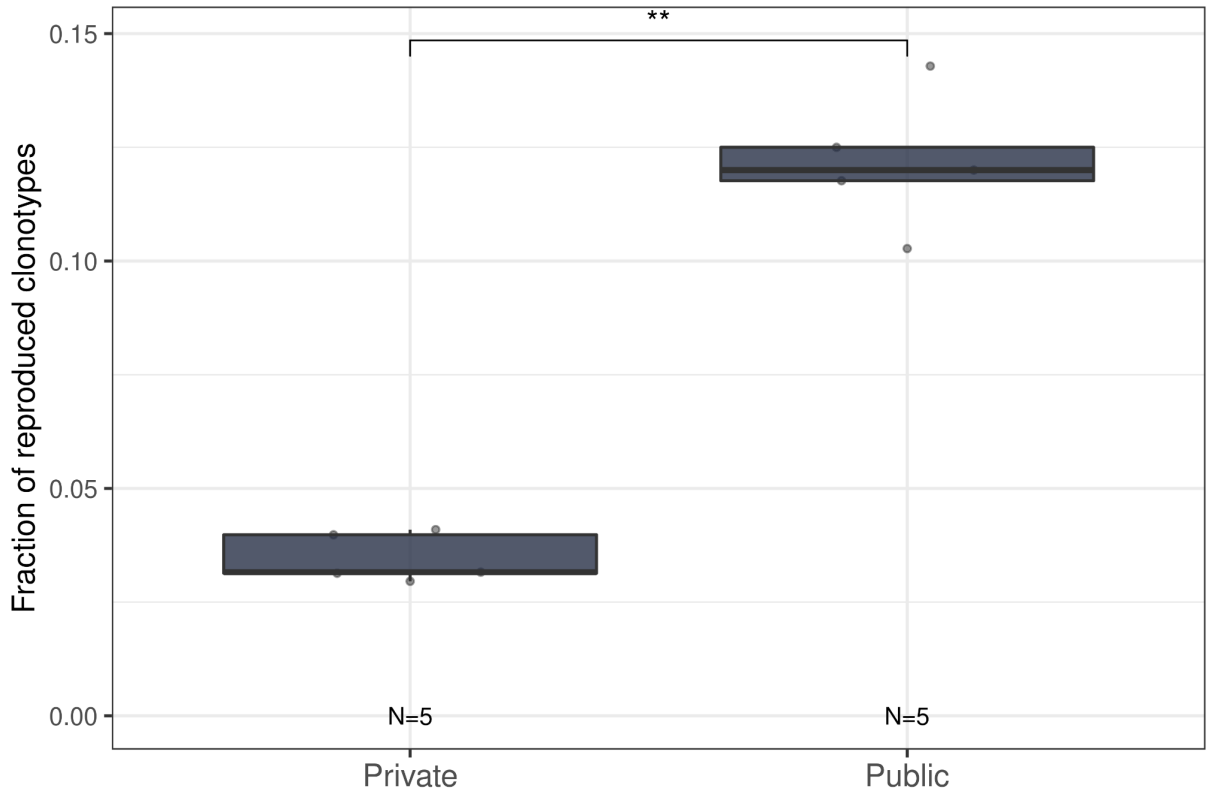


Figure 17. Persistence of private and public clonotypes. Fraction of persistent clonotypes detected at more than one time point among clonotypes detected in repertoires from only one donor (private) or in at least two donors (public). Each dot represents the fraction of persistent clonotypes from one donor. N indicates the number of repertoires in the group. Comparisons were performed with two-sided Mann-Whitney U test. * = $p \leq 0.05$, ** = $p \leq 0.01$, *** = $p \leq 10^{-3}$, **** = $p \leq 10^{-4}$.

4.4 Novel approach for V- and J-gene allele variants inference and genotyping

Inference of V- and J-gene allelic variants from a typical repSeq data remains a challenge; currently available tools lack flexibility and have crucial limitations related to requirements for sampling and sequencing depth. Thus we developed a novel algorithm which allows V- and J-gene allele inference from immune receptor repertoire sequencing data containing both non- and hypermutated receptor sequences using a minimalistic starting with a reference library with only one allelic variant per each gene.

First steps of the algorithm are common for processing repertoire sequencing data and include sequence alignment and annotation with the initial V-, D- and J-gene segment reference library, followed by assembly of clonotypes. By the end of upstream processing each clonotype has an annotation of a V- and J-gene segment and most importantly a defined set of mutations which differentiates that clonotype sequence from the reference sequence of the corresponding V- or J-gene segment.

Then the algorithm separately infers alleles for V- and J-genes. Below we provide the description for V-gene allele inference:

1) Clonotypes are grouped by the V-gene segments.
2) For each mutation within the group, including insertions and deletions, we define a set of clonotypes which contain this mutation.

3) The mutations are filtered based on the diversity of the combinations of J-genes and CDR3-lengths of clonotypes containing that mutation. The mutations that do not have sufficient diversity of J-genes and CDR3-lengths are then removed from each of the clonotype's mutation sets.

4) Then clonotypes are further grouped by these "thinned" mutation sets, including "empty" mutation sets. The diversity of J-genes and CDR3 lengths combinations within each of these groups is determined along with the number of clonotypes containing no mutations in J-gene after filtering at step 3. Mutation sets are then filtered by thresholds of these two parameters, resulting in a list of allele candidates.

5) Clonotypes are then assigned to the closest allele candidates. Candidates are then sorted by the score which represents the linear combination of the number of different J-genes and the number of different CDR3 lengths in clonotypes assigned to these candidates.

6) Candidates with the score not lower than 0.35 of the maximum score are then selected for the subject-specific gene set library.

This stepwise approach based sequential filtering first on the level of individual mutation and then on the level of mutation sets dramatically reduces noise introduced by SHM and sequencing and PCR-errors. The threshold of 0.35 for the final alleles filtering was initially chosen from a theoretical consideration of possible distributions of expressed alleles for a V-gene allowing the presence

of three allelic variants due to possible V-gene duplications. This was then corroborated by examining empirical score distributions for alleles in repertoire sequencing of IGH repertoire of a healthy donor with known genotype; in this case the donor was different from the one in the benchmarking of the algorithm.

This approach allows both to infer novel (undocumented) V- and J-gene alleles and to perform genotyping with high sensitivity and precision.

4.5 Benchmarking of the for V- and J-gene allele variants inference and genotyping

To assess the performance of the developed algorithm we utilized publicly available dataset (Rodriguez et al. 2022) containing both AIRR-seq data and highly reliable genotyping data of the IGH locus reconstructed using Pacific Biosciences HiFi long-read sequencing. The AIRR-seq data was represented by an IGH repertoire of unsorted PBMCs from a healthy individual and therefore we limited our comparison to tools which were suitable for this type of data, i.e. excluding IgDiscover, which requires unmutated datasets for correct performance. As mentioned above ImPre is no longer supported, while the run time of Partis exceeded two weeks on a high-performance cluster which appears to be prohibitive for real-world applications. Thus, for the comparison we utilized TIgGER, which is also the most widely cited tool for the task. Upstream analysis, including sequence alignment to reference V- and J-gene library and defining the full-length clonotypes, was

identical for both tools (see Methods section). For the alignment step and V- and J-gene annotation we used a custom minimalistic gene set library with only one allelic variant per V- and J-gene, derived from a custom public genome reference to match the one used for the long-read assembly (Rodriguez et al. 2020). To test the sensitivity of the tools we also downsampled the dataset to 500 000, 100 000, 50 000 and 10 000 sequencing reads aligned to the VDJ-region. Then we performed allele variants inference and genotyping with both tools and compared the resulting individualized V- and J-gene libraries with the accurate genotype inferred utilizing NGS long-read sequencing (Rodriguez et al. 2022).

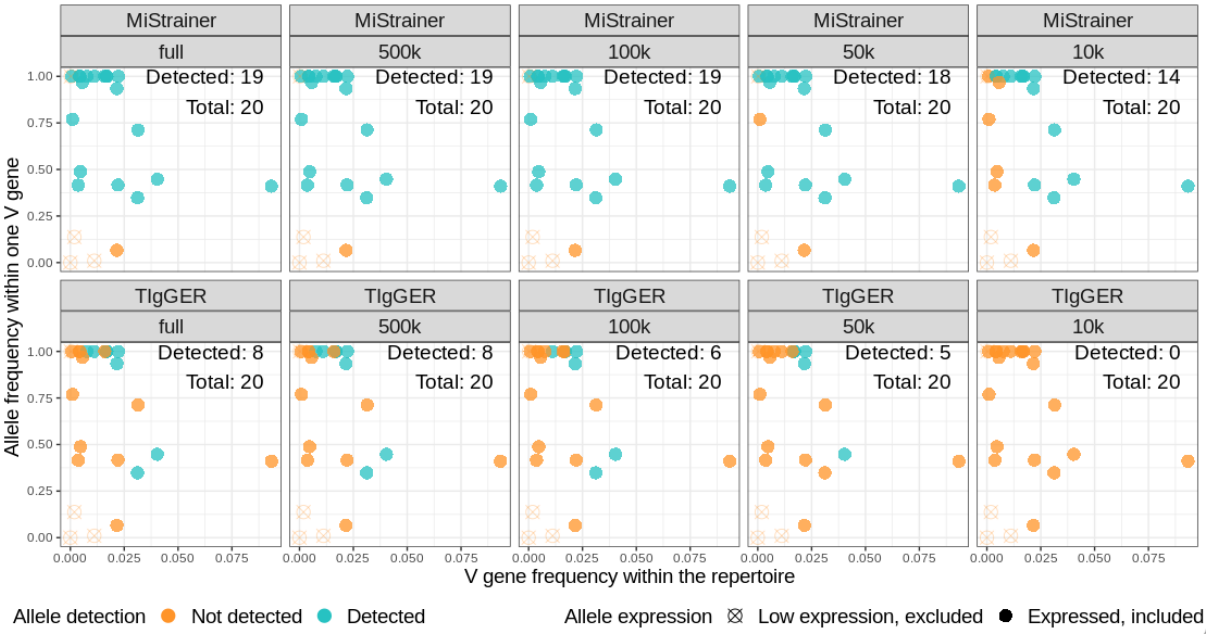


Figure 18. Detection of the allele variants of V-genes depending on expression of the V genes and allelic imbalance. Each dot represents a V-gene allele present in the donor's genotype confirmed by long-read sequencing. The upper row represents detection by the original algorithm, the lower - allele detection by golden standard tool TIgGer. Columns represent different depths of downsampling by number of aligned reads, from right to left: full set (1 071 532), 500 000, 100 000, 50 000, 10 000. V gene, and allele frequencies for each facet were calculated using the full set of reads and allele-resolved V- and J-gene reference library. Alleles, excluded due to low expression (<10 clonotypes), are represented as empty crossed dots.

The AIRR-seq data which we used for the comparison was 5'RACE derived therefore our analysis was limited only to expressed alleles, so we excluded those alleles which had less than 10 IGH clonotypes assigned to it utilizing allele-resolved reference (see Methods section). The allele detection depended on the expression of the V genes and allelic imbalance (Fig. 18). None of the tools yielded false-positive alleles, so no SHM was mistakenly recognized as an allelic mutation.However, out of 20 alleles which were not present in the initial reference library, TIgGER managed to detect only 8 (40%, Table 3), while our algorithm identified 19 (95%, Table 3). While our approached failed to detect the allele with significant allelic imbalance (IGHV1-2*05), TIgGER showed no apparent preference for detecting or missing alleles with low expression or allelic imbalance. Interestingly, downsampling up to less than one tenth of the aligned reads (100 000) had no effect on the allele detection by MiStrainer.

Tool	**Number of reads for downsampling Number of inferred alleles**				
	Full sample (1 071 532 reads)	**500k**	**100k**	**50k**	**10k**
MiStrainer	19 (95%)	19 (95%)	19 (95%)	18 (90%)	14 (70%)
TIgGER	8 (40%)	8 (40%)	6 (30%)	5 (25%)	0 (0%)

Table 3. Detection of the allele variants of V-genes depending on depth of sequencing. Number and fraction of detected alleles out of 20 alleles present in donor's genotype confirmed by long-read sequencing and absent in the initial reference library gene set.

Even when downsampled to 10 000 aligned reads the majority of alleles (70%) were identified (Table 3, row 1). TIgGer on the contrary, was much more sensitive to the number of reads in the input library (Table 3, row 2). Thus we demonstrate superior sensitivity of our approach which allows using datasets with shallow sequencing depth for inference of allelic variants for the majority of the V- and J-genes.

Chapter 5. Discussion and recommendations

Using advanced library preparation technology, we performed a longitudinal study of BCR repertoires of the three main antigen-experienced B cell subsets — memory B cells, plasmablasts, and plasma cells — from peripheral blood of six donors, sampled three times over the course of a year. We analyzed these repertoires from two conceptually different but complementary points of view. We compared various repertoire features between the cell subsets, including clonotype stability in time and convergence between individuals.

Comparative analysis of the cell subsets revealed significant differences in IGH isotype distribution, rate of SHM, and CDR3 length. IgM clonotypes predominated in the Bmem subset, whereas in ASCs the switched isotypes IgA and IgG together represented >80% of repertoire diversity on average. As expected, classical switched isotypes have higher rates of SHM, and the rate of SHM in ASCs is in general higher than in Bmem. This is concordant with previous observation of Bmem being predominantly produced in the early GCs, while LLPCs emerge very late in the GC responses (Weisel et al. 2016), which may explain the observed higher levels of SHM in antibody-secreting cells (ASCs) (Phad et al. 2022) and assume less mutated more broadly reactive BCRs in Bmem compared to LLPCs. The IgD isotype in Bmem cells showed similarities to IgM, where most IgD clonotypes had low levels of SHM, although there was a fraction of heavily-mutated clonotypes. On average, IgD-switched PL and PBL had a comparable level of

SHM with IgG- and IgA-expressing ASC clonotypes. Notably, the level of SHM and CDR3 length in PBL on average exceeded that of PL in IgM, IgA, and IgG isotypes. We hypothesize that such PBLs with heavily hypermutated BCRs could be the subset of B cell progeny that continue to acquire mutations after optimal affinity has been achieved, while another part of the clonal progeny is committed to a long-lived PL fate and acquires the CD138 marker characteristic of this cell subset (Garimilla et al. 2019).

While different in many aspects, immune-experienced B cell subsets are similar — and concordantly distinct from naive B cells — in terms of IGHV gene segment usage. Moreover, we observed that the correlated enrichment or depletion in V segment usage frequency generally coincides with the level of sequence similarity of the V segments. Most IGHV-3 family members were observed more frequently in antigen-experienced B cells compared to naive subsets in all donors and time points, while most of the other V genes that are well-represented in the naive subset decreased in frequency. These differences in V usage frequencies between naive and antigen-experienced B cell subsets have also been reported in several previous studies, even though different FACS gating strategies were used (Mitsunaga and Snyder 2020; Ghraichy et al. 2021). Our findings further support the idea that initial recruitment of B cells to the immune response is in many cases determined by the germline-encoded parts of the BCR, presumably CDR1 and CDR2. Previous studies have shown high levels of convergence in IGHV usage between B cell clonotypes specific for particular pathogens or

self-antigens (Peng et al. 2019; Galson et al. 2015; Bashford-Rogers et al. 2019).

We further analyzed the repertoire similarity of cell subsets over time and between individuals. Intuitively, the Bmem subset is the most stable over time, showing less repertoire divergence and a greater number of shared clonotypes between sampling time points in the same individuals. Our finding expands the recent observation of Bmem subset stability in elderly donors (Phad et al, 2022) on a larger cohort of donors of younger age. Compared to intra-individual sharing, we detected a very small number of common clonotypes in Bmem cells. Those clonotypes have comparable levels of SHM to private ones, assuming a germinal center-dependent origin. Two recent studies on extra-deep repertoires of bulk peripheral blood B cells reported 1–6% (Soto et al. 2019) or ~1% (Briney et al. 2019) shared V-CDR3aa-J clonotypes between pairs of unrelated donors, with lower repertoire convergence for class-switched clonotypes shown in the latter study. Using the same method, we similarly measured 0.06% repertoire overlap in the Bmem subset. Complementing the model proposed by Briney *et al.* — wherein IGH repertoires are initially dissimilar and then homogenize during B cell development before finally becoming highly individualized after immunological exposure — we found a significantly higher number of shared clonotypes between IGH repertoires among the most abundant Bmem clonotypes, indicating functional convergence presumably due to exposure to common environmental antigens. The latter is further supported by the higher number of persisting

Bmem clonotypes observed among public clonotypes compared to private ones.

There were two important limitations in the conducted studies. First, only two of the donors were female, thus there was a considerable gender bias in the study which should be addressed in future studies. Second, having naive B-cell repertoires sequenced would add more credibility in the analysis compared to utilization of external datasets. However in the dataset used, repertoire sequencing was performed using a 5'RACE-based approach very similar to ours, which limits the possible biases introduced by potential batch effects.

Finally, we have developed a novel approach for inferring novel allelic variants of V- and J-genes for construction of individualized reference libraries, which were shown to be important for downstream analysis of the adaptive immune receptor repertoires (Gadala-Maria et al., 2015). Our approach has a number of advantages including the ability to be used on both hypermutated and non-hypermutated data and also low requirements for the completeness of the starting library of V- and J-gene allelic variants. It was tested using an AIRR-seq dataset coupled with the accurate genotype of the donor, reconstructed using a NGS long-read technology, and could perform well even at the depth of 100 000 reads aligned to the reference. Despite promising results we consider further validation of the algorithm necessary, especially utilizing datasets from donors with non-caucasian origin. Further directions of this study would also include investigating the allelic diversity in human population and its impact on development of various diseases.

Bibliography

1. Akkaya, M., Kwak, K. & Pierce, S.K. B cell memory: building two walls of protection against pathogens. *Nat Rev Immunol* **20,** 229–238 (2020). https://doi.org/10.1038/s41577-019-0244-2

2. Allen, Christopher D C, and Jason G Cyster. “Follicular dendritic cell networks of primary follicles and germinal centers: phenotype and function.” *Seminars in immunology* vol. 20,1 (2008): 14-25. doi:10.1016/j.smim.2007.12.001

3. Anderson, Shannon M, Ashraf Khalil, Mohamed Uduman, Uri Hershberg, Yoram Louzoun, Ann M Haberman, Steven H Kleinstein, and Mark J Shlomchik. “Taking Advantage: High-Affinity B Cells in the Germinal Center Have Lower Death Rates, but Similar Rates of Division, Compared to Low-Affinity Cells.” *Journal of Immunology* 183, no. 11 (December 1, 2009): 7314–25. https://doi.org/10.4049/jimmunol.0902452.

4. Bashford-Rogers, R. J. M., L. Bergamaschi, E. F. McKinney, D. C. Pombal, F. Mescia, J. C. Lee, D. C. Thomas, et al. 2019. “Analysis of the B Cell Receptor Repertoire in Six Immune-Mediated Diseases.” *Nature* 574 (7776): 122–26. https://doi.org/10.1038/s41586-019-1595-3.

5. Bolotin, Dmitriy A, Stanislav Poslavsky, Igor Mitrophanov, Mikhail Shugay, Ilgar Z Mamedov, Ekaterina V Putintseva, and Dmitriy M Chudakov. 2015. “MiXCR: Software for Comprehensive Adaptive Immunity Profiling.” *Nature Methods* 12 (5): 380–81. https://doi.org/10.1038/nmeth.3364.

6. Bonsignori, Mattia, Hua-Xin Liao, Feng Gao, Wilton B. Williams, S. Munir Alam, David C. Montefiori, and Barton F. Haynes. 2017. “Antibody-Virus Co-Evolution in HIV Infection: Paths for HIV Vaccine Development.” *Immunological Reviews* 275 (1): 145–60. https://doi.org/10.1111/imr.12509.

7. Briney, Bryan, Anne Inderbitzin, Collin Joyce, and Dennis R. Burton. 2019. “Commonality despite Exceptional Diversity in the Baseline Human Antibody Repertoire.” *Nature* 566 (7744): 393–97. https://doi.org/10.1038/s41586-019-0879-y.

8. Chaudhary, Neha, and Duane R. Wesemann. 2018. “Analyzing Immunoglobulin Repertoires.” *Frontiers in Immunology* 9 (March): 462. https://doi.org/10.3389/fimmu.2018.00462.

9. Corcoran, Martin M. and Phad, Ganesh E. and Bernat, Néstor Vázquez and Stahl-Hennig, Christiane and Sumida, Noriyuki and Persson, Mats A.A. and Martin, Marcel and Karlsson Hedestam, Gunilla B. “Production of individualized V gene databases reveals high levels of immunoglobulin genetic diversity”. *Nature Communications* 7:13642 (2016) https://dx.doi.org/10.1038/ncomms13642

10. Cremasco, Viviana et al. “B cell homeostasis and follicle confines are governed by fibroblastic reticular cells.” *Nature immunology* vol. 15,10 (2014): 973-81. doi:10.1038/ni.2965

11. Crotty, Shane. “T Follicular Helper Cell Biology: A Decade of Discovery and Diseases.” *Immunity* vol. 50,5 (2019): 1132-1148. doi:10.1016/j.immuni.2019.04.011

12. Cyster, Jason G. “B cell follicles and antigen encounters of the third kind.” *Nature immunology* vol. 11,11 (2010): 989-96. doi:10.1038/ni.1946

13. Cyster, Jason G, and Christopher D C Allen. “B Cell Responses: Cell Interaction Dynamics and Decisions.” *Cell* vol. 177,3 (2019): 524-540. doi:10.1016/j.cell.2019.03.016

14. Cyster, Jason G, and Susan R Schwab. “Sphingosine-1-phosphate and lymphocyte egress from lymphoid organs.” *Annual review of immunology* vol. 30 (2012): 69-94. doi:10.1146/annurev-immunol-020711-075011

15. Davydov, Alexey N., Anna S. Obraztsova, Mikhail Y. Lebedin, Maria A. Turchaninova, Dmitriy B. Staroverov, Ekaterina M. Merzlyak, George V. Sharonov, et al. 2018. “Comparative Analysis of B-Cell Receptor Repertoires Induced by Live Yellow Fever Vaccine in Young and Middle-Age Donors.” *Frontiers in Immunology* 9 (October): 2309. https://doi.org/10.3389/fimmu.2018.02309.

16. De Bourcy, Charles F. A., Cesar J. Lopez Angel, Christopher Vollmers, Cornelia L. Dekker, Mark M. Davis, and Stephen R. Quake. 2017. “Phylogenetic Analysis of the Human Antibody Repertoire Reveals Quantitative Signatures of Immune Senescence and Aging.” *Proceedings of the National Academy of Sciences* 114 (5): 1105–10. https://doi.org/10.1073/pnas.1617959114.

17. De Silva, Nilushi S., and Ulf Klein. 2015. “Dynamics of B Cells in Germinal Centres.” *Nature Reviews Immunology* 15 (3): 137–48. https://doi.org/10.1038/nri3804.

18. Edgar, R. C. 2004. “MUSCLE: Multiple Sequence Alignment with High Accuracy and High Throughput.” *Nucleic Acids Research* 32 (5): 1792–97. https://doi.org/10.1093/nar/gkh340.

19. Fujita, Teizo. “Evolution of the lectin-complement pathway and its role in innate immunity.” *Nature reviews. Immunology* vol. 2,5 (2002): 346-53. doi:10.1038/nri800

20. Gadala-Maria, Daniel, Gur Yaari, Mohamed Uduman, and Steven H. Kleinstein. 2015. “Automated Analysis of High-Throughput B-Cell Sequencing Data Reveals a High Frequency of Novel Immunoglobulin V Gene Segment Alleles.” *Proceedings of the National Academy of Sciences* 112 (8): E862–70. https://doi.org/10.1073/pnas.1417683112.

21. Gaebler, Christian, Zijun Wang, Julio C. C. Lorenzi, Frauke Muecksch, Shlomo Finkin, Minami Tokuyama, Alice Cho, et al. 2021. “Evolution of Antibody Immunity to SARS-CoV-2.” *Nature* 591 (7851): 639–44. https://doi.org/10.1038/s41586-021-03207-w.

22. Galson, Jacob D, Elizabeth A Clutterbuck, Johannes Trück, Maheshi N Ramasamy, Márton Münz, Anna Fowler, Vincenzo Cerundolo, Andrew J Pollard, Gerton Lunter, and Dominic F Kelly. 2015. “BCR Repertoire Sequencing: Different Patterns of B-cell Activation after Two Meningococcal Vaccines.” *Immunology & Cell Biology* 93 (10): 885–95. https://doi.org/10.1038/icb.2015.57.

23. Garimilla, Swetha, Doan C. Nguyen, Jessica L. Hallileу, Christopher Tipton, Alexander F. Rosenberg, Christopher F. Fucile, Celia L. Saney, et al. 2019. “Differential Transcriptome and Development of Human Peripheral Plasma Cell Subsets.” *JCI Insight* 4 (9): e126732. https://doi.org/10.1172/jci.insight.126732.

24. Georgiou, George et al. “The promise and challenge of high-throughput sequencing of the antibody repertoire.” *Nature biotechnology* vol. 32,2 (2014): 158-68. doi:10.1038/nbt.2782

25. Ghraichy, Marie, Valentin von Niederhäusern, Aleksandr Kovaltsuk, Jacob D Galson, Charlotte M Deane, and Johannes Trück. 2021. “Different B Cell Subpopulations Show Distinct Patterns in Their IgH Repertoire Metrics.” *ELife* 10 (October): e73111. https://doi.org/10.7554/eLife.73111.

26. Gidoni, Moriah, Omri Snir, Ayelet Peres, Pazit Polak, Ida Lindeman, Ivana Mikocziova, Vikas Kumar Sarna, et al. 2019. “Mosaic Deletion Patterns of the Human Antibody Heavy Chain Gene Locus Shown by Bayesian Haplotyping.” *Nature Communications* 10 (1): 628. https://doi.org/10.1038/s41467-019-08489-3.

27. Gojobori, Nei. 1986. “Simple Methods for Estimating the Numbers of Synonymous and Nonsynonymous Nucleotide Substitutions.” *Molecular Biology and Evolution*, September. https://doi.org/10.1093/oxfordjournals.molbev.a040410.

28. Green, Jesse A et al. “The sphingosine 1-phosphate receptor $S1P_2$ maintains the homeostasis of germinal center B cells and promotes niche confinement.” *Nature immunology* vol. 12,7 672-80. 5 Jun. 2011, doi:10.1038/ni.2047

29. Grimsholm, Ola, Eva Piano Mortari, Alexey N. Davydov, Mikhail Shugay, Anna S. Obraztsova, Chiara Bocci, Emiliano Marasco, et al. 2020. “The Interplay between CD27dull and CD27bright B Cells Ensures the Flexibility, Stability, and Resilience of Human B Cell Memory.” *Cell Reports* 30 (9): 2963-2977.e6. https://doi.org/10.1016/j.celrep.2020.02.022.

30. Gupta, Namita T., Jason A. Vander Heiden, Mohamed Uduman, Daniel Gadala-Maria, Gur Yaari, and Steven H. Kleinstein. 2015. “Change-O: A Toolkit for Analyzing Large-Scale B Cell Immunoglobulin Repertoire Sequencing Data.” *Bioinformatics* 31 (20): 3356–58. https://doi.org/10.1093/bioinformatics/btv359.

31. Heesters, Balthasar A et al. “Follicular dendritic cells: dynamic antigen libraries.” *Nature reviews. Immunology* vol. 14,7 (2014): 495-504. doi:10.1038/nri3689

32. Hoehn, Kenneth B., Jackson S. Turner, Frederick I. Miller, Ruoyi Jiang, Oliver G. Pybus, Ali H. Ellebedy, and Steven H. Kleinstein. 2021. “Human B Cell Lineages

Engaged by Germinal Centers Following Influenza Vaccination Are Measurably Evolving." Preprint. Immunology. https://doi.org/10.1101/2021.01.06.425648.

33. Hoogeboom, Robbert, and Pavel Tolar. "Molecular Mechanisms of B Cell Antigen Gathering and Endocytosis." *Current topics in microbiology and immunology* vol. 393 (2016): 45-63. doi:10.1007/82_2015_476

34. Horns, Felix, Christopher Vollmers, Cornelia L. Dekker, and Stephen R. Quake. 2019. "Signatures of Selection in the Human Antibody Repertoire: Selective Sweeps, Competing Subclones, and Neutral Drift." *Proceedings of the National Academy of Sciences* 116 (4): 1261–66. https://doi.org/10.1073/pnas.1814213116.

35. Kurosaki, Tomohiro et al. "B cell signaling and fate decision." *Annual review of immunology* vol. 28 (2010): 21-55. doi:10.1146/annurev.immunol.021908.132541

36. Laserson, Uri, Francois Vigneault, Daniel Gadala-Maria, Gur Yaari, Mohamed Uduman, Jason A. Vander Heiden, William Kelton, et al. 2014. "High-Resolution Antibody Dynamics of Vaccine-Induced Immune Responses." *Proceedings of the National Academy of Sciences* 111 (13): 4928–33. https://doi.org/10.1073/pnas.1323862111.

37. Lefranc, Marie-Paule, and Gérard Lefranc. "Immunoglobulins or Antibodies: IMGT® Bridging Genes, Structures and Functions." *Biomedicines* vol. 8,9 319. 31 Aug. 2020, doi:10.3390/biomedicines8090319

38. Liu, Wanli et al. "Antigen Receptor Nanoclusters: Small Units with Big Functions." *Trends in immunology* vol. 37,10 (2016): 680-689. doi:10.1016/j.it.2016.07.007

39. Ma, Ke-Yue et al. "Immune Repertoire Sequencing Using Molecular Identifiers Enables Accurate Clonality Discovery and Clone Size Quantification." *Frontiers in immunology* vol. 9 33. 5 Feb. 2018, doi:10.3389/fimmu.2018.00033

40. Mandric, Igor, Jeremy Rotman, Harry Taegyun Yang, Nicolas Strauli, Dennis J. Montoya, William Van Der Wey, Jiem R. Ronas, et al. 2020. "Profiling Immunoglobulin Repertoires across Multiple Human Tissues Using RNA Sequencing." *Nature Communications* 11 (1): 3126. https://doi.org/10.1038/s41467-020-16857-7.

41. Mayer, Christian T, Anna Gazumyan, Ervin E Kara, Alexander D Gitlin, Jovana

Golijanin, Charlotte Viant, Joy Pai, et al. “The Microanatomic Segregation of Selection by Apoptosis in the Germinal Center.” *Science* 358, no. 6360 (October 13, 2017). https://doi.org/10.1126/science.aao2602.

42. McDonald, John H., and Martin Kreitman. 1991. “Adaptive Protein Evolution at the Adh Locus in Drosophila.” *Nature* 351 (6328): 652–54. https://doi.org/10.1038/351652a0.

43. Mitsunaga, Erin M., and Michael P. Snyder. 2020. “Deep Characterization of the Human Antibody Response to Natural Infection Using Longitudinal Immune Repertoire Sequencing.” *Molecular & Cellular Proteomics* 19 (2): 278–93. https://doi.org/10.1074/mcp.RA119.001633.

44. Mora Thierry, Aleksandra M. Walczak. Quantifying lymphocyte receptor diversity. in Systems Immunology: An Introduction to Modeling Methods for Scientists, eds. J. Das, and C. Jayaprakash (2018)

45. Nemazee, David. “Mechanisms of central tolerance for B cells.” *Nature reviews. Immunology* vol. 17,5 (2017): 281-294. doi:10.1038/nri.2017.19

46. Nielsen, Sandra C.A., Fan Yang, Katherine J.L. Jackson, Ramona A. Hoh, Katharina Röltgen, Grace H. Jean, Bryan A. Stevens, et al. 2020. “Human B Cell Clonal Expansion and Convergent Antibody Responses to SARS-CoV-2.” *Cell Host & Microbe* 28 (4): 516-525.e5. https://doi.org/10.1016/j.chom.2020.09.002.

47. Nourmohammad, Armita, Jakub Otwinowski, Marta Łuksza, Thierry Mora, and Aleksandra M Walczak. 2019. “Fierce Selection and Interference in B-Cell Repertoire Response to Chronic HIV-1.” Edited by Thomas Leitner. *Molecular Biology and Evolution* 36 (10): 2184–94. https://doi.org/10.1093/molbev/msz143.

48. Peng, Wujian, Song Liu, Jingye Meng, Jiali Huang, Jianrong Huang, Donge Tang, and Yong Dai. 2019. “Profiling the TRB and IGH Repertoire of Patients with H5N6 Avian Influenza Virus Infection by High-Throughput Sequencing.” *Scientific Reports* 9 (1): 7429. https://doi.org/10.1038/s41598-019-43648-y.

49. Perez-Andres, M., B. Paiva, W. G. Nieto, A. Caraux, A. Schmitz, J. Almeida, R. F. Vogt, et al. 2010. “Human Peripheral Blood B-Cell Compartments: A Crossroad in B-Cell Traffic.” *Cytometry Part B: Clinical Cytometry* 78B (S1): S47–60. https://doi.org/10.1002/cyto.b.20547.

50. Phad, G. E., Pinto, D., Foglierini, M., Akhmedov, M., Rossi, R. L., Malvicini, E., Cassotta, A., Fregni, C. S., Bruno, L., Sallusto, F., & Lanzavecchia, A. (2022). Clonal structure, stability and dynamics of human memory B cells and circulating plasmablasts. *Nature immunology*, *23*(7), 1–10. https://doi.org/10.1038/s41590-022-01230-1

51. R Core Team. 2018. “R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing,.” https://www.R-project.org/.

52. Ralph, Duncan & Matsen IV, Frederick. (2017). Per-sample immunoglobulin germline inference from B cell receptor deep sequencing data. PLOS Computational Biology. 15. 10.1371/journal.pcbi.1007133.

53. Robinson, M.D., Oshlack, A. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol* **11,** R25 (2010). https://doi.org/10.1186/gb-2010-11-3-r25

54. Robinson, M.D., McCarthy, D.J. and Smyth, G.K. (2010) EdgeR: A Bioconductor Package for Differential Expression Analysis of Digital Gene Expression Data. Bioinformatics, 26, 139-140. http://dx.doi.org/10.1093/bioinformatics/btp616

55. Rodda, Lauren B et al. “Single-Cell RNA Sequencing of Lymph Node Stromal Cells Reveals Niche-Associated Heterogeneity.” *Immunity* vol. 48,5 (2018): 1014-1028.e6. doi:10.1016/j.immuni.2018.04.006

56. Rodriguez, O. L., Gibson, W. S., Parks, T., Emery, M., Powell, J., Strahl, M., Deikus, G., Auckland, K., Eichler, E. E., Marasco, W. A., Sebra, R., Sharp, A. J., Smith, M. L., Bashir, A., & Watson, C. T. (2020). A novel framework for characterizing genomic haplotype diversity in the human immunoglobulin heavy chain locus. *Frontiers in Immunology*, *11*, 2136. https://doi.org/10.3389/fimmu.2020.02136

57. Rodriguez, O. L., Safonova, Y., Silver, C. A., Shields, K., Gibson, W. S., Kos, J. T., Tieri, D., Ke, H., Jackson, K. J. L., Boyd, S. D., Smith, M., Marasco, W., & Watson, C. T. (2022). Antibody repertoire gene usage is explained by common genetic variants in the immunoglobulin heavy chain locus. *BioRxiv*. https://doi.org/10.1101/2022.07.04.498729

58. Rubelt F *et al*. AIRR Community Recommendations for Sharing Immune Repertoire Sequencing Data. Nat Immunol 18:1274 (2017). DOI: 10.1038/ni.3873

59. Sakharkar, Mrunal, C. Garrett Rappazzo, Wendy F. Wieland-Alter, Ching-Lin Hsieh, Daniel Wrapp, Emma S. Esterman, Chengzi I. Kaku, et al. 2021. “Prolonged Evolution of the Human B Cell Response to SARS-CoV-2 Infection.” *Science Immunology* 6 (56): eabg6916. https://doi.org/10.1126/sciimmunol.abg6916.

60. Sethna, Zachary, Yuval Elhanati, Curtis G Callan, Aleksandra M Walczak, and Thierry Mora. 2019. “OLGA: Fast Computation of Generation Probabilities of B- and T-Cell Receptor Amino Acid Sequences and Motifs.” Edited by Bonnie Berger. *Bioinformatics* 35 (17): 2974–81. https://doi.org/10.1093/bioinformatics/btz035.

61. Shah, Hemangi B., Kenneth Smith, Jonathan D. Wren, Carol F. Webb, Jimmy D. Ballard, Rebecka L. Bourn, Judith A. James, and Mark L. Lang. 2019. “Insights From Analysis of Human Antigen-Specific Memory B Cell Repertoires.” *Frontiers in Immunology* 9 (January): 3064. https://doi.org/10.3389/fimmu.2018.03064.

62. Elsner, Rebecca A, and Mark J Shlomchik. “Germinal Center and Extrafollicular B Cell Responses in Vaccination, Immunity, and Autoimmunity.” *Immunity* vol. 53,6 (2020): 1136-1150. doi:10.1016/j.immuni.2020.11.006

63. Schroeder, Harry W Jr, and Lisa Cavacini. “Structure and function of immunoglobulins.” *The Journal of allergy and clinical immunology* vol. 125,2 Suppl 2 (2010): S41-52. doi:10.1016/j.jaci.2009.09.046

64. Shugay, Mikhail, Olga V Britanova, Ekaterina M Merzlyak, Maria A Turchaninova, Ilgar Z Mamedov, Timur R Tuganbaev, Dmitriy A Bolotin, et al. 2014. “Towards Error-Free Profiling of Immune Repertoires.” *Nature Methods* 11 (6): 653–55. https://doi.org/10.1038/nmeth.2960.

65. Soto, Cinque, Robin G. Bombardi, Andre Branchizio, Nurgun Kose, Pranathi Matta, Alexander M. Sevy, Robert S. Sinkovits, Pavlo Gilchuk, Jessica A. Finn, and James E. Crowe. 2019. “High Frequency of Shared Clonotypes in Human B Cell Receptor Repertoires.” *Nature* 566 (7744): 398–402. https://doi.org/10.1038/s41586-019-0934-8.

66. Stamatakis, Alexandros. 2014. “RAxML Version 8: A Tool for Phylogenetic Analysis and Post-Analysis of Large Phylogenies.” *Bioinformatics* 30 (9): 1312–13. https://doi.org/10.1093/bioinformatics/btu033.

67. Stavnezer, Janet, Jeroen E.J. Guikema, and Carol E. Schrader. 2008. “Mechanism and Regulation of Class Switch Recombination.” *Annual Review of Immunology* 26 (1): 261–92. https://doi.org/10.1146/annurev.immunol.26.021607.090248.

68. Tonegawa, S. “Somatic generation of antibody diversity.” *Nature* vol. 302,5909 (1983): 575-81. doi:10.1038/302575a0

69. Toyama, Hirochika et al. “Memory B cells without somatic hypermutation are generated from Bcl6-deficient B cells.” *Immunity* vol. 17,3 (2002): 329-39. doi:10.1016/s1074-7613(02)00387-4

70. Turchaninova, M A, A Davydov, O V Britanova, M Shugay, V Bikos, E S Egorov, V I Kirgizova, et al. 2016. “High-Quality Full-Length Immunoglobulin Profiling with Unique Molecular Barcoding.” *Nature Protocols* 11 (9): 1599–1616. https://doi.org/10.1038/nprot.2016.093.

71. Vidarsson, Gestur, Gillian Dekkers, and Theo Rispens. 2014. “IgG Subclasses and Allotypes: From Structure to Effector Functions.” *Frontiers in Immunology* 5 (October). https://doi.org/10.3389/fimmu.2014.00520.

72. Weisel, Florian J et al. “A Temporal Switch in the Germinal Center Determines Differential Output of Memory B and Plasma Cells.” *Immunity* vol. 44,1 (2016): 116-130. doi:10.1016/j.immuni.2015.12.004

73. Wickham, Hadley. 2016. “ggplot2: Elegant Graphics for Data Analysis.” Springer-Verlag New York. doi:10.1007/978-0-387-98141-3_6

74. Wu, Yu-Chang, David Kipling, Hui Sun Leong, Victoria Martin, Alexander A. Ademokun, and Deborah K. Dunn-Walters. 2010. “High-Throughput Immunoglobulin Repertoire Analysis Distinguishes between Human IgM Memory and Switched Memory B-Cell Populations.” *Blood* 116 (7): 1070–78. https://doi.org/10.1182/blood-2010-03-275859.

75. Yang, Jianying, and Michael Reth. “Receptor Dissociation and B-Cell Activation.” *Current topics in microbiology and immunology* vol. 393 (2016): 27-43. doi:10.1007/82_2015_482

76. Yang, Xiujia, Minhui Wang, Jiaqi Wu, Dianchun Shi, Yanfang Zhang, Huikun Zeng, Yan Zhu, et al. 2021. “Large-Scale Analysis of 2,152 Ig-Seq Datasets Reveals Key Features of B Cell Biology and the Antibody Repertoire.” *Cell Reports* 35 (6): 109110. https://doi.org/10.1016/j.celrep.2021.109110.

77. Yu, Guangchuang, David K. Smith, Huachen Zhu, Yi Guan, and Tommy Tsan-Yuk Lam. 2017. “Ggtree: Package for Visualization and Annotation of Phylogenetic Trees with Their Covariates and Other Associated Data.” Edited by Greg McInerny. *Methods in Ecology and Evolution* 8 (1): 28–36. https://doi.org/10.1111/2041-210X.12628.

78. Zhang, W., Wang, I. M., Wang, C., Lin, L., Chai, X., Wu, J., Bett, A. J., Dhanasekaran, G., Casimiro, D. R., & Liu, X. (2016). IMPre: An Accurate and Efficient Software for Prediction of T- and B-Cell Receptor Germline Genes and Alleles from Rearranged Repertoire Data. *Frontiers in immunology*, *7*, 457. https://doi.org/10.3389/fimmu.2016.00457

79. Zvyagin, I V, I Z Mamedov, O V Tatarinova, E A Komech, E E Kurnikova, E V Boyakova, V Brilliantova, et al. 2017. “Tracking T-Cell Immune Reconstitution after TCRαβ/CD19-Depleted Hematopoietic Cells Transplantation in Children.” *Leukemia* 31 (5): 1145–53. https://doi.org/10.1038/leu.2016.321.

Appendices

Supplementary Data SD1.

Sequencing statistics for all samples in the study

ID	Donor	Time point	Cell subset	Replicate label	Number of sorted cells	Total sequencing reads	Total UMIs identified	Oversequencing threshold (MIGEC)	Number of unique IGH cDNA molecules	IGH clonotypes
AT_T1_Bmem_1	AT	T1	Bmem	1	101 400	789 962	183 522	3	45 611	26 264
AT_T1_PBL_1	AT	T1	PBL	1	7 200	4 093 824	266 448	4	120 200	3 304
AT_T1_PL_1	AT	T1	PL	1	1 800	2 375 939	250 581	3	116 700	1 975
AT_T2_Bmem_1	AT	T2	Bmem	1	50 600	1 108 213	91 975	4	44 065	21 311
AT_T2_Bmem_2	AT	T2	Bmem	2	57 400	970 185	72 073	4	34 567	19 426
AT_T2_PBL_1	AT	T2	PBL	1	2 520	1 334 248	83 646	4	47 569	1 391
AT_T2_PL_1	AT	T2	PL	1	800	1 584 476	104 166	4	63 796	686
AT_T3_Bmem_1	AT	T3	Bmem	1	50 000	852 735	55 667	4	31 270	16 480
AT_T3_Bmem_2	AT	T3	Bmem	2	40 800	1 179 541	29 503	8	10 294	7 308
AT_T3_PBL_1	AT	T3	PBL	1	1 000	892 947	37 707	4	20 500	628
AT_T3_PBL_2	AT	T3	PBL	2	1 000	614 693	23 220	6	7 726	981
AT_T3_PL_1	AT	T3	PL	1	400	564 322	14 671	6	7 387	157
AT_T3_PL_2	AT	T3	PL	2	200	559 237	14 668	6	5 720	229
D01_T2_Bmem_1	D01	T2	Bmem	1	50 300	1 213 467	138 253	3	74 269	25 096

D01_T2_Bmem_3	D01	T2	Bmem	3	55 400	1 405 674	126 501	3	62 296	24 946
D01_T2_PBL_1	D01	T2	PBL	1	2 100	1 443 850	189 645	3	86 415	1 083
D01_T2_PBL_2	D01	T2	PBL	2	2 100	1 494 871	192 790	3	76 577	1 330
D01_T2_PL_1	D01	T2	PL	1	1 020	1 088 070	224 037	3	79 751	1 019
D01_T2_PL_2	D01	T2	PL	2	1 010	860 382	252 380	3	67 242	840
D01_T3_Bmem_1	D01	T3	Bmem	1	50 000	501 375	111 232	3	26 642	14 572
D01_T3_Bmem_2	D01	T3	Bmem	2	50 000	657 227	114 142	3	37 675	18 832
D01_T3_PBL_1	D01	T3	PBL	1	1 000	559 732	93 005	3	37 336	720
D01_T3_PBL_2	D01	T3	PBL	2	1 000	1 046 599	87 976	4	39 947	602
D01_T3_PBL_3	D01	T3	PBL	3	1 000	1 686 059	136 467	4	64 365	671
D01_T3_PL_1	D01	T3	PL	1	500	658 225	70 016	3	37 749	297
D01_T3_PL_2	D01	T3	PL	2	500	676 981	100 252	3	47 630	450
IM_T1_Bmem_1	IM	T1	Bmem	1	186 572	4 749 335	168 413	6	107 678	58 370
IM_T1_PL_1	IM	T1	PL	1	129	610 230	11 803	8	4 748	71
IM_T2_Bmem_1	IM	T2	Bmem	1	69 900	770 829	293 287	3	33 415	18 977
IM_T2_Bmem_2	IM	T2	Bmem	2	68 400	1 599 771	333 319	3	80 817	32 575
IM_T2_PBL_1	IM	T2	PBL	1	2 000	2 497 522	576 024	3	167 286	1 809
IM_T2_PBL_3	IM	T2	PBL	3	2 486	1 859 211	433 510	3	138 552	1 519

IM_T2_PL_2	IM	T2	PL	2	920	1 034 501	198 676	3	65 247	883
IM_T3_Bmem_1	IM	T3	Bmem	1	50 000	753 205	76 005	3	31 868	17 241
IM_T3_Bmem_2	IM	T3	Bmem	2	50 000	587 810	81 367	3	28 486	16 374
IM_T3_PBL_1	IM	T3	PBL	1	2 000	1 890 905	73 810	6	37 633	1 140
IM_T3_PBL_2	IM	T3	PBL	2	2 000	1 660 572	64 768	6	31 294	1 074
IM_T3_PL_1	IM	T3	PL	1	1 000	1 323 995	48 189	6	24 131	648
IM_T3_PL_2	IM	T3	PL	2	1 000	1 688 147	40 890	8	17 015	623
IZ_T1_Bmem_1	IZ	T1	Bmem	1	101 800	3 343 261	155 571	6	75 849	45 699
IZ_T1_PBL_1	IZ	T1	PBL	1	3 900	2 524 536	123 589	4	66 047	1 533
IZ_T1_PL_1	IZ	T1	PL	1	850	2 203 302	98 341	4	59 518	725
IZ_T2_Bmem_1	IZ	T2	Bmem	1	50 500	2 038 213	158 150	4	61 519	29 479
IZ_T2_Bmem_2	IZ	T2	Bmem	2	56 300	1 754 528	150 761	4	71 961	32 211
IZ_T2_PBL_1	IZ	T2	PBL	1	1 140	396 294	40 934	3	20 687	670
IZ_T2_PBL_2	IZ	T2	PBL	2	1 840	2 070 821	96 511	4	46 777	1 074
IZ_T2_PL_1	IZ	T2	PL	1	1 050	1 422 734	90 373	4	50 559	647
IZ_T2_PL_2	IZ	T2	PL	2	625	1 927 504	38 262	8	12 112	478
IZ_T3_Bmem_1	IZ	T3	Bmem	1	50 000	1 258 639	140 436	3	69 643	31 358
IZ_T3_Bmem_2	IZ	T3	Bmem	2	50 000	1 267 287	184 493	3	88 524	32 806

IZ_T3_PBL_1	IZ	T3	PBL	1	2 000	1 595 607	108 325	4	47 087	924
IZ_T3_PBL_2	IZ	T3	PBL	2	2 000	1 717 640	112 206	6	25 034	1 049
IZ_T3_PL_1	IZ	T3	PL	1	200	658 983	58 323	3	31 681	205
IZ_T3_PL_2	IZ	T3	PL	2	200	799 307	60 330	4	20 208	227
MRK_T1_Bmem_1	MRK	T1	Bmem	1	143 162	3 497 861	202 489	4	110 734	53 191
MRK_T1_PL_1	MRK	T1	PL	1	251	947 060	18 357	8	7 247	121
MRK_T2_Bmem_1	MRK	T2	Bmem	1	51 700	1 492 681	138 361	3	75 438	31 960
MRK_T2_Bmem_2	MRK	T2	Bmem	2	50 600	1 467 925	116 428	4	43 879	21 619
MRK_T2_PBL_1	MRK	T2	PBL	1	2 130	1 901 745	174 345	3	94 195	1 294
MRK_T2_PBL_3	MRK	T2	PBL	3	2 020	1 864 752	313 948	3	144 008	1 194
MRK_T2_PBL_4	MRK	T2	PBL	4	2 000	1 398 793	54 776	6	16 234	1 430
MRK_T2_PL_1	MRK	T2	PL	1	1 000	1 265 223	106 033	3	61 553	774
MRK_T2_PL_2	MRK	T2	PL	2	1 035	1 273 643	113 246	3	62 750	800
MRK_T2_PL_3	MRK	T2	PL	3	1 000	1 993 922	181 675	3	109 307	1 183
MRK_T3_PL_1	MRK	T3	PL	1	400	912 109	36 495	4	22 164	305
MT_T3_Bmem_1	MT	T3	Bmem	1	50 000	1 399 198	64 019	4	30 257	16 718
MT_T3_Bmem_2	MT	T3	Bmem	2	50 000	1 249 421	96 955	4	33 399	17 639
MT_T3_PBL_1	MT	T3	PBL	1	1 000	1 265 498	34 101	6	16 036	528

MT_T3_PBL_2	MT	T3	PBL	2	1 000	961 889	22 404	6	7 946	554
MT_T3_PL_1	MT	T3	PL	1	400	896 245	29 748	6	16 312	309
IM_T1_PBL_1	IM	T1	PBL	1	2 200	5 344 878	33 943	23	5 308	195
MRK_T1_PBL_1	MRK	T1	PBL	1	5 336	2 709 667	33 059	16	4 212	361
MRK_T3_Bmem_1	MRK	T3	Bmem	1	50 000	1 507 115	14 740	16	2 708	1 490
MRK_T3_Bmem_2	MRK	T3	Bmem	2	50 000	21 137 727	90 400	32	4 211	2 198
MRK_T3_PBL_1	MRK	T3	PBL	1	1 000	1 413 106	7 541	23	812	41
MRK_T3_PBL_2	MRK	T3	PBL	2	1 000	3 264 160	10 592	23	1 334	42
MRK_T3_PL_2	MRK	T3	PL	2	200	796 898	4 262	23	708	9
D01_T1_Btot_1	D01	T1	Btot	1	219 500	3 313 035	789 196	3	139 368	48 527
D01_T1_Btot_2	D01	T1	Btot	2	218 000	2 751 033	708 236	3	108 404	39 786
MT_T3_pbmc_1	MT	T3	pbmc	1	-	62 807	14 850	1	13 460	7 787
MT_T3_pbmc_2	MT	T3	pbmc	2	-	187 107	5 357	8	3 670	2 823
D01_T3_pbmc_1	D01	T3	pbmc	1	-	322 891	60 352	1	51 894	24 715
D01_T3_pbmc_2	D01	T3	pbmc	2	-	283 987	84 417	1	71 370	38 358
					Total	143 099 899	10 641 534	-	4 067 593	836 938

Supplementary Data SD2.

Isotype frequencies per sample

ID	Isotype frequency by number of unique nucleotide sequences (clonotypes)					Isotype frequency by number of identified IGH cDNA molecules				
	IgM	IgD	IgG	IgE	IgA	IgM	IgD	IgG	IgE	IgA
AT_T1_Bmem_1	62,60%	10,50%	8,70%	0,00%	18,10%	66,22%	6,99%	7,17%	0,16%	19,46%
AT_T1_PBL_1	4,03%	0,54%	32,96%	0,12%	62,35%	5,67%	0,07%	22,18%	0,11%	71,99%
AT_T1_PL_1	7,90%	0,50%	36,30%	1,70%	53,60%	15,85%	0,16%	22,03%	0,99%	60,98%
AT_T2_Bmem_1	53,00%	11,10%	12,90%	0,00%	22,90%	60,20%	6,80%	9,30%	0,10%	23,70%
AT_T2_Bmem_2	54,70%	10,10%	12,10%	0,00%	23,00%	57,20%	6,60%	9,30%	0,00%	27,00%
AT_T2_PBL_1	5,10%	0,43%	30,12%	0,22%	64,13%	7,01%	0,24%	20,55%	0,07%	72,14%
AT_T2_PL_1	8,60%	0,44%	31,05%	0,73%	59,18%	12,13%	0,08%	19,78%	0,43%	67,58%
AT_T3_Bmem_1	48,40%	7,00%	13,70%	0,00%	30,90%	52,80%	4,40%	9,50%	0,00%	33,30%
AT_T3_Bmem_2	51,30%	5,70%	11,00%	0,00%	31,90%	47,50%	4,30%	9,40%	0,00%	38,80%
AT_T3_PBL_1	3,98%	0,48%	23,73%	0,16%	71,66%	5,69%	0,01%	17,05%	0,24%	77,00%
AT_T3_PBL_2	5,70%	0,00%	20,30%	0,10%	73,90%	7,20%	0,00%	13,00%	0,00%	79,70%
AT_T3_PL_1	16,60%	0,00%	23,60%	0,00%	59,90%	24,00%	0,00%	14,00%	0,00%	61,00%
AT_T3_PL_2	19,20%	0,40%	17,90%	0,40%	62,00%	25,23%	0,72%	11,54%	0,28%	62,24%
D01_T2_Bmem_1	34,36%	7,83%	24,25%	0,00%	33,56%	46,00%	4,00%	16,00%	0,00%	34,00%
D01_T2_Bmem_3	36,30%	7,50%	23,40%	0,00%	32,80%	40,50%	3,80%	16,70%	0,00%	38,90%

D01_T2_PBL_1	3,90%	0,70%	34,80%	0,00%	60,60%	5,90%	0,40%	26,40%	0,00%	67,30%
D01_T2_PBL_2	3,90%	1,10%	36,00%	0,00%	59,00%	8,10%	0,50%	26,40%	0,00%	65,00%
D01_T2_PL_1	8,90%	0,40%	29,60%	0,00%	61,00%	16,89%	0,39%	16,19%	0,00%	66,52%
D01_T2_PL_2	8,60%	1,00%	32,90%	0,00%	57,60%	17,50%	0,60%	15,30%	0,00%	66,60%
D01_T3_Bmem_1	41,30%	6,60%	16,40%	0,00%	35,70%	44,50%	4,30%	13,00%	0,20%	38,00%
D01_T3_Bmem_2	35,74%	7,10%	22,18%	0,01%	34,98%	40,70%	4,50%	17,50%	0,30%	37,00%
D01_T3_PBL_1	2,22%	0,83%	40,14%	0,14%	56,67%	3,11%	0,33%	31,93%	0,04%	64,58%
D01_T3_PBL_2	2,20%	1,00%	37,20%	0,00%	59,60%	2,00%	0,50%	37,70%	0,00%	59,80%
D01_T3_PBL_3	1,60%	0,60%	40,20%	0,00%	57,50%	1,49%	0,76%	39,58%	0,00%	58,17%
D01_T3_PL_1	4,71%	0,67%	55,56%	1,35%	37,71%	8,00%	0,10%	46,60%	1,10%	44,20%
D01_T3_PL_2	3,33%	0,44%	54,67%	0,89%	40,67%	6,37%	0,20%	44,46%	0,69%	48,28%
IM_T1_Bmem_1	53,21%	8,32%	19,69%	0,00%	18,78%	62,70%	5,20%	15,00%	0,00%	17,20%
IM_T1_PL_1	18,00%	0,00%	35,00%	0,00%	46,00%	39,30%	0,00%	29,20%	0,00%	31,50%
IM_T2_Bmem_1	67,90%	3,20%	6,00%	0,00%	22,90%	68,80%	1,90%	5,00%	0,00%	24,30%
IM_T2_Bmem_2	54,10%	6,20%	14,90%	0,00%	24,70%	59,40%	2,90%	9,70%	0,00%	28,00%
IM_T2_PBL_1	21,30%	0,20%	34,40%	0,00%	44,20%	37,00%	0,00%	16,00%	0,00%	47,00%
IM_T2_PBL_3	22,40%	0,20%	31,90%	0,00%	45,50%	36,00%	0,00%	15,00%	0,00%	48,00%
IM_T2_PL_2	20,00%	0,10%	28,50%	0,00%	51,30%	28,20%	0,00%	19,60%	0,00%	52,20%

IM_T3_Bmem_1	63,20%	6,90%	10,20%	0,00%	19,70%	63,10%	4,10%	9,50%	0,00%	23,30%
IM_T3_Bmem_2	65,90%	6,60%	8,10%	0,00%	19,40%	64,40%	4,10%	8,20%	0,00%	23,40%
IM_T3_PBL_1	4,70%	0,20%	49,60%	0,00%	45,50%	6,60%	0,00%	34,90%	0,00%	58,60%
IM_T3_PBL_2	6,10%	0,00%	51,90%	0,00%	42,10%	9,00%	0,00%	35,00%	0,00%	56,00%
IM_T3_PL_1	9,26%	0,00%	45,06%	0,00%	45,68%	12,00%	0,00%	28,00%	0,00%	60,00%
IM_T3_PL_2	11,10%	0,00%	40,10%	0,00%	48,80%	14,20%	0,00%	23,30%	0,00%	62,50%
IZ_T1_Bmem_1	74,70%	12,40%	1,90%	0,00%	11,00%	76,40%	8,20%	2,10%	0,00%	13,30%
IZ_T1_PBL_1	5,70%	0,40%	10,60%	0,00%	83,40%	6,70%	0,01%	6,12%	0,00%	87,17%
IZ_T1_PL_1	22,30%	1,20%	11,40%	0,10%	64,80%	34,97%	0,02%	5,67%	0,05%	59,29%
IZ_T2_Bmem_1	68,80%	12,10%	3,30%	0,00%	15,80%	72,60%	6,50%	2,90%	0,00%	18,00%
IZ_T2_Bmem_2	68,60%	11,70%	3,50%	0,00%	16,20%	72,20%	5,90%	3,10%	0,00%	18,90%
IZ_T2_PBL_1	8,50%	0,10%	13,00%	0,00%	78,40%	11,90%	0,00%	7,60%	0,00%	80,60%
IZ_T2_PBL_2	9,30%	0,00%	11,70%	0,00%	79,00%	11,50%	0,00%	9,80%	0,00%	78,60%
IZ_T2_PL_1	22,10%	0,62%	15,92%	0,15%	61,21%	30,55%	0,01%	10,42%	0,09%	58,92%
IZ_T2_PL_2	19,90%	0,20%	17,80%	0,20%	61,90%	27,53%	0,01%	13,93%	0,17%	58,36%
IZ_T3_Bmem_1	70,60%	10,00%	3,60%	0,00%	15,70%	76,80%	5,10%	2,40%	0,00%	15,70%
IZ_T3_Bmem_2	70,90%	9,70%	3,10%	0,00%	16,20%	76,80%	4,10%	2,00%	0,00%	17,10%
IZ_T3_PBL_1	14,10%	1,00%	10,00%	0,00%	75,00%	20,00%	0,00%	6,10%	0,00%	73,80%
IZ_T3_PBL_2	15,10%	0,40%	10,60%	0,00%	74,00%	21,70%	0,00%	6,30%	0,00%	72,00%
IZ_T3_PL_1	36,60%	2,40%	4,40%	0,50%	56,10%	44,32%	0,02%	3,30%	0,18%	52,19%
IZ_T3_PL_2	51,50%	0,40%	5,30%	0,00%	42,70%	59,10%	0,00%	2,70%	0,00%	38,20%
MRK_T1_Bmem_1	51,43%	16,87%	15,44%	0,00%	16,26%	63,99%	10,62%	9,78%	0,00%	15,61%

MRK_T1_PL_1	12,40%	5,00%	43,80%	0,00%	38,80%	19,00%	3,00%	34,00%	0,00%	45,00%
MRK_T2_Bmem_1	47,00%	11,80%	21,30%	0,00%	20,00%	48,10%	6,20%	18,20%	0,10%	27,50%
MRK_T2_Bmem_2	54,39%	12,86%	12,90%	0,01%	19,85%	54,80%	7,60%	10,90%	0,00%	26,70%
MRK_T2_PBL_1	6,30%	0,50%	25,50%	0,00%	67,70%	7,50%	0,10%	14,80%	0,00%	77,60%
MRK_T2_PBL_3	5,60%	0,20%	27,10%	0,00%	67,10%	4,70%	0,00%	17,30%	0,00%	78,10%
MRK_T2_PBL_4	4,50%	0,10%	27,20%	0,00%	68,20%	5,10%	0,00%	19,10%	0,00%	75,80%
MRK_T2_PL_1	8,91%	0,78%	39,41%	0,13%	50,78%	13,71%	0,31%	23,70%	0,08%	62,20%
MRK_T2_PL_2	8,00%	0,38%	38,62%	0,50%	52,50%	12,71%	0,34%	21,61%	0,31%	65,03%
MRK_T2_PL_3	8,11%	0,42%	40,66%	0,17%	50,63%	9,90%	0,16%	25,05%	0,06%	64,84%
MRK_T3_PL_1	7,50%	0,30%	32,50%	0,00%	59,70%	12,90%	0,20%	22,90%	0,00%	64,00%
MT_T3_Bmem_1	55,80%	6,00%	12,50%	0,00%	25,70%	53,80%	4,10%	13,20%	0,00%	28,90%
MT_T3_Bmem_2	56,40%	6,10%	12,80%	0,00%	24,60%	57,90%	3,70%	10,60%	0,00%	27,80%
MT_T3_PBL_1	3,22%	2,65%	17,99%	0,00%	76,14%	3,60%	1,30%	18,00%	0,00%	77,10%
MT_T3_PBL_2	3,20%	2,00%	22,90%	0,00%	71,80%	5,70%	1,70%	19,60%	0,00%	73,00%
MT_T3_PL_1	4,90%	2,30%	31,70%	0,00%	61,20%	6,90%	1,30%	23,50%	0,00%	68,20%