MASTER'S THESIS

MAPPING OF THE HEALTHY IMMUNOGLOBULIN REPERTOIRE OVER THE COURSE OF B CELL MATURATION BY DEEP PYROSEQUENCING

von

Hanna Lange

Committee members:

1. Prof. Dr. Hans-Jürgen Mägert
2. Prof. Dr. Christiana Cordes

..

List of contents

ABSTRACT

Healthy immunoglobulin repertoire has not been extensively evaluated reflecting in part the challenge of generating sufficiently robust data sets by conventional clonal sequencing. Deep sequencing has revolutionized the capacity to evaluate the depth and breadth of the Ig repertoire along the B cell developmental pathway, and can be used to pin point defect(s) of primary or acquired B-cell associated diseases. In this study healthy IgM and IgG repertoires were studied by 454-pyrosequencing to establish the healthy controls for diseased repertoires.

Messenger RNA was extracted from peripheral blood mononuclear cells from four healthy young adults. Amplicon library of immunoglobulin heavy chain variable region [IGH] of IgM or IgG was generated from mRNA by a RT-PCR followed by a nested PCR. IgM or IgG-specificity was determined by downstream primers Cμ15 and Cμ2, or Cγ16 and Cγ1 homologous to the IgM or IgG constant region without separation of IgM^+ and IgG^+ B cells by cell surface markers. The upstream primers were an IGHV family-specific primer cocktail. IGHM or IGHG amplicon library, 400 to 500 nucleotides covering the whole IGH, was then gel purified and submitted for 454-pyrosequencing. An average number of 7,100 quality sequences were obtained for each library. A novel IgSEQ software developed by us for automated analysis of IGH pyrosequences through IMGT/V-QUEST and IMGT/JunctionAnalysis was applied to obtain information about somatic hypermutation [SHM], use of IGHV, IGHD and IGHJ alleles and IGHD reading frame, length and clonality in complementarity determining regions [CDR], and junctional modifications. The program ESPRIT was used to evaluate the extent of biodiversity.

IgM repertoire is significantly different from IgG repertoire in healthy young adults. A higher percentage of sequences with SHM was observed in IgG [99%] compared with IgM [98%]. IgG sequences contain more non-silent mutations in CDR1, CDR2 and FR3 regions than IgM sequences. When comparing sequences with SHM, IgM sequences showed significantly greater fraction of sequences with nucleotide [N] insertion in V_H-D-J_H junctions. Frequency distribution of CDRH3 length was Gaussian-like in IgM repertoire but relatively variable among individuals in IgG. Although used preferentially in both IgM and IgG, IGHV3 was expressed at lower frequency in IgG than in IgM. There was an increased use of IGHV1 and IGHV6 in IgG in comparison with IgM. IgM repertoire is significantly different from IgG repertoire in healthy young adults. IgG antibody repertoire comprise more diversity than IgM mostly due to acquisition of greater extent of SHM.

1 Introduction

The immune system is an amazing means of our body to defend countless pathogens that we are continuously exposed to. Every day we come into contact with many non-self molecules and yet become ill only rarely. How does our body protect us? This is one of many questions that are addressed to the research field of immunology.

The underlying study focused on one of the key players of the immune system: the B cells and their immunoglobulins. B cells are a cell type which is able to recognize numerous antigens (Ag) with its specific cell surface–bound immunoglobulin, the so called B cell receptor (BCR). The way of creating such a great amount of Ag specificity in their BCR has been of researchers' interest since decades. Even after the Japanese biologist Tonegawa was honored with the Nobel Prize in 1987, for unraveling the secret of antibody diversity, there are still enough mechanisms on the molecular and cellular level that need to be explored.

1.1 Immunoglobulins

An antibody molecule is the secreted form of the B cell receptor with an identical structure, except for a small region of the constant heavy chain. In the case of the BCR, the C terminus contains a hydrophobic sequence that enables the BCR to be anchored into the cell membrane of the B lymphocyte. For the soluble antibody molecule in contrast the C terminus is hydrophilic (Rogers *et al.* 1980). The antibodies functions can be explained by three principles: neutralization, opsonization and complement activation. Neutralization explains the event of binding a foreign antigen, e.g. a toxin, and thus hinders the interaction with host cells. The binding of immunoglobulins to foreign antigens also marks them for further degradation by macrophages. This process is referred to as opsonization. Complement activation represent/explains the involvement of antibodies in the innate immunity.

The immunoglobulin molecule consists of two identical heavy chains which are joined with 2 identical light chains by disulfide bonds and non-covalent interactions. The presence of two identical heavy and light chains gives the antibody the ability to simultaneously bind two antigens at their specific antigen-binding sites, also referred to as *Fab fragments* (Silverton *et al.* 1977; Edelman 1991). The region where both arms of that Y shaped molecule meet is called hinge region. The hinge region is responsible for the flexibility within the molecule. Altogether the whole antibody molecule has a molecular weight of approximately 150 kDa.

There are five major antibody classes: IgM, IgD, IgG, IgA and IgE that differ in their structure as well as their functions. Every isotype contains a different heavy chain C terminal region. Both, the heavy and the light chains consist of one variable (V) and at least one constant (C) domain (Figure 1). The variable domain of both chain types varies greatly in their amino acid composition between antibodies. The constant region determines the antibody's effector functions. Depending on the isotype, there are three to four constant domains (C_H) in the heavy chain, which are numbered from the amino terminal to the carboxy terminal end (C_H1, C_H2, etc.) (Kabat 1982; Davies & Metzger 1983). The variability of the variable regions is not equally distributed throughout the regions. There are six hypervariable regions which exhibit a higher degree of variability, the so called complementarity-determining regions (CDR). The three CDR's of the light chain and those of the heavy chain together create the antigen-binding site, as their loops are brought into close contact in the folded molecule. In fact, they determine the specificity of the antibody. Each CDR is followed by a less variable framework region (FR). The complementarity-determining region 3 (CDR3) of the heavy chain varies most extensively in length because it is encoded by V, D and J gene segments which are recombined in a process called VDJ recombination. It also displays the highest degree of diversification due to junctional modifications that are created in the event of somatic recombination (Tonegawa 1983; Wu *et al.* 1993). This process will be explained in detail in the following section.

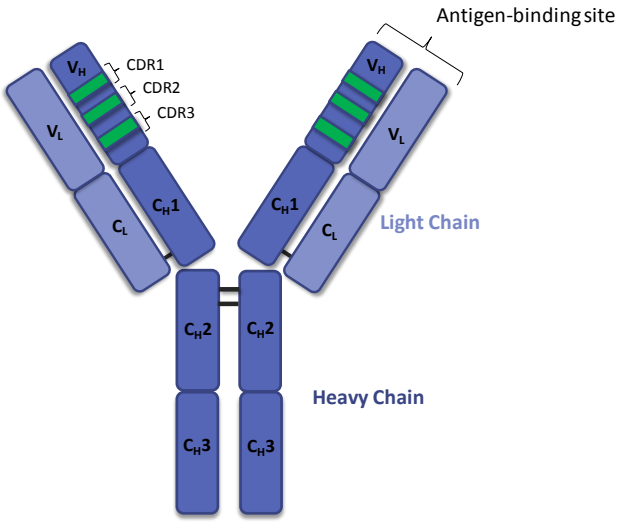


Figure 1: Schematic structure of an IgG antibody molecule. The two heavy chains consist of three constant (C_H) and one variable (V_H) domain. The light chains are composed of one constant and one variable domain. The six hypervariable regions of both the V_H and the V_L form the Antigen-binding site. The most hypervariable region lies in the variable domain of the heavy chain and is called CDR3. Modified from (Kumagai & Tsumoto 2001).

1.2 The generation of B cell receptor diversity

The diversity of the B cell receptor and likewise for secreted antibodies is created by unique genetic mechanisms throughout the development of B lymphocytes. There are several factors and mechanisms involved in the diversification process. The first and probably most important reason for the enormous diversity which can be found in the B cell receptors is the existence of multiple gene segments, which encode the variable region of the heavy and light immunoglobulin chains.

The second factor yielding to a highly diverse BCR repertoire takes place at the very beginning of the lymphocyte development in the bone marrow and comprises the assembly of gene segments mentioned above to form an exon coding for the variable heavy and light chains. This molecular mechanism is called somatic recombination, or V-D-J recombination in the case of a heavy chain. Further diversity is created by the combination of different light with different heavy chains as well as from junctional variability. Finally the process of somatic hypermutation induces point mutations in the DNA sequences that encode for the variable region of the antibody and is therefore able to change its amino acid composition. These factors will be explained in detail in this chapter.

The immunoglobulin variable heavy chain locus is composed of a set of gene segments. There are multiple copies for each type of gene segment. The variable part of the heavy chain of an antibody molecule is encoded by three different gene segments: the V_H, D_H and the J_H gene segments (Tonegawa 1983) (Figure 2). Most of them are gathered in a definite area which is referred to as *cluster* with lengths of up to one (or more) megabase(s) on the chromosome 14q32.3 (Cox *et al.* 1982). Approximately 38-46 functional genes belong to the variable heavy chain locus (V_H). A cluster of 23 functional diversity gene segments (D_H) lies between the V gene segments and the cluster of 6 joining gene segments (J_H). There are also several gene segments belonging to the light chain locus. All of these germline gene segments are thought to evolve from conversion, duplication and diversification (Fukui *et al.* 1983; Lee *et al.* 1993). The heavy chain V gene segments can be grouped into 7 families (IGHV) due to sequence homology. V families can further be grouped into four clans, from which clan IV only contains one pseudogene. A clan comprises gene subgroups that appear to be related on phylogenetic trees. The subclasses IGHV1, IGHV5 and IGHV7 belong to clan I. Clan II includes IGHV2, IGHV4 and IGHV6 families. Subgroup IGHV3 shares the least sequence homology with the other families and therefore belongs to a single clan (Cook & Tomlinson 1995; Matsuda *et al.* 1998; Pallares *et al.* 1999).The 23 functional gene segments of the

human diversity gene locus have lengths of 11 to 37 nucleotides. They are further on classified into seven families (Tomlinson *et al.* 1994; Corbett *et al.* 1997). Due to the addition and deletion of nucleotides during the event of somatic recombination, the D_H gene can occur in six different reading frames (RF), three of them arising from an inverted D_H segment use. It has been reported that those RFs that arise from an inverted D_H gene segment and D-D fusions are practically never observed in healthy human subjects (Ohm-Laursen *et al.* 2006). Considering only productive rearrangements with forward-facing RFs, there is a bias towards one RF. Indeed different theories in literature speculate how and why one RF is preferred over the others. Reading frame usage appears to be selected after the V-D-J rearrangement process occurred, when a B cell with a certain specificity for an antigen is clonally expanded. It is also suggested that individual D_H segments favor different reading frames and that the RF is thus evolutionary conserved (Briney *et al.* 2012; Benichou *et al.* 2013).

The IGHJ gene locus consists of 9 gene segments, all arranged in a single cluster. Only 6 gene segments are functional however (Ruiz *et al.* 1999). The heavy chain locus also contains a large cluster of C_H genes, which encode for the constant heavy chain of the antibody molecule. The most used haplotype in humans contains 9 genes in the IGHC locus, namely μ, δ, γ_3, γ_1, α_1, γ_2, γ_4, ε and α_2 (Flanagan & Rabbitts 1982).

While the V region of the heavy chain is encoded by clusters of the three segments namely V_H-D-J_H, the variable region of the light chain does only contain V and J gene segments, referred to as V_L and J_L. Besides, the arrangement of these segments in the germline DNA depends on the type of light chain locus. There are two distinct types of light chains that differ in the organization of their gene segments. Only one of these two types is used to be assembled in one immunoglobulin molecule. There is no combination of both types in one antibody. The lambda light chain locus is assembled by a set of V_L gene segments followed by a set of J_L segments each linked to a C_L gene. The kappa light chain locus by contrast contains a set of V_L gene segments followed by a set of J_L and then by a single C_L gene segment. The two light chain types are not equally used in the human antibody repertoire. The average κ to λ in humans is 2:1. Distortions in this ration could be used to detect defects in the B cell development (Frippiat *et al.* 1995; Tomlinson *et al.* 1995; Williams *et al.* 1996).

Having described the structure of the immunoglobulin heavy chain locus (IGH) leads to the question of how the B cell receptor is being generated out of those gene segments that are assembled in the germline DNA. Before limited numbers of gene segments were found to encode the immunoglobulin protein it was first believed that each BCR was encoded by a

separate gene. In fact however, every BCR is encoded by the combination of a V_H, D and J_H gene which have been chosen from the gene segment clusters belonging to the germ line DNA. This process is called somatic recombination.

The process of somatic recombination takes place in the bone marrow when a hematopoietic stem cell is becoming an actual B cell. The rearrangement and combination of different gene segments is made possible by the presence of noncoding DNA sequences that flank every single gene segment. These regions are referred to as recombination signal sequences (RSSs). Recombination signal sequences consist of three regions: the highly conserved heptamer, followed by a spacer and a conserved nonamer DNA sequence. The sequence of the spacer is variable, in contrast to its nucleotide length which is either 12 bp or 23 bp long. Gene segments that are flanked by a 23 bp spacer can only be joined to a gene segment flanked by a 12 bp spacer. Following this 12/23 rule V_H gene segments can never be joined with J_H gene segments, because both of them are flanked by 23 bp spacers. There is an enzyme complex called the V(D)J recombinase that carries out the necessary steps to join the segments. First the two enzymes RAG-1 and RAG-2 recognize the signal sequences and bring the segments together according to the 12/23 rule. RAG-1 and RAG-2 are specific lymphoid enzymes that are only expressed in the developing lymphocytes to mediate the recombination event (Nagaoka *et al.* 2000). They are encoded by the recombination-activating genes 1 and 2 respectively. After aligning the two gene segments, the endonuclease activity of the RAG complex creates single-stranded nicks right 5’ behind the RSS of the recombining gene. The OH group overhang which is created after the cleavage at the 3’ end of the gene will then react with a phosphodiester bond of the opposite DNA to form a hairpin loop. The blunt ended signal ends are then joined with the help of the ubiquitous enzymes of the recombinase complex Ku70:Ku80 and a DNA ligase IV that creates a piece of extrachromosomal DNA which gets lost during cell division. The coding ends however are joined in a slightly different mechanism. Also the heterodimer Ku70:Ku80 binds to the aligned ends and forms a ring so that DNA-dependent protein kinase is recruited. The endonuclease Artemis is further on activated by phosphorylation and opens the hairpin by creating nicks at various positions, creating a nucleotide overhang that originally was complementary in the double strand and is therefore called palindromic. This nicking is therefore a means to create sequence variability. At the cut end the enzyme terminal deoxinucleotidyl transferase (TdT) adds randomly non-template encoded nucleotides while at the same time DNA repair enzymes with exonuclease activity delete nucleotides from both ends of the DNA that do not pair. After this modification

the two gene segments are finally joined with the help of a DNA ligase IV (Alt *et al.* 1992; McBlane *et al.* 1995). This process thus creates diversity in the joining of the combined gene segments and makes the joining an important feature to look at when estimating the biodiversity of a given memory B cell repertoire. The fact that there are multiple clusters of genes in the germline DNA and that these can be combined randomly creates already two steps to increase the diversity and is called combinatorial diversity. Additionally the V(D)J recombination event increases the number of events to create diversity by the modification of the junction, which is referred to as junctional diversity. The final event that can shape the BCR repertoire is called somatic hypermutation (SHM), a process that introduces point mutations into the V region genes of mature B cells in the secondary lymphoid organs (Muramatsu *et al.* 1999; Odegard & Schatz 2006). The process mostly proceeds in the germinal centers of secondary lymphoid tissues. Germinal centers (GC) are distinct regions in the follicles that are predominantly seeded with B cells and to a smaller extent with T cells (Ramiscal & Vinuesa 2013). The initiation of SHM in B cells happens with expression of the enzyme activation-induced cytidine deaminase. This enzyme, also called AID, is only expressed in activated B cells, which means B cells that have already encountered antigen (Tomlinson *et al.* 1996). AID modifies only single-stranded DNA, thus in the process of transcription where the DNA is temporarily opened, AID causes a nucleophilic attack on the cytosine ring on both sides of the DNA strand which results in its deamination and forms an uridine (Maul & Gearhart 2010). Since uridine is a foreign base to DNA it causes a mismatch with the guanosine of the opposite DNA strand and thus activates in the mismatch repair mechanism DNA repair enzymes to remove the whole uridine nucleotide and several other nucleotides adjacent to them. The other way of repairing the mismatch is called base-excision repair mechanism. In this case the repair enzyme uracil-DNA glycosylase (UNG) only removes the uracil base from the uridine nucleotide, creating a gap with absent base. In the following DNA replication a new nucleotide will be added in the opposite strand and another enzyme will splice out the abasic nucleotide. The repair of the gap further on leads to gene conversion, which is an event unlikely to happen in humans. Since AID causes carcinogenesis, when active on other than lymphocyte immunoglobulin loci, expression is highly restricted to only centroblast B cells (Li *et al.* 2004). A centroblast is an activated B cell type that undergoes rapid proliferation in the GC. Somatic hypermutation can be explained as a means of the body to enhance the affinity maturation of B cells with specificity for a certain antigen. The mature B cells, which became activated by antigen, most likely

undergo another stage of further diversification in the germinal centers where they compete with other B cells with specific BCR for an encountered antigen. Only B cells with an increased affinity for that antigen continue to receive survival signals from helper T cells and follicular dendritic cells and are thus selected for clonal expansion (Smith *et al.* 1997; Shlomchik & Weisel 2012).

In case the AID enzyme deaminates switch regions of the constant locus, class switch recombination is initiated. Class-switch recombination does not increase the specificity of the B cell receptor repertoire *per se*, but plays an important role in the diversification of its functions. Class-switching allows the BCR to acquire distinct effector functions which are mediated by the type of the constant heavy chain. Isotype-switch only occurs in activated mature B cells. The VDJ segment of the IGH is translocated from its original position upstream of the C_H gene and placed in front of a different C region. It is important to note that the selection for the new C region does not proceed randomly; rather it is the result of various cytokines released from helper T and other cells upon antigen recognition. They enhance the transcription of that constant region gene which BCR isotype should be expressed later. Since AID does only induce mutations in single-stranded DNA, it will be targeted to the actively transcribed regions. The prerequisites of this mechanism are the switch regions which are located in the intron sequence upstream of every constant region gene, except for the Cδ (Manis *et al.* 2002). The Cδ.gene is located right downstream of the Cμ gene and shares the same switch region. The switch occurs between the switch region of the Cμ gene and the switch-region of the gene of the new isotype. Transcription factors that are released from the cell upon cytokine stimulation, bind to the intronic (I) region promoter, a region upstream of each switch region, and initiates transcription. The start of transcription recruits the AID enzyme that now deaminates various cytidine residues which results in the formation of uracil (Xue *et al.* 2006). Two other enzymes further convert the uracils to nicks on both DNA strands. Since the two switch regions that now contain multiple breaks lie in two distinct locations the DNA break repair machinery joins the two switch regions and excises all DNA that lies in between these regions. The rearranged VDJ segment is now located a few kilo bases upstream of the new constant gene. After transcription, the region between is spliced out and the immunoglobulin can be expressed on the cell surface of the B cell (Stavnezer *et al.* 2008). This mechanism ensures the body to create antibodies with the same specificity while changing its effector properties (Snapper *et al.* 1997).

The total diversity of the BCR repertoire is achieved by several distinct molecular mechanisms. The simple fact that the genetic information for the antibody variable region is inherited in several gene segments gives rise to 6,300 different potential recombinations for the heavy chain and approximately 255 for the light chains. These numbers can be multiplied by 1000 to include the diversity that is introduced by junctional modifications. Further the repertoire is diversified by combining two of several light chains to two heavy chains (Figure 2). After the B cells have encountered their cognate antigen they will possibly undergo somatic hypermutation and class-switch recombination in the secondary lymphoid tissues. Thus in total, approximately $2x10^{12}$ different antibodies can be created.

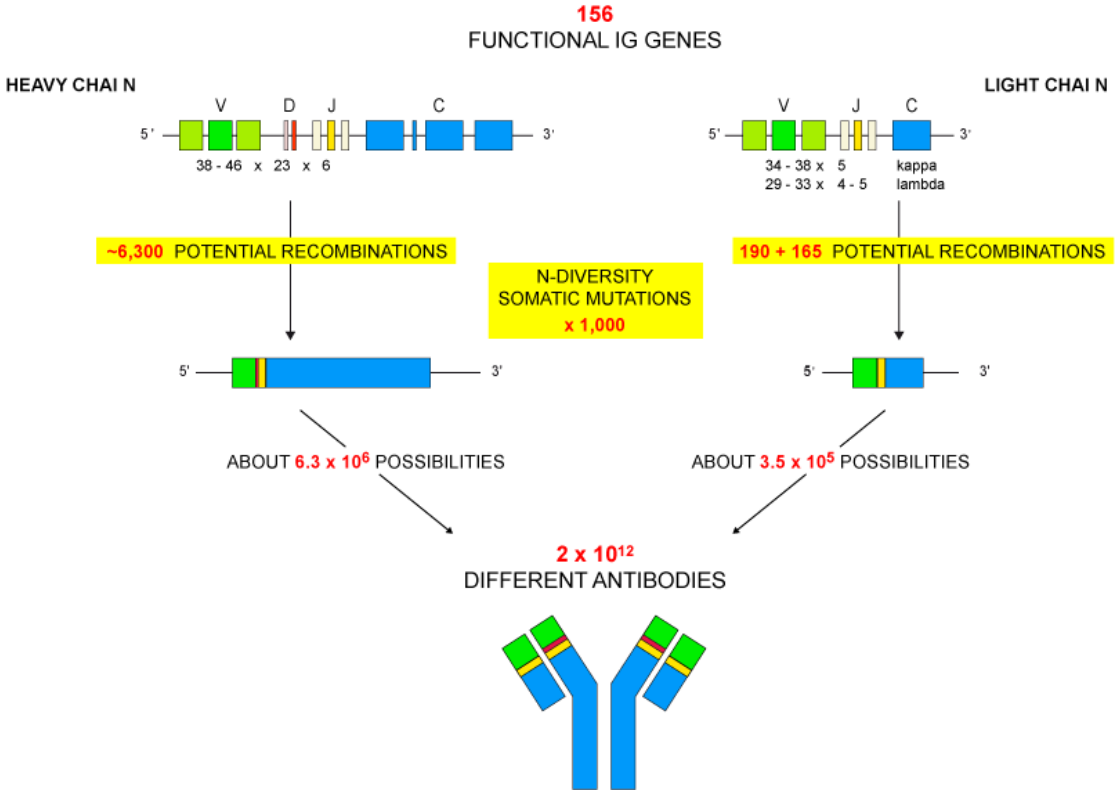


Figure 2: The molecular generation of antibody diversity. The occurrence of multiple variable, diversity and junctional gene segments leads to a great combinatorial and junctional diversity in the B cell receptor repertoire that makes it possible for the body's immune system to respond to an enormous number of different (foreign) antigens. The final estimated number of different antibodies each with different specificities is estimated to be $2x10^{12}$. Figure from www.IMGT.org.

1.3 The developmental pathway of B lymphocytes

Humoral immunity is initiated by naïve mature B cells in the peripheral blood, which enter the secondary lymphoid tissues to become activated. Approximately 5-15% of circulating lymphoid cells are B lymphocytes (Maddaly *et al.* 2010). Naïve B cells are the progeny of hematopoietic stem cells from the bone marrow (Busslinger 2004). Starting from a progenitor cell the B lymphocyte undergoes several stages of maturation until it leaves the bone marrow as immature B cell and migrates to peripheral lymphatic organs. The process of B cell maturation requires the interaction with bone marrow stromal cells for necessary signaling and is also referred to as antigen-independent phase of development (Hystad *et al.* 2007). The bone marrow dependent stages of the B cell development are highly dependent on the functional assembly of gene segments and involve the expression of various surface markers as well as transcription factors (LeBien 2000). Until the early pro-B cell stage the DNA of the B cell is still in germline configuration. The first B cell stage that arises from a common lymphoid progenitor (CLP) is the pro-B cell. Characteristic for the pro-B cell stage is the rearrangement of the heavy chain locus, namely the recombination of a D gene with a heavy chain J_H gene segment (Figure 3). For the required proliferation signals the B cell binds its receptor tyrosine kinase Kit (CD117) to the stem cell factor (SCF) expressed on the surface of the bone marrow stromal cell. The late pro B cell stage is characterized by the rearrangement of the V_H gene segment to the previously rearranged DJ_H segment. A productively rearranged heavy chain locus leads to the expression of a pre-B cell receptor (pre-BCR) inside the cell at the large pre-B cell stage. The pre-B cell receptor forms of the expressed μ-heavy chains with surrogate light chains (λ5, Vpre-B), because the light chain loci gene rearrangement has not taken place yet (Pieper *et al.* 2013). The pre-BCR plays an important role for the formation of only one antigen specificity per B cell, a process called allelic exclusion (Löffert *et al.*). Thus, at the large pre-B cell stage the expression of the RAG1/2 proteins, which are required for VDJ recombination, is reduced to ensure no further rearrangement of the heavy chain locus. As the B cell continues proliferation it undergoes the rearrangement of the light chain locus in the small pre-B cell stage. At this time the RAG1/2 proteins are again expressed to enable the recombination of the light chain gene segments. So from each large pre-B cell many B cells with certain antigen specificities are developing. A productive light chain gene rearrangement eventually leads to the expression of an IgM BCR of the immature B cell, which forms a BCR complex with Igα/Igβ proteins on the cell surface (Perez-Andres *et al.* 2010). Igα/Igβ heterodimer is needed in the BCR complex to transduce signals from the BCR for its

interaction with intracellular tyrosine kinases. The immature B cell is further tested on self-reactivity before it leaves the bone marrow to become a mature B cell, expressing also IgD on its surface (Casellas *et al.* 2001). This important step generates a self tolerance which prevents the B lymphocytes from encountering self-antigens. Each B cell expresses on average 1.5×10^5 membrane-bound antibody molecules (Maddaly *et al.* 2010). Immature B cells leave the bone marrow via sinusoids and are transported with the blood stream to the spleen or other secondary lymphoid tissues. The migration of the B lymphocytes into the spleen or the lymph nodes is initially directed by the release of the cytokines CCL18 and CCL19 from dendritic cells (Ansel & Cyster 2001). Since not all potential self antigens are present in the bone marrow, there is a second phase of testing a B cell's self-reactivity in the periphery. It could be shown in a mouse model that B cells that strongly react with self antigen by cross-linking are removed from the B cell pool by clonal deletion (Goodnow *et al.* 1989). As soon as the B lymphocytes migrate to secondary lymphoid organs the subsequent development is assigned as antigen-dependent phase. There are two different ways for B lymphocytes to be activated by an antigen: the activation upon encounter a T-independent (TI) antigen or by a T-dependent (TD) antigen. Usually the exposure to microbial lipopolysaccharides or bacterial flaggelin elicits a TI activation of B cells. This non-specific activation involves both mature and immature B cells and leads to the expression of only IgM, showing lower antigen affinity. The B cell response to T-dependent antigens requires the help of T helper cells and their cytokines. Soluble protein antigens cause a cross-linking of the B cell's antigen receptors and therefore attract antigen-specific T cell help (Maddaly *et al.* 2010). It is generally believed that mature B cells that have encountered antigen migrate to the germinal centers of the secondary lymphoid tissues and undergo somatic hypermutation as well as class switching. The germinal center stage is known as a phase of rapid proliferation and the continuously competition for survival signals from follicular dendritic cells and T helper cells. B lymphocytes which acquire SHM are expressing rendered surface immunoglobulin with an either increased or decreased affinity for their cognate antigen. Only these cells continue to differentiate into isotype-switched memory B cells or antibody secreting plasma cells that obtain an improved affinity for their antigen presented on a follicular dendritic cell or a $CD4^+$ T cell. The follicular dendritic cells provide essential survival and proliferation signals to the selected B cell. The process of selection for B cells with increased antigen affinity is called affinity maturation. B cells that received survival signals from the antigen-presenting cell

(APC) further differentiate into long-lived plasma cells or to a smaller amount into GC-derived memory B cells. Most of these memory B cells are class switched.

Another fate of B lymphocytes that have already seen antigen is the differentiation into memory B cells right after they experienced antigen exposure without the formation of a germinal center. The expression of CD40 ligand by the APC is critical for the differentiation into GC independent memory B cells. These types of memory B cells are mostly expressing IgM on their surface and do not contain SHM. Although it has been shown that populations of IgM^+ memory B cells exist that carry SHM but did not evolve after germinal center reaction, however from marginal zone B cells. The majority of switched memory B cells are derived from germinal centers. After the exposure to an unknown antigen the earliest memory B cells can be detected after 3 days, still before the germinal center form (Taylor *et al.* 2012).

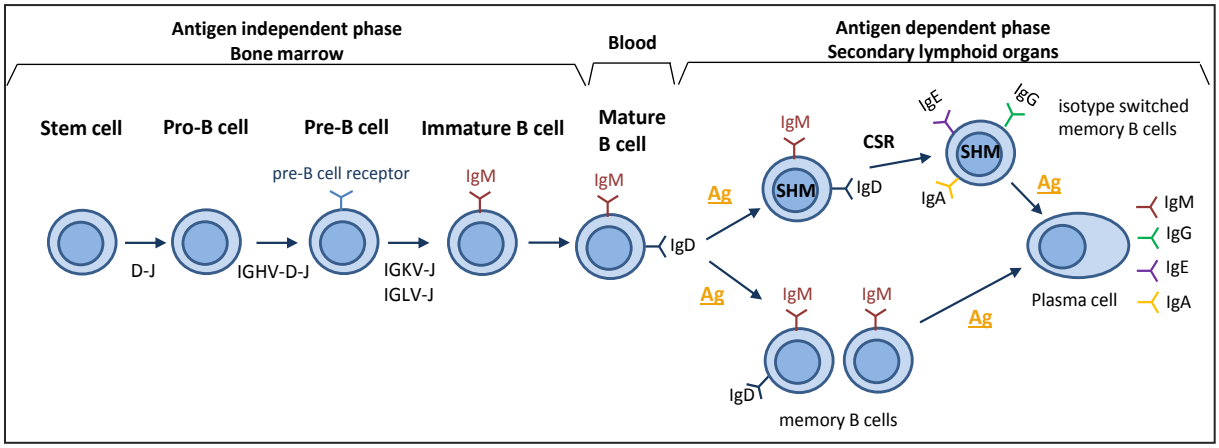


Figure 3: Developmental stages of B lymphocytes. The antigen independent stage of the development proceeds in the bone marrow, versus the antigen dependent stage takes place in the periphery. Modified from Perez-Andres et al. 2010.

1.4 Aim of this investigation

In this study the IgM and IgG transcriptome repertoires of four healthy subjects have been subjected to examination of their genetic and molecular properties with data obtained from deep sequencing. Deep pyrosequencing makes it possible to generate up to 3,000,000 reads per run (Metzker 2010) and thus has revolutionized the capacity to evaluate the immunoglobulin repertoires. Insights in the Ig heavy chain variable region (IGH) repertoire along the B cell developmental pathway might give an important baseline for further investigations of B cell associated disorders. Molecular perturbations in memory B cell repertoires could indicate diseases when comparing with the healthy repertoires.

The primary focus of this study was the contribution of genetic markers, such as V_HDJ_H combinations and the extent of SHM, to the diversity of the given repertoires. It was hypothesized that the IgG repertoire displays greater biodiversity due to its derivation from IgM^+ B cells and thus belonging to a developmentally later stage.

This investigation was conducted using mRNA from whole peripheral blood mononuclear cells (PBMC) and therefore presents a new method to discover genetic properties along the B cell's developmental pathway. The advantage of using mRNA over genomic DNA is that only productive DNA is expressed. No surface marker has been used prior to generating transcriptome repertoires, so the only way to identify the developmental stage was by looking at the presence of somatic hypermutations. Somatic hypermutations are the most precisely means to distinguish naïve from memory B cells.

2 Material and Methods

2.1 Study cohort

A total of four healthy young adults (median age: 20 years) was enrolled in this study. None of them was infected with HIV or suffered from any autoimmune or chronic disease nor received of any kind of vaccination up to 30 days prior to the study.

2.2 Cell sorting for B cell profiles

Peripheral blood mononuclear cells were stained using the whole blood lyses method (Bossuyt *et al.* 1997) and sorted with the LSR2 flow cytometer (BD Biosciences, Franklin Lakes, NJ, USA). B cells were positively selected using the monoclonal PECy7-conjugated anti-CD19 (BD Biosciences) antibody. The separation of IgM+ and IgG+ B cells was done by means of APC-conjugated anti-IgM and anti-IgG antibody (BD Biosciences). Among all subjects B cell percentages ranged from 4.0 to 10.4% of the lymphocytes.

2.3 Isolation of Peripheral blood mononuclear cells and mRNA

Peripheral blood mononuclear cells (PBMC) were extracted from peripheral blood with the aid of Lymphoprep™ density gradient medium (Fisher Scientific, Waltham, MA, USA) using a density gradient centrifugation protocol. Granulocytes and the erythrocytes sediment through the Lymphoprep™ layer due to their higher density leaving a layer of PBMC on the top of their own layer, but beneath the plasma layer. Messenger RNA was extracted from PBMC (without the isolation of IgM^+ and IgG^+ B cells) with the MicroPoly[A]Purist™ kit (Ambion, Austin, TX, USA) following the manufacturer's instructions.

2.4 cDNA synthesis and creation of amplicon libraries

To synthesize cDNA, approximately 10 ng (equivalent to 5×10^6 B cells) of extracted mRNA was used as a template for the reverse transcriptase (RT) PCR.

The first round primers enabled the amplification of the whole rearranged immunoglobulin sequence, as the upstream primer binds to the first framework region of the heavy chain variable antibody sequence and the downstream primer to the corresponding constant part of the sequence (Figure 4).

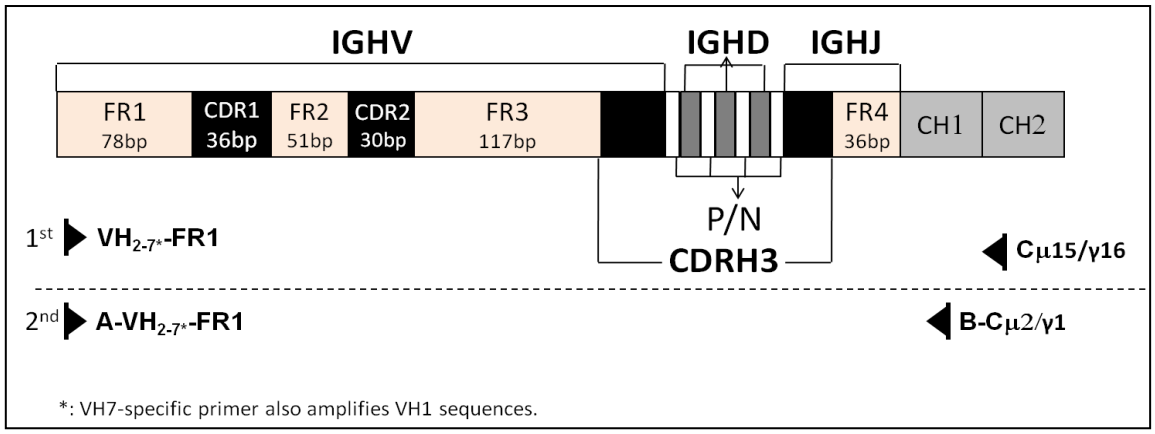


Figure 4: Primers used to amplify the IgM and IgG variable heavy region [IGHM/IGHG] for library construction. The upstream primer cocktail consists of six primers which are specific to IGHV1 to IGHV7, and located in FR1. The downstream primers are located in the constant region of IgM/IgG. PCR products ranged from 400 bp to 500 bp covering the entire heavy chain variable region [IGH]. P/N symbolizes the non-germline encoded and palindromic nucleotides that are added in the junction in the process of somatic recombination.

Specificity for the IgM and IgG transcriptome was archived by using downstream primers homologous to the constant region of the immunoglobulin isotype (Gokmen *et al.* 1998). The Cμ15 primer was used for the amplification of the IgM sequences and the Cγ16 primer for the amplification of the IgG repertoire (Table 1).

Table 1: Downstream primers used in the RT-PCR. The Cμ15 primer binds to the IGHM and the Cγ16 primer to the IGHG constant region. Both were obtained from Invitrogen (Carlsbad, CA, USA).

Primer name	Sequence 5' -> 3'
Cμ15	GACGAAGACGCTCACTTTGGG
Cγ16	CACCTTGGTGTTGCTGGGCTTGT

The upstream primer cocktail contained 6 different primers with binding specificity to all IGHV families (Table 2). To check for binding bias, every primer was tested for its binding capacity in a previous investigation. It could be seen that each primer displayed similar binding capacities to the IGHV family sequences.

Table 2: Forward primer cocktail used in the RT-PCR for IgM and IgG antibody sequences. Primers bind to the framework 1 antibody sequence and enable the amplification of all seven IGHV families. All primers were obtained from Invitrogen.

Primer name	Sequence 5' -> 3'
VH2-FR1	CTCTGGTCCTACGCTGGTGAAACCC
VH3-FR1	CTGGGGGGTCCCTGAGACTCTCCTG
VH4-FR1	CTTCGGAGACCCTGTCCCTCACCTG
VH5-FR1	CGGGGAGTCTCTGAAGATCTCCTGT
VH6-FR1	TCGCAGACCCTCTCACTCACCTGTG
VH7/VH1-FR1	CTGGGGCCTCAGTGAAGGTCTCCTG

The RT-PCR was conducted in a 50 µl reaction volume at 50°C (30 min), 95 °C (2 min), 92 °C (30 sec), 62°C (35 sec), 72°C (1 min), 72°C (10 min) for 20 cycles with the following reagents:

Table 3: Composition of the reaction mix for the reverse-transcriptase PCR. The RT Platinum *Taq* PCR kit was obtained from Invitrogen.

Reagents	Volume per reaction (µl)	Final concentration
2x Reaction mix	25	1x
$MgSO_4$ (50 mM)	1.3	2,8 mM
VHs-FR1 cocktail (5µM/primer)	0.5	50 nM/primer
Cµ15 / Cγ16 (30 µM)	0.5	300 nM
RT Platinum *Taq* mix	1	
dH2O	11.7	
Template mRNA	10	10 ng

The second round PCR was performed to increase the yield of the first PCR products using primers with a template-independent region, which functions as adapter for the subsequent pyrosequencing step. The upstream primer mix contained the same template-specific sequences as the first round PCR primers except for the adapter sequences (Table A 1)
The second round downstream primers were located inside the (nested) first round amplicons (Figure 4) and contained labeled adapters for subsequent *pyrosequencing* (Table A 2).

The amplification of the IGH specific sequences was conducted at 94 °C (2 min), 94 °C (30 sec), 62 °C (35 sec), 68 °C (1 min), 68 °C (5 min) in a cycle with 30 repetitions.
The Q5 High–Fidelity DNA polymerase (NEB, USA) was used to ensure high-fidelity reading rates.

The reaction mix was composed of the following reagents:

Table 4: Composition of the second round PCR reaction mix.

Reagents	Volume per reaction (µl)	Final concentration
A-VHs-FR1 cocktail (5µM/primer)	0.5	50 nM/primer
TiB-MID-Cµ2 / Cγ1 (30 µM)	0.5	300 nM
Q5 High-Fidelity 2x Master Mix	25	1x
dH2O	19	
Template mRNA	5	

Prior to submitting the amplicons to the Interdisciplinary Center for Biotechnology Research (University of Florida, USA) for *454-pyrosequencing*, the PCR products were gel-purified (QIAquick Gel extraction kit, Qiagen). The deep sequencing was done with the Genome Sequencer FLX (454 Life Sciences) according to the manufacturer's protocol.

2.5 Bioinformatics pipeline

To analyze the sequencing data, a bioinformatics pipeline has been developed.

A total of 5000 -15,700 raw sequences per subject could be obtained which were further on processed through a quality control filtering step (Figure 5). Ambiguous nucleotides and sequences that failed to align properly to the reference sequences of the IMGT database or contained any sequencing errors were removed.

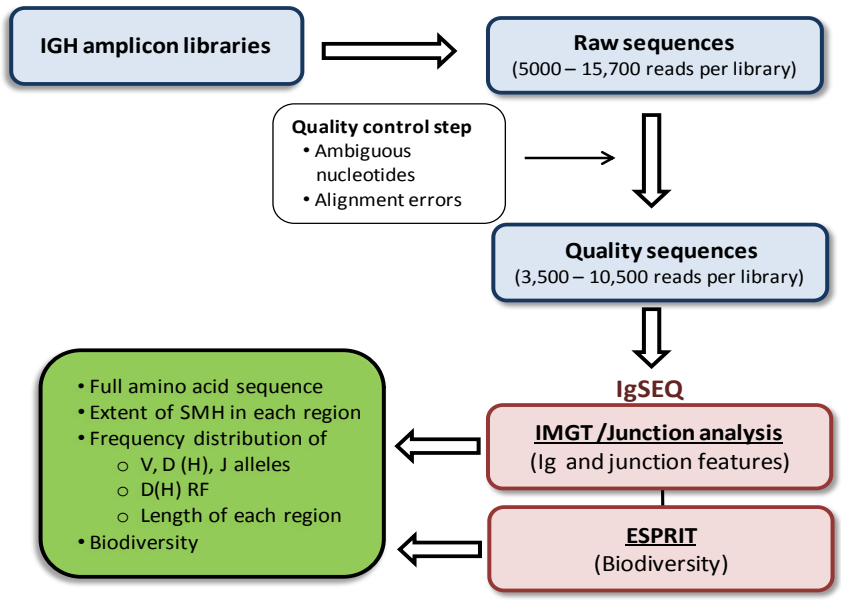


Figure 5: Bioinformatic pipeline overview. After qualifying raw sequences, quality sequences were analyzed by *IMGT/ Junction analysis* and *ESPRIT* to obtain several data (green box).

A novel IgSEQ software, developed by our group, for automated analysis through *IMGT/V-QUEST* and *IMGT/Junction Analysis* was applied to obtain key features of immunoglobulin repertoire, including somatic hypermutation (SHM), use of IGHV, IGHD and IGHJ genes, amino acid lengths of framework regions and complementarity determining regions, as well as junctional modifications (Monod *et al.* 2004; Brochet *et al.* 2008).

The numbers of silent, non-silent and total somatic hypermutations have been determined for all framework and complementarity determining regions. Sequences with one or less than one somatic hypermutation (one nucleotide difference from the reference sequences) were excluded from further investigation to prevent ambiguities (Glanville *et al.* 2009). Furthermore, all sequences harboring two or more mutations over the whole IGHV-D-J antibody sequence were classified as SHM^{+}.

The biodiversity was estimated using the program *ESPRIT*. The study includes factors yielding to biodiversity such as the extent of junctional modifications and somatic hypermutations, but also the variety which is created by V_H-D-J_H combinations.
The values of biodiversity are further influenced by the number of input cells and the distribution and richness of the combinational sequence clusters. Therefore it was weighted by the number of input B cells, to make the data comparable among individuals.

The rarefaction analysis gives information about the diversity of the transcriptome repertoire at the depth of sequences (number of sequences). The run of the curve displays an increase or decrease of the biodiversity. Deeper initial slopes and a left shift of the curve indicate an increase.
Chao1 analysis describes the maximum estimated biodiversity within the input templates (3500 - 10500 sequences). In this study the sequences were clustered at a genetic distance of 10%.

The data was combined and summarized in *Excel* (Microsoft). Subsequent graphic presentation and statistics were implemented with GraphPad *Prism* (GraphPad 5.2 Software, CA, USA). Comparisons of the amount of somatic hypermutation and the frequency of gene alleles among study groups were performed by ANOVA followed by a Bonferroni post test (Yin *et al.* 2013).

3 Results

3.1 Somatic Hypermutations

When comparing the IgM with the IgG transcriptome repertoires initially the total extent of sequences with or without somatic hypermutation was determined. Sequences that do not contain somatic hypermutations are considered to correspond to the naïve B cell population, whereas the mutated sequences are equivalent to the memory B cell population. It was found that the IgG repertoire contains approximately five times more sequences with somatic hypermutations (99.2%) than the IgM repertoire (94.7%) (Figure 6).

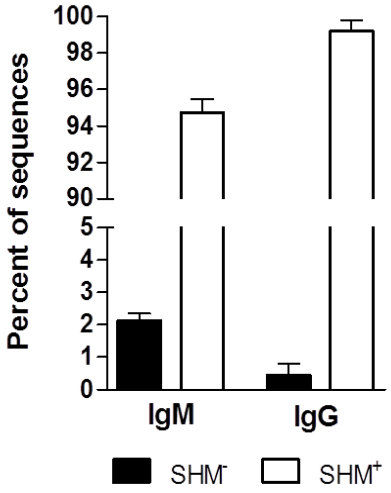


Figure 6: Sequence profile. The IgM transcriptome repertoire possesses about five times more sequences without somatic hypermutations than the IgG repertoire. Error bars indicate SEM among four individuals.

Besides the greater general extent of somatic hypermutations, the IgG transcriptome repertoire contains more nonsilent mutations (21.6 mutations per 100 bp), compared to the IgM repertoire (13.8 mutations per 100 bp) and therefore a different corresponding amino acid composition of the antibody (Figure 7A). Although the difference among both repertoires was not significant, there was a visible trend towards a greater extent of mutations in the IgG repertoire. The amount of sequences with silent mutations was approximately similar in both repertoires (4.6 mutations per 100 bp ±1.5). Drilling down the extent of mutations over the whole IGH sequence the IgG repertoire possesses significantly more silent and nonsilent mutations in the first and second complementary determining region (CDR1 and CDR2) and the third framework region (FR3) than the IgM repertoire (Figure 7B). The framework region one is the only region in the IgM repertoire that harbors significantly more somatic hypermutations than the IgG repertoire. The greatest extent of nonsilent mutations in the IgG

transcriptome was found in the CDR2 region, followed by the CDR1 and the FR1. The FR1 region also displays the greatest amount of silent and nonsilent mutations of the IgM repertoire.

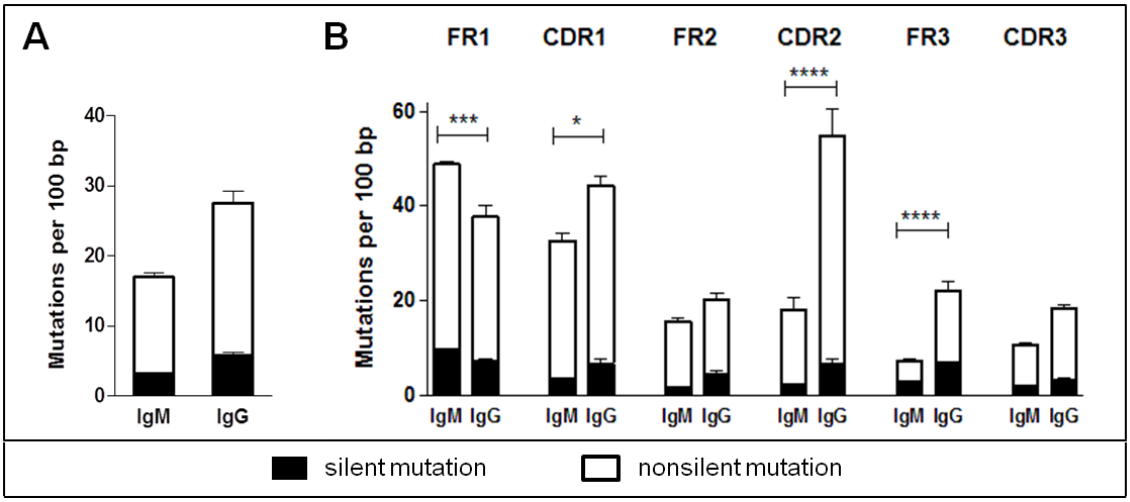


Figure 7: Silent and nonsilent SHM along the whole depth of sequences (A) and in all CDRs and FRs of IGHM and IGHG. A. The IgG transcriptome repertoire contains a greater extent of silent and nonsilent mutations than the IgM repertoire. B. Overall, the extent of either silent or nonsilent SHM is greater in FRs and CDRs in IGHG than in IGHM except for FR1 which has significantly more mutations in IGHM than in IGHG. Error bars indicate SEM among four individuals. * p<0.05, *** p<0.001, **** p<0.0001.

3.2 Frequency of IGHV, IGHD and IGHJ gene families

In this investigation, the frequency of existing IGHV, IGHD and IGHJ gene families in the mutated IgM and IgG expressing B cell repertoires was compared among the four individuals.

Every variable heavy chain immunoglobulin gene segment [IGHV] belongs to one of seven gene families (Cook & Tomlinson 1995).

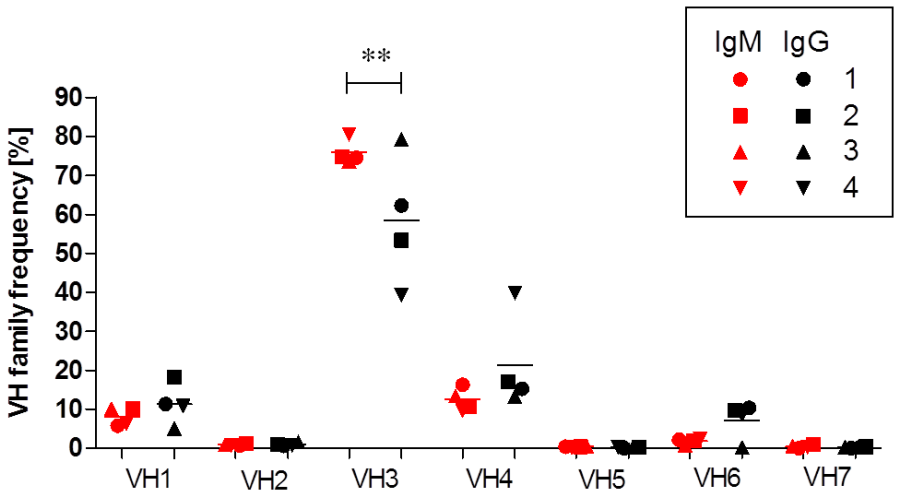


Figure 8: IGHV family usage in IgM and IgG transcriptome repertoires. IgG repertoire predominantly contained IGHV3 but at a reduced frequency and with an increase in use of IGHV4 in comparison to IgM repertoire. Each symbol represents one individual. ** p< 0.01

Figure 8 shows a strong notable bias towards the gene family VH3 (~70%) among all individuals in both the IgM and the IgG repertoire. There is also a slight bias towards gene family VH1 and VH4 among the four individuals. Comparison of the frequency of IGHV gene families between IgM and IgG repertoire showed a significantly elevated frequency of the IGHV3 segments in the IgM repertoire, and a trend towards higher frequency of IGHV4 family in the IgG repertoire. The other gene families appear consistent among all subjects. Although the total frequencies of IGHV families of the IgG repertoire are very similar to those of IgM, except for IGHV3, there is a greater variability of use among each individual; noticeable as the black data points are more spread out vertically.

Biased gene usage does not only apply to IGHV genes. IGHD genes are also unequally distributed within the examined B cell repertoires (Figure 9). The frequency of the seven IGHD families (Corbett *et al.* 1997) was similar in both IgM and IgG repertoires. The gene frequencies of various IGHD genes of the IgM repertoire are clustered together among individuals, whereas the IgG repertoire gene family usage displays some more variation from one person to another.

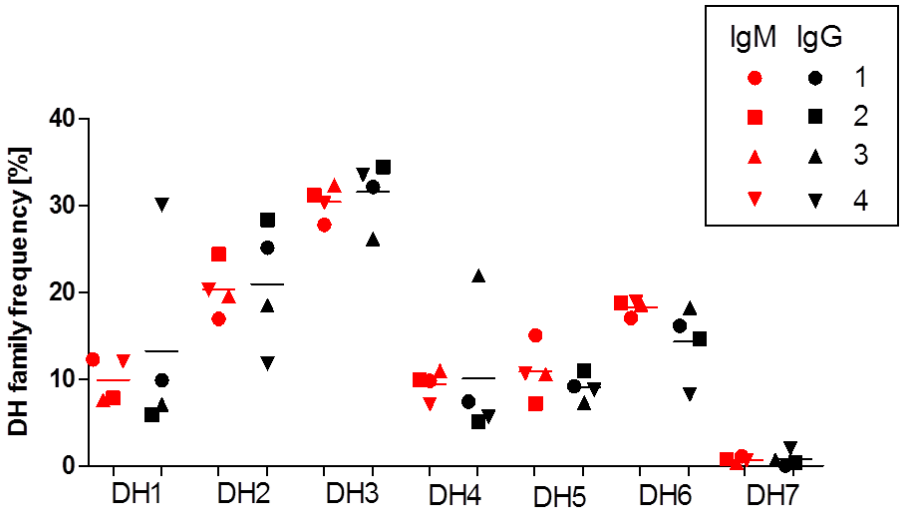


Figure 9: IGHD gene family usage in IgM and IgG transcriptome repertoires. Frequency of IGHD gene families was similar in IgM and IgG B cell repertoires and among all individuals. IGHD3 was used most in both repertoires. Each symbol represents one individual.

From seven IGHD gene families, IGHD3 showed the highest frequency (~30%) among all individuals. The IGHD7 gene segments were underrepresented in all individuals.

There is also considerable variation between the utilization frequencies of IGHJ genes (Figure 10). The IGHJ4 family genes make the greatest contribution (~50%) to both the repertoires of all examined individuals, but with a significant difference in the frequency among IgM and IgG repertoire.

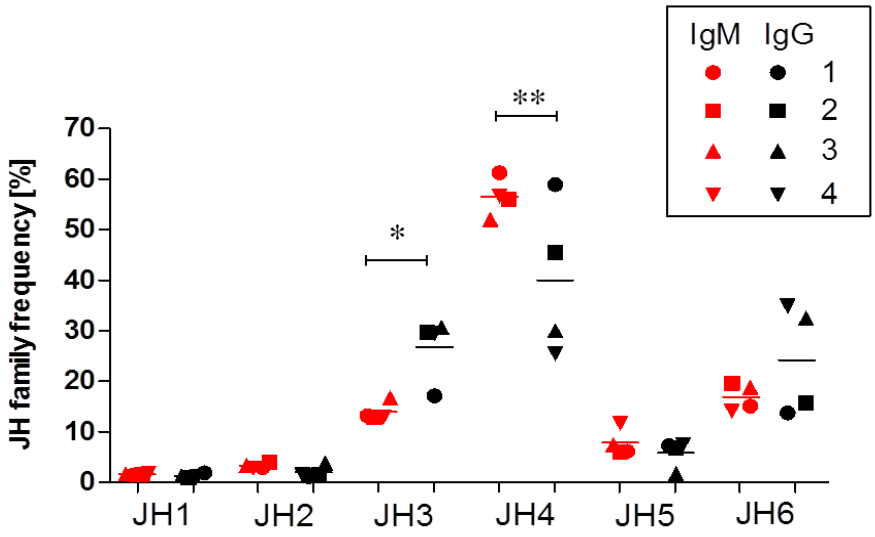


Figure 10: IGHJ gene family usage in IgM and IgG transcriptome repertoires. Except for IGHJ3 and IGHJ4 the frequency was similar in both repertoires. Although there is a significant difference in usage, IGHJ4 was the most abundant family in both repertoires. Each symbol represent one individual.

While the IGHJ4 family is present with 56% in the IgM repertoire, only 40% of gene segments in the IgG repertoire belong to the IGHJ4 family. The second most abundant IGHJ genes of the IgG repertoire belong to the IGHJ3 family, followed by those of the IGHJ6 family. The usage of IGHJ3 differs significantly between IgM and IgG repertoire. The frequency of IGHJ1, IGHJ2, IGHJ5 and IGHJ6 is similar in both examined repertoires.

3.3 D_H Reading frame usage and the extent of non-germline encoded nucleotides

It is known from previous investigations that D_H gene segments of the immunoglobulin heavy chain variable region can occur in three different forward and three inverted reading frames. Due to the addition and deletion of nucleotides between the junction of two gene segments during the process of somatic recombination. Although there are three possible forward reading frames of the D_H segment, reading frames are not equally utilized in the antibody repertoire. As the occurance of inverted RF in productive rearrangements has been considered very unlikely in humans this study focused on the use of the forward reading frames only (Ohm-Laursen *et al.* 2006).

As seen from Figure 11, in this investigation the preferred reading frame two with ~48% in both IgM and IgG repertoires. The frequency of usage of RF1 and RF3 was significantly different from that of RF2.

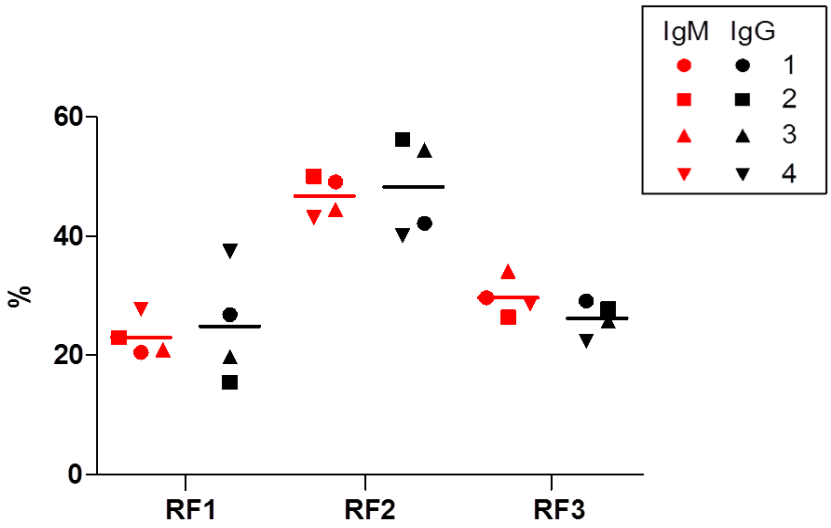


Figure 11: The D_H reading frame usage in IgM and IgG transcriptome repertoires. In both repertoires RF 2 was preferred. Each symbol represents one individual.

The addition of non-template encoded nucleotides [N] between the junction of the V_H-D and D-J_H segments, also occurs during somatic recombination of B cells. In this investigation the amount of sequences with and without N was compared between the mutated and unmutated IgM and IgG repertoires. Figure 12A displays the difference of added N between the mutated IgM and IgG repertoires. 98.7% of sequences of the mutated IgG repertoire contain added N, whereas the IgM repertoire displays approximately 2% less. In the unmutated repertoire as shown in Figure 12B the amount of sequences with N is very similar between the IgM and IgG repertoires. Although every sequence of the unmutated IgG repertoire contains N nucleotides, at least 1% of sequences lack added N in the IgM repertoire.

The extent of non-template encoded nucleotides is slightly higher in the hypermutated repertoires.

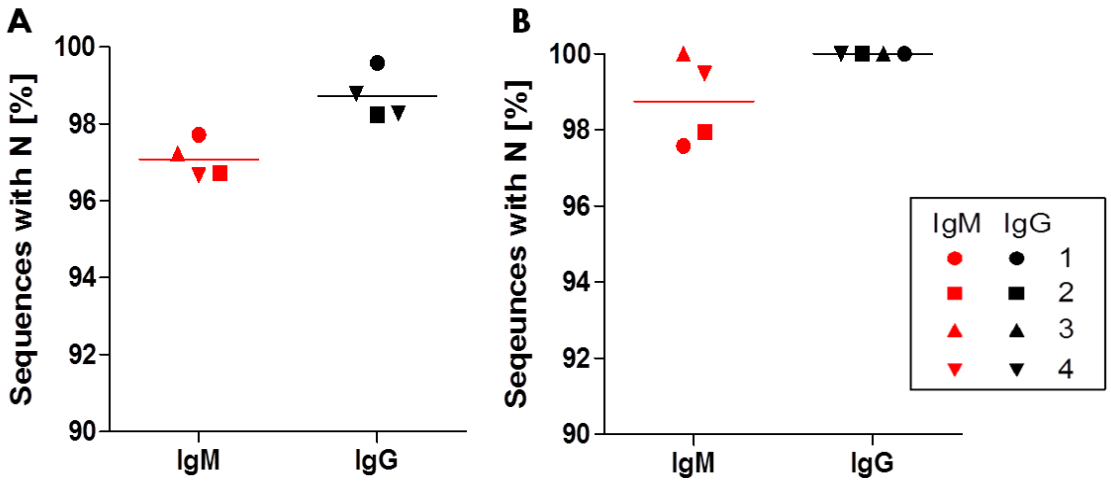


Figure 12: The frequency of sequences with non-germline encoded nucleotides in IgM and IgG repertoires. A: In the repertoires that contain somatic hypermutation, IgG sequences contain approximately two times more N than IgM sequences. B: Looking at sequences that do not contain SHM, the percentage of sequences with N was higher than in the mutated repertoires. All of the few non-mutated IgG sequences contain N. Each symbol represents one individual.

3.4 Complementarity determining region 3 length variation

The complementarity determining region 3 (CDR3) of the heavy chain is one of three highly variable immunoglobulin regions, that displays the highest degree of variability, due to its genetic composition and is therefore a marker for diversity.

Figure 13 depicts the CDR3 length distribution in the IgM (upper panel) and IgG (lower panel) transcriptome repertoires. In both repertoires the amino acid length ranged from five to 32 residues. Differences can be seen in the frequency of the peaks. The length variation of the IgM repertoire follows a bell-shaped curve with a maximum peak of 14 amino acids for all subjects, whereas the graph of the IgG repertoire displays a more irregular (diverse) distribution in which the frequencies of the preferred CDR3 lengths vary among the individuals. In the IgG repertoire subject 1, 3 and 4 preferentially use one particular CDR3 length, subject 3 the CDR3 length of 15 amino acids with a frequency of over 30% and subject 4 the length of 19 amino acids. The CDR3 length distribution for subject 2 reveals two major peaks for the CDR3 length of 10 and 18 amino acid residues with a frequency of ~16%.

In contrast to the highest peaks of the IgM repertoire which reach a frequency of 15%, the highest peaks in the IgG repertoire reach frequencies over 30%. This may be due to the

exposure of a particular antigen that causes a specific B cell clone to proliferate as a response to that antigen.

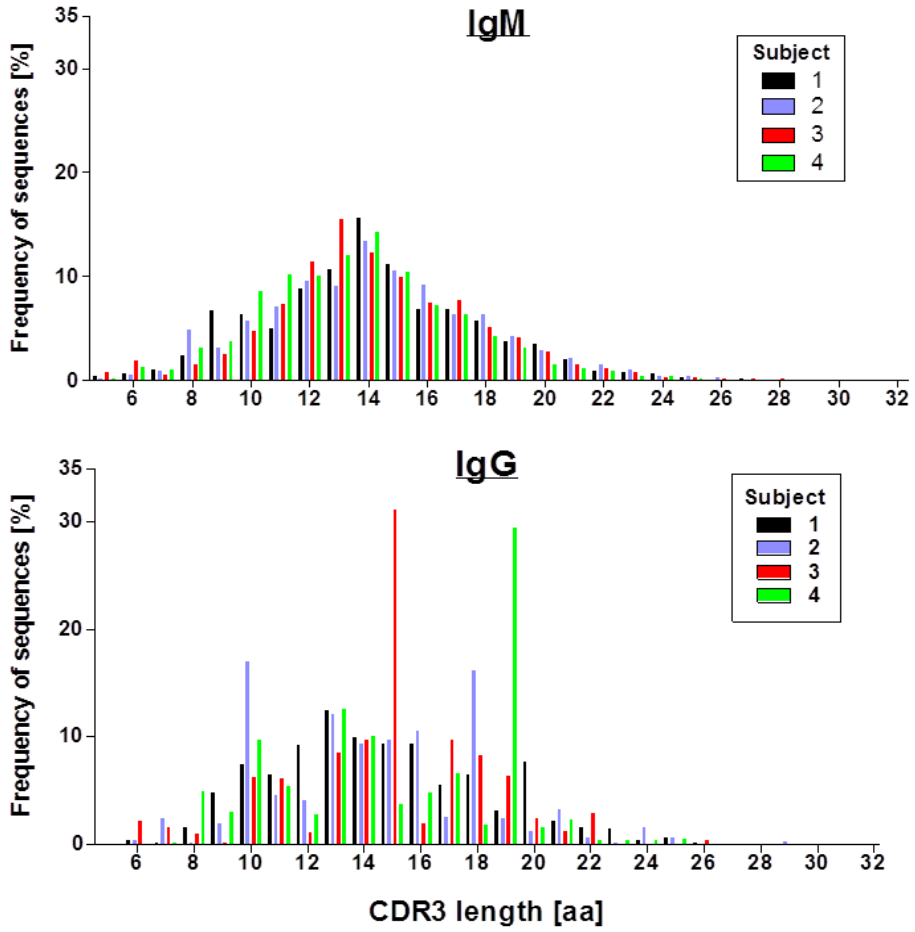


Figure 13: CDR3 amino acid length distribution in IgM and IgG transcriptome repertoires. The IgM repertoire displays a Gaussian-llike distribution, whereas the IgG repertoire looks skewed.

3.5 Clonality in the IgG repertoire

To further investigate if the high frequency of sequences with certain CDR3 length of the IgG repertoire derived from clonal expansion or from other potentially conserved genetic mechanisms, all CDR3 length duplicates were removed from the repertoire of each individual. Only one sequence per cluster of identical amino acid residues for a particular CDR3 length was kept to compose a unique pool of sequences of a certain CDR3 length. Clonally expanded B cell populations all expressing a specific B cell receptor (and therefore contain several sequences with the same CDR3 amino acid composition) indicate an elevated immune response due to antigen exposure. By removing all duplicative sequences possessing the same

V_H-D-J_H amino acid sequence for one CDR3 length, the frequency of the cluster may differ from the normal CDR3 length distribution (Figure 13). The removal of clonality was performed for each of the four individuals separately.

The overall CDR3 length distribution of subject 1 (Figure 14A) displays a Gaussian-like shape except for the outlier peak at a CDR3 length of 20 amino acids. A total of 80% of multiple existing sequences within the junction area was removed to generate a unique sequence pool (Figure 14B). The junction region is defined based on IMGT as ranging from the Cysteine (position 104) encoded from the 3' end of IGHV up to the first Tryptophan or Phenylalanin encoded from the IGHJ gene, spanning the whole CDR3 region (Figure A 1). After removing the duplicates with the same junction amino acid sequence, the distribution looks more Gaussian. There is also a shift of the maximum peak from a CDR3 length of 13 amino acids (Figure 14A) to 15 amino acids (Figure 14B).

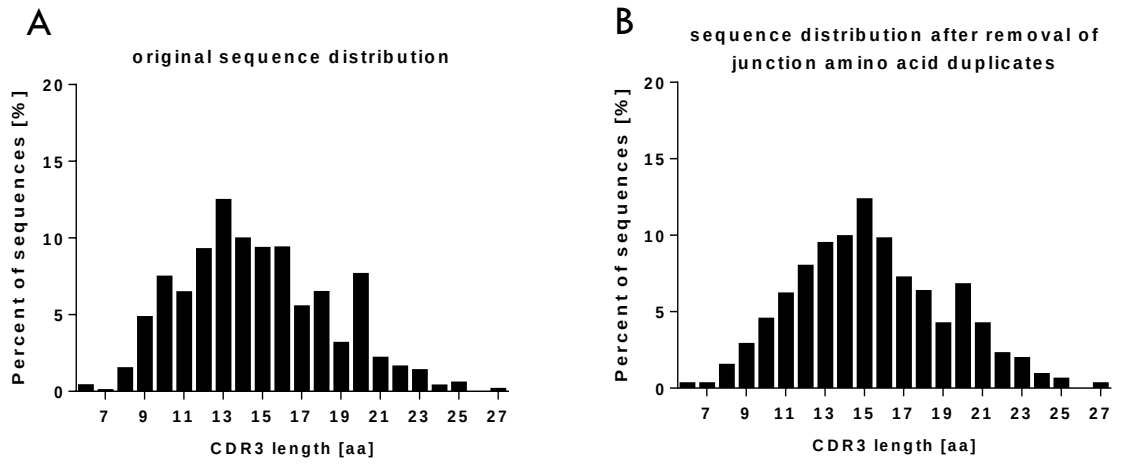


Figure 14: Subject 1 - CDR3 amino acid length distribution considering all quality sequences (A) and all unique sequences which were obtained by removing duplicative sequences with the same junction amino acid composition (B). After removing all multiple sequence copies the original sequence number of 3450 was greatly reduced to 666 sequences. The frequency of sequences in B is slightly different from graph A. The major peak shifted from a CDR3 length of 13 to 15, and the peak at a CDR3 length of 20 amino acids is still noticeable when comparing A and B.

To further estimate the impact of clonality within the sequence clusters with the most used CDR3 amino acid lengths, total sequence numbers for that particular CDR3 length were subjected to the removal of doubled or multiple sequences using the same V_H-D-J_H amino acid residues. This study was done particularly for the CDR3 length of 13, 15 and 20 amino acids. For the CDR3 length of 13 amino acids a total of 430 sequences were observed. After removal of all duplicative sequences with the same IGH VDJ region, 268 unique sequences remain, resulting in 37.7% clonality. To explore whether the sequences among themselves

with this CDR3 length are genetically related, an amino acid alignment of the IGH VDJ region of the four most abundant sequence clusters was performed (Figure 15). The most used sequence cluster with 52 multiples (cluster 1) served as reference sequence for the alignment. Except for the variable and the junction region of sequence cluster 2, all other sequence clusters align very well. The match of the alignment of the IGH VDJ region of sequence cluster 1, 3 and 4 can be confirmed by observing the use of IGHV, D and J families, which are indeed identical. All the three clusters include the V3-7*01, D3-16*01 and J4*02 gene. The second sequence cluster differs from the rest, with only the IGHJ family in common.

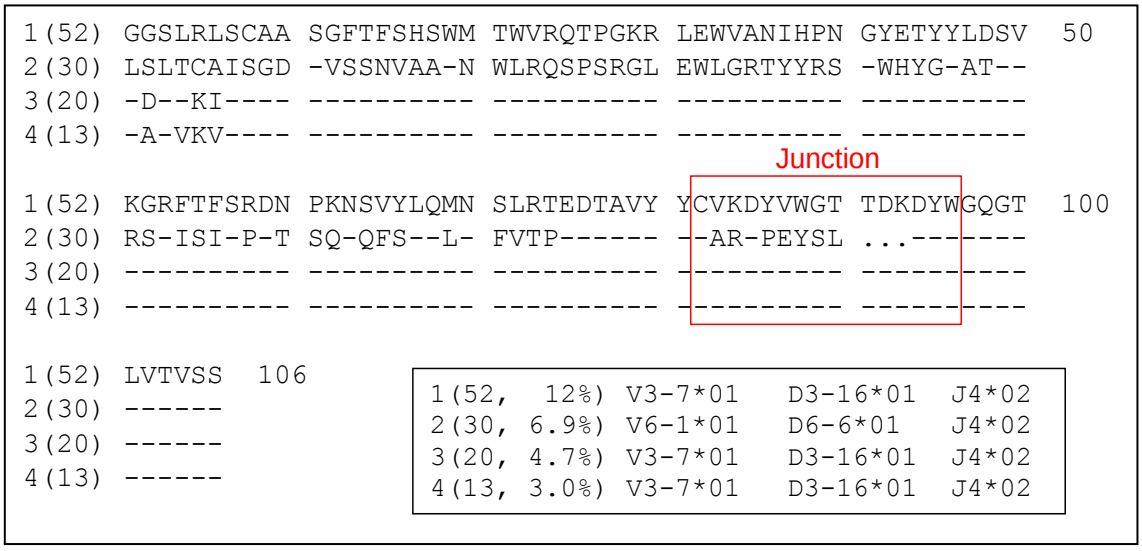


Figure 15: Subject 1: Amino acid alignment of the IGH VDJ antibody region of the four most abundant sequences with a CDR3 length of 13 amino acids. The in red designated region displays the junction region between the corresponding variable and joining gene segments, and comprises the diversity segment and junctional modifications, due to somatic recombination. The first sequence cluster was with 12% the most used cluster for this CDR3 length and was therefore chosen as reference sequence. The lower panel of this figure presents the IGHV, D and J alleles of the underlying sequence clusters.

These results suggest that sequence clusters 1, 3 and 4 are more closely related genetically compared to sequence cluster 2.

To confirm this consideration a phylogenetic neighbor-joining tree was created (Figure 16). As expected, the sequence clusters 1, 3 and 4 group together at one branch of the tree, indicating their genetic proximity. This finding implies that memory B cells carrying an antigen binding site on their IgG B cell receptor with the CDR3 length of 13 amino acids might have developed mostly from the same ancestral cell.

Figure 16: Subject 1: Neighbor-joining tree of the four most abundant sequence clusters with a CDR3 length of 13 amino acids. The sequence clusters 1, 3 and 4 grouped together at one end of the tree branch, indicating a closer phylogenetic proximity. The numbers in brackets indicate the number of sequences belonging to this cluster.

Based on Figure 14B, the highest frequency of sequences after removal of duplicates from the junction region, shifted towards a CDR3 length of 15 amino acids. When looking into detail at the sequences clustering at this CDR3 length, 322 sequences could be found in the whole IgG expressing B cell repertoire of subject 1. With 230 unique sequences, the clonality in this case was about 28.6%, indicating that the majority of B cell receptors were unique. The IGH VDJ amino acid alignment and the neighbor-joining tree of the four most abundant sequence clusters with the CDR3 length of 15 amino acids are shown in Figure A 2 and Figure A 3. The most abundant sequence cluster only comprises 7% of all clusters within this CDR3 length, followed by the second most abundant cluster with 3.4%. Both sequence clusters are composed of different IGH VDJ families. All in all the CDR3 length of 15 amino acids contains many different sequences with distinct amino acid residues.

Figure 14B also displays that the high peak at a CDR3 length of 20 amino acids falls out of the usual distribution pattern of the total sequences as well as that ones with a unique junction region (Figure 14 A and B). From 263 overall sequences with this CDR3 length, 170 were unique, displaying a clonality of 35.4%. Both the IGH VDJ amino acid alignment and the neighbor-joining tree are presented in Figure A 4and Figure A 5.

Altogether the results of this clonality study were quite similar for the three dominant sequence clusters of subject 1. The clonality ranged from 28% to 37%, indicating that most of the sequence clusters that do not fit into the normal distribution of the CDR3 length distribution, do not derive from a clonal expansion of a certain specific B cell.

The CDR3 length distribution of subject 2 (Figure 17) displays an intermittent sequence frequency while having two predominant length of 10 and 18 amino acid residues. After removing the duplicates with the same junction amino acid sequence from the pool, the frequency of sequences with a CDR3 length of 18 amino acids remains similar, yet those with a CDR3 length of 10 amino acid is reduced (Figure 17B). Instead the frequency of CDR3 lengths with 13 amino acids increased. Compared with the original sequence pool, the number of sequences after removing all junction duplicates was reduced to 85%, counting then 709 sequences.

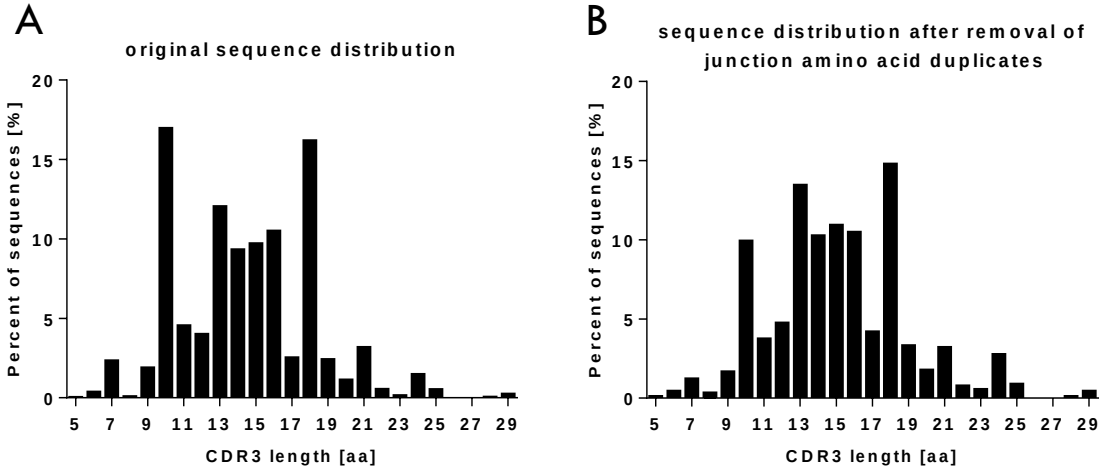


Figure 17: Subject 2: CDR3 amino acid length distribution considering all quality sequences (A) and all unique sequences which were obtained by removing duplicate sequences with the same junction amino acid composition (B). After removing all multiple sequence copies the original sequence number of 5968 is greatly reduced to 907 sequences. The frequency of sequences in Figure B is slightly different from graph A. The major peak shifted from a CDR3 length of 10 to 13, while the peak at a CDR3 length of 18 amino acids remains almost similar.

For the sequence cluster with the CDR3 length of 10 amino acids a total of 1012 sequences were attained. After removing the duplicates with the same IGH VDJ region, the sequence number decreased to 601, including a clonality of 41%.

The amino acid alignment of the four most abundant sequence clusters with a CDR3 length of 10 amino acids revealed that sequence cluster 1, 3 and 4 are rather conform (Figure 18). Especially cluster 1 and 4 have, except for a single amino acid, the same IGH VDJ region.

```
1(92) QTLSLTCAFS GDGFSSNSVA WNWIRQSPSR GLEWLGRTYY WSKWNNDYAV  50
2(46) .GA-VKVSCK TS-YTFTASY LH-V--A-GQ -P--M--HIN PYSGVSKQTQ
3(40) .--------- ------R--- ---------- ---------- --R------G
4(26) E--------- ---------- ---------- ---------- ----------
                                              Junction
1(92) SVKGRITISP DTSKNQFSLQ LKSVTPEDTA VYYCARGRNS GFDIWGQGTM  100
2(46) KFQ--V-MTR ---ISTAF-E -TGL-SD--- -------FGYN LP-S-----L
3(40) --R------- -PP------- ---------- --------N-- A---------A
4(26) ---------- ---------- ---------- ---------- ----------

1(92) VTVSS  105
2(46) ---F-
3(40) -----
4(26) -----

1 (92, 9.1%) V6-1*01  D3-16*01  J3*02
2 (46, 4.6%) V1-2*02  D5-24*01  J5*01
3 (40, 4.0%) V6-1*01  D5-24*01  J3*02
4 (26, 2.6%) V6-1*01  D3-16*01  J3*02
```

Figure 18: Subject 2: Amino acid alignment of the IGH VDJ antibody region of the four most abundant sequences with a CDR3 length of 10 amino acids. The in red designated region displays the junction region between the corresponding variable and joining gene segments, and comprises the diversity segment and junctional modifications, due to somatic recombination. The first sequence cluster was with 9.1% the most used cluster for this CDR3 length and was therefore chosen as reference sequence. The lower panel of this figure presents the IGHV, D and J alleles of the underlying sequence clusters.

Figure 18 also shows that both sequence cluster 1 and 4 have identical V_H-D-J_H usage. Cluster 2 has the least correspondence to the reference sequence. This cluster also uses a different IGHV and IGHJ family. To estimate the genetic relationship, a neighbor-joining tree was created, displaying that indeed sequence cluster 1 and 4 are very closely related, as they stand on the same branch end of the phylogenetic tree (Figure 19). Sequence cluster 2 has the greatest genetic distance to all the other sequences.

Figure 19: Subject 2: Neighbor-joining tree of the four most abundant sequence clusters with a CDR3 length of 13 amino acids. The sequence clusters 1, 3 and 4 grouped together at one end of the tree branch, indicating a closer phylogenetic proximity. The numbers in brackets indicate the number of sequences belonging to this cluster.

Based on Figure 17, the second most frequently occurring CDR3 length in the repertoire of subject 2, was 18 amino acids. This cluster contains 966 sequences, of which 686 are unique, thus having a clonality of 28.9%. The amino acid alignment of the four predominant sequence

clusters with a CDR3 length of 18 reveals that cluster 1, 2 and 3 are matching together in most of their amino acids. (Figure A 6). These three clusters also harbor the same IGHV, IGHD and IGHJ family, which make them very similar to each other. Cluster 4 however does not align with the same similarity as clusters 1-3. Indeed cluster 4 contains different IGHV, IGHD and IGHJ families than the other sequence clusters. It can be assumed that sequence clusters 1, 2 and 3 are closely related genetically. This assumption can be confirmed by reviewing the phylogenetic tree (Figure A 7). It is also important to notice that no sequence clusters were predominant, since even the reference sequence made up only 2.1% of all sequence clusters.

The third most frequently used CDR3 length in the repertoire of subject 2 is 13 amino acids. There are 718 sequences present having this CDR3 length, of which 507 are unique. The clonality within this sequence pool is therefore 29.4%. The most frequent sequence cluster with this CDR3 length has 28 duplicates and was chosen as reference sequence in the IGH VDJ amino acid alignment (Figure A 8). Cluster 2 and 3 display some degree of agreement in the alignment, due to the use of the same IGHV family as cluster 1. The sequence cluster 4 however possesses with IGHV1 a complete different family than the other 3 clusters and therefore does not show similarity in the alignment. The confirmation to this finding can be found by examining the neighbor-joining tree (Figure A 9). Especially cluster 2 and 3 display a short distance among each other.

The CDR3 length distribution of subject 3 shows one main peak at a length of 15 amino acids, both in the sequence distribution with all sequences (Figure 20A) and within the distribution once duplicates were removed from the junction region (Figure 20B). Overall there are not many changes in the peak frequencies among both distributions although the shape of the graph with the unique junction sequences appears more Gaussian-like. The frequency of sequences with a CDR3 length of 15 amino acids decreased. The total number of 8534 sequences in the whole repertoire decreased to 709 sequences after removing the duplicates with the same junction amino acid sequence.

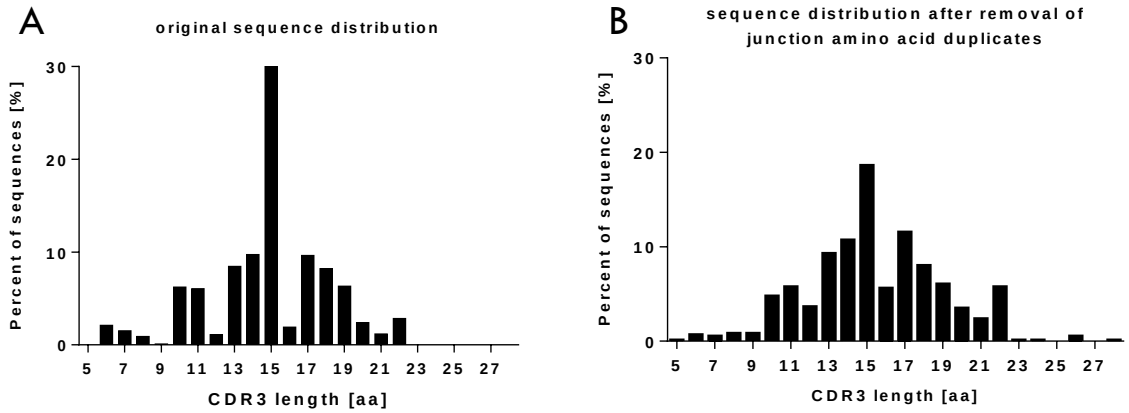


Figure 20 Subject 3: CDR3 amino acid length distribution considering all quality sequences (A) and all unique sequences which were obtained by removing duplicate sequences with the same junction amino acid composition (B). After removing all multiple sequence copies the original sequence number of 8534 is greatly reduced to 709 sequences. The frequency of sequences in B is a little bit different from graph A.

A closer look at the sequence cluster with a CDR3 length of 15 amino acids in the original repertoire reveals a total of 2654 sequences of which 1290 are unique in their IGH VDJ region. For this CDR3 length a clonality of 51.4% can be found. The most used sequence cluster with 17.1%, possessing a CDR3 length of 15 amino acids, was chosen as reference sequence for an amino acid alignment with the four most used sequence clusters (Figure 21). The alignment shows high similarity among sequence cluster 1, 2 and 4. All of them use the same V_H-D-J_H alleles. The clusters differ from sequence 1 only in three and four amino acid residues. Sequence cluster 3 possesses a different IGHD and IGHJ family. Consequently its sequence does not match with the reference sequence in the junction area.

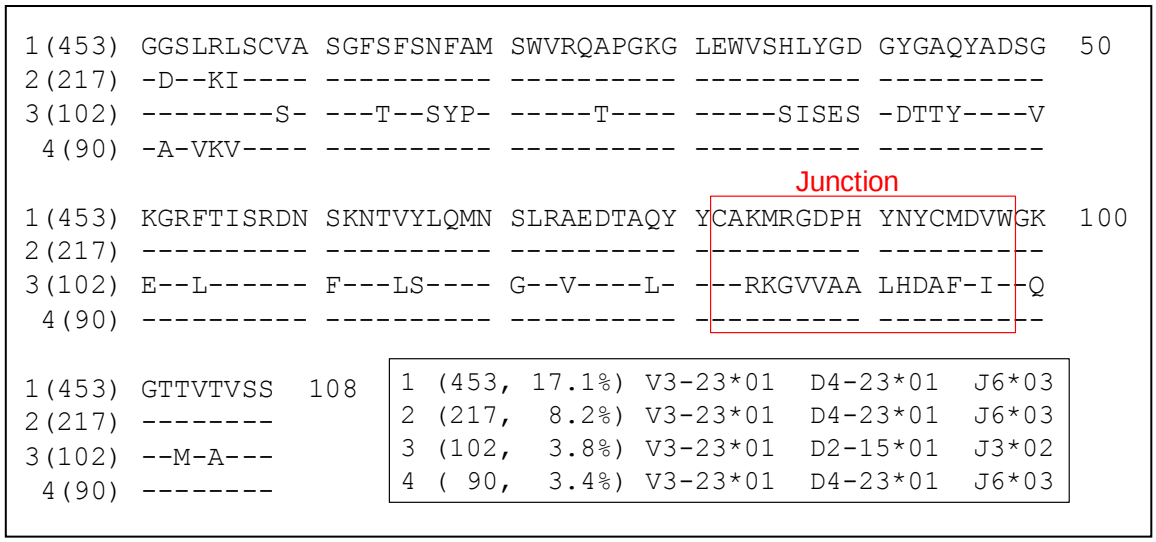


Figure 21: Subject 3: Amino acid alignment of the IGH VDJ antibody region of the four most abundant sequences with a CDR3 length of 15 amino acids. The in red designated region displays the junction region between the corresponding variable and joining gene segments, and comprises the diversity segment and junctional modifications, due to somatic recombination. The first sequence cluster was with 17.1% the most used cluster for this CDR3 length and was therefore chosen as reference sequence. The lower panel of this figure presents the IGHV, D and J alleles of the underlying sequence clusters.

The genetic relationship was further evaluated by means of a phylogenetic tree (Figure 22) Sequence cluster 2 and 4 are assembled together on one end of the tree, sharing the same node and potentially descending from the same progenitor cells like sequences from cluster 1. Sequence cluster 3 with a different D_H and J_H family belongs to another branch of the tree.

Figure 22: Subject 3: Neighbor-joining tree of the four most abundant sequence clusters with a CDR3 length of 15 amino acids. The sequence clusters 2 and 4 group together at one end of the tree branch, indicating a close phylogenetic proximity to each other and cluster 1. The numbers in brackets indicate the number of sequences belonging to this cluster.

The CDR3 length distribution of subject 4 displays a major frequency of sequences with the CDR3 length of 19 amino acid residues in the original distribution (Figure 23A). Other CDR3 lengths do not occur at similar frequencies. After removing the duplicative sequences that contain the same IGH VDJ region, the distribution normalized, because the peak frequency of the CDR3 length with 19 amino acids is reduced (Figure 23B). In this distribution the frequency of sequences with the CDR3 length of 14 amino acids is greater than that of 19 amino acids and therefore the maximum peak in this distribution. This is why these two sequence pools are further described in detail by means of their most existing sequence clusters.

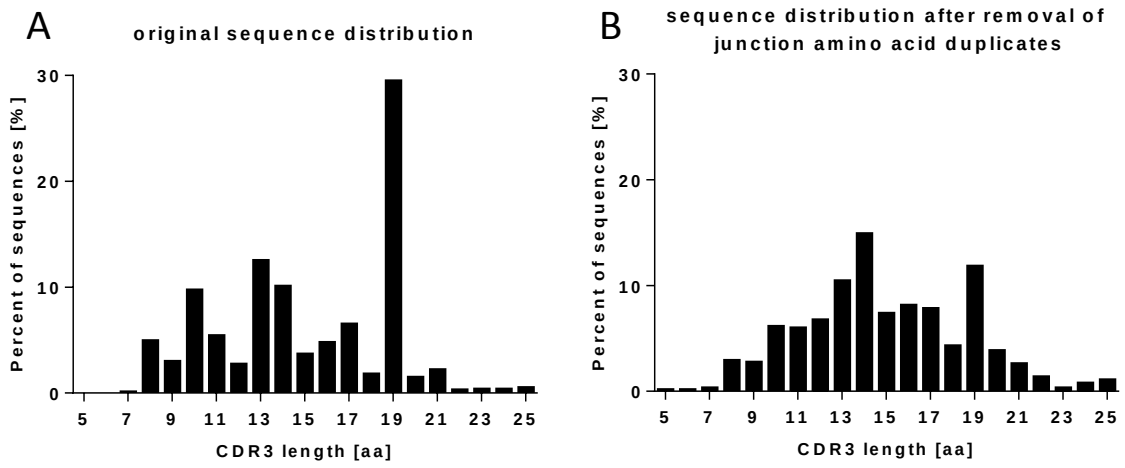


Figure 23: Subject 4: CDR3 amino acid length distribution considering all quality sequences (A) and all unique sequences which were obtained by removing duplicate sequences with the same junction amino acid composition (B). After removing all multiple sequence copies the original sequence number of 4470 is greatly reduced to 651 sequences. The frequency of sequences in B looks like a normalized cluster.

Starting with the 1318 sequences containing a CDR3 length of 19 amino acids, a total of 574 unique sequence clusters were found, displaying a clonality of 56.4%. For the four most abundant sequence clusters an amino acid alignment was performed. Sequence cluster 1 consists of 501 sequences with the same IGH VDJ region and was chosen as reference sequence for the alignment (Figure 24). The alignment shows similarity between sequence cluster 1 and 4, as they contain the same IGH VDJ alleles. Sequence 2 does not match the reference sequence due to its different V_H, D_H and J_H alleles. The largest portion of the sequence pool is made up with sequence cluster 1 (38.1%). Likely there was a strong clonal expansion of B cells specific for a certain antigen in this repertoire.

```
1(501) TLSLTCTVSG GSISGYYWSW IRQPPGKGLE WIGYIYYRGT TNYNPSLKSR  50
 2(46) ---------- ----SGTFF- AWIRQPP-KG LEWIGTIHYS GRTYYNPSLK
 3(22) VPETLLHCLW WLHQWLLLEL DPAA------ ---------- ----------
 4(13) ---------- ---------- ---------- ---------- ---------

                                              Junction
1(501) VTMAIDTSKS QFSLRLNSVT AADTAVYYCA RDFQDTTSGR YYFYYMDVWG  100
 2(46) SRLT-SVDT- KNHFS-SLTS VTAADTAVYY CARSPQYFSD GSG-ANAFDI
 3(22) ---------- ---------- ---------- ---------- ----------
 4(13) ---------- ---------- ---------- ---------- ----------

1(501) QGSTVTVSS  109
 2(46) W-HGTV-TV
 3(22) ---------
 4(13) ---------

 1(501,38.1%) V4-59*01  D1-1*01  J6*03
 2(46, 3.5%)  V4-39*02  D3-22*01 J3*02
 3(22, 1.7%)  V4-61*07  D4-17*01 J6*03
 4(13, 1.0%)  V4-59*01  D1-1*01  J6*03
```

Figure 24: Subject 4: Amino acid alignment of the IGH VDJ antibody region of the four most abundant sequences with a CDR3 length of 19 amino acids. The in red designated region displays the junction region between the corresponding variable and joining gene segments, and comprises the diversity segment and junctional modifications, due to somatic recombination. The first sequence cluster was with 38.1% the most used cluster for this CDR3 length sequence pool and was therefore chosen as reference sequence. The lower panel of this figure presents the IGHV, D and J alleles of the underlying sequence clusters.

The phylogenetic tree from the four most abundant sequence clusters confirms the close genetic relationship of sequence cluster 1 and 4 (Figure 25). Due to the usage of the same IGHJ gene, sequence cluster 3 is also closer related to 1 and 4 than cluster 2.

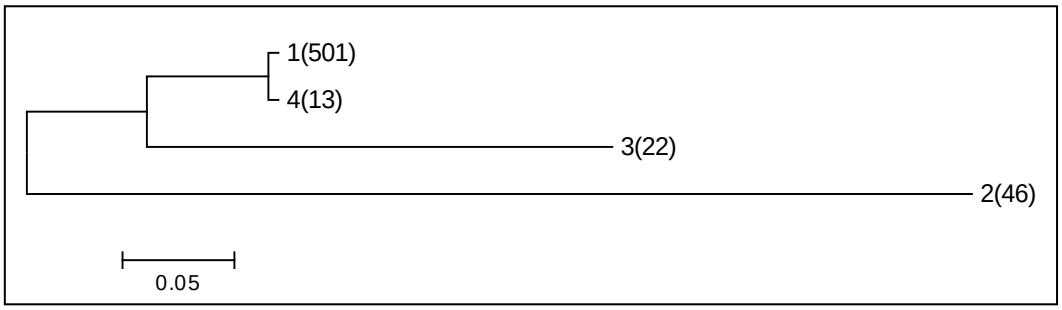


Figure 25: Subject 4: Neighbor-joining tree of the four most abundant sequence clusters with a CDR3 length of 15 amino acids. The sequence clusters 1 and 4 group together at one end of the tree branch, indicating a close phylogenetic proximity to each other. The numbers in brackets indicate the number of sequences belonging to this cluster.

The other sequence pool with a CDR3 length of 14 amino acids that possesses the greatest frequency of sequences after the removal of junctional duplicates (Figure 23) contains 451 sequences. After the removal of duplicative sequences with identical IGH VDJ regions from

the total CDR3 length pool, 331 sequence clusters remained and therefore a clonality of 26.6%. Out of those none is represented over 6.4%, which indicates a highly diverse sequence pool. Looking at the amino acid alignment of the four most often occurring sequence clusters, one find cluster 1 and 2 to be matched very well (Figure A 10) Both clusters use the same IGH VDJ alleles. While cluster 3 matches with the reference sequences to some extent, the alignment of cluster 4 is quite poor. Sequences from cluster 4 contain different IGH VDJ families than cluster 1 and 2.

The neighbor-joining tree of Figure A 11 displays a great distance between sequence cluster 4 and the cluster 1, 2 and 3. Cluster 1 and 2 themselves are located on the same branch end and are considered to be very closely genetically related.

Table 5 summarizes the results of this clonality study.

Table 5: Predominantly used CDR3 lengths with their corresponding number of sequence clusters, unique sequences and clonality.

Subject	Predominant CDR3 length	Nb. of sequences within cluster	Nb. of unique seq.	Clonality [%]
1	13	430	268	37.7
	15	322	230	28.6
	20	263	170	35.7
2	10	1012	601	41.0
	18	966	686	28.9
	13	718	507	29.4
3	15	2654	1290	51.4
4	19	1318	574	56.4
	14	451	331	26.6

3.6 Biodiversity in the IgM and IgG transcriptome repertoires

Biodiversity is a key factor to estimate the effectiveness of the immune system works. A distinct biodiversity is associated with a greater ability to recognize unknown antigens and therefore increased protection. The extent of somatic hypermutations and CDRH3 clonality are two of several factors that contribute to biodiversity.

To evaluate the extent of biodiversity in these IgM and IgG memory B cell repertoires, a rarefaction and Chao1 analysis was performed. Since the biodiversity is influenced by the number of input cells, it was weighted by the absolute number of input B cells. Rarefaction curves display the increase of biodiversity in the repertoires of the participating individuals along the number of analyzed sequences. Likewise, deeper slopes of the rarefaction curves indicate higher biodiversity than lower slopes. In this investigation, the curves of the IgG data show deeper slopes than the IgM curves. Consequently the IgG repertoire contains a higher level of biodiversity (Figure 26A). The Chao1 analysis reveals greater maximum estimated biodiversity in the IgG repertoire (Figure 26B). In addition, the maximum estimated biodiversity among individuals varies in the IgG repertoire. Whereas data points of the IgM repertoire are consistently clustered together, those of the IgG repertoire are more spread out. It can therefore be assumed that the IgG repertoire shows a greater individual variation in terms of biodiversity.

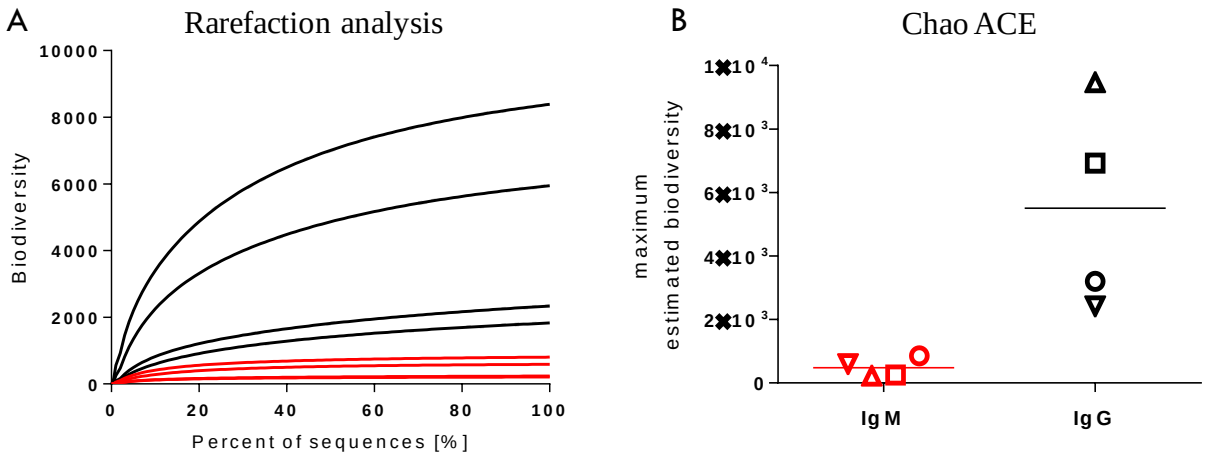


Figure 26: Biodiversity investigation among IgM and IgG transcriptome repertoire. A: The biodiversity along the depth of sequences was evaluated using a rarefaction analysis. B: For the maximum estimated biodiversity a Chao1 investigation has been conducted. Numbers and symbols represent four subjects. Curves and symbols marked in red represent the IgM repertoire. Black curves and symbols correspond to the IgG repertoire. Both graphs show an overall greater biodiversity of the IgG repertoire.

4 Discussion

The goal of this study was to investigate molecular differences between IgM and IgG transcriptome repertoires that contribute to the biodiversity of the B cell immunity. For this purpose a new strategy of using mRNA of total B cell populations for subsequent deep sequencing was applied. These B cells have not been sorted into different subtypes by using cell surface markers prior to the extraction of mRNA. The advantage of this method is that possible incorrect assignments to different B cell subtypes can be avoided. The immunoglobulin subtypes IgM and IgG were assigned based on their heavy chain constant region, which provides undoubted accuracy. Until now, the field of B cell development and diversity still contains a lot of aspects to discover and revise.

Investigations on the assignments of different B cell clonotypes in past years have been shown that there is a great amount of B cell subtypes that cannot be distinguished by means of one cell surface marker. In contrast it has recently been shown that there are several distinct B cell populations using the same surface antigen. That makes it difficult to capture all memory cells, when analyzing immunoglobulin repertoires. Previous studies on assessing immunoglobulin repertoires used DNA from previously sorted B cell subclones.
The cell surface marker CD27 for example has long been thought to identify only somatically hypermutated memory B cells. Although, more recent studies have shown that there is a population of B cells lacking CD27 expression and can still be referred to as memory B cells (Wei *et al.* 2007). The most precise marker to identify memory B cells is the analysis of somatic hypermutations, since they are imprinted in the rearranged IGH variable region genes.

High-throughput sequencing has enabled the generation of almost 16,000 raw reads in this study and therefore gives the possibility to analyze a lot more sequences then conventional Sanger sequencing, which is essential given the fact that the immune system comprises about 10^{12} different antibody specificities.

4.1 Somatic hypermutations as marker for B cell development and biodiversity

Somatic hypermutations are a means to distinguish memory from naïve B cells. The analysis of the IGHM repertoire revealed a percentage of 2.2% sequences without somatic hypermutation and approximately 95% of sequences that contain at least two or more (Figure 6). This high number of sequences with somatic hypermutations might indicate a greater IgM

expression memory B cell population than previously thought. Although not significant the IgG transcriptome repertoire contains about 4% more sequences with somatic hypermutation than the IgM repertoire suggesting that indeed IgG expressing memory B cells are more diversified and originate from a later developmental stage. They have undergone SHM in a germinal center reaction, as well as class-switch recombination to gain new effector functions. Although the extent of somatically mutated sequences in the IgG transcriptome repertoire was greater than in the IgM repertoire that might not definitely indicate a greater pool of mutated IgG expressing memory B cells in the peripheral blood. According to Klein et al. activated B cells such as memory B cells contain a 7 times higher amount of mRNA compared to naïve B cells (Klein *et al.* 1998). Due to this fact, a disproportionally greater amount of mRNA of those cells could have contributed in generating the amplicon library.

The analysis of the total extent of somatic hypermutations (Figure 7A) revealed that the IgG repertoire contains more sequences with both silent and nonsilent mutations compared to the IgM sequence repertoire. The relevant SHM that lead to a more diversity in the immunoglobulin repertoires are the nonsilent mutations, since they possibly alter the amino acid sequence. The more nonsilent mutations a sequence repertoire displays, the greater the BCR was shaped towards better antigen affinity. The detailed look at different IGH regions shows that the two transcriptome repertoires are quite different regarding the locations of the SHM (Figure 7B). The greatest extent of SHM in the IgM repertoire was found in the first framework region, whereas in the IgG repertoire most SHM are located in the CDR2. In the IgG repertoire most nonsilent SHM can be found in the CDR's. The distribution of SHM in the IgM repertoire however reveals the greatest extent in FR1 and CDR1. Framework regions in general do not contribute as much as complementarity determining regions to antibody affinity and antigen binding capacity. However, nonsilent mutations in FR could alter the correlation of amino acid residues in the framework regions of the antibody molecule, so that the whole secondary and tertiary structure of the antibody could change and consequently enhance or impair the binding capacity. The framework regions of the IgM repertoire might be a target for SHM due to a higher transcription rate at the beginning of the antibody DNA sequence. It has been proven that SHM are targeted to regions with higher transcription rates (Odegard & Schatz 2006). My attempted explanation is, because IgM antibodies tend to form pentameric complexes in the body, mutations in the structure-giving regions could affect the three-dimensional protein structure.

Somatic hypermutations are not equally distributed along the IGH chain. There are mutational hotspots with a higher frequency of SHM. In this study especially FR1 of the IgM repertoire was affected by SHM, followed by mutations in the CDR1. This is true for both the silent and the nonsilent mutations. The distribution of SHM is significantly different from the IgG repertoire. IgG transcriptome repertoire displays the greatest extent of SHM in the CDR2 and CDR1 regions. Liberman *.et al*. found that CDR regions of switched memory B cell clones are subjected to a stronger selection process compared to the IgM positive memory B cells (Liberman *et al.* 2013). It was further suggested that IgM expressing B cells leave the germinal center reaction earlier than their class-switched counterparts, resulting in a generally lower extent of SHM (Seifert & Kuppers 2009).

4.2 Characterization of combinatorial and junctional diversity

To evaluate the genetic properties of the rearranged immunoglobulin DNA, the utilization frequencies of IGH VDJ families, DH reading frame as well as non-germline encoded nucleotides were determined.

Figure 8 displays that the IGHV family utilization is consistent among the four analyzed individuals. In both IgM and IgG repertoires the VH3 family was used predominantly. The IgM repertoire shows a significantly higher frequency of VH3 family though. Whereas the individual VH family usage of the IgM repertoire is clustered, the usage is more spread out among individuals in the IgG repertoire. Overall the frequencies of IGHV families are similar among the IgM and the IgG repertoire.

The same conclusion can be reached when observing the IGHD family usage (Figure 9). There are no significant differences when comparing the IgM with the IgG transcriptome repertoires. Yet, also in the case of IGHD families, the individual frequencies of the IgG repertoires are spread out, instead of clustered together as it is case in the IgM repertoires. The most frequently used DH family is DH3, followed by DH2.

Utilization frequencies of the IGHJ families also displayed variation (Figure 10). The most predominantly occurring JH family is JH4 for both the IgM and the IgG repertoire, although with significantly different frequency. As already mentioned above there is some intraindividual variation in the IgG repertoire in contrast to the consistency of in the IgM repertoire.

The reason for this observation might be that the IgM memory B cells give baseline immune protection in every individual with the same most abundant V_H-D-J_H combinations. The

characterization of the IGH VDJ family frequency is means to map the combinatorial diversity of a given repertoire. It has already been reported that the utilization frequency of the different IGH VDJ families and the combination of certain gene segments is highly biased (Volpe & Kepler 2008). The utilization frequencies seem to be genetically determined, since they are influenced by the variation in recombination signal sequences (Jackson *et al.* 2013) There are already investigations to study the frequency of IGH VDJ families, but most of them focused on the detection of malignancies, which makes them not suitable to investigate healthy repertoires (van Dongen *et al.* 2003). An even greater bias in the utilization frequencies is induced by the D_H gene segments. Theoretically they can be used in 6 different reading frames and thus hold the possibility to alter the amino acid sequence, especially of the CDR3 region. Although there are potentially 6 reading frames (RF), three of them arise from an inverted D_H gene segment and are rarely found in any healthy B cell repertoire (Benichou *et al.* 2013). D_H segments can even occur in a double fusion, resulting in V_H-(DD)-J_H recombinations (Briney *et al.* 2012). It has been shown that the use of RF is biased. Studies in mice for example suggest the selection against RF containing a stop codon (Zemlin *et al.* 2008). In contrast to other studies evaluating the D_H reading frames, in this investigation the use of RF2 was preferred in both IgM and IgG repertoires (Figure 11). Studies in mice however have shown that mice forced to use RF2 developed fewer numbers of both immature and mature B cells in the spleen and bone marrow (Schelonka *et al.* 2008). The D_H RF influences the amino acid content of the CDRH3 repertoire, since it represents the central part of the CDRH3 coding DNA. Until recently it was believed that the DH RF choice was strongly regulated by the underlying rearrangement process, but a more recent study claims the means of negative selection instead. There are two selection checkpoints suggested: one at the stage of early B cell development and one occurring in the periphery (Benichou *et al.* 2013). The RF bias is dependent on the individual D_H gene segment. Benichou *et al*. also report a very similar distribution of RF use in IgM and IgG memory B cell repertoires. As a matter of fact, the present investigation could not find any impact of SHM on the DH RF usage. The preferential RF in the unmutated sequences still account for RF2, although the sequence number was too small to be representative (data not shown).

The junctional diversity was evaluated by observing the non-germline encoded nucleotides in the mutated and unmutated repertoires (Figure 12). In both sequence repertoires with and without SHM the IgG repertoire contains more N nucleotides than IgM repertoire. Interestingly, all sequences of the nonmutated IgG repertoire contain N nucleotides, whereas

there are approximately 1.3% of sequences without N in the mutated sequences. It seems that IgG expressing memory B cells are selected to leave the GC prior to acquisition of SHM because they are already sufficiently diversified and probably possess enough antigen affinity. The addition of N nucleotides is intrinsically biased due to the preference of TdT to incorporate G nucleotides (Jackson *et al.* 2007).

4.3 The role of immunoglobulin heavy chain CDR3 region for biodiversity

The heavy chain CDR3 region is the region located in the middle of the binding site, after the antibody molecule has assumed its final shape. It greatly makes contact with antigens and is therefore subjected to diversification along the B cell development. It is consequently an important characteristic to evaluate immunoglobulin repertoires. The CDRH3 region acquires the most diversity because its genetic formation involves three different gene segments. Another aspect that helps to create diversity as well as functional flexibility is the non-germline encoded nucleotides that flank the D_H gene segment (Jackson *et al.* 2013).

It is already known that the CDR3 length distribution displays a Gaussian-like shape in non-stimulated B cell populations (Miqueu *et al.* 2007). Here the IgM and IgG immunoglobulin repertoires from memory B cells were analyzed. Especially the length distribution of the IgM repertoire displays a Gaussian distribution with a predominantly used length of 14 amino acids (Figure 13). The frequencies are very consistent between the four individuals.
In contrast the CDR3 length distribution of the IgG repertoire displays variation between individuals, which alters the Gaussian shape from polyclonal to oligoclonal. At least three of four individual's repertoires prefer amino acid lengths outside the bell shape curve.

The CDRH3 length and amino acid composition influences the shape and folding structure of its loop and the interacting loops of CDRH1 and CDRH2 (Davis *et al.* 1998). The differences between the repertoires of different individuals can be explained by the existence of clonal expansions of B cells with a specific BCR. Obviously there must have been a strong antigen that activated a particular B cell and thus enhanced its proliferation and the differentiation into IgG expression memory B cells. This event is clearly represented in the oligoclonal frequency peaks in the IgG repertoire. Every peak of the length distribution contains multiple distinct sequences, which can be summarized in multiple clusters.

4.4 Measurement of B cell activation by clonality in antibody sequences

Since the IgG repertoire displayed some monoclonal peaks in the CDR3 length distribution, these peaks were subjected to further analysis. Since every peak is composed of several sequences with the same amino acid length, it is interesting to see if all of those are composed of the same amino acid sequence. It is believed that the predominantly occurring CDR3 lengths and outlier indicate B cell activation and in response their clonal expansion (Miqueu *et al.* 2007).

This study cannot confirm the idea that high length frequencies relate to a strong clonal expansion of a specific B cell clone. When observing the predominant CDR3 lengths of a given immunoglobulin B cell repertoire, they do not always show the highest amount of clonality. Yet they are composed of a highly diverse cluster of different sequences with the same CDR3 length, but variable amino acid composition. The most abundant sequence clusters do not automatically add to the majority of sequences belonging to one CDR3 length. This observation is independent of the position of the peak with respect to the length distribution pattern. Outliner peaks as well as predominantly occurring peaks within the Gaussian-like shape do not correlate with an increase in clonality *per se*.

4.5 Factors contributing to biodiversity of the IgM and IgG repertoires

There are several factors influencing the extent of diversity in a given memory B cell population. These factors were already described in detail in the introduction of this thesis. Since IgM is the first antibody type that is produced upon first encounter with an unknown antigen, it is believed to belong to a more primarily immune response that has not acquired full antigen affinity yet (Taylor *et al.* 2012). Consequently it can be assumed that the IgM^+ memory B cell compartment comprises less diversity in its BCR than IgG^+ B cells.

The rarefaction and Chao1 analysis provides the possibility to compare IgM and IgG biodiversity (Figure 26). The rarefaction curve estimates the diversity of the repertoires along with the depth of sequences. The higher the number of analyzed sequences of a given repertoire the more biodiversity can be catched. Rarefaction curves do not rise endlessly however. At a certain number of sequences there is no further increase in diversity possible, since the curve reaches the plateau phase. In this investigation all curves belonging to the IgM repertoires reached the plateau phase indicating that no more biodiversity can be expected

even when the input number of sequences would be greater. The curves of the IgG repertoires have deeper initial slopes and do not fully reach the plateau phase, confirm the hypothesis of a greater extent of biodiversity. This finding is also confirmed by the Chao1 analysis which estimates the maximum biodiversity. It also reveals more variation between subjects in the IgG repertoires, suggesting that IgG immunoglobulins are expressed to precisely respond to individual antigen exposure. Obviously different stages during B cell maturation add to the amount of biodiversity (Arnaout *et al.* 2011). The factor SHM that is shaping the BCR post antigen exposure seems to contribute to the very high extent to the biodiversity of a B cell population.

In conclusion this study contributes to the molecular understanding of creating antibody diversity and presents an important baseline for further investigations in healthy and diseased individuals.

5 References

Alt FW, Oltz EM, Young F, Gorman J, Taccioli G , Chen J (1992). VDJ recombination. *Immunol Today*. **13**, 306-314.

Ansel KM , Cyster JG (2001). Chemokines in lymphopoiesis and lymphoid organ development. *Curr Opin Immunol*. **13**, 172-179.

Arnaout R, Lee W, Cahill P, Honan T, Sparrow T, Weiand M, Nusbaum C, Rajewsky K , Koralov SB (2011). High-Resolution Description of Antibody Heavy-Chain Repertoires in Humans. *PLoS One*. **6**.

Benichou J, Glanville J, Prak ET, Azran R, Kuo TC, Pons J, Desmarais C, Tsaban L , Louzoun Y (2013). The restricted DH gene reading frame usage in the expressed human antibody repertoire is selected based upon its amino acid content. *J Immunol*. **190**, 5567-5577.

Bossuyt X, Marti GE , Fleisher TA (1997). - Comparative analysis of whole blood lysis methods for flow cytometry. *Cytometry*. **30**, 124-133.

Briney BS, Willis JR, Hicar MD, Thomas JW , Crowe JE (2012). Frequency and genetic characterization of V(DD)J recombinants in the human peripheral blood antibody repertoire. *Immunology*. **137**, 56-64.

Brochet X, Lefranc MP , Giudicelli V (2008). IMGT/V-QUEST: the highly customized and integrated system for IG and TR standardized V-J and V-D-J sequence analysis. *Nucleic Acids Res*. **36**, W503-508.

Busslinger M (2004). Transcriptional control of early B cell development. *Annu Rev Immunol*. **22**, 55-79.

Casellas R, Shih TA, Kleinewietfeld M, Rakonjac J, Nemazee D, Rajewsky K , Nussenzweig MC (2001). Contribution of receptor editing to the antibody repertoire. *Science*. **291**, 1541-1544.

Cook GP , Tomlinson IM (1995). The human immunoglobulin VH repertoire. *Immunology Today*. **16**, 237-242.

Corbett SJ, Tomlinson IM, Sonnhammer ELL, Buck D , Winter G (1997). Sequence of the human immunoglobulin diversity (D) segment locus: a systematic analysis provides no evidence for the use of DIR segments, inverted D segments, “minor” D segments or D-D recombination. *Journal of Molecular Biology*. **270**, 587-597.

Cox DW, Markovic VD , Teshima I (1982). Genes for immunoglobulin heavy chains and for |[alpha]|1-antitrypsin are localized to specific regions of chromosome 14q. **297**, 428-430.

Davies DR , Metzger H (1983). Structural basis of antibody function. *Annu Rev Immunol*. **1**, 87-117.

Davis MM, Boniface JJ, Reich Z, Lyons D, Hampl J, Arden B , Chien Y (1998). Ligand recognition by alpha beta T cell receptors. *Annu Rev Immunol*. **16**, 523-544.

Edelman GM (1991). Antibody structure and molecular immunology. *Scand J Immunol*. **34**, 1-22.

Flanagan JG , Rabbitts TH (1982). Arrangement of human immunoglobulin heavy chain constant region genes implies evolutionary duplication of a segment containing gamma, epsilon and alpha genes. *Nature*. **300**, 709-713.

Frippiat JP, Williams SC, Tomlinson IM, Cook GP, Cherif D, Le Paslier D, Collins JE, Dunham I, Winter G , Lefranc MP (1995). Organization of the human immunoglobulin lambda light-chain locus on chromosome 22q11.2. *Hum Mol Genet*. **4**, 983-991.

Fukui K, Noma T, Takeuchi K, Kobayashi N, Hatanaka M , Honjo T (1983). Origin of adult T-cell leukemia virus. Implication for its zoonosis. *Mol Biol Med*. **1**, 447-456.

Glanville J, Zhai W, Berka J, Telman D, Huerta G, Mehta GR, Ni I, Mei L, Sundar PD, Day GMR, Cox D, Rajpal A , Pons J (2009). Precise determination of the diversity of a combinatorial antibody library gives insight into the human immunoglobulin repertoire. *Proceedings of the National Academy of Sciences*. **106**, 20216-20221.

Gokmen E, Raaphorst FM, Boldt DH , Teale JM (1998). Ig Heavy Chain Third Complementarity Determining Regions (H CDR3s) After Stem Cell Transplantation Do Not Resemble the Developing Human Fetal H CDR3s in Size Distribution and Ig Gene Utilization. *Blood*. **92**, 2802-2814.

Goodnow CC, Crosbie J, Jorgensen H, Brink RA , Basten A (1989). Induction of self-tolerance in mature peripheral B lymphocytes. *Nature*. **342**, 385-391.

Hystad ME, Myklebust JH, Bø TH, Sivertsen EA, Rian E, Forfang L, Munthe E, Rosenwald A, Chiorazzi M, Jonassen I, Staudt LM , Smeland EB (2007). Characterization of Early Stages of Human B Cell Development by Gene Expression Profiling. *The Journal of Immunology*. **179**, 3662-3671.

Jackson KJ, Gaeta BA , Collins AM (2007). Identifying highly mutated IGHD genes in the junctions of rearranged human immunoglobulin heavy chain genes. *J Immunol Methods*. **324**, 26-37.

Jackson KJ, Kidd MJ, Wang Y , Collins AM (2013). The Shape of the Lymphocyte Receptor Repertoire: Lessons from the B Cell Receptor. *Front Immunol*. **4**, 263.

Kabat EA (1982). Antibody diversity versus antibody complementarity. *Pharmacol Rev*. **34**, 23-38.

Klein U, Rajewsky K , Kuppers R (1998). Human immunoglobulin (Ig)M+IgD+ peripheral blood B cells expressing the CD27 cell surface antigen carry somatically mutated variable region genes: CD27 as a general marker for somatically mutated (memory) B cells. *J Exp Med*. **188**, 1679-1689.

Kumagai I , Tsumoto K (2001). Antigen-Antibody Bindinged^eds). Encyklopedia of Life Sciences: Nature Publishing Group.

LeBien TW (2000). Fates of human B-cell precursors. *Blood*. **96**, 9-23.

Lee A, Desravines S , Hsu E (1993). IgH diversity in an individual with only one million B lymphocytes. *Dev Immunol*. **3**, 211-222.

Li Z, Woo CJ, Iglesias-Ussel MD, Ronai D , Scharff MD (2004). The generation of antibody diversity through somatic hypermutation and class switch recombination. *Genes Dev*. **18**, 1-11.

Liberman G, Benichou J, Tsaban L, Glanville J , Louzoun Y (2013). Multi Step Selection in Ig H Chains is Initially Focused on CDR3 and Then on Other CDR Regions. *Front Immunol*. **4**.

Löffert D, Ehlich A, Müller W , Rajewsky K Surrogate Light Chain Expression Is Required to Establish Immunoglobulin Heavy Chain Allelic Exclusion during Early B Cell Development. *Immunity*. **4**, 133-144.

Maddaly R, Pai G, Balaji S, Sivaramakrishnan P, Srinivasan L, Sunder SS , Paul SFD (2010). Receptors and signaling mechanisms for B-lymphocyte activation, proliferation and differentiation – Insights from both in vivo and in vitro approaches. *FEBS Letters*. **584**, 4883-4894.

Manis JP, Tian M , Alt FW (2002). Mechanism and control of class-switch recombination. *Trends Immunol*. **23**, 31-39.

Matsuda F, Ishii K, Bourvagnet P, Kuma K, Hayashida H, Miyata T , Honjo T (1998). The Complete Nucleotide Sequence of the Human Immunoglobulin Heavy Chain Variable Region Locus. *J Exp Med*. **188**, 2151-2162.

Maul RW , Gearhart PJ (2010). AID and somatic hypermutation. *Adv Immunol*. **105**, 159-191.

McBlane JF, van Gent DC, Ramsden DA, Romeo C, Cuomo CA, Gellert M , Oettinger MA (1995). Cleavage at a V(D)J recombination signal requires only RAG1 and RAG2 proteins and occurs in two steps. *Cell*. **83**, 387-395.

Metzker ML (2010). Sequencing technologies - the next generation. *Nat Rev Genet*. **11**, 31-46.

Miqueu P, Guillet M, Degauque N, Doré JC, Soulillou JP , Brouard S (2007). Statistical analysis of CDR3 length distributions for the assessment of T and B cell repertoire biases. *Mol Immunol*. **44**, 1057-1064.

Monod MY, Giudicelli V, Chaume D , Lefranc M-P (2004). IMGT/JunctionAnalysis: the first tool for the analysis of the immunoglobulin and T cell receptor complex V–J and V–D–J JUNCTIONs. *Bioinformatics*. **20**, i379-i385.

Muramatsu M, Sankaranand VS, Anant S, Sugai M, Kinoshita K, Davidson NO , Honjo T (1999). Specific expression of activation-induced cytidine deaminase (AID), a novel member of the RNA-editing deaminase family in germinal center B cells. *J Biol Chem*. **274**, 18470-18476.

Nagaoka H, Yu W , Nussenzweig MC (2000). Regulation of RAG expression in developing lymphocytes. *Curr Opin Immunol*. **12**, 187-190.

Odegard VH , Schatz DG (2006). Targeting of somatic hypermutation. *Nat Rev Immunol*. **6**, 573-583.

Ohm-Laursen L, Nielsen M, Larsen SR , Barington T (2006). No evidence for the use of DIR, D-D fusions, chromosome 15 open reading frames or VH replacement in the peripheral repertoire was found on application of an improved algorithm, JointML, to 6329 human immunoglobulin H rearrangements. *Immunology*. **119**, 265-277.

Pallares N, Lefebvre S, Contet V, Matsuda F , Lefranc MP (1999). The human immunoglobulin heavy variable genes. *Exp Clin Immunogenet*. **16**, 36-60.

Perez-Andres M, Paiva B, Nieto WG, Caraux A, Schmitz A, Almeida J, Vogt RF, Jr., Marti GE, Rawstron AC, Van Zelm MC, Van Dongen JJ, Johnsen HE, Klein B , Orfao A (2010). Human peripheral blood B-cell compartments: a crossroad in B-cell traffic. *Cytometry B Clin Cytom*. **78 Suppl 1**, S47-60.

Pieper K, Grimbacher B , Eibel H (2013). B-cell biology and development. *J Allergy Clin Immunol*. **131**, 959-971.

Ramiscal RR , Vinuesa CG (2013). T-cell subsets in the germinal center. *Immunol Rev*. **252**, 146-155.

Rogers J, Early P, Carter C, Calame K, Bond M, Hood L , Wall R (1980). Two mRNAs with different 3' ends encode membrane-bound and secreted forms of immunoglobulin mu chain. *Cell*. **20**, 303-312.

Ruiz M, Pallares N, Contet V, Barbi V , Lefranc MP (1999). The human immunoglobulin heavy diversity (IGHD) and joining (IGHJ) segments. *Exp Clin Immunogenet*. **16**, 173-184.

Schelonka RL, Zemlin M, Kobayashi R, Ippolito GC, Zhuang Y, Gartland GL, Szalai A, Fujihashi K, Rajewsky K , Schroeder HW, Jr. (2008). Preferential use of DH reading frame 2 alters B cell development and antigen-specific antibody production. *J Immunol*. **181**, 8409-8415.

Seifert M , Kuppers R (2009). Molecular footprints of a germinal center derivation of human IgM+(IgD+)CD27+ B cells and the dynamics of memory B cell generation. *J Exp Med*. **206**, 2659-2669.

Shlomchik MJ , Weisel F (2012). Germinal center selection and the development of memory B and plasma cells. *Immunol Rev*. **247**, 52-63.

Silverton EW, Navia MA , Davies DR (1977). Three-dimensional structure of an intact human immunoglobulin. *Proc Natl Acad Sci U S A*. **74**, 5140-5144.

Smith KG, Light A, Nossal GJ , Tarlinton DM (1997). The extent of affinity maturation differs between the memory and antibody-forming cell compartments in the primary immune response. *EMBO J*. **16**, 2996-3006.

Snapper CM, Marcu KB , Zelazowski P (1997). The immunoglobulin class switch: Beyond "accessibility"ed^eds): Immunity, pp. 217-223.

Stavnezer J, Guikema JE , Schrader CE (2008). Mechanism and regulation of class switch recombination. *Annu Rev Immunol*. **26**, 261-292.

Taylor JJ, Pape KA , Jenkins MK (2012). A germinal center-independent pathway generates unswitched memory B cells early in the primary response. *J Exp Med*. **209**, 597-606.

Tomlinson IM, Cook GP, Carter NP, Elaswarapu R, Smith S, Walter G, Buluwela L, Rabbitts TH , Winter G (1994). Human immunoglobulin VH and D segments on chromosomes 15q11.2 and 16p11.2. *Hum Mol Genet*. **3**, 853-860.

Tomlinson IM, Cox JP, Gherardi E, Lesk AM , Chothia C (1995). The structural repertoire of the human V kappa domain. *EMBO J*. **14**, 4628-4638.

Tomlinson IM, Walter G, Jones PT, Dear PH, Sonnhammer EL , Winter G (1996). The imprint of somatic hypermutation on the repertoire of human germline V genes. *J Mol Biol*. **256**, 813-817.

Tonegawa S (1983). Somatic generation of antibody diversity. **302**, 575-581.

van Dongen JJ, Langerak AW, Bruggemann M, Evans PA, Hummel M, Lavender FL, Delabesse E, Davi F, Schuuring E, Garcia-Sanz R, van Krieken JH, Droese J, Gonzalez D, Bastard C, White HE, Spaargaren M, Gonzalez M, Parreira A, Smith JL, Morgan GJ, Kneba M , Macintyre EA (2003). Design and standardization of PCR primers and protocols for detection of clonal immunoglobulin and T-cell receptor gene recombinations in suspect lymphoproliferations: report of the BIOMED-2 Concerted Action BMH4-CT98-3936. *Leukemia*. **17**, 2257-2317.

Volpe JM , Kepler TB (2008). Large-scale analysis of human heavy chain V(D)J recombination patterns. *Immunome Res*. **4**, 3.

Wei C, Anolik J, Cappione A, Zheng B, Pugh-Bernard A, Brooks J, Lee E-H, Milner ECB , Sanz I (2007). A New Population of Cells Lacking Expression of CD27 Represents a Notable Component of the B Cell Memory Compartment in Systemic Lupus Erythematosus. *The Journal of Immunology*. **178**, 6624-6633.

Williams SC, Frippiat JP, Tomlinson IM, Ignatovich O, Lefranc MP , Winter G (1996). Sequence and evolution of the human germline V lambda repertoire. *J Mol Biol*. **264**, 220-232.

Wu TT, Johnson G , Kabat EA (1993). Length distribution of CDRH3 in antibodies. *Proteins*. **16**, 1-7.

Xue K, Rada C , Neuberger MS (2006). The in vivo pattern of AID targeting to immunoglobulin switch regions deduced from mutation spectra in msh2-/- ung-/- mice. *J Exp Med*. **203**, 2085-2094.

Yin L, Hou W, Liu L, Cai Y, Wallet MA, Gardner BP, Chang K, Lowe AC, Rodriguez CA, Sriaroon P, Farmerie WG, Sleasman JW , Goodenow MM (2013). IgM Repertoire Biodiversity is Reduced in HIV-1 Infection and Systemic Lupus Erythematosus. *Front Immunol*. **4**.

Zemlin M, Schelonka RL, Ippolito GC, Zemlin C, Zhuang Y, Gartland GL, Nitschke L, Pelkonen J, Rajewsky K , Schroeder HW, Jr. (2008). Regulation of repertoire development through genetic control of DH reading frame preference. *J Immunol*. **181**, 8416-8424.

6 Abbreviations

Ag	antigen
bp	base pairs
CD	cluster of differentiation
CDR	complementarity determining region
CLP	common lymphoid progenitor
BCR	B cell receptor
FR	framework region
GC	germinal center
H	heavy (chain)
IGHV	immunoglobulin heavy chain variable region
IGHD	immunoglobulin heavy chain diversity region
IGHJ	immunoglobulin heavy chain juncional region
IMGT	international ImMunoGeneTics information system
N	non-germline encoded nucleotide
PCR	polymerase chain reaction
PBMC	peripheral blood mononuclear cells
RAG	recombination activating gene
RF	reading frame
RT-PCR	reverse-transcriptase PCR
SHM	somatic hypermutation

7 Acknowledgements

My particular thanks go to Prof. Dr. Li Yin who continuously supported and advised me during my time in Gainesville, Florida.

I am very grateful to Prof. Dr. Hans-Jürgen Mägert for initiating the contact to my laboratory in Gainesville and thus giving me the opportunity to carry out my master thesis there.

I would like to thank Prof. Dr. Maureen M. Goodenow for giving me the chance to work in her lab.

Further I want to thank Prof. Dr. Christiana Cordes for reviewing my thesis.

I am very grateful for the nice atmosphere and help from the Goodenow lab created by amazing people:

Dr. Julie C Williams, Steve Pomeroy, Kai-Fen Chang, Manju Karki, Xinrui Zhang, Sofia Appelberg, Amanda Lowe.

My special thanks go to David J. Nolan, who was a very nice neighbor and pop corn maker.

Furthermore I would like to thank my friends for making my stay in the U.S. unforgettable: Michelle, Kshitij, Alexandra, Katrina, Thomas, Eric, Frances…

I am grateful to Raghu, for his kindness and the funny conversations beyond work.

I would also like to thank René for his support from home

and finally

my whole family, which is very important to me.

8 Affirmation

Hereby, I declare that all the work presented in this master thesis was prepared by my own, carried out solely with the help of the literature and aids cited.

Göttingen, 13.04.2014

Hanna Lange

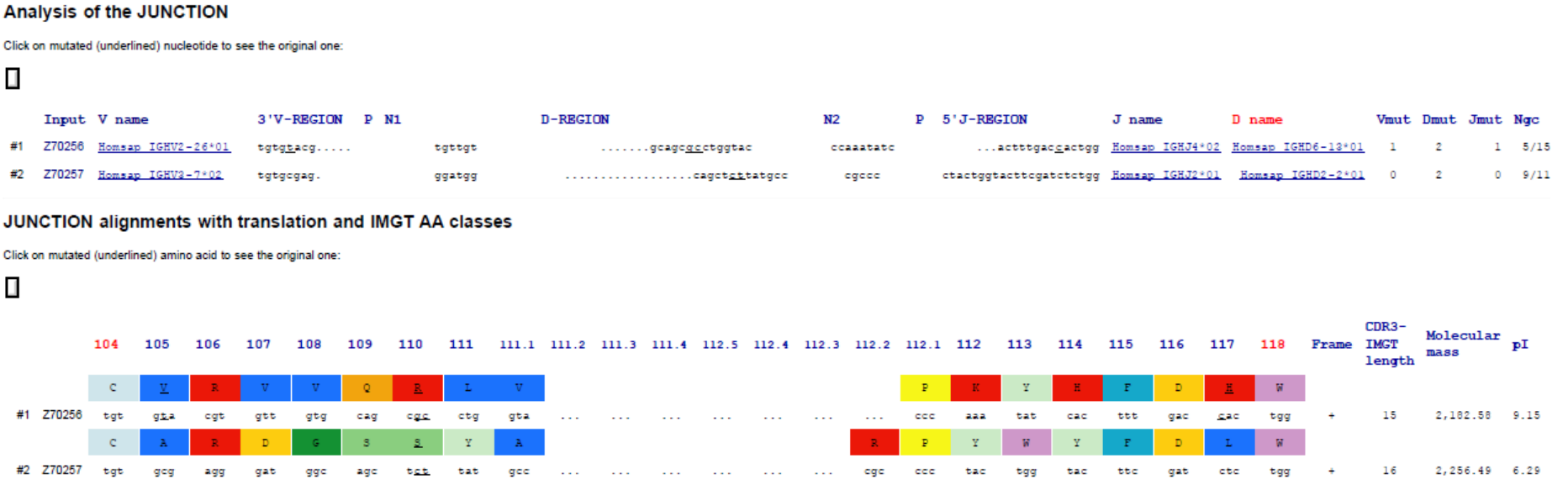
9 Appendix

Table A 1: Upstream primers for the second round PCR containing non-template specific adapter regions. The IGHV specific primer sequences are shown in green. All primers were purchased from Invitrogen.

Primer name	Sequence 5' -> 3'
A-VH2 FR1	CGTATCGCCTCCCTCGCGCCATCAGGTCTGGTCCTACGCTGGTGAAACCC
A-VH3 FR1	CGTATCGCCTCCCTCGCGCCATCAGCTGGGGGGTCCCTGAGACTCTCCTG
A-VH4 FR1	CGTATCGCCTCCCTCGCGCCATCAGCTTCGGAGACCCTGTCCCTCACCTG
A-VH5 FR1	CGTATCGCCTCCCTCGCGCCATCAGCGGGGACTCTCTGAAGATCTCCTGT
A-VH6 FR1	CGTATCGCCTCCCTCGCGCCATCAGTCGCAGACCCTCTCACTCACCTGTG
A-VH7/ VH1 FR1	CGTATCGCCTCCCTCGCGCCATCAGCTGGGGCCTCAGTGAAGGTCTCCTG

Table A 2: Downstream primers with adaptor sequences for the amplification of the IGHM and IGHG region. The IGHC specific primer sequences are displayed in green. All primers were obtained from Invitrogen.

Primer name	Sequence 5' -> 3'
TiB-MID1-Cμ2	CTATGCGCCTTGCCAGCCCGCTCAGACGAGTGCGTGGAATTCTCACAGGAGACG
TiB-MID2- Cμ2	CTATGCGCCTTGCCAGCCCGCTCAGACGCTCGACAGGAATTCTCACAGGAGACG
TiB-MID3- Cμ2	CTATGCGCCTTGCCAGCCCGCTCAGAGACGCACTCGGAATTCTCACAGGAGACG
TiB-MID4- Cμ2	CTATGCGCCTTGCCAGCCCGCTCAGAGCACTGTAGGGAATTCTCACAGGAGACG
TiB-MID5- Cγ1	CTATGCGCCTTGCCAGCCCGCTCAGATCAGACACGAGACCGATGGGACCTTGGTGGAAG
TiB-MID6- Cγ1	CTATGCGCCTTGCCAGCCCGCTCAGATATCGCGAGAGACCGATGGGACCTTGGTGGAAG
TiB-MID7- Cγ1	CTATGCGCCTTGCCAGCCCGCTCAGCGTGTCTCTAAGACCGATGGGACCTTGGTGGAAG
TiB-MID8- Cγ1	CTATGCGCCTTGCCAGCCCGCTCAGCTCGCGTGTCAGACCGATGGGACCTTGGTGGAAG



Analysis of the JUNCTION

Click on mutated (underlined) nucleotide to see the original one:

	Input	V name	3'V-REGION	P	N1	D-REGION	N2	P	5'J-REGION	J name	D name	Vmut	Dmut	Jmut	Ngc
#1	Z70256	Homsap IGHV2-26*01	tgtgtacg.....		tgttgt	.....gcagcgcctggtac	ccaaatatc		...actttgaccactgg	Homsap IGHJ4*02	Homsap IGHD6-13*01	1	2	1	5/15
#2	Z70257	Homsap IGHV3-7*02	tgtgcgag.		ggatgg	.....cagcttttatgcc	cgccc		ctactggtacttcgatctctgg	Homsap IGHJ2*01	Homsap IGHD2-2*01	0	2	0	9/11

JUNCTION alignments with translation and IMGT AA classes

Click on mutated (underlined) amino acid to see the original one:

		104	105	106	107	108	109	110	111	111.1	111.2	111.3	111.4	112.5	112.4	112.3	112.2	112.1	112	113	114	115	116	117	118	Frame	CDR3-IMGT length	Molecular mass	pI
		C	V	R	V	V	Q	R	L	V								P	K	Y	H	F	D	H	W				
#1	Z70256	tgt	gta	cgt	gtt	gtg	cag	cgc	ctg	gta	...	...	...	...	...	...	...	ccc	aaa	tat	cac	ttt	gac	cac	tgg	+	15	2,182.58	9.15
		C	A	R	D	G	S	R	Y	A							R	P	Y	W	Y	F	D	L	W				
#2	Z70257	tgt	gcg	agg	gat	ggc	agc	ttt	tat	gcc	...	...	...	...	...	...	cgc	ccc	tac	tgg	tac	ttc	gat	ctc	tgg	+	16	2,256.49	6.29

Figure A 1: Example of a IMGT Junction Analysis result of two input sequences. The nucleotide sequence of the 3'V, the whole D, the 5'J region and potential palindromic and non-template encoded nucleotides are shown in the upper part. In the bottom panel the corresponding amino acid sequences in the correct reading frame are denoted.

```
1(22) LTCAISGDSV STNTAAWSWI RQSPSRGLEW LGRTLYRSNK WHNEFVVSMR  50
2(11) GSLRL-CTAS GFTFGNFAMT WVRQTP-KGL EWLSTIFGGG FGTYSAD-V-
 3(9) GSLRL-CAAS GFIFTSYAMS WVRQAP-KGL EWVSGINTGG IGTYYAD-VK
 4(9) ---------- -S-L---N-- ---------- ----Y---QW F--DYAS-VK
                                          Junction
1(22) SRININPDTS KNQLSLHLDS VTPEDTAVYY CARDMGSTSP HSLGFWGQGT  100
2(11) G-FT-SR-N- --T-Y-QMN- LRV----I-- -VKYN-MVFQ SYYMDV--K-
 3(9) G-FT-SR-N- --T-Y-QIN- LGAA---I-- -VKHLSPGPN WTPFDY----
 4(9) --LA----A- --HF--L-S- ---D------ ---ERDYGRS ADFD------

1(22) PVTVSS  106
2(11) -T----
 3(9) ------
 4(9) L-----

1(22, 6.9%) V6-1*01  J4*02  D1-26*01
2(11, 3.4%) V3-23*01 J6*03  D2-15*01
3 (9, 2.8%) V3-23*01 J4*02  D2-2*01
4 (9, 2.8%) V6-1*01  J4*02  D4-23*01
```

Figure A 2: Subject 1: Amino acid alignment of the whole IGH VDJ region with a CDR3 amino acid length of 15. Due to the dominance with 22 sequences, cluster 1 was chosen as reference sequence in this alignment. Sequence cluster 1 and 4 align very exact, associating with the same IGHV and IGHJ genes. The bottom panel displays the IGH VDJ genes found for these clusters. The numbers in brackets indicate the number and percentage of sequences belonging to this cluster.

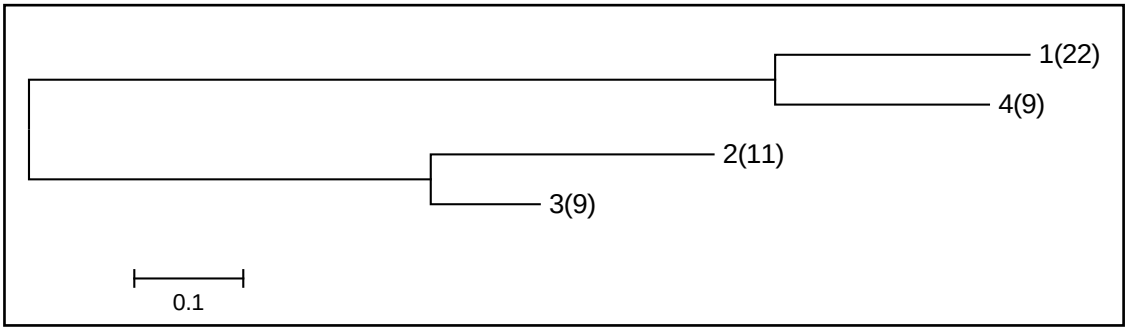


Figure A 3: Subject 1: Neighbor-joining tree of the four most abundant sequence clusters with a CDR3 length of 15 amino acids. Sequence clusters 1 and 4 are genetically more closely related with each other than with cluster 2 and 3. The numbers in brackets indicate the number of sequences belonging to this cluster.

```
1(21)  GGSLRLSCSA SGFTFGDHPM AWIRQTPQKG LEALITTSRN  40
2(13)  --------A- -----S-YA- N-V--A-G-- --WVSSI-G-
3(11)  -A-VKV--KT --YD-NKFAI S-V--A-GR- --WMGWINIY
4(11)  --------A- -----NSYA- H-V--A-G-- --WVAF--Y-

1(21)  AESKHYIASV EGRFTISRDD FRNTVHLQMT SITPDDAGLY  80
2(13)  GG-IYRAD-- K---IT---N S----Y---H -LRAE-TAV-
3(11)  NSDIK-NQKF Q----MTT-T STS-AFMELA -LR---TAI-
4(11)  GSE-YNAD-- K--------N SK--LY---N -LRVE-TAV-
                  Junction
1(21)  FCARNKISPA TQLWLPYDTF DVWGKGTIVS VSS         113
2(13)  Y---EGSKHS STWTVLPNY- -Y--Q--VIT ---
3(11)  Y---DENWDF CVHNCGLGY- -S--Q--L-- ---
4(11)  Y--KVVY-RS YLSDHY-YYM ---R--T-T ---

       1(21, 7.9%) V3-49*01 D2-8*01  J3*01
       2(13, 4.9%) V3-23*01 D6-13*01 J4*02
       3(11, 4.2%) V1-18*01 D3-3*01  J4*02
       4(11, 4.2%) V3-30*03 D6-25*01 J6*03
```

Figure A 4.Subject 1: Amino acid alignment of the IGH VDJ region with a CDR3 length of 20 amino acids. The sequence cluster 1 consists of 21 sequences with the same IGH VDJ region and was therefore considered as reference sequence in this alignment. The numbers in brackets indicate the number and percentage of sequences belonging to this cluster.

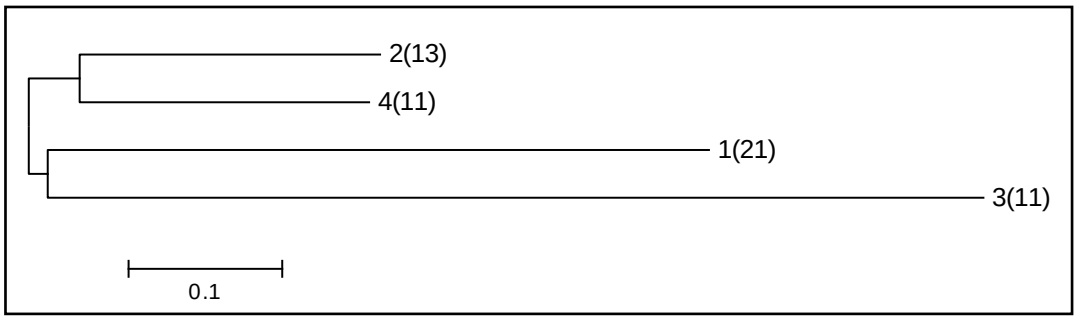


Figure A 5: Subject 1: Neighbor-joining tree of the four most abundant sequence clusters with a CDR3 length of 20 amino acids. Sequence clusters 2 and 4 are genetically more closely related with each other than with cluster 2 and 3. The numbers in brackets indicate the number of sequences belonging to this cluster.

```
1(20) GSLRLSCVVS GFSFSNYAMS WVSQAPGKGL EWVSAISGSD TGTYYTDSVK  50
2(16) -------AA- --T-N----- --R-TQ---- ----TF--GS DT--TA---R
3(16) -------AA- --T-T----N --R----REM ----S-T-GG HN--HAE---
4(14) A-VKV--KA- -YT--F-Y-H --R----Q-- --MGM-NP-G GS-THAQKFQ

                                          Junction
1(20) GRFTISRDNS KNTLYLHMNS LRAEDTAIYY CAKAPAGSCR GRSCYRLDFW  100
2(16) ---------- ------Q-T- --V------- ----TLPT-A -AL--NF-S-
3(16) ---------- ------Q--- -------V-- ---GRLET-S -VV--PF---
4(14) D-V-VT--T- TS-V-MELS- --S----V-- --LRGPYCSG -TCYDAF-I-

1(20) GQGTPVTVSS  110
2(16) ----L-----
3(16) ----L-S---
4(14) ----L-----

 1(20, 2.1%) V3-23*01  D2-15*01  J4*02
 2(16, 1.7%) V3-23*01  D2-8*02   J4*01
 3(16, 1.7%) V3-23*01  D2-21*02  J4*02
 4(14, 1.5%) V1-46*01  D2-15*01  J3*02
```

Figure A 6: Subject 2: Amino acid alignment of the IGH VDJ region with a CDR3 length of 18 amino acids. The sequence cluster 1 consists of 20 sequences with the same IGH VDJ region and was therefore considered as reference sequence cluster in this alignment. The numbers in brackets indicate the number and percentage of sequences belonging to this cluster.

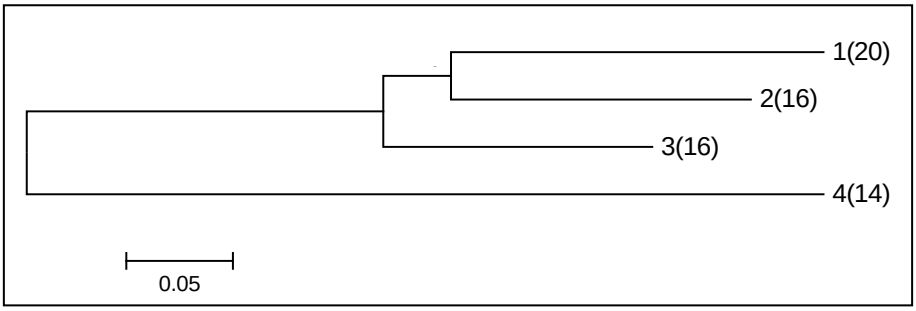


Figure A 7: Subject 2: Neighbor-joining tree of the four most abundant sequence clusters with a CDR3 length of 18 amino acids. Sequence clusters 1 and 2 are genetically more closely related with each other than with cluster 3 and 4. The numbers in brackets indicate the number of sequences belonging to this cluster.

```
1(28) SETLSLTCAV YGGSFSDYYW SWIRQPPGKS LEWIGEIDHS GSTSYNPSLQ  50
2(19) TLS-TC-VSG GSV-SGS--- T-V------G -----Y-HYT -A-N--RA-K
3(13) TLS-TC-VSG GSI-S-S--- G--------G --C--S-YY- -N-Y-----K
 4(9) ASVKVSCK-S GYTFT-YDIN WVRQATGQGL EWMGWMNTD- -N-G-AQNF-
                                          Junction
1(28) SRVTISVASS KSQLSLNLDS VTAADTAVYF CARVRVGCCD AYDLWGQGTV  100
2(19) ---A--ADT- EN-F--S-Y- ---------- -V-SVHYYYY YM-V--K--T
3(13) -------DT- -N-F--K-I- L--------Y ---LNS-PAW YF-Y----L
 4(9) G---MTRDT- I-TAY-E-S- LRSE-----Y -G--TSRPSW YF----R--L

1(28) VTVSS  105
2(19) I----
3(13) -----
 4(9) -----

1(28, 3.9%) V4-34*01 D5-24*01 J3*01
2(19, 2.7%) V4-61*01 D6-6*01  J6*03
3(13, 1.8%) V4-39*01 D5-12*01 J4*02
4( 9, 1.3%) V1- 8*01 D2-15*01 J2*01
```

Figure A 8: Subject 2: Amino acid alignment of the IGH VDJ region with a CDR3 length of 13 amino acids. The sequence cluster 1 consists of 28 sequences with the same IGH VDJ region and was therefore considered as reference sequence cluster in this alignment. The numbers in brackets indicate the number and percentage of sequences belonging to this cluster. While sequence clusters 2 and 3 have some amino acids in common with the reference sequence, cluster 4 displays a very distinct IGH VDJ amino acid composition

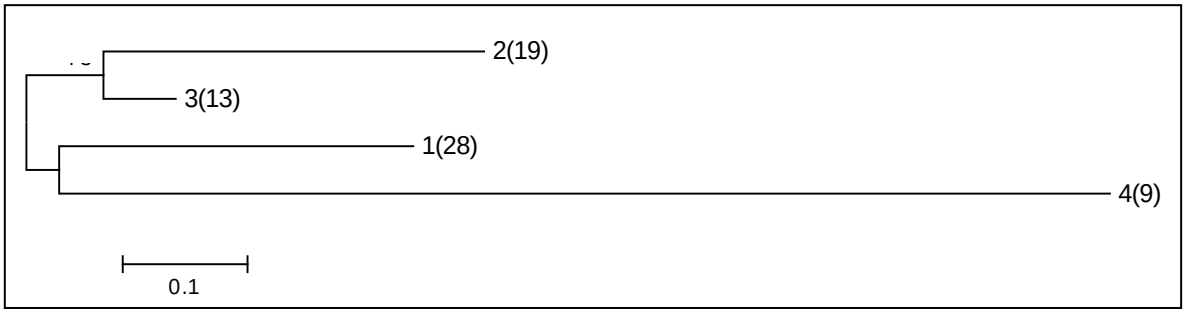


Figure A 9: Subject 2: Neighbor-joining tree of the four most abundant sequence clusters with a CDR3 length of 13 amino acids. Sequence clusters 2 and 3 are genetically more closely related with each other than with cluster 4. The numbers in brackets indicate the number of sequences belonging to this cluster.

```
1(29) GGSLRLSCAA SGFTFSSYAM SWVRQAPGQG LEWVSVISES ATTTSYADSV  50
2(13) -D--KI---- ---------- ---------- ---------- ----------
3(13) ---------- ------N--- A-------K- -----T-TGG SDK-Y-----
 4(9) -A-VKV--K- --Y--TNFGI ---------- ---MGW--GD NGN-N--QKL

                                                  Junction
1(29) KGRSTVSRDN SKNTLYLQMY SLRAEDTAVY YCAKKFRVGG THSFDIWGQG  100
2(13) ---------- ---------- ---------- ---------- ----------
3(13) ---F-I---- ---------N ---------- F---EC-SSS CTQ--P----
 4(9) Q--V-MTT-T -TS-A-MELG ---SD----- ---RE-SGYD GDW--P----

1(29) TMVTVSS  107
2(13) -------
3(13) -V-----
 4(9) -L-----

 1 (29, 6.43%) V3-23*01   D3-10*01  J3*02
 2 (13, 2.88%) V3-23*01   D3-10*01  J3*02
 3 (13, 2.88%) V3-23*01   D2-2*01   J5*02
 4 ( 9, 2.00%) V1-18*01   D5-12*01  J5*02
```

Figure A 10: Subject 4: Amino acid alignment of the IGH VDJ region with a CDR3 length of 14 amino acids. The sequence cluster 1 consists of 29 sequences with the same IGH VDJ region and was therefore considered as reference sequence cluster in this alignment. The numbers in brackets indicate the number and percentage of sequences belonging to this cluster. While sequence clusters 2 and 3 have some amino acids in common with the reference sequence, cluster 4 displays a very distinct IGH VDJ amino acid composition

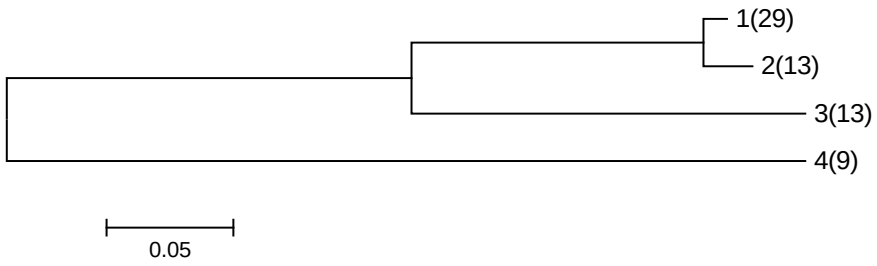


Figure A 11: Subject 4: Neighbor-joining tree of the four most abundant sequence clusters with a CDR3 length of 14 amino acids. Sequence clusters 1 and 2 are genetically closely related with each other. Sequence cluster 4 does not share the same branch and is therefore showing a greater genetic distance. The numbers in brackets indicate the number of sequences belonging to this cluster.