From the Institute of Experimental Dermatology
of the University of Lübeck
Acting Director: Prof. Dr. Ralf Ludwig

Do autoimmune bullous diseases, age and gender influence the B cell receptor repertoire of humans?

Dissertation
for Fulfillment of
Requirements
for the Doctoral Degree
of the University of Lübeck

from the Department of Natural Sciences

Submitted by

Julia Bischof
from Wolgast

Lübeck, September 29, 2016

First referee: Prof. Dr. Saleh Ibrahim

Second referee: Prof. Dr. Rudolf Manz

Date of oral examination: 09.01.2017

Approved for printing: Lübeck, 12.01.2017

Contents

Summary

The immune system is a complex network of cells and organs that mainly defends the body against disease. In particular B lymphocytes are important cellular components of the adaptive immune response. They produce highly diverse immunoglobulins to recognize antigens. The antigen binding sites of these receptors are individually generated for each cell. This leads to an almost unlimited receptor repertoire, which can bind various types of antigens.

Major advances in next generation sequencing opened the era of deep sequencing of B cell receptors (immunoglobulins or antibodies) of whole cell populations to determine the immunoglobulin repertoire. The antibody repertoire includes both hardwired features in the germline that are shaped over evolutionary time and somatic changes that are selected for by microorganisms and environmental antigens during the lifetime of an individual. Investigating the selective pressure, that acts upon B cells to enrich for useful B cell receptor specificities, helps to understand the processes that lead to autoimmune blistering diseases (AIBD), such as pemphigus vulgaris (PV) or bullous pemphigoid (BP).

While previous studies hint at antibody repertoire abnormalities in AIBD, there is no consistent abnormality that has been defined in a majority of PV or BP patients through the analysis of antigen-enriched B cells to date. We therefore wondered if a B cell subset specific approach is the key to understand how B cell tolerance breaks down in both diseases. To address this question I statistically analyzed the variable heavy chain (VH) DNA repertoires of four B cell populations (sorted for naïve, IgM+/IgM- memory B cells, and plasmablasts) from 10 PV patients, 10 BP patients and 20 healthy controls, using IMGT/HighV-QUEST and a self-written R script, *bcRep*. In this package I mainly focused on the establishment of algorithms to assemble sequences into clones and statistically analyze gene usage and gene-gene combinations of V, D and J segments. Further the diversity of hypervariable complementary determining regions, which are important for the antigen binding, as well as mutation patterns can be analyzed using this package.

Although PV and BP are both autoimmune bullous diseases, they are caused by different mechanisms, which can be nicely seen when analyzing the B cell receptor repertoire. Differences between PV patients and healthy controls are mainly based on CDR3 sequence diversity and only few changes in gene usage, predominantly in IgM- memory cells and plasmablasts. However BP patients can be distinguished from healthy controls mainly by different gene usage in IgM+, but also IgM- memory cells and plasmablasts. Thus the B

cell receptors of patients suffering from PV or BP seem not to be fundamentally different in the naïve repertoire, but may be influenced by different reactions during immune response.

Further I wanted to investigate age and gender specific differences in the B cell receptor repertoire of healthy individuals. Does the B cell repertoire change steadily with age or does it basically stay the same? Do females and males show already differences in the naïve repertoire or at least in the memory repertoire, because their repertoire develops and/or reacts differently? Only few details are known about age or gender dependent development of certain B cell subsets. My data shows, that there are stronger differences in age, than in gender. It is not only the memory repertoire that shapes individuals by remembering antigens. But a wider repertoire means also a higher chance of developing autoreactive cells and causing damage.

In addition I identified genes of the variable region that show significantly different abundances in females and males. Some results can confirm findings from previous studies. Furthermore, genes were found that are overrepresented in women and that are already known to be associated with autoimmune diseases, like rheumatoid arthritis. It is already known that women are more prone to autoimmunity and thus a changed gene usage could contribute to trigger this. Further there are genes in women that show age-dependent quadratic trends of gene abundances with a peak at an age of around 55 years that may be associated with hormones, like estrogen or androgen. Such dependencies could not be found in men.

Zusammenfassung

Das Immunsystem ist ein komplexes Netzwerk aus Zellen und Organen, dessen Hauptaufgabe es ist, den Körper vor Krankheitserregern zu schützen. Speziell B-Lymphozyten sind wichtige zelluläre Bestandteile der adaptiven Immunantwort. Sie produzieren Immunoglobuline zur Antigenerkennung. Die Antigenbindungsstelle dieser Rezeptoren ist individuell in jeder Zelle, so dass die Populationen dieser Zellen ein fast unbegrenztes Rezeptorrepertoire ausbilden, welches in der Lage ist jegliche Antigene zu binden.

Durch große Fortschritte in der Sequenzierungstechnologie, ist es mittlerweile möglich B-Zell Rezeptoren (Immunoglobuline oder Antikörper) gleichzeitig in ganzen Zellpopulationen zu sequenzieren und damit ihr Antikörperrepertoire zu bestimmen. Dieses beinhaltet sowohl fest veranlagte Charakteristiken in den Keimbahnsequenzen, die sich über die Zeit hinweg durch Selektion entwickelt haben, als auch somatische Veränderungen, die durch Krankheitserreger und Umweltfaktoren beeinflusst werden. Durch Untersuchungen des selektiven Drucks, der auf B-Zellen wirkt, um nützliche B-Zell-Rezeptoren auszubilden, können Umstände, die zu autoimmunen blasenbildenden Erkrankungen (engl.: *autoimmune bullous disease*, AIBD) führen, besser verstanden werden. In dieser Arbeit wird das B-Zell-Repertoire zweier AIBD-Subtypen, Pemphigus Vulgaris (PV) und Bullous Pemphigoid (BP), vor allem statistisch untersucht.

Während vorhergehende Studien bereits vermuteten, dass es Auffälligkeiten im Antikörperrepertoire bei AIBD-Patienten gibt, konnten bisher keine Unregelmäßigkeiten im B-Zell-Repertoire von PV und BP Patienten festgestellt werden. Deswegen stellte sich die Frage, ob womöglich eine Studie, die spezifische B-Zell-Subgruppen untersucht, helfen könnte, um die Mechanismen zu verstehen, die zur Aufspaltung der B-Zell-Toleranz in beiden Patientengruppen führen. Um dieses Problem zu untersuchen, wurde die DNA der variablen Regionen der schweren Ketten von vier verschiedenen Subpopulationen untersucht: naive B-Zellen, positive und negative IgM Gedächtnis B-Zellen und Plasmablasten. Zur statistischen Analyse von Unterschieden zwischen Patienten und gesunden Kontrollen (PV: n=10, BP: n=10, Kontrollen: n=20), wurden IMGT/HighV-QUEST und ein von mir geschriebenes R Paket namens *bcRep* genutzt. Hierbei lag der Fokus nicht nur in der Etablierung von Algorithmen zur Gruppierung von Sequenzen zu Klonen, sondern auch im Nutzverhalten von Genen der variablen Region und in der Diversität von hypervariablen Sequenzen, die für die Antigenbindung wichtig sind. Auch Mutationen auf Nukleotid- und Aminosäureebene können mit Hilfe des R Pakets untersucht werden.

Obwohl PV und BP beide zu der Gruppe der autoimmunen blasenbildenden Krankheiten gehören, basieren sie doch auf grundlegend verschiedenen Mechanismen. Entsprechend wurden in dieser Arbeit auch spezifische Unterschiede im B-Zell-Repertoire bei diesen Erkrankungen festgestellt. Unterschiede zwischen PV Patienten und gesunden Kontrollen sind vor allem in der Diversität von CDR3 Sequenzen zu sehen. Beide Gruppen scheinen sich nur wenig in der Nutzung spezifischer Gene der variablen Region zu unterscheiden, wenn jedoch, lassen sich hauptsächlich Unterschiede in den IgM negativen Zellen und Plasmablasten finden. In BP hingegen, sind die beiden Gruppen vor allem in der Nutzung einzelner Gene der variablen Region unterscheidbar, auch hier vorzugsweise in IgM positiven und negativen Zellen und Plasmablasten. Die B-Zell-Rezeptoren der beiden Patientengruppen scheinen sich nicht bereits im grundlegenden naïven Repertoire zu unterscheiden, sondern werden erst durch verschiedene Reaktionen im Laufe der Immunantwort verändert.

Zusätzlich wurden alters- und geschlechtsspezifische Unterschiede im B-Zell-Repertoire gesunder Kontrollen untersucht. Hierbei interessierte nicht nur, ob sich das B-Zell-Repertoire im Verlauf des Alters ändert, sondern auch, in welchen B-Zell-Subgruppen mögliche Veränderungen auftreten. Unterscheiden sich Frauen und Männer bereits im naïven Repertoire oder gibt es möglicherweise nur Unterschiede im Gedächtnisrepertoire? Bis jetzt ist nur wenig über die alters- und geschlechtsabhängige Entwicklung von speziellen B-Zell-Populationen bekannt. Meine Daten zeigen jedoch, dass es mehr alters-, als geschlechtsabhängige Veränderungen gibt. Nicht nur das Gedächtnisrepertoire, sondern auch die Natur spontaner Immunantworten verändern sich im Laufe des Lebens. Ein breiteres Repertoire bedeutet jedoch nicht nur, dass umso mehr Krankheitserreger erkannt werden können, sondern auch, dass sich die Wahrscheinlichkeit einer autoreaktiven Reaktion und somit Schädigung von Zellen vergrößert.

Außerdem wurden in dieser Arbeit Gene der variablen Region identifiziert, deren Ausprägungen sich in B-Zell-Rezeptoren beider Geschlechter signifikant unterschieden. Zusätzlich wurden in Frauen überrepräsentierte Gene gefunden, für die bereits Assoziationen mit Autoimmunerkrankungen, wie rheumatoider Arthritis, bekannt sind. Dies passt zu dem Befund, dass es viele Autoimmunerkrankungen gibt, die bei Frauen häufiger auftreten als bei Männern und somit könnte ein verändertes Nutzverhalten dieser Gene zu möglichen Ursachen beitragen. Weiterhin wurden Gene gefunden, die einen annähernd quadratischen Verlauf der Häufigkeit in Abhängigkeit des Alters zeigten, meist mit einem Maximum/Minimum bei einem Alter von ca. 55 Jahren. Speziell diese Gene können auf eine Assoziation mit Geschlechtshormonen hinweisen. Bei Männern wurden solche Verläufe nicht gefunden.

1 Introduction

The immune system is a complex network of cells and organs that defends the body against disease or other potentially damaging foreign microorganisms [**?**]. Lymphocytes, in particular B and T cells, are the major cellular components of the adaptive immune response. Highly diverse immunoglobulins (Ig) and T cell receptors (TCR) provide specific immune reactions due to pathogen recognition.

Major advances in next generation sequencing (NGS) opened the era of deep sequencing of B and T cell receptor repertoires. The antibody repertoire includes both hardwired features in the germline that are shaped over evolutionary time [**?**] and somatic changes that are selected for by microorganisms and environmental antigens during the lifetime of an individual [**?**]. Germline-encoded features of the antibody repertoire include the variable (V), diversity (D) and joining (J) gene segments at the heavy chain locus and V and J segments at the light chain locus. Somatically generated changes in the antibody repertoire arise due to V(D)J recombination and somatic hypermutation (SHM). The expressed antibody repertoire reflects not only these mechanisms of diversification, but also the selective pressures that act upon the B cells themselves to enrich for useful B cell receptor (BCR) specificities while eliminating or inactivating cells with harmful BCR specificities. Causes for selective pressures are pathogens, environmental antigens or chronically present endogenous antigens, such as bacterial flora [**?**]. Investigating these selective pressures helps to understand the processes that lead to autoimmune blistering diseases (AIBD), such as pemphigus vulgaris (PV) or bullous pemphigoid (BP).

Pemphigus vulgaris is a chronic and severe AIBD, characterized by flaccid blistering and erosions of mucous membranes and skin. With some exceptions, the sera of PV patients contain circulating autoantibodies directed to desmoglein 3 and, in case of skin involvement, additionally to desmoglein 1 [**?** **?** **?**]. A multitude of evidence exists that these autoantibodies have a direct pathogenic effect [**?**].

Bullous pemphigoid is a subepidermal AIBD, which is characterized by circulating immunoglobulin G (IgG) autoantibodies against two hemidesmosomal proteins: BP230, which is a cytoplasmic protein, and BP180, which is a transmembrane protein of the cutaneous basement membrane zone [**?** **?**]. BP is currently the most common AIBD and occurs slightly more often in females compared to males and only in the elderly [**?** **?**].

Treatment of PV and BP mainly aims to suppress of the humoral immune response, and remove of pathogenic autoantibodies from the circulation. Glucocorticosteroids are a

mainstay of any treatment regimen [**?**]. In treatment refractory cases, the B cell depleting anti-CD20 antibody rituximab proved to be helpful, and may be combined with removal of serum IgG by protein A immunoapheresis [**?**].

While previous studies hint at antibody repertoire abnormalities in AIBD, there is no consistent abnormality that has been defined in a majority of PV and BP patients through the analysis of antigen-enriched B cells to date. This leads to the question if a fundamentally different approach, which focuses of particular B cell subsets, is needed to understand how B cell tolerance breaks down in both diseases.

Further I wanted to investigate age and gender-specific differences in the B cell receptor repertoire of healthy controls. Only few details are known about age or gender dependent development of certain B cell subsets (in this case naïve, IgM memory cells and plasmablasts) [**?** **?**]. But it is already known, that the immune system of the elderly is less able to respond effectively to infectious challenges [**?**]. Although the efficacy of the human antibody response is decreased with age, the total number of B cells remains constant [**?** **?**]. Only in very old age does it decrease; but long after a loss of functional competence is observed [**?**].

Men and women differ in several aspects of physiology, including immune responses. Women produce more elevated circulating levels of antibodies than men do [**?**]. Consequently their tendency to produce higher levels of autoantibodies and other factors like hormones or incomplete X chromosome inactivation lead to the fact that there are many autoimmune diseases which appear more often in females than males, but men suffer more from infectious diseases, cancer and death compared to women of the same age [**?** **?**]. Men and women experience the same types of aging-related changes to the immune system, but males may experience them earlier than women [**?**].

Immune gender differences have been reported [**?**], but are rarely studied in elderly humans. There are only few studies comparing men and women in old age, when hormone differences are less marked. Furthermore most of these studies analyze pooled B cells, rather than investigating differences in B cell subsets. Therefore I studied not only differences of four different B cell groups in young and old individuals, but also separately in females and males.

1.1 The immune system

1.1.1 Organization of the immune system: the innate and adaptive immunity

Most pathogens trigger immune response by activating the innate immune system [?]. The skin and mucous membranes are the first line of defense against potentially damaging foreign microorganisms, since they build a physical and chemical barrier against infections [? ?]. Microorganisms, which cross that barrier come upon cells, which initiate the innate immune response. Beyond this there are also non-cellular factors including the complement system, which can be activated with or without antibodies [? ?]. Functions of the complement system include the opsonization of pathogens, lysis of infected cells, clearance of immune complexes and activation of other components of the immune system [? ? ?].

Leukocytes play a crucial role in the orchestration of innate and adaptive immune responses. [?]. These cells arise from the bone marrow and some of them also mature there [? ?]. Afterwards, they migrate to the peripheral tissues or circulate in the blood and lymphatic system. All of these cells originate from the same progenitor cells: pluripotent hematopoietic stem cells in the bone marrow [? ?]. Hematopoietic cells develop from stem cells with a restricted potential, like direct progenitor cells of red blood cells, blood platelets and of the lymphatic and myeloid leucocytes [? ?]. Myeloid progenitor cells give rise to macrophages, granulocytes, mast cells and dendritic cells (DCs) of the innate immune system, but also to megakaryocytes and red blood cells [?].

Macrophages appear almost everywhere. They are matured forms of monocytes, which circulate in the blood and migrate to tissues, where differentiation takes place [? ? ?]. Macrophages are relatively long living cells, which engulf and kill microorganisms which could potentially damage the body [?]. Furthermore they contribute to inflammatory processes and secrete signaling molecules, which initiate immune responses [? ?]. These cells recognize microorganisms due to special polysaccharides or glycoproteins, bind them and elaborate proteins including cytokines and chemokines, which, in turn, initiate inflammation [? ?]. Cytokine is a general term for proteins that are secreted by cells and influence the behavior of cells with appropriate receptors [? ?]. Chemokines are a subclass of cytokines that mediates and controls cellular migration [?]. When they initiate an inflammation process, cells and molecules of the innate immune system are directed into the tissue, where they kill pathogens [? ?]. Further lymphocytes become activated and initiate the adaptive immune response [? ?]. Afterwards effectors of the adaptive immune response, including T cells and B cells, are directed to the site of infection [?].

The granulocytic lineage includes neutrophils, eosinophils and basophils [?]. Mature granulocytes are short living cells and are produced in increased numbers during immune

responses, in which they migrate from the blood to the sites of infection [?].

Mast cells are large cells, located in connective tissue all over the body. They differentiate in the tissue and have an important function in allergic processes [?].

One of the most important cells of the immune system is the dendritic cell [?]. These are immature cells which migrate from the bone marrow into the blood and finally into the tissues [? ?]. They eliminate microorganisms and present antigens to T cells [? ?].

Macrophages are located in tissues and are the first line of defense against bacteria, which they can detect due to pattern recognition receptors [? ?]. Activation of these receptors leads to absorption and killing of bacteria [?]. Further they can release cytokines or chemokines to amplify immune responses [? ?]. Macrophages can respond to viruses, parasites and fungi [? ?].

Triggering of the adaptive immune response begins with the absorption of microorganisms by immature dendritic cells in the inflamed tissue [? ?]. These cells have similar receptors on their surface, like macrophages and neutrophils, and detect general characteristics of pathogens, for example bacterial polysaccharides [?]. The function of dendritic cells is not primarily to kill pathogens, but to move antigens of pathogens to the peripheral lymphatic tissues and present them to T cells [? ?]. This leads to stimulation of T cells, proliferation and differentiation [?]. For most types of immune responses B cells require T cell help for full activation; hence stimulation of T cells is an essential component of most immune responses [?]. Activated dendritic cells secrete also cytokines which influence the innate and adaptive immune response [?].

B and T cells differ with respect to their antigen receptors, which have distinct chemical structures. Both types of antigen receptors (BCR and TCR) are generated by gene rearrangement, but pass different stages of development [?]. The BCR recognizes and binds the antigen directly [?]. In contrast, the TCR recognizes fragments of peptides, which are bound to MHC molecules (major histocompatibility complexes) on the surface of antigen presenting cells [?]. T cells expressing CD4 molecules are activated by peptides presented in the context of MHC class II. Once activated, these cells can migrate towards the B cell zone to interact with antigen-activated B lymphocytes, thereby providing additional signals to the activated B cells required to yield an antibody response against protein antigens. Serum Ig efficiently fights extracellular pathogens, but also bocks spreading of intracellular microbes [?].

BCRs, also named membrane bound immunoglobulin or antibody, are expressed on the surface of B cells [?]. By alternative splicing, the membrane anchor of these molecules can be removed, which produces a soluble form, retaining its original antigen specifity, but not serve as a BCR anymore, but as a soluble effector molecule [? ?]. In response to an antigen and adequate co-stimulation, B cells can differentiate into memory B cells or plasma cells. The latter secrete soluble Ig, or antibodies [? ?]. Differentiation into memory cells and plasma cells is accompanied by proliferation, which leads to the

expansion of the original B cell clone. This process is called clonal expansion [? ?].

1.1.2 The structure of a B cell receptor

All antibodies have the same structure and are called Igs [? ?]. They have a Y shape and consist of constant and variable regions. The constant regions comprise the base of the Y and the variable regions are positioned at the tips. The constant regions interact with other cells of the immune system via Fc receptors. The variable regions bind antigen [? ?].

Igs are tetrameric proteins that consist of two different polypeptide chains, each occurring two times, respectively: the heavy chain (H chain) and the light one (L chain) (see Fig. 1.1 A) [? ?]. The heavy chains are connected through disulfide bonds [? ?]. Each heavy chain is also linked to a light chain via disulfide bonds [? ?]. The two H chains and L chains are identical respectively [?]. Thus the antibody consists of two identical binding sites for antigens [? ?].

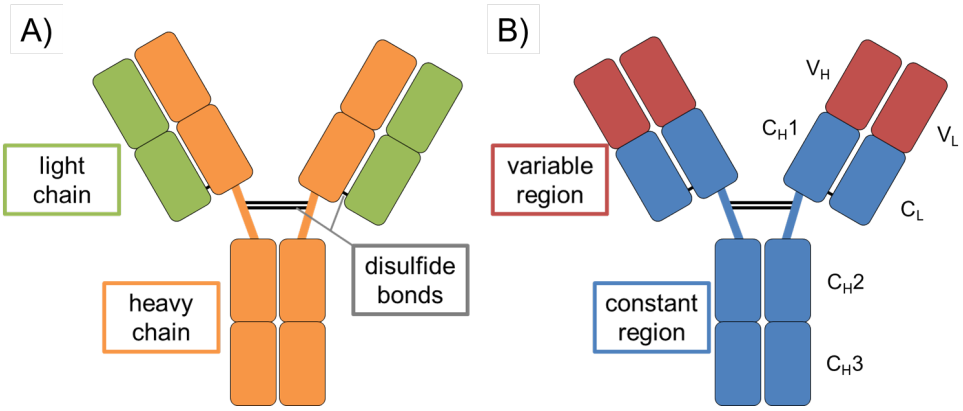

Figure 1.1: Structure of an antibody. A) Immunoglobulins consist of two different types of polypeptide chains: the light (green) and heavy chains (orange), which are joined via disulfide bonds (gray). B) Each heavy and light chain consists of a variable (red) and constant (blue) region. Note: IgG is shown as an example, other immunoglobulins have similar structures, see Fig. 1.6

There are two types of light chains, which are called λ- or κ-chains [? ? ?]. An immunoglobulin only consists of one type, never of both [? ?]. It seems that there are no functional differences between both types and both can appear in all Ig subclasses [? ?]. The ratio of λ- and κ- chains differs between species. In mice the κ/λ ratio is 20:1, in human 2:1, in bovines 1:20 [? ?].

The antibody effector functions are determined by the H chain constant region [? ?]. Heavy chain constant regions are divided into five isotypes, which specify the functional activity of the antibody molecule [? ?]. The five immunoglobulin isotypes are IgM, IgD, IgG, IgA and IgE (see section 1.1.3 Transitional and mature B cells) [? ?]. The corresponding heavy chains are called μ, δ, γ, α and ϵ [?]. IgG is the most abundant

isotype in the blood and can be divided into IgG1, 2, 3 and 4 [?]. Whereas IgA is the most abundant subtype in the whole body and can also be divided into several subclasses.

Further each chain consists of constant and variable regions (Fig. 1.1 B) [? ?]. The constant region can have one of four or five types and determines the effector function [? ?]. The variable region determines the specificity of the antigen binding [? ?].

Both chains of the antibody consist of protein domains of a similar folded structure [?]. There are differences between constant and variable regions. Each domain comprises two beta sheets which are connected via disulfide bonds [? ?]. One of the most important differences between constant and variable regions is in the length of both domains [?]. The variable region is longer and contains one additional loop (see Fig. 1.2) [?]. Most of the amino acids, that are consistent in V- and C-regions, have a central position and are important for the structure stability [?].

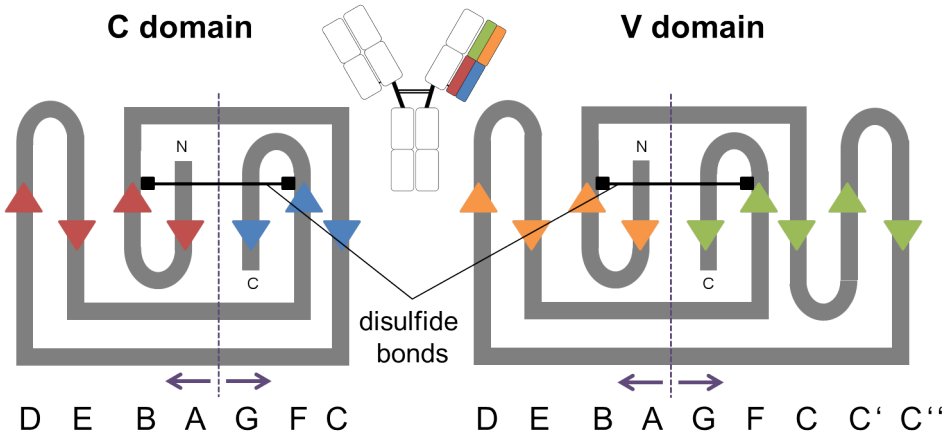

Figure 1.2: Structure of the variable (V) and constant (C) regions of a light chain of Immunoglobulins. Unfolded chains are shown. Each domain consists of several polypeptide chains that are anti-parallel folded to two beta sheets (red and blue for C domain; orange and green for V domain). There are two variable segments (C' and C") which only appear in the V domain, but not in the C domain.

The variable regions of antibodies differ from each other. However the degree of variability is not equally distributed over the whole region, but concentrated in some segments (see Fig. 1.3) [?]. There are three variability hotspots which contain hypervariable regions (HV1, HV2, HV3) [? ?]; these regions are also referred to as the complementary determining regions (CDRs). In H chains the positions are at around amino acids 30-36, 49-65 and 95-103 [?]. In L chains they are positioned at amino acid 28-35, 49-59 and 92-103 [?]. The most variable region is HV3 [? ?]. The sequence segments between those parts are less variable and are called framework regions (FR1, FR2, FR3, FR4) [? ?].

The structural basis of the domain is given through framework sequences that form beta sheets [? ?]. The hypervariable regions build loops at the border of the β barrel, which

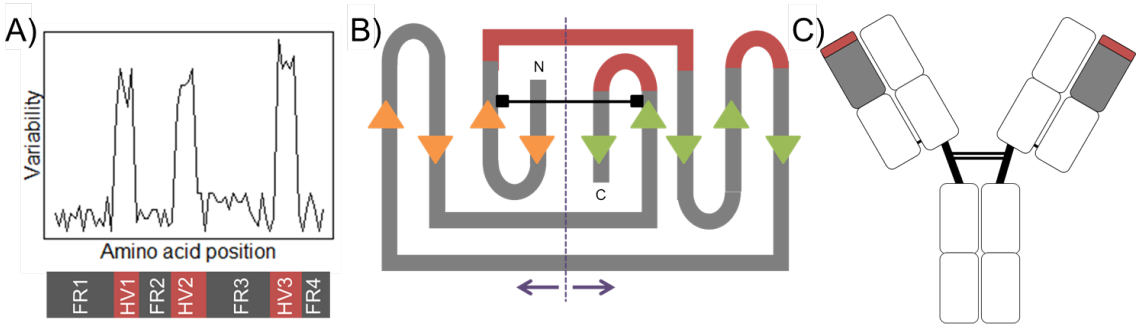

Figure 1.3: Hypervariable regions are restricted to some parts of the folded structure (example: V domain of the light chain). A) Variability plot: the degree of variability at each position equals the ratio of the number of different amino acids at this position and the frequency of the expected number of amino acids at this position. B) An unfolded light chain is shown. The hypervariable regions are close to each other. C) When heavy and light chains are joined, the six hypervariable regions lie next to each other and form the antigen binding site.

lie next to each other in the folded protein [**?**]. Thus the sequence variability is not only limited to some parts of the variable region, but also assigned to a special region on the surface of the molecule (see Fig. 1.3 B, C) [**?**]. When V domains of light and heavy chains are connected, the hypervariable regions of both chains are also connected and comprise the antigen binding site [**?**]. The six hypervariable loops (three of the H chain and three of the L chain) determine the antigen specificity due to a structure, which is complementary to the antigen (complementary determining region; CDR1, CDR2, CDR3) [**?**]. Hence, one possibility to generate antibodies of different specificity underlies the combinations of variable regions of L and H chains, which contributes to the huge diversity of antibodies [**?** **?**]. Additionally the antibody repertoire can increase due to the connection of two D segments [**?**]. This is a rare event, but occurs in humans in around 5% of antibody sequences, which leads to exceptionally long CDR3 loops in heavy chains [**?**].

The interaction of antibodies and antigens can be influenced by high concentrations of salt, extreme pH values or detergents [**?**]. Antibody-antigen binding is a reversible, non-covalent interaction [**?**]. Forces like electrostatic interactions, hydrogen bonds, van-der-Waals bonds and hydrophobic binding all contribute to antigen binding [**?**]. Receptors are cross-reactive, which means that one receptor can bind many antigens and vice versa [**?** **?**].

1.1.3 The maturation of a B cell receptor

There are complex mechanisms leading to the development of a diverse receptor repertoire to defend against pathogens [**?**]. Not every receptor variant can be encoded in the genome [**?**]. This would lead to a much higher number of genes encoding antigen receptors, than other genes. Instead, the variable region is encoded in several segments which are

connected via somatic hypermutation (SHM) during the development of lymphocytes [? ? ?]. This process is called gene rearrangement or V(D)J recombination [? ?]. The total V region than consists of two or three types of gene segments, each existing in several copies in the germline genome [?]. There are primary and secondary rearrangements [?]. The primary ones consist of V(D)J recombinations in order to generate an antibody repertoire [?]. The secondary rearrangements refer to one or more rearrangements after the primary one to rescue unproductive cells or alter BCR specificity when B cell are self-reactive [?]. The selection of gene segments is random and the huge number of possible combinations contributes to the diversity of the receptor repertoire. In humans there are more than 1000 different antibody molecules [?]. But the number of antibodies to a specific antigen at a single point in time is limited due to the total number of B cells in an individual [?].

In non-lymphoid cells the gene segments encoding variable region of the Immunoglobulin are relatively far away from gene segments encoding the constant region [?]. In contrast, following completed gene rearrangement in the mature B cell, the V region is located near the C region [?].

The phases of the B cell development are: pro B cell, large and small pre B cells and mature B cells [? ? ?] (see Fig 1.4). In each step of the development process, the B cell undergoes quality control steps to assure that the antibody rearrangements are functional [? ? ?]. This serves as a signal for the cell to enter the next phase of development. There are several options for such reorganizations, which increase the likelihood of expressing a functional antigen receptor, but there are specific check points that control expression of a B cell to a receptor with only one specificity [? ? ?].

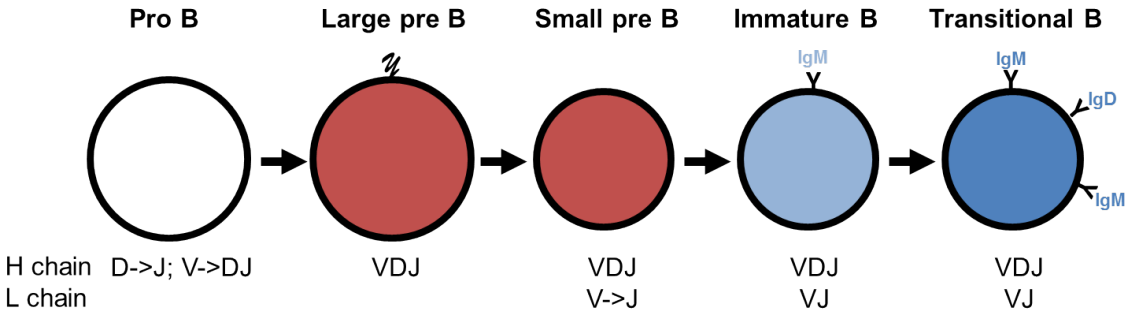

Figure 1.4: Schematic representation of B cell development. In pro B cell VDJ rearrangement of the heavy chain takes place. In large pre B cells H chains are paired with surrogate L chains, followed by a clonal expansion of H chain positive pre B cells and the development of small pre B cells. In this phase light chains are rearranged and production of functional L chains promotes emergence of IgM+ immature B cells, which then differentiate into IgM+ IgD+ transitional cells.

1.1.3.1 Pro B cells

The first cell of the B line, the pro B cell, develops from the common lymphoid progenitor, where first gene rearrangement takes place [**?**]. In early pro B cells, D and J genes of heavy chains are joined first (see Fig. 1.5) [**?** **?**]. This occurs on both alleles of the locus and the cell develops to a late pro B cell [**?**]. In humans most D-J combinations are potentially beneficial due to the translation of D gene segments in all three reading frames, without occurrence of stop codons [**?**]. No mechanisms are necessary to make D-J combinations productive [**?**]. The next step is comprised of the rearrangement of a V gene segment to a D-J sequence of a heavy chain (see Fig. 1.5) [**?**]. This happens only on one chromosome [**?**].

Special transcription factors of the pro B cell (transcription factor 3, TCF3 or E2A; early B cell factor, EBF) induce expression of several proteins, which are important for gene rearrangement, for example the recombination-activating genes RAG-1 and RAG-2 of the V(D)J recombinase [**?** **?**]. If those transcription factors are missing, the connection of the D and J regions of the heavy chain are inhibited [**?** **?**]. Another important protein, which is induced by E2A and EBF, is the transcription factor Pax-5 (paired box protein) [**?** **?**]. This is a B cell specific activator, which influences genes encoding the CD19 component of the B cell co-receptor and Igα [**?** **?**]. Igα is an important component for the pre B cell receptor, as well as the B cell receptor [**?**]. If Pax-5 is missing, development of pro B cells is interrupted, but pro B cells can still be prompted to develop to T cells or myeloid cells [**?** **?**]. Further Pax-5 induces expression of a B cell linker protein (BLNK), which is important for the development of pro B cells and for signaling of the matured B cell antigen receptor [**?** **?**]. It mobilizes proteins that are part of the intracellular signaling pathway of the antigen receptor [**?**].

The diversity of antigen receptor repertoires of B cells is further increased by the enzyme terminal deoxynucleotidyl transferase (TdT), which is expressed by pro B cells and adds nontemplated nucleotides into the junctions between rearranged gene segments [**?**]. This expression is reduced in the pre B cell phase during the rearrangement of the light chain [**?**].

For the correct joining of the gene segments explicit mechanisms exist. Not only the correct position of each segment, but also the joining of V, D and J segments and segments of the same class, are important [**?**]. DNA rearrangement is controlled by conserved DNA sequences that are located next to the recombination loci [**?**]. These sequences are called recombination signal sequences (RSS) and always have the same structure: a conserved region, the coding region, a non-conserved spacer region of 12 or 23 base pairs (bp) and again a conserved sequence [**?** **?**]. Normally a recombination only takes place in gene segments of the same chromosome [**?**]. Rearranegments typically follow the 12/23 rule, where a gene segment flanked by an RSS with a 12 bp spacer can only be joined to a gene

segment flanked by an RSS with a 23 bp spacer [? ?]. For example, in the mouse the V and J gene segments of heavy chains are flanked by RSSs with 23 bp spacers and D segments are flanked by 12 bp spacers [? ?]. That is the reason why D and J or V and D are joined, but V is not directly joined to J.

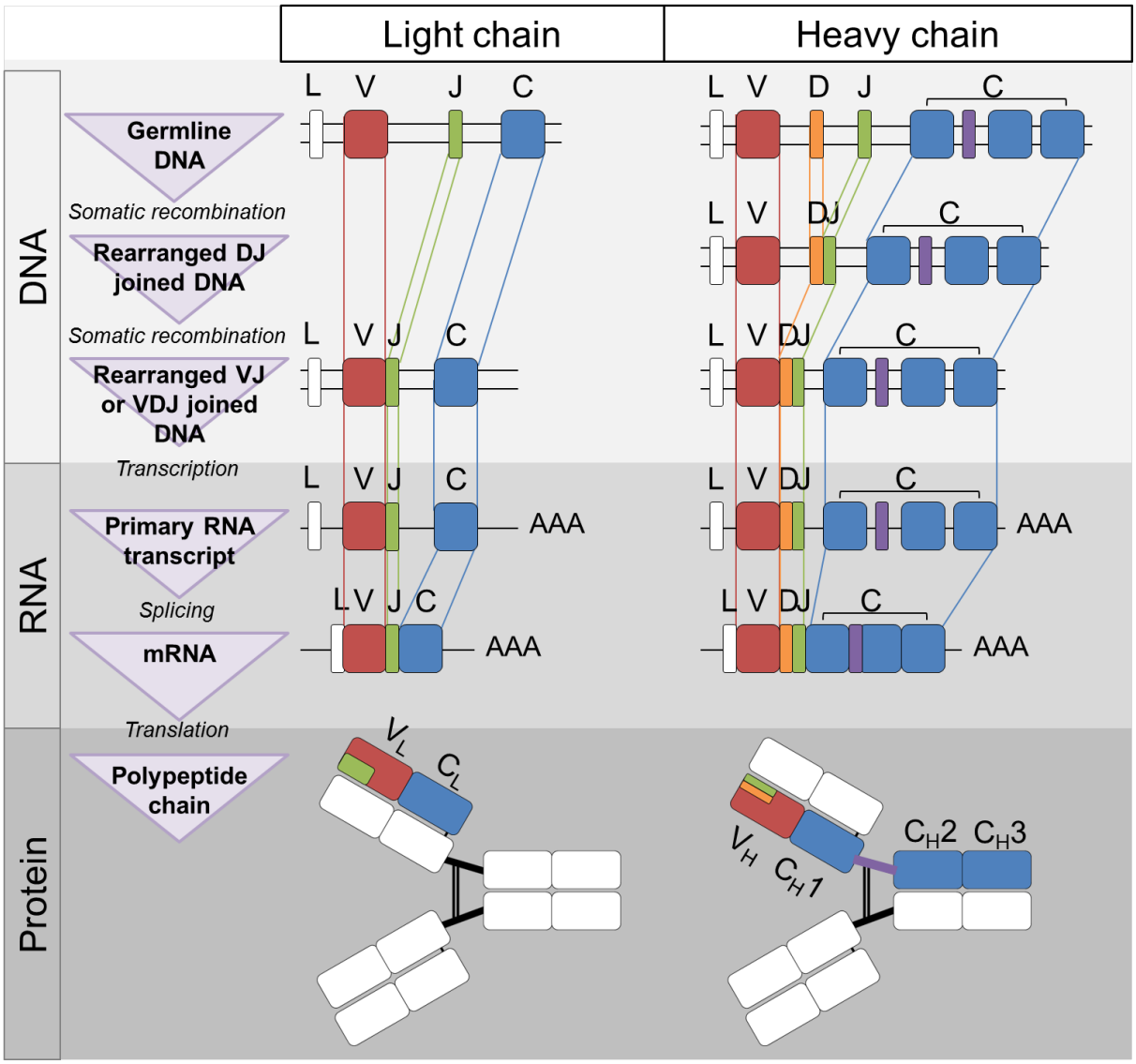

Figure 1.5: Genes of the variable region are arranged by several gene segments. The genes of the light chain consist of two parts: V and J gene segments. When those two parts are joined, an exon for the variable region is build. The leader peptide (L) channels the protein into secretory pathways of the cell. A separate exon encoded the constant region. Splicing of the mRNA of the light chain joins the C with the V region and removes the introns. In the heavy chain first D and J segments are joined, then the DJ with the V segment, which results in a complete VDJ exon. The gene for the C region consists of several exons, which are spliced together with the L peptide and connected to each other. L sequences are removed after translation and disulfide bonds are generated. Hinges are marked in purple.

A successful rearrangement leads to the production of the complete heavy μ-chain [**?**]. After H chain rearrangement is completed and the resulting functional heavy chain is paired with the surrogate light chain and expressed on the cell surface, the cell becomes a pre B cell. Pro B cells that do not produce functional μ-chains are eliminated, which occurs in around 45% of all pro B cells. In two of three cases the first VDJ rearrangement is unproductive, leading to a rearrangement on the second chromosome [**?**]. Further the repertoire of V genes contains pseudogenes, that are sometimes included in rearrangement, leading to unproductive rearrangements [**?**]. Unproductive rearrangements do not lead necessarily to failing pro B cell developments, because other rearrangements can take place on the same or on the other chromosome [**?**].

1.1.3.2 Large pre B cells

V(D)J recombination leads on the one hand to an very diverse repertoire, but on the other hand unproductive rearrangements can occur [**?**]. That is why pro B cells need mechanisms to check functionality of heavy chains. This is tested by inserting the heavy chain into a receptor and adding two invariant substituted proteins, which are similar to the structure of light chains [**?**]. Heavy and surrogate light chains are paired, resulting in a pre B cell receptor [**?** **?**]. This receptor signals a productive rearrangement to the pro B cell [**?** **?**]. The formation of a pre B cell receptor is an important quality control step during the transition of a pro B cell to a pre B cell [**?** **?**]. Defects in this step can drive to immune deficiency [**?**].

Successful rearrangements at both alleles of the heavy chain can lead to a B cell that expresses two different receptors having different antigen specificities [**?**]. To prevent such events, signals of the pre B cell receptor can down-regulate the expression of the recombinase enzyme, reducing the likelihood of allelic inclusion [**?** **?**].

Further the shift from the pro B cell phase to the large pre B cell phase includes many rounds of cell proliferation, leading to an increase in the number of cells with productive rearrangements 30 to 60 fold, before becoming a small resting pre B cell [**?** **?**].

1.1.3.3 Small pre B cell

In the small pre B cell rearrangement of the light chain loci takes place [**?** **?**]. Each of the small pre B cells emerged from the same big pre B cell, can underlie different rearrangements, resulting in different antigen specificities, which also contribute to the huge diversity of B cell receptors [**?**].

The V region of a light chain is encoded by two gene segments, V and J [**?**]. In L chains, V gene segments are localized at amino acid position 95 to 101, followed by an up to 13 amino acid long J (joining) gene segment [**?**]. During development of the V domain the V and J gene segments get connected and build an exon (see Fig. 1.5) [**?** **?**

]. J gene segments lie closer to the C domain than to the V gene segments [**?**]. The C region and the rearranged V region are separated by an intron [**?** **?**]. The V and C regions are joined by splicing of the RNA transcript (see Fig. 1.5) [**?** **?**].

During the rearrangement of the light chain loci, allele exclusion also occurs, but is far less strict than with H chains [**?**]. If a V-J rearrangement does not produce a functional light chain, more rearrangements of the unused V and J segments can take place on the same allele, a process sometimes referred to as leapfrogging [**?**]. After rearranging one chromosome, this process can also occur on the other chromosome or at other loci [**?**]. Typically rearrangement occurs first at the κ locus, and if this fails to generate a productive rearrangement, rearrangement can proceed to the λ locus. This leads to an increased probability of producing productive light chains [**?** **?**].

The genetic loci of heavy and light chains are on different chromosomes [**?** **?**]. Each locus has a different structure. For example the λ locus of the light chain is on chromosome 22 [**?** **?**]. There is a cluster of several V gene segments and groups of four J segments joined to C segments [**?** **?**]. At the κ locus on chromosome 2 exists a group of V genes, followed by a group of J genes and then C genes [**?** **?**]. The organization of the heavy chain locus on chromosome 14 is similar to the κ locus: separate groups of V, D and J genes [**?** **?**]. The only difference is that there is not only one C region, but several C regions belonging to different isotypes [**?** **?**]. B cells first express heavy chains of the isotypes μ and δ, which is a result of alternative RNA splicing [**?**]. Afterwards IgM and IgD are produced [**?**]. The expression of other isotypes appears after DNA rearrangement in a process called class switch [**?**].

Each immunoglobulin locus consists of several V gene segments [**?**]. In reality there is not only one copy per gene segment, but all gene segments exist in at least two copies in the germline DNA. The selection of the gene parts appears randomly [**?**]. In Table 1.1 the number of functional gene segments for H and L chains in the human genome is shown [**?**]. Further the organization of V, D and J genes of a heavy chain on chromosome 14 can be seen in Fig. 1.6.

Table 1.1: Number of functional gene segments for variable regions of heavy and light chains in the human genome [**?**]

Gene segment	**Heavy chain**	**Light chain**	
		κ	λ
V segment	40	40	30
D segment	25	0	0
J segment	6	5	4

There are many more different V segments than D or J segments [**?**]. Due to the huge number of possible combinations the survival pressure of each gene is reduced and the number of mutations to a non-functional gene is increased [**?**]. Such non-functional genes are called pseudogenes [**?**].

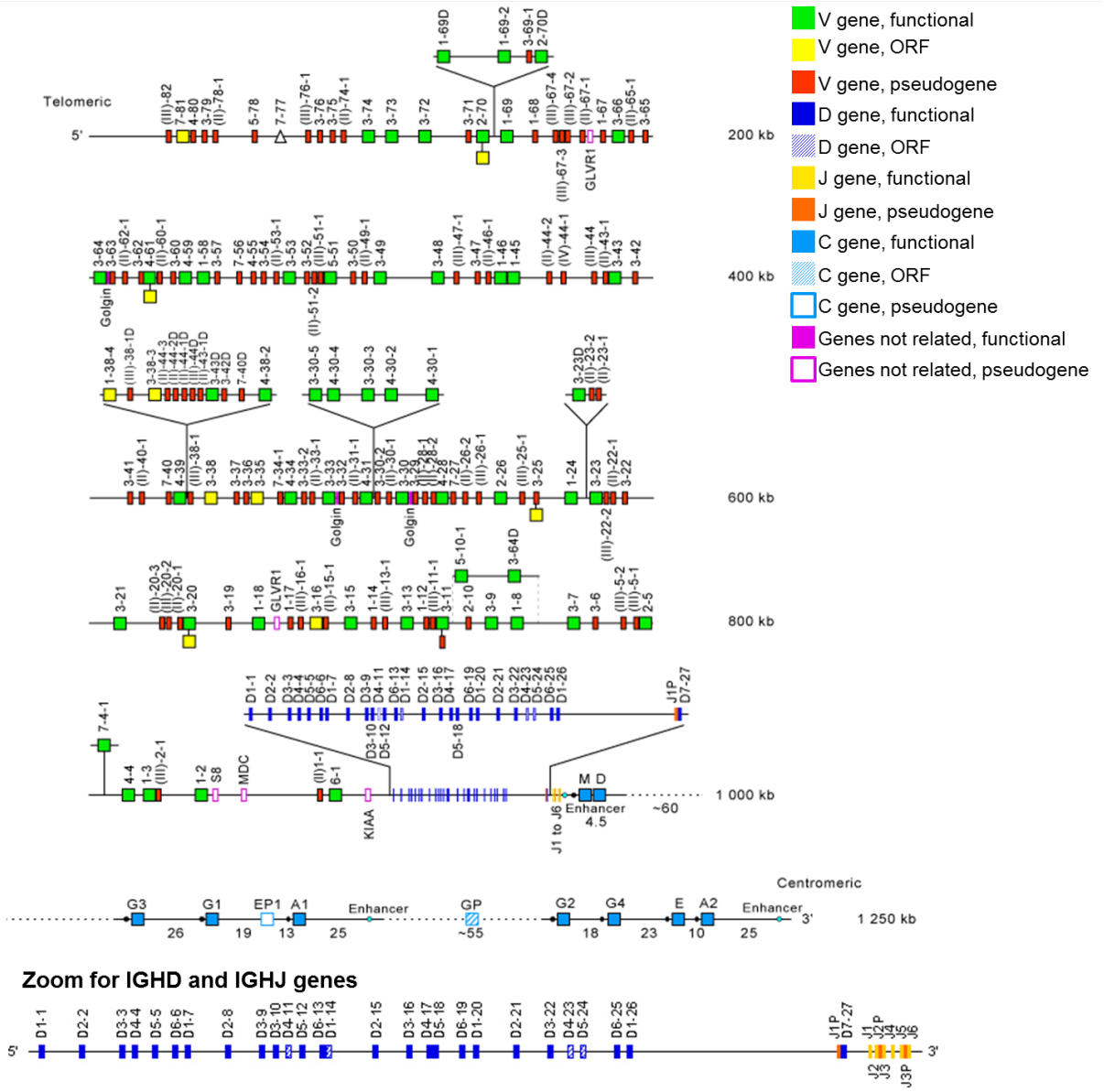

Figure 1.6: Human Ig heavy chain (IGH) locus on chromosome 14. V, D and J genes are color coded. [? ?],[www.imgt.org/IMGTrepertoire/index.php?section=LocusGenes&repertoire=locus&species=human&group=IGH]

1.1.3.4 Immature B cell

Once a rearranged light chain is associated with a heavy chain, IgM can be expressed on the cell surface (surface IgM; sIgM) and the pre B cell becomes an IgM+ immature B cell, which then differentiates into an IgM+ IgD+ transitional B cell [?].

Further in this phase the antigen receptor is tested against tolerance of endogenous antigens first time [?]. This is called central tolerance, because it occurs in the bone marrow, a central lymphatic organ [?]. In the bone marrow the development of the immature B cell is dependent on signals of sIgM [?]. The surface IgM associates with Igα and Igβ and build together a functional B cell receptor complex [?]. Signals of Igα

are very important, because they influence whether B cells can migrate into other tissues and how long they survive [**?**].

Mutations can occur at both levels: DNA and protein [**?**]. Changes in nucleotides can also lead to changes in amino acids [**?**]. Mutations that only appear in nucleotides, but not in amino acids are called silent mutations [**?**]. If also amino acids are changed, they are called non-silent or replacement mutations [**?**].

Immature B cells that do not react strongly to endogenous antigens, can also mature. But once they exhibit autoreactivity, development can be stopped [**?**]. In most cases these cells do not need to be killed (negative selection) [**?** **?**]. Clonal deletion is not the primary mode of autoreactive immature B cell selection, but rather serves as a default mechanism that operates, when receptor editing fails [**?** **?**].

B cells can undergo further antibody gene rearrangement (receptor editing) when they exhibit autoreactivity or if they fail to signal at all [**?** **?**]. These cells still have an affinity for an antigen, but they do not bind with sufficient affinity to signal fully [**?**]. Nevertheless they can be activated when there are high concentrations of autoantigens or inflammatory signal [**?** **?**]. Clonal deletion or apoptosis of cells with special antigen specificity can take place [**?** **?**]. Another alternative would be receptor editing, where the autoreactive receptor is substituted by a non-autoreactive receptor due to gene rearrangements [**?** **?**]. When an immature B cell produces sIgM for the first time, RAG is still expressed [**?**]. If the receptor is not autoreactive, gene rearrangement usually stops [**?**]. B cell development continues and RAG proteins disappear when B cells undergo further development in the spleen [**?**]. On the other hand, if a receptor is autoreactive, a strong crosslinking between sIgM molecules takes place, development is stopped and RAG expression goes on [**?**]. That is why rearrangements of light chains proceed and autoreactive chains can become non-autoreactive, until all V and J segments are exhausted. Cells that are still autoreactive undergo apoptosis [**?** **?**].

There are some studies showing that 50-70% of immature B cells react with self-antigens [**?** **?**]. When about 30-35% of immature B cells undergo receptor editing, the remaining autoreactive immature B cells (15-35%) must undergo clonal deletion, anergy or ignorance [**?**]. But still 30-40% of transitional B cells remain autoreactive after selection in bone marrow; so only few of the autoreactive immature B cells really undergo clonal deletion, but ignorance [**?** **?**].

1.1.3.5 Transitional and mature B cells

Immature B cells migrate from the bone marrow to the spleen and pass through three transitional stages, followed by differentiation into either follicular B cells (also called B-2 cells), marginal zone B cells and a minor population of B-1 cells, depending on signals received through the BCR and other receptors [**?** **?** **?**]. Once differentiated, follicular B cells are now considered mature B cells, or naïve B cells [**?** **?**].

The genes of the C region are placed at the 3'- end of the J gene segment of the heavy chain [?]. Each gene of the constant region can be divided into several exons, which correspond to a special Immunoglobulin domain of the folded protein [?]. The gene, which encodes the μ-C-region is directly located next to the J segment and is the nearest gene after DNA rearrangement [?]. After rearrangement a complete transcript for a heavy μ chain is generated [?]. That is the reason why heavy μ chains are expressed first and IgM is the first isotype that is generated during B cell development [?]. Next to the μ gene the δ gene is located, which encodes the heavy chain of IgD [?]. IgD is expressed together with IgM on almost all matured B cells, but only secreted by plasma cells in low amounts [?]. Cells that express IgM and IgD do not undergo class switch, resulting in irreversible changes of the DNA [? ?]. But they produce a long primary transcript, which can be spliced in different ways and then results in μ or δ chains [?]. On the one hand the VDJ-exon can be combined with the constant region of the μ chain, but on the other hand also with the constant region of the δ chain [?]. Immatured B cells mainly produce μ transcripts; matured B cells mainly produce δ transcripts, in combination with some μ transcripts [?]. After activation of the B cell, coexpression stops due to class switch processes or δ transcripts are excluded from transcription process [? ?].

Activation of naïve B-2 cells occurs in the secondary lymphoid organs (SLO), such as the spleen and lymph nodes [?]. After B-2 cells have matured in the spleen, they circulate through the blood and SLOs, which receive a constant supply of antigen through circulating lymph [?]. Within the SLO, B cell activation begins when the B cell binds to an antigen via its BCR [?]. Once activated, B cells participate in a two-step differentiation process that yields both short-lived plasmablasts for immediate protection and long-lived plasma cells mediating for long-term protection [?]. The first step, known as the extrafollicular response, activated B cells within the B cell zone of follicles immediately migrate into the extrafollicular areas of SLO [?]. During this step activated B cells proliferate, may undergo immunoglobulin class switching, and differentiate into plasmablasts that produce early, low-affinity antibodies mostly of class IgM [? ?]. The second step consists of activated B cells entering a lymphoid follicle and forming a germinal center (GC), which is a specialized microenvironment where B cells undergo extensive proliferation, immunoglobulin class switching, and affinity maturation [? ?]. Isotype switching and somatic hypermutation are both catalyzed by the enzyme activation induced cytidine deaminase (AID) [?]. These processes are facilitated by T follicular helper cells within the GC and generate both high-affinity memory B cells and long-lived plasma cells [? ? ?]. Resultant plasma cells secrete large amounts of antibody and either stay within the SLO or migrate into bone marrow [? ?].

Memory B cell activation begins with the detection and binding of their target antigens. Some memory B cells can be activated without T cell help, such as certain virus-specific memory B cells, but others appear to require T cell help [?]. After activation, the

memory B cell differentiates either into a plasma cell via an extrafollicular response or it enters a GC reaction where plasma cells and more memory B cells are generated [**?** **?**]. It is unclear whether memory B cells undergo further affinity maturation within these secondary GCs [**?** **?**].

A class switch from IgM to one of the other isotypes only appears after stimulation by antigen and is controlled by cytokines [**?** **?**]. The change is based on a non-homologous DNA recombination process, which is controlled by repetitive sequences (switch regions) [**?**]. These regions are located between the J segments of the heavy chain and the constant μ-C-region and on some positions upstream of the genes of the other isotypes [**?**]. After switch from IgM/IgD coexpression to another isotype a DNA recombination event between the VDJ-exon and a special region in front of the corresponding isotype takes place and all of the intervening DNA is excised from the chromosome [**?**].

The five isotypes of Immunoglobulins differ in their constant region of the heavy chains [**?**]. Sequence differences between the heavy chains lead to different characteristics of the isotypes and hence to distinct effector functions [**?**]. One can distinguish between the number and localization of the disulfide bonds, the number of oligosaccharides and the constant domains, as well as the length of the hinge (see Fig. 1.7) [**?**]. Heavy chains of IgM and IgE have no hinge regions, but an additional C region [**?**]. Despite a missing hinge, the chains are still flexible and can change angles [**?**]. The different isotypes differ in their effector function [**?**].

In summary, immunoglobulin diversity is based on several mechanisms. It is not only interesting to study sequence diversity, but also to focus on special V, D or J gene distributions under different conditions. Ig diversity can be a consequence of several mechanisms, including:

1. The gene rearrangements that join two or three segments to an exon of a variable region generate diversity on two ways: first there are several copies of each gene segment and due to different rearrangements different genes can develop. This mechanism is called combinatorial diversity [**?**].

2. Secondly, due to the joining of gene segments nucleotides can be added or deleted. This results in a junctional diversity and mainly influences the diversity of the CDR3 sequence. Palindrome (P) and non-coding (N) nucleotides can be inserted and lead to framework shifts and non-functional sequences [**?**].

3. Lots of different variable regions of the heavy and light chains can be joined together and build different antigen binding sites.

4. Somatic hypermutations can lead to point mutations in the rearranged genes of the V region, which contribute to a selected antigen binding [**?** **?**].

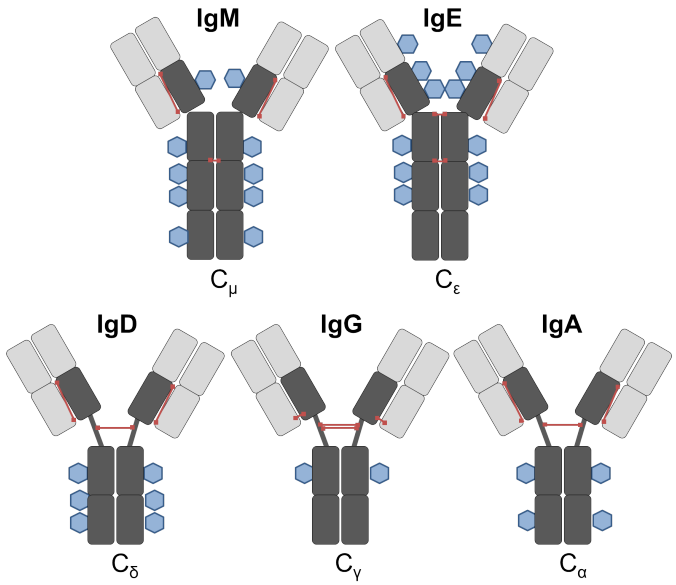

Figure 1.7: Isotypes of Immunoglobulins are encoded by genes of the constant region of the heavy chain. The structure of the secreted forms of different heavy chain isotypes is shown. Note, in addition to its monomeric form presented here, IgA also exists in a dimeric form. Each square represents a domain. Isotypes differ in the number of constant regions (dark gray squares), disulfide bonds (red lines) and carbohydrate side chains (blue hexagons).

1.1.4 The development of autoimmunity

Autoimmune diseases are characterized by activation of the immune system in the absence of an external threat of other organisms [**?**]. Inflammation and tissue damage may occur in the absence of infection, toxin exposure or tumor growth and cause T cell or antibody reactivity to self [**?**]. The innate immune system and inflammatory mediators are activated without any antigen-specific immune response [**?**]. But there are also known cases, which are characterized by the activation of the adaptive immune response, when negative selection is unable to remove all potentially pathogenic B and T cells during affinity maturation [**?** **?**].

Most autoimmune diseases develop spontaneously, while for a few specific genetic modifications targeting a particular pathway in the immune response have been identified [**?**].

There are over 80 defined autoimmune diseases affecting about 5-7% of the population [**?** **?**]. Two critical facts about autoimmune diseases are important in understanding the high frequency of these diseases:

1. Autoreactivity is an aspect of every normal immune system. In fact, the repertoire of immunocompetent lymphocytes that provides protective immunity is selected based on autoreactivity [**?**]. It helps to shape the immune system so that it does not become the pathogenic autoreactivity associated with tissue damage [**?**]. Too little

response leads to potential neglect of danger, whereas an overexuberant response can potentially lead to autoimmunity [**?**].

2. There is a genetic predisposition to autoimmunity and aspects of this may be similar for many different autoimmune diseases [**?**]. They require not only autoreactivity, but are also influenced by target-organ vulnerability [**?**].

The primary event of autoimmune diseases, which triggers the disease process, may be the activation of the innate immune response. It can be activated primary or secondary. Immune complexes containing endogenous Toll-like receptors (TLR) ligands, such as DNA, RNA or citrullinated proteins can activate dendritic cells and other myeloid cells to amplify inflammatory pathways [**?**]. Because tissue injury also leads to activation of innate immune cell networks, thus autoimmune-triggered tissue injury is ongoing, the innate as well as the adaptive immune system in always engaged [**?**].

There is a coordinated interplay among cells of the adaptive immune system with dendritic cells, T cells and B cells interacting to generate the effector response of the immune system [**?**]. DCs and B cells can transport antigen into SLOs. Here, DC are the most important antigen-presenting cell responsible for the activation of T cells during a primary response. Vice versa T cells activate DCs and can help activated B cells to proliferate and undergo class-switch; B cells activate T cells via presentation of peptide antigens and co-stimulatory molecules and cytokines. This leads to an immune response that recognizes a broad spectrum of epitopes and enlists multiple effector mechanisms, relevant to defend against pathogens but also during the pathogenesis of autoimmune and allergic diseases [**?**]. For the immune system to function effectively there must be a sufficient size of both the naïve and memory repertoires of B and T cells that can respond enormous diversity of microbial antigens and a means of regulating those cells that respond to self-antigen [**?**].

A major question in autoimmune disease research is whether the process is autonomous or whether it is driven by antigen, and, if the latter, whether the antigen is a self-antigen or a foreign antigen [**?**]. There are several animal models showing molecular mimicry following activation by microbial antigens, but also that self-antigens drive autoimmune response [**?**]. An excess of apoptotic cells or a problem of their clearance, as well as extensive tissue damage can lead to the presentation of normally sequestered self-antigens in a proinflammatory setting. Alternatively or in addition, post-translational alteration could convert non-immunogenic self-antigens into immunogenic self-antigens [**?**]. Further polymorphisms in autoantigens and environmental exposures may also constitute risk factors for autoimmune disease [**?**].

B cells considerably contribute to most human autoimmune diseases. They are of increased interest due to the therapeutic benefit of B cell depleting therapies in some diseases [**?**]. Anti-CD20 antibodies target CD20, a B cell specific molecule that is first

expressed on the cell surface of all B cell stages starting at the pre B cell stage, but is lost in some plasma cells. B cells contribute to disease pathogenesis not only by production of autoantibodies, but also through their function as cellular adjuvants for $CD4^+$ activation [**?**]. Further they regulate T cell function and inflammation through cytokine production [**?**].

1.1.5 Age and gender dependent development of the B cell receptor repertoire

Only a few studies exist analyzing the age or gender dependent development of B cell receptor repertoires [**?** **?**]. It is already known, that the immune system of the elderly is less able to respond effectively to infectious challenges [**?**]. For instance very young or very old people have less resistance to Streptococcus pneumoniae, with older people being three times more susceptible and having a greater morbidity and risk of mortality [**?**]. Although the efficacy of the human antibody response is decreased with age, the total number of B cells remains constant [**?** **?**]. Only in extreme old age does it decrease; but this quantitative decrease in B cell numbers occurs long after a qualitative decrease (loss of efficacy) [**?**]. Looking at B cell subsets of the innate immune response, studies showed that the number of neutrophils, monocytes, plasmacytoid and myeloid dendritic cells appear to remain stable or moderately decrease with age [**?**]. However, natural killer cells can increase in number with age. However, despite their increased numbers, they exhibit decreased cytotoxicity and produce lower levels of cytokines [**?**].

The quality of the repertoire does deteriorate with age: fewer antigen-specific antibodies with lower antigen-antibody affinity are produced [**?** **?**]. In mice it was also shown that the absolute number of pro and pre B cells decreased with age, thus bone marrow B cell output decreases, along with increased production in myeloid lineages [**?**]. Also a reduced expression of factors that influence B cell differentiation (like E2A or RAG) may be responsible for diminished B cell production [**?**]. Nevertheless, levels of serum and salivary antibodies are quite variable between individuals [**?**].

Decreased IgM and IgD levels in the elderly suggest a shift from the naïve B cell compartment ($CD27^-$) towards memory B cells ($CD27^+$) with increasing age [**?**]. Yet, Dunn-Walters et al. also showed that with age the levels of memory and plasma cells, as defined by CD27 and CD38 surface markers, decrease [**?**]. The IgM memory compartment decreases with age and may be responsible for the decreased reactivity to polysaccharide antigens and increased susceptibility of older people to bacterial infection [**?**]. Further it is known than IgM high/IgD medium $CD27^+$ increases with age, whereas IgM medium/IgD high $CD27^+$ remain the same, which suggests that the repertoire of B cells available to respond to T independent challenges may be comprised in old age [**?**]. In contrast IgD-memory cells (lacking CD27) increase with age, but also during chronic viral infections

or autoimmune conditions [?].

Analyzing SHM, so far no study could indicate that its mechanisms are altered in memory cells during aging in any way [?]. However, it was shown that expanded clones become more numerous with age [?].

Dunn-Walters et al. also showed that vaccine challenges alter the repertoire of all ages [?]. One interesting observation is that the length of the heavy chain CDR3 appears to decrease in antigen-reactive B cells upon boosting [?]. But the selection of repertoires for smaller CDR3 sequences is diminished with age [?]. Further samples from older individuals show greater CDR3 sizes in the naïve repertoire, suggesting that putative negative selection processes are also less effective with age [?]. A CDR3 spectratype analysis has demonstrated a loss of diversity in peripheral blood of some older subjects, which correlated with poor health and survival [? ?].

One of the pitfalls of such studies is that they used blood samples, which are not necessarily the best representation of the lymphocyte repertoire, since cells that are present in blood are usually in transition from one to another tissue [?]. Further, often unsorted sets of B cells are analyzed, which only give overviews about general distributions, but no insights about B cell subset usages.

Men and women differ in several aspects of physiology, including immune responses. Women produce more elevated circulating levels of antibodies than men [?]. Consequently their tendency to produce higher levels of autoantibodies leads to the fact that there are many autoimmune diseases which occur more often in females than males, but men suffer more from infectious diseases, cancer and death than same aged women [? ?]. Men and women experience the same types of aging-related changes to the immune system, but males experience them earlier than women [?].

Immune gender differences have been reported [?], but are relatively under studied in elderly humans. There are only few studies comparing men and women in old age, when hormone differences are less marked. It is known that older women have more circulating B cells than men, whereas men have more CD4 central memory T cells and higher monocyte levels [?]. Also the number, differentiation state and function of innate cells differ dramatically between the sexes. Innate cells of women generally demonstrate a more intense response than those of men [?].

Gender differences in immune cells and immune response can be linked to sex hormone levels, like estrogen, progesterone and testosterone [? ?]. For instance estrogen influences the development and function of B and T cells. Increased estrogen levels lead to IL-17 down-regulation by bone marrow stromal cells [?]. Conversely, estrogen increases B cell production during antigen-specific activation events in response to infections and autoantigens [?].

1.1.6 Next generation sequencing techniques in immunology

As already mentioned, antibody repertoires are very diverse. The theoretical repertoire of TCRs can be estimated at 10^{12} to 10^{20} different receptors that could be generated in humans [**?** **?**]. Due to SHMs the BCR repertoire is probably much more diverse. Therefore, the application of technology with high sensitivity in monitoring specific BCRs is very helpful for immunologists.

Next generation sequencing combines several advantages over current microarray platforms, like lower background noise, better sensitivity, larger dynamic range and greater transcript counts [**?**]. In principle, NGS methods are based on attaching millions of DNA fragments to a surface and simultaneously sequencing all fragments in parallel. Usually, samples are randomly sheared into smaller fragments and used as fragment libraries. Afterwards, specific sequencing primers or adaptor sequences are attached to both sides of the fragments to prepare the fragments for sequencing [**?**]. NGS typically outputs billions of short sequences (25 - 800 bp; called reads), associated with quality scores [**?**].

NGS allows the study of basic phenomena, like clone identification and can provide an estimate of the size of an individual's unique repertoire at any given moment. In clinical practice, this is very important for immunological monitoring of patient's repertoires and monoclonal antibody discovery [**?**].

NGS technology is developing rapidly, but at the moment there are three platforms, which are used in most studies: 454 Roche, Illumina and Ion Torrent [**?** **?** **?**].

- 454 sequencing is based on emulsion PCR and pyrosequencing [**?**]. It can generate longest read lengths (up to 800 bases), compared to Illumina and Ion Torrent [**?** **?**]. The 454 platform can capture a full IgH chain sequence in a single read, which is very helpful when studying patterns of hypermutation in clonally related IgHs [**?**]. However, the throughput is quite small compared to Illumina sequencing and the technique suffers from frame-shift errors [**?**].

- Illumina can analyze greatest number of reads and has the highest throughput per run, but processes only short read lengths (up to 300 bp) and has a relatively high error rate, compared to 454 and Ion Torrent [**?** **?**]. The Illumina platform comprises the HiSeq, MiSeq and NextSeq sequencing systems [**?**]. Currently, the MiSeq technology is able to produce the longest reads using Illumina sequencing-by-synthesis technology. In addition, the rate of homopolymer errors is much lower compared to 454 and IonTorrent sequencing. This makes the MiSeq platform most appropriate to detect open reading frames [**?**].

- Ion Torrent technology is based on standard DNA polymerase sequencing with unmodified nucleotides [**?**]. When a nucleotide is incorporated into neosynthesized DNA strand, a hydrogen ion is released and detected by a hypersensitive ion sen-

sor [**?**]. This method struggles from frame-shift errors and the short read length requires highly multiplexed PCRs, resulting in amplification biases [**?**].

NGS can be used to sequence the highly diverse B cell receptor repertoire. There are many sequence parts, which may help to understand how the repertoire develops and why it might be different in diseased people. For instance, the CDR3 loops of heavy Ig chains can vary in sequence and length, which can influence the ability to recognize diverse antigens. This variation can also cause accumulation of poorly functional or autoreactive immunoglobulins [**?**]. It has already been shown that in mice the average length of IgH CDR3 loops increases during B cell development [**?**]. Therefore the analysis of CDR3 sequences may help to understand better how selection processes balance the benefits of Ig repertoire diversity with the risk of non-functionality and auto-reactivity of highly variable CDR3s.

Further the amino acid residues within CDRs can contribute to antigen binding directly, via contribution of a side group that makes contacts with the antigen [**?**]. Further the amino acids influence the conformation of the peptide backbone indirectly in a manner that facilitates direct interaction of neighboring amino acid side groups [**?**]. Therefore it is important to analyze the amino acid composition along the CDR3 region.

V(D)J segment use and combinations can also be analyzed using NGS. There are particular genes and combinations that are commonly used in healthy individuals. The pairing of V, (D) and J genes does not appear by random [**?**]. When analyzing for instance patients suffering from any disease, there could be a shift in gene usage or even combinations.

1.2 Autoimmune bullous diseases

Autoimmune blistering diseases comprise a large group of potentially devastating diseases which differ in terms of their epidemiological characteristics [**?**]. AIBD is a group of dermatological conditions, which clinically manifest blisters and are caused by the production of pathological autoantibodies against specific components of skin that are involved in adhesion (keratinocytes) [**?** **?**]. In the epidermis, neighboring keratinocytes adhere to each other through organelles known as desmosomes, while dermal-epidermal junction adhesion is mediated by hemidesmosomes [**?**] (see Fig. 1.8). The majority of antigens recognized by AIBD autoantibodies are desmosomal and hemidesmosomal transmembrane glycoproteins involved in epidermal cell-cell and epidermal-dermal adherence [**?**]. Desmosomes contain two parallel intracellular plaques, which are located just beneath the cell membranes of neighboring cells [**?**]. These plaques are composed of plakin family proteins and serve as insertion sites of intracellular keratins, whereas the desmosome core is composed of transmembrane calcium-dependent cell adhesion molecules known as desmosomal cadherins [**?**]. Cadherins include Desmogleins (Dsg 1-4) and desmocollins (Dsc 1-3), which differ in expression throughout the epidermis [**?**].

The hemidesmosomes are located on the dermal pole of the epidermal basal cells and contain an intracellular plaque and a core structure [**?**]. The extracellular space (lamina lucida) separates these cells from the underlying lamina densa, which is composed of collagen IV [**?**]. The hemidesmosomal plaque contains intracellular proteins BP230 and plectin [**?**]. The lamina lucida contains the ectodomains of transmembrane proteins BP180, $\alpha 6\beta 4$ integrin and laminin 5 [**?**] (see Fig. 1.8).

AIBD can be separated into two categories: the skin splits within desmosomes in the epidermis, which is what happens in pemphigus diseases or the skin splits at basement membrane zone (the so called epidermal-dermal junction), as occurs in the pemphigoid diseases [**?** **?**].

There are about 2,000 new cases of AIBD per year in Germany [**?**]. This results in an overall prevalence of 12,000 cases [**?**].

1.2.1 Pemphigus diseases

1.2.1.1 The pemphigus family

Pemphigus diseases are intraepidermal blistering diseases, that effect skin and mucosal membranes [**?**]. They are characterized by the acantholysis caused by circulating autoantibodies directed against Dsg1 and Dsg3, which are components of the keratinocyte desmosomes [**?** **?** **?** **?**] (see Fig. 1.8 and 1.9 d). The binding of these autoantibodies to desmosomes leads to dissolution of these structures and to a loss of keratinocyte adhesion, manifested by the development of blistering, just above the basal layer of the epidermis

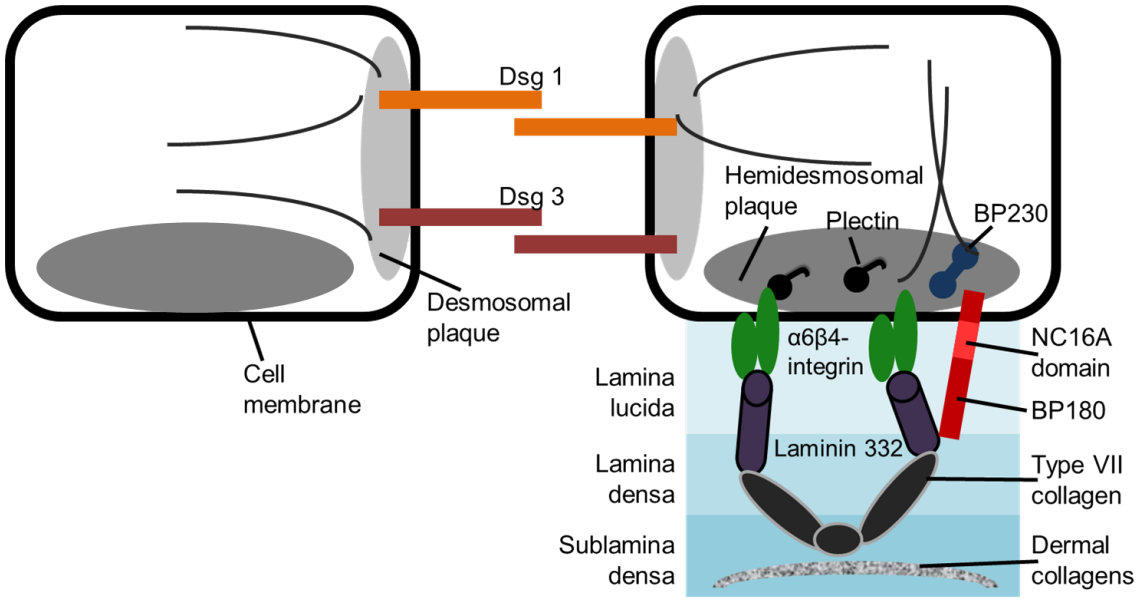

Figure 1.8: A schematic view of desmosomal and hemidesmosomal target antigens in AIBD and the interactions between them. Two neighboring basal keratinocytes are shown schematically. Target antigens of pemphigus diseases are desmosomal structural proteins by means of which neighboring keratinocytes adhere to each other. They include desmosomal plaque proteins and transmembrane proteins of the cadherin group (Dsg 1 and Dsg 3). Hemidesmosomal proteins anchor the epidermis to the dermis and are the target antigens in subepidermal AIBD, in which cleavage occurs between the derma and epidermis. Hemidesmosomal plaque proteins (BP230, plectin) interact with the transmembrane proteins BP180 and $\alpha 6\beta 4$-integrin, which, in turn, are connected by way of laminin 332 to type VII collagen.

[? ? ?] (see Fig. 1.9).

The incidence of pemphigus is globally between one and five million cases per year [?]. The age of onset is variable, but peaks between 60 and 70 years exist [?].

Pemphigus can be subdivided into four families: pemphigus foliaceus (PF), paraneoplastic pemphigus (PNP), IgA pemphigus and pemphigus vulgaris [? ? ?]. PF is caused by Immunoglobulin G (IgG) autoantibodies to Dsg1 only and effects are restricted to skin [? ?]. 20% of western pemphigus cases suffer from pemphigus foliaceus [?]. In contrast, PNP is characterized by not only blisters and erosions, but also rubber-like plaques and pustules [?]. It shares association with certain malignancies and is present with severe blistering of oral and cunjunctival surfaces [?]. The rarest type is the IgA pemphigus, which is only associated with pustule formation, in most cases [?].

Dsg3 is a 130kD glycoprotein and a component of the keratinocyte cell membrane [?]. It is primarily expressed in basal and directly suprabasal layers of the epidermis [?]. Desmoglein 1 is a 160kD glycoprotein, which is less commonly found in PV, but more in PF [?]. It is expressed throughout the epidermis, but is most concentrated within superficial layers [?]. Patients with only Dsg3 autoantibodies typically have only oral

lesions, whereas patients with both Dsg1 and Dsg3 autoantibodies can have both mucosal and skin involvement, like in PV [?]. Patients having only Dsg1 autoantibodies have only skin involvement, for example in PF [?].

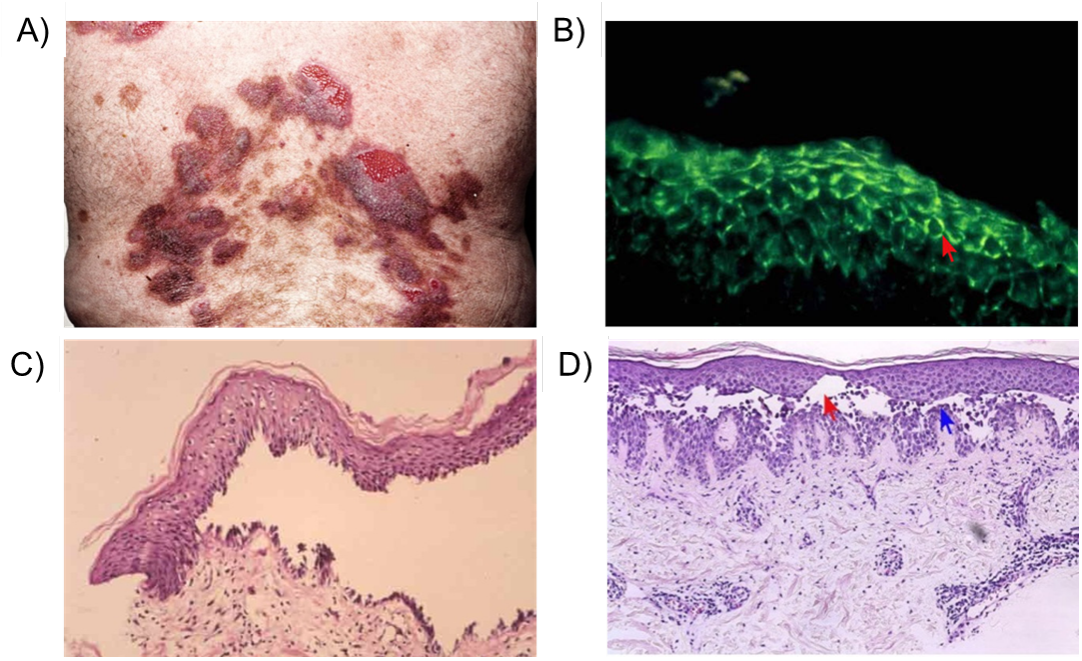

Figure 1.9: Characteristics of pemphigus diseases (in special pemphigus vulgaris). A) Clinical characteristics of pemphigus diseases [?]. PV lesions present with weeping of serous contents, erosions that bleed and crust easily, pain, burning, and tenderness without pruritus. B) The antibodies in pemphigus are against intercellular antigens [?]. Direct immunofluorescence therefore reveals a characteristic intercellular pattern. The red arrow points to an intercellular deposition throughout the epidermis. C) and D) Microscopically, pemphigus vulgaris shows detachment of keratinocytes from each other due to loss of desmosome integrity, causing acantholysis (red arrow) and intraepidermal bullous formation [?]. The blue arrow points to the suprabasal epidermis.

1.2.1.2 Pemphigus vulgaris

PV is the most common subtype of pemphigus: around 70-80% of pemphigus cases in most western countries are comprised of pemphigus vulgaris [? ? ?]. PV can affect patients of any age, but most commonly it arises between 30 and 50 years of age [?]. The incidence is about 1 to 5 cases per million per year [? ?]. A much higher incidence has been found in patients of Jewish ancestry [? ?]. It is more common among women [? ?]. PV is caused by IgG autoantibodies to Dsg1 and Dsg3 [? ?]. The major auto-antigen is Dsg3, but 50-60% of patients also have antibodies against Dsg1 [?]. Clinical manifestations include erosions and blisters of oral mucosa and skin, and may also affect the genital mucosae and eyes [?]. Very little is known about the acute triggers of pemphigus. For example, first degree relatives do not always develop disease, despite sharing genes and

some environmental conditions [?]. However, there is an association between smokers and PV: smoking seems to be beneficial to lower risk of pemphigus vulgaris [?].

PV can be fatal without appropriate treatment due to extensive loss of epidermal barrier function from widespread erosions, potentially leading to infection and metabolic disturbances [?]. The main objectives of treatment are controlling disease, preventing relapse and avoiding adverse events associated with prolonged use of steroids and immunosuppressive agents [?]. Systemic corticosteroids, like Azathioprine, mycophenolate, but also Rituximab, remain the gold standard for clinical therapy [?]. Before widespread availability of systemic corticosteroids, the mortality rate was over 73%, with average survival measuring in only one month after diagnosis [?]. In the 1950s the availability of systemic corticosteroids led to a decrease in mortality rate to 29% [? ?]. In the 1970s it had already decreased to 5% due to better understanding of steroid use, in conjunction with adjuvant immunosuppressant therapy [? ?]. Rituximab is an anti-CD20 monoclonal humanized antibody with the potential to reduce Desmoglein autoantibodies and selectively deplete B cells [?].
It is known, that in the peripheral blood interleukin 10 (IL-10) producing B cells have increased levels after Rituximab treatment in patients with severe pemphigus who achieved long-term complete remission than in those with incomplete remission [?]. Although Rituximab treatment is not available for pemphigus in all countries, B regulatory (Bregs) cells seem to play an important role for long-term remission of patients with pemphigus treated also with intravenous immunoglobulin (IVIG) [?]. The expression of Il-10 defines such a class of B cells with regulatory functions. These cells are dependent on the production of cytokines including IL-10 and transforming growth factor (TGF-β) [?]. Further they may downregulate T cell function [?]. Pan et al. showed, that the number of Bregs is significantly increased in the blood of PV patients with active disease compared with that of patients in remission [?]. B cells from patients were found to secrete less IL-10 and be defective in suppressing interferon gamma (IFN-γ) secretion by T helper cells [?]. Moreover Dsg specific antibody subclasses are considered to be more related to disease activation: PV patients with active disease demonstrate predominantly IgG4, followed by IgG1 autoantibodies to Dsg3 [? ?]. Patients in remission and unaffected relatives can demonstrate anti-Dsg IgG1 [? ?]. In patients with active disease anti-Dsg3 autoantibodies show a significantly higher percentage of total serum IgG4 versus IgG1, indicating selective enrichment of PV autoantibodies in the IgG4 subclass [? ?]. Th1 cells are able to introduce secretion of IgG1 and IgG2, whereas IgG4 and IgE subtypes are mediated by Th2 activation [? ?]. However, IL-10 may also be a double-edged sword of pemphigus vulgaris. It could be delivered to induce Dsg3-specific T cell exhaustion in order to attenuate autoimmune response [?]. But IL-10 could also further stimulate anti-Dsg3 IgG4 production, and this in fact may be the predominant response early on in therapy [?]. Alternatively administration of IL-10 after B cell depletion may promote tolerance

and avoid further stimulation of IgG4 class switch and antibody secretion due to prior depletion of the memory B cell pool [**?**]. This may re-establish the immune regulatory network and prevent disease relapse after B cell depletion, which affects approximately 80% of PV patients treated with Rituximab [**?**].

1.2.2 Pemphigoid diseases

1.2.2.1 The pemphigoid family

The pemphigoid family belongs to subepidermal blistering diseases, where formation of blisters in the epidermal-dermal junction occurs [**?**] (see Fig. 1.10 a, c). In this case blisters have thicker roofs than those of pemphigus diseases and are usually tense [**?**]. There are several subclasses defined, like pemphigoid gestationis, linear IgA dermatosis, mucous membrane pemphigoid or bullous pemphigoid [**?**]. Pemphigoid gestationis manifests itself during pregnancy and is, as well as bullous pemphigoid, associated with severe itch [**?**]. Usually it presents without blisters, but rather with eczematous or urticarial skin changes [**?**]. Linear IgA dermatosis is the most common AIBD in children, where tense blisters are often seen in ring-like arrangements [**?**]. In patients suffering from mucous membrane pemphigoid, mucous membranes near the surface of the body are involved [**?**]. Involvements of eyes can lead to blindness [**?**].

1.2.2.2 Bullous pemphigoid

Bullous pemphigoid is a subepidermal skin blistering disease, which is characterized by the presence of autoantibodies directed against hemidesmosomes of the cutaneous basement membrane zone [**?**]. It is currently the most common AIBD and occurs slightly more common in females and nearly always in the elderly [**?** **?**]. The median age group is about 80 years [**?** **?**]. Bullous pemphigoid is very rare in childhood; about 50 cases have been reported, 15 of which were infants aged less than one year [**?** **?**]. BP is probably one of the only autoimmune diseases in which the incidence increases with age [**?**]. The incidence is between 4.5 and 66 cases per million per year in Europe [**?** **?** **?**]. In the elderly ($>$ 80 years) the incidence is much higher: 150 - 330 new cases per one million per year [**?** **?**]. The one year mortality rate ranges between 15% and 40% and is 2 to 4 fold higher compared to sex and age matched controls [**?** **?**]. The main risk for developing BP is unknown, but in the elderly the increased use of some drugs, like diuretics or psycholeptics, can increase risk [**?**].

Some studies also mentioned patients without blisters [**?**]. This non-bullous form of BP seems to occur in about 20% of patients [**?**]. Leading features, like excoriated, eczematous, popular, urticarial, or even only pigmented lesions associated with moderate to intractable pruritus, persist for weeks or even months [**?**]. In this stage, diagnosis is particularly difficult and challenging [**?**].

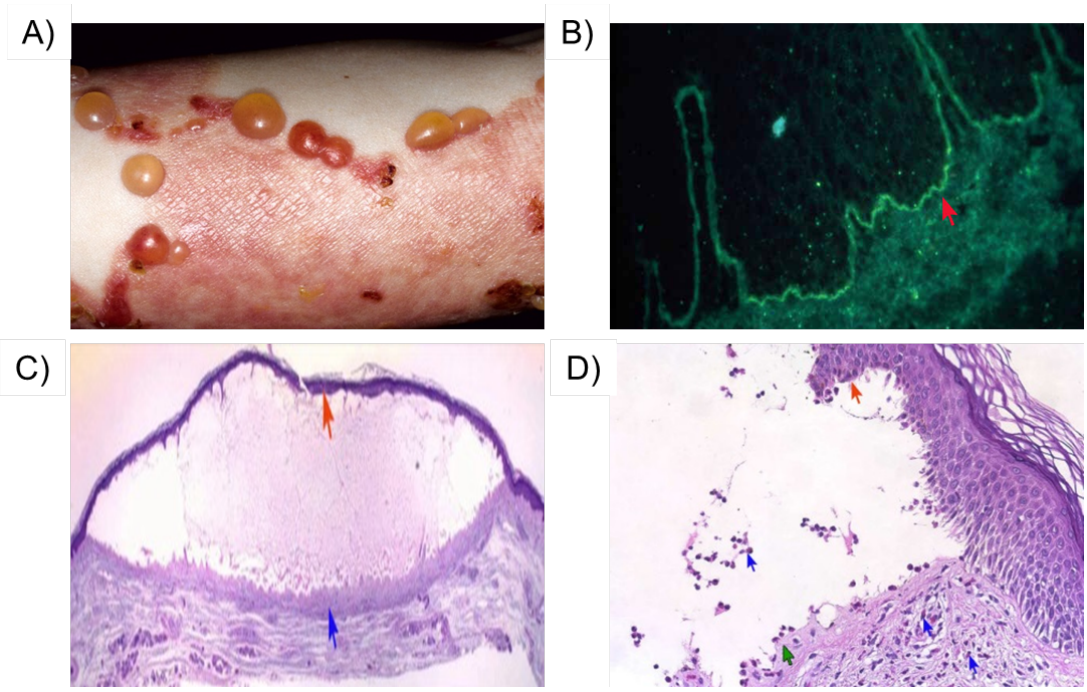

Figure 1.10: Characteristics of Pemphigoid diseases (in special bullous pemphigoid). A) In BP patients, formation of blisters in the epidermal-dermal junction occurs [**?**]. In this case blisters have thicker roofs than those of pemphigus diseases and are usually tense. B) Immunofluorescence for autoantibodies to epidermal basement membrane, forming a bright (flourescent) green line along the epidermal basement membrane [**?**]. The red arrow points to the epidermal basement membrane. C) and D) Microscopically, the lesion shows a subepidermal bulla or vesicle [**?**]. The epidermis thus forms the roof of the blister (red arrow in C) and the papillary dermis forms its floor (blue arrow in C; green arrow in D). The blister and the perivascular infiltrate in the dermis often contain eosinophils (blue arrows in D). The edges of the bullae show degranulating eosinophils (blue arrows in D) close to the basement membrane of the epidermis (red arrow in D).

Patients suffering from BP have circulating IgG autoantibodies against two hemidesmosomal proteins: BP230, which is a cytoplasmic protein, and BP180, which is a transmembrane protein [**?** **?**] (see Fig. 1.8). BP230 and BP180 are also known as BPAG1 and BPAG2. Pathogenic autoantibodies preferentially target the non-collagenous 16A domain of BPAG2 [**?** **?**].

BPAG2 is a type II homotrimeric transmembrane protein with an intracellular N-terminus [**?**]. Each molecule consists of three 180kD collagen alpha-1 (XVII) chains that are characterized by a globular intracellular domain [**?**]. A short transmembrane stretch and extracellular C-terminus domain, composed of 15 collagen repeats separated by 10 non-collagenous (NC) subdomains, is recognized as an immunodominant region by 90% of BP patients [**?**]. The membrane proximal domain NC16A serves as a key part of the flexible collagen-like triple helix [**?**]. It is a critical target for the anti-BPAG2 antibody and contains major epitopes recognized by autoreactive T and B cells in patients

[**?**]. BPAG2 contains multiple binding sites for BPAG1, plectin-1a and integrin-β4 in its cytoplasmic domain and an extracellular binding site for integrin-α6, which help to anchor the protein in the basement membrane, where it further links to collagen VII [**?**] (see Fig. 1.8).

The treatment for BP, as in pemphigus diseases, is based on immunosuppression in combination with elimination of the pathogenic autoantibodies [**?**]. Depending on the severity of the disease, different combinations of plasmapheresis, high dose IVIG, corticosteroids, Rituximab and cyclophosphamide are being used [**?**].

There are several studies describing the important role of neutrophils and T helper cells in the pathogenesis of BP [**?** **?**]. In the majority of patients Th1 and Th2 responses to multiple epitopes in the BPAG2 ectodomain are present [**?** **?**]. A considerable number of patients displayed NC16A-specific peripheral T-cell responses, mainly of the Th2 type or a mixed Th1/Th2 type, which is consistent with the detection of both Th2-regulated IgG4 and Th1-regulated IgG1 autoantibodies in sera of patients [**?**]. In active BP, autoantibodies against NC16A domain are predominantly of the IgG1 class, whereas a dual IgG1 and IgG4 response to this domain is related to more severe skin involvement [**?**]. Further, there are studies showing that T helper and B cell reactivities against BP180 are almost constantly detectable in BP patients and different epitope recognition of BP180 seems to be associated with distinct clinical severity [**?**]. These findings support the concept, that BP180, but not BP230 is the primary autoantigen of BP that is critical for disease development. BP180 is recognized in about 90% of BP patients, whereas BP230 is only found in about 50% of patients [**?**].

2 Aims and Objectives

Little is known about B cell subset specific differences in the antigen-receptor repertoire of patients suffering from PV and BP, compared to controls. Further, only general statements can be given about age and gender specific differences and changes in the repertoire of B cells. The general aim of this study is to investigate:

- Are there any differences in the B cell receptor repertoire between patients suffering from PV and BP and healthy controls?
- Are there already differences in the naïve repertoire between the groups?
- Can we distinguish PV or BP patients from controls considering later stages in B cell development?
- Do we need to consider age and gender specific differences in the BCR repertoire? Does the repertoire change with age and if so, are only later stages of B cell development involved or also the naïve repertoire? Do females and males differ in terms of BCR repertoire and if so, how?
- Can we answer these questions using statistics?

The overarching hypothesis is that patients with PV and BP have B cell tolerance checkpoint defects that will be revealed by B cell subset-specific alterations in the B cell repertoire. While the developmental timing of these defects can vary from one patient to the next, AIBD, in particular PV and BP, patients may have far more B cell abnormalities than control subjects. To test this hypothesis, we decided to study the antibody repertoire more globally using a combination of flow cytometric sorting of B cell subsets and next generation sequencing of IgH gene rearrangements in ten PV, ten BP patients and 20 controls. B cells were sorted into four different subsets as described previously by Sekiguchi et al. [**?**]:

- mature naïve (IgM+, $CD27^{-}$),
- IgM memory cells, that underwent no class switch: IgM+ ($CD27^{+}$),
- IgM memory cells, that underwent class switch: IgM- and ($CD27^{+}$)
- Plasmablasts (PB) ($CD27^{++}$, $CD38^{++}$).

Genomic DNA from flow cytometrically purified B cells was isolated and subjected to two rounds of amplification to generate IgH rearrangements. In PV and controls, IgH rearrangements were amplified using either leader primers or FR1 primers. In BP, only FR1 and JH primers were used. In the second round, IgH amplicons were amplified with primers that contained Illumina adapters. Nextera kits were used for 2x300 bp paired end sequencing on an Illumina MiSeq instrument. In both studies, hundreds of thousands of antibody gene rearrangement sequences were generated in the lab of our collaborator, Dr. Luning Prak, University of Pennsylvania. My task was to contribute to the analysis of these data sets. To analyze such large data sets, sequences were loaded into IMGT/HighV-QUEST database to get information about V(D)J assignment. I wrote an R package, called *bcRep*, to perform automated statistical analyses and visualize the results. To answer the questions named above, I analyzed different issues, considering

- Clone characteristics, like number of clones, size of clones, CDR3 sequence length distribution, etc.
- VH and DH gene usage
- Diversity of clone size and CDR3 sequences
- Mutation pattern on nucleotide and amino acid level.

3 Material and methods

3.1 Patient samples

3.1.1 Ethics statement

All human participants gave written informed consent. Samples and demographic data of patients and controls were collected in adherence to ethics and German privacy protection regulations.

All studies with human materials followed the ethical principles established by the Declaration of Helsinki and were approved by the ethics committee of the University Hospital of Schleswig-Holstein, Campus Lübeck, Lübeck, Germany (No. 07-179).

3.1.2 Patients and controls

Healthy controls from Lübeck, Germany, and PV patients treated in the Department of Dermatology of the University Hospital of Schleswig-Holstein, Campus Lübeck or Campus Kiel, were recruited to participate in this study. The age and gender of patients and controls, as well as concomitant or previous medication were recorded. PV patients who received rituximab previously were not included in this study.

3.1.3 Analysis and separation of B cell subpopulations

Analysis and separation of B cell subpopulations wer done in the Department of Dermatology of the University Hospital of Schleswig-Holstein, Campus Lübeck. Human peripheral mononuclear blood cells (PBMC) were enriched from 45 ml freshly collected K2-EDTA blood by density gradient centrifugation using LymphoPrepTM (Axis-Shield, Dundee, U.K.). Residual erythrocyte contamination was removed by hypotonic lysis in ice-cold 0.2% NaCl solution for 30 sec. PBMCs were stained with Anti-CD-19-PE-Cy7 (clone: IHB19), Anti-CD-14-FITC (clone: M5E2) and Anti-CD-3-CD3 FITC (clone: HIT3a), Anti-CD-27-BV605 (clone: O323), Anti-CD-38-PE (clone: Hb7) and Anti-IgM-BV421 (clone: MHM88, all antibodies by BioLegend, San Diego, CA, USA) for flow-cytometric analysis and sorting on a FACSAriaTM III (Becton Dickinson, San Jose, CA, USA) platform.

For gating of cells, special marker of the cluster of differentiation (CD) can be used. It is a group of monoclonal antibodies, that all detect the same cell surface molecule [**?**]. It can be used as a protocol to identify and investigate cell surface molecules to provide targets for immunophenotyping of cells [**?**]. In this case, cells can be defined based on what molecules are present on their surface [**?**]. Combining several markers allows for cell types with very specific definitions within the immune system. For gating of cells the following strategy was used (Fig. 3.1): Lymphocytes were distinguished by low forward and side scatter and doublets were excluded by pulse width gating. B cells ($CD19^+$, $CD3^-$, $CD14^-$) were separated into four subsets on the basis of CD27 and CD38 expression: mature naive ($CD27^-$, $CD38^+$), plasmablasts ($CD27^{++}$, $CD38^{++}$) and memory B cells ($CD27^+$, $CD38^+$). Memory B cells were further subdivided into a class-switched (IgM-) and a non-class-switched (IgM+) fraction. All cells were collected in DNA stabilizing Cell Lysis Solution (Qiagen, Hilden, Germany). For data analysis on subsets, FlowJo v10 software (Tree Star, Ashland, OR, USA) was used.

3.2 Next generation sequencing and processing of sequence data

3.2.1 IgH sequencing

Immunoglobulin heavy chain sequencing was done by the group of our collaborator, Dr. Luning Prak, University of Pennsylvania. IgH family-specific PCRs were performed on genomic DNA samples from sorted cells. Two different primers were used:

1. Framework region 1 (FR1) primer: The libraries for sequencing of the Illumina MiSeq platform were prepared using a cocktail of VH1, VH2, VH3, VH4, VH5, VH6 from framework region FR1 forward primers, and one J region reverse primer [**?**].

2. VH leader primer: Three multiplexed mixes were employed with forward primers matching VH family leader sequences and reverse primers matching JH gene segment sequences, using a similar strategy to that described in [**?**].

Heavy chain gene rearrangements (VDJ) were amplified with AmpliTaq Gold (Life Technologies, Carlsbad, CA) and 10X buffer at a final concentration of 1.5 mM MgCl2, 0.2 mM dNTPs and 0.5 uM of primer mix. The primers each contained an appropriate adaptor sequence for subsequent Illumina Miseq sequencing. Amplicons were purified by the Agencourt AMPure XP beads system (Beckman Coulter, Inc., Indianapolis, IN). Library quality was evaluated using Bioanalyzer 2100 (Agilent Technologies, Santa Clara, CA)

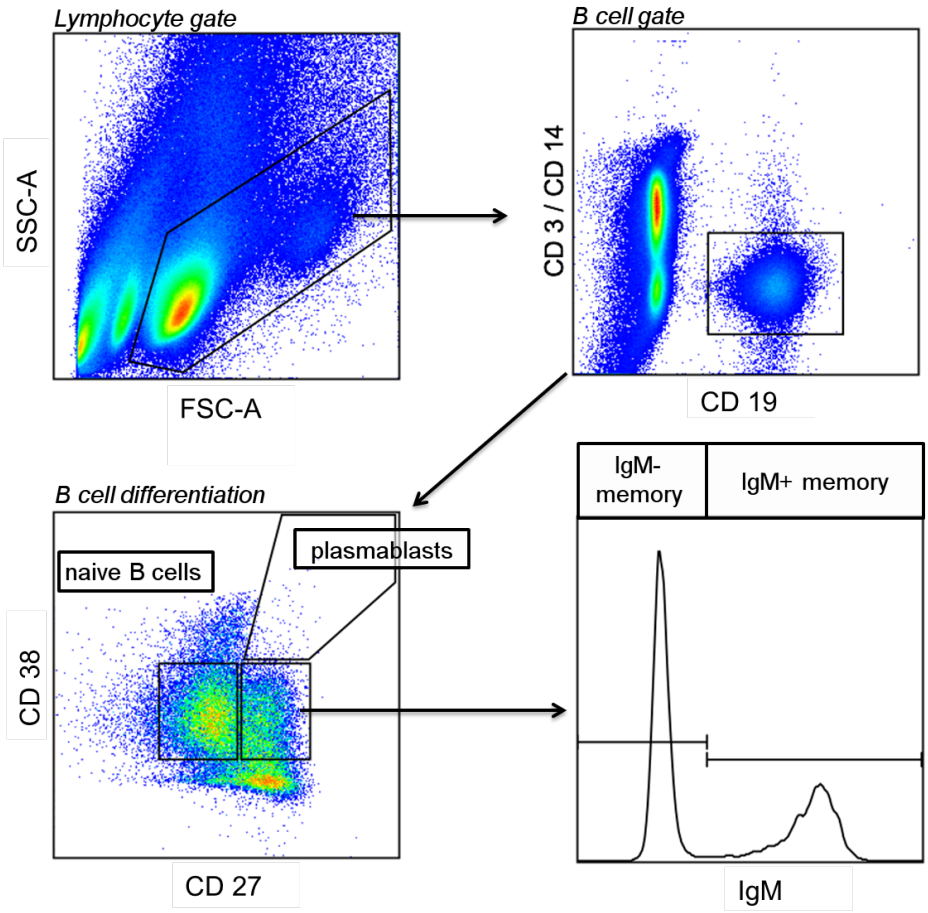

Figure 3.1: Strategy for gating of B cells. Peripheral blood B cells ($CD19^+$, $CD3^-$, $CD14^-$) were separated into four subsets on the basis of CD27 and CD38 expression by fluorescence activated cell sorting: mature naive ($CD27^-$, $CD38^+$), plasmablasts ($CD27^{++}$, $CD38^{++}$) and memory B cells ($CD27^+$, $CD38^+$). Memory B cells were further subdivided into a class-switched (IgM-) and a non-class-switched (IgM+) fraction.

and quantified by Qubit Fluorometric Quantitation (Thermo Fisher Scientific, Grand Island, NY), then loaded onto an Illumina MiSeq using a 2x300 bp paired end kit (Illumina MiSeq Reagent Kit v3, 600 cycle).

3.2.2 Quality control of sequence data and further processing

Quality control of Fastq files, IMGT/HighV-QUEST and statistical analysis of the data were done by myself.

Fastq files were filtered using pRESTO v 0.4 [**?**] and a minimum Phred quality score of 20 was used. Sequences of forward and reverse strands were trimmed, sorted, matched and assembled using Illumina coordinates. Short reads of 10 or less bases were excluded from further analyses.

Fasta files were further processed by IMGT/HighV-QUEST version 3.3.2 and 3.3.5 [**?**] (see 3.3 IMGT/HighV-QUEST) and these outputs were used for further statistical analysis (productive, as well as unproductive sequences).

3.3 IMGT/HighV-QUEST

IMGT/HighV-QUEST is a free available online tool developed by IMGT (the International ImMunoGeneTics Information system, France) to analyze B or T cell receptor sequences of human, mouse or rat in a detailed and accurate way [**?**]. It serves as a bridge between biological and computational spheres in bioinformatics [**?**]. IMGT provides a user friendly interface to upload sequence data and download results. High-quality analyses on V(D)J domains, based on IMGT-Ontology are performed [**?**].

It is possible to upload up to 500,000 Fasta formatted sequences per batch which are then statistically analyzed. IMGT returns 11 output files (txt format) with concepts of identification, description, classification and numerotation (see Tab. 3.1) [**?**]. If less than 150,000 sequences are uploaded, individual result files are provided as well. It identifies V, D and J genes and alleles in rearranged V-J and V-D-J sequences by alignment with the germline IG and TR gene and allele sequences of the IMGT reference directory [**?**]. FR and CDRs are delimited according to the IMGT unique numbering for the V domain [**?**]. Further mutations in V region are described and hot spot positions in the closest germline V gene are identified [**?**]. Insertions and deletions are detected and described by reference to the IMGT unique numbering [**?**]. Most information is provided for both nucleotide (nt) and amino acid (AA) levels.

IMGT/HighV-QUEST result files were used for further statistical analyses.

3.4 Important terms in immunological analyses

3.4.1 Mutations

In IMGT/HighV-QUEST output table 7 (V-REGION-mutation-and-AA-changes-tables) information about silent and replacement mutations in the total V region, FR 1-3 and CDR 1-2 are given. Silent mutations are stated like "g>10a" meaning that there is a mutation at the 10th position in the germline sequence from guanine to adenine. If this mutations also results in an amino acid change (non-silent mutations/replacement mutation), it is marked like "g>10a | V>4I". In this case at codon 4 valine is changed to isoleucine.

Further, there is a table (1 Summary) describing the degree of identity of the V region in percentages, nucleotides, as well as with and without gaps.

Table 3.1: Result files of IMGT/HighV-QUEST analysis [**?**]

	Output file	Description
1	Summary	Provides synthesis of the analysis (sequence functionality; junction frame usage; closest V, D, J genes and alleles; FR and CDR lengths, AA junction; insertions; . . .)
2	IMGT-gapped-nt-sequences	Includes nt sequences of labels that have been gapped according to IMGT unique numbering
3	Nt-sequences	Includes ungapped nt sequences
4	IMGT-gapped-AA-sequences	Includes AA sequences of labels that have been gapped according to IMGT unique numbering
5	AA-sequences	Includes ungapped AA sequences
6	Junction	Includes results of IMGT/JunctionAnalysis
7	V-REGION-mutation-and-AA-changes-tables	Includes list of mutations (nt and AA mutations; AA class identity or change for total V region and FR/CDR sequences
8	V-REGION-nt-mutation-statistics	Includes number of nt positions including IMGT gaps; number of (identical) nt; number of (all) mutations/ silent/ replacement/ transitions/ transversions for V region and FR/CDR sequences
9	V-REGION-AA-mutation-statistics	Includes the number of AA positions including IMGT gaps; number of identical AA; number of (all) AA changes; number of AA changes according to AA-ClassChangeTable/ AAClassSimilarityDegree for V region and FR/CDR sequences
10	V-REGION-mutation-hot-spots	Indicates localization of hot spot motifs detected in closest germline V region with positions in FR and CDR sequences
11	Parameters	Includes date of analysis, IMGT/HighV-QUEST version and parameters used for the analysis

With these investigations one can get information about number of total mutations, silent and replacement mutations for one sequence or as a mean for the total set of sequences/clones. Further so called R/S ratios can be calculated. R/S ratios are ratios of replacement (R) vs. silent (S) mutations. It is known that mutations in particular locations in the FR regions are more likely to be structurally destructive than those in CDR regions [**?** **?**]. However, mutations in the CDRs are more likely to alter the antigen-binding properties. Therefore, it is usually assumed that SHM during antigenic selection tend to result in the accumulation of replacement mutations over silent mutations in the CDRs, whereas the opposite is true for FRs [**?** **?**]. Thus the R/S ratio can be used to get information about antigenic selection within those gene segments (positive selection: R/S ratio » 1) [**?**].

3.4.2 Clone collapsing

Sequences were assembled to clones. Sequences deemed clonally related had to have the same VH and JH gene assignments and the same CDR3 length, including 85% sequence identity, like recommended in [**?**]. Due to possible somatic hypermutation, up to 15% sequence dissimilarity within the CDR3 sequence was accepted. In this case it is unclear whether two similar CDR3 sequences are different because of mutations or due to different germline sequences.

To reduce the consideration of sequencing errors, "clones" with only one sequence appearing once were excluded from further analyses.

3.4.3 Gene usage

IMGT/HighV-QUEST also outputs information about V, D and J genes of each sequence. The nomenclature is always the same: For V and D genes it is "subgroup - gene * allele", for instance IGHV3-30*01, which stands for the third V subgroup of the heavy gene (HV) of an Ig, gene number 30 and allele 1. For J genes the nomenclature is "gene * allele", for example IGHJ1*03, which corresponds to the third allele of the first J gene of the heavy chain (HJ) of an immunoglobulin.

3.4.4 Diversity

3.4.4.1 Gini index

The Gini index measures the inequality of clone size distribution [**?**]. It is bound between zero and one. An index of zero represents a clone set of uniformly distributed clones, all having the same size whereas a Gini index of one would point to a set including only one clone.

3.4.4.2 True diversity

A diversity index is a quantitative measure that reflects how many different types (species) exist in a dataset. Simultaneously it takes into account how evenly the basic entities are distributed among those types. There are several diversity indices, which are simple transformations of the effective number of types, but each index can be interpreted as a measure corresponding to some real phenomenon.

Originally the true diversity arose from ecological analysis, but it is now used in many different fields, including immunology. It depends only on the number of species and an exponent q, and not on the functional form of the index [**?**]. In almost all cases

nonparametric diversity indices are monotonic functions of

$$D^q = (\sum_{i=1}^{n} p_i^q)^{1/(1-q)},$$

or limits of such functions as q approaches unity [**?**]. D is the effective number of types, q the order, p_i the relative abundance of species i and n the total number of species observed [**?**]. This means that when calculating the diversity of a set of sequences, it does not matter whether one uses Simpson concentration, inverse Simpson concentration or Shannon entropy; after conversion all give the same diversity. In Tab. 3.2 conversions of common diversity indices to true diversities are shown [**?**]. Diversities can be transformed in terms of the diversity index itself (x) or the proportions of the species (p_i) [**?**].

The diversity order indicates its sensitivity to common and rare species [**?**]. In the case of immunological research, the term species refers to specific amino acid motifs or to clonally related sequences, for instance. The diversity of order zero ($q = 0$) is completely insensitive to species frequencies and is better known as species richness [**?**]. Orders less than unity give diversities that disproportionately favor rare species, while all values of q greater than unity disproportionately favor the most common ones [**?**]. In the case of $q = 1$, all species are weighted by their frequency without favoring rare or common species [**?**]. Regardless of q it always gives exactly n when applied to a community with n equally-common species.

Table 3.2: Conversion of specific diversity indices to true diversity indices [**?**].

Index x		Diversity in terms of x	Diversity in terms of p_i
Species richness	$x = \sum_{i=1}^{n} p_i^0$	x	$\sum_{i=1}^{n} p_i^0$
Shannon entropy	$x = -\sum_{i=1}^{n} p_i^1 ln(p_i)^1$	$exp(x)$	$exp(-\sum_{i=1}^{n} p_i^1 ln(p_i^1)$
Simpson concentration	$x = \sum_{i=1}^{n} p_i^2$	$1/x$	$1/(\sum_{i=1}^{n} p_i^2)$

3.5 Data analysis

3.5.1 Statistical tests

To test for differences between two groups, Wilcoxon-Mann-Whitney tests were used. In case of more than two groups, the Kruskal-Wallis test was applied, followed by pairwise Wilcoxon-Mann-Whitney tests as post hoc tests. Bonferroni correction was used to correct

p-values for multiple comparison.

When associations between two or more continuous variables, like gene proportions and age were analyzed, generalized linear models were used. Generalized linear models are flexible generalizations of ordinary linear regression and are applied, when the response variables have error distribution models other than a normal distribution.

3.5.2 Constrained Analysis of Principal Coordinates

Ordination methods can be classified in constrained or unconstrained procedures. Unconstrained analyses include for instance principal component analysis or metric dimensional scaling/principal coordinate analysis. In contrast to constrained methods, they are not based on a priori hypothesis testing, but do reduce dimensions on the basis of some general criterion. Constrained Analysis of Principal Coordinates (CAP) uses an a priori hypothesis to relate a matrix of response variables, Y, with some predictor variables, X. It calculates a canonical analysis on principal coordinates based on any symmetric distance matrix, including a test by permutation, as described by Anderson and Willis [**?**].

X may contain the codes of an ANOVA model (a design matrix), yielding a generalized discriminant analysis, or it may contain one or more explanatory (predictor) variables of interest (e.g. environmental variables), yielding a generalized canonical correlation analysis [**?**].

CAP allows not only any distance or dissimilarity measure to be used, but also takes correlation structure among variables in the set of response variables into account. Centering methods and eigenvalue decomposition are applied to calculate the predicted values and the variance they are explaining [**?**].

In this thesis the Bray-Curtis dissimilarity was used as the basis for the constrained analysis. It is used to quantify the dissimilarity of two different sites, based on the counts at each site:

$$BC_{ij} = 1 - \frac{2B_{ij}}{N_i + N_j},$$

where B_{ij} is the sum of the smaller values for those species, which are common between both sites and N_i and N_j are the total number of species count at both sites [**?**]. If abundances at both sites are expressed as proportions, the index reduces to

$$BC_{ij} = 1 - \frac{2B_{ij}}{2} = 1 - B_{ij}.$$

The index is bound between zero and one, where zero means the two sites have the same composition sharing all species, and one means the two sites do not share any species [**?**].

3.5.3 Random forest

The Random Forests algorithm is a machine learning algorithm used to classify large amounts of data with accuracy. It is an ensemble learning method for classification and regression that construct a number of decision trees in the training step and outputting the class that is the mode of the classes output by individual trees [**?**]. Random Forests are a combination of tree predictors where each tree depends on the values of a random vector sampled independently with the same distribution for all trees in the forest [**?**]. Introducing the right kind of randomness makes them accurate classifiers and regressors.

Single decision trees often have high variance or high bias [**?**]. Random Forests attempt to mitigate the problems of high variance and high bias by averaging to find a natural balance between the two extremes [**?**].

The Random Forests algorithm was developed by Leo Breiman and Adele Cutler [**?**]. Random Forests grows many classification trees. Each tree is grown as follows:

1. Choose a training set randomly for the growing tree. Sample a number n of cases at random - but with replacement, from the original data (bootstrapping) [**?** **?**]. Breiman recommends to take 2/3 of the total set as training set [**?**].

2. Let M be the number of input variables. At each node, m predictor variables are selected at random out of M and the best split of these m is used to split the node [**?**]. The value of m is held constant during the forest growing. In other words, instead of choosing the best split among all predictors, m predictors are randomly sampled and the best split among those is chosen. This is a method called baggering [**?**].

3. Each tree is grown to the largest possible extent, without pruning [**?**]. New data are predicted by aggregating the predictions of the n trees [**?**]. After all of the data are run down the tree, proximities are computed for each pair of cases [**?**]. At the end of the run, the proximities are normalized by dividing by the number of trees [**?**].

Error rates can be estimated using the training set. First, predict the data not in the bootstrap sample using the grown tree with the bootstrap sample [**?**]. Then, aggregate these predictions and calculate the so called out-of-bag (OOB) error rate [**?**].

There are some important parameters, which need to be considered for the evaluation of the results:

- The accuracy can be quantitatively assessed by a classification error matrix by comparing relationships between known reference data and the corresponding results of the classification. In short: 1-OOB.

- The kappa coefficient measures the accuracy between classification result and reference data. It takes into account that correct classifications can occur by chance. The higher kappa, the higher is the chance of identifying the factors correctly.
- The importance is estimated by looking at how much prediction error increases when OOB data for that variable is permuted while all others are left unchanged. The higher the level of importance of a factor, the higher is the probability of a correct classification.

4 Results

4.1 R package *bcRep*

Major advances in NGS led to possibilities of deep sequencing of B and T cell receptor repertoires. Existing tools like IMGT/HighV-QUEST (tested version: 3.3.5; [**?**]) process raw IG/TCR NGS data, while extracting V, D and J regions and defining special sequence parts like CDRs or FRs. However, to interpret these sequences and compare them among study groups, further analyses are required. Online tools for B and T cell repertoire analysis are available and include *Change-O* [**?**], *iRAP* [**?**], *IMEX* [**?**], *MiXCR* [**?**] and *VDJtools* [**?**]. Unfortunately, most of them are limited to either the number of input sequences or a limited number of analysis methods. Furthermore, the user is restricted to the output format generated by the program and individual output modifications are usually lacking. Whereas *Change-O* was designed to track somatic hypermutations of BCRs, *iRAP* was developed to characterize repertoire-level dynamics and diversity of B and T cell immune repertoires. *IMEX* analyzes diversity and clones of IMGT/HighV-QUEST data, while *MiXCR* concentrates on processing raw data to quantitative clonotypes. *VDJtools* can use several types of inputs, but also focuses mainly on clonotype data. Tab. 4.1 provides a comparison between *bcRep* [**?**] and other selected IG analysis tools, like *Change-O*, *iRAP* and *IMEX*. *bcRep* comprises many functions in one package, where otherwise several tools are required and it can easily be augmented by existing R functions.

bcRep is a new R package [**?**], which comprises methods to combine and read IMGT/HighV-QUEST output files, and several methods to study not only clones, but also the total set of input sequences or subsets of sequences. Sequences can be filtered for their functionality or junction frame usage, and clones also for their size. Gene usage, as well as (silent and replacement) mutations and diversity can be analyzed. Clonotypes can be classified and compared between different samples. Several dissimilarity and distance measurements are available to analyze relations between gene usage or sequence data of different samples (beta diversity). Samples can not only be analyzed individually, but also compared to each other. Further it has no limitations regarding sequence numbers and is available for Unix, Mac OS X and Windows systems.

bcRep can be used by scientists new to IG repertoire analysis, as well as by advanced users. Functions can be applied without reformatting the input data and most results

can be visualized with implemented plotting routines included in this package. Advanced programmers can use the provided functions as entry for more in depth analyses.

Table 4.1: Comparison of the different B cell receptor repertoire analysis tools and *bcRep*. "+" refers to feature exists, "-" refers to feature does not exist . Information was taken from the documentation of the tools. Abbreviations: R = R package, CL = command line, OT = online tool, GUI = graphical user interface, IMGT = IMGT/HighV-QUEST

feature	*bcRep*	*Change-O*	*iRAP*	*IMEX*
base	R	R, CL	OT	GUI, CL
input	IMGT	IMGT	FASTA	FASTA, IMGT
special function to read input	+	+	-	-
combine several files	+	-	-	+
sequence number limited	-	-	+	-
comparison of samples	+	-	-	+
sequence filtering	+	-	-	-
sequence statistics	+	-	-	+
general mutation statistics	+	+	-	-
advanced mutation statistics	+	+	-	-
lineage trees	-	+	+	-
gene usage	+	-	+	+
gene/gene combinations	+	-	+	-
assemble clonotypes	+	+	+	+
clone filtering	+	-	-	-
clone statistics	+	-	+	+
shared clones	+	-	-	+
clone tracking	-	-	+	-
amino acid distribution	+	+	-	-
diversity	+	+	+	+
dissimilarities/distances on gene usage data	+	-	-	-
dissimilarities/distances on sequence data	+	+	-	-
multidimensional scaling	+	-	-	-
several visualization routines	+	-	+	+
alignment of sequences	-	+	-	-
estimation of repertoire size	-	-	+	-

In the following section I describe data formats used as input and methods implemented in *bcRep*. An overview about all functions can be found in Tab. 4.2. The R package vignette provides a more detailed overview about the functions and their outputs or visualization methods.

Parallel processing is possible for some methods using the *doParallel* package [**?**]. The number of computing cores is set by the user (single core processing by default).

Table 4.2: Functions of the *bcRep* package and their description.

Function	Description
combineIMGT()	Combines several IMGT/HighV-QUEST outputs
readIMGT()	Reads IMGT/HighV-QUEST outputs and filters for sequences without results (optionally; see paragraph "Input data")
sequences.functionality(), sequences.junctionFrame()	Gives information about functionality and junction frame usage of input data
sequences.getAnyFunctionality(), sequences.getProductives(), sequences.getUnproductives()	Filters datasets for productive/unproductive sequences
sequences.getAnyJunctionFrame(), sequences.getInFrames(), sequences.getOutOfFrames()	Filters datasets for in-frame/out-of-frame sequences
sequences.mutation()	Summary statistics about mutations in V-region, FR1-3 or CDR1-2 sequences, like number of all mutations, number of silent/replacement mutations or R/S ratio
sequences.mutation.AA(), plotSequenesMutationAA()	Analyzes all replacement mutations and returns a matrix with proportions of mutations from (germline) amino acid to mutated amino acid + visualization method
sequences.mutation.base(), plotSequencesMutationBase()	Analyzes nucleotide distributions at mutated position and next to silent mutations (positions -3 to +3) + visualization method
clones()	Combines sequences to clonotypes with same V gene and J gene (optional) and a variable CDR3 sequence identity
clones.filterSize(), clones.filterFunctionality(), clones.filterJunctionFrame()	Filters clones for their size, functionality or junction frame usage
clones.CDR3Length(), plotClonesCDR3Length(), plotClonesCopyNumber()	Statistics and visualizations of CDR3 length distribution and copy number of clones

Function	Description
clones.giniIndex()	Gini index of a set of clones
clones.shared(), clones.shared.summary()	Clones shared between at least two samples. Same criteria than in clones()
geneUsage(), plotGeneUsage()	V(D)J gene usage in general or stratified for functionality or junction frame usage (for subgroups, genes or alleles) + visualization method
compare.geneUsage(), plotCompareGeneUsage()	Comparison of gene usage between different samples (for subgroups, genes or alleles) + visualization method
sequences.geneComb(), plotGeneComb()	Gene/gene combinations for V(D)J genes (for subgroups, genes or alleles) + visualization method
aaDistribution(), plotAADistribution()	Amino acid distribution of sequences of the same length + visualization method
compare.aaDistribution(), plotCompareAADistribution()	Comparisons of amino acid distribution of sequences of the same length of different samples + visualization method
trueDiversity(), plotTrueDiversity()	True diversity of sequences of the same length (Richness, Shannon, Simpson) + visualization method
compare.trueDiversity(), plotCompareTrueDiversity()	Comparisons of diversity of sequences of the same length of different samples + visualization method
geneUsage.distance()	Several dissimilarity and distance measurements for gene usage data
sequences.distance()	Several dissimilarity and distance measurements for sequence data
dist.PCoA(), plotDistPCoA()	Multidimensional scaling (principal coordinate analysis) of distances + visualization method

4.1.1 Input data

The input data for *bcRep* are output tables of IMGT/HighV-QUEST. In total, IMGT/HighV-QUEST returns 10 tables (plus a parameter table and in some cases individual files). Tables required as input for the function are described in the corresponding help file. Functions to combine the output from several IMGT/HighV-QUEST output folders and

to read in these tables are provided.

While reading input tables, sequences without any information (marked as "no results" in the "D-GENE and allele" column) can be excluded. IMGT/HighV-QUEST gives no results, if

- the D gene and allele reference directory of the IGH analyzed sequences cannot be managed by the IMGT/GENE database.

- imprecise identification of the 3'V-REGION of the V gene and allele or/and of the 5'J-REGION of the J gene and allele.

- the number of mutations in the V, D and/or J region is higher than a given threshold (set in preferences). [**?**]

4.1.2 Sequence analysis

Functions to analyze features of the sequences from IMGT/HighV-QUEST output are implemented in the package. Information about functionality and junction frame distributions can be retrieved. Furthermore, filtering for subsets of functionality and junction frames is possible. Possibilities to analyze and visualize gene usage, as well as gene-gene combinations on subgroup, gene and allele level are given. For all these functions absolute or relative values can be returned.

4.1.3 Mutation analysis

Basic summary statistics about mutations, like R/S ratios (the ratio of replacement and silent mutations), are provided. IMGT/HighV-QUEST already provides tables containing general information about silent and replacement mutations, but no statistics. Silent mutations can be further analyzed by studying proportions of mutations from one to another nucleotide to find silent mutations that appear more often than others in a given set of sequences. Further methods to investigate nucleotide distributions of the environment of mutated positions. Therefore three positions up- and downstream of the mutated position are considered and ratios of mutation from one nucleotide to another are returned. This helps to get an overview about nucleotides that appear maybe more frequently at positions around the mutations.

Additionally, replacement mutations can be further analyzed. Proportions of mutations resulting in amino acid replacements (reference amino acid according to germline identified by IMGT) are calculated to find substitutions that appear more often than others.

4.1.4 Clone analysis

Clonotypes can be classified using different criteria regarding the CDR3, V and J genes. A threshold for CDR3 sequence identity can be chosen to either allow only identical CDR3 sequences (identity = 100%) or include possible somatic hypermutations (identity < 100%). It is mandatory to have the same V genes criterion. The application to same J genes is optional. The user can select, how strong CDR3 identity should be weighted and if sequences should not only have the same V genes, but also share the same J genes. For instance *iRAP* considers same V, D and J genes and 100% CDR3 amino acid sequence identity. *Change-O* provides several methods to define clones: assigning total Ig sequences into clones, considering same V and J genes and a junction length with a specified substitution distance model or defining clones by specified distance metrics on CDR3 sequences and cutting of hierarchical clustering trees.

A function to look for clones shared between two or more samples is provided, as well. This function uses the same criteria as described above (clones). Additionally, a summary function is implemented. This function returns the number of clones per sample and the number of clones shared between different groups of samples.

Further clone features like copy number, CDR3 length, functionality, junction frames and gene usage can be analyzed and visualized. Filtering methods for clone size, functionality and junction frame usage are provided, as well.

4.1.5 Diversity analysis

Functions for amino acid distributions, as well as diversity measurements are implemented.

True diversity and Gini index are implemented as described in section 3.4.4 Diversity. True diversity (alpha diversity) can be analyzed using order zero (effective number of types (richness) [**?**]), one (Shannon entropy [**?**]) or two (inverse Simpson concentration [**?**]).

Diversity indices are calculated for sequences of the same length. Considering somatic hypermutations, deletions and insertions, it is difficult to assign CDR3 sequences to their native sequence and length. Therefore, diversity indices are calculated for each position. When visualizing the results, figures for each sequence length (x-axis: sequence position, y-axis: diversity index) or one figure including mean diversities and standard deviations (x-axis: sequence length; y-axis: mean diversity index) can be returned.

Further a function calculating the Gini index, which measures the inequality of clone size distribution, is given.

4.1.6 Comparison of different samples

Functions to compare data of different samples are provided. For instance, gene usage, amino acid distribution and diversity can be compared and results visualized across different samples. These functions need an input list containing sequence information from at least two individuals.

Additionally, clone sets of different samples can be compared. This function helps analyzing whether there are so called "public clones" that are shared among several samples or only "private clones" which represent each sample uniquely.

4.1.7 Dissimilarity/distance measurements and multidimensional scaling

For gene usage, as well as for sequence data several dissimilarity and distance functions are provided. With these functions relationships between several samples can be analyzed (beta diversity). Dissimilarity, as well as distance measurements describe numerically how similar two objects are. For example, the Levenshtein distance [**?**], which represents the minimum number of single-character edits between two sequences, would be two for the sequences "AABBCC" and "ABBBBC", because there are two changes (second position A $\rightarrow$ B, fifth position C $\rightarrow$ B). Contrary, the longest common substring algorithm [**?**] returns an index of four (ABBC) for the given example. In the case of distances, higher values describe higher distances/dissimilarities. Small distances are equivalent to many similarities or little dissimilarity.

Studying distances between sequences can be done by either analyzing all input sequences together or analyzing subsets of sequences of the same length. Based on the R package *stringdist* [**?**] dissimilarity or distance indices like Levenshtein, cosine [**?**], q-gram [**?**], Jaccard [**?** **?**], Jaro-Winker [**?**], Damerau-Levenshtein [**?**], Hamming [**?**], optimal string alignment [**?**] and longest common substring can be calculated. The indices are described more detailed in help files of *bcRep* and *stringdist* packages. For instance, Hamming distance only counts character substitutions between two sequences of the same length, whereas the Levenshtein distance also takes deletions and insertions into account. The optimal string alignment also allows for one transposition of adjacent characters, the full Damerau-Levenshtein distance allows for multiple substring edits. The q-gram, cosine, Jaccard and Jaro-Winkler distances underlie more complex algorithms.

For gene usage data a table containing gene proportions of different samples is required as input. When having samples in rows and genes in columns, the distances between the samples, based on the gene usage can be analyzed. Transforming this table will end up in distances between different genes, based on the different samples. Dissimilarity or distance measurements like Bray-Curtis [**?**], Jaccard or cosine are provided using implementations of the R packages *vegan* [**?**] and *proxy* [**?**]. Bray-Curtis is often used for abundance

data, whereas Jaccard distance uses presence/absence data.

Furthermore, these results can be used to perform multidimensional scaling (e.g. principal coordinate analysis (PCoA) or CAP) and to visualize levels of similarity. Ordination methods, like PCoA or CAP can be used to display information contained in a distance matrix.

Availability: The R package *bcRep* is available with all source code, test data and a vignette on the CRAN repository (`https://cran.r-project.org/web/packages/bcRep`).

4.2 The B cell receptor repertoire of PV patients

B cell subsets of ten patients suffering from pemphigus vulgaris (P1-P10) and ten healthy controls (C1-C10) were analyzed. Two experiments were performed, using different sequencing primers: a) FR1 primers and b) VH leader primers, as described in 3.2.1 IgH sequencing. For more uniform comparison between different datasets, only the results of the experiment using FR1 primers are shown. In case of noteworthy results of the VH leader primer experiment, these are mentioned in the text. In most cases both experiments show the same tendencies, but due to more sequences available in experiment b), test statistics might differ between both sets.

Patient and control characteristics can be seen in Tab. 4.3. The diagnosis of PV was confirmed as follows: 1) mucosal-dominant type of PV with oral lesions at the time of blood sampling; 2) detection of an intercellular IgG deposition in the epidermis by direct immunofluorescence on skin specimens; 3) detection of circulating autoantibodies in the serum that bind intercellularly on monkey esophagus by indirect immunofluorescence microscopy; 4) detection of circulating IgG anti-desmoglein 3 autoantibodies by the EUROIMMUN Dsg3 ELISA system. Healthy controls had no history of any severe or relevant chronic disease, nor did they have any symptoms of an acute infection at the time of sampling. PV patients were all receiving some form of immunosuppressive therapy, in contrast to none of the control subjects. While the mean age of patients was 50 ± 19, the mean age of controls was 42 ± 7 years. Two of ten patients and seven of ten controls were females.

After quality control and IMGT/HighV-QUEST analysis, the number of sequences used for further analysis remain as given in Fig. 4.1. Most sequences could be found for PB, followed by IgM-, naïve and IgM+.

Table 4.3: PV patient (P1-10) and healthy control (C1-10) characteristics. Individual ID's, age and sex are shown. All individuals are caucasians, except P7 (indian). F = female, M = male.

PV patients			Healthy controls		
ID	Age	Sex	ID	Age	Sex
P1	69	M	C1	38	F
P2	33	M	C2	39	M
P3	44	F	C3	31	F
P4	NA	M	C4	45	F
P5	51	F	C5	45	M
P6	51	M	C6	45	F
P7	35	M	C7	40	F
P8	69	M	C8	37	F
P9	32	M	C9	55	F
P10	65	M	C10	45	M

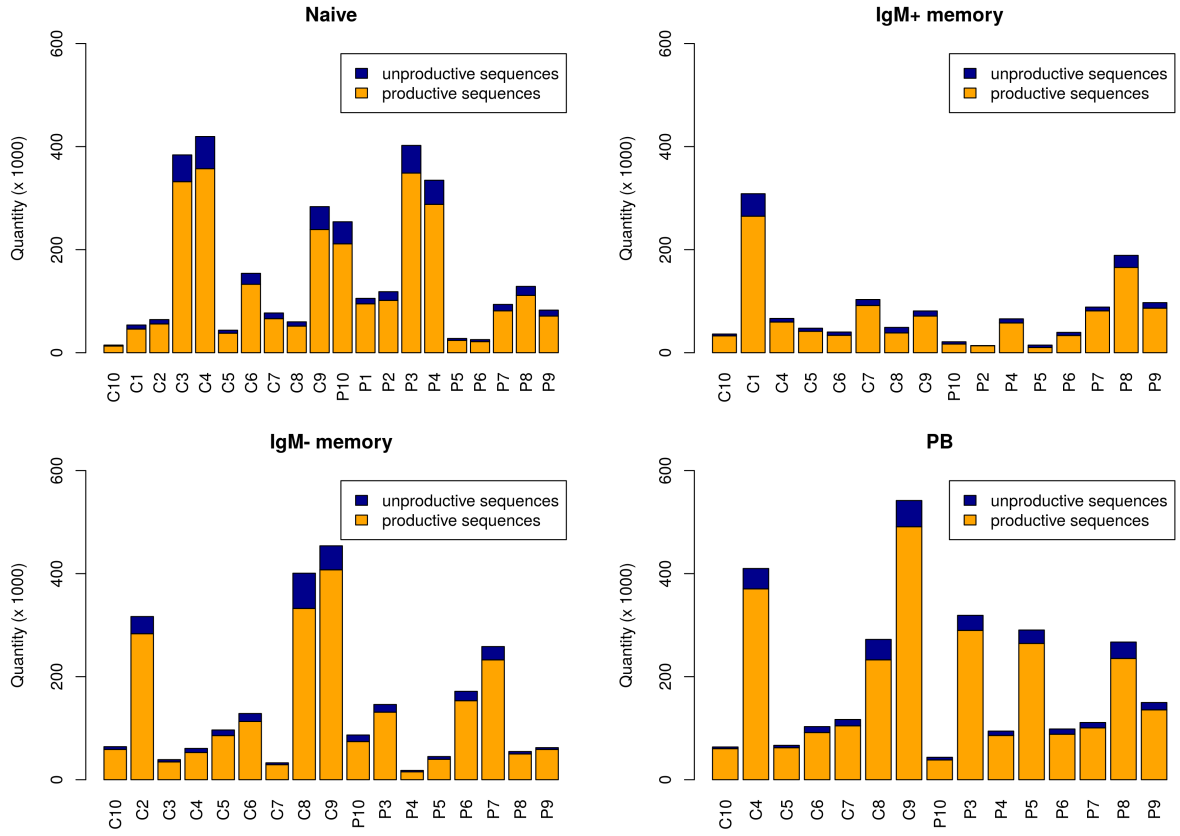

Figure 4.1: Total number of sequences of PV patients and controls resulting from IMGT/HighV-QUEST analysis. Individuals are listed on x-axis (C1-10: controls; P1-10: PV patients), number of sequences on y-axis. The percentage of productive and unproductive sequences in color-coded.

4.2.1 Mutation analysis

Silent and non-silent mutations were analyzed. Studying the average percentage of V gene identity, compared to germline sequence, there are no differences between PV patients and controls (Fig. 4.2). The highest percentages of germline identity can be found for naïve cells, followed by IgM+, IgM- and PB, concluding that most mutations appear in IgM- and PB, consistent with their developmental progression.

Comparing the mean number of mutations per sequence in different parts of the heavy chain (total V, FR1-3, CDR1-2), in most cases PV patients have lower percentages of V gene sequence identity and thus more mutations per sequence than controls (Fig. 4.3, ratio >1). Except for IgM+ memory B cells case/control ratios are very similar (ratio $\approx$ 1) or even below one, indication more mutations per sequence in controls, compared to PV patients.

Further R/S ratios were calculated. There are no significant differences between PV patients and controls, showing that the ratios of replacement and silent mutations are similar in both groups (Supplement Fig. S1). The ratios increase from naïve, to IgM+, IgM- and PB.

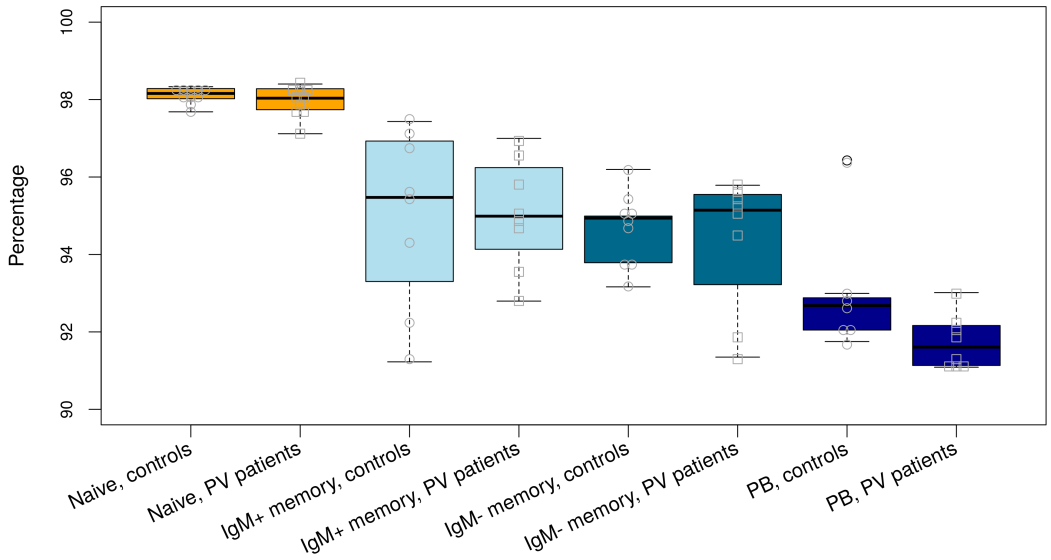

Figure 4.2: Average percentage of V gene identity compared to germline in PV patients in controls. The V gene sequence identity is inversely proportional to the number of mutations. B cell subsets and groups are listed on x-axis; percentage of V gene sequence identity is shown on the y-axis. Highest identity values were found for naïve cells; lowest for PB. PV patients and controls show similar numbers of V gene identity.

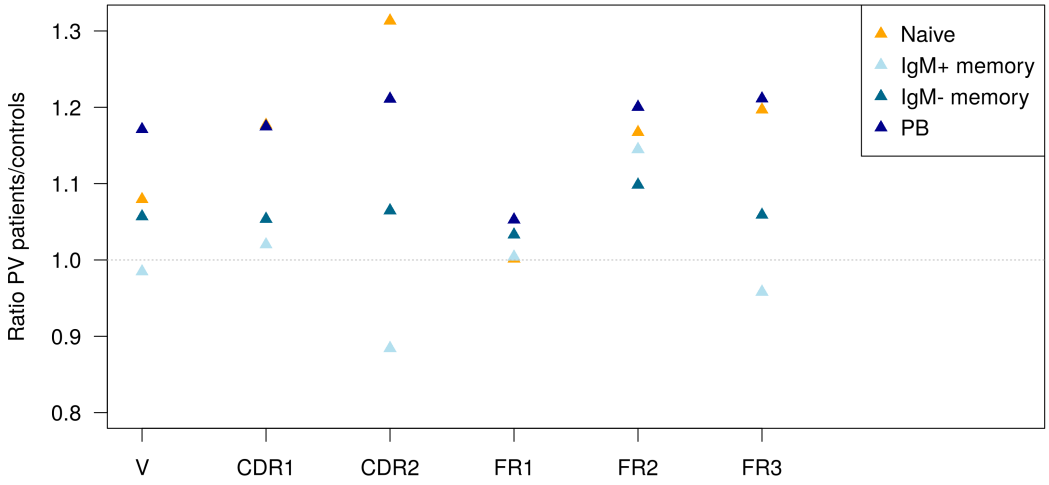

Figure 4.3: Ratio of mean numbers of mutations per sequence in PV patients and controls. For most sequence parts (V, FR1-3, CDR1-2) PV patients have slightly more mutations per sequence than controls (ratio > 1). Except for IgM+ memory cells the ratios are around one or smaller than one, indicating more mutations per sequence in controls, compared to patients.

Analyzing special patterns of nucleotide mutations in PV patients and controls, B cell subset dependent differences can be seen (Fig 4.4). Proportions of mutations from germline to any other nucleotide were calculated and the percentage difference was visu-

alized in a heatmap. Therefore following formula was used:

$$PD_{ij} = \frac{C_{ij} - P_{ij}}{C_{ij}} * 100,$$

were PD_{ij} is the percentage difference matrix with rows i (germline) and columns j (mutation); C_{ij} is the matrix containing proportions of mutations in controls; P_{ij} is the matrix containing proportions of mutations in PV patients.

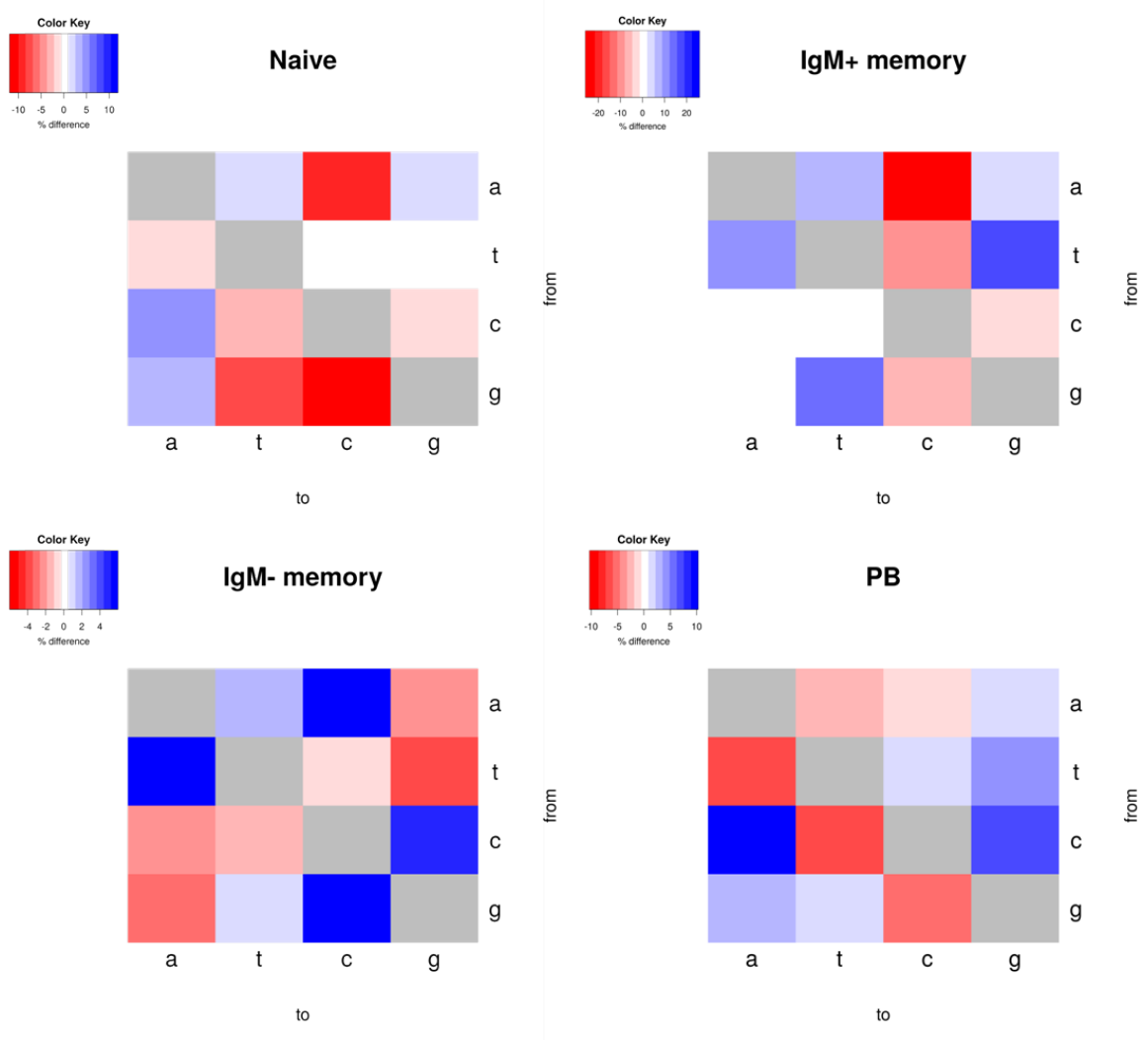

Figure 4.4: Percentages in nucleotide mutation differences between PV patients and controls, from germline (from) to mutated nucleotide (to). Differences are color-coded: red colors show larger proportions in PV patients, compared to controls; blue colors indicate larger proportions in controls. The darker the color, the larger the difference. White fields show no difference; gray ones were not analyzed. Highest differences can be seen for IgM+ cells, lowest ones for IgM-. Naïve, IgM+ memory and PB show similar patterns of mutations, compared to IgM- memory cells.

Most differences can be seen within IgM+, followed by PB, naïve and IgM-. Whereas

adenine (a) to cytosine (c) mutations appear more often in patients than controls in naïve (∼8% difference) and IgM+ memory cells (∼20% difference), they have similar proportions in PB (∼2% difference) and IgM- memory cells (∼4% difference; but slightly higher percentages in controls). Mutations from thymine (t) to guanine (g) appear in IgM+ cells more often in controls (∼10% difference), whereas percentages are similar in the other B cell subsets (0% in naïve, IgM- and PB both ∼3% difference). Further in plasmablasts mutations from thymine to adenine have higher proportions in PV patients, whereas mutations from cytosine to adenine are more abundant in controls (both ∼10%). The same tendencies can be seen in naïve B cells, but conflicting proportions for IgM+ and IgM- memory cells (in much lower amounts).

Next, the same analyses were performed for amino acid changes. Most of the differences were between zero and one percent, indicating a high similarity in replacement mutations in both groups (Supplement Fig. S2).

4.2.2 Clone characteristics

Sequences were clustered into clones, considering same V and J genes and a CDR3 amino acid sequence identity of 85% (and same CDR3 sequence length).

The highest number of clones were found in naïve cells, followed by IgM+, IgM- and PB (decreasing order, see Fig 4.5 A). The maximum number of clones in naïve cells is 30,439. There are two patients having approximately that many clones (P3, P10). For all other subsets, less than 10,000 clones per individual were found. In the FR1 primer experiment no significant differences could be found. Whereas same tendencies exist for the VH leader primer experiment and significant differences in clone number between both groups could be seen in IgM- memory cells ($p=0.035$) and plasmablasts ($p=0.009$), having fewer clones in cases.

B cell subsets having only few clones contained largest clones (Fig 4.5 B). In naïve B cells only small clones (highest number in controls: 456, and in cases: 269) were found. The sizes increase from IgM+ to IgM- memory to PB (max controls: 8,787; max cases: 13,456). Again no significance can be reached in FR1 experiment, but in VH1 leader primer experiment for all B cell subsets, except for naïve B cells, significant differences between both groups can be found ($p<0.001$). In all these cases PV patients have larger clones than controls.

Comparing the number of sequences and the number of clones per individual, only a somewhat linear trend can be seen for naïve cells: with slightly increasing number of sequences, the number of clones also increases (Fig. 4.6). For all other subsets one can see similar number of clones for varying number of sequences. The failure to observe a linear relationship between clone number and unique sequence number in the memory subsets could be due to large clone sizes in these compartments or it could reflect limited

sampling of these subsets.

Further CDR3 amino acid sequence distributions were analyzed. Independently of subsets, CDR3 lengths were similar between cases and controls. Lengths ranged on average between one and 31 amino acids. Some longer sequences (with up to 47 amino acids) were only found in IgM+ memory cells and plasmablasts of PV patients and one control. These were only small clones (with fewer than 15 sequences) and had different V and J genes. In Fig. 4.7 A) the lengths of the longest CDR3 amino acid sequences per individual are shown.

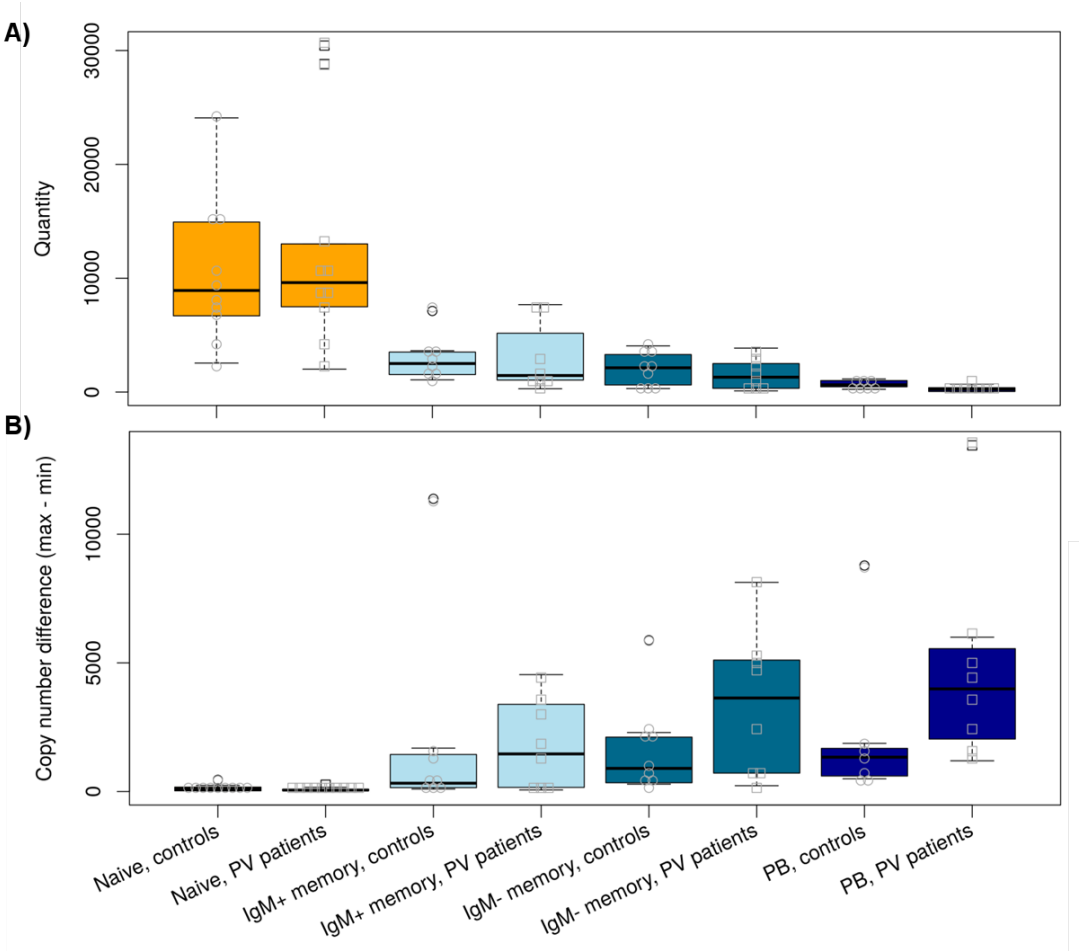

Figure 4.5: Number (A) and size (B) of clones in PV patients and controls. A) In FR1 primer experiment no significant differences could be found. Same tendencies exist for VH leader primer experiment (not shown) and significant differences between both groups could be seen in IgM- memory cells and PB, having fewer clones in cases. B) Generally, clones of PV patients contain more sequences, compared to controls. Again no significance can be reached in FR1 experiment; but in VH1 leader primer experiment (not shown) for all B cell subsets, except for naïve B cells, significant differences between both groups can be found.

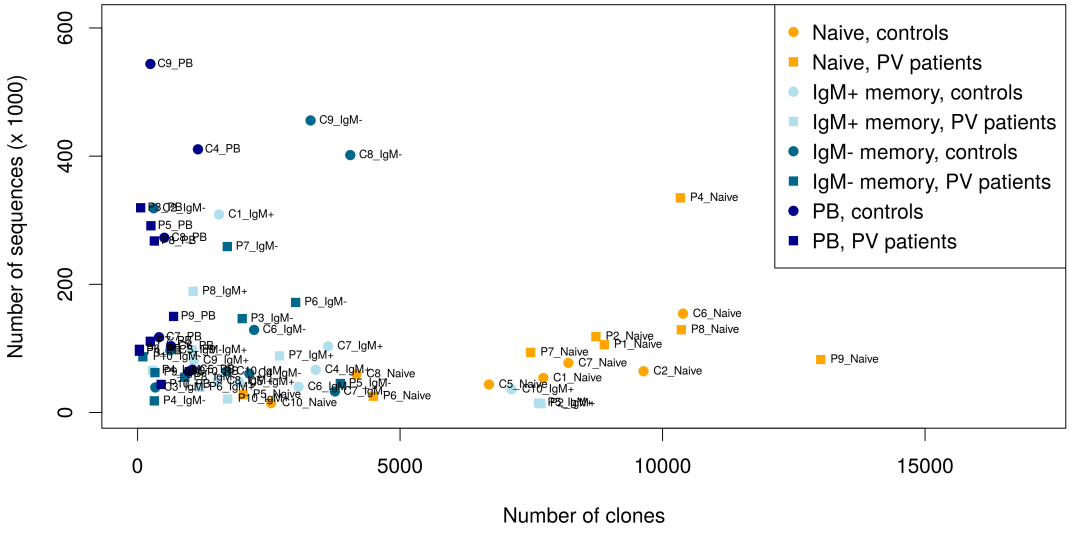

Figure 4.6: Number of clones vs. number of sequences in PV patients and controls. Strong linear dependencies cannot be seen.

Considering length distributions (kernel density), densities between PV patients and controls differed significantly for all B cell subsets ($p<0.001$), except naïve cells (Fig. 4.7 B). Independently of subsets, CDR3 sequences with a length of 12 to 17 amino acids appeared most, having also most of differences between both groups in this interval.

4.2.3 Gene usage in clones

There are many possible combinations of V, D and J genes that can be used to build a heavy chain. It has been proposed that autoantibodies in patients with PV exhibit skewed VH gene usage. Because in principal many different VH genes can be used to mount an immune response, it is hard to understand why such a skewing exists. To find genes, that may influence disease onset and/or progression, I analyzed VH usage in different B cell subsets from PV patients and controls. Note that because molecular calibrators were not used to quantify the absolute frequencies of different VH gene usage, one need to rely on the assumption that the artifactual skewing of VH usage due to PCR amplification biases will be similar across different samples.

To systematically compare global VH usage between different B cell subsets in patients compared to controls, a constrained analysis of principal coordinates was performed (Fig. 4.8). The Bray-Curtis distance (see 3.5.2 Constrained Analysis of Principal Coordinates) was applied on VH and DH gene usage data, containing genes in columns and individuals in rows. Afterwards, a capscale analysis was applied on these distances and the first two axes were plotted (Fig. 4.8). Each dot in Fig. 4.8 represents one individual. For V genes, as well as for D genes, the first two axes (CAP1: $p=0.001$ and CAP2: $p=0.001$), could significantly divide individuals into groups and explain up to 18% of the overall variance. Considering only VH gene usage, naïve cells of both groups are nicely separated from the

other B cell subsets. However gene usage in naïve cells seems to be similar in PV patients and controls. Further there are almost no differences between cases and controls in VH gene usage of both IgM memory subsets. Plasmablast gene distributions in controls are

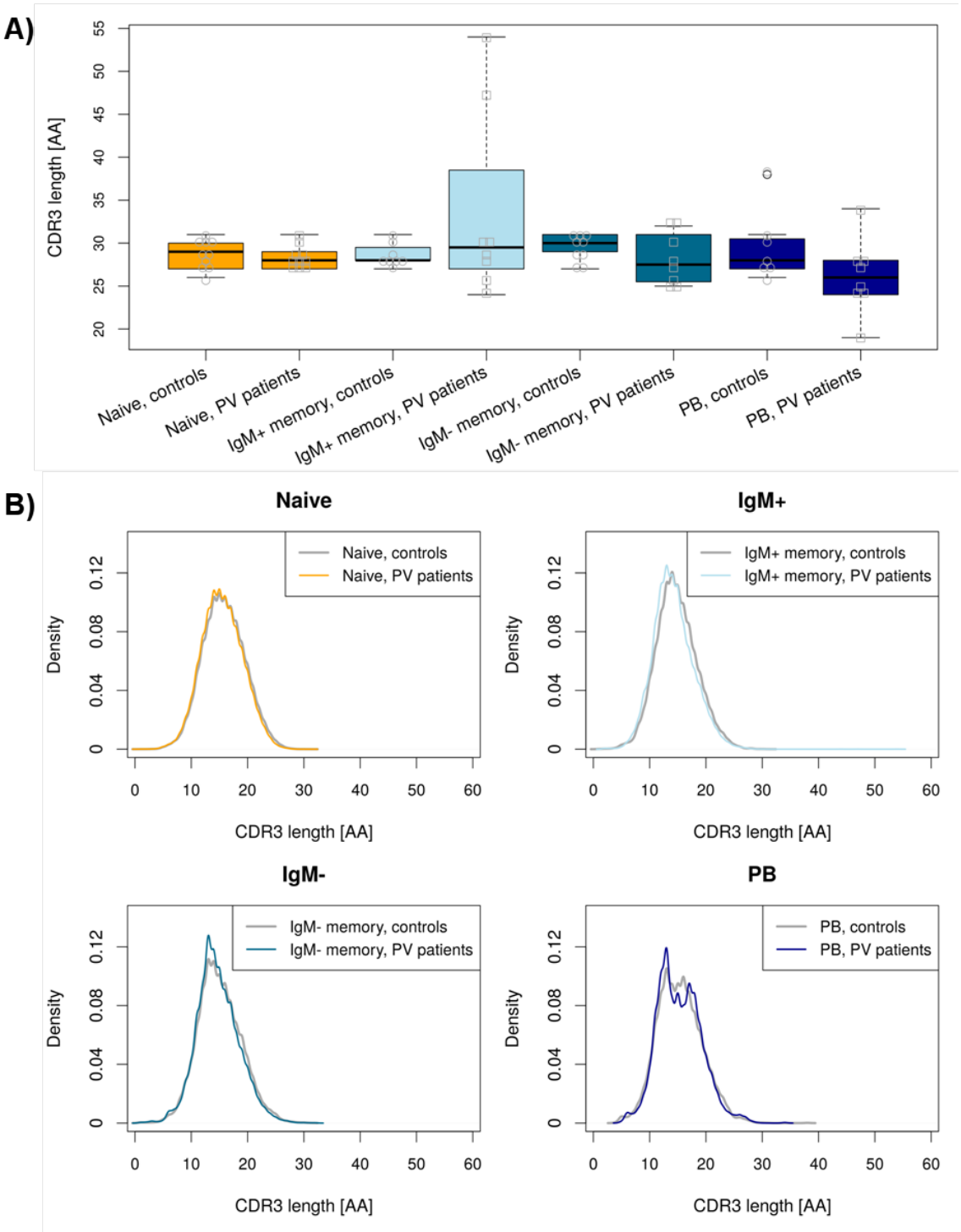

Figure 4.7: CDR3 amino acid sequence length distribution in PV patients and controls. A) Maximal CDR3 sequence length per individual. Only in IgM+ memory cells of PV patients exhibited slightly longer CDR3 sequences. B) Considering length distributions (kernel density, average bandwidth = 0.47) some small differences appear in IgM+ memory, IgM- memory and PB subsets.

similar to IgM memory cells, whereas PBs in PV patients are well separated from all other subsets and show far more variability. These findings suggest that global selection within the PB pool is altered and potentially more relaxed in PV.

Considering DH gene usage, naïve cells can also be distinguished from all other subsets. But there are no strong differences neither between the other subsets or cases and controls. However, the general DH gene usage of plasmablasts of cases is again more variable than in all other groups.

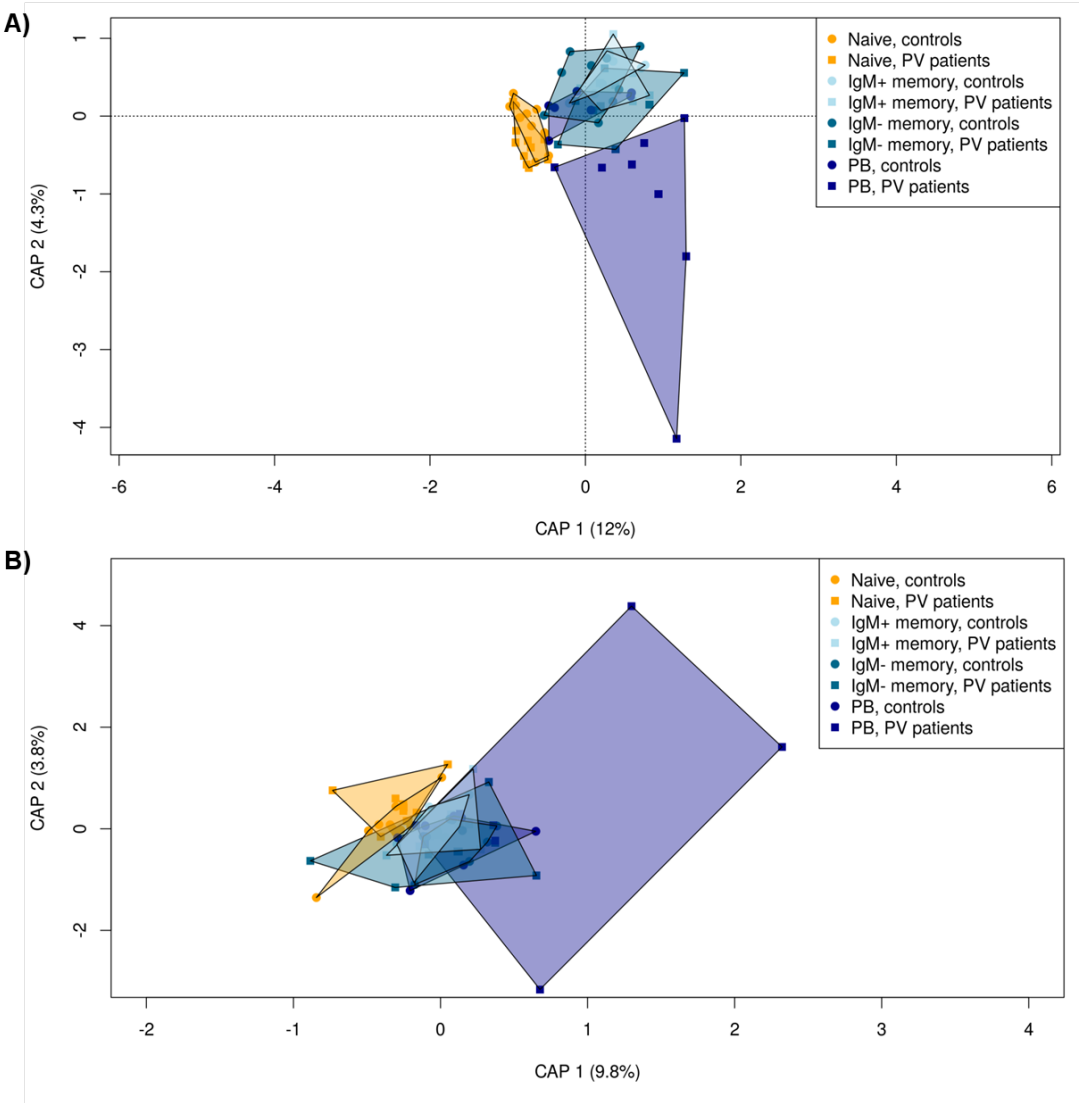

Figure 4.8: Capscale analysis of A) VH and B) DH gene usage in PV patients and controls. For both gene families, the first axis explains around 10-12% of the variance and the second axis acoounts for 4 to 6%. B cell subsets are color-coded. PV patients are shown as squares, controls as dots. Naïve cells can nicely be separated from all other B cell subsets. Gene usage of plasmablasts of PV patients is more diverse than these of other groups.

Further VH and DH gene usage of clones were analyzed. Proportions of genes per individual (Supplement Fig. S3), as well as per group were studied (Fig. 4.9).

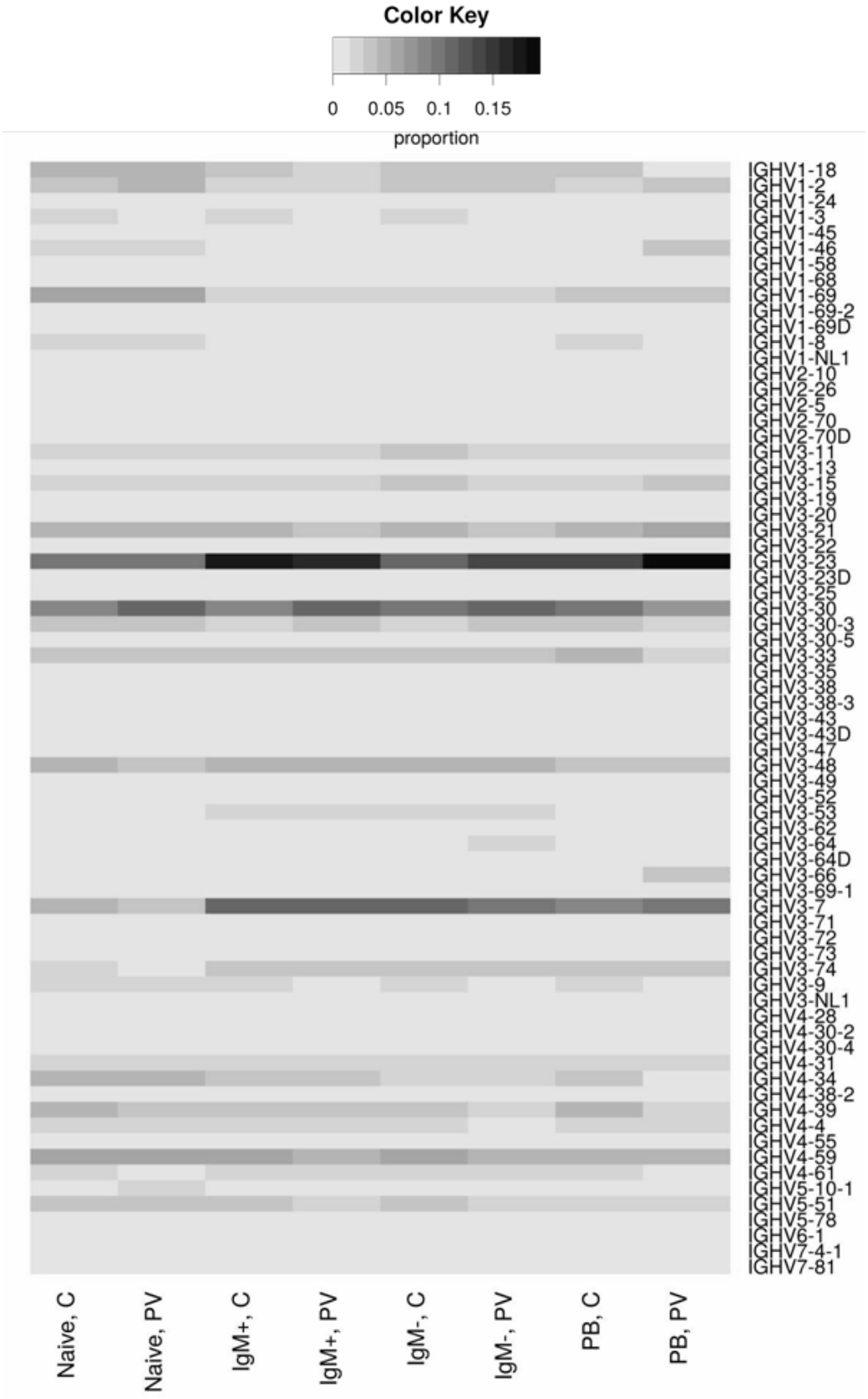

Figure 4.9: Heatmap of VH gene usage in PV patients and controls. Average values of all individuals per group were taken. Proportions of V genes are color-coded. Light colors prefer to small proportions, dark ones to high proportions. Genes that are significantly different expressed ($p<0.05$) between cases and controls, are shown in Tab 4.4

Significant differences were found for IGHV1-18, IGHV1-3, IGHV1-45, IGHV1-46, IGHV1-8, IGHV1-NL1, IGHV2-26, IGHV3-13, IGHV3-47, IGHV3-49, IGHV3-73, IGHV3-74, IGHV4-34, IGHV4-61 ($p<0.05$, see Tab. 4.4).

Table 4.4: Significant differences in VH gene usage between PV patients and controls (Wilcoxon Mann Whitney test, $p<0.05$). B cell subset, VH gene, p value and relative abundance are shown. Most differences in VH gene usage were seen for plasmablasts. Genes belonging to V subgroups 1 and 3 are usually significantly different between both groups. Except IGHV3-13 in IgM+ memory cells, all other genes have higher abundance in controls, compared to patients.

B cell subset	VH gene	p value	abundance
Naïve	IGHV3-74	0.0068	PV patients < controls
	IGHV4-61	0.0039	PV patients < controls
IgM+	IGHV1-18	0.0379	PV patients < controls
	IGHV3-13	0.0281	PV patients > controls
	IGHV3-49	0.0498	PV patients < controls
IgM-	IGHV1-46	0.0479	PV patients < controls
	IGHV3-73	0.0183	PV patients < controls
PB	IGHV1-18	0.0063	PV patients < controls
	IGHV1-3	0.0307	PV patients < controls
	IGHV1-45	0.0213	PV patients < controls
	IGHV1-8	0.028	PV patients < controls
	IGHV1-NL1	0.0213	PV patients < controls
	IGHV2-26	0.043	PV patients < controls
	IGHV3-47	0.0213	PV patients < controls
	IGHV4-34	0.0125	PV patients < controls

In addition to differences in overall usage within individual subsets, some VH genes exhibit altered distributions of VH usage when all subsets are considered. IGHV1-3 appears slightly more often in controls, than in PV patients (except for PB). IGHV3-9 shows same trend, except for naïve B cells. In IgM+ and IgM- memory cells IGHV5-51 is more abundant in controls, than in cases. Whereas IGHV3-30 and IGHV3-30-5 have higher proportions in PV patients, compared to controls, in naïve cells, IgM+ and IgM-. In plasmablasts the ratios of these two genes are vice versa (Fig. 4.9).

To determine if specific V-D combinations are preferentially used in PV patients or controls, I studied the V-D subgroup and gene combinations of each group and generated a matrix containing the differences between PV patients and controls. In Fig. 4.10 differences are shown that are larger or smaller than the mean $\pm$ 2 standard deviations of the difference matrix (difference matrix = V-D combination matrix of PV patients $-$ V-D combination matrix of controls). Of note, there are only gene-gene combinations exceeding this threshold, where proportions in PV patients are higher in controls. For almost all B cell subsets (naïve, IgM+, IgM-) IGHV1 in combination with IGHD3 is found more frequently in PV. Further in naïve cells, IGHV1 together with IGHD5 are

more abundant in cases than controls. There are no differences in plasmablasts exceeding the given threshold, probably due to sampling limitations.

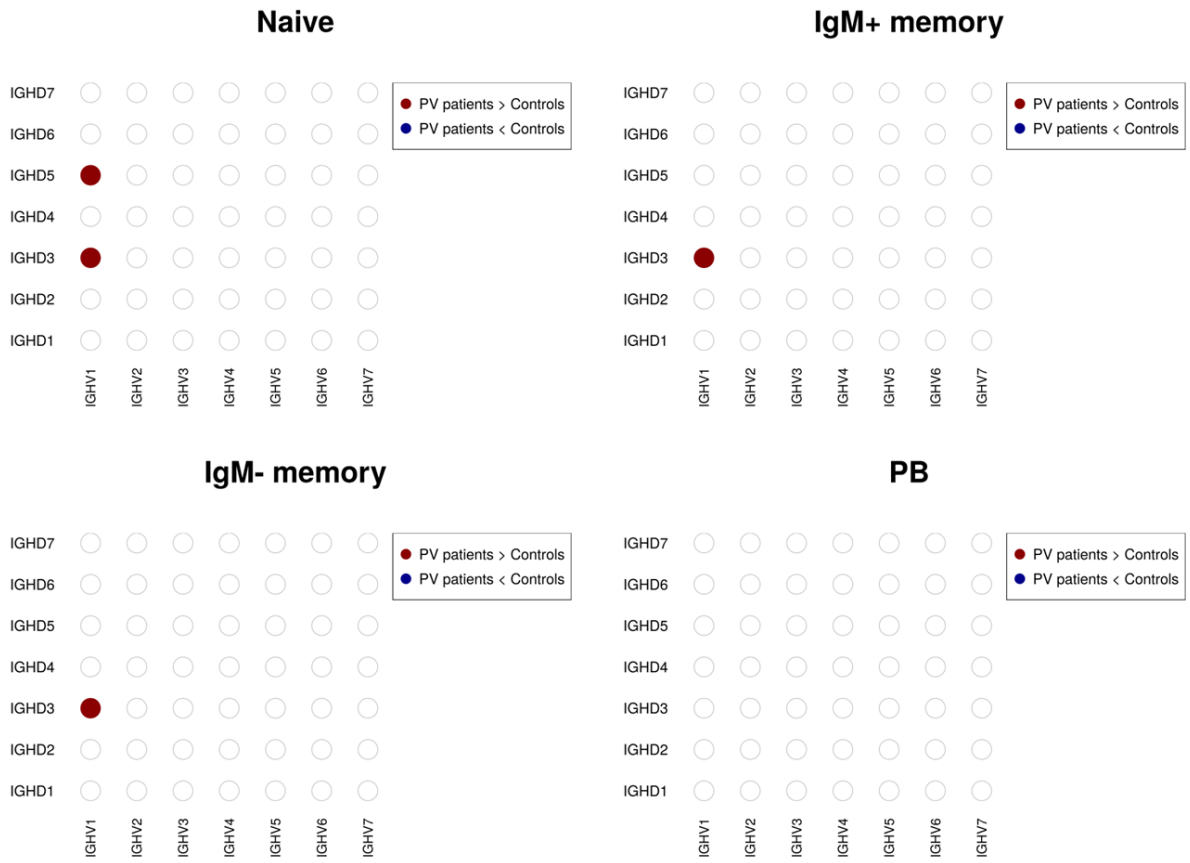

Figure 4.10: V-D subgroup combinations preferentially more abundant in PV patients, than in healthy controls. Combinations that exceed mean ± 2 standard deviations of the difference matrix (cases - controls) are shown. Difference directions are color-coded (red and blue).

Looking at all V-D combinations, regardless of statistical significance (Tab 4.5, Supplement Fig. S4), there are some trends with some of the other gene combinations. Genes of the families IGHV3/5 appear more often together with genes of IGHD1/2/3/6.

In particular, IGHV3-23/IGHV3-23D together with IGHD3-10 is in naïve cells more abundant in PV patients, but in IgM+ memory cells more abundant in controls. Whereas IGHV3-23/IGHV3-23D in combination with IGHD6-19 in both, IgM+ and IgM- memory cells are less abundant in cases compared to controls. Further gene/gene combinations with IGHV1-69 (in naïve cells: IGHD3-22; in PB: IGHD2-2) are more abundant in PV patients, compared to controls. Whereas combinations with IGHV3-7 (in IgM-: IGHD2-2, IGHD2-21, IGHD3-10; in PB: IGHD1-1, IGHD1-14, IGHD2-2, IGHD3-10) show a contrary picture: they appear more often in controls.

Random forest analyses were done to see whether there are predictors that can help to distinguish between PV patients and controls, based on their VH gene usage. Neither

within particular B cell subsets, nor for whole data set, could cases and controls be differentiated with promising parameters. Accuracy, as well as kappa values were less than 0.1 in most analyses.

Table 4.5: V-D gene combinations preferentially more abundant in PV patients, than in healthy controls. Combinations that exceed mean ± 2 standard deviations of the difference matrix (cases - controls) are shown.

B cell subset	IGHV gene	IGHD gene	abundance
Naïve	IGHV1-2	IGHV1-26	PV patients > controls
	IGHV1-2	IGHD3-22	PV patients > controls
	IGHV1-2	IGHD6-19	PV patients > controls
	IGHV1-69	IGHD3-22	PV patients > controls
	IGHV3-23/IGHV3-23D	IGHD3-10	PV patients > controls
	IGHV3-30	IGHD3-10	PV patients > controls
	IGHV3-30-5	IGHD3-10	PV patients > controls
	IGHV4-59	IGHD1-26	PV patients < controls
	IGHV4-59	IGHD3-3	PV patients < controls
IgM+	IGHV3-23/IGHV3-23D	IGHD6-19	PV patients < controls
	IGHV3-23/IGHV3-23D	IGHD3-3	PV patients < controls
	IGHV3-23/IGHV3-23D	IGHD3-10	PV patients < controls
	IGHV3-7	IGHD2-2	PV patients < controls
	IGHV3-7	IGHD2-21	PV patients < controls
	IGHV3-7	IGHD3-10	PV patients < controls
IgM-	IGHV3-23/IGHV3-23D	IGHD6-19	PV patients < controls
PB	IGHV1-69	IGHD2-2	PV patients > controls
	IGHV3-7	IGHD1-1	PV patients < controls
	IGHV3-7	IGHD1-14	PV patients < controls
	IGHV3-7	IGHD2-2	PV patients < controls
	IGHV3-7	IGHD3-10	PV patients < controls

4.2.4 Diversity analyses

We already saw differences between PV patients and controls in clone quantity and size. This could be due to several reasons, like different gene usage, mutations or even differences in CDR3 sequence usage. To further analyze diversity, I used two different approaches. On the one hand the Gini index was measured, giving information about clone size distributions. On the other, hand true diversity indices of order one were calculated for CDR3 sequences of the same length.

Analyzing clone size distributions clearly shows that clones of naïve cells have lowest Gini indices, but show high variances within the group (Fig. 4.11). In naïve B cells of both groups, indices range between 0.27 and 0.65, indicating that clones of these cells are highly variably distributed. Some individuals have clones that are almost equally distributed; some are dominated by a set of larger clones. IgM+ memory cells and PB

have much higher Gini indices, ranging from 0.44 to 0.77 (IgM+) and from 0.57 to 0.79 (PB) and are dominated by large clones. In general PV patients appear to have higher Gini indices than controls. Whereas IgM- memory cells also have high Gini indices (range 0.55 to 0.75), the variability within the group is higher than in IgM+ or PB. Both groups show high variances, but the median in PV patients is higher than the one of controls. Significant differences between PV patients and controls could only be found for naïve cells (p=0.015) in VH leader primer experiment, with controls having higher Gini indices than patients.

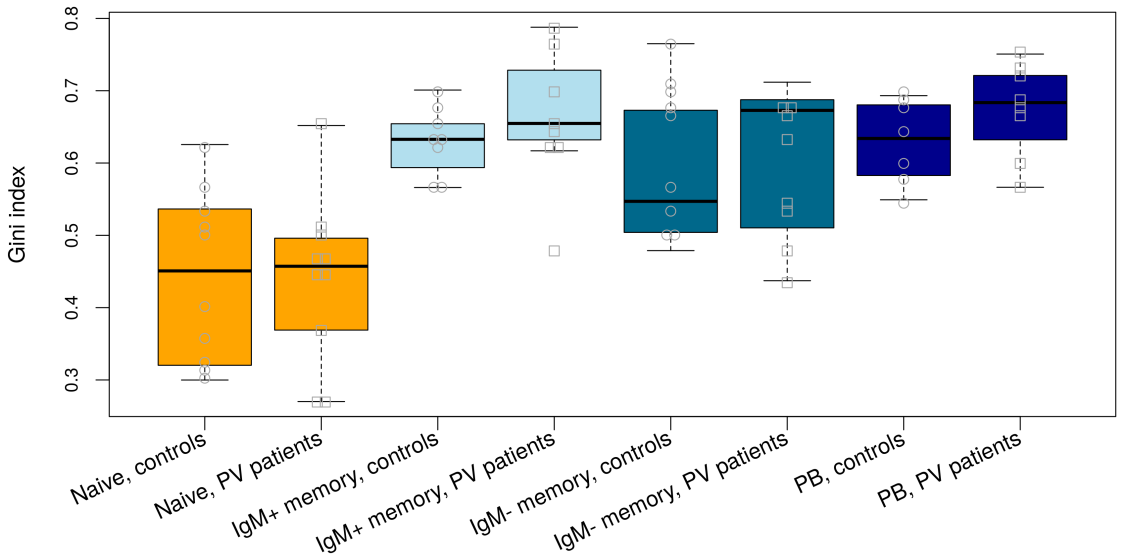

Figure 4.11: Gini index of clones in PV patients and controls. Significant differences (p<0.05) between PV patients and controls could only be found for naïve cells in VH leader primer experiment (not shown), with controls having higher indices than patients and thus being more dominated by large clones.

True diversity can be used to measure variability of amino acids in sequences. The amino acid distribution at each position can be calculated for CDR3 sequences of the same length. In this case, the maximum diversity is at 20 (representing the 20 different amino acids). As shown in Fig. 4.12, diversity indices are smaller for short and long CDR3 sequences, but the same holds true for the beginning and the end of the sequences (not shown). Highest diversity indices can be found for naïve B cells, having similar values for PV patients and controls. For almost all CDR3 amino acid sequence lengths, naïve cells are more diverse than other B cell subsets. For IgM+ (n.s.), IgM- memory cells (n.s.) and PB (p=0.015) controls have more diverse CDR3 sequences than PV patients, in general. In more detail, the order of decreasing diversity is like naïve cases $\approx$ naïve controls > IgM+ controls > IgM- controls $\approx$ IgM+ cases > IgM- cases $\approx$ PB controls > PB cases.

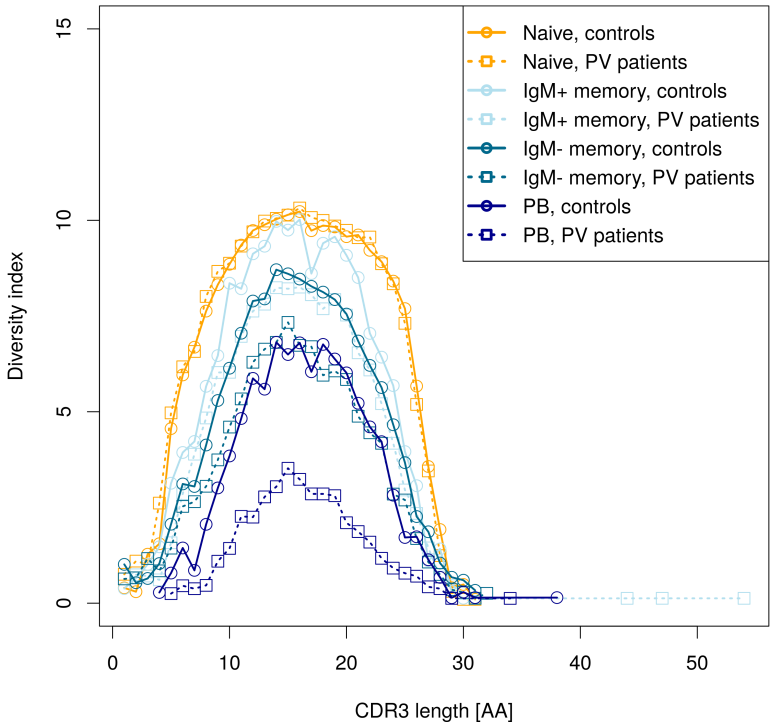

Figure 4.12: Diversity indices for CDR3 sequences of the same length in PV patients and controls. The x-axis represents CDR3 amino acid sequence lengths; the y-axis shows true diversity indices of order one. Patients are represented by squares, controls by circles. Naïve cells of controls and PV patients are more diverse than the other B cell subsets, having almost identical indices in both groups. For IgM+, IgM- memory cells (in general) and PB ($p<0.05$) controls have more diverse CDR3 sequences, than PV patients.

4.3 The B cell receptor repertoire of BP patients

B cell subsets of ten patients suffering from bullous pemphigoid (B1-B10) and ten healthy controls (C11-C20) were analyzed. For this experiment only FR1 primers were used for sequencing. Patient and control characteristics can be seen in Tab. 4.6. All BP patients met the following criteria (inclusion criteria): 1) clinical presentation with prurigo-type lesions, eczema or tense blisters on inflamed skin, 2) IgG deposition at the dermal-epidermal junction by direct and indirect immunofluorescence and 3) microscopy and detection of circulating BP180/NC16A autoantibodies. For the controls, patients with non-inflammatory and non-autoimmune diagnoses were enrolled, mostly basal cell carcinoma or squamous cell carcinoma of the skin. All controls were age- and sex-matched to the BP patients. Mean age of patients was similar to the controls (patients: 79 ± 8 years, controls: 80 ± 7 years). Five of ten individuals in both groups were females.

Table 4.6: BP patient (B1-11) and healthy control (C11-20) characteristics. Individual ID's, age and sex are shown. All individuals are caucasian. F = female, M = male.

BP patients			Healthy controls		
ID	Age	Sex	ID	Age	Sex
B1	61	F	C11	64	F
B2	79	F	C12	87	F
B4	74	F	C13	86	M
B5	89	F	C14	74	F
B6	84	M	C15	76	M
B7	77	M	C16	82	M
B8	77	F	C17	87	M
B9	84	M	C18	82	F
B10	83	M	C19	84	M
B11	82	M	C20	75	F

After quality control and IMGT/HighV-QUEST analysis, the number of sequences used for further analysis are shown in Fig. 4.13. The mean number of sequences was similar in all subsets but tended to be higher in BP patients, compared to controls.

4.3.1 Mutation analysis

As with the PV analysis, silent and replacement mutations in BP patients and controls were analyzed first (Fig. 4.14). Again V gene sequence identity was highest for naïve cells. The number of mutations per V gene sequence increased from IgM+ to PB to IgM- memory cells. For each B cell subset the variance within the groups was always higher in BP patients, compared to controls, whereas the median values of both groups were similar for each subset. For IgM+ memory cells, there were two extreme outliers for both groups, having a mean V gene sequence identity of less than 80%. This suggests that in

BP patients not only more mutations occur, but patients are also more variable within the group.

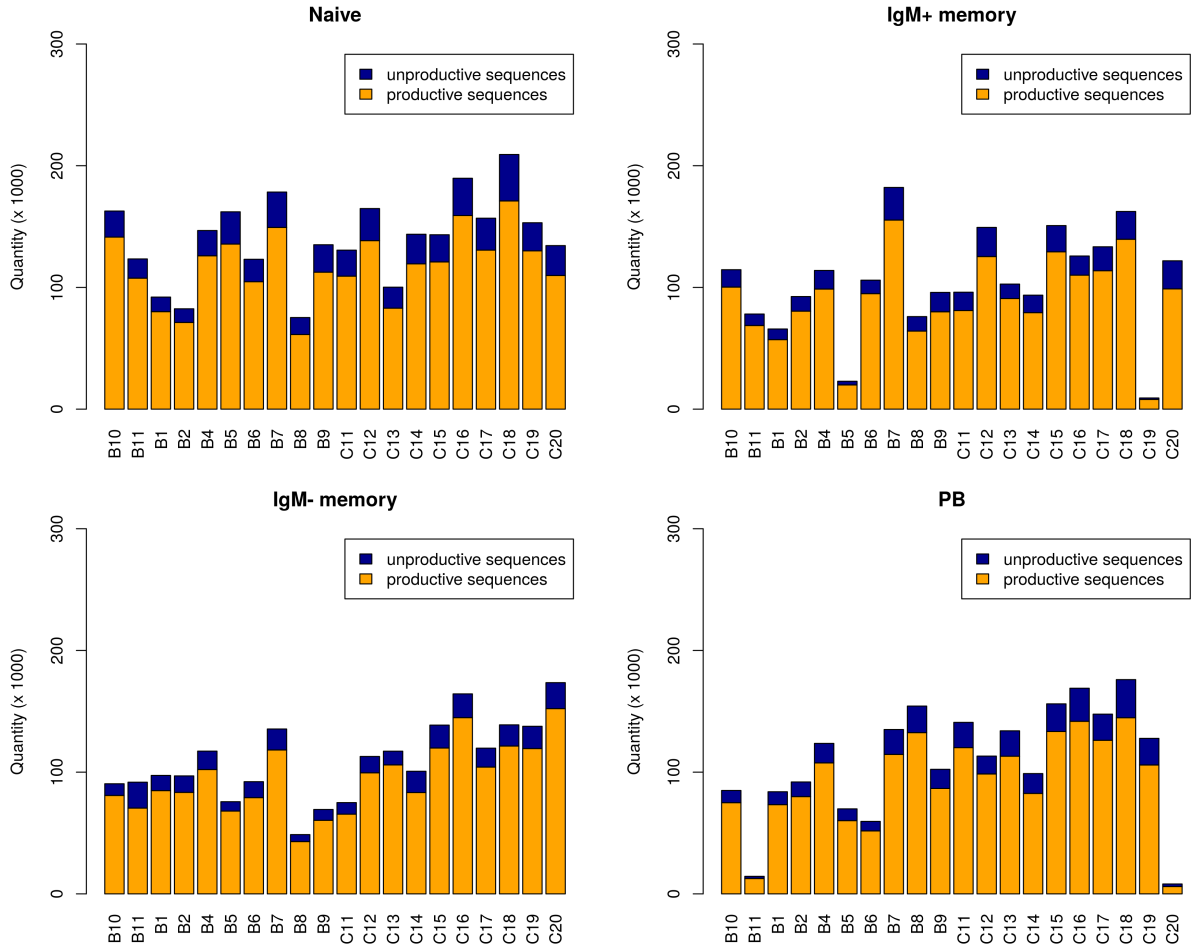

Figure 4.13: Total number of sequences of BP patients and controls resulting from IMGT/HighV-QUEST analysis. Individuals are listed o the x-axis (C11-20: controls; B1-11: BP patients). Percentage of productive and unproductive sequences is color-coded.

Next, the ratio of the mean number of mutations per sequence of BP patients and controls was investigated (Fig. 4.15). First the average number of mutations and then the ratio of cases vs. controls were calculated. Analyses were performed for the total V region and for CDR1/2 and FR1-3 sequences. For all sequence parts the ratios of IgM+ and IgM- memory cells were smaller than one and for naïve cells they were larger than one. Ratios of plasmablasts were larger than one for V, CDR1/2 and FR2; slightly smaller than one for FR1/3. Ratios higher than 1.6 were found for mutations in FR2 of naïve cells and PB. These results indicate that in naïve cells and PB there are more mutations per sequence in BP patients compared to controls on average. In contrast, in IgM+ and IgM- cells, there are fewer mutations per sequence in BP patients. FR1 is the only region, where cases and controls have similar number of mutations, independent of the B cell subset.

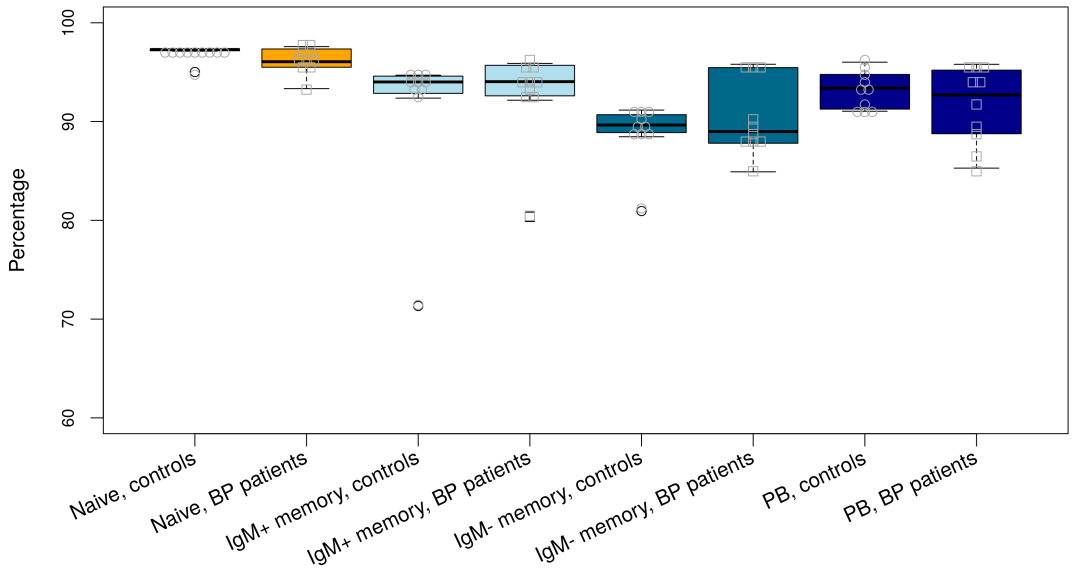

Figure 4.14: Average percentage of V gene identity compared to germline in BP patients and controls. Groups are shown on the x-axis, percentage of V gene sequence identity on the y-axis. Highest percentages were found for naïve cells, followed by IgM+, IgM- memory cells and plasmablasts.

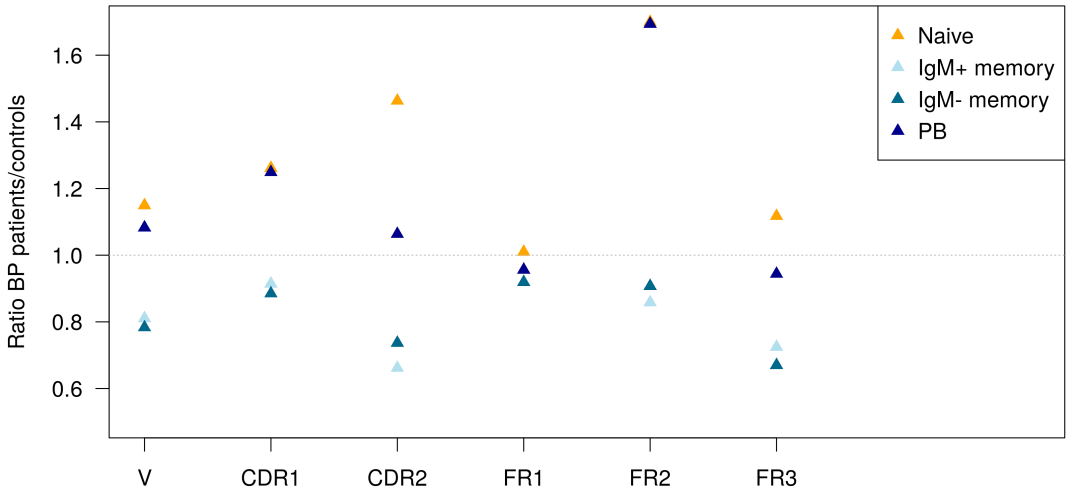

Figure 4.15: Ratio of mean numbers of mutations per sequence in BP patients and controls. B cell subsets are color-coded. Sequence parts are listed on x-axis, ratio of patients vs. controls is given on y-axis. For almost all sequence parts, ratios of both IgM memory B cell subsets are below one. For naïve cells and plasmablasts ratios are almost always greater than one.

Next, the R/S ratios, representing the ratio of replacement to silent mutations, were studied. For most of the sequenced regions (V, CDR1/2, FR1-3) ratios did not exceed 1.5, indicating a slightly higher number of replacement mutations compared to silent mutations in both groups (Supplement Fig. S5).

Looking in more detail, there are different patterns of nucleotide mutations in BP patients and controls (Fig. 4.16). The same approach as in PV analysis (percentage difference) was used. There are more differences in naïve, IgM+ memory cells and PB

(up to 10%) than in IgM- memory cells (up to 5%). Independently of the B cell subset, there are several mutations occurring more frequently in BP patients than in controls (red fields in Fig. 4.16), for instance

- adenine (a) to cytosine (c),
- thymine (t) to guanine (g) or adenine,
- guanine to cytosine or thymine and
- cytosine to guanine or adenine.

In contrast, mutations that are more abundant in controls than cases are mainly concentrated on thymine to cytosine or guanine to adenine (blue fields in Fig. 4.16).

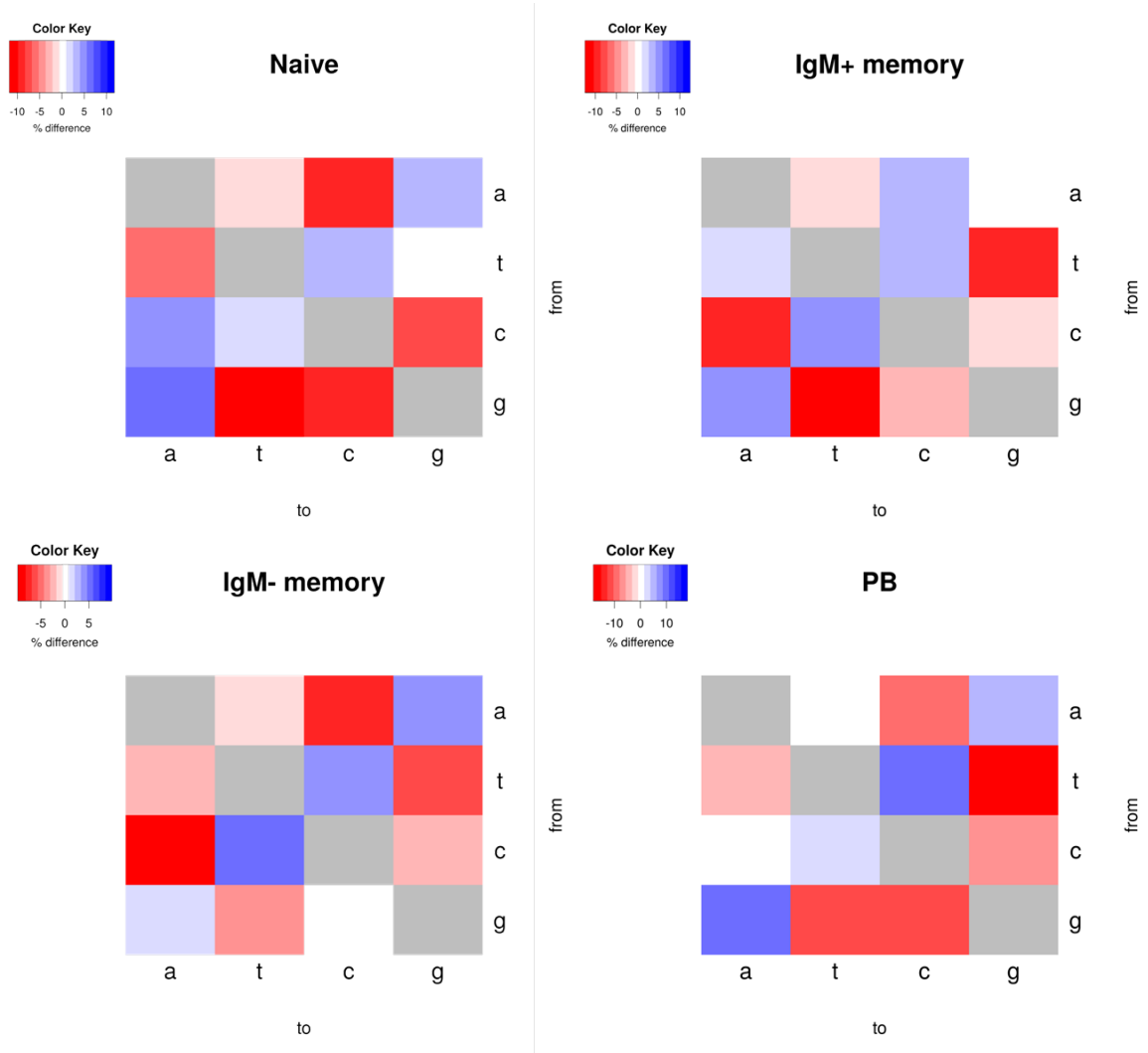

Figure 4.16: Percentages in nucleotide mutation differences between BP patients and controls, from germline (from) to mutated nucleotide (to). There are several similar patterns of mutations for all B cell subsets.

Using the same approach for amino acid changes, no considerable results can be found between BP patients and controls. In almost all B cell subsets, differences between cases and controls are less than 0.5%, indicating similar patterns in replacement mutations (Supplement Fig. S6). Taken together, these findings suggest that there might be different mutation patterns in PV patients, compared to controls, which are based on nucleotide, but not amino acid changes.

4.3.2 Clone characteristics

The same criteria were used to collapse sequences to clones as in the PV study (Fig. 4.17 A). Again most clones were found for sequences of naïve cells (ranging between 6,000 and 15,000), with slightly higher numbers in controls, than in BP patients. In all other B cell subsets, fewer clones per individual were found (in most cases less than 10,000), showing less variance within groups in patients compared to controls. Median values of patients and controls were similar for all B cell subsets.

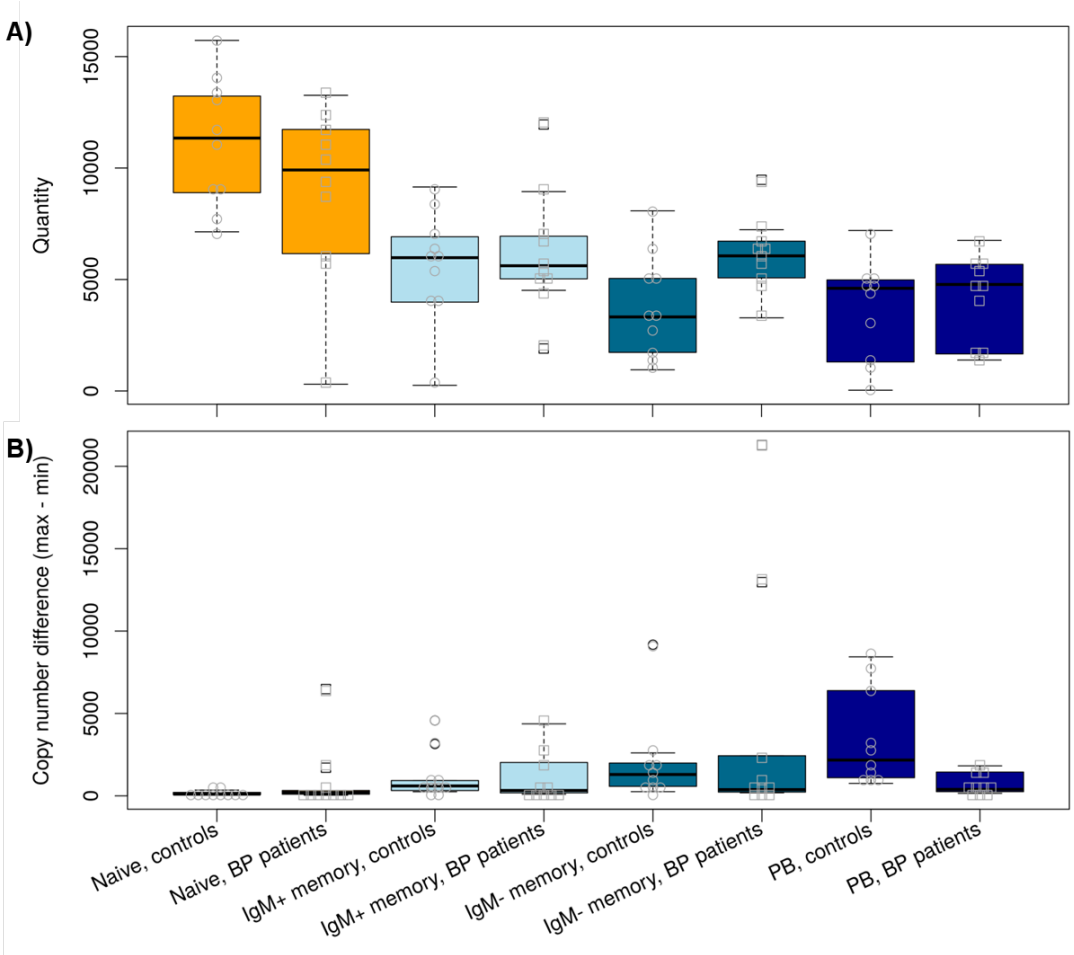

Figure 4.17: Number (A) and size (B) of clones of BP patients and controls. B cell subsets are listed on x-axis, number of clones (A) and size of clones (B) are given on y-axis. Clones in plasmablasts of controls are significantly larger than those of patients ($p<0.05$).

Considering the sizes of the clones, which represents the number of sequences belonging to each clone, the largest clones were found for IgM- cells with up to 21,285 sequences per clone (BP patients) (Fig. 4.17 B). Generally, there are larger clones in naïve and IgM- memory cells of BP patients, compared to controls. Further significant differences can be found for PB (p=0.004), indicating larger clones in controls, than cases.

Analyzing the influence of the number of sequences used for clone collapsing on the number of resulting clones, there are no clear linear trends (Fig. 4.18). For most B cell subsets no clear linear association between the number of clones and the number of sequences can be seen. Only for a few outliers amongst the naïve B cell subsets do a higher number of sequences result in a higher number of clones.

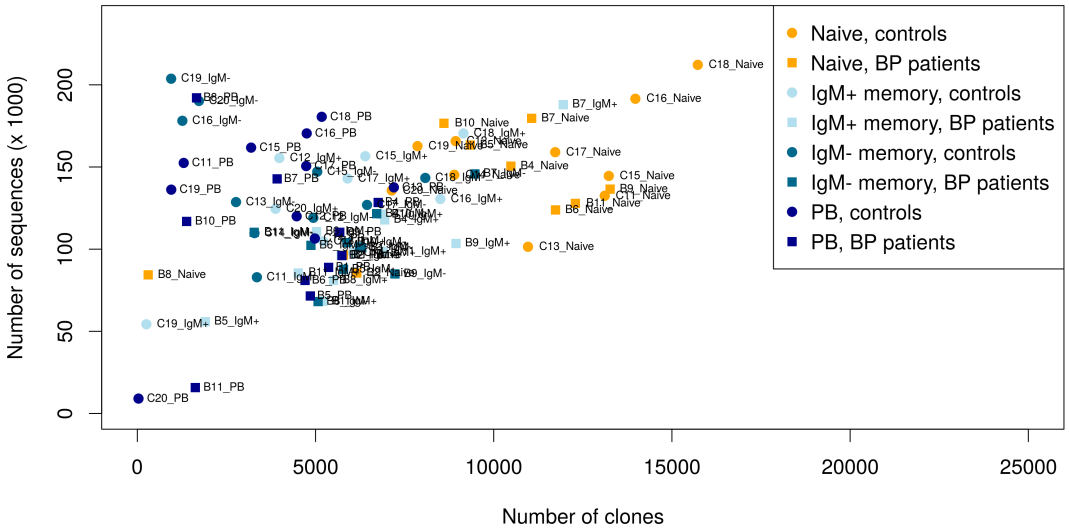

Figure 4.18: Number of clones vs. number of sequences in BP patients. The number of clones is represented on x-axis, whereas the number of sequences is shown on the y-axis. B cell subsets are color-coded, BP patients and controls can be distinguished by different symbols.

Next, the CDR3 amino acid sequence lengths of clones were studied. There were no significant differences in CDR3 lengths in both groups, but there was a trend that BP patients have some clones with smaller CDR3 sequences than controls. Most CDR3 sequences had lengths of 15 to 20 amino acids. The longest sequences of naïve cells had lengths of around 60, whereas for all other B cell subsets even longer CDR3 sequences existed (Fig. 4.19 A). Comparing only the maximum sequence lengths in clones per individual, significant differences could only be found for IgM+ memory cells (p=0.049), where BP patients contained shorter ones, compared to controls. Considering the distribution of these lengths, there were significant differences between both groups for all B cell subsets (p<0.001), except for IgM- memory cells (Fig. 4.19 B).

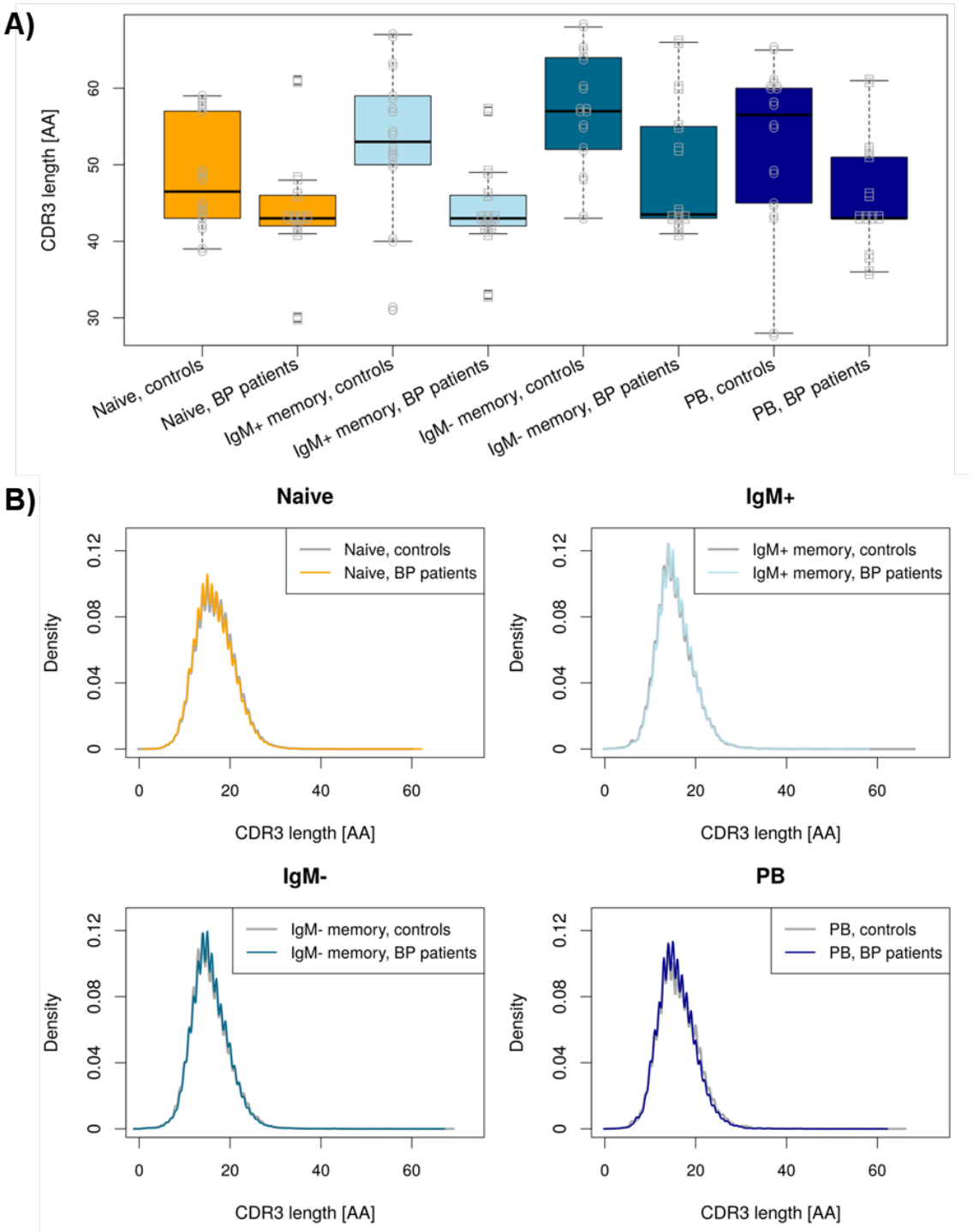

Figure 4.19: CDR3 amino acid sequence length distribution in BP patients and controls. A) Maxima of CDR3 amino acid sequence lengths are shown, with groups on the x-axis and sequence length on y-axis. In IgM+ cells BP patients contain significantly shorter CDR3 sequences, than controls. B) Kernel densities (y-axis) of CDR3 sequence lengths (x-axis) are shown (average bandwidth = 0.39).

4.3.3 Gene usage in clones

Like in the PV study, I wanted to analyze factors leading to a skewing of the repertoire. There may be genes that are overrepresented in BP patients and thus support disease onset and/or progression. But on the other hand, some genes may have a protective role and thus their expression level correlates inversely with disease severity. This leads to the question, if gene distributions in BP patients and controls differ. Therefore first a constrained analysis of principal coordinates was used to study different gene usages in B cell subsets, but also between cases and controls (Fig. 4.20). Bray-Curtis distance was applied on relative abundances of genes (columns) per individual (rows). Afterwards capscale analysis and ANOVA were performed to check for significant separation by axes. For both gene families, VH and DH, first two axes explained at least 12% of variance and significantly separated data into groups ($p<0.005$).

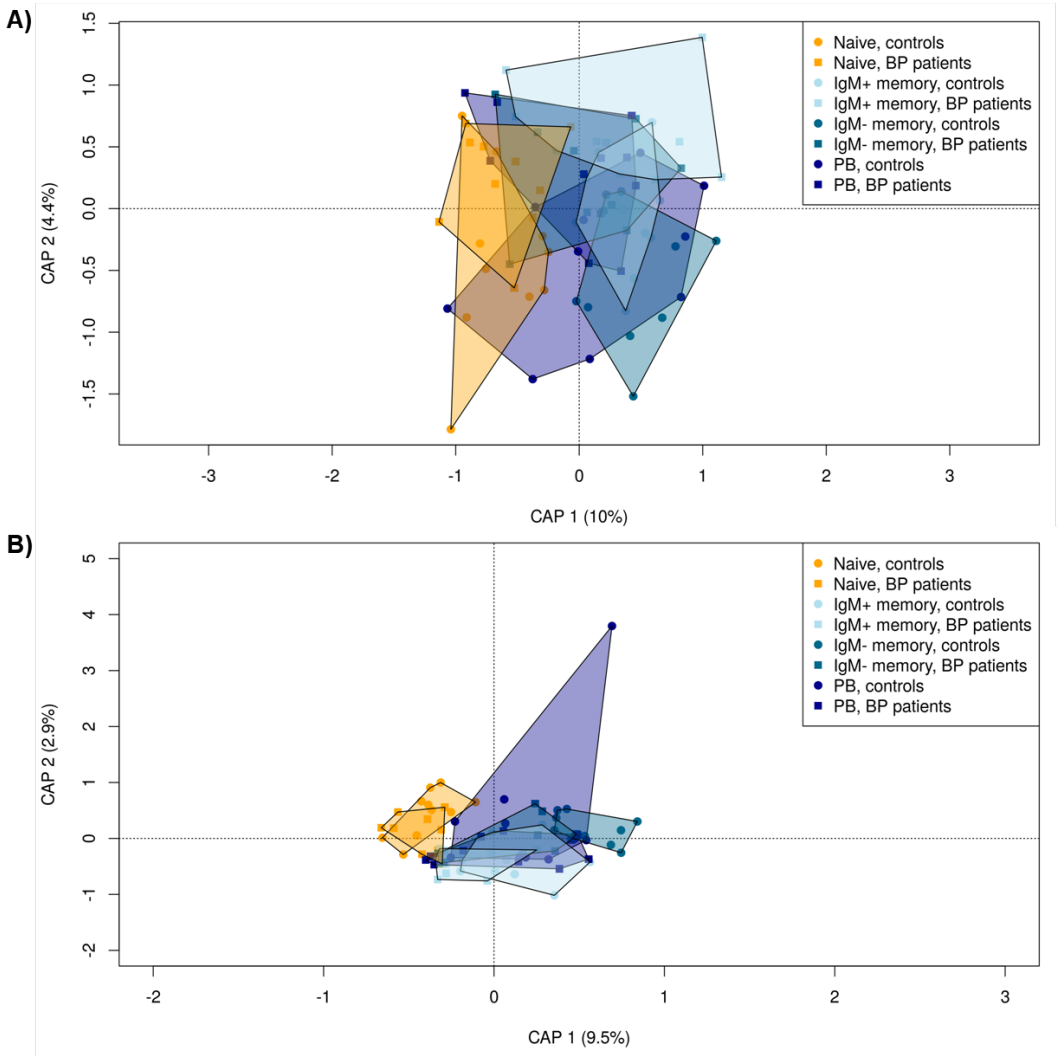

Figure 4.20: Capscale analysis of A) VH and B) DH gene usage in BP patients and controls. For both gene families, first axis explains around 10% of variance and second axis around 3-4%. B cell subsets are color-coded. BP patients are shown as squares, controls as dots.

Analyzing the VH gene usage of all B cell subsets in total, no clearly separated groups can be seen. Naïve B cells of cases and controls can be separated from both IgM memory cell subsets. Separation of BP patients and controls cannot be achieved for any of the B cell subsets, which means, that the overall gene distribution of cases and controls is similar, but there can exist differences in the abundance of individual genes.

Considering DH gene usage similar tendencies than for VH genes can be seen: naïve cells are somehow separated from the other subsets, but all in all the overall gene usage of BP patients is alike to the one of controls.

Further V gene proportions in both groups were analyzed. In Fig. 4.21 average gene proportions for all B cell subsets are shown (see also Supplement Fig. S7 for subject wise gene proportions). Genes that have significantly different abundances in BP patients and controls, are shown in Tab. 4.7. V genes of families 1-4 and 6 appear significantly different. IGHV4-30-4 has significantly different abundances in cases and controls in all four B cell subsets. Whereas genes like IGHV2-5, IGHV2-70, IGHV3-20, IGHV3-53 and IGHV6-1 appear to be significantly different in three of four B cell subsets. It is noteworthy that most of the significant genes appear only in very low abundances, which suggests that some of these differences may be due to limitations in sampling. However, differences between the groups seem not to be due to specific outliers, since numbers within the groups are similar to each other.

Not only the abundance of certain V genes, but also the abundance of V-D gene combinations may be interesting. Therefore all possible gene-gene combinations on subgroup and gene level were analyzed. First all subgroup combinations that exceed the threshold of mean $\pm$ 2 standard deviations of the difference matrix were analyzed. The difference matrix was calculated by subtracting the mean V-D combination matrix of controls from the one of the cases. Only a couple of combinations exceed the threshold (naïve: 1/49, IgM+/IgM-: 0/49, PB: 2/49). In these cases the V-D combinations are more abundant in BP patients, than in controls. Especially in naïve cells, IGHV3 together with IGHD6 seem to be more preferable in BP patients and in plasmablasts IGHV4 in combination with IGHD3 and IGHV1 with IGHD6.

Moreover the same analyses were done on gene level. In Fig. 4.22 all possible V-D gene combinations are shown. There are many combinations having higher proportions in BP patients than controls, and only few showing an inverse picture. In all subsets most of the high abundant combinations are between IGHV3 and IGHD6 subgroups. Only in PB high variability is seen, across almost all V and D genes. Also the number of combinations that pop up is almost equal in both groups (BP $>$ controls vs. BP $<$ controls).

Random forest analyses were done to see whether there are predictors that can help to distinguish between BP patients and controls, based on their V gene usage. Neither within particular B cell subsets, nor for whole data set, could cases and controls be differentiated with promising parameters (accuracy, as well as kappa values were less than 0.1).

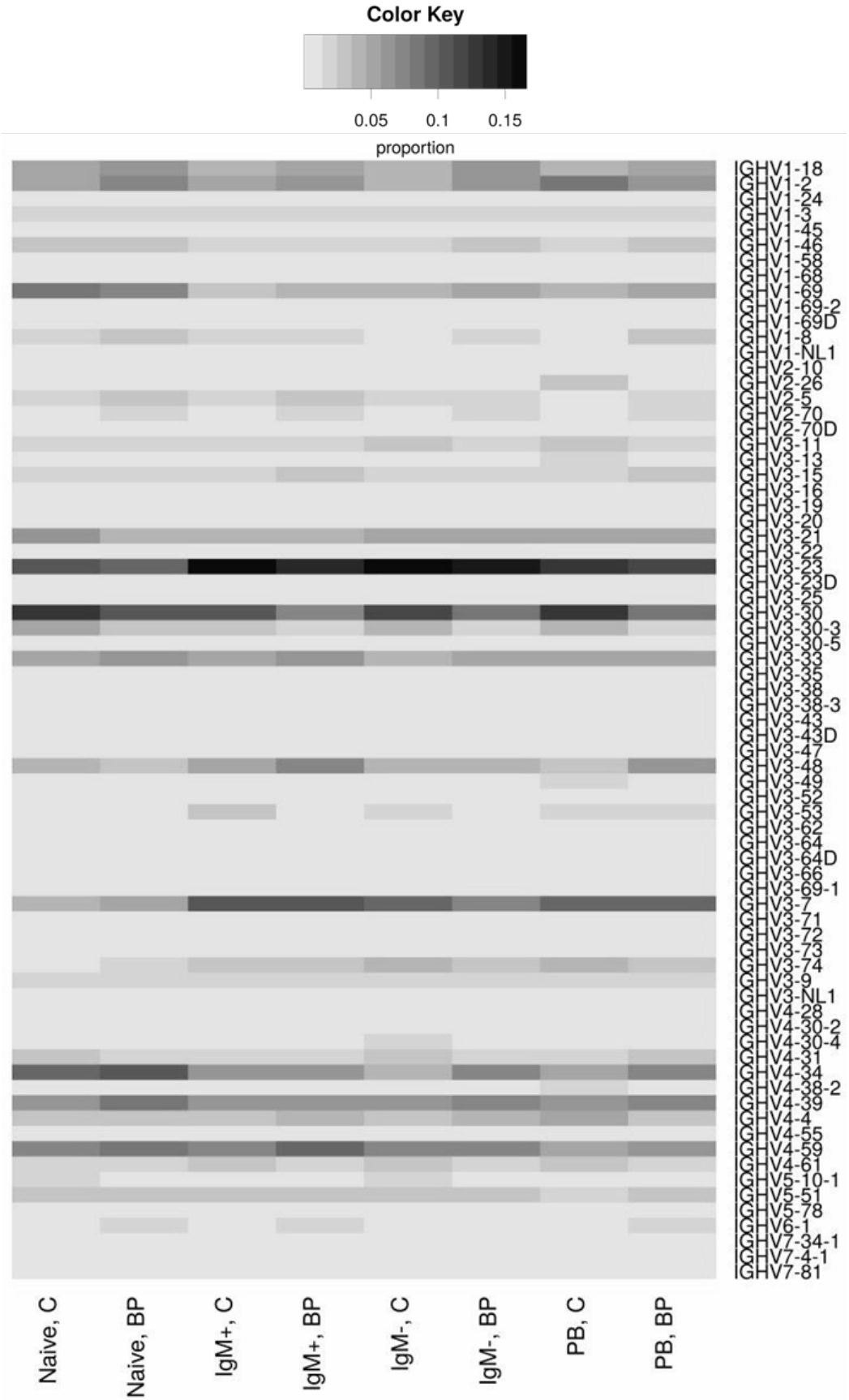

Figure 4.21: Heatmap of VH gene usage in BP patients and controls. Mean values of all individuals per group were taken. Proportions are color-coded. Light colors refer to small proportions, dark ones to high proportions. Significantly different gene proportions in cases and controls are shown in Tab. 4.7

Table 4.7: Significant differences in VH gene usage between BP patients and controls (Wilcoxon Mann Whitney test, $p<0.05$).

B cell subset	VH gene	p value	abundance
Naïve	IGHV1-69D	0.0227	BP patients < controls
	IGHV2-70	0.0052	BP patients > controls
	IGHV3-52	0.0288	BP patients < controls
	IGHV3-53	0.0342	BP patients < controls
	IGHV4-30-4	0.0139	BP patients < controls
	IGHV6-1	0.0147	BP patients > controls
IgM+	IGHV2-5	0.0232	BP patients > controls
	IGHV3-30	0.0232	BP patients < controls
	IGHV3-47	0.0311	BP patients < controls
	IGHV3-53	0.0433	BP patients < controls
	IGHV4-30-4	0.0373	BP patients < controls
	IGHV4-61	0.0288	BP patients < controls
	IGHV6-1	0.0288	BP patients > controls
IgM-	IGHV1-18	0.0068	BP patients > controls
	IGHV1-24	0.0052	BP patients > controls
	IGHV1-69D	0.0063	BP patients < controls
	IGHV2-5	0.0068	BP patients > controls
	IGHV2-70	0.0021	BP patients > controls
	IGHV3-30	0.0355	BP patients < controls
	IGHV3-30-3	0.0185	BP patients < controls
	IGHV3-30-5	0.0185	BP patients < controls
	IGHV3-53	0.0068	BP patients < controls
	IGHV4-30-4	0.0355	BP patients < controls
	IGHV6-1	0.0021	BP patients > controls
PB	IGHV1-18	0.0232	BP patients > controls
	IGHV1-24	0.0376	BP patients > controls
	IGHV2-5	0.0089	BP patients > controls
	IGHV2-70	0.0052	BP patients > controls
	IGHV3-20	0.0211	BP patients > controls
	IGHV3-48	0.0355	BP patients > controls
	IGHV4-30-4	0.0448	BP patients < controls

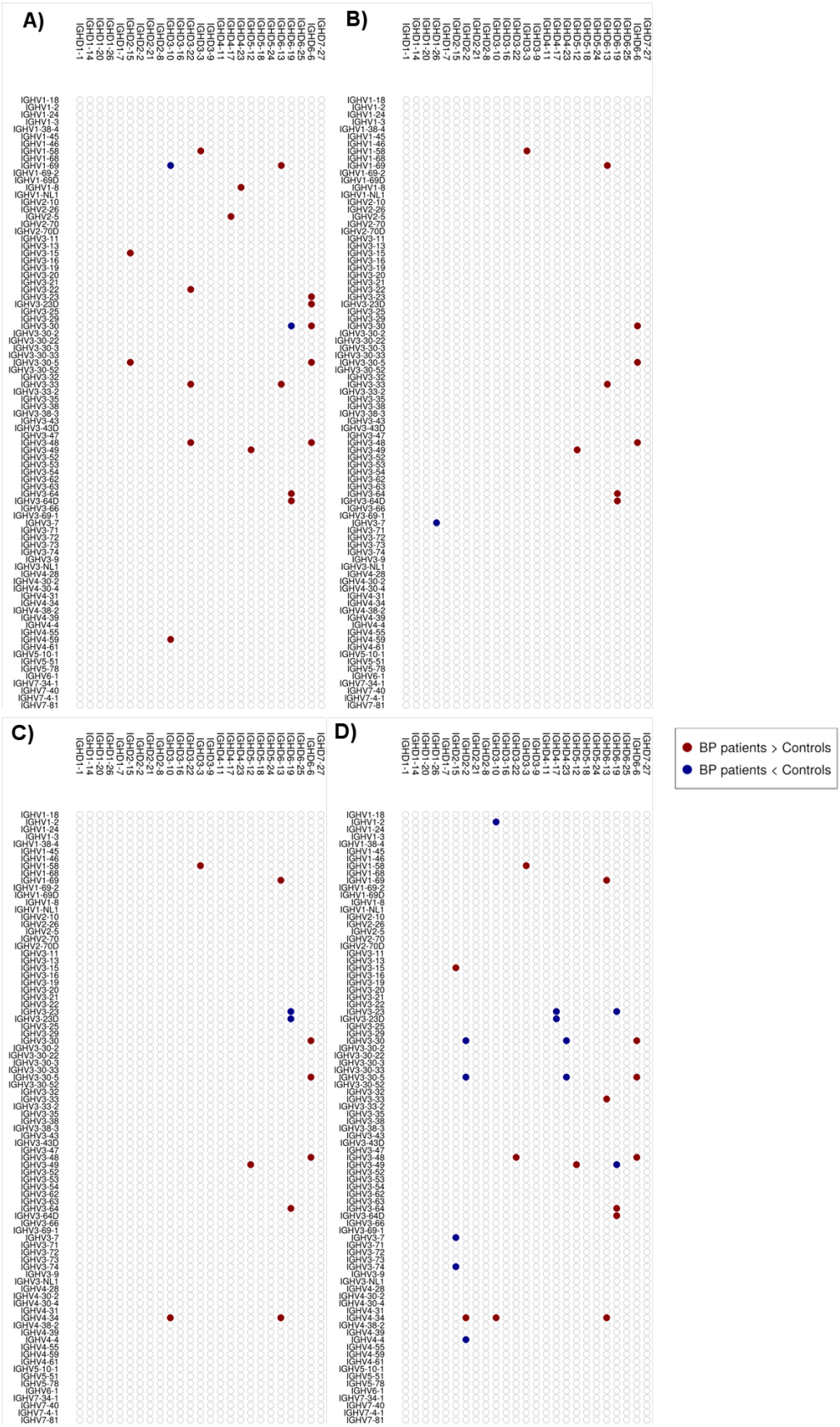

Figure 4.22: V-D gene combinations preferentially more abundant in BP patients, than in healthy controls. Combinations that exceed mean ± 2 standard deviations of the difference matrix (cases - controls) are shown. Different directions are color-coded (red and blue). A) naïve cells, B) IgM+ memory cells, C) IgM- memory cells, D) PB.

4.3.4 Diversity analyses

In the previous analyses, we already saw, that BP patients and controls differ in several aspects. The analyses of clone sizes and VH gene usage already demonstrated tendencies, that there might be differences in diversity. Therefore I analyzed diversity on clone sizes, as well as on CDR3 amino acid sequences. To measure inequality of clone size, the Gini index was used. Generally, Gini indices of BP patients were lower than in controls, except in naïve cells (Fig. 4.23). Significant differences between cases and controls could only be found for IgM+ cells (p=0.003). Whereas indices of naïve cells were relatively low (< 0.6), almost all other subsets had higher Gini values, showing that clones of both IgM memory subsets, as well as plasmablasts are more dominated by larger clones than naïve cells.

To study diversity of CDR3 amino acid sequences of clones, true diversity indices for each position of CDR3 sequences with the same length were calculated. Independent of the B cell subset, BP patients and controls showed almost identical diversity indices (Fig. 4.24). Diversity indices of plasmablasts were slightly smaller than those of the other B cell subsets.

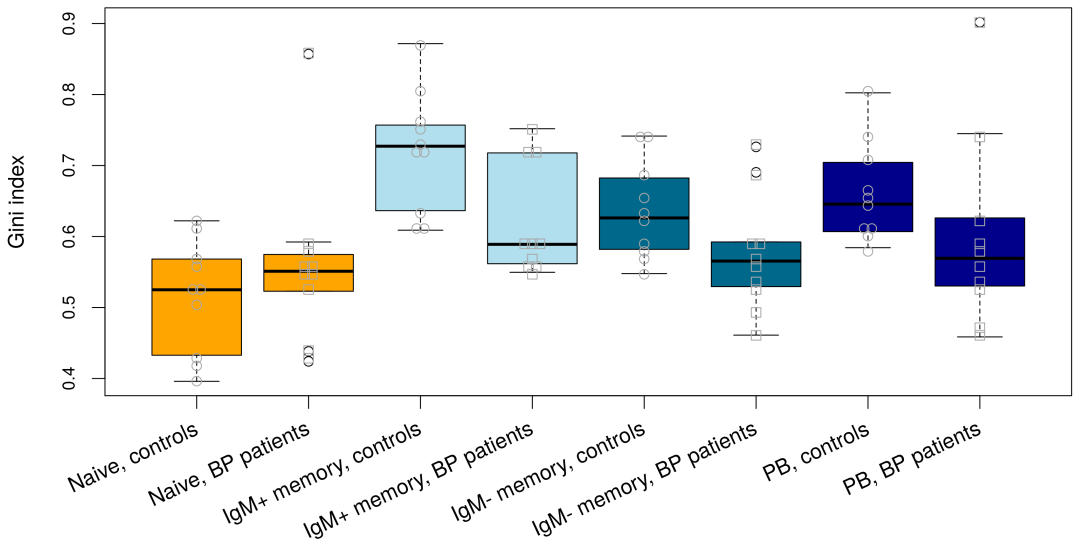

Figure 4.23: Gini index of clones of BP patients and controls. IgM+ cells of controls refer to significantly higher Gini indices than those of BP patients (p=0.003)

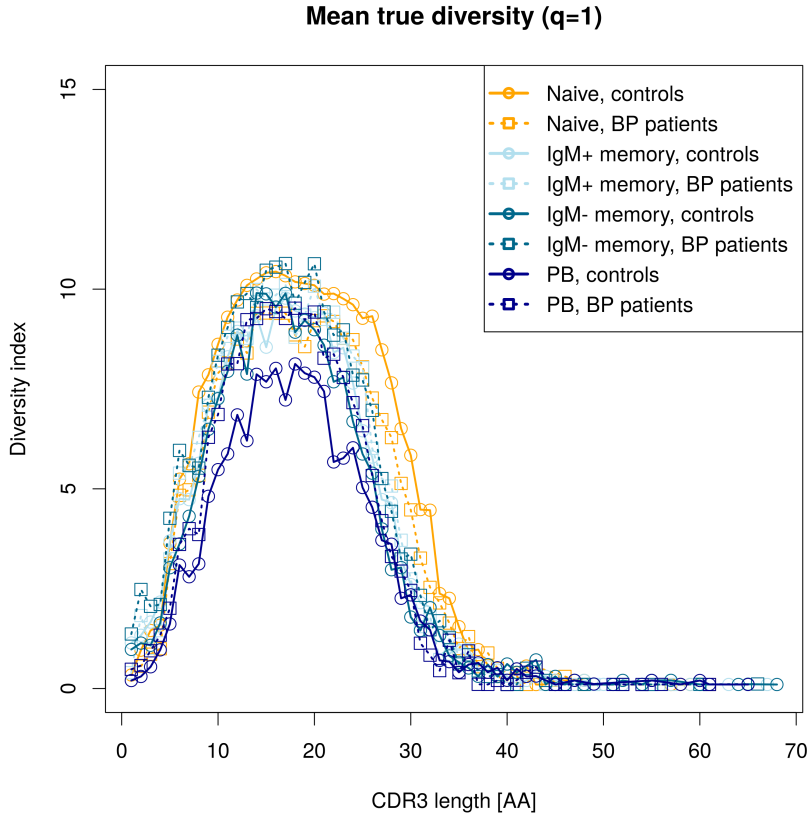

Figure 4.24: Diversity indices for CDR3 sequences of the same length in BP patients and controls. The x-axis represents CDR3 amino acid sequence lengths; the y-axis shows true diversity indices of order one. B cell subsets are color-coded. Patients are represented by squares, controls by dots.

4.4 Age and sex related changes in the B cell receptor repertoire of healthy controls

Samples of 20 healthy controls of the University Hospital of Schleswig-Holstein in Lübeck and Campus Kiel were studied (Tab 4.8). The mean age was 61 ± 21, with a minimum age of 31 and a maximum age of 81. 12 of 20 individuals were female. Individuals were clustered into four groups:

- young females: females with an age < 55 (n = 6)
- old females: females with an age ≥ 55 (n = 6)
- young males: males with an age < 55 (n = 3)
- old males: males with an age ≥ 55 (n = 5)

For dichotomization of age groups, the age of 55 was taken as this is the typical age for the menopause in women which I assumed to be important for the B cell receptor repertoire.

Table 4.8: Sample characteristics of healthy controls. Age, sex and the total number of sequences used for further statistical analyses are shown. Samples with missing data are marked as "-". F = female, M = male.

			Total number of analyzed sequences			
ID	Age	Sex	Naïve	IgM+	IgM-	PB
C1	38	F	47,772	60,598	-	-
C2	39	M	57,843	-	59,730	-
C3	31	F	344,559	-	61,246	-
C4	45	F	372,385	293,127	380,609	270,273
C5	45	M	39,370	35,670	63,555	61,845
C6	45	F	137,734	55,009	94,105	42,509
C7	40	F	68,539	88,452	107,966	35,088
C8	37	F	53,412	116,409	242,567	93,562
C9	55	F	248,744	30,348	503,409	39,171
C10	45	M	13,457	89,315	51,509	33,893
C11	64	F	114,663	85,834	69,235	125,505
C12	87	F	145,480	131,127	103,831	102,261
C13	86	M	87,455	94,262	109,802	119,300
C14	74	F	125,180	83,092	87,823	86,647
C15	76	M	126,612	135,529	124,898	140,591
C16	82	M	166,770	114,817	151,588	150,182
C17	87	M	137,238	119,358	109,277	131,249
C18	82	F	180,779	146,955	127,764	152,058
C19	84	M	13,6517	8,503	12,3502	110,643
C20	75	F	115,734	104,337	161,903	6,850

4.4.1 Mutation analysis

It is already known that the numbers of mutations change with age. Comparing the V gene sequence identity to germline sequence, sequences of naïve B cells have fewer mutations per sequence than all other B cell subsets (Fig. 4.25), as expected. Sequences of IgM- memory cells have more mutations per sequence than IgM+ memory cells, whereas plasmablasts lie in between, with high variances within groups. Furthermore, naïve B cells of young males have significantly fewer mutations than old males (p=0.036).

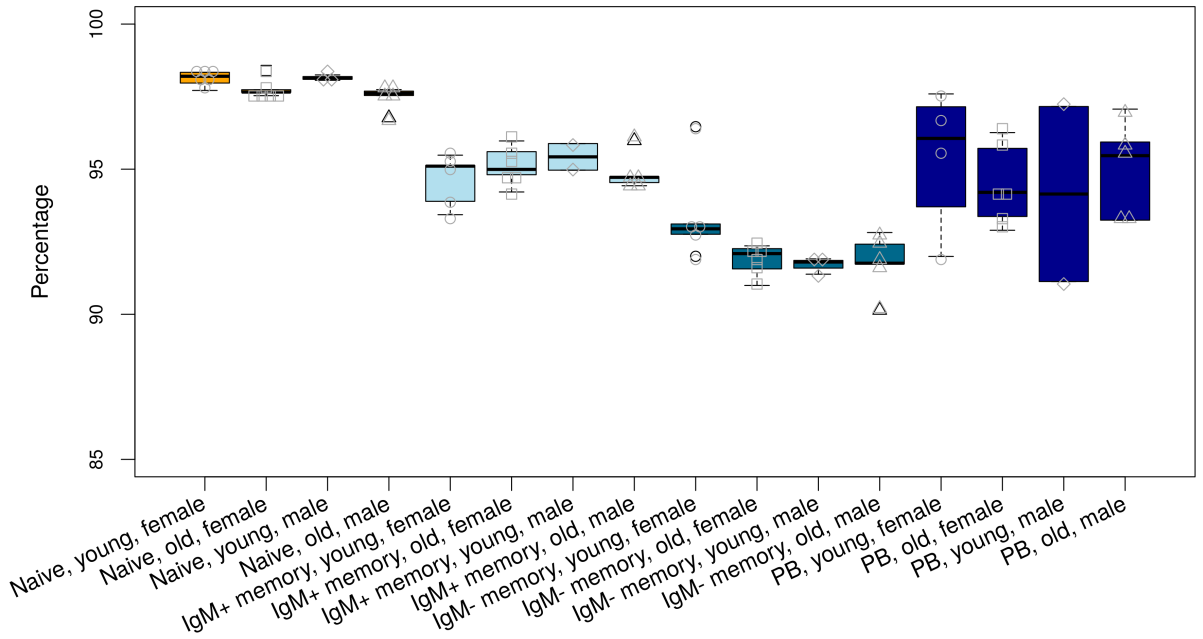

Figure 4.25: Percentage of V gene sequence identity compared to germline sequence in healthy controls. Groups are listed on the x-axis, whereas percentages are on the y-axis. Significance ($p<0.05$) could be reached for naïve cells, considering young and old males.

Next I wanted to have a more detailed look if there were nucleotide changes during aging, but also if there were differences between females and males. Therefore percentages of nucleotide mutations in females and males were analyzed and percentage difference matrices (as explained in PV study) of young vs. old individuals were visualized (Fig. 4.26). There are some mutations that occur more often in old than in young individuals, independent of gender or B cell subset:

- Guanine (g) to cytosine (c) or thymine (t),
- Thymine to adenine (a),
- Adenine to cytosine and
- Cytosine to guanine.

All these mutations are transversions (substitutions from purine to pyrimidine or vice versa). There are no further B cell subset specific mutations with high proportions.

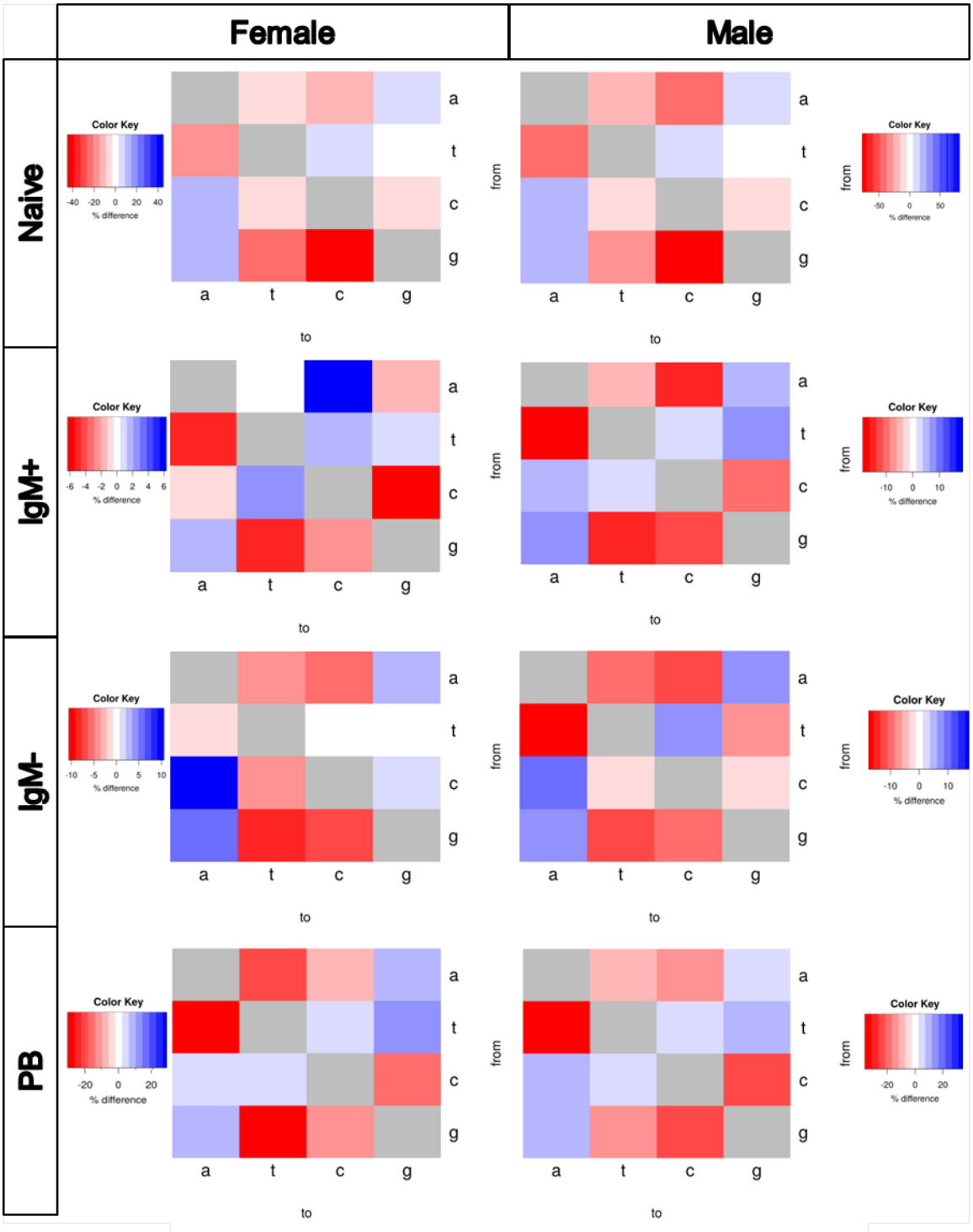

Figure 4.26: Ratios of nucleotide mutations, comparing young and old individuals. Ratios are color-coded: red colors represent mutations that appear more often in old individuals, compared to young ones; blue colors represent mutations that appear more often in young individuals, compared to old ones. White fields represent no difference between young and old samples; gray fields were not analyzed.

4.4.2 Clone characteristics

As before, sequences were collapsed to clones, when they had the same V and J genes and at least 85% CDR3 amino acid sequence identity (and the same CDR3 sequence length). Individuals had most clones, when analyzing naïve cells (median = 10,014) (Fig. 4.27). The median numbers of clones for both IgM memory subsets and plasmablasts were less than 4,000 (IgM+: median = 3,933; IgM-: median = 2,769; PB: median = 1,150). Old individuals tended to have more memory clones, than young ones.

Analyzing clone size, the number of sequences belonging to a clone, significant differences could just be seen for naïve B cells of males, where old males have significantly larger clones than young males (p=0.024) (Fig. 4.27). Also in the other B cell subsets, old men had generally larger clones than young ones.

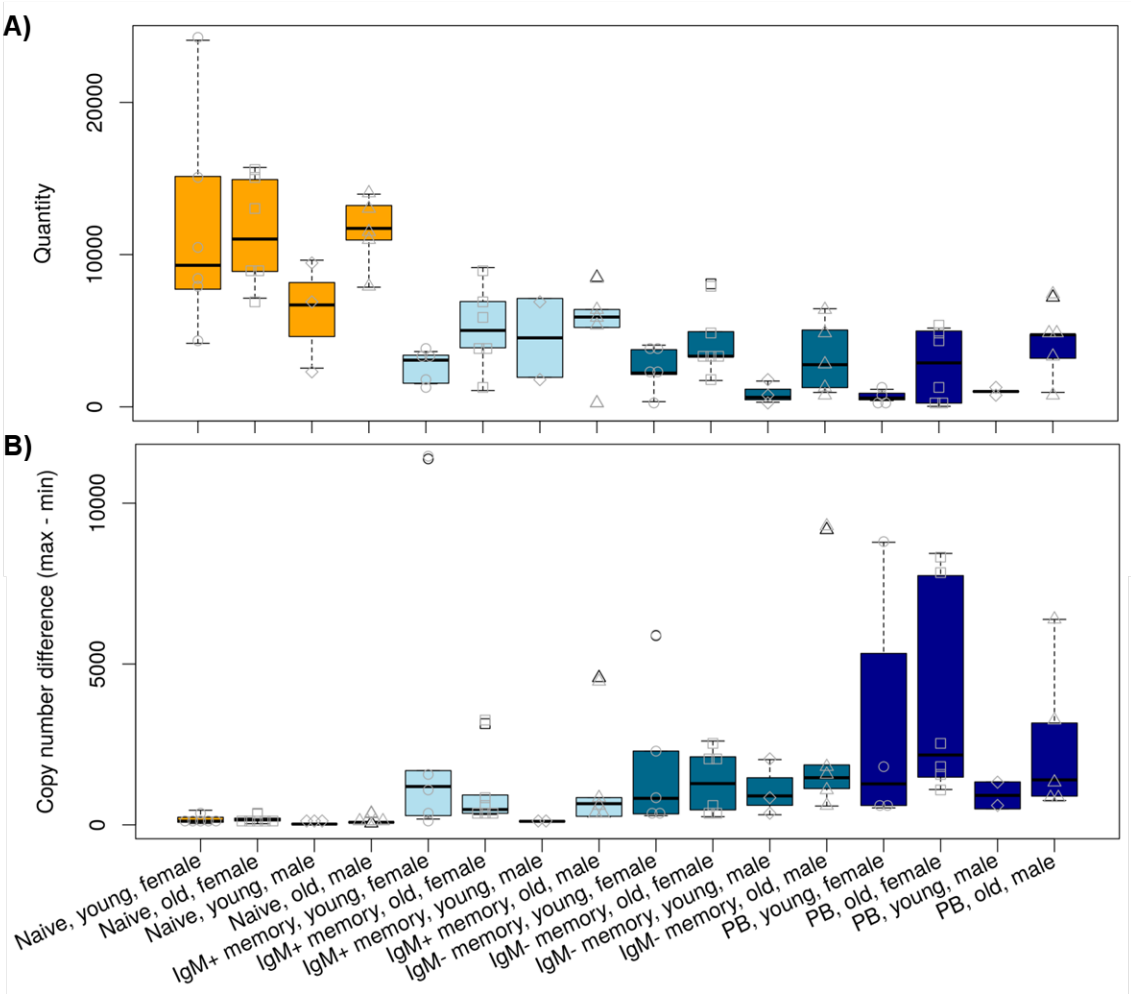

Figure 4.27: A) Number and B) size of clones of healthy controls. Groups are listed on the x-axis, quantities on the y-axis. Naïve cells contained most clones, but smallest ones; PB vice versa. Clones of naïve cells of old individuals are significantly larger than those of young males (p=0.024).

Next CDR3 amino acid sequence lengths were studied. Looking at the maximum CDR3 length per individual, in general old individuals contained longer CDR3 sequences than

young ones, in all B cell subset (Fig. 4.28 A). In young individuals the maximum CDR3 sequence length was at about 31 amino acids, whereas old individuals contained sequences with over 60 amino acids. In naïve cells and IgM- memory cells old individuals contained

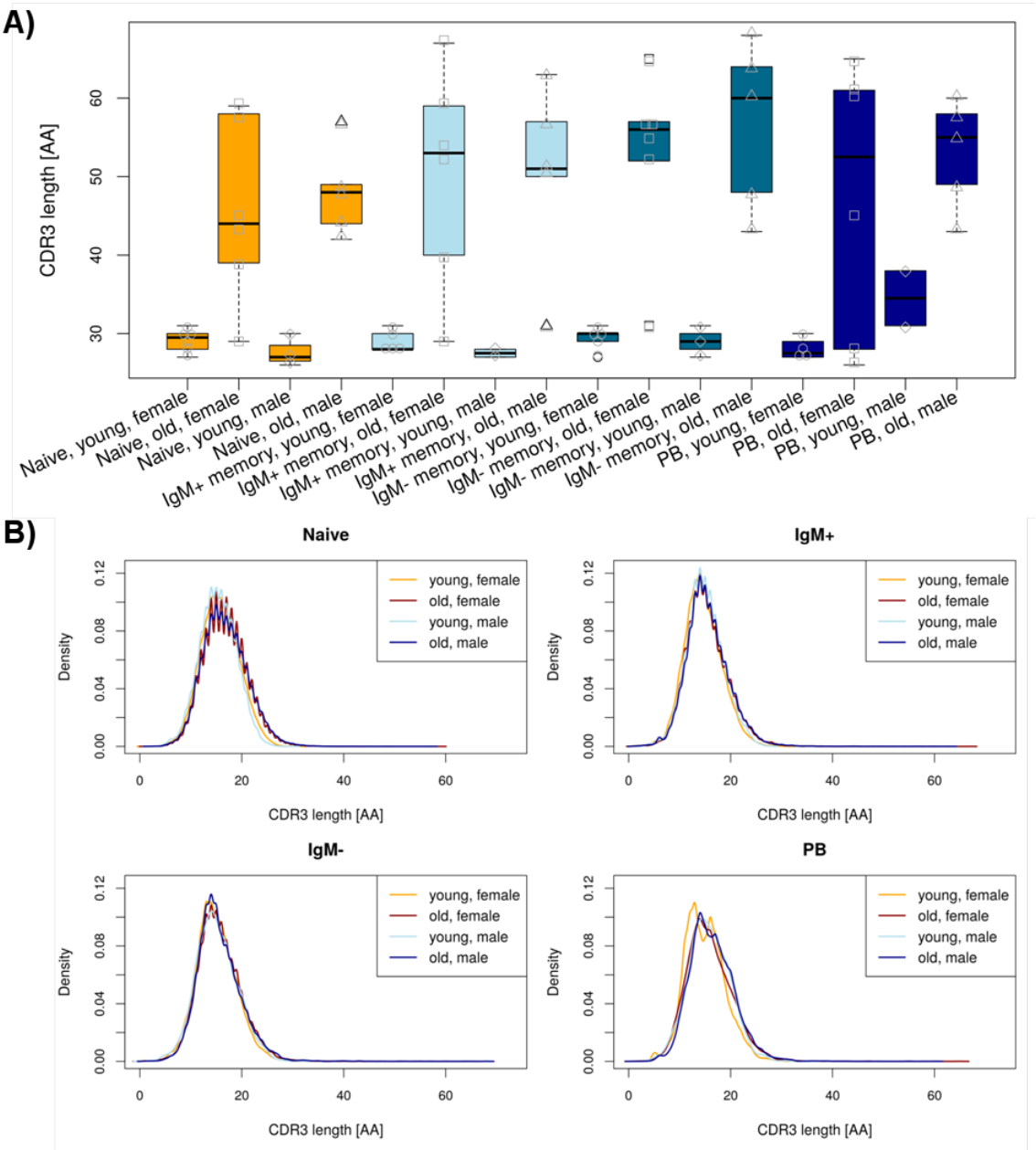

Figure 4.28: CDR3 amino acid sequence length distributions of clones in healthy controls. Maxima of CDR3 sequence lengths of clones are shown in panel A). In naïve cells and IgM- memory cells old individuals contained significantly longer CDR3 sequences ($p<0.05$), than young ones, in both females and males. In IgM+ memory cells only old women had significantly longer CDR3 sequences than young women ($p<0.05$). CDR3 length distributions (kernel density, with average bandwidth = 0.51) are shown in panel B) (CDR3 amino acid length on the x-axis, density of the y-axis; groups are color-coded).

significantly longer CDR3 sequences (naïve: female p=0.024, male p=0.035; IgM-: fe-

male p=0.010, male p=0.035), than young ones, in both females and males. In IgM+ memory cells only old women had significantly longer CDR3 sequences than young women (p=0.021). Moreover I was interested if the sequences with long CDR3's contain special V genes. Within this subset similar VH gene usage pattern for all clones were seen (Supplement Fig. S8).

Analyzing the length distribution using kernel identity with a mean bandwidth of 0.51, for almost all combinations yielded significant differences (p<0.001) amoung young vs. old, as well as female vs. male (Fig. 4.28 B). Only in naïve and IgM- memory cells did young females and males have similar CDR3 sequence lengths. For all other combinations the sequence length (x-axis), as well as the proportions (y-axis) differed significantly between the four groups. For example CDR3 sequences of clones in naïve cells having lengths of 15 to 18 amino acids, were significantly more abundant in young individuals compared to old ones (orange and light blue lines vs. dark red and dark blue lines in Fig. 4.28 B).

4.4.3 Gene usage in clones

Constrained analysis of principal coordinates for VH and DH gene usage containing all B cell subsets showed no real separation of the B cell subsets between young and old or male and female, although the first three axes were significant using ANOVA (p<0.05) (Supplement Fig. S9). For both VH and DH gene usage the first two axes explained about 14-16% of variance.

Going more into detail and analyzing gene usage of the four groups separately for each B cell subset, ANOVA showed significantly different VH gene usage between young and old individuals in naïve cells and IgM+ memory cells (naïve: p=0.047, IgM+: p=0.027) (Fig. 4.29, Supplement Fig. S10). In both subsets separation of the first axis (CAP1) also appears to be significant (naïve: p=0.042, IgM+: p=0.023). Again, the first two axes explain up to 17% of the variance.

These findings led me to a more detailed analysis of gene usage. Differences between age and gender specific groups, as well as linear or bell curved age trends were studied.

Fig. 4.30 represents the average VH gene usage of all groups. Significant differences (p<0.05) could be found for the genes shown in Tab. 4.9. There are more VH genes that are significantly different abundant between gender, than for age. In naïve cells different expressed genes belong to IGHV subgroups 1, 2 and 3. Genes like IGHV1-58, IGHV1-NL1, IGHV2-26, IGHV2-5, IGHV2-70 and IGHV3-7 are significantly different between young and old individuals in females and males. Whereas for IgM+ memory cells only significantly different (p<0.05) VH gene abundances could be found in males, for IgM- cells and PB almost all genes are only significantly different in females. Most of these genes belong to VH subgroup 1 and 3. Considering gender differences, significant differences

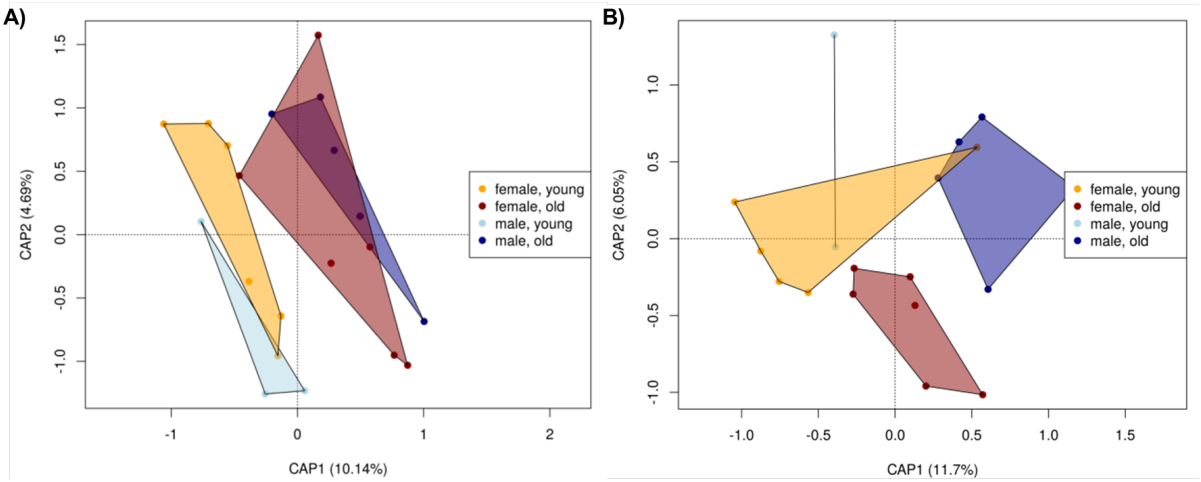

Figure 4.29: Constrained analysis of principal coordinates of VH gene usage in A) naïve and B) IgM+ memory cells, using Bray-Curtis dissimilarity. In both subsets also separation of the first axis (CAP1) appears to be significant ($p<0.05$). Again the first two axes explain up to 17% of the variance. Groups are color-coded. In B) only data of two young males was available (two light blue dots, connected by black line).

between females and males in young individuals, but not old ones, could only be found in naïve B cells and plasmablasts. Whereas the genes in naïve cells mostly belong to IGHV subgroup 2, the ones of PB mostly belong to IGHV subgroup 3. For both IgM memory subsets significant differences between females and males could be found for both groups, young and old individuals. Both B cell subsets are again dominated by genes of subgroup 1 and 3, but showing almost no overlap in significant results.

I further looked for more general associations on VH subgroup level. Here, significant differences ($p<0.05$) in gender could be found for IGHV1 and IGHV5 in IgM+ memory cells. Abundances were higher in males, than females, in both genes.

I was not only interested in differences between age and gender specific groups, but also in age related trends in VH gene usage. Therefore linear and bell curved trends during age were analyzed. Significant linear trends ($p<0.05$) could be found for both, females and males. Bell curved trends could only be seen in females, which leads to the speculation that some VH genes may be under the influence of hormonal differences between males and females. Significant linear associations of gene proportions and age are shown in Tab. 4.10 ($p<0.05$). Not only results of a generalized linear model, showing linear associations of gene proportions and age or the interaction of age and sex, are shown, but also correlation of gene abundances and age in females and males. Most significant linear associations can be seen in naïve B cells, containing genes of VH families 1, 2, 3, 5 and 7. For three of these genes also the interaction term of age and sex appears to be significant: IGHV3-22 has a strong positive trend in males (cor = 0.87), but no significant linear trend

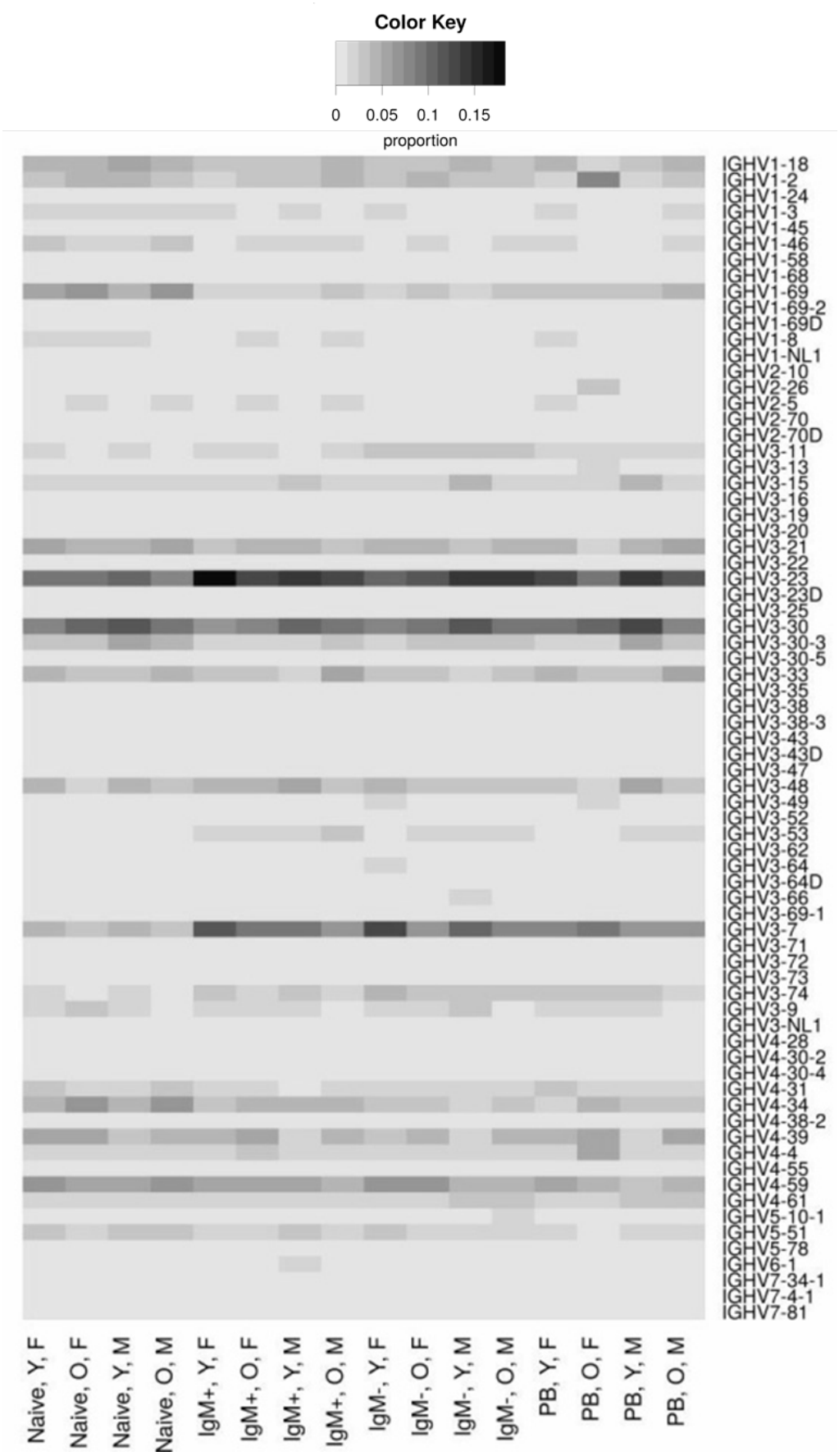

Figure 4.30: Average VH gene usage of healthy controls. Proportions are color-coded: light colors represent low proportions, dark colors represent high proportions. Genes with significant different proportions between groups are listed in Tab. 4.9 ($p<0.05$).

in females. However, IGHV3-30-5 proportions show a strong positive linear trend with age (cor =0.05) and IGHV7-81 a negative linear trend (cor = -0.65) in females, but no significant association in males.

Table 4.9: Gender and age specific significant differences in VH gene usage in healthy controls (Wilcoxon-Mann-Whitney test, p<0.05). Subset = B cell subset; group = groups used for Wilcoxon test; F vs. M = all females vs. males; young, F vs. M = females vs. males in young individuals; old, F vs. M = females vs. males in old individuals; Y vs. O = all young vs. old; female, Y vs. O = young females vs. old females; male, Y vs. O = young males vs. old males; female & male, Y vs. O = young females vs. old females and young males vs. old males.

	gender specific differences		**age specific differences**	
VH gene	subset	group	subset	group
IGHV1-18	IgM+	F < M	IgM-	male, Y > O
	PB	old, F < M		
IGHV1-24			Naïve	Y > O
IGHV1-46	PB	old, F < M		
IGHV1-58			Naïve	Y > O
			Naïve	female & male, Y > O
IGHV1-69D			Naïve	Y < O
			Naïve	male, Y < O
IGHV1-NL1			Naïve Y < O	
			Naïve	female & male, Y < O
IGHV2-10			IgM-	Y < O
IGHV2-26	Naïve	young, F < M	Naïve, IgM+	Y < O
			Naïve	female & male, Y < O
			IgM+	female, Y < O
IGHV2-5			Naïve	Y < O
			Naïve	female & male, Y < O
IGHV2-70	Naïve	young, F > M	Naïve, IgM+, PB	Y < O
			Naïve	female & male, Y < O
			IgM+, IgM-	female, Y < O
IGHV2-70D			Naïve	Y < O
IGHV3-15			IgM-	Y > O
			IgM-	male, Y > O
IGHV3-20	IgM-	young, F < M		
IGHV3-22			IgM-	Y < O
			Naïve, IgM-	male, Y < O
IGHV3-23D	PB	old, F < M	IgM-	Y < O

	gender specific differences		**age specific differences**	
VH gene	subset	group	subset	group
			Naïve	male, Y < O
			IgM-	female & male, Y <O
IGHV3-30-5			Naïve	Y < O
			IgM-	female, Y < O
IGHV3-38	Naïve	F > M		
IGHV3-43	Naïve	young, F > M		
IGHV3-43D			IgM-	Y > O
IGHV3-47	IgM-	F > M		
IGHV3-48			IgM-	Y > O
			Naïve	female, Y > O
IGHV3-69-1			Naïve, IgM-, PB	Y < O
			PB	female, Y < O
IGHV3-7			Naïve	Y > O
			Naïve	female & male, Y > O
IGHV3-72	IgM-	F < M	IgM+	Y > O
	IgM+	old, F > M		
IGHV3-73			IgM+	Y > O
IGHV3-74			Naïve	Y > O
			Naïve	female, Y > O
IGHV3-NL1			IgM-	Y < O
			IgM-	female, Y < O
IGHV4-31			IgM-	Y < O
IGHV4-34			Naïve	Y < O
IGHV4-39	IgM+	old, F > M	IgM-	Y < O
IGHV4-59	IgM-	F > M		
IGHV4-61	PB	F < M		
IGHV5-51			PB	female, Y > O
IGHV6-1			IgM-	Y > O
			IgM-	male, Y > O

Significant linear associations with age in IgM+ memory cells appear within 6 genes of VH families 1, 2, 3 and 4, with two genes having also a significant interaction term: IGHV3-52 has a significant negative correlation with age in females (cor = -0.62), whereas IGHV3-62 correlates negatively with age in males (cor = -0.63). Further significant linear associations can be found for IgM- memory cells. These genes belong to VH families 1, 2 and 3. Five of eleven genes show significant interaction terms, but only three of them significant correlations within females or males: Within males, IGHV3-15 is negatively

correlated with age (cor = 0.77), whereas IGHV3-22 is positively correlated (cor = 0.78). Within females IGHV3-48 shows a negative correlation with age (cor = -0.74). In plasmablasts only two genes are significantly linear associated with age, belonging to VH families 1 and 5. Both do not have significant interaction terms or strong correlations within females or males.

There are some genes which are significantly quadratic associated ($p<0.05$) with age in women, but not in men (Fig. 4.31). This bell shaped progression may show associations with hormones, like estrogen, which has highest levels in women at an age between 50 and 55 and afterwards levels decrease again. In naïve B cells IGHV3-43D and IGHV3-66 show age dependent curves with at minimum at 55 years. In IgM+ memory cells IGHV3-33 shows the same tendency, whereas IGHV3-20 and IGHV3-53 have their maxima at around 55 years. IGHV3-20 shows the age dependent same bell shaped curve in IgM- cells, as well as IGHV3-69-1. In plasmablasts IGHV1-18 and IGHV1-24 levels decrease 55 years and afterwards increase.

Table 4.10: Significant linear associations (p<0.05) of a) gene usage and age and b) gene usage and age dependent of sex in healthy females and males. A generalized linear model (GLM) was used to get linear associations of gene usage and age or interaction of age and sex (age:sex). Correlation test (Pearson's product-moment correlation) was used to get trends of linear association in females and males. n.s. = not significant (p > 0.05); R^2 = coefficient of determination; cor = correlation coefficient

		GLM			Correlation test			
B cell subset	VH gene	p value (age)	p value (age:sex)	R^2	p value (females)	cor (females)	p value (males)	cor (males)
Naïve	IGHV1-58	0.0001	n.s.	0.73	0.0001	-0.88	n.s.	-0.70
	IGHV2-26	0.0002	n.s.	0.77	0.0029	0.78	0.0001	0.97
	IGHV2-5	0.0064	n.s.	0.62	0.0137	0.69	0.0046	0.87
	IGHV2-70	0.0435	n.s.	0.47	0.0108	0.70	n.s.	0.57
	IGHV3-22	n.s.	0.0210	0.45	n.s.	0.02	0.0055	0.87
	IGHV3-30-5	0.0001	0.0223	0.67	0.0005	0.85	n.s.	0.45
	IGHV3-47	0.0225	n.s.	0.45	0.0454	-0.59	n.s.	0.25
	IGHV3-48	0.0083	n.s.	0.38	0.0265	-0.63	n.s.	-0.32
	IGHV3-7	0.0003	n.s.	0.63	0.0021	-0.79	0.0317	-0.75
	IGHV3-74	0.0015	n.s.	0.56	0.0011	-0.82	n.s.	-0.53
	IGHV5-51	0.0034	n.s.	0.49	0.0046	-0.75	n.s.	-0.18
	IGHV7-81	0.0118	0.0184	0.40	0.0128	-0.69	n.s.	0.36
IgM+	IGHV1-NL1	0.0024	n.s.	0.72	0.0090	0.74	n.s.	0.63
	IGHV2-10	0.0331	n.s.	0.33	n.s.	0.55	n.s.	0.34
	IGHV3-52	0.0142	0.0425	0.46	0.0384	-0.63	n.s.	0.42
	IGHV3-62	n.s.	0.0299	0.46	NA	NA	n.s.	-0.63
	IGHV4-39	0.0461	n.s.	0.44	n.s.	0.58	n.s.	0.54

B cell subset	VH gene	GLM p value (age)	GLM p value (age:sex)	GLM R^2	Correlation test p value (females)	Correlation test cor (females)	Correlation test p value (males)	Correlation test cor (males)
IgM-	IGHV1-24	n.s.	0.0257	0.43	n.s.	0.14	n.s.	-0.67
	IGHV1-46	0.0221	n.s.	0.32	0.0290	0.65	n.s.	0.06
	IGHV1-69-2	0.0366	n.s.	0.49	n.s.	0.51	NA	NA
	IGHV2-70	0.0485	n.s.	0.35	0.0067	0.76	n.s.	0.06
	IGHV3-15	n.s.	0.0241	0.58	n.s.	-0.35	0.0251	-0.77
	IGHV3-22	n.s.	0.0356	0.49	n.s.	0.19	0.0236	0.78
	IGHV3-23D	0.0163	n.s.	0.67	0.0139	0.71	0.0120	0.82
	IGHV3-48	0.0028	0.0323	0.51	0.0086	-0.74	n.s.	-0.02
	IGHV3-49	0.0304	n.s.	0.36	n.s.	-0.58	n.s.	-0.07
	IGHV3-66	n.s.	0.0298	0.39	n.s.	0.15	n.s.	-0.69
	IGHV3-7	0.0410	n.s.	0.27	n.s.	-0.54	n.s.	-0.26
PB	IGHV1-69D	0.0321	n.s.	0.42	n.s.	-0.57	n.s.	0.13
	IGHV5-51	0.0140	n.s.	0.52	0.0419	-0.65	n.s.	-0.04

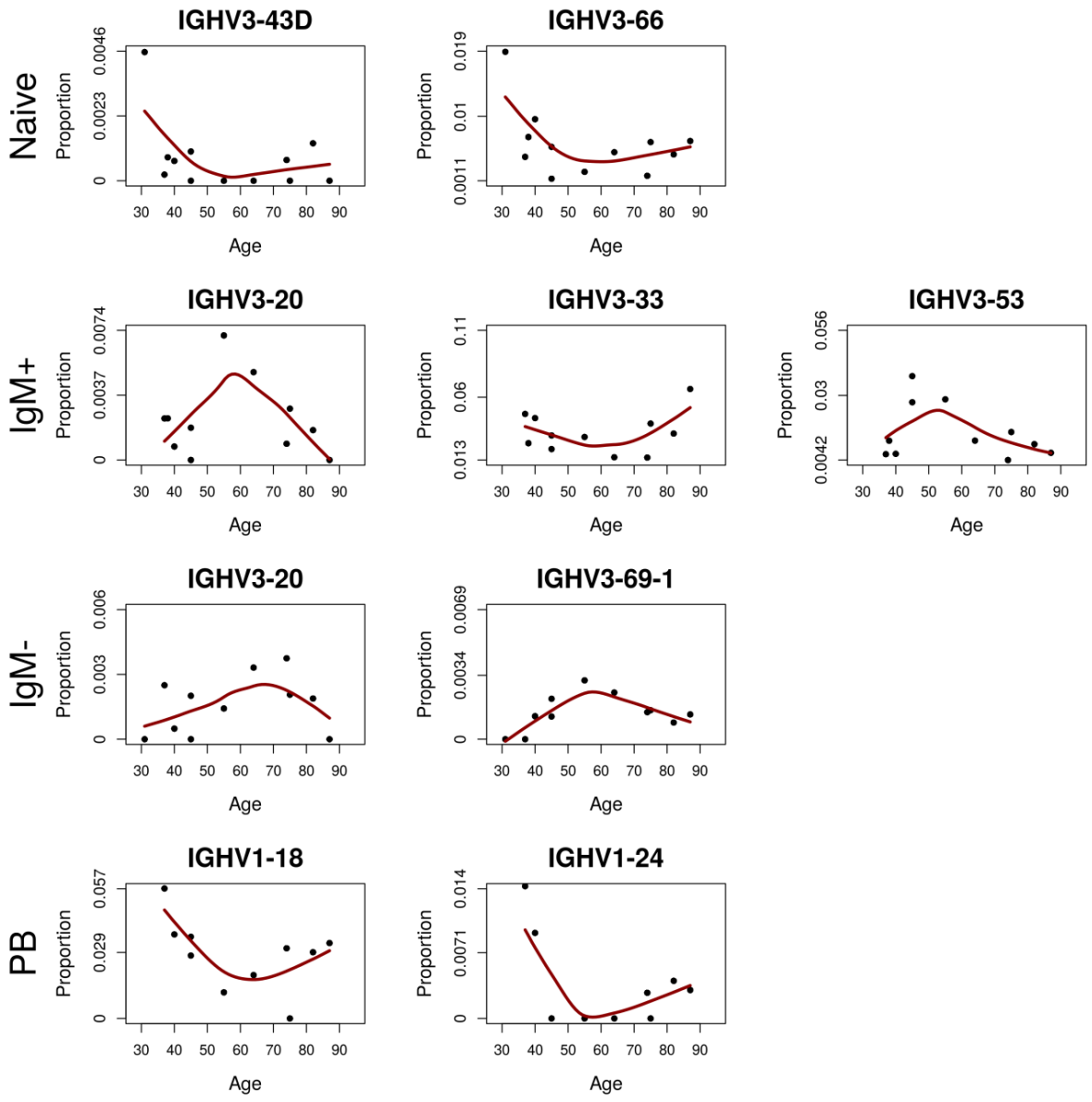

Figure 4.31: Significant quadratic associations ($p<0.05$) of gene usage and age in healthy females. Age is shown on the x-axis, relative gene abundances on the y-axis.

Due to so many differences in gene usage between young and old individuals, but also between females and males, I wondered if there are some genes that may be characteristic for each of those groups. Therefore I used a random forest machine learning approach. First a B cell subset specific analyses on VH and DH gene usages was performed. Due to small sample sizes accuracy and kappa values were not high enough to get informative results. Afterwards all subsets were analyzed together, to increase the sample size. For DH genes the important parameters were still not high enough, which could be due to lacking sample size or due to little importance of DH genes in differentiating between gender or age. But for VH gene usage an accuracy of 0.8 and kappa value of 0.73 could be reached. Therefore a training set of 80% of sample size (60/76 individuals) were randomly chosen for a cross-validated analysis (10 fold, repeated 5 times) to predict important V

genes for the four subclasses (young females, old females, young males and old males).

In the control set (16/76 individuals) the random forest predictions were tested for accuracy. Four of 16 individuals were mismatched. Two were predicted as being young females, but those were old females; two were predicted as being young females, but were young males. Accuracy was at 0.71, kappa at 0.6 and no informative rate of 0.29. The p value for having higher accuracy than no information rate was significant with p=0.0011.

One of the most interesting parts of this analysis is which V genes could serve as predictors for age and sex? The importance parameter helps to analyze which V genes can be distinguished best between the four groups (Fig. 4.32). IGHV2-70 in females (young: 93%, old: 97%) and IGHV2-70D in males (young: 68%, old: 67%) are among the most important genes for all four classes. IGHV2-26 and IGHV2-5 are very important for young females and young vs. old males (IGHV2-26: young females: 85%, young males: 69%, old males: 75%; IGHV2-5: young females: 67%, young males: 79%, old males: 65%). IGHV1-NL1 is very important for young and old females and young males (young females: 77%, old females: 100%, young males: 72%). IGHV4-59 (60%) is important for young females, whereas IGHV3-64 (67%), IGHV3-66 (69%) and IGHV4-39 (75%) are important for old females. However IGHV3-33 (62%) in young males and IGHV5-10-1 (75%) in old males have also high importance values.

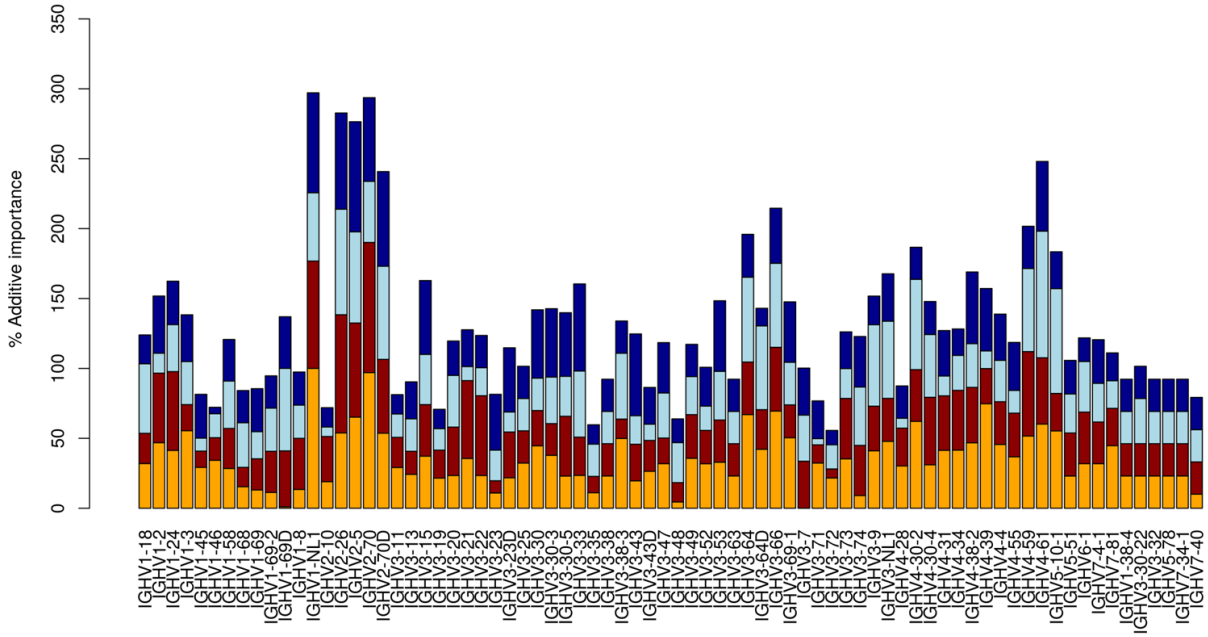

Figure 4.32: Importance values of genes used for random forest analysis. Genes are represented on x-axis; additive percentage of importance is represented on y-axis. The four age and gender specific groups are color-coded: orange (young females), red (old females), light blue (young males), dark blue (old males).

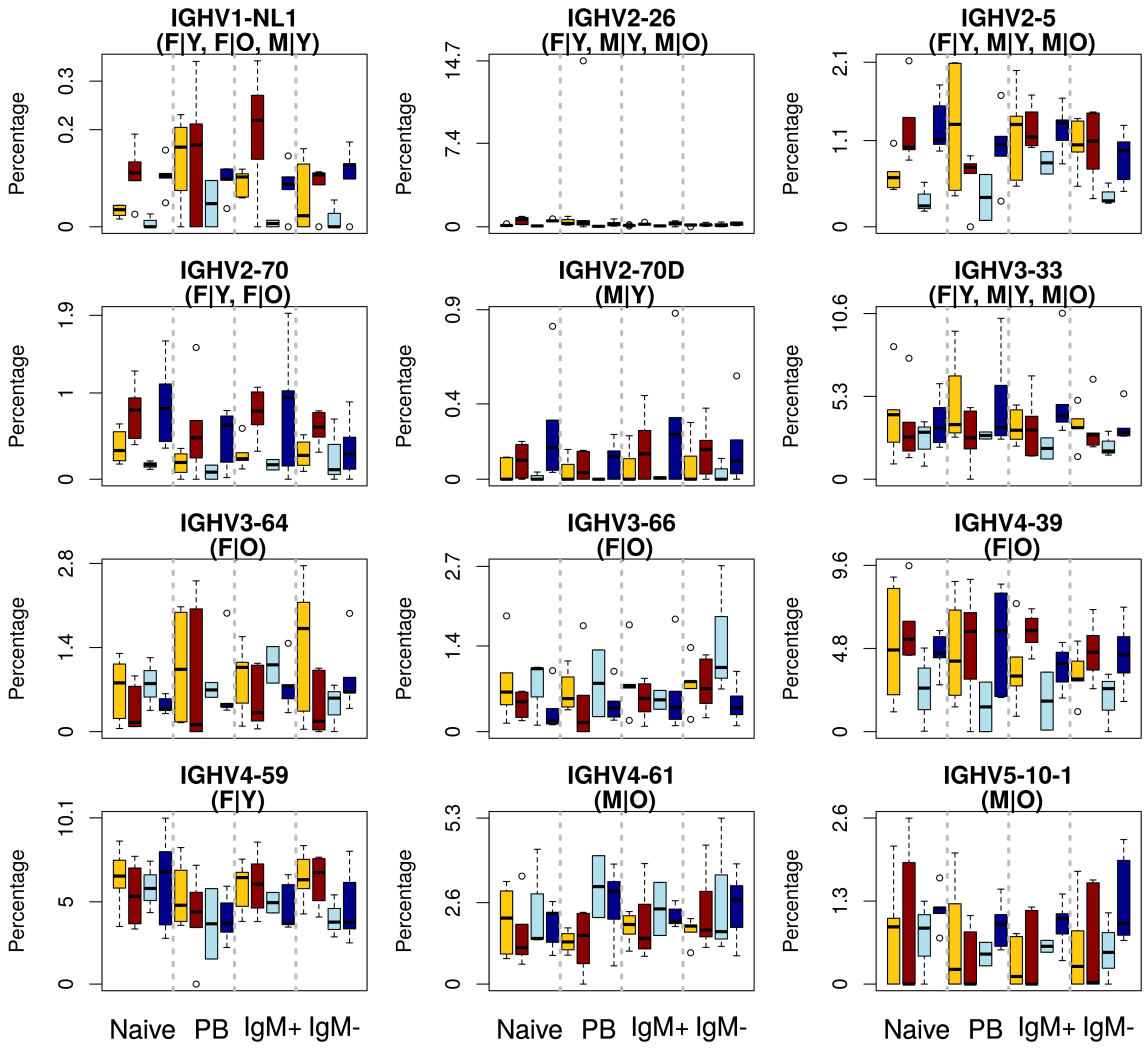

Figure 4.33: Gene proportions of genes with highest importance values. Groups are listed on x-axis, percentages of genes are represented on y-axis. Age and gender groups are color-coded (F|Y = young females: orange; F|O = old females: red; M|Y = young males: light blue; M|O = old males: dark blue).

Further this information was used to see how these genes could be used to predict age or sex. In Fig. 4.33 you can see, how these genes are distributed in the four given classes.

VH genes that help to distinguish between all four groups:

- IGHV1-NL1 is significantly more abundant in young individuals compared to old ones, considering all B cell subsets (females: p=0.003, males: p=0.0005). In young samples females show higher IGHV1-NL1 levels than males (p=0.034). Analyzing subset individually, IgM- memory cells of young individuals are significantly more abundant than those of old ones (females: p=0.041, males: p=0.036) no significant differences for other B cell subsets). In males variability within the groups is lower than in females. Abundance is decreasing from young females to young males to old females to old males.

VH genes that help to distinguish females from males:

- IGHV3-64 is more abundant, but values are also more variable in old females compared to old males, combining all B cell subsets (p=0.022).
- IGHV4-59 has same tendencies than IGHV3-64, but gene levels in plasmablasts are very variable in all for classes. In special, females have significantly higher IGHV4-59 levels than males, combining all B cell subsets (young: p=0.019, old: p=0.004).
- IGHV4-61 is significantly more abundant in males, compared to females, independent of age and B cell subset (young: p=0.031, old: p=0.018). In naÃŕve cells values are highly variable.

VH genes that help to distinguish between young and old individuals:

- IGHV2-26 is significantly more abundant in young individuals, compared to old ones, independent of gender, combining all B cell subsets (females: p=0.0006, males: p=0.0002). Analyzing subset wise, IGHV2-26 is significantly more abundant in young females, compared to old females in both IgM memory subsets (IgM+: p=0.017, IgM-: p=0.0087). Further gene levels of naÃŕve and IgM- memory cells in young males show higher levels compared to old males (naÃŕve/IgM-: p=0.036). Gene levels of IgM- memory cells and plasmablasts are higher than for naÃŕve and IgM+ memory cells.
- IGHV2-5 is significantly more abundant in young individuals than in old ones (females: p=0.0061, males: p= 7.99*10-7) and more abundant in young males compared to young females (p= 0.024), independent of the B cell subset. B cell subset specific, IGHV2-5 levels of naÃŕve cells show higher proportions in young males compared to old males (p=0.036). Further gene proportions in IgM- cells are higher in young individuals, than in old ones, independent of sex (p=0.0022).
- IGHV2-70 and IGHV2-70D are significantly more abundant in young individuals compared to old ones, independent of sex, combining all B cell subsets (IGHV2-70: females: p= 5.65*10-7, males: p=0.0052; IGHV2-70D: females: p=0.00012, males: p=0.0012). IGHV2-70 gene levels in naÃŕve (p= 0.0043) and both IgM memory cell subsets (IgM+: p=0.004, IgM-: p=0.015) are significantly higher in young females, compared to old females. In old males it appears in no or only very few amounts.
- IGHV4-39 has significantly higher gene levels in young individuals than in old ones, independent of gender, combining all B cell subsets (females: p=0.018, males: p=0.0053). Gene levels of young females are also significantly higher than in young males, summarizing all B cell subsets (p=0.022). Highest levels are in young females, followed by old females. Young males have similar levels than old females, old males lower ones.

VH genes that could be markers for only one of the four classes:

- IGHV3-33 shows similar abundance in young and old females and young males. Old males contain significantly lower genes levels than young males (p=0.0038), as well as they show lower gene proportions than old females (p=0.023), independent of the B cell subset.

- IGHV3-66 has higher levels in naÃŕve and IgM- memory cells of old females, compared to all other groups. However in all groups gene levels are very variable.

- IGHV5-10-1 is very variable in females, but has lower variance in males. However this gene appears significantly more often in young males than old ones (p=0.013). Also old females show significantly higher gene proportions than old males (p=0.0018).

4.4.4 Diversity analyses

One very interesting point is, whether there are differences in diversity between females and males or young and old individuals. Therefore I measured the inequality of clone sizes (Gini index) and the amino acid diversity of CDR3 sequences.

Considering clone size distributions, significant differences could be found only in naïve B cells: in females (p=0.024), as well as in males (p=0.036), young individuals have significantly lower Gini indices than old ones (Fig. 4.34). The same tendency can be seen in plasmablasts. In both IgM memory subsets, generally old males had higher Gini indices than young ones, whereas both female groups contained very similar indices. This indicates that old males may be more dominated by large clones than young ones.

Studying the amino acid diversity of CDR3 sequences, only few significant differences could be found (Fig. 4.35). CDR3 sequences of naïve cells are more diverse than the one of IgM memory cells or plasmablasts. IgM+ and IgM- memory cell show similar diversity indices, whereas plasmablasts tend to be less diverse. There was a trend that CDR3 sequences of old individuals are more diverse than the ones of young people, but only for IgM+ memory cells and plasmablasts significance could be reached (IgM+: female: p=0.037, male: $p=6.5 * 10^{-5}$; PB: female: p=0.038, male: n.s.).

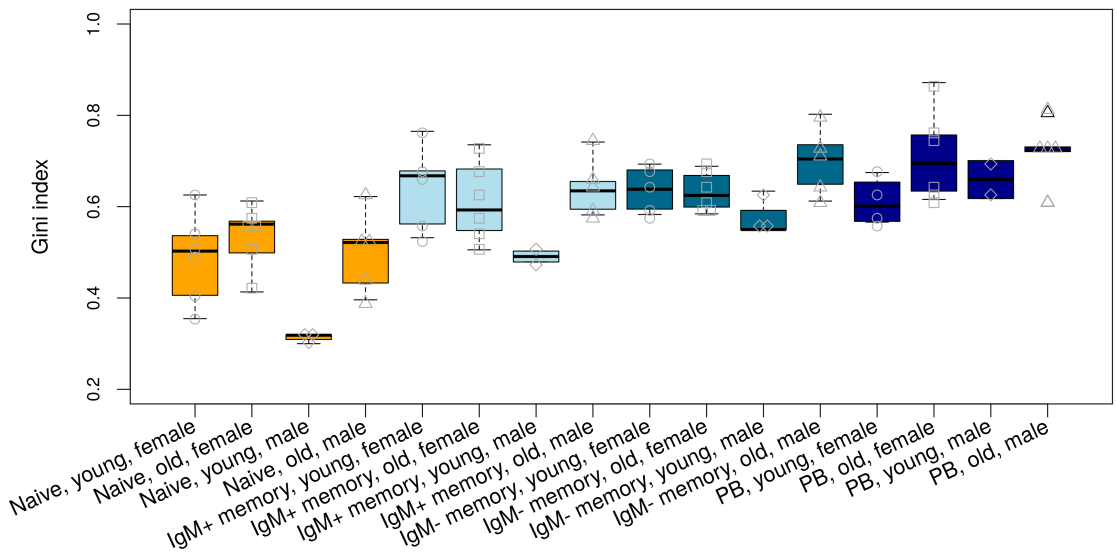

Figure 4.34: Gini index of healthy controls. Groups are listed on x-axis, Gini index is represented on y-axis. The higher the Gini index, the more the sample is dominated by large clones. Significant differences could be found only in naïve B cells: in females and males males, young individuals have significantly lower Gini indices than old ones.

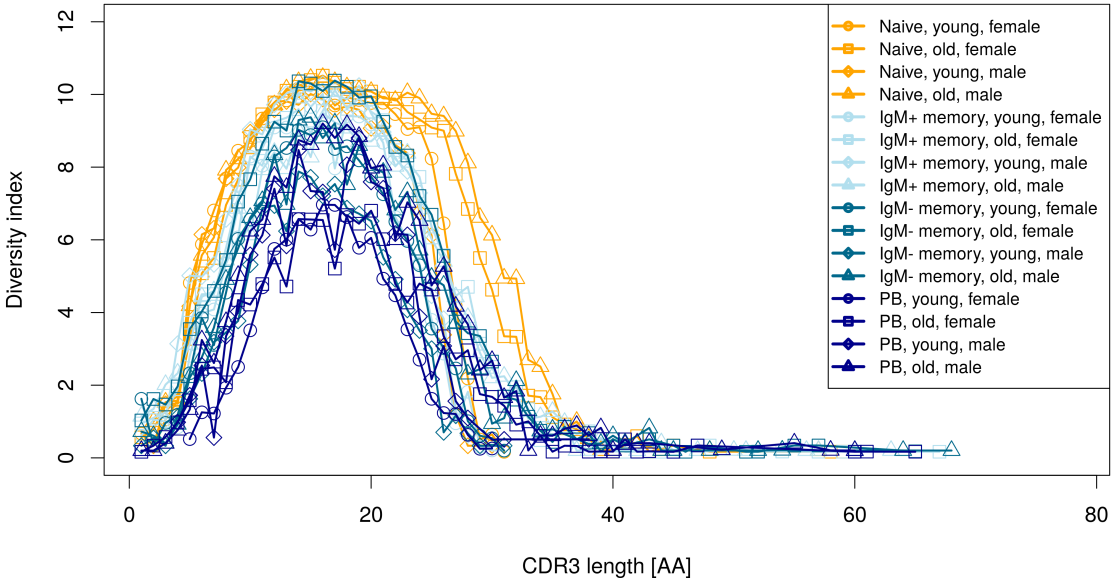

Figure 4.35: Average true diversity of CDR3 amino acid sequences of clones. Diversity was calculated position wise for sequences of the same length and for each group averages were taken. CDR3 sequence lengths are shown on x-axis; diversity indices (true diversity of order 1) are represented on y-axis. B cell subsets are color-coded. Age and gender specific groups can be distinguished by different symbols. CDR3 sequences in clones of IgM+ memory cells are significantly more diverse in the elderly, compared to young subjects. In plasmablasts only old females show significantly higher diversity indices than young ones.

5 Discussion

In this study, I determined the B cell receptor repertoire from 10 PV, 10 BP patients and 20 healthy control subjects (10 paired with PV and 10 with BP). From these study subjects, four subpopulations reflecting different stages of B cell development and adaptation were separately analyzed. The naïve repertoire should be mainly determined by the germline repertoire that is generated during primary B cell maturation in the bone marrow. It is likely to be affected by immunomodulatory treatments, particularly if those treatments comprise primary bone marrow output. The IgM positive and even more the IgM negative (which presumably express IgA or IgG or more rarely IgE) memory B cell pools contain additional information about what the immune system has responded to and been selected for over the lifetime of the individual. Finally, the plasmablasts potentially reflect an ongoing immune response.

5.1 Technical aspects

Limitations of the whole study comprise not only cell sorting, but also the sample size. Human peripheral mononuclear blood cells were taken from PV and BP patients, as well as healthy controls. Blood samples are not necessarily the best representation of lymphocyte repertoire, since cells that are present in blood are usually in transition from one to another tissue. It is unknown to what degree pathologic B cells in PV or BP are present in the circulation. While some autoantibody-producing B cell clones have been isolated, the isolation techniques typically rely on multiple rounds of selection or panning for enrichment. Although the aim in these studies was to analyze the B cell repertoire as a whole, rather than the autoantibody repertoire specifically, the argument can be made that the best place to look for the mechanism of autoimmunity is the autoimmune-prone B cells themselves. Whether selection by antigen is sufficient for recovering such cells from the blood remains an open question. Alternative approaches (sorting by homing receptor expression or sequencing IgE antibodies from RNA) or sample types (from skin), are worth considering in this regard. Further analyzing naïve cells and plasmablasts are good to get insights into the basic repertoire and spontaneous immune response. But since PV and BP are known to be IgG associated diseases, it is difficult to differentiate results regarding IgM- memory cells, into specific IgG, IgA or IgE associated subsets.

All the results I found are limited to a single chain (heavy chain). So the data cannot

provide information about the identity of immune receptor pairs encoded by individual B lymphocytes. To reconstitute BCRs for functional analysis, therapeutic use or modeling of receptor-antigen binding, the heavy and light chains from a complete BCR must be identified as a pair.

Another important aspect is, that we used DNA and not RNA samples. Bashford-Rogers et al. discussed several B cell receptor sequencing methods in [**?**]. They stated that due to BCR allele defective-rearrangements the number of clones in the data may be increased, whereas unequal numbers of RNA molecules per cell may skew the BCR repertoires derived from RNA [**?**]. We thought that the DNA may better represent the current disease situation of the patient and that information about productive and unproductive genes may help in understanding the diseases.

Further the whole process of DNA amplification and next generation sequencing could be slightly improved. First, it is already known that there are errors occurring during DNA amplification using PCR. If the error occurred during IgH PCR, there could be significant amplification of polymerase errors rather than bona fide somatic hypermutations. Further, the Illumina based sequencing method generates sequencing errors at an estimated frequency of 1%. To minimize the contributions of sequencing errors to our analysis, we removed single copy sequences from the analysis. However, given the high depth of sequencing for some of the samples it is highly likely that this cut-off alone is not sufficiently stringent. Sequencing errors can be narrowed by several quality control steps. For example, sequencing errors can be reduced by choosing an appropriate cut-off for the Phred quality score when filtering Fastq files. I chose 20 as a threshold for Phred score, as it corresponds to a 99% probability of a correctly identified base. To identify mutations correctly a higher Phred score may be chosen, which eliminates almost all sequencing errors and thus mutations can be identified as real mutations (for example a Phred score between 30 and 40 would represent a probability of 99.9% to almost 100% of a correct identified base). However, a more stringent quality control leads also to a smaller number of sequences which can be further analyzed. Due to possible PCR and sequencing errors, I did statistical analysis not on sequence level, but on clone level. Here the focus is not on the sequence itself, but on similar sequences, grouping together. Clones were assigned into sequences by sticking to following criteria: 1) same VH and JH genes and 2) same CDR3 amino acid sequence lengths with at least 85% of sequence identity (like recommended in [**?**]).

Another problem is the great variability of sequence quantities used for statistical analysis. We tried to get similar cell numbers for all samples. Only for plasmablasts smaller cell numbers, compared to the other subsets, were provided. Due to this subsampling step, errors may occur and the DNA used for analysis may not represent the total B cell repertoire. Although at least similar cell numbers were used, sequence numbers, and thus the basis for collapsing clones, varies between 6,000 and 500,000 sequences. In particu-

lar, for plasmablasts only few sequences were available. But by taking a close look at each individual after all the analyses, outliers could be detected and results interpreted independently. Samples with only few sequences or a number more than average did not appear as outliers in most cases.

In total I analyzed four different B cell subsets of 10 PV patients, 10 BP patients and 20 healthy controls. BP Patients and the corresponding controls were age and sex matched, whereas the PV dataset contained more variability between patients and controls. The healthy controls of the PV study were a little bit younger (PV: 50 ± 19, controls: 42 ± 7) and included more females (70%), than the PV patients (20%). But further analyses could not confirm gender or age specific biases in IgH diversity, which strongly influence results (data not shown).

Moreover despite the extensive analysis per subject, the numbers of subjects in each disease, age and sex category are still quite small. However, I analyzed two different types of autoimmune bullous diseases, both of which are very rare. Recruitment of patients and age and sex matched controls, as well as getting their agreement to participate on a study like this was very time-consuming. To get more power and reliable results, sample size needs to be increased and results need to be validated by replication experiments. For instance, applying a Wilcoxon-Mann-Whitney test to study differences between two groups, requires at least 27 to 47 individuals per group to reach a power of 80% and seeing middle- or large-scale differences (α= 0.05, Cohan's d = 0.6 or 0.8; calculated by G*Power [**?**]).

Last, but not least, the results are dependent on the current status of the IMGT/HighV-QUEST database. It could be that some of the genes are not identificable due to missing or error-prone annotations in the database. Further genome builds are updated regularly and thus information about gene positions and functionality may change.

5.2 R package *bcRep*

The R package, *bcRep*, can help to analyze and visualize large amounts of NGS-based B cell receptor data. It contains over 40 functions to analyze sequences and clones with respect to mutations, diversity, gene usage and others. IMGT/HighV-QUEST output, but also self-written files can be used as input. It is written in a user friendly manner, includes many help pages and is available on Windows, Linux and Mac OS X. There are several other tools which also help to analyze similar types of data, but in most cases they are limited to a certain type of analysis or to limited numbers of input. *bcRep* comprises functions analyzing many different repertoire features and offers parallel processing for a faster analysis of large data sets. Further analyses can also be performed at individual sample or between-subject levels.

However, there are still some limitations that need to be mentioned. For instance, due

to the script language (R), some functions may need more computational time than the same methods written in compiler based programming languages, like JAVA or C. But one needs also to consider that R scripts are easily accessible and thus replicable, and intermediate results can be easily further processed.

There are some algorithms that are also used by other tools, but are programmed slightly differently and so result in different outputs. These differences can make it difficult to compare run time and output of several tools. There is no clear wrong or right answer regarding some of these issues. For instance almost every tool uses different criteria to cluster sequences into clones. Most tools assume the same V and J gene assignments, but have different approaches to cluster the CDR3 sequences, for example. As long as algorithms are defined clearly, the user can decide on his/her own which tool fits best for his/her requirements.

There are also some topics that need further attention in the future. Not only mutations, but also diversity, as well as rarefaction can be analyzed in more detail. The package lacks functions to study if the set of sequences really represents the total set of B cells or if only a too small number is analyzed, where no biological conclusions can be made. This is very important for datasets with only a few thousand sequences.

The R package was published in October 2015 and has been downloaded over 3000 times since then.

5.3 The B cell receptor repertoire of PV patients

The pathogenic effects of autoantibodies in PV have been very thoroughly investigated over the past decades. Although the exact mode of action is still a matter of debate, it is well established that these IgG autoantibodies cause blistering in PV [**?** **?**]. In contrast, the origin of autoantibodies and the mechanisms that cause specific autoreactivity against Dsg3 and Dsg1 are still not understood.

Genetic association studies suggest a heritable susceptibility to develop autoimmune diseases, in particular, pemphigus vulgaris [**?** **?**]. Another remarkable property of PV is its variable distribution, in contrast to BP, in different populations. While PV is much more frequent than BP in Kuwait [**?**], the opposite is true for Germany and Singapore [**?** **?**]. This suggests the possibility of genetic selection for PV susceptibility, it may even be speculated that there is a specific advantage that goes along with PV risk. The impact of the genes on PV susceptibility may be distinguished into two kinds: genetic variations that generally increase the risk of autoimmunity and variations that are very specifically increasing the risk for PV. These first of the two kinds is reflected by genetic associations with the presence of a quite frequent FcγRIIc variant [**?**], which is also found in other autoimmune diseases [**?**]. FcγRII is an inhibitory Fc receptor. Knock-out studies in mice show that the absence of FcγRII can result in anti-dsDNA antibody production by

plasma cells (in mice with an anti-DNA knock-in) [**?**]. Whether PV or BP patients harbor similar defects in FcγRII or other immunomodulatory or regulatory factors is unknown. The second kind of genetic defect may explain why latent autoantibodies against Dsg3 and Dsg1 are found in first-degree relatives in family studies [**?**]. A plausible candidate for genetic variations of the second kind may be the germline repertoire of antibodies encoded by different VH gene alleles [**?** **?**]. These may provide a more or less optimal template for the development of autoantibodies, and it is very likely that PV might be associated with certain VH gene alleles. Another possibility is that the defect is not in how the immune system responds to the antigen but rather to how the antigen may be delivered to the immune system.

After collapsing sequences into clones, the largest number of clones could be found in naïve cells, the fewest in plasmablasts and numbers in between in both IgM memory sets. The main reason for this difference is that far more naive B cells were sampled than memory B cells. What was far more interesting was that among IgM- cells and plasmablasts controls had significantly more clones than PV patients. This difference did not correlate to the frequencies of the cell subsets, which were similar in both patients and controls.

Clone sizes were estimated using total copies. It should be noted that this measure is inherently flawed because there was no controlling for the depth of sequencing in different samples. The basic observation (that higher copy number clones are present in the more mature B cell pools and in the PBs in particular) is entirely consistent with the possibility that these samples, consisting of smaller cells, are sequenced at greater depth (leading to an artifactual increase in copy numbers per clone). The larger clone sizes in the memory subsets of PV patients could also be due to differences in sequencing depth if fewer B cells were sampled from patients. However, my results suggest that controls have more clones, but therefore smaller ones, tending to a more diverse repertoire, which could be confirmed by further diversity analyses. However, there could be some kind of clonal selection in the patients, but also factors that may suppress development of special clones and lead to a narrowed repertoire. The higher diversity in plasmablast sequences of PV patients could indicate a relaxed control of autoreactive B cells.

The immune repertoire data sets allow inferring the VH gene that make up the germline repertoire of each individual in our data set. Furthermore, it is possible to describe the usage of individual genes. It must be noted, however, that due to the high homology between certain VH genes differentiation may be very difficult. Moreover, the efficiency of VH gene discrimination is affected by sample sizes. Therefore, it is important to note that NGS samples from PV cases and controls have slightly, but not significantly different sequence counts. I found 14 genes to be significantly differentially expressed between PV patients and controls. Most differences were seen in the plasmablast repertoire, which may be too undersampled to be robust. Almost all significantly differentially expressed

VH genes are more abundant in controls, than cases, except IGHV3-13 in IgM+ cells. This observation, coupled with the fact that nearly all of the differentially expressed VH genes are low in abundance, suggests that most of the basis for this observation is that there are fewer clones in the PV memory pools. Thus VH genes that are more common in controls may not reflect true differences between PV and controls. The more interesting examples are of genes that are expressed more frequently in PV (in spite of the potential differences in sampling between PV and controls).

It has been proposed that autoantibodies in patients with PV exhibit skewed VH gene usage. For example, in some PV patients anti-desmoglein 3 antibodies appear to preferentially utilize VH1-46 H chains [**?**]. Because in principal many different VH genes can be used to mount an immune response, it is hard to understand why such a skewing exists specifically for anti-desmoglein 3. There are several possible explanations for VH1-46 skewing in PV: 1) desmoglein 3 has specific properties that severely limit the VH genes that can respond to it (the alternative is that VH1-46 is one of many different H chains that can respond to desmoglein 3 and its relative usage is patient-specific); 2) VH1-46 is actually generated in response to a different antigen and happens, in some unfortunate individuals, to cross react with desmoglein 3 [**?**]; 3) VH1-46 is normally autoreactive but its selection is globally altered in PV due to a tolerance checkpoint defect. To distinguish between these alternatives, I analyzed VH usage in different B cell subsets from PV patients and controls. Given the potential importance that has been described to VH1-46 in the pemphigus literature, I was disappointed with our failure to find significant differences in global VH1-46 usage, subset-specific VH usage in PV patients vs. controls. In our analysis, IGHV1-46 had marginally higher gene levels in IgM- memory cells of controls compared to PV patients; this is the opposite of what one would predict if VH1-46 clones are activated in PV. Cho and colleagues performed site-directed mutagenesis of VH1-46 autoantibodies which suggests that acidic amino-acid residues introduced by somatic mutation or heavy chain VDJ recombination were necessary and sufficient for Dsg3 binding [**? ?**]. In our dataset IGHV1-46 could be detected in only low amounts, in both groups among the four analyzed B cell subsets. In all analyzed subsets it has slightly higher levels in controls, compared to patients. In fact, VH1-46 could be detected in only 75% of IgM- cells and 50% of plasmablasts in patients, compared to 87.5% (IgM-) and 100% (PB) in controls. Taken together, these data suggest that global and subset-specific regulation of VH1-46 is not altered. Of course it is possible that the autoantibody repertoire comprises such a small fraction of the overall B cell repertoire that I have not been able to detect the difference. However, it is also worth pointing out that differences seen in VH1-46 amongst anti-dsg3 antibodies were only observed in some of PV patients and may not be generalizable to all PV patients. The failure to identify a dominant VH in the immune response to a complex antigen is, in and of itself, not unexpected. On the one hand it could underlie different selection processes, which change the ability of generating

autoreactive genes and elimination of those. On the other hand, these differences could appear just due to different genetic backgrounds and selection mechanisms in Americans (Cho et al. [**?** **?**]) and Europeans (our dataset).

Most of the genes that are significantly different expressed between PV patients and controls are functional. Only two of them are pseudogenes: IGHV1-NL1 and IGH3-47. Pseudogenes are known to be DNA segments that are built like all other genes, but they are not translated into functional proteins. Instead they are unproductive due to different reasons. However, there are some studies showing that pseudogenes could have a biological function [**?**]. Pink et al. summarized several papers discussing this problem: pseudogenes may regulate coding genes or tumor suppressors by acting as microRNA decoys. Further they are often deregulated during cancer progression [**?**]. Both genes have lower levels in PV patients compared to healthy controls, which suggests that these pseudogenes may play a protective role. They could influence disease initiation, but also passively by influencing rearrangement and regulation of other genes.

Another notable observation is that some of the VH genes appear to have copy number variation. Unfortunately, our NGS approach does not allow correct determination of the copy numbers of these genes, so I did not further analyze this in the context of a possible association with PV. Further it might be that specific VH genes that are overrepresented in one of both groups, provide less important information than the CDR3 sequences and clones belonging to these genes.

Further I was interested in differences regarding mutation frequencies and patterns. Although there are no significant differences between groups in the average V gene sequence identity in any of the analyzed B cell subsets, it is noteworthy that for all subsets, except for IgM+ memory cells the ratios of the average number of mutations per sequence of patients versus controls are above one, indicating more mutations in cases than controls (for analyzed sequence parts: V, CDR1/2, FR1/2/3). For IgM+ cells they are less or almost equal to one. So even if the number of mutations is quite similar, PV patients show in average slightly higher numbers of mutations per sequence. Further there were few nucleotide substitution patterns amongst mutated bases, which are more abundant in one of both groups. In naïve and IgM+ memory cells mutations from adenine to cytosine and in plasmablasts from thymine to adenine appear more often in PV patients than controls. In contrast, controls have more mutations from thymine to guanine in IgM+ memory cells and from cytosine to adenine in PB. All these mutations are transversions, which occur usually less frequently than transitions. Moreover mutations to guanine or adenine could lead to a translation of stop codons and further to the development of unproductive sequences. However the percentages of unproductive sequences are not different in our dataset, suggesting that negative selection may help to elude the production of unproductive sequences, but maybe using different mechanisms in PV patients and controls.

In summary, no differences in the naïve repertoire, indicate that there is no general problem with the BCR repertoire, causing susceptibility to autoimmunity. These results provide no compelling evidence for a single, dominant tolerance checkpoint defect in PV patients. But it is possible that PV patients nevertheless have tolerance defects that are either masked from detection based upon their low frequency, their heterogeneity, their inter-individual heterogeneity or by time. All of these patients have established disease. It is possible that the patients I should have studied need to be during initial disease onset. Such patients would presumably have BCR repertoires that are less likely to be confounded by therapy and the effects of chronic disease. The breakdown of self-tolerance in PV may, furthermore, be a complex process that requires not only that a maladaptive immune response occurs, but that this response is targeted to specific self-antigens. The presentation of these antigens, in a sufficiently pro-inflammatory, immunogenic context, may be critical for disease initiation. A general hypothesis is that individuals may have a different susceptibility to develop autoimmune diseases; that they have an autoimmune-prone immune system. Keeping an optimal trade-off between a sufficient immune response and the risk of developing autoimmunity is a difficult task, given the continuous competition with infectious organisms. Since I see much more differences in plasmablasts, than both IgM memory sets, I can neither confirm nor reject that PV is an IgG driven disease. But it seems that differences in the B cell receptor repertoire are mainly reflected in the ongoing immune response, instead of the memory pool.

5.4 The B cell receptor repertoire of BP patients

Bullous pemphigoid is characterized by IgG autoantibodies against BP180 and BP230, as well as an inflammatory infiltrate including eosinophils and neutrophils. The critical epitopes necessary for autoantibody mediated disease induction are harbored within the NC16A region of BP180. Zuo et al. showed that the IgG4 anti-NC16A blocks IgG1 and IgG3 induced complement fixation, neutrophil infiltration and blister formation [**?**].

BP patients and healthy controls show similar numbers in clone quantity, size and diversity. Only for IgM+ memory cells and plasmablasts do I see larger clones in controls than in BP patients. This finding suggests that BP patients are generally not different from healthy controls with respect to clonal diversity in the different B cell pools. However there are differences in mutation patterns and VH gene usage.

While the number of mutations in both groups is not different, the kinds of mutations are. Mutations that are more frequent in BP patients than controls are mainly transversions (a $\rightarrow$ c, t $\rightarrow$ g or a, g $\rightarrow$ c or t, c $\rightarrow$ g or a). Whereas transitions are more abundant in controls (t $\rightarrow$ c, g $\rightarrow$ a). In healthy humans transversions are less frequent than transitions due to more complicated structure changes. Transitions result less frequently in amino acid substitutions and are therefore more likely to persist as silent

mutations. This could mean, that there are more mutations in BP patients, changing the amino acid code of the antigen binding sites (complementary determining regions) and resulting in an altered antigen binding affinity. Alexandrov et al. analyzed patterns of mutation signatures in cancer diseases [**?**]. They found tranversions, like t → g or c → a, being overrepresented in patients suffering from lung and oesophageal cancer, but also chronic lymphocytic leukemia (CLL) [**?**]. CLL is a monoclonal disorder characterized by a progressive accumulation of functionally incompetent lymphocytes. It is known to often co-occur with bullous pemphigoid or other autoimmune diseases [**?** **?** **?**] and such those genes may play an important role in disease triggering. Thus there might be some mutation pattern, which preferentially influence development of CLL and BP.

Considering VH gene usage, most differences can be seen in both IgM memory subsets (IgM- > IgM+) and plasmablasts. Genes like IGHV1-69D, IGHV3-30, IGHV3-53 and IGHV4-30-4 are, in almost all analyzed B cell subsets, overrepresented in controls and may play a protective role. However, genes like IGHV2-5, IGHV2-70 and GHV6-1 are more abundant in BP patients. In particular, IGHV2-5 has higher levels in BP patients than controls, in all analyzed B cell subsets, except naïve cells. Stanganelli et al. analyzed VH gene usage and mutations of Argentinian patients suffering from chronic lymphocytic leukemia [**?**]. This group found highest abundances for IGHV3-23, IGHV1-69 and IGHV2-5 and that those genes are associated with CLL.

IGHV2-70 interacts with several other genes in pathways, like *creation of C4 and C2 activators and Interleukin-3, -5* and *GM-CSF (granulocyte macrophage colony-stimulating factor) signaling* and may thus play a role in activating the complement system [**?**]. Jordon already spotted that participation of properdin, a positive regulator of complement activation, in addition to early complement components (C1, C4 and C2) suggest local activation of both complement pathways in BP [**?**]. Further Borrego et al. investigated the deposition of eosinophil granule proteins during blister formation in bullous pemphigoid [**?**]. They showed that IL-3 and IL-5 inhibit the enhancement of eosinophil survival by blister fluids [**?**]. GM-CSF may also partially decrease eosinophil survival. Therefore IGHV2-70 may act as a key gene in BP disease activation.

IGHV6-1 is in naïve, but also both IgM memory B cell subsets overrepresented in BP patients than controls. Edwards et al. studied VH gene usage of IgE antibodies from one patient with atopic dermatitis and revealed that sequence analysis of random clones showed a IGHV6-1 usage of 33%, although this gene is rarely used in the expressed adult repertoire [**?**]. They designed a cyclic protein from the CDR3 of some of these clones, which finally bound to self, but also nonself antigens, like human IgG, tetanus toxoid or human and bovine Willebrandt factor [**?**]. This shows that some IgE antibodies may bind more than one antigen, which would have important implications for understanding the multiple sensitivities seen in conditions such as atopic dermatitis, but also other allergic diseases [**?**]. Atopic dermatitis, as well as bullous pemphigoid are acantholytic disorders,

which are based on a loss of cohesion between keratinocytes [**?**]. High levels of IGHV6-1 may be important for such processes and contribute to blister formation. However this gene is the only existing gene of IGHV family 6 and thus may be just an artifact due to sequencing errors or mutations.

Furthermore, IGHV3-48 is only in plasmablasts significantly more abundant in BP patients, compared to controls. Nevertheless, it is known to be associated with CLL or autoimmune diseases, like celiac disease [**?** **?**]. Hojjat-Farsangi et al. analyzed the VH gene usage of leukemic B cells in Asian CLL patients and found among others IGHV3-48, IGHV4-39, and IGHV1-8 to be most abundant and associated with CLL [**?**]. As already mentioned, CLL is a disease that co-occurs with bullous pemphigoid and such IGHV3-48 may contribute to progression or even onset of disease. Snir et al. studied IgA memory and plasma cells in patients with celiac disease and found IGHV3-48, IGHV4-59, IGHV5-10-1, and IGHV5-51 gene segments to be overrepresented in patients [**?**]. Celiac disease is an autoimmune disorder, where ingestion of gluten leads to damage in the small intestine. Further it is known, that dermatitis herpetiformis, which is a bullous skin manifestation of celiac disease, is also a distinct subcategory of the pemphigoid skin diseases [**?**]. This leads to the suggestion that IGHV3-48 and also other genes may be associated with early pemphigoid skin disease triggering.

In contrast IGHV3-30 is less abundant in BP patients, compared to controls, mainly seen in IgM+ memory cells. IGHV3-30 is known to be regulated in autoimmune diseases. Olee et al. analyzed associations between antibodies and the development of the B cell repertoire [**?**]. They concluded that there are several IGHV3-30-like genes in the human VH gene repertoire and that a complete deletion of these genes is relatively restricted to a subset of autoimmune patients, suffering from diseases like rheumatoid arthritis or systemic lupus erythematosus [**?**]. There might be an evidence for deletion of developmentally regulated autoreactive V genes in autoimmune diseases [**?**].

Summarizing, the B cell receptor repertoires of patients suffering from bullous pemphigoid are not globally different than the ones of healthy controls. The naïve repertoire of both groups is quite similar, as well as characteristics on sequence level. However, there are differences in pattern of silent mutations and in gene usage in both IgM memory subsets and plasmablasts. Many genes that are overrepresented in BP patients, are already known to be associated with other autoimmune or skin diseases. BP may not only be driven by IgG specific features, but also by other components of the memory and the spontaneous immune response. Further complement activation and recruitment of inflammatory cells are important for the development of clinical manifestations, such as blister formation [**?**].

5.5 Age and sex dependent changes in the B cell receptor repertoire of healthy controls

When analyzing the age and gender specific changes of the B cell repertoire of healthy controls, more changes during age, than between females and males could be seen. Our results showed that most clones could be found in naïve cells, followed by IgM+, IgM- memory cells and plasmablasts. Old individuals tended to have more memory clones than young individuals. More clones lead to the suggestion that there is more variability within the set of sequences and thus more variability in the immune response. Consequently the immune system of old healthy individuals has adapted over time and became more widespread, which can be beneficial, but may also result in the generation of autoreactive BCR specificities. Further the naïve immune response shows the highest variability within the groups, whereas the immune memory (IgM+/- cells) and the actively responding B cells (plasmablasts) are less variable within the groups. These inter-individual differences may become magnified with age, as the repertoire continues to be molded by environmental and pathological antigen exposures over time.

On the other hand, B cell subsets containing the most clones, showed smallest clone sizes. In this case largest clones were found in plasmablasts, followed by IgM-, IgM+ and naïve cells. Old males had larger clones than young ones, concluding that there might be a clonal selection in the male elderly. Similar findings were reported by Dunn-Walters et al., suggesting an increase of expanded clones that occur with age [**?**]. In naïve cells this could mean, that the rearrangement of genes may underlie different mechanisms during aging. The control of autoreactivity, but also learning processes and recognition of particular pattern are one of the main aspects which are probably changed when becoming older.

Furthermore our data shows that old individuals had clones with much longer CDR3 amino acid sequences than young ones. This could be on the one hand due to more occurrences of mutations, in particular insertions and deletions, during aging because of infections or environmental factors; but also due to a less effective negative selection. Dunn-Walters et al. suggested that older samples show greater CDR3 sizes in naïve cells [**?**], but same trends for IgM+ and IgM- memory cells and plasmablasts could be seen.

Tabibian-Keissar et al. used CDR3 spectratype analysis to demonstrate a loss of diversity in peripheral blood of old individuals, which correlated with poor health and survival [**?**]. I used Gini and true diversity indices to analyze diversity of clone size distributions and CDR3 amino acid sequences. However our data shows on the one hand that in all analyzed B cell subsets old males are more dominated by large clones, than young ones (in naïve cells this can be also seen in females), but with a median Gini index around 0.6 are still diverse. On the other hand also CDR3 sequences of clones, in particular among the IgM+ memory cells and plasmablasts, in women are more diverse in the elderly, whereas

no differences can be seen in the naive cells. It is known that women produce more elevated circulating levels of antibodies than men [**?**]. Their tendency to produce higher levels of autoantibodies may contribute to autoimmunity in females.

Considering gene usage, again more differences within age groups, than gender could be found. Regarding differences between groups (beta diversity), the most significant results were observed in naïve B cells. One could assume that fundamental differences between young and old individuals can be found already in the basic germline repertoire. Differences in gene usage between females and males are less abundant than between young and old individuals, but appear almost equally distributed in all analyzed B cell subsets. Comparing VH gene family distributions of our list of significant differences within age with the IMGT database, genes of the VH family 2 seem to be overrepresented and genes of family 4 underrepresented in our list (IGHV2: 7% of all genes to IMGT database; 19% of significant results regarding age; IGHV4: 24% of all genes in IMGT database; 11% of significant results regarding age). In the significant list of differences within gender only VH family 2 seems to be overrepresented (7% of all genes in IMGT database; 15% of significant results regarding gender). Wang et al. used Ig VH cDNA libraries to analyze immune response changes, in particular in IgM and IgG cells, in three young and five old individuals [**?**]. They found IGHV3-23 and IGHV1-2 to be the most abundant in both groups [**?**]. Further their data showed that IGHV4-34, IGHV4-59 and IGHV1-69 were expressed at much higher levels in older subjects, compared to young ones [**?**]. In our dataset IGHV3-23 (average = 12%) is also one of the most abundant genes in all groups and subsets, followed by IGHV3-30 (average = 9.5%). IGHV1-2 has an average abundance of 3.5%, but is still in the top 10 most abundant genes. IGHV4-39 is also significantly more abundant in our dataset in naïve cells of old individuals, compared to young ones. Autoantibodies encoded by this gene are associated with cold agglutinin disease [**?**]. This disease occurs primarily in old individuals or after infections. In our dataset it occurs slightly more in elderly compared to young subjects; greater differences can be seen in females, than in males. IGHV1-69 appears not to be significantly differentially expressed, but therefore IGHV1-69D has higher gene levels in naïve cells of old individuals than of young ones. Polymorphisms in IGHV1-69 are known to affect susceptibility to rheumatoid arthritis [**?** **?**], which is also a disease emerging in elderly. IGHV4-59 is not significantly different amongst the different age groups, but IgM- memory cells of females show significantly higher levels than do males. This gene is also known to be associated with rheumatoid arthritis, which occurs more often in females, than males and may play an important role in disease triggering in females [**?**].

Further there were some genes in females showing quadratic associations with age. In particular, there were five genes showing a decrease in gene proportions until the age at around 55 and afterwards an increase (naïve: IGHV3-43D, IGHV3-66; IgM+: IGHV3-33; PB: IGHV1-18, IGHV1-24). On the other hand I found four genes with increasing

abundance until the age of 55 years and afterwards decreasing levels (IgM+: IGHV3-20, IGHV3-53; IgM-: IGHV3-20, IGHV3-69-1). This led us suggest, that at least parts of the B cell repertoire might be influenced by hormones. There are already a few studies assuming that estrogen, but also androgens control the BCR repertoire [**?** **?** **?**].

During childbearing years females experience cyclical elevations of estrogen during the phase of menstrual cycle. In the years prior menopause levels are dramatically increased and after menopause occurring in only very low amounts [**?**]. The average age of menopause is around 50-55 years [**?**]. This progress perfectly fits with our gene expression curves of IGHV3-20 in both IgM memory subsets, IGHV3-53 (IgM+) and IGHV3-69-1 (IgM-) which all belong to the same VH family 3 and seem to positively correlate with estrogen concentration course. On the other hand there are three genes of VH family 3 and two of VH family 1, which show an inverse progression and may be suppressed by estrogen. Among other tasks, estrogen promotes extra-medullary hematopoiesis and T cell lymphopoiesis in the liver, which may contribute to increased escape from negative selection in B and T cells and thus increasing risk of autoimmunity [**?**]. Low concentrations tend to promote Th1 cell proliferation and IFN-γ production; while high concentrations increase Th2 cell production of IL-4, IL-5 and IL-10 and enhance regulatory T cell function, which results in the proliferation of B cells and their maturation into plasma cells [**?** **?**]. Thus estrogen can act as a double-edged sword; in low concentrations it helps to strengthen antibody secretion and acts relatively inflammatory, while in higher concentrations may have more anti-inflammatory effects.

The concentration of androgens in females decreases slightly during aging, but there is no dramatic drop during menopause. It is known that androgens can reduce antibody production and play an important role in B cell homeostasis and tolerance [**?**]. For example, androgen can enhance a Th1 response, specifically leading to increased IL-2 production and activation of CD8+ T cells [**?**].

I further used VH gene usage data of all B cell subsets as input for random forest analysis to get information about important genes, independently from all other analysis. In this analysis many genes could be confirmed to be meaningful for the differentiation of gender or age. Again genes like IGHV3-33, IGHV3-66, IGHV4-39 and IGHV4-39 showed high importance values (60-75%). But also IGHV1-NL1, IGHV3-64, IGHV2-26, IGHV2-5 and some others could help to discriminate between young and old females and males.

Another point I was interested in, were different mutation pattern in females and males or young and old adults. Overall it seems that females and males are quite similar with respect to overall SHM frequencies, but there were differences within age. More SHMs tended to occur in the elderly, particular in males, who were the oldest study participants. Our findings are consistent with those of Wang et al. who also showed an increase in the small fraction of highly mutated VH genes in most of the old subjects they studied [**?**].

I also analyzed the pattern of mutations. There are some nucleotide mutations in

our dataset, which occur more often in the elderly, compared to younger individuals (g → c or t, t → a, c → g, a → c). All these mutations are transversions, which occur less frequently than transitions, in general. In particular mutations to thymine/uracil, adenine or guanine can lead to translations of stop codons, resulting in a non-functional gene. This can be also confirmed by our dataset: in general, as well as for females and males respectively, old individuals had more unproductive genes (in average 4 - 5.2%), than young ones (in average 2.2 to 3.7%). However, one needs to keep in mind, that such differences in mutations can also be influenced by sequencing errors, which can not only substitute nucleotides, but also insert or delete mutations, which we cannot see anymore.

In summary, analyzing several B cell subsets, there are more differences with age, than with gender. Most of the differences can be seen in naïve B cells, but also in IgM+ memory cells. Hence it is not only the memory BCR repertoire which differs between young and old subjects, but also the naive B cell repertoire. Further I found many VH genes show different distributions among age groups, but also some which are influenced by biological sex. There is a trend, that old individuals have a more diverse B cell receptor repertoire, but also more mutations, often resulting in unproductive sequences. The repertoire of old individuals is not only smarter due to different infections the cells underwent, but also more prone to autoreactive cells due to such a diverse repertoire.

... But there is still a question that needs to be answered: Do autoimmune bullous diseases, age and gender influence the B cell repertoire of humans? – Yes, they do. Nevertheless, this answer needs to be differentiated. Although PV and BP are both autoimmune bullous diseases, they are caused by different mechanisms, which can be nicely seen when analyzing the B cell repertoire. I could show that differences between PV patients and healthy controls are mainly based on CDR3 sequence diversity and only few changes on gene usage, predominantly in IgM- memory cells and plasmablasts. There could be some kind of clonal selection in the patients, but also factors that may suppress development of special clones and lead to a narrowed repertoire. The higher diversity in plasmablast sequences of PV patients could indicate a relaxed control of autoreactive B cells. However BP patients can be distinguished from healthy controls mainly by different gene usage in IgM+, but also IgM- memory cells and plasmablasts. Many genes that were found to be overrepresented in BP patients are already known to be associated with other autoimmune or skin diseases. However, there are only few differences in diversity of CDR3 sequences between both groups.

Considering also age and gender differences in B cell repertoire, one can assume that there are stronger differences in age, than in gender. I could show, that it is not only the memory repertoire that shapes individuals by remembering pathogens. But a wider repertoire means also a higher chance of developing autoreactive cells and causing damage.

Also the spontaneous immune response in form of plasmablasts, is altered over time. There are particular VH genes which are overrepresented in old individuals, compared to young ones. Further CDR3 sequences seem to be more diverse in the elderly, compared to young adults. Not only a less effective antigen binding, but also a higher probability of somatic hypermutations may influence a different immune response in the elderly.

Further I found some gender specific differences in the B cell repertoire. Some genes that are overrepresented in women are already known to be associated with autoimmune diseases and it is also known that there are many autoimmune diseases which appear more often in females. Further there are genes in women that show age-dependent quadratic curves with a peak at an age of around 55 years, that may be associated with hormones, like estrogen or androgen. Such dependencies could not be found in men.

Supplement

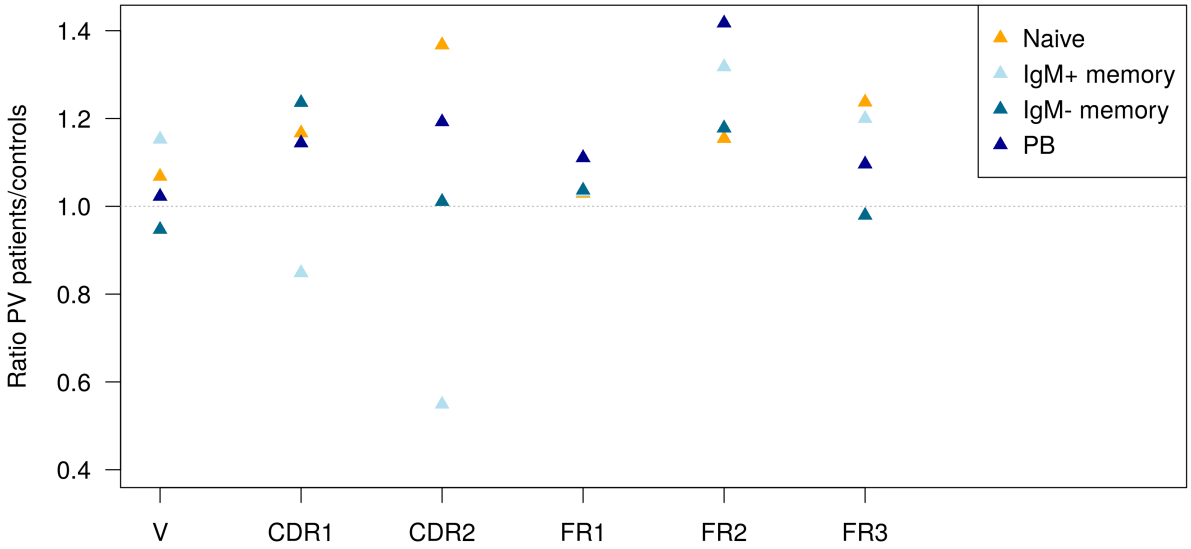

Figure S1: Ratio of mean R/S ratios in PV patients and controls. For most sequence parts (V, FR1-3, CDR1-2) PV patients have higher R/S ratios than controls. Except naïve cells the ratios are less than one in CDR1 and CDR2 regions, indicating higher R/S ratios in controls, compared to patients.

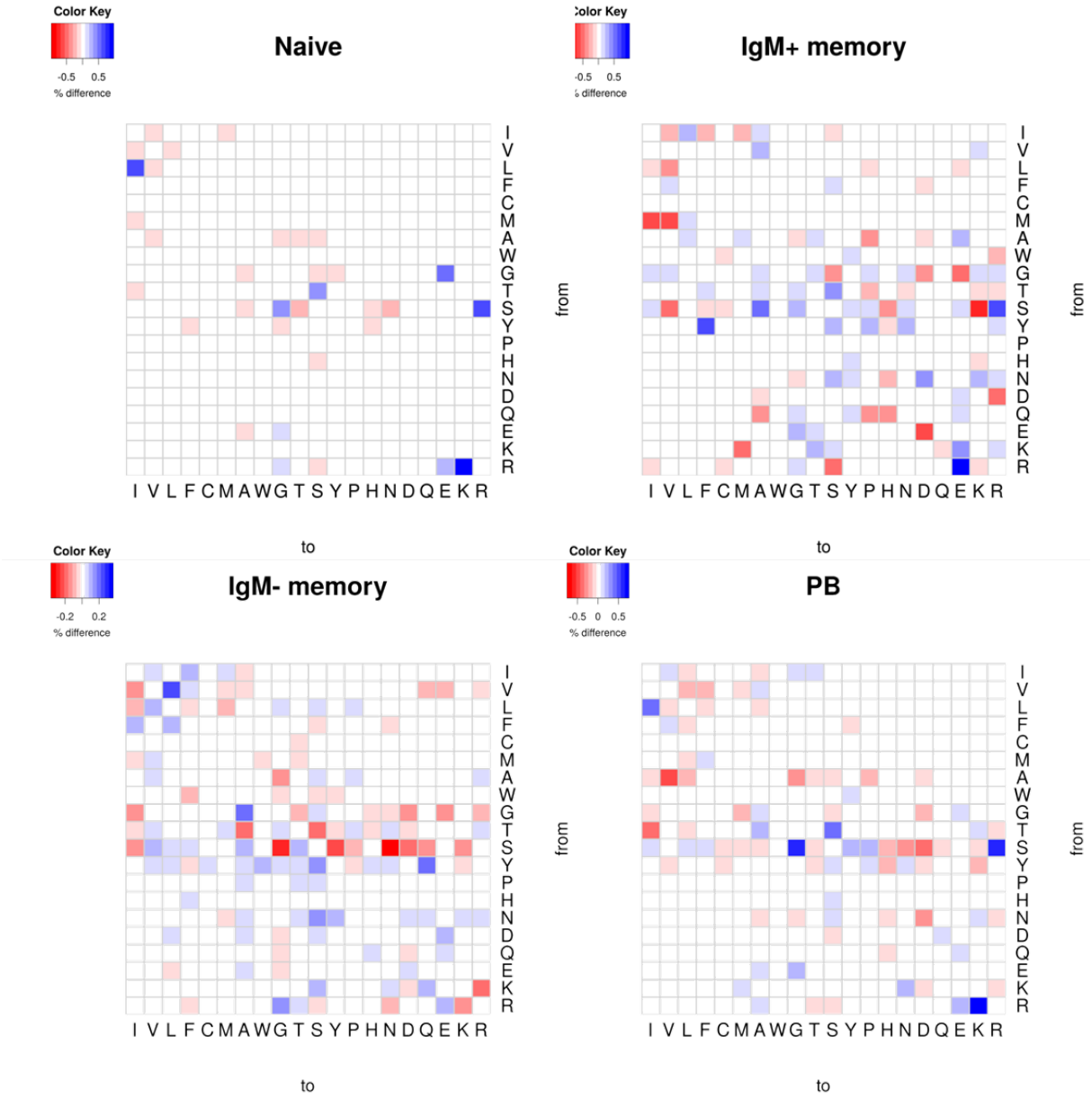

Figure S2: Percentages in amino acid mutation differences between PV patients and controls, from germline (from) to mutated nucleotide (to). Differences are color coded: red colors show higher proportions in PV patients, compared to controls; blue colors indicate higher proportions in controls. The darker the color, the higher the difference. White fields show no difference; gray ones were not analyzed. Most of the differences lay between zero and one percent, indicating high similarity in replacement mutations in both groups.

Figure S3: Heatmap of VH gene usage in PV patients (P1-P10) and controls (C1-C10). Proportions of V genes are color coded. Light colors prefer to small proportions, dark ones to high proportions. White fields mean 'no abundance'.

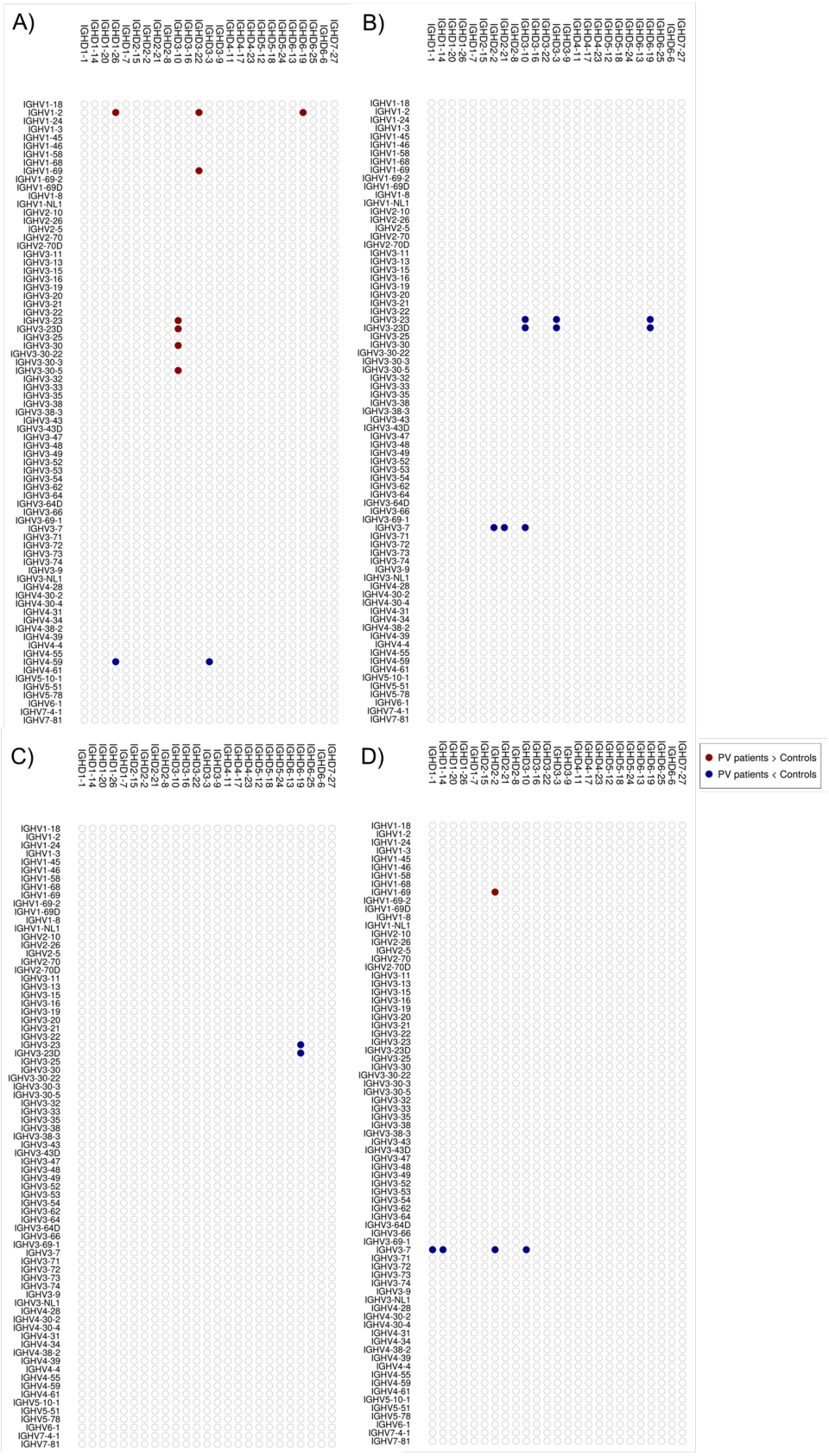

Figure S4: V-D gene combinations preferentially more abundant in PV patients, than in healthy controls. Combinations that exceed mean ± 2 standard deviations of the difference matrix (cases − controls) are shown. Difference directions are color coded (red and blue). A) naïve B cells, B) IgM+ memory cells, C) IgM- memory cells, D) plasmablasts.

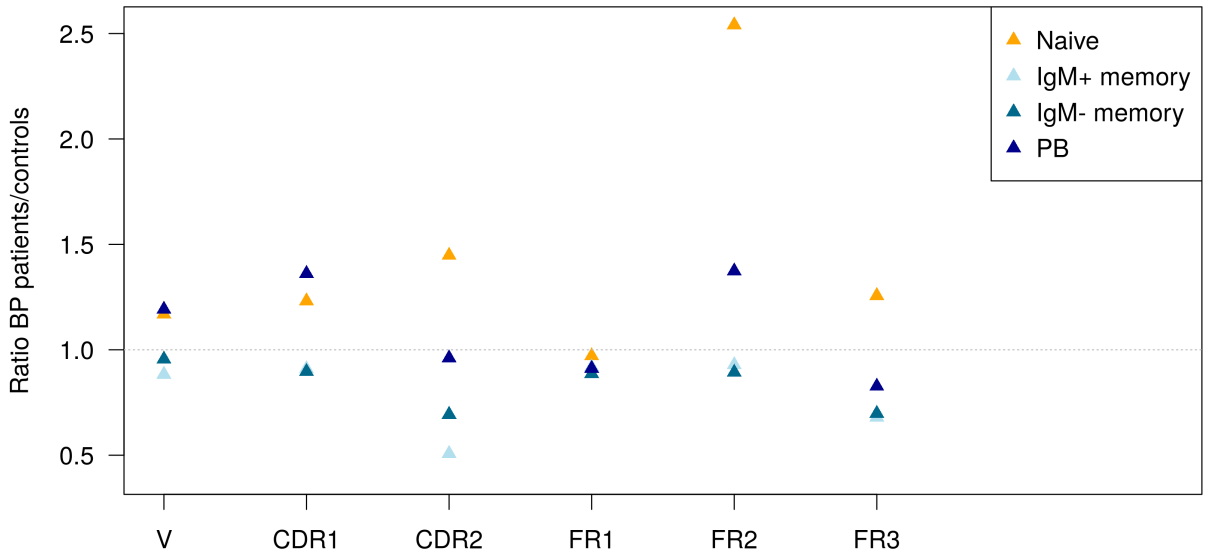

Figure S5: Ratio of mean R/S ratios in BP patients and controls. For most sequence parts (V, FR1-3, CDR1-2) BP patients have higher R/S ratios than controls in naïve cells and plasmablasts. In both IgM memory B cell subsets the ratios are less than one, indicating higher R/S ratios in controls, compared to patients.

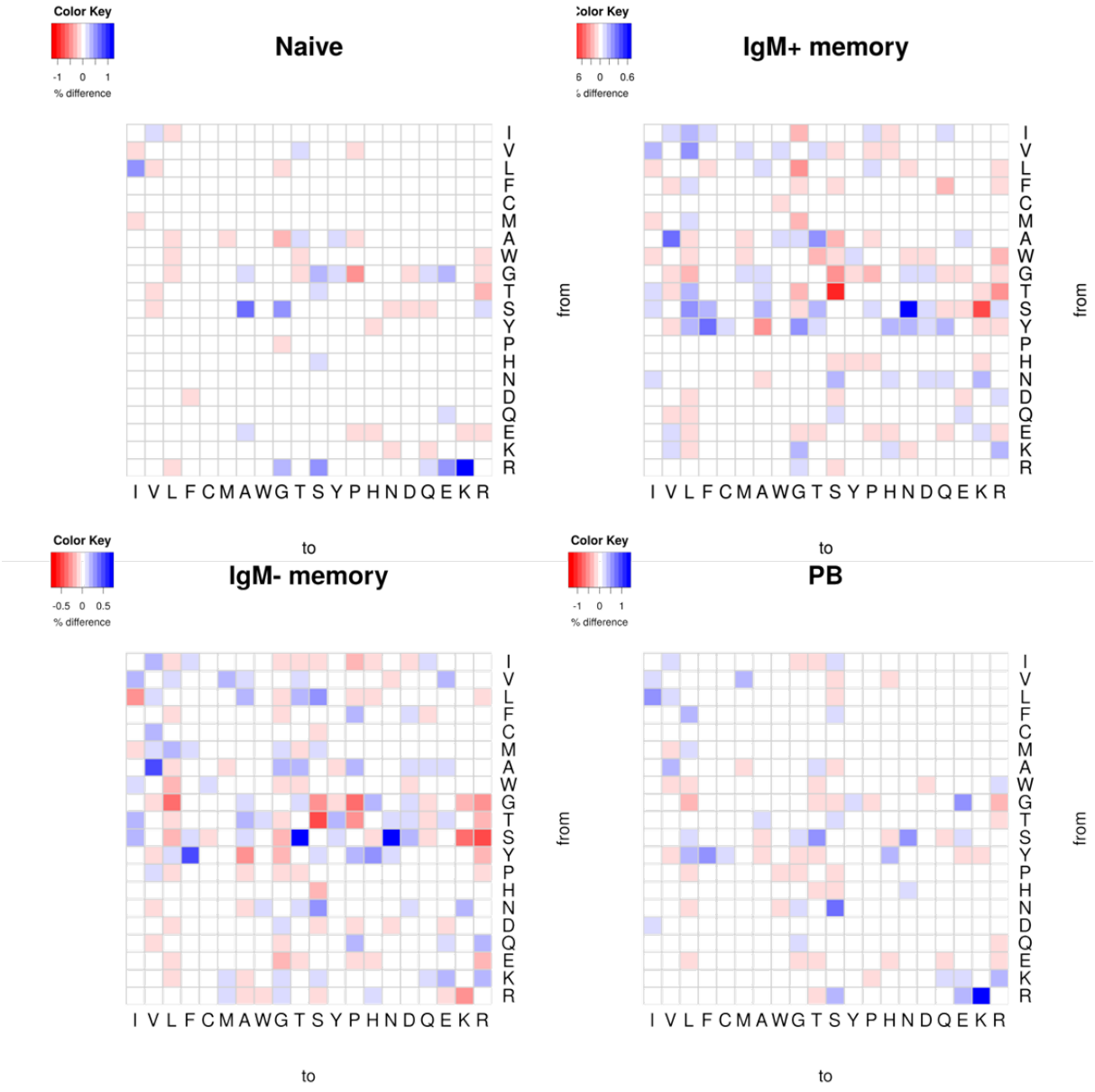

Figure S6: Percentages in amino acid mutation differences between BP patients and controls, from germline (from) to mutated nucleotide (to). Differences are color coded: red colors show higher proportions in BP patients, compared to controls; blue colors indicate higher proportions in controls. The darker the color, the higher the difference. White fields show no difference; gray ones were not analyzed. Most of the differences lay between zero and one percent, indicating high similarity in replacement mutations in both groups.

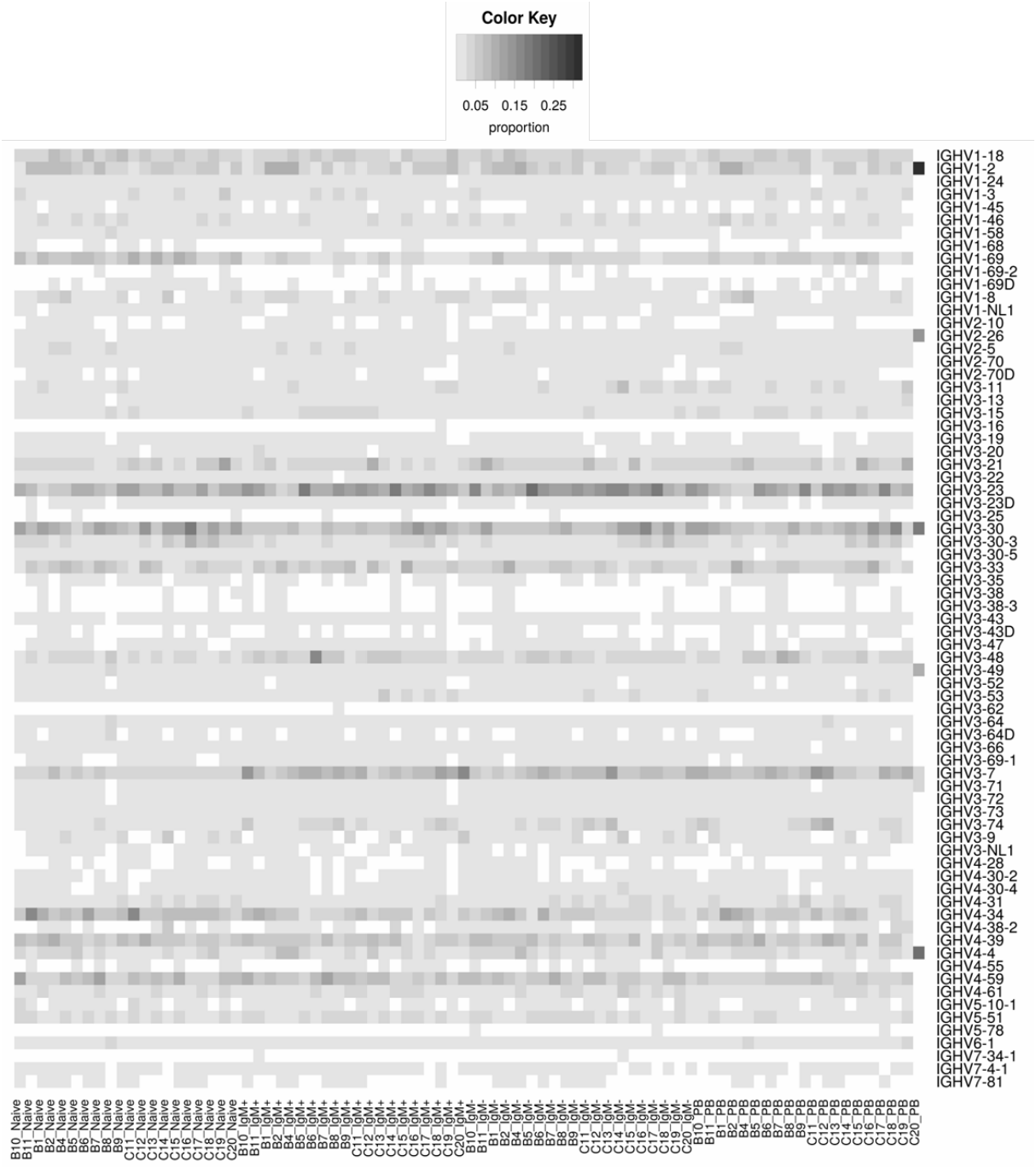

Figure S7: Heatmap of VH gene usage in BP patients (B1-B11) and controls (C11-C20). Proportions of V genes are color coded. Light colors prefer to small proportions, dark ones to high proportions. White fields mean 'no abundance'.

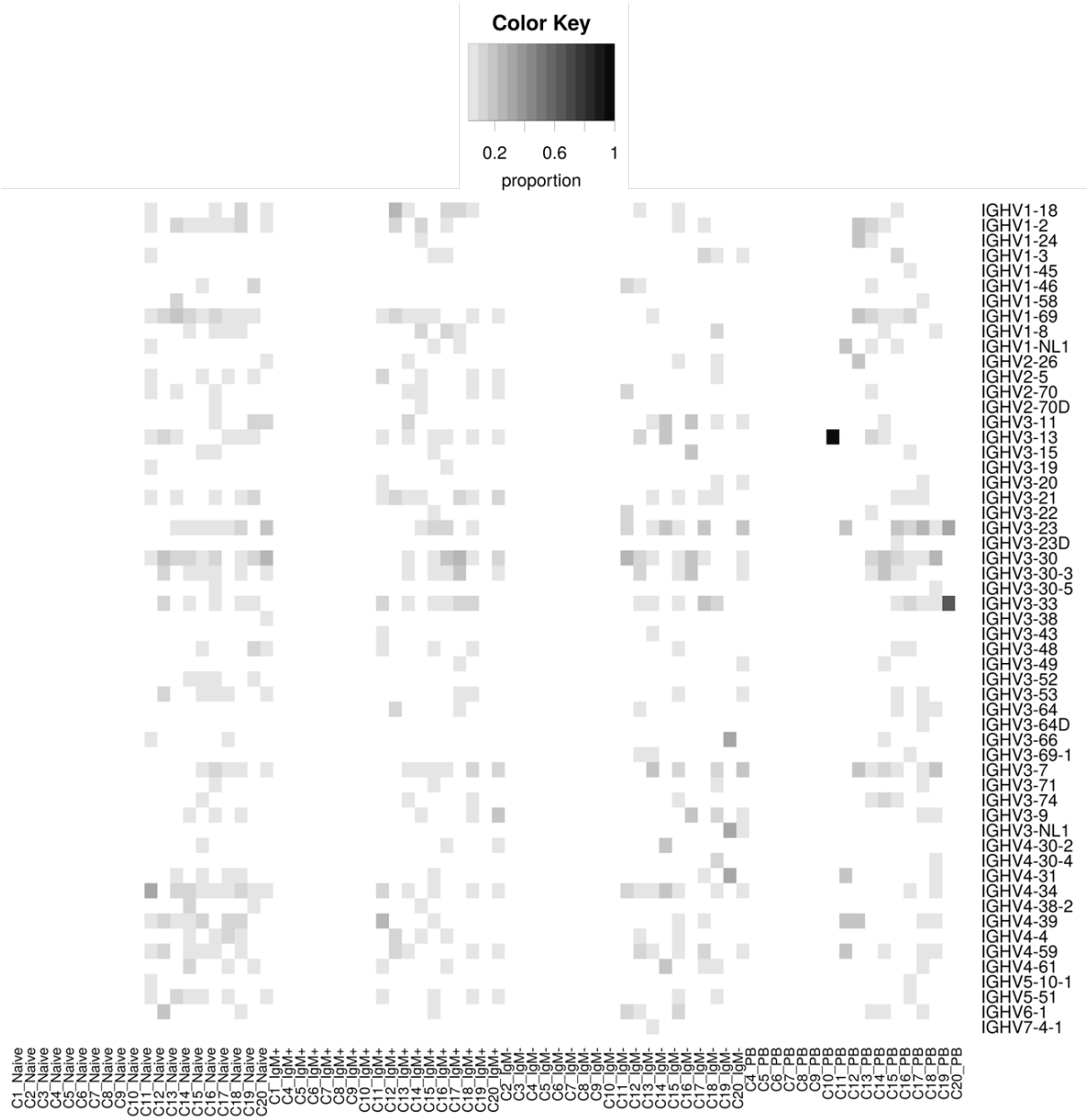

Figure S8: Heatmap of VH gene usage of sequences with CDR3 amino acid sequence lengths greater than 32 in healthy controls. Proportions of V genes are color coded. Light colors prefer to small proportions, dark ones to high proportions. White fields mean 'no abundance'.

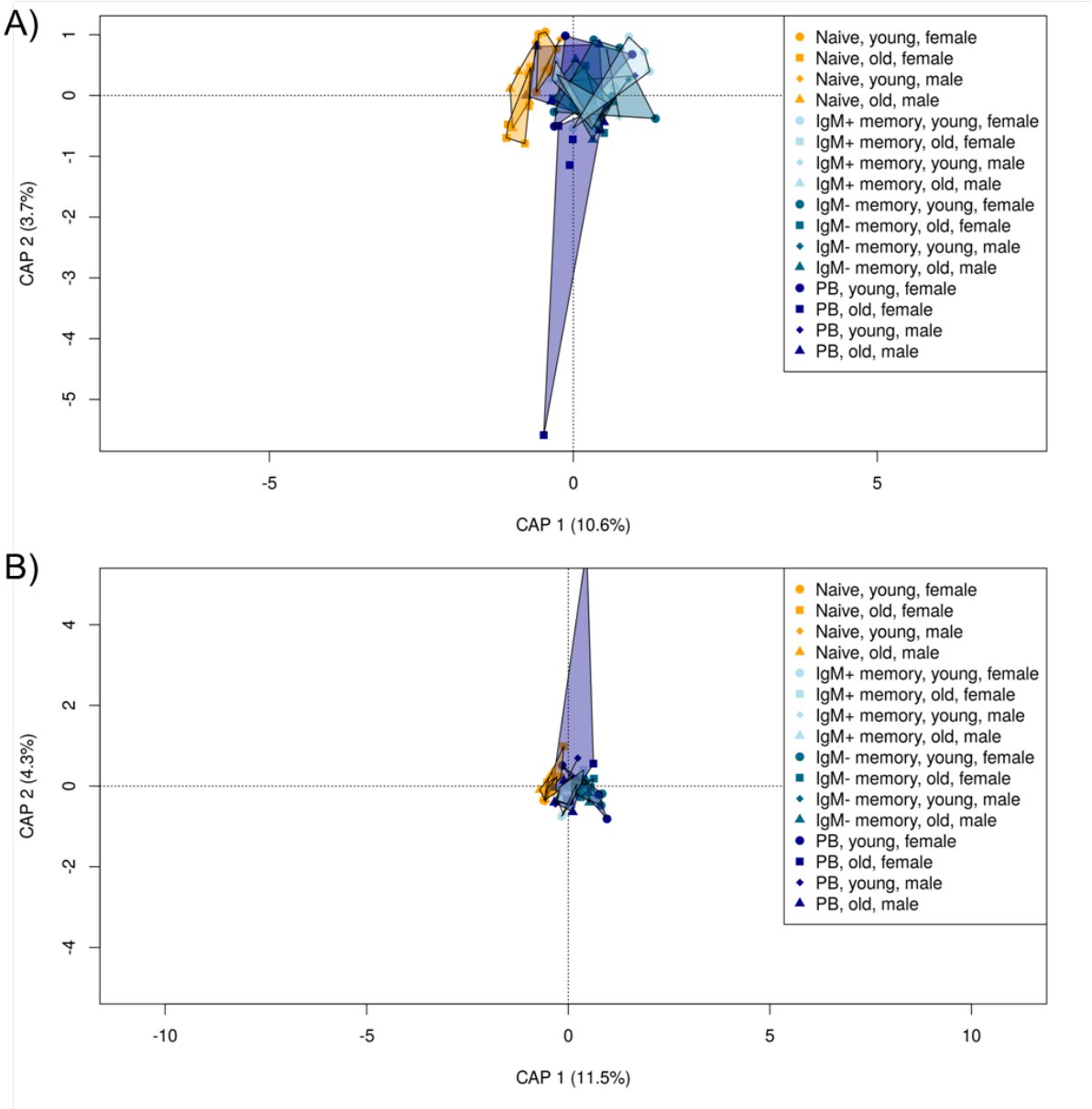

Figure S9: Constrained analysis of principal coordinates of A) VH gene usage and B) DH gene usage in all B cell subsets, using Bray-Curtis dissimilarity. In both panels, separation of the first three axes (CAP1-3) appears to be significant ($p<0.05$). For both VH and DH gene usage the first two axes explained about 14-16% of variance. Groups are color coded.

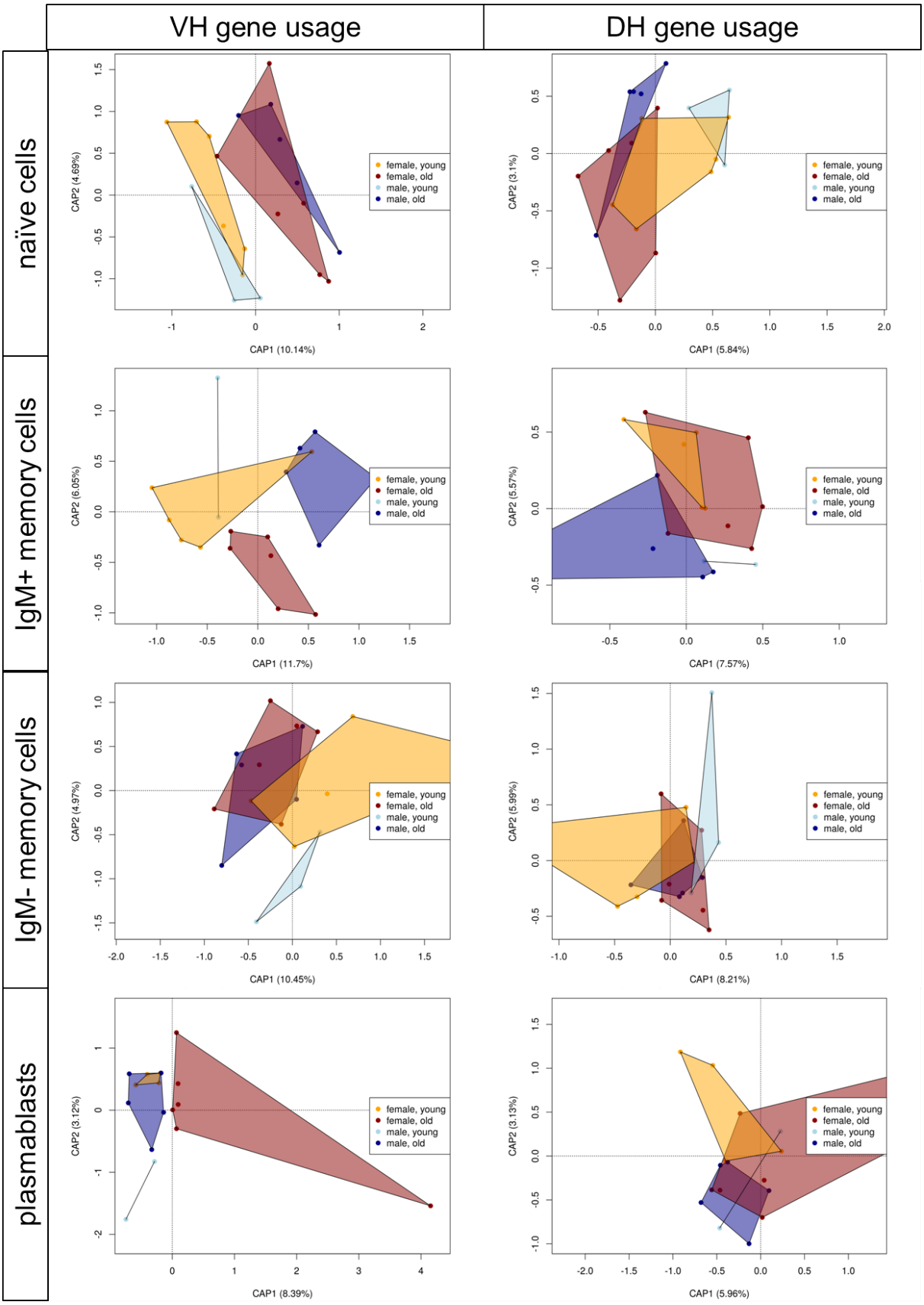

Figure S10: Constrained analysis of principal coordinates of VH and DH gene usage, using Bray-Curtis dissimilarity.Shown are first two axes (CAP1, CAP2). Age and gender groups are color coded.

Abreviations

AA	Amino acid
AIBD	Autoimmune bullous disease
AID	activation induced cytidine deaminase
BCR	B cell receptor
BLNK	B cell linker protein
BP	Bullous pemphigoid
bp	Base pairs
Bregs	B regulatory cells
CAP	Constrained analysis of principal coordinates
CD	Cluster of differentiation
CLL	Chronic lymphocytic leukemia
D / J / V segment	Diversity/Joining/Variable segment
DC	Dendritic cell
Dsg1, Dsg3	Desmoglein 1, Desmoglein 3
E2A	Transcription factor 3, TCF3 or E2A
EBF	Early B cell factor
GC	Germinal center
GM-CSF	Granulocyte macrophage colony-stimulating factor
H chain	Heavy chain
IFN-γ	Interferon gamma
Ig	Immunogloublin
IL-10, IL-17	Interleukin 10, Interleukin 17
IVIG	Intravenous immunoglobulin
L chain	Light chain
MDS	Multidimensional scaling
MHC	Major histocompatibility complexes
NC16A	Non-collagenous 16A domain
nt	Nucleotide
n.s.	Not significant
OOB	Out-of-bag error rate
PBMC	Peripheral blood mononuclear cell
PF	Pemphigus foliaceus

PNP	Paraneoplastic pemphigus
PV	Pemphigus vulgaris
RAG	Recombination-activating genes
R/S ratio	Ratio of replacement vs. silent mutations
Rss	Recombination signaling sequences
sIgM	Surface immunoglobulin M
SLO	Secondary lymphoid organs
TCR	T cell receptor
TdT	Terminal desoxynucleotidyl transferase
TGF-β	Transforming growth factor
Th1	T helper 1
TLR	Toll like receptor

Amino acids:

1-Letter	3-Letter	Amino Acid
A	Ala	Alanine
R	Arg	Arginine
N	Asn	Asparagine
D	Asp	Aspartic acid
C	Cys	Cysteine
E	Glu	Glutamic acid
Q	Gln	Glutamine
G	Gly	Glycine
H	His	Histidine
I	Ile	Isoleucine
L	Leu	Leucine
K	Lys	Lysine
M	Met	Methionine
F	Phe	Phenylalanine
P	Pro	Proline
S	Ser	Serine
T	Thr	Threonine
W	Trp	Tryptophan
Y	Tyr	Tyrosine
V	Val	Valine

Nucleotides:

1-Letter	Nucleotide
a	Adenine
c	Cytosine
g	Guanine
t	Thymine
u	Uracil

List of Figures

List of Tables

Acknowledgement

Firstly, I would like to express my special appreciation and thanks to my advisor Prof. Saleh M. Ibrahim for the continuous support of my PhD study and related research, for your patience, motivation, and immense knowledge. Your guidance helped me in all the time of research and writing of this thesis.

Besides Prof. Ibrahim, I would like to thank my second advisor: Prof. Rudolf A. Manz, for your insightful comments and encouragement, but also for the hard question which incented me to widen my research from various perspectives.

My sincere thanks also goes to Dr. Andreas Recke, Dr. Eline T. Luning Prak, Dr. Wenzhao Meng, Dr. Uri Hershberg and Dr. Axel Künstner for the stimulating discussions and for enlightening me the first glance of research. I also thank all other LIED colleagues. Without their precious support it would not be possible to conduct this research.

Last but not least, I would like to thank my family and friends for supporting me spiritually throughout writing this thesis and my life in general. You have been the best support in the moments when there was no one to answer my queries.