Investigations into the human immunoglobulin repertoire utilizing high-throughput sequencing

Inaugural-Dissertation

To obtain the academic degree

Doctor rerum naturalium (Dr. rer. nat.)

Submitted to the Department of Biology, Chemistry and Pharmacy

of Freie Universität Berlin

by

Florian Rubelt

born in Berlin, Germany

September 2012

I hereby declare that all experiments and writing contained within this thesis were conducted by myself, all references used are cited accordingly and any personal assistance has been acknowledged by name.

All experiments for this thesis were conducted from October 2008 to September 2012 in the working group of Dr. Zoltán Konthur at the Department of Vertebrate Genomics under the supervision of Prof. Dr. Hans Lehrach in the Max Planck Institute for Molecular Genetics.

1. Reviewer: Prof. Dr. Hans Lehrach
Max Planck Institute for Molecular Genetics
Department of Vertebrate Genomics
Ihnestraße 63-73
14195 Berlin, Germany

2. Reviewer: Prof. Dr. Burghardt Wittig
Freie Universität Berlin
Molecular Biology and Bioinformatics
Arnimallee 22
14195 Berlin, Germany

Date of defense: December 13, 2012

Acknowledgment

I would like to thank Prof. Dr. Hans Lehrach, Director at the Max Planck Institute for Molecular Genetics who gave me the great opportunity to conduct my dissertation at his Department in such an interesting research area. Furthermore, I thank Prof. Dr. Burghardt Wittig from Freie Universität Berlin for his kindness to serve as an academic advisor of my thesis.

Foremost, I would like to express my sincere gratitude to my direct supervisor Dr. Zoltán Konthur for the continuous support of my dissertation, for his great tutoring, advice and stimulating discussions as well as to give me the opportunity to travel, meet and collaborate with international scientists during this time. I also would like to thank all the members of the group, current and previous, for their helpful hands and helpful discussions. In particular, Dr. Volker Sievert, Carola Stoschek and Friedericke Braig as well as my two companion Ph.D. students Stephan Klatt and Yuliya Georgieva for good working atmosphere, support and friendship.

My sincere thanks also goes to my closest collaboration partners, Dr. Christian Diener and Florian Knaust for bioinformatics and sequencing advice and helpful discussions.

Additionally I would like to thank my parents for their support in the last years. Furthermore Benjamin Weist for fruitful discussion on immunology.

Last but not the least I would like expresses my deep gratitude to Miriam, my wife, for her love, understanding and belief in me, even in difficult times.

Thank you all for a fascinating time in my life.

Table of contents

List of abbreviations

AID	activation-induced cytidine deaminase
Am	genetic marker of α chains
ATP	Adenosine triphosphate
b-cell	bursa of Fabricius cell (immunoglobulin producing cell)
BCR	B-cell receptor
BMI	body mass index
bp	base pair
BSA	bovine serum albumin
C	constant
CD	cluster of differentiation
cDNA	complementary DNA
CDR	complementarity-determining region
CH1	constant heavy 1
CL	constant light
cm	centimeter
CSR	class switch recombination
D	diversity
DBMS	data base management system
DNA	deoxyribonucleic acid
dNTPs	deoxyribonucleotide triphosphates
dsDNA	double strand DNA
E. coli	*Escherichia coli*

emPCR	Roche’s bead-based emulsion PCR
ePCR	water-in-oil emulsion PCR (for amplicon generation)
ET SSB	extreme thermo stable single strand binding protein
µl	micro liter
Fab	fragment of antigen binding
Fc	fragment crystallizable
Fc fragment	fragment crystallizable region
FcεRI	Fc epsilon receptor type I
FR	framework region
gb	giga base
GC	germinal centers
gDNA	genomic DNA
Gm	genetic marker of γ chains
HC	heavy chain
Ig	immunoglobulin
IgA1	immunoglobulin α 1 (subclass of IgA)
IgA2	immunoglobulin α 2 (subclass of IgA)
IgD	immunoglobulin δ
IgE	immunoglobulin ε
IgG 1	immunoglobulin γ 1 (subclass of IgG)
IgG 2	immunoglobulin γ 2 (subclass of IgG)
IgG 3	immunoglobulin γ 3 (subclass of IgG)
IgG 4	immunoglobulin γ 4 (subclass of IgG)
IGHD	immunoglobulin heavy diversity
IGHJ	immunoglobulin heavy joining

IGHV	immunoglobulin heavy variable
IGKJ	immunoglobulin kappa joining
IGKV	immunoglobulin kappa variable
IGLJ	immunoglobulin lambda joining
IGLV	immunoglobulin lambda variable
IgM	immunoglobulin μ
IgSF	immunoglobulin superfamily
IMGT	international ImMunoGeneTics information system®
J	joining
kDa	kilo Dalton
kg	kilo gram
Km	genetic marker of κ chains
LC	light chain
mg	milli gram
MID	multiplex identifier
ml	milli liter
mRNA	messenger RNA
N	nucleotide
ng	nano gram
NGS	next generation sequencing
NHEJ	non-homologous end joining
PacBio	Pacific Bioscience
PAMP	Pathogen-associated molecular pattern
PBMC	peripheral blood mononuclear cell
PCR	polymerase chain reaction

pmol	pico mol
PTP	PicoTiterPlate
RAG	recombination-activating genes
RE-site	recognition motif
RNA	ribonucleic acid
RSS	recombination signal sequences
RT	reverse transcription
RTases	reverse transcriptases
SHM	somatic hyper mutation
SOLiD	Sequencing by Oligo Ligation Detection
ssDNA	single strand DNA
T-cell	thymus derived lymphocytes
TNF	tumor necrosis factor
V	variable
VH	variable heavy
VL	variable light
κ	kappa
λ	lambda

1. Introduction

The immune system is a very complex and fascinating composition of different cells and organs. Immunology as study of the body's defense against infections came into scientific focus after Edward Jenner 1796 discovered the first vaccination against smallpox [1, 2]. Since then, the knowledge about the immune system has constantly increased. More and more strategies are currently being developed to use this know-how to cure people or prevent certain diseases, particularly by vaccination [3]. Research in this field helps to save and improve the quality of life of millions of people, for instance through deeper understanding of the functionality of B-cells on which the humoral immune response is based upon [4]. Hence, the focus of my thesis centers on a vital part of B-cell biology – the B-cell derived immunoglobulins, which play a major role in the immune defense of the human body.

Once an infectious agent tries to enter our body, the first lines of defense are the physical and chemical barriers that protect us from pathogens [5]. When this border is passed, the innate immune response comes into play, which involves a variety of resistance mechanism and cells, such as macrophages and dendritic cells. This unspecific part of the immune system is present at all time and responds rapidly to foreign antigen exposure. Among other functions, the innate immune cells are able to ingest and kill microbes by producing a variety of toxic chemicals and strong degenerative enzymes [6]. Innate immunity has also the ability to distinguish between self and nonself antigens by recognition of certain molecular structures, like membranes of infectious microorganisms [7]. This antigen recognition is mediated by cell-to-cell contact and/or soluble receptors. The receptors recognize pathogen-associated molecular patterns (PAMPs) [7, 8], such as lipopolysaccharide or bacterial (unmethylated) CpG DNA [9]. Due to their defined structure, which is encoded in the germ line, no cross reactivity with host cells occur [10]. In contrast to the adaptive immune system, innate immunity does not alter by clonal cell expansion when it is repeatedly exposed to a pathogen, but plays a key role in the activation and orientation of adaptive immunity [11]. The adaptive immune system is more specific. It is able to develop an immunological memory and it is efficient in eliminating infections, but its response is delayed [1]. Adaptive immunity emerged during the earliest evolutionary stages of vertebrates and specifically with the appearance of gnathostomata (vertebrates with jaw) [10, 12]. The cell repertoire of this system is able to focus on any specific pathogens and is mainly based on lymphocytes: B lymphocytes (B-cell; from bursa of Fabricius (lymphoid organ in young chicken)) and T lymphocytes (T-cell (thymus derived lymphocytes)) [1]. Each cell type has its ascertained

role and function in the adaptive immune system. Both, B and T lymphocytes originate in the bone marrow; B lymphocytes also mature there whereas T-cells develop further in the thymus [4, 13].

The receptors on B and T-cells responsible for precise binding of antigens are encoded in different gene segments. These segments are rearranged and can be further hyper-mutated to specify and increase affinity [14, 15]. However, some mechanisms are necessary to avoid self-recognition of the receptors as these reactions could by harmful [16].

Much of the basic research in immunology is conducted on laboratory animals, which results in great breakthroughs in understanding the system. Nevertheless, animal models have some limitations when it comes to transforming the obtained findings to humans. The immune system of laboratory animals remains different in many aspects [17]. With new technologies it has become possible to conduct also studies directly on human samples. Though, there are obviously more ethical and technical limitations in the design of such studies [18]. However, the results of this thesis can be directly beneficial for people, for instance in the field of vaccine development.

The research area of immunoglobulins is of great interest to a broad field of scientists. Notably, numerous Noble Prices in Physiology or Medicine were awarded for investigation on antibody structure and their creation.

1972 to Gerald M. Edelman and Rodney R. Porter "*for their discoveries concerning the chemical structure of antibodies*" [19].

1984 to Niels K. Jerne, Georges J.F. Köhler and César Milstein *"for theories concerning the specificity in development and control of the immune system and the discovery of the principle for production of monoclonal antibodies"* [20].

1987 to Susumu Tonegawa *"for his discovery of the genetic principle for generation of antibody diversity"* [21].

However, the research on immunoglobulin repertoire, e.g. of healthy humans and their changes with aging are still under investigation. A novel and powerful tool are new sequencing technologies to investigate natural repertoires of immunoglobulins, recently demonstrated and published by Weinstein *et al.* 2009 using Zebrafish [22].

1.1 Immunoglobulins

During the fight of the body against a wide range of pathogens, the lymphocytes of the adaptive immune system – which evolved to recognize a great variety of different antigens from bacteria, viruses, and other harmful organisms – take a central role [4].

All cellular components of the blood derive from hematopoietic stem cells of the bone marrow [1]. During development – following the basic principles of cell differentiation – cells become more and more determined acquiring the specific function of the mature cells. Lymphocytes first become lymphoid and then develop either into B-cell or T-cell lineages [13]. The lymphoid progenitor gives rise to the earliest B-lineage cells, the pro-B-cells, in which immunoglobulin gene rearrangement begins. Immunoglobulins (or antibodies) are the antigen-recognition molecules and serve as cell receptors on the surface of the B-cell (B-cell receptor (BCR)) [23]. Furthermore, they can be secreted as antibodies with the same antigen specificity by terminally differentiated B-cells, called plasma cells. The secretion of antibodies is the main function of B-cells in adaptive immunity.

1.1.1 Antibody structure

Antibodies or immunoglobulins are Y-shaped proteins produced by B-cells, and belong to the eponymous immunoglobulin superfamily (IgSF) [23, 24]. The arms of the Y are the Fab fragments (fragment of antigen binding) and each of them is able to bind antigens, while the Fc fragment (fragment crystallizable region) presents the body (figure 1.1). This classification of the structure was facilitated by the enzymatic cleavage with papain, which produced three distinct fragments with distinct properties. Antibodies are heterotetrameres consisting of two heavy (HC) and two light chains (LC). The heavy and light chains of the immunoglobulins are each encoded by a separate multigene family [25, 26] and the individual V (variable) and C (constant) domains are each encoded by independent elements. The constant domain is encoded by individual exons and the V(D)J (variable (diversity) joining) gene segments determine the different specificities of affinities for different epitopes of the immunoglobulin. The HC exits in different isotypes and can be subdivided into five major immunoglubulin classes resulting in different effector functions of the antibody. The HC locus is located on chromosome 14q32.33. For specification of the immune reaction the HC can be changed by class switch recombination (CSR) and modify the effector function (see also isotypes). The

LC consist of *κ* (kappa) or *λ* (lambda) chains and cannot be changed by CSR. The *κ* locus is located on chromosome 2p11.2. and is encoded by a single exon [27]. The *λ* chain is encoded by four functional genes on chromosome 22q11.2 [28]. The tips of the Y shaped antibody are the paratopes that typically bind the epitope – the site on the antigen, which is recognized. The binding domain can be divided into 3 hypervariable intervals termed complementarity-determining regions (CDRs) and is accompanied by 4 regions of stable sequences termed framework regions (FRs). The antigen binding side of the antibody is highly diverse and able to bind specific epitopes with a high affinity. The specific recognition of epitopes which presents only parts of the antigens enables the immune system to distinguish between closely related antigens. Next to this, the different allotypes are antigenic determinants after inhibition of serological hemagglutination. They are called Gm, Am and Km as genetic markers of γ, α and κ chains and may also affect the specific antibody responses to infectious agents between individuals [29, 30].

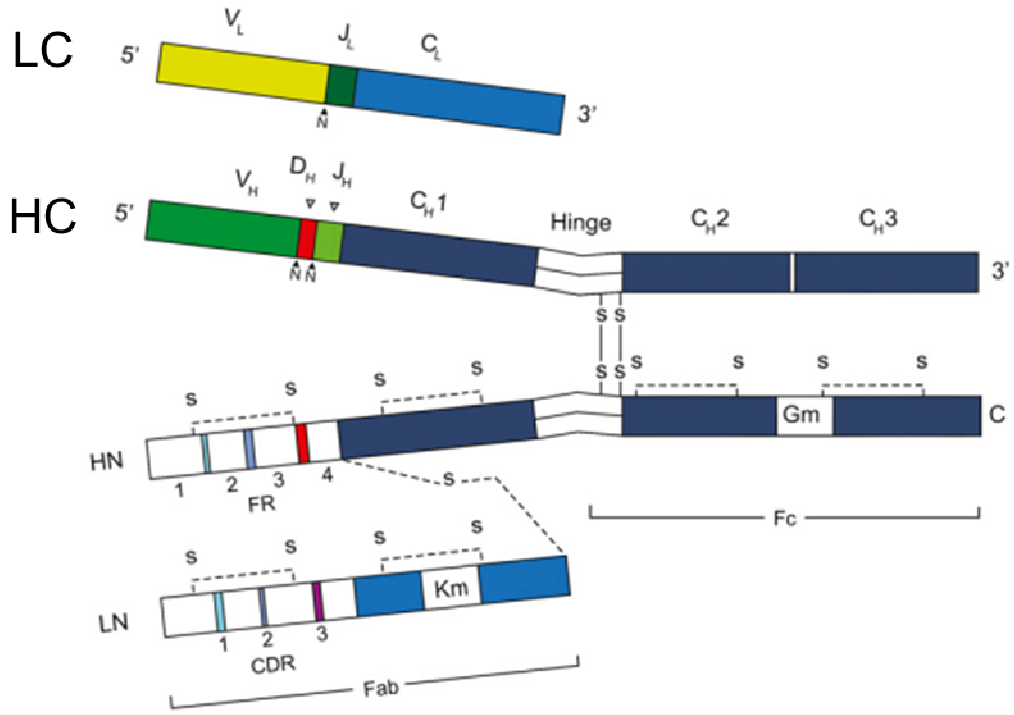


Figure 1.1: Schematic drawing of an IgG molecule. Schematic drawing of an IgG molecule consists of two heavy and two light chains (HC and LC). The upper part represents the HC and LC at nucleotide level, whereas the lower part displays the protein sequence. The LC consists of the antigen binding site consisting of the variable V_L and the joining domain J_L as well as the constant part C_L kappa or lambda. The N indicates nucleotide insertion or deletion in the V_L and J_L joining region. The HC consists of the antigen binding site consisting of the V_H, D_H (diversity) and J_H gene segments linked with a flexible amount of nucleotides (N). The hinge region connects the constant heavy 1 (C_H1) in the Fab arm with the C_H2 in the fragment crystallizable (Fc) region, which determines the effector function of the immunoglobulin isotype. The protein schemata from the amino terminus (N) to the carboxy terminus (C) indicates the disulfide bridges (S--S) and the allotype marker regions Gm and Km (γ and κ chains, respectively). The three complementarity-determining regions (CDR) in the 5' part (V(D)J region) of the Fab arm is colored and the four framework regions are shown in white. Figure modified from Schroeder *et al*. [23].

1.1.2 Isotypes

The light chain isotypes can be subdivided into κ and λ subgroups. Both have their own constant domains: There is one for κ and four functional ones for λ. Each of the constant domains has its own associated V and J genes. Assembly of the light chain to the immunoglobulin occurs in the pre-B-cell phase and cannot be changed in a later process.

The heavy chain has different constant genes, which lead to different antibody classes and induce different effector functions.

Table 1.1: Isotype differences.

	IgM	IgD	IgG1	IgG2	IgG3	IgG4	IgA1	IgA2	IgE
Heavy chain	μ	δ	γ	γ	γ	γ	α1	α2	ε
Serum level mg/ml [1]	1.5	0.03	9	3	1	0.5	3.0	0.5	0.00005
Classical Complement activation	+++		+++	+	+++				
Molecular weight (kDa)	970	184	146	146	165	146	160	160	188
Half-life in serum (days)	10	3	21	20	7	21	6	6	2

+ symbolizes the strength of activation capability of the complement system

IgM is the immunoglobulin class expressed first, consisting of a μ heavy chain and being associated with the primary immune response. IgMs have a low affinity due to their immaturity, because they have not undergone a somatic hypermutation in response to an antigen. Therefore, IgMs tend to be more polyreactive compared to other isotypes.

IgDs are produced by alternative splicing from the same mRNA as IgMs and are also found circulating in the blood in very low levels. The function of IgDs is not yet fully understood but it is proposed that they regulate the B-cell fate at specific development stages [31] and that they could bind specific proteins of bacteria with their Fc part, resulting in B-cell stimulation and activation [32]. IgD can replace IgM on the surface of B-cells (switching back and forth is also possible) or can be found on B-cells together with IgMs.

The remaining Ig-classes IgG, IgA and IgE are generated by CSR and appear to specify the effector function of the immunoglobulins.

IgG is the most prominent isotype with the longest half-life of all immunoglobulin classes and dominates the humoral immune system. This class can be subdivided into four subgroups – IgG1 to IgG4, whereby the naming is according to their finding and relative concentration in the blood of healthy people in Western Europe from high to low. The different subclasses have different effector functions. IgG1 and IgG3 are produced in response to protein antigens whereas the other two are involved in the response to polysaccharide antigens. Like IgM, IgG1, IgG2 and IgG3 are able to activate the complement cascade (see below) while IgG4 is not complement-activating. In general, the function of IgG4 is poorly understood. They have an interesting characteristic – the 'Fab-arm exchange', which leads to asymmetric antibodies with two different antigen binding sites. The generation of monovalent, non-cross-linking IgG4 antibodies results in antibodies with a non-inflammatory effect, while bi-specific antibodies possess a potentially pro-inflammatory effect [33]. Only IgGs can be placentally transferred to the fetus.

IgAs are normally more frequent in the serum than IgM, but less frequent compared to IgG. They dominate the humoral mucosal immunity [34] and are crucial in the immune protection of infants [35]. Two subtypes are known, namely IgA1 and IgA2. IgA1 has a longer hinge region and is more sensitive to bacterial proteases. IgA1 is predominant in serum, whereas IgA2 dominates in mucosal secretions and genital tract: IgA2 is more stable compared to IgA1 due a 13 amino acid deletion in the hinge region that harbors a protease recognition site [36, 37].

The discovery of IgE [38], decades after the discovery of the other immunoglobulin isotypes [39], had a great impact on the understanding of allergic reactions. IgE has the lowest serum concentration and a half-life in plasma of less than 48 hours. The rapid absorption of IgE in tissues and the small numbers of B-cells switched to IgE explain the low serum titer and the late discovery in 1968 [40]. The high titer of IgE in helminthes-infected populations indicates that IgE may defend originally against metazoan parasites. Therefore, protection against parasites is called as one of the major task of this isotype [39]. IgE is associated with prominent problems of the immune system, namely hypersensitivity and allergy. IgEs can be captured on the surface of other immune cells (e.g. mast cells) by binding to the Fc epsilon receptor type I (FcεRI). Once bound, IgEs lend their specificity of antigen recognition to the capturing cells and may persist for several weeks in this state [41, 42].

The complement system – as an example for isotype effector function differences – is important for the clearance of pathogens and apoptotic cells (table 1.1). Upon antigen

binding, some antibodies are capable to activate the complement system through activation of the classical pathway wherein complements act as a rapid and efficient immune observation system. Amongst other functions it also contributes substantially to cell homeostasis by eliminating cellular fragments and infectious microbes [43]. The various Ig isotypes possess different activation capabilities: IgM, IgG1 and IgG3, are very effective in activation, while IgG2 fixes complement relatively poorly. IgG4, IgA, IgD and IgE are incapable to activate the classical complement pathway [44].

1.1.3 V(D)J recombination and somatic hypermutation

The diversity of immunoglobulin paratopes is a result of multiple processes. The major process is the rearrangement of variable (V), diversity (D) and joining (J) gene segments on chromosomal level into complete antibody recombinations (figure 1.2) [26]. VDJ or VJ recombinations in HC or LC, respectively, take place during lymphocyte development. All HC isotypes share the same VDJ genes, whereas kappa and lambda chains have their own distinct combinations of V and J genes. These rearrangements of genes during lymphocyte development lead already to a high diversity of binding sites of the antibody repertoire by only a limited number of genes. Mechanistically, recombination is a series of enzyme-guided specific DNA breakage and rejoining events. Each of the V, D and J segment is flanked by recombination signal sequences (RSS) of different length located at the joining ends of the genes. The RSS is a short DNA stretch containing a conserved heptamer, a 12 or 23 bp spacer and a nonamer at the other side. Joining of segments typically involves a 12-bp and a 23-bp RSS according to the 12/23 rule. The V_λ (23 spacer) can recombine with J_λ (12 bp), and V_κ (12 bp) can recombine with J_κ (23 bp). In the HC V_H and J_H have both a 23 bp spacer avoiding recombination without the D_H, which possess 12 bp RSS on both sides. For recombination to take place, two recombination-activating genes (RAG1 and RAG2) are required [45] which are almost exclusive expressed in developing lymphocytes [46]. In the DNA cleavage phase, a 12 RSS and a 23 RSS come close together and the spacers place the heptamer and nonamer on the same side of the DNA molecule forming a 1 turn helix (12 bp spacer) or 2 turn helix (23 bp spacer). Composition of this process may be mediated through stepwise capture model of assembly, which involves initial RAG binding [14]. RAG1 and RAG2 introduce a DNA double strand break of both RSS. Both ends are then combined in the nonhomologous end joining (NHEJ) process. The joining between the RSS ends is precise and the repair of joining ends is template independently conducted by nucleotide

addition through fill-in DNA synthesis, which also contributes to the diversification of the antibodies.

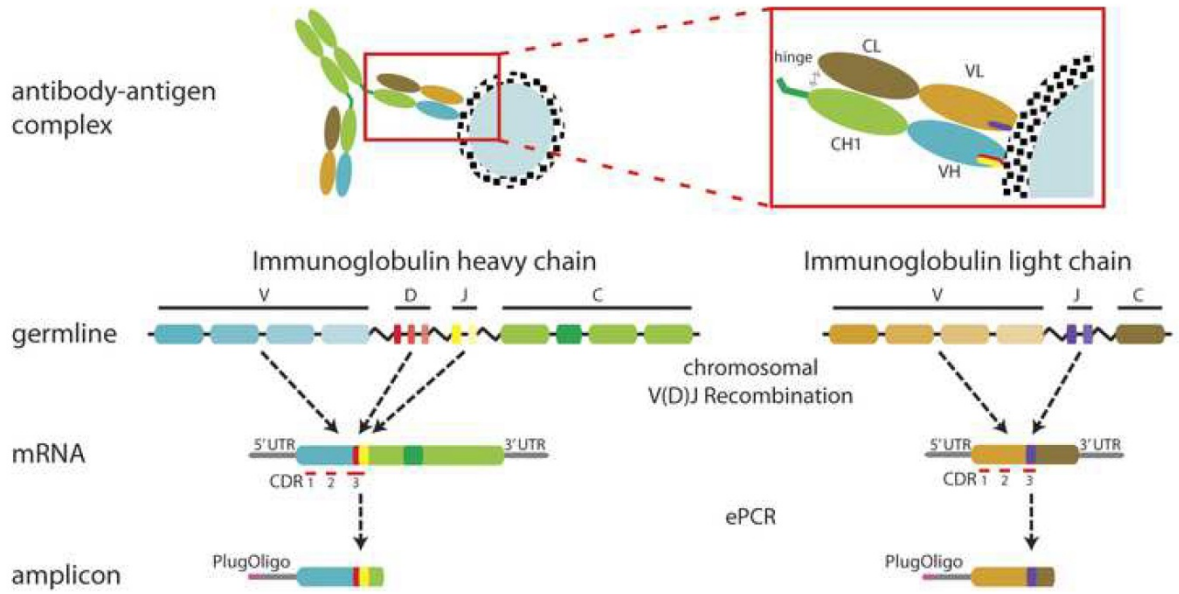


Figure 1.2: Schematic illustration of immunoglobulin IgG and mechanism of V(D)J recombination. (A) Antibody-antigen interaction with zoom on Fab arm highlighting the assembly of the antigen binding site. **(B)** Chromosomal V(D)J rearrangement in the B-cell connects one Variable-gene (V) with one Diversity- (D; only HC) and one Joining-gene (J) out of a pool of different V(D)J genes to enable a high diversity of specifities within the antibody repertoire. Constant (C) domains specify the induced immune reaction. VH: Variable heavy; VL: Variable light; CH1: Constant heavy 1: CL: Constant light. CDR: complementarity determining region. Figure modified from Rubelt *et al*. manuscript IV.

Once the heavy chain has been rearranged, the pro-B-cell tests this recombination before the light chain rearrangement is initiated. The cell produces two constant surrogate proteins, which can be assembled with the μ chain to form the pre-B-cell receptor. Signal from the pre-B-cell receptor enforces allelic exclusion (also later on light chain) to prevent different receptors to arise on one single cell. In the light chain, nonproductive rearrangements can be rescued by further recombination. At first, the κ locus will be rearranged, if this is not productive the λ locus will be rearranged. Eventually the immature IgM receptor is expressed and finally in mature B-cells also IgD is co-expressed through alternative splicing.

Mature B-cells are still able to diversify by somatic hyper mutation (SHM). Once mature B-cells bind to an antigen, they get stimulated to differentiate further. The cells migrate to the dark zone of germinal centers (GC) and develop to centroblasts [47]. GC are compartments within secondary lymphoid organs and compose the site for B-cell clonal expansion, SHM as well as affinity-based selection and T-cell interaction for the production of antibodies with a high affinity for given antigens [48, 49]. SHMs are introduced by activation-induced cytidine deaminase (AID), of which the expression in highly upregulated in GC. AID editing enzyme initiates random SHM by deaminating deoxycytidine residues to deoxyuridines in ssDNA during transcription of the Ig locus [15]. SHM introduced mutations are ultimately allowing affinity maturation to the antigen, as point mutations within V(D)J exons lead to different amino acids and therefore to a modified paratope of the antibody. The competitive selection of B-cells for antigen and T-cell interaction ensures only survival of the cells with the highest affinity for the given antigen [48].

Many of the enzymes needed for SHM (e.g. AID) are also required for CSR also occurring in the centroblasts B-cells [46], though SHM and CSR do not occur simultaneously in a cell [50]. CSR is an end-joining rather than a homologous recombination [50]. Via cytokine signaling of T-cells and other cells, naive B-cells are competent to switch to any isotype [51]. Some of these activated cells become memory B-cells through follicular helper CD4 (cluster of differentiation 4) T interleukin 21 production [52], whereas others differentiate into plasma cells that continuously secrete high levels of antibodies.

1.2 Immune senescence and vaccination

Edward Jenner, a country doctor living in England, published 1798 [2] his findings on the first vaccination against smallpox [53]. Nowadays the world is free of smallpox (except for some strains in high safety level laboratories) demonstrating the great impact of vaccination on mankind. Today, vaccines against many infectious agents are developed and administered to people routinely. However, the efficacy of vaccination differs between targeted infections as well as between younger and older people. The age effect becomes obvious when observing the mortality rates due to infections. About ~ 36,000 influenza-associated deaths occur annually in the United States and approximately 90% of these deaths occur among people aged 65 years or older [54].

Older people suffer from an increasing vulnerability to newly emerging pathogens [55] and there is a correlation between age and decline in vaccination efficacy [56]. The changes in the immune system occurring during an individuals lifespan are described as immune senescence. Immune senescence plays a more and more important role in biomedical research, as the number of older people in the world is constantly increasing. Multiple changes in various parts of the immune system occur during aging. In the innate immune system, such changes are observed in e.g. toll-like receptor expression and cytokine expression [57]. In the adaptive immune system, the B-cells and respective antibody expression, as well as T-cells may be affected. Starting with involution of the thymus in young adults [58], aging affects all stages of T-cell response including generation, maturation, differentiation, activation and functionality of T-cells [59].

The increasing morbidity and mortality rates in people over 65 years of age in respect to infections is of serious concern, especially in the light of reduced vaccination efficacy. Influenza virus infections, for example, regularly lead to serious health problems and in some cases to death. Vaccination against influenza prevents most of the people from illness but necessitates annual immunization to achieve protection against current virus strains. Especially for older people, immunization is highly recommended due to increased vulnerability and severity of health problems after infection. With increasing age (>65 years), however, the efficacy of vaccination decreases [59-61]. Whereas efficacy in young adults is relatively high with 70-90%, in elderly over 65 the efficacy decreases to 17-53% [62, 63]. While older people are still able to produce immunoglobulins with a high affinity for viral antigens, they do not produce these in sufficient amounts [64]. It was also observed that the antibody titers against vaccines in the age group >50 years were already reduced compared to young controls [65, 66]. The age related decrease is not due to a lack of specific antibodies because of reduced V(D)J recombination capability, but rather a problem of the

antibody titer and lack of specificity in the right immunoglobulin class to elicit an adequate response [64]. One explanation for this observation are problems related to CSR, i.e. reduced class switching from IgM to a more specific and potent isotype, such as IgG upon exposure to the antigen/vaccine [67]. Furthermore, the number of antigen specific memory B-cells after vaccination is also reduced in elderly resulting in reduced antibody titers, which is suggested to be due to impaired $CD4^+$ helper T-cell response [61]. Changes in the immune system and respective repertoires in regard to aging have been reported in mice and humans [59, 68]. Although this is of major relevance for the understanding of reduced vaccination efficacy in coherence to aging, no such data is currently available and, therefore, requires further investigation. This coins the topic of my thesis, wherein in-depth analyses of antibody repertoires of healthy donors representing different age groups are analyzed.

1.3 Next generation sequencing

Analysis of the immunoglobulin repertoire based on the expressed mRNA would not be possible without powerful sequencing technology. Since 1977, when the first genome of a bacteriophage was entirely sequenced, the technological possibilities changed completely [69, 70]. High-throughput DNA-sequencing is revolutionizing current experimental methodologies in terms of amounts of obtained DNA-sequences from a single experiment, speed and cost [71]. This technological innovation, which is generally referred to as "next generation sequencing" (NGS) and especially the so-called pyrosequencing method were the basis of the present thesis.

DNA came into focus in 1953, when James D. Watson and Francis Crick discovered the three dimensional structure of the DNA [72] and when finally Marshal Nirenberg decoded the genetic code in 1962 [73]. 15 years later, Frederick Sanger published - next to other groups - his method of chain termination for DNA sequencing [74]. The Sanger method for DNA sequencing dominated the field for the last three decades. In 1995, it was possible to sequence the complete genome of the *Haemophilus influenza* virus for the first time [75], and with continuing improvements of the method, it became possible to sequence the whole human genome [76, 77]. Today, the so called next generation sequencing methods are in place and have completely changed the way of DNA-sequencing as well as the surrounding experimental approaches. At the moment a variety of next generation sequencing systems are available, which all possess different strengths, bottlenecks and characteristics. These methods include among others Illumina/Solexa Genome Analyzer, Applied Biosystems SOLiD System, ION Torrent, Pacific Bioscience SMRT (PacBio) and Roche/454 FLX; each with their own advantages and disadvantages regarding accuracy, read length and amount of sequences (table 1.2).

Table 1.2: Characteristics of selected NGS Technologies.

Platform	Illumina Miseq	Illumina HiSeq	Roche GS FLX Titanium +[5]	Ion Torrent PGM	PacBio RS
Instrument Coast[1]	$128K	$654K		$80K[2]	$695K
Sequence yield per run	1,5-2GB	600GB	700 MB	1GB	100MB
Sequence coast per GB	$502	$41	$6000[6]	$1000	$2000
Run time	27 h[3]	11 days	23 h	2 h	2h
Reported Accuracy	Mostly > Q30	Mostly > Q30		Mostly Q20	<10
Observed Raw Error Rate	0.80%	0,26%	0,003%	1.71%	12,86%
Read length	up to 150 bases	up to 150 bases	up to 1000 bases shotgun	~200 bases	Average 1500 bases[4]
Paired reads	Yes	Yes	Yes	Yes	No
Typical DNA requeirements	50-1000ng	50-1000ng	50-1000ng	100-1000ng	~1µg

[1] All cost calculations are based on list price quotations obtained from the manufacturers and assume expected sequence yield stated; [2] System price including PGM, server, OneTouch and One Touch ES; [3] Including two hours of cluster generation; [4] mean mapped read length includes adapter and reverse strand sequences. Subread lengths, i.e. the individual stretches of sequence originating from the sequenced fragment, are significantly shorter; [5] according official information; [6] per run for our group; Table modified from Quail *et al*. [78]

Illumina technology

One of the most popular NGS systems to date is the Illumina Genome Analyzer (formerly Solexa). It is based on the work of Turcatti [79, 80] and the so-called “Bridge-PCR” [81] for target amplification. The principle of this method is depicted in figure 1.3. At first, the DNA has to be converted into a special sequencing library possessing defined ends, so-called adapters [69, 82]. Next, the dsDNA is melt into ssDNA using sodium hydroxide and the ssDNA is then distributed in low concentration on a flow cell. Each cell contains two types of onlinucleotides on the surface, which are complimentary to the adapter of the ssDNAs. The

ssDNA binds to the complementary oligonucleotides and becomes covalently linked to the surface by reverse synthesis. This principle is referred to as bridge amplification. Next, the newly synthesized strands get cleaved. These results in all strands having the same orientation and after denaturation the ssDNA cannot align. Repetition of bridge amplification-, cleavage- and denaturation-steps provide ssDNA clusters, which contains amplified DNA fragments of a single species (figure 1.3) [69, 82, 83].

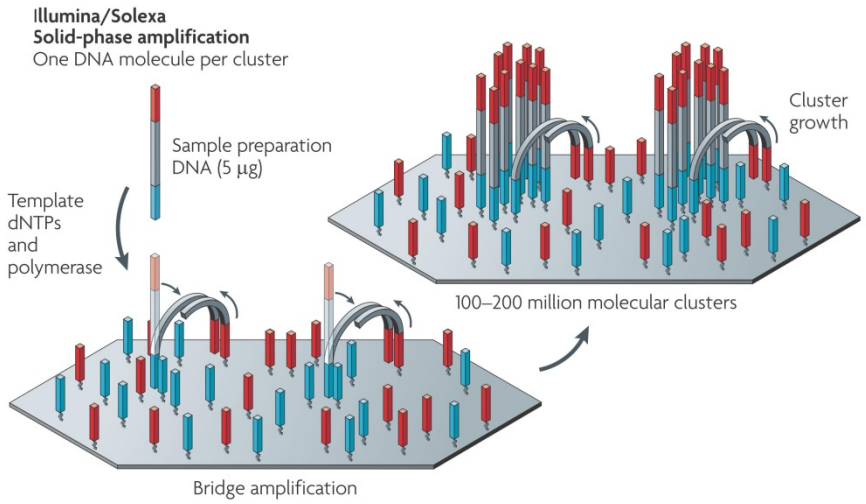


Figure 1.3: DNA immobilization on Illumina/Solexa flow cell and bridge amplification. The ssDNA possessing two different adapters is priming on complementary, surface bound oligonucleotides. Reverse strand synthesis creates a new DNA strand covalently bound to the surface. The free ends can bend to another oligoadapters and synthesis of a new covalently bound strand is carried out. Finally, a cluster of identical ssDNA molecules is created. Figure modified from Metzker *et al.* [84].

The sequencing principle itself is “Sequencing by Synthesis”, first described 1985 [85], which is based on the measurement of dNTP-coupled fluorescence during DNA elongation. For this, all four nucleotides, which are labeled with different fluorescent dyes and removable blocking groups, are added to the cells. One of the nucleotides is incorporated and after washing of the cells, the fluorescence of the corresponding nucleotide is detected. The four color image produced in this process is corresponding to the four different nucleotides. After cleavage of the dye and the terminating groups, a new round of nucleotides is added and the cycle start again (figure 1.4) [84].

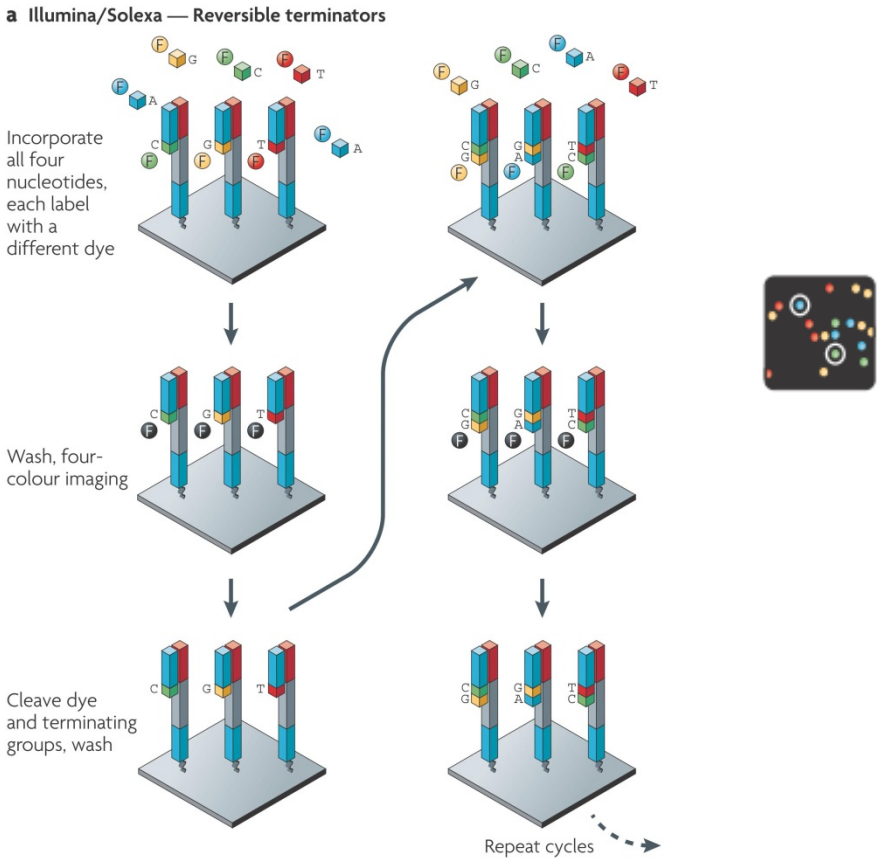


Figure 1.4: Illumina/Solexa four-color reversible termination and imaging method. One of the four different fluorescently-labeled nucleotides is incorporated in the newly synthesized strands. Single templates illustrate here one cluster of clonal amplified ssDNA strands. After washing, a four color image is created. Cleavage of the dye at the terminating group enables a new round of synthesis to commence by adding another nucleotide. Figure modified from Metzker *et al.* [84].

In table 1.2 the benchmarking data of the Illumina HiSeq and the more recent Illumina MiSeq are shown. These sequencers are well-suited for large sequencing projects requiring substantial amounts of data or more defined experiments in clinical settings [86], respectively. While the price per giga base (gb) is low, the sequencing length of ~100 bp (today up to 150 bp) [78] is fairly short. Due to this limitation, Illumina is not the method of choice for antibody sequencing, where read length above 400 bases is required (see below).

SOLiD technology

Another NGS system follows the Sequencing by Oligo Ligation Detection (SOLiD) approach. The Applied Biosystems SOLiD sequencer uses an approach which is in the beginning similar to Roche's (see below), which starts with creation of clonal bead populations in emulsion, bead deposition on a glass slide. Finally, sequencing is carried out by ligation of fluorescently labeled di-base probes, which compete for ligation [87].

In more detail: The DNA molecules are equipped with special adapters, which allow immobilization of the DNA to the bead surface. The clonal bead amplification process is carried out in emulsions and, hence, each bead contains ssDNA of the same template. Afterwards, the beads are immobilized onto glass slides. A sequencing oligonucleotide primer is ligated to the ssDNA molecules in the adapter region and serves as the starting point for sequencing. A mixture of fluorescently-labeled 8-mer probes is added and the probes compete for ligation to the samples. The first two bases of the probes are specific for ligation (out of 16 possible) followed by six degenerated bases and the fluorescent dye. After the fifth base is a defined cleavage site. Probes anneal and get ligated. Unbound probes are washed away and bound probes are identified by their fluorescence. Next, the probe is cleaved and the process starts over and new probes are added to the samples. After five rounds, the newly created strand gets denatured and a new starting primer shifted by one base starts a new ligation round (figure 1.5). Each base is therefore effectively called twice in a sequencing read and increase the accuracy of the approach. Sequences are deduced by interpreting the ligation results of the 16 labeled interrogation probes [69, 83, 88].

The current instrument (SOLiD 3) produces up to 50 gb per slide in the form of 35 to 50 bp reads [89]. The accuracy of the platform is very high achieving up to 99.99%. However, due to the short read length, this method is not suited for sequencing antibody repertoires as planned in this thesis.

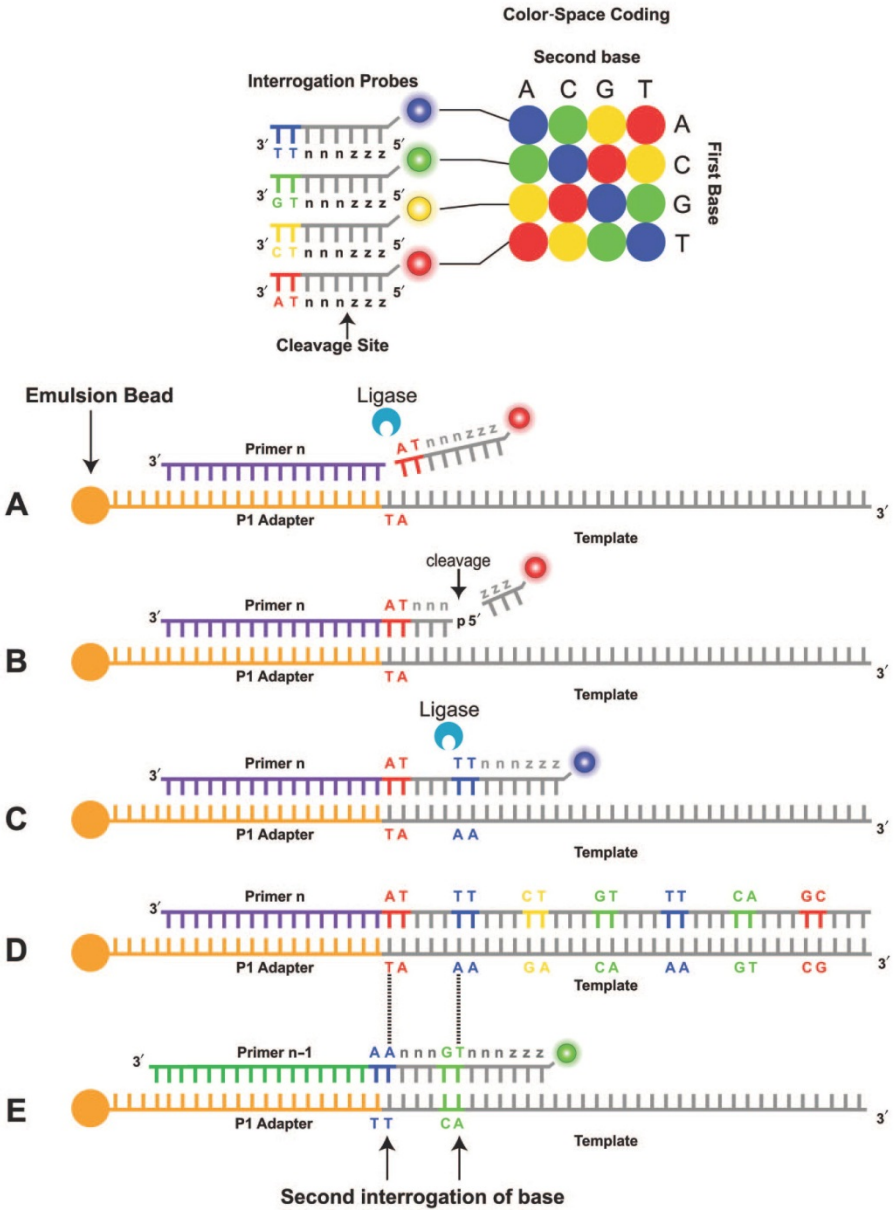


Figure 1.5: SOLiD sequencing by ligation technology. DNA is amplified on beads in emulsion and denatured. Primer n serves as starting point for ligation. **A** Interrogation probes get ligated and fluorescent dye will be detected. **B** Cleavage of the dye and creation of free 5' Phosphate. **C** Annealing and ligation of the next probe. **D** Complete extension of the ligated probes. **E** Denaturation and start with of a new round with primer n-1. Figure from Voelkerding *et al.* [88].

Roche/454 technology

The Roche/454 sequencing technology is also based on the sequencing by synthesis principle. Analogous to the technologies described above, for the sequencing of distinct DNA fragments, it is necessary that they are flanked with defined DNA sequences at their ends – here named Adapter A and B sequences. The target DNA molecules are captured onto the beads through oligonucleotides complementary to the Adapter B sequence, which are covalently linked to the Capture-Beads. Subsequently, the template is amplified in an emulsion PCR. Ideally, only one species of DNA molecule binds to one Capture-Bead and gets amplified. Emulsion based amplification should ensure that each bead is covered with multiple copies from only one template and, finally, 1-5 x 10^7 copies of DNA are bound to each bead (figure 1.6).

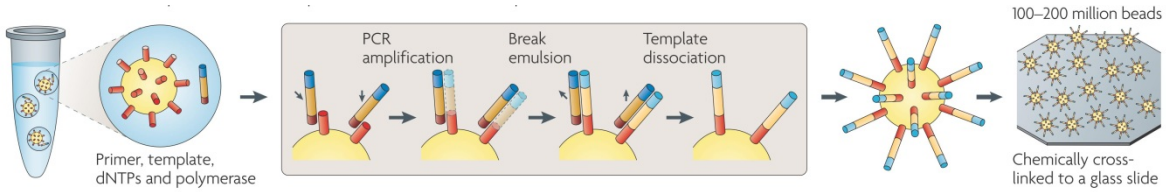


Figure 1.6: Bead loading for Roche/454 Pyrosequencing. DNA is amplified on beads in emulsion. Template DNA is captured onto the surface of a bead by a covalently attached oligo complementary to Adapter B. Monoclonal amplification starts from a single template molecule and, finally, the amplification products cover the complete bead. Template-loaded beads are loaded on PicoTiter™ plates. Figure from Metzker *et al.* [84].

After purification of the beads and template dissociation, the ssDNA-coated beads are loaded on PicoTiter™ plates, which contain cavities fitting only one bead per well (figure 1.7 A). The currently used pyrosequencing method from Roche is based on the release of pyrophosphate during DNA elongation. The complementary primer to Adapter A of the DNA strand (opposite of the bead binding area) is added together with polymerase for DNA elongation. During the sequencing run, each nucleotide group is added separately and consecutively. An enzymatic reaction generates light from inorganic pyrophosphate molecules, which are released by incorporation of nucleotides in the new DNA strand. Small beads cover sulphurylase and luciferase surrounding the template beads. Pyrophosphate is converted by sulphurylase to ATP, which is used by luciferase to generate finally photons [90]. The amount of produced light is detected by a charge-coupled device camera and is directly proportional to the amount of incorporated nucleotides. This feature also allows to distinguish between different quantities of the same nucleotide incorporated at a defined

position in the template DNA chain (figure 1.7 B). Finally, the raw data is bioinformatically quality controlled, analyzed and – if needed – sorted according to individual multiplex identifiers (MIDs). MIDs are 8-10 bases long defined nucleotide sequences at the beginning of each read, which are primer-added. They enable to trace back the sequences to the original sample [71, 84, 91].

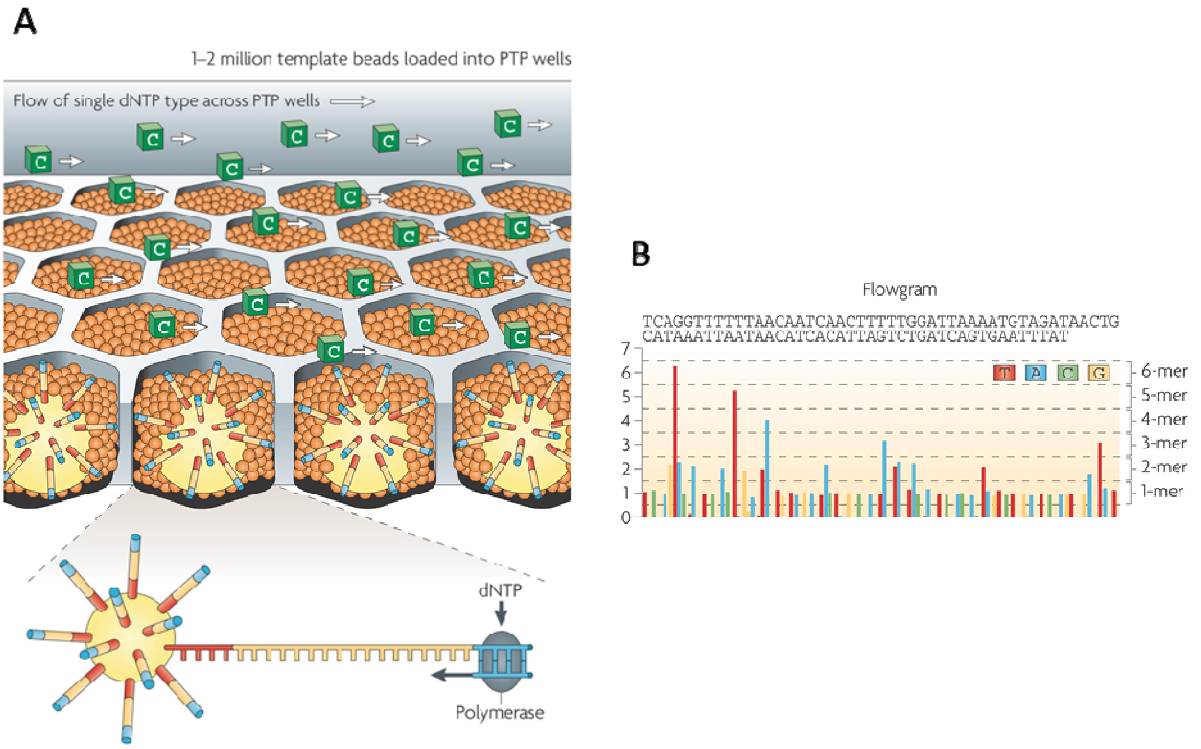


Figure 1.7: PicoTiter plate configuration and description of the pyrosequencing read-out process. (A) Beads with ssDNA amplicons are loaded on a PicoTiter™ Plate device. The surface design only allows one bead to enter per well. **(B)** The incorporated nucleotide at a given position is defined through the amount of enzymatic light reaction measured for each individual cycle during DNA elongation. The quantity of light measured at a given position is directly proportional to the amount of incorporated nucleotides. Figure modified from Metzker *et al.* [84].

Currently, Roche/454 is the most accessible and most robust NGS system available with acceptable read length for immunoglobulin repertoire sequencing studies, such as the one conducted in this thesis. The average read length obtain with the FLX system is >400 bp. With the next generation – FLX plus – read length of ~1000 bp will inevitably also become

available for amplicon sequencing, however, currently the system is only available for shotgun libraries.

Emerging NGS technologies, which can in future offer competitive read length to Roche/454 are Ion Torrent and Pacific Bioscience.

Ion Torrent technology

The Ion Torrent Systems detect protons released during DNA synthesis with a semiconductor sensor [92] and, therefore uses a different approach as Illumina, SOLiD or 454, which all use optical read-out systems. Here, it is possible to monitor the hydrogen ion release directly and not using any optical methodology [92, 93]. The library preparation is similar to other NGS platforms and the sequencing process itself particularly resembles to that of the Roche/454 system. The DNA is fragmented, ligated with adapters and clonal amplification on beads is carried out in an emulsion PCR. Next, beads are loaded in sensor wells located on a wafer. Each of the four bases is added individually and if the nucleotide is incorporated, released protons are detected by an ion-sensitive field-effect transistor due to pH change (0.02 pH units per single base). The number of included nucleotides is directly proportional with the measured shift in pH (figure 1.8). In between the individual cycles, a washing step is introduced to remove remaining nucleotides. During a standard sequencing run, these steps are repeated hundreds of times.

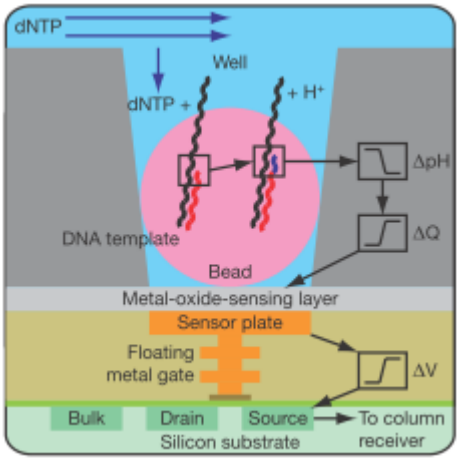


Figure 1.8: Ion Torrent nucleotide incorporation detection system. Simplified drawing of a detection chamber with bead-coupled DNA and subjacent sensor. During the incorporation of a nucleotide into the nascent DNA strand, the released protons (H^+) change the pH, which is subsequently detected by the sensor. Figure from Rothberg *et al.* [92]

Today, the possible reading length is 150 bp [78] but the company has announced that shortly read length of ~400 bp will be reached on ION PGM™ SEQUENCER. The current accuracy is ~98% (see table 1.2).

Pacific Bioscience technology

The sequencing length is ever increasing for many established NGS platforms and new companies, such as Pacific Biosciences, enter the market. They propose a single-molecule sequencing method with a read length of up to 15,000 bp [78]. Pacific Bioscience, or PacBio, has developed a single molecule real time sequencing platform for direct observation of processive DNA polymerization [94]. At the bottom of a nanostructure well, called zero-mode waveguide, a DNA polymerase is coupled via biotin-streptavidin linkage. The Polymerase combines with a single DNA template and starts to synthesize a new strand by incorporating distinctly labeled nucleotides. At the moment when the labeled nucleotide forms a cognate association with the template in the polymerase active site, the distinct fluorescence is recognized due to its longer time spent in the optical detection volume. The distinct fluorescence over the background is measured; the fluorescence-tag is cleaved off during elongation process and diffuses out of the detection area. The process continues by elongation of the new DNA strand through incorporation of the next correspondingly labeled nucleotide (figure 1.9). The sequence can be directly read out during the sequencing process without any delay. Also, chemical modification (e. g. methylation) of the template DNA strand can be detected [71, 94, 95].

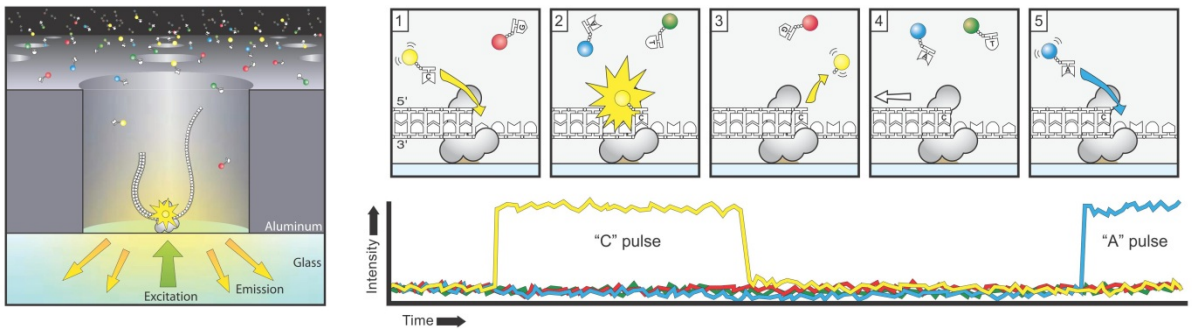


Figure 1.9: Pacific Bioscience real time sequencing methodology. At the bottom of a zero mode waveguide a polymerase is fixed elongation a single DNA template. Fluorescent labeled nucleotides moving around and by incooperating into the new DNA strand the distinct dye is measured. Figure modified from Eid *et al.* [94]

The average read length is stated to be 1,000 bp, with reads up to 15,000 bp being possible. One major drawback of the technology is the high error rate of currently 12,9% [78], which can be reduced by resequencing the template strand multiple times after circularization. Improving the accuracy by circularization reduces the template length but is essential to obtain higher quality reads.

For sequencing antibody cDNAs a long read length is mandatory. Due to the recombination of different genes during V(D)J recombination as well as isotype changes during CSR, it is practically impossible to assemble short sequences into an full antibody-representing sequence. On the one hand, a template for assembly is missing and, therefore, the sequences could end up being artificial. In cases where some parts of the sequences are identical, whereas recombination of other parts, or just fraction of it (like short D segments) are different, may lead to completely artificial assemblies. On the other hand, to ensure correct identification of germline gene segments in affinity matured antibodies, also long read length as well as high accuracy is needed. The high similarities within the different segments also increase the problem of alignment within the immunoglobulin genes themselves and this can only be overcome by covering the whole sequence.

At the time point when the thesis started, the Roche/454 technology was the only one that could offers long enough read length and sequencing accuracy necessary for the project. This technology can cover all information of recombination as well as isotype information in a single read. Therefore, this methodology was chosen for the antibody sequencing in the present work.

1.3.1 Antibody sequencing

In the early days of antibody sequencing, the Sanger method required much experimental effort and individual alignment with germline gene segments from VBASE to identify immunoglobulins [96]. In 1998, Sblattero and Bradbury [97] developed a new degenerate primer set to amplify as many V genes as possible at the time from low amount of template by PCR [98] allowing easier and faster cloning. The analysis of the obtained data also got easier: sequences were analyzed applying online tools and databases, such as Kabat database [99] VBASE2 [100] as well as the IMGT/LIGM database [101, 102]. Each database provides analyses according to specific nomenclatures for V, D and J genes. At this time, the number of analyzed sequences was limited by the speed of sequencing technologies.

Through the introduction of next generation sequencing (NGS) technologies the situation changed and new experimental approaches for the elucidation of immunoglobulin repertoires became possible. NGS allows to obtain thousands or millions of sequences of high quality in a relatively short time at affordable costs. Since its first application in the field of immunoglobulins, notably the analysis of the VDJ recombination patterns of antibodies in the Zebrafish model [22, 103], this novel methodology has been applied to many human studies, such as the monitoring of lymphocyte clonality [104]. Since 2009, also multiple insights into the nature of antibody diversity were provided in an unrivalled depth focusing on specific questions, however, primarily investigating only fractions of the full Ig-repertoire [104-113].

Investigation of the VDJ recombinations lead to a better understanding of the distribution and diversity of antibody recombinations in humans [107, 111, 114]. Paired with the increased throughput of sequencing, online databases and query tools were also adapted to the new requirements of high throughput sequencing. For instance, IMGT/HighV-QUEST was introduced, allowing online processing of NGS data [115]. IMGT/highV-QUEST is especially designed for NGS approaches, is able to analyses up to 150,000 sequences at once and allows insertions/deletions in the sequences, hence, tolerating sequencing errors.

Although the emergence of new sequencing approaches has considerably changed the way B- and T-cell repertoires are analyzed, some restrictions remain [116]. On the one hand, NGS is still very expensive and limits the size and depth experiments can be carried out. On the other hand, DNA sample preparation can introduce biases in the analyzed sample, as currently, standard amplification of Ig-repertoires from mRNA still uses many different V gene specific primers in parallel reactions to ensure completeness [110, 117, 118].

To tackle this issue is one aim of my thesis. In particular, the development of an amplification method for unbiased Ig-repertoires as well as a novel sequencing strategy introducing information of Ig-isotypes is in the center of my efforts.

2. Aims

Aim 1:

Development of a complete and unbiased immunoglobulin amplification method and its adaptation to next generation pyrosequencing technology.

The aim of the thesis was to develop a new method for antibody sequencing. The method has to cover the complete expressed antibody repertoire from human B-cells in peripheral blood. Amplicons ought to contain the recombination pattern from heavy and light chain as well as their isotype affiliation. Further, the method shall avoid any bias or building of chimeric sequences during PCR amplification of the template. A pyrosequencing process to obtained amplicons with minimal amplification steps has to be established and ensure results in a high yield of reads per run with a suitable length for downstream analysis.

Aim 2:

Analysis of the immunoglobulin repertoire in peripheral blood samples of healthy human donors.

A representative collection of human samples has to be analyzed. The obtained sequences have to reflect a cohort of healthy humans of different ages and gender. The analysis of the sequences ought to highlight their V(D)J recombination and their distribution. Furthermore, the new bioinformatic and statistical analysis approach have to take in consideration the isotype assignment to examine the influence of CSR in antibody repertoires.

3. MANUSCRIPT I

V-gene amplification revisited - An optimised procedure for amplification of rearranged human antibody genes of different isotypes.

Lim TS, Mollova S, **Rubelt F**, Sievert V, Dübel S, Lehrach H, Konthur Z.

New Biotechnology. 2010 May 31; 27(2):108-17. Epub 2010 Jan 18.

The analysis and cloning of human antibody repertoire requires a primer set which covers ideally the complete set of immunoglobulin genes. Here, we revisited the known primer sets for V-gene amplification and analyzed whether this set still holds. We found, that some V-genes annotated to be functional are not covered by the existing primer set and, therefore, we extended the set by 1 heavy, 2 kappa and 1 lambda V-gene specific primers. Furthermore, we optimized the amplification conditions and introduced the use of ET SSB (extreme thermo stable single strand binding protein) in the amplification process to reduce possible secondary structures in template cDNA hampering V-gene amplification. The extended primer set covers near to 100% of all functional and putatively functional V genes as annotated in VBASE2.

My contribution to this manuscript was the evaluation of the three novel primers in respect to PCR performance. Furthermore, I conducted the 454 pyrosequencing experiments of the amplified V-genes and contributed to the consecutive sequence analysis verifying amplification success. Moreover, I was involved in writing of the manuscript.

The original article is online available at: http://dx.doi.org/10.1016/j.nbt.2010.01.001

4. MANUSCRIPT II

A streamlined protocol for emulsion polymerase chain reaction and subsequent purification.

Schütze T, **Rubelt F**, Repkow J, Greiner N, Erdmann VA, Lehrach H, Konthur Z, Glökler J.

Analytical Biochemistry. 2011 Mar 1;410(1):155-7. Epub 2010 Nov 25.

In this publication we describe the development of a simple PCR amplification protocol in a water-in-oil emulsion (ePCR). The method is independent of special laboratory equipment and provides a straightforward protocol for ePCR including a fast sample recovery by DNA purification. We show that it is adaptable to many DNA amplification approaches including the creation of cDNA. Additionally, we demonstrate that the application of ePCR on complex mixtures of templates helps to avoid bias during amplification and to reduce the generation of chimeric products, which are frequent problems in conventional PCRs. This is especially a problem in sample preparation for Next Generation Sequencing approaches.

This protocol was later used for the preservation of diversity in enriched SELEX libraries prior NGS sequencing experiments [119] and I have used it to generate antibody amplicon libraries from cDNA.

My contribution to this manuscript was towards method development. I have conducted all experiments showing the application of this protocol on complex cDNA amplification derived from donor material. Furthermore, I was involved in writing of the manuscript.

The original article is online available at: http://dx.doi.org/10.1016/j.ab.2010.11.029

5. MANUSCRIPT III

A universal method for library preparation allowing unidirectional amplicon pyrosequencing

Florian Knaust[§], **Florian Rubelt[§]**, Friederike Braig, Richard Reinhardt and Zoltán Konthur

[§] joint first Authors

This manuscript is currently under consideration at a journal and (in the event of publication) the content of the final article may differ from the here presented version.

In this manuscript a new universal method for unidirectional amplicon sequencing is presented, which is uncoupling the amplicon and the 454-library generation steps. Amplicons are generated in an emulsion PCR and only in a second step specifically designed 454-adaptor sequences are added in a directed fashion by sticky-end ligation. The presented method allows high flexibility in daily lab routine, since direction of sequencing as well as the addition of identifiers needed in sample multiplexing can be decided upon only shortly prior sequencing. We show the application of this method by generating antibody amplicon libraries, for which the standard amplicon generating protocol from Roche has failed. Therefore, we have co-developed and streamlined this novel sequencing protocol for the purpose of antibody sequencing and show antibody cDNA sequencing as the prime example for this sequencing strategy.

My contributions to this manuscript were the preparation of antibody cDNA templates including all optimization steps during amplification, the modifications necessary for optimal purification of the amplicons during library preparation, as well as subsequent evaluation of the sequencing results. Furthermore, I was involved in writing of the manuscript.

A universal method for library preparation allowing unidirectional amplicon pyrosequencing

Florian Knaust[1,§], Florian Rubelt[1,2,§], Friederike Braig[1], Richard Reinhardt[1,3] and Zoltán Konthur[1]*

[1]Max Planck Institute for Molecular Genetics, Ihnestr. 63-73, 14195 Berlin, Germany

*Correspondence: konthur@molgen.mpg.de

ABSTRACT

Background: Amplicon sequencing is the method of choice when large-scale target-specific sequencing information is required. Due to its longer read length, the Roche GS FLX system is preferred. Two standard methods are available for generating amplicon sequencing libraries. In 'rapid-libraries', the amplicons are equipped with 454-adaptor sequences by random ligation resulting in bi-directional sequencing. In '454-amplicon-libraries', the 454-adaptors are incorporated into the amplicons by PCR allowing unidirectional sequencing. While the design of the latter method is simple, the protocol contains various pitfalls and needs optimization for each step and every target region to be addressed.

Results: Here we present an alternative and universal method for directed amplicon sequencing in which the amplicon and the 454-library generation steps are uncoupled, circumventing possible sequence incompatibilities during PCR amplification. The amplicons are generated in an emulsion PCR and only in a second step specifically designed 454-adaptor sequences are added in a directed fashion by sticky-end ligation. As an example, we demonstrate our method by sequencing the variable domains of human immunoglobulin heavy-chain cDNAs. We present data on amplicon and library quality controls and multiplexing results obtained from a GS-FLX Titanium run.

Conclusion: The benefits of our method are multiple. The 454-sequencing-library preparation is largely independent of amplicon-generating primers and is easily adaptable to novel amplicons. Full flexibility in run design is achieved through uncoupling the amplicon-generation from the sequencing step. Through a variety of adaptors, 454-sequencing-libraries can be prepared flexibly in respect to Multiplexidentifier-tagging as well as definition of sequencing orientation – just prior sequencing.

BACKGROUND

With the advance of next generation sequencing technologies (NGS) it is possible to produce up to a billion reads from a single specimen [1]. There are two major approaches for preparation of sequencing libraries: by random shotgun or specific amplification products (amplicon). For determination of whole genomes or transcriptomes the shotgun approach is appropriate. If target-specific sequencing is required, amplicon sequencing is the method of choice. Deep-sequencing of amplicons has been performed in various studies, e.g. focusing on microbial ecology [2], single nucleotide polymorphism discovery [3], mitochondrial DNA heteroplasmy detection [4], or the detection of insertions, deletions and point mutations [5] as well as RNA-editing [6]. While sequencing of amplicons can be performed on various platforms, such as Illumina or SOLID, the Roche 454 GS FLX Instrument utilizing massively parallel pyrosequencing [7] is often chosen due to the longer read length obtained [1].

Roche 454 amplicon sequencing and its limitations

For sequencing amplicons on the GS FLX Instrument, two general approaches are available. During the construction of so-called 'rapid-libraries' the adaptor sequences necessary for 454 pyrosequencing are added to the amplicon by random ligation of sequencing-adaptors resulting in bidirectional sequencing of the amplicon [8]. During the construction of so-called '454-amplicon-libraries' the adaptor sequences are added during amplicon-generation by PCR. Here, the amplicons can be sequenced in a unidirectional or bidirectional manner depending on the adaptor sequences and kits (Roche's Lib-A or Lib L-emPCR-kits) used [9]. Since sequencing-direction in this approach can be defined by primer design, this method is chosen if (i) the amplicon includes problematic areas, such as homopolymer-stretches (e.g. poly-A) on one side of the amplicon, which can dramatically reduce pyrosequencing efficiency, (ii) in cases where the determination of only one end of the amplicon is of scientific interest or (iii) within experiments where a wide diversity of amplicon variants is expected, such as ribosomal 16S RNA sequencing.

For the generation of 454-amplicon-libraries, the target-regions are amplified by PCR with primers consisting target specific sequences at their 3'-end and up to 40 bp long 5'-overhanging ends that include adaptor-sequences necessary for 454-sequencing and Multiplexidentifiers (MIDs) for sample tracking. This approach has several flaws that can spoil successful amplicon-library generation making it hard or even impossible for some applications to obtain a working library. First, the large size of the 5'-overhang can result in amplification difficulties. Second, the 454-adaptor-sequences themselves may cause trouble in the amplification due to incompatibility with the target-specific sequences. Third, the sequencing direction and MID-tagging has to be determined early in the process, already at the amplicon-generation step by PCR. Altogether, this can result in incompatibilities with run-design, for example in cases where not all libraries planned to be sequenced in a single run have passed quality control or where libraries need to be re-sequenced to obtain additional reads and, therefore, multiplexing with new samples would be required. Another major

technical drawback in Roche's amplicon method is the usage of conventional PCR protocols, which are prone to introduce bias of PCR-products [10, 11].

To overcome all these problems, we developed a new method for amplicon sequencing by uncoupling the amplicon and the 454-library generation steps. The amplicons are generated in an emulsion PCR (ePCR)[12] and only in a second step, the assignment of 454-adaptor-sequences (including MIDs) are added by directed ligation just prior sequencing. As an example, we demonstrate our novel method by sequencing the variable domains of human immunoglobulin (Ig) cDNAs. Here we show data on quality controls as well as results obtained from a multiplexed Roche GS-FLX Titanium sequencing run.

Sequencing antibody repertoires – a specific application example

The diversity of the antibody repertoire in humans is primarily a result of a V(D)J recombination mechanism out of several V-, D- and J-genes to form the heavy and light chain sequences of immunoglobulins [13]. For sequencing the Ig heavy chain (IgH) including a short area from the constant part for isotype determination a minimum read length of >380 bp is required and therefore Roche 454 pyrosequencing is the method of choice. Its application results in high numbers of sequences within hours [7] at a lower cost than traditional Sanger-sequencing [14]. To date, all sequencing results published rely on sequence-specific amplification of Ig-repertoires by PCR prior sequencing. One primer is placed in the V-gene, while the other is placed either in the J-gene or the following constant domain. Since the complexity of V- and J-genes is high, normally large sets of degenerate primers need to be used. These can introduce a bias towards certain sequences and at the same time can limit the number of possible V-genes being amplified due to insufficient target-specificity of the primer for certain V-genes [15-18]. To circumvent this problem, we have taken an alternative approach to diminish possible primer-dependent bias and tackled the problem already during first-strand cDNA synthesis and amplification. We used a single V-gene independent 5' end adaptor (PlugOligo) during reverse transcription in combination with 3'-primers derived from the conserved CH1 domain of the Ig. Since PCR-based amplification processes can skew the Ig-repertoire, we applied an emulsion-based method for amplicon-generation to reduce possible bias during amplification of cDNA and maintenance of IgH diversity [19]. In previous studies it was shown that emulsion PCR reduces the synthesis of short, chimeric products and other by-products during amplification. The compartmentalization of template DNA in droplets reduces the competition between fragments of different lengths, thus diminishing the bias for amplifying smaller fragments and decreasing formation of by-products [20-22].

Novel protocol

Here, we present a universal protocol for unidirectional amplicon sequencing as an alternative to the standard amplicon procedure where the 454-adaptor sequences are added by PCR. This method can overcome the above discussed limitations of current strategies. Furthermore, we show that it can

accommodate special requirements, such as the application of a PluginOligo for amplicon generation necessary in Ig-repertoire sequencing set-up. The high flexibility of the new protocol is achieved by uncoupling the amplicon and the 454-library generation steps. The latter step is independent from the amplicon generation step and only requires the amplicons to be equipped with short, *Sfi*I restriction enzyme cleavage sites added by primers. *Sfi*I is a type II restriction enzyme with the recognition motif (RE-site) GGCCNNNN↑NGGCC (where N represents any base and ↑ the point of cleavage) and, hence, can produce differently designed 3'-overhangs of 3 bases length providing flexibility for downstream treatment. The simplistic requirements on primer design for amplicon generation in combination with ePCR can significantly increase the quality of the sample to be sequenced. Additionally, our method can be easily adapted to other scientific tasks and targets, as long as these short primer overhangs are kept constant. The sequencing orientation can be chosen by specific ligation of pre-designed 454-adaptors compatible to one or the other generated overhang by the *Sfi*I enzyme.

METHODS

To allow unidirectional sequencing, in our protocol the 454-adaptor sequences are added to the amplicon by directed ligation, and not by PCR as long 5'-overhangs, such as in Roche's 454-amplicon-library generation protocol. The workflow of our protocol is depicted in Figure 1 and comprises the following steps: First, amplicon generation includes first strand cDNA synthesis, ePCR amplification with *Sfi*I-containing primers and gel-purification. Second, 454-library generation includes enzymatic cleavage of the amplicon DNA, ligation of 454-adaptors and bead-based purification. Finally, the library enters the standard Roche GS-FLX Titanium sequencing procedure.

Template generation and ePCR amplification

Reverse transcription of human mRNA (Figure 1A) from peripheral blood mononuclear cells was performed using the MINT cDNA synthesis kit with an $oligo_d$(T) primer (Evrogen, Russia). Through the application of a poly-G-adaptor (PluginOligo), the mRNA is elongated and, hence, the generated cDNA is elongated by a defined 5'-end (Figure 1B). Next, the amplicon is generated in an ePCR. As forward primer, a *Sfi*I-RE-site elongated PluginOligo (5' TATCAGGCCGAGGCGGCC-PluginOligo 3'), while as reverse primer an equimolar pool of IgH-class specific primers with a *Sfi*I-RE-site (5' AGAGTGGCCATTACGGCC-IgH 3') were used (Figure 1C). ePCRs were assembled using Phusion Hot Start DNA polymerase from Finnzymes (Finland) in combination with the Micellula DNA Emulsion & Purification Kit from Roboklon (Germany) and primers at a final concentration of 10 pmol/µl. ePCRs were conducted on a MJ Research PTC-225 thermocycler with following conditions: initial heating at 98°C for 45 sec and 15 cycles elongation (98°C for 10 sec, 65°C for 20 sec and 72°C for 22 sec) and finally 72°C for 4 min. The ePCR protocol was repeated using the complete elongation product of the first ePCR as template (additional 15 cycles) to increase the amount of DNA. The DNA was separated

by 1.2% agarose gel electrophoresis and subsequently gel-purified (Seqlab, Germany). The resulting PCR-product is now flanked by *Sfi*I-sites (Figure 1D).

_Sfi_I template-restriction

Enzymatic cleavage of at least 100 ng amplicon DNA was performed in a total volume of 100 µl with 40 units of restriction enzyme *Sfi*I in Buffer 4 (New England Biolabs, USA) in the presence of BSA for 1 h at 50°C using a thermocyler. The cleaved DNA (Figure 1E) was purified using the MinElute PCR Purification Kit from Qiagen (Germany) following the manufacturer's instructions and eluted in 15.75 µl EB Buffer.

Assembly and ligation of specifically designed 454-adaptors

454-adaptors containing Roche-454-A- and –B-sequences, sequencing key GACT, MIDs and SfiA- and –B -compatible overhangs were prepared. For adaptor-A oligonucleotides AdA and AdArc were annealed (here with Roche's MID 20 [23]). For adaptor-B the oligonucleotides AdB and AdBrc were annealed (Table 1). Adaptor-annealing and preparation followed Roche's technical bulletin [24]. All oligonucleotides used were from MWG eurofins (Germany). The assembled adaptors contained complementary sticky ends to the *Sfi*I RE-sites of the amplicon and possess 5'-phosphorylated-ends. Next, the adaptors were ligated to the amplicon (Figure 1F). For ligation, all 15.75 µl of the cleaved and purified amplicon DNA from the step above, 1 µl of 5 mM ATP solution, 0.25 µl of 454-adaptors-A and –B (10µM), 2 µl of 10x ligation buffer (Roche) and 1 µl of diluted T4 ligase from Roche (1:4 with 1x ligation buffer) were added to a total volume of 20 µl, incubated in a thermocycler for 16 h at 4°C, followed by an inactivation step at 65°C for 10 min. Further, the ligation reaction was size-selected and purified twice with ampure xp-beads (Beckman Coulter, USA), using a base cut-off of 500 base pairs (bp). DNA was eluted with 20 µl of Buffer EB. Finally, 1 µl of library was run on an 2100 Bioanalyzer using High Sensitivity DNA-Chip and –Kit for quality-control (Agilent Technologies, USA). Note, that only at this step the sequencing direction and a MID for multiplexing needs to be assigned to the Amplicon (Figure 1G).

454-Sequencing and signal processing

Four libraries were prepared, each with a different MID-sequence, titrated and amplified in equal proportion using the 'GS FLX Titanium LV emPCR Kit (Lib-L)' in one large volume cup. Next, two million beads were sequenced using the 'GS FLX Titanium Sequencing Kit XLR70' on one region of a 'GS FLX Titanium PicoTiterPlate Kit' (PTP) on a GS FLX Instrument (Figure 1H). All kits were purchased from Roche 454 Life Sciences and used according to the manufacturer's protocol. For signal-processing, Roche's standard shotgun signal processing protocol was performed.

RESULTS AND DISCUSSION

Amplicon generation

An ideal 454-amplicon-library contains only the target-specific product gained in the amplification flanked by the necessary 454-adaptor-sequences to archive good sequencing results. Therefore, standard 454-amplicon-library protocols need optimization for each and every target region to be addressed. Within each step, one has to ensure that the output of by-products is reduced as far as possible. Here, we present a more simplified alternative method which allows far more flexibility and quick adaptation to any sequencing target. Instead of long primers with 40 bp 5'-overhangs, we only incorporated short 18 bp 5'-overhangs adding specified *Sfi*I RE-sites. The recognition sites attached to the primers contain nucleotide variants resulting in different 3'-overhangs once cut (Figure 1D). The amplicon was generated in an ePCR and the reaction was purified on a column to remove small DNA fragments and components of the emulsion. Subsequently, the products were separated on an agarose gel, the target-specific band was cut out and the DNA was gel-extracted. This should minimize the amount and number of by-products, which can occur even when the primers are highly target-specific. Size-selection by gel-electrophoresis is particularly important, since any smaller or larger PCR-product will enter the next step of 454-library generation.

In our example we applied this protocol to human Ig-repertoires. Previous studies have always relied on the use of V-gene specific primer sets in combination with Roche's standard protocols. Our intention was to develop an amplification method independent of V-gene specific primers to diminish possible primer-dependent bias during amplification discussed above. Hence, we designed a universal 5'-primer amplifying all cDNAs based on the PluginOligo. The combination of the universal 5'-primer and the specific 3'-primers derived from the conserved CH1 domain of the immunoglobulins proved successful for specific amplification of the target (~700 bp). The primers perfectly tolerated the addition of the *Sfi*I RE-sites and the ePCR resulted in Ig-specific amplicons. An aliquot of the amplicons was run on an Agilent Bioanalyzer 2100 High Sensitivity DNA Chip (Figure 2A). Even though impurities are visible, the target specific product (tsp) representing our amplicon yielded the largest peak and verified the success of the amplification of the targeted IgH region.

454-library production

The generation procedure of the 454-library is universal as long as the amplicons are equipped with the *Sfi*I restriction enzyme cleavage sites described above. We found amplicon amounts as low as 100 ng DNA sufficient for library preparation. At first, the amplicon is digested with restriction enzyme *Sfi*I, cleaving off the flanking parts of the amplicon. *Sfi*I is a tetramer in solution and cuts only when two recognition motifs are bound at once. Szczelkun and Halford have reported a preference for *Sfi*I to bind two recognition motifs in *cis* rather than in *trans* forming looped DNA structures [25]. Hence, amplicons should be predominantly cleaved at both ends. After digestion of the amplicon, cleaved-off fragments (below 40 bp) as well as the restriction enzyme were removed with QIAGEN's MinElute purification kit. Stringent removal of the small fragments is important, as they would compete with

adaptors–A and –B for available sticky-ends during the ligation process reducing the quality of the 454-library preparation ultimately resulting in lower numbers of sequences. In the next step, adaptors–A and –B are ligated. To ensure that all amplicons are provided with adaptors A and B, both are added in excess. Unbound adaptors as well as impurities of smaller size were removed by ampure xp-bead-purification adjusted to remove fragments below 500 bp in size. The exclusion of DNA fragments using ampure xp-beads can be regulated and the selection-range is adjustable by altering the volume of the reaction. Efficient ligation and subsequent removal of the adaptors was analysed on an Agilent Bioanalyzer 2100 High Sensitivity DNA Chip (Figure 2B). Efficient ligation was observed due to the complete shift of the target specific peak through the added 454-adaptor sequences flanking the amplicon. At the same time removal of smaller fragments were observed. While the peaks of small fragments are much lower than the tsp, their ration in terms of molarity is important. Therefore, the final protocol was amended to include a second ampure xp-bead-purification step applying the same cut-off. After the second purification the small fragments are below the detection limit (Figure 2C).

Sequencing on the GS-FLX Instrument

We prepared four different libraries according to our new method, each provided with a different 10 bp long MID during adaptor-ligation to allow multiplexing in a single run. Optimal copy per bead ratios for Roche's bead-based emulsion PCR (emPCR) were determined for each library by titration. Next, the amount of used libraries were adjusted to obtain comparable amounts of reads per library. In total, two million beads obtained after bead recovery were sequenced on one region of a PTP. Signal processing was performed with the default shotgun protocol according to Roche's recommendation for unidirectional reads-amplicon sequencing [9]. This way, reads with poor quality at their 3'-end are trimmed rather than discarded, leading in most cases to a higher number of reads, however, reducing the average read length compared to amplicon signal processing. Nevertheless, we obtained about 410,000 reads with an average and median read length of 378 bp and 429 bp, respectively (Figure 3 and Table 2). The number of obtained sequences is higher than the 375,000 reads expected from a standard amplicon run [9] and the average read length generated by this run is well within the stated 350-450 bp for Titanium chemistry [26]. Here, the biggest fraction of reads of all four libraries is between 401 to 500 bp and 75% to 85% of reads per library are longer than 300 bp (Table 3), emphasizing the good run output. Further analysis revealed that 71% of the MID-sorted sequences above 380bp length contained antibody sequences. None the less, we also observed a ~100 bp fragment, which represents short library products with the Adaptor B sequence at the end of the read. This underlines the necessity of several size selection steps. After only single ampure xp-bead purification, we obtained run results with high amounts of short reads and low total yield (data not shown).

Multiplexing sequencing samples is part of daily lab routines. Applying our protocol, tagging of amplicons with MIDs is independent of their generation by ePCR and, therefore, the design of multiplexed sequencing runs can be made easily. For instance, in cases where older libraries need to be re-run to obtain more sequences, one can now add additional new samples easily by assigning

different MIDs just shortly before preparing the run. In Roche's standard amplicon-protocol, the MIDs are assigned to samples early in the amplification process reducing the flexibility in run-design. Here, we show a run of multiplexed libraries generated with four MID-tagged adaptors on a single PTP-region. Passed filter well reads were split by MIDs with allowed mismatch of zero (Table 3). For emPCR, we have adjusted each of the four libraries to obtain equal amount of sequences. All libraries were successfully sequenced and differ only little from the expected ratio. Meanwhile, we have prepared adaptors containing Titanium MIDs 20 to 40 and have applied them successfully (data not shown) underpinning the high degree of flexibility for multiplexing offered by Roche´s large MID set [23]. We show that multiplexing during emPCR is possible with the new method of library generation.

CONCLUSION

Our novel method offers a wide extension and alternative to existing Roche GS FLX 454-pyrosequencing amplicon protocols. It may not only be applicable in sequencing approaches where standard amplicon-sequencing methods have failed, but also where high flexibility in target-selection and run-design is necessary. We present a new universal method for unidirectional amplicon sequencing by uncoupling the amplicon and the 454-library generation steps circumventing possible sequence incompatibilities during PCR amplification. The amplicons are generated in an ePCR and only in a second step, the 454-adaptor sequences including MIDs are assigned. This requires the amplicons to be equipped with short, *Sfi*I restriction enzyme cleavage sites added by primers to allow directed, sticky-end ligation just prior sequencing. Therefore, the sequencing orientation can be chosen by specific ligation of pre-designed 454-adaptors. Through a variety of adaptors, sequencing-libraries can be prepared flexibly in respect to MID-tagging as well as the choice of sequencing orientation. This offers freedom of choice in run-design and makes our protocol universally applicable. Our method has been demonstrated here to work on human IgH cDNA amplicons but can be easily adapted to any other amplicon approaches on the Roche GS FLX platform. Finally we envisage, that our simple but effective approach can be also easily applied on alternative sequencing systems by redesigning the adaptor sequences, respectively.

ACKNOWLEDGEMENT

The authors would like to acknowledge Sven Klages for his help in processing the sequencing data. Work was funded by the Max Planck Society for the Advancement of Sciences.

AUTHOR DETAILS

[1]Max Planck Institute for Molecular Genetics, Ihnestr. 63-73, 14195 Berlin, Germany

[2]Faculty of Biology, Chemistry, and Pharmacy, Freie Universität Berlin, Takustr. 3, 14195 Berlin, Germany

[3]Max Planck Genome Centre Cologne, Max Planck Institute for Plant Breeding Research, Carl-von-Linné-Weg 10, 50829 Köln, Germany

[§]The authors wish it to be known that, in their opinion, the first 2 authors should be regarded as joint First Authors.

AUTHORS' CONTRIBUTIONS

FK, FR, FB, RR and ZK have all contributed to and participated in drafting this manuscript. FK and FR should be regarded as joint First Authors. All authors read and approved the final manuscript.

REFERENCES

1. Gardner AF, Wang J, Wu W, Karouby J, Li H, Stupi BP, Jack WE, Hersh MN, Metzker ML: **Rapid incorporation kinetics and improved fidelity of a novel class of 3'-OH unblocked reversible terminators**. *Nucleic Acids Res* 2012 .
2. Wittekindt NE, Padhi A, Schuster SC, Qi J, Zhao F, Tomsho LP, Kasson LR, Packard M, Cross P, Poss M: **Nodeomics: pathogen detection in vertebrate lymph nodes using meta-transcriptomics**. *PloS one* 2010, **5**(10):e13432.
3. Bundock PC, Eliott FG, Ablett G, Benson AD, Casu RE, Aitken KS, Henry RJ: **Targeted single nucleotide polymorphism (SNP) discovery in a highly polyploid plant species using 454 sequencing**. *Plant biotechnology journal* 2009, **7**(4):347-354.
4. Holland MM, McQuillan MR, O'Hanlon KA: **Second generation sequencing allows for mtDNA mixture deconvolution and high resolution detection of heteroplasmy**. *Croatian medical journal* 2011, **52**(3):299-313.
5. Grossmann V, Schnittger S, Schindela S, Klein HU, Eder C, Dugas M, Kern W, Haferlach T, Haferlach C, Kohlmann A: **Strategy for robust detection of insertions, deletions, and point mutations in CEBPA, a GC-rich content gene, using 454 next-generation deep-sequencing technology**. *The Journal of molecular diagnostics : JMD* 2011, **13**(2):129-136.
6. Stein PE, Leslie AG, Finch JT, Carrell RW: **Crystal structure of uncleaved ovalbumin at 1.95 A resolution**. *J Mol Biol* 1991, **221**(3):941-959.
7. Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, Berka J, Braverman MS, Chen YJ, Chen Z *et al*: **Genome sequencing in microfabricated high-density picolitre reactors**. *Nature* 2005, **437**(7057):376-380.
8. Roche: **cDNA Rapid Library Preparation Method Manual**; 2010.
9. Roche: **454 Sequencing System Guidelines for Amplicon Experimental Design** 2012.
10. Polz MF, Cavanaugh CM: **Bias in template-to-product ratios in multitemplate PCR**. *Applied and environmental microbiology* 1998, **64**(10):3724-3730.

11. Suzuki MT, Giovannoni SJ: **Bias caused by template annealing in the amplification of mixtures of 16S rRNA genes by PCR**. *Applied and environmental microbiology* 1996, **62**(2):625-630.
12. Nakano M, Komatsu J, Matsuura S, Takashima K, Katsura S, Mizuno A: **Single-molecule PCR using water-in-oil emulsion**. *Journal of biotechnology* 2003, **102**(2):117-124.
13. Schatz DG, Ji Y: **Recombination centres and the orchestration of V(D)J recombination**. *Nat Rev Immunol* 2011, **11**(4):251-263.
14. Metzker ML: **Sequencing technologies - the next generation**. *Nat Rev Genet* 2010, **11**(1):31-46.
15. Zhu Z, Dimitrov DS: **Construction of a large naive human phage-displayed Fab library through one-step cloning**. *Methods Mol Biol* 2009, **525**:129-142, xv.
16. Prabakaran P, Chen W, Singarayan MG, Stewart CC, Streaker E, Feng Y, Dimitrov DS: **Expressed antibody repertoires in human cord blood cells: 454 sequencing and IMGT/HighV-QUEST analysis of germline gene usage, junctional diversity, and somatic mutations**. *Immunogenetics* 2011.
17. Lim TS, Mollova S, Rubelt F, Sievert V, Dübel S, Lehrach H, Konthur Z: **V-gene amplification revisited - An optimised procedure for amplification of rearranged human antibody genes of different isotypes**. *N Biotechnol* 2010, **27**(2):108-117.
18. Sblattero D, Bradbury A: **A definitive set of oligonucleotide primers for amplifying human V regions**. *Immunotechnology* 1998, **3**(4):271-278.
19. Schütze T, Rubelt F, Repkow J, Greiner N, Erdmann VA, Lehrach H, Konthur Z, Glokler J: **A streamlined protocol for emulsion polymerase chain reaction and subsequent purification**. *Anal Biochem* 2011, **410**(1):155-157.
20. Williams R, Peisajovich SG, Miller OJ, Magdassi S, Tawfik DS, Griffiths AD: **Amplification of complex gene libraries by emulsion PCR**. *Nat Methods* 2006, **3**(7):545-550.
21. Shao K, Ding W, Wang F, Li H, Ma D, Wang H: **Emulsion PCR: a high efficient way of PCR amplification of random DNA libraries in aptamer selection**. *PloS one* 2011, **6**(9):e24910.
22. Hori M, Fukano H, Suzuki Y: **Uniform amplification of multiple DNAs by emulsion PCR**. *Biochem Biophys Res Commun* 2007, **352**(2):323-328.
23. Roche: **TCB No. 005-2009**. 2009.
24. Roche: **TCB No. 004-2009**. 2009.
25. Szczelkun MD, Halford SE: **Recombination by resolvase to analyse DNA communications by the SfiI restriction endonuclease**. *The EMBO journal* 1996, **15**(6):1460-1469.
26. Roche: **Sequencing Method Manual**; 2009.

FIGURES

Figure 1. Schematic overview of all library-preparation-steps. (**A**) IgH-mRNA consisting of 150 bp long 5'-untranslated region (grey), followed by about 380 bp VDJ-region (orange), 1500 bp C-region (brown) and it's 3'-untranslated region (grey). (**B**) mRNA and its single-stranded complementary DNA (ss cDNA) with introduced sequences during first strand synthesis (violet). (**C**) Primers (small letters) used in ePCR binding with their 3'-ends to the ss cDNA (capital letters). (**D**) Amplicon of the ePCR, flanked by the primer-introduced *Sfi*I-sites (light and dark blue) and their *Sfi*I-recognition-sites (underlined). (**E**) Amplicon enzymatically cleaved with *Sfi*I. (**F**) Selfmade 454-adaptors containing 454-adaptor-A- (dark green, complementary sequence light green) and –B-sequences (red, complementary sequence orange), both including gact-key-sequence at their 3'-end and the different *Sfi*I-compatible overhangs (light and dark blue). Adaptor-A-sequence is followed by a 10 bp Roche-MID. (**G**) Ligation product including all sequences necessary for 454-sequencing. (**H**) Library bound to capture bead during sequencing with sequencing primer (dark green).

Figure 2. Sizing and quantification of amplicons and 454-libraries. Agilent Bioanalyzer 2100 High Sensitivity DNA Chip profiles after (**A**) ePCR, (**B**) agarose-gel size selection and first ampure xp-beads purification and (**C**) second ampure xp-beads purification. All traces show a peak of the target specific product (tsp) and additionally lower marker (lm) and upper marker peaks (um).

Figure 3. Size-distribution of gact-library-reads visualized by Roche's gsRunBrowser-software.

TABLES

Table 1. Oligonucleotides for novel 454-adaptors containing *Sfi*I–A and –B compatible overhangs.

Oligo name	Sequence
AdA:	5' CCATCTCATCCCTGCGTGTCTCCGACGACTACGACTACAGTTA 3'
AdArc:	5' **P-**CTGTAGTCGTAGTCGTCGGAGACACGCAGGGATGAGATGG 3'
AdB:	5' CCTATCCCCTGTGTGCCTTGGCAGTCGACTAGG 3'
AdBrc:	5' **P-**AGTCGACTGCCAAGGCACACAGGGGATAGG-3'

The MID 20 representing sequences are underlined, 5'-phosphorylated ends are bold.

Table 2. Summary of the GS FLX Titanium run on a selected PTP-Region.

GACT-Library Wells	
Raw Wells	1,018,280
Key Pass Wells	956,547
Passed Filter Wells	410,627
Total Bases	155,179,932
% Passed Filter	42.93
Lenght Average	377.91
Median Reads Lenght	429

Size-distribution output of the gsRunBrowser software for gact-library reads.

Table 3. Analysis of read-length-distribution of four libraries from the same GS FLX Titanium run.

	Titanium MID	*Reads*		*Sequence Length Range Distribution (in bp)*					
		count	fraction	1-100	101-200	201-300	301-400	400-500	>500
Library A	28	115022	30%	11%	6%	8%	29%	42%	4%
Library B	35	90971	24%	12%	4%	7%	17%	54%	6%
Library C	39	72313	18%	4%	4%	7%	20%	59%	6%
Library D	32	107515	28%	12%	4%	6%	13%	58%	7%
non matching MIDs	*	24806	-						

Fraction calculation is performed excluding the non-matching MIDs; displayed read-lengths are reduced by 10 bp of MIDs. Reads are derived from the PTP region described in Table 2.

Figure 1

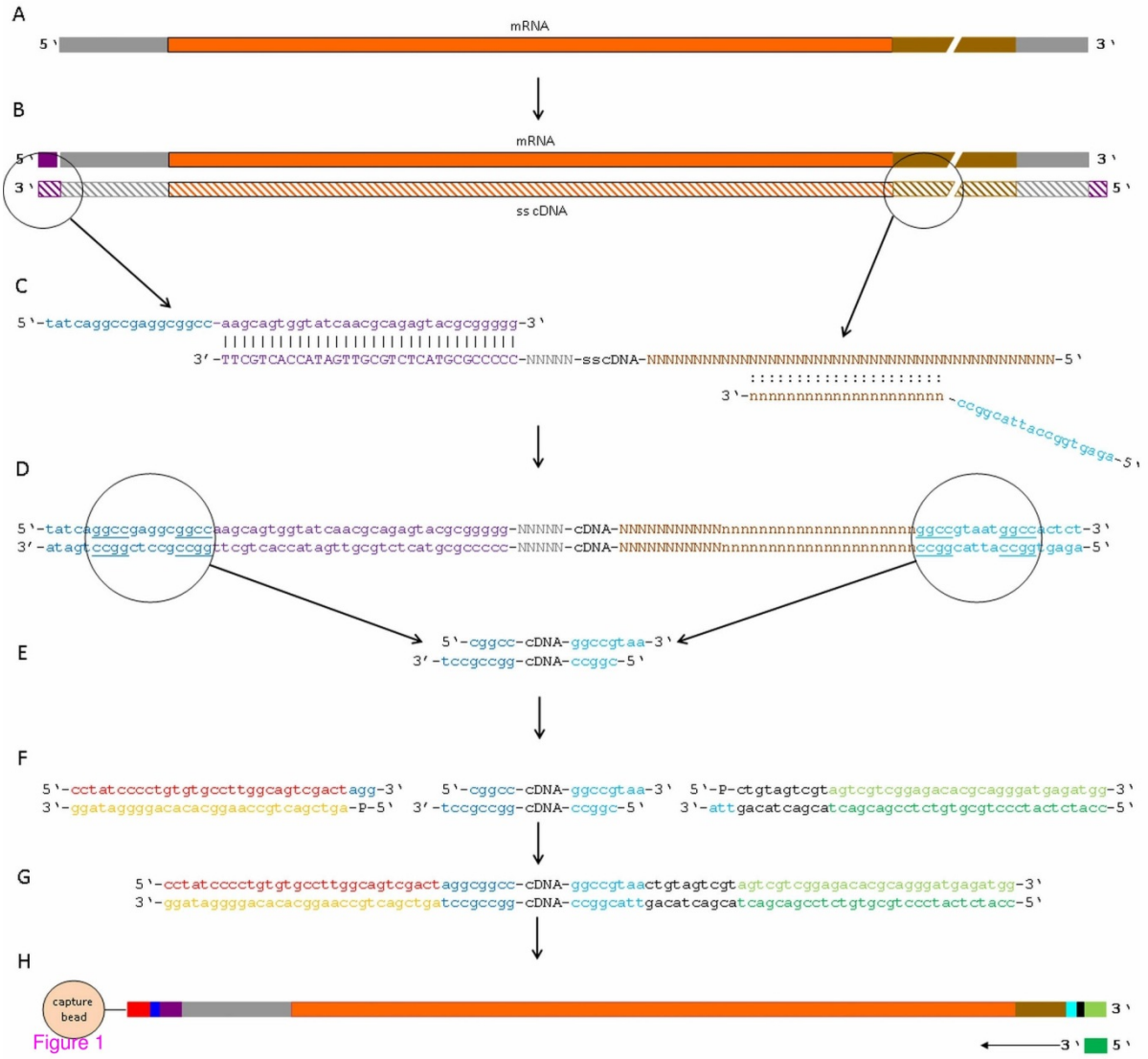

A
mRNA
5'
3'
B
mRNA
ss cDNA
C
5'-tatcaggccgaggcggcc-aagcagtggtatcaacgcagagtacgcggggg-3'
3'-TTCGTCACCATAGTTGCGTCTCATGCGCCCCC-NNNNN-sscDNA-NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN-5'
3'-nnnnnnnnnnnnnnnnnnnn-ccggcattaccggtgaga-5'
D
5'-tatcaggccgaggcggccaagcagtggtatcaacgcagagtacgcggggg-NNNNN-cDNA-NNNNNNNNNNNnnnnnnnnnnnnnnnnnnnnnnnnggccgtaatggccactct-3'
3'-atagtccggctccgccggttcgtcaccatagttgcgtctcatgcgccccc-NNNNN-cDNA-NNNNNNNNNNNnnnnnnnnnnnnnnnnnnnnnnnnccggcattaccggtgaga-5'
E
5'-cggcc-cDNA-ggccgtaa-3'
3'-tccgccgg-cDNA-ccggc-5'
F
5'-cctatcccctgtgtgccttggcagtcgactagg-3'
3'-ggatagggacacacggaaccgtcagctga-P-5'
5'-cggcc-cDNA-ggccgtaa-3'
3'-tccgccgg-cDNA-ccggc-5'
5'-P-ctgtagtcgtagtcgtcggagacacgcagggatgagatgg-3'
3'-attgacatcagcatcagcagcctctgtgcgtccctactctacc-5'
G
5'-cctatcccctgtgtgccttggcagtcgactaggcggcc-cDNA-ggccgtaactgtagtcgtagtcgtcggagacacgcagggatgagatgg-3'
3'-ggatagggacacacggaaccgtcagctgatccgccgg-cDNA-ccggcattgacatcagcatcagcagcctctgtgcgtccctactctacc-5'
H
capture bead
3'
5'
Figure 1

Figure 2

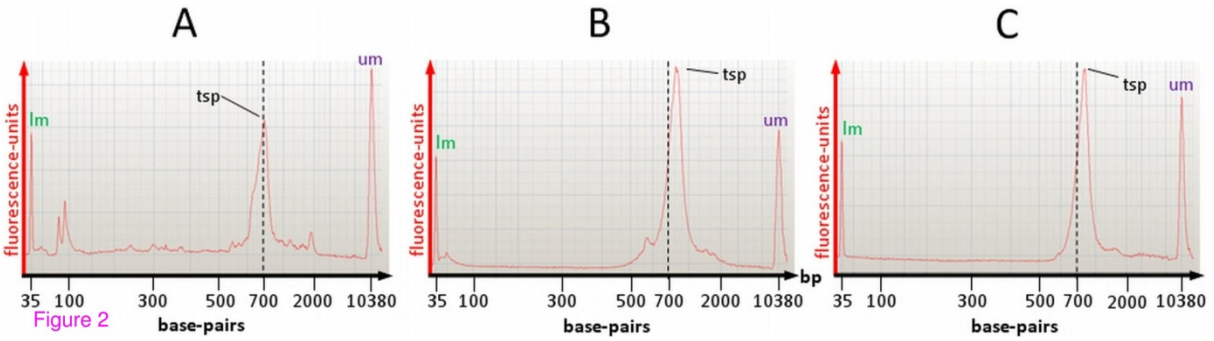
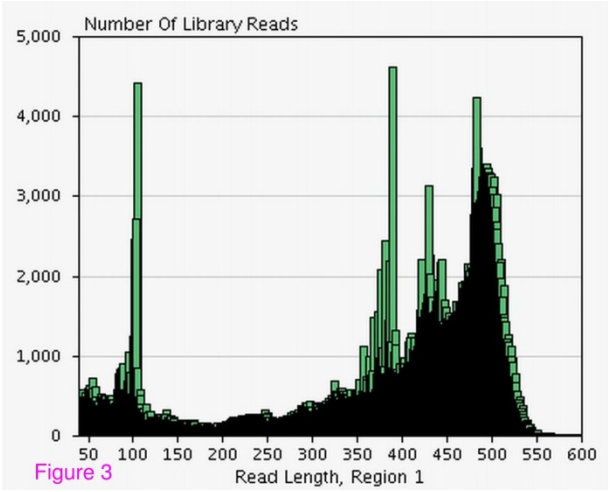


Figure 3



6. MANUSCRIPT IV

Onset of immune senescence defined by unbiased pyrosequencing of human immunoglobulin mRNA repertoires

Florian Rubelt, Volker Sievert, Florian Knaust, Christian Diener, Theam Soon Lim, Karl Skriner, Edda Klipp, Richard Reinhardt, Hans Lehrach, Zoltán Konthur

Here, we present a novel antibody diversity analysis strategy based on 454 pyrosequencing. Our novel avenue of analysis is based not only on information on V(D)J recombination, but also on class switch recombination. For each individual donor, isotype-specific analysis of the antibody recombination patterns were conducted and compared. As a direct consequence, for the first time, donors clustered hierarchically according to age. We could observe changes in immunoglobulin isotype repertoires to be age-dependent and we show that reduction of class switch capability is occurring at a much earlier time point as expected. The age of fifty and beyond already defines the onset of immune senescence and, therefore, demands a redefinition of the term “elderly” in context of the immune system.

My contributions to this manuscript were the preparation of all cDNA and all antibody sequencing libraries necessary for pyrosequencing. I was involved in row data processing and in the statistical analyses of the data together with Dr. Christian Diener as well as the interpretation of the results. Furthermore, I was involved in writing of the manuscript.

The original article is online available since November 30, 2012 at:

http://dx.doi.org/ 10.1371/journal.pone.0049774

Onset of immune senescence defined by unbiased pyrosequencing of human immunoglobulin mRNA repertoires

Florian Rubelt[1,2], Volker Sievert[1], Florian Knaust[1], Christian Diener[1,3], Theam Soon Lim[1,4], Karl Skriner[5], Edda Klipp[3], Richard Reinhard[1,6], Hans Lehrach[1], Zoltán Konthur[1]

[1]Max Planck Institute for Molecular Genetics, Ihnestr. 63-73, 14195 Berlin, Germany

[2]Faculty of Biology, Chemistry, and Pharmacy, Freie Universität Berlin, Takustr. 3, 14195 Berlin, Germany

[3]Theoretische Biophysik, Humboldt-Universität zu Berlin, Invalidenstr. 42, 10115 Berlin, Germany

[4]Institute for Research in Molecular Medicine, Universiti Sains Malaysia, 11800 Penang, Malaysia

[5]Department of Rheumatology and Clinical Immunology, Charité – Universitätsmedizin Berlin, Charitéplatz 1, 10117 Berlin, Germany

[6]Max Planck Genome Centre Cologne, Max Planck Institute for Plant Breeding Research, Carl-von Linné-Weg 10, 50829 Köln, Germany

Abstract

The immune system protects us from foreign substances or pathogens by generating specific antibodies. The variety of immunoglobulin (Ig) paratopes for antigen recognition is a result of the V(D)J rearrangement mechanism, while a fast and efficient immune response is mediated by specific immunoglobulin isotypes obtained through class switch recombination (CSR). To get a better understanding on how antibody-based immune protection works and how it changes with age, the interdependency between these two parameters need to be addressed. Here, we have performed an in depth analysis of antibody repertoires of 14 healthy donors representing different gender and age groups. For this task, we developed a unique pyrosequencing approach, which is able to monitor the expression levels of all immunoglobulin V(D)J recombinations of all isotypes including subtypes in an unbiased and quantitative manner. Our results show that donors have individual immunoglobulin repertoires and cannot be clustered according to V(D)J recombination patterns, neither by age nor gender. However, after incorporating isotype-specific analysis and considering CSR information into hierarchical clustering the situation changes. For the first time the donors cluster according to age and separate into young adults and elderly donors (>50). As a direct consequence, this clustering defines the onset of immune senescence at the age of fifty and beyond. The observed age-dependent reduction of CSR ability proposes a feasible explanation why reduced efficacy of vaccination is seen in the elderly and implies that novel vaccine strategies for the elderly should include the “Golden Agers”.

Introduction

The humoral immune system creates a vast diversity of immunoglobulins (Ig) via rearrangements of variable- (V), diversity- (D; only in heavy chain) and Joining- (J) gene segments [1] to generate a pool of antibodies being able to bind to foreign substances or pathogens (Figure 1). Once an antigen is entering the body, an initial IgM-response is affinity-matured by somatic hypermutation and is finally transferred into an immune response mediated by specific immunoglobulin isotypes obtained through class switch recombination (CSR) [2]. Hence, to get a better understanding of antibody-based immune protection it is not enough to assess V(D)J recombination, but the effector function of an antibody encoded in the isotype is of equal importance. All antibody classes have different functions and the switch from IgM/IgD to a different isotype is a controlled and complex process [3].

In depth analysis of antibody repertoires of healthy donors representing different age groups has not been performed yet although it is of major interest for the understanding of reduced vaccination efficacy in elderly populations [4,5]. Recent findings suggest that the dramatically reduced vaccination efficacy in elderly populations is not because of a lack of specific antibodies due to reduction of V(D)J recombination, but rather a problem in antibody titre and lacking specificity in the right immunoglobulin class to elicit an adequate response [6].

In our study we set out to monitor for the first time V(D)J recombination patterns interrelated with Ig-isotype information on an mRNA level using Next Generation Sequencing (NGS) in an unbiased and quantitative manner. NGS has revolutionized the research on antibody repertoires by providing a before unreached amount of antibody sequences for analysis. NGS was first employed for the analysis of Ig heavy chain repertoires in the Zebrafish model [7,8]. Since then, multiple insights into the nature of antibody diversity has been provided in an unrivalled depth focusing on specific questions, however, primarily investigating only into fractions of the Ig-repertoire [9-18]. Standard amplification of Ig-repertoires from mRNA use many different V-gene specific primers in parallel reactions to ensure completeness [14,19,20]. To diminish possible primer-dependent bias [21], we developed a novel amplification strategy independent of V-gene specific 5’ primers. Further, our novel avenue of analysis is based not only on information on V(D)J recombination but also on CSR profiles of individual donors by incorporating isotype-specific analysis of the antibody sequences. As a direct consequence, donors clustered hierarchically according to age. For the first time we could observe changes in immunoglobulin isotype repertoires to be age-dependent indicating

reduction of class switch recombination ability already occurring at a much earlier time point than expected.

Results and Discussion

Unbiased amplification and sequencing of human Ig-repertoires.

We have developed a novel amplification strategy for heavy and light chain (HC and LC) repertoires starting from total RNA of peripheral blood cells. We used a single V-gene independent 5' end adapter (PlugOligo) during reverse transcription in combination with five HC and two LC PCR primers derived from conserved CH1/CL regions (Figure 1). CH1-specific primers were chosen in such a way that the obtained sequences could be subsequently subdivided into five isotypes with nine subtypes (IgA1, -A2, -D, -E, -G1, -G2, -G3, -G4, -M). Since PCR-based amplification processes can skew the Ig-repertoire, we developed a "single-pot" emulsion-based method for HC and LC amplification to ensure unbiased amplification and maintenance of diversity [21]. DNA sequencing of Ig-repertoires from 14 healthy Caucasians of different age and gender was performed using a Roche Genome Sequencer FLX/454 system [22]. In total 3,566,089 reads were obtained. The raw sequences were analysed according to three criteria (i) over 380 bp length, (ii) unique assignment to Ig-class and (iii) unambiguous assignment of V(D)J rearrangements to V-,D- and J-genes. For assignment of rearrangements we applied the IMGT/High V-Quest tool [23-25] and for class assignment we developed and employed a signature-based method independent of the CH1-specific primers used for amplification. Classification of sequences was performed on a gene rather than allele resolution using regular expression pattern matching of the IMGT/High V-Quest output supplemented with a heuristic approach to handle genes, which could not be identified unambiguously. A relational data model was developed for structural storage retrieval of both raw sequence data and analysis results based on the PostgreSQL RDBMS.

In total, we obtained 1,357,978 (38.08%) sequences with a complete and unambiguous set of V(D)J gene assignments of high quality (1,046,521 HC and 295,555 LC). Our method allows the unbiased analysis of V-gene usage in antibody repertoires and its power is demonstrated by obtaining sequences of several V-genes, e.g. IGHV3-13 and IGHV4-61, which have been missed in previous deep sequencing studies where V-gene specific primers were applied [12]. The analysis of light chain repertoires revealed that the most frequently used kappa and

lambda light chains over all donors are IGKV1-39 and IGKV4-1 (14% and 13%) and IGLV2-14 and IGLV1-40 of (22% and 8%), respectively.

Analysis of V(D)J recombination in healthy donors of different age and gender.

The description of V(D)J recombination frequencies to reflect complete antibody repertoires is inherently complex. Looking only at the combination of individual V-, D- and J-gene segments independent of the resulting antibody sequence, we already found 6685 different VDJ recombination patterns for HC and 240 different VJ recombination patterns in LC. We performed hierarchical clustering of the 14 donors based on the overall distribution and relative frequency of these V(D)J recombination in HC and LC and found neither age nor gender-specific grouping (Figure 2). Next, we included isotype information and repeated the clustering (Text S1). Again, no grouping was observed (Figure S1). At the same time, however, we observed an overall tendency in the usage of certain VDJ recombinations within the 14 individuals of the cohort (Figure 3A) and analysed the most frequent rearrangements; those present 100-times over the median expression of all VDJ recombination (>1.07%; Figure S2). These VDJ recombination patterns were predominantly found in only single isotypes suggesting to have originated from oligoclonal expansion of B-cells and, hence, reflecting natural diversification of specific immune responses. The three most frequent VDJ-rearrangements were analysed in greater detail and strikingly, each of them could almost exclusively be assigned to one major Ig-isotype. Donor I200091-032 showed a distinct recombination pattern (VH4-34/D2-12/J4; 3.4%), of which 96% is expressed as IgA2 (Figure 3B). A detailed analysis of the complementarity determining regions (CDRs) on the amino acid level further revealed that 65% of this VDJ rearrangement can be attributed to two sequences (54% and 11%, respectively) that differ in a single amino acid. This could be either a result of clonal expansion or of a kind of converging maturation in response to an antigen. Donor I200091-030 has an elevated VH2-5/D3-22/J4 recombination pattern (3.8%) with a frequency of 95% in isotype A1 (Figure 3C). Detailed examination of CDR-composition clearly suggests a polyclonal response since an even distribution of seven different amino acid sequences contribute to 60% of this isotype recombination. In the third example (donor I200091-21, Figure 3D), the most frequent recombination VH1-2/D1-26/J3 (3.3%) was observed within IgG1 (89%). At the same time we noticed in this donor a broad usage of VH1-2 (21.6%) with different J and D segments over all V-genes of which the majority was seen in IgG1 (17.6% of all). Altogether, IgG1 expression was 8 times over the median of all

donors in this case indicating a polyclonal or multi-antigenic immune response since no distinct amino acid pattern was observed on the CDR-level. Although none of the donors analysed were vaccinated recently and all were asymptomatic without any recent or long-term medical pre-history according to voluntary disclosure, the determination of the Ig-repertoires by NGS already implies that our method will be suitable to monitor V(D)J recombination patterns in response to specific antigens/vaccines or during the course of certain diseases, as suggested earlier [26].

Ig-isotype frequencies show age-dependency.

Our method results in a quantitative overview of all isotypes and their relative abundance, offering a detailed picture of the immunoglobulin repertoire. At first we calculated for each donor the relative amount of obtained sequences per isotype over the total number of sequences. Most reads belonged to IgM (median 35.7%), whereas IgE- or IgG4-specific reads were rarely obtained. Looking at the cohort as a whole, a correlation between IgM (pval = 0.005) and IgD (pval = 0.029) expression levels and age of the donors could readily be observed (Figure 4A; Table S1). After dividing the donors into young adults (19-30 years) and elderly (49-62 years) we found additionally a significant increase in the IgM (pval = 0.029) and a concomitant decrease in IgG2 (pval = 0.027) levels as well as and an uneven reduction in the other isotypes among the elderly (except IgD and IgG4). At the same time, no correlation between isotype distribution and age was observed in the group of young adults (Tables S2 and S3). We noticed that the antibody repertoire of the 24 years old male resembles more that of elderly with a higher frequency of IgM and IgD than the other young adults.

Differences in CSR-ability are observed between young and elderly.

We calculated the number of unique VDJ recombinations per isotype in proportion to all isotypes for each donor (Figure S3). No age dependency in the young adults was seen and the relative number of unique VDJ recombinations for each isotype is comparable within this group. In the group of elderly, however, age dependency was clearly observed with a gradual increase in relative numbers of VDJ recombination in IgM (pval = 0.006) and IgD (pval = 0.003) and a significant decrease in IgG2 (pval = 0.011). This is in good agreement with the

above finding purely based on isotype analysis and clearly relates to the biological function; i.e. dividing an immune reaction into initial response (IgM/D) and specific response (IgA/E/G) after class switch recombination (Figure 4B). Naïve B-cells initially express IgM or IgD through alternative mRNA splicing and only after stimulation with antigens will the cells undergo CSR of the antigen receptor resulting in IgA/E/G expressing B-cells through a process of DNA rearrangement driven by enzymatic processes [27]. CSR marks the onset of a specific response and results in changes in immunoglobulin effector function while the specificity of the immunoglobulin to the antigen, and hence variable domain usage, remains largely unaffected. We observed in the elderly a strong correlation between age and reduction in CSR ability (correlation 0.95; pval = 0.004) and no correlation in the young adults (pval = 0.663) (Tables S4-S6).

Hierarchical clustering segregates young and elderly.

Finally, we compared the VDJ recombination patterns within the donors with regards to CSR by analysing the overlap between VDJ recombination in the different isotypes. The results of the top 100 VDJ rearrangements between each isotype of a donor were applied to cluster the donors hierarchically. The mean of the overlap significantly differed in the young and elderly groups (p-value = 0.0112; Welch two sample t-test) and a reduction in the elderly of 35.52% (young: 0.1185; elderly: 0.0811) was observed, clearly segregating the two age groups. The heat map in Figure 4C shows that donors clustered according to age but not gender. Below the age of fifty, the donors clustered in pairs age-independently, while above fifty the donors with similar age cluster pairwise suggesting correlation according to reduced CSR ability. Clustering on the basis of single isotypes or group (e.g. IgGs) only, revealed a tendency for age correlation (Figures S4-S7). We conclude that only monitoring the complete repertoire, now possible for the first time, can reveal donor-specific implications of impairment of CSR and shed light into the complexity of immune senescence in the elderly.

The age of fifty and beyond marks the onset of immune senescence.

Age has a strong influence on the repertoire of the isotypes due to onset of reduction of CSR from IgM/IgD to the more specific and effective isotypes IgA and IgG, which are important not only for rapid and efficient immune response to infectious agents [28] but also for the induction of immune protection upon vaccination. Decreased production of vaccine-specific

antibodies rather than VDJ recombination efficiency or avidity is thought to be responsible for overall decrease of vaccine-efficacy [6], presumably due to impairment of $CD4^+$ helper cell responses [5]. We find that immune senescence correlates with reduced CSR ability as a result of reduced transcription of Ig-isotypes and that this process starts around the age of fifty. This is in good agreement with reduced vaccine efficacy in the older population [4-6,29-32]. Our results are also in concordance with recent findings that suggest a decrease in the number of mature activated B-cells and a decreasing ability for CSR in elderly populations above sixty years of age [33]. Mouse experiments suggest the molecular mechanism of ageing in B-cells to be driven by TNF-α and low-grade inflammation increasing with age. Subsequently this has a negative effect on transcription factor E47 and activation-induced cytidine deaminase and hence down-regulates CSR [34]. Noteworthy is the relatively early onset of immune senescence. This change in the immune system suggests itself to be influenced by hormonal changes at this period of life [35]. Our findings on immune senescence starting with the age of fifty calls for further investigation in this direction as well as the development of different treatment and vaccination procedures for the "Golden Ager".

Material and Methods

Medical history of healthy donors

The samples were collected from Caucasian donors living in the Berlin area by in.vent Diagnostica GmbH (Hennigsdorf, Germany) in the course of routine blood donation and represent leftovers from infectious disease screening and, hence, no ethical approval is necessary for this study. According to German Transfusion Law (Transfusionsgesetz - TFG) no ethical approval is necessary if material is collected during such a routine process, since no additional intervention is necessary. However, written informed consent of the donors is mandatory and has been obtained by all individuals for our specific research program. The identity of the donors was made anonymous by in.vent Diagnostica GmbH prior to sample transfer to the Max Planck Institute for Molecular Genetics. Prior blood donation, the donors stated to by free of any symptoms for at least 8 weeks and filled out a questionnaire with 37 questions concerning the following areas: previous vaccinations (< 12 weeks), disease history in the areas of neurology (<10 years), otorhinolaryngology, lung, cardiovascular, liver, gastrointestinal tract (<10 years), pancreas, blood, cancer, kidney, endocrinology, rheumatism, gynecology (<10 years), eyes, serious infections (<10 years), skin, teeth, tropical diseases (e.g. malaria), pregnancy and allergy. Further, medication status was recorded for

antibiotics, heart, blood clotting, diuretics, abstergent agents, glucocorticosteroids, anti-diabetic, thyroid, contraceptive, recreational and other drugs. Additionally, surgical interventions, alcohol consumption and smoking habits were recorded, as well as familial history of severe diseases, such as cancer and autoimmune disorders (Table 1). After questionnaire evaluation, only donors were finally included into the cohort who could answer the majority of these questions with a no. Exclusion criteria were predominantly fixed to disease or medication history, which could influence the immune status of the donor.

Sample preparation and sequencing

Reverse transcription and amplification of the mRNA was performed using the MINT cDNA synthesis kit (Evrogen, Russia). Immunoglobulin amplification was carried out in two independent reactions for heavy and light chain, respectively. Ig-class specific primers were pooled in an equal molar range to a final concentration of 10 pmol/µl to allow chain and donor specific ePCRs. ePCRs conditions were as follows: initial heating at 98°C for 45 sec and 15 cycles elongation (98°C for 10 sec, 65°C for 20 sec and 72°C for 22 sec) and finally 72°C for 4 min. The amplicons were purified according a modified protocol [21] with a DNA purification kit (Roboklon, Germany) and the ePCR was repeated (additional 15 cycles). DNA was purified by 1.2% agarose gel electrophoresis, enzymatically cleaved with SfiI and ligated with self-made Roche454 adaptors containing appropriate SfiI-sites (Knaust et al, submitted). Ligated DNA was repurified by agarose gel electrophoresis and Agencourt AMPure XP. For sequencing, the libraries were treated according to the manufacturer's recommendation: bead-coupled amplification using the "GS FLX Titanium LV emPCR Kit" followed by sequencing applying "GS FLX Titanium Sequencing Kit XLR70" and "GS FLX Titanium PicoTiterPlate Kit".

Pattern search for Ig-isotype assignment

Isotypes were determined using a simplistic pattern matching approach using isotype-specific signatures found in the constant region of the different Ig-class, specifically (Table 2). Sequences of rearranged antibodies were tested against these signatures using the "Fuzznuc" program from the EMBOSS[36] suite (Version 6.1.0) with default parameters. The output was further processed using shell scripts to retrieve sequence identifiers to be saved in our DBMS for further analysis. The pattern IgG2-01-02-03-04 recognizes both IgG2 and IgG4 while IgG4-01-04 is specific for IgG4 alone – the intersection of sequences found by both patterns is considered as IgG4 and the relative complement of IgG2-01-02-03-04 in IgG4-01-04 as IgG2.

V(D)J assignment using IMGT/High V-Quest

Sequences (size selected, >380bp) of the rearranged antibodies obtained were submitted to the IMGT/HighV-QUEST high throughput analysis portal with "allow insertions/deletions" option enabled. Output was filtered using following three steps: (i) We only considered sequences which had a complete set of V(D)J genes with a identity score greater than 85% and were successfully assigned to an Ig-class by pattern matching against isotype-specific signatures. (ii) IMGT/HighV-QUEST generated output at the allele level. For our analyses however, we classified sequences by gene. To convert alleles to genes, we applied a regular expression filter on the IMGT identifiers (Protocol S1). This filter merged different V, D or J-alleles into the respective genes and treated duplicated and unduplicated D-genes as similar. Additionally, a simple heuristic was employed to integrate some genes which were hardly distinguishable from each other and hence were often ambiguously assigned. (iii) As a last quality control, all sequences were excluded, which were assigned to more than one V, D or J-gene, respectively.

Acknowledgements

This work was supported by Max Planck Society for the Advancement of Sciences. TSL acknowledges financial support from the Malaysia Ministry of Higher Education, Higher Institution Center of Excellence (HICoE) Grant (311/CIPPM/4401005). We thank Dr. D. Vanhecke and Dr. J. Woodsmith for critical reading of the manuscript.

References

1. Schatz DG, Ji Y (2011) Recombination centres and the orchestration of V(D)J recombination. Nat Rev Immunol 11: 251-263.
2. Stavnezer J, Guikema JE, Schrader CE (2008) Mechanism and regulation of class switch recombination. Annu Rev Immunol 26: 261-292.
3. Schroeder HW, Jr., Cavacini L (2010) Structure and function of immunoglobulins. J Allergy Clin Immunol 125: S41-52.
4. Frasca D, Blomberg BB (2011) Aging affects human B cell responses. J Clin Immunol 31: 430-435.

5. Aberle JH, Stiasny K, Kundi M, Heinz FX (2012) Mechanistic insights into the impairment of memory B cells and antibody production in the elderly. Age (Dordr).

6. Sasaki S, Sullivan M, Narvaez CF, Holmes TH, Furman D, et al. (2011) Limited efficacy of inactivated influenza vaccine in elderly individuals is associated with decreased production of vaccine-specific antibodies. J Clin Invest 121: 3109-3119.

7. Weinstein JA, Jiang N, White RA, 3rd, Fisher DS, Quake SR (2009) High-throughput sequencing of the zebrafish antibody repertoire. Science 324: 807-810.

8. Jiang N, Weinstein JA, Penland L, White RA, 3rd, Fisher DS, et al. (2011) Determinism and stochasticity during maturation of the zebrafish antibody repertoire. Proc Natl Acad Sci U S A 108: 5348-5353.

9. Wang Y, Jackson KJ, Gaeta B, Pomat W, Siba P, et al. (2011) Genomic screening by 454 pyrosequencing identifies a new human IGHV gene and sixteen other new IGHV allelic variants. Immunogenetics 63: 259-265.

10. Jackson KJ, Wang Y, Gaeta BA, Pomat W, Siba P, et al. (2012) Divergent human populations show extensive shared IGK rearrangements in peripheral blood B cells. Immunogenetics 64: 3-14.

11. Fischer N (2011) Sequencing antibody repertoires: the next generation. MAbs 3: 17-20.

12. Boyd SD, Gaeta BA, Jackson KJ, Fire AZ, Marshall EL, et al. (2010) Individual variation in the germline Ig gene repertoire inferred from variable region gene rearrangements. J Immunol 184: 6986-6992.

13. Glanville J, Zhai W, Berka J, Telman D, Huerta G, et al. (2009) Precise determination of the diversity of a combinatorial antibody library gives insight into the human immunoglobulin repertoire. Proc Natl Acad Sci U S A 106: 20216-20221.

14. Prabakaran P, Chen W, Singarayan MG, Stewart CC, Streaker E, et al. (2012) Expressed antibody repertoires in human cord blood cells: 454 sequencing and IMGT/HighV-QUEST analysis of germline gene usage, junctional diversity, and somatic mutations. Immunogenetics 64: 337-350.

15. Glanville J, Kuo TC, von Budingen HC, Guey L, Berka J, et al. (2011) Naive antibody gene-segment frequencies are heritable and unaltered by chronic lymphocyte ablation. Proc Natl Acad Sci U S A 108: 20066-20071.

16. Wu YC, Kipling D, Leong HS, Martin V, Ademokun AA, et al. (2010) High-throughput immunoglobulin repertoire analysis distinguishes between human IgM memory and switched memory B-cell populations. Blood 116: 1070-1078.

17. Boyd SD, Marshall EL, Merker JD, Maniar JM, Zhang LN, et al. (2009) Measurement and clinical monitoring of human lymphocyte clonality by massively parallel VDJ pyrosequencing. Sci Transl Med 1: 12ra23.

18. Briney BS, Willis JR, Crowe JE, Jr. (2012) Human Peripheral Blood Antibodies with Long HCDR3s Are Established Primarily at Original Recombination Using a Limited Subset of Germline Genes. PLoS One 7: e36750.

19. Zhu Z, Dimitrov DS (2009) Construction of a large naive human phage-displayed Fab library through one-step cloning. Methods Mol Biol 525: 129-142, xv.

20. Lim TS, Mollova S, Rubelt F, Sievert V, Dübel S, et al. (2010) V-gene amplification revisited - An optimised procedure for amplification of rearranged human antibody genes of different isotypes. N Biotechnol 27: 108-117.

21. Schütze T, Rubelt F, Repkow J, Greiner N, Erdmann VA, et al. (2011) A streamlined protocol for emulsion polymerase chain reaction and subsequent purification. Anal Biochem 410: 155-157.

22. Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, et al. (2005) Genome sequencing in microfabricated high-density picolitre reactors. Nature 437: 376-380.

23. Lefranc MP, Giudicelli V, Ginestoux C, Jabado-Michaloud J, Folch G, et al. (2009) IMGT, the international ImMunoGeneTics information system. Nucleic Acids Res 37: D1006-1012.

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

25. Alamyar E, Giudicelli V, Li S, Duroux P, Lefranc MP (2012) IMGT/HighV-QUEST: the IMGT(R) web portal for immunoglobulin (IG) or antibody and T cell receptor (TR) analysis from NGS high throughput and deep sequencing. Immunome Res 8: 26.

26. Arnaout RA (2005) Specificity and overlap in gene segment-defined antibody repertoires. BMC Genomics 6: 148.

27. Maul RW, Saribasak H, Martomo SA, McClure RL, Yang W, et al. (2011) Uracil residues dependent on the deaminase AID in immunoglobulin gene variable and switch regions. Nat Immunol 12: 70-76.

28. Burlington DB, Clements ML, Meiklejohn G, Phelan M, Murphy BR (1983) Hemagglutinin-specific antibody responses in immunoglobulin G, A, and M isotypes as measured by enzyme-linked immunosorbent assay after primary or secondary infection of humans with influenza A virus. Infect Immun 41: 540-545.

29. Gibson KL, Wu YC, Barnett Y, Duggan O, Vaughan R, et al. (2009) B-cell diversity decreases in old age and is correlated with poor health status. Aging Cell 8: 18-25.

30. Weinberger B, Herndler-Brandstetter D, Schwanninger A, Weiskopf D, Grubeck-Loebenstein B (2008) Biology of immune responses to vaccines in elderly persons. Clin Infect Dis 46: 1078-1084.

31. Stiasny K, Aberle JH, Keller M, Grubeck-Loebenstein B, Heinz FX (2012) Age affects quantity but not quality of antibody responses after vaccination with an inactivated flavivirus vaccine against tick-borne encephalitis. PLoS One 7: e34145.

32. Weinberger B, Keller M, Fischer KH, Stiasny K, Neuner C, et al. (2010) Decreased antibody titers and booster responses in tick-borne encephalitis vaccinees aged 50-90 years. Vaccine 28: 3511-3515.

33. Frasca D, Diaz A, Romero M, Landin AM, Blomberg BB (2011) Age effects on B cells and humoral immunity in humans. Ageing Res Rev 10: 330-335.

34. Frasca D, Romero M, Diaz A, Alter-Wolf S, Ratliff M, et al. (2012) A Molecular Mechanism for TNF-alpha-Mediated Downregulation of B Cell Responses. J Immunol 188: 279-286.

35. Gameiro CM, Romao F, Castelo-Branco C (2010) Menopause and aging: changes in the immune system--a review. Maturitas 67: 316-320.

36. Rice P, Longden I, Bleasby A (2000) EMBOSS: the European Molecular Biology Open Software Suite. Trends Genet 16: 276-277.

Figure Legends

Figure 1. Schematic illustration of immunoglobulin IgG and mechanism of V(D)J recombination and our amplification strategy for 454-sequencing. Immunolgobulins recognise antigens via paratopes primarily defined by complementarity determining regions (CDR). CDR 1 and 2 are defined by the V-genes, while CDR 3 is generated through V(D)J recombination. Chromosomal V(D)J rearrangement in the B-cell connect one Variable-gene (V) with one Diversity (D; only HC) and one Joining-gene (J) out of a pool of different V(D)J genes to enable a high diversity of binding affinities within the antibody repertoire. Constant (C) domains specify the induced immune reaction. VH: Variable heavy; VL: Variable light; CH1 Constant heavy 1; CL: Constant light. Amplicon represent fragments generated for pyroseqeucning by emulsion PCR (ePCR). PlugOligo: V-gene independent 5' end adapter for amplification.

Figure 2. Hierarchical clustering of V(D)J recombination pattern distributions. (**A**) Clustering of HC VDJ rearrangements. (**B**) Clustering of LC VJ rearrangements. Heatmaps show relative frequency of V(D)J recombination patterns (columns) versus donors (rows). Blue and pink colours represent male and female, respectively. The age of the donor is recorded on the right. Individual V(D)J counts were normalized by total number of sequences for each donor. Normalized frequencies were log-transformed [F as $\ln(F + 1e\text{-}6)$] and intensity was visualised from black to lime. Row and column dendograms use euclidean distance.

Figure 3. In detail analysis of VDJ rearrangements. (**A**) Overall distribution of VDJ rearrangements in 14 donors. (**B**) IgA2-specific VDJ rearrangements in donor I200091-032. (**C**) IgA1-specific VDJ rearrangements in donor I200091-030. (**D**) IgG1-specific VDJ rearrangements in donor I200091-021. (**E**) Gauge; sphere volumes refer to their respective numerical proportion. Less than 10ppm are represented by a fixed size sphere. Green colour shading indicates number of reads constituting respective recombination. Other colours highlight selected V-genes; blue: IGHV1-2, yellow: IGHV2-5, red: IGHV4-34.

Figure 4. Analyses based on Ig-isotype distributions. (**A**) Relative frequency of isotype-specific sequences within donors sorted left to right according age. (**B**) Relative variability of isotypes on the basis of initial (IgM/D) and specific response (IgA/E/G). Variability: percentage of VDJs covered by a distinct isotype in each donor. Variability V_{AD} for antibody type A and donor D was calculated from the number of occurring VDJs n_{AD} and the total number of occurring VDJs in the donor D n_D as n_{AD}/n_D. (**C**) Clustering of donors (rows) according to coincident appearance of most frequent VDJ rearrangements in their isotypes (column) with age and gender. The hundred most frequently occurring VDJ rearrangements (or less if there were less than hundred) for each donor and isotype were selected. For each donor the overlap between each pair of isotypes was quantified (visualised from black to red) using the formula $n_{both}/\max(n_A, n_B)$, where n_{both} is the number of VDJs present in both sets of VDJs and n_A and n_B represent the sizes of the sets. Blue and pink colours represent male and female, respectively.

Tables

Table 1. Donor details and obtained numbers of sequences for analysis

Donor ID	LC (No. of seq.)	HC (No. of seq.)	age	gender	Height in cm	Weight in kg	BMI[&]	surgeries
I192158 -77	26257	65770	62	m	180	90	27.8	none
I192158-80[*]	15566	63697	49	m	170	70	24.2	none
I192158-95[*]	10763	54175	21	f	164	53	19.7	none
I192158-105[*]	11620	16019	24	m	187	87	24.9	nasal septum, wisdom tooth
I200091-002	755	75863	57	f	175	80	26.1	caecum, biliary
I200091-004	0	62891	30	f	174	65	21.5	ankle
I200091-017	13960	95743	54	f	162	66	25.1	tonsils
I200091-021	19834	128861	20	m	184	83	24.5	none
I200091-023	91528	127206	19	m	187	87	24.9	small surgery at index finger
I200091-024	79339	94201	60	f	167	70	25.1	tonsils, caecum, cyst at ovary
I200091-028	1196	46102	52	m	180	80	24.7	hand, knee, nose, throat, caecum, hernia inguinalis
I200091-030	13105	96308	22	m	180	86	26.5	none
I200091-032	2828	74597	20	f	167	57	20.4	nasal polyposis
I200091-038[§]	8804	45088	28	f	167	70	25.1	none

[&]BMI: body mass index; [*]smoker; [§]familial history of cardiac defect

Table 2. Signature sequences used for the assignment of Ig-classes

Ig-isotype with allele information	DNA-sequence
IgA1-01	GCAGAGGCTCA
IgA2-01-02-03	GTCGAGGCTCA
IgD-01-02	AGCCTTGGTGG
IgE-01-02	GCTCTGTGTGG
IgG2-01-02-03-04	GCTGTGCTCTCGGA
IgG3-01-02	AGAGGTGCTCCTGGAGCA
IgG4-01-04	AGGGCGGCTGTGCTC
IgM-01-03	GCGGATGCACTC
IgKC-01-02-03-04-05	GCAGCCACAGTT
IgLC1-2	GGCGGGAACA
IgLC3-7	GGTGGGAACA

Figure 1

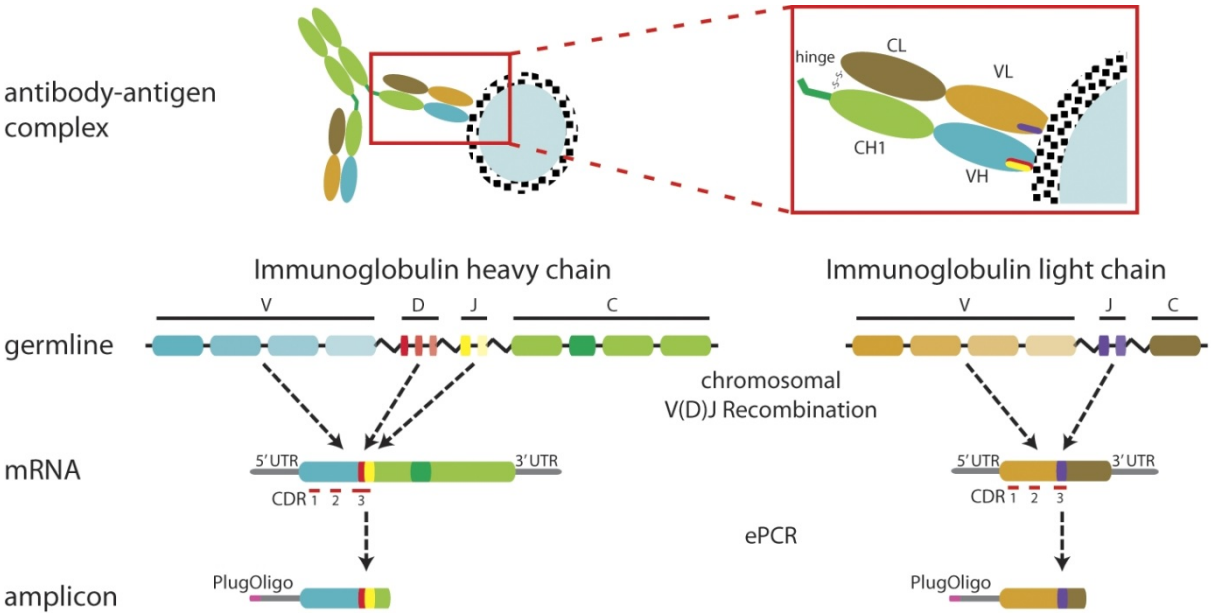

antibody-antigen complex
hinge
CL
VL
CH1
VH
Immunoglobulin heavy chain
Immunoglobulin light chain
V
D
J
C
germline
chromosomal V(D)J Recombination
mRNA
5′UTR
3′UTR
CDR 1 2 3
ePCR
amplicon
PlugOligo

Figure 2

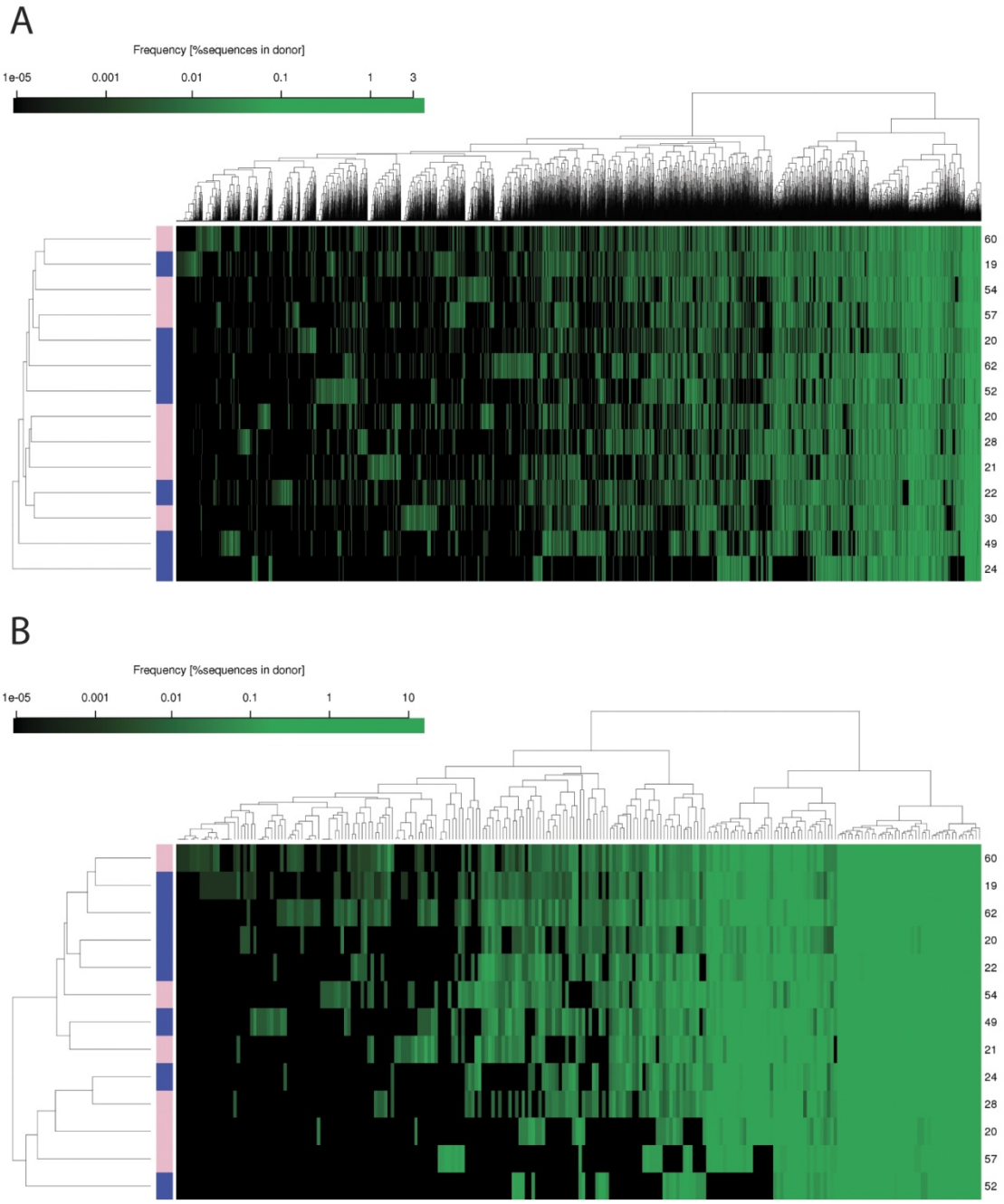

A
Frequency [%sequences in donor]
1e-05
0.001
0.01
0.1
1
3
60
19
54
57
20
62
52
20
28
21
22
30
49
24
B
Frequency [%sequences in donor]
1e-05
0.001
0.01
0.1
1
10
60
19
62
20
22
54
49
21
24
28
20
57
52

Figure 3

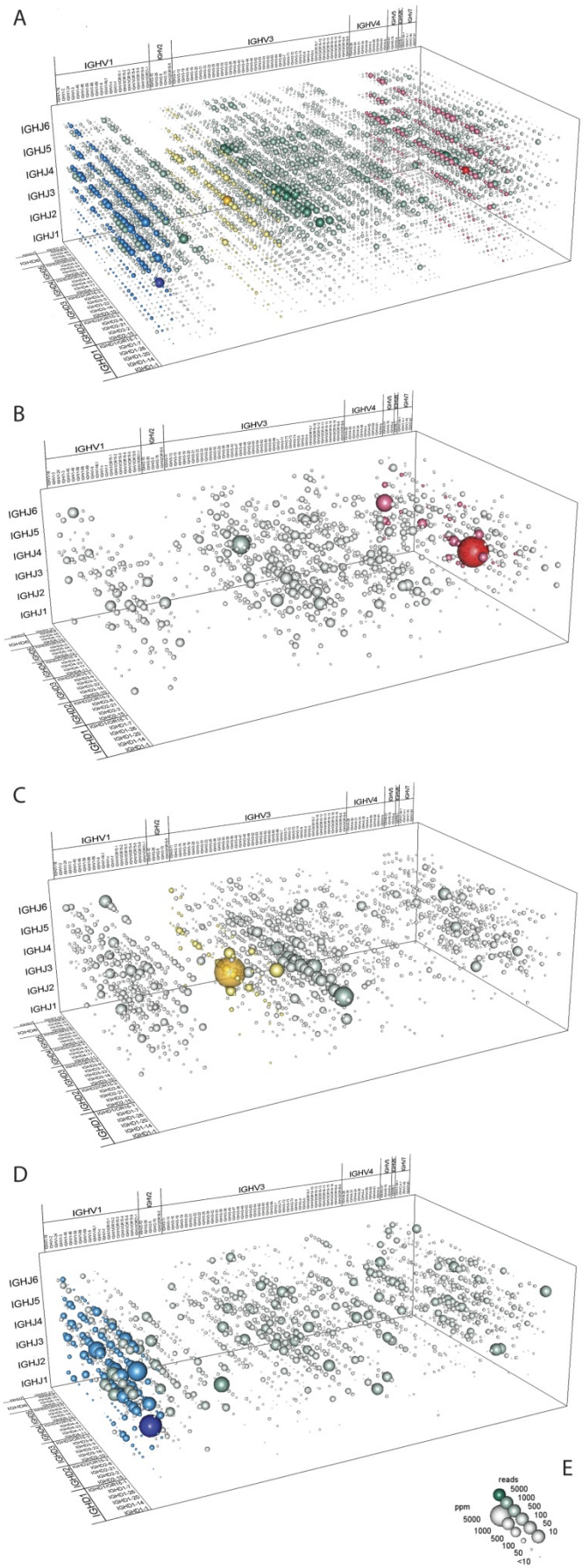
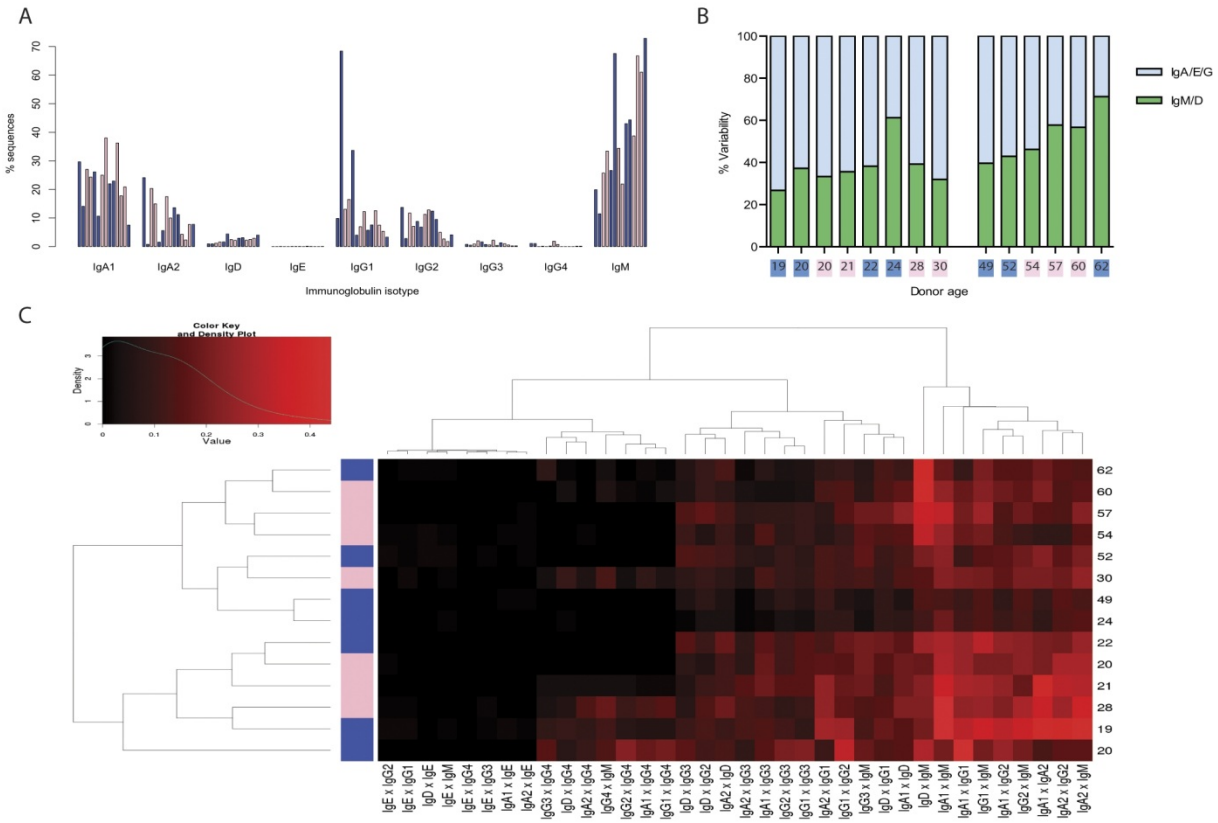

A
B
C
D
E
IGHV1
IGHV2
IGHV3
IGHV4
IGHJ6
IGHJ5
IGHJ4
IGHJ3
IGHJ2
IGHJ1
reads
5000
1000
500
100
50
10
0
ppm
5000
1000
500
100
50
<10

Figure 4


A
% sequences
IgA1
IgA2
IgD
IgE
IgG1
IgG2
IgG3
IgG4
IgM
Immunoglobulin isotype
B
% Variability
IgA/E/G
IgM/D
Donor age
C
Color Key and Density Plot
Density
Value

Supplementary Protocol S1:

Regular expression filter used to integrate IMGT/High V-Quest alleles into genes. Filtering was performed at the database level using the following PL/SQL-function:

```
DECLARE
bfr varchar;
BEGIN
-- this regular expression matches IMGT identifiers without the allele part
bfr := substring(seq,'(IG[KHL][VDJ][0-9]??<1,2??>|IG[KHL][VDJ][0-
9]??<1,2??>D??<0,1??>-[hdcfba0-9]??<1,3??>|IG[KHL][VDJ][0-9]??<1,2??>-[0-
9]??<1,3??>-[0-9]??<0,2??>|IG[KHL][VDJ][0-9]??<1,2??>-NL[0-
9]??<1,3??>|IG[KHL][VDJ][0-9]??<1,2??>/OR[Y0-9]??<0,2??>-[0-9]??<1,3??>)');
-- SPECIAL CASES
-- Some genes turned out to be close variants of others.
-- Such variants were treated as one gene and were merged
-- into their most common "relative" because they often lead to ambigous
-- assignments in IMGT highVQuest runs
IF bfr = 'IGHV3-30-3' THEN
  bfr := 'IGHV3-30';
END IF;
IF bfr = 'IGHV3-NL1' THEN
 bfr := 'IGHV3-30';
END IF;
-- These IGHV4-30 subvariants are hardly distinguishable
-- hence we treat them as generic IGHV4-30
IF bfr = 'IGHV4-30-2' OR bfr = 'IGHV4-30-4' THEN
 bfr := 'IGHV4-30';
END IF;
-- IGHV4/OR15-8 is a variant from papua neuguinea and
-- matches to IGHV4-4 in an imgt search, hence we treat it as IGHV4-4
IF bfr = 'IGHV4/OR15-8' THEN
  bfr := 'IGHV4-4';
END IF;
-- IGHV3/OR16-6 is a variant from papua neuguinea and
-- matches to IGHV3-15 in an imgt search, hence we treat it as IGHV3-15
IF bfr = 'IGHV3/OR16-6' THEN
  bfr := 'IGHV3-15';
END IF;
-- checking for IGKV D variants
-- As we check for presence/absence we do not care for D
-- variants for now.
IF bfr ~ 'IGKV[0-9]*?D.[0-9]+' THEN
   bfr := replace(bfr,'D','');
END IF;
RETURN bfr;
END
```

Supplementary Text S1

Analysis of VDJ recombination patterns with isotype information.

Parallel to the clustering according to heavy chain VDJ recombination patterns described in the main publication, the 14 donors were also clustered with the observed VDJ recombination patterns being previous grouped according to isotypes (Figure S1). This clustering revealed neither grouping according to age nor gender. We found 6,685 unique VDJ recombination patterns (visualised in Figure 3A) of which most occurred only rarely in the different isotypes. Their frequency was analysed and we observed that some VDJ recombinations occurred with more than 100-fold higher frequency within some donors (Figure S2) than the overall median frequency (0.017%). The top three are discussed in the main publication.

Also, the relative number of unique VDJ recombination patterns per isotype in proportion to all isotypes for each donor was calculated (Figure S3). No age dependency in the young adults was seen and the relative number of unique VDJ recombinations for each isotype is comparable within this group. In the elderly however, age dependency was clearly observed with a gradual increase in relative numbers of VDJ recombination in IgM and IgD and a significant decrease in IgG2. The other isotypes in the elderly showed no distinct age-related decrease (see also Tables S4-6). However, a strong correlation became evident by grouping the antibody isotypes in initial response (IgM/D) and specific response (IgA/E/G) after class switch recombination (CSR) as shown in Figure 4B.

Analysis of VDJ recombination patterns with CSR information.

We compared the VDJ recombination patterns within donors under the aspect of CSR by analysing the overlap between VDJ recombinations in the different isotypes. For this, the results of the top 100 VDJ recombination within each isotype were applied to cluster the donors by similarity. First, the overlap between IgM and the other IG classes (IgAs/IgE/IgGs) was analysed (Figure S4). This revealed no gender specific clustering, but a tendency for clustering in young and elderly. Then we compared the VDJ recombination patterns within the donors of the predominant class switch from IgM to IgG (Figure S5). We detected no clustering relation according to gender. In contrast to the previous analysis (Figure S4), age-dependent clustering is not evident. Next, we analysed the overlap between the most frequent 100 VDJ recombination patterns within the donors between all IgG sub-isotypes among themselves and IgM (Figure S6). Again, no clustering relation according to gender was detected, but a similar tendency for clustering in young and elderly was observed as in Figure

S4. Finally, we compared the VDJ recombination patterns within the donors between IgM and IgA1 and IgA2 (Figure S7). Here, clustering the donors by similarity revealed no relation between age or gender.

Frequency comparison of obtained sequences per isotype and their statistical analyses.

With our novel unbiased amplification and sequencing method for the first time a quantitative overview over all isotypes and their relative abundance is possible. We calculated for each donor the relative amount of obtained sequences per isotype over the total number of sequences per donor. Additionally we separated an immune reaction into initial response (IgM/D) and specific response (IgA/E/G) after class switch recombination. To assess possible age dependency, we calculated the relative abundance of all isotypes separately (Figure 4A) and as groups and analysed the relevance of our data statistically (Table S1). Over the total age period only IgM increase is significant and, therefore, also the combination IgM/D. Next, we assessed the age dependency separately in the young and the older adult group (Table S2 and S3, respectively), based on the relative abundance as above. In the young cohort no age dependency was observed and for all isotypes or groups no significant correlation or p-values were obtained. In the older adult group, isotypes IgD, IgM and IgG4 increase with age, while the others decrease. However, only the increase of IgM and IgG4 is significant. Note that the data for IgG4 should be regarded with caution, as the overall number of sequences for this isotype is low in the analysed set. In summary, the relative frequency of IgM sequences in the set is correlated with the increasing age of the donors in the set of elderly.

Changes in the distribution of unique recombination per isotype and their statistical analyses.

The variability as a function of unique VDJ recombinations per isotype in proportion to all isotypes for each donor was calculated (Figure S3). To assess age dependency, the changes of VDJ recombination proportion for each isotype in connection to the age of all donors were calculated as correlation and p-value (Table S4). A significant increase in IgD and IgM and a decrease in IgG2 were observed in the cohort, however, the overall correlation to age is weak. When the analysis was repeated only on the data obtained from the young cohort, no age dependency was observed (Table S5). Next, we analyzed the elderly group on its own (Table S6). This revealed a significant increase and a strong correlation for IgD and IgM and a significant decrease in IgG2 to correlate with increasing age of the donors. The other isotypes

showed mainly a slight, not significant reduction. When analysing the two groups – initial response (IgM/D) and specific response (IgA/E/G) – a strong correlation between increase in IgM/D and decrease in IgA/E/G is seen with age.

Analysis of changes in the VDJ rearrangement pattern distribution by entropy.

Additionally, entropy was used as a measure of dispersion within the distribution of VDJ recombination for individual donors. In order to minimize the influence of sample size (number of available sequences) and the high number of possible recombinations in regimes of small sample sizes entropy was calculated according to the method of Chao and Shen [1], which takes unobserved species into account and performs mostly independent of sample sizes. Because entropy quantifies the maximum information a distribution can comprise, it can be used as another measure for the diversity of the antibody repertoire. Calculations were carried out over all donors as well as over the age groups separately (Tables S7 to S9, respectively). No correlation or significant values were obtained for the young adults. Consistent with the results for the variability, entropy is increasing with age for the IgM and IgD isotypes (IgM, pval = 0.034; IgD, pval = 0.005) and there is a decline in entropy within the IgG2 isotype in the elderly (pval = 0.009). Thus, the diversity of the immune response, particularly in the elderly, is retained within the IgM/D isotype and not adequately transferred to the IgG isotypes, which remain significantly less diverse (smaller entropy). This can be attributed to reduced CSR.

Supplementary Figure Legends

Figure S1. VDJ recombination pattern distributions of 14 donors incorporating heavy chain isotype information. Heatmap shows relative frequency of VDJs recombination patterns in columns versus donors including antibody isotype information in rows. Gender of the donors is represented by blue and pink colours for male and female, respectively. Individual VDJ counts were normalized by total number of sequences for each donor. The distribution of VDJ frequencies showed an exponential distribution. Hence, colouring was applied to a log-transformation of the normalized frequencies F as $\ln(F + 1e\text{-}6)$ and visualised with increasing intensity from black to lime. Row and column dendograms use euclidean distance.

Figure S2. VDJ rearrangements 100-fold over the median frequency of all VDJ recombination patterns within the cohort. Graph represents all individual VDJ rearrangements with a frequency >1.07 % of all donors.

Figure S3. Variability as a function of unique VDJ recombination patterns in each isotype in proportion to all isotypes within donors. Variability was defined as the percentage of VDJs covered by a distinct antibody type in each donor. The variability V_{AD} for antibody type A and donor D was calculated from the number of occurring VDJs n_{AD} and the total number of occurring VDJs in the donor D n_D as n_{AD}/n_D.

Figure S4. Clustering of donors according to coincident appearance of most frequent VDJ rearrangements in IgM with all CSR-dependent isotypes. First, the hundred most frequently occurring VDJ recombination for each donor and antibody isotype (or less if there were less than hundred different VDJs) was determined. For each donor the overlap between each pair of antibodies was quantified using the formula $n_{both}/\max(n_A, n_B)$, where n_{both} is the number of VDJs present in both sets of most frequently occurring VDJs and n_A and n_B are the sizes of the two sets. Gender of the donors is represented by blue and pink colours for male and female, respectively. The age of the donor is recorded on the right. Row and column dendograms use euclidean distance.

Figure S5. Clustering of donors according to coincident appearance of most frequent VDJ rearrangements in IgM and IgG subisotypes. The heatmap was generated as described before, with considering only the shown overlap pairs. Gender of the donors is represented by blue and pink colours for male and female, respectively. The age of the donor is recorded on the right. Row and column dendograms use euclidean distance.

Figure S6. Clustering of donors according to coincident appearance of most frequent VDJ rearrangements in IgG with subisotypes of IgG and with IgM. The heatmap was generated as described before, considering only the shown overlap pairs. Gender of the donors is represented by blue and pink colours for male and female, respectively. The age of the donor is recorded on the right. Row and column dendograms use euclidean distance.

Figure S7. Clustering of donors according to coincident appearance of most frequent VDJ rearrangements in IgM with IgA1 and IgA2. The heatmap was generated as described before, with considering only the shown overlap pairs. Gender of the donors is represented by blue and pink colours for male and female, respectively. The age of the donor is recorded on the right. Row and column dendograms use euclidean distance.

Additional Reference:

1. Chao A, Shen T-J (2003) Nonparametric estimation of Shannon's index of diversity when there are unseen species in sample. Environmental and Ecological Statistics 10: 429-443.

Figure S1

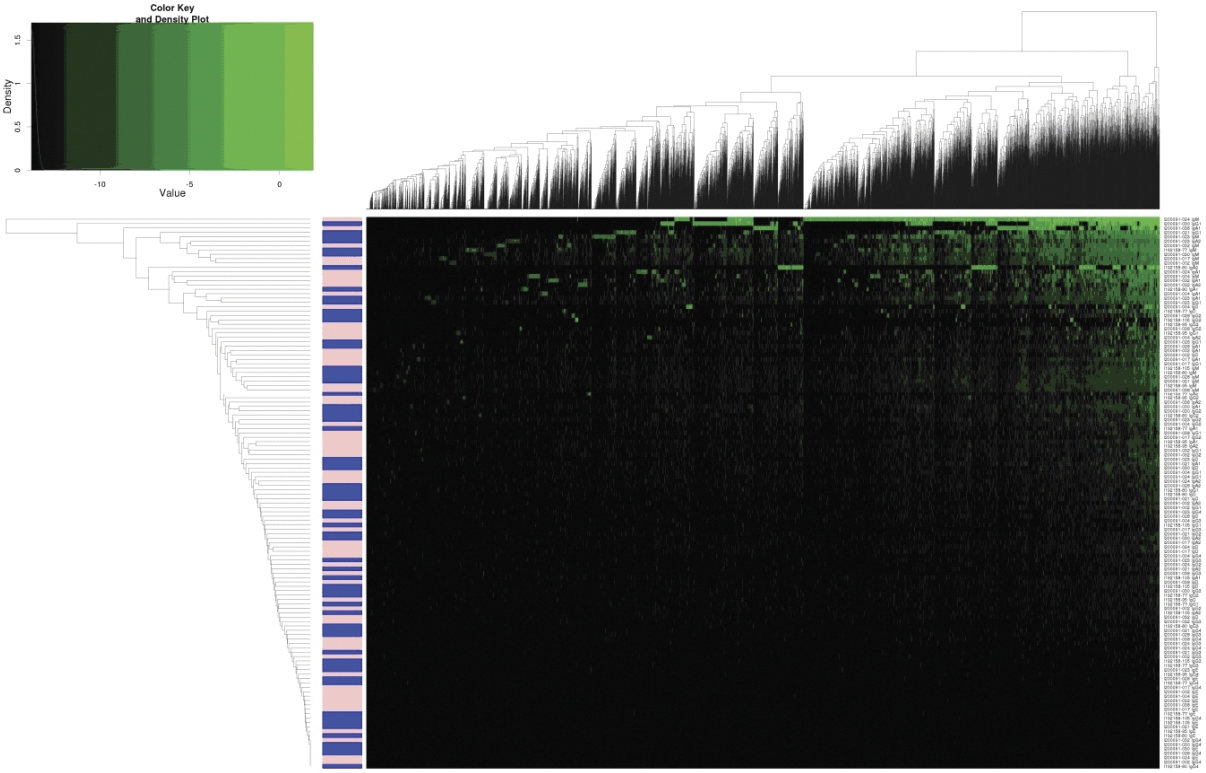


Figure S2

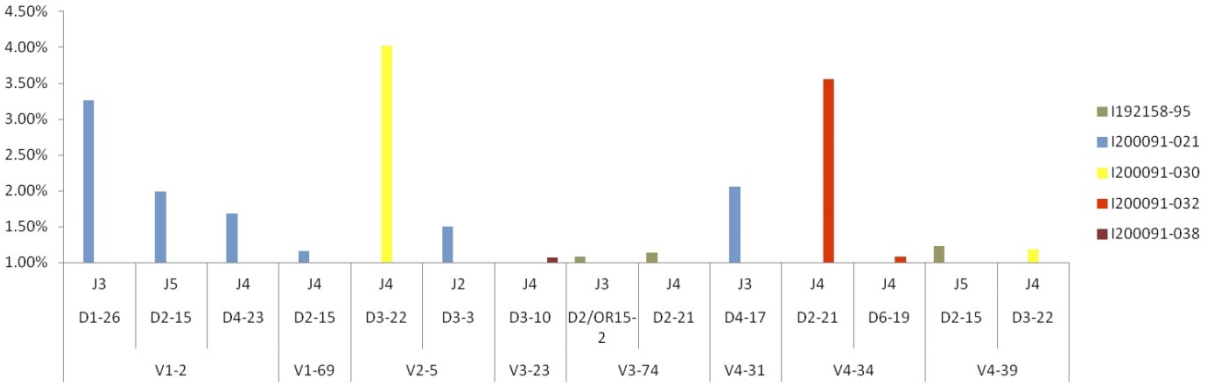


Figure S3

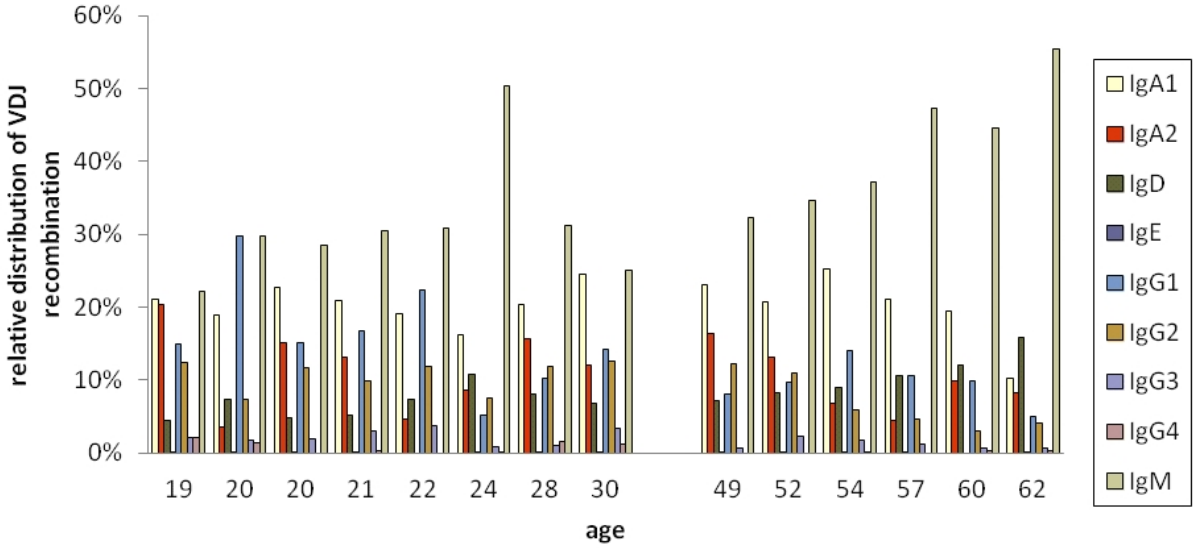

60%
50%
40%
30%
20%
10%
0%
relative distribution of VDJ recombination
19 20 20 21 22 24 28 30 49 52 54 57 60 62
age
IgA1
IgA2
IgD
IgE
IgG1
IgG2
IgG3
IgG4
IgM

Figure S4

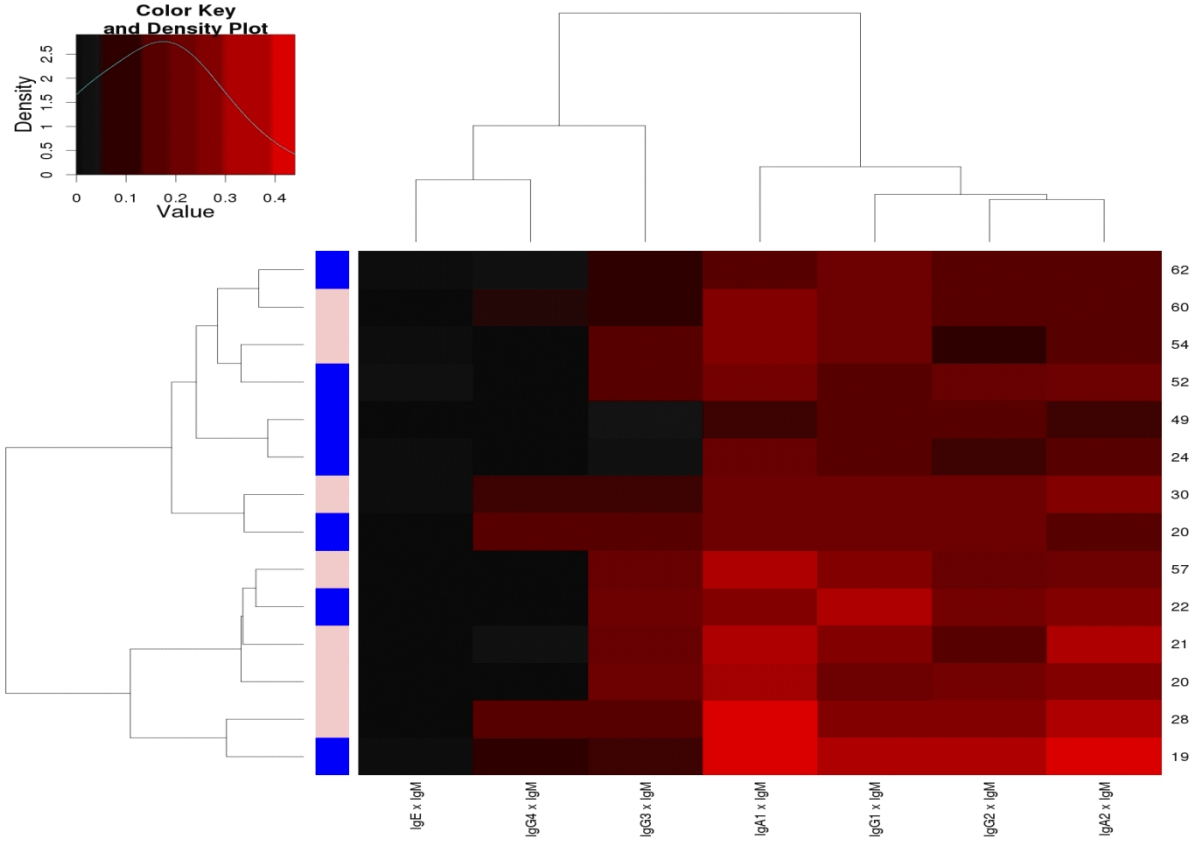

Color Key
and Density Plot
Density
0
0.5
1
1.5
2
2.5
0
0.1
0.2
0.3
0.4
Value
62
60
54
52
49
24
30
20
57
22
21
20
28
19
IgE x IgM
IgG4 x IgM
IgG3 x IgM
IgA1 x IgM
IgG1 x IgM
IgG2 x IgM
IgA2 x IgM

Figure S5

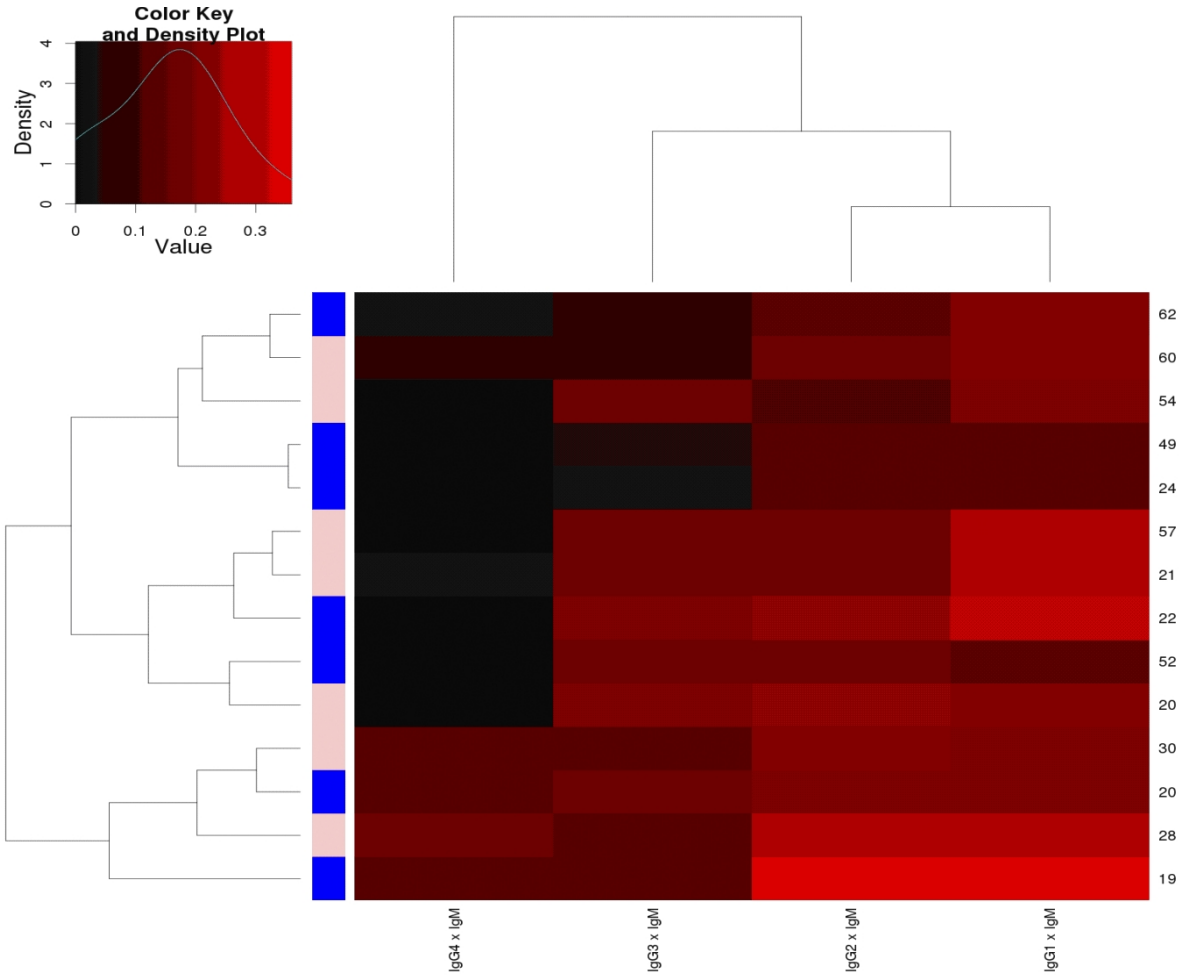

Color Key
and Density Plot
Density
0
1
2
3
4
0
0.1
0.2
0.3
Value
62
60
54
49
24
57
21
22
52
20
30
20
28
19
IgG4 x IgM
IgG3 x IgM
IgG2 x IgM
IgG1 x IgM

Figure S6

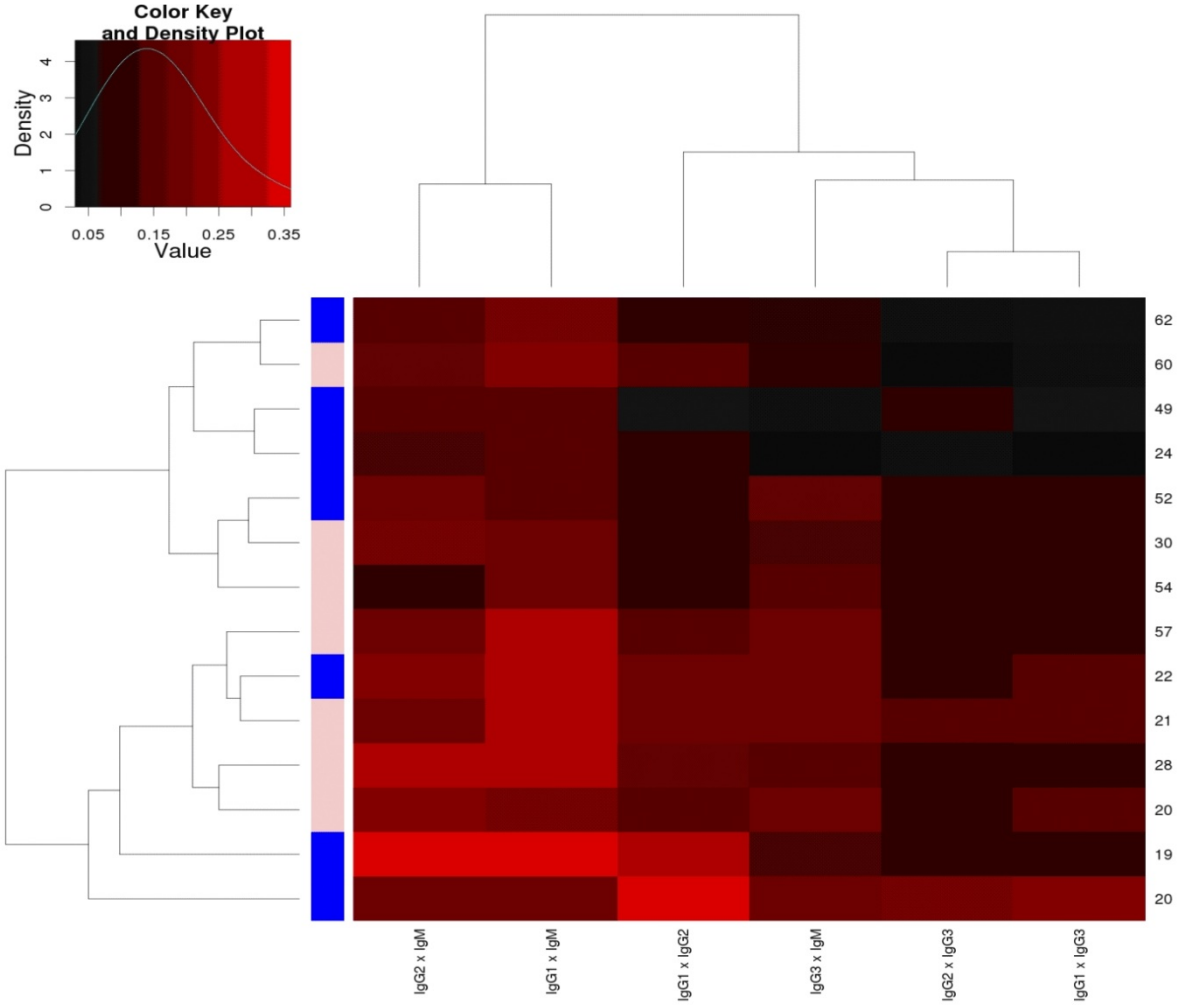

Color Key
and Density Plot
Density
0
1
2
3
4
0.05
0.15
0.25
0.35
Value
62
60
49
24
52
30
54
57
22
21
28
20
19
20
IgG2 x IgM
IgG1 x IgM
IgG1 x IgG2
IgG3 x IgM
IgG2 x IgG3
IgG1 x IgG3

Figure S7

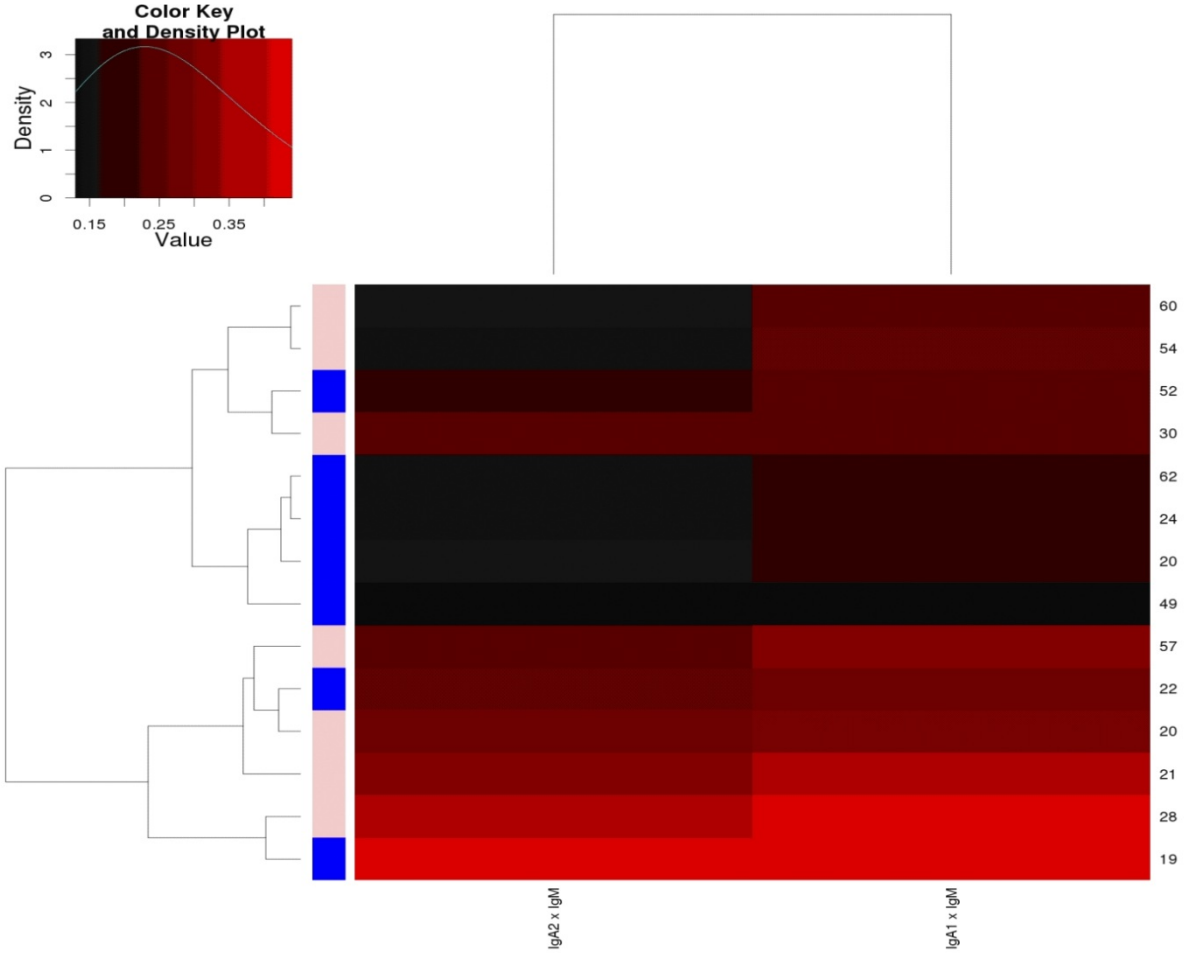

Color Key
and Density Plot
Density
0
1
2
3
0.15
0.25
0.35
Value
60
54
52
30
62
24
20
49
57
22
20
21
28
19
IgA2 x IgM
IgA1 x IgM

Table S1. Statistical analysis of relative amount of obtained sequences per isotype over the total number of sequences from all 14 donors.

isotypes	correlation	p-value
IgA1	-0.20284	0.48675
IgA2	-0.32861	0.25132
IgD	0.58232	0.02889
IgE	0.24094	0.40666
IgG1	-0.45860	0.09908
IgG2	-0.47966	0.08262
IgG3	-0.43607	0.11906
IgG4	-0.44497	0.11086
IgM	0.70830	0.00458
IgM + IgD	0.70644	0.00474
IgA + IgE + IgG	-0.70644	0.00474

For all nine isotypes and groups correlation and p-values according a linear model fit (F-test) were calculated for age dependency.

Table S2. Statistical analysis of relative amount of obtained sequences per isotype over the total number of sequences from the young adult group.

isotype	correlaltion	p-value
IgA1	0.34921	0.39651
IgA2	-0.11493	0.78640
IgD	0.52678	0.17980
IgE	0.12509	0.76790
IgG1	-0.37555	0.35926
IgG2	0.30270	0.46615
IgG3	0.31290	0.45049
IgG4	0.35856	0.38311
IgM	0.22724	0.58837
IgM + IgD	0.24762	0.55434
IgA + IgE + IgG	-0.24762	0.55434

For all nine isotypes and groups correlation and p-values according a linear model fit (F-test) were calculated for age dependency.

Table S3. Statistical analysis of relative amount of obtained sequences per isotype over the total number of sequences from the elderly group.

isotype	correlation	p-value
IgA1	-0.56552	0.24215
IgA2	-0.51642	0.29423
IgD	0.46762	0.34970
IgE	-0.14161	0.78901
IgG1	-0.42357	0.40264
IgG2	-0.86254	0.02704
IgG3	-0.54596	0.26243
IgG4	0.84907	0.03245
IgM	0.85822	0.02873
IgM + IgD	0.85685	0.02927
IgA + IgE + IgG	-0.85685	0.02927

For all nine isotypes and groups correlation and p-values according a linear model fit (F-test) were calculated for age dependency.

Table S4. Unique VDJ recombination per isotype in proportion to all isotypes in all donors.

isotypes	correlation	p-value
IgA1	-0.15983	0.58521
IgA2	-0.23863	0.41131
IgD	0.73313	0.00285
IgE	0.23648	0.26855
IgG1	-0.56822	0.03401
IgG2	-0.64336	0.01305
IgG3	-0.54214	0.04521
IgG4	-0.47527	0.08928
IgM	0.64514	0.01272
IgM + IgD	0.67501	0.00808
IgA + IgE + IgG	-0.67501	0.00808

Correlations were calculated using the Pearson rank method and linear dependencies were evaluated by standard linear model fits. Significance of the intercept term was than quantified with an F-test.

Table S5. Unique VDJ recombination per isotype in proportion to all isotypes in young donors.

isotypes	correlation	p-value
IgA1	0.27138	0.51560
IgA2	0.01206	0.97739
IgD	0.44763	0.26607
IgE	0.57430	0.17752
IgG1	-0.45257	0.26018
IgG2	0.25725	0.53852
IgG3	-0.00097	0.99817
IgG4	-0.02829	0.95758
IgM	0.11277	0.79033
IgM + IgD	0.18371	0.66321
IgA + IgE + IgG	-0.18371	0.66321

Correlations were calculated using the Pearson rank method and linear dependencies were evaluated by standard linear model fits. Significance of the intercept term was than quantified with an F-test.

Table S6. Unique VDJ recombination per isotype in proportion to all isotypes in elderly donors.

isotypes	correlation	p-value
IgA1	-0.73945	0.09298
IgA2	-0.63579	0.17482
IgD	0.95275	0.00330
IgE	-0.08037	0.89777
IgG1	-0.32645	0.52771
IgG2	-0.91184	0.01132
IgG3	-0.41907	0.40820
IgG4	0.92849	0.24221
IgM	0.93712	0.00581
IgM + IgD	0.94895	0.00384
IgA + IgE + IgG	-0.94895	0.00384

Correlations were calculated using the Pearson rank method and linear dependencies were evaluated by standard linear model fits. Significance of the intercept term was than quantified with an F-test.

Table S7. Analysis of changes in the VDJ rearrangement pattern distribution by entropy over all donors.

isotypes	correlation	p-value
IgA1	0.01049	0.97160
IgA2	-0.26130	0.36686
IgD	0.31680	0.26977
IgE	0.18646	0.94504
IgG1	-0.44889	0.10738
IgG2	-0.66682	0.00920
IgG3	-0.54034	0.04607
IgG4	-0.36296	0.41348
IgM	0.33073	0.24810

Table S8. Analysis of changes in the VDJ rearrangement pattern distribution by entropy in the young adults.

isotypes	correlation	p-value
IgA1	0.05655	0.89419
IgA2	0.10641	0.80198
IgD	-0.24102	0.56528
IgE	0.59089	0.40911
IgG1	-0.28189	0.49879
IgG2	-0.05481	0.89744
IgG3	-0.22140	0.59825
IgG4	-0.07953	0.88096
IgM	-0.58855	0.12482

Table S9. Analysis of changes in the VDJ rearrangement pattern distribution by entropy in the elderly.

isotypes	correlation	p-value
IgA1	-0.74272	0.09077
IgA2	-0.64811	0.16395
IgD	0.94074	0.00516
IgE	-0.67353	0.32647
IgG1	-0.31234	0.54673
IgG2	-0.92320	0.00862
IgG3	-0.40495	0.42578
IgG4	NA	NA
IgM	0.84451	0.03439

7. Further analysis of V(D)J gene usage and recombination

Additionally to the results described in manuscripts III and IV, further sequence analyses were conducted. Basis for these analyses were the sequences obtained from antibody repertoires of 14 healthy Caucasians, which were derived from RNA samples of peripheral blood mononuclear cells. With the newly developed method descripted in manuscript III, cDNA was generated and amplified, followed by amplicon library preparation and sequenced on a 454 FLX pyrosequencer. Sequences were first stored in the nextIGbase, a database developed together with Dr. Volker Sievert for antibody sequencing data storage and processing in the group of Dr. Konthur. There, sequences were linked to the donors and for each sequence the isotype was identified by applying a short pattern search in the constant region of the antibody, according to manuscript IV. Sequences were grouped in collections of 150.000 sequences - each over 380 bp length - and send to IMGT/HighV-QUEST [95] for V(D)J gene assignment. Obtained results were integrated into the nextIGbase and further analyzed for quality criteria, such as unique V(D)J gene assignment. Also, alleles were pooled to the corresponding genes, as described in manuscript IV. In total 1,046,521 HC and 295,555 LC sequences, out of 3,566,089 reads, fulfilled all quality criteria and were used for analysis of the immunoglobulin repertoire of the donors.

In combination with an especially selected donor cohort of healthy individuals of different age and gender, the new method offers a novel perspective of analysis and a comprehensive view on the expressed immunoglobulin repertoire in healthy people.

At first, the median frequency for each gene was calculated by normalization of the sequences of all 14 healthy donors, separately. The numbers of sequences for each V gene were divided by the total number of sequences within the donor for HC or LC, respectively. VDJ assignment and nomenclature were provided from IMGT. The nomenclature begins with IGH, which stands for immunoglobulin heavy, followed by V, D or J gene symbols. The number after the four letter code (e.g. IGHV1-1) describes the gene family and the number after the hyphen the specific gene.

7.1 Heavy chain gene usage

First, the obtained sequences were used for V gene analysis. Of all 86 V genes in IMGT, 58 different V genes were found (figure 7.1). The relative frequency of reads differs between the V genes and donors. Most of the V genes are expressed less than 5%. The V gene with the highest median is IGHV3-23 followed by IGHV3-30. Also, IGHV4-34 is frequently expressed. IGHV1-2 is the gene with widest spread of usage. The observed outlier value is due to overexpression of said gene in a single 20 years old male donor (I200091-021), who was already recognized to have an ongoing immune reaction, as described in manuscript IV. The majority of V genes can be found in all 14 donors. The genes with lowest frequencies were not found in all donors. For some of these genes (e.g. IGHV1/OR15-1 and IGHV3/OR16-8) IMGT has currently not finalized the annotation and also their functionality has not been fully assessed.

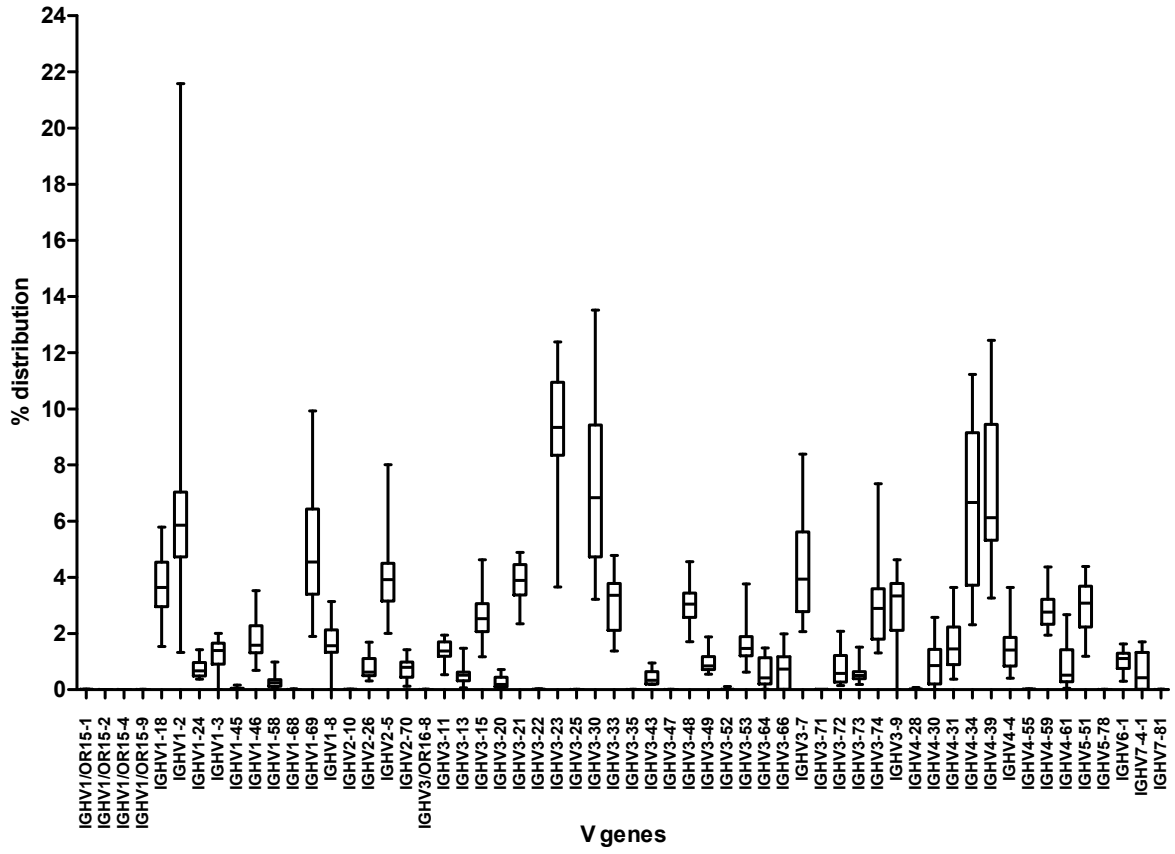


Figure 7.1: IGHV gene distribution within the donor cohort (n=14). The median distribution is shown as lines, 25 to 75 percentile as Boxes and the full range as whiskers. V gene nomenclature was assigned according to IMGT.

In the next step, the IGHD gene usage was assessed. In total 30 different D genes – out of 32 in IMGT – could be assigned (figure 7.2). IGHD3-10, IGHD3-22 and IGHD6-19 are the three most commonly used genes. Most of the D genes show a high variability in usage frequency between the different donors. Especially the genes, which were found in more than 5% of the reads, show great variation. All genes were found in all donors.

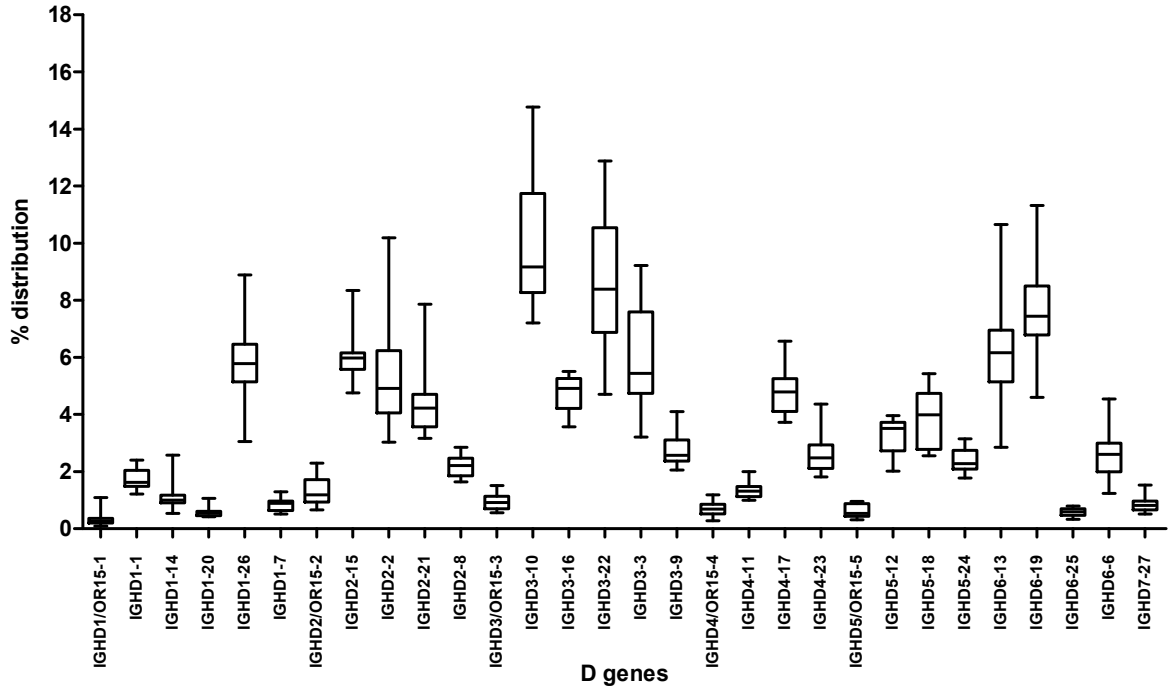


Figure 7.2: HC IGHD gene distribution within the donor cohort (n=14). The median distribution is shown as lines, 25 to 75 percentile as Boxes and the full range as whiskers. D gene nomenclature was assigned according to IMGT.

Finally, the distribution of IGHJ gene usage was analyzed (figure 7.3). All six known IGHJ genes were found in all donors and show a different distribution pattern within the HC of the donors. IGHJ4 is the most frequently found J gene with a median usage of 51%, whereas IGHJ1 and IGHJ2 are only sporadically assigned to reads. IGHJ3, IGHJ5 and IGHJ6 genes are commonly found in a similar frequency.

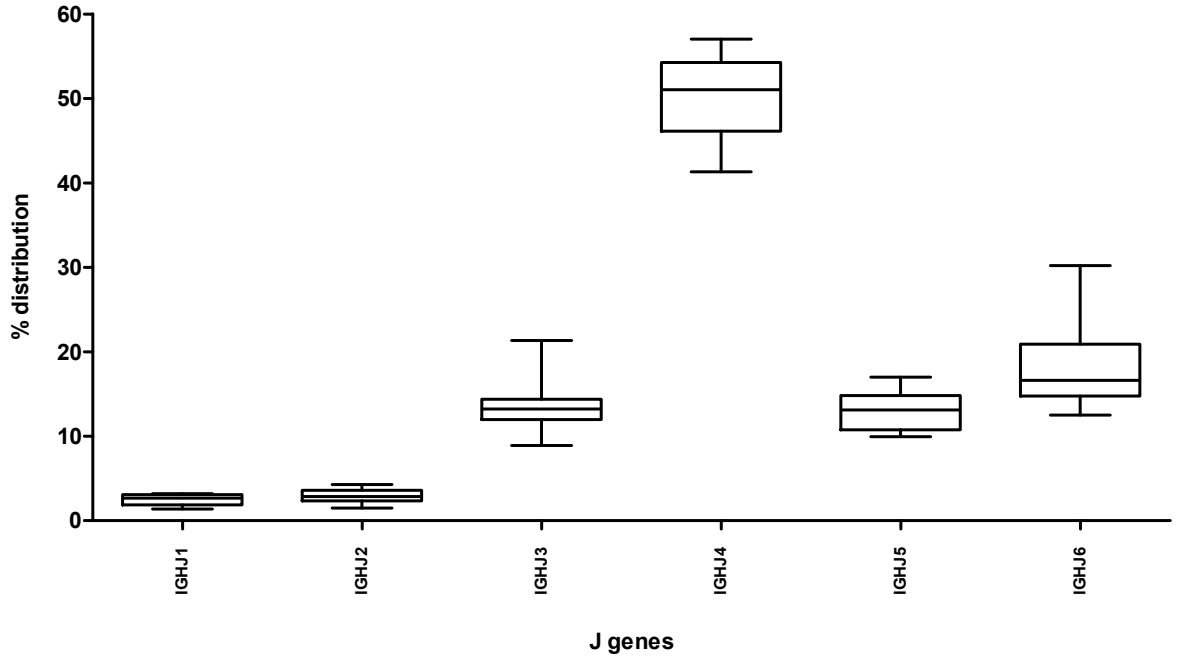


Figure 7.3: IGHJ gene distribution within the donor cohort (n=14). The median distribution is shown as lines, 25 to 75 percentile as Boxes and the full range as whiskers. J gene nomenclature was assigned according to IMGT.

7.2 Distribution of Heavy chain VDJ gene segment genes usage on per isotypes

Applying the new immunoglobulin sequencing method described in this thesis, for the first time it is also possible to assign the recombined immunoglobulin HC genes to the different immunoglobulin isotypes. Assignment was performed on the basis of defined sequence motifs in the CH1 region. In total, nine isotypes can be differentiated: IgA1, IgA2, IgD, IgE, IgG1, IgG2, IgG3, IgG4 and IgM. The isotype classes IgE and IgG4 were found in very little amounts and, therefore, were excluded from further analysis. At first, for each donor the relative distribution of every found gene per isotype was calculated individually. This data was then used to obtain the median distribution of genes per isotope within all donors. Figure 7.4 shows the distribution of V genes and their assigned isotypes for all donors. For most V genes, the distribution between the individual isotypes is widely spread around the median, whereas for some V genes certain tendencies are visible. For example, usage of IGHV3-30 shows similar frequency in gene usage in IgA1, IgA2, IgG1 and IgM. Obviously, as already seen in the overall IGHV distribution (above), IGHV3-23 has the highest median around 10% of all V genes used in all Ig isotypes.

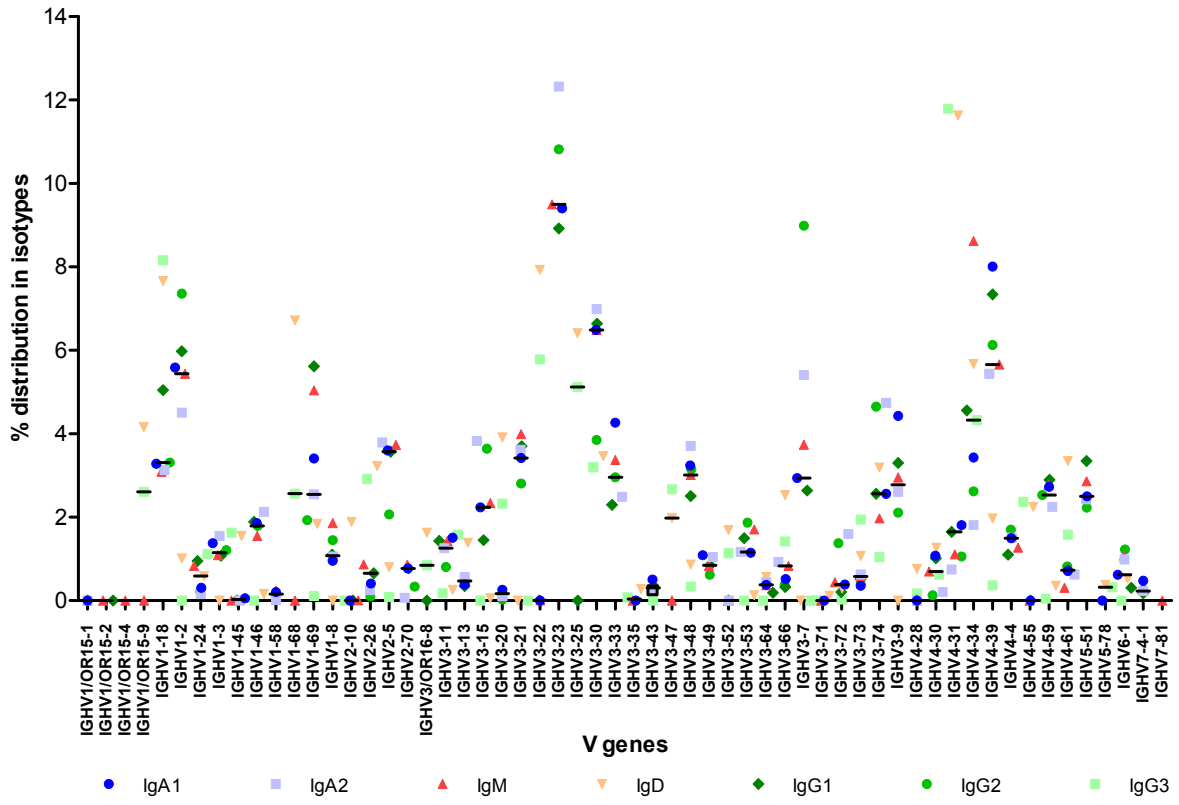


Figure 7.4: Relative distribution of V gene usage per isotypes in the donor cohort. Black line indicates the median of distribution for each gene.

The analysis of IGHD distribution according to isotypes reveals a similar picture (figure 7.5). All 30 observed IGHD genes within the donor cohort were specified with different median frequency in the different Ig isotypes. Their usage is also spread between the antibody classes, but less widely than in the IGHV genes. No tendency for any Ig isotype is observed for any of the genes. IGHD3-10 and IGHD3-22 are the most frequently observed D genes, as expected from the overall distribution analysis.

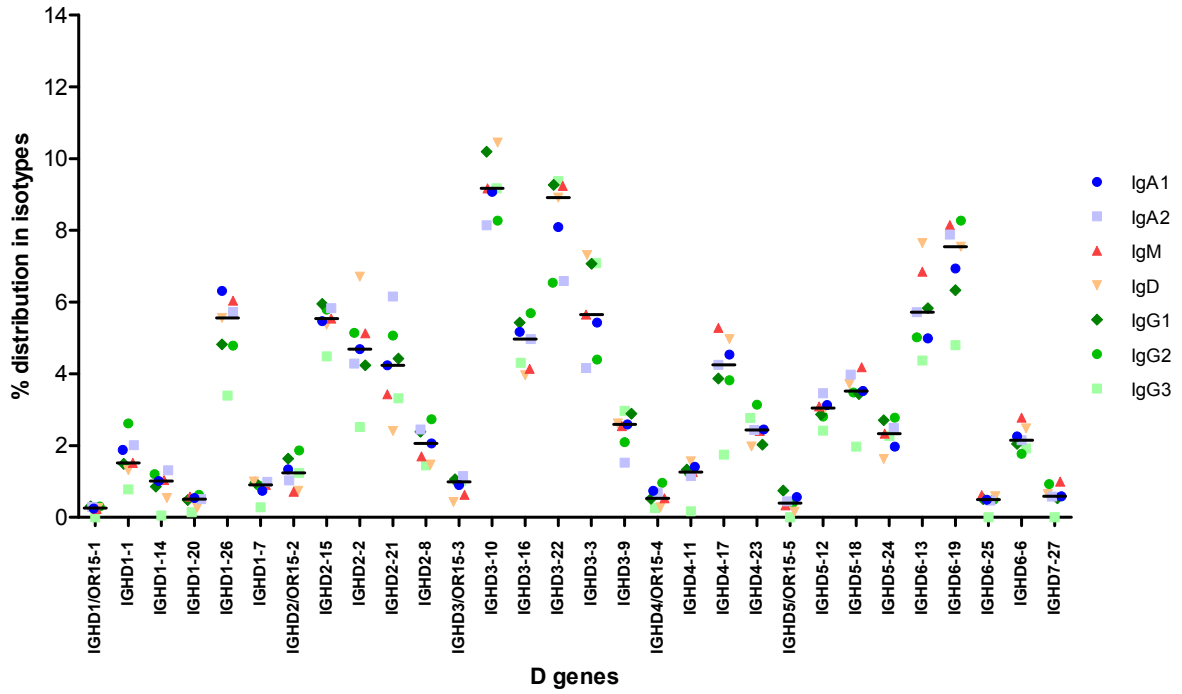


Figure 7.5: Relative distribution of D gene usage per isotypes in the donor cohort. Black line indicates the median of distribution for each gene.

Finally, the distribution of the six IGHJ genes within the isotypes was assessed and the results are shown in Figure 7.6. Interestingly, except IGHJ4 and IGHJ6 in the IgD isotype, all other IGHJ genes frequencies in all isotypes are close to the median of the corresponding IGHJ gene. As expected, IGHJ4 has the highest gene usage frequency in all assigned antibody classes.

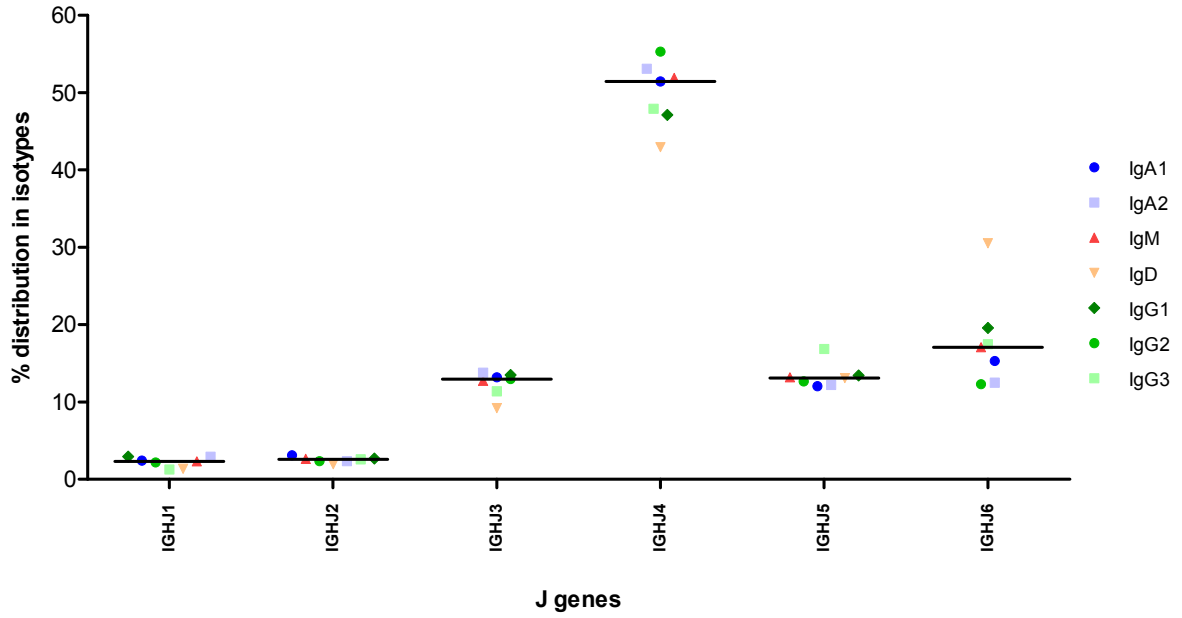


Figure 7.6: Relative distribution of J gene usage per isotypes in the donor cohort. Black line indicates the median of distribution for each gene.

7.3 The top heavy chain VDJ recombination patterns

In total 6685 unique heavy chain VDJ recombination patterns were found in the sequencing analysis. These are 64% of all recombinations possible with the genes actually found and 41% of all theoretically possible. 6303 (94.3%) of them were found more than once and 5131 (76.8%) at least ten times. Only 670 (10%) of the VDJ recombination were seen in all 14 donors and ⅓ of the recombination (2189) in at least 10 donors. We included rare or unique recombinations for the major analysis, but these should be regarded with caution as their annotations may be artifacts due to amplification or sequencing errors. Therefore, for a detail comparison just a fraction of found recombinations were analyzed in more detail.

The calculations to obtain the top 100 most frequently expressed VDJ recombinations (irrespective of isotype) per donor individually were done by normalization of each donor over all obtained sequences and arranging them according to their concurrency in the individual donor. The top 100 recombinations represent roughly ¼ to ⅓ of the obtained sequences per donor with a median frequency of 28.7% (average 30.2%; figure 7.7). Only donor I200091-021 (male 20 years old) has a high proportion of about ~50%. This is the same donor which has been observed on individual IGHV gene level to have highly overexpress IGHV1-2.

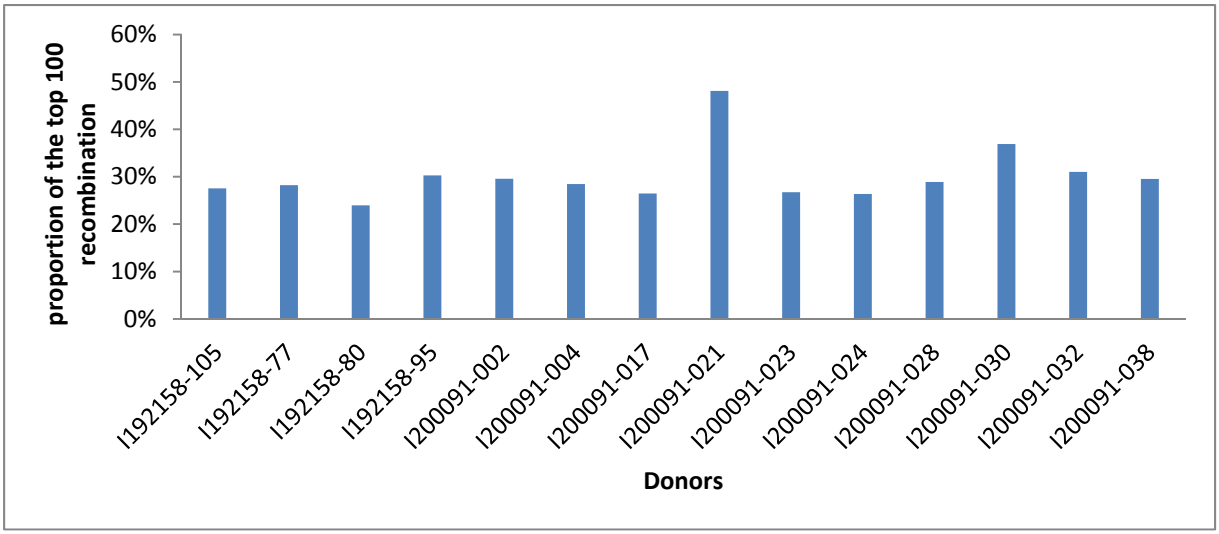


Figure 7.7: Proportion of the top 100 recombinations. Proportion of the top 100 pattern to the total number of recombinations found in each individual donor of the cohort (n=14).

Next we analyzed how many of the top 100 recombinations within a single individual are also found in the other individuals.

The distribution of recombination pattern shows only a small overlap between the different donors (figure 7.8). In total, 567 different recombinations reflect the distribution of the top 100 which is 8.5% of the total number of found recombinations. 57.3% of the recombinations are only found in one donor, 16.0% is shared between two donors and 9.3% in at least seven donors. The figure shows the strong decline of the shared recombinations between the donors. Most of the patterns are only found in a single donor's top expression.

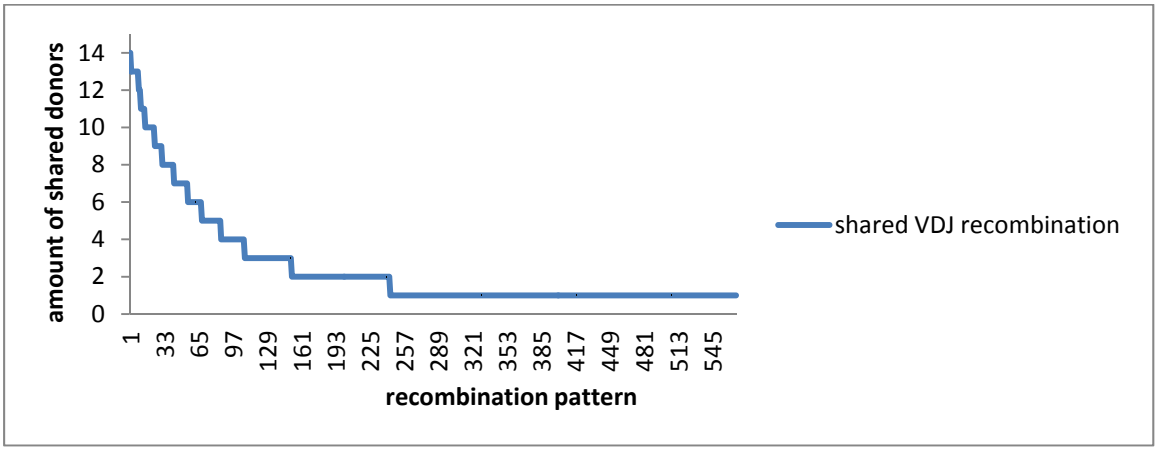


Figure 7.8: Shared recombination pattern in the cohort. Number of shared VDJ recombination between the top 100 recombinations of all individual donors of the cohort (n=14).

Next we summarized the analysis of the VDJ recombination patterns shared in minimum 50% of the donor cohort (n=14) (figure 7.9). Only 53 recombinations appear to be shared in more than seven donors. Of all shared VDJ recombination patterns, the combination IGHV3-23 with IGHD3-22 and IGHJ4 is the only recombination existing in all top 100 recombinations of all donors with a median occurrence of 0.5% per donor.

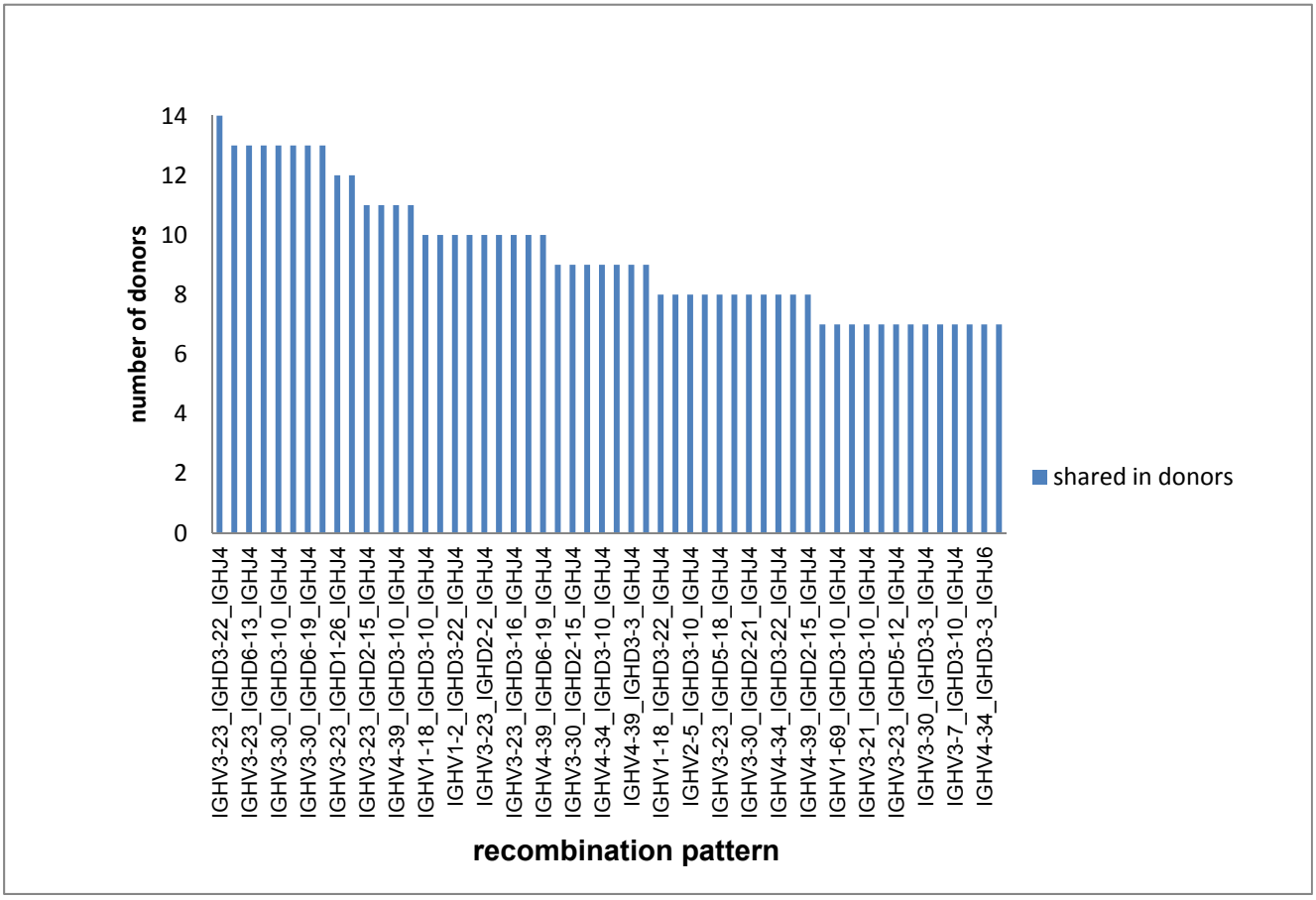


Figure 7.9: Most shared recombinations found in minimum 50% of the donors (n=14).

7.4 Light chain gene segment usage

Human Immunoglobulin light chains can be either of kappa or lambda type. They have each their own set of V and J genes. The kappa gene locus is located on Chromosome 2p11.2, the Lambda genes are encoded on chromosome 22q11.2. There are 54 and 52 V genes and 5 and 7 J-genes known for kappa and lambda, respectively.

Of the known 54 KV genes, 30 were found in our donor cohort of healthy Caucasians (figure 7.10). The two kappa V genes with the highest median are IGKV1-39 and 4-1 and belong to a subset of six genes with median frequencies over 10%. KV1-39 and KV4-1 have the highest median frequencies. The remaining IGKV ranges between 0% and 6 %.

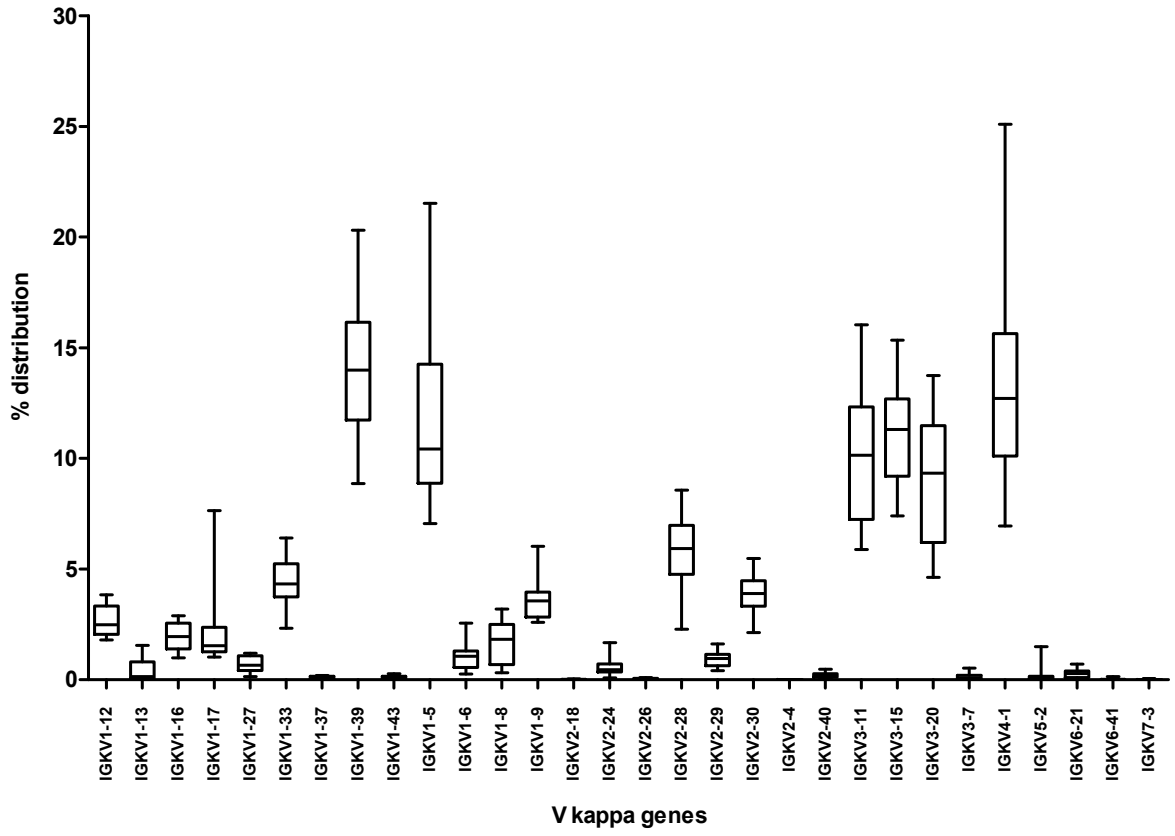


Figure 7.10: IGKV gene distribution within the donor cohort (n=13). The median distribution is shown as lines, 25 to 75 percentile as Boxes and the full range as whiskers. Kappa V gene nomenclature was assigned according to IMGT.

The lambda V gene analysis revealed 34 different genes of the 52 IGLV genes in IMGT (figure 7.11). IGLV2-14 is the most expressed gene in this study. Overall, the expression distribution of the genes varies widely between the donors, while some genes were only rarely observed.

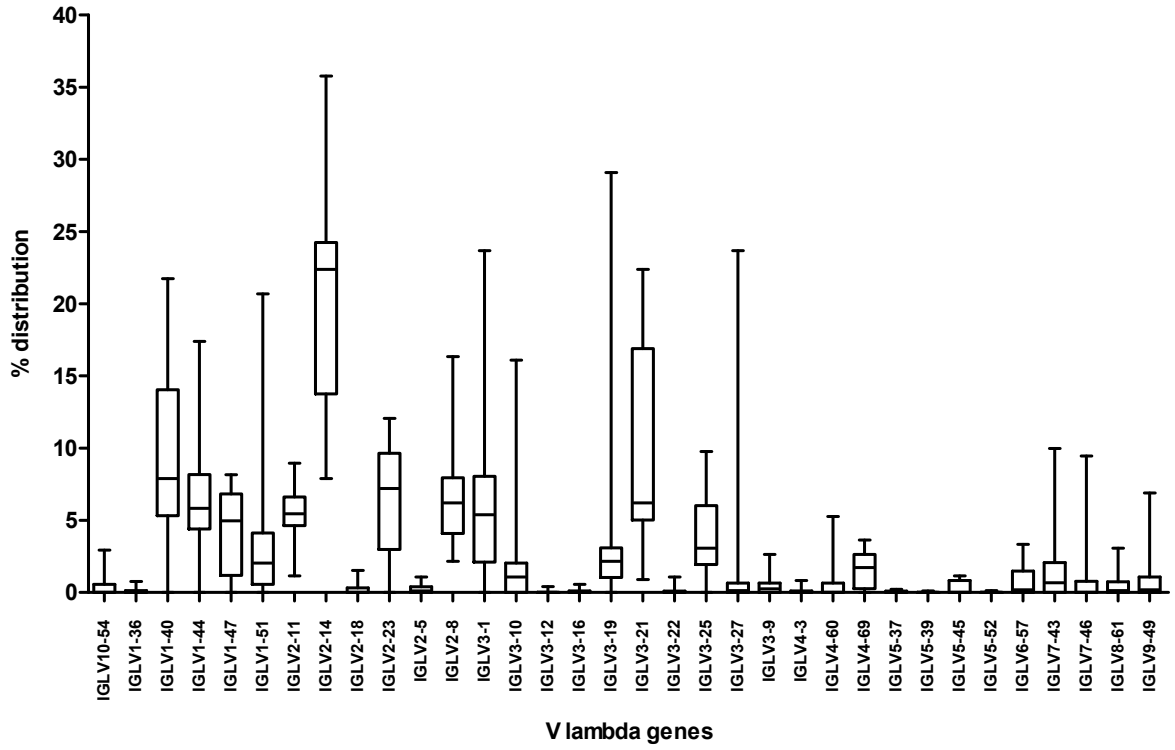


Figure 7.11: IGLV gene distribution within the donor cohort (n=13). The median distribution is shown as lines, 25 to 75 percentile as Boxes and the full range as whiskers. Lambda V gene nomenclature was assigned according to IMGT.

Finally, the gene segment usage of IGKJ and IGLJ were was analyzed (figure 7.12). For kappa light chains, IGKJ1, IGKJ2 and IGKJ4 usage is evenly distributed. IGKJ3 and IGKJ5 genes are less frequent. In lambda light chains, IGLJ1 (56%) and IGLJ3 (42%) are the most prominent J genes contributing 98% of all variants. The frequency of IGLJ5, IGLJ7 and IGLJ6 genes was rare, while IGLJ2 and IGLJ4 have not been found at all in our set of data.

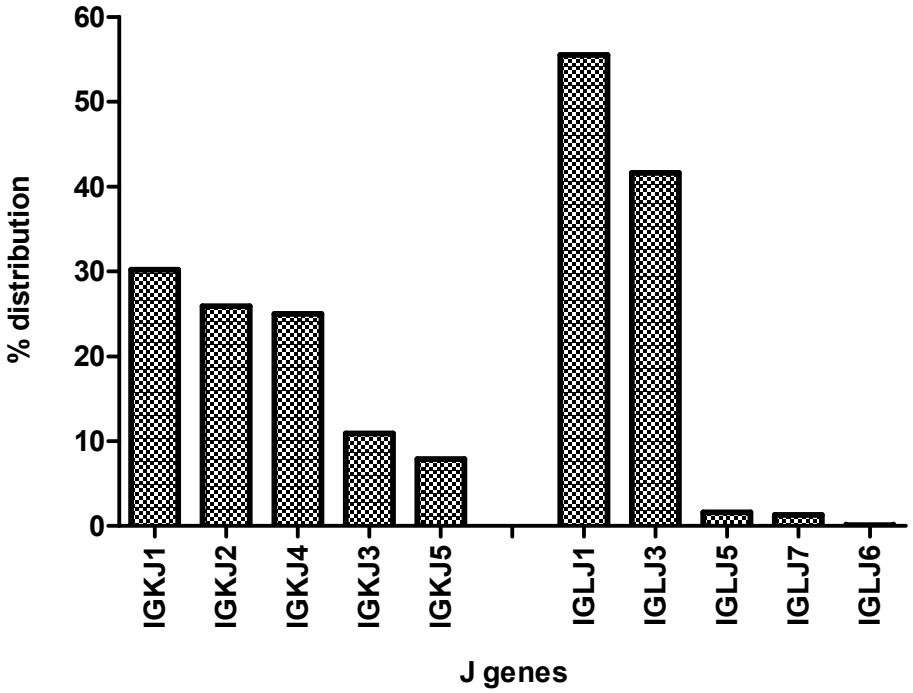


Figure 7.12: IGKJ and IGLJ gene distribution within the donor cohort (n=13). Kappa and lambda J gene nomenclature was assigned according to IMGT.

8. Discussion

In this thesis, a novel immunoglobulin sequencing method was developed. Amplicons were produced with V gene independent primers in a state-of-the-art emulsion PCR, thereby avoiding any bias or building of chimeric sequences during amplification. The amplicons contain the recombination pattern from all heavy and light chains as well as their isotype affiliation. Furthermore, the benefits of this new methodology and its relevance for reliable high throughput sequencing were shown by analyzing the expressed antibody repertoires from peripheral blood mononuclear cells (PBMC) of 14 healthy donors. Hence, the foundation for deeper understanding of the immunoglobulin repertoire as well as immune senescence is given.

Development of a complete and unbiased immunoglobulin amplification method and its adaptation to next generation pyrosequencing technology.

The adaptive immune system with its immunoglobulins is of great importance for the protection of our body. For the first time ever, the diversity of V(D)J rearrangements in combination with complete isotype assignment has been shown in this thesis and shads light into the nature of the immune system (manuscript VI).

The first obstacle for the establishment of a sequencing procedure of immunoglobulin repertoires is the amplification of the rearranged antibodies genes. Hence, it was necessary to evaluate a common primer set for the amplification of V(D)J recombinations (manuscript I). The standard way of antibody amplification and cloning is by the use of primers binding at the 5'of the V gene. The annealing site for the second primer is either at the end of the J gene, or at the end of the constant heavy 1 (CH1), respectively constant light (CL) domain. For this approach we revisited the established V gene primer set from Sblattero and Bradbury [97] and analyzed their binding efficiency with all known V genes in VBASE2 (manuscript I) [118]. The existing primers cover theoretically 89.2% of all recorded 576 V genes in VBASE2. In the course of the analysis it was found, that some of the primer sequences proposed did not fully cover certain V gene families, as earlier reported. Therefore, primers were partially renamed corresponding to their V gene families they can really bind to and novel primers were designed to increase the coverage. With the newly added primers it is possible to amplify 100% (previously, one gene was missing) of all class 1

V genes, which describes a group of V genes displaying genomic as well as rearranged references and are proven functional. At the same time, the coverage of all functional and putatively functional V genes in class 2 V genes (only genomic reference) could be increased to 97.8%. Overall, theoretically 91.5% of all 576 V genes including orphans and pseudogenes are covered by the extended primer set. Thus, the increased primer set is useful to include more V genes into the amplification and cloning procedure.

To evaluate the efficiency of amplification of the extended primer set, pyrosequencing was conducted. For the analysis of the obtained sequences a novel in house database called nextIGbase was developed (together with Dr. Volker Sievert), which allows to evaluate the sequencing results based on VBASE2, in a high throughput manner. The development of tools for the analysis of immunoglobulin data is of crucial importance. For manuscript I, 1000 antibody sequences were statistically analyzed and verified 52 (out of 61) class 1 and 84 (out of 206) class 2 V genes. These results are very interesting, because they illustrate the limited knowledge about the functionality of many V genes at the current time. Since the amplified recombinations were all obtained from mRNA of PBMC of human donors applying a primer binding to the CH1 region, all obtained V genes can be seen as functional. Focusing on class 1 genes alone does not seem sufficient anymore and it can be estimated that many class 2 genes may be functional and frequently expressed as well. Due to technological limitations in the past, the functionality of many V genes has been misinterpreted, since insufficient data was available due to lack of high throughput sequencing. For the planned NGS approach for antibody sequencing in this thesis, it was concluded that one of the primary requirements is to cover all V genes and not only the annotated functional genes. A selection of particular genes would not reflect the real expression profile and, therefore, limits the informative value of the study.

Additionally to the extended primer set, we have optimized the amplification protocol. This included the evaluation of several reverse transcriptases for cDNA synthesis, several polymerases for amplification and additives, such as the commercially available extreme thermo stable single strand binding protein (ET SSB). SSB plays an essential role in DNA replication [120] by stabilizing the ssDNA intermediates generated during DNA processing in *E. coli* [121]. Additionally, the use of SSB in PCR amplification can improve the quality of the products by increasing the fidelity of the polymerase [122-124]. In *E. coli*, SSB is expressed from a single gene and works as a homotetramer during DNA replication. Notably, SSBs were identified in organisms ranging from prokaryotes to eukaryotes (Replication protein A) as well as in archaea. It's function is to protect the DNA, to keep it unfolded and to coordinate DNA-replication proteins [125]. The ET SSB, isolated from a hyperthermophilic

microorganism, is stable to be used in a PCR and our results show a beneficial effect on amplification efficacy. Next to ET SSB, the use of the Phusion Polymerase, with its characteristic high fidelity [126], could be identified as best choice for amplification of rearranged human antibody genes.

To reflect the natural expression of immunoglobulins, any bias or creation of chimera during amplification of the templates has to be minimized. Chimera are species of artificial molecules originating from a mixture of templates during a PCR reaction. To achieve this, we developed a streamlined protocol for emulsion PCR (ePCR) and subsequent purification of the amplified DNA (manuscript II) [127]. The aqueous droplets in the water in oil emulsion act as micro reaction compartments in a bulk oil phase. Chemical or biochemical reactions in emulsion systems [128] have been investigated and adopted for single template PCR amplification [129]. It is known that the compartmentalization of the template DNA including all compounds for the PCR has many advantages. The bias of amplification is reduced by “blind” amplification, whereby the specific template has no competition from other templates. Furthermore, the production of chimera is avoided if only a single template is available for amplification in a droplet and, last but not least, the generation of byproducts like primer-dimers is recorded to be reduced [129-132].

Within this thesis, the adaption and verification of cDNA amplification in emulsion PCR was successfully established. In manuscript II it is shown that the amplification of cDNA is different between “open” (conventional) and ePCR. Notably, a distortion is observed between the amplification size distributions in the open PCR and ePCR using 18 and 21 cycles of elongation (manuscript II figure 2a). In the ePCR, the differences are much smaller and artifact creation is reduced, which approves the benefits of the emulsion procedure. The newly developed and streamlined protocol is easy to use in lab routine and requires no special equipment. Of note, the vortex-formed homogenous emulsion can be used in a standard thermo cycler for amplification. The establishment of a subsequent purification protocol applying standard DNA purification kit components further simplified the ePCR procedure and enables the standardization of amplification. The results of this work were directly picked up by Roboklon GmbH (Germany) and led to the development of the “Micellula DNA Emulsion & Purification Kit”, emphasizing the impact and applicability of these technical changes in routine lab work.

One limitation of ePCR is that the maximum number of cycles is restricted to 15 to 20, dependent on the length of the amplified target. The reason for the limited number of cycles,

is anticipated to be the plateau phase in PCR, where the increasing concentration of amplified DNA within a vesicle becomes inhibitory for the functioning of the polymerase [133]. If a higher amplification rate is required, a second ePCR amplification step can be simply performed based on the purified DNA. Overall, the application of ePCR is a major milestone in obtaining high quality amplified DNA without distortion of the DNA population and well suitable for antibody amplification.

To cover the complete expression repertoire of immunoglobulins and to avoid primer dependent bias during amplification, a new approach was established which is independent of the use of V gene specific primers. This was achieved by elongating the 5-prime end of the nascent cDNA by a defined stretch of DNA, which could serve as an annealing site for a newly designed primer amplification. Therefore, a single new 5' primer replaces the degenerate V gene primer set for heavy and light chains and is applicable for all isotype amplification, as long as a specific 3' primer is added to the system. In 1989, a similar approach was developed by the group of M. Davis for the analysis of T-cell receptor δ chain, naming the added sequence "anchor" [134]. The combination of the anchor sequence with specific 3' primers located in the constant region of immunoglobulins allows the amplification of all immunoglobulin cDNA from heavy and light chains. The method covers all known or unknown V genes, without constraints (manuscript III and IV). Thus, it is feasible to reduce amplification differences originating from different efficiency of diverse sets and/or degenerate primers. The CH1/CL region specific 3' primers were chosen to have a high similarity and equal annealing temperature. To circumvent that possible cross-reactivity of the primers skew the amplification process, the isotype assignment was performed by using a pattern search approach in the amplified CH1/CL part and was not based on the used primers per se. Finally, this amplification method for immunoglobulin genes was adapted to the use of ePCR and the conditions were further optimized to ensure minimal distortion of the repertoire.

In the next step, this approach needed to be transferred to Roche's 454 pyrosequencing platform. As the sequencing reaction is dependent on specifically introduced primers, they need to be incorporated during template generation. Commonly, these primer sequences are simply added by PCR. However, this approach failed in this case due to primer incompatibility of Roche's Adapter A and the anchor sequence used for template amplification. Another standard procedure for 454 pyrosequencing is the addition of Adapters A and B by random blunt-end ligation to the PCR product. This method could also not be applied, because the sequencing orientation needs to be 3' – 5' in respect to the amplified immunoglobulin gene. If sequencing is conduct from the 5' end, starting with the anchor

sequence, the current read length of ~400 bp at that time would have been too short to cover the whole antibody genes rendering the approach insufficient. The bioinformatic analysis of immunoglobulin genes requires a minimum of 380 bp to efficiently allow the determination of V, D, and J segments as well as to recover CH1/CL information necessary for isotype assignment. Hence, an alternative approach was chosen, based on the specific ligation of modified Adapters A and B to the PCR product (manuscript III). Therefore, SfiI restriction enzyme recognition sequences were introduced at both end of the amplicon during cDNA amplification. SfiI is a type II restriction enzyme which can make different sticky ends as it tolerates nucleotide variants in its recognition motif. As a direct consequence, unidirectional amplicon sequencing is possible by ligating modified Adaptor A and B sequences, accordingly, to the end of the PCR product ultimately solving the limitations of the Roche standard protocol described (manuscript III).

The analysis of the first 454-sequencing results for our publication 2010 (manuscript I) revealed many V genes that were annotated with genomic evidence only. Therefore, the functionality of these V genes remained uncertain. The found class 2 V genes (41% of all class 2 genes) were more frequent than expected, considering the fact that just 1000 sequences were analyzed in total. Hence, the reference sequences from VBASE2 were compared with the annotation in IMGT [135] and differences in their V gene annotation became evident. All found class 2 V genes were in IMGT annotated as functional and, consequently, confirmed our results. Next to these results, the more extensive set of reference sequences available in IMGT as well as the frequent updates of the database with newly found genes and annotations lead us to switch our bioinformatical tools for the V(D)J assignment from VBASE2 to IMGT. Despite the change of the reference database, some limitations are still possible. For instance, in the event of the amplification of a completely unknown gene, which the new V gene independent method is able to cover, this gene might not or might be falsely assigned by IMGT and presumably rejected according to our quality control criteria. Such case is however regarded as very rare, since IMGT is frequently updated and even very rare genes should be included. Despite this fact, in 2011 a new kappa V gene was discovered [105] and possible additional findings cannot be ruled out. Also D-D fusions cannot be detected by our approach, because IMGT is limited in this respect. This type of recombination is expected to occur in 1 out of 800 native cells [136]. While the frequency is quite low, it may by detect in future – once such an analytical tool is available and the current datasets are being reanalyzed. To overcome limitations associated with the sequencing accuracy of the 454 pyrosequencing technology or short read length, the obtained sequences are not annotated to alleles. The allele output of IMGT is pooled to the

corresponding genes, since the small differences in nucleotide sequences between the alleles cannot be reliably resolved by single reads. Especially the error rate of the sequencing procedure at the end of the read corresponding to the 5' end of the V gene may lead to false assignment. Despite these limitations, the new approach presented here is one of the most complex and comprehensive immunoglobulin sequencing approaches to date. The method allows qualitative and quantitative analyses and enables to gain new insights into the immunoglobulin repertoire.

Analysis of the immunoglobulin repertoire in peripheral blood samples of healthy human donors.

For the main sequencing project, peripheral blood samples from 14 healthy Caucasians were collected and analyzed. All donors gave written informed consent for their blood to be used in this project and an ethic vote was obtained from the "Freiburger Ethik Kommission International" under the code 012/1679. Additionally, all donors were asked to fill out a questionnaire to exclude individuals with recent diseases or interventions, which could influence the immune status, such as infections or vaccination. Details of the selected donors can be found in manuscript IV, table 1. The RNA of the donors was reverse transcribed in cDNA and the immunoglobulin repertoire was amplified and sequenced on a 454 FLX pyrosequencer with the newly developed method described in manuscript III. All obtained sequences were stored in the nextIGbase, where every read is annotated with the donors information. Primary VDJ assignment was conducted using the IMGT/HighV-QUEST query tool [137]. In total 3,566,089 reads were gained from the sequencer, of which 1,357,978 satisfied all subsequent quality criteria.

First, the VDJ-gene distribution was analyzed (see chapter 7). Of note is that the distribution of expression for the most genes is very high, reflecting the differences between the donors immune systems. This underpins the need of numerous donors to be included into the study design to see the natural distribution in a cohort.

The obtained V gene sequences could be mapped to 58 different IGVH-genes, of which 37 are common in all donors. It is not surprising that not all V genes were found, because some genes may occur only very rarely and many pseudogenes exist, which will not be expressed. Nevertheless, the method is not limited in finding genes. For instance, two V genes which were missed in a previous sequencing study [108] were found here: IGHV3-13 (median frequency 0.52%) and IGHV4-61 (median frequency 0.52%), underlining the power of the V gene independent amplification approach. Interestingly, antibodies encoded by IGHV4-34 have been associated with the occurrence of autoimmune diseases [138-141]. However, on

transcript level, we found this gene in all our healthy individuals in a relatively abundant fashion (median 6.7%, range 2.3% to 11.2%). Of note, it has a preference for certain isotypes, mainly IgD and IgG3 and to a certain extent also IgM (figure 7.4). If this V gene really has any predictive value for autoimmune diseases only future studies can show, where the immune repertoire of autoimmune patients is compared to a healthy cohort.

With respect to D genes, IGHD3 is predominant (33%), and yet, many genes are found in comparable amounts (figure 7.2). Studies observing the distribution of D-gens are just coming into focus. Thereby, attention should be given to the relatively short nucleotide length and the high similarity between the genes, which increases the possibility of false assignment. The shown distribution is one of the most complex available today and shows great variation between the individual donors. Nevertheless, the most frequently expressed genes here were also found in other studies to be highly expressed [111]. With regards to isotype distribution, the D gene usage shows similar behavior in all Ig classes.

Focusing on the J-segment, a clear preference in recombination is visible. IGHJ4 with a rate of more than 50% reveals to be the most preferred gene. Interestingly, J gene usage in the different Ig classes shows very similar distributions, except for IGHJ6, which is overrepresented in IgD compared to the other classes.

Nowadays, the analysis of the light chain also gets more and more attention (chapter 7.3). It has been recently postulated that they may have a similarly significant role in antigen binding like heavy chains [142]. Light chains only have V and J gene recombinations and lack class switch. Therefore, the expected complexity in respect to recombination patterns is significantly lower than that of the heavy chain. Again, the superiority of the novel amplification method is seen. Previous publications did not analyze kappa IGKV4 to IGKV6 [143], since the used primers do not amplify these genes. Consequently, the kappa repertoire was not completely covered and in particular, IGKV4-1 (representing 12.7% of this repertoire) was missed. This fact certainly highlights the advantage of the V gene independent amplification methodology applied here. The J-gene usage in kappa chains is distributed evenly, with a slight preference for IGKJ1, IGKJ2 and IGKJ4 (out of 5 genes), which contribute in total more than 80% to all kappa sequences obtained.

The lambda repertoire is typically missed in antibody studies due to their lower frequency of usage. In case of the lambda repertoire, IGLV2-14 was found in all donors with an overall frequency of 22.4%. The second most abundant V-gene IGLV1-40 with (7.9%) was found in 12 donors, supporting the findings of Stamatopoulus *et al*. [144], who reported IGLV2-14 to be the most frequently expressed IGLV in normal cells. However, they questioned this finding due to missing representative reference data which is now provided in this study. In lambda

chains, IGLJ1 (56%) and IGLJ3 (42%) are the most prominent J genes, of all 5 variants found. Our data suggests that we have missed the IGLJ 2 gene from the lambda repertoire, even though IGLJ1, IGLJ2, IGLJ3 and IGLJ7 are all recorded as functional [23]. The obtained IGLJ5 and IGLJ6 sequences could be therefore falsely assigned. This situation will need further evaluation, especially the fact that IGLJ2 has been completely missed in this dataset. This could be due to the design of the lambda constant primer, but could also represent a bioinformatic problem. At least one IGLJ2 sequence should have been observed in our set due to false priming or false gene assignment. Therefore the here presented data on the lambda chain should be regarded with caution, as it might not be final.

In summary, the basic analysis of V(D)J frequencies observed within our cohort reveals no major differences to those of other studies, even though this is one of the most extensive immunoglobulin sequencing studies carried out so far. The major difference to all other studies is that the immunoglobulin repertoires were amplified without any V gene specific primer. The observed discrepancies to the other publications support the quality of our method and stress its advantage of wider coverage of V genes. Nevertheless, the cohort of analyzed donors as well as the number of reads per donor should be increased in future studies.

This study of immunoglobulin gene usage distribution marks a good starting point for further scientific questions. For instance, the comparison of this data set with data from diseased individuals could enable us to determine, whether disease-specific variances in gene usage really exist.

The next level of analyses integrated the entire patterns of rearrangements of the described genes. In total 6685 unique VDJ recombination patterns were found. These are 64% of all possible recombination of the genes currently found, and 41% of all theoretically possible recombinations. The number of recombinations observed could be obviously enhanced by increasing the sequencing depth. It is expected that especially the discovery of rare recombination events is proportional to the sequencing depth. This is also represented in the data at hand: 6303 (94.3%) of all 6685 recombinations were found more than once and 5131 (76.8%) at least ten times. Unique recombinations were also included here, but these should be regarded with some caution, as their annotation may be faulty due to amplification or sequencing errors.

Interestingly, only 670 (10%) of the found VDJ recombinations were observed in all 14 donors and ⅓ of the recombinations (2189) in at least 10 donors. This shows that the current sequencing depth is far from complete coverage of the immunoglobulin repertoire of each

individual. Theoretically, if unlimited sequencing depth were possible, the overlap between found VDJ recombinations in the individual donors should be approaching the total number of possible recombinations with the available set of genes.

To minimize the influence of sequencing depth, one possibility is to look only at the most abundant recombinations for each individual. Therefore, here only the top 100 most expressed VDJ recombinations (not considering isotypes) from each donor were analyzed individually and taken into consideration for comparison (chapter 7.3). Even by increasing the amount of sequences per donor, the individual composition of the top 100 recombinations it is unlikely that the top 100 recombinations will change.

Intriguingly, 56.0 % of the recombination patterns were observed only once, underpinning the high variance between the donors (figure 7.8). This observation can be explained by the differences of preferred gene usage in the donors as well as the strong influence of the environment. In particular, environmental influence should be seen in class switched B-cells, due to different antigen exposure. Especially when extrapolated to lifetime, environmental influence has to be rated higher than the influence of the differences in preferred gene usage – as Jiang *et al*. estimated [103]. Nevertheless, only more comprehensive studies and monitoring of the immunoglobulin-repertoire at different time points during the life-span of the same individual would allow to give a more precise explanation.

The recombination pattern of IGHV3-23, IGHD3-22 and IGHJ4 (median occurrence of 0.5% per donor) is the only recombination existing in all top 100 recombinations. This is not surprising, since the recombination is a product of the most frequent individual gene segments. The IGHV and IGHJ genes represent the most common, and the IGHD gene the second most frequently observed gene segment in this data set (chapter 7.1). This finding is underpinning the assumption of unpreferential recombination of VDJ genes as Arnaout and colleagues reported recently [114].

In summary, the top 100 found recombiantions represent roughly ¼ to ⅓ of the obtained sequences per donor (median 28.7%; average 30.2%), in which only a small overlap of recombinations between individual donors is seen (figure 7.8). Consequently, the analysis at hand displays the huge diversity of the immunglobulin repertoire as well as the high variability between the donors.

For a better understanding of how the antibody-based adaptive immune system works it is inadequate to cover the V(D)J gene distribution and recombination patterns only. Notably,

the effector function is of importance and has also to be considered. In manuscript IV, the results of the high throughput sequencing of 14 healthy donors of different age and gender using 454 pyrosequencing technology is presented. Here, one major question was to determine whether the donors can be clustered by age or gender. First, the clustering algorithm was used only by focusing on the V(D)J recombination frequency and donors did neither cluster by age nor by gender. In a second step, the isotype information was introduced as an additional parameter. Still, no clustering of the donors according to age of gender was observed.

However, the dominance of some recombinations found in all donors has become quite obvious, especially for the light chain. This discrepancy is surely dependent on the sequencing depth – especially in the light chain, where the overall recombination possibilities are much more limited, as discussed above. In manuscript IV, the most dominant recombinations of all donors was analyzed (manuscript IV figure S2). Three donors have striking over-represented pattern profiles. These could be regarded as a clear indicator for clonal expansion of a distinct B-cell in response to an antigen. Even though, the donors have declared to be free of any symptoms of illnesses, the eventuality of a beginning immune response to a recent infection cannot be excluded. Therefore, the possibility of clonal expansion was address (manuscript IV, figure 3). For all three donors, the over-representation of some recombinations were almost exclusively found within a single isotype and allows not only speculation on a beginning immune reaction of the donors, but also on the kind of pathogen and/or location of infection. While the over-representation of certain sequences could also be due to PCR over-amplification, despite the effort to avoid such bias, it is very unlikely in the light of the following in-depth analyses.

Pattern expansion assigned to isotype A1 and A2 (manuscript IV, figure 3 b and c) were further analyzed on CDR level. Both expansion clearly show high similarities in the CDRs and therefore are very likely to be caused by clonal expansion or converging maturation to an antigen. The IgA1 immune response is assumed to be located in the serum, where it can activate the alternative complement pathway for clearance of pathogens and apoptotic cells [145]. Furthermore, the IgA1 Fc part cans recruit phagocytes for antibody-mediated detection of pathogens and their clearance. IgA2 dominates in mucosal secretions for instance in genital tract and may be directed against a bacterial infection, providing a comparable target for phagocyte-clearance as described for IgA1. A bacterial infection can be assumed, due to the high stability of IgA2 against bacterial proteases. In the third example a frequent recombination pattern in IgG1 is observed, showing a preference for a certain V gene in combination with different J and D segments. Hence, a starting immunological response can be estimated. Notably, the amount of IgG1 response in this donor is over-represented in

comparison to the other donors. The high IgG1 titer and the described recombination pattern indicate a humoral response to a pathogen in the serum.

While the above interpretations are feasible, to determine whether they can be really applied to establish a reference for the prediction of an immunological response, further evaluations have to be done in the future. A good possibility to do so is to conduct sequencing experiments on donors undergoing vaccination. There, the status can be monitored prior, during and after the immunization process and could therefore shed some light on the underlying mechanism of immune response. Possibly a similar approach could also be chosen by investigating certain types of diseases and compare the results with those obtained in this study. In the long run, it may then even be possible to apply this technology of immunoglobulin sequencing as a diagnostic tool, as suggested elsewhere [146].

The real benefits of the novel approach become evident by analyzing the data from even a broader view: the aspect of immune senescence, which is issued in manuscript IV as well. Immune senescence describes the changes in the immune system occurring during an individual lifespan, and manuscript IV describes for the first time that this can be also observed by antibody sequencing.

Dividing the donor in a young (19-30 years) and elderly (49-62 years) cohort enabled to observe age effects. At first, the relative amount of obtained sequences per isotype within each donor was monitored. The proportion of the isotype classes within the individual is of interest, because of the different function of the isotypes. For an effective immune response it is not only important that the immunoglobulins have a high affinity for the antigen. Additionally, the effector functions of the isotypes are essential for an adequate immune response. In the analysis of the isotype distribution, the young cohort shows no age-effect, but in the elderly group the IgM proportion is augmenting significantly (p-value: 0.029) with the increasing age of the donors (manuscript IV Figure 4A; table S3).

Next, the relative variability of isotypes was analyzed by calculating the number of unique VDJ recombinations per isotype for each donor. These were then grouped into an initial (IgM/D) and a specific immune-response (IgA/E/G) discriminating native and class switched isotypes (manuscript IV Figure 4B and S3). With this approach it is expected to account for variations in numbers of obtained sequences, as all classes should be equally affected by the sequencing depth. The variability, as calculated here, allows the comparison of the diversity in isotypes and the groups within donors and between them. In the young cohort a more or less equal distribution of the variability between IgM/D and IgA/E/G was observed. In contrast, in the elderly cohort the relative number of VDJ recombinations of IgM/D increased.

This observed change is attributed to the change in the class switch ability of B-cells, which is proposed to be influenced by the age of the donors [67].

Following on from the idea that differences between initial and effector isotypes may be based on reduction in CSR, the overlap of VDJ recombinations between isotypes of each donor were analyzed (manuscript IV figure 4C). Analogous to previous analyses above, only the top 100 expressed recombinations per isotype were considered to minimize the influence of sequencing depth. Here, the hypothesis was that a reduction in CSR should lead to a decrease in the overlap between the most expressed VDJ recombination patterns within one donor in the different isotypes. Indeed, this was the case. Further, the overlap-analysis approach also confirmed that the CSR is significantly different between the two age groups. This is emphasized by the strong reduction of overlaps of about ⅓ in the elderly group compared to the young cohort.

Finally, the overlap of VDJ recombination between the different immunoglobulin classes within each individual was used for hierarchical clustering of all donors in the cohort. In this analysis, the donors segregated for the first time according to age into young and elderly. Intriguingly, the younger donors of the elderly cohort seem to connect both groups. This phenomenon could be explained by the starting senescence of the immune system at the age of 50 and beyond and can be seen to some extend also in sub-analyses of the data set whereby the overlap-analysis was conducted only in selected Ig classes (manuscript IV; figure S4-S7). These findings underline the benefit of analyzing the full immunoglobulin repertoire.

As last confirmation of the age related changes of the immunoglobulin repertoire the entropy was analyzed (manuscript IV; table S9). Entropy measures the dispersion within the distribution of VDJ recombination for individual donors. Assuming that the CSR is affected only in the elderly, the young should not be influenced. Indeed, in this group, no significant age related changes were found. In the elderly, however, in the not class switched isotypes IgM and IgD the entropy was increased. This fact, and a slight decrease in the other isotypes, supports the explanation that the CSR is already affected in people in the “golden age”. In IgM and IgD isotypes the dispersion of VDJ patterns increase through insufficient CSR compared to other immunoglobulin classes and, therefore, leads to an increased entropy.

It has to be mentioned that in the isotype distribution, the variability and the recombination-overlap analysis, the 24 year old male donor in the cohort showed rather characteristics of an elderly than a young individual. Ideally, it would be interesting to analyze this donor in more

detail in future studies, for instance when the donor has underwent an immunization process. In principle, his response to vaccines should be comparable to that of an elderly individual.

Age related changes in relation to CSR came into focus just recently and until now, the changes were expected to arise in people older than 60 or 65 years. Therefore, the majority of vaccination-efficacy studies related to immune senescence did not include people within the age range of 50 to 60 years. Intriguingly, a resent tick bone study in Austria revealed that the immune response to vaccines is already impaired at an age of fifty [66], stressing the importance of changing the focus of immunological research to this age as well.

In summary, the integration of Ig class analysis as a function of age in combination with the newly introduced overlap-analysis approach for the investigation of VDJ recombination identified that 50 years of age and beyond marks already the onset of immune senescence. This is earlier as commonly discussed in the field and emphasizes that the approaches developed within my thesis could contribute to the analysis of age-related research in the field of immunology and vaccination in particular. The reduction of vaccination efficacy in concordance with increasing age is posing a serious health problem in an aging society. The newly developed method of immunoglobulin amplification and high throughput sequencing, as well as the new statistical analyses for immunoglobulin sequences, represent a huge advancement in immunoglobulin research. The possibilities to measure the changes occurring in a relative young age and the expected influence for successful vaccination opens up a new filed in antibody research and allow the connection to other coherent immunological processes.

9. Summary

The main role of the adaptive immune system is to protect the body against pathogens. Immunoglobulins are an essential part of this system. The great diversity of immunoglobulins is obtained by a complex mechanism of genetic reorganization from a predetermined set of gene segments on chromosomal level and subsequent affinity maturation. Thereby a specific affinity for a particular pathogen is finally achieved. For a successful immune response to occur, the effector functions of the constant region of the immunoglobulin are also essential - which are different in all immunoglobulin classes. The latest innovations in DNA sequencing technology allow nowadays to analyze and evaluate immunological questions in respect to cell-bound (i.e. B-cell receptors) and secreted immunoglobulins from human donors on cDNA level.

The technological innovation, which is generally referred to as "next generation sequencing" and especially the so-called pyro-sequencing laid the basis to the present study. The aim was to develop a method for sequencing antibody cDNA by pyro-sequencing, which allows all genes and all immunoglobulin classes to be quantitatively and qualitatively measurable. With this novel approach the immunoglobulin repertoire from peripheral blood samples of healthy donors of different ages and gender ought to be analyzed.

The newly developed immunoglobulin specific cDNA amplification method established during this thesis is independent of the use of V-gene specific primers. It further allows the verification of the corresponding immunoglobulin classes and sub-groups in the obtained sequences. Through the introduction of a newly developed emulsion PCR method, the influences of the sample preparation and the cDNA amplification procedure on the representative quality of the obtained sequences were minimized. Furthermore, it was necessary to create a new amplicon processing protocol for the 454 Roche GS FLX sequencer to achieve compatibility with the amplicons generated in the previous antibody amplification process.

Using the novel immunoglobulin-amplification and sequencing method, more than 3.5 million sequences were obtained from a representative group of 14 healthy individuals of different age and gender. At first, a comprehensive analysis of the distribution of V(D)J genes in the donor was performed. Some genes, which were not covered in previous studies, could be included in the study and their relative occurrence evaluated. Moreover, for the first time, the VDJ gene distribution in conjunction with their immunoglobulin class has been analyzed. The results of the analyses show that there are differences in the immunoglobulin repertoire of

young (19-30 years) and elderly people (> 50 years). A reduction in the ability to change the classes of antibodies is observed and this correlates with age. This is in line with current studies, where the reduction in class switch is discussed as a cause for decreased vaccine efficacy in the elderly population. In this work, it is shown that the influence of age on the ability to change the immunoglobulin classes can be analyzed by cDNA sequencing. The results also strongly suggest that the senescence of the immune system begins in an age range between 50 and 60 years. A correlation between gender and the ability to change the classes of antibodies could, however, not be detected in this study.

The developed methods for sequencing immunoglobulin repertoire allow entirely new insights into the immune system. Due to the improved experimental set-up, the complexity of the variables that can be analyzed from a single sample increases considerably. In future, changes of the immune system in healthy and diseased individuals can be measured and analyzed on a new, unrivaled level of complexity.

The methods developed during my PhD and the obtained results provide a solid basis for future research addressing the analysis and verification of disease-specific changes in the immune system.

10. Zusammenfassung

Die wichtigste Rolle des adaptiven Immunsystems ist der Schutz des Körpers vor Pathogenen. Immunglobuline sind ein wesentlicher Bestandteil dieses Systems. Die große Vielfalt von Immunglobulinen wird durch einen komplexen Mechanismus der genetischen Reorganisation aus einem vorgegebenen Satz von Gensegmenten auf chromosomaler Ebene und anschließender Affinitätsreifung erhalten. Dadurch wird letztlich eine spezifische Affinität für ein bestimmtes Pathogen erreicht. Für eine erfolgreiche Immunantwort sind außerdem die Effektor-Funktionen der konstanten Region des Immunoglobulins, die in allen Immunglobulin-Klassen unterschiedlich sind, entscheidend. Die jüngsten Innovationen in der Sequenz-Analyse von DNA erlauben seit kurzem, zellgebundene (B-Zell-Rezeptoren) und sezernierte Immunglobuline von menschlichen Spendern im großen Umfang auf cDNA-Ebene zu sequenzieren und immunologisch zu bewerten.

Diese technologischen Innovationen, die generell als „Next-Generation-Sequencing“ bezeichnet werden und speziell die der sogenannten Pyro-Sequenzierung bildeten die Basis der vorliegenden Untersuchung. Ziel war die Entwicklung eines Verfahrens zur Sequenzierung von Antikörpern durch Pyro-Sequenzierung von cDNA, das alle Gene sowie alle Immunglobulin-Klassen quantitativ und qualitativ messbar macht. Dafür sollte das Immunglobulin-Repertoire aus peripheren Blutproben von gesunden Spendern unterschiedlichen Alters und Geschlechts analysiert werden.

Die in dieser Arbeit neu entwickelte Immunglobulin-spezifische cDNA-Amplifikationsmethode ist unabhängig von der Verwendung von Primern, die für V-Gene spezifisch sind. Sie ermöglicht zudem die Verifizierung der zugehörigen Immunglobulin-Klassen und ihrer Untergruppen in den erhaltenen Sequenzen. Der Einfluss der Vorbereitung auf die repräsentative Qualität der Proben und deren cDNA-Vervielfältigung wird durch eine entwickelte Emulsions-PCR-Methode minimiert. Des Weiteren war es notwendig, ein neues Amplicon-Aufarbeitungsverfahren für den 454 Roche GS FLX Sequenzer zu erstellen, um Kompatibilität mit den vorangegangenen Schritten der Amplicon-Generierung zu erreichen.

Mittels der neuartigen Immunglobulin-Amplifikations- und Sequenzier-Methode konnten aus einer repräsentativen Gruppe von 14 gesunden Individuen unterschiedlichen Alters und Geschlechts mehr als 3,5 Millionen Sequenzen gewonnen werden. Im ersten Schritt wurde eine umfassende Analyse der Verteilung der V(D)J-Genen innerhalb der Spender durchgeführt. Dabei konnten auch Gene, die in früheren Studien nicht abgedeckt wurden, in die Untersuchung einbezogen werden und deren relatives Vorkommen evaluiert werden.

Überdies wurde erstmalig die VDJ-Gen-Verteilung im Zusammenhang mit deren Immunglobulin-Klasse analysiert.

Das Ergebnis der Analyse zeigt, dass junge (19-30 Jahre) und ältere Menschen (>50 Jahre) sich in Bezug auf ihr jeweiliges Immunglobulin-Repertoir unterschiedlich verhalten. Die Reduktion der Fähigkeit zum Wechsel der Antikörperklassen korreliert mit dem Alter – in aktuellen Untersuchungen wird dies als Ursache für verringerte Impfstoffwirksamkeit bei älteren Menschen diskutiert. In dieser Arbeit wird deutlich, dass der Einfluss des Alters auf die Fähigkeit zum Wechsel der Antikörperklassen durch Immunglobulin-Sequenzierung analysiert werden kann. Die Ergebnisse lassen zudem darauf schließen, dass in einem Alter zwischen 50 und 60 Jahren die Seneszenz des Immunsystems beginnt. Eine Korrelation zwischen Geschlecht und Fähigkeit zum Wechsel der Antikörperklassen konnte in dieser Untersuchung dagegen nicht nachgewiesen werden.

Durch die hier entwickelten Verfahren zur Immunglobulin-Sequenzierung ergeben sich ganz neue Erkenntnisse in Bezug auf das Immunsystem. Aufgrund des verbesserten Versuchsaufbaus hat sich die Komplexität an Messgrößen, die aus einer einzigen Probe analysiert werden können, deutlich erhöht. In Zukunft können die Veränderungen des Immunsystems in Gesunden und Erkrankten in einer neuen, vorher unerreichten Komplexität gemessen und analysiert werden.

Die im Rahmen meiner Promotion entwickelten Verfahren und die damit erhaltenen Ergebnisse bieten eine solide Grundlage für zukünftige wissenschaftliche Fragestellungen, die der Entdeckung und Verifizierung krankheitsspezifischer Veränderungen des Immunsystems dienen.

11. References

1. Ken Murphy, P.T., Mark Walport *Janway´s Immunobiology seventh edition*. 2008.
2. Edward, J., *An Inquiry into the Causes and Effects of the Variolae Vaccinae, a Disease Discovered in Some of the Western Countries of England, Particularly Gloucestershire, and Known by the Name of "The Cow Pox".* Vol. 1923:84. 1798: Reprinted by Milan: R Lier & Co.
3. Pulendran, B. and R. Ahmed, *Immunological mechanisms of vaccination.* Nat Immunol, 2011. **12**(6): p. 509-17.
4. Blom, B. and H. Spits, *Development of human lymphoid cells.* Annu Rev Immunol, 2006. **24**: p. 287-320.
5. Metz, M. and M. Maurer, *Innate immunity and allergy in the skin.* Curr Opin Immunol, 2009. **21**(6): p. 687-93.
6. Akira, S., S. Uematsu, and O. Takeuchi, *Pathogen recognition and innate immunity.* Cell, 2006. **124**(4): p. 783-801.
7. Janeway, C.A., Jr. and R. Medzhitov, *Innate immune recognition.* Annu Rev Immunol, 2002. **20**: p. 197-216.
8. Janeway, C.A., Jr., *Approaching the asymptote? Evolution and revolution in immunology.* Cold Spring Harb Symp Quant Biol, 1989. **54 Pt 1**: p. 1-13.
9. Hacker, G., V. Redecke, and H. Hacker, *Activation of the immune system by bacterial CpG-DNA.* Immunology, 2002. **105**(3): p. 245-51.
10. Rodriguez, R.M., A. Lopez-Vazquez, and C. Lopez-Larrea, *Immune systems evolution.* Adv Exp Med Biol, 2012. **739**: p. 237-51.
11. Bottazzi, B., et al., *An integrated view of humoral innate immunity: pentraxins as a paradigm.* Annu Rev Immunol, 2010. **28**: p. 157-83.
12. Danilova, N., *The evolution of adaptive immunity.* Adv Exp Med Biol, 2012. **738**: p. 218-35.
13. Koch, U. and F. Radtke, *Mechanisms of T cell development and transformation.* Annu Rev Cell Dev Biol, 2011. **27**: p. 539-62.
14. Schatz, D.G. and P.C. Swanson, *V(D)J recombination: mechanisms of initiation.* Annu Rev Genet, 2011. **45**: p. 167-202.
15. Peled, J.U., et al., *The biochemistry of somatic hypermutation.* Annu Rev Immunol, 2008. **26**: p. 481-511.
16. Meffre, E. and H. Wardemann, *B-cell tolerance checkpoints in health and autoimmunity.* Curr Opin Immunol, 2008. **20**(6): p. 632-8.
17. Davis, M.M., *A prescription for human immunology.* Immunity, 2008. **29**(6): p. 835-8.
18. Davis, M.M., *Immunology taught by humans.* Sci Transl Med, 2012. **4**(117): p. 117fs2.
19. Nobelprize.org. *The Nobel Prize in Physiology or Medicine 1972*. 2 Aug 2012 Available from: http://www.nobelprize.org/nobel_prizes/medicine/laureates/1972/.
20. Nobelprize.org. *The Nobel Prize in Physiology or Medicine 1984*. 2 Aug 2012 Available from: http://www.nobelprize.org/nobel_prizes/medicine/laureates/1984/.
21. Nobelprize.org. *The Nobel Prize in Physiology or Medicine 1987*. 2 Aug 2012 Available from: http://www.nobelprize.org/nobel_prizes/medicine/laureates/1987/.
22. Weinstein, J.A., et al., *High-throughput sequencing of the zebrafish antibody repertoire.* Science, 2009. **324**(5928): p. 807-10.
23. Schroeder, H.W., Jr. and L. Cavacini, *Structure and function of immunoglobulins.* J Allergy Clin Immunol, 2010. **125**(2 Suppl 2): p. S41-52.
24. Williams, A.F. and A.N. Barclay, *The immunoglobulin superfamily--domains for cell surface recognition.* Annu Rev Immunol, 1988. **6**: p. 381-405.
25. Leder, P., *The genetics of antibody diversity.* Sci Am, 1982. **246**(5): p. 102-15.
26. Tonegawa, S., *Somatic generation of antibody diversity.* Nature, 1983. **302**(5909): p. 575-81.

27. Lefranc, M.P., *Nomenclature of the human immunoglobulin kappa (IGK) genes.* Exp Clin Immunogenet, 2001. **18**(3): p. 161-74.
28. Lefranc, M.P., *Nomenclature of the human immunoglobulin lambda (IGL) genes.* Exp Clin Immunogenet, 2001. **18**(4): p. 242-54.
29. Pandey, J.P., *Immunoglobulin GM and KM allotypes and vaccine immunity.* Vaccine, 2000. **19**(6): p. 613-7.
30. Lefranc, M.P. and G. Lefranc, *Human Gm, Km, and Am allotypes and their molecular characterization: a remarkable demonstration of polymorphism.* Methods Mol Biol, 2012. **882**: p. 635-80.
31. Geisberger, R., M. Lamers, and G. Achatz, *The riddle of the dual expression of IgM and IgD.* Immunology, 2006. **118**(4): p. 429-37.
32. Riesbeck, K. and T. Nordstrom, *Structure and immunological action of the human pathogen Moraxella catarrhalis IgD-binding protein.* Crit Rev Immunol, 2006. **26**(4): p. 353-76.
33. Aalberse, R.C., et al., *Immunoglobulin G4: an odd antibody.* Clin Exp Allergy, 2009. **39**(4): p. 469-77.
34. Macpherson, A.J., et al., *The immune geography of IgA induction and function.* Mucosal Immunol, 2008. **1**(1): p. 11-22.
35. Woof, J.M. and J. Mestecky, *Mucosal immunoglobulins.* Immunol Rev, 2005. **206**: p. 64-82.
36. Brandtzaeg, P. and F.E. Johansen, *Mucosal B cells: phenotypic characteristics, transcriptional regulation, and homing properties.* Immunol Rev, 2005. **206**: p. 32-63.
37. Cerutti, A., *The regulation of IgA class switching.* Nat Rev Immunol, 2008. **8**(6): p. 421-34.
38. Stanworth, D.R., *The discovery of IgE.* Allergy, 1993. **48**(2): p. 67-71.
39. Burton, O.T. and H.C. Oettgen, *Beyond immediate hypersensitivity: evolving roles for IgE antibodies in immune homeostasis and allergic diseases.* Immunol Rev, 2011. **242**(1): p. 128-43.
40. Bennich, H.H., et al., *Immunoglobulin E: a new class of human immunoglobulin.* Immunology, 1968. **15**(3): p. 323-4.
41. Kubo, S., et al., *Long term maintenance of IgE-mediated memory in mast cells in the absence of detectable serum IgE.* J Immunol, 2003. **170**(2): p. 775-80.
42. Gould, H.J., et al., *The biology of IGE and the basis of allergic disease.* Annu Rev Immunol, 2003. **21**: p. 579-628.
43. Ricklin, D., et al., *Complement: a key system for immune surveillance and homeostasis.* Nat Immunol, 2010. **11**(9): p. 785-97.
44. Bindon, C.I., et al., *Human monoclonal IgG isotypes differ in complement activating function at the level of C4 as well as C1q.* J Exp Med, 1988. **168**(1): p. 127-42.
45. Oettinger, M.A., et al., *RAG-1 and RAG-2, adjacent genes that synergistically activate V(D)J recombination.* Science, 1990. **248**(4962): p. 1517-23.
46. Dudley, D.D., et al., *Mechanism and control of V(D)J recombination versus class switch recombination: similarities and differences.* Adv Immunol, 2005. **86**: p. 43-112.
47. Kelsoe, G., *The germinal center: a crucible for lymphocyte selection.* Semin Immunol, 1996. **8**(3): p. 179-84.
48. Victora, G.D. and M.C. Nussenzweig, *Germinal centers.* Annu Rev Immunol, 2012. **30**: p. 429-57.
49. MacLennan, I.C., *Germinal centers.* Annu Rev Immunol, 1994. **12**: p. 117-39.
50. Stavnezer, J., J.E. Guikema, and C.E. Schrader, *Mechanism and regulation of class switch recombination.* Annu Rev Immunol, 2008. **26**: p. 261-92.
51. Stavnezer, J., *Antibody class switching.* Adv Immunol, 1996. **61**: p. 79-146.
52. Crotty, S., *Follicular helper CD4 T cells (TFH).* Annu Rev Immunol, 2011. **29**: p. 621-63.
53. Brunham, R.C. and K.M. Coombs, *In celebration of the 200th anniversary of Edward Jenner's Inquiry into the causes and effects of the variolae vaccinae.* Can J Infect Dis, 1998. **9**(5): p. 310-3.

54. Thompson, W.W., et al., *Estimating influenza-associated deaths in the United States.* Am J Public Health, 2009. **99 Suppl 2**: p. S225-30.
55. McElhaney, J.E. and R.B. Effros, *Immunosenescence: what does it mean to health outcomes in older adults?* Curr Opin Immunol, 2009. **21**(4): p. 418-24.
56. Chen, W.H., et al., *Vaccination in the elderly: an immunological perspective.* Trends Immunol, 2009. **30**(7): p. 351-9.
57. Liu, W.M., et al., *Aging and impaired immunity to influenza viruses: implications for vaccine development.* Hum Vaccin, 2011. **7 Suppl**: p. 94-8.
58. Aspinall, R. and D. Andrew, *Thymic involution in aging.* J Clin Immunol, 2000. **20**(4): p. 250-6.
59. Reber, A.J., et al., *Immunosenescence and Challenges of Vaccination against Influenza in the Aging Population.* Aging Dis, 2012. **3**(1): p. 68-90.
60. Frasca, D. and B.B. Blomberg, *Aging affects human B cell responses.* J Clin Immunol, 2011. **31**(3): p. 430-5.
61. Aberle, J.H., et al., *Mechanistic insights into the impairment of memory B cells and antibody production in the elderly.* Age (Dordr), 2012.
62. Goodwin, K., C. Viboud, and L. Simonsen, *Antibody response to influenza vaccination in the elderly: a quantitative review.* Vaccine, 2006. **24**(8): p. 1159-69.
63. Vu, T., et al., *A meta-analysis of effectiveness of influenza vaccine in persons aged 65 years and over living in the community.* Vaccine, 2002. **20**(13-14): p. 1831-6.
64. Sasaki, S., et al., *Limited efficacy of inactivated influenza vaccine in elderly individuals is associated with decreased production of vaccine-specific antibodies.* J Clin Invest, 2011. **121**(8): p. 3109-19.
65. Stiasny, K., et al., *Age affects quantity but not quality of antibody responses after vaccination with an inactivated flavivirus vaccine against tick-borne encephalitis.* PLoS One, 2012. **7**(3): p. e34145.
66. Weinberger, B., et al., *Decreased antibody titers and booster responses in tick-borne encephalitis vaccinees aged 50-90 years.* Vaccine, 2010. **28**(20): p. 3511-5.
67. Frasca, D., et al., *Age effects on B cells and humoral immunity in humans.* Ageing Res Rev, 2011. **10**(3): p. 330-5.
68. Linton, P.J. and K. Dorshkind, *Age-related changes in lymphocyte development and function.* Nat Immunol, 2004. **5**(2): p. 133-9.
69. Kircher, M. and J. Kelso, *High-throughput DNA sequencing--concepts and limitations.* Bioessays, 2010. **32**(6): p. 524-36.
70. Sanger, F., et al., *Nucleotide sequence of bacteriophage phi X174 DNA.* Nature, 1977. **265**(5596): p. 687-95.
71. Stranneheim, H. and J. Lundeberg, *Stepping stones in DNA sequencing.* Biotechnol J, 2012. **7**(9): p. 1063-73.
72. Watson, J.D. and F.H. Crick, *Molecular structure of nucleic acids: a structure for deoxyribose nucleic acid. J.D. Watson and F.H.C. Crick. Published in Nature, number 4356 April 25, 1953.* Nature, 1974. **248**(5451): p. 765.
73. Nirenberg, M.W. and J.H. Matthaei, *The dependence of cell-free protein synthesis in E. coli upon naturally occurring or synthetic polyribonucleotides.* Proc Natl Acad Sci U S A, 1961. **47**: p. 1588-602.
74. Sanger, F., S. Nicklen, and A.R. Coulson, *DNA sequencing with chain-terminating inhibitors.* Proc Natl Acad Sci U S A, 1977. **74**(12): p. 5463-7.
75. Fleischmann, R.D., et al., *Whole-genome random sequencing and assembly of Haemophilus influenzae Rd.* Science, 1995. **269**(5223): p. 496-512.
76. Lander, E.S., et al., *Initial sequencing and analysis of the human genome.* Nature, 2001. **409**(6822): p. 860-921.
77. Venter, J.C., et al., *The sequence of the human genome.* Science, 2001. **291**(5507): p. 1304-51.
78. Quail, M., et al., *A tale of three next generation sequencing platforms: comparison of Ion torrent, pacific biosciences and illumina MiSeq sequencers.* BMC Genomics, 2012. **13**(1): p. 341.

79. Fedurco, M., et al., *BTA, a novel reagent for DNA attachment on glass and efficient generation of solid-phase amplified DNA colonies.* Nucleic Acids Res, 2006. **34**(3): p. e22.
80. Turcatti, G., et al., *A new class of cleavable fluorescent nucleotides: synthesis and optimization as reversible terminators for DNA sequencing by synthesis.* Nucleic Acids Res, 2008. **36**(4): p. e25.
81. Adessi, C., et al., *Solid phase DNA amplification: characterisation of primer attachment and amplification mechanisms.* Nucleic Acids Res, 2000. **28**(20): p. E87.
82. Bentley, D.R., et al., *Accurate whole human genome sequencing using reversible terminator chemistry.* Nature, 2008. **456**(7218): p. 53-9.
83. Liu, L., et al., *Comparison of next-generation sequencing systems.* J Biomed Biotechnol, 2012. **2012**: p. 251364.
84. Metzker, M.L., *Sequencing technologies - the next generation.* Nat Rev Genet, 2010. **11**(1): p. 31-46.
85. Melamede, R.J., *Automatable process for sequencing nucleotide.*, 1985: USA.
86. Loman, N.J., et al., *Performance comparison of benchtop high-throughput sequencing platforms.* Nat Biotechnol, 2012. **30**(5): p. 434-9.
87. McKernan, K.J., et al., *Sequence and structural variation in a human genome uncovered by short-read, massively parallel ligation sequencing using two-base encoding.* Genome Res, 2009. **19**(9): p. 1527-41.
88. Voelkerding, K.V., S.A. Dames, and J.D. Durtschi, *Next-generation sequencing: from basic research to diagnostics.* Clin Chem, 2009. **55**(4): p. 641-58.
89. Koboldt, D.C., et al., *Massively parallel sequencing approaches for characterization of structural variation.* Methods Mol Biol, 2012. **838**: p. 369-84.
90. Margulies, M., et al., *Genome sequencing in microfabricated high-density picolitre reactors.* Nature, 2005. **437**(7057): p. 376-80.
91. Rothberg, J.M. and J.H. Leamon, *The development and impact of 454 sequencing.* Nat Biotechnol, 2008. **26**(10): p. 1117-24.
92. Rothberg, J.M., et al., *An integrated semiconductor device enabling non-optical genome sequencing.* Nature, 2011. **475**(7356): p. 348-52.
93. Nawy, T., *Speed-reading DNA in the dark.* Nat Methods, 2011. **8**(9): p. 708-9.
94. Eid, J., et al., *Real-time DNA sequencing from single polymerase molecules.* Science, 2009. **323**(5910): p. 133-8.
95. McCarthy, A., *Third generation DNA sequencing: pacific biosciences' single molecule real time technology.* Chem Biol, 2010. **17**(7): p. 675-6.
96. *VBASE*. 2003 - 2004; Available from: http://vbase.mrc-cpe.cam.ac.uk/index.php?&MMN_position=1:1.
97. Sblattero, D. and A. Bradbury, *A definitive set of oligonucleotide primers for amplifying human V regions.* Immunotechnology, 1998. **3**(4): p. 271-8.
98. Saiki, R.K., et al., *Primer-directed enzymatic amplification of DNA with a thermostable DNA polymerase.* Science, 1988. **239**(4839): p. 487-91.
99. Johnson, G. and T.T. Wu, *Kabat Database and its applications: future directions.* Nucleic Acids Res, 2001. **29**(1): p. 205-6.
100. Retter, I., et al., *VBASE2, an integrative V gene database.* Nucleic Acids Res, 2005. **33**(Database issue): p. D671-4.
101. Lefranc, M.P., et al., *IMGT, the international ImMunoGeneTics database.* Nucleic Acids Res, 1999. **27**(1): p. 209-12.
102. Lefranc, M.P., *IMGT, the international ImMunoGeneTics database.* Nucleic Acids Res, 2003. **31**(1): p. 307-10.
103. Jiang, N., et al., *Determinism and stochasticity during maturation of the zebrafish antibody repertoire.* Proc Natl Acad Sci U S A, 2011. **108**(13): p. 5348-53.
104. Boyd, S.D., et al., *Measurement and clinical monitoring of human lymphocyte clonality by massively parallel VDJ pyrosequencing.* Sci Transl Med, 2009. **1**(12): p. 12ra23.

105. Wang, Y., et al., *Genomic screening by 454 pyrosequencing identifies a new human IGHV gene and sixteen other new IGHV allelic variants.* Immunogenetics, 2011. **63**(5): p. 259-265.
106. Jackson, K.J., et al., *Divergent human populations show extensive shared IGK rearrangements in peripheral blood B cells.* Immunogenetics, 2012. **64**(1): p. 3-14.
107. Fischer, N., *Sequencing antibody repertoires: the next generation.* MAbs, 2011. **3**(1): p. 17-20.
108. Boyd, S.D., et al., *Individual variation in the germline Ig gene repertoire inferred from variable region gene rearrangements.* J Immunol, 2010. **184**(12): p. 6986-92.
109. Glanville, J., et al., *Precise determination of the diversity of a combinatorial antibody library gives insight into the human immunoglobulin repertoire.* Proc Natl Acad Sci U S A, 2009. **106**(48): p. 20216-21.
110. Prabakaran, P., et al., *Expressed antibody repertoires in human cord blood cells: 454 sequencing and IMGT/HighV-QUEST analysis of germline gene usage, junctional diversity, and somatic mutations.* Immunogenetics, 2012. **64**(5): p. 337-50.
111. Glanville, J., et al., *Naive antibody gene-segment frequencies are heritable and unaltered by chronic lymphocyte ablation.* Proc Natl Acad Sci U S A, 2011. **108**(50): p. 20066-71.
112. Wu, Y.C., et al., *High-throughput immunoglobulin repertoire analysis distinguishes between human IgM memory and switched memory B-cell populations.* Blood, 2010. **116**(7): p. 1070-8.
113. Briney, B.S., J.R. Willis, and J.E. Crowe, Jr., *Human Peripheral Blood Antibodies with Long HCDR3s Are Established Primarily at Original Recombination Using a Limited Subset of Germline Genes.* PLoS One, 2012. **7**(5): p. e36750.
114. Arnaout, R., et al., *High-resolution description of antibody heavy-chain repertoires in humans.* PLoS One, 2011. **6**(8): p. e22365.
115. Alamyar, E., et al., *IMGT/HighV-QUEST: the IMGT(R) web portal for immunoglobulin (IG) or antibody and T cell receptor (TR) analysis from NGS high throughput and deep sequencing.* Immunome Res, 2012. **8**(1): p. 26.
116. Pareek, C.S., R. Smoczynski, and A. Tretyn, *Sequencing technologies and genome sequencing.* J Appl Genet, 2011. **52**(4): p. 413-35.
117. Zhu, Z. and D.S. Dimitrov, *Construction of a large naive human phage-displayed Fab library through one-step cloning.* Methods Mol Biol, 2009. **525**: p. 129-42, xv.
118. Lim, T.S., et al., *V-gene amplification revisited - An optimised procedure for amplification of rearranged human antibody genes of different isotypes.* N Biotechnol, 2010. **27**(2): p. 108-17.
119. Schutze, T., et al., *Probing the SELEX process with next-generation sequencing.* PLoS One, 2011. **6**(12): p. e29604.
120. Chase, J.W. and K.R. Williams, *Single-stranded DNA binding proteins required for DNA replication.* Annu Rev Biochem, 1986. **55**: p. 103-36.
121. Liu, J., et al., *Novel, fluorescent, SSB protein chimeras with broad utility.* Protein Sci, 2011. **20**(6): p. 1005-20.
122. Chou, Q., *Minimizing deletion mutagenesis artifact during Taq DNA polymerase PCR by E. coli SSB.* Nucleic Acids Res, 1992. **20**(16): p. 4371.
123. Kunkel, T.A., R.R. Meyer, and L.A. Loeb, *Single-strand binding protein enhances fidelity of DNA synthesis in vitro.* Proc Natl Acad Sci U S A, 1979. **76**(12): p. 6331-5.
124. Perales, C., et al., *Enhancement of DNA, cDNA synthesis and fidelity at high temperatures by a dimeric single-stranded DNA-binding protein.* Nucleic Acids Res, 2003. **31**(22): p. 6473-80.
125. Prakash, A. and G.E. Borgstahl, *The structure and function of replication protein a in DNA replication.* Subcell Biochem, 2012. **62**: p. 171-96.
126. B., F.B.F.a.S., *Demonstration of the Expand PCR System's greater fidelity and higher yields with a lacI-based fidelity assay.* Biochemica, 1995. **Biochemica**(2): p. 34,35
127. Schütze, T., et al., *A streamlined protocol for emulsion polymerase chain reaction and subsequent purification.* Anal Biochem, 2011. **410**(1): p. 155-7.

128. Tawfik, D.S. and A.D. Griffiths, *Man-made cell-like compartments for molecular evolution.* Nat Biotechnol, 1998. **16**(7): p. 652-6.
129. Nakano, M., et al., *Single-molecule PCR using water-in-oil emulsion.* J Biotechnol, 2003. **102**(2): p. 117-24.
130. Shao, K., et al., *Emulsion PCR: a high efficient way of PCR amplification of random DNA libraries in aptamer selection.* PLoS One, 2011. **6**(9): p. e24910.
131. Hori, M., H. Fukano, and Y. Suzuki, *Uniform amplification of multiple DNAs by emulsion PCR.* Biochem Biophys Res Commun, 2007. **352**(2): p. 323-8.
132. Williams, R., et al., *Amplification of complex gene libraries by emulsion PCR.* Nat Methods, 2006. **3**(7): p. 545-50.
133. Kainz, P., *The PCR plateau phase - towards an understanding of its limitations.* Biochim Biophys Acta, 2000. **1494**(1-2): p. 23-7.
134. Loh, E.Y., et al., *Polymerase chain reaction with single-sided specificity: analysis of T cell receptor delta chain.* Science, 1989. **243**(4888): p. 217-20.
135. Lefranc, M.P., et al., *IMGT, a system and an ontology that bridge biological and computational spheres in bioinformatics.* Brief Bioinform, 2008. **9**(4): p. 263-75.
136. Briney, B.S., et al., *Frequency and genetic characterization of V(DD)J recombinants in the human peripheral blood antibody repertoire.* Immunology, 2012.
137. Alamyar, E., et al. *IMGT/HighV-QUEST: A High-Throughput System and Web Portal for the Analysis of Rearranged Nucleotide Sequences of Antigen Receptors*. in *11émes Journées Ouvertes de Biologie, Informatique et Mathématiques (JOBIM)*. 2010. Montpellier.
138. Ruzickova, S., et al., *Chronic lymphocytic leukemia preceded by cold agglutinin disease: intraclonal immunoglobulin light-chain diversity in V(H)4-34 expressing single leukemic B cells.* Blood, 2002. **100**(9): p. 3419-22.
139. Bhat, N.M., et al., *VH4-34 encoded antibody in systemic lupus erythematosus: effect of isotype.* J Rheumatol, 2002. **29**(10): p. 2114-21.
140. Bhat, N.M., et al., *Recognition of auto- and exoantigens by V4-34 gene encoded antibodies.* Scand J Immunol, 2000. **51**(2): p. 134-40.
141. van Vollenhoven, R.F., et al., *VH4-34 encoded antibodies in systemic lupus erythematosus: a specific diagnostic marker that correlates with clinical disease characteristics.* J Rheumatol, 1999. **26**(8): p. 1727-33.
142. Hadzidimitriou, A., et al., *Evidence for the significant role of immunoglobulin light chains in antigen recognition and selection in chronic lymphocytic leukemia.* Blood, 2009. **113**(2): p. 403-11.
143. Jackson, K.J., et al., *Benchmarking the performance of human antibody gene alignment utilities using a 454 sequence dataset.* Bioinformatics, 2010. **26**(24): p. 3129-30.
144. Stamatopoulos, K., et al., *Immunoglobulin light chain repertoire in chronic lymphocytic leukemia.* Blood, 2005. **106**(10): p. 3575-83.
145. Daha, N.A., et al., *Complement activation by (auto-) antibodies.* Mol Immunol, 2011. **48**(14): p. 1656-65.
146. Arnaout, R.A., *Specificity and overlap in gene segment-defined antibody repertoires.* BMC Genomics, 2005. **6**: p. 148.

12. Curriculum vitae

The curriculum vitae is not contained in the online version for reasons of data protection.