Characterization of Immunoglobulin Repertoires after Vaccination in OmniRat^{TM} and Monoclonal CD5^{+} B Cell Expansion in A20^{BKO}sCYLD^{BOE} Mice using an Ion Torrent PGM High-Throughput Sequencing Platform

Dissertation zur Erlangung der naturwissenschaftlichen

Doktorwürde durch den Fachbereich I – Psychobiologie der

UNIVERSITÄT TRIER

Vorgelegt von Dipl. Chem. Jean-Philippe Bürckert

Gutachter:

Prof. Dr. C. P. Muller

Prof. Dr. J. Meyer

Luxembourg, Dezember 2017

Dissertationsort: **Trier**

This doctoral thesis has been performed at the Vaccinology and B Cell Immunology Unit, Department of Infection and Immunity, Luxembourg Institute of Health, Luxembourg

Under the guidance of

Prof. Dr. Claude P. Muller, Department of Infection and Immunity, Luxembourg Institute of Health, Luxembourg

and

Prof. Dr. Jobst Meyer, Department of Neurobehavioural Genetics, University of Trier, Germany

with funding from the

Fonds National de la Recherche (FNR) Luxembourg under the individual “Aides à la Formation Recherche” (AFR) PhD grant No. 7039209

Content

List of Figures

List of Tables

Abbreviations

AID	Activation-induced cytidine deaminase
Abs	Antibodies
ALUM	alum hydroxide
B-ALL	B Cell acute lymphoblastic leukemia
BAFF	B cell activating factor of the TNF family
BaP	Benzo[a]Pyrene
BaP-TT	Benzo[a]Pyrene-Tetanus toxoid
B-CLL	B cell chronic lymphocytic leukemia
BCR	B cell receptor
BM	bone marrow
C	Immunoglobulin Constant region
CDR	Complementary-determining region
CDR1	Complementary-determining region 1
CDR2	Complementary-determining region 2
CDR3	Complementary-determining region 3
CLL	Chronic lymphocytic leukemia
CMV	cytomegalovirus
CSR	Class switch recombination
D	Immunoglobulin Diversity gene segment
DUBs	Deubiquitinating enzymes
EBV	Epstein-Barr virus
EPT	Endpoint titers
F	Fusion glycoprotein of the measles virus
Fab	Antigen binding fragment
FDR	False discovery rate
FL	Full length
FR	Framework region
FR1	Framework region 1
FR2	Framework region 2
FR3	Framework region 3
GC	Germinal center
H	Hemagglutinin glycoprotein of the measles virus
HF	Hemagglutinin and fusion proteins of the measles virus
Hib	*Haemophilus influenzae* type B
HTS	High-throughput sequencing
IAP	Inhibitor of apoptosis
IG	Immunoglobulin
IgA	Immunoglobulin A
IgD	Immunoglobulin D
IgE	Immunoglobulin E
IgG	Immunoglobulin G
IgM	Immunoglobulin M
IGH	Immunoglobulin heavy chain
IGHV	Immunoglobulin heavy chain V gene segment
IGHD	Immunoglobulin heavy chain D gene segment
IGHJ	Immunoglobulin heavy chain J gene segment
IGL	Immunoglobulin light chain
i.m.	intra muscular
IMGT	ImMunoGeneTics database

Indel	Insertions and deletions of nucleotides
i.p.	intraperitoneal
i.m.	intramuscular
i.v.	intravenously
J	Immunoglobulin Joining gene segment
K48	Lysine 48
LEF-1	Lymphoid enhancer-binding factor 1
LN	Lymph nodes
M-CLL	Mutated form of chronic lymphocytic leukemia
MenC	Group C meningococcal
MID	Multiplex identifier
MRD	Minimal residual disease
MV	Measles virus
MVA	Modified Vaccinia virus Ankara
NF-κB	Nuclear factor kappa B
nt	Nucleotide
PB	Peripheral blood
PBMC	peripheral blood B cell
PerC	Peritoneal cavity
PGM	(Ion Torrent) Personal Genome Machine
ROSIE	Rosetta Online Server that Includes Everyone
RSV	Respiratory syncytial virus
sCYLD	short splice variant of CYLD
SHM	Somatic hypermutation
ssUID	Single side unique molecular identifier
TCL1	T-cell leukemia 1
TIgGer	Tool for Ig Genotype Elucidation via Rep-Seq
TT	Tetanus toxoid
U-CLL	Unmutated form of chronic lymphocytic leukemia
UID	Unique identifier
V	Variable
VST-counts	Variance stabilizing transformed count data
WT	Wild type

General Abstract

With the advent of high-throughput sequencing (HTS), profiling immunoglobulin (IG) repertoires has become an essential part of immunological research. The dissection of IG repertoires promises to transform our understanding of the adaptive immune system dynamics. Advances in sequencing technology now also allow the use of the Ion Torrent Personal Genome Machine (PGM) to cover the full length of IG mRNA transcripts. The applications of this bench-top scale HTS platform range from identification of new therapeutic antibodies to the deconvolution of malignant B cell tumors. In the context of this thesis, the usability of the PGM is assessed to investigate the IG heavy chain (IGH) repertoires of animal models. First, an innovate bioinformatics approach is presented to identify antigen-driven IGH sequences from bulk sequenced bone marrow samples of transgenic humanized rats, expressing a human IG repertoire (OmniRat^{TM}). We show, that these rats mount a convergent IGH CDR3 response towards measles virus hemagglutinin protein and tetanus toxoid, with high similarity to human counterparts. In the future, databases could contain all IGH CDR3 sequences with known specificity to mine IG repertoire datasets for past antigen exposures, ultimately reconstructing the immunological history of an individual. Second, a unique molecular identifier (UID) based HTS approach and network property analysis is used to characterize the CLL-like CD5^{+} B cell expansion of A20^{BKO} mice overexpressing a natural short splice variant of the CYLD gene (A20^{BKO}sCYLD^{BOE}). We could determine, that in these mice, overexpression of sCYLD leads to unmutated subvariant of CLL (U-CLL). Furthermore, we found that this short splice variant is also seen in human CLL patients highlighting it as important target for future investigations. Third, the UID based HTS approach is improved by adapting it to the PGM sequencing technology and applying a custom-made data processing pipeline including the ImMunoGeneTics (IMGT) database error detection. Like this, we were able to obtain correct IGH sequences with over 99.5% confidence and correct CDR3 sequences with over 99.9% confidence. Taken together, the results, protocols and sample processing strategies described in this thesis will improve the usability of animal models and the Ion Torrent PGM HTS platform in the field if IG repertoire research.

Chapter 1 - General Introduction

1.1. The structure of the adaptive immune system

B cells and T cells form the two pillars of the adaptive immune system. Together they exhibit the key feature of our body's defensive system against pathogens, infections, adverse proteins and cancer. The first detailed description of the B cell population structures and dynamics was awarded with the Nobel prize of Medicine to Susumu Tonegawa in 1983 (1). This doctoral thesis builds upon the more than 50,000 scientific publications that have been published in this field ever since (2). Utilizing high-throughput sequencing (HTS), it aims to provide a deeper understanding of B cell population dynamics in vaccination and lymphocytic leukemia using the $A20^{BKO}sCYLD^{BOE}$ mouse model as well as OmniRat™, a transgenic rat with human B Cell genes.

1.1.1. The B cell receptor

The B cell receptor (BCR) is the surface variant of the Immunoglobulin (IG) molecule. IG are glycoproteins that can bind in a key-lock like principle to pathogens, proteins and even small peptides, either neutralizing or opsonizing them in the process. Structurally, the IG molecule consists of two identical heavy chains (IGH) and two identical light chains (IGL) that are linked by disulfide bonds (**Figure 1a**). The antigen binding fragments (Fab regions) at the tip of both chains are each characterized by the three highly variable complementary determining regions (CDR1-3) and the structurally important, conserved framework regions (FR1-3). The CDR3 of the IGH exhibits the largest variability and diversity of the IG molecules and presents the key determinant for antigen binding (3, 4). As a consequence, information about the CDR3 proves to be sufficient to uniquely characterize an IG molecule (3). The IGH extents into the Constant (C or Fc) region, defining the IG Isotype and exhibiting surface and complement tissue (Fc receptor) binding capacities (5). There exist five different isotypes, IgM, IgG, IgA, IgE and IgD, which are either expressed as B cell receptor or secreted as antibody molecules differing by carboxy-terminal sequence of the IGH generated through alternative splicing of the same mRNA transcript (6). The isotypes have different biological functions and home different tissues in the body (**Table 1**).

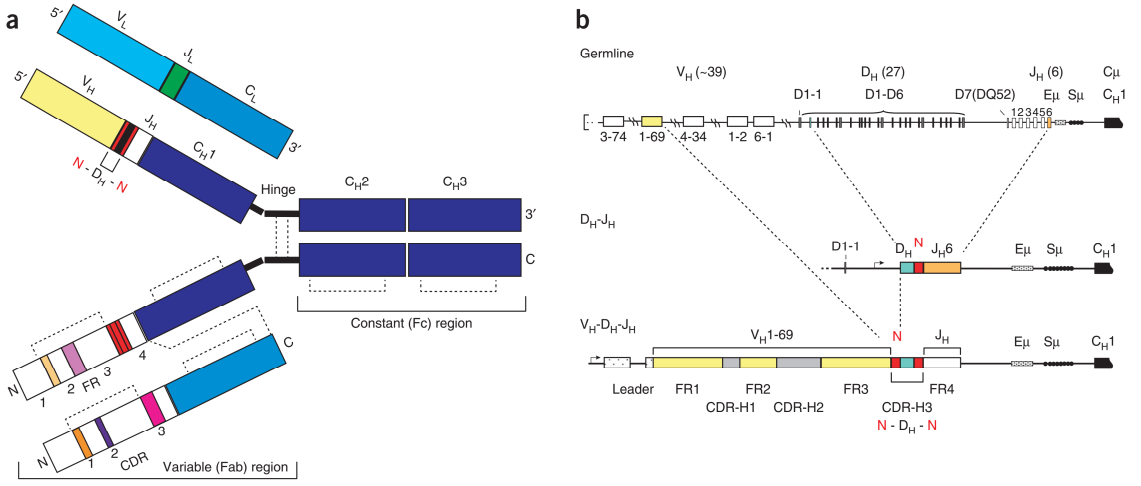


Figure 1 Schematic structure of the IG molecule and primary IG repertoire diversification mechanism. (a) Representative structure of an IgG molecule. The domains encoded by IGH Variable (IGHV), Diversity (IGHD) and Joining (IGHJ) as well as IGL Variable (IGLV) and Joining (IGHJ) gene segments are shown for the top chains. The lower chains indicate the IG FR1-3 and CDR1-3. Dashed lines represent the disulfide bonds. (b) Steps of primary IG repertoire diversification. The diversity of the primary IG repertoire is mediated by somatic recombination of IGHV, IGHD, IGHJ and IGLV, IGLJ gene segments. Additional diversity is generated by random, nontemplated addition of N-nucleotides. An IGHD and IGHJ gene are combined first into a IGH DJ segment. Afterwards an IGHV and the IGH DJ are recombined into a full length IGH VDJ chain. The IGL is recombined after the IGH. Heterodimeric pairing of the IGH and IGL then forms a complete IgM IG molecule on the surface of a naïve B cell. Eμ: IgM intronic enhancer; Sμ: tandem repeats for class-switch recombination. Numbers in parentheses refer to estimates of human germline IGHV, IGHD and IGHJ segments. Reprinted by permission from Macmillan Publishers Ltd: Nature Biotechnology (56), copyright 2014.

Table 1 Properties of IG isotypes. Adapted from (7, 8)

IG Isotype	Structure as antibody molecule	Half-life (days)	Serum conc. (mg/mL)	Functions
IgG	Monomer	0.5-9	7-23	Antiviral, allergy, toxin neutralization
IgM	Pentamer/Hexamer	1.5	5	Primary response
IgA	Mononer/Dimer	0.5-3	6	Direct toxin neutralization, viruses and bacteria
IgD	Monomer	0.04	3	Homeostasis
IgE	Dimer	0.0003	0.5	Allergy

1.1.2. The primary IG repertoire

During the process of maturation in the primary lymphoid organs (bone marrow, spleen), the hematopoietic stem cell derived Early Pro-B cells first arrange IG heavy chain Diversity (IGHD) and Joining (IGHJ) DNA gene segments, choosing from a variety of IGH loci (**Figure 1b**). In a next step, the Late Pro-B cells join this rearranged IGHDJ structure with a Variable (IGHV) gene segment. Heavy chain VDJ rearrangement is mediated by recombination activating genes 1 and 2 (*RAG1* and *RAG2*) through DNA segment deletion (9). Both processes are inherent to nucleotide deletions from DNA exonucleases, and insertions of templated palindromic nucleotides by DNA polymerases as well as non-templated nucleotides from transferases (1). At both stages the rearrangements are checked for productivity, with the possibility of allelic exclusion. About 50% of all Pre-B Cells are subsequently signaled to die by apoptosis (10). In a next step, the Pre-B cells express their B cell receptor as rearranged IGH together with a surrogate IGL to perform initial tests for self-reactivity. Such autoreactive B cells are either clonally eliminated or receive additional receptor modification (11). Afterwards, the Immature B cells rearrange IGLV and IGLJ genes, using either κ or λ light chain loci, with the possibility of loci exclusion in the case of unproductive rearrangements (5). These B cells then express a fully functional B cell IgM receptor on their surface. After passing an additional checkpoint to prevent self-reactivity, the B cells are released as immature or naïve B cells into the periphery expressing IgM alongside IgD receptors. The process described above, is referred to as the primary B cell repertoire. It forms a broad, fully functional IG repertoire, capable of reacting to almost any possible antigen, albeit only with low affinity. The diversity at this stage stems from choosing different IGH V(D)J combinations from 44-60 IGHV (12), 27 IGHD and 6 IGHJ genes (13), and 324 different IGL (200 Igκ and 124 Igλ) genes, allowing 2.5 x 10^6 different combinations (**Table 2**). Alternative IGHD reading frames, IGHD to IGHD gene-fusion, and imprecise IGHD to IGHJ and IGHV to IGHD junctions (14) further diversify the primary IG repertoire, leading to theoretically 10^{14} different possible IG molecules.

Table 2 Number of possible human IG receptor combinations

IG chain	Number of gene segments	potential combinations
IGLV κ locus	40	200 IGL kappa chains
IGLJ κ locus	5	
IGLV λ locus	31	124 IGL lambda chains
IGLJ λ locus	4	
IGHV	44	7,128 IGH chains
IGHD	27	
IGHJ	6	
Total number of combinations		**2.452×10^6**

1.1.3. The secondary IG repertoire

To overcome the affinity limitations of the primary repertoire, the adaptive immune system is able to enhance target-binding of IG molecules in an iterative process. This results in highly specialized IG molecules of superior affinity and avidity to their target antigen. When a B cell encounters its cognate antigen, and receives the necessary co-stimulations from $CD4^+$ helper T Cells (MHC class II binding and CD40-CD40L co-stimulation), it initiates the formation of a Germinal Center (GC) reaction (**Figure 2**) (15, 16). These highly organized areas coordinate the affinity maturation of B cell receptors against the specific antigen in a two-zone iteration set-up within secondary lymphoid organs (spleen, lymph nodes, Payer's patches). In the Dark Zone of the GC, B Cells undergo rapid proliferation with by random point-mutations of the V genes from both IGH and IGL, called somatic hypermutation (SHM, (1, 17)). The high mutation rate of 10^{-3} mutations per base pair (18, 19) is mediated by the activation-induced cytidine deaminase (AID) and occurs at significantly higher levels within the CDR, than in the FR of both the IGH and IGL V segments (20). In the Light Zone of the GC mutated B Cells compete for improved

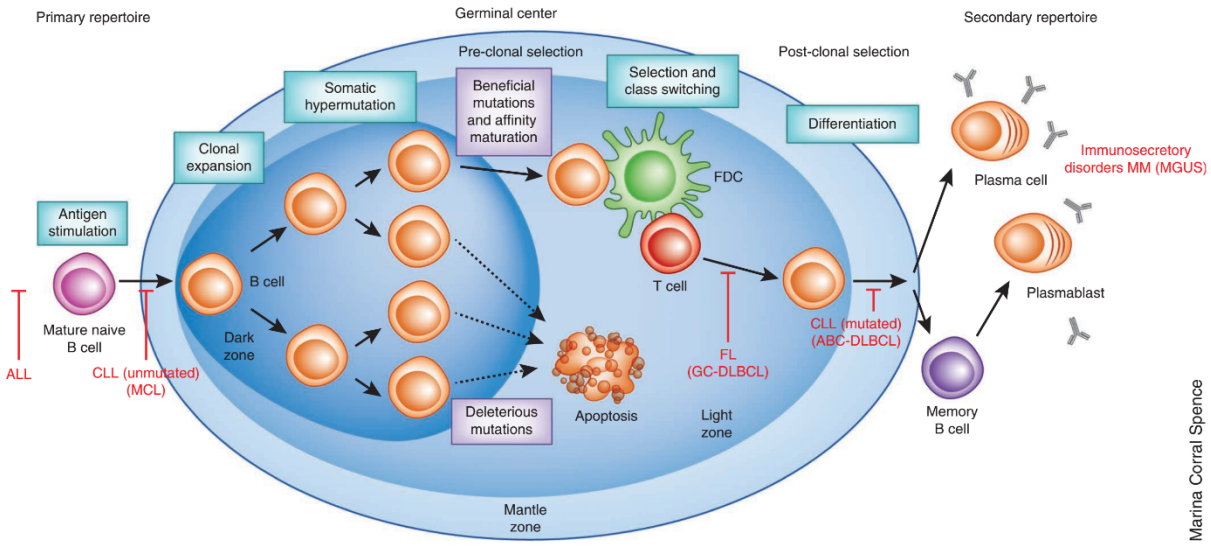


Figure 2 Diversification process of the secondary IG repertoire. Naïve IgM and IgD expressing B cells were generated in the bone marrow and form the primary IG repertoire. Steps in the formation of secondary IG repertoire diversity are pointed out in boxes. Upon cognate antigen activation, in the presence of T cell mediated co-stimulation, they form a germinal center (GC) reaction. The GC is divided into a dark and a light zone. In the dark zone, B cells rapidly proliferate rapidly resulting in clonal expansion with contemplate somatic hypermutation (SHM) initiated by activation-induced cytidine deaminase (AID). B cells with high affinity IG receptors can enter the light zone of the GC where they undergo class switch recombination (CSR) to IgG, IgA or IgE IG isotypes. They further differentiate into memory B cells, plasma cells and plasma blasts. Steps that can lead to abnormal B cell proliferation and malignancies are indicated in red. ALL, acute lymphoblastic leukemia; CLL, chronic lymphocytic leukemia; MCL, mantle cell lymphoma, GD-DLBCL, germinal center diffuse large B cell lymphoma; FL, follicular lymphoma; ABC-DLBCL, activated B cell-like DLBCL; MGUS, monoclonal gammopathy of undetermined significance; MM, multiple myeloma. Malignancies not investigated by HTS at the time of the review (2014) are shown in parentheses. Reprinted by permission from Macmillan Publishers Ltd: Nature Biotechnology (56), copyright 2014.

antigen-binding capacities, mediated by follicular dendritic cells (FDC) and T Cells (21–23). In addition, AID-induced class-switch recombination occurs to change the low affinity IgM receptor to the more specialized IgG/A/E isotypes, with subsequent selection against self-reactivity. After these GC reactions, consisting of several iterative rounds of proliferation, mutation and positive-negative selection, memory B Cells (memBC) and terminally differentiated Plasma Cells (PC) emerge. The former mediate rapid recall upon encountering the same antigen, whereas the latter transit to the bone marrow (11). Within the bone marrow PC can secrete antibodies into the periphery at unprecedented rate of up to 20,000 IG molecules per second until clearance of the antigen is achieved (24–26). Most PC stay in the bone marrow for weeks and up to 6 months (27). Some remain as long-lived PC with a half-life times of, for example 11 years in the case of tetanus toxoid but also much longer, for example against measles virus (MV), over 3000 years of half-life time has been reported (28). B Cell memory recalls or secondary adaptive immune responses bypass the multi-stage activation necessary to stimulate naïve B cells. They lead to the direct generation of antibody producing plasma blasts, but also enable the initiation of additional affinity maturation through new GC formations (29, 30). All B cells originating from the same GC evolved from the same progenitor and are thus clonally related, characterized by similar SHM-profiles (16). In healthy individuals, typically 80% of the peripheral blood B cells (PBMC) are naïve and clonally unrelated. The other 20% are memBCs of past immunological encounters (31). Taken together, the enormous diversity of the secondary IG repertoire stems from somatic hypermutation and class-switch recombination resulting in high specificity and avidity of the IG molecules.

1.2. Studying the IG repertoire in Health and Disease

1.2.1. Low-throughput Sequencing of the IG repertoire

With the arrival of Sanger sequencing it became possible to assess IG molecules on the transcriptional level, albeit at low throughput (32). Like this, using limiting dilution, antibodies and IG receptors of determined specificity could be functionally described, cloned and expressed in immortalized B cells, yielding pathogen-neutralizing antibodies (33–38). Following this, numerous IG sequences encoding for antibodies against clinically important pathogens, such as SARS coronavirus (39), Influenza (40), HIV (41–43) and dengue (44), were determined and B cell related autoimmune reactions were described on a functional level (45–47). **Table 3** provides a brief overview over a selection of low-throughput sequencing vaccination studies.

Table 3 Low-throughput sequencing IG repertoire studies, adapted from (48)

Vaccine	Cells used	Methodology	Key findings	Ref.
TIV (Influenza)	IgG PB 7 d.p.v.	Single cell IGH and IGL PCR with Sanger sequencing	Study of 50 mAbs produced from 14 individuals against three different influenza strains, showing that influenza-specific antibody response is pauci-clonal, with extensive SHM-derived intraclonal diversification of the influenza-specific lineages	(40)
TT	PB 6 d.p.v.	Single cell IGH and IGL PCR, cloned into E. coli and Sanger sequencing of TT+ clones	The level of SHM were similar between individuals, and did not increase through the study, suggesting the limit had already been reached through previous routine vaccinations.	(49)
TT	TT-specific PB 7 d.p.v.	Single cell isotype specific IGH and IGL PCR with Sanger sequencing	CDR3 length, IGH VDJ gene usage, and distribution of SHMs were similar among TT-specific PB and memBC cells.	(50)
PS (23 valent)	IgG PB 7 d.p.v.		137 mAbs against 19 of the 23 vaccine serotypes from four individuals were cloned, and it was found that most antibodies were serotype-specific, but 12% cross reacted with two or more serotypes	(51)
PS (23 valent)	PPS4 or PPS14 specific B cells 6 w.p.v.	Single cell culture with IGH PCR and Sanger sequencing of pooled cells	More than 1300 sequences from 40 individuals. significant differences in antibody repertoires between young and elderly individuals. the latter had higher clonality with lower levels of SHM	(52)
PS or PS-DT or OC-CRM	Lymphocytes, 7 d.p.v.	Fusion of lymphocytes to mouse myeloma cells with IGH and IGL PCR and Sanger sequencing	15 cell lines that secreted antibody against Hib PS were sequenced from 10 individuals. These mAbs had undergone SHM and demonstrated increased B-cell clonality after vaccination and bias towards use of the IGHV3 gene family.	(53)
PS-DT			4 cell lines that secreted antibody against Hib PS from four individuals were sequenced, where all used IGHV3 genes, but 2 unique IGHD-IGHJ gene segments, indicating that the four cell lines were from two different lineages.	(54)

Abbreviations: DT = diphtheria toxoid; Hib = *Haemophilus influenza* type b; OC = oligosaccharide; PPV23 = 23-valent pneumococcal polysaccharide vaccine; PS = polysaccharide; TIV = trivalent inactivated influenza vaccine; PB = Plasma blast; mAb = monoclonal antibody; d/w. p v. = days/weeks post vaccination.

1.2.2. High-throughput sequencing the IG repertoire

With the invention of high-throughput sequencing (HTS) technologies it became possible to probe an individual's IG repertoire at unprecedented depth (55). Typically, IG repertoire sequencing of human samples uses either B cell DNA or (m)RNA of IGH rearrangements isolated from donor PBMCs (56). Sequencing B cell DNA favors determination of diversity as the number of sequences will be proportional to the number of DNA molecules (57). Using (m)RNA templates enables the estimation of relative expression levels of IG molecules and requires initial reverse transcription into cDNA (56). As plasma cells produce >100 times more IG RNA transcripts than naïve B Cells or memBC, this cell type will be overrepresented by such an approach (56, 58). In a typical human sequencing library set-up, the templates are subsequently amplified in a multiplex PCR approach using either IGHJ gene or C-Region specific forward primers combined with the BIOMED-2 primer set of FR1, FR2 or FR3 targeting reverse primers (**Figure 3**) (59). These primers, designed by van Dongen and colleagues 2003 (59), were validated by several independent researchers (60–62), and became the gold standard for human IG repertoire HTS experiments. The FR1 BIOMED-2 primer set generates the longest amplicons and is the most widely used set, due to the superior information gained about the IGHV region (63, 64). In general, the utilization of several primers in a competing PCR set-up can lead to an uneven target amplification

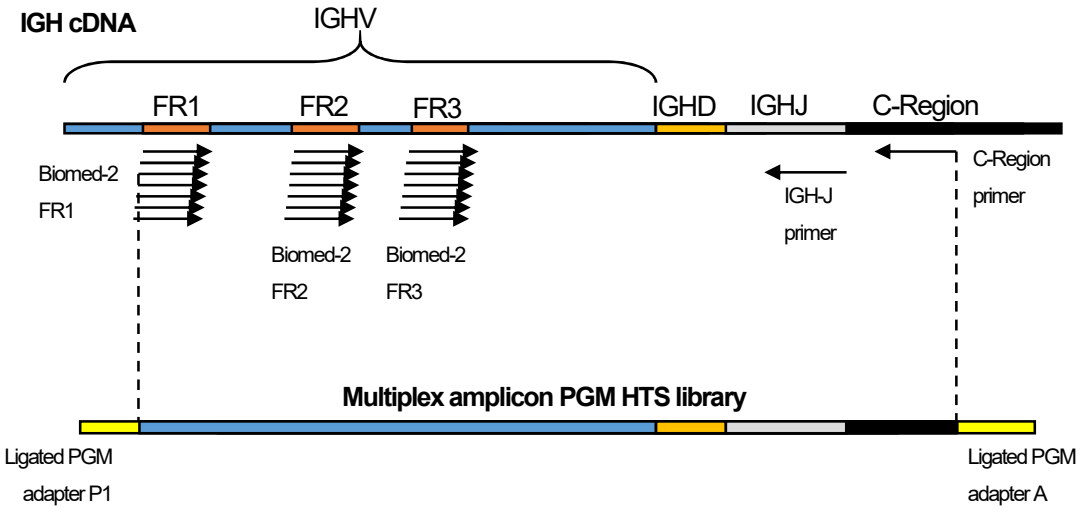


Figure 3 Schematic overview of the PGM multiplex primer library preparation approach. IGH cDNA is amplified with a Biomed-2 FR-specific primerset consisting of 7 different primers. IGHJ or C-region specific primers are used for the reverse strand. Templates are amplified with 25-35 cycles of PCR. PGM sequencing adapter are ligated to blunt-ended purified templates and the amplicon reaction mixture is again amplified in a 12 cycle PCR reaction using adapter P1 and A as Primer (not shown). Afterwards libraries are ready to be sequenced on an IonTorrent PGM.

(65). Such a bias may skew the IG repertoire readout and falsely label sequences as originating from highly expanded clones. To overcome this technical problem, approaches such as 5'RACE PCR or unique molecular identifier (UID) tagging of every single RNA molecule in a PCR mixture have been applied (48, 56, 57). 5'RACE uses IGHJ or C-Region specific primers in a cDNA reverse transcription set-up wit subsequent template switching (66–68). The template switching oligonucleotide is then targeted in a conventional PCR amplification avoiding the use of multiple primers. While this technique eliminates the PCR bias, it bears the pitfalls of low-efficiency and false-template switching (69, 70). In addition to that, the amplicons from 5'RACE PCR usually extent to 500-600 base pairs, which exceeds the sequencing length of most HTS systems, enforcing the use of Roche's 454 platform due to its superior read length (67). While being the instrument of choice for almost all early HTS studies on the IG repertoire, 454 sequencers were expensive and more error-prone compared to other available systems (**Table 4**). More recently, Vollmers and co-workers adapted an RNA barcoding approach to the Illumina sequencing platform (71). By adding 8 random nucleotides (unique identifiers – UIDs) to each primer prior to amplification and HTS, the group combined HTS reads into UID families that represented PCR copies of the same original RNA molecule (71). This allowed to completely reverse the amplification of RNA transcripts and provided the means for thorough error correction through building of consensus sequences from each UID family (71–74). Requiring at least 5 reads per UID (71–73), this method considerably lowers the throughput of HTS approaches and depends on complex bioinformatic pipelines backed by sophisticated computer hardware, to allow smooth processing of larger HTS datasets. However, UID-based sequencing has ever since become the gold standard for IG repertoire HTS approaches and was adapted to other sequencing platforms (see also **Figure 26, Chapter 4**) (68, 75, 76). Another crucial subject of HTS studies is the separate assessment of IGH and IGL transcripts (56, 57). The identification of a usable antibody sequence (i.e. it can be cloned, expressed, and tested for target specificity) requires the information of both IG chain sequences. While it is possible to sequence each separately at high-throughput, it remains elusive which transcripts belonged together (77). In addition, the relatively low-throughput of current cloning strategies renders subsequent trial-and-error approaches not very cost-effective. DeKosky and his group published their approach on sequencing artificially linked IGH:IGL transcripts, preserving the original pairing (78). In addition, recent advances in single-cell cloning and sequencing achieve depths comparable to early HTS approaches, also keeping the original IGH:IGL combinations (79).

Table 4 Error-rates per 100 bp of the different sequencing platforms, adapted from (83)

Platform	Substitution	SD	Indels	SD	Total	SD
454 GS FLX	0.09000	N/A*	0.90000	N/A*	0.99000	N/A*
454 GS Junior	0.05430	N/A*	0.39055	N/A*	0.45540	N/A*
Illumina HiSeq	0.26400	0.11238	0.02561	0.02351	0.28467	0.11875
Illumina MiSeq	0.24551	0.11079	0.00905	0.01436	0.29652	0.18867
Ion Torrent PGM	0.16985	0.44761	1.45793	1.21924	1.63112	1.24217

Data from (75, 80–82), * only one sample available, Indels = Insertions and Deletions

1.2.3. B Cell research: aims and scopes

HTS approaches on IG repertoires have been applied for antibody discovery in the context of vaccination and infectious diseases, to investigate IG repertoire dynamics and development, to understand immune dysregulation and B cell malignancies.

1.2.3.1. Antibody discovery

The core of employing HTS to study post-vaccine and post-infection IG repertoires is the identification of sequences that originate from antigen-induced, antigen-driven and antigen-specific IG molecules. For experiments based on human peripheral blood samples, it is crucial to first separate sequences of the antigen-driven from naïve B cells. This can be achieved *ipso facto* by determining their mutation status in relation to the prospected rearranged germline genes, as antigen-driven B Cells must have passed GC reactions thus accumulated IGHV region mutations (84). However, with the constant challenges of the adaptive immune system, these sequences might not have been solely raised in response to the applied vaccine or infection of interest. Another important aspect is the critical time window in which such antigen-driven B cells can be found in the periphery (55). While most studies examined blood B cells harvested 7 days after infection or vaccination, the memory-recall Plasma blasts can already be found as early as 3 days after challenge (28). To distinguish the sequences of antigen-specific B cells from the pool of mutated, post-GC B cell sequences, several approaches have been developed. One method is to search for sequences of known specificity within the assessed IG repertoire (85–89). Another approach is to identify shared and public IG molecules in response to the same antigen, also referred to as IG convergence (**Figure 4**). With an estimated 10^{18} possible unique receptors (55), the theoretical repertoire surpasses by far the capacities of human B cell populations (10^{10}-10^{11}). Therefore, it is stochastically unlikely, that certain individuals would share the same IG receptors (84). However,

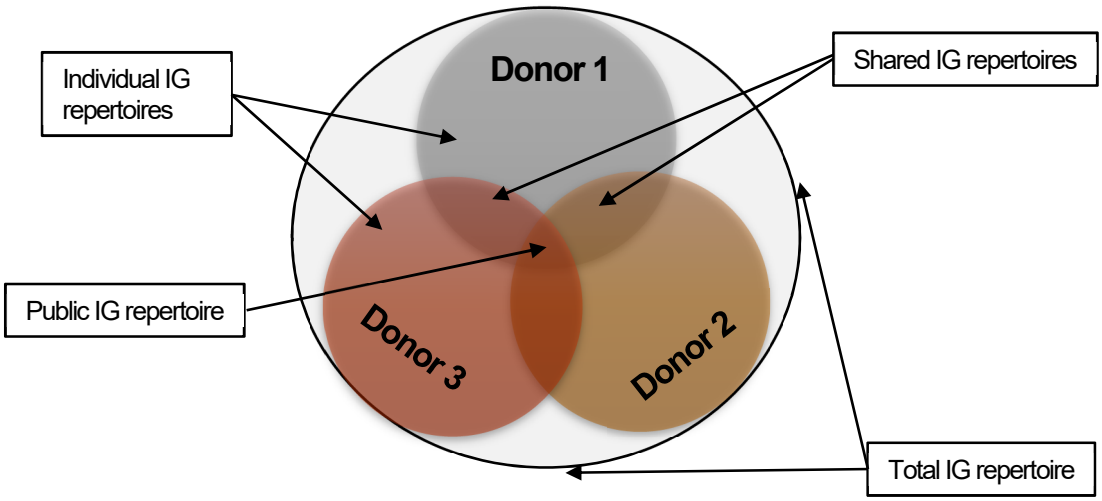


Figure 4 Schematic representation of IG repertoire overlap. Individual repertoires are presented by circles per donor. Shared repertoires are shown as IG sequences shared between 2 donors. Public repertoires are described as sequences shared between 3 and more donors. All donors form a theoretical total IG repertoire.

antigens impose structural restrictions for IG interaction, which results in a shared structural homology in the elicited IG molecules. This homology can, to some extent, be mapped to the underlying amino acid. Additionally, the high variability and deterministic binding function of the IGH CDR3 amino acid sequence has been proven sufficient to investigate IG convergence (3, 90–92). Following this concept, several groups were able to identify such shared and public IG sequences (**Figure 4**) in response to acute dengue infections (90), *Haemophilus influenzae* type B (Hib), tetanus toxoid polysaccharides and group C meningococcal (MenC) polysaccharides, using a trivalent vaccine (92). In the future, databases could contain all IGH CDR3 sequences with known specificity to mine IG repertoire datasets for past antigen exposures ultimately reconstructing the immunological history of an individual (93–99).

1.2.3.2. IG repertoire dynamics and development

The potential number of different IG gene segment recombinations in humans is about 2.5×10^{6}. Junctional diversity through nucleotide insertions and deletions increases this number to approximately 10^{18} different IG molecules, which is referred to as the theoretical naïve or total IG repertoire (**Figure 4**). Using HTS, Arnaout and Glanville estimated the depth of the human peripheral blood B cell repertoire to be 3×10^{9}-3×10^{10} unique IG receptors per individual (100, 101). However, little is known at this resolution about B cell turnover dynamics or the contributions of different tissues, which are difficult to access in humans (102).

1.2.3.3. Immune repertoires with age

The adaptive immune system is impaired in human infants and elderly individuals alike, making these individuals more susceptible to certain infectious diseases such as influenza (103). IG repertoire development has been found to be deterministic and programmed in murine and human *feti* (104–106). The IGHV gene usage of human adults differs from that of fetal IG repertoires. The latter preferentially express IGHV1, 3 and 4 gene families in response to respiratory syncytial virus (RSV) infections, instead of the immunodominant IGHV genes found in response by adult individuals (IGHV3–23, IGHV3–30, IGHV3–33 and IGHV4–04) (107). In addition, young individuals exhibit less somatic hypermutation in these responses, producing less optimized or weaker immune responses than adults. A major underlying factor of these observations could be the suppressed expression of the terminal deoxyribonucleotidyl transferase enzymes responsible for the nontemplated random nucleotide insertions and deletions during IGHD and IGHJ recombination (108). Alterations of IG repertoire diversity and structure were associated with age in multiple studies (109–112), showing increased clonality and delays in the immune responses of the elderly.

1.2.3.4. Chronic lymphocytic leukemia

With an incidence rate of 4.92 per 100,000 individuals per year in Europe, chronic lymphocytic leukemia (CLL) is the most common form of leukemia (113). Incidence rates are higher for men than for women, and two thirds of CLL patients are over 60 years old (114). There is evidence, for a population specific difference in CLL to leukemia rates, with 35-40% of all leukemia being CLL in Denmark but only 3-5% in Japan and China (115). The clinical outcome of CLL varies strongly between patients. While many patients remain without specific symptoms and don't require treatment, others become increasingly susceptible to infections (116). Early diagnosis and characterization of the malignant B cell clones are essential for personalized treatment, and thus of high clinical importance.

CLL is determined by high (> 5×10^9 cells/L) peripheral blood clonal B cell count over at least 3 months. With these B cells expressing κ:λ IGL at a ratio over 1:3 or below 1:1.03, and showing a small cell phenotype with dense nuclei and partially aggregated chromatin (117–121). The malignant lymphocytes express CD19, CD5 and CD23, with no or weak expression of surface IG (122). An accumulation of $CD5^+$ B cells in secondary lymphoid organs is typical for CLL, but unlike other malignancies, accumulation originates from resistance to apoptosis rather than increased proliferation rates (123). The anti-apoptotic profile is characterized by increased expression of the Bcl-2 survival protein and other

microenvironmental factors, causing rapid apoptosis of CLL lymphocytes *in vitro* (124, 125). About 25% of CLL patients are asymptomatic, complicating diagnosis or detection of disease onset (124). Most CLL cases are diagnosed during routine examinations of elevated lymphocyte counts at physicians (126). CLL symptoms include persistent lymphocytosis, splenomegaly, and mild cervical, supraclavicural and/or axillary nodes lymphadenopathy (114, 127, 128). About 50% of CLL patients experience mild anemia and thrombocytopenia is observed in approximately 25% of the patients (129, 130). Skin symptoms include exfoliative dermatitis, erythroderma and secondary skin infections, which are observed in approximately 5% of the patients (131, 132). The function of the immune system in CLL patients is usually impaired, manifesting itself in immunodeficiency despite increased B cell count (130). Autoimmunity against red blood cells, leading to anemia, platelets and neutrophils is observed in about 25% of the patients during the disease (133, 134). Mortality and morbidity of patients is highly affected by frequently occurring bacterial infections of the skin, respiratory, as well as urinary tract. Additionally, patients show poor responses to vaccination (130, 135). The vaccine response can be correlated with clinical outcome and survival (130).

Two different staging systems have been developed to allow clinical prognosis, relying either on a combination of lymphadenopathy, organomegaly and cytopenias (5 Rai stages, (136)), or on the number of affected lymphal areas and cytopenias (3 Binet stages, (137)). While both provide prognosis and treatment timing suggestions, the heterogeneity of the different categories and outcomes warrants the identification of more accurate clinical markers. One of the most reliable prognosis markers is the IGHV mutation status. Post-antigenic stimulation (**Figure 2**) and thus mutated IGHV regions (>2% divergence from germline) in the malignant clone (M-CLL) resulting in improved survival rates (138, 139). This can be partially explained by the increased occurrence of polyreactive autoantibodies in unmutated CLL (U-CLL) clones, of which the patients usually require early treatment (140–143). The mutation status of CLL clones mostly assessed by Sanger sequencing of PBMC cDNA or BM aspirates with a commercially available assay. Extensive work has been performed to characterize the B cell receptors in CLL patients (94, 144–146). Malignant clones are found to be highly biased in IGHV gene usage compared to IG repertoires from healthy individuals (96, 141, 146–151). According to these studies, similar IG transcripts are produced by different CLL patients, using dominantly IGHV 1-69 paired with IGHJ 6 genes and IGHV 4-34 in U-CLL and M-CLL respectively. Furthermore, they report stereotypical CDR3s, with similar length and amino acid composition. Taken together this provides evidence for an involvement of B cell

receptor reactivity in the onset of CLL leukemogenesis, possibly through a common antigen or auto-antigen (151–153).

1.2.4. OmniRat™ – transgenic rats with human IG repertoire

The underlying modulating mechanisms and dynamics of the IG repertoire are exceptionally conserved among jawed vertebrates (154, 155). While these mechanisms bear the possibility to generate >10^{18} different IG molecules (55), the number of circulating B cells limits this theoretical diversity to a maximum of 10^{10}-10^{11} in humans (101), 10^{9} in mice (2) and 300,000 IG receptors in Zebrafish (156). In addition, clonally related B cells targeting the same antigen further limit the expressed diversity of the IG repertoire. Compared to early studies in the IG repertoire or other fields in immunology, the utilization of animal models has been largely ignored. While some of the first studies on HTS IG repertoires were conducted using zebrafish and mice (77, 156–158), during the following years, studies focused mainly on human B cells. Thus, due to the technical limitations inherent to human studies, our knowledge on deep-sequenced repertoires is built almost solely on B cells circulating in the peripheral blood. However, the vaccine- and disease-elicited antibody responses are generated in inaccessible GC reactions, and thereafter transit into the bone marrow within a narrow 1d time-window (28). Consequently, human IG repertoires are blurred by a large accumulation of naïve B cell 'noise' making it difficult to, for instance, identify reoccurring shared or public clones. Unfortunately, no primer set with quality and verification equal to BIOMED-2 was ever released to amplify any non-human IG transcripts. In addition, existing databases on non-human IG repertoire germlines proved to be incomplete or inaccurate with uncontrolled release of partially unproductive or even false IG transcripts (Communication of the Adaptive Immune Receptor Repertoire community, AIRR, unpublished). The availability of secondary lymphoid organs, combined with lower repertoire diversity, highly warrants rodent animal models to decipher the immune response to vaccination and disease at unprecedented depth.

Several transgenic rodents have been developed to bridge the gap between human limitations and animal models (159–161). Yet, their performance proved to be suboptimal, suffering from imperfect interaction of human constant region with the endogenous rodent cellular signaling machinery (162). In the presented work, we utilized the OmniRat™, a transgenic rat expressing human IG genes (163, 164). These rats utilize a chimeric human/rat IGH locus with 22 human IGHV, 27 IGHD and 6 IGHJ germline genes, which are linked to the rat IGH C-region segments (**Figure 5**). For the light chains, 12 IGLV and 5 IGLJ human gene segments were introduced into the κ locus, whereas 16 IGLV and 5 IGLJ human

gene segments were introduced in the λ locus. The endogenous rat IG loci were silenced by zinc finger nucleases (165, 166). OmniRat™ express a fully diversified IG repertoire with human fab fragments and rat constant regions, and have been successfully used for the production of high affinity chimeric monoclonal antibodies (164).

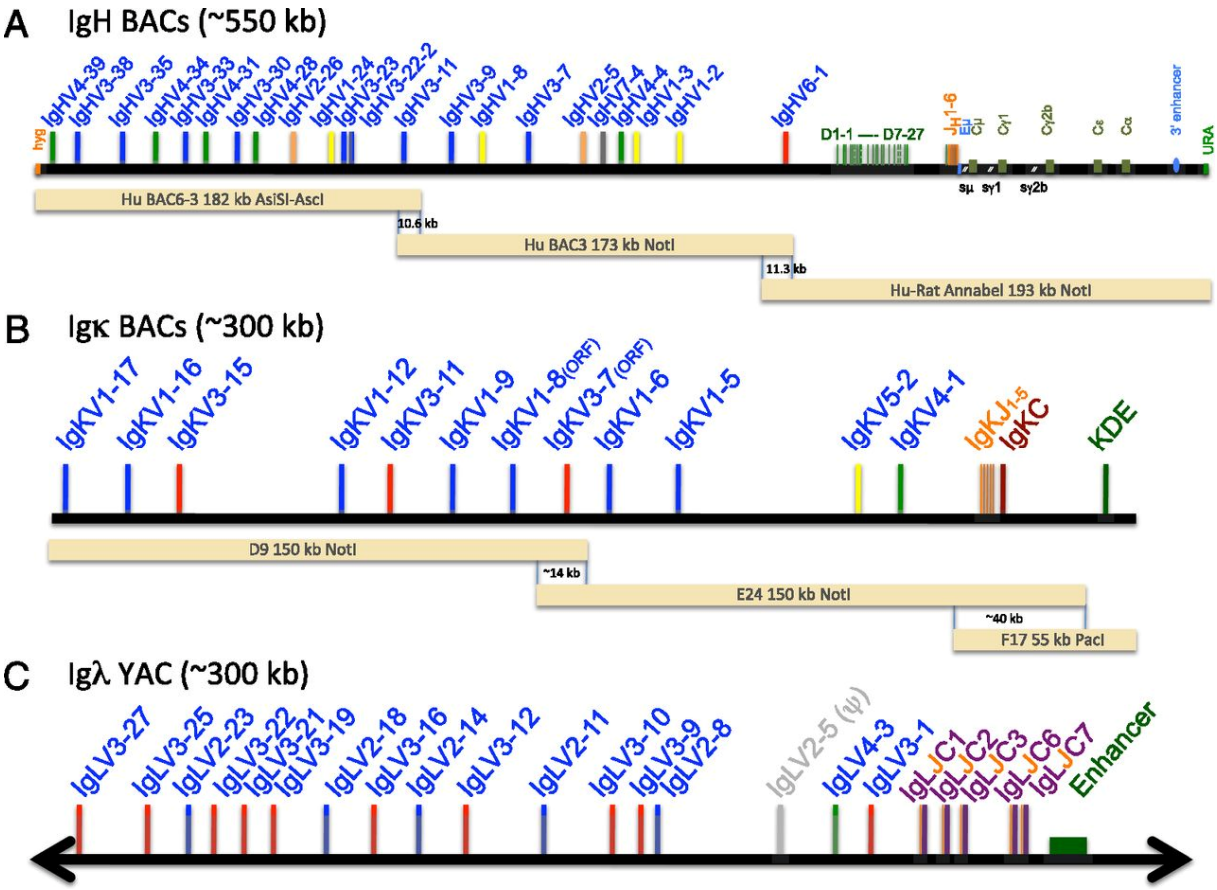


Figure 5 Schematic overview of the integrated human IG loci in OmniRat™. (A) The chimeric human/rat IGH region contains three overlapping bacterial artificial chromosomes (BAC) with 22 different and potentially functional human IGHV segments. BAC6-3 has been extended with IGHV 3-11 to provide a 10.6-kb overlap to BAC3, which overlaps 11.3 kb via IGHV 6-1 with the C-region human/rat. The latter is chimeric and contains all human IGHD and IGHJ segments followed by the rat C-region (Cμ, Cγ1, Cγ2b, Cε, Cα) with full enhancer sequences. (B) The human IGL κ BACs with 12 IGLV (κ) and all IGLJ (κ) provide an ~14-kb overlap in the IGL κ region and ~40 kb in C-region (κ) to include the kappa-deleting element (KDE). (C) The human IGL λ region with 17 IGLV (λ) and all IGLJ and C-regions (λ), including the 3′ enhancer, is from a yeast artificial chromosome (167). Figure used from (164) with permission.

Thesis aims and hypotheses

Dissection of the IG repertoire with HTS in health and disease promises to transform our understanding of the adaptive immune system dynamics. The applications range from identification of novel (therapeutic) antibodies to the deconvolution of malignant B cell development. In this context, the development of robust HTS and data processing methods also for animal models is crucial. The purpose of this thesis is to establish IG repertoire sequencing of mouse models and OmniRat^{TM} on the Ion Torrent PGM HTS platform. The specific aims were as follows:

1. Development of a bioinformatics framework to identify antigen-driven IGH sequences from bulk sequenced bone marrow B cell RNA transcripts.
2. Characterization of the CLL-like CD5^{+} B cell expansion observed in A20^{BKO}sCYLD^{BOE} mouse model using a custom-made bioinformatics network property analysis approach.
3. Improve PGM IGH sequencing quality and throughput by developing a laboratory and bioinformatics sample processing protocol using a UID barcoding approach tailored to the PGM platform.

Chapter 2

Functionally convergent B cell receptor sequences in transgenic rats expressing a human B cell repertoire in response to tetanus toxoid and measles antigens

Jean-Philippe **BÜRCKERT**†, Axel R.S.X. **DUBOIS**†, William J. **FAISON**, Sophie **FARINELLE**, Regina **SINNER**, Anke **WIENECKE-BALDACCHINO**, Emily **CHARPENTIER**, and Claude P. **MULLER**

† Both authors contributed equally to the work

Department of Infection and Immunity, Luxembourg Institute of Health / Laboratoire National de la Santé

[This chapter is based on a manuscript accepted Dec 05, 2017 in Frontiers in Immunology, DOI: 10.3389/fimmu.2017.01834, a previous version of it was part of the doctoral thesis of A.R.S. X. Dubois]

Authors' contributions: J-P.B. and A.R.S.X.D contributed equally to the work. J-P.B. designed and developed the bioinformatics approach, interpreted data, performed data processing and wrote the manuscript. A.R.X.S.D designed and carried out wetlab research, prepared samples, interpreted data and wrote the manuscript. W.J.F. supported bioinformatics approaches and data processing, corrected the manuscript. A.W-B. set up the raw data processing script and performed raw data processing. S.F. and E.C. provided technical assistance with immunizations, ELISA and virus culture. R.S. performed PGM sequencing. C.P.M. designed research, interpreted data, corrected the manuscript and supervised work.

2.1. Summary

The identification and tracking of antigen-specific immunoglobulin (IG) sequences within total IG repertoires is central to high-throughput sequencing (HTS) studies of infections or vaccinations. In this context, public IG sequences shared by different individuals exposed to the same antigen could be valuable markers for tracing back infections, measuring vaccine immunogenicity, and perhaps ultimately allow the reconstruction of the immunological history of an individual. Here, we immunized groups of transgenic rats expressing human IG against diverse sets of bacterial, viral and chemically defined antigens. We showed that these antigens impose a selective pressure causing the IG heavy chain (IGH) repertoires of the rats to converge towards the expression of antibodies with highly similar IGH CDR3 amino acid sequences. We implemented a computational approach, similar to differential gene expression analysis, that selects for clusters of CDR3s with 80% similarity that are significantly overrepresented within the different groups of immunized rats. These clusters represent complex antigen-specific signatures exhibiting stereotypic amino acid patterns that include previously described IG sequences specific for tetanus toxoid and measles virus proteins. Our results highlight the potential use of the transgenic IG rats as a model to readily identify convergent signatures to large numbers of antigens that could potentially be used to draw an antigenic map of past immune exposures for humans as well.

2.2. Introduction

Immunoglobulin (IG) molecules are the primary effectors of the humoral immune response. In theory, IG can bind to every possible antigen through the large variety of immunoglobulin V (variable), D (diversity) and J (joining) gene rearrangements in the bone marrow and target-oriented affinity maturation in germinal centers (1). All B cells of a germinal center are clonally related to a common ancestor and target the same antigen with varying affinities, iteratively selecting for improved affinity and avidity (15). The IG molecules of the emerging B cells bind to the target epitope in a lock-and-key principle which is mediated mainly by the heavy chain complementary-determining region 3 (CDR3) loop on top of the IG (1, 3). The CDR3 is the most variable part of the IG sequence and the main antigen-binding determinant. The repertoire of CDR3s sufficiently describes the entire functional immunoglobulin heavy chain (IGH) repertoire of an individual (3, 4).

High-throughput sequencing (HTS) has been widely applied to study the IGH repertoire in response to vaccination and infection (90). With this technique, it has become possible to investigate the evolutionary affinity maturation processes after antigenic challenge and to compare their outcome across individuals (168). The IGH repertoire is essentially private (84), but it appears that individuals also produce a public response to a common antigenic stimulus characterized by a certain degree of similarity at the CDR3 sequence level (44, 169–171). Public CDR3s were notably identified in human in response to dengue infection, H1N1 seasonal influenza vaccination and repetitive polysaccharide antigens (90, 92, 171). Such CDR3s provided signatures of past immunological exposures allowing for sequence-based monitoring of vaccination or infectious diseases, and perhaps ultimately to reconstruct an individual's antigenic history. Studies investigating this concept of public IG CDR3s mainly used human blood-derived PBMCs. These represent only a miniscule part of the complete IG repertoire (172) and it is critical to capture the affinity-matured B cells during their brief transit from the germinal centers through peripheral blood to the bone marrow. The large heterogeneity of human B cell repertoires composed through past exposure to a plethora of antigens further complicates the identification of antigen-induced IG sequences in the context of single antigen challenge or vaccination (173). The usage of an animal model provides ready access to secondary lymphoid organs after restricted antigen exposure, enabling a focused investigation of antigen experienced plasma cells (28, 174, 175).

Here we applied HTS on class switched, bone marrow B cells, rich in serum antibody producing plasma cells, from rats carrying human germline IGH and light chain (IGL) loci, the OmniRat™ (164–166, 176). These transgenic rats were immunized with viral (Modified Vaccinia virus Ankara, MVA), protein (measles virus hemagglutinin and fusion proteins, HF and tetanus toxoid, TT) and chemically defined hapten-conjugate antigens (Benzo[a]Pyrene-TT, BaP-TT) to study the evolution of convergent CDR3 amino acid sequences. We showed that OmniRat™ mount convergent IG responses characterized by CDR3s with high amino acid sequence similarity. The level of similarity was consistent for all investigated antigens. We applied an approach similar to differential expression analysis to identify overrepresented clusters of highly similar, antigen-driven CDR3s. These could be grouped into antigen-associated signatures matching previously described measles virus-specific OmniRat™ hybridomas (177) and human TT-specific antibodies (92, 178–180). Our results suggested that humanized IG transgenic rats can be used as a model to study human-like IG repertoire dynamics and to determine antigen-associated CDR3 signatures to characterize the history of antigen exposure in human individuals.

2.3. Materials and Methods

2.3.1. Animals and immunizations

Humanized IG transgenic rats (OmniRat™, Open Monoclonal Technology Inc., Palo Alto, USA) were developed and bred as previously described (164–166, 176). OmniRat™ carry a chimeric human/rat IGH locus, where 22 human IGHV genes and all human IGHD and IGHJ genes are linked to the rat C region genes in germline configuration as well as fully human, IGL λ and κ loci (164) (**Figure 5**). Animals were separated into 6 groups of 4 to 6 individuals (**Table 5**). They received 3 intraperitoneal injections at 2-weeks intervals and were sacrificed 7 days after the last injection. Injections either contained 100 µg of tetanus toxoid (TT group, n = 4; Serum Institute of India, Pune, IN) or of a benzo[a]pyrene-TT conjugate construct (BaP-TT group, n = 5) (181), both formulated with 330 µg of aluminum hydroxide. Other rats were injected with 10^7 p.f.u. of a recombinant Modified Vaccinia virus Ankara (MVA) expressing the hemagglutinin (H) and fusion (F) glycoproteins of the measles virus (MVA-HF group, n = 6) or the MVA viral vector only (MVA group, n = 6) without adjuvant. The control animals received either 330 µg of aluminum hydroxide alone (ALUM group, n = 6) or were left untouched (NEG group, n

= 5). Antigen-specific IgG responses were monitored by ELISA 10 days after immunizations and at sacrifice. All animal procedures were in compliance with the rules described in the Guide for the Care and Use of Laboratory Animals (182) and accepted by the 'Comité National d'Éthique de Recherche' (CNER, Luxembourg)

Table 5 Study design: Antigen- and vaccination groups

Antigen	TT group		MVA group		Controls	
	TT (n=4)	BaP-TT (n=5)	MVA (n=6)	MVA-HF (n=6)	ALUM (n=6)	NEG (n=5)
BaP	-	x	-	-	-	-
TT	x	x*	-	-	-	-
MV H + F	-	-	-	x	-	-
MVA	-	-	x	x	-	-
Alum	x	x	-	-	x	-

* indicates that TT was chemically modified in the hapten coupling process for the BaP-TT group and does therefore not possess the same antigenic surface as the native TT used in the TT group.

2.3.2. Antigens for immunization and ELISA

BaP was coupled to ovalbumin (OVA, Sigma-Aldrich) for ELISA and to purified tetanus toxoid as previously described (181). The recombinant Modified Vaccinia virus Ankara (MVA) and the recombinant MVA carrying measles virus H and F proteins of the Edmonston strain (MV vaccine strain, clade A) viruses were propagated on BHK-21 cells (ATTC™ CCL-10™) as previously described (183–185). Antigen-specific IgG antibody levels in sera were determined in 384-well microtiter plates (Greiner bio-one, Wemmel, BE), coated overnight at 4°C with either 250 ng of MV antigen (Measles grade 2 antigens, Microbix Biosystems, Mississauga, USA), 2.5×10^5 PFU of sonicated MVA (~314 ng), 187.5 ng of TT or 0.25 µM of BaP-OVA in carbonate buffer (100 mM, pH9.6). Free binding sites were saturated with 1% bovine serum albumin (BSA) in Tris-buffered saline at room temperature for 2h. Serial dilutions of the sera were added for 90 min at 37°C, and developed with alkaline phosphatase-conjugated goat anti-rat IgG (1/750 dilution, ImTec Diagnostics, Antwerp, BE) and the appropriate substrate. Absorbance was measured at 405 nm. Endpoint titers (EPT) were determined as the serum dilutions corresponding to 5 times the background (**Figure 6**).

2.3.3. Sample preparation, amplification and Ion Torrent PGM Sequencing

Lymphocytes were isolated from bone marrow samples by density-gradient centrifugation (ficoll® Paque Plus, Sigma-Aldrich). Total RNA was extracted with an RNeasy midi kit following the manufacturer's protocol (Qiagen). cDNA was prepared using dT_{18} primers and Superscript III reverse transcriptase

(Thermo Fisher Scientific) at 50°C for 80 min. Recombined IGH fragments were subsequently amplified by PCR using primers for human IGHV regions and rat Cγ region with Q5 Hot Start High Fidelity polymerase (NEB, Ipswich, USA) as described previously (177). Amplicons were size selected on a 2% agarose gel and quantified. Quality was checked with a Bioanalyzer (High Sensitivity DNA, Agilent Technologies, Diegem, BE). Four randomly-selected libraries were pooled in equimolar concentrations and sequenced on a 318™ Chip v2 (Thermo Fisher Scientific) using multiple identifiers (MIDs) with the Ion OneTouch™ Template OT2 400 Kit and the Ion PGM Sequencing 400 Kit (Thermo Fisher Scientific) on the Ion Torrent Ion Personal Genome Machine (PGM™) System (Thermo Fischer Scientific).

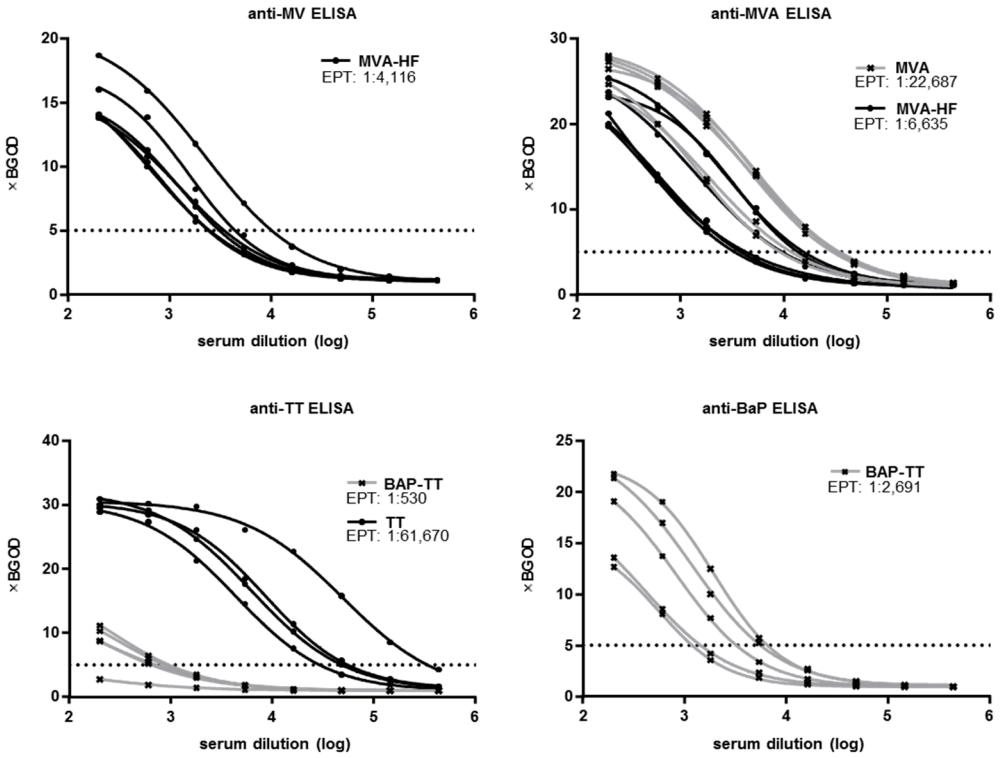


Figure 6 OmniRat™ ELISAs. Antigen-specific serum IgG in immunized rats were measured by indirect ELISA. The serum dilution was plotted (log scale) against binding, measured by absorbance expressed as multiples of the average optical density of the background (BG OD). Average endpoint titers (EPT) were determined as the serum dilution corresponding to five times the BG OD (dotted line).

2.3.4. Quality control and sequence annotation

BAM files were extracted from the Torrent Suite ™ software (version 4.0.2, standard settings) and demultiplexed by multiplex identifiers (MID). Only reads with an unambiguously assigned MID (0 mismatches), identified primers at both ends (2 mismatches allowed) and more than 85% of the bases with a quality score above 25 were considered for further analysis (**Table 6**). After clipping MIDs and primers, sequences were collapsed and submitted to the ImMunoGeneTics database (IMGT) HighV-

QUEST webserver (www.imgt.org, (186)) for IGHV gene annotation and CDR3 delineation (187). IGHV and IGHJ genes for the in-frame, productive sequences were subsequently assigned using a local installation of IgBlast (188), including only the genes present in the genome of the OmniRat™ as references. Only sequences with an unambiguously assigned IGHV and IGHJ gene were considered for further analysis. Human Ig sequences (IgG and IgM) were obtained from the Sequence Read Archive database (https://www.ncbi.nlm.nih.gov/sra) and processed as described (Accession number: SRP068407; (189)). Samples taken at day 7 post immunization were excluded to avoid skewing of data distributions by the applied vaccination. Samples were annotated using IMGT and post-processed using Change-O framework (v 0.2.4, (190)). Only functional sequences present at least three times per dataset considered for assessing CDR3 length distributions and somatic hypermutation level.

Table 6 HTS sequencing and data processing. The first column indicates the vaccination group including the number of animals. The second column is the study-number of the animal. The third column shows the raw reads per animal as identified per MID. The fourth column shows the quality filtered reads, collapsed to unique sequences. The fifth column shows the number of unique CDR3 amino acid sequences after IMGT processing. The last two rows show the total, mean and SD per column.

Vaccination	ID	Raw-reads w. MID	Filtered unique nt sequences	Unique CDR3s
MVA-HF n=6	1	1,365,752	75,632	4,007
	2	1,072,907	87,632	5,680
	3	1,573,425	97,070	5,050
	4	1,579,703	105,168	5,609
	5	1,110,541	53,913	3,536
	6	1,073,485	68,542	4,469
MVA n=6	7	929,621	75,643	4,333
	8	995,002	91,300	5,084
	9	937,392	65,840	3,288
	10	1,084,088	129,977	7,232
	11	998,174	105,000	4,980
	12	1,198,061	129,219	5,691
BaP-TT n=5	13	1,177,010	94,218	3,191
	14	990,929	91,500	5,838
	15	989,221	80,886	5,538
	16	1,126,831	89,527	4,399
	17	1,646,696	50,234	2,379
TT n=4	18	1,184,407	105,527	3,912
	19	1,119,688	65,630	4,228
	20	1,036,147	54,763	4,471
	21	1,380,809	137,371	6,756
Alum n=6	22	1,431,185	65,363	3,586
	23	1,879,372	85,192	7,202
	24	1,004,133	78,375	8,013
	25	896,017	68,058	6,602
	26	879,714	46,409	3,946
	27	850,298	86,623	4,365
NEG n=5	28	929,883	60,509	5,619
	29	1,121,113	100,287	5,885
	30	938,306	106,170	6,422
	31	1,380,088	135,281	6,809
	32	1,593,984	84,948	4,026
Total		37,473,982	2,771,807	162,146
Mean± SD		1,171,062 ±257,905	86,619 ±24,056	5,067 ±1,335

2.3.5. CDR3 similarity threshold for public immune responses

The number of matches for the 200 most frequent CDR3 (top 200) of a rat A in a rat B was obtained for a series of similarity thresholds and returned as ratio from 0 to 1 (i.e. all top 200 CDR3s of rat A have a match in rat B). Ratios were determined from 50% to 100% sequence similarity in one percent increments. The averages for the top 200 matching ratios at each increment were then calculated for all rats within a vaccination group and all rats vaccinated with unrelated antigens. Rats with related antigens were excluded in the pairwise comparison (e.g. MVA as intra-group for MVA-HF). The average top 200 matching ratios were plotted against sequence similarity along with the first derivatives in GraphPad Prism 5 (www.graphpad.com).

2.3.6. Identification of antigen-driven CDR3 clusters

Only CDR3s longer than 4 amino acids were considered for analysis. CDR3s with a minimum of 80% amino acid similarity, with one amino acid length difference allowed, were considered as relatives. Length differences were penalized the same as a substitution. For each CDR3, the cumulative count of all its 80% relatives per rat (CDR3-count) was calculated and stored in a fuzzy match count table. Data was imported and analyzed with DESeq2 according to the standard workflow for RNA-seq, treating CDR3-counts as expression values (191). Briefly, data was imported as a count-data matrix and converted into a DESeq2-object with conditions according to the antigens used for vaccination. Correct sample grouping was confirmed using variance stabilizing transformed count data (VST-counts). Euclidian distance computation was performed on VST-counts as described in the DESeq2 vignette (192). Principle component analysis plots were generated using the 'PlotPCA' function on VST-counts of the DESeq2 package. P-values were adjusted for multiple testing and to determine the false discovery rate (FDR) using Benjamini-Hochberg correction (193). Based on an FDR of 1%, over-represented CDR3 sequences were extracted if their adjusted p-values were lower than 0.01. Log2-fold change cut-offs were determined manually per antigen group. The extracted CDR3 sequences were grouped using single-seed iterative clustering based on maximum difference of 80% sequence similarity. All analytical scripts were written in Python 2.7 and R 3.2.3 (194).

2.3.7. 3D modeling

Selected IG nucleotide sequences were uploaded to IMGT for annotation. Sequences were elongated to full length by adding the missing nucleotides from the closest germline gene as predicted by the IMGT

algorithm. Full length sequences were submitted to the “Rosetta Online Server that Includes Everyone” (ROSIE, http://rosie.rosettacommons.org/, (195–197)) with enabled H3 loop modeling option. ROSIE-output PDB files of the grafted and relaxed models were visualized using PyMol (version 1.7.4, http://pymol.org, (198)).

2.4. Results

2.4.1. High-throughput sequencing of OmniRat™ IGH mRNA transcripts

To study convergent IGH repertoires in response to vaccination, 32 transgenic IG humanized rats (OmniRat™) were immunized with different antigens (**Table 5**; TT, BaP-TT, MVA, MVA-HF). Aluminum hydroxide (ALUM) was used as an adjuvant for TT and BaP-TT. Two control groups received either the adjuvant alone or were left untouched (NEG). All animals exhibited a specific antibody response against the immunizations and mock immunized (ALUM group) and non-immunized animals (NEG group) showed no detectable antigen-specific antibodies (**Figure 7**). MVA-HF and BaP-TT vaccinated animals exhibited a specific immune response against the MVA vector or the TT carrier protein respectively, albeit at lower levels than the animals immunized with these antigens only (**Figure 7**). Rearranged heavy chain IgG genes were amplified from mRNA extracted from bone marrow (BM) lymphocytes and sequenced on a high-throughput sequencing (HTS) Ion Torrent PGM platform. A total of 37,473,982 raw reads with MID were obtained (range: 850,298 – 1,879,372 per animal, **Table 6**). After quality control and annotation, on average 86,619 sequences per animal were retained for analysis. The rats expressed a diverse IGH repertoire, including varying frequencies of all human IGHV and IGHJ genes. All possible IGHV-IGHJ combinations were found in all vaccination groups with no obvious bias in IGHV, IGHJ genes or IGHV-IGHJ recombination usage. The CDR3 length distribution of rats was comparable to that of observed in human IgG B cells (189) and somatic hypermutation resulted in an average germline similarity of 97.27% (± 1.9%).

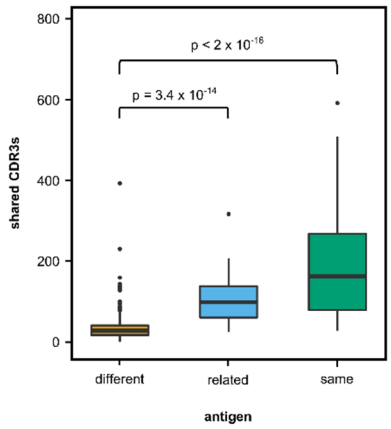


Figure 7 Shared CDR3s in OmniRat™–pairs. Box-whisker plots represent the number of identical CDR3s shared between pairs of rats from different (369 pairs, orange), related (56 pairs, lightblue), or the same vaccination group (71 pairs, green). More CDR3s were shared between rats from the same (p-value 3.4 x 10^{-14}) or related (p-value < 2 x 10^{-16}) antigen group than between rats of different antigen groups (Kruskal-Wallis test followed by Nemenyi post hoc test).

2.4.2. Highly similar CDR3 sequences in response to the same antigen

We first investigated to what extent rats that received the same antigen expressed the same CDR3 amino acid sequences. Pairs of rats from different vaccination groups (369 pairs) shared less CDR3s with each other than pairs of rats within the same group (71 pairs, p-value < 2×10^{-16}, Kruskal-Wallis with Nemenyi post-hoc test) or immunized with related antigens (56 pairs, p-value = 3.4×10^{-14}), indicating that mutual CDR3s are essentially induced by the immunizations (**Figure 7**). Among a total of 11,643 identical CDR3s (i.e. 100% similarity) that were shared by any set of two or more rats, irrespective of the antigen, 5,346 CDR3s (45.9%) were shared exclusively by animals of the same group and 1,912 (16.4%) were shared between animals immunized with a related antigen (TT and BaP-TT, MVA and MVA-HF). Most of the CDR3s shared within groups were common to only 2 animals of the same group (6,467; 89.1% of CDR3s shared within groups only). CDR3s present in all animals of a group were rare (**Table 7**). For instance, only a single CDR3 was shared between all 6 rats immunized with MVA-HF, and 3 CDR3s were shared between all the 12 animals exposed to the MVA vector (combined MVA and MVA-HF group). However, multiple CDR3s which differed only by one or two amino acids were shared by all animals within a vaccination group but not by animals from other groups (**Table 8**). Interestingly, these differences occurred preferentially at certain positions of a CDR3 amino acid sequence. This suggested that the vaccinations seemed to have induced identical CDR3s as well as clusters of highly similar CDR3s.

Table 7 CDR3s shared between rats in the same vaccination group. The first column shows the vaccination group, the last two rows are the combined vaccination groups with shared antigens. The second column shows the number of unique CDR3s not shared with any animal, i.e. the combined 'private' CDR3 repertoire if considering only identical CDR3s. The third column shows how many animals per group share unique CDR3s. Most CDR3s are not shared by all the animals of a vaccination group, considering only identical CDR3s as shared.

GROUP	Not shared	CDR3s shared by numbers of animals per group										
		n = 2	3	4	5	6	7	8	9	10	11	12
BaP-TT	27,167	510	74	12	0	-	-	-	-	-	-	-
TT	27,317	988	75	15	-	-	-	-	-	-	-	-
MVA	61,139	1,123	168	40	1	0	-	-	-	-	-	-
MVA-MV HF	43,778	686	51	4	0	1	-	-	-	-	-	-
NEG	33,675	455	37	3	0	-	-	-	-	-	-	-
BaP-TT + TT	/	258	101	24	11	0	1	1	0	-	-	-
MVA + MVA-MV HF	/	924	316	102	72	35	29	15	8	8	4	3

Table 8 Similarity of selected CDR3 sequences shared by rats in the MVA HF group. The first column shows a representative CDR3s and their relatives in animals within the vaccination group. Unchanged amino acids are described by points and changing amino acids are shown. The second column shows the number of animals in the group that express the different CDR3s. The CDR3s tend to vary only at certain positions.

MVA HF associated CDR3s																No. of animals
A	R	H	R	T	Y	Y	Y	G	S	G	S	P	L	F	D	Y	4
-	-	-	-	-	F	-	-	-	-	-	-	-	-	-	-	-	4
-	-	-	-	-	-	-	-	-	-	-	-	-	P	-	-	-	3
-	-	-	-	-	-	-	-	-	-	-	-	-	H	-	-	-	2
-	-	-	-	-	H	-	-	-	-	-	-	-	-	-	-	-	2
-	-	-	-	-	-	-	F	-	-	-	-	-	-	-	-	-	2
-	-	-	Q	-	-	-	-	-	-	-	-	-	R	-	-	P	2
-	-	-	-	-	F	-	F	-	-	-	-	-	-	-	-	-	2
-	-	-	-	-	H	-	-	-	-	-	-	-	I	-	-	-	2
-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	P	2
-	-	-	-	-	F	-	F	-	-	-	-	-	R	-	-	P	2
-	-	-	K	-	F	-	-	-	-	-	-	-	R	-	-	-	2
-	-	-	-	-	-	-	-	-	-	-	-	-	I	-	-	-	2
-	-	-	-	-	-	-	-	-	-	-	-	-	R	-	-	-	2

2.4.3. Shared antigen-related CDR3s at 80% sequence similarity.

We compared CDR3s within and across the different vaccination groups to estimate the degree of similarity between these antigen-related clusters. We determined which of the top 200 CDR3s, representing on average 72.1% (± 7.6%) of the repertoire of the rats. of any rat A had a related CDR3 in a rat B either within the same group (intra-group comparison) or between groups (inter-group comparison) allowing for a single amino acid substitution. The same analysis was repeated for two, three and up to eight amino acid substitutions. The number of top 200 CDR3s found to be present inter- and intra-group were plotted against the amino acid substitutions expressed as percent of CDR3 length

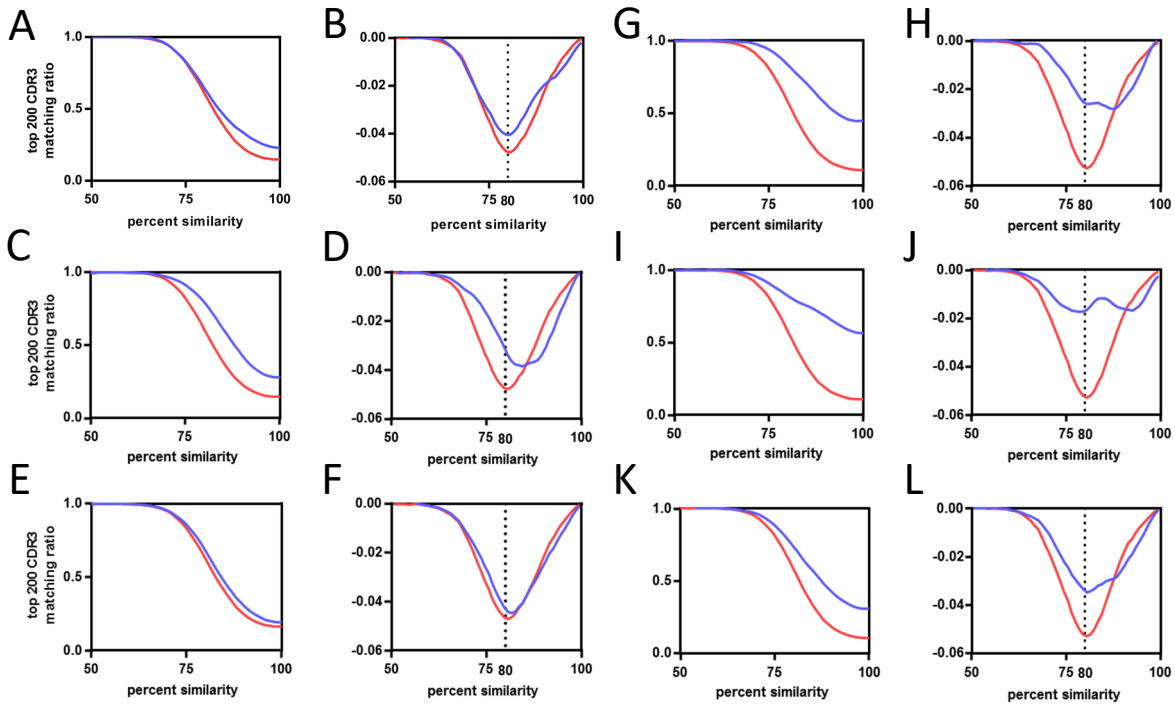


Figure 8 Influence of CDR3 sequence similarity on CDR3 repertoire overlap between rats. (A) Average fractions of top 200 CDR3s of the MVA-HF vaccination group shared with all CDR3s of other samples. Samples were divided into two groups having either the same antigen (HF group samples, blue curve) or different antigens (ALUM, BaP-TT, TT and NEG samples, red curve). Both curves follow a similar sigmoidal behavior. (B) First derivative of both curves. Inflection points align at 80% CDR3 amino acid similarity. (C, E, G, I, K) similar to (A) but for the MVA group, the combined groups MVA-HF + MVA, the BaP-TT group, the TT group and the combined groups BaP-TT and TT respectively. (D, F, H, J, L) Corresponding first derivatives similar to (B).

(**Figure 8**). The resulting sigmoidal curves showed a similar shape for all vaccination groups. In the exponential phase between 100% and 90-95%, intra-group overlap was higher than inter-group overlap. In the linear phase between 90-95% and 75%, overlap increased faster for the inter-group comparison. In the asymptotic phase below 75% similarity, both inter- and intra-group overlap leveled off towards 1, indicating that all top 200 CDR3s of a rat had relatives in any other rat, irrespective of the antigen administered. The first derivatives of the curves clearly showed that in all cases the inflection point was at around 80% (**Figure 8**). Thus, at this similarity threshold a maximum number of related CDR3s can be found within the same group while keeping the number of related CDR3s between groups at a minimum. In conclusion, all antigens induced in these rats a public IGH response that can best be characterized by clusters of CDR3s with at least 80% similarity.

2.4.4. Hierarchical clustering of CDR3 repertoires at 80% sequence similarity

Based on the above observation, we identified antigen-driven CDR3s using a workflow developed for differential gene expression analysis of RNA-seq data (27). For each CDR3 within a rat, counts of CDR3 sequences with 80% similarity (CDR3-counts) were used analogous to RNA-seq read counts. Rats of the same vaccination group were considered as replicates. The CDR3-counts followed a negative binomial distribution (**Figure 9A**). Compared to RNA-seq data, CDR3s usually lack a baseline

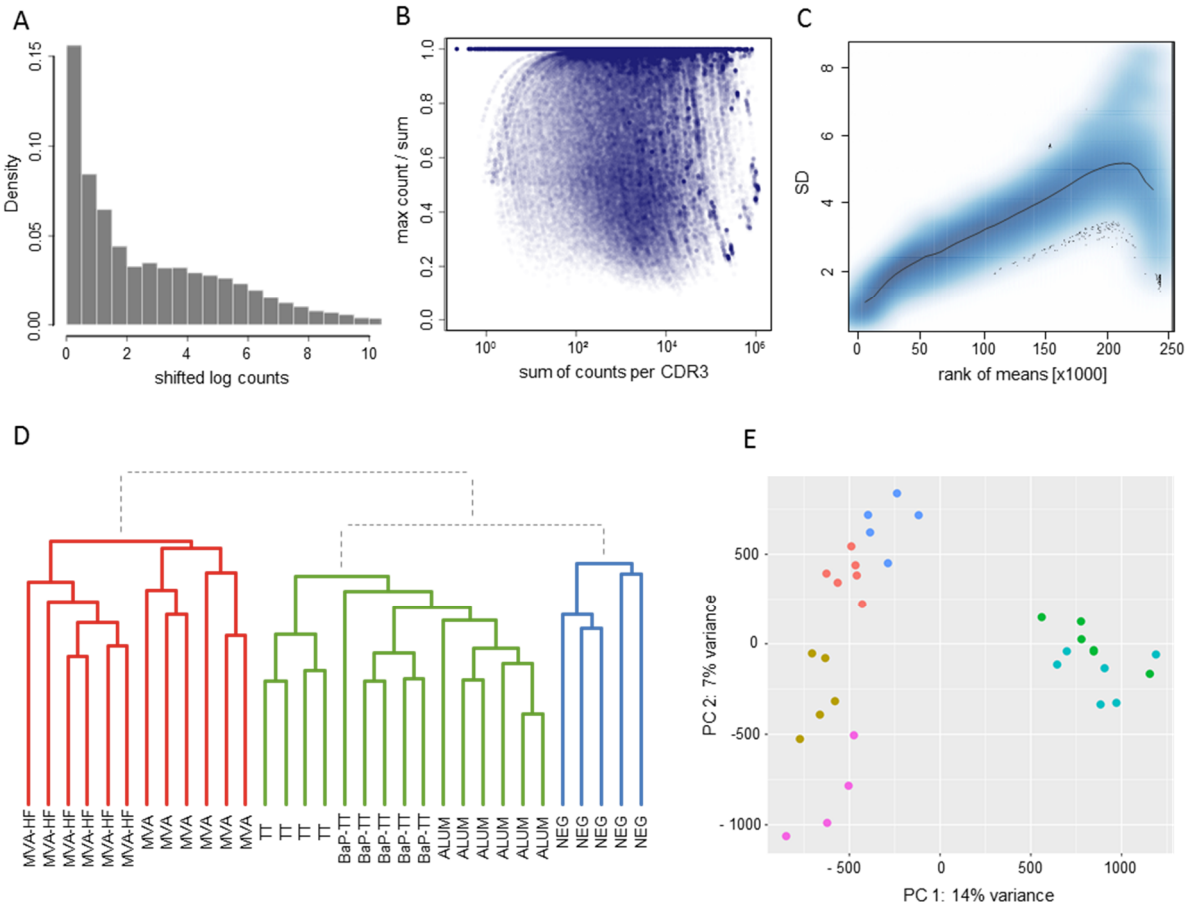


Figure 9 DESeq2 statistics and sample grouping. (A) Density-histogram representing the distribution of CDR3-counts (log(x+1) transformed). The CDR3-counts follow a negative binomial distribution (B) Sparsity-plot displaying the count distribution per CDR3. The sum of counts for every CDR3 is plotted in log10-scale against the highest count for the CDR3 divided by the sum of all counts for the CDR3. Density of data is indicated by hue. (C) CDR3-wise standard deviation of ranked means of counts after variance stabilizing transformation (VST-counts). The black line shows the standard deviation for all ranked means of VST-counts across all samples, the blue area indicates the data distribution and density by hue. (D) Dendrogram of the Euclidian sample distances calculated for VST-counts. Three main clusters are indicated by coloration (Cluster I: red, Cluster II: green, Cluster III: blue). (E) Scatterplot for the first two principal components of VST-counts. Samples are colored by vaccination-group (MVA: light blue, MVA-HF: green, TT: pink, BaP-TT: gold, ALUM: red, NEG: blue).

expression and are essentially private, resulting mostly in zero-counts for individuals across the study, while some shared CDR3s have very high counts in a single animal (**Figure 9B**). To account for this distribution, we applied variance stabilizing transformation (VST) to the CDR3-counts reducing the variance of the standard deviations over ranked means (**Figure 9C**). Hierarchical clustering of VST-

counts revealed three clusters (**Figure 9D**). Cluster I included all animals immunized with MVA (with or without MV HF protein expression). Interestingly, within this cluster, animals of the MVA-HF group and of the MVA group emerged from two separate branches indicating, that additional presentation of HF antigens leaves a distinct imprint in the CDR3 repertoire. Cluster II contained the three groups of animals that received alum as an adjuvant (TT, BaP-TT and ALUM). Again, each of the three groups clustered on separated sub-branches. Cluster III contained only untreated animals (NEG group) and was distinct from all immunized animals. The low variance and the specific grouping of the samples through both principle components showed that the VST-counts cluster the data by vaccination group. This indicated, that the different antigens had distinct impact on a subset of the IG repertoire of the rats (**Figure 9E**). When the data were reanalyzed applying an 85% or 75% threshold, the clear clustering of rats by vaccination group was lost (**Figure 10**), thus confirming that the 80% similarity threshold was optimal to identify antigen-associated responses on the CDR3 repertoire of the rats. Additionally, it showed that VST CDR3-count data can be analyzed analogously to RNA-seq count data.

2.4.5. Large numbers of antigen-driven CDR3s form stereotypic signatures

Similar to RNA-seq expression experiments, we aimed to identify CDR3s that are differentially represented between groups of rats. Based on a false discovery rate (FDR) of 1%, 16,727 of the 249,657 (6.9%) unique CDR3s across all groups were found to be overrepresented. One hundred-fold differences in numbers of overrepresented CDR3s were identified in each of the six antigen-groups

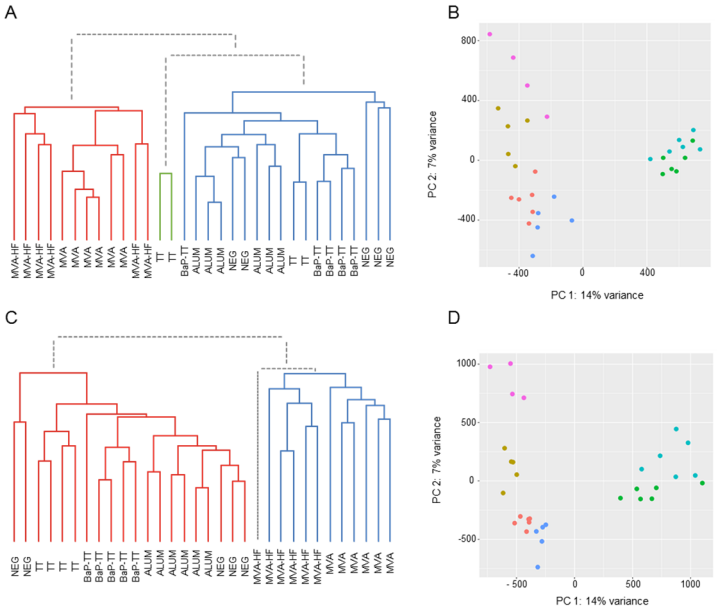


Figure 10 Sample grouping for 75% and 85% CDR3 similarity counts. (A) Dendrogram of the Euclidian sample distances calculated for VST-counts based on 75% CDR3 similarity. Three main clusters are indicated by coloration (Cluster I: red, Cluster II: green, Cluster III: blue). (B) Scatterplot for the first two principal components of VST-counts based on 75% CDR3 similarity. Samples are colored by vaccination-group (MVA: light blue, MVA-HF: green, TT: pink, BaP-TT: gold, ALUM: red, NEG: blue). (C) Dendrogram of the Euclidian sample distances calculated for VST-counts based on 85% CDR3 similarity. Three main clusters are indicated by coloration (Cluster I: red, Cluster II: green, Cluster III: blue). (D) Scatterplot for the first two principal components of VST-counts based on 85% CDR3 similarity. Samples are colored like in (B).

(**Table 9**). The highest number of overrepresented CDR3s was found in the two combined groups MVA

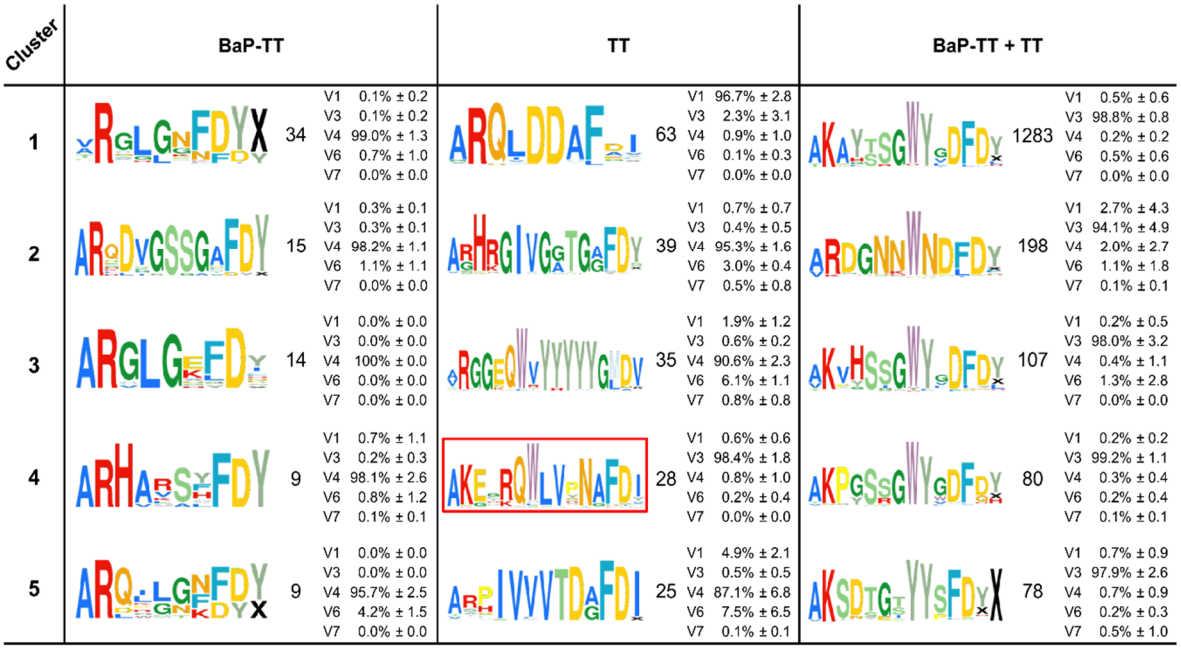
and MVA-HF (n=11,080, 10.4% of the unique CDR3s for this combined group), and TT and BaP-TT (n=2,451, 4.4%), which reflected the high immunogenicity of the antigens TT and MVA common within these groups. Less overrepresented CDR3s were found in the MVA (n=1,689, 2.6%), the MVA-HF (n=804, 1.8%) and TT group (n=540, 1.8%). The lowest number of overrepresented CDR3s was found in the BaP-TT group (n=163, 0.6%). These overrepresented CDR3s could be considered group-specific, and thus immunization induced.

Table 9 Antigen-driven sequences and 80% similarity clusters. The first column is the vaccination group. The second column describes the antigen(s) that these animals received with the immunization. The TT in the BaP-TT group is chemically altered in the conjugation process and therefore not identical with the one from the TT group. The third column is the number of unique CDR3s extracted as overexpressed by DESeq2. The fourth column contains the number of 80% similarity CDR3-clusters with more than 10 unique CDR3s. These are considered as the most important. The last column is the number of total clusters.

Vaccination group	Antigen	Overexpressed CDR3s	No. of Clusters (n>10)	Total clusters
BaP-TT	BaP + (TT)	163	3	20
TT	TT	540	14	46
MVA-HF	MV H+F	804	13	79
MVA	MVA	1689	28	99
BaP-TT & TT	TT backbone	2451	25	63
MVA-HF & MVA	MVA backbone	11080	233	419

Overrepresented CDR3s were grouped into clusters of 80% sequence similarity (**Figure 11**). The larger the antigen, the more clusters were found. For instance, 20 clusters were found for the BaP-hapten while 109 clusters were found for the TT protein (46 for TT alone and 63 for TT + BaP-TT combined). The largest number of clusters was found for the MVA virus antigen (518, 99 for MVA alone and 419 for MVA and MVA-HF combined). These complex antigen-driven clusters of CDR3s, typical for each group, represented up to 46.5% of the bone marrow IGH repertoire of the rats (**Figure 12**). The fraction of the repertoire corresponding to these CDR3 clusters varied between the groups but was relatively consistent among animals of the same antigen-group. Sequences encoding the overrepresented CDR3s were surprisingly diverse in IGHV gene usage. IGHV genes belonging to one family were largely predominant for each cluster (average 93.3% ± 9.8 of genes belonging to one family per cluster) and the

A



B

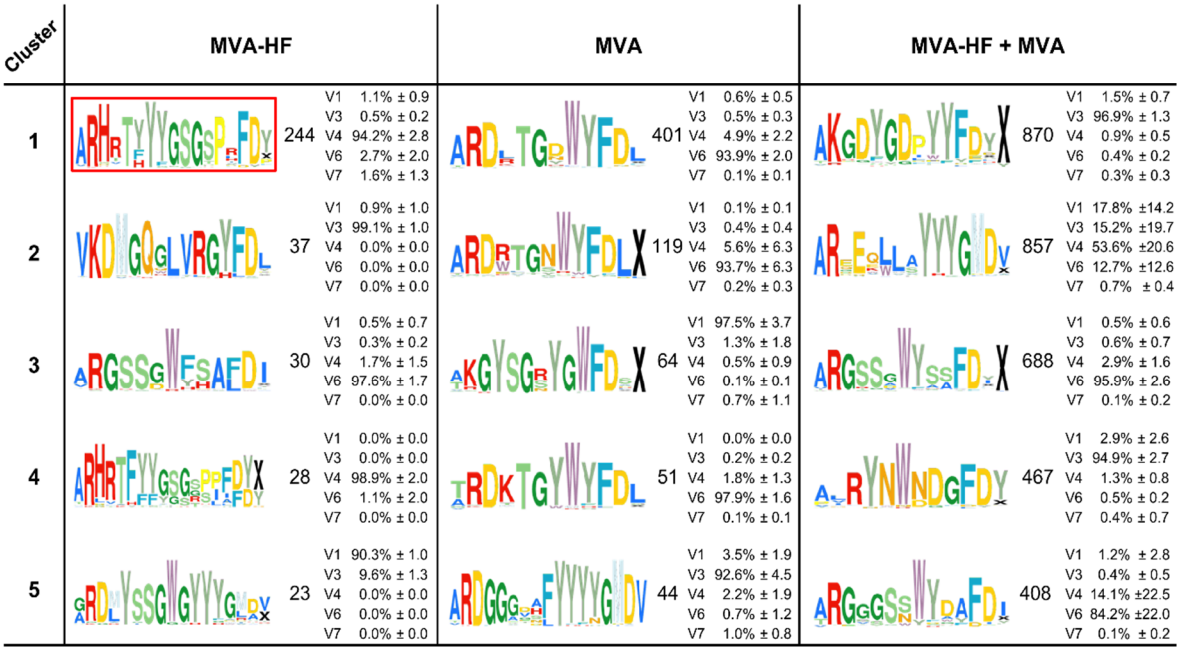


Figure 11 Antigen-associated CDR3-similarity clusters. The top 5 clusters of 80% similar CDR3s overexpressed in response to the antigens are shown as Weblogos. Coloration follows IMGT amino acid coloration scheme (372). Numbers represent the unique CDR3s in each cluster. Average IGHV gene family usage (± SD) is indicated as percentage of sequences across all rats per cluster (see also supplemental Table S3). IGHV2 was excluded, as no sequences in the clusters were derived from this family. (A) Clusters associated with the antigens BaP-TT, TT and the combined antigen groups BaP-TT and TT. The red box indicates TT-associated OmniRat™ CDR3s bearing an amino acid pattern also found in human anti-TT PBMC CDR3 sequences from for independent studies. (B) Clusters associated with the antigens MVA-HF, MVA and the combined antigens MVA and MVA-HF. The red box indicates MV-HF associated CDR3-signature also identified in OmniRat™ hybridomas generated in an independent experiment in response to MV antigens.

level of associated IGHV genes were comparable between rats of each cluster (**Figure 11**). Interestingly, *cluster 2* of the combined group MVA + MVA-HF showed an elevated IGHV gene family repertoire. The dominant IGHV4 gene family accounted only for 57.5% (± 20.4%) of the CDR3s and the IGHV gene families IGVH1 (16.3% ± 13.4%), IGVH3 (14.2% ± 18.5%), and IGVH6 (11.3% ± 11.6%) were predominating the *cluster 2* repertoire of one, two and one rat, respectively. All together we showed that OmniRat™ exhibited large fractions of highly similar, stereotypic CDR3s in response to the applied vaccinations, even across groups with shared antigens. All together, we showed that OmniRat™ exhibited large fractions of highly similar, stereotypic CDR3s in response to the applied vaccinations, even across groups with shared antigens.

2.4.6. Stereotypic signatures match MV-specific and TT-specific CDR3s

MA-HF signatures were compared to the previously described CDR3s of MV-specific hybridoma clones derived from an independent set of OmniRat™ immunized with whole MV antigens (15). The largest of the identified HF associated clusters (244 members) matched three CDR3s of MV-specific hybridoma cells, suggesting that this cluster is an MV-H or F protein induced CDR3 signature (**Figure 13**). Similarly, our TT-associated clusters were compared to known human TT-specific IGH sequences (9, 16–18). The CDR3s from the TT-associated *cluster 4* matched 12 published human CDR3s (**Figure 14A**). This signature and the human CDR3s consisted of 15-mer CDR3s following the same amino acid pattern. Both humans and rats elicited a conserved paratope defined by a static motif '+QWLV' ('+' = R/K) at the center of the CDR3, flanked by variable positions that are connected to the torso of the CDR3 (**Figure**

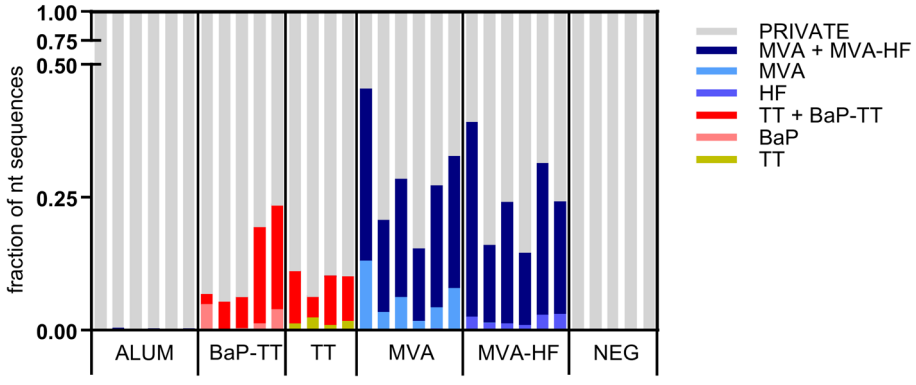


Figure 12 Fractions of the nucleotide IG repertoire encoding for CDR3 signatures. The IG repertoire per sample is displayed using numbers of full length nucleotide sequences. Nucleotide sequences encoding for CDR3s that are part of a signature are colored by associated antigen.

14B). This indicates that similar key are used even across species. The sequence similarity between the human and rat CDR3s ranged from 67% to 87% resulting from different torso amino acid compositions at the positions flanking the conserved binding motif (**Figure 14C+D**). To compare the structures of these CDR3s from human and rat origin, we performed 3D homology modeling on their Fab-fragments. Four human antibodies with available heavy and light chain sequences (17) and four selected OmniRat™ heavy chain sequences paired with the human light chains were modeled with Rosetta Antibody. Within the OmniRat™-human chimeric Fab-fragments, the CDR3s formed torso structures ranging from unconstrained amino acid formations over short beta-sheets to rigid beta-sheet hairpin constructs (**Figure 14C**). Like the rats, human CDR3s exposed the key binding residues at the very tip of the CDR3 loop by a rigid beta-sheet hairpin formation of the torso that protruded out of the IGH core structure (**Figure 14D and 15**). Together our results corroborate the evolution of functionally convergent CDR3s in different individuals and by different vaccines delivering the same antigen. Also, this strongly indicates that OmniRat™ and humans, albeit the lower sequence similarity between their TT-associated CDR3s, produce antibodies with highly homologous CDR3s in response to the same antigen.

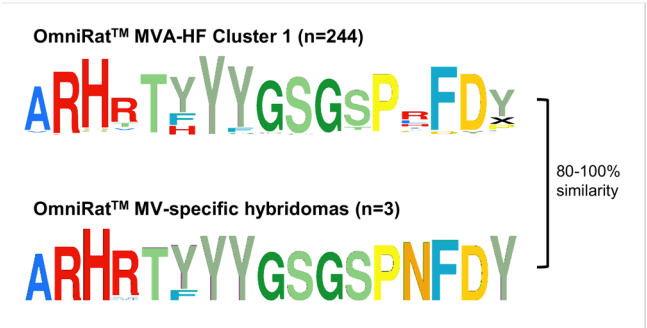


Figure 13 OmniRat™ MV-specific CDR3 signature. The clusters of CDR3s overrepresented in response to MVA-HF (see also **Figure 11B**) and the CDR3s from 3 monoclonal hybridomas specific for MV-H protein are shown as weblogos. The differences between the sequences were calculated as Levenshtein distances in percent of CDR3 length.

2.5. Discussion

We analyzed more than 2,700,000 functional IGH sequences derived from the bone marrow of transgenic rats expressing human B cell receptor genes immunized with different antigens. Our study showed that these rats produced identical as well as highly similar CDR3 amino acid sequences in response to common antigenic challenges. When shared CDR3 repertoire fractions were investigated at different levels of sequence similarity, overlaps between rats from the same vaccination group were optimal around 80% CDR3 amino acid similarity. Applying a differential gene expression workflow to the counts of 80% similar CDR3s, we presented a novel way to identify convergent, stereotypic CDR3 sequences in response to an antigenic stimulus. These included known CDR3s induced by different

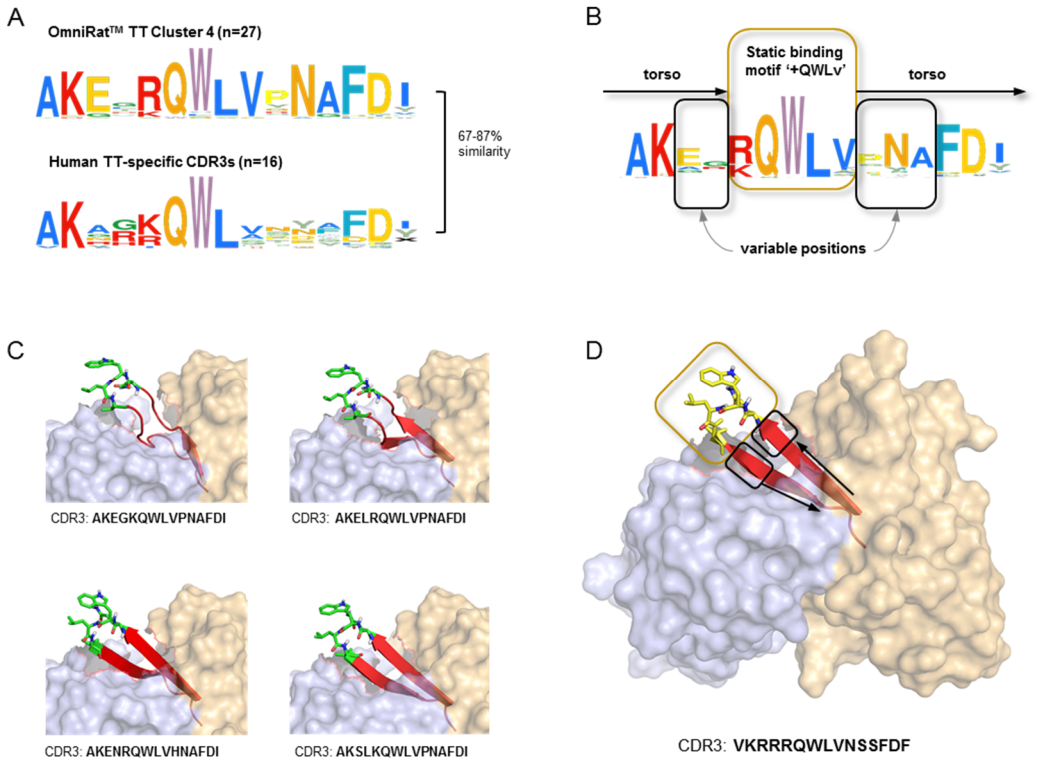


Figure 14 Omnirat™ and human antibodies against tetanus toxoid with similar properties and structures. (A) Sequence similarity range between OmniRat™ TT-associated *cluster 4* (Fig. 4B) CDR3s and human TT-specific CDR3s (Levenshtein distance as percent of sequence length). (B) Amino acid pattern for the combined TT-specific human and TT-associated rat CDR3s. The weblogo highlights the conserved binding motif '+QWLV' and torso amino acids with variable positions are highlighted. (C) 3D-homology models of four OmniRat™-HC-human-LC chimeric antibody fab fragments. Heavy chains are colored in orange and light chains in blue both visualized with 50% transparent surface. CDR3 torsos are shown as cartoon and colored in red. Binding motifs are displayed as sticks and colored in green with only polar hydrogens shown. Views were enlarged to focus on the CDR3 structure (D) Complete human fab-fragment visualized as described for (C). Motif, variable and torso structures are highlighted with boxes and arrows as described in (B).

measles antigens, indicating that the identified CDR3s are specific for the measles virus H or F proteins which were shared in both immunizations. In addition, our approach also identified CDR3s in response to tetanus toxoid that were remarkably similar to known tetanus-specific CDR3s from human samples. Our findings highlighted the presence of convergent IGH transcripts at high levels in the bone marrow of the transgenic rats and that these sequences are highly similar to those of humans.

Pairs of rats within the same group shared more identical CDR3s than pairs from different groups, but very few CDR3s were shared among all rats of a group. Given the tremendous size and diversity of the IG repertoire, finding identical sequences in several individuals is indeed unlikely (84). Because of private processes during B cell development including stochastic affinity maturation of the IG molecules, a certain variability in CDR3s converging towards reactivity with the same antigen is to be expected (56, 158, 168). Galson and coworkers found that for the identification of public repertoires in humans an 87.5-91.6% cut-off (1 in 12 to 1 in 8 amino acids) was optimal to identify TT- and influenza-related CDR3

clusters (199). In the present study, we explored the relation between CDR3 sequence similarity and the overlap between CDR3 repertoires, by inter- and intra-group cross-comparisons at different levels of sequence similarity. Our data showed that 80% amino acid similarity optimized the intra-group overlap between CDR3 repertoires while keeping the inter-group overlap at a minimum. The identified antigen-associated clusters were absent in rats outside immunization groups, which provides a strong support for their underlying biological relevance.

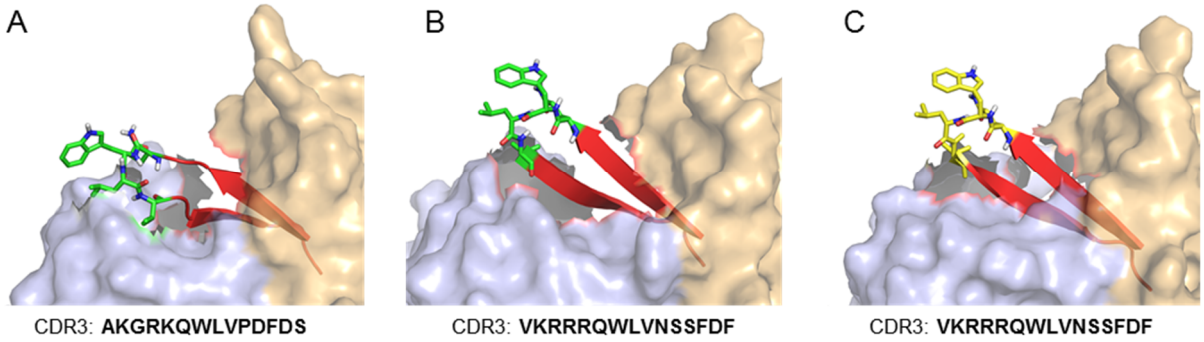


Figure 15 Homology models of three human antibodies against tetanus toxoid bearing the '+QWLV' binding motif. Heavy chains are colored in orange and light chains in blue both visualized with 50% transparent surface. CDR3 torsos are shown as cartoon and colored in red. Binding motifs are displayed as sticks and colored in green with only polar hydrogens shown. Views were enlarged to focus on the CDR3 structure.

We showed that between 6% and 46% of the bone marrow IGH repertoire correspond to convergent CDR3 sequences. Similar proportions (15-50%) of antigen-specific CDR3 sequences were reported in peripheral blood B cells of patients with acute dengue infections (90). In contrast, in the reconvalescent dengue patients as well as in influenza patients, convergent sequences represented only to less than 1% of peripheral B cell sequences. Such human studies are normally restricted to peripheral blood where only a small fraction of the repertoire can be found and assessed (172, 200). Blood is the only readily available source for sampling B cells in humans. However, the high turnover rate of 5×10^{11} B cells per day (201) and low sampling depth makes it difficult to capture a significant fraction for exhaustive analysis of antigen experienced Ig repertoires in a vaccination context (173, 202). In contrast, high levels of antigen selected B cells can be found within the bone marrow, where about 17% of all B cells reside, making this tissue a preferable target for studying the antigen-specific B cell response after vaccination. In this regard, elevated levels (37%) of public clones were observed in mice in response to hepatitis B surface antigen and less to 4-Hydroxy-3-nitrophenylacetyl hapten conjugated to Hen Egg Lysozyme (22%) and OVA (14%), when examining bone marrow derived long-lived plasma cells ($CD138^+$ $CD22^-$ $MHCII^-$ $CD19^-$ IgM^- PI^-) (173). In the present study, we analyzed IG mRNA from bulk

rat bone marrow cell isolates, an organ rich in serum antibody-producing plasma cells (28, 174, 175). Because plasma cells express significantly higher levels of IG mRNA (estimated 500:5:2 compared to memory or naïve B cells, (72)) and only class switched BCR (IgG) were targeted, those are overrepresented in our datasets (56, 199). This explains the high fractions of converging CDR3s we found in the assessed bone marrow IgG repertoire. We thus primarily targeted antigen-associated effector B cells, facilitating the tracking of antigen-specific sequences induced by similar antigens.

The IGH repertoire of OmniRat™ displayed a CDR3 length distribution comparable to that of human IgG B cells isolated from peripheral blood and lower than that of the corresponding IgM B cells (189). This indicates, that OmniRat™ are generating CDR3 diversity in an equivalent way as humans. Interestingly, the average mutation rate of OmniRat™ IgG sequences was much lower as human blood-derived IgG or IgM B cells. This is most likely due to the much younger Ig repertoire of the rats having not been exposed to a large number of antigens as compared to humans, leaving their repertoires relatively unchallenged and thus unmutated (173).

The clusters of antigen-induced CDR3s exhibited a diverse IGHV gene usage from mainly one predominant IGHV gene family. Only CDR3 *cluster 2* of the animals in the combined vaccination group MVA-HF + MVA exhibited significant differences in IGHV gene family usage across the rats. Similarly, different IGHV gene families were found in convergent CDR3 responses to dengue infections and anti-polysaccharide vaccination in humans implying a convergent evolution (90, 92). In an antigen-antibody binding scenario the IGHV gene encodes for a structural scaffold while the CDR3 must precisely fit the antigen surface (3). Hoogenboom and Winter concluded from their synthetic antibody library that substitution of the CDR3 alone can create entirely different antibody specificities (203). However, in contrast to the other 29 clusters, the 15-amino acid long signature of MVA-HF + MVA cluster 2 was largely composed by the IGHJ6 gene segment resulting in a tyrosine-rich 'YYYGMDV' tail motif, with a shorter section from the more diverse D/N region. Similarly, the signatures reported by Parameswaran an colleagues in response to dengue infections were composed largely of a long, tyrosine-rich tail motif resulting from the IGHJ region with very short D/N sections (90). We expect that such CDR3 signatures exhibit a rather polyreactive binding mediated by the abundant tyrosine residues. Yet, the absence of any 80% relatives in the other vaccination group underlines their presence as a result of the antigen exposure through the applied vaccination.

The varying IGHV gene assignments in each cluster can be explained by the relatively short read length of our approach. In this regard, IGHV genes belonging to the same family (e.g. IGHV4-39 and IGHV4-

34) share a high level of similarity, only differing by few nucleotides. As the amplicons of the rat IG VDJ transcripts did not span the whole IGHV sequence, information encoded in the FR1 and CDR1 are lost. This could easily lead to varying gene assignments, explaining the IGHV gene variability of the described CDR3 clusters.

Potential influence on the repertoire composition could result from PCR amplification biases introduced during library preparation as well as sequencing errors (57). We did not account for potential errors and sequencing bias by using molecular barcodes or similar methods (67, 71–73). However, the data analysis of the present study was based on collapsed, unique nucleotide Ig sequences, minimizing the influence of potential PCR amplification bias. Analysis of the nucleotide sequences before and after collapsing to unique nucleotide sequences revealed no major difference in our findings, indicating that PCR amplification bias did not falsifying our results. The Ion Torrent PGM sequencing platform is prone to insertion and deletion errors, especially within homopolymer repeats (75). Such errors cause frameshifts within the IG sequence which are detected by IMGT with 98% efficiency in a benchmarking setup, missing only indels at the beginning and end of the sequence or if placed in close proximity to each other masking the resulting frameshift (see **Chapter 4**)(204). Sequences with detected indels are marked by IMGT as productive with detected errors and were not included in the described analysis. Furthermore, our analysis is based on the CDR3 amino acid sequence. An insertion or deletion within the CDR3 encoding nucleotides results in the sequence being labelled as unproductive, with no correction attempts undertaken by IMGT (204). Less than 1% of indel combinations remain undetected by IMGT and could be present within the CDR3 encoding nucleotides (204). These rare combination of sequencing errors would then result in artefactual CDR3s either covered by the applied 80% sequence similarity clustering threshold or missed because of higher sequence variation. Therefore, such CDR3 artefacts can be expected to induce only a small underrepresentation of CDR3s by lowering CDR3-counts. In conclusion, the presented workflow is well protected from potential sequencing errors or PCR bias, that could impact our conclusions.

We found that certain CDR3s have high counts of 80% relatives within a group but very few to none in the unrelated groups. This is in principle comparable to differential gene expression in RNA-seq data. The CDR3-counts followed a negative binomial distribution but, unlike in RNA-seq experiments, our data contained large amounts of CDR3s with zero-counts over different samples. These correspond to private CDR3s that are absent in other rats of the same or other groups. On the other hand, some CDR3s exhibited very high counts of 80% relatives within an animal. While such a data distribution is uncommon

in RNA-seq, they were nevertheless compatible with our computational approach (DESeq2, (191)) as demonstrated by negative binomial data distribution and perfect sample grouping after variance stabilizing transformation of the CDR3-counts. Interestingly, the Euclidian distance grouping of MVA-HF and MVA rats remained unchanged for 85% and even for 75% CDR3-counts in contrast to TT-associated rats. Similarly, Trück and colleagues found highly similar (≤ 2 mismatches) Hib- and TT-related sequences enriched seven days post-vaccination, but could not identify H1N1- and MenC-related sequences at the same threshold (92). Correspondingly, statistical evidence of convergent CDR3s in pairs of donors against influenza with a mean genetic distance of ~75% were reported (205). Together with our data this indicated that the identification of convergent Ig repertoire responses using amino acid similarity thresholds was applicable. Future research will tell, to what extent the 80% threshold can be applied to other antigens.

Identified convergent CDR3 matched to sequences of previously described human monoclonal antibodies against TT protein (92, 178–180). Despite the relatively low sequence similarity (67% to 87%) between OmniRat™ and human TT-specific CDR3s, they shared a common sequence and structural motif at the center of the CDR3. The center part of the CDR3 is exposed at the tip of the loop structure which directly interacts with the antigen while the adjacent amino acids act as a supporting scaffold. Similarly, Greiff and coworkers observed stereotypical motifs at the center of the CDR3 amino acid sequences in specific antibodies following 4-Hydroxy-3-nitrophenylacetyl vaccination in mice (173). While structural similarity cannot readily be used to determine antibody specificity, algorithms to identify convergent CDR3s could be further improved by including structural parameters drawn from the expanding amount of available crystal structures.

In conclusion, we demonstrated a strong public IGH response with converging and overlapping CDR3 repertoires in animals exposed to the same antigens. These converging repertoires consisted of similar CDR3 sequences that can be best described using an 80% amino acid similarity threshold. Additionally, we presented an approach to identify such CDR3s by adopting a group-wise expression analysis, similar to RNA-seq approaches. This provides also a valuable tool for large-scale HTS data-mining to identify potential candidates for high-affinity targeted antibody design.

Chapter 3

Sustained NF-kB Signaling Synergizes with sCYLD Expression in Progression from Monoclonal B Cell Lymphocytosis to Chronic Lymphocytic Leukemia

Matthias **HAHN**[1], Jean-Philippe **BÜRCKERT**[2], Carina **LUTTENBERGER**[1], Sabrina **KLEBOW**[1], Moritz **HESS**[3], Mona **AL-MAARRI**[4], Merly **VOGT**[4], Sonja **REIßIG**[1], Michael **HALLEK**[5], Thorsten **BUCH**[6], Anke **WIENECKE-BALDACCHINO**[2], Claude P. **MULLER**[2], Christian **PALLASCH**[5], Thomas **WUNDERLICH**, Ari **WAISMAN**[1] and Nadine **HÖVELMEYER**[1]

[1]*Institute for Molecular Medicine, University Medical Centre of the Johannes Gutenberg-University of Mainz, 55131 Mainz, Germany*

[2]*Department of Infection and Immunity, Luxembourg Institute of Health, Esch-Sur-Alzette L-4354, Grand-Duchy of Luxembourg*

[3]*Institute of Medical Biostatistics, Epidemiology and Informatics, University Medical Centre of the Johannes Gutenberg-University of Mainz, 55131 Mainz, Germany*

[4] *Max Planck Institute for Metabolism Research, CECAD, CMMC, Institute for Genetics, 50931 Cologne, Germany*

[5]*Department I of Internal Medicine, CMMC, CECAD, University of Cologne, 50935 Cologne, Germany*

[6]*Institut of Laboratory Animal Science, University of Zürich, Zürich, Switzerland*

**Corresponding author: Dr. Nadine Hövelmeyer (Email: hoevelme@uni-mainz.de)*

[This chapter is based on a manuscript accepted May 22, 2017 in Leukemia, DOI: 10.1038/leu.2017.168, a previous version of it was part of the doctoral thesis of M. Hahn]

Authors' contributions: J.-P. B. designed the HTS-based approach of this manuscript and developed the bioinformatics data analysis; developed the HTS sample preparation protocol; adapted the HTS raw data processing script; wrote the HTS data analysis script; prepared the HTS samples; processed, analyzed and interpreted the HTS data; made the corresponding figures and wrote the corresponding sections in the manuscript. A. W.-B. designed and developed the original HTS raw data processing script. C. P. M. supervised HTS work and corrected the corresponding manuscript sections. All other research was conducted by M. H. under the supervision of N. H. and A. W. The detailed contributions of the other authors can be assessed in the original manuscript.

3.1. Summary

NF-κB activation has been implicated in the pathogenesis of chronic lymphocytic leukemia (CLL) and clinically asymptomatic monoclonal B-cell lymphocytosis (MBL). We demonstrate that enhanced canonical NF-κB activation, driven by B-cell specific expression of a natural splice variant of the tumor suppressor CYLD, sCYLD, lacking domains required for NF-κB inhibition, leads to the development of high penetrance, indolent clonal B cell lymphoproliferation, resembling MBL. Additional B-cell-specific deletion of the deubiquitinase A20, reinforced NF-κB activity, thereby accelerating monoclonal expansion of $CD5^+$ B cells due to their accelerated survival and proliferation, recapitulating hallmarks of human CLL. Importantly, we show sCYLD expression associated with human CLL, suggesting that inhibition of alternative splicing of this key regulator is essential for keeping B cells nonmalignant.

3.2. Introduction

Chronic lymphocytic leukemia (CLL) represents the most common type of leukemia among adults in the Western world and is characterized by the accumulation of abnormal monoclonal $CD5^+$ B cells in blood, bone marrow, and secondary lymphoid tissues (206). Several biological parameters have been introduced to characterize the heterogeneity of CLL and to evaluate the prognosis of CLL patients (118, 207). Leukemic clones with unmutated immunoglobulin heavy chain (IGHV) genes (U-CLL) correlate with poor clinical outcome while clones with frequently mutated IGHV (M-CLL) genes exhibit an indolent asymptomatic course (139, 150, 208, 209).

Many factors have been shown to contribute to the etiology of CLL, including genetic predisposition, auto antigen stimulation, microRNAs and cytogenetic abnormalities (210). A number of signaling pathways, such as the nuclear factor kappa B (NF-κB) pathway, show constitutive activity in CLL cells (211, 212), which is suggestive of a pathogenic role. However the underlying molecular mechanisms leading to its constitutive activation remain largely unknown (213–215). In contrast to other mature B cell lymphomas only a few recurrently mutated genes involved in canonical or non-canonical NF-κB activation have been identified in CLL (e.g. BIRC3, MYD88, and NFKBIE) (216).

To maintain immune cell homeostasis and to prevent persistent NF-κB activation, a tight control of this pathway is required. One mechanism to regulate NF-κB pathway activity is ubiquitination. Whereas linkage with polyubiquitin chains via lysine 48 (K48) results in proteasomal degradation of the target protein, polyubiquitin chains linked via K63 have non-degradative, regulatory functions and serve in the recruitment of various kinase complex platforms (217). Ubiquitination is a reversible process mediated by deubiquitinating enzymes (DUBs), such as CYLD and A20, which prevent persistent NF-κB pathway activation by deubiquitinating their target proteins (218, 219).

CYLD, originally identified as a tumor suppressor gene mutated in familial cylindromatosis (220), has been shown to regulate diverse biological functions, including immune cell development, activation, inflammation and tumorigenesis (221). Previously, we have identified a short isoform of CYLD (sCYLD), which is encoded by a natural splice variant of the *Cyld* mRNA that retains DUB activity but lacks the domain harboring the binding sites for TRAF2 and IKKy/NEMO, required to dampen NF-κB activation. Mice lacking full-length (FL)-CYLD but overexpressing exclusively sCYLD, demonstrate enlarged lymphoid organs resulting from an expanded B2 B cell compartment. This increase is a consequence of a B cell-intrinsic survival advantage, which has been associated

with higher expression levels of Bcl-2 (222).

Downregulation of FL-CYLD expression was found in various types of cancers, including CLL (223). CLL B cells fail to undergo TNFα-induced cell death due to LEF1-mediated downregulation of CYLD and RIPK3. Reduced CYLD protein levels in CLL cells resulted in constant K63 ubiquitination of the kinase RIPK1 leading to permanent activation of the canonical NF-κB pathway and resistance to apoptosis (215).

Similar to CYLD, A20 (TNFAIP3) is capable of removing activating K63, but also K48 linked ubiquitin chains from its target proteins (219). B cell-specific deletion of A20 results in sustained canonical NF-κB activity in B cells and predisposes mice to autoimmunity (224–226). The importance of A20 regulatory activity is further emphasized by the finding that mutations in the A20 locus are associated with various human B cell lymphomas (217) although not to CLL (227, 228).

Here, we report that mice with accelerated canonical NF-κB activation, driven by sCYLD overexpression in B cells, spontaneously develop a $CD5^+$ B cell lymphoproliferative disorder. Further enhancement of NF-κB activation, by additional B cell-specific deletion of A20, reinforces clonal accumulation of $CD5^+$ B cells ultimately leading to a late-onset CLL-like disease, recapitulating hallmarks of human CLL. Importantly, we found that a substantial number of malignant cells of CLL patients express the sCYLD variant. These findings demonstrate that alternative splicing of the NF-κB regulator *Cyld* predisposes accumulating $CD5^+$ B cell to malignant transformation, while further accelerated NF-κB activation promotes the progression to a CLL-like disease in mice. Thus, interfering with ubiquitination-driven NF-κB activation represents a promising therapeutic target for the treatment of CLL and identifies sCYLD as a potential risk factor of this disease.

3.3. Materials and Methods

3.3.1. Mice

Cyld$^{F/F}$ and A20$^{F/F}$ mice have been described previously (222, 225). A20^{BKO}sCYLD^{BOE} were generated by crossing A20^{FF}CD19-Cre mice to CYLD^{FF} mice. CYLD$^{full\text{-}KO}$ mice (229) were bred to A20^{FF}CD19-Cre mice to obtain A20^{BKO}CYLD^{KO} mice. CD45.1 mice were obtained from the central breeding (TARC, Mainz). All mouse experiments were approved (G 11-1-026).

3.3.2. CLL patient samples

Primary CLL cells were obtained from the peripheral blood of patients after written informed consent to the Declaration of Helsinki and with Institutional Review Board approval (#11-319) at the University of Cologne.

3.3.3. B Cell Isolation, Proliferation and Survival Analysis

Spleens were harvested mice. For B cell isolation, homogenized splenocytes were incubated with anti-CD19 or anti-CD43 microbeads for 15 min at 4°C, washed with FACS buffer, and separated on LS or LD columns (Miltenyi Biotec). Purity was determined by flow cytometry being 95-98%. Complete media consisted of RPMI-1640, 10% FCS, 10 mM HEPES, 0.05 mM β-mercaptoethanol and L-glutamine. For in vitro proliferation assay 3×10^5 VioletCellTracer labeled B cells were stimulated with a final concentration of Fab anti-mouse IgM [10µg/ml] (Jackson ImmunoResearch), anti-CD40 [10µg/ml] (BioXCell, West Lebanon), LPS [20µg/ml] (Sigma) or CpG [0,1µM] (InvivoGen). Cells were acquired on FACSCanto II and analyzed with FlowJo software. B cell survival was determined by 7AAD staining using FlowJo software for data analysis and by counting live cells using Trypan blue.

3.3.4. In Vivo BrdU-Labeling

To analyze BrdU incorporation in vivo, mice were fed with 1mg/ml bromdesoxyuridin (BrdU) (Sigma) and 1 % sucrose in the drinking water for 5 and 10 days. BrdU uptake of splenic and peritoneal cavity B cells was determined by FACS analysis. Therefore 2-4×10^6 cells per sample were surface stained according to normal FACS. Afterwards, stained cells were washed once with PBS-FCS 2%, resuspended in 200µl PBS and subsequently injected into 1ml ice-cold 70% ethanol and incubated 30 min on ice for fixation. For permeabilization an equal volume of PBS-1%paraformaldehyd- 0,01%Tween-

20 (v/v) was added and cells were incubated for 1h on ice or overnight at 4°C. On the next day, cells were DNaseI treated (1ml PBS^{++} plus 300µg/ml DNaseI (Roche)) for 10min at RT and washed once with PBS-FCS 2%. Then cells were stained with BrdU antibody for 20min at RT. BrdU incorporation was compared to splenic cells of non-BrdU treated mice. The percentage of BrdU- positive cells was detected by FACS.

3.3.5. RNA isolation and real-time PCR

RNA was isolated (QIAGEN RNeasy Mini Kit) and reverse transcribed (Promega Kit) for quantitative real-time polymerase chain reaction (qRT) using probes and primers from Qiagen as described on their homepage: http://www.quiagen.com/poducts/pcr/quantitect/primerassays.aspx

3.3.6. Array Analysis

Total RNA (1µg) was amplified and labeled using the Affymetrix One-Cycle Target Labeling Kit (Freiburg, Germany) according to the manufacturer's recommendations. As newly transcribed RNA mainly consists of mRNA, it was amplified and labeled according to the manufacturer's protocol for mRNA. The amplified and fragmented biotinylated complementary RNA (cRNA; 15µg) was hybridized to Affymetrix. Mouse Gene 1.0 ST Arrays using standard procedures. Arrays were assessed for quality and robust multi-array average (RMA)-normalized. Quality assessment consisted of RNA degradation plots, Affymetrix quality control metrics, sample crosscorrelation, and probe-level visualizations. Normalization incorporated (separately for each RNA-type data set) background correction, quantile normalization, and probe-level summation by RMA. The data were analyzed for differential gene expression using an empirical Bayes moderated t-test, implemented in the Bioconductor package Linear Models for Microarray Data (LIMMA). The results were sorted by adjusted p-value and exported in tab-delimited format.

3.3.7. Ion Torrent PGM library preparation and sequencing

Libraries of immunoglobulin heavy chains (IGH) were prepared using an adapted version of the unique molecular identifiers described elsewhere (71). 500ng purified RNA from peritoneal cavity and spleen were reverse transcribed using Superscript III (Life technologies) according to the manufacturer's protocol and a mouse IgM specific constant region primer (IGHCµ). This primer contained an 8-random-nucleotide unique identifier (8N-UID) and partial Ion Torrent PGM sequencing adapter A (**Table 10**).

Second strand synthesis was performed using Phusion polymerase (NEB) with a mouse IGHV region specific multiplex primer set (**Table 10**) containing 8N-UIDs and the partial sequencing adapter P1 (1 cycle with 98°C for 2min, 50°C for 2min and 72°C for 10min). The double stranded cDNA libraries were purified twice using AMPureXP Beads (Agencourt, 1:1 v/v) and subsequently amplified with Q5 polymerase (NEB) and full sequencing adapters as primers (**Table 10**; 98°C for 5min, 20 cycles of 98°C for 10sec, 65°C for 20sec and 72°C for 30sec, and 72°C for 2min). Amplified libraries were purified twice with AMPureXP Beads (1:1, v/v) and quality was controlled using an Agilent 2100 Bioanalyzer (Agilent Technologies). Libraries were sequenced on the Ion Torrent PGM platform (400bp protocol, 318v2 chip, 5 libraries/chip) according to the manufacturer's instructions (all trimming options disabled in the Torrent Suite™, version 4.4).

3.3.8. Computational analysis of Ion Torrent PGM data

The two 8N-UIDs of each read were used to reverse the PCR amplification and sequencing error-correction on the datasets (see (71, 73) for a detailed description of similar approaches, possible errors and necessary correction measures). Raw reads were demultiplexed, quality filtered (80% bp, QSC > 20) and grouped into UID families based on their dual 8N-UIDs. With an in-house IgBLAST installation ((188), http://www.ncbi.nlm.nih.gov/igblast/, version 1.2.0) sequences were checked for a consistent IGHV and IGHJ region assignment (IMGT classification scheme) inside a UID family and productivity (e.g. no stop codons). From productive, unambiguously assigned sequences, a consensus sequence was built for each UID family using the pagan multiple sequence aligner ((230), http://code.google.com/p/pagan-msa/, version 0.47). Consensus sequences were again checked for productivity with IgBLAST to account for incorrect consensus builds. The remaining sequences, each representing one original IGH mRNA molecule, were collapsed and uploaded to IMGT (www.imgt.org, HighV-QUEST version 1.3.1) for additional error correction (IMGT indel detection algorithm) and final classification. IMGT readout was filtered for productive sequences and collapsed. To remove remaining artificial variants (mostly undetected early PCR or reverse transcription errors), only sequences with at least two identical nucleotide copies were considered. Data analysis was performed with custom-made python and R scripts in a Linux environment. Plots were generated using Graphpad Prism (version 5) and R (version 3.2.3).

3.3.9. Network analysis of BCR repertoire data

A custom-made R script based on the iGraph module in R (http://igraph.sourceforge.net/index.html) was used to compute and display vertex cluster networks: Sequences were displayed as vertices, with log-transformed normalized counts (ratio of total count) providing the vertex size. Vertices were connected if they differed by one nucleotide (calculated as hamming distance of 1). A vertex cluster is defined as group of connected vertices. Vertex cluster networks were compared by their Gini coefficients (231) calculated for vertex and cluster distributions using the “ineq” package in R. The Gini coefficient provides an unevenness measure ranging from 0 (even distribution) to 1 (maximum unevenness of population).

3.3.10. Protein Isolation and Western Blotting

Total protein lysates, cytoplasmatic and nuclear fraction were isolated as previously described (232). Equal amounts of protein lysates were separated on a 4-12% bis-tris polyacrylamyd gradient gel (Invitrogen) and transferred onto a polyvinylidene fluoride membrane. Membranes were subsequently blocked for 1hr with 5% milk powder (MP) in TBS containing 0,1 % Tween at RT and incubated over night with different primary antibodies: anti-αβ-Tubulin, anti-HDAC-1, anti-NF-κB2 (Cell Signaling) and anti-RelA, anti-RelB, anti-NF-κB1, anti-I-κBα, anti-CYLD (SantaCruz) in 5% BSA or 5% milk powder over night at 4°C. After incubation with horseradish peroxidaseconjugated anti-rabbit or anti-mouse secondary antibody at RT (Amersham or Santa Cruz) proteins were detected using an ECL Plus kit (GE Healthcare).

3.3.11. Flow Cytometry

Single-cell suspensions were prepared from bone marrow, spleen, lymph node, peritoneal cavity and treated with Fc-Block (BioXcell) and surface or intracellular stained with monoclonal antibodies: B220 (BioL RA3-6B2), CD19 (BioL 6D5), CD23 (BD B3B4), CD90.2 (BioL 30-H12), IgM (eBio II/41), IgD (eBio 11-26c), CD5 (eBio 53-7.3), Bcl-2 (BioL BCL/10C4), Zap-70 (eBio 1E7.2), CD20 (eBio AISB12), CD22 (BioL OX-97), CD29 (eBio eBioHMb1-1), CD54 (BioL YN1/1.7.4), CD49d (eBio R1-2), CD45.1 (eBio A20), CD45.2 (eBio 104) For intranuclear staining FoxP3 staining kit was used with respective intracellular antibodies (eBioscience) according to the manufacturer's instructions. Dead cells were excluded with fixable viability dye V506 (eBioscience) or ef780 (eBioscience). Samples were acquired on FACSCanto II and FACS Aria (BD) machines, and analyzed with FlowJo Version 8.87.

3.3.12. IGHV Gene Rearrangement Analysis

DNA was purified from FACS sorted $CD19^+/CD90^-/CD5^+$ or CD5- B cells from spleen or peritoneal cavity from mice that were diagnosed with CD5+ B cell lymphoproliferations. IGHV sequences were amplified by PCR, using forward primers that anneal to the framework region I of the mouse IGHV families: IGHVJ558a: CAGGTGCAGCTGAARCAGTCA; GTGAAGCCTGGAGGGTCCC; GAGGTGAAGCTKGYGGAGTCT; IGHVQ52: IGHV7183b: IGHV6/7: CAGATCCAGTTGGTRCAGTCT and reverse SAGGTCCAGCTGCAGCAGTCTGG; IGHV36-60/4: IGHV7183c: IGHVGam3.8: GAGGTGMAGCTTCYSGAGTC; GGCTTAGTGMAGCCTGGAGG; primer positioned downstream of the IGHJ4 segment IGHJ4int2: ACTATCCCTCCAGCCATAGG. Cycling conditions were the following: 1 cycle 95°C for 5 minutes, 35 cycles of 95°C for 30 sec, 61°C for 30 sec, 72°C for 2 min. PCR products were separated on a 1.5% agarose gel.

3.3.13. Histology

For histological analysis spleen, liver and lung were fixed by immersion in 4% paraformaldehyde (PFA) and subsequently embedded in paraffin, cut into 4 µm thick tissue sections and stained with hematoxylin and eosin (H&E). Sections of the experimental groups were blinded and reviewed by an expert. Blood and bone marrow smears were stained with Giemsa staining.

3.3.14. Quantification of Western Blots

For quantification of Western blots ImageJ was used.

3.3.15. Statistical analysis

All statistical differences are presented as mean +/- SD and statistical significance was determined by two-tailed t-test using Prism 5.

Table 10 Primer sequences for 8N-UID layout murine HTS library preparation

Vprimer	Sequence
IGHV1a	AGRTYCAGCTGCARCAGTCT
IGHV1b	AGGTCCAACTGCAGCAGCC
IGHV1c	TCAGTGAAGATGTCCTGCAAG
IGHV1d	AACTGGGTGAAGCAGAGGCCT
IGHV1e	AAGTTGTCCTGCACAGCTTCT
IGHV1f	AAGCTCAGCTGCAAGGCTTCT
IGHV2a	CCTCACAGAGCCTGTCCA
IGHV2b	CAGCCATCACAGACTCTGTCTC
IGHV3	GTGCAGCTTCAGGAGTCAG
IGHV4	GGAGGTGGCCTGGTGCAG
IGHV5a	AGCCTGGAGGGTCCCTGAA
IGHV5b	GCTTAGTGCAGCCTGGA
IGHV6	GAGGAGTCTGGAGGAGGCTT
IGHV7	TCTGGAGGAGGCTTGGTACA
IGHV8	CTGGGATATTGCAGCCCTCC
IGHV9	CAGTCTGGACCTGAGCTGAAG
IGHV10	GTGAGGTGCAGCTTGTTGAG
IGHV11	GAAGTGCAGCTGTTGGAGAC
IGHV12a	CCTGGTGAAACCCTCACAG
IGHV12b	GCTGTCATCAAGCCATCACAG
IGHV13	AGGCTTGGTGAGGCCTGGA
IGHV14	GAGGTTCAGCTGCAGCAGT
IGHV15	CAGGTTCACCTACAACAGTCTG
IGHV16	GTGCAGCTGGTGGAATCT
IGHCμ	AGACATTTGGGAAGGACTGAC
A	CCATCTCATCCCTGCGTGTCTCCGACTCAG
P1	CCTCTCTATGGGCAGTCGGTGAT

Example Layout:	
pA-UID-IGHV4	**GCGTGTCTCCGACTCAG**NNNNNNNNGGAGGTGGCCTGGTGCAG
pP1-UID-IGHCμ	**CTATGGGCAGTCGGTGAT**NNNNNNNNAGACATTTGGGAAGGACTGAC
	bold= partial A/P1 adapter
	italic=UID bases

3.4. Results

4.3.1. Increased number of CD5$^+$ B cells in mice overexpressing a natural splice variant of CYLD

To study the role of sCYLD on the distribution of B cell subsets in a B cell-specific manner, we have generated mice that express sCYLD and lack the expression of the (FL)-CYLD in B cells (sCYLD^{BOE} mice), as described previously (222). Interestingly, at three months of age FACS analysis revealed the presence of an expanded CD5$^+$ B cell population in peripheral blood (PB) (**Figure 16A**), and to a lower extent also in the lymph nodes (LN) and spleens compared to CYLD$^{full\text{-}KO}$ mice (which lack expression of all CYLD splice variants) or to wild type (WT) and CD19-Cre control mice. CD5 expression on B cells defines the B1a subset which is almost absent in the spleen and preferentially localizes in the pleural and peritoneal cavities (PerC) (233). FACS analysis of B cell subsets in the peritoneum revealed that the proportion of B1a, B1b and B2 B cells as well as the total cell count was not significantly changed in sCYLD^{BOE} mice compared to controls at this age **(Figure 16B**).

At 12 months of age, sCYLD^{BOE} mice displayed a significant expansion of CD5$^+$ B cells, reaching around 20% of all B cells in PB whereas in controls this population represented only about 2% (**Figure 16C, upper panel**). CD5$^+$ B cells were also significantly expanded in the LN and spleen of sCYLD^{BOE} mice (**Figure 16C, middle and lower panel**), resulting in splenomegaly (data not shown).

Thus, sCYLD overexpression drives the expansion of CD5$^+$ B cells leading to a B cell lymphoproliferative disorder characterized by the accumulation of mature B cells in the BM, PB and lymphoid tissues.

4.3.2. Loss of A20 expression in B cells accelerates CD5$^+$ B cell expansion

CD5 is not only a marker of B1a cells, but also expressed by B cell lymphoma and leukemia, such as in B cell chronic lymphocytic leukemia (B-CLL) cells (234). Since sCYLD overexpression in B cells failed to progress to overt CLL, we reinforced NF-κB signaling, by crossing sCYLD^{BOE} mice to mice with B cell-specific A20 deficiency (A20^{BKO}sCYLD^{BOE} mice). We used A20^{BKO} mice as a tool to achieve sustained activation of canonical NF-κB signaling that in CLL patients occurs through other mechanisms, including B cell receptor (BCR) or micro-environmental activation (146, 235). In addition, these mice on their own harbor a dramatic decrease of CD5$^+$ B1a cells in the peritoneal cavity and spleen (225).

Additional deletion of A20 in B cells doubled the percentage of CD5$^+$ B cells in the blood as evident in A20^{BKO}sCYLD^{BOE} mice at 3 months of age (**Figure 17A**). Importantly, this accumulation was also evident

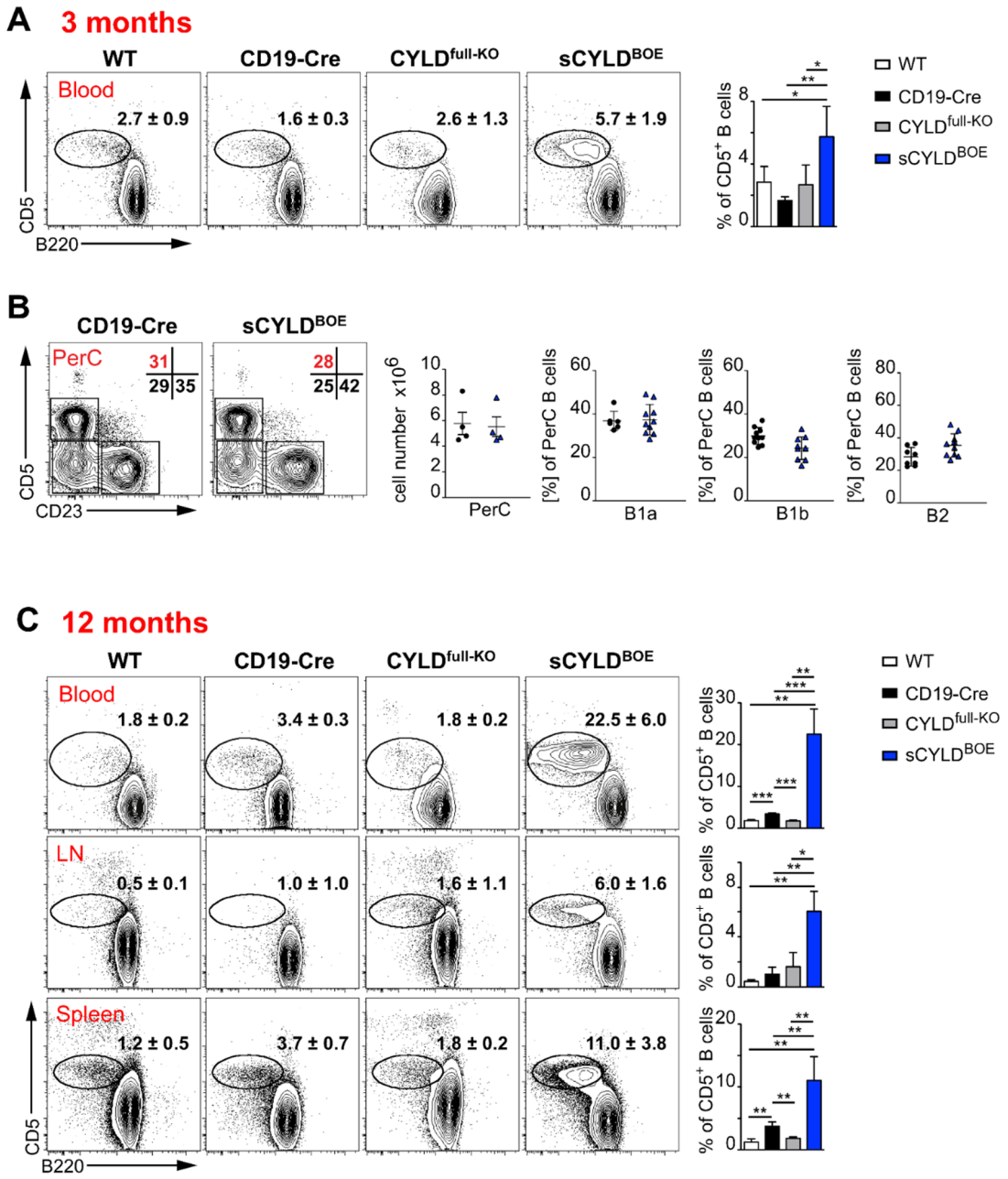


Figure 16 sCYLD expression leads to the expansion of CD5⁺ B cells. (A) Representative flow cytometry of B220^{low}CD5⁺ cells among CD19⁺ B cells in blood and right graph percentage of B220^{low}CD5⁺ B cells from mice with the indicated genotypes. sCYLD^{BOE} mice display a distinct CD5⁺ B cell population that is absent in the controls. (B) Representative flow cytometry of B cell subpopulations in the peritoneal cavity (PerC) of sCYLD^{BOE} mice and age-matched CD19-Cre controls. Numbers in the plots indicate the proportion of CD5⁺ (B1a), CD5⁻ (B1b) and CD5⁻ CD23⁺ (B2) B cells. Right panel: Total cell counts of PerC lymphocytes and percentages of B1a, B1b and B2 B cells in PerC of sCYLD^{BOE} mice compared to CD19-Cre controls. Data are shown as actual values (filled symbols) and as mean and standard deviation (± SEM). (C) Representative flow cytometry of B220^{low}CD5⁺ cells among CD19⁺ B cells in peripheral blood (PB, upper panel), lymph node (LN, middle panel), and spleen (lower panel). Numbers in the plots represent percentage (± SEM). Right graphs show percentages of CD5⁺ B cells from sCYLD^{BOE}, CYLD$^{full-KO}$ and the appropriate age-matched wild type (WT) and CD19-Cre controls. Data are represented as mean and standard deviation (± SD). Statistical significance was calculated using t-test. P values are indicated, * $p < 0.05$; ** $p < 0.005$; *** $p < 0.001$.

in mice with heterozygous sCYLD expression (A20^{BKO}sCYLD$^{BOE/WT}$ mice) (**Figures 17A lower panel and 17B**) demonstrating a role of sCYLD in regulating the pool of CD5⁺ B cells. CD5⁺ lymphocytes were not increased in the blood of A20^{BKO} mice and mice lacking both, A20 in B cells, and all variants of CYLD

protein (A20^{BKO}CYLD$^{full-KO}$) (**Figure 17A, B**). CD5^{+} B cells steadily increased with age in A20^{BKO}sCYLD$^{BOE/WT}$ and A20^{BKO}sCYLD^{BOE} mice reaching up to 80% in PB at the age of 12 months (**Figure 17B**). CD5^{+} B cells accumulated also in the BM, LN and spleen reaching up to 60% of all B cells in A20^{BKO}sCYLD^{BOE} mice (**Figure 17C**).

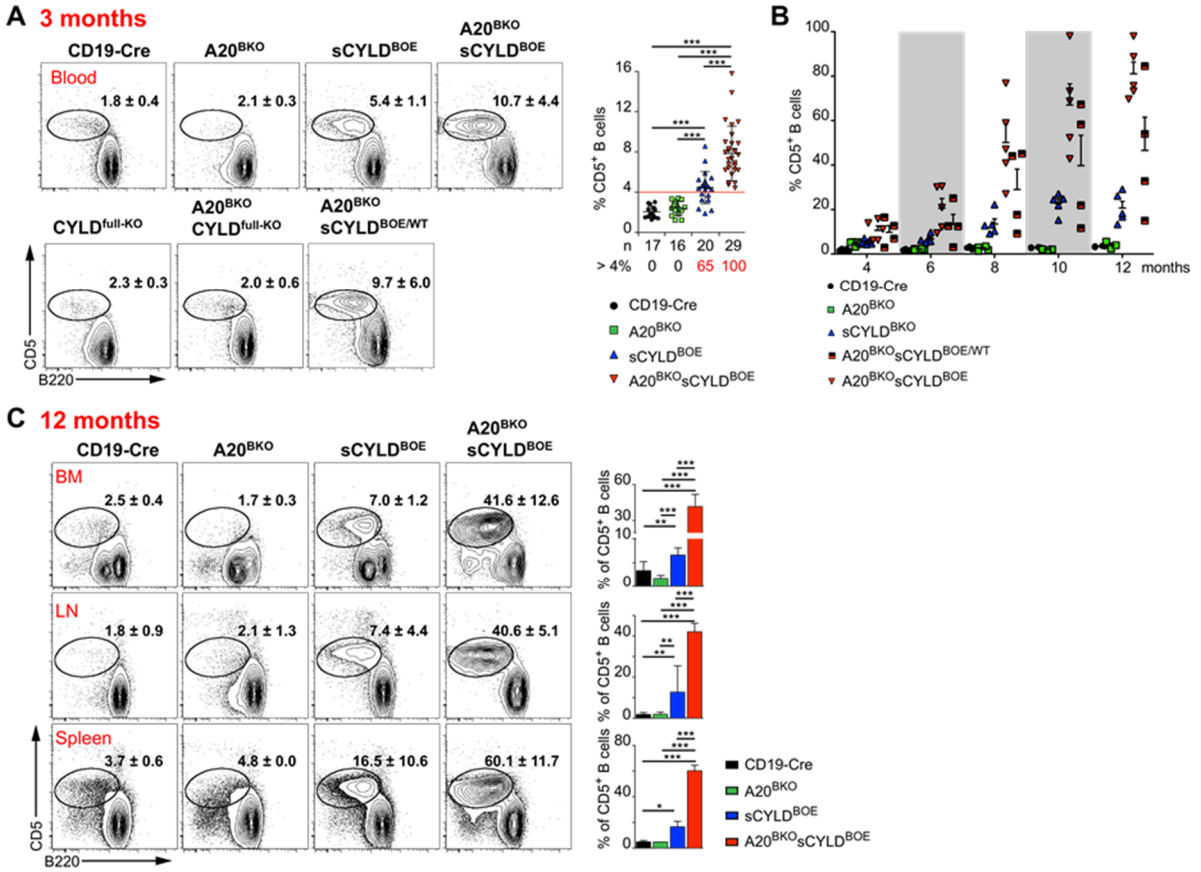


Figure 17 Accumulation of CD5^{+} B cells in A20^{BKO}sCYLD^{BOE} mice. (A) Representative flow cytometry of B220^{low}CD5^{+} B cells among CD19^{+} B cells in peripheral blood of the indicated mouse strains. Numbers in the plots indicate the proportion of CD5 expressing B cells ± SD. Right graph: Percentage of B220^{low}CD5^{+} B cells among CD19^{+} B cells in PB of the indicated mice represented as dots analyzed by flow cytometry; red line marks the upper threshold for normal percentages of B220^{low}CD5^{+} B cells in PB of control mice; n, number of mice analyzed; ">4%" shows the percent of tested mice that are considered positive for a beginning CD5^{+} lymphocytosis. Data are shown both as actual values (filled symbols) and as mean standard deviation (± SD). (B) Percentage of B220^{low}CD5^{+} B cells in PB analyzed by flow cytometry over time up to 12 months of age. Data of each mouse is shown as the actual value with the indicated dots representing the genotype. (C) Representative flow cytometry of B220^{low}CD5^{+} B cells in bone marrow (BM), lymph node (LN) and spleen of mice with the indicated genotypes at the age of 12 months. Numbers indicate the proportion of B220^{low}CD5^{+} B cells ± SD. Right graphs represent percentage of B220^{low}CD5^{+} B cells among the CD19^{+} B cells in lymphatic organs of the indicated mouse strains. Data are shown as bar graphs with mean ± SD. Statistical significance was calculated using t-test. P values are indicated, * $p < 0.05$; ** $p < 0.005$; *** $p < 0.001$.

While the number of CD5^{+} B1a cells in the peritoneal cavity was not increased in sCYLD mice at 3 months of age (**Figure 16B**) and virtually absent in A20^{BKO} mice (**Figures 18A**) this population was dramatically expanded at the age of 16 months in sCYLD mice as well as in A20^{BKO}sCYLD^{BOE} mice (97% in A20^{BKO}sCYLD^{BOE} mice compared to maximal 60% in the controls, **see Figure 18A**). At this age the total cell count of the peritoneal cavity was 10-20-fold increased in A20^{BKO}sCYLD^{BOE} compared to

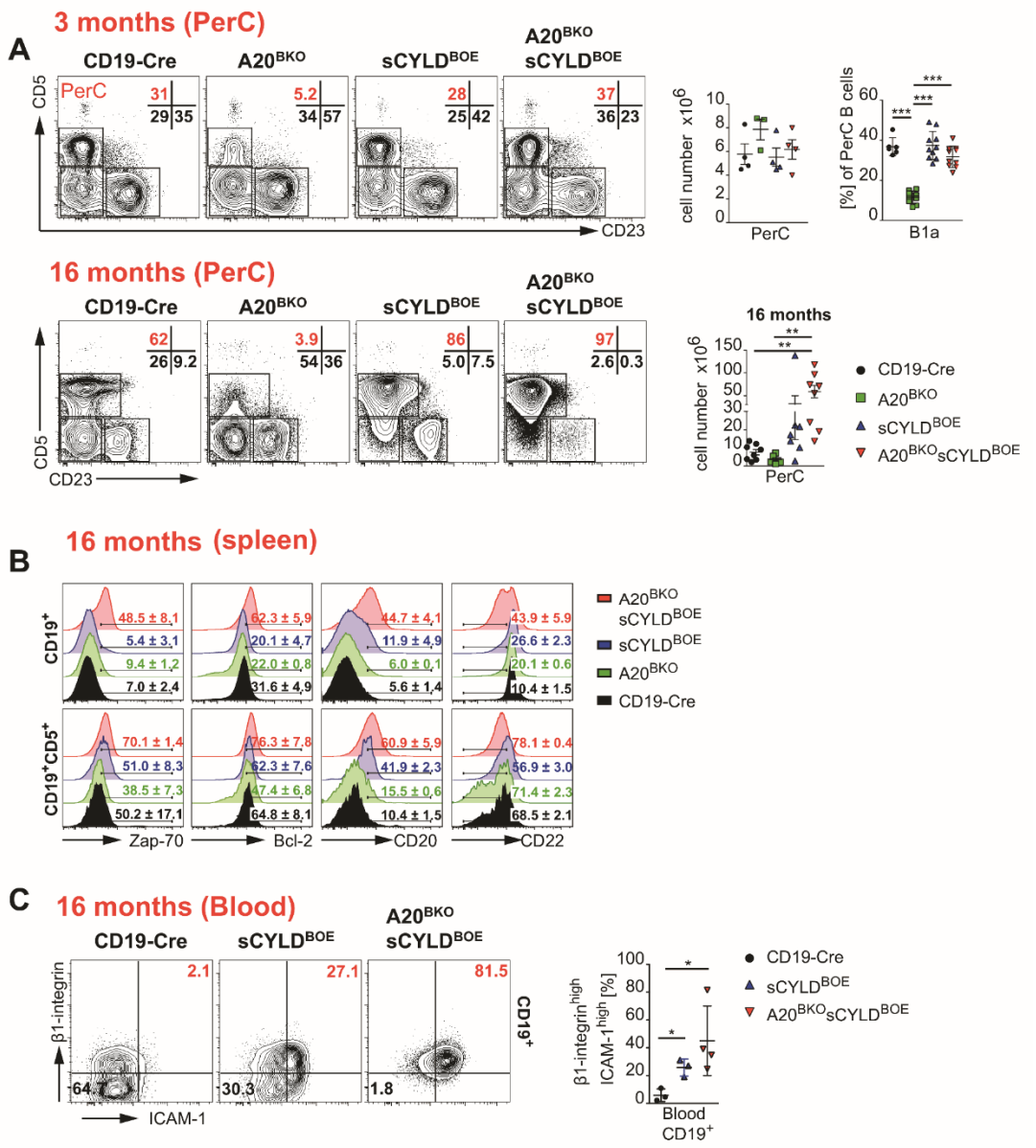


Figure 18 Dramatic expansion of CD5⁺ B cells in the peritoneal cavity of aged mice expressing sCYLD. (A, B) Representative flow cytometry of CD19⁺ B cell subpopulations in the peritoneal cavity (PerC) of the indicated mouse strains (A, upper panel) 3-month-old, and (B, lower panel) 16-month-old. Numbers in the plots indicate the percentage of B1a, B1b and B2 B cells. Graphs represent total cell number of PerC from mice with the indicated genotypes and percentage of B1a B cells (A) Representative flow cytometry of (top) CD19⁺ B cells and (bottom) CD19⁺CD5⁺ B cells for the expression level of the surface marker genes Zap-70, Bcl-2, CD20 and CD22. Numbers represent the mean percentage of the high expressing cell fraction with standard deviation (± SD) from 2 mice, 3 independent experiments. (C) Representative flow cytometry analysis of CD19⁺ B cells from blood of 16-month-old mice with the indicated genotypes for the expression of adhesion molecules using antibodies against ICAM-1 and β1-integrin. Red numbers in the quadrants represent the mean percentage of the high expressing cells ± SD. Right graph represents percentage of ICAM-1 and β1-integrin high expressing cells. Statistical significance was calculated using student´s t test. P values are indicated, * $p < 0.05$. Data are shown both as actual values (filled symbols) and as mean standard deviation (± SD).

controls (**Figure 18A**). sCYLD expression compensates for the decrease of B1a cells in A20^{BKO} mice. Thus, sCYLD expression but not FL-CYLD or A20 deletion drives the expansion of CD5⁺ B cells.

Next, we tested whether the cells of the expanded B cell population also expressed other markers typical for tumor cells found in CLL patients. Indeed, we found increased expression of ZAP-70, Bcl-2 and

CD20 in CD19^{+} B cells as well as in CD19^{+}CD5^{+} splenic B cells of aged A20^{BKO}sCYLD^{BOE} mice compared to control cells (**Figure 18B**). Moreover, expression of the BCR-inhibitory molecule CD22 was reduced, as previously reported for malignant cells of CLL patients **(Figure 18B**) (236). Homing to secondary lymphoid organs and to the BM is a central aspect of leukemic pathophysiology. Increased expression of adhesion molecules by malignant cells has been associated with their emigration and invasiveness into different tissues (237). Analysis of B cells isolated from blood of aged A20^{BKO}sCYLD^{BOE} mice showed increased expression of CD54 (ICAM-1) (**Figure 18C**). Thus, upregulation of CLL markers presumably leads to the evasiveness of these cells.

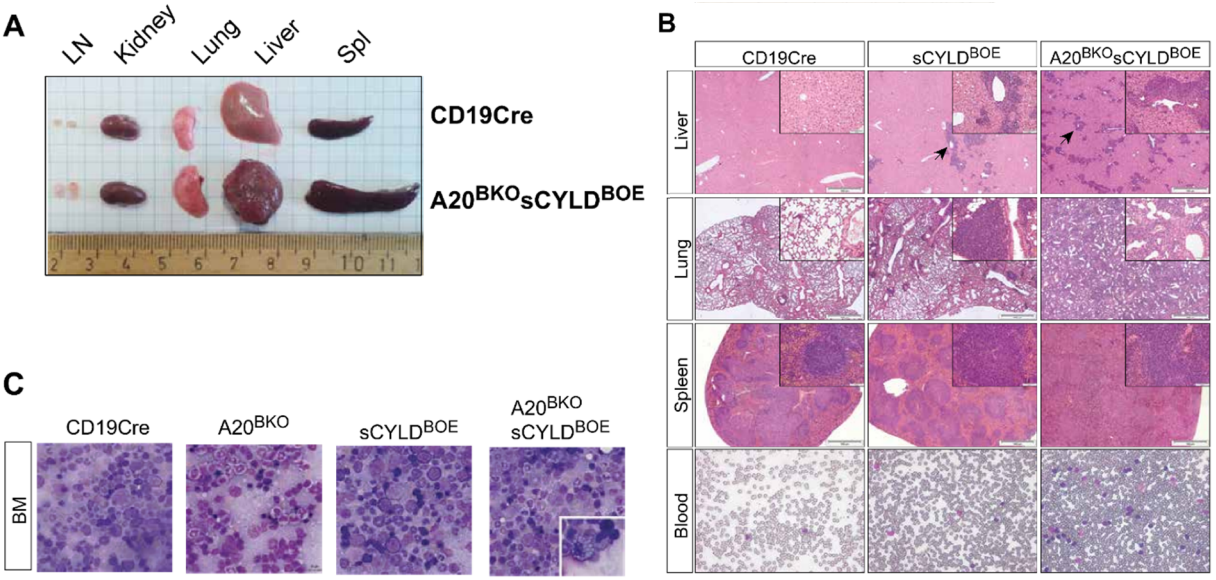


Figure 19 Infiltration of cells into non-lymphoid organs. (A) Comparison of lymph nodes (LNs), kidney, lung, liver and spleen (Spl) dissected from (top) aged CD19-Cre and (bottom) aged matched A20^{BKO}sCYLD^{BOE} mice (18-22-month-old). Ruler indicates the size of the organs (centimeters). Pictures shown are from one representative experiment out of 6. (B) Representative H&E staining of liver, lung, spleen sections and blood smears of 18-22-month-old A20^{BKO}sCYLD^{BOE} mice and age matched controls. Sections are shown at 10x (upper right corner) and 40x magnification, showing infiltrating cells. (C) Bone marrow (BM) smears of 12 months old A20^{BKO}sCYLD^{BOE} mice and age matched controls stained with May-Grünwald-Giemsa staining (x40 magnification) Lower right corner: Mott cell appearing in the BM of aged A20^{BKO}sCYLD^{BOE} mice.

4.3.3. Extensive cell infiltration into non-lymphoid organs in mice with sCYLD expression

The LNs and spleens of A20^{BKO}sCYLD^{BOE} mice increased with age (**Figure 19A**). Moreover, we found that the splenic architecture of mice overexpressing sCYLD was disorganized and revealed a distorted white and red pulp with irregular lymphoid follicles (**Figure 19B**). Importantly, also non-immune organs, including the lung and liver were enlarged, recapitulating hallmarks of human CLL (**Figure 19B**). Furthermore, lung and liver were massively infiltrated by lymphocytes in both strains of mice overexpressing sCYLD but not in control animals (**Figure 19B**). PB and BM smears displayed increased

abundance of mature lymphocytes and occasionally Mott cells which are associated with various pathological conditions, such as lymphoma and multiple myeloma (**Figure 19B, C**). Overall, 100% of the analyzed aged A20^{BKO}sCYLD^{BOE} mice developed B cell-lymphoproliferation. Collectively, we demonstrate that sCYLD overexpression combined with B cell-specific A20 deletion leads to a dramatic accumulation of CD5^{+} B cells and cell infiltration into non-lymphoid tissues similar to what has been observed in human CLL.

4.3.4. Clonal B cell expansion in A20^{BKO}sCYLD^{BOE} mice

The increase in CD5^{+} B cell numbers seen in A20^{BKO}sCYLD^{BOE} mice could result from either the outgrowth of neoplastic B cell clones or a gradual accumulation of clonal B cells. To examine clonality we performed PCR from sorted peritoneal cavity and splenic CD5^{+} B cells of aged (12 and 22 months) A20^{BKO}sCYLD^{BOE} and control mice. This analysis showed an impaired IGHV gene usage and a reduced diversity of IGH VDJ rearrangements (**Figure 20A, B, C**).

For a more detailed analysis, we sequenced the splenic IGHV repertoire of young (3 months) and aged (22 months) A20^{BKO}sCYLD^{BOE} mice along with the corresponding controls. We used Eμ-TCL1 mice, which express a transgene for human T-cell leukemia 1 (TCL1) under the control of the IGHV promoter and the Eμ enhancer as a positive control for murine CLL (238).

Skewing of the IGHV repertoires can be characterized by their Gini coefficient (239). Coefficients were determined for the cluster size distribution (cluster Gini coefficient) and the vertex size distribution (vertex Gini coefficient) of each sample. Large cluster Gini coefficients indicate ongoing clonal expansion with somatic mutations, whereas large vertex Gini coefficients indicate large expansion of a single clone. All young animals, including CD19-Cre control mice exhibited normal Gini coefficients (**Figure 21, box c**) and similar networks of expanding clones (**Figure 21, upper panel**). In contrast, aged mice separated into two distinct groups. One group (**Figure 21, box b**) consisted of all CD19-Cre control mice and one sCYLD^{BOE} mouse, all of which exhibited skewed repertoires characterized by elevated vertex Gini coefficients and limited clonal expansion (**Figure 21, lower panel**). This clearly shows that these mice have large numbers of similar or identical B cells indicating extensive proliferation. IGHV sequences that are less than 98% homologous to germline are considered to have undergone somatic hypermutation, therefore a mutated status is assigned to CLL clones when it displays more than 2% deviation from the germline IGHV sequence (240). In-depth sequence analyses revealed elevated mutation rates in the

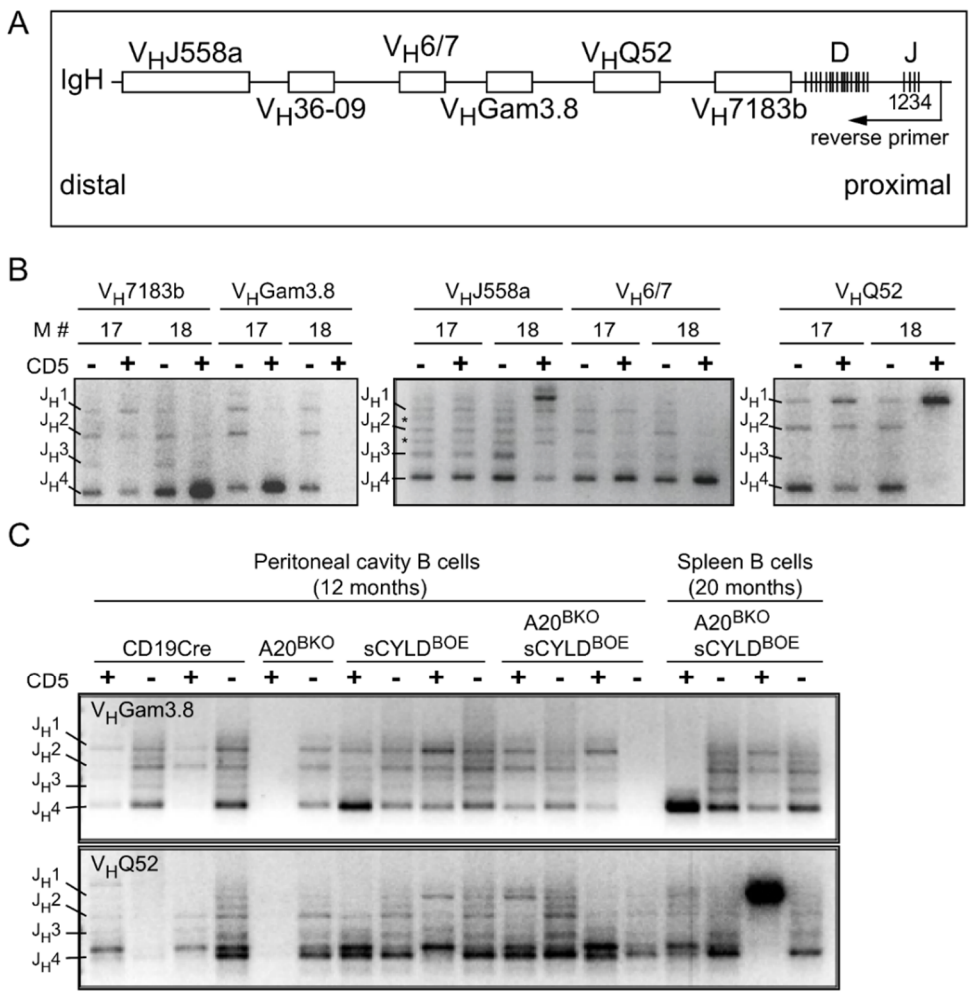


Figure 20 IGH VDJ recombination analysis of CD5^{+}B220^{low} and CD5^{-} B2 B cells. (A) Schematic overview of the genomic locus for IGHV gene families and IGHD and IGHJ genes. Primer locations for the IGH VDJ recombination PCR are indicated. (B) IGH VDJ recombination PCR of FACS sorted splenic CD5^{+} and CD5^{-} B cells isolated from 20 months old A20^{BKO}sCYLD^{BOE} mice. Mouse numbers and used forward primers for the respective IGHV gene families are indicated. PCR products with the respective IGHJ gene expression are marked on the left side. * indicates unspecific amplified PCR products. (C) IGH VDJ recombination PCR of FACS sorted PerC (left side) and splenic (right side) CD5^{+} and CD5^{-} B cells isolated from 12 and 20 months old A20^{BKO}sCYLD^{BOE} mice and age matched controls. Mouse numbers and forward primers used for the respective IGHV gene families are indicated. PCR products with the respective IGHJ gene expression are marked on the left side. * indicates unspecific amplified PCR products. No PCR products were obtained for CD5^{+} B cells in the lane of A20^{BKO} sample, due to the lack of these cells. A20^{BKO}sCYLD^{BOE} mice lack CD5^{-} B cells.

IGHV regions of B cells from CD19-Cre mice (**Figure 21C**) probably emanating from germinal center reactions. Hence, we conclude that the mutations in the CD19-Cre B cells are attributed to normal antigen responses. All old A20^{BKO}sCYLD^{BOE} animals clustered in the second group (**Figure 21, box a**) exhibiting highly skewed IGHV repertoires, with vertex Gini coefficients ranging from 0.85 to 1 and resulting from large vertices with only few direct relatives to the main vertex (**Figure 21, lower panel**). B cells from Eμ-TCL1 mice also clustered in this group with vertex plots similarly to aged A20^{BKO}sCYLD^{BOE} mice (**Figure 21, lower panel**). Interestingly, also two out of three sCYLD^{BOE} mice clustered in this group indicating that expression of sCYLD drives expansion of unmutated B cell clones.

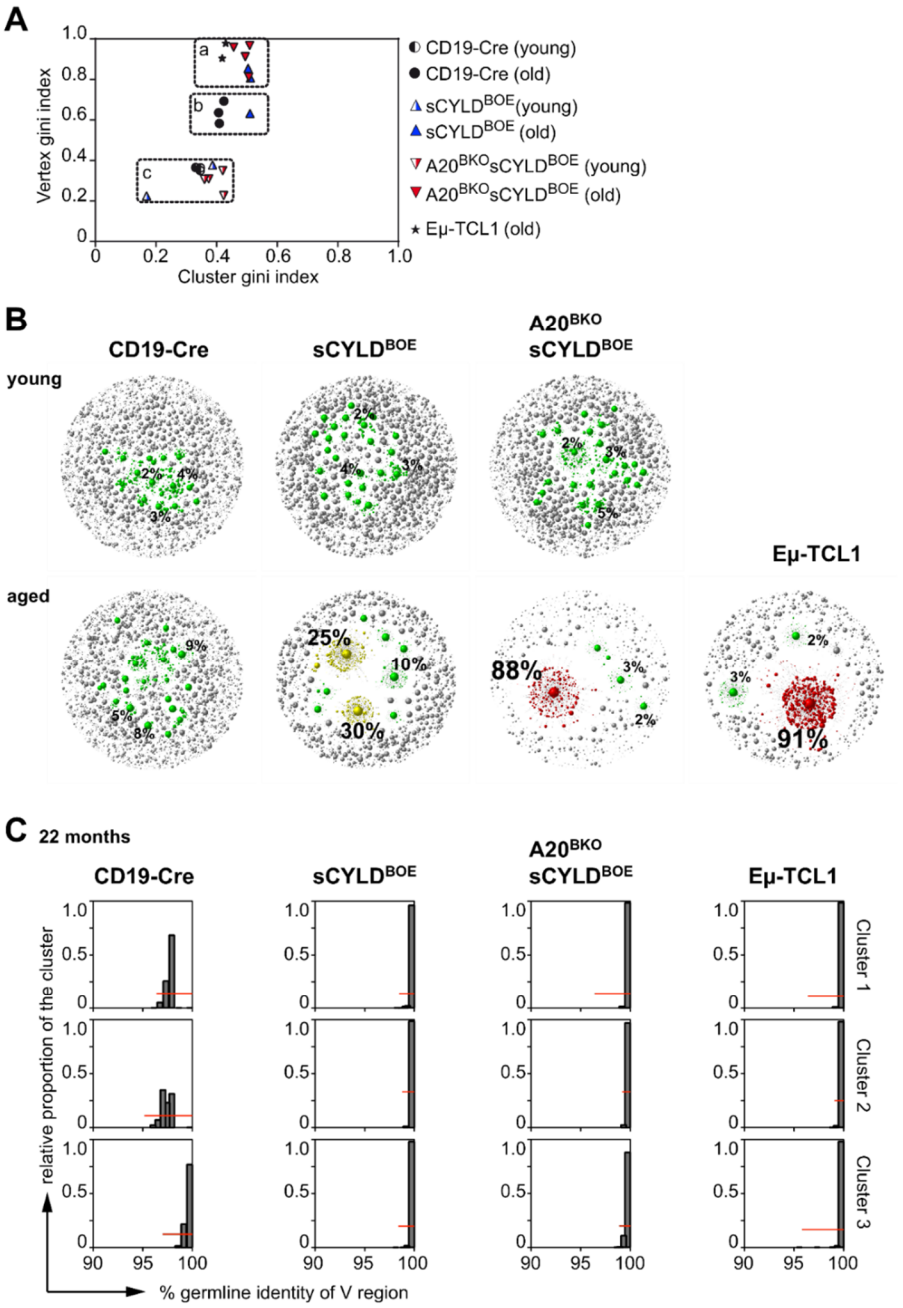


Figure 21 sCYLD^{BOE} cells engraft and outgrow wild type host cells. (A) Schematic overview of the experimental setup of the cell transfer experiment. Splenocytes were isolated from 12 months old mice with the indicated genotypes and transferred intravenously (i.v.) into mildly irradiated CD45.1+ wild type mice. (B) Percentage of expanded CD45.2+/CD19+ B cells in CD45.1+ host mice in blood, bone marrow (BM), peritoneal cavity (PerC) and spleen (Spl) 8 months after transfer. Data of each mouse is shown as the actual value with the indicated dots depending on the genotype together with mean ± SD of each experimental group. P-values are indicated, * $p < 0.05$; ** $p < 0.005$; *** $p < 0.001$ (2-tailed unpaired t-test). (C) Total cell numbers of CD45.1+ and CD45.2+ B cells of indicated transferred experimental groups in lymphatic organs of host mice 8 months after transfer. Spleen cells isolated from A20^{BKO}sCYLD^{BOE} mice and sCYLD^{BOE} mice replaced CD45.1+ wild type host cells and expanded in the PerC and spleen.

In support of these findings, we found that B cells of all aged A20^{BKO}sCYLD^{BOE} mice had a highly skewed IGHV region usage. The large CLL-like clones in these mice were all derived from IGHV 1-55 or IGHV 1-52, which were also detected in expanded clones of aged Eμ-TCL1 mice (data not shown).

Furthermore, also one of the sCYLD^{BOE} mice contained a larger B cell clone that utilized the IGHV 1-52 dsgment. Thus, sCYLD overexpression drives the expansion of B cells with unmutated IGHV gene rearrangements.

4.3.5. A20^{BKO}sCYLD^{BOE} CD5^{+} B cells expand in wild type hosts

Murine CLL cells as well as CD5^{+} B1a cells were shown to survive and proliferate when transferred to another host (241, 242). To test whether sCYLD expressing B cells are transferable and expand in healthy wild type recipients, we performed adoptive transfer experiments of splenic cells from the different experimental groups into congenic CD45.1 C57BL/6 mice (**see scheme Figure 22A**).

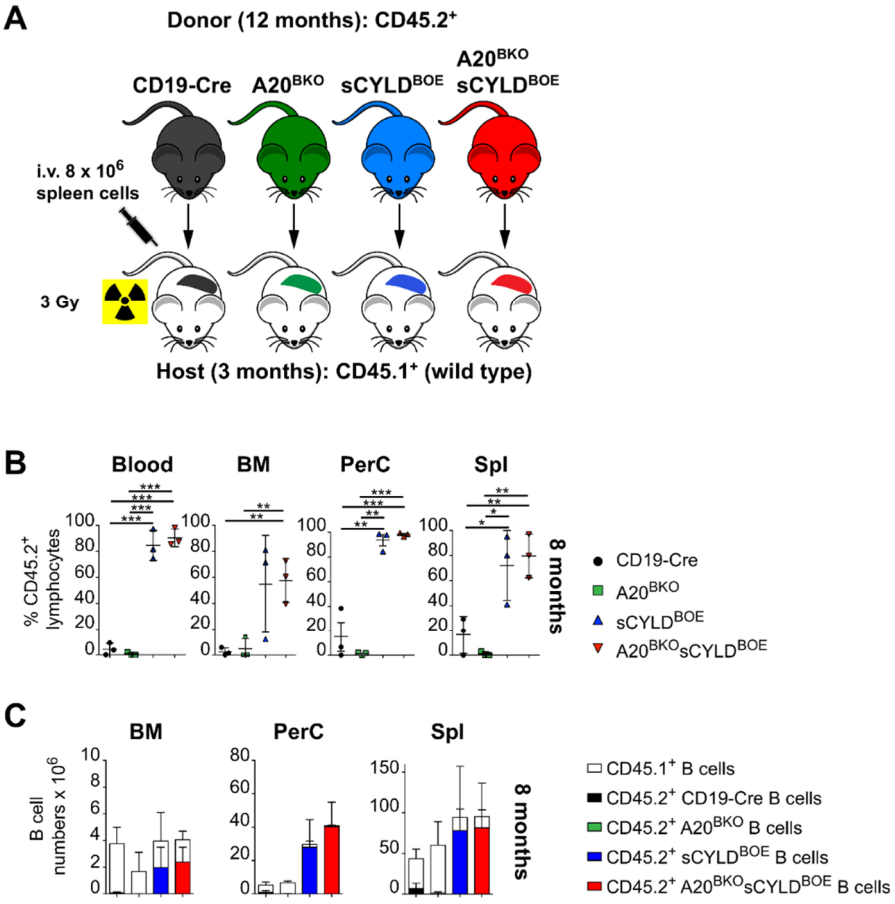


Figure 22 HTS IG repertoire analysis of CLL mouse models and controls. (A) Cluster Gini coefficients plotted against vertex Gini coefficients of samples from 3 months old ("young") and 22 months ("old") CD19-Cre (n=3), sCYLD^{BOE} (n=3), A20^{BKO}sCYLD^{BOE} (n=3), Eμ-TCL1 (n=2) mice. The 3 dashed boxes group mice with normal (diverse) repertoires (a), repertoires with moderate (B) and repertoires with strong (C) monoclonal expansion (single, large vertex clusters). (D) Representative vertex-cluster networks (log scale) for CD19-Cre, sCYLD^{BOE}, A20^{BKO}sCYLD^{BOE} and Eμ-TCL1 mice. The cluster network of the A20^{BKO}sCYLD^{BOE} repertoire shows one large clone comprising 88% of the sequences. In the Eμ-TCL1 sample, the largest clone comprises 91% of all sequences. In the sCYLD^{BOE} network, several clones have undergone expansion. The networks of CD19-Cre mice show multiple small clones of similar size. Clusters are color-coded by their proportion of the total repertoire (<1% = grey; >1% - <15% = green; >15 - 85% = yellow; >85% = red). (C) IGHV region identity of the 3 largest IGH sequence clusters of the 22 months old mice expressed in terms of % IGHV region identity compared to germline (as derived from IMGT High-V Quest output). Red lines indicate the total range of the mutation rate.

Six weeks after transfer, we detected a distinct population of donor-derived $CD5^+$ B cells in the blood of hosts that received either $sCYLD^{BOE}$ or $A20^{BKO}sCYLD^{BOE}$ cells which significantly increased over time (data not shown). Analysis of peripheral organs showed that in the peritoneal cavity, host B cells were completely outgrown by sCYLD as well as$A20^{BKO}sCYLD^{BOE}$ donor $CD5^+$ B cells (79% - 98%) five months post transfer. Eight months post transfer all host mice transplanted with spleen cells from $sCYLD^{BOE}$ or $A20^{BKO}sCYLD^{BOE}$ mice showed a dramatic expansion of the $CD5^+$ B cells also in the blood, bone marrow (BM), peritoneal cavity (PerC) and spleen (**Figure 22B, C**). Collectively, these data demonstrate that B cells from mice that overexpress sCYLD outcompete B cells of host mice similar to what was shown for CLL cells.

4.3.6. Enhanced NF-κB activation drives clonal $CD5^+$ B cell accumulation

To understand the molecular mechanism leading to the expansion of $CD5^+$ B cells in $A20^{BKO}sCYLD^{BOE}$ mice we performed global gene expression profiling from purified B cells of the indicated genotypes at the age of 8 weeks. We found that in B cells from $A20^{BKO}sCYLD^{BOE}$ mice 35 genes involved in the NF-κB pathway were differentially regulated compared to CD19-Cre control B cells (**Figure 23A**). We verified that IκBa and RelB were significantly upregulated on RNA level (**Figure 23B**) whereas on protein level IκBα was constantly degraded in unstimulated B cells isolated from $A20^{BKO}sCYLD^{BOE}$ mice compared to control B cells. Furthermore, in naive B cells sCYLD expression leads to an increase of cytoplasmic and nuclearp100 (NF-κB2) and RelB (**Figure 23D**), both substrates of the alternative NF-κB pathway and transcriptional targets of the canonical NF-κB pathway (243, 244). However, in agreement with our previous observations (222), the processing of p100 to p52 was not enhanced but rather decreased in B cells of mice overexpressing sCYLD (**Figure 23C**). In addition, naïve B cells isolated from these mice showed significantly increased spontaneous translocation of RelA to the nucleus probably as a result of constant IκBα degradation (**Figure 23C**). Taken together, these data show that expression of sCYLD results in overactivation of the canonical NF-κB pathway in naïve B cells. Moreover, sustained canonical NF-κB signaling in sCYLD expressing B cells was associated with the expression of pro-survival and proliferation genes including XIAP, survivin, cyclin D2 and p53 (**Figure 23B**), thereby providing the $A20^{BKO}sCYLD^{BOE}$ B cells excessive pro-survival signals that presumably drive the expansion of $CD5^+$ B cells. RelB expression was shown to rescue CLL cells from apoptosis (245). Consistently, B cells from $sCYLD^{BOE}$ and $A20^{BKO}sCYLD^{BOE}$ mice that comprise higher

RelB levels, exhibited a survival advantage compared to A20^{BKO} and control B cells *ex vivo* (**Figure 24A**).

To examine whether the increase of B cells was also due to increased proliferation, we fluorescence labeled B cells isolated from the different mice and tested their proliferative capacity. Here, we found that B cell-specific deletion of A20 resulted in increased proliferation when stimulated with anti-CD40 (**Figure 24B**). This proliferative advantage was irrespective of sCYLD expression, which failed to contribute to excessive proliferation (**Figure 24B**). To verify these findings *in vivo,* we tested B cells for

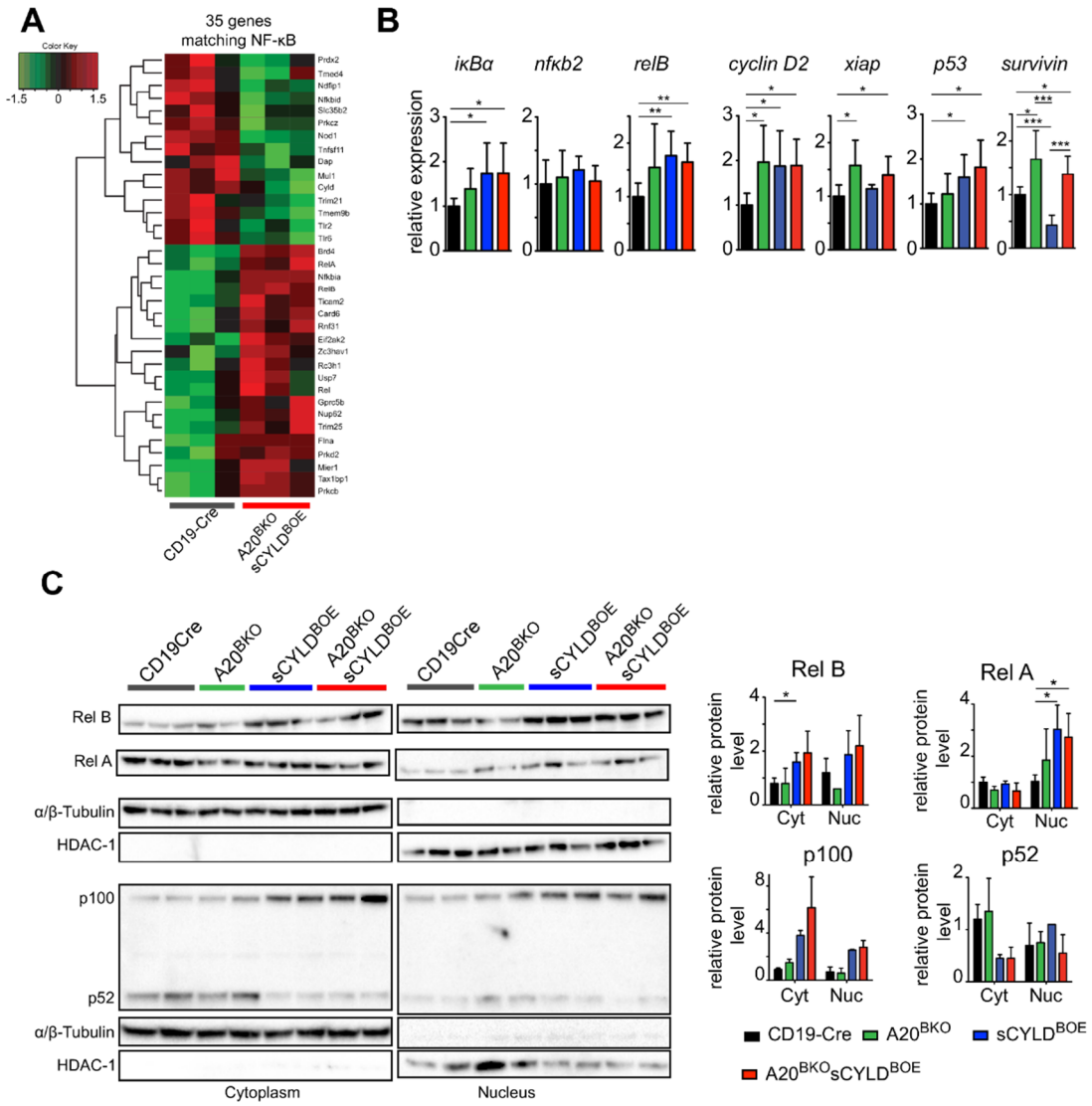


Figure 23 Increased canonical NF-κB activation synergizes with sCYLD expression in the development of CLL. (A) Heat map of differently expressed NF-κB matching transcripts in splenic B cells sorted from A20^{BKO}sCYLD^{BOE} and CD19-Cre control mice at the age of 8-weeks (n=3), assessed with a cut-off of a change in expression of 1.5-fold change and p-value = 0.5. (B) Relative expression of NF-κB transcription factors and relevant NF-κB target genes from MACS purified splenic B cells by RT-PCR analysis of the indicated genotypes. Gene expression levels were normalized to HPRT expression. (C) Western blot analysis of cytoplasmic (left panel) and nuclear protein extracts (right panel) of MACS purified naïve B cells of mice with the indicated genotypes using the indicated antibodies. α/β-Tubulin and HDAC-1 were used as loading controls. Right graphs represent quantification of cytoplasmic und nuclear location of NF-κB transcription factors from indicated genotypes.

incorporation of BrdU. We noted a higher percentage of BrdU positive B cells in the spleens of A20^{BKO} and A20^{BKO}sCYLD^{BOE} mice, corroborating the *in vitro* data (**Figure 24C**). Interestingly, in peritoneal cavity B cells, BrdU incorporation was the highest in A20^{BKO}sCYLD^{BOE} mice, probably representing the site of CD5^{+} B cell expansion (**Figure 24C**). Together, the increase in canonical NF-κB activation, driven by A20 deletion, results in higher proliferation while increased survival is due to sCYLD overexpression which in combination most likely contributes to a CLL-like disease.

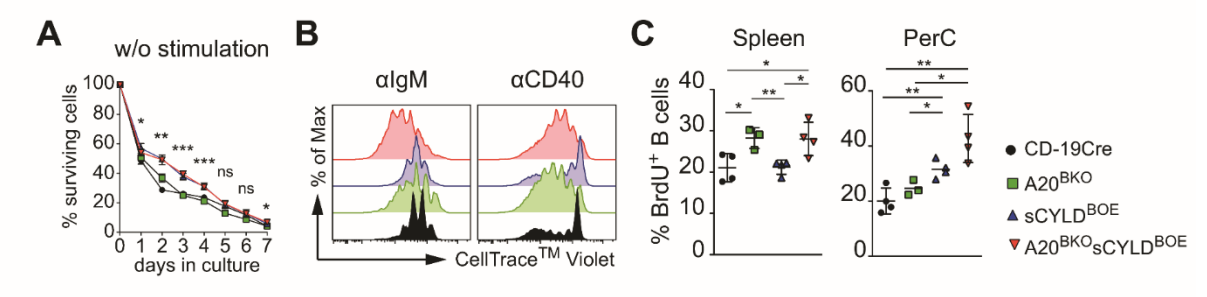


Figure 24. Increased proliferation and survival of A20^{BKO}sCYLD^{BOE} B cells. (A) MACS purified B cells of the indicated genotypes were cultured without stimulation. Percentage of living cells was determined daily by counting and FACS analysis. Each time point represents the mean of 5 different samples per experimental group ± SD. (B) Assessment of proliferation by violet cell tracer dilution assay. Histogram shows violet cell tracer dilution in MACS purified B cells from LN after 3 days in culture with the indicated stimuli. A20^{BKO}sCYLD^{BOE} B cells display increased proliferation rate after anti-IgM or anti-CD40 stimulation compared to CD19-Cre control B cells. Data are representatives of at least 3 independent experiments. (C) *In vivo* BrdU incorporation. Mice of indicated genotypes were fed with BrdU drinking water for 12 days. BrdU uptake was measured by flow cytometry. B cells were stained with CD19 antibody and an intracellular anti-BrdU antibody. Shown are percentages of BrdU^{+} B cells for each mouse in spleen and peritoneal cavity.

4.3.7. sCYLD expression in human CLL patients

We found that complete CYLD deficiency leaves CD5^{+} B cells unaffected whereas heterozygous expression of full length CYLD and sCYLD is sufficient to drive the accumulation of these cells. To address whether sCYLD expression is associated with human CLL we performed PCR of cDNA derived from CLL patient samples using primers located in the exons adjacent to exons 6-8, which are absent in the murine sCYLD. We detected, in addition to the band resembling FL-CYLD a band of 700 bp, representing human sCYLD. To verify that, we indeed detected sCYLD, we sequenced the two predominant products. Sequence analysis revealed that the 1240bp fragment resembles FL-CYLD product whereas the shorter band represents CYLD lacking the exons containing the binding sites for TRAF2 and NEMO equivalent to mouse sCYLD (**Figure 25A**). In total 32.7% (17 patients) of 52-tested CLL samples were positive for sCYLD expression as detected by PCR. Taken together, these results demonstrate that aberrant splicing of sCYLD occurs also in human CLL samples. However, the molecular mechanism causative for this aberrant splicing still needs to be investigated. Nevertheless,

sCYLD expression inhuman B-CLL patient cells implicates an important role for this truncated protein in human CLL pathology possibly representing a new risk factor for this disease. Overall, we show that alternative splicing of the tumor suppressor CYLD regulates the pool of $CD5^+$ B cells in mice through accelerated activation of NF-κB signaling. Reinforced activation of this pathway leads to the development of B1 cell-associated tumor formation in aging mice. sCYLD expression in human CLL samples implicates a role for this truncated protein also in human CLL pathology.

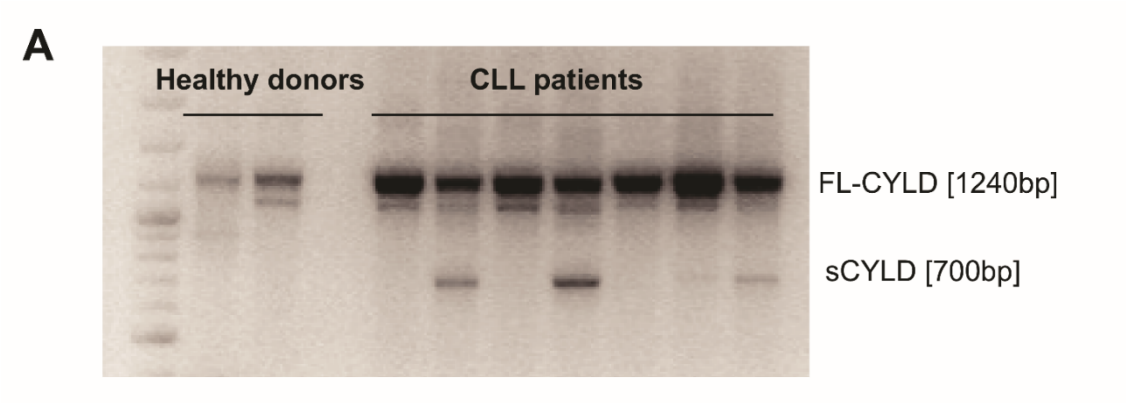


Figure 25. sCYLD expression in human CLL patient samples. (A) Representative PCR analysis of cDNA from B-CLL patients. PCR sizes of FL-CYLD and sCYLD are indicated.

3.5. Discussion

Accumulating evidence shows the importance of canonical NF-κB pathway activity in chronic lymphocytic leukemia (CLL). However, the underlying mechanisms were largely unknown (216, 246). Here we demonstrate that mice overexpressing a naturally occurring splice variant of the tumor suppressor CYLD (sCYLD) develop $CD5^+$ B cell lymphoproliferative disease. Heterozygous sCYLD expression but not FL-CYLD deletion is sufficient to drive the B cell expansion by modulating their survival rather than by affecting their proliferative capacity. We further show that the development of a full-blown CLL-like phenotype in sCYLD mice requires additional alteration such as enforced canonical NF-κB activation, which we achieved by inactivating A20 specifically in B cells. Although A20 was reported not to be involved in human CLL pathogenesis (227, 228), we used these mice as a tool to accelerate canonical NF-κB activation, mimicking activation of this pathway, which is normally driven by sustained BCR activation or the micro-environment (235, 247, 248). This enforced canonical NF-κB activation drives the progression to an indolent late onset CLL-like disease due to acquired proliferative skills of the $CD5^+$ B cell clones. These data are consistent with the TRAF2DN/Bcl2 CLL mouse model demonstrating that overcoming a certain threshold of NF-□B activation resulted in a B cell leukemia

resembling CLL (249). It has been shown that aging of C57BL/6 mice can result in B1 cell neoplasia. Age-related B-1 cell neoplasia in mice has been compared to human B-CLL, as both diseases develop as a result of slowly expanding $CD5^+$ B cell populations. Indeed, several mouse models for B-CLL show B1 cell expansion as a result of a survival benefit (250). $A20^{BKO}sCYLD^{BOE}$ mice show a gradual increase in $CD5^+$ B cell population in all lymphoid organs, which at final stages are also spreading to non-lymphoid organs owing to a survival as well as a proliferative advantage.

The biological factors that contribute to the development of CLL are not completely understood, but decreased susceptibility to apoptosis and deregulated proliferation driven by NF-κB was shown to play an important role in disease pathogenesis. For example, the transgenic expression of APRIL, a protein involved in NF-κB activation, resulted in expansion of $CD5^+$ B1 cells (251) and Eμ-TCL1 overexpression directly enhances NF-κB activity resulting in CLL pathogenesis (252). These observations underline a critical role for hyper-activated NF-κB signaling in CLL. A heterogeneous set of genetic lesions associated with CLL seems to constantly activate positive regulators or disrupt negative regulators of numerous signaling pathways converging in the activation of the NF-κB transcription complex (253–257). In particular, Akt activation, B cell receptor (BCR) signaling, CD40 ligation and B-cell activating factor of the TNF family (BAFF) and APRIL increase NF-κB activity and thereby enhance CLL cell survival (258–260). $A20^{BKO}sCYLD^{BOE}$ B cells showed increased levels of p100, a well-known NF-κB target. Interestingly it was shown that high protein levels of p100 protected CLL cells from apoptosis due to an elevated expression of the anti-apoptotic protein Bcl-2 (261). Another study using CLL patient samples showed that RelB activity together with RelA was associated with increased CLL cell survival in such a way that the strength of RelB activity correlated with the prognosis of the CLL patients (262). RelA was further shown to be an independent biomarker of clinical outcome in CLL (263). Thus, the survival advantage of B cells expressing sCYLD is at least partially driven by canonical NF-κB activation presumably via up-regulation of anti-apoptotic proteins. Moreover, the canonical NF-κB pathway also regulates expression of endogenous inhibitors of apoptosis (IAPs) such as survivin (264), which was also significantly increased in B cells of $A20^{BKO}sCYLD^{BOE}$ mice.

The processing of p100 to its active form p52 was not changed in naïve B cells from mice expressing sCYLD indicating that the non-canonical NF-κB pathway, which regulates p100 processing via NIK/IKK1, does not contribute to the accumulation of $CD5^+$ B cells. Thus, sCYLD expression regulates the number of $CD5^+$ B cells most likely due to an oncogenic function accelerating canonical NF-κB signaling. Deregulation of CYLD expression was previously shown to be involved in the development of

human B-CLL, since a fraction of CLL cases overexpressed the CYLD repressor lymphoid enhancer-binding factor 1 (LEF-1), decreasing CYLD levels. Consequently, the cell death program of the malignant cells in the CLL patients was impaired due to a constitutive ubiquitination of RIP1 (215). RIP1 is a target protein of CYLD and has been shown to act as a scaffold to either activate NF-κB or to initiate programmed cell death, depending on its ubiquitination status (265). Moreover, CYLD was shown to be considerably down regulated in human B-CLL cells compared to normal B cells (223). In our CLL-like model, $CD5^+$ B cell accumulation also occurs in the presence of FL-CYLD, suggesting that sCYLD mediates a dominant effect over FL-CYLD in CLL disease pathogenesis. Nevertheless, sCYLD expression in $CD5^+$ tumor formation in humans is persuasively supported by the detection of sCYLD message in a large fraction of CLL patients. Further analysis will be needed to unravel the molecular mechanism of sCYLD function in human and mouse CLL pathology.

Together, we show that sCYLD expression regulates the pool of $CD5^+$ B cells by hyper activation of the canonical NF-κB pathway. Reinforced NF-κB signaling leads to a late onset CLL-like disease in mice. We propose that our new CLL mouse model represents a remarkable model for human CLL to study NF-κB driven disease progression and as a preclinical model for testing new therapeutics for this disease.

Chapter 4

Robust sequencing of immunoglobulin heavy chain transcripts from Balb/C mice using single side unique molecular identifiers on an Ion Torrent PGM.

Jean-Philippe **BÜRCKERT**, William J. **FAISON**, Axel R.S.X. **DUBOIS**, Regina **SINNER**, Oliver **HUNEWALD**, Anke **WIENECKE-BALDACCHINO**, Anne **BRIEGER**, and Claude P. **MULLER**

Department of Infection and Immunity, Luxembourg Institute of Health / Laboratoire National de la Santé

[This chapter is an adapted version of the manuscript under revision in Oncotarget, manuscript #038998, submitted Aug 31, 2017. The latest version is available at biorXiv.org as preprint under the DOI: 10.1101/219568]

Authors' contributions: J.-P.B. designed research project, cultivated hybridomas, performed library preparation, developed and supervised development of bioinformatics approaches, performed data processing and analysis, interpreted data and wrote the manuscript. A.R.S.X.D designed research project and interpreted data. W.J.F. and O.H. supported and developed bioinformatics approaches and performed data processing. A.W.-B. developed and wrote the raw data processing bioinformatics pipeline. R.S. performed Ion Torrent PGM sequencing. A.B. designed research project, supervised work, assisted library preparation and hybridoma cultivation and interpreted data. C.P.M. supervised work, interpreted data and corrected the manuscript.

4.1. Summary

With the advent of high-throughput sequencing (HTS), profiling immunoglobulin (IG) repertoires has become an essential part of immunological research. Advances in sequencing technology now also allow the use of the Ion Torrent Personal Genome Machine (PGM) to cover the full length of IG mRNA transcripts. Nucleotide insertions and deletions (indels) are the dominant errors of the PGM sequencing platform and can critically falsify IG repertoire assessments. Here we use a set of artificially falsified murine IG heavy chain (IGH) sequences to benchmark the indel detection and correction measures of ImMunoGeneTics (IMGT) database, the most commonly used sequence alignment database for IG sequences. We could show that IMGT efficiently detects 98% of the artificially introduced indels through gene-segment frameshifts. The remaining undetected indels are either located at the beginning and end of the sequences or produce frameshifts that cancel each other out, e.g. with an insertion and deletion in close proximity. IMGTs indel correction algorithm corrects up to 87% of the tested insertions. Deletions and mixed indels result in mostly false sequences, as deleted nucleotides are not inferred by IMGT. The most important part of the IGH sequence, the complementary determining region 3 (CDR3) is covered by a conservative detection and culling algorithm of IMGT, returning 100% correct CDR3s for up to 3 insertions or 3 deletions. We further show that HTS datasets from a PGM sequencer can be highly accurate if combined with a tailored single side unique molecular identifiers (ssUID) library preparation and the appropriate data processing steps. In this regard, considering consensus sequences with at least two copies from datasets with UID families of minimum 3 reads results in correct IGH nucleotide and amino acid sequences with over 99% confidence and correct CDR3 amino acid sequences with over 99.9% confidence. The protocol and sample processing strategies described in this study will help to establish benchtop-scale sequencing of IG heavy chain transcripts in the field of IG repertoire research.

4.2. Introduction

The diversity of the immunoglobulin (IG) repertoire is the key feature of the adaptive immune system, enabling it to theoretically combat every possible antigen encountered during an individual's lifetime (1). With the development of high-throughput sequencing (HTS) it became possible to analyze the IG repertoire at high depth (64, 77, 156, 266–268). Studies, almost a decade ago, established Roche's 454 sequencer as the first tool of choice for exhaustive characterization of IG repertoires due to its superior read-length (269). More recently, Illumina's MiSeq and HiSeq sequencers as well as the Ion Torrent Personal Genome Machine (PGM, Thermo Fisher Scientific) provided an improved sequencing technologies which can reach across the full V(D)J nucleotide sequence span (68). The different technologies of the sequencers result each in their specific error-rates and -types (57, 269–275). Illumina's optical sequencing produces mostly nucleotide (nt) transversions and transitions, which can be corrected by building consensus sequences (56). The 454's pyrosequencing chemistry and the PGMs semiconductor technique mainly introduce homopolymer repeats resulting in insertions and deletions of bases, which can be corrected by gene segment-wise reference alignment (88).

Most sequencing approaches use IG isotype specific constant (C) region primers to translate IG heavy-chain (IGH) (m)RNA into cDNA, which are subsequently amplified using a set of IGHV region specific primers in a multiplex PCR approach. However, this can result in skewed repertoire read-outs due to biased PCR efficacy (57, 68, 276). In addition, sequencing errors can falsify somatic hypermutation profiles, IGH VDJ germline gene assignment and clonal grouping (68, 76). Unique identifiers (UID) which tag individual RNA molecules at cDNA transcription level have been used to obtain an unbiased view on the IG repertoire (277–280). This method also allows thorough error-correction by building consensus sequences, albeit at the cost of sequencing depth. In all cases, complex bioinformatic approaches are necessary to perform raw-read processing (281). Subsequent alignments to germline genes to assign IGH VDJ genes are usually conducted using the ImMunoGeneTics (IMGT) database, which applies an error correction algorithm for insertions and deletions in the process (187, 282).

After the initial proof-of-concept studies, the use of animal models to study the IG repertoire dynamics has been largely ignored (77, 268). One major factor being the lack of a suitable IGHV region primer set comparable to BIOMED-2, developed for the human IG repertoire (59). Yet, animal models offer advantages over human studies, as they are not limited to peripheral blood and have a lower B cell

diversity (283–286). As IMGT provides repertoires for various species, we chose to develop a method to profile the IG repertoire of Balb/C mice, one of the most commonly used animal models.

In the present study, the performance of the PGM sequencing platform together with the IMGT database for the assessment of murine IGH repertoires is evaluated. In this context, several novel aspects are examined: first, the IMGT database's indel detection and correction algorithm is benchmarked with a set of artificially falsified sequences. Second, a 16-nucleotide single side UID (ssUID) barcoding technique tailored to the PGM sequencing chemistry is provided together with a swift 1-day library preparation protocol. Third, the PGM's error-rate for sequencing murine IG transcripts with our barcoding strategy and customized data processing is determined.

4.3. Materials and Methods

4.3.1. RNA extraction

RNA was extracted with Trizol LS/chloroform (Thermo Fisher Scientific, Waltham, USA) method from seven monoclonal hybridoma cell lines (produced in house) with 10^6 cells each. DNA was digested using the DNAfree kit (Thermo Fisher Scientific), RNA was further purified using Agencourt® RNAclean XP beads (Analis, Suarlée, BE) and quantified on a NanoDrop® Spectrophotometer (ND1000, Isogen Life Science, De Meern, NL). RNA was either directly used for library preparation or stored at -80°C.

4.3.2. Reference sequences

Hybridoma cDNA transcripts were obtained using mouse constant region IgG primer (**Table 11**) in a Superscript III (Thermo Fisher Scientific) reverse transcription following the manufacturer's instructions for templates with high GC content. Transcripts were Sanger-sequenced (3100 Avant, Thermo Fisher Scientific) using constant region IgG and IGHV region primers (**Table 11**). Forward and reverse sequences were aligned and submitted to IMGT V-QUEST (http://www.imgt.org, (287)) to verify the nucleotide sequence and to translate into amino acids. These sequences were subsequently used as reference sequences in alignments and artificial error insertion experiments.

4.3.3. Datasets with artificial insertions and deletions

Artificial datasets were generated using the Biopieces indel_seq package (http://www.biopieces.org). For each of the original 7 hybridoma sequences, 2500 error-containing sequences were generated by combining 0-3 insertions and 0-3 deletions, obtaining a total of 37500 artificial sequences per hybridoma. For every set, indel-type and -position were determined by alignment to the original sequence to ensure homogenous error distributions. All artificial datasets were uploaded to IMGT HighV-QUEST and sorted by annotation: IMGT annotates correct sequences as productive. Sequences with a detected indel (frameshift, stop codon) are marked as “productive (see comment)” if the error can be corrected (referred to as “productive with detected errors”). Sequences with uncorrectable errors are classified as “unproductive”. If no fitting germline can be found sequences are marked as “unknown” or “no result” (referred to as “unknown/else”). The remaining indels on nucleotide level and amino acid changes were determined using the SeqAn library (288) in a custom-made C++ reference alignment program. For datasets with one insertion and one deletion (i1d1) the positions of the indels were determined by position-wise mismatch detection using a custom made Biopython (289) script. Upon detection, the nucleotide positions were returned and the process repeated with reverse complement sequences.

4.3.4. Library preparation and HTS

Approximately 100ng (as determined by Nanodrop) of total RNA per hybridoma was used for library preparation. We adapted the UID labeling method developed by Vollmers and colleagues (71) to the PGM sequencing platform (**Figure 26**). RNA was reverse transcribed using Superscript III reverse transcriptase, according to the manufacturer’s instructions, using multiplex identifiers (MID) and UID tagged mouse constant region (IGHγ) primers elongated by partial PGM sequencing adapter pA (**Table 11**). The MID tag allowed multiplexing of several samples on one sequencing chip. The UID tag consists of two times 8 random nucleotides separated by a “GATC” spacer (N8-GATC-N8). With this UID tag each RNA molecule targeted by the primer is uniquely labeled (see (71, 73) for detailed theoretical descriptions). The RT reaction mixtures were split into two equal second strand synthesis reactions using Phusion High-Fidelity DNA polymerase (NEB, Massachusetts, USA) with a mouse IGHV region primer mix (**Table 11**). The reaction conditions were as follows: 98°C 2min, 50°C 2min, 72°C 10 min in a single cycle reaction. Both reaction aliquots were combined and purified twice using Agencourt AMPure XP beads (Analis) in a 1:1 (v/v) ratio to remove primer traces. Libraries were subsequently amplified with a Q5 Hot Start High-Fidelity DNA polymerase (NEB) using the full-length Ion Torrent PGM

sequencing adapters A and P1 as primers (**Table 11**) with the following conditions: 98°C for 1min, 20 cycles of 98°C for 10s, 65°C for 20s, 72°C for 30 seconds. Final elongation was done at 72°C for 2 min. Amplified libraries were purified twice using equal volumes of AMPure XP beads. Quality of the libraries as well as size of the amplicon and concentrations were determined using Agilent 2100 Bioanalyzer (Agilent Technologies, Diegem, BE) with the High Sensitivity DNA Kit (Agilent Technologies). 10 libraries were pooled equimolar on an Ion 316™ Chip (Thermo Fisher Scientific) and sequenced on a PGM sequencer, with all quality trimming options disabled on the Torrent Suite™ v4.0.2

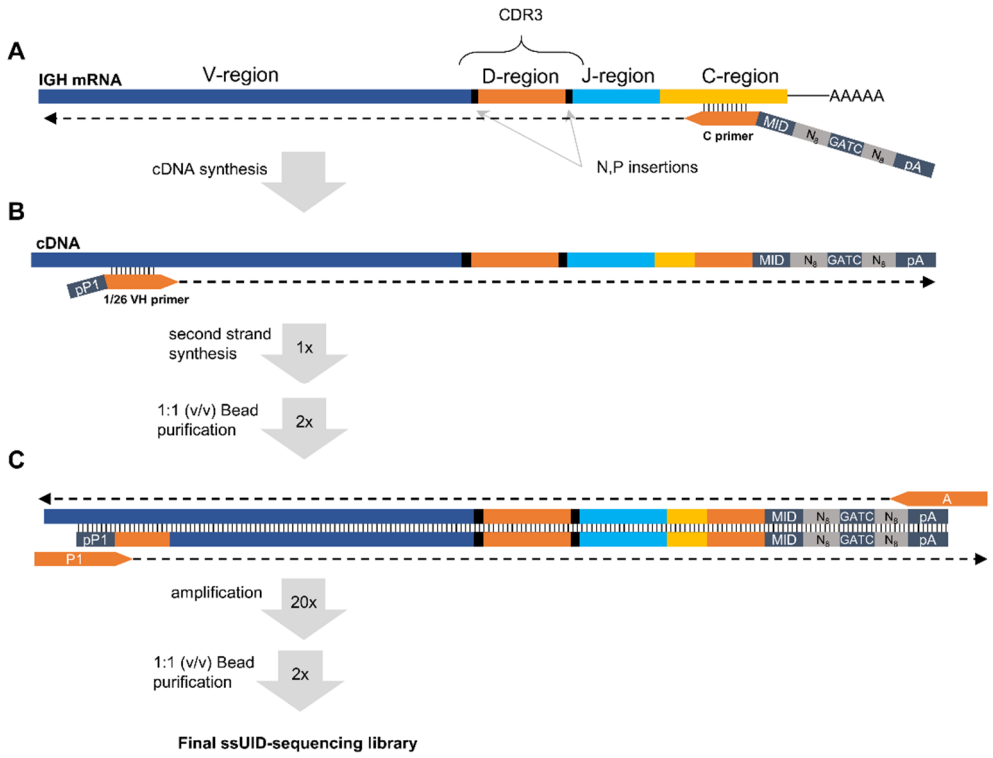


Figure 26 3-step PGM ssUID sequencing library preparation. (A) In a first step, purified mRNA is used in a Superscript III reverse transcription. The Primer for the reverse transcription is specific for the murine IGH C region and elongated by an MID for sample multiplexing as well as a UID consisting of 2 x 8 random nucleotides (N_8) separated by a 4-nucleotide spacer ('GATC'). The primer ends with the partial PGM sequencing adapter pA. (B) In the second step a mix of 26 IGHV region targeting primers (elongated by the partial PGM sequencing adapter pP1) is used in a single cycle PCR reaction to avoid amplification. The product of this reaction is purified twice with Agencourt AMPureXP beads to remove the IGHV primers from the reaction mixture. (C) In the final step, the purified reaction mixture is amplified using the full-length P1 and A adapters as primers in a 20 cycle PCR reaction. The product is as well purified twice to obtain the ssUID-tagged sequencing library.

4.3.5. Data processing pipeline for the HTS datasets

Untrimmed raw reads were demultiplexed by their MIDs, retaining only sequences containing the full UID primer sequence for further analysis, with no mismatches allowed. The UID sequence was extracted in relation to the starting position of the detected primer including the 'GATC' spacer and stored in the

sequence identifier. After clipping the MID, UID and constant region primer, the trimmed reads were quality controlled (80% of the bases Phred-like quality score above 20) and grouped into UID families. Using pagan-msa (230), a consensus sequence was generated for each UID-family containing more than 2 members. Subsequently, sequences were collapsed to unique reads, storing counts in the read identifier, and uploaded to IMGT for error detection, correction, annotation and translation into amino acids. Post-IMGT datasets were separated into four categories (“productive”, “productive with detected errors”, “unproductive” and “unknown/else”) and processed separately. Data processing was performed using custom-made Python scripts (Python v2.7) employed in a parallelizing bash wrapper script using gnu-parallel (290) and the Biopieces framework (http://www.biopieces.org/).

4.3.6. Graphs and statistics

All graphs and statistical analyses were performed using R base packages or GraphPad Prism 6. Average numbers are reported as mean ± standard deviation (SD) unless specified otherwise.

Table 11 Primer sequences for ssUID N_8-GATC-N_8 layout murine HTS library preparation

Vprimer	Sequence
IGHV1a	AGRTYCAGCTGCARCAGTCT
IGHV1b	AGGTCCAACTGCAGCAGCC
IGHV1c	TCAGTGAAGATGTCCTGCAAG
IGHV1d	AACTGGGTGAAGCAGAGGCCT
IGHV1e	AAGTTGTCCTGCACAGCTTCT
IGHV1f	AAGCTCAGCTGCAAGGCTTCT
IGHV2a	CCTCACAGAGCCTGTCCA
IGHV2b	CAGCCATCACAGACTCTGTCTC
IGHV3	GTGCAGCTTCAGGAGTCAG
IGHV4	GGAGGTGGCCTGGTGCAG
IGHV5a	AGCCTGGAGGGTCCCTGAA
IGHV5b	GCTTAGTGCAGCCTGGA
IGHV6	GAGGAGTCTGGAGGAGGCTT
IGHV7	TCTGGAGGAGGCTTGGTACA
IGHV8	CTGGGATATTGCAGCCCTCC
IGHV9	CAGTCTGGACCTGAGCTGAAG
IGHV10	GTGAGGTGCAGCTTGTTGAG
IGHV11	GAAGTGCAGCTGTTGGAGAC
IGHV12a	CCTGGTGAAACCCTCACAG
IGHV12b	GCTGTCATCAAGCCATCACAG
IGHV13	AGGCTTGGTGAGGCCTGGA
IGHV14	GAGGTTCAGCTGCAGCAGT
IGHV15	CAGGTTCACCTACAACAGTCTG
IGHV16	GTGCAGCTGGTGGAATCT
IGHCγ	GGCCAGTGGATAGACHGATG
A	CCATCTCATCCCTGCGTGTCTCCGACTCAG
P1	CCTCTCTATGGGCAGTCGGTGAT

Example Layout:	
pP1-IGHV4	**GCGTGTCTCCGACTCAG**GGAGGTGGCCTGGTGCAG
pA-N_8-GATC-N_8-IGHCγ	**CTATGGGCAGTCGGTGAT***NNNNNNNNGATCNNNNNNNN*AGACATTTGGGAAGGACTGAC
	bold= partial A/P1 adapter italic=16N ssUID

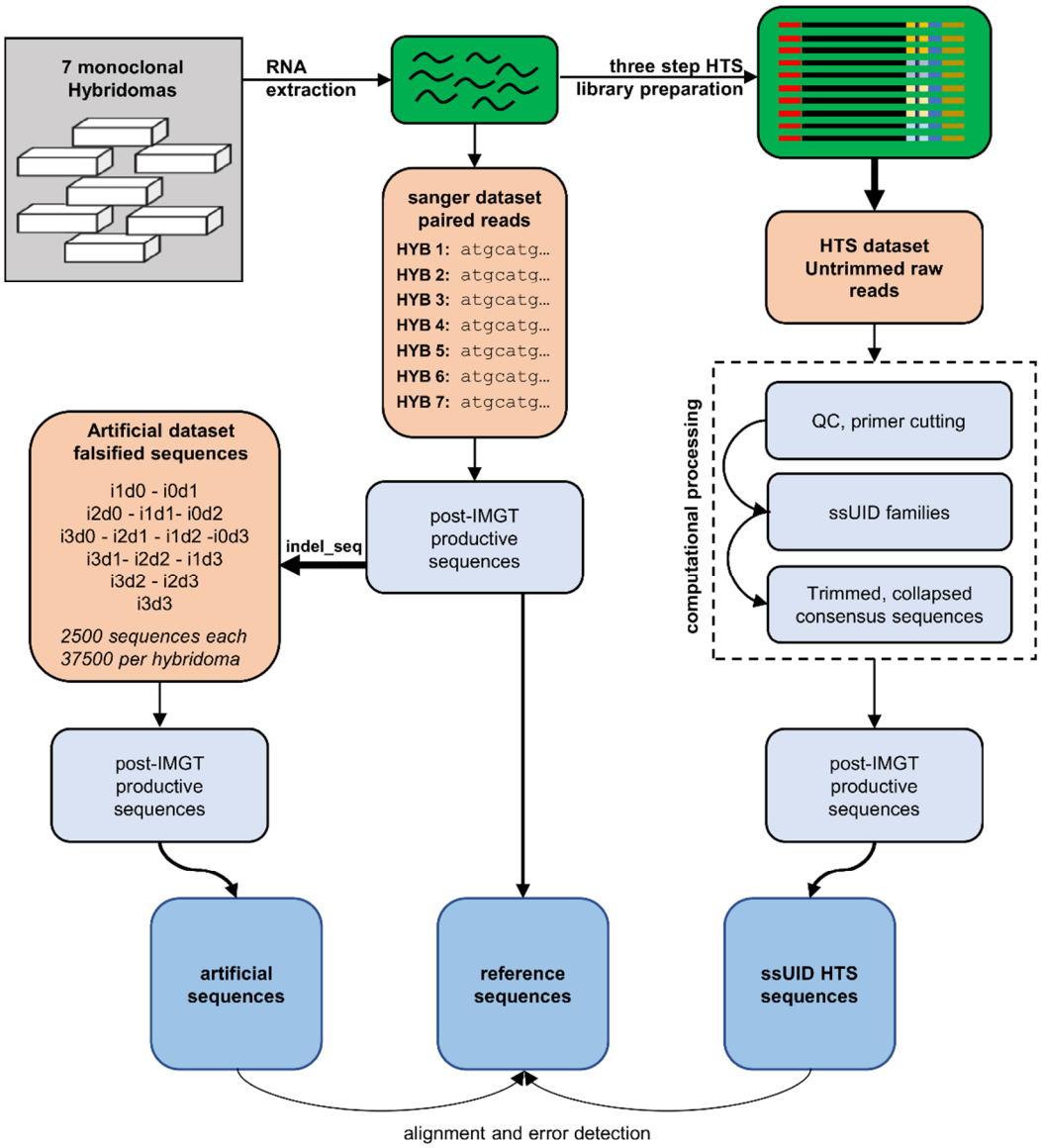


Figure 27 Study design and data processing sheet. RNA was extracted from 7 monoclonal hybridoma cell lines and reverse transcribed into cDNA. cDNA sequences were determined by Sanger sequencing and submitted to IMGT to determine reference sequences. Reference sequences were artificially falsified using the indel_seq program, introducing up to 3 insertions and 3 deletions. 2500 artificial sequences were generated for each permutation and hybridoma and processed by IMGT. Post-IMGT sequences were aligned to the references to determine error detection and correction. RNA was also used to generate high-throughput sequencing (HTS) libraries in a three-step library preparation protocol. Single side unique identifiers (ssUID) were introduced during reverse transcription to tag each RNA molecule individually. Libraries were sequenced on an Ion Torrent PGM sequencer with all quality trimming options disabled in the Torrent Suite software. Untrimmed raw sequences were processed with a custom-made bioinformatics pipeline generating consensus sequences per UID family. Collapsed consensus sequences were submitted to IMGT and post-IMGT sequences aligned to the reference sequences to determine error detection and correction.

4.4. Results

4.4.1. Reference Sequences

A set of 7 monoclonal mouse hybridoma cell lines was used to investigate the distribution and influence of insertions and deletions (indels) produced by the Ion Torrent PGM sequencing technology on murine IGH repertoire sequencing (**Figure 27**)**.** Reference sequences were obtained from Sanger sequenced cDNA transcripts of hybridoma RNA subsequently annotated and translated into amino acids by IMGT V-QUEST.

4.4.2. Distribution of artificial insertions and deletions

To investigate the influence of indels on IMGT processing at all possible positions of an IGH sequence, we generated a benchmark dataset from the reference sequences that contained artificially introduced indels at random positions. To cover each position within a 300 nt sequence with minimum 90% certainty, at least 2398 erroneous variants are required (291). Thus, we generated 2500 artificial, randomly flawed sequences for each permutation of 0-3 insertions and/or deletions (indels, annotated as i1d0, i0d1, i1d1 … i3d3), resulting in a total of 37500 artificial sequences per hybridoma with indels ranging from 1 to 6 events. Indels were homogenously present as determined by graphical reference alignment (**Figure 28A**). Uncovered positions resulted from indels within homopolymer stretches which were always assigned to the beginning of such a nucleotide repeat region (**Figure 28B**).

4.4.3. IGHVDJ nt error detection

As each sequence of the benchmark system contained indel errors, all sequences marked by IMGT as productive were falsely categorized as error free. In general, IMGT correctly recognized 97.9% (± 2.9%) of the introduced indels over all datasets and categorized the sequences then either as productive with detected indels, unproductive or unknown (**Figure 28C**). Interestingly, only the sets with one insertion and/or deletion (i1d0, i0d1 and i1d1) exhibited elevated numbers of unrecognized indels. For these IMGT falsely returned 8% (±1.8%) of the sequences as productive, whereas for all other datasets it was only 0.7% (± 0.4%). Such undetected indels were found at the beginning and the end of the sequence or across the whole sequence for i1d1 datasets due to indels in close proximity to each other masking the frame-shifts (**Figure 28D, E and Fig**ure **29)**. The number of unproductive sequences increased with the number of indel events, regardless of their composition. Accordingly, the number of productive

sequences with detected indels decreased. Less than 50% of sequences with more than 3 indels, were retained. Indels were homogenously distributed in the uncorrected productive sequences with detected errors until about 4/5th of the sequence lengths while the opposite is true for the uncorrected unproductive sequences (**Figure 28D, E, Figure 29**). This section of the sequence coincides with the IMGT IGH junction which encodes for the CDR3 (292). Accordingly, upon detecting an indel in the IGH junction, IMGT categorized the sequence as unproductive and no corrective attempts were made.

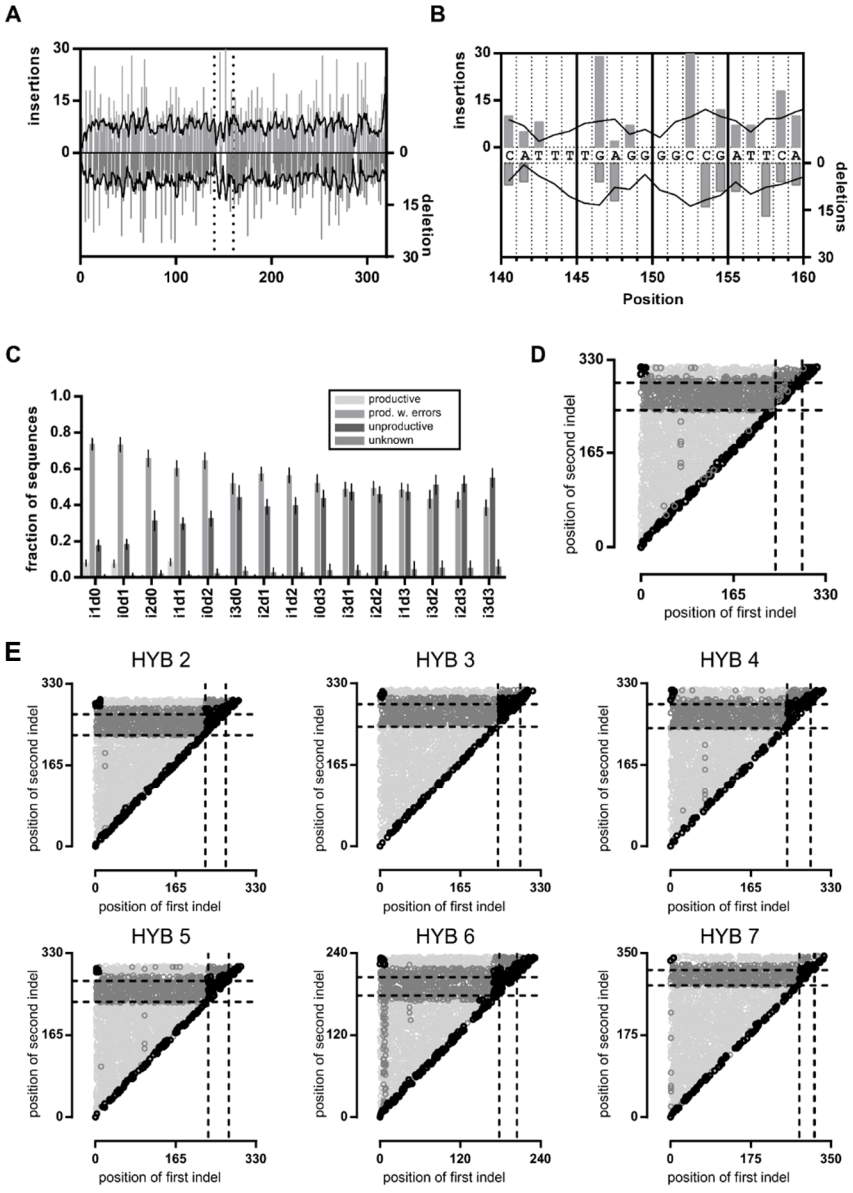


Figure 28 Indels in the artificial dataset. (A) Insertion and deletion events displayed as determined by graphical alignments of the reference sequence to the i1d0 and i0d1 dataset of hybridoma 1. Grey bars represent the actual detected indel and the black line presents the moving average over 4 neighbors. The dotted lines vertical present the segment that is magnified in (B) to visualize the problem of determining the position of indels in homopolymer repeats. (C) Indel detection rates by IMGT processing shown as bar chart with error bars indicating the SD over all 7 datasets (D) Visualization of indel proximity. The distances between the first and second indel before correction in the i1d1 dataset of Hybridoma 1 are shown as scatterplot. Dotted lines indicate the position of the IMGT junction (i.e. the nt encoding for the CDR3). Productive sequences with detected indels are shown in light grey, unproductive sequences are shown in dark grey. Sequences without detected errors are shown in black. (E) Like (D) but for Hybridomas 2-7.

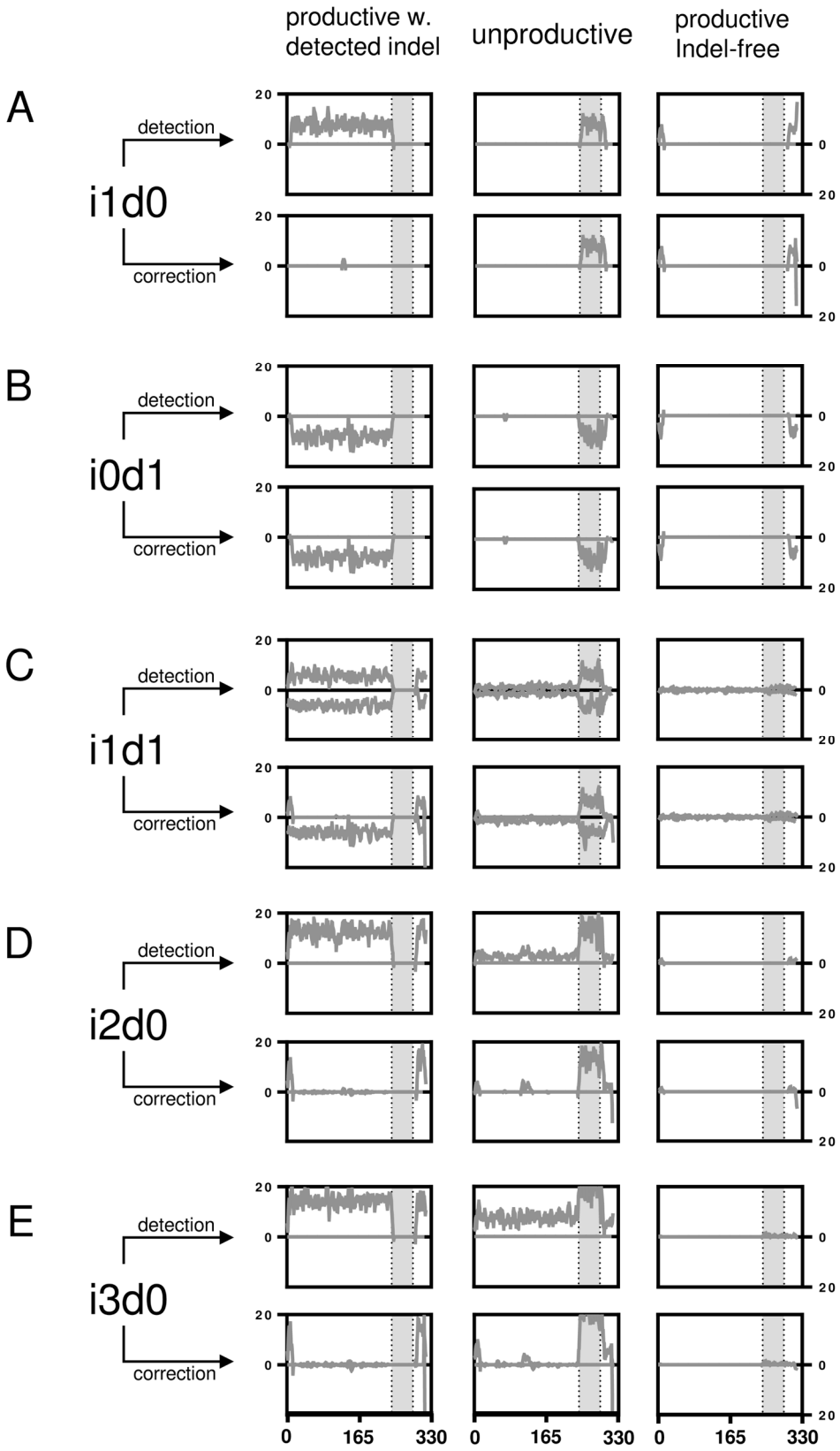


Figure 29 Selected alignments of artificially falsified datasets from Hybridoma 2. Indel positions are shown before and after IMGT error correction for hybridoma 1 separated by productivity. (A) The indels for the i1d0 dataset are shown per nucleotide position as line plot (smoothened over 4 neighbors). The grey area marks the IGH VDJ junction. (B-E) like (A) but with different, indicated permutations.

4.4.4. Nucleotide error correction

Upon detection of an indel, IMGT tried to correct it by alignment to its closest germline. The efficacy of this process was investigated by aligning the sequences with detected indels to determine the number of correct sequences (**Figure 29 and Figure 30**). A thorough error reduction was observed for up to three insertion errors in datasets without deletions, returning 87% ± 3.2% (i1d0), 72% ± 5.5% (i2d0) and 56% ± 7.0% (i3d0) of productive sequences (including productive with detected errors) as correct (**Figure 30**). Within these sequences indels that were not corrected by the IMGT were mainly found at the beginning and end of the sequence (**Figure 29**). In the case of deletions, the IMGT correction introduced a gap for the missing nucleotide as the original nucleotide was unknown. Consequently, the number of correct sequences found in mixed datasets is very low (i1d1: 1% ± 0.3%, i2d1: 2% ± 0.3%, i3d1: 2% ± 0.6%, i2d2 and i3d2 <1%). In datasets with insertions and deletions, the number of insertions within the sequences was always reduced (**Figure 29C**). No correct sequence could be identified in deletion-only datasets (**Figure 30**).

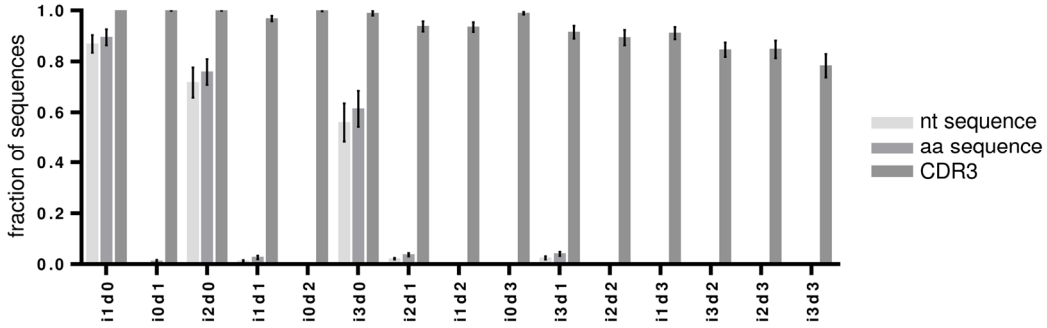


Figure 30 Indel correction by IMGT. The fraction of correct sequences shown as bar charts of nucleotide (nt), amino acid (aa) and CDR3 amino acid sequences. Error bars indicate SD over all 7 datasets.

4.4.5. Amino acid error correction

Theoretically, translated amino acids are less influenced by sequencing errors because of the redundancy of the genetic code. Thus, most amino acid translations were returned correctly in the case of insertion-only datasets and with slightly higher numbers compared to the nucleotide datasets (mean correct amino acid sequences for i1d0: 89% ± 2.9%, i2d0: 76% ± 4.7%, i3d0: 61% ± 6.5%, **Figure 30**). Although also very low, higher numbers of correct translations were observed in mixed indel datasets than for the corresponding nucleotide datasets (i1d1: 3% ± 0.7%, i2d1: 4% ± 0.6%, i3d1: 4% ± 0.8%, i2d2 and i3d2 <1%, **Figure 30**). Interestingly, some amino acid translations were found to be correct for the i0d1 datasets (1% ± 0.5%, **Figure 30**). Deletion-affected datasets were usually returned with the

wrong amino acid sequence by the IMGT algorithm. During IMGT processing, nucleotide deletions rendered the whole codon triplet elusive and were translated as gaps in the amino acid sequence.

Remarkably, the CDR3 proved to be protected chiefly from insertions and deletions through a more conservative correction approach of the IMGT algorithm for this part of the sequence. As mentioned above, detected indels within the IGH junction, and thus the CDR3, corrupted the entire sequence as unproductive (**Figure 29**). Correction and culling attempts by IMGT turned out to be largely successful (100% correct CDR3s for up to 3 insertions or deletions). Even for the i3d3 indel permutation, IMGT returned 78% ± 4.3% correct CDR3s (**Figure 30**). Datasets with simultaneous insertions and deletions showed in general lower numbers of correct CDR3 sequences (range 78-97%). This resulted from sequences where indels were introduced in close proximity of each other, producing no detectable frameshift within the IGH junction (**Fig 28D, E**). While invisible for the IMGT algorithm, they were observed as variants of the correct CDR3 amino acid sequence.

Taken together the above data show, that IMGT processing exhibits adequate detection of indels through frame-shifts in mouse IGH nt sequences. Consequently, frame-shift masking errors cannot be detected and result in amino acid changes in the translations. IMGTs indel correction proved to be reliable for single insertions. However, the impossibility to correct for deletions and larger indel permutations makes consideration of sequences categorized as “productive with detected indels” unfavorable.

4.4.6. HTS of hybridoma ssUID libraries

Next, the IMGT database and a PGM-tailored data processing pipeline developed by our group were tested using real HTS datasets derived from 7 monoclonal hybridomas (**Figure 26, Figure 27**). The HTS libraries were prepared using an Ion Torrent PGM tailored single-side UID approach (**Figure 26**) allowing for error correction through building consensus sequences from all reads within a UID family. The ssUID barcodes together with the C-region primer and appropriate ‘GATC’ spacer were correctly identified at the sequencing start site of 99.12% ± 0.56% of the usable reads containing a sample specific MID (**Table 12**). Between 146,010 and 739,854 reads were obtained per sample, with varying UID family size distributions (**Figure 31A**). After raw data processing, 1,431 to 47,169 consensus sequences were retained per hybridoma (**Table 12**) and uploaded to IMGT HighV-QUEST.

Table 12 HTS datasets of monoclonal hybridomas, pre-IMGT.

Set	CDR3	Chip	reads with MID	reads with primer & UID	consensus sequences
HYB1	SRWDYRYVYYPLDY	A	207,753	206,929	4,159
HYB2	ARTYYGSYGFDY	A	147,634	146,010	7,760
HYB3	ARQWLILWLGFAY	A	222,929	222,100	1,431
HYB4	ARWDYRYVYYPLDY	A	882,242	877,823	16,643
HYB5	TRGYYRYDGGFY	B	747,827	733,258	7,319
HYB6	APKGLAY	B	743,465	739,854	47,169
HYB7	ASRTTATGY	B	204,348	201,619	5,426

4.4.7. IMGT processing of HTS datasets

The majority of the post-IMGT sequences were categorized as productive (75.8% ± 22.6%) and 10.9% (± 9.6%) were categorized as productive with detected indels (**Table 13**). The remaining sequences were either categorized as unproductive or unknown/else. To investigate the undetected or uncorrected errors within the two productive categories, sequences were aligned to their corresponding references. For Hybridoma 3, which had the poorest UID distribution (**Figure 31A**), only 26.8% of the sequences were classified as productive and 68.8% unproductive (**Table 13**). This hybridoma was therefore excluded from further analysis.

In the group of productive sequences with detected errors, IMGT's indel correction algorithm improved the number of correct sequences by 54.1% to on average 55.3% (32.2%, **Figure 31B**). As expected, IMGT corrected most sequences that contained single insertions efficiently, reducing these errors from average 25.2% (± 24.3%) to 0.48% (± 0.72%, p-value = 0.0027, two-tailed t-test in Graphpad Prism, using Holm-Sidak's method (293) to account for multiple testing with alpha = 5%). Single deletions were found in 29.9% of the sequences (± 24.3%). They increased slightly after IMGT error correction (31.6% ± 24.1%), as insertions of higher indel permutations were corrected, leaving only deletions in the sequences. Accordingly, such higher permutations were found in 33.8% (± 23.8%) of the sequences before error-correction and in 8.8% (± 6.3%) afterwards. Accordingly, these were found in 33.8% (± 23.8%) of the sequences before and in 8.8% (± 6.3%) after error correction. While the detection of indel errors in the sequences by IMGT was efficient, the remaining errors after correction affected 44.7% ± 32.2% of the sequences and, as described for the benchmarking sequences above, makes further consideration of sequences marked as "productive with detected indels" inadvisable.

Table 13 HTS datasets of monoclonal hybridomas post-IMGT

Set	prod. seq	%	prod. w. det. indel	%	unprod	%	unknown/ else	%
HYB1	3,328	79.6%	622	14.9%	127	3.0%	102	2.4%
HYB2	4,866	62.7%	2,449	31.6%	250	3.2%	195	2.5%
HYB3*	381	26.6%	62	4.3%	984	68.8%	4	0.3%
HYB4	13,515	81.2%	2,215	13.3%	329	2.0%	584	3.5%
HYB5	6,697	91.5%	281	3.8%	51	0.7%	290	4.0%
HYB6	43,767	92.8%	3,009	6.4%	287	0.6%	106	0.2%
HYB7	5,216	96.1%	111	2.0%	15	0.3%	84	1.5%
Mean	11,110	75.8%	1,250	10.9%	292	11.2%	195	2.1%
SD	13,842	22.6%	1,165	9.6%	303	23.5%	180	1.4%

Sequences marked as productive without detected indels are not modified by IMGT but can nonetheless contain indel and nucleotide substitution errors. IMGT does not detect ambiguous nucleotides as errors but marks them as silent mutations. On average 2.2% (± 1.6%) of the consensus sequences in the productive dataset without detected indels contained ambiguous nucleotides (**Table 14**), which were discarded from the datasets.

Table 14 Ambiguous nucleotides in productive sequences without detected indels. * Indicates that Hybridoma 3 was excluded from the calculation of mean and SD.

	HYB1	HYB2	HYB3*	HYB4	HYB5	HYB6	HYB7	**Mean***	**SD***
Amb nt	26	135	46	97	90	2289	148	**464**	**817**
%	0.8	2.6	12.0	0.7	1.3	5.2	2.8	**2.2**	**1.6**

Most of the remaining sequences were indeed error-free (98.8% ± 0.5%, **Fig. 31C**). The other 1.2% contained on average, 0.2% (± 0.1%) i1d1 in close proximity to each other, masking frameshifts. Some sequences showed single insertions (0.1% ±0.2%) and deletions (0.15% ± 0.13%), found at the

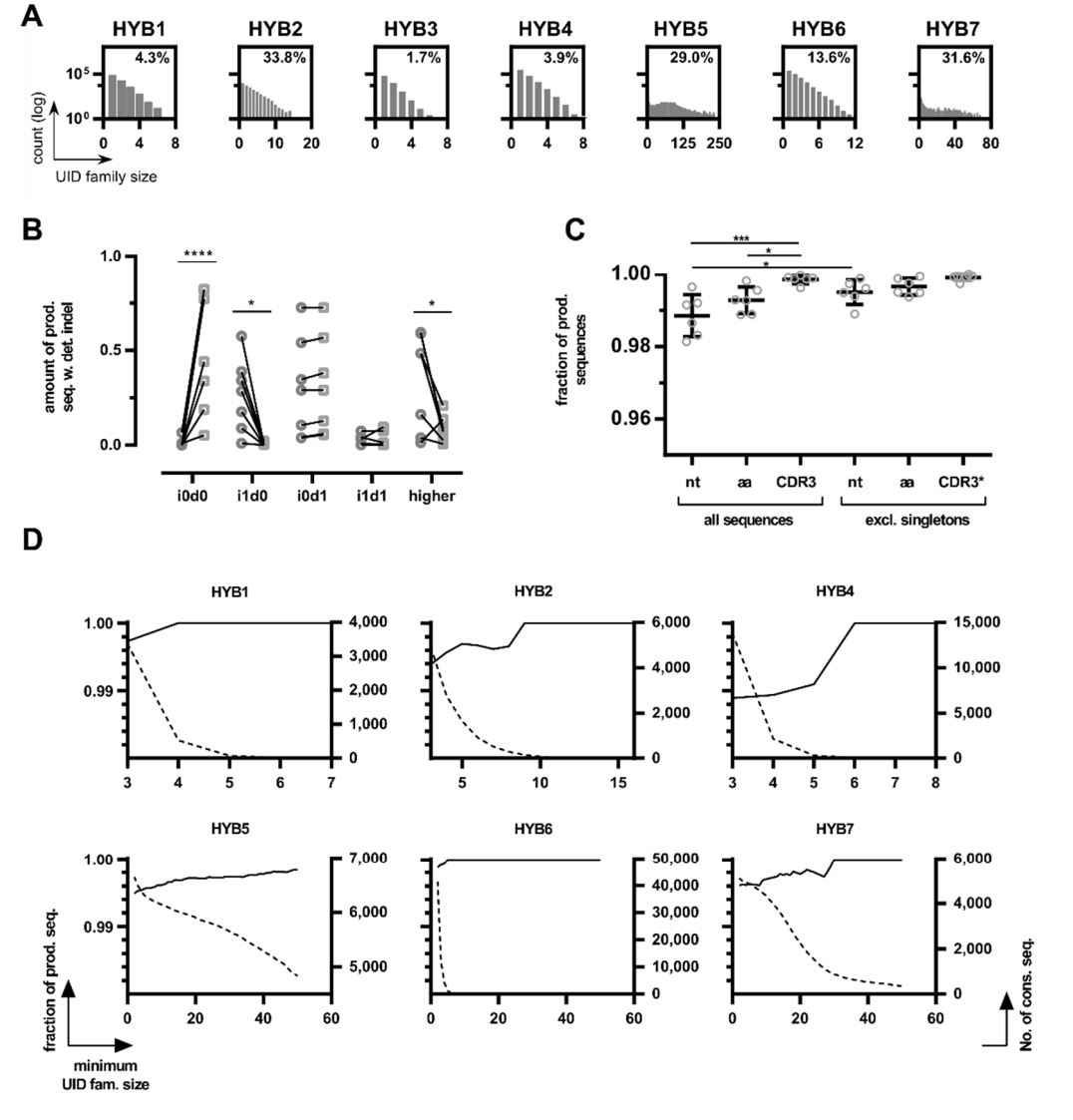


Figure 31 HTS data on monoclonal hybridomas. (A) UID family size distributions per sample. The number of UID families (log transformed) is plotted by the number of reads assigned to a ssUID per hybridoma. The amount of UID families containing a minimum of 3 reads are indicated as percentage value. (B) Indel distributions on productive sequences with detected errors. The amount of indel-free (i0d0), single insertions (i1d0), single deletions (i0d1), one single insertion and deletion (i1d1) and higher permutations are shown as fraction of productive reads with detected indels before (circles) and after (squares) IMGT error-correction. P values are indicated **** $p < 0.0001$, * $p < 0.05$, multiple two-tailed t-test with Holm-Sidak's method to account for multiple testing. All other differences were not statistically significant. (C) The number of error-free sequences in the productive dataset without detected indels are shown as scatterplot with mean and ± SD. Data are shown for all nucleotide sequences (nt), amino acid sequences (aa) and CDR3s for all sequences and data without singleton sequences. For CDR3s singleton exclusion was performed on the basis of full-length amino acid sequences. P values are indicated *** $p < 0.001$, * < 0.05, One-way ANOVA with Sidak's post-hoc test. All other differences were not statistically significant. (D) Influence of UID family size on amount of correct sequences. The number of correct sequences are shown as line per minimum UID family size (left y-axis). The number of consensus sequences are shown as dotted line per minimum family size (right y-axis).

beginning or the end and therefore not causing a detectable frameshift. The remaining false sequences contained mainly nucleotide substitutions, with the majority being transversions (0.5% ± 0.3%) and very few transitions (< 0.1%). As described by Shugay and coworkers, such substitutions originate from

dominating polymerase errors occurring early during the amplification (73). As polymerase errors are occurring at relatively random positions, it is stochastically unlikely, that the same errors are found repeatedly within a dataset and can thus be accounted for by considering only consensus sequences that appear more than once in the final dataset (72, 73). Following this approach, the data was reassessed, excluding singleton consensus sequences which reduced the number of sequences by 0.8% (± 0.4%). The number of transversions was reduced significantly by 0.3% to 0.16% (± 0.19%, p-value = 0.008, two-tailed t-test in Graphpad Prism, using Holm-Sidak's method to account for multiple testing with alpha = 5%, data not shown). Consequently, the amount of error-free sequences improved significantly by 0.7% to 99.5% (±0.3%, p-value < 0.0001, two-tailed t-test, using Holm-Sidak's method to account for multiple testing with alpha = 5%).

The number of reads per UID family is crucial to obtain reliable consensus sequences Increasing the minimum number of required reads per UID family improved the amount of correct sequences, reaching 100% for all hybridomas, except Hybridoma 5, albeit with different UID family sizes (**Figure 31D**). However, with increasing minimum UID family sizes, the number of sequences decreased exponentially. Consequently, at the point of reaching 100% correct sequences, on average only 7.9% (± 7.1%, excl. Hybridoma 5) of the sequences remained (**Figure 31D**). According to our data, keeping a minimum UID family size of 3 provided adequate accuracy and throughput when using an Ion Torrent PGM.

As expected, the amount of correct amino acid sequences was higher (99.3% ± 0.3%) than the amount of correct nucleotide sequences (**Figure 31C**). An average of 0.6% (± 0.4%) of the sequences was subject to amino acid changes. Excluding singleton amino acid sequences increased the number of correct amino acid sequences to 99.7% (± 0.2%), but this increase was not statistically significant. CDR3 amino acid sequences were returned almost entirely correct (99.85% ± 0.11%, **Figure 31C**), increasing to 99.91% (± 0.08%) when singleton full length amino acid sequences were excluded.

In summary, combining the described ssUID library preparation with UID family consensus sequences followed by indel detection through IMGT presented a highly reliable IGH repertoire sequencing approach on the Ion Torrent PGM.

4.1. Discussion

Investigation of IG repertoires by HTS is challenging both with respect to the library preparation as well as sequencing error assessment and data processing. Using artificially falsified sequences, we show here that the IMGT indel detection algorithm is efficient while the IMGT indel correction algorithm only corrects single insertions efficiently. We confirm the utility of the Ion Torrent PGM to assess murine IGH repertoires with high confidence, using a dedicated library preparation protocol with a PGM-tailored 16 nt single side unique identifier (ssUID) barcoding technique. Our data show, that appropriate data processing reduced the error rate of PGM-sequenced IGH repertoires to less than 0.5% false nucleotide and amino acid sequences, and to less than 0.01% false CDR3 sequences per dataset.

Sequencing of IGH repertoires requires a thorough assessment and correction of platform inherent sequencing errors (57, 269, 270, 273–275). Using the IMGT database for reference alignment, the typical Ion Torrent PGMs indel errors can theoretically be detected as codon frame-shifts (88). The VDJ structure of the IGH sequence facilitates indel detection by frame-shift, since gene segments can be aligned separately. In our study, the IMGT algorithm successfully detects 97.9% of all indels, regardless of their composition and only single insertions or deletions at the beginning or the end of the sequences (7.9% and 7.5%, respectively), or i1d1 compositions in close proximity to each other could not be identified (8.5%). IMGT tries to correct detected insertions subsequently by removing the false nucleotide(s) according to the predicted germline sequence. In the artificially falsified datasets of our study, insertion-only errors were corrected by the IMGT algorithm with 87% (i1d0), 72% (i2d0) and 56% (i3d0) efficiency. Deletions, on the other hand, are more difficult to recover since the missing nucleotide cannot necessarily be inferred from the germline sequence with sufficient confidence. Consequently, artificially introduced deletions were not corrected by IMGT. Also, for sequences with mixed insertions and deletions only the insertions were corrected by IMGT leaving the sequence erroneous. Taken together, these data indicate that detection of indels by IMGT is highly efficient and sequences categorized as “productive” without detected errors are almost entirely indel-free. The low efficiency of the indel correction algorithm makes it inadvisable to take productive sequences with detected indels, which in our study correspond to about 10% of the final HTS consensus sequences, into account for any downstream analysis.

Mixed events of adjacent insertions and deletions are the most difficult to detect and can remain unnoticed by the IMGT algorithm. The resulting nucleotide substitutions can falsify somatic

hypermutation profiles (56, 281). UID barcoded RNA transcripts allow to this problem (68, 71–73). B cells contain up to several thousands of identical IG RNA molecules that are individually tagged by a UID (71, 294). Therefore, a HTS run provides a snapshot of the relative abundance of RNA transcripts (56). As for SNP identification, single occurrences of nucleotide substitutions can be ruled out as artifacts and only transcripts above a certain copy threshold should be retained (294). Our data show, that considering sequences with at least 2 copies in the final dataset improves the proportion of correct sequences by 0.7% to 99.5%. In this regard, as our sampling material are monoclonal hybridomas, all derived sequences (between 1,431 and 47,169) represent identical RNA-molecules, making it stochastically more likely, that the same indel error appears several times. Thus, it is expectable, that the positive influence of excluding singletons would be even higher in bulk B cell derived datasets, where less sequences are derived from identical RNA molecule.

Template amplification with multiple primers during library preparation can significantly bias the repertoire composition (57, 76). This bias is essentially removed by UID barcoding but reduces sequencing depth at the same time (72, 74, 295, 296). In our study, the raw sequencing depth does not influence the relative number of correct sequences while the average UID family size proved to be crucial. For instance, Hybridoma 3, although having only the 3rd lowest amount of raw-reads, lacked eligible UID family sizes (> 2 sequences per UID). For this Hybridoma 3, less than 0.5% of the consensus sequences were built from UID families with more than 2 members, resulting in the poorest error correction rate during sample processing. Consequently, IMGT returned only 26.6% of the consensus sequences as productive. We therefore conclude from our data, that for applying a UID family-wise consensus building approach, samples with less than 0.5% eligible consensus reads after pre-IMGT processing do not have enough coverage to achieve sufficient confidence and depth for the post-IMGT sequences and should be discarded from further analysis.

For grouping reads by UID families, it is essential to identify the UID tags correctly (72, 74). The PGM sequencing chemistry is unidirectional, starting with the sequencing adapter A. Comparable protocols for the Illumina sequencing platforms usually consist of UID tags at the beginning and the end of the amplicon sequence (71). We chose to introduce the 16 random nucleotides of the UID tag at the sequencing start site as the PGM semiconductor technology is significantly less accurate towards the end of the sequence (80). We included a 4-nucleotide spacer as junction into the UID tag resulting in the N8-GATC-N8 ssUID layout of this study. Like this we address that the PGM indel rate increases in homopolymer stretches with their length (75), in particular when homopolymers are longer than 8nt (81).

While breaking potential homopolymer patterns within the UID, this design also reduces the number of mistakes during primer synthesis and allows to generate sets of primers with individual spacers that could be used to tag different experiments.

In conclusion, we have demonstrated that using our ssUID library preparation in combination with the IMGT database, the PGM sequencing platform can be efficiently used to assess murine IGH repertoires. Considering only consensus sequences with at least two copies improved the sequence quality considerably. Taken together, this approach allowed to obtain highly reliable IGH sequences, with more than 99% confidence in general and 99.9% confidence for the correct CDR3 sequences. The protocol and sample processing strategies described in this study will help to establish the benchtop-scale Ion Torrent sequencing technology of animal models in the field of immunoglobulin repertoire research.

Chapter 5 - General Discussion

4.2. IG repertoire convergence

The IG repertoire of an individual is determined by several factors, including genotype and chromatin structure (84, 297), age (85, 111, 298–301) and history of immunogenic events (111). Recent advances in high-throughput sequencing have raised the possibility to investigate the modulation and dynamics of IGH repertoires after vaccination and infection. Together with low-resolution studies, there is an increasing body of information available characterizing the immune response to vaccination against influenza (40, 71, 78, 85, 89, 298, 302), tetanus (49, 50, 303), *Haemophilus influenza* type b (Hib) (53, 304), *Streptococcus pneumoniae* (298, 305) and hepatitis B (306), as well as infectious diseases including influenza (37, 307–310), rotavirus (311–314), HIV (42, 87, 88, 273, 315–318), hepatitis C (319), cytomegalovirus (CMV) (111), Epstein-Barr virus (EBV) (111), *Staphylococcus aureus* (320) and dengue virus (44, 90, 321–323). In the context of this thesis, a similar characterization of IGH repertoire responses is provided for the first time, for measles virus (MV) antigens, Modified Vaccina virus Ankara (MVA) and tetanus toxoid hapten conjugated with Benzo[a]pyrene (TT-BaP).

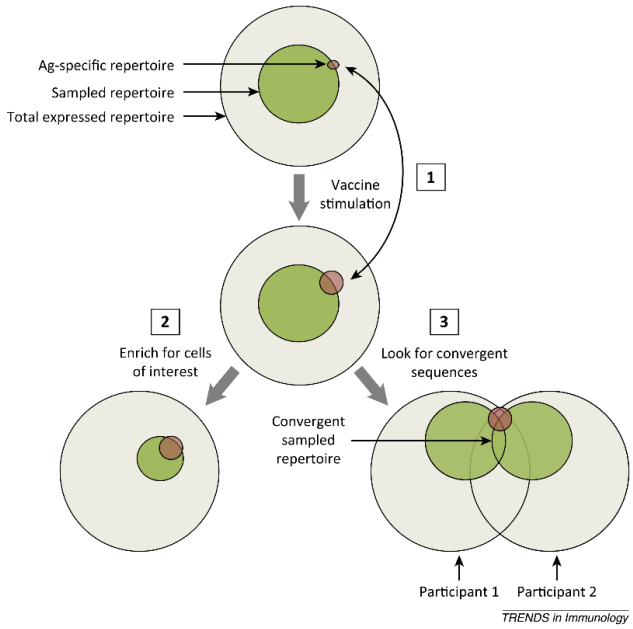


Figure 32. Investigation of the antigen-induced repertoire by vaccination. (1) The proportion of the antigen-induced repertoire is enriched by the clonal expansion following vaccination. (2) The ratio between antigen-induced repertoire and sampled antigen-unrelated repertoire can be physically enriched by isolating plasma cells or sorting cells for antigen-specificity prior HTS. (3) Antigen-induced sequences can be identified by stimulating several individuals with the same antigen and screening for convergent responses. Reprinted from (48), copyright 2014, with permission from Elsevier.

Selecting antigen-induced IG sequences from the total repertoire is crucial for investigating the response of IG repertoires to antigenic stimuli (48). An *a priori* method for identifying such sequences, is the investigation of the shared fraction of IG repertoires across individuals after recent exposure to a common antigen (**Figure 6** (48, 56)). This is referred to as a convergent IG response or public IG repertoires. They can either be identified by screening datasets for sequences of known specificity and with a certain degree of similarity, or by searching for sequences shared across donors (90, 92). In this regard, the large variability of the IGH CDR3 sequences can be exploited as barcode for antigen-specificity (3). Low throughput studies revealed that the IG repertoires of 19 individuals, after vaccination with small, biochemically defined Hib polysaccharides, induced similar clones using predominantly IGHV 3-23 germline genes and IGH CDR3s with a common 'GYGMD' amino acid motif (324). For more complex antigens, such as TT or influenza, the IG responses differed across individuals (40, 49, 50). Even within a single individual, repeated influenza vaccinations only resulted in one-third of the sequenced clones being shared between the samples (40). However, these studies were limited to less than 500 sequences per individual. The first human HTS based studies showed, that although the usage of IGH gene recombinations can follow a similar pattern between individuals, the IGH CDR3 sequence is mostly private (85, 158). Nevertheless, the CDR3 'ARLDYYYYYGMDL' was shown to be overexpressed across 60 subjects suffering of an acute dengue infection, but remained undetected in healthy controls (90). Therefore, it seems plausible, that (certain) antigenic exposure can limit the possible CDR3 diversity, facilitating IG repertoire convergence. However, a major concern with such studies is the unavailability of cloneable full IG heavy and light chain information to perform subsequent verification of the IG specificity. Following the hypothesis of public CDR3 sequences, we investigated IGH repertoire convergence in response to several different and overlapping antigens (TT, TT-BaP, MVA, MVA-HF) in the OmniRat™ model. To account for the high variability and private nature of CDR3 sequences, it is common to allow for some level of similarity when investigating about shared and public CDR3 repertoires (90, 92, 306). Most studies arbitrarily defined thresholds of global amino acid mismatches (92) or sequence similarity (306), and little is known about the applicability of such measures. We present for the first time a thorough investigation of this link, by using a series of moving CDR3 sequence similarity thresholds, with subsequent determination of the public repertoire across 4 vaccination and 2 control groups. Our data show that an 80% sequence similarity linked between 6% and 46% of IGH repertoires to the underlying vaccinations. Although we were unable to verify the specificity of these identified sequences, due to lacking information on the corresponding IGL

sequences, there are two aspects supporting their link to the applied vaccinations. First, all identified CDR3 sequences were virtually absent in unrelated vaccination groups or controls, but were shared across animals belonging to the same group and even across groups for overlapping antigens (e.g. TT and TT-BaP or MVA and MVA-HF). Second, we identified CDR3s induced by TT and MVA-HF that were present in human-based studies after TT vaccination (92, 179, 325, 326), or a study using whole MV antigens administered to OmniRat™ through different vaccination routes (177). The latter is particularly interesting, as the specific characteristics of the IG response to MV have not yet been investigated using HTS. Yet, there is strong evidence of a converging molecular IG response, as 96% of all neutralizing serum antibodies can be mapped to two dominant epitopes on the MV hemagglutinin (H) protein (327, 328). While this might be one of the reasons of the anti-MV vaccine's success, it certainly also warrants the study of IG response to this antigen with regard to IG repertoire convergence. However, the high vaccination rate of the population might impede the control recruitment in human studies.

We provided a ready-to-use bioinformatics pipeline for IG repertoire analysis, utilizing the RNA-seq framework DESeq2 to identify antigen-induced sequences as overrepresented clusters of 80% similar CDR3s. Albeit successfully applied for the Omnirat™ model, the identification of similar responses in a human vaccination studies, which followed MV vaccination in MV naïve people from the Lao People's Democratic Republic (LAO PDR), or when using published datasets of a hepatitis B vaccination study (306), with this data processing approach was not possible (data not shown). In this regard, the use of OmniRat™ animal model allowed to sequence a much larger fraction of the IGH repertoire, due to the limited amount of circulating B cells compared to humans. In addition, the animals experienced considerably less antigenic exposures than human counterparts, reducing the complexity of their pre-shaped IG repertoire. Furthermore, using an animal model allowed access to the bone marrow, where antigen specific plasma cells producing the neutralizing serum antibodies reside (28, 286). The human studies are limited to PBMCs. The plasma cells must be obtained within a narrow 1d time window in which they travel from the GC through the periphery to the bone marrow (329). While most human studies focused on sampling on day 7 post vaccination, the plasma cell bursts can be seen on days 4-10 following a challenge, with different kinetics between primary and secondary response (91, 329, 330). In addition, little is known about the connection between the transiting plasma cells and the neutralizing antibodies in the serum, detectable by ELISA (28). Lee and coworkers developed an elegant approach to solve these problems inherent to human studies (202). By combining high-resolution proteomics analysis of the serum antibodies using LC-MS/MS with a database developed from HTS of blood plasma

cells B cell receptor transcripts, they were able to link the serum response of individuals to the elicited B cells after influenza vaccination. They showed that the anti-influenza serum repertoire was comprised of 40 to 147 clonotypes (i.e. IG molecules of the same origin with different mutation patterns) and that ~60% of these were elicited from pre-existing clones (202). This is in line with the results presented in chapter 2, where the applied algorithm selected on average 121 (range 20-419) CDR3 clusters (i.e. clonotypes) with an 80% similarity from OmniRat™ bone marrow IG transcripts.

Transgenic human IG loci might limit the variability and affinity maturation processes of the OmniRat™. Wardemann and coworkers showed, that 55-75% of the antibodies they cloned from human, immature B cells exhibited autoreactivy *in vitro* (331). While these B cells were subsequently absent in the pool of naïve B cells released to the periphery in humans, the pool of naïve B cells in the transgenic animals most likely differed in composition. It could potentially contain some of these otherwise deleted self-reactive cells, making them available for antigen selection. Therefore, selected IG molecules with OmniRat™ origin cannot be transferred to human individuals without careful revision of their specificity and reactivity. Yet, in previous studies, the rats expressed a diverse IG repertoire and normal SHM profiles (164). Taken together, our data supported the hypothesis of converging CDR3 repertoires and warranted the use of animal models for future investigation on IG repertoire conversion (see also section 4.5).

4.3. HTS of B cell malignancies

The diversification and selection processes of IG repertoires in health and disease have important clinical implications. For instance, the IG repertoire in B cell malignancies is often dominated by a single, clonal IG sequence (63, 332) and intraclonal mutations via SHM have been observed in B cell lymphomas (333–335). In addition, the level of mutation in the tumor clone can be linked to the clinical outcome of chronic lymphocytic leukemia (CLL) patients, with mutated malignant clones (M-CLL) having inferior survival rates than those of unmutated malignant clones (U-CLL) (142).

HTS approaches on B cell malignancies are still at an early stage of development and until now, only two thorough studies have been published using network properties to characterize CLL and B Cell acute lymphoblastic leukemia (B-ALL) (336, 337). Nevertheless, these studies already showed several advantages for clinical monitoring and diagnosis purposes. Typically, B cell malignancies are detected and described using qPCR assays with human IGHV gene specific primers. This set-up complicates the detecting of multiple disease subclones, or independent B cell malignancies (64, 239). HTS of the same

samples, can provide such information, as well as follow the disparate dynamics of cellular clones hereby identifying individual responses to therapy. Furthermore, relapse of B-ALL can be predicted by early clonal dynamics and detection of minimal residual disease (MRD) which is often linked to chemo-resistant malignant clones (338, 339). These MRD clones may either be present at low levels at diagnosis or acquire the resistance by surviving the initial chemotherapy (340). In both cases, superior sensitivity of HTS approaches allows MRD clone detections already at diagnosis and thus much earlier than comparable qPCR assays.

In chapter 3 we described a similar network property analysis using a mouse model with designated $CD5^{+}$ B cell expansion in the peritoneal cavity. Hereby expanding the usage of Ion Torrent PGM sequencing to the development of disease-specific animal models. Typically, the characterization of an animal model with putative malignant B cell expansion would require extensive, time-consuming verification of the underlying malignant mechanisms. HTS on IG repertoires allowed to fully describe the clonal relationship, extensive expansion, SHM mutation profiles and inter-animal comparison of the B cell population in these mice at high depth. Enabling the determination of the CLL-like B cell expansion of the $A20^{BKO}sCYLD^{BOE}$ in the mouse model as an unmutated subvariant (U-CLL-like) and highlighting the expression of short CYLD gene variants as important target for future CLL investigations.

4.4. Technical issues

4.4.1. Error rates and sample preparation approaches

HTS sequencing has emerged as revolutionary method to characterize IG repertoires (48, 55–57, 281). However, there are several technical limitations of the sequencing technologies that require careful assessment. For instance, current sequencing error rates range from ~0.29 errors per 100 sequenced bases in Illumina based systems to as high as 0.99 and 1.63 in 454 GLS and Ion Torrent PGM systems, respectively (**Table 4**, as summarized by (83)). These rates were determined with whole genome and amplicon approaches on various species' and templates (75, 80, 81, 341–344). In contrast to the other sequencing platforms, the Ion Torrent PGM error rate is relatively high, but varies strongly with the different sequencing kits and chips used. For instance, Salipante and co-workers reported 0.015 errors per base (as ~1.5 errors per 100 bases) sequencing 16s-RNA with an Ion 400bp Sequencing kit and Ion 318^{TM} v2 Chips (345). Whereas Ross and colleagues report 0.011 to 0.020 errors per base (~1.1-

2.0 errors per 100 bases) in a whole genome approach, using 200bp Sequencing kits and Ion 318™ v1 chips (82). In both cases, the majority of errors were insertions and deletions (indels) at homopolymer regions of the sequences, with insertions being slightly more abundant. The lack of data combined with different sample preparation, sequencing and analysis protocols, impeded a genuine error rate determination of the Ion Torrent PGM platform (83). Nevertheless, it can be readily assumed that a typical PGM run produces roughly ~1 false basecall per 100 base pairs. To obtain sufficient information from IG sequences, an amplicon should consist of 200-400 bp, covering IG FR3, FR2 or even FR1 gene segments down to into the constant region to determine the IG isotype. Nucleotide substitutions, on the other hand, are occurring one order of magnitude less frequently than indels because of the PGM sequencing technology (75, 81, 82). Therefore, it can be expected, that every single raw read from a PGM sequenced IG transcript contains between one and three indels, with varying permutation.

A standard approach to account for sequencing errors is to increase the reads per sequence template (coverage), at the detriment of sequencing depth (346). Like this, by combining similar reads into a pile-up alignment, most sequencing errors can be ruled out efficiently (346, 347). In the case of IG repertoire sequencing, this method cannot be applied without additional preparation steps, as nucleotide substitutions may arise from sequencing errors or somatic hypermutation at similar rates (19, 348). Several closely related library preparation methods, using single molecule barcoding (referred to here as unique identifier, UID) strategies, have been developed to overcome this problem (73, 167, 294, 349–351). In such approaches, random UIDs of 10-16nt are used to label each RNA or DNA molecule in the reaction prior to amplification, e.g. during reverse transcription (71–73). Subsequent PCR then amplifies the UID, and reads can be grouped into UID-families post-sequencing (73). A UID family contains multiple copies (i.e. coverage) of the same original molecule, enabling sequencing and PCR error corrections by building a consensus sequence (**Figure 33**) (72). However, recent studies have identified several theoretical and practical limitations of these approaches (72, 74).

First, the reads within one UID family are not independent copies of the original sequence. PCR errors occur as branching processes, therefore early mistakes during PCR will be amplified exponentially and cannot be corrected through consensus building (73, 352). This is also valid for errors introduced during reverse transcription of the RNA molecules, part of most IG repertoire HTS sample preparation strategies. However, such errors can be accounted for by only considering consensus sequences that appear several times in the final dataset (72). In the PGM data presented in **Chapter 4**, the number of correct sequences increased to over 99.5% by considering only consensus sequences with at least 2

copies. With the remaining 0.5% erroneous sequences containing mostly a single mismatch or indel over a ~300 bp amplicon, this translated roughly into about 1.6 × 10^{-5} errors per base (calculated as follows: 0.5% of 77,095 sequences contain 1 error over 300bp, which approximates to 390 total false bases divided by 77,095 × 300 bases). This provides a 600-fold improvement over the original 0.01 errors per base. Illumina sequencing of UID-based samples normally require a consensus sequence to be present at least 5 times to be genuine (71), strongly increasing sequencing accuracy with error rates as low as 10^{-8} errors per base (167). However, this is only feasible with the higher throughput of Illumina sequencing platforms.

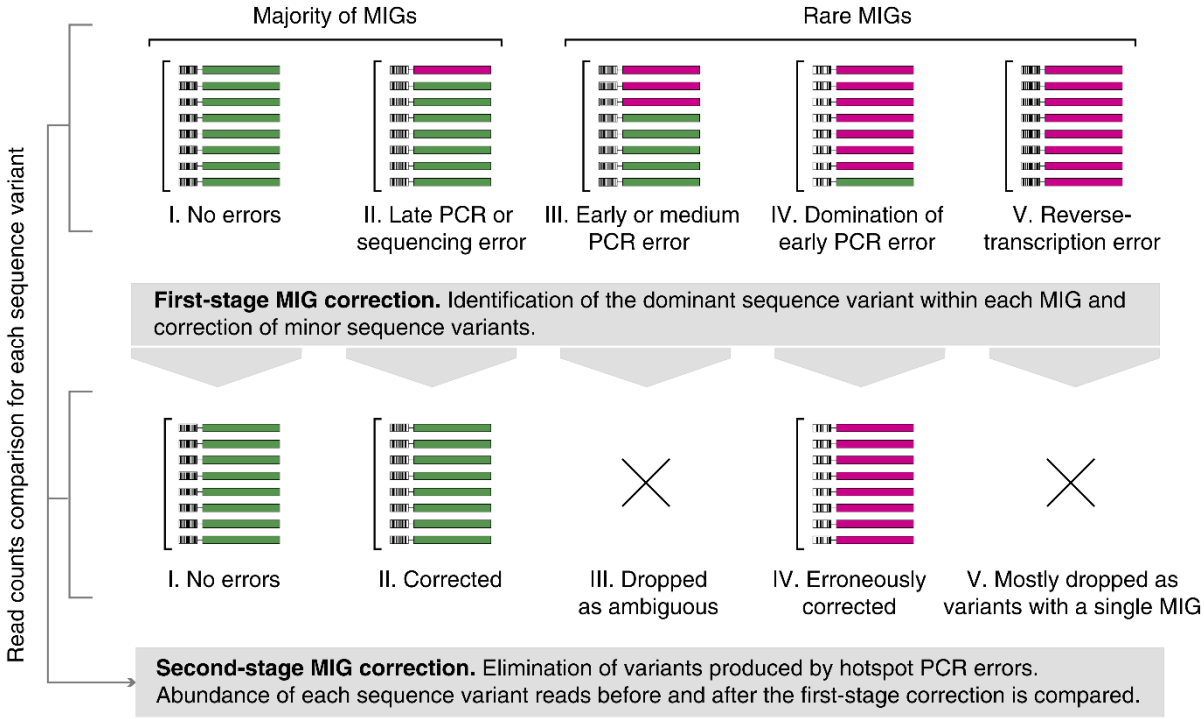


Figure 33 HTS sequencing error correction by building UID family consensus sequences. UID families are referred to as Molecular Identifier Groups (MIG). The majority of UID families are error free or contain few sequencing and/or PCR errors which are corrected through consensus building. Other MIGs contain errors occurring during reverse transcription, or so early in the PCR reaction that the subsequent amplification leaves them as dominant sequence in the UID family. These sequences can be not corrected through consensus building but by excluding consensus sequences with ambiguous nucleotides and considering only consensus sequences represented several times. Reprinted by permission from Macmillan Publishers Ltd: Nature Methods (73), copyright 2014.

Second, while the UID consensus approach allows to retrospectively remove PCR amplification during data analysis, it does not remove the actual bias during the reaction. In this regard, templates may be amplified with different efficiency due to differences in the UID or target sequences (349). Together with sampling variance, this can result in substantial differences in UID family size distribution, strongly altering sequencing depth and UID mediated error correction (72). This was also observed for the library preparation presented in **Chapter 4**. Although all samples were processed identically, i.e. 100ng starting material and in total 20 cycles of PCR amplification, the UID family size distribution varied significantly.

For instance, samples from the two Hybridomas 5 and 6 displayed a similar sequencing depth (747,827 and 743,465 reads respectively). Yet, they presented a more than 6-fold difference in the final consensus sequence count (7,529 and 48,084, respectively), and a 22-fold difference in average UID-family size (78 and 3.5, respectively). Consequently, the number of correct sequences after full processing was higher (99.4%) for Hybridoma 6 than in Hybridoma 5 (97.9%). Similar observations have been reported for Illumina-based UID sequencing strategies and a UID-family size of ~10 has been warranted optimal for error correction, sequencing depth and accuracy (72).

Third, sequencing and PCR based errors can also affect the UID sequence itself, resulting in an overestimation of the original molecule count during the data analysis process (74, 76). Egorov and co-workers explain that for example "[…], if one [UID] of 12 nt length is amplified and sequenced 10^4 times, it may routinely produce ~10-20 erroneous [UID] subvariants generated during PCR amplification, which may be represented altogether by >100-200 sequencing reads […]" (74) based on spike-in experiments (73, 74, 275). They also note, that such UID subvariants were reproducibly generated, mostly during late PCR cycles (cycles 27-35) when a large number of molecules was amplified (74). It is therefore advisable, to keep the number of amplification cycles during HTS library preparation to an absolute minimum, and rather adjust the input material. However, this presents a complex task for bulk B cell experiments, as the RNA abundance ratio from peripheral blood B cells typically varies strongly [estimated 500:5:2] between plasma cells, memory B cells to naïve B cells (72). Therefore, when seeking information about memory or naïve B cells in (m)RNA-based approaches, FACS sorting should be applied to separate these cells from the otherwise overrepresented plasma cells.

Forth, the diversity of UIDs tends to be lower than estimated by theoretical calculations. For instance, a 12 nt UID should theoretically have a diversity of 1.7×10^7 unique variants, but the actual observed diversity was only 1.4×10^7 (74), putatively resulting from uneven primer synthesis. PCR and sequencing errors in the UID sequence further reduce the actual diversity, resulting in $>3\times10^4$ natural UID collisions, i.e. the same UID for different molecules, in the case of 10^6 starting molecules (74, 351, 353). In the method presented in **Chapter 4**, a 16nt UID was chosen, resulting in 4.3×10^9 possible variants to tag the RNA molecules of 10^6 cells, rendering natural collisions unlikely. To reduce the influence of sequencing- and polymerase-errors in the UID sequence, the UID was split into two times 8 nt with a 'GATC' spacer. This spacer breaks potential homopolymer patterns within the UID, reducing the amount of possible sequencing errors. The utilized Ion Torrent PGM sequencing platform is prone to indels in

homopolymer regions with probability correlated to their length, becoming especially pronounced with more than 8 repeated nucleotides (75, 80–82).

4.4.2. Study design and biological constraints

One of the applications for HTS approaches on IG repertoires is antibody discovery, providing an economical advantage over time and cost-intensive conventional approaches. Within **Chapter 2**, we described a method, that allows selection for antigen-driven IGH sequences from the total sequenced repertoire of OmniRat™, using public repertoire analysis. However, due to the lack of information concerning the corresponding IGL, the actual specificity of these IGH sequences remains elusive. This has been a common problem to most studies aiming on identification of convergent IG repertoires or antibody discovery (48, 90).

The level of somatic hypermutation within the IGHV gene of a sequence can be used *ipso facto* to distinguish between antigen-experienced and naïve B cells (84). However, especially in human-based experiments, these sequences may be elicited by numerous immunological incidents without link to the applied vaccination. Furthermore, several studies revealed a high degree of polymorphism in the human IG loci (12, 64, 354–356), complicating a correct assessment of somatic hypermutation levels. The sequences identified as antigen-driven in **Chapter 2** showed no difference in SHM profiles compared to the unselected sequences (data not shown). However, the library preparation used in our study is based on RNA, resulting in a strong overrepresentation of plasma cell transcripts in the dataset (58). In addition to that, samples were amplified with a 30 cycle PCR, further increasing the ratio of plasma cell sequences compared to those from other B cell subtypes. It can therefore be readily assumed, that the majority of sequences in this experiment were derived from antigen-experienced plasma cells, explaining the homogenous level of SHM across all sequences.

Recently, DeKosky and colleagues developed a technology for sequencing paired IGH-IGL transcripts, using a single-cell emulsion PCR (303). This method achieved a throughput of ~10^6 B cells per experiment, with a pairing precision of 97%. Another interesting approach, published by Howie and coworkers, employed combinatorics rather than physical linking of corresponding IGH and IGL information (357). Briefly, this “pairSEQ” protocol described the sorting of a fixed number of cells into 96-well plates, and an in-well cell lysis and library preparation with barcodes labeling sequences by their well. Post-sequencing reads are then de-multiplexed to map each sequence back to the original wells. Because of the hIGHDiversity of the utilized T cells, it is highly improbable, that two different clones

would occupy the exact same set of wells. Consequently, a pair of IGH and IGL sequences found in several wells, is highly likely originating from the same original clone. As multiple cells can be sequences per well, the throughput of this method is very high, reaching up to 2,000,000 pairs from a single 96-well plate. So far the pairSEQ technique has only been used for T cell repertoire sequencing (357), which exhibits considerably lower diversity than B cell repertoires (55). However, as the method itself stems from diversity, it is certainly possible to apply it also for IG repertoire sequencing.

4.5. Future work

In the course of this thesis the Ion Torrent PGM was established as HTS platform to assess IGH repertoire transcripts of mouse and OmniRat™. In addition, novel sample and data processing methods were developed and applied to characterize the $CD5^{+}$ B cell expansion of $A20^{BKO}sCYLD^{BOE}$ mouse model and to investigate IG repertoire responses to vaccination. While these approaches paved the path to analyze and dissect the IG repertoire, further work is required to improve our understanding the IG repertoire in health and disease.

Chapter 2 describes an innovative approach to select antigen-driven IGH from bulk sequenced bone marrow B cell transcripts. In this regard, antigen-driven IGH were identified by their CDR3 amino acids, overrepresented in animals of a vaccination group and absent in control and unrelated vaccination groups. According to Xu and coworkers, the sequence of a CDR3 is sufficient to describe the specificity of an IG molecule (3). Therefore, these antigen-linked CDR3s can be seen as signatures for the underlying antigenic exposure. One of these signatures, consisting of 3 highly similar CDR3s varying at two positions (‘ARH[M/R]T[F/Y]YYGSGSPNFDY‘) was found in the vaccination group receiving MVA expressing measles virus (MV) hemagglutinin (H) and fusion (F) proteins via intra peritoneal (i.p.) injection. The same three CDR3s were also found in hybridoma cell-lines prepared from OmniRat™ lymph node B cells after vaccinating them with whole MV antigens via footpad immunization (177). This is a strong confirmation, that this CDR3 is specific for measles virus H or F protein, which are the only overlapping antigens in both studies. In a recent experiment, our group vaccinated OmniRat™ intra muscular (i.m.) with an experimental anti-MV vaccine using DNA vaccination carrier ICANtibodies™ (358) causing in vivo transfected cells to express only the MV-H proteins (359–362). The study contained vaccines prepared from two genetically distant MV clades, the Edmonston strain or clade A, also used in the current measles vaccine, and clade D11 which expressed a structurally different H protein (363). While the data is still being processed, a first assessment revealed, animals from both

MV-H vaccination groups but none of the two control groups in this study produced IGH sequences with the same CDR3 signature (‘ARH[M/R]T[F/Y]YYGSGSPNFDY‘). It can be therefore concluded, that first, the signature is induced by the MV-H protein, second, that the expression is clade unspecific, third, the signature is expressed in B cells from different lymphatic organs (lymph nodes, bone marrow) and upon vaccination through different administration routes (footpad, i.p., i.m.). In addition to that, the neutralizing serum response against MV is mainly targeting the MV-H protein, with > 96% of all antibodies directed against a certain epitope (327, 328). Taken together this strongly underlines the existence of a convergent cellular and serum IG response against MV-H protein. Future work could determine, if the antibodies carrying the determined MV-H CDR3 signature (‘ARH[M/R]T[F/Y]YYGSGSPNFDY‘) are indeed targeting a dominant epitope on the MV-H protein, providing a biological foundation for the IG repertoire convergence. Furthermore, the immune reaction against MV is known to be cross-protective against all genotypes (364), but sera from vaccines and naturally infected individuals differ in their neutralization activity towards wild-type MV strains (365–367). The in-depth investigation of IG repertoires in response to MV vaccination and infection with HTS could determine, if there are distinct IG elicited against specific genotypes causing the different serum reactions of infected and vaccinated individuals. Consequently, MV signatures have to be identified in humans. As a first step, it should be determined if the OmniRat™ signatures are also present in humans after MV vaccination, which requires MV-naïve donors as control. While in the western world, the high rate of MV vaccination makes this task unfeasible, our group has recruited a group of LAO PDR individuals, which were determined naive for measles virus serum antibody levels. PBMCs were collected before and after administering a standard MV vaccine and the IG repertoire determined by HTS on our PGM platform. The samples are currently being processed through our bioinformatics pipeline. In this regard, it is of primary interest to first determine the full germline repertoire of the Lao PDR study subjects. During IG repertoire data processing, the level of SHM, and thus antigen-exposure is determined through reference alignment as divergence form the germline genes. However, there is increased evidence, that individuals express a very private germline repertoire (354–356, 368), which can also be linked to their vaccine and disease response, as well as reduced IG clonal diversity, immune senescence and malignant B cell clone formation (368–370). For instance, Feeney and coworkers found that the Navajo population is more susceptible to *Haemophilius influenzae*, because of a 4.5-fold reduced recombination efficacy of an otherwise common IGL κ gene (371). Further studies linked individual IG germline repertoires to vaccine and disease response, as well as reduced IG clonal diversity, immune senescence, and malignant B

cell clone formation (111, 354, 355, 368, 370). To determine individual germline repertoires during HTS data analysis, the Kleinstein Lab recently released a promising bioinformatics “Tool for Ig Genotype Elucidation via Rep-Seq” (TIgGER, (348)). TIgGER enables to identify new IG germline alleles by detecting reoccurring mutation patterns from known germline genes (348). In addition to the LAO PDR MV vaccination follow-up introduced above, a cooperative project was carried out recently by our group in Laos: 26 healthcare workers were vaccinated against hepatitis B and their IG repertoire sequenced before and after the first, second and third vaccination. Together with the LAO PDR MV study, this offers the opportunity to assess whether the germline genes of Lao-people differs from the standard germline genes as provided by IMGT, and to discover new human IG germline gene alleles. This requires a thorough sample preparation and data analysis method for the utilized Ion Torrent PGM sequencing platform. Our group developed a PGM-tailored RNA barcoding approach, allowing to retrieve IGH sequences with sufficient confidence (**Chapter 4**). We showed that considering only those sequences from datasets with sufficient sequencing depth, adequate ssUID distribution and copy allowed to obtain correct IGH sequences with over 99.5% confidence. Our pipeline relies on the error detection and correction of IMGT, which was proven to be highly efficient when detecting indels through frameshifts compared to germline genes. As TIgGer also employs IMGT reference alignments as input, the data produced by out approach is ready to use for identification of possible new IG germline alleles.

Canonical NF-κB activation is one of the hallmarks of the pathogenesis of CLL, promoting excessive proliferation of $CD5^+$ B cells via up-regulation of anti-apoptotic proteins. In **Chapter 3** describes our network analysis approach on HTS splenocyte RNA transcripts revealed a unmutated CLL-like expansion of B cells within the $A20^{BKO}sCYLD^{BOE}$ mouse model. In the context of this project, the expression of the short alternative splicing variant of tumor suppressor gene CYLD could be linked to the excessive $CD5^+$ B cell expansion within the peritoneal cavity and spleen of mice. The B1 cell-associated tumor formations these mice show high similarity of human CLL and there is evidence, that the sCYLD expression also plays a role in human CLL (215, 223, 265). Future research should determine the underlying mechanism of sCYLD mediated constitutive NF-kappa activation and their role in human B cell. In this regard, the $A20^{BKO}sCYLD^{BOE}$ mouse model could provide the means to understand the link between RIP1 ubiquination in the context of deregulated CYLD expression and NF-κB activation, which is the proposed mechanism for ubiquitination-related CLL (265). Furthermore, the mouse model can be used as a preclinical model to test new approaches and therapies for CLL treatment and monitoring.

References

1. Tonegawa, S. 1983. Somatic generation of antibody diversity. *Nature* 302: 575–581.

2. Paige, C. J., and G. E. Wu. 1989. The B cell repertoire. *FASEB J.* 3: 1818–1824.

3. Xu, J. L., and M. M. Davis. 2000. Diversity in the CDR3 Region of V H Is Sufficient for Most Antibody Specificities. *Immunity* 13: 37–45.

4. Ippolito, G. C., R. L. Schelonka, M. Zemlin, I. I. Ivanov, R. Kobayashi, C. Zemlin, G. L. Gartland, L. Nitschke, J. Pelkonen, K. Fujihashi, K. Rajewsky, and H. W. Schroeder. 2006. Forced usage of positively charged amino acids in immunoglobulin CDR-H3 impairs B cell development and antibody production. *J. Exp. Med.* 203: 1567–78.

5. Woof, J. M., and D. R. Burton. 2004. Human antibody-Fc receptor interactions illuminated by crystal structures. *Nat. Rev. Immunol.* 4: 89–99.

6. Alt, F. W., A. L. M. Bothwell, M. Knapp, E. Siden, E. Mather, M. Koshland, and D. Baltimore. 1980. Synthesis of secreted and membrane-bound immunoglobulin mu heavy chains is directed by mRNAs that differ at their 3' ends. *Cell* 20: 293–301.

7. Schroeder, H. W. J., and L. Cavacini. 2010. Structure and Function of Immunoglobulins (author manuscript). *J. Allergy Clin. Immunol.* 125: S41–S52.

8. Burton, D. R., and J. M. Woof. 1992. Human antibody effector function. *Adv. Immunol.* 51: 1–84.

9. Schatz, D. G., and P. C. Swanson. 2011. V(D)J Recombination: Mechanisms of Initiation. *Annu. Rev. Genet.* 45: 167–202.

10. Brady, B. L., N. C. Steinel, and C. H. Bassing. 2010. Antigen Receptor Allelic Exclusion: An Update and Reappraisal. *J. Immunol.* 185: 3801–3808.

11. Janeway, C. A., P. Travers, M. Walport, and M. Shlomchik. 2001. *Immunobiology: The Immune System In Health And Disease*,. Garland Science, New York.

12. Zehnder, L., A. M. Collins, K. C. Nadeau, M. Egholm, D. B. Miklos, J. Birgitte, B. Simen, B. Hanczaruk, K. D. Nguyen, L. N. Zhang, B. Sahaf, C. D. Jones, A. Z. Fire, E. L. Marshall, J. D. Merker, J. Scott, D. Boyd, B. A. Gaëta, K. J. Jackson, S. D. Boyd, J. M. Maniar, B. B. Simen, and J. L. Zehnder. 2010. Gene Rearrangements Repertoire Inferred from Variable Region Individual Variation in the Germline Ig Gene Individual Variation in the Germline Ig Gene Repertoire Inferred from Variable Region Gene Rearrangements. *J. Immunol.* 184: 6986–6992.

13. Lefranc, M.-P. 2004. IMGT, the international ImMunoGeneTics information system(R). *Nucleic Acids Res.* 33: D593–D597.

14. Kalinina, O., C. M. Doyle-Cooper, J. Miksanek, W. Meng, E. L. Prak, and M. G. Weigert. 2011. Alternative mechanisms of receptor editing in autoreactive B cells. *Proc. Natl. Acad. Sci. U. S. A.* 108: 7125–30.

15. MacLennan, I. C. M. 1994. Germinal Centers. *Ann Rev Immunol* 12: 117–39.

16. Shlomchik, M. J., and F. Weisel. 2012. Germinal center selection and the development of memory B and plasma cells. *Immunol. Rev.* 247: 52–63.

17. McHeyzer-Williams, L. J., and M. G. McHeyzer-Williams. 2004. Antigen-Specific Memory B Cell Development. *Annu. Rev. Immunol.* 23: 487–513.

18. McKean, D., K. Huppi, M. Bell, L. Staudt, W. Gerhard, and M. Weigert. 1984. Generation of antibody diversity in the immune response of BALB/c mice to influenza virus hemagglutinin. *Proc. Natl. Acad. Sci.* 81: 3180–3184.

19. Kleinstein, S. H., Y. Louzoun, and M. J. Shlomchik. 2003. Estimating hypermutation rates from clonal tree data. *J. Immunol.* 171: 4639–4649.

20. Lin, M. M., M. Zhu, and M. D. Scharff. 1997. Sequence dependent hypermutation of the

immunoglobulin heavy chain in cultured B cells. In *Proceedings of the National Academy of Sciences of the United States of America* vol. 94. 5284–9.

21. Gojobori, T., and M. Nei. 1986. Relative contributions of germline gene variation and somatic mutation to immunoglobulin diversity in the mouse. *Mol Biol Evol* 3: 156–167.

22. Griffiths, G. M., C. Berek, M. Kaartinen, and C. Milstein. 1984. Somatic mutation and the maturation of immune responses to 2-phenyl oxazolon. *Nature* 312: 271–275.

23. Eisen, H. N., and G. W. Siskind. 1964. Variations in Affinities of Antibodies During the Immune Response. *Biochemistry* 3: 996–1008.

24. Manz, R. A., A. Thiel, and A. Radbruch. 1997. Lifetime of plasma cells in the bone marrow. *Nature* 388: 133–134.

25. Bernasconi, N. L. 2010. Activation of Human Memory B Cells Maintenance of Serological Memory by Polyclonal Activation of Human Memory B Cells. *Science (80).* 2199: 10–14.

26. Hibi, T., H. M. Dosch, and T. Ig. 1986. Limiting dilution analysis of the B cell compartment in human bone marrow. *Eur. J. Immunol.* 16: 139–45.

27. Bhoj, V. G., D. Arhontoulis, G. Wertheim, J. Capobianchi, C. A. Callahan, C. T. Ellebrecht, A. E. Obstfeld, S. F. Lacey, J. J. Melenhorst, F. Nazimuddin, W.-T. Hwang, S. L. Maude, M. A. Wasik, A. Bagg, S. Schuster, M. D. Feldman, D. L. Porter, S. A. Grupp, C. H. June, and M. C. Milone. 2016. Persistence of long-lived plasma cells and humoral immunity in individuals responding to CD19-directed CAR T cell therapy. *Blood* 128: blood-2016-01-694356.

28. Halliley, J. L., C. M. Tipton, J. Liesveld, A. F. Rosenberg, J. Darce, I. V Gregoretti, L. Popova, D. Kaminiski, C. F. Fucile, I. Albizua, S. Kyu, K.-Y. Chiang, K. T. Bradley, R. Burack, M. Slifka, E. Hammarlund, H. Wu, L. Zhao, E. E. Walsh, A. R. Falsey, T. D. Randall, W. C. Cheung, I. Sanz, and F. E.-H. Lee. 2015. Long-Lived Plasma Cells Are Contained within the CD19(-)CD38(hi)CD138(+) Subset in Human Bone Marrow. *Immunity* 43: 132–45.

29. Liu, Y. J., S. Oldfield, and I. C. MacLennan. 1988. Memory B cells in T cell-dependent antibody responses colonize the splenic marginal zones. *Eur. J. Immunol.* 18: 355–62.

30. Tangye, S. G., Y. J. Liu, G. Aversa, J. H. Phillips, and J. E. de Vries. 1998. Identification of functional human splenic memory B cells by expression of CD148 and CD27. *J Exp Med* 188: 1691–1703.

31. Tangye, S. G., and K. L. Good. 2007. Human IgM+CD27+ B Cells: Memory B Cells or “Memory” B Cells? *J. Immunol.* 179: 13–19.

32. Seidman, J. G., A. Leder, M. H. Edgell, F. Polsky, S. M. Tilghman, D. C. Tiemeier, and P. Leder. 1978. Multiple related immunoglobulin variable-region genes identified by cloning and sequence analysis. *Proc. Natl. Acad. Sci. U. S. A.* 75: 3881–5.

33. Ehlich, A., V. Martin, W. Müller, and K. Rajewsky. 1994. Analysis of the B-cell progenitor compartment at the level of single cells. *Curr. Biol.* 4: 573–583.

34. Klein, U., K. Rajewsky, and R. Küppers. 1998. Human Immunoglobulin (Ig)M+ IgD+ Peripheral Blood B Cells Expressing the CD27 Cell Surface Antigen Carry Somatically Mutated Variable Region Genes: CD27 as a General Marker for Somatically Mutated (Memory) B Cells. *J. Exp. Med* 188: 1679–1689.

35. Küppers, R., M. Zhao, M. L. Hansmann, and K. Rajewsky. 1993. Tracing B cell development in human germinal centres by molecular analysis of single cells picked from histological sections. *EMBO J.* 12: 4955–4967.

36. Tiller, T., M. Tsuiji, S. Yurasov, K. Velinzon, M. C. Nussenzweig, and H. Wardemann. 2007. Autoreactivity in human IgG+ memory B cells. *Immunity* 26: 205–13.

37. Corti, D., J. Voss, S. J. Gamblin, G. Codoni, A. Macagno, D. Jarrossay, S. G. Vachieri, D. Pinna, A. Minola, F. Vanzetta, C. Silacci, B. M. Fernandez-Rodriguez, G. Agatic, S. Bianchi, I. Giacchetto-Sasselli, L. Calder, F. Sallusto, P. Collins, L. F. Haire, N. Temperton, J. P. M. Langedijk, J. J. Skehel, and A. Lanzavecchia. 2011. A Neutralizing Antibody Selected from Plasma Cells That Binds to Group 1 and Group 2 Influenza A Hemagglutinins. *Science (80).* 333: 850–6.

38. Corti, D., and A. Lanzavecchia. 2013. Broadly neutralizing antiviral antibodies. *Annu. Rev. Immunol.*, 31: 705-42.

39. Traggiai, E., S. Becker, K. Subbarao, L. Kolesnikova, Y. Uematsu, M. R. Gismondo, B. R. Murphy, R. Rappuoli, and A. Lanzavecchia. 2004. An efficient method to make human monoclonal antibodies from memory B cells: potent neutralization of SARS coronavirus. *Nat. Med.* 10: 871–875.

40. Wrammert, J., K. Smith, J. Miller, W. A. Langley, K. Kokko, C. Larsen, N.-Y. Zheng, I. Mays, L. Garman, C. Helms, J. James, G. M. Air, J. D. Capra, R. Ahmed, and P. C. Wilson. 2008. Rapid cloning of high-affinity human monoclonal antibodies against influenza virus. *Nature* 453: 667–71.

41. Walker, L. M., S. K. Phogat, P.-Y. Chan-Hui, D. Wagner, P. Phung, J. L. Goss, T. Wrin, M. D. Simek, S. Fling, J. L. Mitcham, J. K. Lehrman, F. H. Priddy, O. A. Olsen, S. M. Frey, P. W. Hammond, P. G. P. Protocol G Principal Investigators, S. Kaminsky, T. Zamb, M. Moyle, W. C. Koff, P. Poignard, and D. R. Burton. 2009. Broad and potent neutralizing antibodies from an African donor reveal a new HIV-1 vaccine target. *Science* 326: 285–9.

42. Scheid, J. F., H. Mouquet, N. Feldhahn, M. S. Seaman, K. Velinzon, J. Pietzsch, R. G. Ott, R. M. Anthony, H. Zebroski, A. Hurley, A. Phogat, B. Chakrabarti, Y. Li, M. Connors, F. Pereyra, B. D. Walker, H. Wardemann, D. Ho, R. T. Wyatt, J. R. Mascola, J. V Ravetch, and M. C. Nussenzweig. 2009. Broad diversity of neutralizing antibodies isolated from memory B cells in HIV-infected individuals. *Nature* 458: 636–640.

43. Burton, D. R. 2012. A Blueprint for HIV Vaccine Discovery. *Cell Host Microbe* 12: 396–407.

44. Beltramello, M., K. L. Williams, C. P. Simmons, A. MacAgno, L. Simonelli, N. T. H. Quyen, S. Sukupolvi-Petty, E. Navarro-Sanchez, P. R. Young, A. M. De Silva, F. A. Rey, L. Varani, S. S. Whitehead, M. S. Diamond, E. Harris, A. Lanzavecchia, and F. Sallusto. 2010. The human immune response to dengue virus is dominated by highly cross-reactive antibodies endowed with neutralizing and enhancing activity. *Cell Host Microbe* 8: 271–283.

45. Mesin, L., L. M. Sollid, and R. Di Niro. 2012. The intestinal B-cell response in celiac disease. *Front. Immunol.* 3: 313.

46. Meffre, E., and H. Wardemann. 2008. B-cell tolerance checkpoints in health and autoimmunity. *Curr. Opin. Immunol.* 20: 632–638.

47. Scheid, J. F., H. Mouquet, J. Kofer, S. Yurasov, M. C. Nussenzweig, and H. Wardemann. 2011. Differential regulation of self-reactivity discriminates between IgG+ human circulating memory B cells and bone marrow plasma cells. *Proc. Natl. Acad. Sci.* 108: 18044–18048.

48. Galson, J. D., A. J. Pollard, J. Trück, and D. F. Kelly. 2014. Studying the antibody repertoire after vaccination: Practical applications. *Trends Immunol.* 35: 319–331.

49. Poulsen, T. R., A. Jensen, J. S. Haurum, S. Andersen, T. R. Poulsen, A. Jensen, J. S. Haurum, and P. S. Andersen. 2011. Limits for Antibody Affinity Maturation and Repertoire Diversification in Hypervaccinated Humans. *J. Immunol.* 187.

50. Frolich, D., C. Giesecke, H. E. Mei, K. Reiter, C. Daridon, P. E. Lipsky, and T. Dorner. 2010. Secondary Immunization Generates Clonally Related Antigen-Specific Plasma Cells and Memory B Cells. *J. Immunol.* 185: 3103–3110.

51. Smith, K., J. J. Muther, A. L. Duke, E. McKee, N.-Y. Y. Zheng, P. C. Wilson, and J. A. James. 2013. Fully human monoclonal antibodies from antibody secreting cells after vaccination with Pneumovax®23 are serotype specific and facilitate opsonophagocytosis. *Immunobiology* 218: 745–754.

52. Kolibab, K., S. L. Smithson, A. K. Shriner, S. Khuder, S. Romero-Steiner, G. M. Carlone, and M. A. J. Westerink. 2005. Immune response to pneumococcal polysaccharides 4 and 14 in elderly and young adults. I. Antibody concentrations, avidity and functional activity. *Immun. Ageing* 2: 10.

53. Adderson, E. E., P. G. Shackelford, A. Quinn, P. M. Wilson, M. W. Cunningham, R. A. Insel, and W. L. Carroll. 1993. Restricted immunoglobulin VH usage and VDJ combinations in the human response to haemophilus influenzae type b capsular polysaccharide. *J.Clin.Invest.* 91: 2734–2743.

54. Pinchuck, G. V., C. Nottenburg, and E. C. B. Milner. 1995. Predominant V-region gene configurations in the human antibody response to Haemophilus influenzae capsule polysaccharide. *Scand. J. Immunol.*

41: 324–330.

55. Benichou, J., R. Ben-Hamo, Y. Louzoun, and S. Efroni. 2012. Rep-Seq: Uncovering the immunological repertoire through next-generation sequencing. *Immunology* 135: 183–191.

56. Georgiou, G., G. C. Ippolito, J. Beausang, C. E. Busse, H. Wardemann, and S. R. Quake. 2014. The promise and challenge of high-throughput sequencing of the antibody repertoire. *Nat. Biotechnol.* 32: 158–168.

57. Baum, P. D., V. Venturi, and D. A. Price. 2012. Wrestling with the repertoire: The promise and perils of next generation sequencing for antigen receptors. *Eur. J. Immunol.* 42: 2834–2839.

58. Klein, U., R. Kuppers, and K. Rajewsky. 1997. Evidence for a large compartment of IgM-expressing memory B cells in humans. *Blood* 89: 1288–1298.

59. van Dongen, J. J. M., A. W. Langerak, M. Brüggemann, P. A. S. Evans, M. Hummel, F. L. Lavender, E. Delabesse, F. Davi, E. Schuuring, R. García-Sanz, J. H. J. M. van Krieken, J. Droese, D. González, C. Bastard, H. E. White, M. Spaargaren, M. González, A. Parreira, J. L. Smith, G. J. Morgan, M. Kneba, and E. A. Macintyre. 2003. Design and standardization of PCR primers and protocols for detection of clonal immunoglobulin and T-cell receptor gene recombinations in suspect lymphoproliferations: report of the BIOMED-2 Concerted Action BMH4-CT98-3936. *Leukemia* 17: 2257–2317.

60. van Krieken, J. H. J. M., a W. Langerak, E. a Macintyre, M. Kneba, E. Hodges, R. G. Sanz, G. J. Morgan, A. Parreira, T. J. Molina, J. Cabeçadas, P. Gaulard, B. Jasani, J. F. Garcia, M. Ott, M. L. Hannsmann, F. Berger, M. Hummel, F. Davi, M. Brüggemann, F. L. Lavender, E. Schuuring, P. a S. Evans, H. White, G. Salles, P. J. T. a Groenen, P. Gameiro, C. Pott, and J. J. M. Van Dongen. 2007. Improved reliability of lymphoma diagnostics via PCR-based clonality testing: report of the BIOMED-2 Concerted Action BHM4-CT98-3936. *Leukemia* 21: 201–206.

61. Evans, P. A. S., C. Pott, P. J. T. A. Groenen, G. Salles, F. Davi, F. Berger, J. F. Garcia, J. H. J. M. van Krieken, S. Pals, P. Kluin, E. Schuuring, M. Spaargaren, E. Boone, D. González, B. Martinez, R. Villuendas, P. Gameiro, T. C. Diss, K. Mills, G. J. Morgan, G. I. Carter, B. J. Milner, D. Pearson, M. Hummel, W. Jung, M. Ott, D. Canioni, K. Beldjord, C. Bastard, M. H. Delfau-Larue, J. J. M. van Dongen, T. J. Molina, and J. Cabeçadas. 2007. Significantly improved PCR-based clonality testing in B-cell malignancies by use of multiple immunoglobulin gene targets. Report of the BIOMED-2 Concerted Action BHM4-CT98-3936. *Leukemia* 21: 207–214.

62. Brüggemann, M., H. White, P. Gaulard, R. Garcia-Sanz, P. Gameiro, S. Oeschger, B. Jasani, M. Ott, G. Delsol, a Orfao, M. Tiemann, H. Herbst, a W. Langerak, M. Spaargaren, E. Moreau, P. J. T. a Groenen, C. Sambade, L. Foroni, G. I. Carter, M. Hummel, C. Bastard, F. Davi, M.-H. Delfau-Larue, M. Kneba, J. J. M. van Dongen, K. Beldjord, and T. J. Molina. 2007. Powerful strategy for polymerase chain reaction-based clonality assessment in T-cell malignancies Report of the BIOMED-2 Concerted Action BHM4 CT98-3936. *Leuk. Off. J. Leuk. Soc. Am. Leuk. Res. Fund, U.K* 21: 215–221.

63. Campbell, P. J., E. D. Pleasance, P. J. Stephens, E. Dicks, R. Rance, I. Goodhead, G. a Follows, A. R. Green, P. A. Futreal, and M. R. Stratton. 2008. Subclonal phylogenetic structures in cancer revealed by ultra-deep sequencing. In *Proceedings of the National Academy of Sciences* vol. 105. 13081–13086.

64. Boyd, S. D., E. L. Marshall, J. D. Merker, J. M. Maniar, L. N. Zhang, B. Sahaf, C. D. Jones, B. B. Simen, B. Hanczaruk, K. D. Nguyen, K. C. Nadeau, M. Egholm, D. B. Miklos, J. L. Zehnder, and A. Z. Fire. 2009. Measurement and clinical monitoring of human lymphocyte clonality by massively parallel VDJ pyrosequencing. *Sci Transl Med* 1: 12ra23.

65. Polz, M. F., and C. M. Cavanaugh. 1998. Bias in template-to-product ratios in multitemplate PCR. *Appl. Environ. Microbiol.* 64: 3724–3730.

66. Sambrook, J., and D. W. Russell. 2006. Rapid Amplification of 5′ cDNA Ends (5′-RACE). *Cold Spring Harb. Protoc.* 2006: pdb.prot3989.

67. Bashford-Rogers, R. J., A. L. Palser, S. F. Idris, L. Carter, M. Epstein, R. E. Callard, D. C. Douek, G. S. Vassiliou, G. a Follows, M. Hubank, and P. Kellam. 2014. Capturing needles in haystacks: a comparison of B-cell receptor sequencing methods. *BMC Immunol.* 15: 29.

68. He, L., D. Sok, P. Azadnia, J. Hsueh, E. Landais, M. Simek, W. C. Koff, P. Poignard, D. R. Burton,

and J. Zhu. 2015. Toward a more accurate view of human B-cell repertoire by next-generation sequencing, unbiased repertoire capture and single-molecule barcoding. *Sci. Rep.* 4: 6778.

69. Freeman, J. D., R. L. Warren, J. R. Webb, B. H. Nelson, and R. A. Holt. 2009. Profiling the T-cell receptor beta-chain repertoire by massively parallel sequencing. *Genome Res.* 19: 1817–1824.

70. Warren, R. L., J. D. Freeman, T. Zeng, G. Choe, S. Munro, R. Moore, J. R. Webb, and R. A. Holt. 2011. Exhaustive T-cell repertoire sequencing of human peripheral blood samples reveals signatures of antigen selection and a directly measured repertoire size of at least 1 million clonotypes. *Genome Res.* 21: 790–7.

71. Vollmers, C., R. V Sit, J. a Weinstein, C. L. Dekker, and S. R. Quake. 2013. Genetic measurement of memory B-cell recall using antibody repertoire sequencing. *Proc. Natl. Acad. Sci. U. S. A.* 110: 13463–8.

72. Turchaninova, M. A., A. Davydov, O. V Britanova, M. Shugay, V. Bikos, E. S. Egorov, V. I. Kirgizova, E. M. Merzlyak, D. B. Staroverov, D. A. Bolotin, I. Z. Mamedov, M. Izraelson, M. D. Logacheva, O. Kladova, K. Plevova, S. Pospisilova, and D. M. Chudakov. 2016. High-quality full-length immunoglobulin profiling with unique molecular barcoding. *Nat Protoc.* 11: 1599–1616.

73. Shugay, M., O. V Britanova, E. M. Merzlyak, M. a Turchaninova, I. Z. Mamedov, T. R. Tuganbaev, D. a Bolotin, D. B. Staroverov, E. V Putintseva, K. Plevova, C. Linnemann, D. Shagin, S. Pospisilova, S. Lukyanov, T. N. Schumacher, and D. M. Chudakov. 2014. Towards error-free profiling of immune repertoires. *Nat. Methods* 11: 653–655.

74. Egorov, E. S., E. M. Merzlyak, A. A. Shelenkov, O. V Britanova, G. V Sharonov, D. B. Staroverov, D. A. Bolotin, A. N. Davydov, E. Barsova, Y. B. Lebedev, M. Shugay, and D. M. Chudakov. 2015. Quantitative Profiling of Immune Repertoires for Minor Lymphocyte Counts Using Unique Molecular Identifiers. *J. Immunol.* 194: 6155–6163.

75. Bragg, L. M., G. Stone, M. K. Butler, P. Hugenholtz, and G. W. Tyson. 2013. Shining a Light on Dark Sequencing: Characterising Errors in Ion Torrent PGM Data. *PLoS Comput. Biol.* 9: e1003031.

76. Khan, T. A., S. Friedensohn, A. R. G. de Vries, J. Straszewski, H.-J. Ruscheweyh, and S. T. Reddy. 2016. Accurate and predictive antibody repertoire profiling by molecular amplification fingerprinting. *Sci. Adv.* 2: e1501371–e1501371.

77. Reddy, S. T., X. Ge, A. E. Miklos, R. A. Hughes, S. H. Kang, K. H. Hoi, C. Chrysostomou, S. P. Hunicke-Smith, B. L. Iverson, P. W. Tucker, A. D. Ellington, and G. Georgiou. 2010. Monoclonal antibodies isolated without screening by analyzing the variable-gene repertoire of plasma cells. *Nat. Biotechnol.* 28: 965–9.

78. DeKosky, B. J., G. C. Ippolito, R. P. Deschner, J. J. Lavinder, Y. Wine, B. M. Rawlings, N. Varadarajan, C. Giesecke, T. Dörner, S. F. Andrews, P. C. Wilson, S. P. Hunicke-Smith, C. G. Willson, A. D. Ellington, and G. Georgiou. 2013. High-throughput sequencing of the paired human immunoglobulin heavy and light chain repertoire. *Nat Biotechnol* 31: 166–169.

79. Busse, C., I. Czogiel, and P. Braun. 2013. Single-cell based high-throughput sequencing of full-length immunoglobulin heavy and light chain genes. *Eur. J. Immunol.* 1–23.

80. Loman, N. J., R. V Misra, T. J. Dallman, C. Constantinidou, S. E. Gharbia, J. Wain, and M. J. Pallen. 2012. Performance comparison of benchtop high-throughput sequencing platforms. *Nat. Biotechnol.* 30: 434–9.

81. Quail, M. M., M. E. Smith, P. Coupland, T. D. T. Otto, S. R. S. Harris, T. R. Connor, A. Bertoni, H. H. P. Swerdlow, Y. Gu, J. Rothberg, W. Hinz, T. Rearick, J. Schultz, W. Mileski, M. Davey, J. Leamon, K. Johnson, M. Milgrew, M. Edwards, J. Eid, A. Fehr, J. Gray, K. Luong, J. Lyle, G. Otto, P. Peluso, D. Rank, P. Baybayan, B. Bettman, D. Bentley, S. Balasubramanian, H. H. P. Swerdlow, G. Smith, J. Milton, C. Brown, K. Hall, D. Evers, C. Barnes, H. Bignell, I. Kozarewa, Z. Ning, M. M. Quail, M. Sanders, M. Berriman, D. Turner, M. M. Quail, T. D. T. Otto, Y. Gu, S. R. S. Harris, T. Skelly, J. McQuillan, H. H. P. Swerdlow, S. Oyola, F. Syed, H. Grunenwald, N. Caruccio, H. Lam, M. Clark, R. Chen, R. Chen, G. Natsoulis, M. O'Huallachain, F. Dewey, L. Habegger, T. Carver, S. R. S. Harris, M. Berriman, J. Parkhill, J. McQuillan, N. Ponsting, Z. Ning, T. D. T. Otto, M. Sanders, M. Berriman, C. Newbold, K. Nakamura, T. Oshima, T. Morimoto, S. Ikeda, H. Yoshikawa, Y. Shiwa, S. Ishikawa, M. Linak, A. Hirai, H. Takahashi, B. Diep, S. Gill, R. Chang, T. Phan, J. Chen, M. Davidson, F. Lin, J. Lin, H. Carleton, E. Mongodin, E.

Achidi, M. Gardner, N. Hall, E. Fung, O. White, M. Berriman, R. Hyman, J. Carlton, A. Pain, K. Nelson, S. Bowman, M. Choi, U. Scholl, W. Ji, T. Liu, I. Tikhonova, P. Zumbo, A. Nayir, A. Bakkaloglu, S. Ozen, S. Sanjad, T. Down, V. Rakyan, D. Turner, P. Flicek, H. Li, E. Kulesha, S. Graf, N. Johnson, J. Herrero, E. Tomazou, P. Giresi, J. Kim, R. McDaniell, V. Iyer, J. Lieb, D. Johnson, A. Mortazavi, R. Myers, B. Wold, G. Langridge, M. Phan, D. Turner, T. Perkins, L. Parts, J. Haase, I. Charles, D. Maskell, S. Peters, G. Dougan, D. Licatalosi, A. Mele, J. Fak, J. Ule, M. Kayikci, S. Chi, T. Clark, A. Schweitzer, J. Blume, X. Wang, L. Mamanova, R. Andrews, K. James, E. Sheridan, P. Ellis, C. Langford, T. Ost, J. Collins, D. Turner, S. Myllykangas, J. Buenrostro, G. Natsoulis, J. Bell, H. Ji, N. Shao, H. Hu, Z. Yan, Y. Xu, H. Hu, C. Menzel, N. Li, W. Chen, P. Khaitovich, Z. Wang, M. Gerstein, M. Snyder, S. Gnerre, I. Maccallum, D. Przybylski, F. Ribeiro, J. Burton, B. Walker, T. Sharpe, G. Hall, T. Shea, S. Sykes, J. Levin, M. Yassour, X. Adiconis, C. Nusbaum, D. Thompson, N. Friedman, A. Gnirke, A. Regev, A. Adey, Asan, X. Xun, J. Kitzman, E. Turner, B. Stackhouse, A. MacKenzie, N. Caruccio, X. Zhang, B. Flusberg, D. Webster, J. Lee, K. Travers, E. Olivares, T. Clark, J. Korlach, S. Turner, T. Holden, J. Lindsay, C. Corton, M. M. Quail, J. Cockfield, S. Pathak, R. Batra, J. Parkhill, S. Bentley, J. Edgeworth, H. Li, R. Durbin, H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin, S. Angiuoli, and S. Salzberg. 2012. A tale of three next generation sequencing platforms: comparison of Ion torrent, pacific biosciences and illumina MiSeq sequencers. *BMC Genomics* 13: 341.

82. Ross, M. G., C. Russ, M. Costello, A. Hollinger, N. J. Lennon, R. Hegarty, C. Nusbaum, and D. B. Jaffe. 2013. Characterizing and measuring bias in sequence data. *Genome Biol.* 14: R51.

83. Laehnemann, D., A. Borkhardt, and A. C. McHardy. 2016. Denoising DNA deep sequencing data-high-throughput sequencing errors and their correction. *Brief. Bioinform.* 17: 154–179.

84. Glanville, J., T. C. Kuo, H.-C. von Büdingen, L. Guey, J. Berka, P. D. Sundar, G. Huerta, G. R. Mehta, J. R. Oksenberg, S. L. Hauser, D. R. Cox, A. Rajpal, and J. Pons. 2011. Naive antibody gene-segment frequencies are heritable and unaltered by chronic lymphocyte ablation. *Proc. Natl. Acad. Sci. U. S. A.* 108: 20066–71.

85. Jiang, N., J. He, J. a Weinstein, L. Penland, S. Sasaki, X. S. He, C. L. Dekker, N. Y. Zheng, M. Huang, M. Sullivan, P. C. Wilson, H. B. Greenberg, M. M. Davis, D. S. Fisher, and S. R. Quake. 2013. Lineage Structure of the Human Antibody Repertoire in Response to Influenza Vaccination. *Sci. Transl. Med.* 5: 171ra19-171ra19.

86. Zhu, J., X. Wu, B. Zhang, K. McKee, S. O'Dell, C. Soto, T. Zhou, J. P. Casazza, J. C. Mullikin, P. D. Kwong, J. R. Mascola, L. Shapiro, J. Becker, B. Benjamin, R. Blakesley, G. Bouffard, S. Brooks, H. Coleman, M. Dekhtyar, M. Gregory, X. Guan, J. Gupta, J. Han, A. Hargrove, S. -l. Ho, T. Johnson, R. Legaspi, S. Lovett, Q. Maduro, C. Masiello, B. Maskeri, J. McDowell, C. Montemayor, J. Mullikin, M. Park, N. Riebow, K. Schandler, B. Schmidt, C. Sison, M. Stantripop, J. Thomas, P. Thomas, M. Vemulapalli, and A. Young. 2013. De novo identification of VRC01 class HIV-1-neutralizing antibodies by next-generation sequencing of B-cell transcripts. *Proc. Natl. Acad. Sci.* 110: E4088–E4097.

87. Wu, X., T. Zhou, J. Zhu, B. Zhang, I. Georgiev, C. Wang, X. Chen, N. S. Longo, M. Louder, K. McKee, S. O'Dell, S. Perfetto, S. D. Schmidt, W. Shi, L. Wu, Y. Yang, Z.-Y. Z. Yang, Z.-Y. Z. Yang, Z. Zhang, M. Bonsignori, J. A. Crump, S. H. Kapiga, N. E. Sam, B. F. Haynes, M. Simek, D. R. Burton, W. C. Koff, N. A. Doria-Rose, M. Connors, J. C. Mullikin, G. J. Nabel, M. Roederer, L. Shapiro, P. D. Kwong, and J. R. Mascola. 2011. Focused evolution of HIV-1 neutralizing antibodies revealed by structures and deep sequencing. *Science* 333: 1593–602.

88. Zhu, J., G. Ofek, Y. Yang, B. Zhang, M. K. Louder, G. Lu, K. McKee, M. Pancera, J. Skinner, Z. Zhang, R. Parks, J. Eudailey, K. E. Lloyd, J. Blinn, S. M. Alam, B. F. Haynes, M. Simek, D. R. Burton, W. C. Koff, J. C. Mullikin, J. R. Mascola, L. Shapiro, P. D. Kwong, N. C. S. NISC Comparative Sequencing Program, J. C. Mullikin, J. R. Mascola, L. Shapiro, P. D. Kwong, J. Becker, B. Benjamin, R. Blakesley, G. Bouffard, S. Brooks, H. Coleman, M. Dekhtyar, M. Gregory, X. Guan, J. Gupta, J. Han, A. Hargrove, S. Ho, T. Johnson, R. Legaspi, S. Lovett, Q. Maduro, C. Masiello, B. Maskeri, J. McDowell, C. Montemayor, J. C. Mullikin, M. Park, N. Riebow, K. Schandler, B. Schmidt, C. Sison, M. Stantripop, J. Thomas, P. Thomas, M. Vemulapalli, and A. Young. 2013. Mining the antibodyome for HIV-1-neutralizing antibodies with next-generation sequencing and phylogenetic pairing of heavy/light chains. *Proc. Natl. Acad. Sci. U. S. A.* 110: 6470–5.

89. Krause, J. C., T. Tsibane, T. M. Tumpey, C. J. Huffman, B. S. Briney, S. A. Smith, C. F. Basler, and J. E. Crowe. 2011. Epitope-Specific Human Influenza Antibody Repertoires Diversify by B Cell Intraclonal Sequence Divergence and Interclonal Convergence. *J. Immunol.* 187: 3704–3711.

90. Parameswaran, P., Y. Liu, K. M. Roskin, K. K. L. Jackson, V. P. Dixit, J. Y. Lee, K. L. Artiles, S. Zompi, M. J. Vargas, B. B. Simen, B. Hanczaruk, K. R. McGowan, M. A. Tariq, N. Pourmand, D. Koller, A. Balmaseda, S. D. Boyd, E. Harris, and A. Z. Fire. 2013. Convergent antibody signatures in human dengue. *Cell Host Microbe* 13: 691–700.

91. Mitchell, R., D. F. Kelly, A. J. Pollard, and J. Trück. 2014. Polysaccharide-specific B cell responses to vaccination in humans. *Hum. Vaccines Immunother.* 10: 1661–1668.

92. Trück, J., M. N. Ramasamy, J. D. Galson, R. Rance, J. Parkhill, G. Lunter, A. J. Pollard, and D. F. Kelly. 2015. Identification of Antigen-Specific B Cell Receptor Sequences Using Public Repertoire Analysis. *J. Immunol.* 194: 252–261.

93. Li, Y., Y. Liu, H. Chen, P. Wei, and F. Li. 2012. Three-dimensional structure-activity relationship modeling of cross-reactivities of a polyclonal antibody against pyrene by comparative molecular field analysis. *J Mol Biochem*. 1, 206-211.

94. Agathangelidis, A., N. Darzentas, A. Hadzidimitriou, X. Brochet, F. Murray, X. J. Yan, Z. Davis, E. J. Van Gastel-Mol, C. Tresoldi, C. C. Chu, N. Cahill, V. Giudicelli, B. Tichy, L. B. Pedersen, L. Foroni, L. Bonello, A. Janus, K. Smedby, A. Anagnostopoulos, H. Merle-Beral, N. Laoutaris, G. Juliusson, P. F. Di Celle, S. Pospisilova, J. Jurlander, C. Geisler, A. Tsaftaris, M. P. Lefranc, A. W. Langerak, D. G. Oscier, N. Chiorazzi, C. Belessi, F. Davi, R. Rosenquist, P. Ghia, and K. Stamatopoulos. 2012. Stereotyped B-cell receptors in one-third of chronic lymphocytic leukemia: A molecular classification with implications for targeted therapies. *Blood* 119: 4467–4475.

95. Darzentas, N., and K. Stamatopoulos. 2013. The significance of steroetyped B-cell receptors in chronic lymphocytic leukemia. *Hematol. Oncol. Clin. North Am.* 27: 237–250.

96. Messmer, B. T., E. Albesiano, D. G. Efremov, F. Ghiotto, S. L. Allen, J. Kolitz, R. Foa, R. N. Damle, F. Fais, D. Messmer, K. R. Rai, M. Ferrarini, and N. Chiorazzi. 2004. Multiple distinct sets of stereotyped antigen receptors indicate a role for antigen in promoting chronic lymphocytic leukemia. *J. Exp. Med.* 200: 519–525.

97. Rossi, D., and G. Gaidano. 2010. Biological and clinical significance of stereotyped B-cell receptors in chronic lymphocytic leukemia. *Haematologica* 95: 1992–1995.

98. Warren, E. H., F. A. Matsen IV, and J. Chou. 2013. High-throughput sequencing of B- And T-lymphocyte antigen receptors in hematology. *Blood* 122: 19–22.

99. Hoi, K. H., and G. C. Ippolito. 2013. Intrinsic bias and public rearrangements in the human immunoglobulin Vλ light chain repertoire. *Genes Immun.* 14: 1–6.

100. Arnaout, R., W. Lee, P. Cahill, T. Honan, T. Sparrow, M. Weiand, C. Nusbaum, K. Rajewsky, and S. B. Koralov. 2011. High-resolution description of antibody heavy-chain repertoires in humans. *PLoS One* 6: e22365.

101. Glanville, J., W. Zhai, J. Berka, D. Telman, G. Huerta, G. R. Mehta, I. Ni, L. Mei, P. D. Sundar, G. M. R. Day, D. Cox, A. Rajpal, and J. Pons. 2009. Precise determination of the diversity of a combinatorial antibody library gives insight into the human immunoglobulin repertoire. *Proc. Natl. Acad. Sci.* 106: 20216–20221.

102. Dimitrov, D. S. 2010. Therapeutic antibodies, vaccines and antibodyomes. *MAbs* 2: 347–356.

103. Feeney, a J., a Tang, and K. M. Ogwaro. 2000. B-cell repertoire formation: role of the recombination signal sequence in non-random V segment utilization. *Immunol. Rev.* 175: 59–69.

104. Perlmutter, R. M., J. F. Kearney, S. P. Chang, and L. E. Hood. 1985. Developmentally controlled expression of immunoglobulin VH genes. *Science* 227: 1597–601.

105. Berman, J. E., K. G. Nickerson, R. R. Pollock, J. E. Barth, R. K. B. Schuurman, D. M. Knowles, L. Chess, and F. W. Alt. 1991. VH gene usage in humans: biased usage of the VH6 gene in immature B lymphoid cells. *Eur. J. Immunol.* 21: 1311–1314.

106. Kalled, S. L., and P. H. Brodeur. 1990. Preferential rearrangement of V kappa 4 gene segments in pre-B cell lines. *J Exp Med* 172: 559–566.

107. Williams, J. V., J. H. Weitkamp, D. L. Blum, B. J. LaFleur, and J. E. Crowe. 2009. The human

neonatal B cell response to respiratory syncytial virus uses a biased antibody variable gene repertoire that lacks somatic mutations. *Mol. Immunol.* 47: 407–414.

108. Schroeder, H. W., L. Zhang, and J. B. Philips. 2001. Slow, programmed maturation of the immunoglobulin HCDR3 repertoire during the third trimester of fetal life. *Blood* 98: 2745–2751.

109. Gibson, K. L., Y.-C. C. Wu, Y. Barnett, O. Duggan, R. Vaughan, E. Kondeatis, B.-O. O. Nilsson, A. Wikby, D. Kipling, and D. K. Dunn-Walters. 2009. B-cell diversity decreases in old age and is correlated with poor health status. *Aging Cell* 8: 18–25.

110. Weksler, M. E. 2000. Changes in the B-cell repertoire with age. *Vaccine* 18: 1624–1628.

111. Wang, C., Y. Liu, L. T. Xu, K. J. L. Jackson, K. M. Roskin, T. D. Pham, J. Laserson, E. L. Marshall, K. Seo, J. Y. Lee, D. Furman, D. Koller, C. L. Dekker, M. M. Davis, a Z. Fire, and S. D. Boyd. 2014. Effects of Aging, Cytomegalovirus Infection, and EBV Infection on Human B Cell Repertoires. *J. Immunol.* 192: 603–611.

112. Weksler, M. E., and P. Szabo. 2000. The effect of age on the B-cell repertoire. *J. Clin. Immunol.* 20: 240–249.

113. Sant, M., C. Allemani, C. Tereanu, R. De Angelis, R. Capocaccia, O. Visser, R. Marcos-Gragera, M. Maynadi, A. Simonetti, J. M. Lutz, F. Berrino, M. Hackl, J. Holub, M. Maynadie, B. Holleczek, L. Tryggvadottir, H. Comber, F. Bell??, A. Giacomin, S. Ferretti, E. Crocetti, D. Serraino, M. Vercelli, M. Federico, M. Fusco, M. Michiara, R. Tumino, L. Mangone, F. Falcini, A. Iannelli, M. Budroni, R. Zanetti, S. Piffer, F. La Rosa, P. Zambon, S. Sowe, K. England, F. Langmark, J. Rachtan, R. Mezyk, M. Zwierko, M. Ondrusova, M. Primicakelj, S. Khan, G. Jundt, M. Usel, S. M. Ess, A. Bordoni, R. Otter, J. W. Coebergh, S. Siesling, D. Greenberg, N. Easey, M. Roche, G. Lawrence, A. Gavin, D. H. Brewster, and J. Steward. 2010. Incidence of hematologic malignancies in Europe by morphologic subtype: Results of the HAEMACARE project. *Blood* 116: 3724–3734.

114. Zenz, T., H. Döhner, and S. Stilgenbauer. 2007. Genetics and risk-stratified approach to therapy in chronic lymphocytic leukemia. *Best Pract. Res. Clin. Haematol.* 20: 439–453.

115. Redaelli, A., P. Corporation, B. L. Laskin, J. M. Stephens, M. F. Botteman, and P. C. L. European. 2004. The clinical and epidemiological burden of chronic lymphocytic leukaemia. *Eur. J. Cancer* 13: 279–287.

116. Morabito, F., F. R. De, L. Laurenti, K. Zirlik, A. G. Recchia, M. Gentile, E. Morelli, E. Vigna, V. Gigliotti, R. Calemma, B. Amoroso, A. Neri, G. Cutrona, M. Ferrarini, S. Molica, P. G. Del, C. Tripodo, and A. Pinto. 2011. The cumulative amount of serum free light chain is a strong prognosticator in chronic lymphocytic leukemia. *Blood* 118: 6353–6361.

117. Eichhorst, B., M. Dreyling, T. Robak, E. Montserrat, and M. Hallek. 2011. Chronic lymphocytic leukemia: ESMO clinical practice guidelines for diagnosis, treatment and follow-up. *Ann. Oncol.* 22: 50–54.

118. Hallek, M., B. D. Cheson, D. Catovsky, F. Caligaris-cappio, G. Dighiero, and H. Do. 2008. Guidelines for the diagnosis and treatment of chronic lymphocytic leukemia : a report from the International Workshop on Chronic Lymphocytic Leukemia updating the National Cancer Institute – Working Group 1996 guidelines. *Blood* 111: 5446–5456.

119. Lin, T. S., F. T. Awan, and J. C. Byrd. 2008. Chapter 83: Chronic lymphocytic leukemia. *Hematol. Basic Princ. Pract.* 333: 1327–1347.

120. Cheson, B., J. Bennett, M. Grever, N. Kay, M. Keating, S. O'Brien, and K. Rai. 1996. National Cancer Institute-sponsored Working Group guidelines for chronic lymphocytic leukemia: revised guidelines for diagnosis and treatment. *Blood* 87: 4990--7.

121. Kilo, M. N., and D. M. Dorfman. 1996. The utility of flow cytometric immunophenotypic analysis in the distinction of small lymphocytic lymphoma/chronic lymphocytic leukemia from mantle cell lymphoma. *Am. J. Clin. Pathol.* 105: 451–457.

122. Zhang, S., and T. J. Kipps. 2014. The pathogenesis of chronic lymphocytic leukemia. *Annu. Rev. Pathol.* 9: 103–18.

123. Ferrarini, M., and N. Chiorazzi. 2004. Recent advances in the molecular biology and

immunobiology of chronic lymphocytic leukemia. *Semin. Hematol.* 41: 207–223.

124. Inamdar, K. V., and C. E. Bueso-Ramos. 2007. Pathology of chronic lymphocytic leukemia: an update. *Ann. Diagn. Pathol.* 11: 363–389.

125. Mauro, B. F. R., R. Foa, D. Giannarelli, I. Cordone, S. Crescenzi, and E. Pescarmona. 2015. Clinical Characteristics and Outcome of Young Chronic Lymphocytic Leukemia Patients: A Single Institution Study of 204 Cases. *Blood* 94(2): 448–454.

126. Bhatwadekar, S., and T. J. Kipps. 2000. Chronic lymphocytic leukemia. *Curr. Opin. Hematol.* 7: 223–234.

127. Cuneo, A., G. M. Rigolin, R. Bigoni, C. De Angeli, A. Veronese, F. Cavazzini, A. Bardi, M. G. Roberti, E. Tammiso, P. Agostini, M. Ciccone, M. Della Porta, A. Tieghi, L. Cavazzini, M. Negrini, and G. Castoldi. 2004. Chronic lymphocytic leukemia with 6q- shows distinct hematological features and intermediate prognosis. *Leukemia* 18: 476–83.

128. Sanchez, M. L., J. Almeida, D. Gonzalez, M. Gonzalez, M. A. Garcia-Marcos, A. Balanzategui, M. C. Lopez-Berges, J. Nomdedeu, T. Vallespi, M. Barbon, A. Martin, P. De la Fuente, G. Martin-Nuñez, J. Fernandez-Calvo, J. M. Hernandez, J. F. San Miguel, and A. Orfao. 2003. Incidence and clinicobiologic characteristics of leukemic B-cell chronic lymphoproliferative disorders with more than one B-cell clone. *Blood* 102: 2994–3002.

129. Diehl, L. F., and L. H. Ketchum. 1998. Autoimmune disease and chronic lymphocytic leukemia: autoimmune hemolytic anemia, pure red cell aplasia, and autoimmune thrombocytopenia. *Semin. Oncol.* 25: 80–97.

130. Dearden, C. 2008. Disease-specific complications of chronic lymphocytic leukemia. *Hematology* 2008: 450–456.

131. Bonvalet, D., C. Foldes, and J. Civatte. 1984. Cutaneous manifestations in chronic lymphocytic leukemia. *J. Dermatol. Surg. Oncol.* 10: 278–82.

132. Cerroni, L., P. Zenahlik, G. Höfler, S. Kaddu, J. Smolle, and H. Kerl. 1996. Specific Cutaneous Infiltrates of B-cell Chronic Lymphocytic Leukemia. *Am. J. Surg. Pathol.* 20: 1000–1010.

133. Dameshek, W., and R. S. Schwartz. 1959. Leukemia and auto-immunization- some possible relationships. *Blood* 14: 1151–1158.

134. Pisciotta, A. V, and J. S. Hirschboeck. 1957. Therapeutic considerations in chronic lymphocytic leukemia; special reference to the natural course of the disease. *AMA Arch Intern Med* 99: 334–335.

135. Tsiodras, S., G. Samonis, M. J. Keating, and D. P. Kontoyiannis. 2000. Infection and immunity in chronic lymphocytic leukemia. *Mayo Clin. Proc.* 75: 1039–1054.

136. Rai, K. R. 1975. Clinical staging of chronic lymphocytic leukemia. *Blood* 46: 219–235.

137. Binet, J. L., M. Leporrier, G. Dighiero, D. Charron, G. Vaugier, H. M. Beral, J. C. Natali, M. Raphael, B. Nizet, and J. Y. Follezou. 1977. A clinical staging system for chronic lymphocytic leukemia. Prognostic significance. *Cancer* 40: 855–864.

138. Cai, J., C. Humphries, A. Richardson, and P. W. Tucker. 1992. Extensive and selective mutation of a rearranged VH5 gene in human B cell chronic lymphocytic leukemia. *J. Exp. Med.* 176: 1073–81.

139. Hamblin, T. J., Z. Davis, A. Gardiner, D. G. Oscier, and F. K. Stevenson. 1999. Unmutated Ig V(H) genes are associated with a more aggressive form of chronic lymphocytic leukemia. *Blood* 94: 1848–1854.

140. Martin, T., S. F. Duffy, D. a Carson, and T. J. Kipps. 1992. Evidence for somatic selection of natural autoantibodies. *J. Exp. Med.* 175: 983–991.

141. Hervé, M., K. Xu, Y. S. Ng, H. Wardemann, E. Albesiano, B. T. Messmer, N. Chiorazzi, and E. Meffre. 2005. Unmutated and mutated chronic lymphocytic leukemias derive from self-reactive B cell precursors despite expressing different antibody reactivity. *J. Clin. Invest.* 115: 1636–1643.

142. Caligaris-Cappio, F., and P. Ghia. 2008. Novel insights in chronic lymphocytic leukemia: Are we getting closer to understanding the pathogenesis of the disease? *J. Clin. Oncol.* 26: 4497–4503.

143. Mouquet, H., and M. C. Nussenzweig. 2012. Polyreactive antibodies in adaptive immune responses to viruses. *Cell. Mol. Life Sci.* 69: 1435–1445.

144. Sutton, L. A., E. Kostareli, A. Hadzidimitriou, N. Darzentas, A. Tsaftaris, A. Anagnostopoulos, R. Rosenquist, and K. Stamatopoulos. 2009. Extensive intraclonal diversification in a subgroup of chronic lymphocytic leukemia patients with stereotyped IGHV4-34 receptors: Implications for ongoing interactions with antigen. *Blood* 114: 4460–4468.

145. Kern, W., U. Bacher, S. Schnittger, F. Dicker, T. Alpermann, T. Haferlach, and C. Haferlach. 2014. Flow cytometric identification of 76 patients with biclonal disease among 5523 patients with chronic lymphocytic leukaemia (B-CLL) and its genetic characterization. *Br. J. Haematol.* 164: 565–569.

146. Stevenson, F. K., and F. Caligaris-Cappio. 2004. Chronic lymphocytic leukemia: Revelations from the B-cell receptor. *Blood* 103: 4389–4395.

147. Kipps, T. J., E. Tomhave, L. F. Pratt, S. Duffy, P. P. Chen, and D. a Carson. 1989. Developmentally restricted immunoglobulin heavy chain variable region gene expressed at high frequency in chronic lymphocytic leukemia. In *Proceedings of the National Academy of Sciences of the United States of America* vol. 86. 5913–5917.

148. Schroeder, H. W., and G. Dighiero. 1994. The pathogenesis of chronic lymphocytic leukemia: Analysis of the antibody repertoire. *Immunol. Today* 15: 288–294.

149. Fais, F., F. Ghiotto, S. Hashimoto, B. Sellars, A. Valetto, S. L. Allen, P. Schulman, V. P. Vinciguerra, K. Rai, L. Z. Rassenti, T. J. Kipps, G. Dighiero, H. W. Schroeder, M. Ferrarini, and N. Chiorazzi. 1998. Chronic lymphocytic leukemia B cells express restricted sets of mutated and unmutated antigen receptors. *J. Clin. Invest.* 102: 1515–1525.

150. Chiorazzi, N., and M. Ferrarini. 2003. B CELL CHRONIC LYMPHOCYTIC LEUKEMIA : Lessons Learned from Studies of the B Cell Antigen Receptor Biases in V Gene Use. *Annu. Rev. Immunol.* 21: 841–94.

151. Ghiotto, F., F. Fais, and A. Valetto. 2004. Remarkably similar antigen receptors among a subset of patients with chronic lymphocytic leukemia. *J. Clin. Invest.* 113: 1008–1016.

152. Tobin, G., O. Soderberg, U. Thunberg, and R. Rosenquist. 2004. V(H)3-21 gene usage in chronic lymphocytic leukemia--characterization of a new subgroup with distinct molecular features and poor survival. *Leuk Lymphoma* 45: 221–228.

153. Tobin, G., U. Thunberg, K. Karlsson, F. Murray, A. Laurell, K. Willander, G. Enblad, M. Merup, J. Vilpo, G. Juliusson, C. Sundström, O. Söderberg, G. Roos, and R. Rosenquist. 2004. Subsets with restricted immunoglobulin gene rearrangement features indicate a role for antigen selection in the development of chronic lymphocytic leukemia. *Blood* 104: 2879–2885.

154. Marchalonis, J. J., M. K. Adelman, S. F. Schluter, and P. A. Ramsland. 2006. The antibody repertoire in evolution: Chance, selection, and continuity. *Dev. Comp. Immunol.* 30: 223–247.

155. Litman, G. W., J. P. Cannon, and L. J. Dishaw. 2005. Reconstructing immune phylogeny: new perspectives. *Nat. Rev. Immunol.* 5: 866–879.

156. Weinstein, J. A. 2010. High-Throughput Sequencing of the Zebrafish Antibody Repertoire Joshua A. Weinstein,. *Zebrafish* 807: 807–810.

157. Jiang, N., J. a Weinstein, L. Penland, R. a White, D. S. Fisher, and S. R. Quake. 2011. Determinism and stochasticity during maturation of the zebrafish antibody repertoire. *Proc. Natl. Acad. Sci. U. S. A.* 108: 5348–53.

158. Ippolito, G. C., K. H. Hoi, S. T. Reddy, S. M. Carroll, X. Ge, T. Rogosch, M. Zemlin, L. D. Shultz, A. D. Ellington, C. L. VanDenBerg, and G. Georgiou. 2012. Antibody repertoires in humanized NOD-scid-IL2Rnull mice and human B cells reveals human-like diversification and tolerance checkpoints in the mouse. *PLoS One* 7.

159. Lonberg, N. 2008. Human monoclonal antibodies from transgenic mice. *Handb. Exp. Pharmacol.* 181: 69–97.

160. Brüggemann, M., M. J. Osborn, B. Ma, J. Hayre, S. Avis, B. Lundstrom, and R. Buelow. 2015.

Human Antibody Production in Transgenic Animals. *Arch. Immunol. Ther. Exp. (Warsz).* 63: 101–108.

161. Green, L. L. 2014. Transgenic mouse strains as platforms for the successful discovery and development of human therapeutic monoclonal antibodies. *Curr. Drug Discov. Technol.* 11: 74–84.

162. Pruzina, S., G. T. Williams, G. Kaneva, S. L. Davies, A. Martín-López, M. Brüggemann, S. M. Vieira, S. A. Jeffs, Q. J. Sattentau, and M. S. Neuberger. 2011. Human monoclonal antibodies to HIV-1 gp140 from mice bearing YAC-based human immunoglobulin transloci. *Protein Eng. Des. Sel.* 24: 791–799.

163. Ma, B., M. J. Osborn, S. Avis, L. H. Ouisse, S. Ménoret, I. Anegon, R. Buelow, and M. Brüggemann. 2013. Human antibody expression in transgenic rats: Comparison of chimeric IgH loci with human VH, D and JH but bearing different rat C-gene regions. *J. Immunol. Methods* 400–401: 78–86.

164. Osborn, M. J., B. Ma, S. Avis, J. Dilley, X. Yang, K. Lindquist, A.-L. Iscache, L.-H. Ouisse, I. Anegon, M. S. Neuberger, M. Brüggemann, M. J. Osborn, B. Ma, S. Avis, A. Binnie, J. Dilley, X. Yang, K. Lindquist, S. Ménoret, A.-L. Iscache, L.-H. Ouisse, A. Rajpal, I. Anegon, M. S. Neuberger, R. Buelow, and M. Brüggemann. 2013. High-affinity IgG antibodies develop naturally in Ig-knockout rats carrying germline human IgH/Igκ/Igλ loci bearing the rat CH region. *J. Immunol.* 190: 1481–90.

165. Geurts, A. M., G. J. Cost, Y. Freyvert, B. Zeitler, C. Jeffrey, V. M. Choi, S. S. Jenkins, A. Wood, X. Cui, X. Meng, A. Vincent, S. Lam, M. Michalkiewicz, R. Schilling, S. Kalloway, H. Weiler, S. Ménoret, I. Anegon, G. D. Davis, L. Zhang, E. J. Rebar, P. D. Gregory, F. D. Urnov, H. J. Jacob, and R. Buelow. 2010. Knockout Rats Produced Using Designed Zinc Finger Nucleases. *Science (80).* 325: 2009–2011.

166. Ménoret, S., A.-L. Iscache, L. Tesson, S. Rémy, C. Usal, M. J. Osborn, G. J. Cost, M. Brüggemann, R. Buelow, and I. Anegon. 2010. Characterization of immunoglobulin heavy chain knockout rats. *Eur. J. Immunol.* 40: 2932–41.

167. Schmitt, M. W., S. R. Kennedy, J. J. Salk, E. J. Fox, J. B. Hiatt, and L. A. Loeb. 2012. Detection of ultra-rare mutations by next-generation sequencing. *Proc. Natl. Acad. Sci.* 109: 14508–14513.

168. Jiang, N., J. He, J. a Weinstein, L. Penland, S. Sasaki, X.-S. He, C. L. Dekker, N.-Y. Zheng, M. Huang, M. Sullivan, P. C. Wilson, H. B. Greenberg, M. M. Davis, D. S. Fisher, and S. R. Quake. 2013. Lineage structure of the human antibody repertoire in response to influenza vaccination. *Sci. Transl. Med.* 5: 171ra19.

169. Henry Dunand, C. J., and P. C. Wilson. 2015. Restricted, canonical, stereotyped and convergent immunoglobulin responses. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* 370: 20140238-.

170. Galson, J. D., E. A. Clutterbuck, J. Trück, M. N. Ramasamy, M. Münz, A. Fowler, V. Cerundolo, A. J. Pollard, G. Lunter, and D. F. Kelly. 2015. BCR repertoire sequencing: different patterns of B-cell activation after two Meningococcal vaccines. *Immunol. Cell Biol.* 93: 885–95.

171. Jackson, K. J. L., Y. Liu, K. M. Roskin, J. Glanville, R. A. Hoh, K. Seo, E. L. Marshall, T. C. Gurley, M. A. Moody, B. F. Haynes, E. B. Walter, H. X. Liao, R. A. Albrecht, A. García-Sastre, J. Chaparro-Riggers, A. Rajpal, J. Pons, B. B. Simen, B. Hanczaruk, C. L. Dekker, J. Laserson, D. Koller, M. M. Davis, A. Z. Fire, and S. D. Boyd. 2014. Human responses to influenza vaccination show seroconversion signatures and convergent antibody rearrangements. *Cell Host Microbe* 16: 105–114.

172. Trepel, F. 1974. Number and distribution of lymphocytes in man. A critical analysis. *Klin. Wochenschr.* 52: 511–515.

173. Greiff, V., U. Menzel, E. Miho, C. Weber, R. Riedel, S. Cook, A. Valai, T. Lopes, A. Radbruch, T. H. Winkler, and S. T. Reddy. 2017. Systems Analysis Reveals High Genetic and Antigen-Driven Predetermination of Antibody Repertoires throughout B Cell Development. *Cell Rep.* 19: 1467–1478.

174. McMillan, R., R. L. Longmire, R. Yelenosky, J. E. Lang, V. Heath, and C. G. Craddock. 1972. Immunoglobulin synthesis by human lymphoid tissues: normal bone marrow as a major site of IgG production. *J. Immunol.* 109: 1386–94.

175. Slifka, M. K., R. Antia, J. K. Whitmire, and R. Ahmed. 1998. Humoral immunity due to long-lived plasma cells. *Immunity* 8: 363–372.

176. Muellenbeck, M. F., B. Ueberheide, B. Amulic, A. Epp, D. Fenyo, C. E. Busse, M. Esen, M. Theisen, B. Mordmüller, and H. Wardemann. 2013. Atypical and classical memory B cells produce Plasmodium

falciparum neutralizing antibodies. *J. Exp. Med.* 210: 389–99.

177. Dubois, A. R. S. X., J. P. Buerckert, R. Sinner, W. J. Faison, A. M. Molitor, and C. P. Muller. 2016. High-resolution analysis of the B cell repertoire before and after polyethylene glycol fusion reveals preferential fusion of rare antigen-specific B cells. *Hum. Antibodies* 24: 1–15.

178. de Kruif, J., A. Kramer, T. Visser, C. Clements, R. Nijhuis, F. Cox, V. van der Zande, R. Smit, D. Pinto, M. Throsby, and T. Logtenberg. 2009. Human Immunoglobulin Repertoires against Tetanus Toxoid Contain a Large and Diverse Fraction of High-Affinity Promiscuous VH Genes. *J. Mol. Biol.* 387: 548–558.

179. Meijer, P. J., P. S. Andersen, M. Haahr Hansen, L. Steinaa, A. Jensen, J. Lantto, M. B. Oleksiewicz, K. Tengbjerg, T. R. Poulsen, V. W. Coljee, S. Bregenholt, J. S. Haurum, and L. S. Nielsen. 2006. Isolation of Human Antibody Repertoires with Preservation of the Natural Heavy and Light Chain Pairing. *J. Mol. Biol.* 358: 764–772.

180. Poulsen, T. R., P.-J. Meijer, A. Jensen, L. S. Nielsen, and P. S. Andersen. 2007. Kinetic, Affinity, and Diversity Limits of Human Polyclonal Antibody Responses against Tetanus Toxoid. *J. Immunol.* 179: 3841–3850.

181. Grova, N., E. J. F. Prodhomme, M. T. Schellenberger, S. Farinelle, and C. P. Muller. 2009. Modulation of carcinogen bioavailability by immunisation with benzo[a]pyrene-conjugate vaccines. *Vaccine* 27: 4142–4151.

182. Edition, E. 2011. *Guide*, 8th ed. National Academies Press (US), Washington (DC).

183. Carroll, M. W., and B. Moss. 1997. Host range and cytopathogenicity of the highly attenuated MVA strain of vaccinia virus: propagation and generation of recombinant viruses in a nonhuman mammalian cell line. *Virology* 238: 198–211.

184. Drexler, I., K. Heller, B. Wahren, V. Erfle, and G. Sutter. 1998. Highly attenuated modified vaccinia virus Ankara replicates in baby hamster kidney cells, a potential host for virus propagation, but not in various human transformed and primary cells. *J. Gen. Virol.* 79: 347–352.

185. Staib, C., and G. Sutter. 2003. Live Viral Vectors: Vaccinia Virus. *Vaccine Protoc.* 87: 51–68.

186. Lefranc, M.-P. 2004. IMGT, The International ImMunoGeneTics Information System, http://imgt.cines.fr. *Methods Mol. Biol.* 248: 27–49.

187. Alamyar, E., P. Duroux, M. P. Lefranc, and V. Giudicelli. 2012. IMGT® tools for the nucleotide analysis of immunoglobulin (IG) and t cell receptor (TR) V-(D)-J repertoires, polymorphisms, and IG mutations: IMGT/V-QUEST and IMGT/HighV-QUEST for NGS. In *Methods in Molecular Biology* vol. 882. 569–604.

188. Ye, J., N. Ma, T. L. Madden, and J. M. Ostell. 2013. IgBLAST: an immunoglobulin variable domain sequence analysis tool. *Nucleic Acids Res.* 41: W34--40.

189. Galson, J. D., J. Trück, A. Fowler, E. A. Clutterbuck, M. Münz, V. Cerundolo, C. Reinhard, R. van der Most, A. J. Pollard, G. Lunter, and D. F. Kelly. 2015. Analysis of B Cell Repertoire Dynamics Following Hepatitis B Vaccination in Humans, and Enrichment of Vaccine-specific Antibody Sequences. *EBioMedicine* 2: 2070–2079.

190. Gupta, N. T., J. A. Vander Heiden, M. Uduman, D. Gadala-Maria, G. Yaari, and S. H. Kleinstein. 2015. Change-O: A toolkit for analyzing large-scale B cell immunoglobulin repertoire sequencing data. *Bioinformatics* 31: 3356–3358.

191. Love, M. I., W. Huber, and S. Anders. 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* 15: 550.

192. Love, M. I., S. Anders, and W. Huber. 2014. *Differential analysis of count data - the DESeq2 package*.

193. Benjamini, Y., and Y. Hochberg. 1995. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *R. Stat. Soc.* 57: 289–300.

194. Team, R. D. C. 2004. R: A language and environment for statistical computing. *Vienna, Austria R*

Found. Stat. Comput. .

195. Sivasubramanian, A., A. Sircar, S. Chaudhury, and J. J. Gray. 2009. Toward high-resolution homology modeling of antibody F v regions and application to antibody-antigen docking. *Proteins Struct. Funct. Bioinforma.* 74: 497–514.

196. Lyskov, S., F. C. Chou, S. Ó. Conchúir, B. S. Der, K. Drew, D. Kuroda, J. Xu, B. D. Weitzner, P. D. Renfrew, P. Sripakdeevong, B. Borgo, J. J. Havranek, B. Kuhlman, T. Kortemme, R. Bonneau, J. J. Gray, and R. Das. 2013. Serverification of Molecular Modeling Applications: The Rosetta Online Server That Includes Everyone (ROSIE). *PLoS One* 8: 5–7.

197. Sircar, A., E. T. Kim, and J. J. Gray. 2009. RosettaAntibody: Antibody variable region homology modeling server. *Nucleic Acids Res.* 37: W474–W479.

198. Schrödinger, L. L. C. 2015. The PyMOL molecular graphics system, version 1.8.

199. Galson, J. D., J. Trück, A. Fowler, M. Münz, V. Cerundolo, A. J. Pollard, G. Lunter, and D. F. Kelly. 2015. In-depth assessment of within-individual and inter-individual variation in the B cell receptor repertoire. *Front. Immunol.* 6: 531.

200. Apostoaei, I. A., and J. R. Trabalka. 2010. Review, Synthesis, and Application of Information on the Human Lymphatic System to Radiation Dosimetry for Chronic Lymphocytic Leukemia. 1–51.

201. Westermann, J., and R. Pabst. 1992. Distribution of lymphocyte subsets and natural killer cells in the human body. *Clin. Investig.* 70: 539–544.

202. Lee, J., D. R. Boutz, V. Chromikova, M. G. Joyce, C. Vollmers, K. Leung, A. P. Horton, B. J. DeKosky, C.-H. Lee, J. J. Lavinder, E. M. Murrin, C. Chrysostomou, K. H. Hoi, Y. Tsybovsky, P. V Thomas, A. Druz, B. Zhang, Y. Zhang, L. Wang, W.-P. Kong, D. Park, L. I. Popova, C. L. Dekker, M. M. Davis, C. E. Carter, T. M. Ross, A. D. Ellington, P. C. Wilson, E. M. Marcotte, J. R. Mascola, G. C. Ippolito, F. Krammer, S. R. Quake, P. D. Kwong, and G. Georgiou. 2016. Molecular-level analysis of the serum antibody repertoire in young adults before and after seasonal influenza vaccination. *Nat Med* 22: 1456–1464.

203. Hoogenboom, H. R., and G. Winter. 1992. By-passing immunisation. Human antibodies from synthetic repertoires of germline VH gene segments rearranged in vitro. *J. Mol. Biol.* 227: 381–388.

204. Bürckert, J.-P., W. J. Faison, A. R. S. X. Dubois, R. Sinner, O. Hunewald, A. Wienecke-Baldacchino, A. Brieger, and C. P. Muller. 2017. High-throughput sequencing of murine immunoglobulin heavy chain transcripts using single side unique molecular identifiers on an Ion Torrent PGM. *bioRxiv* .

205. Strauli, N. B., and R. D. Hernandez. 2016. Statistical inference of a convergent antibody repertoire response to influenza vaccine. *Genome Med.* 8: 60.

206. Kipps, T. T. J. 2003. Immunobiology of chronic lymphocytic leukemia. *Curr. Opin. Hematol.* 10: 312–318.

207. Gentile, M., F. R. Mauro, A. Guarini, and R. Foà. 2005. New developments in the diagnosis, prognosis and treatment of chronic lymphocytic leukemia. *Curr. Opin. Oncol.* 17: 597–604.

208. Damle, B. R. N., T. Wasil, F. Fais, F. Ghiotto, A. Valetto, S. L. Allen, A. Buchbinder, D. Budman, K. Dittmar, J. Kolitz, S. M. Lichtman, P. Schulman, V. P. Vinciguerra, K. R. Rai, M. Ferrarini, and N. Chiorazzi. 1999. Ig V Gene Mutation Status and CD38 Expression As Novel Prognostic Indicators in Chronic Lymphocytic Leukemia. *Blood* 94: 1840–1847.

209. Rossi, D., V. Spina, R. Bomben, S. Rasi, M. Dal-Bo, A. Bruscaggin, F. M. Rossi, S. Monti, M. Degan, C. Ciardullo, R. Serra, A. Zucchetto, J. Nomdedeu, P. Bulian, A. Grossi, F. Zaja, G. Pozzato, L. Laurenti, D. G. Efremov, F. Di-Raimondo, R. Marasca, F. Forconi, G. Del Poeta, G. Gaidano, and V. Gattei. 2013. Association between molecular lesions and specific B-cell receptor subsets in chronic lymphocytic leukemia. *Blood* 121: 4902–4905.

210. Fabbri, G., and R. Dalla-Favera. 2016. The molecular pathogenesis of chronic lymphocytic leukaemia. *Nat. Rev. Cancer* 16: 145–162.

211. Rosati, E., R. Sabatini, G. Rampino, A. Tabilio, M. Di Ianni, K. Fettucciari, A. Bartoli, S. Coaccioli, I. Screpanti, P. Marconi, W. Dc, E. Rosati, R. Sabatini, G. Rampino, A. Tabilio, M. Di Ianni, K. Fettucciari,

A. Bartoli, S. Coaccioli, I. Screpanti, and P. Marconi. 2009. Constitutively activated Notch signaling is involved in survival and apoptosis resistance of B-CLL cells Constitutively activated Notch signaling is involved in survival and apoptosis resistance of B-CLL cells. *Blood* 113: 856–865.

212. Cuni, S., P. Perez-Aciego, G. Perez-Chacon, J. A. Vargas, A. Sanchez, F. M. Martin-Saavedra, S. Ballester, J. Garcia-Marco, J. Jorda, and A. Durantez. 2004. A sustained activation of PI3K/NF-kappaB pathway is critical for the survival of chronic lymphocytic leukemia B cells. *Leukemia* 18: 1391–1400.

213. Schuh, K., a Avots, H. P. Tony, E. Serfling, and C. Kneitz. 1996. Nuclear NF-ATp is a hallmark of unstimulated B cells from B-CLL patients. *Leuk. Lymphoma* 23: 583–92.

214. Furman, R. R., Z. Asgary, J. O. Mascarenhas, H. C. Liou, and E. J. Schattner. 2000. Modulation of NF-kappa B activity and apoptosis in chronic lymphocytic leukemia B cells. *J. Immunol.* 164: 2200–2206.

215. Liu, P., B. Xu, W. Shen, H. Zhu, W. Wu, Y. Fu, H. Chen, H. Dong, Y. Zhu, K. Miao, W. Xu, and J. Li. 2012. Dysregulation of TNFα-induced necroptotic signaling in chronic lymphocytic leukemia: suppression of CYLD gene by LEF1. *Leukemia* 26: 1293–1300.

216. Mansouri, L., N. Papakonstantinou, S. Ntoufa, K. Stamatopoulos, and R. Rosenquist. 2016. NF-kB activation in chronic lymphocytic leukemia: A point of convergence of external triggers and intrinsic lesions. *Semin. Cancer Biol.* 39: 40–48.

217. Malynn, B. A., and A. Ma. 2010. Ubiquitin makes its mark on immune regulation. *Immunity* 33: 843–52.

218. Massoumi, R. 2010. Ubiquitin chain cleavage: CYLD at work. *Trends Biochem. Sci.* 35: 392–399.

219. Hymowitz, S. G., and I. E. Wertz. 2010. A20: from ubiquitin editing to tumour suppression. *Nat. Rev. Cancer* 10: 332–341.

220. Bignell, G. R., W. Warren, S. Seal, M. Takahashi, E. Rapley, R. Barfoot, H. Green, C. Brown, P. J. Biggs, S. R. Lakhani, C. Jones, J. Hansen, E. Blair, B. Hofmann, R. Siebert, G. Turner, D. G. Evans, C. Schrander-Stumpel, F. a Beemer, a van Den Ouweland, D. Halley, B. Delpech, M. G. Cleveland, I. Leigh, J. Leisti, and S. Rasmussen. 2000. Identification of the familial cylindromatosis tumour-suppressor gene. *Nat. Genet.* 25: 160–165.

221. Mathis, B. J., Y. Lai, C. Qu, J. S. Janicki, and T. Cui. 2015. CYLD-mediated signaling and diseases. *Curr. Drug Targets* 16: 284–94.

222. Hövelmeyer, N., F. T. Wunderlich, R. Massoumi, C. G. Jakobsen, J. Song, M. a Wörns, C. Merkwirth, A. Kovalenko, M. Aumailley, D. Strand, J. C. Brüning, P. R. Galle, D. Wallach, R. Fässler, and A. Waisman. 2007. Regulation of B cell homeostasis and activation by the tumor suppressor gene CYLD. *J. Exp. Med.* 204: 2615–2627.

223. Wu, W., H. Zhu, Y. Fu, W. Shen, J. Xu, K. Miao, M. Hong, W. Xu, P. Liu, and J. Li. 2014. Clinical significance of down-regulated cylindromatosis gene in chronic lymphocytic leukemia. *Leuk. Lymphoma* 55: 588–94.

224. Chu, Y., J. C. Vahl, D. Kumar, K. Heger, A. Bertossi, E. Wójtowicz, V. Soberon, D. Schenten, B. Mack, M. Reutelshöfer, R. Beyaert, G. Van Loo, M. Schmidt-supprian, W. Dc, E. Wo, M. Reutelsho, and K. Amann. 2011. differentiation and hyperactivation and cause inflammation and autoimmunity in aged mice B cells lacking the tumor suppressor TNFAIP3 / A20 display impaired differentiation and hyperactivation and cause inflammation and autoimmunity in aged mice. *Blood* 117: 2227–2236.

225. Hövelmeyer, N., S. Reissig, N. Thi Xuan, P. Adams-Quack, D. Lukas, A. Nikolaev, D. Schlüter, and A. Waisman. 2011. A20 deficiency in B cells enhances B-cell proliferation and results in the development of autoantibodies. *Eur. J. Immunol.* 41: 595–601.

226. Tavares, R. M., E. E. Turer, C. L. Liu, R. Advincula, P. Scapini, L. Rhee, J. Barrera, C. A. Lowell, P. J. Utz, B. A. Malynn, and A. Ma. 2010. The ubiquitin modifying enzyme A20 restricts B cell survival and prevents autoimmunity. *Immunity* 33: 181–91.

227. Philipp, C., J. Edelmann, A. Bühler, D. Winkler, S. Stilgenbauer, and R. Küppers. 2011. Mutation analysis of the TNFAIP3 (A20) tumor suppressor gene in CLL. *Int. J. Cancer* 128: 1747–1750.

228. Frenzel, L. P., R. Claus, N. Plume, J. Schwamb, C. Konermann, C. P. Pallasch, J. Claasen, R. Brinker, B. Wollnik, C. Plass, and C. M. Wendtner. 2011. Sustained NF-kappaB activity in chronic lymphocytic leukemia is independent of genetic and epigenetic alterations in the TNFAIP3 (A20) locus. *Int. J. Cancer* 128: 2495–2500.

229. Massoumi, R., K. Chmielarska, K. Hennecke, A. Pfeifer, and R. Fässler. 2006. Cyld inhibits tumor cell proliferation by blocking Bcl-3-dependent NF-kappaB signaling. *Cell* 125: 665–77.

230. Löytynoja, A., A. J. Vilella, and N. Goldman. 2012. Accurate extension of multiple sequence alignments using a phylogeny-aware graph algorithm. *Bioinformatics* 28: 1684–1691.

231. Morrow, J. S. 1977. Toward a more normative assessment of maldistribution: the Gini index. *Inquiry* 14: 278–292.

232. Sasaki, Y., E. Derudder, E. Hobeika, R. Pelanda, M. Reth, K. Rajewsky, M. Schmidt-Supprian, M. Matsumoto, F. Beermann, J. Tschopp, and P. Schneider. 2006. Canonical NF-κB Activity, Dispensable for B Cell Development, Replaces BAFF-Receptor Signals and Promotes B Cell Proliferation upon Activation. *Immunity* 24: 729–739.

233. Kantor, A. B., and L. A. Herzenberg. 1993. Origin of murine B cell lineages. *Annu. Rev. Immunol.* 11: 501–38.

234. Pritsch, O., C. Magnac, G. Dumas, C. Egile, and G. Dighiero. 1993. V gene usage by seven hybrids derived from CD5+ B-cell chronic lymphocytic leukemia and displaying autoantibody activity. *Blood* 82: 3103–12.

235. Herishanu, Y., P. Pérez-Galán, D. Liu, A. Biancotto, S. Pittaluga, B. Vire, F. Gibellini, N. Njuguna, E. Lee, L. Stennett, N. Raghavachari, P. Liu, J. P. McCoy, M. Raffeld, M. Stetler-Stevenson, C. Yuan, R. Sherry, D. C. Arthur, I. Maric, T. White, G. E. Marti, P. Munson, W. H. Wilson, and A. Wiestner. 2011. The lymph node microenvironment promotes B-cell receptor signaling, NF-kappaB activation, and tumor proliferation in chronic lymphocytic leukemia. *Blood* 117: 563–74.

236. Matutes, E., and a Polliack. 2000. Morphological and immunophenotypic features of chronic lymphocytic leukemia. *Rev. Clin. Exp. Hematol.* 4: 22–47.

237. Drillenburg, P., and S. T. Pals. 2000. Cell adhesion receptors in lymphoma dissemination. *Blood* 95: 1900–1910.

238. Bichi, R., S. A. Shinton, E. S. Martin, A. Koval, G. A. Calin, R. Cesari, G. Russo, R. R. Hardy, and C. M. Croce. 2002. Human chronic lymphocytic leukemia modeled in mouse by targeted TCL1 expression. *Proc. Natl. Acad. Sci.* 99: 6955–6960.

239. Bashford-rogers, R. J. M., A. L. Palser, B. J. Huntly, R. Rance, G. S. Vassiliou, G. a Follows, and P. Kellam. 2013. Network properties derived from deep sequencing of human B-cell receptor repertoires delineate B-cell populations. 1874–1884.

240. Ghia, P., K. Stamatopoulos, C. Belessi, C. Moreno, S. Stilgenbauer, F. K. Stevenson, F. Davi, and R. Rosenquist. 2007. ERIC recommendations on IGHV gene mutational status analysis in chronic lymphocytic leukemia. *Leuk. Off. J. Leuk. Soc. Am. Leuk. Res. Fund, U.K* 21: 1–3.

241. Förster, I., and K. Rajewsky. 1987. Expansion and functional activity of Ly1+ B cells upon transfer of peritoneal cells into allotype congenic, newborn mice. *Eur. J. Immunol.* 17: 521–528.

242. Enzler, T., A. P. Kater, W. Zhang, G. F. W. Ii, H. Chuang, J. Lee, E. Avery, C. M. Croce, M. Karin, T. J. Kipps, and W. Dc. 2012. Chronic lymphocytic leukemia of E μ - TCL1 transgenic mice undergoes rapid cell turnover that can be offset by extrinsic CD257 to accelerate disease progression Chronic lymphocytic leukemia of E -TCL1 transgenic mice undergoes rapid cell turnover that can be offset by extrinsic CD257 to accelerate disease progression. *Blood* 114: 4469–4476.

243. Dejardin, E., N. M. Droin, M. Delhase, E. Haas, Y. Cao, C. Makris, Z.-W. Li, M. Karin, C. F. Ware, and D. R. Green. 2002. The lymphotoxin-beta receptor induces different patterns of gene expression via two NF-kappaB pathways. *Immunity* 17: 525–35.

244. Fusco, A. J., O. V. Savinova, R. Talwar, J. D. Kearns, A. Hoffmann, and G. Ghosh. 2008. Stabilization of RelB requires multidomain interactions with p100/p52. *J. Biol. Chem.* 283: 12324–12332.

245. Mineva, N. D., T. L. Rothstein, J. A. Meyers, A. Lerner, and G. E. Sonenshein. 2007. CD40 ligand-mediated activation of the de novo RelB NF-kappaB synthesis pathway in transformed B cells promotes rescue from apoptosis. *J. Biol. Chem.* 282: 17475–17485.

246. Lopez-Guerra, M., and D. Colomer. 2010. NF-kappaB as a therapeutic target in chronic lymphocytic leukemia. *Expert Opin Ther Targets* 14: 275–288.

247. Duhren-von Minden, M., R. Ubelhart, D. Schneider, T. Wossning, M. P. Bach, M. Buchner, D. Hofmann, E. Surova, M. Follo, F. Kohler, H. Wardemann, K. Zirlik, H. Veelken, and H. Jumaa. 2012. Chronic lymphocytic leukaemia is driven by antigen-independent cell-autonomous signalling. *Nature* 489: 309–312.

248. Buggins, A. G. S., C. Pepper, P. E. M. Patten, S. Hewamana, S. Gohil, J. Moorhead, N. Folarin, D. Yallop, N. S. B. Thomas, G. J. Mufti, C. Fegan, and S. Devereux. 2010. Interaction with vascular endothelium enhances survival in primary chronic lymphocytic leukemia cells via NF-kappaB activation and de novo gene transcription. *Cancer Res.* 70: 7523–33.

249. Zapata, J. M., M. Krajewska, H. C. Morse, Y. Choi, and J. C. Reed. 2004. TNF receptor-associated factor (TRAF) domain and Bcl-2 cooperate to induce small B cell lymphoma/chronic lymphocytic leukemia in transgenic mice. *Pnas* 101: 16600–16605.

250. Klein, U., and R. Dalla-Favera. 2010. New insights into the pathogenesis of chronic lymphocytic leukemia. *Semin. Cancer Biol.* 20: 377–383.

251. Planelles, L., C. E. Carvalho-Pinto, G. Hardenberg, S. Smaniotto, W. Savino, R. Gómez-Caro, M. Alvarez-Mon, J. De Jong, E. Eldering, C. Martínez-A, J. P. Medema, and M. Hahne. 2004. APRIL promotes B-1 cell-associated neoplasm. *Cancer Cell* 6: 399–408.

252. Pekarsky, Y., A. Palamarchuk, V. Maximov, A. Efanov, N. Nazaryan, U. Santanam, L. Rassenti, T. Kipps, and C. M. Croce. 2008. Tcl1 functions as a transcriptional regulator and is directly involved in the pathogenesis of CLL. *Proc. Natl. Acad. Sci. U. S. A.* 105: 19643–8.

253. Puente, X. S., S. Beà, R. Valdés-Mas, N. Villamor, J. Gutiérrez-Abril, J. I. Martín-Subero, M. Munar, C. Rubio-Pérez, P. Jares, M. Aymerich, T. Baumann, R. Beekman, L. Belver, A. Carrio, G. Castellano, G. Clot, E. Colado, D. Colomer, D. Costa, J. Delgado, A. Enjuanes, X. Estivill, A. A. Ferrando, J. L. Gelpí, B. González, S. González, M. González, M. Gut, J. M. Hernández-Rivas, M. López-Guerra, D. Martín-García, A. Navarro, P. Nicolás, M. Orozco, Á. R. Payer, M. Pinyol, D. G. Pisano, D. A. Puente, A. C. Queirós, V. Quesada, C. M. Romeo-Casabona, C. Royo, R. Royo, M. Rozman, N. Russiñol, I. Salaverría, K. Stamatopoulos, H. G. Stunnenberg, D. Tamborero, M. J. Terol, A. Valencia, N. López-Bigas, D. Torrents, I. Gut, A. López-Guillermo, C. López-Otín, and E. Campo. 2015. Non-coding recurrent mutations in chronic lymphocytic leukaemia. *Nature* 526: 519–524.

254. Landau, D. A., S. L. Carter, P. Stojanov, A. McKenna, K. Stevenson, M. S. Lawrence, C. Sougnez, C. Stewart, A. Sivachenko, L. Wang, Y. Wan, W. Zhang, S. A. Shukla, A. Vartanov, S. M. Fernandes, G. Saksena, K. Cibulskis, B. Tesar, S. Gabriel, N. Hacohen, M. Meyerson, E. S. Lander, D. Neuberg, J. R. Brown, G. Getz, and C. J. Wu. 2013. Evolution and impact of subclonal mutations in chronic lymphocytic leukemia. *Cell* 152: 714–726.

255. Quesada, V., A. J. Ramsay, and C. Lopez-Otin. 2012. Chronic lymphocytic leukemia with SF3B1 mutation. *N. Engl. J. Med.* 366: 2530.

256. Rossi, D., V. Spina, F. Forconi, D. Capello, M. Fangazio, S. Rasi, M. Martini, V. Gattei, A. Ramponi, L. M. Larocca, F. Bertoni, and G. Gaidano. 2012. Molecular history of Richter syndrome: Origin from a cell already present at the time of chronic lymphocytic leukemia diagnosis. *Int. J. Cancer* 130: 3006–3010.

257. Mansouri, L., L.-A. Sutton, V. Ljungström, S. Bondza, L. Arngården, S. Bhoi, J. Larsson, D. Cortese, A. Kalushkova, K. Plevova, E. Young, R. Gunnarsson, E. Falk-Sörqvist, P. Lönn, A. F. Muggen, X.-J. Yan, B. Sander, G. Enblad, K. E. Smedby, G. Juliusson, C. Belessi, J. Rung, N. Chiorazzi, J. C. Strefford, A. W. Langerak, S. Pospisilova, F. Davi, M. Hellström, H. Jernberg-Wiklund, P. Ghia, O. Söderberg, K. Stamatopoulos, M. Nilsson, and R. Rosenquist. 2015. Functional loss of IκBε leads to NF-κB deregulation in aggressive chronic lymphocytic leukemia. *J. Exp. Med.* 212: 833–43.

258. Petlickovski, A., L. Laurenti, X. Li, S. Marietti, P. Chiusolo, S. Sica, G. Leone, and D. G. Efremov. 2005. Sustained signaling through the B-cell receptor induces Mcl-1 and promotes survival of chronic

lymphocytic leukemia B cells. *Blood* 105: 4820–4827.

259. Barragán, M., B. Bellosillo, C. Campàs, D. Colomer, G. Pons, and J. Gil. 2002. Involvement of protein kinase C and phosphatidylinositol 3-kinase pathways in the survival of B-cell chronic lymphocytic leukemia cells. *Blood* 99: 2969–2976.

260. Endo, T., M. Nishio, T. Enzler, H. B. Cottam, T. Fukuda, F. Danelle, M. Karin, T. J. Kipps, W. Dc, and D. F. James. 2012. BAFF and APRIL support chronic lymphocytic leukemia B-cell survival through activation of the canonical NF- κ B pathway BAFF and APRIL support chronic lymphocytic leukemia B-cell survival through activation of the canonical NF- B pathway. *Blood* 109: 703–710.

261. Viatour, P., M. Bentires-Alj, A. Chariot, V. Deregowski, L. de Leval, M.-P. Merville, and V. Bours. 2003. NF-κB2/p100 induces Bcl-2 expression. *Leukemia* 17: 1349–1356.

262. Xu, J., P. Zhou, W. Wang, A. Sun, and F. Guo. 2014. RelB, together with RelA, sustains cell survival and confers proteasome inhibitor sensitivity of chronic lymphocytic leukemia cells from bone marrow. *J. Mol. Med.* 92: 77–92.

263. Hewamana, S., T. T. Lin, C. Rowntree, K. Karunanithi, G. Pratt, R. Hills, C. Fegan, P. Brennan, and C. Pepper. 2009. Rel a is an independent biomarker of clinical outcome in chronic lymphocytic leukemia. *J. Clin. Oncol.* 27: 763–769.

264. Tracey, L., A. Pérez-Rosado, M. J. Artiga, F. I. Camacho, A. Rodríguez, N. Martínez, E. Ruiz-Ballesteros, M. Mollejo, B. Martinez, M. Cuadros, J. F. Garcia, M. Lawler, and M. Á. Piris. 2005. Expression of the NF-kappaB targets BCL2 and BIRC5/Survivin characterizes small B-cell and aggressive B-cell lymphomas, respectively. *J. Pathol.* 206: 123–134.

265. Declercq, W., T. Vanden Berghe, and P. Vandenabeele. 2009. RIP Kinases at the Crossroads of Cell Death and Survival. *Cell* 138: 229–232.

266. Reddy, S. T., and G. Georgiou. 2011. Systems analysis of adaptive immunity by utilization of high-throughput technologies. *Curr. Opin. Biotechnol.* 22: 584–589.

267. Fischer, N. 2011. Sequencing antibody repertoires: The next generation. *MAbs* 3: 17–20.

268. Weinstein, J. A., N. Jiang, R. A. White, D. S. Fisher, and S. R. Quake. 2009. High-Throughput Sequencing of the Zebrafish Antibody Repertoire. *Science (80).* 324: 807–810.

269. Metzker, M. L. 2010. Sequencing technologies - the next generation. *Nat. Rev. Genet.* 11: 31–46.

270. Huse, S. M., J. a Huber, H. G. Morrison, M. L. Sogin, and D. M. Welch. 2007. Accuracy and quality of massively-parallel DNA pyrosequencing. *Genome Biol.* 8: R143.

271. Nguyen, P., J. Ma, D. Pei, C. Obert, C. Cheng, and T. L. Geiger. 2011. Identification of errors introduced during high throughput sequencing of the T cell receptor repertoire. *BMC Genomics* 12: 106.

272. Fuellgrabe, M. W., D. Herrmann, H. Knecht, S. Kuenzel, M. Kneba, C. Pott, and M. Brüggemann. 2015. High-Throughput, Amplicon-Based Sequencing of the CREBBP Gene as a Tool to Develop a Universal Platform-Independent Assay. *PLoS One* 10: e0129195.

273. Zhu, J., S. O'Dell, G. Ofek, M. Pancera, X. Wu, B. Zhang, Z. Zhang, J. C. Mullikin, M. Simek, D. R. Burton, W. C. Koff, L. Shapiro, J. R. Mascola, and P. D. Kwong. 2012. Somatic populations of PGT135-137 HIV-1-neutralizing antibodies identified by 454 pyrosequencing and bioinformatics. *Front. Microbiol.* 3: 315.

274. Deng, W., B. S. Maust, D. H. Westfall, L. Chen, H. Zhao, B. B. Larsen, S. Iyer, Y. Liu, and J. I. Mullins. 2013. Indel and Carryforward Correction (ICC): A new analysis approach for processing 454 pyrosequencing data. *Bioinformatics* 29: 2402–2409.

275. Bolotin, D. A., I. Z. Mamedov, O. V. Britanova, I. V. Zvyagin, D. Shagin, S. V. Ustyugova, M. A. Turchaninova, S. Lukyanov, Y. B. Lebedev, and D. M. Chudakov. 2012. Next generation sequencing for TCR repertoire profiling: Platform-specific features and correction algorithms. *Eur. J. Immunol.* 42: 3073–3083.

276. Carlson, C. S., R. O. Emerson, A. M. Sherwood, C. Desmarais, M.-W. Chung, J. M. Parsons, M. S. Steen, M. a LaMadrid-Herrmannsfeldt, D. W. Williamson, R. J. Livingston, D. Wu, B. L. Wood, M. J.

Rieder, and H. Robins. 2013. Using synthetic templates to design an unbiased multiplex PCR assay. *Nat. Commun.* 4: 2680.

277. Best, K., T. Oakes, J. M. Heather, J. Shawe-Taylor, and B. Chain. 2015. Computational analysis of stochastic heterogeneity in PCR amplification efficiency revealed by single molecule barcoding. *Sci. Rep.* 5: 14629.

278. Cocquet, J., A. Chong, G. Zhang, and R. A. Veitia. 2006. Reverse transcriptase template switching and false alternative transcripts. *Genomics* 88: 127–131.

279. Fu, G. K., J. Wilhelmy, D. Stern, H. C. Fan, and S. P. A. Fodor. 2014. Digital encoding of cellular mRNAs enabling precise and absolute gene expression measurement by single-molecule counting. *Anal. Chem.* 86: 2867–2870.

280. Choi, N. M., S. Loguercio, J. Verma-Gaur, S. C. Degner, A. Torkamani, A. I. Su, E. M. Oltz, M. Artyomov, and A. J. Feeney. 2013. Deep sequencing of the murine IgH repertoire reveals complex regulation of nonrandom V gene rearrangement frequencies. *J. Immunol.* 191: 2393–402.

281. Yaari, G., and S. H. Kleinstein. 2015. Practical guidelines for B-cell receptor repertoire sequencing analysis. *Genome Med.* 7: 121.

282. Lefranc, M.-P., V. Giudicelli, P. Duroux, J. Jabado-Michaloud, G. Folch, S. Aouinti, E. Carillon, H. Duvergey, A. Houles, T. Paysan-Lafosse, S. Hadi-Saljoqi, S. Sasorith, G. Lefranc, and S. Kossida. 2015. IMGT(R), the international ImMunoGeneTics information system(R) 25 years on. *Nucleic Acids Res.* 43: D413–D422.

283. Mestas, J., and C. C. W. Hughes. 2004. Of mice and not men: differences between mouse and human immunology. *J. Immunol.* 172: 2731–2738.

284. Simonetti, G., M. Teresa, S. Bertilaccio, P. Ghia, and U. Klein. 2014. Perspectives Mouse models in the study of chronic lymphocytic leukemia pathogenesis and therapy. *Blood* 124: 1010–1019.

285. Schroeder, H. W. 2006. Similarity and divergence in the development and expression of the mouse and human antibody repertoires. *Dev. Comp. Immunol.* 30: 119–135.

286. Janeway, C. A., P. Travers, M. Walport, and M. J. Shlomchik. 2001. *Immunobiology*,.

287. Brochet, X., M. P. Lefranc, and V. Giudicelli. 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: 503–508.

288. Iacobuzio-Donahue, C. A., R. Ashfaq, A. Maitra, N. V. Adsay, G. L. Shen-Ong, K. Berg, M. A. Hollingsworth, J. L. Cameron, C. J. Yeo, S. E. Kern, M. Goggins, and R. H. Hruban. 2003. Highly Expressed Genes in Pancreatic Ductal Adenocarcinomas: A Comprehensive Characterization and Comparison of the Transcription Profiles Obtained from Three Major Technologies. *Cancer Res.* 63: 8614–8622.

289. Cock, P. J. A., T. Antao, J. T. Chang, B. A. Chapman, C. J. Cox, A. Dalke, I. Friedberg, T. Hamelryck, F. Kauff, B. Wilczynski, and M. J. L. De Hoon. 2009. Biopython: Freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* 25: 1422–1423.

290. Tange, O. 2011. GNU Parallel: the command-line power tool. *USENIX Mag.* 36: 42–47.

291. Hildebrand, M. V. 1993. The birthday problem. *Am. Math. Mon.* 100: 643.

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

293. Holm, S. 1979. A simple sequentially rejective multiple test procedure. *Scand. J. Stat.* .

294. Kinde, I., J. Wu, N. Papadopoulos, K. W. Kinzler, and B. Vogelstein. 2011. Detection and quantification of rare mutations with massively parallel sequencing. *Proc. Natl. Acad. Sci. U. S. A.* 108: 9530–5.

295. Shiroguchi, K., T. Z. Jia, P. A. Sims, and X. S. Xie. 2012. Digital RNA sequencing minimizes sequence-dependent bias and amplification noise with optimized single-molecule barcodes. *Proc. Natl.*

Acad. Sci. U. S. A. 109: 1347–52.

296. Mamedov, I. Z., O. V. Britanova, I. V. Zvyagin, M. A. Turchaninova, D. A. Bolotin, E. V. Putintseva, Y. B. Lebedev, and D. M. Chudakov. 2013. Preparing unbiased T-cell receptor and antibody cDNA libraries for the deep next generation sequencing profiling. *Front. Immunol.* 4: 456.

297. Shih, H.-Y., and M. S. Krangel. 2013. Chromatin architecture, CCCTC-binding factor, and V(D)J recombination: managing long-distance relationships at antigen receptor loci. *J. Immunol.* 190: 4915–21.

298. Wu, Y.-C. B., D. Kipling, and D. K. Dunn-Walters. 2012. Age-Related Changes in Human Peripheral Blood IGH Repertoire Following Vaccination. *Front. Immunol.* 3: 193.

299. Ademokun, A., Y.-C. Wu, V. Martin, R. Mitra, U. Sack, H. Baxendale, D. Kipling, and D. K. Dunn-Walters. 2011. Vaccination-induced changes in human B-cell repertoire and pneumococcal IgM and IgA antibody at different ages. *Aging Cell* 10: 922–930.

300. Dunn-Walters, D. K., and A. A. Ademokun. 2010. B cell repertoire and ageing. *Curr. Opin. Immunol.* 22: 514–20.

301. Boyd, S. D., Y. Liu, C. Wang, V. Martin, and D. K. Dunn-Walters. 2013. Human lymphocyte repertoires in ageing. *Curr. Opin. Immunol.* 25: 511–5.

302. Tan, Y.-C., L. K. Blum, S. Kongpachith, C.-H. Ju, X. Cai, T. M. Lindstrom, J. Sokolove, and W. H. Robinson. 2014. High-throughput sequencing of natively paired antibody chains provides evidence for original antigenic sin shaping the antibody response to influenza vaccination. *Clin. Immunol.* 151: 55–65.

303. Dekosky, B. J., T. Kojima, A. Rodin, W. Charab, G. C. Ippolito, A. D. Ellington, and G. Georgiou. 2014. In-depth determination and analysis of the human paired heavy- and light-chain antibody repertoire. *Nat. Med.* 21: 1–8.

304. Liu, L., and A. H. Lucas. 2003. IGH V3-23*01 and its allele V3-23*03 differ in their capacity to form the canonical human antibody combining site specific for the capsular polysaccharide of Haemophilus influenzae type b. *Immunogenetics* 55: 336–8.

305. Rohatgi, S., D. Dutta, S. Tahir, and D. Sehgal. 2009. Molecular dissection of antibody responses against pneumococcal surface protein A: evidence for diverse DH-less heavy chain gene usage and avidity maturation. *J. Immunol.* 182: 5570–85.

306. Galson, J. D., J. Trück, E. A. Clutterbuck, A. Fowler, V. Cerundolo, A. J. Pollard, G. Lunter, and D. F. Kelly. 2016. B-cell repertoire dynamics after sequential hepatitis B vaccination and evidence for cross-reactive B-cell activation. *Genome Med.* 8: 68.

307. Wrammert, J., D. Koutsonanos, G.-M. Li, S. Edupuganti, J. Sui, M. Morrissey, M. McCausland, I. Skountzou, M. Hornig, W. I. Lipkin, A. Mehta, B. Razavi, C. Del Rio, N.-Y. Zheng, J.-H. Lee, M. Huang, Z. Ali, K. Kaur, S. Andrews, R. R. Amara, Y. Wang, S. R. Das, C. D. O'Donnell, J. W. Yewdell, K. Subbarao, W. A. Marasco, M. J. Mulligan, R. Compans, R. Ahmed, and P. C. Wilson. 2011. Broadly cross-reactive antibodies dominate the human B cell response against 2009 pandemic H1N1 influenza virus infection. *J. Exp. Med.* 208: 181–193.

308. Yu, X., T. Tsibane, P. A. McGraw, F. S. House, C. J. Keefer, M. D. Hicar, T. M. Tumpey, C. Pappas, L. A. Perrone, O. Martinez, J. Stevens, I. A. Wilson, P. V Aguilar, E. L. Altschuler, C. F. Basler, and J. E. Crowe. 2008. Neutralizing antibodies derived from the B cells of 1918 influenza pandemic survivors. *Nature* 455: 532–6.

309. Krause, J. C., T. Tsibane, T. M. Tumpey, C. J. Huffman, R. Albrecht, D. L. Blum, I. Ramos, A. Fernandez-Sesma, K. M. Edwards, A. Garcia-Sastre, C. F. Basler, and J. E. Crowe. 2012. Human Monoclonal Antibodies to Pandemic 1957 H2N2 and Pandemic 1968 H3N2 Influenza Viruses. *J. Virol.* 86: 6334–6340.

310. Simmons, C. P., N. L. Bernasconi, A. L. Suguitan, K. Mills, J. M. Ward, N. V. V. Chau, T. T. Hien, F. Sallusto, D. Q. Ha, J. Farrar, M. D. de Jong, A. Lanzavecchia, and K. Subbarao. 2007. Prophylactic and therapeutic efficacy of human monoclonal antibodies against H5N1 influenza. *PLoS Med.* 4: e178.

311. Weitkamp, J.-H., N. Kallewaard, K. Kusuhara, E. Bures, J. V Williams, B. LaFleur, H. B. Greenberg,

and J. E. Crowe. 2003. Infant and Adult Human B Cell Responses to Rotavirus Share Common Immunodominant Variable Gene Repertoires. *J. Immunol.* 171: 4680–4688.

312. Weitkamp, J.-H. H., N. Kallewaard, K. Kusuhara, D. Feigelstock, N. Feng, H. B. Greenberg, and J. E. Crowe. 2003. Generation of recombinant human monoclonal antibodies to rotavirus from single antigen-specific B cells selected with fluorescent virus-like particles. *J. Immunol. Methods* 275: 223–237.

313. Weitkamp, J.-H., B. J. Lafleur, H. B. Greenberg, and J. E. Crowe. 2005. Natural evolution of a human virus-specific antibody gene repertoire by somatic hypermutation requires both hotspot-directed and randomly-directed processes. *Hum. Immunol.* 66: 666–76.

314. Tian, C., G. K. Luskin, K. M. Dischert, J. N. Higginbotham, B. E. Shepherd, and J. E. Crowe. 2008. Immunodominance of the VH1-46 antibody gene segment in the primary repertoire of human rotavirus-specific B cells is reduced in the memory compartment through somatic mutation of nondominant clones. *J. Immunol.* 180: 3279–88.

315. Zhou, T., I. Georgiev, X. Wu, Z.-Y. Yang, K. Dai, A. Finzi, Y. Do Kwon, J. F. Scheid, W. Shi, L. Xu, Y. Yang, J. Zhu, M. C. Nussenzweig, J. Sodroski, L. Shapiro, G. J. Nabel, J. R. Mascola, and P. D. Kwong. 2010. Structural basis for broad and potent neutralization of HIV-1 by antibody VRC01. *Science* 329: 811–7.

316. Bonsignori, M., K.-K. Hwang, X. Chen, C.-Y. Tsao, L. Morris, E. Gray, D. J. Marshall, J. A. Crump, S. H. Kapiga, N. E. Sam, F. Sinangil, M. Pancera, Y. Yongping, B. Zhang, J. Zhu, P. D. Kwong, S. O'Dell, J. R. Mascola, L. Wu, G. J. Nabel, S. Phogat, M. S. Seaman, J. F. Whitesides, M. A. Moody, G. Kelsoe, X. Yang, J. Sodroski, G. M. Shaw, D. C. Montefiori, T. B. Kepler, G. D. Tomaras, S. M. Alam, H.-X. Liao, and B. F. Haynes. 2011. Analysis of a Clonal Lineage of HIV-1 Envelope V2/V3 Conformational Epitope-Specific Broadly Neutralizing Antibodies and Their Inferred Unmutated Common Ancestors. *J. Virol.* 85: 9998–10009.

317. Corti, D., J. P. M. Langedijk, A. Hinz, M. S. Seaman, F. Vanzetta, B. M. Fernandez-Rodriguez, C. Silacci, D. Pinna, D. Jarrossay, S. Balla-Jhagjhoorsingh, B. Willems, M. J. Zekveld, H. Dreja, E. O'Sullivan, C. Pade, C. Orkin, S. A. Jeffs, D. C. Montefiori, D. Davis, W. Weissenhorn, Á. McKnight, J. L. Heeney, F. Sallusto, Q. J. Sattentau, R. A. Weiss, and A. Lanzavecchia. 2010. Analysis of Memory B Cell Responses and Isolation of Novel Monoclonal Antibodies with Neutralizing Breadth from HIV-1-Infected Individuals. *PLoS One* 5: e8805.

318. Scheid, J. F., H. Mouquet, B. Ueberheide, R. Diskin, F. Klein, T. Y. K. Oliveira, J. Pietzsch, D. Fenyo, A. Abadir, K. Velinzon, A. Hurley, S. Myung, F. Boulad, P. Poignard, D. R. Burton, F. Pereyra, D. D. Ho, B. D. Walker, M. S. Seaman, P. J. Bjorkman, B. T. Chait, and M. C. Nussenzweig. 2011. Sequence and structural convergence of broad and potent HIV antibodies that mimic CD4 binding. *Science* 333: 1633–7.

319. Racanelli, V., C. Brunetti, V. De Re, L. Caggiari, M. De Zorzi, P. Leone, F. Perosa, A. Vacca, and F. Dammacco. 2011. Antibody V(h) repertoire differences between resolving and chronically evolving hepatitis C virus infections. *PLoS One* 6: e25606.

320. Lu, D. R., Y.-C. Tan, S. Kongpachith, X. Cai, E. A. Stein, T. M. Lindstrom, J. Sokolove, and W. H. Robinson. 2014. Identifying functional anti-Staphylococcus aureus antibodies by sequencing antibody repertoires of patient plasmablasts. *Clin. Immunol.* 152: 77–89.

321. Smith, S. a, Y. Zhou, N. P. Olivarez, A. H. Broadwater, A. M. de Silva, and J. E. Crowe. 2012. Persistence of Circulating Memory B Cell Clones with Potential for Dengue Virus Disease Enhancement for Decades following Infection. *J. Virol.* 86: 2665–2675.

322. de Alwis, R., M. Beltramello, W. B. Messer, S. Sukupolvi-Petty, W. M. P. B. Wahala, A. Kraus, N. P. Olivarez, Q. Pham, J. Brian, W. Y. Tsai, W. K. Wang, S. Halstead, S. Kliks, M. S. Diamond, R. Baric, A. Lanzavecchia, F. Sallusto, and A. M. de Silva. 2011. In-depth analysis of the antibody response of individuals exposed to primary dengue virus infection. *PLoS Negl. Trop. Dis.* 5: e1188.

323. Dejnirattisai, W., A. Jumnainsong, N. Onsirisakul, P. Fitton, S. Vasanawathana, W. Limpitikul, C. Puttikhunt, C. Edwards, T. Duangchinda, S. Supasa, K. Chawansuntati, P. Malasit, J. Mongkolsapaya, and G. Screaton. 2010. Cross-Reacting Antibodies Enhance Dengue Virus Infection in Humans. *Science (80).* 328: 745–748.

324. Lucas, A. H., and D. C. Reason. 1999. Polysaccharide vaccines as probes of antibody repertoires in man. *Immunol. Rev.* 171: 89–104.

325. Sundling, C., Y. Li, N. Huynh, C. Poulsen, R. Wilson, S. O'Dell, Y. Feng, J. R. Mascola, R. T. Wyatt, and G. B. Karlsson Hedestam. 2012. High-Resolution Definition of Vaccine-Elicited B Cell Responses Against the HIV Primary Receptor Binding Site. *Sci. Transl. Med.* 4: 142ra96 LP-142ra96.

326. Poulsen, T. R., P.-J. Meijer, A. Jensen, L. S. Nielsen, and P. S. Andersen. 2007. Kinetic, affinity, and diversity limits of human polyclonal antibody responses against tetanus toxoid. *J. Immunol.* 179: 3841–3850.

327. Ertl, O. T., D. C. Wenz, F. B. Bouche, G. A. M. Berbers, and C. P. Muller. 2003. Immunodominant domains of the Measles virus hemagglutinin protein eliciting a neutralizing human B cell response. *Arch. Virol.* 148: 2195–2206.

328. Tahara, M., J. P. Bürckert, K. Kanou, K. Maenaka, C. P. Muller, and M. Takeda. 2016. Measles virus hemagglutinin protein epitopes: The basis of antigenic stability. *Viruses* 8: 1–15.

329. Blanchard-rohner, G., A. S. Pulickal, C. M. Jol-van der Zijde, M. D. Snape, A. J. Pollard, C. M. J. Der Zijde, M. D. Snape, and A. J. Pollard. 2009. Brief report Appearance of peripheral blood plasma cells and memory B cells in a primary and secondary immune response in humans IgG-PC IgA-PC IgM-PC IgG-MC. *Vaccine* 114: 4998–5002.

330. Henn, A. D., S. Wu, X. Qiu, M. Ruda, M. Stover, H. Yang, Z. Liu, S. L. Welle, J. Holden-Wiltse, H. Wu, and M. S. Zand. 2013. High-Resolution Temporal Response Patterns to Influenza Vaccine Reveal a Distinct Human Plasma Cell Gene Signature. *Sci. Rep.* 3: 2327.

331. Wardemann, H., S. Yurasov, A. Schaefer, J. W. Young, E. Meffre, and M. C. Nussenzweig. 2003. Predominant autoantibody production by early human B cell precursors. *Science* 301: 1374–7.

332. Arber, D. A. 2000. Molecular Diagnostic Approach to Non-Hodgkin's Lymphoma. *J. Mol. Diagnostics* 2: 178–190.

333. Stamatopoulos, K., C. Kosmas, N. Stavroyianni, and D. Loukopoulos. 1996. Evidence for immunoglobulin heavy chain variable region gene replacement in a patient with B cell chronic lymphocytic leukemia. *Leukemia* 10: 1551–6.

334. Bagnara, D., V. Callea, and C. Stelitano. 2006. IgV gene intraclonal diversification and clonal evolution in B-cell chronic lymphocytic leukaemia. *Br. J. Heamatol.* 133(1): 50-8 .

335. Volkheimer, A., J. Weinberg, and B. Beasley. 2007. Progressive immunoglobulin gene mutations in chronic lymphocytic leukemia: evidence for antigen-driven intraclonal diversification. *Blood* 109(4):1559-67.

336. Bashford-Rogers, R. J. M., K. A. Nicolaou, J. Bartram, N. J. Goulden, L. Loizou, L. Koumas, J. Chi, M. Hubank, P. Kellam, P. A. Costeas, and G. S. Vassiliou. 2016. Eye on the B-ALL: B-cell receptor repertoires reveal persistence of numerous B-lymphoblastic leukemia subclones from diagnosis to relapse. *Leukemia* 30: 1–10.

337. Bashford-rogers, R. 2014. Analysing the B-cell repertoire : Investigating B-cell population dynamics in health and disease. *Dissertation*. University of Cambridge (uk.bl.ethos.708880).

338. Pui, C.-H., D. Pei, E. Coustan-Smith, S. Jeha, C. Cheng, W. P. Bowman, J. T. Sandlund, R. C. Ribeiro, J. E. Rubnitz, H. Inaba, D. Bhojwani, T. A. Gruber, W. H. Leung, J. R. Downing, W. E. Evans, M. V Relling, and D. Campana. 2015. Clinical utility of sequential minimal residual disease measurements in the context of risk-based therapy in childhood acute lymphoblastic leukaemia: a prospective study. *Lancet Oncol.* 16: 465–474.

339. Dhedin, N., A. Huynh, S. Maury, R. Tabrizi, K. Beldjord, V. Asnafi, X. Thomas, P. Chevallier, S. Nguyen, V. Coiteux, J.-H. Bourhis, Y. Hichri, M. Escoffre-Barbe, O. Reman, C. Graux, Y. Chalandon, D. Blaise, U. Schanz, V. Lheritier, J.-Y. Cahn, H. Dombret, N. Ifrah, and GRAALL group. 2015. Role of allogeneic stem cell transplantation in adult patients with Ph-negative acute lymphoblastic leukemia. *Blood* 125: 2486–2496.

340. Ma, X., M. Edmonson, D. Yergeau, D. M. Muzny, O. A. Hampton, M. Rusch, G. Song, J. Easton, R. C. Harvey, D. A. Wheeler, J. Ma, H. Doddapaneni, B. Vadodaria, G. Wu, P. Nagahawatte, W. L.

Carroll, I.-M. Chen, J. M. Gastier-Foster, M. V. Relling, M. A. Smith, M. Devidas, J. M. G. Auvil, J. R. Downing, M. L. Loh, C. L. Willman, D. S. Gerhard, C. G. Mullighan, S. P. Hunger, and J. Zhang. 2015. Rise and fall of subclones from diagnosis to relapse in pediatric B-acute lymphoblastic leukaemia. *Nat. Commun.* 6: 6604.

341. Ono, Y., K. Asai, and M. Hamada. 2013. PBSIM: PacBio reads simulator--toward accurate genome assembly. *Bioinformatics* 29: 119–121.

342. Minoche, A. E., J. C. Dohm, and H. Himmelbauer. 2011. Evaluation of genomic high-throughput sequencing data generated on Illumina HiSeq and Genome Analyzer systems. *Genome Biol.* 12: R112.

343. Gilles, A., E. Meglécz, N. Pech, S. Ferreira, T. Malausa, and J.-F. Martin. 2011. Accuracy and quality assessment of 454 GS-FLX Titanium pyrosequencing. *BMC Genomics* 12: 245.

344. Jünemann, S., F. J. Sedlazeck, K. Prior, A. Albersmeier, U. John, J. Kalinowski, A. Mellmann, A. Goesmann, A. von Haeseler, J. Stoye, and D. Harmsen. 2013. Updating benchtop sequencing performance comparison. *Nat. Biotechnol.* 31: 294–296.

345. Salipante, S. J., T. Kawashima, C. Rosenthal, D. R. Hoogestraat, L. A. Cummings, D. J. Sengupta, T. T. Harkins, B. T. Cookson, and N. G. Hoffman. 2014. Performance comparison of Illumina and Ion Torrent next-generation sequencing platforms for 16S rRNA-based bacterial community profiling. *Appl. Environ. Microbiol.* 80: 7583–7591.

346. Roach, J. C., G. Glusman, A. F. A. Smit, C. D. Huff, R. Hubley, P. T. Shannon, L. Rowen, K. P. Pant, N. Goodman, M. Bamshad, J. Shendure, R. Drmanac, L. B. Jorde, L. Hood, and D. J. Galas. 2010. Analysis of Genetic Inheritance in a Family Quartet by Whole-Genome Sequencing. *Science (80).* 328: 636–639.

347. Reumers, J., P. De Rijk, H. Zhao, A. Liekens, D. Smeets, J. Cleary, P. Van Loo, M. Van Den Bossche, K. Catthoor, B. Sabbe, E. Despierre, I. Vergote, B. Hilbush, D. Lambrechts, and J. Del-Favero. 2011. Optimized filtering reduces the error rate in detecting genomic variants by short-read sequencing. *Nat. Biotechnol.* 30: 61–68.

348. Gadala-Maria, D., G. Yaari, M. Uduman, and S. H. Kleinstein. 2015. Automated analysis of high-throughput B-cell sequencing data reveals a high frequency of novel immunoglobulin V gene segment alleles. *Proc. Natl. Acad. Sci.* 112: 201417683.

349. Jabara, C. B., C. D. Jones, J. Roach, J. A. Anderson, and R. Swanstrom. 2011. Accurate sampling and deep sequencing of the HIV-1 protease gene using a Primer ID. *Proc. Natl. Acad. Sci.* 108: 20166–20171.

350. Hiatt, J. B., R. P. Patwardhan, E. H. Turner, C. Lee, and J. Shendure. 2010. Parallel, tag-directed assembly of locally derived short sequence reads. *Nat. Methods* 7: 119–122.

351. Casbon, J. A., R. J. Osborne, S. Brenner, and C. P. Lichtenstein. 2011. A method for counting PCR template molecules with application to next-generation sequencing. *Nucleic Acids Res.* 39: e81–e81.

352. Lou, D. I., J. A. Hussmann, R. M. Mcbee, A. Acevedo, R. Andino, and W. H. Press. 2013. High-throughput DNA sequencing errors are reduced by orders of magnitude using circle sequencing. *Proc. Natl. Acad. Sci.* 110: 19872–19877.

353. Deakin, C. T., J. J. Deakin, S. L. Ginn, P. Young, D. Humphreys, C. M. Suter, I. E. Alexander, and C. V. Hallwirth. 2014. Impact of next-generation sequencing error on analysis of barcoded plasmid libraries of known complexity and sequence. *Nucleic Acids Res.* 42.

354. Watson, C. T., K. M. Steinberg, J. Huddleston, R. L. Warren, M. Malig, J. Schein, A. J. Willsey, J. B. Joy, J. K. Scott, T. A. Graves, R. K. Wilson, R. A. Holt, E. E. Eichler, and F. Breden. 2013. Complete Haplotype Sequence of the Human Immunoglobulin Heavy-Chain Variable, Diversity, and Joining Genes and Characterization of Allelic and Copy-Number Variation. *Am. J. Hum. Genet.* 92: 530–546.

355. Kidd, M. J., Z. Chen, Y. Wang, K. J. Jackson, L. Zhang, S. D. Boyd, A. Z. Fire, M. M. Tanaka, B. A. Gaëta, and A. M. Collins. 2012. The inference of phased haplotypes for the immunoglobulin H chain V region gene loci by analysis of VDJ gene rearrangements. *J. Immunol.* 188: 1333–40.

356. Wang, Y., K. J. Jackson, B. Gäeta, W. Pomat, P. Siba, W. A. Sewell, and A. M. Collins. 2011. Genomic screening by 454 pyrosequencing identifies a new human IGHV gene and sixteen other new

IGHV allelic variants. *Immunogenetics* 63: 259–265.

357. Howie, B., A. M. Sherwood, A. D. Berkebile, J. Berka, R. O. Emerson, D. W. Williamson, I. Kirsch, M. Vignali, M. J. Rieder, C. S. Carlson, and H. S. Robins. 2015. High-throughput pairing of T cell receptor α and β sequences. *Sci. Transl. Med.* 7: 301ra131.

358. Arnaoty, A., V. Gouilleux-Gruart, S. Casteret, B. Pitard, Y. Bigot, and T. Lecomte. 2013. Reliability of the nanopheres-DNA immunization technology to produce polyclonal antibodies directed against human neogenic proteins. *Mol. Genet. Genomics* 288: 347–363.

359. Ferraro, B., M. P. Morrow, N. A. Hutnick, T. H. Shin, C. E. Lucke, and D. B. Weiner. 2011. Clinical Applications of DNA Vaccines: Current Progress. *Clin. Infect. Dis.* 53: 296–302.

360. Li, L., F. Saade, and N. Petrovsky. 2012. The future of human DNA vaccines. *J. Biotechnol.* 162: 171–82.

361. Rajčáni, J., T. Moško, and I. Režuchová. 2005. Current developments in viral DNA vaccines: shall they solve the unsolved? *Rev. Med. Virol.* 15: 303–325.

362. Liu, M. A. 2003. DNA vaccines: a review. *J. Intern. Med.* 253: 402–10.

363. Rota, P. A., D. A. Featherstone, and W. J. Bellini. 2009. Molecular epidemiology of measles virus. *Curr. Top. Microbiol. Immunol.* 330: 129–50.

364. Tahara, M., Y. Ito, M. a Brindley, X. Ma, J. He, S. Xu, H. Fukuhara, K. Sakai, K. Komase, P. a Rota, R. K. Plemper, K. Maenaka, and M. Takeda. 2013. Functional and Structural Characterization of Neutralizing Epitopes of Measles Virus Hemaggltinin Protein. *J. Virol.* 87: 666–675.

365. Klingele, M., H. K. Hartter, F. Adu, W. Ammerlaan, W. Ikusika, and C. P. Muller. 2000. Resistance of recent measles virus wild-type isolates to antibody-mediated neutralization by vaccinees with antibody. *J. Med. Virol.* 62: 91–8.

366. Tamin, A., P. A. Rota, Z. D. Wang, J. L. Heath, L. J. Anderson, and W. J. Bellini. 1994. Antigenic analysis of current wild type and vaccine strains of measles virus. *J. Infect. Dis.* 170: 795–801.

367. De Swart, R. L., S. Yü Ksel, C. N. Langerijs, C. P. Muller, and A. D. M. E. Osterhaus. Depletion of measles virus glycoprotein-specific antibodies from human sera reveals genotype- specific neutralizing antibodies. *J. Gen. Virol.* 90(12):2982-9.

368. Boyd, S. D., B. A. Gaeta, K. J. Jackson, A. Z. Fire, E. L. Marshall, J. D. Merker, J. M. Maniar, L. N. Zhang, B. Sahaf, C. D. Jones, B. B. Simen, B. Hanczaruk, K. D. Nguyen, K. C. Nadeau, M. Egholm, D. B. Miklos, J. L. Zehnder, and A. M. Collins. 2010. Individual Variation in the Germline Ig Gene Repertoire Inferred from Variable Region Gene Rearrangements. *J. Immunol.* 184: 6986–6992.

369. Zuckerman, N. S., K. J. McCann, C. H. Ottensmeier, M. Barak, G. Shahaf, H. Edelman, D. Dunn-Walters, R. S. Abraham, F. K. Stevenson, and R. Mehr. 2010. Ig gene diversification and selection in follicular lymphoma, diffuse large B cell lymphoma and primary central nervous system lymphoma revealed by lineage tree and mutation analyses. *Int. Immunol.* 22: 875–887.

370. Watson, C. T., and F. Breden. 2012. The immunoglobulin heavy chain locus: genetic variation, missing data, and implications for human disease. *Genes Immun.* 13: 363–373.

371. Feeney, A. J., M. J. Atkinson, M. J. Cowan, G. Escuro, and G. Lugo. 1996. A defective Vkappa A2 allele in Navajos which may play a role in increased susceptibility to haemophilus influenzae type b disease. *J. Clin. Invest.* 97: 2277–82.

372. Pommié, C., S. Levadoux, R. Sabatier, G. Lefranc, and M.-P. Lefranc. 2004. IMGT standardized criteria for statistical analysis of immunoglobulin V-REGION amino acid properties. *J. Mol. Recognit.* 17: 17–32.

Presentations and Meeting Participations

2013

Participated in the International Workshop "Zebrafish: an animal model in biomedical research" in London, United Kingdom; 23.05.2013 – 24.05.2013

Poster presentation "Probing the B cell repertoire with low molecular weight hapten-conjugates" at the ESF-EMBO Symposium with support from EFIS "B Cells from Bedside to Bench and Back again" in Pultusk, Poland; 02.09.2013 – 07.09.2013.

Poster presentation "Probing the B cell repertoire with low molecular weight hapten-conjugates" at the Life Science PhD Days in Luxembourg, Luxembourg; 10.09.2013 – 11.09.2013.

Poster presentation "Probing the B cell repertoire with low molecular weight hapten-conjugates" at the 43th Annual Meeting of the German Society for Immunology 2013 in Mainz, Germany; 11.09.2013 – 14.09.2013.

2014

Poster presentation and talk "Probing the B cell repertoire with low molecular weight hapten-conjugates" at the Gordon Research Conference and Seminar on "Antibody Biology and Engineering" in Lucca, Italy, 23.03.2014 – 28.03.2014.

Invited talk "New approaches to vaccine design: How next-generation sequencing can help to overcome current vaccine limitations" at the Séance de Communications Courtes Printemps de 2014 in Luxembourg, Luxembourg; 19.05.2014

Talk "What can the in-depth analysis of the B cell repertoire tell us about the past exposure to immunological challenges?" at the Annual Meeting of the AK Vakzine in Göttingen, Germany; 22.05.2014 – 23.05.2014.

Participated in the ECCB 2014: IMGT, the Global Reference in Immunogenetics and Immunoinformatics in Strasbourg, France; 07.09.2014

Talk "Identification of antigen specific signatures in the B cell repertoire of OmniRat™ and humans" at the 17th SarLorLux Meeting on Virus Research in Nancy, France; 11.09.2014.

Poster Presentation "In-depth analysis of the B cell repertoire in response to an immune-prophylactic strategy against environmental carcinogens" at the 44th Annual Meeting of the German Society for Immunology 2014 in Bonn, Germany; 17.09.2014 – 20.09.2014

2015

Invited talk "Approaching the B cell repertoire with NGS: from sample preparation to network clustering" at the Department Seminar of the Institute for Molecular Medicine, Johannes-Gutenberg University of Mainz in Mainz, Germany; 29.01.2015

Participated in the Nature Conference: Immune Profiling in Health and Disease; Seattle, WA, United States; 09.09.2015 – 11.09.2015

Talk "Genetic and Structural Determinants of the B cell Repertoire in Response to Antigenic Challenges" at the Department Seminar of the Department of Infection and Immunity of the Luxembourg Institute of Health in Esch-sur-Alzette, Luxembourg; 16.09.2015

2016

Poster presentations and talk "Exploiting functional IG repertoire convergence to determine vaccination elicited IG sequences *in silico*" and "Anchored lineage reconstruction reveals b cell repertoire convergence on measles virus hemagglutinin protein in three independent vaccination cohorts" at the Gordon Research Conference and Seminar on "Antibody Biology and Engineering" in Galveston, TX, United states; 19.03.2016 – 25.03.2016

Poster presentations and talk "Exploiting functional IG repertoire convergence to determine vaccination elicited IG sequences *in silico*" and "Anchored lineage reconstruction reveals b cell repertoire convergence on measles virus hemagglutinin protein in three independent vaccination cohorts" at the 2nd Adaptive Immune Receptor Repertoire (AIRR) Community Meeting in Rockville, MD, United States; 27.06.2016 – 30.06.2016.

Talk "Identification of Antigen-driven B cell Receptor Sequences from HTS Datasets using DESeq2" at the Department Seminar of the Department of Infection and Immunity of the Luxembourg Institute of Health in Esch-sur-Alzette, Luxembourg; 12.09.2016

2017

Invited talk "Wrestling with the B cell repertoire – Using High-Throughput Sequencing to understand B cell repertoire dynamics in health and disease" at the Department of Immunology of the University of Toronto in Toronto, Canada; 12.05.2017.

Poster presentation and Talk "*In silico* identification of Antigen-driven B cell Receptor Sequences from HTS Datasets of transgenic rats carrying human Ig-genes using DESeq2" at the Batsheva de Rothschild Seminar on: Stochasticity and Control in the Dynamics and Diversity of Immune Repertoires held at the Israel Institute for Advanced Studies in Jerusalem, Israel; 18.06.2017 – 23.06.2017.

Participated in the 3rd Adaptive Immune Receptor Repertoire (AIRR) Community Meeting in Rockville, MD, United States; 03.12.2017 – 06.12.2017.

Publications

Dubois, A.R.S.X.; **Bürckert, J.-P.**; Sinner, R.; Faison, W.J.; Molitor, A.M.; Muller, C.P. High-Resolution Analysis of the B Cell Repertoire Before and After Polyethylene Glycol Fusion Reveals Preferential Fusion of Rare Antigen-Specific B Cells *Hum. Antibodies*, **2016**, *24 (1-2)*, 1-15.

Tahara, M.; **Bürckert, J.-P.**; Kanou, K.; Maenaka, K.; Muller, C.P.; Takeda, M. Measles Virus Hemagglutinin Protein Epitopes: The Basis of Antigenic Stability. *Viruses,* **2016**, *8*, 216.

Hahn, M.; **Bürckert, J.-P.**; Klebow, S.; Luttenberger, C.; Hess, M.; Al-Maarri, M.; Vogt, M.; Reißig, S.; Hallek, M.; Buch, T.; Wienecke-Baldacchino, A.; Muller, C.P.; Pallasch, C.; Wunderlich, T.; Waisman, A.; Hövelmeyer, N. Aberrant splicing of the tumor suppressor CYLD promotes the development of chronic lymphocytic leukemia via sustained NF-κB signaling, *Leukemia, accepted Jun 01, 2017*.

Bürckert, J.-P.; Dubois, A.R.S.X.; Faison, W.J.; Farinelle, S.; Charpentier, E.; Sinner, R.; Wienecke-Baldacchino, A.; Muller, C.P. Functionally Convergent Immunoglobulin Heavy Chain Sequences in Transgenic Rats Expressing Human B Cell Receptors in Response to Tetanus Toxoid and Measles Virus Antigens, *Frontiers in Immunology, accepted Dec 05, 2017*.

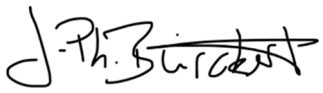
Rubelt, F.; Busse, C.E.; Bukhari, S.A.C.; **Bürckert J.-P.**; Mariotti-Ferrandiz, E.; Cowell, L.G.; Watson, C.T.; Marthandan, N.; Faison, W.J.; Hershberg, U.; Laserson, U.; Corrie, B.D.; Davis, M.D.; Peters, B.; Lefranc, M.-P.; Scott, J.K.; Breden, F.; The AIRR Community; Luning Prak, E.T.; Kleinstein, S.H. AIRR Community Recommendations for Sharing Immune Repertoire Sequencing Data. *Nature Immunol*, **2017**,18 (12), 1274-1278.

Bürckert, J.-P.; Faison, W.J.; Dubois, A.R.S.X.; Sinner, R.; Hunewald, O.; Wienecke-Baldacchino, A.; Brieger, A.; Muller, C.P.; Robust sequencing of immunoglobulin heavy chain transcripts from Balb/C mice using single side unique molecular identifiers on an Ion Torrent PGM. *Manuscript under revision in Oncotarget, submitted Aug 31, 2017.*

Erklärung

Hiermit versichere ich, dass ich die vorliegende Dissertationsschrift selbstständig verfasst, und keine anderen als die angegebenen Hilfsquellen verwendet habe. Die Arbeit wurde bisher weder im Inland, noch im Ausland in gleicher oder ähnlicher Form einer anderen Prüfungsbehörde vorgelegt.

Trier, 16.12.2017



Jean-Philippe Bürckert