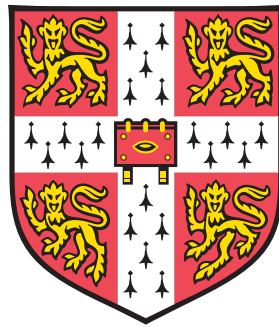# Analysing the B-cell repertoire:

Investigating B-cell population dynamics in health and disease.

University of Cambridge
Jesus College

A thesis submitted for the degree of
Doctor of Philosophy

Rachael Bashford-Rogers

The Wellcome Trust Sanger Institute,
Wellcome Trust Genome Campus,
Hinxton, Cambridge, CB10 1SA,
United Kingdom.

August 2014

Declaration

This thesis describes work carried out between January 2011 and August 2014 under the supervision of Prof. Paul Kellam and Prof. Allan Bradley at the Wellcome Trust Sanger Institute, while member of Jesus College, University of Cambridge. This thesis is the result of my own work and includes nothing that is the outcome of work done in collaboration except where specifically indicated in the text.

This thesis does not exceed the specified length limit of 300 pages as defined by the Biology Degree Committee at approximately 43,046 words long, 207 pages. This thesis has been typeset in 12pt font according to the specifications defined by the Board of Graduate Studies and the Biology Degree Committee.

Rachael Bashford-Rogers
August 2014.

Abstract

The adaptive immune response selectively expands B- and T-cell clones following antigen recognition by B- and T-cell receptors (BCR and TCR) respectively. Next-generation sequencing of these extensive, sequence-diverse repertoires is a powerful tool for dissecting these cell populations at high-resolution. In this thesis, we develop novel, robust, sensitive and reproducible computational approaches for analysing B-cell populations using high-throughput BCR sequencing.

We show that BCR sequences can be organised into networks based on sequence diversity, with differences in network connectivity clearly distinguishing between diverse repertoires of healthy individuals and clonally expanded repertoires from individuals with clonal B-cell disorders, such as chronic lymphocytic leukaemia (CLL) and B-cell acute lymphocytic leukaemia (B-ALL). Network population measures quantify the BCR clonality status and are robust to sampling and sequencing depths. The detection of BCR sequences at levels as low as 1 in 10^7 RNA molecules highlights the clinical utility of BCR sequencing in both detecting and monitoring dynamics of malignant cells throughout treatment with exquisite sensitivity. We show that time-dependent evolution of BCR repertoire provides a powerful means of assessing B-cell tumor clone evolution and response to therapy, as well as revealing insights into the biology of these diseases through phylogenetic methods.

Using this data, we integrated both theoretical and experimental frameworks of BCR sequencing to assess the biases and reproducibilities of different sequencing depths and technologies, amplification methods and starting material to confirm the biological insights gained from data interpretation. Mapping BCR and TCR repertoires promises to transform our understanding of adaptive immunity, with applications ranging from exploring infection and vaccination dynamics to determining evolutionary pathways for haematological malignancies and monitoring of minimal residual disease following chemotherapy.

Acknowledgements

Contents

List of Figures

List of Tables

Nomenclature

5' RACE	5' ended Rapid Amplification of cDNA Ends
AID	Activation-induced DNA-cytosine deaminase
ALL	Acute lymphoblastic leukaemia
BCR	B-cell receptor
BLAST	Basic Local Alignment Search Tool
cDNA	Complementary DNA (DNA synthesised from mRNA template)
CDR1, 2, 3	Complementary determining region 1, 2, 3
CLL	Chronic lymphocytic leukaemia
DNA	Deoxyribonucleic acid
FL	Follicular lymphoma
FWR1, 2, 3	Framework region 1, 2, 3
GAPDH	Glyceraldehyde 3-phosphate dehydrogenase
Ig	Immunoglobulin
IgH	Heavy chain immunoglobulin
IgHD	Heavy chain diversity immunoglobulin gene
IgHJ	Heavy chain joining immunoglobulin gene
IgHV	Heavy chain variable immunoglobulin gene
IgK	Kappa (light) chain Immunoglobulin
IgL	Lambda (light) chain Immunoglobulin
LCL	Lymphoblastoid cell line
LDA	Linear discriminant analysis
mRNA	Messenger RNA
MRD	Minimal residual disease
PCR	Polymerase chain reaction
qPCR	Quantitative real-time PCR
RNA	Ribonucleic acid
RT-PCR	Reverse transcription polymerase chain reaction
SHM	Somatic hypermutation
SLL	Small lymphocytic lymphoma
TCR	T-cell receptor
WBC	White blood count

Chapter 1

1. Introduction

1.1. Structure of the adaptive immune system

B-cells and T-cells are key to the immune response, and are crucial to the human body's ability to protect against infection and cancer by producing antibodies that can bind to pathogens and removing infected cells. Dysregulation of B- and T-cells can lead to life-threatening disorders, where understanding B- and T-cell population structures and dynamics are of considerable clinical importance (Tonegawa, 1983), particularly in response to infection (Foster, 2008), malignancies (Mao et al., 2007) and autoimmunity.

The adaptive immune response selectively expands B- and T- cell clones following antigen recognition by B- or T- cell receptors respectively. B-cell receptors (BCRs) mediate the humeral adaptive immune response by the binding antigens licensing B-cell clonal proliferation and antibody production. B-cells play a pivotal role in preventing and clearing infection as well as offering protection against antigen re-challenge. T-cells also play an important role in the adaptive immune response by an number of mechanisms, including co-stimulation of B-cells to differentiate and produce antibodies, stimulating clearance of antigen by other cells of the immune system, direct killing of infected cells, and regulation of immune responses. This section describes antibody structures and functions, B-cell development and the generation of B-cell BCR diversity.

1.1.1. Structure of antibodies

The main function of a B-cell is to produce and secrete immunoglobulin (Ig). Immunoglobulins are glycoproteins that bind antigens with high specificity to facilitate the clearance of antigen either by binding other parts of the immune system or by direct binding of antigen thus inhibiting antigen activity, known as neutralisation. The basic structural units of all immunoglobulins are very similar, consisting of two identical heavy chain (IgH) and two identical light (IgL) chain proteins, linked by disulphide bridges (**Figure 1.1**). The sites at the tip of the antigen-binding (Fab) regions are highly diversified and formed from the variable domains of the heavy (IgH) and light chains (IgL), both generated during B-cell development by

highly regulated gene rearrangements in the B-cell receptor gene loci, addressed in detail in Section 1.1.3 (Woof and Burton, 2004b, Lydyard et al., 2000, Tonegawa, 1983). The trunk of the heavy chain protein is known as the constant region, and is defined by the antibody isotype. Although the different isotypes of immunoglobulin have distinct biological activities, structures and distributions throughout the body, and trigger different effector mechanisms, all isotypes of immunoglobulin (IgA, IgD, IgE, IgG, and IgM) can be expressed as a membrane-associated form on the surface of the B-cell (B-cell receptor) or as a secreted form (antibody). The membrane-associated and secreted Ig forms differ only at the carboxy-terminus of the heavy chain, where a hydrophobic anchor sequence forms part of the membrane-associated Ig protein, and a hydrophilic sequence forms part of the secreted Ig sequence. Differential RNA splicing of the same RNA transcript, known as alternative splicing, generates these two Ig forms (Alt et al., 1980).
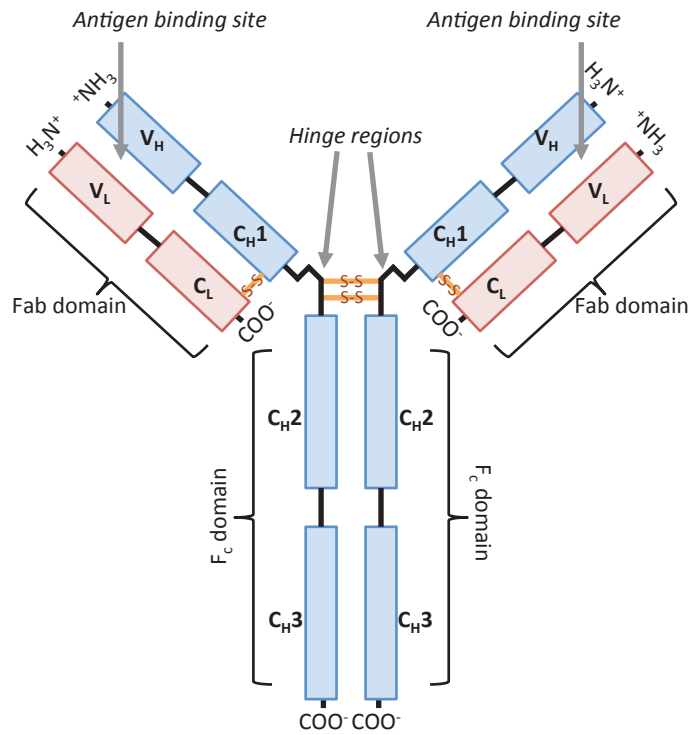
Figure 1.1. Representative structure of an antibody.

V_H indicates the heavy chain variable regions, generated from rearrangement of the Ig gene locus, and C_H1-3 are the constant regions of the antibody in the F_c domain. Likewise, V_L indicates the light chain variable region comprised of IgLV-J recombined genes. S-S indicated regions with disulphide bonds. Fab domains denote the antigen-binding regions. Adapted from Lydyard et al. (Lydyard et al., 2000).

1.1.2. Antibody isotypes

The structure of the heavy chain constant (C_H) gene defines the effector function of the immunoglobulin, and the paired IgH and IgL variable regions define the antigen specificity. The five isotypes of immunoglobulin (IgA, IgG, IgD, IgE, and IgM corresponding to α, γ, δ, ε and μ chains in the IgH gene locus) each form different effector functions. Despite the amino-acid differences between the isotypes, each C_H gene folds into similar structures consisting of β sheet linked together by inter-chain disulphide bonds. Each C_H gene is divided into domains defined as C_H1, C_H2 and C_H3 for IgA, IgD and IgG), and C_H1, C_H2, C_H3 and C_H4 for IgM and IgE. The C_H2-C_H3(-C_H4) domains comprise the region of the antibody, known as the F_c fragment, that mediates effector function by the binding of F_c receptors (FcRs) on effector cells or by activating other immune pathways such as complement activation. Each isotype differs in terms of size, complement fixation and receptor binding, such as to FcRs (Woof and Burton, 2004a). The immunoglobulin isotype also influences binding kinetics of the antibody by different binding efficiencies to the different FcRs. A summary of isotype properties is given in Table 1.1.

Table 1.1. Properties of immunoglobulin isotypes.

Adapted from (Schroeder and Cavacini, 2010) and (Burton and Woof, 1992).

Immunoglobulin isotype	Structure	Serum concentration (mg mL^{-1})	Half-life (days)	Placental transfer*	Complement activation*	Other functions
IgG1	Monomer	9	23	+++	+	Antiviral or secondary response
IgG2	Monomer	3	23	+	+	Neutralize toxins
IgG3	Monomer	1	7	+++	+++	Viral response
IgG4	Monomer	0.5	23	+++	-	Allergy
IgM	Pentamer/hexameric**	1.5	5	-	+++	Primary response
IgA1	Monomer or dimer	3	6	-	+	Direct neutralisation of toxins, viruses and bacteria
IgA2	Monomer or dimer	0.5	6	-	-	Direct neutralisation of toxins, viruses and bacteria
IgD	Monomer	0.04	3	-	-	Homeostasis
IgE	Monomer	0.0003	0.5	-	-	Allergy

*Major effector functions of each isotype are denoted by +++, lesser functions are denoted by +, and – denotes lack of corresponding function.
** (Davis and Shulman, 1989).

- *IgM*

Monomeric IgM is expressed on the surface of naïve B-cells. After maturation and antigen stimulation, a pentameric form of IgM is secreted, where each unit is linked by disulphide bonds in the C_H4 region. This pentameric form of IgM is linked to joining chains, known as J-chains, by disulphide bonds, which helps mucosal surface secretion. IgM functions by binding antigen for destruction, known as opsonisation, and fixing complement. The monomeric form of IgM generally has low affinity for antigen as the B-cells that produce IgM are early in differentiation and the V(-D)-J regions have not undergone somatic hypermutation. However, the pentameric form may have a high total binding strength, known as avidity, due to the multimeric interactions, which is particularly enhanced if the antigen itself has multiple repeating units, thus is very efficient at opsonisation (Matsuda et al., 1998).

- *IgD*

Low levels of circulating IgD are found in the serum, and the half-life of serum IgD is short. IgD antibodies have no known effector function, but IgD can react with specific bacterial proteins, such as *Moraxella catarrhalis* outer membrane

protein MID (Riesbeck and Nordstrom, 2006). Most B-cells expressing surface IgD also express surface IgM. It is thought that membrane-bound IgD contributes to the regulation of B-cell fate at particular developmental stages (Geisberger et al., 2006).

- **_IgG_**

The predominant immunoglobulin isotype is IgG, which has the longest serum half-life. There are four subclasses of IgG (IgG1, IgG2, IgG3, and IgG4), numbered in order of their serum levels in the blood of healthy individuals. Each subclass differs in terms of their antibody flexibility, as shown by crystal structure analysis, and their affinities to different F_c receptors, and ability to fix complement. For example, the different subtypes of IgG also differ in terms of their disulfide bond structures, where multiple disulfide bond structures have been observed for IgG2 and IgG4 subtypes (Liu and May, 2012). These subclasses also differ in terms of their trans-placental transport and participation in secondary immune responses (summarised in Table 1.1). Response to protein antigens is generally facilitated by IgG1 and IgG3, whereas response to polysaccharide antigens is typically facilitated by IgG2 and IgG4. IgG antibodies directly neutralise toxins and viruses as well as activating other parts of the immune system, such as the classical pathway, which involves a cascade of immune protein production leading to antigen elimination (Cavacini et al., 2003, Scharf et al., 2001).

- **_IgA_**

Although IgA antibodies have relatively high levels in the serum, they are predominantly observed on mucosal surfaces and in secretions, such as saliva and breast milk (Woof and Mestecky, 2005). Serum IgA generally exists as a monomer, but at the mucosal surfaces, secretary IgA is a dimer. The dimeric form associates with two other proteins, a J-chain and a secretary component chain, all linked by disulphide bonds. There are two subclasses of IgA (IgA1 and IgA2) that differ mainly in the hinge regions (indicated on **Figure 1.1**). The shorter hinge region in IgA2 decreases sensitivity to bacterial proteases, and predominates in many mucosal secretions such as the genital tract, whereas over 90% of serum IgA is of the IgA1 form. The main function of IgA is direct neutralisation of toxins, viruses and bacteria and the prevention of binding to mucosal surfaces. Intracellular IgA is thought to be

important in the prevention of bacterial and viral infections (Corthesy, 2007), where intracellularisation is thought to be mediated through polymeric immunoglobulin receptor (pIgR)-mediated endocytosis at the basolateral surfaces of epithelial cells followed by transcytosis (Mazanec et al., 1993, Lamm, 1997, Corthesy and Kraehenbuhl, 1999).

- ***IgE***

IgE is a very potent immunoglobulin even though it has the lowest serum levels and the shortest half-life. This immunoglobulin is made in response to parasitic worm infections, but also associated with hypersensitivity and allergic reactions. The high potency is, in part, due to the high affinity to the FcεRI receptor that is expressed on mast cells, eosinophils, basophils, and Langerhans cells (Corthesy, 2007).

1.1.3. Generation of antibody diversity

There are two different mechanisms for generating somatic diversity in B-cell receptor sequences: DNA rearrangement of the V, D and J germline segments and somatic mutation.

1.1.4. B-cell development

1.1.4.1. Immunoglobulin gene rearrangements

B-cells develop from hematopoietic stem cells and differentiate through several maturation stages in the bone marrow, after which they migrate through the peripheral blood to the secondary lymphoid organs (**Figure 1.2**). Survival and maturation of B-cells at each stage of development depends on signals transmitted through cell-surface ligands (Koopman et al., 1994, Mackay et al., 2005) and B-cell development requires the joint action of many cytokines and transcription factors that positively and negatively regulate gene expression (Milne and Paige, 2006, Hardy et al., 2007).

Early progenitor B cells (pro-B cells) are the first stage of differentiation, that rearrange DNA segments of the *Ig* loci to generate unique IgH sequence. The germline IgH chain gene locus encodes for multiple distinct copies of the variable (V), diversity (D) and joining (J) genes, which are separated by over 100 kbp from a much smaller number of DNA segments encoding the constant genes (**Figure 1.3** and **Figure 1.4**) (Lydyard et al., 2000). The total number of reportedly functional IgHV

genes in humans currently range from between 45 to 60 due to variable levels of gene loci heterozygosity (Boyd et al., 2010b), 27 IgHD genes and 6 IgHJ genes (Lefranc et al., 2009). This organisation is maintained in most somatic cell types, but in each individual mature B-cell a unique DNA rearrangement event has occurred. Functional immunoglobulin genes are generated by the process of site-specific recombination by recombination activating genes 1 (RAG1) and 2 (RAG2) through the deletion of intervening DNA (Schatz and Swanson, 2010), creating a IgH gene containing one V, one D and one J gene (VDJ) encoding the protein sequence for the antigen binding region of the IgH protein (**Figure 1.5**) (Lydyard et al., 2000, Latchman, 2005). The order of recombination events is highly regulated, where first the IgHD and IgHJ genes are brought together forming the pro-B-cell. Then the IgHV is recombined to form IgHV-D-J gene recombination of the pre-B-cell. The imprecise joining of the V, D and J gene segments leads to the introduction of random deletions and insertions of nucleotides during recombination events, resulting in sequence diversification at the junctional regions (Tonegawa, 1983). Further mechanisms that contribute to the generation of diversity include alternative IgHD reading frames, IgHD-IgHD fusions, and imprecise joining at the IgHD-J and IgHV-D junctions (Kalinina et al., 2011). These pre-B-cells are selected for functional heavy chain by IgV-D-J μ exon protein expression and IgH assembly by pairing and cell-surface expression with a surrogate light chain protein and Igα/Igβ (Vettermann and Schlissel, 2010, ten Boekel et al., 1998). This complex, along with the participation of the BCR signalling cascade, such as Syk, B-cell linker, and phosphoinositide 3-kinase, give the pre-B-cell signals for survival and proliferation (Fuentes-Panana et al., 2004).

Likewise, each IgL chain locus encodes for multiple distinct copies of variable (V) gene segments and joining (J) gene segments, in addition to a gene segment encoding the λ chain and the κ chain for the λ and the κ gene loci respectively, creating an immature B-cell (Woof and Burton, 2004a). When a cell has successfully rearranged a IgH gene, the B-cell begins to rearrange the κ light chain genes to bring together a κV and κJ. If this produces a functional κ light chain, the B-cell expresses and transcribes the heavy and κ light chains, else it then attempts to rearrange the κ light chain genes on the other chromosome. If the cell is unsuccessful at producing a functional κ light chain BCR from both chromosome κ light chain loci, then the B-cell attempts to rearrange the λ light chain genes to bring together a λV and λJ.

Likewise, if this produces a functional λ light chain, the B-cell expresses and transcribes the heavy and λ light chains.

Each mature, but antigen naïve, B-cell expresses a single BCR sequence, where the B-cell population has sufficient BCR diversity for initial recognition of all potential antigens in their environment (Dorner et al., 1998, Lydyard et al., 2000). After functional V-(D)-J recombination of IgH and IgL chain genes, naïve B-cells transcribe the IgH and IgL genes and are able to produce IgD and IgM immunoglobulin isotype by alternative splicing of the transcript to fuse the μ and δ exon to the IgHJ exon respectively (Geisberger et al., 2006). Later in development and in response to stimulation, B-cells can be signalled to produce other isotypes (IgA, IgG and IgE) (Schroeder and Cavacini, 2010) by alternative transcript splicing and class switch recombination (Section 1.1.6). These naïve B-cells typically migrate to the s such as spleen and lymph nodes where they may encounter antigens (Honjo et al., 2002). Each recombination creates a B-cell clone that can expand to form a lineage expressing a clonal B-cell receptor (a plasma-membrane anchored IgH/L complex) consisting of a heavy and light (either λ or κ) chains.
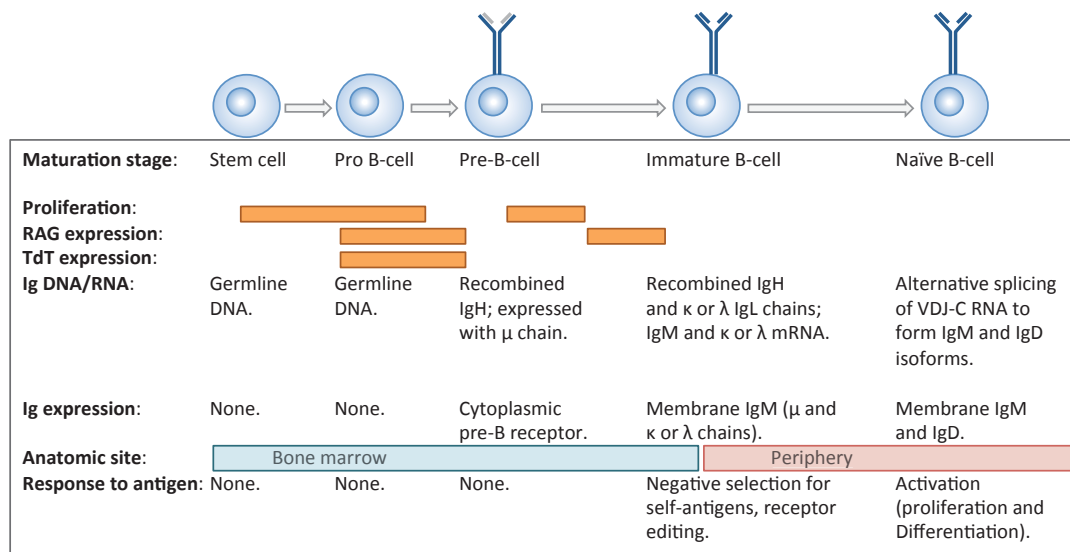
Figure 1.2. Stages of B-cell maturation.

Each stage is indicated by their Ig status, anatomical site and cell-surface marker expression, and the ability to respond to antigen. The haematopoietic stem cell develops into a naïve B-cell through a number of differentiation steps, where the IgH germline DNA is recombined in the pre-B-cell, leading to IgV-D-J μ exon protein expression and IgH assembly with a variant surrogate light chain protein. The light chain (λ or κ chain) then recombines to create an immature B-cell, with membrane assembly of IgM. Naïve B-cells are able to alternatively splice the constant region mRNA to express IgM and IgD. Increased proliferation rates, RAG and TdT expression is highly regulated and occur at specific points in differentiation, indicated by the orange bars, and the sites where each B-cell differentiation stage is indicated by the blue and red bars for the bone marrow and periphery respectively. B-cell Ig rearrangement status, Ig expression and ability to respond to antigen binding is indicated for each differentiation stage. Adapted from Edwards *et al*. (Edwards and Cambridge, 2006).
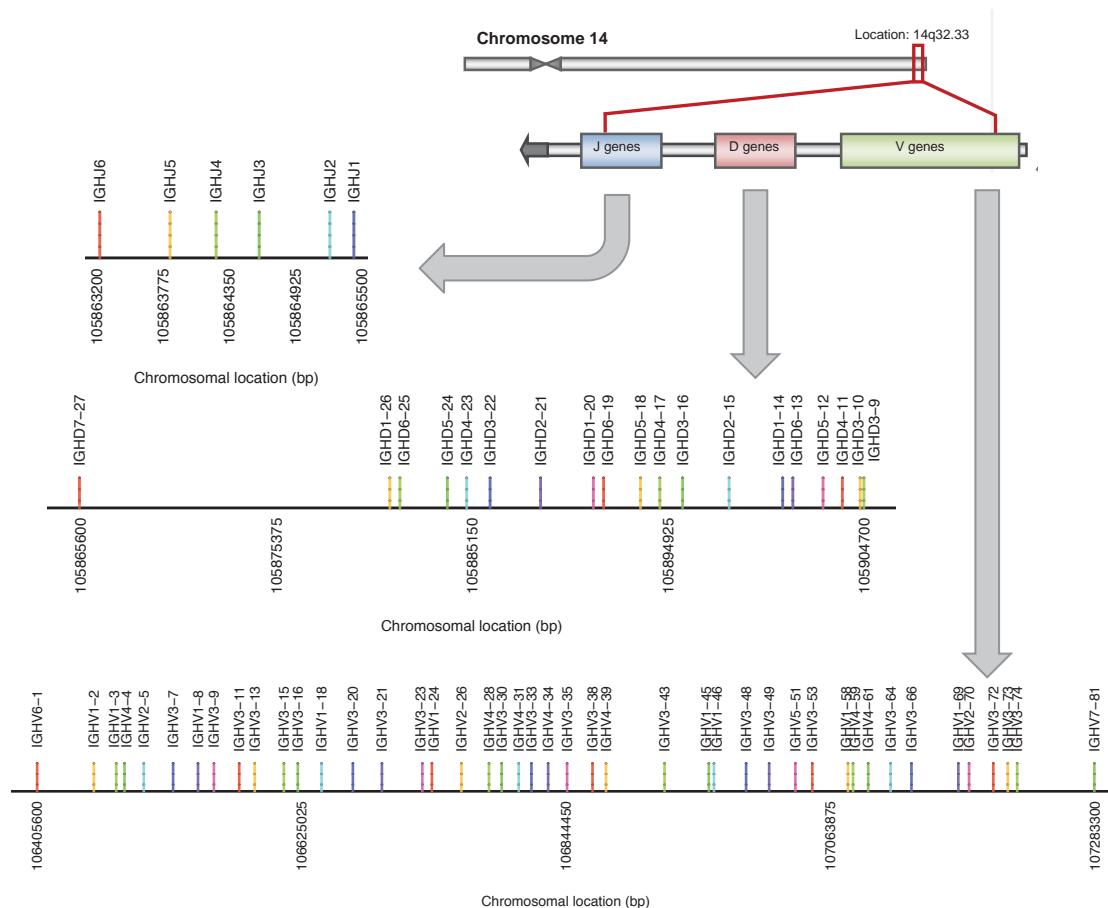
Figure 1.3. Arrangement of the human IgH gene locus on chromosome 14.

Schematic diagram of human chromosome 14, with the Ig locus marked in the red box. Within the locus are the multiple IgHJ, IgHD, and IgHV genes, where the chromosomal locations are marked to scale on the chromosomal sections at the top, middle and bottom respectively. Each coloured line along the chromosomal scale bar represents an IgH gene, with the name indicated below. Generated from gene coordinates from Ensembl.
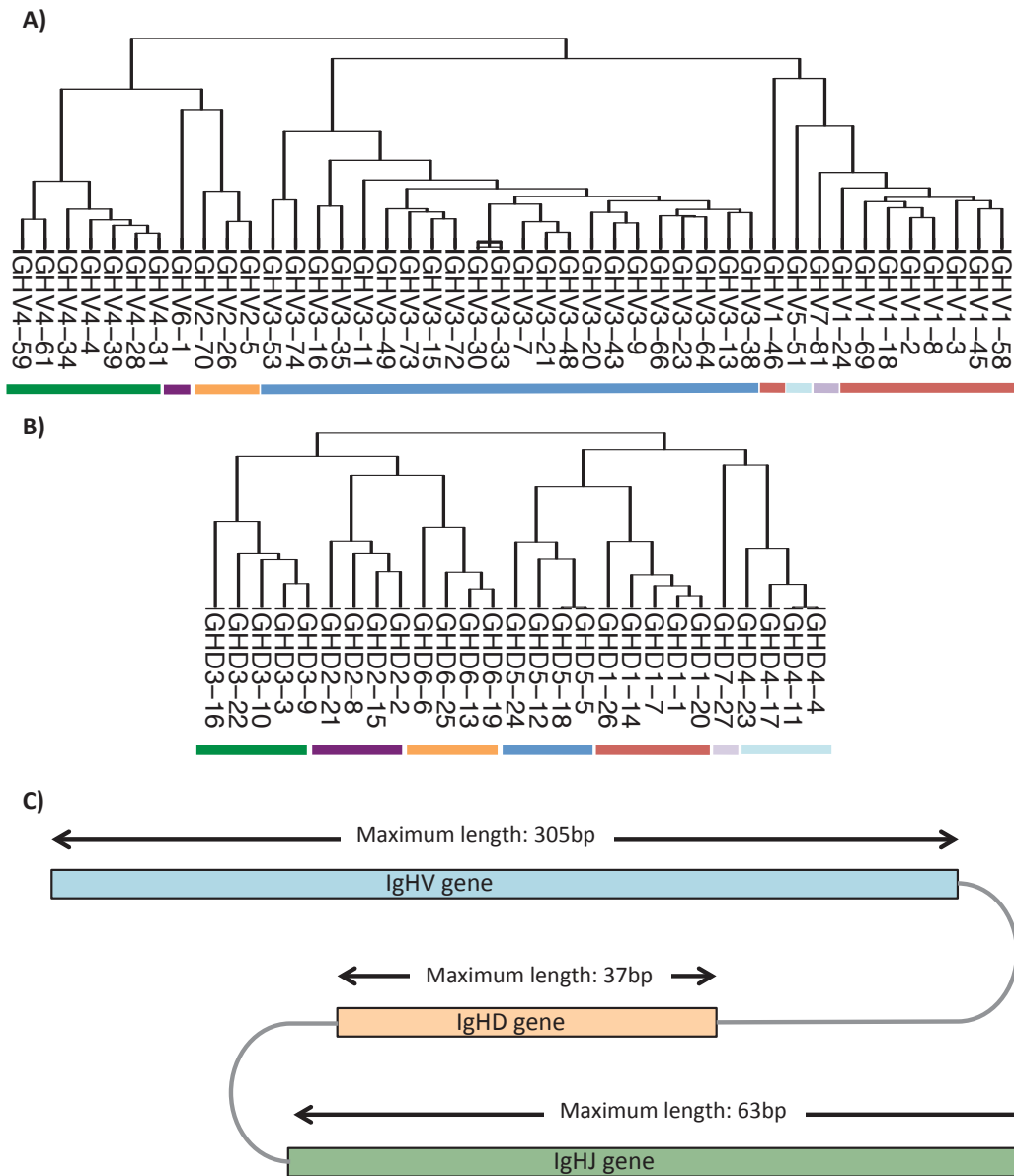
Figure 1.4. Phylogenetic sequence relationships between the human A) IgHV and B) IgHD genes.

Each IgHV or IgHD gene sequences was aligned using Mafft (Katoh and Standley, 2013) and a neighbour joining tree was fitted using BIONJ in R (Gascuel, 1997). The branch lengths represent the estimated evolutionary distance between gene sequences, and the tips are named by the corresponding gene names. Different coloured bars underneath each tree represent different IMGT gene groups (Lefranc et al., 2009) defined by (Giudicelli and Lefranc, 1999). As there are only 6 human IgHJ genes, phylogenetic sequence relationships of the IgHJ genes were not included. **C)** Schematic diagram of a functional BCR, with the maximum human IgHV, IgHD and IgHJ gene lengths indicated.
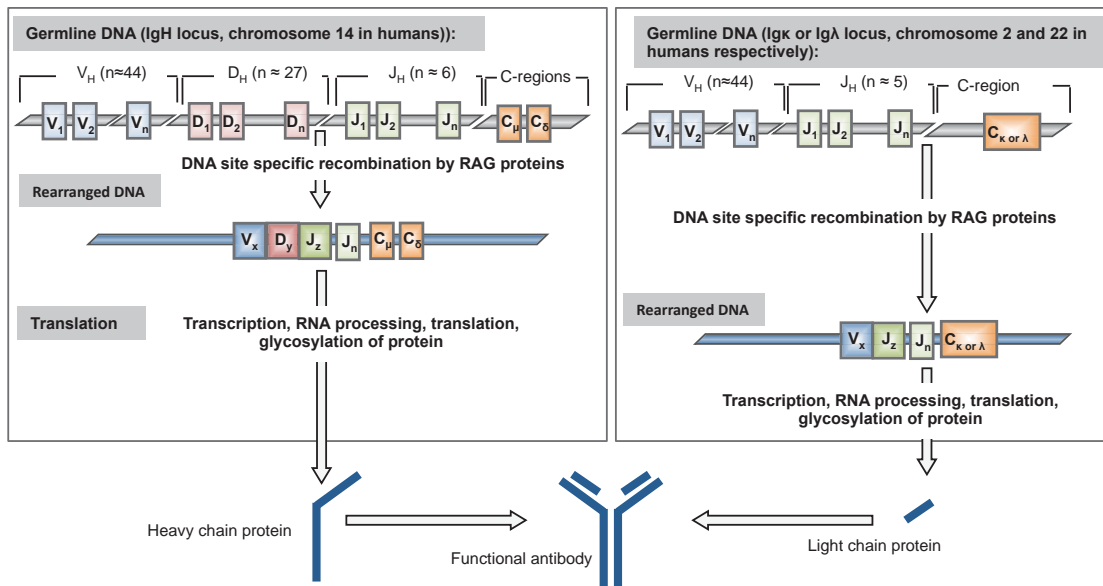
Figure 1.5. Stages of immunoglobulin gene rearrangement.

Gene rearrangements and assembly for the heavy and light to form a functional B-cell receptor. V_x, D_y, J_z denote the IgHV, IgHD, and IgHJ genes. C_μ and C_δ denote the μ and δ constant chains respectively, and $C_{\lambda\ or\ \kappa}$ denotes the λ or κ chains constant chain. Adapted from Jackson et al (Jackson et al., 2013).

1.1.4.2. B-cell receptor editing and allelic exclusion

If a maturing naïve B-cell has high affinity for self-antigens or does not form a functional BCR, the cells are removed by induced programmed cell death in the bone marrow, known as negative selection. B-cells committed to cell death can be rescued by modifying the V-J light chain recombination so that the B-cell receptor no longer recognises self-antigens or creates a functional reading frame (Dorner et al., 1998). This occurs by the process of receptor editing where renewed IgHV-D-J rearrangement can result in expression of a functional or non-auto-reactive BCR, rescued further by expression of a different IgL chain. Receptor editing is under genetic control, where PLCγ2 is thought to play a role in regulating the recombination-activating (*rag*) genes, and therefore receptor editing (Verkoczy et al., 2007, Benschop et al., 1999, Derudder et al., 2009). Each mature B-cell typically expresses a single heavy chain and light chain allele. The expression of productive functional heavy and light chains suppresses subsequent immunoglobulin gene rearrangements as well as expression of other rearranged alleles, a process known as allelic exclusion (Kitamura and Rajewsky, 1992).

1.1.5. B-cell response to antigens

Naïve B-cells require multiple signals to become activated: the first signal is delivered through the binding of the IgM B-cell receptor to an antigen (a protein, peptide, carbohydrate or other substance that the immune system perceives as being foreign or harmful). IgM cross-linking on the cell surface causing localised IgM clustering. This assembly provides intracellular signalling to the B-cell through communication of the BCR complex via the Igα and Igβ complex. For T-cell dependent antigens (detailed in Section 1.1.5.1), the second signal is delivered through T-helper cell recognition of peptide fragments of antigen bound to MHC class II molecules on the B-cell surface, and the interaction between CD40 ligand (CD40L) on the T-cell surface and CD40 on the B-cell surface. For T-cell independent antigens (detailed in Section 1.1.5.2), the second signal is by interactions between the antigen itself and B-cell surface, or by non-T-cell accessory cells. The third signal is given by the binding of Toll-like receptors (TLRs), that are up-regulated in naïve B-cells upon BCR activation, as well as other co-receptors, such as the CD19:CD21:CD81 protein complex. An example of this is the T-cell independent protein LPS, which binds LPS-binding protein and CD14, that subsequently associates with the receptor TLR-4 on the B-cell, leading to increased B-cell activation. A fourth signal can be delivered through cytokines (LeBien and Tedder, 2008).

1.1.5.1. T-cell dependent B-cell responses

Antigens can be classified as either T-cell dependent (T_{dep}) or T-cell independent (T_{indep}), depending on whether T-cell stimulation is required. The differences between T-cell dependent or independent immune responses are based on antigen size, structure, and nature. The majority of antigens are T_{dep} antigens that cannot induce B-cell proliferation without T-cell help, (i.e. activation signals from T-helper cells that respond to the same antigen). T_{dep} antigen responses lead to the generation of high-affinity class-switched B-cell responses (i.e. antibodies with heavy chains classes of IgM to IgG, IgA or IgE, detailed in Section 1.1.6). However, naïve T-cells require co-stimulatory signals from professional antigen presenting cells (APCs), such as dendritic cells, B-cells and macrophages. For example, dendritic cells, on encounter with a pathogen or antigen, endocytose and display the processed antigen fragments or peptides on their cell surface complexed with MHC proteins.

These cells carry the peptides to local lymph nodes or organs, and undergo maturation in order to be able to active T-cells. However, activation of T-cells requires three protein signals from the APC. Firstly the MHC molecules with bound antigen or antigen fragment must be able to bind the T-cell receptor. Secondly, co-stimulatory proteins of the APC must be able to bind complementary receptors on the T-cell surface, and thirdly, the action of cell adhesion molecules of the T-cells and APCs to enable contact between the T-cell and APC for long enough for the T-cell to become active. However, if the T-cell does not receive all three signals, it may be triggered to apoptose or the activation suppressed, a process known as immunological tolerance. T-helper cells also act to regulate the immune response by cytokine secretion (Korn et al., 2009)

During a T_{dep} response, a small proportion of activated B-cells differentiate into short-lived low-affinity plasma cells within the B-cell regions of the secondary lymphoid organs. Recruitment of the remaining activated B-cells to the B-cell follicular regions of the secondary lymphoid tissues lead to formation of germinal centres (GCs). GCs are micro-anatomical structures that support antigen specific B-cell clonal expansion, positive selection based on antigen affinity and BCR diversification by somatic hypermutation (SHM) (McHeyzer-Williams and McHeyzer-Williams, 2005). SHM is a process that introduces point mutations and, occasionally, insertions and deletions into the variable regions of the heavy chain immunoglobulin, where some of the resulting populations are expanded through positive selection for higher affinity antigen binding (Gojobori and Nei, 1986). These lead to some B-cells improving their antigen specificity and affinity to the antigen, often by several orders of magnitude (Griffiths et al., 1984, Eisen and Siskind, 1964). These hypermutations occur only in B-cells expressing cell type-specific activation-induced cytosine deaminase (AID) and actively transcribed Ig genes. AID is thought to act on both IgH and IgL strands of DNA by deaminating cytosines to uracils. The resulting uracils therefore base-pair with adenines during the next round of B-cell genome and cell division, leading to C to T, or G to A conversions. The additional process of uracil excision by uracil glycosylases and error prone repair of replication of abasic sites leads to transition and transversion mutations at C/G bases (Batrak et al., 2011). The immunoglobulin genes in B-cells are diversified by hypermutation at a significantly higher rate compared to non-immunoglobulin genes in B-cells, and have been found to occur at a significantly higher level within the complementary

determining regions (CDRs) compared to the framework regions (FWRs) in the BCR (Lin et al., 1997). The estimated average rate of somatic hypermutation is 1.51% per nucleotide site, or 3.09% per amino acid (Gojobori and Nei, 1986), which is high enough for the mutation rate to play a significant role in generating antibody diversity (Baltimore, 1981, Tonegawa, 1983). However, if the B-cell does not receive required activation signals after SHM, it may be triggered to apoptose or the activation suppressed as part of immunological tolerance to select only B-cells with the optimal antigen binding properties and to reduce the risk of generation of auto-reactive B-cells. High-affinity B-cells in the GC are positively selected on the basis of antigen-binding affinity. These cells rapidly proliferate to expand the size of the antigen-reactive B-cell pool with more than 2×10^5 cells in the dark zone and differentiate into either long-lived plasma cells or memory B-cells. Plasma cells typically migrate to the bone marrow and spleen, and secrete high-affinity antibodies for extended periods of time leading to clearance of antigen (Manz et al., 1997, Bernasconi et al., 2002).

1.1.5.2.   T-cell independent B-cell responses

T_{indep} antigens can induce B-cell responses directly. Two main types of T_{indep} antigens exist: type I (T_{indep}-I) polyclonal B-cell stimulants, typically soluble antigen, and type II (T_{indep}-II) large non-protein polymeric molecules with repeated epitopes, typically cell-bound antigen. The T_{indep}-II antigens are able to cross-link multiple B-cell receptors on naïve B-cells leading to activation and stimulation of antibody production in the absence of T-cell help. However, TLR stimulation or complement activation with CD21 stimulation is typically additionally required for maximal stimulation. The development of long-term memory B-cells activated against T_{indep} antigens is limited (reviewed in Section 1.1.7.2) (Mond et al., 1995, Adderson, 2001), but T_{indep} antigens, such as polysaccharides, can be modified to produce T-cell dependent B-cell responses via conjugation to protein carriers, resulting in the initiation of longer-lived antibody memory responses (Kelly et al., 2006, Pollard et al., 2009).

1.1.6. Class switch recombination

B-cell activation and isotype switching from IgM to IgG, IgA or IgE through recombination and deletion process is initiated by the encounter of antigens. This process is achieved through deletional recombination via the introduction of DNA

double-stranded breaks in two participating switch regions, rejoining of the broken regions to each other, and deletion of the intervening sequences containing the various C_H genes, in a process known as class switch recombination (CSR) (Chaudhuri and Alt, 2004). The immunoglobulin heavy chain constant region locus consists of an ordered array of Ig C_H region genes each flanked at the 5' side by switch (S) regions. These S regions are composed of tandem repeat units. CSR occurs through the initiation of AID by looping and deletion of the genomic DNA. This generates an extra-chromosomal DNA recombination product, known as the switch circle (SC) (Muramatsu et al., 2000, Manis et al., 2002, Okazaki et al., 2002). As the $C\mu$ gene is located in the most proximal to the IgHV-D-J exon, CSR between the $S\mu$ and another S region at the 5' side brings another C_H gene adjacent to the IgHV-D-J exon (**Figure 1.6**). The specificity of CSR is determined by the chromatin accessibility of the target regions (Muramatsu et al., 2000). An alternative mechanism of CSR has been shown to occur through inter-chromosomal exchange between the target S regions in stimulated B-cells, which would give non-circular chromosomal products (Dougier et al., 2006).
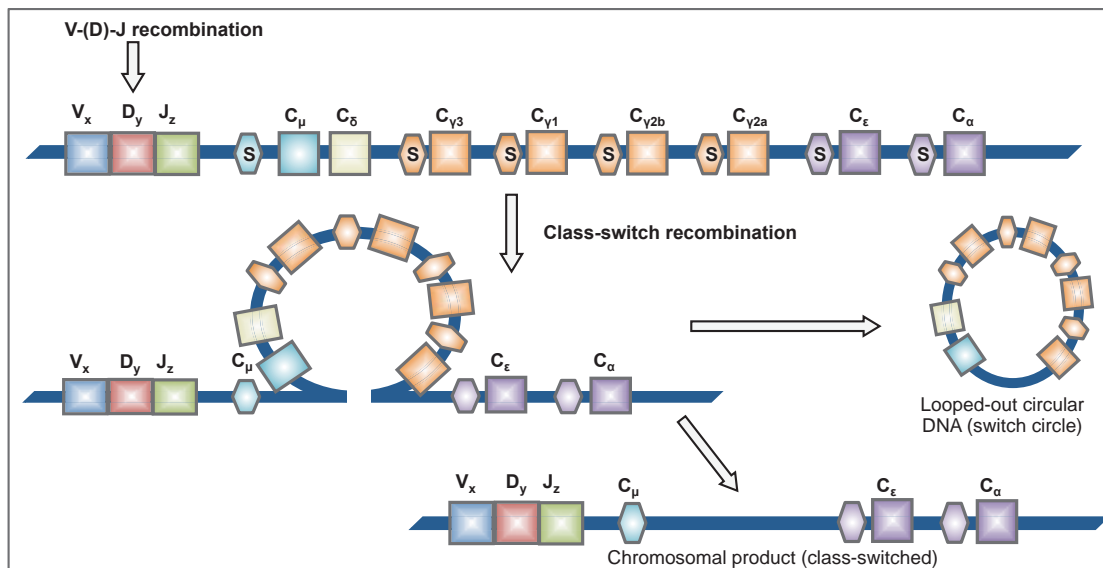
Figure 1.6. Mechanism of class-switch recombination.

The rectangles and hexagons represent exons and switch (S) regions respectively. V-(D)-J recombination occurs in the bone marrow, after which somatic hypermutation and class-switch recombination occurs in the peripheral lymph organs. Class-switch recombination involves the bringing together of heavy chain V-D-J exon with the target constant gene S region, and subsequent removal of intervening DNA. This results in a different chromosomal product and the looped out circular DNA. Adapted from Kinoshita *et al.* (Kinoshita and Honjo, 2001).

1.1.7. B-cell memory responses

The first exposure (the primary exposure) of a pathogen or antigen leads to the activation of naïve B- and T-cells. However, naïve B-cells require multiple signals to become activated (outlined in Section 1.1.5), but leads to the differentiation of antigen-specific antibody producing plasmablasts and memory B-cells, and differentiation of naïve T-cells to memory T-cells. These cells can persist for many years and maintained in a resting state in the absence of sustained antigen, thus immunological memory is established. Immunological memory allows the immune system to respond more rapidly to subsequent re-encounter to the same antigen. Resting memory B-cells are thought to have a low proliferation rate and the number of memory B-cells is highly regulated.

1.1.7.1. Generating T-cell dependent antigen immunological memory

Primary T_{dep} responses result in the interaction of antigen-stimulated B-cells with T-cells and other accessary cells (reviewed in Section 1.1.5.1), leading to the generation of short-lived plasma cells (PCs), GC-independent "early" memory B-cells and/or a GC reaction. The primary GC reactions persist following immunization for long periods (more than 8 months after initial antigen exposure) for certain types of antigen, as shown by monitoring memory B-cells over extended periods of time through the labelling of AID-expressing cells with yellow fluorescent protein (YFP) (Dogan et al., 2009). During this time in the GC, SHM and class-switch recombination can occur, resulting in the generation of high-affinity antigen-specific GC B-cells, that can differentiate into memory B-cells or long-lived PCs.

High-affinity antibody-producing long-lived PCs are thought to be integral to immunological memory, and reside in the in paracortical areas (immediately surrounding the medulla of the lymph nodes) to mature (Mohr et al., 2009). These cells migrate to the medullary regions, where CD93 expression is required for plasma cell survival in the bone marrow (Chevrier et al., 2009). Circulating antibodies from post-GC plasma cells therefore contribute to ongoing immune protection (Bernasconi et al., 2002). These plasma cells can also engage in antigen-specific immune regulation by negatively regulating the expression of BCL-6 and IL-21 in antigen-specific T_{FH} cells (Pelletier et al., 2010), and therefore modulating T-helper cell responses. It is thought that signalling through the BCR or MHC class II molecules in these post-GC plasma cells regulate the ongoing production of serum high-affinity

antibodies, but the mode of long-term antigen-presentation or regulation of post-GC plasma cells is not fully understood.

Humans and mice have been shown to generate memory B-cells expressing surface IgM (IgM$^+$ memory B-cells) as well as class-switched memory B-cells (expressing immunoglobulin isotypes other than IgM) (Weller et al., 2004, Tangye and Good, 2007). By tracking the murine memory B-cells during T$_{dep}$ responses against sheep red blood cells (Dogan et al., 2009) and phycoerythrin (PE, a fluorescent T$_{dep}$ antigen) (Pape et al., 2011), it was shown than IgM$^+$ memory B-cells persist longer than IgG1$^+$ memory B-cells. It has been shown in mice that, although the IgG memory B-cell population reduces in number over time, the number of IgM$^+$ memory B-cells remains constant after resolution of the primary response (Pape et al., 2011). In addition, IgM$^+$ memory B-cells have a slower turnover rate and typically contain lower levels of SHM than IgG$^+$ memory B-cells. Although IgM$^+$ memory B-cells are stimulated during subsequent antigen exposures, class-switched memory B-cells more rapidly differentiate into plasmablasts. IgM$^+$ memory B-cells can also be generated in T$_{indep}$ responses in the presence of different adjuvants (molecular components that magnify or modulate response to antigen).

1.1.7.2. Generating T-cell independent antigen immunological memory

T$_{indep}$-II antigen can generate long-lived PCs that secrete IgM or IgG antibodies from secondary lymphoid organs (Hsu et al., 2006) and the bone marrow (Taillardet et al., 2009, Foote et al., 2012). However, long-lived PCs generated from T$_{indep}$-II antigen have been shown to secrete lower antibody levels compared to their T$_{dep}$ counterparts (Taillardet et al., 2009). Although immunological memory can be employed against T$_{indep}$-II antigen, generated memory B-cells exhibit shorter longevity and different cell-surface phenotypes in T$_{indep}$ responses compared to that of the T$_{dep}$ response.

B-1 B-cells are a minor B-cell population and are able to contribute to the immune response against T$_{indep}$-II antigen. B1 B-cells express IgM and are thought to self-renew unlike "conventional" B2 B-cells. It has been shown in mice that B-1 B-cell populations expressing BCRs consisting of IgHV12 and IgHV11 in combination with IgHJ1, with no or low levels of SHM, is thought to be responsible for major natural antibody response against phosphatidylcholine (PtC), a ubiquitously expressed

membrane phospholipid found in both bacteria and mammalian cells (Mercolino et al., 1988, Yoshikawa et al., 2009, Arnold et al., 1994, Popi et al., 2009, Rowley et al., 2007). It is thought that anti-PtC antibodies comprise up to 15% of B-1 cells in the peritoneal cavity in most mouse strains (Arnold and Haughton, 1992), and the majority of B-1 B-cells are generated during foetal or neonatal development, and undergo self-renewal throughout life, thus considered a germline "memorised" B-cell response (Hardy and Hayakawa, 2001, Berland and Wortis, 2002). B1 B-cells can be divided into two subtypes, where B1-a B-cells make up the majority of the B-1 B-cells population and express CD5, and B1-b B-cells are the minor B-1 B-cell population that do not express CD5. It has been shown that, a population of IgM$^+$ mouse B1-b B-cells persists that confers protection against *Borrelia hermsii* infection on transfer to antigen naïve mice, suggesting that these are memory B-cells (Alugupalli et al., 2004). In addition, after immunisation with (4-hydroxy-3-nitrophenyl)-acetyl (NP)-Ficoll, a T_{indep}-II antigen, mouse IgG$^+$ and IgG$^-$ B1-b B-cells were shown to persist for more than 4 months, where these cells rapidly divided on adoptively transfer into antigen naïve mice (Obukhanych and Nussenzweig, 2006). However, it is unclear whether the precursors to these anti-NP-Ficoll B-cells were of B1 B-cell or B2 B-cell origin, whether aborted GCs could have been generated (de Vinuesa et al., 2000), and the human counterpart to the B1 B-cell population.

1.1.7.3. Immunological memory recall

The positioning of memory B-cells in the antigen-draining sites of secondary lymphoid tissues, such as the tonsil mucosal epithelium and splenic marginal zone, and the enhanced expression of co-stimulatory molecules assists rapid presentation of antigen to specific T-cells, thus promoting strong secondary adaptive immune responses (summarised in **Figure 1.7**) (Liu et al., 1988, Tangye et al., 1998, Liu et al., 1995). Indeed, enhanced reactivity of memory B-cells over naïve B-cells is thought to be, in part, conveyed by the cytoplasmic domain of surface IgG, thus contribute to rapid secondary immune responses (Martin and Goodnow, 2002). Upon rechallenge with the same antigen, antigen-specific memory B-cells can return to the GCs, undergo further clonal expansion, and differentiate into effector cells, such as plasma cells that secrete high-affinity antibodies.
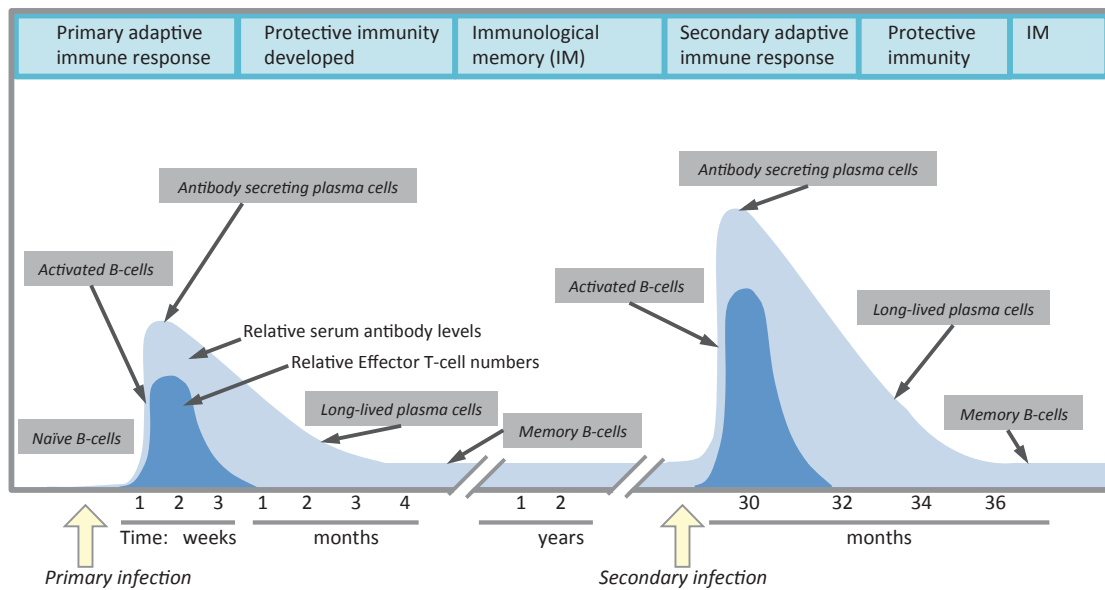
Figure 1.7. Features of primary and secondary response.

Primary exposure to antigen activates naïve B-cells with antigen specificity, resulting in clonal expansion of antigen-specific long-lived memory B-cells, plasma cells and memory T-cells. Plasma cells generate large amounts of serum antibody against the antigen, thus providing protective immunity. The majority of plasma cells in the primary response live for up to a few months, thus the serum antibody levels decline. Subsequent exposure to this antigen leads to the reactivation and proliferation of antigen-specific long-lived memory B- and T-cells and differentiation into effector cells. As memory B- and T-cells are sensitive to activation, and also may have already undergone class switch recombination and affinity maturation, secondary responses typically occur more rapidly and to a larger degree than primary exposure, and producing more effective antigen-specific antibodies. Adapted from McHeyzer-Williams *et al.* (McHeyzer-Williams et al., 2012).

1.2. Measuring B-cell population structure

Healthy humans have approximately 3×10^9 B-cells in the peripheral blood, where the population of B-cells in an individual reflects both current B-cell responses and historical immune encounters from memory B-cell and plasma cell populations. As a B-cell clone expresses a unique BCR, the B-cell population structure can effectively be probed by analysing the repertoire of BCR sequences in a given B-cell sample, for example from peripheral blood or bone marrow sample. This section details the main advances in understanding B-cell population structures and dynamics in health and disease by B-cell BCR repertoire sequencing.

1.2.1. Low-throughput B-cell receptor analyses

Low-throughput analysis of the heavy and light chains in the 1990s has illuminated biological mechanisms involved in the generation of specific immune responses. The functional characterisation of antibodies was made possible by the cloning of immunoglobulin genes from single B-cells and the isolation of specific antibodies. An alternative route for expression of antibodies was made possible through B-cell immortalisation (Tiller et al., 2007, Corti et al., 2011, Corti and Lanzavecchia, 2013). These methods have led to the isolation of neutralising antibodies to a range of pathogens. A summary of six vaccine studies based on low-resolution B-cell repertoire characterisation is given in Table 1.2.

Table 1.2. Summary of vaccine studies based on low-resolution B-cell repertoire characterisation.

Adapted from (Galson et al., 2014)

Vaccine*	Cells used	Methodology	Key findings	References
Influenza				
TIV	IgG plasmablasts 7 days after vaccination	Single cell heavy and light chain PCR followed by 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.	(Wrammert et al., 2008)
TIV	IgA and IgM plasmablasts 7 days after vaccination	Single cell heavy and light chain PCR followed by 454 sequencing	384–768 sequences analysed from three individuals. eight mAbs from large clonal sequence families, and 12 mAbs from singleton sequences were cloned, where 75% of these mAbs bound and neutralized influenza. Most effective binding was from sequences from large clonal families, three of which bound more effectively to the HA from the previous influenza season than the vaccine strain.	(Tan et al., 2014)
Tetanus				
TT	Plasmablasts 6 days after three consecutive vaccinations, separated by at least 1.5 years	Single cell heavy and light chain linkage PCR, cloning into *Escherichia coli* and Sanger sequencing of TT-positive clones	The level of SHM in the BCR sequences were similar between individuals, and did not increase through the study, suggesting the limit had already been reached through previous routine vaccinations.	(Poulsen et al., 2011)
TT	TT-specific plasmablasts 7 days, and TT-specific memory B cells 9 days after vaccination	Single cell isotype-specific heavy and light chain PCR followed by Sanger sequencing	CDR3 length, IgHV-D-J gene usage, and distribution of SHMs were similar among TT-specific plasmablasts and memory cells.	(Frolich et al., 2010)
S. pneumoniae				
PS (23 valent)	IgG plasmablasts 7 days after vaccination	Single cell heavy and light chain PCR followed by Sanger sequencing	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.	(Smith et al., 2013)
PS (23 valent)	PPS4 or PPS14 specific B cells 6 weeks after vaccination	Single cell culture followed by IgH PCR and Sanger sequencing of pooled, cultured cells	More than 1300 sequences from 40 individuals were analysed, showing significant differences in antibody repertoires between young and elderly individuals, where the latter had significantly more clonal with lower levels of SHM.	(Kolibab et al., 2005)
Hib				
PS or PS-DT or OC-CRM	Lymphocytes 7 days after vaccination	Fusion of lymphocytes to mouse myeloma cells followed by culture, heavy and light chain PCR and Sanger sequencing	15 cell lines that secreted antibody against Hib PS were sequenced from 10 individuals, where it was found that these mAbs had undergone SHM and demonstrated increased B-cell clonality after vaccination and bias towards use of the IgHV3 gene family.	(Adderson et al., 1993)
PS-DT	Lymphocytes 7 days after vaccination	Fusion of lymphocytes to mouse myeloma cells followed by culture, heavy and light chain PCR and Sanger sequencing	4 cell lines that secreted antibody against Hib PS from four individuals were sequenced, where all used IgHV3 genes, but 2 unique IgHD-IgHJ and IgHKV gene segments, indicating that the four cell lines were from two different lineages.	(Pinchuk et al., 1995)
PS or PS-DT	Lymphocytes 7 days after vaccination	Fusion of lymphocytes to mouse myeloma cells followed by culture, IgH PCR and Sanger sequencing	5 cell lines that secreted antibody against Hib PS from four individuals were sequenced, where all used IgHV3 genes, but unique IgHD-IgHJ gene combinations.	(Adderson et al., 1991)

*Abbreviations: DT = diphtheria toxoid; Hib = *Haemophilus influenzae* type b; OC = oligosaccharide; PPV23 = 23-valent pneumococcal polysaccharide vaccine; PS = polysaccharide; TIV = trivalent inactivated influenza vaccine.

1.2.2. High-throughput B-cell receptor analyses

As healthy peripheral blood contains 0.07-0.53 x10^6 B-cells per ml, high-throughput sequencing, currently able to produce >10^7 sequencing reads per run covering the variable BCR gene sequence, is well suited for sampling this BCR repertoire (Dimitrov, 2010, Benichou et al., 2012). To ensure that the maximum number of sequencing reads correspond to fully rearranged BCRs and reduce the number of non-specific sequencing reads, B-cell DNA samples require PCR amplification, and B-cell RNA samples require both cDNA synthesis and PCR amplification. The three main BCR amplification methods for sequencing BCR repertoires are PCR using IgH-specific multiplex primers (van Dongen et al., 2003), 5' rapid amplification of cDNA ends (5'RACE) (Freeman et al., 2009, Bertioli, 1997, Warren et al., 2011, Varadarajan et al., 2011) and RNA-capture using RNA bait/capture probes (Choi et al., 2009, Mamanova et al., 2010) (summarised in **Figure 1.8**).

Three sets of human IgH-specific multiplex PCR primers have been designed (van Dongen et al., 2003), and validated (van Krieken et al., 2007, Evans et al., 2007, Vargas et al., 2008, Lukowsky et al., 2006, Bruggemann et al., 2007), known as BIOMED-2 FR1, FR2 and FR3 primer sets. These primer sets use a single IgHJ specific primer that can potentially bind any IgHJ gene, and 6-7 IgHV primers that can potentially bind any of the reference IgHV genes. The annealing sites of the BIOMED-2 FR1, FR2 and FR3 IgHV primers are in the highly conserved FR1, FR2 and FR3 regions of the IgHV genes respectively (**Figure 1.9**). PCR amplification using the BIOMED-2 FR1 primer set gives the longest PCR product, therefore is the most popularly used primer set for biological studies (Campbell et al., 2008, Boyd et al., 2009, Sanchez et al., 2003, Maletzki et al., 2012, Boyd et al., 2010a, Lev et al., 2012, Jager et al., 2012, Krause et al., 2011, Bashford-Rogers et al., 2013). Similar primers have also been designed to amplify the B-cell light chains (van Dongen et al., 2003). This multiplex PCR method can be performed on either RNA or DNA and sensitive enough to amplify BCRs from even single cells.

RNA-capture is based around the methods used for human exome sequencing that uses RNA bait/capture probes and subsequent universal PCR amplification (Choi et al., 2009, Mamanova et al., 2010). Briefly, the cDNA is generated from RNA, typically using primers that allow for sample indexing and sequencer specific

adaptors. The resulting sampling is hybridised with 120bp biotinylated RNA-capture baits, designed to bind to any IgHV or IgH constant region (as well as, potentially, other regions, such as the light chains and T-cell receptors, **Figure 1.8**B). The hybridised sequences are specifically bound to magnetic streptavidin beads, after which the sequences are universally amplified and sequenced. This allows for enrichment, amplification and sequencing of TCRs (α, β, γ and δ chains) and BCRs (heavy and light chains) simultaneously. PCR and RNA-capture methods can use RNA or DNA, but have the potential for sequence-based differential annealing and biased capture.

5'RACE uses a single Ig-specific primers, either to the heavy or light chain J genes or constant regions, for first strand Ig cDNA synthesis and subsequent sequence-independent template switching primer for second strand cDNA synthesis (**Figure 1.8**C). This eliminates potential multiplex primer bias, but can have low efficiency, high non-specific amplification, and short fragment contamination from RNA degradation or incomplete cDNA synthesis and template switching (Freeman et al., 2009, Bertioli, 1997, Warren et al., 2011, Varadarajan et al., 2011). Also, as the RNA bait probes and multiplex PCR primers are generated from reference Ig and TCR gene databases, they may lack the same efficiency as 5'RACE for capturing certain human allelic variants of TCR or BCR segments that are not represented in the reference database.
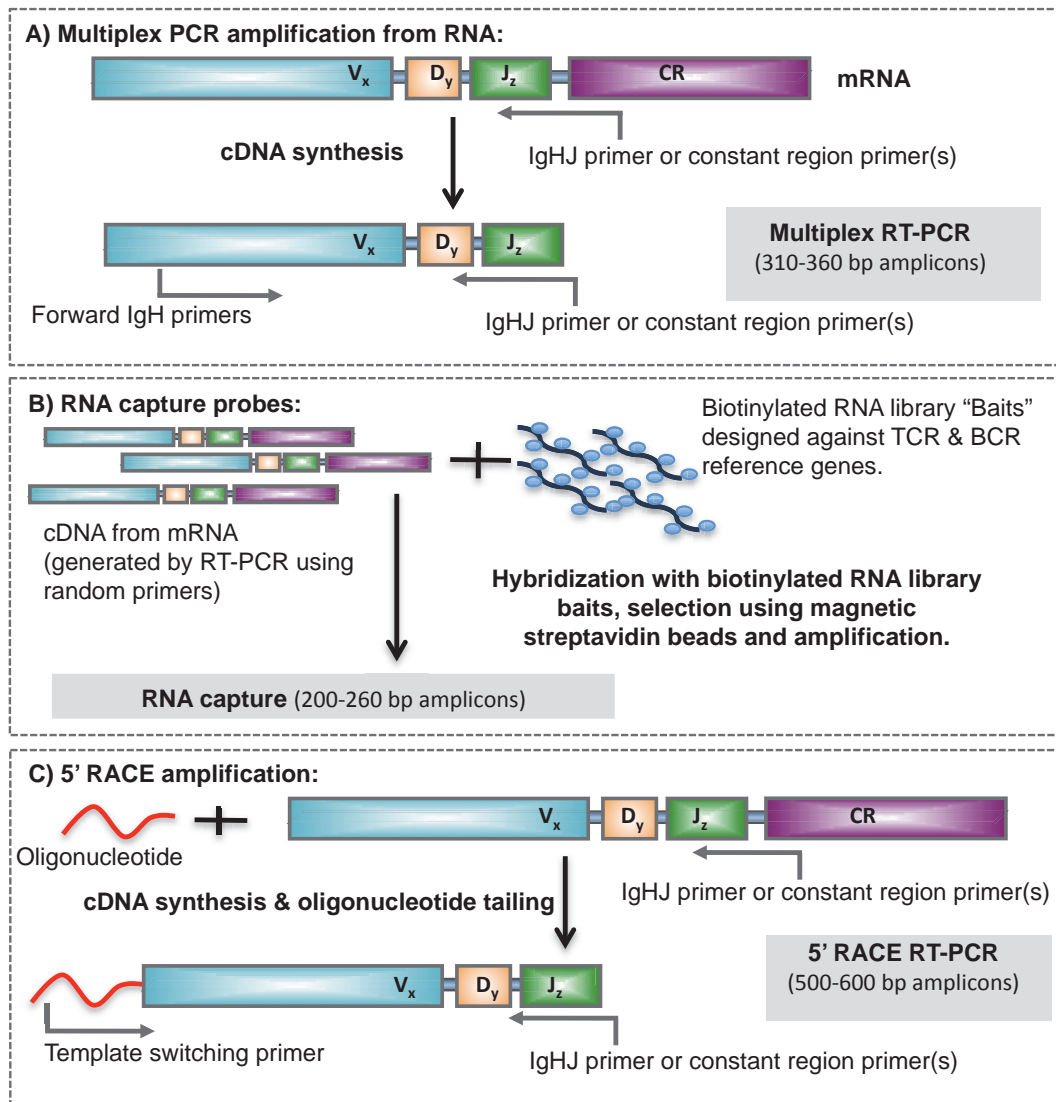
Figure 1.8. Different IgH RNA sequencing methods.

A schematic diagram of the different BCR amplification methods: RNA was extracted from peripheral blood samples and multiplex PCR, 5'RACE and RNA-capture. **A)** Multiplex RT-PCR of RNA uses degenerate primers located in conserved regions of the IgHV and IgHJ or constant region primer, where lengths of PCR products are based on the use of BIOMED-2 FR1 primer set. **B)** RNA-capture uses RNA bait probes and subsequent PCR amplification. **C)** 5' Rapid amplification of cDNA ends (5'RACE) of RNA uses a single IgHJ or constant region primer and a template switching primer.
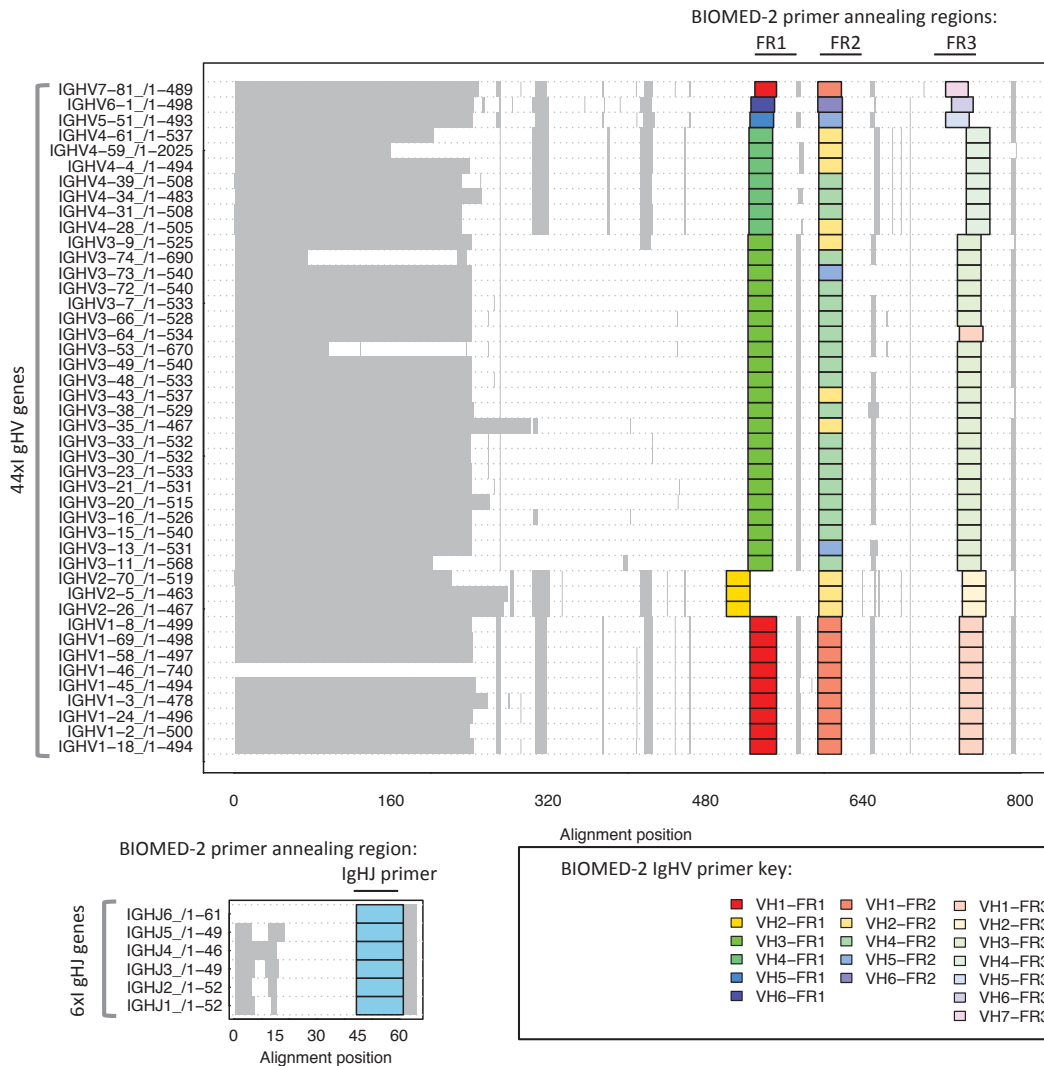
Figure 1.9. Alignment of human IgHV and J genes with BIOMED-2 primer annealing locations.

Each row of the figure represents the alignment of the corresponding IgHV or IgHJ gene sequence by CLustalW, with the gene name given on the left of the alignment box. Grey represents gaps in the alignment, and the coloured boxes represent the best annealing locations of the BIOMED-2 FR1, FR2 and FR3 primers. For the IgHV region, the colours of the primer annealing locations represent the primer that is most likely to anneal (with a maximum of 3 base-pair mismatches between the gene sequence and the primer sequence). For the IgHJ region, the light-blue boxes represent the primer annealing location of the BIOMED-2 IgHJ primer.

The limitation of sequencing of paired heavy and light chains from bulk cells in independent reactions is that information of the heavy and light pairings are lost. A solution to this is high-throughput single cell linkage PCR. This method can currently sequence more than 50,000 cells in a single experiment by depositing single cells in a high-density microwell plate and *in situ* lysis (DeKosky et al., 2013). mRNA is then captured in magnetic beads, on which RT-PCR is performed by emulsion VH:VL linkage PCR. These methods can be used to characterise antibodies of interest by generation of recombinant antibody by cloning the paired heavy and light chains into expression vectors, such as antibody variants of an isolated anti-HIV broadly neutralising antibody (Zhu et al., 2013a).

1.2.3. B-cell receptor repertoires

1.2.3.1. B-cell repertoires in model species

One of the first studies of the B-cell repertoire was performed on zebrafish (Weinstein et al., 2009). Zebrafish are ideal organisms for trialling high-throughput sequencing methods as they have recognisable adaptive immune system similar to humans that undergo IgHV-D-J recombination to generation functional BCRs with junctional diversity and the potential for somatic hypermutation. However, the zebrafish immune system contains only about 300,000 B-cells that produce antibodies, thus can be exhaustively sampled in a single sequencing run. Weinstein *et al.* found that not all possible IgHV-D-J combinations were used per zebrafish, but IgHV-D-J frequency distributions were highly correlated between individual zebrafish. A summary of studies of B-cell repertoires in model species is given in Table 1.3.

Table 1.3. Summary of studies of B-cell repertoires in model species.

Vaccine/disease	Cells used	Methodology	Key findings	References
None	14 zebrafish	Multiplex PCR of IgH and 454 sequencing	Only between 50-86% of all possible IgHV-D-J combinations were used per zebrafish, where IgHV-D-J frequency distributions were highly correlated between individuals zebrafish. Zebrafish exhibited unique BCRs that were shared between different individuals.	(Weinstein et al., 2009)
None	3 healthy homozygous isogenic Teleost fish and 4 Viral Hemorrhagic Septicemia Virus (VHSV)-infected Teleost fish	Cloning of IgHs into pCR2.1 vector and 454 sequencing	IgM, IgD and IgT repertoires were distinct (where the IgT class is specific to fish and encoded by the τ constant region gene). Clonal expansions dominated by a small number of large public and private clones were observed in infected fish. Differences were seen between the mucosal IgT and IgM repertoires, indicating that both IgM$^+$ and IgT$^+$ splenic B-cells responded to systemic infection but to different degrees.	(Castro et al., 2013)
None	PB from two crossbred calves (Brown Swiss × Red Angus-Simmental) and two purebred Holstein calves	Multiplex PCR of IgH and 454 sequencing	The bimodal length distribution of unique CDR3 sequences, with common lengths of 5-6bp and 21-25bp amino acids. 19 extremely long CDR3 sequences (up to 62 amino acids in length) within IgG transcripts were observed, a phenomenon that is rarely observed in other species. In addition, there was a high number of cysteine residues in the CDR3 regions compared to human BCRs.	(Larsen and Smith, 2012)
Finegoldia magna	4 adult C57BL/6 mice received injections of the V$_L$-targeting superantigen, protein L (PpL), a product of the common commensal bacterial species, *Finegoldia magna*	5' RACE of mouse light chains and 454 sequencing	Recurrent and consistent patterns of IgKV-J gene pairing was observed, where PpL exposure resulted in a significant shift in the repertoire in all exposed mice. Specifically, significant reductions in the frequencies of 14 specific IgKV gene combinations were observed. This suggests that microbial protein may modulate the composition of the B-cell repertoire, with potential implications for the relationship between the host-microbiome.	(Gronwall et al., 2012)
None	2 NOD-*scid-IL2Rγ*^{null} mice were engrafted with UCB CD34$^+$ hematopoietic progenitors and CD19$^+$ immature, naïve or total splenic B-cells were sampled at either 16 or 18 weeks of age	5' RACE of mouse heavy and light chains and 454 sequencing	Naïve B-cell repertoires in humanised mice are indistinguishable from those in human PB B-cells, with high correlations between heavy and light V-(D)-J gene usage frequencies, and similar CDR3 length distributions and charged amino-acid content, and hydropathy. However, the CDR3 region was found to be highly diversified in these mice, and specific for each individual.	(Ippolito et al., 2012)

1.2.3.2. Diversity of the immune repertoire

The potential number of different IgHV-D-J combinations in humans is 7128 (44 IgHVs x 27 IgHD x 6 IgHJs), and the number of light chain combinations is 200 and 124 for the Igκ chains and Igλ chains respectively (Table 1.4). However, with the junctional diversity between the gene recombination regions comprising of non-template additions and deletions greatly increases the number of potential unique BCRs that an individual can produce to over approximately 10^{14}. Further diversification can occur during the process of somatic hypermutation. However, as each B-cell can only encode for a single BCR, the true number of unique BCRs in an individual is bounded by the number of B-cells present. With the healthy peripheral blood B-cell population contains approximately 80% naïve B-cells and 20% memory B-cells (Tangye and Good, 2007), where each naïve B-cell is antigen inexperienced so each naïve B-cell BCR is considered to be unique (i.e. not clonally expanded). Sequencing BCRs from only naïve B-cells therefore theoretically results in a diverse BCR population with all BCRs represented with equal probability. In fact, the number of unique BCRs in two healthy individuals was estimated to be $3x10^9$ -$3x10^{10}$ by high-throughput sequencing of the CDR3 regions (Arnaout et al., 2011, Glanville et al., 2009a). However, little is currently known about B-cell turnover, dynamics or the different tissue distributions or B-cell repertoires (Dimitrov, 2010).

Table 1.4. Number of potential human BCR gene segment combinations.

Ig chain	Number of gene Segments	Number of potential combinations
IgκV	40	
IgκJ	5	200 Igκ chains
IgλV	31	
IgλJ	4	124 Igλ chains
IgHV	44	
IgHD	27	
IgHJ	6	7,128 IgH chains
Total number of potential combinations:		$2.452x10^6$

High-throughput BCR sequencing of different B-cell subsets can distinguish between human transitional, naive repertoires, switched memory B-cell and IgM memory populations (Wu et al., 2010). Previous studies have qualitatively shown diverse IgH repertoires in healthy patients contrasting with clonal populations in

malignancies (Boyd et al., 2009, Campbell et al., 2008, Logan et al., 2011, Sanchez et al., 2003, Maletzki et al., 2012, Carulli et al., 2011, Bashford-Rogers et al., 2013), and that distinct subsets of B-cells, defined by difference cell surface markers, within the same individual have distinct repertoires (Wu et al., 2010). Other significant studies of healthy BCR repertoires by high-throughput sequencing are summarised in Table 1.5.

Table 1.5. Summary of studies of B-cell repertoires from healthy individuals.

Cells used	Methodology	Key findings	References
Using IgD, CD27 and CD10 to sort cells into transitional, naive, switched memory, and IgM memory populations	Multiplex PCR of IgH and 454 sequencing	Memory B-cell repertoires differ from transitional and naïve repertoires, where the IgM memory repertoire is distinct from that of class-switched memory B-cells.	(Wu et al., 2010)
654 healthy human donors	Multiplex PCR of heavy and light chains and 454 sequencing	Although length and sequence variability in the CDR3 regions, substantial contributions to somatic diversity were found in the CDR1 and CDR2 regions. Estimation of the total number of unique BCRs from this healthy donor was estimated to be at least 3.5×10^{10}.	(Glanville et al., 2009b)
10 different human tissues using RNA samples derived from a large number of individuals	Multiplex PCR of IgH and 454 sequencing	Unique B-cell gene repertoires were observed in mucosal tissues, such as stomach, intestine and lung that differed significantly from those found in lymphoid tissues or PB. Mucosal tissue BCRs were distinct from peripheral blood and lymphoid tissue repertoires in their mutation level, frequency of both insertions and deletions, and CDR3 region lengths.	(Briney et al., 2014)
Isolated naive, IgM memory and IgG memory B cells from 4 healthy individuals	Multiplex PCR of IgH and 454 sequencing	IgHV-D-D-J recombinations were present in approximately 1 in 800 circulating B-cells, where this frequency was significantly reduced in memory B-cell subsets. These recombinations were shown to occur with all IgHD genes, suggesting that all recombination signal sequences that flank the IgHD genes are able to undergo IgHV-D-D-J recombination.	(Briney et al., 2012)

Historically, it was thought that BCRs and TCRs would not typically be shared between individuals due to the potential number of unique BCRs and TCRs compared to the limited number of B- and T-cells. However, it has been shown in multiple studies that certain BCRs and TCRs are shared significantly between individuals, known as public BCRs and TCRs respectively (exemplified in **Figure 1.10**), and thought to be a result of germline encoded BCRs (i.e. gene combinations with no somatic hypermutations (Li et al., 2012, Agathangelidis et al., 2012, Darzentas and Stamatopoulos, 2013, Messmer et al., 2004, Rossi and Gaidano, 2010, Warren et al., 2013, Hoi and Ippolito, 2013).
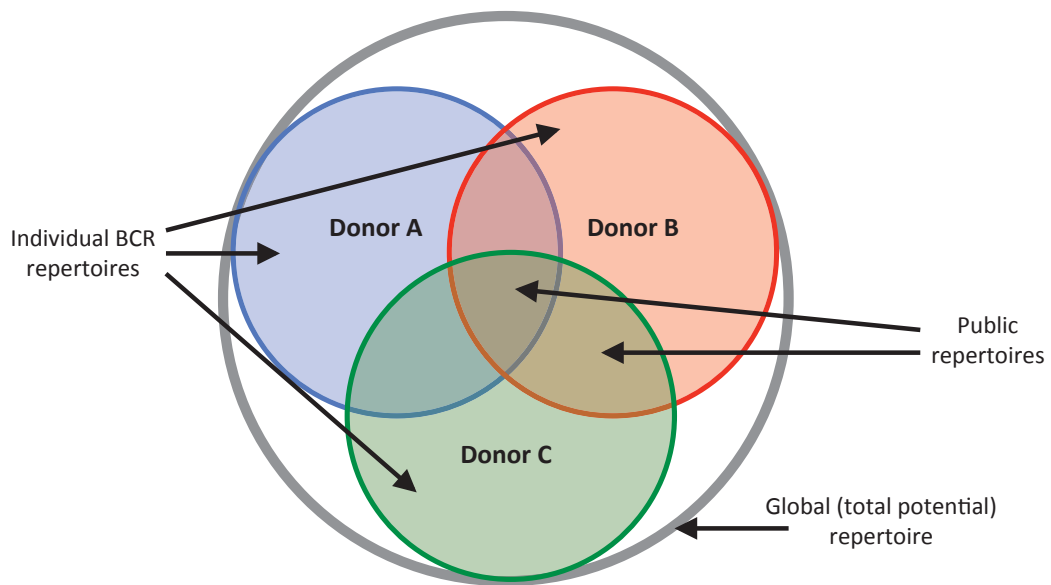
Figure 1.10. Schematic diagram of the different types of BCR repertoire.

Individual repertoires from each individual, where the BCRs shared between two or more donors are the public repertoires. The global repertoire is the set of all potential BCR sequences. Not drawn to scale, and adapted from (Finn and Crowe, 2013).

1.2.3.3. Immune repertoire variation with age

The adaptive immune system is not fully functional in human infants, and therefore infants can receive maternal antibodies (IgG) through their mother's milk. However, young infants are at increased risk of infectious diseases, such as influenza (Feeney et al., 2000). Although the Ig diversity is primarily thought to be a random process, evidence for deterministic, programmed repertoire development in foetal repertoires has been shown by overrepresentation of certain V segments in both mouse and humans (Perlmutter et al., 1985, Berman et al., 1991, Kalled and Brodeur, 1990). Preferential IgHV gene use in the adult B-cell repertoire is distinct from that of foetal, and young infant B-cell repertoires. For example, very young infant respiratory syncytial virus (RSV)-specific B-cells (<3 years of age, purified by immunoaffinity purification using RSV F protein) exhibited a biased repertoire with preferential by use of the IgHV1, IgHV3, and IgHV4 gene families, and less common use of the four immunodominant genes seen in the adult RSV F-specific B-cell response (IgHV3–23, IgHV3–30, IgHV3–33 and IgHV4–04) (Williams et al., 2009). The BCRs from children under three months of age possessed significantly fewer somatic mutations than those of adults, thus suggesting that younger children produce a different and potentially less optimised or weaker immune response than adults. The most frequently observed rearranged BCRs in healthy adult humans included IgHV4-59, IgHV4-61, IgHV3-23, and IgHV3-48 genes, where only 10 different IgHV genes account for more than half of all observed BCRs (Arnaout et al., 2011, Glanville et al., 2009b, Lloyd et al., 2009). Similarly, IgHD2-2, IgHD3-3, IgHD3-10, and IgHD3-22 were the highest observed IgHD genes and IgHJ4 and IgHJ6 the highest observed IgHJ genes in rearranged BCRs from healthy individuals, where this pattern of IgH gene recombination bias were shown to be consistent between multiple unrelated healthy adult individuals and in separate studies (Arnaout et al., 2011, Brezinschek et al., 1997).

The mechanisms so far implicating determinism versus stochasticity in the foetal repertoire are two-fold. Firstly, variation in recombination signal sequences which flank the V, D and J genes leads to favoured gene segments to be recognised by the recombinase (Feeney et al., 2000). Secondly, the observation that the expression of the terminal deoxyribonucleotidyl transferase enzyme is supressed in infants, therefore reducing the diversity generated through non-templated random nucleotide

insertions and deletions at IgHV-D and IgHD-J junctions) (Schroeder et al., 2001). Changes in B-cell repertoire structure have been associated with age in multiple studies, where increases in clonality and delays in immune response correlate with age and immunosenescence (summarised in Table 1.6).

Table 1.6. Summary of studies of immune repertoire variation with age.

Adapted from (Galson et al., 2014).

Vaccine*	Cells used	Methodology	Key findings	References
None	Zebrafish (ZF)	Multiplex PCR of IgH and 454 sequencing	At 2 weeks, ZF have highly preferential use of a small number of IgHV-D-J gene combinations, but this stereotypy decreases as the zebrafish mature by 1 month old. Evidence of complex diversification processes of antibody maturation observed due to clonal expansion during the affinity maturation process.	(Jiang et al., 2011)
None	14 healthy donors representing different gender and age groups	5' RACE of heavy and light chains and 454 sequencing	Donor B-cell repertoires separate into clusters of young adults and elderly donors (>50), thus suggests that clustering defines the onset of immune senescence at the age of fifty and beyond.	(Rubelt et al., 2012)
Influenza vaccine				
TIV or LAIV	Naïve B-cells, and plasmablasts on the day of vaccination and on days 7 or 8 and day 28 after vaccination from 17 participants from three age groups	IgH-specific multiplex PCR, and 454 sequencing	Higher clonality and SHM level was observed in the influenza-specific antibody repertoire in older individuals compared to younger individuals. In twins, the SHM level of the IgM repertoire was similar, but diverged for the IgG repertoire, indicating that the naïve repertoire is more influenced by individual genetics, but the memory repertoire is more influenced by environmental stimuli.	(Jiang et al., 2013)
Co-administered vaccines				
TIV and PPV23	PBMCs on the day of vaccination and on day 7 and day 28 after vaccination from 14 participants from two age groups	Semi-nested isotype and IgH-specific multiplex PCR, and 454 sequencing	At day 7 post-vaccination the repertoire changed, but returned to a baseline-like state after 28 days. Clonal expansion after vaccination is delayed in older individuals compared to young individuals, and age-related differences in IgA and IgM repertoire dynamics were observed.	(Wu et al., 2012)

*Abbreviations: DT=diphtheria toxoid; OC=oligosaccharide; PPV23=23-valent pneumococcal polysaccharide vaccine; PS=polysaccharide; TIV=trivalent inactivated influenza vaccine.

1.2.3.4. B-cell repertoire responses to vaccines and natural infections

Sequencing the B-cell immune response to infectious diseases and vaccines, such as human immunodeficiency virus and influenza, have been used to understand better the development of an antigen specific immune response and for identification of antigen-specific antibodies. Key studies and their findings are summarised in Table 1.7 and Table 1.8.

Table 1.7. Summary of studies of antigen-specific antibody repertoires.

Disease	Cells used	Methodology	Key findings	References
Human immunodeficiency virus (HIV)				
Natural HIV infection	Sera and PBMCs from a HIV-1 infected donor	5' RACE of IgH and 454 sequencing	Heavy and light-chain phylogenetic trees of identified anti-HIV 10E8 variants displayed similar architectures, where 10E8 variants reconstituted from matched and unmatched phylogenetic branches displayed significantly lower autoreactivity when matched.	(Zhu et al., 2013a)
Natural HIV infection	4 healthy controls, 4 subjects with SLE, and 4 HIV-1 infected therapy-naïve individuals and 4 HIV-1 infected individuals receiving combination antiretroviral therapy	Multiplex PCR of IgH and 454 sequencing	HIV or SLE subjects have increased clonality within their IgHM repertoire compared to healthy individuals. Antiretroviral therapy failed to reduce IgHM clonality in HIV-infected individuals, but IgHV-D-J gene combinations within the IgHM repertoire were found to be similar to known broadly neutralising HIV-1 antibodies.	(Yin et al., 2013)
Natural HIV infection	An acute HIV-1 infection donor was followed for more than 3 years starting from approximately 4 weeks after HIV-1 infection	Multiplex PCR of IgH and 454 sequencing	Study of the evolution and structure of a broadly neutralizing antibody from an African donor followed from the time of infection. This antibody, CH103, neutralized approximately 55% of HIV-1 isolates, which co-crystallised with the HIV-1 envelope protein gp120, thus revealing a new loop-based mechanism of CD4-binding-site recognition. This study showed virus and antibody co-evolution and maturation.	(Liao et al., 2013)
Plasmodium falciparum (Pf)				
Natural Pf infection	Classical memory B-cells and atypical memory B-cells from peripheral blood.	Single cell heavy and light chain PCR followed by Sanger sequencing	Natural Pf infection induces the development of memory B-cells from 67 healthy adults with neutralizing serum IgG activity against asexual blood stage parasites from a highly endemic area in Gabon that produce broadly neutralizing antibodies against blood stage Pf parasites, but only atypical memory B-cells, rather than classical memory B-cells, show signs of active antibody secretion	(Muellenbeck et al., 2013)

Table 1.8. Summary of studies of B-cell repertoires from vaccinations.

Vaccine*	Cells used	Methodology	Key findings	References
Influenza vaccine				
TIV	1 healthy volunteer vaccinated against seasonal influenza, hepatitis A, and hepatitis B, and 2 healthy volunteers vaccinated against seasonal flu vaccine, sampled at multiple time points before and after vaccinations	Multiplex PCR of IgH and 454 sequencing	IgHV and J gene usage differs between individuals, and is conserved within individuals over time. CDR3 clustering into clonal groups showed clonal expansion and contraction in response to the vaccine with different participants exhibiting different dynamics. A small number of highly mutated, persistent clones were found within all individuals, potentially corresponding to long-lived B-cell memory or indicative of chronic infection.	(Laserson et al., 2014)
TIV	Memory B-cells 14 days after vaccination from 1 participant	High-throughput single cell heavy and light chain linkage PCR and 2x250bp Illumina sequencing	The accuracy of heavy and light chain pairings identified using a high-throughput method was validated. Identification of 240 putatively influenza-specific heavy and light chain CDR3 pairings.	(DeKosky et al., 2013)
TIV or LAIV	PBMCs on the day of vaccination, and on day 7 and 28 after two vaccinations given a year apart from 28 individuals	IgH-specific RT and second strand synthesis, and PCR using barcoded priemrs. Custom 100x120bp Illumina sequencing.	Demonstrated different repertoire dynamics after TIV and LAIV vaccination. TIV induced a stronger response, with more abundant IgG lineages than LAIV. Shared antibody sequences on day 7 after two TIV vaccinations were found, where these lineages are present after the second vaccination potentially due to memory B-cell recall. Suggested that this method could be used to identify cross-reactive antibodies.	(Vollmers et al., 2013)
TIV	PBMCs on 18 time-points around two vaccinations given a year apart from 1 participant, and 10 time-points around one vaccination from 2 participants).	IgH-specific multiplex PCR, and 454 sequencing	IgHV and J gene usage differs between individuals, and is conserved within individuals over time. CDR3 clustering into clonal groups showed clonal expansion and contraction in response to the vaccine with different participants exhibiting different dynamics. A small number of highly mutated, persistent clones were found within all individuals, potentially corresponding to long-lived B-cell memory or indicative of chronic infection.	(Laserson et al., 2014)
Tetanus vaccine				
TT	Plasmablasts 7 days after vaccination (one participant)	High-throughput single cell heavy and light chain linkage PCR and 2x250bp Illumina sequencing	Identified 86 putatively TT-specific heavy and light chain pairings, 10 of which where cloned into HEK293K cells followed by competitive ELISA of the antibodies produced showed them to be TT-specific.	(DeKosky et al., 2013)
TT	Bulk plasmablasts, memory B-cells, and antigen-specific plasmablasts 7 days and 3 months after vaccination from 2 participants	Heavy and light chain-specific multiplex PCR, and 454 sequencing. High-throughput single cell of heavy and light chain linkage PCR from day	Analysed the serum antibody repertoire by using the IgH sequence database to interpret results from high-resolution liquid chromatography tandom mass spectrometry of the serum antibodies. This showed that ~5% of the plasmablast clonotypes identified by sequencing at day 7 could subsequently also be detected in the serological response 9 months after vaccination	(Lavinder et al., 2014)

*Abbreviations: LAIV=live attenuated influenza vaccine, TIV=trivalent influenza vaccine, TT=tetanus vaccine

1.2.3.5.   *In vivo* B-cell evolutionary processes

The immune system is capable of continually learning and memorising immunological experiences. The study of B-cell dynamics by B-cell cell receptor sequencing has been useful in the understanding of affinity maturation and selection of resulting mutants. Initially, using small sequence datasets per clone, lineage analysis became a popular analytical tool to understand mutational processes (**Figure 1.11**) (Steiman-Shimony et al., 2006a, Dunn-Walters et al., 2004, Barak et al., 2008, Frumkin et al., 2005, Uduman et al., 2014, McIntyre et al., 2014, Messmer et al., 2005b, Steiman-Shimony et al., 2006b, Bankoti et al., 2014, Sok et al., 2013, Green et al., 2013, Bergqvist et al., 2013, von Budingen et al., 2012, Seifert and Kuppers, 2009, Spencer et al., 2009). Lineage trees describe the clonal relationships between related cells within a lineage, where the root of the tree is assumed to be the germline sequence. Both the internal nodes as well as the leaves can represent sequences, as intermediate sequences can be included in the sample. Furthermore, lineage trees are not necessarily binary as a single B-cell can produce a population of identical cells that can produce mutations (Barak et al., 2008).
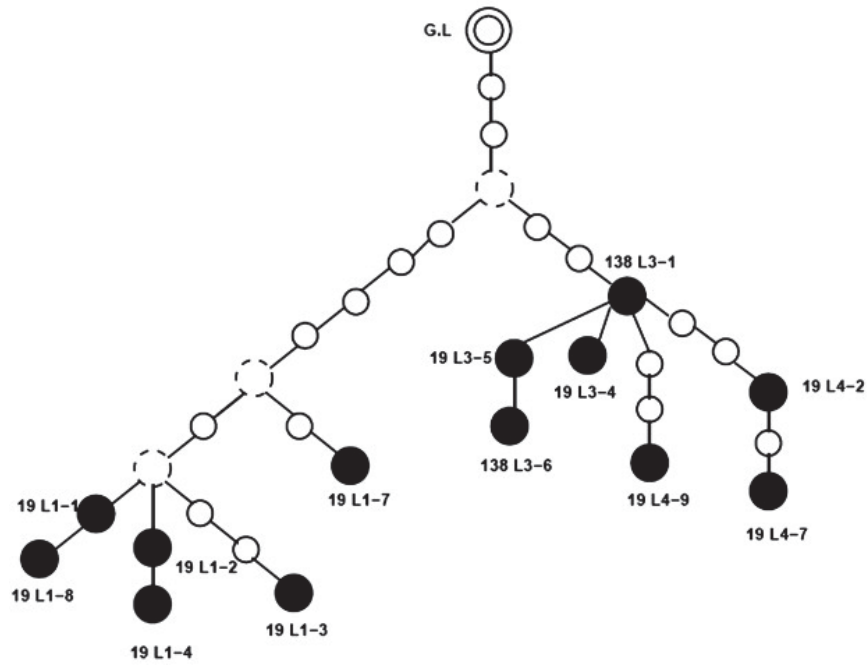
Figure 1.11. Lineage tree constructed by IgTree.

BCR sequences were from a patient with light chain amyloidosis (Manske et al., 2006). The trees are rooted with the nearest germline sequence (double circle). The filled circles represent the sampled sequences (named beside the nodes), dashed circles represent nodes that have more than one descendant, and solid white cycles represent nodes that have only a single descendant, inferred by IgTree©. Figure from (Barak et al., 2008).

Different phylogenetic methods are employed to analyse larger sequencing datasets, and using different hypotheses of evolution. Maximum likelihood methods estimate phylogenetic relationships by determining the theoretical likelihood of query sequences arising from a given ancestor by somatic hypermutation (Kepler et al., 2014, Kepler, 2013, Liao et al., 2013, Zhu et al., 2013b, Wu et al., 2011, Doria-Rose et al., 2014). Neighbour-joining methods use agglomerative clustering to generate trees representing sequencing relationships, and is typically faster for large datasets (Liao et al., 2013, Wu et al., 2011, Logan et al., 2011). Maximum parsimony methods assume that populations of cells, such as tumour cells or B-cell clones, develop in a parsimonious manner, such that the evolutionary process to create the population is minimal. Maximum parsimony assumes minimal number of explicit assumptions, thus useful when the true evolutionary process in B-cells is unknown or temporally variable (Campbell et al., 2008, Sutton et al., 2009, Rossi et al., 2012, Dagklis et al., 2012, Benichou et al., 2013). For example, Campbell *et al.* fitted unrooted parsimony models to generate phylogenetic trees for the malignant clones in 2 chronic lymphocytic leukaemia (CLL) samples to show the evolutionary relationships among the subclones and dominant clone of CLL cells (**Figure 1.12**) (Campbell et al., 2008). Although bootstrapping shows uncertainty in the ancestral relationships between individual subclones, there is strong support for 3 different classes of subclones: B-cells representing the intermediate stages between germline and the dominant clone, blind alleys representing divergent evolution away from the germline sequence, and ongoing evolution from the dominant clone. The persistence of B-cells from the intermediate stages suggest that initiating driver mutation(s) may have led to leukaemogenesis at the earliest branch-point of the tree.
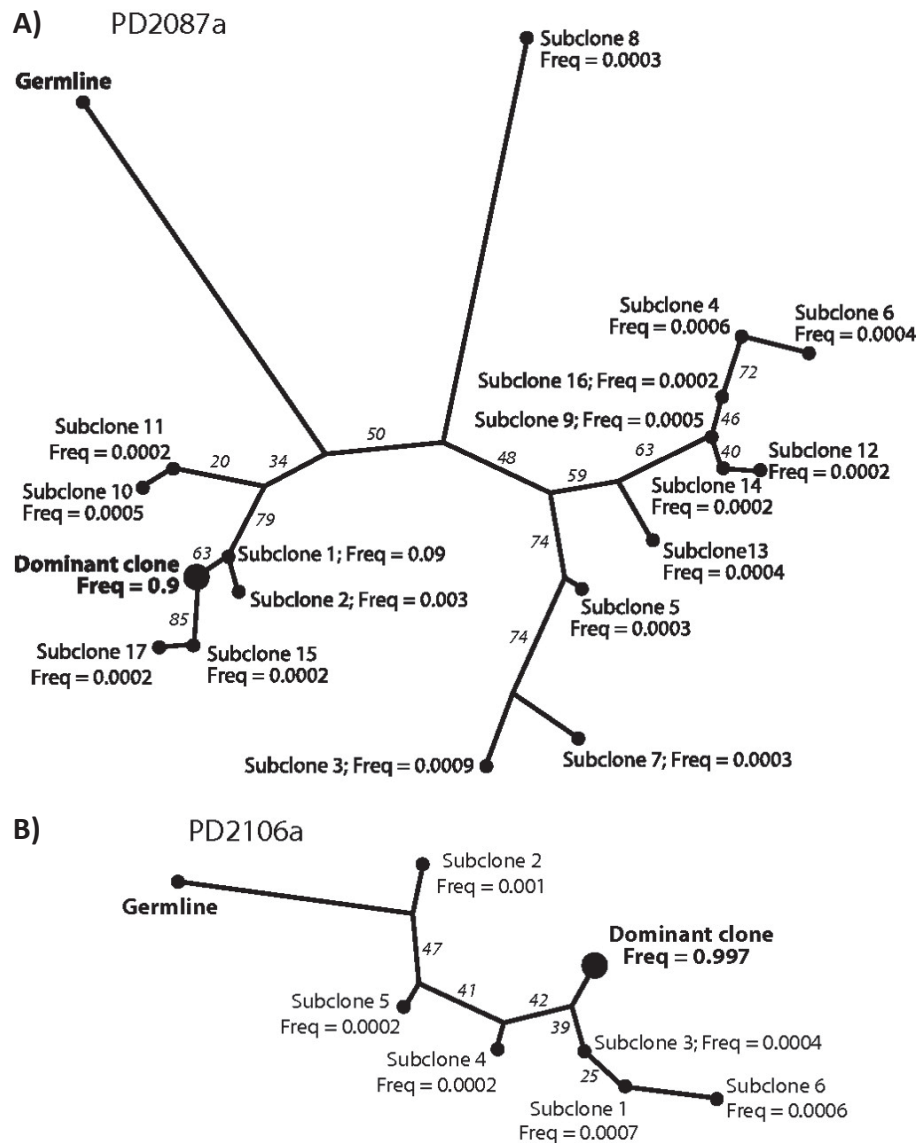
Figure 1.12. Maximum parsimony trees of B-clones.

The phylogenetic interrelationships between BCR clones for 2 chronic lymphocytic leukaemia patients, where the length of each branch proportional to the number of varying bases (evolutionary distance). The percentage support across 1000 bootstrap samples is given beside each intermediate branch. From (Campbell et al., 2008).

1.3. Chronic lymphocytic leukaemia (CLL)

1.3.1. Aetiology and epidemiology

Chronic lymphocytic leukaemia (CLL) is the most common form of leukaemia, representing 30% of all leukaemias. The incidence rate for CLL is 4.92 per 100,000 per year in Europe (Sant et al., 2010). The rates of CLL vary between populations; 35-40% of all leukaemia in Denmark is CLL, but only 3-5% in Chinese and Japanese populations (Redaelli et al., 2004). Incidence rates are higher for men than women, and increase with age, with two thirds of patients older than 60 years of age (Zenz et al., 2007). The clinical course of CLL is highly heterogeneous across individual patients (Morabito et al., 2011). Many CLL patients are asymptomatic, and remain treatment free for many decades, while an aggressive form of the disease can affect others. Patient conditions may deteriorate with the disease or may suffer from therapy related treatments (Morabito et al., 2011). Therefore, biological indicators of disease progression and prognosis are of great clinical importance. Identifying the risk factors associated with requirement early treatment or better prognosis estimation will decrease the treatments given to patients with the non-aggressive disease, with the majority of treatments carry significant toxicities.

1.3.2. Biology, pathogenesis and diagnosis of CLL

The diagnosis of CLL is made on two criteria. Firstly, if greater than 5×10^9 cells/L peripheral blood B-cells for at least 3 months, where clonality of circulating B-cells needs to be confirmed by flow-cytometry (Eichhorst et al., 2011, Hallek et al., 2008). CLL typically has preferential kappa or lambda immunoglobulin light chain usage at a ratio of greater than 3:1 or less that 1:0.3 respectively (Rozman and Montserrat, 1995, Cheson et al., 1996, Kilo and Dorfman, 1996). Secondly, leukaemia cells found in blood smears are small, mature B-cells with a narrow border of cytoplasm and dense nucleus with partially aggregated chromatin and lacking distinct nucleoli. CLL B-lymphocytes co-express CD19, CD5 and CD23, with weak or no expression of CD20, CD79b, FMC7 and surface immunoglobulin.

Monoclonal B-cell lymphocytosis (MBL) is thought to be a pre-clinical manifestation of CLL. The diagnostic criteria for MBL is exhibiting less than 5×10^9 cells/L peripheral blood B-cells for at least 3 months (Eichhorst et al., 2011, Hallek et al., 2008) along with either (a) kappa or lambda immunoglobulin light chain usage at

a ratio of greater than 3:1 or less that 1:0.3 respectively, (b) greater than 25% of B-cells expressing low-level or no surface immunoglobulin, or (c) a disease-specific immunophenotype, such as CD5+.

CLL manifests as an increasing collection of B-cells with related BCRs (malignant B-cell clone) that exhibit a wide range of phenotypic states, illustrated by the expression of different cell-surface proteins. Typical CLL is characterised by the accumulation of mature CD5+ B-cells in the blood, bone-marrow and secondary lymphoid organs (Chiorazzi et al., 2005). Unlike most tumour entities, only a small proportion of CLL cells proliferate, potentially acting as tumour stem cells (Messmer et al., 2005a), suggesting accumulation of CLL cells *in vivo* is not due to increased proliferation rates, but rather due to resistance to apoptosis (Chiorazzi et al., 2005). Evidence for CLL resistance to apoptosis includes both an anti-apoptotic expression profile, such as high expression of Bcl-2 protein (Inamdar and Bueso-Ramos, 2007, Mauro et al., 1999), and micro-environmental signals. Evidence for the latter is that CLL cells cultured without support *in vitro* rapidly undergo apoptosis, which can be prevented by co-culture with supporting stromal cells. Different types of stromal cells assist in survival of CLL cells *in vitro* and thought to be an integral part of the CLL microenvironment. These include monocyte-derived nurse-like cells (NLCs, a subset of large, round, adherent cells (CD14+ cells) that differentiate *in vitro* on co-culture with CLL or healthy B-cells) (Burger et al., 2000, Bhattacharya et al., 2011), CXCL12-expressing mesenchymal stromal cell (MSCs)(Burger et al., 2000, Eisele et al., 2009), or follicular dendritic cells (FDCs) (Pedersen et al., 2002). However, normal B-cells are not supported in this manner. It has recently been established that a signalling pathway for CLL B-cell survival and apoptotic resistance is activated by upregulation of protein kinase C (PKC)-βII expression on contact with stromal cells. Unregulated stromal PKC-βII in biopsies from patients with CLL, acute lymphoblastic leukaemia, and mantle cell lymphoma suggests that this pathway may commonly be activated in a variety of haematological malignancies (Lutzny et al., 2013).

The signs and symptoms of CLL gradually develop, therefore the onset of disease is difficult to identify. The disease is often discovered accidentally as a result of elevated lymphocyte counts during routine physician visits (Andritsos and Khoury, 2002). Asymptomatic CLL is seen in about 25% of patients, where the duration of the asymptomatic phase is highly variable (Inamdar and Bueso-Ramos, 2007). The early

signs of disease include persistent lymphocytosis, mild cervical, supraclavicular, and/or axillary nodes lymphadenopathy and splenomegaly. Thrombocytopenia and mild anaemia is seen in approximately 25% and 50% of patients respectively. Nodular and diffuse skin infiltrations, exfoliative dermatitis, erythroderma, and secondary skin infections are seen in about 5% of patients (Bonvalet et al., 1984, Cerroni et al., 1996). Disease progression can lead to organ infiltration, adenopathy with splenomegaly, hypersplenism, and peripheral cytopenias. These patients can present with weight loss, fever and night sweats. Advanced disease exhibits extensive bone marrow infiltration by neoplastic cells. Due to replacement of marrow by tumour cells, symptoms include severe anaemia, thrombocytopenia, and neutropenia (Inamdar and Bueso-Ramos, 2007).

CLL patients have increased frequency of abnormal immune manifestations, including immunodeficiency and autoimmunity despite the increased number of B-cells (Dearden, 2008). Approximately half of CLL patients have hypogammaglobulinemia (Hudson and Wilson, 1960). Bacterial infections are responsible for the majority of illnesses in patients with CLL, particularly infections of the respiratory tract, urinary tract, and skin, as well as viral infections. These infections contribute highly to patient morbidity and mortality. Many patients have poor responses to vaccination (Dearden, 2008, Shaw et al., 1960), where vaccine response is correlated with better CLL patient outcome and treatment-free survival (Dearden, 2008).

CLL is frequently associated with autoimmune conditions. Coombs' positive autoimmune haemolytic anaemia is seen in up to 25% of patients at some point during the course of the disease (Dameshek and Schwartz, 1959, Pisciotta and Hirschboeck, 1957). This condition involves the production of antibodies against red blood cells during or after developing CLL. Approximately 6% of patients develop red cell aplasia, and a subset of CLL patients develop auto-antibodies against platelets and neutrophils leading to thrombocytopenic purpura and neutropenia. Bence Jones paraproteinemia is seen in 65% of patients (Diehl and Ketchum, 1998).

1.3.3. Monoclonal B lymphocytosis as a possible pre-leukemic phase

Monoclonal B-cell lymphocytosis (MBL) is thought to be a pre-clinical manifestation of CLL, characterised by asymptomatic B-cell clonal expansions with surface phenotypes similar to that of CLL (Marti et al., 2007, Marti et al., 2005).

MBL has been detected in older adults in good health (Shim et al., 2014). The prevalence has been reported in a number of studies, ranging from <1% (Rachel et al., 2007, Shim et al., 2007) to 18%, depending on the detection methods and tested populations (Shim et al., 2010). MBL is more frequent in males, with prevalence significantly higher in individuals with relatives with CLL, and increases with age (Rawstron et al., 2002). However the incidence of MBL is approximately 100 times greater than that of CLL, and therefore cannot be taken to be a definitive sign of genuine neoplastic transformation (Ghia et al., 2000). Some CLL-like MBL clones can be present at much higher frequencies in the blood, with a 1-2% per year rate of progressing to symptomatic CLL (Rawstron et al., 2008, Shanafelt et al., 2009). The natural history of MBL is not well understood.

1.3.4. Disease staging in CLL

The Rai stage was first prognostic staging process to be developed for CLL, using a combination of lymphadenopathy (abnormal enlargement of lymph nodes), organimegaly (abnormal enlargement of organs), and cytopenias (anaemia and thrombocytopenia (platelet number reduction)) to determine five prognostic groups with median survivals given in Table 1.9 (Rai et al., 1975).

Table 1.9. Rai stage median survival.

Rai Stage	Risk level	Prognosis factors	Median survival
Stage 0	Low	Lymphocytosis	> 150 months
Stage 1	Intermediate	Lymphocytosis + Lymph node enlargement	101 months
Stage 2	Intermediate	Lymphocytosis + Spleen/liver enlargement	71 months
Stage 3	High	Lymphocytosis + anaemia	19 months
Stage 4	High	Lymphocytosis + thrombocytopenia	19 months

Adapted from Rai *et al.* (Rai et al., 1975).

The Binet staging system was also developed for CLL, which relied on the number of affected lymphal areas and cytopenias, summarized in Table 1.10 (Binet et al., 1977). The Rai and Binet staging systems provide prognosis for the patient as well as the appropriate time for patient therapy. However, there is significant heterogeneity of outcomes at the different stages, so new and more accurate prognostic markers in CLL are of great clinical interest.

Table 1.10. Binet stage median survival.

Binet Stage	Risk level	Prognosis factors	Median survival
Stage A	Low	Lymphocytosis + less than 3 enlarged lymphal areas	> 12 years
Stage B	Intermediate	Lymphocytosis + more than 3 enlarged lymphal areas	7 years
Stage C	Intermediate	Lymphocytosis + anaemia or thrombocytopenia	2 years

Adapted from Binet *et al.* (Binet et al., 1977).

1.3.5. Prognostic markers in CLL

Recurring genomic abnormalities with prognostic significance have been identified in genetic studies using interphase fluorescence *in situ* hybridisation (FISH) and chromosomal analysis in CLL (Oscier et al., 2002, Juliusson et al., 1990). Many reports have associated CLL prognosis with genomic aberrations, summarised in Table 1.11.

Table 1.11. Genomic markers in CLL associated with prognosis.

Marker type	Genomic/chromosomal markers	Relative prognosis	Reference
Deletions	Deletions in 11q, 17p	Poor	(Krober et al., 2006)
Deletions	Deletions in 13q	Good	(Krober et al., 2006)
Deletions	Deletion in 6q	Intermediate	(Cuneo et al., 2004)
Mutations	TP53, ATM (tumour suppressor genes)	Poor	(Zenz et al., 2010)
Mutations	IRF4, Bcl-2 polymorphism	Good	(Havelange et al., 2011)
Mutations	Bcl-6 mutation	Poor	(Sarsotti et al., 2004)
Mutations	MDM2 SNP	Poor	(Gryshchenko et al., 2008)
IgVH mutational status	IgVH mutated	Good	(Schroeder and Dighiero, 1994, Fais et al., 1998, Damle et al., 1999, Hamblin et al., 1999)
	IgVH unmutated	Poor	
Gene expression	ZAP-70 (correlates with mutational status)	Poor	(Krober et al., 2006)
	V3-21 gene usage	Poor	(Krober et al., 2006)
Micro RNAs	Micro RNA signature associated with prognosis	-	(Calin et al., 2005)
Telomere length	Longer telomere length (correlates with mutational status)	Good	(Grabowski et al., 2005)

There are four prognosis markers that are currently widely in clinical use:

1.3.5.1.   IgHV mutational status

Studies of CLL in some patients have shown that CLL cells do not possess any somatic hypermutations in the complementary determining regions (CDRs) of the immunoglobulin genes, whereas other patients have highly mutated BCRs (Cai et al., 1992). It has been suggested that the two different mutational statuses of CLL patients may be derived from two different stages of B-cell ontology, with the unmutated CLL cases corresponding to pre-antigenic stimulation, and the mutated cases corresponding to post-antigenic stimulation (Hamblin et al., 1999, Damle et al., 1999). The examination of IgHV genes in CLL patients have shown that the two subsets of CLL that are prognostically significant, with studies suggesting an inferior survival and high likelihood of requiring early treatment in patients with unmutated IgHV. For example, Hamblin *et al.* found that the median survival for stage A patients with mutated CLL was 293 months (n=46) compared to 95 months for unmutated CLL (n=38) (p-value=0.0008) (Hamblin et al., 1999). Furthermore, leukemic cells from unmutated patients tend to produce polyreactive antibodies (Martin et al., 1992, Herve et al., 2005b), similar to natural autoantibodies (Caligaris-Cappio and Ghia, 2008, Mouquet and Nussenzweig, 2011).

Commercially available assays are used to determine mutational status of the CLL, which relies on capillary sequencing of reverse transcribed/PCR products of peripheral blood and bone marrow aspirate from each patient. The percentage BCR mutation is calculated by comparing the IgHV sequences to the germline sequence database (difference of > 2% from germline counterpart is classified as mutated BCR).

1.3.5.2.   Interphase fluorescence in-situ hybridization

Interphase fluorescence in-situ hybridization (iFISH) has shown that trisomy 12, and deletions in 13q14 are correlated with IgHV mutational status in CLL (Hamblin et al., 1999, Damle et al., 1999). Trisomy 12, and chromosome 13 and 14 abnormalities are the most common genomic aberrations associated with CLL.

Juliusson *et al.* found that CLL patient with a normal karyotype (n=173) had a median overall survival of greater than 15 years, compared to 7.7 years for karyotypic abnormalities (n=218) (Juliusson et al., 1990). In addition, patients with single karyotypic abnormalities (n=113) had better prognosis than those with complex karyotypes (p-value<0.001). Within the subset of patients with single karyotypic abnormalities, patients with trisomy 12 (n=67) had poorer survival than patients with chromosome 13q (n = 51) (p-value=0.01), where the latter has the same survival as those with a normal karyotype. Patients with chromosome 14q had significantly poorer survival than those with abnormalities of chromosome 13q (p-value<0.05).

1.3.5.3.   CD38 expression on CLL B-cells

CD38 is a single chain type II transmembrane glycoprotein, which is expressed in a discontinuous manner in normal B-cell development. CD38 can be detected in high levels in B-cell precursors, germinal centre cells and plasma cells, and lower expression is usually seen in the peripheral blood and tonsillar B-cells of health individuals. CD38 function is thought to include complex ectoenzymatic activity and signal transduction for regulation of cell proliferation and survival (Kumagai et al., 1995). However, CD38 expression has been seen in a proportion of CLL patients, and is correlated with survival outcomes and a increase probability of requiring treatment, including continuous chemotherapy or chemotherapy with two or more agents or regimens (Durig et al., 2002). Specifically a CD38-negative patient group required minimal or no treatment, remained treatment-free for a longer time period and had prolonged survival (p-value<0.05 between CD38-negative CLL subgroup (<20% of the CLL cells expressed membrane CD38, n = 77) and CD38-positive CLL subgroup (≥20% of the CLL cells expressed membrane CD38, n = 56)).

1.3.5.4.   Zeta-associated protein-70 expression on CLL B-cells

Zeta-associated protein (ZAP-70) expression is involved in T-cell receptor signal transduction (Elder et al., 1994, Iwashima et al., 1994). Normal B-cells do not express ZAP-70, but has been shown to be overexpressed in IgHV unmutated CLL by microarray analyses, and serves as a surrogate marker for IgHV mutational status (Rosenwald et al., 2001, Klein et al., 2001, Wiestner et al., 2003). *In vitro*, ZAP-70 is

involved in the signal transduction cascade initiated by BCR stimulation in IgHV unmutated CLL (Chen et al., 2002, Chen et al., 2005, Crespo et al., 2003). Importantly, there is a statistically significant inferior clinical course for CLL patients with ZAP-70 expression (Krober et al., 2006, Schroers et al., 2005).

1.3.5.5. CLL proliferation centres

B-cell signalling pathways have been found to be associated with proliferative potential in neoplastic cells in CD38-positive, ZAP-70-positive and unmutated CLL patients, and it has been suggested that stimulation may occur in the "pseudo-follicular" proliferation centres (PCs) (Ferrarini and Chiorazzi, 2004). PCs are focal aggregates of variable sizes scattered in the lymph nodes, and their presence are observed in CLL (Ratech et al., 1988, Granziero et al., 2001, Schmid and Isaacson, 1994, Soma et al., 2006). Similar proliferation centres have also been seen in the inflamed tissues of patients with systematic autoimmune and inflammatory disorders, such as rheumatoid arthritis (Takemura et al., 2001) and multiple sclerosis (Corcione et al., 2005). It is thought that the clustering of pro-lymphocytes and CLL cells forms pseudo-follicular proliferation centres, where small lymphocytes accumulate and overflow into the peripheral blood. The PC microenvironment consists of pro-lymphocytes and CLL cells intermixed and surrounded by CD3+ T-cells (most of which are CD40 and CD40+), which are in close contact with the proliferating malignant B-cells (Ghia et al., 2002). Follicular dendritic cells have been observed in some PC (Ratech et al., 1988, Schmid and Isaacson, 1994), along with stromal cells and accessary cells interspersed with the small lymphocytes (Caligaris-Cappio, 2003, Burger and Kipps, 2006).

1.3.6. Current treatments for CLL

In CLL, the decision to treat is guided by clinical staging, symptoms, and disease activity. Patients in early stages of disease (Rai 0-II or Binet A) are generally only monitored but not treated unless associated CLL symptoms occur. It has only been shown that treatment is beneficial for patients at later stages (Rai III-IV or Binet B-C), but no statistical difference in outcome has been found by treating patients at earlier stages (Rai I-II or Binet A). Disease activity is typically monitored by

lymphocyte doubling time (the time it takes for the number of lymphocytes to double) of less than 6 months or by the rapid growth of lymph nodes, and is often an indication to commence treatment (Hallek and German, 2005). Although the aim for treatment is disease eradication, most patients who have a complete response typically have minimal residual disease. The term minimal residual disease (MRD) refers to low-level disease, often after incompletely effective chemotherapy (Paietta, 2002). The general consensus for MRD level is between 0.01% and 0.035% leukemic cells within a morphologically normal appearing bone marrow, as the detection of less than 0.01% of leukaemic cells by flow cytometry may not be reliably reached due to variability of technical expertise in different clinical laboratories (Campana, 2010, Coustan-Smith et al., 2000, Paietta, 2002). Detection of disease relapse after therapy is of great clinical importance, particularly to determine if further CLL therapy is required. Flow-cytometry and real-time quantitative polymerase chain reaction techniques are typically used for clinical monitoring of MRD (Moreton et al., 2005). The main therapies available to CLL patients are:

- *Alkylating agents ± prednisone (chlorambucil, cyclophosphamide)*

Chlorambucil and other alkylating agents can bind to cellular structures such as membranes, RNA, proteins and DNA. It is thought that DNA cross-linking is the most important mechanism for anti-tumour activity. Despite the relative benefit to some patients with the use of chlorambucil, drug resistance and relapse remains a problem. Furthermore, CLL cells are typically not highly proliferative, therefore raising questions about the anti-tumour activity of the drug (Begleiter et al., 1996). Prednisone is a corticosteroidal immunosuppressant drug which acts by repressing the activity of transcription factors such as activating protein-1 (AP-1) and nuclear factor-κB (NF-κB), thus inhibiting cytokine production, changes the expression of various oncogenes, induces cell-cycle arrest and apoptosis (Inaba and Pui, 2010).

- *Combination chemotherapy with alkylating agents (COP, R-CHOP)*

R-CHOP is composed of rituximab, cyclophosphamide, hydoxydaunorubicin, oncovin, prednisone respectively for the letter abbreviations, and COP is composed of cyclophosphamide, vincristine, prednisone. Hydoxydaunorubicin prevents DNA and RNA from replicating, oncovin inhibits during the M phase of the cell cycle and

prednisone is anti-inflammatory. The synthetic corticosteroid drug, prednisone, is an immunosuppressant used in the treatment of some inflammatory diseases, such as allergies, as well as cancer in higher doses. Prednisone is converted via hepatic metabolism to prednisolone, which irreversibly binds to the alpha and beta glucocorticoid receptors. The glucocorticoid receptor-prednisone complexes dimerise, and interact with nucleic DNA leading to gene transcription alterations. However, the long-term use of prednisone and other steroids has been associated with development of osteoporosis, where bone loss is observed in approximately 50% of patients taking 7.5mg prednisone for more than 3 months, and 25% of patients developed osteoporotic fractures, and further patients developed osteonecrosis (Van Staa et al., 2000). Prednisone and other steroids have been shown to promote apoptosis of osteoblasts and osteoclasts, as well as reducing the recruitment of osteoblasts and osteoclasts from progenitor cells (Weinstein et al., 1998). Therefore, the suitability of these combination therapies need to assessed on an individual basis, particularly in reference to the risk of developing osteo-related complications.

In a randomised study of 287 stage B CLL patients, treatment response was improved with CHOP (n=147) compared to chlorambucil plus prednisone (n=140) (p-value=0.007, chi-square test), but showed no difference in survival (p-value=0.33, score test). However, for stage C CLL patients, there were no significant differences treatment response and survival between CHOP (n=44) or CHOP plus methotrexate (n=46). Therefore, even though CHOP has been shown to improve therapy response, questions remain about its effectiveness at treating advanced CLL patients (Binet, 1994). These treatments are used in other B-cell malignancies, where 2-year and 5-year follow-up studies have shown that the outcome of elderly diffuse large B-cell lymphoma (DLBCL) patients on R-CHOP therapy regimens have shown significant increases the rate of complete response, decreases the rates of treatment failure and relapse, better event-free survival and overall survival compared to CHOP alone (Coiffier et al., 2002, Feugier et al., 2005).

- *Purine analogs*

Fludarabine, pentostatin, and cladribine are the three purine analogs currently used in CLL. Pentostatin inhibits adenosine deaminase by mimicking adenosine, thus reducing the cell's capability to process DNA (Sauter et al., 2008) and typically used in patients who have relapsed as well as those with acute graft-versus-host disease

(Bolanos-Meade et al., 2005). Cladribine works by a similar manner as pentostatin, with complete response and overall response rates similar to fludarabine (Robak, 2001), although 18-42% of patients experience fever side-effects after cladribine infusion (Van Den Neste et al., 1996, Saven et al., 1999). Cladribine is also used in the treatment of treatment of symptomatic hairy cell leukaemia, and is in clinical trials for use in the treatment of multiple sclerosis (Giovannoni et al., 2010). Fludarabine inhibits DNA synthesis by hindering ribonucleotide reductase and DNA polymerase. Fludarabine affects both resting and dividing cells, therefore works on both cancerous and healthy cells. Fludarabine monotherapy produces the best longer overall survival rates, but combination with Chlorambucil has shown some increased benefit (Rai et al., 2000, Johnson et al., 1996). Fludarabine has been combined with purine analogs (such as low-dose fludarabine with cyclophosphamide ± mitoxantrone), that have been shown to be effective in a subset of elderly CLL patients while with low infectious complications and negligible toxic side-effects (Marotta et al., 2000).

- *Monoclonal antibodies (campath-1H, rituximab)*

Campath-1H (alemtuzumab) is a humanised anti-CD52 antibody, an antigen on the surface of normal and malignant lymphocytes. This treatment has also been approved for the treatment of multiple sclerosis. However, the exact mechanism of campath-1H is not fully defined (Hu et al., 2009).

Rituximab is a CD20-specific monoclonal antibody that causes potent antibody-mediated B-cell cytotoxicity. Depletion of circulating B-cells from the pre-B-cell stage to the pre-plasma cell stage can lead to reduction and, in some cases, remission of CLL (Grillo-Lopez et al., 2002, Cragg et al., 2005). However, germinal centre B-cells have been found to be resistant to killing, potentially due to poor tissue penetration by rituximab (Grillo-Lopez et al., 2002). Rituximab is now being used in patients with autoimmune disease (Buch et al., 2011). Sometimes, a combination of chemotherapy and immunotherapy is used, such as fludarabine, cyclophosphamide, rituximab, fludarabine and campath-1H.

- *Transplantation (auto, allo, RIC)*

The use of auto-grafting in CLL is being an increasingly frequent treatment option. However, complications can be caused by fludarabine in stem-cell harvesting.

Hematopoietic stem cell transplantation (SCT) has been explored in clinical trials in younger patients with associated adverse disease risk factors. Although autologous SCT is not curative, it has a low treatment-associated mortality rate. Only a small number of patients are offered myeloablative allogeneic SCT due to high treatment-associated morbidity and mortality (Gribben, 2009).

1.3.7. B-cell receptors in CLL

B-cell malignancies have been found to typically express dominant clonal IgH receptors (Arber, 2000), and a variety of assays have been developed to assess B-cell clonality for diagnosis of B-cell cancers, such as in CLL and mantle cell lymphoma (MCL) (Campbell et al., 2008).

The suggestion that CLL B-cells are selected by antigenic pressure is reinforced by a number of studies showing highly restricted and biased IgHV gene usage in the B-cell repertoire of CLL patients compared to normal adult repertoire (Kipps et al., 1989, Herve et al., 2005a, Schroeder and Dighiero, 1994, Fais et al., 1998, Chiorazzi and Ferrarini, 2003, Stevenson and Caligaris-Cappio, 2004, Tobin et al., 2004a, Ghiotto et al., 2004, Messmer et al., 2004, Widhopf et al., 2004, Tobin et al., 2004b). Similar CLL BCRs are expressed between different CLL patients arising from common V-(D-)J gene usage in the heavy and light chains that share structural features such as CDR3 length, amino acid composition and joining regions, such as IgHV1-69 with IgHJ6 in the unmutated CLL, and IgHV4-34 in the mutated CLL. These stereotyped BCRs in CLL supports the hypothesis that BCR reactivity may play an important role in the CLL leukaemogenesis, potentially through activation by common antigen or auto-antigen (Tobin et al., 2004a, Ghiotto et al., 2004, Messmer et al., 2004, Widhopf et al., 2004, Tobin et al., 2004b).

CLL B-cells have been shown to express more than one IgHV allele in about 3.1% patients (Visco et al., 2013). This phenomenon can be explained either by the expression of two productive BCRs in a monoclonal CLL clone, or the presence of two distinct clonal expansions, known as bi-clonal CLL (Langerak et al., 2011). The prevalence of dual BCR expression in a single CLL clone has been reported in up to 5% of CLL cases, and thought to be due to incomplete allelic exclusion or secondary rearrangements of the IgH locus (Visco et al., 2013, Katayama et al., 2001, Rassenti and Kipps, 1997). Bi-clonal CLL is defined as the presence of two or more

phenotypically or morphologically distinct leukemic populations (Sanchez et al., 2003).

Multiple B-cell neoplasms are frequently encountered in patients, with associations of CLL with small lymphocytic lymphoma (SLL) follicular lymphoma (FL) (Boiocchi et al., 2012, Sanchez et al., 2006), and hairy cell leukaemia with CLL and SLL (Gine et al., 2002). Indolent B-cell lymphomas can develop into more aggressive disease, such as by Richter transformation or the transformation of FL to diffuse large B-cell lymphoma (DLBCL) (Boiocchi et al., 2012). Composite neoplasms can be clonally related, as suggested by related IgV gene rearrangements in cells from two lymphomas. Many of such cases have been shown to exhibit both shared and distinct somatic mutations, suggesting separate development of the lymphomas from a common premalignant precursor (Rosenquist et al., 2004b, Rosenquist et al., 2004a, Tinguely et al., 2003, van den Berg et al., 2002, Kuppers et al., 2001, Marafioti et al., 1999, Brauninger et al., 1999, Schmitz et al., 2005). Therefore, multiple B-cell neoplasms represent models to understand the transformation process in tumourigenesis and development of heterogeneous tumour populations from shared cancer precursors. Therefore, detection and monitoring of B-cell populations in lymphoid malignancies is of great clinical importance.

1.4. B-cell Acute lymphoblastic leukaemia

1.4.1. Aetiology and epidemiology

Acute lymphoblastic leukaemia (ALL) is the most common childhood leukaemia, where children account for two thirds of all ALL cases. Typically, children with ALL have a better prognosis than adult patients with ALL. Through the use of combinations of drug therapies, outlined in Section 1.4.4, between 80-90% of children are cured (Pui et al., 2008, Fielding, 2008), but the cure rate in adults is 30-40% (Pui et al., 2008). Relapse remains the leading cause of morbidity and mortality in children. The reasons for the difference in cure rates between children and adults is not fully understood, but thought to comprised of multiple factors including different therapeutic protocols between these groups and differences in biology between the disease groups.

1.4.2. Biology, pathogenesis and diagnosis of ALL

ALL is thought to develop from a single leukaemic progenitor cell with the capability of indefinite clonal expansion. Different subtypes of ALL are based on the stage of lymphoid differentiation at which leukaemogenesis occurred, either in the committed lymphoid B-cell (1-2%) or T-cell lineage (15-20%), or an early precursor B-cell (80%) or early precursor T-cell (~2%) (Reaman, 2002).

Diagnosis is confirmed by the presence of lymphoblasts on a bone marrow biopsy or aspirate, or peripheral blood smear, containing more than 20-25% of cells with the immunophenotype for ALL (Sabattini et al., 2010). Lymphoid lineage cells can be confirmed by immunophenotyping, which also distinguishes between B-cell and T-cell lineages as well as stage of differentiation (Huh and Ibrahim, 2000). Distinguishing acute myeloid leukemia (AML) from ALL is routinely achieved by staining leukemic cells for myeloperoxidase (MPO), where ALL is typically MPO-negative (Bennett et al., 1981, Bennett et al., 1976). Additional risk stratification and prognostic estimation of patients presenting with ALL include complete blood count, bone marrow and CNS involvement, cytogenetic studies, and tests for additional infections.

Common symptoms of ALL include fatigue (50%), fever (60%), pallor (skin paleness, 25%) and weight loss (26%). Bone pain caused by infiltration of blast cells

into the marrow cavity and periosteum occurs in 23% of patients (Dworzak and Panzer-Grumayer, 2003, Silverman and Sallan, 2003). The large burden of leukemic cells in patients with B- or T- cell ALL or B-cell precursor leukaemia commonly results in blood hyperkalemia (excess potassium), hyperuricemia (excess uric acid), and hyperphosphatemia (phosphate excess) with secondary hypocalcemia (low serum calcium). Therefore, intravenous hydration and sodium bicarbonate are often used to alkalise the urine, and hyperuricemia is treated with allopurinol, and hyperphosphatemia is treated with aluminum hydroxide or calcium carbonate (Pui et al., 1997). The peripheral blast-cell count can be reduced before chemotherapy by allopurinol, a purine synthesis inhibitor (Masson et al., 1996). Infiltration and involvement of the central nervous system (CNS) is found in <5% of children with ALL at presentation. The symptoms of CNS involvement include vomiting, headache, papilledema (swelling of the optic disc) and abducens nerve palsy (cranial nerve VI dysfunction) (Craig, 2003, Ma et al., 1997, Downing and Shannon, 2002). Fever is presented in at least half of ALL patients, either due to pyrogenic cytokines released from leukemic cells, including IL-1, IL-6 and tumor necrosis factor (Dinarello and Bunn, 1997) or from infection. These symptoms are often treated with broad-spectrum antibiotics until infection can be excluded (Hughes et al., 1987).

1.4.3. Prognostic markers in ALL

Individuals with Down's syndrome or ataxia telangiectasia have an increased risk of developing ALL (Hasle et al., 2000, Morrell et al., 1986). A number of prognostic factors in ALL have been determined, where some are co-associated with age (summarised in Table 1.12). Furthermore, additional genetic modifications have been found in relapsed ALL different to that seen in patients at presentation. Of significance are mutations in the histone acetyl transferase domain of cyclic adenosine monophosphate (cAMP) response element-binding protein found in approximately 20% of relapsed ALL cases, particularly in hyperdiploid ALL (seen in 60% of relapsed patients), thought to be a result of clonal selection during disease course rather than clonal evolution (Inthal et al., 2012).

Table 1.12. Genomic and cell-based prognostic factors in ALL.

Genomic aberration	Risk association*	Reference
Hyperdiploidy	Younger age and better prognosis	(Aguiar et al., 1996, Burmeister et al., 2010)
t(12;21) [ETV6/RUNX1] translocation	Younger age and better prognosis	(Aguiar et al., 1996, Burmeister et al., 2010)
t(9;22) [BCR/ABL1]	Older age and worse prognosis	(Secker-Walker et al., 1991)
Complex karyotype	Older age and worse prognosis	(Secker-Walker et al., 1991)
Hypodiploidy	Older age and worse prognosis	(Secker-Walker et al., 1991)
Janus kinase 1 and 2 mutations	Poor prognosis and associated with T-cell precursor ALL in adults	(Mullighan et al., 2009)
Ikaros family zinc finger protein 1 mutations	Worse prognosis	(Kuiper et al., 2010)
Cytokine receptor-like factor 2 translocations	Older age and worse prognosis	(Chen et al., 2012)
BCR-ABL1 translocations	Worse prognosis	(Roberts et al., 2012)
Intra-chromosomal amplifications of chromosome 21 (the gain of at least three copies of the RUNX1 region)	Worse prognosis	(Moorman et al., 2007)
Philadelphia chromosome	Worse prognosis	(Fielding et al., 2009)

Other risk factors	Risk association*	Reference
T-cell ALL	Worse prognosis	(Neumann et al., 2012)
B-cell ALL	Better prognosis	(Neumann et al., 2012)
Early T-cell precursor ALL (CD3$^+$, CD5^{weak}, CD8$^-$, CD1a$^-$ expression)	Worse prognosis	(Neumann et al., 2012)

* All studies based on comparing the outcomes of multiple patients with or without each corresponding risk factor, where statistically significant prognostic associations have p-values <0.05.

1.4.4. Current treatments for ALL

The ALL treatment regimen is typically determined by patient age and the Philadelphia chromosome status. Philadelphia chromosome is a chromosomal abnormality with the reciprocal translocation between chromosome 9 and 22 that has a statistically poor prognosis (Fielding et al., 2009). Children and young adults are treated with paediatric regimens, and Philadelphia chromosome-positive ALL patients receive tyrosine kinase inhibitor (TKI, such as such as imatinib) in addition to chemotherapy. Typically there are three treatment phases: (a) induction phase, (b) consolidation phase and (c) maintenance phase. The outline of these phases are discussed below (Larson et al., 1995, Kantarjian et al., 2004, Thomas et al., 2004, Rowe et al., 2005, Cortes et al., 1995).

Induction phase treatment

The aim of the induction phase is patient remission, defined by healthy blood cell counts, the absence of leukemic cells in the bone marrow and repopulation of the bone marrow with healthy cells. Combinations of chemotherapy drugs are used in this stage according to patient risk profile (such as defined in Table 1.12), but typically including vincristine (a mitotic inhibitor), dexamethasone or prednisone (as an anti-inflammatory and immunosuppressant drug), and doxorubicin, daunorubicin, or another anthracycline drug (DNA intercalating agent that inhibits DNA and RNA synthesis). Treatment of leukemic cells that have entered the CNS or to prevent leukaemic cells from entering CNS includes intrathecal chemotherapy often involving methotrexate (an antimetabolite). Radiation therapy may be used directly to the brain or spinal cord.

Consolidation phase treatment

For patients that achieve remission, a short course of chemotherapy is performed lasting a few months. Typically the same drugs are used as in the induction phase given in high doses. For high relapse risk patients (as defined in Table 1.12) and those with poor prognostic factors, allogeneic or autologous stem cell transplant can be given. CNS prophylaxis may be continued.

Maintenance phase treatment

A maintenance chemotherapy program of methotrexate and 6-mercaptopurine (an immunosuppressive drug) is given to patients after the consolidation phase. This phase usually lasts about 2 years. Additional drugs, such as imatinib, are given to Philadelphia chromosome-positive ALL patients, and CNS prophylaxis may be continued.

1.4.5. Monitoring minimal residual disease in ALL

MRD testing is routinely used in most paediatric ALL protocols and an increasing number of adult ALL trial protocols. The risk stratification and clinical significance of MRD depends intrinsically on the MRD assays used and time points tested (Bruggemann et al., 2010, Borowitz et al., 2003).

The clinical evaluation of treatment responses in ALL patients is achieved with a range of MRD assays. B-ALL and T-ALL cells have distinct clonal rearrangements in their B- or T-cell receptors respectively, and are often associated with the expression of gene fusions and leukaemia-associated immunophenotypes. Assays based on PCR or flow cytometry have the sensitivity to detect one ALL cell in at least 10^4-10^5 healthy cells from clinical samples (summarised in Table 1.13) (Campana, 2010). However, this may not be sensitive enough to detect MRD considering only a single or small number of B-cells is required for disease relapse and >10^6 B-cells are taken in a typical 10ml blood sample.

MRD is currently most frequently quantified using real-time quantitative PCR (qPCR) (van der Velden et al., 2007). Using immunoglobulin rearrangements in B-ALL patients or TCR rearrangements in T-ALL patients has proven to be sensitive and quantitative. However, in some patients, ongoing immunoglobulin or TCR rearrangements occur generating leukaemic subclones with distinct sequences, which can be undetected at diagnosis but become the dominant clone subsequently. Therefore, recommendations have been made to monitor two or more different rearrangement from diagnosis or to use additional MRD assays such as flow cytometry (van der Velden et al., 2007, Faham et al., 2012).

Genetic abnormalities are carried in most ALL cells. qPCR of gene fusions are most frequently used for MRD detection, such as BCR-ABL1, ETV6-RUNX1, MLL-AFF1, and TCF3-PBX1. These recurrent abnormalities are present and suitable for MRD monitoring in about 40% of ALL patients (Campana, 2010, Bruggemann et al., 2010). The benefits of qPCR are the rapidity of the procedure that does not require sequencing or patient-specific primer design, and the stable association of the gene fusion and the ALL clone. Although the use of mRNA for qPCR is typically more sensitive than DNA due to the higher copy number per cell, mRNA is prone to degradation leading to potential false negative results and the number of transcripts per cell for a fusion gene may be variable between patients, thus quantification is difficult (Gabert et al., 2003).

ALL cells express cell surface markers that resemble closely the origin of B- or T-lymphoid precursors. Healthy T-lymphoid precursors typically do not circulate but instead occupy the thymus. Therefore blood or bone marrow cells with cell surface markers resembling that of T-lymphoid precursors is sufficient to identify T-

ALL (Coustan-Smith et al., 2006). Detection of B-ALL cells typically relies on the aberrant expression of cell markers known as the leukaemia-associated immunophenotypes, which can be identified in more than 95% of B-ALL cases and typical for each ALL disease subtype (Campana, 2010).

An alternative MRD assay is next-generation sequencing of the BCR or TCR repertoires in B-ALL and T-ALL respectively. ALL cells typically arise from the leukaemic transformation of a single lymphoid precursor, therefore each B- or T-ALL cell has a unique B- or T-cell receptor rearrangement respectively that is a unique marker for the leukaemic clone in high-throughput sequencing data (Bruggemann et al., 2010). Markers for MRD in B-ALL can be determined at the time of diagnosis when the leukaemic B-cell load is greatest and the peripheral blood exhibits significant B-cell clonal expansion. B-cell BCR repertoire analysis has been used to identify IgHV-D-J gene usage in the leukaemic B-cell clone(s). Subsequent monitoring of the patient can be used to follow the leukemic B-cell population during and after therapy, enabling early detection of minimal residual disease (Brisco et al., 2009). This approach overcomes the requirement of patient-specific reagents while achieving a sensitivity of >1 in 10^6 cells (Logan et al., 2011, Faham et al., 2012, Gawad et al., 2012). Furthermore, this approach allows for the assessment of the total diversity of the B- or T-cell populations, thus able to follow the ALL sub-clonal evolution as well as the major malignant clone (Ladetto et al., 2013).

Table 1.13. The main clinical assays used to monitor MRD in acute lymphoblastic leukaemia.

Adapted from (Campana, 2010).

Molecular target	Method	% of patients that can be monitored	Sensitivity for the detection of MRD	Strengths	Weaknesses
BCR and TCR gene rearrangements	qPCR	~90%	0.01%-0.001%	• High sensitivity • Consensus protocols established • Generally accurate quantification	• Laborious • Clonal evolution and secondary rearrangements may lead to false negatives • The requirement for more than one target reduces the applicability of the assay
Fusion transcripts	qPCR	~40%	0.01%-0.001%	• Rapid • Unambiguous association of fusion transcript with leukaemic/pre-leukaemic clone • Stable throughout therapy/clonal evolution	• Quantification not certain • RNA instability may lead to false negatives
Leukaemic immunophenotype	Flow cytometry	~95%	0.01%	• Widely applicable • Rapid • Accurate quantification	• False positive/negatives without expertise in sample processing and data interpretation • Phenotype shifts during course of disease means multiple sets of molecular markers are required

1.5. Aims and hypotheses

Mapping of BCR and TCR repertoires promises to transform our understanding of adaptive immune dynamics, with applications ranging from identifying novel antibodies and determining evolutionary pathways for haematological malignancies to monitoring of minimal residual disease following chemotherapy (Weinstein et al., 2009, Woof and Burton, 2004b, Tonegawa, 1983). The aim of this thesis is to investigate B-cell diversity in health and disease as follows:

1. Investigating and developing robust methods for analysing high-throughput B-cell receptor sequencing repertoires.

2. Determining robust methods for deep-sequencing B-cell receptor populations.

3. Investigating B-cell repertoire dynamics following treatment of acute lymphoblastic leukaemia and investigating MRD and relapse.

Chapter 2

2. Materials and methods

2.1. Samples

Peripheral blood mononuclear cells (PBMCs) were isolated from 10ml of whole blood from healthy volunteers and CLL patients using Ficoll gradients (GE Healthcare), summarised in Table 2.1. For the B-ALL peripheral blood samples, DNA and RNA extraction was performed by incubation with erythrolysis buffer for 15 minutes, centrifugation, discarding supernatant (repeated twice), and resuspension of cells in PBS. Total RNA was isolated using TRIzol® and purified using RNeasy Mini Kit (Qiagen) including on-column DNase digestion according to manufacturer's instructions. Total RNA was also isolated from 1×10^6 cells from Human lymphoblastoid cell lines (LCLs) from the HapMap project (Frazer et al., 2007), where the number of passages was unknown. DNA extraction was isolated using TRIzol® and the MiniPrep kit (Qiagen) according to manufacturer's instructions. CLL and B-ALL samples were approved by the relevant institutional review boards and ethics committees (07/MRE05/44 and EEBK/EΠ/2014/15 respectively).

Table 2.1. Table of samples used.

Sample ID in chapter	Sample ID in chapter 4	Patient type	Age, years	Gender	Time since CLL	Genomic abnormality	Phenotype
CLL 1	-	CLL	77	Male	7	13q deletion	Unknown
CLL 2	-	CLL	58	Male	2	Stable trisomy 12	Atypical: CD23 negative
CLL 3	-	CLL	78	Male	1.5	No FISH performed	CD38+
CLL 4	-	CLL + HCC	77	Male	2.5	No FISH performed	Unknown
CLL 5	Pat. 2	CLL	59	Female	1.25	No abnormalities seen	Unknown
CLL 6	-	CLL	67	Male	2	11q deletion	Unknown
CLL7	-	CLL	69	Male	13	No FISH performed	Recurrent haemolysis in the past
CLL 8	-	CLL	64	Male	4.5	No FISH performed	Unknown
CLL 9	-	CLL	77	Male	5.25	No FISH performed	Unknown
CLL 10	-	CLL	81	Male	8	13q deletion	Lambda light chain restricted, CD38-
CLL 11	Pat. 6	CLL	81	Male	10	No FISH performed	Unknown
-	Pat. 1	CLL + prostate carcinoma	67	Male	8	No FISH performed	Unknown
-	Pat. 3	CLL	80	Female	7	13q deletion	Unknown
-	Pat. 4	CLL	82	Female	5	No FISH performed	Unknown
-	Pat. 5	CLL	82	Male	0.75	No FISH performed	Unknown
-	Pat. 7	CLL	72	Female	4.5	13q14.3 deletion	Unknown
-	Pat. 8	CLL	64	Male	9.5	13q14.3 deletion	Unknown
-	Pat. 9	CLL	71	Male	8	No FISH performed	Unknown
-	Pat. 10	CLL	80	Male	5	No FISH performed	Unknown
-	Pat. 11	CLL	56	Male	0.75	No FISH performed	Unknown
-	Pat. 12	CLL	81	Male	1	13q deletion	CD5 negative
-	Pat. 13	CLL	63	Male	1.8	No FISH performed	Unknown
Healthy 1	Healthy 1	Age matched control 1	74	Female	-	Mix of t(11;14), 11q deletion, 13q addition	Anaemia
Healthy 2	Healthy 2	Age matched control 2	62	Female	-	NA	NA
Healthy 3	Healthy 3	Age matched control 3	75	Female	-	NA	NA
Healthy 4	Healthy 4	Age matched control 4	67	Female	-	NA	NA
Healthy 5	Healthy 5	Age matched control 5	68	Female	-	NA	NA
Healthy 6	Healthy 6	Healthy 6	55	Male	-	NA	NA
Healthy 7	Healthy 7	Healthy 7	23	Male	-	NA	NA
Healthy 8	Healthy 8	Healthy 8	23	Male	-	NA	NA
Healthy 9	Healthy 9	Healthy 9	25	Male	-	NA	NA
Healthy 10	Healthy 10	Healthy 10	24	Female	-	NA	NA
Healthy 11	Healthy 11	Healthy 11	24	Female	-	NA	NA
Healthy 12	Healthy 12	Healthy 12	24	Female	-	NA	NA
Healthy 13	Healthy 13	Healthy 13	24	Female	-	NA	NA

2.2. B-cell methods

2.2.1. RT-PCR

RT-PCR reagents were purchased from Invitrogen. The FR1 and FR2 primer sets used (supplied by Sigma Aldrich) are described by Van Dongen *et al.* (van Dongen et al., 2003) and in Table 2.1. Reverse transcription was performed using 500ng of total RNA mixed with 1μl JH reverse primer (10μM), 1μl dNTPs (0.25mM), and RNase free water added to make a total volume of 11μl. This was incubated for 5 minutes at 65°C, and 4μl First strand buffer, 1μl DTT (0.1M), 1μl RNaseOUT™ Recombinant Ribonuclease Inhibitor and 1μl SuperScript™ III reverse transcriptase (200units/μl) was added. RT was performed at 50°C for 60 minutes before heat-inactivation at 70°C for 15 minutes (**Figure 2.1**). PCR amplification of cDNA (5μl of the RT product) was performed with the JH reverse primer and the FR1 or FR2 forward primer set pools (0.25 μM each), using 0.5μl Phusion® High-Fidelity DNA Polymerase (Finnzymes), 1μl dNTPs (0.25mM), 1μl DTT (0.25mM), per 50μl reaction. For multiplex PCR amplification of DNA, 30ng of DNA was mixed with the JH reverse primer and the FR1 forward primer set (0.25 μM each), using 0.5μl Phusion® High-Fidelity DNA Polymerase (Finnzymes), 1μl dNTPs (0.25mM), 1μl DTT (0.25mM), per 50μl reaction. The following PCR program was used: 3 minutes at 94°C, 35 cycles of 30 seconds at 94°C, 30 seconds at 60°C and 1 minute at 72°C, with a final extension cycle of 7 minutes at 72°C on an MJ Thermocycler.
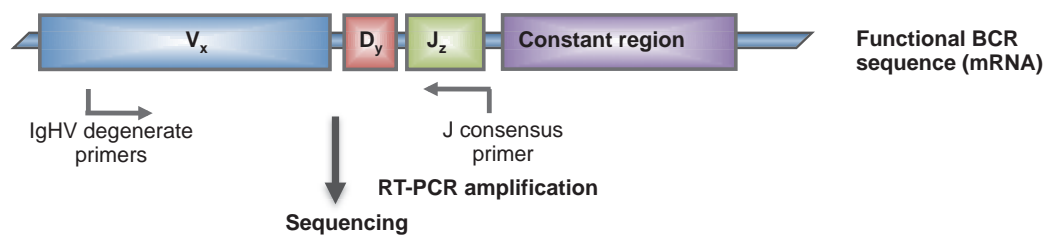
Figure 2.1. Sequencing of B-cell receptor repertoires.

Representation of the genomic rearrangement process during V-D-J recombination to generate the heavy chain B-cell receptor. B-cell receptor amplification was performed by reverse transcription on total RNA by single J region primer, and subsequent multiplex PCR amplification.

Table 2.1. Human B-cell receptor PCR primers.

Primer	Sequence	
JH reverse	CTTACCTGAGGAGACGGTGACC	
VH1-FR1 forward	GGCCTCAGTGAAGGTCTCCTGCAAG	
VH2-FR1 forward	GTCTGGTCCTACGCTGGTGAAACCC	
VH3-FR1 forward	CTGGGGGGTCCCTGAGACTCTCCTG	FR1 primer set*
VH4-FR1 forward	CTTCGGAGACCCTGTCCCTCACCTG	
VH5-FR1 forward	CGGGGAGTCTCTGAACATCTCCTGT	
VH6-FR1 forward	TCGCAGACCCTCTCACTCACCTGTG	
VH1-FR2 forward	CTGGGTGCGACAGGCCCCTGGACAA	
VH2-FR2 forward	TGGATCCGTCAGCCCCCAGGGAAGG	
VH3-FR2 forward	GGTCCGCCAGGCTCCAGGGAA	
VH4-FR2 forward	TGGATCCGCCAGCCCCCAGGGAAGG	FR2 primer set*
VH5-FR2 forward	GGGTGCGCCAGATGCCCGGGAAAGG	
VH6-FR2 forward	TGGATCAGGCAGTCCCCATCGAGAG	
VH7-FR2 forward	TTGGGTGCGACAGGCCCCTGGACAA	
B-actin forward	CGCCTTTGCCGATCCGCCG	
B-actin reverse	CTTCTCGCGGTTGGCCTTGGG	
GAPDH forward	GAAGGTGAAGGTCGGAGTC	
GAPDH reverse	GAAGATGGTGATGGGATTTC	
B-globin forward	CTGCCGTTACTGCCCTGTGGG	
B-globin reverse	GGACAGCAAGAAAGCGAGCTTAGTG	

* The FR1 primers were pooled to give the FR1 primer set, and the FR2 primers were pooled to give the FR2 primer set. Primer JH reverse was used to prime cDNA synthesis. The expected amplicon sizes for the IgHV PCR products for JH reverse/FR1 primer set is 310-360bp, and the expected size ranges for the IgHV PCR products for JH reverse/FR2 primer set is 260-295bp. The expected amplicon sizes for beta-actin and beta-globin PCR products are 150bp and 340bp respectively.

2.2.2. RNA capture for sequencing BCR repertoires

Total RNA was initially processed for target enrichment using the NEBNext kit (NEB) according to manufacturers protocol. Briefly, mRNA was isolated by polyA+ selection and converted to cDNA. cDNA at 0.3 to 0.7ng/μl was fragmented to 200bp (Covaris), ligated to sequencing adaptors (Illumina) and size selected at 200bp (Life Technologies E-Gel). Samples were then indexed for pre-capture pooling (NEBNext Multiplex Oligos for Illumina Index Primers 1 to 12). A pre-capture library was generated using 12 cycles of PCR (KAPA Biosystems Library Amplification Kit). Libraries were pooled and hybridised to biotinylated RNA-capture baits (custom design (Fisher et al., 2014), full protocol available on request), Agilent SureSelect) at 65°C for 24 h. Hybridised fragments were selected using streptavidin magnetic beads, washed and eluted for multiplexed sequencing on Illumina Miseq.

2.2.3. 5' Rapid amplification of cDNA ends (5'RACE) of B-cell receptors

5'RACE was performed using SMARTer™ Pico PCR cDNA Synthesis Kit (Clontech) according to Clontech protocols, using the JH-reverse primer (Table S3) and SMARTer 5' primer for PCR amplification.

2.2.4. Sequencing methods

454-libraries were prepared using standard Roche-454 Rapid Prep protocols incorporating 10-base multiplex identifier (MID) tags and sequenced using an FLX Titanium Genome Sequencer (Roche/454 Life Sciences). MiSeq libraries were prepared using Illumina protocols and sequenced by 250bp or 300bp paired-ended MiSeq (Illumina) as indicated. Raw 454 or MiSeq reads were filtered for base quality (median >32) using the QUASR program (http://sourceforge.net/projects/quasr/) (Watson et al., 2013). MiSeq forward and reverse reads were merged together if they contained identical overlapping region of >65bp, or otherwise discarded. The 250bp reads from the 5'RACE experiment were retained if they contained a JH-reverse primer sequence and orientated to begin with IgHV gene. Reads from RNA-capture were BLAST aligned to reference IgH genes, and trimmed if the reads extended outside the IgHV-D-J region, and filtered for length (>160bp). Non-immunoglobulin

sequences were removed and only reads with significant similarity to reference IgHV genes from the IMGT database (Lefranc et al., 2009) using BLAST (Altschul et al., 1990) were retained ($1x10^{-10}$ E-value threshold). Primer sequences were trimmed from the reads, and sequences retained for analysis only if both primer sequences were identified and if sequence lengths were greater than 255bp or 195bp for FR1 and FR2 primed samples respectively for 454, or both forward and reverse reads greater than 110 bp for MiSeq. FR1 primed PCR samples from CLL patients were also Sanger-sequenced.

2.2.5. Per-base error quantification

The same PCR protocol and read quality filtering was used to amplify beta-actin, beta-globin and GAPDH genes from two healthy individuals (amplicon sizes of 150bp, 340bp respectively). The sequence representing the majority of the reads for each sample was classified as the 'true' gene sequence for that individual to account for individual allelic variation. Any differences between this sequence and the reads were considered to be PCR and/or sequencing error and classified as homopolymeric indels (occurring in a region of two or more consecutive identical bases), non-homopolymeric indels, or mismatches.

2.2.6. Reference-based V-D-J assignment

BLAST (Altschul et al., 1990) was used to align the 454 sequences against known BCR sequences from the ImMunoGeneTics (IMGT) database (Lefranc et al., 2009). Due to the difference in length of the different gene families, different BLAST e-value thresholds were used for the IgHV, IgHD, and IgHJ-genes (10^{-70}, 10^{-3} and 10^{-20} respectively).

2.2.7. Network assembly and analysis

The network generation algorithm is summarised in **Figure 2.2**. Briefly, each vertex represents a unique sequence, where the relative size of the vertex is proportional to the number of sequence reads identical to the vertex sequence. Edges were calculated between vertices that differed by single nucleotide non-indel differences. The network generation was performed using custom Python scripts

using CD-Hit (Li and Godzik, 2006) and analyses were performed using igraph implemented in R (http://igraph.sourceforge.net/index.html). The distribution of mismatches within a single network cluster were determined by aligning the sequence representing the largest vertex with the sequences to which it is connected and the positions of mismatches were determined along the sequences. Two-sided t-tests were performed in R.
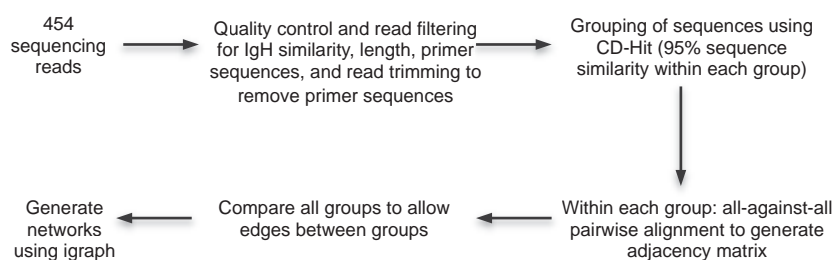
Figure 2.2. Outline of network generation method.

The sequencing reads were initially filtered for base quality (median >32) using QUASR quality control program (http://sourceforge.net/projects/quasr/). To filter for non-immunoglobulin sequences, only reads that had significant similarity to a reference IgHV gene from the IMGT database (Lefranc et al., 2009) using BLAST (Altschul et al., 1990) (E-value threshold of 1×10^{-10}) were retained. The primer sequences were trimmed from the reads, and sequences were retained for analysis only if both primer sequences were identified and the sequence lengths were greater than 255bp for FR1 primed samples, or 195 bp for FR2 primed samples. Fast algorithms were used to cluster the reads into groups of similar sequences (with greater than 95% sequence identity) using the CD-HIT program (Li and Godzik, 2006). Within each group, all-against-all pairwise alignments were performed, using customized python scripts, to determine all intra-group edges within the network, after which each group was compared to determine inter-group edges. Due to the reported prevalence of 454 derived sequencing errors in homopolymeric regions (Albers et al., 2011), homopolymeric indels were not included in the total number of mismatches in pairwise comparisons. The network analyses were performed using igraph implemented in R (http://igraph.sourceforge.net/index.html).

2.2.8. Diversity measure calculations

The Gini index was calculated by ordering the cluster sizes from largest to smallest and creating a cumulative frequency distribution, where $R = \{r_1, r_2, \dots, r_n\}$, r_i is the cumulative size of the all the largest clusters until the ith largest cluster and normalized such that $r_n = 1$. The Gini index is $Gini\ index\ (g) = \sum_{i=1}^{N} \frac{(r_i - (i/N))}{N}$, where N is the number of clusters (Morrow, 1977).

2.2.9. Estimation of cluster sizes due to sequencing error

The Poisson distribution can estimate the expected number of reads containing i errors from the (central) vertex of size n reads, given an estimated error rate. The expected number of sequences with i errors is $n.p_i$, where $p_i = P(X = i) = \frac{\lambda\ e^{-\lambda}}{i!}$, and λ is the expected number of mutations per read. A cluster is defined as a set of interconnected vertices, in which edges are generated between vertices that differ by a single base. A vertex v is only included in a cluster when the minimum distance from v to any of the sequences in the cluster containing the central vertex is one. Thus, all the sequencing errors at $i=1$ generate vertices that have edges connecting to the central vertex. At $i >1$, a vertex with set of mutations M_x will be connected to the cluster only if there exists a vertex in the cluster with a set of mutations M_y such that $\left| \frac{M_x}{M_y} \right| = |\{x \in M_x | x \notin M_y\}| = 1$ (i.e. there is only one mutation in M_x that is not in M_y). Therefore the probability of vertices due to i sequencing errors is estimated by drawing $S[n, i]$ samples from a multinomial distribution, for which the probability of the possible vertices that could connect to the cluster is given by $S[n, i] = \prod_{j=1}^{i-1} \frac{E[n, j-1]}{l} . p_i$, where l is the length of the sequence and $E[n, j]$ is the estimated number of vertices that are in the cluster which are at distance of j from the central node. 1000 independent samples were drawn from the multinomial distribution to estimate the average number of vertices at distances i from the central vertex, and therefore the cluster size due to sequencing error can be estimated by summing over the expected number of vertices at all i, $1 \leq i \leq \infty$.

2.2.10. Phylogenetic analysis of BCR sequences

BCR sequences related to the largest cluster were aligned using Mafft (Katoh and Standley, 2013) and a maximum parsimony tree was fitted using Paup* (Wilgenbusch and Swofford, 2003). The branch lengths represent the evolutionary distance between BCR sequences and bootstrapping was performed to evaluate the reproducibility of the trees, showing strong tree support (>95% certainty for all branches) as determined by *phangorn* in R (Schliep, 2011).

2.2.11. Linear discriminant analysis of BCR repertoire parameters

Each numerical B-cell repertoire feature was normalised to sum to one over all samples. The *lda* function in R was performed to find a linear combination of features that best separates sample types (Rindskopf, 1997), projected over the first and second LDA dimensions. Hierarchical clustering of samples was performed using *hclust* in R (Murtagh and Contreras, 2012), where the distance measures between any two samples *i* and *j* was determined by:

$$d = \sqrt{\left(LDA_1^i - LDA_1^j\right)^2 + \left(LDA_2^i - LDA_2^j\right)^2}$$

Where LDA_1^i and LDA_2^i are the first and second LDA dimension values for sample *i* respectively.

Chapter 3

3. Developing computational methods for assessing B-cell receptor populations from next-generation sequencing

3.1. Introduction

To date next-generation sequencing (NGS) of BCRs have primarily focused on classifying the IgHV, D and J recombination frequencies to understand the diversity of the BCR repertoire (Boyd et al., 2009, Campbell et al., 2008, Maletzki et al., 2012, Lev et al., 2012, Jager et al., 2012, Weinstein et al., 2009). However, computational assignment of V-D-J sequences to reference databases results in many incompletely identified IgHV, D and J genes even when the germline alleles are known (Weinstein et al., 2009). This is most likely due to somatic hypermutation (SHM) masking the identity of the germline genes present in the NGS, or the existence of new diverse IgH gene alleles not present in the reference database. Further, investigation of V-D-J gene usage frequencies utilises only part of the BCR sequence diversity with important information about the V-D-J joining regions and somatic hypermutations not considered.

This chapter describes the development of analysis methods for BCR sequence data using the full BCR V-D-J sequence variation and that does not rely on prior V-D-J gene classification. It was previously shown that zebrafish BCR repertoire diversity can be interpreted through full V-D-J genotype diversity using BCR networks, and that these are an intuitive way for understanding B-cell repertoires (Ben-Hamo and Efroni, 2011). In such networks, the lowest level of organisation in a population of B-cells, namely unique B-cells, are represented by sparse networks whereas highly developed (connected) networks most likely result from clonal expansions of B-cells, arising through antigenic exposure or B-cell malignancies (Ben-Hamo and Efroni, 2011). However, such analyses have never been applied to mammals, or during infection or disease. In this chapter, network methods were developed to provide a robust framework for analysing vast NGS sequencing repertoires from B-cell populations. This chapter aimed to distinguish between

diverse B-cell populations and clonal B-cell populations both qualitatively and quantitatively.

3.2. Results

3.2.1. Next-generation sequencing of IgH variable genes

RT-PCR amplification of the expressed rearranged IgHV-D-J loci from mRNA from human B-cell populations was performed using the consensus IgHJ primer and FR1 or FR2 IgHV family primers (**Figure 2.1** and Table 2.1) (van Dongen et al., 2003). Peripheral blood (PB) samples from thirteen healthy individuals, eleven CLL patients, and eight LCLs yielded PCR products of expected sizes (310-360bp for FR1 and 250-295bp for FR2 primed samples) and were 454 sequenced (Table 3.1). Samples yielded an average of 42,324 sequencing reads after filtering for quality and presence of IgH sequence (Table 3.2). Briefly, only reads were retained with median base quality Phred scores of greater than 32, with significant similarity to reference IgHV genes (E-value $<1x10^{-10}$), and with identifiable primer sequences. Two additional samples from CLL patient A (pre and post treatment) were sequenced on the MiSeq platform (Table 3.2). The BCR 454 sequence datasets from *Boyd et al.* (Boyd et al., 2009) were also analysed, which includes three healthy individuals and five patients with clonal blood disorders (Table 3.3).

Table 3.1. Patient sample information.

Sample	Patient type	Age, years	Gender	Time since CLL diagnosis, years
CLL 1	CLL	77	Male	7
CLL 2	CLL	58	Male	2
CLL 3	CLL	78	Male	1.5
CLL 4	CLL+HCC	77	Male	2.5
CLL 5	CLL	59	Female	1.25
CLL 6	CLL	67	Male	2
CLL7	CLL	69	Male	13
CLL 8	CLL	64	Male	4.5
CLL 9	CLL	77	Male	5.25
CLL 10	CLL	81	Male	8
CLL 11	CLL	81	Male	10
Healthy 1	Age matched control 1	74	Female	-
Healthy 2	Age matched control 2	62	Female	-
Healthy 3	Age matched control 3	75	Female	-
Healthy 4	Age matched control 4	67	Female	-
Healthy 5	Age matched control 5	68	Female	-
Healthy 6	Healthy 6	55	Male	-
Healthy 7	Healthy 7	23	Male	-
Healthy 8	Healthy 8	23	Male	-
Healthy 9	Healthy 9	25	Male	-
Healthy 10	Healthy 10	24	Female	-
Healthy 11	Healthy 11	24	Female	-
Healthy 12	Healthy 12	24	Female	-
Healthy 13	Healthy 13	24	Female	-

* Abbreviations: CLL=chronic lymphocytic leukemia, HCC=Hepatocellular carcinoma.

Table 3.2. Sample information and number of sequencing reads.

Primer	Type*	ID	Platform	Number of reads	Number of reads (after filtering**)	Average read length (bp)	Multiplex
FR1	CLL	CLL 1	454	58700	51311	290.4	Multiplex half plate C
FR1	CLL	CLL 2	454	54937	31694	290.6	Multiplex half plate C
FR1	CLL	CLL 3	454	46657	26828	310.2	Multiplex half plate C
FR1	CLL	CLL 4	454	45632	27126	287.9	Multiplex half plate C
FR1	CLL	CLL 5	454	40780	26086	294.6	Multiplex 7/8 plate D
FR1	CLL	CLL 6	454	59847	54761	310.6	Multiplex 7/8 plate D
FR1	CLL	CLL 7	454	22036	18273	303.9	Multiplex 7/8 plate D
FR1	CLL	CLL 8	454	44079	37208	308.5	Multiplex 7/8 plate D
FR1	CLL	CLL 9	454	34139	29401	305.9	Multiplex 7/8 plate D
FR1	CLL	CLL 10	454	55331	51018	311.9	Multiplex 7/8 plate D
FR1	CLL	CLL 11	454	33950	27650	301.8	Multiplex 7/8 plate D
FR1	Healthy	Healthy 1	454	56105	28638	288.4	Multiplex half plate A
FR1	Healthy	Healthy 2	454	77698	40556	288.5	Multiplex half plate A
FR1	Healthy	Healthy 3	454	45539	23848	286.6	Multiplex half plate A
FR1	Healthy	Healthy 4	454	132359	59456	286.5	Multiplex half plate A
FR1	Healthy	Healthy 5	454	53350	40435	315.9	Multiplex half plate C
FR1	Healthy	Healthy 6	454	60637	41878	292.5	Multiplex half plate C
FR1	Healthy	Healthy 7	454	50600	35852	291.8	Multiplex half plate C
FR1	Healthy	Healthy 8	454	35163	25454	296.5	Multiplex half plate C
FR1	Healthy	Healthy 9	454	34796	26849	289.3	Multiplex half plate C
FR1	Healthy	Healthy 10	454	44991	34248	291.4	Multiplex half plate C
FR1	Healthy	Healthy 11	454	33085	25083	291.9	Multiplex half plate C
FR1	Healthy	Healthy 12	454	45134	36828	299.2	Multiplex 7/8 plate D
FR1	Healthy	Healthy 13	454	40984	33792	296.2	Multiplex 7/8 plate D
FR1	LCL	LCL 1	454	65182	58117	290.5	Multiplex whole plate B
FR1	LCL	LCL 2	454	64483	53894	305	Multiplex whole plate B
FR1	LCL	LCL 3	454	24473	17285	302	Multiplex whole plate B
FR1	LCL	LCL 4	454	101156	82317	295.4	Multiplex whole plate B
FR1	LCL	LCL 5	454	53964	45325	298.3	Multiplex whole plate B
FR1	LCL	LCL 6	454	47691	40233	301	Multiplex whole plate B
FR1	LCL	LCL 7	454	43047	32340	290.4	Multiplex whole plate B
FR1	LCL	LCL 8	454	59503	50617	308.1	Multiplex whole plate B
FR2	Healthy	Healthy 1	454	43209	33628	229.3	Multiplex half plate A
FR2	Healthy	Healthy 2	454	27379	20904	228.3	Multiplex half plate A
FR2	Healthy	Healthy 3	454	23379	19009	228.1	Multiplex half plate A
FR2	Healthy	Healthy 4	454	36846	27756	226.9	Multiplex half plate A
FR2	LCL	LCL 1	454	81271	55741	239.4	Multiplex whole plate B
FR2	LCL	LCL 2	454	106236	88253	257.3	Multiplex whole plate B
FR2	LCL	LCL 3	454	117359	107230	247.4	Multiplex whole plate B
FR2	LCL	LCL 4	454	96943	88771	249.6	Multiplex whole plate B
FR2	LCL	LCL 5	454	69621	61840	240.1	Multiplex whole plate B
FR2	LCL	LCL 6	454	55010	48408	234.2	Multiplex whole plate B
FR2	LCL	LCL 7	454	57697	50834	222.1	Multiplex whole plate B
FR2	LCL	LCL 8	454	50789	45501	250.9	Multiplex whole plate B
FR1	CLL	Patient A Pre-treatment	MiSeq	56864	40414	264.3	Multiplex 1/74 lane
FR1	CLL	Patient A Post-treatment	MiSeq	42053	36197	265.4	Multiplex 1/74 lane

* Abbreviations: LCL = Human lymphoblastoid cell line, CLL = chronic lymphocytic leukemia.

** Reads were filtered for complete primer sequences and length (where reads shorter than 255bp are removed for FR1 primed samples, and reads shorter than 195bp for FR2 primed samples).

Table 3.3. Sample information and number of sequencing reads from the Boyd et al. dataset.

Primer	Type* **	ID	Number of reads (after filtering)	Average read length (bp)
FR2	Healthy donor 1, time 0	Healthy 12 a1	12316	228
FR2	Healthy donor 1, time 0	Healthy 12 a2	17943	227.9
FR2	Healthy donor 1, time 14 months	Healthy 12 b1	13189	227.3
FR2	Healthy donor 1, time 14 months	Healthy 12 b2	10361	227.6
FR2	Patient 1; CLL/SLL time 0 months	CLL/SLL 1a	2774	216.4
FR2	Patient 1; CLL/SLL time 3 months	CLL/SLL 1b	2353	213.9
FR2	Patient 2; FL	FL1	11293	228.4
FR2	Patient 3; FL and SLL in Lymph node	FL/SLL	30391	215.5
FR2	Patient 4; CLL/SLL	CLL/SLL 2	31201	227.2
FR2	Healthy donor 2	Healthy 13	24545	226
FR2	Patient 6; CLL	CLL 12	17438	225.9
FR2	Healthy donor 3	Healthy 14	29883	223
FR2	Patient 6 CLL diluted 1:10	CLL 12 1:10	13362	223.2
FR2	Patient 6 CLL diluted 1:100	CLL 12 1:100	26966	222.9
FR2	Patient 6 CLL diluted 1:1000	CLL 12 1:1000	22063	222.9
FR2	Patient 6 CLL diluted 1:10000	CLL 12 1:10000	26464	222.7
FR2	Patient 6 CLL diluted 1:100000	CLL 12 1:100000	26635	222.8

* Abbreviations: CLL = chronic lymphocytic leukemia, SLL =Small lymphocytic lymphoma, FL= Follicular lymphoma.
**From (Boyd et al., 2009).

3.2.2. Next-generation sequencing error rate

Firstly, the 454 NGS error rate was determined to assess the number of sequencing errors to expect in BCR sequencing. To achieve this, reverse transcription and PCR was performed to amplify universally expressed genes, beta-actin, beta-globin and GAPDH, from two healthy individuals (amplicon sizes of 150bp, 150bp and 340bp respectively, **Table 3.4**). After sequencing by either 454 or MiSeq, the same read quality filtering was performed as with the BCR sequences. The sequence representing the majority of the reads for each sample was classified as the 'true' gene sequence for that individual to account for individual allelic variation. Any differences between this sequence and the reads were considered to be RT-PCR and/or sequencing error and classified as homopolymeric indels (occurring in a region of two or more consecutive identical bases), non-homopolymeric indels, or mismatches. The distribution of mismatches and indels was random across the genes. By counting the base-pair differences between the true gene sequence and sequence variants, the combined per-base error-rate for the RT-PCR and sequencing process for the 454 platform was 1.74×10^{-4} (**Table 3.5**, of which homopolymeric indels and non-homopolymeric errors accounted for 59.7% (1.04×10^{-4}) and 40.3% (7.04×10^{-5}) of the total error-rate respectively). These error rates were consistent between repeats of the same genes. Of note is the high homopolymeric error-rate, which has been previously reported with 454 sequencing at similar levels (Luo et al., 2012, Wang et al., 2007, Boyd et al., 2009, Gall et al., 2013). Similarly the combined per-base error-rate for RT-PCR and MiSeq sequencing was 1.70×10^{-4} (**Table 3.6**), where, again, the error rates are consistent between repeats of the same genes and similar to previously reported error rates of 5.9×10^{-4} (Lou et al., 2013). This means for every 4,070bp sequenced, there is a 50% chance that there is at least one sequencing error using MiSeq sequencing, and for every 3,980bp sequenced there is a 50% chance that there is at least one sequencing error using 454 sequencing.

Table 3.4. Sample information and number of sequencing reads for control genes.

Sample name**	Gene	Platform	Number of gene specific reads	Number of reads after filtering*	% of original reads retained after filtering	Multiplexing
Healthy 1	Beta-Actin	454	7673	7671	99.97	Multiplex 1/8th plate D
Healthy 2	Beta-Actin	454	2109	2105	99.81	Multiplex 1/8th plate D
Healthy 1	Beta-Globin	454	6983	4871	69.76	Multiplex 1/8th plate D
Healthy 2	Beta-Globin	454	5361	3387	63.18	Multiplex 1/8th plate D
Healthy 1	GAPDH	MiSeq	93213	89821	96.36	Multiplex 1/74 Lane
Healthy 2	GAPDH	MiSeq	50386	48242	95.74	Multiplex 1/74 Lane
Healthy 1	Beta-Actin	MiSeq	21551	13696	63.55	Multiplex 1/74 Lane
Healthy 2	Beta-Actin	MiSeq	86899	56909	65.49	Multiplex 1/74 Lane
Healthy 1	GAPDH	MiSeq	187923	179831	95.69	Multiplex 1/74 Lane
Healthy 2	GAPDH	MiSeq	181150	172914	95.45	Multiplex 1/74 Lane

* Reads were filtered for homology with the corresponding target gene and subsequently filtered for intact primer sequences and for complete primer sequences and length (reads shorter than 150bp for beta-actin, and shorter than 340bp for beta-globin were removed).
High read filtering in the beta-globin samples are due to non-specific PCR amplifications.

** Amplification of beta-actin, beta-globin and GAPDH genes from two healthy individual samples.

Table 3.5. Estimated average per-base 454 error frequencies by type.

Sample	Gene	Number of reads	Average read length, bp	Overall non-homopolymeric error rate	Type Insertions	Deletions	Mismatches	Insertions and deletions Non-homopolymeric	Homopolymeric
Healthy 1	Beta Actin	7671	105	3.97E-05	5.09E-05	3.73E-06	2.11E-05	1.86E-05	3.60E-05
Healthy 2	Beta Actin	2105	104	3.17E-05	1.36E-05	4.52E-06	3.17E-05	0.00E+00	1.81E-05
Healthy 1	Beta Globin	4871	297.9	1.03E-04	1.43E-04	1.18E-04	6.00E-05	4.34E-05	2.17E-04
Healthy 2	Beta Globin	3385	294	1.06E-04	1.75E-04	1.00E-05	6.63E-05	4.02E-05	1.45E-04
Average				7.03E-05	9.55E-05	3.40E-05	4.48E-05	2.56E-05	1.04E-04

*Amplicon lengths were 104bp for beta-actin, and 295bp for beta-globin.

Table 3.6. Estimated average per-base MiSeq error frequencies.

Sample	Gene	Overall error rate
Healthy 1	GAPDH	2.78E-04
Healthy 2	GAPDH	2.71E-04
Healthy 1 (repeat)	GAPDH	2.80E-04
Healthy 2 (repeat)	GAPDH	2.81E-04
Healthy 1	Beta-Actin	6.34E-05
Healthy 2	Beta-Actin	6.30E-05
	Average	1.70E-04

3.2.3. Percentage of identical BCR reads between samples

The percentage of reads identical to the most abundant BCR sequence in each sample was determined to make an initial assessment of the differences in B-cell clonality of the PB and LCL samples. The percentage of reads corresponding to the most abundant BCR sequence in each of the CLL and LCL samples (range 39.3%-87.8% and 35.2%-78.7% respectively) were significantly higher than that of PB from healthy individuals (range 0.10% -14.0%) with a p-value <0.001 (**Figure 3.1**). There was no significant difference in the percentage of identical reads between the LCL and CLL patient samples (p-value=0.0594). Therefore, the healthy individuals represent diverse BCR populations, whereas the LCL and CLL samples represent more restricted or clonal BCR populations. Sanger and MiSeq sequencing confirmed that the dominant clonal sequences from the CLL samples were identical to that from 454 sequencing (excluding homopolymeric indels) indicating that there are no significant differences between sequencing platforms for high abundance sequences.
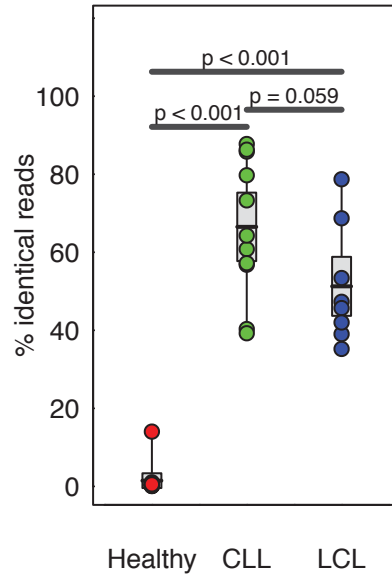
Figure 3.1. Sequencing of B-cell receptor repertoires.

The percentage of reads corresponding to the most abundant B-cell receptor sequence for each sample, separated into sample type: healthy individuals, CLL patients and LCLs. Two-sided t-tests were performed between the sample subsets, with the p-values indicated above.

3.2.4. Limitations of V-D-J gene classification

To determine the proportion of BCRs that cannot be classified in terms of IgHV, D and J gene usage, each BCR sequence was aligned to the germline sequences from the ImMunoGeneTics database (IMGT) (Lefranc et al., 2009) by BLAST (**Figure 3.2**). Due to the difference in length of the different gene families, different BLAST E-value thresholds were used for the IgHV, IgHD, and IgHJ-genes (10^{-70}, 10^{-3} and 10^{-20} respectively). The majority of sequences could be classified to their most closely related reference sequences for IgHV and IgHJ genes (an average of 99.8% and 96.1% of BCR sequences were classified respectively). Substantially fewer IgHD were identifiable (average of 40.5%) due to the shorter sequence length and potential insertions and deletions within the joining regions between the V-D-J boundaries, which has been noted in previous studies (Weinstein et al., 2009). Incomplete V-D-J gene classification may be due to SHM masking the identity of the germline genes present in individuals and/or the existence of allelic variants of reference IgH (Boyd et al., 2010a). There was no significant difference between the percentage of classified V, D and J genes of our dataset compared to that of *Boyd et al. (2009)* (**Figure 3.2**). To overcome the limitations of IgH V-D-J gene classification, the use of sequence-based network analysis is proposed next. Network analysis makes use of complete V-D-J sequence information and mutational relationships. Without the requirement of reference gene classification, network analysis is proposed to be more informative and robust framework for BCR repertoire analysis.
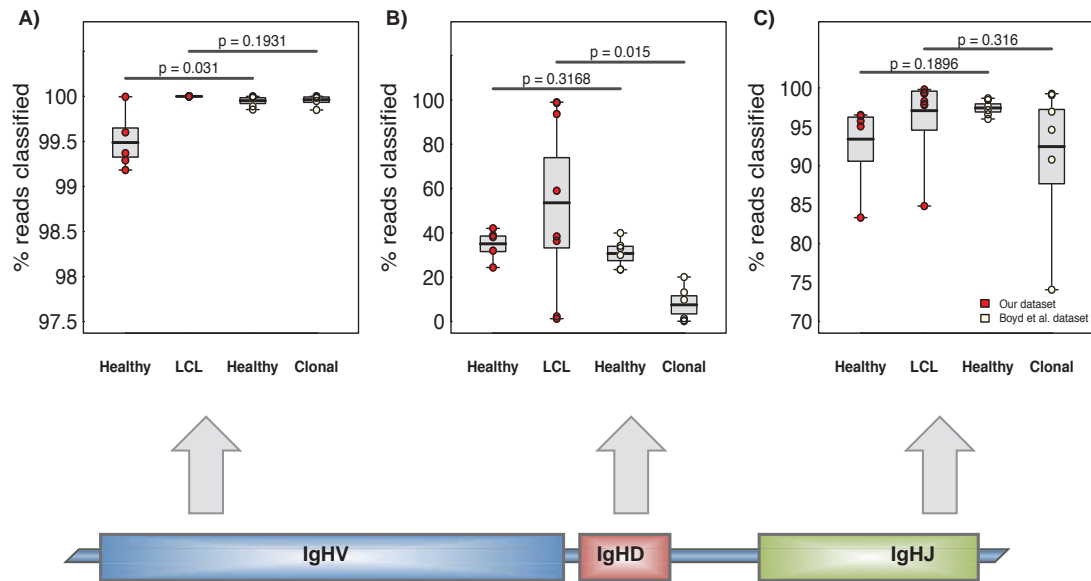
Figure 3.2. Percentage of reference sequences matched to 454 reads.

For **A)** IgHV, **B)** IgHD and **D)** IgHJ genes. Samples from healthy individuals or human lymphoblastoid cell lines from this study (red) and the dataset *Boyd et al. (2009)* (white). The clonal samples in *Boyd et al.*'s dataset refer to patients with CLL, small lymphocytic lymphoma and/or follicular lymphoma. The BLAST e-value thresholds for the IgHV, IgHD, and IgHJ-genes were 10^{-20}, 10^{-3} and 10^{-4} respectively. P-values were calculated using two-sided t-test in R.

3.2.5. BCR sequences organise into networks based on sequence diversity

It is reasonable to consider each different BCR sequence as a distinct product from amplification of a rearranged BCR from a B-cell. Therefore the B-cell repertoire can be represented as a network representing BCR sequence space. Networks are powerful tools for understanding the overall structure of large multidimensional datasets, where information is represented in the form of vertices and edges between vertices. Here, networks are able to represent the BCR sequence repertoire in the following way: a vertex represents a different sequence, and the number of identical BCR sequences defines the vertex size. Edges are created between vertices that differ by one nucleotide, i.e. highly related BCR sequences. Clusters are groups of interconnected vertices, where any two vertices in a cluster are related by the set of point mutations indicated by the edge-path between them (**Figure 3.3**). Therefore, code was developed here in python to analyse high-throughput BCR sequencing data to generate the networks.
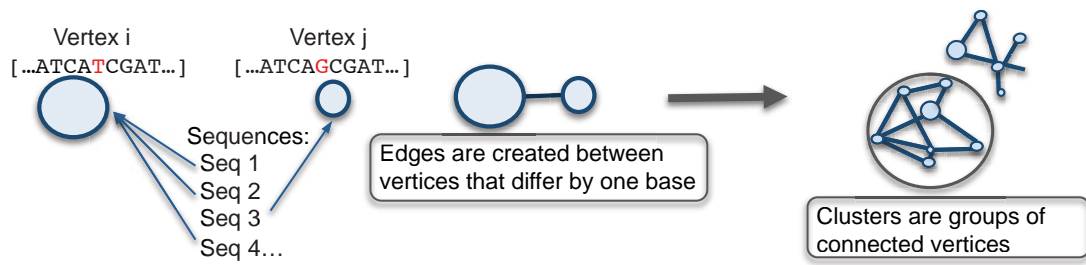
Figure 3.3. Generation of B-cell receptor sequence networks.

Schematic diagram showing the method by which the sequencing networks are generated: each vertex represents a unique sequence, where the relative size of the vertex is proportional to the number of sequencing reads that were identical to the vertex sequence. Edges are created between vertices that differ by one base (indel or substitution).

To test the analysis of high-throughput BCR sequencing data by networks, filtered and trimmed 454 or MiSeq sequences for each sample were used directly to generate a sequence network (**Figure 3.4**). Differences in network architectures are clearly seen by comparing B-cell populations from healthy individuals and LCLs. In LCLs, the majority of BCR sequences fall within a small number of clusters (greater than 40% of all sequencing reads form the largest cluster in each sample), as these samples are predominantly comprised of a small number of large B-cell clone types (**Figure 3.4**Ai). In contrast, healthy individuals have sparsely connected networks where most sequences are unique, thus yielding small vertices indicative of high overall BCR sequence diversity in the sampled repertoire (**Figure 3.4**Aii). From healthy individuals, the largest cluster representing 16.7% (4023 reads) of the total population in healthy individual 10.

The maximum CLL vertex sizes differ between samples (39.2-99.5% of total sequences) suggesting that large but variable-sized subsets of B-cells express the predominant BCR sequence(s), surrounded by BCR variants (including total process errors) of the dominant sequence. Of note, the extent of cluster size diversity is different between CLL samples, with some displaying extensive clonal enlargement (**Figure 3.4**Bi) whereas others have more limited clonal expansion (**Figure 3.4**Bii) or expansion of two dominant clusters (**Figure 3.4**C). Although the clinical relevance of dual clonal expansions are not known, previous studies have shown that the presence of two expanded IgH rearrangements can be either due to multiple productive gene rearrangements or the co-existence of two expanded clones with the CLL phenotype (Plevova et al., 2014). Therefore, the magnitude of connectivity of different samples varies between individual patients with CLL. However, in all cases, the CLL sequence networks are clearly distinct from the sparsely connected age-matched healthy individual BCR networks.

A i)
LCL sample 1
58,117 sequences total
99.6 % maximum cluster

ii)
Healthy control 1
40,435 sequences total
0.16 % maximum cluster

B i)
CLL sample 1
51,311 sequences total
96.1 % maximum cluster

ii)
CLL sample 2
31,694 sequences total
43.5 % maximum cluster

C)
CLL sample 5
26,086 sequences total
48.0 % maximum cluster

D i)
Healthy donor 1
12,316 sequences total
0..49 % maximum cluster

ii)
Patient 2:
FL lymph node
11,293 sequences total
61.1 % maximum cluster

iii)
Patient 3:
FL and SLL lymph node
30,391 sequences total
73.29 % maximum cluster

Figure 3.4. B-cell receptor repertoires from different samples.

A) Comparison of BCR sequence networks between i) a typical LCL sample and ii) a typical healthy individual. **B)** BCR sequence networks of CLL patients with i) extensive clonal enlargement and ii) limited clonal expansion. **C)** BCR sequence networks of CLL patient 5 showing expansion of two dominant clusters. **D)** Networks generated from sequencing dataset from *Boyd et al.* (Boyd et al., 2009) of i) healthy donor 1, ii) patient 2 with follicular lymphoma (FL) and iii) patient 3 with FL and small lymphocytic lymphoma (SLL). The vertex colors correspond to the relative abundance of the corresponding sequences, where red, orange and yellow indicates observation of a sequence in >90%, between 40-90% and <40% of the reads in the sample respectively.

It is proposed that sequences within a cluster are most likely related to a single rearranged, unique IgHV-D-J BCR progenitor that has undergone proliferation and somatic hypermutation, but also potentially contains BCRs with sequencing error(s). Somatic hypermutation has been found previously to preferentially occur within the CDRs of the BCR compared to the FRW regions (Lin et al., 1997), whereas sequencing errors would be distributed randomly along the length of the BCR. This could be due to either preferential AID targeting to the CDRs, or to selection of B-cells with fewer mutations in the FRW regions, such as those that would negatively change the BCR structure. To test this, sequences within clusters were aligned, and the distribution of base-pair changes was determined (**Figure 3.5**). Although base-pair differences are distributed along the length of the 454 sequences (**Figure 3.5**A-C), in all the healthy individual samples, mutations significantly occur within the CDR regions, known to be hotspots for somatic hypermutation (Lin et al., 1997), compared to the FWR regions (p-value=0.000338, **Figure 3.5**D), suggesting that these are a result of SHM rather than errors.
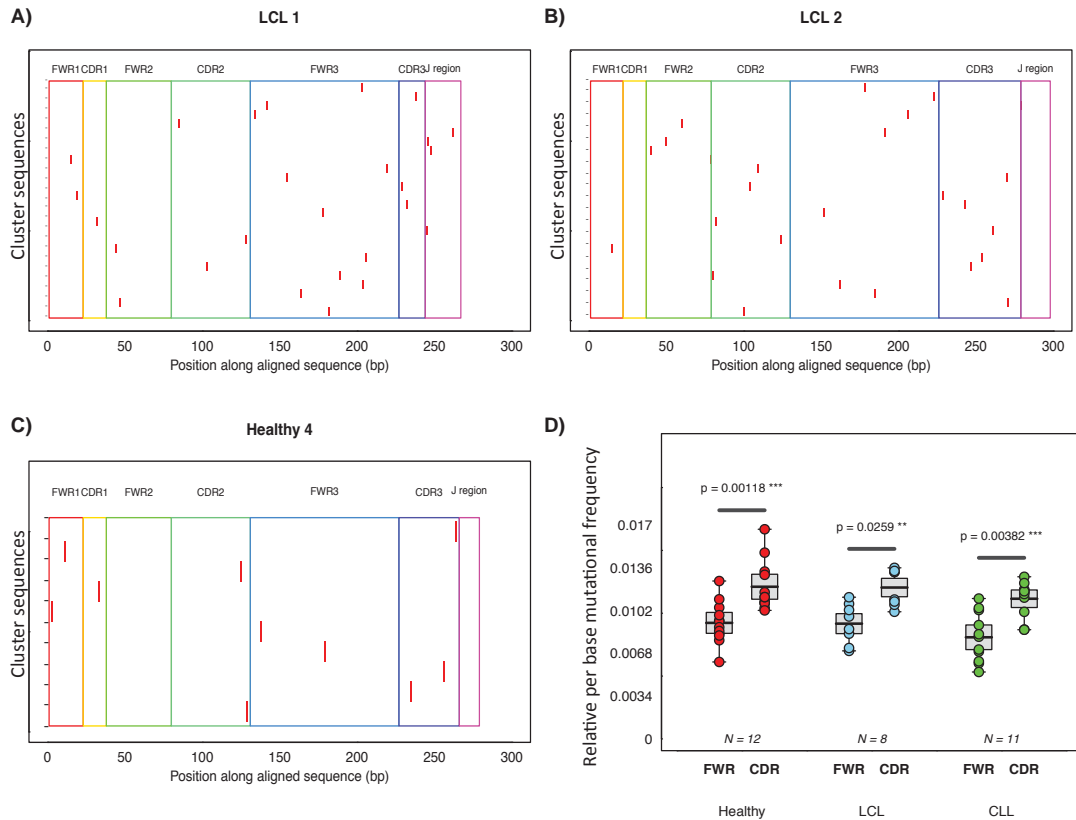
Figure 3.5. Distribution of mutations between connected vertex sequences.

Locations of mutational differences between vertices for **A-B)** two LCL samples and **C)** a representative healthy individual sample. For each sample, the sequence representing the highest connectivity was aligned with up to 25 randomly selected sequences to which it is connected, and the spatial distribution of mismatches was calculated. Each row on the figure represents a 454 sequence, and the red lines mark positions where the sequence differs in terms of mismatches. These are overlaid with the different structural regions of the BCR, as defined by IMGT/V-Quest (Giudicelli et al., 2011). **D)** The relative proportion of mutations found in either the FWR or CDR regions for all the 12 healthy, 8 LCL and 11 CLL samples. Mutations were determined by aligning all sequences separated by a single edge within each network. P-values calculated by paired T-test.

3.2.6. Population measures capture network and sample diversity

Several parameters were investigated to describe the quantitative features of the sequenced BCR repertoire from B-cell populations, including the Gini Index, maximum and second maximum cluster sizes. The Gini index is an unevenness measure, that can take a value between 0 and 1. A Gini index of 0 reflects complete equality and Gini index values close to 1 indicates high inequality or unevenness. When applied to the vertex size distribution for a given sample, these measures quantify the overall clonal nature of a sample. When the Gini index is applied to the cluster size distributions, these measures quantify the overall clustering of the sample. As shown in the previous section (Section 3.2.5), clustered sequences represent highly related BCRs with the hallmarks of SHM. Therefore, the cluster Gini index relates to the overall SHM of a sample. A high cluster Gini index is indicative of some clusters with high numbers of connected and related vertices, whereas a cluster Gini index suggests that all the clusters have more equal and lower numbers of connected and related vertices. The maximum cluster size measure is the percentage of reads corresponding to the largest cluster and indicates the degree of clonal expansion of a sample. To assess the possibility of dual clonal expansions, a measure of the second maximum cluster size as a percentage of reads in a sample was also included.

The LCL samples, due to the more restricted BCR repertoires and highly connected clusters yield high cluster and vertex Gini Indices (averages of 0.94 and 0.80, range 0.91-0.97 and 0.62-0.91 respectively) (**Figure 3.6**A) showing high unevenness of the size distributions. By contrast, B-cell networks of healthy individuals occupy a distinct region of Gini Index vertex and cluster space (averages of 0.21 and 0.05, range 0.10-0.39 and 0.03-0.11 respectively). The low vertex Gini indices shows that the healthy samples have more even distributions of vertex sizes, where each unique BCR sequence is observed a small number of times, and no BCR sequences dominate the repertoire. The low vertex Gini indices shows that the healthy samples have no clusters that dominate the repertoire. The CLL samples occupy a spatial range between healthy individuals and LCL B-cell population extremes with low vertex (between 0.62 and 0.97), and cluster Gini Indices (between 0.15 and 0.83), due to their B-cell clonal expansions. There is however considerable variation between the cluster Gini Indices, with CLL patients 1, 10 and 11 having low cluster Gini Indices, indicative of a highly expanded dominant cluster or dominant clones.

Of note, one healthy individual (healthy individual 10) has a more developed network as defined by an increase in connectivity and vertex sizes resulting in higher vertex and cluster Gini Indices (**Figure 3.6**A point (a)). This increased clonality was verified by independent sequencing using the BIOMED-2 FR2 primer set (strong linear correlation between BIOMED-2 FR1 and FR2 primed samples, R^2-value>0.996, **Figure 3.7** and **Figure 3.8**). These strong correlations also indicate no significant primer amplification bias, which has been the major caution of PCR based approaches. Further, the highest expressed BCR sequence for healthy individual 10 has 90.6% sequence identity with the closest germline IgHV gene (16 mismatches in 243bp of alignment) suggesting that this B-cell clone has undergone SHM, therefore could potentially be antigen driven.

Networks were generated from the sequences derived from *Boyd et al.* (Boyd et al., 2009) to validate these population measures on independent BCR sequence data. This showed that the clonal populations of the patients with CLL, small lymphocytic lymphoma (SLL) and/or follicular lymphoma (FL) are distinct from the diverse populations of healthy individuals (**Figure 3.6**B), occupying equivalent regions of the cluster and vertex Gini Index graphs to CLL samples within this study. Therefore, the Gini Index population measure robustly separates distinct B-cell populations into different regions based on the clonal nature of the sample and is applicable to data from other laboratories.
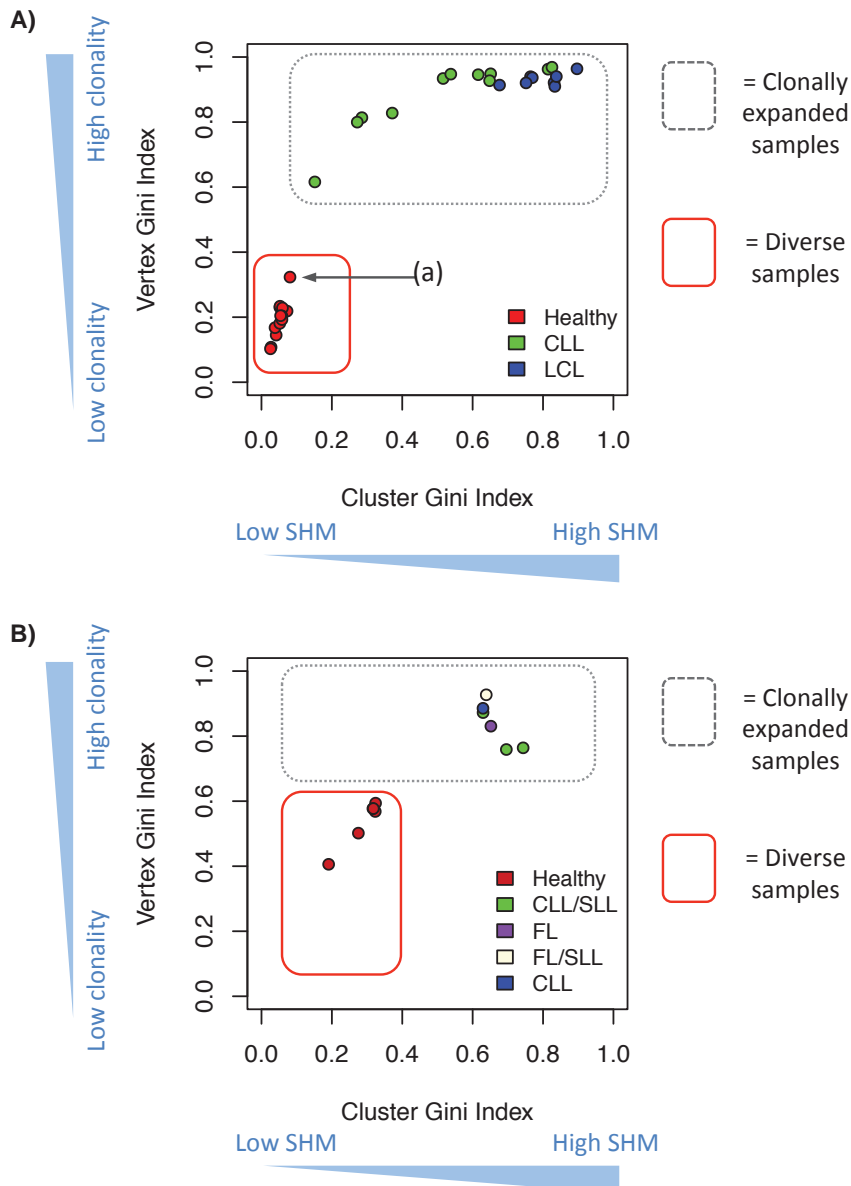
Figure 3.6. Measures differentiating between B-cell receptor populations.

The cluster Gini Index plotted against vertex Gini Index for **A)** thirteen healthy individual samples, eleven chronic lymphocytic leukemia (CLL), and eight human lymphoblastoid cell line (LCL) samples and **B)** six healthy individual samples, two samples from patients with CLL and small lymphocytic lymphoma (SLL), one sample from a patient with follicular lymphoma (FL), and one sample from a patient with FL and SLL using the dataset from *Boyd et al.* **(Boyd et al., 2009)**. The red box and grey dashed box distinguish between the regions occupied between diverse and clonal populations respectively. Point (a) corresponds with healthy individual 10.
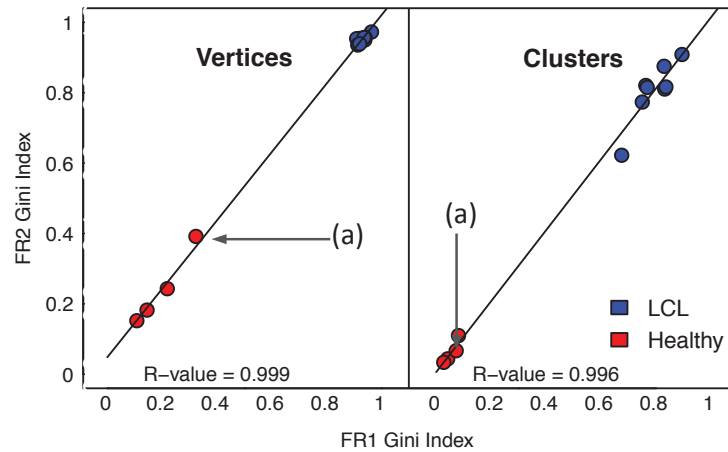
Figure 3.7. Comparison of diversities from FR1 and FR2 primer sets.

Correlation between the Gini Indices of BIOMED-2 FR1 or FR2 primed samples for vertex sizes and cluster sizes, with the corresponding Pearson R-value. LCL represented in blue and healthy individual samples in red. Point (a) corresponds with healthy individual 10.

Ai) **Healthy donor 10**
Multiplex PCR using FR1 primers
28,638 sequences total
16.79 % maximum cluster

ii) **Healthy donor 10**
Multiplex PCR using FR2 primers
33,628 sequences total
8.246 % maximum cluster

Bi) **Healthy donor 11**
Multiplex PCR using FR1 primers
40,556 sequences total
0.067 % maximum cluster

ii) **Healthy donor 11**
Multiplex PCR using FR2 primers
20,904 sequences total
0.048 % maximum cluster
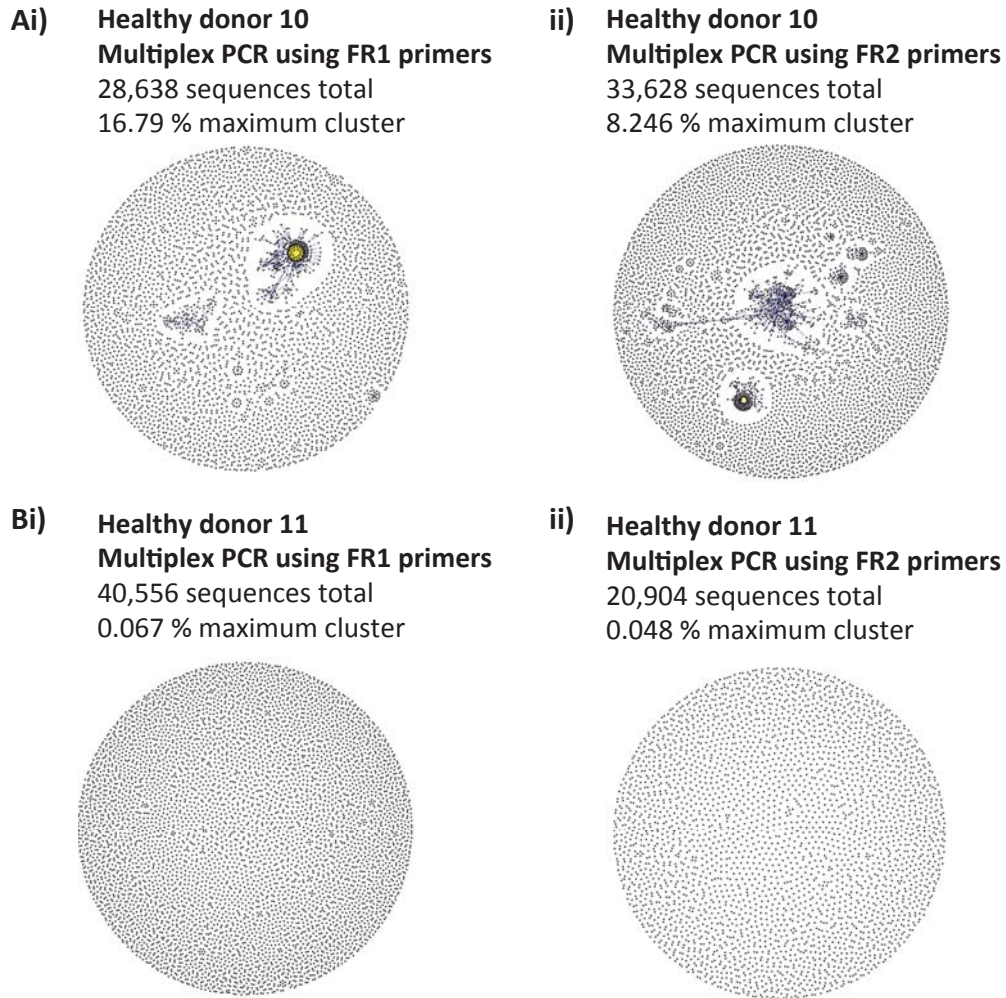
Figure 3.8. B-cell receptors networks for FR1 and FR2 primer amplified healthy donors.

A) B-cell receptor repertoires from healthy individual 10 amplified using **i)** the FR1 primer set and **ii)** the FR2 primer set. **B)** B-cell receptor repertoires from healthy individual 11 amplified using **i)** FR1 primer set and **ii)** the FR2 primer set. The vertex colors correspond to the relative abundance of the corresponding sequences, where red, orange and yellow indicates observation of a sequence in >90%, between 40-90% and <40% of the reads in the sample respectively.

Next, separation of monoclonal expansions, biclonal expansions and diverse B-cell populations was investigated using the maximum cluster sizes and second maximum cluster sizes (**Figure 3.9**A). The CLL and LCL samples have maximum cluster sizes >30% of the total reads compared to maximum cluster sizes of healthy individual samples of <20%. However, the LCLs and CLLs collectively occupy two distinct regions in this space. One group exhibits a single dominant clonal sequence (monoclonal), where all remaining clusters are <5% of the total reads (**Figure 3.9**A surrounded by the dashed line).

The second group of samples has two dominant clusters above 40% and 20% of the total reads respectively (bi-clonal). To determine whether the two dominant clusters are derived from the same B-cell lineage, alignments between the cluster sequences can be used. Firstly, if the two clusters derived from the same B-cell progenitor, they would exhibit the same IgHV-D-J rearrangement. If the two clusters came from different B-cell progenitors but have undergone the same IgHV-D-J rearrangement, the joining regions between the rearranged genes should be different. Therefore, to test whether the two dominant clusters in CLL patient 5 (**Figure 3.4**D) originate from the same B-cell progenitor the IgHV-D-J combinations were determined using IgBLAST. The two dominant clusters use different V-D-J genes ([IGHV3-66*03/IGHD6-19*01/IGHJ3*02] and [IGHV6-1*01/IGHD3-3*01/IGHJ4*02] respectively), and the alignment between the most abundant BCR sequences within these clusters show poor sequence similarity (**Figure 3.10**). Together this indicates that the two dominant clusters in CLL patient 5 originate from two different B-cell progenitors, or secondary rearrangements within the CLL clone. This could potentially be clarified by determining whether the B-cells from these two clusters share identical light chain sequences.

Limited polyclonal expansions were observed also in 5/8 of the LCL samples reflecting that EBV transformation of peripheral B-cells frequently results in polyclonal LCLs. Using the dataset from *Boyd et al.* (Boyd et al., 2009), the same phenomenon of polyclonal expansions in a subset of samples was shown (patients with CLL/SLL and FL/SLL, **Figure 3.4**Eiii) where the maximum cluster sizes are >35% and second maximum cluster sizes are >19% of the total reads (**Figure 3.9**B). Therefore the polyclonal status of the tumour samples can be determined using B-cell network reconstruction and analysis.
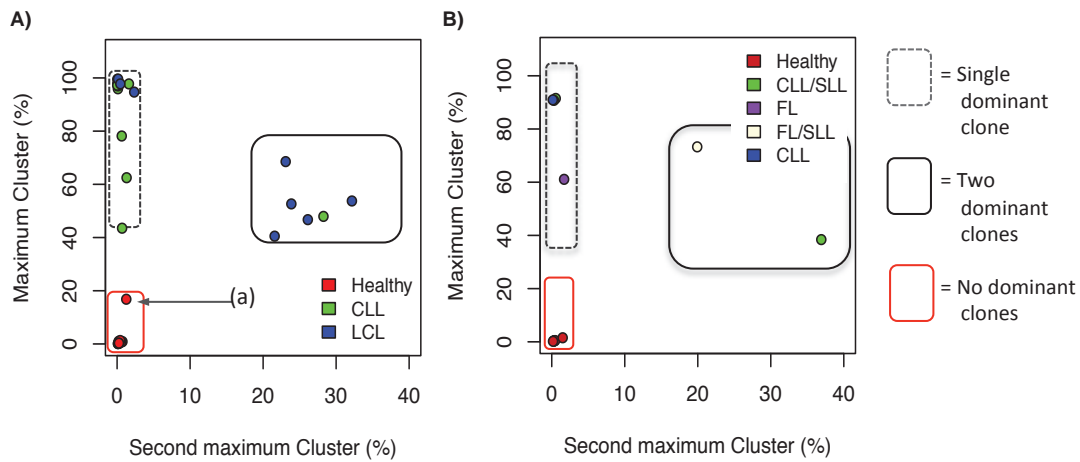
Figure 3.9. Measures differentiating between B-cell receptor dominant clusters.
The second maximum cluster sizes plotted against the maximum cluster sizes for **A)** thirteen healthy individual samples, eleven chronic lymphocytic leukemia (CLL), and eight human lymphoblastoid cell line (LCL) samples and **B)** six healthy individual samples, two samples from patients with CLL and small lymphocytic lymphoma (SLL), one sample from a patient with follicular lymphoma (FL), and one sample from a patient with FL and SLL using the dataset from *Boyd et al.* **(Boyd et al., 2009)**. The red, grey dashed and black solid boxes distinguish between the regions occupied between unexpanded populations, monoclonal expanded populations and biclonally expanded populations respectively. Point (a) corresponds with healthy individual 10.
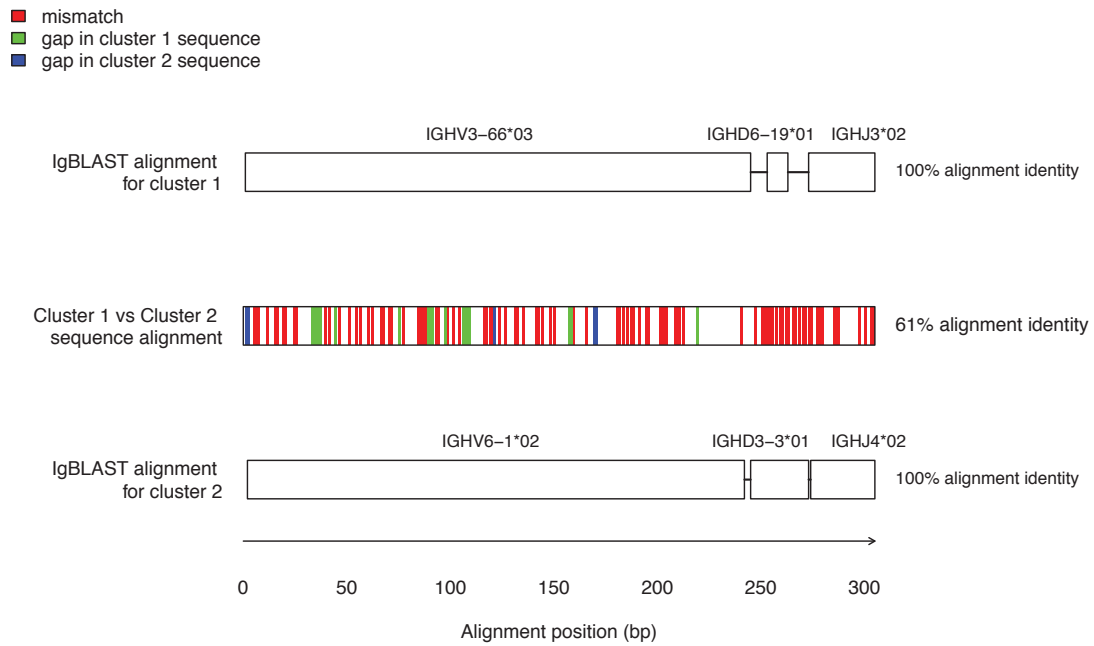
Figure 3.10. Comparison of cluster 1 and cluster 2 sequences for CLL patient 5.

Sequence alignment of the most highly expressed sequences in the two dominant clones for CLL patient 5 by ClustalW to reference IgHV genes and to each other. Cluster 1 and cluster 2 refer to the largest and second largest clonal clusters in the BCR network for CLL patient 5 respectively (representing 48% and 28.3% of the reads respectively). The cluster 1 and 2 sequences were aligned to each other, and the positions of mismatches and gaps are indicated by the coloured boxes in the corresponding alignment positions in the middle row. IgBLAST was used to identify the most similar reference IgHV, D and J genes to the cluster 1 and 2 sequences, shown in the corresponding rows, with the regions of alignments indicated by the boxes. These showed 100% sequence identity between the CLL cluster sequences and the reference germline sequences, as indicated by the text to the right of the alignments.

3.2.7. Network property sensitivity to sequencing depth and edge lengths

To robustly compare B-cell populations between samples, the population measures used must reflect differences in population structure rather than variations in depth of sequencing (scale invariant) and volume of PB sample. If a given diversity measure is scale invariant for B-cell networks then the network diversity measure should be the same regardless of the depth of sampled sequences, i.e. a subset of sequences should yield the same network diversity measure as the full set of sequences. All the proposed population measures were tested as a function of sequencing depth by randomly sampling different proportions of the sequence data for each sample followed by calculation of the corresponding network parameters for both the vertex and cluster size distributions for the LCL, CLL and healthy samples. All the proposed measures showed little variation at different sample sizes even when sub-sampling as low as 20% of the original total data (**Figure 3.11**A-D). Below 20%, small deviations in the Gini Index measures are seen, which is due to low sampling depth leading to higher relative sampling stochasticity. Therefore, this suggests a minimal read depth of ~8,000 for use in comparing populations. As these network measures had minimal standard deviation over all sub-sampling ranges, they are therefore robust parameters for inter-sample comparison.

Secondly, it was hypothesised that generating networks to allow edges to join BCR sequences with greater than one base-pair difference would not greatly influence the network architecture. This hypothesis is based on the assertion that any two B-cells derived from different progenitor B-cells would yield IgHV-D-J rearrangements or non-template insertions and deletions that would differ by only a few base-pairs. To test this, networks were generated to include edge lengths of up to 5 base-pair changes (**Figure 3.12**). It is shown that networks with edges between BCR sequences that differ by up to 5 base-pairs faithfully retain the network architecture for both the clonal and diverse samples (from LCLs and healthy individuals respectively).
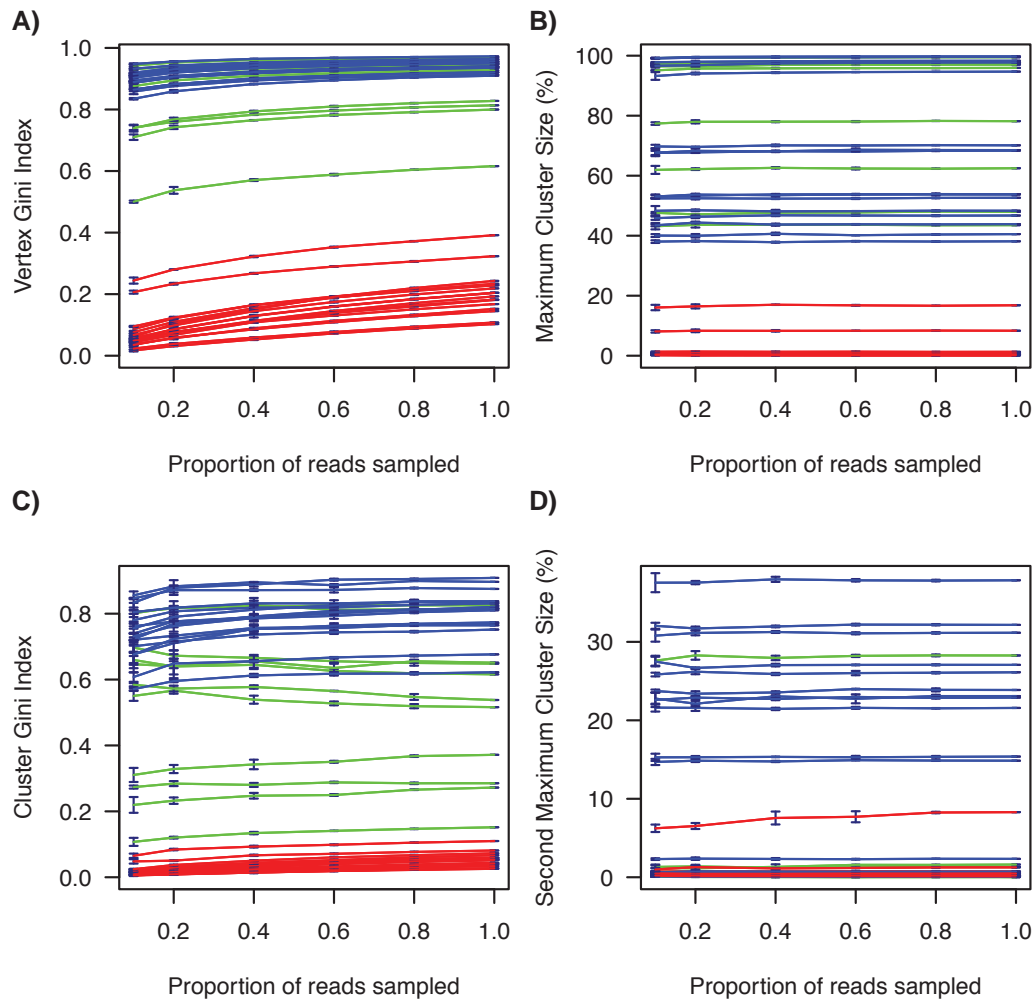
Figure 3.11. Variation of BCR population measures with sampling depth.

The 454 sequences were randomly sampled at a range of proportions of the overall number of reads for eight human lymphoblastoid cell line (LCL) (blue) and thirteen healthy individual samples (red) and eleven patients with chronic lymphocytic leukaemia (CLL) (green). The variation of the diversity measures against varying sequencing depth using, respectively, the Gini index for **A)** vertices and **B)** clusters, and **C)** maximum and **D)** second maximum cluster sizes. An average of 4 subsamples was taken at each proportion, and the error bars give the standard deviation from the mean for each network measure.
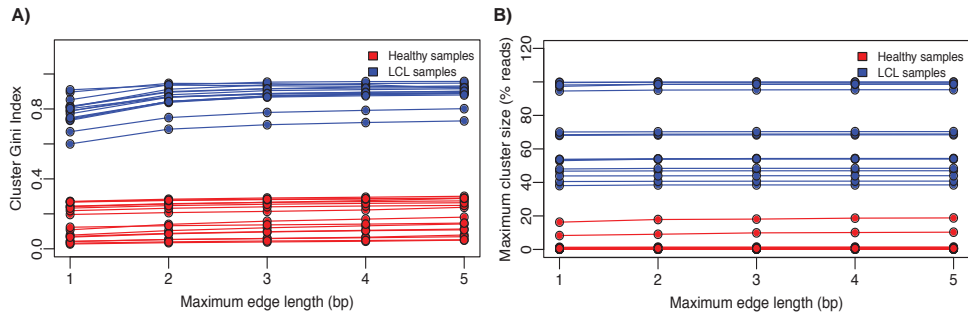
Figure 3.12. Network structure variation with edge length.

A) Cluster Gini Indices and **B)** maximum cluster sizes for networks with different maximum edge lengths. For each sample, networks were generated allowing generation of edges between vertices that differ by at most the corresponding number of mismatches (edge lengths). The corresponding network parameters were calculated for each network, and plotted.

3.2.8. Minimal effect of sequencing errors on network properties

Next, it was investigated whether the diversity of sequences within clusters were likely to be due to the process of somatic hypermutation or sensitive to or generated through sequencing error of a unique amplified BCR sequences. For a given BCR sequenced multiple times, such as when multiple B-cells express identical BCRs, the expected number of vertices comprising a cluster that could be due to sequencing error was estimated, given the experimentally derived PCR and sequencing error-rates (described in Section 2.12). All the samples have cluster sizes greater than that expected due to per-base error alone of 1.74×10^{-4} (Table 3.5), even at twice the measured error-rate (**Figure 3.13**). Therefore, the connectivity patterns of networks predominantly reveal differences in clonal expansions of B-cell populations rather than total sequencing errors. The clusters identified in BCR networks are derived from B-cells that share a common pro-B-cell progenitor with rearranged V-D-J that have subsequently expanded and diversified.
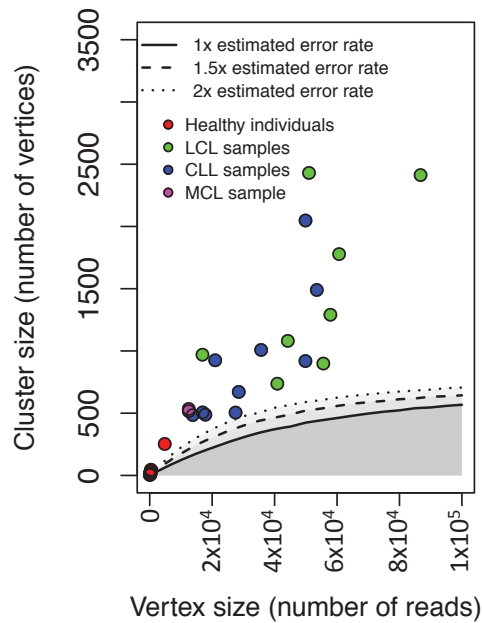
Figure 3.13. Assessment of error in BCR networks.

Comparison of the cluster sizes of the BCR networks with the expected cluster sizes that would result from sequencing and PCR error at three different error-rates. For a given BCR cDNA that is sequenced multiple times, the estimates of the number of vertices making up a cluster that may be due to sequencing error at a given PCR and sequencing error-rates, where the dotted, dashed and solid lines refer to 1x, 1.5x and 2x the experimentally determined non-homopolymeric 454 error rate of 7.04×10^{-5}, assuming the central BCR sequence is otherwise unconnected to any other vertices in the BCR network. The circles show the cluster sizes of the largest clusters in the networks for the different samples, where red, blue and green correspond to healthy individuals, CLL patients and LCL samples respectively.

3.2.9. BCR repertoire network parameters relate to CLL development

To assess the sensitivity of BCR sequencing using multiplex PCR amplification, the titration experiment from *Boyd et al.* (Boyd et al., 2009) in which serial 10-fold dilutions of a known clonal CLL PB sample into normal peripheral blood was used. 90.9% of all reads in the undiluted sample fall within the leukemic cluster (**Figure 3.14**A-B). Using these methods, the leukemic clonal sequences can be detected at dilutions as low as 1:100,000 when the sequence is known and pre-defined. (A MiSeq BCR dilution series was also performed in chapter 7 giving sensitivity of >1:10^7). When the leukemic cluster sequences are unknown, detection of expanded clones relies on detecting the maximum cluster size that is significantly different from that of healthy individuals. Significant increases in maximum cluster size were seen above that of the healthy individual in CLL dilutions of 1:100 or less.

The relationship between the BCR population measures and the CLL clinical information for each patient was next determined. Interestingly, there was a strong correlation between the length of time since CLL diagnosis with the vertex Gini Index (**Figure 3.15**A) and the maximum cluster size (**Figure 3.15**B). This suggests longer disease times lead to larger vertices representing larger tumor clonal populations, in agreement with previous studies (Hayes et al., 2010, Kelly et al., 2002).
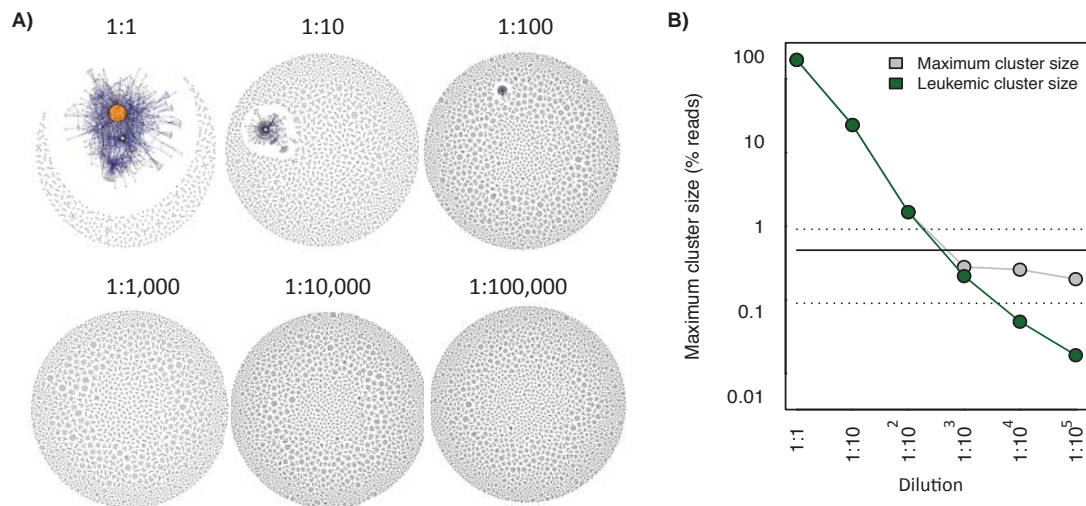
Figure 3.14. Variation of B-cell receptor populations.

A) B-cell receptor networks for the titration of a chronic lymphocytic leukemia clonal sample into healthy peripheral blood from the dataset from *Boyd et al. (2009)* and **B)** the corresponding number of reads corresponding to the leukemic clone (green) and the maximum cluster size of each dilution (grey). The solid horizontal line shows the mean maximum cluster size for healthy individuals from this dataset (0.52% of total reads), and the dashed horizontal lines show the mean +/- standard deviation of maximum cluster size for healthy individuals for this dataset.
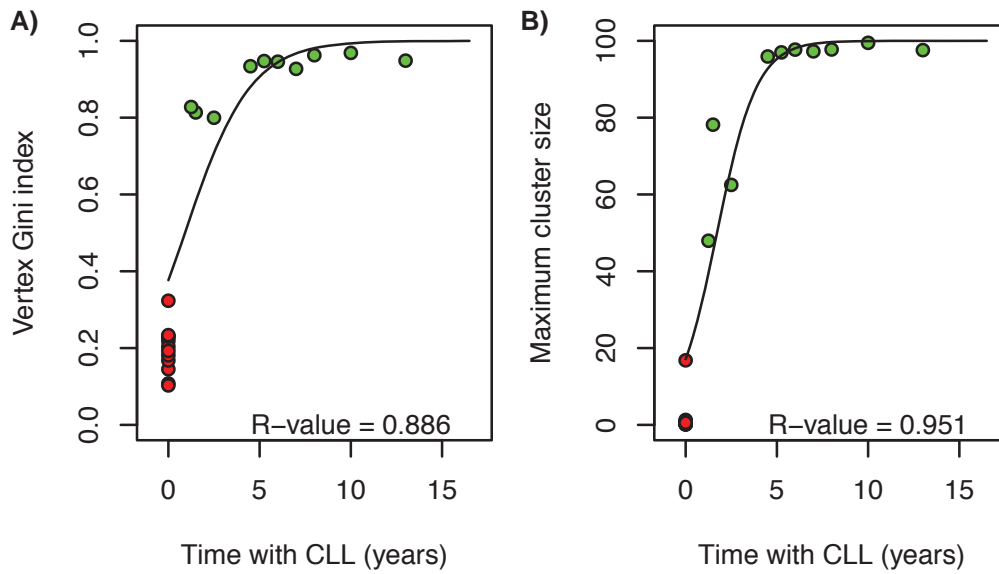
Figure 3.15. BCR diversity variation with time since CLL diagnosis.

Correlation between the **A)** vertex Gini Index and **B)** maximum cluster size with the length of time since chronic lymphocytic leukemia (CLL) diagnosis for each patient. The R^2-value for the logistic regression is given. The red and green circles correspond to healthy individuals and CLL patients respectively.

3.2.10. Following malignant B-cell clonal dynamics by BCR sequencing

It was hypothesised that BCR sequencing can be used to follow the dynamics of B-cell clones in samples taken multiple time points from patients. To test this, samples were taken from patients separated by a period of time in which the patient had either (a) not undergone any treatment prior to or during sampling (12 samples, denoted no treatment samples), (b) before and after a round of Chlorambucil treatment (5 samples, denoted during treatment samples) or (c) patients who had previously undergone Chlorambucil treatment, but not treatment was given during sampling (4 samples, denoted after treatment samples), summarised in Table 3.7 and **Figure 3.16**. The BCRs from these samples were amplified by multiplex PCR and sequenced by MiSeq sequencing.

Table 3.7. Filtered BCR depths for temporal CLL patient samples.

ID	Number of filtered BCR reads		Clinical condition	Stage	Treatment history	Patient ID	Time between samples (days)
	Time 1	Time 2					
C1	85858	81882	No treatment	A	Untreated	Pat. 1	182
C2	149400	154339	No treatment	C	Untreated	Pat. 5	255
C3	45354	113142	No treatment	A	Untreated	Pat. 6	168
C4	113142	144814	No treatment	A	Untreated	Pat. 6	91
C5	132877	127714	No treatment	A	Untreated	Pat. 8	56
C6	127714	109907	No treatment	A	Untreated	Pat. 8	63
C7	109907	112753	No treatment	A	Untreated	Pat. 8	92
C8	112753	121805	No treatment	A	Untreated	Pat. 8	125
C9	221127	138806	No treatment	B	Untreated	Pat. 9	57
C10	138806	120070	No treatment	B	Untreated	Pat. 9	84
C11	161873	123582	No treatment	A	Untreated	Pat. 11	151
C12	112722	87626	No treatment	A	Untreated	Pat. 13	245
C13	58446	241118	During treatment	A	Chlorambucil	Pat. 2	623
C14	99029	91332	During treatment	B	Chlorambucil	Pat. 4	28
C15	91332	149131	During treatment	B	Chlorambucil	Pat. 4	91
C16	138208	154211	During treatment	B	Chlorambucil	Pat. 7	28
C17	90864	98690	During treatment	C	Chlorambucil	Pat. 10	140
C18	146188	175261	After treatment	Atypical	Chlorambucil	Pat. 3	182
C19	149131	123046	After treatment	B	Chlorambucil	Pat. 4	98
C20	126241	151777	After treatment	C	Chlorambucil, rituximab	Pat. 12	163
C21	151777	164127	After treatment	C	Chlorambucil, rituximab	Pat. 12	152
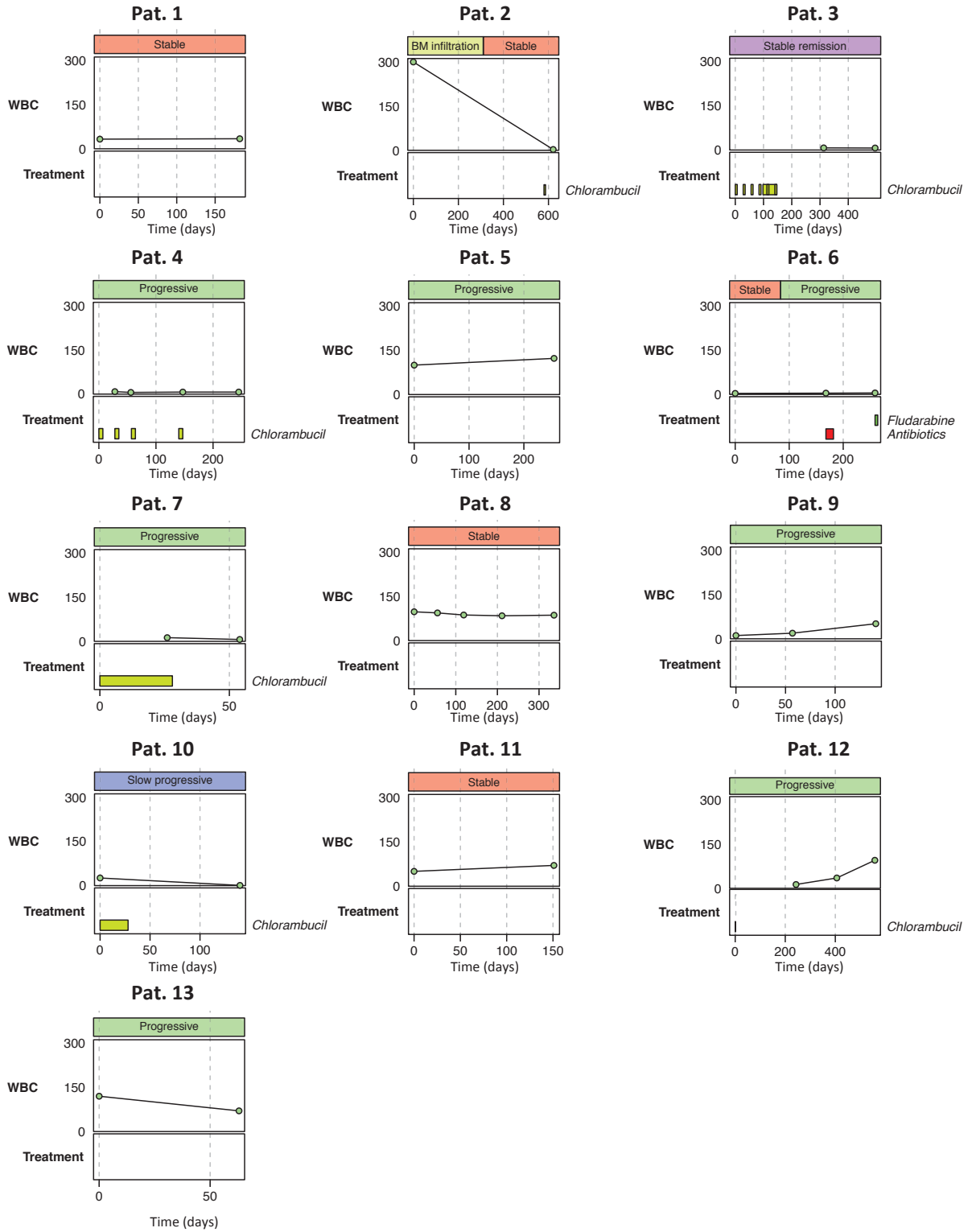
Figure 3.16. Treatment times and white blood cell count over time for temporal CLL samples.

Diagrams showing the relative sampling times with respect to treatment times and white blood cell count (WBC). In each panel, the top bar indicates the disease progression status of the patient at the corresponding time points, the middle section indicates the WBC at the sampled time points, which corresponded to the times that PBMC samples were taken for the BCR analyses, and the lower section indicated if and when treatment was given to the patients. Abbreviation: BM is bone marrow.

The largest (malignant) cluster was the same at each time point, where the higher frequency sequences are always retained between time points. However, the maximum (malignant) cluster size changes over the time intervals were distinct between clinical groups (range 2.52-95.77% of total sequences, **Figure 3.17**Ai). Notably, only the malignant clusters in samples taken during treatment significantly decreased in size over the time interval of sampling (p-value<0.005), whereas the no treatment and after treatment samples did not significantly increase or decrease in cluster sizes. The maximum cluster sizes from samples taken during treatment were significantly reduced compared to the no treatment samples (p-value=0.00445, **Figure 3.17**Aii). This reflects the white blood cell count (WBC) for these patients, where the WBC was significant reduced during treatment compared to the no treatment samples (p-value=0.00812, **Figure 3.17**B). Interestingly, the after therapy samples exhibited a mixed response to therapy, where the change in maximum cluster sizes range from a reduction by 10.14% to an increase by 10.60%, reflecting the change in WBC, which ranged from a reduction of -0.4×10^9 cells/L to an increase of 60.2×10^9 cells/L.

In addition, the vertex and cluster Gini indices derived in Section 3.2.6 that describe the B-cell clonalities of the samples were shown to change in a similar fashion (**Figure 3.18**). In patients that were undergoing active treatment, the vertex Gini indices significantly decreased over time, suggesting that the overall clonality of these patients were decreasing (p-value<0.005, **Figure 3.18**Ai). However, in patients that were not undergoing active treatment, the vertex Gini indices did not significantly increase or decrease (p-values>0.0759, **Figure 3.18**Ai) suggesting stable overall clonality in these patients. In fact, the vertex Gini indices from samples taken during treatment were significantly reduced compared to the no treatment samples (p-value=1.71×10^{-5}, **Figure 3.18**Aii), suggesting significant changes in the overall B-cell population clonality during treatment. For samples taken after treatment (i.e. no active treatment given between sampling, p-value = 0.288, **Figure 3.18**Ai), there was no significant increase or decrease in vertex Gini index, and the changes were not significantly different from that of the no treatment samples (p-value=0.813, **Figure 3.18**Aii). This suggests that when active treatment is discontinued in CLL patients, the overall PB B-cell clonality remains stable.

The cluster Gini index indicates the overall sample SHM, where an increase in the cluster index means a higher unevenness of the number of unique BCRs in

between the clusters in a sample. Therefore, the significantly increase in the cluster Gini indices for the patients who had never undergone therapy (p-value<0.005, **Figure 3.18**Bi), indicating that there is significant diversification in these patients even though the CLL clone is not significantly enlarging, as indicated by **Figure 3.17**Ai and **Figure 3.18**Ai. During treatment, the cluster Gini index typically reduces (p-value=0.0177, **Figure 3.18**Bi), most likely as a result of reducing the cluster size. However, this does not reach significance, therefore may be indicative of clonal diversification even when the clone is actively being reduced in size. However, the change in cluster Gini indices from samples taken during treatment were significantly reduced compared to the no treatment samples (p-value=0.0077, **Figure 3.18**Bii), suggesting that therapy significantly reduces CLL clonal. After therapy, 3/4 of the patients had stable cluster Gini indices, and only a single patient had 2-fold increase in cluster Gini index (**Figure 3.18**Bi), corresponding to an increase in WBC from 12.9×10^9cells/L to 35.1×10^9cells/L, thus further confirming a mixed post-therapy CLL response.

Together these data can be interpreted as, in the absence of treatment, the clone sizes are stable in frequency and undergo CLL clonal diversification. However, during active Chlorambucil treatment, the dominant CLL clone reduces in frequency, with suppressed diversification. This means that Chlorambucil treatment not only reduces the WBC, but also the proportion of the white blood cell population consisting of CLL cells. Once therapy is removed, there appears to be a mixed outcome, where some patients retain a stably low WBC and clonality, whereas others exhibit re-expansion of the CLL clone.
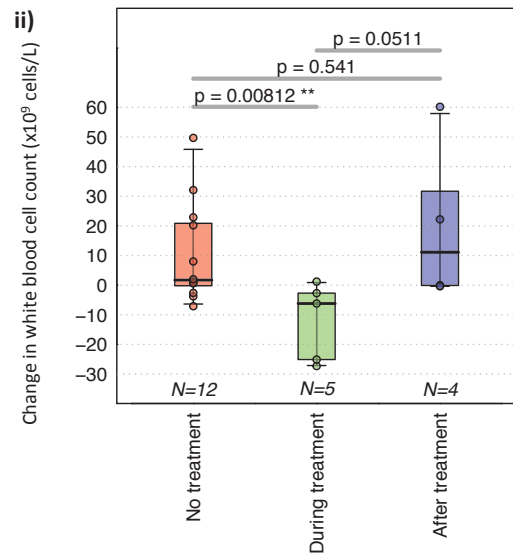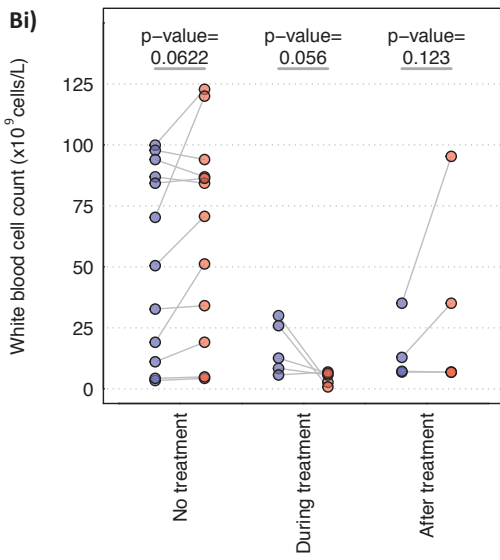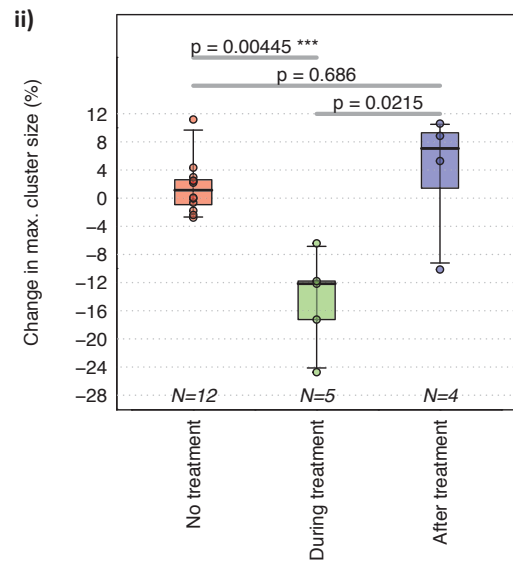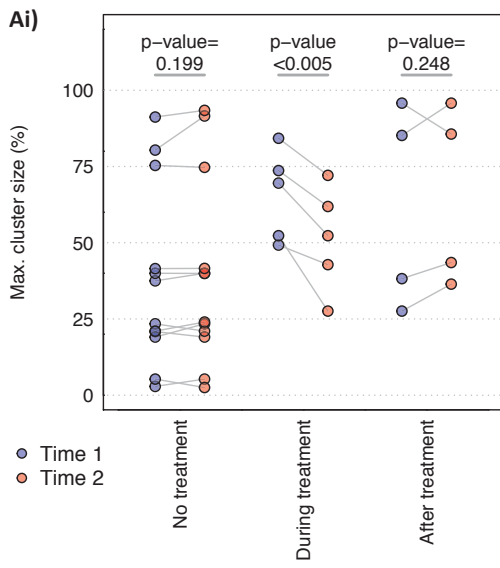
Figure 3.17. Dynamics of CLL BCR repertoires and white blood cell counts.

Samples were taken from patients separated by a period of time in which the patient had either (a) not undergone any treatment prior to or during sampling (12 samples, denoted no treatment samples), (b) before and after a round of Chlorambucil treatment (5 samples, denoted during treatment samples) or (c) patients who had previously undergone Chlorambucil treatment, but not treatment was given during sampling (4 patients, denoted after treatment samples). **Ai)** The plot of the maximum (malignant) B-cell cluster sizes for first and second samples taken, where the blue and red points represent the first and second samples respectively. The grey lines join together adjacent samples from the same patient, and two-side paired t-test p-values between the first and second samples are given above each group. **ii)** Boxplots of the changes in the maximum cluster sizes between the first and second samples for each patient, where the p-values of the significance of the changes between groups is given above, calculated by two-sided unpaired t-tests. **Bi)** The plot of the white blood cell counts (WBCs) for first and second samples taken, where the blue and red points represent the first and second samples respectively. The grey lines join together adjacent samples from the same patient, and two-side paired t-test p-values between the first and second samples are given above each group. **ii)** Boxplots of the changes in the WBCs between the first and second samples for each patient, where the p-values of the significance of the changes between groups is given above, calculated by two-sided unpaired t-tests.
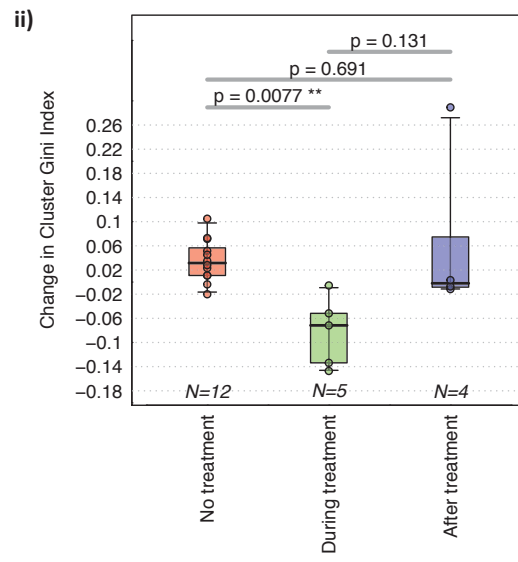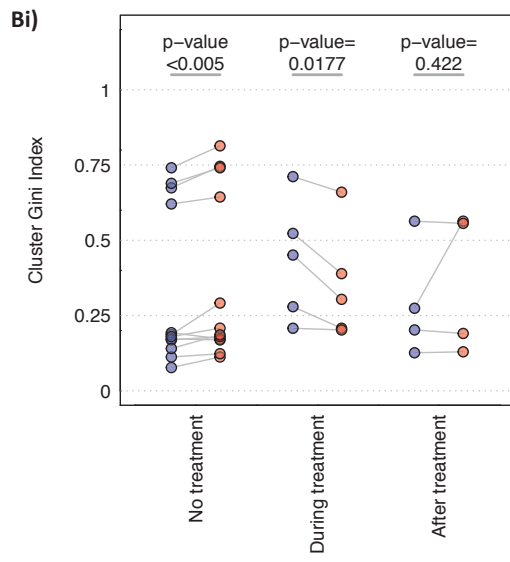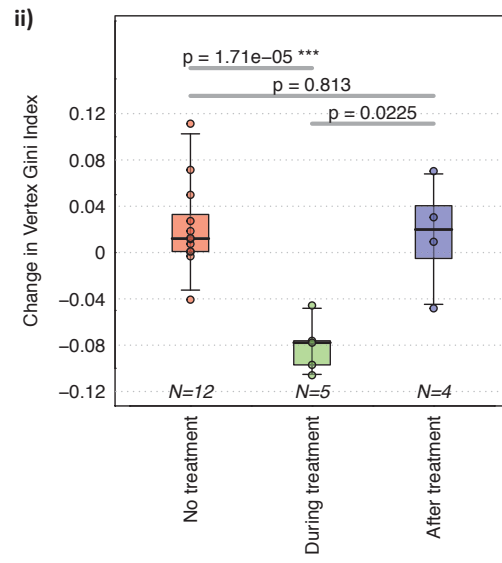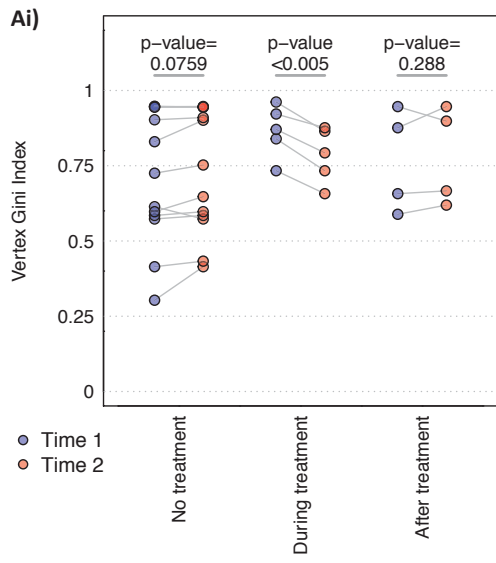
Figure 3.18. Dynamics of CLL BCR repertoires properties.

Samples were taken from patients separated by a period of time as in **Figure 3.17**. **Ai)** The plot of the vertex Gini indices for first and second samples taken, where the blue and red points represent the first and second samples respectively. The grey lines join together adjacent samples from the same patient, and two-side paired t-test p-values between the first and second samples are given above each group. **ii)** Boxplots of the changes in the vertex Gini indices between the first and second samples for each patient, where the p-values of the significance of the changes between groups is given above, calculated by two-sided unpaired t-tests. **Bi)** The plot of the cluster Gini indices for first and second samples taken, where the blue and red points represent the first and second samples respectively. The grey lines join together adjacent samples from the same patient, and two-side paired t-test p-values between the first and second samples are given above each group. **ii)** Boxplots of the changes in the cluster Gini indices between the first and second samples for each patient, where the p-values of the significance of the changes between groups is given above, calculated by two-sided unpaired t-tests.

3.2.11. Phylogenetic analysis of B-cell clones

Clonal evolution in CLL as exemplified by the presence of mutations in the genome and by multiple BCRs related to the dominant CLL BCR sequence (Landau et al., 2013, Schuh et al., 2012). Mutations in the BCR may be used to infer the mutational route from a CLL B-cell ancestor to the rest of the leukaemic clone by phylogenetic analysis. Phylogenetic analysis can be used to reconstruct the evolutionary history of organisms (Pybus et al., 2002). However, to date, no B-cell specific evolutionary model of BCR diversification have been developed, hampered primarily by (a) a BCR evolutionary tree is not strictly bifurcating due to the expansion of multiple B-cells with identical BCRs, that can each independently diversify, (b) non-constant mutation rate, dependent on co-stimulation from multiple sources, such as T-cell activation, and (c) ongoing or secondary rearrangements can lead to the replacement of IgHV gene segment while retaining the same IgHD-J region, thus leading to different evolutionary histories in different regions in the BCR (Marshall et al., 1995, Steenbergen et al., 1993, Gawad et al., 2012, Choi et al., 1996).

However, despite these drawbacks, the phylogenetic relationships between BCR sequences can inform about the process of B-cell clonal diversification, such as by the tree shape, such as star shapes of each of phylogenetic trees suggest unselected clonal expansion, compared to antigenic drift seen in, for example, influenza virus (Steinbruck and McHardy, 2012, Bedford et al., 2014). Therefore, using the patient samples in section 3.2.10, all the sequences from the dominant clusters were extracted, to determine the maximum parsimony phylogenetic tree structures of the leukaemic clone, and to infer the process of diversification. Maximum parsimony was chosen as the phylogenetic model as this imposes the fewest number of explicit assumptions on the data. For each patient, all the BCR sequences related to the CLL clone were aligned using Mafft (Katoh and Standley, 2013) and a maximum parsimony tree was fitted using Paup* (Wilgenbusch and Swofford, 2003). The branch lengths represent the evolutionary distance between BCR sequences and bootstrapping was performed to evaluate the reproducibility of the trees, showing strong tree support (>95% certainty for all branches). The majority of the trees from patients have a star-like structure (**Figure 3.19**), suggesting that the CLL clone emerged from a single common ancestor (Martins and Housworth, 2002), represented by the central BCR, which was the most frequently observed BCR. The concentric

rings of BCR variants represent incremental increases in base pair differences from the central dominant BCR sequence. However, the phylogenetic trees in patients 6, 10 and 13 show small outgrowths, suggesting potential growth and diversification advantages in the B-cells corresponding to these branches, potentially reflecting genomic variations in these B-cells.

Patient 3
Cluster size: 27.78% of reads

Patient 2
Cluster size: 49.00% of reads

Patient 1
Cluster size: 81.57% of reads

Patient 6
Cluster size: 77.51% of reads

Patient 5
Cluster size: 93.15% of reads

Patient 4
Cluster size: 70.51% of reads

Patient 9
Cluster size: 38.04% of reads

Patient 12
Cluster size: 86.41% of reads

Patient 8
Cluster size: 21.14% of reads

Patient 11
Cluster size: 42.03% of reads

Patient 7
Cluster size: 85.29% of reads

Patient 10
Cluster size: 69.78% of reads

Patient 13
Cluster size: 77.5% of reads
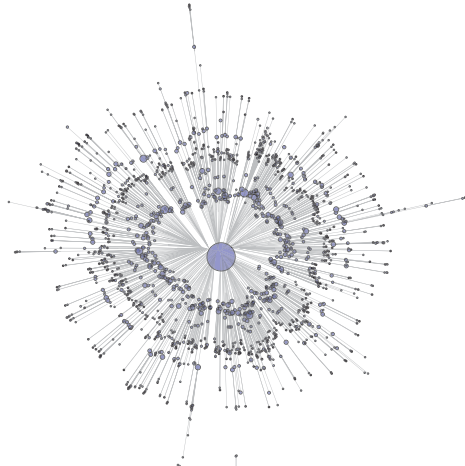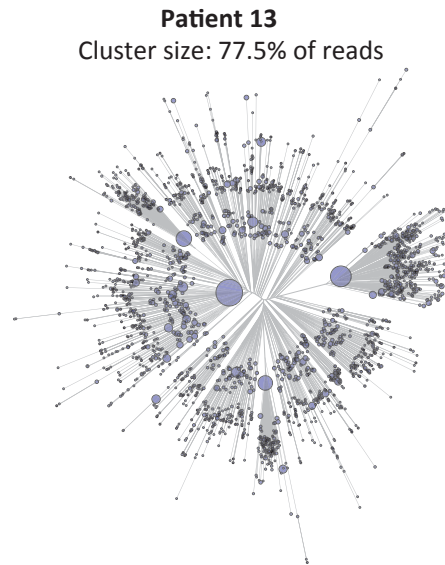
Figure 3.19. Unrooted maximum parsimony trees of the malignant CLL clusters.

For each patient, all sequences in the maximum cluster were aligned using Mafft (Katoh and Standley, 2013) and a maximum parsimony tree was fitted using Paup* (Wilgenbusch and Swofford, 2003). The branch lengths represent the evolutionary distance between BCR sequences and bootstrapping was performed to evaluate the reproducibility of the trees, showing strong tree support (>95% certainty for all branches). The branch lengths are proportional to the number of varying bases (evolutionary distance) and the tips represent unique BCR sequences within the malignant CLL cluster. The sizes of the tips are proportional to the number of sequencing reads corresponding to the BCR.

3.3. Conclusions

The aim of this chapter is to discriminate between healthy and malignant B-cell expansions through BCR repertoire sequencing. To do this, methods must be robust to noise, such as PCR and sequencing error, as well as sequencing depth. The effects of amplification and sequencing error are often of concern for BCR deep sequencing. However, the strong linear correlations of the network parameters between samples that have been amplified using independent primer sets suggest limited amplification bias and no significant effect on the overall population structure. In all samples tested from healthy and haematological cancer patients, the cluster sizes are notably greater than that expected due to the process error alone, suggesting that the network structures represent the population structures of the B-cell sample.

The observation of frequent multiple identical BCR sequences in malignant B-cell samples and only low frequency identical BCR sequences from healthy individuals suggests that multiple identical RNA molecules from a single B-cell are rarely sequenced. Therefore, clusters of related sequences are likely to represent BCRs from clonal expansions of evolutionarily related B-cells, whereas naïve B-cell populations form singletons in sparsely connected networks. The probabilities of resampling BCRs from a population is revisited in more detail in Chapter 4.

If the B-cell network from limited sequencing is a random sample of the entire circulating peripheral blood BCR repertoire, then a scale invariant diversity measure should also capture the predominant structure of the unsampled network. Here, it has been shown that network structures, combined with these population measures discriminate between B-cell repertoires of different clonalities in health and disease. These measures are robust to variations in sequencing and sampling depth and different filtering strategies and are applicable to independently produced datasets (Boyd et al., 2009). Using different primer sets, sequencing depths and sequencing technologies, the samples still cluster according to the clonal nature of the samples, occupying the equivalent distinct regions of Gini Index and maximum/second maximum graphs. Therefore this analytical strategy is applicable to any BCR deep sequencing technology.

Deep sequencing of BCR repertoires potentially allows the detection of a clonal lymphoid population in a background of polyclonal cells without prior knowledge of the leukemic sequence (Sayala et al., 2007). Here, the limit of *de novo*

detection of malignant clonality is at least 1 in 100 dilution of CLL cells into healthy blood. In addition, the vertex Gini Index is strongly correlated with the time an individual has been living with CLL. This has potential applications in the detection of clonal B-cell disorders and malignancies, particularly as the early stages of these diseases are asymptomatic, such as in CLL. When there is prior knowledge of a BCR of interest, such as in leukaemia, the limit of detection is much greater (>1 in 10^5 cells). In practice, this has important potential uses in monitoring disease during therapy (addressed in Chapter 4) and minimal residual disease detection (addressed in Chapter 5).

An important result of this framework to assess B-cell repertoire structure is to understand the changes involved in a healthy immune repertoire, such as during vaccination, compared to malignant B-cell expansion. There was variation between the network-based diversity measures of a "normal" BCR repertoires between the healthy individuals, where a larger-scaled assessment of the primary immune response compared to early stage leukaemia could provide clinically important early diagnostic or prognostic information to patients. For example, one healthy individual (healthy individual 10) exhibited a more clonal BCR repertoire compared to the other healthy individuals, defined by an increase in connectivity. Further work could be performed to determine the likelihood of this clonality resulting from an antigen specific memory B-cell expansion or an undiagnosed malignant transformation in an otherwise asymptomatic individual.

Similarly, the presence of more than one BCR clonal expansion in CLL and other blood cancers has unknown clinical implications. These enlarged clusters representing BCRs with different V-D-J gene combinations may be due to either the expansion of two distinct malignant B-cell transformations, or separate antigen-stimulated B-cell clonal expansion unrelated to CLL. These methods used in time-series may allow the distinction between antigen-driven positive selections in CDRs compared to malignant-driven expansion.

B-cells form dynamic populations of cells. Here it is shown that these populations expand and potentially evolve over time. For the first time it is possible to observe the specifics of a short-term effect of therapy on the B-cell repertoire in CLL, and demonstrates how networks lend themselves to phylogenetic approaches. During therapy, there was a significant reduction in B-cell clonality and the percentage of BCRs relating to the malignant B-cell cluster. Work here therefore provides a

framework for analysing deep high-throughput BCR sequencing datasets to probe B-cell population changes between serial samples or individuals.

Chapter 4

4. Comparison of BCR amplification and sequencing methods

4.1. Introduction

For immune repertoire sequencing to be useful, it is therefore vital that sample preparation and sequencing approaches give reproducible, unbiased and sensitive representations of BCR repertoires. However, there is concern over the validity and biases of biological insights gained from the different BCR and TCR enrichment, amplification and sequencing methods, particularly whether the sequencing data truly represents the corresponding B-cell populations. As the B-cell receptor is highly diversified, there is potential for some immunoglobulin rearrangements to be preferentially captured and amplified, leading to biased sequencing data.

This chapter integrates both the theoretical and experimental frameworks for BCR sequencing to determine whether the B-cell sequencing data represents that expected theoretically. Then, the utilities, biases and reproducibilities of different sequencing depths, sequencing technologies, amplification methods, read lengths and starting material are assessed using samples of diverse B-cell populations from healthy peripheral blood (PB), clonal B-cell populations from lymphoblastoid cell lines (LCL) and PB from chronic lymphocytic leukaemic (CLL) patients.

4.2. Results

4.2.1. Generation of BCR sequencing datasets for comparative studies

Experimental BCR sequencing datasets were generated through the amplification of LCLs and PB B-cell BCRs from healthy individuals, and CLL patients by the three main BCR amplification methods; multiplex PCR, 5' Rapid amplification of cDNA ends (5'RACE), and RNA-capture, and sequenced by 454 Roche and Illumina MiSeq (summarised in Table 4.1). Each sample generated an average of 40,763 reads (summarised in Table 4.2). For each sample, reads were filtered for immunoglobulin similarity and length, and, where relevant, primer sequences were removed according to Methods (Section 2.2.4). IgHV classifications were performed on each BCR sequence by determining the best alignment to the ImMunoGeneTics (IMGT) database (Lefranc et al., 2009) using BLAST (Altschul et

al., 1990). For each sample, the reference IgHV gene frequencies and clonality measures developed in Chapter 3, namely the vertex and cluster Gini indices and maximum cluster sizes, were determined for the comparisons in this chapter. These diversity measures correspond to that seen in equivalent sample types in previous studies (Table 4.3, (Bashford-Rogers et al., 2013)).

Table 4.1. Samples used in this study for each amplification method.

Sample type*	ID	Multiplex (454)	Multiplex (MiSeq)	5' RACE (MiSeq)	RNA capture (MiSeq)
CLL	Sample 1	Y	Y	Y	Y
CLL	Sample 2	Y	Y	Y	-
CLL	Sample 3	Y	Y	-	-
CLL	Sample 4	Y	Y	Y	-
CLL	Sample 5	Y	Y	Y	-
CLL	Sample 6	Y	Y	Y	-
CLL	Sample 7	Y	Y	Y	-
CLL	Sample 8	Y	Y	Y	-
Healthy	Sample A	Y	-	Y	-
Healthy	Sample B	Y	-	Y	-
Healthy	Sample C	Y	Y	Y	-
Healthy	Sample D	Y	-	Y	-
Healthy	Sample E	Y	Y	Y	-
Healthy	Sample F	Y	Y	-	-
Healthy	Sample G	Y	Y	-	-
Healthy	Sample H	Y	Y	-	-
Healthy	Sample I	Y	Y	-	Y
LCL	LCL 1	Y	-	-	-
LCL	LCL 2	Y	-	-	-
LCL	LCL 3	Y	-	-	-
LCL	LCL 4	Y	-	-	-
LCL	LCL 5	Y	-	-	-
LCL	LCL 6	Y	-	-	-
LCL	LCL 7	Y	-	-	-
LCL	LCL 8	Y	-	-	-
LCL	LCL 9	Y	-	-	-
LCL	LCL 10	Y	-	-	-

* Abbreviations: CLL = chronic lymphocytic leukaemia, healthy = PBMC from healthy blood donor, LCL = human lymphoblastoid cell line.

Table 4.2. Mean and standard deviation of read depths per sample.

Technology	Mean read depth per sample (after filtering)	Average number of multiplexed samples per lane/run	Average % BCR sequences after filtering*
Multiplex PCR (454)	33,413	12	76.10
Multiplex PCR (MiSeq)	31,118	50	60.30
5' RACE (MiSeq)	72,586	95	55.09
RNA capture (MiSeq)	58,015	2	1.53**

* Percentage of reads after filtering for open reading frames rearranged BCR sequences from the whole read set.

** RNA capture has lower percentage of filtered reads due to the designed simultaneous capture of immunoglobulin heavy and light chains as well as T-cell receptors.

Table 4.3. Mean diversity measures for each sample type.

Sample type	Mean maximum cluster size (% of total BCR sequences)	Mean vertex Gini Index	Mean cluster Gini Index
Healthy	0.581	0.182	0.047
Chronic lymphocytic leukaemia (CLL)	95.117	0.931	0.612
Human lymphoblastoid cell line (LCL)	65.205	0.934	0.790

4.2.2. Theoretical framework for sampling and sequencing BCR repertoires

As exhaustive sampling of total B-cells is not possible in humans, the "true" extent of the total BCR repertoire in humans can only be estimated. To understand the BCR sequencing data properly, it is first important to estimate the types of B-cells sampled, and the proportion of total B-cells from a patient in each sample. This will give a theoretical estimate for the expected percentage of BCRs to be shared between technical repeats and the expected sampling stochasticity in any given BCR sample, which will then be tested experimentally.

A typical PB sample (10-20ml) accounts for ~0.4% of the total PB (average of 5L of blood in a healthy adult), from which only a fraction is used in current BCR sequencing methods with approximately 0.012% of all B-cells being represented in the material that is sequenced, Table 4.4. The healthy peripheral blood B-cell population contains approximately 80% naïve B-cells and 20% memory B-cells (Tangye and Good, 2007). As naïve B-cells are antigen inexperienced, each naïve B-cell BCR is often considered to be unique. This means that sequencing BCRs from only naïve B-cells theoretically result in a diverse BCR population, with all BCRs represented with equal probability. Therefore the distribution of BCR frequencies should follow approximately a binomial distribution, where parameters depend on the number of RNA molecules per cell, the number of B-cells represented and efficiencies of the RT-PCR, PCR and sequencing steps. Under the assumption that each naïve B-cell is unique, it therefore is theoretically impossible to resample identical BCRs from a population of naïve cells. The memory B-cell population, however, consists of cells that have undergone proliferation and potentially somatic hypermutation. This means that it is possible to sample multiple memory B-cells exhibiting identical BCR sequences or highly related BCRs after somatic mutation that originate from the same pre-B-cell.

Table 4.4. Estimation of number and percentage of sampled peripheral blood B-cells.

	Number of cells	% of total B-cell repertoire	Notes
Total B-cells in blood*	1,500,000,000	100	Average adult has 5L of blood
B-cells in sample	3,000,000	0.4	10 ml blood sample
RNA extraction	3,000,000	0.4	Assume 100 % efficiency
RT-PCR**	300,000	0.04	Average 10ug RNA extracted per sample, 1000ng RNA used in RT-PCR
PCR	90,000	0.012	6ul out of 20ul of RT-product used in PCR

* (www.stemcell.com)

** Average of 10ug RNA extracted per healthy PB sample.

To achieve a theoretical estimation of the percentage of BCRs that should overlap between any two samples from the same peripheral blood aliquot from a single individual, simulations were generated as follows: the total B-cells in the sample is N, from which each RT-PCR sample taken contains n B-cells equivalent. The proportion of memory B-cells in the peripheral blood is given by p_m, and the proportion of naïve B-cells is $1- p_m$ (assuming no plasma B-cell in the blood). Therefore the number of memory B-cells in the total population is $N* p_m$. The number of memory B-cells per unique BCR can be modelled as a normal distribution $N(\mu, \sigma)$, with a mean μ and standard deviation σ. Each simulation draws a random sample of $N* p_m$ BCRs where the probability of resampling a specific BCR follows $N(\mu, \sigma)$. From this, two random samples from this simulated total B-cell populations are drawn, and the percentage overlap between the samples is determined.

Three such simulations were generated, where the parameters are summarised in Table 4.5. All three simulations begin with the estimation of total and sampled B-cells from Table 4.4 and **Figure 4.1**. Simulations A and B assume that the proportion of memory B-cells is 20%, whereas simulation C models the proportion of memory B-cells as a normal distribution with mean of 20% and standard deviation of 8% to reflect inter-individual differences in the memory-to-naïve B-cell ratios (Tangye and Good, 2007). The percentage of overlapping BCRs between samples in simulations A, B and C determined for 1000 simulation repeats was 6.185%, 18.29%, and 19.71% respectively (**Figure 4.2**, blue box plots). The lower overlap in simulation A is

explained by the lower number of memory B-cells per BCR, therefore a lower probability of re-sampling B-cells with the same BCR. The higher variance of overlapping BCR percentages in simulation C is explained by the higher variance of percentage of memory B-cells in the PB.

Table 4.5. B-cell sampling simulation parameters.

Parameter	Simulation A	Simulation B	Simulation C
N	3000000	3000000	3000000
n	90000	90000	90000
p_m	0.2	0.2	$N(20, 8)$
μ	1	1	1
σ	10	100	100
Number of simulation repeats	1000	1000	1000
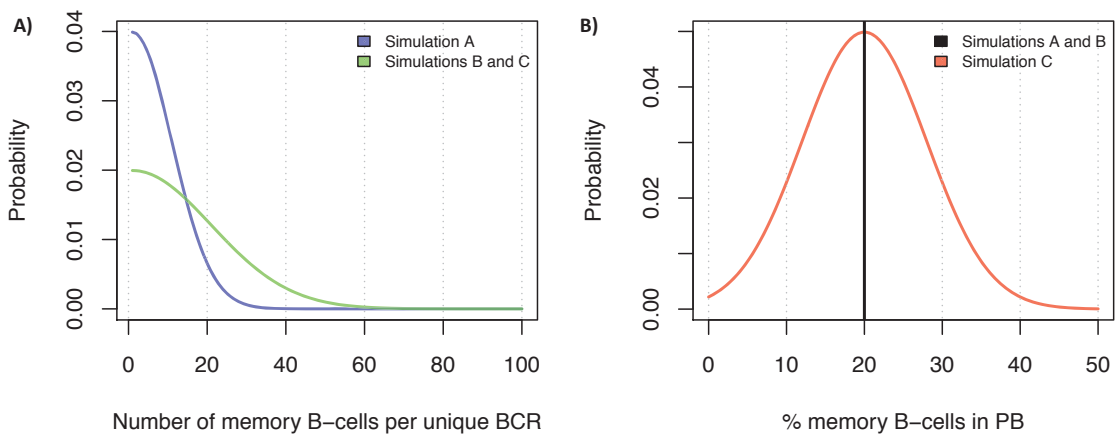
Figure 4.1. Simulation distributions.

Simulation distributions for **A)** the number of memory B-cells per unique BCR sequence (varying σ) and **B)** the percentage of memory B-cells in the peripheral blood (PB) (varying p_m). The colours of the distributions correspond to the simulations indicated in the key.

Chapter 5

5. Minimal residual disease in B-acute lymphoblastic leukaemia

5.1. Introduction

Many of the therapies for ALL cause significant toxicities and carry the potential of long-term complications including secondary malignancies. Additionally, the majority of treatment failures occur as a result of disease relapse occurring either during or after completion of treatment. Therefore, improved detection and monitoring of minimal residual disease in B-cell ALL (B-ALL) is of great clinical importance, particularly for tailoring therapeutic dosing and strategies (Biondi and Masera, 1998). Here, the BCR repertoire was sequenced in a set of B-ALL patients to determine whether BCR sequencing can be used to (a) monitor B-ALL residual disease load and (b) decipher the ontogeny and B-cell population dynamics in relapse patients?

5.2. Results

5.2.1. BCR sequencing of longitudinal samples from B-ALL patients

Longitudinal samples from six B-ALL patients over the course of therapy were analysed for the presence of residual leukaemic cells by both a routine clinical MRD monitoring method of quantifying qPCR transcript levels of fusion genes associated with individual leukaemias (performed by the molecular diagnostic laboratory of the Karaiskakio Foundation), and also by sequencing the BCR repertoire and mining leukaemia-specific BCR sequences. For each patient, a "primary sample" was studied with high leukemic load, as indicated by a qPCR T/C transcript (T/C) ratio greater than 1.66; a ratio which was reduced to zero in subsequent samples taken over the course of therapy (summarised in Table 5.1). Additionally, BCR sequencing was performed on peripheral blood samples from 18 healthy individuals within the range of 20-75 years of age. BCR sequencing yielded 124,302 to 2,972,494 filtered BCR sequences per sample (Table 5.1). BCR network analysis was applied to the sequencing datasets, to identify clusters representing groups of highly related BCR sequences (Bashford-Rogers et al., 2013). Clonality was observed in all B-ALL primary samples, as indicated by largest cluster sizes greater than 2.796% of the total

BCR repertoire. In comparison, the largest cluster sizes from the 18 healthy individuals averaged 0.618% (standard deviation of 0.641%, range 0.14-2.5%).

Table 5.1. B-ALL patient sample information.

Patient ID	Sample ID	Time since first sample (days)	Target/ control transcript ratio	Total BCR sequences in sample	Target transcript type**	Sample source*	Largest cluster (% of BCR sequences)
527	527_A	0	13.95	124,302	E2A-PBX1	BM	43.733
527	527_B	8	0.02	270,572	E2A-PBX1	BM	0.696
527	527_C	15	0.00	756,674	E2A-PBX1	BM	0.320
527	527_D	30	0.00	698,592	E2A-PBX1	BM	0.097
527	527_E	109	0.00	2,320,485	E2A-PBX1	BM	6.818
527	527_F	889	0.00	2,301,914	E2A-PBX1	BM	0.580
859	859_A	0	1.66	454,071	TEL-AML1	PB	2.796
859	859_B	7	0.03	786,283	TEL-AML1	BM	0.179
859	859_C	84	0.00	737,736	TEL-AML1	BM	0.738
859	859_D	374	0.00	1,929,858	TEL-AML1	BM	2.159
859	859_E	1241	0.00	2,025,955	TEL-AML1	BM	0.219
1592	1592_A	0	34.60	259,439	E2A-PBX1	BM	26.600
1592	1592_B	12	12.98	264,698	E2A-PBX1	BM	26.105
1592	1592_C	33	0.02	216,356	E2A-PBX1	BM	0.192
1592	1592_D	554	0.00	129,923	E2A-PBX1	PB	1.040
1611	1611_A	0	35.04	189,634	E2A-PBX1	BM	27.843
1611	1611_B	12	0.00	264,128	E2A-PBX1	BM	0.448
1611	1611_D	510	0.00	284,526	E2A-PBX1	BM	0.175
1611	1611_F	944	0.00	346,134	E2A-PBX1	PB	1.751
1703	1703_A	0	0.12	2,972,494	TEL-AML1	PB	0.390
1703	1703_B	18	0.00	2,209,688	TEL-AML1	BM	1.049
1703	1703_C	336	0.00	1,861,228	TEL-AML1	BM	0.131
1703	1703_D	567	0.00	1,475,750	TEL-AML1	BM	0.353
1703	1703_E	567	3.12	1,237,270	TEL-AML1	CSF	3.833
3243	3243_A	0	1.75	297,165	BCR-ABL	BM	10.196
3243	3243_B	20	0.02	372,194	BCR-ABL	BM	0.194
3243	3243_C	31	0.01	340,706	BCR-ABL	BM	0.331
3243	3243_D	56	0.00	315,718	BCR-ABL	BM	0.320
3243	3243_E	91	0.00	319,850	BCR-ABL	BM	0.451

Samples highlighted in orange denote the primary samples for each patient.

* Abbreviations: BM is bone marrow, PB is peripheral blood and CSF is cerebrospinal fluid.

** E2A-PBX1: gene fusion between the transcription factor *E2A* with the homeodomain protein *PBX1*. TEL-AML1: gene fusion between the transcription factor *TEL* with the transcription factor *AML1*. BCR-ABL: gene fusion between the "breakpoint cluster region" (*BCR*), a 5.8 kbp region of DNA on chromosome 22 (22q11), with the tyrosine kinase *ABL1*.

5.2.2. Comparison of ALL and CLL repertoires

B-ALL is thought to arise from a malignant transformation of immature hematopoietic progenitor at one of several stages of early B-cell development. The BCR repertoire in B-ALL has been shown to be distinct from that of later-stage B-cell leukaemias, such as chronic lymphocytic leukaemia (CLL), where the preferential IgHV-J gene usage in B-ALL has been shown to reflect early B-cell repertoires in B-ALL (Duke et al., 2003). To determine whether B-ALL B-cells have undergone less maturation and diversification than both the mutated and unmutated subtypes of CLL, we compared the BCR clusters in the B-ALL samples to those seen in 9 CLL patients from Chapter 3.

Firstly, the mutational distance of the dominant leukaemic BCR sequences in each CLL and B-ALL patient were compared to confirm that the B-ALL sequences represent an earlier stage of B-cell development than CLL. For each patient the dominant BCR sequence of the B-ALL or CLL cluster was aligned to the IMGT reference database and the percentage sequence identity to reference IgHV genes was determined using IgBLAST (Ye et al., 2013) (**Figure 5.1**A). The dominant BCR sequences in each B-ALL patient were either identical or within 3bp from a reference germline sequence (mean 99.52% of nucleotides identical to reference), supporting the hypothesis that B-ALL arises from B-cells that have not undergone somatic hypermutation (SHM). The sequences that were not identical to the reference germline sequences may be accounted for by allelic variation of the IgHV locus not present in the reference BLAST database. CLL can be defined into two subtypes, where two different mutational statuses of CLL patients are thought to be derived from two different stages of B-cell ontology, with the unmutated CLL cases corresponding to pre-antigenic stimulation, and the mutated cases corresponding to post-antigenic stimulation (Hamblin et al., 1999, Damle et al., 1999). Therefore, the CLL patients were subgrouped into unmutated (where the dominant BCR had >98% sequence similarity with reference germline IgHV-D-J sequences) or mutated CLL (dominant BCR <98% sequence similarity with reference germline IgHV-D-J sequences). The unmutated subtype CLL patients exhibited no significant difference in sequence similarity to the reference IgHV BLAST database (mean 98.9% identical to reference, p-value=0.478), whereas the mutated subtype CLL patients had

significantly lower sequence similarity to the reference IgHV BLAST database (mean 92.5% identical to reference).

Secondly, the diversification of the malignant clusters in B-ALL and CLL were compared to determine whether B-ALL has a lower propensity to diversify than CLL. For each patient, all BCR sequences within the malignant cluster were aligned and for the sequences not identical to the dominant vertex of the cluster, the numbers of mutations away from this highest-observed BCR sequence were determined (**Figure 5.1**B). The B-ALL malignant clusters showed a lower mutational distances from the dominant BCR sequence than CLL (means distances of 2.017bp, 2.277 and 2.694bp for B-ALL unmutated CLL and mutated CLL respectively, p-values<0.005), suggesting lower levels of SHM within the B-ALL B-cell population compared to both CLL mutational subtypes. Together, this data supports the idea that B-ALL arises from earlier stages of B-cell differentiation than the mutated CLL subtype, and as such displays lower, albeit detectable, levels of SHM and clonal diversification than both CLL mutational subtypes.
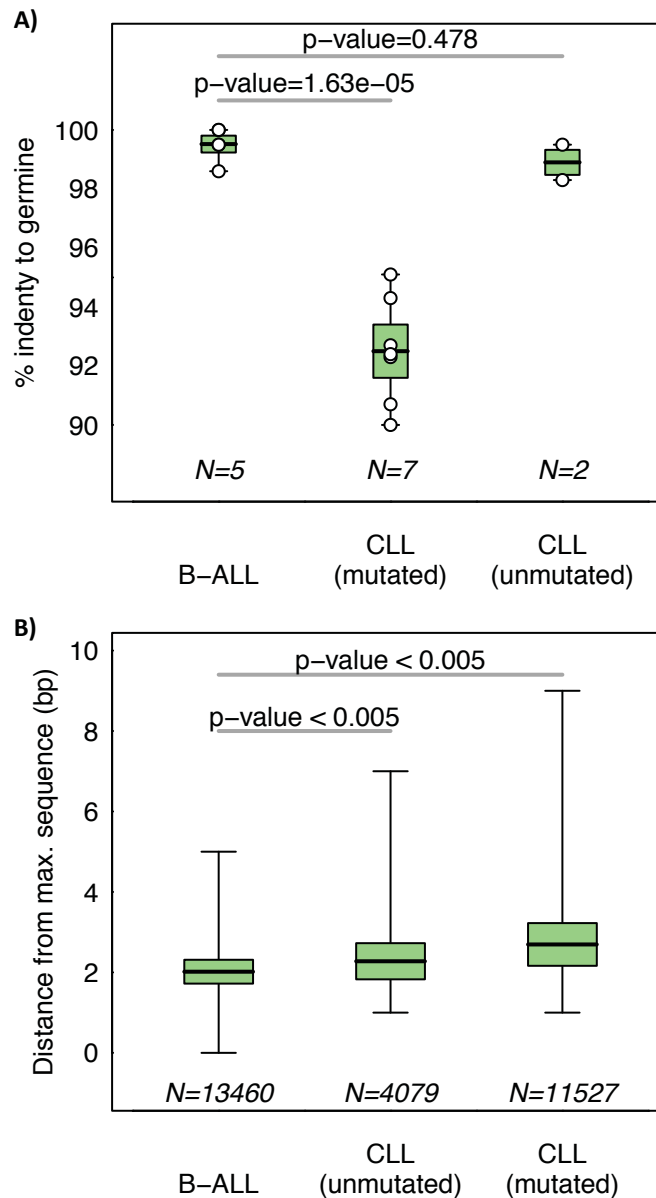
Figure 5.1. Comparing the B-cell repertoire in B-ALL with CLL.

A) The percentage sequence identity of the dominant clonal sequence compared to reference germline sequences from B-ALL and CLL patients with either unmutated CLL (dominant BCR >98% sequence similarity with reference germline IgHV-D-J sequences) or mutated CLL (dominant BCR <98% sequence similarity with reference germline IgHV-D-J sequences) and **B)** the distribution of base-pair distances of all unique sequences in the malignant clusters away from the dominant clonal sequence from B-ALL and unmutated CLL or mutated CLL patients. P-values for comparisons between the distributions are indicated (two-sided T-test).

5.2.3. BCR sequencing sensitivity to detect B-ALL clones

Minimal residual disease detection by BCR sequencing requires the accurate detection of B-ALL- associated BCR sequences within a large sequencing dataset. Therefore a Python code, named MRD Assessment and Retrieval Code in pYthon (MRDARCY), was developed to identify malignant BCR sequences from a diagnostic B-ALL sample represented by sequences in the largest cluster, and to search samples at later time points for identical or related BCRs allowing for a specified number of base-pair mismatches (here a threshold of 8 bp is used).

To assess the sensitivity of BCR sequencing for detecting specific B-cell clones from RNA, we performed a titration experiment using serial 10-fold dilutions of a known clonal B-ALL PB sample RNA (sample 1592_A) into normal peripheral blood RNA. The IgH multiplex PCR primer set used in the PCR amplification of the B-cell repertoire consists of six primers, with each primer binding to a different subset of IgHV genes. Therefore it was hypothesized that the primer from this set that binds best to the leukaemic BCRs in the dilution series, denoted IgHV-specific primer, would amplify the leukaemic BCRs preferentially, thus increasing sensitivity of detecting leukaemic BCR compared to the multiplex approach. To test this, multiplex PCR amplification and singleplex IgHV-specific PCR amplification were performed on these samples. Each dilution series sample yielded an average of 125,642 filtered BCR sequences (range of 18,970-294,354, **Figure 5.2**A-B). 31.41% of all BCR sequences in the undiluted sample are related to the leukaemic cluster as identified by MRDARCY, where the percentages of leukaemic BCRs detected approximated to a log-log correlation with dilution. Leukaemia-specific BCR sequences were detected in dilutions as low as 1 in 10^7 RNA molecules for both the multiplex and singleplex IgHV-specific PCR strategies (**Figure 5.2**B when the BCR identity of the tumour clone was known *a priori*). In contrast to this, qPCR has been shown to have a sensitivity of 1 cell in 10^5-10^6 (Campana, 2010). Interestingly, there was an increase in sensitivity of an average of 13.57x using the singleplex IgHV-specific PCR strategy across the dilution range, suggesting that this patient specific MRD monitoring approach, where multiplex BCR sequencing is used on the initial sample and followed by specific clonotypic IgHV primer, could be adapted into a powerful clinical MRD monitoring tool. In fact, with this sensitivity, if only 1 B-ALL cell is present in a typical 5ml blood sample containing ~$1.5x10^6$ B-cells, a read depth of

only 4.5×10^6 is required to give a >95% probability of detection (using the Poisson distribution and assuming that all BCRs were amplified). Therefore, BCR sequencing has unparalleled sensitivity to capture specific sequences with an important application in MRD monitoring of B-ALL and potentially other B-cell leukaemias.

However, when the leukemic cluster BCR sequences are unknown, detection of expanded clones relies on detecting the maximum cluster size that is significantly different from that of healthy individuals, i.e. when there the leukaemic B-cell population represents 1 in 100-500 RNA molecules (light green line, **Figure 5.2**B), and is consistent with the dilution series in Section 3.2.9. Therefore, the sequencing of BCR repertoires at diagnosis of B-ALL may be critical to the subsequent detection and tracking of small clonal lymphoid populations in a background of polyclonal cells.

A)

No dilution   1:10   1:100

$1:10^3$ $1:10^5$ $1:10^7$

B)

Legend:
- % ALL sequences (multiplex primer)
- % ALL sequences (IgHV spec. primer)
- % max. cluster size (multiplex primer)
- % max. cluster size (IgHV spec. primer)

Y-axis: % of reads

X-axis: Dilution
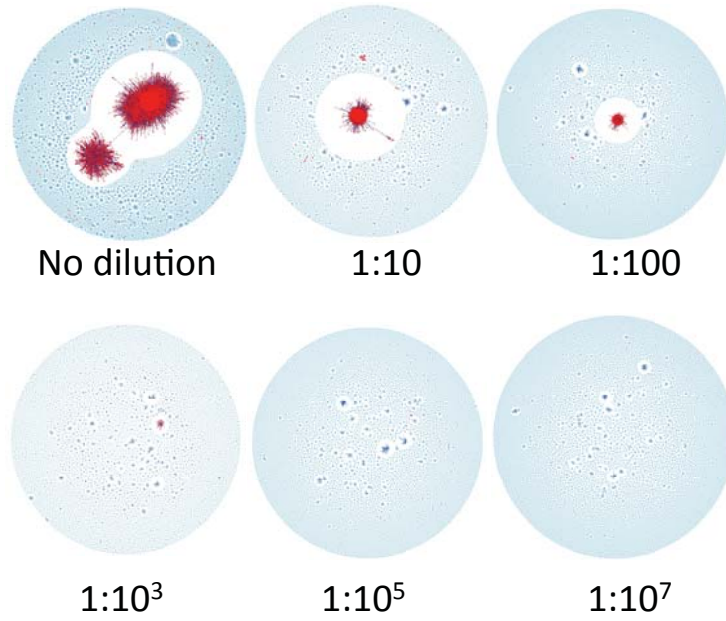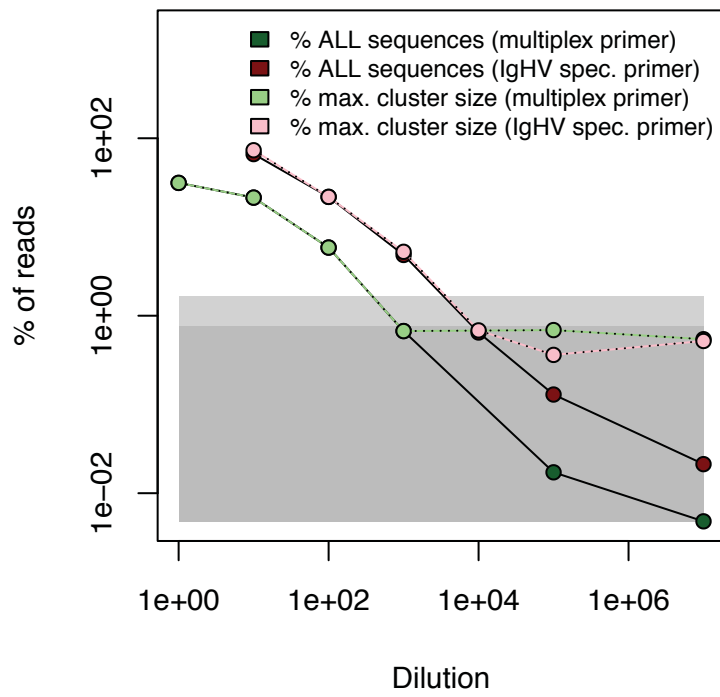
Figure 5.2. BCR sequencing sensitivity.

RNA from a clonal B-ALL patient sample was mixed with RNA from healthy peripheral blood PBMCs at different ratios. BCR sequencing using the full set of multiplex primers or a single PCR primer chosen from the multiplex primer set with the best alignment (i.e. annealing potential) to the malignant B-ALL BCR sequence (IgHV specific primer). **A)** Network diagrams showing sequential dilution of B-ALL BCR population into healthy blood using the multiplex primers, where vertices within 5bp sequence similarity to the B-ALL cluster are marked in red at each dilution, otherwise coloured blue. **B)** The percentages of BCR sequences corresponding to the B-ALL BCR population at each dilution into healthy RNA using multiplex primers (dark-green) and IgHV specific primer (dark-red). Overlaid with the percentage of BCR sequences in the largest cluster for multiplex primers (light-green) and IgHV specific primer (light-red).

It is possible, but unlikely, that the same IgHV-D-J rearrangement and joining regions can be generated by chance in independent B-cell clones, particularly as the B-ALL clonal BCRs are typically unmutated. To determine the false positive-rate for B-ALL BCR sequence detection, MRDARCY was used to detect B-ALL BCR sequences from the 6 B-ALL patients in 13 unrelated healthy BCR sequencing datasets using the same parameters (Table 5.2). A total of 23,480,661 BCR sequences were tested from unrelated samples, with only a single BCR match to a B-ALL cluster in B-ALL patient 5. This sequence was unmutated with short non-template additions (4bp) with 100% identity to a minor BCR clone in the B-ALL cluster (observed 219 times on the B-ALL patient). Therefore the presence of unrelated sequences matching the B-ALL-specific BCR sequence by chance occurs at a rate of 1 in $2x10^7$ BCR sequences/cells.

Table 5.2. False positive rate for detecting B-ALL MRD.

B-ALL patient	Number of unrelated healthy BCRs tested against B-ALL cluster	Number of reads matched*
B-ALL 1	3,730,269	0
B-ALL 2	4,098,690	0
B-ALL 3	4,097,093	0
B-ALL 4	3,836,054	0
B-ALL 5	3,922,068	1
B-ALL 6	3,796,487	0
Total	23,480,661	1

* By matching of B-ALL BCR sequences in 13 unrelated healthy sample BCR datasets.

5.2.4. Detecting B-ALL BCRs in clinical samples

Having shown the sensitivity of BCR sequencing, it was hypothesised that B-ALL clonal sequences will be detected in all the samples that were defined as qPCR T/C ratio MRD positive. Therefore, for each B-ALL patient, MRDARCY was used to identify BCR sequences in the largest cluster in the primary qPCR positive samples (highlighted in Table 5.1) and the percentage of matched BCR sequences in longitudinal samples was determined (allowing a maximum of 8 bp mismatches, **Figure 5.3**). Each of the six patients' samples showed a strong correlation between the fusion qPCR transcript levels (blue lines, **Figure 5.3**) and the frequencies of B-ALL sequences related to the largest cluster, known as clonotypic sequences (red lines, **Figure 5.3**), where the Pearson product-moment correlation coefficients between the percentage of B-ALL BCRs matched per sample and T/C ratios are >0.87 (Table 5.3). All samples that were qPCR positive were also positive for B-ALL BCR sequences. As the BCR sequencing sensitivity for detecting BCR sequences in RNA is greater than 1 in 107 and the qPCR result was very low, the lack of detection of B-ALL in patient 859 day 84 is likely to be due to the lack of sampling a B-ALL cell in the BM RNA aliquot used for PCR rather than failure of RNA detection. To determine whether detecting low-level B-ALL sequences is subject to sampling stochasticity, the low-level B-ALL RNA samples were re-amplified and re-sequenced (Table 5.4). Detection of B-ALL sequences were reproducible in samples where the number of B-ALL matched sequences was greater than 0.0016% of the BCR repertoire confirming that MRD above this level can be reliably detected using BCR sequencing. However, below this level detection of very low-level B-ALL sequences was subject to sampling stochasticity. Furthermore, some patient samples were positive for B-ALL BCR sequences where MRD was undetected using qPCR, such as patient 527 (day 15), indicating that the sensitivity of the BCR sequencing method equal to or better than that of qPCR, with the additional advantage that BCR sequencing based MRD monitoring can be done without gene fusion knowledge.

A) Patient 527

B) Patient 859

C) Patient 1611

D) Patient 1703

E) Patient 3243

F) Patient 1592

Legend:
- % BCRs in Cluster 1
- % BCRs in Cluster 2
- Target transcript level
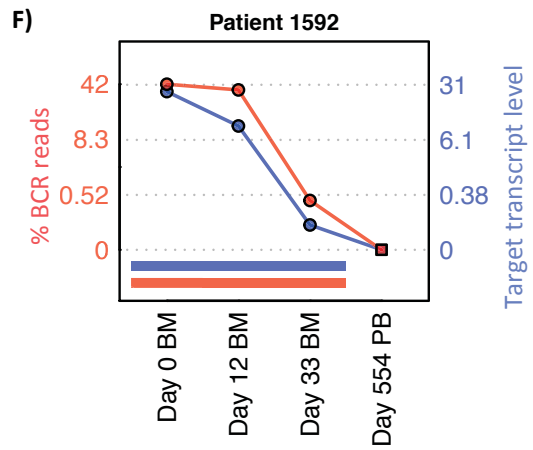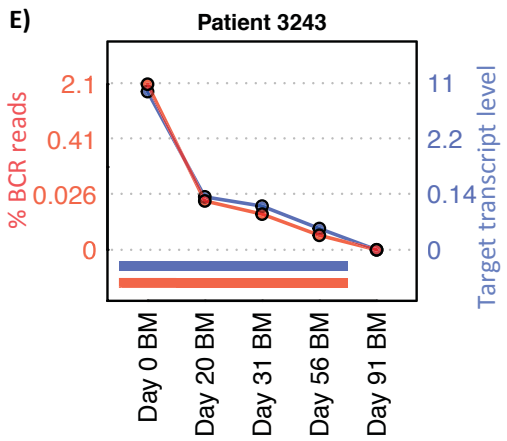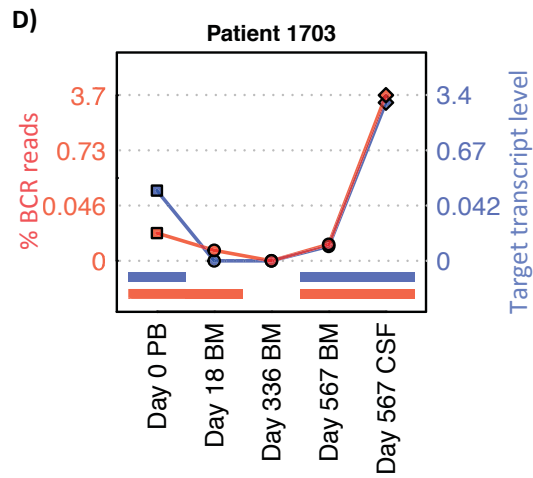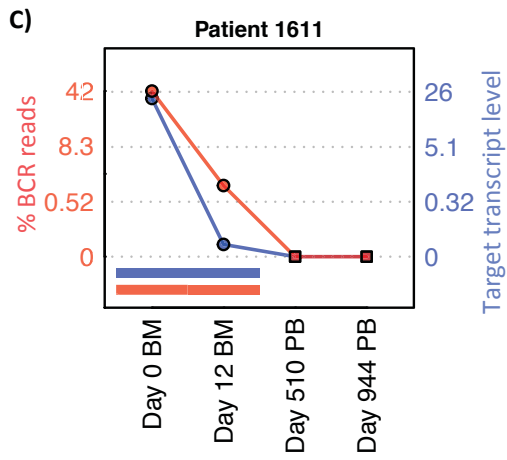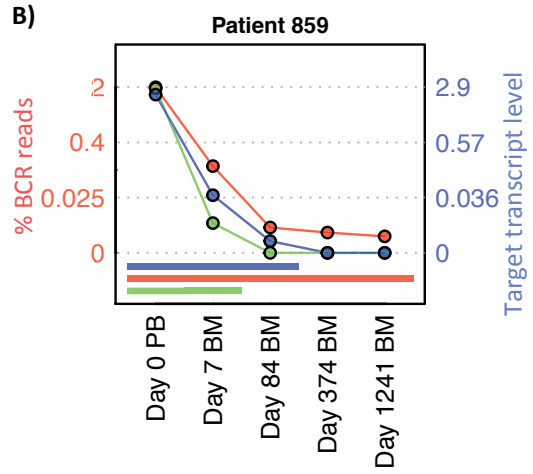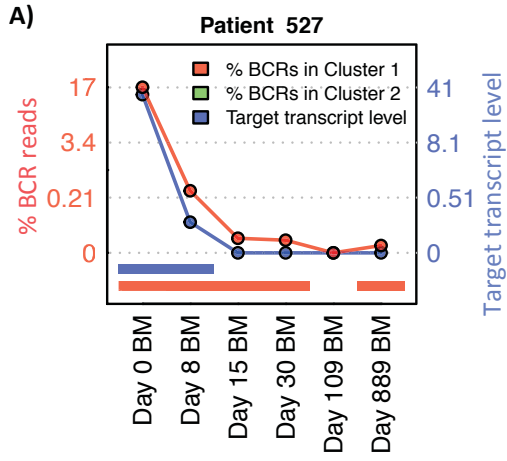
Figure 5.3. B-ALL BCR populations.

Variation of T/C qPCR transcript ratios (blue) and percentage of clonotypic B-ALL BCR reads over time for each patient (red and green for largest and second largest clusters respectively). The blue axis on the right of each plot corresponds to the T/C qPCR transcript ratios levels and the red axis on the left of each plot refers to the percentage of sequences in the corresponding clusters, both of which have a square scale to highlight lower frequency observations. Blue and red bars under each plot indicate time-points that are positive for B-ALL transcripts and B-ALL BCR reads respectively.

Table 5.3. Correlations between the percentage of B-ALL BCRs matched and qPCR levels.

Patient ID	Linear gradient between % BCRs matched and T/C ratio*	R^2-value*
527	0.3384	0.9997
859	0.5917	0.9997
1611	1.3216	0.9988
1703	0.9229	0.9986
3243	0.1627	1.0000
1592	0.8312	0.8782

* Linear gradients and Pearson product-moment correlation coefficients (R^2-values) between the percentage of B-ALL BCRs matched per sample and qPCR target to control transcript (T/C) ratios.

Table 5.4. Percentages of B-ALL clonotypic BCR sequences in repeated samples.

Patient ID	qPCR T/C level	Time since first sample (days)	BCR sequencing (initial sample)* % of B-ALL sequences	BCR sequencing (re-amplified)** % of B-ALL sequences
527	13.9510	0	41.21494	-
527	0.0197	8	0.81457	-
527	0.0000	15	0.00249	0.00056
527	0.0000	30	0.00140	0.00000
527	0.0000	109	0.00000	0.00000
527	0.0000	889	0.00016	0.00000
859	1.6612	0	2.89096	-
859	0.0292	7	0.21739	0.18325
859	0.0001	84	0.00159	0.00028
859	0.0000	374	0.00065	0.00032
859	0.0000	1241	0.00029	0.00031
1592	34.6048	0	31.45017	-
1592	12.9828	12	27.33152	-
1592	0.0211	33	0.24774	-
1592	0.0000	554	0.00000	-
1611	35.0403	0	26.48259	-
1611	0.0013	12	0.90122	0.12890
1611	0.0000	19	0.06560	0.00000
1611	0.0000	33	0.00000	0.00000
1611	0.0000	510	0.00000	-
1611	0.0000	944	0.00000	-
1703	0.1211	0	0.00266	0.00329
1703	0.0000	18	0.00005	0.00000
1703	0.0000	336	0.00000	0.00000
1703	0.0002	567	0.00033	0.00148
1703	3.1218	567	3.38261	-
3243	1.7453	0	10.73040	-
3243	0.0219	20	0.08141	-
3243	0.0102	31	0.02319	-
3243	0.0006	56	0.00063	-
3243	0.0000	91	0.00000	-

 * The initial BCR sequencing dataset.

** Where the RNA was re-amplified and sequenced independently.

For patient 859, the two largest clusters had similar sizes (2.807% and 2.891% of total reads) corresponding to IgHV gene rearrangements of [IgHV4-34, IgHD4-11, IgHJ6] and [IgHV1-2, IgHD4-11, IgHJ6] (**Figure 5.3**E, red and green lines respectively). The identical IgHD-IgHJ gene usage may be indicative of IgH secondary rearrangements. Although the second largest cluster (indicated in green) became undetectable after day 7, the largest cluster (indicated in red) was never fully eradicated over the 1241 days of sampling. Ongoing IgHV rearrangements have been shown to occur as the result of either of two processes. Firstly an ancestral B-ALL clone may undergo partial *IgH* gene rearrangement firstly of the IgHD-J genes, with multiple B-cells in this clone able to recombine the IgHD-J with different IgHV segments to become fully rearranged, thus generating multiple IgHV-D-J combinations sharing the same IgHD-J region. Secondly, in a secondary rearrangement, an existing IgHV in a full IgHV-D-J rearrangement may be exchanged for a 5' germline IgHV while retaining the same IgHD-J region (Marshall et al., 1995, Steenbergen et al., 1993, Gawad et al., 2012, Choi et al., 1996, Liu et al., 2013). Therefore, to assess whether these clusters may have originated from secondary rearrangements of a single ancestral BCR, the most frequently observed BCR sequence from both clusters were aligned to each other (**Figure 5.4**B). Although there is only 61% alignment identity between the two BCR sequences representing the two clusters, the 55 nucleotides spanning the IgHD-IgHJ region and, notably, 3pb of the 3' end of the IgHV gene in the cluster 2 BCR sequence is identical to IgHV-D joining region (consisting of random nucleotide additions during *IgH* gene rearrangement) were identical, which is consistent with the hypothesis of secondary rearrangements. In addition, these BCR sequences show no mutations in the IgHV genes compared to the reference germline database, thus reinforcing the hypothesis that these two clonal B-ALL BCRs are indeed from the same progenitor B-ALL B-cells from early stages of B-cell differentiation that have not undergone SHM but where a secondary rearrangement of the IgHV has occurred. This could potentially be determined through the sequencing of the light chain BCR sequences.

A) Patient 859 (Day 0)

454,071 total BCR reads

Cluster 1

Cluster 2

B)

Identical sequence

IgBLAST alignment for cluster 1 — IGHV4–34 — IGHD4–11 — IGHJ6

Cluster 1 vs Cluster 2 sequence alignment — 61% alignment identity

IgBLAST alignment for cluster 2 — IGHV1–2 — IGHD4–11 — IGHJ6

Alignment position (bp)

C)


```
                                            T  A  A  D  T  A  V  F  R  G  L  A  T  T  V  I  T  T  T  T  T  W  T  S  G  A  K  G  P
Cluster 1 BCR sequence         181  ACCGCCGCGGACACGGCTGTGTTCCGGGGGCTCGCAACTACAGTAATTACTACTACTACTACATGGACGTCTGGGGCCAAGGGACCA  267
V  100.0% (202/202)  IGHV4-34*01  256                                                                                       277
D  100.0% (10/10)    IGHD4-11*01    3                                                                                        12
J   97.6% (40/41)    IGHJ6*03       6                                                                                        45

                                            L  R  S  D  D  T  A  V  Y  Y  C  A  T  T  V  I  T  T  T  T  T  W  T  S  G  A  K  G  P
Cluster 2 BCR sequence         151  CTGAGATCTGACGACACGGCCGTGTATTACTGTGCAACTACAGTAATTACTACTACTACTACATGGACGTCTGGGGCCAAGGGACCA  236
V  100.0% (185/185)  IGHV1-2*02     286                                                                                      290
D  100.0% (10/10)    IGHD4-11*01      3                                                                                       12
J   97.4% (38/39)    IGHJ6*03         6                                                                                       44
```

Identical sequence

Figure 5.4. Bi-clonal B-cell expansion in B-ALL patient 859.

A) Network diagram for B-ALL patient 859 at day 0, where vertices within the largest cluster (Cluster 1) are coloured in red and vertices within the second largest cluster (Cluster 2) are coloured in green, otherwise vertices coloured in blue. **B)** BCR sequence alignment of the dominant sequences from the two dominant clusters in patient 859. Cluster 1 and cluster 2 refer to the largest and second largest clusters in the BCR sequence network for patient 859 respectively (representing 2.81% and 2.89% of BCRs respectively). The cluster 1 and 2 sequences were aligned to each other, and the positions of differences between sequences are indicated by the coloured boxes in the corresponding positions in the middle row, using red for mismatches, green for gaps in cluster 1 BCR and blue for gaps in cluster 2 BCR. The percentage identities of each alignment are indicated at the right of each sequence depiction. The cluster 1 and 2 sequences had 100% alignment identity with IgHV gene rearrangements of [IgHV4-34, IgHD4-11, IgHJ6] and [IgHV1-2, IgHD4-11, IgHJ6] respectively, where the red, blue and green boxes for IgHV, D and J genes mark the gene boundaries respectively. **C)** Alignments of Cluster 1 and cluster 2 BCR sequences with closest reference IgHV (highlighted in red), IgHD (highlighted in blue) and IgHJ (highlighted in green) genes, where . denotes alignment similarity between the cluster sequences and the reference genes, A/T/G/C denotes a different base to the reference, and - denotes the region outside of the gene alignments. The 55pg region of the BCR sequence that is identical between the cluster 1 and cluster 2 sequences is highlight in the orange box and yellow text.

In addition, these clusters display similar properties, including the mean distance from the most frequently observed BCR within each cluster (2.281bp and 2.135bp for clusters 1 and 2 respectively, Table 5.5). However, MRD was observed only for cluster 1 throughout the 1241 days of sampling, suggesting that these clusters were differentially affected by therapy. Therefore, BCR sequencing can detect multiple disease subclones irrespective of their composition of driver mutations and individual proliferative properties.

Table 5.5. Table of the properties of the largest two clusters in patient 859

	Cluster 1	Cluster 2
Cluster size (% of total sequences)	2.469	2.379
N reads	11211	10801
Number of unique sequences in cluster	2858	2037
IgHV gene	IGHV4-34*01	IGHV1-2*02
IgHJ gene	IGHJ6*03	IGHJ6*03
Number of sequences representing most frequently observed BCR	5625	6603
Mean distance from most frequently observed BCR	2.281	2.135

5.2.5. Detecting B-ALL BCRs in RNA and DNA

The BCR RNA expression in mature B-cells is greater than that of pre-B-cells or immature B-cells (Hoffmann et al., 2002). To account for the possibility that B-cell receptor expression in B-ALL cells/samples may be lower than in non-malignant mature B-cells, which may lead to the under-estimation of the number of malignant B-cells in a given sample, the DNA and RNA BCR repertoires were compared in three patient samples (**Figure 5.5**). For every patient time point, B-ALL-derived BCR sequences were detected in the DNA sample at a higher percentage of total BCR sequences compared to the percentage derived from studying the matched RNA sample. Therefore, although BCR sequencing is highly sensitive for the detection of B-ALL-derived sequences, the RNA BCR repertoire may be significantly underestimating the true percentage of B-ALL cells in the sample and the use of DNA repertoires in B-ALL may further increase the sensitivity for MRD detection. However, DNA is more stable in plasma than RNA so detection of plasma DNA may

be more indicative of lysed or dead cells, whereas plasma RNA is more readily degraded (El-Hefnawy et al., 2004, Garcia-Olmo et al., 2013). As the difference is very striking between the RNA and DNA clonotype frequencies in B-ALL samples, DNA BCR sequencing should be used as an MRD marker rather than RNA.
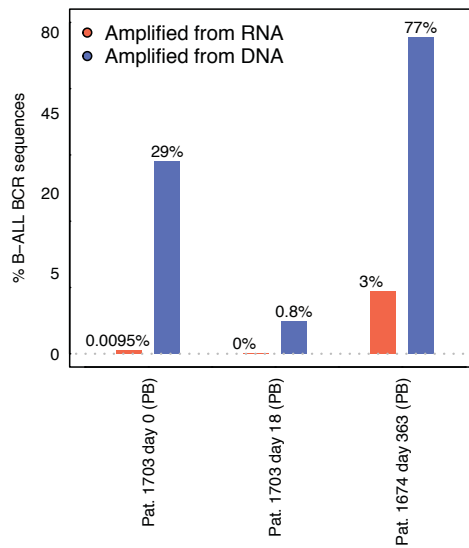
Figure 5.5. Detection of B-ALL BCR sequences in RNA and DNA samples.

Bar-graph showing the percentages of B-ALL sequences from BCR datasets generated from either the RNA or DNA from B-ALL patients (red and blue bars respectively).

5.2.6. Distinguishing between B-ALL and healthy samples

Increased clonality is observed in B-ALL samples with high levels of leukaemic load (i.e. when the qPCR T/C transcript ratio is greater than 1.66, Table 5.1). However, it is possible that the B-cell populations in B-ALL patients after therapy would still be distinct from healthy B-cell populations. If so, features of the B-cell repertoire would distinguish between B-ALL patient samples with high leukaemic loads (B-ALL high, T/C qPCR transcript ratio>1), B-ALL patient samples with low levels of leukaemic loads (B-ALL low, T/C qPCR transcript ratio<1), B-ALL patient samples with undetectable MRD after therapy (B-ALL undetectable, T/C qPCR transcript ratio=0) and healthy B-cell samples. Therefore, for each B-ALL sample and the 18 healthy individual samples, nine features of the B-cell sequencing data were calculated to distinguish between these different sample types, namely:

(a) *The vertex and cluster Gini index:* measurements of overall clonality and cluster size heterogeneity respectively.

(b) *The largest cluster size (as a percentage)*: to distinguish between samples with different maximum cluster sizes.

(c) *The sum of the largest two cluster sizes (as a percentage):* measurement to incorporate the second largest cluster size, which may distinguish between samples with secondary rearrangements.

(d) *The percentage of unique BCRs in largest cluster*: to distinguish between samples with different levels of SHM in the largest cluster.

(e) *The percentage of sequences representing the most frequently observed BCR sequence*: to distinguish between samples with or without dominant BCR sequences.

(f) *The percentage of sequences representing the first and second most frequently observed IgHV-J rearrangement:* measurement to distinguish between samples with specific rearrangements, irrespective of the largest cluster sizes.

(g) *The ratio of the number of unique CDR3 sequences to unique full length BCR sequences*: as the CDR3 length is shorter than the full length BCR sequence, but B-cells sharing the same CDR3 sequence are likely to originate from a single pre-B-all precursor, then lower ratios of unique

CDR3 sequences to unique full length BCR sequences suggests lower B-cell clonal complexity.

Each of these features describes different aspects of the B-cell repertoire. Linear discriminant analysis (LDA) was performed to find a linear combination of features that best separates sample types (**Figure 5.6**A) (Rindskopf, 1997). The first LDA dimension (LDA 1) separates the B-ALL high samples from the B-ALL low, B-ALL undetectable and healthy samples. The features that contribute most to distinguishing between these sample groups are the largest cluster size (contribution: -219.4), the sum of the largest two cluster sizes (denoted 1st + 2nd largest cluster (%), contribution: 302.8), the percentage of sequences corresponding the most frequently observed BCR sequence (denoted max. vertex, contribution: -189.0), and the percentage of sequences corresponding the first and second most frequently observed IgHV-J rearrangement (denoted Max. VJ Gene freq and 2nd max. VJ Gene freq contribution respectively: contributions of 204.7 and -165.4). The second LDA dimension (LDA 2) separates the healthy samples from the B-ALL low/undetectable samples. The features that contribute most to distinguishing between these sample groups are the vertex Gini index (contribution: 517.2), and cluster Gini index (contribution: -293.7), as indicated by the highest magnitude of the corresponding variable contributions for LDA2 (**Figure 5.6**B). Therefore, two-dimensional LDA successfully distinguishes B-ALL high, B-ALL low/undetectable and healthy samples.

To test whether the resulting LDA 1 and LDA2 linear combinations can be used as a linear classifier of sample type, hierarchical clustering was performed using the Euclidean distances between the LDA 1 and LDA 2 coordinates of each sample (as defined in Section 2.2.11, **Figure 5.6**C). This shows clear separation of B-ALL high samples (branch A, **Figure 5.6**C) from healthy samples (branch B, **Figure 5.6**C). The B-ALL low/undetectable samples were indistinguishable by these methods, but, interestingly, distinct from the other two groups (branch C, **Figure 5.6**C). 2 out of 18 healthy samples were misclassified into branch C, indicating that some healthy individuals may exhibit a range of B-cell repertoire features that can overlap with that of B-ALL low/undetectable.

These data show that patient B-ALL B-cell repertoires differ significantly from those of healthy individuals during maximum tumour burden, which is

unsurprising. Notably however, patient B-ALL B-cell repertoires during and after maximum tumour removal by therapy differ significantly from those of healthy individuals. Such a difference of low or undetectable B-ALL BCR repertoires may represent an effect of the prior presence of a large B-ALL clone or an effect of anti-leukaemic therapy, as patients remain on maintenance treatment for 2-3 years after diagnosis including lymphotoxic drugs such as corticosteroids and antimetabolites (e.g. Methotrexate). Overall, two-dimensional LDA in BCR sequencing repertoires successfully distinguished between B-ALL high samples, B-ALL low/undetectable samples and healthy samples and can effectively classify such samples. Whether subsets of the B-ALL low/undetectable clusters of patients, perhaps those without a "healthy" cluster member (branch C', **Figure 5.6**C), are more likely to relapse would be interesting to pursue. Alternatively, these "healthy" individuals that co-cluster with the B-ALL patients may have more clonal features of their B-cell repertoires for reasons that are unclear.
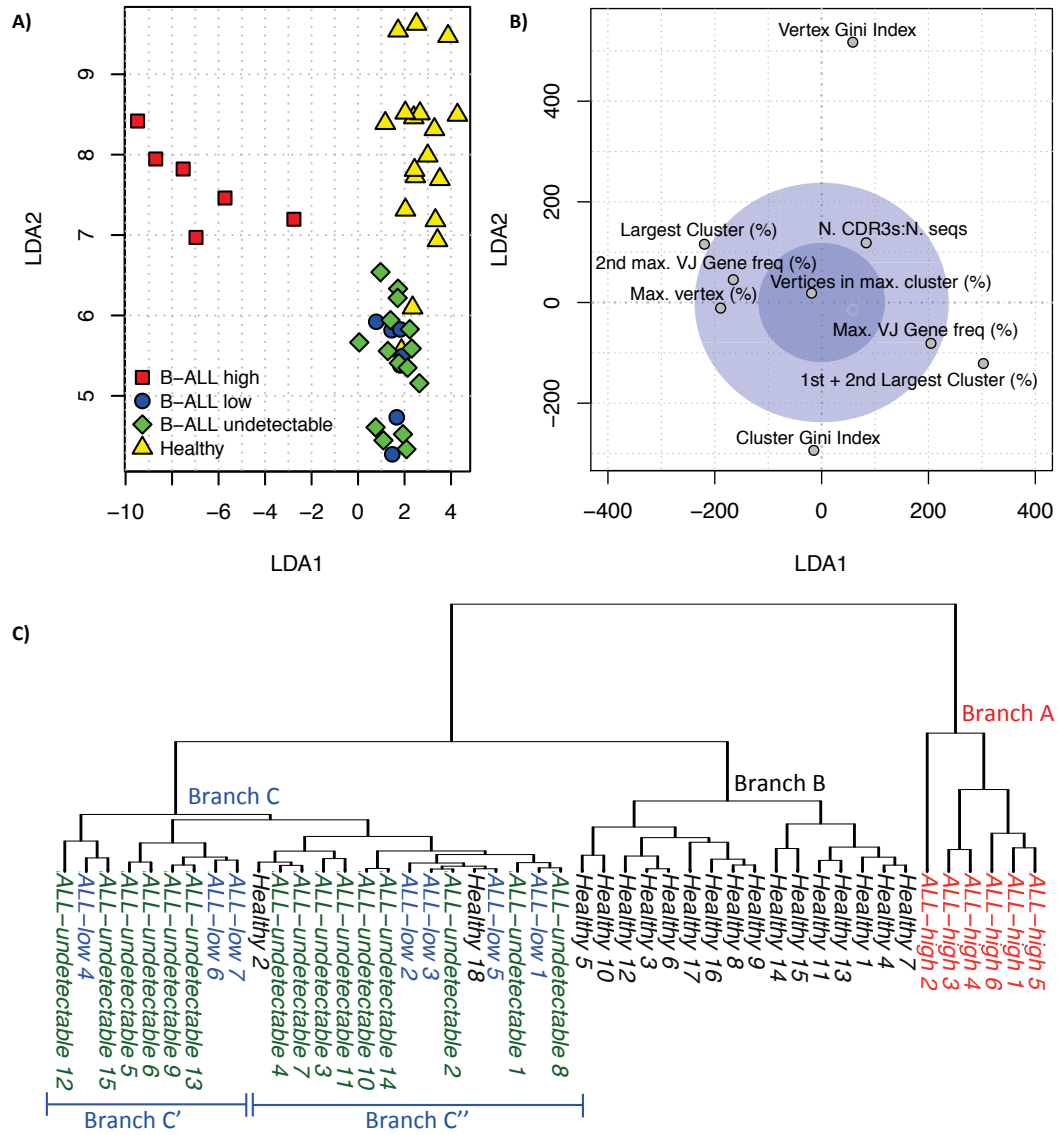
Figure 5.6. Distinguishing between B-ALL and healthy B-cell populations.

A) Linear discriminant analysis (LDA) performed on all samples to differentiate between diagnostic B-ALL samples (B-ALL-high, red (T/C qPCR transcript ratio>1)), B-ALL samples after treatment with detectable MRD (B-ALL-low, blue (T/C qPCR transcript ratio<1)), B-ALL samples with undetectable MRD (B-ALL-undetectable, green (T/C qPCR transcript ratio=0)), and healthy individuals (yellow). **B)** LDA variable contributions for the first two dimensions, where the blue circle indicates the mean scalar variable contributions for the first two dimensions, and variables outside this region indicate greatest contribution to the separation of classes. **C)** Hierarchical clustering tree of the patient samples using the distance measures derived from LDA 1 and 2. Branches (A), (B) and (C) refer to branches of the hierarchical tree corresponding to B-ALL-high, healthy and B-ALL-low/undetectable samples respectively, and where branches C' and C'' are sub-branches in branch C.

5.2.7. ALL Relapse: a case study of CSF relapse

One of the patients in this cohort, patient 1703, unfortunately developed CSF relapse after more than 2 years from initial therapy (summarised in). The sample taken on day 0 was taken more than one week after therapy had started and likely after a significant reduction in disease bulk, therefore the B-ALL qPCR T/C transcript level was relatively low. B-ALL was undetectable by day 18 (by qPCR MRD monitoring), but re-emerged at day 567 predominantly in the CSF, although it was also detectable in the BM. In this patient, both the B-cell were amplified and sequenced to understand the adaptive immune dynamics of relapse in B-ALL.

Table 5.6. Detection of B-ALL cells in patient 859.

Source*	Target/control transcript level**	% B-ALL BCR reads (from RNA)	% B-ALL BCR reads (from DNA)
Day 0, PB	0.121	0.00266	28.63019
Day 18, BM	0	5.42E-05	0.804093
Day 336, BM	0	0	-
Day 567, BM	0.000222	0.000332	-
Day 567, CSF	3.122	3.38	-

* Abbreviations: BM is bone marrow, PB is peripheral blood and CSF is cerebrospinal fluid.
** Target transcript: TEL/AML1 translocation.

The largest clone in the patient 1703 day 0 DNA sample, representing 28.63% of all BCR sequences, was identified as the B-ALL clone. This clone was detected as the largest cluster in the day 567 CSF sample (from RNA), representing 3.38% of BCR sequences (Table 5.6). However, 80% of cells in this sample resembled the leukaemia-associated immunophenotype of lymphoblasts (CD10$^+$, CD19$^+$, CD45$^{low/-}$) by flow cytometry. The reason for a low representation of B-ALL BCRs in the RNA sample compared to flow-cytometry is unclear but could be explained by the lower expression of immunoglobulin in B-ALL cells compared to mature B-cells (addressed in Section 5.2.5).

Clonal evolution has been observed in B-ALL as exemplified by the presence of tumour mutations in the genome (Mullighan, 2012) and by multiple BCRs related to the dominant B-ALL BCR sequence. Although expression of AID, which is

required for somatic hypermutation, has been detected only in some B-ALL patients (Feldhahn et al., 2007, Messina et al., 2011, Iacobucci et al., 2010, Hardianti et al., 2005), the accumulation of non-AID-mediated or mutations caused by low-level AID expression in these cells can result in clonal diversification in B-ALL (Jiao et al., 2014). These mutations may be used to infer the mutational route from a B-ALL B-cell ancestor to the rest of the leukaemic clone by phylogenetic analysis. To infer the phylogenetic relationships between B-ALL sequences before and after relapse, all the BCR sequences related to the B-ALL clone at day 0 derived by combining both RNA and DNA sequencing datasets and day 567 relapse (from RNA sequencing dataset) were identified (including identical or related BCRs within a threshold of 8 bp of the using MRDARCY) and aligned using Mafft (Katoh and Standley, 2013) and a maximum parsimony tree was fitted using Paup* (Wilgenbusch and Swofford, 2003). The branch lengths represent the evolutionary distance between BCR sequences. Bootstrapping was performed to evaluate the reproducibility of the trees, showing strong tree support (>95% certainty for all branches), and the tree tips were coloured according to whether the BCRs were observed at day 0 (BM) and/or day 567 (CSF) (**Figure 5.7**A). The tree has a star-like structure, suggesting that the original B-ALL BCR clone emerged from a single common ancestor (Martins and Housworth, 2002), represented by the central BCR, which was the most frequently observed BCR at day 0 (BM) (making up 40.0% and 74.6% of total related B-ALL sequences for BCR repertoires derived from RNA and DNA respectively) and day 567 (CSF) (40.0% and 63.0% of total related B-ALL sequences for BM and CSF respectively). Interestingly, there was high BCR sequence overlap between the day 0 (BM) and day 567 (CSF) samples (86.08%), even at distances of 7 nucleotides from the central BCR (**Figure 5.7**B). Furthermore, there is a strong linear correlation between the B-ALL BCR frequencies the day 0 (BM) and day 567 (CSF) samples (R^2-value=0.9993 for all BCRs), suggesting that some of the population structure of this B-ALL cluster is

Chapter 6

6. Overall summary and future work

6.1. Overall summary

Healthy humans have approximately 3×10^9 B-cells in the peripheral blood and this population encompasses the repertoire of distinct B-cells expressing different B-cell receptors (BCRs) necessary to bind diverse antigens and produce an effective humoral immune response. B-cells are dynamic populations of immune cells that evolve over time. The aim of this thesis was to investigate B-cell population diversity and dynamics in health and disease using the sequence diversity and population structure of the B-cell BCR repertoire. This required the development of novel, robust, sensitive and reproducible high-throughput B-cell receptor sequencing methods.

This thesis demonstrates that human BCR repertoire diversity can be interpreted through full V-D-J genotype diversity using networks. BCR sequences can be organised into networks based on sequence diversity, with differences in network connectivity providing clinically useful B-cell repertoire structure information. An important result of this framework is the ability to determine how B-cell repertoire structures differ between health and disease. Samples from clonal B-cell populations, such as from CLL, B-ALL and other clonal blood disorders, can readily be distinguished from healthy samples by an increase in BCR clonality and decrease in BCR diversity. For example, different features of the B-cell repertoire can be used to distinguish between patients with B-ALL patients with high leukaemic cell loads, B-ALL patients with low or undetectable levels of leukaemic cell loads and healthy individuals, and can be used as a sample classifier. Interestingly, the B-ALL samples remain largely distinct from healthy B-cell repertoires even after years of undetectable disease, suggesting a long-term B-cell repertoire impact of either the disease or, more likely, the anti-leukaemic therapy. Similarly, even though CLL therapy by Chlorambucil results in significant reduction in peripheral blood B-cell clonality, CLL patient samples remain distinct from equivalent samples from healthy individuals. Long-term effects of B-cell depletion therapy have been observed in previous studies, where SHM rate is reduced in rheumatoid arthritis patients even 6 years after

rituximab treatment (Dorner et al., 2010, Muhammad et al., 2009, Stolz and Schuler, 2009). There was variation between the diversity measures of the BCR repertoires between the healthy individuals, thus indicating a range representing healthy B-cell clonality and diversity. A larger-scaled assessment of primary immune responses compared to early stage leukaemias could provide clinically important diagnostic or prognostic information to patients.

The utility of this method can extended to clinical monitoring of disease, MRD and relapse. Unparalleled sensitivity of BCR sequencing for detecting MRD was demonstrated here compared to conventional clinical methods, where detection of leukaemic BCR RNA is greater than 1 in 10^7 RNA molecules, which is increased 13.57-fold by using only a single IgHV-specific primer corresponding to the specific BCR of interest. In practice, when there is prior knowledge of a BCR of interest, such as in leukaemia, the limit of detection is dependent on the number of cells sampled and the sequencing depth. However, the limit of *de novo* detection of malignant clonality is at least 1 in 100 dilution of CLL or ALL cells into healthy blood. When the clone of interest is small (i.e. less than 1 in 100 cells), diversity measures alone cannot directly be used to distinguish from healthy samples.. Therefore, detection of MRD is achievable as long as the clinical sample contains malignant cells, sequencing is performed at an adequate depth and the malignant clonal sequence is known *a priori*. In addition to increased sensitivity, the ability to detect multiple subclones in leukaemias by BCR sequencing highlights its advantages over qPCR methods, thought to occur between 1.38-2.70% in CLL (Plevova et al., 2014, Kern et al., 2014), 19.35-27% in ALL (Beishuizen et al., 1991, Kitchingman et al., 1986), and 10% in lymphomas (Sklar et al., 1984). This is particularly relevant in diseases where B-cell clones can undergo secondary rearrangements or in cases of two independent B-cell malignancies (Boyd et al., 2009, Bashford-Rogers et al., 2013). Enlarged clusters representing BCRs with different IgHV-D-J gene combinations may be due to either the expansion of two distinct malignant B-cell transformations, or separate antigen-stimulated B-cell clonal expansion unrelated to the malignancy. The presence of more than one BCR clonal expansion has unknown clinical implications in CLL and B-ALL, but with the risk of secondary malignancies in these patients, monitoring these bi-clonal B-cell disorders is of great clinical importance.

The utility of these methods could extend further to autoimmunity, immunodeficiency, response to infection and vaccination, thus potentially improving the understanding and clinical practices of a vast realm of diseases.

6.2. Future work

Using this thesis as a framework for immune repertoire analysis, it is apparent that there are many biological and clinical applications to the methods described here. The utility of these methods extend beyond malignancy to autoimmunity, immunodeficiency, response to infection and vaccination. However, this section will cover directions of future work directly derived from the findings in this thesis.

Firstly the full human allelic variation in the heavy and light Ig V, (D) and J genes is still unknown, where population differences in gene sequences may result in differential susceptibility of diseases. Biases in immunoglobulin gene recombination patterns have been shown to affect influenza susceptibility, where a polymorphism in the recombination signal sequence of IgKV locus in the Navajo population is associated with increased influenza susceptibility. This polymorphism reduces recombination of a commonly used IgKV gene by about 4.5-fold (Feeney et al., 1996). Therefore, future work should include determining the association between allelic variation in the heavy and light IgV, (D) and J genes and corresponding promoter regions. This could potentially be achieved through the analysis of the immunoglobulin loci of large-scale datasets, such as exome or whole genome sequencing of large numbers of individuals from the UK10K and 1000 Genome datasets and by a large scale analysis of IgH and IgL productive rearrangement frequencies in the peripheral blood of diverse populations of people.

Further experiments should include determining the B-cell repertoire differences between different anatomical locations within an individual, such as between lymph nodes and peripheral blood. Model systems, such as mice, can be used to investigate the development and spatial structure of immune responses during vaccination or infectious challenge. However, the availability of some anatomical regions from humans is limited, for example, bone marrow biopsies are typically only taken in individuals with blood abnormalities, such as anaemia, leukopenia, thrombocytopenia and leukaemia. However, even these samples, when paired with peripheral blood, could give valuable information on the spatial arrangement of specific B-cell populations. For example, this thesis has shown important potential uses of BCR sequencing in monitoring disease during therapy (Chapter 4) and minimal residual disease detection (in Chapter 5). However, it is of great clinical benefit to defining optimal anatomical locations for detecting minimal residual B-cell

and, potentially, T-cell populations during leukaemia therapy. In particular, it would be clinically useful to determine if single or multiple peripheral blood draws are more effective at sampling malignant B-cells for detection of MRD compared to bone marrow biopsies in B-ALL. Additionally, the mode of relapse remains a question in many leukaemias, such as where B-ALL MRD cells reside during therapy and what circumstances lead to relapse, particularly to certain anatomical sites such as CSF. Therefore, multiple sampling of different anatomical sites may give information on which regions are less readily accessible to therapy and potential reservoirs of cancer cells.

Multiple B-cell clonal expansions were observed in some of the CLL and B-ALL patients in this thesis, which opens the question of whether these clones are distinct malignant B-cell transformations, or separate antigen-stimulated B-cell clonal expansion unrelated to the malignancy. This may be answered by two different approaches. Firstly, BCR sequencing of longitudinal samples from these patients may be used to determine whether there is reduction in the sizes of any of the clones in the absence of therapy, suggesting antigen-driven clonal expansion and subsequent reduction. Secondly, single-cell whole-genome or exome sequencing may be performed on cells from the expanded clones to determine whether there are shared genomic features or aberrations that may be indicative of single or multiple malignant expansions. An alternative to this would be single-cell transcriptomic analysis, which would give information on the differences between cells from different clonal expansions on an RNA-expression level, potentially shedding light on the similar or different processes that have led to their clonal growth.

Previous studies have shown that non-B- and non-T-cell malignancies are often marked by profound defects in B-cell and T-cell function, such as in melanoma and solid tumors in mice (Baitsch et al., 2011, Ahmadzadeh et al., 2009, Sakuishi et al., 2010)). B- and T-cell exhaustion prevents optimal control of infection and malignancy, and therefore understanding the structure and dynamics of the normal B- and T-cell repertoires in patients with different malignancies may help identify common underlying principles of immune-dysfunction, to assess potential for diagnostic or prognostic marker for disease development and to identify therapeutic opportunities. This is achievable by cell sorting of activated memory B-cells or plasma cells, paired heavy and light chain sequencing and screening for reactivity against the malignant cell populations of interest. Next, the question of which B-cell

subsets produce these anti-malignant BCRs could be addressed by BCR high-throughput sequencing of flow sorted B-cell populations. This may determine whether anti-malignant B-cells are indeed prone to immunological exhaustion, and the dynamics of such a process may be determined using longitudinal samples. Importantly, an exhaustive phenotype of B-cells that specifically bind malignant cells may be a useful biomarker for relapse risk.

References

http://www.stemcell.com. Frequencies of Cell Types in Human Peripheral Blood.

ADDERSON, E. E. 2001. Antibody repertoires in infants and adults: effects of T-independent and T-dependent immunizations. *Springer Semin Immunopathol,* 23**,** 387-403.

ADDERSON, E. E., SHACKELFORD, P. G., QUINN, A. & CARROLL, W. L. 1991. Restricted Ig H chain V gene usage in the human antibody response to Haemophilus influenzae type b capsular polysaccharide. *J Immunol,* 147**,** 1667-74.

ADDERSON, E. E., SHACKELFORD, P. G., QUINN, A., WILSON, P. M., CUNNINGHAM, M. W., INSEL, R. A. & CARROLL, W. L. 1993. Restricted immunoglobulin VH usage and VDJ combinations in the human response to Haemophilus influenzae type b capsular polysaccharide. Nucleotide sequences of monospecific anti-Haemophilus antibodies and polyspecific antibodies cross-reacting with self antigens. *J Clin Invest,* 91**,** 2734-43.

AGATHANGELIDIS, A., DARZENTAS, N., HADZIDIMITRIOU, A., BROCHET, X., MURRAY, F., YAN, X. J., DAVIS, Z., VAN GASTEL-MOL, E. J., TRESOLDI, C., CHU, C. C., CAHILL, N., GIUDICELLI, V., TICHY, B., PEDERSEN, L. B., FORONI, L., BONELLO, L., JANUS, A., SMEDBY, K., ANAGNOSTOPOULOS, A., MERLE-BERAL, H., LAOUTARIS, N., JULIUSSON, G., DI CELLE, P. F., POSPISILOVA, S., JURLANDER, J., GEISLER, C., TSAFTARIS, A., LEFRANC, M. P., LANGERAK, A. W., OSCIER, D. G., CHIORAZZI, N., BELESSI, C., DAVI, F., ROSENQUIST, R., GHIA, P. & STAMATOPOULOS, K. 2012. Stereotyped B-cell receptors in one-third of chronic lymphocytic leukemia: a molecular classification with implications for targeted therapies. *Blood,* 119**,** 4467-4475.

AGUIAR, R. C., SOHAL, J., VAN RHEE, F., CARAPETI, M., FRANKLIN, I. M., GOLDSTONE, A. H., GOLDMAN, J. M. & CROSS, N. C. 1996. TEL-AML1 fusion in acute lymphoblastic leukaemia of adults. M.R.C. Adult Leukaemia Working Party. *Br J Haematol,* 95**,** 673-7.

AHMADZADEH, M., JOHNSON, L. A., HEEMSKERK, B., WUNDERLICH, J. R., DUDLEY, M. E., WHITE, D. E. & ROSENBERG, S. A. 2009. Tumor antigen-specific CD8 T cells infiltrating the tumor express high levels of PD-1 and are functionally impaired. *Blood,* 114**,** 1537-44.

ALBERS, C. A., LUNTER, G., MACARTHUR, D. G., MCVEAN, G., OUWEHAND, W. H. & DURBIN, R. 2011. Dindel: accurate indel calls from short-read data. *Genome research,* 21**,** 961-73.

ALT, F. W., BOTHWELL, A. L., KNAPP, M., SIDEN, E., MATHER, E., KOSHLAND, M. & BALTIMORE, D. 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.

ALTSCHUL, S. F., GISH, W., MILLER, W., MYERS, E. W. & LIPMAN, D. J. 1990. Basic local alignment search tool. *Journal of molecular biology,* 215**,** 403-10.

ALUGUPALLI, K. R., LEONG, J. M., WOODLAND, R. T., MURAMATSU, M., HONJO, T. & GERSTEIN, R. M. 2004. B1b lymphocytes confer T cell-independent long-lasting immunity. *Immunity,* 21**,** 379-90.

ANDRITSOS, L. & KHOURY, H. 2002. Chronic lymphocytic leukemia. *Curr Treat Options Oncol,* 3**,** 225-31.

ARBER, D. A. 2000. Molecular diagnostic approach to non-Hodgkin's lymphoma. *The Journal of molecular diagnostics : JMD,* 2**,** 178-90.

ARNAOUT, R., LEE, W., CAHILL, P., HONAN, T., SPARROW, T., WEIAND, M., NUSBAUM, C., RAJEWSKY, K. & KORALOV, S. B. 2011. High-resolution description of antibody heavy-chain repertoires in humans. *PloS one,* 6**,** e22365.

ARNOLD, L. W. & HAUGHTON, G. 1992. Autoantibodies to phosphatidylcholine. The murine antibromelain RBC response. *Annals of the New York Academy of Sciences,* 651**,** 354-9.

ARNOLD, L. W., PENNELL, C. A., MCCRAY, S. K. & CLARKE, S. H. 1994. Development of B-1 cells: segregation of phosphatidyl choline-specific B cells to the B-1 population occurs after immunoglobulin gene expression. *The Journal of experimental medicine,* 179**,** 1585-95.

BAITSCH, L., BAUMGAERTNER, P., DEVEVRE, E., RAGHAV, S. K., LEGAT, A., BARBA, L., WIECKOWSKI, S., BOUZOURENE, H., DEPLANCKE, B., ROMERO, P., RUFER, N. & SPEISER, D. E. 2011. Exhaustion of tumor-specific CD8(+) T cells in metastases from melanoma patients. *J Clin Invest,* 121**,** 2350-60.

BALTIMORE, D. 1981. Somatic mutation gains its place among the generators of diversity. *Cell,* 26**,** 295-6.

BANKOTI, J., APELTSIN, L., HAUSER, S. L., ALLEN, S., ALBERTOLLE, M. E., WITKOWSKA, H. E. & VON BUDINGEN, H. C. 2014. In multiple sclerosis, oligoclonal bands connect to peripheral B-cell responses. *Ann Neurol,* 75**,** 266-76.

BARAK, M., ZUCKERMAN, N. S., EDELMAN, H., UNGER, R. & MEHR, R. 2008. IgTree: creating Immunoglobulin variable region gene lineage trees. *J Immunol Methods,* 338**,** 67-74.

BASHFORD-ROGERS, R. J., PALSER, A. L., HUNTLY, B. J., RANCE, R., VASSILIOU, G. S., FOLLOWS, G. A. & KELLAM, P. 2013. Network properties derived from deep sequencing of human B-cell receptor repertoires delineate B-cell populations. *Genome Res*.

BATRAK, V., BLAGODATSKI, A. & BUERSTEDDE, J. M. 2011. Understanding the immunoglobulin locus specificity of hypermutation. *Methods Mol Biol,* 745**,** 311-26.

BEDFORD, T., SUCHARD, M. A., LEMEY, P., DUDAS, G., GREGORY, V., HAY, A. J., MCCAULEY, J. W., RUSSELL, C. A., SMITH, D. J. & RAMBAUT, A. 2014. Integrating influenza antigenic dynamics with molecular evolution. *Elife,* 3**,** e01914.

BEGLEITER, A., MOWAT, M., ISRAELS, L. G. & JOHNSTON, J. B. 1996. Chlorambucil in chronic lymphocytic leukemia: mechanism of action. *Leuk Lymphoma,* 23**,** 187-201.

BEISHUIZEN, A., HAHLEN, K., HAGEMEIJER, A., VERHOEVEN, M. A., HOOIJKAAS, H., ADRIAANSEN, H. J., WOLVERS-TETTERO, I. L., VAN WERING, E. R. & VAN DONGEN, J. J. 1991. Multiple rearranged immunoglobulin genes in childhood acute lymphoblastic leukemia of precursor B-cell origin. *Leukemia,* 5**,** 657-67.

BEN-HAMO, R. & EFRONI, S. 2011. The whole-organism heavy chain B cell repertoire from Zebrafish self-organizes into distinct network features. *BMC systems biology,* 5**,** 27.

BENICHOU, J., BEN-HAMO, R., LOUZOUN, Y. & EFRONI, S. 2012. Rep-Seq: uncovering the immunological repertoire through next-generation sequencing. *Immunology,* 135**,** 183-191.

BENICHOU, J., GLANVILLE, J., PRAK, E. T., AZRAN, R., KUO, T. C., PONS, J., DESMARAIS, C., TSABAN, L. & LOUZOUN, Y. 2013. The restricted DH gene reading frame usage in the expressed human antibody repertoire is selected based upon its amino acid content. *J Immunol,* 190**,** 5567-77.

BENNETT, J. M., CATOVSKY, D., DANIEL, M. T., FLANDRIN, G., GALTON, D. A., GRALNICK, H. R. & SULTAN, C. 1976. Proposals for the classification of the acute leukaemias. French-American-British (FAB) co-operative group. *Br J Haematol,* 33**,** 451-8.

BENNETT, J. M., CATOVSKY, D., DANIEL, M. T., FLANDRIN, G., GALTON, D. A., GRALNICK, H. R. & SULTAN, C. 1981. The morphological classification of acute lymphoblastic leukaemia: concordance among observers and clinical correlations. *Br J Haematol,* 47**,** 553-61.

BENSCHOP, R. J., MELAMED, D., NEMAZEE, D. & CAMBIER, J. C. 1999. Distinct signal thresholds for the unique antigen receptor-linked gene expression programs in mature and immature B cells. *J Exp Med,* 190**,** 749-56.

BERGQVIST, P., STENSSON, A., HAZANOV, L., HOLMBERG, A., MATTSSON, J., MEHR, R., BEMARK, M. & LYCKE, N. Y. 2013. Re-utilization of germinal centers in multiple Peyer's patches results in highly synchronized, oligoclonal, and affinity-matured gut IgA responses. *Mucosal Immunol,* 6**,** 122-35.

BERLAND, R. & WORTIS, H. H. 2002. Origins and functions of B-1 cells with notes on the role of CD5. *Annual review of immunology,* 20**,** 253-300.

BERMAN, J. E., NICKERSON, K. G., POLLOCK, R. R., BARTH, J. E., SCHUURMAN, R. K., KNOWLES, D. M., CHESS, L. & ALT, F. W. 1991. VH gene usage in humans: biased usage of the VH6 gene in immature B lymphoid cells. *Eur J Immunol,* 21**,** 1311-4.

BERNASCONI, N. L., TRAGGIAI, E. & LANZAVECCHIA, A. 2002. Maintenance of serological memory by polyclonal activation of human memory B cells. *Science,* 298**,** 2199-202.

BERTIOLI, D. 1997. Rapid amplification of cDNA ends. *Methods in molecular biology,* 67**,** 233-8.

BHATTACHARYA, N., DIENER, S., IDLER, I. S., BARTH, T. F., RAUEN, J., HABERMANN, A., ZENZ, T., MOLLER, P., DOHNER, H., STILGENBAUER, S. & MERTENS, D. 2011. Non-malignant B cells and chronic lymphocytic leukemia cells induce a pro-survival phenotype in CD14(+) cells from peripheral blood. *Leukemia,* 25**,** 722-726.

BINET, J.-L. 1994. Is the CHOP regimen a good treatment for advanced CLL? Results from two randomized clinical trials. French Cooperative Group on Chronic Lymphocytic Leukemia. *Leuk Lymphoma,* 13**,** 449-56.

BINET, J. L., LEPOPRIER, M., DIGHIERO, G., CHARRON, D., D'ATHIS, P., VAUGIER, G., BERAL, H. M., NATALI, J. C., RAPHAEL, M., NIZET, B. & FOLLEZOU, J. Y. 1977. A clinical staging system for chronic lymphocytic leukemia: prognostic significance. *Cancer,* 40**,** 855-64.

BIONDI, A. & MASERA, G. 1998. Molecular pathogenesis of childhood acute lymphoblastic leukemia. *Haematologica,* 83**,** 651-9.

BOIOCCHI, L., WITTER, R. E., HE, B., SUBRAMANIYAM, S., MATHEW, S., NIE, K., CERUTTI, A., COLEMAN, M., KNOWLES, D. M., ORAZI, A. & TAM, W. 2012. Composite chronic lymphocytic leukemia/small lymphocytic lymphoma and follicular lymphoma are biclonal lymphomas: a report of two cases. *Am J Clin Pathol,* 137**,** 647-59.

BOLANOS-MEADE, J., JACOBSOHN, D. A., MARGOLIS, J., OGDEN, A., WIENTJES, M. G., BYRD, J. C., LUCAS, D. M., ANDERS, V., PHELPS, M., GREVER, M. R. & VOGELSANG, G. B. 2005. Pentostatin in steroid-refractory acute graft-versus-host disease. *J Clin Oncol,* 23**,** 2661-8.

BONVALET, D., FOLDES, C. & CIVATTE, J. 1984. Cutaneous manifestations in chronic lymphocytic leukemia. *J Dermatol Surg Oncol,* 10**,** 278-82.

BOROWITZ, M. J., PULLEN, D. J., SHUSTER, J. J., VISWANATHA, D., MONTGOMERY, K., WILLMAN, C. L., CAMITTA, B. & CHILDREN'S ONCOLOGY GROUP, S. 2003. Minimal residual disease detection in childhood precursor-B-cell acute lymphoblastic leukemia: relation to other risk factors. A Children's Oncology Group study. *Leukemia,* 17**,** 1566-72.

BOYD, S. D., GAETA, B. A., JACKSON, K. J., FIRE, A. Z., MARSHALL, E. L., MERKER, J. D., MANIAR, J. M., ZHANG, L. N., SAHAF, B., JONES, C. D., SIMEN, B. B., HANCZARUK, B., NGUYEN, K. D., NADEAU, K. C., EGHOLM, M., MIKLOS, D. B., ZEHNDER, J. L. & COLLINS, A. M. 2010a. Individual variation in the germline Ig gene repertoire inferred from variable region gene rearrangements. *Journal of immunology,* 184**,** 6986-92.

BOYD, S. D., GAETA, B. A., JACKSON, K. J., FIRE, A. Z., MARSHALL, E. L., MERKER, J. D., MANIAR, J. M., ZHANG, L. N., SAHAF, B., JONES, C. D., SIMEN, B. B., HANCZARUK, B., NGUYEN, K. D., NADEAU, K. C., EGHOLM, M., MIKLOS, D. B., ZEHNDER, J. L. & COLLINS, A. M. 2010b. Individual variation in the germline Ig gene repertoire inferred from variable region gene rearrangements. *J Immunol,* 184**,** 6986-92.

BOYD, S. D., MARSHALL, E. L., MERKER, J. D., MANIAR, J. M., ZHANG, L. N., SAHAF, B., JONES, C. D., SIMEN, B. B., HANCZARUK, B., NGUYEN, K. D., NADEAU, K. C., EGHOLM, M., MIKLOS, D. B., ZEHNDER, J. L. & FIRE, A. Z. 2009. Measurement and clinical monitoring of human lymphocyte clonality by massively parallel VDJ pyrosequencing. *Science translational medicine,* 1**,** 12ra23.

BRAUNINGER, A., HANSMANN, M. L., STRICKLER, J. G., DUMMER, R., BURG, G., RAJEWSKY, K. & KUPPERS, R. 1999. Identification of common germinal-center B-cell precursors in two patients with both Hodgkin's disease and non-Hodgkin's lymphoma. *N Engl J Med,* 340**,** 1239-47.

BREZINSCHEK, H. P., FOSTER, S. J., BREZINSCHEK, R. I., DORNER, T., DOMIATI-SAAD, R. & LIPSKY, P. E. 1997. Analysis of the human VH gene repertoire. Differential effects of selection and somatic hypermutation on human peripheral CD5(+)/IgM+ and CD5(-)/IgM+ B cells. *J Clin Invest,* 99**,** 2488-501.

BRINEY, B. S., WILLIS, J. R., FINN, J. A., MCKINNEY, B. A. & CROWE, J. E., JR. 2014. Tissue-specific expressed antibody variable gene repertoires. *PLoS One,* 9**,** e100839.

BRINEY, B. S., WILLIS, J. R., HICAR, M. D., THOMAS, J. W., 2ND & CROWE, J. E., JR. 2012. Frequency and genetic characterization of V(DD)J recombinants in the human peripheral blood antibody repertoire. *Immunology,* 137**,** 56-64.

BRISCO, M. J., LATHAM, S., SUTTON, R., HUGHES, E., WILCZEK, V., VAN ZANTEN, K., BUDGEN, B., BAHAR, A. Y., MALEC, M., SYKES, P. J., KUSS, B. J., WATERS, K., VENN, N. C., GILES, J. E., HABER, M., NORRIS, M. D., MARSHALL, G. M. & MORLEY, A. A. 2009. Determining the repertoire of IGH gene rearrangements to develop molecular markers for minimal residual disease in B-lineage acute lymphoblastic leukemia. *J Mol Diagn,* 11**,** 194-200.

BRUGGEMANN, M., SCHRAUDER, A., RAFF, T., PFEIFER, H., DWORZAK, M., OTTMANN, O. G., ASNAFI, V., BARUCHEL, A., BASSAN, R., BENOIT, Y., BIONDI, A., CAVE, H., DOMBRET, H., FIELDING, A. K., FOA, R., GOKBUGET, N., GOLDSTONE, A. H., GOULDEN, N., HENZE, G., HOELZER, D., JANKA-SCHAUB, G. E., MACINTYRE, E. A., PIETERS, R., RAMBALDI, A., RIBERA, J. M., SCHMIEGELOW, K., SPINELLI, O., STARY, J., VON STACKELBERG, A., KNEBA, M., SCHRAPPE, M., VAN DONGEN, J. J., EUROPEAN WORKING GROUP FOR ADULT ACUTE LYMPHOBLASTIC, L. & INTERNATIONAL BERLIN-FRANKFURT-MUNSTER STUDY, G. 2010. Standardized MRD quantification in European ALL trials: proceedings of the Second International Symposium on MRD assessment in Kiel, Germany, 18-20 September 2008. *Leukemia,* 24**,** 521-35.

BRUGGEMANN, M., WHITE, H., GAULARD, P., GARCIA-SANZ, R., GAMEIRO, P., OESCHGER, S., JASANI, B., OTT, M., DELSOL, G., ORFAO, A., TIEMANN, M., HERBST, H., LANGERAK, A. W., SPAARGAREN, M., MOREAU, E., GROENEN, P. J., SAMBADE, C., FORONI, L., CARTER, G. I., HUMMEL, M., BASTARD, C., DAVI, F., DELFAU-LARUE, M. H., KNEBA, M., VAN DONGEN, J. J., BELDJORD, K. & MOLINA, T. J. 2007. Powerful strategy for polymerase chain reaction-based clonality assessment in T-cell malignancies Report of the BIOMED-2 Concerted Action BHM4 CT98-3936. *Leukemia : official journal of the Leukemia Society of America, Leukemia Research Fund, U.K,* 21**,** 215-21.

BUCH, M. H., SMOLEN, J. S., BETTERIDGE, N., BREEDVELD, F. C., BURMESTER, G., DORNER, T., FERRACCIOLI, G., GOTTENBERG, J. E., ISAACS, J., KVIEN, T. K., MARIETTE, X., MARTIN-MOLA, E., PAVELKA, K., TAK, P. P., VAN DER HEIJDE, D., VAN VOLLENHOVEN, R. F., EMERY, P. & RITUXIMAB CONSENSUS EXPERT, C. 2011. Updated consensus

statement on the use of rituximab in patients with rheumatoid arthritis. *Ann Rheum Dis,* 70**,** 909-20.

BURGER, J. A. & KIPPS, T. J. 2006. CXCR4: a key receptor in the crosstalk between tumor cells and their microenvironment. *Blood,* 107**,** 1761-7.

BURGER, J. A., TSUKADA, N., BURGER, M., ZVAIFLER, N. J., DELL'AQUILA, M. & KIPPS, T. J. 2000. Blood-derived nurse-like cells protect chronic lymphocytic leukemia B cells from spontaneous apoptosis through stromal cell-derived factor-1. *Blood,* 96**,** 2655-2663.

BURMEISTER, T., GOKBUGET, N., SCHWARTZ, S., FISCHER, L., HUBERT, D., SINDRAM, A., HOELZER, D. & THIEL, E. 2010. Clinical features and prognostic implications of TCF3-PBX1 and ETV6-RUNX1 in adult acute lymphoblastic leukemia. *Haematologica,* 95**,** 241-6.

BURTON, D. R. & WOOF, J. M. 1992. Human antibody effector function. *Adv Immunol,* 51**,** 1-84.

CAI, J., HUMPHRIES, C., RICHARDSON, A. & TUCKER, P. W. 1992. Extensive and selective mutation of a rearranged VH5 gene in human B cell chronic lymphocytic leukemia. *The Journal of experimental medicine,* 176**,** 1073-81.

CALIGARIS-CAPPIO, F. 2003. Role of the microenvironment in chronic lymphocytic leukaemia. *British journal of haematology,* 123**,** 380-8.

CALIGARIS-CAPPIO, F. & GHIA, P. 2008. Novel insights in chronic lymphocytic leukemia: are we getting closer to understanding the pathogenesis of the disease? *Journal of clinical oncology : official journal of the American Society of Clinical Oncology,* 26**,** 4497-503.

CALIN, G. A., FERRACIN, M., CIMMINO, A., DI LEVA, G., SHIMIZU, M., WOJCIK, S. E., IORIO, M. V., VISONE, R., SEVER, N. I., FABBRI, M., IULIANO, R., PALUMBO, T., PICHIORRI, F., ROLDO, C., GARZON, R., SEVIGNANI, C., RASSENTI, L., ALDER, H., VOLINIA, S., LIU, C. G., KIPPS, T. J., NEGRINI, M. & CROCE, C. M. 2005. A MicroRNA signature associated with prognosis and progression in chronic lymphocytic leukemia. *N Engl J Med,* 353**,** 1793-801.

CAMPANA, D. 2010. Minimal residual disease in acute lymphoblastic leukemia. *Hematology Am Soc Hematol Educ Program,* 2010**,** 7-12.

CAMPBELL, P. J., PLEASANCE, E. D., STEPHENS, P. J., DICKS, E., RANCE, R., GOODHEAD, I., FOLLOWS, G. A., GREEN, A. R., FUTREAL, P. A. & STRATTON, M. R. 2008. Subclonal phylogenetic structures in cancer revealed by ultra-deep sequencing. *Proceedings of the National Academy of Sciences of the United States of America,* 105**,** 13081-6.

CARULLI, G., MARINI, A., CIANCIA, E. M., BRUNO, J., VIGNATI, S., LAMBELET, P., CANNIZZO, E., OTTAVIANO, V., GALIMBERTI, S., CARACCIOLO, F., FERRERI, M. I., CIABATTI, E. & PETRINI, M. 2011. Discordant lymphoma consisting of splenic mantle cell lymphoma and marginal zone lymphoma involving the bone marrow and peripheral blood: a case report. *Journal of medical case reports,* 5**,** 476.

CASTRO, R., JOUNEAU, L., PHAM, H. P., BOUCHEZ, O., GIUDICELLI, V., LEFRANC, M. P., QUILLET, E., BENMANSOUR, A., CAZALS, F., SIX, A., FILLATREAU, S., SUNYER, O. & BOUDINOT, P. 2013. Teleost fish mount complex clonal IgM and IgT responses in spleen upon systemic viral infection. *PLoS Pathog,* 9**,** e1003098.

CAVACINI, L. A., KUHRT, D., DUVAL, M., MAYER, K. & POSNER, M. R. 2003. Binding and neutralization activity of human IgG1 and IgG3 from serum of HIV-infected individuals. *AIDS Res Hum Retroviruses,* 19**,** 785-92.

CERRONI, L., ZENAHLIK, P., HOFLER, G., KADDU, S., SMOLLE, J. & KERL, H. 1996. Specific cutaneous infiltrates of B-cell chronic lymphocytic leukemia: a clinicopathologic and prognostic study of 42 patients. *Am J Surg Pathol,* 20**,** 1000-10.

CHAUDHURI, J. & ALT, F. W. 2004. Class-switch recombination: interplay of transcription, DNA deamination and DNA repair. *Nat Rev Immunol,* 4**,** 541-52.

CHEN, I. M., HARVEY, R. C., MULLIGHAN, C. G., GASTIER-FOSTER, J., WHARTON, W., KANG, H., BOROWITZ, M. J., CAMITTA, B. M., CARROLL, A. J., DEVIDAS, M., PULLEN, D. J., PAYNE-TURNER, D., TASIAN, S. K., RESHMI, S., COTTRELL, C. E., REAMAN, G. H., BOWMAN, W. P., CARROLL, W. L., LOH, M. L., WINICK, N. J., HUNGER, S. P. & WILLMAN, C. L. 2012. Outcome modeling with CRLF2, IKZF1, JAK, and minimal residual disease in pediatric acute lymphoblastic leukemia: a Children's Oncology Group study. *Blood,* 119**,** 3512-22.

CHEN, L., APGAR, J., HUYNH, L., DICKER, F., GIAGO-MCGAHAN, T., RASSENTI, L., WEISS, A. & KIPPS, T. J. 2005. ZAP-70 directly enhances IgM signaling in chronic lymphocytic leukemia. *Blood,* 105**,** 2036-41.

CHEN, L., WIDHOPF, G., HUYNH, L., RASSENTI, L., RAI, K. R., WEISS, A. & KIPPS, T. J. 2002. Expression of ZAP-70 is associated with increased B-cell receptor signaling in chronic lymphocytic leukemia. *Blood,* 100**,** 4609-14.

CHESON, B. D., BENNETT, J. M., GREVER, M., KAY, N., KEATING, M. J., O'BRIEN, S. & RAI, K. R. 1996. National Cancer Institute-sponsored Working Group guidelines for chronic lymphocytic leukemia: revised guidelines for diagnosis and treatment. *Blood,* 87**,** 4990-7.

CHEVRIER, S., GENTON, C., KALLIES, A., KARNOWSKI, A., OTTEN, L. A., MALISSEN, B., MALISSEN, M., BOTTO, M., CORCORAN, L. M., NUTT, S. L. & ACHA-ORBEA, H. 2009. CD93 is required for maintenance of antibody secretion and persistence of plasma cells in the bone marrow niche. *Proc Natl Acad Sci U S A,* 106**,** 3895-900.

CHIORAZZI, N. & FERRARINI, M. 2003. B cell chronic lymphocytic leukemia: lessons learned from studies of the B cell antigen receptor. *Annu Rev Immunol,* 21**,** 841-94.

CHIORAZZI, N., RAI, K. R. & FERRARINI, M. 2005. Mechanisms of disease: Chronic lymphocytic leukemia. *New England Journal of Medicine,* 352**,** 804-815.

CHOI, M., SCHOLL, U. I., JI, W. Z., LIU, T. W., TIKHONOVA, I. R., ZUMBO, P., NAYIR, A., BAKKALOGLU, A., OZEN, S., SANJAD, S., NELSON-WILLIAMS, C., FARHI, A., MANE, S. & LIFTON, R. P. 2009. Genetic diagnosis by whole exome capture and massively parallel DNA sequencing. *Proceedings of the National Academy of Sciences of the United States of America,* 106**,** 19096-19101.

CHOI, Y., GREENBERG, S. J., DU, T. L., WARD, P. M., OVERTURF, P. M., BRECHER, M. L. & BALLOW, M. 1996. Clonal evolution in B-lineage acute lymphoblastic leukemia by contemporaneous VH-VH gene replacements and VH-DJH gene rearrangements. *Blood,* 87**,** 2506-12.

COIFFIER, B., LEPAGE, E., BRIERE, J., HERBRECHT, R., TILLY, H., BOUABDALLAH, R., MOREL, P., VAN DEN NESTE, E., SALLES, G., GAULARD, P., REYES, F., LEDERLIN, P. & GISSELBRECHT, C. 2002. CHOP chemotherapy plus rituximab compared with CHOP alone in elderly patients with diffuse large-B-cell lymphoma. *N Engl J Med,* 346**,** 235-42.

CORCIONE, A., ALOISI, F., SERAFINI, B., CAPELLO, E., MANCARDI, G. L., PISTOIA, V. & UCCELLI, A. 2005. B-cell differentiation in the CNS of patients with multiple sclerosis. *Autoimmunity reviews,* 4**,** 549-54.

CORTES, J., O'BRIEN, S. M., PIERCE, S., KEATING, M. J., FREIREICH, E. J. & KANTARJIAN, H. M. 1995. The value of high-dose systemic chemotherapy and intrathecal therapy for central nervous system prophylaxis in different risk groups of adult acute lymphoblastic leukemia. *Blood,* 86**,** 2091-7.

CORTHESY, B. 2007. Roundtrip ticket for secretory IgA: role in mucosal homeostasis? *J Immunol,* 178**,** 27-32.

CORTHESY, B. & KRAEHENBUHL, J. P. 1999. Antibody-mediated protection of mucosal surfaces. *Curr Top Microbiol Immunol,* 236**,** 93-111.

CORTI, D. & LANZAVECCHIA, A. 2013. Broadly neutralizing antiviral antibodies. *Annu Rev Immunol,* 31**,** 705-42.

CORTI, D., VOSS, J., GAMBLIN, S. J., CODONI, G., MACAGNO, A., JARROSSAY, D., VACHIERI, S. G., PINNA, D., MINOLA, A., VANZETTA, F., SILACCI, C., FERNANDEZ-RODRIGUEZ, B. M., AGATIC, G., BIANCHI, S., GIACCHETTO-SASSELLI, I., CALDER, L., SALLUSTO, F., COLLINS, P., HAIRE, L. F., TEMPERTON, N., LANGEDIJK, J. P., SKEHEL, J. J. & LANZAVECCHIA, A. 2011. A neutralizing antibody selected from plasma cells that binds to group 1 and group 2 influenza A hemagglutinins. *Science,* 333**,** 850-6.

COUSTAN-SMITH, E., RIBEIRO, R. C., STOW, P., ZHOU, Y., PUI, C. H., RIVERA, G. K., PEDROSA, F. & CAMPANA, D. 2006. A simplified flow cytometric assay identifies children with acute lymphoblastic leukemia who have a superior clinical outcome. *Blood,* 108**,** 97-102.

COUSTAN-SMITH, E., SANCHO, J., HANCOCK, M. L., BOYETT, J. M., BEHM, F. G., RAIMONDI, S. C., SANDLUND, J. T., RIVERA, G. K., RUBNITZ, J. E., RIBEIRO, R. C., PUI, C. H. & CAMPANA, D. 2000. Clinical importance of minimal residual disease in childhood acute lymphoblastic leukemia. *Blood,* 96**,** 2691-6.

CRAGG, M. S., WALSHE, C. A., IVANOV, A. O. & GLENNIE, M. J. 2005. The biology of CD20 and its potential as a target for mAb therapy. *Curr Dir Autoimmun,* 8**,** 140-74.

CRAIG, F. E. 2003. Bone marrow evaluation in pediatric patients. *Semin Diagn Pathol,* 20**,** 237-46.

CRESPO, M., BOSCH, F., VILLAMOR, N., BELLOSILLO, B., COLOMER, D., ROZMAN, M., MARCE, S., LOPEZ-GUILLERMO, A., CAMPO, E. & MONTSERRAT, E. 2003. ZAP-70 expression as a surrogate for immunoglobulin-variable-region mutations in chronic lymphocytic leukemia. *N Engl J Med,* 348**,** 1764-75.

CUNEO, A., RIGOLIN, G. M., BIGONI, R., DE ANGELI, C., VERONESE, A., CAVAZZINI, F., BARDI, A., ROBERTI, M. G., TAMMISO, E., AGOSTINI, P., CICCONE, M., DELLA PORTA, M., TIEGHI, A., CAVAZZINI, L., NEGRINI, M. & CASTOLDI, G. 2004. Chronic lymphocytic leukemia with 6q- shows distinct hematological features and intermediate prognosis. *Leukemia,* 18**,** 476-83.

DAGKLIS, A., PONZONI, M., GOVI, S., CANGI, M. G., PASINI, E., CHARLOTTE, F., VINO, A., DOGLIONI, C., DAVI, F., LOSSOS, I. S., NTOUNTAS, I., PAPADAKI, T., DOLCETTI, R., FERRERI, A. J., STAMATOPOULOS, K. & GHIA, P. 2012. Immunoglobulin gene repertoire in ocular adnexal lymphomas: hints on the nature of the antigenic stimulation. *Leukemia,* 26**,** 814-21.

DAMESHEK, W. & SCHWARTZ, R. S. 1959. Leukemia and auto-immunization- some possible relationships. *Blood,* 14**,** 1151-8.

DAMLE, R. N., WASIL, T., FAIS, F., GHIOTTO, F., VALETTO, A., ALLEN, S. L., BUCHBINDER, A., BUDMAN, D., DITTMAR, K., KOLITZ, J., LICHTMAN, S. M., SCHULMAN, P., VINCIGUERRA, V. P., RAI, K. R., FERRARINI, M. & CHIORAZZI, N. 1999. Ig V gene mutation status and CD38 expression as novel prognostic indicators in chronic lymphocytic leukemia. *Blood,* 94**,** 1840-7.

DARZENTAS, N. & STAMATOPOULOS, K. 2013. The Significance of Stereotyped B-Cell Receptors in Chronic Lymphocytic Leukemia. *Hematology-Oncology Clinics of North America,* 27**,** 237-+.

DAVIS, A. C. & SHULMAN, M. J. 1989. Igm - Molecular Requirements for Its Assembly and Function. *Immunology Today,* 10**,** 118-&.

DE VINUESA, C. G., COOK, M. C., BALL, J., DREW, M., SUNNERS, Y., CASCALHO, M., WABL, M., KLAUS, G. G. & MACLENNAN, I. C. 2000. Germinal centers without T cells. *J Exp Med,* 191**,** 485-94.

DEARDEN, C. 2008. Disease-specific complications of chronic lymphocytic leukemia. *Hematology / the Education Program of the American Society of Hematology. American Society of Hematology. Education Program*, 450-6.

DEKOSKY, B. J., IPPOLITO, G. C., DESCHNER, R. P., LAVINDER, J. J., WINE, Y., RAWLINGS, B. M., VARADARAJAN, N., GIESECKE, C., DORNER, T., ANDREWS, S. F., WILSON, P. C., HUNICKE-SMITH, S. P., WILLSON, C. G., ELLINGTON, A. D. & GEORGIOU, G. 2013. High-throughput sequencing of the paired human immunoglobulin heavy and light chain repertoire. *Nat Biotechnol, 31*, 166-9.

DERUDDER, E., CADERA, E. J., VAHL, J. C., WANG, J., FOX, C. J., ZHA, S., VAN LOO, G., PASPARAKIS, M., SCHLISSEL, M. S., SCHMIDT-SUPPRIAN, M. & RAJEWSKY, K. 2009. Development of immunoglobulin lambda-chain-positive B cells, but not editing of immunoglobulin kappa-chain, depends on NF-kappaB signals. *Nat Immunol, 10*, 647-54.

DIEHL, L. F. & KETCHUM, L. H. 1998. Autoimmune disease and chronic lymphocytic leukemia: autoimmune hemolytic anemia, pure red cell aplasia, and autoimmune thrombocytopenia. *Semin Oncol, 25*, 80-97.

DIMITROV, D. S. 2010. Therapeutic antibodies, vaccines and antibodyomes. *mAbs, 2*, 347-56.

DINARELLO, C. A. & BUNN, P. A., JR. 1997. Fever. *Semin Oncol, 24*, 288-98.

DOGAN, I., BERTOCCI, B., VILMONT, V., DELBOS, F., MEGRET, J., STORCK, S., REYNAUD, C. A. & WEILL, J. C. 2009. Multiple layers of B cell memory with different effector functions. *Nat Immunol, 10*, 1292-9.

DORIA-ROSE, N. A., SCHRAMM, C. A., GORMAN, J., MOORE, P. L., BHIMAN, J. N., DEKOSKY, B. J., ERNANDES, M. J., GEORGIEV, I. S., KIM, H. J., PANCERA, M., STAUPE, R. P., ALTAE-TRAN, H. R., BAILER, R. T., CROOKS, E. T., CUPO, A., DRUZ, A., GARRETT, N. J., HOI, K. H., KONG, R., LOUDER, M. K., LONGO, N. S., MCKEE, K., NONYANE, M., O'DELL, S., ROARK, R. S., RUDICELL, R. S., SCHMIDT, S. D., SHEWARD, D. J., SOTO, C., WIBMER, C. K., YANG, Y., ZHANG, Z., PROGRAM, N. C. S., MULLIKIN, J. C., BINLEY, J. M., SANDERS, R. W., WILSON, I. A., MOORE, J. P., WARD, A. B., GEORGIOU, G., WILLIAMSON, C., ABDOOL KARIM, S. S., MORRIS, L., KWONG, P. D., SHAPIRO, L., MASCOLA, J. R., BECKER, J., BENJAMIN, B., BLAKESLEY, R., BOUFFARD, G., BROOKS, S., COLEMAN, H., DEKHTYAR, M., GREGORY, M., GUAN, X., GUPTA, J., HAN, J., HARGROVE, A., HO, S. L., JOHNSON, T., LEGASPI, R., LOVETT, S., MADURO, Q., MASIELLO, C., MASKERI, B., MCDOWELL, J., MONTEMAYOR, C., MULLIKIN, J., PARK, M., RIEBOW, N., SCHANDLER, K., SCHMIDT, B., SISON, C., STANTRIPOP, M., THOMAS, J., THOMAS, P., VEMULAPALLI, M. & YOUNG, A. 2014. Developmental pathway for potent V1V2-directed HIV-neutralizing antibodies. *Nature, 509*, 55-62.

DORNER, T., BREZINSCHEK, H. P., FOSTER, S. J., BREZINSCHEK, R. I., FARNER, N. L. & LIPSKY, P. E. 1998. Delineation of selective influences shaping the mutated expressed human Ig heavy chain repertoire. *Journal of immunology, 160*, 2831-41.

DORNER, T., KINNMAN, N. & TAK, P. P. 2010. Targeting B cells in immune-mediated inflammatory disease: a comprehensive review of mechanisms of action and identification of biomarkers. *Pharmacol Ther, 125*, 464-75.

DOUGIER, H. L., REYNAUD, S., PINAUD, E., CARRION, C., DELPY, L. & COGNE, M. 2006. Interallelic class switch recombination can reverse allelic exclusion and allow trans-complementation of an IgH locus switching defect. *Eur J Immunol, 36*, 2181-91.

DOWNING, J. R. & SHANNON, K. M. 2002. Acute leukemia: a pediatric perspective. *Cancer Cell, 2*, 437-45.

DUKE, V. M., GANDINI, D., SHERRINGTON, P. D., LIN, K., HEELAN, B., AMLOT, P., MEHTA, A. B., HOFFBRAND, A. V. & FORONI, L. 2003. V(H) gene usage differs in germline and mutated B-cell chronic lymphocytic leukemia. *Haematologica, 88*, 1259-71.

DUNN-WALTERS, D. K., EDELMAN, H. & MEHR, R. 2004. Immune system learning and memory quantified by graphical analysis of B-lymphocyte phylogenetic trees. *Biosystems,* 76**,** 141-55.

DUPUIS, A., GAUB, M. P., LEGRAIN, M., DRENOU, B., MAUVIEUX, L., LUTZ, P., HERBRECHT, R., CHAN, S. & KASTNER, P. 2013. Biclonal and biallelic deletions occur in 20% of B-ALL cases with IKZF1 mutations. *Leukemia,* 27**,** 503-7.

DURIG, J., NASCHAR, M., SCHMUCKER, U., RENZING-KOHLER, K., HOLTER, T., HUTTMANN, A. & DUHRSEN, U. 2002. CD38 expression is an important prognostic marker in chronic lymphocytic leukaemia. *Leukemia : official journal of the Leukemia Society of America, Leukemia Research Fund, U.K,* 16**,** 30-5.

DWORZAK, M. N. & PANZER-GRUMAYER, E. R. 2003. Flow cytometric detection of minimal residual disease in acute lymphoblastic leukemia. *Leuk Lymphoma,* 44**,** 1445-55.

EDWARDS, J. C. & CAMBRIDGE, G. 2006. B-cell targeting in rheumatoid arthritis and other autoimmune diseases. *Nat Rev Immunol,* 6**,** 394-403.

EICHHORST, B., DREYLING, M., ROBAK, T., MONTSERRAT, E., HALLEK, M. & GROUP, E. G. W. 2011. Chronic lymphocytic leukemia: ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-up. *Ann Oncol,* 22 Suppl 6**,** vi50-4.

EISELE, L., HADDAD, T., SELLMANN, L., DUHRSEN, U. & DURIG, J. 2009. Expression levels of CD38 on leukemic B cells but not on non-leukemic T cells are comparably stable over time and predict the course of disease in patients with chronic lymphocytic leukemia. *Leukemia Research,* 33**,** 775-778.

EISEN, H. N. & SISKIND, G. W. 1964. Variations in Affinities of Antibodies during the Immune Response. *Biochemistry,* 3**,** 996-1008.

EL-HEFNAWY, T., RAJA, S., KELLY, L., BIGBEE, W. L., KIRKWOOD, J. M., LUKETICH, J. D. & GODFREY, T. E. 2004. Characterization of amplifiable, circulating RNA in plasma and its potential as a tool for cancer diagnostics. *Clinical Chemistry,* 50**,** 564-573.

ELDER, M. E., LIN, D., CLEVER, J., CHAN, A. C., HOPE, T. J., WEISS, A. & PARSLOW, T. G. 1994. Human severe combined immunodeficiency due to a defect in ZAP-70, a T cell tyrosine kinase. *Science,* 264**,** 1596-9.

EVANS, P. A. S., POTT, C., GROENEN, P. J. T. A., SALLES, G., DAVI, F., BERGER, F., GARCIA, J. F., VAN KRIEKEN, J. H. J. M., PALS, S., KLUIN, P., SCHUURING, E., SPAARGAREN, M., BOONE, E., GONZALEZ, D., MARTINEZ, B., VILLUENDAS, R., GAMEIRO, P., DISS, T. C., MILLS, K., MORGAN, G. J., CARTER, G. I., MILNER, B. J., PEARSON, D., HUMMEL, M., JUNG, W., OTT, M., CANIONI, D., BELDJORD, K., BASTARD, C., DELFAU-LARUE, M. H., VAN DONGEN, J. J. M., MOLINA, T. J. & CABECADAS, J. 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.

FAHAM, M., ZHENG, J., MOORHEAD, M., CARLTON, V. E., STOW, P., COUSTAN-SMITH, E., PUI, C. H. & CAMPANA, D. 2012. Deep-sequencing approach for minimal residual disease detection in acute lymphoblastic leukemia. *Blood,* 120**,** 5173-80.

FAIS, F., GHIOTTO, F., HASHIMOTO, S., SELLARS, B., VALETTO, A., ALLEN, S. L., SCHULMAN, P., VINCIGUERRA, V. P., RAI, K., RASSENTI, L. Z., KIPPS, T. J., DIGHIERO, G., SCHROEDER, H. W., JR., FERRARINI, M. & CHIORAZZI, N. 1998. Chronic lymphocytic leukemia B cells express restricted sets of mutated and unmutated antigen receptors. *J Clin Invest,* 102**,** 1515-25.

FEENEY, A. J., ATKINSON, M. J., COWAN, M. J., ESCURO, G. & LUGO, G. 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.

FEENEY, A. J., TANG, A. & OGWARO, K. M. 2000. B-cell repertoire formation: role of the recombination signal sequence in non-random V segment utilization. *Immunol Rev,* 175**,** 59-69.

FELDHAHN, N., HENKE, N., MELCHIOR, K., DUY, C., SOH, B. N., KLEIN, F., VON LEVETZOW, G., GIEBEL, B., LI, A., HOFMANN, W. K., JUMAA, H. & MUSCHEN, M. 2007. Activation-induced cytidine deaminase acts as a mutator in BCR-ABL1-transformed acute lymphoblastic leukemia cells. *J Exp Med,* 204**,** 1157-66.

FERRARINI, M. & CHIORAZZI, M. 2004. Recent advances in the molecular biology and immunobiology of chronic lymphocytic leukemia. *Seminars in Hematology,* 41**,** 207-223.

FEUGIER, P., VAN HOOF, A., SEBBAN, C., SOLAL-CELIGNY, P., BOUABDALLAH, R., FERME, C., CHRISTIAN, B., LEPAGE, E., TILLY, H., MORSCHHAUSER, F., GAULARD, P., SALLES, G., BOSLY, A., GISSELBRECHT, C., REYES, F. & COIFFIER, B. 2005. Long-term results of the R-CHOP study in the treatment of elderly patients with diffuse large B-cell lymphoma: a study by the Groupe d'Etude des Lymphomes de l'Adulte. *J Clin Oncol,* 23**,** 4117-26.

FIELDING, A. K. 2008. The treatment of adults with acute lymphoblastic leukemia. *Hematology Am Soc Hematol Educ Program***,** 381-9.

FIELDING, A. K., ROWE, J. M., RICHARDS, S. M., BUCK, G., MOORMAN, A. V., DURRANT, I. J., MARKS, D. I., MCMILLAN, A. K., LITZOW, M. R., LAZARUS, H. M., FORONI, L., DEWALD, G., FRANKLIN, I. M., LUGER, S. M., PAIETTA, E., WIERNIK, P. H., TALLMAN, M. S. & GOLDSTONE, A. H. 2009. Prospective outcome data on 267 unselected adult patients with Philadelphia chromosome-positive acute lymphoblastic leukemia confirms superiority of allogeneic transplantation over chemotherapy in the pre-imatinib era: results from the International ALL Trial MRC UKALLXII/ECOG2993. *Blood,* 113**,** 4489-96.

FINN, J. A. & CROWE, J. E., JR. 2013. Impact of new sequencing technologies on studies of the human B cell repertoire. *Curr Opin Immunol,* 25**,** 613-8.

FISHER, J., YAN, M., HEUIJERJANS, J., CARTER, L., ABOLHASSANI, A., FROSCH, J., WALLACE, R., FLUTTER, B., HUBANK, M., KLEIN, N., CALLARD, R., GUSTAFSSON, K. & ANDERSON, J. 2014. Neuroblastoma killing properties of V-delta 2 and V-delta2 negative gamma delta T cells following expansion by artificial antigen presenting cells. *Clin Cancer Res*.

FOOTE, J. B., MAHMOUD, T. I., VALE, A. M. & KEARNEY, J. F. 2012. Long-term maintenance of polysaccharide-specific antibodies by IgM-secreting cells. *J Immunol,* 188**,** 57-67.

FOSTER, H. D. 2008. Host-pathogen evolution: Implications for the prevention and treatment of malaria, myocardial infarction and AIDS. *Medical hypotheses,* 70**,** 21-5.

FRAZER, K. A., BALLINGER, D. G., COX, D. R., HINDS, D. A., STUVE, L. L., GIBBS, R. A., BELMONT, J. W., BOUDREAU, A., HARDENBOL, P., LEAL, S. M., PASTERNAK, S., WHEELER, D. A., WILLIS, T. D., YU, F., YANG, H., ZENG, C., GAO, Y., HU, H., HU, W., LI, C., LIN, W., LIU, S., PAN, H., TANG, X., WANG, J., WANG, W., YU, J., ZHANG, B., ZHANG, Q., ZHAO, H., ZHOU, J., GABRIEL, S. B., BARRY, R., BLUMENSTIEL, B., CAMARGO, A., DEFELICE, M., FAGGART, M., GOYETTE, M., GUPTA, S., MOORE, J., NGUYEN, H., ONOFRIO, R. C., PARKIN, M., ROY, J., STAHL, E., WINCHESTER, E., ZIAUGRA, L., ALTSHULER, D., SHEN, Y., YAO, Z., HUANG, W., CHU, X., HE, Y., JIN, L., LIU, Y., SUN, W., WANG, H., WANG, Y., XIONG, X., XU, L., WAYE, M. M., TSUI, S. K., XUE, H., WONG, J. T., GALVER, L. M., FAN, J. B., GUNDERSON, K., MURRAY, S. S., OLIPHANT, A. R., CHEE, M. S., MONTPETIT, A., CHAGNON, F., FERRETTI, V., LEBOEUF, M., OLIVIER, J. F., PHILLIPS, M. S., ROUMY, S., SALLEE, C., VERNER, A., HUDSON, T. J., KWOK, P. Y., CAI, D., KOBOLDT, D. C., MILLER, R. D., PAWLIKOWSKA, L., TAILLON-MILLER, P., XIAO, M., TSUI, L. C., MAK, W., SONG, Y. Q., TAM, P. K., NAKAMURA, Y.,

KAWAGUCHI, T., KITAMOTO, T., MORIZONO, T., NAGASHIMA, A., OHNISHI, Y., SEKINE, A., TANAKA, T., TSUNODA, T., et al. 2007. A second generation human haplotype map of over 3.1 million SNPs. *Nature,* 449**,** 851-61.

FREEMAN, J. D., WARREN, R. L., WEBB, J. R., NELSON, B. H. & HOLT, R. A. 2009. Profiling the T-cell receptor beta-chain repertoire by massively parallel sequencing. *Genome research,* 19**,** 1817-24.

FROLICH, D., GIESECKE, C., MEI, H. E., REITER, K., DARIDON, C., LIPSKY, P. E. & DORNER, T. 2010. Secondary immunization generates clonally related antigen-specific plasma cells and memory B cells. *J Immunol,* 185**,** 3103-10.

FRUMKIN, D., WASSERSTROM, A., KAPLAN, S., FEIGE, U. & SHAPIRO, E. 2005. Genomic variability within an organism exposes its cell lineage tree. *PLoS computational biology,* 1**,** e50.

FUENTES-PANANA, E. M., BANNISH, G., SHAH, N. & MONROE, J. G. 2004. Basal Igalpha/Igbeta signals trigger the coordinated initiation of pre-B cell antigen receptor-dependent processes. *J Immunol,* 173**,** 1000-11.

GABERT, J., BEILLARD, E., VAN DER VELDEN, V. H., BI, W., GRIMWADE, D., PALLISGAARD, N., BARBANY, G., CAZZANIGA, G., CAYUELA, J. M., CAVE, H., PANE, F., AERTS, J. L., DE MICHELI, D., THIRION, X., PRADEL, V., GONZALEZ, M., VIEHMANN, S., MALEC, M., SAGLIO, G. & VAN DONGEN, J. J. 2003. Standardization and quality control studies of 'real-time' quantitative reverse transcriptase polymerase chain reaction of fusion gene transcripts for residual disease detection in leukemia - a Europe Against Cancer program. *Leukemia,* 17**,** 2318-57.

GALL, A., KAYE, S., HUE, S., BONSALL, D., RANCE, R., BAILLIE, G. J., FIDLER, S. J., WEBER, J. N., MCCLURE, M. O., KELLAM, P. & INVESTIGATORS, S. T. 2013. Restriction of V3 region sequence divergence in the HIV-1 envelope gene during antiretroviral treatment in a cohort of recent seroconverters. *Retrovirology,* 10**,** 8.

GALSON, J. D., POLLARD, A. J., TRUCK, J. & KELLY, D. F. 2014. Studying the antibody repertoire after vaccination: practical applications. *Trends Immunol,* 35**,** 319-331.

GARCIA-OLMO, D. C., PICAZO, M. G., TOBOSO, I., ASENSIO, A. I. & GARCIA-OLMO, D. 2013. Quantitation of cell-free DNA and RNA in plasma during tumor progression in rats. *Molecular Cancer,* 12.

GASCUEL, O. 1997. BIONJ: an improved version of the NJ algorithm based on a simple model of sequence data. *Molecular biology and evolution,* 14**,** 685-95.

GAWAD, C., PEPIN, F., CARLTON, V. E., KLINGER, M., LOGAN, A. C., MIKLOS, D. B., FAHAM, M., DAHL, G. & LACAYO, N. 2012. Massive evolution of the immunoglobulin heavy chain locus in children with B precursor acute lymphoblastic leukemia. *Blood,* 120**,** 4407-17.

GEISBERGER, R., LAMERS, M. & ACHATZ, G. 2006. The riddle of the dual expression of IgM and IgD. *Immunology,* 118**,** 429-37.

GHIA, P., MELCHERS, F. & ROLINK, A. G. 2000. Age-dependent changes in B lymphocyte development in man and mouse. *Experimental gerontology,* 35**,** 159-65.

GHIA, P., STROLA, G., GRANZIERO, L., GEUNA, M., GUIDA, G., SALLUSTO, F., RUFFING, N., MONTAGNA, L., PICCOLI, P., CHILOSI, M. & CALIGARIS-CAPPIO, F. 2002. Chronic lymphocytic leukemia B cells are endowed with the capacity to attract CD4+, CD40L+ T cells by producing CCL22. *European journal of immunology,* 32**,** 1403-13.

GHIOTTO, F., FAIS, F., VALETTO, A., ALBESIANO, E., HASHIMOTO, S., DONO, M., IKEMATSU, H., ALLEN, S. L., KOLITZ, J., RAI, K. R., NARDINI, M., TRAMONTANO, A., FERRARINI, M. & CHIORAZZI, N. 2004. Remarkably similar antigen receptors among a subset of patients with chronic lymphocytic leukemia. *J Clin Invest,* 113**,** 1008-16.

GINE, E., BOSCH, F., VILLAMOR, N., ROZMAN, M., COLOMER, D., LOPEZ-GUILLERMO, A., CAMPO, E. & MONTSERRAT, E. 2002. Simultaneous diagnosis of hairy cell leukemia

and chronic lymphocytic leukemia/small lymphocytic lymphoma: a frequent association? *Leukemia,* 16**,** 1454-9.

GIOVANNONI, G., COMI, G., COOK, S., RAMMOHAN, K., RIECKMANN, P., SOELBERG SORENSEN, P., VERMERSCH, P., CHANG, P., HAMLETT, A., MUSCH, B., GREENBERG, S. J. & GROUP, C. S. 2010. A placebo-controlled trial of oral cladribine for relapsing multiple sclerosis. *N Engl J Med,* 362**,** 416-26.

GIUDICELLI, V., BROCHET, X. & LEFRANC, M. P. 2011. IMGT/V-QUEST: IMGT standardized analysis of the immunoglobulin (IG) and T cell receptor (TR) nucleotide sequences. *Cold Spring Harbor protocols,* 2011**,** 695-715.

GIUDICELLI, V. & LEFRANC, M. P. 1999. Ontology for immunogenetics: the IMGT-ONTOLOGY. *Bioinformatics,* 15**,** 1047-54.

GLANVILLE, J., ZHAI, W., BERKA, J., TELMAN, D., HUERTA, G., MEHTA, G. R., NI, I., MEI, L., SUNDAR, P. D., DAY, G. M., COX, D., RAJPAL, A. & PONS, J. 2009a. Precise determination of the diversity of a combinatorial antibody library gives insight into the human immunoglobulin repertoire. *Proc Natl Acad Sci U S A,* 106**,** 20216-21.

GLANVILLE, J., ZHAI, W., BERKA, J., TELMAN, D., HUERTA, G., MEHTA, G. R., NI, I., MEI, L., SUNDAR, P. D., DAY, G. M., COX, D., RAJPAL, A. & PONS, J. 2009b. Precise determination of the diversity of a combinatorial antibody library gives insight into the human immunoglobulin repertoire. *Proceedings of the National Academy of Sciences of the United States of America,* 106**,** 20216-21.

GOJOBORI, T. & NEI, M. 1986. Relative contributions of germline gene variation and somatic mutation to immunoglobulin diversity in the mouse. *Mol Biol Evol,* 3**,** 156-67.

GRABOWSKI, P., HULTDIN, M., KARLSSON, K., TOBIN, G., ALESKOG, A., THUNBERG, U., LAURELL, A., SUNDSTROM, C., ROSENQUIST, R. & ROOS, G. 2005. Telomere length as a prognostic parameter in chronic lymphocytic leukemia with special reference to VH gene mutation status. *Blood,* 105**,** 4807-12.

GRANZIERO, L., GHIA, P., CIRCOSTA, P., GOTTARDI, D., STROLA, G., GEUNA, M., MONTAGNA, L., PICCOLI, P., CHILOSI, M. & CALIGARIS-CAPPIO, F. 2001. Survivin is expressed on CD40 stimulation and interfaces proliferation and apoptosis in B-cell chronic lymphocytic leukemia. *Blood,* 97**,** 2777-83.

GREEN, M. R., GENTLES, A. J., NAIR, R. V., IRISH, J. M., KIHIRA, S., LIU, C. L., KELA, I., HOPMANS, E. S., MYKLEBUST, J. H., JI, H., PLEVRITIS, S. K., LEVY, R. & ALIZADEH, A. A. 2013. Hierarchy in somatic mutations arising during genomic evolution and progression of follicular lymphoma. *Blood,* 121**,** 1604-11.

GRIBBEN, J. G. 2009. Stem cell transplantation in chronic lymphocytic leukemia. *Biol Blood Marrow Transplant,* 15**,** 53-8.

GRIFFITHS, G. M., BEREK, C., KAARTINEN, M. & MILSTEIN, C. 1984. Somatic mutation and the maturation of immune response to 2-phenyl oxazolone. *Nature,* 312**,** 271-5.

GRILLO-LOPEZ, A. J., HEDRICK, E., RASHFORD, M. & BENYUNES, M. 2002. Rituximab: ongoing and future clinical development. *Semin Oncol,* 29**,** 105-12.

GRONWALL, C., KOSAKOVSKY POND, S. L., YOUNG, J. A. & SILVERMAN, G. J. 2012. In vivo VL-targeted microbial superantigen induced global shifts in the B cell repertoire. *J Immunol,* 189**,** 850-9.

GRUHN, B., TAUB, J. W., GE, Y. B., BECK, J. F., ZELL, R., HAFER, R., HERMANN, F. H., DEBATIN, K. M. & STEINBACH, D. 2009. Prenatal origin of childhood acute lymphoblastic leukemia, in Children of all age groups. *International Journal of Molecular Medicine,* 24**,** S62-S62.

GRYSHCHENKO, I., HOFBAUER, S., STOECHER, M., DANIEL, P. T., STEURER, M., GAIGER, A., EIGENBERGER, K., GREIL, R. & TINHOFER, I. 2008. MDM2 SNP309 is associated with poor outcome in B-cell chronic lymphocytic leukemia. *J Clin Oncol,* 26**,** 2252-7.

HALLEK, M., CHESON, B. D., CATOVSKY, D., CALIGARIS-CAPPIO, F., DIGHIERO, G., DOHNER, H., HILLMEN, P., KEATING, M. J., MONTSERRAT, E., RAI, K. R., KIPPS, T. J. & INTERNATIONAL WORKSHOP ON CHRONIC LYMPHOCYTIC, L. 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**,** 5446-56.

HALLEK, M. & GERMAN, C. L. L. S. G. 2005. Chronic lymphocytic leukemia (CLL): first-line treatment. *Hematology Am Soc Hematol Educ Program***,** 285-91.

HAMBLIN, T. J., DAVIS, Z., GARDINER, A., OSCIER, D. G. & STEVENSON, F. K. 1999. Unmutated Ig V(H) genes are associated with a more aggressive form of chronic lymphocytic leukemia. *Blood,* 94**,** 1848-54.

HARDIANTI, M. S., TATSUMI, E., SYAMPURNAWATI, M., FURUTA, K., SUZUKI, A., SAIGO, K., KAWANO, S., TAKENOKUCHI, M., KUMAGAI, S., MATSUO, Y., KOIZUMI, T. & TAKEUCHI, M. 2005. Presence of somatic hypermutation and activation-induced cytidine deaminase in acute lymphoblastic leukemia L2 with t(14;18)(q32;q21). *Eur J Haematol,* 74**,** 11-9.

HARDY, R. R. & HAYAKAWA, K. 2001. B cell development pathways. *Annual review of immunology,* 19**,** 595-621.

HARDY, R. R., KINCADE, P. W. & DORSHKIND, K. 2007. The protean nature of cells in the B lymphocyte lineage. *Immunity,* 26**,** 703-14.

HASLE, H., CLEMMENSEN, I. H. & MIKKELSEN, M. 2000. Risks of leukaemia and solid tumours in individuals with Down's syndrome. *Lancet,* 355**,** 165-9.

HAVELANGE, V., PEKARSKY, Y., NAKAMURA, T., PALAMARCHUK, A., ALDER, H., RASSENTI, L., KIPPS, T. & CROCE, C. M. 2011. IRF4 mutations in chronic lymphocytic leukemia. *Blood,* 118**,** 2827-9.

HAYES, G. M., BUSCH, R., VOOGT, J., SIAH, I. M., GEE, T. A., HELLERSTEIN, M. K., CHIORAZZI, N., RAI, K. R. & MURPHY, E. J. 2010. Isolation of malignant B cells from patients with chronic lymphocytic leukemia (CLL) for analysis of cell proliferation: Validation of a simplified method suitable for multi-center clinical studies. *Leukemia Research,* 34**,** 809-815.

HERVE, M., XU, K., NG, Y. S., WARDEMANN, H., ALBESIANO, E., MESSMER, B. T., CHIORAZZI, N. & MEFFRE, E. 2005a. Unmutated and mutated chronic lymphocytic leukemias derive from self-reactive B cell precursors despite expressing different antibody reactivity. *J Clin Invest,* 115**,** 1636-43.

HERVE, M., XU, K., NG, Y. S., WARDEMANN, H., ALBESIANO, E., MESSMER, B. T., CHIORAZZI, N. & MEFFRE, E. 2005b. Unmutated and mutated chronic lymphocytic leukemias derive from self-reactive B cell precursors despite expressing different antibody reactivity. *The Journal of clinical investigation,* 115**,** 1636-43.

HOFFMANN, R., SEIDL, T., NEEB, M., ROLINK, A. & MELCHERS, F. 2002. Changes in gene expression profiles in developing B cells of murine bone marrow. *Genome Res,* 12**,** 98-111.

HOI, K. H. & IPPOLITO, G. C. 2013. Intrinsic bias and public rearrangements in the human immunoglobulin Vlambda light chain repertoire. *Genes Immun,* 14**,** 271-6.

HONJO, T., KINOSHITA, K. & MURAMATSU, M. 2002. Molecular mechanism of class switch recombination: linkage with somatic hypermutation. *Annual review of immunology,* 20**,** 165-96.

HSU, M. C., TOELLNER, K. M., VINUESA, C. G. & MACLENNAN, I. C. 2006. B cell clones that sustain long-term plasmablast growth in T-independent extrafollicular antibody responses. *Proc Natl Acad Sci U S A,* 103**,** 5905-10.

HU, Y., TURNER, M. J., SHIELDS, J., GALE, M. S., HUTTO, E., ROBERTS, B. L., SIDERS, W. M. & KAPLAN, J. M. 2009. Investigation of the mechanism of action of alemtuzumab in a human CD52 transgenic mouse model. *Immunology,* 128**,** 260-70.

HUDSON, R. P. & WILSON, S. J. 1960. Hypogammaglobulinemia and chronic lymphatic leukemia. *Cancer,* 13**,** 200-4.

HUGHES, W. T., RIVERA, G. K., SCHELL, M. J., THORNTON, D. & LOTT, L. 1987. Successful intermittent chemoprophylaxis for Pneumocystis carinii pneumonitis. *N Engl J Med,* 316**,** 1627-32.

HUH, Y. O. & IBRAHIM, S. 2000. Immunophenotypes in adult acute lymphocytic leukemia. Role of flow cytometry in diagnosis and monitoring of disease. *Hematol Oncol Clin North Am,* 14**,** 1251-65.

IACOBUCCI, I., LONETTI, A., MESSA, F., FERRARI, A., CILLONI, D., SOVERINI, S., PAOLONI, F., ARRUGA, F., OTTAVIANI, E., CHIARETTI, S., MESSINA, M., VIGNETTI, M., PAPAYANNIDIS, C., VITALE, A., PANE, F., PICCALUGA, P. P., PAOLINI, S., BERTON, G., BARUZZI, A., SAGLIO, G., BACCARANI, M., FOA, R. & MARTINELLI, G. 2010. Different isoforms of the B-cell mutator activation-induced cytidine deaminase are aberrantly expressed in BCR-ABL1-positive acute lymphoblastic leukemia patients. *Leukemia,* 24**,** 66-73.

INABA, H. & PUI, C. H. 2010. Glucocorticoid use in acute lymphoblastic leukaemia. *Lancet Oncol,* 11**,** 1096-106.

INAMDAR, K. V. & BUESO-RAMOS, C. E. 2007. Pathology of chronic lymphocytic leukemia: an update. *Ann Diagn Pathol,* 11**,** 363-89.

INTHAL, A., ZEITLHOFER, P., ZEGINIGG, M., MORAK, M., GRAUSENBURGER, R., FRONKOVA, E., FAHRNER, B., MANN, G., HAAS, O. A. & PANZER-GRUMAYER, R. 2012. CREBBP HAT domain mutations prevail in relapse cases of high hyperdiploid childhood acute lymphoblastic leukemia. *Leukemia,* 26**,** 1797-803.

IPPOLITO, G. C., HOI, K. H., REDDY, S. T., CARROLL, S. M., GE, X., ROGOSCH, T., ZEMLIN, M., SHULTZ, L. D., ELLINGTON, A. D., VANDENBERG, C. L. & GEORGIOU, G. 2012. Antibody repertoires in humanized NOD-scid-IL2Rgamma(null) mice and human B cells reveals human-like diversification and tolerance checkpoints in the mouse. *PLoS One,* 7**,** e35497.

IWASHIMA, M., IRVING, B. A., VAN OERS, N. S., CHAN, A. C. & WEISS, A. 1994. Sequential interactions of the TCR with two distinct cytoplasmic tyrosine kinases. *Science,* 263**,** 1136-9.

JACKSON, K. J., KIDD, M. J., WANG, Y. & COLLINS, A. M. 2013. The Shape of the Lymphocyte Receptor Repertoire: Lessons from the B Cell Receptor. *Front Immunol,* 4**,** 263.

JAGER, U., FRIDRIK, M., ZEITLINGER, M., HEINTEL, D., HOPFINGER, G., BURGSTALLER, S., MANNHALTER, C., OBERAIGNER, W., PORPACZY, E., SKRABS, C., EINBERGER, C., DRACH, J., RADERER, M., GAIGER, A., PUTMAN, M. & GREIL, R. 2012. Rituximab serum concentrations during immuno-chemotherapy of follicular lymphoma correlate with patient gender, bone marrow infiltration and clinical response. *Haematologica,* 97**,** 1431-8.

JIANG, N., HE, J., WEINSTEIN, J. A., PENLAND, L., SASAKI, S., HE, X. S., DEKKER, C. L., ZHENG, N. Y., HUANG, M., SULLIVAN, M., WILSON, P. C., GREENBERG, H. B., DAVIS, M. M., FISHER, D. S. & QUAKE, S. R. 2013. Lineage structure of the human antibody repertoire in response to influenza vaccination. *Sci Transl Med,* 5**,** 171ra19.

JIANG, N., WEINSTEIN, J. A., PENLAND, L., WHITE, R. A., 3RD, FISHER, D. S. & QUAKE, S. R. 2011. Determinism and stochasticity during maturation of the zebrafish antibody repertoire. *Proc Natl Acad Sci U S A,* 108**,** 5348-53.

JIAO, W., VEMBU, S., DESHWAR, A. G., STEIN, L. & MORRIS, Q. 2014. Inferring clonal evolution of tumors from single nucleotide somatic mutations. *BMC Bioinformatics,* 15**,** 35.

JOHNSON, S., SMITH, A. G., LOFFLER, H., OSBY, E., JULIUSSON, G., EMMERICH, B., WYLD, P. J. & HIDDEMANN, W. 1996. Multicentre prospective randomised trial of fludarabine versus cyclophosphamide, doxorubicin, and prednisone (CAP) for treatment of advanced-stage chronic lymphocytic leukaemia. The French Cooperative Group on CLL. *Lancet,* 347**,** 1432-8.

JULIUSSON, G., OSCIER, D. G., FITCHETT, M., ROSS, F. M., STOCKDILL, G., MACKIE, M. J., PARKER, A. C., CASTOLDI, G. L., GUNEO, A., KNUUTILA, S., ELONEN, E. & GAHRTON, G. 1990. Prognostic subgroups in B-cell chronic lymphocytic leukemia defined by specific chromosomal abnormalities. *N Engl J Med,* 323**,** 720-4.

JUNEMANN, S., SEDLAZECK, F. J., PRIOR, K., ALBERSMEIER, A., JOHN, U., KALINOWSKI, J., MELLMANN, A., GOESMANN, A., VON HAESELER, A., STOYE, J. & HARMSEN, D. 2013. Updating benchtop sequencing performance comparison. *Nat Biotechnol,* 31**,** 294-6.

KALININA, O., DOYLE-COOPER, C. M., MIKSANEK, J., MENG, W., PRAK, E. L. & WEIGERT, M. G. 2011. Alternative mechanisms of receptor editing in autoreactive B cells. *Proc Natl Acad Sci U S A,* 108**,** 7125-30.

KALLED, S. L. & BRODEUR, P. H. 1990. Preferential rearrangement of V kappa 4 gene segments in pre-B cell lines. *J Exp Med,* 172**,** 559-66.

KANTARJIAN, H., THOMAS, D., O'BRIEN, S., CORTES, J., GILES, F., JEHA, S., BUESO-RAMOS, C. E., PIERCE, S., SHAN, J., KOLLER, C., BERAN, M., KEATING, M. & FREIREICH, E. J. 2004. Long-term follow-up results of hyperfractionated cyclophosphamide, vincristine, doxorubicin, and dexamethasone (Hyper-CVAD), a dose-intensive regimen, in adult acute lymphocytic leukemia. *Cancer,* 101**,** 2788-801.

KATAYAMA, Y., SAKAI, A., KATSUTANI, S., TAKIMOTO, Y. & KIMURA, A. 2001. Lack of allelic exclusion and isotype switching in B cell chronic lymphocytic leukemia. *Am J Hematol,* 68**,** 295-7.

KATOH, K. & STANDLEY, D. M. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol,* 30**,** 772-80.

KELLY, D. F., SNAPE, M. D., CLUTTERBUCK, E. A., GREEN, S., SNOWDEN, C., DIGGLE, L., YU, L. M., BORKOWSKI, A., MOXON, E. R. & POLLARD, A. J. 2006. CRM197-conjugated serogroup C meningococcal capsular polysaccharide, but not the native polysaccharide, induces persistent antigen-specific memory B cells. *Blood,* 108**,** 2642-7.

KELLY, L. M., YU, J. C., BOULTON, C. L., APATIRA, M., LI, J., SULLIVAN, C. M., WILLIAMS, I., AMARAL, S. M., CURLEY, D. P., DUCLOS, N., NEUBERG, D., SCARBOROUGH, R. M., PANDEY, A., HOLLENBACH, S., ABE, K., LOKKER, N. A., GILLILAND, D. G. & GIESE, N. A. 2002. CT53518, a novel selective FLT3 antagonist for the treatment of acute myelogenous leukemia (AML). *Cancer cell,* 1**,** 421-32.

KEPLER, T. B. 2013. Reconstructing a B-cell clonal lineage. I. Statistical inference of unobserved ancestors. *F1000Res,* 2**,** 103.

KEPLER, T. B., MUNSHAW, S., WIEHE, K., ZHANG, R., YU, J. S., WOODS, C. W., DENNY, T. N., TOMARAS, G. D., ALAM, S. M., MOODY, M. A., KELSOE, G., LIAO, H. X. & HAYNES, B. F. 2014. Reconstructing a B-Cell Clonal Lineage. II. Mutation, Selection, and Affinity Maturation. *Front Immunol,* 5**,** 170.

KERN, W., BACHER, U., SCHNITTGER, S., DICKER, F., ALPERMANN, T., HAFERLACH, T. & HAFERLACH, C. 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**,** 565-9.

KILO, M. N. & DORFMAN, D. M. 1996. The utility of flow cytometric immunophenotypic analysis in the distinction of small lymphocytic lymphoma/chronic lymphocytic leukemia from mantle cell lymphoma. *American journal of clinical pathology,* 105**,** 451-7.

KINOSHITA, K. & HONJO, T. 2001. Linking class-switch recombination with somatic hypermutation. *Nat Rev Mol Cell Biol,* 2**,** 493-503.

KIPPS, T. J., TOMHAVE, E., PRATT, L. F., DUFFY, S., CHEN, P. P. & CARSON, D. A. 1989. Developmentally restricted immunoglobulin heavy chain variable region gene expressed at high frequency in chronic lymphocytic leukemia. *Proceedings of the National Academy of Sciences of the United States of America,* 86**,** 5913-7.

KITAMURA, D. & RAJEWSKY, K. 1992. Targeted disruption of mu chain membrane exon causes loss of heavy-chain allelic exclusion. *Nature,* 356**,** 154-6.

KITCHINGMAN, G. R., MIRRO, J., STASS, S., ROVIGATTI, U., MELVIN, S. L., WILLIAMS, D. L., RAIMONDI, S. C. & MURPHY, S. B. 1986. Biologic and prognostic significance of the presence of more than two mu heavy-chain genes in childhood acute lymphoblastic leukemia of B precursor cell origin. *Blood,* 67**,** 698-703.

KLEIN, U., TU, Y., STOLOVITZKY, G. A., MATTIOLI, M., CATTORETTI, G., HUSSON, H., FREEDMAN, A., INGHIRAMI, G., CRO, L., BALDINI, L., NERI, A., CALIFANO, A. & DALLA-FAVERA, R. 2001. Gene expression profiling of B cell chronic lymphocytic leukemia reveals a homogeneous phenotype related to memory B cells. *J Exp Med,* 194**,** 1625-38.

KOLIBAB, K., SMITHSON, S. L., RABQUER, B., KHUDER, S. & WESTERINK, M. A. 2005. Immune response to pneumococcal polysaccharides 4 and 14 in elderly and young adults: analysis of the variable heavy chain repertoire. *Infect Immun,* 73**,** 7465-76.

KOOPMAN, G., KEEHNEN, R. M., LINDHOUT, E., NEWMAN, W., SHIMIZU, Y., VAN SEVENTER, G. A., DE GROOT, C. & PALS, S. T. 1994. Adhesion through the LFA-1 (CD11a/CD18)-ICAM-1 (CD54) and the VLA-4 (CD49d)-VCAM-1 (CD106) pathways prevents apoptosis of germinal center B cells. *J Immunol,* 152**,** 3760-7.

KORN, T., BETTELLI, E., OUKKA, M. & KUCHROO, V. K. 2009. Il-17 and Th17 Cells. *Annual Review of Immunology,* 27**,** 485-517.

KRAUSE, J. C., TSIBANE, T., TUMPEY, T. M., HUFFMAN, C. J., BRINEY, B. S., SMITH, S. A., BASLER, C. F. & CROWE, J. E., JR. 2011. Epitope-specific human influenza antibody repertoires diversify by B cell intraclonal sequence divergence and interclonal convergence. *Journal of immunology,* 187**,** 3704-11.

KROBER, A., BLOEHDORN, J., HAFNER, S., BUHLER, A., SEILER, T., KIENLE, D., WINKLER, D., BANGERTER, M., SCHLENK, R. F., BENNER, A., LICHTER, P., DOHNER, H. & STILGENBAUER, S. 2006. Additional genetic high-risk features such as 11q deletion, 17p deletion, and V3-21 usage characterize discordance of ZAP-70 and VH mutation status in chronic lymphocytic leukemia. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology,* 24**,** 969-75.

KUIPER, R. P., WAANDERS, E., VAN DER VELDEN, V. H., VAN REIJMERSDAL, S. V., VENKATACHALAM, R., SCHEIJEN, B., SONNEVELD, E., VAN DONGEN, J. J., VEERMAN, A. J., VAN LEEUWEN, F. N., VAN KESSEL, A. G. & HOOGERBRUGGE, P. M. 2010. IKZF1 deletions predict relapse in uniformly treated pediatric precursor B-ALL. *Leukemia,* 24**,** 1258-64.

KUMAGAI, M., COUSTAN-SMITH, E., MURRAY, D. J., SILVENNOINEN, O., MURTI, K. G., EVANS, W. E., MALAVASI, F. & CAMPANA, D. 1995. Ligation of CD38 suppresses human B lymphopoiesis. *The Journal of experimental medicine,* 181**,** 1101-10.

KUPPERS, R., SOUSA, A. B., BAUR, A. S., STRICKLER, J. G., RAJEWSKY, K. & HANSMANN, M. L. 2001. Common germinal-center B-cell origin of the malignant cells in two composite

lymphomas, involving classical Hodgkin's disease and either follicular lymphoma or B-CLL. *Mol Med,* 7**,** 285-92.

LADETTO, M., BRUGGEMANN, M., MONITILLO, L., FERRERO, S., PEPIN, F., DRANDI, D., BARBERO, D., PALUMBO, A., PASSERA, R., BOCCADORO, M., RITGEN, M., GOKBUGET, N., ZHENG, J., CARLTON, V., TRAUTMANN, H., FAHAM, M. & POTT, C. 2013. Next-generation sequencing and real-time quantitative PCR for minimal residual disease detection in B-cell disorders. *Leukemia*.

LAMM, M. E. 1997. Interaction of antigens and antibodies at mucosal surfaces. *Annu Rev Microbiol,* 51**,** 311-40.

LANDAU, D. A., CARTER, S. L., STOJANOV, P., MCKENNA, A., STEVENSON, K., LAWRENCE, M. S., SOUGNEZ, C., STEWART, C., SIVACHENKO, A., WANG, L., WAN, Y., ZHANG, W., SHUKLA, S. A., VARTANOV, A., FERNANDES, S. M., SAKSENA, G., CIBULSKIS, K., TESAR, B., GABRIEL, S., HACOHEN, N., MEYERSON, M., LANDER, E. S., NEUBERG, D., BROWN, J. R., GETZ, G. & WU, C. J. 2013. Evolution and impact of subclonal mutations in chronic lymphocytic leukemia. *Cell,* 152**,** 714-26.

LANGERAK, A. W., DAVI, F., GHIA, P., HADZIDIMITRIOU, A., MURRAY, F., POTTER, K. N., ROSENQUIST, R., STAMATOPOULOS, K., BELESSI, C. & EUROPEAN RESEARCH INITIATIVE ON, C. L. L. 2011. Immunoglobulin sequence analysis and prognostication in CLL: guidelines from the ERIC review board for reliable interpretation of problematic cases. *Leukemia,* 25**,** 979-84.

LANGERAK, A. W. & DONGEN, J. J. M. 2011. Multiple clonal Ig/TCR products: implications for interpretation of clonality findings. *Journal of Hematopathology,* 5**,** 35-43.

LARIMORE, K., MCCORMICK, M. W., ROBINS, H. S. & GREENBERG, P. D. 2012. Shaping of human germline IgH repertoires revealed by deep sequencing. *J Immunol,* 189**,** 3221-30.

LARSEN, P. A. & SMITH, T. P. 2012. Application of circular consensus sequencing and network analysis to characterize the bovine IgG repertoire. *BMC Immunol,* 13**,** 52.

LARSON, R. A., DODGE, R. K., BURNS, C. P., LEE, E. J., STONE, R. M., SCHULMAN, P., DUGGAN, D., DAVEY, F. R., SOBOL, R. E., FRANKEL, S. R. & ET AL. 1995. A five-drug remission induction regimen with intensive consolidation for adults with acute lymphoblastic leukemia: cancer and leukemia group B study 8811. *Blood,* 85**,** 2025-37.

LASERSON, U., VIGNEAULT, F., GADALA-MARIA, D., YAARI, G., UDUMAN, M., VANDER HEIDEN, J. A., KELTON, W., TAEK JUNG, S., LIU, Y., LASERSON, J., CHARI, R., LEE, J. H., BACHELET, I., HICKEY, B., LIEBERMAN-AIDEN, E., HANCZARUK, B., SIMEN, B. B., EGHOLM, M., KOLLER, D., GEORGIOU, G., KLEINSTEIN, S. H. & CHURCH, G. M. 2014. High-resolution antibody dynamics of vaccine-induced immune responses. *Proc Natl Acad Sci U S A,* 111**,** 4928-33.

LATCHMAN, D. 2005. Gene Regulation (Advanced Texts)

LAVINDER, J. J., WINE, Y., GIESECKE, C., IPPOLITO, G. C., HORTON, A. P., LUNGU, O. I., HOI, K. H., DEKOSKY, B. J., MURRIN, E. M., WIRTH, M. M., ELLINGTON, A. D., DORNER, T., MARCOTTE, E. M., BOUTZ, D. R. & GEORGIOU, G. 2014. Identification and characterization of the constituent human serum antibodies elicited by vaccination. *Proc Natl Acad Sci U S A,* 111**,** 2259-64.

LEBIEN, T. W. & TEDDER, T. F. 2008. B lymphocytes: how they develop and function. *Blood,* 112**,** 1570-80.

LEFRANC, M. P., GIUDICELLI, V., GINESTOUX, C., JABADO-MICHALOUD, J., FOLCH, G., BELLAHCENE, F., WU, Y., GEMROT, E., BROCHET, X., LANE, J., REGNIER, L., EHRENMANN, F., LEFRANC, G. & DUROUX, P. 2009. IMGT, the international ImMunoGeneTics information system. *Nucleic acids research,* 37**,** D1006-12.

LEV, A., SIMON, A. J., BAREKET, M., BIELORAI, B., HUTT, D., AMARIGLIO, N., RECHAVI, G. & SOMECH, R. 2012. The kinetics of early T and B cell immune recovery after bone marrow transplantation in RAG-2-deficient SCID patients. *PloS one,* 7**,** e30494.

LI, H., YE, C., JI, G. & HAN, J. 2012. Determinants of public T cell responses. *Cell Res,* 22**,** 33-42.

LI, W. & GODZIK, A. 2006. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics,* 22**,** 1658-9.

LIAO, H. X., LYNCH, R., ZHOU, T., GAO, F., ALAM, S. M., BOYD, S. D., FIRE, A. Z., ROSKIN, K. M., SCHRAMM, C. A., ZHANG, Z., ZHU, J., SHAPIRO, L., PROGRAM, N. C. S., MULLIKIN, J. C., GNANAKARAN, S., HRABER, P., WIEHE, K., KELSOE, G., YANG, G., XIA, S. M., MONTEFIORI, D. C., PARKS, R., LLOYD, K. E., SCEARCE, R. M., SODERBERG, K. A., COHEN, M., KAMANGA, G., LOUDER, M. K., TRAN, L. M., CHEN, Y., CAI, F., CHEN, S., MOQUIN, S., DU, X., JOYCE, M. G., SRIVATSAN, S., ZHANG, B., ZHENG, A., SHAW, G. M., HAHN, B. H., KEPLER, T. B., KORBER, B. T., KWONG, P. D., MASCOLA, J. R. & HAYNES, B. F. 2013. Co-evolution of a broadly neutralizing HIV-1 antibody and founder virus. *Nature,* 496**,** 469-76.

LIN, M. M., ZHU, M. & SCHARFF, M. D. 1997. Sequence dependent hypermutation of the immunoglobulin heavy chain in cultured B cells. *Proceedings of the National Academy of Sciences of the United States of America,* 94**,** 5284-9.

LIU, H. & MAY, K. 2012. Disulfide bond structures of IgG molecules: structural variations, chemical modifications and possible impacts to stability and biological function. *MAbs,* 4**,** 17-23.

LIU, J., LANGE, M. D., HONG, S. Y., XIE, W., XU, K., HUANG, L., YU, Y., EHRHARDT, G. R., ZEMLIN, M., BURROWS, P. D., SU, K., CARTER, R. H. & ZHANG, Z. 2013. Regulation of VH replacement by B cell receptor-mediated signaling in human immature B cells. *J Immunol,* 190**,** 5559-66.

LIU, Y. J., BARTHELEMY, C., DE BOUTEILLER, O., ARPIN, C., DURAND, I. & BANCHEREAU, J. 1995. Memory B cells from human tonsils colonize mucosal epithelium and directly present antigen to T cells by rapid up-regulation of B7-1 and B7-2. *Immunity,* 2**,** 239-48.

LIU, Y. J., OLDFIELD, S. & MACLENNAN, I. C. 1988. Memory B cells in T cell-dependent antibody responses colonize the splenic marginal zones. *Eur J Immunol,* 18**,** 355-62.

LLOYD, C., LOWE, D., EDWARDS, B., WELSH, F., DILKS, T., HARDMAN, C. & VAUGHAN, T. 2009. Modelling the human immune response: performance of a 1011 human antibody repertoire against a broad panel of therapeutically relevant antigens. *Protein Eng Des Sel,* 22**,** 159-68.

LOGAN, A. C., GAO, H., WANG, C., SAHAF, B., JONES, C. D., MARSHALL, E. L., BUNO, I., ARMSTRONG, R., FIRE, A. Z., WEINBERG, K. I., MINDRINOS, M., ZEHNDER, J. L., BOYD, S. D., XIAO, W., DAVIS, R. W. & MIKLOS, D. B. 2011. High-throughput VDJ sequencing for quantification of minimal residual disease in chronic lymphocytic leukemia and immune reconstitution assessment. *Proceedings of the National Academy of Sciences of the United States of America,* 108**,** 21194-9.

LOMAN, N. J., MISRA, R. V., DALLMAN, T. J., CONSTANTINIDOU, C., GHARBIA, S. E., WAIN, J. & PALLEN, M. J. 2012. Performance comparison of benchtop high-throughput sequencing platforms. *Nat Biotechnol,* 30**,** 434-9.

LOU, D. I., HUSSMANN, J. A., MCBEE, R. M., ACEVEDO, A., ANDINO, R., PRESS, W. H. & SAWYER, S. L. 2013. High-throughput DNA sequencing errors are reduced by orders of magnitude using circle sequencing. *Proc Natl Acad Sci U S A,* 110**,** 19872-7.

LUKOWSKY, A., MARCHWAT, M., STERRY, W. & GELLRICH, S. 2006. Evaluation of B-cell clonality in archival skin biopsy samples of cutaneous B-cell lymphoma by

immunoglobulin heavy chain gene polymerase chain reaction. *Leukemia & lymphoma,* 47**,** 487-93.

LUO, C., TSEMENTZI, D., KYRPIDES, N., READ, T. & KONSTANTINIDIS, K. T. 2012. Direct Comparisons of Illumina vs. Roche 454 Sequencing Technologies on the Same Microbial Community DNA Sample. *PloS one,* 7**,** e30087.

LUTZNY, G., KOCHER, T., SCHMIDT-SUPPRIAN, M., RUDELIUS, M., KLEIN-HITPASS, L., FINCH, A. J., DURIG, J., WAGNER, M., HAFERLACH, C., KOHLMANN, A., SCHNITTGER, S., SEIFERT, M., WANNINGER, S., ZABORSKY, N., OOSTENDORP, R., RULAND, J., LEITGES, M., KUHNT, T., SCHAFER, Y., LAMPL, B., PESCHEL, C., EGLE, A. & RINGSHAUSEN, I. 2013. Protein Kinase C-beta-Dependent Activation of NF-kappa B in Stromal Cells Is Indispensable for the Survival of Chronic Lymphocytic Leukemia B Cells In Vivo. *Cancer Cell,* 23**,** 77-92.

LYDYARD, P. M., WHELAN, A. & FANGER, M. W. 2000. Instant Notes Series; Instant Notes in Immunology. i-x, 1-318.

MA, S. K., CHAN, G. C., HA, S. Y., CHIU, D. C., LAU, Y. L. & CHAN, L. C. 1997. Clinical presentation, hematologic features and treatment outcome of childhood acute lymphoblastic leukemia: a review of 73 cases in Hong Kong. *Hematol Oncol,* 15**,** 141-9.

MACKAY, F., SIERRO, F., GREY, S. T. & GORDON, T. P. 2005. The BAFF/APRIL system: an important player in systemic rheumatic diseases. *Curr Dir Autoimmun,* 8**,** 243-65.

MALETZKI, C., JAHNKE, A., OSTWALD, C., KLAR, E., PRALL, F. & LINNEBACHER, M. 2012. Ex-vivo clonally expanded B lymphocytes infiltrating colorectal carcinoma are of mature immunophenotype and produce functional IgG. *PloS one,* 7**,** e32639.

MAMANOVA, L., COFFEY, A. J., SCOTT, C. E., KOZAREWA, I., TURNER, E. H., KUMAR, A., HOWARD, E., SHENDURE, J. & TURNER, D. J. 2010. Target-enrichment strategies for next-generation sequencing (vol 7, pg 111, 2010). *Nature Methods,* 7**,** 479-479.

MANIS, J. P., TIAN, M. & ALT, F. W. 2002. Mechanism and control of class-switch recombination. *Trends in immunology,* 23**,** 31-9.

MANSKE, M. K., ZUCKERMAN, N. S., TIMM, M. M., MAIDEN, S., EDELMAN, H., SHAHAF, G., BARAK, M., DISPENZIERI, A., GERTZ, M. A., MEHR, R. & ABRAHAM, R. S. 2006. Quantitative analysis of clonal bone marrow CD19+ B cells: use of B cell lineage trees to delineate their role in the pathogenesis of light chain amyloidosis. *Clin Immunol,* 120**,** 106-20.

MANZ, R. A., THIEL, A. & RADBRUCH, A. 1997. Lifetime of plasma cells in the bone marrow. *Nature,* 388**,** 133-4.

MAO, Z. R., QUINTANILLA-MARTINEZ, L., RAFFELD, M., RICHTER, M., KRUGMANN, J., BUREK, C., HARTMANN, E., RUDIGER, T., JAFFE, E. S., MULLER-HERMELINK, H. K., OTT, G., FEND, F. & ROSENWALD, A. 2007. IgVH mutational status and clonality analysis of Richter's transformation - Diffuse large B-cell lymphoma and Hodgkin lymphoma in association with B-cell chronic lymphocytic leukemia (B-CLL) represent 2 different pathways of disease evolution. *American Journal of Surgical Pathology,* 31**,** 1605-1614.

MARAFIOTI, T., HUMMEL, M., ANAGNOSTOPOULOS, I., FOSS, H. D., HUHN, D. & STEIN, H. 1999. Classical Hodgkin's disease and follicular lymphoma originating from the same germinal center B cell. *J Clin Oncol,* 17**,** 3804-9.

MARGULIES, M., EGHOLM, M., ALTMAN, W. E., ATTIYA, S., BADER, J. S., BEMBEN, L. A., BERKA, J., BRAVERMAN, M. S., CHEN, Y. J., CHEN, Z., DEWELL, S. B., DU, L., FIERRO, J. M., GOMES, X. V., GODWIN, B. C., HE, W., HELGESEN, S., HO, C. H., IRZYK, G. P., JANDO, S. C., ALENQUER, M. L., JARVIE, T. P., JIRAGE, K. B., KIM, J. B., KNIGHT, J. R., LANZA, J. R., LEAMON, J. H., LEFKOWITZ, S. M., LEI, M., LI, J., LOHMAN, K. L., LU, H., MAKHIJANI, V. B., MCDADE, K. E., MCKENNA, M. P., MYERS, E. W., NICKERSON, E.,

NOBILE, J. R., PLANT, R., PUC, B. P., RONAN, M. T., ROTH, G. T., SARKIS, G. J., SIMONS, J. F., SIMPSON, J. W., SRINIVASAN, M., TARTARO, K. R., TOMASZ, A., VOGT, K. A., VOLKMER, G. A., WANG, S. H., WANG, Y., WEINER, M. P., YU, P., BEGLEY, R. F. & ROTHBERG, J. M. 2005. Genome sequencing in microfabricated high-density picolitre reactors. *Nature,* 437**,** 376-80.

MAROTTA, G., BIGAZZI, C., LENOCI, M., TOZZI, M., BOCCHIA, M. & LAURIA, F. 2000. Low-dose fludarabine and cyclophosphamide in elderly patients with B-cell chronic lymphocytic leukemia refractory to conventional therapy. *Haematologica,* 85**,** 1268-70.

MARSHALL, G. M., KWAN, E., HABER, M., BRISCO, M. J., SYKES, P. J., MORLEY, A. A., TOOGOOD, I., WATERS, K., TAURO, G., EKERT, H. & ET AL. 1995. Characterization of clonal immunoglobulin heavy chain and I cell receptor gamma gene rearrangements during progression of childhood acute lymphoblastic leukemia. *Leukemia,* 9**,** 1847-50.

MARTI, G., ABBASI, F., RAVECHE, E., RAWSTRON, A. C., GHIA, P., AURRAN, T., CAPORASO, N., SHIM, Y. K. & VOGT, R. F. 2007. Overview of monoclonal B-cell lymphocytosis. *Br J Haematol,* 139**,** 701-8.

MARTI, G. E., RAWSTRON, A. C., GHIA, P., HILLMEN, P., HOULSTON, R. S., KAY, N., SCHLEINITZ, T. A., CAPORASO, N. & INTERNATIONAL FAMILIAL, C. L. L. C. 2005. Diagnostic criteria for monoclonal B-cell lymphocytosis. *Br J Haematol,* 130**,** 325-32.

MARTIN, S. W. & GOODNOW, C. C. 2002. Burst-enhancing role of the IgG membrane tail as a molecular determinant of memory. *Nat Immunol,* 3**,** 182-8.

MARTIN, T., DUFFY, S. F., CARSON, D. A. & KIPPS, T. J. 1992. Evidence for somatic selection of natural autoantibodies. *The Journal of experimental medicine,* 175**,** 983-91.

MARTINS, E. P. & HOUSWORTH, E. A. 2002. Phylogeny shape and the phylogenetic comparative method. *Syst Biol,* 51**,** 873-80.

MASSON, E., SYNOLD, T. W., RELLING, M. V., SCHUETZ, J. D., SANDLUND, J. T., PUI, C. H. & EVANS, W. E. 1996. Allopurinol inhibits de novo purine synthesis in lymphoblasts of children with acute lymphoblastic leukemia. *Leukemia,* 10**,** 56-60.

MATSUDA, F., ISHII, K., BOURVAGNET, P., KUMA, K., HAYASHIDA, H., MIYATA, T. & HONJO, T. 1998. The complete nucleotide sequence of the human immunoglobulin heavy chain variable region locus. *J Exp Med,* 188**,** 2151-62.

MAURO, F. R., FOA, R., GIANNARELLI, D., CORDONE, I., CRESCENZI, S., PESCARMONA, E., SALA, R., CERRETTI, R. & MANDELLI, F. 1999. Clinical characteristics and outcome of young chronic lymphocytic leukemia patients: a single institution study of 204 cases. *Blood,* 94**,** 448-54.

MAZANEC, M. B., NEDRUD, J. G., KAETZEL, C. S. & LAMM, M. E. 1993. A three-tiered view of the role of IgA in mucosal defense. *Immunol Today,* 14**,** 430-5.

MCHEYZER-WILLIAMS, L. J. & MCHEYZER-WILLIAMS, M. G. 2005. Antigen-specific memory B cell development. *Annu Rev Immunol,* 23**,** 487-513.

MCHEYZER-WILLIAMS, M., OKITSU, S., WANG, N. & MCHEYZER-WILLIAMS, L. 2012. Molecular programming of B cell memory. *Nat Rev Immunol,* 12**,** 24-34.

MCINTYRE, D., ZUCKERMAN, N. S., FIELD, M., MEHR, R. & STOTT, D. I. 2014. The V(H) repertoire and clonal diversification of B cells in inflammatory myopathies. *Eur J Immunol,* 44**,** 585-96.

MERCOLINO, T. J., ARNOLD, L. W., HAWKINS, L. A. & HAUGHTON, G. 1988. Normal mouse peritoneum contains a large population of Ly-1+ (CD5) B cells that recognize phosphatidyl choline. Relationship to cells that secrete hemolytic antibody specific for autologous erythrocytes. *The Journal of experimental medicine,* 168**,** 687-98.

MESSINA, M., CHIARETTI, S., IACOBUCCI, I., TAVOLARO, S., LONETTI, A., SANTANGELO, S., ELIA, L., PAPAYANNIDIS, C., PAOLONI, F., VITALE, A., GUARINI, A., MARTINELLI, G. &

FOA, R. 2011. AICDA expression in BCR/ABL1-positive acute lymphoblastic leukaemia is associated with a peculiar gene expression profile. *Br J Haematol,* 152**,** 727-32.

MESSMER, B. T., ALBESIANO, E., EFREMOV, D. G., GHIOTTO, F., ALLEN, S. L., KOLITZ, J., FOA, R., DAMLE, R. N., FAIS, F., MESSMER, D., RAI, K. R., FERRARINI, M. & CHIORAZZI, N. 2004. Multiple distinct sets of stereotyped antigen receptors indicate a role for antigen in promoting chronic lymphocytic leukemia. *J Exp Med,* 200**,** 519-25.

MESSMER, B. T., MESSMER, D., ALLEN, S. L., KOLITZ, J. E., KUDALKAR, P., CESAR, D., MURPHY, E. J., KODURU, P., FERRARINI, M., ZUPO, S., CUTRONA, G., DAMLE, R. N., WASIL, T., RAI, K. R., HELLERSTEIN, M. K. & CHIORAZZI, N. 2005a. In vivo measurements document the dynamic cellular kinetics of chronic lymphocytic leukemia B cells. *Journal of Clinical Investigation,* 115**,** 755-764.

MESSMER, B. T., MESSMER, D., ALLEN, S. L., KOLITZ, J. E., KUDALKAR, P., CESAR, D., MURPHY, E. J., KODURU, P., FERRARINI, M., ZUPO, S., CUTRONA, G., DAMLE, R. N., WASIL, T., RAI, K. R., HELLERSTEIN, M. K. & CHIORAZZI, N. 2005b. In vivo measurements document the dynamic cellular kinetics of chronic lymphocytic leukemia B cells. *J Clin Invest,* 115**,** 755-64.

MILNE, C. D. & PAIGE, C. J. 2006. IL-7: a key regulator of B lymphopoiesis. *Semin Immunol,* 18**,** 20-30.

MOHR, E., SERRE, K., MANZ, R. A., CUNNINGHAM, A. F., KHAN, M., HARDIE, D. L., BIRD, R. & MACLENNAN, I. C. 2009. Dendritic cells and monocyte/macrophages that create the IL-6/APRIL-rich lymph node microenvironments where plasmablasts mature. *J Immunol,* 182**,** 2113-23.

MOND, J. J., LEES, A. & SNAPPER, C. M. 1995. T cell-independent antigens type 2. *Annu Rev Immunol,* 13**,** 655-92.

MOORMAN, A. V., RICHARDS, S. M., ROBINSON, H. M., STREFFORD, J. C., GIBSON, B. E., KINSEY, S. E., EDEN, T. O., VORA, A. J., MITCHELL, C. D., HARRISON, C. J. & PARTY, U. K. M. R. C. N. C. R. I. C. L. W. 2007. Prognosis of children with acute lymphoblastic leukemia (ALL) and intrachromosomal amplification of chromosome 21 (iAMP21). *Blood,* 109**,** 2327-30.

MORABITO, F., DE FILIPPI, R., LAURENTI, L., ZIRLIK, K., RECCHIA, A. G., GENTILE, M., MORELLI, E., VIGNA, E., GIGLIOTTI, V., CALEMMA, R., AMOROSO, B., NERI, A., CUTRONA, G., FERRARINI, M., MOLICA, S., DEL POETA, G., TRIPODO, C. & PINTO, A. 2011. The cumulative amount of serum free light chain is a strong prognosticator in chronic lymphocytic leukemia. *Blood*.

MORETON, P., KENNEDY, B., LUCAS, G., LEACH, M., RASSAM, S. M., HAYNES, A., TIGHE, J., OSCIER, D., FEGAN, C., RAWSTRON, A. & HILLMEN, P. 2005. Eradication of minimal residual disease in B-cell chronic lymphocytic leukemia after alemtuzumab therapy is associated with prolonged survival. *J Clin Oncol,* 23**,** 2971-9.

MORRELL, D., CROMARTIE, E. & SWIFT, M. 1986. Mortality and cancer incidence in 263 patients with ataxia-telangiectasia. *J Natl Cancer Inst,* 77**,** 89-92.

MORROW, J. S. 1977. Toward a more normative assessment of maldistribution: the Gini Index. *Inquiry : a journal of medical care organization, provision and financing,* 14**,** 278-92.

MOUQUET, H. & NUSSENZWEIG, M. C. 2011. Polyreactive antibodies in adaptive immune responses to viruses. *Cellular and molecular life sciences : CMLS*.

MUELLENBECK, M. F., UEBERHEIDE, B., AMULIC, B., EPP, A., FENYO, D., BUSSE, C. E., ESEN, M., THEISEN, M., MORDMULLER, B. & WARDEMANN, H. 2013. Atypical and classical memory B cells produce Plasmodium falciparum neutralizing antibodies. *J Exp Med,* 210**,** 389-99.

MUHAMMAD, K., ROLL, P., EINSELE, H., DORNER, T. & TONY, H. P. 2009. Delayed acquisition of somatic hypermutations in repopulated IGD+CD27+ memory B cell receptors after rituximab treatment. *Arthritis Rheum,* 60**,** 2284-93.

MULLIGHAN, C. G. 2012. Molecular genetics of B-precursor acute lymphoblastic leukemia. *J Clin Invest,* 122**,** 3407-15.

MULLIGHAN, C. G., ZHANG, J., HARVEY, R. C., COLLINS-UNDERWOOD, J. R., SCHULMAN, B. A., PHILLIPS, L. A., TASIAN, S. K., LOH, M. L., SU, X., LIU, W., DEVIDAS, M., ATLAS, S. R., CHEN, I. M., CLIFFORD, R. J., GERHARD, D. S., CARROLL, W. L., REAMAN, G. H., SMITH, M., DOWNING, J. R., HUNGER, S. P. & WILLMAN, C. L. 2009. JAK mutations in high-risk childhood acute lymphoblastic leukemia. *Proc Natl Acad Sci U S A,* 106**,** 9414-8.

MURAMATSU, M., KINOSHITA, K., FAGARASAN, S., YAMADA, S., SHINKAI, Y. & HONJO, T. 2000. Class switch recombination and hypermutation require activation-induced cytidine deaminase (AID), a potential RNA editing enzyme. *Cell,* 102**,** 553-63.

MURTAGH, F. & CONTRERAS, P. 2012. Algorithms for hierarchical clustering: an overview. *Wiley Interdisciplinary Reviews-Data Mining and Knowledge Discovery,* 2**,** 86-97.

NAKAMURA, K., OSHIMA, T., MORIMOTO, T., IKEDA, S., YOSHIKAWA, H., SHIWA, Y., ISHIKAWA, S., LINAK, M. C., HIRAI, A., TAKAHASHI, H., ALTAF-UL-AMIN, M., OGASAWARA, N. & KANAYA, S. 2011. Sequence-specific error profile of Illumina sequencers. *Nucleic Acids Res,* 39**,** e90.

NEUMANN, M., HEESCH, S., GOKBUGET, N., SCHWARTZ, S., SCHLEE, C., BENLASFER, O., FARHADI-SARTANGI, N., THIBAUT, J., BURMEISTER, T., HOELZER, D., HOFMANN, W. K., THIEL, E. & BALDUS, C. D. 2012. Clinical and molecular characterization of early T-cell precursor leukemia: a high-risk subgroup in adult T-ALL with a high frequency of FLT3 mutations. *Blood Cancer J,* 2**,** e55.

OBUKHANYCH, T. V. & NUSSENZWEIG, M. C. 2006. T-independent type II immune responses generate memory B cells. *J Exp Med,* 203**,** 305-10.

OKAZAKI, I. M., KINOSHITA, K., MURAMATSU, M., YOSHIKAWA, K. & HONJO, T. 2002. The AID enzyme induces class switch recombination in fibroblasts. *Nature,* 416**,** 340-5.

ONCIU, M. 2009. Acute lymphoblastic leukemia. *Hematol Oncol Clin North Am,* 23**,** 655-74.

OSCIER, D. G., GARDINER, A. C., MOULD, S. J., GLIDE, S., DAVIS, Z. A., IBBOTSON, R. E., CORCORAN, M. M., CHAPMAN, R. M., THOMAS, P. W., COPPLESTONE, J. A., ORCHARD, J. A. & HAMBLIN, T. J. 2002. Multivariate analysis of prognostic factors in CLL: clinical stage, IGVH gene mutational status, and loss or mutation of the p53 gene are independent prognostic factors. *Blood,* 100**,** 1177-84.

PAIETTA, E. 2002. Assessing minimal residual disease (MRD) in leukemia: a changing definition and concept? *Bone Marrow Transplant,* 29**,** 459-65.

PAPE, K. A., TAYLOR, J. J., MAUL, R. W., GEARHART, P. J. & JENKINS, M. K. 2011. Different B cell populations mediate early and late memory during an endogenous immune response. *Science,* 331**,** 1203-7.

PEDERSEN, I. M., KITADA, S., LEONI, L. M., ZAPATA, J. M., KARRAS, J. G., TSUKADA, N., KIPPS, T. J., CHOI, Y. S., BENNETT, F. & REED, J. C. 2002. Protection of CLL B cells by a follicular dendritic cell line is dependent on induction of Mcl-1. *Blood,* 100**,** 1795-1801.

PELLETIER, N., MCHEYZER-WILLIAMS, L. J., WONG, K. A., URICH, E., FAZILLEAU, N. & MCHEYZER-WILLIAMS, M. G. 2010. Plasma cells negatively regulate the follicular helper T cell program. *Nat Immunol,* 11**,** 1110-8.

PERLMUTTER, R. M., KEARNEY, J. F., CHANG, S. P. & HOOD, L. E. 1985. Developmentally controlled expression of immunoglobulin VH genes. *Science,* 227**,** 1597-601.

PINCHUK, G. V., NOTTENBURG, C. & MILNER, E. C. 1995. Predominant V-region gene configurations in the human antibody response to Haemophilus influenzae capsule polysaccharide. *Scand J Immunol,* 41**,** 324-30.

PISCIOTTA, A. V. & HIRSCHBOECK, J. S. 1957. Therapeutic considerations in chronic lymphocytic leukemia; special reference to the natural course of the disease. *AMA Arch Intern Med,* 99**,** 334-5.

PLEVOVA, K., FRANCOVA, H. S., BURCKOVA, K., BRYCHTOVA, Y., DOUBEK, M., PAVLOVA, S., MALCIKOVA, J., MAYER, J., TICHY, B. & POSPISILOVA, S. 2014. Multiple productive immunoglobulin heavy chain gene rearrangements in chronic lymphocytic leukemia are mostly derived from independent clones. *Haematologica,* 99**,** 329-38.

POLLARD, A. J., PERRETT, K. P. & BEVERLEY, P. C. 2009. Maintaining protection against invasive bacteria with protein-polysaccharide conjugate vaccines. *Nat Rev Immunol,* 9**,** 213-20.

POPI, A. F., MOTTA, F. L., MORTARA, R. A., SCHENKMAN, S., LOPES, J. D. & MARIANO, M. 2009. Co-ordinated expression of lymphoid and myeloid specific transcription factors during B-1b cell differentiation into mononuclear phagocytes in vitro. *Immunology,* 126**,** 114-22.

POULSEN, T. R., JENSEN, A., HAURUM, J. S. & ANDERSEN, P. S. 2011. Limits for antibody affinity maturation and repertoire diversification in hypervaccinated humans. *J Immunol,* 187**,** 4229-35.

PUI, C. H., RELLING, M. V., LASCOMBES, F., HARRISON, P. L., STRUXIANO, A., MONDESIR, J. M., RIBEIRO, R. C., SANDLUND, J. T., RIVERA, G. K., EVANS, W. E. & MAHMOUD, H. H. 1997. Urate oxidase in prevention and treatment of hyperuricemia associated with lymphoid malignancies. *Leukemia,* 11**,** 1813-6.

PUI, C. H., ROBISON, L. L. & LOOK, A. T. 2008. Acute lymphoblastic leukaemia. *Lancet,* 371**,** 1030-43.

PYBUS, O. G., RAMBAUT, A., HOLMES, E. C. & HARVEY, P. H. 2002. New inferences from tree shape: numbers of missing taxa and population growth rates. *Syst Biol,* 51**,** 881-8.

QUAIL, M. A., SMITH, M., COUPLAND, P., OTTO, T. D., HARRIS, S. R., CONNOR, T. R., BERTONI, A., SWERDLOW, H. P. & GU, Y. 2012. A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. *BMC Genomics,* 13**,** 341.

QUINCE, C., LANZEN, A., CURTIS, T. P., DAVENPORT, R. J., HALL, N., HEAD, I. M., READ, L. F. & SLOAN, W. T. 2009. Accurate determination of microbial diversity from 454 pyrosequencing data. *Nat Methods,* 6**,** 639-41.

RACHEL, J. M., ZUCKER, M. L., FOX, C. M., PLAPP, F. V., MENITOVE, J. E., ABBASI, F. & MARTI, G. E. 2007. Monoclonal B-cell lymphocytosis in blood donors. *Br J Haematol,* 139**,** 832-6.

RAI, K. R., PETERSON, B. L., APPELBAUM, F. R., KOLITZ, J., ELIAS, L., SHEPHERD, L., HINES, J., THREATTE, G. A., LARSON, R. A., CHESON, B. D. & SCHIFFER, C. A. 2000. Fludarabine compared with chlorambucil as primary therapy for chronic lymphocytic leukemia. *N Engl J Med,* 343**,** 1750-7.

RAI, K. R., SAWITSKY, A., CRONKITE, E. P., CHANANA, A. D., LEVY, R. N. & PASTERNACK, B. S. 1975. Clinical staging of chronic lymphocytic leukemia. *Blood,* 46**,** 219-34.

RASSENTI, L. Z. & KIPPS, T. J. 1997. Lack of allelic exclusion in B cell chronic lymphocytic leukemia. *J Exp Med,* 185**,** 1435-45.

RATECH, H., SHEIBANI, K., NATHWANI, B. N. & RAPPAPORT, H. 1988. Immunoarchitecture of the "pseudofollicles" of well-differentiated (small) lymphocytic lymphoma: a comparison with true follicles. *Human pathology,* 19**,** 89-94.

RAWSTRON, A. C., BENNETT, F. L., O'CONNOR, S. J., KWOK, M., FENTON, J. A., PLUMMER, M., DE TUTE, R., OWEN, R. G., RICHARDS, S. J., JACK, A. S. & HILLMEN, P. 2008.

Monoclonal B-cell lymphocytosis and chronic lymphocytic leukemia. *N Engl J Med,* 359**,** 575-83.

RAWSTRON, A. C., YUILLE, M. R., FULLER, J., CULLEN, M., KENNEDY, B., RICHARDS, S. J., JACK, A. S., MATUTES, E., CATOVSKY, D., HILLMEN, P. & HOULSTON, R. S. 2002. Inherited predisposition to CLL is detectable as subclinical monoclonal B-lymphocyte expansion. *Blood,* 100**,** 2289-90.

REAMAN, G. H. 2002. Pediatric oncology: current views and outcomes. *Pediatr Clin North Am,* 49**,** 1305-18, vii.

REDAELLI, A., LASKIN, B. L., STEPHENS, J. M., BOTTEMAN, M. F. & PASHOS, C. L. 2004. The clinical and epidemiological burden of chronic lymphocytic leukaemia. *Eur J Cancer Care (Engl),* 13**,** 279-87.

RIESBECK, K. & NORDSTROM, T. 2006. Structure and immunological action of the human pathogen Moraxella catarrhalis IgD-binding protein. *Crit Rev Immunol,* 26**,** 353-76.

RINDSKOPF, D. 1997. Modern applied statistics with S-plus - Venables,WN, Ripley,BD. *Journal of Educational and Behavioral Statistics,* 22**,** 244-245.

ROBAK, T. 2001. Cladribine in the treatment of chronic lymphocytic leukemia. *Leuk Lymphoma,* 40**,** 551-64.

ROBERTS, K. G., MORIN, R. D., ZHANG, J., HIRST, M., ZHAO, Y., SU, X., CHEN, S. C., PAYNE-TURNER, D., CHURCHMAN, M. L., HARVEY, R. C., CHEN, X., KASAP, C., YAN, C., BECKSFORT, J., FINNEY, R. P., TEACHEY, D. T., MAUDE, S. L., TSE, K., MOORE, R., JONES, S., MUNGALL, K., BIROL, I., EDMONSON, M. N., HU, Y., BUETOW, K. E., CHEN, I. M., CARROLL, W. L., WEI, L., MA, J., KLEPPE, M., LEVINE, R. L., GARCIA-MANERO, G., LARSEN, E., SHAH, N. P., DEVIDAS, M., REAMAN, G., SMITH, M., PAUGH, S. W., EVANS, W. E., GRUPP, S. A., JEHA, S., PUI, C. H., GERHARD, D. S., DOWNING, J. R., WILLMAN, C. L., LOH, M., HUNGER, S. P., MARRA, M. A. & MULLIGHAN, C. G. 2012. Genetic alterations activating kinase and cytokine receptor signaling in high-risk acute lymphoblastic leukemia. *Cancer Cell,* 22**,** 153-66.

ROSENQUIST, R., MENESTRINA, F., LESTANI, M., KUPPERS, R., HANSMANN, M. L. & BRAUNINGER, A. 2004a. Indications for peripheral light-chain revision and somatic hypermutation without a functional B-cell receptor in precursors of a composite diffuse large B-cell and Hodgkin's lymphoma. *Lab Invest,* 84**,** 253-62.

ROSENQUIST, R., ROOS, G., ERLANSON, M., KUPPERS, R., BRAUNINGER, A. & HANSMANN, M. L. 2004b. Clonally related splenic marginal zone lymphoma and Hodgkin lymphoma with unmutated V gene rearrangements and a 15-yr time gap between diagnoses. *Eur J Haematol,* 73**,** 210-4.

ROSENWALD, A., ALIZADEH, A. A., WIDHOPF, G., SIMON, R., DAVIS, R. E., YU, X., YANG, L., PICKERAL, O. K., RASSENTI, L. Z., POWELL, J., BOTSTEIN, D., BYRD, J. C., GREVER, M. R., CHESON, B. D., CHIORAZZI, N., WILSON, W. H., KIPPS, T. J., BROWN, P. O. & STAUDT, L. M. 2001. Relation of gene expression phenotype to immunoglobulin mutation genotype in B cell chronic lymphocytic leukemia. *J Exp Med,* 194**,** 1639-47.

ROSSI, D. & GAIDANO, G. 2010. Biological and clinical significance of stereotyped B-cell receptors in chronic lymphocytic leukemia. *Haematologica-the Hematology Journal,* 95**,** 1992-1995.

ROSSI, D., SPINA, V., FORCONI, F., CAPELLO, D., FANGAZIO, M., RASI, S., MARTINI, M., GATTEI, V., RAMPONI, A., LAROCCA, L. M., BERTONI, F. & GAIDANO, G. 2012. Molecular history of Richter syndrome: origin from a cell already present at the time of chronic lymphocytic leukemia diagnosis. *Int J Cancer,* 130**,** 3006-10.

ROWE, J. M., BUCK, G., BURNETT, A. K., CHOPRA, R., WIERNIK, P. H., RICHARDS, S. M., LAZARUS, H. M., FRANKLIN, I. M., LITZOW, M. R., CIOBANU, N., PRENTICE, H. G., DURRANT, J., TALLMAN, M. S., GOLDSTONE, A. H., ECOG & PARTY, M. N. A. L. W. 2005. Induction therapy for adults with acute lymphoblastic leukemia: results of

more than 1500 patients from the international ALL trial: MRC UKALL XII/ECOG E2993. *Blood,* 106**,** 3760-7.

ROWLEY, B., TANG, L., SHINTON, S., HAYAKAWA, K. & HARDY, R. R. 2007. Autoreactive B-1 B cells: constraints on natural autoantibody B cell antigen receptors. *Journal of autoimmunity,* 29**,** 236-45.

ROZMAN, C. & MONTSERRAT, E. 1995. Chronic lymphocytic leukemia. *The New England journal of medicine,* 333**,** 1052-7.

RUBELT, F., SIEVERT, V., KNAUST, F. & DIENER, C. E. A. 2012. Onset of Immune Senescence Defined by Unbiased Pyrosequencing of Human Immunoglobulin mRNA Repertoires. *PLOS One,* 7.

SABATTINI, E., BACCI, F., SAGRAMOSO, C. & PILERI, S. A. 2010. WHO classification of tumours of haematopoietic and lymphoid tissues in 2008: an overview. *Pathologica,* 102**,** 83-7.

SAKUISHI, K., APETOH, L., SULLIVAN, J. M., BLAZAR, B. R., KUCHROO, V. K. & ANDERSON, A. C. 2010. Targeting Tim-3 and PD-1 pathways to reverse T cell exhaustion and restore anti-tumor immunity. *J Exp Med,* 207**,** 2187-94.

SANCHEZ, M. L., ALMEIDA, J., GONZALEZ, D., GONZALEZ, M., GARCIA-MARCOS, M. A., BALANZATEGUI, A., LOPEZ-BERGES, M. C., NOMDEDEU, J., VALLESPI, T., BARBON, M., MARTIN, A., DE LA FUENTE, P., MARTIN-NUNEZ, G., FERNANDEZ-CALVO, J., HERNANDEZ, J. M., SAN MIGUEL, J. F. & ORFAO, A. 2003. Incidence and clinicobiologic characteristics of leukemic B-cell chronic lymphoproliferative disorders with more than one B-cell clone. *Blood,* 102**,** 2994-3002.

SANCHEZ, M. L., ALMEIDA, J., LOPEZ, A., SAYAGUES, J. M., RASILLO, A., SARASQUETE, E. A., BALANZATEGUI, A., TABERNERO, M. D., DIAZ-MEDIAVILLA, J., BARRACHINA, C., PAIVA, A., GONZALEZ, M., SAN MIGUEL, J. F. & ORFAO, A. 2006. Heterogeneity of neoplastic cells in B-cell chronic lymphoproliferative disorders: biclonality versus intraclonal evolution of a single tumor cell clone. *Haematologica,* 91**,** 331-9.

SANT, M., ALLEMANI, C., TEREANU, C., DE ANGELIS, R., CAPOCACCIA, R., VISSER, O., MARCOS-GRAGERA, R., MAYNADIE, M., SIMONETTI, A., LUTZ, J. M., BERRINO, F. & GROUP, H. W. 2010. Incidence of hematologic malignancies in Europe by morphologic subtype: results of the HAEMACARE project. *Blood,* 116**,** 3724-34.

SARSOTTI, E., MARUGAN, I., BENET, I., TEROL, M. J., SANCHEZ-IZQUIERDO, D., TORMO, M., RUBIO-MOSCARDO, F., MARTINEZ-CLIMENT, J. A. & GARCIA-CONDE, J. 2004. Bcl-6 mutation status provides clinically valuable information in early-stage B-cell chronic lymphocytic leukemia. *Leukemia,* 18**,** 743-6.

SAUTER, C., LAMANNA, N. & WEISS, M. A. 2008. Pentostatin in chronic lymphocytic leukemia. *Expert Opin Drug Metab Toxicol,* 4**,** 1217-22.

SAVEN, A., BURIAN, C., ADUSUMALLI, J. & KOZIOL, J. A. 1999. Filgrastim for cladribine-induced neutropenic fever in patients with hairy cell leukemia. *Blood,* 93**,** 2471-7.

SAYALA, H. A., RAWSTRON, A. C. & HILLMEN, P. 2007. Minimal residual disease assessment in chronic lymphocytic leukaemia. *Best practice & research. Clinical haematology,* 20**,** 499-512.

SCHARF, O., GOLDING, H., KING, L. R., ELLER, N., FRAZIER, D., GOLDING, B. & SCOTT, D. E. 2001. Immunoglobulin G3 from polyclonal human immunodeficiency virus (HIV) immune globulin is more potent than other subclasses in neutralizing HIV type 1. *J Virol,* 75**,** 6558-65.

SCHATZ, D. G. & SWANSON, P. C. 2010. V(D)J Recombination: Mechanisms of Initiation. *Annual review of genetics*.

SCHLIEP, K. P. 2011. phangorn: phylogenetic analysis in R. *Bioinformatics,* 27**,** 592-3.

SCHMID, C. & ISAACSON, P. G. 1994. Proliferation centres in B-cell malignant lymphoma, lymphocytic (B-CLL): an immunophenotypic study. *Histopathology,* 24**,** 445-51.

SCHMITZ, R., RENNE, C., ROSENQUIST, R., TINGUELY, M., DISTLER, V., MENESTRINA, F., LESTANI, M., STANKOVIC, T., AUSTEN, B., BRAUNINGER, A., HANSMANN, M. L. & KUPPERS, R. 2005. Insights into the multistep transformation process of lymphomas: IgH-associated translocations and tumor suppressor gene mutations in clonally related composite Hodgkin's and non-Hodgkin's lymphomas. *Leukemia,* 19**,** 1452-8.

SCHROEDER, H. W., JR. & CAVACINI, L. 2010. Structure and function of immunoglobulins. *J Allergy Clin Immunol,* 125**,** S41-52.

SCHROEDER, H. W., JR. & DIGHIERO, G. 1994. The pathogenesis of chronic lymphocytic leukemia: analysis of the antibody repertoire. *Immunol Today,* 15**,** 288-94.

SCHROEDER, H. W., JR., ZHANG, L. & PHILIPS, J. B., 3RD 2001. Slow, programmed maturation of the immunoglobulin HCDR3 repertoire during the third trimester of fetal life. *Blood,* 98**,** 2745-51.

SCHROERS, R., GRIESINGER, F., TRUMPER, L., HAASE, D., KULLE, B., KLEIN-HITPASS, L., SELLMANN, L., DUHRSEN, U. & DURIG, J. 2005. Combined analysis of ZAP-70 and CD38 expression as a predictor of disease progression in B-cell chronic lymphocytic leukemia. *Leukemia,* 19**,** 750-8.

SCHUH, A., BECQ, J., HUMPHRAY, S., ALEXA, A., BURNS, A., CLIFFORD, R., FELLER, S. M., GROCOCK, R., HENDERSON, S., KHREBTUKOVA, I., KINGSBURY, Z., LUO, S., MCBRIDE, D., MURRAY, L., MENJU, T., TIMBS, A., ROSS, M., TAYLOR, J. & BENTLEY, D. 2012. Monitoring chronic lymphocytic leukemia progression by whole genome sequencing reveals heterogeneous clonal evolution patterns. *Blood,* 120**,** 4191-6.

SECKER-WALKER, L. M., CRAIG, J. M., HAWKINS, J. M. & HOFFBRAND, A. V. 1991. Philadelphia positive acute lymphoblastic leukemia in adults: age distribution, BCR breakpoint and prognostic significance. *Leukemia,* 5**,** 196-9.

SEIFERT, M. & KUPPERS, R. 2009. Molecular footprints of a germinal center derivation of human IgM+(IgD+)CD27+ B cells and the dynamics of memory B cell generation. *J Exp Med,* 206**,** 2659-69.

SHANAFELT, T. D., KAY, N. E., RABE, K. G., CALL, T. G., ZENT, C. S., MADDOCKS, K., JENKINS, G., JELINEK, D. F., MORICE, W. G., BOYSEN, J., SCHWAGER, S., BOWEN, D., SLAGER, S. L. & HANSON, C. A. 2009. Brief report: natural history of individuals with clinically recognized monoclonal B-cell lymphocytosis compared with patients with Rai 0 chronic lymphocytic leukemia. *J Clin Oncol,* 27**,** 3959-63.

SHAW, R. K., SZWED, C., BOGGS, D. R., FAHEY, J. L., FREI III, E., MORRISON, E. & UTZ, J. P. 1960. Infection and immunity in chronic lymphocytic leukemia. *Arch Intern Med,* 106**,** 467-478.

SHIM, Y. K., MIDDLETON, D. C., CAPORASO, N. E., RACHEL, J. M., LANDGREN, O., ABBASI, F., RAVECHE, E. S., RAWSTRON, A. C., ORFAO, A., MARTI, G. E. & VOGT, R. F. 2010. Prevalence of monoclonal B-cell lymphocytosis: a systematic review. *Cytometry B Clin Cytom,* 78 Suppl 1**,** S10-8.

SHIM, Y. K., RACHEL, J. M., GHIA, P., BOREN, J., ABBASI, F., DAGKLIS, A., VENABLE, G., KANG, J., DEGHEIDY, H., PLAPP, F. V., VOGT, R. F., MENITOVE, J. E. & MARTI, G. E. 2014. Monoclonal B-cell lymphocytosis in healthy blood donors: an unexpectedly common finding. *Blood,* 123**,** 1319-26.

SHIM, Y. K., VOGT, R. F., MIDDLETON, D., ABBASI, F., SLADE, B., LEE, K. Y. & MARTI, G. E. 2007. Prevalence and natural history of monoclonal and polyclonal B-cell lymphocytosis in a residential adult population. *Cytometry B Clin Cytom,* 72**,** 344-53.

SILVERMAN, L. B. & SALLAN, S. E. 2003. Newly diagnosed childhood acute lymphoblastic leukemia: update on prognostic factors and treatment. *Curr Opin Hematol,* 10**,** 290-6.

SKLAR, J., CLEARY, M. L., THIELEMANS, K., GRALOW, J., WARNKE, R. & LEVY, R. 1984. Biclonal B-cell lymphoma. *N Engl J Med,* 311**,** 20-7.

SMITH, K., MUTHER, J. J., DUKE, A. L., MCKEE, E., ZHENG, N. Y., WILSON, P. C. & JAMES, J. A. 2013. Fully human monoclonal antibodies from antibody secreting cells after vaccination with Pneumovax(R)23 are serotype specific and facilitate opsonophagocytosis. *Immunobiology,* 218**,** 745-54.

SOK, D., LASERSON, U., LASERSON, J., LIU, Y., VIGNEAULT, F., JULIEN, J. P., BRINEY, B., RAMOS, A., SAYE, K. F., LE, K., MAHAN, A., WANG, S., KARDAR, M., YAARI, G., WALKER, L. M., SIMEN, B. B., ST JOHN, E. P., CHAN-HUI, P. Y., SWIDEREK, K., KLEINSTEIN, S. H., ALTER, G., SEAMAN, M. S., CHAKRABORTY, A. K., KOLLER, D., WILSON, I. A., CHURCH, G. M., BURTON, D. R. & POIGNARD, P. 2013. The effects of somatic hypermutation on neutralization and binding in the PGT121 family of broadly neutralizing HIV antibodies. *PLoS Pathog,* 9**,** e1003754.

SOMA, L. A., CRAIG, F. E. & SWERDLOW, S. H. 2006. The proliferation center microenvironment and prognostic markers in chronic lymphocytic leukemia/small lymphocytic lymphoma. *Human pathology,* 37**,** 152-9.

SPENCER, J., BARONE, F. & DUNN-WALTERS, D. 2009. Generation of Immunoglobulin diversity in human gut-associated lymphoid tissue. *Semin Immunol,* 21**,** 139-46.

STEENBERGEN, E. J., VERHAGEN, O. J., VAN LEEUWEN, E. F., VON DEM BORNE, A. E. & VAN DER SCHOOT, C. E. 1993. Distinct ongoing Ig heavy chain rearrangement processes in childhood B-precursor acute lymphoblastic leukemia. *Blood,* 82**,** 581-9.

STEIMAN-SHIMONY, A., EDELMAN, H., BARAK, M., SHAHAF, G., DUNN-WALTERS, D., STOTT, D. I., ABRAHAM, R. S. & MEHR, R. 2006a. Immunoglobulin variable-region gene mutational lineage tree analysis: application to autoimmune diseases. *Autoimmun Rev,* 5**,** 242-51.

STEIMAN-SHIMONY, A., EDELMAN, H., HUTZLER, A., BARAK, M., ZUCKERMAN, N. S., SHAHAF, G., DUNN-WALTERS, D., STOTT, D. I., ABRAHAM, R. S. & MEHR, R. 2006b. Lineage tree analysis of immunoglobulin variable-region gene mutations in autoimmune diseases: chronic activation, normal selection. *Cell Immunol,* 244**,** 130-6.

STEINBRUCK, L. & MCHARDY, A. C. 2012. Inference of genotype-phenotype relationships in the antigenic evolution of human influenza A (H3N2) viruses. *PLoS Comput Biol,* 8**,** e1002492.

STEVENSON, F. K. & CALIGARIS-CAPPIO, F. 2004. Chronic lymphocytic leukemia: revelations from the B-cell receptor. *Blood,* 103**,** 4389-95.

STOLZ, C. & SCHULER, M. 2009. Molecular mechanisms of resistance to Rituximab and pharmacologic strategies for its circumvention. *Leuk Lymphoma,* 50**,** 873-85.

SUTTON, L. A., KOSTARELI, E., HADZIDIMITRIOU, A., DARZENTAS, N., TSAFTARIS, A., ANAGNOSTOPOULOS, A., ROSENQUIST, R. & STAMATOPOULOS, K. 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**,** 4460-8.

TAILLARDET, M., HAFFAR, G., MONDIERE, P., ASENSIO, M. J., GHEIT, H., BURDIN, N., DEFRANCE, T. & GENESTIER, L. 2009. The thymus-independent immunity conferred by a pneumococcal polysaccharide is mediated by long-lived plasma cells. *Blood,* 114**,** 4432-40.

TAKEMURA, S., BRAUN, A., CROWSON, C., KURTIN, P. J., COFIELD, R. H., O'FALLON, W. M., GORONZY, J. J. & WEYAND, C. M. 2001. Lymphoid neogenesis in rheumatoid synovitis. *Journal of immunology,* 167**,** 1072-80.

TAN, Y. C., BLUM, L. K., KONGPACHITH, S., JU, C. H., CAI, X., LINDSTROM, T. M., SOKOLOVE, J. & ROBINSON, W. H. 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.

TANGYE, S. G. & GOOD, K. L. 2007. Human IgM+CD27+ B cells: memory B cells or "memory" B cells? *J Immunol,* 179**,** 13-9.

TANGYE, S. G., LIU, Y. J., AVERSA, G., PHILLIPS, J. H. & DE VRIES, J. E. 1998. Identification of functional human splenic memory B cells by expression of CD148 and CD27. *J Exp Med,* 188**,** 1691-703.

TEN BOEKEL, E., MELCHERS, F. & ROLINK, A. G. 1998. Precursor B cells showing H chain allelic inclusion display allelic exclusion at the level of pre-B cell receptor surface expression. *Immunity,* 8**,** 199-207.

THOMAS, X., BOIRON, J. M., HUGUET, F., DOMBRET, H., BRADSTOCK, K., VEY, N., KOVACSOVICS, T., DELANNOY, A., FEGUEUX, N., FENAUX, P., STAMATOULLAS, A., VERNANT, J. P., TOURNILHAC, O., BUZYN, A., REMAN, O., CHARRIN, C., BOUCHEIX, C., GABERT, J., LHERITIER, V. & FIERE, D. 2004. Outcome of treatment in adults with acute lymphoblastic leukemia: analysis of the LALA-94 trial. *J Clin Oncol,* 22**,** 4075-86.

TILLER, T., TSUIJI, M., YURASOV, S., VELINZON, K., NUSSENZWEIG, M. C. & WARDEMANN, H. 2007. Autoreactivity in human IgG+ memory B cells. *Immunity,* 26**,** 205-13.

TINGUELY, M., ROSENQUIST, R., SUNDSTROM, C., AMINI, R. M., KUPPERS, R., HANSMANN, M. L. & BRAUNINGER, A. 2003. Analysis of a clonally related mantle cell and Hodgkin lymphoma indicates Epstein-Barr virus infection of a Hodgkin/Reed-Sternberg cell precursor in a germinal center. *Am J Surg Pathol,* 27**,** 1483-8.

TOBIN, G., SODERBERG, O., THUNBERG, U. & ROSENQUIST, R. 2004a. 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**,** 221-8.

TOBIN, G., THUNBERG, U., KARLSSON, K., MURRAY, F., LAURELL, A., WILLANDER, K., ENBLAD, G., MERUP, M., VILPO, J., JULIUSSON, G., SUNDSTROM, C., SODERBERG, O., ROOS, G. & ROSENQUIST, R. 2004b. Subsets with restricted immunoglobulin gene rearrangement features indicate a role for antigen selection in the development of chronic lymphocytic leukemia. *Blood,* 104**,** 2879-85.

TONEGAWA, S. 1983. Somatic generation of antibody diversity. *Nature,* 302**,** 575-81.

UDUMAN, M., SHLOMCHIK, M. J., VIGNEAULT, F., CHURCH, G. M. & KLEINSTEIN, S. H. 2014. Integrating B cell lineage information into statistical tests for detecting selection in Ig sequences. *J Immunol,* 192**,** 867-74.

VAN DEN BERG, A., MAGGIO, E., RUST, R., KOOISTRA, K., DIEPSTRA, A. & POPPEMA, S. 2002. Clonal relation in a case of CLL, ALCL, and Hodgkin composite lymphoma. *Blood,* 100**,** 1425-9.

VAN DEN NESTE, E., DELANNOY, A., VANDERCAM, B., BOSLY, A., FERRANT, A., MINEUR, P., MONTFORT, L., MARTIAT, P., STRAETMANS, N., FILLEUL, B. & MICHAUX, J. L. 1996. Infectious complications after 2-chlorodeoxyadenosine therapy. *Eur J Haematol,* 56**,** 235-40.

VAN DER VELDEN, V. H., CAZZANIGA, G., SCHRAUDER, A., HANCOCK, J., BADER, P., PANZER-GRUMAYER, E. R., FLOHR, T., SUTTON, R., CAVE, H., MADSEN, H. O., CAYUELA, J. M., TRKA, J., ECKERT, C., FORONI, L., ZUR STADT, U., BELDJORD, K., RAFF, T., VAN DER SCHOOT, C. E., VAN DONGEN, J. J. & EUROPEAN STUDY GROUP ON, M. R. D. D. I. A. L. L. 2007. Analysis of minimal residual disease by Ig/TCR gene rearrangements: guidelines for interpretation of real-time quantitative PCR data. *Leukemia,* 21**,** 604-11.

VAN DONGEN, J. J., LANGERAK, A. W., BRUGGEMANN, M., EVANS, P. A., HUMMEL, M., LAVENDER, F. L., DELABESSE, E., DAVI, F., SCHUURING, E., GARCIA-SANZ, R., VAN KRIEKEN, J. H., DROESE, J., GONZALEZ, D., BASTARD, C., WHITE, H. E., SPAARGAREN, M., GONZALEZ, M., PARREIRA, A., SMITH, J. L., MORGAN, G. J., KNEBA, M. & MACINTYRE, E. A. 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 : official journal of the Leukemia Society of America, Leukemia Research Fund, U.K,* 17**,** 2257-317.

VAN KRIEKEN, J. H., LANGERAK, A. W., MACINTYRE, E. A., KNEBA, M., HODGES, E., SANZ, R. G., MORGAN, G. J., PARREIRA, A., MOLINA, T. J., CABECADAS, J., GAULARD, P., JASANI, B., GARCIA, J. F., OTT, M., HANNSMANN, M. L., BERGER, F., HUMMEL, M., DAVI, F., BRUGGEMANN, M., LAVENDER, F. L., SCHUURING, E., EVANS, P. A., WHITE, H., SALLES, G., GROENEN, P. J., GAMEIRO, P., POTT, C. & DONGEN, J. J. 2007. Improved reliability of lymphoma diagnostics via PCR-based clonality testing: report of the BIOMED-2 Concerted Action BHM4-CT98-3936. *Leukemia : official journal of the Leukemia Society of America, Leukemia Research Fund, U.K,* 21**,** 201-6.

VAN STAA, T. P., LEUFKENS, H. G., ABENHAIM, L., ZHANG, B. & COOPER, C. 2000. Use of oral corticosteroids and risk of fractures. *J Bone Miner Res,* 15**,** 993-1000.

VARADARAJAN, N., JULG, B., YAMANAKA, Y. J., CHEN, H., OGUNNIYI, A. O., MCANDREW, E., PORTER, L. C., PIECHOCKA-TROCHA, A., HILL, B. J., DOUEK, D. C., PEREYRA, F., WALKER, B. D. & LOVE, J. C. 2011. A high-throughput single-cell analysis of human CD8(+) T cell functions reveals discordance for cytokine secretion and cytolysis. *The Journal of clinical investigation,* 121**,** 4322-31.

VARGAS, R. L., FELGAR, R. E. & ROTHBERG, P. G. 2008. Detection of clonality in lymphoproliferations using PCR of the antigen receptor genes: Does size matter? *Leukemia Research,* 32**,** 335-338.

VENERI, D., ORTOLANI, R., FRANCHINI, M., TRIDENTE, G., PIZZOLO, G. & VELLA, A. 2009. Expression of CD27 and CD23 on peripheral blood B lymphocytes in humans of different ages. *Blood Transfus,* 7**,** 29-34.

VERKOCZY, L., DUONG, B., SKOG, P., AIT-AZZOUZENE, D., PURI, K., VELA, J. L. & NEMAZEE, D. 2007. Basal B cell receptor-directed phosphatidylinositol 3-kinase signaling turns off RAGs and promotes B cell-positive selection. *J Immunol,* 178**,** 6332-41.

VETTERMANN, C. & SCHLISSEL, M. S. 2010. Allelic exclusion of immunoglobulin genes: models and mechanisms. *Immunol Rev,* 237**,** 22-42.

VISCO, C., MORETTA, F., FALISI, E., FACCO, M., MAURA, F., NOVELLA, E., NICHELE, I., FINOTTO, S., GIARETTA, I., AVE, E., PERBELLINI, O., GUERCINI, N., SCUPOLI, M. T., TRENTIN, L., TRIMARCO, V., NERI, A., SEMENZATO, G., RODEGHIERO, F., PIZZOLO, G. & AMBROSETTI, A. 2013. Double productive immunoglobulin sequence rearrangements in patients with chronic lymphocytic leukemia. *Am J Hematol,* 88**,** 277-82.

VOLLMERS, C., SIT, R. V., WEINSTEIN, J. A., DEKKER, C. L. & QUAKE, S. R. 2013. Genetic measurement of memory B-cell recall using antibody repertoire sequencing. *Proceedings of the National Academy of Sciences of the United States of America,* 110**,** 13463-13468.

VON BUDINGEN, H. C., KUO, T. C., SIROTA, M., VAN BELLE, C. J., APELTSIN, L., GLANVILLE, J., CREE, B. A., GOURRAUD, P. A., SCHWARTZBURG, A., HUERTA, G., TELMAN, D., SUNDAR, P. D., CASEY, T., COX, D. R. & HAUSER, S. L. 2012. B cell exchange across the blood-brain barrier in multiple sclerosis. *J Clin Invest,* 122**,** 4533-43.

WANG, C., MITSUYA, Y., GHARIZADEH, B., RONAGHI, M. & SHAFER, R. W. 2007. Characterization of mutation spectra with ultra-deep pyrosequencing: application to HIV-1 drug resistance. *Genome research,* 17**,** 1195-201.

WANG, Y. H., STEPHAN, R. P., SCHEFFOLD, A., KUNKEL, D., KARASUYAMA, H., RADBRUCH, A. & COOPER, M. D. 2002. Differential surrogate light chain expression governs B-cell differentiation. *Blood,* 99**,** 2459-67.

WARREN, E. H., MATSEN, F. A. T. & CHOU, J. 2013. High-throughput sequencing of B- and T-lymphocyte antigen receptors in hematology. *Blood,* 122**,** 19-22.

WARREN, R. L., FREEMAN, J. D., ZENG, T., CHOE, G., MUNRO, S., MOORE, R., WEBB, J. R. & HOLT, R. A. 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 research,* 21**,** 790-7.

WATSON, S. J., WELKERS, M. R. A., DEPLEDGE, D. P., COULTER, E., BREUER, J. M., DE JONG, M. D. & KELLAM, P. 2013. Viral population analysis and minority-variant detection using short read next-generation sequencing. *Philosophical Transactions of the Royal Society B-Biological Sciences,* 368.

WEINSTEIN, J. A., JIANG, N., WHITE, R. A., 3RD, FISHER, D. S. & QUAKE, S. R. 2009. High-throughput sequencing of the zebrafish antibody repertoire. *Science,* 324**,** 807-10.

WEINSTEIN, R. S., JILKA, R. L., PARFITT, A. M. & MANOLAGAS, S. C. 1998. Inhibition of osteoblastogenesis and promotion of apoptosis of osteoblasts and osteocytes by glucocorticoids. Potential mechanisms of their deleterious effects on bone. *J Clin Invest,* 102**,** 274-82.

WELLER, S., BRAUN, M. C., TAN, B. K., ROSENWALD, A., CORDIER, C., CONLEY, M. E., PLEBANI, A., KUMARARATNE, D. S., BONNET, D., TOURNILHAC, O., TCHERNIA, G., STEINIGER, B., STAUDT, L. M., CASANOVA, J. L., REYNAUD, C. A. & WEILL, J. C. 2004. Human blood IgM "memory" B cells are circulating splenic marginal zone B cells harboring a prediversified immunoglobulin repertoire. *Blood,* 104**,** 3647-54.

WIDHOPF, G. F., 2ND, RASSENTI, L. Z., TOY, T. L., GRIBBEN, J. G., WIERDA, W. G. & KIPPS, T. J. 2004. Chronic lymphocytic leukemia B cells of more than 1% of patients express virtually identical immunoglobulins. *Blood,* 104**,** 2499-504.

WIEMELS, J. L., CAZZANIGA, G., DANIOTTI, M., EDEN, O. B., ADDISON, G. M., MASERA, G., SAHA, V., BIONDI, A. & GREAVES, M. F. 1999. Prenatal origin of acute lymphoblastic leukaemia in children. *Lancet,* 354**,** 1499-1503.

WIESTNER, A., ROSENWALD, A., BARRY, T. S., WRIGHT, G., DAVIS, R. E., HENRICKSON, S. E., ZHAO, H., IBBOTSON, R. E., ORCHARD, J. A., DAVIS, Z., STETLER-STEVENSON, M., RAFFELD, M., ARTHUR, D. C., MARTI, G. E., WILSON, W. H., HAMBLIN, T. J., OSCIER, D. G. & STAUDT, L. M. 2003. ZAP-70 expression identifies a chronic lymphocytic leukemia subtype with unmutated immunoglobulin genes, inferior clinical outcome, and distinct gene expression profile. *Blood,* 101**,** 4944-51.

WILGENBUSCH, J. C. & SWOFFORD, D. 2003. Inferring evolutionary trees with PAUP*. *Curr Protoc Bioinformatics,* Chapter 6**,** Unit 6 4.

WILLIAMS, J. V., WEITKAMP, J. H., BLUM, D. L., LAFLEUR, B. J. & CROWE, J. E., JR. 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**,** 407-14.

WOOF, J. M. & BURTON, D. R. 2004a. Human antibody-Fc receptor interactions illuminated by crystal structures. *Nat Rev Immunol,* 4**,** 89-99.

WOOF, J. M. & BURTON, D. R. 2004b. Human antibody-Fc receptor interactions illuminated by crystal structures. *Nature reviews. Immunology,* 4**,** 89-99.

WOOF, J. M. & MESTECKY, J. 2005. Mucosal immunoglobulins. *Immunol Rev,* 206**,** 64-82.

WRAMMERT, J., SMITH, K., MILLER, J., LANGLEY, W. A., KOKKO, K., LARSEN, C., ZHENG, N. Y., MAYS, I., GARMAN, L., HELMS, C., JAMES, J., AIR, G. M., CAPRA, J. D., AHMED, R. & WILSON, P. C. 2008. Rapid cloning of high-affinity human monoclonal antibodies against influenza virus. *Nature,* 453**,** 667-71.

WU, X., ZHOU, T., ZHU, J., ZHANG, B., GEORGIEV, I., WANG, C., CHEN, X., LONGO, N. S., LOUDER, M., MCKEE, K., O'DELL, S., PERFETTO, S., SCHMIDT, S. D., SHI, W., WU, L., YANG, Y., YANG, Z. Y., YANG, Z., ZHANG, Z., BONSIGNORI, M., CRUMP, J. A., KAPIGA, S. H., SAM, N. E., HAYNES, B. F., SIMEK, M., BURTON, D. R., KOFF, W. C., DORIA-ROSE, N. A., CONNORS, M., PROGRAM, N. C. S., MULLIKIN, J. C., NABEL, G. J., ROEDERER, M., SHAPIRO, L., KWONG, P. D. & MASCOLA, J. R. 2011. Focused

evolution of HIV-1 neutralizing antibodies revealed by structures and deep sequencing. *Science,* 333**,** 1593-602.

WU, Y. C., KIPLING, D. & DUNN-WALTERS, D. K. 2012. Age-Related Changes in Human Peripheral Blood IGH Repertoire Following Vaccination. *Front Immunol,* 3**,** 193.

WU, Y. C., KIPLING, D., LEONG, H. S., MARTIN, V., ADEMOKUN, A. A. & DUNN-WALTERS, D. K. 2010. High-throughput immunoglobulin repertoire analysis distinguishes between human IgM memory and switched memory B-cell populations. *Blood,* 116**,** 1070-8.

YE, J., MA, N., MADDEN, T. L. & OSTELL, J. M. 2013. IgBLAST: an immunoglobulin variable domain sequence analysis tool. *Nucleic Acids Res,* 41**,** W34-40.

YIN, L., HOU, W., LIU, L., CAI, Y., WALLET, M. A., GARDNER, B. P., CHANG, K., LOWE, A. C., RODRIGUEZ, C. A., SRIAROON, P., FARMERIE, W. G., SLEASMAN, J. W. & GOODENOW, M. M. 2013. IgM Repertoire Biodiversity is Reduced in HIV-1 Infection and Systemic Lupus Erythematosus. *Front Immunol,* 4**,** 373.

YOSHIKAWA, S., KAWANO, Y., MINEGISHI, Y. & KARASUYAMA, H. 2009. The skewed heavy-chain repertoire in peritoneal B-1 cells is predetermined by the selection via pre-B cell receptor during B cell ontogeny in the fetal liver. *International immunology,* 21**,** 43-52.

ZENZ, T., DOHNER, H. & STILGENBAUER, S. 2007. Genetics and risk-stratified approach to therapy in chronic lymphocytic leukemia. *Best practice & research. Clinical haematology,* 20**,** 439-53.

ZENZ, T., EICHHORST, B., BUSCH, R., DENZEL, T., HABE, S., WINKLER, D., BUHLER, A., EDELMANN, J., BERGMANN, M., HOPFINGER, G., HENSEL, M., HALLEK, M., DOHNER, H. & STILGENBAUER, S. 2010. TP53 mutation and survival in chronic lymphocytic leukemia. *J Clin Oncol,* 28**,** 4473-9.

ZHU, J., OFEK, G., YANG, Y., ZHANG, B., LOUDER, M. K., LU, G., MCKEE, K., PANCERA, M., SKINNER, J., ZHANG, Z., PARKS, R., EUDAILEY, J., LLOYD, K. E., BLINN, J., ALAM, S. M., HAYNES, B. F., SIMEK, M., BURTON, D. R., KOFF, W. C., PROGRAM, N. C. S., MULLIKIN, J. C., MASCOLA, J. R., SHAPIRO, L. & KWONG, P. D. 2013a. 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.

ZHU, J., WU, X., ZHANG, B., MCKEE, K., O'DELL, S., SOTO, C., ZHOU, T., CASAZZA, J. P., PROGRAM, N. C. S., MULLIKIN, J. C., KWONG, P. D., MASCOLA, J. R. & SHAPIRO, L. 2013b. De novo identification of VRC01 class HIV-1-neutralizing antibodies by next-generation sequencing of B-cell transcripts. *Proc Natl Acad Sci U S A,* 110**,** E4088-97.

Appendix A

Published works

Molecular Evolution of Broadly Neutralizing Llama Antibodies to the CD4-Binding Site of HIV-1.

PLoS Pathogens 2014**.**

McCoy LE, Rutten L, Frampton D, Anderson I, Granger L, **Bashford-Rogers R**, Dekkers G, Strokappe NM, Seaman MS, Koh W, Grippo, V., Kliche, A., Verrips, T., Kellam, P., Fassati, A., Weiss, R. A.

Capturing needles in haystacks: comparison of B-cell receptor sequencing methods.

BMC Immunology 2014**.**

Bashford-Rogers, R., Palser, A, Idris, S, Carter, L, Epstein, M, Callard, R, Douek, D., Vassiliou, G., Follows, G., Hubank, M. and Kellam, P.

Network properties derived from deep sequencing of the human B-cell receptor repertoires delineates B-cell populations.

Genome Research 2013.

Bashford-Rogers, R. Palser AL, Huntly BJ, Rance R, Vassiliou GS, Follows GA, Kellam P.

Transmission and evolution of the Middle East respiratory syndrome coronavirus in Saudi Arabia: a descriptive genomic study.

Lancet 2013.

Cotten M, Watson SJ, Kellam P, Al-Rabeeah AA, Makhdoom HQ, Assiri A, Al-Tawfiq JA, Alhakeem RF, Madani H, Alrabiah FA, Hajjar, S. A., Al-Nassir, W. N., Albarrak, A., Flemban, H., Balkhy, H. H., Alsubaie, S., Palser, A. L., Gall, A., **Bashford-Rogers, R**., Rambaut, A., Zumla, A. I. & Memish, Z. (2013).