Using Antibody Next Generation Sequencing Data to Aid Antibody Engineering

Aleksandr Kovaltsuk
St Hugh's College
University of Oxford

A thesis submitted for the degree of
Doctor of Philosophy

Michaelmas 2020

Acknowledgements

This has been a great and pleasant journey that taught me many valuable lessons and skills, and connected me with some of the smartest and kindest people of this generation.

The completion of this DPhil would be impossible without excellent support, advice and patience from Prof. Charlotte Deane. Her depth of scientific knowledge and curiosity motivated me to be a better researcher from day one. I would also like to thank Prof. Deane for assembling the group of the nicest fellow DPhil researchers.

I wish to thank all OPIG members, especially those who shared room 2.17 with me in the last 4 years. Special shout-outs go Konrad, Claire, Jin, Susan, Catherine, honourable dames (Eve and Sarah) and Fergi. It has been a pleasure to work with you all as well as to share many good laughs. I would like to express my sincere gratitude to Matt for being a good friend and for his unwavering support throughout this DPhil.

I had a great pleasure working with my industrial supervisors from UCB, Pharma (James and Seb). I am very grateful to Dr. Johannes Trück and to his group members (Valentin and Marie). They made my visits to their lab very welcoming. I must also thank BBSRC, Royal Commission for the exhibition of 1851 and EMBO for making this DPhil a truly exceptional and unique opportunity.

I am deeply indebted to my family for their continuous support of my journey. Especially, I would like to thank my grandad who always believed in me and wanted me to excel academically.

I cannot begin to express my thanks to Ruth, who was by side throughout this challenging time period. Her care, kindness and love makes me forget about any worries I have.

List of Publications

The work performed during this DPhil resulted in the following research papers:

Peer-reviewed publications

1. **Kovaltsuk** A, Krawczyk K, Galson JD, Kelly DF, Deane CM, Trück J. (2017) *How B-Cell Receptor Repertoire Sequencing Can Be Enriched with Structural Antibody Data*, Frontiers in Immunology 8:1753. doi: 10.3389/fimmu.2017.01753
2. Krawczyk K, Kelm S, **Kovaltsuk** A, Galson DJ, Kelly D, Trück J, Regep C, Leem J, Wong WK, Nowak J, Snowden J, Wright M, Starkie L, Scott-Tucker A, Shi J, Deane CM. (2018) *Structurally Mapping Antibody Repertoires*, Frontiers in Immunology , 9:1698
3. **Kovaltsuk** A, Krawczyk K, Kelm S, Snowden J, Deane CM. (2018) *Filtering Ig-seq data using antibody structural information*, Journal of Immunology, 201(12):3694-3704
4. **Kovaltsuk** A, Leem J, Kelm S, Snowden J, Deane CM, Krawczyk K. (2018) *Observed Antibody Space: a resource for data mining next generation sequencing antibody repertoires*, Journal of Immunology, 201(7): 2502-2509
5. Krawczyk, K, Raybould MIJ, **Kovaltsuk** A and Deane CM. (2019) *Looking for Therapeutic Antibodies in Next Generation Sequencing Repositories*, mAbs, 11(7):1197-1205
6. **Kovaltsuk** A, Raybould MIJ, Wong WK, Marks C, Kelm S, Snowden J, Trück J and Deane CM. (2020) *Structural Diversity of B-cell Receptor Repertoires along the B-cell Differentiation Axis in Humans and Mice*, PLoS Computational Biology, 16(2):e1007636
7. Ghraichy M, Galson JD, **Kovaltsuk** A, von Niederhäusern V, Schmid JM, Miho E, Kelly DF, Deane CM and Trück J. (2020) *Maturation of the Human Immunoglobulin Heavy Chain Repertoire With Age*, Frontiers in Immunology , 11:1734
8. Raybould MIJ, **Kovaltsuk** A, Marks C and Deane CM. (2020) *CoV-AbDab: the Coronavirus Antibody Database*, Bioinformatics, ():btaa739
9. Galson JD, Schaetzle S, Bashford-Rogers RJM, Raybould MIJ, **Kovaltsuk** A, Kilpatrick GJ, Minter R, Finch DK, Dias J, James L, Thomas G, Lee WYJ, Betley J, Cavlan O, Leech A, Deane CM, Seoane J, Caldas C, Pennington D, Pfeffer P and Osbourn J. (2020) *Deep sequencing of B cell receptor repertoires from COVID-19 patients reveals strong convergent immune signatures*, Frontiers in Immunology, (Accepted)

10. Raybould MIJ, Marks C, **Kovaltsuk** A, Lewis AP, Shi J and Deane CM. (2021) *Public Baseline and Shared Response Structures Support the Theory of Antibody Repertoire Functional Commonality*, PLoS Computational Biology, 17(3):e1008781

Statement of Work Ownership

The use of the first personal plural in the thesis is stylistic. The work presented herein was performed by me, unless otherwise stated. If the work was carried out by other researchers, their contribution is appropriately acknowledged.

Abstract

Future successful exploitation of antibodies as diagnostic and therapeutic agents will greatly benefit from an increased understanding of natural B-cell receptor (BCR) repertoire diversities. The advent of next-generation sequencing of immunoglobulin genes (Ig-seq) has made it possible to sequence large snapshots of BCR repertoires in a single experiment.

In the results chapters of this thesis, we begin by describing a method (AntiBOdy Sequence Selector, "ABOSS") for filtering BCR repertoire data, which considers the structural viability of each sequence and is orthogonal to all other current methods (Chapter 2). ABOSS leverages the presence/absence of a conserved disulphide bridge found in antibodies as a way of both identifying structurally-viable BCR sequences and estimating the sequencing error rate. We show that this method is able to identify structurally impossible sequences missed by common error-correction methods.

Next, we describe the development of Observed Antibody Space (OAS), the first resource that curates BCR sequences from publicly available studies. As of October 2020, OAS contains more than 1.9 billion sequences from 85 studies. In OAS, all BCR repertoire sequences are annotated and profiled for structural viability.

We next describe the development of a novel method (SAAB+) to interrogate complete BCR repertoires at the structural level (Chapter 4). SAAB+ annotates large portions of BCR repertoires with three-dimensional information by mapping sequences to crystallographically solved antibody structures. By applying SAAB+ to BCR repertoires in OAS we, for the first time, document repertoire structural changes along the B-cell maturation axis in humans and mice.

In the final experimental chapter, we describe our work in COVID-19 research where we have compared the structural and sequence diversities of SARS-CoV-2 BCR repertoires to healthy repertoires deposited in OAS. We also outline the development of the first organised database (CoV-AbDab) that curates all publicly available anti-SARS-CoV-2 antibodies in a standardised format.

Finally, we discuss how recent developments in paired-chain Ig-seq platforms and deep learning algorithms could have a lasting impact on established Ig-seq analysis pipelines. We also outline how the tools described in this thesis can be combined with these field-disruptive technologies to advance our understanding of the immune system and improve computational antibody engineering.

Abbreviations

ABOSS AntiBOdy Sequence Selector

ADA Anti-drug antibody

AID Activation-Induced Cytidine Deaminase

AIRR Adaptive Immune Receptor Repertoire

BCR B-Cell Receptor

cDNA complementary DNA

CDR Complementary Determining Regions

CH Constant heavy

CL Constant light

COVID-19 Coronavirus disease 2019

CPU Central processing unit

CSR Class switching recombination

DBSCAN Density-Based Spatial Clustering of Applications with Noise

DTW Dynanamic Time Warping

ENA European Nucleotide Archive

ESS Environment-specific Substitution Score

Fab domain Fragment antigen binding domain

Fc region Fragment crystallisable region

FDA Food and Drug Administration

FO Follicular

Fv Variable Fragment

GC Germinal Center

gDNA genomics DNA

HMM Hidden Markov Model

HSC Hematopoietic precursor cell

Ig-seq Next-Generation Sequencing of immunoglobulin gene repertoires

IMGT International ImMunoGeneTics information system

MiAIRR Minimal standards adopted by the Adaptive Immune Receptor Repertoire

MZ marginal zone

NCBI National Center for Biotechnology Information

OAS Observed Antibody Space

PC Plasma B-cell

PCA Principal Component Analysis

PDB Protein databank

PPi Pyrophosphate

PSSM Position specific substitution matrix

RAG Recombination-activating gene

RBD Receptor binding domain

RMSD Root-mean-square-deviation

SAbDab Structural antibody database

SARS-CoV-2 Severe acute respiratory syndrome coronavirus 2

SHM Somatic hypermutation

TCR T-Cell Receptor

Contents

1 Introduction

Contents

This chapter contains reproduced material from the following publication:

1. **Kovaltsuk, A.**, Krawczyk, K., Galson, J.D., Kelly, D.F., Deane, C.M. & Trück, J. (2017) How B-Cell Receptor Repertoire Sequencing Can Be Enriched with Structural Antibody Data, Frontiers in Immunology, 8:1753

1.1 Antibody Biology

1.1.1 B-cells and the adaptive immune system

B-cells are the primary components of the adaptive immune system in jawed vertebrates. Their main function is to recognise and neutralise pathogens (antigens) *via* membrane-anchored B-cell receptors (BCR) or antibodies, a secreted version of BCRs. In this thesis, we will use BCRs and antibodies interchangeably to describe general B-cell encoded binding sites. Antibodies facilitate antigen neutralisation *via* three main pathways. These include direct antibody opsonisation, the recruitment of the complement system or activating effector cells of the immune system [1]. The majority of B-cells reside in the secondary lymphoid tissues where they profile the surrounding environment against pathogenic molecules. Only 2% of all peripheral B-cells belong to antibody secreting plasma cells [2].

The collection of all somatically recombined BCR sequences expressed in the body at a point in time is known as a BCR repertoire. The theoretical diversity of a human BCR repertoire is estimated to be in the range of 10^{13}-10^{18} unique molecules [3, 4]. This enormous sequence diversity confers the capacity for recognising potentially any given antigen [5]. No organism can realistically produce all theoretically-possible BCRs at once due to the body size constraints, but rather relies on a much smaller diversity that is hopefully sufficient to bind to the majority of antigens. In the human body, it is estimated that only 10^{11} B-cells are present at a single time point [6].

Studying BCR repertoire activation, maturation and expansion in response to antigenic stimulation has proved instrumental in understanding the dynamics of the adaptive immune system. For example, cross-repertoire clonal sequence convergence and enrichment can be used to isolate antigen-specific antibody sequences [7]. This is now facilitated by the advent of next-generation sequencing of immunoglobulin genes (Ig-seq), which allows scientists to gather comprehensive snapshots of BCR repertoire diversity across organisms, individuals and immune states [6]. Ig-seq technology has been used in vaccine design, drug discovery and immunodiagnostics [8, 9].

1.1.2 Antibody composition

Antibodies (BCRs) are Y-shaped immunoglobulin (Ig) proteins with an average mass of 150 kDa. They are composed of a pair of heterodimeric (heavy and light) polypeptide chains (Figure 1.1). The heavy (HC) and light (LC) chains are products of two independent mRNA transcripts that co-assemble into full-length immunoglobulin molecules in the endoplasmic reticulum of the B-cell. The full-length antibody structure can be divided into two major portions based on biological functionalities. These regions are the crystallisable fragment (Fc region) and the antigen binding fragment (Fab domain) (Figure 1.1A).

1.1.2.1 Antibody Fc region

The Fc region is a homodimer that is located between the hinge region and the C-terminus of the two heavy chains (Figure 1.1A). There are five possible main Fc portions (isotypes) in humans (IGHM, IGHA, IGHD, IGHG and IGHE), and which one is found on a particular antibody is governed by the maturity and function of that antibody. In humans, IGHG and IGHA isotypes are further stratified into subclasses (four IGHG and two IGHA) that differ in the hinge region composition, the number of glycosylation sites and the effector function [10]. Depending on the Fc region expressed, antibodies can exist as monomers (IGHG, IGHD, IGHE), dimers (IGHA) or pentamers (IGHM) in humans [11]. The main biological processes that are regulated by the Fc regions include recruitment of effector cells (natural killer cells, macrophages) and the complement system [1], regulation of antibody serum half-life [12, 13] and antigen binding avidity [11], and antibody localisation within the body [14]. The naïve (e.g. antigen-unexperienced) B-cells display BCRs with IGHM isotypes, which are usually low affinity and high avidity binders against their cognate antigens [15]. The biological functions of the Fc region can be amended for therapeutic applications using rational antibody engineering [16].

1.1.2.2 Antibody Fab region

The Fab region consists of both the light chain and the portion of the heavy chains that lies between the hinge region and the N-terminus (Figure 1.1). The part consisting of the variable domains of the heavy (VH) and light (VL) chains, which contain the area responsible for binding to the antigen, is known as the variable fragment (Fv). Within each B-cell, the antibody Fv domains are assembled by somatic recombination between V(D)J gene segments from heavy and one of the two light chain loci (λ

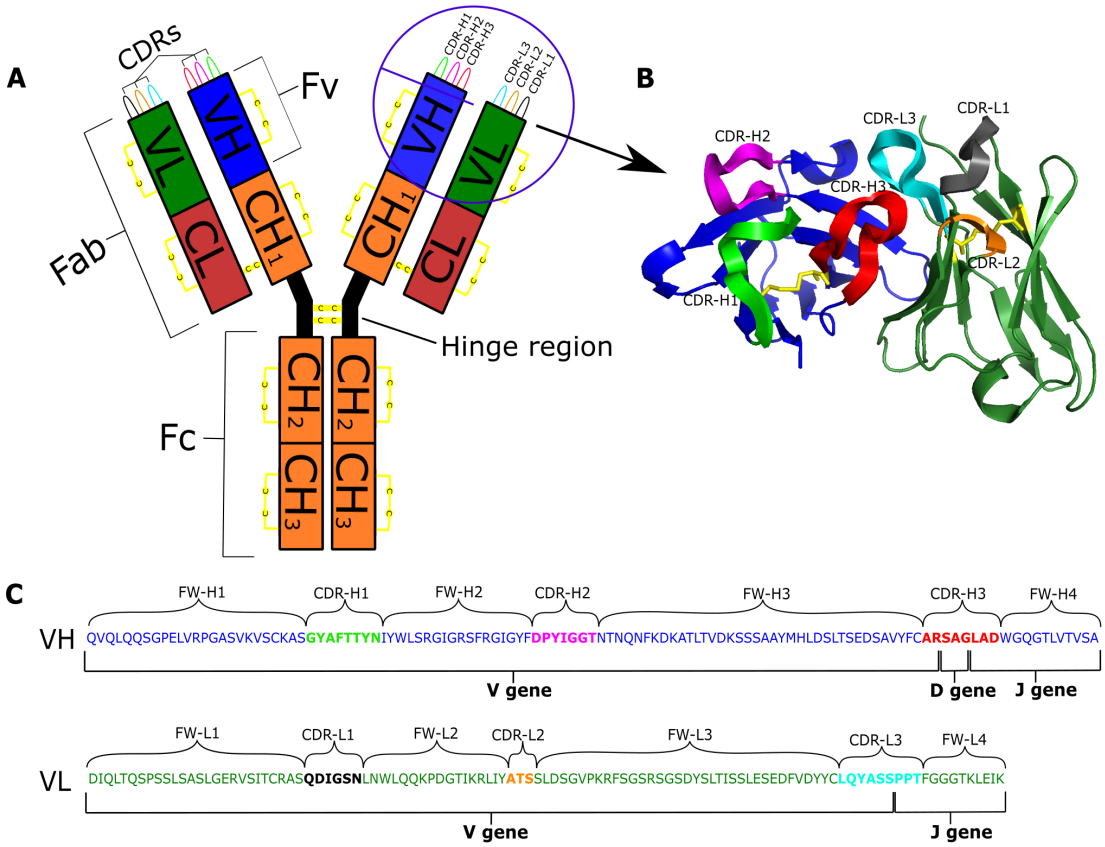


Figure 1.1: **Antibody structure and sequence composition.** **(A)** Schematic representation of an antibody IGHG structure. **(B)** Structural representation of the Fv region. **(C)** Genetic composition of VH domian and VL domian chains [IMGT numbering [17]]: VH is coloured blue; VL is green; CDRs are labelled and depicted in different colours; and disulphide bonds are in yellow.

and κ) (see Section 1.1.5 for more details) [18]. The light chain loci lack D gene segments, reducing its overall sequence diversity. In human, there are 56 V, 23 D and 6 J identified productive heavy chain germline genes, and 41 V and 5 J light chain germline genes in the λ locus, and 33 V and 5 J germline genes in the κ locus [19].

1.1.2.3 Antibody CDR regions

Within each of the VH and VL chains lie three hypervariable loops, the complementarity determining regions (CDR), which are the most diverse parts of the antibody. These loops form the majority of chemical interactions with antigens [5]. The CDR3 of the heavy chain (CDR-H3) is the most sequence and structurally diverse of the CDRs as it is formed at the joining region between the V, D, and J gene segments, includes non-templated insertions and is subjected to high levels of somatic hypermutation [15]. As a result of this diversity, CDR-H3 plays a key role in antigen recognition and binding [20, 21]. For this reason, CDR-H3 sequences are often taken as a proxy to assess antibody binding properties [22, 23]. However, sequence changes outside CDR-H3 can alter the structure and chemistry of the antigen-binding pocket, which could lead to changes in antibody function [24]. Furthermore, about 20% of antigen interacting residues are found outside the antibody CDR regions [25]. It is not yet studied whether non-CDR antigen interacting residues are positively selected against a particular set of antigens [25].

1.1.2.4 Antibody framework regions

The non-CDR sections of the Fv domain are called the framework regions. These regions form a series of structurally rigid anti-parallel β-sheets that define the immunoglobulin fold [26]. Framework positions proximal to CDRs are known as framework anchor regions. These residues can influence the structural orientation of the CDR loops [27].

1.1.2.5 Antibody inter- and intra-domain disulphide bonds

Several disulphide bonds help maintain the immunoglobulin fold (Figure 1.1). One set of disulphide bonds holds the heavy constant domains together in the hinge region and another set connects the light and heavy chains. Intra-variable domain cysteine pairs play a crucial part in shaping the antibody Fv region and artificial disruption of these bonds leads to impaired stability, folding and antigen recognition (Figure 1.2) [28] [29]. While there is evidence that some antibodies can still fold when the disulphide bond within the heavy or light chain is ablated [30–32], such antibodies

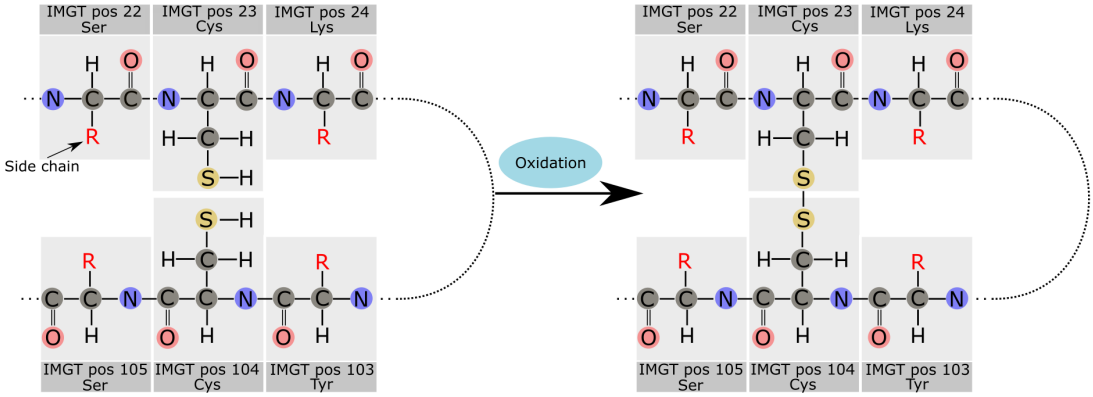


Figure 1.2: **Antibody intra-domain disulphide bond formation**. Several disulphide bonds help stabilise the immunoglobulin fold. All these bonds are formed between two closely-spaced cysteines in the oxidising environment of the endoplasmic reticulum. In the VH and VL chain, an intra-domain disulphide bond is formed between a pair of cysteines in framework 1 (IMGT position 23) and framework 3 (IMGT position 104) (see Section 1.1.3 for IMGT numbering details).

have been found *via* rational protein engineering where the conserved cysteines are mutated alongside further compensating modifications to the rest of the antibody sequence that stabilise the overall structure [30–32].

1.1.3 Antibody Sequence Annotation

Antibody sequence analysis is usually concentrated on the Fv domain, as it is the most sequence variable region and the major determinant of antibody binding properties. The Fv domain is not fixed in length owing to somatic recombination and junctional diversity between the V(D)J gene loci. To confer consistency and standardisation on antibody sequence annotation of alignments, several sequence numbering schemes have been developed [33, 34]. Such consistent annotation frameworks allow for the direct calculation of antibody sequence similarities and to formally define the most relevant sets of residues for antibody binding property assessment [34]. The IMGT scheme is an increasingly commonly-used method to annotate BCR repertoires [35]. This numbering was built considering both structural and sequence information [36]. The IMGT scheme supports symmetrical amino acid insertions inside CDRs, which ensures that structurally equivalent residues are assigned with matching positional identifiers regardless of CDR length (Figure 1.3). In contrast, the Chothia numbering is often used by structural biologists due to its simple CDR loop indel management and inherently structural focus [33, 37]. However, the Chothia numbering does not annotate structurally equivalent residues identically (Figure 1.4).

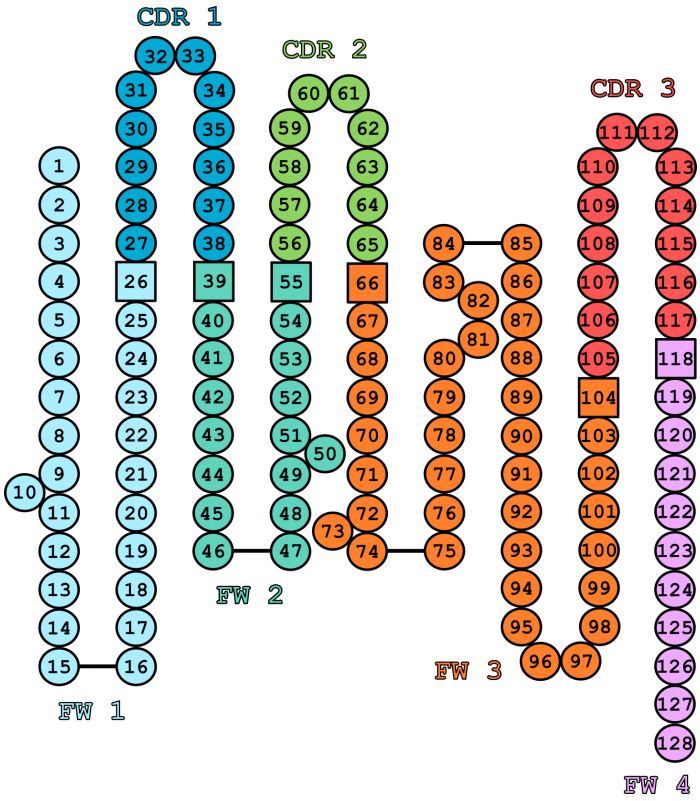


Figure 1.3: **IMGT/Collier de Perles representation [38] of an antibody VH and VL domain sequence**. Antibody regions are visualised in different colours. Both antibody VH and VL domains share the same antibody region definitions. Framework (FW) 1 region lies between positions 1 and 26, CDR 1 between 27 and 38; FW 2 between 39 and 55; CDR 2 between 56 and 65; FW 3 between 66 and 104; CDR 3 between 105 and 117; FW 4 between 118 and 128. A small number of framework residues were known to be absent in at least one of the three antibody gene loci [35]. For instance, IMGT positions 10 and 73 are not present in the VH germline genes in humans. Such positions are depicted as protruding nodes in the IMGT/Collier de Perles representation. Squares represent the framework anchor positions that surround the CDR loops.

One of the principal differences between numbering schemes is how they define CDR regions. Wu and Kabat [39] were the first to discover and define CDRs as portions of Fv chains that display high-sequence entropy, but (as with numbering schemes), there is not a single CDR definition and different schemes are used for legacy reasons or for specific features (such as insertion management in IMGT). The different numbering schemes define antibody CDR positions very consistently with the exception of CDR-H1 and CDR-H2 [40].

IMGT numbered BCR repertoire sequences can reveal the usage of amino acids at structurally equivalent positions. This analysis will enable scientists to interrogate

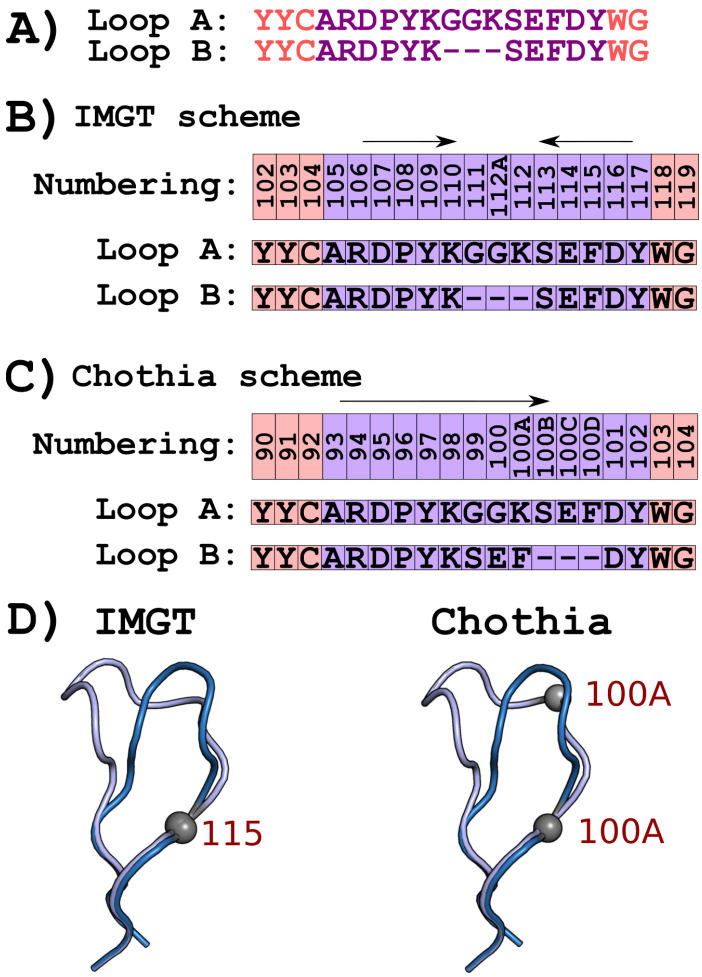


Figure 1.4: **Comparison of IMGT and Chothia numbering schemes**. **(A)** Two varying length CDR-H3 loops are aligned. Salmon colour depicts framework regions and violet colour indicates the CDR-H3 region. **(B)** Loop numbering according to the IMGT scheme. The IMGT scheme provides inward symmetrical loop numbering. If the loop is longer than 13 residues (Loop A), additional positions are introduced at the loop apex in the following order: 112A→111A→112B, *etc.* [36]. If a CDR-H3 loop is shorter than 13 amino acids, gaps are created in the middle of the loop (Loop B). **(C)** The Chothia scheme supports unidirectional CDR-H3 loop numbering. For loops longer than 10 residues, additional positions are sequentially created at the distal loop end (position 100). **(D)** Spatial localisation of equivalently numbered residues according to the (left) IMGT or (right) Chothia schemes. The models were generated using the ABodyBuilder web-server [41] and the resultant structures were aligned in pymol [42].

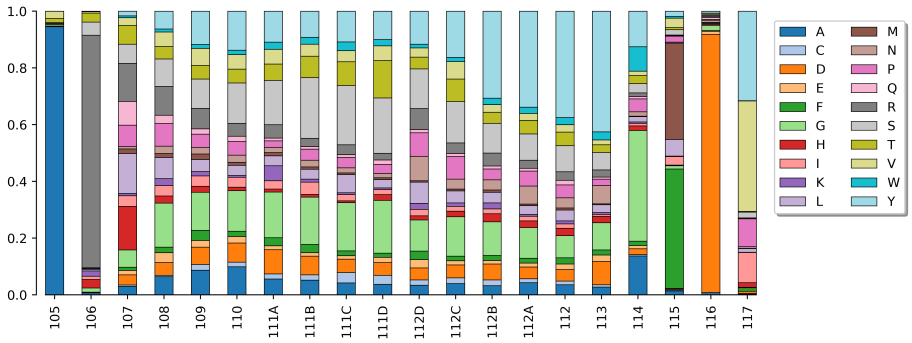


Figure 1.5: **Amino acid distribution across positions in CDR-H3 loops in a healthy BCR repertoire.** Positional amino acid usages are not uniform in the CDR-H3 loop. Distal positions are more sequence conserved (as being encoded by V and J germline genes), whilst central positions are highly sequence entropic. This agrees with our knowledge of antigen recognition, where the central positions are often incorporated into paratope structure, whilst the distal positions play a structurally supportive role [21]. Plotted along the X-axis are CDR-H3 positions according the IMGT numbering scheme. The Y-axis depicts amino acid proportional usages. Soto *et al.,* [3] study was employed as a source of a healthy BCR repertoire. IMGT numbered sequences used in Soto *et al.,* study [3] were downloaded from the Observed Antibody Space (OAS) resource [43].

positional entropy in antibody sequences that can shed light on the potential role of each position in antibody function. Highly conserved positions are more likely to have a role in antibody scaffolding, whilst highly entropic positions are more likely to contribute to the formation of the antigen binding site (Figure 1.5).

Several tools have been developed to assign numbering schemes to millions of BCR sequences. The two most commonly used tools are Abnum [33] and ANARCI [34]. Abnum uses several conserved positions in a sequence to identify relationships between antibody regions [33]. However, numbering outputs with Abnum can be inaccurate when the conserved positions are missing or mutated. Furthermore, Abnum does not take into account the amino acid composition of the antibody sequence outside these conversed positions. This is particularly important when BCR repertoires are analysed, as these are known to accumulate a large number of sequencing errors [44]. For instance, the Illumina platform is predominantly selected for BCR sequencing [43]. This state-of-the-art technology suffers from incorrect base calling, which could lead to sequencing errors in the conserved antibody positions and corrupted annotation by Abnum. The somatic hypermutation (SHM) machinery of antigen-experienced B-cells further diversify BCR repertoires by recursively introducing point mutations into the recombined V(D)J gene [45]. This could potentially cause a mutation in

the conserved antibody position, rendering the ability of Abnum to annotate of the sequence.

ANARCI is a more recently developed numbering tool that supports a wider range of numbering schemes compared to other tools [34]. In contrast to Abnum, ANARCI considers the amino acid composition of the entire antibody sequence for numbering rather than relying on a small set of conserved residues. First, ANARCI pre-builds Hidden Markov Model (HMM) profiles of a typical antibody sequence for different species based on the combination of available V and J germline genes. These genes are automatically downloaded from the IMGT website upon ANARCI installation [36]. As a result, ANARCI can build different antibody HMM profiles depending on the current availability of the IMGT germline genes. Next, each sequence is scored against the HMM profiles and the highest scoring alignment is identified. This allows for more accurate numbering, identification of indels as well as removal of unaligned sequences (poor quality/erroneous data).

1.1.4 Antibody CDR structure

Structural superimposition of CDR loops has suggested that all CDRs, except for CDR-H3, adopt a restricted number of conformations, termed canonical classes [46, 47]. The canonical classes link sequence patterns to a defined structure [47–49]. This enables the accurate prediction of canonical class structure from sequence. Over the last 30 years there have been several attempts to cluster CDR sequences/structures [37, 40, 46–51]. On the sequence level, the presence of certain cluster defining key residues can indicate the shape the loop can adopt [37, 47, 50]. Hence, some sequence changes to the canonical CDRs can be tolerated with no explicit change to loop conformations. The canonical CDR cluster information is constantly revisited as more antibody crystallographically-solved structures become available [49].

CDR-H3 shows a high degree of sequence, length and structure variation [52]. Due to this diversity, it has so far proven impossible to classify CDR-H3 loops into canonical classes in the manner of the other CDRs. It has been proposed that the base section of CDR-H3 can be categorised into ‘bulged’ or ‘extended’ conformations based on the presence of asparagine at position 116 (IMGT numbering) [53, 54]. However, increasing knowledge of CDR-H3 structural diversity has shown that the CDR-H3 bulged/extended configuration is difficult to predict solely from sequence [55].

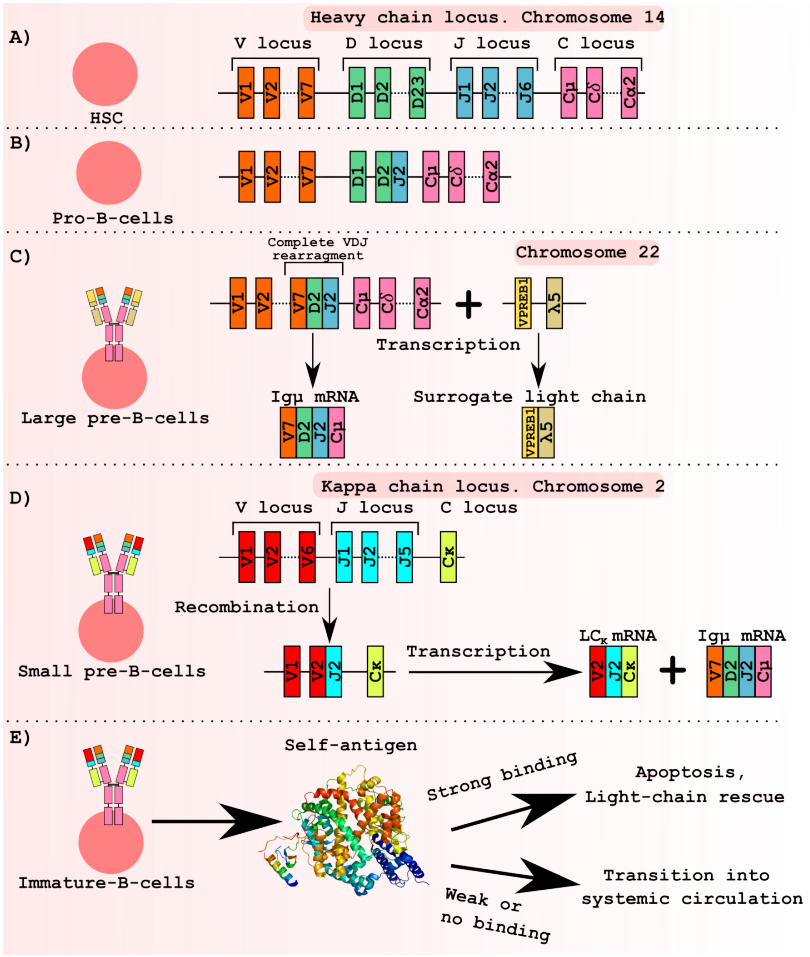


Figure 1.6: **Mechanism of BCR formation**. **(A)** RAG-1,2 enzymes initiate D-J gene recombination in the heavy chain locus of HSCs. **(B)** The V_H gene segment is joined with the recombined D_H-J_H to form the Igμ receptor in pro-B-cells. **(C)** The surrogate light chain is co-expressed with Igμ to form pre-BCR on the plasma membrane of large pre-B-cells. **(D)** Somatic recombination takes place between V_κ and J_κ light chain genes. The recombined light chain replaces the surrogate light chain to express a functional BCR on the small pre-B-cell surface. If the BCR is non-functional, several other V_κ and J_κ combinations are interrogated, before V_λ and J_λ light chain genes are tested (not shown). **(E)** Immature B-cells are profiled for self-reactivity before leaving the bone marrow for the first time. Self-reactive immature B-cells undergo apoptosis or light chain switching ("light chain rescue").

1.1.5 B-cell receptor formation

All B-cells originate from hematopoietic precursor cells (HSC). The prenatal development of B-cells takes place in the liver and bone marrow, and continues exclusively in the bone marrow after birth. B-cells can be broadly grouped into immature and mature developmental stages based on their function, body localisation and the func-

tional formation of their BCRs [56]. B-cells are regarded as immature before they exit the primary lymphoid organs into the systematic circulation for the first time. The immature B-cells can be subdivided into four developmental stages based on the assembly of their productive BCRs (Figure 1.6). The full description of the BCR formation process can be found in Appendix Section A.

1.1.6 B-cell maturation

Antigen-inexperienced transitional naive B-cells that exit the primary lymphoid organs for the first time migrate to the secondary lymphoid organs such as the spleen or lymph nodes. There they can undergo antigen-induced activation via one of two possible proliferation pathways.

1.1.6.1 Extra-follicular pathway

Naive B-cells that reside in the marginal zone (MZ) of the spleen proliferate into MZ B-cells (Figure 1.7B). These B-cells are often considered "innate"-like cells as they play a central role in early detection and elimination of blood-borne pathogens [57, 58]. Upon continuous antigenic stimulation, MZ B-cells terminally differentiate into short-lived plasma B-cells (PC) that secrete low-affinity polyreactive IGHM antibodies [57]. It was also shown that some MZ B-cells can enter the germinal centers to follow T-cell dependent activation.

1.1.6.2 Germinal center reaction

Activated naive B-cells that do not enter to the MZ proliferate into follicular (FO) B-cells. The FO cells undergo antigen-induced B-cell activation through differentiation into short-lived plasma or germinal center (GC) B-cells (Figure 1.7C) [59]. The GC consists of a light and a dark zone. The GC reaction starts in the dark zone, where B-cells undergo rapid expansion and class switching recombination (CSR) in their C_H chain, followed by somatic hypermuation (SHM) in their Fv domain [59, 60]. The mutated B-cells enter the light zone where B-cell clonal selection takes place. B-cells whose BCRs have higher affinities against their cognate antigen are positively selected, others undergo apoptosis. GC B-cells are able to circulate between the dark and light zones where multiple rounds of SHMs and selections are performed. This BCR paratope structural refining process is called affinity maturation. B-cells that are positively selected differentiate into memory or long-lived plasma B-cells [56].

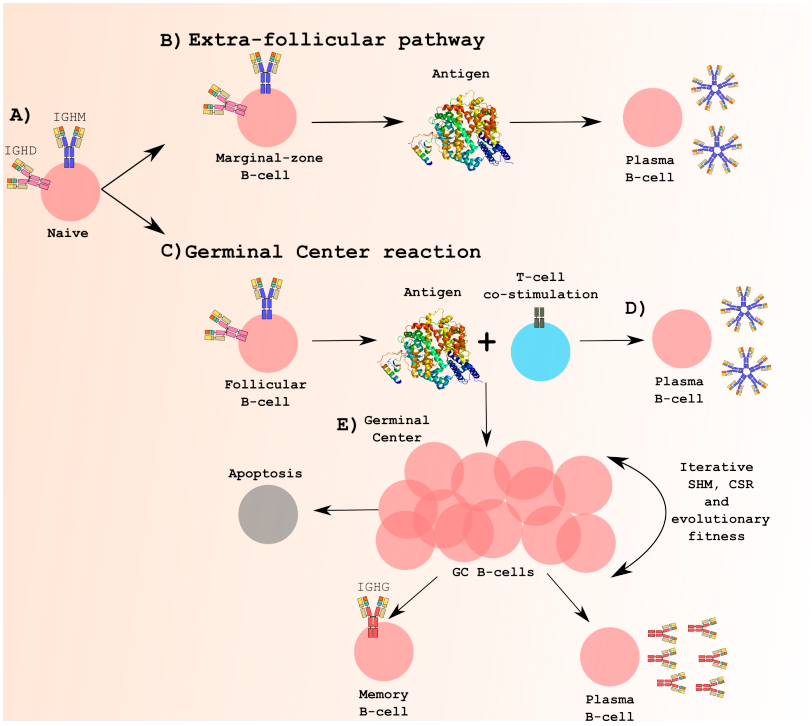


Figure 1.7: **B-cell maturation pathways in humans. (A)** Mature naive B-cells migrate into the secondary lymphoid organs. **(B)** B-cells that migrate to the marginal zone (MZ) of the spleen differentiate into MZ B-cells. Upon antigenic stimulation, MZ cells quickly turn into short-lived plasma B-cells that secrete low-affinity and high-avidity IGHM antibodies. **(C)** Naive B-cells that do not migrate to the MZ differentiate into follicular (FO) B-cells. Upon antigen recognition and T-cell co-stimulation FO cells can either become **(D)** short-lived antibody secreting B-cells or **(E)** form germinal centers (GC). Within GCs, B-cells undergo successive rounds of somatic hypermutations and class switch recombination. B-cells with favourable evolutionary profiles differentiate in memory B-cells or long-lived IGHG antibody secreting B-cells.

Antibodies generated via the GC reaction have higher specificity and affinity for their cognate antigens [57].

1.1.6.3 Somatic hypermutation

Somatic hypermutation (SHM) is the main component of affinity maturation in antigen-responding B-cells [61]. The SHM process starts by recruiting the Activation-Induced Cytidine Deaminase (AID) enzyme to the V(D)J segment in the B-cell genome. AID deaminates cytosines to highly mutagenic uracil (C→U). This base pair mismatch recruits B-cell's DNA repair machinery to excise or repair the mismatching nucleotides. If not repaired this creates C-T and G-A transitions. The repair ma-

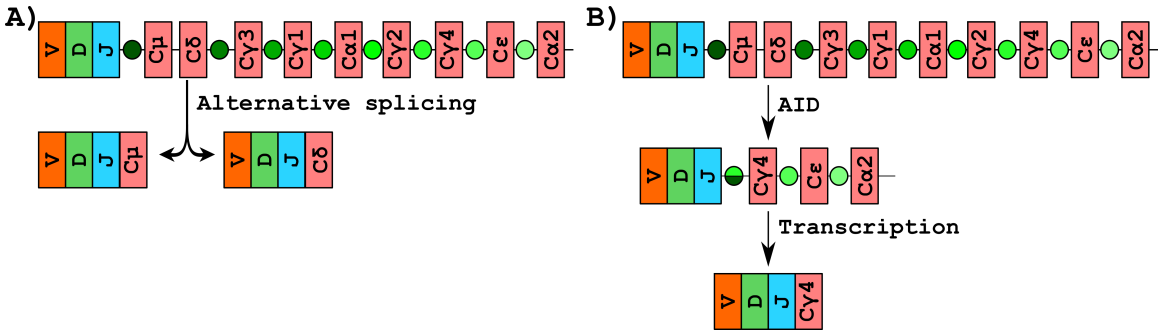


Figure 1.8: **Heavy chain class switch recombination (CSR)**. **(A)** The C_H locus of naive B-cells contains nine isotype exons separated by switch regions (circles), with the exception of IGHM and IGHD. Naive B-cells display BCRs with two separate isotypes (IGHM and IGHD) *via* alternative splicing. **(B)** Upon antigenic-stimulation, the AID enzyme migrates to the C_H locus where it recombines two switch (S) regions. The first S region lies upstream of the C_H locus, whilst the second S region is selected downstream the locus. As a result, a BCR changes expression of its Fc domain. Next, newly synthesised BCR is profiled for evolutionary fitness in the dark zone of germinal centers. The CSR process is iterative, where BCRs can change their Fc region several times until optimal evolutionary profiles are achieved.

chinery itself is error-prone which introduces further mutations at the original lesion site [62]. The mutation rate caused by AID is 10^6 times greater per base pair per cell division in the V(D)J segment than the background genomic DNA mutation rate in non-AID targeted genes [45].

AID genomic DNA targeting is not uniform across the V(D)J segment [62]. AID preferentially mutates two hot-spot motifs WRCY/RGYW and WA/TW (where W = A/T, R = G/A, Y = C/Y), whilst its activity is suppressed at the SYC/GRS cold-spot motif (where S = C/G) [62]. The distribution of these motifs also varies across antibody regions. An increased presence of the hot-spot motifs is recorded in CDRs when compared to framework regions. This is concomitant with higher levels of recorded SHMs in the CDR regions [62, 63].

1.1.6.4 Class switch recombination

B-cells change the expression of antibody isotypes to regulate their effector function and spatial distribution throughout the body [64]. Similarly to SHM, the AID enzyme initiates class switch DNA recombination (CSR) in antigen-experienced B-cells. The human C_H locus encodes for 9 isotype exons (Figure 1.8). These isotype genes are separated by switch regions, with the exception of IGHM and IGHD [65].

CSR performs the intra-chromosomal deletion of one or more antibody isotype

gene segments by recombining two switch regions. The first participating switch region is often located upstream of IGHM, whilst the second region can be any of the downstream regions in the C_H locus [64]. As a result, the AID activity excises the IGHM/IGHD exons from the B-cell genome, and the cell switches to expression of another isotype type (Figure 1.8B) [64].

1.1.7 Antibody formats across jawed vertebrate species

The adaptive immune system developed in jawed fish approximately 500 million years ago [66]. Two major events are thought to give rise to the adaptive immune system: whole genome duplication and introduction of RAG transposon motifs. Despite varying mechanisms of somatic recombination between V(D)J genes, secreted antibodies display a remarkable structural conservatism across the majority of jawed vertebrate species [66]. Most notable dissonance from the conventional antibody format is observed in cows, chicken, rabbits, sharks and camelids. For instance, alongside conventional antibody production, Camelids also secrete non-conventional antibodies (nanobodies) that consist of a homodimer of two heavy IGHG chains [67]. Nanobodies lack the CH1 region in their IGHG domain causing ablation of light chain association. As a result, the nanobody VH domain is solely responsible for antigen recognition and binding. To compensate for the loss of the light chain and the reduced antigen interaction interface, the average CDR-H3 loop length is longer in nanobody repertoires [68]. To support this extended loop configuration, nanobodies often form an additional intra-domain disulphide bridge between the CDR-H1 and CDR-H3 regions [67].

The exact evolutionary purpose of camelid nanobody generation still remains elusive [67]. Their favorable biochemical properties (size, stability, affinity) have made nanobodies a successful class of biotherapeutic, diagnostic and research agents [69].

1.1.8 Antibodies as therapeutic agents

The properties of antibodies, in particular their antigen recognition specificity and binding affinity, have made them useful as diagnostics and research agents as well as the most successful class of biopharmaceuticals [70]. The modular nature of antibodies means that novel antibody formats can be engineered to expand on antibody functionalities [70]. Although small molecules constitute the largest proportion of potential therapeutics in clinical trials, the antibody market is steadily growing, with new antibody approvals at a rate of about five per year. As of 2020, five out of the

10 best-selling drugs worldwide in terms of revenue were recombinant monoclonal antibodies [71].

1.1.8.1 Clinically approved antibodies

The first antibody that received its FDA approval was muromonab in 1985 [72]. As of October 2020, 93 therapeutic antibodies were approved for clinical use [73]. Several publicly available databases have been created to keep track of clinical trial progression of each antibody that is filed for the therapeutic application [73, 74]. These databases are a crucial source of successful examples of therapeutic antibody designs.

Antibodies are large proteins, which means that they can only interact with extracellular targets such as plasma membrane proteins or cytokines [75]. For instance, trastuzumab is a highly successful therapy for breast cancer treatment where the HER2 receptor is over-expressed on the cell surface leading to uncontrolled B-cell proliferation. Trastuzumab binding downregulates HER2 expression which restores the normal cell cycle [76].

The protein nature of antibodies also limits the potential routes of drug administration. Sub-cutaneous and intravenous injections minimise initial antibody enzymatic proteolysis during drug administration, whilst the oral route compromises molecular integrity [75].

1.1.8.2 Antibody discovery platforms

Hybridoma and display technologies (phage, ribosome and yeast) are the two best-established antibody discovery platforms. These technologies strive to develop high affinity and high specificity binders against the target of interest.

Hybridoma is an *in vivo* technology that leverages the immune system of various animals models to isolate target-specific monoclonal antibodies (Figure 1.9A). First, an animal (usually mouse or rabbit) is immunised with a desired antigen. After the immune response is mounted (two weeks), the animal is sacrificed and its splenic B-cells are harvested. B-cells are then immortalised by fusion with cancer myeloma cells [77]. Immortalised B-cells are seeded into individual wells where they are induced to secrete antibodies. In each well, antibody containing medium is screened in functionality assays. The isolated target-specific antibodies act as initial lead candidates for therapeutic antibody engineering campaigns. Further amino acid mutations are introduced to these antibodies to improve affinities and to attain favourable developability profiles (solubility, immunogenicity, expression and physical stability) [28].

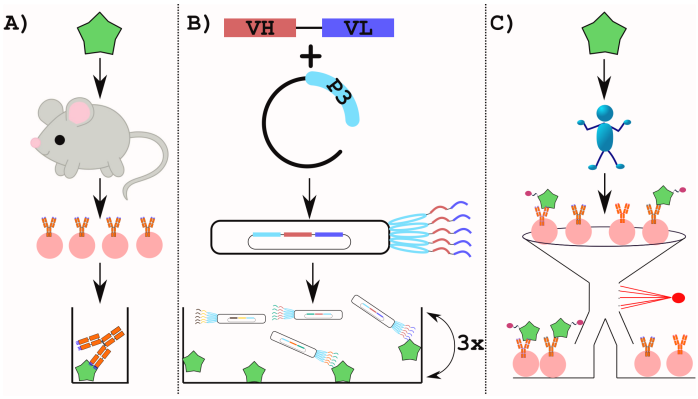


Figure 1.9: **Antibody discovery platforms. (A)** Hybridoma technology raises antigen-specific antibodies using *in vivo* animal models. Splenic B-cells are harvested from a convalescent animal model such as a mouse. Harvested B-cells are then induced and their secreted antibodies are screened in functional assays. **(B)** Phage display allows the synthesis of massive human scFv or Fab libraries. Each phage displays up-to five copies of the same VH/VL combination on its surface. These libraries are panned multiple times to enrich for antigen-specific phage. **(C)** Antigen baiting allows isolation of antigen specific B-cells directly from seropositive patients. Collected B-cells are incubated with the fluorescently tagged antigen. Then, fluorescent activated cell sorting (FACS) identifies B-cells whose BCRs are antigen complementary.

Phage display is an *in vitro* technology that leverages a surface minor coat protein (pIII) of M13 filamentous phage to display antibody single-chain Fv (scFv) regions [78] (Figure 1.9B). To create a display library, human somatically recombined or synthetic VH and VL genes are randomly joined and ligated into the M13 plasmid DNA next to the pIII gene. A size of a typical phage display library is 10^{11} unique VH/VL pairs [79]. In the library preparation process, phage pIII-scFv proteins are multiplied inside the bacterium *Escherichia coli*. The helper phage is then added to *E.coli* to finalise functional M13 phage assembly [78]. Phages, whose scFvs are target specific, are iteratively enriched in 2-3 rounds of biopanning selections [78]. Since there is a direct link between phage-displayed antibody genotype and phenotype, the sequence composition of scFv regions can be easily deciphered. The sequenced Fv sequences can now be grafted onto human IGHG (or other) scaffolds for developability and functional profiling [78]. As of 2020, the world best selling drug in terms of revenue (adalimumab) was discovered using phage display [71].

1.1.8.3 Antibody immunogenicity

Although antibody structural configurations are structurally conserved across mammals, any non-human antibodies are still non-native proteins to the human immune

system. Hence, antibody discovery by conventional hybridoma technology will be immunogenic in the human body [75]. This is a lesser issue with phage display where human VH and VL genes are used for the display library creation [79]. However, even fully human therapeutic antibodies can elicit an anti-drug antibody (ADA) response in a small number of patients [80].

A large part of improving antibody developability profiles is focused on minimising the ADA reaction. Multiple strategies have been developed to "humanise" non-human antibodies whilst preserving antigen binding. Antibody modular structure allows grafting an entire non-human Fv region onto human IGHG scaffolds creating chimeras. However, chimeric antibodies still elicit strong ADA response. To further reduce the ADA response, researchers often graft only CDR regions onto human antibody scaffolds [81].

Apart from the phage display technology, it is possible to generate fully human antibodies using genetically modified animals models or antigen baiting techniques. Kymouse is one of the successful mouse models that contains all human heavy and kappa light V(D)J loci [82]. The Kymouse model can be combined with the hybridoma technology to yield fully human antibodies. The mechanism of antigen baiting is based on single B-cell sorting [83]. B-cells whose BCRs are complementary against the fluorescently labelled antigen are separated (Figure 1.9C), which is followed by V(D)J gene segment sequencing [84].

1.2 Sequencing B-cell receptor repertoires

Since antibodies are secreted products of plasma B-cells, their diversity can only be assessed on the amino acid level using techniques such as liquid chromatography tandem-mass spectroscopy (LC-MS/MS) [85]. These approaches are expensive, time-consuming, provide limited sequence coverage and only focused on a subset of antibody secreting B-cells. Therefore, researchers often resort to next-generation sequencing of re-arranged immunoglobulin (Ig-seq) genes in BCR repertoires. Clonal diversity and sequence convergences across multiple B-cell donors is often used to isolate antigen-specific monoclonal antibodies [8, 22].

1.2.1 Ig-seq Technologies

With the development of Ig-seq platforms, immense volumes of immunoglobulin information can be generated at the nucleotide sequence level. This allows for the

interrogation of snapshots of BCR repertoire diversity at great depth. By controlling the number of replicates that are combined and the number of B-cells contained therein, it is possible to obtain a large fraction of a BCR repertoire from a blood sample [3, 4]. Datasets generated by Ig-seq have improved our understanding of immune systems across numerous species and have already been applied in vaccine development, drug discovery and immunodiagnostics [86–88]. The success of Ig-seq in antibody discovery relies on the ability to accurately identify the population of BCRs specific to antigens. By leveraging Ig-seq technologies, expansion and sequence convergence across B-cell clonal families have been observed in the response to a variety of antigens and pathogens [22]. This may serve as an additional way together with the well-established discovery platforms (e.g. phage display) to isolate antigen-specific antibodies through identifying sequences common amongst several individuals exposed to the same antigen [8, 89].

Several Ig-seq technologies have been developed in the last decade. These sequencing approaches yield varying depth and costs per base as well as suffer from platform-specific errors and biases.

1.2.1.1 Roche 454 pyrosequencing

In 2009, Roche 454 pyrosequencing was the first Ig-seq platform which was successfully applied to generate BCR repertoire diversity data [90, 91]. The early success of Roche 454 pyrosequencing was attributed to its great read length which was sufficient to cover full-length V(D)J amplicons. Roche 454 pyrosequencing was discontinued in 2013 as advances in short read Illumina sequencing and computational assembly methods of these reads provided overall superior data quality [92].

Roche 454 pyrosequencing works by isolating each single-stranded V(D)J gene amplicon onto a single sequencing bead via adapter annealing (Figure 1.10A). Roche 454 pyrosequencing employs the "sequencing by synthesis" method where enzymatic reactions are carried out to simultaneously extend and sequence DNA strands which are complementary to bead-immobilised amplicons (Figure 1.10B). Roche 454 pyrosequencing is a cyclic process, where only one of the four nucleotide types (A, G, T, C) is present in the sequencing reaction at a single time. Nucleotide incorporation into the complementary strand releases pyrophosphate (PPi) which is enzymatically converted into a chemiluminescent signal. Before the next nucleotide is added, all NTPs are washed away from the sequencing reaction. Recording the intensity of the

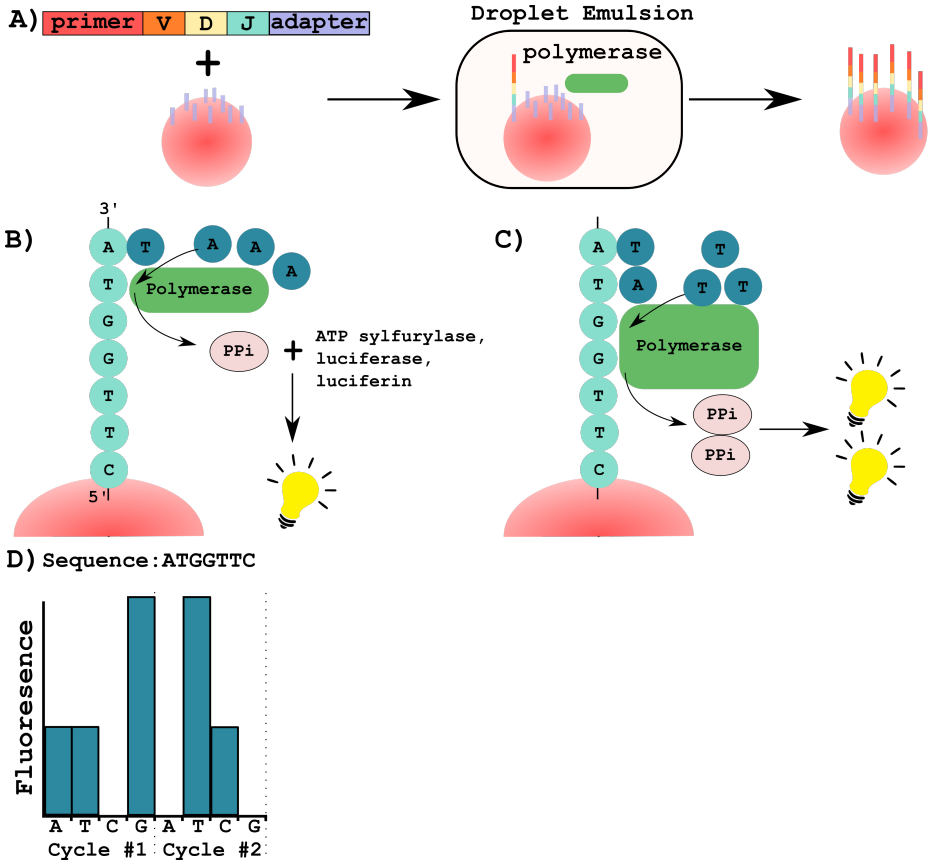


Figure 1.10: **Roche 454 pyrosequencing.** **(A)** A sequencing amplicon is prepared for Roche 454 pyrosequencing by attaching adapter and primer sequences to recombined V(D)J gene segments. The adapter sequence is required for amplicon attachment to sepharose beads that are covered with hybridisation DNA oligos. Amplicons and beads are mixed at defined concentrations to ensure than only a single amplicon is found in the surface of each bead. The beads are encapsulated with DNA polymerase into individual emulsion droplets to enrich the number of amplicon copies via PCR amplification. **(B)** During the "sequencing by synthesis" step, nucleotides of a single type are sequentially added to the reaction. If the added nucleotide type is complementary to the residue next in the sequence, it is then incorporate into the complementary sequence releasing pyrophosphate (PPi). The PPi molecule interacts with ATP sylfurylase, luciferase and luceferin to yield a chemifluorescent signal. **(C)** The intensity of the fluorescent signal is proportional to the number of nucleotides added to the complementary strand. **(D)** An exemplary output from Roche 454 pyrosequencing on ARGGTTC ssDNA. Roche 454 operates in cycles where one of the four nucleotides is sequentially added and then washed away from the reaction. Plotting along the X-axis are two sequencing cycles. The Y-axis shows the fluorescence readout proportional to the number of nucleotides added to the complementary strand.

chemiluminescent signal at each nucleotide addition step reveals the amplicon nucleotide sequence composition (Figure 1.10D). The estimated sequencing error rate of Roche 454 pyrosequencing is between 0.25%-1% per base [93].

1.2.1.2 Illumina sequencing

As of October 2020, Illumina MiSeq remains the most popular Ig-seq platform [4, 43, 94]. It provides the most optimal performance for BCR repertoire generation when compared to other available platforms: 2x300 base-pair read length, high read depth, low error rates, and low costs per base [4, 95, 96].

Similarly to Roche 454 pyrosequencing, Illumina technology also leverages "sequencing by synthesis" to decipher the nucleotide sequence composition of V(D)J amplicons (Figure 1.11). Single stranded V(D)J amplicons are hybridised to the Illumina flowcell which is covered in single stranded DNA (ssDNA) oligos. PCR "bridge amplification" forms dense homogeneous clusters for each amplicon. The reverse strands are enzymatically cleaved and washed away from the flowcell. Sequencing begins by adding fluorescently tagged nucleotides to the flowcell (Figure 1.11B). Each time a nuclotide is added to the growing complementary chain, a light signal is emitted. The fluorescent tags also serve as terminators preventing more than one nucleotide addition per cycle. In each amplicon cluster, a base call is made judging on signal intensity and homogeneity measurements (Figure 1.11C). The tags are enzymatically cleaved and the sequencing reaction continues to the next cycle. Once the forward read is sequenced, the sequencing product is washed away and the reverse strand is synthesised through "bridge amplification". The forward strand is now enzymatically cleaved and the "sequencing by synthesis" method is performed on the reverse strand. In this manner, millions of V(D)J amplicons can be sequenced in parallel. Illumina technology supplies a varying number of cycles depending on the platform. Illumina 2x300 MiSeq which is popular used for full-length BCR sequencing provides 300 cycles for both forward and reverse strands.

The Illumina platform outputs forward (R1) and reverse (R2) reads in the FASTQ format [97] for each sequenced amplicon cluster. The FASTQ format represents a text file containing nucleotide sequences, unique identifiers, and phred quality scores for each called base within the individual sequence. Illumina-outputted phred scores range between 2 and 40, where higher scores indicate a higher probability of the correct base call. Since V(D)J amplicons are ∼500 base pairs, and Illumina MiSeq R1 and R2 reads only span 300 base pairs, assembly methods are required to form BCR consensus sequences based on R1-R2 overlapping regions.

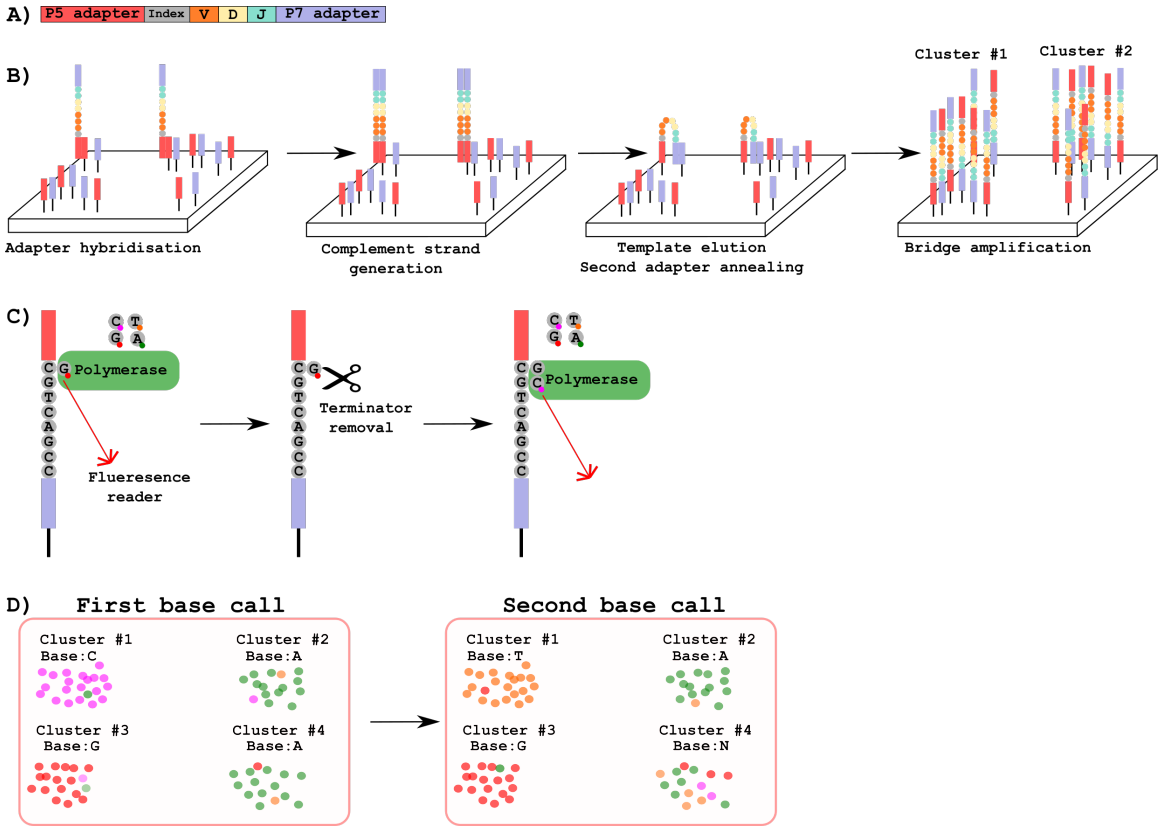


Figure 1.11: **Illumina Sequencing platform.** **(A)** Forward and reverse Illumina adapters are attached to respective ends of V(D)J transcripts to create sequencing amplicons. To allow simultaneous sequencing of several separate samples, short DNA indexing motifs are ligated to the amplicons. These indexes are unique to each sample. **(B)** The forward adapter sequence of the amplicons is first hybridised to its complimentary single stranded DNA (ssDNA) oligo found on the surface of the Illumina flowcell. A DNA polymerase creates amplicon's complementary sequence which is now covalently linked to the flowcell. Now the reverse adapter of the newly synthesised strand hybridises to its neighbouring oligos forming a "bridge". Next, PCR is applied repeatedly to enrich individual amplicons into dense homogeneous clusters. This process is called "bridge amplification". **(C)** "Sequencing by synthesis" ensures that only one nucleotide is added (and sequenced) to the growing complementary strand. All nucleotides participating in the reaction are fluorescently labelled. This label is unique to each nucleotide type. Upon nucleotide incorporation, a fluorescence reader records identities of the newly added nucleotides. The fluorescent label also acts as a reaction terminator preventing incorporation of more than one nucleotide in a single cycle. Next, all reaction present nucleotides are washed away and the nucleotide label is enzymatically removed, and the reaction proceeds to the next cycle. **(D)** Bridge amplification forms dense homogeneous amplicon clusters that are scattered across the flowcell. During each nucleotide addition cycle, a fluorescence reader makes a base call based on the intensity and homogeneity of the signal. If a base cannot be deciphered, it is designated as N.

1.2.1.3 Other Ig-seq technologies

Several other Ig-seq platforms are being actively developed to overcome Illumina sequencing limitations such as short read lengths. Two of these platforms include Oxford Nanopore [98] and PacBio [99] sequencing. In recent work, PacBio has been shown to be a robust research tool to identify yet uncharacterised murine V gene alleles, as this Ig-seq platform provides the high read length needed to profile V_H locus genomic DNA [99]. Longer read lengths also enable identification of BCR isotype information from V_H locus genomic DNA.

1.2.2 Paired VH/VL Ig-seq methods

Commonly-used methods for BCR repertoire generation rely on the bulk lysis of B-cells in order to extract somatically recombined V(D)J gene segments. However, this leads to ablation of native antibody heavy and light chain pairing information. As the CDR-H3 loop plays the key role in antigen-recognition [20], the majority of Ig-seq studies focus primarily on interrogation of unpaired heavy chain BCR repertoires [43]. The decision to avoid paired sequencing of natively associated VH/VL domain is primarily based on the complexity of high-throughput experimental setups needed to identify VH/VL mRNA molecules originating from the same B-cell [100]. However, native VH and VL pairing information is required to obtain deeper insights into antibody repertoire diversity and dynamics as well as to build structural representation of complete Fv binding sites [101, 102]. To circumvent this limitation, several paired VH-VL Ig-seq methods have been developed. All these approaches rely on individual B-cell encapsulation into oil emulsion droplets prior to DNA labelling or physical chaining of cognate VH and VL cDNA molecules.

1.2.2.1 Physical VH/VL cDNA linkage

In 2013, DeKosky *et al.,* developed the first technology for paired BCR repertoire sequencing [100]. Their method relies on encapsulation of single B-cells into droplets containing poly(dT) beads. Upon cell lysis, poly-A-tails of VH and VL mRNA molecules are captured onto the beads. A linkage reverse transcription is then performed to covalently link VH and VL cDNA. This method generates ∼850 base pair constructs that span (in 5'→3' direction) CH, VH, linker, VL and CL regions. Due to limited read length of current Illumina technology, the analysis is restricted to a 500 base pair fragment covering only CDR-H3, CDR-L3, and neighboring framework 4 and proximal positions of framework 3 of the respective chains.

1.2.2.2 10xGenomics V(D)J sequencing

In 2019, a new paired BCR repertoire sequencing technology, 10xGenomics V(D)J, emerged. 10xGenomics overcomes Illumina short read length limitations enabling scientists to sequence full-length paired VH/VL transcripts [103]. In contrast to the intricate experimental setup used by DeKosky *et al.,* [100], the 10xGenomics platform is commercially available, which provides an easy access to paired data generation.

10xGenomics technology relies on single B-cell isolation into separate emulsion droplets, where each droplet contains a single 10xGenomics gel bead. These beads act as a delivery system of barcoded DNA oligo, where each bead contains a unique version of a 10xGenomics barcode (Figure 1.12A). B-cells are lysed and the barcoded oligos are added to cognate VL and VH cDNA molecules within each emulsion droplet during reverse transcription (Figure 1.12B). After PCR amplification of the barcoded cDNA molecules, the products are subjected to enzymatic fragmentation in the V(D)JC segment. Sequencing adapters are then ligated to the length varying amplicons followed by Illumina sequencing. On average, each V(D)J re-arrangement is covered with one thousand reads. Short-read genome assemblers can then be used to generate full-length BCR sequences from raw data, where natively associated VH and VL sequences share identical 10xGenomics barcodes [103, 104]. The efficiency of various short-read genome assembly pipelines has not yet been systematically benchmarked on 10xGenomics data.

A 10xGenomics sequencing chip contains eight separate channels which enable multiple sample processing in parallel. Each channel can take up-to 10,000 B-cells, which is currently the main bottleneck to scaling up paired BCR repertoire sequencing. As a result, the repertoire depth attained with paired sequencing technology is significantly smaller when compared to unpaired sequencing platforms [3, 4, 103, 105]).

1.2.3 Sequencing sample preparation

Since Ig-seq was first developed in 2009, multiple protocols for sequencing sample preparation have been created. This variety of experimental setups is mainly linked to technological Ig-seq advances, reagent and equipment availability, and the scientific questions being investigated. Regardless of experimental setup, all these approaches share four key steps: 1) blood sample acquisition, 2) B-cell isolation 3) extraction of somatically re-arranged V(D)J gene segments (either mRNA or gDNA), and 4) ligation of sequencing adapters and index DNA barcodes. Due to high experimental

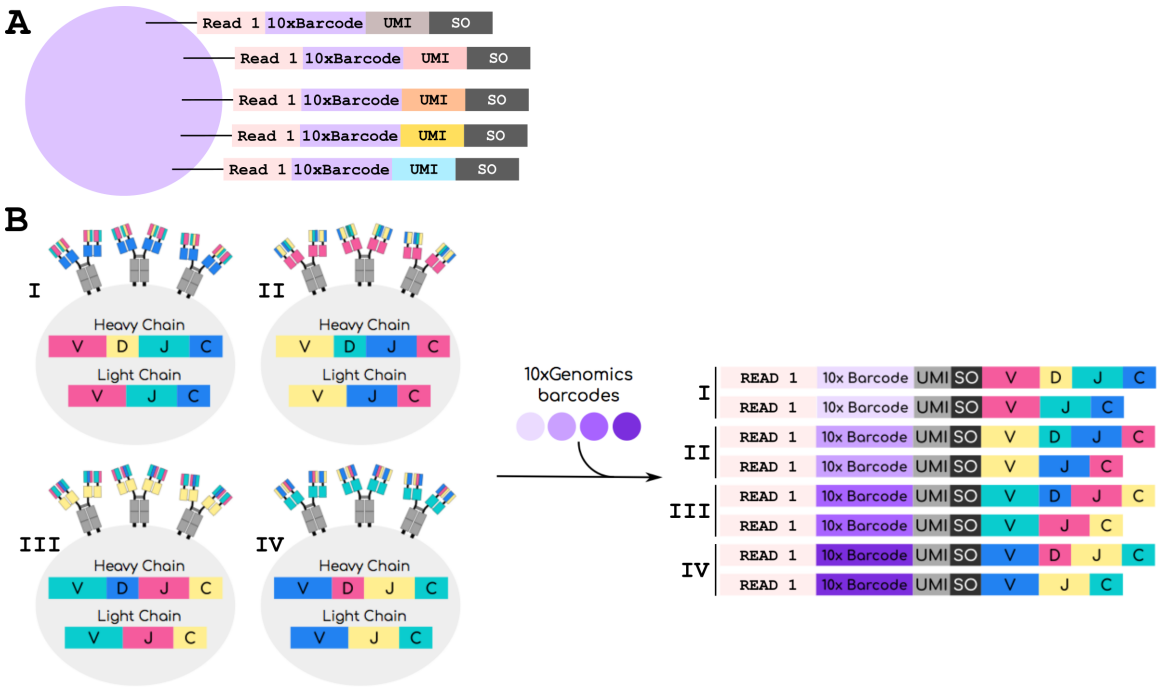


Figure 1.12: **10xGenomics V(D)J sequencing method. (A)** A 10xGenomics gel bead acts as a DNA oligo delivery agent. Each bead contains millions of oligos that share identical 16 nucleotide 10xGenomics barcode. The total number of unique 10xGenomics barcodes is estimated to be ∼750k. Each DNA oligo consists of (5'→3') read 1, 10xGenomics barcode, unique molecular identifier (UMI) and switch region (SO). The read 1 motif is required for sequencing adapter ligation. The switch region allows for the attachment of the DNA oligo to V(D)J transcripts. UMI motifs are unique across oligos found on the same bead. Be leveraging UMI motifs, it is possible to count how many different V(D)J cDNA molecules contributed to BCR contig generation. **(B)** Single B-cells are encapsulated into emulsion droplets with a single copy of 10xGenomics beads. Upon B-cell lysis, unique 10xGenomics barcodes are ligated to natively associated VH and VL transcripts. The constructs are then subjected to PCR amplification, enzymatic fragmentation and adapter sequence ligation prior to Illumina sequencing (not shown).

and labour costs, B-cells are not routinely sorted into separate B-cell sub-populations [43], which, unavoidably, leads to the loss of experimental resolution. Most Ig-seq protocols interrogate nearly full-length V(D)J segments, although some methods exclusively operate on the CDR-H3 region [106].

1.2.3.1 Choice of genetic material

The advantage of using mRNA over gDNA as the source of genetic information is the ability to sequence BCR isotype information as the switch region is removed during the primary transcription step (Figure 1.13). Reverse transcription of mRNA also allows the incorporation of unique molecular identifiers (UMIs) to correct for any potential sequencing errors and amplification biases [107].

There are two main approaches that create sequencing libraries using V(D)JC mRNA. In the first method, all V(D)JC mRNA molecules are converted to DNA with a mix of primers specific for constant heavy domains (Figure 1.13A). In the next step, resultant cDNA molecules are PCR amplified with a set of six BIOMED-2 V-gene specific primers and a universal 3'end binding primer [107, 108].

Another common technique to create sequencing libraries from mRNA is rapid amplification of cDNA ends, or 5'RACE. All mRNA molecules reverse-transcribed into cDNA with an oligo(dT) primer specific to the mRNA poly-A tail. The 5'RACE method also adds a universal priming site to the 3'end of cDNA (Figure 1.13B). Next, a nested PCR is performed to amplify V(D)JC cDNA molecules. A mix of C-gene primers is used to target the 3'end, whereas a single 5'end primer is employed to bind the universal priming site [109]. Hence, 5'RACE minimises any potential V-gene primer amplification biases.

Usage of gDNA as the source of V(D)J recombination information alleviates mRNA expression biases across B-cells. This is particularly important when working with plasma B-cells, where the mRNA expression rate is increased several fold compared to other B-cell subclasses [110]. At this moment, Illumina technology does not allow simultaneous coverage of full-length recombined V(D)J segments, switch regions and C-gene segments. Hence, isotype information data is lost during V(D)J gDNA sequencing. Similarly to mRNA amplification, a set of six BIOMED-2 primers are used to target V-genes and a single universal reverse primer is used to bind to the J-gene region [108, 111] (Figure 1.13C).

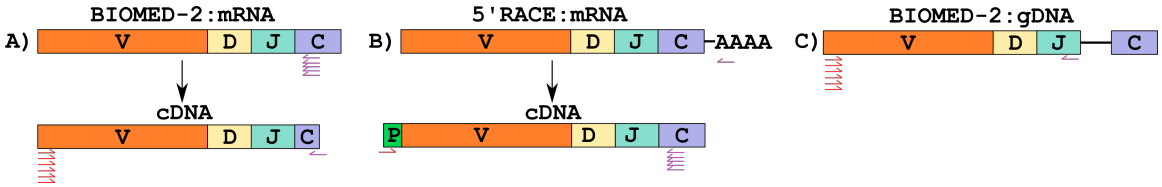


Figure 1.13: **BCR sequencing library preparation. (A)** A set of constant region specific primers is used to reverse-transcribe mRNA into cDNA. Next, a mix of BIOMED-2 primers and a single universal 3'end primer are employed to PCR amplify cDNA molecules [108]. **(B)** 5'RACE is used to reverse-transcribe all mRNA molecules using a poly-A tail specific primer. 5'RACE also adds a universal priming (P) site to all cDNA molecules. Next, a nested PCR is performed to amply V(D)J cDNA molecules. There, a set of constant region specific primers and a single P site specific primer are used. **(C)** To amplify recombined V(D)J gDNA, a set of BIOMED-2 V-gene primers and an universal J-gene primer are used.

1.2.4 Sequence-based analysis pipelines for BCR repertoire data

1.2.4.1 Unpaired BCR data processing

Advances in Ig-seq technologies have caused an explosion in BCR repertoire data availability. This, in turn, has prompted the creation of multiple start-to-finish BCR analysis pipelines. The two most commonly used pipelines are Immcantation [112, 113] and ImmuneDB [114, 115]. The output formats produced by these tools were used as a scaffold for the development of minimal standards adopted by the Adaptive Immune Receptor Repertoire (AIRR, "MiAIRR standards") community [116, 117]. The current MiAIRR standards mandate usage of flat tab-delimited text files with a predefined and consistent set of data descriptor columns. These requirements ensure that the wide range of BCR repertoire tools can be readily stacked to benefit from the cross-fertilisation of sequence interrogation methods.

The first step in BCR repertoire analysis is to pre-process outputs from Ig-seq platforms, which usually come in the FASTQ format [97]. In this step, reads are quality checked (Phred score), primer sequences are removed, consensus sequences are assembled. Then, V(D)J genes are identified for each BCR sequence usually using a local installation of IgBlastn [118] or the IMGT High/V-Quest web server [119]. As the IMGT server only permits separate submissions of 250,000 BCR sequences, IgBlastn is preferred to annotate millions of V(D)J recombinations [3]. IMGT GeneDB is usually used for the download of the latest version of germline genes [19]. These genes are used as a germline reference by the AIRR community tools [117]. In addition

to germline gene identification, both IgBlastn and IMGT High/V-Quest return other valuable BCR sequence descriptors such as the number of somatic hypermutations, isolated CDR and framework regions, nucleotide indel information and IMGT-scheme aligned nucleotide sequence.

In some BCR repertoire studies, all V(D)J cDNA molecules are labelled with unique molecular identifiers (UMIs) [107, 109, 120]. These barcodes provide a higher resolution on potential PCR amplification and sequencing biases. Bioinformatics tools have been developed to "de-bias" repertoires by grouping BCR sequences into separate clusters based on the UMI barcodes [113]. Within each cluster, the most redundant sequence is assigned as a consensus sequence. All other cluster sequences are collapsed to match the consensus.

1.2.4.2 Paired 10xGenomics BCR data processing

Pipelines of processing paired 10xGenomics BCR data are still in their nascent state [103, 121]. The majority of published studies used the CellRanger software, developed by 10xGenomics, as a short-read assembler. CellRanger also performs contig annotation with V(D)J gene origin information. Contigs with incomplete annotations are automatically discarded. CellRanger outputs contigs in the FASTA format which can then be annotated with IgBlastn to comply with the MiAIRR standards [117, 118].

1.2.4.3 Clonotyping

Assignment of BCR sequences into clonal lineage families (also known as clonotyping) is a fundamental technique to identify functionally related BCRs. Clonotyping is a simple, fast and powerful approach to study repertoire dynamics and diversity. Common clonotyping protocols work on the premise of clustering together antibodies deriving from the same V and J genes, having the same CDR-H3 loop length, and having over a minimum threshold CDR-H3 loop sequence identity. This minimum threshold is not yet standardised but typically ranges between 75% to 100% sequence identity [3, 4, 122]. The difficulty of standardising the threshold is caused by the wide range of the BCR research performed within the AIRR community. For instance, higher thresholds are selected to construct more accurate B-cell phylogenetic trees, whilst the lower threshold better groups B-cells with similar binding properties [109, 122]. All BCR sequences found in the same clonotype are assumed to originate from the same common ancestor B-cell and hence, to bear similar antigen reactivity [113]. Clonal assignment opens new avenues for BCR repertoire interrogation. For example,

a reduction in clonal diversity is often observed in elderly individuals and cancer patients, and could signal a past or ongoing repertoire response against a pathogen [123, 124]. Greiff *et al.,* showed that clonal diversity significantly differs between healthy and diseased individuals [88]. Hence, BCR repertoire clonal diversities can potentially be used as a feature in immunodiagnostics.

The theoretical probability of finding similar BCR sequences across different individuals is very low when considering the massive diversity of recombined V(D)J segments and a relative low depth of Ig-seq [125]. However, Briney *et al.,* demonstrated that healthy human B-cell donors share a small number of clonotypes ($\sim 1\%$ between two individuals) which were defined as "public clones". Identification of private and shared clonotypes reveals information on BCR repertoire predetermination across individuals, further advancing our understanding of immunology [94].

1.2.4.4 Other common analyses

B-cells undergo an iterative process of affinity maturation in response to antigenic stimulation. An assignment of BCR sequences into the clonal families does not reveal the directionality of the repertoire evolution. To combine clonotypes with B-cell evolutionary information, scientists have developed multiple tools for B-cell phylogenetic tree generation [86]. These tools utilise patterns of SHMs and class switching within individual clonotypes [86] or even across an entire BCR repertoire [126] to predict the most probable B-cell clonal tree topology. Accurate reconstruction of the phylogenetic trees is of high importance to study vaccine-driven BCR sequence maturation [127].

Transforming BCR repertoires into a network representation has also shown to be a powerful research tool to study BCR repertoire architecture [128, 129]. The major bottleneck of studying BCR repertoire architecture with the network representation analysis is the amount of computational resources (CPU time and RAM) required to calculate the distance matrix between all repertoire sequences [129].

Other common BCR repertoire analysis techniques include identification of novel genes/alleles [99, 130], studying CDR-H3 amino acid composition [131], length distributions [122, 132], and V and J gene usages [122, 133].

1.3 Structural interrogation of BCR repertoires

1.3.1 Antibody modelling

The structural antibody database (SAbDab) [134] iterates through the protein data bank (PDB) to identify recently solved antibody structures each week [135]. As of October 2020, more than 4,210 antibodies were found in the public domain. Both human and mouse antibodies constitute the majority of all structures in SAbDab (44% and 33% respectively). These antibodies do not represent the normal state of the immune system, as they were engineered/selected to bind to antigens of interest. This steadily increasing number of experimentally determined antibody structures has enabled researchers to rapidly and accurately model antibodies by leveraging homology modelling methods [41, 136]. Below we review current antibody modelling approaches and their applications.

1.3.1.1 Antibody modelling workflow

The standard antibody modelling workflow includes four steps (Figure 1.14) [41, 136, 137]. The first step is homology modelling of the VH and VL frameworks that will act as a grafting scaffold for the future model. The framework template can either be selected by sequence identity to the full-length chain [136] or to individual framework regions [41, 138]. The assessment of the modelling accuracy is performed in the *ex post facto* analysis, where distances are calculated between the selected templates and the true structure coordinates [41]. Due to framework structural and sequence invariance, current computational tools can model framework structures very accurately (sub-Ångstrom predictions) [139]. The second step is determining the VH/VL orientation, which can be achieved by copying the orientation angle from structures with high Fv sequence identity using VH/VL orientation methods such as ABAngle [140], analytical estimation of the angle using energy functions [141], tailored protein–protein docking [142] or structure-trained machine learning [143]. Once the VH/VL orientation is set, it constrains the geometry of the binding site, allowing for the third step, which is modelling of non-H3 CDRs. At this stage, either the canonical classes are used [144] or template-based modeling such as FREAD [27, 145]. FREAD maps an input loop sequence onto a known loop structure ("template") based on the compatibility of their anchor region orientation. The selected templates are then ranked by calculating their sequence/structural similarity scores (Environment-specific Substitution Score (ESS)) with the target loop sequence [146]. ESS scores are

devised based on observed amino acid substitution scores in a protein evolutionary environment [147]. Although non-H3 CDRs can change their canonical class when mutated, FREAD was shown to predict non-H3 structures within 1.5 Å in ∼90% of cases [49].

In the final step, the CDR-H3 loop is modelled using homology, *ab initio* or a combination of both techniques [148]. The resultant antibody model is refined for feasibility of dihedral angles from the Ramachandran distribution, side chain orientations and side-chain clashes [141, 149, 150].

1.3.1.2 CDR-H3 modelling

Homology modelling approaches are fast at generating models if a template structure is available. Homology models can be created using online services: PIGSpro [137], Kotai Antibody Builder [151], Repertoire Builder [152] and ABodyBuilder [41]. Homology modelling is highly dependent on the quality of the solved antibody structures and the availability of a similar template structure in current databases, which can be a problem for CDR-H3 where suitable templates for longer loop lengths are often unavailable [148, 153]. This lack of templates is primarily due to the huge diversity of CDR-H3 shapes, increasing structural freedom for longer CDR-H3 loops [52]. An alternative to homology modelling is *ab initio* modelling, which does not rely on knowledge of already solved structures. These methods create a large number of potential conformations which makes them computationally expensive compared with homology methods [154]. *Ab initio* approaches include RosettaAntibody [155] and PLOP [156]. Hybrid loop modelling methodologies leverage the advantages of both modelling paradigms. For instance, Accelrys creates an initial loop model with a knowledge-based approach followed by *ab initio* loop refinement [157]. Sphinx is a novel CDR-H3 modelling tool [138] that was inspired by the length-independent canonical CDR clustering of Nowak *et al.* [48]. Sphinx outperformed all modelling tools on CDR-H3 structure prediction in an *ex post facto* comparison to the antibody modelling assessment [139].

CDR-H3 *ab initio* and hybrid modelling tools propose hundreds of potential shapes ("decoys") for a given loop. These decoys are then scored and ranked based on geometric and statistical potentials [155]. However, the top ranked decoy is not always the closest to the native loop configuration [138]. Hence, development of robust decoy scoring functions is as important as optimisation of CDR-H3 modelling software. Recently, a bespoke CDR-H3 decoy scoring tool (DeepH3) has been released [158]. DeepH3 is a deep learning method which was trained to predict inter-residue

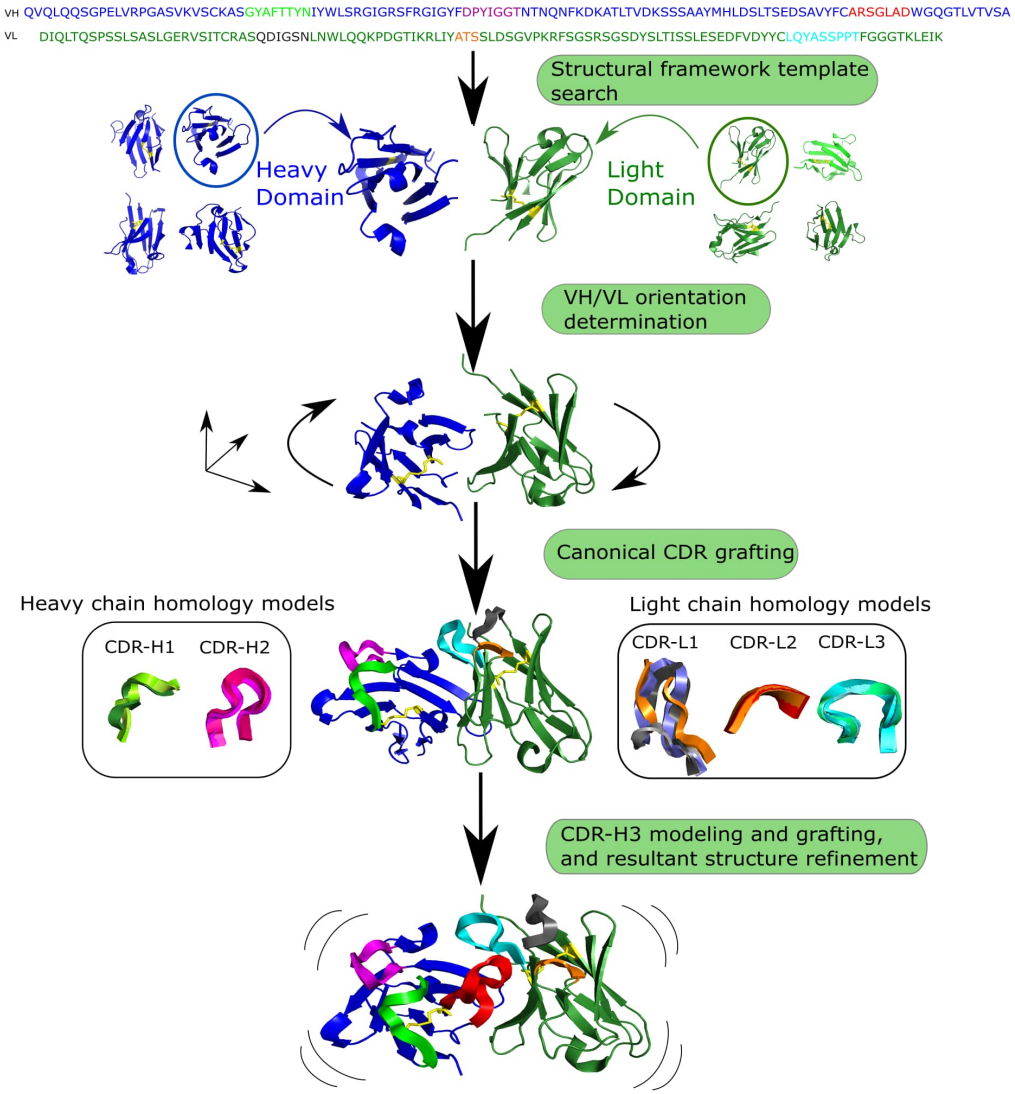


Figure 1.14: **Generalised workflow of antibody modeling.** First, heavy and light chain frameworks are determined by homology modeling using templates from known structures. Next, the VH/VL orientation is calculated. The third step is modeling non-H3 complementarity determining regions (CDRs), followed by modeling and grafting of CDR-H3 onto the pre-assembled scaffold. Finally, sidechains are added to the resultant structure and it is refined.

distances and orientations in the antibody Fv domain. DeepH3 outperformed earlier scoring functions implemented in RosettaAntibody by selecting a set of decoys that had lower root-mean-square-deviation (RMSD) values to corresponding native CDR-H3 structures. The RMSD calculation of the protein backbone atom positions provides a pairwise measurement of the three-dimensional dissimilarity between two sets of coordinates where solved or predicted structures are available. The RMSD is usually calculated on the backbone atoms, but this calculation can be performed including the side chain atoms as well [152]. Sub-Ångstrom RMSD indicates structurally identical shapes, while an RMSD value greater than 2 Å for a short segment indicates structurally a distinct conformation.

Despite the development of different approaches, no single tool currently exists that is able to reliably model native CDR-H3 structure. Accurate predictions of the CDR-H3 specifically and other CDRs in general are crucial to structurally characterise the antibody paratope and to more accurately predict the antibody–antigen binding complex [159].

1.3.1.3 Antibody modelling assessment

The performance of antibody modelling tools has been assessed in two blind studies, AMA-I and AMA-II [139, 160, 161], where several computational tools were benchmarked against a small number of crystallographically solved but unpublished antibody crystal structures. Models of frameworks and canonical CDRs are usually accurate within 1–1.5 Å RMSD, respectively, which is very close to native structure. However, CDR-H3 prediction remains the biggest hurdle for computational antibody modelling as average accuracies for this step ranged between 2.5 and 3 Å RMSD, indicating a decidedly different structure to the native fold. Predictions of this quality are usually not suitable for rational design applications [162].

AMA-II suggested that antibody modelling tools on average produce models of approximately similar accuracies with higher RMSD for longer loop lengths. However, the time required is radically different between homology and *ab initio* approaches [139, 161]. Homology modelling can produce a model on average in under a minute [ABodyBuilder [41], PIGSpro [137]], whereas *ab initio* approaches may require up to tens of CPU hours per model [RosettaAntibody takes 482 CPU hours on average per model [163]]. To be able to use a fast homology method a suitable template is needed. Such templates are becoming more frequently available as the number of solved antibody structures increases, with modelling performance in line with the *ab initio* methods [41, 134]. In order to model millions of sequences in a typical BCR

dataset, speed is crucial. Modelling at such high throughput can currently be achieved by tools such as ABodyBuilder, which is able to generate a model within 30 seconds [41]. However, further increasing the rate and accuracy of antibody modelling, and developing new ways of speeding up CDR-H3 prediction, are needed if we are to structurally characterise complete BCR datasets.

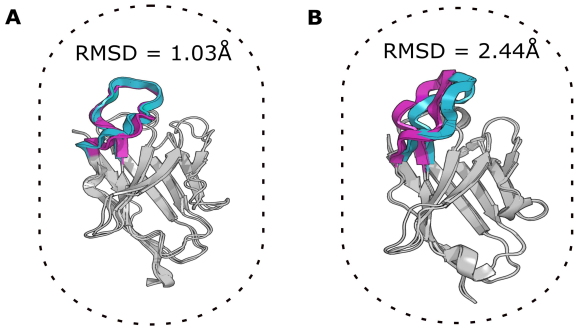
1.3.2 Repertoire modelling

The accuracy and speed of some computational tools mean that thousands of sequences from BCR datasets can be modelled. Such structurally annotated BCR datasets allow more relevant comparisons of binding sites and thus a more accurate functional grouping of molecules (Figure 1.15). The improved capacity to compare and group antibodies allows us to better visualise the antibody structure space and to investigate structural convergences of paratopes, which can be important for vaccine development [22, 102, 164, 165]. In addition, modelled BCR repertoires can be used as input for several computational tools, which annotate structure-derived antibody properties, such as the therapeutic viability of the molecule [166, 167].

Common approaches to delineate BCR repertoires usually employ nucleotide data analysis tools, remaining firmly within the remit of information that can be derived from primary sequences [86, 168]. These rapid methods of calculating BCR repertoire descriptors are highly scalable, an important property as BCR datasets become ever larger and more numerous [3, 4, 43]. However, the decision to avoid paratope structural descriptors leads to the loss of valuable information [48, 101, 169], as it is known that similar sequences can have markedly different epitope complementarity and *vice versa* [170]. Therefore, a computationally efficient structure-based BCR repertoire method should augment current Ig-seq analysis pipelines to deliver a clearer understanding of the processes that govern BCR repertoire development.

1.3.2.1 RossettaAntibody repertoire modelling

One of the first structural analyses of BCR repertoires was that of DeKosky *et al.,* [101]. Using high-throughput RosettaAntibody modelling, more than 2,000 models of naïve and antigen-experienced BCR repertoire sequences from three individuals were analysed. These models helped to obtain a set of structural descriptors such as net charge, surface hydrophobicity of solvent accessible surface area for computationally determined paratopes. However, the choice of methodologies for this study imposed several limitations. Paired VH/VL data did not contain information about the full-length Fv region. Hence, all paired reads had to be completed using respective parent



Pair	PDB	V gene	J gene	CDRH3	Sequence identity	RMSD
A	4NZU	IGHV3-30*11	IGHJ4*01	ARAPDCADADCHKGAFGY	27.7%	1.03Å
	4S1S	IGHV1-2*04	IGHJ1*01	VRTADCERDPCKGWVFPH		
B	3U7W	IGHV1-2*02	IGHJ1*01	TRGKYCTARDYYNWDFEH	88%	2.44Å
	4JDV	IGHV1-2*02	IGHJ1*01	ARGKYCTARDYYNWDFQH		

Figure 1.15: **Two aligned pairs of VH chains extracted from SAbDab [134].** The CDR-H3 sequences in pair **(A)** belong to different CDR-H3 clonotypes but adopt very similar structural configurations with a RMSD value of ~ 1 Å. Pair **(B)** includes germline precursor (4JDV) and matured (3U7W) anti-gp120 HIV antibodies [171, 172]. Although CDR-H3 sequences of pair **(B)** are members of the same clonotype, the RMSD shows that their CDR-H3 shapes are structurally distinct (RMSD > 2 Å).

V gene germline sequences. Moreover, the slow speed of RosettaAntibody modelling only permitted the prediction of the structures for 1% (or 2,000 sequences) of the total BCR repertoire in 570k CPU hours. Finally, the paired reads with CDR-H3 sequences longer than 16 amino acids were not included in the structural analysis as the modelling accuracy of such loops is currently low due to the limited availability of longer loop templates. This emphasizes the challenges of modelling longer CDR-H3 configurations [52, 148]. Hence, novel fast and reliable CDR-H3 structure prediction as well as technologically optimised paired VH/VL gene Ig-seq methods are urgently needed for improved BCR repertoire modelling and interrogation.

1.3.2.2 SCALOP repertoire modelling

RosettaAntibody [163] is a well-established antibody modeling tool and is able to structurally model sequence data; however, its run times make it difficult to structurally characterise the millions of sequences that are gathered during a typical Ig-seq experiment. For this reason, streamlined approaches are being developed to tackle the structural annotation of BCR repertoires. For instance, Wong *et al.,* [49] developed SCALOP to rapidly annotate millions of non-CDR-H3 loops (e.g. canonical loops) with structural information. SCALOP operates on the pre-built position specific substitution matrix (PSSM), which is calculated on the frequencies of amino acids

found at equivalent positions within structural clusters. These structural clusters are devised by clustering respective non-CDR-H3 loops based on pre-defined backbone RMSD value thresholds [48, 49]. These non-CDR-H3 loops were downloaded from SAbDab, a database that curates all publicly available antibody structures [134]. SCALOP was run on 5 million heavy and 8 million light chain sequences. The expected BCR repertoire coverage and canonical class prediction accuracy for SCALOP are 93% and 89% respectively [49].

1.3.2.3 ABodyBuilder repertoire modelling

Structural characterisation of large sequence sets can be extended to the entire Fv region. ABodyBuilder was used to predict structures of 6,000 paired antibody sequences from public repositories [41]. The average modelling time per 1,000 antibody sequences was 567 CPU hours compared with 285,000 CPU hours using RossettaAntibody [101]. ABodyBuilder produces model accuracies that are in line with other tools tested in the AMA-II study [161]. Using tools such as ABodyBuilder, one can perform large-scale structural modelling of BCR repertoires. Such structural characterisation of BCR repertoire similarity/difference would allow more accurate inter-molecule comparisons and assessment of developability [167].

1.3.2.4 Future directions in repertoire modelling

Most publicly available BCR repertoires are unpaired, only covering either the heavy or light variable domain [43]. This precludes the generation of models with natively paired VH/VL domains. In order to harvest the structural information concealed in publicly available BCR repertoires, new computational tools are requires to either artificially pair VH and VL chains and/or to extract information from the single chain sequences.

Krawczyk *et al.,* developed a method (SAAB) for complete structural annotation of unpaired BCR repertoires [169]. There, the researchers used a bespoke version of FREAD loop homology modelling [27] to predict antibody CDR shapes. Despite the limited availability of crystallographically solved antibody data, FREAD modelling found structural matches for the $\sim$ 77% of CDR loops in complete BCR repertoires [169].

Raybould *et al.,* [102] used a combinatorial approach to computationally pair separate VH and VL repertoires which originated from the same sequencing sample. Only VH/VL pairs with high identity to crystallographically solved antibody

interfaces were retained. By employing a bespoke version of ABodyBuilder [41], Raybould *et al.,* demonstrated a strong structurally convergence response across BCR repertoires in response to vaccination [102].

1.4 Thesis outline

Chapter 2

In chapter 2, we present our original work on the development of a novel tool - the AntiBody Sequence Selector (ABOSS) - to identify and flag potential sequencing errors in BCR repertoire data. ABOSS leverages the presence/absence of the conserved cysteines in a BCR sequence to create a structure-based estimate of the sequencing error rate. Since it uses information rather than sequence homology clustering, ABOSS is orthogonal to all other computational methods for error estimation.

Chapter 3

In chapter 3, we present our original work on the development of the Observed Antibody Space (OAS) resource. It is the first curated database that curates publicly available BCR repertoires. As of October 2020, we have collected Ig-seq outputs from 80 unpaired and 5 paired BCR studies, covering 1.9 billion sequences across diverse immune states, organisms (primarily human and mouse), and individuals. We have sorted, cleaned, annotated, translated, and numbered these sequences and make the data available at `http://opig.stats.ox.ac.uk/webapps/oas/`.

Chapter 4

In chapter 4, we describe our original work on studying structural diversity in BCR repertoires along the B-cell differentiation axis in humans and mice. We introduce a novel rapid structural annotation software - SAAB+ - to predict CDR loop shapes in complete BCR repertoires. The distribution of these shapes is then leveraged to calculate repertoire structural profiles. The SAAB+ pipeline reveals unprecedented insights into both the structural predetermination and dynamics of the adaptive immune response further advancing our understanding of immunology.

Chapter 5

In this chapter, we describe the development and application of a wide range of tools and analyses to help the scientific community in the global COVID-19 research

effort. We process all publicly available SARS-CoV-2 binding antibodies including the binders isolated from COVID-19 patients in a consistent format into the CoV-AbDab database. Profiling CoV-AbDab entries against the OAS resource showed that some individuals might already posses SARS-CoV-2 binding antibodies, potentially elicited by seasonal coronavirus infections. We also find strong convergent sequence and structural responses between BCR repertoires from SARS-CoV-2 infected individuals.

Future work

In the last chapter of the thesis, we briefly discuss future directions of our research. Some of these analyses were originally planned to be carried in the last year of this DPhil work, but were postponed due to the ongoing COVID-19 pandemic.

2

Filtering Ig-seq data using antibody structural information

Contents

This chapter is based on the material from the following paper:

1. **Kovaltsuk, A.**, Krawczyk, K., Kelm, S., Snowden, J. & Deane, C.M. (2018) Filtering Next-Generation Sequencing of the Ig Gene Repertoire Data Using Antibody Structural Information Journal of Immunology, 201(12):3694-3704

I carried out all the work described in this chapter.

2.1 Introduction

Next-generation sequencing of the immunoglobulin gene repertoire (Ig-seq) produces large volumes of information at the nucleotide sequence level, allowing interrogation of snapshots of the B-cell receptor (BCR) diversity. Such data have improved our understanding of immune systems across numerous species and have already been integrated into vaccine development and drug discovery [e.g., [6, 9, 87, 173]]. However, the high-throughput nature of Ig-seq means that it is afflicted by high error rates, which makes it difficult to distinguish between Ig-seq artifacts and true nucleotide alterations introduced by V(D)J recombination processes and the somatic hypermutation (SHM) machinery of B-cells. Raw Ig-seq data could also contain a distorted diversity of BCR repertoire sequences caused by recursive nucleotide error introduction across tens of sequencing cycles as well as biased PCR amplification of sequencing library amplicons in the sample preparation step.

Several experimental Ig-seq error-correction approaches have been proposed; however, an agreed standard does not yet exist [96]. This is the result of the discordance in experimental protocols employed in the sequencing library preparation step [3, 4, 174]. Existing experimental approaches for error correction include taking invariant sequence portions as a proxy for estimating error or DNA barcoding sequences that should be identical if they originate from the same amplicon. For instance, Galson *et al.,* [175] performed sequencing of the constant portions of the BCR heavy chain. As this region should be sequence invariant (lack of SHMs), it offered an estimated error rate on the variable heavy (VH) domains sequenced in the course of the same study. This error estimate was then used as a bespoke cut-off to group these BCR repertoire sequences into separate clonal families. Khan *et al.,* [120] barcoded individual BCR cDNA transcripts with unique molecular identifiers (UMI) prior to PCR. The resultant pool of genetic data was sequenced, and identically UMI barcoded sequences were put into separate clusters in which a consensus sequence was devised based on the positional nucleotide frequencies. All other members of the cluster were collapsed

with respect to this consensus sequence. Error can be introduced even in this method in the early steps of sequencing sample preparation, such as reverse transcription and PCR [176, 177]. Devising a correct sequence within the clusters is heavily dependent on sequence redundancies, which precludes correction of singleton clusters using the UMI barcoding approach. To attain sufficient amplicon coverage, scientists could input fewer B-cells into the experiment and/or increase the number of sequencing cycles to achieve a higher sequencing depth [107].

Techniques such as UMI barcoding or sequencing constant portions are time consuming and require specialised experimental setups as well as bespoke bioinformatics analysis pipelines. To address such issues, several computational error-correction tools have been developed [96]. These applications all operate by building consensus sequences using primary sequence homology clustering. The majority of these tools work only in the remit of CDR-3 of the VH domain (CDR-H3) [178, 179], largely ignoring the rest of the sequence. MiXCR is the most commonly used Ig-seq error-correction tool to date [180]. It supports the analysis of entire VH or variable light (VL) chains and performs sequencing error correction. MiXCR works by aligning sequences from an BCR dataset to reference V, J, and C genes followed by identifying gene feature sequences. This is a k-mer of residues identical across multiple sequences and is found in CDR-H3 by default. These gene feature sequences are then used to sort BCR sequences into sets of separate clonotypes. The number of unique clonotypes is always overestimated because of PCR and sequencing errors. To overcome this, "correct" sequences are found by performing heuristic multilayer clustering on these clonotypes, in which the most redundant clonotypes are treated as correct. This could lead to false identification of representative clusters due to large PCR or related plasma B-cell expansions. A more recently developed Ig-seq construction tool, IgReC [181], takes a different approach. It uses Hamming graphs to identify correct sequences. Benchmark analysis on barcoded BCR data shows that the IgReC pipeline is as accurate as experimental error-correction approaches [181]. This suggests that advances in algorithm development can potentially alleviate the need for experimental Ig-seq correction. All currently available computational methods consider sequence information alone. The omission of structural information inputs into a BCR repertoire analysis can lead to incorrect structural repertoire diversity estimations [153] as well as lower the success rate of experimental antibody production from the sequencing data. In this paper, we consider how knowledge of an antibody structure may help identify sequencing errors by finding sequences that are not structurally viable,

as structural viability is crucial for the correct functioning of an antibody. We then use this structural information to estimate sequencing error rates.

A typical antibody structure requires the presence of a disulphide bridge within each of the variable chains. This covalent bond helps maintain the immunoglobulin fold. Cysteines at positions 23 and 104 [ImMunoGeneTics (IMGT) numbering [36]] must be present in structurally viable natural antibody/BCR sequences [29, 35, 182, 183]. There is evidence that some antibodies can still fold when the disulphide bond is ablated [30–32]. ABPC48 is the only example of an antibody that naturally lacks cysteine at position 104 [182]. APBC48 is a mouse antibody derived from plasmacytoma [184]. Although ABPC48 antibody is able to fold, restoration of cysteine at position 104 significantly improves its stability [32].

In this work, we describe a novel computational tool, Antibody Sequence Selector (ABOSS) that uses the presence/absence of the conserved cysteines in a BCR sequence to create a structure-based estimate of the sequencing error rate in a given BCR repertoire. As opposed to other error-correction tools that operate at the nucleotide sequence level, ABOSS uses amino acid sequences, as they relate directly to protein structure. ABOSS both filters out amino acid sequences that are not structurally viable as well as flags those likely to contain erroneous residue/positions. Because of its use of structural information rather than homology clustering, ABOSS is orthogonal to all other computational methods for error estimation.

Examining ABOSS performance on simulated BCR datasets indicated that ABOSS successfully isolates ∼99% of structurally viable sequences whilst preserving most of the SHM-generated diversity. We tested ABOSS on six separate BCR datasets and found that our error calculations based on structural viability were in line with error estimates declared in other recently published studies.

2.2 Materials and Methods

2.2.1 Data

We have tested ABOSS on six BCR repertoires datasets (Table 2.1). Two datasets are from Khan *et al.,* [120], the raw sequences (Khan_R) and the UMI error-corrected sequences (Khan_C); each of these datasets was composed of three immunised datasets of a single mouse that were pooled together (2.4 million sequences). The Galson *et al.,* [175] dataset (HEPB) consists of sequences from hepatitis B studies [185, 186] at a time point before the 11 participants were vaccinated (9.9 million sequences). The third and fourth datasets are proprietary UCB Pharma datasets of 5.6 million

VH (UCB_H) and 9.3 million VL (UCB_L) chain sequences [169]. The UCB data was generated from a pool of 494 non-antigen challenged B-cell donors. The Vander Heiden *et al.,* [109] datasets (Healthy_H and Healthy_L) include sequences from four healthy B-cell donors. A mixture of VH and VL gene primers were used in sequencing material preparation, which produced pooled VH/VL BCR datasets. Healthy_H and Healthy_L are the sorted VH and VL sequences, respectively. These BCR datasets were employed to test ABOSS across heterogeneous sequencing setups.

2.2.2 ANARCI Parsing

ABOSS supports several input formats. These can be amino acid sequences in the FASTA file or raw IgBlastn outputs [118].

The first step of ABOSS is to parse every sequence through ANARCI [34], an antibody/TCR-numbering program. ANARCI parsing acts as a pre-filtering step, removing sequences that 1) contain unusual insertions/deletions in the framework and canonical CDR regions with the reference to the IMGT scheme, 2) do not align to the respective species Hidden Markov Models (HMMs) of the IMGT germline, and 3) have a J gene sequence identity of 50% to the IMGT germline (of the respective species) or truncated framework 4 region. Calculation of the J gene sequence identity allows us to remove sequences in which indels have occurred in CDR-3 and framework 4. At this point, ABOSS also removes sequences in human and mouse BCR datasets that have CDR-3s longer than 37 amino acids. This cutoff is in place to remove sequences with abnormally long and potentially erroneous CDR-H3s [132]. These are chimeric sequences that arise as a result of PCR error. Separate cut-offs should be implemented for bovine antibodies as these pose ultralong CDR-H3 loops [187]. Sequences that pass these initial tests are numbered using the IMGT scheme [36]. This provides a consistent frame of reference for sequences and defines CDR and framework regions. We employ the IMGT numbering scheme in ABOSS because it assigns length-mismatched CDRs located in roughly structurally equivalent space to identical residue numbers [36, 170].

2.2.3 Residue Error Rate Estimation

IMGT numbering enables the calculation of amino acid distributions by position (Equation 2.1).

$$P_{i,j} = \frac{M_{i,j}}{\sum_{i=1}^{n} M_{i,j}} \tag{2.1}$$

In Equation 2.1, the proportion of an amino acid at a position $P_{i,j}$ is calculated by dividing the occurrence of each amino acid type at the given position $M_{i,j}$ by the total number all amino acid occurrences at respective positions.

IMGT positions 23 and 104 are used to estimate the error rate in the data. In all naturally occurring antibodies, both these positions are always a cysteine residue [29, 182]. Some antibody pseudo V genes encode for a non-cysteine residue at position 23 and 104, but these antibodies are not structurally viable [35]. Therefore, under a conservative model, any amino acid divergence from cysteine at these positions can be treated as error. We define the error to be equal to the largest non-cysteine amino acid proportion (or normalised occurrence, see Equation 2.1) found at either of the two positions.

Using the sum of all non-cysteine amino acid proportion overestimates the residue error rate (Figure 2.1). The dataset used in Figure 2.1 would be ascribed an error rate of 0.0027 (the occurrence of C104G). Thus, all amino acids at a position that occurs with proportion of <0.0027 across the dataset are considered erroneous. This will flag all the non-cysteine residue types at position 23 and 104 in the data as they all occur <0.0027, but will also indicate several other positions in which residues may be erroneous. In this fashion, we provide an error estimate for individual residues, which can be extrapolated to the entire sequence. Hence, the mutations that are observed repeatedly (like SHMs) will not be flagged.

2.2.4 Structure based filtering of BCR repertoire data

The next stage is ABOSS filtering. In this step, if the proportion of an amino acid at a position is below the residue error rate, amino acids of that type at that position are flagged as potentially erroneous. The type of the amino acid at these positions is not fixed and varies across BCR repertoires.

ABOSS creates a reference matrix, which contains the "allowed" amino acids at each IMGT position. The allowed amino acids are those whose proportion in the BCR dataset are greater than the residue error rate at the respective position (see previous section). The reference matrix also contains the amino acids from IMGT germline sequences as they represent structurally viable antibodies. We did not include any antibody sequences from PDBs, as these are often subjected to a high degree of protein engineering. If <20 entries (as the total count of unique amino acid types) are used to calculate amino acid proportions at a position, this position is not included in the reference matrix.

Once the reference matrix is calculated, every sequence from the BCR dataset is compared with it. For a given position in a sequence from the BCR dataset, a flag is placed if the amino acid in the sequence is not present in the reference matrix. ABOSS outputs a comma-separated values file of the ANARCI-parsed sequences, their redundancies, CDR-H3 regions, flagged residue/positions, V and J genes, and query names of the original raw nucleotide sequences from the IgBlastn output. The ABOSS filtered dataset refers to the set of sequences with zero flagged residue/positions.

2.2.5 *In silico* PCR simulation

We created *in silico* PCR error simulated BCR repertoires which were based on two separate ABOSS-filtered datasets, UCB_H and Khan_R. The simulations were performed at the nucleotide level. The nucleotide sequences that corresponded to the amino acid sequences that passed ABOSS with zero flagged residue/positions (see ABOSS-filtered dataset in Table 2.2) acted as starting points for our simulation. During the simulation, each sequence in the starting dataset was subjected to randomised nucleotide mutations. The distribution of the number of nucleotide mutations was proportional to the distribution of flagged residue/positions in the respective redundant BCR data determined by ABOSS analysis (see ABOSS analysis on raw BCR data), whereas mutation positions were stochastically selected (PCR enzyme biases were not considered) along the VH domain. Only sequences in which random mutations were introduced were added to the final simulation dataset. As both UCB_H and Khan_R datasets were generated using Illumina sequencing technology, only nucleotide substitutions were considered in the error simulations.

To assess the robustness of ABOSS, we varied both residue error rate and dataset size in our error simulations. To increase the residue error rate, every entry from the originally calculated distribution of flagged residue/positions was amplified by an error multiplier. Separate simulations were carried out for individual values of the error rate multiplier that ranged between one and eight. The simulation final dataset sizes were equivalent to the size of the respective ANARCI-parsed BCR dataset (see ANARCI-parsed dataset in Table 2.2). Separate simulations were also performed in which the size of the simulation final dataset was varied to be between one and eight times smaller than the respective ANARCI-parsed BCR dataset.

2.2.6 BCR SHM lineage tree simulation

We carried out two separate SHM simulations on the nucleotide sequences of the Healthy_H and UCB_H datasets. Nucleotide sequences, whose translated amino acids had zero ABOSS-flagged residue/positions in Healthy_H, were assigned as the most-recent common ancestors (MRCA) in the simulations. Two different clonal lineage trees (Lineage_A and Lineage_B) were employed for the number of progenitor sequences and SHM substitutions (see Appendix B.1 for more lineage tree details). We used the human HH_S5F–targeting model [62] from the SHazaM package (`https://shazam.readthedocs.io/`) to perform SHM substitutions in the lineage trees. All MRCA and progeny sequences were added to the final SHM simulation datasets.

2.3 Results

2.3.1 ABOSS Algorithm

ABOSS is a computational method that leverages structural antibody information to calculate the sequencing error rate and flag potentially erroneous residue/positions in BCR datasets. Specifically, we exploit the knowledge of the conserved cysteines at IMGT positions 23 and 104, which shape and stabilise the conformation of the antibody variable chains. The presence of these conserved cysteines can be used as a way of both identifying structurally viable sequences and estimating the sequencing error rate.

ABPC48 is the only characterized natural antibody that lacks either cysteine at either position [30]. A small number of structurally stable antibodies with pairwise substitutions of the conserved cysteines based on the ABPC48 antibody scaffold have been engineered [31, 32, 188]. These pairwise substitutions require further stabilising mutations to the antibody structure, often to the opposite variable chain [31, 32, 188]. The known structurally viable non-cysteine pairs seen at positions 23 and 104 are summarised in Hagihara *et al.,* [182]. In our BCR datasets, we rarely observe the pairwise substitution of cysteines. For instance, the total number of instances when the substitution of both cysteines was observed in the UCB_H data were 811, which corresponded to ∼0.015% of UCB_H. Of these 811 pairwise substitutions, the potentially viable substitutions as described in Hagihara *et al.,* [182] were C23S-C104S, C23S-C104A, C23A-C104S, and C23Y-C104V, which appeared 24, 2, 1, and 1 times,

respectively. The six amino acids that constitute the largest proportions of non-cysteine residue types at positions 23 and 104 in our six raw BCR datasets are always the amino acids 1 nucleotide edit distance from the cysteine codons (Figure 2.1). The top non-cysteine residue type at positions 23 and 104 varies between our BCR datasets. This demonstrates the stochastic nature of this amino acid substitution as the datasets were sources from different research groups. It was previously demonstrated that SHM substitutions are significantly reduced at positions 23 and 104 in gene-specific amino acid substitution profiles of SHM [63]. This must be attributed to negative structural selection, as SHM substitution still takes place at these positions in passenger alleles and using the HH_S5F computational model [63]. This evidence suggests that the substitutions in conserved cysteines seen in BCR datasets are highly likely to be attributed to sequencing and/or errors.

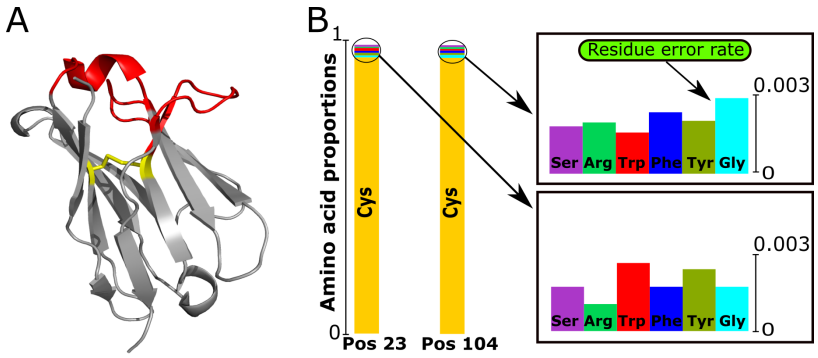


Figure 2.1: **Calculation of the residue error rate in terms of structural viability. (A)** Three-dimensional structure of the VH chain (PDB code: 5WUV) with the conserved disulphide bridge shown. Framework (grey) and CDRs regions (red), the cysteine bond between positions 23 and 104 in yellow. **(B)** The distribution of amino acid types found at positions 23 and 104 for a BCR dataset. Since both positions in natural antibodies should be cysteines, the non-cysteine occurrence indicates possible sequencing error.

In the first step of the ABOSS protocol, all sequences are parsed using ANARCI [34], which numbers the sequences in accordance with the IMGT scheme [36]. Antibody sequences with low sequence identities to ANARCI pre-built Hidden Markov Model (HMM) species profiles; unusual insertions/deletions along the antibody chain are discarded. Next, ABOSS calculates the residue error rate using the ANARCI parsed sequences. The residue error rate is taken as the largest non-cysteine amino acid proportion found at IMGT position 23 or 104 (Figure 2.1). This is a conservative approach as other positions such as W41 and W118 also display a high level

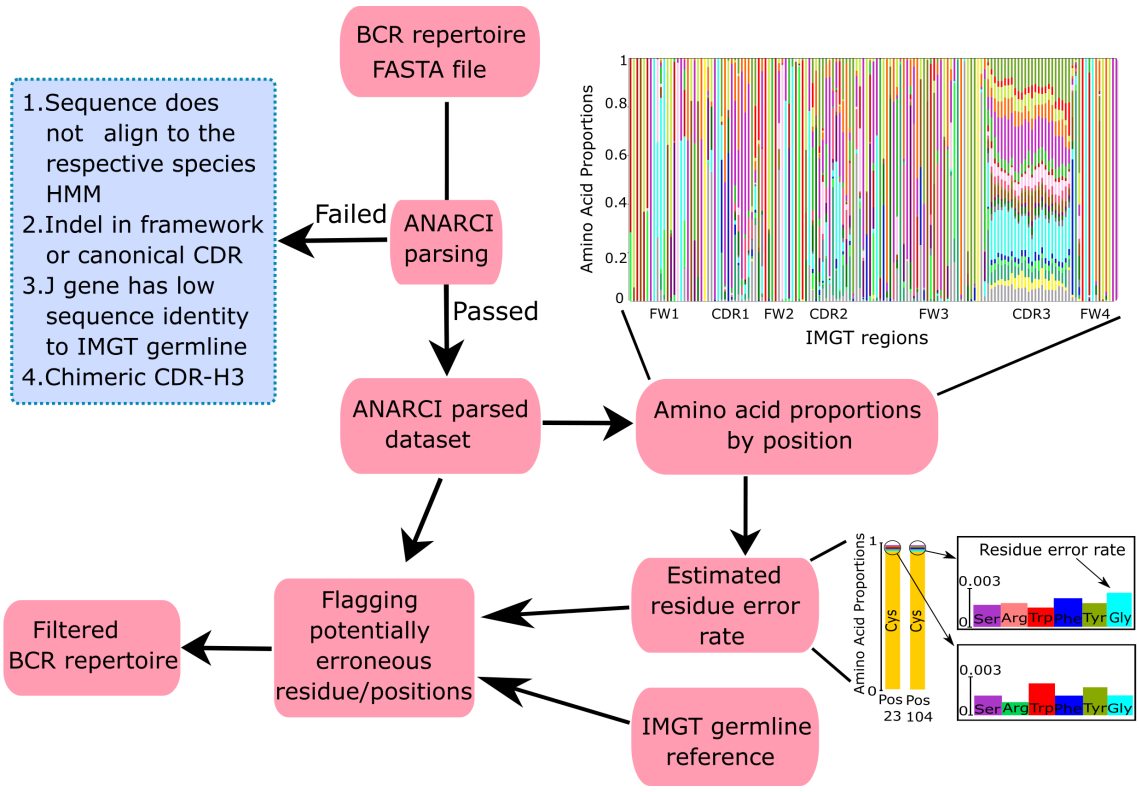


Figure 2.2: **Workflow of ABOSS**. ABOSS input is antibody amino acid sequences in the FASTA format. Every sequence from the input file is IMGT-numbered with ANARCI (ANARCI parsing). The amino acid distribution by IMGT position is calculated for successfully ANARCI parsed sequences. The residue error rate is estimated based on the amino acid distributions at positions 23 and 104 (see Figure 2.1 for more details). Synonymous mutations were not considered in the residue error rate calculation. The estimated residue error rate together with the ANARCI numbered IMGT germline genes are used to flag potentially erroneous residue/positions in individual antibody sequences. Filtered BCR dataset refers to a collection of sequences that pass ABOSS analysis with zero flagged residues/positions.

of structural preservation [36]. The residue error rate is then used to flag specific residue/positions in individual sequences.

The workflow of the algorithm is summarised in Figure 2.2. ABOSS analysis takes <10 h wall-clock time for 5 million unique antibody amino acid sequences on a standard eight-core desktop computer (intel i7-6700). ABOSS is parallelised, allowing for shorter runtimes on more powerful machines. ABOSS is available via `http://opig.stats.ox.ac.uk//resources`.

2.3.2 ABOSS analysis on raw BCR data

We ran ABOSS on six BCR datasets (Tables 2.1, 2.2). We consider two sequences Sequence redundancy if they have identical length and identical amino acid composition. ANARCI parsing removed between 3–23% of sequences in the BCR datasets (Table 2.2). The ANARCI parsing step removed the largest proportion of sequences from Healthy_L, followed by the Healthy_H, UCB_L, Khan_R, UCB_H, and HEPB datasets, respectively. In the second step, ABOSS filtering and residue/position in the sequences are flagged as potential errors. The Khan_R was the dataset with the smallest proportion of sequences with zero ABOSS-flagged residue/positions (26.6%) (Table 2.2). The HEPB dataset had the highest proportion of zero ABOSS-flagged residue/positions (65.9%), followed by Health_L, UCB_L (37.3%), Healthy_H, and UCB_H (33.7%).

Dataset name	Study description	Total dataset size	Antibody chain	Dataset average redundancy	Participants
Khan_R	Raw sequences of immunised mouse 1 from Khan *et al.,* [120]	2.4m	Heavy	3.74	1 (mouse)
Khan_C	Raw sequences of immunised mouse 1 from Khan *et al.,* [120]	2.4m	Heavy	45.3	1 (mouse)
HEPB	Human hepatitis B vaccination from Galson *et al.,* [175]	9.9m	Heavy	1.93	11
UCB_H	Proprietary UCB BCR data of the VH chain	5.6m	Heavy	1.15	494
UCB_L	Proprietary UCB BCR data of the VL chain	9.3m	Light	1.12	494
Healthy_H	VH chains from healthy human B cell donors from Vander Heiden *et al.,* [109]	1.4m	Heavy	1.9	4
Healthy_L	VH chains from healthy human B cell donors from Vander Heiden *et al.,* [109]	6.3m	Light	2.96	4

Table 2.1: **Summary of the datasets used**. The seven datasets (Khan_R, Khan_C, HEPB, UCB_H, UCB_L, Healthy_H and Healthy_L) were obtained from different sequencing methodologies, organisms and immunisation protocol. Peripheral blood mononuclear cells (PBMC) were used as a source of the genetic information for all datasets, except for Khan_R and Khan_C where murine splenic B-cells were extracted. The Khan_R and Khan_C datasets are the immunised mouse 1 dataset of the Khan *et al.,* [120] study before and after the barcode correction approach. These datasets are from repeated Ig-seq of the same mouse. The majority of sequences in this BCR dataset start at IMGT position 8. The Khan_R and Khan_C datasets consist of antibody amino acid and corresponding nucleotide sequences. The Khan_R dataset has the highest redundancy amongst the interrogated non-corrected datasets. We have removed the roughly 10% synthetic spike-ins in the Khan_R and Khan_C datasets as their sequence design did not yield structurally viable antibodies. The HEPB dataset from Galson *et al.,* [175] is from 11 participants. Standard MiSeQ 250bp Illumina Ig-seq was performed. The reads were gene-aligned and processed using IMGT/HighV-Quest [189]. Due to selection of PCR primers, most of the sequences start at position 17. This dataset contains amino acid sequences only. The dataset's redundancy is almost two times lower than the Khan_R data. The UCB proprietary BCR datasets were obtained from 494 participants. The UCB_H and UCB_L datasets comprise 5.6 million and 9.3 million sequences respectively. The UCB_H and UCB_L datasets contain both antibody amino acid and corresponding nucleotide sequences. The UCB datasets were aligned with IgBlastn [118], V and J genes identified, and pre-filtered for stop codons, they contain full-length variable chain sequences as described in Krawczyk *et al.,* [169]. The UCB_H and UCB_L datasets are the least redundant amongst the datasets. The Healthy_H and Healthy_L datasets come from four healthy human B cell donors from the Vander Heiden *et al.,* [109] study. In this study, sequencing primers for both heavy and light chain genes were used at the same time forming pooled raw nucleotide samples. The raw nucleotide BCR datasets were obtained from the OAS resource [44] followed by translating sequences into amino acids and antibody chain separation using IgBlastn [118].

Current Ig-seq error-correction pipelines assign greater confidence to highly redundant sequences and manipulate the nucleotide sequences of rare sequences [112, 120, 180, 190]. In contrast, ABOSS does not have a direct link between sequence redundancy and correct sequences. To examine the performance overlap of ABOSS and redundancy based Ig-seq error-correction tools, we compared the number of ABOSS-flagged residue/positions with the sequence redundancy for our six datasets (Figure 2.3). In every dataset, sequences that are more redundant (e.g. have copies of the same sequence after data filtering) tend to have fewer ABOSS-flagged residue/positions. This suggests that even though ABOSS is not a redundancy-based technique, its results are still in line with the widely adopted methodology based on sequence re-

Data source	Starting dataset	ANARCI parsed dataset	ABOSS filtered dataset	Sequence percentage with zero flags	ABOSS Residue Error Rate (%)
HEPB	9,985,575 (5,175,036)	9,700,893 (4,932,588)	6,579,118 (3,226,473)	65.9 %	0.22
Khan_R	2,445,354 (653,520)	2,247,761 (521,675)	649,685 (47,593)	26.6%	1.5741
UCB_L	9,371,465 (8,380,540)	8,021,407 (7,120,100)	3,494,319 (2,983,103)	37.3%	0.4674
UCB_H	5,645,304 (4,925,532)	5,277,305 (4,587,918)	1,903,703 (1,561,082)	33.7%	0.5892
Healthy_H	1,422,405 (745,276)	1,135,185 (558,171)	486,437 (176,012)	34.2%	0.5427
Healthy_L	6,317,736 (2,135,745)	4,860,389 (1,372,804)	2,667,263 (386,165)	42.2%	0.4121

Table 2.2: **ABOSS analysis of six BCR datasets.** In the table, dataset sizes are given as the number of redundant sequences, the number of non-redundant sequences are shown in parentheses. Starting datasets are the inputs for ABOSS. ANARCI parsed datasets contain sequences that are successfully IMGT-numbered. ABOSS filtered datasets are the number of sequences that contain zero flagged residues. The percentage of sequences with zero flags are calculated as a percentage of redundant ABOSS passed sequences over the total number of starting redundant sequences. Residue error rates are calculated as described in Figure 2.1.

dundancy. ABOSS does flag residues as erroneous in a number of highly redundant clones, which might be flagged as correct by redundancy-reliant methods. If a sequence was highly redundant, it could, in theory, avoid any of its residues being flagged by ABOSS, as every residue/position in these sequences would be present more times than the residue error rate. The horizontal dashed lines in Figure 2.3 shows the redundancy necessary to achieve this. Only a single sequence from the Healthy_L dataset reached such a level of redundancy (Figure 2.3F).

2.3.3 Ig-seq error simulation to estimate sequence volumes and error rates tolerated by ABOSS

To investigate the types of BCR datasets that ABOSS can successfully analyse, we benchmarked ABOSS with respect to dataset redundancies, sequencing error rates, and input sequence volumes. We tested ABOSS on two datasets with contrasting depth and breadth of coverage: the UCB_H and Khan_R datasets (see Table 2.1).

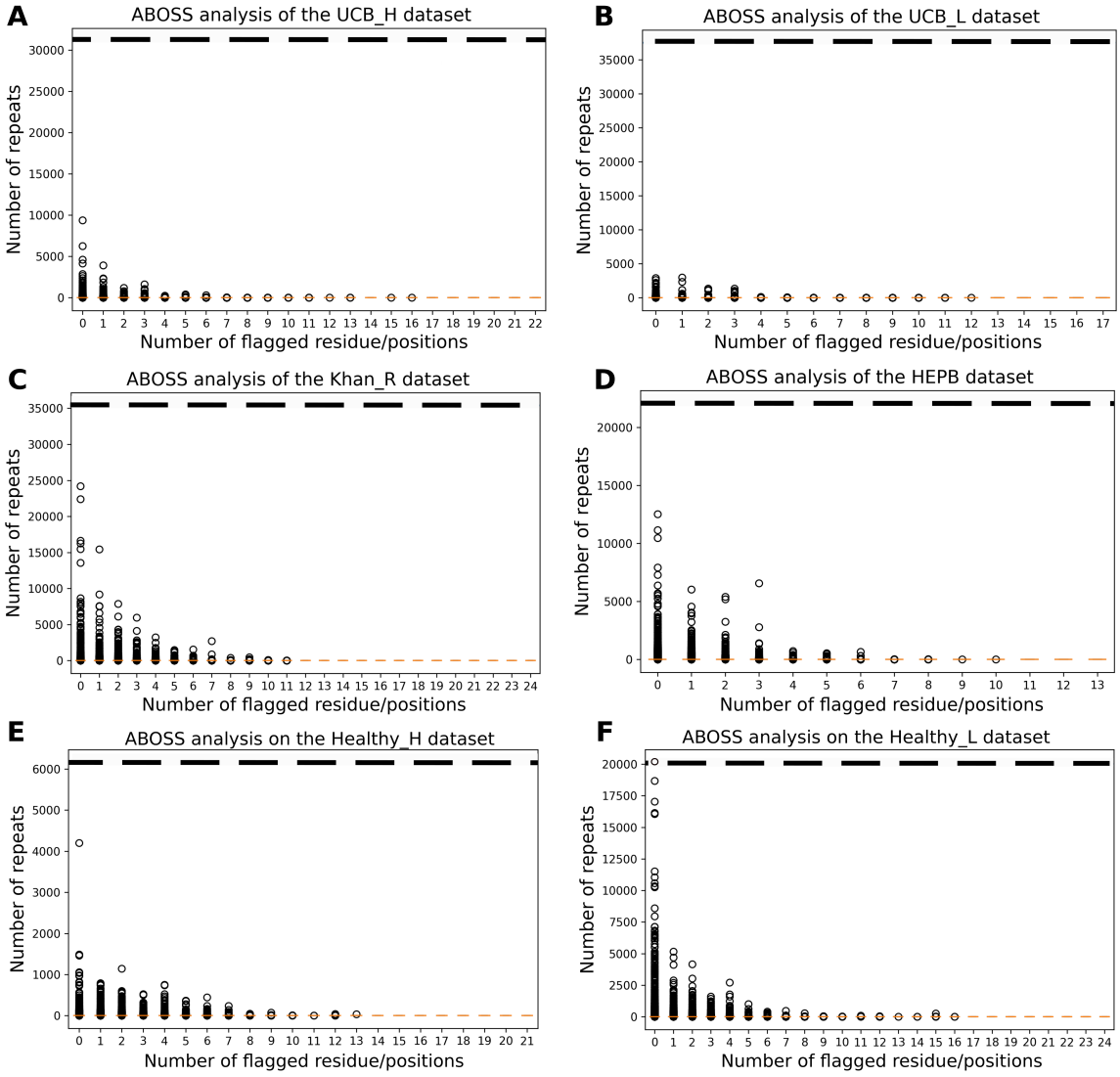


Figure 2.3: **Sequence redundancy relative to the number of ABOSS flagged residue/positions in the sequences of our six datasets (see Table 2.1 for more details): UCB_H (A), UCB_L (B), Khan_R (C), HEPB (D), Healthy_H (E) and Healthy_L (F).** The ABOSS filtering step outputs the number of flagged residue/positions for every sequence in the ANARCI parsed BCR dataset. Zero flagged residue/positions indicates that the sequence is structurally viable. The general trend in each BCR dataset is that the more redundant the sequence the fewer ABOSS flagged residue/positions it has. The horizontal dashed line represents the residue error rate in terms of the number of entries required for a residue/position to be identified as structurally viable.

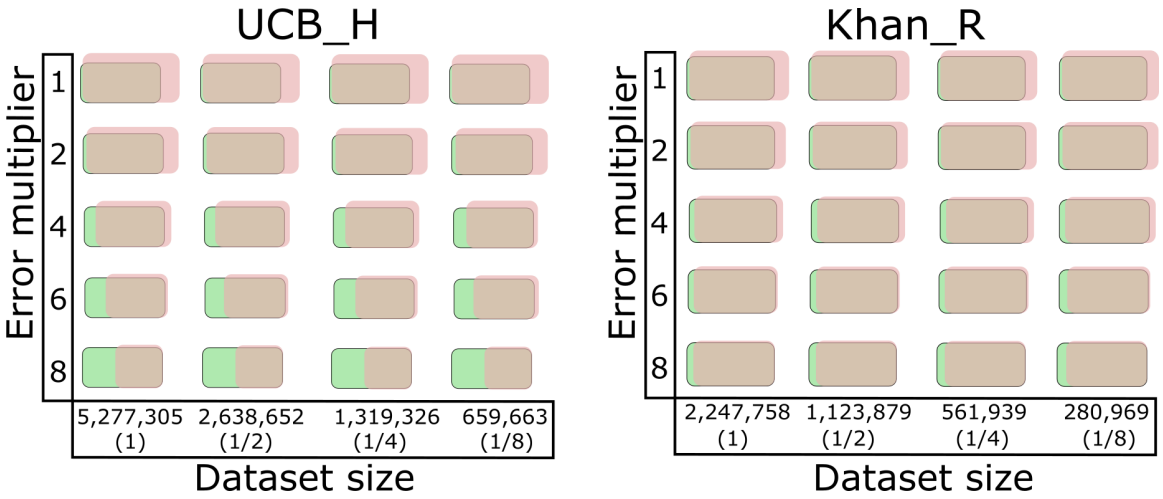


Figure 2.4: **Examination of ABOSS performance on BCR error simulation datasets.** BCR data simulation was carried out based on the previously calculated numbers of ABOSS flagged residue/positions in two BCR datsets, UCB_H and Khan_R (see Figure 2.3). The X-axis corresponds to UCB_H_Sim and Khan_R_Sim dataset sizes used for simulation (the percentages relative to the sizes of respective datasets that passed ANARCI are shown in parentheses). The Y-axis shows the multiplier of the original distribution of erroneous residue/positions in the BCR datasets (see Figure 2.3). The total number of ABOSS filtered sequences (pink) and the number of the correct sequences (green) are pictured (The corresponding percentages are given in Figure 2.5). The overlapping region indicates the proportion of the correct sequences that passed ABOSS relative to the total number of ABOSS filtered sequences.

The starting datasets for the simulation consisted of sequences that passed ABOSS analysis with zero flagged residue/positions. The sizes of the simulation final datasets (UCB_H_Sim and Khan_R_Sim) were based on the number of sequences that passed the ANARCI-parsing step (see Table 2.2). We used the distribution of the number of flagged residue/positions in the UCB_H and Khan_R datasets as calculated by ABOSS (see Figure 2.3) to introduce erroneous residue/positions into our simulation-starting datasets. The mutation substitution positions were stochastically selected along the VH chain. From these starting points, the simulations were performed as described in Materials and Methods (see *In silico* PCR simulation).

The simulation results are shown in Figure 2.4 and the corresponding detailed percentage breakdowns are visualised in Figure 2.5. Using sizes and error rates that match the original data, ABOSS recovered 99.6% and 99% of the correct sequences incorporated into the UCB_H_Sim and Khan_R_Sim datasets, respectively. Reducing the UCB_H_Sim and Khan_R_Sim dataset sizes does not appear to influence the

percentage of correct sequences recovered by ABOSS analysis. Increasing the error rates has a minor effect on the recovered number of correct sequences from the Khan_R_Sim dataset and a much larger effect on the recovery of correct sequences from the UCB_H_Sim dataset. This difference is due to the far lower initial redundancy of the UCB_H data.

ABOSS also retained small numbers of sequences from UCB_H_Sim and Khan_R_Sim that were not present in the simulation-starting datasets (Figure 2.4). These sequences are still structurally viable. The number of these sequences was larger in the UCB_H_Sim dataset (∼30%) than in the Khan_R_Sim dataset (∼17%) (Figure 2.5). As the residue error rate was increased, the simulation-starting datasets constituted larger proportions of the ABOSS-filtered UCB_H_Sim and Khan_R_Sim datasets (Figures 2.4, 2.5).

The outputs from these error simulations suggest that ABOSS performance becomes robust when either the BCR data are redundant or more than ∼600k of nonredundant sequences are available.

2.3.4 ABOSS analysis on SHM-generated diversity

The SHM machinery of B-cells increases BCR sequence/structural diversity by introducing amino acid substitutions in the variable region [191]. SHM helps to fine-tune the geometric and chemical complementary of BCR receptors against their target epitopes [192]. These substitutions are known to exhibit uneven frequencies along the variable region [62, 63]. We exploited the 5-mer nucleotide, HH_S5F-targeting model of SHM [62] to examine the ability of ABOSS to flag errors whilst preserving SHM-generated diversity in BCR datasets. The HH_S5F-targeting model computes probabilities of potential mutations for a given sequence. The HH_S5F-targeting model was developed on the previously observed patterns of SHMs across BCR repertoires in humans [62].

The model requires a clonal tree reference to estimate rates of substitutions. We used two distinct architectures of BCR clonal lineage trees (Lineage_A and Lineage_B) to construct such substitution matrixes. The detailed description of the Lineage_A and Lineage_B models is found in Appendix B.1. We used these two lineages to have coverage of the spectrum of SHM mutations, as Lineage_A has a low substitution rate, whereas Lineage_B has a high one. Using the HH_S5F model combined with either of the Lineage_A or Lineage_B SHM references, the simulations were performed on Healthy_H and UCB_H. These two datasets were selected to test ABOSS performance on low- (UCB_H) and high- (Healthy_H) redundancy data. The sequences with zero

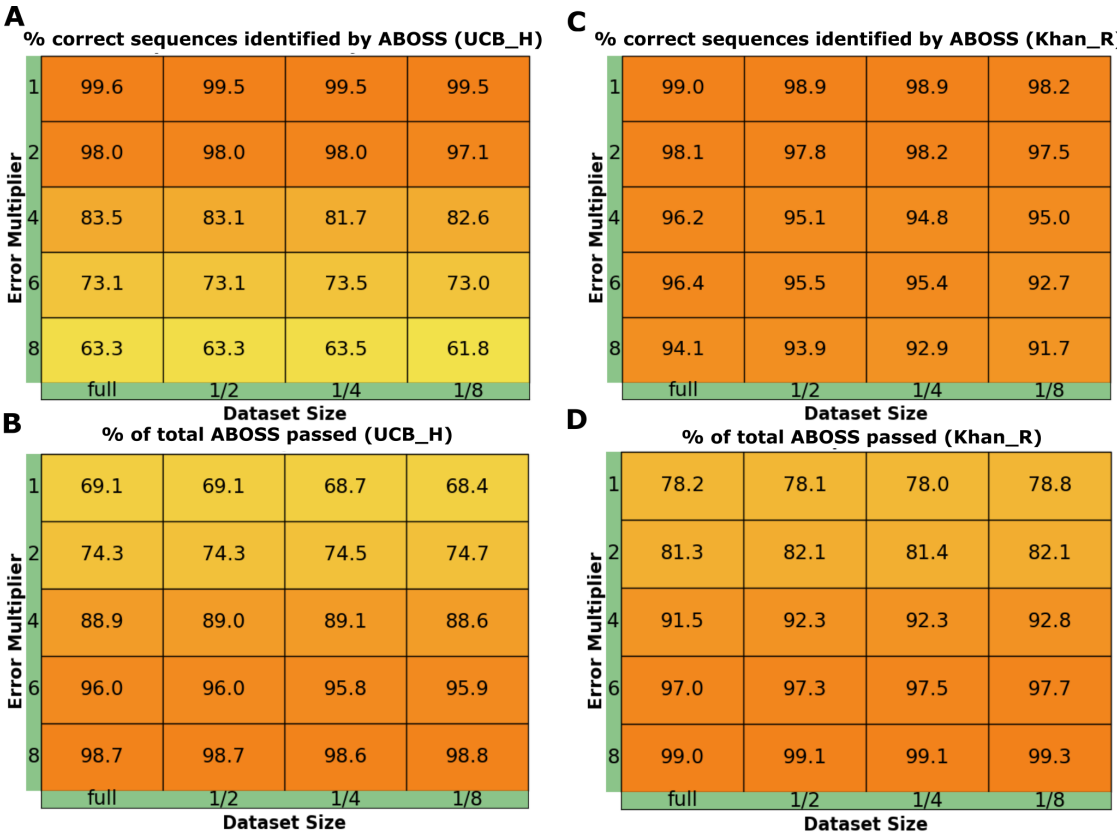


Figure 2.5: **BCR error simulation to assess sequences volumes and error rate tolerance of ABOSS.** The percentage outputs from BCR error simulations of UCB_H (A-B) and Khan_R (C-D) datasets. The X-axis corresponds to the proportions of UCB_H_Sim and Khan_R_Sim dataset sizes used for simulation relative to the size of the dataset that passed ANARCI. The Y-axis shows the multiplier of the original distribution of flagged residue/positions in the BCR datasets (see Figure 2.3). Plots (A,C) indicate the percentage of correct sequences that were incorporated in the UCB_H_Sim and Khan_R_Sim dataset that passed the ABOSS analysis, whilst plots (B,D) present the percentage of these sequences relative to the total number of UCB_H_Sim and Khan_R_Sim sequences respectively that passed ABOSS.

ABOSS-flagged residue/positions were used as the MRCA that were then employed as templates to which SHM mutations were introduced. The HH_S5F-targeting model [62] introduced roughly the same SHM substitution ratios along the VH region in the two lineage trees (Figure 2.6). There was a biased increase in SHM substitutions in framework 3 and CDR regions and positions flanking the CDRs, similar to previous results [62, 63]. As the HH_S5F model does not consider structural selection pressure on the heavy chain positions, the conserved cysteines were mutated, which resulted in the residue error rates of 0.002567 (Lineage_A) and 0.008 (Lineage_B) in UCB_H, and 0.002513 (Lineage_A) and 0.0076 (Lineage_B) in Healthy_H simulation datasets,

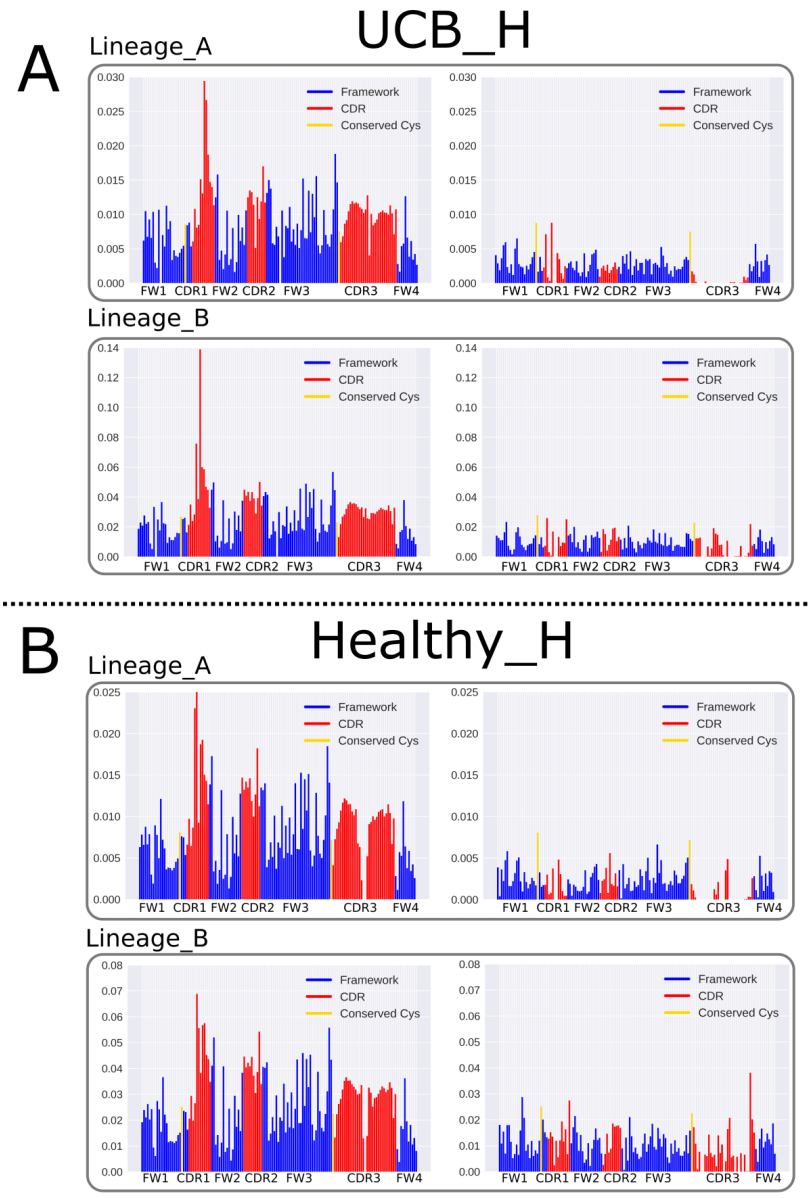


Figure 2.6: **ABOSS performance on SHM-simulated BCR data diversity**. Two BCR clonal lineage trees (Lineage_A and Lineage_B) were employed to provide the background mutational reference to introduce SHM substitutions into the ABOSS-filtered **A)** UCB_H and **B)** Healthy_H datasets using the human HH_S5F–targeting model [62]. The X-axis shows positions along the VH domain, and the Y-axis shows the proportions of residue/positions in the simulation datasets. The figures on the left depict the proportion of SHM substitutions introduced at positions in the VH domain. The figures on the right represent the proportions of ABOSS-flagged residue/positions in the simulation datasets.

respectively. The Lineage_A produced residue error rates were within the observed range of human BCR repertoire data, whereas the Lineage_B-generated residue error

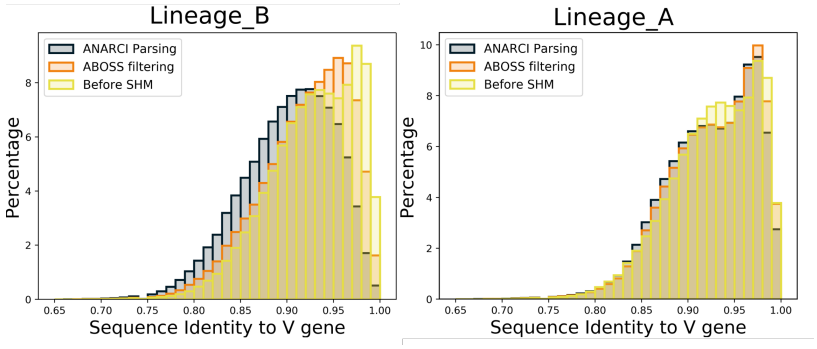


Figure 2.7: **Relationship between ABOSS filtering and V gene sequence identity**. The sequence identities to the closest V gene for ABOSS filtered UCB_H data (yellow), the SHM simulated (Lineage_A and Lineage_B) UCB_H data (grey) and ABOSS filtered SHM simulated UCB_H data (orange). The X-axis gives the sequence identity of the BCR data to the closest V gene sequence, whilst the Y-axis displays the percentage with that sequence identity. The majority of sequences in the ABOSS filtered data had germline mismatching residue/positions, less than 4% of the total sequences were identical to the V genes. Lineage_B had a higher substitution rate than Lineage_A. This produced a higher percentage of sequences in the BCR repertoire data with lower identities to the closest V gene as well as a higher residue error rate. As these sequences harbored an increased number of SHM substitutions, there was an increased probability for some of these substitutions to be found below the residue error rate in ABOSS analysis.

rates exceeded this range (Table 2.2). ABOSS exhibited no preferential selection of unmutated germline V gene sequences over sequences that harbor SHM mutations in the BCR data (Figure 2.7).

Most of the HH_S5F-generated diversity was preserved by ABOSS analysis (Figure 2.6). ABOSS flagged residue/positions uniformly along the VH domain, with the exception of CDR-H3, in which far fewer residue/positions were flagged. These proportions of ABOSS flags are unrelated to the pattern of generated SHM substitutions, which has a strong bias toward framework 3 and CDR loops. ABOSS algorithm was able to discern between structurally conserved regions (framework and canonical CDRs) and hypervariable CDR-H3 loops where a much higher positional residue diversity is permitted.

These results demonstrate that ABOSS is able to flag structurally nonviable residue/positions whilst preserving the majority of SHM substitutions. However, some true rare SHM substitutions may still be removed by ABOSS, as their positional presence in the variable region is below the residue error rate. Therefore, highly SHM-altered BCR datasets may have a higher proportion of true mutations (e.g. sorted plasma B-cell repertoires) incorrectly flagged as erroneous.

Dataset	Outputs	% of sequences found in the original dataset
ABOSS filtered UCB_H	1,903,703 (1,561,082)	100 (100)
IgReC-corrected UCB_H	5,572,963 (4,069,318)	51.4 (43.3)
IgReC on ABOSS filtered UCB_H	1,894,860 (1,320,438)	57.3 (47.9)
ABOSS on Healthy_H	486,437 (176,012)	100 (100)
IgReC-corrected Healthy_H	1,303,128 (367,235)	71.6 (60.4)
IgReC on ABOSS filtered Healthy_H	476,072 (61,281)	93.1 (87.9)

Table 2.3: **Interrogation of IgReC performance on UCB_H and Healthy_H**. IgReC was run on the raw nucleotide UCB_H and Healthy_H datasets as well as the ABOSS-filtered data. IgReC-constructed datasets derived from the raw data contained roughly 50 and 30% of sequences that were different to ones found in the UCB_H and Healthy_H datasets, respectively. When tested on the ABOSS-filtered datasets, IgReC was unable to find V and J germline references for 8,843 sequences in ABOSS-filtered UCB_H and 10,365 in ABOSS-filtered Healthy_H. IgReC also generated ∼42 and ∼9% of sequences that were not present in the original UCB_H and Healthy_H datasets. The default parameters were used to run IgReC: ./igrec.py –s reads.fasta –l IGH –o output_dir. The numbers and percentage of nonredundant sequences are shown in parentheses.

2.3.5 ABOSS and IgReC, an Ig-seq computational error correction tool

We compared ABOSS to IgReC, a computational Ig-seq error correction tool [181]. IgReC clusters and corrects PCR and sequencing errors in Ig-seq datasets based on sequence redundancy and homology. IgReC was recently benchmarked alongside other commonly used tools to error correct Ig-seq data [181]. Its performance was considered comparable, if not better, than all other tools tested. IgReC relies on identification of clonotypes and sequence clustering.

We ran IgReC on the UCB_H and Healthy_H datasets, as IgReC requires full-length VH or VL sequences, and the HEPB and Khan_R datasets have truncated framework 1 regions. IgReC uses positional nucleotide frequencies to modify Ig-seq dataset sequences, making it difficult to carry out an overlap comparison with ABOSS. IgReC removed ∼1.5% of UCB_H and 8% of Healthy_H, but modified nearly 50 and 30% of the sequences to ones not seen in the original UCB_H and Healthy_H datasets,

respectively (Table 2.3).

For both datasets, roughly 30% of the sequences in the IgReC corrected set contained ABOSS-flagged residue/positions. The redundancy of sequences that did not pass ABOSS but are found in the IgReC-corrected data are lower than the average of the IgReC corrected data (Table 2.4).

As the data above suggest that IgReC and ABOSS remove different sequences, ABOSS was run on the IgReC-corrected UCB_H and Healthy_H datasets. ABOSS filtered out 3,327,793 sequences (59.7%) from the IgReC-corrected UCB_H with a residue error rate of 0.0055. This error rate was very similar to that given by ABOSS for the original UCB_H dataset (see Table 2.2). Among the IgReC-corrected UCB_H sequences filtered out by ABOSS, 37,671 (1.1%) sequences failed to pass ANARCI, whereas the rest contained ABOSS-flagged residue/positions, of which 120,264 (3.6%) sequences lacked conserved cysteines. Applying ABOSS to the IgReC-corrected Healthy_H dataset yielded a residue error rate of 0.0041, which filtered 685,536 sequences (52.6%). Of these filtered sequences, 143,862 (21%) sequences failed ANARCI parsing, and of the rest with flagged residue/positions, 33,529 (4.9%) lacked cysteines at positions 23 and/or 104. IgReC analysis does not appear to correct stop codons, as at least one was identified in ∼85.7% of the sequences that failed to pass ANARCI parsing from the IgRec-corrected Healthy_H dataset.

We then tested the reverse protocol, running IgReC on the ABOSS-filtered UCB_H and Healthy_H datasets. IgReC generates structurally incorrect sequences (∼0.01%) when it is run on the ABOSS filtered UCB_H dataset. Many of these sequences had a stop codon introduced by IgReC. IgReC also altered the sequences of over 40 and 9% of the data to ones that were absent in the original UCB_H and Healthy_H datasets, respectively (Table 2.3).

Given these results, we would suggest that IgReC analysis can be enhanced by first using ABOSS to filter out structurally impossible BCR repertoire data.

2.3.6 Comparison to experimental BCR error correction methods

We also compared the results of ABOSS to two different experimental approaches, first to the work of Galson et al.,[175]. Their methodology of residue error estimation employs an analogous approach to ours. It is based on a calculation of the proportion of nucleotide mismatches to the germline in the sequence-invariant C region, which is adjacent to the framework 4 region of the heavy chain (FW-H4). ABOSS analysis on their HEPB dataset estimated the residue error rate to be 0.2276%. This is in

Dataset	Dataset size	Found in ABOSS filtered data	Found in ABOSS filtered data (%)
IgReC-corrected UCB_H	5,572,963 (4,069,318)	1,693,246 (1,136,620)	30%
IgReC-corrected Healthy_H	1,303,131 (367,238)	437,652 (166,842)	33%

Table 2.4: **Study of sequence overlaps between ABOSS filtered out data and IgReC-corrected data.** To investigate whether IgReC misses structurally incorrect sequences, ABOSS filtered UCB_H and Healthy_H sequences were searched in the respective IgReC-corrected data. In both the IgReC-corrected UCB_H and IgReC-corrected Healthy_H datasets, roughly 30% of the sequences were found in the ABOSS filtered out data. The numbers of non-redundant sequences in the data are shown in parenthesis.

the agreement with the residue error rates estimated by Galson *et al.,* [175], which ranged between 0.19 and 0.79%.

Secondly, we contrasted ABOSS with the experimental/computational error-correction protocol of Khan *et al.,* [120]. This method considers the entirety of the VH domain by attaching UMI barcodes to cDNA prior to Illumina sequencing, followed by clustering of identically UMI-barcoded sequences and error correction. The Khan_C dataset is the experimentally corrected version of the Khan_R dataset. In the process of this error-correction protocol, sequences are computationally modified (∼34% of Khan_C sequences have been altered from the sequences experimentally determined in the Khan_R dataset). This may be due to another sequence present in the Khan_R dataset (increasing redundancy from 3.7 in Khan_R to 45.3 in Khan_C) or to a new sequence altogether. The new sequence can be generated when the cluster representative sequence is yielded based on the positional amino acid frequencies of all cluster members. This modification means that the redundancy of sequence changes and that 0.5% of nonredundant (0.02% of redundant) sequences in the Khan_C dataset are not present in the original Khan_R dataset. These sequence changes make comparison with ABOSS difficult, as within the ABOSS protocol, no sequences are altered.

ABOSS analysis on Khan_R selects a similar number of nonredundant sequences to Khan_C (∼50,000), but only ∼6000 of these sequences are directly observed in the Khan_C dataset (Table 2.5). In terms of redundant sequences, ABOSS selects a far smaller set. This reflects the fact that sequences have been modified to increase the redundancy of specific sequences in the Khan_C dataset. The redundant over-

Data source	Dataset size	Redundancy	Overlap
Khan_C	2,385,080 (52,623)	45.3	36.8%
ABOSS filtered Khan_R	649,685 (47,593)	13.7	89.6%

Table 2.5: **Comparison analysis of ABOSS and the barcode approach of Khan *et al.,* [120].** ABOSS was run on the Khan_R dataset. The ABOSS outputs were contrasted with the Khan_C dataset (see Table 2.1 for dataset information). The overlap represents the percentage of total sequences that are shared between the Khan_C and ABOSS-filtered Khan_R datasets. ABOSS appears to be more conservative than the barcode approach. The numbers of nonredundant sequences are shown in parentheses.

laps between the ABOSS filtered Khan_R dataset and the Khan_C dataset are 36.8 and 89.6%, respectively (Table 2.5). Around 60% of Khan_C sequences are not seen in the ABOSS-filtered Khan_R dataset of these ∼1% that fail the ANARCI-parsing step (suggesting they would not produce viable antibodies), ∼0.04% are not found in Khan_R, the others contain residue/positions that are ABOSS flagged as below the residue error rate (Figure 2.8). Those flagged by ABOSS include ∼0.3% redundant and ∼8% nonredundant sequences that lack a cysteine at either position 23 or 104. About 10% of the ABOSS filtered Khan_R dataset are not found in the Khan_C dataset (Figure 2.8). Sequence modifications introduced to unprocessed Khan_R to create error-corrected Khan_C are recorded in detail in Khan *et al.,* [120] enabling tracing of the sequence changes. About 98.5% of the non-overlapping ABOSS filtered Khan_R with Khan_C datasets were no longer shared between these two datasets, whilst ∼1.5% contained ABOSS flagged residue/positions. Interrogation of steps performed in Khan *et al.,* [120] pipeline showed that the 98.5% of ABOSS non-overlapping sequences were clustered, error-corrected and collapsed to already existing sequences in ABOSS filtered Khan_R. This indicates that a small proportion of ABOSS filtered Khan_R sequences might actually be error harbouring sequences in the BCR dataset, but these sequences do not violate our knowledge of immunoglobulin folding and hence, are filtered as structurally viable. Interestingly, the UMI barcoding approach also alters viable antibody sequences found in the Khan_R dataset to incorrect found in Khan_C (1.5%) (Figure 2.8) (Discussed in more detail in Section 2.3.7).

ABOSS provides orthogonal functionality to BCR data error correction and can

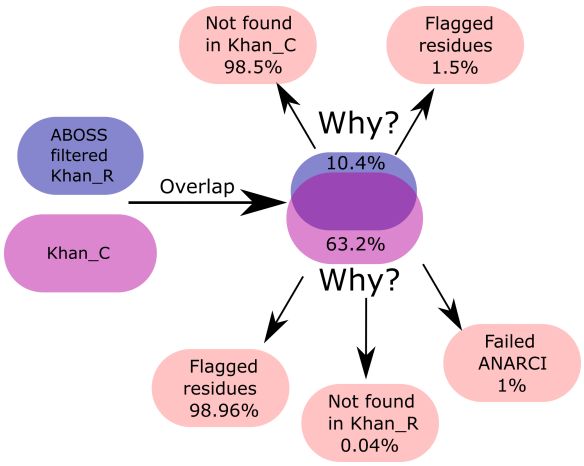


Figure 2.8: **Overlap analysis between ABOSS and the UMI-barcoding approach from Khan *et al.*, [120]** ABOSS was run on the uncorrected Khan_R dataset. The ABOSS outputs were contrasted with the error-corrected Khan_C dataset (see Table 2.1 for the dataset information. About 63% of sequences from Khan_C are not found in ABOSS filtered Khan_R. The majority of these sequences contain ABOSS flagged residue/positions, while ∼1% fail the ANARCI parsing step and ∼0.04% are new sequences that are not found in original Khan_R. Modifications between the Khan_R and Khan_C datasets are catalogued in detail which allows tracing amino acid changes introducing using the UMI barcode correction. About 10% of the ABOSS filtered Khan_R dataset do not overlap with the Khan_C dataset. Tracking sequence mutations from Khan_R to Khan_C shows that 98.5% these non-overlapping sequences are modified to ones present in the ABOSS filtered Khan_R dataset whilst ∼1.5% of the non-overlapping ABOSS filtered Khan_R sequences are changed to the sequences that have ABOSS flagged residue/positions.

be used to complement the UMI barcoding approach, an common practice in Ig-seq data analysis [86]. Performance of the barcode approach is heavily dependent on drawing a consensus sequence from a pool of identically barcoded sequences. Two common problems in the barcoding approach are when a large number of the barcoded sequences are singletons or several identically barcoded sequences share the highest redundancies in a cluster. These problems hamper the ability of the approach to correct data efficiently. ABOSS can be used prior to clustering to prevent all structurally nonviable sequences from becoming consensus sequences.

UMI barcodes are also used for accurate detection of template amplification and quantification biases in Ig-seq datasets [193]. This allows for the precise calculation of the amount and diversity of sequencing templates [120]. In this scenario, ABOSS should not be run prior to the barcode-correction approach, as it is a conservative tool that always reduces the dataset size and never alters BCR sequences.

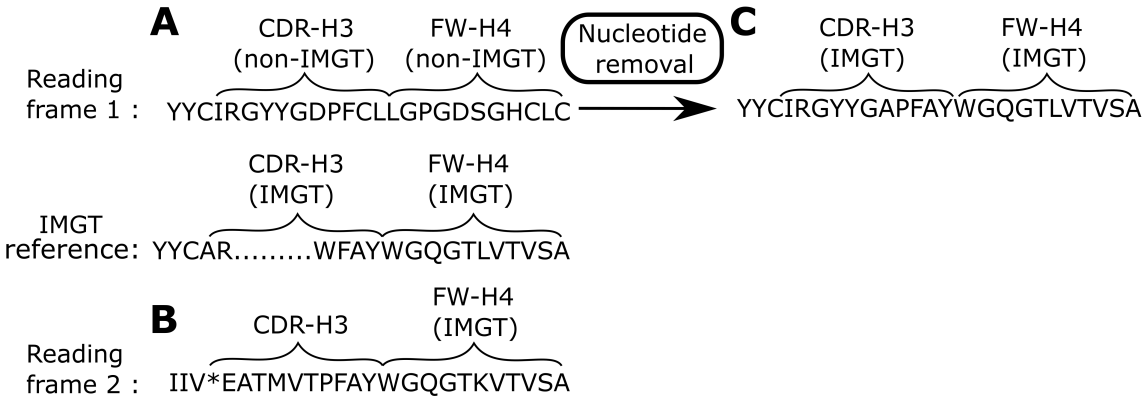


Figure 2.9: **ABOSS flags structurally non-viable antibody sequences.**. **(A)** The distal part of the VH chain sequence of an antibody that is selected as correct in the Khan_C dataset. The closest germline matches of the V and J genes for this sequence are IGHV5-4*02 and IGHJ3*01 respectively. This sequence is shown in the first reading frame. The FW-H4 region and the distal end of CDR-H3 of this sequence do not align to the IMGT amino acid germline. **(B)** Translating this antibody sequence into the second reading frame creates FW-H4 and the distal end of CDR-H3 that align to the IMGT amino acid germline. **(C)** An arbitrary deletion of a single nucleotide from the middle of the CDR-H3 region generates a structurally viable antibody sequence when it is translated in the first reading frame.

2.3.7 The orthogonality of ABOSS

As an example of how ABOSS identifies potentially non-structurally viable sequences that are not picked up by other techniques, Figure 2.9 shows an example of an antibody sequence from the UMI barcode corrected Khan_C dataset [120]. This sequence is translated into amino acids in the first reading frame. This sequence cannot be structurally viable, as FW-H4 and the distal end of CDR-H3 do not align to the known IMGT amino acid germline. Translating this sequence into the second reading frame reveals the correct translation of FW-H4 and a distal end of CDR-H3 that now align to the IMGT amino acid germline. This suggests that a single nucleotide insertion was introduced into the CDR-H3 region.

If we run ABOSS on Khan_C (the experimentally/computationally error-corrected set of Khan_R), the ANARCI-parsing step in conjunction with the check for conserved cysteines at IMGT positions 23 and 104 removed 11.5% of the unique sequences. These structurally impossible sequences correspond to 0.9% of the total redundant dataset (Figure 2.10). The inability of ANARCI to align the full-length FW-H4 to the IMGT germline was the main cause for sequences from the Khan_C dataset to fail ANARCI parsing, as these sequences were considered to have a truncated FW-H4

region.

A

Total number of sequences in Khan_C		2,385,080
Sequences that lack conserved cysteines		6,762
Sequences removed by ANARCI parsing	Non-amino acid residue in sequence	926
	Framework of canonical CDR indel	14,150
	No IMGT HMM match	13

B

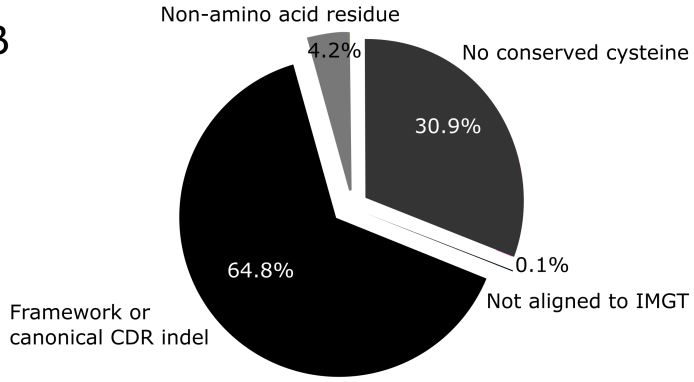


Figure 2.10: **Identification of structurally nonviable antibody sequences using first steps of the ABOSS pipeline on the Khan_C datasets.** Each sequence from the Khan_C (Table 2.1) dataset is examined for structural viability using ABOSS. **(A)** The tabulated output gives the total number sequences that did not pass the ANARCI parsing step. **(B)** The pie chart shows the percentage of sequences that fail the ANARCI step and those that lack a cysteine at position 23 and 104. The reasons include the following: 1) a sequence lacks a cysteine at position 23 or 104, 2) an indel is present in the framework or canonical CDR regions, 3) a non–amino acid residue is present in a sequence, or 4) a sequence does not align to the IMGT amino acid germlines of V or J genes.

These results demonstrate how leverage of our knowledge of immunoglobulin folding can help to filter data, even those that had been generated by a UMI-barcoding approach.

We have examined the robustness of the ABOSS protocol by running it on a dataset parsed by either ANARCI or annotated with IgBlastn [118], a sequence-centered Ig-seq data processing and annotation tool. ANARCI parsing removed ∼9% more sequences than IgBlastn (Table 2.6). However, if ABOSS is run on the ANARCI-parsed data or on data already passed by IgBlastn approximately, the same number of

Dataset name	Starting dataset	ANARCI parsed dataset	ABOSS filtered dataset	ABOSS Residue Error Rate (%)
Healthy_H	1,422,405 (745,276)	1,135,185 (558,171)	486,437 (176,012)	0.5427
Healthy_H_IgBlastn	1,228,129 (597,976)	1,135,116 (558,129)	486,429 (176,008)	0.5427
Healthy_L	6,317,736 (2,135,745)	4,860,389 (1,372,804)	2,667,263 (386,165)	0.4121
Healthy_L_IgBlastn	5,361,955 (1,539,964)	4,859,848 (1,372,519)	2,776,176 (386,146)	0.4118

Table 2.6: **IgBlastn and ANARCI parsing.** Performance of IgBlastn and ANARCI parsing was investigated on two datasets, Healthy_H and Healthy_L (see Table 2.1 for more details). IgBlastn analysis was performed on the nucleotide sequences that were downloaded from the Observed Antibody Space resource [43]. IgBlastn productive–called sequences were put into Healthy_H_IgBlastn and Healthy_L_IgBlastn datasets, respectively. ANARCI parsing was performed on the translated amino acid version of Healthy_H and Healthy_L (Table 2.1). All four datasets were then subjected to ABOSS analysis.

sequences are obtained. Examination of the sequences ANARCI removes and IgBlastn does not reveals that these sequences tend to not have a full-length framework 4 region or nothing at position 23 or had unusual indels in canonical CDR and framework regions. The ANARCI-parsed Healthy_H and Healthy_L datasets contained almost all (>99.98%) sequences that IgBlastn called productive. ABOSS analysis generated almost identical outputs for Healthy_H and Healthy_H_IgBlastn, whereas there was an increase (∼4%) in the number of Healthy_L_IgBlastn sequences compared with Healthy_L as a result of the slightly smaller residue error rate.

2.4 Discussion

ABOSS is an orthogonal redundancy–neutral method that uses structural information to calculate sequencing error rate estimates for BCR datasets. The novelty of our approach is founded in the application of current knowledge of immunoglobulin folding to identify and flag potential errors in BCR repertoire data.

ABOSS has been tested on six different BCR repertoires, ranging from 1,422,405 sequences to 9,985,575 sequences, which were generated by a variety of sequencing protocols. The ABOSS analysis pipeline is rapid and takes 10 hours to analyse 5

million unique BCR sequences on a standard desktop computer. ABOSS-calculated residue error rates agree well with experimental error rates that were available in Galson *et al.,* [175] where the researchers applied a similar approach for error estimation on the sequenced antibody constant region. ABOSS outputs original BCR repertoire data with a structural viability report for each submitted sequence together with IMGT amino acid numbering [36].

ABOSS identified 99% of correct antibody sequences in an *in silico* simulated BCR repertoire datasets when using dataset sizes and error rates matching those in the experiments. Since ABOSS considers each position individually, some structurally viable sequences with multiple compensatory mutations can be flagged as incorrect. Decreasing the size of the simulation BCR data did not affect the percentage of correct sequences recovered. Our simulation results suggest that even at far higher error rates, ABOSS performs well as long as either the repertoire redundancy is high or the dataset size is large enough (∼600,000 unique sequences).

The model selected to introduce *in silico* sequencing errors was based on Illumina technologies, in which nucleotide substitutions can happen stochastically along the antibody variable domain. For Roche 454 datasets [used to be one of the most common sources of BCR repertoire data [43]], nucleotide indel introduction along the sequenced amplicon is the main origin of sequencing errors. Further simulation experiments might be necessary to understand the behavior of ABOSS on Roche 454 BCR repertoires.

We also ran ABOSS on BCR repertoire data with computationally simulated SHM diversity to assess its ability to preserve viable mutations. These simulations indicated that ABOSS was able to spot structurally incorrect residue/positions whilst preserving the SHM-generated diversity. Most of the ABOSS flagged SHMs as structurally non-viable were located the antibody conserved regions (e.g. framework and canonical CDRs), whilst keeping the majority of the SHMs in the CDR-H3 loop. This confirms the structural basis of the ABOSS algorithm. It is also hard to assess the accuracy of SHM substitutions introduced by the HH_S5F model in the structural context. In the functional antibody repertoire, there are a number of positions where SHM substitutions are not observed, in particular, positions 23 and 104. Since the HH_S5F model is a probabilistic model, all positions have a chance of being mutated [62], as it was seen in our simulated datasets. Therefore, SHM in functional genes has a negative reinforcement effect on the residue error rate, which will mean ABOSS is less likely to flag positions that harbour SHM substitutions.

The nature of ABOSS analysis is orthogonal to current Ig-seq correction techniques; in particular, it does not alter sequences but rather removes those it considers to contain impossible structural features. Comparison to leading experimental and computational Ig-seq error-correction methods that do alter sequences shows that these approaches retain as well as create antibody sequences that are likely to be structurally nonviable (i.e., lack of cysteines at conserved position 23 and 104 or antibody regions that are out of the correct reading frame). For instance, the IgReC error-correction protocol [181] did not change the proportions of non-cysteine residues at IMGT positions 23 and 104 after it was run across several BCR repertoires. Similarly, more than 12% of unique sequences contained structurally unfavourable features after the UMI-barcoding approach carried out in Khan *et al.,* [120]. These results suggest that ABOSS should be used alongside current state-of-art error-correction protocols to increase confidence of structural viability of BCR repertoire sequences.

IgBlastn [118] is a highly popular Ig-seq data annotation software, which has been established as one of the core tools by Adaptive Immune Receptor Repertoire (AIRR) community [117]. IgBlastn enables researchers to generate a rich sequence annotation (e.g. V(D)J gene alleles) as well as to filter out non-productive sequences. IgBlastn leverages a bespoke version of BLAST algorithm [194] to query BCR sequences against the corresponding species IMGT germlines. We have demonstrated that the ANARCI parsing step is as robust as IgBlastn annotation to identify non-productive BCR sequences. In addition, ANARCI parsing is also able to spot structurally nonviable sequences that are missed by IgBlastn.

Although the ABOSS algorithm was solely developed on BCR repertoire data, it could be easily adapted to TCR sequencing data as variable domains of the TCR structure also obeys the canonical immunoglobulin configuration: three hyper-variable CDR loops that are structurally supported by four beta-sheet framework regions [195]. Similarly to BCRs, two cysteine residues form a disulphide bridge within each of the two variable domains of TCRs. This covalent bond is necessary for productive receptor formation [36]. Since T-cells lack the functional SHM machinery [45], any observed amino acid divergences from the germline in the variable region must be considered as a sequencing error. It comes with the assumption that all V(D)J genes have been discovered across all species populations as well as ethnic groups. To estimate the residue error rate in TCR repertoires, the ABOSS algorithm can be extrapolated across entire V and J gene regions enabling a more powerful calculation of the estimate.

In the next chapter, we will describe the development of Observed Antibody Space (OAS), the first resource that curates 1.8 billion BCR sequences gathered across 85 independent studies.

3

Observed Antibody Space: a resource for data mining of next-generation sequencing of antibody gene repertoires

Contents

This chapter is based on the following papers:

1. **Kovaltsuk, A.**, Leem, J., Kelm, S., Snowden, J., Deane, C.M. & Krawczyk, K. (2018) Observed Antibody Space: a resource for data mining next generation sequencing of antibody repertoires Journal of Immunology, 201(7):2502-2509

2. Krawczyk, K., Raybould, M.I.J., **Kovaltsuk, A.** & Deane, C.M. (2019) Looking for Therapeutic Antibodies in Next Generation Sequencing Repositories mAbs, 11(7):1197-1205

I carried out all the work described in this chapter unless noted otherwise.

3.1 Introduction

Technical advances in Ig-seq technology have outpaced storage and analysis pipelines [116, 117, 196]. This has meant that the outputs of B-cell receptor (BCR) studies are scattered across repositories, making it difficult to conduct large-scale data mining of BCR repertoires or inter-study comparisons [116]. Metadata, such as heavy chain isotype, B-cell subpopulation, B-cell donor age, or subject identifiers, are not typically standardised upon deposition of a BCR study into a public repository. Therefore, extraction of specific subsets of BCR repertoires for comparative analyses is challenging and always requires bespoke scripts for metadata standardisation. Furthermore, the sequencing data are typically deposited as raw nucleotide reads [197]. Nontrivial *ad hoc* efforts are needed to convert such raw reads to amino acid sequences that ultimately dictate molecular structure and antigen recognition. In addition, due to a wide range of constantly evolving experimental setups used in Ig-seq, multiple bespoke analysis pipelines are required to process raw reads [e.g. [4, 109, 198–200]]. Some of these issues are addressed by services that provide BCR/T-cell receptor (TCR)–specific data deposition and analysis pipelines such as the B-T.CR wiki (https://b-t.cr), ImmPort (http://immport.org) [201, 202], immunoSEQ Analyzer (http://clients.adaptivebiotech.com/), iReceptor (http://ireceptor.irmacs.sfu.ca/) [203], or VDJServer (http://vdjserver.org) [204]. iReceptor and VDJServer are the main resources that fall under the umbrella of the organized effort by the Adaptive Immune Receptor Repertoire (AIRR) community to provide standardised deposition and analysis pipelines for Ig-seq outputs [117]. These services chiefly focus on facilitating bulk deposition of raw data to perform standardised sequencing analyses. Recently, the AIRR community published a set of

minimal requirements (MiAIRR) in order to streamline submissions of BCR/TCR-data to iReceptor [117]. However, all these repositories primarily focus on nucleotide data curation. This prompts further data processing and installation of additional software packages to convert raw nucleotide data into a format suitable for BCR amino acid sequence/structural analyses. In the first part of this chapter, we describe the methods we developed to identify, clean and annotate both unpaired and paired BCR repertoire data, and make them available as a starting point for both sequence and structural immunoinformatics analyses.

To streamline large-scale data mining of publicly-available BCR repertoires, we have created the Observed Antibody Space (OAS) resource. In OAS, we provide access to both unpaired and paired BCR repertoires. We have collected the raw outputs of 80 unpaired BCR studies, covering over 1.8 billion sequences (as of October, 2020). Paired heavy/light chain BCR sequencing is an emerging technology, and so far, we have identified five studies that we have included in OAS (October 2020). As described in Chapter 1 (Section 1.2.2.2) paired sequencing technology requires the encapsulation of individual B-cells into separate emulsion droplets, the throughput of the number of sequenced B-cells is significantly lower than with unpaired sequencing technology [4, 103]. In paired OAS, we benchmarked different pipelines for short read BCR assembly to select the one that generated the highest quality contigs. All BCR repertoire sequences were converted into amino acids whilst preserving the link to the respective original raw nucleotide sequences. The amino acid sequences were numbered using the International ImMunoGeneTics information system (IMGT) scheme [36]. As of January 2020, all the paired and newly added unpaired studies contain a revised format of sequence annotation to ensure full compliance with the MiAIRR standards [117]. The data are available for querying or bulk download at `http://opig.stats.ox.ac.uk/webapps/oas/`. We believe that OAS will facilitate data-mining BCR repertoires for improved understanding of the dynamics of the immunesystem and, thus, better engineering of biotherapeutics.

The constantly increasing volume of publicly deposited BCR repertoires opens opportunity for orthogonal methods in therapeutic antibody discovery. Sequenced BCR repertoires chart the sequence space which is naturally allowed. This data can be exploited by machine learning and deep learning methods to extract immunological signals relevant for therapeutic antibody engineering [205–207]. However, it still remains unknown to what degree naturally occurring BCR sequences are similar to therapeutically purposed antibodies. Correlating therapeutic antibody sequence

similarity to natural BCR repertoires with their safety profiles will lead to the identification of sequence liability features such as anti-drug antibodies or poor developability profiles. This knowledge can be leveraged to accelerate discovery of safer ("more natural") biotherapeutics. In the second half of this chapter, we describe our analysis of sequence overlaps between all current clinical-stage therapeutic (CST) antibodies against the public BCR datasets curated in the OAS database.

3.2 Materials and Methods

3.2.1 Data Accession

A list of study accession codes of publicly available BCR datasets were obtained *via* a literature review. The majority of raw reads were downloaded in the FASTQ format from the European Nucleotide Archive (ENA) [208] and in the SRA format from the National Center for Biotechnology Information (NCBI) website [197]. In a small number of cases, another public BCR repository was specified [e.g [4, 94, 209, 210]]. Metadata were manually extracted from the deposited datasets and arranged in a reproducible format.

3.2.2 Raw unpaired nucleotide data preprocessing

3.2.2.1 Sequence assembly

If a raw nucleotide sequence file was downloaded in the SRA format, it was first unpacked into forward (R1) and reverse (R2) FASTQ reads. The downloaded FASTQ files were processed depending on the sequencing platform. Raw R1 and R2 Illumina reads were assembled into sequences with FLASH [211]. The assembled BCR sequences were converted to the FASTA format using FASTX-Toolkit [212]. All Roche 454 reads were downloaded exclusively in the FASTQ format. As raw reads from Roche 454 are not paired (*i.e.* only contain R1), these FASTQ files were directly converted to the FASTA format with the FASTX-Toolkit.

3.2.2.2 Isotype identification

The VH chain sequences were automatically annotated with isotype information unless such data were given in the corresponding publication. No antibody subclass information was obtained. Automatic isotype annotation was performed by aligning first 21 nucleotides of the 5' constant heavy domain 1 (CH1) of any given BCR sequence against the IMGT isotype reference [36] of the respective species using the

Smith–Waterman algorithm [213]. We assigned a score of 2 for a nucleotide match and a score of -1 for a nucleotide mismatch or a gap in the local alignment. The IMGT isotype references comprised 21-nucleotide–long fragments of the CH1 domain of the antibody isotypes. To ensure a high confidence of correct isotype assignment, we employed a conservative threshold of 30 in the Smith–Waterman algorithm scoring function.

Sequences whose Smith–Watermann algorithm score was below the threshold for all isotypes were assigned as "bulk". The robustness of this protocol was confirmed on the author-annotated BCR datasets [23, 185, 214], on which it achieved 99% accurate annotations. Around 1% of the BCR repertoire sequences had a very short (or missing) CH1 domain sequence. Such sequences were also assigned as bulk.

3.2.2.3 ANARCI annotation

IgBlastn [118] was used to convert the FASTA files of BCR nucleotide sequences to amino acids. The amino acid sequences were then parsed with ANARCI [34] using the IMGT scheme [36]. In this step, every sequence is IMGT-numbered and inspected for compliance based on the following rules derived from our knowledge of immunoglobulin folding. Amino acid sequences that harbor unusual indels in canonical CDR and framework regions or stop codons are removed as these are considered structurally nonviable. ANARCI does not number a sequence if its V and J genes do not align to a Hidden Markov Model [215] built on its respective species amino acid IMGT germlines [36]. We also filter out potentially chimeric sequences by detecting duplicated CDR-H3 regions in every amino acid sequence, checking for the complete sequence residue annotation, checking for the full-length framework 4 region, and imposing a length cutoff of 37 residues for CDR-H3 in human, mouse, rat, rabbit, alpaca, and rhesus antibodies.

The technical limitations of sequencing platforms means that certain reads were missing significant portions of the V region (e.g., portions of CDR1); sequences that did not have all three CDRs were discarded as incomplete. Missing portions of V/J genes in the OAS sequences were not extended with the germline information.

The V and J gene annotation available in OAS is obtained using ANARCI, which identifies the germline genes with the highest amino acid identity [34]. The ANARCI germline database was compiled using V/J genes downloaded from the IMGT website [19]. As the V and J genes of camels have not yet been well characterised [216], we employed the alpaca (the closest relative available) immunoglobulin genes in camel BCR data annotation, as these two species belong to the same biological family

(*Camelidae*). As data from other poorly cataloged species are added to OAS, we will use the closest available relative for V and J gene annotation.

3.2.2.4 Automation of new BCR repertoire identification

Using the protocol above, we have so far annotated unpaired BCR repertoires from 80 independent studies. To streamline updating OAS with new data, we have generated a procedure to automatically identify BCR datasets from raw sequence read archives. We apply our BCR annotation protocol to each raw nucleotide dataset deposited in the NCBI/ENA repositories; if we find more than 10,000 BCR sequences in any given dataset, it is set aside for manual inspection. Manual inspection is still necessary to efficiently assign metadata, as these are currently deposited in a non-standardised manner. This procedure allows for automatic identification of new BCR datasets and semi-automatical updating of OAS.

3.2.2.5 Compliance with the AIRR community standards

As of January 2020, the format of BCR repertoires added to OAS has been amended to comply with the MiAIRR requirements mandated by the AIRR community [117]. To meet these standards, the sequence annotation and ANARCI parsing steps (Section 3.2.2.3) have been modified accordingly.

IgBlastn is the cornerstone of many AIRR-data analysis pipelines [118]. It provides a rich annotation of AIRR-data such as nucleotide indel information, V(D)J gene allele matches. The MiAIRR standards mandate that all AIRR-tools should work directly with the IgBlastn outputs [117]. Previously IgBlastn tool was integrated into the OAS sequence processing pipeline to solely translate BCR sequences into amino acids in the correct reading frame. Now all newly OAS deposited BCR repertoires contain the full IgBlastn sequence annotation. This includes more than 100 unique BCR sequence descriptors such as the length of antibody regions, percentage identity to closest V and J genes *etc.*

AIRR-data analysis pipelines work primarily within the remit of annotated V(D)J gene alleles. It enables researchers to decipher patterns of somatic hypermutations (SHMs) [23], study topology of clonal lineage trees [126, 217] or compare distributions of V(D)J gene usages [122]. The AIRR-tools can work with sequences that we would consider structurally non-viable (e.g. indels or missing antibody regions; see Chapter 2) as long as V(D)J genes can be identified. For instance, a missing complementarity-determining-region 1 of the heavy chain (CDR-H1) would only have a minor impact on correct V and J gene identification. Therefore, simply filtering out all structurally

non-viable BCR sequences from OAS would lead to an undesirable loss of information in the AIRR-data analysis pipelines. To meet the data quality requirements of both structural biology and AIRR communities, instead of just removing all structural non-viable antibody sequences, each OAS sequence now contains a short structural viability report. This gives flexibility to OAS users to adjust sequence quality based on the needs of their analyses. The short viability report also allows users to revert back to the original OAS sequence quality setup ensuring the same standard across all OAS deposited studies. The only quality controls implemented prior to uploading BCR repertoires are based on IgBlastn sequence annotation. Here, sequences that are productive, in the correct reading frame and do not harbour any stop codons are added to the OAS resource.

To filter out poorly sequenced BCR amplicons, we rely on IgBlastn sequence annotations. We only keep BCR sequences that are productive, in the correct reading frame, and do not have stop codons. In each dataset, only unique antibody sequences are retained based on identical amino acid sequences of the V(D)J region.

3.2.3 Raw paired nucleotide data preprocessing

3.2.3.1 Raw sequence assembly and annotation

The downloaded forward and reverse 10xGenomics Illumina reads were assembled into contigs using CellRanger v3.1, the official 10xGenomics software (https://support.10xgenomics.com/single-cell-vdj). CellRanger outputs the assembled contigs in the FASTA format as well as providing BCR nucleotide sequence annotation. IgBlastn is then employed both to annotate the contig sequences according to the AIRR community standards and to convert nucleotide sequences into amino acids as described in Section 3.2.2.5. Next, ANARCI parsing is used to IMGT-number [36] and check structural viability of each amino acid BCR sequence. This step was identical to one used in the unpaired January 2020 version of OAS (Section 3.2.2.5). Similar to the unpaired OAS version, VH contigs were annotated with the isotype information based on the constant domain sequence. The processed BCR repertoires were then combined with the CellRanger sequence annotation and author-provided sequencing sample metadata to form an OAS Data Unit. All these steps are visualised in Figure 3.6.

3.2.3.2 Linking VH and VL sequences

During B-cell droplet compartmentalisation, unique 10xGenomics DNA barcodes are attached to VH and VL transcripts within each individual B-cell. Hence, in the ideal case scenario, any given 10xGenomics barcode should only be associated with one unique VH and one unique VL contigs ("1-to-1 pairing") that originated from the same B-cell. However, in many cases more than one unique VH and/or VL chain sequences harbour identical 10xGenomics DNA barcodes (Figure 3.9). Linking such VH and VL chain sequences combinatorially can lead to artificial inflation of the real sequence number and incorrect estimation of the repertoire diversity (Figure 3.7).

To address this combinatorial sequence inflation problem, we supply two separate output formats of paired Data Units in OAS: linked and unlinked. In the linked version, we explicitly filter for contigs whose 10xGenomics barcodes are associated with only one unique VH and one VL sequences. These contigs are then merged on the 10xGenomics barcode information resulting in "1-to-1" VH-VL linked contigs. However, some B-cells can express multiple functional BCRs that originate from separate somatic V(D)J recombination processes (multi-immunoglobulin specificity) [218]. To account for this multi-immunoglobulin B-cell specificity, we supply the unpaired version of OAS Data Units. There, we do not merge VH and VL contigs on the 10xGenomics barcode information, but provide all contigs that passed CellRanger, IgBlastn and ANARCI annotations as shown in Figure 3.6. This format gives extended flexibility to computationally skilled OAS users to define their own cutoffs for 10xGenomics barcode sharedness between VH and VL contigs which widens the scope of potential bioinformatics analyses.

3.2.4 10xGenomics contig assembly with SSAKE and CellRanger softwares

CellRanger and SSAKE [104] were benchmarked for the quality of BCR contig assembly from raw 10xGenomics Illumina reads. For this assessment, raw BCR repertoires were downloaded from Goldstein et al., [103]. We employed the same CellRanger assembly pipeline as described in paired OAS (Section 3.2.3.1) to process the raw BCR repertoires.

We created a new BCR sequence assembly pipeline to scrutinise V(D)J contigs generated with the SSAKE assembler using raw 10xGenomics Illumina reads. The pipeline encompasses five major steps which are visualised in Figure 3.10. First, we derive 10xGenomics barcodes that have sufficient read coverage using the same

approach as described in Goldstein et al., [103]. In each BCR repertoire, all 10xGenomics barcodes are isolated from the forward reads and counted. The barcodes are sorted by their number of reads. The read number cut-off for sufficient coverage was calculated as one tenth of the number of reads of the 150^{th} most frequent barcode (see Figure 3.10). For instance, if the 150^{th} most frequent barcode was found in 3000 reads, then any reads whose barcodes appear in less than 300 reads are removed before any subsequent contig assembly steps. Raw reads are then written to new FASTA files for each unique 10xGenomics barcode. SSAKE is used to assemble the filtered reads into contigs employing the same parameters as in Goldstein et al., [103]. SSAKE generates tens of varying-length contigs, where most of them do not span the full variable chain. Hence, IgBlastn is used to annotate the contigs by aligning them to IMGT germlines [36, 118]. Based on the IgBlastn outputs, the following contig filtering criteria are applied: 1) V and J genes can be annotated, 2) contig amino acid length is at least 100 amino acids, 3) contigs are productive and in the correct reading frame, 4) there are no stop codons. In some cases, more than two filtered contigs associate with the same 10xGenomics barcode. If more than one of these contigs share the same VJ gene rearrangements, we only retain the contig with the highest count of unique molecular identifiers (UMIs).

3.2.5 Looking for therapeutic antibodies in the OAS database

3.2.5.1 OAS Database

The OAS resource is constantly updated as new BCR repertoires are released into the public domain. Our study was conducted in March 2019, at this time only unpaired BCR datasets were curated in OAS. Furthermore, only 60 out of the current 80 BCR studies were available. The total sequence number was approximately 1 billion (∼960 million heavy chain and ∼60 million light chain sequence). These BCR studies still covered multiple organisms (primarily human, mouse, rhesus, rabbit, camel and rat), individuals and immune states. At this time (March 2019), Briney et al., [4] was by far the largest BCR study that accounted for one third of all heavy chain sequences in OAS.

3.2.5.2 Formatting clinical stage-therapeutic antibody sequences

Clinical stage-therapeutic (CST) antibody sequences were obtained from Raybould et al., [167]. To comply with the OAS format and enable length independent antibody sequence comparisons, the CST were numbered with ANARCI [34] according to the

IMGT scheme [36]. The CST sequences were classified into four groups (chimeric, humanised, human, mouse), based on their international non-proprietary names [219]. Sequences with names containing '-xizumab' or '-ximab' were labeled as 'chimeric'. Sequences not matching this criterion but containing '-zumab' in their name were classified as 'humanised'. Sequences that contained only '-umab' in their name were labelled as 'fully human'. Three mouse antibodies (muromonab, abagovomab and racotumomab), were labeled as 'mouse'.

3.2.5.3 Aligning CSTs to OAS data

We separately aligned the heavy chain, light chain, the combination of the three heavy or light chains IMGT-defined CDRs, and the IMGT-defined CDR-H3 of CSTs to each of the sequences in OAS. We noted a match if an IMGT position in a "query" CST is also found in a "template" sequence from OAS, and they have the same amino acid residue. For the full sequence alignments, the number of matches is divided by the length of the query and by the length of the template, producing two sequence identities. The final sequence identity is the average between these two. Calculating the sequence identity in this way prevents the scenario when one sequence is a substring of another, creating an artificially high sequence identity with a large length discrepancy. The CDR alignments were performed only when the IMGT-defined loop lengths matched.

3.3 Results

The OAS resource provides access to both paired and unpaired BCR repertoires. So far we have collected raw sequencing outputs from 80 unpaired and 5 paired BCR studies. OAS data can easily be accessed from the main OAS web page (`http://opig.stats.ox.ac.uk/webapps/oas`) (Figure 3.1).

3.3.1 Unpaired version of OAS

3.3.1.1 Unpaired data annotation

All raw BCR nucleotide reads were annotated and translated into amino acids using IgBlastn [118]. Within OAS, it is possible to link back from the translated amino acid sequences to the raw nucleotide data. The full amino acid sequences were then IMGT numbered using ANARCI [34]. As well as providing IMGT and gene annotations, ANARCI acts as a broad-brush filter of BCR sequences that are likely to be erroneous

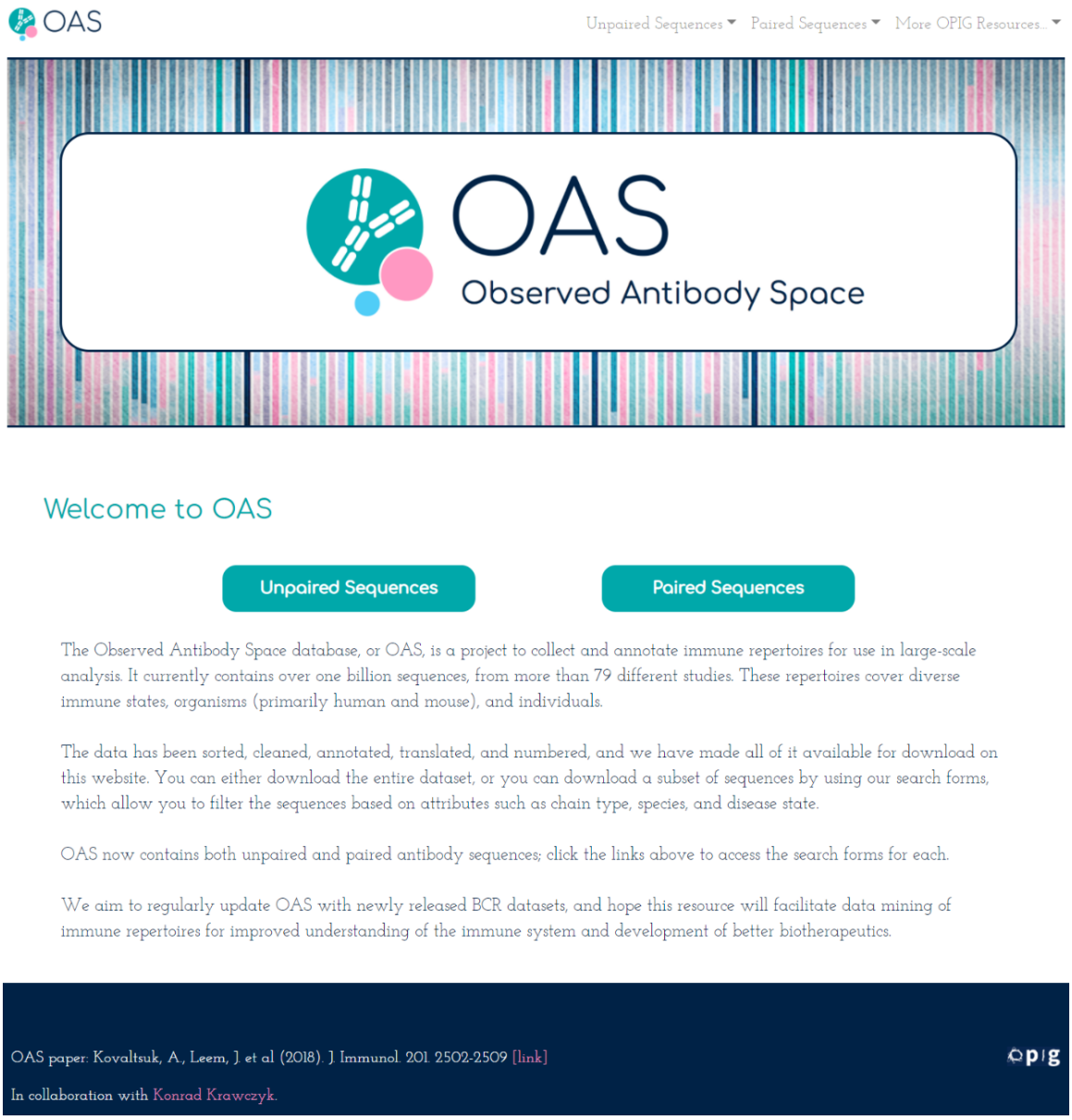


Figure 3.1: **Main OAS web page**. The welcome page provides a concise introduction about the data deposited in OAS. The users can easily locate two buttons in the center of the page that will redirect them to unpaired and paired BCR data searches respectively. Users can also find more information on Search, Download and Help by clicking on *unpaired sequences* and *paired sequences* tabs at the top of the page.

(see Section 3.2.2.3). For each BCR dataset, we provide the total number of amino acids that were retrieved from IgBlastn outputs as well as after ANARCI parsing. These numbers may be useful as proxies for dataset quality assessment.

Applying the same retrieval, amino acid conversion, gene annotation, and numbering protocol to all sequences assures the same point of reference across the 80 heterogeneous BCR studies. This protocol produces the full IMGT-numbered sequences together with gene annotations for each of the 80 studies.

3.3.1.2 Implementing the AIRR standards into OAS

Researchers within the AIRR community developed a large arsenal of tools that primarily work within the scope of nucleotide data [86]. To streamline cross-communication between those tools, the AIRR community issued the recommended minimal standards (MiAIRR) for working with BCR/TCR-repertoire (often referred as AIRR-data) files. The format is a tab-delimited text file with a pre-defined set of key columns [117]. Nucleotide BCR data is also deposited in OAS. However, OAS lacks the MiAIRR annotation, as the resource was originally developed to provide numbered amino acid sequences for structural BCR repertoire interrogation [102, 153]. In January 2020 the MiAIRR standards were fully integrated into OAS to benefit from cross-fertilisation of structural biology and AIRR community analysis approaches. (see Methods Section 3.2.2.5).

3.3.1.3 OAS Data Units

The processed sequences in each BCR repertoire are annotated with metadata (e.g., individuals, age, vaccination regimen, B-cell type, and source, etc.) and split into individual files based on the isotype information (Figure 3.2). Deposition of such metadata is currently not standardised amongst researchers and requires *ad hoc* manual curation for each dataset. In an effort to organise the BCR sequences using such metadata, we have grouped the sequences within each dataset into Data Units. Each Data Unit represents a group of sequences within a given dataset with a unique combination of metadata values. The metadata values are summarised in Table 3.1.

3.3.1.4 Unpaired OAS statistics

As of October 2020, 80 BCR studies are included in unpaired OAS, totaling 1,908,875,113 sequences (1,604,110,219 VH and 304,764,894 VL sequences). The deposition of BCR sequences into the public domain markedly increased in the last four years (Figure 3.3). Between 2010 and 2016 the total number of OAS deposited

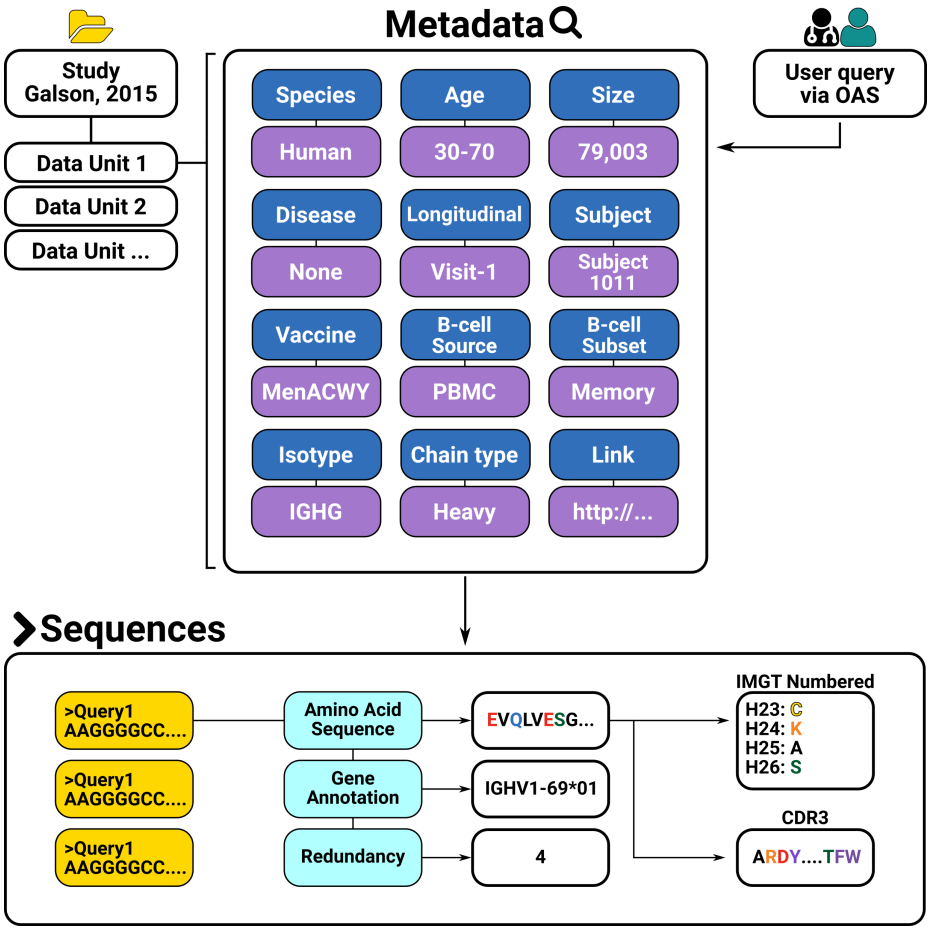


Figure 3.2: **Description of an OAS Data Unit**. The data from 80 BCR studies are sorted into Data Units. Each Data Unit is a set of BCR sequences that share the same set of metadata. Each sequence in a Data Unit is further associated with sequence-specific annotations generated by ANARCI and IgBlastn.

BCR sequences is ∼241 million, whereas every subsequent year contains at least 250 million sequences. In particular, high expansions of BCR repertoire sequence availability can be observed in 2017 and 2019 (Figure 3.3). In 2017, Greiff et al., [94] released more than 240 million mouse BCR sequences - the largest BCR repertoire study at that time. In 2019, two independent studies interrogated human healthy BCR repertoire diversities [3, 4], where leukapheresis was employed as a technique to collect unprecedented numbers of B-cells from subjects for Ig-seq. These two studies totalled in more than 500 million annotated BCR sequences.

The unpaired OAS resource harbours BCR sequence information generated by two separate sequencing technologies: Roche 454 pyrosequencing and Illumina. Roche 454 was the first sequencing technology successfully applied to glean sequence diversity information from BCR repertoires [90, 91]. Roche 454 is the sole source of BCR data

Metadata Name	Metadata Description
Chain	Heavy chain/Light chain annotation
Isotype	Identified or deposited isotype information
Age	Information on the age of the human B cell donors
Disease	Indication of whether the donor was sick at the time of B cell extraction
Vaccine	Indication if the B cell donor was purposely immunized prior to B cell extraction
B-cell subset	Indication if a particular B cell subset was sorted for Ig-seq
Species	Organism of the B cell donor
B-cell source	Organ/tissue from which the B cells were extracted
Subject	Indication of a particular B cell donor from whom the B cells were sourced
Longitudinal	If the study was longitudinal, an indicator of the time point
Size	Number of redundant amino acid sequences in the Data Unit
Size_igblastn	Number of redundant amino acid sequences extracted from Igblastn outputs prior to ANARCI parsing
Unique sequences	Number of unique amino acid sequences in the Data Unit
Link	Link to the source publication
MiAIRR	Indication if the Data Unit format complies with the minimal AIRR community standards
Processed	Indication if the corresponding authors of the study made their processed data available

Table 3.1: **Metadata descriptors of each Data Unit in OAS.** Each Data Unit is uniquely identified by the study and a collection of the metadata values.

in OAS between 2010 and 2013. With the rapid advances in paired short read Illumina sequencing, the official support for Roche 454 was discontinued in 2013. Since 2014 Illumina technology is the dominant source of BCR sequences. (Figure 3.4). Several research groups still used Roche 454 for Ig-seq until 2017. As of October 2020, 98.5% of all OAS sequences were generated using Illumina and only 1.5% using Roche 454.

The availability of BCR studies with access to sequencing data has significantly increased in recent years (Figure 3.5). In the early days of Ig-seq (years 2010 - 2012), only a single study per year made its BCR repertoires public[1]. Since then, the number of publicly available BCR studies has ranged between 6 and 13 each year. Assuming that approximately the same yearly percentage of the total BCR studies allow public access to their data, it can be concluded that Ig-seq has become an increasingly popular technique to study BCR repertoires (Figure 3.5).

[1]Only BCR repertoires that passed the OAS sequence processing pipeline were counted.

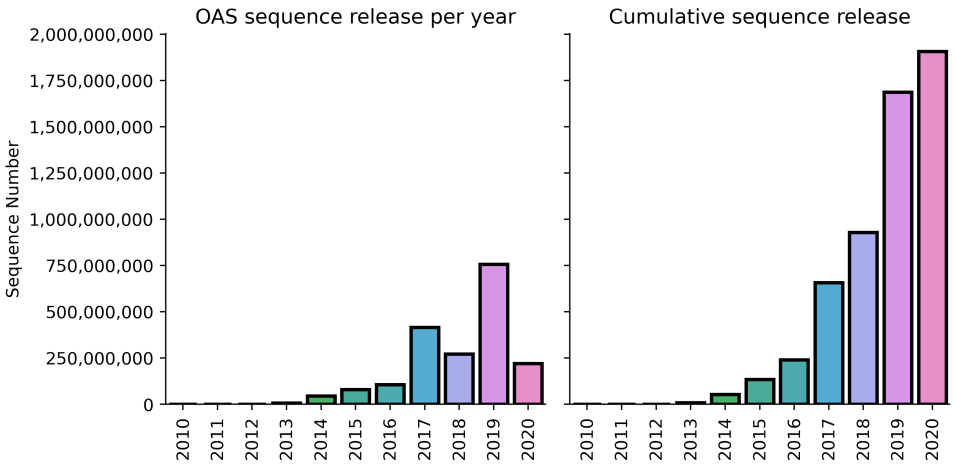


Figure 3.3: **Number of deposited sequences in unpaired OAS.** All OAS deposited BCR sequences were grouped by the publication year. The number of publicly deposited BCR sequences per year are depicted on the left, and the corresponding cumulative distribution is shown on the right. The earliest study contained in OAS dates back to 2010 [220]. Between 2010 and 2016, yearly depositions of BCR sequences rose steadily from 30,000 to 100 million sequences. Since 2017, advances in sequencing technologies and experimental protocols allowed scientists to obtain more comprehensive snapshots of BCR repertoires with more than 250 million sequences released publicly each year [3, 4, 94].

The majority of the sequences deposited in OAS are human (∼82.3%) followed by mouse (∼17%). Forty-three (or 53.8%) of the BCR studies interrogate the immune system of diseased individuals, the most common ailment being HIV which contributes to 20 studies. In 2020, the research focus of the AIRR community rapidly shifted to decipher the biological footprint of the novel SARS-COV-2 virus on the human adaptive immune system [221]. Currently, OAS provides access to five SARS-COV-2 BCR studies where patients displayed a wide range of clinical outcomes. The OAS database also contains 55 (or 68.8%) BCR studies of antigen-non-stimulated BCR repertoires (the collection of B-cells from donors who are healthy and not purposefully vaccinated).

The main source of B-cells in the OAS database is peripheral blood (∼858 million of sequences), followed by leukapheresis-filtered PBMCs (∼539 million), splenocytes (∼216 million) and bone marrow (∼164 million). The database holds isotype information for each individual heavy sequence, and the two most common isotypes are IGHM (∼676 million) and IGHG (∼334 million). For ∼427 million IGH sequences, we were not able to assign isotypes with high confidence. The large number of these isotype non-annotated sequences came from Soto et al., [3], where the corresponding authors

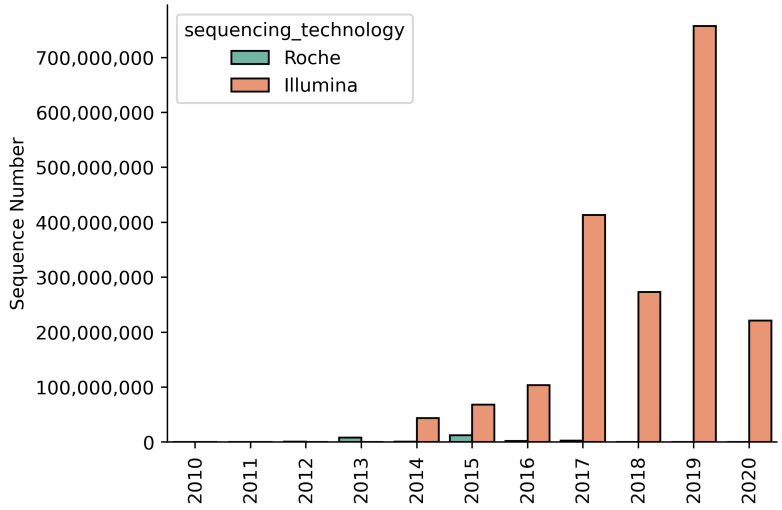


Figure 3.4: **Number of sequences deposited in unpaired OAS each year stratified by sequencing technology used.** Currently unpaired OAS incorporates BCR sequencing outputs from two sequencing platforms: Roche 454 and Illumina. Since long read Roche 454 sequencing was developed before Illumina, Roche 454 sequencing was the sole source of BCR sequences in OAS between year 2010 and 2013. In 2014 the first Illumina sequenced BCR repertoires were made publicly available which provided better sequence quality and higher sequencing depth than Roche 454. Although Roche 454 technology was discontinued in 2013, some research groups continued using this technology to generate BCR repertoires until 2017.

supplied BCR sequences with removed CH1 domains (∼141 million sequences).

The median redundant size of the BCR studies in the OAS database is 5.4 million sequences, whereas the largest BCR study was that by Briney et al., [4] (∼319 redundant sequences). Two sequences are redundant if they are of identical length, identical amino acid composition and identical isotype information. Detailed statistics on each dataset are given in Appendix Table C.1. All the data may be bulk downloaded or individual Data Units queried at `http://opig.stats.ox.ac.uk/webapps/oas/oas`.

3.3.2 Paired version of OAS

3.3.2.1 Paired data annotation

All OAS deposited paired BCR repertoires were generated using 10xGenomics technology. Since it involves paired short read (150 base pairs) Illumina sequencing, computational read assembly is required to obtain full length variable chain sequences. We tested two different assembler softwares (see Section 3.3.3) and found that 10xGenomics CellRanger creates higher quality BCR contigs than general short-

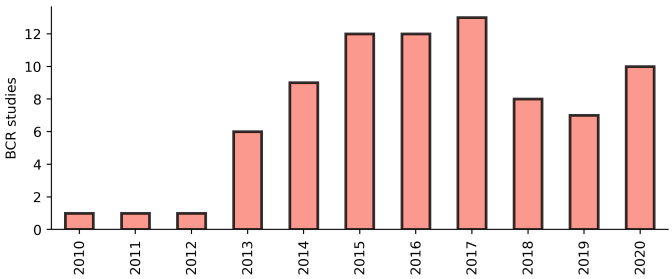


Figure 3.5: **Number of BCR studies in OAS by year**. All BCR repertoires with public data access were downloaded. Only BCR studes that passed OAS quality filters were counted and added to OAS.

read SSAKE assembler [104]. Therefore, CellRanger was used to assemble all raw forward and revearse BCR reads into contigs in OAS (Figure 3.6). The assembled contigs were annotated and translated into amino acids according to the MiAIRR standards [118]. These annotated contigs were combined with the author provided metadata and ANARCI antibody IMGT numbering to form an unlinked OAS Data Unit (Figure 3.6). For each unlined Data Unit, an accompanying linked version was created where 10xGenomics barcode sharedness was scrutinised to prevent undesirable combinatorial inflation of unique VH-VL contigs (Figure 3.7). Only contigs whose 10xGenomics barcodes were associated with one unique VH and one unique VL chains are retained in the linked OAS Data Unit version.

3.3.2.2 Paired OAS statistics

As of October 2020, five BCR studies are included in paired OAS. Two of those were released in 2019, the remaining three were published in 2020. Although only five studies are currently deposited in OAS, they cover BCR repertoires from three separate species (Figure 3.8). Detailed information about these BCR repertoires can be found in Table 3.2. Since paired OAS employs two separate VH-VL linking approaches for each Data Unit (linked and unlinked), below are two subsections about paired OAS statistics.

3.3.2.3 Linked Data Unit

The total number of unique "1-to-1" VH-VL contigs in the linked version of paired OAS is 121,838. More than 50% of these pairs came from a single study [103] (Figure 3.8). Sixty percent of all BCR sequences are from non-antigen stimulated B-cell donors. The most common disease in paired OAS is tonsillitis (27,744 sequences

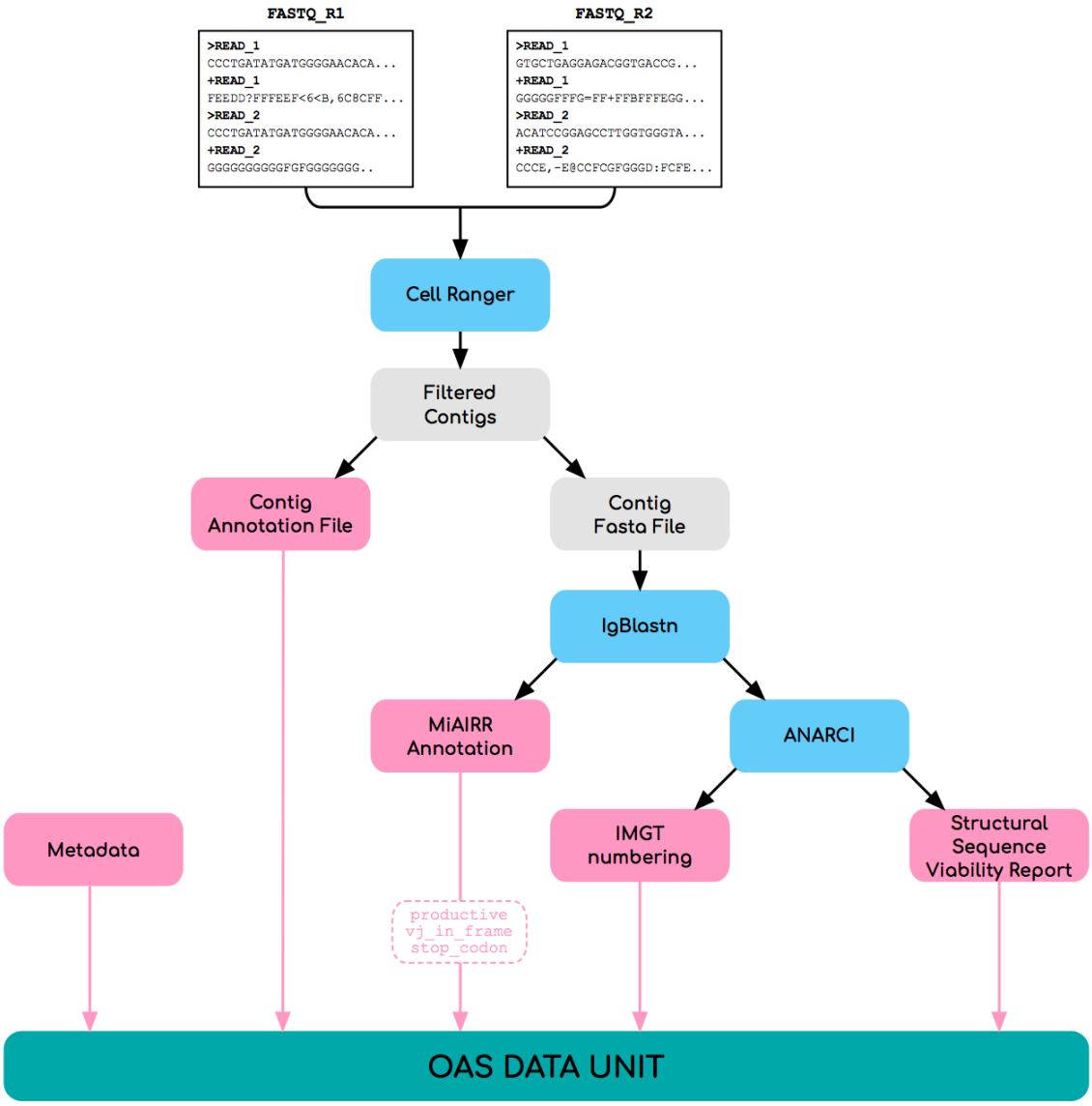


Figure 3.6: **Processing 10xGenomics raw outputs for OAS**. Raw forward and reverse Illumina reads are obtained from the NCBI website [197]. 10xGenomics CellRanger is employed to assemble these reads into variable chain contigs. The contig fasta file is then annotated according to the MiAIRR standards with IgBlastn [118]. Only sequences that are productive, in the correct reading frame, and do not contain stop codons are retained. Next, ANARCI is employed to provide a structural viability report and IMGT numbering for each variable chain sequence. Finally, IgBlastn, CellRanger and ANARCI annotations are combined with the author-provided metadata to form an OAS data unit.

or 22.7%), followed by SARS-COV-2 (11,388 sequences or 9.3%) and HIV (5,547 or 4.5%).

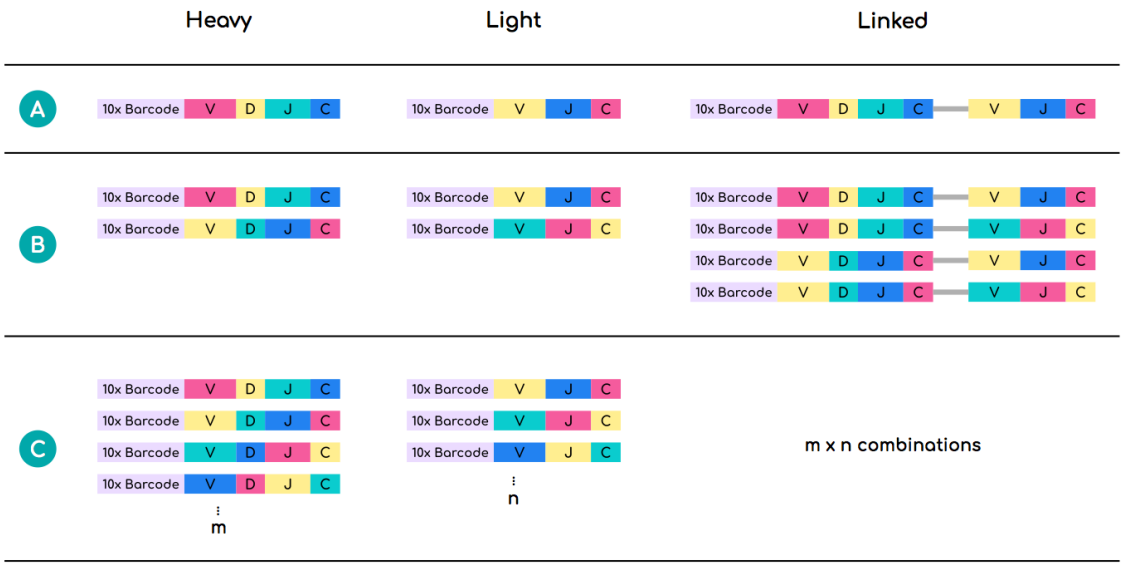


Figure 3.7: **Combinatorial linking of heavy and light chain sequences. (A)** In the case, when the same 10xGenomics barcode is only associated with one unique VH and one unique VL sequences, only one VH-VL combination is possible. **(B)** In some cases more than one VH and/or VL sequences share the same 10xGenomics barcode. If two VH and VL sequences share the same barcode, the total number of unique combinations would be four. **(C)** As the number of unique VH and VL sequences that share the same 10xGenomics barcode increases, the total number of potential VH-VL combination is equal to the number of unique VH times the number of unique VL sequences.

Study	Species	Disease	Vaccine	B-cell source	B-Cell Subset	Sequence #
[222]	mouse	SARS-COV-2	None	Lymph	Plasmablast	11,388 (28,648)
[223]	human	None	None	PBMC	RV+B-cells	771 (1,854)
[103]	mouse rat	None	OVA None	Lymph	Unsorted-B-cells	73,061 (296,891)
[121]	human	Tonsillitis Apnea	None	Tonsillectomy	Memory-B-cells Unsorted-B-cells	31,071 (84,376)
[224]	human	HIV	None	PBMC	Unsorted-B-cells	5,547 (15,672)

Table 3.2: **Summary of paired BCR repertoires in OAS.** Five major metadata descriptors (species, disease, vaccine, B-cell source, and B-cell type) of OAS deposited paired BCR studies are given in this table. The "Sequence #" column indicates the number of "1-to-1" linked VH-VL sequences in Data Units, with the total unlinked numbers in parentheses. RV+B-cells, Rhinovirus specific B-cell; OVA, Ovalbumin.

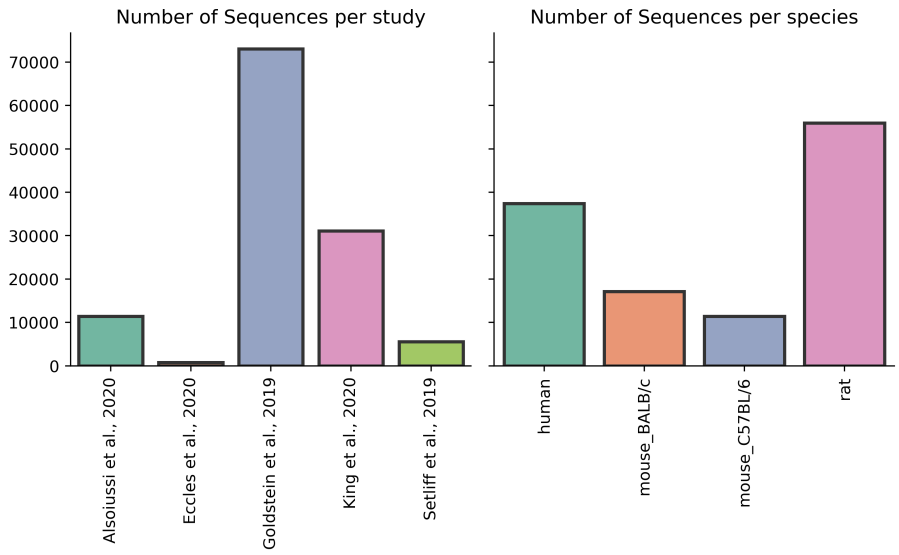


Figure 3.8: **Summary of BCR studies deposited in paired OAS.** So far, five paired-sequencing BCR studies have been added to OAS. These studies also hold BCR data from three different species: human, rat and mouse (mouse_BALB/c and mouse_C57BL/6). The Y-axis shows the number of VH/VL sequences in the linked OAS Data Units.

3.3.2.4 Unlinked Data Unit

Paired OAS incorporates 427,441 unlinked BCR sequences. This indicates that one unique VH and one unique VL ("1-to-1") pairing is only observed in ∼60% of all sequences ($121,838 * 2/427,441 = 0.59$). This is a surprisingly low number as it suggests that ∼40% of B-cells express more than one functional BCR. To investigate the distribution of the number of functional variable chain contigs produced by a single B-cell, we calculated frequencies of identical 10xGenomics barcode usages in all unlinked Data Units (Figure 3.9). We found that ∼35% of contigs shared the same 10xGenomics barcode with two or more other unique contigs. On rare occasions, the same barcode was found across more than eight unique contigs. This high degree of barcode sharedness could be indicative of a potential experimental error, where multiple B-cells were incorporated into a single emulsion droplet prior to 10xGenomics barcode ligation. The remaining 5% of contigs did not share their barcodes with any other contigs. This suggests that either the barcode ligation is not 100% efficient and/or CellRanger fails to assemble high quality contigs in a small number of identically barcoded reads.

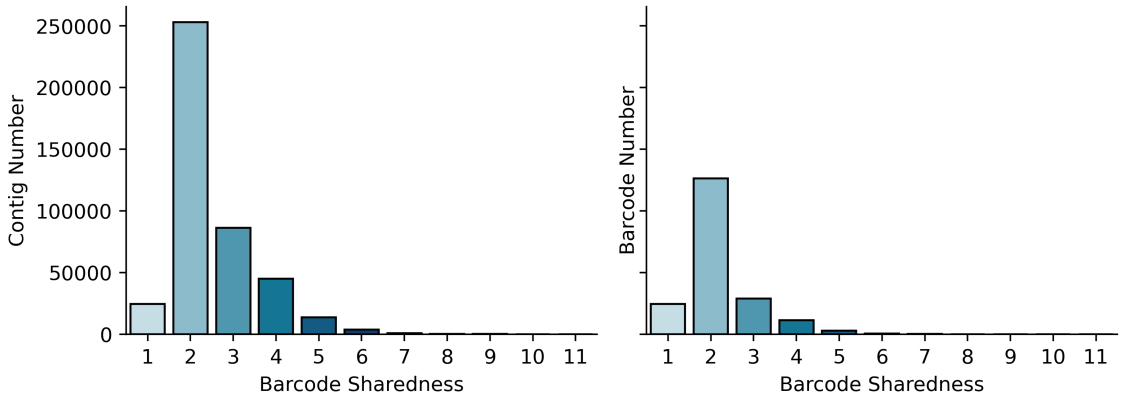


Figure 3.9: **Sharedness of 10xGenomics barcodes in unlinked paired OAS.** Barcode sharedness is defined as the number of times a 10xGenomics barcode is found across all unique contigs in a given BCR repertoire (X-axis). The Y-axis depicts the number of variable chain contigs **(left)** or unique 10xGenomics barcodes **(right)** in paired OAS. For instance, if a given barcode is only associated with a single contig, the barcode sharedness is equal to 1. **(left)** All contigs were binned and counted based on the barcode sharedness. **(right)** All contigs were binned based on the barcode sharedness and the number of unique barcodes at each sharedness value was calculated.

3.3.3 Comparison of contig assembly with SSAKE and CellRanger

Several 10xGenomics data processing pipelines have already been developed to study paired BCR repertoire diversities [103, 225]. However, it still remains unknown whether these pipelines result in the same BCR contigs. In order to select the best existing assembler for integration into the sequence processing pipeline of paired OAS, we scrutinised contigs generated by two tools (SSAKE [104] and 10xGenomics original CellRanger), which were previously reported to work with BCR 10xGenomics data [103, 226]. For CellRanger, we used the same sequence processing pipeline as described in paired OAS (Section 3.2.3.1). BCR sequence assembly with SSAKE was largely based on the Goldstein et al., [103] method, except for the contig annotation and quality filtering step. For this, we used the AIRR community recommended IgBlastn [117] instead of the Genentech-developed software, Absolve (`https://github.com/Genentech/Absolve`), as both tools have been shown to yield very similar BCR V(D)J gene annotations [103]. The SSAKE based BCR assembly pipeline is visualised in Figure 3.10. These two assembly pipelines were run on all raw BCR repertoires from Goldstein et al., [103].

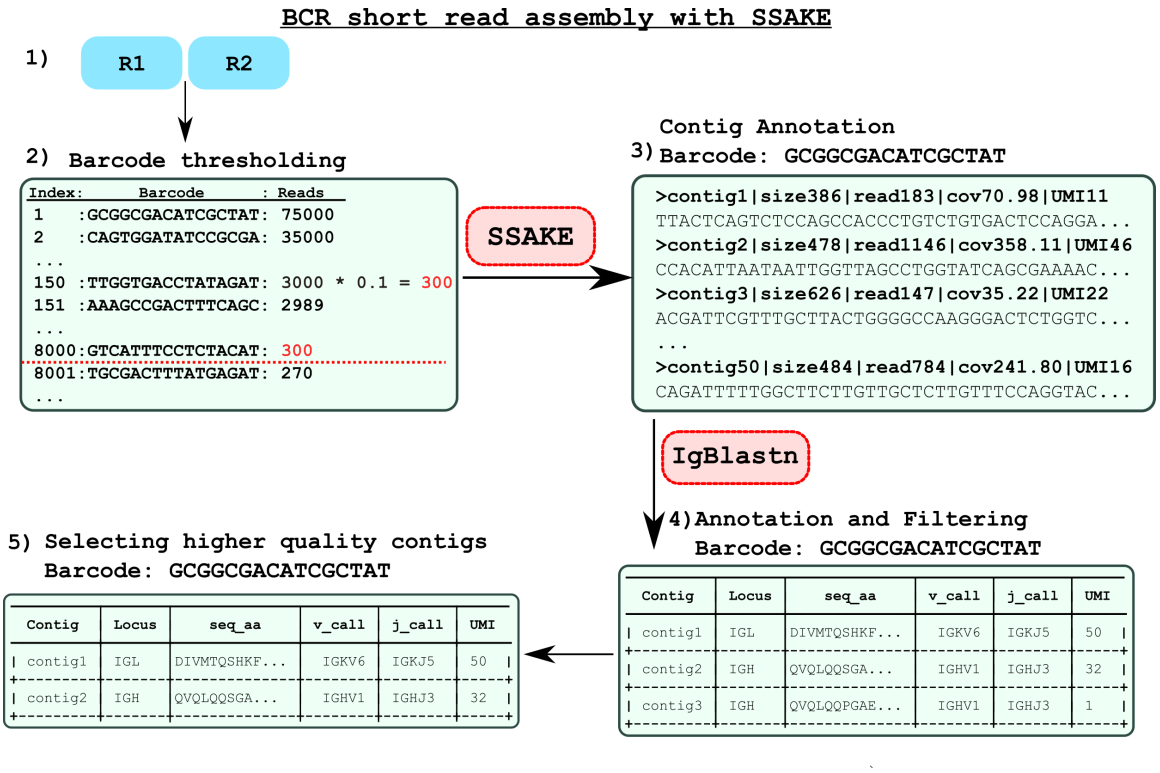


Figure 3.10: **BCR sequence assembly pipeline with SSAKE.** A five step BCR sequence assembly pipeline was developed to process short 10xGenomics reads with SSAKE. **(1)** Forward (R1) and reverse (R2) reads are downloaded from NCBI [197]. **(2)** Next, a read coverage cutoff is devised. This cutoff is based on the frequency of 10xGenomics barcodes as described in Goldstein et al., [103]. The barcodes are extracted from the R1 reads and counted. The read coverage cutoff is calculated as 10% of the frequency of the 150th most abundant barcode. Any reads whose barcode frequency lies below this cut-off are removed from R1 and R2. Next, the raw reads are written to separate files according to 10xGenomics barcodes. **(3)** SSAKE is run on each of these files separately. SSAKE usually generates tens of contigs, where most of those are non-functional BCR sequences. **(4)** These contigs are annotated and filtered with IgBlastn. We only allowed sequences that were productive, in the correct reading frame, had no stop codons, and were at least 100 amino acids long. **(5)** In a small number of cases, multiple contigs had the same combination of V and J genes. This could potentially be an assembly error, as multi-specific B-cells were shown to express different combinations of V and J genes [218]. Thus, if more than one identically barcoded contig had the same VJ combination, we only kept the contig with the highest number of UMIs (e.g. more mRNA transcripts that contributed to this contig assembly).

Whilst SSAKE consistently generated a larger number of unique contigs than CellRanger from the respective raw BCR reads (Figure 3.11A), the number of contigs created by these two pipelines were strongly correlated (Spearman's $\rho = 0.97$). This suggests that both approaches extract similar information from raw BCR reads. The SSAKE pipeline also outputted a higher number of unique 10xGenomics barcodes from raw reads (Figure 3.11B). The number of unique contigs was at least two times larger than the number of unique barcodes in each BCR repertoire. This suggests that an individual barcode, on average, is shared between two or more contigs within each assembled BCR repertoire (Figure 3.11).

A number of unique contigs associated with each 10xGenomics barcode can be used as a proxy to study assembly quality. For instance, in most cases we expect only two contigs to share the same barcode (*i.e.* one B-cell - one BCR). It was previously shown that some B-cells might display multi-immunoglobulin specificity [218]. Thus we also expect that a small number of the barcodes could be associated with more than two unique contigs. However, recording either a single contig or more than

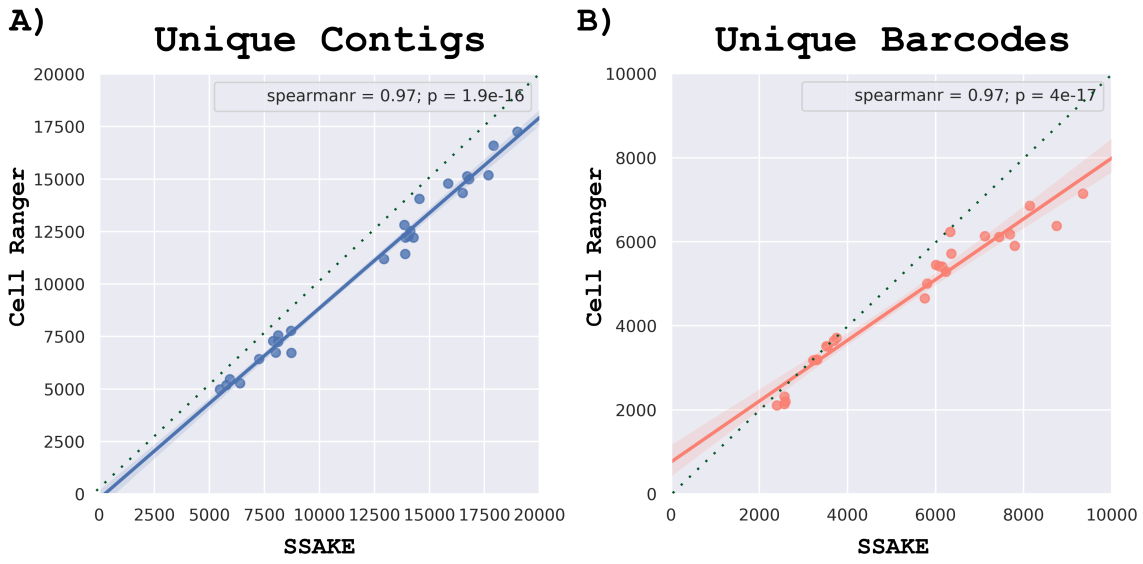


Figure 3.11: **Correlation between SSAKE and CellRanger outputs.** SSAKE and CellRanger assembly pipelines were run on 28 raw 10xGenomics BCR repertoires from Goldstein et al., [103]. The numbers of unique contigs **(A)** and unique 10xGenomics barcodes **(B)** outputted by these pipelines were compared. These numbers showed a very strong correlation (Spearman's $\rho = 0.97$). SSAKE consistently outputted both a higher number of unique contigs and unique barcodes. The dotted diagonal line indicates an equivalent output scenario. The X-axis shows SSAKE output numbers, and the Y-axis depicts CellRanger output numbers.

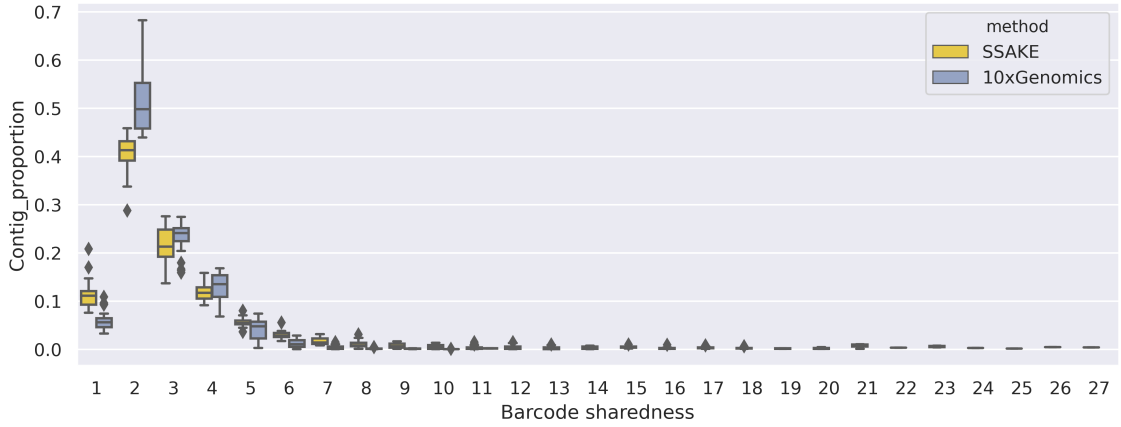


Figure 3.12: **10xGenomics barcode frequencies in SSAKE and CellRanger outputs In each assembled BCR repertoires.** 10xGenomics barcodes were binned based on the number of contigs associated with them ("Barcode sharedness"). The X-axis shows the barcode sharedness. The Y-axis shows the proportion of the total contig number in a given BCR repertoire. Colours represent different assembly pipelines.

ten contigs per barcode could be indicative of experimental and/or assembly error. Here, the barcode sharedness analysis was conducted to investigate the distribution of barcode association across contigs. We found that in the majority of cases only two unique contigs shared the same barcode (Figure 3.12). This percentage of "1-to-1" pairing was higher with CellRanger (∼52%) than SSAKE (∼41%). Furthermore, SSAKE also extracted a larger number of barcodes that only associated with a single contig than CellRanger (11.4% vs 5.9% respectively). Interestingly, both SSAKE and CellRanger contained a small number of contigs that shared their barcodes with more than ten other unique contigs. However, the number of such contigs was also higher in SSAKE assemblies (Figure 3.12).

Since SSAKE and CellRanger pipelines were run on the same raw BCR reads, we should expect to obtain a high level of assembled contig overlaps. Overlapping contigs were counted if their variable chain length, 10xGenomics barcode, V and J genes, and full amino acid sequences matched. We found that only ∼56% of SSAKE and ∼63% of CellRanger contigs were shared (Figure 3.13A). The lower overlap percentage in SSAKE-assembled BCR repertoires is concomitant with the higher number of total contigs. This surprisingly low degree of overlap between SSAKE and CellRanger assemblers could potentially be caused by a single mismatching or missing amino acid residue.

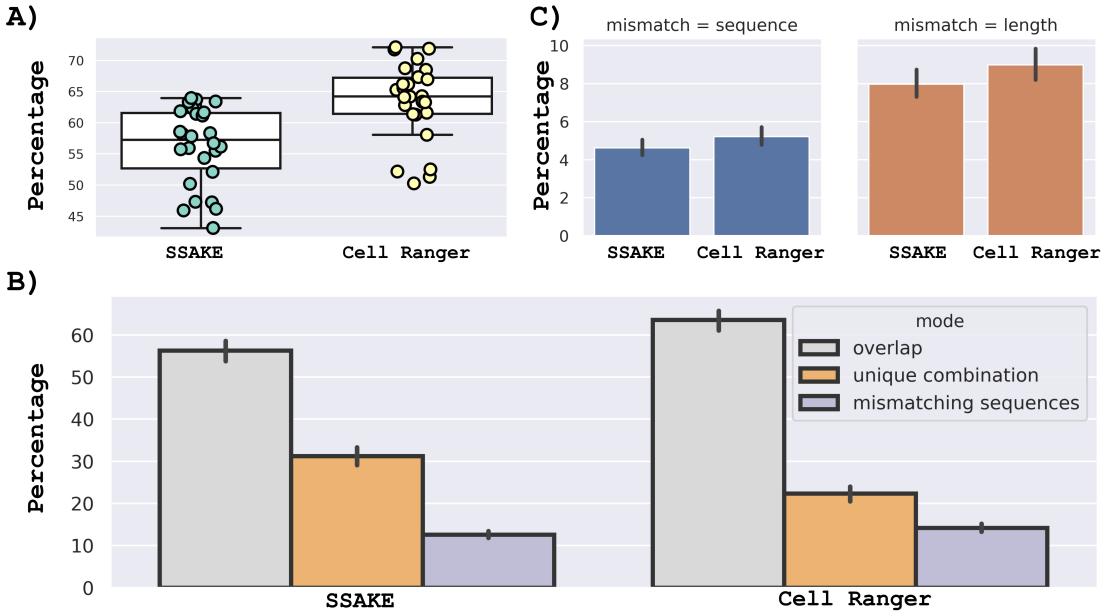


Figure 3.13: **Assembled contig overlap between SSAKE and CellRanger.** SSAKE and CellRanger were run on 28 BCR repertoires from Goldstein et al., [103]. Contig overlaps were then calculated between corresponding pairs of the assembled BCR repertoires. The overlap was counted when contigs shared the same 10xGenomics barcode, V and J genes and were 100% amino acid sequence identical. **(A)** The percentage of contig overlap between BCR repertoires assembled with the SSAKE and CellRanger pipelines. **(B)** The percentage of overlapping and non-overlapping contigs between SSAKE and CellRanger. The non-overlapping contigs were subdivided into two groups. The "Unique combination" group contains contigs that have unique combinations of 10xGenomics barcode, V and J genes that are not seen in the corresponding BCR repertoires assembled by the opposing pipeline. The "Mismatching sequences" group encompasses contigs that share identical 10xGenomics barcodes, V and J genes with contigs assembled by the opposing pipeline. However, these contigs were not 100% sequence identical either due to length or positional residue mismatches. **(C)** The "Mismatching sequences" group was further stratified into sequence and length mismatching contigs respectively. The percentage values were calculated based on the total number of contigs found in the corresponding BCR repertoires.

Contrasting the non-overlapping contigs between the SSAKE and CellRanger pipelines can reveal assembly errors. To assess the degree of dissonance in assembly outputs, we checked whether SSAKE and CellRanger captured the same VJ gene recombination events in the non-overlapping contigs. Surprisingly, we found that $\sim 31\%$ of all SSAKE and $\sim 22\%$ of all CellRanger contigs had non-matching V and J gene combinations at a given 10xGenomics barcode (Figure 3.13B). The remaining 12.5% of SSAKE and 14.2% of CellRanger contigs had matching 10xGenomics barcodes, V and J genes, but were not 100% amino acid sequence identical (Figure 3.13B). Since these contigs originated from the same VJ gene recombination event, we used the Needleman-Wunsch pairwise sequence alignment [227] to identify any mismatching and missing residues. The analysis of the pairwise alignments revealed that $\sim$5% of all contigs had mismatching residues and $\sim$8.5% differed in length. Although it is hard to judge which of two residue mismatching contigs is correct, it is straightforward to identify a better assembly in the length-mismatching contigs. In more than 99% of the length-mismatched contigs, CellRanger generated complete variable chain assemblies, whilst SSAKE assemblies missed a few 5' and/or 3' residues.

These results suggest that CellRanger outputs fewer contigs than SSAKE, but these contigs have higher assembly quality. Therefore, we integrated CellRanger into the sequence processing pipeline of paired OAS.

3.3.4 Looking for therapeutic antibody sequences in OAS

In the last section of this chapter, we describe an application of unpaired OAS in drug discovery. We studied sequence overlaps between all clinical-stage therapeutic (CST) antibodies and BCR repertoires deposited in the OAS resource (March 2019). Understanding of the relationship between successful drug design and the overlap with natural BCR repertoires can be leveraged to make more informed decisions in the antibody discovery campaigns. I was a co-author of this research, where I was responsible for formatting OAS datasets for the analysis, validating the results, contributing to the experiment design and editing the manuscript. This research was published in mAbs in 2019 [125].

3.3.4.1 CST antibody sequence alignment to natural BCR repertoires

We used a set of 242 CST antibodies [167], all of which have completed Phase 1 clinical trials. We separately aligned the CST VH or VL domains, combination of the three complementarity-determining regions (CDRs) from VH or VL and CDR-H3s to

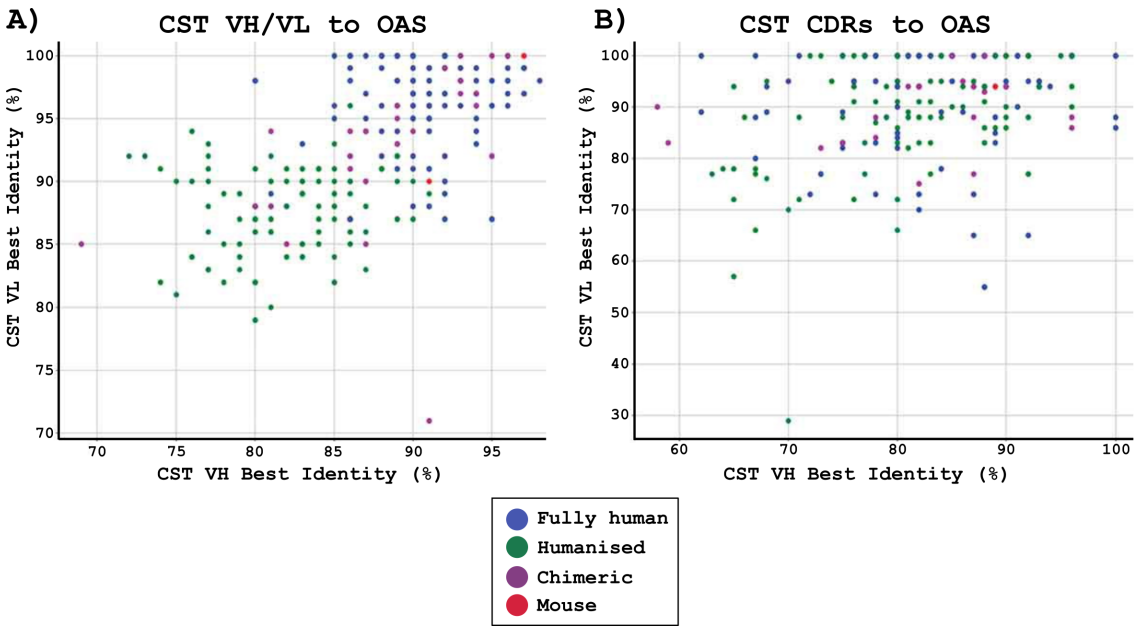


Figure 3.14: **Highest sequence identity matches between 242 CSTs and natural BCR repertoires from OAS**. Heavy and light chain (**A**) variable regions or (**B**) all three CDR loops of the CST sequences were aligned to the respective BCR sequence regions in OAS. Fully human CST sequences are denoted by blue dots, humanised by green, chimeric by magenta and mouse in red. In a small number of cases where CSTs had the same identity values and different antibody type, we report the antibody type by majority vote of proximal CSTs.

all the sequences in OAS (see Method Section 3.2.5). We performed the search across all organisms, individuals and immune states to be comprehensive and to reflect the myriad of CST types: fully human, humanised, chimeric or fully mouse [228]. The individual identities of the CSTs with respect to the best match from OAS are given in Figure 3.14 and their distributions are plotted in Figure 3.15.

The best sequence identity matches of CST variable regions to natural BCR repertoire deposited in OAS are given in Figure 3.14A. Ninety (or 37.1%) CST VH chains have matched to OAS with sequence identity (seqID) $\geq$ 90%; 18 (or 7.4%) CSTs with seqID $\geq$ 95%. We find 158 (or 65.2%) therapeutic VL chains with $\geq$ 90% seqID to an OAS sequence, with 96 (or 39.7%) $\geq$ 95% seqID, and 28 (or 11.5%) with 100% seqID. For 16 (or 6.6%) of the CSTs, we find both VH and VL chain matches $\geq$ 95% seqID. In the most extreme case, enfortumab, we were able to find both VH and VL chain matches of 98% seqID (the differences are H38:N-S, H88:S-Y, L37:G-S, L52:F-L, where the first amino acid comes from enfortumab and the second from an unpaired OAS sequence).

The largest discrepancy between the CSTs and OAS BCR sequences is typically

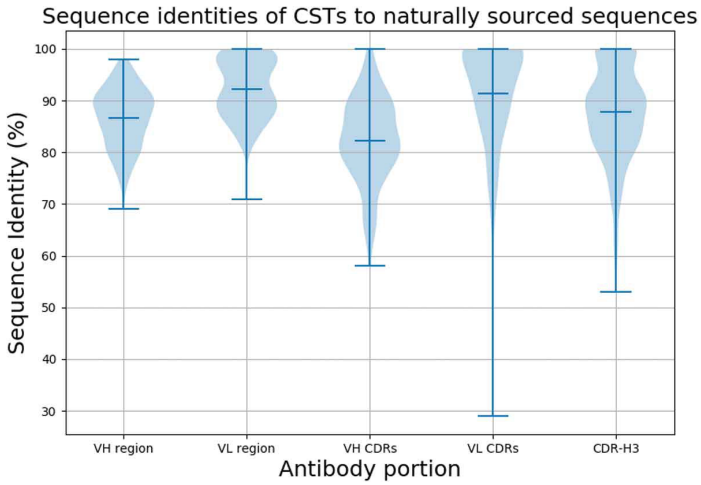


Figure 3.15: **Distribution of the highest sequence identity matches of CSTs to naturally observed antibody sequences in OAS.** The violin plots show the distribution of sequence identities of the variable heavy (VH) and light (VL) chains, heavy and light CDR regions and CDR-H3 of CSTs to the highest matches in OAS.

concentrated in the CDR regions that are the main contributors to cognate antigen complementarity [21]. It still remains unclear the extent to which the highly mutable CDR loops of engineered therapeutics differ from those that are expressed naturally by the adaptive immune system. We searched for the best CST matches to the CDR regions in OAS, as they form the majority of the antibody paratope. The sequence identity was calculated across the entire CDR region. The identity calculations were only allowed if all three CDR lengths matched between a given CST and BCR sequence pair. The search was performed using the IMGT-defined [36] CDR triplets from the VH or VL chain, disregarding the framework regions (*i.e.* we concatenated the sequences of the CDRH1-3 loops, or CDRL1-3 loops) (Figure 3.14 and Figure 3.15). We found 46 (or 19.0%) of CST VH chain CDR triplets to have matches to an OAS CDR triplet with $\geq$ 90% seqID, 15 (or 6.1%) with $\geq$ 95% seqID and 4 (or 1.6%) with 100% seqID. There were 156 (or 64.4%) CST VL CDR triplets with $\geq$ 90% seqID to an OAS CDR triplet, with 110 (or 45.4%) $\geq$ 95% seqID, and 90 (or 37.1%) with 100% seqID. Strikingly, for two CSTs (obiltoxaximab and zanolimumab), we found OAS sequences where all three VH and VL chain CDRs were identical. Obiltoxaximab and zanolimumab are engineered to bind to *Bacillus anthracis* spores and the human CD4 receptor.

Of the six CDRs, CDR-H3 is the most sequence and structurally diverse [3, 169]. Due to its key role in binding, it is subjected to extensive *in vitro* antibody engineering [229, 230]. We checked how likely it is to find CST-derived CDR-H3s in naturally

observed sequences. To assess this, we searched for the best CST CDR-H3 matches in OAS, regardless of the framework region and remaining CDRs (Figure 3.15). Of our 242 CST CDR-H3s, we found 54 (or 22%) perfect matches in 1 billion of OAS deposited sequences. The perfect matches tended to be for shorter CDR-H3s, but some longer loops with perfect matches were also found (Figure 3.16). Twenty nine perfect matches were found in just one large BCR repertoire sequencing study of Briney et al., [4]. This study sampled the diversity of the human BCR gene repertoires of ten individuals at an unprecedented depth. The large proportions of matches from this single study suggest that substantial CDR-H3 diversity can be found in a very limited number of individuals. Forty seven (or 19.4%) perfect matches were found in OAS datasets other than that of Briney et al., showing that some artificial CDR-H3 sequences can be independently observed across multiple natural BCR repertoires. Twenty two (or 9%) CDR-H3 matches were found in both Briney et al., data and other BCR studies in OAS (Figure 3.16). These 22 shared CST sequences come from 9 humanised and 13 fully human antibodies. The 54 perfect CDR-H3 matches were distributed among all antibody types, with 23 humanised, 22 fully human, 8 chimeric and 1 mouse (21.9%, 22.0%, 22.8% and 33.3% of each category, respectively). These results show that, despite the large theoretical sequence space accessible to the CDR-H3 region [3, 4], therapeutically exploitable CDR-H3 loops are found in just ∼960 million VH chain sequences from 60 BCR studies. This convergence, coupled with the fact that CDR-H3 loops often mediate antibody specificity [231] and binding affinity, could suggest intrinsically driven biases in antigen recognition [5], independent of artificial discovery methods (Appendix Table C.2).

3.3.4.2 CST matches to OAS by antibody type

The discovery platform/antibody type has a direct influence on the percentage of the CST VH/VL match to OAS sequences (Figure 3.14). A closer analysis of the CST types showed that antibodies produced *via* more artificial protocols such as humanisation have lower variable region sequence identities to sequences in OAS from those of fully human molecules (Figure 3.17). For the majority of the fully human sequences we find matches of 90% seqID or better, whereas matches to the majority of humanised CSTs fall below 90% seqID. Chimeric antibodies appear to have seqID values intermediate between the two classes.

The CST antibody type also reflects the organism that produced the best BCR repertoire seqID match. Of the 100 fully human CSTs, the 90 most similar VH chains, all 100 most similar VL chains, and 55 most similar CDR-H3 loops come

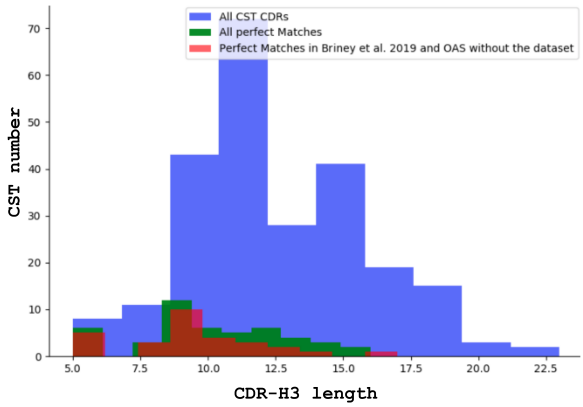


Figure 3.16: **Distribution of CDR-H3 lengths of CSTs perfectly matched OAS.** CST CDR-H3 lengths (blue) are overlaid on the lengths of the perfect matches (green). The red histogram shows the length distribution of CST CDR-H3 that are found in both Briney et al., [4] and all other OAS deposited BCR studies.

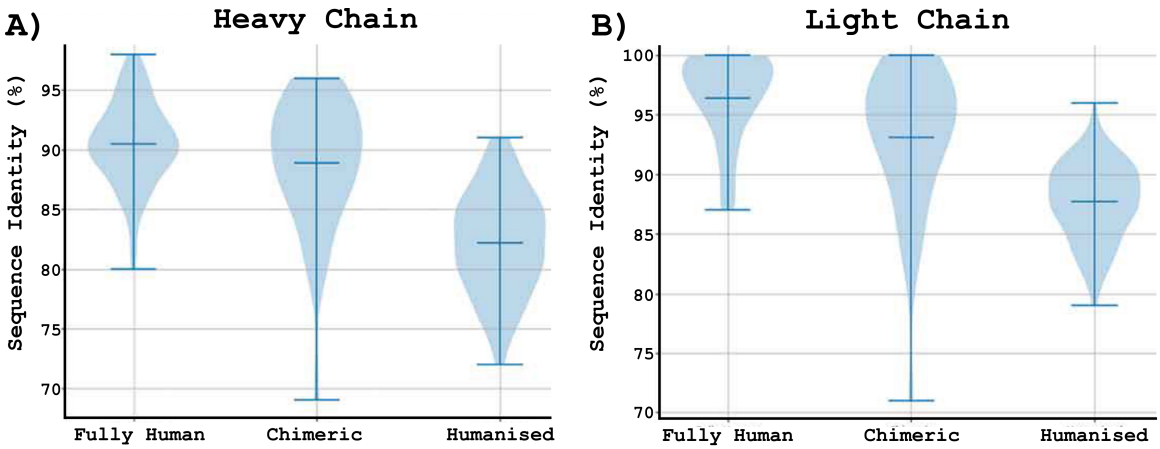


Figure 3.17: **Distribution of the highest sequence identity matches of VH/VL domains of CSTs to natural BCR repertoires from OAS stratified by CST antibody type.** CST **(A)** VH chain and **(B)** VL chain sequence identities to OAS stratified by fully human, chimeric and humanised antibody types. Only three mouse antibodies were present in our CST library. Hence, these murine CSTs were excluded from this analysis, as the sample size was deemed insufficient for distribution interrogation.

from human BCR studies in OAS. Of the 105 humanised antibodies, 82 (or 78.0%) of VH chains, and 79 (or 75.2%) of VL chains found closest matches in human BCR studies, whilst 71 (or 67.6%) of the best CDR-H3s matches were identified in mouse BCR studies. This further reflects the dominance of CDR-H3 in binding, as biotech companies often graft this loop from binding mouse antibodies to transfer specificity

and binding affinity. It also suggests that mining a dataset such as OAS could provide a more accurate measure of antibody 'humanness' than our current metrics [232].

3.4 Discussion

In this study, we describe the OAS database, a unified repository to facilitate large-scale data mining of annotated BCR repertoires in both their amino acid and nucleotide forms. The need for a well-established repository of BCR repertoire data prompted us to perform a combination of literature search and manual curation of the datasets to organise the data into OAS. The current lack of widely adopted deposition standards hampers automatic updating of OAS, as datasets in which we find a large number of BCR studies still require manual curation to perform metadata annotation consistently. Hopefully, continued efforts such as those by the AIRR community will result in standardisation of Ig-seq outputs and will further streamline deposition procedures facilitating large-scale data mining of BCR repertoires [196]. Devising unified BCR repertoire repositories is challenging because of both the size of the datasets as well as the diverse data descriptors and analytical pipelines desired by bioinformaticians, wet lab scientists, and clinicians [127, 233].

To our knowledge, OAS was the first organised collection of a large body of BCR sequencing outputs that is designed for continuous expansion as more and more BCR data become available. The basic data files are stored in an efficiently compressed format and are searchable by light-weight metadata entries. As of 2020, OAS complies with the AIRR community standards to benefit from cross-fertilisation of sequence annotation methods in both the AIRR and structural biology fields. To allow comparative bioinformatics analyses across different subsets of BCR repertoires, we have annotated the datasets by commonly-used metadata descriptions, such as organism, isotype, B-cell type and source, and the immune state of B-cell donors. To facilitate research about particular BCR sequences or regions, we make full IMGT-numbered, high-quality amino acid sequences available together with a rich sequence annotation by IgBlastn.

These BCR data should aid in-depth comparative analyses across different studies to discern the commonalities observed between independent samples as well as directing Ig-seq experiments on not-yet interrogated BCR repertoires. Revealing shared preferences can be invaluable in identifying the portions of the theoretically allowed BCR space that are strategically used to start immune responses [4, 102, 153]. Furthermore, such comparative studies can offer a way of deconvoluting the various dif-

ference of immune repertoires, such as differences between diversity of isotypes [234] or organisms [235]. Charting the differences between repertoires of human/mouse is of particular interest for engineering better humanised biotherapeutics [132, 153].

Paired BCR chain sequence information provides an enhanced view on BCR biology [103, 105]. Recent breakthroughs in sequencing technology allow for the complete delineation of BCR variable domains [236]. As of October 2020, OAS already includes paired high quality sequences from five independent studies. As the paired data provides an enhanced view of the BCR paratope residues, we expect an exponential increase in availability of this kind of data in the upcoming years.

Currently 10xGenomics technology is the only source of the full length paired BCR repertoires in OAS. We tested two separate 10xGenomics assembly pipeline and found that 10xGenomics original CellRanger provides the highest contig assembly quality. Thus, CellRanger was integrated into the paired OAS sequence processing pipeline.

Recently, an alternative paired BCR technology (CelliGo) was published [237]. As OAS is set for continuous expansion, a bespoke sequence processing pipeline will need to be developed and validated to incorporate CelliGo sequencing outputs.

Beyond identifying broad commonalities across repertoires, data mining Ig-seq outputs provides novel avenues for designing better antibody based therapeutics. The plethora of currently available BCR data offers a glimpse at a set of sequences that should be able to fold and function in an organism. Aligning therapeutic candidates to sequences in BCR repertoires can reveal mutational choices that might be naturally acceptable, hence providing shortcuts for antibody engineering such as humanisation [207, 238, 239]. Further, contrasting the naturally observed BCR sequences with therapeutic antibodies can offer insights as to the naturally favoured biophysical properties of these molecules [167, 228]. All these immediate applications rely on the availability of well-structured datasets that can offer a unified point of reference for bioinformatics analyses. We hope OAS will facilitate data mining BCR repertoires, help identify strategic preferences of our immune systems, and will ultimately improve how we engineer antibodies into better therapeutics.

As an example of one of the immediate applications of OAS in drug discovery, we performed a systematic search of Clinical-Stage Thearapeutic (CST) antibody sequences in naturally observed BCR sequences from unpaired OAS (March 2019 version). We found that despite the theoretically large primary sequence diversity accessible to antibodies ($\sim 10^{13}$) [3, 4], non-trivial convergences between artificially developed CSTs and naturally observed BCR repertoire sequences were recorded.

The closest repertoire matches to CSTs were sourced from 48 of the 60 independent studies available in OAS (as of March 2019), indicating that finding a close match to at least one CST is likely in most BCR repertoire studies. We identified 16 CSTs that shared more than 95% VH/VL sequence identity with at least one sequence deposited in OAS. Correlating the sequence identity matches with immunogenicity profiles of corresponding CSTs can help identify sequence/structural features to engineer safer biotherapeutics. CSTs can also be employed as baits to study structural/sequence convergence in BCR repertoires which were pre-immunised with the same antigen. This will advance our understanding of molecular mechanisms that dictate BCR maturation and selection.

Our sequence matching to the CSTs analysis was limited to the unpaired OAS version as of March 2019. Since then the size of OAS almost doubled. This prompts a repeat of the analysis as more hits with even higher sequence identity matches are expected to be identified.

The incorporation of paired BCR sequences into OAS now permits a more comprehensive analysis of paratope overlaps between CSTs and naturally observed BCR sequences. At the moment the number of the paired sequences remains low ($\sim 120,000$) covering only a minuscule portion of the sequence space accessible to BCRs. Hence, as more paired BCR datasets become publicly available, the statistical power of our overlap analysis will increase to draw even stronger conclusions about observed convergences across the entire paratope and VH/VL regions.

To democratise access to BCR repertoire data, the OAS resource does not require any login credentials. Google analytics is employed to study the resource use traffic and impact.

In the next thesis chapter, we will describe the development of a novel analysis technique which we applied to study structural dynamics of BCR CDR loops. We used two studies from OAS [23, 94] as the source of BCR data for this analysis.

4

Structural Diversity of B-Cell Receptor Repertoires along the B-cell Differentiation Axis in Humans and Mice

Contents

This chapter contains reproduced material from the following publications:

1. **Kovaltsuk, A.**, Raybould, M.I.J., Wong, W.K., Marks, C., Kelm, S., Snowden, J., Trück, J. & Deane, C.M. (2020) Structural Diversity of B-cell Receptor Repertoires along the B-cell Differentiation Axis in Humans and Mice, PLoS Computational Biology, 16(2):e1007636

2. Ghraichy, M., Galson, J.D., **Kovaltsuk, A.**, von Niederhäusern, V., Schmid, J.M., Miho, E., Kelly, D.F., Deane, C.M. & Trück, J. (2020) Maturation of Naïve and Antigen-experienced B-cell Receptor Repertoires with Age, Frontiers in Immunology, 11:1734

I carried out all the work described in this chapter unless noted otherwise.

4.1 Introduction

Next-generation sequencing of immunoglobulin genes (Ig-seq) has become an essential technique in immunology [240, 241]. Most Ig-seq data analysis tools work within the remit of B-cell receptor (BCR) primary sequence information [86, 170]. These rapid methods of measuring BCR repertoire diversity are highly scalable, an important property as BCR datasets become ever larger and more numerous [3, 4, 43]. However, the decision to avoid BCR paratope structural descriptors could lead to inaccuracies [48, 101, 169], as it is known that similar sequences can have markedly different epitope complementarity and vice versa [170]. Therefore, a computationally-efficient structure-based BCR repertoire method should augment current Ig-seq analysis pipelines to deliver a clearer understanding of the process of BCR development.

One of the first structural analyses of BCR repertoire data was that of DeKosky *et al.,* [101]. They demonstrated that antibody models from paired-chain naive and memory BCR repertoires displayed different physicochemical properties. However, their analysis was limited to 2,000 antibody models from three B-cell donors [101]. Most publicly-available BCR repertoires are unpaired, only covering either the heavy (VH) or light (VL) variable domain [43] precluding the generation of refined antibody models. Raybould *et al.,* [102] developed a pipeline to combinatorially pair VH and VL sequences based on their structural compatibility. This approach yields a large number of combinations, which hinders our ability to model complete BCR repertoires. Krawczyk *et al.,* [169] showed that it was possible to annotate complete unpaired BCR repertoires with structural information by mapping loop sequences individually onto crystallographically-solved antibody structures.

In the first part of this chapter, we present a novel pipeline (SAAB+) which expands functionalities of the previously published approach by Krawczyk *et al.,* [169] to investigate structural BCR repertoire diversities along the B-cell differentiation axis in humans and mice. We show that structurally annotating BCR repertoires yields unprecedented insights into both the structural predetermination and dynamics of the adaptive immune response. By approximating BCR repertoire structures with rapid homology modelling techniques, we find that different B-cell types can be distinguished by their usage of CDR loop structures. Our analysis reveals that BCR repertoires of naive B-cells tend to contain a conserved "public" CDR structure profile, whilst those of more differentiated B-cell types become more personalised due to the proprietary SHM events.

In the second part of this chapter, we investigate structural maturation using unsorted BCR repertoires across different age groups in humans using the SAAB+ pipeline. We find an increasingly higher degree of structural divergence from germline encoded structures in older participants (between 6 months and 50 years), which sheds light on individual's complex history of antigenic stimulations. Interestingly, antigen-unexperienced B-cells show similar structural features from birth to adulthood highlighting BCR repertoire fitness to bind to a broad range of yet unseen antigens. Our results provide crucial information about the structural changes in antibody CDRs during B-cell differentiation, with a plethora of prospective applications in immunodiagnostics and rational immunotherapeutic engineering.

4.2 Methods

4.2.1 Data

4.2.1.1 Structural diversity along the maturation axis in humans and mice

Human BCR repertoire data from Galson *et al.,* [23] and mouse (C57BL/6 inbred strain) BCR data from Greiff *et al.,* [94] were used. Galson *et al.,* ("human") is a longitudinal vaccination study across nine healthy human B-cell donors, in which the heavy chain of naive, marginal zone, memory, and plasma B-cell types were interrogated [23]. Greiff *et al.,* ("mouse") is a high depth sequencing study of the murine adaptive immune system in response to antigenic stimulation, containing heavy chain BCR repertoires from pre, naive and plasma B-cells [94]. Both studies used FACS to sort B-cells into subpopulations according to their differentiation stages.

The BCR amino acid sequences were downloaded from the Observed Antibody Space (OAS) [43] resource, retaining their Data Unit information. Each Data Unit is a sequencing sample from a single B-cell donor with a defined combination of B-cell type and isotype information, and contains sequences that are IMGT-numbered [36] and filtered for antibody structural viability. Henceforth, OAS Data Units will be referred to as B-cell receptor (BCR) repertoires.

To investigate structural changes along the B-cell differentiation axis, BCR repertoires with defined B-cell type and isotype information were downloaded. Only IGHG and IGHM sequences were considered as these were the most abundant. The total number of BCR repertoires in the human and mouse data were 85 and 82 respectively.

These human and mouse datasets were employed in the first part of the results section.

4.2.1.2 Structural diversity across different age groups in humans

To study structural BCR repertoire diversities across different ages in humans, we used B-cell type unsorted BCR repertoires from 53 healthy human blood donors whose ages ranged between 6 months and 50 years ("Healthy (Ghr.)") [107]. The BCR repertoires were directly sent to us by the corresponding authors. These Healthy (Ghr.) repertoires were also used in Chapter 5 (Section 5.3.3).

Since Unique Molecular Identifiers (UMI) DNA barcodes were attached to BCR cDNA, we therefore undertook a preprocessing pipeline which consisted of pRESTO package to group sequences tagged with the same UMIs in order to correct for any potential PCR and sequencing biases. Within each cluster, a consensus sequence was devised based on sequence redundancy [112]. All remaining cluster sequences were collapsed to match the respective cluster consensus sequences. The bias corrected BCR repertoire sequences were annotated with Change-O to comply with the recommended standards of the AIRR community [113, 117]. Stampy was then used to identify BCR subtype information based on the constant domain sequence [242]. Amino acid sequences of the processed BCR repertoires were numbered with the IMGT scheme [36] and filtered for structural viability using "ANARCI parsing" [34] as per the first steps of the ABOSS algorithm [44]. The full ABOSS pipeline was not run as the number of processed sequences was low for the optimal ABOSS performance. Sequences were filtered out that (i) could not be aligned to the human Hidden Markov Model (HMM) profile of an IMGT germline (ii) had a J gene sequence identity of <50% to a human IMGT germline or (iii) contained non-amino acid entries in CDRs. Since the primer masking step in pRESTO [112] can remove the first

framework region and positions 127 and 128 of some sequences, ANARCI parsing was customized to account for these exceptions. To retain as many sequences as possible for structural annotation, we substituted undetermined residues in the framework region with the residues from their respective parent germline genes. Finally, each BCR repertoire was split into two groups based on the isotype information: "MD" (IGHM, IGHD) and "AEG" (IGHA, IGHE, IGHG).

The Healthy (Ghr.) datasets were used in the final part of the results section.

4.2.2 SAAB+ pipeline

To annotate the BCR repertoire data with structural information, we developed SAAB+, a customised version of our previously-published SAAB pipeline [169]. In original SAAB, FREAD homology modelling is employed to predict shapes for all three CDR loops in unpaired BCR sequences [27]. FREAD outputs a Protein Data Bank (PDB) code of a crystallographically-solved CDR structure (template) from its pre-built library that has the highest Environment-specific Substitution Score (ESS) (*i.e.* structural homology) [146] to the target loop, and favourable anchor region orientations with the selected framework structure (see Introduction, Section 1.3.1.1 for more details).

The major differences/improvements of SAAB+ over SAAB are the following. SAAB+ scrutinises each antibody sequence for structural viability. In this step, sequences are passed through multiple structural filters (alignment to Hidden Markov Model (HMM) profiles, indel and conserved residue identification, chimeric sequence removal, presence of all 3 CDR loops according to the IMGT numbering [36] definition) [43, 44].

Non-CDR-H3 loops adopt a limited number of structural configurations, known as canonical classes (see Introduction, section 1.1.4) [37, 47, 48]. Hence, SAAB+ employs SCALOP [49] which can rapidly and accurately identify canonical loop classes from sequence alone. This gives a significant boost to the analysis rate without compromising prediction accuracy. SCALOP performance is described in Introduction (Section 1.3.2.2). The number of canonical classes is known to change as more antibody structures become available. SCALOP auto-updates its canonical class database every month. The June 2019 SCALOP database was used in this study.

In addition to CDR-H3 shape prediction, the SAAB+ pipeline also performs structural clustering of CDR-H3 templates. To find structural templates with similar CDR-H3 loop shapes (analogous to canonical loop shapes), SAAB+ structurally clusters them based on their backbone atom RMSD values (see Section 4.2.4 in Methods).

SAAB+ outputs a tab delimited text file that can easily be merged with the current AIRR-seq standard format file [116].

4.2.3 Validating FREAD for use in SAAB+

Accurately modelling all the CDR-H3s in an BCR repertoire dataset is challenging, owing to the vastness of structural space accessible to these loops [52, 55, 148], relative to the small number of publicly-available crystallographically-solved antibodies (many of which are highly sequence redundant) [134]. In addition, structurally-solved antibodies have a CDR-H3 length distribution and sequence diversity that is different from natural BCR data (Figure 4.3). We tested the performance of FREAD on the BCR data and, at the parameters used, the expected average RMSD of FREAD CDR-H3 template predictions for both human and mouse data is 2.5 Å (see Section 4.3.2). This is in line with current state-of-the-art CDR-H3 modelling software tools that were benchmarked in the antibody modelling assessment (mean RMSD of 2.8 Å) [139, 161]. In a similar manner to DeKosky *et al.,* [101], we limited our CDR-H3 analysis to loop lengths of 16 amino acids or shorter, as far fewer structures with longer CDR-H3 loops are available and longer loops have increased structural freedom. We also excluded CDR-H3 loops shorter than five amino acids from our analysis, as only three CDR-H3 templates covered these lengths. FREAD templates were downloaded from SAbDab (14th November 2018) [134], and consisted of all X-ray crystal structures of antibodies with a resolution better than 2.9 Å.

4.2.4 CDR-H3 clustering

To identify similar CDR-H3 loop structures, we used the Dynamic Time Warping (DTW) algorithm [48] to cluster FREAD CDR-H3 templates by backbone RMSD. The Dynamic Time Warping (DTW) algorithm enables rapid backbone atom RMSD calculations between length-mismatched CDR loops. The CDR-H3 loops found within 0.6 Å RMSD were placed in the same cluster, reducing our 2,943 FREAD CDR-H3 templates to 1,169 CDR-H3 structural cluster. Only length matched CDR-H3 templates were put into the same cluster at this RMSD cut-off.

4.2.5 Filtering BCR repertoires

The mouse and human BCR repertoires used in the first part of this chapter were generated using different experimental protocols. The data downloaded from the OAS resource is already pre-filtered for antibody structural viability, however, some

BCR repertoires might still have sequence biases. As PCR sequencing can lead to variable amplicon amplification, we removed any BCR repertoire if its two most redundant CDR-H3 clusters contained more than 80% of all repertoire sequences. We also discarded any BCR repertoire that contained fewer than 10,000 sequences with predicted CDR-H3 structures — this cut-off was selected to allow for adequate sampling of CDR-H3 template usages, whilst retaining as many BCR repertoires as possible. This reduced the number of repertoires for all subsequent structural analysis to 81 (human) and 73 (mouse). CDR-H1 and CDR-H2 loops were not taken into account in determining BCR repertoire quality, since canonical class coverage was ∼95% and ∼99% for the human and mouse data respectively (Table 4.4).

4.2.6 Patterns of CDR-H3 cluster usage

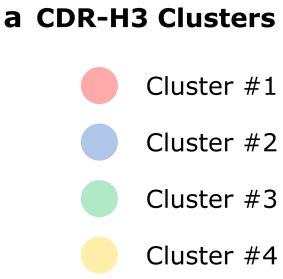
We analysed the pattern of CDR-H3 cluster frequencies in the human and mouse data, to identify clusters whose usages were over-represented (Structural Stems), random (Randomly-Used) and under-represented (Under-Represented) within a given B-cell type.

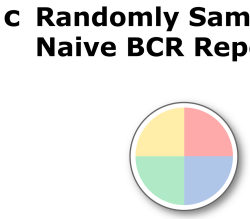
The structurally-annotated human and mouse BCR datasets were split into individual groups based on unique B-cell type and isotype combinations (e.g. "Naive_IGHM", "Memory_IGHG"). Within these groups, we calculated the CDR-H3 length distributions and the proportion modellable by FREAD for each CDR-H3 length. Next, we randomly selected CDR-H3 templates from our FREAD library (with replacement) according to these distributions, to generate a randomised dataset for each BCR repertoire. Sampling was performed across the set of FREAD templates already present in each BCR repertoire (Figure 4.1). The randomised dataset sizes were set to one million sequences and the total number of randomised datasets was matched to the number of the BCR repertoires within the corresponding groups (Table 4.1).

A one-sided Mann-Whitney rank test ($p = 0.05$) was performed on the relative usage of each CDR-H3 cluster in the grouped BCR repertoires and the corresponding randomised datasets, to categorize them as Structural Stem, Random-Usage or Under-Represented CDR-H3 clusters (Figure 4.1).

4.2.7 Shannon Entropy calculation to investigate the structural diversity of CDR-H3 clusters

$$Entropy = -\sum_{i=1}^{n} p_i log p_i \tag{4.1}$$





d Cluster Usage

Cluster ID	RS	Observed				Classification
#1	25	41	33	30	45	Structural Stem
#2	25	30	29	30	37	Structural Stem
#3	25	27	27	16	8	Random-Usage
#4	25	2	11	24	10	Under-Represented

Figure 4.1: **Pattern of CDR-H3 cluster usage within a specific B-cell type.** A schematic representation of how we grouped CDR-H3 clusters based on their pattern of usage. **(a)** In this mock example, only four CDR-H3 clusters are found in **(b)** four naive BCR repertoires. **(c)** In the case of random uniform sampling, each of these clusters would constitute approximately 25% of a simulated BCR repertoire. **(d)** Structural Stems are defined as CDR-H3 clusters, which are over-represented across BCR repertoires when compared to random cluster usage. Under-represented are clusters that are under-represented across repertoires. CDR-H3 clusters, which usages are not significantly different from random sampling, were termed Random-Usage.

Equation 4.1 shows a Shannon Entropy formula where n is equal to the number of represented CDR-H3 clusters ($0 < n \leq 1,169$) and p_i is the proportion of individual cluster usage in a given BCR repertoire. High entropy demonstrates a high diversity of CDR-H3 structures, whilst low entropy indicates the over-representation of one or more CDR-H3s. To account for the decreasing number of represented CDR-H3 structures along the B-cell differentiation axis, we calculated the proportion of theoretical maximum entropy for each BCR repertoire to yield a normalised estimate of the diversity of CDR-H3 clusters used (Figure 4.14) (Equation 4.2).

$$\begin{aligned} Entropy_{normalised} &= \frac{\sum_{i=1}^{n} p_i log p_i}{\sum_{i=1}^{n} (\frac{1}{n}) log(\frac{1}{n})} \\ &\Rightarrow \frac{\sum_{i=1}^{n} p_i log p_i}{log(\frac{1}{n})} \\ &\Rightarrow \frac{Entropy}{log(\frac{1}{n})} \end{aligned} \tag{4.2}$$

In Equation 4.2, normalised Shannon entropy of CDR-H3 cluster usage is calculated by dividing repertoire's Shannon entropy by the theoretical maximum entropy of the same repertoire. Maximum Shannon entropy is obtained when all repertoire-present CDR-H3 clusters (n) are used at the same proportions ($\frac{1}{n}$).

4.2.8 Statistical Analysis

Statistical analyses were performed in Python using the scikit-learn [243] and scipy packages. Detailed information on statistical tests is outlined in the figure legends. Data visualization was performed with the seaborn package.

4.2.9 Data availability

All data are within the manuscript and supporting information files, except for our software tool, which is available at `https://github.com/oxpig/saab_plus` under BSD 3-Clause license.

4.3 Results

4.3.1 Structural annotation of human and mouse BCR repertoires

We searched the Observed Antibody Space (OAS) resource [43] for heavy chain BCR studies that contained at least three different B-cell types, had sequences with defined isotype information and consisted of at least 50 BCR repertoires, and identified two studies: Galson *et al.,* ("human") [23] and Greiff *et al.,* ("mouse") [94]. Table 4.1 provides a summary of the number of BCR repertoires with accompanying B-cell type and isotype information in the human and mouse data.

BCR repertoire	B-cell type	OAS Data Unit Number
human	Naive_IGHM	9
human	Marginal_IGHM	9
human	Memory_IGHM	9
human	Memory_IGHG	9
human	Plasma_IGHM	20
human	Plasma_IGHG	25
mouse	Pre_IGHM	20
mouse	Naive_IGHM	24
mouse	Plasma_IGHM	20
mouse	Plasma_IGHG	9

Table 4.1: **Number of human and mouse BCR repertoires used in the SAAB+ analysis.** The human and mouse BCR repertoires were downloaded from the OAS resource. The human data was generated in Galson *et al.,* [23] and the mouse data was from Greiff *et al.,* [94]. These repertoires were used in all subsequent analyses, except for Section 4.3.8 where SAAB+ was applied to healthy human BCR repertoire data from Ghraichy *et al.,* (Healthy (Ghr.)) [107].

Annotating the antibody CDR sequences in these human and mouse BCR studies with structural information allows us to investigate how the three-dimensional shape of CDR-H1, CDR-H2 and CDR-H3 loops vary across BCR repertoires (Figure 4.2). To achieve this, we developed the SAAB+ pipeline that uses FREAD homology modelling for CDR-H3 shape prediction [27] and SCALOP for canonical loop class identification [49] (see Section 4.2.2).

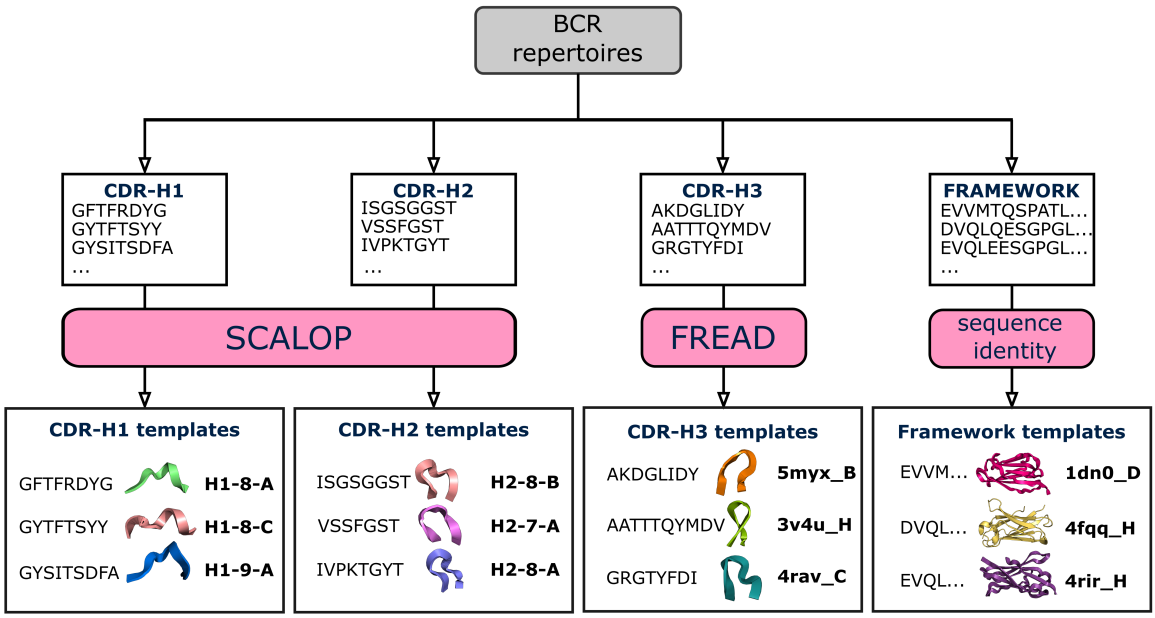


Figure 4.2: **Structural annotation of BCR repertoires.** BCR repertoires are sourced from the OAS resource. For each BCR sequence, CDR loop sequences are extracted, and the closest structural framework match is found, which is used in CDR-H3 loop grafting [169]. Next, SCALOP is used to identify canonical classes for non-CDR-H3 sequences, and FREAD is used to identify whether a CDR-H3 sequence shares a structure with any FREAD crystallographically-solved structures (templates). SCALOP returns a canonical class cluster identification (e.g. H1-8-A); FREAD returns the PDB code of an antibody structure with a protein chain specified (e.g. 5myx_B) [135], a CDR-H3 structural template.

4.3.2 FREAD Performance Assessment

Antibody CDR-H3 loop structure prediction is a high dimensional problem, where loops have access to a diverse set of structural configurations (e.g. coordinates in three-dimensional space). Successful structural interrogation of an entire BCR repertoire relies on our ability to accurately predict CDR-H3 loop shapes. Hence, we assessed the performance of our FREAD CDR-H3 loop modelling for human and mouse BCR repertoires.

As described in Introduction (Section, 1.3.1.2), CDR-H3s exhibit high sequence and length diversity in BCR repertoire data [4]. Our FREAD library contains CDR-H3 loop structures derived from SAbDab [134]. The sequence diversity of these CDR-H3 templates may not be representative of BCR repertoire data, since many of these structures are not natural antibodies but rather rationally engineered [167]. Their length distribution is also different from natural BCR repertoire data (Figure 4.3). These differences might lead to relatively low FREAD ESS (*i.e.* structural homology)

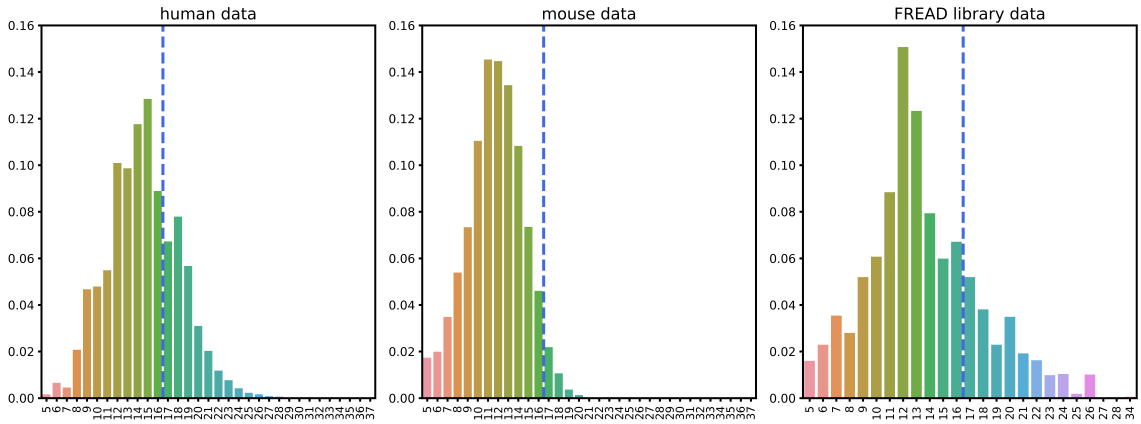


Figure 4.3: **Normalised CDR-H3 length distribution in the human and mouse data, and our FREAD CDR-H3 template library.** Normalised CDR-H3 length distribution was calculated for all BCR repertoires in the human and mouse data, and all CDR-H3 templates that were in our FREAD library. Mouse CDR-H3s were on average shorter than human. The distribution of FREAD CDR-H3 lengths is different that found in the human and mouse data. The vertical blue line shows our CDR-H3 length cutoff (16 residues). Sequences and FREAD templates whose CDR-H3 lengths were longer than the cutoff were not considered in our analysis.

scores between an BCR repertoire sequence and its best hit in the FREAD library.

To accurately estimate FREAD performance for CDR-H3 structure prediction on a given BCR repertoire dataset, we used the following three-step method which is visualised in Figure 4.4. First, we used FREAD to predict templates for all SAbDAb CDR-H3 sequences, retaining all suggested templates alongside their ESS scores ('all versus all'). Here, we allowed FREAD to suggest templates with identical CDR-H3 sequences, since these sequences can be observed in natural BCR repertoire data [169]. We binned CDR-H3 templates into three separate groups (5 to 12, 13 and 14, and 15 and 16) depending of CDR-H3 loop length and FREAD template availabilities. To evaluate the structural similarity between each FREAD template and the native loop, we measured backbone RMSD using the Dynamic Time Warping (DTW) algorithm [48]. This yielded a distribution of ESS scores with accompanying probabilities of RMSD values. Aggregation of the ESS scores into the respective CDR-H3 length bins in the 'all versus all' experiment forms a bimodal distribution (Figure 4.5). The smaller peak associates with much higher ESS scores since it is generated by identical or close-to-identical CDR-H3 sequence matches to the FREAD library. The strong first peak encompasses most of the ESS scores that signifies a low structural homology between the majority of the FREAD templates. This also confirms that the current state of solved antibody CDR-H3 loops has incomplete and biased coverage of the

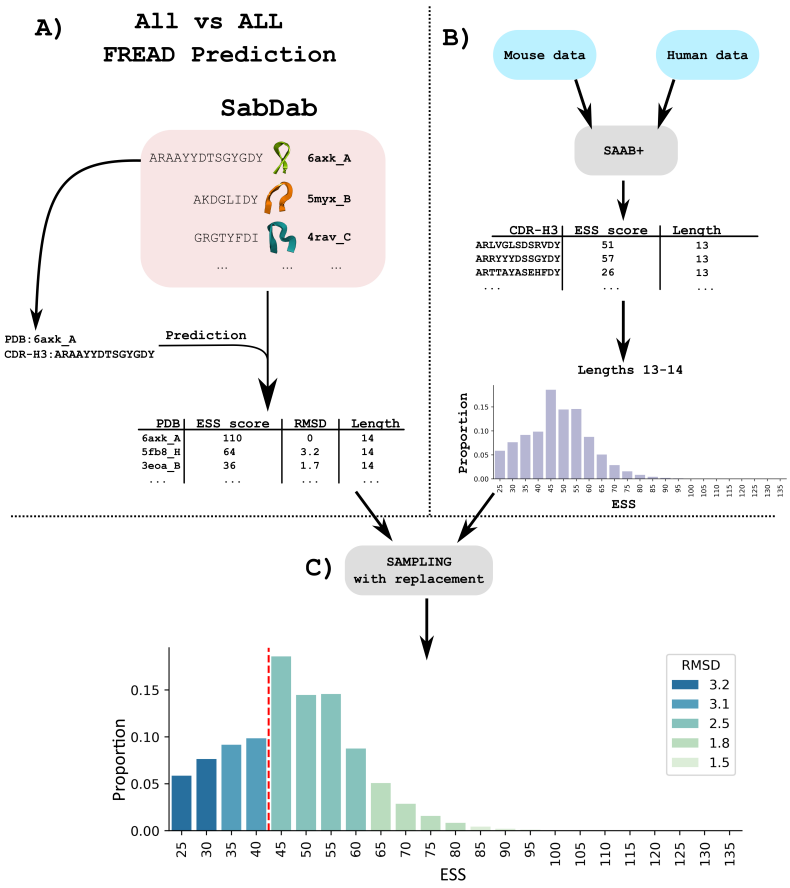


Figure 4.4: **FREAD CDR-H3 modelling assessment**. A schematic summary of the three-step method that was implemented to accurately assess CDR-H3 loop modelling in BCR repertoires. **A)** The first step is to obtain distributions of RMSD scores for a given pair of ESS scores and CDR-H3 loop lengths. This kind of information can readily be generated by leveraging structural antibody databases. 3,769 antibody structures were downloaded from the SAbDab database [134] and applied to "all versus all" cross-validation analysis, where FREAD was used to predict all potential CDR-H3 Antibody PDB template with accompanying ESS and RMSD scores to true structures. **B)** Since all structural databases contain biased antibody sequence data, we used the human and mouse data to obtain naturally observed diversity of CDR-H3 sequences and ESS scores. The SAAB+ pipeline was used to generate ESS scores which were aggregated into separate distributions based on the CDR-H3 length bin. Here, for simplicity, we only show a single length bin. **C)** In the last step, the naturally observed distribution of ESS scores **(B)** is enriched with the distribution of RMSD scores **(A)** across the entire ESS score range. This allows us to estimate average RMSD values within each segment of a five ESS score range. Our pre-defined quality modelling control was set to 3 Å, hence, the red vertical dashed line indicates an ESS cut-off below which CDR-H3 predictions were deemed to be low quality (> 3Å).

structural space.

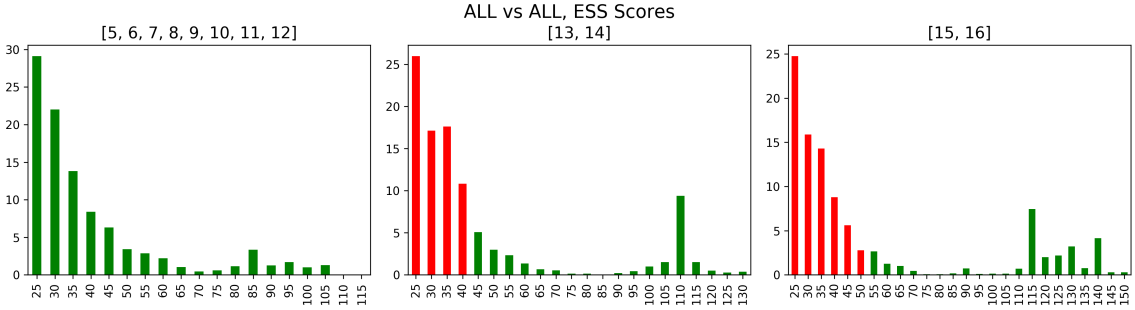


Figure 4.5: **Probability mass function of FREAD ESS scores in the 'all versus all' experiment.** CDR-H3 loop sequences were split into three length bins (5-16, 13-14 and 15-16) to account for the loop accessible structural space and availability of CDR-H3 structural templates. A higher ESS score indicated a closer structural match to the true structure. Red colour highlights ESS scores where average CDR-H3 loop RMSD values are below 3 Å or less than 15% of natural BCR repertoire data is retained in a given length bin. Since longer CDR-H3 loops have access to a wider range of structural configurations and FREAD predictions generate on average higher ESS scores for longer loops, bespoke ESS cut-offs are needed to retain only high quality predictions in each of the three length bins. The X-axis shows a sorted range of ESS scores and the Y-axis shows a percentage of BCR repertoire sequences.

Next, we generated the combined distribution of ESS scores across the top FREAD predictions in the human and mouse BCR repertoire data (ESS_TOP). Since the mouse data contains ∼66 times more sequences, it was subsampled to match the number of sequences found in the human data. To ensure that representative murine ESS scores were picked, this subsampling procedure was repeated 100 times and the average ESS score for each subsample was recorded. The aggregated average ESS scores in each length bin adopt the shape of unimodal Poisson distribution (Figure 4.6). This confirms a high structural diversity of CDR-H3 sequence found in natural BCR repertoires, which naturally results in fewer identical CDR-H3 matches to our FREAD library when compared to the 'all versus all' experiment.

Finally, we randomly picked 1000 FREAD predictions using ESS score values (with replacement) from our 'all versus all' assessment to match the ESS_TOP distribution. This was repeated 100 times and the average RMSD and precision scores were calculated across different loop length bins (5 to 12, 13 and 14, and 15 and 16) and ESS values. This generated the distribution of CDR-H3 length bins with accompanying average RMSD (bins_RMSD) and precision (bins_precision) values.

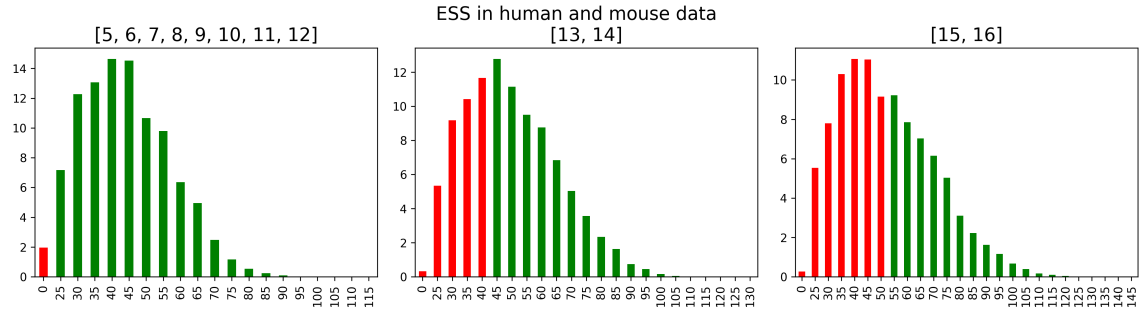


Figure 4.6: **Probability mass function of ESS scores in natural BCR repertoires.** FREAD was run on the human and mouse data yielding only the highest FREAD ESS scores. The human and mouse data were size matched and their combined ESS scores were recorded. Next, ESS scores were grouped into three separate bins based on the CDR-H3 lengths (5-16, 13-14 and 15-16). As expected, average ESS scores increased with CDR-H3 loop lengths as the ESS score is a cumulative value along the sequence and most of pairwise substitution values are positive integers in the ESS look up matrix. FREAD was unable to generate any predictions (ESS score = 0) only in a small number of cases. Higher ESS score indicate closer structural matches to true structures (Figure 4.5). Since longer CDR-H3 loops have access to a wider range of structural configurations and FREAD predictions generate on average higher ESS scores for longer loops, bespoke ESS cut-offs are needed to retain only high quality predictions in each of the three length bins. Red colour highlights ESS scores where average CDR-H3 loop RMSD values are below 3 Å or less than 15% of natural BCR repertoire data is retained in a given length bin. The X-axis shows a sorted range of ESS scores and the Y-axis shows a percentage of BCR repertoire sequences.

We chose ESS cutoffs to achieve an average RMSD better than 3 Å or at least 15% of coverage for every length bin in the human and mouse data (Figure 4.6). Sequences found below the ESS cut-off were not considered in any CDR-H3 related structural analyses. As expected, the higher ESS cut-offs were needed for longer CDR-H3 loops to retain high confidence FREAD predictions. This was concomitant with poorer structural coverage of the binned sequences with longer CDR-H3s (Figure 4.6). Table 4.2 shows the average RMSD and precision we estimate FREAD will achieve on the BCR repertoire data.

4.3.3 Structural CDR-H3 coverage and template usage

We investigated the structures of CDR-H3s used across BCR repertoires of different B-cell types in the human and mouse data. Although generating three-dimensional models can provide a better resolution on BCR repertoire functional grouping, the

Data	Metric	Score
Human	RMSD	2.5 Å
	Precision	68.8%
Mouse	RMSD	2.5 Å
	Precision	68.4%

Table 4.2: **Estimated FREAD average RMSD and precision on the human and mouse BCR repertoire data.** FREAD performance of CDR-H3 structure prediction was validated on the human and mouse data across three separate CDR-H3 length bins: 5 to 12, 13 and 14, and 15 and 16. For each length bin, ESS cutoffs were selected to achieve an average RMSD better than 3 Å or a coverage greater than 15%. The same ESS cutoffs were selected for both human and mouse data. Precision was defined as the percentage of FREAD predictions within 3 Å over the total number of predictions within the ESS cutoff.

Data	Total sequences	CDR-H3 template predicted	Mean coverage with std
Human	5,712,939	2,750,469 (48.1%)	47.2±11%
Mouse	206,680,496	182,309,575 (88%)	88.1±4%

Table 4.3: **FREAD coverage of BCR data.** The human data contained 5.7 million sequences with CDR-H3 loop lengths of 16 amino acids or shorter (see Methods). FREAD generated predictions for 48.1% of CDR-H3s in the human data, with an average coverage of 47.2% across BCR repertoires. The total number of mouse sequences was ∼207 million, of which 88% were structurally-annotated with FREAD. The average structural coverage across mouse BCR repertoires was 88.1%.

time required to model complete repertoires renders the approach very slow [102]. Hence, predicting approximate shapes of CDRs rather than creating refined three-dimensional antibody models allows the SAAB+ pipeline to structurally annotate complete BCR repertoires rapidly. SAAB+ can analyse ∼4.5 million BCR sequences a day on a 40 core computing cluster (Intel Xeon E5-2699 v4 @ 2.20GHz). SAAB+ structurally annotated the human data within two days and the mouse data in 4–5 weeks. Table 4.3 shows the coverage achieved by FREAD for each species.

CDR-H3 structural coverages of BCR repertoires were similar across different B-cell types in the human data (Kruskal-Wallis test, $p = 0.37$), but varied in the mouse data (Kruskal-Wallis test, $p < 0.001$). In both species, the variance of coverage was lower in the BCR repertoires of antigen-unexperienced B-cells (Figure 4.7). The mean structural coverage was higher for mouse CDR-H3s than for human CDR-H3s (Table 4.3). Differences in length distributions could be a major cause of this discrepancy,

as CDR-H3 structures are harder to predict for longer lengths, and the most common lengths were 11 and 12 residues in the mouse data, compared to 15 residues in the human data (Figure 4.3).

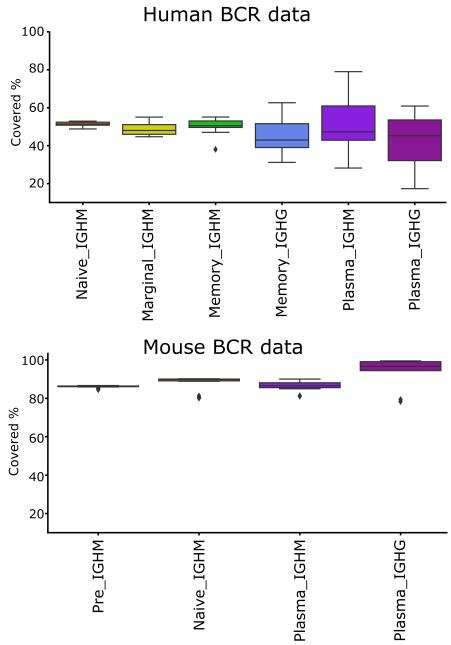


Figure 4.7: **CDR-H3 structural coverage in human (A) and mouse (B) BCR repertoires.** FREAD CDR-H3 modellability was calculated across BCR repertoires of different B-cell types. FREAD modellability is defined as the percentage of sequences with predicted CDR-H3 structures over the total number of sequences in a given BCR repertoire.

Human and mouse BCR repertoires are the effector products of two different sets of germline genes. We therefore investigated whether species germline genes might also translate into preferred CDR-H3 structure usage. We used reported species origin information from SAbDab [134] to calculate the usages of different species CDR-H3 templates across our BCR repertoires (Figure 4.8). As expected, the human and mouse BCR repertoires data used different frequencies of species CDR-H3 templates. The human BCR repertoires tended to use more human CDR-H3 templates as compared to uniform CDR-H3 template sampling, with mouse CDR-H3 templates appearing about as often as would be expected at random. In the mouse data, usage of mouse CDR-H3 templates was enriched, whilst usage of human CDR-H3 templates was reduced. These usages were roughly similar across B-cell types in both human and mouse data, suggesting a species bias towards CDR-H3 structural sampling largely independent of B-cell maturation. To identify portions of structural space that were

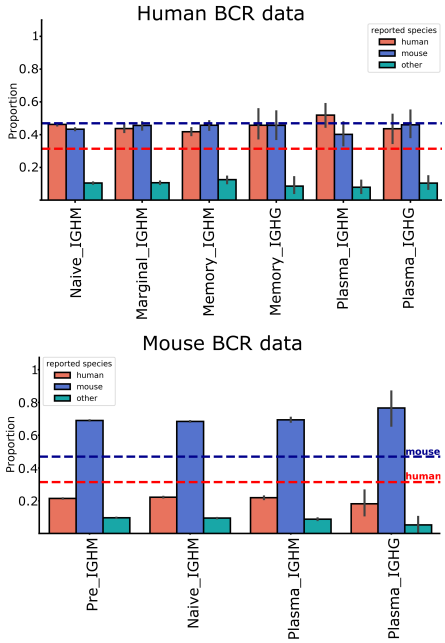


Figure 4.8: **Reported species origin of CDR-H3 templates in the human and mouse data.** We calculated the proportion of CDR-H3 template reported species of origin for every BCR repertoire across different B-cell types. Reported species origin information was extracted from SAbDab [134]. Orange bars show the proportion of human CDR-H3 templates, blue bars represent mouse CDR-H3 templates, while cyan bars depict the proportion of other than human or mouse CDR-H3 templates. The horizontal lines represent the expected outcome of uniform sampling of human (orange) and mouse (blue) CDR-H3 templates. Uniform sampling was calculated as the number of human or mouse CDR-H3 templates over the total number CDR-H3 templates found in our FREAD library.

never seen in the human or mouse BCR repertoire data, we searched for CDR-H3 clusters that were never utilised in the human and mouse data, recording all CDR-H3 templates belonging to these clusters. The number of such templates was 109 (or ∼4% of all FREAD templates). Eighty-eight of the 109 unused CDR-H3 templates derived from nanobodies, which constituted ∼32% of all nanobody CDR-H3 loops in our FREAD library. A further six unused templates belonged to engineered human single heavy domain antibodies. The remaining 15 templates were from conventional antibodies.

Together, these results suggest that different species may engage different epitopes on the same antigen through inherent structural biases. Hence, as more antibody structures become available it will possible to use only PDBs of the species of interest to minimise any potential biases whilst preserving a high repertoire structural coverage.

4.3.4 CDR-H3 cluster profiles along the B-cell differentiation axis

The adaptive immune system responds to antigen exposure by selecting and optimising the most efficacious BCRs. Therefore, B-cells at different maturation stages may possess discrete paratope structural properties.

Galson *et al.,* [23] demonstrated that different B-cell types could be separated using three heterogeneous sequence descriptors (clonality, average CDR-H3 loop length and percentage of V gene mutations) in a principal component analysis (PCA). We repeated their experiment on our human and mouse data (Figure 4.9A and 4.9B). In the human data, their sequence descriptors distinguished B-cell types. In the mouse data, pre, naive, and plasma IGHM BCR repertoires clustered together, whilst plasma IGHG were clearly distinguishable from other B-cell types. This poor separation of murine B-cell types is caused by inherently low levels of SHMs and CDR-H3 length variations in mice.

We investigated whether the structural annotation of CDR-H3s on its own could distinguish the BCR repertoires of different B-cell types, by performing PCA on

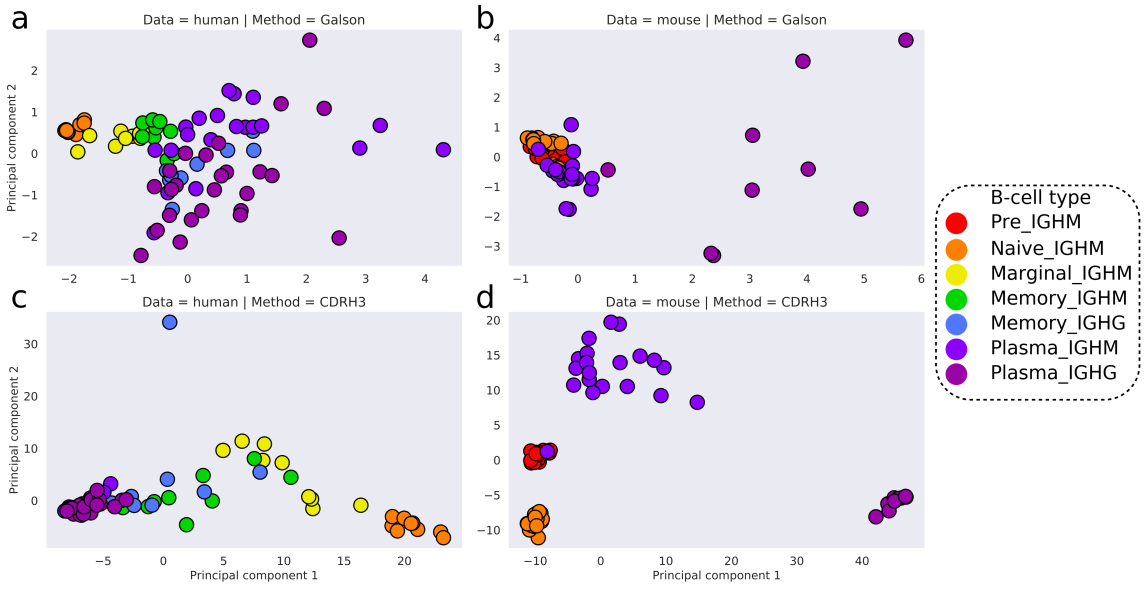


Figure 4.9: **Principal component analysis on the human and mouse BCR repertoire.** Features included in the PCA were either average CDR-H3 length, clonality and percentage of SHMs in V genes **(A, B)** or CDR-H3 cluster usages **(C, D)**. The human data is shown in **A** and **C**, whilst the mouse data is in **B** and **D**. The first two principal components are used to visualise the separation of BCR repertoires. Colours represent different B-cell types.

CDR-H3 cluster usages across BCR repertoires. We found a clear separation of B-cell types in both the human and mouse data (Figure 4.9C and 4.9D), with a sequential pattern of B-cell differentiation in the human data (Naive → Marginal → Memory → Plasma). Mouse IGHM and IGHG plasma BCR repertoires can be distinguished by CDR-H3 cluster usages, whereas neither we nor Galson *et al.,* [23] observe the same separation in the human plasma BCR repertoires. The variance of CDR-H3 cluster usages in plasma IGHM were, in fact, more similar to antigen-unexperienced than to plasma IGHG BCR repertoires in the mouse data. Inaccuracies arising during B-cell sorting could cause improper B-cell labelling, adding noise to the B-cell type separation seen in Figure 4.9. In laboratory mice, the range and degree of antigen exposure is limited by pathogen-free housing conditions and low organism ages. This "purity" could account for the finer separation of B-cell types.

To quantify the behaviour seen in Figure 4.9, we employed the DBSCAN clustering algorithm [244] with increasing maximum distance to closest neighbours (ϵ) to interrogate the densities of CDR-H3 cluster usages across BCR repertoires. Clustering at lower ϵ distances indicates a more similar distribution of CDR-H3 cluster usages. In the human data, all naive BCR repertoires clustered at low ϵ distances along with one marginal zone BCR repertoire. As the value of ϵ was increased, all marginal zone BCR repertoires merged with the naive BCR repertoire cluster, followed by memory and finally plasma BCR repertoires (Figure 4.10). In the mouse data, pre and naive BCR repertoires initially formed two separate clusters at low ϵ distances. As ϵ was increased, antigen-unexperienced (pre and naive) BCR repertoire merged into a single cluster, followed by plasma IGHM and plasma IGHG repertoires respectively (Figure 4.11).

BCR repertoires of different B-cell types are known to have their own characteristic distributions of CDR-H3 lengths [23, 132]. To see whether this alone was driving the separation, we repeated our PCA experiment at specific lengths of CDR-H3, again employing DBSCAN to interrogate the densities of CDR-H3 cluster usages. For each length, we observed the same patterns, confirming that our separation of BCR repertoires was not solely an artefact of CDR-H3 loop length (Figure 4.12).

These findings give structural confirmation to our understanding of B-cell development from antigen-unexperienced to terminally-differentiated plasma B-cells. The collection of CDR-H3s in a terminally-differentiated BCR repertoire should be reflective of individual's complex history of antigenic stimulations yielding highly specialised, high-affinity antibodies [3, 4]. These results demonstrate a mode of structural BCR repertoire ontogeny, where antigen-unexperienced BCR repertoires have

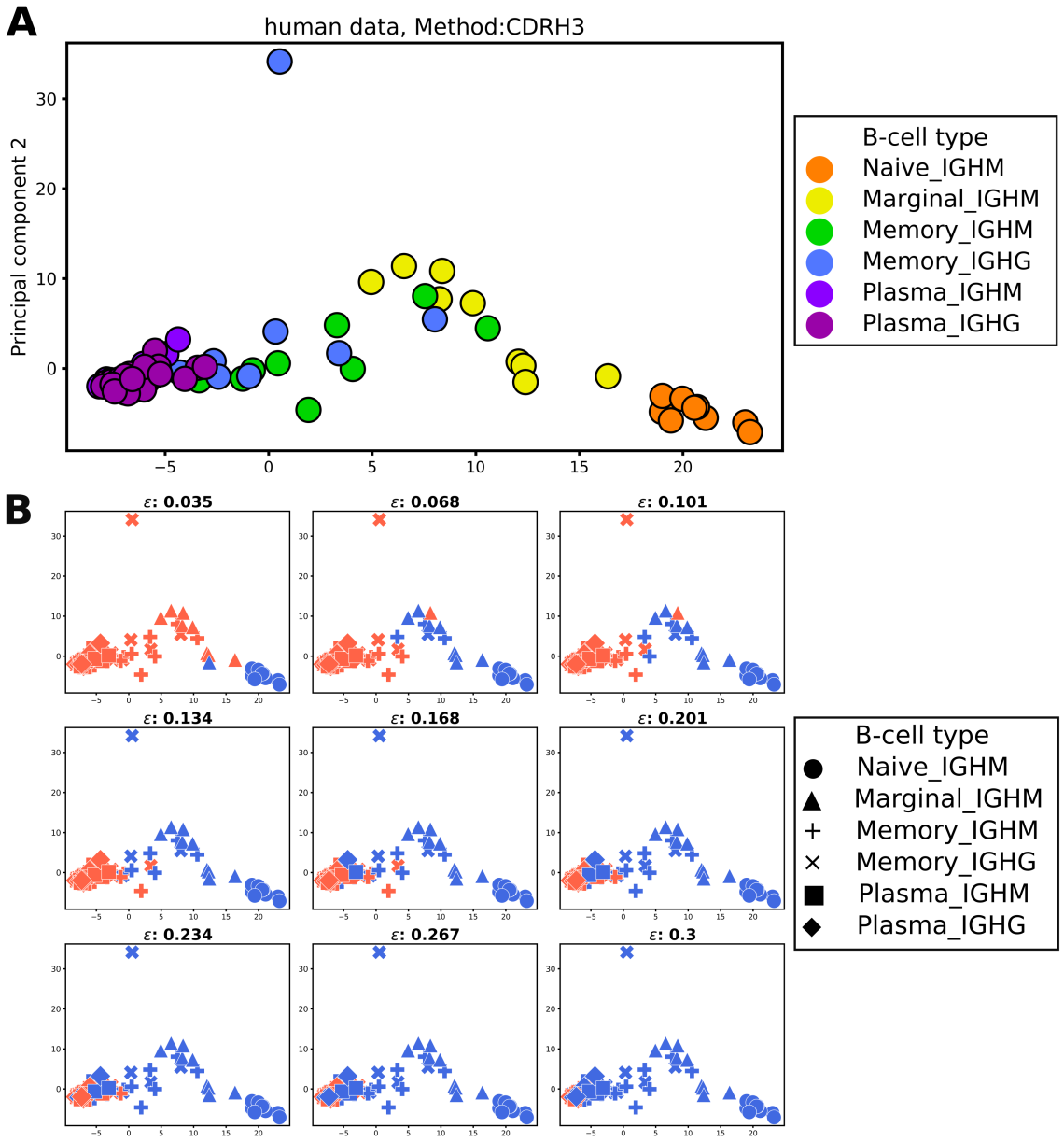


Figure 4.10: **Analysis of densities of CDR-H3 cluster usage in the human data. (a)** PCA was performed on the human BCR repertoires (circles) with CDR-H3 cluster usages used as the features. The first two principal components were used to visualise any separation. Colours represent different B-cell types. **(b)** DBSCAN analysis with increasing maximum distance (ϵ) was employed to interrogate CDR-H3 cluster usage densities across human BCR repertoires. PCA analysis (as in **a**) was then used to visualise the DBSCAN clustering. The parameter ϵ was increased left-to-right, top-to-bottom. Marker shapes indicate different B-cell types; blue colour represents BCR repertoires that clustered with antigen-unexperienced BCR repertoires (naive); orange colour shows DBSCAN-unclustered BCR repertoires at that ϵ value.

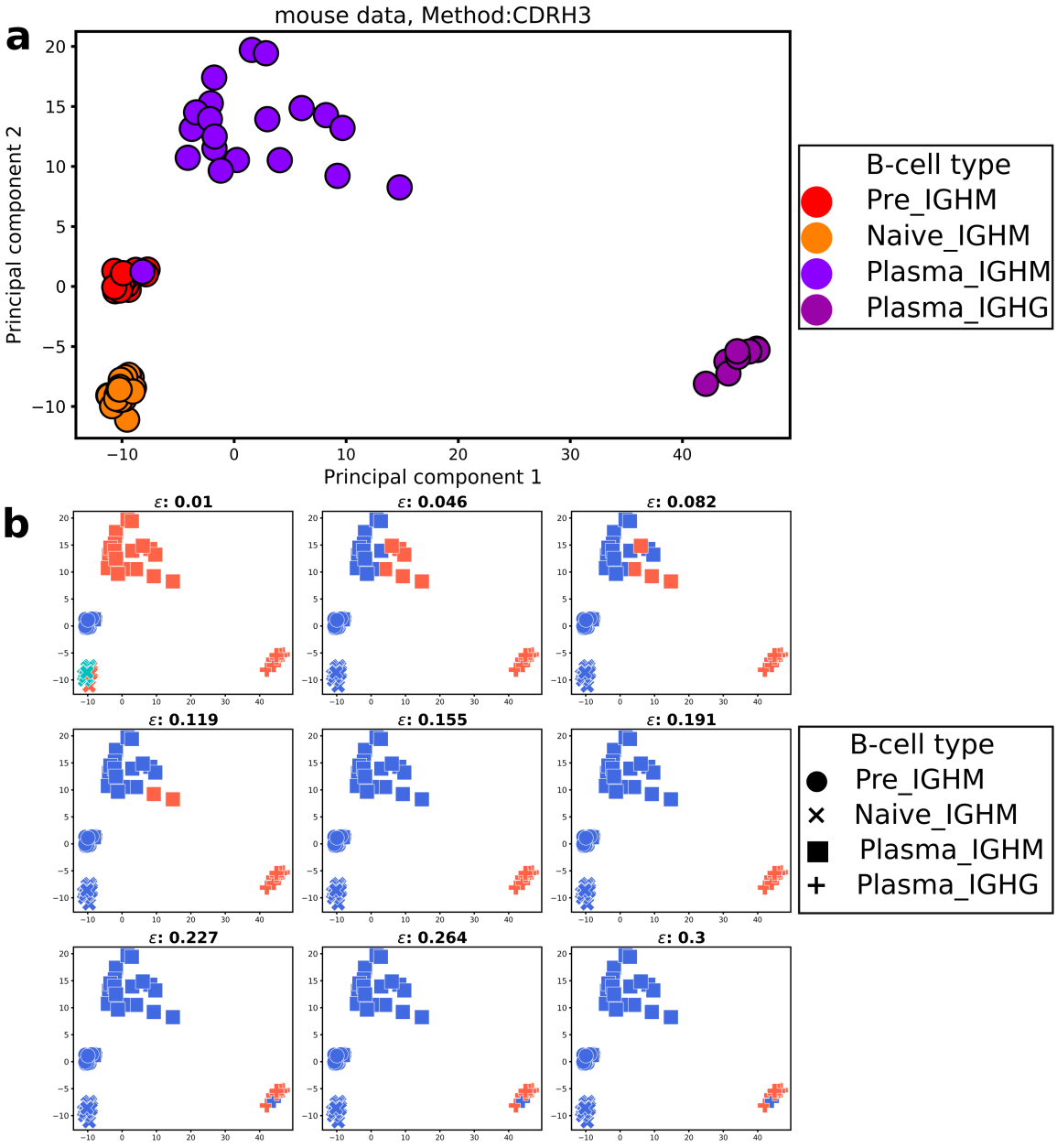


Figure 4.11: **Analysis of densities of CDR-H3 cluster usage in the mouse data. (a)** PCA was performed on the mouse BCR repertoires (circles), with CDR-H3 cluster usages used as the features. The first two principal components were used to visualise any separation. Colours represent different B-cell types. **(b)** DBSCAN analysis with increasing maximum distance (ϵ) was employed to interrogate CDR-H3 usage densities across mouse BCR repertoires. PCA analysis (as in **a**) was then used to visualise the DBSCAN clustering. The parameter ϵ was increased left-to-right, top-to-bottom. Marker shapes indicate different B-cell types; cyan colour (in the top left subplot) represents naive BCR repertoires, blue colour represents BCR repertoires that clustered with antigen-unexperienced BCR repertoires (pre and naive); orange colour shows DBSCAN-unclustered BCR repertoires at that ϵ value.

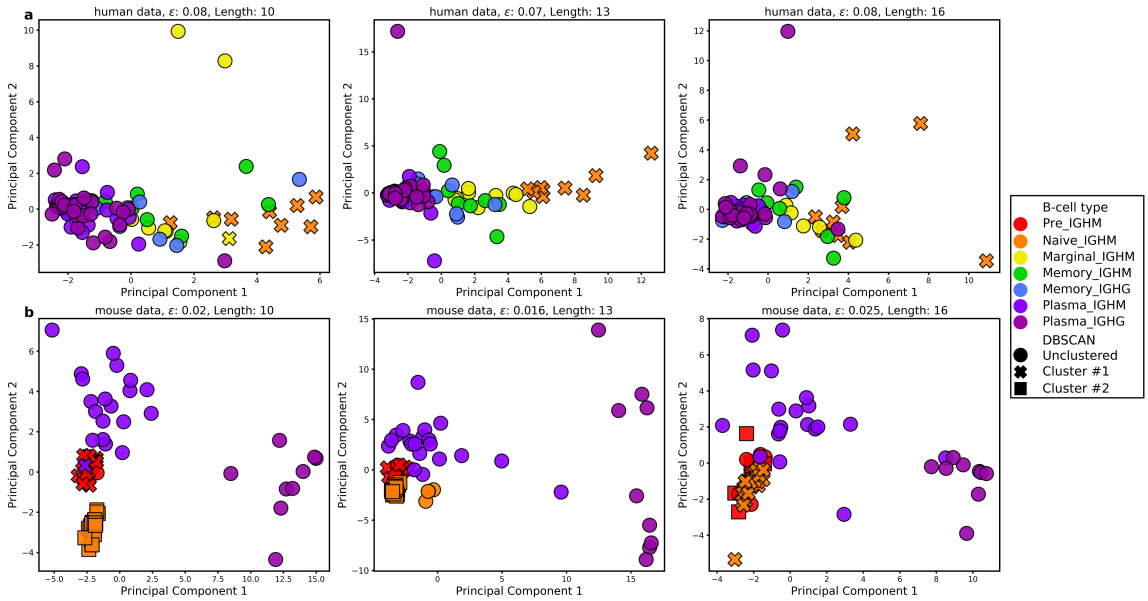


Figure 4.12: **Structural interrogation of the human and mouse data at separate CDR-H3 lengths.** **(a)** PCA was performed on **(a)** human and **(b)** mouse BCR repertoires, with CDR-H3 cluster usages selected as the features. The first two principal components were used to visualise any separation. Colours represent different B-cell types. DBSCAN was employed to quantify densities of CDR-H3 cluster usages across the repertoires. Marker shapes illustrate DBSCAN cluster information. Circle markers indicate DBSCAN unclustered BCR repertoires; other marker shapes show individual DBSCAN clusters. We considered a B-cell type separation if naive and pre (antigen-unexperienced) BCR repertoires displayed the closest densities of CDR-H3 cluster usages at lower ϵ values in DBSCAN. In both **(a)** human and **(b)** mouse repertoires, antigen-unexperienced B-cell types cluster first regardless of CDR-H3 lengths. This confirms that BCR repertoires of different B-cell types have different patterns of CDR-H3 cluster usage.

the most conserved "public" frequencies of CDR-H3 structural clusters across individuals. Upon antigenic stimulation, the somatic hypermutation (SHM) machinery of B-cells recursively introduces point mutations, primarily to the antibody CDR regions [62, 63]. Our DBSCAN analysis shows that BCR repertoires of different B-cell types do not use equal frequencies of CDR-H3 clusters, suggesting that affinity maturation leads to discernible structural changes in the paratope. As B-cells differentiate to the next developmental stage, their repertoires become more personalised through multiple rounds of positive selections; a fine-tuning of antibody CDR structures along the differentiation axis.

Next, we checked whether above results were caused by varying numbers of utilised CDR-H3 clusters. We evaluated the total number of CDR-H3 clusters represented across different B-cell types in the human and mouse data (Figure 4.13A and 4.13B). None of the BCR repertoires used the maximum number of CDR-H3 clusters (which is 1,169), and the numbers varied between BCR repertoires, with antigen-unexperienced repertoires using the most. The average number of CDR-H3 clusters in plasma IGHG BCR repertoires was 3–4 times smaller than in naive repertoires.

This difference could potentially be explained by a varying number of sorted B-cells. To account for the varying sizes of BCR repertoires, we subsampled 10,000 sequences from each of them 100 times and recorded the average number of CDR-H3 clusters. The subsampling gave a similar pattern to the complete data, with the average number of CDR-H3 clusters being highest in antigen-unexperienced BCR repertoires (Figure 4.13C and 4.13D), and total numbers of represented clusters decreasing along the B-cell differentiation axis. This drop in the number of CDR-H3 clusters is not caused by poorer structural coverage of more differentiated BCR repertoires, as we have already shown that the coverage is not significantly different across B-cell types in the human data, and increases for more differentiated cells in the mouse data (Figure 4.7). Therefore, we suspect that this decrease in the number of represented CDR-H3 clusters along the differentiation axis was the result of only specific CDR-H3 structures transitioning to the next development stage.

To confirm this hypothesis, we investigated whether the decreased numbers of CDR-H3 clusters in antigen-experienced BCR repertoires are also accompanied by structural specialisation i.e. personalised CDR-H3 cluster usage. We employed Shannon entropy (Equation 4.1) to investigate the structural diversity of CDR-H3s in different B-cell types.

The entropy analysis showed that the structural diversity of CDR-H3 gradually decreased along the B-cell differentiation axis. Antigen-unexperienced BCR reper-

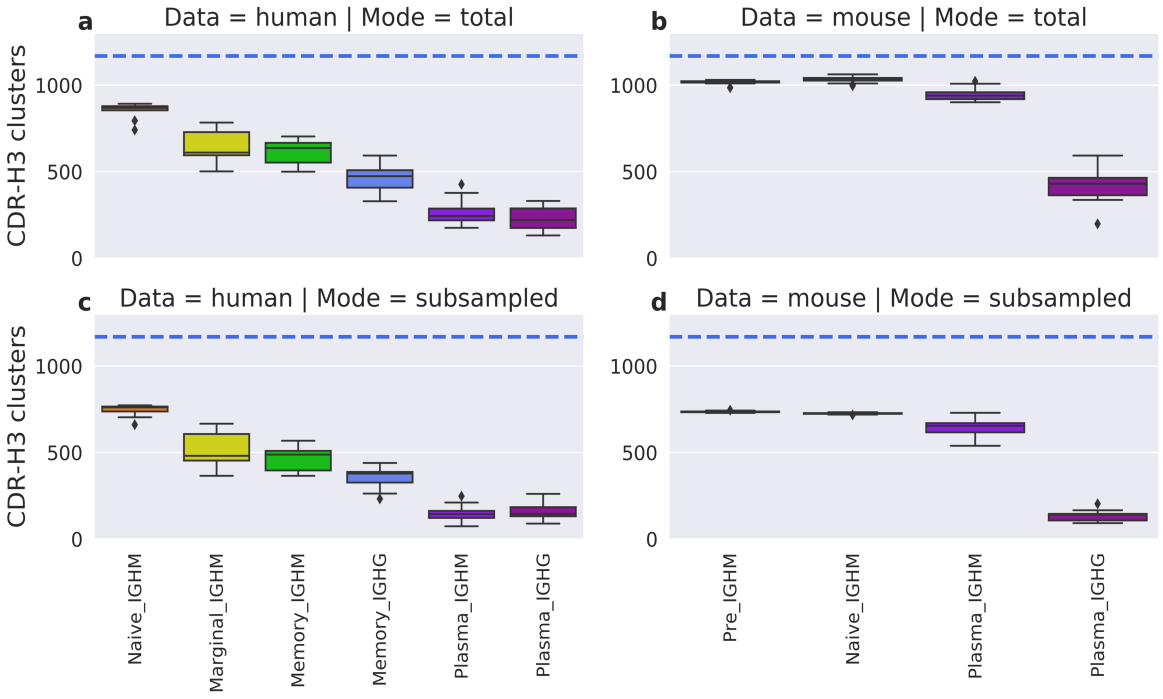


Figure 4.13: **Average number of CDR-H3 clusters in the human and mouse data.** The top boxplots depict the total number of CDR-H3 clusters in human **(a)** and mouse **(b)** BCR repertoires. In the bottom boxplots, every human **(c)** and mouse **(d)** BCR repertoire was subsampled 100 times for 10,000 sequences, with the average number of CDR-H3 clusters recorded. Colours represent different B-cell types. The horizontal blue line shows the total number of CDR-H3 clusters in our FREAD library, and therefore the theoretical maximum.

toires had the highest structural diversity of CDR-H3s, as well as the lowest variance in entropy across B-cell types. Marginal and memory IGHM BCR repertoires utilised the same number of CDR-H3 structures ($p = 0.66$, Mann-Whitney U-Test), whilst the structural diversity was significantly lower in memory B-cells ($p = 0.005$, Mann-Whitney U-Test). Our results again give structural confirmation of the affinity maturation process, where only CDR structures that are specific to cognate antigens are retained.

Overall, the above results demonstrate that B-cell types can be distinguished based on the profile of CDR-H3 structural descriptors alone and that antigen-unexperienced BCR repertoires utilised the highest number and the highest entropy of CDR-H3 clusters. Cluster frequencies in naive BCR repertoires were conserved across different B-cell donors. As B-cells differentiate, their CDR-H3 cluster usage becomes narrower and more distinct between individuals, which is reflective of both affinity maturation and a personalised history of B-cell selection in both inside and outside of the germinal

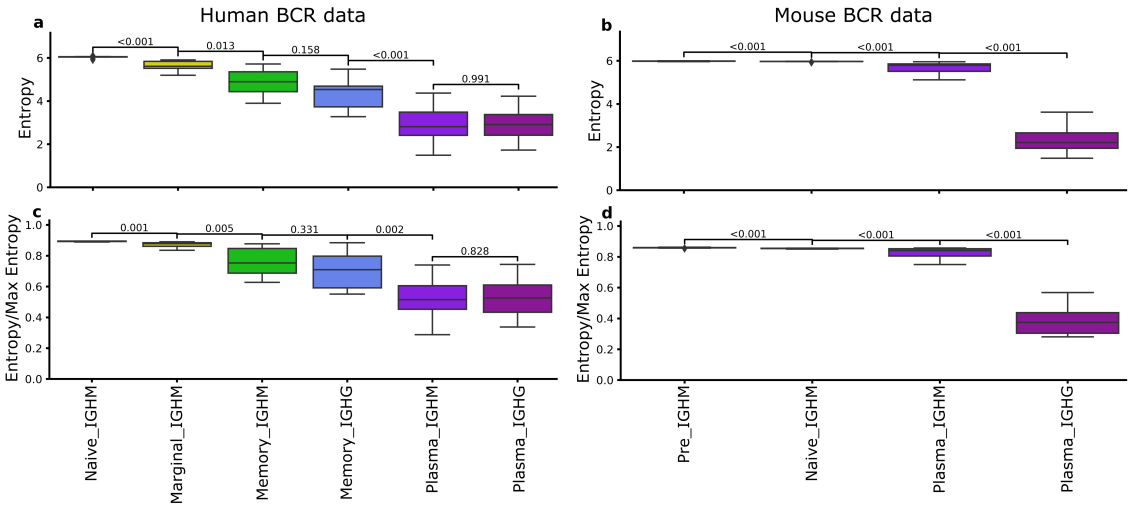


Figure 4.14: **CDR-H3 structural diversity in the human and mouse BCR repertoire data.** Shannon entropy was calculated for CDR-H3 cluster usage in human **(a)** and mouse **(b)** BCR repertoire data. To account for the varying numbers of CDR-H3 clusters across B-cell types, structural diversity of CDR-H3s was expressed as a proportion of entropy over theoretical maximum entropy in the human **(c)** and mouse **(d)** data (Equation 4.2). Theoretical maximum entropy was found for each BCR repertoire by using CDR-H3 structures represented in the given repertoire in equal proportions in entropy calculations. Higher values indicate higher diversities of CDR-H3 cluster usage. The Mann-Whitney U-test was used for statistical analysis and p-values are reported.

centers. These results provide us with the first structural insight into fundamental processes that govern BCR repertoire differentiation across B-cell donors.

4.3.5 Canonical class characterisation

Our analysis so far has focused on CDR-H3, but CDR-H1 and CDR-H2 also play a key role in shaping the antibody paratope [245]. Most CDR-H1 and CDR-H2 loops are found in a small set of structures known as canonical classes. This allows prediction of their structure from sequence with high confidence [49].

A single V gene encodes for both CDR-H1 and CDR-H2 loops and it is known that SHMs preferentially take place in these loops during B-cell differentiation [62, 63, 246]. As the level of SHMs increases with B-cells differentiation, the number of mutations in the V gene has often been used as a proxy to study B-cell development [107, 247].

Here, we investigated whether SHMs in the V gene lead to structural changes in CDR-H1 and CDR-H2 in humans and mice. We calculated the percentage of

Data	Total sequences	CDR-H1 annotated	CDR-H2 annotated
Human	5,712,939	5,598,599 (97.7%)	5,425,279 (95.4%)
Mouse	206,680,496	204,805,604 (99%)	206,592,576 (∼100%)

Table 4.4: **SCALOP annotation of BCR repertoire data.**. Annotation was performed on the human and mouse data. The human data contained 5.7 million sequences with CDR-H3 loop lengths of 16 amino acids or shorter. SCALOP predicted CDR-H1 loop shapes in 97.7% of sequences and CDR-H2 loop shapes in 95.4% in the human data. The total number of mouse sequences was ∼207 million, of which 99% of CDR-H1 and ∼100% of CDR-H2 loop shapes were annotated.

sequences across BCR repertoires where either the CDR-H1 or CDR-H2 canonical class diverged from its parent germline. Sequences with unassigned canonical class information were retained in the analysis as their number was low (Table 4.4), and SHMs can still change loop conformation to a yet uncharacterized canonical class. As of June 2019, only one human and six mouse V genes contained either a CDR-H1 or a CDR-H2 shape that did not fall into a SCALOP canonical classes [49].

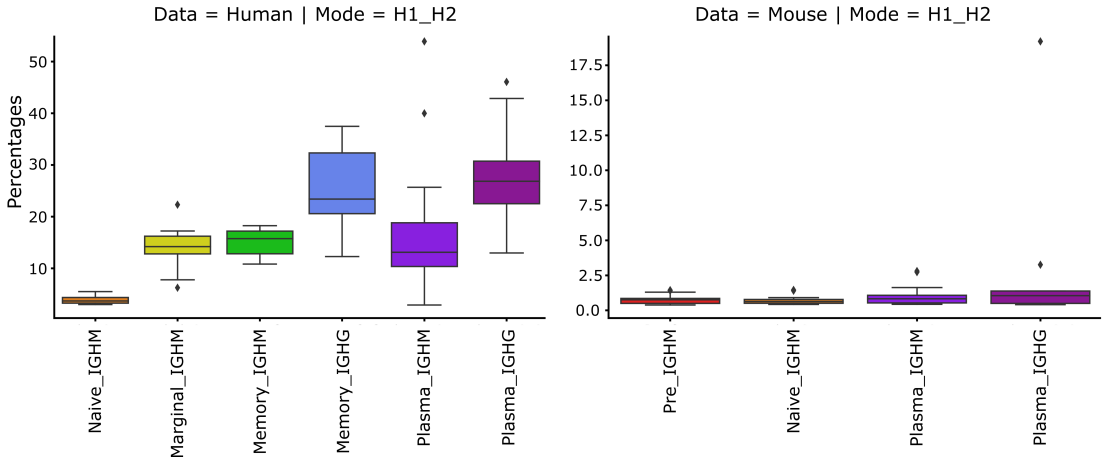


Figure 4.15: **Canonical class divergence from parent germline class in the human and mouse BCR repertoire data.** Canonical class divergence was defined as a mismatch in either the CDR-H1 or CDR-H2 canonical class from the germline canonical class. Percentages were calculated as the number of sequences with canonical class divergence over the total number of sequences in a given BCR repertoire. Colours represent different B-cell types.

Canonical class divergence from germline occurred in all B-cell types, but was observed to increase along the B-cell differentiation axis in the human data (Figure 4.15). This was less clear in the mouse data. Pre- and naive B-cells had less canonical

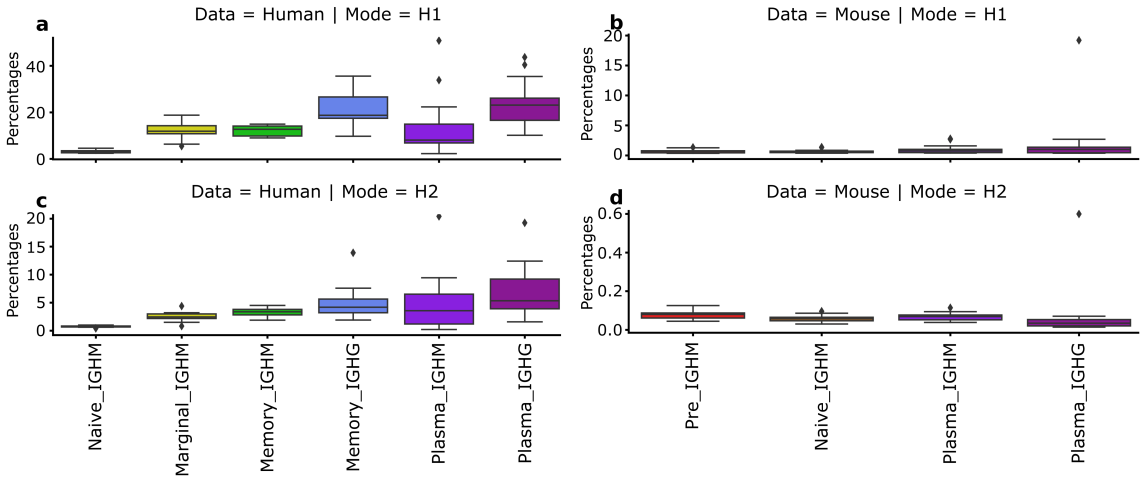


Figure 4.16: **CDR-H1 and CDR-H2 canonical class divergence from parent germline class in the human and mouse BCR repertoire data.** The top boxplots show canonical class divergence in CDR-H1 canonical class from the germline canonical class in the **(a)** human and **(b)** mouse data. Percentages were calculated as the number of sequences with canonical class divergence over the total number of sequences in a given BCR repertoire. The bottom boxplots show the same analysis on CDR-H2 canonical class in the **(c)** human and **(d)** mouse data. Colours represent different B-cell types.

class divergence from the germline, whereas memory and plasma B-cells had a higher divergence. These results place structural information on the knowledge that the percentage of V gene mutations increases with B-cell differentiation [23]. The average percentage of canonical class divergence across B-cell types was consistently higher in human than mouse data. This is in agreement with previously-reported results showing that human V genes tend to accumulate a larger number of SHMs than mouse [248]. This could potentially be caused by the sterile conditions where mouse are housed as well as the short life span of the animal.

CDR-H1 and CDR-H2 loops had different levels of canonical class divergence in both human and mouse data, with CDR-H1s changing their germline loop shapes more often than CDR-H2s (Figure 4.16). This can probably be directly attributed to the different number of canonical classes accessible to CDR-H1 and CDR-H2 (7 versus 4), which implies CDR-H1 loops have a greater degree of structural freedom.

Both Galson *et al.,* [23] and Greiff *et al.,* [94] studies showed that the V gene usages varied across B-cell types. Here, we investigated whether canonical class usages could provide a structural explanation for the observed alterations in V gene utilisation during B-cell differentiation. As with CDR-H3, we performed PCA on combinations of canonical class usages across BCR repertoires (Figure 4.17). In the human data,

we found that naive B-cells utilise very similar canonical class usage, whilst more differentiated B-cell type had higher variance in the class usage. In the mouse data, BCR repertoires were separated into different B-cell types with the clear sequential pattern of B-cell differentiation.

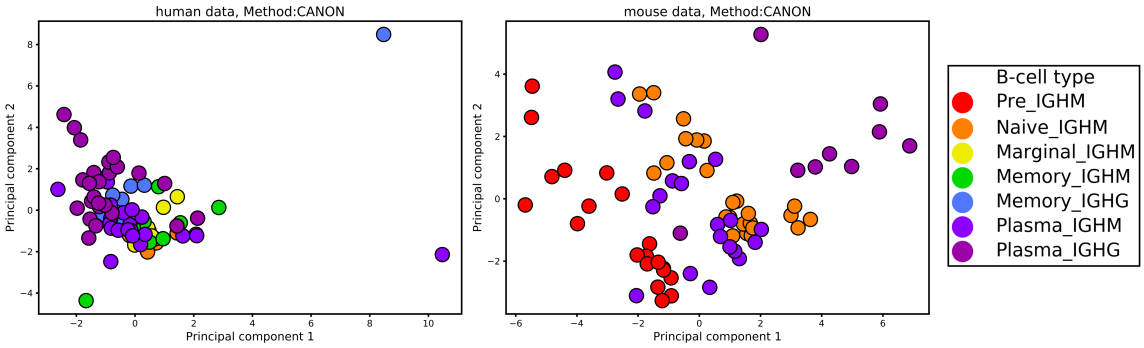


Figure 4.17: **PCA on canonical CDR loop frequencies in the human and mouse BCR repertoires.** Features included in the PCA were frequencies of CDR-H1 and CDR-H2 combinations in BCR repertoires. The first two principal components were used to visualise the separation of BCR repertoires. Colours represent different B-cell types.

Our results demonstrate that canonical class usages are not static during B-cell differentiation, with more mature B-cells exhibiting a higher level of canonical class divergence from the parent germline. This confirms that CDR-H1 and CDR-H2 structures are modulated to help refine the antibody paratope configuration against the cognate antigen in both humans and mice.

4.3.6 Canonical class usages in humans and mice

Current BCR repertoire analysis techniques are largely limited to the intra-species analyses [86, 249], as different species have proprietary sets of V and J germline genes [250]. This makes it difficult to perform any meaningful functional cross-species analyses of the germline genes. Interrogation of antibody solved structures has shown that canonical loops which are encoded by the V genes can still adopt similar shapes regardless of the species origin [49]. As CDR-H1 and CDR-H2 loops play a key role in paratope structure formation, annotating these canonical CDR loops with structural information could be used as a proxy to draw meaningful insights about V gene functioning across different species.

In Section 4.3.4, we have shown that utilisation of CDR-H3 templates displayed strong species biases. Since the antibody paratope primarily (but not exclusively)

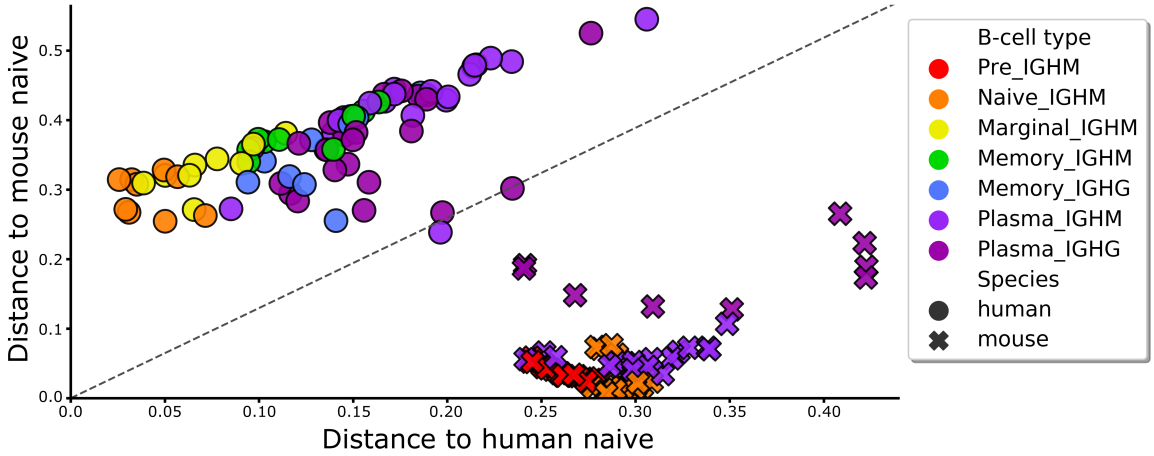


Figure 4.18: **Similarity between canonical loop usages in the human and mouse BCR repertoires.** To find similarities in canonical class usages between the human and mouse data, all BCR repertoires were compared to the average human and mouse repertoires. The average canonical loop proportions were devised based on the human and mouse naive BCR data respectively. The Euclidean distances were then calculated between each human and mouse BCR repertoire and the average naive human (X-axis) and mouse (Y-axis) repertoires. Smaller distances indicate a higher similarity between canonical loop usages. Different colours represent B-cell types; marker style depicts different species.

consists of six CDR regions, here we fill in some of the missing parts in intra-species paratope and V gene functionality comparison by calculating similarities in canonical loop usages between the human and mouse heavy chain data. Since the naive BCR repertoires act as the starting point for structural diversification, we created separate average profiles of canonical loop usages for human and mouse naive BCR repertoires. All other BCR repertoires were then individually compared against these average naive repertoires by calculating repertoire similarity scores (Figure 4.18). The Euclidean distance was selected as the metric to quantify the distances (or "similarity") between a given pair of BCR repertoires (Equation 4.3).

$$distance = \sqrt{\sum_{i=1}^{n} (O_i - A_i)^2} \tag{4.3}$$

In Equation 4.3, we calculate the Euclidean distances based on proportions of canonical loop ($n = 11$) usages between human/mouse datasets (O_i) and the average human/mouse naive repertoires (A_i).

As expected, naive BCR repertoires had the smallest distances to the respective species average naive repertoires (Figure 4.18). Strikingly, all mouse repertoires and

most of human (except for two plasma) BCR repertoires were closer to the respective species average naive repertoires. For the first time, this experiment revealed inter-species variations in canonical loop structural diversities and hence, divergent V gene functioning in BCR repertoires. This indicates that structure frequencies vary across all CDR loops of the VH chain between humans and mice. Therefore, BCR repertoires from different species might preferentially engage separate epitopes of the same antigen. Interestingly, the similarity scores also varied across B-cell types (Figure 4.18). The BCR repertoires had increasingly larger distances to the average naive repertoires along the maturation axis in both the human and mouse data. This agrees with our results in Section 4.3.5, where we described the dynamics of canonical loop shapes in response to antigenic stimulation. The combined evidence from these two experiments show that canonical loop shapes are not static, but readily change their backbone geometries to improve epitope complimentary upon antigenic stimulation.

4.3.7 Patterns of CDR-H3 cluster usage

Biased usage of CDR-H3 clusters is observed in different BCR repertoires along the differentiation axis. As shown in Figure 4.9, antigen-unexperienced B-cells share the closest patterns of CDR-H3 cluster usage. A detailed understanding of biased CDR-H3 structure usage would significantly advance our knowledge of the adaptive immune system development and maturation.

To investigate patterns of biased CDR-H3 cluster usage, we split CDR-H3 clusters into three groups for each B-cell type based on frequencies of CDR-H3 clusters used across our human and mouse BCR repertoires (see Methods 4.2.6 for details). "Structural Stems", which were defined as CDR-H3 clusters, whose frequencies were significantly over-represented across the BCR repertoires of a given B-cell type, "Under-Represented" which describes under-represented CDR-H3 clusters. And CDR-H3 clusters, whose frequencies were not significantly different from random uniform sampling -"Random-Usage" (Figure 4.1).

First, we looked at the average number of CDR-H3 clusters found in our three groups (Structural Stems, Random-Usage and Under-Represented) across the different B-cell types. In all BCR repertoires, Under-Represented always contained the largest number of CDR-H3 clusters (Figure 4.19), however, this does not translate to dominance in terms of coverage (Figure 4.20). This is because, in most cases, Under-Represented CDR-H3 clusters tend to have only a few sequences in a repertoire that share that shape, whereas Structural Stems will have far higher numbers.

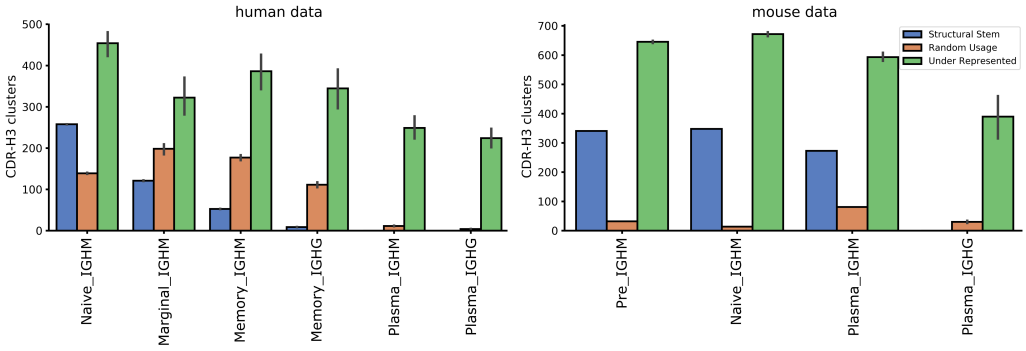


Figure 4.19: **Number of CDR-H3 clusters based on their pattern of usage across different B-cell types the in human and mouse data.** Structural Stems (blue bars) were defined as CDR-H3 clusters, which were over-represented across BCR repertoires of the same B-cell type. Under-Represented (green bars) were under-represented CDR-H3 clusters. CDR-H3 clusters, whose usages were not significantly different from random sampling, were termed Random-Usage (orange bars). The X-axis shows different B-cell types in the order of the B-cell maturation axis. The Y-axis shows the number of CDR-H3 clusters.

In the human data, the number of Structural Stems was largest in naive BCR repertoires and gradually decreased along the B-cell differentiation axis. The number of Random-Usage CDR-H3 clusters was lowest in the naive repertoires. This number increased in marginal BCR repertoires followed by a gradual decline along the B-cell differentiation axis. Similar to the human data, the number of Structural Stems was the highest in antigen-unexperienced BCR repertoires in the mouse data. The number of Structural Stems declined in plasma IGHM and were completely absent in plasma IGHG repertoires.

Next, we investigated the proportional composition of BCR repertoires across B-cell types with Structural Stem, Random-Usage and Under-Represented CDR-H3 clusters. The distribution of repertoire coverages differed between B-cell types in both human and mouse data (Figure 4.20). Structural Stems cover ∼70–80% of antigen-unexperienced BCR repertoires, with coverage declining along the B-cell differentiation axis. In contrast, coverage with Under-Represented clusters gradually increased as B-cells matured. Pre and naive BCR repertoires were least covered with Random-Usage CDR-H3 clusters (only 5–10%). In the human data, coverage with Random-Usage CDR-H3 clusters showed a transient increase in memory BCR repertoires followed by a decline in plasma repertoires, though this trend was less evident in the mouse data.

We investigated conservatism of Structural Stem cluster usage between naive and

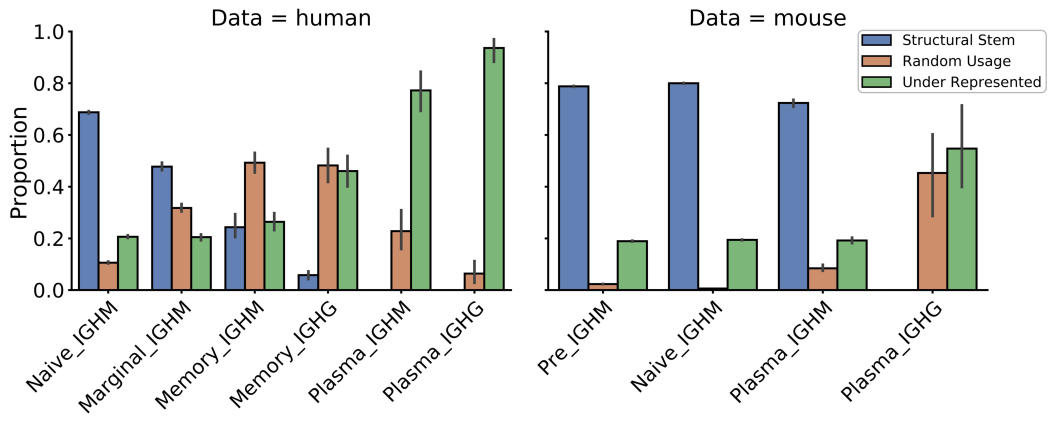


Figure 4.20: **Coverage of BCR repertoires with CDR-H3 clusters based on their pattern of usage in the human and mouse data.** The X-axis shows different B-cell types in the order of the B-cell differentiation axis. The Y-axis shows the proportion coverage of BCR repertoire sequences with CDR-H3 clusters.

antigen-experienced BCR repertoires. We defined Structural Stem conservatism as the number of shared clusters between Structural Stem clusters in B-cell types (Figure D.1). In the human data, ∼98% of Structural Stem CDR-H3 clusters from naive BCR repertoires were found in antigen-experienced BCR repertoires. An analogous pattern was seen with the mouse data. Approximately 99% of Structural Stem CDR-H3 clusters in naive BCR repertoires were found in plasma IGHM BCR repertoires. Our results demonstrates that the same CDR-H3 clusters are preferentially over-represented across different B-cell types, with the number of these over-represented CDR-H3 clusters diminishing to none along the B-cell development axis. This again reinforces our findings that usage of CDR-H3 clusters becomes increasing different in BCR repertoires along the B-cell differentiation axis as only a small number of new over-represented CDR-H3 clusters are shared between antigen-experienced BCR repertoires. These over-represented clusters can be a product of antigen-specific clonally expanded B-cells.

These results demonstrate that antigen-unexperienced BCR repertoires display CDR-H3 structural conservatism. Naive BCR repertoires are largely composed of "public" sets of over-represented CDR-H3 clusters. During B-cell selection, CDR-H3 cluster usages become less conserved across individuals as the coverage with Random-Usage and Under-Represented CDR-H3 clusters rise. In terminally-matured plasma IGHG BCR repertoires, none of CDR-H3 clusters was significantly over-represented across individuals. This reflects how the history of antigenic stimulations structurally shapes BCR repertoires, which become increasingly specialised as B-cells differentiate.

4.3.8 Structural interrogation of healthy human BCR repertoires across different age groups

Sequence diversity in BCR repertoires is known to change in response to numerous factors [6, 88, 251, 252]. One of these factors is the age of B-cell donors [123, 253, 254]. For instance, researchers found an increased load of SHMs in BCR transcripts [255] and an abbarant repertoire isotype composition in older subjects [253]. However, structural dynamics of human BCR repertoire loop shapes across age groups still remains unknown, as all the previous experiments were performed using the primary sequence information only [123]. As it is antibody three-dimensional configuration that dictates antigen binding, interrogation of structural profiles of BCR repertoires should provide an explanation as to why some B-cell donors have less efficient immune responses than others. To study immunosenescence on a structural level, we applied SAAB+ to BCR repertoires from 53 healthy B-cell donors aged between 6 months and 50 years from Ghraichy *et al.,* [107].

BCR repertoires from the Graichy *et al.,* [107] study ("Healthy (Ghr.)") were generated from peripheral blood B-cells that were not experimentally sorted into separated B-cell subsets prior to Ig-seq. Each bulk sequenced BCR repertoire was stratified into two groups based on the annotated isotype information: "MD" (IGHM, IGHD) and "AEG" (IGHA, IGHE, IGHG). Since 60–70% of all B-cells found in the peripheral blood are naive $IGHM^{+}/IGHD^{+}$ B-cells [2], the "MD" group was selected to provide an approximate representation of an antigen-unexperienced B-cell compartment. BCR sequences in the "AEG" group were regarded as antigen-experienced. We note that some antigen-experienced IGHM B-cell types (marginal, memory and plasma IGHM B-cells) are allocated into the MD group. However, these B-cell types are much less abundant in the peripheral blood than naive IGHM B-cells (10% versus 70%) [2].

An analysis of SHMs in the Graichy *et al.,* [107] study showed a significant increase in the number of mutations associated with the age of B-cell donors, especially in first 10 years of life. To check whether these accumulated mutations in the V gene also translate into structural shape changes, we carried out the germline divergence experiment as previously described in Section 4.3.5. All Healthy (Ghr.) repertoires regardless of the isotype information contained some sequences whose structurally annotated canonical loops diverged from the germline encoded structures (Figure 4.21). Antigen-experienced (AEG) BCR repertoires displayed a significantly higher percentage of sequences with diverging canonical loop structures than the MD group. This confirms that antigen-experienced B-cells actively change their BCR paratope

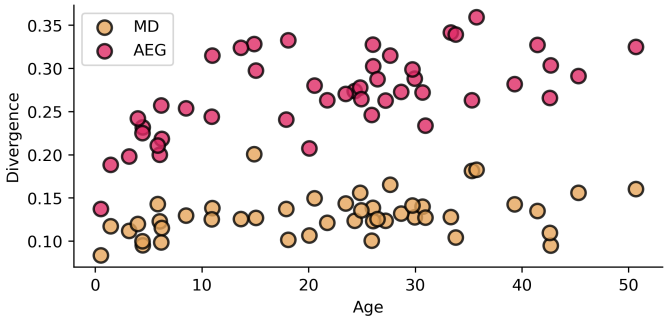


Figure 4.21: **Canonical class divergence in healthy human BCR repertoires with age.** The MD and AEG grouped healthy (Ghr.) repertoires were interrogated for canonical loop divergence from germline encoded loop structures. All germline encoded canonical loops were already pre-annotated with SCALOP [49]. A divergent BCR sequence was counted when either of its canonical loops did not match to its corresponding human germline encoded canonical loop classes. The proportion of divergence was defined as the ratio between the number of the divergent BCR sequences and the total number of BCR sequences in a given repertoire.

configurations to improve chemical and geometrical complimentary for their target antigen. Interestingly, few AEG BCR repertoires consisted of sequences of which more than one third contained germline structure mismatching canonical loops. The proportion of divergence also increased with age in the AEG group, especially in first 10 years of life. This agrees with the previous primary sequence SHM analyses, and the fact that marginal and memory B-cells are not present at birth [107, 255]. A similar pattern was observed in the MD group, although it was less pronounced (Figure 4.21).

The CDR-H3 loop was also found to exhibit abnormal sequence features in BCR repertoires taken from elderly subjects when compared to healthy B-cell donors [123, 256]. These include a small number of highly expanded and persistent clonotypes as well as an increase in highly mutated BCR V gene transcripts with skewed CDR-H3 loop lengths and the higher average CDR-H3 loop hydrophobicity [257]. Therefore, we hypothesise that the structural diversity of CDR-H3 loops could also become compromised with age, providing the basis to some of the clinical outcomes seen in elderly patients [256]. We employed SAAB+ annotated CDR-H3 clusters from Healthy (Ghr.) repertoires to probe structural profiles of these loops across age groups. Since the number of BCR repertoire sequences was not sufficient to provide coverage for each CDR-H3 structural cluster (Figure 4.22), we used a repeated subsampling technique to normalise and de-bias our estimates. We subsampled without replacement all BCR repertoires to 1,000 sequences 100 times and then recorded mean estimates

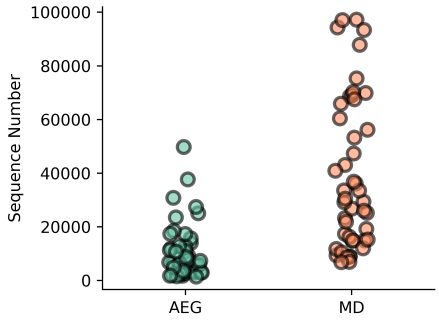


Figure 4.22: **Distribution of SAAB+ annotated Healthy (Ghr.) repertoire sizes.** The original BCR repertoires were obtained directly from the corresponding authors of the study. Each repertoire was structurally annotated with the SAAB+ pipeline. Only successfully annotated sequences were retained for CDR-H3 structural analyses. MD corresponds to antigen-unexperienced Healthy (Ghr.) repertoires, whilst AEG are antigen-experienced repertoires.

of CDR-H3 cluster numbers and normalised Shannon entropy (Equation 4.2) (Figure 4.23). This sample size was selected in order to retain as many AEG repertoires as possible (only 3 out of 53 were found below this threshold), whilst permitting meaningful CDR-H3 cluster analysis.

The AEG and MD repertoires exhibited different CDR-H3 structural profiles which agrees with our previously reported results in Section 4.3.4. The antigen-unexperienced BCR repertoires contained a significantly larger number of unique CDR-H3 clusters and higher cluster entropies than the AEG repertoires (Figure 4.23). This reinforces our previous findings where we showed that structural specialisation in antigen-experienced BCR repertoires causes biased CDR-H3 cluster usages [153]. Remarkably, the MD repertoires exhibited little variance in both the number and entropy of represented CDR-H3 clusters across the interrogated age groups. This demonstrates that the naive B-cell compartment is structurally stable from the time of birth to adulthood, which must be pivotal in mounting an efficient immune response against new antigens. The AEG group exhibited much higher variance in both the number of utilised CDR-H3 clusters and cluster entropies than the MD group. This high variance in the antigen-experience BCR repertoires can be explained with ongoing or recent antigenic stimulations leading to positive selection of antigen binding CDR-H3 loop configurations. Given the sample size and the age range, no obvious patterns were observed in CDR-H3 cluster usage in the AEG group (Figure 4.23). Interestingly, in young adults many AEG repertoires had the same entropy values as the MD repertoires. Such pattern was not seen in children where the AEG repertoires are

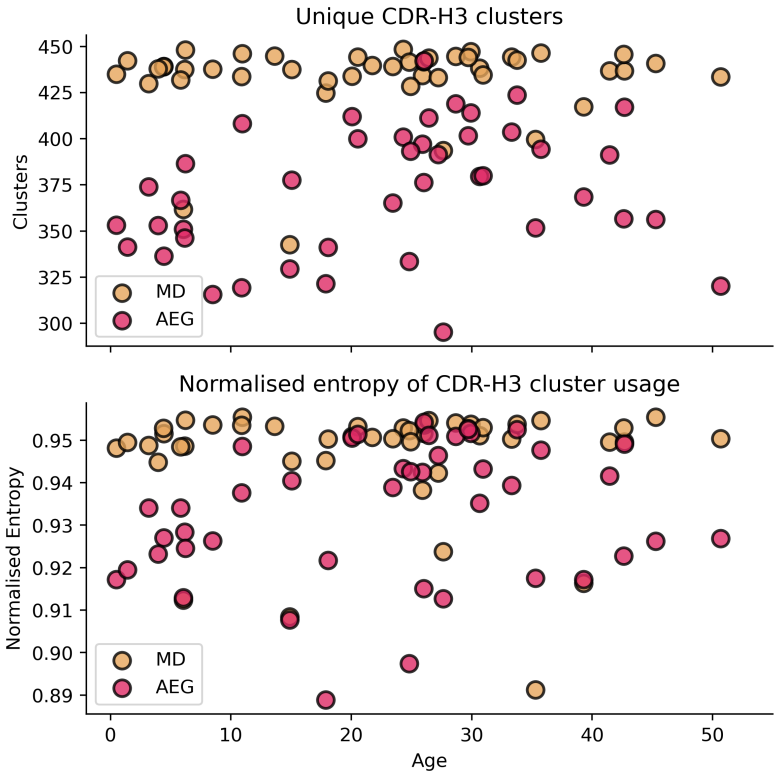


Figure 4.23: **Structural profiles of CDR-H3 loops in healthy BCR repertoires across different ages.** The MD and AEG grouped Healthy (Ghr.) repertoires were interrogated for the number of CDR-H3 cluster utilised and normalised CDR-H3 cluster usage entropy. All BCR repertoires were subsampled to 1000 sequences 100 times followed by calculation of the average values of the estimate. This procedure standardises BCR repertoire sizes and eliminates any potential sampling biases.

consistently structurally less entropic (Figure 4.23). The increase in entropy in BCR repertoires of young adults can potentially be caused by an accumulated exposure to a variety of unique antigens prompting expansions of a wider range of CDR-H3 structural clusters.

4.4 Discussion

We have carried out the first systematic study of structural diversity in the BCR repertoires of multiple donors and species along the B-cell differentiation axis as well as across different ages of human B-cell donors. By mapping sequences to solved antibody structures, we show the structural transformation occurring as BCR repertoires develop in humans and mice.

Our data show that sorted B-cell subsets can be distinguished based solely on the

structural diversity of CDR-H3 loops. Antigen-unexperienced (pre and naive) BCR repertoires contain conserved "public" CDR-H3 cluster frequencies across individuals. As B-cells differentiate, their structural repertoires become increasingly personalised, as a reflection of each individual's history of antigen exposure. Antigenic stimulation induces marked changes in the pattern of CDR-H3 cluster usage in BCR repertoires. The repertoires utilise a smaller number of available CDR-H3 configurations, CDR-H3 structural diversity is reduced, and CDR-H3 cluster usage becomes increasingly divergent from naive BCR repertoires. Structural changes also take place in non-CDR-H3 loops, highlighting the importance of canonical loops in paratope shaping. This shows how structure changes as B-cells, whose CDRs are complementary to cognate antigens, are positively selected.

The human and mouse BCR data used in this work came from two studies that deciphered BCR repertoire sequence diversities following antigenic stimulation [23, 94]. In their original publications, the sequence diversity was shown to decrease along the B-cell differentiation axis, with the most marked decline in the plasma B-cell repertoires in response to antigen exposure [23, 94]. In our work, we find evidence to support these diversity changes on the structural level, which can be indicative of affinity selection of antigen specific BCRs. The SAAB+ pipeline was able to capture both these sequence/structural convergence paradigms, where similar CDR-H3 templates are more likely to be selected in less sequence diverse BCR repertoires as well as identification of sequence dissimilar CDR-H3 loops located in the same structural space.

Most publicly available BCR studies interrogate BCR repertoires either in disease or antigenic stimulation settings [43]. This could potentially lead to sequence/structural diversity measurements that are different from the resting BCR repertoire state. Recently, two BCR studies released cumulatively more than 7 billion BCR reads of unsorted B-cells from healthy individuals [3, 4]. These studies provided crucial insights into the BCR sequence diversity and dynamics amongst healthy individuals. However, analogous large-scale studies have not yet been performed on sorted BCR repertoires due to high costs concomitant with experimental setups, and labour requirements. In order to develop a robust immunodiagnostics pipeline, it is important to understand sequence/structural diversity on the level of the whole repertoire as well as individual B-cell types.

In this work we considered only the three CDRs encoded by heavy chain genes. However, the light chain also plays an important role in shaping the BCR paratope [103]. Therefore, it is anticipated that the diversity of light chain CDR configurations

would also change along the B-cell differentiation axis. To advance our understanding of the role of the light chain in paratope shaping, further investigations are required to study joined structural diversities of heavy and light chains within individual B-cell donors.

Inclusion of cognate light chain pairing information into our analysis would also facilitate generation of refined antibody models. Increased availability of paired heavy/light BCR data [101] and improvements in antibody modelling speed [170] will pave the way to a new frontier of antibody structure usage analysis at the scale of an entire BCR repertoire. Structural descriptors harvested from these models will push forward the resolution of our current work, enabling calculations of paratope charge and hydrophobicity, as well as antibody developability profiles [167].

Structural coverage of CDR-H3s of the human and mouse data was ∼48% and ∼88% respectively. In this analysis, we did not consider BCR sequences with CDR-H3 loop lengths greater than 16 amino acids to ensure high prediction accuracies. This modelling quality filter removed 3.9% (±2 s.t.d) of the mouse and 29% (±11 s.t.d) of the human BCR datasets from the analysis, meaning that the structural diversity information of longer CDR-H3 loops remains unexplored (Figure 4.3). Despite these marked differences in BCR repertoire coverages, both human and mouse data showed similar patterns of structural diversities along the differentiation axis. These findings agree with and provide structural reasons for the sequence diversity measurements calculated across all lengths in the original studies of the human and mouse data. As the mouse repertoires were already ∼ 90% covered with structural information, it is unlikely that the unexplored portions of human BCR repertoires would significantly alter structural diversity profiles of sorted B-cell subsets.

More than 650 antibody structures have been solved since the start of 2019 [134], which constitutes more than 15% of our CDR-H3 template library. With this steadily increasing rate of antibody structure availability and continuous improvements in homology modelling technologies, further studies will be soon necessary to investigate trade-offs between BCR repertoire coverage and prediction accuracies.

Structural characterization of BCR repertoire data can augment existing analysis pipelines [170]. Current BCR repertoire data clustering approaches work on the premise that CDR-H3 sequence identity together with matching V and J genes can be used as a proxy to study paratope convergences [170]. However, sequences with low CDR-H3 sequence identity can adopt close shapes and vice versa [170]. Hence, the development of structure-aware clustering methods such as SAAB+ allows for

the direct grouping of structurally/functionally related BCR sequences [258], as well as enables structural changes to be traced within individual B-cell linages.

A set of CDR-H3 clusters was consistently over-represented across all B-cell donors ("Structural Stems") within the specific B-cell types. These clusters encompassed 70–80% of all sequences in antigen-unexperienced BCR repertoires. This shows that humans and mice largely rely on a conserved "public" set of CDR-H3 clusters to initiate antigen recognition. This knowledge could be leveraged to study immune system disorders, including immunosenescence, where distortions in the conserved public pattern of CDR-H3 cluster usage in antigen-unexperienced BCR repertoires could signal disease states. Furthermore, the knowledge of over-represented CDR-H3 clusters in naive BCR repertoires could be applied in rational phage display library engineering, with Structural Stem cluster sequences used as starting points for library diversity generation.

Recently, transgenic mouse models with human adaptive immune system have been created to raise "naturally human" antibodies in non-human systems [82]. However, their BCR repertoires are shaped inside the murine environment, which could potentially select for BCR paratopes non-native to the human body. Hence, our structural diversity analysis could also be employed in the paratope "humanness" assessment of BCR repertoires derived from transgenic animals.

The human and mouse BCR repertoires displayed biased CDR-H3 template as well as canonical loop shape usages. This was consistent across all B-cell types. This highlights a fundamental dissonance in structural paratope diversity in these two species. Hence, separate species might display preferential binding to different structural epitope motifs found on the same antigens.

For the first time, rapid and accurate structural interrogation of BCR repertoires has enabled us to advance our understanding of the immune system development from birth to adulthood in humans. By applying the SAAB+ pipeline to healthy BCR repertoires from the Ghraichy *et al.,* [107] study, we found that antigen-unexperienced BCR repertoires exhibited conserved structural properties across all B-cell donors regardless of the age group. The antigen-unexperienced BCR repertoires consistently utilised a wider range of CDR-H3 cluster when compared to the antigen-experienced repertoires. Furthermore, CDR-H3 cluster usage was less entropic in the antigen-experienced BCR repertoires, thus highlighting the structural specialisation in response to antigenic stimulation. No apparent signs of immunosenescence were identified in BCR repertoires of healthy adults. This potentially could be caused by a relatively small sample size (53 individuals) and the limited age range

(between 6 months and 50 years). Incorporating BCR repertoires from elderly participants (>70 years) into the analysis as well as preserving the same experimental setup should reveal a clearer signal about immunosenescence on both sequence and structural levels.

Prior body of knowledge showed that human V gene transcripts harbour an increasing number of SHMs with age [107, 255]. For the first time, we deciphered these SMHs on the structural level. We compared CDR-H1 and CDR-H2 loop shapes to the respective germline-encoded shapes to record structural deviations. The antigen-experienced BCR repertoires showed an increasingly higher degree of the structural divergence with the age of the B-cell donors. This pattern was also seen in the antigen-unexperienced BCR repertoires, although at significantly lower levels. These findings reinforce the fact that all antibody CDR loops respond to antigen stimulations by improving paratope-epitope structural complimentary. The increased degree of structural divergence with age is caused by individual's accumulated history of antigenic stimulations.

All the experiments we conducted in this chapter were exclusively focused on BCR repertoire data. However, T-cells also play a key role in mounting an efficient and timely immune response against pathogens [259]. T-cell receptors (TCRs) share common features of the immunoglobulin class with BCRs [26] including somatic V(D)J gene rearrangement to produce an enormous sequence diversity. This enables TCRs to bind to virtually any peptides presented on the surface of the Major-Histocompatibility-Complex (MHC) protein [260]. A large body of research has been performed to decipher TCR sequence diversity to find correlations between expanded TCR clonal sequences and disease prognosis [261, 262]. Annotating entire TCR repertoires with structural information will further advance our understanding of paratope structural convergences otherwise missed by primary sequence analysis techniques. The functionality of our SAAB+ pipeline can easily be extended to perform the structural analysis of TCR repertoires. Five out of six TCR loops have shown to adopt a limited number of structural configurations (e.g. canonical classes), which can be rapidly and reliably predicted from primary sequence [263]. As with BCRs, the hardest task in structural TCR repertoire interrogation is accurate modelling of CDR3β loops [264]. A separate assessment of FREAD CDR3β modelling will be required as the natural sequence diversity in TCR repertoires could differ from BCR repertoires (leading to a different shape of ESS score distributions) and the number of crystallographically solved TCR structures is much lower than BCR structures (448 versus 4,249 as of July 2020) [265].

In the next chapter, we present our contributions to the global COVID-19 research effort. There, we employ OAS and SAAB+ to interrogate sequence and structural diversities in SARS-COV-2 BCR repertoires.

5

Deciphering the immunological footprint of SARS-CoV-2 on the human adaptive immune system

Contents

This chapter is based on the material from the following papers:

1. Raybould, M.I.J., **Kovaltsuk, A.**, Marks, C. & Deane, C.M. (2020) CoV-AbDab: the Coronavirus Antibody Database, Bioinformatics, ():btaa739

2. Galson, J.D., Schaetzle, S., Bashford-Rogers, R.J.M., Raybould, M.I.J., **Kovaltsuk, A.**, Kilpatrick, G.J., Minter, R., Finch, D.K., Dias, J., James, L., Thomas, G., Lee, W.Y.J., Betley, J., Cavlan, O., Leech, A., Deane, C.M., Seoane, J., Caldas, C., Pennington, D., Pfeffer, P. & Osbourn, J. (2020) Deep sequencing of B cell receptor repertoires from COVID-19 patients reveals strong convergent immune signatures, Frontiers in Immunology (Accepted)

3. Ghraichy, M., Galson, J.D., **Kovaltsuk, A.**, von Niederhäusern, V., Schmid, J.M., Miho, E., Kelly, D.F., Deane, C.M. & Trück, J. (2020) Maturation of Naïve and Antigen-experienced B-cell Receptor Repertoires with Age, Frontiers in Immunology, 11:1734

I carried out all the work described in this chapter unless noted otherwise.

5.1 Introduction

SARS-CoV-2 is a new strain of betacoronavirus that was first identified in Wuhan, China in December 2019 [266]. The SARS-CoV-2 virus has rapidly spread across the globe causing the coronavirus disease (COVID-19-19) pandemic. As of October 2020, COVID-19 has infected more than 41 million people and claimed more than 1.2 million human lives [267]. The gathered epidemiological data shows that the mortality rate of COVID-19 is not uniform across the human population. Chronic co-morbidities, gender, race and age have been shown to correlate with the severity of COVID-19 [268, 269].

The first ever recorded outbreak of coronavirus (SARS-CoV-1) was seen in Asia in 2002 [270]. The outbreak was contained within two years, however, more than 8,000 people were infected of which ∼10% died. SARS-CoV-1 and SARS-CoV-2 share 79% of genomic information [271] and were shown to utilise the same mechanism for host cell entry [272]. Sarbecovirus (SARS-CoV-1 and SARS-CoV-2) strains leverage their receptor binding domain (RBD) of the surface spike glycoprotein (S protein) to translocate into host cells by binding to plasma membrane embedded angiotensin converting enzyme 2 (ACE-2). SARS-CoV-2 has been shown to target

ACE-2-expressing cells (endothelial and epithelial cells, and macrophages) in the airways [273]. The damaged cells release pro-inflammatory cytokines, which, in turn, recruit more immune cells to the site. This establishes a pro-inflammatory feedback loop resulting in the loss of pulmonary tissue function that could eventually lead to body hypoxia [273, 274].

Researchers all around the world have been racing to isolate monoclonal COVID-19 neutralising antibodies as potential therapeutics. This kind of intervention is urgently needed to tread severely affected patients as neither a vaccine nor small molecule treatment currently exists (as of October 2020). COVID-19 research has already yielded a large number of SARS-CoV-2 binding antibodies whose sequence information is scattered across an ever-increasing body of scientific literature. Collating these sequences could reveal molecular descriptors of SARS-CoV-2 specific antibodies. Scientists could then leverage this knowledge to isolate promising structural paratope and epitope motifs for therapeutic exploitation. In the first part of this chapter, we outline the development of CoV-AbDab [275], a database that curates all recorded anti-SARS-CoV-1, anti-SARS-CoV-2 and anti-MERS-CoV antibodies with accompanying metadata. Since many surface proteins of both MERS-CoV and SARS-CoV-1 show a high degree of sequence homology to the COVID-19 virus [271], we decided to record their cognate binding/neutralising antibodies as they could potentially harbour betacoronavirus cross-reactive paratope motifs [221].

Multiple studies have already employed sequencing of B-cell receptor (BCR) repertoires to investigate an antibody polyclonal response to COVID-19 in humans [8, 276]. These analyses have helped isolate potent monoclonal COVID-19 antibodies as well as advance our understanding of the viral footprint on the repertoire sequence diversity [277, 278]. In the second part of this chapter, we compliment the existing body of research by performing an inter-study assessment of clonal overlaps between SARS-CoV-2 and healthy BCR repertoires deposited in the Observed Antibody Space (OAS) database [43].

Current COVID-19 BCR repertoire analysis approaches rely exclusively on the primary sequence information to decipher BCR paratope diversities [9]. Supplementing these pipelines with structure-aware techniques could provide insights into BCR paratope dynamics otherwise missed by conventional sequence methods [170]. A precise description of BCR repertoire paratope convergences in response to COVID-19 is of paramount importance in vaccine efficiency assessment. In the last section of this chapter, we apply SAAB+ [153] to characterise the structural dynamics of all CDR loop geometries in COVID-19 BCR repertoires.

5.2 Methods

5.2.1 CoV-AbDab database

5.2.1.1 Data acquisition

Academic preprints, papers and patents containing amino acid as well as nucleotide sequences of coronavirus-binding antibodies/nanobodies were sourced by querying PubMed, BioRxiv, MedRxiv, GenBank, and Google Patents with relevant search terms. ANARCI was used to number antibody variable domain (Fv) sequences according to the IMGT scheme [36], and to determine V and J genes. Our Structural Antibody Database (SabDab) [134], which tracks all antibody structures deposited in the Protein Data Bank (PDB) [135], was mined to identify solved structures of coronavirus-binding antibodies/nanobodies. If a reported coronavirus-binding antibody did not have accompanying structural information in SabDab, we used our antibody/nanobody homology modelling tool, ABodyBuilder [41], to generate a full Fv region structural model.

5.2.1.2 Calculating the closest CoV-AbDab hits to OAS

More than 1.4 billion human BCR IGH sequences were downloaded from the OAS database [43]. These sequences were divided into two groups based on the SARS-CoV-2 status of B-cell donors. For every SARS-CoV-2 binding IGH sequence from CoV-AbDab, a closest hit was calculated to each of these two OAS splits. A hit was defined as a proportion of CDR-H3 sequence identity between IGH sequences with matching CDR-H3 length, V and J genes. This approach is analogous to common clonotyping techniques [86], where hits ≥ 0.85 to a template sequence are often regarded as possessing similar paratope configurations and epitope binding properties [86]. For each CoV-AbDab antibody, we retained only the top scoring hits to each of the OAS splits.

5.2.2 Estimating sequence convergences in COVID-19 BCR repertoires

5.2.2.1 Clonotype datasets

BCR repertoire sequences were placed in the same clonotype if they had identical CDR-H3 lengths, V and J genes, and whose CDR-H3 sequences were within 1 amino acid mismatch per 10 residues.

This clonotyping method was applied to BCR repertoires generated in two separate studies. In our collaborative investigation (Galson *et al.,* [8]), we sequenced BCR repertoires from 31 COVID-19 infected patients using unsorted PBMCs. The mean time since the symptom onset was 10.8 days in the patient group. To maximise the capture of the COVID-19-specific ("functional") response, we only retained clonotypes that were shared across four or more B-cell donors. This resulted into 1,254 "Alchemab" Alchemab clones. A second size-matched set of "random" clonotypes was derived from the Ghraichy *et al.,* [107] study, in which, BCR repertoires were sequenced from 53 healthy human B-cell donors. Clonotype representative sequences were randomly selected from these datasets to match the number of the Alchemab dataset clones disregarding clonal overlaps between B-cell donors.

5.2.2.2 COVID-19 BCR repertoires from Nielsen *et al.,* [198]

We used the Nielsen dataset *et al.,* [198] study ('CoV-2 (Nie.)') as an additional source of SARS-CoV-2 BCR data. Processed and annotated CoV-2 (Nie.) BCR repertoires from six COVID-19 infected patients were downloaded from the Observed Antibody Space (OAS) resource [43]. The average time since the symptom onset was 11.3 days in the patient group. In CoV-2 (Nie.), separate BCR cDNA and gDNA sequencing libraries were prepared and sequenced for each patient. To circumvent the disparity in dataset sizes between pairs of technical replicates, we only selected cDNA replicates for the downstream analysis as they provided better repertoire sequence coverage (Figure 5.6). For one patient (ID:7453) in this dataset, separate blood samples were drawn two days apart (before and after seroconversion to COVID-19).

5.2.2.3 Healthy BCR repertoires from Briney *et al.,* [4]

To estimate the number of "public" clones amongst the Alchemab and random clonotypes, we used data from Briney *et al.,* [4] ('Healthy (Bri.)') as a large scale source of healthy BCR repertoire background data. The pre-processed BCR repertories from nine healthy B-cell donors including all 18 replicates were downloaded from the OAS resource [43].

5.2.2.4 Estimating the functional clonal overlap between the Nielsen and Briney data

The CoV-2 (Nie.) and Healthy (Bri.) BCR repertoires were profiled for cluster centre matches in the Alchemab and random clonotypes. A BCR repertoire sequence was determined as a match if it had identical CDR-H3 loop length, V and J genes, and

was within 1 amino acid mismatch per 10 CDR-H3 residues to a convergent clonotype representative sequence.

To determine whether Alchemab and random clonotype matches were enriched in CoV-2 (Nie.) and/or Healthy (Bri.) repertoires, we performed a cross-repertoire analysis. First, the Healthy (Bri.) datasets were subsampled to match the size of the largest CoV-2 (Nie.) repertoire, ignoring those datasets that were already smaller. This repertoire subsampling was repeated 20 times and the mean number of clonal Alchemab and random matches was recorded and compared to the number of matches in the CoV-2 (Nie.) datasets.

5.2.3 Structural interrogation of COVID-19 BCR repertoires

5.2.3.1 Human BCR repertoire data

All BCR repertoires employed for structural analysis are processed separately from those studied in the clonal overlap analysis (Section 5.2.2).

Human BCR repertoires were sourced from five separate studies, three on disease-responding repertoires and two on healthy repertoires, which we used as controls of the resting repertoire structural diversity. Galson *et al.,* ('CoV-2 (Gal.)') and Nielsen *et al.,* ('CoV-2 (Nie.)') are studies of BCR repertoire responses to SARS-CoV-2 virus across 25 and 6 patients respectively [8, 198]. Only BCR repertoires from the seropositive blood samples of Nielsen *et al.,* were used [198]. We also analysed Davis *et al.,* [247] ('Ebola'), a study of BCR repertoires across four patients that recovered from Ebola infection. Ghraichy *et al.,* [107] ('Healthy (Ghr.)') interrogated BCR repertoire profiles across 53 healthy subjects. Finally, we commissioned blood samples from nine healthy individuals, sequencing their BCR repertoires (See Methods Section 5.2.3.2). These nine repertoire samples represented our second healthy control set ('Healthy (OPIG)'). All five datasets were chosen with consistent experimental approaches in mind — they each focused only on BCR IGH sequences derived from sub-population unsorted B-cells.

Previously-published BCR repertoires were downloaded from the OAS database [43] in the form of IMGT-numbered amino acid IGH sequences [36]. Within each BCR repertoire, we only retained sequences that contained all three CDR regions, had defined isotype and were seen at least twice, as a quality control measure.

5.2.3.2 'Healthy (OPIG)' Repertoire Data Generation

The methods described in this section were developed and performed in Dr. Johannes Trück's group, University of Zürich, Switzerland. The sequencing strategy was largely based on the recently published Ghraichy *et al.,* [107] method with few modifications to accommodate fluorescence-activated cell sorting (FACS) of B-cells. I visited the Trück's group to observe and record all sequencing sample preparation steps.

Nine healthy participants (aged 18-50) were recruited with informed consent. Blood samples were acquired in the form of 40 mL buffy coats. The blood samples were diluted with PBS-EDTA at room temperature. The diluted blood was layered on top of the Ficoll-Paque Plus (Sigma-Aldrich) and centrifuged at 1000xg for 20 minutes. Peripheral blood mononuclear cells (PBMCs) were extracted from the top layer of the separated buffy coat and washed with 50 mL of PBS-EDTA at 300xg for 10 minutes. The supernatant was removed, and the PBMC containing pellet was resuspended in 8 mL of PBS-EDTA. The count and viability of PBMCs was assessed with Trypan blue (Sigma). $CD19^+$ B-cells were magnetically pulled using CD19+ microbeads (Miltenyi Biotec) using the AutoMACS Pro. The cells were lysed with RLT buffer (Qiagen) and stored at -80°C for further use.

In this work, we only used unsorted $CD19^+$B-cells in order to be consistent with methods used across all interrogated BCR studies. Total RNA was isolated from 500,000 unsorted B-cells using RNeasy kits (Qiagen). BCR IGH mRNA was used as a template for first strand cDNA using SuperScript III/IV (Invitrogen). For each cDNA sample, two separate reverse transcriptions reactions were performed based on information of the constant heavy (CH) domain: 1) mix of IGH(M/D)-specific primers 2) mix of IGH(A/E/G)-specific primers. All primers included 14-nucleotide unique molecular identifiers (UMIs) and partial p7 Illumina adapter. The cDNA templates of IGH re-arrangements were PCR amplified using a mix of six framework 1 (FW1) specific primers followed by a second PCR round where p5 Illumina adapters were ligated. PCR conditions for the first round were 95°C for 5 minutes, either 8 cycles (IGHD/IGHM) or 12 cycles (IGHA/E/G) of 98°C for 20 seconds, 60°C for 45 seconds and 72°C for 1 minute, and 72°C for 5 minutes. The PCR conditions for the second round were 95°C for 5 minutes, 22 cycles of 98°C for 20 seconds, 69°C for 20 seconds and 72°C for 15 seconds, and 72°C for 5 minutes. PCR amplicons were separated using gel electrophoresis. DNA bands between 500 and 600 base pairs were excised from the gel and purified. DNA concentration was normalised across all sequencing libraries and 2x300 base pair Illumina MiSeq was used to sequence these samples in parallel.

A pResto toolkit [112] was employed to quality filter and assemble raw Illumina FASTQ reads. Reads with a mean Phread score below 20 were discarded. Next, forward and reverse reads were assembled into sequences if they had a minimum 10 nucleotide overlap. Sequences tagged with the same UMIs were put into separate groups. Within each group, a consensus sequence was defined as the most redundant sequence and all other sequences were collapsed to it. Furthermore, UMI-mismatching BCR sequences with identical full length nucleotide sequences and isotype information were also collapsed. Next, BCR repertoires were annotated according to the minimum standards mandated by the Adaptive Immune Receptor Repertoire (MiAIRR) community with IgBlastn [117, 118]. Sequences that were productive, in the correct reading frame, without stop codons and contained all three CDR loops were retained. ANARCI was employed to number each BCR sequence according to the IMGT scheme and check its structural viability as described in OAS Chapter 3 (Section 3.2.2.5).

5.2.3.3 BCR repertoire structural annotation with SAAB+

SAAB+ was employed to structurally annotate IMGT-defined CDR loops in IGH sequences as described in [153] and Chapter 4. Briefly, SAAB+ applies SCALOP, a method that accurately and rapidly predicts a loop's canonical class directly from its sequence [49], to structurally classify CDR-H1 and CDR-H2. CDR-H3 loops have much higher sequence/structural diversity as a result of the VDJ gene recombination process. Therefore, SAAB+ implements a bespoke version of FREAD tailored for antibody loop modelling [27, 145] to predict a solved CDR-H3 structure ('template') that accurately characterises the given CDR-H3 loop sequence.

Longer CDR-H3 loops in particular have access to an exceptionally large conformational space which is relatively poorly sampled by the solved antibody template structures in the PDB [153], making accurate homology modelling challenging. Our analysis therefore focused on the loop lengths between 5 and 16 inclusive (the "SAAB+ modellability range") (see Chapter 4, section 4.3.2 for more details). The CDR-H3 structural template library was built from a snapshot of SAbDab taken on June 14th, 2020 [134]. We filtered for template CDR-H3 loops from antibody X-ray crystal structures with a resolution $\leq$ 2.9Å. We made an exception for two anti-SARS-CoV-2 CDR-H3s (PDB IDs: 6w41 and 6w7y), whose crystal resolutions were above the standard threshold.

BCR CDR-H3 sequences that could not be modelled by FREAD were classed as "SAAB+ unmodellable" and were removed from any subsequent analyses. All

repertoires with fewer than 10,000 SAAB+ modellable sequences were considered to have inadequate structural sampling and so were precluded from further analysis. Repertoires with greater than 10,000 annotated sequences were randomly sub-sampled to 10,000 datapoints for CDR-H3 structural profiling, ensuring that all structural comparisons were made between size-matched samples.

As some CDR-H3 templates can adopt similar loop backbone geometries, the Dynamic Time Warping (DTW) algorithm described in Nowak *et al.,* [48] was employed to measure the backbone RMSD values between all CDR-H3 template structures. All templates within 0.6Å RMSD of one-another were placed into the same CDR-H3 structural cluster. This method aggregated 3,468 individual CDR-H3 template structures into 1,368 CDR-H3 structural clusters. Repertoire CDR-H3 structural profiles are described in terms of these cluster usages. The total number of unique clusters containing crystallographically-solved SARS-CoV-2-binding CDR-H3 loops ("structural templates") was 3.

5.2.3.4 SARS-CoV-2 over-represented CDR-H3 clusters

The usages of each SAAB+ annotated CDR-H3 cluster were compared between all SARS-CoV-2 and all healthy or Ebola BCR repertoires. We employed the one-tailed Mann-Whitney U-test ($p<0.05$) to check each structural cluster for usage biases by the discrete repertoire state. Depending on the hypothesis, individuals clusters whose usages were found below the significance value were deemed either over- or under-represented respectively.

5.3 Results

5.3.1 CoV-AbDab

The results of this section are based the publication of Raybould *et al.,* [275] publication in Bioinformatics. I was a co-author on this work, where I assisted with data acquisition and data management. I carried out the profiling of CoV-AbDab deposited antibodies against the OAS database.

Since the onset of the COVID-19 pandemic, research groups around the world have been generating data on SARS-CoV-2 binding/neutralising antibodies. Each such dataset contains valuable molecular descriptors of coronavirus-binding antibodies. However, a more detailed analysis of COVID-19 neutralising paratope/epitope motifs can only be achieved once all publicly available COVID-19 binding antibodies are

assimilated in a standardised format. The CoV-AbDab database is the first resource that collates all these experimental results in an attempt to accelerate COVID-19 treatment development. All CoV-AbDab sequences and corresponding structures (if available) were obtained by manual curation of pre-prints, publications, structural databases and patents on SARS-CoV-1, SARS-CoV-2 and MERS-CoV antibodies.

CoV-AbDab deposited antibodies are annotated in a consistent format with accompanying rich metadata. Each sequence contains information on antibody format (nanobody or conventional antibody), coronavirus strain binding and neutralisation information, specified epitope, antibody origin (phage display, hybridoma, BCR repertoire biopanning), V and J gene annotation. This information can be used as search attributes to focus on the subset of anti-coronavirus antibodies. CoV-AbDab users also have access to multiple bulk download options: 1) amino acid Fv sequence with annotated V and J genes, 2) antibody IMGT numbering, 3) PDB structures or homology models. Lastly, CoV-AbDab also enables researchers to find closest "query" matches to the database deposited sequences (Figure 5.1).

5.3.1.1 CoV-AbDab statistics

As of October 13th, CoV-AbDab contained 1,567 anti-coronavirus antibodies of which 1,291 were reported to bind to SARS-CoV-2. 377 (or 30%) of all SARS-CoV-2 binding antibodies were reported to be cross-reactive with SARS-CoV-1. The most popular technique for SARS-CoV-2 antibody isolation was "antigen baiting" where FACS is employed to isolate antigen bound B-cells from seropositive patients (∼80%) (e.g. [84]) followed by display technologies (∼15%) and hybridoma (∼5%). About 30% of all reported SARS-CoV-2 binding antibodies were neutralising.

Since the RBD domain is responsible for initiating viral host cell translocation *via* ACE-2 receptor binding, RBD domain targeting antibodies have a greater chance being SARS-CoV-2 neutralising. As the human adaptive immune system mounts a polyclonal response to the virus, this necessitates additional experimental epitope validation. About 58% of published SARS-CoV-2 antibodies were experimentally validated to target the RBD domain, and in 5% of the data the N-terminal domain was targeted. In the remaining 37% of the data no epitope information was specified.

Human BCR repertoires have been shown to display biased V gene frequencies in response to disease [22]. Thus, we investigated whether V gene usages of COVID-19 antibodies in CoV-AbDab differed from those found in healthy human BCR repertoires (Figure 5.2). We identified a marked gene usage dissonance between CoV-AbDab and a healthy BCR repertoire from Soto *et al.,* [3]. Usages of the following

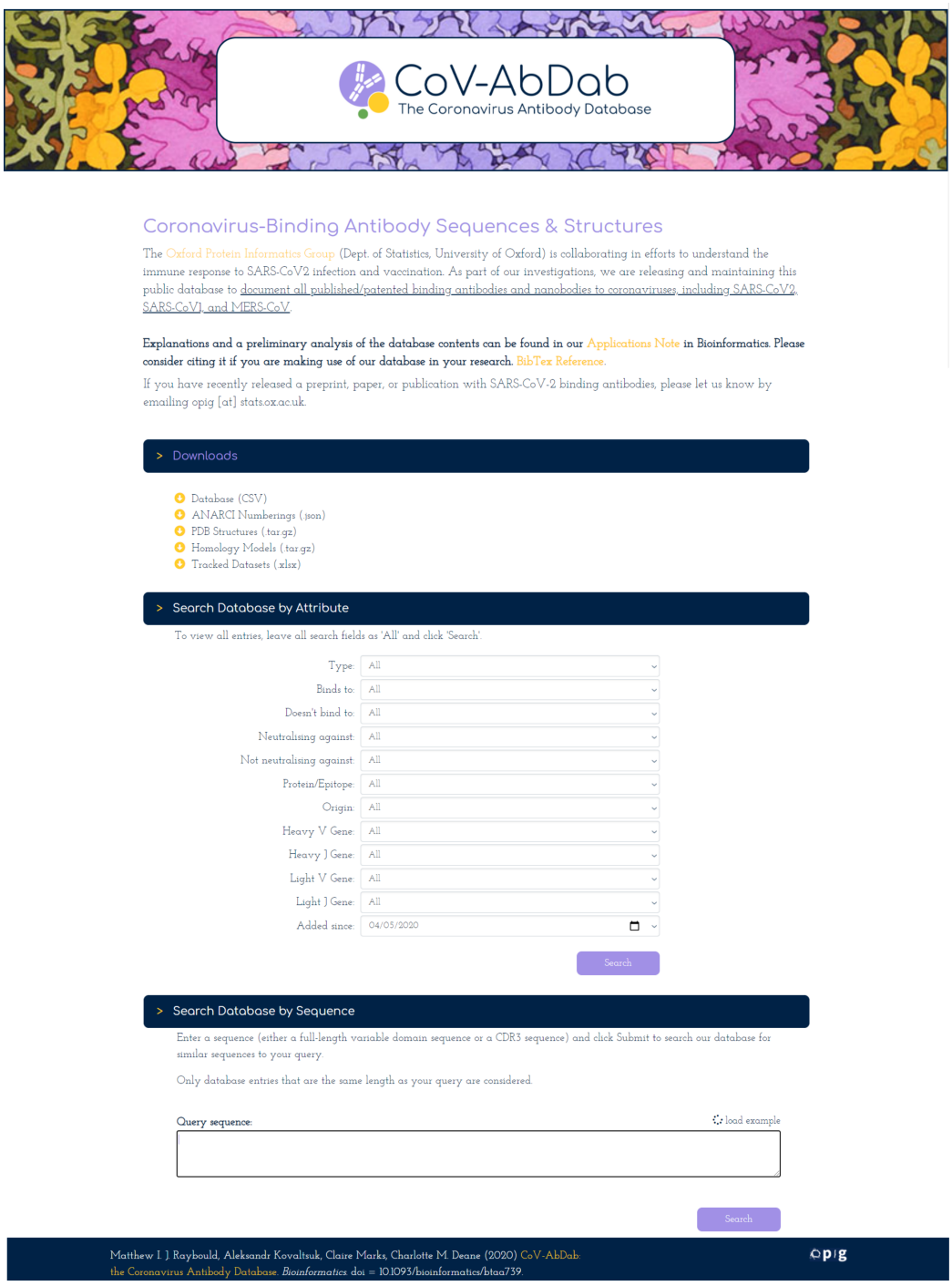


Figure 5.1: **The homepage of CoV-AbDab.** The main page provides several functionalities: Downloads, Search by Attribute and Search by Sequence. Users can download all sequences in the comma-separated format with accompanying metadata. CoV-AbDab also provides separate files containing ANARCI numbering, PDB structures and ABodyBuilder homology models.

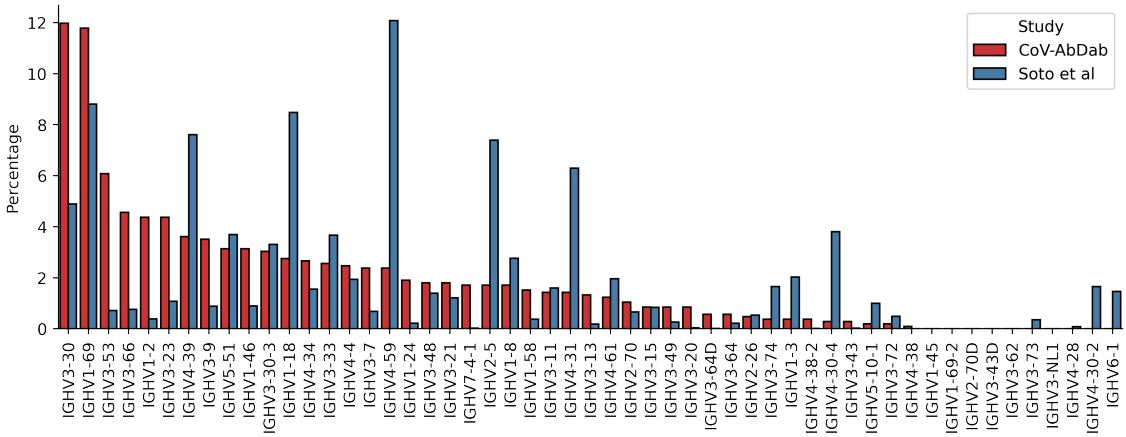


Figure 5.2: **Human V gene usages in CoV-AbDab and healthy BCR repertoires.** V genes from all human SARS-CoV-2 binding antibodies in CoV-AbDab were isolated and counted. As a control of V gene usage in healthy human BCR repertoires we used a single "Donor 3" repertoire from Soto *et al.,* [3]. The annotated Soto dataset was downloaded from OAS [43].

V genes were notably increased in CoV-AbDab: IGHV3-30, IGHV1-69, IGHV3-53, IGHV3-66, IGHV1-2, IGHV3-23, IGHV3-9, IGHV1-46, IGHV3-7, IGHV1-24. Eight of these ten V genes were also reported to be significantly over-represented in other SARS-CoV-2 BCR repertoires [198]. As the majority of CoV-AbDab deposited SARS-CoV-2 antibody originates from antigen baiting experiments suggesting that this technique yields a panel of antibodies whose V gene usages closely mirror a natural BCR repertoire response to COVID-19.

5.3.1.2 CoV-AbDab sequence overlap with OAS

In the previous section, we showed that the majority of CoV-AbDab SARS-CoV-2 antibodies are derived directly from seropositive human patients. However, it is not yet known whether these antibodies are also present in SARS-CoV-2 naive B-cell donors. To perform this analysis we downloaded all human BCR sequences from the OAS database [153]. The OAS repertoires were split into two groups based on SARS-CoV-2 status of B-cell donors. SARS-CoV-2 repertoires only accounted for ∼4.5% of all OAS deposited human IGH sequences. For each CoV-AbDab deposited sequence, we recorded the closest sequence hit to each OAS split. All CoV-AbDab sequences matched to at least one sequence in each OAS split. However, the mean match score was higher in non-SARS-CoV-2 than SARS-CoV-2 BCR repertoires (0.76 and 0.7 respectively) (Figure 5.3A). The closest hit score for each individual CoV-AbDab

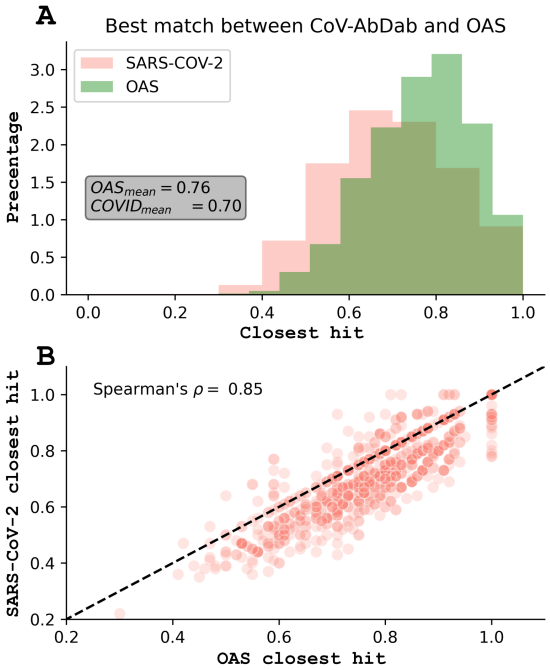


Figure 5.3: **CoV-AbDab sequence overlap with OAS.** All human SARS-CoV-2 binding/neutralising antibodies were downloaded from CoV-AbDab. All naturally observed human BCR repertoires were downloaded from OAS. These repertoires were split into two groups based on the SARS-CoV-2 status of B-cell donors. The hit was defined as a proportion of CDR-H3 sequence identity between IGH sequences in CoV-AbDab and OAS with matching CDR-H3 loop length, V and J genes. **(A)** A distribution of the closest hit scores in the two OAS splits. **(B)** Spearman's correlation of the closest hits between the two OAS groups was calculated. The black dashed line shows the scenario of equivalent closest hit scores. All CoV-AbDab entries found above the dashed line had higher closest hit scores in SARS-CoV-2 BCR repertoires than in the rest of OAS.

entry showed a strong correlation between two OAS splits (Spearman's $\rho = 0.85$) (Figure 5.3B). This indicates that similar sequence information is aggregated in both groups. Several perfect hit matches were recorded in both SARS-CoV-2 (21) and non-SARS-CoV-2 (34) OAS sequences. This suggests that BCR repertoires of some non-COVID-19 challenged individuals might already possess SARS-CoV-2 binding antibodies, potentially elicited by seasonal coronavirus infection [279]. The majority of these matches were found in two large studies [3, 4] where only population unsorted B-cells were made available. Although SARS-CoV OAS repertoires constituted only ~4.5% of all human OAS IGH sequences, 29.6% of the recorded CoV-AbDab antibodies were equally close, or more proximal, to the SARS-CoV-2 BCR repertoires than to healthy/unrelated disease repertoires. This implies a non-trivial convergent

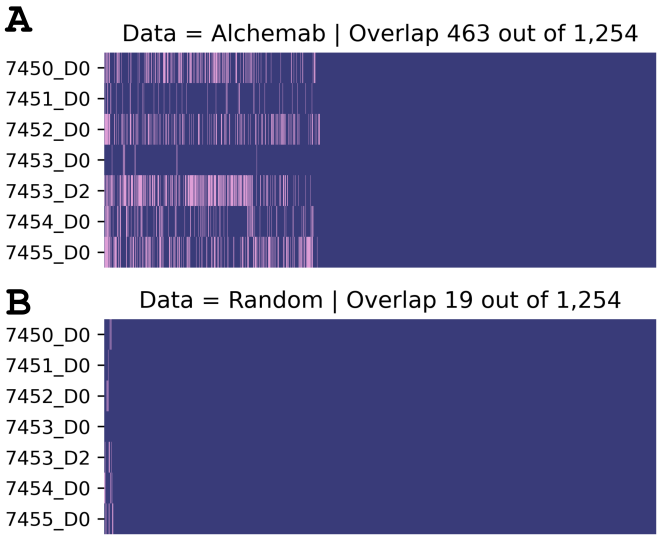


Figure 5.4: **Clonotype match profile to BCR sequence data from the six COVID-19 patients from the Nielsen *et al.*, [198] study.** Plotted along the X-axis are **(A)** the 463 Alchemab and **(B)** 19 Random clonotype matches represented in at least one dataset in the CoV-2 (Nie.) data. Each row represents separate BCR datasets in the CoV-2 (Nie.) data. Pink colour shows a clonal match. In **(B)**, Subject ID 7453_D0 had zero clonal matches.

response to COVID-19 exists across BCR repertoires. In this analysis, only the binding response to SARS-CoV-2 was investigated as all COVID-19 specific antibodies were downloaded from CoV-AbDab.

5.3.2 Sequence convergences across COVID-19 BCR studies

In the previous section we have demonstrated a strong sequence overlap between SARS-CoV-2 specific antibodies from CoV-AbDab and naturally observed BCRs collated in OAS. However, the extent of functional clonal sharing between a pair of independent COVID-19 BCR studies still remains unknown. Here, we investigated functional and public clonotype overlaps between four separate BCR repertoire studies: two SARS-CoV-2 and two healthy. The section is largely based on the Galson *et al.*, pre-print [8] which I co-authored. I was in charge of designing and executing the clonotype overlap analysis.

5.3.2.1 Alchemab clonal overlap with the Nielsen data [198]

We derived a set of 1,254 SARS-CoV-2 "Alchemab" convergent clonotypes by calculating clonal overlaps across 31 COVID-19 patients [8]. These clonotypes harboured

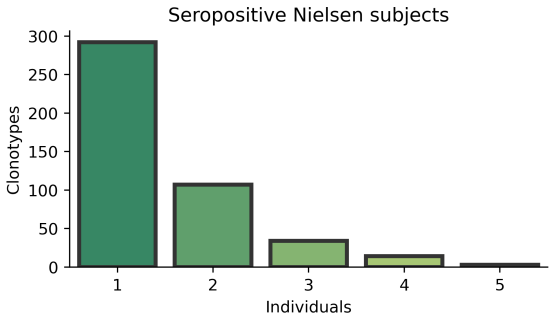


Figure 5.5: **Alchemab clonotype sharedness in the CoV-2 (Nie.) data.** Plotted along the X-axis are the number of the seroconverted individuals from the CoV-2 (Nie.) datasets that share identical Alchemab clonotypes. The majority of clonal matches were associated only with a single individual. Interestingly, two Alchemab clonotypes were shared across all five B-cell donors.

a small number of somatic hypermutations (SHMs) from the closest V gene (∼2.3 nucleotides). This is a surprisingly low count, since only class switched IGHG and IGHA BCR sequences were considered. The majority of SHMs was found in the CDR regions. This suggests that the naive BCR compartment is being stimulated in response to COVID-19 leading to early class switching and clonal expansion of unmutated BCR sequences. A size-matched set of "random" clonotypes was selected from 53 healthy B-cell donors from the Ghraichy *et al.,* [107] study.

An enriched presence of the Alchemab clonotypes in BCR repertoires could be indicative of the disease state. To validate this, the Alchemab and random clonotypes were profiled against seven SARS-CoV-2 BCR repertoires collected from six patients in the Nielsen *et al.,* study [198] ('CoV-2 (Nie.)'). Our analysis showed that at least one BCR sequence in the Nielsen data matched to 463 of our 1,254 Alchemab clonotypes (Figure 5.4A). The average number of clonotype matches to the CoV-2 (Nie.) data was 102.5, but this varied considerably between patients and timepoints. Two of the six patients were seronegative at the day of blood draw (ID:7451 and ID:7453), and these two patients had the fewest clonotype matches (8 and 31 respectively). Patient 7453 had an additional blood sample taken two days later (following seroconversion), and at this point had a large increase in the number of clonotype matches to 201. There was one of the 1,254 Alchemab clonotypes that was found across all six CoV-2 (Nie.) patients, and 2 clonotypes that were found in all five samples from the seroconverted patients, but not found in the seronegative patients (Figure 5.5). In contrast, only 19 of the total 1,254 healthy random clonotypes overlapped with any sequence in the CoV-2 (Nie.) data (Figure 5.4B). This confirms the high level of

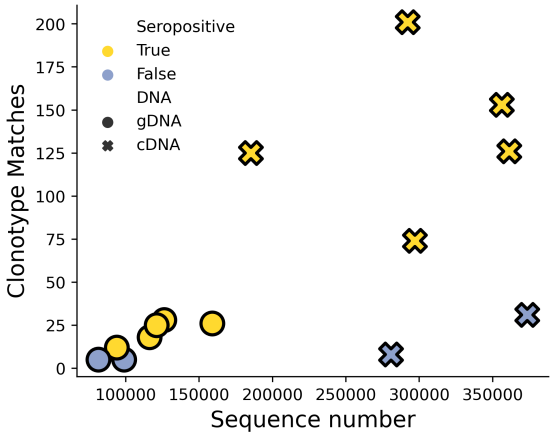


Figure 5.6: **Alchemab clonotype overlaps between cDNA and gDNA sample replicates in the CoV-2 (Nie.) data.** Each CoV-2 (Nie.) dataset contains two technical replicates depending on the DNA template: cDNA and gDNA. Plotted along the X-axis is the sequence number in BCR repertoires. The Y-axis shows the number of matches with the Alchemab clonotypes. Yellow colour corresponds to seropositive and blue colour to seronegative subjects. Circles and Xs represent gDNA and cDNA samples respectively. As the number of sequences increases along the x axis, we observe a proportional increase in the number of clonotype matches. This confirms that the sequence diversity in cDNA samples is not skewed by large plasma B-cell expansions.

functional clonotype convergence between independent SARS-CoV-2 BCR studies.

In this work, we only used BCR repertoire replicates generated from cDNA and not gDNA in the CoV-2 (Nie.) data as these samples provided better sequence coverage which was concomitant with a proportionally higher number of clonal matches (Figure 5.6).

5.3.2.2 Estimating functional COVID-19 overlap in the Alchemab clonotypes

BCR repertoire studies have shown that a small number of clonotypes can always be identified between several healthy B-cell donors [3, 4, 131]. These shared clones are often referred as "public" clones [131].The percentage of clonotype overlaps between two individuals was found to be between 0.2% and 1.6% [4]. Hence, the overlap seen between the Alchemab clonotypes and the CoV-2 (Nie.) data (from the previous section) could potentially be driven by public and not SARS-CoV-2 specific ("functional") clones. To estimate the degree of functional clone sharing, the Alchemab clones were further profiled against BCR repertoires from healthy donors from Briney

et al., [4] ('Healthy (Bri.)'). Matches to the Alchemab clones were found in both studies, however, the number of matches was significantly higher in SARS-CoV-2 than healthy BCR repertoires ($p<0.001$, Mann-Whitney U-test) (Figure 5.7A). The mean clonal match to the Healthy (Bri.) datasets was 62.5 against 135.8 to the seropositive patients in the CoV-2 (Nie.) data. This confirms that SARS-CoV-2 BCR repertoires from seropositive patients are expected to have a higher degree of the functional clonal overlap than healthy BCR repertoires. Since some Alchemab clonotypes were located in the Healthy (Bri.) data, it strongly suggests that the Alchemab set contains a mix of both COVID-19 binding and public clones. The random clones were also found in the Healthy (Bri.) study, but the number of matches was very low (1-20) (Figure 5.7B). This confirms that randomly picked clones bear no valuable information about antigen binding or public clonotype sharedness.

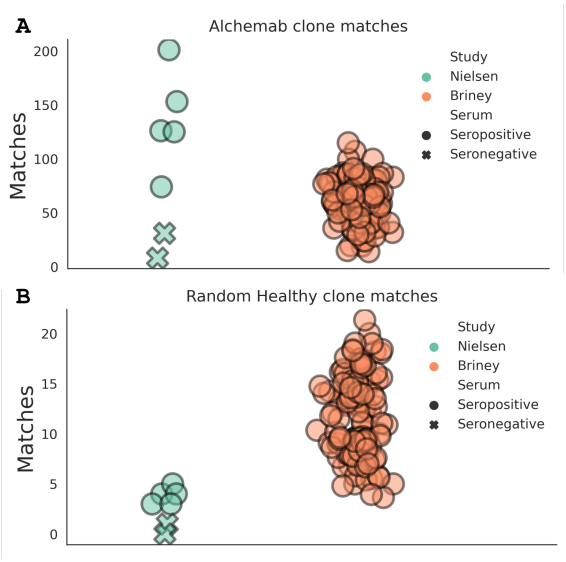


Figure 5.7: **Number of Alchemab and random clonal matches in the Nielsen [198] and Briney[4] studies.** Each entry represents a separate BCR dataset. Along the Y-axis are the number of clonal matches with the **(A)** Alchemab and **(B)** Random clonotypes. Green colour depicts the CoV-2 (Nie.) datasets; Orange colour shows the Healthy (Bri.) datasets. The X marker style represents seronegative samples in the CoV-2 (Nie.) study.

The studies that yielded the Alchemab clones and CoV-2 (Nie.) datasets were performed in the UK and US respectively. At that time two separate SARS-CoV-2 strains were dominant in these two distant geographic locations [280]. Our results, therefore, suggest that the COVID-19 patients mount a convergent SARS-CoV-2 binding response despite the existence of different viral strains suggesting the development of COVID-19 cross-strain immunity.

5.3.3 Structural profiles of COVID-19 BCR repertoires

We have a manuscript in preparation based on the results described in this section. I was in charge of performing and designing all the analyses in this section. We collaborated with Dr. Johannes Trück to generate BCR repertoires from nine healthy B-cell donors. I visited the Trück's group at University of Zürich, Switzerland to record all steps performed for BCR repertoire generation.

As of October 2020, multiple studies have already been performed to decipher primary sequence convergence in BCR repertoires in response to COVID-19 [9, 276]. However, knowledge of repertoire structural convergence still remains unknown. Since all three IGH CDRs are involved in binding to the SARS-CoV-2 antigen [281], this selection pressure ought to be exerted upon the geometries of maturing CDR sequences in SARS-CoV-2 responding BCR repertoires.

We employed our SAAB+ software [153, 169] to probe whether COVID-19 influences the distribution of repertoire CDR loop geometries. For an input IGH sequence, SAAB+ both predicts the canonical class of the CDR-H1/CDR-H2 loops using SCALOP [49], and the shape of the CDR-H3 loop using FREAD [27] (see Chapter 4 for more details). SAAB+ was run on five separate human BCR repertoire studies, all of which sequenced unsorted B-cells. We used two investigations of SARS-CoV-2 BCR repertoires — Galson *et al.,* ('CoV-2 (Gal.)') [8] and Nielsen *et al.,* ('CoV-2 (Nie.)') [198]. For comparison to an unrelated disease, we used a previous study of Ebola BCR repertoires — Davis *et al.,* ('Ebola') [247]. Finally, for comparison to unstimulated healthy BCR repertoires, we used data from Ghraichy *et al.,* (Healthy (Ghr.)') [107] and samples we recently commissioned from nine healthy individuals ('Healthy (OPIG)').

5.3.3.1 Structural annotation of CDR-H3 loops

First, we looked at structural diversity profiles of CDR-H3 loops across IGH repertoires. CDR-H3 loop modellability was not uniform across studies ($p<0.001$, Kruskal-Wallis H-test), with mean values ranging from 51% (CoV-2 (Gal.)) to 58% (Ebola) (Appendix Table E.2). The most likely reason for this variance is the difference in average CDR-H3 length over the SAAB+ modellability range (loop lengths 5-16) of each repertoire (Appendix Table E.2). Repertoire CDR-H3 length distribution can be affected by B-cell subset and isotype compositions [23], and these parameters were not held constant across the different studies. A weak negative correlation was found

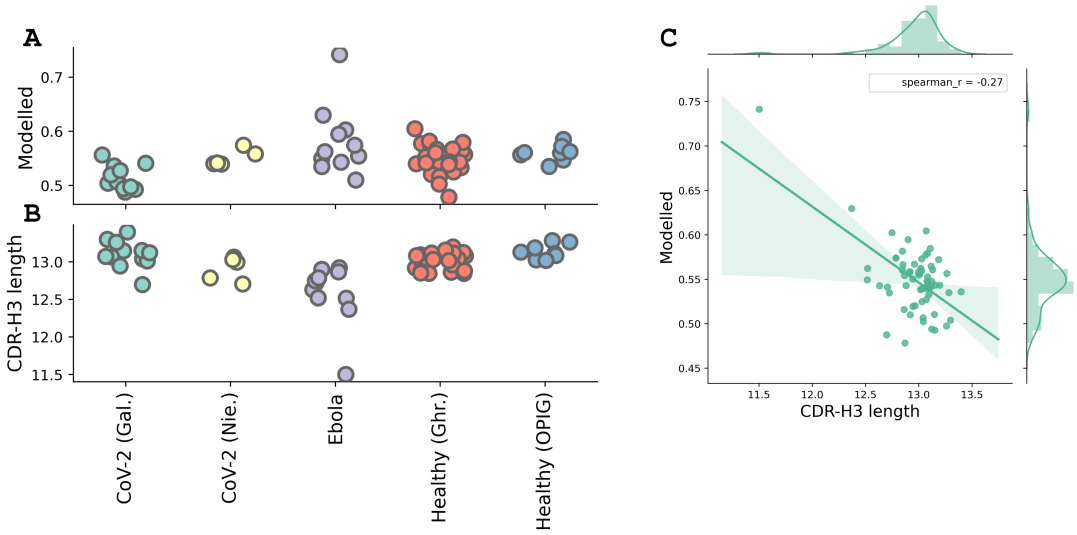


Figure 5.8: **Analysis of SAAB+ CDR-H3 modellability across five independent studies.** SAAB+ was run only on sequences which CDR-H3 loop lengths were between 5 and 16 amino acids (SAAB+ modellability range). **A**) Proportions of modelled CDR-H3 loops in BCR repertoires. **B**) CDR-H3 length distributions within the SAAB+ modellability range across BCR repertoires. **C**) Spearman's rank correlation between BCR repertoire average CDR-H3 lengths and SAAB+ modellability in the SAAB+ modellable range.

between repertoire CDR-H3 length and modellability (Pearson's $\rho = -0.27$) (Figure 5.8). This shows that the repertoire average CDR-H3 loop length only plays a minor role in the structural coverage across our BCR repertoires. Hence, other features such as biases in repertoire sequence and structural diversities are likely to be strong contributors to repertoire structural modellability. However, this disparity in modellability should only have a small impact on overall CDR-H3 structural profiles, as coverage across all BCR repertoires was roughly in the same range ($\text{Mean}_{\text{Modellability}} = 54.9 \pm 3.8\%$).

Together, the difference in annotation proportion and disparity in an initial sequencing sample depth resulted in SAAB+ annotated datasets ranging from 1,827 to 183,743 sequences (Figure 5.9). Annotated sets with fewer than 10,000 data points were removed from further analysis to ensure adequate CDR-H3 loop shape sampling. Each annotated dataset was subsampled to yield a single sample of 10,000 SAAB+ profiled sequences. A total of 17 SARS-CoV-2 repertoires, 11 Ebola repertoires, and 36 healthy repertoires were carried forward for downstream CDR-H3 profile analyses (Appendix Table E.1).

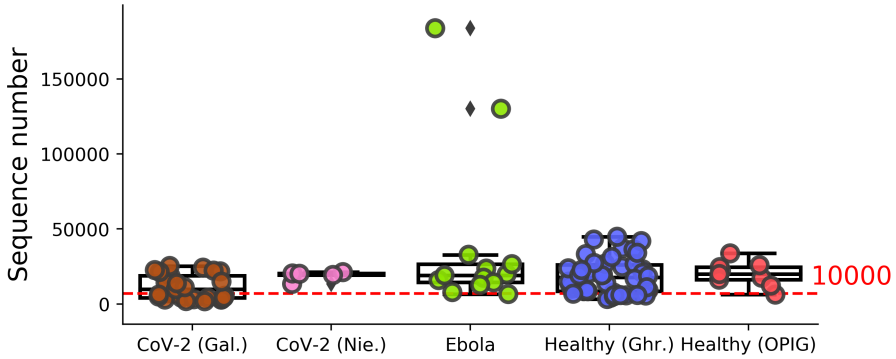


Figure 5.9: **Number of SAAB+ modelled sequences in BCR repertoires in five independent studies.** Only sequences that were within the SAAB+ modellability range (5-16 amino acids) were counted. The red dashed line indicates the cutoff of the number of SAAB+ modelled sequences. BCR repertoires with fewer than 10,000 modelled sequences were not included in the downstream analyses.

5.3.3.2 Structural profile of CDR-H3 loops

The structural space accessible to CDR-H3 loops is very large [52, 148]. We previously showed that it is possible to extract meaningful statistics about structural CDR-H3 loop diversity by grouping FREAD CDR-H3 template predictions into CDR-H3 clusters based on the distance of their backbone atoms (Chapter 4 and [44]). The number of utilised CDR-H3 clusters and entropy of those clusters can then be used as a proxy to study structural repertoire proliferation. A reduction in the number of used CDR-H3 clusters and/or entropies in virus stimulated BCR repertoires could potentially be indicative of repertoire structural convergence against a common antigen.

Performing a *post-hoc* Dunn test ($p<0.05$) showed that healthy BCR repertoires utilised a significantly higher number of available CDR-H3 clusters when compared to the virus stimulated datasets (Appendix Table E.3). The number of CDR-H3 structural clusters sampled across virus stimulated and healthy datasets did not vary significantly in spite of different repertoire structural coverages (Figure 5.10). On average, 90 more clusters were present in the healthy BCR repertoires (∼920 clusters) compared to the virus stimulated BCR repertoires (∼830 clusters). This provides a strong evidence of the repertoire selection, as these datasets were sources from various research groups using different sequencing sample preparation strategies.

Next, we studied the structural entropy of CDR-H3 loop usage in healthy and virus stimulated repertoires. A reduction in the repertoire structural entropy could signal repertoire activation through increased structural similarity of antibody loop

backbone geometries. The analysis of SAAB+ annotated repertoires showed that the diversities of CDR-H3 shapes were significantly different across healthy and disease states ($p<0.001$, Kruskal-Wallis H-Test) (Figure 5.10B). Similar to available CDR-H3 cluster utilisation, CDR-H3 entropy was significantly higher in healthy repertoires than in SARS-CoV-2 or Ebola repertoires ($p<0.05$, Post-hoc Dunn test). The entropy and the number of utilised CDR-H3 clusters for each repertoire can be found in Appendix Table E.3.

5.3.3.3 Structural commonalities across BCR repertoires

We have demonstrated that healthy and virus stimulated BCR repertoires have distinct CDR-H3 structural profiles. Healthy repertoires utilise larger numbers and have higher diversities of available CDR-H3 shapes when compared to SARS-CoV-2 or convalescent Ebola BCR datasets. This is indicative of a repertoire structural response to the immunogens. However, these metrics ignore the actual pattern of the repertoire CDR-H3 loop structure usage by reducing them to a single score. Next, we compared the patterns of CDR-H3 shape usages across BCR repertoires and checked whether structural commonalities exist across BCR repertoires of the same health states.

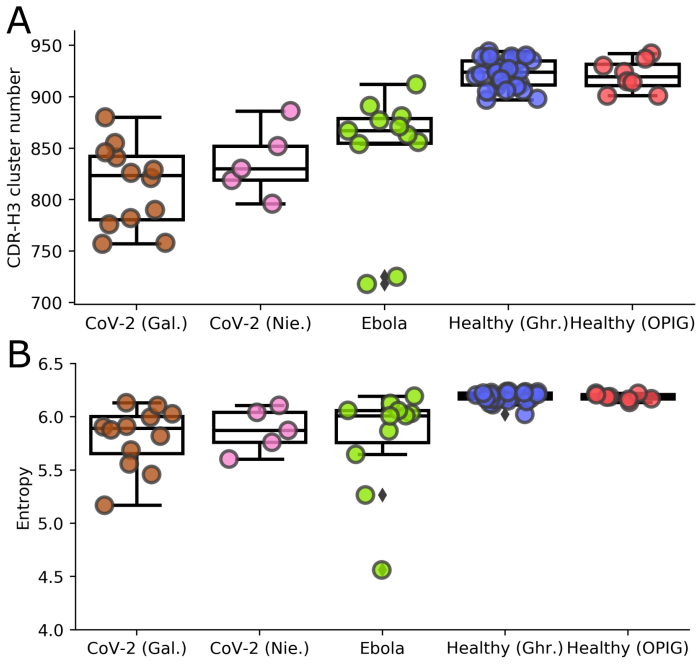


Figure 5.10: **CDR-H3 structural profiles in BCR repertoires.** In each BCR repertoire, CDR-H3 cluster frequencies were reduced to (**A**) the total unique number of utilised CDR-H3 clusters and (**B**) Shannon entropy values.

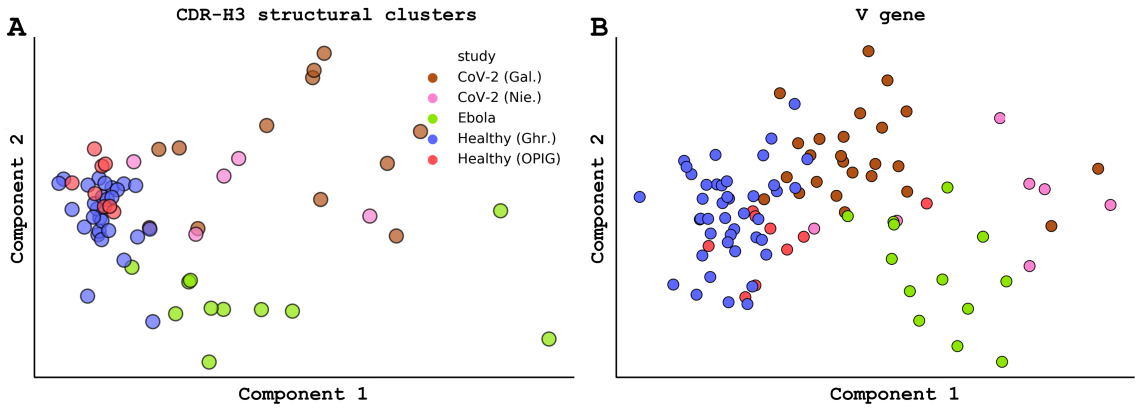


Figure 5.11: **Separation of BCR repertoires based on structural and sequence properties.** Principal component analysis (PCA) was performed on **(A)** CDR-H3 cluster and **(B)** V gene frequencies in SAAB+ annotated BCR repertoires across five independent studies. First two principal components were used to visualise any repertoire separations. Repertoires that have more similar CDR-H3 cluster or V gene usages lie closer together in the PCA space.

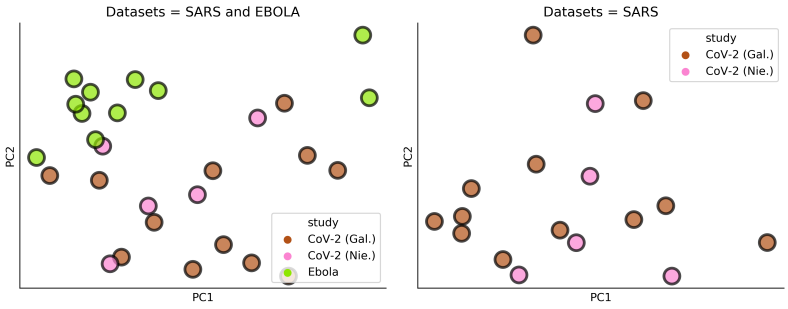


Figure 5.12: **PCA on subsets of BCR studies.** PCA was performed on CDR-H3 cluster frequencies in BCR repertoires across (**A**) SARS-CoV-2 repertoires and Ebola studies or (**B**) two SARS-CoV-2 studies.

We employed principal component analysis (PCA) to separate the SAAB+ annotated datasets based on the pattern of their CDR-H3 shape utilisation. PCA showed that different health states can be effectively distinguished based on the variance of CDR-H3 shape frequencies (Figure 5.11A). Principal component 1 separated healthy from virus stimulated BCR repertoires, whilst principal component 2 distinguished SARS-CoV-2 and Ebola repertoires. Healthy repertoires from two separated studies overlapped and clustered separately from the virus stimulated repertoires. CoV-2 (Nie.) and CoV-2 (Gal.) repertoires also clustered together which is indicative of a confluent structural response observed in these two independent studies. In spite of having similar CDR-H3 structural diversity profiles, Ebola responding BCR reper-

toires clearly separated from healthy and SARS-CoV-2. This suggests a structurally distinct response taking place in the BCR repertoires against Ebola as opposed to SARS-CoV-2 virus. Hence, structural profiling of BCR repertoires could be used to classify datasets based on the past or ongoing immune response.

Previous studies have shown that SARS-CoV-2 BCR repertoires display biased V gene frequencies [8, 198, 275]. Therefore, V gene usage alone could be leveraged as a feature to distinguish between different health states. We confirm that PCA indeed separates BCR studies on V gene usage, however, CDR-H3 cluster usages employed as PCA features generated a finer separation (Figure 5.11B). Incorporating J gene information did not improve the separation, as the gene usage matrix was very sparse.

To check that the observed intra-health state structural commonalities of CDR-H3 loop frequencies were not simply sequencing biases, we performed PCA on SARS-CoV-2 and Ebola repertoires, as well as just SARS-CoV-2 repertoires (Figure 5.12). We observed a clear separation between SARS-CoV-2 and Ebola datasets, whilst PCA could not distinguish between two independent SARS-CoV-2 studies. This confirms the confluent structural response in BCR repertoires against respective immunogens.

Another potential reason for the dataset separations could be the difference in repertoire average CDR-H3 loop lengths. We have already demonstrated that PCA was unable to distinguish between CoV-2 (Nie.) and CoV-2 (Gal.) BCR repertoires in spite of their average CDR-H3 length biases (see Appendix Table E.2). Here, we repeated PCA on all BCR repertoires limiting the analysis to a single CDR-H3 loop length (Appendix Figure E.1). PCA was still able to discriminate between different health states, regardless of the reduced structural space. This evidence reinforces our findings of the converging structural response to SARS-CoV-2 versus Ebola antigens.

5.3.3.4 Structural convergence in SARS-CoV-2 repertoires

We have demonstrated that BCR repertoires respond to the antigenic stimulation by altering their CDR-H3 structure usage and that the pattern of this response is also confluent between BCR repertoires responding to the same immunogen. Here, we investigated which CDR-H3 shapes are over- and under-represented in SARS-CoV-2 BCR repertoire compared to repertoires from healthy individuals and individuals recovering from Ebola. This knowledge of over-represented CDR-H3 geometries can be exploited in *de novo* antibody design and immunodiagnostics.

We found that 23 CDR-H3 shapes were over-represented across SARS-CoV-2 relative to healthy BCR repertoires (Figure 5.13), 17 of which were also over-represented

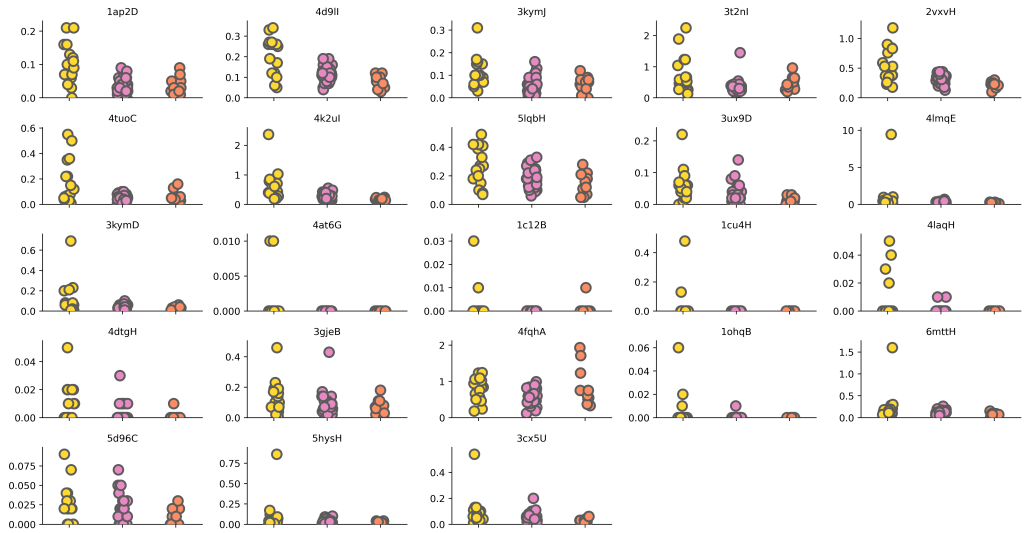


Figure 5.13: **The percentage presence of the over-represented CDR-H3 clusters in the combined SARS-CoV-2, Healthy and Ebola BCR repertoires.** Each subplot shows one of the 23 over-represented CDR-H3 clusters. The x-axis depicts different health states. The y-axis shows the percentage of repertoire sequences whose CDR-H3s were predicted to take the shape of the respective over-represented CDR-H3 cluster. Yellow: SARS-CoV-2; Purple: Healthy; Orange: Ebola

against Ebola repertoires (p-value<0.05, Mann-Whithney U-test). This poses a potential structurally distinct response against the SARS-CoV-2 immunogen. Interestingly, 422 CDR-H3 shapes were under-represented in SARS-CoV-2 repertoires when compared to healthy repertoires. This again confirms the structural specialisation, as the repertoires prioritise utilisation of a smaller number of CDR-H3 structures.

In our CDR-H3 structural library only three clusters contained loops from SARS-CoV-2 binding antibodies. None of these clusters were present in the over-represented CDR-H3 cluster set. This confirms that the polyclonal antibody response is elicited against COVID-19 across multiple patients, where different coronavirus epitopes are targeted. Further, a sample size of just three CDR-H3 clusters is not representative of all SARS-COV-2 antigen complementary CDR-H3 geometries. Hence, as more solved SARS-CoV-2 binding antibodies become available, the accuracy of this analysis will improve.

5.3.3.5 Structural divergence in canonical CDR loops

During B-cell affinity maturation in response to antigenic stimulation, nucleotide SHMs are iteratively introduced primarily to antibody CDR loop regions [63]. These mutations could induce structural changes in CDRs, generating a diverse library of

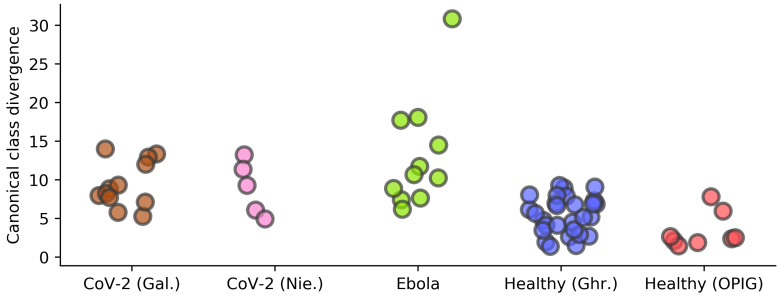


Figure 5.14: **The degree of canonical CDR loop divergence across five independent studies.** Canonical class divergence was defined as a mismatch in either CDR-H1 or CDR-H2 with the respective loops encoded by the same germline V gene as the target sequence. To obtain the degree of canonical class divergence, the number of sequences with loop class mismatches were divided by the total number of sequences in the IGH repertoire multiplied by 100.

paratope configurations in BCR repertoires. Positive selection then filters and expands B-cells whose BCR structures are both chemically and geometrically complementary against the immunogen. Thus, understanding of structural diversity of the canonical CDR-H1/CDR-H2 is also important, as the canonical CDR residues form key chemical interactions with target antigens [21].

Here, we investigated whether backbone geometries of canonical CDR loops in BCR repertoire sequences differed from those encoded by their patent germline V genes as described in Chapter 4 (Section 4.3.5). The degree of this structural divergence can be used as a proxy to study BCR repertoire maturation [153]. All interrogated repertoires contained some sequences whose canonical classes differed from those encoded by their parent germline genes (Figure 5.14). However, the degree of the canonical class divergence was not uniform across the health states. The average degree of the divergence was the lowest in healthy repertoires (Healthy (OPIG) and Healthy (Ghr.)) (Appendix Table E.4). Virus stimulated datasets had a significantly higher degree of the divergence when compared to the healthy repertoires. This provides structural evidence that antigenic stimulation induces structural changes in canonical CDR loops in IGH repertoires to improve their chemical complementary against the immunogen.

5.4 Discussion

Since the first reports of a newly-emerged SARS-CoV-2 virus in December 2019 [267], a colossal body of research has been carried out to study the impact of the virus on the human body [268] and to find potential therapeutic interventions [221]. To help the research community in this global effort, we have developed and applied a wide range of tools and databases to investigate the molecular characteristics of SARS-CoV-2-binding antibodies as well as SARS-CoV-2 stimulated BCR repertoires.

As of October 2020, we have collated more than 1,200 SARS-CoV-2-binding antibodies from the scientific literature. Each of these binders holds a valuable piece of molecular descriptor information about SARS-CoV-2 antigen complementarity. Annotating all these antibodies in a standardised manner could potentially reveal which paratope motifs drive neutralisation of the virus. The development of CoV-AbDab has enabled researchers to perform a wider range of analyses. One of the immediate applications of CoV-AbDab is the assessment of the adaptive immune response elicited by COVID-19 vaccine candidates. As CoV-AbDab contains both COVID-19 binding and neutralising antibodies, a unified enrichment of the known neutralising sequences in BCR repertoires across all post-vaccinated human donors would be indicative of an effective and protective immune response.

Deciphering human COVID-19 BCR repertoires will shed light on the response commonality shared across multiple seropositive patients. If linked with a good prognosis, this shared response holds potential to be a good starting point for a therapeutic antibody discovery campaign [282]. Furthermore, correlating clonal sequence presence with patient's clinical outcome could be used in initial patient stratification, allowing more appropriate treatment to be administered earlier in the response [8, 276]. Here, we investigated the immune response similarity to COVID-19 in patients from distinct geographic locations. The clonal overlap was leveraged as a measure of repertoire functional convergence. We found a significantly increased clonal overlap between two independent SARS-CoV-2 BCR studies [8, 198] when compared to antigenically-non-stimulated healthy BCR repertoires suggesting the development a COVID-19 cross-strain immunity. This knowledge can now be used as a reference to focus on a narrower set of clones for the development of a potential COVID-19 therapy.

Finally, we also studied structural dynamics of canonical CDR and CDR-H3 loops in COVID-19 BCR repertoires. By mapping CDR-H3 loop sequences to a set of crystallographically solved antibodies, we have shown that SARS-CoV-2 BCR repertoires

utilise more similar CDR-H3 shape frequencies than in healthy or other unrelated disease states. To our knowledge this is the first evidence of structural convergence in CDR-H3 backbone geometries in response to COVID-19 at the BCR repertoire level. We have also confirmed that SARS-CoV-2 repertoires display similar general CDR-H3 structural features normally associated with disease stimulated repertoires: a reduced number of sampled CDR-H3 structures and lowered entropy amongst these loop structure usage [153]. We identified 23 CDR-H3 structures that were significantly over-represented in SARS-CoV-2 BCR repertoires. Hence, BCRs adopting these shapes have a higher chance of being antigen specific. We have also shown that canonical CDRs changed their shapes in response to COVID-19 infection, and the rate of this change was significantly higher than in non-stimulated BCR repertoires. Hence, researchers may benefit from considering all IGH CDR loops when engineering COVID-19 neutralising antibodies.

As our structural analysis was performed in June 2020, only a handful of SARS-CoV-2 binding antibodies were available at that time. As of October 13$^{\text{th}}$, 28 of such antibodies are now deposited in SAbDab [134]. This prompts the repeat of structural annotation analysis as it should provide an even clearer picture on CDR structural convergence in SARS-CoV-2 repertoires.

In the final chapter of this thesis, I will present future directions in my research.

6 Future work

In this DPhil work, we have described the development and validation of two novel computational pipelines (ABOSS and SAAB+) for orthogonal structure-focused BCR repertoire error-profiling and analysis. We have also reported the creation of two novel databases (OAS and CoV-AbDab). OAS is the first database that contains more than 1.9 billion naturally observed BCR sequences which were gathered, cleaned and annotated from 85 studies. CoV-AbDab curates all publicly released SARS-CoV and SARS-MERS specific antibodies providing the molecular description of virus recognition by the adaptive immune system. We have then used these tools to investigate structural and sequence diversities across SARS-CoV-2 repertoires.

Below we outline several research avenues that can be further explored based on the results reported in this thesis work.

In silico B-cell sub-population sorting

Experimental B-cell sub-population sorting is expensive and labour intensive. Hence, the development of computational B-cell sorting tools would alleviate experimental costs and improve analysis resolution. Recent advances in deep learning methods enabled researchers to successful classify antigen-stimulated immune repertoires [205, 283]. In collaboration with Dr. Johannes Trück, we generated high quality sub-population sorted BCR repertoires. These datasets bear high quality sequence representations of each B-cell developmental compartment. There is the potential to use these datasets to successfully train current state of the art deep learning algorithms to distinguish different B-cell types or repertoire functional groupings. Sub-population sorted data which is deposited in OAS could be used as an additional source to minimise undesired model over-fitting. The performance of trained models could then be

tested against simple sequence features such as SHM cut-offs.

Paired BCR data

Structural analyses performed throughout this thesis were focused on deciphering unpaired heavy chain BCR data. The availability of natively paired VH/VL repertoire data opens new avenues for structural diversity analyses. As the current sequencing depth of the 10xGenomics V(D)J technology only covers up-to eighty thousand sequences, it permits structural transformation of complete BCR repertoire samples with rapid homology modelling tools such as ABodyBuilder [41]. These structural models will enable better functional repertoire grouping. The immediate applications of this work will be structural profiling of BCR repertoires across different species including genetically modified to expand our understanding of repertoire "humanness" as well as studying antibody structural integrity when κ and λ chains are swapped.

Upgrading OAS

Current OAS updates are not automated due to non-standardised deposition of BCR repertoire study metadata. Hence, bespoke scripts are required to incorporate each additional study into OAS. To automate OAS, non-trivial efforts will be needed to extract key metadata information by introducing and validating state of the art text mining algorithms.

Current OAS architecture only provides BCR data unit downloads. To expand OAS functionalities, we will need to re-organise OAS data deposition into non-relational database formats such as MongoDB. This will enable rapid sequence match queries across the entire resource. Further, repertoire statistics (e.g. CDR-H3 lengths, V and J gene usages) will be available for each OAS Data Unit. In the paired version of OAS, we will integrate ABodyBuilder [41] to provide a quick way to structurally visualise and compare naturally observed BCRs.

A

B-cell developmental stages

A.1 Pro-B-cells

In the first stage, hematopoietic precursor cells (HSCs) give rise to pro-B-cells in the red bone marrow. Pro-B-cells do not yet have functional BCRs on their plasma membrane. Two kinases (Recombination-activating gene (RAG-1) and RAG-2) start somatic VDJ gene recombination in the heavy chain locus by initially joining a D_H and a J_H gene segment (Figure 1.6B). All DNA bases that lie between these two genes are deleted from the B-cell's genome using RAG transposon motifs [284].

A.2 Large Pre-B-cells

Pro-B-cells transition into the large pre-B-cell stage when a single gene from the variable (V) heavy chain locus recombines with the D_H-J_H gene assembly. The resultant VDJ complex now represents B-cell's VH domain. During primary transcription, VDJ complex is joined with the first gene from the constant (C) heavy locus encoding for the IGHM isotype to yield antibody heavy chain (HC) mRNA. This newly recombined HC is denoted as Igμ. The pre-BCR complex can now be formed on the B-cell plasma membrane by combining the Igμ protein with the surrogate light chain (Figure 1.6C) [285]. All pre-B-cells utilise the same surrogate light chain for each Igμ molecule. The surrogate light chain is the product of VpreB1 and λ5 genes that share high sequence homology to human V_λ and J_λ -C_λ genes respectively [286]. Large pre-B-cells require a proliferative stimulus in the form of an antigen-binding signal via pre-BCRs in order to survive. However, a pre-BCR is unable to transmit this signal alone due to its short cytoplasmic Igμ domain [287]. Hence, the pre-BCR forms a complex with a heterodimeric signalling molecule consisting of Ig-α and Ig-β

proteins. B-cells, whose pre-BCRs are responsive to foreign antigens, are positively selected and proliferate into small pre-B-cells [56].

A.3 Small Pre-B-cells

In the small pre-B-cell developmental stage, RAG-1 and RAG-2 recombine V_L and J_L gene segments to form the variable light (VL) domain. The C gene is then added to the VL segment during primary transcription to generate light chain (LC) mRNA (Figure 1.6D). Once translated, this LC replaces the surrogate light chain on the pre-BCR complex. A functional human LC can be encoded by either κ or λ loci found on two separate chromosomes: 2 and 22 respectively. First, LC_κ is added to Igμ to form a functional IGHM/IGK BCR. If the resultant complex is non-productive, several other V_κ and J_κ re-arrangements are attempted. If this still fails, the λ locus is employed to synthesise the LC_λ. If the HC/LC_λ assembly is productive, it leads to IGHM/IGL BCR formation. B-cells with functional BCRs are now called immature B-cells. B-cells whose HC and LC molecules fail to combine into functional BCRs undergo apoptosis.

A.4 Immature B-cells

The fine-tuned regulation by RAG-1 and RAG-2 ensures than only one unique HC and LC are somatically recombined in each individual B-cell (allelic exclusion) [284]. A recent research suggested that many B-cells can yield multiple productive BCR that originate from different somatic V(D)J recombination processes [218]. However, the exact molecular mechanism of multi-immunoglobulin specificity is not yet understood [218].

The immature B-cells are profiled against self-reactivity before exiting the bone marrow (Figure 1.6E). A B-cell, whose BCR is non- or weakly-reactive against self-antigens transition to secondary lymphoid organs as mature naive B-cells, whilst other B-cells undergo apoptosis or "light chain rescue" [56]. At this developmental stage, the immature B-cells also co-express both IGHD and IGHM BCR isotypes on the plasma membrane through alternative splicing.

Appendix Chapter 2

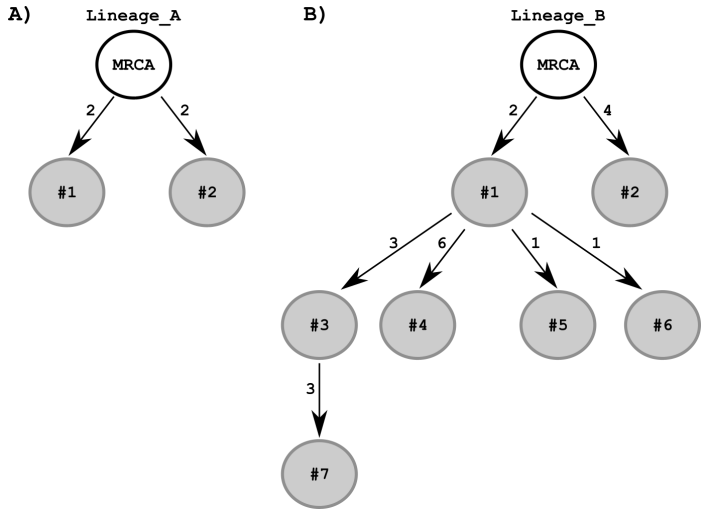


Figure B.1: **Architecture of Lineage A and Lineage B.** Two clonal tree references (Lineage_A and Lineage_B) from the HH_S5F-targeting model [62] of SHMs were used to introduce mutations into the BCR sequences. Both lineages require the most recent common ancestor (MRCA) sequence as the input. Running the HH_S5F-targeting model outputs progenitor B-cells (grey circles), whose number differ between the lineage architectures. The number of introduced SHMs is indicated by the edge value. **A)** Lineage_A yields two progenitor B-cells. Both these cells harbour extra two nucleotide mutations in their V(D)J region with respect to the MRCA. **B)** Lineage_B outputs seven progenitor B-cells in total. The progenitor B-cell #1 further undergoes a second round of SHMs yielding four progenitors (#3 - #6). The progenitor #3 undergoes the final round of SHMs producing the progenitor #7. The average mutational load in Lineage_A and Lineage_B with respect to the MRCA is 2 and 4.7 nucleotides respectively.

C

Appendix Chapter 3

Study	Species	Disease	Vaccine	B Cell Source	B Cell Subset	Sequence #
[220]	mouse_C57BL/6J	None	None	Spleen Lymph	Unsorted-B-Cells	22414 (12041)
[288]	human	HIV	None	PBMC	Unsorted-B-Cells	396793 (203672)
[289]	human	HIV	None	PBMC	Unsorted-B-Cells	876704 (175907)
[290]	human	HIV	None	PBMC	Unsorted-B-Cells	1294536 (759867)
[291]	human	HIV	None	PBMC	Unsorted-B-Cells	1424818 (631637)
[124]	human	CLL None	None	PBMC	Unsorted-B-Cells	188030 (133750)
[256]	human	None	Flu	PBMC	Naive-B-Cell/Plasmablast	3508010 (2328253)
[292]	mouse	None	NP-CGG	Bone-Marrow Spleen Small-Intestinal	Unsorted-B-Cells	149842 (51047)
[293]	human	HIV	None	PBMC	Unsorted-B-Cells	1544776 (468277)
[294]	mouse_BALB/c	None	NP-CGG	Spleen/Bone-Marrow	ASC	14372771 (7853042)
[22]	human	Dengue None	None	PBMC	Unsorted-B-Cells	42889 (38360)
[214]	mouse_BALB/c	None	NP-CGG	Spleen/Bone-Marrow	Plasmablast/Plasma-B-Cells	7961636 (2902204)
[295]	human	MS	None	CSF PBMC	Unsorted-B-Cells	783703 (298904)

Continued on next page

Table C.0 – continued from previous page

Study	Species	Disease	Vaccine	B Cell Source	B Cell Subset	Sequence #
[296]	human	MS	None	Cervical-Lymph-Node White-Matter-Lesion Cortex Choroid-Plexus Pia-Mater Spleen	Unsorted-B-Cells	8569409 (3394167)
[200]	human	Allergic-Rhinitis None	None	PBMC/Nasal-Biopsy	Unsorted-B-Cells	37754 (25734)
[234]	human	None	None	PBMC	Mature-B-Cells Immature-B-Cells Transitional-B-Cells Plasmacytes Memory-B-cell-IgD-negative Memory-B-cell-IgD-positive	107811 (90290)
[209]	human	HIV None	None	PBMC	Unsorted-B-Cells	12739693 (5570289)
[297]	rhesus	None	HIV	PBMC	Unsorted-B-Cells	44408 (29033)
[298]	mouse_C57BL/6	None	None	Splenocytes	Unsorted-B-Cells	815484 (248620)
[299]	mouse_outbred C57BL/6	None	E.Coli	Biopsy-Small-Intestine	Unsorted-B-Cells	1690732 (547518)
[185]	human	None	HepB None	PBMC	HepB+B-cells HLA-DR+-Plasma-B-Cells Unsorted-B-Cells	27438826 (15256345)
[300]	human	None	Tetanus/Flu	Bone-Marrow	Plasma-B-Cells	2414404 (1282996)
[301]	human	HIV	None	PBMC	Unsorted-B-Cells	724633 (295924)

Continued on next page

Table C.0 – continued from previous page

Study	Species	Disease	Vaccine	B Cell Source	B Cell Subset	Sequence #
[23]	human	None	MenACWY-conjugate MenACWY-polysaccharide None	PBMC	Naive-B-Cells Memory-B-Cells HLA-DR–Plasma-B-Cells HLA-DR+-Plasma-B-Cells Marginal-Zone-B-Cells	8259743 (3552745)
[302]	human	HIV	None	PBMC	Unsorted-B-Cells	5564693 (1502752)
[303]	human	SLE None	Tetanus Flu None	PBMC	Unsorted-B-Cells	29411784 (14240874)
[88]	mouse_BALB/c	None	NP-CGG	Spleen	ASC Plasma-B-Cells Naive-B-Cells	789450 (529867)
[304]	human	Allergy+NoSIT Allergy+SIT	None	Nasal-Biopsy PBMC	Unsorted-B-Cells	668937 (474150)
[305]	human	HIV	None	PBMC	Unsorted-B-Cells	791973 (190087)
[306]	human	HIV	None	PBMC	Unsorted-B-Cells	2172582 (562706)
[307]	human	HIV	None	PBMC	Memory-B-Cells	12304225 (6763202)
[308]	human	None	pH1N1-AS03 pH1N1 TIV	PBMC	Plasma-B-Cells	14583670 (5474719)

Continued on next page

Table C.0 – continued from previous page

Study	Species	Disease	Vaccine	B Cell Source	B Cell Subset	Sequence #
[309]	human	None	Flu None	PBMC	Unsorted-B-Cells HA-Negative-Activated-B-Cells Naive-B-Cells Memory-B-Cells ASC Activated-B-Cells Resting-Memory-B-Cells HA-Positive-Activated-B-Cells HA-Positive-ASC HA-Negative-ASC HA-Positive-Memory-B-cells HA-Negative-Memory-B-cells	13910585 (7281539)
[310]	human	None	None	PBMC	Naive-B-Cells Memory-B-Cells	2415261 (1778662)
[120]	mouse_BALB/c	None	None OVA	Spleen	Unsorted-B-Cells	24190036 (8415955)
[176]	human	None	None	PBMC	Plasma-B-Cells Naive-B-Cells Memory-B-Cells	192002 (183248)
[311]	human	HIV	None	PBMC	Unsorted-B-Cells	1965275 (542812)
[312]	human	HIV None	None	PBMC	Unsorted-B-Cells Memory-B-Cells	230658 (69574)
[313]	human	None	None	PBMC	Unsorted-B-Cells	2765530 (1470267)
[314]	rhesus human mouse_C57BL/6 mouse_BALB/c	None	None	PBMC	Unsorted-B-Cells	5345396 (2887322)
[315]	mouse_BALB/c	None	NP-CGG None	Splenocytes-Non-Germinal Splenocytes-Germinal	Memory-B-Cells	5523153 (1033896)

Continued on next page

Table C.0 – continued from previous page

Study	Species	Disease	Vaccine	B Cell Source	B Cell Subset	Sequence #
[89]	human	None	HepB None	PBMC	HepB+B-cells HLA-DR+-Plasma-B-Cells Unsorted-B-Cells	22612035 (11124229)
[316]	rabbit	None	HIV	PBMC Spleen	Unsorted-B-Cells	4360370 (3007885)
[109]	human	MuSK-MG None/ AChR-MG	None	PBMC	Unsorted-B-Cells Naive-B-Cells Memory-B-Cells	13993268 (5469768)
[94]	mouse_C57BL/6J mouse_BALB/c mouse_outbred	None	HepB OVA NP-HEL None	Bone-Marrow Spleen	Plasma-B-Cells Pre-B-Cells Naive-B-Cells	246988884 (137343535)
[68]	camel	None	None	PBMC	Unsorted-B-Cells	1211314 (1188969)
[317]	mouse_BALB/c	None	Plasmodium None	Spleen	Unsorted-B-Cells	178197 (123514)
[318]	mouse_Ighe/e mouse_Ighg/g mouse_Igh/wt	None	OVA	Bone-Marrow Spleen	Follicular-B-Cells Pro-B-Cells	93132 (57121)
[319]	human	Allergy+NoSIT	None	Bone-Marrow PBMC	Unsorted-B-Cells	32892614 (12615839)
[111]	human	CMV/EBV EBV None	None	Mesenteric-lymph-node Colon Spleen Bone-Marrow Jejunum Ileum PBMC Lung	Unsorted-B-Cells	56040120 (26957547)

Continued on next page

Table C.0 – continued from previous page

Study	Species	Disease	Vaccine	B Cell Source	B Cell Subset	Sequence #
[320]	human	None	HepB/HepA/Flu Flu None	PBMC	Unsorted-B-Cells	25285974 (10842819)
[321]	human	HIV	None	PBMC	Unsorted-B-Cells	11970928 (6397501)
[322]	human	None	None	PBMC	Unsorted-B-Cells	14201361 (6007017)
[63]	human	None	None	PBMC	Unsorted	2860653 (859131)
[323]	rat	None	DNP HuD	Splenocytes	Unsorted-B-Cells	6359396 (4292287)
[210]	mouse_C57BL/6J	None	None	Splenocyte Spleen	Unsorted-B-Cells	41908 (27864)
[324]	human	None HCV	None	PBMC	Unsorted-B-Cells	4445345 (462817)
[325]	mouse_C57BL/6	None	None	Spleen Peritoneum	B-2 B-1a B-1b Marginal-zone-B-Cell Follicular-B-Cells	345388 (207521)
[326]	human	Asthma None	None	PBMC Biopsy	Unsorted-B-Cells	24778181 (7088032)
[327]	human HIS-mouse	HIV None	Sheep-erythrocytes Plasmodium None	Cord-Blood PBMC Splenocytes	Unsorted-B-Cells	154568609 (59857938)
[122]	human	HIV None	None	PBMC	Unsorted-B-Cells	46179809 (19790661)
[328]	human	HIV None	None	PBMC	Unsorted-B-Cells Naive-B-Cells Memory-B-Cells	9620172 (7758765)

Continued on next page

Table C.0 – continued from previous page

Study	Species	Disease	Vaccine	B Cell Source	B Cell Subset	Sequence #
[329]	human	Allergy	None	Bone-Marrow PBMC	Unsorted-B-Cells	32952700 (12613741)
[330]	human	Healthy/celiac-disease	None	PBMC	Naive-B-Cells	52740686 (23759161)
[331]	human	HIV	TIV	PBMC	Unsorted-B-Cells	110018455 (15442884)
[107]	human	None	None	PBMC	Unsorted-B-Cells	28905062 (6187148)
[3]	human	None	None	LeukoPak	Unsorted-B-Cells	220776486 (161856393)
[332]	human	None	None	PBMC	Unsorted-B-Cells	3849765 (739696)
[4]	human	None	None	LeukoPak	Unsorted-B-Cells	318693668 (227728314)
[247]	human	Ebola	None	ASC PBMC	Unsorted-B-Cells	22306805 (12201598)
[198]	human	SARS-COV-2	None	PBMC	Unsorted-B-Cells	8995778 (2929692)
[278]	human	None SARS-COV-2	None	PBMC	Unsorted-B-Cells	7392066 (4720776)
[174]	human	HIV-Broad-Neutralizing HIV-Non-Neutralizing None	None	PBMC	Unsorted-B-Cells	132733654 (75175738)
[333]	human	SARS-COV-2	None	PBMC	Unsorted-B-Cells	3898647 (2065516)
[121]	human	Tonsillitis obstructive sleep apnea	None	Tonsillectomy	Unsorted-B-Cells Naive-B-Cells Memory-B-Cells Germlinal-Center-B-Cells Plasmablast	1457867 (1155955)

Continued on next page

Table C.0 – continued from previous page

Study	Species	Disease	Vaccine	B Cell Source	B Cell Subset	Sequence #
[334]	human	HIV None	None	PBMC	Unsorted-B-Cells	3585977 (2098787)
[276]	human	None SARS-COV-2	None	PBMC	Unsorted-B-Cells	15616914 (8753737)
[190]	human	Light Chain Amyloidosis	None	Bone-Marrow	Unsorted-B-Cells	10351232 (2691246)
[335]	human	POEMS	None	Bone-Marrow Bone-Lession	Unsorted-B-Cells	6306038 (1831518)
[9]	human	None SARS-COV-2	None	PBMC	Unsorted-B-Cells	31030626 (10030156)

Table C.1: The datasets are organized into studies related to a given Ig-seq experiment. Each study in the OAS database is subdivided into Data Units. Each Data Unit is a collection of IMGT-numbered amino acid sequences uniquely identified by the metadata descriptors given in Table 3.1; five of which (species, disease, vaccine, B-cell source, and B-cell type) are given in this table. The "Sequence #" column indicates the total number of redundant sequences in our database, with the nonredundant numbers in parentheses.
ABC, activated B-cell; ASC, Antibody secreting cell; CLL, chronic lymphocytic leukemia; DNP, keyhole limpet hemocyanine modified with dinitrophenyl; Flu, influenza; HuD, paraneoplastic encephalomyelitis antigen; MG, myasthenia gravis; MS, multiple sclerosis; NP-CGG, chicken gamma globulin; NP-HEL, hen egg lysozyme; SLE, systemic lupus erythematosus.

CST Name	Target	CST Name	Target
Enfortumab	Poliovirus-receptor-like 4	Amatuximab	Mesothelin
Tabalumab	Tumor necrosis factor ligand superfamily member 13B	Lorvatuzumab	CD56
Ascrinvacumab	Serine/threonine-protein kinase receptor R3	Bimagrumab	Activin receptor type-2B
Rilotumumab	Hepatocyte growth factor	Solanezumab	Beta amyloid
Brodalumab	IL17A	Camrelizumab	Programmed cell death protein 1
Ramucirumab	Kinase insert domain receptor	Crenezumab	Beta amyloid
Zanolimumab	CD4	Quilizumab	IGHE
Foravirumab	Rabies virus glycoprotein	Obiltoxaximab	Anthrax Protective Antigen
Dusigitumab	Insulin-like growth factor 2	Coltuximab	CD19
Ublituximab	B-lymphocyte antigen CD20	Landogrozumab	growth differentiation factor 8
Dectrekumab	IL13	Daclizumab	IL2RA
Sifalimumab	Interferon alfa	Visilizumab	CD3
Brentuximab	CD30	Lintuzumab	CD33
Radretumab	Fibronectin	Simtuzumab	Lysyl oxidase homolog 2
Canakinumab	IL1B	Tildrakizumab	IL23A
Utomilumab	CD137	Sacituzumab	Tumor-associated calcium signal transducer 2
Tesidolumab	Complement pathway protein C5	Etrolizumab	$\alpha 4\beta 7$
Glembatumumab	Transmembrane glycoprotein NMB	Andecaliximab	Gelatinase B
Iratumumab	CD30	Palivizumab	Respiratory syncytial virus
Abagovomab	Cancer antigen 125	Vadastuximab	CD33
Sarilumab	IL6R	Natalizumab	Alpha-4 integrin
Timolumab	Amine oxidase, copper containing 3	Pinatuzumab	CD22
Evolocumab	Proprotein convertase subtilisin/kexin type 9	Eptinezumab	Calcitonin
Elotuzumab	CD319	Inotuzumab	CD22
Emibetuzumab	Hepatocyte growth factor receptor	Certolizumab	Tumor necrosis factor alpha
Nivolumab	Programmed cell death protein 1	Reslizumab	IL5
Indatuximab	Syndecan 1	Farletuzumab	Folate receptor 1

Table C.2: **Targets of Clinical stage therapeutics (CST) with 100% CDR-H3 sequences matches to the OAS data.**

Appendix Chapter 4

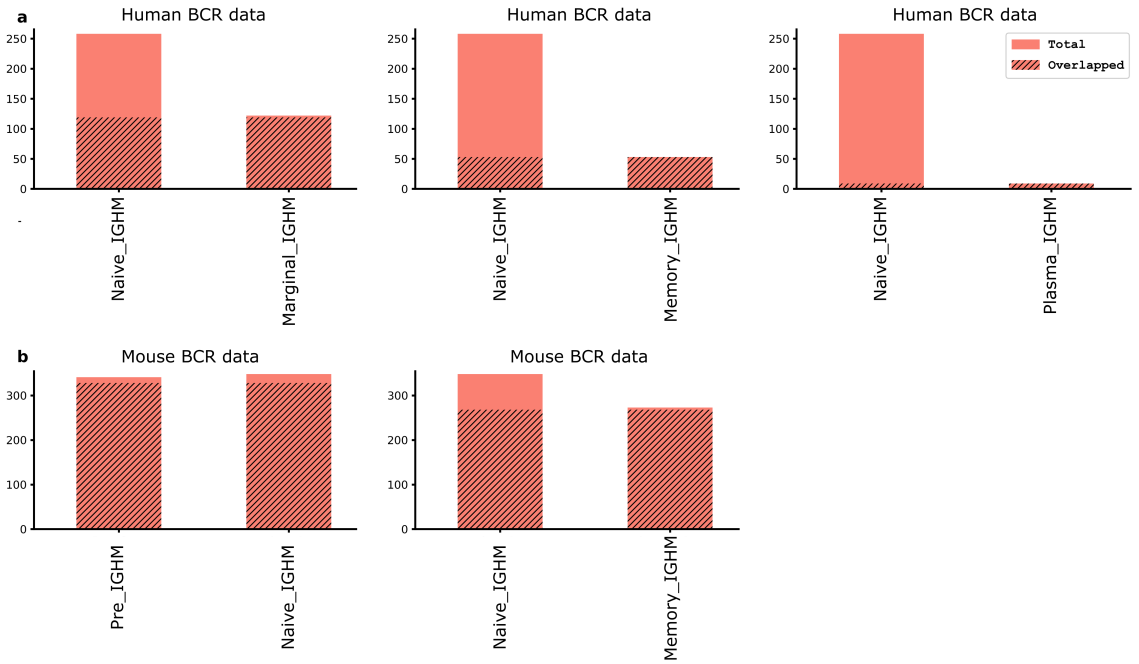


Figure D.1: **Overlap of Structural Stem CDR-H3 clusters between naive and antigen experienced BCR repertoires in the human and mouse data.** naive and antigen experienced BCR repertoires were investigated for the Structural Stem overlap in the human **(row A)** and mouse **(row B)** data. The overlap was defined as a number of shared clusters between Structural Stem CDR-H3 clusters in BCR repertoires found in two different B-cell types. The X-axis shows the B-cell types. The Y-axis shows the total number of Structural Stem CDR-H3 clusters. The stripped pattern indicates the number of overlapped Structural Stem CDR-H3 clusters.

E

Appendix Chapter 5

Study	BCR repertoire sequenced	BCR repertoire analysed
CoV-2 (Gal.)	25	12
CoV-2 (Nie.)	5	5
Ebola	13	11
Healthy (Ghr.)	42	28
Healthy (OPIG)	9	8

Table E.1: **Number of BCR repertoires used in the structural analysis.** Each SAAB+ annotated BCR repertoire was check for sufficient structural coverage. Only repertoires containing $\geq$ 10,000 SAAB+ annotated CDR-H3 sequences were retained.

Study	Percentage CDR-H3 modellable	Mean CDR-H3 length
Healthy (Ghr.)	54.89 ± 2.58	12.98 ± 0.12
Healthy (OPIG)	55.98 ± 1.50	13.12 ± 0.1
SARS-CoV-2 (Gal.)	51.44 ± 2.17	13.17 ± 0.17
SARS-CoV-2 (Nie.)	55.1 ± 1.53	12.89 ± 0.16
Ebola	58.14 ± 6.28	12.6 ± 0.37

Table E.2: **CDR-H3 structural modellability.** The percentage CDR-H3 modellability for each analysed BCR repertoire study and the mean CDR-H3 length over this SAAB+ modellability range (± 1 standard deviation).

Study	Cluster Number	Shannon Entropy
CoV-2 (Gal.)	813±40	5.8±0.29
CoV-2 (Nie.)	837±34	5.88±0.2
Ebola	847±64	5.80±0.49
Healthy (Ghr.)	923±13	6.19±0.04
Healthy (OPIG)	921±15	6.19±0.03

Table E.3: **Summary of CDR-H3 structural profiles in BCR repertoires.** In each BCR repertoire, CDR-H3 shape frequencies were reduced to (**A**) the number of utilised CDR-H3 clusters and (**B**) Shannon entropy values. The values shown in the table are the averages for each metric plus standard deviation in respective BCR studies.

Study	Divergence
CoV-2 (Gal.)	9.37±2.98
CoV-2 (Nie.)	8.97±3.49
Ebola	13.08±7.12
Healthy (Ghr.)	5.43±2.37
Healthy (OPIG)	3.33±2.27

Table E.4: **Summary of canonical class divergence.** Canonical class divergence from the germline encoded CDR shapes was calculated across BCR studies. Mean divergence estimates (±s.t.d) were calculated for each study.

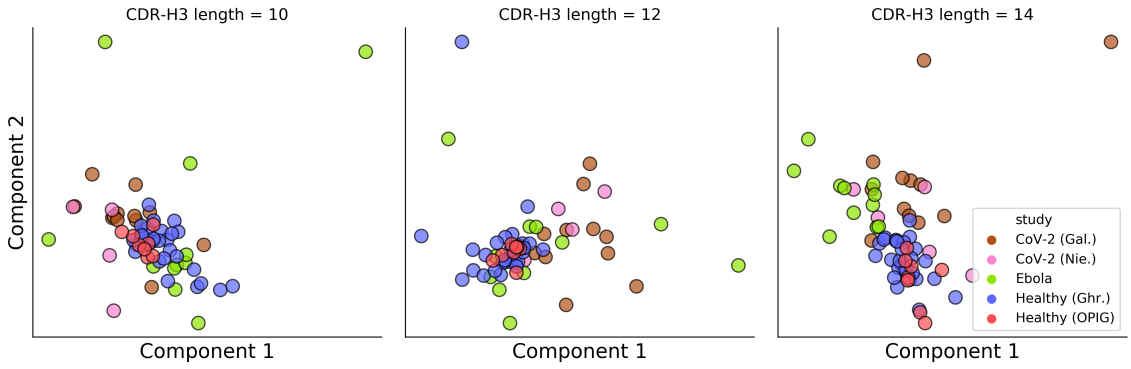


Figure E.1: **PCA on CDR-H3 cluster frequencies in BCR repertoires with sequences filtered for a singular CDR-H3 loop length.** Each BCR repertoire was stratified into three subrepertoires which were limited to a single CDR-H3 loop length: 10, 12 or 14.

Glossary

Alchemab clones A set of 1,254 convergent B-cell clonotypes that were found in at least four COVID-19 patients in the Alchemab dataset. .

Alchemab dataset A set of BCR repertoires that was obtained from 31 COVID-19 infected patients (Galson *et al.,* [8]). The mean time since the symptom onset was 10.8 days in the patient group. V(D)J cDNA was extracted from unsorted PBMCs (20mL blood sample) and used to sequence BCR repertoires.

Antibody PDB template A structure of an antibody region which is obtained from the PDB and used by the homology modelling software to predict antibody 3D conformation. Each template contains antibody atom positions in the 3D space.

CDR-H3 structural cluster A set of CDR-H3 loop structures whose backbone residue RMSD distances are found within 0.6 Å. The pairwise RMSD distances are calculated using the DTW algorithm.

Dynamic Time Warping (DTW) An algorithm that calculates the pairwise similarity between temporal sequences that have the same starting and end coordinates but the varying number of time points. DTW can be applied to study the structural similarity (RMSD) between antibody loops with mismatching lengths, as the loops have the same starting and end point coordinates (anchor residues), and the length of the amino acid sequence can discretised into separate time points.

ESS score A score that is assigned to a selected PDB loop template by FREAD based on its sequence/structural homology to the target loop. Higher scores indicate better matches to the pre-build library of the PDB loop templates. FREAD yields ESS scores by leveraging the pre-computed matrix of observed amino acid substitution frequencies in a protein evolutionary environment.

Nielsen dataset A set of BCR repertoires that was obtained from 6 COVID-19 infected patients (Nielsen *et al.,* [198]). The mean time since the symptom onset was 11.3 days in the patient group. Unsorted PBMCs were used as the source of genetic information to take a snapshot of the BCR repertoire diversity. Only repertoires generated from V(D)J cDNA were used in our work.

OAS data unit A set of processed BCR sequences that share the same set of metadata: isotype, B-cell source and type, donor's health information, *etc.* Each sequence in a Data Unit is further associated with sequence-specific annotations generated by ANARCI and IgBlastn.

RMSD calculation The RMSD of the antibody atom positions provides a pairwise measurement of the three-dimensional similarity between two sets of coordinates where solved or predicted structures are available. The RMSD is usually calculated using the Euclidean distance on the backbone atoms, but this calculation can be performed including the side chain atoms as well.

Sequence redundancy Two sequences are considered redundant if they have identical length and identical amino acid composition in the processed BCR dataset.

References

1. Wang, W., Erbe, A. K., Hank, J. A., Morris, Z. S. & Sondel, P. M. *NK cell-mediated antibody-dependent cellular cytotoxicity in cancer immunotherapy* 2015.
2. Perez-Andres, M. *et al.* Human peripheral blood B-cell compartments: a crossroad in B-cell traffic. *Cytometry Part B: Clinical Cytometry* **78,** S47–S60 (2010).
3. Soto, C. *et al.* High frequency of shared clonotypes in human B cell receptor repertoires. *Nature* **566,** 398–402 (2019).
4. Briney, B., Inderbitzin, A., Joyce, C. & Burton, D. R. *Commonality despite exceptional diversity in the baseline human antibody repertoire* Feb. 2019.
5. Collis, A. V., Brouwer, A. P. & Martin, A. C. Analysis of the antigen combining site: correlations between length and sequence composition of the hypervariable loops and the nature of the antigen. *Journal of molecular biology* **325,** 337–354 (2003).
6. Georgiou, G. *et al. The promise and challenge of high-throughput sequencing of the antibody repertoire* Feb. 2014.
7. Reddy, S. T. *et al.* Monoclonal antibodies isolated without screening by analyzing the variable-gene repertoire of plasma cells. *Nature biotechnology* **28,** 965–969 (2010).
8. Galson, J. D. *et al.* Deep sequencing of B cell receptor repertoires from COVID-19 patients reveals strong convergent immune signatures. *bioRxiv* (2020).
9. Montague, Z. *et al.* Dynamics of B-cell repertoires and emergence of cross-reactive responses in COVID-19 patients with different disease severity. *arXiv preprint arXiv:2007.06762* (2020).
10. Steffen, U. *et al.* IgA subclasses have different effector functions associated with distinct glycosylation profiles. *Nature communications* **11,** 1–12 (2020).
11. Janeway Jr, C. A., Travers, P., Walport, M. & Shlomchik, M. J. The distribution and functions of immunoglobulin isotypes (2001).
12. Vieira, P. & Rajewsky, K. The half-lives of serum immunoglobulins in adult mice. *European journal of immunology* **18,** 313–316 (1988).

13. Hinton, P. R. *et al.* An engineered human IgG1 antibody with longer serum half-life. *The Journal of Immunology* **176,** 346–356 (2006).

14. Palmeira, P., Quinello, C., Silveira-Lessa, A. L., Zago, C. A. & Carneiro-Sampaio, M. *IgG placental transfer in healthy and pathological pregnancies* 2012.

15. Teng, G. & Papavasiliou, F. N. Immunoglobulin somatic hypermutation. *Annu. Rev. Genet.* **41,** 107–120 (2007).

16. Higel, F., Seidl, A., Sörgel, F. & Friess, W. *N-glycosylation heterogeneity and the influence on structure, function and pharmacokinetics of monoclonal antibodies and Fc fusion proteins* 2016.

17. Lefranc, M. *et al.* IMGT, the international ImMunoGeneTics information system. *Nucleic Acids Res.* **37,** D1006–12 (2009).

18. Tonegawa, S. Somatic generation of antibody diversity. *Nature* **302,** 575 (1983).

19. Giudicelli, V., Chaume, D. & Lefranc, M.-P. IMGT/GENE-DB: a comprehensive database for human and mouse immunoglobulin and T cell receptor genes. *Nucleic acids research* **33,** D256–D261 (2005).

20. Xu, J. L. & Davis, M. M. Diversity in the CDR3 region of VH is sufficient for most antibody specificities. *Immunity* **13,** 37–45 (2000).

21. Sela-Culang, I., Kunik, V. & Ofran, Y. The structural basis of antibody-antigen recognition. *Frontiers in immunology* **4,** 302 (2013).

22. Parameswaran, P. *et al.* Convergent antibody signatures in human dengue. *Cell host & microbe* **13,** 691–700 (2013).

23. Galson, J. D. *et al.* BCR repertoire sequencing: different patterns of B-cell activation after two Meningococcal vaccines. *Immunology and cell biology* **93,** 885–895 (2015).

24. D'Angelo, S. *et al.* Many routes to an antibody heavy-chain CDR3: Necessary, yet insufficient, for specific binding. *Frontiers in Immunology* **9** (2018).

25. Kunik, V., Peters, B. & Ofran, Y. Structural consensus among antibodies defines the antigen binding site. *PLoS Comput Biol.* **8,** e1002388 (2012).

26. Bork, P., Holm, L. & Sander, C. The immunoglobulin fold. *J. Mol. Biol* **242,** 309–320 (1994).

27. Choi, Y. & Deane, C. M. FREAD revisited: Accurate loop structure prediction using a database search algorithm. *Proteins: Structure, Function and Bioinformatics* **78,** 1431–1440 (2010).

28. Wang, W., Singh, S., Zeng, D. L., King, K. & Nema, S. *Antibody structure, instability, and formulation* 2007. arXiv: z0024.

29. Glockshuber, R., Schmidt, T. & Plueckthun, A. The disulfide bonds in antibody variable domains: effects on stability, folding in vitro, and functional expression in Escherichia coli. *Biochemistry* **31,** 1270–1279 (1992).

30. Rudikoff, S. & Pumphrey, J. G. Functional antibody lacking a variable-region disulfide bridge. *Proceedings of the National Academy of Sciences* **83,** 7875–7878 (1986).

31. Wörn, A. & Plückthun, A. Mutual stabilization of VL and VH in single-chain antibody fragments, investigated with mutants engineered for stability. *Biochemistry* **37,** 13120–13127 (1998).

32. Proba, K., Honegger, A. & Plückthun, A. A natural antibody missing a cysteine in VH: consequences for thermodynamic stability and folding. *Journal of molecular biology* **265,** 161–172 (1997).

33. Abhinandan, K. & Martin, A. C. Analysis and improvements to Kabat and structurally correct numbering of antibody variable domains. *Molecular immunology* **45,** 3832–3839 (2008).

34. Dunbar, J. & Deane, C. M. ANARCI: antigen receptor numbering and receptor classification. *Bioinformatics* **32,** 298–300 (2016).

35. Lefranc, M.-P. IMGT, the international ImMunoGeneTics database®. *Nucleic acids research* **31,** 307–310 (2003).

36. Lefranc, M.-P. *et al.* IMGT unique numbering for immunoglobulin and T cell receptor variable domains and Ig superfamily V-like domains. *Developmental & Comparative Immunology* **27,** 55–77 (2003).

37. Al-Lazikani, B., Lesk, A. M. & Chothia, C. Standard conformations for the canonical structures of immunoglobulins. *Journal of molecular biology* **273,** 927–948 (1997).

38. Ehrenmann, F., Giudicelli, V., Duroux, P. & Lefranc, M.-P. IMGT/Collier de Perles: IMGT standardized representation of domains (IG, TR, and IgSF variable and constant domains, MH and MhSF groove domains). *Cold Spring Harbor Protocols* **2011,** pdb–prot5635 (2011).

39. Wu, T. T. & Kabat, E. A. An analysis of the sequences of the variable regions of Bence Jones proteins and myeloma light chains and their implications for antibody complementarity. *Journal of Experimental Medicine* **132,** 211–250 (1970).

40. Honegger, A. & PluÈckthun, A. Yet another numbering scheme for immunoglobulin variable domains: an automatic modeling and analysis tool. *Journal of molecular biology* **309,** 657–670 (2001).

41. Leem, J., Dunbar, J., Georges, G., Shi, J. & Deane, C. M. *ABodyBuilder: Automated antibody structure prediction with data–driven accuracy estimation* in *MAbs* **8** (2016), 1259–1268.

42. DeLano, W. L. *et al.* Pymol: An open-source molecular graphics tool. *CCP4 Newsletter on protein crystallography* **40,** 82–92 (2002).

43. Kovaltsuk, A. *et al.* Observed Antibody Space: A Resource for Data Mining Next-Generation Sequencing of Antibody Repertoires. *The Journal of Immunology* **201,** 2502–2509 (Oct. 2018).

44. Kovaltsuk, A., Krawczyk, K., Kelm, S., Snowden, J. & Deane, C. M. Filtering Next-Generation Sequencing of the Ig Gene Repertoire Data Using Antibody Structural Information. *The Journal of Immunology* (2018).

45. Pilzecker, B. & Jacobs, H. Mutating for good: DNA damage responses during somatic hypermutation. *Frontiers in Immunology* **10,** 438 (2019).

46. Chothia, C., Lesk, A. M., *et al.* Canonical structures for the hypervariable regions of immunoglobulins. *Journal of molecular biology* **196,** 901–917 (1987).

47. North, B., Lehmann, A. & Dunbrack Jr, R. L. A new clustering of antibody CDR loop conformations. *Journal of molecular biology* **406,** 228–256 (2011).

48. Nowak, J. *et al.* Length-independent structural similarities enrich the antibody CDR canonical class model. **8,** 751–760 (2016).

49. Wong, W. K. *et al.* SCALOP: Sequence-based antibody canonical loop structure annotation. *Bioinformatics* **35,** 1774–1776 (May 2019).

50. Martin, A. C. & Thornton, J. M. Structural families in loops of homologous proteins: automatic classification, modelling and application to antibodies. *Journal of molecular biology* **263,** 800–815 (1996).

51. Long, X., Jeliazkov, J. R. & Gray, J. J. Non-H3 CDR template selection in antibody modeling through machine learning. *PeerJ* **7,** e6179 (2019).

52. Regep, C., Georges, G., Shi, J., Popovic, B. & Deane, C. M. The H3 loop of antibodies shows unique structural characteristics. *Proteins: Structure, Function and Bioinformatics* **85,** 1311–1318 (2017).

53. Morea, V., Tramontano, A., Rustici, M., Chothia, C. & Lesk, A. M. Conformations of the third hypervariable region in the VH domain of immunoglobulins. *Journal of molecular biology* **275,** 269–294 (1998).

54. Kuroda, D., Shirai, H., Kobori, M. & Nakamura, H. Structural classification of CDR-H3 revisited: a lesson in antibody modeling. *Proteins: Structure, Function, and Bioinformatics* **73,** 608–620 (2008).

55. Weitzner, B. D., Dunbrack Jr, R. L. & Gray, J. J. The origin of CDR H3 structural diversity. *Structure* **23,** 302–311 (2015).

56. Pieper, K., Grimbacher, B. & Eibel, H. B-cell biology and development. *Journal of Allergy and Clinical Immunology* **131,** 959–971 (2013).

57. Cerutti, A., Cols, M. & Puga, I. Marginal zone B cells: virtues of innate-like antibody-producing lymphocytes. *Nature Reviews Immunology* **13,** 118–132 (2013).

58. Yam-Puc, J. C., Zhang, L., Zhang, Y. & Toellner, K.-M. Role of B-cell receptors for B-cell development and antigen-induced differentiation. *F1000Research* **7** (2018).

59. LeBien, T. W. & Tedder, T. F. B lymphocytes: how they develop and function. *Blood* **112,** 1570–1580 (2008).

60. Roco, J. A. *et al.* Class-switch recombination occurs infrequently in germinal centers. *Immunity* **51,** 337–350 (2019).

61. Peled, J. U. *et al.* The biochemistry of somatic hypermutation. *Annu. Rev. Immunol.* **26,** 481–511 (2008).

62. Yaari, G. *et al.* Models of somatic hypermutation targeting and substitution based on synonymous mutations from high-throughput immunoglobulin sequencing data. *Frontiers in immunology* **4,** 358 (2013).

63. Sheng, Z. *et al.* Gene-specific substitution profiles describe the types and frequencies of amino acid changes during antibody somatic hypermutation. *Frontiers in Immunology* **8** (2017).

64. Stavnezer, J., Guikema, J. E. & Schrader, C. E. Mechanism and regulation of class switch recombination. *Annual review of immunology* **26** (2008).

65. Xu, Z., Zan, H., Pone, E. J., Mai, T. & Casali, P. Immunoglobulin class-switch DNA recombination: induction, targeting and beyond. *Nature Reviews Immunology* **12,** 517–531 (2012).

66. Flajnik, M. F. & Kasahara, M. Origin and evolution of the adaptive immune system: genetic events and selective pressures. *Nature Reviews Genetics* **11,** 47–59 (2010).

67. Muyldermans, S. Nanobodies: natural single-domain antibodies. *Annual review of biochemistry* **82,** 775–797 (2013).

68. Li, X. *et al.* Comparative analysis of immune repertoires between bactrian Camel's conventional and heavy-chain antibodies. *PLoS ONE* **11** (2016).

69. Steeland, S., Vandenbroucke, R. E. & Libert, C. Nanobodies as therapeutics: big opportunities for small antibodies. *Drug discovery today* **21,** 1076–1113 (2016).

70. Kaplon, H., Muralidharan, M., Schneider, Z. & Reichert, J. M. *Antibodies to watch in 2020* in *MAbs* **12** (2020), 1703531.

71. Urquhart, L. Top product forecasts for 2020. *Nature reviews. Drug Discovery* **19,** 86–86 (2020).

72. Smith, S. L. Ten years of Orthoclone OKT3 (muromonab-CD3): a review. *Journal of Transplant Coordination* **6,** 109–121 (1996).

73. Raybould, M. I. *et al.* Thera-SAbDab: the Therapeutic Structural Antibody Database. *Nucleic Acids Research* **48,** D383–D388 (2020).

74. Poiron, C. *et al.* IMGT/mAb-DB: the IMGT® database for therapeutic monoclonal antibodies. *Poster no101* **11** (2010).

75. Lu, R.-M. *et al.* Development of therapeutic antibodies for the treatment of diseases. *Journal of biomedical science* **27,** 1–30 (2020).

76. Piccart-Gebhart, M. J. *et al.* Trastuzumab after adjuvant chemotherapy in HER2-positive breast cancer. *New England Journal of Medicine* **353,** 1659–1672 (2005).

77. Köhler, G. & Milstein, C. Continuous cultures of fused cells secreting antibody of predefined specificity. *nature* **256,** 495–497 (1975).

78. Smith, G. P. & Petrenko, V. A. Phage display. *Chemical reviews* **97,** 391–410 (1997).

79. Bazan, J., Całkosiński, I. & Gamian, A. Phage display—A powerful technique for immunotherapy: 1. Introduction and potential of therapeutic applications. *Human vaccines & immunotherapeutics* **8,** 1817–1828 (2012).

80. Cohen, S. B. *et al.* Similar efficacy, safety and immunogenicity of adalimumab biosimilar BI 695501 and Humira reference product in patients with moderately to severely active rheumatoid arthritis: results from the phase III randomised VOLTAIRE-RA equivalence study. *Annals of the rheumatic diseases* **77,** 914–921 (2018).

81. Vaughan, T. J., Osbourn, J. K. & Tempest, P. R. Human antibodies by design. *Nature biotechnology* **16,** 535–539 (1998).

82. Lee, E. C. *et al.* Complete humanization of the mouse immunoglobulin loci enables efficient therapeutic antibody discovery. *Nature Biotechnology* **32,** 356–363 (2014).

83. Wardemann, H. *et al.* Predominant autoantibody production by early human B cell precursors. *Science* **301,** 1374–1377 (2003).

84. Chi, X. *et al.* A neutralizing human antibody binds to the N-terminal domain of the Spike protein of SARS-CoV-2. *Science* **369,** 650–655 (2020).

85. Lavinder, J. J. *et al.* Identification and characterization of the constituent human serum antibodies elicited by vaccination. *Proceedings of the National Academy of Sciences* **111,** 2259–2264 (2014).

86. Miho, E. *et al.* Computational strategies for dissecting the high-dimensional complexity of adaptive immune repertoires. *Frontiers in immunology* **9,** 224 (2018).

87. Parola, C., Neumeier, D. & Reddy, S. T. Integrating high-throughput screening and sequencing for monoclonal antibody discovery and engineering. *Immunology* **153,** 31–41 (2018).

88. Greiff, V. *et al.* A bioinformatic framework for immune repertoire diversity profiling enables detection of immunological status. *Genome medicine* **7,** 1–15 (2015).

89. Galson, J. D. *et al.* B-cell repertoire dynamics after sequential hepatitis B vaccination and evidence for cross-reactive B-cell activation. *Genome medicine* **8,** 1–13 (2016).

90. Weinstein, J. A., Jiang, N., White, R. A., Fisher, D. S. & Quake, S. R. High-throughput sequencing of the zebrafish antibody repertoire. *Science* **324,** 807–810 (2009).

91. Boyd, S. D. *et al.* Measurement and clinical monitoring of human lymphocyte clonality by massively parallel VDJ pyrosequencing. *Science translational medicine* **1,** 12ra23–12ra23 (2009).

92. Luo, C., Tsementzi, D., Kyrpides, N., Read, T. & Konstantinidis, K. T. Direct comparisons of Illumina vs. Roche 454 sequencing technologies on the same microbial community DNA sample. *PloS one* **7,** e30087 (2012).

93. Shao, W. *et al.* Analysis of 454 sequencing error rate, error sources, and artifact recombination for detection of Low-frequency drug resistance mutations in HIV-1 DNA. *Retrovirology* **10,** 1–16 (2013).

94. Greiff, V. *et al.* Learning the High-Dimensional Immunogenomic Features That Predict Public and Private Antibody Repertoires. *The Journal of Immunology* **199,** 2985–2997 (Oct. 2017).

95. Quail, M. A. *et al.* A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. *BMC genomics* **13,** 1–13 (2012).

96. Friedensohn, S., Khan, T. A. & Reddy, S. T. Advanced methodologies in high-throughput sequencing of immune repertoires. *Trends in biotechnology* **35,** 203–214 (2017).

97. Cock, P. J., Fields, C. J., Goto, N., Heuer, M. L. & Rice, P. M. The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. *Nucleic acids research* **38,** 1767–1771 (2009).

98. Singh, M. *et al.* High-throughput targeted long-read single cell sequencing reveals the clonal and transcriptional landscape of lymphocytes. *Nature communications* **10,** 1–13 (2019).

99. Watson, C. T. *et al.* A comparison of immunoglobulin IGHV, IGHD and IGHJ genes in wild-derived and classical inbred mouse strains. *Immunology and cell biology* (2019).

100. DeKosky, B. J. *et al.* High-throughput sequencing of the paired human immunoglobulin heavy and light chain repertoire. *Nature biotechnology* **31,** 166–169 (2013).

101. DeKosky, B. J. *et al.* Large-scale sequence and structural comparisons of human naive and antigen-experienced antibody repertoires. *Proceedings of the National Academy of Sciences of the United States of America* **113,** E2636–E2645 (May 2016).

102. Raybould, M. I. *et al.* Evidence of antibody repertoire functional convergence through public baseline and shared response structures. *BioRxiv* (2020).

103. Goldstein, L. D. *et al.* Massively parallel single-cell B-cell receptor sequencing enables rapid discovery of diverse antigen-reactive antibodies. *bioRxiv,* 660472 (2019).

104. Warren, R. L., Sutton, G. G., Jones, S. J. & Holt, R. A. Assembling millions of short DNA sequences using SSAKE. *Bioinformatics* **23,** 500–501 (2007).

105. DeKosky, B. J. *et al.* In-depth determination and analysis of the human paired heavy-and light-chain antibody repertoire. *Nature medicine* **21,** 86 (2015).

106. Chang, Y.-H. *et al.* Network signatures of IgG immune repertoires in hepatitis B associated chronic infection and vaccination responses. *Scientific reports* **6,** 26556 (2016).

107. Ghraichy, M. *et al.* Maturation of the Human Immunoglobulin Heavy Chain Repertoire With Age. *Frontiers in Immunology* **11,** 1734 (2020).

108. Van Dongen, J. *et al.* Design and standardization of PCR primers and protocols for detection of clonal immunoglobulin and T-cell receptor gene recombinations in suspect lymphoproliferations: report of the BIOMED-2 Concerted Action BMH4-CT98-3936. *Leukemia* **17,** 2257–2317 (2003).

109. Vander Heiden, J. A. *et al.* Dysregulation of B cell repertoire formation in myasthenia gravis patients revealed through deep sequencing. *The Journal of Immunology* **198,** 1460–1473 (2017).

110. Nutt, S. L., Hodgkin, P. D., Tarlinton, D. M. & Corcoran, L. M. The generation of antibody-secreting plasma cells. *Nature Reviews Immunology* **15,** 160–171 (2015).

111. Meng, W. *et al.* An atlas of B-cell clonal distribution in the human body. *Nature Biotechnology* **35,** 879–886 (2017).

112. Vander Heiden, J. A. *et al.* pRESTO: a toolkit for processing high-throughput sequencing raw reads of lymphocyte receptor repertoires. *Bioinformatics* **30,** 1930–1932 (2014).

113. Gupta, N. T. *et al.* Change-O: a toolkit for analyzing large-scale B cell immunoglobulin repertoire sequencing data. *Bioinformatics* **31,** 3356–3358 (2015).

114. Rosenfeld, A. M., Meng, W., Luning Prak, E. T. & Hershberg, U. ImmuneDB: a system for the analysis and exploration of high-throughput adaptive immune receptor sequencing data. *Bioinformatics* **33,** 292–293 (2017).

115. Rosenfeld, A. M., Meng, W., Prak, L., Tjetske, N. & Hershberg, U. ImmuneDB: a novel tool for the analysis, storage, and dissemination of high-throughput immune repertoire sequencing data. *Frontiers in immunology* **9,** 2107 (2018).

116. Breden, F. *et al.* Reproducibility and reuse of adaptive immune receptor repertoire data. *Frontiers in immunology* **8,** 1418 (2017).

117. Vander Heiden, J. A. *et al.* AIRR community standardized representations for annotated immune repertoires. *Frontiers in immunology* **9,** 2206 (2018).

118. Ye, J., Ma, N., Madden, T. L. & Ostell, J. M. IgBLAST: an immunoglobulin variable domain sequence analysis tool. *Nucleic acids research* **41,** W34–W40 (2013).

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

120. Khan, T. A. *et al.* Accurate and predictive antibody repertoire profiling by molecular amplification fingerprinting. *Science advances* **2,** e1501371 (2016).

121. King, H. W. *et al.* Antibody repertoire and gene expression dynamics of diverse human B cell states during affinity maturation. *bioRxiv* (2020).

122. Setliff, I. *et al.* Multi-Donor Longitudinal Antibody Repertoire Sequencing Reveals the Existence of Public Antibody Clonotypes in HIV-1 Infection. *Cell Host and Microbe* (2018).

123. Dunn-Walters, D. K. The ageing human B cell repertoire: a failure of selection? *Clinical & Experimental Immunology* **183,** 50–56 (2016).

124. Bashford-Rogers, R. J. *et al.* Network properties derived from deep sequencing of human B-cell receptor repertoires delineate B-cell populations. *Genome research* **23,** 1874–1884 (2013).

125. Krawczyk, K., Raybould, M. I., Kovaltsuk, A. & Deane, C. M. Looking for therapeutic antibodies in next-generation sequencing repositories. **11,** 1197–1205 (2019).

126. Hoehn, K. B. *et al.* Repertoire-wide phylogenetic models of B cell molecular evolution reveal evolutionary signatures of aging and vaccination. *Proceedings of the National Academy of Sciences* **116,** 22664–22672 (2019).

127. Yaari, G. & Kleinstein, S. H. Practical guidelines for B-cell receptor repertoire sequencing analysis. *Genome medicine* **7,** 121 (2015).

128. Bashford-Rogers, R. J. *et al.* Network properties derived from deep sequencing of human B-cell receptor repertoires delineate B-cell populations. *Genome research* **23,** 1874–1884 (2013).

129. Miho, E., Greiff, V., Reddy, S. T., *et al.* Large-scale network analysis reveals the sequence space architecture of antibody repertoires. *Nature communications* **10,** 1321 (2019).

130. Gadala-Maria, D., Yaari, G., Uduman, M. & Kleinstein, S. H. Automated analysis of high-throughput B-cell sequencing data reveals a high frequency of novel immunoglobulin V gene segment alleles. *Proceedings of the National Academy of Sciences* **112,** E862–E870 (2015).

131. Greiff, V. *et al.* Systems Analysis Reveals High Genetic and Antigen-Driven Predetermination of Antibody Repertoires throughout B Cell Development. *Cell Reports* **19,** 1467–1478 (2017).

132. Zemlin, M. *et al.* Expressed murine and human CDR-H3 intervals of equal length exhibit distinct repertoires that differ in their amino acid composition and predicted range of structures. *Journal of molecular biology* **334,** 733–749 (2003).

133. Bashford-Rogers, R. *et al.* Analysis of the B cell receptor repertoire in six immune-mediated diseases. *Nature,* 1–5 (2019).

134. Dunbar, J. *et al.* SAbDab: the structural antibody database. *Nucleic acids research* **42,** D1140–D1146 (2014).

135. Berman, H., Henrick, K., Nakamura, H. & Markley, J. L. The worldwide Protein Data Bank (wwPDB): ensuring a single, uniform archive of PDB data. *Nucleic acids research* **35,** D301–D303 (2007).

136. Weitzner, B. D. *et al.* Modeling and docking of antibody structures with Rosetta. *Nature protocols* **12,** 401 (2017).

137. Lepore, R., Olimpieri, P. P., Messih, M. A. & Tramontano, A. PIGSPro: prediction of immunoGlobulin structures v2. *Nucleic acids research* **45,** W17–W23 (2017).

138. Marks, C. *et al.* Sphinx: merging knowledge-based and ab initio approaches to improve protein loop prediction. *Bioinformatics* **33,** 1346–1353 (2017).

139. Teplyakov, A. *et al.* Antibody modeling assessment II. Structures and models. *Proteins: Structure, Function and Bioinformatics* **82,** 1563–1582 (2014).

140. Dunbar, J., Fuchs, A., Shi, J. & Deane, C. M. ABangle: characterising the VH–VL orientation in antibodies. *Protein Engineering, Design & Selection* **26,** 611–620 (2013).

141. Zhu, K. *et al.* Antibody structure determination using a combination of homology modeling, energy-based refinement, and loop prediction. *Proteins: Structure, Function, and Bioinformatics* **82,** 1646–1655 (2014).

142. Marze, N. A., Lyskov, S. & Gray, J. J. Improved prediction of antibody VL–VH orientation. *Protein Engineering, Design and Selection* **29,** 409–418 (2016).

143. Bujotzek, A. *et al.* Prediction of VH-VL domain orientation for antibody variable domain modeling. *Proteins: Structure, Function and Bioinformatics* **83,** 681–695 (2015).

144. Marcatili, P., Rosi, A. & Tramontano, A. PIGS: automatic prediction of antibody structures. *Bioinformatics* **24,** 1953–1954 (2008).

145. Deane, C. M. & Blundell, T. L. CODA: a combined algorithm for predicting the structurally variable regions of protein models. *Protein Science* **10,** 599–612 (2001).

146. Hill, J. R., Kelm, S., Shi, J. & Deane, C. M. Environment specific substitution tables improve membrane protein alignment. *Bioinformatics* **27,** i15–i23 (2011).

147. Choi, Y. & Deane, C. M. Predicting antibody complementarity determining region structures without classification. *Molecular BioSystems* **7,** 3327–3334 (2011).

148. Marks, C. & Deane, C. Antibody H3 structure prediction. *Computational and Structural Biotechnology Journal* **15,** 222–231 (2017).

149. Leem, J., Georges, G., Shi, J. & Deane, C. M. Antibody side chain conformations are position-dependent. *Proteins: Structure, Function, and Bioinformatics* **86,** 383–392 (2018).

150. Wang, Q., Canutescu, A. A. & Dunbrack Jr, R. L. SCWRL and MolIDE: computer programs for side-chain conformation prediction and homology modeling. *Nature protocols* **3,** 1832 (2008).

151. Yamashita, K. *et al.* Kotai Antibody Builder: Automated high-resolution structural modeling of antibodies. *Bioinformatics* **30,** 3279–3280 (2014).

152. Schritt, D. *et al.* Repertoire Builder: high-throughput structural modeling of B and T cell receptors. *Molecular Systems Design & Engineering* (2019).

153. Kovaltsuk, A. *et al.* Structural diversity of B-cell receptor repertoires along the B-cell differentiation axis in humans and mice. *PLOS Computational Biology* **16,** e1007636 (2020).

154. Zhu, K. & Day, T. Ab initio structure prediction of the antibody hypervariable H3 loop. *Proteins: Structure, Function, and Bioinformatics* **81,** 1081–1089 (2013).

155. Sircar, A., Kim, E. T. & Gray, J. J. RosettaAntibody: antibody variable region homology modeling server. *Nucleic acids research* **37,** W474–W479 (2009).

156. Jacobson, M. P. *et al.* A hierarchical approach to all-atom protein loop prediction. *Proteins: Structure, Function, and Bioinformatics* **55,** 351–367 (2004).

157. Fasnacht, M. *et al.* Automated antibody structure prediction using Accelrys tools: results and best practices. *Proteins: Structure, Function, and Bioinformatics* **82,** 1583–1598 (2014).

158. Ruffolo, J. A., Guerra, C., Mahajan, S. P., Sulam, J. & Gray, J. J. Geometric Potentials from Deep Learning Improve Prediction of CDR H3 Loop Structures. *bioRxiv* (2020).

159. Krawczyk, K., Liu, X., Baker, T., Shi, J. & Deane, C. M. Improving B-cell epitope prediction and its application to global antibody-antigen docking. *Bioinformatics* **30,** 2288–2294 (2014).

160. Almagro, J. C. *et al.* Antibody modeling assessment. *Proteins: Structure, Function, and Bioinformatics* **79,** 3050–3066 (2011).

161. Almagro, J. C. *et al.* Second Antibody Modeling Assessment (AMA-II). *Proteins: Structure, Function and Bioinformatics* **82,** 1553–1562 (2014).

162. Kuroda, D., Shirai, H., Jacobson, M. P. & Nakamura, H. Computer-aided antibody design. *Protein engineering, design & selection* **25,** 507–522 (2012).

163. Lyskov, S. *et al.* Serverification of molecular modeling applications: the Rosetta Online Server that Includes Everyone (ROSIE). *PloS one* **8,** e63906 (2013).

164. Trück, J. *et al.* Identification of antigen-specific B cell receptor sequences using public repertoire analysis. *The Journal of Immunology* **194,** 252–261 (2015).

165. Richardson, E. *et al.* A computational method for immune repertoire mining that identifies novel binders from different clonotypes, demonstrated by identifying anti-Pertussis toxoid antibodies. *bioRxiv* (2020).

166. Lauer, T. M. *et al.* Developability index: a rapid in silico tool for the screening of antibody aggregation propensity. *Journal of pharmaceutical sciences* **101,** 102–115 (2012).

167. Raybould, M. I. *et al.* Five computational developability guidelines for therapeutic antibody profiling. *Proceedings of the National Academy of Sciences of the United States of America* **116,** 4025–4030 (2019).

168. Galson, J. D., Pollard, A. J., Trück, J. & Kelly, D. F. Studying the antibody repertoire after vaccination: practical applications. *Trends in immunology* **35,** 319–331 (2014).

169. Krawczyk, K. *et al.* Structurally mapping antibody repertoires. *Frontiers in immunology* **9,** 1698 (2018).

170. Kovaltsuk, A. *et al.* How B-cell receptor repertoire sequencing can be enriched with structural antibody data. *Frontiers in Immunology* **8,** 1753 (2017).

171. Scharf, L. *et al.* Structural basis for HIV-1 gp120 recognition by a germ-line version of a broadly neutralizing antibody. *Proceedings of the National Academy of Sciences* **110,** 6049–6054 (2013).

172. Diskin, R. *et al.* Increasing the potency and breadth of an HIV antibody by using structure-based rational design. *Science* **334,** 1289–1293 (2011).

173. Davide, F. R. *et al.* Convergent Antibody Responses to SARS-CoV-2 in Convalescent Individuals. *Nature* (2020).

174. Roskin, K. M. *et al.* Aberrant B cell repertoire selection associated with HIV neutralizing antibody breadth. *Nature immunology* **21,** 199–209 (2020).

175. Galson, J. D. *et al.* In-depth assessment of within-individual and inter-individual variation in the B cell receptor repertoire. *Frontiers in immunology* **6,** 531 (2015).

176. Turchaninova, M. *et al.* High-quality full-length immunoglobulin profiling with unique molecular barcoding. *Nature protocols* **11,** 1599–1616 (2016).

177. Shugay, M. *et al.* Towards error-free profiling of immune repertoires. *Nature methods* **11,** 653–655 (2014).

178. Kuchenbecker, L. *et al.* IMSEQ—a fast and error aware approach to immunogenetic sequence analysis. *Bioinformatics* **31,** 2963–2971 (2015).

179. Yu, Y., Ceredig, R. & Seoighe, C. LymAnalyzer: A tool for comprehensive analysis of next generation sequencing data of T cell receptors and immunoglobulins. *Nucleic Acids Research* **44** (2015).

180. Bolotin, D. A. *et al.* MiXCR: software for comprehensive adaptive immunity profiling. *Nature methods* **12,** 380–381 (2015).

181. Shlemov, A. *et al.* Reconstructing antibody repertoires from error-prone immunosequencing reads. *The Journal of Immunology* **199,** 3369–3380 (2017).

182. Hagihara, Y. & Saerens, D. Engineering disulfide bonds within an antibody. *Biochimica et Biophysica Acta (BBA)-Proteins and Proteomics* **1844,** 2016–2023 (2014).

183. Koenig, P. *et al.* Mutational landscape of antibody variable domains reveals a switch modulating the interdomain conformational dynamics and antigen binding. *Proceedings of the National Academy of Sciences* **114,** E486–E495 (2017).

184. Auffray, C., Sikorav, J., Ollo, R. & Rougeon, F. Correlation between D region structure and antigen-binding specificity: evidences from the comparison of closely related immunoglobulin VH sequences. *Ann Immunol (Paris)* **132,** 77–88 (1981).

185. Galson, J. D. *et al.* Analysis of B cell repertoire dynamics following hepatitis B vaccination in humans, and enrichment of vaccine-specific antibody sequences. *EBioMedicine* **2,** 2070–2079 (2015).

186. Galson, J. D. *et al.* B-cell repertoire dynamics after sequential hepatitis B vaccination and evidence for cross-reactive B-cell activation. *Genome medicine* **8,** 1–13 (2016).

187. Haakenson, J. K., Huang, R. & Smider, V. V. Diversity in the cow ultralong CDR H3 antibody repertoire. *Frontiers in immunology* **9,** 1262 (2018).

188. Proba, K., WoÈrn, A., Honegger, A. & PluÈckthun, A. Antibody scFv fragments without disulfide bonds, made by molecular evolution. *Journal of molecular biology* **275,** 245–253 (1998).

189. Alamyar, E., Giudicelli, V., Duroux, P. & Lefranc, M. IMGT/HighV-QUEST: A high-throughput system and Web portal for the analysis of rearranged nucleotide sequences of antigen receptors-High-throughput version of IMGT/V-QUEST. *V-QUEST 11èmes Journées Ouvertes en Biologie, Informatique et Mathématiques (JOBIM),* 7–9 (2010).

190. Chen, E. C. *et al.* Diverse patterns of antibody variable gene repertoire disruption in patients with amyloid light chain (AL) amyloidosis. *PloS one* **15,** e0235713 (2020).

191. Peled, J. U. *et al.* The biochemistry of somatic hypermutation. *Annu. Rev. Immunol.* **26,** 481–511 (2008).

192. Di Noia, J. M. & Neuberger, M. S. Molecular mechanisms of antibody somatic hypermutation. *Annu. Rev. Biochem.* **76,** 1–22 (2007).

193. Friedensohn, S. *et al.* Synthetic standards combined with error and bias correction improve the accuracy and quantitative resolution of antibody repertoire sequencing in human naıve and memory B cells. *Frontiers in immunology* **9,** 1401 (2018).

194. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *Journal of molecular biology* **215,** 403–410 (1990).

195. Dunbar, J., Knapp, B., Fuchs, A., Shi, J. & Deane, C. M. Examining Variable Domain Orientations in Antigen Receptors Gives Insight into TCR-Like Antibody Design. *PLoS Computational Biology* **10** (2014).

196. Rubelt, F. *et al. Adaptive Immune Receptor Repertoire Community recommendations for sharing immune-repertoire sequencing data* 2017.

197. Agarwala, R. *et al.* Database Resources of the National Center for Biotechnology. *Nucleic Acids Research* **45,** D12–D17 (2017).

198. B cell clonal expansion and convergent antibody responses to SARS-CoV-2. *Cell and host* (2020).

199. Hoehn, K. B. *et al.* Dynamics of immunoglobulin sequence diversity in HIV-1 infected individuals. *Philosophical Transactions of the Royal Society B: Biological Sciences* **370,** 20140241 (2015).

200. Wu, Y.-C. B. *et al.* Influence of seasonal exposure to grass pollen on local and peripheral blood IgE repertoires in patients with allergic rhinitis. *Journal of allergy and clinical immunology* **134,** 604–612 (2014).

201. Bhattacharya, S. *et al.* ImmPort: disseminating data to the public for the future of immunology. *Immunologic research* **58,** 234–239 (2014).

202. Bhattacharya, S. *et al.* ImmPort, toward repurposing of open access immunological assay data for translational and clinical research. *Scientific Data* **5** (2018).

203. Corrie, B. D. *et al.* iReceptor: A platform for querying and analyzing antibody/B-cell and T-cell receptor repertoire data across federated repositories. *Immunological reviews* **284,** 24–41 (2018).

204. Christley, S. *et al.* VDJServer: a cloud-based analysis portal and data commons for immune repertoire sequences and rearrangements. *Frontiers in immunology* **9,** 976 (2018).

205. Widrich, M. *et al.* Modern Hopfield networks and attention for immune repertoire classification. *arXiv preprint arXiv:2007.13505* (2020).

206. Widrich, M. *et al.* DeepRC: Immune repertoire classification with attention-based deep massive multiple instance learning. *bioRxiv* (2020).

207. Wollacott, A. M. *et al.* Quantifying the nativeness of antibody sequences using long short-term memory networks. *Protein Engineering, Design and Selection* **32,** 347–354 (2019).

208. Leinonen, R. *et al.* The European nucleotide archive. *Nucleic acids research* **39,** D28–D31 (2010).

209. Schanz, M. *et al.* High-throughput sequencing of human immunoglobulin variable regions with subtype identification. *PLoS ONE* **9** (2014).

210. Rettig, T. A., Ward, C., Bye, B. A., Pecaut, M. J. & Chapes, S. K. Characterization of the naive murine antibody repertoire using unamplified high-throughput sequencing. *PLoS ONE* **13** (2018).

211. Magoč, T. & Salzberg, S. L. FLASH: fast length adjustment of short reads to improve genome assemblies. *Bioinformatics* **27,** 2957–2963 (2011).

212. Gordon, A., Hannon, G., *et al.* Fastx-toolkit. *FASTQ/A short-reads preprocessing tools (unpublished) http://hannonlab. cshl. edu/fastx_toolkit* **5** (2010).

213. Smith, T. F., Waterman, M. S., *et al.* Identification of common molecular subsequences. *Journal of molecular biology* **147,** 195–197 (1981).

214. Greiff, V. *et al.* Quantitative assessment of the robustness of next-generation sequencing of antibody variable gene repertoires from immunized mice. *BMC immunology* **15,** 40 (2014).

215. Eddy, S. R. *et al.* Multiple alignment using hidden Markov models. **3,** 114–120 (1995).

216. Arbabi-Ghahroudi, M. Camelid single-domain antibodies: historical perspective and future outlook. *Frontiers in immunology* **8,** 1589 (2017).

217. Hoehn, K. B., Lunter, G. & Pybus, O. G. A phylogenetic codon substitution model for antibody lineages. *Genetics* **206,** 417–427 (2017).

218. Shi, Z. *et al.* More than one antibody of individual B cells revealed by single-cell immune profiling. *Cell Discovery* **5,** 1–13 (2019).

219. Jones, T. D. *et al. The INNs and outs of antibody nonproprietary names* in *MAbs* **8** (2016), 1–9.

220. Ota, M. *et al.* Regulation of the B cell receptor repertoire and self-reactivity by BAFF. *The Journal of Immunology* **185,** 4128–4136 (2010).

221. Davide, F. R. *et al.* Convergent Antibody Responses to SARS-CoV-2 in Convalescent Individuals. *Nature* (2020).

222. Alsoussi, W. B. *et al.* A potently neutralizing antibody protects mice against SARS-CoV-2 infection. *The Journal of Immunology* **205,** 915–922 (2020).

223. Eccles, J. D. *et al.* T-bet+ Memory B Cells Link to Local Cross-Reactive IgG upon Human Rhinovirus Infection. *Cell reports* **30,** 351–366 (2020).

224. Setliff, I. *et al.* High-throughput mapping of B cell receptor sequences to antigen specificity. *Cell* **179,** 1636–1646 (2019).

225. Boland, B. S. *et al.* Heterogeneity and clonal relationships of adaptive immune cells in ulcerative colitis revealed by single-cell analyses. *Science immunology* **5** (2020).

226. Adamo, L. *et al.* Myocardial B cells are a subset of circulating lymphocytes with delayed transit through the heart. *JCI insight* **5** (2020).

227. Needleman, S. B. & Wunsch, C. D. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of molecular biology* **48,** 443–453 (1970).

228. Jain, T. *et al.* Biophysical properties of the clinical-stage antibody landscape. *Proceedings of the National Academy of Sciences* **114,** 944–949 (2017).

229. Knappik, A. *et al.* Fully synthetic human combinatorial antibody libraries (HuCAL) based on modular consensus frameworks and CDRs randomized with trinucleotides. *Journal of molecular biology* **296,** 57–86 (2000).

230. De Kruif, J., Boel, E. & Logtenberg, T. Selection and application of human single chain Fv antibody fragments from a semi-synthetic phage antibody display library with designed CDR3 regions. *Journal of molecular biology* **248,** 97–105 (1995).

231. Tsuchiya, Y. & Mizuguchi, K. The diversity of H 3 loops determines the antigen-binding tendencies of antibody CDR loops. *Protein Science* **25,** 815–825 (2016).

232. Abhinandan, K. & Martin, A. C. Analyzing the "degree of humanness" of antibody sequences. *Journal of molecular biology* **369,** 852–862 (2007).

233. Greiff, V., Miho, E., Menzel, U. & Reddy, S. T. Bioinformatic and statistical analysis of adaptive immune repertoires. *Trends in immunology* **36,** 738–749 (2015).

234. Mroczek, E. S. *et al.* Differences in the composition of the human antibody repertoire by B cell subsets in the blood. *Frontiers in immunology* **5,** 96 (2014).

235. Schroeder Jr, H. W. Similarity and divergence in the development and expression of the mouse and human antibody repertoires. *Developmental & Comparative Immunology* **30,** 119–135 (2006).

236. Horns, F., Dekker, C. L. & Quake, S. R. Memory B cell activation, broad anti-influenza antibodies, and bystander activation revealed by single-cell transcriptomics. *Cell Reports* **30,** 905–913 (2020).

237. Gérard, A. *et al.* High-throughput single-cell activity-based screening and sequencing of antibodies using droplet microfluidics. *Nature Biotechnology* **38,** 715–721 (2020).

238. Olimpieri, P. P., Marcatili, P. & Tramontano, A. Tabhu: tools for antibody humanization. *Bioinformatics* **31,** 434–435 (2015).

239. Schmitz, S., Soto, C., Crowe Jr, J. E. & Meiler, J. *Human-likeness of antibody biologics determined by back-translation and comparison with large antibody variable gene repertoires* in *Mabs* **12** (2020), 1758291.

240. Chaudhary, N. & Wesemann, D. R. Analyzing immunoglobulin repertoires. *Frontiers in immunology* **9,** 462 (2018).

241. Marks, C. & Deane, C. M. How repertoire data is changing antibody science. *Journal of Biological Chemistry,* jbc–REV120 (2020).

242. Lunter, G. & Goodson, M. Stampy: a statistical algorithm for sensitive and fast mapping of Illumina sequence reads. *Genome research* **21,** 936–939 (2011).

243. Pedregosa FABIANPEDREGOSA, F. *et al. Scikit-learn: Machine Learning in Python Gaël Varoquaux Bertrand Thirion Vincent Dubourg Alexandre Passos PEDREGOSA, VAROQUAUX, GRAMFORT ET AL. Matthieu Perrot* tech. rep. (2011), 2825–2830. `http://scikit-learn.sourceforge.net.`.

244. Ester, M., Kriegel, H. P., Sander, J. & Xu, X. A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. *Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining,* 226–231. arXiv: `10.1.1.71.1980`.
`https://www.aaai.org/Papers/KDD/1996/KDD96-037.pdf` (1996).

245. Olimpieri, P. P., Chailyan, A., Tramontano, A. & Marcatili, P. Prediction of site-specific interactions in antibody-antigen complexes: the proABC method and server. *Bioinformatics* **29,** 2285–2291 (2013).

246. Yeap, L.-S. *et al.* Sequence-intrinsic mechanisms that target AID mutational outcomes on antibody genes. *Cell* **163,** 1124–1137 (2015).

247. Davis, C. W. *et al.* Longitudinal analysis of the human B cell response to Ebola virus infection. *Cell* **177,** 1566–1582 (2019).

248. Shi, B. *et al.* Comparative analysis of human and mouse immunoglobulin variable heavy regions from IMGT/LIGM-DB with IMGT/HighV-QUEST. *Theoretical Biology and Medical Modelling* **11,** 30 (2014).

249. Corcoran, M. M. *et al.* Production of individualized V gene databases reveals high levels of immunoglobulin genetic diversity. *Nature communications* **7,** 1–14 (2016).

250. Lefranc, M.-P., Giudicelli, V., Regnier, L. & Duroux, P. IMGT, a system and an ontology that bridge biological and computational spheres in bioinformatics. *Briefings in bioinformatics* **9,** 263–275 (2008).

251. Jackson, K. J. *et al.* Human responses to influenza vaccination show seroconversion signatures and convergent antibody rearrangements. *Cell host & microbe* **16,** 105–114 (2014).

252. De Bourcy, C. F. *et al.* Phylogenetic analysis of the human antibody repertoire reveals quantitative signatures of immune senescence and aging. *Proceedings of the National Academy of Sciences* **114,** 1105–1110 (2017).

253. Wu, Y.-C. B., Kipling, D. & Dunn-Walters, D. K. Age-related changes in human peripheral blood IGH repertoire following vaccination. *Frontiers in immunology* **3,** 193 (2012).

254. Martin, V., Wu, Y.-C., Kipling, D. & Dunn-Walters, D. Ageing of the B-cell repertoire. *Philosophical Transactions of the Royal Society B: Biological Sciences* **370,** 20140237 (2015).

255. IJspeert, H. *et al.* Evaluation of the antigen-experienced B-cell receptor repertoire in healthy children and adults. *Frontiers in immunology* **7,** 410 (2016).

256. Jiang, N. *et al.* Lineage structure of the human antibody repertoire in response to influenza vaccination. *Science translational medicine* **5,** 171ra19–171ra19 (2013).

257. Wang, C. *et al.* Effects of aging, cytomegalovirus infection, and EBV infection on human B cell repertoires. *The Journal of Immunology* **192,** 603–611 (2014).

258. Raybould, M. I. J., Wong, W. K. & Deane, C. M. Antibody–antigen complex modelling in the era of immunoglobulin repertoire sequencing. *Molecular Systems Design & Engineering* **4,** 679–688 (2019).

259. Emerson, R. O. *et al.* Immunosequencing identifies signatures of cytomegalovirus exposure history and HLA-mediated effects on the T cell repertoire. *Nature genetics* **49,** 659–665 (2017).

260. Soto, C. *et al.* High Frequency of Shared Clonotypes in Human T Cell Receptor Repertoires. *Cell reports* **32,** 107882 (2020).

261. Gomez-Tourino, I., Kamra, Y., Baptista, R., Lorenc, A. & Peakman, M. T cell receptor β-chains display abnormal shortening and repertoire sharing in type 1 diabetes. *Nature communications* **8,** 1–15 (2017).

262. Rosati, E. *et al.* Overview of methodologies for T-cell receptor repertoire analysis. *BMC biotechnology* **17,** 61 (2017).

263. Wong, W. K., Leem, J. & Deane, C. M. Comparative analysis of the CDR loops of antigen receptors. *Frontiers in immunology* **10,** 2454 (2019).

264. Wong, W. K. *et al.* TCRBuilder: multi-state T-cell receptor structure prediction. *Bioinformatics* **36,** 3580–3581 (2020).

265. Leem, J., de Oliveira, S. H. P., Krawczyk, K. & Deane, C. M. STCRDab: the structural T-cell receptor database. *Nucleic acids research* **46,** D406–D412 (2018).

266. Huang, C. *et al.* Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. *The lancet* **395,** 497–506 (2020).

267. Dong, E., Du, H. & Gardner, L. An interactive web-based dashboard to track COVID-19 in real time. *The Lancet infectious diseases* **20,** 533–534 (2020).

268. Guan, W.-j. *et al.* Clinical characteristics of coronavirus disease 2019 in China. *New England journal of medicine* **382,** 1708–1720 (2020).

269. Golestaneh, L. *et al.* The association of race and COVID-19 mortality. *EClinicalMedicine* **25,** 100455 (2020).

270. Feng, D. *et al.* The SARS epidemic in mainland China: bringing together all epidemiological data. *Tropical Medicine & International Health* **14,** 4–13 (2009).

271. Lu, R. *et al.* Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding. *The Lancet* **395,** 565–574 (2020).

272. Yan, R. *et al.* Structural basis for the recognition of SARS-CoV-2 by full-length human ACE2. *Science* **367,** 1444–1448 (2020).

273. Tay, M. Z., Poh, C. M., Rénia, L., MacAry, P. A. & Ng, L. F. The trinity of COVID-19: immunity, inflammation and intervention. *Nature Reviews Immunology,* 1–12 (2020).

274. Huertas, A. *et al. Endothelial cell dysfunction: a major player in SARS-CoV-2 infection (COVID-19)?* 2020.

275. Raybould, M. I., Kovaltsuk, A., Marks, C. & Deane, C. M. CoV-AbDab: the Coronavirus Antibody Database. *BioRxiv* (2020).

276. Kuri-Cervantes, L. *et al.* Comprehensive mapping of immune perturbations associated with severe COVID-19. *Science Immunology* **5** (2020).

277. Brouwer, P. *et al.* Potent neutralizing antibodies from COVID-19 patients define multiple targets of vulnerability. *bioRxiv* (2020).

278. Schultheiß, C. *et al.* Next-generation sequencing of T and B cell receptor repertoires from COVID-19 patients showed signatures associated with severity of disease. *Immunity* **53,** 442–455 (2020).

279. Edridge, A. W. *et al.* Seasonal coronavirus protective immunity is short-lasting. *Nature Medicine,* 1–3 (2020).

280. Chiara, M., Horner, D. S., Gissi, C. & Pesole, G. Comparative genomics suggests limited variability and similar evolutionary patterns between major clades of SARS-Cov-2. *bioRxiv* (2020).

281. Piccoli, L. *et al.* Mapping neutralizing and immunodominant sites on the SARS-CoV-2 spike receptor-binding domain by structure-guided high-resolution serology. *Cell* (2020).

282. Liao, M. *et al.* Single-cell landscape of bronchoalveolar immune cells in patients with COVID-19. *Nature medicine,* 1–3 (2020).

283. Hu, X. & Liu, X. S. DeepBCR: Deep learning framework for cancer-type classification and binding affinity estimation using B cell receptor repertoires. *bioRxiv,* 731158 (2019).

284. Löffert, D., Ehlich, A., Müller, W. & Rajewsky, K. Surrogate light chain expression is required to establish immunoglobulin heavy chain allelic exclusion during early B cell development. *Immunity* **4,** 133–144 (1996).

285. Sakaguchi, N. & Melchers, F. λ5, a new light-chain-related locus selectively expressed in pre-B lymphocytes. *Nature* **324,** 579–582 (1986).

286. Kudo, A. & Melchers, F. A second gene, VpreB in the lambda 5 locus of the mouse, which appears to be selectively expressed in pre-B lymphocytes. *The EMBO journal* **6,** 2267–2272 (1987).

287. Pleiman, C. M., D'Ambrosio, D. & Cambier, J. C. The B-cell antigen receptor complex: structure and signal transduction. *Immunology today* **15,** 393–399 (1994).

288. Wu, X. *et al.* Focused evolution of HIV-1 neutralizing antibodies revealed by structures and deep sequencing. *Science* **333,** 1593–1602 (2011).

289. Sundling, C. *et al.* Single-cell and deep sequencing of IgG-switched macaque B cells reveal a diverse Ig repertoire following immunization. *The Journal of Immunology* **192,** 3637–3644 (2014).

290. Zhu, J. *et al.* De novo identification of VRC01 class HIV-1–neutralizing antibodies by next-generation sequencing of B-cell transcripts. *Proceedings of the National Academy of Sciences* **110,** E4088–E4097 (2013).

291. Liao, H.-X. *et al.* Co-evolution of a broadly neutralizing HIV-1 antibody and founder virus. *Nature* **496,** 469–476 (2013).

292. Wesemann, D. R. *et al.* Microbial colonization influences early B-lineage development in the gut lamina propria. *Nature* **501,** 112–115 (2013).

293. Zhou, T. *et al.* Multidonor analysis reveals structural elements, genetic determinants, and maturation pathway for HIV-1 neutralization by VRC01-class antibodies. *Immunity* **39,** 245–258 (2013).

294. Menzel, U. *et al.* Comprehensive evaluation and optimization of amplicon library preparation methods for high-throughput antibody sequencing. *PloS one* **9,** e96727 (2014).

295. Palanichamy, A. *et al.* Immunoglobulin class-switched B cells form an active immune axis between CNS and periphery in multiple sclerosis. *Science translational medicine* **6,** 248ra106–248ra106 (2014).

296. Stern, J. N. *et al.* B cells populating the multiple sclerosis brain mature in the draining cervical lymph nodes. *Science translational medicine* **6,** 248ra107–248ra107 (2014).

297. Sundling, C. *et al.* Single-cell and deep sequencing of IgG-switched macaque B cells reveal a diverse Ig repertoire following immunization. *The Journal of Immunology* **192,** 3637–3644 (2014).

298. Collins, A. M., Wang, Y., Roskin, K. M., Marquis, C. P. & Jackson, K. J. The mouse antibody heavy chain repertoire is germline-focused and highly variable between inbred strains. *Philosophical Transactions of the Royal Society B: Biological Sciences* **370,** 20140236 (2015).

299. Lindner, C. *et al.* Diversification of memory B cells drives the continuous adaptation of secretory antibodies to gut microbiota. *Nature Immunology* **16,** 880–888 (2015).

300. Halliley, J. L. *et al.* Long-lived plasma cells are contained within the CD19-CD38hiCD138+ subset in human bone marrow. *Immunity* **43,** 132–145 (2015).

301. Zhou, T. *et al.* Structural repertoire of HIV-1-neutralizing antibodies targeting the CD4 supersite in 14 donors. *Cell* **161,** 1280–1292 (2015).

302. Wu, X. *et al.* Maturation and diversity of the VRC01-antibody lineage over 15 years of chronic HIV-1 infection. *Cell* **161,** 470–485 (2015).

303. Tipton, C. M. *et al.* Diversity, cellular origin and autoreactivity of antibody-secreting cell population expansions in acute systemic lupus erythematosus. *Nature Immunology* **16,** 755–765 (2015).

304. Levin, M. *et al.* Persistence and evolution of allergen-specific IgE repertoires during subcutaneous specific immunotherapy. *Journal of Allergy and Clinical Immunology* **137,** 1535–1544 (2016).

305. Bhiman, J. N. *et al.* Viral variants that initiate and drive maturation of V1V2-directed HIV-1 broadly neutralizing antibodies. *Nature medicine* **21,** 1332–1336 (2015).

306. Doria-Rose, N. A. *et al.* Developmental pathway for potent V1V2-directed HIV-neutralizing antibodies. *Nature* **509,** 55–62 (2014).

307. Huang, J. *et al.* Identification of a CD4-Binding-Site Antibody to HIV that Evolved Near-Pan Neutralization Breadth. *Immunity* **45,** 1108–1121 (2016).

308. Galson, J. D., Trück, J., Kelly, D. F. & Van Der Most, R. Investigating the effect of AS03 adjuvant on the plasma cell repertoire following pH1N1 influenza vaccination. *Scientific Reports* **6** (2016).

309. Ellebedy, A. H. *et al.* Defining antigen-specific plasmablast and memory B cell subsets in human blood after viral infection or vaccination. *Nature Immunology* **17,** 1226–1234 (Sept. 2016).

310. Rubelt, F. *et al.* Individual heritable differences result in unique cell lymphocyte receptor repertoires of naıve and antigen-experienced cells. *Nature communications* **7,** 11112 (2016).

311. Soto, C. *et al.* Developmental pathway of the MPER-directed HIV-1-neutralizing antibody 10E8. *PloS one* **11,** e0157409 (2016).

312. Bonsignori, M. *et al.* Maturation pathway from germline to broad HIV-1 neutralizer of a CD4-mimic antibody. *Cell* **165,** 449–463 (2016).

313. Joyce, M. G. *et al.* Vaccine-induced antibodies that neutralize group 1 and group 2 influenza A viruses. *Cell* **166,** 609–623 (2016).

314. Corcoran, M. M. *et al.* Production of individualized V gene databases reveals high levels of immunoglobulin genetic diversity. *Nature communications* **7,** 1–14 (2016).

315. Cui, A. *et al.* A model of somatic hypermutation targeting in mice based on high-throughput Ig sequencing data. *The Journal of Immunology* **197,** 3566–3574 (2016).

316. Banerjee, S. *et al.* Evaluation of a novel multi-immunogen vaccine strategy for targeting 4E10/10E8 neutralizing epitopes on HIV-1 gp41 membrane proximal external region. *Virology* **505,** 113 (2017).

317. Fisher, C. R. *et al.* T-dependent B cell responses to Plasmodium induce antibodies that form a high-avidity multivalent complex with the circumsporozoite protein. *PLoS Pathogens* **13** (2017).

318. Tong, P. *et al.* IgH isotype-specific B cell receptor expression influences B cell fate. *Proceedings of the National Academy of Sciences* (2017).

319. Levin, M., Levander, F., Palmason, R., Greiff, L. & Ohlin, M. Antibody-encoding repertoires of bone marrow and peripheral blood—a focus on IgE. *Journal of Allergy and Clinical Immunology* **139,** 1026–1030. arXiv: `arXiv:1011.1669v3` (2017).

320. Gupta, N. T. *et al.* Hierarchical clustering can identify B cell clones with high confidence in Ig repertoire sequencing data. *The Journal of Immunology* **198,** 2489–2499 (2017).

321. Landais, E. *et al.* HIV envelope glycoform heterogeneity and localized diversity govern the initiation and maturation of a V2 apex broadly neutralizing antibody lineage. *Immunity* **47,** 990–1003 (2017).

322. Vergani, S. *et al.* Novel method for high-throughput full-length IGHV-DJ sequencing of the immune repertoire from bulk B-cells with single-cell resolution. *Frontiers in immunology* **8,** 1157 (2017).

323. VanDuijn, M. M., Dekker, L. J., Van IJcken, W. F., Sillevis Smitt, P. A. & Luider, T. M. Immune repertoire after immunization as seen by next-generation sequencing and proteomics. *Frontiers in immunology* **8,** 1286 (2017).

324. Eliyahu, S. *et al.* Antibody repertoire analysis of Hepatitis C virus infections identifies immune signatures associated with spontaneous clearance. *Frontiers in immunology* **9,** 3004 (2018).

325. Prohaska, T. A. *et al.* Massively parallel sequencing of peritoneal and splenic B cell repertoires highlights unique properties of B-1 cell antibodies. *The Journal of Immunology* **200,** 1702–1717 (2018).

326. Ohm-Laursen, L. *et al.* Local clonal diversification and dissemination of B lymphocytes in the human bronchial mucosa. *Frontiers in immunology* **9,** 1976 (2018).

327. Waltari, E. *et al.* 5' rapid amplification of cDNA ends and illumina MiSeq reveals B cell receptor features in healthy adults, adults with chronic HIV-1 infection, cord blood, and humanized mice. *Frontiers in immunology* **9,** 628 (2018).

328. Johnson, E. L. *et al.* Sequencing HIV-neutralizing antibody exons and introns reveals detailed aspects of lineage maturation. *Nature communications* **9,** 1–13 (2018).

329. Thörnqvist, L. & Ohlin, M. Data on the nucleotide composition of the first codons encoding the complementary determining region 3 (CDR3) in immunoglobulin heavy chains. *Data in brief* **19,** 337–352 (2018).

330. Gidoni, M. *et al.* Mosaic deletion patterns of the human antibody heavy chain gene locus shown by Bayesian haplotyping. *Nature communications* **10,** 1–14 (2019).

331. Sevy, A. M., Soto, C., Bombardi, R. G., Meiler, J. & Crowe, J. E. Immune repertoire fingerprinting by principal component analysis reveals shared features in subject groups with common exposures. *BMC bioinformatics* **20,** 1–10 (2019).

332. Vázquez Bernat, N. *et al.* High-quality library preparation for NGS-based immunoglobulin germline gene inference and repertoire expression analysis. *Frontiers in immunology* **10,** 660 (2019).

333. Kim, S. I. *et al.* Stereotypic Neutralizing VH Clonotypes Against SARS-CoV-2 RBD in COVID-19 Patients and the Healthy Population. *bioRxiv* (2020).

334. Simonich, C. A. *et al.* Kappa chain maturation helps drive rapid development of an infant HIV-1 broadly neutralizing antibody lineage. *Nature communications* **10,** 1–12 (2019).

335. Bender, S. *et al.* Immunoglobulin variable domain high-throughput sequencing reveals specific novel mutational patterns in POEMS syndrome. *Blood, The Journal of the American Society of Hematology* **135,** 1750–1758 (2020).